Bayesian Regularization in Regression Models for Survival Data

Dissertation

zur Erlangung des akademischen Grades
eines Doktors der Naturwissenschaften
am Institut für Statistik
an der Fakultät für Mathematik, Informatik und Statistik
der Ludwig-Maximilians-Universität München

Eingereicht von
Susanne Konrath
am 16. April 2013
in München

Erstgutachter:	Prof. Dr. Ludwig Fahrmeir, Ludwig-Maximilians-Universität München
Zweitgutachter:	Prof. Dr. Thomas Kneib, Georg-August-Universität Göttingen
Tag der Disputation:	20.06.2013

ZUSAMMENFASSUNG

Diese Arbeit beschäftigt sich mit der Entwicklung flexibler zeitstetiger Überlebenszeitmodelle, die auf dem Accelerated Failure Time (AFT) Modell für die Überlebenszeit und dem Cox Relative Risk (CRR) Modell für die Hazardrate basieren. Die Flexibilisierung betrifft zum einen die Erweiterung des Prädiktors, um gleichzeitig eine Vielfalt von verschiedenartigen Kovariableneffekten zu berücksichtigen. Zum anderen werden die oftmals zu restriktiven parametrischen Annahmen über die Verteilung der Überlebenszeit durch semiparametrische Ansätze ersetzt, die flexiblere Formen der Überlebenszeitverteilung ermöglichen. Wir verwenden die Bayesianische Methodik für die Inferenz. Die auftretenden Probleme, wie zum Beispiel die Penalisierung der hochdimensionalen linearen Effekte, die Glättung nicht linearer Effekte und die Glättung der Basis-Überlebenszeit Verteilung, werden durch Regularisierungs-Prioris gelöst, die für die jeweilige Anforderung speziell angepaßt werden.

Durch die betrachtete Erweiterung der beiden Modellklassen können verschiedene Herausforderungen, die in der praktischen Analyse von Lebensdauerdaten auftreten, bewältigt werden. Beispielsweise können die Modelle mit hochdimensionalen Merkmalsräumen umgehen (z. B. Genexpressionsdaten), sie ermöglichen die Variablenselektion aus der Menge oder einer Teilmenge der verfügbaren Kovariablen und erlauben gleichzeitig die Modellierung irgendeiner Art nicht linearer Effekte für Kovariable, die immer in das Modell eingeschlossen werden sollen. Die Möglichkeit der nichtlinearen Modellierung von Kovariableneffekten, ebenso wie die semiparametrische Modellierung der Überlebenszeitverteilung, ermöglichen darüber hinaus die visuelle Prüfung der Linearitätsannahme für Kovariableneffekte beziehungsweise der parametrischer Annahmen über die Überlebenszeitverteilung.

In dieser Arbeit wird gezeigt, wie das $p > n$ Paradigma, die Relevanz von Untersuchungsmerkmalen, die semiparametrische Inferenz für funktionale Effektformen und die semiparametrische Inferenz für die Überlebenszeitverteilung in einem vereinheitlichten Bayesianischen Rahmen behandelt werden können. Wegen der Möglichkeit, die Stärke der Regularisierung bei den betrachteten Prioris für die linearen Regressionskoeffizienten zu kontrollieren, ist es nicht notwendig, konzeptionell zwischen den Fällen $p \leq n$ und $p > n$ zu unterscheiden. Um die gewünschte Regularisierung durchzuführen, werden die Regressionskoeffizienten mit entsprechenden Schrumpfungs-, Selektions- oder Glättungs-Prioris verbunden. Da die verwendeten Regularisierungs-Prioris alle eine hierarchische Darstellung unterstützen, ermöglicht die resultierende modulare Priori Struktur, in Kombination mit angemessenen Unabhängigkeitsannahmen für die Parameter der Prioris, die Schaffung eines einheitlichen Bayesianischen Rahmens und die Möglichkeit, effiziente MCMC Ziehungsschemen für die gemeinsame Schrumpfung, Selektion oder Glättung in flexiblen Klassen von Lebensdauermodellen zu konstruieren. Die Bayesianische Formulierung ermöglicht somit die gleichzeitige Schätzung aller Modellparameter ebenso wie die Prädiktion und Unsicherheitsaussagen über die Modellspezifizierung.

Die dargelegten Methoden wurden durch den flexiblen und allgemeinen Ansatz der strukturiert additiven Regression (STAR) für Zielvariable aus einer Exponentialfamilie und

Überlebenszeitmodelle vom CRR-Typ angeregt. Derartige systematische und flexible Erweiterungen sind im allgemeinen für AFT Modelle nicht verfügbar. Ein Ziel dieser Arbeit ist, die Klasse der AFT Modelle zu erweitern, um eine ebenso reichhaltige Klasse von Modellen bereitzustellen wie die, die aus den STAR Ansatz resultieren, wobei das Hauptaugenmerk auf der Schrumpfung von linearen Effekten, der Selektion von Kovariablen mit linearen Effekten und der Glättung von nichtlinearen Effekten stetiger Kovariablen, als typischem Bespiel einer nicht-linearen Modellierung, liegt. Im Speziellen werden der Bayesianische Lasso, der Bayesianische Ridge und der Bayesianische NMIG (eine Art Spike-and-Slab Priori) Ansatz zur Regularisierung der linearen Effekte kombiniert mit dem P-Spline Ansatz der die Glättung der nichtlinearen Effekte und der Basiszeitverteilung regularisiert. Um die Fehlerverteilung im AFT Modell flexibel zu gestalten, werden die parametrischen Annahmen über die Basis-Fehlerverteilung durch die Annahme einer endliche Gauss-Mischverteilung ersetzt. Für den Spezialfall der Spezifizierung einer einzigen Mischungskomponente reduziert sich das Schätzproblem auf die Schätzung eines log-normalen AFT Modells mit STAR Prädiktor. Zusätzlich wird die bestehende Klasse von CRR survival Modellen mit STAR Prädiktor, bei der ebenfalls die Basis-Hazardfunktion durch P-Splines approximiert wird, erweitert, um die Regularisierung der linearen Effekte mit den genannten Prioris zu ermöglichen, was den Anwendungsbereich dieser reichhaltigen Klasse von CRR Modellen weiter verbreitert. Schließlich wird der kombinierte Schrumpfungs-, Selektions- und Glättungsansatz auch in das semiparametrische CRR Modell eingeführt, bei dem die Basis-Hazardfunktion unspezifiziert bleibt und die Inferenz auf der Partiellen Likelihood basiert.

Neben der Erweiterung der beiden Überlebenszeit Modellklassen werden die verschiedenen Regularisierungseigenschaften der betrachteten Schrumpfungs- und Selektions-Prioris untersucht. Die entwickelten Methoden und Algorithmen sind in der öffentliche verfügbaren Software `BayesX` und in `R`-Funktionen implementiert und die Leistungsfähigkeit der Methoden und Algorithmen wird umfangreich in Simulationsstudien getestet und anhand von drei realen Datensätzen dargestellt.

ABSTRACT

This thesis is concerned with the development of flexible continuous-time survival models based on the accelerated failure time (AFT) model for the survival time and the Cox relative risk (CRR) model for the hazard rate. The flexibility concerns on the one hand the extension of the predictor to take into account simultaneously for a variety of different forms of covariate effects. On the other hand, the often too restrictive parametric assumptions about the survival distribution are replaced by semiparametric approaches that allow very flexible shapes of survival distribution. We use the Bayesian methodology for inference. The arising problems, like e. g. the penalization of high-dimensional linear covariate effects, the smoothing of nonlinear effects as well as the smoothing of the baseline survival distribution, are solved with the application of regularization priors tailored for the respective demand.

The considered expansion of the two survival model classes enables to deal with various challenges arising in practical analysis of survival data. For example the models can deal with high-dimensional feature spaces (e. g. gene expression data), they facilitate feature selection from the whole set or a subset of the available covariates and enable the simultaneous modeling of any type of nonlinear covariate effects for covariates that should always be included in the model. The option of the nonlinear modeling of covariate effects as well as the semiparametric modeling of the survival time distribution enables furthermore also a visual inspection of the linearity assumptions about the covariate effects or accordingly parametric assumptions about the survival time distribution.

In this thesis it is shown, how the $p > n$ paradigm, feature relevance, semiparametric inference for functional effect forms and the semiparametric inference for the survival distribution can be treated within a unified Bayesian framework. Due the option to control the amount of regularization of the considered priors for the linear regression coefficients, there is no need to distinguish conceptionally between the cases $p \leq n$ and $p > n$. To accomplish the desired regularization, the regression coefficients are associated with shrinkage, selection or smoothing priors. Since the utilized regularization priors all facilitate a hierarchical representation, the resulting modular prior structure, in combination with adequate independence assumptions for the prior parameters, enables to establish a unified framework and the possibility to construct efficient MCMC sampling schemes for joint shrinkage, selection and smoothing in flexible classes of survival models. The Bayesian formulation enables therefore the simultaneous estimation of all parameters involved in the models as well as prediction and uncertainty statements about model specification.

The presented methods are inspired from the flexible and general approach for structured additive regression (STAR) for responses from an exponential family and CRR-type survival models. Such systematic and flexible extensions are in general not available for AFT models. An aim of this work is to extend the class of AFT models in order to provide such a rich class of models as resulting from the STAR approach, where the main focus relies on the shrinkage of linear effects, the selection of covariates with linear effects together with the smoothing of nonlinear effects of continuous covariates as representative of a nonlinear modeling. Combined are in particular the Bayesian lasso, the Bayesian ridge and the Bayesian NMIG (a kind of spike-and-slab prior) approach to regularize the linear effects and the P-spline approach to regularize the smoothness of the nonlinear effects and the baseline survival time distribution. To model a flexible error distribution for the AFT model, the parametric assumption for the baseline error distribution is replaced by the assumption of a finite Gaussian mixture distribution. For the special case of specifying one basis mixture component the estimation problem essentially boils down to estimation of log-normal AFT model with STAR predictor. In addition, the existing class of CRR survival models with STAR predictor, where also baseline hazard rate is approximated by a P-spline, is expanded to enable the regularization of the linear effects with the mentioned priors, which broadens further the area of application of this rich class of CRR models. Finally, the combined shrinkage, selection and smoothing approach is also introduced to the semiparametric version of the CRR model, where the baseline hazard is unspecified and inference is based on the partial likelihood.

Besides the extension of the two survival model classes the different regularization properties of the considered shrinkage and selection priors are examined. The developed methods and algorithms are implemented in the public available software `BayesX` and in `R`-functions and the performance of the methods and algorithms is extensively tested by simulation studies and illustrated through three real world data sets.

ACKNOWLEDGEMENTS

I would like to thank my Ph.D. supervisor, Ludwig Fahrmeir, and Thomas Kneib for this opportunity and their excellent guidance and support.

Further acknowledgements:

To the German Research Foundation (DFG) for the financial support of the research project “Bayesianische Regularisierung in Regressionsmodellen mit hochdimensionalen Prädiktoren”.

To Peter.

CONTENTS

INTRODUCTION

1. Introduction to the basic concepts

1.1. Basic concepts of survival analysis

In continuous-time survival analysis the focus of attention is on a nonnegative random variable T, that is defined as the time to a predefined event, i. e. the duration time, where an individual is under a special unique risk (in contrast to competing risk models) until the interesting event occurs. T is usually called survival or failure time. As a generic example, the risk could be the diagnosis of infection with a deadly disease and the corresponding event is the death of an individual in the study who is infected. The survival time in this example is the duration from the diagnosis until death and as also reflected by this example, the interesting event has only two complementary states, 0 = “event not occurred” and 1 = “event occurred”, and transitions from the state 1 to the state 0 are excluded.

Symptomatic for the collected survival data set is its incompleteness due to the fact that the exact survival time of some individuals is unknown, *censored*, and the only available information is that the event occurred in a certain period of time. A special and most common censoring scheme is the *right censoring*, where an individual’s survival time becomes incomplete at the right side of the observation period, i. e., the only available information is, that the event happens at any time after the follow up. Reasons that hinder the observation of the exact survival time are, for example, that the event doesn’t occur during the end of the finite follow-up period in the study or the individual is lost during the study or withdrawn due to an event that is not of interest (e. g., cured or another competing risk). As a consequence of the censoring summary statistics of survival time distributions, such as the sample mean or the standard error for the mean, do not have desired statistical properties like unbiasedness as an example. Therefore, to accommodate for censoring, numerous methods have been developed for handling these incompletely observed survival times adequately, and survival analysis became a special topic in statistical research with applications in many fields of study like economics, medicine, biology, public health or epidemiology.

There is a great variety of literature devoted to the analysis of survival data. A detailed introduction to survival analysis from a frequentist perspective and description of the possible censoring and truncation schemes can be found, e. g., in Klein and Moeschberger (2003) or Kalbfleisch and Prentice (2002). A powerful tool for a unified, efficiently handling of survival and event history data arises using the counting process representation of the corresponding models, which is exposed, e. g., in Andersen et al. (1993). For a general introduction and overview for full parametric and nonparametric Bayesian approaches for survival models we refer to Ibrahim et al. (2001), who give also a comprehensive review on Bayesian survival analysis.

1.1.1. Survival quantities

Let the absolutely continuous, nonnegative random variable $T \geq 0$ represent the survival time. For simplicity we assume in this subsection that the survival times T_i, $i = 1,...,n$, of all patients follow the same general distribution $T_i \sim T$. Besides the *probability density function* (p.d.f.)

$$f_T(t) = \lim_{\Delta t \to 0+} \frac{1}{\Delta t} \mathbb{P}(t \leq T < t + \Delta t), \quad t \geq 0, \tag{1.1}$$

and the corresponding *cumulative distribution function* (c.d.f.)

$$F_T(t) = \mathbb{P}(T \leq t) = \int_0^t f_T(s) ds, \quad t \geq 0,$$

there are also some other quantities available to describe the probability distribution of the survival time T. In the survival analysis context it is common to use the *survival function*

$$S_T(t) = 1 - F_T(t) = \mathbb{P}(T > t), \quad t \geq 0,$$

which is the probability that an individual will survive till time $t \geq 0$ and the *hazard rate function* $\lambda_T(t) \geq 0$, which is defined by

$$\lambda_T(t) = \lim_{\Delta t \to 0+} \frac{1}{\Delta t} \mathbb{P}(t \leq T < t + \Delta t \mid T \geq t), \quad t \geq 0, \tag{1.2}$$

and interpreted as the instantaneous risk of failure in the interval $[t, t + \Delta t)$, given survival up to time $t \geq 0$. In general the interpretation of the hazard rate as probability is not valid, but for small $\Delta t > 0$ the hazard rate expression $\lambda_T(t)\Delta t$ is approximately the conditional probability of failure in the interval $[t, t + \Delta t)$ given survival up to time t, i. e., $\mathbb{P}(t \leq T < t + \Delta t \mid T \geq t) \approx \lambda_T(t)\Delta t$. Finally the *cumulative hazard function* is given as

$$\Lambda_T(t) = \int_0^t \lambda_T(s) ds, \quad t \geq 0. \tag{1.3}$$

While each of the functions $f_T(t)$, $F_T(t)$, $S_T(t)$, $\lambda_T(t)$ and $\Lambda_T(t)$ illustrate different aspects of the survival distribution, they separately provide mathematically equivalent full specifications of the survival distribution. Therefore, there exist some important one-to-one relationships of these quantities. In particular the connection

$$\lambda_T(t) = \frac{f_T(t)}{S_T(t)}, \tag{1.4}$$

which is derived immediately from the definition of the hazard function and

$$S_T(t) = \exp\left(-\Lambda_T(t)\right) \tag{1.5}$$

are mainly used in the following.

1.1.2. Data structure

To accommodate censoring in the data, in the statistical model and in the methods, an additional positive and continuous random variable $C \geq 0$ is introduced to describe the censoring process, where C_i, $i = 1,...,n$, denote the corresponding potential censoring times of each individual. An individual's *observed survival time* $\tilde{T}_i$ is said to be *right censored* at time $C_i \geq 0$, if the exact value T_i is not known and we only know, that it is greater than or equal to C_i. The observed survival time for each subject in the sample is then given as the minimum of the true survival time and the censoring time

$$\tilde{T}_i = \min\left(T_i, C_i\right), \quad i = 1,...,n \ ,$$

and the so called *censoring indicator*

$$D_i = I\left(C_i \leq T_i\right), \quad i = 1,...,n \, ,$$

reports, if an observation is right censored ($D_i = 0$) or not ($D_i = 1$) using the indicator function $I(\cdot)$ for definition. Beside the survival times there are usually sets of covariates collected, which may have an individual specific influence on the survival times. In summary, the *observed right censored survival data* is represented as

$$\mathfrak{D} = \left\{(\tilde{t}_i, d_i, \mathbf{v}_i'), i = 1,...,n\right\},$$

where $\tilde{t}_i \geq 0$ is the observed survival time, $d_i \in \{0,1\}$ is the censoring indicator and $\mathbf{v}_i = (v_{i,1},...,v_{i,p})'$ is the p-dimensional vector of observed covariates for the n individuals of the sample.

1.1.3. Survival regression models

The distribution over the survival times $T_i \geq 0$ is no longer independent of individual specific characteristics, if additional covariates $\mathbf{v}_i$ are available, where some of them are suspected to have an influence on the individual's survival times. Influential individual specific characteristics cause heterogeneity in the population and require conditioning the survival distribution on the associated parameters, yielding a separate survival distribution for each individual in the sample. Heterogeneity in the population is addressed by the formulation of regression models to describe the functional dependence between the distribution of the survival times and the set of covariates with the task, to build a model that adequately describes the available data in terms of explanation and prediction. We consider two major approaches in continuous-time survival regression, which address different aspects of the survival distribution. For simplicity we take account for linear effects $\boldsymbol{\beta}$ of time-independent covariates $\mathbf{x}_i \subset \mathbf{v}_i$, which build a subset of the observed covariates $\mathbf{v}_i$ in the collected data $\mathfrak{D}$. This assumption is abandoned in the later sections.

Cox relative risk model (CRR model)

A popular survival regression model is the relative risk model of Cox (1972). In contrast to the AFT model, introduced below, the relationship of the covariates and the survival time $T_i \geq 0$ is implicitly defined by the specification of the hazard function as

$$\lambda_i(t \mid \boldsymbol{\beta}, \lambda_0) = \lambda_0(t)\exp(\mathbf{x}_i'\boldsymbol{\beta}), \qquad (1.6)$$

where $\lambda_0(\cdot) \geq 0$ is an unspecified, arbitrary baseline hazard function and $\boldsymbol{\beta} = (\beta_1,...,\beta_{p_x})'$ denotes the p_x-dimensional vector of regression coefficients associated to the time-independent covariates $\mathbf{x}_i = (x_{i1},...,x_{ip_x})' \subset \mathbf{v}_i$. The impact of the covariates is subsumed in the predictor $\eta_i = \eta_i(\boldsymbol{\beta}) = \mathbf{x}_i'\boldsymbol{\beta}$, which acts through the exponential function (to ensure a nonnegative hazard function) as individual specific modifier at the common baseline hazard function in the population. In addition, the model formulation (1.6) separates the effects of the covariates completely from the baseline hazard, i. e. from the underlying baseline survival distribution of the population.

In particular, the factor $\exp(\eta_i) = \exp(\mathbf{x}_i'\boldsymbol{\beta})$, also called *relative risk* , summarizes the effects of different personal characteristics and scales the baseline hazard individual specific, while the single covariate effect $\exp(\beta_k)$ corresponds to the unit change of the hazard function with respect to a unit change in the covariate x_{ik}. The famous property of the CRR model arises, when the hazard rate ratio of two individuals with covariates $\mathbf{x}_i, \mathbf{x}_j$, $i \neq j$ is considered

$$\frac{\lambda_i(t \mid \boldsymbol{\beta}, \lambda_0)}{\lambda_j(t \mid \boldsymbol{\beta}, \lambda_0)} = \exp\left(\eta_i - \eta_j\right) = \exp\left((\mathbf{x}_i - \mathbf{x}_j)'\boldsymbol{\beta}\right).$$

For time-independent covariates the hazard ratio is constant for any two covariate combinations leading to *proportional hazard rates*. The crucial, rather strong property, that the *hazard rate functions* of different individuals *can not cross*, must be seriously verified to hold in practice. Another special and remarkable feature of the CRR model is the presence of the *partial likelihood*, compare Subsection 1.1.4, which enables suitable likelihood inference for the regression coefficients without the need to specify baseline hazard function. A possible parametric specification for inference arises from Weibull regression model, where the hazard is given by

$$\lambda_i(t \mid \alpha, \boldsymbol{\beta}) = \alpha t^{\alpha-1} \exp(\mathbf{x}_i'\boldsymbol{\beta}),$$

with shape parameter $\alpha > 0$. The Weibull model is adequate, if the baseline hazard $\lambda_0(t) = \alpha t^{\alpha-1}$ is assumed to be monotone increasing ($\alpha > 1$), monotone decreasing ($\alpha < 1$) or constant ($\alpha = 1$), where in the latter case the survival times have an exponential distribution $T_i \sim \text{Exp}(\mathbf{x}_i'\boldsymbol{\beta})$.

Accelerated failure time model (AFT model)

A regression model that specifies the direct impact of the covariates on the survival time $T_i \geq 0$ is the accelerated failure time model, also introduced by Cox (1972). The functional relationship in the AFT model is described by

$$T_i = T_{0,i} \exp(\mathbf{x}_i'\boldsymbol{\beta}), \tag{1.7}$$

where $T_{0,i} \geq 0$ are covariate independent baseline survival times and $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_{p_x})'$ is a p_x-dimensional vector of regression parameters that represents the linear effects of time-independent covariates $\mathbf{x}_i = (x_{i1}, \ldots, x_{ip_x})' \subset \mathbf{v}_i$. In contrast to the CRR model the predictor $\eta_i = \eta_i(\boldsymbol{\beta}) = \mathbf{x}_i'\boldsymbol{\beta}$ determines the so-called *acceleration factor* $\exp(\eta_i) = \exp(\mathbf{x}_i'\boldsymbol{\beta})$ for the baseline survival time $T_{0,i}$, where a negative value of the predictor $\eta_i = \mathbf{x}_i'\boldsymbol{\beta} < 0$ causes an acceleration and a positive value $\eta_i = \mathbf{x}_i'\boldsymbol{\beta} > 0$ a deceleration (= negative acceleration) of the baseline survival time $T_{0,i}$. Through the exponential link function in (1.7) each single covariate causes a multiplicative change of $T_{0,i}$ and in particular $\exp(\beta_j)$ reflects the unit change of the survival time T_i with respect to a unit change in the covariate x_j. The baseline survival times $T_{0,i} \geq 0$ can be interpreted as the individual lifespan if $\mathbf{x}_i = \mathbf{0}$, but in general the baseline survival time T_0 is an unobservable model component. For parametric inference the baseline survival times $T_{0,i}$, $i = 1, \ldots, n$, are assumed to be independent and identical distributed (i.i.d.) with subject to the baseline survival time distribution of $T_0 \geq 0$ in the population. Under this assumption, the ratio of the mean survival times of two individuals with observed covariates $\mathbf{x}_i, \mathbf{x}_j$, $i \neq j$,

$$\frac{\mathbb{E}(T_i \mid \boldsymbol{\beta})}{\mathbb{E}(T_j \mid \boldsymbol{\beta})} = \exp(\eta_i - \eta_j) = \exp\big((\mathbf{x}_i - \mathbf{x}_j)'\boldsymbol{\beta}\big),$$

is constant, for any two time independent covariate combinations leading to *proportional changes of the survival time means*, and especially $\exp(\beta_k)$ quantifies this proportion with respect to a unit change in the covariate x_{ik} compared to x_{jk}. The generic form of the hazard rate function is given by

$$\lambda_i\big(t \mid \boldsymbol{\beta}\big) = \lambda_0\big(t \cdot \exp(-\mathbf{x}_i'\boldsymbol{\beta})\big)\exp(-\mathbf{x}_i'\boldsymbol{\beta}), \tag{1.8}$$

where $\lambda_0(\cdot) \geq 0$ denotes the baseline hazard function that describes the covariate independent baseline survival time distribution. In contrast to the hazard function in the CRR model the covariates affect also the baseline hazard $\lambda_0(\cdot)$ and **Figure 1.1** visualizes the different impact of a binary covariate on the baseline hazard in the CRR and AFT model.

An alternative and often used representation of the AFT model is obtained, when the logarithmic transformation is applied to (1.7). On the log-scale the AFT model gets an additive structure

$$Y_i := \log(T_i) = \mathbf{x}_i'\boldsymbol{\beta} + \sigma\varepsilon_i, \tag{1.9}$$

that is much closer to conventional regression models with response $Y_i := \log(T_i)$ and random baseline error term $\sigma\varepsilon_i := \log(T_{0,i})$. The interpretation of the covariate effects in the log-linear version of the AFT model is straightforward in terms of Y_i. The random baseline error term is further decomposed in a fixed scale factor $\sigma > 0$ and random error terms $\varepsilon_i \in \mathbb{R}$ which are assumed to be i.i.d. with density $f_\varepsilon(\cdot)$. In the later sections we use the definition $Y_0 := \beta_0 + \sigma\varepsilon$, including the common intercept β_0, to describe the common baseline error distribution of the population.

The error ε_i is often assumed to have a density from a standard location-scale family, where the location parameter is equal to zero and scale parameter is equal to one. Since the location-scale distribution family is invariant for linear (affine) transformations, the location parameter of the log-survival time Y_i in (1.9) is modeled by the predictor $\eta_i = \mathbf{x}_i'\boldsymbol{\beta}$ and the scale parameter is given by $\sigma > 0$. Using for example i.i.d. baseline errors $\varepsilon_i \sim N(0,1)$ from the standard Gaussian distribution in the log-linear representation, the log-survival times $Y_i \mid \boldsymbol{\beta}, \sigma$ also have a Gaussian distribution, where the location parameter $\eta_i = \mathbf{x}_i'\boldsymbol{\beta}$ determines the mean and the scale parameter $\sigma > 0$ determines the standard deviation. On the associated time-scale we get a lognormal distribution for the survival times, $T_i \mid \boldsymbol{\beta}, \sigma \sim \text{LogN}(\mathbf{x}_i'\boldsymbol{\beta}, \sigma)$, with shape parameter σ and scale parameter $\eta_i = \mathbf{x}_i'\boldsymbol{\beta}$. Using alternatively i.i.d. baseline errors from a standard extreme value distribution with density $f_\varepsilon(\varepsilon) = \exp(\varepsilon - \exp(\varepsilon))$, the popular and widely used Weibull regression is obtained. The resulting distribution of the log-survival times $Y_i \mid \boldsymbol{\beta}, \sigma$ is also an extreme value distribution, where the location parameter $\eta_i = \mathbf{x}_i'\boldsymbol{\beta}$ corresponds to the mode. Returning to time-scale, the associated survival times have a Weibull distribution, $T_i \mid \boldsymbol{\beta}, \sigma \sim \text{WB}(\alpha, \lambda)$, with shape parameter $\alpha = 1/\sigma$, scale parameter $\lambda = \exp(-\mathbf{x}_i'\boldsymbol{\beta}/\sigma)$ and the hazard function

$$\lambda_i(t \mid \boldsymbol{\beta}, \sigma) = \frac{1}{\sigma} t^{\frac{1}{\sigma}-1} \exp\big(-\mathbf{x}_i'\boldsymbol{\beta}\big)^{\frac{1}{\sigma}}.$$

The unique feature of the Weibull regression model is that it can either be viewed as special case of the AFT model or as special case of the CRR model. Note that in the CRR context the Weibull regression model has another parameterization as in the AFT context with the one-to-one connections $\alpha = \sigma^{-1}$ and $\beta_{j,\text{CRR}} = -\sigma^{-1}\beta_{j,\text{AFT}}$. To simplify the notation and disburden the common treatment of the

AFT and CRR model in this introductory section, the scale parameter $\sigma > 0$ of the AFT model is from now on assumed to be known, so that $\boldsymbol{\beta}$ is the parameter of primary interest. Inference for the scale parameter is outlined in the subsequent sections.

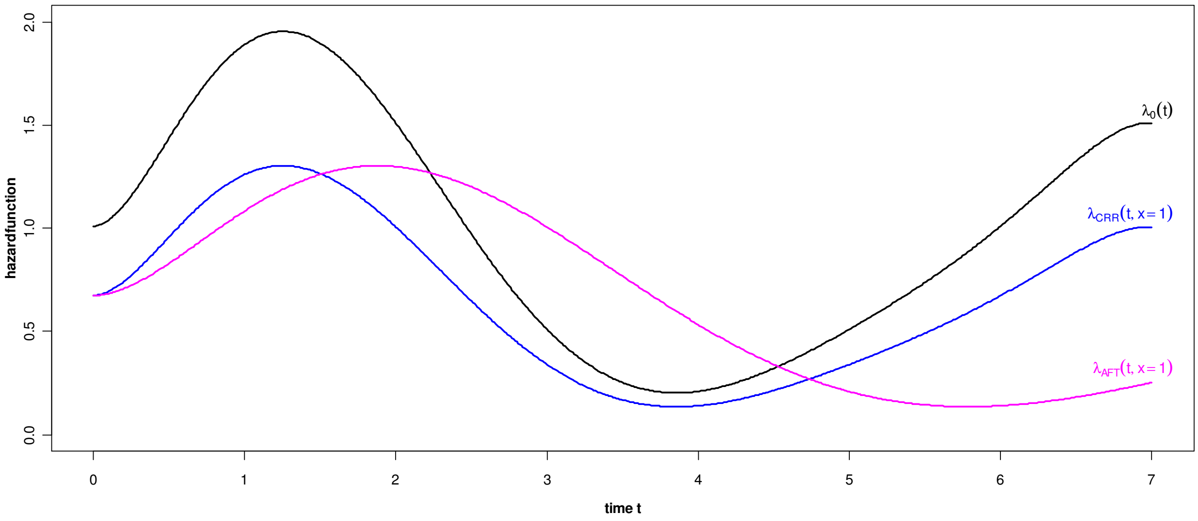


Figure 1.1: Hazard functions of the AFT and CRR model. The baseline hazard $\lambda_0(t)$ (black line) corresponds to $x = 0$ and is compared to the hazard function for $x = 1$ under the CRR model (blue line) with $\beta = 2/3$ and an AFT model (magenta line) with $\beta = -2/3$.

This subsection is concluded with the short remark that also semiparametric versions of the AFT model, with unspecified baseline survival times/errors, can be considered for inference of the regression parameters $\boldsymbol{\beta}$, similar to the semiparametric version of the CRR model. The methods are based on censored rank statistics and there is a lot of literature dealing with the development of these statistics and adequate inferential methods, so that meanwhile the AFT model can also be viewed as practical semiparametric alternative to the CRR model, even in the context with time-dependent covariates. However, the methods are numerical challenging and computationally intensive and there is no inferential pendant to the *partial likelihood* of the CRR model. We refer to Kalbfleisch and Prentice (2002) for a comprehensive treatment of parametric and nonparametric parametric AFT models and to Wei (1992) for a review of inference procedures for nonparametric models in the frequentist setting.

1.1.4. Likelihood structure

Full likelihood

For estimation of parametric survival regression models it becomes also necessary to model explicitly the introduced censoring mechanism. In general the censoring time $C_i \geq 0$ is treated as a survival time, where the interesting event is the censoring and $C_i \mid \boldsymbol{\psi}$ denotes the corresponding distribution which depends on a set of parameters $\boldsymbol{\psi} \in \Psi$. The survival distribution $T_i \mid \boldsymbol{\theta}$ is assumed to depend on the parameters $\boldsymbol{\theta} \in \Theta$ which are the parameters of main interest. For simplicity we can think about $\boldsymbol{\theta} \equiv \boldsymbol{\beta}$, but in parametric models, like e. g. the Weibull model, we have generally more parameters of interest, i. e. $\boldsymbol{\theta} = (\boldsymbol{\beta}', \alpha)'$. To derive an adequate likelihood, further assumptions are useful to simplify the likelihood structure of a survival regression model. Often the censoring process is assumed to be *noninformative*, so that the distribution of the lifetimes $T_i \,|\, \boldsymbol{\theta}$ and censoring times $C_i \,|\, \boldsymbol{\psi}$ of each individual $i = 1,\dots,n$, do not share common parameters of interest, i. e. the intersection of $\boldsymbol{\theta}$ and $\boldsymbol{\psi}$ is empty. Further assumptions are that given the covariates the lifetimes $T_i \,|\, \boldsymbol{\theta}$ are conditional independent, the censoring times $C_i \,|\, \boldsymbol{\psi}$ are conditional independent and the lifetimes and censoring times are conditional independent of each other. At least the likelihood contribution for a possibly right censored observation is given by the joint distribution of the observable quantities $\tilde{T}_i$ and D_i as

$$L_i\left(\boldsymbol{\theta},\boldsymbol{\psi} \,|\, \mathfrak{D}\right) = f_{\tilde{T}_i,D_i}\left(\tilde{t}_i, d_i \,|\, \boldsymbol{\theta},\boldsymbol{\psi}\right) = \left(f_{T_i}(\tilde{t}_i \,|\, \boldsymbol{\theta})^{d_i} S_{T_i}(\tilde{t}_i \,|\, \boldsymbol{\theta})^{1-d_i}\right)\left(f_{C_i}(\tilde{t}_i \,|\, \boldsymbol{\psi})^{1-d_i} S_{C_i}(\tilde{t}_i \,|\, \boldsymbol{\psi})^{d_i}\right). \qquad (1.10)$$

The noninformative censoring induces that the components concerning the censoring process act as constants and can be neglected if the focus relies on $\boldsymbol{\theta}$. Finally, the likelihood contribution for the observation of a true survival time ($\tilde{T}_i = T_i$) given the data $\mathfrak{D}$ is simply $L_i(\boldsymbol{\theta} \,|\, \mathfrak{D}) = f_i(\tilde{t}_i \,|\, \boldsymbol{\theta})$ and for a right censored observation ($\tilde{T}_i = C_i$) the contribution is given by $L_i(\boldsymbol{\theta} \,|\, \mathfrak{D}) = \mathbb{P}(T_i > \tilde{t}_i \,|\, \boldsymbol{\theta}) = S_i(\tilde{t}_i \,|\, \boldsymbol{\theta})$. In summary, the full likelihood for right censored survival data is represented by

$$L\left(\boldsymbol{\theta} \,|\, \mathfrak{D}\right) = \prod_{i=1}^{n} L_i\left(\boldsymbol{\theta} \,|\, \mathfrak{D}\right) = \prod_{i=1}^{n} f_i\left(\tilde{t}_i \,|\, \boldsymbol{\theta}\right)^{d_i} S_i\left(\tilde{t}_i \,|\, \boldsymbol{\theta}\right)^{1-d_i}. \qquad (1.11)$$

In terms of the hazard function using the relationships (1.4) and (1.5) the full likelihood is expressed as

$$L\left(\boldsymbol{\theta} \,|\, \mathfrak{D}\right) = \prod_{i=1}^{n} L_i\left(\boldsymbol{\theta} \,|\, \mathfrak{D}\right) = \prod_{i=1}^{n} \lambda_i\left(\tilde{t}_i \,|\, \boldsymbol{\theta}\right)^{d_i} \exp\left(-\Lambda_i\left(\tilde{t}_i \,|\, \boldsymbol{\theta}\right)\right). \qquad (1.12)$$

If there is sufficient evidence for a parametric specification of the survival distribution, maximum likelihood methods based on the full likelihood can be used to estimate the model parameters $\boldsymbol{\theta} \in \Theta$ leading to usual properties like asymptotic normality and unbiasedness of the estimates.

Partial Likelihood

As proposed in Cox (1972) and further discussed in Cox (1975), the inference of the regression coefficients $\boldsymbol{\beta}$ in the semiparametric CRR model (1.6) can be carried out in terms of the *partial likelihood*

$$pL\left(\boldsymbol{\beta} \,|\, \mathfrak{D}\right) = \prod_{i=1}^{n} \left\{ \frac{\exp\left(\mathbf{x}_i'\boldsymbol{\beta}\right)}{\sum_{k=1}^{n} 1_{(\tilde{t}_k \geq \tilde{t}_i)} \exp\left(\mathbf{x}_k'\boldsymbol{\beta}\right)} \right\}^{d_i}. \qquad (1.13)$$

The indicator function $1_{(\tilde{t}_k \geq \tilde{t}_i)}$ in the denominator is used to describe the *risk set* $R(\tilde{t}_i) = \{k : \tilde{t}_k \geq \tilde{t}_i\}$ at the observed survival time $\tilde{t}_i$, which consists of all individuals who are event-free and still under observation just prior to time $\tilde{t}_i$. In contrast to the full likelihood, e. g. (1.11), there is no separate contribution to the partial likelihood for a censored observation $d_i = 0$ and information from censored individuals enters the likelihood only via the risk set. To practice the estimation, the partial likelihood is treated as a usual likelihood function and the maximum partial likelihood estimator of $\boldsymbol{\beta}$ is shown to be consistent and asymptotically normal, compare, e. g., Andersen and Gill (1982). The estimation in the CRR model is often continued by the estimation of the cumulative baseline hazard function in terms of the Breslow estimate $\hat{\Lambda}_o^{BR}(\cdot)$, Breslow (1972, 1974), which is given by the step function

$$\hat{\Lambda}_0^{Br}(t) = \sum_{i=1}^{n} \frac{1_{(\tilde{t}_i \leq t)}\, d_i}{\sum_{k=1}^{n} 1_{(\tilde{t}_k \geq \tilde{t}_i)} \exp\left(\mathbf{x}_k'\hat{\boldsymbol{\beta}}\right)} \qquad (1.14)$$

and depends on the estimator $\hat{\boldsymbol{\beta}}$ from the maximization of the partial likelihood. The Breslow estimator can be jointly derived with the partial likelihood from a profile likelihood approach, assuming a piecewise constant baseline hazard between two consecutive distinct uncensored failure times, compare, e. g., Breslow (1972, 1974), Murphy and Van der Vaart (2000) or van Houwelingen et al. (2006) for details. The asymptotic properties of this estimator were also established by Andersen and Gill (1982).

Since the partial likelihood only depends on the observed order, not on the exact values of the failure times, corrections are required if *ties* (identical survival times) are present to take account for the permutation of those individuals with identical survival times, because if more than one individual has its event at the same time, the ordering is no longer unique. For a moderate number of ties among the uncensored survival times, so that the use of the continuous time Cox model is still justified, there are several suggestions to approximate the partial likelihood, compare, e. g., Therneau and Grambsch (2000), Klein and Moeschberger (2003). The correction proposed by Breslow (1972, 1974) arises naturally from the profile likelihood approach by treating the tied observations at a given time as distinct contributions to the likelihood, and in particular the formulation of the partial likelihood in (1.13) results in the Breslow correction in the presence of ties. The partial likelihood approach can also be applied for extensions of the Cox model, e. g., with nonlinear covariate effects, Sleeper and Harrington (1990), Gray (1992), time-varying effects, Verweij and van Houwelingen (1995), frailties, Therneau and Grambsch (2000), or time-varying covariates, Klein and Moeschberger (2003).

1.1.5. Bayesian Inference

An alternative concept to the likelihood inference is the Bayesian inference. Bayesian inference relies on the posterior distribution of the model parameters $\boldsymbol{\theta} \in \Theta$ given the observed data $\mathfrak{D}$ and the operational core is the Bayes theorem, where the density of the posterior distribution $p(\boldsymbol{\theta} \mid \mathfrak{D})$ is defined as

$$p(\boldsymbol{\theta} \mid \mathfrak{D}) = \frac{L(\boldsymbol{\theta} \mid \mathfrak{D}) p(\boldsymbol{\theta})}{\int_\Theta L(\boldsymbol{\theta} \mid \mathfrak{D}) p(\boldsymbol{\theta}) d\boldsymbol{\theta}} \propto L(\boldsymbol{\theta} \mid \mathfrak{D}) p(\boldsymbol{\theta}) . \tag{1.15}$$

The posterior distribution is expressed in terms the prior density $p(\boldsymbol{\theta})$, which represents the prior knowledge of the complete set of model parameters $\boldsymbol{\theta} \in \Theta$ and the likelihood $L(\boldsymbol{\theta} \mid \mathfrak{D})$, that may also depend only on a subset of $\boldsymbol{\theta}$. The so called marginal likelihood in the denominator does not depend on model parameters and acts as normalization constant of the posterior density. This causes the annotated proportionality of the posterior density to the product of the prior density and the likelihood.

For posterior maximization, the normalizing constant is negligible, and finding the mode of the posterior density is equivalent to the maximization of the right hand side of (1.15). The corresponding optimization problem has the general form

$$\hat{\boldsymbol{\theta}} = \arg\max_{\boldsymbol{\theta}} \{\log p(\boldsymbol{\theta} \mid \mathfrak{D})\} = \arg\max_{\boldsymbol{\theta}} \{\log L(\mathfrak{D} \mid \boldsymbol{\theta}) + \log p(\boldsymbol{\theta})\} . \tag{1.16}$$

and $\hat{\boldsymbol{\theta}}$ denotes the maximum a posteriori (MAP) estimate. If weakly informative priors are used, the prior term acts like a proportionality constant and the Bayesian optimization problem for finding the mode is equivalent to the optimization of log-likelihood, and hence the posterior mode estimate coincides with the maximum likelihood estimate of $\boldsymbol{\theta} \in \Theta$. Despite this interesting connection to the likelihood inference, the posterior mode is not in general the unique or best choice to obtain a

Bayesian point estimate. In the Bayesian-risk sense, for example under the squared error loss function, the optimal choice for a point estimate of the regression parameters is given by the posterior mean instead of the posterior mode. However, Bayesian inference is rather based on the access to whole posterior distribution than just finding its mode and under a full Bayesian approach, the evaluation of the posterior provides a probabilistic basis to consider the uncertainty of a model.

In practice the entailed integral calculations to evaluate the normalizing constant in the denominator (1.15) are often not feasible and as a consequence the posterior density has no closed analytical form. In such situations the posterior can be explored by generating samples from the posterior distribution by Markov Chain Monte Carlo (MCMC) techniques. The main goal of MCMC methods is to generate (dependent) samples $\boldsymbol{\theta}^{(s)}$, $s = 1,2,...,S$, from a given distribution, in particular the posterior distribution. By utilizing MCMC integration, e. g., with

$$\int_{\Theta} g(\boldsymbol{\theta}) p(\boldsymbol{\theta} \mid \mathfrak{D}) d\boldsymbol{\theta} = E_{p(\boldsymbol{\theta}|\mathfrak{D})}\left(g(\boldsymbol{\theta})\right) \approx \frac{1}{K}\sum_{k=1}^{K} g(\boldsymbol{\theta}^{(k)}),$$

it is possible to approximate the mean of a functional $g(\cdot)$ using the generated sample of the parameters $\boldsymbol{\theta}^{(1)},...,\boldsymbol{\theta}^{(S)}$. For example, the posterior mean of $\boldsymbol{\theta}$ is estimated using the identity function $g(\cdot) = id(\cdot)$. Uncertainty about the model parameters is considered by the corresponding empirical counterparts of the standard deviation or credible intervals. We refer at this point, e. g., to Gelman et al. (2004) or Gilks et al. (1996) for a detailed illustration of the basic concepts of Bayesian analysis and posterior inference based on MCMC methods and in the special context of survival analysis to Ibrahim et al. (2001). Bayesian analysis of the CRR model has also been studied in terms of the partial likelihood, where the full likelihood $L(\boldsymbol{\theta} \mid \mathfrak{D})$ in (1.15) is replaced by the partial likelihood $pL(\boldsymbol{\theta} \mid \mathfrak{D})$. This approach is often justified by showing that the posterior, based on the partial likelihood, approximates the full marginal posterior of the regression coefficients with a very diffuse prior on the cumulative baseline hazard function. We sketch the idea in Section 7.4.2 and refer for details to Kalbfleisch (1978), Sinha et al. (2003) and Kim and Kim (2009).

1.2. Basic concepts of regularization

Regularized estimation approaches have emerged as a general tool to address different problems in applied regression analysis like shrinkage of highly correlated covariate effects to uniquely solve underdetermined estimation equation systems, selection of important covariates from the set of available covariates or for smoothing of nonlinear effects to reflect a more complex influence of the covariates. As an example consider gene expression data. With today's analytical methods, thousands of genes can be analyzed simultaneously for any given patient, but acquisition of suitable patients is often difficult and time consuming and so sparse data sets arise with huge feature spaces, but only very few data points. One of the resulting problems is to compensate *identification problems* of an estimator, if a lot of parameters have to be estimated and/or heavy correlations inducing multicollinearity are present. In such situations the estimation equation system is often underdetermined and as a consequence, there is no unique solution available and the optimization procedure becomes numerically unstable. Regularization is used to find unique solutions by introducing additional constraints supporting the identification of the regression parameters. Also the prediction can be enhanced by constructing estimators with a little bit of bias to obtain a smaller variance, known e. g. from ridge regression. Another goal is the separation of influential variables and nuisance covariates that are not associated with the response. Also *variable selection*, as a form of model selection in which the class of considered models is represented by subsets of the available covariates in the data, becomes an important task especially in high-dimensional feature spaces, where a lot of covariates are suspected to be rather unimportant. To answer questions concerning the relevance of individual features, regularization methods are utilized that shrink the regression coefficient estimates toward zero and simultaneously enforce some coefficients to be set equal to zero, which are then interpreted as unimportant nuisance variables. A prominent representative is given by the lasso regression, Tibshirani (1996). Beside the gene expressions often additional patient specific characteristics, like age or weight, are available, and we want to enable more flexible shapes to reflect the impact of such covariates on the survival time. In modeling nonlinear effects, *smoothness* penalties have a long tradition in semiparametric regression, with smoothing splines and penalized polynomial splines as the most prominent examples, see Wood (2006) or Ruppert et al. (2003) for overviews. In this case, the penalty represents a roughness measure for unknown functions that avoids overfitting induced by overly flexible function estimates.

1.2.1. Frequentist regularization

In summary, the general idea of regularized regression relies on the incorporation of additional assumptions about the model parameters into the estimation problem. In practice, a *penalty term* is added to the estimation function to enforce that the solutions are determined with respect to these constraints. The resulting optimization problem is reflected by the *penalized (log-) likelihood*

$$\log \mathrm{L}_{\mathrm{pen}}(\boldsymbol{\beta},\lambda) = \log \mathrm{L}(\boldsymbol{\beta} \mid \mathfrak{D}) - \mathrm{pen}(\boldsymbol{\beta};\lambda), \tag{1.17}$$

where $\log \mathrm{L}(\boldsymbol{\beta} \mid \mathfrak{D})$ denotes the logarithm of the model specific likelihood $\mathrm{L}(\boldsymbol{\beta} \mid \mathfrak{D})$ and $\mathrm{pen}(\boldsymbol{\beta};\lambda)$ is the penalty term that splits into two components $\mathrm{pen}(\boldsymbol{\beta};\lambda) = \lambda \mathrm{pen}(\boldsymbol{\beta})$. The term $\mathrm{pen}(\boldsymbol{\beta})$ defines the form of the penalty and $\lambda \geq 0$ is the regularization parameter, which determines the impact of $\mathrm{pen}(\boldsymbol{\beta})$ at the solution of the regularized optimization problem

$$\hat{\boldsymbol{\beta}}(\lambda) = \arg\max_{\boldsymbol{\beta}} \{\log \mathrm{L}_{\mathrm{pen}}(\boldsymbol{\beta},\lambda)\}. \tag{1.18}$$

For the special case of $\lambda = 0$ the regularized solution coincides with the maximum likelihood estimate $\hat{\boldsymbol{\beta}}(0) = \hat{\boldsymbol{\beta}}_{\mathrm{ML}}$. Otherwise the estimate is, e. g., shrunken towards zero and the various values of $\lambda \geq 0$ trace out a path of solutions, where the resulting bias of the estimate is due to the associated size of the penalty term incorporated in the likelihood. The behavior at the limit $\lambda \to \infty$ depends on the specific selected penalty, but for shrinkage-towards-zero penalties we obtain $\hat{\boldsymbol{\beta}}(\lambda) \to \mathbf{0}$. A particular solution of (1.18) is often determined by crossvalidation, where λ is chosen to minimize the prediction error. The selection of a special type of the penalty term allows to handle the before mentioned demands on the resulting estimate. Some well-known examples include the *ridge penalty* $\mathrm{pen}(\boldsymbol{\beta}) := \mathrm{L}_2(\boldsymbol{\beta}) = \sum_j \beta_j^2$, Hoerl and Kennard (1970), which is used to find a unique estimate for an underdetermined estimation equation system. The topic of variable selection is addressed e. g. by the *lasso penalty* $\mathrm{pen}(\boldsymbol{\beta}) := \mathrm{L}_1(\boldsymbol{\beta}) = \sum_j |\beta_j|$ proposed by Tibshirani (1996). Due to the special shape of the contours of both penalty functions, the covariate estimates are shrunken towards zero. In contrast to the ridge penalty, the square-cut contours of the lasso penalty enable that small covariates can be estimated to be exactly zero, when maximizing the penalized likelihood, so that the solution to the lasso regularized optimization problem is sparse and simultaneously accomplishes the goals of estimation and model selection. We provide more details to this topic in Sections 4.1 to 4.3. The ridge and lasso penalty are special cases of the more general L_q-penalty with $L_q(\boldsymbol{\beta}) = \sum_j |\beta_j|^q$, $q > 0$. Another topic that can be addressed is the smoothing of unknown functions $f(\cdot)$ of continuous covariates x, which are approximated, e. g., by linear combinations of basis functions $B_k(\cdot)$, i. e. $f(x) \approx \sum_{k=1}^{g} \beta_k B_k(x)$, where the regression coefficients β_k represent the corresponding weights of the basis functions. Besides the selection of the basis functions, especially choosing the right number $g \in \mathbb{N}$ of basis functions is a hard task, since it determines the flexibility in the shape of the linear combination and therefore the fit to the unknown function. Using only few basis functions may be too restrictive to reflect possible shape variations of the unknown function. A large number of basis functions enables a high flexibility to fit the function, but coincides with the problem of interpolating the data or overfitting. A penalty based on the squared differences of the coefficients, like $\text{pen}(\boldsymbol{\beta}) = \sum_j (\beta_j - \beta_{j-1})^2$, can be used to avoid overfitting and to enforce a smooth estimate of the unknown function, compare, e. g., Eilers and Marx (1996) for details.

Regularization based regression methods are primarily explored in the context of the classical linear model. In survival regression based on the CRR model, regularization is considered by several authors, e. g. Verweij and van Houwelingen (1994) and van Houwelingen et al. (2006) proposed a *ridge* regularized CRR model, where the partial likelihood is used to form the penalized partial likelihood in (1.17) and the shrinkage parameter is determined by minimizing the cross-validated partial likelihood, Verweij and van Houwelingen (1993). Tibshirani (1997), Gui and Li (2005) and Park and Hastie (2007) applied the *lasso* penalty to the partial likelihood and Zhang and Lu (2007) use the *adaptive lasso*, Zou (2006), to handle the variable selection and model estimation simultaneously. Under some mild conditions the estimator is shown to have sparse and oracle properties. They use the Bayesian Information Criterion (BIC) for tuning parameter selection and a bootstrap variance approach for standard error. The *adaptive lasso*, the *elastic net*, Zou and Hastie (2005), and the *SCAD* penalty, Fan and Li (2001), are used for high-dimensional Cox models by Benner et al. (2010). Their article also provides a good comparative review of these penalized partial likelihood approaches. Fan and Li (2002) applied the *SCAD* penalty to the CRR model considering also gamma frailties. Gray (1992) used an additive model for the predictor to take account for smooth nonlinear covariate effects, modeled by penalized splines, covariate interactions and time-varying effects.

Several authors investigated also the regularization of the AFT model, e. g. Huang et al. (2006) considered variable selection via the *lasso* penalty and Huang and Ma (2010) via the *bridge* penalty, Fu (1998), in the semiparametric AFT model with unspecified error distribution, where inference is carried out in terms of weighted least squares with Kaplan-Meier weights. Johnson et al. (2008) use the *lasso*, *elastic net*, *SCAD* and *adaptive lasso* penalty for variable selection in the semiparametric AFT model, where inference is based on the penalized Buckley-James estimator, Buckley and James (1979). Wang et al. (2008) and Engler and Li (2009) apply the *elastic net* regularization to gene expression data. Datta et al. (2007) considered the *lasso* in the high-dimensional parametric AFT model with Gaussian and log-Weibull errors using partial least squares for estimation.

1.2.2. Bayesian regularization

From a Bayesian perspective there is a natural close relationship to the frequentist regularization, since, under certain conditions, the penalty terms correspond to log-prior terms that express specific information about the regression coefficients. Using the Bayesian formula (1.15) with an *informative* *prior* $p(\boldsymbol{\beta}\,|\,\lambda)$ for the regression coefficients given the tuning parameter $\lambda > 0$ and an additional (independent) hyperprior $p(\lambda)$ for the shrinkage parameter, the posterior for an observation model $L(\mathfrak{D}\,|\,\boldsymbol{\beta})$ is given as

$$p(\boldsymbol{\beta},\lambda\,|\,\mathfrak{D}) \propto L(\mathfrak{D}\,|\,\boldsymbol{\beta})p(\boldsymbol{\beta}\,|\,\lambda)p(\lambda) \tag{1.19}$$

with $\boldsymbol{\theta} = (\boldsymbol{\beta}',\lambda)'$ and $p(\boldsymbol{\theta}) = p(\boldsymbol{\beta}\,|\,\lambda)p(\lambda)$. If the regularization parameter λ is assumed to be known or fixed, the prior $p(\lambda)$ is negligible and the resulting maximization problem (1.16) becomes

$$\hat{\boldsymbol{\beta}}(\lambda) = \arg\max_{\boldsymbol{\beta}}\{\log L(\mathfrak{D}\,|\,\boldsymbol{\beta}) + \log p(\boldsymbol{\beta}\,|\,\lambda)\}\,. \tag{1.20}$$

Comparing the Bayesian optimization problem (1.20) with the frequentist optimization problem (1.18) shows, that the posterior mode estimate $\hat{\boldsymbol{\beta}}(\lambda)$ is equivalent to regularized maximum likelihood estimate, if the negative log-prior $-\log p(\boldsymbol{\beta}\,|\,\lambda)$ is proportional to the regularization term $\text{pen}(\boldsymbol{\beta};\lambda)$. Under this conjunction the penalized log-likelihood can be interpreted as the logarithm of the posterior distribution density $p(\boldsymbol{\beta}\,|\,\mathfrak{D},\lambda) \propto \log L_{\text{pen}}(\boldsymbol{\beta},\lambda)$ and consequently the frequentist regression parameter estimates (1.18) can be interpreted as mode of the posterior distribution. With the exception of the *SCAD* penalty all of the previous mentioned penalties comprise Bayesian versions of priors. E. g., the ridge and lasso estimates have a Bayesian interpretation as MAP estimates formulating i.i.d. Gaussian priors $p(\beta_j\,|\,\lambda) \propto \exp(-\lambda\sum_j \beta_j^2)$ or double exponential priors $p(\beta_j\,|\,\lambda) \propto \exp(-\lambda\sum_j |\beta_j|)$ on the regression coefficients, which are both special cases of the exponential power prior $p(\beta_j\,|\,\lambda) \propto \exp(-\lambda\sum_j |\beta_j|^q)$.

Besides the close connection between the Bayesian and the frequentist regularization approach also some differences and advantages arise from the Bayesian perspective. One difference is that the tuning parameter λ, which controls the regularization, is in general not assumed to be fixed and there is also a prior $p(\lambda)$ specified. Full Bayesian inference enables that all model parameters are simultaneously estimated and in particular the regression parameters $\boldsymbol{\beta}$ and the tuning parameter λ are jointly estimated. This offers new methods to estimate the complexity parameter λ by using the usual point estimates like the mode, mean or median of the marginal posterior $p(\lambda\,|\,\mathfrak{D})$ or the corresponding empirical counterparts from the MCMC sample of λ. In frequentist regularization crossvalidation is a popular method to determine reasonable values of the tuning parameter λ. Compared to the burden, which crossvalidation can cause for complex models in practice, the Bayesian approach provides a comparatively easy access to an estimate $\hat{\lambda}$. Further, the recruited prior $p(\lambda)$ incorporates uncertainty about the tuning parameter λ into the model, and uncertainty in estimating the tuning parameter can be addressed by the marginal posterior $p(\lambda\,|\,\mathfrak{D})$. In addition, integrating over the tuning parameter creates marginal priors for the regression coefficients $\boldsymbol{\beta}$, which differ from those when the tuning parameter is assumed to be fixed and induce a different kind of regularization behavior of the corresponding marginal penalty of the regression coefficients. A further challenge of some frequentist variable selection approaches like the lasso is the estimation of the standard error associated to the zero estimated regression coefficients, compare e. g. Tibshirani (1996) or Kyung et al. (2010). In MCMC based inference, standard errors for regression coefficients or other model parameters are a byproduct from the sampling based approach to the posterior.

Several authors have investigated the Bayesian regularization concept (mainly for Gaussian responses), proposing a lot of priors to address the before mentioned regression tasks and connections. In particular Lindley and Smith (1972) showed that using i.i.d. Gaussian priors for the regression coefficients $\boldsymbol{\beta}$ is leading to the *ridge* regression estimate as posterior mode. Tibshirani (1996) and Park and Casella (2008) showed that the *lasso* estimate results as posterior mode, if i.i.d. Laplace priors for the regression coefficients $\boldsymbol{\beta}$ are selected. Park and Casella (2008) provide also a full Bayesian version of the lasso by assuming an additional gamma prior for the squared shrinkage parameter. Griffin and Brown (2005) investigated various regularization priors that support a scale mixture of normal representation for the regression coefficients, West (1987). Under certain conditions, compare Section 4.1 to 4.3, such priors induce an adaptive, covariate specific shrinkage which avoids the overshrinkage of large regression coefficients. Armagan and Zaretzki (2010) use the scale mixture of normal representation to derive an *adaptive ridge* prior for posterior mode estimation in the linear regression model. Recently Polson and Scott (2011) describe the corresponding prior distribution that results in the *bridge* regression estimate, and Li and Lin (2010) and Hans (2011) investigate the prior associated to the *elastic net* penalty. Other Bayesian approaches for variable selection are based on bimodal spike-and-slab priors for the regression coefficients, where the spike-mode is exactly or close around zero to remove unimportant variables and the slab-mode is rather flat and differs form zero to retain important variables, compare George and McCulloch (1993), Smith and Kohn (1996), Ishwaran and Rao (2005b) and Li and Zhang(2010). The squared difference penalty, typically applied in penalized spline smoothing, Eilers and Marx (1996), is related to a Gaussian random walk assumption for the polynomial spline coefficients as shown in Lang and Brezger (2004), Brezger and Lang (2006).

Although the Bayesian regularization approaches for possibly high-dimensional linear predictors can be carried straightforward to the survival context the Bayesian literature dealing with these topics is quite sparse. In the framework of the CRR regression model Kaderali (2006) used a time-constant baseline hazard with a Normal-Gamma prior, Griffin and Brown (2005), for the regression coefficients. Recently Tachmazidou et al. (2010) used the *Bayesian lasso*, Park and Casella (2008), in combination with an exponential distribution of the survival times. Joint estimation of the baseline hazard and unregularized linear covariate effects in the CRR model has also been considered by Sinha (1993), who suggests a gamma process prior for the cumulative baseline hazard function. Lee et al. (2011) developed a semiparametric model for handling high-dimensional data by extending the *Bayesian lasso* to the CRR model, where the cumulative baseline hazard function is modeled nonparametrically by a discrete gamma process, compare Kalbfleisch (1978). Rockova et al. (2012) review hierarchical Bayesian formulations of various regularization and selection priors and apply them to Probit and Weibull survival regression models. Fahrmeir et al. (2010), Kneib et al. (2011) and Konrath et al. (2013) provide a unified approach to combined shrinkage, selection and smoothing in the framework of exponential family and hazard regression. The AFT model has not received much attention in the Bayesian regularization framework. Sha et al. (2006) propose for AFT models with log-normal and log-t distributional assumptions a Bayesian variable selection approach based on mixture priors for the regression coefficients, in the spirit of George and McCulloch (1993). There are several approaches to model the baseline survival quantities in order to get more flexible shapes for the survival time distribution. An example that fits in the Bayesian regularization framework is given by Komárek et al. (2005) who replaced the error distribution by a semiparametric penalized Gaussian mixture and Komárek et al. (2007) who extended this approach to interval censored data AFT with random effects.

1.3. Outline

Approaches for combined regularization with respect to shrinkage, selection and smoothing have a direct application to possibly high-dimensional regression problems. For example, in the presence of influential clinical predictors we may want to select important microarray features, while the clinical effects are assumed to be linear or nonlinear. Although regularization of high-dimensional coefficient vectors or smoothing of nonlinear effects or the development of flexible semiparametric versions of the CRR or AFT model have gained a lot of attention in the recent years, publications on the combination of the approaches are very rare.

The aim of this work is to derive flexible classes of AFT and CRR survival models by casting various regularization approaches into one general, unified Bayesian framework. The presented methods are based on the flexible and general approach for structured additive regression (STAR) for responses from exponential family models, Fahrmeir et al. (2004), and CRR-type survival models, Hennerfeind et al. (2006). On the one hand flexibility is addressed in terms of an extended version of the predictor, where various effect types are additively combined, each equipped with a suitable regularization prior. The structured additive modeling of the predictor is convenient for both, the inference and the interpretation of the different covariate effects. On the other hand flexibility is addressed in terms of the baseline survival distribution, which is modeled nonparametrically and smoothness priors are used to prevent overfitting. Each extension separately and both in combination provide large classes of flexible AFT-type and CRR-type regression models.

The unified Bayesian approach relies on the *hierarchical model representation* combined with suitable conditional independence assumption about the model parameters to support a modular structure. One major building block is the hierarchical formulation of the regularization priors for linear effects obtained through the representation as *scale mixture of normals*, West (1987). Auxiliary latent variance parameters enable a reformulation of the prior in terms of the product of a conditionally Gaussian prior given the variance parameter and a prior for the variance parameter given further hyperparameters. Besides the advantageous hierarchical representation, additional priors for the hyperparameters entail marginally a modification of the regularization prior for the regression coefficients. Such hyperpriors are very useful to enforce an adaptive (covariate-specific) shrinkage of the regression coefficients and hence to avoid the overshrinkage of large regression coefficients, as observed e. g. under the lasso penalty. In particular we consider the Bayesian lasso and ridge prior and a Normal Mixture of Inverse Gamma (NMIG) prior. Another major building block is given by the *basis function representation* of the various non-linear model components. The basis function representation preserves the linear structure for the non-linear predictor components and *random walk priors* for the basis function coefficients allow also a hierarchical reformulation with (improper) conditional Gaussian densities given variance or smoothing parameters as shown e. g. in Brezger and Lang (2006). In particular we consider smooth effects of continuous covariates as one representative of the various effect-types which support a basis function representation. Also the flexible extensions of the baseline quantities are also expressed by linear combinations of basis functions with random walk smoothness priors. In the AFT model the baseline error is modeled as penalized Gaussian mixture, Komárek et al. (2007) and in the CRR model the logarithm of baseline hazard rate is approximated by penalized B-splines, Hennerfeind et al. (2006). Besides the full likelihood specification, inference in the CRR model is also carried out in terms of the partial likelihood, where the baseline hazard is left unspecified.

The full Bayesian framework has the advantage that it facilitates a joint modeling and estimation of the baseline quantities and the regression coefficients of the extended predictor. No asymptotic assumptions or conjectures are needed for finite sample inference and the case $n < p$ is automatically covered. In addition, the *Markov chain Monte Carlo* simulation techniques build a versatile tool for the joint estimation. The derived MCMC samplers are based on *Gibbs sampling* or *Metropolis-Hastings within Gibbs sampling*. In particular in the CRR regression model, the full conditionals of the regression coefficients in the predictor are non-Gaussian. Samples from non-Gaussian full conditionals can be drawn in a unified computationally efficient way from *IWLS proposals*, introduced by Gamerman (1997). The general idea of IWLS proposals is to obtain a Gaussian proposal by matching the mode and the curvature of the full conditional at the current state of parameter vector in each update step. Metropolis-Hastings-steps with these multivariate Gaussian IWLS proposals have several advantages. The proposal can be used with multivariate coefficient vectors to take correlations into account. In addition, the proposal automatically adapts to the form of the full conditional and thereby avoids a manual tuning of the proposal density and typically leads to samplers with satisfactory mixing and convergence properties. Due to the hierarchical prior representation and conjugate hyperprior specification, Gibbs sampling remains possible for variance parameters and the model parameters on hierarchical stages below, like the shrinkage parameter. For AFT regression models, or models with a latent Gaussian structure such as Probit models, the conjugate conditional Gaussian priors for the regression coefficients induce full conditionals of the regression coefficients that are also Gaussian and facilitate Gibbs sampling. Finally, the provided modular hierarchical framework supports the *extensibility* of our approaches. In particular we can link the priors for combined regularization straightforward to various kinds of observation models arising e. g. from exponential family regression models. Due to the resulting modular structure of MCMC algorithms, it is also easy to extend the model at some places without having to re-implement the rest of the estimation algorithm.

Under the MCMC sampling approach for a full Bayesian inference the sharp *variable selection property* of some regularization priors gets alleviated. This is due to the fact, that the proposed MCMC techniques provide samples from the (marginal) posterior distribution, but they do not maximize the posterior. As a consequence, there is no exact zero estimate of a regression coefficient obtainable, even for a set of samples close to zero. From the theoretical point of view using the posterior mean instead of the posterior mode is not a drawback, since the posterior mode does not play the central role in Bayesian inference and Park and Casella (2008) or Hans (2009) give realistic examples, where the lasso posterior mean outerperforms the posterior mode in prediction and estimation. However, still regularization of the regression coefficients takes place and coefficients corresponding to covariates with minor effect are even so shrunken close to zero. Variable selection is supported through the inspection of the posterior distribution of individual regression coefficients or through posterior inclusion probabilities as provided by the NMIG prior and carried out in a post inferential step by hard shrinkage selection. In our simulations and applications we consider several empirical thresholding procedures as used in Konrath (2007) and recently proposed in Li and Lin (2010) with respect to their predictive performance.

The developed and described procedures are implemented in public available software, like `BayesX` for the extended CRR model based on the full likelihood and exponential family regression or in `R`-functions for the extended AFT model and the extended CRR model based on the partial likelihood.

The performance of the developed methods and algorithms is extensively tested by simulation studies and illustrated by three real world data sets.

1.4. Organization

The rest of this work is organized as follows: Part I is devoted to the extension of the AFT model. The considered extensions and their modeling are provided in Section 2, and Section 3 to Section 5 provide the associated priors for the model components. In particular in Section 4, we introduce the utilized Bayesian regularization priors for the joint shrinkage, selection and smoothing and investigate and illustrate their specific shrinkage properties. Finally, Section 6 addresses the posterior inference for model parameters based on MCMC simulations. Part II considers the extension of the CRR model and in particular Section 7 introduces the model extensions. Prior specification and posterior inference is carried out in Section 8 and Section 9. Simulations to test and demonstrate the flexibility and applicability of the proposed methodology are provided in Sections 10 (AFT model) and Section 11 (CRR model) of Part III and the data applications in Section 12 to Section 14 of Part IV. Optional results for the simulations and applications will be provided in an electronic supplement. Finally, the concluding Sections 15 and 16 contain a summary and comments on directions of future research.

PART I. BAYESIAN REGULARIZATION IN THE AFT MODEL

2. Extended AFT model

2.1. Basic AFT model

Let $T_i \geq 0$, $i = 1,...,n$, denote the random variable representing the non negative, continuous survival time of an individual i from a heterogeneous population. This heterogeneity of the population is caused by individual-specific characteristics that effect the individual's survival time, like sex, age or medical treatment of a patient in a clinical study. The functional dependence of the survival time on the covariates is determined in terms of the transformation $Y_i := \log(T_i) \in \mathbb{R}$ by the log-linear representation of the AFT regression model as introduced in (1.9)

$$Y_i := \log(T_i) = \eta_i + \sigma\varepsilon_i, \tag{2.1}$$

where $\eta_i \in \mathbb{R}$ denotes the predictor that summarizes the covariate effects and $\sigma > 0$ is a scale factor for the covariate independent random error terms $\varepsilon_i \in \mathbb{R}$. The errors are assumed to be independent and identically distributed with absolutely continuous density $f_\varepsilon(\cdot)$. This implies that the log-survival times Y_i, $i = 1,...,n$, are conditional independent given the covariates. In parametric AFT models the error term is often assumed to belong to a specific location-scale family, like the Gaussian or extreme value distribution for example. The observed right censored survival data is given as

$$\mathfrak{D} = \{(\tilde{y}_i, d_i, \mathbf{v}_i'), i = 1,...,n\}, \tag{2.2}$$

where $\tilde{y}_i = \log(\tilde{t}_i)$ is the logarithm of the observed survival time, $d_i \in \{0,1\}$ the censoring indicator and $\mathbf{v}_i = (v_{i1},...,v_{ip})'$ is the p-dimensional vector of the observed covariates for the n individuals of the sample.

This thesis considers two extensions of the AFT model to enable a more refined and flexible formulation of this model: On the one hand the predictor η_i is additively expanded to enable the regularization of some or all covariates with linear effects. Further nonlinear effects are considered, where functional forms of the effects are utilized to represent flexible relationships between the response and the corresponding covariates. In summary, the predictor gets a structured additive form, where each summand reflects the specific form of the covariate impact on the log-survival time. The covariate-specific predictor components are equipped with informative regularization priors, compare Section 4, to enforce the desired shrinkage of linear effects or the smoothing of the nonlinear effects. On the other hand the parametric assumptions of the error distribution are replaced by flexible semiparametric assumptions, where the error distribution is modeled by a penalized Gaussian mixture distribution.

2.2. Extended predictor

To attain a higher flexibility in modeling various functional relationships between the covariates and the response, the predictor is partitioned into three subgroups that represent the specific assumption about the functional form of the impact of the covariates. Accordingly, the vector of explanatory covariates is partitioned as $\mathbf{v}_i = (\mathbf{u}_i', \mathbf{x}_i', \mathbf{z}_i')'$ to reflect by notation the different ways how the covariates are treated. In particular, the predictor η_i is assumed to summarize the different functional forms of the covariates in a structured additive form given by

$$\eta_i = \mathbf{u}_i'\boldsymbol{\gamma} + \mathbf{x}_i'\boldsymbol{\beta} + f_1(z_{i1}) + ... + f_{p_z}(z_{ip_z}) . \tag{2.3}$$

The components of the predictor are used to describe

- *Linear effects* $\mathbf{u}_i'\boldsymbol{\gamma}$ of a moderate low number of unregularized, time-independent, categorical or continuous covariates $\mathbf{u}_i = (u_{i0,}u_{i1,}...,u_{ip_u})' \subset \mathbf{v}_i$, $p_u \ll n$, that are forced into the model. The regression coefficients $\boldsymbol{\gamma} = (\gamma_0, \gamma_1, ..., \gamma_{p_u})'$ model at least the global intercept term defined by γ_0 (with $u_{i0} = 1$, $i = 1,...,n$), which is in general not regularized and is also required for the identifiability of the optional nonlinear terms.

- *Regularized linear effects* $\mathbf{x}_i'\boldsymbol{\beta}$ of possibly high-dimensional categorical or continuous time-independent covariates $\mathbf{x}_i = (x_{i1,}...,x_{ip_x})' \subset \mathbf{v}_i$, with $p_x \le n$ or $p_x > n$. The regression coefficients $\boldsymbol{\beta} = (\beta_1, ..., \beta_{p_x})'$ are equipped with an informative shrinkage- or selection-type prior to identify those effects with the highest impact on the response.

- *Smooth nonlinear effects* $f_j(z_{ij})$, $j = 1,...,p_z$, which are defined by smooth functions $f_j(\cdot)$ of time-independent continuous covariates z_{ij} that need to be regularized to avoid overfitting. A suitable tool to model unknown functional forms of covariates is provided by semiparametric basis function approaches, where each of the unknown functions $f_j(\cdot)$ in the predictor (2.3) is represented in terms of a linear combination

$$f_j(z) = \sum_{k=1}^{g_j} \alpha_{jk} B_{jk}(z) = \mathbf{b}_j'(z)\boldsymbol{\alpha}_j \tag{2.4}$$

  of a finite number $g_j < \infty$ of known basis functions $\mathbf{b}_j(\cdot) = (B_{j1}(\cdot), ..., B_{jg_j}(\cdot))'$ and a vector of coefficients $\boldsymbol{\alpha}_j = (\alpha_{j1}, ..., \alpha_{jg_j})'$. In particular, the Bayesian penalized splines (P-splines) approach, as developed by Lang and Brezger (2004) is considered, which builds the Bayesian counterpart of the P-splines proposed by Eilers and Marx (1996). In this approach the numerical advantageous B-splines of De Boor (2001) are picked as basis functions and placed using a set of (inner, equidistant) knots $\xi_1, ..., \xi_{s_j}$, with $\min(z_j) = \xi_1 < ... < \xi_{s_j} = \max(z_j)$, from the support of the j-th covariate z_j. The number of B-spline basis functions with degree q_j is determined as $g_j = s_j + q_j - 1$. Since the B-splines are bounded and have local support over the range of a few knots, the corresponding design matrices are sparse (as well as the associated penalty matrices) and computational efficient matrix inversion is possible. In practice we often use cubic B-splines, i. e. $q_j = 3$. As trade-off for the number of basis functions the use of a moderate large number is proposed, that provides sufficient flexibility in the shape for a well suited approximation. This is combined with a Bayesian regularization of the distances between adjacent basis coefficients $\boldsymbol{\alpha}_j = (\alpha_{j1}, ..., \alpha_{jg_j})'$ by utilizing a Gaussian random walk prior that enforces the desired smoothness of the approximation and avoids overfitting. For identifiability reasons it is necessary to center all functions horizontally about zero and include an intercept term in the linear component of the predictor.

Further effect-types like varying coefficients, random effects, spatial effects, time-dependent effects or interactions can also be included in the predictor and cast into the unified modeling via basis function expansions as shown, e. g., in Brezger and Lang (2006) for exponential family regression, Kneib and Fahrmeir (2007), Hennerfeind et al. (2006) for geoadditive Cox-type survival regression models or in Fahrmeir and Kneib (2011) for both regression model types. The focus here relies on smooth nonlinear effects to demonstrate the methodological principle, but the implementation of other effect types is straightforward. A note about this generalization and the inclusion of time-dependent covariates is given in the Outlook Section 16.

Generic notation: Due to the linear structure of the basis function approach (2.4), the vector $\mathbf{f}_j = (f_j(z_{1j}),...,(z_{nj}))'$ of function evaluations at the observed values z_{ij}, $i = 1,..,n$, of covariate z_j can be expressed as the matrix product $\mathbf{f}_j = \mathbf{Z}_j\boldsymbol{\alpha}_j$, where the design matrix $\mathbf{Z}_j$ has the elements $b_{ik} = B_{jk}(z_{ij})$, $1 \le k \le g_j$, $1 \le i \le n$. In summary, with the design matrices $\mathbf{X}$ and $\mathbf{U}$ of the linear effects $\boldsymbol{\gamma}$ and $\boldsymbol{\beta}$, that have rows $\mathbf{x}_i'$ and $\mathbf{u}_i'$, it turns out, that the vector $\boldsymbol{\eta} = (\eta_1,...,\eta_n)'$ of extended predictors can always be represented in generic matrix form

$$\boldsymbol{\eta} = \mathbf{U}\boldsymbol{\gamma} + \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}_1\boldsymbol{\alpha}_1 + ... + \mathbf{Z}_{p_z}\boldsymbol{\alpha}_{p_z} \tag{2.5}$$

with components

$$\eta_i = \mathbf{u}_i'\boldsymbol{\gamma} + \mathbf{x}_i'\boldsymbol{\beta} + \mathbf{b}_1'(z_{i1})\boldsymbol{\alpha}_1 + ... + \mathbf{b}_{p_z}'(z_{ip_z})\boldsymbol{\alpha}_{p_z} \tag{2.6}$$

and a possibly high-dimensional parameterization of the predictor.

2.3. Extended error distribution

Assuming that the distribution of the error stems, e. g., from a location-scale family leads to models, where only a few number of parameters are used to describe this distribution. Picking up, e. g., the Weibull model from Section 1.1.3, where the error has a standard extreme value distribution, shows that there are only monotone baseline hazard functions possible. Checking parametric assumptions for the error is in general difficult in the presence of censoring. For this reasons the conventional parametric assumption of the error is replaced by a semiparametric distribution that is defined in terms of a finite penalized Gaussian mixture (PGM) as in Komárek et al. (2005). The density of the error distribution in the log-linear specification of the AFT model is approximated by a flexible continuous mixture distribution $\varepsilon \sim \sum_{k=1}^{g_0} w_k N(m_k, s_k^2)$ with density

$$f_\varepsilon(\varepsilon \mid \mathbf{w}) = \sum_{k=1}^{g_0} w_k \varphi(\varepsilon \mid m_k, s_k^2)\,, \tag{2.7}$$

where $\mathbf{w} = (w_1,...,w_{g_0})'$ is the vector of mixture weights corresponding to the finite set of Gaussian mixture densities $\varphi(\varepsilon \mid m_k, s_k^2)$, $k = 1,...,g_0$, with fixed means $\mathbf{m} = (m_1,...,m_{g_0})'$, $m_1 < ... < m_{g_0}$, and fixed variances $\mathbf{s}^2 = (s_1^2,...,s_{g_0}^2)'$, $s_k > 0$. The mean and variance of the error distribution are given by

$$\mu_\varepsilon = \mathbb{E}(\varepsilon \mid \mathbf{w}) = \sum_{k=1}^{g_0} w_k m_k, \quad \sigma_\varepsilon^2 = \mathbb{V}\mathrm{ar}(\varepsilon \mid \mathbf{w}) = \sum_{k=1}^{g_0} w_k \left(m_k^2 + s_k^2\right) - \mu_\varepsilon^2\,, \tag{2.8}$$

compare Appendix A.1.1.

Additional conditions are required to ensure that (2.7) is a probability density and to identify the location and scale parameter of the AFT model (2.1). To guarantee that (2.7) is a probability density with $\int f_\varepsilon(e \mid \mathbf{w}) de = 1$, the weights have to be positive $w_k > 0$ with $w_1 + ... + w_{g_0} = 1$. To fulfill both constraints, the generalized logit-reparametrization of the weights

$$w_k = w_k(\boldsymbol{\alpha}_0) = \frac{\exp(\alpha_{0,k})}{\sum_{j=1}^{g_0} \exp(\alpha_{0,j})}, \quad k = 1, ..., g_0, \tag{2.9}$$

in terms of unrestricted basis coefficients $\boldsymbol{\alpha}_0' = (\alpha_{0,1}, ..., \alpha_{0,g}) \in \mathbb{R}^g$ is used. Since this reparametrization is not unique, $w_j(\boldsymbol{\alpha}_0 + c) = w_j(\boldsymbol{\alpha}_0)$ for any scalar c, one of the g_0 unrestricted coefficients in $\boldsymbol{\alpha}_0$ is set equal to zero

$$\alpha_{0,k} := 0, \quad k \in \{1, ..., g_0\}. \tag{2.10}$$

To guarantee identifiability, the location and scale parameter of the error distribution need to be fixed or at least standardized with

$$\mu_\varepsilon = \sum_{k=1}^{g_0} w_k(\boldsymbol{\alpha}_0) m_k := 0, \quad \sigma_\varepsilon^2 = \sum_{k=1}^{g_0} w_k(\boldsymbol{\alpha}_0)\left(m_k^2 + s_k^2\right) := 1.$$

As shown in the Appendix A.1.2, standardization can be achieved by expressing two of the unrestricted coefficients $\boldsymbol{\alpha}_0$, e. g. $\alpha_{0,g-1}$ and $\alpha_{0,g-2}$ through the remaining coefficients. The weights have to match the constraints

$$\alpha_{g_0-1} = \log\left(\sum_{k=1}^{g_0-3} \exp(\alpha_{0,k}) c_{k,g_0-1} + c_{g_0,g_0-1}\right),$$
$$\alpha_{g_0-2} = \log\left(\sum_{k=1}^{g_0-3} \exp(\alpha_{0,k}) c_{k,g_0-2} + c_{g_0,g_0-2}\right),$$

with

$$c_{k,g_0-1} = -\frac{m_k - m_{g_0-2}}{m_{g_0-1} - m_{g_0-2}} \frac{1 - s^2 - m_k m_{g_0-1}}{1 - s^2 - m_{g_0-1} m_{g_0-2}}, \quad k = 1, ..., g_0 - 3, g_0,$$
$$c_{k,g_0-2} = -\frac{m_k - m_{g_0-1}}{m_{g_0-2} - m_{g_0-1}} \frac{1 - s^2 - m_k m_{g_0-2}}{1 - s^2 - m_{g_0-1} m_{g_0-2}}, \quad k = 1, ..., g_0 - 3, g_0,$$

when equal basis variances $s_k^2 = s^2$, $k = 1, ..., g_0$, and the identifiability constraint $\alpha_{0,g_0} = 0$ are used. Since these restrictions are hard to implement in the Bayesian context, we use an alternative strategy to standardize the error distribution in the constructed MCMC sampler, compare Section 6.2.1.

Similar to the Bayesian P-spline approach, used to extend the predictor, the error density (2.7) can be viewed as a basis function expansion, where the set of mixture densities $\varphi(\cdot \mid m_k, s_k^2)$, $k = 1, ..., g_0$, acts as basis functions positioned at the mean values $\mathbf{m} = (m_1, ..., m_{g_0})'$ that may be denoted as the knots of the basis, and the mixture weights $\mathbf{w} = (w_1, ..., w_{g_0})'$ correspond to the basis coefficients. In the spirit of Bayesian P-spline smoothing, a moderate large number g_0 of basis functions is used to guarantee the flexibility of the approximation in combination with an imposed random walk regularization prior, which controls the variation to achieve the desired smoothness. With respect to the reparametrization in (2.9), the regularization prior is finally formulated for the unrestricted coefficients $\boldsymbol{\alpha}_0' = (\alpha_{0,1}, ..., \alpha_{0,g_0})$. The grid points m_k, the basis variances s_k^2 as well as the constraints of $\boldsymbol{\alpha}_0'$ for standardization can be chosen independently from the location and the scale of the true distribution of Y_i. Komárek et al. (2005) recommend placing the knots on an equidistant grid in the interval

$[-4.5,4.5]$ with the distance $m_k - m_{k-1} = 0.3$ between consecutive knots and the use of common variances $s_k^2 = \frac{2}{3}(m_k - m_{k-1}) = 0.2$. This implies $g_0 = 31$ as number of basis function and that the mixture density is practically zero outside the interval $(-6.6,6.6)$. If only one mixture density is used, $g_0 = 1$, the mixture distribution collapses to the parametric case with Gaussian error. In principle, any mixture density can be used to specify the error distribution and there is no need using Gaussian mixtures, but sampling of truncated observations from the mixture components should be feasible to impute the survival times, compare Section 3.

Distribution of the log-survival time

Due to the structure of the model, the log-survival times also follow a mixture distribution, since one can associate to each observation $(y_i, \mathbf{v}_i)$ a latent error quantity $\varepsilon_i = (y_i - \eta_i)/\sigma$. The density of the log-survival time $Y_i \mid \boldsymbol{\theta}$ is in general given by

$$f_i(y_i \mid \boldsymbol{\theta}) = \frac{1}{\sigma} f_\varepsilon\left(\frac{y_i - \eta_i}{\sigma} \mid \boldsymbol{\alpha}_0 \right).$$

Using the mixture representation of the error we get the mixture distribution density of the log-survival time as

$$\begin{aligned} f_i\left(y_i \middle| \boldsymbol{\theta}\right) &= \sum_{k=1}^{g_0} w_k\left(\boldsymbol{\alpha}_0\right) \frac{1}{\sigma} \varphi\left(\frac{y_i - \eta_i}{\sigma} \mid m_k, s_k^2 \right) \\ &= \sum_{k=1}^{g_0} \frac{w_k\left(\boldsymbol{\alpha}_0\right)}{\sqrt{2\pi}\sigma s_k} \exp\left(-\frac{1}{2\sigma^2 s_k^2} \left(y_i - \eta_i - \sigma m_k\right)^2 \right) \\ &= \sum_{k=1}^{g_0} w_k\left(\boldsymbol{\alpha}_0\right) \varphi\left(y_i \mid \eta_i + \sigma m_k, \sigma^2 s_k^2\right) \end{aligned} \tag{2.11}$$

with the extended predictor η_i of (2.6) and the corresponding parameters $\boldsymbol{\theta} = (\boldsymbol{\alpha}', \boldsymbol{\beta}', \boldsymbol{\gamma}', \sigma)'$, with $\boldsymbol{\alpha} = (\boldsymbol{\alpha}_0', \boldsymbol{\alpha}_1', \ldots, \boldsymbol{\alpha}_{p_z}')'$, $\boldsymbol{\alpha}_j = (\alpha_{0,j}, \ldots, \alpha_{0,g_j})'$. The resulting conditional mean and conditional variance of the response y_i given the model parameters resp. covariates are

$$\mu_{Y_i} = \mathbb{E}(Y_i \mid \boldsymbol{\theta}) = \eta_i + \sigma\mu_\varepsilon, \quad \sigma_{Y_i}^2 = \mathbb{V}\mathrm{ar}(Y_i \mid \boldsymbol{\theta}) = \sigma^2 \sigma_\varepsilon^2.$$

If the error distribution ε is standardized with $\mu_\varepsilon = 0$ and $\sigma_\varepsilon^2 = 1$, one gets

$$\mu_{Y_i} = \mathbb{E}(Y_i \mid \boldsymbol{\theta}) = \eta_i, \quad \sigma_{Y_i}^2 = \mathbb{V}\mathrm{ar}(Y_i \mid \boldsymbol{\theta}) = \sigma^2. \tag{2.12}$$

Distribution of the baseline error

We introduce the notation Y_0 to describe the *baseline error* $Y_0 := \gamma_0 + \sigma\varepsilon$, with

$$\mu_{Y_0} = \mathbb{E}(Y_0 \mid \boldsymbol{\theta}) = \gamma_0 + \sigma\mu_\varepsilon, \quad \sigma_{Y_0}^2 = \mathbb{V}\mathrm{ar}(Y_0 \mid \boldsymbol{\theta}) = \sigma^2 \sigma_\varepsilon^2, \tag{2.13}$$

as the associated location and squared scale of the baseline error distribution. Since the standardization is in general not implemented in the software, we can compute these expressions from the posterior samples of the involved quantities to verify the convergence or mixing.

These expressions reduce to

$$\mu_{Y_0} = \mathbb{E}(Y_0 \mid \boldsymbol{\theta}) = \gamma_0, \quad \sigma_{Y_0}^2 = \mathbb{V}\mathrm{ar}(Y_0 \mid \boldsymbol{\theta}) = \sigma^2, \tag{2.14}$$

if the error distribution ε is standardized.

2.4. Likelihood

The parameterization $\boldsymbol{\theta}=(\boldsymbol{\alpha}',\boldsymbol{\beta}',\boldsymbol{\gamma}',\sigma)'$ of the extended model terms enables a full likelihood specification with respect to the partial knowledge caused by the right censoring of the survival times of some individuals. Based on the generic formulation in (1.11) the full likelihood of the extended AFT model under non-informative right censoring is given as

$$L(\mathfrak{D}\,|\,\boldsymbol{\theta})=\prod_{i=1}^{n}L_i(\mathfrak{D}_i\,|\,\boldsymbol{\theta})=\prod_{i=1}^{n}f_i(\tilde{y}_i\,|\,\boldsymbol{\theta})^{d_i}\,S_i(\tilde{y}_i\,|\,\boldsymbol{\theta})^{1-d_i}\,.$$

To complete the AFT regression model from the Bayesian point of view, all model parameters have to be equipped with suitable prior distributions. In the next section we consider at first the priors associated to the data augmentation. The priors of the predictor components will be derived in the subsequent section.

3. Data augmentation priors

It is often advantageous for inference to introduce additional latent model parameters, which simplify the structure of complex models, compare Tanner and Wong (1987). In the extended AFT model the censoring and the formulation of the error as mixture distribution complicate the inference. The individual likelihood contributions to the model likelihood have the complex form $L_i(\mathfrak{D}_i\,|\,\boldsymbol{\theta})=f_i(\tilde{y}_i\,|\boldsymbol{\theta})^{d_i}S_i(\tilde{y}_i\,|\boldsymbol{\theta})^{1-d_i}$, where for an uncensored observation ($d_i=1$) the mixture density

$$f_i(\tilde{y}_i\,|\boldsymbol{\theta})=\sum_{k=1}^{g_0}w_k(\boldsymbol{\alpha}_0)\varphi(\tilde{y}_i\,|\,\eta_i-\sigma m_k,\sigma^2 s_k^2)\,,$$

and for a censored observation ($d_i=0$) the survival function

$$S_i(\tilde{y}_i\,|\boldsymbol{\theta})=\int_{\tilde{y}_i}^{\infty}f_i(s\,|\boldsymbol{\alpha},\boldsymbol{\beta},\boldsymbol{\gamma},\sigma^2)ds$$

need to be evaluated. To bypass the mixture density representation and the evaluation of the integral in the survival function, three further groups of latent quantities are introduced to augment the likelihood. The problem of censoring can be overcome by treating the unobserved true survival times as latent data, and the imputation of these latent quantities is leading to an uncensored regression model that is fitted in each MCMC iteration. Considered are in particular the vector of latent exact survival times $\mathbf{t}=(t_1,...,t_n)'$ and the vector of exact censoring times $\mathbf{c}=(c_1,...,c_n)'$, which are both partially unobserved since we observe under right censoring either an exact survival or an exact censoring time ($\tilde{t}_i=\min(t_i,c_i)$). To solve the task concerning the mixture representation, we rewrite the likelihood in terms of latent mixture component labels $\mathbf{r}=(r_1,...r_n)'$, $r_i\in\{1,...,g_0\}$, which is leading to conditional Gaussian likelihood contributions. In summary, the *complete data* containing the latent quantities is denoted as

$$\mathfrak{D}^{comp}=\{(\tilde{t}_i,d_i,t_i,c_i,r_i,\mathbf{v}_i),i=1,...,n\}\,.$$

With respect to the complete data we obtain a likelihood-prior structure that simplifies the derivation of the conditional posterior distributions and enforces Gibbs sampling for almost all model parameters.

Augmented survival times

For the moment we disregard the latent component labels $\mathbf{r}=(r_1,...r_n)'$ and consider the partially latent survival and censoring times. The first augmentation concerns the possibly right censored observations of the survival times $\tilde{T}_i=\min(T_i,C_i)$, $i=1,...,n$. The sample population is split into two groups, those individuals for which a survival time $\tilde{T}_i=T_i$, $D_i=1$ is observed and those for which a censoring time $\tilde{T}_i=C_i$, $D_i=0$ is observed. Let $\boldsymbol{\theta}$ denote the parameters of the survival time distribution $T_i \mid \boldsymbol{\theta}$ and $\boldsymbol{\psi}$ the parameters of the censoring time distribution $C_i \mid \boldsymbol{\psi}$. Since $\tilde{T}_i$ is either T_i or C_i the joint distribution of $\tilde{T}_i,D_i,T_i,C_i$ is given as

$$f_{\tilde{T}_i,D_i,T_i,C_i}(\tilde{t}_i,d_i,t_i,c_i \mid \boldsymbol{\theta},\boldsymbol{\psi})=f_{D_i,T_i,C_i}(d_i,t_i,c_i \mid \boldsymbol{\theta},\boldsymbol{\psi}).$$

For a censored observation $D_i=0,\tilde{T}_i=C_i$ we obtain with $T_i>C_i$ the relationship

$$\begin{aligned}&\mathbb{P}\big(D_i=0,T_i\in[t_i,t_i+h_1),C_i\in[c_i,c_i+h_2)\mid\boldsymbol{\theta},\boldsymbol{\psi}\big)\\&\quad=\mathbb{P}\big(T_i>C_i,T_i\in[t_i,t_i+h_1),C_i\in[c_i,c_i+h_2)\mid\boldsymbol{\theta},\boldsymbol{\psi}\big)\\&\quad\approx\mathbb{P}\big(T_i>c_i\mid T_i\in[t_i,t_i+h_1),C_i\in[c_i,c_i+h_2)\big)\mathbb{P}\big(T_i\in[t_i,t_i+h_1)\mid\boldsymbol{\theta}\big)\mathbb{P}\big(C_i\in[c_i,c_i+h_2)\big|\boldsymbol{\psi}\big),\end{aligned}$$

where the last equality utilizes the conditional independence of the survival and censoring times and that c_i is a fixed number . Using further the relationships $f_{T_i}(t_i|\boldsymbol{\theta})=\lim_{h_1\to0}\mathbb{P}(T_i\in[t_i,t_i+h_1)\mid\boldsymbol{\theta})/h_1$ and $f_{C_i}(C_i|\boldsymbol{\psi})=\lim_{h_2\to0}\mathbb{P}(C_i\in[c_i,c_i+h_2)|\boldsymbol{\psi})/h_2$, compare (1.1), the joint distribution reads

$$\begin{aligned}f_{D_i,T_i,C_i}(0,t_i,c_i\mid\boldsymbol{\theta},\boldsymbol{\psi})&=\lim_{\substack{h_1\to0\\h_2\to0}}\frac{\mathbb{P}\big(D_i=0,T_i\in[t_i,t_i+h_1),C_i\in[c_i,c_i+h_2)\mid\boldsymbol{\theta},\boldsymbol{\psi}\big)}{h_1h_2}\\&=\mathbb{P}(T_i>c_i\mid T_i=t_i,C_i=c_i)f_{T_i}(t_i|\boldsymbol{\theta})f_{C_i}(c_i|\boldsymbol{\psi})\\&=1_{[c_i,\infty)}(t_i)f_{T_i}(t_i|\boldsymbol{\theta})f_{C_i}(c_i|\boldsymbol{\psi}).\end{aligned}$$

Similar steps for an uncensored observation $D_i=1,\tilde{T}_i=T_i$ with $T_i<C_i$ are leading to

$$f_{D_i,T_i,C_i}(1,t_i,c_i\mid\boldsymbol{\theta},\boldsymbol{\psi})=1_{[t_i,\infty)}(c_i)f_{C_i}(c_i|\boldsymbol{\psi})f_{T_i}(t_i\mid\boldsymbol{\theta}).$$

Consequently the complete data likelihood contribution for the i-th observation can be written as

$$L_i\big(\mathfrak{D}^{\text{comp}}\mid\boldsymbol{\theta},\boldsymbol{\psi}\big)=\big\{1_{[t_i,\infty)}(c_i)f_{C_i}(c_i|\boldsymbol{\psi})f_{T_i}(t_i\mid\boldsymbol{\theta})\big\}^{d_i}\big\{1_{[c_i,\infty)}(t_i)f_{T_i}(t_i|\boldsymbol{\theta})f_{C_i}(c_i|\boldsymbol{\psi})\big\}^{1-d_i}$$

and inserting $t_i=\tilde{t}_i$ if $d_i=1$ and $c_i=\tilde{t}_i$ if $d_i=0$ is leading to

$$L_i\big(\mathfrak{D}^{\text{comp}}\mid\boldsymbol{\theta},\boldsymbol{\psi}\big)=\big\{1_{[\tilde{t}_i,\infty)}(c_i)f_{C_i}(c_i|\boldsymbol{\psi})f_{T_i}(\tilde{t}_i\mid\boldsymbol{\theta})\big\}^{d_i}\big\{1_{[\tilde{t}_i,\infty)}(t_i)f_{T_i}(t_i|\boldsymbol{\theta})f_{C_i}(\tilde{t}_i|\boldsymbol{\psi})\big\}^{1-d_i}. \tag{3.1}$$

The marginalization over the i-th latent quantity, which is either in the censoring case a true survival time, or vice versa a censoring time for an uncensored observation, results in

$$\begin{aligned}L_i\big(\mathfrak{D}\mid\boldsymbol{\theta},\boldsymbol{\psi}\big)&=\Big(f_{T_i}(\tilde{t}_i\mid\boldsymbol{\theta})\cdot\int_{\mathbb{R}^+}1_{[\tilde{t}_i,\infty)}(c)f_{C_i}(c\mid\boldsymbol{\psi})dc\Big)^{d_i}\Big(f_{C_i}(\tilde{t}_i\mid\boldsymbol{\psi})\cdot\int_{\mathbb{R}^+}1_{[\tilde{t}_i,\infty)}(t)f_{T_i}(t\mid\boldsymbol{\theta})dt\Big)^{1-d_i}\\&=\big(f_{T_i}(\tilde{t}_i\mid\boldsymbol{\theta})\cdot S_{C_i}(\tilde{t}_i\mid\boldsymbol{\psi})\big)^{d_i}\big(f_{C_i}(\tilde{t}_i\mid\boldsymbol{\psi})\cdot S_{T_i}(\tilde{t}_i\mid\boldsymbol{\theta})\big)^{1-d_i},\end{aligned}$$

which coincides with the i-th likelihood contribution from (1.10), so that in summary the (observed) data likelihood $L(\boldsymbol{\theta},\boldsymbol{\psi}\mid\mathfrak{D})$ can be interpreted as the marginal likelihood of the complete data likelihood $L(\boldsymbol{\theta},\boldsymbol{\psi}\mid\mathfrak{D}^{\text{comp}})$ of the latent exact survival and censoring times. If the censoring is independent and noninformative, one can neglect the components of the censoring process for inference on the parameter $\boldsymbol{\theta}$ of main interest and we get from (3.1)

$$L\left(\mathfrak{D}^{\text{comp}} \mid \boldsymbol{\theta}\right) \propto \prod_{i=1}^{n} 1_{[\tilde{t}_i,\infty)}(t_i)^{1-d_i} f_{T_i}(t_i \,|\, \boldsymbol{\theta}) \propto L\left(\mathfrak{D}^{\text{comp}} \mid \boldsymbol{\theta}, \boldsymbol{\psi}\right)$$

or in terms of the log-transformed survival times $Y_i = \log(T_i)$

$$L\left(\mathfrak{D}^{\text{comp}} \mid \boldsymbol{\theta}\right) \propto \prod_{i=1}^{n} 1_{[\tilde{y}_i,\infty)}(y_i)^{1-d_i} f_{Y_i}(y_i \,|\, \boldsymbol{\theta}).$$

Due to these assumptions, the true censoring times are not required for inference about the parameter $\boldsymbol{\theta}$ and we only need to impute the partially latent, exact survival times corresponding to the censored observations.

Augmented mixture distribution

The second augmentation concerns the mixture representation of the baseline error. In general a mixture model, defined by mixture distribution with g_0 components like in (2.11), can be viewed as an incomplete data problem, when the allocation of each observation to one of the components $\{1,...,g_0\}$ is treated as missing data, compare e. g. Frühwirth-Schnatter (2006). Let R denote a discrete indicator variable with values in the set $\{1,...,g_0\}$ that labels the g_0 components of the mixture distribution. For each observation of an exact log-survival time $y_i = \log(t_i)$, $i = 1,...,n$, the realization $r_i \in \{1,...,g_0\}$ of the discrete allocation variable R_i indicates from which of the g_0 mixture components the i-th observation is assumed to arise. Conditional on knowing the mixture component with label r_i, the distribution of $Y_i \mid r_i, \boldsymbol{\theta}$ is Gaussian with mean $\mathbb{E}(Y_i \mid r_i, \boldsymbol{\theta}) = \eta_i - \sigma m_{r_i}$ and variance $\mathbb{V}\text{ar}(Y_i \mid r_i, \boldsymbol{\theta}) = \sigma^2 s_{r_i}^2$, i. e.

$$p(y_i \mid r_i, \boldsymbol{\theta}) := \varphi(y_i \mid \eta_i - \sigma m_{r_i}, \sigma^2 s_{r_i}^2),$$

and the probability, that Y_i belongs to the r_i-th mixture component, is discrete with

$$p(r_i \mid \boldsymbol{\theta}) := \mathbb{P}(R_i = r_i \mid \boldsymbol{\theta}) = w_{r_i}(\boldsymbol{\alpha}_0).$$

The resulting joint density of the completed data (Y_i, R_i), $i = 1,...,n$, is displayed as

$$p(y_i, r_i \mid \boldsymbol{\theta}) = p(y_i \mid r_i, \boldsymbol{\theta}) p(r_i \mid \boldsymbol{\theta}) = \varphi(y_i \mid \eta_i - \sigma m_{r_i}, \sigma^2 s_{r_i}^2) w_{r_i}(\boldsymbol{\alpha}_0)$$

and at last, the finite mixture arises as marginal distribution over the component labels, if it is not possible to record the group indicator R_i and only the random variable $Y_i = \log(T_i)$ is observed

$$p(y_i \mid \boldsymbol{\theta}) = \sum_{r_i=1}^{g_0} p(y_i, r_i \mid \boldsymbol{\theta}) = \sum_{r_i=1}^{g_0} w_{r_i}(\boldsymbol{\alpha}_0) \varphi(y_i \mid \eta_i - \sigma m_{r_i}, \sigma^2 s_{r_i}^2).$$

Thereby and with the argumentation of the last subsection the augmented likelihood contribution with respect to the complete data is finally given as

$$L_i\left(\mathfrak{D}^{\text{comp}} \mid \boldsymbol{\theta}\right) \propto 1_{[\tilde{y}_i,\infty)}(y_i)^{1-d_i} p(y_i \,|\, r_i, \boldsymbol{\theta}) p(r_i \mid \boldsymbol{\theta}).$$

In this complete data representation we associate the components $p(y_i \,|\, r_i, \boldsymbol{\theta})$ and $p(r_i \mid \boldsymbol{\theta})$ to the prior-part of the Bayesian model and identify $1_{[\tilde{y}_i,\infty)}(y_i)^{1-d_i}$ as the likelihood-part.

In summary, we obtain the degenerated augmented likelihood

$$L\left(\mathfrak{D}^{\text{comp}} \mid \boldsymbol{\theta}\right) = \prod_{i=1}^{n} 1_{[\tilde{y}_i,\infty)}(y_i)^{1-d_i} =: L\left(\mathfrak{D} \mid \mathbf{y}\right) \tag{3.2}$$

and, given the independence of $Y_i, R_i \mid \boldsymbol{\theta}$, $i = 1,\dots,n$, the joint prior of the latent data $(\mathbf{y},\mathbf{r})$ is

$$p(\mathbf{y},\mathbf{r} \mid \boldsymbol{\theta}) = p(\mathbf{y} \mid \mathbf{r},\boldsymbol{\theta})p(\mathbf{r} \mid \boldsymbol{\theta}) = \prod_{i=1}^{n} p(y_i \mid r_i,\boldsymbol{\theta}) \prod_{i=1}^{n} p(r_i \mid \boldsymbol{\theta}).$$

With the definitions $\boldsymbol{\Sigma}_y := \sigma^2 \mathbf{S}_r$ and $\boldsymbol{\mu}_y := \boldsymbol{\eta} - \sigma \mathbf{m}_r$, with $\mathbf{S}_r := \operatorname{diag}(s_{r_1}^2,\dots,s_{r_n}^2)$ and $\mathbf{m}_r := (m_{r_1},\dots,m_{r_n})'$, we get for the first component of the joint prior a multivariate Gaussian distribution

$$\mathbf{Y} \mid \mathbf{r},\boldsymbol{\theta} \sim N(\boldsymbol{\mu}_y, \boldsymbol{\Sigma}_y) \tag{3.3}$$

with mean vector $\boldsymbol{\mu}_y$, covariance matrix $\boldsymbol{\Sigma}_y$ and density

$$p\left(\mathbf{y} \mid \mathbf{r},\boldsymbol{\theta}\right) \propto \frac{1}{\left|\boldsymbol{\Sigma}_y\right|} \exp\left(-\frac{1}{2}\left(\mathbf{y}-\boldsymbol{\mu}_y\right)' \boldsymbol{\Sigma}_y^{-1} \left(\mathbf{y}-\boldsymbol{\mu}_y\right)\right).$$

The second component is the product of n discrete multinomial distributions,

$$R_i \mid \boldsymbol{\alpha}_0 \sim \text{MulNom}(1, \mathbf{w}(\boldsymbol{\alpha}_0)), \tag{3.4}$$

$\mathbf{w}(\boldsymbol{\alpha}_0) = (w_1(\boldsymbol{\alpha}_0),\dots,w_{g_0}(\boldsymbol{\alpha}_0))'$, with probability

$$p\left(\mathbf{r} \mid \boldsymbol{\theta}\right) = p\left(\mathbf{r} \mid \boldsymbol{\alpha}_0\right) = \prod_{i=1}^{n} w_{r_i}\left(\boldsymbol{\alpha}_0\right) = \prod_{j=1}^{g_0} w_j^{n_j}\left(\boldsymbol{\alpha}_0\right) = \left(\sum_{j=1}^{g_0} \exp\left(\alpha_{0,j}\right)\right)^{-n} \prod_{j=1}^{g_0} \exp\left(n_j \alpha_{0,j}\right),$$

where $n_j = \sum_{i=1}^{n} I\left(r_i = j\right)$ is the number of observations for which the component label $\mathbf{r}$ equals j. In the last equation we use the reparametrization of the mixture weights (2.9). Because marginalization over the latent variables is leading to the original (marginal) posterior

$$\int L\left(\mathfrak{D}^{\text{comp}} \mid \boldsymbol{\theta}\right) d\mathbf{y} d\mathbf{r} \propto L\left(\mathfrak{D} \mid \boldsymbol{\theta}\right) p\left(\boldsymbol{\theta}\right),$$

marginal characteristics of the parameter $\boldsymbol{\theta}$ are the same, irrespective if they are obtained from the complete or marginal posterior.

4. Regularization priors

In this section continuous regularization priors for shrinkage, selection and smoothing of the various regression model components are considered and compared. The presented priors support a hierarchical reformulation in terms of conditional Gaussian densities given variance parameters, where the variance parameters play the central role to control the desired kind of regularization. The various Bayesian regularization strategies for shrinkage, selection and smoothing are generated by varying the prior specifications on several stages of the hierarchical model.

Shrinkage or selection of linear predictor components, compare Subsections 4.1 to 4.3, relies on the interpretation of the associated priors as *scale mixtures of normal distributions*, West (1987). The prior distribution for the regression coefficients β_j, $j = 1,\dots,p_x$, is represented as

$$p(\beta_j \mid \cdot) = \int \varphi(\beta_j \mid 0, \tau^2_{\beta_j}) p(\tau^2_{\beta_j} \mid \cdot) d\tau^2_{\beta_j} , \tag{4.1}$$

where $\varphi(\cdot \mid m, s^2)$ denotes the density of a Gaussian distribution with mean m and variance s^2 and $p(\tau^2_{\beta_j} \mid \cdot)$ is the density of the mixing distribution of the variance parameter $\tau^2_{\beta_j}$. The switch from the marginal to the conditional prior representation leads to a hierarchical prior formulation in terms of the conditional Gaussian distribution $\beta_j \mid 0, \tau^2_{\beta_j}$ assigned to the regression coefficients and the distribution $\tau^2_{\beta_j} \mid \cdot$ assigned to the variance parameter. In this formulation the variance parameters control the concentration of the Gaussian prior for the regression coefficients around zero and determine the amount of shrinkage, where small variances induce a strong concentration and a heavy shrinkage towards zero. In addition, the scale mixture representation allows for a flexible kind of regularization, since each regression coefficient is equipped with its own variance parameter. A covariate-specific shrinkage is advantageous to avoid the overshrinkage of larger regression coefficients. E. g. Zou (2006) showed this in terms of the adaptive lasso penalty, $\text{pen}(\boldsymbol{\beta};\boldsymbol{\lambda}) = \sum_{j=1}^{p_x} \lambda_j |\beta_j|$, where covariate-specific penalties λ_j are introduced to support the unbiasedness for larger regression coefficient estimates. In contrast the lasso penalty, $\text{pen}(\boldsymbol{\beta};\lambda) = \sum_{j=1}^{p_x} \lambda |\beta_j|$, with its uniform shrinkage of all coefficients, produces biased estimates also for large regression coefficients, because all regression coefficients share a common regularization parameter λ. The flexibility of the regularization is further supported by the option to utilize additional priors for the hyperparameters of the variance distribution $\tau^2_{\beta_j} \mid \cdot$, which leads marginally to a modification of the mixing variance distribution. Finally, this modifies marginally the regularization prior of the regression coefficients and enables more sophisticated types of shrinkage and selection priors.

Under the conditional independence assumption for the regression coefficients given the variance parameters, the prior hierarchy is represented by the multivariate Gaussian prior distribution $\boldsymbol{\beta} \mid \boldsymbol{\tau}^2_\beta \sim N(\boldsymbol{\mu}_\beta, \boldsymbol{\Sigma}_\beta)$ with zero mean vector $\boldsymbol{\mu}_\beta = \mathbf{0}$ and covariance matrix $\boldsymbol{\Sigma}_\beta = \text{diag}(\tau^2_{\beta_1}, \ldots, \tau^2_{\beta_{p_x}})$ and the joint prior distribution of the mixing variances $\boldsymbol{\tau}^2_\beta \mid \cdot$, $\boldsymbol{\tau}^2_\beta = (\tau^2_{\beta_1}, \ldots, \tau^2_{\beta_{p_x}})'$, given further hyperparameters. Under conditional independence given the hyperparameters, the prior $\boldsymbol{\tau}^2_\beta \mid \cdot$ is the product of the priors of the single variance parameters. Since large variance parameters induce less shrinkage, the priors for the unregularized linear regression coefficients $\boldsymbol{\gamma}$ can also be cast in this representation. We write $\boldsymbol{\gamma} \sim N(\boldsymbol{\mu}_\gamma, \boldsymbol{\Sigma}_\gamma)$, with $\boldsymbol{\mu}_\gamma = \mathbf{0}$ and $\boldsymbol{\Sigma}_\gamma = c\mathbf{I}$, $c > 0$ large, or $\boldsymbol{\Sigma}_\gamma^{-1} \to \mathbf{0}$ to denote the assigned weakly informative Gaussian priors. These priors cause virtual no regularization and are appropriate for low-dimensional numbers of covariates that should always enter the model. Smoothing of nonlinear predictor terms or model components, like the baseline error distribution density in the AFT model or the log-baseline hazard function in the CRR model, rely on the basis function representation of these components. The recruited random walk smoothing priors for these semiparametric model components, compare Subsection 4.6, allow a similar hierarchical reformulation in terms of (partially improper) conditional Gaussian densities for the basis function coefficients $\boldsymbol{\alpha}_j$, $j = 0,1,\ldots,p_z$, given variance parameters $\tau^2_{\alpha_j}$, i. e. $\boldsymbol{\alpha}_j \mid \tau^2_{\alpha_j} \sim N(\boldsymbol{\mu}_{\alpha_j}, \boldsymbol{\Pi}_{\alpha_j})$ with $\boldsymbol{\mu}_{\alpha_j} = \mathbf{0}$ and appropriate precision matrix $\boldsymbol{\Pi}_{\alpha_j}$ that depends on $\tau^2_{\alpha_j}$.

Finally, the conditional Gaussian priors of the predictor components act as intermediate quantities in the joint prior to separate the priors of the associated variances and further parameters from lower stages of the hierarchy from the likelihood. This means, that the likelihood is not involved in the MCMC update of the variance parameters and the hyperparameters on hierarchical stages below and, as a consequence, the associated full conditionals have a closed form and enable fast Gibbs sampling. In addition, for Gaussian or latent Gaussian observation models, due to self-conjugacy, Gibbs sampling can be applied to update the predictor components. Also for non Gaussian observation models the quadratic structure of the conditional Gaussian prior of the regression coefficients simplifies the structure of the constructed IWLS-proposal distribution as shown, e. g. in CRR inference Section 9.

4.1. Bayesian ridge prior

4.1.1. Prior hierarchy

A well known penalty to deal with multicollinearity or the problem of $p_x > n$ in classical regression is the ridge penalty. In ridge regression the penalized least squares criterion is minimized with respect to the penalty $\text{pen}(\boldsymbol{\beta};\lambda) = \lambda \sum_{j=1}^{p_x} \beta_j^2$, $\lambda \geq 0$. The Bayesian version of the ridge penalty is given by the assumption of i.i.d. (conditional) Gaussian priors for the regression coefficients

$$\beta_j \mid \lambda \sim_{\text{iid}} N(0,1/2\lambda), \quad j = 1,\dots,p_x, \tag{4.2}$$

that leads to the joint prior density

$$p(\boldsymbol{\beta} \mid \lambda) = \prod_{j=1}^{p_x} p(\beta_j \mid \lambda) = \left(\sqrt{\frac{\lambda}{\pi}}\right)^{p_x} \exp\left\{-\lambda \sum_{j=1}^{p_x} \beta_j^2\right\}. \tag{4.3}$$

For a given value of the shrinkage parameter $\lambda \geq 0$, posterior mode estimation corresponds to maximum penalized likelihood estimation, compare (1.20). The prior (4.3) has the scale mixture of normals representation $\beta_j \mid \tau_{\beta_j}^2 \sim N(0;\tau_{\beta_j}^2)$ with $\tau_{\beta_j}^2 \mid \lambda \sim \delta_{1/2\lambda}(\tau_{\beta_j}^2)$. The symbol $\delta_a(t)$ denotes the Kronecker function which equals 1, if $t = a$, and 0, if $t \neq a$. A full Bayesian specification is obtained, when additionally the shrinkage parameter λ is assumed to be a random variable and is equipped with an appropriate prior. Due to conjugacy to the Gaussian family using a gamma prior,

$$\lambda \sim \text{Gamma}(h_{1,\lambda}, h_{2,\lambda}); \quad h_{1,\lambda}, h_{2,\lambda} > 0, \tag{4.4}$$

is convenient to support a Gibbs update for this parameter. The deterministic connection $\delta_{1/2\lambda}(\tau_{\beta_j}^2)$ between the shrinkage parameter and the variance parameters is leading to identical variance parameters $\tau_\beta^2 = \tau_{\beta_j}^2$ and an identical proportion of shrinkage for all regression coefficients. This somehow artificial notation of the hierarchy, with a gamma prior for the shrinkage parameter λ instead of an inverse gamma prior for the variance parameter τ_β^2, prevents the interpretation of λ as shrinkage parameter similar to the lasso prior. In summary we obtain, due to the identical variance parameters, a multivariate scale mixture of normals, compare e. g. Eltoft et al. (2006), and we express the hierarchy as

$$\beta_j \mid \tau_\beta^2 \sim N(0;\tau_\beta^2), \quad \tau_\beta^2 \mid \lambda \sim \delta_{1/2\lambda}(\tau_\beta^2), \tag{4.5}$$

to reflect the identical shrinkage also in the notation.

To obtain the univariate scale mixture of normals representation, we utilize regression coefficient specific shrinkage parameters λ_j resulting in the hierarchy

$$\beta_j \mid \tau_{\beta_j}^2 \sim N\left(0;\tau_{\beta_j}^2\right), \quad \tau_{\beta_j}^2 \mid \lambda_j \sim \delta_{1/2\lambda_j}\left(\tau_{\beta_j}^2\right), \tag{4.6}$$

with

$$\lambda_j \sim_{\text{iid}} \text{Gamma}\left(h_{1,\lambda}, h_{2,\lambda}\right); \quad h_{1,\lambda}, h_{2,\lambda} > 0. \tag{4.7}$$

In this representation each regression coefficient has a representation as scale mixture of normal distributions (4.1) with individual inverse gamma mixing distribution, which bypasses the identical proportion of shrinkage for all regression coefficients. We consider in the following the more general case (4.6) with (4.7), since the properties of the multivariate scale mixture (4.5) can be derived as a special case. More details are provided in Subsection 4.1.3.

4.1.2. Shrinkage properties

Marginal priors

To investigate the shrinkage properties, we consider the univariate marginal priors of the regression coefficients β_j and the associated variance parameter $\tau^2_{\beta_j}$, induced by the hierarchical prior structure given above. For the mixing variance parameter we obtain

$$\tau^2_{\beta_j} \mid h_{1,\lambda}, h_{1,\lambda} \sim \text{IGamma}\left(h_{1,\lambda}, \tfrac{1}{2} h_{2,\lambda}\right), \tag{4.8}$$

and further marginalization over the variance parameter, compare Appendix B.1, is leading to a scaled Student t-distribution as marginal distribution of the regression coefficients given the hyperparameters $h_{1,\lambda}, h_{2,\lambda}$

$$p\left(\beta_j \mid h_{1,\lambda}, h_{2,\lambda}\right) = \int_0^\infty N\left(\beta_j \mid 0, \tau^2_{\beta_j}\right) \text{IGamma}\left(\tau^2_{\beta_j} \mid h_{1,\lambda}, \tfrac{1}{2} h_{2,\lambda}\right) d\tau^2_{\beta_j} = t\left(\beta_j \mid 2h_{1,\lambda}, \sqrt{h_{2,\lambda}/2h_{1,\lambda}}\right), \tag{4.9}$$

with densities $p(\beta_j \mid d, s) = \Gamma(\tfrac{1}{2}(d+1)) / (\pi d s^2)^{\frac{1}{2}} \Gamma(\tfrac{1}{2} d) \left(1 + \beta_j^2 / d s^2\right)^{-\frac{1}{2}(d+1)}$, where $d = 2h_{1,\lambda}$ are the degrees of freedom and $s = \sqrt{h_{2,\lambda}/2h_{1,\lambda}}$ is the scale parameter. For $d = s^2 = 1$, i. e., $h_{1,\lambda} = 0.5$ and $h_{2,\lambda} = 1$, the standard Cauchy distribution is obtained as special case. Finally, the full Bayesian specification is leading to a marginal distribution of the regression coefficients, which has a representation as scale mixture of normal distributions with an inverse gamma mixing distribution.

The additional prior assumptions about the shrinkage parameters are leading to a more flexible modeling of our prior knowledge and a refinement of the prior tuning, i. e., the shrinkage of the regression coefficients is controlled by the two hyperparameters of the scaled Student t-distribution $p(\beta_j \mid h_{1,\lambda}, h_{2,\lambda})$ in comparison to the single parameter Gaussian prior $p(\beta_j \mid \lambda)$. The associated penalty function simply incorporates the term in the logarithm of the prior, $-\log p(\beta_j \mid h_{1,\lambda}, h_{2,\lambda})$, that depends on β_j and is given by $\text{pen}(\beta_j; h_{1,\lambda}, h_{2,\lambda}) = (h_{1,\lambda} + 0.5) \log(1 + \beta_j^2 / h_{2,\lambda})$. From the optimization perspective, the penalized ML or MAP estimate solves the penalized log-likelihood equations with respect to the penalty term $\text{pen}(\boldsymbol{\beta}; h_{1,\lambda}, h_{2,\lambda}) = (h_{1,\lambda} + 0.5) \sum_{j=1}^{p_x} \log(1 + \beta_j^2 / h_{2,\lambda})$.

Shrinkage properties in terms of the marginal prior of the regression coefficients

The shrinkage behavior of a prior distribution is determined by the specific shape of the density. A concentration of the probability mass around the origin enforces a strong shrinkage of the regression coefficients, while more mass in the tails of the density, enables larger values of the regression coefficients and supports the unbiasedness of the larger regression coefficient estimates. At the limit, for very noninformative, diffuse priors, the regression coefficients are distributed around their ML estimates. In the case of a scaled Student t-distribution the scale parameter equals 1, if $h_{2,\lambda} = 2h_{1,\lambda}$ in our parameterization. If $h_{2,\lambda} > 2h_{1,\lambda}$, the scale parameter is larger than one, $s > 1$, and more probability mass is allocated to the tails of the t-distribution and vice versa, if $h_{2,\lambda} < 2h_{1,\lambda}$, we have $s < 1$ and the probability mass is more concentrated around zero. Also the degrees of freedom $d = 2h_{1,\lambda}$ determine the shrinkage, since larger degrees of freedom concentrate the t-distribution around zero and induce a stronger shrinkage. In summary, the constellation of the two hyperparameters $h_{1,\lambda}$ and $h_{2,\lambda}$ controls the amount shrinkage of smaller regression coefficients and the size of the bias for larger regression coefficients.

Figure 4.1 shows the univariate marginal log-priors of one regression coefficient, $\log(p(\beta_j \mid \cdot))$, associated to the various regularization priors considered in this section. The used hyperparameters are selected to obtain $q_{0.05} = -4$ as the lower 5% quantile of each prior distribution, but in the case of more than one hyperparameter the selection is not unique. In particular the left panel of **Figure 4.1** contains, amongst others, the marginal Gaussian prior (4.2), denoted with $\text{BR}(\lambda = 0.169)$, and the marginal Student t-prior (4.9), denoted with $\text{BR}(h_{1,\lambda} = 0.45, h_{1,\lambda} = 0.248)$. In contrast to the light-tailed Gaussian distribution, the Student t-distribution has a more beneficial shape for regularization, since it shows a distinctive peak at zero and a strong decline, similar like the heavy tailed Cauchy distribution, $\text{BR}(h_{1,\lambda} = 0.5, h_{1,\lambda} = 1)$, if the absolute values of β increase. Therefore, the shrinkage of large coefficients towards 0 is only moderate, whereas shrinkage of small coefficients towards 0 is encouraged.

This is also shown in **Figure 4.2** in terms of the associated penalty functions.

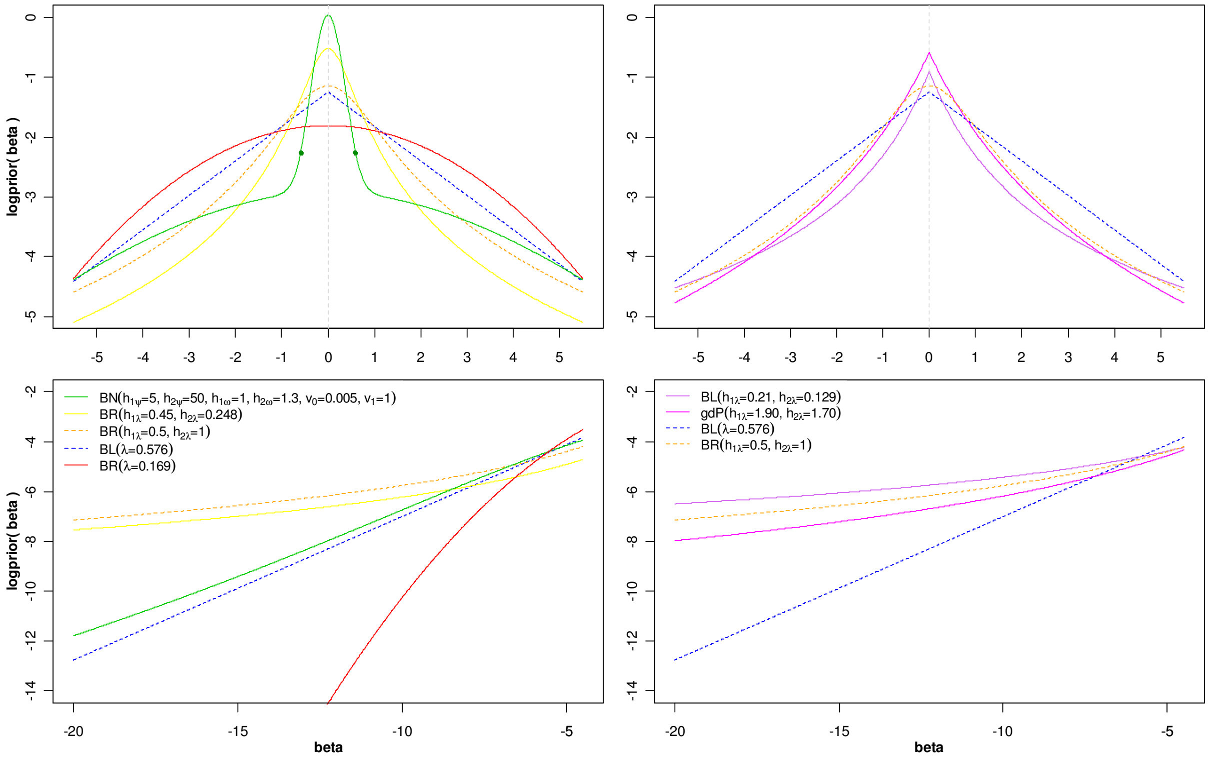


Figure 4.1: Univariate marginal log-priors of the regression coefficients $\log p(\beta \mid \cdot)$ resulting from the various regularization schemes. Upper panel: Log-priors in the range $[-5, 5]$. Lower panel: Log-priors in the left margin $[-20, -5]$. The hyperparameters given in the legends are selected to obtain $q_{0.05} = -4$ as the lower 5% quantile of the marginal prior of the regression coefficient.

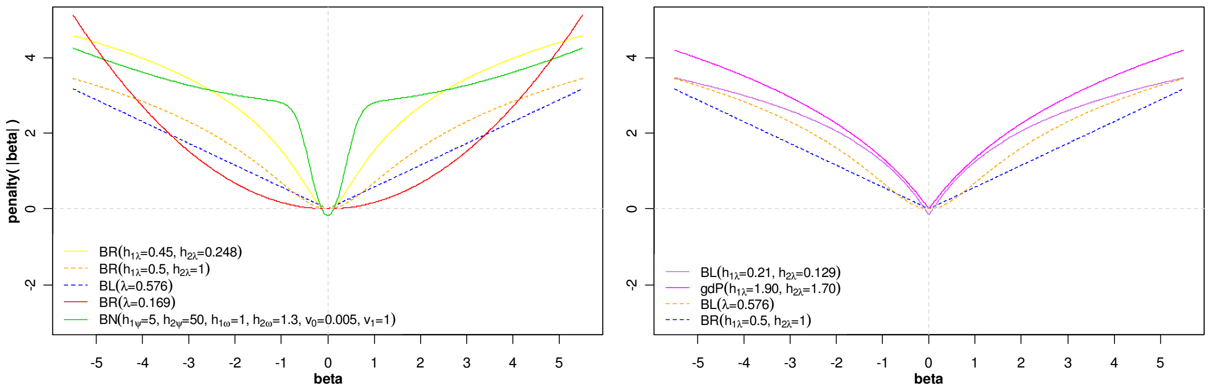


Figure 4.2: Univariate marginal penalty of the regression coefficients $\mathrm{pen}(|\beta||\cdot)$ resulting from the various regularization schemes. The hyperparameters given in the legends are selected to obtain $q_{0.05} = -4$ as the lower 5% quantile of the marginal prior of the regression coefficient.

Especially in the linear regression model with Gaussian error the amount of shrinkage can be quantified and expressed in terms of the ML-estimate. For simplicity we assume in this context orthogonal predictors. Solving the penalized score equations $\partial \log L_{pen}(\hat{\boldsymbol{\beta}}_{pen},\cdot)/\partial\boldsymbol{\beta} = \mathbf{0}$ is leading to the connection $\hat{\beta}_{j,ML} - \hat{\beta}_{j,pen} = \sigma^2 \mathrm{sign}(\hat{\beta}_{j,pen})\mathrm{pen}'(|\hat{\beta}_{j,pen}|;\cdot)$, where $\hat{\beta}_{j,pen}$ denotes the penalized estimate, $\mathrm{pen}'(|\beta_j|;\cdot) = -d\log p(|\beta_j||\cdot)/d\beta_j$ is the first derivate of the penalty function with respect to β_j and σ is the standard error of the ML estimate $\hat{\beta}_{j,ML}$. It follows, that the derivate of the penalty controls the amount of shrinkage and, if the derivate tends to zero for large values of $\hat{\beta}_{j,ML}$, the penalized estimate gets close to the ML estimate $\hat{\beta}_{j,pen} \approx \hat{\beta}_{j,ML}$ and is nearly unbiased. Fan and Li (2001) formulated several conditions to define a good penalty function. The *unbiasedness* of the resulting estimator, when the true unknown parameter is large, is one of the conditions and it is sufficient to show $\mathrm{pen}'(|\beta_j|;\cdot) \to 0$ for $|\beta_j|$ large. The resulting estimator is said to be a *thresholding rule* for the MAP estimation, if the minimum of the thresholding function $T(\beta_j) := |\beta_j| + \sigma^2\mathrm{pen}'(|\beta_j|)$ is positive, that is $TP := \min_{\beta_j \neq 0} T(\beta_j) > 0$. In this case the penalty is *sparse*, since the penalized estimate is set to $\hat{\beta}_{j,pen} = 0$, if $|\hat{\beta}_{j,ML}| \leq TP$, and the model complexity is reduced. For *continuity* of the penalized estimator in the data a necessary and sufficient condition is that the threshold TP is attained at 0. This avoids instability of the estimators as e. g. resulting from all subset selection. In summary, the penalty function must be singular at the origin.

Figure 4.3 shows the derivate of the univariate marginal log-priors of the regression coefficients $d\log p(\beta_j|\cdot)/d\beta_j$ at the positive x-axis and the derivate of penalty function $\mathrm{pen}'(|\beta_j|;\cdot)$ is obtained by reflection across the positive half of the x-axis. The penalty of the Gaussian prior (4.2) is given by $\mathrm{pen}(|\beta_j|;\lambda) = \lambda|\beta_j|^2$ with derivate $\mathrm{pen}'(|\beta_j|;\lambda) = 2\lambda|\beta_j|$. From this formula we can easily see the well known results, that the estimators resulting from this penalty are always biased, since the penalty does not converge towards zero for large $|\beta_j|$, and it is obvious that the minimum TP of the thresholding-function $T(\beta_j) = |\beta_j| + 2\lambda|\beta_j|\sigma^2(\mathbf{X}'\mathbf{X})^{-1}$ with respect to β_j is attained at zero, so this penalty does not produce sparse solutions. The penalty of Student t-prior (4.9), is leading to the derivate $\mathrm{pen}'(|\beta_j|;h_{1,\lambda},h_{,\lambda}) = (2h_{1,\lambda}+1)|\beta_j|(h_{2,\lambda}+\beta_j^2)^{-1}$, compare right panel of **Figure 4.3**, which converges to zero if $|\beta_j|$ gets large and, as a consequence, we obtain less biased estimates. But, due to the smoothness of the Student t-prior at the origin, the derivate of the log-prior is continuous at the origin, so that the minimum TP is attained at zero ($TP = 0$), and consequently there are no sparse solutions obtainable with this penalty resp. prior.

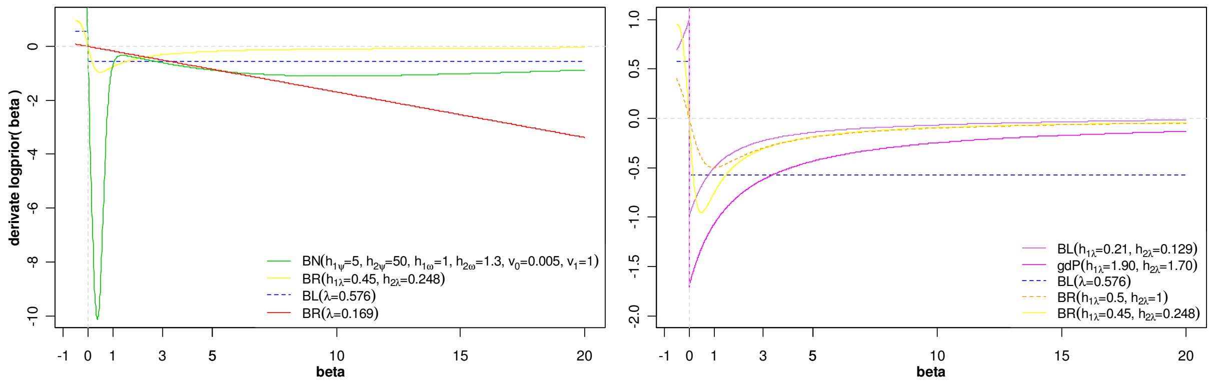


Figure 4.3: First derivate of the univariate marginal log-priors of the regression coefficients $d\log p(\beta\,|\,\cdot)/d\beta$ resulting from the various regularization schemes.The hyperparameters given in the legends are selected to obtain $q_{0.05}=-4$ as the lower 5% quantile of the marginal prior of the regression coefficient. The derivate of the penalty function is obtained by reflection across the x-axis.

Shrinkage properties in terms of the marginal prior of the variance parameters

An intuitive way to understand the shrinkage properties is provided by the analysis of the marginal prior of variance parameters $\tau^2_{\beta_j}$, since small variances $\tau^2_{\beta_j}\to 0$ induce a strong and large variances $\tau^2_{\beta_j}\to\infty$ induce a weak shrinkage of the regression coefficients β_j, with respect to the scale mixture representation. Carvalho et al. (2010) suggested the use of the standardized constraint parameter $\kappa_j := 1/(1+\tau^2_{\beta_j})\in[0,1]$ instead of the variance parameters to improve the comparison of various priors. The associated prior distribution of κ_j is derived by the density transformation $p(\kappa_j)=p_{\tau^2}(\frac{1}{\kappa_j}-1)\kappa_j^{-2}$, where $p_{\tau^2}(\cdot)$ denotes the marginal prior of the variance parameters $\tau^2_{\beta_j}$. The behavior of the prior $p(\kappa_j)$ close to $\kappa_j=1$ ($\tau^2_{\beta_j}\to 0$) controls the shrinkage of the smaller regression coefficients, while prior $p(\kappa_j)$ close to $\kappa_j=0$ ($\tau^2_{\beta_j}\to\infty$) controls the tail robustness of the prior.

For the inverse gamma distribution of the variance parameter in (4.8) we obtain the density

$$p(\kappa_j)=\frac{(0.5h_{2,\lambda})^{h_{1,\lambda}}}{\Gamma(h_{1,\lambda})}\frac{(\kappa_j)^{h_{1,\lambda}-1}}{(1-\kappa_j)^{h_{1,\lambda}+1}}\exp\left(-0.5h_{2,\lambda}\frac{\kappa_j}{1-\kappa_j}\right).$$

At the right limit $\kappa_j\to 1$ the prior is always zero, $p(\kappa_j)\to 0$, and at the left limit $\kappa_j\to 0$ the prior behavior depends on the hyperparameters $h_{1,\lambda}$ and $h_{2,\lambda}$. We obtain for $\kappa_j\to 0$ that $p(\kappa_j)\to 0$, if $h_{1,\lambda}>1$, $p(\kappa_j)\to 0.5h_{2,\lambda}$, if $h_{1,\lambda}=1$, and $p(\kappa_j)\to\infty$, if $h_{1,\lambda}<1$.

Figure 4.4 shows the prior of the parameter κ_j under various hyperparameter constellations. From the upper left panel to the lower right panel the hyperparameter $h_{1,\lambda}$ decreases, within the panels we sweep through the scale of the prior by decreasing the hyperparameter $h_{2,\lambda}$ with fixed value for $h_{1,\lambda}$. The magenta lines mark the values, where the scale parameter of the Student t-prior for the regression coefficients equals one. Within each panel, if the hyperparameter $h_{2,\lambda}$ decreases, more prior mass is assigned to the neighborhood of $\kappa_j\approx 1$ and shrinkage is enforced. The prior $p(\kappa_j)$ becomes strongly peaked near $\kappa_j=0$, if $h_{2,\lambda}$ is small enough. On the other hand, decreasing the hyperparameter $h_{1,\lambda}$ places more probability mass in the neighborhood of $\kappa_j\approx 0$ which promotes the tail robustness and we obtain infinite spikes at $\kappa_j=0$ for $h_{1,\lambda}<1$ in the lower panel. The density $p(\kappa_j)$ associated to the heavy tailed Cauchy prior is displayed in the lower left panel (magenta line). Here the prior $p(\kappa_j)$ equals zero at $\kappa_j=1$ and is unbounded at $\kappa_j=0$ with prior mass $\mathbb{P}(\kappa_j\in[0,0.25])\approx 0.53$ and $\mathbb{P}(\kappa_j\in[0.25,0.75])\approx 0.43$.

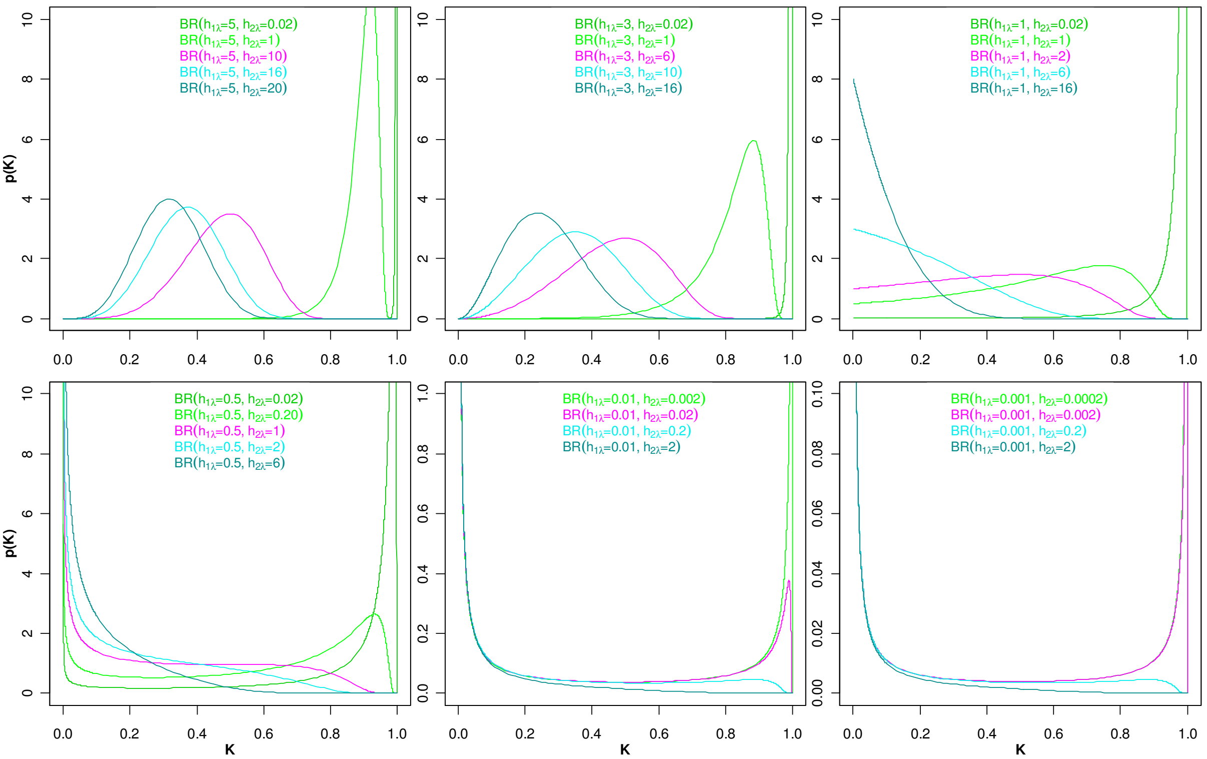


Figure 4.4: Prior densities of the standardized constraint parameter κ_j for the marginal variance prior (4.8) under various hyperparameter combinations given in the legends.The magenta line in the lower left panel corresponds to the Cauchy density as marginal prior of the regression coefficients.

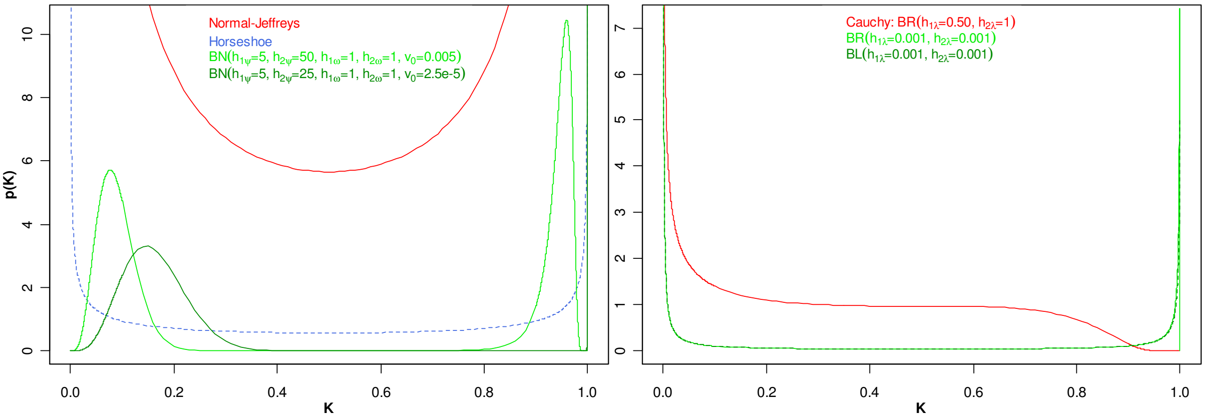


Figure 4.5: Prior densities of the standardized constraint parameter κ_j for various regularization priors. The hyperparameter combinations given in the legends are used in the simulations and applications.

Due to the trade-off between the magnitude of shrinkage and tail robustness, the hyperparameters have to be selected carefully. To enable data driven estimates, we specify diffuse gamma priors for the shrinkage parameter with small values of the hyperparameters $h_{1,\lambda}$ and $h_{2,\lambda}$. For our general settings $h_{1,\lambda} = h_{2,\lambda} = 0.01$ (with $\mathbb{P}(\kappa_j \in [0,0.25]) \approx 0.943$ and $\mathbb{P}(\kappa_j \in [0.25,0.75]) \approx 0.023$), $h_{1,\lambda} = h_{2,\lambda} = 0.001$ (with $\mathbb{P}(\kappa_j \in [0,0.25]) \approx 0.991$ and $\mathbb{P}(\kappa_j \in [0.25,0.75]) \approx 0.0025$) used in the simulations and applications, we obtain a lot of mass in the tails and an enhanced shrinkage for $\kappa_j \approx 1$. For the various shrinkage priors used in the simulations and applications, the distributions of the standardized constraint parameters are compared in **Figure 4.5** together with the horseshoe, Carvalho et al. (2010), and the Normal-Jeffrey prior.

4.1.3. Extensions

In Subsection 4.1.1 we have introduced two versions of the Bayesian ridge prior. One version corresponds to the prior hierarchy in (4.6) and (4.7)

$$\beta_j \mid \tau^2_{\beta_j} \sim \mathrm{N}\left(0;\tau^2_{\beta_j}\right), \quad \tau^2_{\beta_j} \mid \lambda_i \sim \delta_{1/2\lambda_j}\left(\tau^2_{\beta_j}\right), \quad \lambda_j \sim \mathrm{Gamma}\left(h_{1,\lambda}, h_{2,\lambda}\right) \tag{A}$$

and the other corresponds to (4.5) and (4.4)

$$\beta_j \mid \tau^2_{\beta} \sim \mathrm{N}\left(0;\tau^2_{\beta}\right), \quad \tau^2_{\beta} \mid \lambda \sim \delta_{1/2\lambda}\left(\tau^2_{\beta}\right), \quad \lambda \sim \mathrm{Gamma}\left(h_{1,\lambda}, h_{2,\lambda}\right). \tag{B}$$

The first version (A) leads to a marginal Student t-distribution for each regression coefficient β_j, as described above, and enables an individual shrinkage of each regression coefficient via the coefficient-specific variance parameter $\tau^2_{\beta_j}$. The joint prior of the regression coefficients $\boldsymbol{\beta}$ is the product of univariate t-densities. The second version (B) leads to a multivariate Student t-distribution as marginal prior of the regression coefficients $\boldsymbol{\beta}$, with $d = 2h_{1,\lambda}$ degrees of freedom and scale matrix $\boldsymbol{\Sigma}^{\frac{1}{2}} = \sqrt{h_{2,\lambda}/2h_{1,\lambda}}\,\mathbf{I}_{p_x}$, compare Appendix B.1, and induces an identical proportion of shrinkage for all regression coefficients, due to the common variance parameter τ^2_{β}. Both versions differ, because the product of univariate t-densities is not equivalent to a multivariate t-density. But, since the distributions of the marginal variance parameter(s) are in both cases inverse gamma distributions, $\mathrm{IGamma}(h_{1,\lambda}, \frac{1}{2}h_{1,\lambda})$, the analysis of the distribution of the standardized constraint parameter κ can also be applied to analyze the shrinkage behavior under version (B). We use version (B) throughout in the simulations and the applications.

In the upper panel of **Figure 4.6** the different shapes of both versions are shown in terms of the 2-dimensional log-priors of the regression coefficients. The adaptive ridge version (A) behaves similar as the lasso prior (4.15) with ridges at the axes, but conversely we have rounded edges at the axes under the ridge prior (A). In contrast the ridge version (B) produces elliptical contours, compare lower panel of **Figure 4.6**. In the software both versions of the ridge prior are implemented. Version (A) is specified as *"adaptive ridge"* method and (B) simply as *"ridge"*. The term *adaptive* indicates in general, that covariate-specific complexity parameters are used, where each can be equipped in addition with its own hyperparameters if desired. Covariate-specific hyperparameters enable to "stretch" or "compress" the marginal priors of the regression coefficients covariate-specific which increases again the flexibility of the joint prior. Such adaptations may be useful to take into account correlations of the covariates or various covariate scales. In contrast, e. g. the lasso prior (compare Subsection 4.2) is leading to covariate-specific shrinkage, even if a common prior for the shrinkage parameter is introduced.

Group priors

The simultaneous selection of associated covariate groups, arising e. g. from categorical covariates or from pathways representing predefined sets of interconnected genes, is also an important feature of regularization priors. Group sparsity can be handled by assuming identical variance parameters for the associated subgroups of regression coefficients, similar to the ridge version (B), to induce an identical proportion of shrinkage for all regression coefficients within the subgroup. More formal, let $\tilde{\boldsymbol{\beta}}_j = (\beta_{j,1},\ldots,\beta_{j,k_j})'$ denote the $j = 1,\ldots,p_x$ associated subgroups of regression coefficients with group size $k_j \geq 1$. For each group $j \in \{1,\ldots,p_x\}$ we use the hierarchical structure $\tilde{\boldsymbol{\beta}}_j \mid \tau^2_{\tilde{\beta}_j} \sim \mathrm{N}(0;\tau^2_{\tilde{\beta}_j}\mathbf{I}_{k_j})$, $\tau^2_{\tilde{\beta}_j} \mid \lambda_j \sim \delta_{1/2\lambda_j}(\tau^2_{\tilde{\beta}_j})$, $\lambda_j \sim_{\mathrm{iid}} \mathrm{Gamma}\left(h_{1,\lambda}, h_{2,\lambda}\right)$, where $\mathbf{I}_{k_j}$ denotes the k_j-dimensional identity matrix

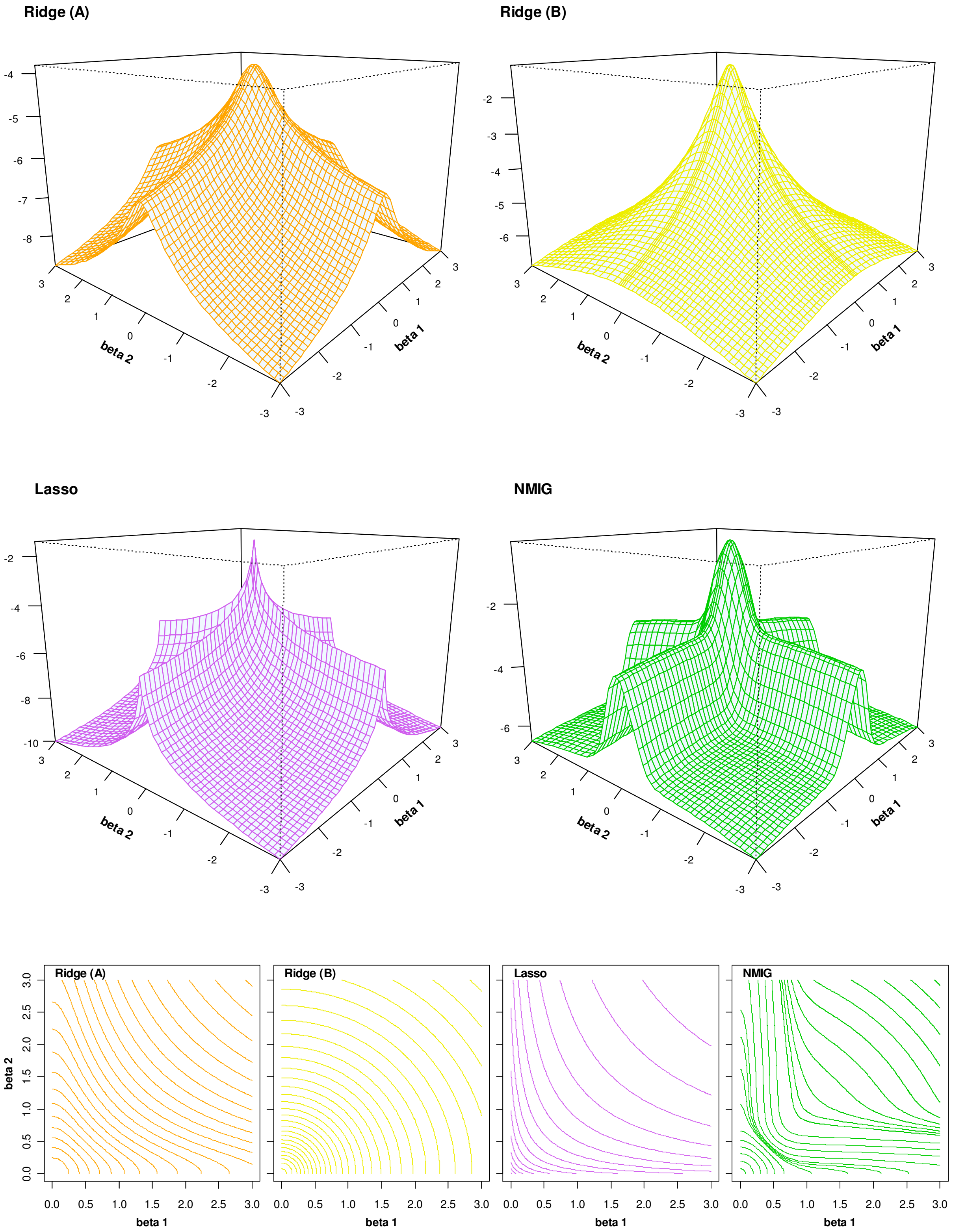


Figure 4.6: Marginal 2-dimensional log-priors of the regression coefficients, $\log p(\beta_1, \beta_2 \mid \cdot)$, and equicontours, $\log p(\beta_1, \beta_2 \mid \cdot) = \text{const}$, resulting from the Bayesian ridge version (A) and (B), the lasso and the NMIG regularization scheme.

and $\tau^2_{\beta_j}$ is the group-specific variance parameter. Concordantly to the ridge version (B), we obtain multivariate Gaussian scale mixtures for each group of associated regression coefficients $\tilde{\boldsymbol{\beta}}_j$ with marginal k_j-dimensional Student t-distributions, each having $d = 2h_{1,\lambda}$ degrees of freedom and a scale matrix $\boldsymbol{\Sigma}_j^{\frac{1}{2}} = \sqrt{h_{2,\lambda}/2h_{1,\lambda}}\mathbf{I}_{k_j}$. This version is not implemented in the software yet.

4.2. Bayesian lasso prior

4.2.1. Prior hierarchy

Just as well known as the ridge regression in the context of collinearity, is the lasso regression, Tibshirani (1996), if simultaneous variable selection and estimation should be achieved. The Bayesian version of the lasso penalty $\text{pen}(\boldsymbol{\beta},\lambda) = \lambda\sum_{j=1}^{p_x}|\beta_j|$ can be formulated with i.i.d. centered Laplace priors

$$\beta_j \mid \lambda \sim_{\text{iid}} \text{Laplace}(0,\lambda), \quad j = 1,...,p_x, \tag{4.10}$$

where $\lambda > 0$ represents the inverse scale parameter of the Laplace distribution, and joint density

$$p(\boldsymbol{\beta} \mid \lambda) = \prod_{j=1}^{p_x} p(\beta_j \mid \lambda) = \left(\frac{\lambda}{2}\right)^{p_x} \exp\left(-\lambda\sum_{j=1}^{p_x}|\beta_j|\right), \tag{4.11}$$

compare, e. g., Park and Casella (2008). **Figure 4.1** shows the Laplace prior, $\text{BL}(\lambda = 0.576)$, in the univariate case. As in ridge regression, for given values of λ, posterior mode estimation corresponds to penalized likelihood estimation.

The Laplace density $p(\beta_j \mid \lambda)$ is expressed as scale mixture of normals (4.1), with an exponential prior on the mixing variances

$$\beta_j \mid \tau^2_{\beta_j} \sim \text{N}\left(0;\tau^2_{\beta_j}\right), \quad \tau^2_{\beta_j} \mid \lambda^2 \sim_{\text{iid}} \text{Exp}\left(\tfrac{1}{2}\lambda^2\right). \tag{4.12}$$

For full Bayesian inference, we use in addition a gamma prior for the squared shrinkage parameter λ^2

$$\lambda^2 \sim \text{Gamma}\left(h_{1,\lambda},h_{2,\lambda}\right), \quad h_{1,\lambda},h_{2,\lambda} > 0, \tag{4.13}$$

where small values of the hyperparameters $h_{1,\lambda} > 0, h_{2,\lambda} > 0$ define diffuse gamma priors and allow data driven estimates of the model parameters.

4.2.2. Shrinkage properties

Marginal priors

The introduction of the hyperprior for the shrinkage parameter is leading to the following marginal density for the mixing variance parameter $\tau^2_{\beta_j}$

$$p\left(\tau^2_{\beta_j} \mid h_{1,\lambda},h_{2,\lambda}\right) = \int \text{Exp}\left(\tau^2_{\beta_j} \mid \tfrac{1}{2}\lambda^2\right)\text{Gamma}\left(\lambda^2 \mid h_{1,\lambda},b_{2,\lambda}\right)d\lambda^2 = \frac{h_{1,\lambda}}{2h_{2,\lambda}}\left[\frac{\tau^2_{\beta_j}}{2h_{2,\lambda}}+1\right]^{-(h_{1,\lambda}+1)}. \tag{4.14}$$

This is the density of generalized Pareto (gP) distribution ($p_{gP}(x;a,s,m) = s^{-1}(1+(x-m)/as)^{-a-1}$, $a > 0$, $x \geq m$), with zero location parameter $m = 0$, scale parameter $s = 2h_{2,\lambda}/h_{1,\lambda}$ and shape parameter $a = h_{1,\lambda}$. As mentioned before, conditionally on the variance parameter $\tau^2_{\beta_j}$, the prior for β_j

is Gaussian, but the marginal density of the regression coefficients is non Gaussian and can be expressed as

$$
\begin{aligned}
p\left(\beta_j \mid h_{1,\lambda}, h_{2,\lambda}\right) &= \int N\left(\beta_j \mid 0, \tau^2_{\beta_j}\right) \mathrm{gPareto}\left(\tau^2_{\beta_j} \mid h_{1,\lambda}, h_{2,\lambda}\right) d\tau^2_{\beta_j} \\
&= \frac{h_{1,\lambda}}{\sqrt{\pi}} \frac{2^{h_{1,\lambda}}}{\sqrt{2h_{2,\lambda}}} \Gamma\left(h_{1,\lambda} + 1/2\right) \exp\left(\frac{1}{4}\frac{\beta_j^2}{2h_{2,\lambda}}\right) D_{-2(h_{1,\lambda}+1/2)}\left(\frac{|\beta_j|}{\sqrt{2h_{2,\lambda}}}\right)
\end{aligned}
\qquad (4.15)
$$

with the parabolic cylinder Function $D_v(\cdot)$, compare Appendix B.2 for details to the derivation. In summary, the derived marginal distribution can be expressed as scale mixture of normals with a generalized Pareto mixing distribution. The hyperparameter $h_{2,\lambda}$ plays the role of a scale parameter in the marginal distribution of the regression coefficients, in particular the scale factor is given by $s = \sqrt{2h_{2,\lambda}}$. With respect to the unscaled distribution ($h_{2,\lambda} = 0.5$), smaller values $h_{2,\lambda} < 0.5$ concentrate more mass around zero and enforce the shrinkage, while larger values $h_{2,\lambda} > 0.5$ shift more mass to the tails of the distribution. Using the connection $D_v(0) = 2^{\frac{v}{2}} \pi^{\frac{1}{2}} \Gamma^{-1}(\frac{1-v}{2})$, we obtain at the origin $p(\beta_j = 0 \mid h_{1,\lambda}, h_{2,\lambda}) = h_{1,\lambda}(4h_{2,\lambda})^{-\frac{1}{2}} \Gamma(h_{1,\lambda} + \frac{1}{2}) \Gamma^{-1}(h_{1,\lambda} + 1)$. As reflected by this expression, the hyperparameter $h_{1,\lambda}$ determines the level of the prior at the abscissa $\beta_j = 0$ and larger values are leading to higher ordinates, which also enforce the shrinkage.

In contrast, e. g., to the ridge prior, the marginal prior (4.15) lacks a simple analytic form and the theoretical properties of the resulting shrinkage estimators are hard to access in terms of the parabolic cylinder function. Armagan et al. (2013) utilize a gamma prior for the shrinkage parameter, $\lambda \sim \mathrm{Gamma}(h_{1,\lambda}, h_{2,\lambda})$, which leads marginally to a generalized double Pareto (gdP) distribution as prior for the regression coefficients, i. e.

$$
p_{gdP}\left(\beta_j \mid h_{1,\lambda}, h_{2,\lambda}\right) = \int \mathrm{Laplace}\left(\beta_j \mid 0, \lambda\right) \mathrm{Gamma}\left(\lambda \mid h_{1,\lambda}, h_{2,\lambda}\right) d\lambda = \frac{h_{1,\lambda}}{2h_{2,\lambda}} \left[\frac{|\beta_j|}{h_{2,\lambda}} + 1\right]^{-(h_{1,\lambda}+1)}, \qquad (4.16)
$$

and the mixing scheme is interpreted as scale mixture of Laplace distributions. The simple analytical expression of the marginal prior enables the formulation of a compact penalty function with summands $\mathrm{pen}(\beta_j; h_{1,\lambda}, h_{2,\lambda}) = (h_{1,\lambda} + 1)\log(|\beta_j| + h_{2,\lambda})$ and $\mathrm{pen}'(|\beta_j|; h_{1,\lambda} h_{2,\lambda}) = (h_{1,\lambda} + 1)/(|\beta_j| + h_{2,\lambda})$ as first derivate to study the properties of the resulting posterior mode estimator. Armagan et al. (2013) show, in the spirit of Fan and Li (2001), that the MAP estimator resulting from this penalty function is continuous in the data, nearly unbiased, if the absolute value of the true parameter $|\beta_j|$ is large, and that small estimated coefficients are set to zero, if $h_{2,\lambda} < 2\sqrt{h_{1,\lambda} + 1}$, i. e. the prior reduces the model complexity. Lee et al. (2012) showed that the gdP-prior, the exponential power prior and the Student t-prior can be viewed as special cases of a generalized t-prior with four hyperparameters and investigate the shrinkage and selection properties in this general framework.

In the frequentist context, Zou (2006) shows similar oracle properties for the adaptive lasso, where covariate-specific weights w_j are introduced in the penalization term $\mathrm{pen}(|\beta_j|; \lambda, w_j) = \lambda w_j |\beta_j|$ of the regression coefficients. This leads to coefficient-specific penalties $\lambda_j = \lambda w_j$ in comparison to the frequentist lasso, $\mathrm{pen}(|\beta_j|; \lambda) = \lambda|\beta_j|$, with its uniform shrinkage of all coefficients. The author state, that under an appropriate choice of the weights w_j, e. g. as the inverse ML estimates, the adaptive lasso can asymptotically perform as well as if the correct submodel was known.

Shrinkage properties in terms of the marginal prior of the regression coefficients

The right panel of **Figure 4.1** shows the marginal priors of the regression coefficients for the Bayesian lasso prior (4.15), $\mathrm{BL}(h_{1,\lambda}=0.21, h_{2,\lambda}=0.129)$, and the generalized double Pareto prior (4.16), $\mathrm{gdP}(h_{1,\lambda}=1.90, h_{2,\lambda}=1.70)$ in the univariate case. By trend, both priors behave very similar. Compared to the one parameter Laplace-prior, $\mathrm{BL}(\lambda=0.576)$, we obtain also peaks around zero and non continuous first derivates at the origin, but the two hyperparameters enable shapes, which assigns more probability mass to the tails. Under the selected hyperparameter constellations the gdP-prior is slightly more concentrated around zero, which results in (marginally) lighter tails. But both prior tails are almost comparable to the Cauchy- or the Bayesian ridge prior tails. Comparing the marginal log-priors from the Bayesian lasso hierarchy, $\mathrm{BL}(h_{1,\lambda}=0.21, h_{2,\lambda}=0.129)$, and the Bayesian ridge hierarchy, $\mathrm{BR}(h_{1,\lambda}=0.45, h_{2,\lambda}=0.248)$, we see that both approaches are not so far from each other as in their one-parameter versions, $\mathrm{BL}(\lambda=0.576)$ and $\mathrm{BR}(\lambda=0.169)$, with marginal Laplace and Gaussian prior. **Figure 4.6** displays the 2-dimensional shape of the lasso prior (4.15) and the associated equicontours. The contours are similar the L_q-penalty with $q<1$.

In the univariate case the contribution to the lasso penalty arising from the Laplace prior (4.10) is given by $\mathrm{pen}(\beta_j;\lambda)=\lambda|\beta_j|$, see **Figure 4.2**, with derivate $\mathrm{pen}'(|\beta_j|;\lambda)=\lambda$, see **Figure 4.3**. The Bayesian lasso penalty has the contributions $\mathrm{pen}\left(\beta_j;h_{1,\lambda}h_{2,\lambda}\right)=\beta_j^2/8h_{2,\lambda}+\log D_{-2(h_{1,\lambda}+1/2)}\left(|\beta_j|/\sqrt{2b_\lambda}\right)$, with derivate

$$\mathrm{pen}'\left(|\beta_j|;h_{1,\lambda},h_{2,\lambda}\right)=\frac{\left(2h_{1,\lambda}+1\right)}{\sqrt{2h_{2,\lambda}}}\frac{D_{-2\left(h_{1,\lambda}+1\right)}\left(\left|\beta_j\right|/\sqrt{2h_{2,\lambda}}\right)}{D_{-2\left(h_{1,\lambda}+1/2\right)}\left(\left|\beta_j\right|/\sqrt{2h_{2,\lambda}}\right)},$$

compare Appendix B.2. The right panel of **Figure 4.3** shows the first derivate of the univariate marginal log-priors of the regression coefficients, $d\log p(\beta_j|\cdot)/d\beta_j$, for the lasso variants. In contrast to the Laplace prior, the Bayesian lasso and gdP prior do not inherit the problem of overshrinking large coefficients, since the derivates vanish if $|\beta_j|$ increases, resulting in a reduction of bias. In contrast to the Bayesian ridge or NMIG (Subsection 4.3) regularization, the Bayesian lasso regularization (also Laplace and gdP) produce a nonzero derivate of the penalty at the origin $\beta_j=0$, because the priors are not continuous differentiable there. With respect to the thresholding function $T(\beta_j)$, the derivate of the penalty evaluated at the origin, $\mathrm{pen}'\left(|\beta_j|=0|\cdot\right)$, determines the threshold $TP>0$ in the linear model with orthogonal predictors, and the MAP resp. penalized ML estimates with $|\hat{\beta}_{ML,j}|<TP$ are set to zero. We obtain at the origin for the Bayesian lasso prior $\mathrm{pen}'\left(0|h_{1,\lambda},h_{2,\lambda}\right)=\sqrt{h_{2,\lambda}^{-1}}\Gamma(h_{1,\lambda}+1)\Gamma^{-1}(h_{1,\lambda}+\frac{1}{2})$, for the gdP prior $\mathrm{pen}'(0;h_{1,\lambda}h_{2,\lambda})=(h_{1,\lambda}+1)/h_{2,\lambda}$ and $\mathrm{pen}'(0;\lambda)=\lambda$ for the Laplace prior.

Shrinkage properties in terms of the marginal prior of the variance parameters

For the generalized Pareto distribution of the variance parameter $\tau_{\beta_j}^2$ in (4.14) we obtain the density

$$p(\kappa_j)=\frac{h_{1,\lambda}}{2h_{2,\lambda}}\frac{1}{\kappa_j^2}\left[\frac{1}{2h_{2,\lambda}}\frac{1-\kappa_j}{\kappa_j}+1\right]^{-(h_{1,\lambda}+1)}=\frac{h_{1,\lambda}}{2h_{2,\lambda}}\kappa_j^{h_{1,\lambda}-1}\left[\frac{1-\kappa_j}{2h_{2,\lambda}}+1\right]^{-(h_{1,\lambda}+1)}$$

for the standardized constraint parameter κ_j. At the right margin $\kappa_j\to1$ we have always finite nonzero values $p(\kappa_j)\to h_{1,\lambda}/2h_{2,\lambda}$. At the left margin $\kappa_j\to0$ we obtain for the prior $p(\kappa_j)\to0$, if $h_{1,\lambda}>1$, $p(\kappa_j)\to(2h_{2,\lambda}(1+1/2h_{2,\lambda})^2)^{-1}$, if $h_{1,\lambda}=1$ and $p(\kappa_j)\to\infty$, if $h_{1,\lambda}<1$. The influence of the hyperparameters $h_{1,\lambda}$ and $h_{2,\lambda}$ is visualized in **Figure 4.7**, which shows the prior of the parameter κ_j under various hyperparameter constellations. From the upper left panel to the lower right panel the hyperparameter $h_{1,\lambda}$ decreases, within the panels the hyperparameter $h_{2,\lambda}$ varies with constant value for $h_{1,\lambda}$. Increasing $h_{2,\lambda}$ shifts more probability mass to the right support of κ_j. The magenta colored densities result if $h_{1,\lambda} = h_{2,\lambda}$. For $h_{1,\lambda} < 1$ and small values $h_{2,\lambda}$ we obtain a horseshoe like shape for the prior $p(\kappa_j)$, i. e. we have high probabilities at the right margin, which determine the shrinkage, and at the left margin, which determine the tail behavior. The shapes of $p(\kappa_j)$ under the ridge and the lasso prior are almost comparable for small values of the hyperparameters, with the exception that under the ridge prior $p(\kappa_j)$ vanishes at $\kappa_j = 1$. We obtain for the two hyperparameter settings $h_{1,\lambda} = h_{2,\lambda} = 0.01$ ($\mathbb{P}(\kappa_j \in [0,0.25]) \approx 0.951$ and $\mathbb{P}(\kappa_j \in [0.25,0.75]) \approx 0.023$) and $h_{1,\lambda} = h_{2,\lambda} = 0.001$ ($\mathbb{P}(\kappa_j \in [0,0.25]) \approx 0.992$ and $\mathbb{P}(\kappa_j \in [0.25,0.75]) \approx 0.0025$), that are used in the simulations and applications, a lot of mass in the tails and an enhanced shrinkage near $\kappa_j \approx 1$. The resulting distributions of the standardized constraint parameters are compared in **Figure 4.5** and we see that the densities $p(\kappa_j)$ under the lasso and ridge prior almost coincide.

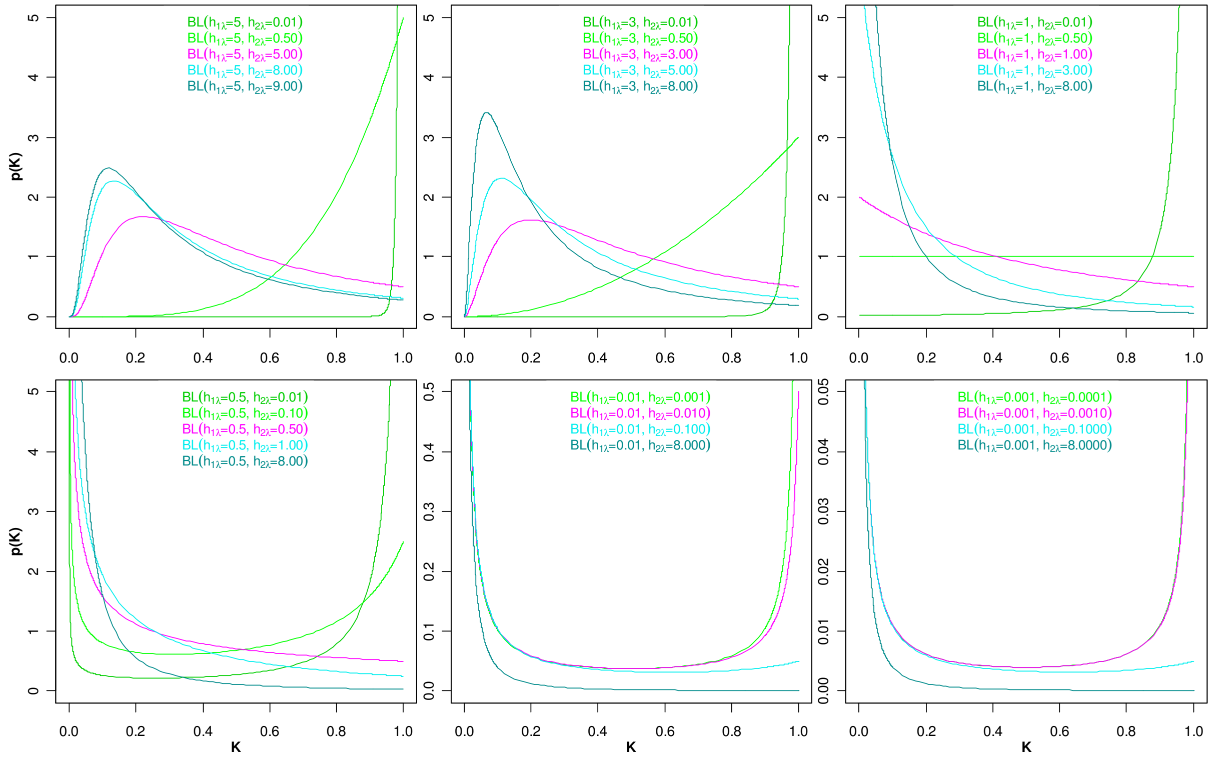


Figure 4.7: Prior densities of the standardized constraint parameter κ_j for the marginal variance prior (4.14) under various hyperparameter combinations given in the legends.

4.2.3. Extensions

Group regularization: We can modify the hierarchical structure of the Bayesian lasso, similar as in Subsection 4.1.3, to obtain a common regularization for an associated group of covariates. Assuming a common variance parameter within each k_j-dimensional group results in a multivariate Gaussian scale mixture representation $\tilde{\boldsymbol{\beta}}_j \sim N(0, \tau^2_{\tilde{\beta}_j} \mathbf{I}_{k_j})$ with $\tau^2_{\tilde{\beta}_j} \mid \lambda^2 \sim_{\text{iid}} \text{Gamma}(\tfrac{1}{2}(k_j + 1), \tfrac{1}{2}\lambda^2)$ for the j-th group of associated regression coefficients $\tilde{\boldsymbol{\beta}}_j$ and marginally in a multivariate Laplace-distribution of the regression coefficients $\tilde{\boldsymbol{\beta}}_j$, compare Kyung et al. (2010) for details. The provided MCMC sampling methods for Bayesian inference can easily be extended to consider group sparsity, but this not implemented yet.

Adaptive priors: To achieve more flexibility, we can equip the hierarchical models above with covariate-specific shrinkage parameters and the resulting models are additionally named with *"adaptive"*. For example, the adaptive version of the lasso prior is given through $\tau_j^2 \mid \lambda_j^2 \sim \mathrm{Exp}\left(\lambda_j^2/2\right)$ with $\lambda_j^2 \sim_{\mathrm{iid}} \mathrm{Gamma}\left(h_{1,\lambda}, b_{2,\lambda}\right)$. It is straightforward to use also covariate-specific hyperparameters, $h_{1,\lambda_j}, b_{2,\lambda_j}$, which can e. g. be utilized, if the covariates are not standardized, to take account for different scales. However, one should keep in mind, that the number of parameters to estimate is increased in the adaptive versions, which can cause problems in situations with low sample sizes. The adaptive versions can be specified in the software if desired, compare Appendix D.3 to D.5.

4.3. Bayesian NMIG prior

4.3.1. Prior hierarchy

Finally, we consider a normal mixture of inverse gamma distributions, shortly named as NMIG prior. This prior has been suggested by Ishwaran and Rao (2003) for the regularization of high-dimensional linear regression models. The conditional prior distribution for the regression coefficients is Gaussian, as in the lasso and ridge case,

$$\beta_j \mid I_j, \psi_j^2 \sim \mathrm{N}\left(0; \tau_{\beta_j}^2 = I_j \psi_j^2\right), \tag{4.17}$$

but in contrast, the variance parameters $\tau_{\beta_j}^2$ are specified through a spike and slab mixture distribution, modeled by the product of the two components

$$I_j \mid v_0, v_1, \omega \sim_{\mathrm{iid}} \mathrm{Bernoulli}(\omega; v_0, v_1), \quad \psi_j^2 \mid h_{1,\psi}, h_{1,\psi} \sim_{\mathrm{iid}} \mathrm{IGamma}\left(h_{1,\psi}, h_{2,\psi}\right). \tag{4.18}$$

The first component in (4.18) is a Bernoulli distributed indicator variable I_j with point mass at the values $v_0 > 0$ and $v_1 > 0$. In particular the parameter v_0 should have a positive value close to zero, to induce small variances, but we assume $v_0 \neq 0$ to avoid degenerated priors. The value of v_1 should be large compared to v_0 and we can use, e. g., $v_1 = 1$. The binary indicator variable takes the value v_0 with probability $\mathbb{P}(I_j = v_0) = 1 - \omega$ and v_1 with probability $\mathbb{P}(I_j = v_1) = \omega$. Since the parameter ω controls how likely the binary variable I_j equals v_1 or v_0, it takes on the role of a complexity parameter which controls the size of the models. The assumptions in (4.18) are leading to a continuous, bimodal distribution for the variance parameter $\tau_{\beta_j}^2 := I_j \psi_j^2$ given the hyperparameters v_0, v_1, $h_{1,\psi}$, $h_{2,\psi}$, ω. In particular, we obtain a mixture of scaled inverse gamma distributions

$$\tau_{\beta_j}^2 \mid v_0, v_1, h_{1,\psi}, h_{2,\psi}, \omega \sim (1-\omega) \cdot \mathrm{IGamma}\left(h_{1,\psi}, v_0 h_{2,\psi}\right) + \omega \cdot \mathrm{IGamma}\left(h_{1,\psi}, v_1 h_{2,\psi}\right), \tag{4.19}$$

with common shape parameter $h_{1,\psi}$ and scale parameters $v_0 h_{2,\psi}$ and $v_1 h_{2,\psi}$, compare Appendix B.3. The priors in (4.18) can alternatively be derived on the base of the mixture distribution (4.19) using the data augmentation approach depicted in Section 3. In the first mixture component, the so-called *spike*, probability mass is strongly concentrated on small values of the variances and in the second mixture component, the so-called *slab*, we obtain a more diffuse distribution with probability mass on a wide support for larger variance values. Variance parameters arising from the *spike* component induce a strong shrinkage of the regression coefficients, while variance parameters from the *slab* component enforce a reduced shrinkage. The prior locations of the two modes of the inverse gamma mixture components are independent of ω and fixed at

$$\mathrm{mode}_{v_0} = v_0 \frac{h_{2,\psi}}{h_{1,\psi}+1}, \quad \mathrm{mode}_{v_1} = v_1 \frac{h_{2,\psi}}{h_{1,\psi}+1}.$$

The shape and scale hyperparameter of the inverse gamma distributed variances ψ_j^2 determine a basic location of the mode, which is then adjusted by the values of v_0 and v_1. We select the hyperparameters $h_{1,\psi}$ and $h_{2,\psi}$ with respect to $h_{1,\psi} < h_{2,\psi}$, to enforce a basic mode $h_{2,\psi}/(h_{1,\psi}+1) \gg 1$. In addition, we assume a beta prior for the complexity parameter ω

$$\omega \sim \mathrm{Beta}\left(h_{1,\omega}, h_{2,\omega}\right), \tag{4.20}$$

with mean $\mathbb{E}(\omega \mid h_{1,\omega}, h_{2,\omega}) = h_{1,\omega}/(h_{1,\omega}+h_{2,\omega}) =: H_\omega$. The beta prior reduces to the uniform prior in the special case $h_{1,\omega} = h_{2,\omega} = 1$, which enables to express an indifferent prior knowledge about the model complexity. With an appropriate choice of the hyperparameters $h_{1,\omega}, b_{2,\omega} > 0$ it is possible to favor more or less sparse models. In particular, for overparameterized models sparse solutions can be enforced by choosing $h_{1,\omega} < b_{2,\omega}$.

In the context of Bayesian variable selection George and McCulloch (1993) use a mixture prior (SSVS prior) at the higher level of the regression coefficients, i. e. $\beta_j \mid I_j \sim (1-I_j)N(0,\tau_{0,j}^2) + I_j N(0,\tau_{1,j}^2)$, with $I_j \in \{0,1\}$ and pre-specified small and large values of the variance parameters $\tau_{0,j}^2$ and $\tau_{1,j}^2$. The NMIG prior mimics this variable selection strategy, since variances $\tau_{\beta_j}^2 = v_0\psi_j^2$ from the spike component of the mixture (4.19) induce a strong shrinkage similar $\tau_{0,j}^2$, whereas variances $\tau_{\beta_j}^2 = v_1\psi_j^2$ from the slab component of the mixture (4.19) support less biased estimates for relevant covariates similar to $\tau_{1,j}^2$. The full Bayesian model specification avoids the direct selection of the values for the variance parameters $\tau_{0,j}^2$ and $\tau_{1,j}^2$ due to utilizing hyperpriors for ψ_j^2. It facilitates an absolutely continuous prior, due to $v_0 \neq 0$, and straightforward Gibbs sampling to update the components I_j and ψ_j^2 simultaneously with the update of the complexity parameter ω. A common feature of the NMIG prior and the SSVS prior is that the regression coefficients can be rated due to their relevance for prediction. Since the sampled indicator values $I_j = v_1$ contain the main information for variable selection, we can utilize the posterior relative frequencies of the state $I_j = v_1$ as relevance measure for the covariate rating. Finally, variable selection is practiced by utilizing a threshold rule for the posterior relative frequencies of the indicators $I_j = v_1$, compare Subsection 4.4. Nevertheless, we have, in contrast to the lasso and the ride prior, an increased number of six hyperparameters to manipulate the shape of the marginal prior of the regression coefficients, and the costs for the obtained flexibility is an enhanced tuning effort. To guide the specification of the hyperparameters, we investigate in the following how the hyperparameters affect the shape of the marginal priors and the shrinkage properties. Further we analyze the behavior of the inclusion probability for a covariate with $I_j = v_1$ in terms of the associated full conditional.

4.3.2. Shrinkage properties

Marginal priors

The marginal density for the mixing variance parameters, after integrating out the parameter ω, is the mixture of two inverse gamma distributions

$$\tau^2_{\beta_j} \mid \cdot \sim \frac{h_{2,\omega}}{h_{1,\omega}+h_{2,\omega}} \cdot \mathrm{IGamma}(h_{1,\psi}, v_0 h_{2,\psi}) + \frac{h_{1,\omega}}{h_{1,\omega}+h_{2,\omega}} \cdot \mathrm{IGamma}(h_{1,\psi}, v_1 h_{2,\psi}), \tag{4.21}$$

which corresponds to the conditional variance density (4.19) for ω fixed at the prior mean ($\omega = H_\omega$). Due to the high number or hyperparameters, we use the dot abbreviation to denote the set $\{v_0, v_1, h_{1,\psi}, h_{2,\psi}, h_{1,\omega}, h_{2,\omega}\}$ in the notation of the marginal distributions. The marginal distribution for the regression coefficients β_j is a mixture of two scaled Student t-distributions

$$\beta_j \mid \cdot \sim \frac{h_{2,\omega}}{h_{1,\omega}+h_{2,\omega}} t\left(2h_{1,\psi}, \sqrt{\frac{v_0 h_{2,\psi}}{h_{1,\psi}}}\right) + \frac{h_{1,\omega}}{h_{1,\omega}+h_{2,\omega}} t\left(2h_{1,\psi}, \sqrt{\frac{v_1 h_{2,\psi}}{h_{1,\psi}}}\right), \tag{4.22}$$

with $d = 2h_{1,\psi}$ degrees of freedom and the scale parameters $s_0 = \sqrt{v_0 h_{2,\psi}/h_{1,\psi}}$ and $s_1 = \sqrt{v_1 h_{2,\psi}/h_{1,\psi}}$, compare Appendix B.3 for the derivation. We obtain as marginal inclusion probability of a covariate given β_j the expression

$$\mathbb{P}(I_j = v_1 \mid \beta_j, \cdot) = \frac{1}{1 + \frac{h_{2,\omega}}{h_{1,\omega}} \left(\frac{v_0}{v_1}\right)^{h_{1,\psi}} \left(\frac{\beta_j^2 + 2v_0 h_{2,\psi}}{\beta_j^2 + 2v_1 h_{2,\psi}}\right)^{-(h_{1,\psi}+0.5)}}. \tag{4.23}$$

Shrinkage properties in terms of the marginal prior of the regression coefficients

The left side of **Figure 4.1** shows the univariate marginal log-prior of the regression coefficients resulting from the NMIG regularization scheme. We have a finite rounded "spike" at the origin and observe a clear concentration of the log-prior around zero, which is stronger than under the other regularization priors. The slope of the prior gets large within in the region $[-1,1]$ and outside the prior it is comparatively flat and initially indicates a reduction of the shrinkage, separately and compared to the other priors. An inspection of the tails exposes over a broad range a prior behavior similar to the Laplace prior, but the differences become larger as $|\beta|$ increases and the log-prior is flattened. In this region the NMIG prior indicates a stronger shrinkage compared to the Bayesian lasso or ridge prior.

The two points on the log-prior mark the intersection points, where the weighted "spike" component of the mixture distribution (4.22) coincides with the weighted "slab" component. Both intersection points are located at the roots

$$ISP_\beta = \pm \sqrt{\frac{2v_0 h_{2,\psi} \left(h_{1,\omega}\right)^{\frac{2}{2h_{1,\psi}+1}} \left(v_1\right)^{\frac{2h_{1,\psi}}{2h_{1,\psi}+1}} - 2v_1 b_{2,\psi} \left(h_{2,\omega}\right)^{\frac{2}{2h_{1,\psi}+1}} \left(v_0\right)^{\frac{2h_{1,\psi}}{2h_{1,\psi}+1}}}{\left(h_{2,\omega}\right)^{\frac{2}{2h_{1,\psi}+1}} \left(v_0\right)^{\frac{2h_{1,\psi}}{2h_{1,\psi}+1}} - \left(h_{1,\omega}\right)^{\frac{2}{2h_{1,\psi}+1}} \left(v_1\right)^{\frac{2h_{1,\psi}}{2h_{1,\psi}+1}}}}.$$

This expression is clearly simplified, if $v_1 = 1$ and a uniform prior for the complexity parameter is used, i. e. $h_{1,\omega} = h_{2,\omega} = 1$. Within the range $[ISP_\beta, -ISP_\beta]$ the spike component dominates the slab component and the location of the intersection points can guide the hyperparameter selection. In addition, it turns out that at the intersection points ISP_β the marginal prior inclusion probability (4.23) always equals $1/2$, i. e. $\mathbb{P}(I_j = v_1 \mid \beta_j = \mp ISP_\beta, \cdot) = 0.5$. That means, regression coefficients outside the interval $[ISP_\beta, -ISP_\beta]$ have a higher prior inclusion probability than $1/2$ and those within have a lower inclusion probability. At the origin $\beta_j = 0$ we obtain $\mathbb{P}(I_j = v_1 \mid \beta_j = 0, \cdot) = (1 + h_{2,\omega}\sqrt{v_1}/h_{1,\omega}\sqrt{v_0})^{-1}$.

Figure 4.6 shows the 2-dimensional marginal log-prior of the regression coefficients and the associated equicontours in comparison to the lasso and ridge prior. Close to the origin we observe typical ridge (A) type contours, since the spike part of the log-prior dominates here. Then moving along the diagonal, if $\beta_1 = \beta_2 > 0.5$, we observe near the axes contours similar to L_q-penalty, with $q \ll 1$ and rounded corners. Regression coefficients in this areas are mainly shrunken towards one of the two axes, i. e. the β_1-axis, if $\beta_2 < \beta_1$, or vice versa towards the β_2-axis, if $\beta_2 > \beta_1$, but the effective direction depends on the shape of the log-likelihood contours, the size of the regression coefficients and their correlation, compare e. g. Konrath (2007) for a detailed 2-dimensional visualization in terms of the lasso and ridge prior. Moving further along the diagonal, if $\beta_1 = \beta_2 > 1$, we observe the area of the t-distributed slab part with initially convex contours with a reduced ridge (A) type shrinkage, when both regression coefficient components are not too close to the axes. The transition to the region, where the contours become concave, is outside the plotting area. In the univariate case we obtain the first derivate on the marginal log-prior as

$$\frac{d\log p(\beta_j;\cdot)}{d\beta_j} = -\frac{\dfrac{(1-H_\omega)(2h_{1,\psi}+1)\beta_j}{\sqrt{2v_0h_{2,\psi}}^{\,3}}\left(1+\dfrac{\beta_j^2}{2v_0h_{2,\psi}}\right)^{-\frac{2h_{1,\psi}+3}{2}} + \dfrac{H_\omega(2h_{1,\psi}+1)\beta_j}{\sqrt{2v_1h_{2,\psi}}^{\,3}}\left(1+\dfrac{\beta_j^2}{2v_1h_{2,\psi}}\right)^{-\frac{2h_{1,\psi}+3}{2}}}{\dfrac{1-H_\omega}{\sqrt{2v_0h_{2,\psi}}}\left(1+\dfrac{\beta_j^2}{2v_0h_{2,\psi}}\right)^{-\frac{2h_{1,\psi}+1}{2}} + \dfrac{H_\omega}{\sqrt{2v_1h_{2,\psi}}}\left(1+\dfrac{\beta_j^2}{2v_1h_{2,\psi}}\right)^{-\frac{2h_{1,\psi}+1}{2}}}. \tag{4.24}$$

We can easily proof that the derivate of the penalty $\text{pen}'(|\beta_j|;\cdot) = -d\log p(|\beta_j||\cdot)/d\beta_j$ converges towards zero for large regression coefficients $|\beta_j| \to \infty$, so that large estimates are less biased, but the convergence is slow compared to the other regularization schemes as shown in the right panel of **Figure 4.3** in terms of the derivate of the log-prior. For small coefficients we observe the clear regularization from the spike component, which quickly decreases with increasing values of $|\beta_j|$. For medium values $|\beta_j|$ around 1 the penalization is reduced, but the amount of penalization increases again with the transition to the slab component and reduces there slowly for increasing values $|\beta_j| \to \infty$.

Figure 4.8 to **Figure 4.10** display the marginal prior of the inclusion probabilities (4.23), the corresponding marginal log-prior of the regression coefficients (4.22) and the first derivate (4.24) on the right half for $\beta_j \in [0,3]$ under the variation of the hyperparameters. In the upper panels we have $h_{1,\psi} = 2.5$ and in the lower panels $h_{1,\psi} = 5$. From the left to the right $h_{2,\psi}$ varies with values $h_{2,\psi} = 25, 50, 100$. Within each panel the parameter v_0 increases with values from $v_0 = 2.5\text{e-}5$ (yellow) to $v_0 = 0.005$ (magenta). Overall the value of v_1 is fixed to 1 and we set $h_{1,\omega} = h_{2,\omega} = 1$. Within each panel of **Figure 4.8** we observe that the prior inclusion probability becomes larger, at fixed values of the regression coefficients, with decreasing values of v_0. Simultaneously, the prior inclusion probability of very small or zero effects becomes smaller and we obtain at the origin the value $\mathbb{P}(I_j = v_1 | \beta_j = 0, \cdot) = (1 + 1/\sqrt{v_0})^{-1}$. This implies in terms of the log-prior of the regression coefficients, $\log p(\beta_j | \cdot)$, that the intersection point of the two mixture components is shifted towards the origin, if v_0 is decreased and the log-prior becomes more and more concentrated, compare **Figure 4.9**. The derivate $d\log p(\beta_j|\cdot)/d\beta_j$ in **Figure 4.10** shows accordingly that lager effects get less penalized and that the penalty of small effects increases. If we move from the left to the right panels in the figures we observe that the decrease of the scale component $h_{2,\psi}$ is leading to the same effects on the displayed quantities as for decreasing values of v_0. In contrast the increase of the degrees of freedom $h_{1,\psi}$, from the top to the bottom panels, induces more concentrated prior inclusion probabilities at the origin with converse effects. Both changes have no impact on the prior inclusion probability at the origin since $\mathbb{P}(I_j = v_1 | \beta_j = 0, \cdot)$ does not depend on $h_{1,\psi}$ and $h_{2,\psi}$. **Figure 4.11** shows the impact on the 3 displayed quantities, if the prior mean of the complexity parameter

$H_{\omega} = h_{1,\omega}/h_{1,\omega} + h_{2,\omega}$ is increased. The larger the prior mean H_{ω}, the smaller is the size of the regression coefficient with prior inclusion probability equal to larger than $\frac{1}{2}$, i. e. the regularization from the spike component is reduced and we observe a clear impact on the inclusion probabilities of small and zero effects which increase.

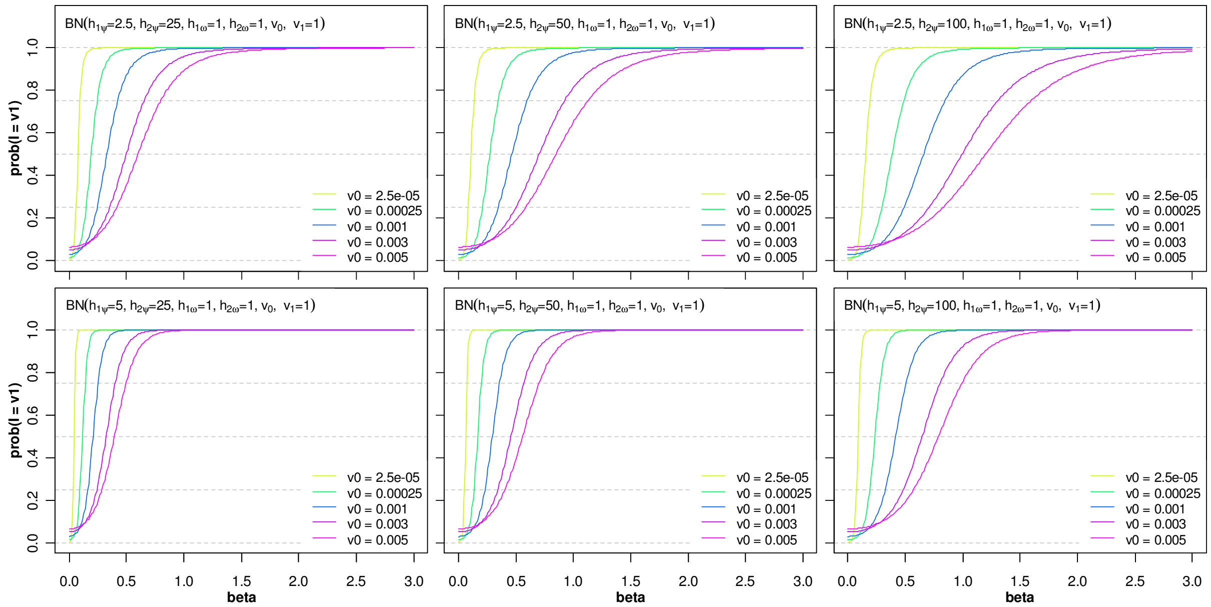


Figure 4.8: Marginal prior inclusion probability of the indicator variable I_j, $\mathbb{P}(I_j = v_1 \mid \beta_j, \cdot)$ given in (4.23), as function of the regression coefficient β_j with the hyperparameters given in the upper left legend. In the upper panel the hyperparameter $h_{1,\psi}$ is set to $h_{1,\psi} = 2.5$ and in the lower panel to $h_{1,\psi} = 5$. From the left side to the right side the hyperparameter $h_{2,\psi}$ varies with values $h_{2,\psi} = 25, 50, 100$. The values v_0 are given in the legend.

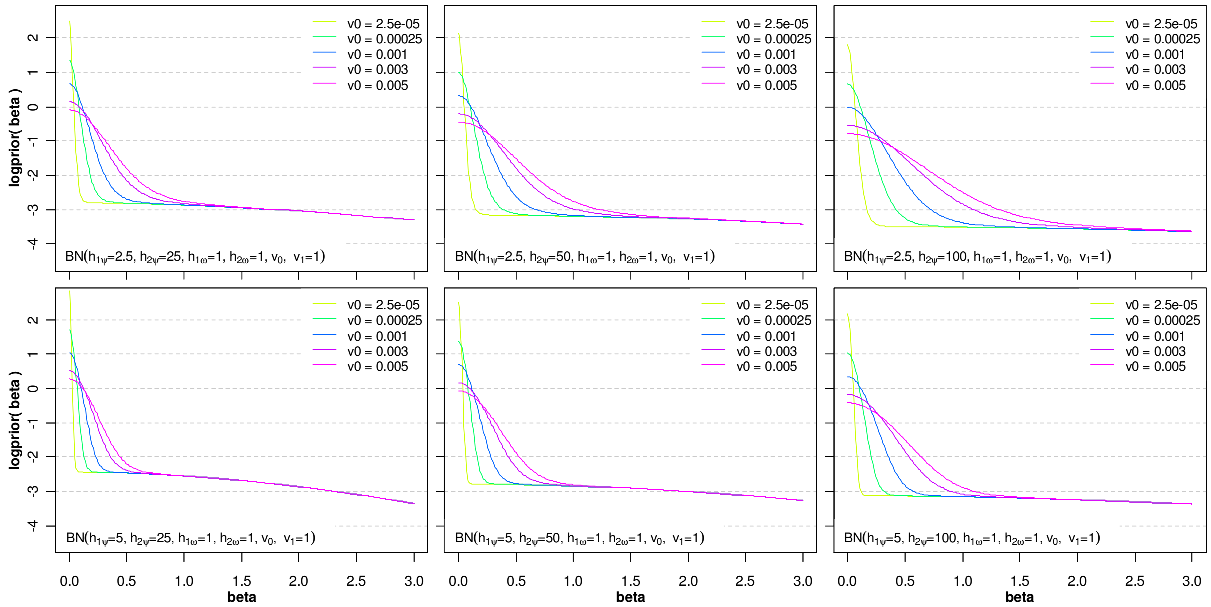


Figure 4.9: Marginal log-prior of the regression coefficients, $\log p(\beta_j \mid \cdot)$ given in (4.22), as function of the regression coefficient β_j with the hyperparameters given in the lower left legend. In the upper panel the hyperparameter $h_{1,\psi}$ is set to $h_{1,\psi} = 2.5$ and in the lower panel to $h_{1,\psi} = 5$. From the left side to the right side the hyperparameter $h_{2,\psi}$ varies with values $h_{2,\psi} = 25, 50, 100$. The values v_0 are given in the legend.

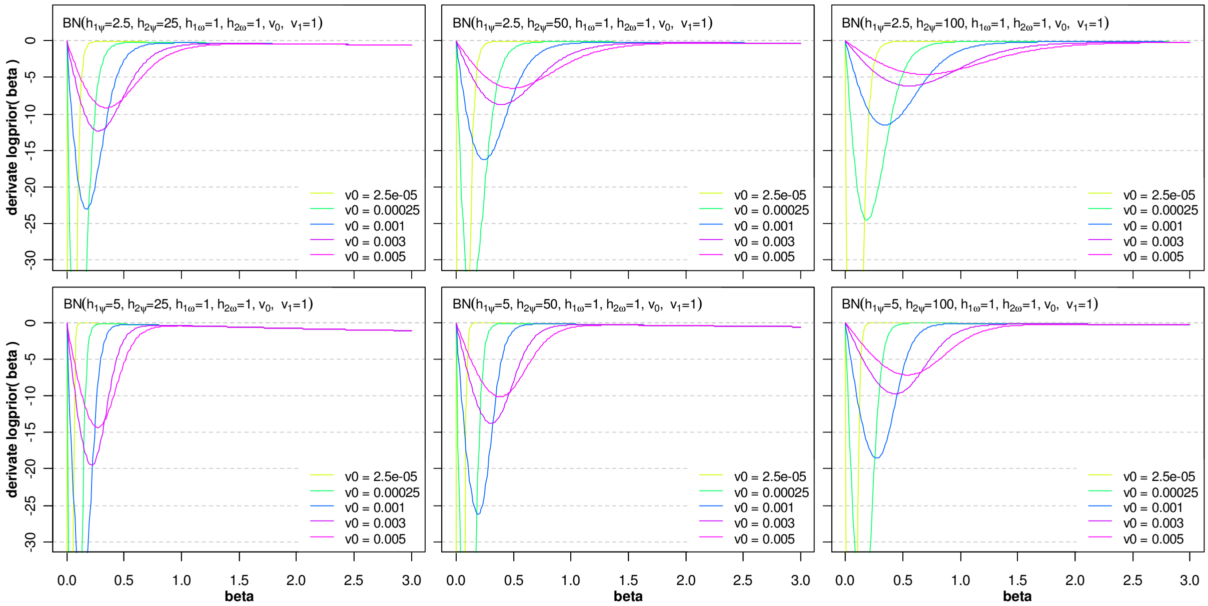


Figure 4.10: Derivate of the marginal log-prior of the regression coefficients, $d\log p(\beta_j \mid \cdot)/d\beta_j$ given in (4.24), as function of the regression coefficient β_j with the hyperparameters given in the upper left legend. In the upper panel the hyperparameter $h_{1,\psi}$ is set to $h_{1,\psi} = 2.5$ and in the lower panel to $h_{1,\psi} = 5$. From the left side to the right side the hyperparameter $h_{2,\psi}$ varies with values $h_{2,\psi} = 25, 50, 100$. The values v_0 are given in the legend.

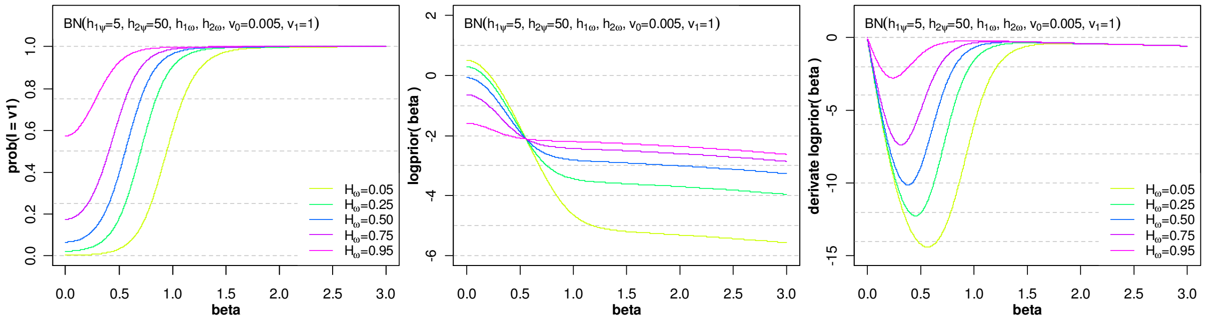


Figure 4.11: Marginal prior inclusion probability of the indicator variable I_j, $\mathbb{P}(I_j = v_1 \mid \beta_j, \cdot)$ given in (4.23), (left side), marginal log-prior of the regression coefficients, $\log p(\beta_j \mid \cdot)$ given in (4.22), (middle), and derivate of the marginal log-prior of the regression coefficients, $d\log p(\beta_j \mid \cdot)/d\beta_j$ given in (4.24), (right side), as function of the regression coefficient β_j with the prior mean of the complexity parameter $H_\omega := h_{1,\omega}/h_{1,\omega} + h_{2,\omega}$ given in the legend.

Shrinkage properties in terms of the marginal prior of the variance parameters

We highlight now the shrinkage properties in terms of the standardized constraint parameter κ_j. For the mixture distribution of the variance parameter $\tau^2_{\beta_j}$ in (4.14) we obtain for κ_j the mixture density

$$p(\kappa_j) = \frac{h_{2,\omega}}{h_{1,\omega}+h_{2,\omega}} \frac{(v_0 h_{2,\psi})^{h_{1,\psi}}}{\Gamma(h_{1,\psi})} \frac{(\kappa_j)^{h_{1,\psi}-1}}{(1-\kappa_j)^{h_{1,\psi}+1}} \exp\left(-\frac{v_0 h_{2,\psi}\kappa_j}{1-\kappa_j}\right)$$
$$+ \frac{h_{1,\omega}}{h_{1,\omega}+h_{2,\omega}} \frac{(v_1 h_{2,\psi})^{h_{1,\psi}}}{\Gamma(h_{1,\psi})} \frac{(\kappa_j)^{h_{1,\psi}-1}}{(1-\kappa_j)^{h_{1,\psi}+1}} \exp\left(-\frac{v_1 h_{2,\psi}\kappa_j}{1-\kappa_j}\right).$$

Due to the similarity, we can use the results from the ridge section to derive the limiting behavior of the mixture density $p(\kappa_j)$ at the margins. At the limit $\kappa_j \to 1$ both prior mixture components are always zero and in summary $p(\kappa_j) \to 0$. For $\kappa_j \to 0$ we obtain concordantly $p(\kappa_j) \to 0$, if $h_{1,\psi} > 1$,

$p(\kappa_j) \to (h_{2,\omega} v_0 h_{2,\psi} + h_{1,\omega} v_1 h_{2,\psi})/(h_{1,\omega} + h_{2,\omega})$, if $h_{1,\psi} = 1$ and $p(\kappa_j) \to \infty$ if $h_{1,\psi} < 1$. **Figure 4.12** shows the prior of the parameter κ_j under various hyperparameter constellations. In the left panel the parameters $h_{1,\omega}$ and $h_{2,\omega}$ are varied keeping the constellation of the remaining parameters fixed. In the middle panel the parameters v_0 and v_1 are varied and in right panel we vary the parameters $h_{1,\psi}$ and $h_{2,\psi}$. The magenta density has identical hyperparameters in each panel. Increasing the hyperparameter $h_{1,\omega}$ enforces that more probability mass is assigned to the left component of the mixture $p(\kappa_j)$. This supports the tail robustness, but leads simultaneously to a reduction of the shrinkage. Vice versa, increasing the hyperparameter $h_{2,\omega}$ is leading to an enhanced probability mass in the right mixture component of $p(\kappa_j)$, which reduces the tail robustness and enforces further shrinkage. Increasing v_1 causes mainly that the prior mass from the left component of the reference mixture $p(\kappa_j)$ is shifted towards $\kappa_j = 0$, the right component coincides almost with the right component of reference distribution. In terms of the decreased parameter v_0, the prior mass from the right component of the reference mixture is shifted towards $\kappa_j = 1$. In summary, the parameters $h_{1,\omega}$ and $h_{2,\omega}$ determine the amount of probability assigned to the left and right mixture component of $p(\kappa_j)$ and the parameters v_1 and v_0 determine the location of the probability at the right and left margins $\kappa_j = 0$ and $\kappa_j = 1$. The hyperparameters $h_{1,\psi}$ and $h_{2,\psi}$ determine the shape and scale of the mixture components similar as outlined under the ridge prior in Subsection 4.1.2. Increasing $h_{2,\psi}$ enhances the mass near $\kappa_j \approx 0$ and increasing $h_{1,\psi}$ enhances the mass near $\kappa_j \approx 1$. The values of the hyperparameter $h_{1,\psi}$ determine weather the limit of the prior at $\kappa_j = 0$ is finite or infinite. From this point of view it seems to be appealing to work also with values $h_{1,\psi} < 1$ to support the tail robustness, e. g. $h_{1,\psi} < 0.9$ as displayed in the right panel of **Figure 4.12**, and adjusting v_0 to a smaller value to emphasize the shrinkage.

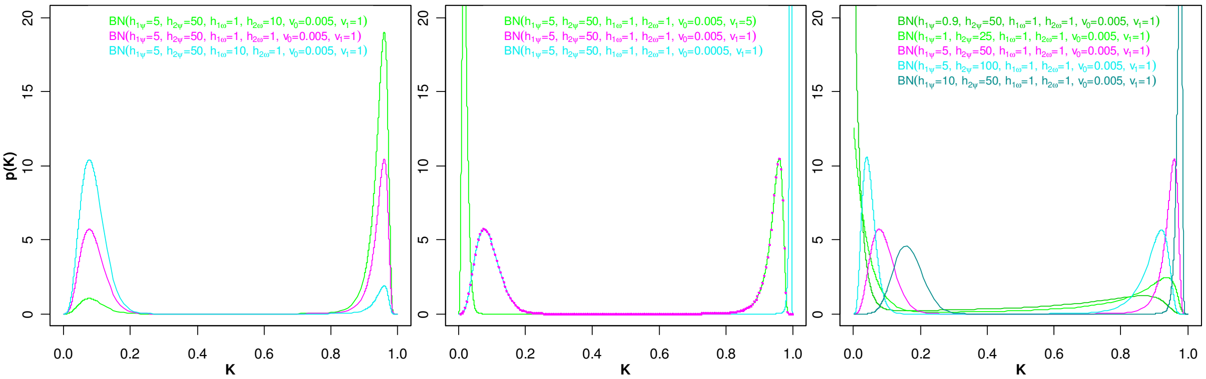


Figure 4.12: Prior densities of the standardized constraint parameter κ_j for the marginal variance prior (4.21) under various hyperparameter combinations given in the legends. In the middle panel the reference density (magenta) is given by the dotted line.

In our simulations and applications we use the two parameter constellations $h_{1,\psi} = 5$, $h_{2,\psi} = 50$, $v_0 = 0.005$, with $\mathbb{P}(\kappa_j \in [0, 0.25]) \approx 0.499$ and $\mathbb{P}(\kappa_j \in [0.25, 0.75]) \approx 0.003$, or $h_{1,\psi} = 5$, $h_{2,\psi} = 25$, $v_0 = 2.5\text{e-}5$, with $\mathbb{P}(\kappa_j \in [0, 0.25]) \approx 0.458$ and $\mathbb{P}(\kappa_j \in [0.25, 0.75]) \approx 0.126$), both in combination with $v_1 = 1$ and $h_{1,\omega} = h_{2,\omega} = 1$. The first setting is suggested by Ishwaran and Rao (2005b) for standardized covariates and rescaled responses in the linear model. This setting is used in the simulations of Section 10 based on the extended AFT model. In the second hyperparameter constellation the selection of smaller effects is emphasized. We use this in the simulations of Section 11 with the extended Cox model and the applications. For both settings the density of the parameter κ_j is shown in **Figure 4.5.** The prior mass in the left and right component of $p(\kappa_j)$ is in both cases close to 0.5 and we have wide areas between the components with close to zero probability mass. In Subsection 4.5 we compare the shrinkage properties of the different shrinkage priors with various hyperparameter constellations in terms of the Weibull model.

Conditional posterior inclusion probability

In the applications we consider the evolution of the parameter estimates if the complexity parameter is varied. The Bayesian paths of the estimates are computed by fixing the complexity parameter ω to the initial value and skip the update of ω within the MCMC sampler. We observed that the parameter paths, in particular the path of the indicator variable, are wiggly and become very unstable for larger regression coefficients at small values of the complexity parameter ω. To clarify this behavior, we consider the full conditional of the indicator variable, compare (6.11), to analyze the probability of sampling $I_j = v_1$ as new state of the Markov chain. The probability is given by

$$\mathbb{P}(I_j = v_1 \mid \cdot) = \left(1 + \frac{1-\omega}{\omega}\sqrt{\frac{v_1}{v_0}}\exp\left\{-\frac{1}{2}\frac{(v_1 - v_0)}{v_0 v_1}\frac{\beta_j^2}{\psi_j^2}\right\}\right)^{-1}$$

and depends on the hyperparameters v_0 and v_1, the current states of the regression coefficient β_j, the variance parameter ψ_j and the complexity parameter ω. In the following we use the fixed hyperparameter constellation $h_{1,\psi} = 5$, $h_{2,\psi} = 50$, $v_0 = 0.005$ and $v_1 = 1$ for demonstration purposes. The hyperparameters $h_{1,\omega}, h_{2,\omega}$ affect only the update of the complexity parameter and may be assumed to be $h_{1,\omega}, h_{2,\omega} = 1$.

- Due to $\mathbb{P}(I_j = v_1 \mid \cdot) \to 1$, if $\beta_j \to \infty$, larger regression coefficients are associated with higher probabilities of sampling $I_j = v_1$ as new state. At the origin $\beta_j = 0$ we obtain $\mathbb{P}(I_j = v_1 \mid \cdot) = (1 + (1-\omega)\sqrt{v_1}/\omega\sqrt{v_0})^{-1}$ and the inclusion probability of a zero effect depends only on the current state of ω.
- When the complexity parameter varies in its range $\omega \in [0,1]$, we obtain an inclusion probability $\mathbb{P}(I_j = v_1 \mid \cdot) \to 1$ at the right margin $\omega \to 1$ and $\mathbb{P}(I_j = v_1 \mid \cdot) \to 0$ at the left margin $\omega \to 0$.
- Finally, for small variances $\psi_j^2 \to 0$ the probability converges to 1, $\mathbb{P}(I_j = v_1 \mid \cdot) \to 1$, and for larger variances $\psi_j^2 \to \infty$ follows $\mathbb{P}(I_j = v_1 \mid \cdot) \to 0$.

The last point may be at the first sight somehow contra-intuitive, since we associate larger variances ψ_j^2 with the slab component and suppose a higher probability to sample $I_j = v_1$. We can clarify this by considering the full conditional of the variance parameter, compare (6.12), which is given by $\psi_j^2 \mid \cdot \sim \text{IGamma}(h_{1,\psi} + 0.5, h_{2,\psi} + 0.5 I_j^{-1}\beta_j^2)$, where the scale parameter depends on the current state of the indicator variable and the regression coefficient. The smaller value $I_j = v_0$ causes an increase in the scale parameter, compared to $I_j = v_1$, and we obtain by trend larger sampled variances. **Figure 4.13** shows the full conditional of ψ_j^2 as function of ψ_j^2 and β_j under the given hyperparameter constellation. At the left side we see the full conditional given $I_j = v_0$ and at the right side given $I_j = v_1$. If we condition on $I_j = v_1$, the scale parameter $h_{2,\psi} + 0.5\beta_j^2$ increases for larger values of the regression coefficients, but in the shown range $\beta_j \in [0,3]$ the changes in the full conditionals for fixed values β_j are only marginal. The marked 5% quantiles and the mean vary marginally within the range of the two marked full conditionals at $\beta_j = 0.5$ and $\beta_j = 1$. In contrast, we observe a strong dependence on the values β_j if we condition on $I_j = v_0$. For larger values β_j the scale parameter $h_{2,\psi} + 0.5 v_0^{-1}\beta_j^2 = h_{2,\psi} + 100\beta_j^2$ becomes very large and more probability mass is shifted to larger values of ψ_j^2. The marked 5% quantiles and the mean vary clearly and the upper 5% quantile exceeds the plotting area for $\beta_j = 1$. The classification of a larger regression coefficient to the component $I_j = v_0$ is leading to larger variance ψ_j^2 compared to $I_j = v_0$ and to a decrease of the sampling probability $\mathbb{P}(I_j = v_1 \mid \cdot)$, and vice versa the classification of a smaller regression coefficient to the component $I_j = v_1$ is leading to smaller variance ψ_j^2 compared to $I_j = v_1$ and an increased sampling probability $\mathbb{P}(I_j = v_1 \mid \cdot)$.

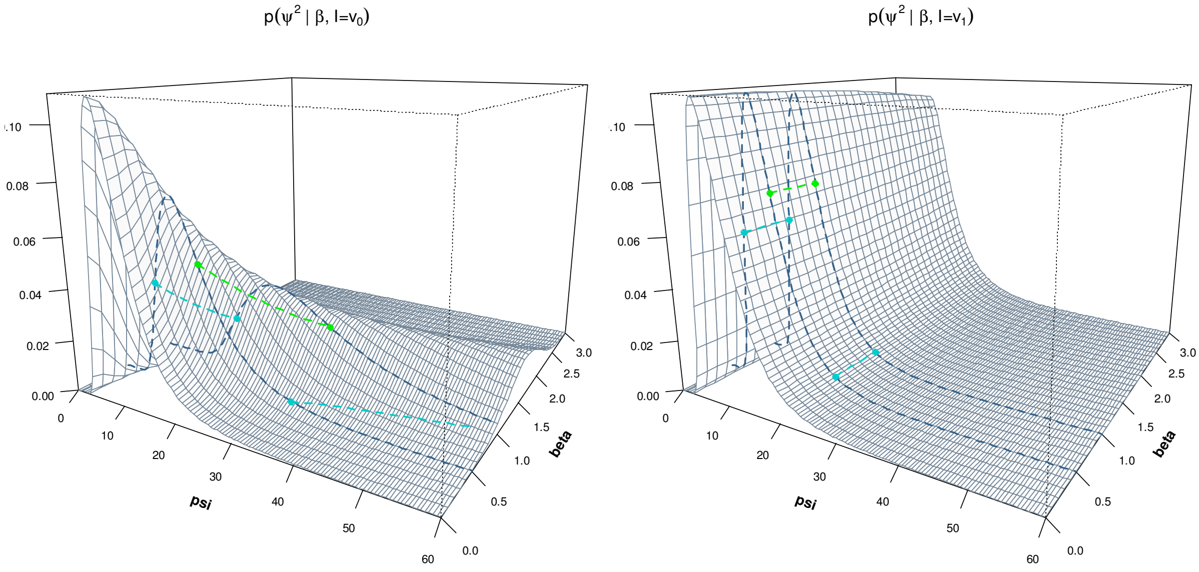


Figure 4.13: Full conditional of the variance parameter $p(\psi_j^2 \mid \beta_j, I_j, h_{1,\psi} = 5, h_{2,\psi} = 50)$ as function of ψ_j^2 and β_j. At the left side we condition on $I_j = v_0$ with $v_0 = 0.005$, at the right side we condition on $I_j = v_1$ with $v_1 = 1$. The bold dashed lines mark the full conditional for the fixed values $\beta_j = 0.5$ and $\beta_j = 1$. The blue points mark the upper and lower 5% quantile and the green point the mean of the full conditional at the both fixed values $\beta_j = 0.5, 1$. The dashed colored lines show the movement of the quantiles and the mean if β_j varies in the interval $[0.5, 1]$.

Figure 4.14 shows the resulting probability for sampling the value v_1, $\mathbb{P}(I_j = v_1 \mid \cdot)$, as a function of the current value of the regression coefficient in the range $\beta_j \in [0,2]$. From the left to the right panel the complexity parameter increases with values $\omega = 0.1, 0.5, 0.9$. Variability in the complexity parameter can be considered by analyzing the changes caused by the variation of ω. In the upper panel we assume that the current state of the indicator is $I_j = v_1$, i. e. the variance parameter is sampled from the full conditional of ψ_j^2 given $I_j = v_1$ and in the lower panel we assume that the current state of the indicator is $I_j = v_0$ with corresponding variance ψ_j^2. Within the panels the inclusion probability is shown in dependence on various values of the current variance parameter ψ_j^2. We use the mean, the mode and the 5% upper and lower quantiles of the associated inverse gamma full conditionals evaluated at the current state β_j, to compute the ranges of the inclusion probability in order to get an impression of the variability caused by the variance parameter ψ_j^2. In the upper and lower panel the inclusion probability increases at the origin $\beta_j = 0$ with values 0.008, 0.07, 0.39 from the left to the right with increasing ω. For large values of the complexity parameter, e. g. $\omega = 0.9$, the probability for sampling $I_j = v_1$ for a zero regression coefficient is about 0.39, irrespective from which component the current state of the variance ψ_j^2 is obtained. The sampling probabilities of $I_j = v_1$ obtained with lower and upper 5% quantile of the full conditional of ψ_j^2 (left and right blue line) provides a kind of upper and lower bound of the inclusion probability for a regression coefficient. In the upper panel, with the current value $I_j = v_1$, we see that the upper and lower bounds of the inclusion probability are both close to 1, for regression coefficients larger than $|\beta_j| > 1.5$ if $\omega = 0.1$ and for regression coefficients larger than $|\beta_j| > 1.0$ if $\omega = 0.9$. For $\omega = 0.1$ we obtain for smaller regression coefficients $|\beta_j| < 0.25$ close to zero sampling probabilities $\mathbb{P}(I_j = v_1 | \cdot)$. Dependent on ω we obtain bounds for regression coefficients, given a certain hyperparameter constellation, which are almost assigned to the spike or the slab component. For some values of the regression coefficients within the area, where the upper and lower bound are close to 0 or 1, we have a high variability in the sampling probability of $I_j = v_1$. If we consider, e. g. for $\omega = 0.1$, the inclusion probability at the mode of the full conditional of ψ_j^2, we find that regression coefficients $|\beta_j| \approx 0.6$ have a sampling probability close to 0.5, but the sampling probability can vary within the range $\mathbb{P}(I_j = v_1 | \cdot) \in [0.05, 0.95]$. For each fixed value of ω we find such regions of the regression coefficients with a high variability in the sampling probability of $I_j = v_1$. For adjacent values of fixed complexity parameters and regression coefficient states that move within this high variability area, the resulting adjacent posterior inclusion probabilities can show a high variability and the parameter paths become wiggly. If the value of the complexity parameter increases, the variability in the sampling probability is reduced, since also the bounds of the sampling probability of $I_j = v_1$ increase and the parameter paths become more stable at the right side. For smaller values of ω, the area of regression coefficients with high variability in the sampling probability of $I_j = v_1$ becomes a wider range and is shifted to larger values of the regression coefficients. So, at the left side the parameter paths of such medium sized regression coefficients become very unstable.

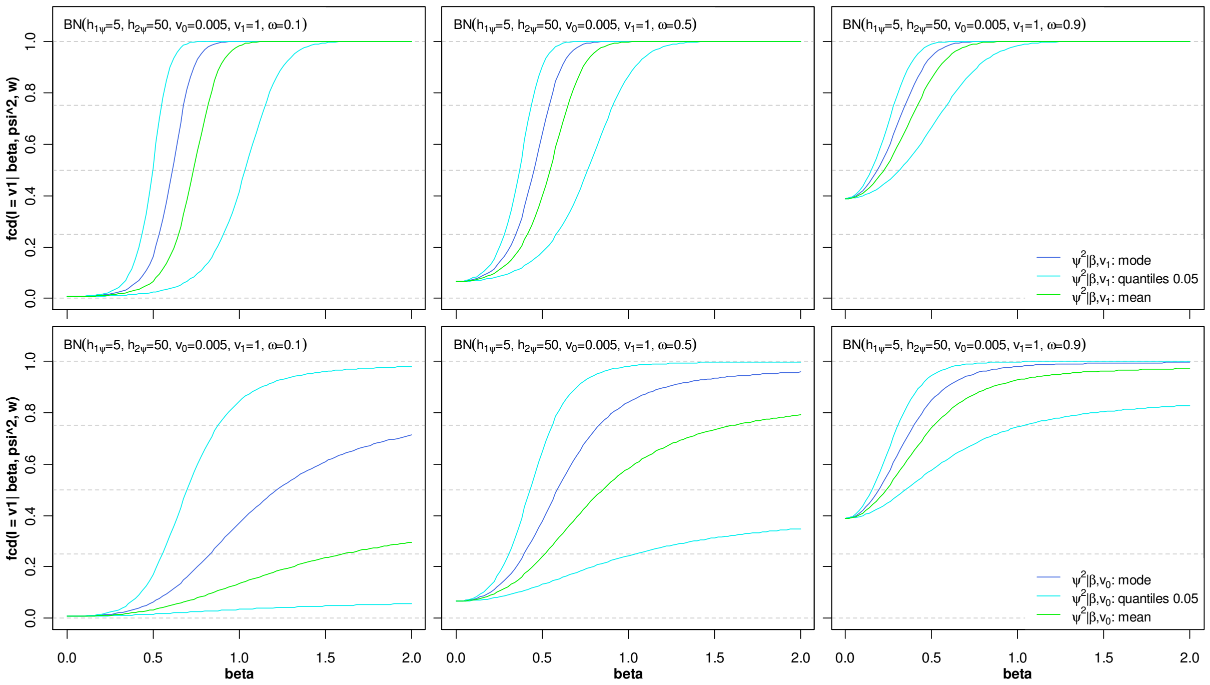


Figure 4.14: Sampling probability from the full conditional density of the indicator variable, $\mathbb{P}(I_j = v_1 | \beta_j, \psi_j^2, \omega)$, as function of the regression coefficient β_j for the hyperparameters given in the upper left legend. From the left side to the right side the complexity parameter ω varies from $\omega = 0.1$ over $\omega = 0.5$ to $\omega = 0.9$. In the upper panel the blue solid lines mark the mode and the lower and upper 5% quantiles of the full conditional $p(\psi_j^2 | \beta_j, I_j = v_1)$ at the corresponding value of β_j and the green line marks the mean. In the lower panel the blue solid lines mark the mode and the lower and upper 5% quantiles of the full conditional $p(\psi_j^2 | \beta_j, I_j = v_0)$ at the corresponding value of β_j and the green line marks the mean.

We further see, that even for small, close to zero, values of the complexity parameter, larger effects can have sampling probability of $I_j = v_1$ close to 1. Such influential coefficients have over a wide range of the complexity parameter high posterior inclusion probabilities. Therefore, the parameter paths under the NMIG penalty show different behavior as e. g. the lasso paths, where for small values of the shrinkage parameter also influential coefficients are strongly shrunken close to zero. In the lower panel, with the current value $I_j = v_0$, the sampling probability $\mathbb{P}(I_j = v_1 \mid \cdot)$ for small regression coefficients is close to the sampling probability in the upper panel. With respect to the mean of the full conditional of ψ_j^2 we have clearly smaller sampling probabilities for larger regression coefficients of comparable size than in the upper panel. The sampling probability is not as close to zero for larger regression coefficients and a current state $I_j = v_0$ can be left. The results shown in **Figure 4.14** are almost comparable for the second hyperparameter constellation ($h_{1,\psi} = 5$, $h_{2,\psi} = 25$, $v_0 = 2.5\text{e-}5$, $v_1 = 1$) used in the simulations and applications with respect to a modified smaller range of the regression coefficients.

4.3.3. Extensions

Similar to the derivation in the Bayesian ridge Section 4.1.3, we obtain for groups of associated regression coefficients, using an identical amount of shrinkage for regression coefficients $\tilde{\boldsymbol{\beta}}_j$ within the groups, mixtures of multivariate Student t-distributions with $d = 2h_{1,\psi}$ degrees of freedom and scale matrices $\boldsymbol{\Sigma}_0^{\frac{1}{2}} = \mathbf{I}\sqrt{v_0 h_{2,\psi} h_{1,\psi}^{-1}}$, $\boldsymbol{\Sigma}_1^{\frac{1}{2}} = \mathbf{I}\sqrt{v_0 h_{2,\psi} h_{1,\psi}^{-1}}$ as marginal distributions of the regression coefficients $\tilde{\boldsymbol{\beta}}_j$, compare Appendix B.3. The adaptive version of the NMIG prior enables the specification of covariate-specific inclusion probabilities $\mathbb{P}(I_j = v_1) = \omega_j$ through covariate-specific hyperparameters utilized in the prior distributions $\omega_j \sim \text{Beta}\left(h_{1,\omega}, h_{2,\omega}\right)$.

4.4. Variable selection

In contrast to the optimization based methods for feature selection, the presented sampling based Bayesian MCMC methods do not eliminate features completely. Sampling based summary statistics from the posterior, like the mean or the median, are never exactly zero even under the Bayesian NMIG prior. Hence, selection of important variables relies on the inspection of the posterior. To build sparse final models, we consider hard shrinkage selection rules to accomplish variable selection.

A first interval criterion is constructed using the empirical standard deviation $\hat{s}_{\hat{\beta}_j}$ of the sampled regression coefficients β_j. We eliminate a covariate x_j from the predictor of the final model, if the zero lies outside the one standard deviation interval around the estimated regression coefficient $\hat{\beta}_j$ and otherwise the covariate is retained, i. e.

$$\textbf{HS.STD}: \quad \hat{\beta}_j := 0 \quad \text{if} \quad 0 \in [\hat{\beta}_j - \hat{s}_{\hat{\beta}_j}, \hat{\beta}_j + \hat{s}_{\hat{\beta}_j}] .$$

The second rule is similar, but based on the 95% credible interval with the empirical quantiles $\hat{q}_{\beta_j,0.025}$ and $\hat{q}_{\beta_j,0.975}$ from the sample of the regression coefficients, i. e.

$$\textbf{HS.CRI}: \quad \hat{\beta}_j := 0 \quad \text{if} \quad 0 \in [\hat{q}_{\beta_j,0.025}, \hat{q}_{\beta_j,0.975}] .$$

By trend the HS.CRI interval has a wider range compared to the HS.STD interval and is leading to sparser final models. In Konrath (2007) these selection rules are utilized in context of regularized exponential family regression, recently Li and Lin (2010) utilized similar rules, where the margins of the intervals are determined via ROC curves. In contrast to the Bayesian lasso and ridge prior, the NMIG prior provides a natural criterion to select covariates on the base of the MCMC samples of the indicator variables I_j. Covariates with considerable influence should be frequently assigned to the mixing distribution component corresponding to the indicator with values $I_j = v_1$. The higher the percentage of the values v_1 in the sample, the larger is the evidence that the corresponding covariate has a non negligible effect. In our simulations and applications we use the intuitive cut-off value of 0.5 as selection threshold and covariates with higher relative frequency of the associated indicator variable value $I_j = v_1$ are included in the final model. In summary, the third criterion is given by

$$\textbf{HS.IND}: \quad \hat{\beta}_j := 0 \quad \text{if} \quad \hat{\mathbb{P}}(I_j = v_1) \leq 0.5,$$

where $\hat{\mathbb{P}}(I_j = v_1)$ denotes the estimated inclusion probability based on posterior relative frequencies of the Bayesian NMIG indicator variable value $I_j = v_1$. In the Simulation Section 11.5 we consider some variations of the threshold value.

4.5. Simulation

In the following we demonstrate the properties of the presented shrinkage priors in a simple setting for various hyperparameter constellations. In the later simulation and application sections we use only a reduced number of methods and hyperparameter constellations with very different settings for the AFT and CRR model. Since there is no connection between the extended versions of AFT and CRR model, the results there are not directly comparable with each other.

Data generation

We use $p_x = 10$ covariates $\mathbf{x}_i = (x_{i,1},...,x_{i,10})'$ which are randomly drawn from a multivariate Gaussian distribution with zero mean, unit variance and no correlation between the covariates. The survival times T_i, $i = 1,...,n$, are generated from an exponential hazard model with constant baseline hazard $\lambda_0(t) = 1$, i. e.

$$\lambda_i(t_i) = \exp(\mathbf{x}_i'\boldsymbol{\beta}), \quad \boldsymbol{\beta} = (0, 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9)',$$

while the censoring variables C_i, $i = 1,...,n$, are generated as i.i.d. draws from the uniform distribution $U[0, c_0]$ with c_0 chosen to obtain the desired censoring rates in each dataset.

We use $R = 50$ replicated datasets with

- $n = 50, 100, 200$ observations,
- with 0% and 25% censored observations in the data.

Simulation setting

We fit with the software package `BayesX` a Bayesian Weibull model, compare Section 9.1.1 for details, to the data with 15000 iterations, a burnin of 5000 iterations and we thin the chain by 10 which results in an MCMC sample of size 1500. Posterior parameter estimates are in general based on the empirical mean of the associated sample from the posterior.

The hyperparameters of the regularization priors are set to the following values.

Bayesian ridge, compare (4.5) and (4.4):

- *BR1*: $h_{1,\lambda} = h_{2,\lambda} = 0.001$, to allow a great amount of adaptiveness to the data.
- *BR2*: $h_{1,\lambda} = 5$, $h_{2,\lambda} = 0.5$, to induce a stronger shrinkage, compare **Figure 4.4.**
- *BR3*: $h_{1,\lambda} = 0.5$, $h_{2,\lambda} = 1$, to obtain a marginal Cauchy prior.

Bayesian adaptive ridge, compare (4.6) and (4.7):

- *ABR1*: $h_{1,\lambda} = h_{2,\lambda} = 0.001$, to allow a great amount of adaptiveness to the data.
- *ABR2*: $h_{1,\lambda} = 5$, $h_{2,\lambda} = 0.5$, to induce a stronger shrinkage, compare **Figure 4.4.**
- *ABR3*: $h_{1,\lambda} = 0.5$, $h_{2,\lambda} = 1$, to obtain a marginal Cauchy prior.

Bayesian lasso, compare (4.12) and (4.13):

- *BL1*: $h_{1,\lambda} = h_{2,\lambda} = 0.001$, to allow a great amount of adaptiveness to the data.
- *BL2*: $h_{1,\lambda} = 5$, $h_{2,\lambda} = 0.5$, to induce a stronger shrinkage compared to BL1, see **Figure 4.7.**

Bayesian NMIG, compare (4.17), (4.18) and (4.20):

- *BN1*: $v_1 = 1$, $v_0 = 0.005$, $h_{1,\psi} = 5$, $h_{2,\psi} = 50$, $h_{1,\omega} = 1$ and $h_{2,\omega} = 1$ ($\text{ISP}_\beta \approx 0.558$).
- *BN2*: $v_1 = 1$, $v_0 = 2.5\text{e-}5$, $h_{1,\psi} = 5$, $h_{2,\psi} = 25$, $h_{1,\omega} = 1$ and $h_{2,\omega} = 1$ to induce a stronger regularization of small regression coefficients ($\text{ISP}_\beta \approx 0.045$).

Results

Regression coefficients

Figure 4.15 displays the median and the interquartile range of the estimated regression coefficients for $n = 50$ observation in the upper panel and $n = 200$ observations in the lower panel. The left panel shows the results for the uncensored data and the right panel those with 25 % censored observations. Within the panels the unregularized estimates (B) are compared with the Bayesian lasso (BL1, BL2), ridge (BR1, BR2) and NMIG (BN1, BN2) estimates.

For $n = 50$ we observe a clear shrinkage of the regression coefficients under the shown regularization priors. Since for the smaller regression coefficients $\beta_j < 0.3$ the unpenalized estimates (B) are close to the true effects, we obtain by trend an overshrinkage of the smaller regression coefficients. This is reversed for the larger effects, where the unpenalized estimates overestimate the true effects. Comparing the NMIG results under the BN1 and BN2 hyperparameter setting, the stronger shrinkage of the smaller effects is confirmed for BN2, with more concentrated boxes for $\beta_1 = 0$ and $\beta_2 = 0.1$. The amount of shrinkage is almost comparable with respect to the median for the regression coefficient $\beta_3 = 0.2$ and the larger regression coefficients are less penalized with BN2, similar to the BR2 estimates. The strongest shrinkage is obtained with the NMIG prior followed by the lasso and the ridge prior. In the presence of censored observations the uncertainty is increased and we observe wider interquartile ranges of the boxes for all model parameters. If more information in terms of an increased number of observations is available, e. g. $n = 200$, the influence of the likelihood to the posterior gets more pronounced, compared to the prior contribution, and the shrinkage of the regression coefficients is reduced.

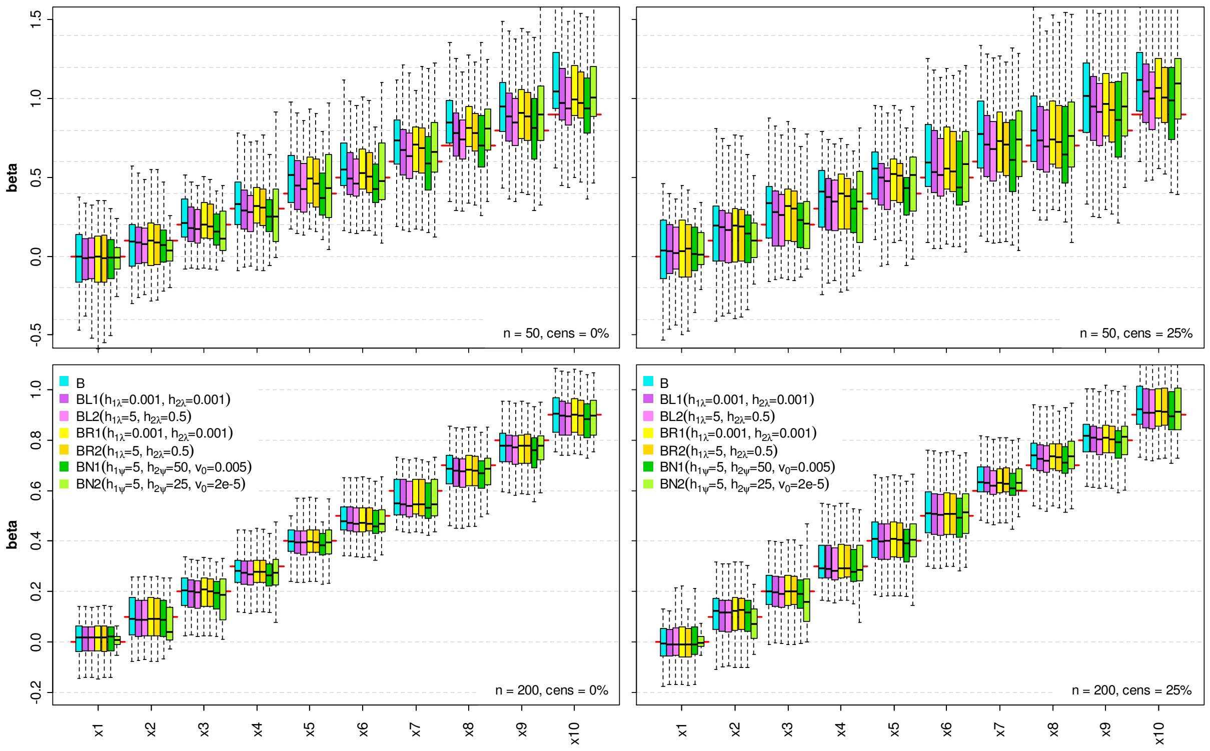


Figure 4.15: Regression coefficient estimates $\hat{\beta}_j$, $j=1,...,10$, under the regularization priors given in the legends. The red horizontal lines mark the true value of the regression coefficient corresponding to the covariates given at the x- axis. Upper panel: Replications with $n = 50$ observations under no censoring (left side) and 25% censoring (right side). Lower panel: Replications with $n = 200$ observations under no censoring (left side) and 25% censoring (right side).

Figure 4.16 compares the estimates obtained with the adaptive versions of the ridge prior (ABR1, ABR2, ABR3), the lasso priors (BL1, BL2) and the Cauchy-prior setting (BR3) from the data with $n = 50$ observations. The estimates under the three adaptive ridge priors are almost comparable to each other with exception for $\beta_1 = 0$, where the concentration of the boxes decreases from ABR2 to ABR3. The adaptive ridge priors induce for smaller regression coefficients a stronger shrinkage compared to the lasso priors, but the difference in the shrinkage is reduced for lager regression coefficients and the adaptive ridge versions become comparable to BL1.

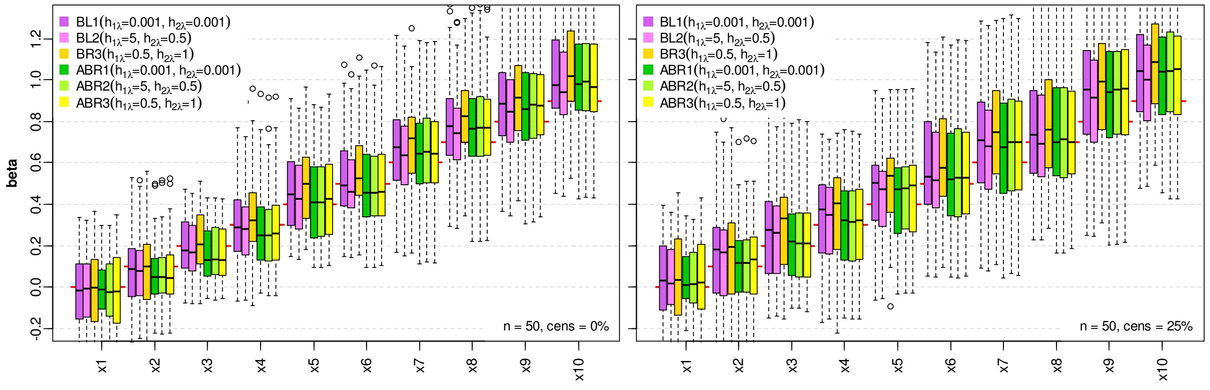


Figure 4.16: Regression coefficient estimates $\hat{\beta}_j$, $j=1,...,10$, under the regularization priors given in the legends. The red horizontal lines mark the true value of the regression coefficient corresponding to the covariates given at the x- axis. Both sides show the replications with $n = 50$ observations under no censoring (left side) and 25% censoring (right side) in the data.

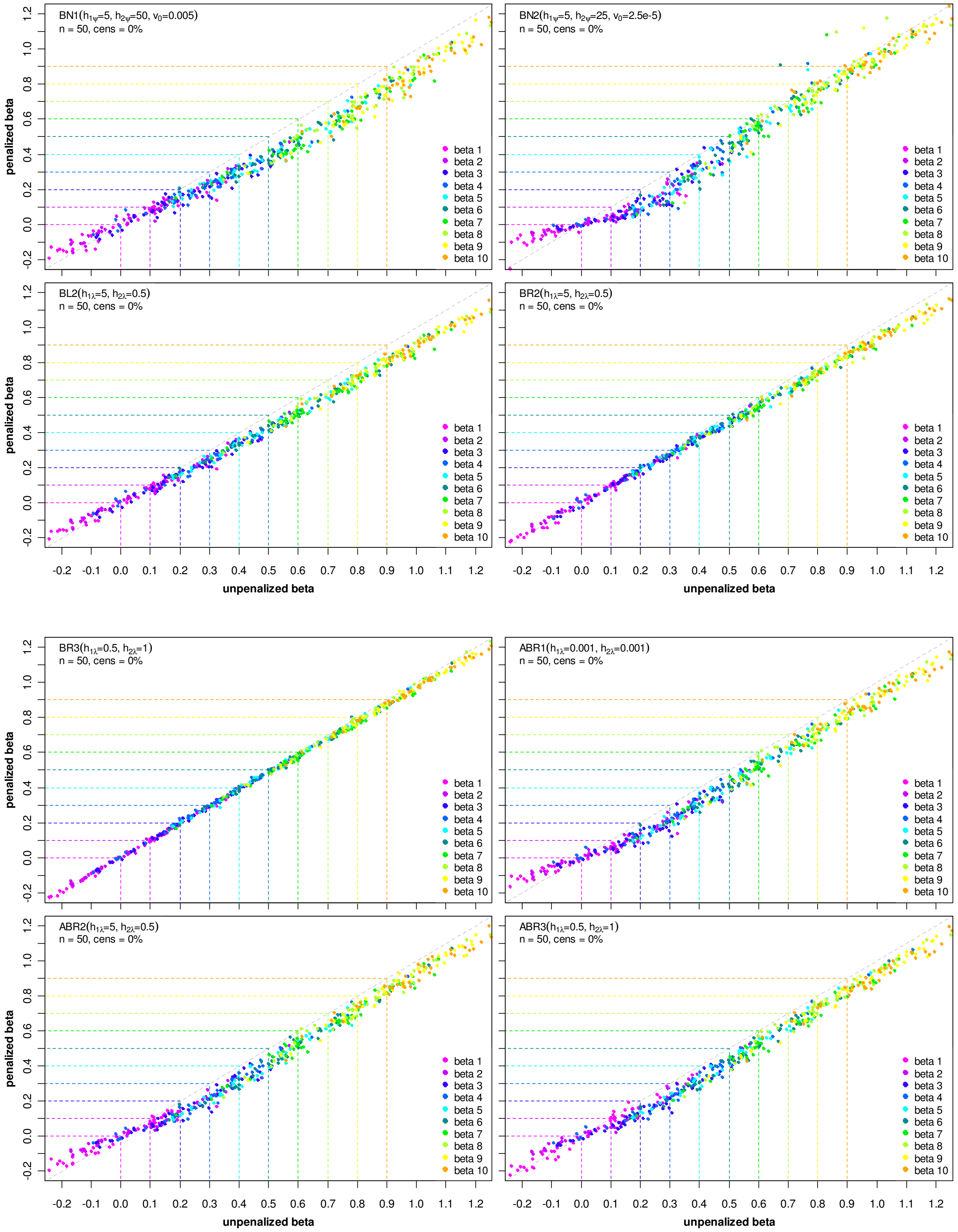


Figure 4.17: Shrinkage of the regularized regression coefficient estimates $\hat{\beta}_j$, $j = 1,...,10$, from the replications with $n = 50$ observations and no censoring in the data under the various regularization priors given in the upper left legends. The x-coordinates represent the unpenalized estimate and the y-coordinates the penalized estimate of the corresponding effect coded by the colors given in the lower right legends. The colored dashed lines mark the associated true values of the regression coefficients on both axes.

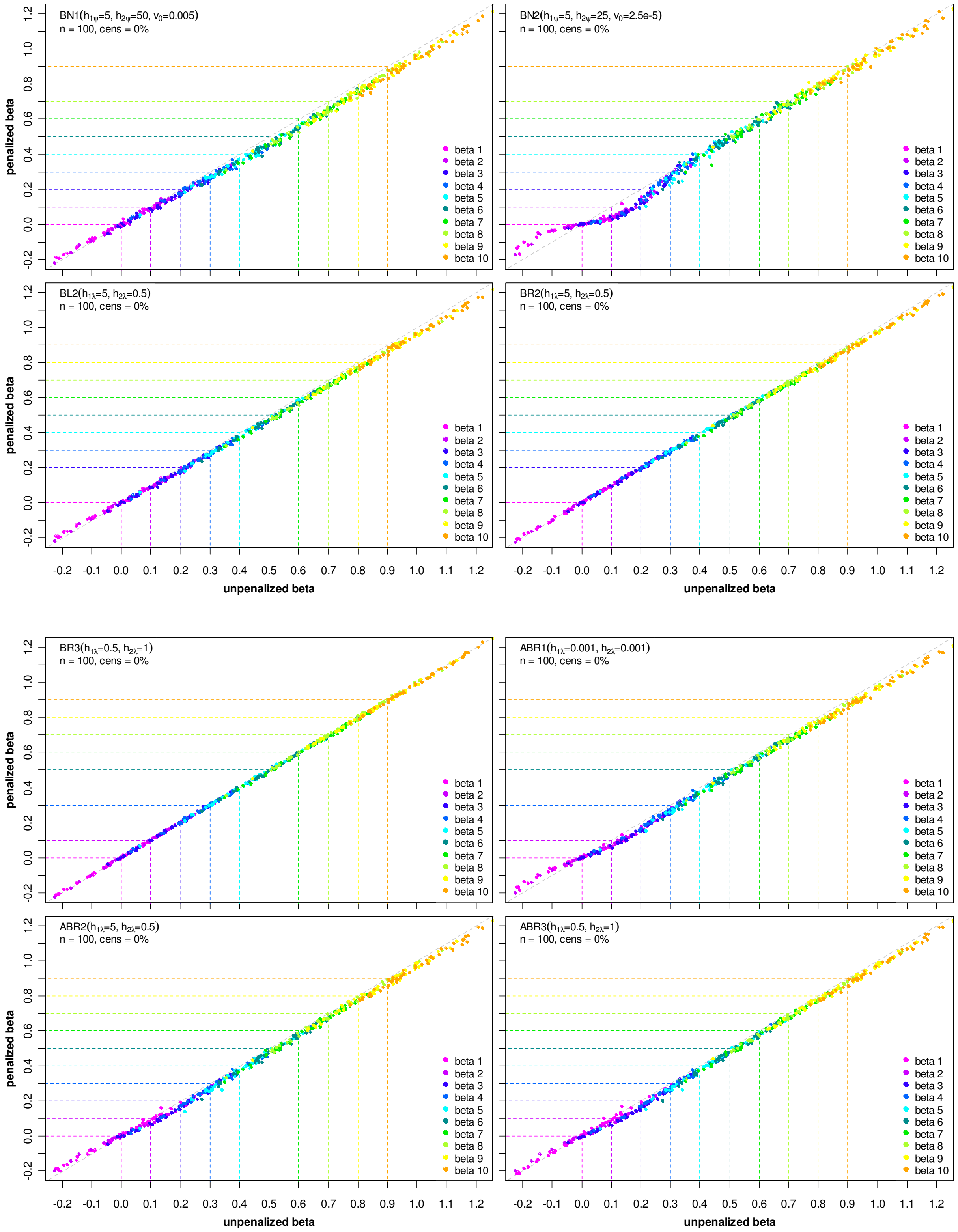


Figure 4.18: Shrinkage of the regularized regression coefficient estimates $\hat{\beta}_j$, $j = 1,...,10$, from the replications with $n = 100$ observations and no censoring in the data under the various regularization priors given in the upper left legends. The x-coordinates represent the unpenalized estimate and the y-coordinates the penalized estimate of the corresponding effect coded by the colors given in the lower right legends. The colored dashed lines mark the associated true values of the regression coefficients on both axes.

To visualize the amount of shrinkage from another point of view, the regularized estimates under the various priors are plotted against the unregularized Bayesian estimates, compare **Figure 4.17** ($n = 50$) and **Figure 4.18** ($n = 100$). The true effect sizes are coded by the colors given in the legends. In particular the amount of shrinkage under the NMIG prior reminds somehow on the SCAD penalty, Fan and Li (2001).

Penalty

In the hierarchical representation of the shrinkage priors, the variance parameters $\tau^2_{\beta_j}$ determine the concentration of the conditional Gaussian distribution of the regression coefficients around zero and smaller variances $\tau^2_{\beta_j}$ induce a stronger shrinkage of the regression coefficients. The regularization of the regression coefficients is reported in **Figure 4.19** (for BN1, BN2, BL1, BR1) and **Figure 4.20** (for BR3 ABR1, ABR2, ABR3) in terms of the logarithm of the inverse variance parameter $\log(\hat{\tau}^{-2}_{\beta_j})$. Displayed are the results obtained with the estimated empirical posterior mean, median, 10% and 90% quantile of $\tau^2_{\beta_j}$ over the replicated uncensored data with $n = 50$ observations.

Under the NMIG prior we observe a great variation in the various location parameter estimates, but each location parameter reflects the decreasing penalty for increasing sizes of the regression coefficients. The penalty is clearly increased for smaller effects in setting BN2, but the variability is clearly reduced for the larger effects. Note that the ridge priors BR induce the same proportion of shrinkage for all regression coefficients. We obtain a covariate-specific penalty under the ABR, BN and BL regularization with a stronger penalization of the smaller and a weaker penalization of the larger regression coefficients. In general we construct the parameter estimates in terms of the empirical mean of the posterior sample, and consequently we report in the simulation and application sections the estimated log-penalty also on the base of the posterior mean of $\tau^2_{\beta_j}$. By the displayed variability in the location parameter estimates we have to keep in mind that the mean estimate under the NMIG prior is rather a lower bound to get an impression about the strength of the regularization. Compared e. g. to the lasso BL1, the log-penalties from the adaptive ridge priors indicate the stronger shrinkage of smaller effects and the weaker regularization of the larger effects. Under the settings ABR2 and ABR3 we observe a concentration of the location parameters for $\beta_1 = 0$, further the resulting marginal posteriors of the variance parameters seem to be extremely skewed, since the mean estimates fall within the region of the 10 % quantile. With respect to this result, ranking the covariates on the basis of the variance parameters $\tau^2_{\beta_j}$ should be rather based on the posterior median than on the posterior mean.

Figure 4.21 and **Figure 4.22** show the associated results to **Figure 4.19** and **Figure 4.20** in terms of the estimated regression coefficients, i. e., we see the empirical posterior mean, median, 10% and 90% quantile of β_j over the replicated uncensored data with $n = 50$ observations under the priors BN1, BN2, BL1 and BR1. In general the clear differences observed in the estimated location parameters of the variances $\tau^2_{\beta_j}$ are less pronounced for the regression coefficients β_j. The empirical mean and median estimates of β_j are almost comparable for most of the regularization priors, even for the Bayesian NMIG prior under the hyperparameter setting BN1. Larger differences for the smaller regression coefficients ($\beta_1, \beta_2, \beta_3$) are obtained with the setting BN2, where the median estimates are much more concentrated around zero than the mean estimates. The relative positions of the shown estimated location parameters indicate almost symmetric marginal posterior distribution for the regression coefficients in contrast to the estimates of the variance parameters.

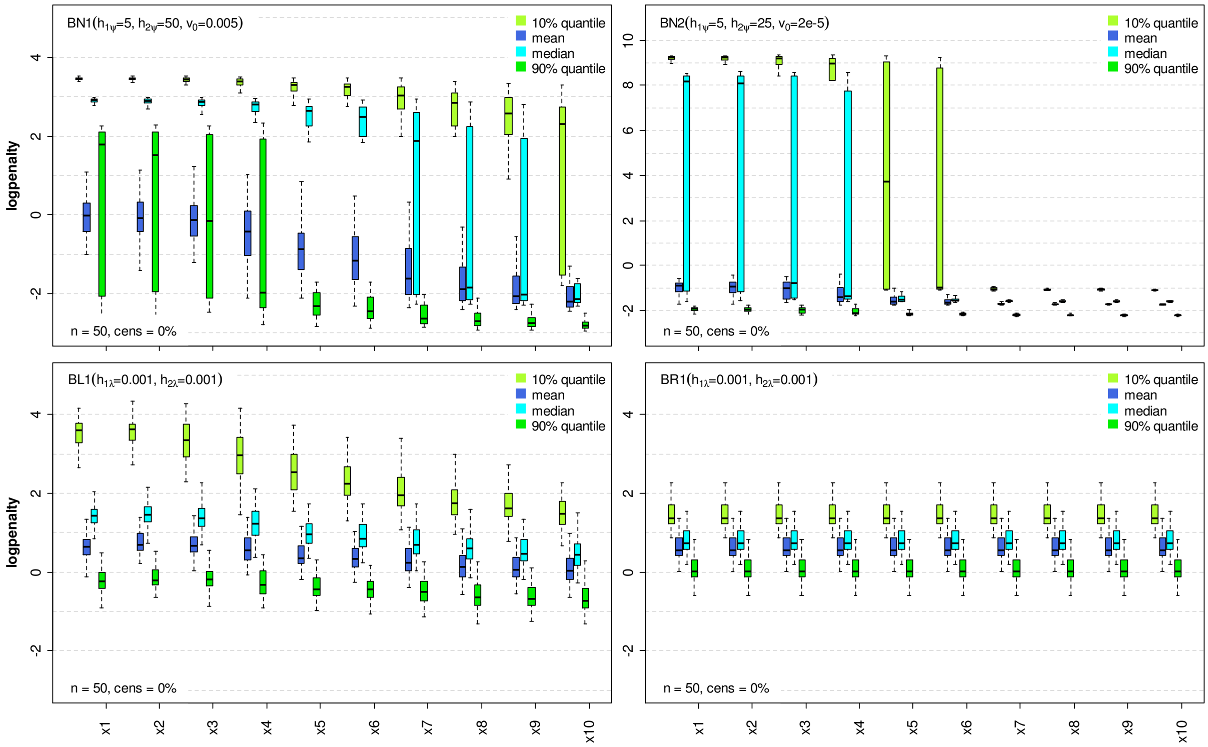


Figure 4.19: Log-penalty estimates $\log(\hat{\tau}_{\beta_j}^{-2})$, $j=1,\ldots,10$, for the replications with $n=50$ observations and no censoring in the data under the regularization priors given in the upper left legends. Shown are the upper and lower 10% quantiles, the median and the mean of the log-penalty.

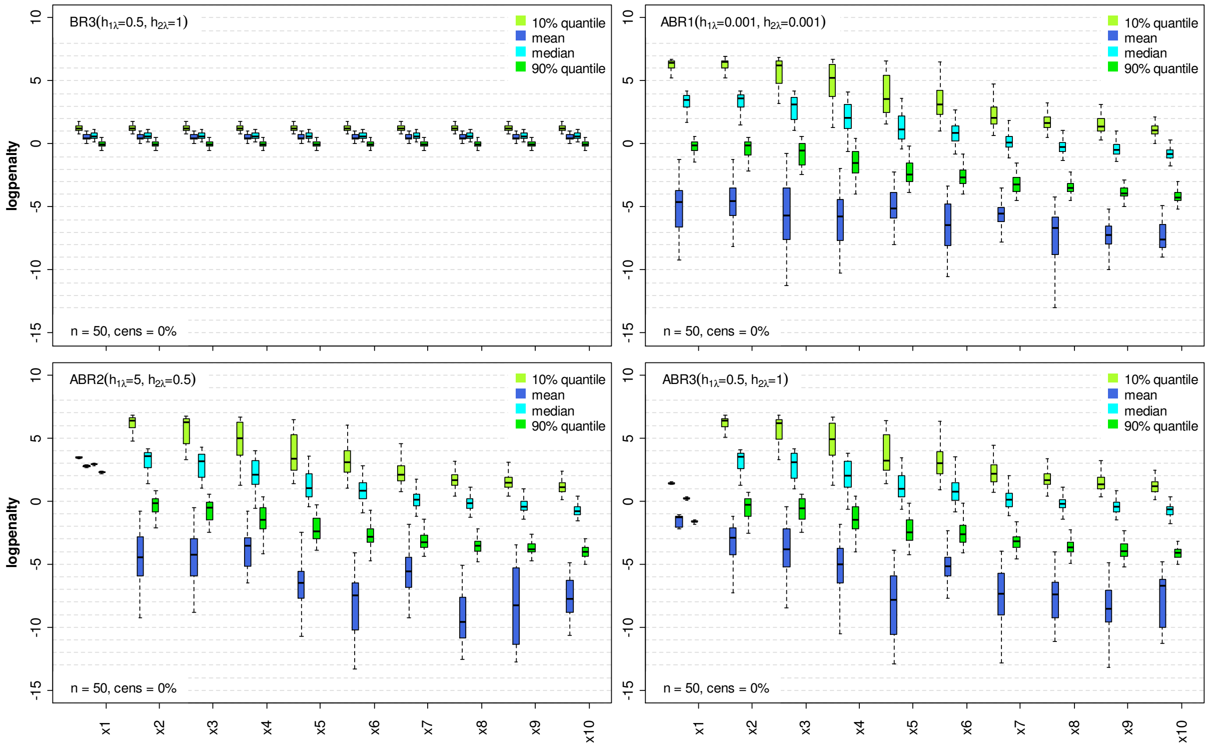


Figure 4.20: Log-penalty estimates $\log(\hat{\tau}_{\beta_j}^{-2})$, $j=1,\ldots,10$, for the replications with $n=50$ observations and no censoring in the data under the regularization priors given in the upper left legends. Shown are the upper and lower 10% quantiles, the median and the mean of the log-penalty.

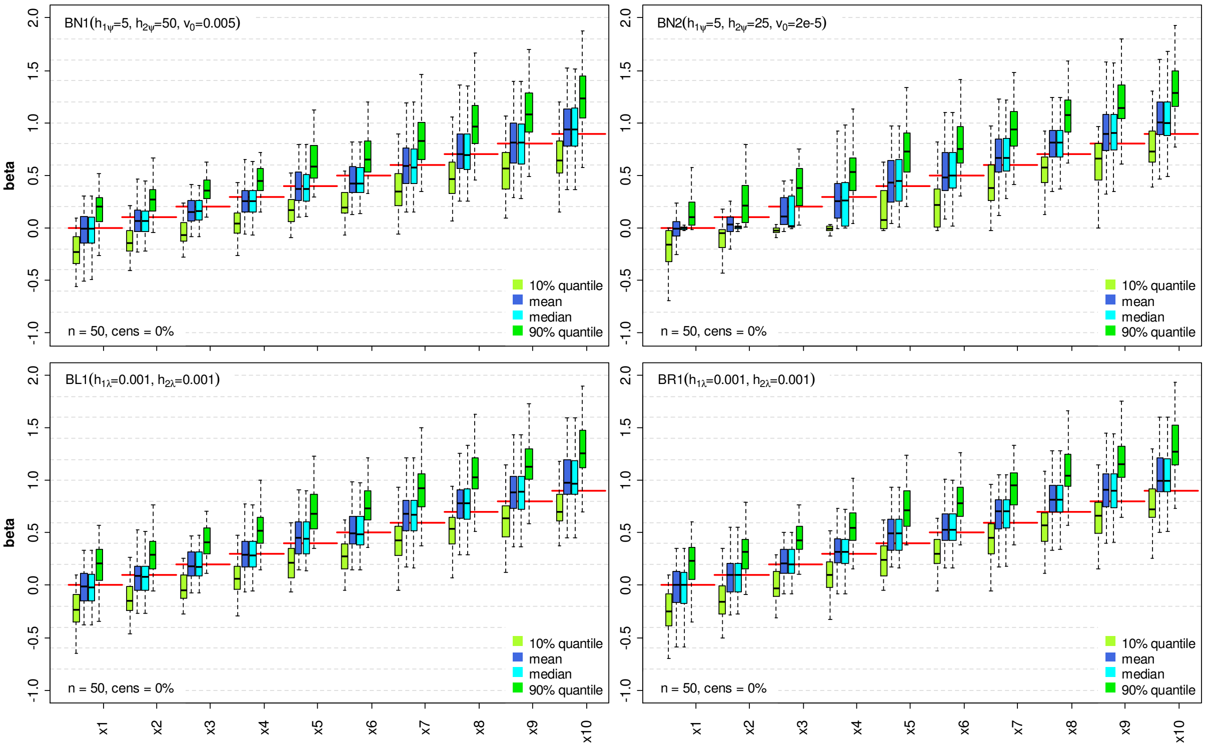


Figure 4.21: Regression coefficient estimates $\hat{\beta}_j$, $j=1,...,10$, for the replications with $n=50$ observations and no censoring in the data under the regularization priors given in the upper left legends. Shown are the upper and lower 10% quantiles, the median and the mean of the regression coefficients.

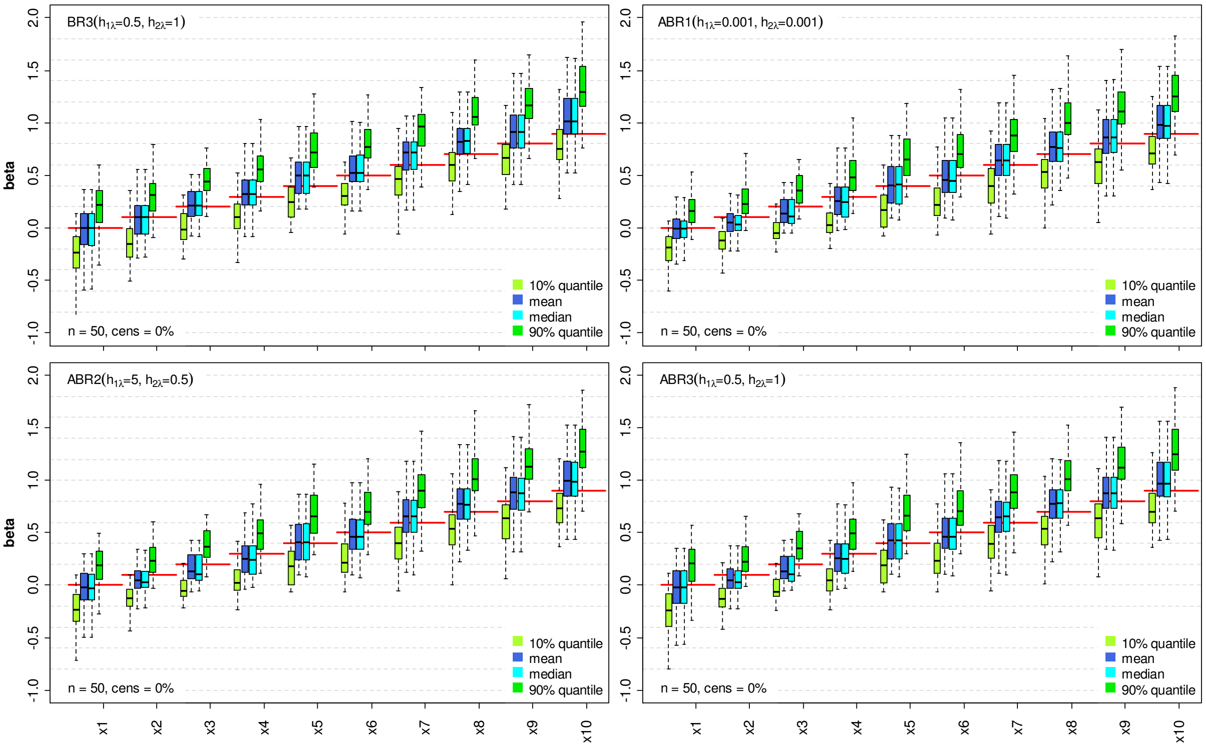


Figure 4.22: Regression coefficient estimates $\hat{\beta}_j$, $j=1,...,10$, for the replications with $n=50$ observations and no censoring in the data under the regularization priors given in the upper left legends. Shown are the upper and lower 10% quantiles, the median and the mean of the regression coefficients.

Indicator variables

Figure 4.23 shows the estimated posterior inclusion probabilities for the two settings of the NMIG prior BN1 and BN2. The estimates are based on posterior relative frequencies of the Bayesian NMIG indicator variable value $I_j = v_1$ in the replications with $n = 50$ and $n = 200$ observations, each with and without censoring in the data.

Under the setting BN1 we observe a slow increase of the estimated inclusion probabilities for increasing size of the regression coefficients. The smaller regression coefficients $(\beta_1, \beta_2, \beta_3)$ have inclusion probabilities close to zero and the largest regression coefficient β_{10} obtains an inclusion probability about 0.8. The uncertainty in the classification to the "spike" and "slab" mixture component, as shown in **Figure 4.14**, is indicated by the larger box-widths for increased regression coefficients. The uncertainty in the classification decreases for larger values of the regression coefficients and the boxes become smaller. Due to the range of the regression coefficients, this behavior can clearly be observed in terms of the setting BN2, but the boxes get also smaller with the setting BN1 for larger coefficients than 0.9. We obtain estimated complexity parameters ω with median ≈ 0.37 in the uncensored data with $n = 50$ observations. Censoring in the data increases the uncertainty and we have larger box-widths. The inclusion probabilities increase by trend and we obtain an increase of the estimated complexity parameter (median ≈ 0.4) with the data containing 25% censored observations. Conversely, when the number of observations is increased to $n = 200$, the estimated inclusion probabilities commonly decrease about a small amount, leading also to a reduction in the estimated complexity parameter (median of the estimates ≈ 0.31).

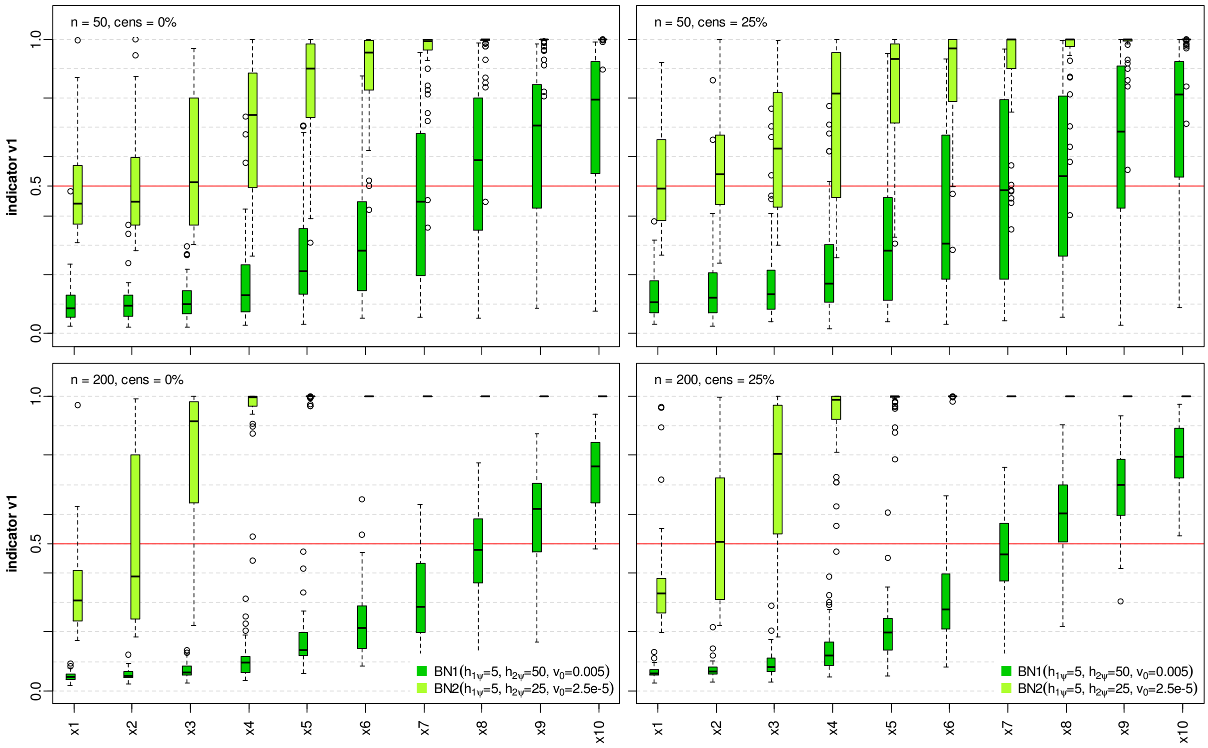


Figure 4.23: Estimated inclusion probabilities, $\hat{\mathbb{P}}(I_j = v_1)$, $j = 1,...,10$, based on posterior relative frequencies of the Bayesian NMIG indicator variable value $I_j = v_1$ for the two hyperparameter settings given in the legends. Upper panel: Replications with $n = 50$ observations under no censoring (left side) and 25% censoring (right side). Lower panel: Replications with $n = 200$ observations under no censoring (left side) and 25% censoring (right side). The red horizontal line marks the cut off value 0.5 of the hard shrinkage selection criterion HS.IND.

With the NMIG prior setting BN2 the estimated inclusion probabilities also increase with increasing effect size, but there we observe basically different and larger sizes of the estimated inclusion probabilities and a rapidly increase compared to the setting BN1. In the data with $n = 50$ observations, effects larger than $\beta_8 = 0.7$ have an inclusion probability about ≈ 1 and the smaller effects around zero about ≈ 0.45 with respect to the median. We obtain estimated complexity parameters ω in the replications with median ≈ 0.75 that reflect the higher model complexity. The impact of the censoring is transferable from the setting BN1, but with increased sample size $n = 200$ we observe a somewhat different effect. The uncertainty of the classification concerns mainly the effects β_2 and β_3, the inclusion probabilities for the effects larger than β_2 are clearly increased and we have a reduced inclusion probability for β_1. This is also reflected in the estimates of the complexity parameter which increase the median ≈ 0.8 for $n = 200$ uncensored observations. It may be somehow confusing at the first sight that the estimated inclusion probability of the zero effect β_1 is not as close to zero as one may possibly expect due to the strong shrinkage of this effect. But we have seen in the right panels of **Figure 4.14** that large values of the complexity parameter ω increase the sampling probability of the indicator value $I_j = v_1$ for zero effects. On the other side we observe sizes of the zero regression coefficients in a broad range around zero, compare **Figure 4.22,** and such effects have a broad range of possible nonzero sampling probabilities also shown in **Figure 4.14.** This explains the high uncertainty in the classification observed for the zero effect under the setting BN2.

If the HS.IND selection rule of Subsection 4.4 was applied, we would remove by trend the covariates x_1 to x_6 in almost all replications from the final models under the prior setting BN1 and the covariate x_{10} with true effect $\beta_{10} = 0.9$ would always be included. Under the setting BN2 the covariates x_2 to x_{10} would frequently appear in the final models while the covariate x_1 would be often removed.

4.6. Random walk prior

Prior hierarchy

Nonparametric model components, like smooth effects, the error term in the flexible AFT model or the log-baseline hazard in the Cox model, are represented as linear combinations of basis functions defined by B-Splines or Gaussian densities. To guarantee smoothness, we assume Bayesian random walk priors of d_j-th order for the basis function weights $\boldsymbol{\alpha}_j$, $j = 0,1,...,p_z$, as suggested in Lang and Brezger (2004), to counterbalance the flexibility provided by utilizing a large number of basis functions. In particular, the first or second order random walk priors are given by,

$$\alpha_{j,k} = \alpha_{j,k-1} + u_{j,k} \quad \text{or} \quad \alpha_{j,k} = 2\alpha_{j,k-1} - \alpha_{j,k-2} + u_{j,k}\,, \tag{4.25}$$

with i.i.d. Gaussian errors $u_{j,k} \sim N(0, \tau^2_{\alpha_j})$, $j = 0,1,...,p_z$, and diffuse priors for the initial values $p(\alpha_{j,1}) \propto \text{const}$ or $p(\alpha_{j,1}) = p(\alpha_{j,2}) \propto \text{const}$. The first order random walk prior controls abrupt jumps in the differences $\Delta^{(1)}\alpha_{j,k} := \alpha_{j,k} - \alpha_{j,k-1}$, while the differences corresponding to the second order random walk prior $\Delta^{(2)}\alpha_{j,k} := \alpha_{j,k} - 2\alpha_{j,k-1} + \alpha_{j,k-2}$ penalizes deviations from a linear trend. Higher order differences with $d_j > 2$ are recursively defined via $\Delta^{(d_j)}\alpha_{j,k} := \Delta^{(d_j-1)}\alpha_{j,k} - \Delta^{(d_j-1)}\alpha_{j,k-1}$ and diffuse priors for the d_j coefficients $p(\alpha_{j,1}) \propto \text{const}, ..., p(\alpha_{j,d_j}) \propto \text{const}$. The variance parameter $\tau^2_{\alpha_j}$ controls the smoothness with the connection, that large values of the variance parameter allow a clear variation in the basis function weights corresponding to wiggly function estimates, while small variances implicate smoothed curves as estimates.

The joint prior for the parameter $\boldsymbol{\alpha}_j$, $j=0,1,...,p_z$, is derived as the product of the conditional Gaussian densities (4.25) and has the form

$$p\left(\boldsymbol{\alpha}_j \mid \tau^2_{\alpha_j}\right) \propto \left(\frac{1}{\tau^2_{\alpha_j}}\right)^{\frac{k_j}{2}} \exp\left(-\frac{1}{2\tau^2_{\alpha_j}}\boldsymbol{\alpha}_j'\mathbf{K}_j\boldsymbol{\alpha}_j\right), \quad j=0,1,...,p_z, \tag{4.26}$$

where $\mathbf{K}_j$ denotes the penalty matrix with $k_j = \text{rank}\left(\mathbf{K}_j\right) = g_j - d_j$. The term $\boldsymbol{\alpha}_j'\mathbf{K}_j\boldsymbol{\alpha}_j$ represents the sum of the quadratic differences $\boldsymbol{\alpha}_j'\mathbf{K}_j\boldsymbol{\alpha}_j = \sum_{k=d_j+1}^{g_j}(\Delta^{(d_j)}\alpha_{j,k})^2$ and the penalty matrix can be written as $\mathbf{K}_j = \mathbf{D}_j'\mathbf{D}_j$, where $\mathbf{D}_j$ is the corresponding d_j-th order difference matrix of dimension $(g_j - d_j)\times g_j$. In general, the penalty matrix $\mathbf{K}_j$ does not have full rank, i. e. $k_j < \dim(\boldsymbol{\alpha}_j)$, and this rank deficiency represents the fact, that specific parts of the function remains unpenalized. For example, a polynomial of order $(d_j - 1)$ remains unpenalized by the d_j-th order penalty matrix. In particular a second order penalty applied to smooth predictor terms $f_j(\cdot)$ is leading to a linear modeled effect in the limiting case, when the variance parameter decreases $\tau^2_{\alpha_j} \to 0$. As a consequence, the conditional Gaussian prior (4.26) is partially improper with precision matrix $\boldsymbol{\Pi}_j := \tau^{-2}_{\alpha_j}\mathbf{K}_j$, and with respect to the partial impropriety the covariance matrix is written as $\boldsymbol{\Sigma}^-_{\alpha_j} := \tau^2_{\alpha_j}\mathbf{K}^-_j$, where $\mathbf{K}^-_j$ denotes a generalized inverse of the penalty matrix $\mathbf{K}_j$. Theoretical results to the propriety of the resulting posterior are given, e. g., in Fahrmeir and Kneib (2009) in the context of structured additive exponential family and hazard regression and in Hennerfeind et al. (2006) in the context of geoadditive survival models.

The conditional distributions $\alpha_{j,k} \mid \boldsymbol{\alpha}_{j,-k}, \tau^2_{\alpha_j}$ of the single basis function weights $\alpha_{j,k}$ given the remaining weights $\boldsymbol{\alpha}_{j,-k} = (\alpha_{j,1},...,\alpha_{j,k-1},\alpha_{j,k+1},...,\alpha_{j,g_j})'$ are also Gaussian with mean and variance

$$\mathbb{E}\left(\alpha_{j,k} \mid \boldsymbol{\alpha}_{j,-k}, \tau^2_{\alpha_j}\right) = -\frac{\sum_{\ell \neq k}\mathbf{K}_j[k,\ell]\alpha_\ell}{\mathbf{K}_j[k,k]}, \quad \mathbb{V}\text{ar}\left(\alpha_{j,k} \mid \boldsymbol{\alpha}_{j,-k}, \tau^2_{\alpha_j}\right) = \frac{\tau^2_{\alpha_j}}{\mathbf{K}_j[k,k]}, \tag{4.27}$$

where $\mathbf{K}_j[k,\ell]$ denotes the element of the penalty matrix in the k-th row and ℓ-th column. In a full Bayesian approach the variance parameters $\tau^2_{\alpha_j}$, $j=0,1,...,p_z$, are commonly equipped with conjugate inverse gamma priors

$$\tau^2_{\alpha_j} \sim \text{IGamma}(h_{1,\tau_j}, h_{2,\tau_j}). \tag{4.28}$$

To specify almost diffuse inverse gamma priors, we select small values of the hyperparameters $h_{1,\tau_j} > 0, h_{2,\tau_j} > 0$. Common choices in this work are $h_{1,\tau_j} = 1$ and a small values $h_{2,\tau_j} \in \{0.01, 0.001\}$ or also $h_{1,\tau_j} \in \{0.01, 0.001\}$. The choice of diffuse, but proper, inverse gamma priors is usually not crucial for smoothing variances to obtain proper full conditionals, compare, e. g., Fahrmeir and Kneib (2009) who provide conditions for propriety of posteriors.

5. Priors for the extended AFT model

In the following the priors of the extended AFT model (2.1) with predictor (2.7) and error distribution (2.6) are briefly summarized. In addition to the estimation of the parameter vector $(\boldsymbol{\alpha}',\boldsymbol{\beta}',\boldsymbol{\gamma}',\sigma)'$ from the extended AFT model, the (partially) latent log-survival times $\mathbf{y}$ and the vector of latent component labels $\mathbf{r}$ need to be imputed. Together with the hierarchical representation of the regularization priors for the predictor components we obtain a beneficial hierarchical model representation to derive fast

MCMC update schemes for posterior inference since most of the priors have a closed and conjugate form.

Joint prior distribution

The complete parameter vector $\boldsymbol{\theta} = (\boldsymbol{\alpha}', \boldsymbol{\beta}', \boldsymbol{\gamma}', \boldsymbol{\tau}_\alpha^{2\prime}, \boldsymbol{\tau}_\beta^{2\prime}, \boldsymbol{\rho}', \sigma)'$ consists of the regularized regression coefficients $\boldsymbol{\alpha} = (\boldsymbol{\alpha}_0', \boldsymbol{\alpha}_1', \ldots, \boldsymbol{\alpha}_{p_z}')'$, $\boldsymbol{\alpha}_j = (\alpha_{0,j}, \ldots, \alpha_{0,g_j})'$, representing the basis function weights of the nonlinear modeled covariate effects in the predictor and the transformed mixture weights of the error distribution as well as the associated smoothness parameters $\boldsymbol{\tau}_\alpha^2 = (\tau_{\alpha_0}^2, \tau_{\alpha_1}^2, \ldots, \tau_{\alpha_{p_z}}^2)'$ and the scale parameter σ. Further contained are the unregularized linear effects $\boldsymbol{\gamma} = (\gamma_0, \ldots, \gamma_{p_u})'$ and the regularized linear effects $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_{p_x})'$ with the associated variance parameters $\boldsymbol{\tau}_\beta^2 = (\tau_{\beta_1}^2, \ldots, \tau_{\beta_{p_x}}^2)'$ and $\boldsymbol{\rho}$, which is the generic notation for prior-specific hyperparameters from further stages of the hierarchical formulation, like the shrinkage parameters.

With the independence and distributional assumptions of the latent quantities from Section 3 we obtain the hierarchical prior structure $p(\mathbf{y}, \mathbf{r}, \boldsymbol{\theta}) = p(\mathbf{y} \mid \mathbf{r}, \boldsymbol{\theta}) p(\mathbf{r} \mid \boldsymbol{\theta}) p(\boldsymbol{\theta})$. The mean and covariance matrix of the log-survival times $\mathbf{y}$ depend on the regression parameters of the predictor and the scale parameter, hence we write $p(\mathbf{y} \mid \mathbf{r}, \boldsymbol{\theta}) = p(\mathbf{y} \mid \mathbf{r}, \boldsymbol{\alpha}_1, \ldots, \boldsymbol{\alpha}_{p_z}, \boldsymbol{\beta}, \boldsymbol{\gamma}, \sigma)$. Accordingly, we have for the component labels the dependence on the (transformed) mixture weights $p(\mathbf{r} \mid \boldsymbol{\theta}) = p(\mathbf{r} \mid \mathbf{w}) = p(\mathbf{r} \mid \boldsymbol{\alpha}_0)$. In summary, the joint prior is given by

$$p(\mathbf{y}, \mathbf{r}, \boldsymbol{\theta}) = p(\mathbf{y} \mid \mathbf{r}, \boldsymbol{\alpha}_1, \ldots, \boldsymbol{\alpha}_{p_z}, \boldsymbol{\beta}, \boldsymbol{\gamma}, \sigma) p(\mathbf{r} \mid \boldsymbol{\alpha}_0) p(\boldsymbol{\theta}), \tag{5.1}$$

with

$$p(\boldsymbol{\theta}) = \prod_{j=0}^{p_z} p(\boldsymbol{\alpha}_j \mid \tau_{\alpha_j}^2) p(\tau_{\alpha_j}^2) \cdot p(\boldsymbol{\beta} \mid \boldsymbol{\tau}_\beta^2) p(\boldsymbol{\tau}_\beta^2 \mid \boldsymbol{\rho}) p(\boldsymbol{\rho}) p(\boldsymbol{\gamma}) p(\sigma^2), \tag{5.2}$$

where the factorization in $p(\boldsymbol{\theta})$ reflects the implied independence assumptions formulated in Section 4. The joint prior consists of the following conditional priors for the parameter components.

Prior of the survival times

At the first stage of the prior hierarchy the joint distribution of the log-survival times $\mathbf{Y}$ is multivariate Gaussian

$$\mathbf{Y} \mid \mathbf{r}, \boldsymbol{\alpha}_1, \ldots, \boldsymbol{\alpha}_{p_z}, \boldsymbol{\beta}, \boldsymbol{\gamma}, \sigma \sim N(\boldsymbol{\mu}_y, \boldsymbol{\Sigma}_y), \tag{5.3}$$

with mean vector $\boldsymbol{\mu}_y = \boldsymbol{\eta} + \sigma \mathbf{m}_r$, $\mathbf{m}_r = (m_{r_1}, \ldots, m_{r_n})'$ and covariance matrix $\boldsymbol{\Sigma}_y = \sigma^2 \mathbf{S}_r$, $\mathbf{S}_r = \operatorname{diag}(s_{r_1}^2, \ldots, s_{r_n}^2)$, compare Section 3. The components m_{r_i} and $s_{r_i}^2$, $r_i \in \{1, \ldots, g_0\}$, are the mean and variance of the r_i-th error mixture component.

Prior of the latent component labels

On the second stage the distribution of the component labels $\mathbf{r} \mid \boldsymbol{\alpha}_0$ is the product of n discrete multinomial distributions $R_i \sim \text{MulNom}(1, \mathbf{w}(\boldsymbol{\alpha}_0))$ with density

$$p(\mathbf{r} \mid \boldsymbol{\alpha}_0) = \prod_{j=1}^{g_0} w_j^{n_j}(\boldsymbol{\alpha}) = \left(\sum_{j=1}^{g_0} \exp(\alpha_{0,j}) \right)^{-n} \prod_{j=1}^{g_0} \exp(n_j \alpha_{0,j}). \tag{5.4}$$

The probability of a single component label is $p(r_i \,|\, \boldsymbol{\alpha}_0) = w_{r_i}(\boldsymbol{\alpha}_0)$ and the last term in expression (5.4) results with the reparametrization of the mixture weights (2.9).

Prior of the unregularized linear effects

The prior distribution of the unregularized regressions coefficients $\boldsymbol{\gamma} = (\gamma_{0,}...,\gamma_{p_u})'$ in the predictor can be taken to be the product of independent diffuse priors $p(\gamma_j) \propto \text{const.}$, $j = 1,...,p_u$, or the product of independent zero mean, highly dispersed Gaussian priors. In the second case we obtain a multivariate Gaussian prior

$$\boldsymbol{\gamma} \,|\, \boldsymbol{\mu}_\gamma, \boldsymbol{\Sigma}_\gamma \sim N\left(\boldsymbol{\mu}_\gamma, \boldsymbol{\Sigma}_\gamma\right), \tag{5.5}$$

with $\boldsymbol{\mu}_\gamma = \mathbf{0}$ and $\boldsymbol{\Sigma}_\gamma^{-1} \to \mathbf{0}$. In both cases the prior of $\boldsymbol{\gamma}$ is denoted with $p(\boldsymbol{\gamma})$. The general form in (5.5) is used to derive the general structure of the full conditionals when the prior is a multivariate Gaussian distribution.

Prior of the regularized linear effects

As seen in the Sections 4.1 to 4.3, the general form of the priors for the regularized regression coefficients $\boldsymbol{\beta} = (\beta_1,...,\beta_{p_x})'$ are zero mean Gaussian distributions $\beta_j \,|\, \tau_{\beta_j}^2 \sim N(0, \tau_{\beta_j}^2)$, where the variance parameters $\tau_{\beta_j}^2$ in combination with further hyperparameters $\boldsymbol{\rho}$ drive the specific kind of shrinkage or variable selection. Under the conditional independence assumption we obtain a multivariate Gaussian prior

$$\boldsymbol{\beta} \,|\, \boldsymbol{\tau}_\beta^2 \sim N\left(\mathbf{0}, \boldsymbol{\Sigma}_\beta\right) \tag{5.6}$$

with diagonal covariance matrix $\boldsymbol{\Sigma}_\beta = \mathbf{D}_{\tau_\beta} = \text{diag}(\tau_{\beta_1}^2,...,\tau_{\beta_{px}}^2)$, that determines the shrinkage of the regression coefficients towards the mean $\boldsymbol{\mu}_\beta = \mathbf{0}$.

The associated priors for the variance and shrinkage parameters are:

Bayesian ridge version (A) ($\boldsymbol{\rho} = (\lambda_1,...,\lambda_{p_x})$)

$$\tau_{\beta_j}^2 \,|\, \lambda \sim \delta_{1/2\lambda_j}(\tau_{\beta_j}^2)\,,\, j = 1,...,p_x\,, \tag{5.7}$$

$$\lambda_j \sim_{iid} \text{Gamma}\left(h_{1,\lambda}, h_{1,\lambda}\right); \quad h_{1,\lambda}, h_{1,\lambda} > 0\,,\, j = 1,...,p_x\,. \tag{5.8}$$

Bayesian ridge version (B) ($\boldsymbol{\rho} = \lambda$)

$$\tau_\beta^2 \,|\, \lambda \sim \delta_{1/2\lambda}(\tau_\beta^2)\,,\, j = 1,...,p_x\,, \tag{5.9}$$

$$\lambda \sim \text{Gamma}\left(h_{1,\lambda}, h_{1,\lambda}\right); \quad h_{1,\lambda}, h_{1,\lambda} > 0\,. \tag{5.10}$$

Bayesian lasso ($\boldsymbol{\rho} = \lambda$)

$$\tau_{\beta_j}^2 \,|\, \lambda^2 \sim_{iid} \text{Exp}\left(\frac{\lambda^2}{2}\right),\; j = 1,...,p_x\,, \tag{5.11}$$

$$\lambda^2 \sim \text{Gamma}\left(h_{1,\lambda}, h_{1,\lambda}\right); \quad h_{1,\lambda}, h_{1,\lambda} > 0\,. \tag{5.12}$$

Bayesian NMIG with $\tau^2_{\beta_j} = I_j \psi^2_j$ $(\boldsymbol{\rho} = \omega)$

$$I_j \mid v_0, v_1, \omega \sim_{\text{iid}} \text{Bernoulli}(\omega; v_0, v_1),\ j = 1, \ldots, p_x\,, \tag{5.13}$$

$$\psi^2_j \mid h_{1,\psi}, h_{2,\psi} \sim_{\text{iid}} \text{IGamma}(h_{1,\psi}, h_{2,\psi}), \quad h_{1,\psi}, h_{2,\psi} > 0\,, j = 1, \ldots, p_x\,, \tag{5.14}$$

$$\omega \sim \text{Beta}\left(h_{1,\omega}, h_{2,\omega}\right); \quad h_{1,\omega}, h_{2,\omega} > 0\,. \tag{5.15}$$

Prior of the nonlinear effects and the transformed mixture weights

For the basis function weights $\boldsymbol{\alpha}_j = (\alpha_{j,1}, \ldots., \alpha_{j,g_j})'$, $j = 0,1,\ldots,p_z$, of the nonlinear predictor components $f_j(\cdot)$ and the mixture error density $f_\varepsilon(\cdot)$ the priors are specified by random walks of d_j-th order. This is leading to an intrinsic Gaussian Markov Random Field (GMRF) prior as defined in the Section 4.6

$$\boldsymbol{\alpha}_j \mid \tau^2_{\alpha_j} \sim \text{N}\left(\mathbf{0}, \boldsymbol{\Sigma}^-_{\alpha_j}\right),\ j = 0,1,\ldots,p_z\,, \tag{5.16}$$

with covariance matrix $\boldsymbol{\Sigma}^-_{\alpha_j} := \tau^2_{\alpha_j} \mathbf{K}^-_j$, where $\mathbf{K}^-_j$ denotes a generalized inverse of the penalty matrix $\mathbf{K}_j$ with rank $\text{rg}(\mathbf{K}_j) = g_j - q_j$. In addition diffuse priors are used for the q_j coefficients $p(\alpha_{j,1}) \propto \text{const.}, \ldots,\ p(\alpha_{j,q_j}) \propto \text{const}$. The smoothness controlling variances $\tau^2_{\alpha_j}$ are equipped with inverse gamma distributions

$$\tau^2_{\alpha_j} \sim \text{IGamma}\left(h_{1,\tau_j}, h_{2,\tau_j}\right), h_{1,\tau_j}, h_{2,\tau_j} > 0 \quad j = 0,1,\ldots,p_z\,. \tag{5.17}$$

Prior of the scale parameter

Finally, for the scale parameter the prior is specified as an inverse gamma distribution

$$\sigma^2 \sim \text{InvGamma}\left(h_{\sigma,1}, h_{\sigma,2}\right),\ h_{1,\sigma}, h_{2,\sigma} > 0\,. \tag{5.18}$$

6. MCMC inference for the extended AFT model

In the following subsections the update of the model parameters and the sampling algorithm are described. Bayesian Inference of the model parameters is carried out with MCMC techniques by consecutively updating conditional posterior distributions (full conditionals) of single parameters or blocks of parameters given the rest of the parameters and the data. The full conditional for a group of parameters is proportional to the posterior distribution density and derived by disregarding all factors that are independent of the considered parameter group. The derived MCMC sampler is based on Gibbs sampling or Metropolis-Hastings (MH) within Gibbs sampling. *Gibbs sampling* is used, if the full conditional of the considered parameter or parameter block given the current values of the remaining parameters has a standard form. Sampling from standard distributions is also a very efficient way to achieve a new state of the Markov chain. Another general way is to perform a *Metropolis-Hastings* (MH) update, where at first a new candidate state is drawn from a proposal distribution and this candidate is then accepted or rejected as new state of the Markov chain based on the ratio of probability densities of the candidate and the current state of the chain. This method is often applied, if the full conditional has no closed form. E. g. for the update of the mixture weights a version of the Metropolis-Hastings algorithm based on *IWLS proposals*, as proposed in Gamerman (1997) and described in Brezger and Lang (2006), is used. With symmetric (Gaussian) proposals that are centered at the current state of the chain, we obtain the so called *Metropolis* update, a simplified special case of the MH update scheme. For univariate full conditionals from nonstandard distributions we use alternatively the *slice sampling* method of Neal (2003). The described inferential procedure is implemented in the function `baftpgm()`, compare the Appendix D.5 for a description and the usage.

6.1. Conditional posterior densities

From the Bayesian theorem (1.15) the joint posterior distribution of the model and augmented parameters is obtained as

$$\mathrm{p}(\mathbf{y},\mathbf{r},\boldsymbol{\theta}\,|\,\mathfrak{D}) \propto \mathrm{L}(\mathfrak{D}\,|\,\mathbf{y},\mathbf{r},\boldsymbol{\theta})\mathrm{p}(\mathbf{y},\mathbf{r},\boldsymbol{\theta}),$$

with the likelihood from (3.2)

$$\mathrm{L}(\mathfrak{D}\,|\,\mathbf{y},\mathbf{r},\boldsymbol{\theta}) = \mathrm{L}(\mathfrak{D}\,|\,\mathbf{y}) = \prod_{i=1}^{n} 1_{[\tilde{y}_i,\infty)}(y_i)^{1-d_i},$$

and the prior from (5.1) and (5.2)

$$\mathrm{p}(\mathbf{y},\mathbf{r},\boldsymbol{\theta}) = \mathrm{p}(\mathbf{y}\,|\,\mathbf{r},\boldsymbol{\alpha}_1,\ldots,\boldsymbol{\alpha}_{p_z},\boldsymbol{\beta},\boldsymbol{\gamma},\sigma)\mathrm{p}(\mathbf{r}\,|\,\boldsymbol{\alpha}_0)\prod_{j=0}^{p_z}\mathrm{p}(\boldsymbol{\alpha}_j\,|\,\tau^2_{\alpha_j})\mathrm{p}(\tau^2_{\alpha_j})\mathrm{p}(\boldsymbol{\beta}\,|\,\boldsymbol{\tau}^2_{\beta})\mathrm{p}(\boldsymbol{\tau}^2_{\beta}\,|\,\boldsymbol{\rho})\mathrm{p}(\boldsymbol{\rho})\mathrm{p}(\boldsymbol{\gamma})\mathrm{p}(\sigma^2),$$

with specifications (5.4) to (5.18). The shown hierarchical structure of the priors with the implied independence assumptions simplifies in the following the derivation of the full conditionals for the model parameters of the extended AFT.

6.1.1. Full conditionals of the predictor components

Unregularized linear regression coefficients $\boldsymbol{\gamma}$

The full conditional of the unregularized linear regression coefficients $\boldsymbol{\gamma}$ is obtained by using the proportionality of the posterior to the product of the multivariate Gaussian prior of the log-survival times $\mathrm{p}(\mathbf{y}\,|\,\mathbf{r},\boldsymbol{\alpha}_1,\ldots,\boldsymbol{\alpha}_{p_z},\boldsymbol{\beta},\boldsymbol{\gamma},\sigma) = \mathrm{p}(\mathbf{y}\,|\,\boldsymbol{\mu}_y,\boldsymbol{\Sigma}_y)$ given in (5.3) and the multivariate Gaussian prior of the regression coefficients $\mathrm{p}(\boldsymbol{\gamma}) = \mathrm{p}(\boldsymbol{\gamma}\,|\,\boldsymbol{\mu}_\gamma,\boldsymbol{\Sigma}_\gamma)$ from (5.5). To simplify the notation, working observations $\tilde{\mathbf{y}}_\gamma := \mathbf{y} - (\boldsymbol{\eta} - \mathbf{U}\boldsymbol{\gamma}) - \sigma\mathbf{m}_r = \mathbf{y} - \boldsymbol{\mu}_y + \mathbf{U}\boldsymbol{\gamma}$, with $\boldsymbol{\mu}_y = \boldsymbol{\eta} + \sigma\mathbf{m}_r$, are introduced by deleting the component $\mathbf{U}\boldsymbol{\gamma}$ from the predictor. In summary we get

$$\begin{aligned}
\mathrm{p}(\boldsymbol{\gamma}\,|\,\cdot) &\propto \mathrm{p}(\mathbf{y}\,|\,\boldsymbol{\mu}_y,\boldsymbol{\Sigma}_y)\mathrm{p}(\boldsymbol{\gamma}\,|\,\boldsymbol{\mu}_\gamma,\boldsymbol{\Sigma}_\gamma) \\
&\propto \exp\left(-\frac{1}{2}(\mathbf{y}-\boldsymbol{\mu}_y)'\boldsymbol{\Sigma}_y^{-1}(\mathbf{y}-\boldsymbol{\mu}_y) - \frac{1}{2}(\boldsymbol{\gamma}-\boldsymbol{\mu}_\gamma)'\boldsymbol{\Sigma}_\gamma^{-1}(\boldsymbol{\gamma}-\boldsymbol{\mu}_\gamma)\right) \\
&\propto \exp\left(-\frac{1}{2}(\tilde{\mathbf{y}}_\gamma-\mathbf{U}\boldsymbol{\gamma})'\boldsymbol{\Sigma}_y^{-1}(\tilde{\mathbf{y}}_\gamma-\mathbf{U}\boldsymbol{\gamma}) - \frac{1}{2}(\boldsymbol{\gamma}-\boldsymbol{\mu}_\gamma)'\boldsymbol{\Sigma}_\gamma^{-1}(\boldsymbol{\gamma}-\boldsymbol{\mu}_\gamma)\right) \\
&\propto \exp\left(-\frac{1}{2}\left(\boldsymbol{\gamma}'(\mathbf{U}'\boldsymbol{\Sigma}_y^{-1}\mathbf{U}+\boldsymbol{\Sigma}_\gamma^{-1})\boldsymbol{\gamma} - 2\boldsymbol{\gamma}'(\mathbf{U}'\boldsymbol{\Sigma}_y^{-1}\tilde{\mathbf{y}}_\gamma+\boldsymbol{\Sigma}_\gamma^{-1}\boldsymbol{\mu}_\gamma)\right)\right).
\end{aligned} \tag{6.1}$$

This full conditional has the kernel of a multivariate Gaussian distribution, $\boldsymbol{\gamma}\,|\,\cdot \sim \mathrm{N}(\boldsymbol{\mu}_{\gamma|\cdot},\boldsymbol{\Sigma}_{\gamma|\cdot})$, with mean vector $\boldsymbol{\mu}_{\gamma|\cdot} = \mathbb{E}(\boldsymbol{\gamma}\,|\,\cdot)$ and covariance matrix $\boldsymbol{\Sigma}_{\gamma|\cdot} = \mathbb{C}\mathrm{ov}(\boldsymbol{\gamma}\,|\,\cdot)$ given by

$$\boldsymbol{\mu}_{\gamma|\cdot} = \boldsymbol{\Sigma}_{\gamma|\cdot}(\mathbf{U}'\boldsymbol{\Sigma}_y^{-1}\tilde{\mathbf{y}}_\gamma + \boldsymbol{\Sigma}_\gamma^{-1}\boldsymbol{\mu}_\gamma), \quad \boldsymbol{\Sigma}_{\gamma|\cdot} = (\mathbf{U}'\boldsymbol{\Sigma}_y^{-1}\mathbf{U} + \boldsymbol{\Sigma}_\gamma^{-1})^{-1}. \tag{6.2}$$

It turns out, that the multivariate Gaussian distribution is the general form of the full conditionals for all predictor components and, as a consequence, Gibbs sampling can be employed. The derivation (6.1) serves as the building block to obtain the full conditionals of the remaining predictor components, where the specific mean and covariance structure of (6.2) results from the specific prior structure. Since the linear effects $\boldsymbol{\gamma}$ are assumed to be unregularized, we simply set $\boldsymbol{\mu}_\gamma = \mathbf{0}$ and the precision to $\boldsymbol{\Sigma}_\gamma^{-1} = \mathbf{0}$ and get

$$\boldsymbol{\mu}_{\gamma|\cdot} = \boldsymbol{\Sigma}_{\gamma|\cdot}\mathbf{U}'\boldsymbol{\Sigma}_y^{-1}\tilde{\mathbf{y}}_\gamma, \quad \boldsymbol{\Sigma}_{\gamma|\cdot} = (\mathbf{U}'\boldsymbol{\Sigma}_y^{-1}\mathbf{U})^{-1}. \tag{6.3}$$

Regularized linear regression coefficients $\boldsymbol{\beta}$

With (5.3) and (5.6) the full conditional of the regularized linear effects is proportional to the product $p(\boldsymbol{\beta}\,|\,\cdot) \propto p(\mathbf{y}\,|\,\boldsymbol{\mu}_y, \boldsymbol{\Sigma}_y)p(\boldsymbol{\beta}\,|\,\boldsymbol{\mu}_\beta, \boldsymbol{\Sigma}_\beta)$. The mean vector of the Gaussian prior of the regularized regression coefficients equals zero, $\boldsymbol{\mu}_\beta = \mathbf{0}$, and the covariance matrix is given by $\boldsymbol{\Sigma}_\beta = \mathrm{diag}(\tau_{\beta,1}^2, \ldots \tau_{\beta,p_u}^2)$. Adapting the general form in (6.2) to the working observations $\tilde{\mathbf{y}}_\beta := \mathbf{y} - \boldsymbol{\mu}_y + \mathbf{X}\boldsymbol{\beta}$, we get for the regularized linear effects also a multivariate Gaussian full conditional distribution, $\boldsymbol{\beta}\,|\,\cdot \sim N(\boldsymbol{\mu}_{\beta|\cdot}, \boldsymbol{\Sigma}_{\beta|\cdot})$, with mean vector $\boldsymbol{\mu}_{\beta|\cdot} = \mathbb{E}(\boldsymbol{\beta}\,|\,\cdot)$ and covariance matrix $\boldsymbol{\Sigma}_{\beta|\cdot} = \mathbb{C}\mathrm{ov}(\boldsymbol{\beta}\,|\,\cdot)$ defined by

$$\boldsymbol{\mu}_{\beta|\cdot} = \boldsymbol{\Sigma}_{\beta|\cdot}\mathbf{X}'\boldsymbol{\Sigma}_y^{-1}\tilde{\mathbf{y}}_\beta, \quad \boldsymbol{\Sigma}_{\beta|\cdot} = (\mathbf{X}'\boldsymbol{\Sigma}_y^{-1}\mathbf{X} + \boldsymbol{\Sigma}_\beta^{-1})^{-1}. \tag{6.4}$$

Regularized nonlinear regression coefficients $\boldsymbol{\alpha}_j$

In completely concordance to the former derivations we obtain with (5.3) and (5.16) the full conditional for the basis coefficients $\boldsymbol{\alpha}_j$, $j = 1, \ldots, p_z$, via $p(\boldsymbol{\alpha}_j\,|\,\cdot) \propto p(\mathbf{y}\,|\,\boldsymbol{\mu}_y, \boldsymbol{\Sigma}_y)p(\boldsymbol{\alpha}_j\,|\,\boldsymbol{\mu}_{\alpha_j}, \boldsymbol{\Sigma}_{\alpha_j}^-)$, where $\boldsymbol{\Sigma}_{\alpha_j}^- = \boldsymbol{\Pi}_{\alpha_j}$ denotes the generalized inverse of the prior precision matrix $\boldsymbol{\Pi}_{\alpha_j} = \tau_j^{-2}\mathbf{K}_j$. Since the prior mean is $\boldsymbol{\mu}_{\alpha_j} = \mathbf{0}$, we get with (6.2) and the working observations $\tilde{\mathbf{y}}_{\alpha_j} = \mathbf{y} - \boldsymbol{\mu}_y + \mathbf{Z}_j\boldsymbol{\alpha}_j$ multivariate Gaussian distributions $\boldsymbol{\alpha}_j\,|\,\cdot \sim N(\boldsymbol{\mu}_{\boldsymbol{\alpha}_j|\cdot}, \boldsymbol{\Sigma}_{\boldsymbol{\alpha}_j|\cdot})$, $j = 1, \ldots, p_z$, as full conditionals. The mean vector $\boldsymbol{\mu}_{\boldsymbol{\alpha}_j|\cdot} = \mathbb{E}(\boldsymbol{\alpha}_j\,|\,\cdot)$ and covariance matrix $\boldsymbol{\Sigma}_{\boldsymbol{\alpha}_j|\cdot} = \mathbb{C}\mathrm{ov}(\boldsymbol{\alpha}_j\,|\,\cdot)$ are given by

$$\boldsymbol{\mu}_{\boldsymbol{\alpha}_j|\cdot} = \boldsymbol{\Sigma}_{\boldsymbol{\alpha}_j|\cdot}\mathbf{Z}_j'\boldsymbol{\Sigma}_y^{-1}\tilde{\mathbf{y}}_{\alpha_j}, \quad \boldsymbol{\Sigma}_{\boldsymbol{\alpha}_j|\cdot} = (\mathbf{Z}_j'\boldsymbol{\Sigma}_y^{-1}\mathbf{Z}_j + \boldsymbol{\Pi}_{\alpha_j})^{-1}. \tag{6.5}$$

Centering of the spline: In general the mean levels of the unknown functions $f_j(\cdot)$ are not identifiable. To ensure identifiability of the model, we center the estimated functions $f_j(\cdot)$ in every iteration of the sampler to have zero mean $\overline{f}_j^* = n^{-1}\sum_{i=1}^n f_j^*(z_{i,j}) = 0$. To do so, we recompute the basis function weights $\boldsymbol{\alpha}_j$ as

$$\alpha_{j,k}^* = \alpha_{j,k} - c_j, \tag{6.6}$$

where $c_j = n^{-1}\sum_{i=1}^n f_j(z_{i,j}) = n^{-1}\sum_{i=1}^n \sum_{l=1}^{g_j} \alpha_{j,k}B_{j,k}(z_{i,j})$ denotes the mean of the function evaluations $f_j(\cdot)$ at the observed data points. We verify that $\overline{f}_j^* := n^{-1}\sum_{i=1}^n \sum_{j=1}^{g_j} \alpha_{j,k}^* B_{j,k}(z_{i,j})$ has zero mean since $\sum_{i=1}^n \sum_{j=1}^{g_j} \alpha_{j,k}^* B_{j,k}(z_{i,j}) = \sum_{i=1}^n \sum_{j=1}^{g_j} \alpha_{j,k}B_{j,k}(z_{i,j}) - \sum_{i=1}^n c_j \sum_{j=1}^{g_j} B_{j,k}(z_{i,j}) = \sum_{i=1}^n f_j(z_{i,j}) - nc_j = 0$, where we used in the fourth equation the fact $\sum_{k=1}^{g_j} B_{j,k}(z) = 1$.

The shift of the basis function weights does not affect the penalty, since the differences $\Delta^{(d_j)}\alpha_{j,k}$ are invariant, i. e. $\Delta^{(d_j)}\alpha_{j,k} = \Delta^{(d_j)}(\alpha_{j,k} - c)$ for any scalar c, and the addition of the subtracted means $c_j \in \mathbb{R}$, $j = 1, \ldots, p_z$, to the intercept γ_0 avoids that the posterior is changed.

6.1.2. Full conditionals of the regularization parameters

Ridge regularization

Version (A): With the deterministic connection $\tau^2_{\beta_j} = 1/2\lambda_j$ in (5.7) the Gaussian prior of the regression coefficients from (5.6) becomes $p(\beta_j \mid \lambda_j) \propto \sqrt{\lambda_j}\exp(-\lambda_j\beta_j^2)$. The full conditional of the shrinkage parameter λ_j is proportional to the product of this prior and the gamma prior of the shrinkage parameter $p(\lambda_j) \propto \lambda_j^{h_{1,\lambda}-1}\exp(-h_{2,\lambda}\lambda_j)$ from (5.8). We simply see that we obtain gamma densities

$$\lambda_j \mid \cdot \sim \text{Gamma}\left(h_{1,\lambda} + \frac{1}{2}, h_{2,\lambda} + \beta_j^2\right),\ j = 1, \ldots p_x, \tag{6.7}$$

as full conditionals.

Version (B): With a similar argumentation we obtain the full conditional of the shrinkage parameter λ. Using the product of $p(\boldsymbol{\beta} \mid \lambda) \propto \sqrt{\lambda_j}^{p_x}\exp(-\lambda\sum_{j=1}^{p_x}\beta_j^2)$ from (5.6) and the gamma prior of the shrinkage parameter $p(\lambda) \propto \lambda^{h_{1,\lambda}-1}\exp(-h_{2,\lambda}\lambda)$ from (5.10) we obtain also a gamma density

$$\lambda \mid \cdot \sim \text{Gamma}\left(h_{1,\lambda} + \frac{p_x}{2}, h_{2,\lambda} + \sum_{j=1}^{p_x}\beta_j^2\right). \tag{6.8}$$

Lasso regularization

We derive the full conditionals for the variance parameters $\tau^2_{\beta_j}$ from the product of the prior for the regression coefficients $p(\beta_j \mid \tau^2_{\beta_j}) \propto \tau^{-1}_{\beta_j}\exp(-\beta_j^2/2\tau^2_{\beta_j})$, compare (5.6), and the exponential prior for the variance parameters $p(\tau^2_{\beta_j} \mid \lambda^2) \propto \lambda^2\exp(-\lambda^2\tau^2_{\beta_j}/2)$ from (5.11)

$$p(\tau^2_{\beta_j} \mid \cdot) \propto \tau^{-1}_{\beta_j}\exp\left\{-\beta_j^2/2\tau^2_{\beta_j} - \lambda^2\tau^2_{\beta_j}/2\right\} = \frac{1}{\sqrt{\tau^2_{\beta_j}}}\exp\left\{-\frac{\lambda^2\tau^2_{\beta_j}}{2}\frac{\beta_j^2}{\lambda^2}\left((\tau^{-2}_{\beta_j})^2 + \lambda^2\beta_j^{-2}\right)\right\}$$

$$\propto \sqrt{\frac{1}{\tau^2_{\beta_j}}}\exp\left\{-\frac{\lambda^2\tau^2_{\beta_j}}{2}\frac{\beta_j^2}{\lambda^2}\left(\frac{1}{\tau^2_{\beta_j}} - \frac{\lambda}{|\beta_j|}\right)^2\right\}.$$

Using the definition $\mu_j := \sqrt{\lambda^2/\beta_j^2}$ and applying the change of variables $t_j^{\,2} = 1/\tau^2_{\beta_j}$ is leading to $p(t_j^2 \mid \cdot) \propto (t_j^2)^{-3/2}\exp(-\lambda^2(2\mu_j^2 t_j^2)^{-1}(t_j^2 - \mu_j)^2)$, which is the kernel of an inverse Gaussian density with mean $\mu_j > 0$ and shape parameter λ^2. Finally the full conditionals for the variance parameters $\tau^2_{\beta_j}$ are inverse Gaussian distributions

$$\frac{1}{\tau^2_{\beta_j}} \mid \cdot \sim \text{InvGauss}\left(\frac{\sqrt{\lambda^2}}{|\beta_j|}, \lambda^2\right)\quad, j = 1, .., p_x\,. \tag{6.9}$$

The full conditional for the quadratic shrinkage parameter is proportional to the product of the gamma prior $p(\lambda^2) \propto (\lambda^2)^{h_{1,\lambda}-1}\exp(-h_{2,\lambda}\lambda^2)$ from (5.12) and the product of the exponential priors of the variance parameters $p(\boldsymbol{\tau}^2_\beta \mid \lambda^2) \propto (\lambda^2)^{p_x}\exp(-\lambda^2\sum_{j=1}^{p_x}\tau^2_{\beta_j}/2)$, compare (5.11). We obtain as full conditional the gamma density

$$\lambda^2 \mid \cdot \sim \text{Gamma}\left(h_{1,\lambda} + p_x, h_{2,\lambda} + \frac{1}{2}\sum_{j=1}^{p_x}\tau^2_{\beta_j}\right). \tag{6.10}$$

NMIG regularization

Under the Bayesian NMIG prior the variance parameter is the product of two variance components $\tau^2_{\beta_j} = I_j \psi^2_j$. The full conditionals of the covariate-specific binary indicator variables I_j are Bernoulli distributions

$$p(I_j \mid \cdot) = \left(1 - \frac{1}{1 + A_j/B_j}\right)^{\delta_{v_0}(I_j)} \left(\frac{1}{1 + A_j/B_j}\right)^{\delta_{v_1}(I_j)}, \tag{6.11}$$

with

$$\frac{A_j}{B_j} = \frac{1-\omega}{\omega} \frac{\sqrt{v_1}}{\sqrt{v_0}} \exp\left\{ \frac{(v_0 - v_1)}{v_0 v_1} \frac{\beta_j^2}{2\psi_j^2} \right\},$$

which is derived from the product of the indicator prior $p(I_j \mid \omega, v_0, v_1) \propto \omega^{\delta_{v_1}(I_j)} (1-\omega)^{\delta_{v_0}(I_j)}$ in (5.13) and the prior of the regression coefficients $p(\beta_j \mid \tau^2_{\beta_j}) \propto \sqrt{I_j^{-1}} \exp(-\beta_j^2 / 2 I_j \psi_j^2)$ in (5.6).

For the second variance parameter component ψ_j^2 the full conditionals are proportional to the product of the Gaussian prior of the regression coefficients $p(\beta_j \mid \tau^2_{\beta_j}) \propto \sqrt{\psi_j^{-2}} \exp(-\beta_j^2 / 2 I_j \psi_j^2)$ in (5.13) and the inverse gamma prior in (5.14) $p(\psi_j^2) \propto (\psi_j^2)^{-h_{\psi,1}-1} \exp(-h_{\psi,2} / \psi_j^2)$, which results in inverse gamma densities

$$\psi_j^2 \mid \cdot \sim \text{InvGamma}\left(h_{1,\psi} + \frac{1}{2}, h_{2,\psi} + \frac{\beta_j^2}{2 I_j} \right), j = 1,..,p_x . \tag{6.12}$$

With the beta prior (5.15) for the complexity parameter $p(\omega) \propto \omega^{h_{1,\omega}-1} (1-\omega)^{h_{2,\omega}-1}$ and the product of the indicator priors (5.13) $p(\mathbf{I} \mid \omega, v_0, v_1) \propto \omega^{n_1} (1-\omega)^{n_0}$, with $n_0 := \#\{j : I_j = v_0\}$ and $n_1 := \#\{j : I_j = v_1\}$, the full conditional for the mixing parameter is also a beta density

$$\omega \mid \cdot \sim \text{Beta}\left(h_{1,\omega} + n_1 ; h_{2,\omega} + n_0\right). \tag{6.13}$$

Smoothing parameters

The full conditionals of the smoothing parameters are proportional to the product of the inverse gamma prior $p(\tau^2_{\alpha_j}) \propto (\tau^2_{\alpha_j})^{-h_{1,\tau_j}-1} \exp(-h_{2,\tau_j} / \tau^2_{\alpha_j})$ from (5.17) and the partial improper multivariate Gaussian prior of the basis function weights $p(\boldsymbol{\alpha}_j \mid \tau^2_{\alpha_j}) \propto (\tau^2_{\alpha_j})^{\text{rank}(\mathbf{K}_j)/2} \exp(-\boldsymbol{\alpha}_j' K_j \boldsymbol{\alpha}_j / 2\tau^2_{\alpha_j})$ from (5.16). We can easy reproduce that the full conditionals are all proper inverse gamma distributions

$$\tau^2_{\alpha_j} \mid \cdot \sim \text{InvGamma}\left(h_{1,\tau_j} + \frac{\text{rank}(\mathbf{K}_j)}{2}, h_{2,\tau_j} + \frac{1}{2} \boldsymbol{\alpha}_j' \mathbf{K}_j \boldsymbol{\alpha}_j \right), \ j = 0,1,...,p_z . \tag{6.14}$$

6.1.3. Full conditional of the mixture weights

In this section we derive several alternatives for the update of the (transformed) mixture weights. Besides a single-update of the mixture weights we consider several block-update schemes to investigate the impact on the convergence of the mixture weights in combination with a standardization of the error distribution within the MCMC sampler, compare Subsection 6.2.1. The particular method is specified in the function `baftpgm()` through the argument `method.alpha` within the `errorpri` list, compare Appendix D.5.

Update scheme "mhcond" (Metropolis-Hastings based block update)

To update the block of transformed mixture weights, we use Metropolis-Hastings steps with IWLS proposals as in detail described e. g. in Brezger and Lang (2006) and shortly summarized in the Appendix C. The general idea of IWLS proposals is to obtain a Gaussian proposal by matching the mode and the curvature of the full conditional at the current state of parameter vector in each update step. The proposal distribution is constructed by a second order Taylor expansion of the logarithm of the full conditional at the current state of the chain. The full conditional for the transformed mixture weights is proportional to the product of the smoothing prior (5.16) and the prior of the component labels (5.4)

$$p\left(\boldsymbol{\alpha}_0 \mid \cdot\right) \propto p\left(\mathbf{r} \mid \boldsymbol{\alpha}_0\right) p\left(\boldsymbol{\alpha}_0 \mid \tau^2_{\alpha_0}\right) \propto \exp\left(\sum_{j=1}^{g_0} n_j \log\left(w_j(\boldsymbol{\alpha}_0)\right) - \frac{1}{2\tau^2_{\alpha_0}} \boldsymbol{\alpha}_0' \mathbf{K}_0 \boldsymbol{\alpha}_0 \right).$$

Due to the identifiability constraint $\alpha_{0,k} = 0$, for one $k \in \{1,\ldots,g_0\}$, the mixture weights $w_j(\boldsymbol{\alpha}_0)$ depend effectively on the parameters $\tilde{\boldsymbol{\alpha}}_0 := (\alpha_{0,1},\ldots,\alpha_{0,k-1},\alpha_{0,k+1},\ldots,\alpha_{0,g})'$ and also the full conditional depends effectively on $\tilde{\boldsymbol{\alpha}}_0$. With respect to this constraint we write the full conditional as

$$p\left(\tilde{\boldsymbol{\alpha}}_0 \mid \cdot\right) \propto \exp\left(\sum_{\ell=1,\ell\neq k}^{g_0} n_\ell \alpha_{0,\ell} - n \cdot \log\left(1 + \sum_{\ell=1,\ell\neq k}^{g_0} \exp(\alpha_{0,\ell}) \right) - \frac{1}{2\tau^2_{\alpha_0}} \tilde{\boldsymbol{\alpha}}_0' \tilde{\mathbf{K}}_0 \tilde{\boldsymbol{\alpha}}_0 \right), \tag{6.15}$$

where $\tilde{\mathbf{K}}_0$ denotes the reduced difference matrix, if the k-th row and k-th column are removed from $\mathbf{K}_0$. To construct the IWLS proposal, the score vector $\mathbf{s}_{\tilde{\alpha}_0}(\tilde{\boldsymbol{\alpha}}_0)$ and the Hessian matrix $\mathbf{H}_{\tilde{\alpha}_0}(\tilde{\boldsymbol{\alpha}}_0)$ of the logarithm of the full conditional, $f(\tilde{\boldsymbol{\alpha}}_0) = \log(p(\tilde{\boldsymbol{\alpha}}_0 \mid \cdot))$, are required. With

$$g(\tilde{\boldsymbol{\alpha}}_0) = \sum_{\ell=1,\ell\neq k}^{g_0} n_\ell \alpha_{0,\ell} - n \cdot \log\left(1 + \sum_{\ell=1,\ell\neq k}^{g_0} \exp(\alpha_{0,\ell}) \right),$$

we derive the first and second order partial derivates of $g(\cdot)$ as

$$\frac{\partial g(\tilde{\boldsymbol{\alpha}}_0)}{\partial \alpha_{0,j}} = n_j - n\, w_j(\boldsymbol{\alpha}_0), \quad j \neq k,$$

$$\frac{\partial^2 g(\tilde{\boldsymbol{\alpha}}_0)}{\partial \alpha_{0,j} \partial \alpha_{0,i}} = -n \frac{\partial w_j(\boldsymbol{\alpha}_0)}{\partial \alpha_{0,i}} = \begin{cases} n\left(w_j(\boldsymbol{\alpha}_0)^2 - w_j(\boldsymbol{\alpha}_0)\right) & i = j, \text{ and } i,j \neq k \\ n w_j(\boldsymbol{\alpha}_0) w_i(\boldsymbol{\alpha}_0) & i \neq j, \text{ and } i,j \neq k. \end{cases} \tag{6.16}$$

With the definitions $\tilde{\mathbf{n}} := (n_1,\ldots,n_{k-1},n_{k+1}\ldots,n_{g_0})'$, $\tilde{\mathbf{w}}(\boldsymbol{\alpha}_0) := (w_1(\boldsymbol{\alpha}_0),\ldots,w_{k-1}(\boldsymbol{\alpha}_0),w_{k+1}(\boldsymbol{\alpha}_0)\ldots,w_{g_0}(\boldsymbol{\alpha}_0))'$ and $\tilde{\mathbf{W}}(\boldsymbol{\alpha}_0) := \operatorname{diag}\left(\tilde{\mathbf{w}}(\boldsymbol{\alpha}_0)\right) - \tilde{\mathbf{w}}(\boldsymbol{\alpha}_0)\tilde{\mathbf{w}}'(\boldsymbol{\alpha}_0)$ the score vector and Hessian matrix of the function $f(\tilde{\boldsymbol{\alpha}}_0) := \log(p(\tilde{\boldsymbol{\alpha}}_0 \mid \cdot))$ are

$$\mathbf{s}_{\tilde{\alpha}_0}\left(\tilde{\boldsymbol{\alpha}}_0\right) = \tilde{\mathbf{n}} - n \cdot \tilde{\mathbf{w}}(\tilde{\boldsymbol{\alpha}}_0) - \frac{1}{\tau^2_{\alpha_0}} \tilde{\mathbf{K}}_0 \tilde{\boldsymbol{\alpha}}_0, \quad \mathbf{H}_{\tilde{\alpha}_0}\left(\tilde{\boldsymbol{\alpha}}_0\right) = -n\tilde{\mathbf{W}}(\boldsymbol{\alpha}_0) - \frac{1}{\tau^2_{\alpha_0}} \tilde{\mathbf{K}}_0.$$

We note from (6.16) that in general the matrix of the first derivates of the weights $\mathbf{w}(\boldsymbol{\alpha}_0)$ with respect to $\boldsymbol{\alpha}_0$ is given by $\mathbf{W}(\boldsymbol{\alpha}_0) := \partial \mathbf{w}(\boldsymbol{\alpha}_0)/\partial \boldsymbol{\alpha}_0 = (\partial w_i(\boldsymbol{\alpha}_0)/\partial \alpha_{0,j})_{i,j=1,\ldots,g_0} = \operatorname{diag}\left(\mathbf{w}(\boldsymbol{\alpha}_0)\right) - \mathbf{w}(\boldsymbol{\alpha}_0)\mathbf{w}'(\boldsymbol{\alpha}_0)$, and we obtain the representation $\tilde{\mathbf{W}}(\boldsymbol{\alpha}_0)$ by removing the k-th row and k-th column from $\mathbf{W}(\boldsymbol{\alpha}_0)$, i. e. $\tilde{\mathbf{W}}(\boldsymbol{\alpha}_0) = \mathbf{W}(\boldsymbol{\alpha}_0)[-k,-k]$. The second order Taylor expansion of $f(\tilde{\boldsymbol{\alpha}}_0) = \log(p(\tilde{\boldsymbol{\alpha}}_0 \mid \cdot))$ around the current state of the parameter vector $\boldsymbol{\alpha}_0^{(c)}$ has the general form

$$\hat{f}\left(\tilde{\boldsymbol{\alpha}}_0\right) \approx f\left(\tilde{\boldsymbol{\alpha}}_0^{(c)}\right) + \left(\tilde{\boldsymbol{\alpha}}_0 - \tilde{\boldsymbol{\alpha}}_0^{(c)}\right)' \mathbf{s}_{\tilde{\alpha}_0}\left(\tilde{\boldsymbol{\alpha}}_0^{(c)}\right) + \frac{1}{2}\left(\tilde{\boldsymbol{\alpha}}_0 - \tilde{\boldsymbol{\alpha}}_0^{(c)}\right)' \mathbf{H}_{\tilde{\alpha}_0}\left(\tilde{\boldsymbol{\alpha}}_0^{(c)}\right)\left(\tilde{\boldsymbol{\alpha}}_0 - \tilde{\boldsymbol{\alpha}}_0^{(c)}\right).$$

Taking the exponential $\exp(\hat{\mathrm{f}}(\tilde{\boldsymbol{\alpha}}_0))$ and neglecting the components that do not depend on $\tilde{\boldsymbol{\alpha}}_0$ leads to a multivariate Gaussian distribution density

$$\varphi(\tilde{\boldsymbol{\alpha}}_0 \mid \hat{\boldsymbol{\mu}}_{\tilde{\alpha}_0}^{(c)}, \hat{\boldsymbol{\Sigma}}_{\tilde{\alpha}_0}^{(c)}) \propto \exp\left(\frac{1}{2} \tilde{\boldsymbol{\alpha}}_0' \mathbf{H}_{\tilde{\alpha}_0}\left(\tilde{\boldsymbol{\alpha}}_0^{(c)}\right)\tilde{\boldsymbol{\alpha}}_0 + \tilde{\boldsymbol{\alpha}}_0'\left(\mathbf{s}_{\tilde{\alpha}_0}\left(\tilde{\boldsymbol{\alpha}}_0^{(c)}\right) - \mathbf{H}_{\tilde{\alpha}_0}\left(\tilde{\boldsymbol{\alpha}}_0^{(c)}\right)\tilde{\boldsymbol{\alpha}}_0^{(c)}\right)\right), \tag{6.17}$$

with mean vector $\hat{\boldsymbol{\mu}}_{\tilde{\alpha}_0}^{(c)} = -\mathbf{H}_{\tilde{\alpha}_0}^{-1}(\tilde{\boldsymbol{\alpha}}_0^{(c)})\left(\mathbf{s}_{\tilde{\alpha}_0}(\tilde{\boldsymbol{\alpha}}_0^{(c)}) - \mathbf{H}_{\tilde{\alpha}_0}(\tilde{\boldsymbol{\alpha}}_0^{(c)})\tilde{\boldsymbol{\alpha}}_0^{(c)}\right) = \tilde{\boldsymbol{\alpha}}_0^{(c)} - \mathbf{H}_{\tilde{\alpha}_0}^{-1}(\tilde{\boldsymbol{\alpha}}_0^{(c)})\mathbf{s}_{\tilde{\alpha}_0}(\tilde{\boldsymbol{\alpha}}_0^{(c)})$ and covariance matrix $\hat{\boldsymbol{\Sigma}}_{\tilde{\alpha}_0}^{(c)} = -\mathbf{H}_{\tilde{\alpha}_0}^{-1}(\tilde{\boldsymbol{\alpha}}_0^{(c)})$. The representation of mean vector in the second equality shows, that the mean of the Gaussian proposal $\varphi(\cdot \mid \hat{\boldsymbol{\mu}}_{\tilde{\alpha}_0}^{(c)}, \hat{\boldsymbol{\Sigma}}_{\tilde{\alpha}_0}^{(c)})$ can be interpreted as one step approximation to the mode of the full conditional obtained by a single Fisher scoring step from the current state. To update the transformed mixture weights based on the current state $\tilde{\boldsymbol{\alpha}}_0^{(c)}$ of the chain, the new value $\tilde{\boldsymbol{\alpha}}_0^{(p)}$ is proposed by drawing a random number from the multivariate Gaussian proposal distribution $\mathrm{N}(\hat{\boldsymbol{\mu}}_{\tilde{\alpha}_0}^{(c)}, \hat{\boldsymbol{\Sigma}}_{\tilde{\alpha}_0}^{(c)})$ with density $\varphi(\cdot \mid \hat{\boldsymbol{\mu}}_{\tilde{\alpha}_0}^{(c)}, \hat{\boldsymbol{\Sigma}}_{\tilde{\alpha}_0}^{(c)})$, where the mean vector $\hat{\boldsymbol{\mu}}_{\tilde{\alpha}_0}^{(c)} = \mathbb{E}(\tilde{\boldsymbol{\alpha}}_0 \mid \tilde{\boldsymbol{\alpha}}_0^{(c)})$ and covariance matrix $\hat{\boldsymbol{\Sigma}}_{\tilde{\alpha}_0}^{(c)} = \mathbb{C}\mathrm{ov}(\tilde{\boldsymbol{\alpha}}_0 \mid \tilde{\boldsymbol{\alpha}}_0^{(c)})$ are given by

$$\hat{\boldsymbol{\mu}}_{\tilde{\alpha}_0}^{(c)} = \hat{\boldsymbol{\Sigma}}_{\tilde{\alpha}_0}^{(c)}\left(\tilde{\mathbf{n}} - \mathrm{n}\tilde{\mathbf{w}}\left(\boldsymbol{\alpha}_0^{(c)}\right) + \mathrm{n}\tilde{\mathbf{W}}\left(\boldsymbol{\alpha}_0^{(c)}\right)\tilde{\boldsymbol{\alpha}}_0^{(c)}\right), \quad \hat{\boldsymbol{\Sigma}}_{\tilde{\alpha}_0}^{(c)} = \left(\mathrm{n}\tilde{\mathbf{W}}\left(\boldsymbol{\alpha}_0^{(c)}\right) + \tau_{\alpha_0}^{-2}\tilde{\mathbf{K}}_0\right)^{-1}. \tag{6.18}$$

Finally, the proposed state $\tilde{\boldsymbol{\alpha}}_0^{(p)}$ is accepted as new state of the chain with probability

$$\mathrm{p}_{\text{accept}}\left(\tilde{\boldsymbol{\alpha}}_0^{(p)}, \tilde{\boldsymbol{\alpha}}_0^{(c)}\right) = \min\left\{1, \frac{\mathrm{p}\left(\tilde{\boldsymbol{\alpha}}_0^{(p)} \mid \cdot\right) \cdot \varphi\left(\tilde{\boldsymbol{\alpha}}_0^{(c)} \mid \hat{\boldsymbol{\mu}}_{\tilde{\alpha}_0}^{(p)}, \hat{\boldsymbol{\Sigma}}_{\tilde{\alpha}_0}^{(p)}\right)}{\mathrm{p}\left(\tilde{\boldsymbol{\alpha}}_0^{(c)} \mid \cdot\right) \cdot \varphi\left(\tilde{\boldsymbol{\alpha}}_0^{(p)} \mid \hat{\boldsymbol{\mu}}_{\tilde{\alpha}_0}^{(c)}, \hat{\boldsymbol{\Sigma}}_{\tilde{\alpha}_0}^{(c)}\right)}\right\}. \tag{6.19}$$

The nominator and denominator contain the full conditional (6.15) and proposal density (6.17), each evaluated at the current state $\tilde{\boldsymbol{\alpha}}_0^{(c)}$ and proposed state $\tilde{\boldsymbol{\alpha}}_0^{(p)}$, whereat $\hat{\boldsymbol{\mu}}_{\tilde{\alpha}_0}^{(p)}$ and $\hat{\boldsymbol{\Sigma}}_{\tilde{\alpha}_0}^{(p)}$ are obtained by inserting the proposal into (6.17) and computing the resulting mean and covariance. We specify this method with `method.alpha="mhcond"` in the function `baftpgm()`.

Update scheme “mhmarg” (Metropolis-Hastings based block update):

This update scheme is based on the marginal likelihood $\mathrm{f}_i(\mathrm{y}_i \mid \boldsymbol{\theta}) = \mathbf{w}'(\boldsymbol{\alpha}_0)\boldsymbol{\varphi}(\mathrm{y}_i)$ from (2.11), with $\boldsymbol{\varphi}(\mathrm{y}_i) := (\varphi_1(\mathrm{y}_i), \ldots, \varphi_{g_0}(\mathrm{y}_i))' = (\varphi(\mathrm{y}_i \mid \eta_i - \sigma \mathrm{m}_1, \sigma^2 \mathrm{s}_1^2), \ldots, \varphi(\mathrm{y}_i \mid \eta_i - \sigma \mathrm{m}_{g_0}, \sigma^2 \mathrm{s}_{g_0}^2))'$. With the smoothing prior (5.16) the full conditional is given as

$$\mathrm{p}\left(\boldsymbol{\alpha}_0 \mid \cdot\right) \propto \mathrm{p}\left(\mathbf{y} \mid \boldsymbol{\theta}\right)\mathrm{p}\left(\boldsymbol{\alpha}_0 \mid \tau_{\alpha_0}^2\right) \propto \exp\left(\sum_{i=1}^{n} \log\left(\mathbf{w}'(\boldsymbol{\alpha}_0)\boldsymbol{\varphi}(\mathrm{y}_i)\right) - \frac{1}{2\tau_{\alpha_0}^2}\boldsymbol{\alpha}_0'\mathbf{K}_0\boldsymbol{\alpha}_0\right).$$

Since one of the transformed mixture weights fulfills the identifiability constraint, i. e. $\alpha_{0,k} = 0$, $k \in \{1, \ldots, g_0\}$, we use from above the reduced vector $\tilde{\boldsymbol{\alpha}}_0 := (\alpha_{0,1}, \ldots, \alpha_{0,k-1}, \alpha_{0,k+1}, \ldots, \alpha_{0,g_0})'$ to construct the IWLS proposal, compare Appendix C. From the previous subsection the first derivate of the weights $\mathbf{w}(\boldsymbol{\alpha}_0)$ with respect to $\boldsymbol{\alpha}_0$ is given by $\mathbf{W}(\boldsymbol{\alpha}_0) := \partial \mathbf{w}(\boldsymbol{\alpha}_0)/\partial \boldsymbol{\alpha}_0 = \operatorname{diag}\left(\mathbf{w}(\boldsymbol{\alpha}_0)\right) - \mathbf{w}(\boldsymbol{\alpha}_0)\mathbf{w}'(\boldsymbol{\alpha}_0)$. With respect to the identifiability constraint we remove now the k-th column of the matrix $\mathbf{W}(\boldsymbol{\alpha}_0)$ that contains the first derivate of the weights $\mathbf{w}(\boldsymbol{\alpha}_0)$ with respect to $\alpha_{0,k}$, and define the resulting $g_0 \times (g_0 - 1)$ dimensional matrix as $\tilde{\mathbf{W}}(\boldsymbol{\alpha}_0) := \mathbf{W}(\boldsymbol{\alpha}_0)[,-k]$. With this representation we write the score vector of the function

$$\mathrm{g}\left(\tilde{\boldsymbol{\alpha}}_0\right) := \sum_{i=1}^{n} \log\left(\mathbf{w}'(\boldsymbol{\alpha}_0)\boldsymbol{\varphi}(\mathrm{y}_i)\right)$$

as

$$\mathbf{s}_{\tilde{\alpha}_0}(\tilde{\boldsymbol{\alpha}}_0) = \sum_{i=1}^{n} \frac{\tilde{\mathbf{W}}'(\boldsymbol{\alpha}_0)\boldsymbol{\varphi}(y_i)}{\mathbf{w}'(\boldsymbol{\alpha}_0)\boldsymbol{\varphi}(y_i)}.$$

Using the Fisher information matrix as an approximation to the negative Hessian matrix, i. e. $\mathbb{E}(-\mathbf{H}_\theta(\boldsymbol{\theta})) = \mathbb{C}\mathrm{ov}(\mathbf{s}_\theta(\boldsymbol{\theta})) = \mathbb{E}(\mathbf{s}_\theta(\boldsymbol{\theta})\mathbf{s}'_\theta(\boldsymbol{\theta}))$, we obtain the Hessian matrix as

$$\mathbf{H}_{\tilde{\alpha}_0}(\tilde{\boldsymbol{\alpha}}_0) \approx -\sum_{i=1}^{n} \frac{\tilde{\mathbf{W}}'(\boldsymbol{\alpha}_0)\boldsymbol{\varphi}(y_i)\boldsymbol{\varphi}'(y_i)\tilde{\mathbf{W}}(\boldsymbol{\alpha}_0)}{\left(\mathbf{w}'(\boldsymbol{\alpha}_0)\boldsymbol{\varphi}(y_i)\right)^2}.$$

Taking into account the first and second order derivates of the penalty term, like in the previous subsection, the second order Taylor expansion of $f(\tilde{\boldsymbol{\alpha}}_0) = \log(p(\tilde{\boldsymbol{\alpha}}_0 \mid \cdot))$ with respect to the current state of the chain $\tilde{\boldsymbol{\alpha}}_0^{(c)}$ results in a multivariate Gaussian proposal distribution $\tilde{\boldsymbol{\alpha}}_0 \mid \tilde{\boldsymbol{\alpha}}_0^{(c)} \sim N\left(\hat{\boldsymbol{\mu}}_{\tilde{\alpha}_0}^{(c)}, \hat{\boldsymbol{\Sigma}}_{\tilde{\alpha}_0}^{(c)}\right)$ with mean vector $\hat{\boldsymbol{\mu}}_{\tilde{\alpha}_0}^{(c)} = \mathbb{E}(\tilde{\boldsymbol{\alpha}}_0 \mid \tilde{\boldsymbol{\alpha}}_0^{(c)})$ and covariance matrix $\hat{\boldsymbol{\Sigma}}_{\tilde{\alpha}_0}^{(c)} = \mathbb{C}\mathrm{ov}(\tilde{\boldsymbol{\alpha}}_0 \mid \tilde{\boldsymbol{\alpha}}_0^{(c)})$ given by

$$\begin{aligned}\hat{\boldsymbol{\mu}}_{\tilde{\alpha}_0}^{(c)} &= \hat{\boldsymbol{\Sigma}}_{\tilde{\alpha}_0}^{(c)} \left(\sum_{i=1}^{n} \frac{\tilde{\mathbf{W}}'\left(\boldsymbol{\alpha}_0^{(c)}\right)\boldsymbol{\varphi}(y_i)}{\mathbf{w}'(\boldsymbol{\alpha}_0^{(c)})\boldsymbol{\varphi}(y_i)} + \sum_{i=1}^{n} \frac{\tilde{\mathbf{W}}'\left(\boldsymbol{\alpha}^{(c)}\right)\boldsymbol{\varphi}(y_i)\boldsymbol{\varphi}(y_i)'\,\tilde{\mathbf{W}}\left(\boldsymbol{\alpha}^{(c)}\right)}{\left(\mathbf{w}'(\boldsymbol{\alpha}_0^{(c)})\boldsymbol{\varphi}(y_i)\right)^2} \tilde{\boldsymbol{\alpha}}_0^{(c)} \right), \\ \hat{\boldsymbol{\Sigma}}_{\tilde{\alpha}_0}^{(c)} &= \left(\sum_{i=1}^{n} \frac{\tilde{\mathbf{W}}'\left(\boldsymbol{\alpha}_0^{(c)}\right)\boldsymbol{\varphi}(y_i)\boldsymbol{\varphi}'(y_i)\tilde{\mathbf{W}}\left(\boldsymbol{\alpha}_0^{(c)}\right)}{\left(\mathbf{w}'(\boldsymbol{\alpha}_0^{(c)})\boldsymbol{\varphi}(y_i)\right)^2} + \frac{1}{\tau_{\alpha_0}^2}\tilde{\mathbf{K}}_d \right)^{-1},\end{aligned} \tag{6.20}$$

with $\boldsymbol{\alpha}_0^{(c)} = (\alpha_{0,1}^{(c)}, \ldots, \alpha_{0,k-1}^{(c)}, 0, \alpha_{0,k+1}^{(c)}, \ldots, \alpha_{0,g_0}^{(c)})'$. The proposed state $\tilde{\boldsymbol{\alpha}}_0^{(p)}$ is then accepted as new state of the chain with probability given in (6.19). We specify this method with `method.alpha="mhmarg"` in the function `baftpgm()`.

Update scheme "mcondblock" (Metropolis based block update)

To achieve higher acceptance rates, Brezger and Lang (2006) suggest using the posterior mode of the previous iteration $\hat{\boldsymbol{\mu}}_{\tilde{\alpha}_0}^{(c-1)}$ for computing the IWLS proposal. More precisely the mean vector and covariance matrix in (6.18) are evaluated by replacing the current state $\tilde{\boldsymbol{\alpha}}_0^{(c)}$ with $\hat{\boldsymbol{\mu}}_{\tilde{\alpha}_0}^{(c-1)}$, i. e.

$$\hat{\boldsymbol{\mu}}_{\alpha_0}^{(c)} = \hat{\boldsymbol{\Sigma}}_{\alpha_0}^{(c)}\left(\tilde{\mathbf{n}} - n\tilde{\mathbf{w}}\left(\hat{\boldsymbol{\mu}}_{\tilde{\boldsymbol{\alpha}}_0}^{(c-1)}\right) + n\tilde{\mathbf{W}}\left(\hat{\boldsymbol{\mu}}_{\tilde{\boldsymbol{\alpha}}_0}^{(c-1)}\right)\tilde{\boldsymbol{\alpha}}_0^{(c)}\right), \quad \hat{\boldsymbol{\Sigma}}_{\alpha_0}^{(c)} = \left(n\tilde{\mathbf{W}}\left(\hat{\boldsymbol{\mu}}_{\tilde{\boldsymbol{\alpha}}_0}^{(c-1)}\right) + \tau_{\alpha_0}^{-2}\tilde{\mathbf{K}}_0\right)^{-1}.$$

With this modification the proposal distribution becomes independent from the current state $\tilde{\boldsymbol{\alpha}}_0^{(c)}$ of the chain and we bypass the recomputation of the mean $\hat{\boldsymbol{\mu}}_{\tilde{\alpha}_0}^{(p)}$ and the covariance $\hat{\boldsymbol{\Sigma}}_{\tilde{\alpha}_0}^{(p)}$ to calculate the proposal density $\varphi(\tilde{\boldsymbol{\alpha}}_0^{(c)} \mid \hat{\boldsymbol{\mu}}_{\tilde{\alpha}_0}^{(p)}, \hat{\boldsymbol{\Sigma}}_{\tilde{\alpha}_0}^{(p)})$ at the current state $\tilde{\boldsymbol{\alpha}}_0^{(c)}$. This decreases the computational effort for the evaluation the acceptance probability (6.19) and increases the speed of the algorithm. If in addition the mean vector of the proposal is exchanged by the current state of the chain, i. e. $\hat{\boldsymbol{\mu}}_{\tilde{\alpha}_0}^{(c)} = \tilde{\boldsymbol{\alpha}}_0^{(c)}$, we do a simpler Metropolis update since the proposal becomes symmetric and the proposal ratio equals 1.

We have implemented a Metropolis update, where the mean vector and covariance matrix of the Gaussian proposal distribution are given by

$$\hat{\boldsymbol{\mu}}_{\tilde{\alpha}_0}^{(c)} = \tilde{\boldsymbol{\alpha}}_0^{(c)}, \quad \hat{\boldsymbol{\Sigma}}_{\tilde{\alpha}_0}^{(c)} = \hat{\boldsymbol{\Sigma}}_{\tilde{\alpha}_0}^{(c-1)} = \left(n\tilde{\mathbf{W}}\left(\hat{\boldsymbol{\mu}}_{\tilde{\boldsymbol{\alpha}}_0}^{(c-1)}\right) + \tau_{\alpha_0}^{-2}\tilde{\mathbf{K}}_0\right)^{-1}. \tag{6.21}$$

The update is done e. g. via the Cholesky decomposition of the covariance matrix $\hat{\boldsymbol{\Sigma}}_{\tilde{\alpha}_0}^{(c-1)} = \mathbf{L}\mathbf{L}'$. With the sample $\mathbf{x} = (x_1, \ldots, x_{g_0-1})' \sim N(\mathbf{0}, \mathbf{I})$ from a standard multivariate Gaussian distribution we compute the proposal via $\tilde{\boldsymbol{\alpha}}_0^{(p)} = \tilde{\boldsymbol{\alpha}}_0^{(c)} + \mathbf{L}\mathbf{x}$, compare computational detail 2 in Subsection 6.1.7, and accept this proposal with probability

$$\operatorname{accept}\left(\tilde{\boldsymbol{\alpha}}_0^{(p)},\tilde{\boldsymbol{\alpha}}_0^{(c)}\right)=\min\left\{1,\frac{p\left(\tilde{\boldsymbol{\alpha}}_0^{(p)}\mid\cdot\right)}{p\left(\tilde{\boldsymbol{\alpha}}_0^{(c)}\mid\cdot\right)}\right\}. \tag{6.22}$$

We specify this method with `method.alpha="mcondblock"` in the function `baftpgm()`.

Update scheme "mcondstep" (Metropolis based block update)

In a further version we use multiple Metropolis acceptance steps within the update of the transformed mixture weights. The mean vector and covariance matrix of the Gaussian proposal distribution are given by (6.21) and we use again the Cholesky decomposition of the covariance matrix $\hat{\boldsymbol{\Sigma}}_{\tilde{\alpha}_0}^{(c-1)}=\mathbf{L}\mathbf{L}'$. Let $\mathbf{x}=(x_1,...,x_{g_0-1})'\sim N(\mathbf{0},\mathbf{I})$ denote a sample from a multivariate standard Gaussian distribution and $\mathbf{e}_j$ denotes the ℓ-th unit vector of dimension (g_0-1). We can represent the sample $\mathbf{v}=\mathbf{L}\mathbf{x}$ from the multivariate Gaussian distribution $N(\mathbf{0},\hat{\boldsymbol{\Sigma}}_{\tilde{\alpha}_0}^{(c-1)})$ as sum of componentwise updates $\mathbf{L}\mathbf{x}=\sum_{\ell=1}^{g_0-1}\mathbf{L}\mathbf{e}_\ell x_\ell$, where the ℓ-th summand $\mathbf{L}\mathbf{e}_\ell x_\ell$ is the ℓ-th column of the lower triangular matrix $\mathbf{L}$ multiplied with x_ℓ. It is obvious, that the proposal $\tilde{\boldsymbol{\alpha}}_0^{(p)}=\tilde{\boldsymbol{\alpha}}_0^{(c)}+\mathbf{L}\mathbf{x}$ from the previous subsection is also obtained with the componentwise update, i. e. $\tilde{\boldsymbol{\alpha}}_0^{(p)}=\tilde{\boldsymbol{\alpha}}_0^{(c)}+\sum_{\ell=1}^{g_0-1}\mathbf{L}\mathbf{e}_\ell x_\ell$, and we can rewrite the update iteratively as $\tilde{\boldsymbol{\alpha}}_{0,\ell}^{(p)}=\tilde{\boldsymbol{\alpha}}_{0,\ell-1}^{(p)}+\mathbf{L}\mathbf{e}_\ell z_\ell$, $\ell=1,...,g_0-1$, starting with the current state $\tilde{\boldsymbol{\alpha}}_{0,0}^{(p)}:=\tilde{\boldsymbol{\alpha}}^{(c)}$. Based on this representation, the proposal arises as sequential modification of the $\ell+1,...,g_0-1$ components of the current value $\tilde{\boldsymbol{\alpha}}^{(c)}=\hat{\boldsymbol{\mu}}_{\tilde{\alpha}_0}^{(c)}$.

In the *"mcondstep"* update scheme we use this iterative construction of the proposal with an additional acceptance step after each iteration. In summary, the proposal is iteratively computed in (g_0-1)-steps via $\tilde{\boldsymbol{\alpha}}_{0,\ell}^{(p)}=\tilde{\boldsymbol{\alpha}}_{0,\ell-1}^{(p)}+\mathbf{L}\mathbf{e}_\ell z_\ell$, $\ell=1,...,g_0-1$, starting with the current state $\tilde{\boldsymbol{\alpha}}_{0,0}^{(p)}:=\tilde{\boldsymbol{\alpha}}^{(c)}$ and we accept the new components of the proposal in each iteration with probability

$$\operatorname{accept}\left(\tilde{\boldsymbol{\alpha}}_{0,\ell}^{(p)}\mid\tilde{\boldsymbol{\alpha}}_{0,\ell-1}^{(p)}\right)=\min\left\{1,\frac{p\left(\tilde{\boldsymbol{\alpha}}_{0,\ell}^{(p)}\mid\cdot\right)}{p\left(\tilde{\boldsymbol{\alpha}}_{0,\ell-1}^{(p)}\mid\cdot\right)}\right\},\quad \ell=1,...,g_0-1. \tag{6.23}$$

The *"mcondblock"* update scheme is obtained as special case of the *"mcondstep"* scheme if the acceptance probabilities all equal 1. We specify this method with `method.alpha="mcondstep"` in the function `baftpgm()`. In addition we have several options to vary the order of the update of the transformed mixture weights by specifying the argument `order.alpha`, compare Subsection 6.2.2.

Update scheme "slice" (single parameter update)

For a single update of the transformed weights $\alpha_{0,j}$, $j=1,...,g_0$, we require the conditional distribution $p(\alpha_{0,j}\mid\boldsymbol{\alpha}_{0,-j},\tau_{\alpha_j}^2)$ of the weight $\alpha_{0,j}$ given the remaining weights $\boldsymbol{\alpha}_{0,-j}:=(\alpha_1,...,\alpha_{j-1},\alpha_{j+1},...,\alpha_{g_0})'$. As pointed out in the Regularization Section 4.6 the conditional distribution $\alpha_{0,j}\mid\boldsymbol{\alpha}_{0,-j},\tau_{\alpha_0}^2$, depends only on the nearest neighbors and is Gaussian with mean and variance given as

$$\mathbb{E}\left(\alpha_{0,j}\mid\boldsymbol{\alpha}_{0,-j}\right)=-\frac{\sum_{k\neq j}\mathbf{K}_0[j,k]\alpha_{0,k}}{\mathbf{K}_0[j,j]},\quad \mathbb{V}\mathrm{ar}\left(\alpha_{0,j}\mid\boldsymbol{\alpha}_{0,-j}\right)=\frac{\tau_{\alpha_0}^2}{\mathbf{K}_0[j,j]}, \tag{6.24}$$

where $\mathbf{K}_0$ denotes the penalty matrix.

In summary, the full conditional distributions of the transformed weights $\alpha_{0,j}$ are proportional to the product of this Gaussian prior and the prior of the component labels (5.4)

$$p(\alpha_{0,j}\,|\,\boldsymbol{\alpha}_{0,-j},\cdot) \propto \frac{\exp\left(n_j\alpha_{0,j}\right)}{\left(\sum_{k=1}^{g_0}\exp\left(\alpha_{0,k}\right)\right)^n}\exp\left(-\frac{1}{2}\frac{\left(\alpha_{0,j}-\mathbb{E}\left(\alpha_{0,j}\,|\,\boldsymbol{\alpha}_{0,-j},\cdot\right)\right)^2}{\mathbb{V}\mathrm{ar}\left(\alpha_{0,j}\,|\,\boldsymbol{\alpha}_{0,-j},\cdot\right)}\right),\quad j=1,...,g_0. \tag{6.25}$$

Due to the identifiability constraint $\alpha_{0,k}=0$ for one $k\in\{1,...,g_0\}$, we update only the parameters $\alpha_{0,1},...,\alpha_{0,k-1},\alpha_{0,k+1},...,\alpha_{0,g}$ and keep $\alpha_{0,k}$ fixed at zero in each iteration. As for the update of the scale parameter (next subsection) we use the univariate slice sampling of Neal (2003). Since this density is log-concave also adaptive rejection sampling, Gilks and Wilde (1992), is a possible alternative. We specify this method with `method.alpha="slice"` in the function `baftpgm()`. In addition we have several options to vary the order of the update of the transformed mixture weights by specifying the argument `order.alpha`, compare Subsection 6.2.2.

Update scheme "dirichlet" (unregularized block update)

To compare the performance of smoothing the baseline error, also an unregularized approach for the weights $\mathbf{w}=(w_1,...,w_{g_0})'$ of the baseline error mixture distribution is considered. In contrast to the smoothing penalty a conjugate Dirichlet prior for the component weights is utilized, compare Frühwirth-Schnatter (2006), i. e.

$$\mathbf{w}\sim\mathrm{Dirichlet}\left(n_{01},...,n_{0g_0}\right),$$

with density

$$p(\mathbf{w})=\frac{\Gamma\left(\sum_{j=1}^{g_0}n_{0j}\right)}{\prod_{j=1}^{g_0}\Gamma\left(n_{0j}\right)}w_1^{n_{01}-1}\cdot...\cdot w_{g_0}^{n_{0g_0}-1}.$$

In this case the full conditional for the weights is

$$p(\mathbf{w}\,|\,\cdot)\propto p(\mathbf{r}\,|\,\boldsymbol{\theta})p(\mathbf{w})=\prod_{j=1}^{g_0}w_j^{n_j}\cdot\prod_{j=1}^{g_0}w_j^{n_{0j}-1}=\prod_{j=1}^{g_0}w_j^{n_j+n_{0j}-1},$$

which is also the density of a Dirichlet distribution:

$$\mathbf{w}\,|\,\cdot\sim\mathrm{Dirichlet}\left(n_1+n_{01},...,n_{g_0}+n_{0g_0}\right). \tag{6.26}$$

Since there is no smoothing in this case, we do not require an update of the parameter $\tau_{\alpha_0}^2$.

6.1.4. Full conditional of the scale parameter

With the multivariate conditional Gaussian prior of the log-survival times $p(\mathbf{y}\,|\,\boldsymbol{\mu}_y,\boldsymbol{\Sigma}_y)$ from (5.3) and the inverse gamma prior $p(\sigma^2)\propto(\sigma^2)^{-h_{1,\sigma}-1}\exp(-h_{2,\sigma}/\sigma^2)$ for the scale parameter σ^2 from (5.18) the full conditional is given as

$$p(\sigma^2\,|\,\cdot)\propto\left(\frac{1}{\sigma^2}\right)^{\frac{n}{2}+h_{\sigma,1}+1}\exp\left(-\frac{1}{2}(\mathbf{y}-\boldsymbol{\mu}_y)'\boldsymbol{\Sigma}_y^{-1}(\mathbf{y}-\boldsymbol{\mu}_y)-\frac{h_{\sigma,2}}{\sigma^2}\right).$$

To separate the dependence of scale parameter, we use the identity

$$\frac{1}{2}(\mathbf{y}-\boldsymbol{\mu}_y)'\boldsymbol{\Sigma}_y^{-1}(\mathbf{y}-\boldsymbol{\mu}_y)=\frac{1}{2\sigma^2}(\mathbf{y}-\boldsymbol{\eta})'\mathbf{S}_r^{-1}(\mathbf{y}-\boldsymbol{\eta})-\frac{1}{\sigma}(\mathbf{y}-\boldsymbol{\eta})'\mathbf{S}_r^{-1}\mathbf{m}_r+\frac{1}{2}\mathbf{m}_r'\mathbf{S}_r^{-1}\mathbf{m}_r$$

with $\mathbf{m}_r = (m_{r_1},...,m_{r_n})'$ and $\mathbf{S}_r = \mathrm{diag}(s^2_{r_1},...,s^2_{r_n})$. The last summand in the identity does not depend on the scale parameter and is omitted. Inserting this identity into the expression of the full conditional and using further the definitions

$$A_\sigma := \frac{n}{2} + h_{\sigma,1} + 1, \quad B_\sigma := \frac{1}{2}(\mathbf{y}-\boldsymbol{\eta})'\mathbf{S}_r^{-1}(\mathbf{y}-\boldsymbol{\eta}) + h_{\sigma,2}, \quad C_\sigma := (\mathbf{y}-\boldsymbol{\eta})'\mathbf{S}_r^{-1}\mathbf{m}_r \tag{6.27}$$

the full conditional is finally built as

$$p\left(\sigma^2 \mid \cdot\right) \propto \left(\frac{1}{\sigma^2}\right)^{A_\sigma} \exp\left(-\frac{1}{\sigma^2}B_\sigma + \frac{1}{\sigma}C_\sigma\right), \tag{6.28}$$

which is obviously not an inverse gamma density or of any standard form. It is shown in the Appendix A.2 that this density is unimodal and the univariate slice sampler, as described in Neal (2003) and implemented in the R-function `uni.slice{R}`, is used to update the scale parameter.

6.1.5. Full conditionals of latent component labels

Since the allocations to the mixture components are not known, we need to impute this latent component labels. The discrete allocation indicator R_i associates each observation with a certain component of the mixture distribution. The classification of the log-survival times y_i, $i = 1,...,n$, via the allocation variable $r_i \in \{1,...,g_0\}$ is obtained from the product of the Gaussian density of the log-survival times $p(y_i \mid \eta_i - \sigma m_{r_i}, \sigma^2 s^2_{r_i}) = \varphi\left(y_i \mid \eta_i - \sigma m_{r_i}, \sigma^2 s^2_{r_i}\right)$, (5.3), and the multinomial prior of the component labels $p\left(r_i \mid \boldsymbol{\alpha}_0\right) = w_{r_i}(\boldsymbol{\alpha}_0)$, (5.4), i. e.

$$p\left(r_i \mid \cdot\right) \propto \varphi\left(y_i \mid \eta_i - \sigma m_{r_i}, \sigma^2 s^2_{r_i}\right) w_{r_i}(\boldsymbol{\alpha}_0).$$

Thus the full conditional of each allocation variable r_i, $i = 1,...,n$, is discrete with the normalized probability

$$p_{ij} := P\left(r_i = j \mid \cdot\right) = \frac{w_j(\boldsymbol{\alpha}_0)\varphi\left(y_i \mid \eta_i - \sigma m_j, \sigma^2 s^2_j\right)}{\sum_{k=1}^{g_0} w_k(\boldsymbol{\alpha}_0)\varphi\left(y_i \mid \eta_i - \sigma m_k, \sigma^2 s^2_k\right)}, \quad j = 1,...,g_0, \tag{6.29}$$

which is the special case of a multinomial distribution $r_i \sim \mathrm{MNom}(1, p_{i1},...,p_{ig_0})$.

6.1.6. Full conditional of the censored log-survival times

As shown in the previous sections, the vector of exact (log-) survival times $\mathbf{y}$ is required to update the remaining model parameters. The exact survival times are only partially known, in particular for the uncensored individuals. The survival times of the right censored individuals with censoring time $\tilde{y}_i$ have to be imputed in each update step. Using the likelihood contribution for a censored observation $L(\mathfrak{D}_i \mid y_i) = 1_{[\tilde{y}_i,\infty)}(y_i)^{1-d_i}$, $d_i = 1$, and the associated prior component of the exact survival time $p(y_i \mid r_i, \boldsymbol{\alpha}_1,...,\boldsymbol{\alpha}_{p_z}, \boldsymbol{\beta}, \boldsymbol{\gamma}, \sigma)$ from (5.3), the full conditional for a censored log-survival time is given as

$$p\left(y_i \mid \cdot\right) \propto 1_{[\tilde{y}_i,\infty)}\left(y_i\right)\frac{1}{\sigma s_{r_i}}\exp\left(-\frac{1}{2\sigma^2 s^2_{r_i}}\left(y_i - \eta_i - \sigma m_{r_i}\right)^2\right),$$

which is the density of a truncated Gaussian distribution with location parameter $\eta_i + \sigma m_{r_i}$, squared scale parameter $\sigma^2 s^2_{r_i}$ and support $[\tilde{y}_i, \infty)$

$$p(y_i \mid \cdot) \sim \mathrm{TN}_{[\tilde{y}_i,\infty)}\left(\eta_i + \sigma m_{r_i}, \sigma^2 s^2_{r_i}\right). \tag{6.30}$$

Sampling from a truncated normal distribution is described in Robert (1995) and practiced using the R-function `rtnorm{msm}`.

6.1.7. Computational details

Detail 1: The generic way to update the predictor components is described in terms of the unregularized linear effects $\boldsymbol{\gamma}$, where we assume $\boldsymbol{\mu}_\gamma \neq 0$ and $\boldsymbol{\Pi}_\gamma = \boldsymbol{\Sigma}_\gamma^{-1} \neq 0$ for the moment to keep the generality of the derivation. To draw the random samples efficiently from a possibly high-dimensional multivariate Gaussian distribution $\boldsymbol{\gamma} \mid \cdot \sim \mathrm{N}(\boldsymbol{\mu}_{\gamma|\cdot}, \boldsymbol{\Pi}_{\gamma|\cdot}^{-1})$ with precision matrix $\boldsymbol{\Pi}_{\gamma|\cdot}$, we follow Rue (2001) and start with computing the Cholesky decomposition of the precision matrix $\boldsymbol{\Pi}_{\gamma|\cdot} = (\mathbf{U}'\boldsymbol{\Sigma}_y^{-1}\mathbf{U} + \boldsymbol{\Sigma}_\gamma^{-1}) = \mathbf{L}\mathbf{L}'$ such that $\mathbf{L}$ denotes the corresponding lower triangular matrix. With the p_u-dimensional sample $\mathbf{x} = (x_1, \ldots, x_{p_u})'$ from a standard Gaussian distribution $x_i \sim \mathrm{N}(0,1)$ we solve the equation $\mathbf{L}'\boldsymbol{\gamma}^* = \mathbf{x}$ via backward substitution to get a sample from $\boldsymbol{\gamma}^* \mid \cdot \sim \mathrm{N}(\mathbf{0}, \boldsymbol{\Pi}_{\gamma|\cdot}^{-1})$. Finally, the sum $\boldsymbol{\mu}_{\gamma|\cdot} + \boldsymbol{\gamma}^*$ is the desired sample of $\mathrm{N}(\boldsymbol{\mu}_{\gamma|\cdot}, \boldsymbol{\Pi}_{\gamma|\cdot}^{-1})$. We can also compute the mean $\boldsymbol{\mu}_{\gamma|\cdot}$ in terms of the Cholesky decomposition. Via the connection $\boldsymbol{\Pi}_{\gamma|\cdot}\boldsymbol{\mu}_{\gamma|\cdot} = \mathbf{L}\mathbf{L}'\boldsymbol{\mu}_{\gamma|\cdot} = (\mathbf{U}'\boldsymbol{\Sigma}_y^{-1}\tilde{\mathbf{y}}_\gamma + \boldsymbol{\Sigma}_\gamma^{-1}\boldsymbol{\mu}_\gamma)$, we solve at first the equation $\mathbf{L}\mathbf{v} = (\mathbf{U}'\boldsymbol{\Sigma}_y^{-1}\tilde{\mathbf{y}}_\gamma + \boldsymbol{\Sigma}_\gamma^{-1}\boldsymbol{\mu}_\gamma)$ via forward substitution and then the equation $\mathbf{L}'\boldsymbol{\mu}_{\gamma|\cdot} = \mathbf{v}$ via backward substitution. Since the precision matrix of the nonlinear terms has band structure, the Cholesky decomposition can be computed by sparse matrix operations.

Detail 2: An alternative is to use the methods implemented in the R function `rmvnorm{mvtnorm}` to draw in terms of the covariance matrix $\boldsymbol{\Sigma}_{\gamma|\cdot}$, instead of the precision matrix, a new state from a multivariate Gaussian distribution $\boldsymbol{\gamma} \mid \cdot \sim \mathrm{N}(\boldsymbol{\mu}_{\gamma|\cdot}, \boldsymbol{\Sigma}_{\gamma|\cdot})$. The procedure based on the e Cholesky decomposition of the covariance matrix uses the steps described in Rue (2001), i. e. with the decomposition $\boldsymbol{\Sigma}_{\gamma|\cdot} = \mathbf{L}\mathbf{L}'$ and the p_u-dimensional sample $\mathbf{x} = (x_1, \ldots, x_{p_u})'$, $x_i \sim \mathrm{N}(0,1)$, the vector $\mathbf{L}\mathbf{x} = \boldsymbol{\gamma}^*$ is computed to obtain the sample from $\boldsymbol{\gamma}^* \mid \cdot \sim \mathrm{N}(\mathbf{0}, \boldsymbol{\Sigma}_{\gamma|\cdot})$. Finally, the mean vector is added and the sum $\boldsymbol{\mu}_{\gamma|\cdot} + \boldsymbol{\gamma}^*$ is the desired sample of $\mathrm{N}(\boldsymbol{\mu}_{\gamma|\cdot}, \boldsymbol{\Sigma}_{\gamma|\cdot})$.

Detail 3: For large parameter vectors we can partition the vector of regression coefficients randomly in blocks of fix size and update sequentially the blocks until each coefficient is updated instead of updating the whole coefficient vector at once. The random sample is then generated from a multivariate Gaussian distribution with the corresponding subvector of the mean and the submatrix of the covariance matrix conditional on the remainder of the regression coefficients and other parameters (argument `blocksize` in the function `baftpgm()`).

6.2. Algorithmic variants

6.2.1. Standardization of the baseline error distribution

As outlined in Subsection 2.3 we require constraints on the transformed weights $\boldsymbol{\alpha}_0$ to achieve a standardized baseline error distribution to enforce the interpretation of the predictor η_i as the mean $\mathbb{E}(Y_i \mid \boldsymbol{\theta}) = \mu_{Y_i}$ and the scale parameter σ^2 as the variance $\mathbb{V}\mathrm{ar}(Y_i \mid \boldsymbol{\theta}) = \sigma^2_{Y_i}$ of the conditional distribution $Y_i \mid \boldsymbol{\theta}$, compare (2.12) and (2.14). In this case the trace plots of the samples of the global intercept parameter γ_0, the scale parameter σ^2 and the samples of the mixture weights $\mathbf{w}(\boldsymbol{\alpha}_0)$ of the corresponding baseline error distribution $Y_0 \mid \gamma_0, \sigma$ indicate the desired convergence. With an unconstrained estimation of the transformed mixture weights $\boldsymbol{\alpha}_0$ the mean $\mu_\varepsilon = \sum_{k=1}^{g_0} w_k(\boldsymbol{\alpha}_0) m_k$ and the variance $\sigma_\varepsilon^2 = \sum_{k=1}^{g_0} w_k(\boldsymbol{\alpha}_0)(m_k^2 + s_k^2) - \mu_\varepsilon^2$, of the error distribution $\varepsilon \,|\, \boldsymbol{\alpha}_0$, (2.8), can take any values $\mu_\varepsilon \in \mathbb{R}$ and $\sigma_\varepsilon^2 > 0$. The corresponding mean and variance of the baseline error distribution $Y_0 \,|\, \gamma_0, \sigma$ are given by $\mu_{Y_0} = \gamma_0 + \sigma\mu_\varepsilon$ and $\sigma_{Y_0}^2 = \sigma^2\sigma_\varepsilon^2$, compare (2.13), and the interpretation of the parameters γ_0 and σ^2 as global intercept and variance of the conditional distribution $Y_i \,|\, \boldsymbol{\theta}$ is not feasible. As a further consequence of the unconstrained update of $\boldsymbol{\alpha}_0$, the parameters γ_0 and μ_ε of the mean $\mu_{Y_0} = \gamma_0 + \sigma\mu_\varepsilon$ as well as the parameters σ^2 and σ_ε^2 of the variance $\sigma_{Y_0}^2 = \sigma^2\sigma_\varepsilon^2$ of the conditional distribution $Y_0 \,|\, \gamma_0, \sigma$ are not identifiable. This is due to the fact, that for any scalar $c \in \mathbb{R}$, $\tilde{\gamma}_0 = \gamma_0 + \sigma c$ in combination with $\tilde{\mu}_\varepsilon = \mu_\varepsilon - c$, and $\tilde{\sigma}^2 = c^2\sigma^2$ in combination with $\tilde{\sigma}_\varepsilon^2 = c^{-2}\sigma_\varepsilon^2$, does not change the mean μ_{Y_0} and the variance $\sigma_{Y_0}^2$. It follows that only the trace plots of the mean μ_{Y_0} and variance $\sigma_{Y_0}^2$ of the baseline error distribution show the desired stationarity, but stationarity is not indicated by the trace plots of the single components of these expressions and the mixture weights $\mathbf{w}(\boldsymbol{\alpha}_0)$.

The unconstrained estimation requires also an adjustment of the hyperparameters of the intercept γ_0 and the scale parameter σ in the algorithm, if informative instead of diffuse priors are used for these parameters. Consider as an example the scale parameter σ^2. In the case of a standardized error distribution $\varepsilon \sim (0,1)$, our prior knowledge of the baseline variance $\mathbb{V}\text{ar}(Y_0 \,|\, \gamma_0, \sigma)$ of the conditional distribution $Y_0 \,|\, \gamma_0, \sigma$ is reflected by the hyperparameters $h_{1,\sigma}$ and $h_{2,\sigma}$ of an inverse gamma distribution, compare (5.18). For an unstandardized baseline error distribution our knowledge concerns the product of the scale parameter and the variance of the baseline error distribution, i. e. $\mathbb{V}\text{ar}(Y_0 \,|\, \gamma_0, \sigma) = \sigma^2\sigma_\varepsilon^2 \sim \text{InvGamma}(h_{1,\sigma}, h_{2,\sigma})$. Due to the identification problem of the variance, the single components σ_ε^2 and σ^2 can take any positive value in each iteration and we have to adjust the hyperparameters of the prior for the scale parameter σ^2 with respect to the values of σ_ε^2 in the iterations according to $\sigma^2 \sim \text{InvGamma}(h_{1,\sigma}, h_{2,\sigma}/\sigma_\varepsilon^2)$. The same argumentation holds for the intercept and is leading to the adjustment $\gamma_0 \sim N(h_{1,\gamma_0} - \sigma\mu_\varepsilon, h_{2,\gamma_0}^2)$ in every iteration of the sampler.

We advocate the following strategy to obtain a standardized baseline error distribution, which avoids the direct implementation of constraints in the update of the (transformed) mixture weights. Let $\mu_\varepsilon = \sum_{k=1}^{g_0} w_k(\boldsymbol{\alpha}_0) m_k$ and $\sigma_\varepsilon^2 = \sum_{k=1}^{g_0} w_k(\boldsymbol{\alpha}_0)(m_k^2 + s_k^2) - \mu_\varepsilon^2$ denote the mean and the variance of the unstandardized error distribution with density $f_\varepsilon(\cdot \,|\, \boldsymbol{\alpha}_0) = \sum_{k=1}^{g_0} w_k(\boldsymbol{\alpha}_0)\varphi(\cdot \,|\, m_k, s_k^2)$ from (2.7). The density of the standardized baseline error distribution $\tilde{\varepsilon}_0 \,|\, \boldsymbol{\alpha}_0$ is obtained via the transformation $\tilde{\varepsilon}_0 = (\varepsilon - \mu_\varepsilon)/\sigma_\varepsilon$ as $f_{\tilde{\varepsilon}_0}(\cdot \,|\, \boldsymbol{\alpha}_0) = \sum_{k=1}^{g_0} w_k(\boldsymbol{\alpha}_0)\varphi(\cdot \,|\, \tilde{m}_{0,k}, \tilde{s}_{0,k}^2)$ with basis knots $\tilde{m}_{0,k}$ and basis variances $\tilde{s}_{0,k}^2$ given by

$$\tilde{m}_{0,k} = \frac{m_k - \mu_\varepsilon}{\sigma_\varepsilon}, \quad \tilde{s}_{0,k}^2 = \frac{s_k^2}{\sigma_\varepsilon^2}, \quad k = 1, \ldots, g_0. \tag{6.31}$$

This transformation shifts and scales only the basis knots and variances to match the zero mean and unit variance condition, but leaves the mixture weights unchanged. To avoid a change of the posterior we have to ensure, that the mean $\mu_{Y_0} = \gamma_0 + \sigma\mu_\varepsilon$ and the variance $\sigma_{Y_0}^2 = \sigma^2\sigma_\varepsilon^2$ of the baseline error distribution $Y_0 \,|\, \gamma_0, \sigma$, now expressed in terms of the standardized error distribution $\tilde{\varepsilon}_0 \,|\, \tilde{\boldsymbol{\alpha}}_0$, does not change. With the relationship $Y_0 = \tilde{\gamma}_0 + \tilde{\sigma}\tilde{\varepsilon}_0 = \gamma_0 + \sigma\mu_\varepsilon + \sigma\sigma_\varepsilon\tilde{\varepsilon}_0$ follows that we have to adjust the intercept and the scale parameter according to

$$\tilde{\gamma}_0 = \gamma_0 + \sigma\mu_\varepsilon, \quad \tilde{\sigma}^2 = \sigma^2\sigma_\varepsilon^2. \tag{6.32}$$

To reformulate the standardized baseline error density in terms of the initial basis knots m_k and basis variances s_k^2, we have to recompute the weights at the basis knots m_k by solving the linear equation system

$$\sum_{k=1}^{g_0} \tilde{w}_k \varphi(m_\ell \mid m_k, s_k^2) = f_{\tilde{\varepsilon}_0}(m_\ell \mid \boldsymbol{\alpha}_0), \quad \ell = 1,...,g_0, \tag{6.33}$$

with respect to the constraints $\sum_{k=1}^{g_0} \tilde{w}_k = 1$ and $\tilde{w}_k > 0$. The standardized baseline error density is then given by $f_{\tilde{\varepsilon}_0}(\cdot \mid \tilde{\boldsymbol{\alpha}}_0) = \sum_{k=1}^{g_0} \tilde{w}_k(\tilde{\boldsymbol{\alpha}}_0)\varphi(\cdot \mid m_k, s_k^2)$, where $\tilde{\boldsymbol{\alpha}}_0$ denotes the corresponding transformed mixture weights.

In the simulations we try two alternatives. We run the MCMC iterations without the standardization of the baseline error density within the algorithm described in Section 5.2.3 and standardize the error density in post-processing steps with the listed corrections (6.32) applied to the samples of the involved parameters. Alternatively, we standardize the baseline error density according to (6.32) within the sampling algorithm, after every update of the mixture weights, but we leave the weights unchanged to avoid in the next iteration sampling from possibly unfavorable conditional posterior regions arising probably from the recomputation of the mixture weights according to (6.33). In this version the knot positions and basis variances change in each iteration of the sampler and need to be stored in addition to the samples of the parameters. The option for a within-standardization is selected in the function `baftpgm()` with the argument `scalebasis=TRUE`. The weights are optionally recomputed after the simulation to show the convergence also in terms of the mixture weights.

6.2.2. Varying the update order of the transformed mixture weights

If we specify the methods `method.alpha="mcondstep"` or `method.alpha="slice"` in the function `baftpgm()`, we have the following options to vary the order of the update of the transformed mixture weight in every loop of the sampler (we assume e. g. $\alpha_{g_0} := 0$ for identifiability):

- `order.alpha="fix1"`. The order of the indices is fixed to $(1,2,3,....,g_0-1)$.
- `order.alpha="fix2"`. The fixed order of the coefficients is determined in the way, that the coefficient j, which is just updated, does not depend on the coefficients used for the update of the previous coefficient $j-1$. If d_0 denotes the used difference order, the update order is $j \to j+(d_0+1) \to j+2(d_0+1) \to, \dots, j = 1,...,d_0+1$. This is the default setting.
- `order.alpha="random1"`. In each update step a random permutation of the indices $(1,2,3,....,g_0-1)$ is used.
- `order.alpha ="random2"`. In the order of the option `order.alpha="fix2"` in each step one random cut is used to exchange the update order.

6.2.3. Varying the update of the component labels

We have several alternatives to classify the observations to the mixture components.

- `method.rlabel="gibbs"`: Random assignment by sampling the labels with p_{ij} from (6.29), which is the default option.
- `method.rlabel="fix-maxprob"`: Hard assignment to the class with maximum probability $p_{i,\max} = \max\{p(r_i = k) : k = 1,...,g_0\}$.

- `method.rlabel="fix-interval"`: Hard assignment to intervals I_j around the knots that build a partition of the domain of the error distribution. E. g. for equidistant knots and homoscedastic basis variances these intervals are $I_j = (0.5(m_{j-1} + m_j), 0.5(m_j + m_{j+1})]$, $j = 1,...,g_0$, and $m_{-1} := -\infty, m_{g_0+1} := \infty$.

6.2.4. Scale dependent implementation

In addition a variant with scale-dependent covariance matrices in the priors of the regularized predictor components (5.6) and (5.16) is implemented, i. e. , $\mathbf{\Sigma}_\beta = \sigma^2 \mathbf{D}_{\tau_\beta}$ and $\mathbf{\Sigma}_{\alpha_j}^- = \sigma^2 \tau_{\alpha_j}^2 \mathbf{K}_j^-$. With this parametrization the values of the scale parameter strengthen or relax the regularization. The derivation of the associated full conditionals is straightforward. We can select this option with the argument `scaledpri=TRUE`.

6.3. Update of the parameters

The Markov chain is generated via MCMC simulations based on drawing from the full conditionals of parameters or parameter blocks given the remaining parameters and the data as derived in the previous sections. The methods are implemented in the `R`-function `baftpgm()` which will be provided from the author on request. The usage of the function is described in the Appendix D.5.

6.3.1. Preprocessing

Standardization: To ensure that comparable regression coefficient sizes imply comparable effect sizes, the covariates are standardized in advance. This avoids the extensive covariate-specific tuning of the priors for different covariate scales. We standardize covariates with linear effects to zero empirical mean and unit empirical variance. To obtain that smooth covariates taking values in $[-1,1]$, we can apply the transformation

$$z_{ij}^* = \frac{2(z_{ij} - z_{j,\min})}{z_{j,\max} - z_{j,\min}} - 1 .$$

Starting values: In general we avoid preprocessing steps to fit the model in order to obtain suitable starting values. An automatic computation of starting values is not implemented in the function `baftpgm()` and in our simulations and applications we start with weakly specified models. The accurately starting values and prior specifications are given corresponding sections. If desired, starting values can e. g. be computed by a view iterations with the `R`-function `bayessurvreg2{bayesSurv}`.

6.3.2. Pseudocode

[1] *Initialization*:

Specify the PGM: Set number g_0 of Gaussian basis functions and choose the location of the means m_j, the scales s_j^2 and the order d_0 of the random walk prior for the (transformed) mixture weights. Select the hyperparameters $h_{1,\sigma}, h_{2,\sigma}$ to specify the inverse gamma prior for the scale parameter σ.

Specify the regularization priors of the linear effects: Set the values of the hyperparameters $h_{1,\lambda}, h_{2,\lambda}$ to specify the gamma prior for the shrinkage parameter $\lambda(\lambda^2)$ in the Bayesian ridge or lasso prior. For the Bayesian NMIG prior set the values v_0, v_1 of the indicator I_j, set the values of the hyperparameters $h_{1,\psi}, h_{2,\psi}$ of the inverse gamma prior for the variance parameter ψ_j^2 and set the hyperparameters $h_{1,\omega}, h_{2,\omega}$ of the beta prior for the complexity parameter ω.

Specify the non-linear effects: Set number g_j of B-spline basis functions and choose the order d_j of the random walk prior for basis function weights.

Select optionally variants described in Subsection 6.2.

Standardize the covariates according to Section 6.3.1 and choose appropriate starting values for the parameters $\boldsymbol{\theta} = (\boldsymbol{\alpha}', \boldsymbol{\beta}', \boldsymbol{\gamma}', \boldsymbol{\tau}_\alpha^{2\prime}, \boldsymbol{\tau}_\beta^{2\prime}, \boldsymbol{\rho}', \sigma)'$.

Set the number C of iterations, set $c = 0$ and repeat the following steps until $c < C$.

[2] *Update of the unregularized linear regression coefficients*:

Draw a new value $\boldsymbol{\gamma}^{(c+1)}$ from a multivariate Gaussian full conditional with mean vector and covariance matrix given in (6.3).

[3] *Update of the regularized linear regression coefficients*:

Draw a new value $\boldsymbol{\beta}^{(c+1)}$ from a multivariate Gaussian full conditional with mean vector and covariance matrix given in (6.4).

[4] *Update of the shrinkage- and selection-prior components:*

Bayesian ridge (A): Draw a new value of the complexity parameter $\lambda_j^{(c+1)}$ from the conditional gamma distribution given by (6.7) and set the variance parameter $\tau_{\beta_j}^{2,(c+1)} = 1/2\lambda_j^{(c+1)}$, $j = 1,...,p_x$.

Bayesian ridge (B): Draw a new value of the complexity parameter $\lambda^{(c+1)}$ from the conditional gamma distribution given by (6.8) and set the variance parameter $\tau_\beta^{2,(c+1)} = 1/2\lambda^{(c+1)}$.

Bayesian lasso: Draw a new value of the variance parameter $\tau_{\beta_j}^{2,(c+1)}$, $j = 1,...,p_x$, from the conditional inverse Gaussian distribution given by (6.9). Draw a new value of the complexity parameter $\lambda^{2,(c+1)}$ from the conditional gamma distribution given by (6.10).

Bayesian NMIG: Draw a new value of the indicator $I_{\beta_j}^{(c+1)}$, $j = 1,...,p_x$, from the conditional Bernoulli distribution given in (6.11). Draw a new value of the variance parameter $\psi_{\beta_j}^{2,(c+1)}$, $j = 1,...,p_x$, from the conditional inverse gamma distribution given in (6.12). Draw a new value of the complexity parameter $\omega^{(c+1)}$ from the conditional beta distribution given in (6.13).

[5] *Update of the regularized spline coefficients of the nonlinear effects*:

Draw a new value $\boldsymbol{\alpha}_j^{(c+1)}$, $j = 1,...,p_z$, from a multivariate Gaussian full conditional with mean vector and covariance matrix given by (6.5).

To center the functions, compute the mean of the function evaluations at the observed data points $c_j^{(c+1)} = n^{-1}\sum_{i=1}^n \alpha_{j,k}^{(c+1)} B_{j,k}(z_{ij})$.

Adjust the current states of $\boldsymbol{\alpha}_j^{(c+1)}$ by $\boldsymbol{\alpha}_j^{(c+1)} - c_j^{(c+1)}$, $j = 1,...,p_z$, and adjust the intercept $\gamma^{(c+1)}$ by $\gamma^{(c+1)} + c_1^{(c+1)} + ... + c_{p_z}^{(c+1)}$.

[6] *Update of the smoothing variances associated to the spline coefficients*:

Draw a new value of the variance parameters $\tau_{\alpha_j}^{2,(c+1)}$, $j = 1,...,p_z$, from the conditional inverse gamma distribution given by (6.14).

[7] *Update of the scale parameter*:

Draw with slice sampling a new value of the variance parameters $\sigma^{2,(c+1)}$ from the conditional inverse gamma distribution given by (6.28) with (6.27).

[8] *Update of the transformed mixture weights*:

Option 1: (Method "`mhcond`") Draw a new value $\tilde{\boldsymbol{\alpha}}_0^{(p)} = (\alpha_{0,1}^{(p)},\ldots,\alpha_{0,k-1}^{(p)},\alpha_{0,k+1}^{(p)},\ldots,\alpha_{0,g}^{(p)})'$ from a multivariate Gaussian proposal distribution with mean vector and covariance matrix given by (6.18). Accept the proposed state as new state of the chain with the acceptance probability given in (6.19). If the proposal is accepted, set $\boldsymbol{\alpha}_0^{(c+1)} = \boldsymbol{\alpha}_0^{(p)}$, else set $\boldsymbol{\alpha}_0^{(c+1)} = \boldsymbol{\alpha}_0^{(c)}$.

Option 2: (Method "`mhmarg`") Draw a new value $\tilde{\boldsymbol{\alpha}}_0^{(p)} = (\alpha_{0,1}^{(p)},\ldots,\alpha_{0,k-1}^{(p)},\alpha_{0,k+1}^{(p)},\ldots,\alpha_{0,g}^{(p)})'$ from a multivariate Gaussian proposal distribution with mean vector and covariance matrix given by (6.20). Accept the proposed state as new state of the chain with the acceptance probability given in (6.19). If the proposal is accepted, set $\boldsymbol{\alpha}_0^{(c+1)} = \boldsymbol{\alpha}_0^{(p)}$, else set $\boldsymbol{\alpha}_0^{(c+1)} = \boldsymbol{\alpha}_0^{(c)}$.

Option 3: (Method "`mcondblock`") Draw a new value $\tilde{\boldsymbol{\alpha}}_0^{(p)} = (\alpha_{0,1}^{(p)},\ldots,\alpha_{0,k-1}^{(p)},\alpha_{0,k+1}^{(p)},\ldots,\alpha_{0,g}^{(p)})'$ from the Gaussian proposal distribution with mean vector and covariance matrix given by (6.21) and accept the proposed state as new state of the chain with the acceptance probability given in (6.22). If the proposal is accepted, set $\boldsymbol{\alpha}_0^{(c+1)} = \boldsymbol{\alpha}_0^{(p)}$, else set $\boldsymbol{\alpha}_0^{(c+1)} = \boldsymbol{\alpha}_0^{(c)}$.

Option 4: (Method "`mcondstep`") Draw a new value $\tilde{\boldsymbol{\alpha}}_{0,\ell}^{(p)}$ from the multivariate Gaussian proposal distribution with mean vector and covariance matrix given by (6.21) and the stepwise update. Accept in each step $\ell = 1,\ldots,g_0 - 1$, the proposed state as new state of the chain with the acceptance probability given in (6.23). If the proposal is accepted, set $\boldsymbol{\alpha}_0^{(c+1)} = \boldsymbol{\alpha}_{0,\ell}^{(p)}$, else set $\boldsymbol{\alpha}_0^{(c+1)} = \boldsymbol{\alpha}_{0,\ell-1}^{(c)}$.

Option 5: (Method "`slice`") Draw a new value with slice sampling to update each component $\alpha_{0,j}^{(c+1)}$, $j = 1,\ldots,k-1,k+1,\ldots,g_0$, from the conditional distribution given by (6.25).

Option 6: (Method "`dirichlet`") Draw a new value state of the component weights $\mathbf{w}^{(c)}$ from the Dirichlet distribution (6.26).

Skip the update of the smoothing variance parameter in [9].

Option 7: (`scalebasis=TRUE`) To standardize the error distribution to zero mean and unit variance, compute the mean and variance of the error distribution at the current values of the basis knots $m_j^{(c)}$ and basis variances $s_j^{2,(c)}$

$$\mu_\varepsilon^{(c+1)} = \sum_{j=1}^{g_0} w_j(\boldsymbol{\alpha}_0^{(c+1)}) m_j^{(c)}, \quad \sigma_\varepsilon^{2,(c+1)} = \sum_{j=1}^{g_0} w_j(\boldsymbol{\alpha}_0^{(c+1)})\left(m_j^{2,(c)} + s_j^{2,(c)}\right) - \mu_\varepsilon^{2,(c+1)}.$$

Update the basis means and variances according to (6.31) with

$$m_j^{(c+1)} = \frac{m_j^{(c)} - \mu_\varepsilon^{(c+1)}}{\sigma_\varepsilon^{(c+1)}}, \quad s_j^{2,(c+1)} = \frac{s_j^{2,(c)}}{\sigma_\varepsilon^{2,(c+1)}}, \; j = 1,\ldots,g_0.$$

Finally, adjust the intercept $\gamma_0^{(c+1)} \to \gamma_0^{(c+1)} + \sigma^{2,(c+1)}\mu_\varepsilon^{(c+1)}$ and the scale parameter $\sigma_\varepsilon^{2,(c+1)} \to \sigma^{2,(c+1)}\sigma_\varepsilon^{2,(c+1)}$ according to (6.32).

[9] *Update of the smoothing variance of the mixture weights*:

Draw a new value of the variance parameter $\tau_{\alpha_0}^{2,(c+1)}$ from the conditional distribution given by (6.14).

[10] *Update of the latent component labels*:

Draw a new value of the latent component labels $r_i^{(c+1)} \in \{1,...,g_0\}$, $i=1,...,n$, from the conditional multinomial distribution with probabilities given by (6.29). Update the number of observations in the mixture components $n_j^{(c+1)} = \#\{i : r_i^{(c+1)} = j\}$, $j=1,...,g_0$.

[11] *Update of the latent exact log-survival times*:

For the censored observations $i \in \{1,...,n\}$ draw a new value of the latent log-survival time $y_i^{(c+1)}$ from the truncated Gaussian distribution given by (6.30).

6.3.3. Postprocessing

Standardization of the error: As outlined in Subsection 6.2.1 the unconstrained estimation of the baseline error is leading to the non-identifiability of the parameters defining the baseline error distribution $Y_0 = \gamma_0 + \sigma\varepsilon$ and the involved parameters do not show the desired convergence. When the basis function means m_k and standard deviations s_k are fixed during the sampler (`scalebasis=FALSE`), we have to compute the standardized error density in post-processing steps utilizing the obtained MCMC sample. We do this by applying the same transformation of the mean m_k and standard deviation s_k as described in Option 7 to the sampled values of the location and scale parameter. The resulting, adjusted sample values of the intercept γ_0 and the scale parameter σ show the desired convergence.

Recomputation of the weights: With the procedure described in Option 7 we achieve a standardized error distribution, $\varepsilon_0 \sim \text{PGM}(0,1)$, but the resulting variability in the locations $\tilde{m}_k$ and scales $\tilde{s}_k$ of the Gaussian basis function sometimes prevents the direct detection of the convergence of basis function weights. Finally, to show in addition the convergence of the weights, we compute the estimated standardized version of error density $f_{\varepsilon_0}(\cdot)$ at a fixed number of grid points. For the standardized densities we can e. g. use the starting knots $\mathbf{m} = (m_1,...,m_{g_0})'$ of the Gaussian basis densities, but any set of grid points is possible. With respect to (6.33), we have to solve the constrained linear equation system $\mathbf{B}^{(s)}\mathbf{w}^{(s)} = \mathbf{f}_{\varepsilon_0}^{(s)}$ subject to $\tilde{w}_k^{(s)} > 0$ and $\sum_{k=1}^{g_0} \tilde{w}_k^{(s)} = 1$, where $\mathbf{B}^{(s)} = (\varphi(m_\ell \mid \tilde{m}_k^{(s)}, \tilde{s}_k^{(s)2}))_{1 \le i,j \le g_0}$ denotes the matrix of the Gaussian basis functions $\varphi(\cdot \mid \tilde{m}_k^{(s)}, \tilde{s}_k^{(s)2})$ with adjusted mean $\tilde{m}_k^{(s)}$ and standard deviation $\tilde{s}_k^{(s)}$, $k=1,....,g_0$, evaluated at each grid point m_ℓ, $\ell=1,...,g_0$. Further $\tilde{\mathbf{w}}^{(s)} = (\tilde{w}_1^{(s)},...,\tilde{w}_{g_0}^{(s)})'$ is the vector of recomputed basis function weights and $\mathbf{f}_{\varepsilon_0}^{(s)} = (f_{\varepsilon_0}^{(s)}(e_1),...,f_{\varepsilon_0}^{(s)}(e_{g_0}))'$ is the vector of the standardized error density computed with the parameter values of the s-th iteration. Since some of the border weights are often very close to zero, solving the constrained equation system becomes often numerically instable. To approximate the solution, we replace one component of the system $\mathbf{B}^{(s)}\mathbf{w}^{(s)} = \mathbf{f}_0^{(s)}$ to satisfy $\sum_{k=1}^{g_0} \tilde{w}_k^{(s)} = 1$ and use the `optim()` optimization method in `R` to minimize the problem $(\mathbf{B}^{(s)}\mathbf{w}^{(s)} - \mathbf{f}_{\varepsilon_0}^{(s)})'(\mathbf{B}^{(s)}\mathbf{w}^{(s)} - \mathbf{f}_{\varepsilon_0}^{(s)})$ with respect to the positivity constraint of the weights. If some of the basis function weights have close to zero values, e. g., at the border knots, the associated recomputed weight often match the lower bound of the box constraints specified in `optim()`.

PART II. BAYESIAN REGULARIZATION IN THE CRR MODEL

7. Extended CRR model

The second popular regression model for continuous right-censored data treated in this work is the semiparametric relative risk model of Cox (1972), where the functional dependence of the survival time on the covariates is specified through the hazard rate function.

A very general, broad and flexible class of Cox type hazard regression models is already proposed by Hennerfeind et al. (2006). The authors extended the Cox model in two directions. On the one hand the logarithm of the nonparametric baseline survival hazard function is modeled by penalized B-splines that allow flexible, smooth shapes for the baseline hazard in the Cox model. On the other hand the predictor is extended to a structured additive predictor in the spirit of GAMs to model, in addition to the linear effects of some covariates, further covariates with other effect types, like smooth effects, time-varying effects, varying coefficients, nonlinear covariate interactions, random effects or spatial effects. The unified Bayesian modeling approach for this rich class of survival models is based on the fact that non-linear effects, including also the logarithm of the baseline hazard, can be expressed or approximated as linear combination of basis functions, where the basis function weights act as linear regression coefficients. This representation forces a purely linear structure of the predictor with appropriate defined design matrices. A common hierarchical model structure results, since all regression parameters in the predictor are equipped with conditional Gaussian priors given variance parameters, which drive the various forms of covariate-specific regularization, like smoothing of nonlinear or spatial effects. Finally, hyperpriors are assigned to the variance parameters to enable full Bayesian inference based on the *full likelihood*. The variance parameters, as an integral part of the model, are estimated jointly with the different covariate effects and the baseline hazard by MCMC simulation techniques.

In the subsequent section, the previous work of Hennerfeind et al. (2006) is expanded to take into account linear regularized effects utilizing informative shrinkage- and selection-type priors to consider also possibly high-dimensional covariates arising, e. g., in microarray-based survival studies. It is shown, that the inference of the regularized linear effects can be treated within the provided unifying framework, since the presented shrinkage priors also enable a hierarchical representation in terms of conditional Gaussian priors given variance parameters that drive the shrinkage towards zero. Therefore, inference for regularized linear effects is only described in combination with smooth effects of continuous covariates, because the inference is straightforward for model terms reflecting the previous mentioned other kinds of effects. The inferential procedures are implemented in the free software `BayesX`.

Bayesian analysis of the Cox model has also been studied in terms of the *partial likelihood*, where the estimation of the covariate effects is of primary interest and the baseline hazard is treated as a nuisance parameter. In the framework of the partial likelihood we consider also an extended predictor to model jointly regularized linear effects and nonlinear smooth effects of continuous covariates as prototype. Since the unified building block for representing the various kinds of effects does not change, if inference is based on the partial instead the full likelihood, the extensions to consider the manifold of effects listed above are managed identically. In summary, the unified modeling approach of the various types of covariate effects, as described in Hennerfeind et al. (2006), is also applicable to model the predictor and the corresponding priors under the partial likelihood. The inferential procedure is implemented in the R-function `bcoxpl()`.

7.1. Basic CRR model

Let $T_i \geq 0$, $i = 1,...,n$, denote the random variables representing the non negative, continuous survival times of the individuals from a heterogeneous population with the assumption that the survival times T_i, $i = 1,...,n$, are conditionally independent given covariates and parameters. The observed right censored survival data is given as $\mathfrak{D} = \{(\tilde{t}_i, d_i, \mathbf{v}_i'), i = 1,...,n\}$, where $\tilde{t}_i = \min(t_i, c_i)$ is the observed survival time, $d_i = 1(t_i \leq c_i) \in \{0,1\}$ is the observed censoring indicator and $\mathbf{v}_i = (v_{i1},...,v_{ip})'$ is the p-dimensional vector of the observed covariates for the n individuals of the sample. As pointed out in (1.6.) the hazard function $\lambda_i(\cdot)$ for individual i is assumed to be built as the product on an unspecified, covariate independent baseline hazard function $\lambda_0(\cdot) \geq 0$ and the exponential link $\exp(\eta_i)$ of the predictor $\eta_i \in \mathbb{R}$ that imports the summarized covariate effects, i. e.

$$\lambda_i(t \mid \boldsymbol{\vartheta}) = \lambda_0(t)\exp(\eta_i), \tag{7.1}$$

where $\boldsymbol{\vartheta}$ is an appropriate vector of regression parameters which will be specified in the following.

7.2. Extended predictor

As in Section 2.2 we partition the vector of explanatory covariates into three different treated subgroups of covariates $\mathbf{v}_i = (\mathbf{u}_i', \mathbf{x}_i', \mathbf{z}_i')'$ and consider a semiparametric form of the predictor $\eta_i = \eta_i(\boldsymbol{\vartheta})$ given by

$$\eta_i = \mathbf{u}_i'\boldsymbol{\gamma} + \mathbf{x}_i'\boldsymbol{\beta} + f_1(z_{i1}) + ... + f_{p_z}(z_{ip_z}) \tag{7.2}$$

that summarizes the different functional forms of the covariates.

The first component of the predictor describes the *linear effects* $\boldsymbol{\gamma} = (\gamma_0, \gamma_1, ..., \gamma_{p_u})'$ of a moderate low number of time-independent, categorical or continuous covariates $\mathbf{u}_i = (u_{i0,} u_{i1,} ..., u_{ip_u})' \subset \mathbf{v}_i$ with $p_u \ll n$, that are forced into the model and should not be regularized. In general it is not necessary to model an intercept term as regression parameter, because this parameter is common to all individuals and is therefore included in the baseline hazard. But for identifiability reasons, with respect to the level of the optional nonlinear terms, at least the global intercept term γ_0 with $u_{i0} = 1$, $i = 1,...,n$, is modeled. The second component describes *regularized linear effects* $\boldsymbol{\beta} = (\beta_1, ..., \beta_{p_x})'$ of possibly high-dimensional categorical or continuous time-independent covariates $\mathbf{x}_i = (x_{i1,} ..., x_{ip_x})' \subset \mathbf{v}_i$, with $p_x \leq n$ or $p_x > n$. The regression coefficients $\boldsymbol{\beta}$ are equipped with an informative shrinkage- or selection-type prior as provided in Sections 4.1 to 4.3 to identify those effects with the highest impact on the response. The remaining functions $f_j(\cdot)$, $j = 1,...,p_z$, are *smooth nonlinear effects* of time-independent continuous covariates z_j that need to be regularized to avoid overfitting. As outlined in Section 2.2 modeling of these unknown functions $f_j(\cdot)$ is based on Bayesian P-splines, compare Lang and Brezger (2004), where each function is approximated as a linear combination

$$f_j(z) = \sum_{k=1}^{g_j} \alpha_{j,k} B_{j,k}(z) = \mathbf{b}_j'(z)\boldsymbol{\alpha}_j$$

of B-spline basis functions $\mathbf{b}_j(\cdot) = (B_{j,1}(\cdot),...,B_{j,g_j}(\cdot))'$ and basis coefficients $\boldsymbol{\alpha}_j = (\alpha_{j,1},...,\alpha_{j,g_j})'$. The basis functions with degree q_j are defined on a sequence of equally spaced (inner) knots $\min(z_j) = \xi_1 < ... < \xi_{s_j} = \max(z_j)$, $g_j = s_j + q_j - 1$, from the domain of the j-th covariate z_j with additional boundary knots. A moderate number of knots is used to maintain the flexibility of the approximations in combination with Gaussian random walk priors for the basis coefficients that control the smoothness. For identification, the functions are centered horizontally about zero.

Further predictor components, like random effects to model unit- or cluster-specific heterogeneity, can be treated in a natural way in the Bayesian framework, where all parameters per se are considered as random variables. In particular, the distributional assumption about the random effects acts as the prior and can be cast into the regularized regression context, e. g. the Gaussian random intercept model with $\delta_i \sim N(0, \tau_\delta^2)$ in combination with an inverse gamma prior for the variance parameter τ_δ^2 corresponds to the Bayesian ridge prior of Section 4.1. Other optional components like *time-dependent covariates* ($z_{ij}'(t)\alpha_j$), *varying coefficients* ($f_j(z_{ik})z_{ij}$, where $f_j(\cdot)$ is a function of covariate z_k that modifies the effect of the covariate z_j), *time-varying effects* ($f_j(t)z_{ij}$, where $f_j(\cdot)$ is a time-dependent function that modifies the effect of the covariate z_j) or *spatial effects* (defined by smooth functions $f_{spat}(\cdot)$ of spatial indices of the geographical areas) can also be included in the predictor and cast into the unified modeling via penalized basis function expansions as shown e. g. in Brezger and Lang (2006) for exponential family regression, Kneib and Fahrmeir (2007), Hennerfeind et al. (2006) for geoadditive Cox-type survival regression models. By incorporating smooth effects or time-dependent covariate effects into the predictor the proportional hazards property is relaxed and the application of the resulting structured additive CRR regression models is not longer restricted to the assumption of proportional hazards.

7.3. Extended baseline hazard function

To obtain a flexible baseline survival distribution in the Cox model, the rather strong parametric assumptions for the baseline hazard function, like e. g. in the Weibull model, are relaxed by placing a P-spline model for the logarithm of the baseline hazard as suggested by Hennerfeind et al. (2006). In the similar manner like the smooth effects of the predictor, the log-baseline hazard is approximated by a linear combination of B-spline basis functions, i. e.

$$f_0(t) := \log \lambda_0(t) = \mathbf{b}_0'(t)\boldsymbol{\alpha}_0,$$

where $\boldsymbol{\alpha}_0 = (\alpha_{01},...,\alpha_{0g_0})'$ denotes the vector of basis function weights corresponding to an appropriate set of B-spline basis functions $\mathbf{b}_0(\cdot) = (B_{0,1}(\cdot),...,B_{0,g_0}(\cdot))'$ evaluated at the observed survival times $\tilde{t}_i$, $i = 1,...,n$. In particular, the piecewise exponential model, which states a step function for the baseline hazard function, is included as a special case when B-splines of degree zero are used. In this case the random walk prior prevents too large jumps between adjacent values of the baseline hazard pieces.

Modeling the log-transformation of the baseline hazard is advantageous, since it allows specifying the smoothness prior for the basis coefficients without any non-negativity restrictions for the regression coefficients $\boldsymbol{\alpha}_0$ to ensure the condition $\lambda_0(t) \geq 0$, $t \in \mathbb{R}_0^+$.

Generic notation: Rewriting the hazard function as $\lambda_i(t) = \exp(f_0(t) + \eta_i)$, the time-independent semiparametric predictor η_i in (7.2) can be further extended to take into account the time-dependent baseline hazard function

$$\eta_i(t) = \mathbf{u}_i'\boldsymbol{\gamma} + \mathbf{x}_i'\boldsymbol{\beta} + \mathbf{b}_0'(t)\boldsymbol{\alpha}_0 + \mathbf{b}_1'(z_{i1})\boldsymbol{\alpha}_1 + \ldots + \mathbf{b}_{p_z}'(z_{ip_z})\boldsymbol{\alpha}_{p_z} . \tag{7.3}$$

Due to the linear structure, the vector of predictors $\boldsymbol{\eta}(\tilde{\mathbf{t}}) = (\eta_1(\tilde{t}_1), \ldots, \eta_n(\tilde{t}_n))'$, evaluated at observed lifetimes $\tilde{t}_i$, $i = 1, \ldots, n$, can always be represented in generic matrix form

$$\boldsymbol{\eta}(\tilde{\mathbf{t}}) = \mathbf{X}\boldsymbol{\beta} + \mathbf{U}\boldsymbol{\gamma} + \mathbf{Z}_0\boldsymbol{\alpha}_0 + \ldots + \mathbf{Z}_m\boldsymbol{\alpha}_m ,$$

using appropriately defined design matrices $\mathbf{X}$ and $\mathbf{U}$, with rows $\mathbf{x}_i'$ and $\mathbf{u}_i'$, for the linear effects and the $n \times g_j$-dimensional design matrices $\mathbf{Z}_j$, with rows $\mathbf{b}_0'(\tilde{t}_i)$, $j = 0$, and $\mathbf{b}_j'(z_{ij})$, $j = 1, \ldots, p_z$, $i = 1, \ldots, n$, representing the evaluations of the basis functions.

7.4. Likelihood

7.4.1. Full likelihood

Joint inference for covariate effects and the baseline hazard is based on the full likelihood. For right censored data the full likelihood is given in (1.12). Inserting the hazard function with representation $\lambda_i(t) = \eta_i(t)$ from (7.3) into the expression of the full likelihood we obtain the following log-likelihood expression

$$l(\boldsymbol{\vartheta} \mid \mathfrak{D}) = \log L(\boldsymbol{\vartheta} \mid \mathfrak{D}) = \sum_{i=1}^{n} \left(d_i \eta_i(\tilde{t}_i) - \int_0^{\tilde{t}_i} \exp\left(\eta_i(s)\right) ds \right), \tag{7.4}$$

where $\boldsymbol{\vartheta} = (\boldsymbol{\alpha}', \boldsymbol{\beta}', \boldsymbol{\gamma}')'$, with $\boldsymbol{\alpha} = (\boldsymbol{\alpha}_0', \boldsymbol{\alpha}_1', \ldots, \boldsymbol{\alpha}_{p_z}')'$, denotes the vector of regression parameters.

The evaluation of the log-likelihood (7.4) requires the computation of the cumulative hazard function $\Lambda_i(t_i \mid \boldsymbol{\vartheta}) = \int_0^{t_i} \exp(\eta_i(s)) ds$ by integration over all time-dependent terms in the predictor. Because in our considerations the log-baseline hazard function $f_0(t) = \log(\lambda_0(t))$ is the only time-dependent function, the expression of the cumulative baseline hazard simplifies to $\Lambda_i(t \mid \boldsymbol{\vartheta}) = \Lambda_0(t) \cdot \exp(\eta_i)$, where η_i is the time-independent part of the predictor $\eta_i = \eta_i(t_i) - \mathbf{b}_0'(t_i)\boldsymbol{\alpha}_0$ given in (7.2) and $\Lambda_0(t) := \Lambda_0(t \mid \boldsymbol{\alpha}_0)$ denotes the cumulative baseline hazard function defined by

$$\Lambda_0(t) = \int_0^t \lambda_0(s) ds = \int_0^t \exp\left(\mathbf{b}_0'(s)\boldsymbol{\alpha}_0\right) ds . \tag{7.5}$$

Using the special functional form of the cumulative hazard function, the log-likelihood can finally be written as

$$l(\boldsymbol{\vartheta} \mid \mathfrak{D}) = \sum_{i=1}^{n} \left(d_i \eta_i(\tilde{t}_i) - \Lambda_0(\tilde{t}_i) \exp(\eta_i) \right). \tag{7.6}$$

Apart from simple parametric forms or using B-splines of degree 0 or 1 to model the log-baseline hazard $f_0(t)$, the integral in (7.5) can not be solved analytically and has to be evaluated numerically using, e. g., the trapezoidal rule.

7.4.2. Partial likelihood

A special feature of the CRR model (7.1) is the possibility to estimate the relationship between the hazard rate and the explanatory variables by treating the baseline hazard function $\lambda_0(\cdot)$ as nuisance parameter. Inference is carried out on the base of the partial likelihood (1.13), where $\lambda_0(\cdot)$ is discarded and hence we do not have to worry about the shape of the baseline hazard function. The logarithm of the partial likelihood is given as

$$\mathrm{pl}(\boldsymbol{\vartheta}\,|\,\mathfrak{D}) = \log \mathrm{pL}(\boldsymbol{\vartheta}\,|\,\mathfrak{D}) = \sum_{i=1}^{n} d_i \left[\eta_i - \log\left(\sum\nolimits_{k=1}^{n} 1_{(\tilde{t}_k \geq \tilde{t}_i)} \exp(\eta_k) \right) \right], \qquad (7.7)$$

where η_i denotes the extended predictor from (7.2). The indicator function in the log-likelihood (7.7) is used to describe the risk set $R(\tilde{t}_i) = \{k : \tilde{t}_k \geq \tilde{t}_i\}$ at the observed survival time $\tilde{t}_i$, which consists of all individuals who are event-free and still under observation just prior to time $\tilde{t}_i$. To estimate the distribution associated with the baseline hazard function (7.1), Breslow (1972, 1974) proposed to estimate for the cumulative baseline hazard $\Lambda_0(t)$ in a post-inferential step by

$$\hat{\Lambda}_0^{Br}(t) = \sum_{i=1}^{n} \frac{1_{(\tilde{t}_i \leq t)}\, d_i}{\sum\nolimits_{k=1}^{n} 1_{(\tilde{t}_k \geq \tilde{t}_i)} \exp(\hat{\eta}_k)}, \qquad (7.8)$$

where the Breslow estimate $\hat{\Lambda}_0^{BR}(\cdot)$ is computed on the base of the regression parameter estimates $\hat{\eta}_k = \eta_k(\hat{\boldsymbol{\vartheta}})$ from the partial likelihood.

Bayesian justification of the partial likelihood

While the partial likelihood is a widespread tool for frequentist inference of the CRR model, it is in general not clear, if the partial likelihood is valid for posterior analysis based on the Bayesian theorem (1.15), where commonly the full likelihood is used. The Bayesian partial likelihood approach is often justified by showing, that the posterior based on the partial likelihood approximates the full likelihood based marginal posterior of the regression coefficients, if a very diffuse prior for the baseline cumulative hazard function is assumed. We sketch the idea in the following and refer for details, e. g., to Kalbfleisch (1978), Sinha et al. (2003) and Kim and Kim (2009).

For simplicity we consider the case of linear effects $\boldsymbol{\vartheta} = \boldsymbol{\gamma}$. In this case Bayesian inference is based on the posterior density $p_{PL}(\boldsymbol{\gamma}\,|\,\mathfrak{D}) \propto PL(\boldsymbol{\gamma}\,|\,\mathfrak{D})p(\boldsymbol{\gamma})$, which is proportional to the product of the partial likelihood $PL(\boldsymbol{\gamma}\,|\,\mathfrak{D})$ and an arbitrary prior $p(\boldsymbol{\gamma})$ of $\boldsymbol{\gamma}$. The Bayesian justification of $p_{PL}(\boldsymbol{\gamma}\,|\,\mathfrak{D})$ for continuous univariate survival data and time-constant covariates is due to Kalbfleisch (1978). Under the assumption of a very diffuse gamma process prior used for the cumulative baseline hazard $\Lambda_0(t) \sim GP(c\Lambda_0^*(t), c)$ he showed, that the posterior density $p_{PL}(\boldsymbol{\gamma}\,|\,\mathfrak{D})$ can be viewed as an approximation of the marginal posterior of $\boldsymbol{\gamma}$

$$p(\boldsymbol{\gamma}\,|\,\mathfrak{D}) \propto \int L(\boldsymbol{\gamma}, \Lambda_0\,|\,\mathfrak{D})p(\boldsymbol{\gamma})p(\Lambda_0)d\Lambda_0 = p(\boldsymbol{\gamma})\int L(\boldsymbol{\gamma}, \Lambda_0\,|\,\mathfrak{D})p(\Lambda_0)d\Lambda_0\,, \qquad (7.9)$$

where $L(\boldsymbol{\gamma}, \Lambda_0\,|\,\mathfrak{D})$ denotes the full joint likelihood of $\boldsymbol{\gamma}$ and $\Lambda_0(t)$, and $p(\Lambda_0)$ denotes the gamma process distribution density. As expressed in (7.9), the prior for the regression parameter $p(\boldsymbol{\gamma})$ is assumed to be independent of $p(\Lambda_0)$. The distribution parameter $\Lambda_0^*(t)$ can be interpreted as an initial guess for the cumulative baseline hazard $\Lambda_0(t)$ and is often assumed to be a known differentiable parametric function depending on further hyperparameters, while the positive real number $c > 0$ is a weight attached to this guess. It is shown, that the corresponding marginal density $\tilde{L}(\gamma \mid \mathfrak{D}) = \int L(\gamma, \Lambda_0 \mid \mathfrak{D}) p(\Lambda_0) d\Lambda_0$, which depends on the hyperparameters $c > 0$ and those of $\Lambda^*(t)$, can be interpreted as likelihood function for γ given the data. In addition $\tilde{L}(\gamma \mid \mathfrak{D})$ provides for different choices of the weight parameter c a spectrum of likelihoods, where two limiting cases are of particular interest. For the very diffuse case $c \to 0$, placing a little faith on the prior guess $\Lambda_0^*(t)$, this marginal likelihood $\tilde{L}(\gamma \mid \mathfrak{D})$ is proportional to the partial likelihood $pL(\gamma \mid \mathfrak{D})$, so that the marginal posterior of γ in (7.9) is approximated by $p_{PL}(\gamma \mid \mathfrak{D}) \propto PL(\gamma \mid \mathfrak{D}) p(\gamma)$. On the other hand, for a strong trust in $\Lambda_0^*(t)$ with $c \to \infty$, the full likelihood $L(\gamma, \Lambda_0^* \mid \mathfrak{D})$ results with $\Lambda_0(t) = \Lambda_0^*(t)$. The examination of this marginal likelihood for varying values of the parameter $c > 0$ enables the evaluation how assumption-dependent the analysis is. Kalbfleisch (1978) showed also that, if the value c is small, the mean of the posterior distribution of the cumulative hazard is approximated by the Breslow estimate (7.8).

Sinha et al. (2003) picked up the approach and extend the results to take into account (external) time-dependent covariates, time-dependent effects, multivariate survival data (if frailties are modeled) and grouped survival data. Since the partial likelihood only depends on the observed order not on the exact values of the failure times, corrections are required if *ties* are present to take into account the permutation of those individuals with identical survival times. This is due to the fact, that the partial likelihood considers only the observed order of the survival times and, if more than one individual has its event at the same time, the ordering is no longer unique. In the Bayesian framework Kim and Kim (2009) investigate corrections of the partial likelihood when many ties are present and they provide a Bayesian justification of using the exact partial likelihood of Peto (1972) in such situations.

8. Priors for the extended CRR model

To complete the Bayesian formulation of the CRR regression model, the regression parameters are equipped with more or less informative regularization priors as presented in the Section 4. The priors are identical to the priors used in the extended AFT model, which emphasizes again the uniformity of the Bayesian approach. We use again $\boldsymbol{\rho}$ as generic notation for the shrinkage prior-specific hyperparameters from further stages of the hierarchical formulation.

Prior of the unregularized linear effects

The prior for the low-dimensional linear effects $\boldsymbol{\gamma} = (\gamma_{0,}..., \gamma_{p_u})'$, which are forced into the model, is assumed to be weakly informative Gaussian

$$\boldsymbol{\gamma} \mid \boldsymbol{\mu}_\gamma, \boldsymbol{\Sigma}_\gamma \sim N(\boldsymbol{\mu}_\gamma, \boldsymbol{\Sigma}_\gamma), \tag{8.1}$$

with $\boldsymbol{\mu}_\gamma = \mathbf{0}$ and $\boldsymbol{\Sigma}_\gamma^{-1} \to \mathbf{0}$. Alternatively we use the product of independent diffuse priors $p(\gamma_j) \propto \text{const.}$, $j = 0, 1, ..., p_u$. In general we use the formulation (8.1) as blueprint to derive the conditional posterior densities, because the remaining regularization priors are also conditional Gaussian and differ only in the specification of the mean vector and covariance matrix.

Prior of the regularized linear effects

For possibly high-dimensional regularized linear effects $\boldsymbol{\beta} = (\beta_1,...,\beta_{p_x})'$ we use the shrinkage or selection priors corresponding to the Bayesian lasso, Bayesian ridge or Bayesian NMIG hierarchy as described in Sections 4.1 to 4.3. In particular, all priors are conditional Gaussian

$$\boldsymbol{\beta} \mid \boldsymbol{\tau}_\beta^2 \sim N(\mathbf{0}, \boldsymbol{\Sigma}_\beta), \tag{8.2}$$

with diagonal covariance matrix $\boldsymbol{\Sigma}_\beta = \mathbf{D}_{\tau_\beta} = \text{diag}(\tau_{\beta_1}^2,...,\tau_{\beta_{p_x}}^2)$, where the variance parameters $\tau_{\beta_j}^2$ drive the covariate-specific shrinkage of the regression coefficients towards the mean $\boldsymbol{\mu}_\beta = \mathbf{0}$. The associated priors for the variance and shrinkage parameters are:

Bayesian ridge version (A) ($\boldsymbol{\rho} = (\lambda_1,...,\lambda_{p_x})$)

$$\tau_{\beta_j}^2 \mid \lambda \sim \delta_{1/2\lambda_j}(\tau_{\beta_j}^2)\,,\, j=1,...,p_x\,, \tag{8.3}$$

$$\lambda_j \sim_{\text{iid}} \text{Gamma}(h_{1,\lambda}, h_{1,\lambda});\quad h_{1,\lambda}, h_{1,\lambda} > 0\,,\, j=1,...,p_x\,. \tag{8.4}$$

Bayesian ridge version (B) ($\boldsymbol{\rho} = \lambda$)

$$\tau_\beta^2 \mid \lambda \sim \delta_{1/2\lambda}(\tau_\beta^2)\,,\, j=1,...,p_x\,, \tag{8.5}$$

$$\lambda \sim \text{Gamma}(h_{1,\lambda}, h_{1,\lambda});\quad h_{1,\lambda}, h_{1,\lambda} > 0\,. \tag{8.6}$$

Bayesian lasso ($\boldsymbol{\rho} = \lambda$)

$$\tau_{\beta_j}^2 \mid \lambda^2 \sim_{\text{iid}} \text{Exp}\left(\frac{\lambda^2}{2}\right),\ j=1,...,p_x\,, \tag{8.7}$$

$$\lambda^2 \sim \text{Gamma}(h_{1,\lambda}, h_{1,\lambda});\quad h_{1,\lambda}, h_{1,\lambda} > 0\,. \tag{8.8}$$

Bayesian NMIG with $\tau_{\beta_j}^2 = I_j \psi_j^2$ ($\boldsymbol{\rho} = \omega$)

$$I_j \mid v_0, v_1, \omega \sim_{\text{iid}} \text{Bernoulli}(\omega; v_0, v_1),\ j=1,...,p_x\,, \tag{8.9}$$

$$\psi_j^2 \mid h_{1,\psi}, h_{2,\psi} \sim_{\text{iid}} \text{IGamma}(h_{1,\psi}, h_{2,\psi}),\quad h_{1,\psi}, h_{2,\psi} > 0\,,\, j=1,...,p_x\,, \tag{8.10}$$

$$\omega \sim \text{Beta}(h_{1,\omega}, h_{2,\omega});\quad h_{1,\omega}, h_{2,\omega} > 0\,. \tag{8.11}$$

Prior of the nonlinear effects and the log-baseline hazard

The priors for the basis function coefficients $\boldsymbol{\alpha} := (\boldsymbol{\alpha}_0', \boldsymbol{\alpha}_1', ..., \boldsymbol{\alpha}_{p_z}')'$ of the nonlinear effects and the log-baseline hazard are specified by random walks of d_j-th order. We obtain conditional, partially improper Gaussian smoothing priors as defined in Section 4.6 with

$$\boldsymbol{\alpha}_j \mid \tau_{\alpha_j}^2 \sim N(\mathbf{0}, \boldsymbol{\Sigma}_{\alpha_j}^-),\ j=0,1,...,p_z\,, \tag{8.12}$$

where $\boldsymbol{\Sigma}_{\alpha_j}^- := \tau_{\alpha_j}^2 \mathbf{K}_j^-$ denotes the covariance matrix and $\mathbf{K}_j^-$ is a generalized inverse of the penalty matrix $\mathbf{K}_j$ with rank $\text{rank}(\mathbf{K}_j) = g_j - q_j$. Diffuse priors are initially used for the q_j coefficients $p(\alpha_{j,1}) \propto \text{const}$,..., $p(\alpha_{j,q_j}) \propto \text{const}$ and the smoothness controlling variance parameters $\boldsymbol{\tau}_\alpha^2 := (\tau_{\alpha_0}^2, \tau_{\alpha_1}^2, ... \tau_{\alpha_{p_z}}^2)'$ are equipped with inverse gamma distributions

$$\tau^2_{\alpha_j} \sim \text{InvGamma}\left(h_{1,\tau_j}, h_{2,\tau_j}\right),\ j = 0,1,...,p_z . \tag{8.13}$$

Joint prior distribution

With the independence assumptions implied in the prior definitions, i. e. that all priors are conditionally and mutually independent, the joint prior distribution of the set of model parameters $\boldsymbol{\theta} = (\boldsymbol{\vartheta}, \boldsymbol{\tau}^2_\alpha, \boldsymbol{\tau}^2_\beta, \boldsymbol{\rho})$ is given by the product

$$p(\boldsymbol{\theta}) = \prod_{j=0}^{p_z} p(\boldsymbol{\alpha}_j \mid \tau^2_{\alpha,j}) p(\tau^2_{\alpha,j}) \cdot p(\boldsymbol{\beta} \mid \boldsymbol{\tau}^2_\beta) p(\boldsymbol{\tau}^2_\beta \mid \boldsymbol{\rho}) p(\boldsymbol{\rho}) p(\boldsymbol{\gamma}), \tag{8.14}$$

where $\boldsymbol{\rho}$ is a generic notation for the shrinkage prior-specific hyperparameters from further stages of the hierarchical formulation. If inference is based on the partial likelihood, the factors $p(\boldsymbol{\alpha}_0 \mid \tau^2_{\alpha,0}) p(\tau^2_{\alpha,0})$, modeling the assumptions of baseline hazard, are discarded.

9. MCMC inference in extended CRR model

Bayesian Inference via MCMC simulation is based on updating full conditionals of single parameters or blocks of parameters given the rest of the parameters and the data. The unified prior structure for functions and parameters is leading to full conditionals with a similar unified structure. For non Gaussian responses, Gibbs sampling as for the regression parameters in the AFT model, is no longer feasible and more general Metropolis-Hastings (MH) algorithms are required. We construct a Markov chain with MH steps using IWLS fashioned proposals, as suggested by Gamerman (1997) and shortly described in Appendix C. Due to the beneficial hierarchical structure of the model, Gibbs sampling for the regularization parameters is still feasible. We first describe MCMC inference for the extended model based on the full likelihood (7.6) with the predictor (7.3), where joint shrinkage and smoothing of covariate effects together with a smooth estimation of the log-baseline hazard are of primary interest. Inference for shrinkage and smoothing of covariate effects based on the partial likelihood (7.7) with the extended predictor (7.2) is outlined subsequently.

9.1. Conditional posterior densities based on the full likelihood

Using the Bayes theorem, the joint posterior $p(\boldsymbol{\theta} \mid \mathfrak{D})$ is proportional to the product of the model likelihood $L(\boldsymbol{\theta} \mid \mathfrak{D})$ and the joint prior density of the model parameters $p(\boldsymbol{\theta})$. Based on the full log-likelihood (7.6) of the extended Cox model with the extended predictor given in (7.3) and the prior (8.14) the posterior has the general form

$$p\left(\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\gamma}, \boldsymbol{\tau}^2_\alpha, \boldsymbol{\tau}^2_\beta, \boldsymbol{\rho} \mid \mathfrak{D}\right) \propto \exp\left(l(\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\gamma} \mid \mathfrak{D})\right) \prod_{j=0}^{p_z} p(\boldsymbol{\alpha}_j \mid \tau^2_{\alpha,j}) p(\tau^2_{\alpha,j}) \cdot p(\boldsymbol{\beta} \mid \boldsymbol{\tau}^2_\beta) p(\boldsymbol{\tau}^2_\beta \mid \boldsymbol{\rho}) p(\boldsymbol{\rho}) p(\boldsymbol{\gamma}). \tag{9.1}$$

9.1.1. Full conditionals of the predictor components

Unregularized linear regression coefficients $\boldsymbol{\gamma}$

In the following we derive the general structure of the full conditionals and the proposal distributions for the predictor components $\boldsymbol{\vartheta} = (\boldsymbol{\alpha}', \boldsymbol{\beta}', \boldsymbol{\gamma}')'$ in terms of the unpenalized regression coefficients $\boldsymbol{\gamma}$

assuming for a while the prior (8.1), $\boldsymbol{\gamma} \mid \boldsymbol{\mu}_\gamma, \boldsymbol{\Sigma}_\gamma \sim \mathrm{N}(\boldsymbol{\mu}_\gamma, \boldsymbol{\Sigma}_\gamma)$, with $\boldsymbol{\mu}_\gamma \neq \mathbf{0}$ and $\boldsymbol{\Sigma}_\gamma^{-1} \neq \mathbf{0}$ to preserve the generality. Discarding the factors in the posterior (9.1) that do not depend on $\boldsymbol{\gamma}$, the full conditional of the regression parameter $\boldsymbol{\gamma}$ is given by

$$p(\boldsymbol{\gamma} \mid \cdot) \propto \exp\left\{ l(\boldsymbol{\vartheta} \mid \mathfrak{D}) - \frac{1}{2}\boldsymbol{\gamma}'\boldsymbol{\Sigma}_\gamma^{-1}\boldsymbol{\gamma} + \boldsymbol{\gamma}'\boldsymbol{\Sigma}_\gamma^{-1}\boldsymbol{\mu}_\gamma \right\}. \tag{9.2}$$

To construct the Gaussian IWLS proposal, we apply a second order Taylor expansion to the logarithm of the full conditional $f(\boldsymbol{\gamma}) = \log(p(\boldsymbol{\gamma} \mid \cdot))$ at the current state of the parameter vector $\boldsymbol{\gamma}^{(c)}$, compare Appendix C for details. Differentiating $f(\boldsymbol{\gamma})$ with respect to $\boldsymbol{\gamma}$ gives the score vector

$$\mathbf{s}_\gamma(\boldsymbol{\gamma}) = \frac{\partial l(\boldsymbol{\vartheta} \mid \mathfrak{D})}{\partial \boldsymbol{\gamma}} - \boldsymbol{\Sigma}_\gamma^{-1}\boldsymbol{\gamma} + \boldsymbol{\Sigma}_\gamma^{-1}\boldsymbol{\mu}_\gamma \tag{9.3}$$

and the hessian matrix

$$\mathbf{H}_\gamma(\boldsymbol{\gamma}) = \frac{\partial^2 l(\boldsymbol{\vartheta} \mid \mathfrak{D})}{\partial \boldsymbol{\gamma} \partial \boldsymbol{\gamma}'} - \boldsymbol{\Sigma}_\gamma^{-1}, \tag{9.4}$$

with the following contributions from the differentiation of the log-likelihood

$$\frac{\partial l(\boldsymbol{\vartheta} \mid \mathfrak{D})}{\partial \boldsymbol{\gamma}} = \sum_{i=1}^{n} d_i \mathbf{u}_i - \int_0^{\tilde{t}_i} \mathbf{u}_i \exp(\eta_i(s))ds = \sum_{i=1}^{n} d_i \mathbf{u}_i - \Lambda_0(\tilde{t}_i)\eta_i \mathbf{u}_i \ ,$$

$$\frac{\partial^2 l(\boldsymbol{\vartheta} \mid \mathfrak{D})}{\partial \boldsymbol{\gamma} \partial \boldsymbol{\gamma}'} = -\sum_{i=1}^{n} \int_0^{\tilde{t}_i} \mathbf{u}_i \mathbf{u}_i' \exp(\eta_i(s))ds = -\sum_{i=1}^{n} \Lambda_0(\tilde{t}_i)\eta_i \mathbf{u}_i \mathbf{u}_i' .$$

The second order Taylor expansion of $f(\boldsymbol{\gamma})$ around the current state of the parameter vector $\boldsymbol{\gamma}^{(c)}$ has the form

$$\hat{f}(\boldsymbol{\gamma}) \approx f(\boldsymbol{\gamma}^{(c)}) + (\boldsymbol{\gamma} - \boldsymbol{\gamma}^{(c)})' \mathbf{s}_\gamma(\boldsymbol{\gamma}^{(c)}) + \frac{1}{2}(\boldsymbol{\gamma} - \boldsymbol{\gamma}^{(c)})' \mathbf{H}_\gamma(\boldsymbol{\gamma}^{(c)})(\boldsymbol{\gamma} - \boldsymbol{\gamma}^{(c)}), \tag{9.5}$$

where $\mathbf{s}_\gamma(\boldsymbol{\gamma}^{(c)})$ and $\mathbf{H}_\gamma(\boldsymbol{\gamma}^{(c)})$ denote the score vector (9.3) and the Hessian matrix (9.4) evaluated at the current state $\boldsymbol{\gamma}^{(c)}$ and the current states of the remaining involved parameters $\boldsymbol{\vartheta}_{-\gamma} = (\boldsymbol{\alpha}^{(c)}, \boldsymbol{\beta}^{(c)})$. Building the exponential of approximation (9.5) and neglecting the components that do not depend on $\boldsymbol{\gamma}$ provides the following structure of the proposal density

$$\varphi(\boldsymbol{\gamma}^{(p)} \mid \hat{\boldsymbol{\mu}}_\gamma^{(c)}, \hat{\boldsymbol{\Sigma}}_\gamma^{(c)}) \propto \exp\left\{ \frac{1}{2}\boldsymbol{\gamma}'\mathbf{H}_\gamma(\boldsymbol{\gamma}^{(c)})\boldsymbol{\gamma} + \boldsymbol{\gamma}'\left(\mathbf{s}_\gamma(\boldsymbol{\gamma}^{(c)}) - \mathbf{H}_\gamma(\boldsymbol{\gamma}^{(c)})\boldsymbol{\gamma}^{(c)}\right) \right\},$$

which represents the kernel of a multivariate Gaussian distribution density $\varphi(\cdot \mid \hat{\boldsymbol{\mu}}_\gamma^{(c)}, \hat{\boldsymbol{\Sigma}}_\gamma^{(c)})$ with mean vector and covariance matrix

$$\hat{\boldsymbol{\mu}}_\gamma^{(c)} = \hat{\boldsymbol{\Sigma}}_\gamma^{(c)}\left(\mathbf{s}_\gamma(\boldsymbol{\gamma}^{(c)}) - \mathbf{H}_\gamma(\boldsymbol{\gamma}^{(c)})\boldsymbol{\gamma}^{(c)}\right), \quad \hat{\boldsymbol{\Sigma}}_\gamma^{(c)} = \left(-\mathbf{H}_\gamma(\boldsymbol{\gamma}^{(c)})\right)^{-1}. \tag{9.6}$$

As already mentioned in the context of the extended AFT model, the reformulation of the mean as $\hat{\boldsymbol{\mu}}_\gamma^{(c)} = \boldsymbol{\gamma}^{(c)} - \mathbf{H}_\gamma(\boldsymbol{\gamma}^{(c)})\mathbf{s}_\gamma(\boldsymbol{\gamma}^{(c)})$ enables the interpretation as one-step-approximation to the mode of the full conditional obtained by a single Fisher scoring step from the current state.

Using the notation $\boldsymbol{\Lambda}(\tilde{\mathbf{t}} \mid \boldsymbol{\vartheta}) := (\Lambda(\tilde{t}_1 \mid \boldsymbol{\vartheta}), \ldots, \Lambda(\tilde{t}_n \mid \boldsymbol{\vartheta}))'$ as the vector of the cumulative baseline hazards evaluated at the observed survival times $\tilde{t}_i$ and $\mathbf{d} := (d_1, \ldots, d_n)'$ as the vector of censoring indicators, the score vector has the compact form

$$\mathbf{s}_\gamma\left(\boldsymbol{\gamma}^{(c)}\right) = \mathbf{U}'\left(\mathbf{d} - \boldsymbol{\Lambda}(\tilde{\mathbf{t}} \mid \boldsymbol{\gamma}^{(c)}, \boldsymbol{\vartheta}_{-\gamma})\right) - \boldsymbol{\Sigma}_\gamma^{-1}\boldsymbol{\gamma}^{(c)} + \boldsymbol{\Sigma}_\gamma^{-1}\boldsymbol{\mu}_\gamma \tag{9.7}$$

and the Hessian matrix reads

$$\mathbf{H}_\gamma\left(\boldsymbol{\gamma}^{(c)}\right) = -\mathbf{U}'\mathrm{diag}\left(\boldsymbol{\Lambda}(\tilde{\mathbf{t}} \mid \boldsymbol{\gamma}^{(c)}, \boldsymbol{\vartheta}_{-\gamma})\right)\mathbf{U} - \boldsymbol{\Sigma}_\gamma^{-1}. \tag{9.8}$$

Finally, a new proposed value $\boldsymbol{\gamma}^{(p)}$ of the Markov chain is obtained by drawing a random number from the Gaussian distribution with density $\varphi(\cdot \mid \hat{\boldsymbol{\mu}}_\gamma^{(c)}, \hat{\boldsymbol{\Sigma}}_\gamma^{(c)})$. The new state is accepted with the probability

$$\mathrm{p}_{\mathrm{accept}}\left(\boldsymbol{\gamma}^{(p)}, \boldsymbol{\gamma}^{(c)}\right) = \min\left\{1, \frac{\mathrm{p}\left(\boldsymbol{\gamma}^{(p)} \mid \cdot\right)\varphi(\boldsymbol{\gamma}^{(c)} \mid \hat{\boldsymbol{\mu}}_\gamma^{(p)}, \hat{\boldsymbol{\Sigma}}_\gamma^{(p)})}{\mathrm{p}\left(\boldsymbol{\gamma}^{(c)} \mid \cdot\right)\varphi(\boldsymbol{\gamma}^{(p)} \mid \hat{\boldsymbol{\mu}}_\gamma^{(c)}, \hat{\boldsymbol{\Sigma}}_\gamma^{(c)})}\right\},$$

where $\mathrm{p}(\boldsymbol{\gamma}^{(p)} \mid \cdot)$ and $\mathrm{p}(\boldsymbol{\gamma}^{(c)} \mid \cdot)$ denote the evaluations of the full conditional (9.2) at the proposed state $\boldsymbol{\gamma}^{(p)}$ and the current state $\boldsymbol{\gamma}^{(c)}$ with respect to the current states of the remaining involved model parameters $\boldsymbol{\vartheta}_{-\gamma}$. The mean vector $\hat{\boldsymbol{\mu}}_\gamma^{(p)}$ and the covariance matrix $\hat{\boldsymbol{\Sigma}}_\gamma^{(p)}$, appearing in the acceptance probability, are computed by the evaluation of the expressions in (9.6) at the proposed value $\boldsymbol{\gamma}^{(p)}$ keeping the remaining parameters $\boldsymbol{\vartheta}_{-\gamma}$ fixed at their current states.

In particular, since we assume a flat Gaussian prior for the unregularized effects $\boldsymbol{\gamma}$ that corresponds to the limiting case $\boldsymbol{\mu}_\gamma = \mathbf{0}$ and $\boldsymbol{\Sigma}_\gamma^{-1} = \mathbf{0}$, the mean and covariance matrix of the Gaussian proposal for the unregularized effects $\boldsymbol{\gamma}$ are given as

$$\begin{aligned} \hat{\boldsymbol{\mu}}_\gamma^{(c)} &= \hat{\boldsymbol{\Sigma}}_\gamma^{(c)}\left[\mathbf{U}'\mathbf{d} - \mathbf{U}'\boldsymbol{\Lambda}\left(\tilde{\mathbf{t}} \mid \boldsymbol{\gamma}^{(c)}, \boldsymbol{\vartheta}_{-\gamma}\right) + \mathbf{U}'\mathrm{diag}\left(\boldsymbol{\Lambda}\left(\tilde{\mathbf{t}} \mid \boldsymbol{\gamma}^{(c)}, \boldsymbol{\vartheta}_{-\gamma}\right)\right)\mathbf{U}\boldsymbol{\gamma}^{(c)}\right], \\ \hat{\boldsymbol{\Sigma}}_\gamma^{(c)} &= \left(\mathbf{U}'\mathrm{diag}\left(\boldsymbol{\Lambda}\left(\tilde{\mathbf{t}} \mid \boldsymbol{\gamma}^{(c)}, \boldsymbol{\vartheta}_{-\gamma}\right)\right)\mathbf{U}\right)^{-1}. \end{aligned} \tag{9.9}$$

Regularized linear regression coefficients β

For the remaining regression coefficients $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ the IWLS proposal densities can conceptually be carried out in the same way as above for $\boldsymbol{\gamma}$. We obtain the corresponding expressions for the mean and the covariance matrix of the Gaussian proposal by replacing the design matrix $\mathbf{U}$, the precision matrix $\boldsymbol{\Sigma}_\gamma^{-1}$ and the mean $\boldsymbol{\mu}_\gamma$ in (9.7) and (9.8) with the associated quantities from the priors of the regularized linear effects and the smooth effects. Proceeding as before, using the conditional Gaussian prior $\boldsymbol{\beta} \mid \boldsymbol{\tau}_\beta^2 \sim \mathrm{N}(\mathbf{0}, \boldsymbol{\Sigma}_\beta)$ from (8.2) the full conditional of $\boldsymbol{\beta}$ is given as

$$\mathrm{p}\left(\boldsymbol{\beta} \mid \cdot\right) \propto \exp\left\{\mathrm{l}(\boldsymbol{\vartheta} \mid \mathfrak{D}) - \frac{1}{2}\boldsymbol{\beta}'\mathbf{D}_\tau^{-1}\boldsymbol{\beta}\right\}. \tag{9.10}$$

Given the current state $\boldsymbol{\beta}^{(c)}$ and the current states of the remaining regression coefficients $\boldsymbol{\vartheta}_{-\beta} = (\boldsymbol{\alpha}^{(c)}, \boldsymbol{\gamma}^{(c)})$, proposals are drawn from a Gaussian density with mean vector and covariance matrix

$$\begin{aligned} \hat{\boldsymbol{\mu}}_\beta^{(c)} &= \hat{\boldsymbol{\Sigma}}_\beta^{(c)}\left[\mathbf{X}'\mathbf{d} - \mathbf{X}'\boldsymbol{\Lambda}\left(\tilde{\mathbf{t}} \mid \boldsymbol{\beta}^{(c)}, \boldsymbol{\vartheta}_{-\beta}\right) + \mathbf{X}'\mathrm{diag}\left(\boldsymbol{\Lambda}\left(\tilde{\mathbf{t}} \mid \boldsymbol{\beta}^{(c)}, \boldsymbol{\vartheta}_{-\beta}\right)\right)\mathbf{X}\boldsymbol{\beta}^{(c)}\right], \\ \hat{\boldsymbol{\Sigma}}_\beta^{(c)} &= \left(\mathbf{X}'\mathrm{diag}\left(\boldsymbol{\Lambda}\left(\tilde{\mathbf{t}} \mid \boldsymbol{\beta}^{(c)}, \boldsymbol{\vartheta}_{-\beta}\right)\right)\mathbf{X} - \mathbf{D}_\tau^{-1}\right)^{-1}. \end{aligned} \tag{9.11}$$

Regularized nonlinear regression coefficients $\boldsymbol{\alpha}_j$

Using the conditional Gaussian priors $\boldsymbol{\alpha}_j \mid \tau^2_{\alpha_j} \sim \mathrm{N}\left(\mathbf{0}, \boldsymbol{\Sigma}^-_{\alpha_j}\right)$ from (9.12), the full conditionals of the basis function weights coefficients $\boldsymbol{\alpha}_j$ are

$$p\left(\boldsymbol{\alpha}_j \mid \cdot\right) \propto \exp\left\{ l\left(\boldsymbol{\vartheta} \mid \mathfrak{D}\right) - \frac{1}{2\tau^2_{\alpha_j}} \boldsymbol{\alpha}_j' \mathbf{K}_j \boldsymbol{\alpha}_j \right\}, \quad j = 0,1,\ldots,p_z . \tag{9.12}$$

Given the current state $\boldsymbol{\alpha}_j^{(c)}$ and the current states of the remaining regression coefficients $\boldsymbol{\vartheta}_{-\alpha_j} = (\boldsymbol{\alpha}_0^{(c)},\ldots,\boldsymbol{\alpha}_{j-1}^{(c)},\boldsymbol{\alpha}_{j+1}^{(c)},\ldots,\boldsymbol{\alpha}_{p_z}^{(c)},\boldsymbol{\gamma}^{(c)},\boldsymbol{\beta}^{(c)})$, proposals are drawn from a Gaussian distribution with mean vector and covariance matrix

$$\begin{aligned} \hat{\boldsymbol{\mu}}^{(c)}_{\alpha,j} &= \hat{\boldsymbol{\Sigma}}^{(c)}_{\alpha,j}\left[\mathbf{Z}_j'\mathbf{d} - \mathbf{Z}_j'\boldsymbol{\Lambda}\left(\tilde{\mathbf{t}} \mid \boldsymbol{\alpha}_j^{(c)}, \boldsymbol{\vartheta}_{-\alpha_j}\right) + \mathbf{Z}_j'\mathrm{diag}\left(\boldsymbol{\Lambda}\left(\tilde{\mathbf{t}} \mid \boldsymbol{\alpha}_j^{(c)}, \boldsymbol{\vartheta}_{-\alpha_j}\right)\right)\mathbf{Z}_j\boldsymbol{\alpha}_j^{(c)}\right] \\ \hat{\boldsymbol{\Sigma}}^{(c)}_{\alpha,j} &= \left(\mathbf{Z}_j'\mathrm{diag}\left(\boldsymbol{\Lambda}\left(\tilde{\mathbf{t}} \mid \boldsymbol{\alpha}_j^{(c)}, \boldsymbol{\vartheta}_{-\alpha_j}\right)\right)\mathbf{Z}_j - \frac{1}{\tau^2_{\alpha_j}}\mathbf{K}_j\right)^{-1} . \end{aligned} \tag{9.13}$$

In particular the band structure of the precision matrices of the Gaussian proposal distributions for the basis coefficients $\boldsymbol{\alpha}_j$ enables an efficient computation of the Cholesky decomposition and a fast implementation of the algorithm of Rue (2001), compare Section 6.1.7, to draw a new proposal and compute the acceptance probability. To identify the model, the B-spline coefficients are centered as described in Section 6.1.1.

The computational efforts increase for the update of the baseline hazard coefficients $\boldsymbol{\alpha}_0$, because the time-dependent cumulative baseline hazard function $\Lambda_0(t) = \int_0^t \exp(\mathbf{b}_0'(s)\boldsymbol{\alpha}_0 ds$ and the associated time-dependent derivates are involved, complicating the computation of the score function and the Hessian matrix of the likelihood. We obtain

$$\frac{\partial l(\boldsymbol{\vartheta} \mid \mathfrak{D})}{\partial \boldsymbol{\alpha}_0} = \sum_{i=1}^n d_i \mathbf{b}_0(t_i) - \exp(\eta_i)\int_0^{t_i} \mathbf{b}_0(s)\exp(\mathbf{b}_0'(s)\boldsymbol{\alpha}_0)ds ,$$

$$\frac{\partial^2 l(\boldsymbol{\vartheta} \mid \mathfrak{D})}{\partial \boldsymbol{\alpha}_0 \partial \boldsymbol{\alpha}_0'} = -\sum_{i=1}^n \exp(\eta_i)\int_0^{t_i} \mathbf{b}_0(s)\mathbf{b}_0'(s)\exp(\mathbf{b}_0'(s)\boldsymbol{\alpha}_0)ds ,$$

and we have to evaluate this time-dependent expressions concerning the log-baseline hazard P-spline model by numerical integration in every iteration of the sampler. As computationally more efficient alternative, one may use MH steps with conditional prior proposals, as developed in Knorr-Held (1999) for state space models and applied to geoadditive hazard rate models in Hennerfeind et al. (2006), which require only the evaluation of the log-likelihood and not evaluation of the derivates.

Weibull baseline hazard model

In addition to the P-spline based approach, we consider a simple parametric Weibull model to model the baseline hazard with $\lambda_0(t) := \alpha_0 t^{\alpha_0 - 1}$ and $\lambda_i(t \mid \boldsymbol{\theta}) = \alpha_0 t^{\alpha_0} \exp(\eta_i(t))$ as competitor. In the extended predictor (7.3) the nonlinear log-baseline component $f_0(t) = \log(\lambda_0(t))$ has in this case the special form $f_0(t) = \log(\alpha_0) + (\alpha_0 - 1)\log(t)$.

Typically a gamma prior is employed to model the prior knowledge about the shape parameter α_0, compare e. g. Ibrahim et al. (2001), i. e.,

$$\alpha_0 \sim \mathrm{Gamma}\left(h_{1,\alpha_0}, h_{2,\alpha_0}\right), \quad h_{1,\alpha_0} > 0, h_{2,\alpha_0} > 0 \tag{9.14}$$

with $\mathbb{E}(\alpha_0) = h_{1,\alpha_0}/h_{2,\alpha_0}$ and $\mathbb{V}ar(\alpha_0) = h_{1,\alpha_0}/h_{2,\alpha_0}^2$. If identical hyperparameters like $h_{1,\alpha_0} = 0.01$ and $h_{1,\alpha_0} = 0.01$ are used, the prior mean of α_0 equals one, which corresponds to a constant hazard over time with a large variance of 100. The update of the log-baseline hazard in this parametric case is achieved by the update of the single parameter α_0. With the prior assumption (9.14), and the likelihood contributions $l_i(\boldsymbol{\vartheta} | \mathfrak{D}) = d_i \log(\alpha_0) + d_i(\alpha_0 - 1)\log(\tilde{t}_i) + d_i\eta_i - \Lambda_i(\tilde{t}_i \,|\, \boldsymbol{\vartheta})$ the full conditional of α_0 is given by

$$p(\alpha_0 \,|\, \cdot) \propto \exp\left\{ \sum_{i=1}^{n} d_i \log(\alpha_0) + d_i(\alpha_0 - 1)\log(t_i) - \Lambda_i(t_i \,|\, \boldsymbol{\vartheta}) + (h_{1,\alpha_0} - 1)\log(\alpha_0) - h_{2,\alpha_0}\alpha_0 \right\}, \quad (9.15)$$

where in our case of time-independent covariates in the predictor the expression for the cumulative baseline hazard is simplified to $\Lambda_i(\tilde{t}_i \,|\, \boldsymbol{\vartheta}) = \tilde{t}_i^{\alpha_0} \exp(\eta_i)$. Since the full conditional does not have a closed form, we use again MH steps to update the shape parameter α_0. The proposal $\alpha_0^{(p)}$ is drawn from a gamma distribution

$$q\left(\cdot \,|\, \alpha_0^{(c)}, d_{\alpha_0}\right) \sim \text{Gamma}\left(d_{\alpha_0}\alpha_0^{(c)}, d_{\alpha_0}\right) \quad (9.16)$$

based on the current value $\alpha_0^{(c)}$, which leads to the acceptance probability

$$p_{\text{accept}}\left(\alpha_0^{(p)}, \alpha_0^{(c)}\right) = \min\left\{ 1, \frac{p\left(\alpha_0^{(p)} \,|\, \cdot\right) q\left(\alpha_0^{(c)} \,|\, \alpha_0^{(p)}, d_{\alpha_0}\right)}{p\left(\alpha_0^{(c)} \,|\, \cdot\right) q\left(\alpha_0^{(p)} \,|\, \alpha_0^{(c)}, d_{\alpha_0}\right)} \right\}.$$

The value of d_{α_0} is determined during the burnin to achieve reasonable acceptance rates. In addition slice sampling is possible, because the full conditional is log-concave, see Ibrahim et al. (2001), and thereby also adaptive rejection sampling can be applied to update the shape parameter of the Weibull model. However, the update of the remaining model parameters is practiced as in the case when the baseline is modeled by a P-spline. In particular, we only have to replace the logarithm of the baseline hazard and the cumulative baseline hazard by the corresponding expressions of the Weibull model.

9.1.2. Full conditionals of the regularization parameters

Due to the stage of the hierarchical model structure, there is no direct connection between the regularization parameters $\boldsymbol{\tau}_\alpha^2$, $\boldsymbol{\tau}_\beta^2$, $\boldsymbol{\rho}$ and the likelihood and, as a consequence, the full conditionals have a closed form to draw directly a new state of the MCMC chain by Gibbs sampling. The same holds, if the partial likelihood is used for inference. The full conditionals of the regularization parameters are derived as in Section 6.1.2 and we shortly summarize here only the results.

Bayesian ridge

Version (A): We have for the variance parameters the deterministic connection $\tau_{\beta_j}^2 = 1/2\lambda_j$ and the full conditionals of the shrinkage parameters are gamma distributions

$$\lambda_j \,|\, \cdot \sim \text{Gamma}\left(h_{1,\lambda} + \frac{1}{2}, h_{2,\lambda} + \beta_j^2 \right), \quad j = 1, \ldots p_x. \quad (9.17)$$

Version (B): We have for the variance parameters the deterministic connection $\tau_\beta^2 = 1/2\lambda$ and the full conditional of the shrinkage parameter is a gamma distribution

$$\lambda \,|\, \cdot \sim \text{Gamma}\left(h_{1,\lambda} + \frac{p_x}{2}, h_{2,\lambda} + \sum_{j=1}^{p_x} \beta_j^2 \right). \quad (9.18)$$

Bayesian lasso

The full conditional of the variance parameters $\tau^2_{\beta,j}$ are inverse Gaussian distributions

$$\frac{1}{\tau^2_{\beta_j}} \mid \cdot \sim \text{InvGauss}\left(\frac{\sqrt{\lambda^2}}{|\beta_j|}, \lambda^2 \right) \quad , j = 1,..,p_x \,, \tag{9.19}$$

and we have a gamma full conditional of the complexity parameter

$$\lambda^2 \mid \cdot \sim \text{Gamma}\left(h_{1,\lambda} + p_x, h_{2,\lambda} + \frac{1}{2}\sum_{j=1}^{p_x} \tau^2_{\beta_j} \right). \tag{9.20}$$

Bayesian NMIG

Under the Bayesian NMIG prior the variance parameter is the product of two variance components $\tau^2_{\beta_j} = I_j \psi_j^2$. The full conditionals of the covariate-specific binary indicator variables I_j have Bernoulli distributions

$$p(I_j \mid \cdot) = \left(1 - \frac{1}{1 + A_j/B_j} \right)^{\delta_{v_0}(I_j)} \left(\frac{1}{1 + A_j/B_j} \right)^{\delta_{v_1}(I_j)} \quad , j = 1,..,p_x \,, \tag{9.21}$$

with

$$\frac{A_j}{B_j} = \frac{1-\omega}{\omega} \frac{\sqrt{v_1}}{\sqrt{v_0}} \exp\left\{ \frac{(v_0 - v_1)}{v_0 v_1} \frac{\beta_j^2}{2\psi_j^2} \right\}$$

and the variance parameters have inverse gamma distributions

$$\psi_j^2 \mid \cdot \sim \text{InvGamma}\left(h_{1,\psi} + \frac{1}{2}, h_{2,\psi} + \frac{\beta_j^2}{2I_j} \right) \quad , j = 1,..,p_x \,. \tag{9.22}$$

The full conditional for the mixing parameter is a beta density

$$\omega \mid \cdot \sim \text{Beta}\left(h_{1,\omega} + n_1 ; h_{2,\omega} + n_0 \right) \tag{9.23}$$

with $n_0 := \#\{ j : I_j = v_0 \}$, $n_1 := \#\{ j : I_j = v_1 \}$.

Smoothing variances:

The full conditionals for the variance parameters $\tau^2_{\alpha_j}$ are (proper) inverse gamma distributions

$$\tau^2_{\alpha_j} \mid \cdot \sim \text{InvGamma}\left(h_{1,\tau_j} + \frac{\text{rank}(\mathbf{K}_j)}{2}, h_{2,\tau_j} + \frac{1}{2} \boldsymbol{\alpha}_j' \mathbf{K}_j \boldsymbol{\alpha}_j \right), \; j = 0,1,...,p_z \,. \tag{9.24}$$

9.2. Conditional posterior densities based on the partial likelihood

Based on the partial log-likelihood (7.7) with the extended predictor (7.2) the posterior has the general form

$$p\left(\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\gamma}, \boldsymbol{\tau}^2_\alpha, \boldsymbol{\tau}^2_\beta, \boldsymbol{\rho} \mid \mathfrak{D}\right) \propto \exp\left(\text{pl}(\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\gamma} \mid \mathfrak{D})\right) \prod_{j=1}^{p_z} p(\boldsymbol{\alpha}_j \mid \tau^2_{\alpha,j}) p(\tau^2_{\alpha,j}) \cdot p(\boldsymbol{\beta} \mid \boldsymbol{\tau}^2_\beta) p(\boldsymbol{\tau}^2_\beta \mid \boldsymbol{\rho}) p(\boldsymbol{\rho}) p(\boldsymbol{\gamma}) \,.$$

9.2.1. Full conditionals of the predictor components

Unregularized linear regression coefficients $\boldsymbol{\gamma}$

To construct the IWLS proposals of the regression coefficients, we proceed as in the previous Section 9.1.1, but we replace the log-likelihood $\mathrm{l}(\boldsymbol{\vartheta}\,|\,\mathfrak{D})$ in the expressions through the partial log-likelihood $\mathrm{pl}(\boldsymbol{\vartheta}\,|\,\mathfrak{D})$ as well as the score vector and Hessian matrix of the full likelihood matrix through corresponding derivatives of the partial likelihood. For the unpenalized linear effects $\boldsymbol{\gamma}$ with Gaussian prior distribution (8.1), $\boldsymbol{\gamma}\,|\,\boldsymbol{\mu}_\gamma,\boldsymbol{\Sigma}_\gamma \sim \mathrm{N}(\boldsymbol{\mu}_\gamma,\boldsymbol{\Sigma}_\gamma)$, with $\boldsymbol{\mu}_\gamma \neq \mathbf{0}$ and $\boldsymbol{\Sigma}_\gamma^{-1} \neq \mathbf{0}$, the full conditional is given by

$$p(\boldsymbol{\gamma}\,|\,\cdot) \propto \exp\left\{\mathrm{pl}(\boldsymbol{\vartheta}\,|\,\mathfrak{D}) - \frac{1}{2}\boldsymbol{\gamma}'\boldsymbol{\Sigma}_\gamma^{-1}\boldsymbol{\gamma} + \boldsymbol{\gamma}'\boldsymbol{\Sigma}_\gamma^{-1}\boldsymbol{\mu}_\gamma\right\}.$$

In particular the score vector $\mathbf{s}_\gamma^{\mathrm{pl}}(\boldsymbol{\gamma}) = \partial \mathrm{pl}(\boldsymbol{\vartheta}\,|\,\mathfrak{D})/\partial\boldsymbol{\gamma}$ and the Hessian matrix $\mathbf{H}_\gamma^{\mathrm{pl}}(\boldsymbol{\gamma}) = \partial^2 \mathrm{pl}(\boldsymbol{\vartheta}\,|\,\mathfrak{D})/\partial\boldsymbol{\gamma}\partial\boldsymbol{\gamma}'$ of the logarithm of the partial likelihood are represented by

$$\mathbf{s}_\gamma^{\mathrm{pl}}(\boldsymbol{\gamma}) = \left(\sum_{i=1}^{n} d_i \left[u_{ij} - \frac{\sum_{k=1}^{n} 1_{(\tilde{t}_k \geq \tilde{t}_i)} \exp(\eta_k) u_{kj}}{\sum_{k=1}^{n} 1_{(\tilde{t}_k \geq \tilde{t}_i)} \exp(\eta_k)}\right]\right)_{1 \leq j \leq p_u}$$

and

$$\mathbf{H}_\gamma^{\mathrm{pl}}(\boldsymbol{\gamma}) = \left(-\sum_{i=1}^{n} d_i \left[\frac{\sum\limits_{k=1}^{n} 1_{(\tilde{t}_k \geq \tilde{t}_i)} \exp(\eta_k) u_{kj} u_{km}}{\sum\limits_{k=1}^{n} 1_{(\tilde{t}_k \geq \tilde{t}_i)} \exp(\eta_k)} - \frac{\sum\limits_{k=1}^{n} 1_{(\tilde{t}_k \geq \tilde{t}_i)} \exp(\eta_k) u_{kj} \cdot \sum\limits_{k=1}^{n} 1_{(\tilde{t}_k \geq \tilde{t}_i)} \exp(\eta_k) u_{km}}{\sum\limits_{k=1}^{n} 1_{(\tilde{t}_k \geq \tilde{t}_i)} \exp(\eta_k) \cdot \sum\limits_{k=1}^{n} 1_{(\tilde{t}_k \geq \tilde{t}_i)} \exp(\eta_k)}\right]\right)_{1 \leq j,m \leq p_u}.$$

The penalized score vector and the penalized Hessian matrix in the second order Taylor expansion of $f(\boldsymbol{\gamma}) = \log p(\boldsymbol{\gamma}\,|\,\cdot)$ are according to (9.3) and (9.4) written as

$$\mathbf{s}_\gamma(\boldsymbol{\gamma}) = \mathbf{s}_\gamma^{\mathrm{pl}}(\boldsymbol{\gamma}) - \boldsymbol{\Sigma}_\gamma^{-1}\boldsymbol{\gamma} + \boldsymbol{\Sigma}_\gamma^{-1}\boldsymbol{\mu}_\gamma, \quad \mathbf{H}_\gamma(\boldsymbol{\gamma}) = \mathbf{H}_\gamma^{\mathrm{pl}}(\boldsymbol{\gamma}) - \boldsymbol{\Sigma}_\gamma^{-1}.$$

Under a flat Gaussian prior we set $\boldsymbol{\mu}_\gamma = \mathbf{0}$ and $\boldsymbol{\Sigma}_\gamma^{-1} = \mathbf{0}$, and the resulting mean vector and covariance matrix of the multivariate Gaussian proposal distribution of regression coefficients $\boldsymbol{\gamma}$ are

$$\hat{\boldsymbol{\mu}}_\gamma^{(c)} = \hat{\boldsymbol{\Sigma}}_\gamma^{(c)}\left[\mathbf{s}_\gamma^{\mathrm{pl}}(\boldsymbol{\gamma}^{(c)}) - \mathbf{H}_\gamma^{\mathrm{pl}}(\boldsymbol{\gamma}^{(c)})\boldsymbol{\gamma}^{(c)}\right], \quad \hat{\boldsymbol{\Sigma}}_\gamma^{(c)} = \left(-\mathbf{H}_\gamma^{\mathrm{pl}}(\boldsymbol{\gamma}^{(c)})\right)^{-1}, \tag{9.25}$$

where $\mathbf{s}_\gamma^{\mathrm{pl}}(\boldsymbol{\gamma}^{(c)})$ and $\mathbf{H}_\gamma^{\mathrm{pl}}(\boldsymbol{\gamma}^{(c)})$ denote the score vector and the Hessian matrix of the partial log-likelihood evaluated at the current state $\boldsymbol{\gamma}^{(c)}$ and the actual states of the remaining model parameters $\boldsymbol{\vartheta}_{-\gamma} = (\boldsymbol{\alpha}^{(c)},\boldsymbol{\beta}^{(c)})$ in the predictor. To compute the mean vector $\hat{\boldsymbol{\mu}}_\gamma^{(p)}$ and covariance matrix $\hat{\boldsymbol{\Sigma}}_\gamma^{(p)}$ of the proposed new value $\boldsymbol{\gamma}^{(p)}$, which are required to compute the acceptance probability, the score vector $\mathbf{s}_\gamma^{\mathrm{pl}}(\boldsymbol{\gamma}^{(p)})$ and Hessian matrix $\mathbf{H}_\gamma^{\mathrm{pl}}(\boldsymbol{\gamma}^{(p)})$ are evaluated at the proposed state $\boldsymbol{\gamma}^{(p)}$ keeping the remaining model parameters of the predictor fixed at their actual states $\boldsymbol{\vartheta}_{-\gamma}$.

Regularized linear regression coefficients $\boldsymbol{\beta}$

Straightforward, the full conditional of the regularized linear effects $\boldsymbol{\beta}$ with the prior (8.2) reads

$$p(\boldsymbol{\beta}\,|\,\cdot) \propto \exp\left\{\mathrm{pl}(\boldsymbol{\vartheta}\,|\,\mathfrak{D}) - \frac{1}{2}\boldsymbol{\beta}'\boldsymbol{\Sigma}_\beta^{-1}\boldsymbol{\beta}\right\}$$

with $\boldsymbol{\Sigma}_\beta^{-1} = \mathrm{diag}(\boldsymbol{\tau}_\beta^{-2})$ and the corresponding multivariate Gaussian proposal distribution of the regularized effects has mean and covariance matrix

$$\hat{\boldsymbol{\mu}}_\beta^{(c)} = \hat{\boldsymbol{\Sigma}}_\beta^{(c)}\left[\mathbf{s}_\beta^{\mathrm{pl}}\left(\boldsymbol{\beta}^{(c)}\right) - \mathbf{H}_\beta^{\mathrm{pl}}\left(\boldsymbol{\beta}^{(c)}\right)\boldsymbol{\beta}^{(c)}\right], \quad \hat{\boldsymbol{\Sigma}}_\beta^{(c)} = \left(-\mathbf{H}_\beta^{\mathrm{pl}}\left(\boldsymbol{\beta}^{(c)}\right) + \boldsymbol{\Sigma}_\beta^{-1}\right)^{-1} \tag{9.26}$$

with score vector $\mathbf{s}_\beta^{\mathrm{pl}}(\boldsymbol{\beta}) = \partial \mathrm{pl}(\boldsymbol{\vartheta} \mid \mathfrak{D})/\partial\boldsymbol{\beta}$ and Hessian matrix $\mathbf{H}_\beta^{\mathrm{pl}}(\boldsymbol{\beta}) = \partial^2 \mathrm{pl}(\boldsymbol{\vartheta} \mid \mathfrak{D})/\partial\boldsymbol{\beta}\partial\boldsymbol{\beta}'$.

Regularized nonlinear regression coefficients $\boldsymbol{\alpha}_j$

Finally, using the prior (8.12) the full conditionals of the basis functions coefficients $\boldsymbol{\alpha}_j$ are

$$p\left(\boldsymbol{\alpha}_j \mid \cdot\right) \propto \exp\left\{\mathrm{pl}\left(\boldsymbol{\vartheta} \mid \mathfrak{D}\right) - \frac{1}{2\tau_{\alpha_j}^2}\boldsymbol{\alpha}_j'\mathbf{K}_j\boldsymbol{\alpha}_j\right\}, \quad j = 1,\dots,p_z\,,$$

with mean and covariance matrix of the multivariate Gaussian proposal distribution given by

$$\hat{\boldsymbol{\mu}}_{\alpha_j}^{(c)} = \hat{\boldsymbol{\Sigma}}_{\alpha_j}^{(c)}\left[\mathbf{s}_{\alpha_j}^{\mathrm{pl}}\left(\boldsymbol{\alpha}_j^{(c)}\right) - \mathbf{H}_{\alpha_j}^{\mathrm{pl}}\left(\boldsymbol{\alpha}_j^{(c)}\right)\boldsymbol{\alpha}_j^{(c)}\right], \quad \hat{\boldsymbol{\Sigma}}_\beta^{(c)} = \left(-\mathbf{H}_{\alpha_j}^{\mathrm{pl}}\left(\boldsymbol{\alpha}_j^{(c)}\right) + \frac{1}{\tau_{\alpha_j}^2}\mathbf{K}_j\right)^{-1}, \quad j = 1,\dots,p_z\,, \tag{9.27}$$

with $\mathbf{s}_{\alpha_j}^{\mathrm{pl}}(\boldsymbol{\alpha}_j) = \partial \mathrm{pl}(\boldsymbol{\vartheta} \mid \mathfrak{D})/\partial\boldsymbol{\alpha}_j$ and $\mathbf{H}_{\alpha_j}^{\mathrm{pl}}(\boldsymbol{\alpha}_j) = \partial^2 \mathrm{pl}(\boldsymbol{\vartheta} \mid \mathfrak{D})/\partial\boldsymbol{\alpha}_j\partial\boldsymbol{\alpha}_j'$.

9.2.2. Full conditionals of the regularization parameters

The rest of the model parameters $\boldsymbol{\tau}_\alpha^2 = (\tau_{\alpha_1}^2,\dots,\tau_{\alpha_{p_z}}^2)$, $\boldsymbol{\tau}_\beta^2$, $\boldsymbol{\rho}$ and are updated by Gibbs steps with the full conditionals as listed above in Subsection 9.1.2.

9.2.3. Computational details

Detail 1: The generic way to sample a new proposal from multivariate Gaussian proposals is based on the algorithms described in Rue (2001), and shortly sketched in Subsection 6.1.7.

Detail 2: The most costly computations in running the whole MCMC samplers are the inversions of the precision matrices within the IWLS parts of the corresponding parameter vectors. To reduce the running time in the case of high-dimensional parameters, we can update these parameters in blocks of smaller size than the size of the whole parameter vector. We use per default a maximal block size of 20 covariates per block. This option can be specified with the `blocksize` argument.

9.3. Update of the parameters

The Markov chain is generated via MCMC simulations based on drawing from the full conditionals of parameters or parameter blocks given the remaining parameters and the data as derived in the previous sections. The methods are implemented in the following software. The inferential procedures for fitting the parametric and nonparametric models based on the full likelihood with P-spline and Weibull baseline are implemented in the `regress` method of the free `BayesX` software available from http://www.stat.uni-muenchen.de/~bayesx/. The procedures based on the partial likelihood are implemented in the `R`-function `bcoxpl()` which will be provided from the author on request. The usage of both functions is described in the Appendix D.3 to D.4.

9.3.1. Preprocessing

Standardization: To ensure that comparable regression coefficient sizes imply comparable effect sizes, covariates are standardized in advance. This avoids the extensive covariate-specific tuning of the priors for different covariate scales. We standardize covariates with linear effects to zero empirical mean and unit empirical variance. To obtain that smooth covariates taking values in $[-1,1]$, we can apply the transformation

$$z_{ij}^{*} = \frac{2(z_{ij} - z_{j,\min})}{z_{j,\max} - z_{j,\min}} - 1 .$$

Starting values in the BayesX method regress: In `BayesX` the starting values for the regression coefficients $\boldsymbol{\alpha},\boldsymbol{\beta},\boldsymbol{\gamma}$ are computed via backfitting within Fisher scoring using the fixed variance parameters $\boldsymbol{\tau}_{\alpha}^{2(0)},\boldsymbol{\tau}_{\beta}^{2(0)}$ initially specified, and finally the resulting estimates are used as initial states $\boldsymbol{\alpha}^{(0)},\boldsymbol{\beta}^{(0)},\boldsymbol{\gamma}^{(0)}$ of the chain.

Starting values in the R-function bcoxpl(): An automatic computation of starting values is not implemented in the function `bxoxpl()` and in our simulations and applications we start with weakly specified models. If preprocessing is desired, the starting values for the regression coefficients $\boldsymbol{\alpha}^{(0)},\boldsymbol{\beta}^{(0)},\boldsymbol{\gamma}^{(0)}$ can optionally be computed with the R function `coxph{survival}` using e. g. the `ridge{survival}` and `pspline{survival}` terms in the formula with fixed penalty parameter.

9.3.2. Pseudocode

[1] *Initialization*:

Specify the regularization priors of the linear effects: Set the values of the hyperparameters $h_{1,\lambda}, h_{2,\lambda}$ to specify the gamma prior for the shrinkage parameter $\lambda(\lambda^2)$ in the Bayesian ridge or lasso prior. For the Bayesian NMIG prior set the values v_0, v_1 of the indicator I_j, set the values of the hyperparameters $h_{1,\psi}, h_{2,\psi}$ of the inverse gamma prior for the variance parameter ψ_j^2 and set the hyperparameters $h_{1,\omega}, h_{2,\omega}$ of the beta prior for the complexity parameter ω.

Specify the non-linear effects: Set number g_j of B-spline basis functions and choose the order d_j of the random walk penalty for basis function weights.

Standardize the covariates according to Subsection 9.3.1.

Set the number C of iterations, set $c = 0$ and repeat the following steps until $c < C$.

[2] *Update of the unregularized regression coefficients*:

Draw a new value $\boldsymbol{\gamma}^{(p)}$ from the Gaussian proposal distribution $\varphi(\cdot \mid \hat{\boldsymbol{\mu}}_{\gamma}^{(p)}, \hat{\boldsymbol{\Sigma}}_{\gamma}^{(p)})$ with mean vector and covariance matrix given in (9.9). Accept the proposed state as new state of the chain with acceptance probability

$$p_{\text{accept}}\left(\boldsymbol{\gamma}^{(p)}, \boldsymbol{\gamma}^{(c)}\right) = \min\left\{1, \frac{p\left(\boldsymbol{\gamma}^{(p)} \mid \cdot\right) \varphi\left(\boldsymbol{\gamma}^{(c)} \mid \hat{\boldsymbol{\mu}}_{\gamma}^{(p)}, \hat{\boldsymbol{\Sigma}}_{\gamma}^{(p)}\right)}{p\left(\boldsymbol{\gamma}^{(c)} \mid \cdot\right) \varphi\left(\boldsymbol{\gamma}^{(p)} \mid \hat{\boldsymbol{\mu}}_{\gamma}^{(c)}, \hat{\boldsymbol{\Sigma}}_{\gamma}^{(c)}\right)}\right\}.$$

If the proposal is accepted, set $\boldsymbol{\gamma}^{(c+1)} = \boldsymbol{\gamma}^{(p)}$, else set $\boldsymbol{\gamma}^{(c+1)} = \boldsymbol{\gamma}^{(c)}$.

[3] *Update of the regularized regression coefficients*:

Draw a new value $\boldsymbol{\beta}^{(p)}$ from the Gaussian proposal distribution $\varphi(\cdot \mid \hat{\boldsymbol{\mu}}_{\beta}^{(p)}, \hat{\boldsymbol{\Sigma}}_{\beta}^{(p)})$ with mean vector and covariance matrix given in (9.11). Accept the proposed state as new state of the chain with acceptance probability

$$p_{\text{accept}}\left(\boldsymbol{\beta}^{(p)},\boldsymbol{\beta}^{(c)}\right)=\min\left\{1,\frac{p\left(\boldsymbol{\beta}^{(p)}\mid\cdot\right)\varphi\left(\boldsymbol{\beta}^{(c)}\mid\hat{\boldsymbol{\mu}}_{\beta}^{(p)},\hat{\boldsymbol{\Sigma}}_{\beta}^{(p)}\right)}{p\left(\boldsymbol{\beta}^{(c)}\mid\cdot\right)\varphi\left(\boldsymbol{\beta}^{(p)}\mid\hat{\boldsymbol{\mu}}_{\beta}^{(c)},\hat{\boldsymbol{\Sigma}}_{\beta}^{(c)}\right)}\right\}.$$

If the proposal is accepted, set $\boldsymbol{\beta}^{(c+1)}=\boldsymbol{\beta}^{(p)}$, else set $\boldsymbol{\beta}^{(c+1)}=\boldsymbol{\beta}^{(c)}$.

[4] *Update of the shrinkage- and selection-prior components*:

Bayesian ridge (A): Draw a new value of the complexity parameter $\lambda_j^{(c+1)}$ from the conditional gamma distribution given by (9.17) and set the variance parameter $\tau_{\beta_j}^{2,(c+1)}=1/2\lambda_j^{(c+1)}$, $j=1,...,p_x$.

Bayesian ridge (B): Draw a new value of the complexity parameter $\lambda^{(c+1)}$ from the conditional gamma distribution given in (9.18) and set the variance parameter $\tau_{\beta}^{2,(c+1)}=1/2\lambda^{(c+1)}$.

Bayesian lasso: Draw a new value of the variance parameter $\tau_{\beta_j}^{2,(c+1)}$, $j=1,...,p_x$, from the conditional inverse Gaussian distribution given in (9.19). Draw a new value of the complexity parameter $\lambda^{2,(c+1)}$ from the conditional gamma distribution given in (9.20).

Bayesian NMIG: Draw a new value of the indicator $I_{\beta_j}^{(c+1)}$, $j=1,...,p_x$, from the conditional Bernoulli distribution given in (9.21). Draw a new value of the variance parameter $\psi_{\beta_j}^{2,(c+1)}$, $j=1,...,p_x$, from the conditional inverse gamma distribution given in (9.22). Draw a new value of the complexity parameter $\omega^{(c+1)}$ from the conditional beta distribution given in (9.23).

[5] *Update of the regularized spline coefficients*:

Draw a new value $\boldsymbol{\alpha}_j^{(p)}=(\alpha_{j,1}^{(p)},...,\alpha_{j,g_j}^{(p)})'$, $j=0,1,...,p_z$ from the Gaussian proposal distribution $q(\cdot\mid\hat{\boldsymbol{\mu}}_{\alpha,j}^{(p)},\hat{\boldsymbol{\Sigma}}_{\alpha,j}^{(p)})$ with mean vector and covariance matrix given by (9.13). Accept the proposed state as new state of the chain with acceptance probability

$$\text{accept}\left(\boldsymbol{\alpha}_j^{(p)},\boldsymbol{\alpha}_j^{(c)}\right)=\min\left\{1,\frac{p\left(\boldsymbol{\alpha}_j^{(p)}\mid\cdot\right)\cdot\varphi\left(\boldsymbol{\alpha}_j^{(c)}\mid\hat{\boldsymbol{\mu}}_{\alpha,j}^{(p)},\hat{\boldsymbol{\Sigma}}_{\alpha,j}^{(p)}\right)}{p\left(\boldsymbol{\alpha}_j^{(c)}\mid\cdot\right)\cdot\varphi\left(\boldsymbol{\alpha}_j^{(p)}\mid\hat{\boldsymbol{\mu}}_{\alpha,j}^{(c)},\hat{\boldsymbol{\Sigma}}_{\alpha,j}^{(c)}\right)}\right\}.$$

If the proposal is accepted, set $\boldsymbol{\alpha}_j^{(c+1)}=\boldsymbol{\alpha}_j^{(p)}$, else set $\boldsymbol{\alpha}_j^{(c+1)}=\boldsymbol{\alpha}_j^{(c)}$. To center the functions compute the mean of function evaluations at the observed data points $c_j^{(c+1)}=n^{-1}\sum_{i=1}^{n}\alpha_{j,k}^{(c+1)}B_{j,k}(z_{ij})$. Adjust the current states of $\boldsymbol{\alpha}_j^{(c+1)}$ by $\boldsymbol{\alpha}_j^{(c+1)}-c_j^{(c+1)}$ and adjust the intercept $\gamma^{(c+1)}$ by $\gamma^{(c+1)}+c_1^{(c+1)}+...+c_{p_z}^{(c+1)}$.

[6] *Update of the smoothing variances*:

Draw a new value of the variance parameters $\tau_{\alpha_j}^{2,(c+1)}$, $j=0,1,...,p_z$ from the conditional inverse gamma distribution given by (9.24).

Modifications for the Weibull model

For the Weibull model we replace in step [5] the update of the log-baseline hazard coefficients $\boldsymbol{\alpha}_0$ by the update of the shape parameter α_0. We use the proposal distribution given in (9.16) and the full conditional from (9.15), compare Section 9.1.1. The update of $\tau_{\alpha_0}^2$ is dropped out in step [6].

Modifications for the partial likelihood

Using the partial likelihood for inference we have to exchange the mean vector and covariance matrix expressions in steps [2], [3] and [5] based on the full likelihood by those based on the partial likelihood, (9.25), (9.26) and (9.27) from Section 9.2.1. The updates of $\boldsymbol{\alpha}_0$ and $\tau^2_{\alpha_0}$ are skipped.

PART III. SIMULATIONS

10. AFT-type models

In this section we investigate the performance of the Bayesian regularization priors in the extended accelerated failure time model (AFT) as described in Section 2. At first, in Subsection 10.1, we consider the regularization in the AFT model concerning the smooth estimation of the error distribution density modeled as penalized Gaussian mixture (PGM). In particular the different variants for the update of the mixture weights, presented in Section 6.1.3, are explored (neglecting initially the effect of covariates) and compared with selected competing Bayesian and frequentist procedures available in the `R` software. In the subsequent sections covariates are also included in the simulation models. We regard the Bayesian regularization priors for the linear effects, as described in Sections 4.1 to 4.3, in the low-dimensional case $p_x < n$, where the number of covariates $p_x \in \mathbb{N}$ does not exceed the number of observations $n \in \mathbb{N}$. In particular in Subsection 10.2 the number of covariates is fixed to $p_x = 25$ and the observations vary from $n = 100$ to $n = 500$. Further, in Subsection 10.3, the linear covariates are modeled by Bayesian P-splines which induce a high-dimensional predictor. Finally, the high-dimensional case with respect to the number of covariates is considered by increasing step by step the number of covariates until it exceeds the number of observations $p_x > n$. We focus here mainly on the impact under the Bayesian NMIG prior.

While the continuous regularization priors for the linear effects cause the shrinkage of these effects toward zero, the used MCMC estimation methods do not directly enforce simultaneous variable selection, like, e. g., the algorithms of the frequentist lasso do. To build a final model with a subset of the available covariates, we use the heuristic selection criteria based on the 95% credible interval and the one standard deviation interval as described in Section 4.4. In particular the Bayesian NMIG regularization prior provides the additional opportunity to access the importance of the linearly modeled features by utilizing the posterior relative frequencies of the two indicator variable values v_0 and v_1. We investigate the reliability of these procedures to identify important features and compare the performance with those from frequentist feature selection based on the AIC criterion under Gaussian error assumption. In the various situations we focus further on the question, which constellation of p_x versus n enables reasonable estimates of the parameters, since the number of model parameters (including the latent ones) is comparatively high in the Bayesian AFT model with PGM error and extended predictor.

Functions and methods

The Bayesian algorithms to estimate the extended AFT model are implemented in the `R`-function `baftpgm()`, which is available from the author by request. In Appendix D.5 we describe the usage of this function. As Bayesian competitor for the extended AFT model we use the `R`-function `bayessurvreg2()` of the package `{bayesSurv}` by A. Komárek, where the baseline error distribution is also modeled as PGM. Besides various censoring schemes, like, e. g., right or interval censoring, this function supports estimation of unpenalized linear effects and random effects in the predictor. As frequentist competitor, in particular in the case without covariates and censoring in the data, we utilize the function `pendensity()` as implemented in the correspondent `R`-package `{pendensity}` by Schellhase and Kauermann (2011). The authors provide a penalized basis function approach with B-spline or Gaussian basis functions to approximate the baseline error density. Frequentist maximum likelihood estimation of the AFT model with parametric error distributions is carried out with the R-function `survreg()` of the `R`-package `{survival}`. Variable selection is practiced by forward-backward-stepwise procedures based on the AIC (Akaike-Information-Criterion) criterion and accomplished by the R-function `step()`. Also ridge regularization of linear effects and combined estimation of nonlinear effects is possible within the function `survreg()`, but the values of the shrinkage resp. smoothing parameters need to be predefined.

Estimation accuracy

In the simulation studies the mean squared error (MSE) of an estimate is used as performance criterion to measure the estimation accuracy within each dataset of $R \in \mathbb{N}$ replications. For example the MSE of the estimated regularized linear effects $\boldsymbol{\beta}$ in the r-th replication, $r = 1,...,R$, is given by

$$\mathrm{MSE}_{(r)}(\hat{\boldsymbol{\beta}}) = \frac{1}{n}(\hat{\boldsymbol{\beta}}^{(r)} - \boldsymbol{\beta})'\mathbf{X}^{(r)\prime}\mathbf{X}^{(r)}(\hat{\boldsymbol{\beta}}^{(r)} - \boldsymbol{\beta}) ,$$

where $n \in \mathbb{N}$ is the common number of observations in each simulation setting, $\hat{\boldsymbol{\beta}}^{(r)}$ is the vector of estimated regression coefficients and $\mathbf{X}^{(r)}$ denotes the associated design matrix of the regularized predictor component in r-th replication. For P-spline-based nonlinear effects $f_j(\cdot)$ of covariates z_j, with observations $z_{i,j}^{(r)}$, $i = 1,...,n$, in the r-th replication, we have

$$\mathrm{MSE}_{(r)}(\hat{f}_j) = \frac{1}{n}(\hat{\mathbf{f}}_j^{(r)} - \mathbf{f}_j)'(\hat{\mathbf{f}}_j^{(r)} - \mathbf{f}_j) , \tag{10.1}$$

where $\hat{\mathbf{f}}_j^{(r)} = (\hat{f}_j^{(r)}(z_{1,j}^{(r)}),...,\hat{f}_j^{(r)}(z_{n,j}^{(r)}))'$ denotes the vector of function evaluations of the estimator $\hat{f}_j^{(r)}(z)$ in the r-th replication and $\mathbf{f}_j = (f_j(z_{1,j}^{(r)}),...,f_j(z_{n,j}^{(r)}))'$ is the corresponding vector of the “true” underlying nonlinear effect $f_j(\cdot)$. The computation of the MSEs of function estimates representing the baseline quantities, like the logarithm of the baseline hazard function $f_0(t) = \log(\lambda_0(t)) + \gamma_0$ in the CRR model or the distribution density $f_{Y_0}(\cdot)$ of the baseline error $Y_0 = \gamma_0 + \sigma\varepsilon$ in the AFT model, is straightforward in terms of (10.1). In the CRR model the baseline hazard function $\lambda_0(\cdot)$ and the associated estimate are evaluated at the observed survival times $\tilde{t}_i^{(r)}$, $i = 1,...,n$, of each replication and the baseline error density $f_{Y_0}(\cdot)$ and the associated estimate in the AFT model are evaluated on a predefined number of equidistant grid points $(e_1,....,e_k)$ that cover uniformly the margins of the “true” underlying density $f_{Y_0}(\cdot)$.

In this work the Bayesian point estimates $\hat{\boldsymbol{\theta}}$ of the model parameters $\boldsymbol{\theta}$ are generally based on the mean of the marginal posterior distribution, approximated by the component specific empirical mean of the generated MCMC sample $\boldsymbol{\theta}^{(s)}$, $s = 1,...,S$. Further summary statistics of the parameter specific marginal posterior distributions like the median, standard deviation or quantiles are also approximated by their empirical counterparts. In particular function estimates are given as the mean of the sample function evaluations $f_j^{(s)}(\cdot)$ at each of the considered grid points. For nonlinear model components $f_j(\cdot)$ formed by P-spline basis functions $\mathbf{b}_j'(z)$ this results in $\hat{f}_j(\cdot) = \mathbf{b}_j'(\cdot)\hat{\boldsymbol{\alpha}}_j^{(r)}$, where $\hat{\boldsymbol{\alpha}}_j$ is the mean vector of the sampled basis function weights $\boldsymbol{\alpha}_j^{(s)} \subset \boldsymbol{\theta}$. With respect to the identifiability of the predictor components, function estimates are horizontally centered around zero. The baseline error density $f_{Y_0}(\cdot \mid \boldsymbol{\alpha}_0, \gamma_0, \sigma)$ is computed for all iterations of the MCMC sampler using the current sampled states of the associated parameters $\boldsymbol{\alpha}_0^{(s)} \subset \boldsymbol{\theta}$, $\gamma_0^{(s)} \in \boldsymbol{\theta}$, $\sigma^{(s)} \in \boldsymbol{\theta}$. Due to the non-identifiability of the location and the scale parameter, the resulting density estimate of the error ε is standardized - during or after the iterations - to achieve zero mean and unit variance, inducing a simultaneous adjustment of the location and scale parameter $\gamma_0^{(s)}$ and $\sigma^{(s)}$ as described in Section 6.2.1. Finally, the estimate of the baseline error density $\hat{f}_{Y_0}(\cdot)$ is computed as the average of the function evaluations $f_{Y_0}^{(s)}(\cdot) = f_{Y_0}(\cdot \mid \boldsymbol{\alpha}_0^{(s)}, \gamma_0^{(s)}, \sigma^{(s)})$ at the grid points.

To visually compare the performance of the different methods, the MSEs of the interesting parameters in the replications are summarized utilizing box plots. In the low-dimensional cases, we additionally report the average number of correctly and incorrectly classified zero and nonzero regression coefficients of the final models obtained after applying one of the hard shrinkage selection rules and compare them for the different shrinkage priors. The used abbreviations to denote various models and the different update schemes of the error weights are summarized in the Reference Section.

10.1. Baseline error density estimation

Error models

In the simulation studies we use the following four target baseline error distributions (BED) to assess the performance of the PGM approach for the error density in the log-linear version of the AFT model:

- ***BED 1***: $Y_0 \sim \text{Gumbel}(\mu = 3, \sigma = 1.5)$
- ***BED 2***: $Y_0 \sim 0.75 \cdot N(\mu = -3, \sigma^2 = 1) + 0.25 \cdot N(\mu = 2, \sigma^2 = 1)$
- ***BED 3***: $Y_0 \sim 0.4 \cdot N(\mu = -3, \sigma^2 = 1) + 0.6 \cdot N(\mu = 0, \sigma^2 = 3.5)$
- ***BED 4***: $Y_0 \sim 0.5 \cdot N(\mu = 0, \sigma^2 = 1) + 0.5 \cdot N(\mu = 0, \sigma^2 = 3.5)$

Figure 10.1 displays the densities of the four baseline error models. The first baseline error model BED 1 uses the Gumbel (maximum extreme value) distribution with cumulative distribution function $F_{Y_0}(y) = \exp(-\exp(-[(y-\mu)/\sigma]))$, where $\mu \in \mathbb{R}$ and $\sigma > 0$ denote the location and scale parameter. The mean and variance of the Gumbel distribution is given by $\mathbb{E}(Y_0) = \mu - \sigma\gamma_E$ and $\mathbb{V}\text{ar}(Y_0) = \pi^2\sigma^2/6$, with γ_E as the Euler constant ($\gamma_E \approx 0.577$). With the parameters specified in error model BED 1 we obtain $\mathbb{E}(Y_0) \approx -2.124$ and $\mathbb{V}\text{ar}(Y_0) \approx 3.701$. The remaining baseline error models are represented as mixtures of two Gaussian distributions with mean and variance given by $\mathbb{E}(Y_0) = -1.75$, $\mathbb{V}\text{ar}(Y_0) = 5.6875$ (BED 2), $\mathbb{E}(Y_0) = -1.2$, $\mathbb{V}\text{ar}(Y_0) = 9.91$ (BED 3) and $\mathbb{E}(Y_0) = 0$, $\mathbb{V}\text{ar}(Y_0) = 6.625$ (BED 4).

Data generation

For each error model we generate $R = 50$ replicated simulation datasets with $n = 500$ observations, on the one hand with 0% and on the other hand with 25% censored observations. In particular, the log-transformed survival times $y_i = \log(t_i)$, $i = 1, \ldots, n$, are generated by drawing i.i.d. random numbers $y_{0,i}$ from the specific target baseline error distribution BED 1 to BED 4, i. e. $y_{0,i} \sim_{\text{iid}} \text{BED}\,\ell$,

$\ell \in \{1,...,4\}$. To explore the performance of the several update schemes for the transformed error weights, no covariates are considered here, so the log-survival times for baseline error model $\text{BED}\,\ell$ are directly given as

$$y_i = y_{0,i}, \quad y_{0,i} \sim_{iid} \text{BED}\,\ell, \quad i = 1,...,n .$$

Censoring times c_i are generated as draws from uniform distributions $C_i \sim_{iid} \text{Uni}[q_{\text{BED}\ell,0.001}, q_{\text{BED}\ell,0.999}]$, where $q_{\text{BED}\ell,0.001}$ and $q_{\text{BED}\ell,0.999}$ denote the 0.001- and 0.999-quantile of the respective target baseline error distribution. After the first run the resulting observed survival times are given as $\tilde{y}_i = \min(y_i, c_i)$. To achieve the desired percentage of censored observations, we generate in additional runs censoring times for the uncensored observations of the previous run until the percentage of right censored observations fits.

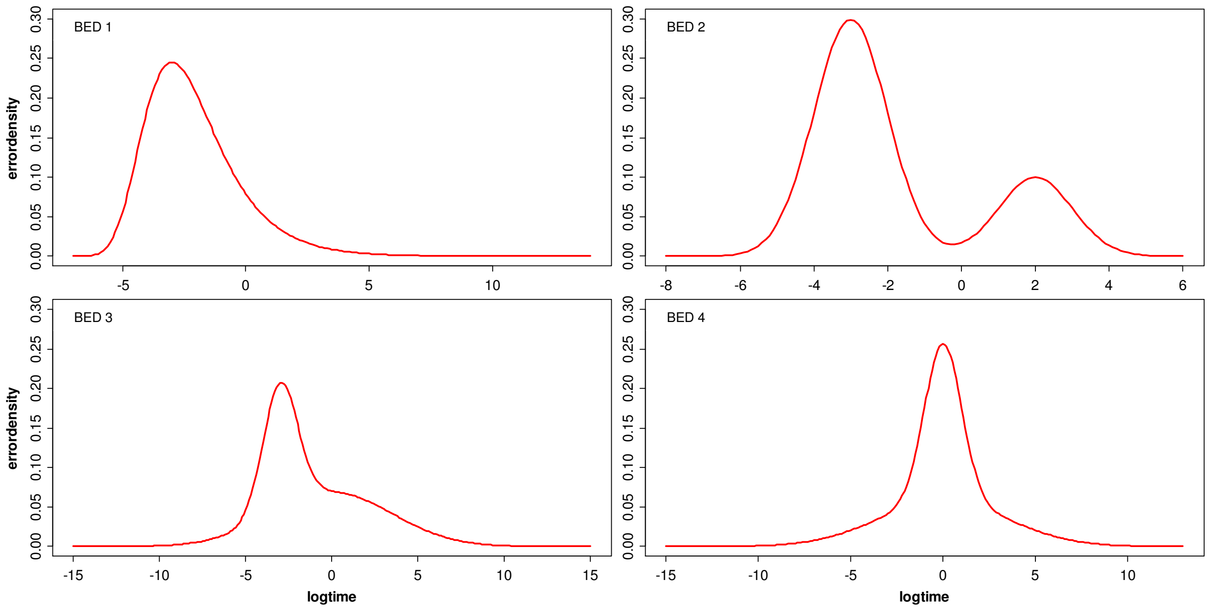


Figure 10.1: Densities of the four baseline error distributions BED 1 (upper left panel), BED 2 (upper right panel), BED 3 (lower left panel) and BED 4 (lower right panel).

Function and parameter specification

Methods: For the Bayesian estimation of the error density we use the function `bayessurvreg2()` and `baftpgm()`. In both functions the PGM error is specified through $g_0 = 21$ equidistant knots m_j placed in the interval $[-4.5, 4.5]$, i. e., $m_1 = -4.5, ..., m_{21} = 4.5$ with distance 0.45. The variance of the Gaussian basis functions is uniformly set to $s_j^2 = 0.25^2$, $j = 1,...,g_0$, and we use the third-order random walk prior to control the smoothness of the estimate. In particular for the *"dirichlet"* update scheme of the error weights we select a reduced number of $g_0 = 7$ equidistant knots in the interval $[-4.5, 4.5]$ with basis variances $s_j^2 = 0.35^2$. Within the function `bayessurvreg2()` we use the slice sampler as default update scheme for the error weights. In the function `baftpgm()` we utilize the option `scalebasis` to specify the standardization of error density within the loops of the sampler, compare Section 6.2.1. The standardization within the sampler (`scalebasis=TRUE`) causes a varying positioning of the knots m_j while the knots are fixed if the standardization is suppressed (`scalebasis=FALSE`). In the annotation of the figures the method names assigned with the suffix *"FK"* indicate the fixed knots, i. e. suppressed standardization. For some methods we vary the update order of the error weights. The suffixes *"R0"*, *"R1"* and *"R2"* indicate the specifications `order.alpha="fix2"`, `"order.alpha=random1"` and `" order.alpha=random2"`.

The function `pendensity()` is applied to the uncensored data without covariates to get a non Bayesian flexible estimate of the baseline error density. We specify here the optional Gaussian basis with 21 knots and with the third-order differences penalty for the basis function weights. The estimates with the `survreg()` procedure are carried out utilizing the extreme value, Gaussian and a logistic error distributions.

Hyperparameters: In general the prior hyperparameters of the scale parameter σ^2 are set to $h_{1,\sigma} = h_{2,\sigma} = 0.01$ and those of the smoothing variance to $h_{1,\tau_0} = 1$, $h_{2,\tau_0} = 0.01$. For the update schemes *"mhcond"* and *"mhmarg"* these basic values are sometimes modified in order to increase the smoothing (justification follows below). We specify a zero mean, diffuse Gaussian prior with variance 100^2 for the fixed effects - especially for the location parameter γ_0 in the use of the function `bayessurvreg2()`.

Starting values: We pass on pre-estimation of the model parameters to find appropriate initial values of the chain in regions close to the parameter estimates. For the transformed error weights $\alpha_{0,j}$, $j = 1,...,21$, with exception of the middle weight $\alpha_{0,11} := 0$, each starting value is set to $\alpha_{0,j}^{(0)} = 0.01$ resulting in a flat error density in the range $[-4.5, 4.5]$. The location and scale parameter start in $\gamma_0^{(0)} = 1$, $\sigma^{2(0)} = 1$ and the smoothing variance is initially set to $\tau_{\alpha_0}^{2(0)} = 0.01$. The component labels $r_i^{(0)}$ are randomly assigned to one of the g_0 error basis densities.

Estimation: For the MCMC algorithms we use 30000 iterations, where the first 15000 iterations are discarded as burnin of the Markov chain and the remaining iterations are thinned using a step width of 15. The resulting 1000 states of the chain build the sample of the posterior distribution and the empirical basis to compute the estimates. The simulations ran on various PCs and Servers with different specifications. For this reason and due to the variety of the update schemes we present in the following only the range of the observed running times for orientation. In general, shortest running times are obtained with the *"dirichlet"* and the longest under the *"mhmag"* update scheme. In the simulations of this section we observed 7–20 min (pro replication) in the data without censoring and 10–23 min in the data with 25% censoring, in particular under the under the *"mhmag"* update scheme we have about 60 min. In the following, the main results of the simulations are summarized and presented.

Results

MSE of the baseline error density

Figure 10.2 shows the MSEs of the estimated error densities, $\text{MSE}(\hat{f}_{Y_0})$, for the two baseline error models BED 2 and BED 3 resulting under the different single and block update schemes for the error weights as described in Section 6.1.3. The upper panel contains the results from the data with no censoring and the lower panel the corresponding results with 25% censored observations in the data. The MSE results from the frequentist *"survreg"* procedure are omitted due to the poor performance (particularly with a Gaussian error the MSEs exceed always the value of $4e^{-05}$).

Apart from some exceptional cases with comparably poor performance (*"mhmarg"* and *"mhcond"* in BED 2 with 25 % censoring), none of the Bayesian update schemes of the transformed error weights appear to be uniformly superior across the error models. Compared to the simulations with the uncensored data, the censoring increases the level of the MSE across all applied methods. But, with exception of the update scheme *"mhcond"* in BED2, the MSE pattern given by the boxes does not clearly vary.

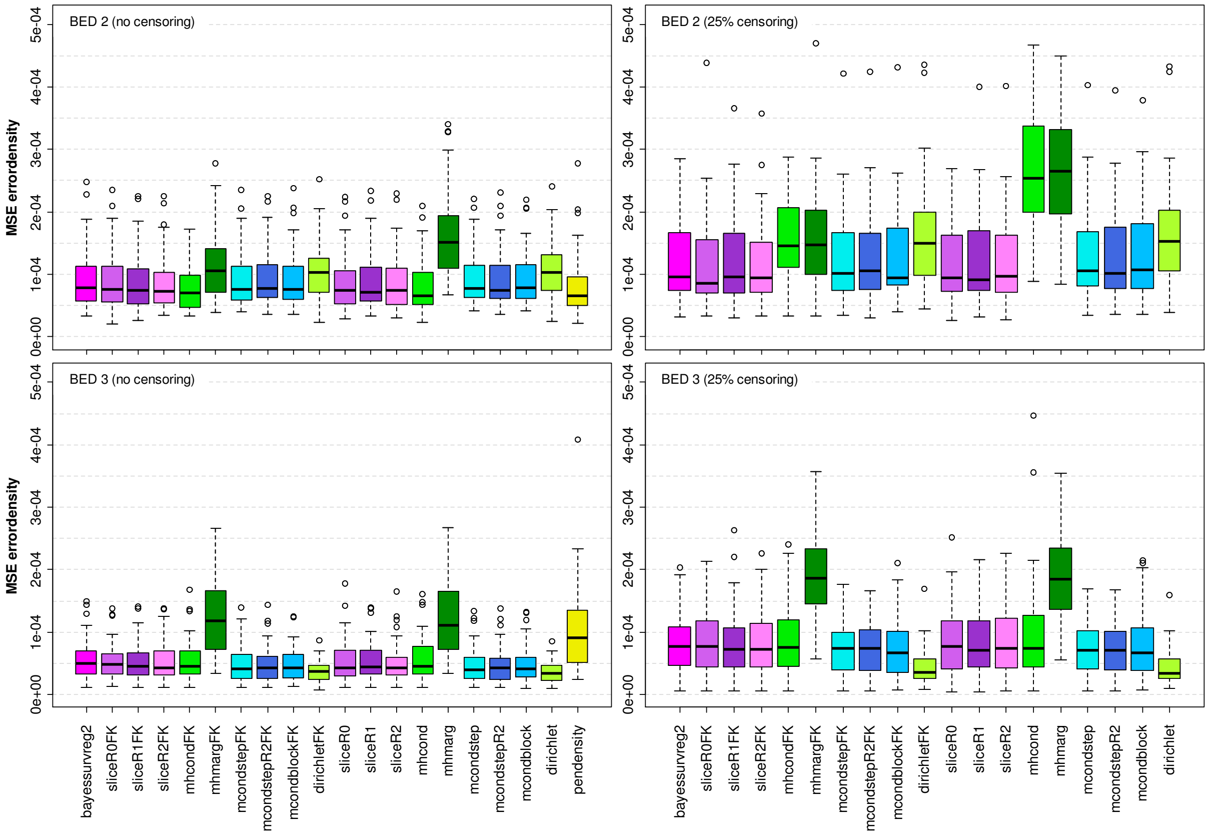


Figure 10.2: Mean squared errors of the estimated baseline error density, $\text{MSE}(\hat{f}_{Y_0})$, in the AFT model with baseline error distribution BED 2 (upper panel) and BED 3 (lower panel), without censoring (left panel) and under 25% censoring (right panel) in the simulation data.

Focusing on the update schemes *"slice"* and *"mcondstep"* under varying update order (*"R0"*, *"R1"*, *"R2"*) of the transformed error weights, we found also no systematic benefit in the estimates and the results in terms of the MSE are almost comparable. In particular the *"dirichlet"* update scheme, where no penalty controls the smoothness of the error density, performs surprisingly well in two of the four error distribution models. Especially in the settings BED 3 and BED 4 (not shown), it achieves the best performance within the uncensored and censored data. In the first two error settings, BED 1 (not shown) and BED 2, we observe conversely an increased MSE compared to the other methods. The frequentist competitor *"pendensity"*, only used in the simulations with the uncensored data, performs best in the estimation of the bimodal distribution BED 2.

As mentioned before, we modify the tuning of smoothing variance prior for the update schemes *"mhmarg"* and *"mhcond"* in some of the simulations. The outstanding comparatively poor performance with the update scheme *"mhmarg"* and the update scheme *"mhcond"* in error setting BED 2 with the censored observations is explained by the induced stronger smoothing of the error density. In particular under error model BED 2 with censoring, the regularization is further increased for the cross over from the fixed to the flexible knot option. We observed that, if the smoothing penalty is too weak, the weights of the basis densities at the right and/or left border get close to zero. As a consequence, the number of occupied classes of the mixture density decreases during the sampling, so that finally the component labels r_i are assigned to a few of the g_0 error basis densities. This also affects the acceptance rates that decrease too. Further, the option to standardize the error density during the sampler boosts this effect. Standardization shifts the border knots with close to zero weights towards $\pm\infty$ (i. e. we get a large distance between the knots) and the number of occupied classes of the mixture decreases further and the component labels r_i are assigned to at least one or two of the g_0 error basis densities. As consequence of the small border weights and the optional standardization, the associated sum of squared differences of the transformed error weights increases and we counterbalance with a stronger penalization through the smoothing variance. We observed the described effects in general for all update methods, if the smoothing regularization is too weak, but in particular the Metropolis-Hastings update schemes required an enhanced regularization. For both update schemes *"mhmarg"* and *"mhcond"* we adapt the first hyperparameter h_{1,τ_0} of the inverse gamma prior of the smoothing variance $\tau^2_{\alpha_0}$ to varying values in the range of 5 to 15. Further the acceptance rates of the transformed error weights profit from the stronger regularization. The values are consistently close to 80% for both update schemes under comparable regularization across the error models. Under weaker regularization, used e. g. in the error models BED 3 and BED 4, the acceptance rates of the *"mhcond"* scheme are generally smaller with values in the range of 50% to 20%, and censoring decreases the acceptance rates further.

We observed also, that the acceptance rates of the transformed error weights in the Metropolis update schemes are very different. For the update scheme *"mcondblock"* we always obtain very low values around 5%, but the few accepted new states are uniformly distributed over the sample. Since the model parameters converge and show also a good mixing (except the transformed error weights), we utilize this scheme without further adaption of the smoothing hyperparameters to consider the impact of these low acceptance rates. In contrast, the acceptance rates of the update scheme *"mcondstep"*, with the iteratively updated transformed error weights, are in general relatively high with values around 70%.

Penalty of the transformed error weights

The induced stronger regularization, if applied, is reflected in the log-penalty term $-\tau^2_{\alpha_0}\boldsymbol{\alpha}_0'\mathbf{K}_0\boldsymbol{\alpha}_0$. **Figure 10.3** shows the resulting log-error penalties for the error model BED 2 and BED 3 under 25 % censoring in the data. Due to the increased regularization under the update scheme *"mhmarg"* we observe very small values for the sum of (third order squared) differences $\boldsymbol{\alpha}_0'\mathbf{K}_0\boldsymbol{\alpha}_0$, compare **Figure 10.4**, and values of the smoothing variance $\tau^2_{\alpha_0}$ close to zero (not shown) are leading in summary to the smaller log-penalty values, compared to the other update schemes. The basic regularization ($h_{1,\tau_0} = 1$) under the update scheme *"mhcond"*, as in the error models BED3 and BED 4, comes along with increased values for the sum of differences. But the larger value of the associated smoothing variance $\tau^2_{\alpha_0}$ causes at last that the penalty has the same range as, e. g., the single update schemes. The same holds for the *"slice"* update scheme, which has under the basic regularization by trend a higher sum of differences $\boldsymbol{\alpha}_0'\mathbf{K}_0\boldsymbol{\alpha}_0$ compared to the block update schemes. Under the basic setting the regularization with the *"mcond"* schemes is by trend weaker as e. g. with the *"slice"* update scheme, but the associated differences in the penalty are only marginally reflected in the MSEs of the baseline error density estimation.

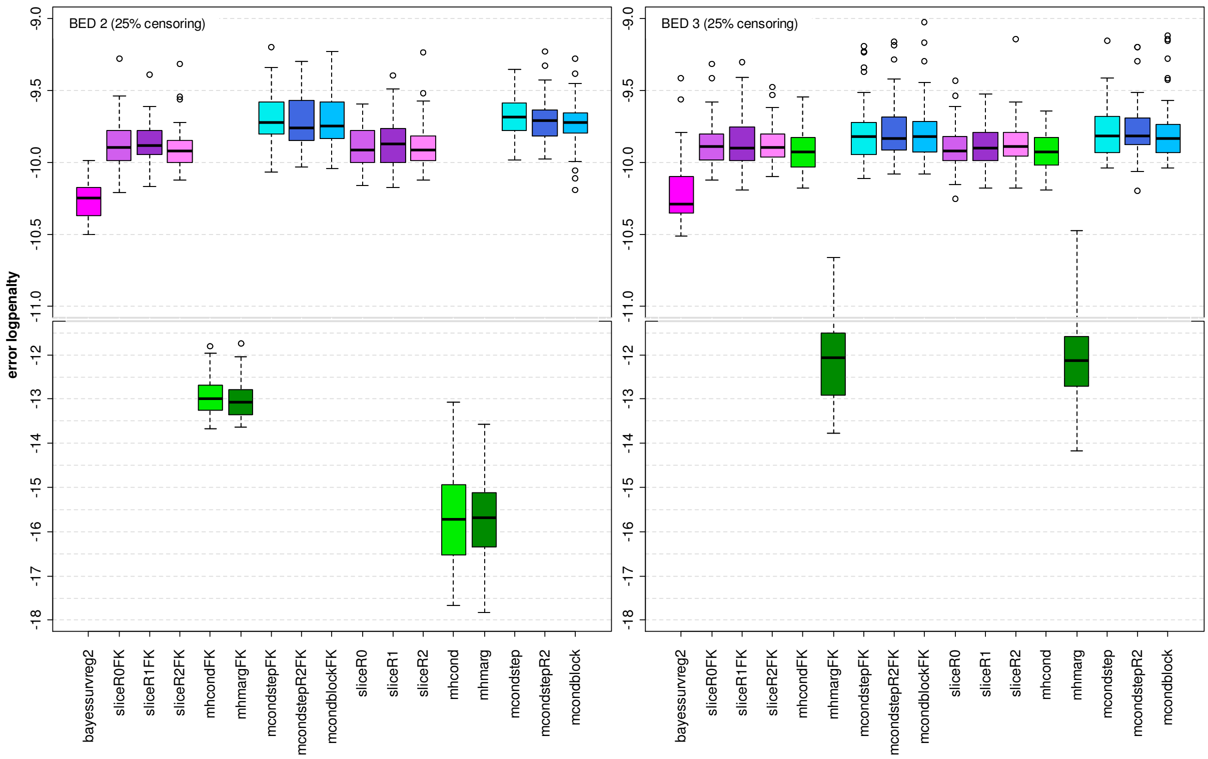


Figure 10.3: Logarithm of the estimated error density penalty term , $-\tau^{-2}_{\alpha_0}\boldsymbol{\alpha}_0'\mathbf{K}_0\boldsymbol{\alpha}_0$, in the AFT model with baseline error distribution BED 2 (left side) and BED 3 (right side) under 25% censoring in the simulation data. The scale of the y-axis changes at the tick mark within the interval $[-12,-11]$.

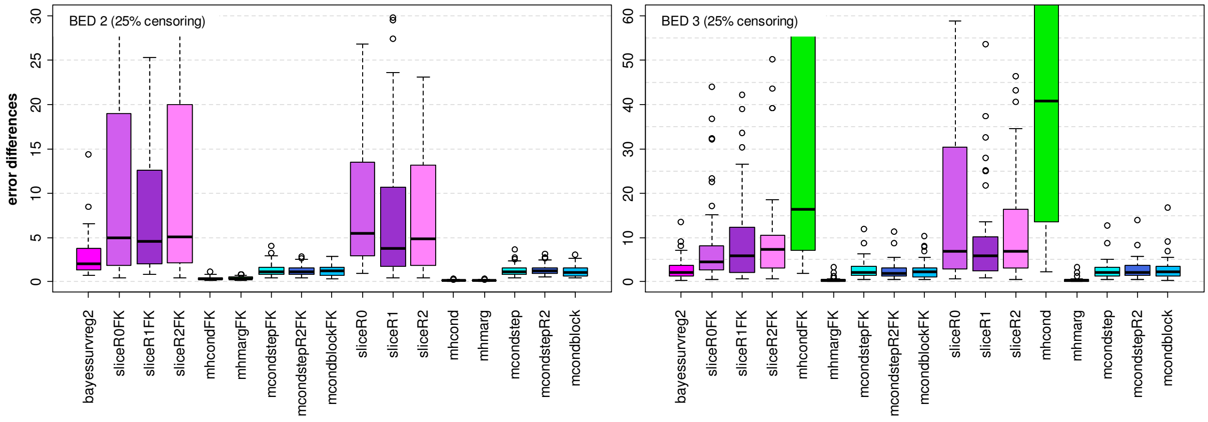


Figure 10.4: Estimated sum of the error penalty differences, $\boldsymbol{\alpha}_0'\mathbf{K}_0\boldsymbol{\alpha}_0$, in the AFT model with baseline error distribution BED 2 (left side) and BED 3 (right side) under 25% censoring in the simulation data.

Baseline error density

Figure 10.5 shows a couple of the estimated baseline error densities under the error model BED 2 with 25% censoring in the simulation data. Displayed are estimates obtained via the update schemes *"mhmarg"* and *"mhcond"*, with increased regularization in BED 2, together with the estimates under the unregularized *"dirichlet"* update scheme and the estimates from *"bayessurvreg2"* as competitor. We note that the increased regularization corrupts the fit in some error density regions which causes finally the increase in the associated MSEs. In particular the fit in the cavity between the two modes and the right mode declines compared e. g. to the method *"bayessurvreg2"*.

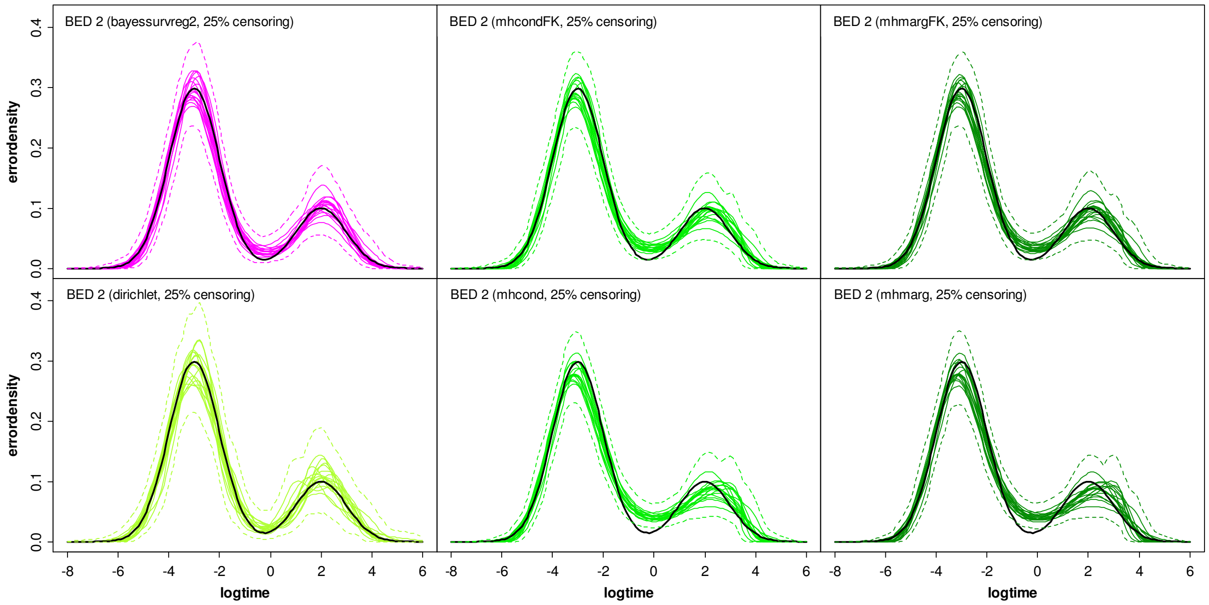


Figure 10.5: Estimated baseline error densities in the AFT model with baseline error distribution BED 2 under 25% censoring in the simulation data for six selected update schemes of the error weights. Displayed are the posterior mean estimates of the error density (colored lines) together with the true error density (black line) for *"bayessurvreg2"* (upper left panel) and under the update schemes *"dirichlet"* (lower left panel), *"mhcond"* (middle panel) and *"mhmarg"* (right panel) for the error weights. The dashed lines mark the minimum of the lower 2.5% quantile and maximum of upper 97.5% quantile in the replications.

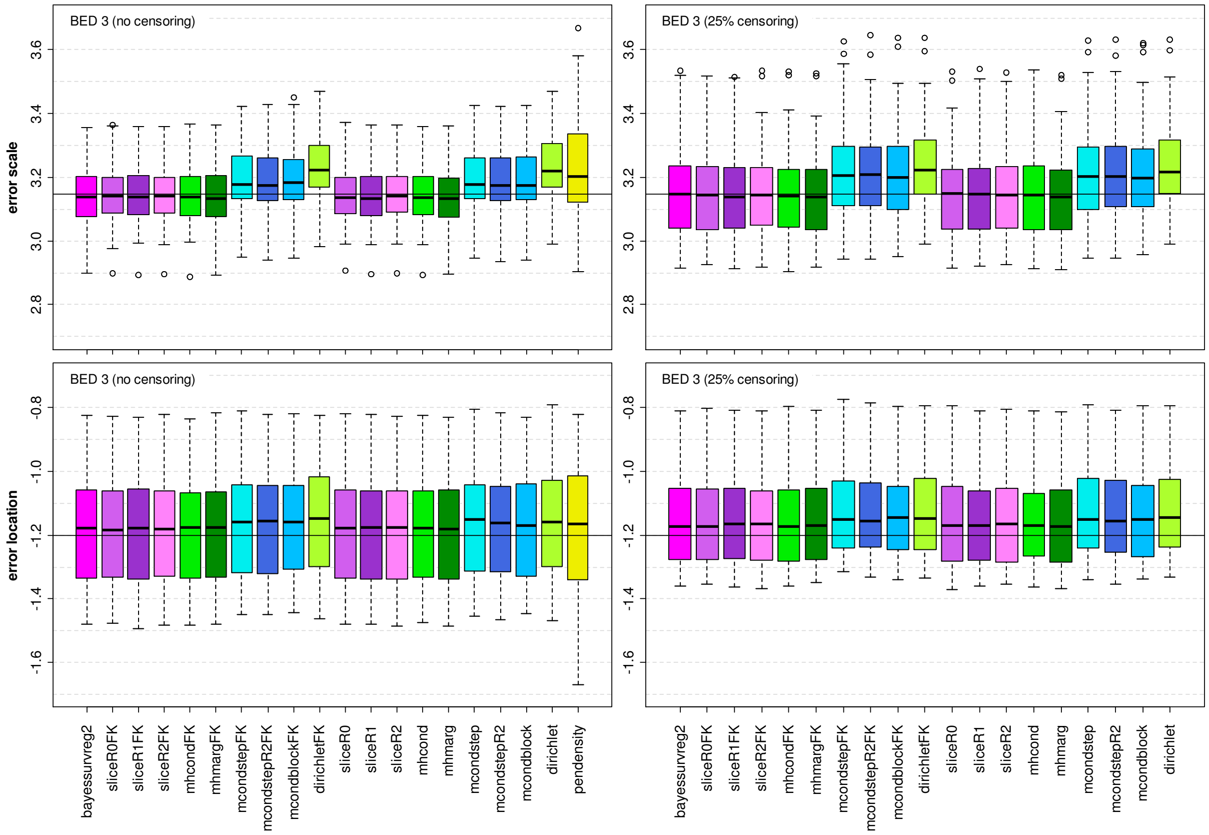


Figure 10.6: Estimated scale (upper panel) and location parameter (lower panel) of the baseline error in the AFT model with baseline error distribution BED 3, without censoring (left panel) and under 25% censoring (right panel) in the simulation data. The black horizontal lines mark the true scale σ_{Y_0} and location μ_{Y_0} parameters under BED 3.

Baseline error location and scale

The estimated location and scale parameter of the baseline error distribution model BED 3 under the different update schemes are given in the lower and upper panel of **Figure 10.6**. We observe in general a higher variability in the estimates of the scale σ than in the location γ_0 parameter and often also wider interquartile ranges (IQR) of the boxes under censoring, reflecting the increased uncertainty.

In general the estimates under the Metropolis schemes *"mcondstep"* and *"mcondblock"* differ from the remaining update schemes. While the differences in location parameter are rather marginal, they are more obvious for the scale parameter especially in the error settings BED 3 and BED 4. By trend we observe smaller absolute values for the location parameter and larger values for the scale parameter.

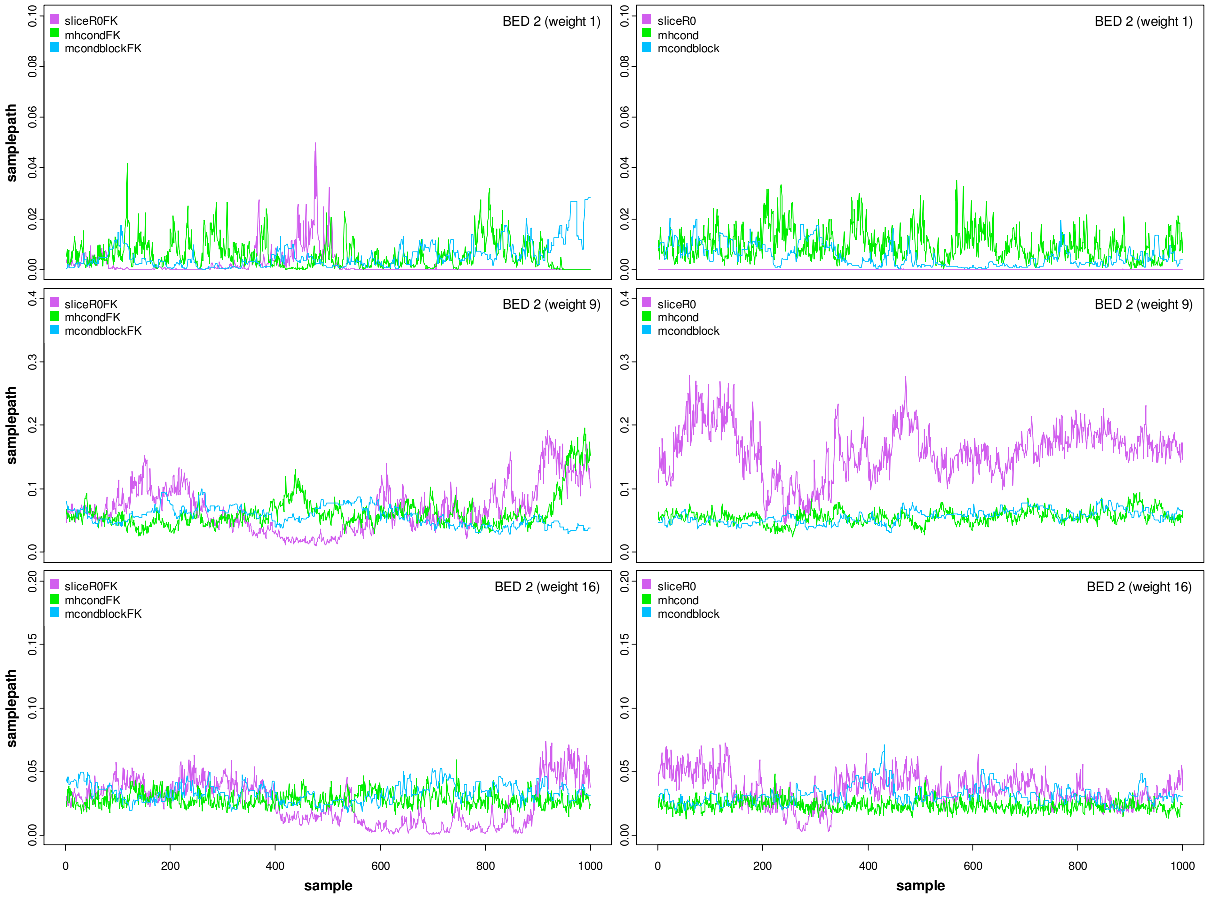


Figure 10.7: Trace plots of three selected sampled error density weights w_1, w_9 and w_{16} in the AFT model with baseline error distribution BED 2. Displayed are the results achieved with the three update schemes *"sliceR0"*, *"mhcond"* and *"mcondblock"* for the error weights, if the error density is not scaled (right panel) or scaled (left panel) within the sampler.

With respect to the MSE performance of the baseline error density estimates we found, that, e. g., the MSE superiority of the *"pendensity"* procedure under BED2 or the *"dirichlet"* update scheme under BED 3 and BED 4 is not reflected in an improved fit to the location and scale parameter. Vice versa also the procedures with poor MSE performance, like, e. g., the *"mhmarg"* update, show a comparable fit to both parameters. Especially the *"survreg"* procedure with Gaussian error provides location and scale estimates with comparable boxes like, e. g., those of *bayessuarvreg2*. So, comparing the fit to the location and scale parameter of the various update schemes enables very limited conclusions about the associated fit to the baseline error density (and reversed), and we present in the following sections only the MSE of the baseline error density estimate.

Baseline error weights

Figure 10.7 displays the sample paths of three selected error weights for one selected simulation dataset from baseline error model BED 2 without (left panel) and with (right panel) the sampler internal standardization of the error density. The option, to standardize the error density during the sampling, introduces more stability in the paths of the larger error weights for the block update schemes, as shown e. g. in the second row of the figure. But nevertheless, to show the desired stationarity of the error weights a recomputation of the weights, as described in Section 6.3.3, is essential.

With the described approximative method we compute the paths given in **Figure 10.8**. Also the low acceptance rates of the update scheme *"mcondblock"* are reflected by the piecewise constant values of the (thinned) sampled weights. In the displayed replication we have an acceptance rate of 3.8 %. The acceptance rates of the displayed *"mcondblock"* update scheme are higher than 80%.

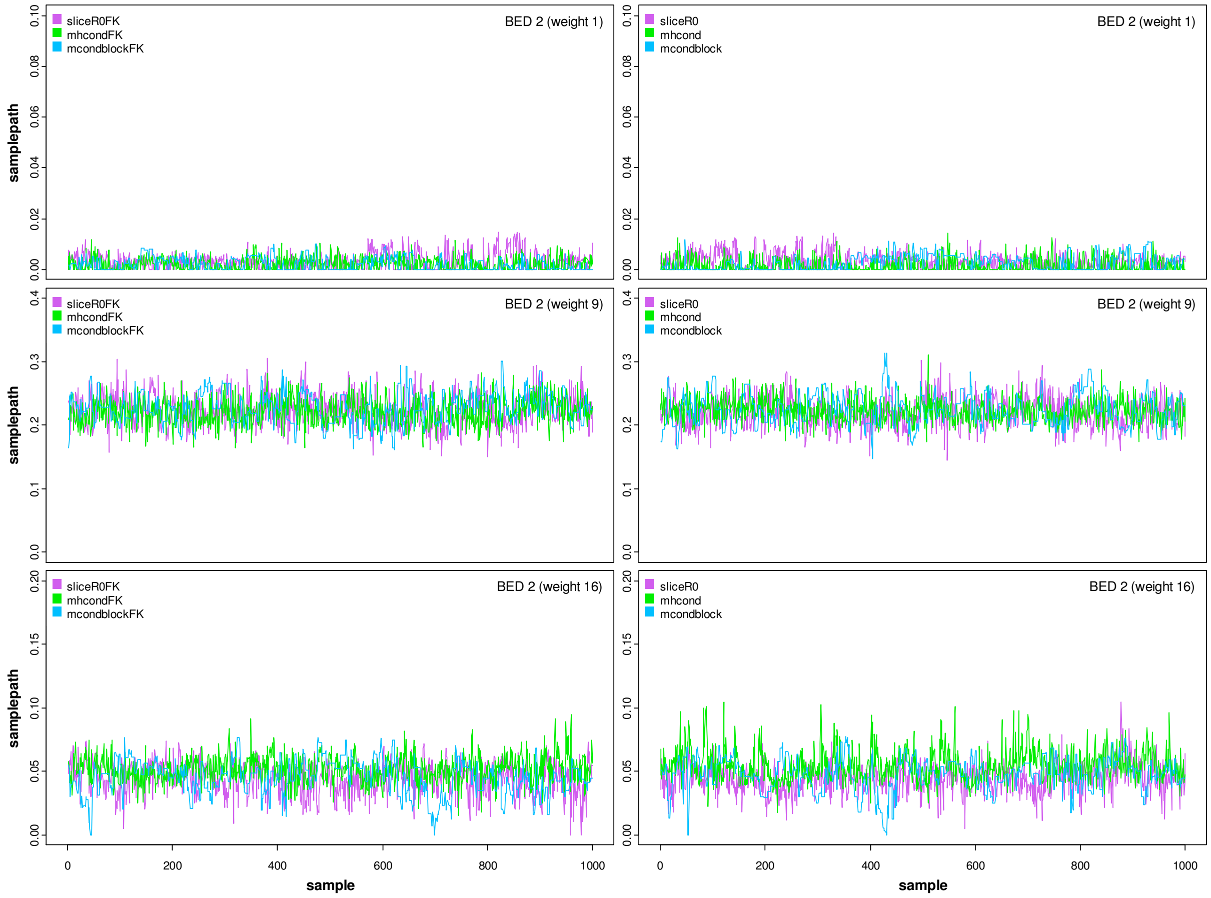


Figure 10.8: Trace plots of three selected recomputed error density weights w_1, w_9 and w_{16} in the AFT model with baseline error distribution BED 2. Displayed are the results achieved with the three update schemes *"sliceR0"*, *"mhcond"* and *"mcondblock"* for the error weights, if the error density is not scaled (right panel) or scaled (left panel) within the sampler.

In summary, none of the considered update schemes has shown a uniformly superiority across the four different error models. Also the option to standardize the error within the sampling or the variation of the update order of the error mixture weights has not shown any systematic impact on the performance of the density estimate, and the same holds for the *"mcondblock"* update scheme with the low acceptance rates. The main influence is caused by the regularization, where a stronger regularization induces a loss of performance.

10.2. Low-dimensional predictor

Data generation

Now covariates are added to investigate additionally the shrinkage properties of the regularization priors and the impact on the baseline error density estimation. We include $p_x = 25$ linear covariate effects ranging from three to zero, in particular

$$y_i = \mathbf{x}_i'\boldsymbol{\beta} + y_{0,i}, \quad y_{0,i} \sim_{iid} \text{BED}\,\ell, \quad i = 1,...,n \ ,$$

with

$$\boldsymbol{\beta} = (3,3,0,0,2,2,0,0,1,1,0,0,0.5,0.5,0,0,0.2,0.2,0,0,0.1,0.1,0,0,0)', \tag{10.2}$$

where $y_{0,i} \sim_{iid} \text{BED}\,\ell$, $\ell \in \{1,...,4\}$, denotes the associated baseline error distribution from models BED 1 to BED 4. The corresponding covariates are generated with zero mean, unit variance and the correlation between $\mathbf{x}_j$ and $\mathbf{x}_k$ is set to $\text{corr}(x_{i,j}, x_{i,k}) = \rho^{|j-k|}$ with $\rho = 0.5$. The log-survival times y_i are generated by adding the individual specific, covariate dependent value of the predictor $\eta_i = \mathbf{x}_i'\boldsymbol{\beta}$ to the random errors $y_{0,i}$, that are drawn respectively from BED1 to BED 4, i. e. $y_i = \eta_i + y_{0,i}$. Further, the censoring times and the desired percentage of censored survival times are generated as described in the previous Subsection 10.1. We use again $R = 50$ replicated datasets for each of the four baseline error models with $n = 500$ observations and 25% censoring in the data.

To explore the performance of the applied estimation methods, when the number of observations and parameters varies, this basic setting is modified. In particular in this subsection we consider in the following

- $p_x = 25$ linear modeled covariates in combination with $n = 500$ observations under the four error models BED 1 to BED 4,
- $p_x = 25$ linear modeled covariates in combination with a decreasing number of observations $n = 400, 300, 200, 100$ under the error model BED 2.

In the next Subsection 10.3 we increase also the number of parameters by modeling the $p_x = 25$ covariates as nonlinear and by increasing the number of covariates p_x.

Function and parameter specification

Methods: For the Bayesian estimation of the error distribution density with the function `baftpgm()` we use commonly the selected update schemes *"sliceR0"* (as single update), *"mcondblock"* (due to low acceptance rates and weaker penalty under standard smoothing prior configuration), *"mcondstep"* (as iterative block update with higher acceptance rates), *"mhcond"* (due to the stronger smoothness regularization), *"dirichlet"* (due to no smoothness regularization) together with the Gaussian error assumption, *"gauss"* (to consider the impact of the miss-specification of the baseline error). We present the results obtained with the option `scalebasis=TRUE` that force the standardization of the baseline error density within the sampler.

The linear effects of the predictor are estimated unregularized and regularized by utilizing the Bayesian ridge (non-adaptive version B), the lasso and the NMIG shrinkage prior. Further, we estimate the linear effects unregularized with the full predictor (PGM.B) and true predictor structure (PGM.BT), where the covariates with zero effects are omitted, using the function `bayessurvreg()`.

Hyperparameters: The hyperparameters of the error priors are set to the same values as in the previous Subsection 10.1. With the same reasoning we tune the smoothness prior of the error density for the update schemes *"mhmarg"* and *"mhcond"* to enforce a stronger regularization. Particularly under error model BED 2 with decreasing number of observations we use $h_{1,\tau_0} = 5$ $(n = 500)$, $h_{1,\tau_0} = 15$ $(100 < n \leq 400)$ and $h_{1,\tau_0} = 20$ $(n = 100)$. Under the basic setting $h_{1,\tau_0} = 1$ we obtained with the update scheme *"mhcond"* and $n = 500$ observations reasonable results in combination with the Bayesian NMIG regularization and the unregularized estimation of the linear effects and the presented results are obtained with the basic setting. With $n = 100$ observations the sampler frequently stucks under *"mhmarg"* update scheme and the results are omitted.

The hyperparameters of the regularization priors for the linear effects are set to the following values: For the shrinkage parameter prior of the Bayesian lasso and Bayesian ridge regularization we set $h_{1,\lambda} = h_{2,\lambda} = 0.01$ to enable data driven estimates of the associated model components. Due to the selected sizes of the regression coefficients we use the NMIG prior setting $v_1 = 1$, $v_0 = 0.005$, $h_{1,\psi} = 5$, $h_{2,\psi} = 50$ for the components of the variance parameter together with $h_{1,\omega} = 1$ and $h_{2,\omega} = 1$ for the complexity parameter. With respect to the results from Section 4.5 effects with absolute value larger than 1 should be less regularized. The second alternative NMIG hyperparameter setting is considered in the CRR simulations. We use a block size of 25 for the regression coefficients, which entails that 25 effects are simultaneously updated.

Starting values: The parameters associated to the error component start with the values listed in the previous Subsection 10.1. For the additional starting values of the linear effects we choose values close to zero, i. e. $\beta_j^{(0)} = 0.01$, $j = 1, ..., p_x$. The Bayesian NMIG prior components start with $I_j^{(0)} = v_0$, $\psi_j^{2(0)} = 0.0416$, which corresponds to the left mode of the bimodal variance prior, and $\omega^{(0)} = 0.5$, while the shrinkage parameter for the Bayesian lasso and ridge prior starts in $\lambda^{(0)} = 1$.

Estimation: For the MCMC algorithms we use again 30000 iterations, where the first 15000 iterations are discarded as burnin of the Markov chain and the remaining iterations are thinned using a step width of 15. We observed running times of the sampler within the range 12–17 min ($p_x = 25, n = 100$) and 12–40 min ($p_x = 25, n = 500$).

Results

MSE of the baseline error density

Results with n = 500 observations: **Figure 10.9** presents the MSEs of the estimated error densities, $\text{MSE}(\hat{f}_{Y_0})$, under the Bayesian lasso regularization of the linear effects for the case with $n = 500$ observations. The shown error model specific MSE pattern, induced by the various update schemes of the error weights, is almost identical under all three regularization priors and if the linear effects are estimated unregularized. Further, the upper panel of **Figure 10.10** shows the results for error model BED 2 with $n = 500$ observations under the three shrinkage priors and for the unregularized effects.

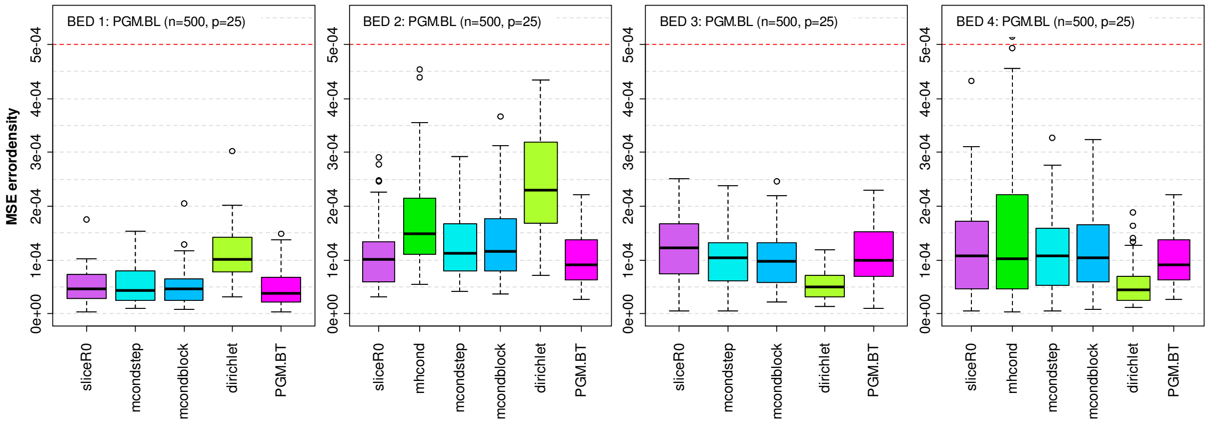


Figure 10.9: Mean squared errors of the estimated baseline error density, MSE($\hat{f}_{Y_0}$), in the AFT model with baseline error distribution BED 1 (left side) to BED 4 (right side), $p_x = 25$ covariates and $n = 500$ observations under the Bayesian lasso regularization of the linear effects. Displayed are the estimates under the update schemes *"sliceR0"*, *"mhcond"*, *"mcondstep"*, *"mcondblock"* and *"dirichlet"* for the error weights. PGM.BT denotes the corresponding results from the model using the true predictor structure. The red dotted line marks the value 5e-5.

Under the error models BED 1 and BED 2 the results from **Figure 10.9** are almost comparable, with respect to the median, to those when no covariates are included in the models (**Figure 10.2**). Under the error models BED 3 and BED 4 we generally observe an increase of the MSEs when covariates are added, but the *"dirichlet"* update scheme still has the best performance, even if compared to the model using the true predictor (PGM.BT). Further, under error model BED 1, the *"dirichlet"* update scheme seems to profit from the inclusion of the covariates, since the median MSE is decreased. At the opposite, under the error model BED 2, a clear increase for the *"dirichlet"* scheme is shown except for the Bayesian NMIG regularization, see upper right panel of **Figure 10.10**. The variability in the MSEs of the scheme *"mhcond"* under model BED 2, as shown in the upper panel of **Figure 10.10** (second column), is caused by the different amounts of regularization of the baseline error. As mentioned in the function and parameter specification, we use the standard setting, $h_{1,\tau_0} = 1$, for the error only in combination with the PGM.B and PGM.BN regularization of the linear effects and $n = 500$ observations. In both cases the resulting MSEs are comparable to those obtained with the update scheme *"sliceR0"*, where the basic setting is generally used.

In summary, in the case of $n = 500$ observations the various regularization methods of the linear effects cause no systematic differences in the (update scheme specific) MSE of the baseline error, as shown in the upper panel of **Figure 10.10**. With exception of the *"dirichlet"* and the Metropolis update scheme *"mhcond"* the MSEs are almost comparable to the MSE resulting from the model with the true predictor structure (PGM.BT). The stronger smoothness regularization under the *"mhcond"* update scheme causes a loss in the performance as already observed in the models without covariates.

Results with n < 500 observations: When the sample size decreases from $n = 500$ to $n = 100$ observations, the MSEs of the error density estimate generally increase. But also a change in the MSE performance, caused by the specific regularization method of the linear effects, is exposed. This is shown in the lower panel of **Figure 10.10** by means of the error model BED 2 with $n = 100$ observations, where the performance under the three shrinkage priors is clearly improved in comparison to the unregularized estimation of the linear effects. Especially the MSEs under the Bayesian lasso and NMIG prior are lower as under the Bayesian ridge prior, with an advance for the Bayesian NMIG regularization.

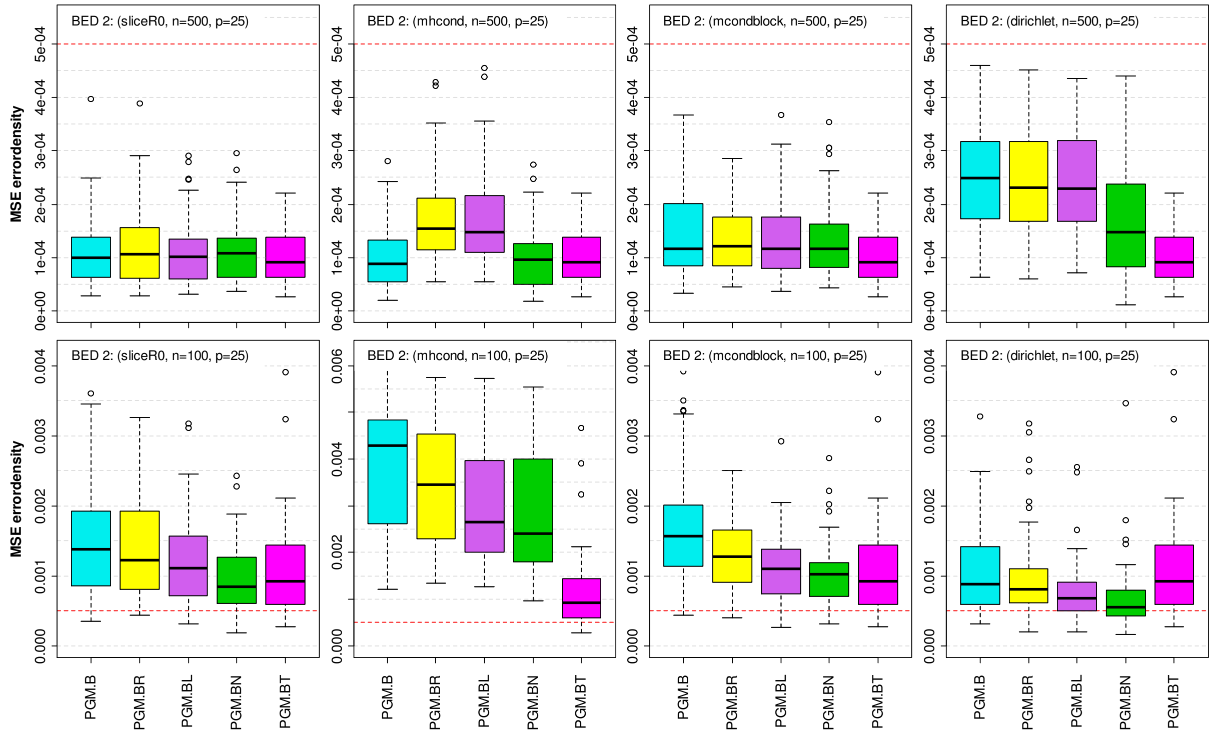


Figure 10.10: Mean squared errors of the estimated baseline error density, $\text{MSE}(\hat{f}_{Y_0})$, in the AFT model with baseline error distribution BED 2, $p_x = 25$ covariates and $n = 500$ (upper panel) to $n = 100$ (lower panel) observations under various shrinkage priors with the update schemes *"sliceR0"* (first column), *"mhcond"* (second column), *"mcondblock"* (third column) and *"dirichlet"* (fourth column) for the error weights. Displayed are the estimates under no, Bayesian ridge, Bayesian lasso and Bayesian NMIG regularization of the linear effects. PGM.BT denotes the results from the model using the true predictor structure. The red dotted line marks the value 5e-5.

Again some specific behavior is observable for the update schemes *"mhcond"* and *"dirichlet"*. As before, the poor performance results under the update scheme *"mhcond"* are explained by the stronger smoothing of the error density across the shrinkage priors, but the hierarchy of the performances, with respect to the regularization variant of the linear effects, is identical to the other update schemes and the best results are obtained under the Bayesian NMIG prior. Especially the MSE performance of the *"dirichlet"* update scheme gets increasingly better, with respect to the other update schemes, when the sample size decreases from $n = 500$ (**Figure 10.9**). Already with $n = 300$ observations (results not shown) the MSE of the *"dirichlet"* update scheme is comparable to the MSEs of the other update schemes and decreases further to the low values shown in **Figure 10.10**. In particular for $n = 100$ observations the MSE performance under the Bayesian lasso and NMIG prior is higher as for the model with the true predictor.

Result with Gaussian error: Finally, we consider the results under the Gaussian error assumption, see **Figure 10.11**. The MSEs of the frequentist AFT models with Gaussian error (AFT), stepwise selection

(AFT.Step) and the true predictor structure (AFT.T) are almost comparable to each other and we show only the results for the stepwise selection.

With respect to the Bayesian methods, the frequentist approaches yield only with error model BED 4 lower MSEs. Within the specific baseline error model the performance of the density estimation under the various Bayesian shrinkage priors is almost comparable in the case of $n = 500$ observations and with decreasing sample size the shrinkage of the linear effects improves the performance, but by a smaller amount as under the PGM error model. In the case of $n = 100$ observations the previously observed performance hierarchy of the regularization priors is reversed and the best results are obtained with the ridge prior followed by the lasso and the NMIG prior, but the differences are marginal.

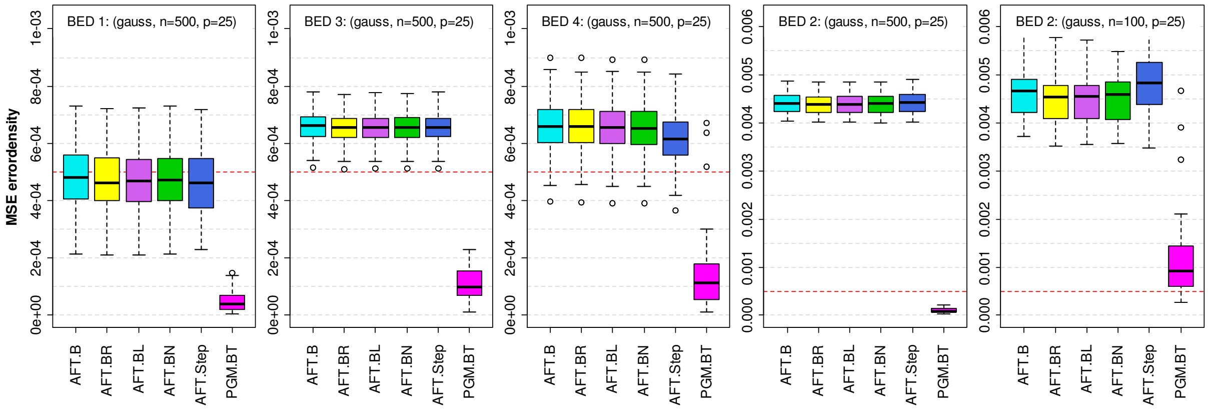


Figure 10.11: Mean squared errors of the estimated baseline error density, $\mathrm{MSE}(\hat{f}_{Y_0})$, in the AFT model with baseline error distribution BED 1 to BED 4, $p_x = 25$ covariates and $n = 500$ or $n = 100$ observations under various shrinkage priors and the Gaussian error assumption. Displayed are the estimates under no, Bayesian ridge, Bayesian lasso and Bayesian NMIG regularization of the linear effects and the frequentist stepwise selection, AFT.Step. PGM.BT denotes the results from the model using the true predictor structure. The red dotted line marks the value 5e-5.

We have also observed a stronger deviation in the estimated location and scale parameter, when the sample size decreases. But, as before in Section 10.1, the differences in the MSE of the estimated error densities are in general not reflected in the location and scale parameter estimates.

Baseline error density

Figure 10.12 shows the estimated error densities under model BED 2 for three update schemes of the error weights with the Bayesian NMIG regularization of the linear effects. From the upper to the lower panel the number of observations is decreased. Compared to **Figure 10.13**, that shows the corresponding results for the unregularized linear effects, the estimates are often more concentrated around the true baseline error density. In the middle column we see the impact of the stronger regularization under the *"mhcond"* scheme (if $n = 200,100$). With $n = 200$ observations the right mode is shifted towards the right border and with $n = 100$ observations the stronger regularization often avoids the adaptation of the estimates to the two modes and the cavity between the two modes and the estimates are often unimodal. These effects are less pronounced under the NMIG prior as under the unregularized linear effects.

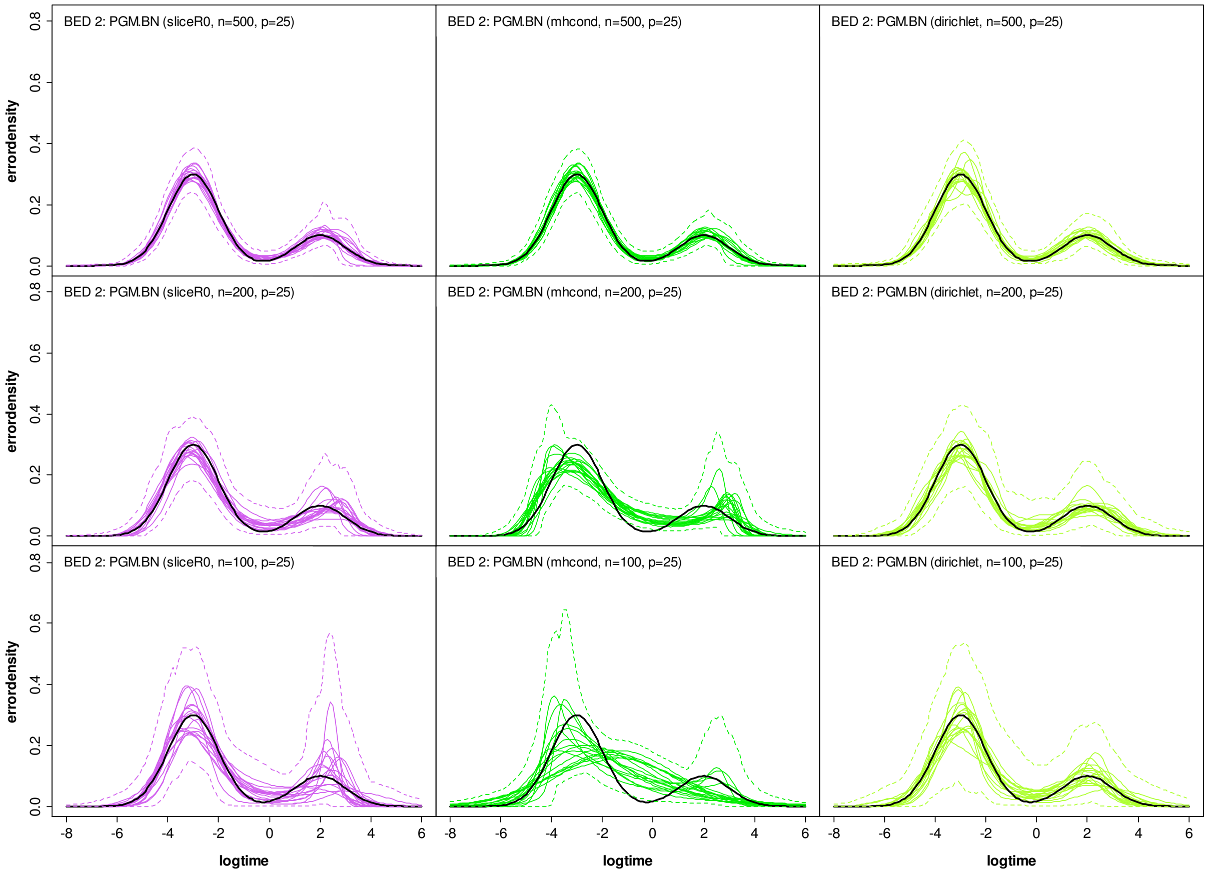


Figure 10.12: Estimated error distribution densities in the AFT model with baseline error distribution BED 2, $p_x = 25$ covariates and $n = 500$ (upper panel), $n = 200$ (middle panel) or $n = 100$ (lower panel) observations under the Bayesian NMIG prior for selected update methods of the error weights. Displayed are the posterior mean estimates of the error density (colored lines) together with the true error density (black line) under the update schemes *"sliceR0"* (left panel), *"mhcond"* (middle panel) and *"dirichlet"* (right panel) for the error weights. The dashed lines mark the minimum of the lower 2.5% quantile and maximum of upper 97.5% quantile in the replications.

MSE of the regression coefficients

Results with n = 500 observations: The results for the cases with $n = 500$ observations are given in **Figure 10.14**, which shows the MSEs of the estimated regularized regression coefficients, $\text{MSE}(\hat{\boldsymbol{\beta}})$, in the error models BED 1 to BED 4 under the Bayesian lasso prior, and in the upper panel of **Figure 10.10**, which shows the results for error model BED 2 with $n = 500$ observations under the three shrinkage priors and for the unregularized effects.

If we compare **Figure 10.14** and **Figure 10.9,** we see that the error model specific differences in the MSEs of the baseline error density, caused by the various update schemes, are less pronounced in terms of the MSEs of the regression coefficients. In particular the outstanding high or low baseline error performances, observed e. g. under the *"dirichlet"* and *"mhcond"* update scheme, are only marginally reflected, but in general we can recognize a similar MSE structure as in **Figure 10.9** with weaker differences. With the given structure of the underlying effects (10.2) we do not reach MSEs comparable to the model using the true predictor structure (PGM.BT), but we notice that the MSE decreases under all error models from the Bayesian ridge over the Bayesian lasso to the Bayesian NMIG regularization, with sometimes marginal differences under the last two priors, compare upper panel of **Figure 10.15**.

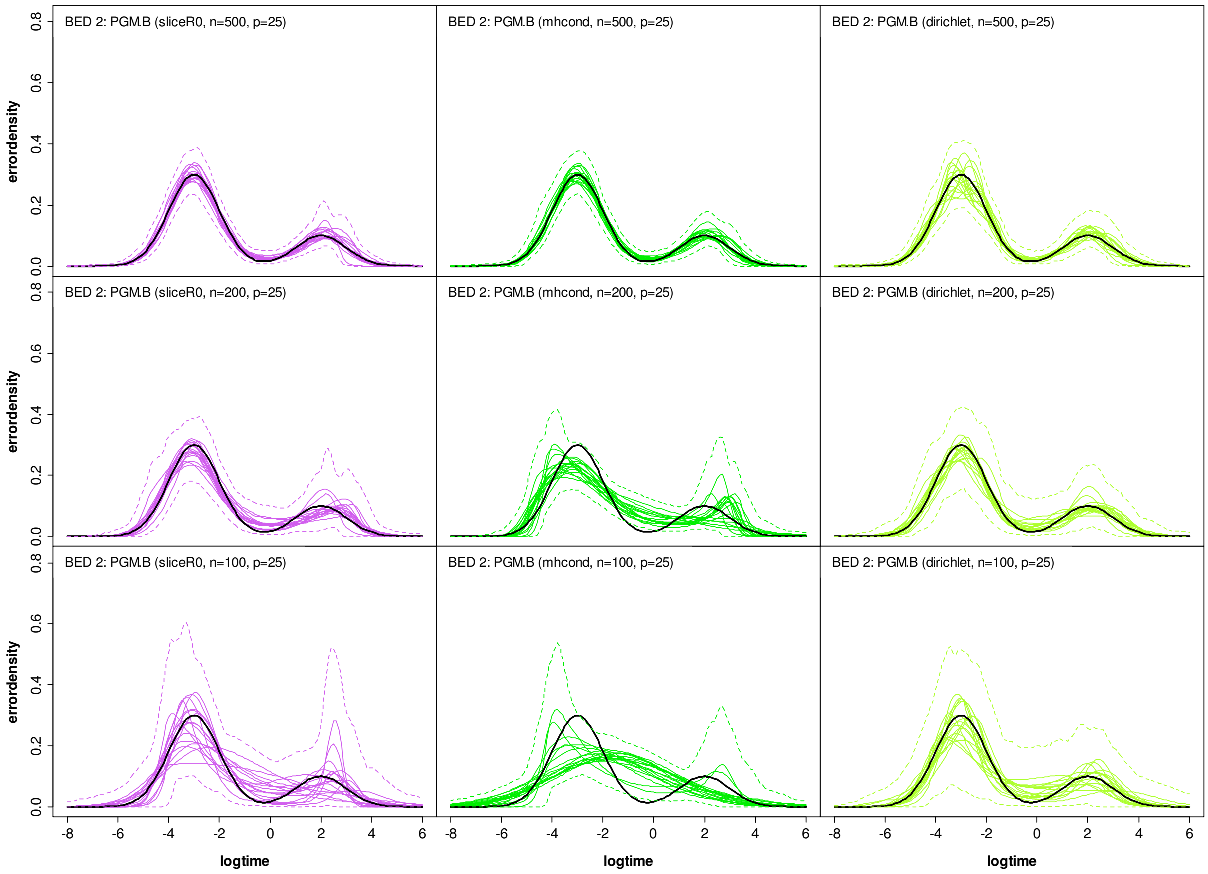


Figure 10.13: Estimated error distribution densities in the AFT model with baseline error distribution BED 2, $p_x = 25$ covariates and $n = 500$ (upper panel), $n = 200$ (middle panel) or $n = 100$ (lower panel) observations under unregularized linear effects for selected update methods of the error weights. Displayed are the posterior mean estimates of the error density (colored lines) together with the true error density (black line) under the update schemes *"sliceR0"* (left panel), *"mhcond"* (middle panel) and *"dirichlet"* (right panel) for the error weights. The dashed lines mark the minimum of the lower 2.5% quantile and maximum of upper 97.5% quantile in the replications.

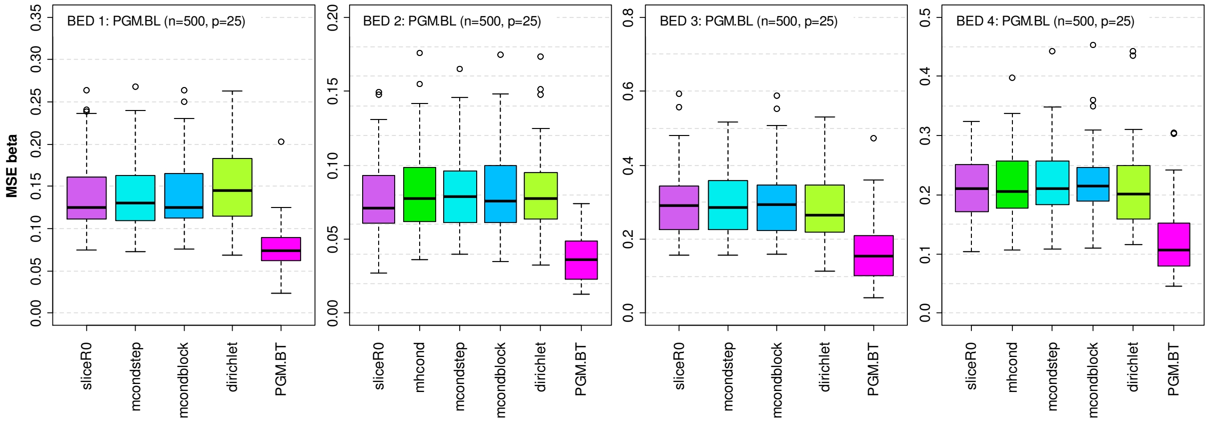


Figure 10.14: Mean squared errors of the estimated regression coefficients, $\text{MSE}(\hat{\boldsymbol{\beta}})$, in the AFT model with baseline error distribution BED 1 (left side) to BED 4 (right side), $p_x = 25$ covariates and $n = 500$ observations under the Bayesian lasso regularization of the linear effects. Displayed are the estimates under the update schemes *"sliceR0"*, *"mhcond"*, *"mcondstep"*, *"mcondblock"* and *"dirichlet"* for the error weights. PGM.BT denotes the corresponding results from the model using the true predictor structure.

Results with n < 500 observations: With decreasing sample size also the MSE of the estimated regression coefficients increases, but the observed MSE trend under the various Bayesian regularization priors with $n = 500$ observations is retained and becomes more obvious, compare lower panel of **Figure 10.15**. Further, the observed MSE pattern of the regression coefficients coincides with the previously seen pattern in terms of the MSE of the baseline error density, compare **Figure 10.10**, so that the specific regularization of the regression coefficients affects also the performance of the baseline error density estimation. Vice versa also the dependence on the baseline error density fit becomes more pronounced with decreasing sample size, e. g. the MSEs under the update scheme *"mhcond"*, with stronger regularized error density, are clearly increased compared to the other update schemes by preserving the specific hierarchy induced by the different shrinkage priors. In addition the MSE of the *"dirichlet"* update becomes more and more comparable to the MSEs under the other update schemes and from $n < 300$ observations the performance is even higher. In summary, the performance of the baseline error is connected to the performance of the predictor, where the basic level of the performance is rather determined by the fit of the baseline error and improvements are possible with an improved fit of the predictor.

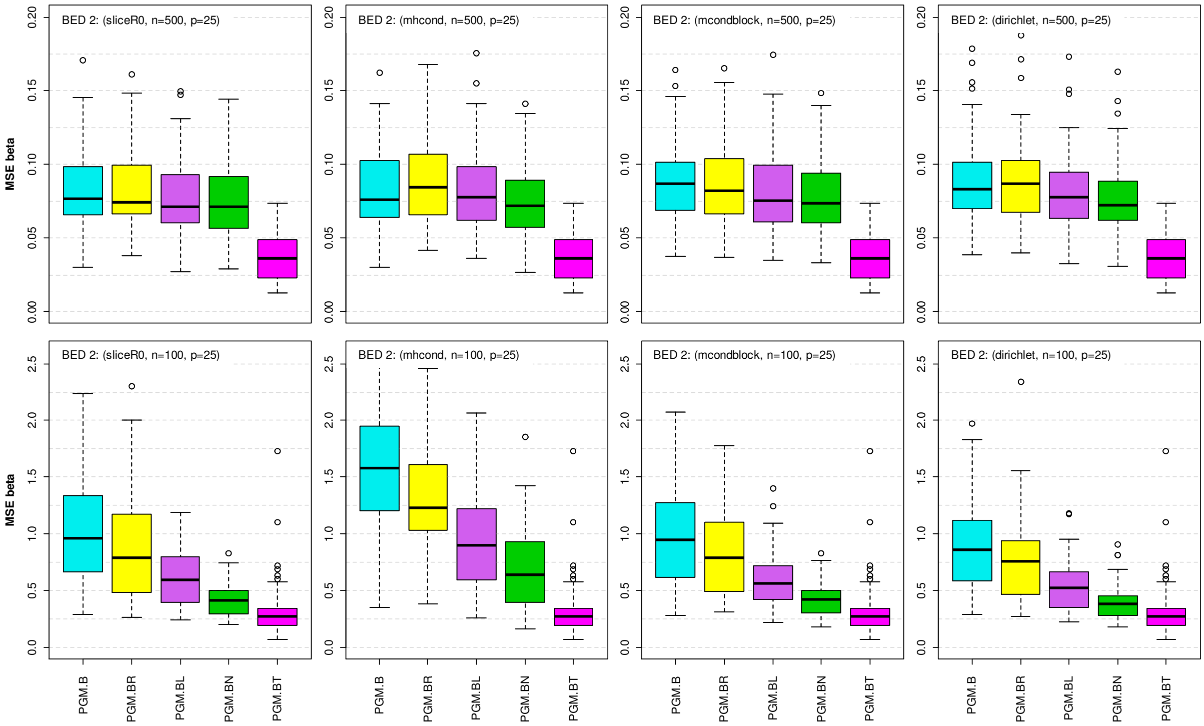


Figure 10.15: Mean squared errors of the estimated regression coefficients, $\text{MSE}(\hat{\boldsymbol{\beta}})$, in the AFT model with baseline error distribution BED 2, $p_x = 25$ covariates and $n = 500$ (upper panel) to $n = 100$ (lower panel) observations under various shrinkage priors with the update schemes *"sliceR0"* (first column), *"mhcond"* (second column), *"mcondblock"* (third column) and *"dirichlet"* (fourth column) for the error weights. Displayed are the estimates under no, Bayesian ridge, Bayesian lasso and Bayesian NMIG regularization of the linear effects. PGM.BT denotes the results from the model using the true predictor structure.

Result with Gaussian error: The observed improved performance induced by the shrinkage priors, is also observable under the Gaussian error assumption, see **Figure 10.16**. The best performance is obtained under the Bayesian NMIG regularization followed by the Bayesian lasso and ridge, where in particular the Bayesian NMIG always outperforms the results from the frequentist stepwise selection.

With respect to the results of the error density estimation, see **Figure 10.11,** we found that the clearly improved performance in the predictor is not notably reflected in the performance of the error distribution in contrast to the results with the PGM error.

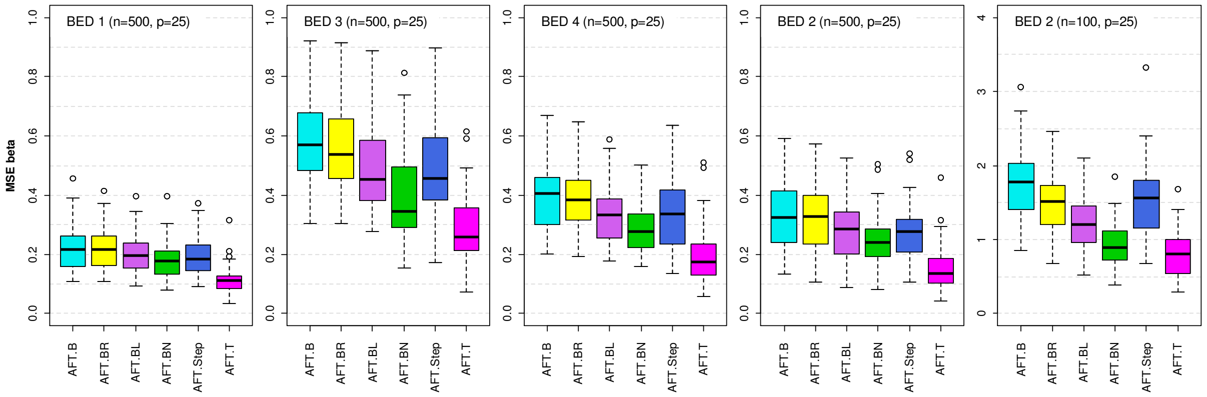


Figure 10.16: Mean squared errors of the estimated regression coefficients, MSE($\hat{\boldsymbol{\beta}}$), in the AFT model with baseline error distribution BED 1 to BED 4, $p_x = 25$ covariates and $n = 500$ or $n = 100$ observations under various shrinkage priors and the Gaussian error assumption. Displayed are the estimates under no, Bayesian ridge, Bayesian lasso and Bayesian NMIG regularization of the linear effects and the frequentist stepwise selection, AFT.Step. AFT.T denotes the results from the frequentist model using the true predictor structure.

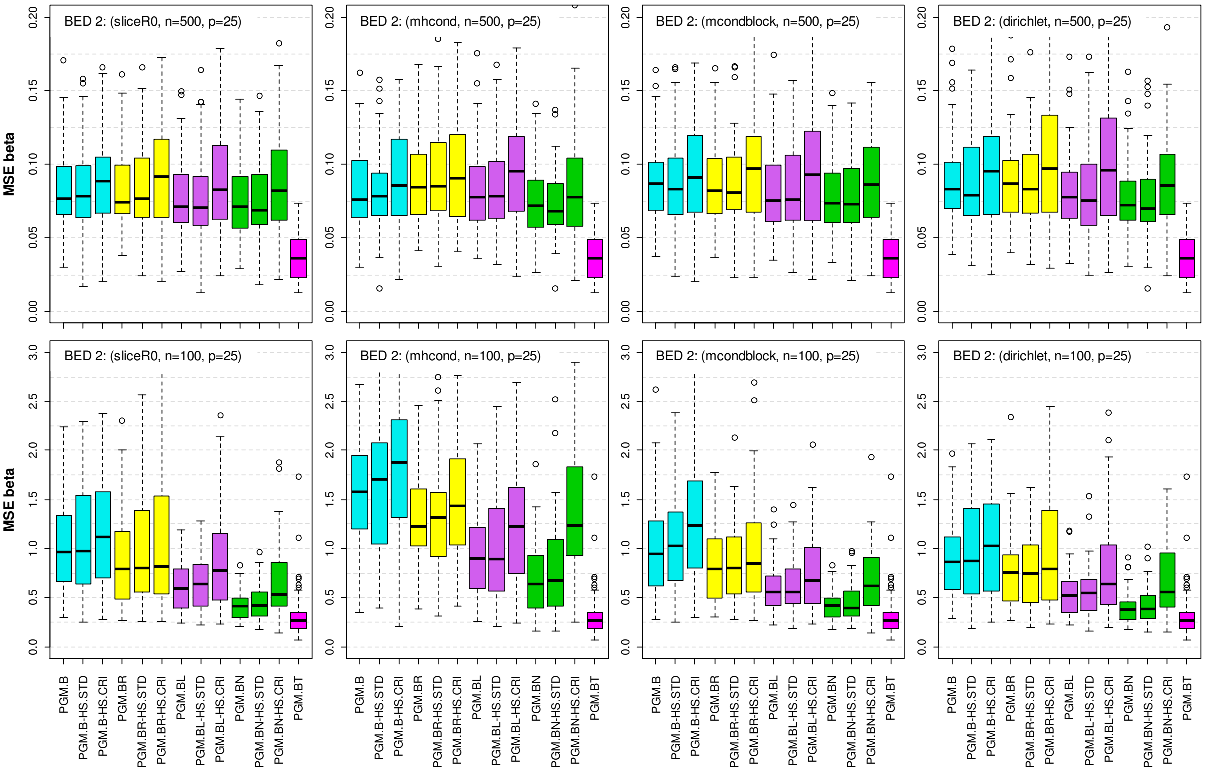


Figure 10.17: Mean squared errors of the estimated regression coefficients, MSE($\hat{\boldsymbol{\beta}}$), in the AFT model with baseline error distribution BED 2, $p_x = 25$ covariates and $n = 500$ (upper panel) or $n = 100$ (lower panel) observations under the update schemes *"sliceR0"* (first column), *"mhcond"* (second column), *"mcondblock"* (third column) and *"dirichlet"* (fourth column) for the error weights. Displayed are the estimates under no, Bayesian ridge, Bayesian lasso and Bayesian NMIG regularization of the linear effects together with the corresponding MSEs resulting from the hard shrinkage variable selection criteria. PGM.BT denotes the results from the model using the true predictor structure.

Results under variable selection: **Figure 10.17** summarizes the MSEs of the regression coefficients obtained under error model BED 2 (**Figure 10.15**) together with the resulting MSEs, if the hard shrinkage selection rules HS.STD, HS.CRI and HS.IND, as described in Section 4.4, are applied. In the figures the MSEs under the HS.IND selection rule are omitted, we obtain with $n = 500$ median values about $MSE \approx 1$ and if $n = 100$ the median MSEs are comparable to those under the HS.CRI criterion.

We find in general that variable selection does not improve the predictive performance with respect to the corresponding prior-specific full model. With $n = 500$ observations, upper panel of **Figure 10.17**, the selection criterion based on one standard deviation interval (HS.STD) is leading to sparse final models with MSE comparable to the full models that include all 25 covariates. The criterion based on the 95% credible interval (HS.CRI) and the NMIG indicator frequencies (HS.IND) set too many nonzero coefficients to zero, compare **Table 10.1**, which increases the MSE of the associated final sparse models. With reduced sample size, e. g. $n = 100$ observations, the performance of the Bayesian NMIG models increases further, relative to the Bayesian ridge and lasso prior models, and the HS.STD selection rule still yields sparse models with comparable MSE as the associated full models, compare **Figure 10.17** (lower panel). Similar results are obtained under the Gaussian error assumption.

NMIG indicators

Results with n = 500 observations: The variable importance feature of the Bayesian NMIG prior is highlighted in the upper panel of **Figure 10.18**, where estimated inclusion probabilities based on posterior relative frequencies of the NMIG indicator variable value $I_j = v_1$ are shown under three selected update schemes for the error weights and under the Gaussian error assumption.

As induced by the specific configuration of the NMIG prior in this section, the inclusion probability for covariates with absolute effect sizes within the range from 1.5 to 0.1 decreases monotonically, where by trend effects larger than 0.7 reach inclusion probabilities that exceed the cut off value 0.5 of the hard shrinkage selection criterion HS.IND, compare also Section 4.5. Based on this threshold the effects $\beta_9 = \beta_{10} = 1$ are separated from the effects $\beta_{13} = \beta_{14} = 0.5$ in the sense, that the hard shrinkage selection rule HS.IND removes the estimated effects with (absolute) size smaller than or equal to 0.5 from the final model. In our specific simulation setting (10.2) six nonzero effects $(\beta_9, \beta_{10}, \beta_{13}, \beta_{14}, \beta_{17}, \beta_{18})$ with sizes 0.5, 0.2 and 0.1 are affected by this decision rule and, as a consequence of ignoring these effects, the mean squared errors of the regression coefficients included in the resulting final models increase considerably. Especially the Gaussian error assumption in the bimodal error model BED 2 increases the uncertainty in the classification (larger box-widths) and the classification of larger effects (β_9, β_{10}) to the component $I_j = v_0$, compared PGM error representation. Nevertheless, the variable separation with HS.IND-threshold of 0.5 is not affected here and the number of correctly classified zero and nonzero effects is almost comparable with the PGM error models.

Results with n < 500 observations: With decreasing sample size the separation of the effects gets blurred, since the interquartile distances of the nonzero effects frequencies increase. The lower panel of **Figure 10.18** shows the results for the case of $n = 100$ observations. The reduced information in the data enhances the classification uncertainty and we observe an increase in the classification of larger effects to the component $I_j = v_0$ and conversely the classification of smaller effects to the component $I_j = v_1$. Besides the effect at the interquartile range also the number of extreme values and outliers increases. As observed before under the Gaussian error assumption, also the stronger smoothness regularization used in the *"mhcond"* update scheme increases further classification of larger effects to the component $I_j = v_0$. The shown inclusion probabilities of the *"mhcond"* update scheme are comparable to those under the Gaussian assumption with $n = 100$ observations. Finally, reduced sample sizes cause that larger effects are stronger and smaller effects are weaker regularized, compare also right column of **Figure 10.20**. Nevertheless, the separation with the threshold 0.5 is still only marginally affected due to the (conveniently selected) effect sizes in (10.2), compare **Table 10.1.**

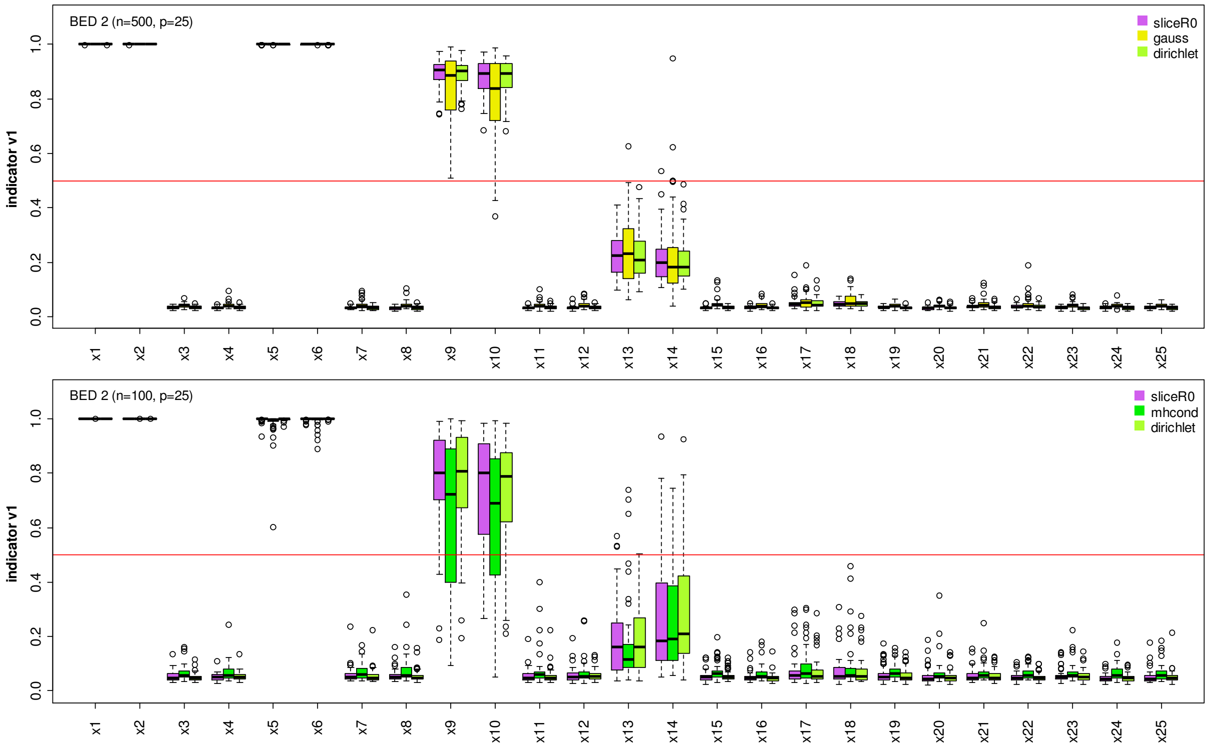


Figure 10.18: Estimated inclusion probabilities based on posterior relative frequencies of the Bayesian NMIG indicator variable value $I_j = v_1$ in the AFT model with baseline error distribution BED 2, $p_x = 25$ covariates and $n = 500$ (upper panel) or $n = 100$ (lower panel) observations. Displayed are the relative frequencies under the update schemes *"sliceR0"*, *"mhcond"* and *"dirichlet"* for the error weights and under the Gaussian error assumption (*"gauss"*). The red horizontal line marks the cut off value 0.5 of the hard shrinkage selection criterion HS.IND.

Linear effects

Figure 10.19 and **Figure 10.20** show the estimates of four selected regression coefficients $\beta_1 = 3$, $\beta_9 = 1$, $\beta_{13} = 0.5$ and $\beta_{23} = 0$ under baseline error model BED 2 with the various shrinkage priors for the linear effects. The results presented in **Figure 10.19** are obtained with the *"sliceR0"* update scheme for the PGM weights and under the Gaussian error assumption with $n = 500$ observations and **Figure 10.20** shows the results under the *"sliceR0"* and the *"mhcond"* update scheme for $n = 100$ observations.

We can observe the specific shrinkage property of the Bayesian NMIG prior (PGM.BN) in the sense of the weaker regularization of larger effects, like $\beta_1 = 3$, where the estimates are close to the unregularized estimates (PGM.B) and a stronger regularization of smaller effects, like $\beta_{13} = 0.5$, compared to the Bayesian lasso (PGM.BL) and ridge (PGM.BR) prior. As shown in **Figure 10.18** the shrinkage begins to increase for effects with absolute values smaller than 1.5, but in the cases with $n = 500$ observations the likelihood dominates the prior information and the shrinkage is only marginal. Under the Gaussian error the deviations to the true effects increase, as reflected by the increased interquartile ranges and the location of the median and the resulting MSE of the regression coefficients. Especially for the zero effects the differences are enlarged. With decreasing sample size the interquartile ranges of the estimates increase and the shrinkage gets more pronounced. In particular the stronger concentration of the estimates around zero for the zero effects under the Bayesian NMIG prior is more emphasized. Under the stronger smoothness regularization, used in the *"mhcond"* scheme with $n = 100$ observations, we observe a stronger regularization under the NMIG prior for the effects β_9 and β_{10} explained by the enhanced variation of the inclusion probabilities.

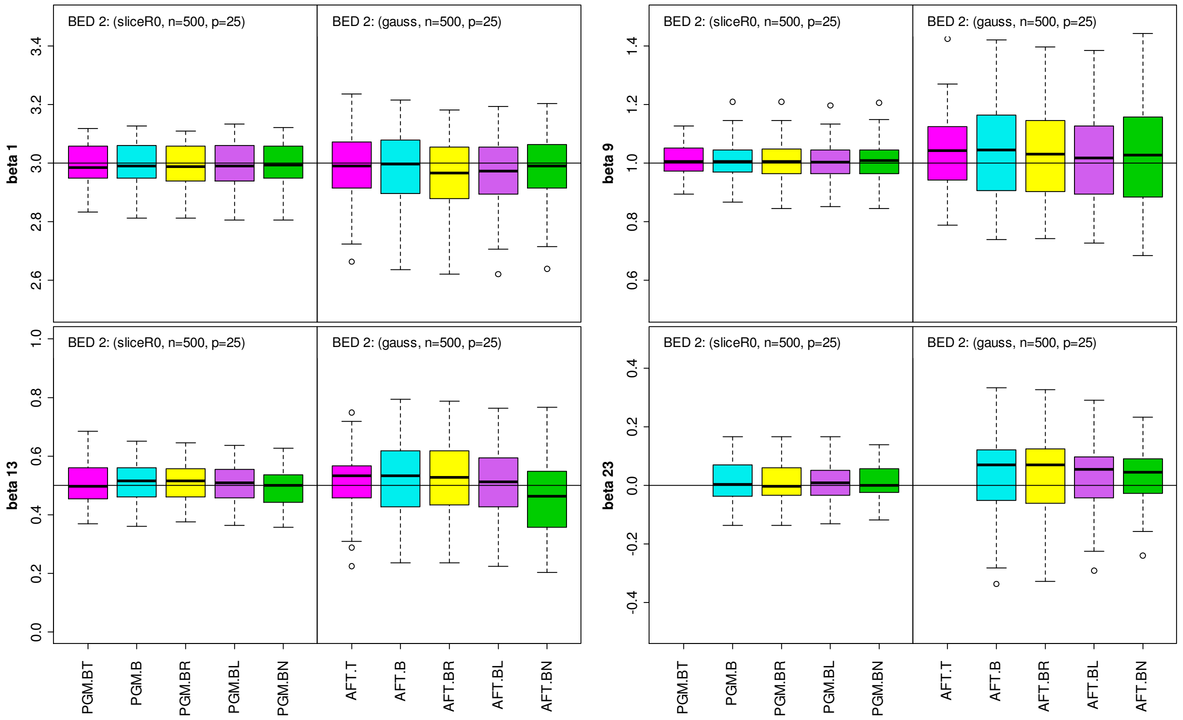


Figure 10.19: Regression coefficient estimates $\hat{\beta}$ for four selected estimated regression coefficients $\beta_1 = 3$ (upper left panel), $\beta_9 = 1$ (upper right panel), $\beta_{13} = 0.5$ (lower left panel) and $\beta_{23} = 0$ (lower right panel) in the AFT model with baseline error distribution BED 2, $p_x = 25$ covariates and $n = 500$ observations. Displayed are the estimates resulting from *"sliceR0"* update scheme of the error weights (left sides) and the Gaussian error assumption *"gauss"* (right sides). The black horizontal lines mark the true values of the regression coefficients.

Classification

Table 10.1 and **Table 10.2** show the obtained average number of the correctly classified nonzero coefficients ($\hat{\beta} \neq 0, \beta \neq 0$) and correctly classified zero coefficients ($\hat{\beta} = 0, \beta = 0$) for the 50 simulation datasets under the different variable selection methods in the AFT model. **Table 10.1** contains the results under baseline error distribution BED 1 to BED 4 with $n = 500$ observations and the results under baseline error distribution BED 2 with decreasing number of observations $n = 400$ to $n = 100$ are given in **Table 10.2**.

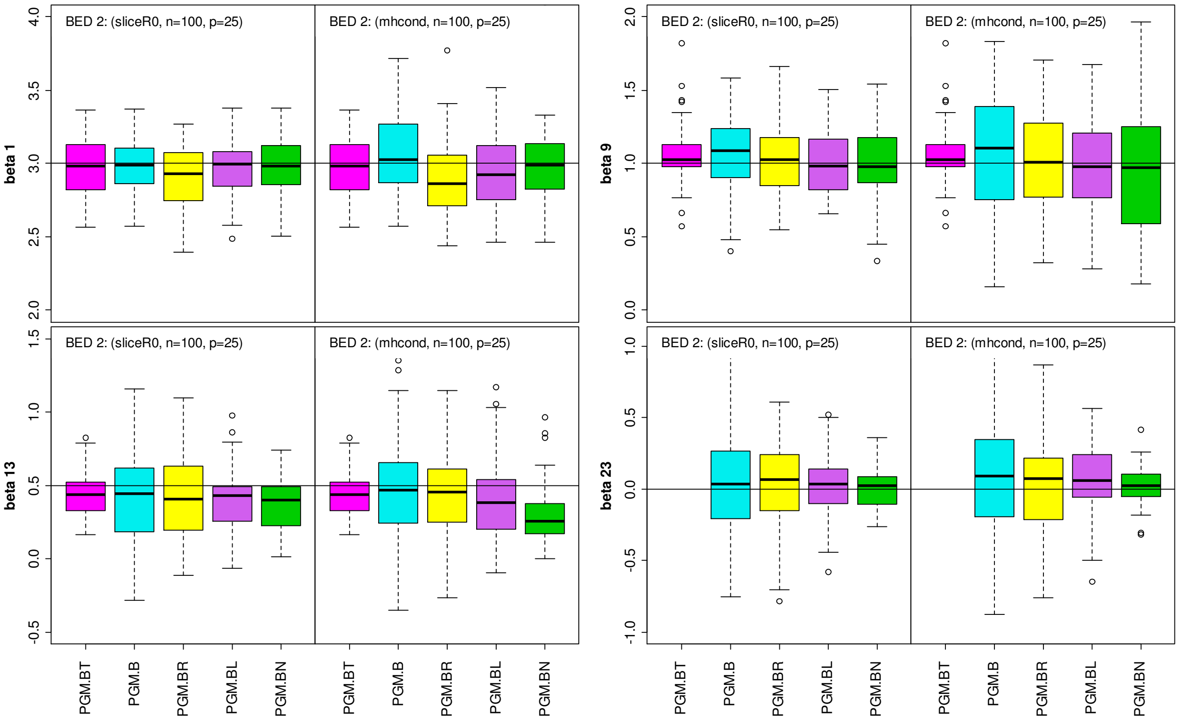


Figure 10.20: Regression coefficient estimates $\hat{\boldsymbol{\beta}}$ for four selected estimated regression coefficients $\beta_1 = 3$ (upper left panel), $\beta_9 = 1$ (upper right panel), $\beta_{13} = 0.5$ (lower left panel) and $\beta_{23} = 0$ (lower right panel) in the AFT model with baseline error distribution BED 2, $p_x = 25$ covariates and $n = 100$ observations. Displayed are the estimates resulting from *"sliceR0"* (left sides) and *"mhcond"* (right sides) update scheme of the error weights. The black horizontal lines mark the true values of the regression coefficients.

With increased length of the hard shrinkage selection interval the average number of correctly classified zero effects increases and the average number of correctly classified nonzero effects decreases. This is reflected by the results from the application of the HS.STD rule, based on the standard deviation, and the HS.CRI rule, based on the 95 % credible region of the estimated regression coefficients. The best results in terms of the MSE of the regression coefficients are obtained under the NMIG prior combined with the HS.STD criterion and the associated final models have a higher average number of correctly classified nonzero effects as under the HS.CRI criterion. By trend, the highest correct classification of the nonzero effects is achieved under BED 2 and the lowest under BED 3 across the update schemes of the error weights, but the difference is about 1 coefficient. Especially under the HS.STD criterion, most of the true nonzero effects are detected with only a marginal, negligible benefit in combination with the Bayesian NMIG prior. The same structure, as shown in **Table 10.1,** results under the Gaussian error assumption with exception that the highest correct classification of the nonzero effects is achieved under BED 1.

For an effect structure like (10.2) with exact zero effects, the selection-type shrinkage of the Bayesian NMIG prior in combination with the HS.IND criterion detects them all, resulting in the optimal value of 13 correctly classified zero effects. Induced by the prior tuning and the selection of the HS.IND-threshold 0.5, only the six largest effects are included in the final model which is reflected by the comparably low average number of 6.02 correctly classified nonzero coefficients and an increased MSE of the associated sparse final model. In Section 11.5 we consider variations of the HS.IND-threshold. With the therein obtained results we can conclude that a smaller value of the HS.IND-threshold, e. g. 0.1, increases the correctly classified nonzero coefficients (obvious) and the predictive performance of the associated final models (since the MSE gets closer to the MSE of the full model).

sliceR0	BED 1		BED 2		BED 3		BED 4	
	n = 500, p_x = 25		n = 500, p_x = 25		n = 500, p_x = 25		n = 500, p_x = 25	
	$\hat{\beta} \neq 0$ $\beta \neq 0$	$\hat{\beta} = 0$ $\beta = 0$	$\hat{\beta} \neq 0$ $\beta \neq 0$	$\hat{\beta} = 0$ $\beta = 0$	$\hat{\beta} \neq 0$ $\beta \neq 0$	$\hat{\beta} = 0$ $\beta = 0$	$\hat{\beta} \neq 0$ $\beta \neq 0$	$\hat{\beta} = 0$ $\beta = 0$
BEST	12	13	12	13	12	13	12	13
AFT.Step	10.02	10.70	9.58	10.66	9.08	10.40	9.52	10.58
PGM.B-HS.STD	10.64	9.00	11.24	8.60	9.86	8.78	10.14	9.40
PGM.BL-HS.STD	10.66	9.64	11.24	8.90	9.86	9.60	10.06	9.92
PGM.BR-HS.STD	10.68	9.18	11.22	8.28	9.98	8.72	10.16	9.28
PGM.BN-HS.STD	10.76	9.80	11.30	8.98	9.90	10.00	10.12	10.00
PGM.B-HS.CRI	9.28	12.54	9.92	12.24	8.44	12.54	8.68	12.38
PGM.BL-HS.CRI	9.32	12.64	9.84	12.54	8.28	12.72	8.62	12.54
PGM.BR-HS.CRI	9.30	12.46	9.88	12.30	8.40	12.50	8.74	12.40
PGM.BN-HS.CRI	9.32	12.70	9.80	12.38	8.24	12.84	8.64	12.64
PGM.BN-HS.IND	6.06	13.00	6.02	13.00	5.98	13.00	6.02	13.00

Table 10.1: Average number of correctly classified coefficients for the AFT models under baseline error distributions BED 1 to BED 4 with $n = 500$ observations after variable selection. Displayed are the results under the *"sliceR0"* update scheme for the transformed error weights. Especially $\hat{\beta} \neq 0, \beta \neq 0$ denotes the case that the estimated effect is nonzero ($\hat{\beta} \neq 0$) when the corresponding true effect is nonzero ($\beta \neq 0$), and $\hat{\beta} = 0, \beta = 0$ denotes the case that the estimated effect is zero ($\hat{\beta} = 0$) when the corresponding true effect is zero ($\beta = 0$). AFT.Step: AFT model with Gaussian error assumption.

With decreasing sample size the average number of correctly classified regression coefficients decreases for the HS.STD and HS.CRI criterion and is reduced about two regression coefficients from the simulations with $n = 500$ to $n = 100$ observations. There is hardly any variation in the classification observable for the HS.IND rule.

sliceR0	BED 2		BED 2		BED 2		BED 2	
	p_x = 25, n = 400		p_x = 25, n = 300		p_x = 25, n = 200		p_x = 25, n = 100	
	$\hat{\beta} \neq 0$ $\beta \neq 0$	$\hat{\beta} = 0$ $\beta = 0$	$\hat{\beta} \neq 0$ $\beta \neq 0$	$\hat{\beta} = 0$ $\beta = 0$	$\hat{\beta} \neq 0$ $\beta \neq 0$	$\hat{\beta} = 0$ $\beta = 0$	$\hat{\beta} \neq 0$ $\beta \neq 0$	$\hat{\beta} = 0$ $\beta = 0$
BEST	12	13	12	13	12	13	12	13
AFT.Step	9.64	10.54	9.34	10.48	8.82	10.56	8.24	9.44
PGM.B-HS.STD	11.08	8.54	10.80	8.96	10.26	9.06	9.32	8.52
PGM.BL-HS.STD	11.08	9.30	10.74	9.38	10.18	9.74	9.18	9.24
PGM.BR-HS.STD	11.14	8.58	10.82	8.88	10.28	9.04	9.32	8.60
PGM.BN-HS.STD	11.16	9.26	10.76	9.52	10.26	9.90	9.00	10.24
PGM.B-HS.CRI	9.94	12.48	9.32	12.38	9.00	12.46	7.20	11.86
PGM.BL-HS.CRI	9.92	12.54	9.32	12.62	8.90	12.68	7.20	12.54
PGM.BR-HS.CRI	9.94	12.36	9.34	12.48	8.92	12.44	7.34	11.96
PGM.BN-HS.CRI	9.92	12.52	9.30	12.64	8.88	12.74	7.28	12.84
PGM.BN-HS.IND	6.02	13.00	6.06	13.00	6.06	13.00	5.96	13.00

Table 10.2: Average number of correctly classified coefficients for the AFT models under baseline error distributions BED 2 with $n = 400$ to $n = 100$ observations after variable selection. Displayed are the results under the *"sliceR0"* update scheme for the transformed error weights. Especially $\hat{\beta} \neq 0, \beta \neq 0$ denotes the case that the estimated effect is nonzero ($\hat{\beta} \neq 0$) when the corresponding true effect is nonzero ($\beta \neq 0$), and $\hat{\beta} = 0, \beta = 0$ denotes the case that the estimated effect is zero ($\hat{\beta} = 0$) when the corresponding true effect is zero ($\beta = 0$). AFT.Step: AFT model with Gaussian error assumption.

Penalties of the linear effects

Finally, **Figure 10.21** and **Figure 10.22** show the covariate specific penalties expressed in terms of the inverse variance parameters $\tau_{\beta_j}^{-2}$ of four selected regression coefficients $\beta_1 = 3$, $\beta_9 = 1$, $\beta_{13} = 0.5$ and $\beta_{23} = 0$ under baseline error model BED 2. The figures show almost similar results for $n = 500$ and $n = 100$ observations and are associated to **Figure 10.19** and **Figure 10.20**. The Bayesian ridge penalty is constant across the regression coefficients, while the Bayesian lasso penalty is smaller for larger regression coefficients and increases for smaller effects. Under the NMIG prior the penalty is close to zero for the lager effects, resulting in a clearly reduced shrinkage, and for smaller effects the penalty increases. The results from Section 4.5 have shown that in particular under the NMIG prior the posterior mean estimate of $\tau_{\beta_j}^{-2}$ for smaller effects covers only a small range of applied penalization, so that the displayed penalties represents rather a lower bound for the penalization of small effects under the NMIG prior.

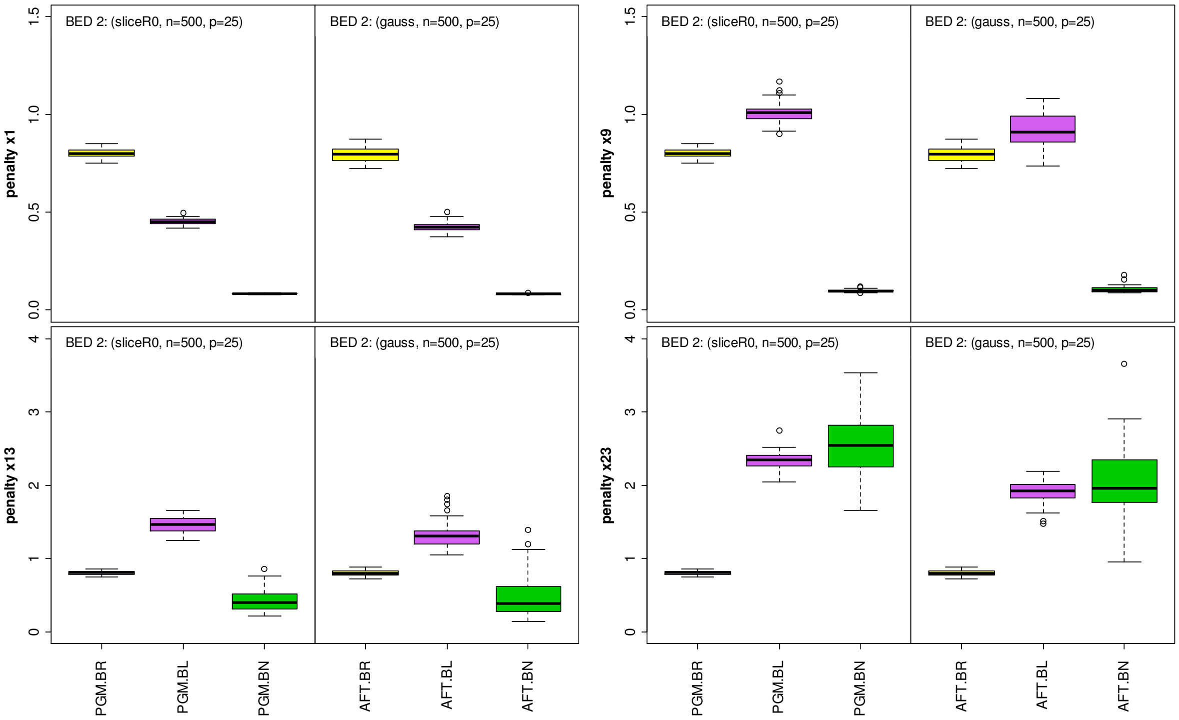


Figure 10.21: Estimates of the covariate specific penalty $\hat{\tau}_{\beta_j}^{-2}$ under the Bayesian ridge, lasso and NMIG regularization in the AFT model with baseline error distribution BED 2, $p_x = 25$ covariates and $n = 500$ observations. Displayed are the Bayesian estimates associated to the four selected estimated regression coefficients $\beta_1 = 3$ (upper left panel), $\beta_9 = 1$ (upper right panel), $\beta_{13} = 0.5$ (lower left panel) and $\beta_{23} = 0$ (lower right panel) resulting from the *"sliceR0"* update scheme of the error weights (left sides) and the Gaussian error assumption *"gauss"* (right sides).

The estimated shrinkage parameters of the Bayesian regularization priors are almost comparable under the four baseline error models and vary marginally with decreasing sample size. We obtain in the data with $n = 500$ observations median estimates about 0.45 for the Bayesian ridge and 1.7 for the Bayesian lasso shrinkage parameter. The Bayesian NMIG complexity parameter has median values about 0.29. When the sample size decreases the shrinkage parameters of the Bayesian ridge and lasso prior decrease marginally and the complexity parameter of the NMIG prior increases marginally.

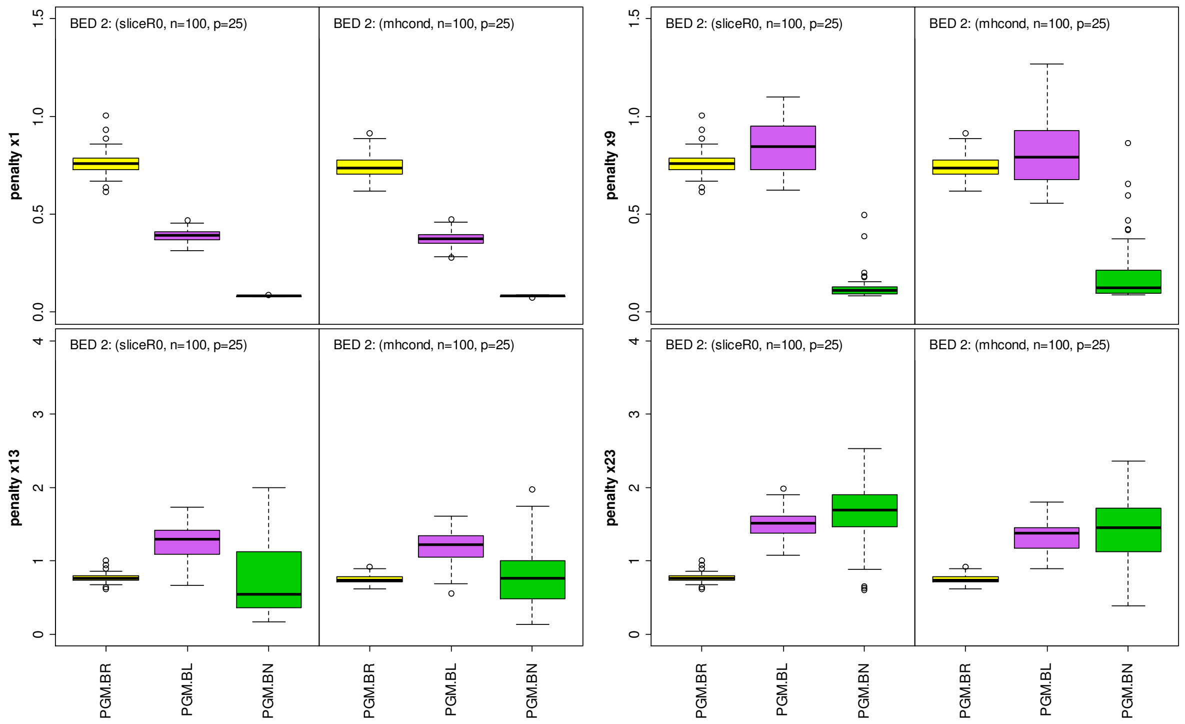


Figure 10.22: Estimates of the covariate specific penalty $\hat{\tau}_{\beta_j}^{-2}$ under the Bayesian ridge, lasso and NMIG regularization in the AFT model with baseline error distribution BED 2, $p_x = 25$ covariates and $n = 100$ observations. Displayed are the Bayesian estimates associated to the four selected estimated regression coefficients $\beta_1 = 3$ (upper left panel), $\beta_9 = 1$ (upper right panel), $\beta_{13} = 0.5$ (lower left panel) and $\beta_{23} = 0$ (lower right panel) resulting from the *"sliceR0"* (left sides) and *"mhcond"* (right sides) update scheme of the error weights.

10.3. High-dimensional predictor

In this section we consider the impact of an increased number of model parameters by modeling the $p_x = 25$ covariates in the simulation data of the previous subsection as nonlinear and by increasing the number of covariates p_x with linear effects.

10.3.1. Nonlinear predictor

In general it is not clear, if the effect of a covariate is really linear, and we can use the nonlinear modeling of covariate effects for a visual inspection of the shape of the influence on the response. If continuous covariates are modeled as nonlinear, e. g. via P-splines, the number of parameters to estimate increases clearly. In addition, the AFT model with flexible PGM error model consists of a high number of parameters to estimate the error distribution density. We investigate in this subsection the performance of the baseline density estimation and the behavior of the regularization priors in the framework, when the number of parameters exceeds the number of observations.

In the following we reconsider the simulation data of the previous Subsection 10.2, where the $p_x = 25$ linear effects are assumed to be smooth functions $f_j(\cdot)$, $j = 1,...,25$, of the covariates, i. e. we state the following predictor structure

$$\eta_i = f_1(x_{i,1}) + ... + f_{25}(x_{i,25}), \quad i = 1,...,n\,.$$

The smooth functions are modeled via Bayesian P-splines $f_j(\cdot)=\sum_{k=1}^{g_j}\alpha_{k,j}b_k(\cdot)$, where we use $g_j = 20$ cubic B-spline basis functions $b_k(\cdot)$ in each representation and the associated basis function weights $\boldsymbol{\alpha}_j = (\alpha_{1,j},...,\alpha_{20,j})'$ are equipped with second-order random walk priors to control the smoothness. In summary, the predictor consists of 500 basis function weights to estimate, and we consider this high-dimensional predictor structure under the baseline error models BED 1 to BED 4 with $n = 500$ observations and under the baseline error model BED 2 with decreasing number of observations $n = 400,300,200,100$.

Function and parameter specification

We use same methods as before in Subsection 10.2 with the given hyperparameter specification of the error priors. The hyperparameter h_{1,τ_0} of the smoothness prior for the PGM error weights is still increased under the *"mhcond"* update scheme and we use this update scheme mainly in error model BED 2. In addition the hyperparameters of the inverse gamma smoothing variance prior for the nonlinear effects are set to $h_{1,\tau_j} = h_{2,\tau_j} = 0.001$ and we use $\alpha_{1,j}^{(0)} = ... = \alpha_{20,j}^{(0)} = 0.01$ and $\tau_{\alpha_j}^{2(0)} = 1$ as initial states of the regularized components. With the nonlinear predictor we observed running times of the sampler within the range of 1 hour – 1 hour 47 min ($p_x = 25, n = 100$) and 60 min – 2 hours 20 min ($p_x = 25, n = 500$).

Results

MSE of the baseline error density

Results with n = 500 observations: **Figure 10.23** shows the MSEs of the estimated baseline error density under the four error models BED 1 to BED 4 from the replications with $n = 500$ observations. The results under the Gaussian error assumption are not visualized, but the median MSEs are given in the annotations of **Figure 10.23** and **Figure 10.24** for comparison. With exception of the Gaussian error results, the increase in the MSE is still rather moderate compared to the models with the strictly linear predictor. So far, the best performances in the error models BED 3 and BED 4 are obtained with the *"dirichlet"* update scheme. Here, with the nonlinear predictor, this result is not approved, but in BED 1 and BED 2 the MSEs are still comparable to the models with linear predictor. Under the *"slice"* update-scheme we observe an increased MSE, compared to the MSEs of the block update schemes *"mcondstep"* and *"mcondblock"*.

Results with n < 500 observations: As shown in **Figure 10.24** the reduction of the sample size increases step by step the MSE of the baseline error density. As previously observed, the stronger smoothing lets the *"mhcond"* scheme sand out with an enhanced MSE compared to the other update schemes, but in particular with $n = 100$ observations the stronger smoothness regularization is leading to a benefit, because the interquartile range is clearly decreased (compared to the other update schemes) and the median is now comparable to the median of *"sliceR0"*. Nevertheless, the performances under the *"sliceR0"* and *"mhcond"* update schemes are really poor and the MSEs act, with respect to the median, on a level comparable to the MSE under the Gaussian error assumption. As previously observed under the strictly linear predictor, the performance obtained with the *"dirichlet"* update scheme increases relative to the other update schemes, when the number of observations is reduced and the best performance for $n = 100$ observations is obtained with this update scheme. In general the MSEs of the baseline error densities cross the marked value 5e-4 between $n = 300$ and

$n = 200$ observations, under the linear predictor the crossing happens between $n = 200$ and $n = 100$ observations.

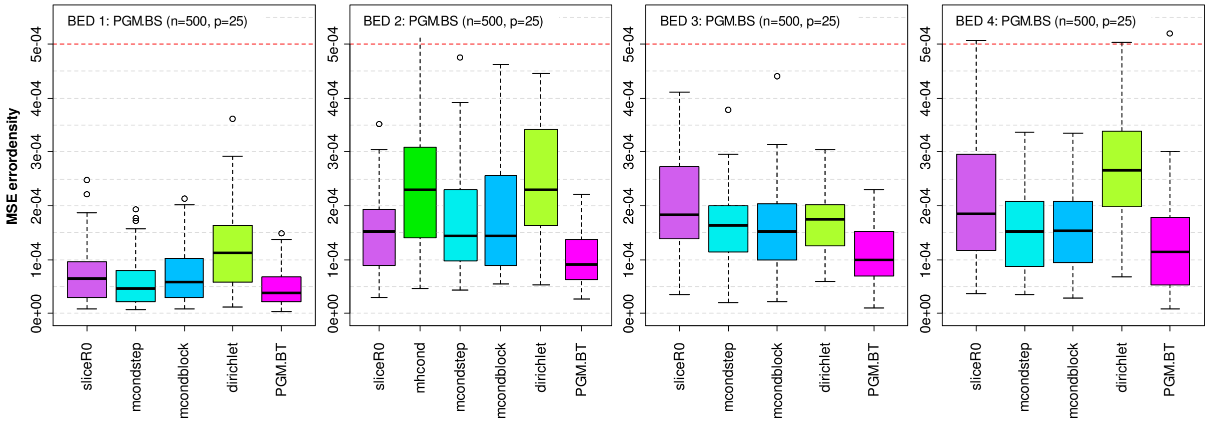


Figure 10.23: Mean squared errors of the estimated baseline error density, $\text{MSE}(\hat{f}_{Y_0})$, in the AFT model with baseline error distribution BED 1 (left side) to BED 4 (right side), $p_x = 25$ covariates and $n = 500$ observations, if the covariate effects are modeled by cubic P-splines. Displayed are the estimates under the update schemes *"sliceR0"*, *"mhcond"*, *"mcondstep"*, *"mcondblock"* and *"dirichlet"* for the error weights. PGM.BT denotes the corresponding results from the model using the true linear predictor structure. The red dotted line marks the value 5e-5. The corresponding median MSEs under the Gaussian error assumption are for are for BED 1: $\text{MSE}_{\text{AFT.BS}}(\hat{f}_{Y_0}) \approx 12.1\text{e-}4$, for BED 2: $\text{MSE}_{\text{AFT.BS}}(\hat{f}_{Y_0}) \approx 47.9\text{e-}4$, for BED 3: $\text{MSE}_{\text{AFT.BS}}(\hat{f}_{Y_0}) \approx 9.2\text{e-}4$ and for BED 4: $\text{MSE}_{\text{AFT.BS}}(\hat{f}_{Y_0}) \approx 13.7\text{e-}4$.

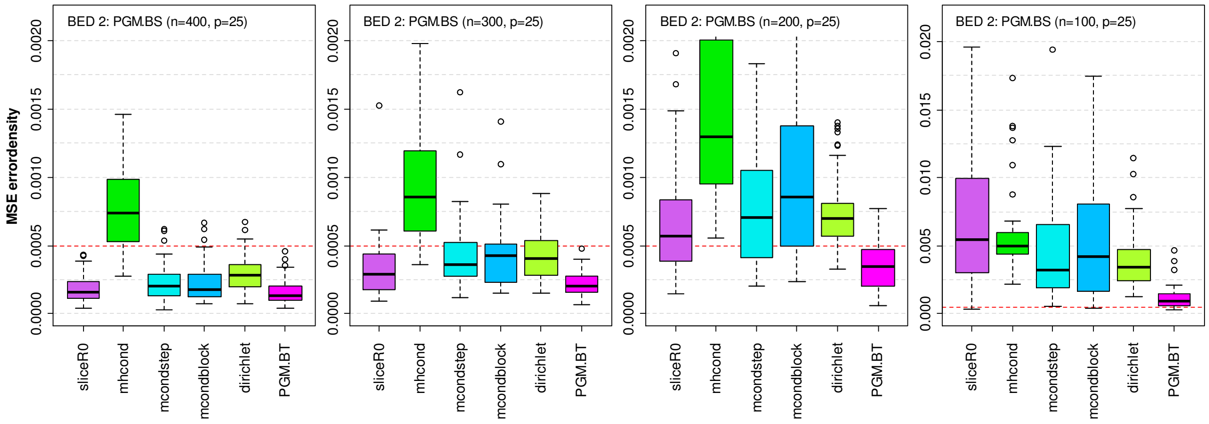


Figure 10.24: Mean squared errors of the estimated baseline error density, $\text{MSE}(\hat{f}_{Y_0})$, in the AFT model with baseline error distribution BED 2, $p_x = 25$ covariates and $n = 400$ (left side) to $n = 100$ (right side) observations, if the covariate effects are modeled by cubic P-splines. Displayed are the estimates under the update schemes *"sliceR0"*, *"mhcond"*, *"mcondstep"*, *"mcondblock"* and *"dirichlet"* for the error weights. PGM.BT denotes the corresponding results from the model using the true linear predictor structure. The red dotted line marks the value 5e-5. The corresponding median MSEs under the Gaussian error assumption with $n = 100$ observations is $\text{MSE}_{\text{AFT.BS}}(\hat{f}_{Y_0}) \approx 58.9\text{e-}4 = 0.0058$.

Baseline error density

The resulting estimates of the baseline error density under error model BED 2 with decreasing number of observations are displayed in **Figure 10.25.** If the sample size is reduced, the fit gets poorer, but with sample sizes $n \geq 200$ the information in the data is still sufficient to reflect the bimodal nature of the baseline error in the estimates. With less than $n = 200$ observations the bimodal shape of the baseline error density is rarely detected and the estimates under the *"mhcond"* update are almost always unimodal, due to the stronger smoothness regularization. Under the *"sliceR0"* update the weaker regularization enables often a bimodal estimate with extremely varying locations of the modes, and the estimates under the *"dirichlet"* scheme are rather undulating and the cavity between the two modes is not detected.

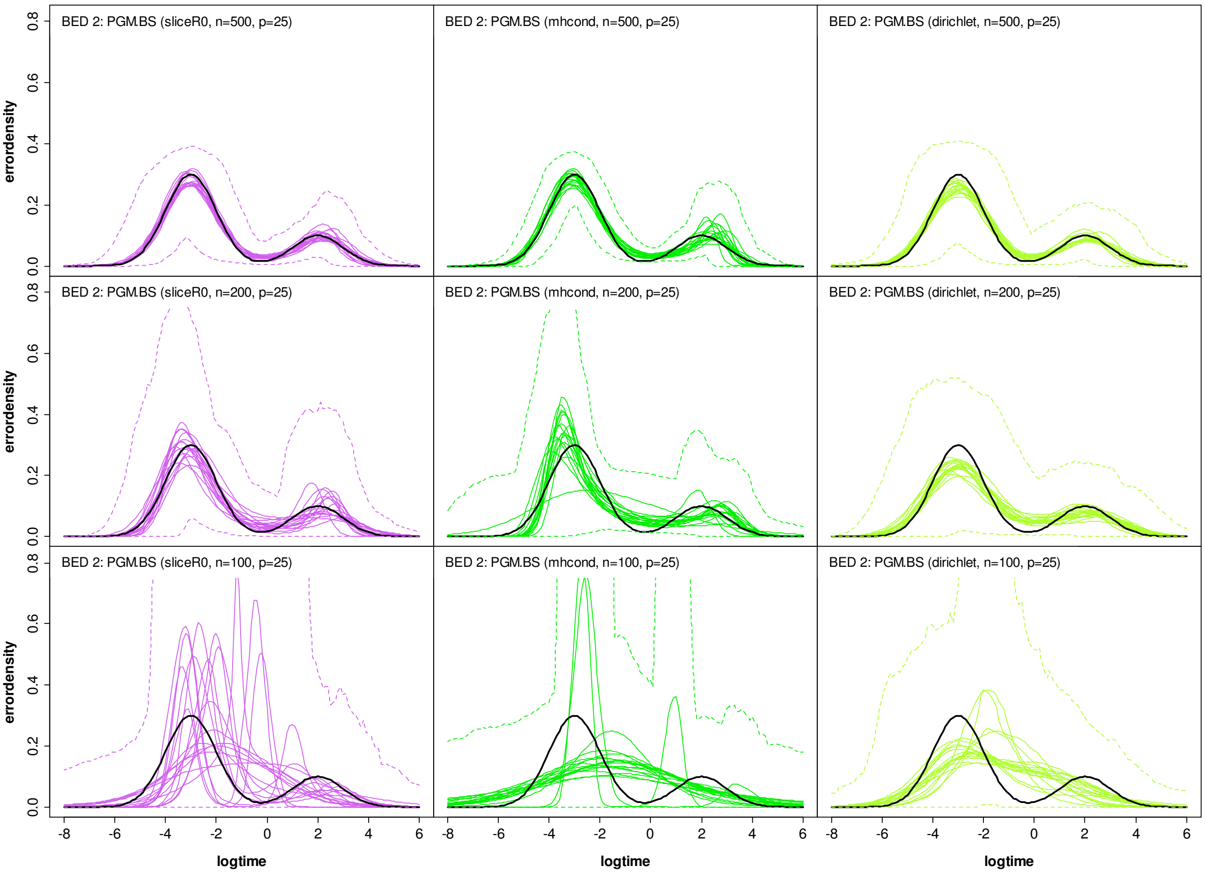


Figure 10.25: Estimated error distribution densities in the AFT model with baseline error distribution BED 2, $p_x = 25$ covariates and $n = 500$ (upper panel), $n = 200$ (middle panel) or $n = 100$ (lower panel) observations when the covariates are modeled as P-splines for selected update methods of the error weights. Displayed are the posterior mean estimates of the error density (colored lines) together with the true error density (black line). The dashed lines mark the minimum of the lower 2.5% quantile and maximum of upper 97.5% quantile in the replications.

MSE of the nonlinear effects

Figure 10.26 shows the resulting sum of the spline individual MSEs, i. e. $\text{MSE}(\hat{f}) = \sum_{j=1}^{25} \text{MSE}(\hat{f}_j)$, under the error models BED 1 to BED 4 with $n = 500$ observations and **Figure 10.27** shows the corresponding results under error model BED 2 with decreasing number of observations. The MSE is clearly increased compared to the strictly linear modeling of the effects, as indicated by the increased differences to the $\text{MSE}(\hat{\boldsymbol{\beta}})$ of the model with the true predictor structure (PGM.BT). As previously observed with the strictly linear predictor, the differences in the performance of the baseline error density estimation, caused by the various update schemes, are again less pronounced in terms of the MSEs of the regression coefficients. Even, the MSEs of the *"dirichlet"* and *"mhcond"* update scheme are almost comparable to the other update schemes. If the sample size decreases, the performance of the *"dirichlet"* scheme is again improved in comparison to the remaining update schemes and the performance is best in the case of $n = 100$ observations. In particular, for the sample size of $n = 100$ the MSE under the Gaussian error assumption is almost in the same range as the MSE with the PGM error. In the case of $n = 200$ observations the spline MSE ($\text{MSE}(\hat{f})$) is comparable to the MSE with the unregularized linear predictor for $n = 100$ observations.

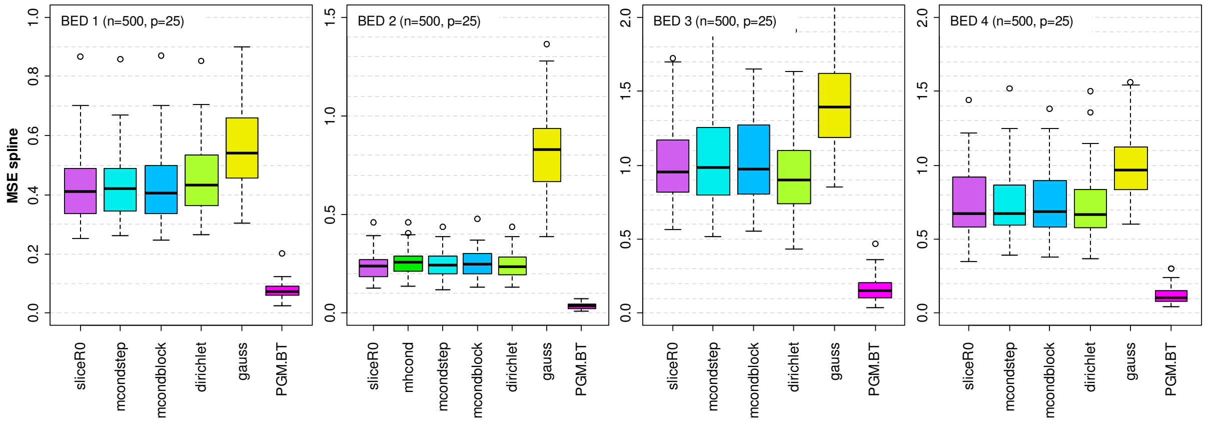


Figure 10.26: Sum of the mean squared errors of the nonlinear effects, $\text{MSE}(\hat{f})$, in the AFT model with baseline error distribution BED 1 (left side) to BED 4 (right side), $p_x = 25$ covariates and $n = 500$ observations, where the effects are modeled by cubic P-splines. Displayed are the estimates under the update schemes *"sliceR0"*, *"mhcond"*, *"mcondstep"*, *"mcondblock"* and *"dirichlet"* for the error weights and the Gaussian error assumption (*"gauss"*). PGM.BT denotes the corresponding results from the model using the true linear predictor structure.

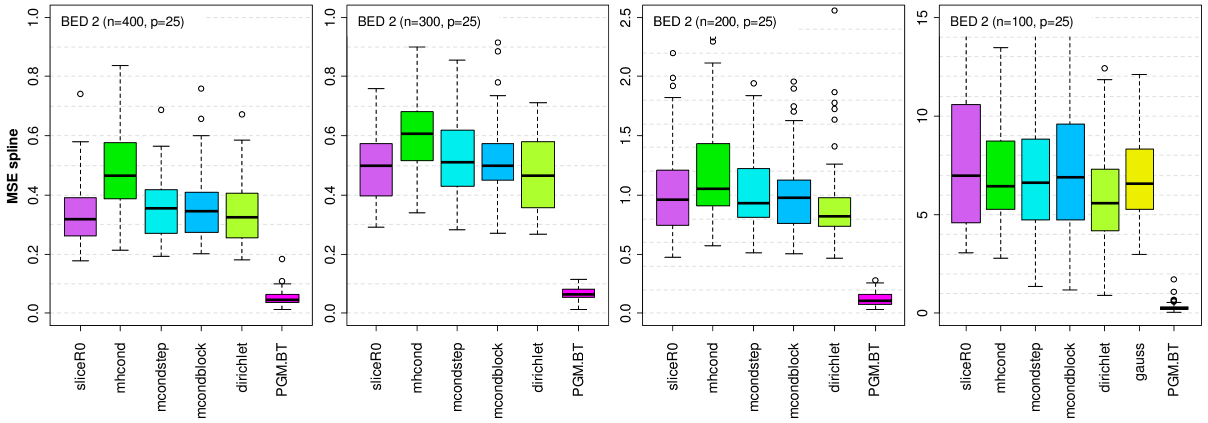


Figure 10.27: Sum of the mean squared errors of the nonlinear effects, $\text{MSE}(\hat{f})$, in the AFT model with baseline error distribution BED 2, $p_x = 25$ covariates and $n = 400$ (left side) to $n = 100$ (right side) observations, where the effects are modeled by cubic P-splines. Displayed are the estimates under the update schemes *"sliceR0"*, *"mhcond"*, *"mcondstep"*, *"mcondblock"* and *"dirichlet"* for the error weights and the Gaussian error assumption (*"gauss"*). PGM.BT denotes the corresponding results from the model using the true linear predictor structure.

Nonlinear effects

Finally, **Figure 10.28** shows the nonlinear function estimates $\hat{f}_{14}$, $\hat{f}_{18}$, and $\hat{f}_{25}$ in the AFT model with baseline error distribution BED 2 and $n = 300$ observations (upper panel) and $n = 100$ observations (lower panel) for one replicated dataset. The black dashed line marks the associated true linear effect.

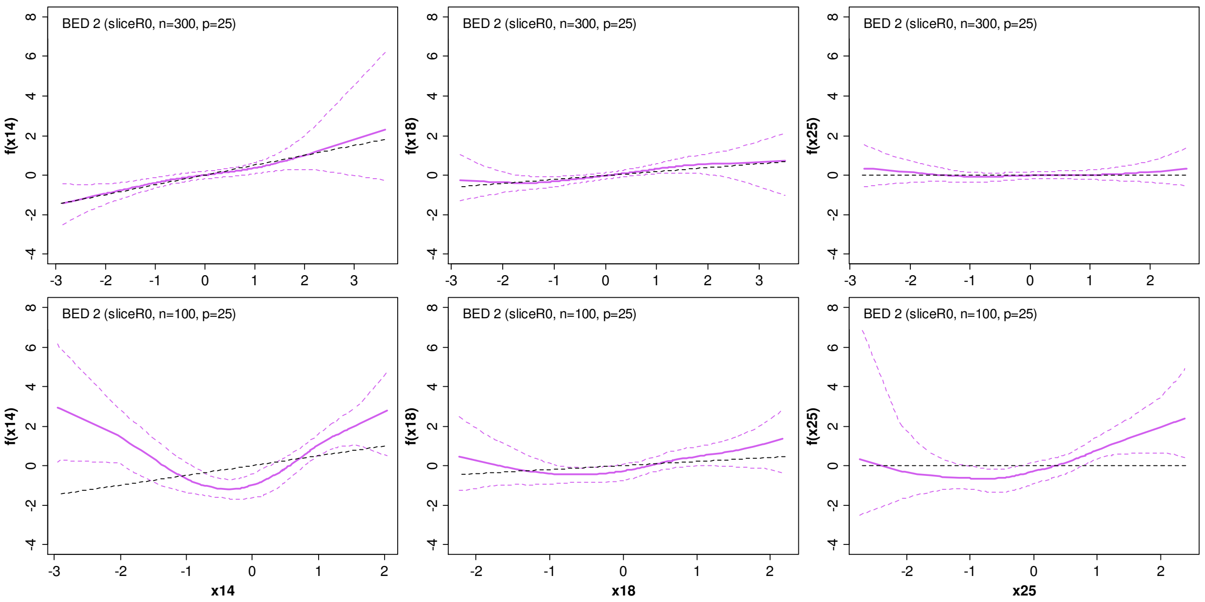


Figure 10.28: Estimations of the nonlinear effects of the covariates x_{14} (first column), x_{18} (second column), and x_{25} (third column), in the AFT model with baseline error distribution BED 2, $p_x = 25$ covariates and $n = 300$ (upper panel) and $n = 100$ (lower panel) observations. Displayed are the posterior mean estimates of the coefficients (colored solid lines) together with the corresponding 95% pointwise credible bands (colored dashed lines) of one selected dataset and the true effect (black dotted line).

10.3.2. Bayesian NMIG prior

Data generation

In this section we investigate the high-dimensional case, where the number of covariates is increased with respect to the previous sections. We consider in particular the AFT model with baseline error model BED 2 and the number of covariates increases from $p_x = 100$ to $p_x = 600$. The covariates, log-survival times and the 25% censoring times are generated as described before in the Subsection 10.2. The so far used vector $\boldsymbol{\beta} = (3,3,0,0,2,2,0,0,1,1,0,0,0.5,0.5,0,0,0.2,0.2,0,0,0.1,0.1,0,0,0)'$ is pasted back-to-back repeatedly until the desired number of effects p_x is attained. Particularly we use the predictor

$$\eta_i = x_{i,1}\beta_1 + x_{i,2}\beta_2 + \ldots + x_{i,p_x}\beta_{p_x}, \quad i = 1,\ldots,n, \tag{10.3}$$

with

- $p_x = 250, 300, 400, 500, 600$ in combination with $n = 500$ observations,
- $p_x = 100, 200, 300$ in combination with $n = 200$ observations,
- $p_x = 400$ in combination with $n = 300$ observations.

The regularization of the linear effects is carried out by means of the Bayesian NMIG prior and we summarize in the following the results for the combinations $n = 500, p_x = 250$ and $n = 200, p_x = 100$, where the number of covariates is still smaller than the number of observations, and for the combinations $n = 200, p_x = 300$ and $n = 300, p_x = 400$, where the number of covariates exceeds the number of observations.

Function and parameter specification

We use the function and parameter settings as given in the previous Sections 10.1 and 10.2. For the high-dimensional combinations $n = 200, p_x = 300$ and $n = 300, p_x = 400$ we observe under the so far used constellation of the hyperparameters a critical convergence in the sample paths of the parameters associated to baseline error density, and a further adaption of the smoothing prior has also shown no improvement. In general the paths of regression coefficients and shrinkage prior components are less concerned from the convergence problems. However, we summarize shortly the results under the so far used prior tuning and interpret the associated results carefully. For some of the combinations we use additional runs without the standardization of the error distribution within the sampler (`scalebasis=FALSE`) and we present these results for the respective combinations. With the higher-dimensional predictors we observed running times of the sampler about 30 min ($p_x = 100, n = 200$), 50 min ($p_x = 300, n = 200$), 1h 40min ($p_x = 400, n = 300$) and 1h 30min ($p_x = 250, n = 500$).

Results

Baseline error location and scale

If the number of covariates is increased with respect to the number of observations, we observe in general a larger deviation of the estimated location and scale parameters from the true values of the underlying baseline error distribution. Until now, the option to standardize the error density during the iterations of the MCMC sampler has shown no obvious impact onto the results. Here, with an increased number of covariates, the standardization is leading to a higher concentration of the estimated moments around true moments of the baseline error density, as shown in **Figure 10.29**. The larger deviations in the high-dimensional combination $n = 300, p_x = 400$ are explained by the weak convergence of the error parameters.

MSE of the baseline error density

As observed in the previous sections the fit to the error moments does not affect the performance of the baseline error density, **Figure 10.30**. With respect to comparable sample sizes, the MSEs of the error density clearly increase with the increasing number of covariates in the model. As previously noticed the *"dirichlet"* update scheme performs well (relative to the other update schemes) in situations with a low sample size, and also here with increased numbers of covariates the best performances are achieved with this unregularized update scheme. In the simulations with $n = 200$ observations the MSE of the *"dirichlet"* update scheme is comparable to the MSE of the PGM error model with the true predictor structure (PGM.BT).

Baseline error density

Figure 10.31 shows the estimates of the baseline error density for the combinations $p_x = 250, n = 500$ (upper panel) and $n = 300, p_x = 400$ (lower panel). Even in the cases with more covariates than observations the estimates, e. g. under the *"sliceR0"* or *"mcondblock"* update scheme, are still smooth and a stronger regularization of the smoothness (results not shown) can not counterbalance the obvious lack of fit due to the weak information in the data, which is also observable for the models with the true predictor structure.

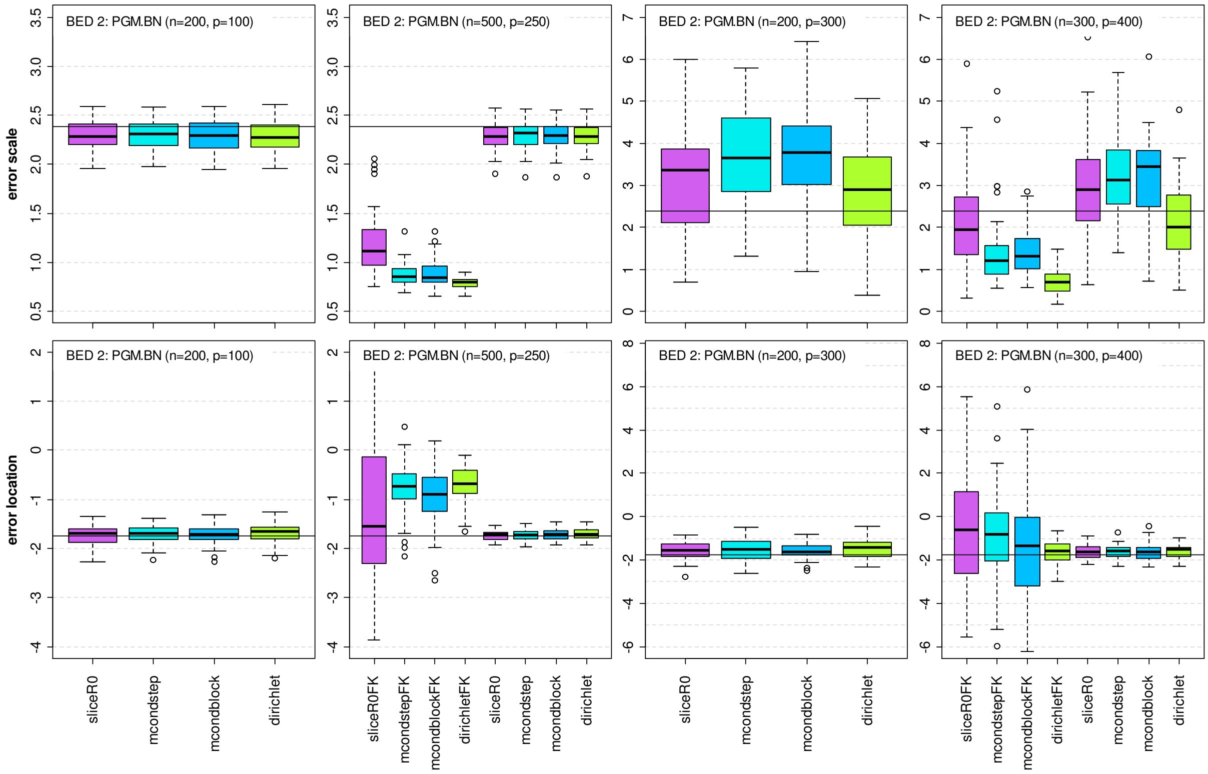


Figure 10.29: Estimated scale (upper panel) and location parameter (lower panel) in the AFT model with baseline error distribution BED 2, $p_x = 100$ covariates and $n = 200$ observations (first column), $p_x = 250$ covariates and $n = 500$ observations (second column), $p_x = 300$ covariates and $n = 200$ observations (third column) and $p_x = 400$ covariates and $n = 300$ observations (last column) under the Bayesian NMIG regularization of the regression coefficients. Displayed are the estimates under various update schemes for the error weights. The black horizontal lines mark the true scale σ_{Y_0} and location μ_{Y_0} of the associated baseline error distribution.

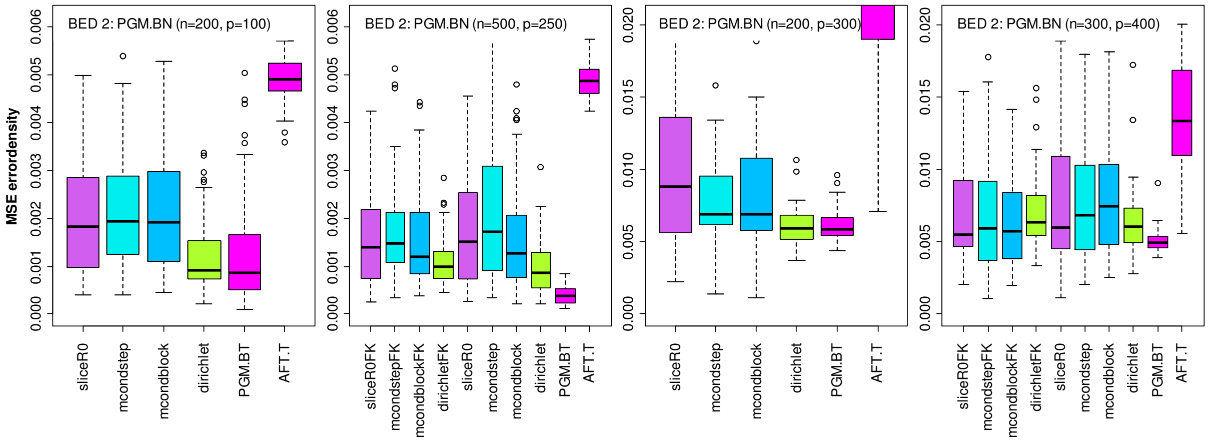


Figure 10.30: Mean squared errors of the estimated baseline error density, $\mathrm{MSE}(\hat{f}_{Y_0})$, in the AFT model with baseline error distribution BED 2, $p_x = 100$ covariates and $n = 200$ observations (first panel), $p_x = 250$ covariates and $n = 500$ observations (second panel), $p_x = 300$ covariates and $n = 200$ observations (third panel) and $p_x = 400$ covariates and $n = 300$ observations (last panel) under the Bayesian NMIG regularization of the regression coefficients. Displayed are the estimates under various update schemes for the error weights. AFT.T denotes the results from the frequentist AFT model with Gaussian error using the true predictor structure and PGM.BT denotes the corresponding results from the Bayesian AFT model with PGM error.

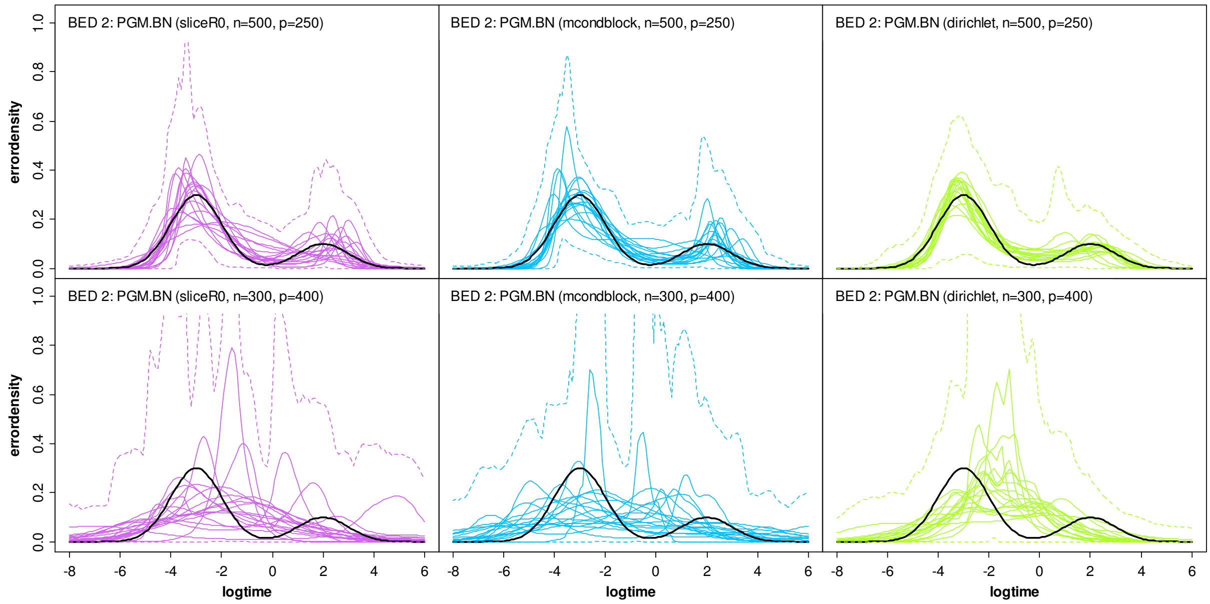


Figure 10.31: Estimated baseline error densities in the AFT model with baseline error distribution BED 2, $p_x = 250$ covariates and $n = 500$ observations (upper panel) and $p_x = 400$ covariates and $n = 300$ observations (lower panel) under the Bayesian NMIG regularization of the regression coefficients. Displayed are the posterior mean estimates of the error density (colored lines) together with the true error density (black line). The dashed lines mark the minimum of the lower 2.5% quantile and maximum of upper 97.5% quantile in the replications.

MSE of the regression coefficients

The MSEs of the estimated regression coefficients, $\text{MSE}(\hat{\boldsymbol{\beta}})$, are standardized to reflect the portion of 25 regression coefficients, e. g. in the case of $p_x = 100$ covariates we divide the MSE of the regression coefficients with 4. Under comparable sample sizes the MSE of the estimated regression coefficients increases with increasing number of covariates in the model. As previously observed, the loss of performance in the error density estimation is also reflected in the level of the performance of the regression coefficient estimates, shown in **Figure 10.32.**

With the given structure of the underlying effects (10.2) we do not reach MSEs comparable to the model using the true predictor structure (PGM.BT), as already observed under the NMIG prior with the low-dimensional linear predictor. If in addition the hard shrinkage selection rules are applied to find sparse final models, compare **Figure 10.33**, we observe a similar, but more pronounced trend of the resulting MSEs as in the previous Subsection 10.2.

For the high-dimensional cases, with still more observations than covariates ($n > p_x$, **Figure 10.33** upper panel), the MSE increases clearly from the HS.STD over the HS.CRI to HS.IND selection rule. The low performance trend of HS.IND criterion is reversed in the high-dimensional case, with less observations than covariates ($n < p_x$, **Figure 10.33** lower panel). In that case, the range of the associated MSE is comparable close to the MSEs resulting form the HS.STD rule and we will find similar results with the high-dimensional simulations in the CRR model, compare Section 11.4. But in general, all hard shrinkage rules yield sparse final models with lower performance than those including the full covariate set, especially in the $n < p_x$ setting. **Table 10.3** shows the associated classification results of the estimated effects under the hard shrinkage variable selection.

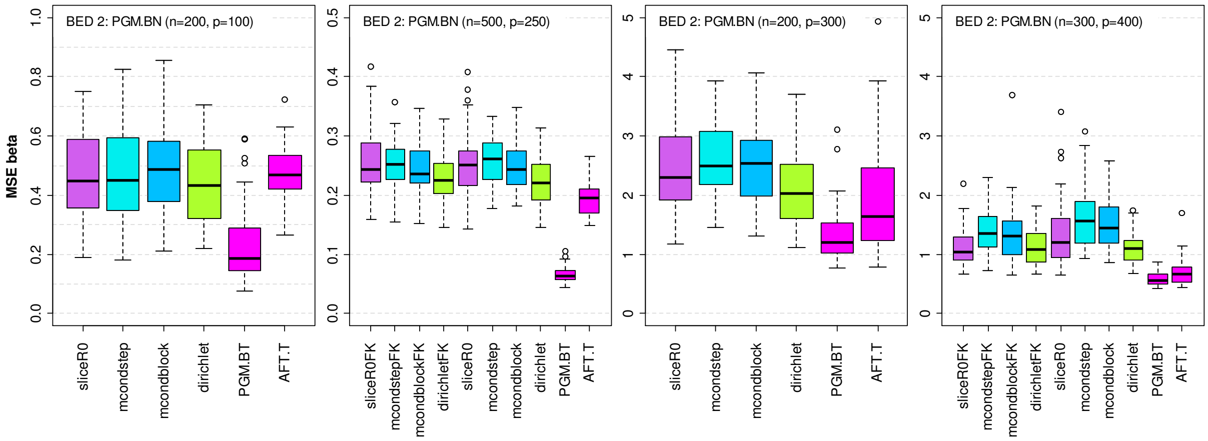


Figure 10.32: Mean squared errors of the estimated regression coefficients, MSE($\hat{\boldsymbol{\beta}}$), in the AFT model with baseline error distribution BED 2, $p_x = 100$ covariates and $n = 200$ observations (first panel), $p_x = 250$ covariates and $n = 500$ observations (second panel), $p_x = 300$ covariates and $n = 200$ observations (third panel) and $p_x = 400$ covariates and $n = 300$ observations (last panel) under the Bayesian NMIG regularization of the regression coefficients. Displayed are the estimates under various update schemes for the error weights. AFT.T denotes the results from the frequentist AFT model with Gaussian error using the true predictor structure and PGM.BT denotes the corresponding results from the Bayesian AFT model with PGM error.

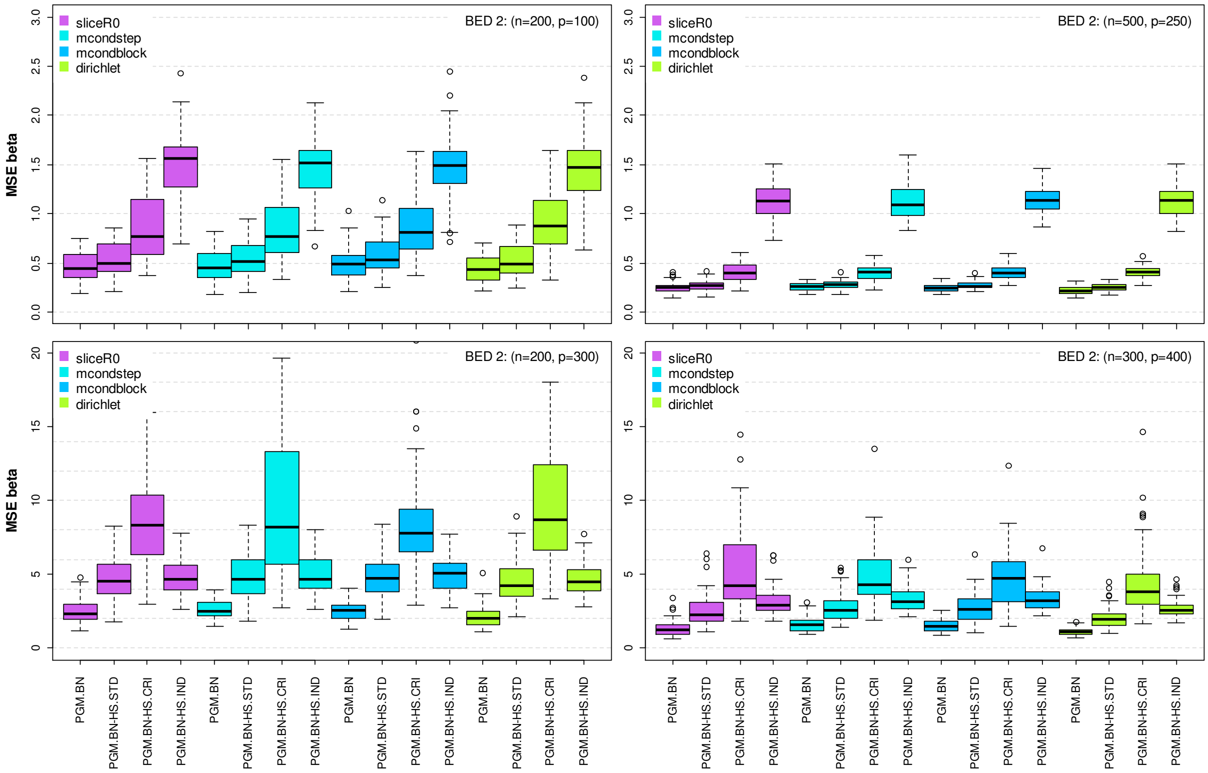


Figure 10.33: Mean squared errors of the estimated regression coefficients, MSE($\hat{\boldsymbol{\beta}}$), in the AFT model with baseline error distribution BED 2 together with the MSEs resulting from the hard shrinkage variable selection criteria. Displayed are the MSEs under various update schemes for the error weights with $p_x = 100$ covariates and $n = 200$ observations (upper left panel), $p_x = 250$ covariates and $n = 500$ observations (upper right panel), $p_x = 300$ covariates and $n = 200$ observations (lower left panel) and $p_x = 400$ covariates and $n = 300$ observations (lower right panel).

NMIG indicators

Concordantly, the constellation between the number of observations and the number of covariates affects the classification of the covariate specific binary variance component I_j to the values v_0 and v_1. The lack of definition, that comes along with an increased number of covariates or reduced number of observations, is indicated by larger interquartile ranges of the estimated inclusion probabilities, based on posterior relative frequencies of the Bayesian NMIG indicator variable value $I_j = v_1$, followed by and an increased number of extreme values.

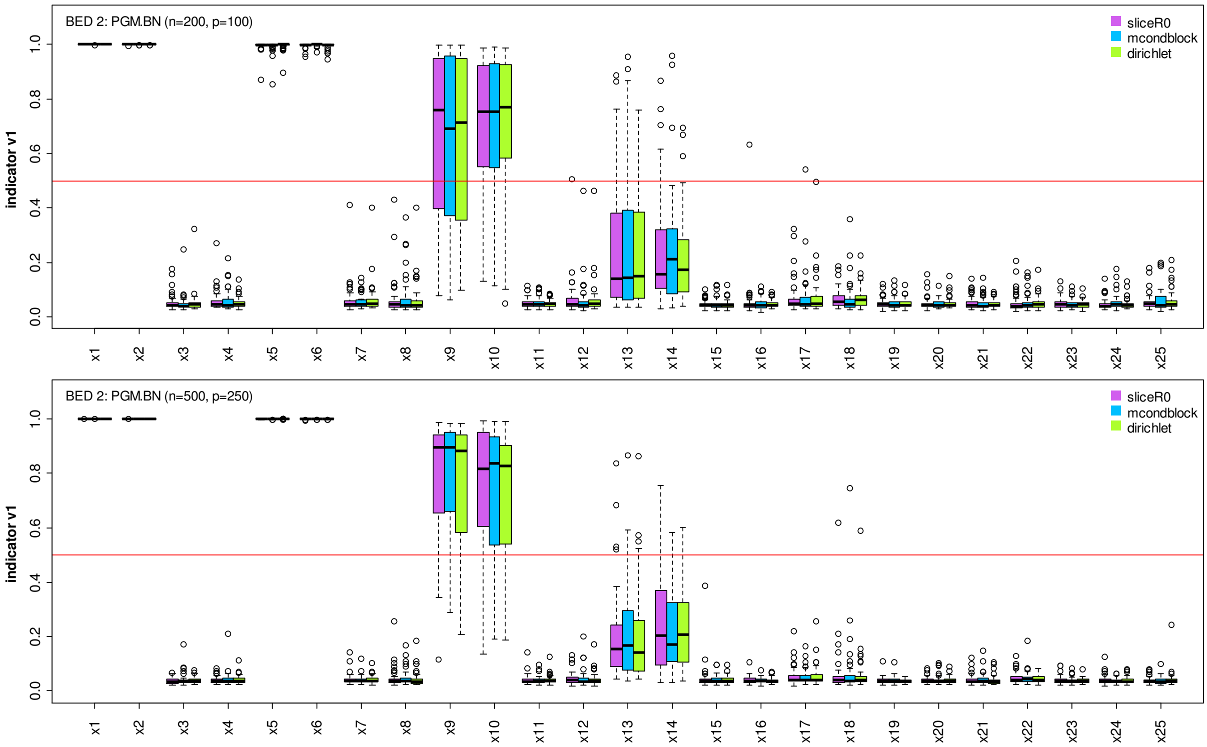


Figure 10.34: Estimated inclusion probabilities based on posterior relative frequencies of the Bayesian NMIG indicator variable value $I_j = v_1$ in the AFT model with baseline error distribution BED 2, $p_x = 100$ covariates and $n = 200$ observations (upper panel) or $p_x = 250$ covariates and $n = 500$ observations (lower panel). Displayed are the frequencies corresponding to the selected effects of the covariates annotated at the x-axis via three different update schemes for the error weights. The red horizontal line marks the cut off value 0.5 of the hard shrinkage selection criterion HS.IND.

If still more observations than covariates are available, ($n > p_x$), the selected threshold 0.5, used in the HS.IND criterion, separates the effects that are larger or equal than $\beta = 1$ from those smaller or equal than $\beta = 0.5$, compare **Figure 10.34**. As in the simulations with the low-dimensional predictor, covariates with effects $\beta_{13} = \beta_{14} = 0.5$ have still higher inclusion probabilities compared to the inclusion probabilities of the covariates with smaller or the zero effects. But, as shown in **Figure 10.35**, this separation gets blurred if the sample size is smaller than the number of covariates ($n < p_x$) and is shifted to larger coefficients. Due to the decreased inclusion probabilities, the HS.IND-threshold separates now covariate effects $\beta = 2$ and $\beta = 1$, and in particular in the cases with lower sample sizes, like $n = 200$ (upper panel), this effect is pushed since the inclusion probabilities of the larger effects $\beta = 2$ are further shifted towards the cut of value of 0.5. In addition the inclusion probabilities of the covariates with effects $\beta_{13} = \beta_{14} = 0.5$ do not longer differ from the inclusion probabilities of the covariates with smaller or the zero effects. With respect to the results from Section 11.5 the decrease of the HS.IND-threshold value increases the number of correctly classified nonzero effects and improves the performance of the sparse final models obtained with the HS.IND criterion. The improvement is mainly caused by the inclusion of the covariates with larger effects, since the inclusion probabilities of the covariates with smaller effects are not distinguishable from the inclusion probabilities of the covariates with zero effects.

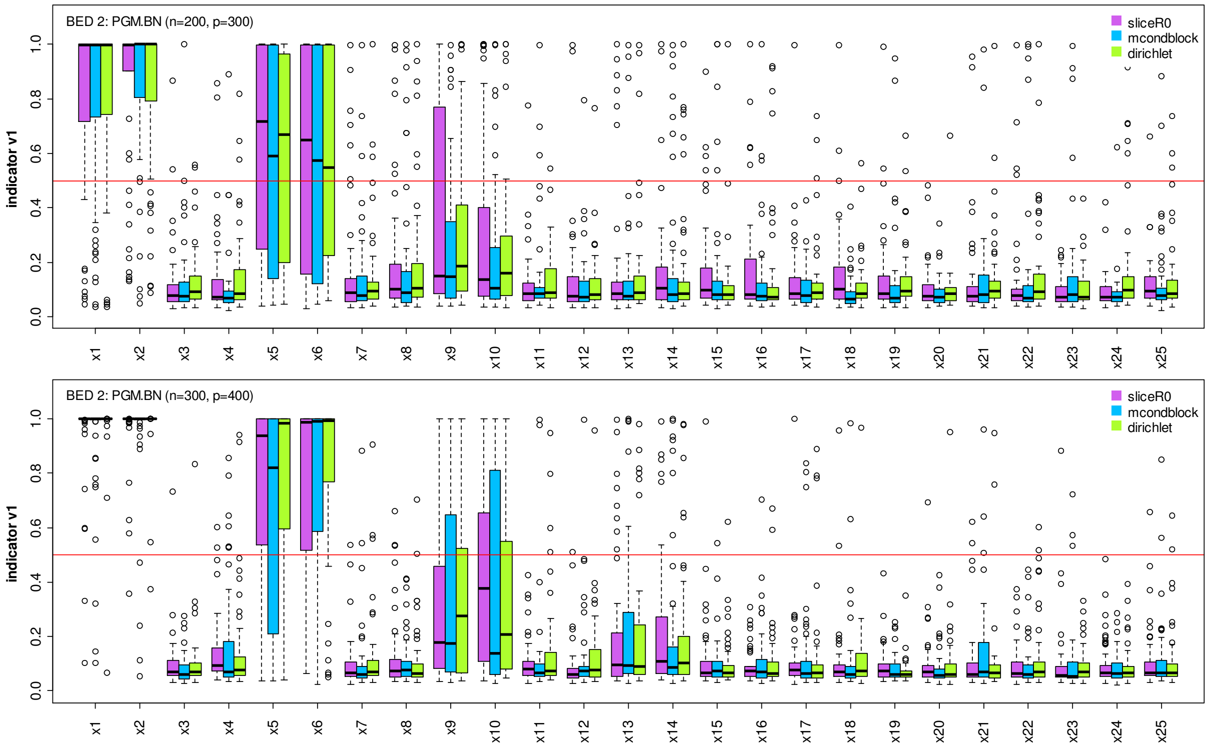


Figure 10.35: Estimated inclusion probabilities based on posterior relative frequencies of the Bayesian NMIG indicator variable value $I_j = v_1$ in the AFT model with baseline error distribution BED 2, $p_x = 300$ covariates and $n = 200$ (upper panel) or $p_x = 400$ covariates and $n = 300$ (lower panel) observations. Displayed are the frequencies corresponding to the selected effects of the covariates annotated at the x-axis via three different update schemes for the error weights. The red horizontal line marks the cut off value 0.5 of the hard shrinkage selection criterion HS.IND.

Classification

As indicated by the posterior inclusion probabilities, the HS.IND selection rule detects almost always all true zero effects reliable and as a consequence the proportion of correctly classified zero effects ($\hat{\beta}=0,\beta=0$) matches well the optimal value of 0.52, compare **Table 10.3**. Mainly affected by the transition from the $n > p_x$ to the $n \leq p_x$ case, is the proportion of correctly classified nonzero effects ($\hat{\beta}\neq 0,\beta\neq 0$) which decreases from 0.23 ($p_x = 100$) to 0.14 ($p_x = 300$) with sample size $n = 200$ and from 0.24 ($p_x = 250$) to 0.21 ($p_x = 600$) with sample size $n = 500$. Since the proportion of correctly classified zero effects is almost constant the number of misclassifications increase.

Linear effects

Lower inclusion probabilities enhance the relative frequencies of the Bayesian NMIG indicator values $I_j = v_0$ and induce a stronger regularization of the associated effects β_j. The impact of the stronger regularization of the regression coefficients is shown in **Figure 10.36** by means of four selected effects with different size in the data with $n = 200$ observations, where at the right sides the predictor includes $p_x = 100$ and at the left sides $p_x = 300$ covariates. If we switch from the $n > p_x$ (left side) to the $n < p_x$ (right side) case, we see the increased shrinkage of the larger regression coefficients. The shrinkage is clearly increased for the effects $\beta_9 = 1$ and $\beta_{13} = 0.5$, where the estimates are very close to zero.

sliceR0	BED 2 n = 200, p_x = 100		BED 2 n = 500, p_x = 250		BED 2 n = 200, p_x = 300		BED 2 n = 300, p_x = 400	
	$\hat\beta \neq 0$ $\beta \neq 0$	$\hat\beta = 0$ $\beta = 0$	$\hat\beta \neq 0$ $\beta \neq 0$	$\hat\beta = 0$ $\beta = 0$	$\hat\beta \neq 0$ $\beta \neq 0$	$\hat\beta = 0$ $\beta = 0$	$\hat\beta \neq 0$ $\beta \neq 0$	$\hat\beta = 0$ $\beta = 0$
BEST	0.48	0.52	0.48	0.52	0.48	0.52	0.48	0.52
AFT.Step	0.38	0.29	-	-	-	-	-	-
PGM.BN-HS.STD	0.36	0.42	0.39	0.39	0.15	0.48	0.20	0.48
PGM.BN-HS.CRI	0.28	0.51	0.32	0.50	0.11	0.51	0.15	0.51
PGM.BN-HS.IND	0.23	0.52	0.24	0.52	0.14	0.49	0.17	0.51

Table 10.3: Average fraction of correctly classified coefficients for the AFT models under baseline error distributions BED 2 after variable selection. Displayed are the results under the *"sliceR0"* update scheme. Especially $\hat\beta \neq 0, \beta \neq 0$ denotes the case that the estimated effect is nonzero ($\hat\beta \neq 0$) when the corresponding true effect is nonzero ($\beta \neq 0$), and $\hat\beta = 0, \beta = 0$ denotes the case that the estimated effect is zero ($\hat\beta = 0$) when the corresponding true effect is zero ($\beta = 0$). AFT.Step: AFT model with Gaussian error assumption.

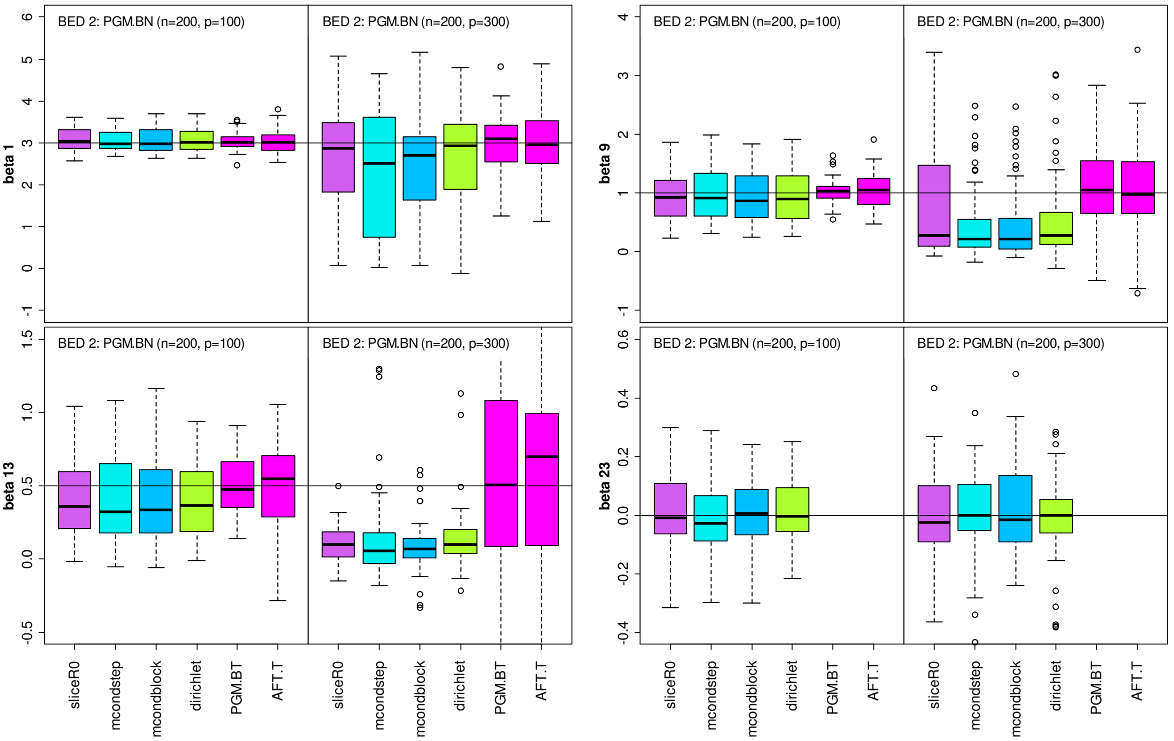


Figure 10.36: Regression coefficient estimates $\hat{\boldsymbol\beta}$ for four selected estimated regression coefficients $\beta_1 = 3$ (upper left panel), $\beta_9 = 1$ (upper right panel), $\beta_{13} = 0.5$ (lower left panel) and $\beta_{23} = 0$ (lower right panel) in the AFT model with baseline error distribution BED 2, $p_x = 100$ covariates and $n = 200$ observations (left sides) or $p_x = 300$ covariates and $n = 200$ observations (right sides). The black horizontal lines mark the true values of the regression coefficients. AFT.T denotes the results from the frequentist AFT model with Gaussian error using the true predictor structure and PGM.BT denotes the corresponding results from the Bayesian AFT model with PGM error.

Penalties of the linear effects and shrinkage parameters

Figure 10.37 shows the associated covariate specific penalties expressed in terms of the inverse variance parameters $\tau_{\beta_j}^{-2}$ that indicate also the increased shrinkage if the number of covariates in the predictor increases. The increased shrinkage causes also a decrease in the estimated complexity parameter ω as shown in **Figure 10.38.** The adaption of the hyperparameters $h_{1,\omega}$ and $h_{2,\omega}$ to force a higher model complexity ω leads to higher inclusion probabilities for all covariates and does not solve the problem that especially in higher-dimensional covariate cases the inclusion probabilities of moderate effects are not separable from the inclusion probabilities covariates with smaller or zero effects, compare also Section 11.4.

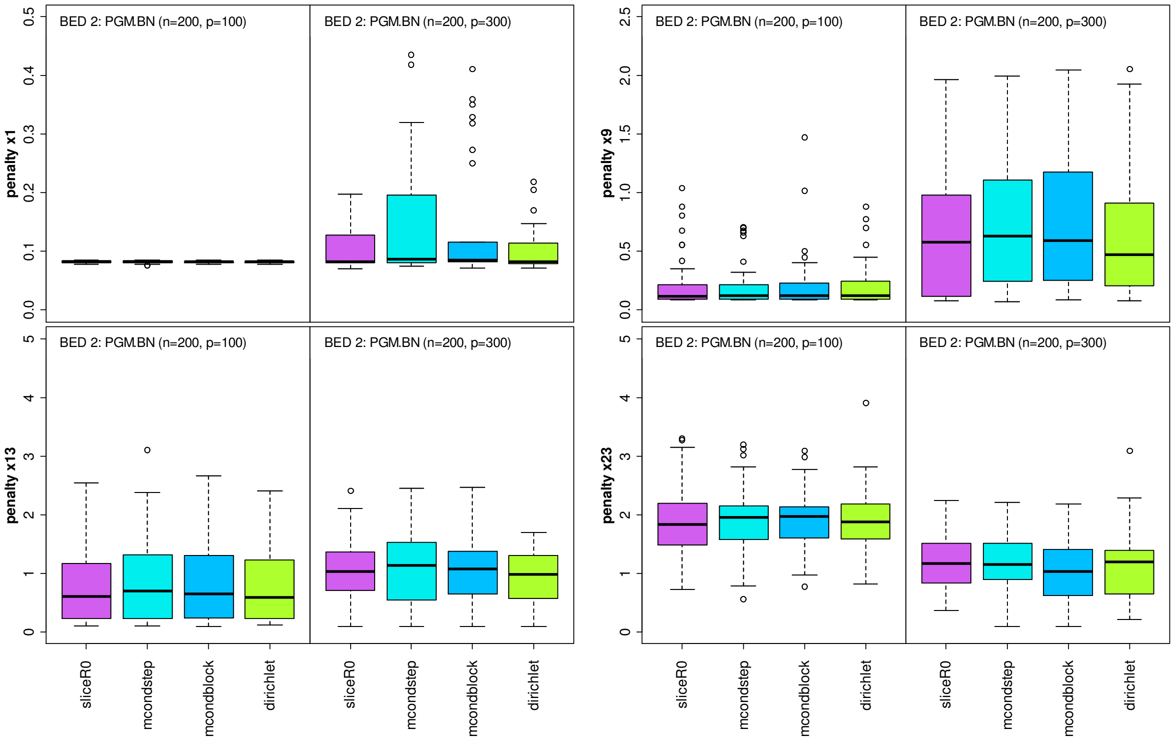


Figure 10.37: Covariate specific penalty $\hat{\tau}_{\beta_j}^{-2}$ for the Bayesian NMIG prior in the AFT model with baseline error distribution BED 2, $p_x = 100$ covariates and $n = 200$ observations (left sides) or $p_x = 300$ covariates and $n = 200$ observations (right sides). Displayed are the Bayesian estimates associated to the four selected estimated regression coefficients $\beta_1 = 3$ (upper left panel), $\beta_9 = 1$ (upper right panel), $\beta_{13} = 0.5$ (lower left panel) and $\beta_{23} = 0$ (lower right panel) under various update schemes for the error weights.

Final remarks

In summary, the performance of the AFT is considered in terms of the performance of the baseline error density and the predictor. We have seen that the performance of both model components is connected and an improved performance of the baseline error induces an improved performance of the predictor and vice versa.

The several strategies applied in the estimation of the baseline error have shown limited effects on the performance. Across the four used baseline error models none of the used update schemes of the error mixture weights has shown superiority. We have seen that the unregularized *"dirichlet"* update scheme performs very well (and best compared to the other update schemes) in some of the four error models and that the performance is improved relative to the other update schemes, when the sample size decreases or the number of covariates increases. Nevertheless, due to the lack of information in the higher-dimensional cases with lower sample sizes, the estimated densities do not reflect the underlying error density, even if the performance of *"dirichlet"* update scheme is higher than the performance of the other methods. Due to the long running times of the sampler, the enhanced tuning effort and the required increased regularization of the smoothness that causes a loss of performance we found no benefits in using the Metropolis-Hastings based update schemes. Finally, the low acceptance rates for the *"mcondstep"* scheme have shown no impact on the performance of the model component estimates. Also the standardization of mixture error density within the sampler has shown no benefit for the estimation of the model components. Possibly in other frameworks, like e. g. in quantile regression, where the scale parameter is modeled covariate-dependent $(\sigma_i(\mathbf{x}) = \mathbf{x}_i'\zeta)$ in combination with informative priors for ζ it may be of any importance (e. g. with respect to the hyperparameter specification, compare Section 6.2.1).

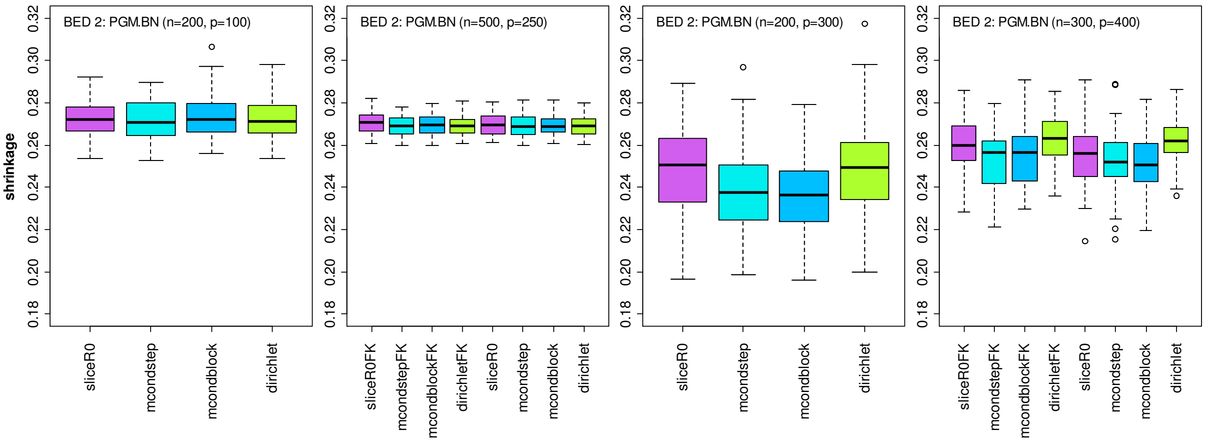


Figure 10.38: Estimated shrinkage parameter $\hat{\omega}$ in the AFT model with baseline error distribution BED 2, p_x = 100 covariates and n = 200 observations (left figure), p_x = 250 covariates and n = 500 observations (second figure), p_x = 200 covariates and n = 300 observations (third figure) and p_x = 400 covariates and n = 300 observations (right figure) under the Bayesian NMIG regularization of the regression coefficients. Displayed are the estimates under various update schemes for the error weights.

If we consider the estimation of the predictor components, we have seen that the application of the regularization priors to the linear covariate effects increases in general the performance compared to the unregularized estimation. In particular, the best performance results (with respect to the used effect model) are obtained with the specific shrinkage of the Bayesian NMIG prior, even with enough information in the data, where the impact of the likelihood dominates the impact of the regularization priors on the estimates, and even in the models with the Gaussian error assumption, where the error model is miss-specified. But, in the miss-specified Gaussian error model, the improved performance resulting from the regularization of the linear effects has only a marginal impact on the performance of the error density. In general, variable selection has shown no benefits for the improvement of the predictive performance, but with the HS.STD criterion we often found sparse models, under all three regularization priors, with a comparable performance as the full models. We have seen, that the posterior inclusion probabilities for the covariates, as provided by the NMIG prior, reflect very well the importance of the covariates. Nevertheless, also variable selection guided by the ranking of the covariates, with respect to the inclusion probabilities, shows in general no improvement of the predictive performance. In the high-dimensional-covariate or low-sample-size cases the inclusion probabilities are shifted towards zero and the separation of the covariates with moderate effects from the covariates with small or zero effects vanishes.

From the variations of the sample size and the number of covariates we found that the AFT model with PGM error can be applied with sample sizes about $n \approx 200$ and a low number of covariates e. g. $p_x \approx 25$, where some of them can also be modeled as nonlinear. With respect to the results of the baseline error estimation we do not recommend the use of the PGM error in higher-dimensional cases with $p_x \geq n/2$, because regularization can compensate only limited the lack of information in the data to estimate reliably the high-parametric AFT model with PGM error.

11. CRR-type models

In this section we investigate the performance of the Bayesian ridge, lasso and NMIG prior under the extended Cox relative risk (CRR) model as described in Section 7. The results obtained from the Bayesian methods are compared with the associated frequentist versions of the ridge or lasso penalty and a backward-stepwise procedure based on the AIC criterion. In addition to the semiparametric approach and the P-spline based modeling, we consider also a parametric Weibull model $\lambda_0(t) = \alpha t^{\alpha-1}$ for the baseline hazard as competitor.

We start in Subsection 11.1 with the case $n > p_x$, where more observations $n \in \mathbb{N}$ than covariates $p_x \in \mathbb{N}$ are available and assume a simple linear shape of the baseline hazard in the data generation process. The models in Subsection 11.2 consider more complex shapes of the baseline hazard and additional nonlinear covariate effects. One of these models is revisited in Subsection 11.3, where we utilize an AFT model with PGM error for inference to explore the consequences, if the survival model is miss-specified. We proceed with the higher-dimensional case in Subsection 11.4, where the number of linear modeled covariates p_x is sequentially increased until it exceeds the sample size n. Finally, in Subsection 11.5, this section is concluded by considering modifications of the hard shrinkage selection criterion based on posterior relative frequencies of the Bayesian NMIG indicator values.

Functions and methods

Frequentist inference for the CRR model relies on the partial likelihood and we utilize the `R`-functions `coxph()` of the package `{survival}` and `penalized()` of the `{penalized}` `R`-package from J. Goeman for estimation. Frequentist variable selection, based on the AIC criterion, is practiced with `coxph()` in combination with the backward-stepwise search as provided by the `step()` function. The function `penalized()` enables the estimation of lasso- and ridge-regularized linear regression coefficients, where the optimal shrinkage parameter λ is determined by n-fold (leave-one out) generalized cross validation. To select genes that are related to the patient's survival, Gui and Li (2005) proposed also a LARS-COX procedure which uses L_1-penalized estimation for the CRR model as well. In this procedure, the least angle regression method, Efron et al. (2004), was applied to solve the computational difficulty in high-dimensional-covariate and low-sample-size cases. Further the `R`-package `{glmpath}`,Park and Hastie (2007), is available as competitor, where the coefficients are computed on a grid of values for λ at which the set of non-zero coefficients changes. As typical for the frequentist lasso, all these methods can select at most n variables. Due to the similarity, we restrict the presentation of the regularized frequentist results to those achieved with the `penalized()` function.

Bayesian inference for the CRR model is either based on the partial likelihood or on the full likelihood, if a parametric Weibull or nonparametric P-spline baseline hazard function is assumed. The MCMC sampling based inference algorithms for the model components are described in Section 9. In particular the partial likelihood based algorithms are implemented in the `R`-function `bcoxpl()`, which is available from the author by request. Inference with the full likelihood is carried out with the method `regress` as implemented in the free software `BayesX` (available from http://www.stat.uni-muenchen.de/~bayesx). The usage of both functions is described in the Appendix D.3 and D.4. The Bayesian point estimates of model parameters are based on the empirical mean of the associated, generated sample from the marginal posterior distribution. Further summary statistics, like the standard deviation or quantiles, are also computed using their empirical counterparts.

Estimation accuracy

We measure the estimation accuracy in terms of the mean squared errors (MSE) as defined in Section 10 over $R = 50$ runs. Further, we report the average number of correctly and incorrectly classified zero and nonzero coefficients after applying the hard shrinkage rules presented in Section 4.4. To compute the cumulative baseline hazard function $\hat{\Lambda}_0(t) = \int_0^t \hat{\lambda}_0(u)du$ associated to the estimated P-spline baseline hazard $\hat{\lambda}_0(\cdot)$, the trapezoidal rule is used so that the results become comparable to the corresponding Breslow estimates $\hat{\Lambda}_0^{Br}(t)$ from the partial likelihood. A selection of the main results is presented in the next sections. The used abbreviations that describe the models and inferential methods are summarized in the Reference Section.

11.1. Low-dimensional linear predictor

Data generation

For our first simulations we use the configuration of the data generating process from Tibshirani (1997). Nine covariates $\mathbf{x}_i = (x_{i,1},...,x_{i,9})'$ are randomly drawn from a multivariate Gaussian distribution with zero mean, unit variance and covariance matrix chosen such that the correlation between $\mathbf{x}_j$ and $\mathbf{x}_k$ is $\operatorname{corr}(x_{i,j}, x_{i,k}) = \rho^{|j-k|}$ with $\rho = 0.5$. The survival times T_i, $i = 1,...,n$, are generated from an exponential hazard model with constant baseline hazard $\lambda_0(t) = 1$, i. e.

$$\lambda_i(t_i) = \exp(\mathbf{x}_i'\boldsymbol{\beta}),$$

while the censoring variables C_i, $i = 1,...,n$, are generated as i.i.d. draws from the uniform distribution $U[0, c_0]$ with c_0 chosen to obtain censoring rates about 25 % in each dataset. For the various CRR models with the following nine regression coefficients

$$\text{CRR 1:} \quad \boldsymbol{\beta} = (-0.7, -0.7, 0, 0, 0, -0.7, 0, 0, 0)',$$

$$\text{CRR 2:} \quad \boldsymbol{\beta} = (0.1, 0.1, 0.1, 0.1, 0.1, 0.1, 0.1, 0.1, 0.1)',$$

$$\text{CRR 3:} \quad \boldsymbol{\beta} = (-0.4, -0.3, 0, 0, 0, -0.2, 0, 0, 0)',$$

we produced $R = 50$ datasets with $n = 200$ life times in each case. The first and the second model were used in Tibshirani (1997), who compared the frequentist lasso regularization in the CRR model with the stepwise procedure, and the first and third model were used in Zhang and Lu (2007) in the context of the adaptive lasso with covariate-specific penalties.

Function and parameter specification

Methods: We us the functions `coxph()` and `step()` for frequentist estimation and `penalized ()` for the frequentist lasso and ridge regularization. For Bayesian inference we use the `R`- function `bcoxpl()`and the `BayesX` method `regress` with the ridge, lasso and NMIG regularization of the linear effects.

Hyperparameters: The hyperparameters for the shrinkage parameter of the Bayesian lasso and ridge prior are set to the weakly informative values $h_{1,\lambda} = h_{2,\lambda} = 0.01$ to enable a greater amount of adaptiveness for the shrinkage parameter depending on the data. The hyperparameters for the two variance parameter components of the Bayesian NMIG prior are $v_1 = 1$, $v_0 = 0.000025$, $h_{1,\psi} = 5$ and $h_{2,\psi} = 25$ in combination with $h_{1,\omega} = 1$ and $h_{2,\omega} = 1$ to define a uniform prior for the complexity parameter ω.

Starting values: In `BayesX` the starting values for the regression coefficients are computed via backfitting within Fisher scoring. In the function `bxoxpl()` we avoid preprocessing steps to fit the model in order to obtain suitable starting values and start with a weakly specified model. For the starting values of the linear effects we select values close to zero, i. e. $\beta_j^{(0)} = 0.01$, $j = 1,...,p_x$. The Bayesian NMIG regularization components starts with $I_j^{(0)} = v_0$, $\psi_j^{2(0)}$ corresponding to the value left mode dependent on the specification of the variance prior for ψ_j^2 and $\omega^{(0)} = 0.5$. The shrinkage parameter for the Bayesian lasso and ridge prior starts in $\lambda^{(0)} = 1$.

Estimation: For the Bayesian MCMC methods based on the full and partial likelihood we use 10000 iterations with a burnin of 2000 and thin the chain by 8, which results in an MCMC sample of size 1000. On a system with quad-core CPU (Intel Quad9550, 2.83 GHz) we need about 5 minutes for the Bayesian partial likelihood based models estimated in `R`, and about 15 seconds for the full likelihood based models in `BayesX`.

Results for model CRR 1

MSE of the linear effects

Figure 11.1 shows the mean squared errors for the estimated regression coefficients, $\text{MSE}(\hat{\boldsymbol{\beta}})$, under the different regularization priors for the linear effects, when inference is based on the partial likelihood (CPL) and the full likelihood (CFL) with P-spline baseline hazard. In addition we show the MSEs obtained after applying the hard shrinkage selection criteria (HS.STD, HS.CRI, HS.IND) to the Bayesian estimates of the regression coefficients as described in Section 4.4. Due to the similarity of the results from the Weibull (WB) and the P-spline baseline hazard model (with exception of the baseline hazard performance) the Weibull model results are often omitted in the following.

We note that the Bayesian NMIG model (CPL.BN, CFL.BN) performs best within each group of survival models (CPL, CFL) and outperforms the stepwise procedure (CPL.Step) as well as the frequentist lasso (CPL.PenL) and the Bayesian lasso (CPL.BL, CFL.BL). The MSEs of the Bayesian NMIG estimates are close to the MSE of the maximum partial likelihood estimates, if the predictor with the true covariate structure under model CRR 1 is used (CPL.T). The MSEs of the corresponding unpenalized Bayesian methods using the true predictor are comparable to the MSE of CPL.T and are omitted in the figures. A marginal improvement in the MSE performance can be observed for the sparse models resulting from hard shrinkage selection criterion based on the standard deviation. In particular the MSE under the Bayesian lasso (CPL.BL-HS.STD, CFL.BL-HS.STD) and Bayesian ridge (CPL.BR-HS.STD, CFL.BR-HS.STD) prior is reduced, compared to the associated models that include all covariates. The MSE under the Bayesian lasso and ridge prior is further improved, if the hard shrinkage criterion based on the 95% credible region is applied to the P-spline model (CFL.BL-HS.CRI, CFL.BR-HS.CRI), but the high performance of the Bayesian NMIG models is not reached. Furthermore, the HS.IND criterion only slightly changes the MSE of the resulting Bayesian NMIG model, since the estimates of the zero effects, compare **Figure 11.2**, are very close to zero anyway, i. e. it is negligible, if they are removed from the final model or not.

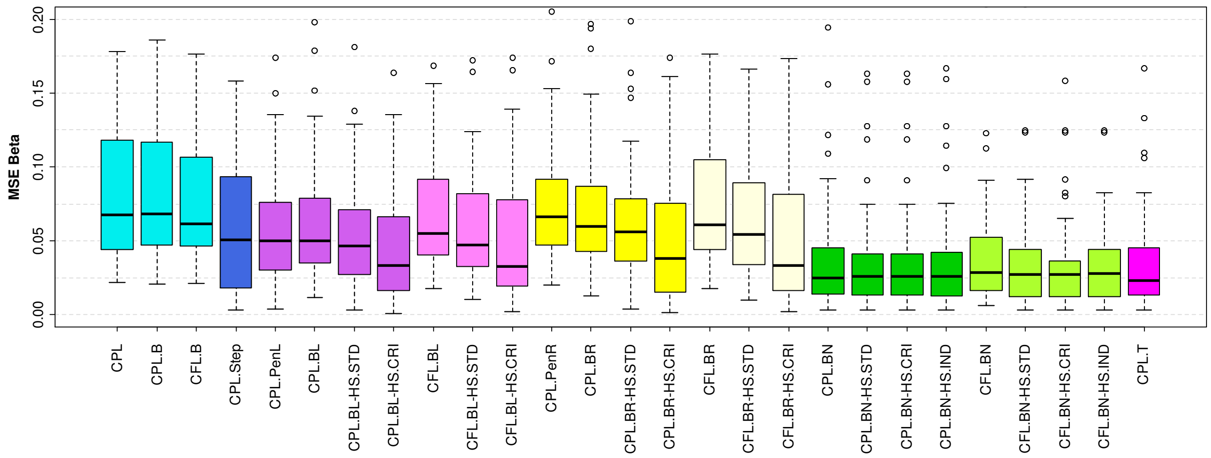


Figure 11.1: Mean squared errors of the regression coefficient estimates, MSE($\hat{\boldsymbol{\beta}}$), under the different regularization and variable selection methods in simulation model CRR 1. The right box (CPL.T) shows the MSE($\hat{\boldsymbol{\beta}}$) for the maximum partial likelihood estimations when the true predictor structure is used.

Classification

The second and third column in **Table 11.1** show the resulting average number of the correctly classified nonzero coefficients ($\hat{\beta} \neq 0, \beta \neq 0$) and correctly classified zero coefficients ($\hat{\beta} = 0, \beta = 0$) for the 50 simulation datasets under the different variable selection methods. Column four displays the frequencies of the final models (MF) with the true predictor structure, i. e. correctly specified zero and nonzero coefficients.

While all methods reach the optimal value of 3 correctly classified nonzero regression coefficients, the optimum of 6 correctly classified zero regression coefficients is only achieved with the Bayesian NMIG regularization in combination with the HS.CRI criterion. The associated final models recover in all 50 cases the true model. High numbers of correctly classified zero coefficients result also for the models obtained with the HS.STD and HS.IND criterion, in particular the partial likelihood models recover in 49 resp. 47 of 50 cases the true model. We note generally for all methods, that the average number of true estimated zero effects tends to smaller values under the Weibull and P-spline model of the baseline hazard as under the partial likelihood, where the baseline is left unspecified. Under comparable prior specification we often observe that the regression coefficients, obtained with the full likelihood, are less regularized as those obtained with the partial likelihood, compare **Figure 11.2**. This various amounts of shrinkage explain the variability in the performance under the different hard shrinkage selection criteria. In summary, in model CRR 1, with clearly separable zero and nonzero effects, the best results in terms of the MSE and the classification are obtained with Bayesian NMIG prior.

	Model 1			Model 2			Model 3		
	$\hat{\beta} \neq 0$ $\beta \neq 0$	$\hat{\beta} = 0$ $\beta = 0$	MF	$\hat{\beta} \neq 0$ $\beta \neq 0$	$\hat{\beta} = 0$ $\beta = 0$	MF	$\hat{\beta} \neq 0$ $\beta \neq 0$	$\hat{\beta} = 0$ $\beta = 0$	MF
BEST	3	6	50	9	0	50	3	6	50
CPL.Step	3	4.90	19	3.94	0	0	2.66	4.58	6
CPL.PenL	3	3.60	2	6.42	0	2	2.88	3.86	7
CFL.B-HS.STD	3	4.18	5	4.58	0	0	2.80	3.90	2
WB.B-HS.STD	3	4.28	4	4.52	0	0	2.78	3.94	3
CPL.B-HS.STD	3	4.20	4	4.54	0	0	2.78	3.86	2
CFL.BL-HS.STD	3	4.38	6	4.38	0	0	2.78	4.26	4
WB.BL-HS.STD	3	4.56	9	4.36	0	0	2.78	4.40	8
CPL.BL-HS.STD	3	4.70	8	4.04	0	0	2.74	4.96	10
CFL.BR-HS.STD	3	4.30	4	4.66	0	0	2.80	3.94	2
WB.BR-HS.STD	3	4.26	3	4.52	0	0	2.80	4.08	4
CPL.BR-HS.STD	3	4.42	6	4.72	0	0	2.76	4.52	7
CFL.BN-HS.STD	3	5.82	42	2.04	0	0	2.12	5.74	12
WB.BN-HS.STD	3	5.82	42	1.94	0	0	2.06	5.78	14
CPL.BN-HS.STD	3	5.98	49	1.10	0	0	1.68	5.96	7
CFL.B.HS-CRI	3	5.68	38	1.62	0	0	2.26	5.66	15
WB.B-HS.CRI	3	5.64	36	1.48	0	0	2.22	5.66	14
CPL.B-HS.CRI	3	5.64	36	1.40	0	0	2.26	5.64	15
CFL.BL-HS.CRI	3	5.74	40	1.32	0	0	2.20	5.76	15
WB.BL-HS.CRI	3	5.68	38	1.32	0	0	2.16	5.78	13
CPL.BL-HS.CRI	3	5.84	43	0.92	0	0	1.98	5.90	13
CFL.BR-HS.CRI	3	5.68	37	1.44	0	0	2.30	5.68	15
WB.BR-HS.CRI	3	5.62	35	1.34	0	0	2.18	5.68	13
CPL.BR-HS.CRI	3	5.76	41	1.04	0	0	2.16	5.82	15
CFL.BN-HS.CRI	3	6.00	50	0.60	0	0	1.54	5.98	3
WB.BN-HS.CRI	3	6.00	50	0.56	0	0	1.52	5.98	4
CPL.BN-HS.CRI	3	6.00	50	0.40	0	0	1.04	5.98	0
CFL.BN-HS.IND	3	5.74	40	2.30	0	0	2.14	5.70	12
WB.BN-HS.IND	3	5.80	41	2.20	0	0	2.12	5.72	14
CPL.BN-HS.IND	3	5.94	47	1.20	0	0	1.68	5.94	7

Table 11.1: Average number of correctly classified regression coefficients for the models CRR 1, CRR 2 and CRR 3 after variable selection. Especially $\hat{\beta} \neq 0, \beta \neq 0$ denotes the case that the estimated effect is nonzero ($\hat{\beta} \neq 0$) when corresponding true effect is nonzero ($\beta \neq 0$), and $\hat{\beta} = 0, \beta = 0$ denotes the case that the estimated effect is zero ($\hat{\beta} = 0$) when corresponding true effect is zero ($\beta = 0$). The columns (MF) display the frequencies of the final models that recover the true model.

Linear effects

Figure 11.2 presents the estimated values of two selected regression coefficients, $\beta_5 = 0$ and $\beta_6 = -0.7$, obtained with the different estimation and regularization methods. If we focus on the Bayesian NMIG prior, we see that the estimates (CPL.BN, CFL.BN) for the zero effects are much more concentrated around zero, similar to the stepwise selection (CPL.Step) and the variable selection under frequentist lasso (CPL.PenL), while the estimates of the nonzero effects are close to the unregularized estimates. The estimates reflect the adaptive selection-type shrinkage property of the Bayesian NMIG prior, with strong shrinkage of smaller and at the same time weak shrinkage of larger regression coefficients, where the separation between “large” and “small” coefficients depends on the specification of the NMIG prior hyperparameters. Obviously, the Bayesian ridge and lasso prior cause less shrinkage of the nonzero regression coefficients in the P-spline model for the baseline hazard (CFL.BR, CFL.BL) as in the semiparametric frequentist and Bayesian version, where inference is based on the partial likelihood (CPL.BR, CPL.BL), compare also **Figure 11.3**. This demonstrates, as previously mentioned in the Section 4.3, that the prior-specific real term regularization depends on the shape of the likelihood. We will see this interdependency again in Subsection 11.3 and the application sections, where we fit the CRR and the AFT model to the data. Due to the large values of the cross validated shrinkage parameters λ, displayed at the left side of **Figure 11.4**, we observe in general a stronger regularization of the frequentist lasso (CPL.PenL) and ridge (CPL.PenR) estimates compared to the Bayesian counterparts.

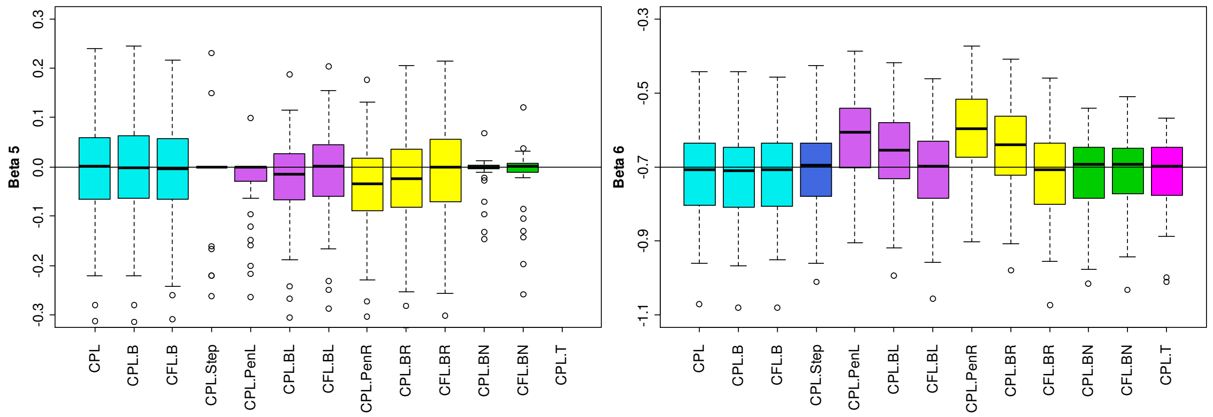


Figure 11.2: Regression coefficient estimates $\hat{\boldsymbol{\beta}}$ for two selected coefficients under different regularization priors in simulation model CRR 1. The right box (CPL.T) shows the estimations when the true predictor structure is used. The black horizontal lines in the box plots mark the values of the true regression coefficients $\beta_5 = 0, \beta_6 = -0.7$.

Penalties and shrinkage parameters

If we take a look at **Figure 11.3**, we see the different amount of the covariate-specific penalization for the nine regression coefficients expressed in terms of the inverse variance parameters $\tau_{\beta_j}^{-2}$, $j = 1,...,9$. Shown are the results under the Bayesian lasso, ridge and NMIG prior with the partial (left side) and full likelihood (right side). With the results from Section 4.5 we keep in mind that in particular under the NMIG prior the posterior mean estimate of $\tau_{\beta_j}^{-2}$ for smaller effects covers only a small range of applied penalization and represents rather a lower bound for the penalization.

The penalization of the nonzero effects β_1, β_2, β_6 induced by the Bayesian NMIG prior leads to much smaller values than those of the Bayesian lasso and Bayesian ridge prior. In contrast to the ridge prior, the adaptive shrinkage, i. e. the small penalization for nonzero effects and larger penalization of zero effects, is reflected by both, the Bayesian lasso and the Bayesian NMIG prior, but the NMIG penalty values for the nonzero effects are very close to zero, so that the resulting regression coefficient estimates are almost unregularized. The Bayesian ridge penalty is within the range of the Bayesian lasso penalty, i. e. smaller effects are less and larger effects are stronger regularized (compared to the lasso). In the case of the P-spline (or Weibull) baseline hazard, when inference is carried out with the full likelihood, we observe by trend a smaller penalization across the priors compared to the partial likelihood approach, with less pronounced differences under the Bayesian NMIG prior.

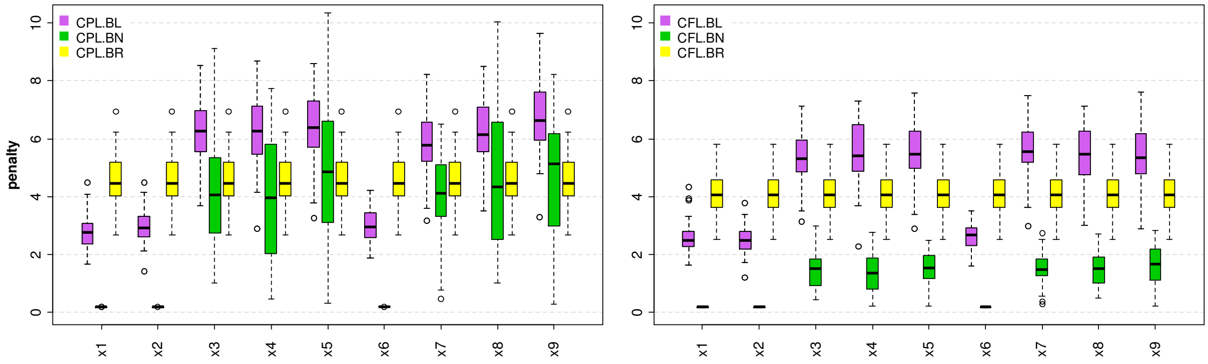


Figure 11.3: Estimates of the covariate-specific penalty $\hat{\tau}_{\beta_j}^{-2}$ for the Bayesian lasso (BL), NMIG (BN) and ridge (BR) prior in simulation model CRR 1. Left side: The partial likelihood (CPL) is used for inference. Right side: The full likelihood with baseline modeled as P-spline (CFL) is used for inference.

The estimated shrinkage parameters are given in **Figure 11.4**. In particular for the frequentist lasso and ridge regularization the shrinkage parameter reflects the amount of penalization that is uniformly applied to all regression coefficients. The penalty from the frequentist lasso is located within the range of the covariate-specific penalty values of the Bayesian lasso, while the penalty from the frequentist ridge clearly exceeds the Bayesian counterpart. The impact of the different amount of penalization induced under the various methods is directly reflected in the estimates of the regression coefficients shown in **Figure 11.2**. Finally, the estimated complexity parameter ω of the Bayesian NMIG prior is displayed at the right side of **Figure 11.4**. We obtain very concentrated values around 0.4, and the weaker regularization of the small effects, observed with the full likelihood approach, increases (marginally) the model complexity.

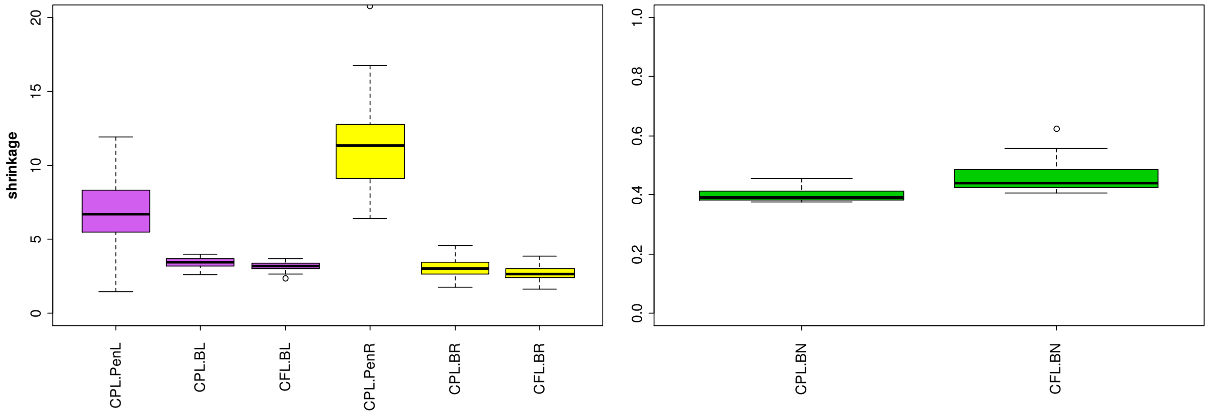


Figure 11.4: Estimated shrinkage parameters under the different regularization methods in simulation model CRR 1. Left side: Shrinkage parameter λ^2 and λ of the frequentist and Bayesian lasso and ridge prior. Right side: Shrinkage parameter ω of the Bayesian NMIG prior.

NMIG indicators

The variable selection feature of the Bayesian NMIG prior is highlighted at the left side in **Figure 11.5**, where the estimated inclusion probabilities, based on posterior relative frequencies of the NMIG indicator variable value $I_j = v_1$, are shown under the partial likelihood and the full likelihood with P-spline baseline hazard. The inclusion probabilities of the nonzero effects are nearly one, with a very small standard deviation. For the zero effects the inclusion probabilities are shifted towards zero and clearly fall below the selection threshold 0.5 of the HS.IND criterion. Although inclusion probabilities for the zero effects resulting from the full likelihood approach tend to be higher than those from the partial likelihood, they provide a good resource to select the important covariates in both cases. At the right side in **Figure 11.5** the acceptance rates of the regression coefficients in the CRR model based on the partial likelihood are shown. In general we achieved high acceptance rates in all simulation models, also under the full likelihood, and often the rates under the Bayesian NMIG prior stand out with notable high values.

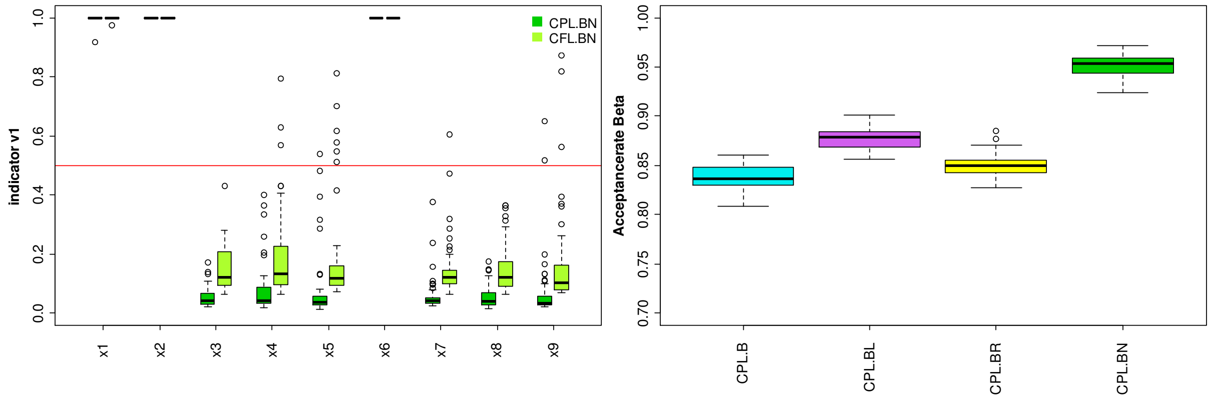


Figure 11.5: Left side: Estimated inclusion probabilities based on posterior relative frequencies of the Bayesian NMIG indicator variable value $I_j = v_1$ for simulation model CRR 1 based on the partial likelihood (CPL.BN) and the full likelihood (CFL.BN). The red horizontal line marks the cut off value 0.5 of the hard shrinkage selection criterion HS.IND. Right side: Acceptance rates of the regression coefficients in the CRR model based on the partial likelihood.

MSE of the baseline quantities

A view at the MSEs of the estimated cumulative baseline hazards, $\mathrm{MSE}(\hat{\Lambda}_0)$, under the different model classes shows the high performance of the baseline estimates produced with the full likelihood approaches, compare **Figure 11.6**. In particular, the higher performance of the Weibull model compared to the nonparametric P-spline model is plausible, since the underlying exponential baseline is a special Weibull baseline with $\alpha = 1$.

Under the partial likelihood approach the Breslow estimators, as a step function, cause a loss in the performance. This becomes apparent by considering the right side of **Figure 11.7**, which shows the estimates of the cumulative baseline hazard for one selected simulation dataset if the Bayesian NMIG prior is applied. In the time interval $[0,1]$, where most of the observations occur, the P-spline based estimate (CFL.BN) and the Breslow estimate (CPL.BN) approximate the true cumulative baseline very well. When time increases, the less observations are available and the deviations get larger, which results in an increasing MSE, in particular for the Breslow estimate. If we restrict the calculations of

$\text{MSE}(\hat{\Lambda}_0)$ to the interval [0,1] that contains most of the observations, the MSEs of the estimated baselines as well as the MSE of the estimated cumulative baselines are very similar across all models.

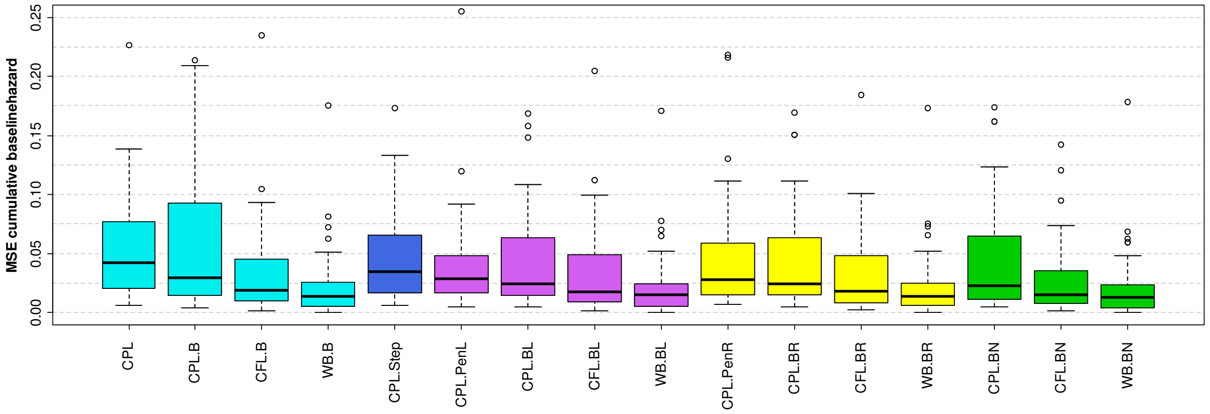


Figure 11.6: Mean squared errors for the estimated cumulative baseline hazard, $\text{MSE}(\hat{\Lambda}_0)$, under the different regularization priors in simulation model CRR 1.

Finally, the left side of **Figure 11.7** shows the P-spline estimates of the log-baseline hazard for one selected simulation dataset under the Bayesian NMIG and unregularized estimation of the linear effects. Both estimates are very close to each other and in summary we found that the specific regularization of the linear effects induces only negligible differences in the global shape of the baseline hazard estimates. Nevertheless, the shrinkage-type and shrinkage-strength of the linear effects affect the estimate of the baseline hazard function, but this is often hard to detect. For a demonstration we refer to the Application Section 12.3.2, Figure (12.10), where the impact of the regularization on the baseline hazard estimate is shown in terms of the Bayesian lasso and NMIG prior with varying shrinkage parameter.

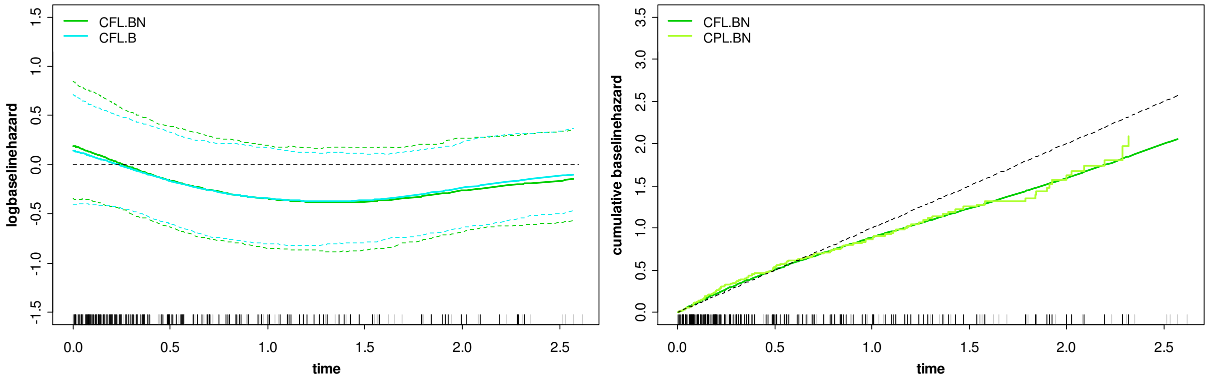


Figure 11.7: Estimation of the log-baseline hazard $\log \hat{\lambda}_0(t)$ (left side) and the cumulative baseline hazard $\hat{\Lambda}_0(t)$ (right side) for one selected dataset under simulation model CRR 1. Left side: Posterior mean estimate of the log-baseline hazard (solid lines) based on the full likelihood together with the 2.5% and 97.5% pointwise credible bands (dashed lines) for the CFL.B and CFL.BN model when the baseline is modeled as P-spline. Right panel: Posterior mean estimate of the cumulative baseline hazard under the Bayesian NIMG prior. In both figures the black dashed line marks the true exponential log-/cumulative baseline hazard and the vertical rugs at the time axis mark the observed event times (black) and censoring times (gray).

Results for model CRR 2

This subsection considers the results for the various estimation methods under the different regularization priors if all nine covariates in the predictor are assigned with small but nonzero effects, i. e. $\beta_j = 0.1,\ j = 1,...,9$.

MSE of the linear effects

Figure 11.8 summarizes the MSEs of the estimated regression coefficients of model CRR 2 under the various applied methods. Best performances in this particular situation are obtained under the frequentist and Bayesian ridge regularization and under Bayesian lasso regularization, where especially the partial likelihood approaches (CPL.PenR, CPL.BR, CPL.BL) outperform all remaining approaches, even the performance of the models utilizing the true predictor structure (CPL, CPL.B, CFL.B). Especially the Bayesian lasso (CPL.BL) estimates achieve a slightly better performance than the sparse estimates from frequentist lasso (CPL.PenL), with values that are comparable to those of the ridge regularization. Also variable selection in terms of the hard shrinkage selection rules causes a clear loss in the predictive performance. In particular the specific, selection-like regularization property of the Bayesian NMIG prior, with enhanced shrinkage of smaller regression coefficients, has negative effects on the MSE and the estimated models have a clearly poor performance.

Classification

The classification results for model CRR 2 are subsumed in **Table 11.1**, where the average number of correctly identified nonzero coefficients ($\hat{\beta} \neq 0, \beta \neq 0$) is recorded. Of course, it is not possible to reach the optimal value of nine when variable selection is applied, but comparably high values result for the frequentist lasso and the Bayesian ridge and lasso in combination with the HS.STD criterion. The identification of the true predictor is out of reach, so all MF values are zero.

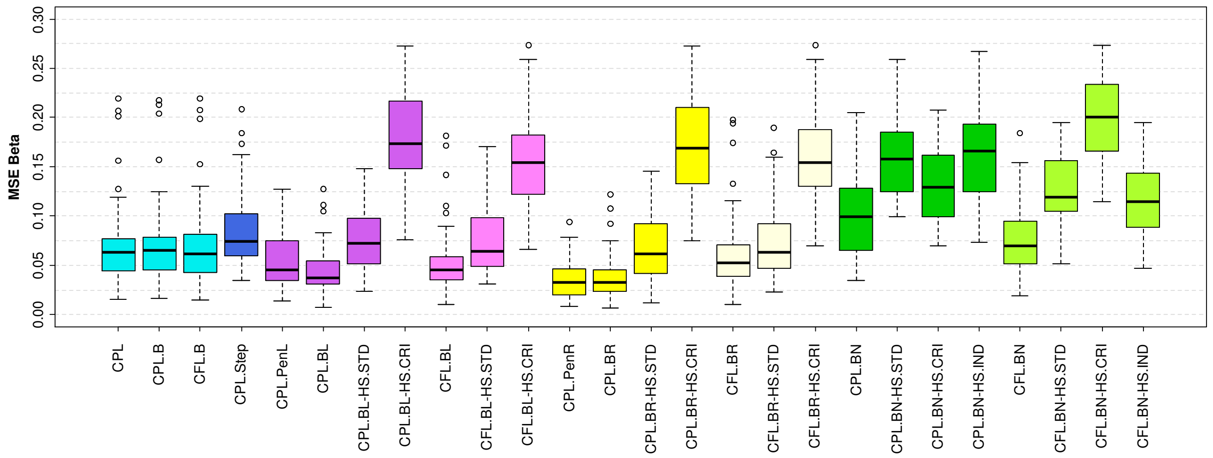


Figure 11.8: Mean squared errors of the regression coefficient estimates, $\mathrm{MSE}(\hat{\boldsymbol{\beta}})$, under the different regularization and variable selection methods in simulation model CRR 2.

NMIG indicators and penalties

The increased loss of performance under the Bayesian NMIG prior is further explained, when taking a look at the left side of **Figure 11.9**, where the estimated inclusion probabilities based on posterior relative frequencies of the NMIG indicator variable value $I_j = v_1$ are shown. Almost all inclusion probabilities are of comparable size and tend to be closer to zero than to one, which induces the heavy penalization of all regression coefficients, compare **Figure 11.9** right side.

Under the various regularization priors we found again a difference between the full and partial likelihood estimates with a trend for less penalization for the full likelihood estimates, which explains the benefit in the MSE under the Bayesian NMIG prior with the full likelihood. But, the weaker regularization obviously causes a drawback in the MSE under the Bayesian lasso and ridge prior.

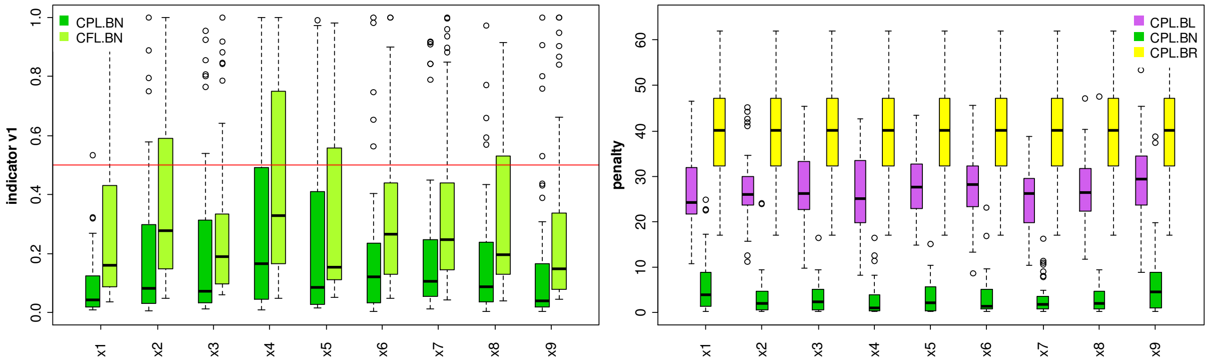


Figure 11.9: Left side: Estimated inclusion probabilities based on posterior relative frequencies of the Bayesian NMIG indicator variable value $I_j = v_1$ for simulation model CRR 2 based on the partial likelihood (CPL.BN) and the full likelihood (CFL.BN). The red horizontal line marks the cut off value 0.5 of the hard shrinkage selection criterion HS.IND. Right side: Estimates of the covariate-specific penalty $\hat{\tau}_{\beta_j}^{-2}$ for the Bayesian lasso (CPL.BL), NMIG (CPL.BN) and ridge (CPL.BR) prior in simulation model CRR 2 under the partial likelihood.

Linear effects

According to the penalization results the point estimates for the regression coefficients are more or less shrunken towards zero and do not reflect the true model, compare **Figure 11.10**. The stepwise procedure behaves similar as the Bayesian NMIG prior by producing zero estimates for the small nonzero effects.

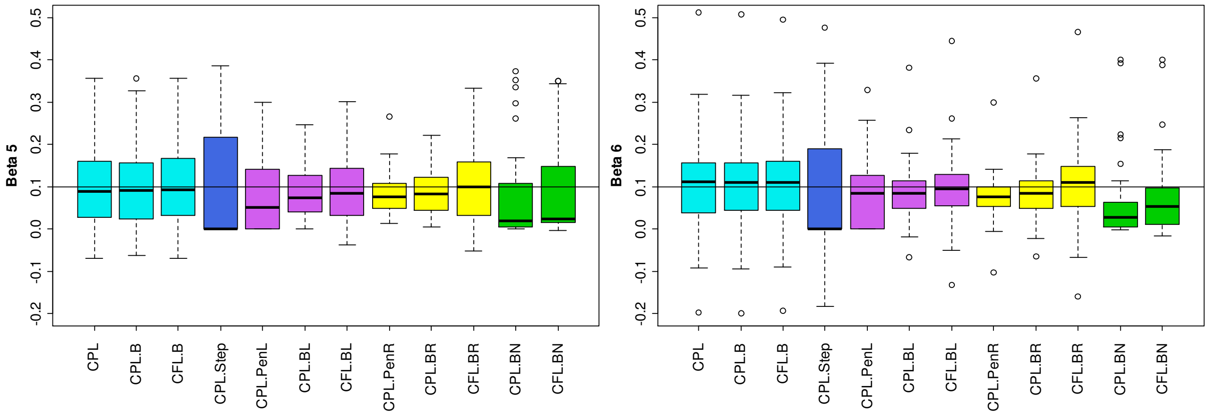


Figure 11.10: Regression coefficient estimates $\hat{\boldsymbol{\beta}}$ for two selected coefficients under different regularization priors in simulation model CRR 2. The black horizontal lines in the box plots mark the values of the true regression coefficients $\beta_5 = -0.1, \beta_6 = -0.1$.

MSE of the baseline quantities

Finally, the estimation of the baseline and cumulative baseline hazard functions leads to comparable results as in model CRR 1 and the results are summarized in terms of the MSE for the cumulative baseline hazard in **Figure 11.11**. Best results are again obtained with the Weibull model and further the estimates under the Bayesian NMIG prior (WB.BN) show a marginal better performance than the remaining Weibull hazard estimates with respect to the median and the box-width.

In summary, it is shown that in the setting of model CRR 2, if all effects have small but nonzero values, variable selection or selection-like shrinkage causes a loss in the MSE performance. Further, it seems to be advantageous to keep all effects regularized in the predictor, in combination with a Bayesian or frequentist ridge or a Bayesian lasso penalty, and that a moderately stronger regularization can improve the predictive performance. From the practical perspective we obtain similar results, compare Application Section 14.3, in particular Figure 14.5, where the predictive performance of the models (measured in terms of the IBS) is shown in dependence on varying values of the shrinkage parameter.

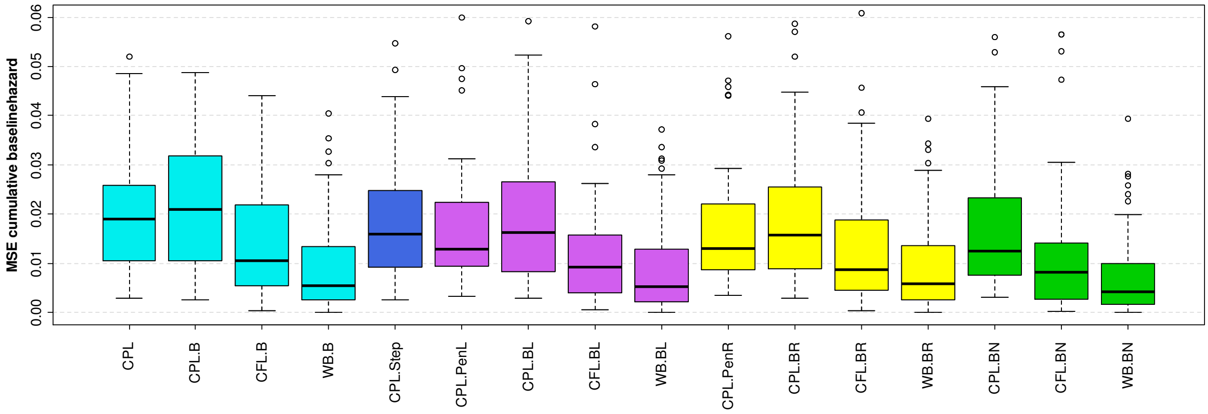


Figure 11.11: Mean squared errors for the estimated cumulative baseline hazard, $\text{MSE}(\hat{\Lambda}_0)$, under the different regularization priors in simulation model CRR 2.

Results for model CRR 3

This subsection considers the results for the various estimation methods under the different regularization priors, if the zero and nonzero effects are distributed to the nine covariates as in model CRR 1, but assigned with smaller nonzero values, i. e. $\boldsymbol{\beta} = (-0.4, -0.3, 0, 0, 0, -0.2, 0, 0, 0)'$.

MSE of the linear effects

Figure 11.12 shows the MSEs achieved for the regression coefficient estimation in the model CRR 3.

Again the MSEs of the maximum partial likelihood estimators with the true predictor structure are recorded as a benchmark result.

As in the previous model CRR 2, the regularization based approaches achieve lower MSEs than those without penalization and the best MSEs are again derived with the lasso and ridge penalty in combination with the partial likelihood. We obtain a similar result also from the Application Section 14, where the predictive performance of the models is assessed in terms of the integrated Brier score.

In contrast to model CRR 2 the Bayesian NMIG models achieve a better performance with less pronounced differences to the lasso and ridge type models. But, none of the methods obtain the high performance of the model utilizing the true predictor structure (CPL.T). Variable selection, as automatically resulting with the frequentist lasso (CPL.PenL) or artificially enforced with the hard shrinkage selection rules for the Bayesian estimates provides sparse models with comparable MSE to the full models including all covariates.

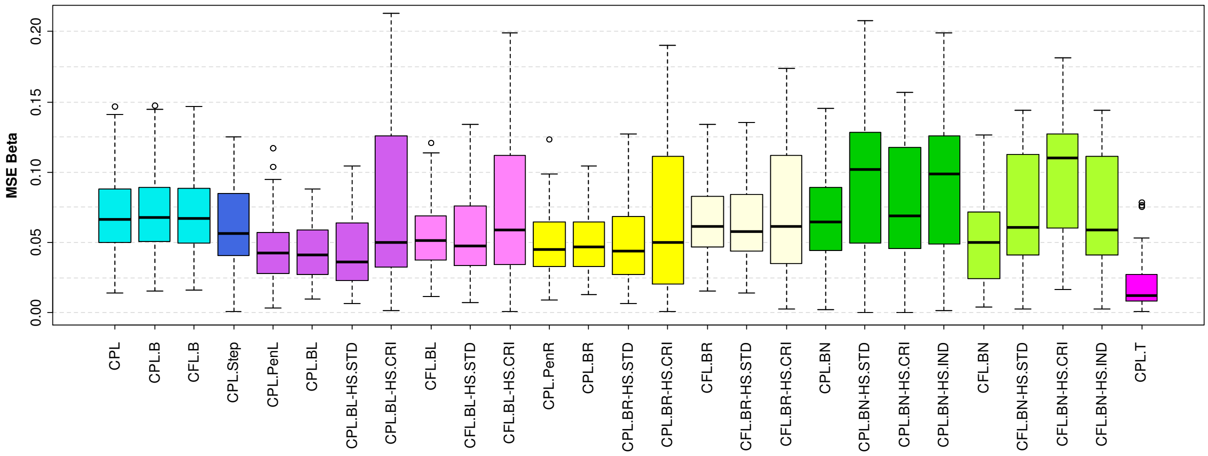


Figure 11.12: Mean squared errors the regression coefficient estimates $\hat{\boldsymbol{\beta}}$ under the different regularization and variable selection methods in simulation model CRR 3. The right box (CPL.T) shows the $\text{MSE}(\hat{\boldsymbol{\beta}})$ for the maximum partial likelihood estimations when the true predictor structure is used.

Classification

As a further consequence we observe in **Table 11.1** a decrease of the correctly classified nonzero effects for model CRR 3 compared to model CRR 1. The best methods detect only in about 15 of 50 cases the true predictor structure.

NMIG indicators and penalties and linear effects

The improvement in the MSE using the NMIG prior (with resp. to model CRR 2) is revealed by taking a look at left side of **Figure 11.13**, where the relative frequencies of the indicator variable value $I_j = v_1$ from the Bayesian NMIG model are displayed. The inclusion probabilities for the largest effect with value $\beta_1 = -0.4$ is nearly to one and for the second largest effect with value $\beta_2 = -0.2$ the HS.IND cut-off value of 0.5 is still frequently passed. Further simulations (not all presented in this work) have shown that, if the same prior settings as noted above are used for comparable models, effects with absolute values larger than 0.3 can be separated from the zero effects very well by the cut-off value 0.5.

The Bayesian penalties are displayed at the right side of **Figure 11.13**. In contrast to the results from model CRR 1 the Bayesian ridge penalty is here in the upper range of the Bayesian lasso penalty. This is leading to a stronger regularization of the nonzero effects with the ridge prior compared to the lasso prior and correspondingly a comparable regularization of the zero effects. Under the Bayesian NMIG prior the penalty of the nonzero effects is again very small and close to zero so that these effects are weakly regularized. The amount of penalization on the zero effects is only limited reflected by the penalty but exhibits in the associated regression coefficient estimates. Finally, the impact of the various amounts of penalization on the regression coefficient estimates is displayed in **Figure 11.14**.

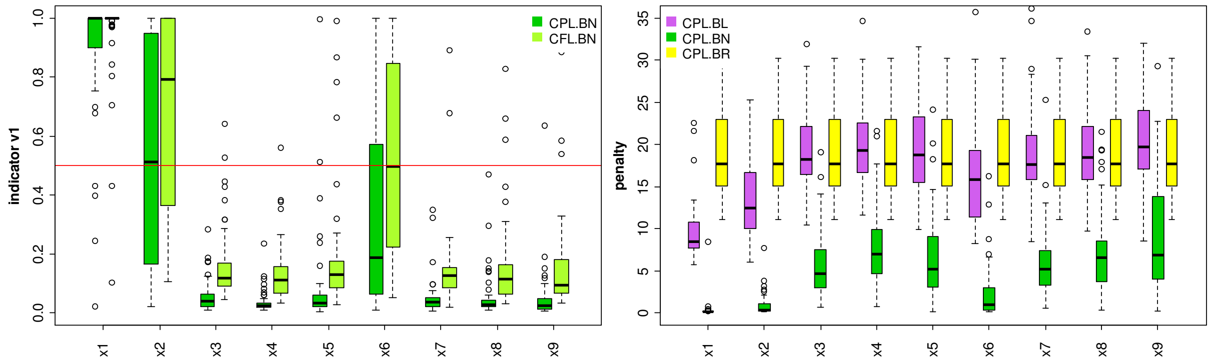


Figure 11.13: Left side: Estimated inclusion probabilities based on posterior relative frequencies of the Bayesian NMIG indicator variable value $I_j = v_1$ for simulation model CRR 3 based on the partial likelihood (CPL.BN) and the full likelihood (CFL.BN). The red horizontal line marks the cut off value 0.5 of the hard shrinkage selection criterion HS.IND. Right side: Estimates of the covariate-specific penalty $\hat{\tau}_{\beta_j}^{-2}$ for the Bayesian lasso (CPL.BL), NMIG (CPL.BN) and ridge (CPL.BR) prior in simulation model CRR 3 under the partial likelihood.

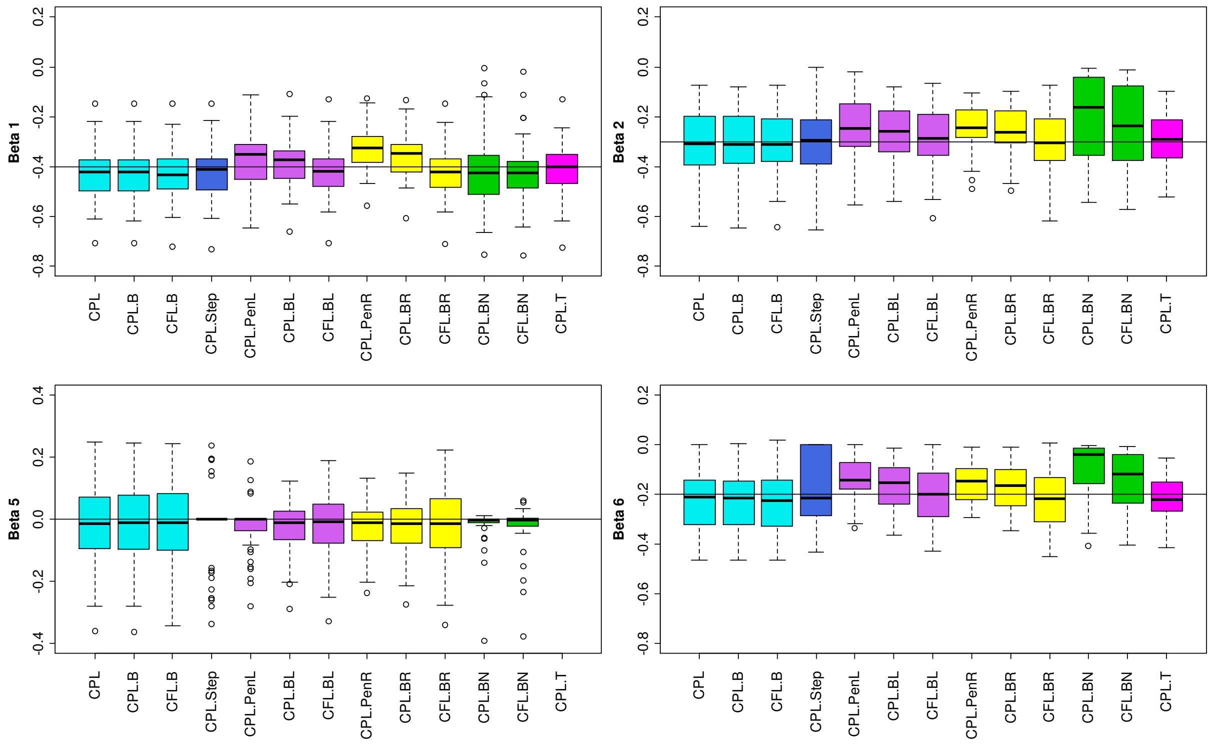


Figure 11.14: Regression coefficient estimates $\hat{\boldsymbol{\beta}}$ for four selected coefficients under different regularization priors in simulation model CRR 3. The right box (CPL.T) shows the estimations when the true predictor structure is used. The black horizontal lines in the box plots mark the values of the true regression coefficients $\beta_1 = -0.4, \beta_2 = -0.3, \beta_5 = 0, \beta_6 = -0.2$.

11.2. Low-dimensional nonlinear predictor

Data generation

With the subsequent simulations we explore the changes caused by the inclusion of a nonlinear effect in the predictor and more complex shapes of baseline hazard. The settings are similar to those in Hennerfeind et al. (2006). Again we consider $R = 50$ datasets, but now with an increased sample size of $n = 1000$ life times. Ten covariates are generated independently as random draws from an uniform $U[-3,3]$ distribution and the lifetimes are generated via the inversion method, compare Bender et al. (2005), from the model

$$\text{CRR 4}: \quad \lambda(t) = \lambda_0(t)\exp\left(\mathbf{x}'\boldsymbol{\beta} + \sin(x_{10})\right),$$

with the sinusoidal nonlinear effect $f_1(x_{10}) = \sin(x_{10})$ of covariate x_{10} and the linear effects

$$\boldsymbol{\beta} = (-0.7, -0.7, 0, 0, 0, -0.7, 0, 0, 0)'.$$

To model more flexible baseline hazards, a linear (but non-Weibull) baseline hazard of the form

$$\text{CRR 4.a}: \quad \lambda_0(t) = 0.25 + 2t$$

and a bathtub-shaped baseline hazard

$$\text{CRR 4.b}: \quad \lambda_0(t) = \begin{cases} 0.75\left(\cos(t) + 1.5\right), & t \le 2\pi \\ 0.75\left(1 + 1.5\right), & t > 2\pi \end{cases}$$

are chosen. The latter assumes an initially high baseline risk that decreases after some time and increases again later on until time $t = 2\pi$ from where the hazard stays constant.

Censoring times are generated in two steps. First, a random proportion of 17% of the generated observations T_i is assigned to be censored. Then, in the second step, the censoring times for this random selection are drawn from the corresponding uniform distributions $U[0, T_i]$. The difference to model CRR 1 is the additional inclusion of a nonlinear effect and the more complex shape of the baseline hazard function.

Function and parameter specification

Methods: Inference is based on the full likelihood and carried out with the `regress` method as implemented in `BayesX`. The logarithm of the baseline hazard as well as the nonlinear effect of covariate x_{10} are modeled with 20 cubic B-spline basis functions equipped with a second-order random walk prior for the associated basis function weights to control the smoothness.

Hyperparameters: The corresponding hyperparameters of both smoothing variances are set to the default values $h_{1,\tau_0} = h_{2,\tau_0} = 0.001$. In the Bayesian lasso and ridge prior for the linear effects the hyperparameters of the shrinkage parameters are set to $h_{1,\lambda} = h_{2,\lambda} = 0.001$ and those for the Bayesian NMIG prior are set as in the section before to $v_1 = 1$, $v_0 = 0.000025$, $h_{1,\psi} = 5$, $h_{2,\psi} = 25$, $h_{1,\omega} = 1$ and $h_{2,\omega} = 1$.

Estimation: We use an increased number of 30000 iterations with a burnin of 10000 and thin the chain by 20 which results in an MCMC sample of size 1000. The running times are about 6 minutes.

Results for model CRR 4.b

We briefly summarize the results for the models CRR 4 by means of model CRR 4.b with bathtub-shaped baseline due to the similarity of the results to each other and to the results of model CRR 1. We further restrict ourselves to the Bayesian methods based on the full likelihood with P-spline approximation for the log-baseline hazard. For the CRR model there are no R-packages available to perform frequentist lasso regression in combination with nonlinear effects. Ridge regression in combination with spline estimation is possible within the function `coxph()`, but the regularization parameters have to be specified in advance.

MSE linear effects

In **Figure 11.15** the MSEs of the estimated regression coefficients, $\mathrm{MSE}(\hat{\boldsymbol{\beta}})$, under the different regularization priors for the linear effects are shown together with the MSEs after the hard shrinkage selection criteria are applied. The MSE pattern is similar to those in **Figure 11.1**, which shows the corresponding results for model CRR 1. As before in model CRR 1, the Bayesian NMIG model performs better than the Bayesian lasso and ridge model regardless of whether hard shrinkage is applied or not, and the MSEs are close to the model estimated with the true predictor structure (CFL.BT).

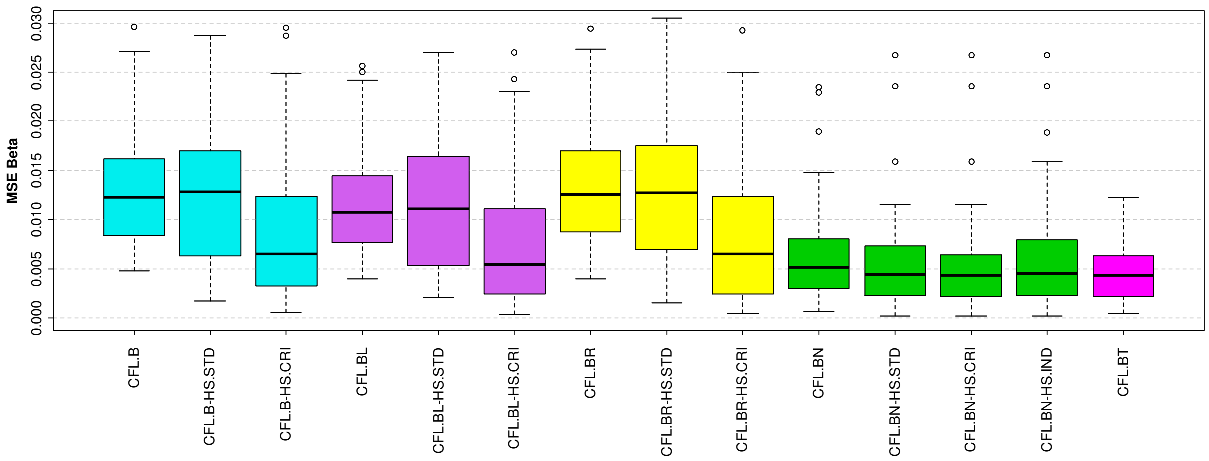


Figure 11.15: Mean squared errors of the regression coefficient estimates, $\mathrm{MSE}(\hat{\boldsymbol{\beta}})$, under the different regularization and variable selection methods in simulation model CRR 4.b. The right box (CFL.BT) shows the $\mathrm{MSE}(\hat{\boldsymbol{\beta}})$ for the estimations when the true predictor structure is used.

Linear effects

The estimates of three selected regression coefficients β_2, β_5 and β_6 are presented in **Figure 11.16**. As before the boxes are very similar under the different regularization methods, also the zero coefficient β_5 shows a higher concentration around zero under the Bayesian NMIG prior. Consequently the hard shrinkage rules applied to the Bayesian NMIG estimates are leading to comparable results with respect to the MSE since the non-influential covariates are assigned with effects very close to zero anyway.

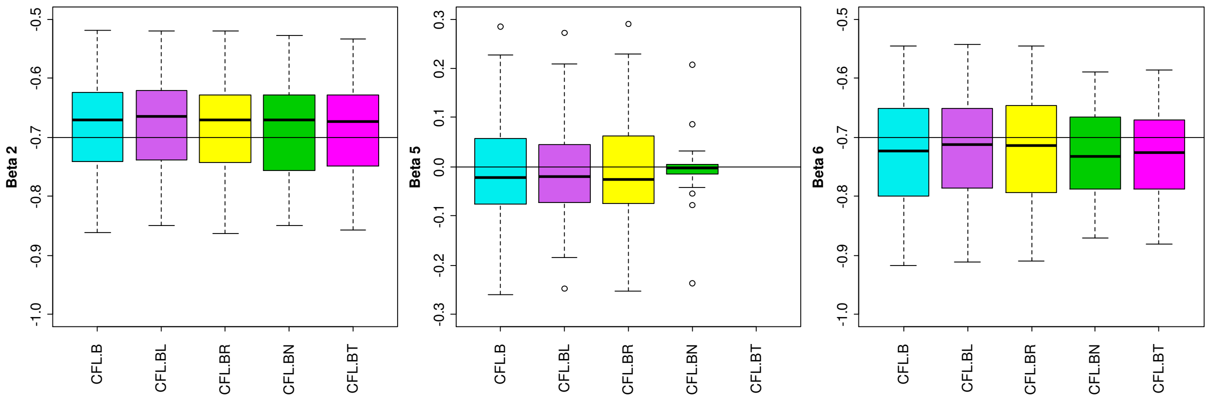


Figure 11.16: Regression coefficient estimates $\hat{\boldsymbol{\beta}}$ for three selected coefficients under different regularization priors in simulation model CRR 4.b. The right box (CFL.BT) shows the estimations when the true predictor structure is used. The black horizontal lines in the box plots mark the values of the true regression coefficients $\beta_2 = -0.7$, $\beta_5 = 0$, $\beta_6 = -0.7$.

NMIG indicators and penalties

As shown in **Figure 11.17,** also the comparison of the regularization-specific penalty and the Bayesian NMIG indicator variables are leading to similar results as in model CRR 1, when inference is based on the full likelihood. We see also that the posterior inclusion probabilities reflect the number of true nonzero and true zero effects very well, so that the HS.IND selection threshold 0.5 yields a sharp separation of the zero from the nonzero effects.

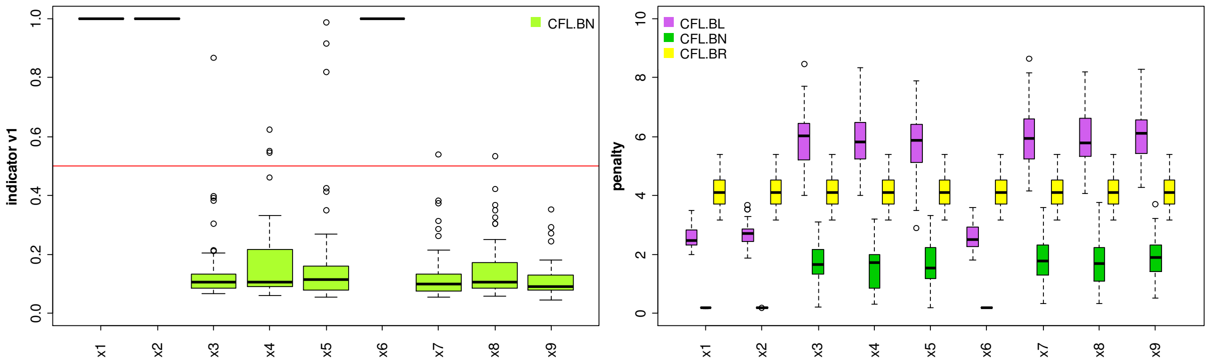


Figure 11.17: Left side: Estimated inclusion probabilities based on posterior relative frequencies of the Bayesian NMIG indicator variable value $I_j = v_1$ for simulation model CRR 4.b based on the full likelihood (CFL.BN). The red horizontal line marks the cut off value 0.5 of the hard shrinkage selection criterion HS.IND. Right side: Estimates of the covariate-specific penalty $\hat{\tau}^{-2}_{\beta_j}$ for the Bayesian lasso (CFL.BL), NMIG (CFL.BN) and ridge (CFL.BR) prior in simulation model CRR 4.b under the full likelihood.

Nonlinear effect and baseline hazard

The MSE results for the estimation of the log-baseline hazard and the nonlinear effect are displayed in **Figure 11.18** and the results are again almost comparable to each other with respect to the different regularization priors.

The shrinkage methods for the linear effects do not clearly affect the performance of the estimates of the nonlinear effect. In **Figure 11.19** (right panel), the estimated P-spline is visualized together with the 2.5% and 97.5% empirical quantiles for one selected dataset under the Bayesian NMIG prior as representative. **Figure 11.19** (left panel) shows the corresponding results for the baseline hazard estimation with the same selected dataset. We observe that the P-spline approximation of the log-baseline hazard performs very well in the time region where most of the events occur.

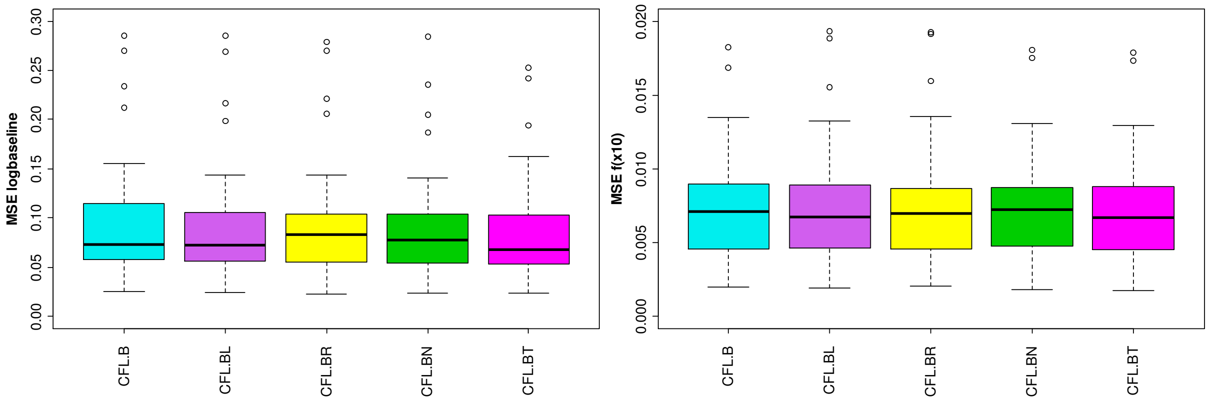


Figure 11.18: Mean squared errors for the estimated lo-baseline hazard, $\mathrm{MSE}(\log\hat{\lambda}_0)$, (left side) and the nonlinear effect, $\mathrm{MSE}(\hat{f}(x_{10}))$, (right side) under the different regularization priors in simulation model CRR 4.b.

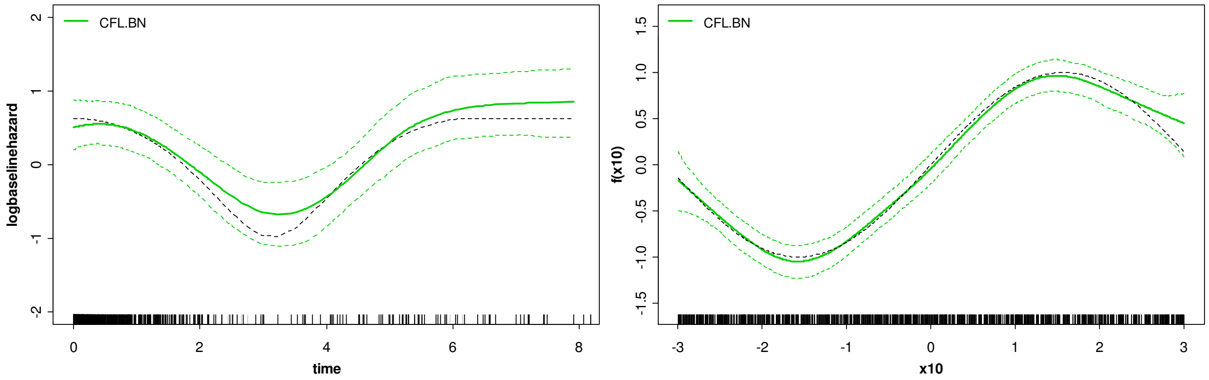


Figure 11.19: Estimate of the log-baseline hazard $\log\hat{\lambda}_0(t)$ (left side) and the nonlinear effect $\hat{f}_1(x_{10})$ (right side) under the Bayesian NMIG regularization in simulation model CRR 4.b for one selected data set. Left side: Estimation of the log-baseline hazard (solid green line) together with the 95% pointwise credible bands (dashed green lines). Right side: Estimation of the nonlinear effect (solid green line) together with the 95% pointwise credible bands (dashed green lines). In both figures the black dashed line marks the true log-baseline hazard and nonlinear effect and the vertical rugs at the time axis mark the observed event times (black) and censoring times (gray).

Deviance information criterion

If we take a look at the Deviance Information Criterion (DIC) and the effective number of parameters (pD), Spiegelhalter et al. (2002), that are given in **Figure 11.20**, all regularization priors yield a comparable DIC, but the Bayesian NMIG has the lowest effective number of parameters with value close to the effective number of parameters of the model with the “true predictor” structure (CFL.BT).

Classification

The frequencies of final models (MF) with the true predictor structure and the number of correctly classified zero and nonzero coefficients, when we apply the hard shrinkage selection rules, are collected in **Table 11.2.** Besides the results from the CRR model the results from the AFT model, described in the following subsection, are displayed. In summary the CFL results of model CRR 4.b are very close to those of model CRR 1 given in **Table 11.1** and the again the highest frequencies of models with the true predictor structure are obtained under the NMIG prior.

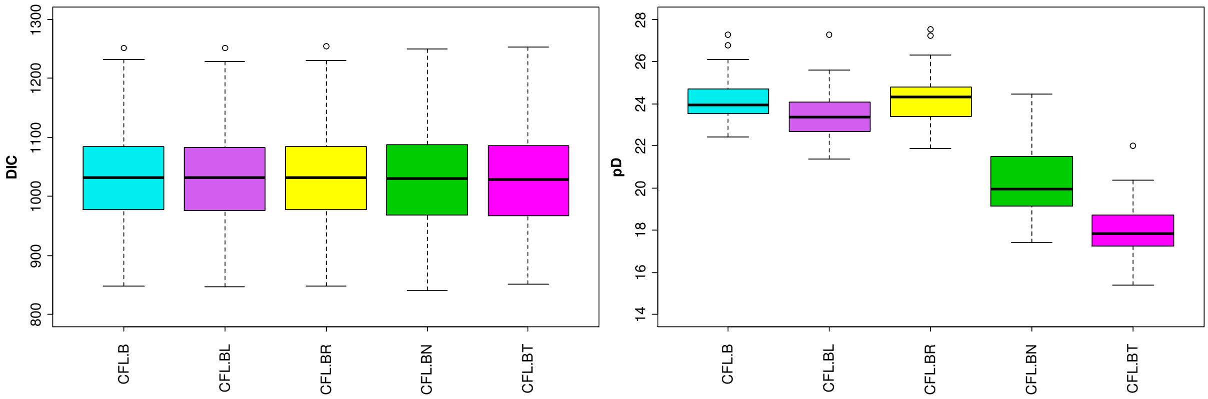


Figure 11.20: Deviance Information Criterion DIC (left side) and the effective number of parameters pD (right side) under the different regularization priors for simulation model CRR 4.b. The right box (CFL.BT) shows the results when the true predictor structure is used.

	Model 4.b (CFL)			Model 4.b (PGM)		
	$\hat{\beta} \neq 0$ $\beta \neq 0$	$\hat{\beta} = 0$ $\beta = 0$	MF	$\hat{\beta} \neq 0$ $\beta \neq 0$	$\hat{\beta} = 0$ $\beta = 0$	MF
BEST	3	6	50	3	6	50
B.HS-STD	3	3.94	4	3	3.90	6
BL.HS-STD	3	4.06	5	3	4.32	10
BR.HS-STD	3	3.96	5	3	3.94	8
BN.HS-STD	3	5.90	45	3	5.92	47
B.HS-CRI	3	5.62	35	3	5.56	35
BL.HS-CRI	3	5.70	37	3	5.72	38
BR.HS-CRI	3	5.58	35	3	5.68	36
BN.HS-CRI	3	5.96	48	3	6.00	50
BN.HS-IND	3	5.82	43	3	5.92	47

Table 11.2: Average number of correctly classified coefficients for the models CRR 4.b after variable selection. CFL marks the estimates based on the full likelihood of the CRR model and PGM marks the estimates of the AFT model with PGM error distribution. Especially $\hat{\beta} \neq 0, \beta \neq 0$ denotes the case that the estimated effect is nonzero ($\hat{\beta} \neq 0$) when the corresponding true effect is nonzero ($\beta \neq 0$), and $\hat{\beta} = 0, \beta = 0$ denotes the case that the estimated effect is zero ($\hat{\beta} = 0$) when the corresponding true effect is zero ($\beta = 0$). The columns (MF) display the average frequencies of the final models that recover the true model.

11.3. Miss-specification using the AFT model

To investigate the loss of performance when the AFT model with penalized Gaussian mixture (PGM) as baseline error distribution is used to fit data generated from a CRR model, we revisit the simulation scenario of the CRR model 4.b. Due to the parameterization of the AFT model, the estimates of linear and nonlinear effects are multiplied with -1 to simplify the visual comparison of the results from the CRR and AFT model.

Function and parameter specification

Methods: For the Bayesian estimation of the error distribution density we use the function `baftpgm()` with update scheme *"sliceR0"*, where the error density is specified as in the Simulation Section 10.2 through $g_0 = 21$ equidistant knots m_j, $j = 1,...,g_0$, that are placed in the interval $[-4.5, 4.5]$. The variance of the Gaussian basis functions is uniformly set to $s_j^2 = 0.25^2$. We use again the third-order random walk penalty to control the smoothness of the baseline error distribution.

The linear effects of the covariates $x_1,...,x_9$ are regularized with the Bayesian lasso, ridge and NMIG prior. Further, the nonlinear effect of covariate x_{10} is modeled by a Bayesian P-spline with $g_1 = 20$ cubic B-spline basis functions and a second-order random walk prior, which matches the setting used before in Section 11.2.

Hyperparameters: The hyperparameters of the prior associated to the scale parameter σ^2 are set to $h_{1,\sigma} = h_{2,\sigma} = 0.001$, and those of the smoothing variance $\tau^2_{\alpha_0}$ are $h_{1,\tau_0} = 1$, $h_{2,\tau_0} = 0.01$. For the regularization priors of the covariate effects we use the hyperparameter setting used in the previous Section 11.2 for model CRR 4.b.

Starting values: The starting values are set as in the Simulation Section 10.2, i. e. for the transformed error weights $\alpha_{0,j}$, $j = 1,...,21$, with exception of the middle weight $\alpha_{0,11} := 0$, each starting value is set to $\alpha^{(0)}_{0,j} = 0.01$. The location and scale parameter start in $\gamma_0^{(0)} = 1$, $\sigma^{2(0)} = 1$ and the smoothing variance is set to $\tau^{2(0)}_{\alpha_0} = 1$. The component labels $r_i^{(0)}$ are randomly assigned to one of the g_0 error basis densities and the starting values of the linear effects are set to $\beta_j^{(0)} = 0.01$, $j = 1,...,9$. For the Bayesian NMIG regularization the sampler starts with $I_j^{(0)} = v_0$, $\psi_j^{2(0)} = 0.0416$, which is the left mode of the bimodal NMIG variance parameter prior and $\omega^{(0)} = 0.5$. The shrinkage parameter for the Bayesian lasso and ridge starts with $\lambda^{(0)} = 1$. The nonlinear effect starts with $\alpha^{(0)}_{1,10} = ... = \alpha^{(0)}_{20,10} = 0.01$ and $\tau^{2(0)}_{\alpha_{10}} = 1$.

Estimation: To fit the models, we use 20000 iterations, where the first 10000 iterations are discarded as burnin of the Markov chain and the remaining iterations are thinned using a step width of 10. The resulting 1000 states of the chain build the sample of the posterior distribution and the empirical basis to compute the estimates. The running times of the sampler are about 32 minutes.

Results

MSE of the linear effects

Figure 11.21 illustrates the MSE of the estimated linear effects under the different regularization priors together with the resulting MSEs, when the hard shrinkage selection rules are applied to obtain sparse final models. While the level of the MSEs is generally larger than in the CRR 4.b model, the results here show a similar MSE structure as those in **Figure 11.15** or **Figure 11.1**.

The application of the hard shrinkage selection rules improves the performance with respect to all priors with only small improvement for the Bayesian NMIG model that is almost close to the model with the true predictor PGM.BT. Especially in the AFT model, the Bayesian lasso prior in combination with the HS.CRI rule performs very well with lower MSEs compared to PGM.BT model based on the true predictor structure.

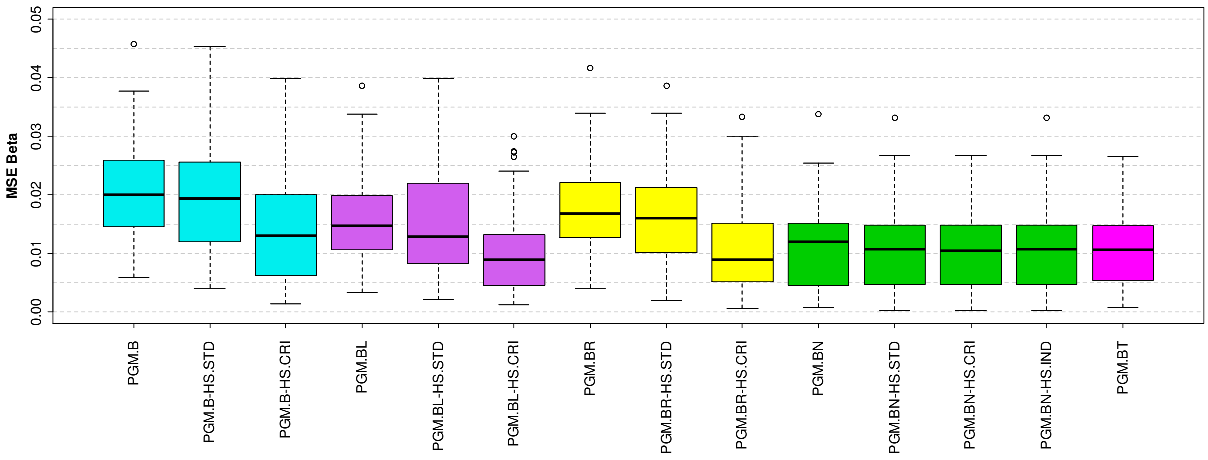


Figure 11.21: Mean squared errors of the regression coefficient estimates, $\mathrm{MSE}(\hat{\boldsymbol{\beta}})$, under the different regularization and variable selection methods in simulation model CRR 4.b under the AFT model with PGM error. The right box (CFL.BT) shows the $\mathrm{MSE}(\hat{\boldsymbol{\beta}})$ for the estimations when the true predictor structure is used.

Linear effects

The basic increase in the MSE of the regression coefficients is explained by **Figure 11.22** that shows the box plots of the three selected estimated linear effects $\beta_2 = -0.7$, $\beta_5 = 0$ and $\beta_6 = -0.7$ under the different regularization priors.

We observe larger deviations from the true values of the effects as when the CRR model is used for inference, compare **Figure 11.16**, even for the model PGM.BT with the true predictor structure. Further, the differences caused by the various regularization priors are more pronounced in the AFT model. The absolute values of the estimates for the regression coefficient β_2 are by trend larger than the true value in the AFT model, and smaller than the true value in the CRR model (**Figure 11.16**). This highlights again the dependence of the regularized estimates on both, the prior and the likelihood, and that identical prior specifications can lead to a different shrinkage behavior if the regression model is exchanged. In particular, this fact complicates the tuning of the NMIG prior in terms of the absolute sizes of the regression coefficients that should fall into strong regularized prior area around origin, compare threshold (ISP_β) in Section 4.3.2.

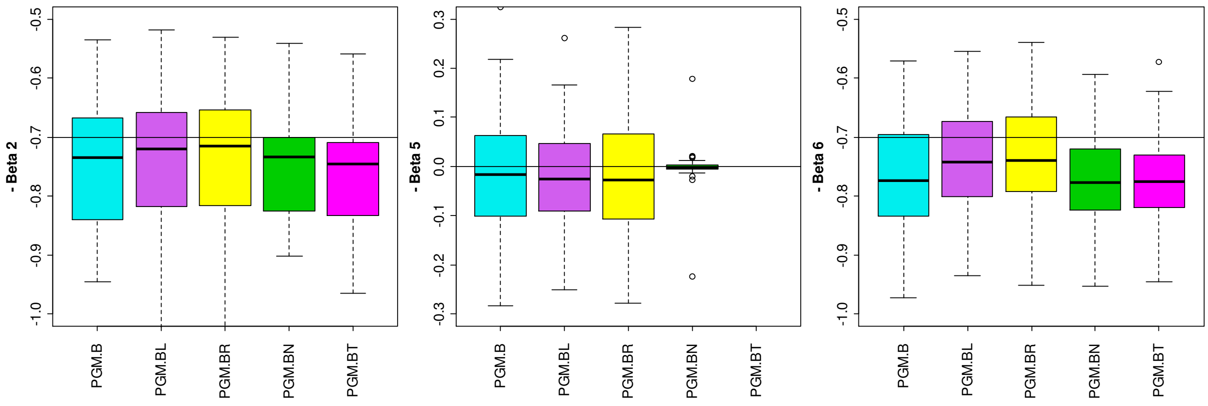


Figure 11.22: Regression coefficient estimates $\hat{\boldsymbol{\beta}}$ for three selected coefficients under different regularization priors in simulation model CRR 4.b under the AFT model with PGM error. The right box (PGM.BT) shows the estimations when the true predictor structure is used. The black horizontal lines in the box plots mark the values of the true regression coefficients $\beta_2 = -0.7$, $\beta_5 = 0$, $\beta_6 = -0.7$.

NMIG indicators and penalties

Figure 11.23 shows the posterior relative frequencies of the Bayesian NMIG indicator value $I_j = v_1$ (left side) and the covariate-specific penalties under the three regularization priors (right side). Similar to the previous simulation results for model CRR 1 and CRR 4 we obtain a clear separation of the nonzero effects from the zero effects if the AFT model is used to fit the data.

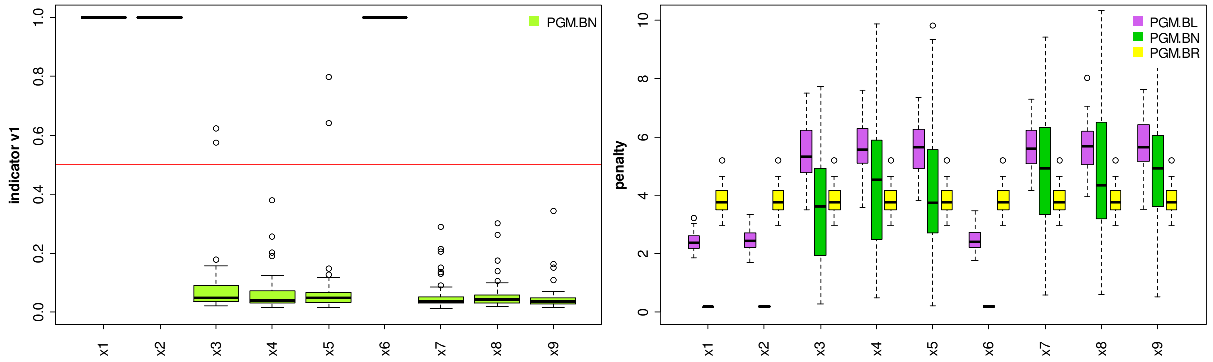


Figure 11.23: Left side: Estimated inclusion probabilities based on posterior relative frequencies of the Bayesian NMIG indicator variable value $I_j = v_1$ for simulation model CRR 4.b under the AFT model with PGM error. The red horizontal line marks the cut off value 0.5 of the hard shrinkage selection criterion HS.IND. Right side: Estimates of the covariate-specific penalty $\hat{\tau}_{\beta_j}^{-2}$ for the Bayesian lasso (PGM.BL), NMIG (PGM.BN) and ridge (PGM.BR) prior in simulation model CRR 4.b under the AFT model with PGM error.

Classification

The classification results, if the three hard shrinkage selection rules are applied, are displayed in **Table 11.2**, which are in summary comparable to those, if a CRR model is used to fit the data. Again, the best performances are obtained with the Bayesian NMIG prior in combination with the three hard shrinkage selection rules and the optimal value is reached with the HS.CRI criterion.

MSE of the nonlinear effect

Further, there is no observable impact on the performance of the estimated nonlinear effect, caused by the different regularization priors for the linear effects. As already noticed for the linear effects, also the level of the $\text{MSE}(\hat{f}_1(x_{10}))$ under the AFT model is generally larger than under the CRR model. We obtain across the unregularized and regularized methods values with lower and upper quartiles in the range of 0.011 and 0.026 and the median values are in the range 0.018.

Nonlinear effects and baseline quantities

Finally, **Figure 11.24** shows for one selected dataset the estimated nonlinear effect (right side) and the associated estimated baseline error distribution density (left side), each with the 95% pointwise credible bands.

In summary we have seen, that the miss-specification of the survival model causes a loss of performance in terms of the MSE of the estimated predictor components. Nevertheless, the regularization priors for the predictor components together with the hard shrinkage rules are leading, under a setting like in model CRR 1 or CRR 4, to comparable results if an AFT or CRR model is used for fitting.

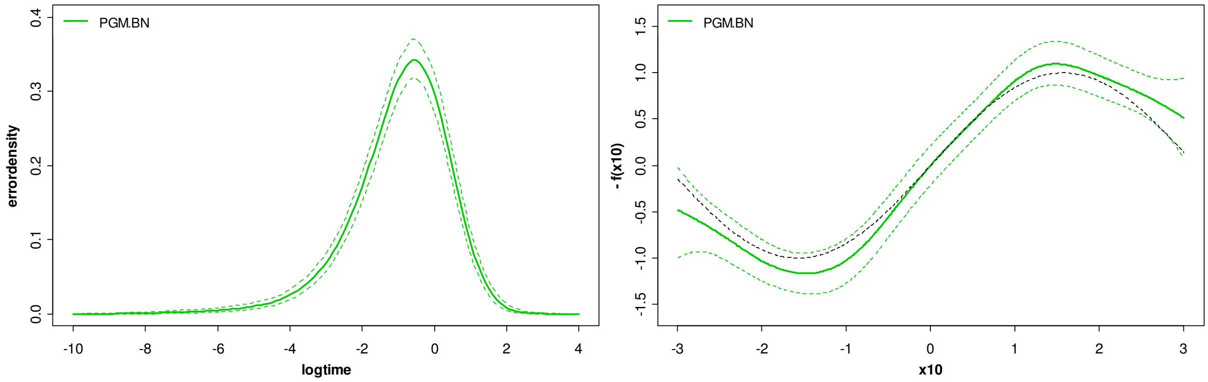


Figure 11.24: Estimate of the baseline error density $\hat{f}_0(y)$ (left side) and the nonlinear effect $\hat{f}_1(x_{10})$ (right side) under the AFT model with Bayesian NMIG regularization in simulation model CRR 4.b for one selected data set. Left side: Estimation of the baseline error density (solid green line) together with the 95% pointwise credible bands (dashed green lines). Right side: Estimation of the nonlinear effect (solid green line) together with the 95% pointwise credible bands (dashed green lines). In both figures the black dashed line marks the true log-baseline hazard and nonlinear effect.

11.4. High-dimensional predictor

Data generation

To investigate the performance of the Bayesian regularization priors in the high-dimensional case, where the number of covariates exceeds the number of observations, we consider again the CRR model with exponential baseline hazard $\lambda_0(t)=1$ as used in the first two simulations of this section. As before, covariates are generated with zero mean, unit variance and $\text{corr}(x_{i,j},x_{i,k})=\rho^{|j-k|}$ with $\rho=0.5$ as correlation between $\mathbf{x}_j$ and $\mathbf{x}_k$. In the basic setting, survival times T_i, $i=1,...,n$, are generated from the model

$$\lambda(t)=\exp(\mathbf{x}'\boldsymbol{\beta}),$$

with the $p_x=20$ regression coefficients

$$\boldsymbol{\beta}=(-0.7,-0.7,0,0,-0.5,-0.5,0,0,-0.3,-0.3,0,0,-0.2,-0.2,0,0,-0.1,-0.1,0,0)'. \qquad (11.1)$$

The number of covariates is increased to $p_x=60,160,200$ and the vector of linear effects $\boldsymbol{\beta}$ in (11.1) is repeatedly pasted back-to-back until the associated number of linear effects is attained. We fix the number of observations to $n=160$ and use again $R=50$ replicated datasets. Censoring times are generated as i.i.d. draws from a uniform $U[0,6]$ distribution until 25% censored observations in the data are achieved.

Function and parameter specification

Methods: Bayesian and frequentist inference is carried out in terms of the regularized partial likelihood. Bayesian inference is practiced with the function `bcoxpl()`. As competitor we use the frequentist lasso and ridge regularization, carried out with the `penalized()` function, together with the frequentist model CPL.T, utilizing `coxph()`, that includes only the covariates with true nonzero effects of the predictor.

Estimation: The number of iterations in the Bayesian MCMC sampler is set to 20000 with a burnin of 5000 and a thinning by 15, resulting in 1000 samples from the posterior distribution. We observe, e. g. for Bayesian lasso, the following average runtimes: 6 minutes ($p_x = 20$), 11 minutes ($p_x = 60$), 25 minutes ($p_x = 160$) and 35 minutes ($p_x = 200$) on a system with quad-core CPU (Intel Quad9550, 2.83 GHz).

Hyperparameters: In the lower-dimensional cases, $p_x = 20, 60$, the previous setting of the prior hyperparameters is used, i. e., in the Bayesian lasso and ridge prior we set $h_{1,\lambda} = h_{2,\lambda} = 0.01$ and the Bayesian NMIG prior is specified with $v_1 = 1$, $v_0 = 0.000025$, $h_{1,\psi} = 5$, $b_{2,\psi} = 25$, $h_{1,\omega} = 1$ and $h_{2,\omega} = 1$. The block size is set to p_x in each simulation run.

To achieve convergence in the higher-dimensional cases, $p_x = 160, 200$, the shrinkage priors require a tuning to control the regularization. If $p_x = 160$, we set the hyperparameters of the inverse gamma prior for the shrinkage parameter to $h_{1,\lambda} = 1000$, $h_{2,\lambda} = 10$ for the Bayesian lasso and to $h_{1,\lambda} = 440$, $h_{2,\lambda} = 20$ for the Bayesian ridge prior. For the Bayesian NMIG complexity parameter ω we specify a beta prior with $h_{1,\omega} = 300$ and $h_{2,\omega} = 1200$. If $p_x = 200$, we set for the Bayesian lasso $h_{1,\lambda} = 6400$ and $h_{2,\lambda} = 40$, for the Bayesian ridge $h_{1,\lambda} = 900$ and $h_{2,\lambda} = 30$, and for the Bayesian NMIG $h_{1,\omega} = 300$ and $b_{2,\omega} = 1500$.

These hyperparameters are found in several runs with various hyperparameter constellations. E. g. with the initial hyperparameter setting, the estimated values of the NMIG complexity parameter ω decrease with increased number of covariates, and the estimates $\hat{\omega}$ are close to zero in the models with $p_x \geq 160$ covariates. Consequently, the posterior inclusion probabilities and regression coefficient estimates are close to zero, too. To counterbalance the strong regularization, we use hyperparameter constellations $h_{1,\omega}, h_{2,\omega}$ which are leading to a prior mean of $H_\omega \approx 0.2$ and an estimated value $\hat{\omega}$ of the same magnitude. With the described hyperparameters we obtain for the estimates $\hat{\omega}$ at last the following median values: $\hat{\omega} \approx 0.36$ (if $p_x = 20$), $\hat{\omega} \approx 0.30$ (if $p_x = 60$), $\hat{\omega} \approx 0.20$ (if $p_x = 160$) and $\hat{\omega} \approx 0.18$ (if $p_x = 200$). Another line of action is used to determine the hyperparameters of the shrinkage parameter in the Bayesian lasso and ridge prior. With the initial setting the sample-paths of the shrinkage parameters are very wiggly, but the paths do not diverge. We select the hyperparameters to obtain shrinkage parameter estimates close to the mean estimate that results from the initial setting, and stable sample-paths of the parameter estimates. Due to the resulting high informative prior setting, the estimates of the shrinkage or complexity parameter show a clearly decreased variability in the replications.

Starting values: In the lower-dimensional cases ($p_x = 20, 60$) the starting values are set as in Subsection 11.1. In the higher-dimensional cases ($p_x = 160, 200$) we use the modified values $\omega^{(0)} = 0.2$ (NMIG prior), $\lambda^{2(0)} = 10$ (lasso prior) and $\lambda^{(0)} = 20$ (ridge prior).

Results

MSE of the linear effects

Figure 11.25 shows the resulting mean squared errors of the estimated regression coefficients, $\text{MSE}(\hat{\boldsymbol{\beta}})$, under the lasso, ridge and NMIG regularization, when the number of covariates included in the regression model increases from $p_x = 20$ (upper left panel) to $p_x = 200$ (lower right panel) together with the resulting MSEs, when the hard shrinkage selection criteria HS.STD, HS.CRI and HS.IND are applied to the Bayesian estimates. The MSEs are standardized by division with the number of covariates in the model.

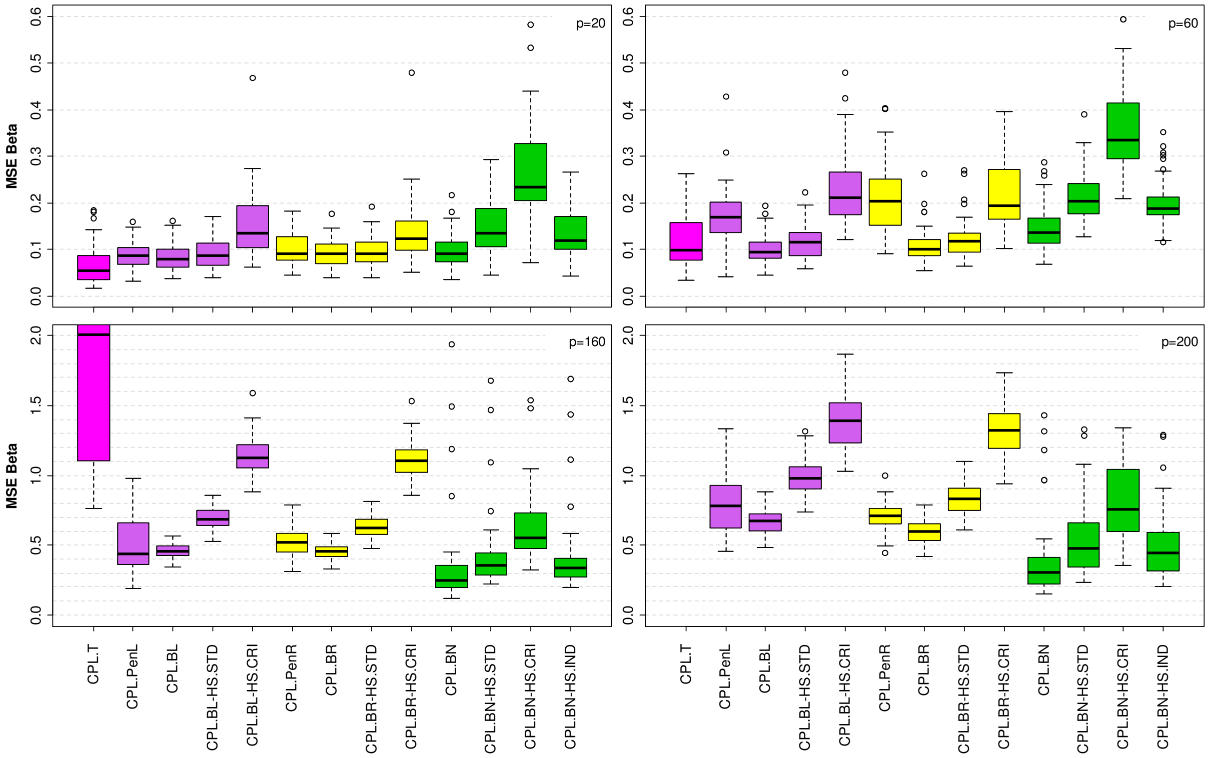


Figure 11.25: Mean squared errors of the regression coefficient estimates, $\text{MSE}(\hat{\boldsymbol{\beta}})$, under the different regularization and variable selection methods in the CRR model with increasing number of covariates. The right box (CPL.T) shows the $\text{MSE}(\hat{\boldsymbol{\beta}})$ for the maximum partial likelihood estimations when the true predictor structure is used.

As to be expected, we have an increased loss of MSE performance across the regularization methods, when the number of effects increases. We also observe that variable selection in the Bayesian models does not improve the predictive performance, and that the MSEs of the sparse final models CPL.BN-HS.STD and CPL.BN-HS.IND are almost comparable. The loss in the predictive performance induced by the variable selection increases as the number of covariates is increased. A similar result is obtained for the frequentist lasso (CPL.PenL) that always provides sparse models. If we compare the frequentist lasso and ridge models (CPL.PenR), we find also that the MSE performance of the ridge models, which include all covariates in the predictor, is almost comparable to those of the lasso models.

In the low-dimensional case ($p_x = 20$) the performance of the regularized models is almost comparable, but we observe a marginal higher performance for the Bayesian models (CPL.BL, CPL.BR, CPL.BN). Nevertheless, all MSEs are larger than the MSE of the frequentist model with the true predictor structure (CPL.T). Increasing the number of covariates ($p_x = 60$) is leading to clearly higher performances of the Bayesian models compared to the frequentist lasso and ridge models, and the MSEs of the Bayesian models are close to the MSE of the CPL.T model. Within the Bayesian models the lasso model (CPL.BL) has the best performance followed by the ridge (CPL.BR) and

NMIG model (CPL.BN), and the HS.STD criterion, applied to the Bayesian lasso and ridge models, is leading to sparse models with only a marginal loss of predictive performance.

While the performance of the Bayesian ridge and lasso models slightly dominates the performance of the Bayesian NMIG in the lower-dimensional cases ($p_x = 20,60$), this result is reversed in the higher-dimensional cases ($p_x = 160,200$), where the Bayesian NMIG models achieve the lowest MSE values. Interestingly, in the high-dimensional case the variable selection based on posterior relative frequencies of the Bayesian NMIG indicator variable value $I_j = v_1$ (CPL.BN-HS.IND) is leading to final models with lower MSE values, compared to the models from the frequentist and Bayesian lasso and ridge regularization.

NMIG indicators

Figure 11.26 displays for the first 20 covariate effects β_j, $j=1,...,20$, the estimated inclusion probabilities given by the posterior relative frequencies of the associated Bayesian NMIG indicator variable values $I_j = v_1$, $j=1,...,20$, when the dimension p_x increases.

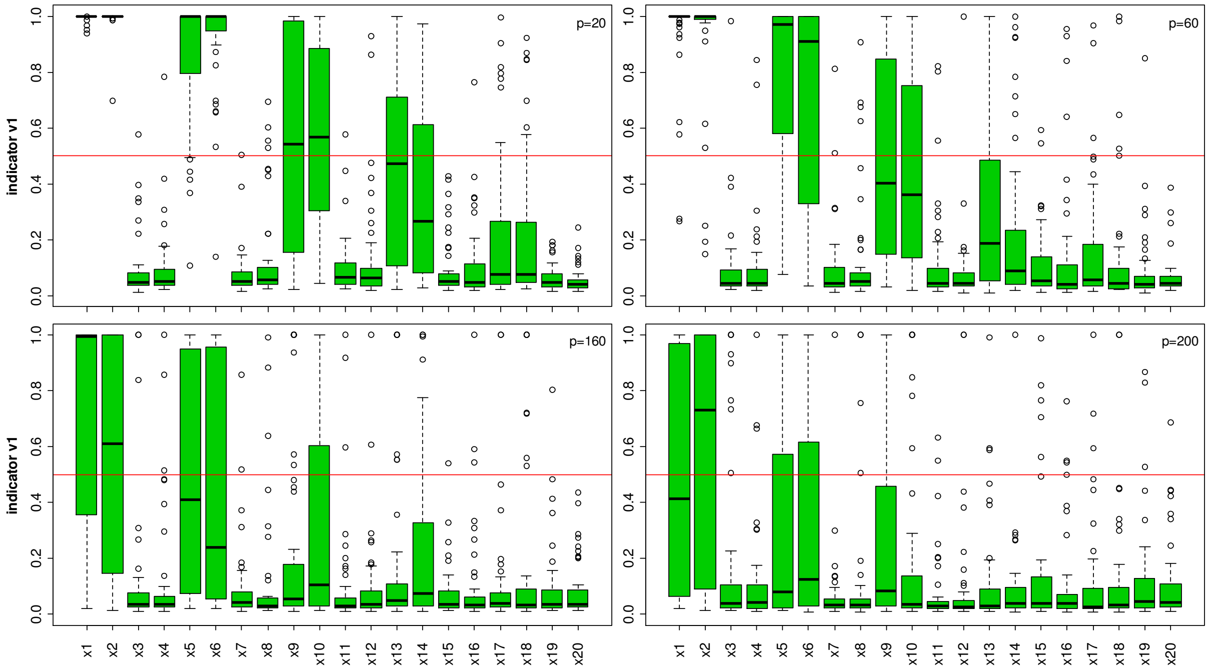


Figure 11.26: Estimated inclusion probabilities based on posterior relative frequencies of the Bayesian NMIG indicator variable value $I_j = v_1$ for CRR model with increasing number of covariates. The red horizontal line marks the cut off value 0.5 of the hard shrinkage selection criterion HS.IND.

In the lower-dimensional cases ($p_x = 20,60$) the structure of the inclusion probabilities fits well to the effect sizes, i. e. the inclusion probabilities decrease if the size of the effects decrease and the inclusion probabilities are small for the zero effects. In particular, in the case $p_x = 20$ (upper left panel) we can clearly separate, in terms of the median inclusion probability, the effects $\beta_{13} = \beta_{14} = 0.3$ from the zero effects, and the cut off value 0.5 of the HS.IND selection rule separates nonzero effects in the range of 0.2 (β_9, β_{10}) to 0.3 (β_{13}, β_{14}). When the number of covariates increases, the inclusion probabilities of larger effects decrease. Especially when the number of covariates exceeds the number of observations ($p_x \geq 160$), even the inclusion probabilities of the comparably large effects $\beta_1 = \beta_2 = -0.7$ and

$\beta_5 = \beta_6 = -0.5$ fall below of our standard selection threshold of 0.5. Possibly an adaption of the HS.IND-threshold to smaller values improves the predictive performance of the CPL.BN-HS.IND models. We consider such adaptations in the following Subsection 11.5 and we will find, compare **Figure 11.31**, that an adjustment to the threshold 0.2 indeed improves the predictive performance. This holds also in the higher-dimensional cases, where we can hardly separate the smaller from the zero effects, but here we force mainly the inclusion of the lager effects $\beta_1 = \beta_2 = -0.7$.

In each of the four simulation models the fraction and size of the nonzero effects in the predictor is identical, and we would expect a comparable model complexity. But, the decrease of the inclusion probabilities with increased number of covariates is also reflected in the decreased estimated values of the complexity parameter ω. In the lower-dimensional cases $p_x \le 60$ we used an uniform prior for the parameter ω. The estimated values $\hat{\omega}$ are concentrated at 0.3, if $p_x = 20$, and at 0.2, if $p_x = 60$. With the hyperparameter setting in the higher-dimensional cases $p_x \ge 160$ we set with the beta prior the focus on complexity parameter values in the range of 0.2, but we observe a further decrease of the inclusion probabilities with almost comparable values of the complexity parameter estimate. So, in the low-dimensional case the adjustment of the HS.IND selection threshold value to smaller values than 0.5 may be an ad hoc solution to improve the detection of the true nonzero effects, but in the higher-dimensional case an adjustment of the prior is required to enhance the detection. We obtain a marginal improvement by a further adjustment of the hyperparameters $h_{1,\omega}$, $h_{2,\omega}$ to force a higher model complexity, but the improvement is limited, since larger values of the complexity parameter ω increases also the inclusion probability of the smaller effect and blurs the separation of small and zero effects. E. g., if we fix the value $\omega = 0.5$ in the case of $p_x = 200$ covariates, we observe for some of the zero effects inclusion probabilities of the same magnitude as the larger effects and variable selection on the base of the HS.IND criterion leads to very low rates of correctly classified regression coefficients ($\hat{\beta} \ne 0, \beta \ne 0$, $\hat{\beta} = 0, \beta = 0$). Nevertheless, if performance is measured in terms of the MSE instead of a high classification rate, we achieve good results with the used prior specifications without variable selection.

Penalties

The observed trends in the evolution of the NMIG inclusion probabilities are also reflected in associated penalties $\tau_{\beta_j}^{-2}$ that are displayed in **Figure 11.27** (green boxes). We observe an increase in the penalty values for the larger effects, when the number of covariates increases, and in the resulting estimates of the larger regression coefficients, **Figure 11.29** (green boxes), are stronger shrunken towards zero. In the two lower-dimensional cases the amount of penalization increases under the Bayesian lasso and ridge prior only marginally if the number of covariates is increased from $p_x = 60$ to $p_x = 160$. In the higher-dimensional cases the penalty is mainly determined by our informative hyperparameter setting and is clearly increased.

Figure 11.28 shows the penalty values λ of the frequentist lasso and ridge regularization. We observe that the penalty values of the frequentist ridge regression tend for each dimension to larger values compared to the Bayesian counterpart. For the frequentist lasso the penalty varies in the lower-dimensional cases within the range of the covariate-specific Bayesian lasso penalties, and in the higher-dimensional cases the frequentist lasso penalty is clearly smaller than the penalties of the Bayesian counterpart.

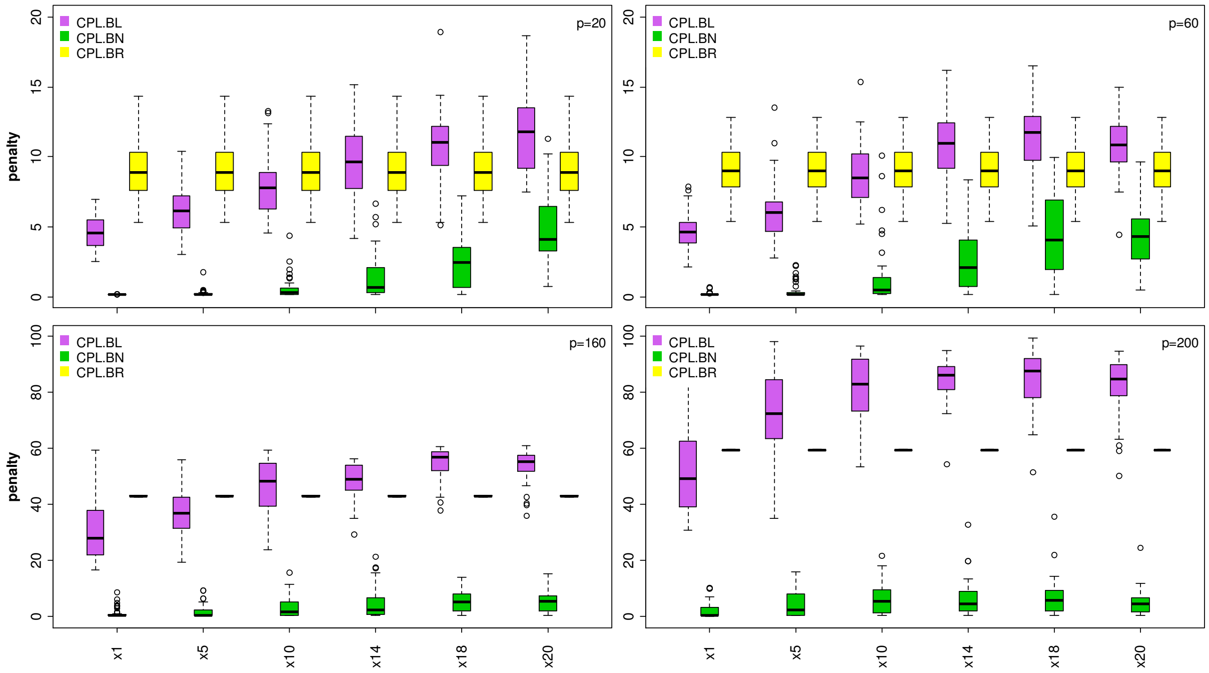


Figure 11.27: Estimates of the covariate-specific penalty $\hat{\tau}^{-2}_{\beta_j}$, $j = 1,5,10,14,18,20$, of six selected covariates under the different Bayesian regularization priors in the CRR models with increasing number of covariates.

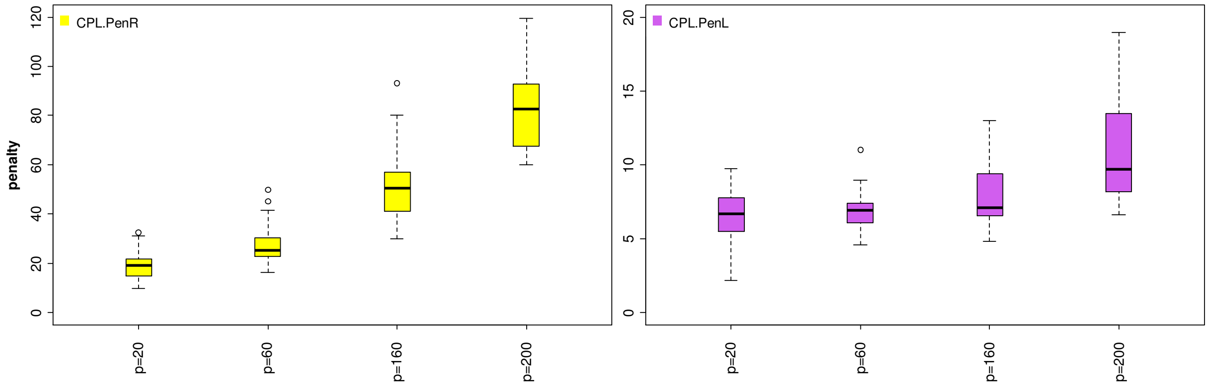


Figure 11.28: Estimated penalty parameter $\hat{\lambda}$ under the frequentist ridge (left side) and lasso (right side) regularization in the CRR models with increasing number of covariates.

Linear effects

The impact on the shrinkage of the regression coefficient estimates, induced by the different ranges of the penalty values, is summarized in **Figure 11.29** by means of four selected regression coefficients ($\beta_1 = -0.7, \beta_{10} = -0.3, \beta_{14} = -0.2, \beta_{20} = 0$) of different sizes.

If we compare the estimates of the largest coefficient β_1 in the higher-dimensional cases, we see the reduced shrinkage of this effect under the Bayesian NMIG prior. In contrast, the shrinkage of the other effects is more pronounced as under the remaining regularization methods. In summary, the resulting smaller deviations of the estimates from the true value -0.7 of the larger effects in the predictor are the main reason for the lower MSE of the NIMG models.

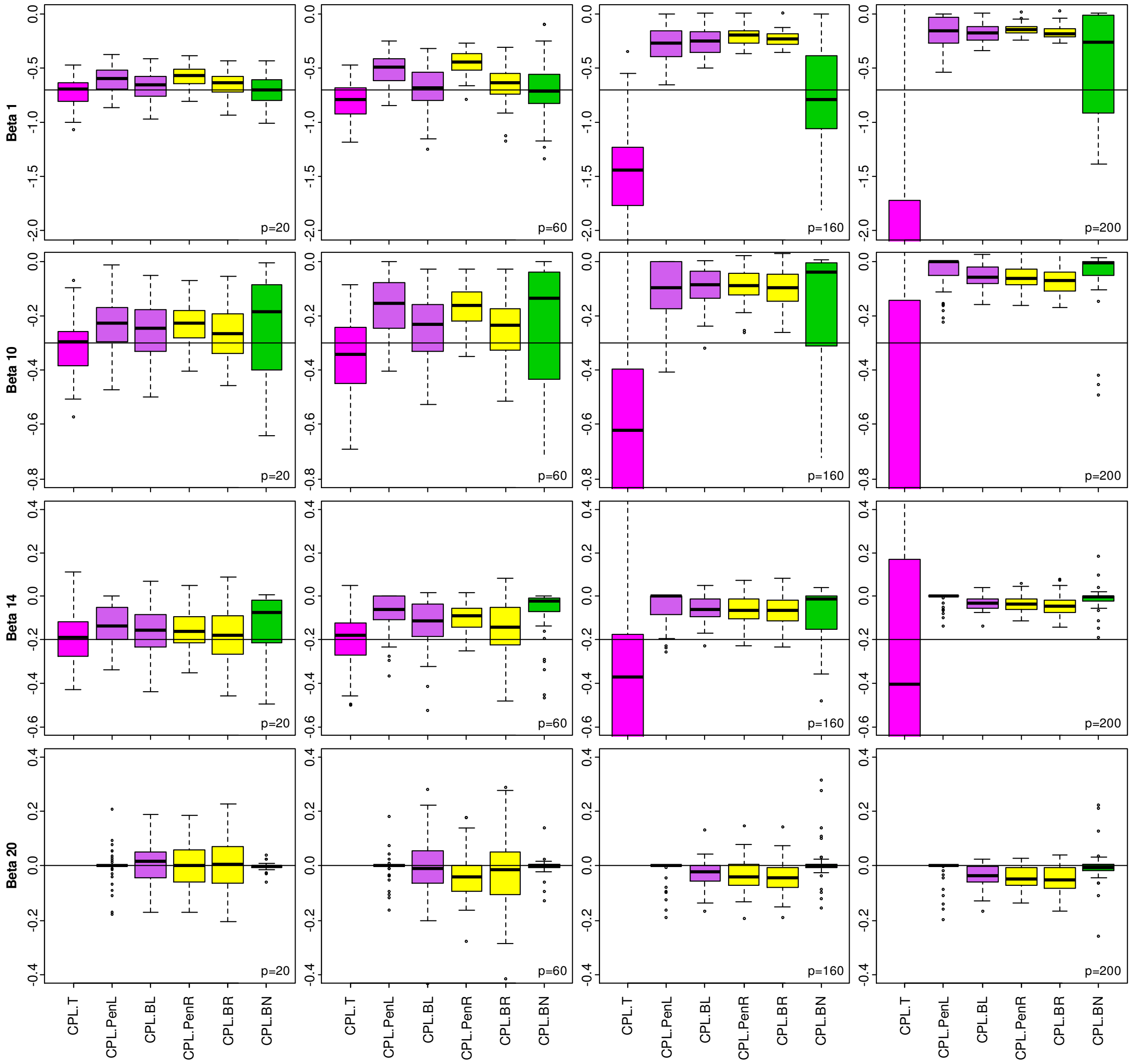


Figure 11.29: Regression coefficient estimates $\hat{\boldsymbol{\beta}}$ of four selected covariates under different regularization priors for the CRR models with increasing number of covariates. The left box (CPL.T) shows the estimations when the true predictor structure is used. The black horizontal lines in the figures mark the values of the true regression coefficients $\beta_1 = -0.7, \beta_{10} = -0.3, \beta_{14} = -0.2, \beta_{20} = 0$.

Classification

When the hard shrinkage selection rules are applied, we observe an impact on the average fraction of correctly classified nonzero effects, $\hat{\beta} \neq 0, \beta \neq 0$, which decreases clearly under all regularization priors if the number of covariates increases, compare classification **Table 11.3**. The highest fractions of correctly classified nonzero effects are achieved with the frequentist lasso regularization (CPL.PenL), followed by Bayesian ridge regularization in combination with the standard deviation based rule (CPL.BR-HS.STD), but the associated sparse final models are in general not the models with the best performance in terms of the MSE of the regression coefficients. E. g. in the high-dimensional case $p_x = 200$ the frequentist lasso CPL.PenL detects on average twice as much true nonzero effects than Bayesian NMIG prior in combination with the HS.IND selection rule. But, the resulting estimated model yields a larger value of the $\text{MSE}(\hat{\boldsymbol{\beta}})$, with a range twice as large as the range of the final model achieved with CPL.BN-HS.IND.

	$p_x = 20$		$p_x = 60$		$p_x = 160$		$p_x = 200$	
	$\hat{\beta} \neq 0$ $\beta \neq 0$	$\hat{\beta} = 0$ $\beta = 0$	$\hat{\beta} \neq 0$ $\beta \neq 0$	$\hat{\beta} = 0$ $\beta = 0$	$\hat{\beta} \neq 0$ $\beta \neq 0$	$\hat{\beta} = 0$ $\beta = 0$	$\hat{\beta} \neq 0$ $\beta \neq 0$	$\hat{\beta} = 0$ $\beta = 0$
BEST	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5
CPL.PenL	0.445	0.288	0.411	0.332	0.299	0.380	0.220	0.416
CPL.BL-HS.STD	0.393	0.396	0.358	0.406	0.187	0.477	0.120	0.486
CPL.BR-HS.STD	0.404	0.348	0.383	0.346	0.238	0.437	0.196	0.450
CPL.BN-HS.STD	0.287	0.492	0.229	0.485	0.142	0.473	0.103	0.474
CPL.BL-HS.CRI	0.284	0.490	0.232	0.488	0.045	0.499	0.011	0.499
CPL.BR-HS.CRI	0.301	0.483	0.264	0.470	0.068	0.498	0.031	0.499
CPL.BN-HS.CRI	0.208	0.500	0.162	0.497	0.094	0.490	0.060	0.490
CPL.BN-HS.IND	0.301	0.489	0.241	0.484	0.151	0.468	0.111	0.470

Table 11.3: Average fraction of correctly classified coefficients for the CRR models after variable selection with increasing number of covariates. Especially $\hat{\beta} \neq 0, \beta \neq 0$ denotes the case that the estimated effect is nonzero ($\hat{\beta} \neq 0$) when the corresponding true effect is nonzero ($\beta \neq 0$), and $\hat{\beta} = 0, \beta = 0$ denotes the case that the estimated effect is zero ($\hat{\beta} = 0$) when the corresponding true effect is zero ($\beta = 0$).

11.5. Adaption of the Bayesian NMIG selection criterion

Finally, some variations of the hard shrinkage selection (HS.IND) criterion defined in Section 4.4 are considered to improve the predictive performance of the resulting final models. Reconsidered are the simulation results with the low-dimensional predictor from Subsection 11.1 and those with the high-dimensional predictor from Subsection 11.4 based on the partial likelihood.

The hyperparameter values $v_0 = 0.000025, v_1 = 1, h_{1,\psi} = 5, h_{2,\psi} = 25$ in the hierarchical representation of the Bayesian NMIG prior were originally chosen to separate in sparse CRR models (ω not too large) effects in the range from 0.3 to 0.2, in the sense that "large" effects, with values larger than 0.3, are less regularized and "small" effects, with values smaller than 0.2, are strong regularized. As shown in the previous simulations, we achieve with the associative HS.IND threshold value 0.5 reasonable results - in terms of the MSE performance and the misclassification rates - when $n > p_x$ and small and large effects are clearly separable, as e. g. in simulation model CRR 1. But, we have also seen that for fixed sample size with increasing number of covariates the separation of "small" and "large" effects gets blurred, so that an adaption of the NMIG prior to the number p_x of covariates is indicated. Further, in cases with "small" or "moderate" effects, as considered in simulation models CRR 2 and CRR 3, the performance of Bayesian NMIG models also decreases under the basic hyperparameter constellation. In summary, there are a lot of situations, where we have to consider a modification of the basic hyperparameter setting in the Bayesian NMIG prior. Nevertheless, in the following we try several strategies to improve the MSE of the regression coefficients under the HS.IND selection criterion without changing the hyperparameters.

- At first we adapt (i. e. decrease), after visual inspection of the posterior inclusion probabilities, the threshold of the HS.IND selection rule to capture also smaller effects (HS.IND.1).
- At second we consider for covariate effects with an inclusion probability larger than 0.5 a modified estimate, defined as the empirical mean over the MCMC subsample, where the associated indicators equal v_1 (HS.IND.2).
- At last we combine both strategies HS.IND.1 and HS.IND.2 (HS.IND.3).

Li and Lin (2010) utilize in the context of the Bayesian elastic net prior the receiver operating characteristic (ROC) curve to adapt the α-level of the credible interval in the HS.CRI criterion. They improve the variable selection accuracy by plotting the correct inclusion rate (sensitivity) against the false inclusion rate (1-specificity) along the range of α in simulations and suggest using $\alpha = 0.5$ in practice, because a higher level of α results in a higher sensitivity but a lower specificity with the elastic net prior. Besides an adjustment of our HS.CRI region this ROC based method provides also another method to determine the HS.IND threshold, but we did not investigate this topic so far.

Results

Figure 11.30 and **Figure 11.31** show the impact of these modifications on the MSE of the estimated regression coefficients. **Figure 11.30** summarizes the results for models CRR 1 to CRR 3 from Subsection 11.1 and **Figure 11.31** those with the higher-dimensional predictor from Subsection 11.4.

As expected, none of the modified selection rules does improve the MSE performance in the simulation model CRR 1 with clearly separable large and small effects and estimated posterior inclusion probabilities close to 1 and 0, compare left panel of **Figure 11.30**. For the remaining simulations CRR 2 (middle panel) and CRR 3 (right panel), the largest improvements are achieved with the first modification HS.IND.1, i. e. by adapting the selection threshold to the lower value 0.1.

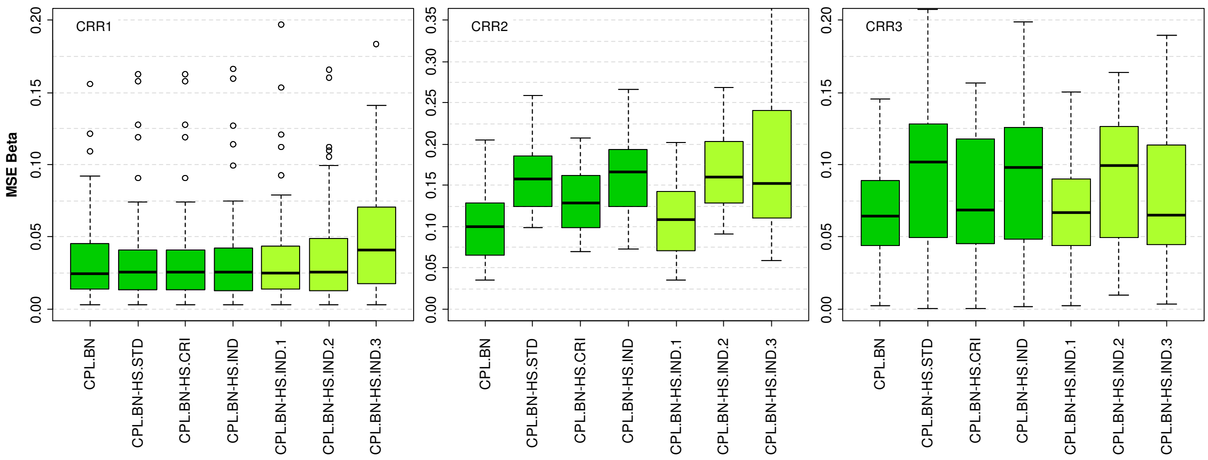


Figure 11.30: Mean squared errors of the regression coefficient estimates, $\mathrm{MSE}(\hat{\boldsymbol{\beta}})$, under the Bayesian NMIG prior and the associated variable selection methods in the simulation models CRR 1 to CRR 3. The additional boxes show the results under the modified HS.IND selection rule. HS.IND.1: Selection threshold 0.1. HS.IND.2: Selection threshold 0.5 and the values of the nonzero regression coefficient estimates are computed using the subsample where the indicator equals v_1. HS.IND.3: Combination of HS.IND.1 and HS.IND.2.

With decreasing value of the HS.IND-threshold the MSE of the resulting final model moves in direction of the MSE of the model CPL.BN, which includes all covariates in the predictor. In models CRR 1 and CRR2, where the effects are smaller and not clearly separated, we have seen that the application of the HS.IND criterion clearly decreases the MSE performance. In such situation it turns out that adapting the threshold value to the smaller observed posterior inclusion probabilities is a reasonable strategy to improve the predictive performance. With this line of action it is possible to get sparse final models with comparable good performance as the CPL.BN model. But, the improvement of the adaptation of the HS.IND-threshold is always limited by the MSE of the full CPL.BN model, and in particular in models CRR 2 and CRR 3 we obtain smaller MSE values with other regularization methods, like the ridge regularization. In the high-dimensional simulations, compare **Figure 11.31**, we obtain a similar result as for models CRR 2 and CRR 3. Due to the decreased estimated inclusion probabilities, the MSE of the CPLBN-HS.IND model is clearly increased in comparison to the CPL.BN model, and decreasing the threshold moves the MSE of the CPLBN-HS.IND in direction of the MSE of the full CPL.BN model.

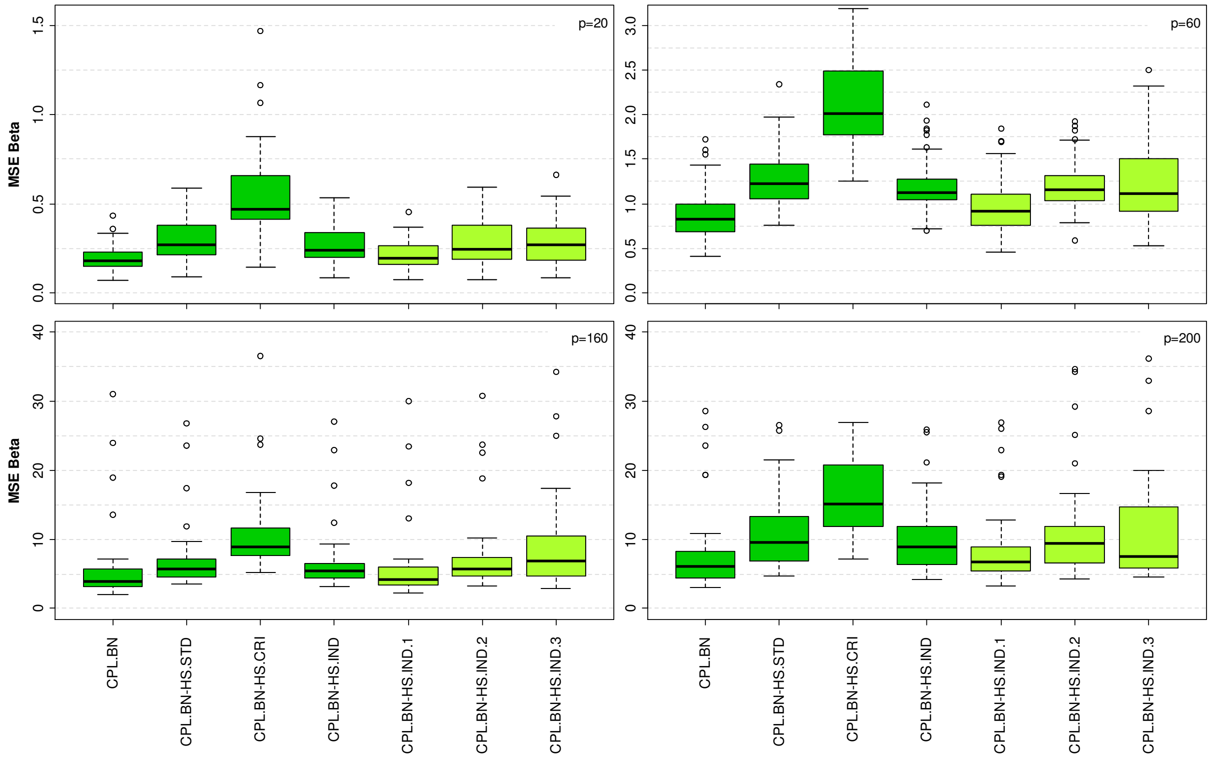


Figure 11.31: Mean squared errors of the regression coefficient estimates, MSE($\hat{\boldsymbol{\beta}}$), under the Bayesian NMIG prior and the associated variable selection methods in the CRR model with increasing number of covariates. The additional boxes show the results under the modified HS.IND selection rule. HS.IND.1: Selection threshold 0.2. HS.IND.2: Selection threshold 0.5 and the values of the nonzero regression coefficient estimates are computed using the subsample where the indicator equals v_1. HS.IND.3: Combination of HS.IND.1 and HS.IND.2.

The absolute values of the coefficient estimates, constructed under the HS.IND2 and HS.IND3 modification, are in general larger, since the samples with associated value $I_j = v_0$ are ignored. Both modifications do not improve the MSE, and the MSE resulting from the HS.IND2 criterion is clearly increased in almost all models (not in CRR 1). So, we note that also the shrinkage of the "larger" effects improves the predictive performance, in particular in the higher-dimensional cases. We refer again to the application in Section 14 which shows similar results from the practical perspective.

Final remarks

In summary, the Bayesian NMIG prior performs best in sparse models, where covariates have mainly "small" and "large" effects as in model type CRR 1. In the higher dimensions also the reduced shrinkage of "larger" effects, if present, causes an improvement of the predictive performance. In models with "moderate" or "smaller" effect sizes, like model types CRR 2 and CRR 3, the Bayesian ridge or lasso prior achieve the best performance results. We have seen that in models with various effect sizes the posterior inclusion probabilities for the covariates, as provided by the NMIG prior, reflect very well the importance of the covariates. But, as previously observed with the AFT simulations, variable selection guided by the induced ranking of the covariates shows in general no improvement of the predictive performance, even in models of CRR 1 type. We have also seen that variable selection may improve the predictive performance, but often full models yield comparable or higher performances as sparse final models.

PART IV. APPLICATIONS

To illustrate the presented methods in applications, we analyze three survival datasets and compare the results with those from available frequentist alternatives. We use the Bayesian methods to fit the extended AFT and CRR models to the data. Inference for the CRR model is based on the partial likelihood and the full likelihood using a P-spline model for the baseline hazard function. In the AFT model the error distribution is specified as penalized Gaussian mixture (PGM) or assumed to be Gaussian. For both survival model types we estimate on the one hand models that assume a strictly linear impact of the available covariates on the patient's survival time, utilizing the Bayesian lasso, ridge and NMIG prior to shrink the effects towards zero. Additionally the hard shrinkage selection rules of Section 4.4 are applied to the Bayesian approaches, in order to identify sparse final models containing only the covariates with the strongest influence on the patient's survival. On the other hand, to take into account possibly nonlinear shapes of some effects, continuous covariates are modeled by P-splines, each equipped with a random walk smoothing prior, in combination with the Bayesian shrinkage priors for the remaining linear effects. Further, this extended setting of the predictor allows an investigation of the variable selection stability under increasing model complexity.

As in the previous simulation sections Bayesian inference for the extended AFT model is carried out with the `R`-function `baftpgm()`. Correspondingly, we utilize the `R`-function `bcoxpl()` for the extended CRR model, if inference is based on the partial likelihood, and the `BayesX` internal `regress` method, if inference is based on the full likelihood. The various Bayesian results are compared to the results from frequentist methods. We use the `coxph()` function to fit the frequentist CRR model based on the partial likelihood and the `survreg()` function to fit an AFT model with Gaussian error. Nonlinear covariate effects are modeled with the `pspline()` term within the formula specification. Both functions are combined with the `step()` function for variable selection based on the AIC criterion in a stepwise-backward procedure. The frequentist lasso and ridge regression in the CRR model with strictly linear predictor is carried out with the function `penalized()`, Goeman (2010). Cumulative baseline hazards associated with the partial likelihood estimates are computed via the Breslow estimator and the cumulative baseline hazards associated with the P-spline estimates of the baseline hazard are computed with the trapezoidal rule.

12. Primary biliary cirrhosis of the liver

12.1. Data

The presented methods for the extended AFT and CRR model are applied to the primary biliary cirrhosis data, provided for example in the `R`-package `{survival}` or the book-homepage of Therneau and Grambsch (2000). Primary biliary cirrhosis (PBC) is an autoimmune disease of the liver, marked by the slow progressive destruction of the small bile ducts (bile canaliculi) within the liver. When these ducts are damaged, bile builds up in the liver and damages over time the tissue. This can lead to scarring, fibrosis and cirrhosis inducing a liver failure and finally to the death of the patient.

In the following we give a short description of the data and refer to Therneau and Grambsch (2000) for a more detailed presentation and an extended frequentist analysis based on the partial likelihood. In the CRR regularization context the PBC data is also used in Tibshirani (1997), who compared the variable selection property of the lasso penalty with a backward-forward stepwise procedure based on p-values. Further, Zhang and Lu (2007) applied the adaptive lasso on this data, where, in contrast to the lasso penalty, covariate-specific weights in the penalization term enable the coefficient-specific shrinkage. In the context of regression spline models, e. g. Sleeper and Harrington (1990) used this data for a sophisticated analysis, where some covariate effects are assumed to have a nonlinear shape modeled by B-spline basis functions. Finally, the data is also analyzed in the context of regularized semiparametric AFT models, in particular Johnson (2008) and Johnson (2009) apply the lasso, the adaptive lasso and the elastic net penalization to ten preselected covariates.

time	number of days between registration and the earlier event of death or transplantation
status	status at endpoint, 0 = censored, 1= transplant or dead
age	age in years
alb	albumin in gm/dl
alkphos	alkaline phosphatase in U/liter
ascites	presence of ascites (0 = no, 1 = yes)
bili	serum bilirubin in mg/dl
chol	serum cholesterol in mg/dl
copper	urine copper in ug/day
edtrt	presence of edema (0.0 = no edema and no diuretic therapy for edema; 0.5 = edema present without diuretics, or edema resolved by diuretics; 1.0 = edema despite diuretic therapy)
hepmeg	presence of hepatomegaly, i. e. enlarged liver (0 = no, 1 = yes)
platelet	platelets per cubic ml / 1000
protime	standardized blood clotting time, prothrombin time in seconds
sex	sex (0 = male, 1 = female)
sgot	liver enzyme SGOT in U/ml
spiders	blood vessel malformations in the skin, presence of spiders (0 = no, 1 = yes)
stage	histologic stage of disease
trig	triglicerides in mg/dl
trt	treatment/drug (1= D-penicillamine, 2 = placebo)

Table 12.1: List of available covariates used in the analysis of the PBC data.

The data has been collected from the Mayo Clinic trial in primary biliary cirrhosis of the liver conducted between 1974 and 1984. A total of 424 PBC patients, referred to the Mayo Clinic during that ten-year interval, met eligibility criteria for the randomized placebo controlled trial of the drug D-penicillamine. The first 312 cases in the data set participated in the randomized trial and contain largely complete data. The additional 112 cases did not participate in the clinical trial, but consented to have basic measurements recorded and to be followed for survival. Six of those cases were lost to follow-up shortly after the diagnosis, so the data considered here consists of the additional 106 cases as well as the 312 randomized participants. Discarding observations with missing values leaves finally

$n = 276$ observations with 58.42 % censoring and a mean survival time of 1889 days, ranging from 41 to 4191 days. The covariates used for analysis are listed in **Table 12.1**. To make our results comparable to those in Tibshirani (1997), covariates were standardized to have zero mean and unit variance. We have not applied log-transformations for covariates that have somehow skewed distributions.

12.2. Analysis

We consider two structures for the predictor. The first is assumed to be strictly linear

$$\begin{aligned}\eta_i = \; & \gamma_0 + \beta_1 \text{acites}_i + \beta_2 \text{edtrt}_i + \beta_3 \text{hepmeg}_i + \beta_4 \text{sex}_i + \beta_5 \text{spiders}_i + \beta_6 \text{stage}_i \\ & + \beta_7 \text{trt}_i + \beta_8 \text{age}_i + \beta_9 \text{alb}_i + \beta_{10} \text{alkphos}_i + \beta_{11} \text{bili}_i + \beta_{12} \text{chol}_i + \beta_{13} \text{platelet}_i \\ & + \beta_{14} \text{protime}_i + \beta_{15} \text{sgot}_i + \beta_{16} \text{trig}_i + \beta_{17} \text{copper}_i ,\end{aligned} \tag{12.1}$$

and in the second version, the continuous covariates are modeled nonlinear

$$\begin{aligned}\eta_i = \; & \gamma_0 + \beta_1 \text{acites}_i + \beta_2 \text{edtrt}_i + \beta_3 \text{hepmeg}_i + \beta_4 \text{sex}_i + \beta_5 \text{spiders}_i + \beta_6 \text{stage}_i \\ & + \beta_7 \text{trt}_i + f_1(\text{age}_i) + f_2(\text{alb}_i) + f_2(\text{alkphos}_i) + f_4(\text{bili}_i) + f_5(\text{chol}_i) \\ & + f_6(\text{platelet}_i) + f_7(\text{protime}_i) + f_8(\text{sgot}_i) + f_9(\text{trig}_i) + f_{10}(\text{copper}_i),\end{aligned} \tag{12.2}$$

where $f_j(\cdot)$, $j = 1,...,10$, are smooth functions of the 10 continuous covariates *age*, *alb*, *alkphos*, *bili*, *chol*, *platelet*, *protime*, *sgot*, *trig* and *copper*, which are modeled by cubic P-splines. In the AFT model

$$y_i = \eta_i + \sigma\varepsilon_i , \tag{12.3}$$

when the error density is specified as penalized Gaussian mixture, i. e. $\varepsilon_i \sim \sum_{j=1}^{g_0} w_j N(m_j, s_j^2)$, we use the same specification of the error as described in the simulation setting of Section 10.1. In summary, $g_0 = 21$ basis functions with equidistant knots m_j, placed in the interval $[-4.5, 4.5]$, and uniform variances $s_j^2 = 0.25^2$ are used to model the error density. A random walk prior with difference order $d_0 = 3$ controls the smoothness of the PGM. The hyperparameters of the scale parameter σ^2 are set to $h_{1,\sigma} = h_{2,\sigma} = 0.01$ and those of the smoothing variance to $h_{1,\tau_0} = 1$, $h_{2,\tau_0} = 0.01$. We utilize the *"sliceR0"*, *"mcondstep"* and *"mcondblock"* update schemes for the transformed mixture weights. Due to negligible differences, the presented results are based on the update scheme *"sliceR0"* as representative. As further competitor the error is assumed to be purely Gaussian.

The hyperparameters of the Bayesian lasso and ridge (version B) gamma prior for the shrinkage parameter λ are set, as in Section 11.1, to weakly informative values $h_{1,\lambda} = h_{2,\lambda} = 0.01$. The hyperparameters of the Bayesian NMIG variance parameter components $\tau_j^2 = I_j \psi_j^2$ are $v_1 = 1$, $v_0 = 0.000025$, $h_{1,\psi} = 5$ and $h_{2,\psi} = 25$ in combination with $h_{1,\omega} = 1$ and $h_{2,\omega} = 1$ to define a uniform prior for the complexity parameter ω. For the representation of the nonlinear effects $f_j(\cdot) = \sum_{k=1}^{g_j} \alpha_{k,j} B_k(\cdot)$ in predictor (12.2) we use $g_j = 20$ cubic B-spline basis functions $B_k(\cdot)$ in combination with second-order random walk priors, $d_j = 2$, for the associated basis function weights $\boldsymbol{\alpha}_j = (\alpha_{1,j}, ..., \alpha_{g_j,j})'$. The hyperparameters of the inverse gamma prior for the smoothness controlling variances $\tau_{\alpha_j}^2$ are set to $h_{1,\tau_j} = h_{2,\tau_j} = 0.001$.

The MCMC algorithms for inference in the AFT model ran with 30000 iterations, where 15000 iterations discarded as burnin and a thinning of 15. The running time is about 80 minutes with strictly linear predictor and 120 minutes with the nonlinear predictor on a system with quad-core CPU (Intel Quad9550, 2.83 GHz).

To fit the CRR models

$$\lambda_i(t_i) = \lambda_0(t_i)\exp(\eta_i), \tag{12.4}$$

we use with the Bayesian methods 20000 iterations with a burnin of 5000 and thin the chain by 10 which results in an MCMC sample of size 1500 (Running times: CFL: 2 -3 minutes, CPL linear predictor: 20 minutes, CPL spline predictor: 6 hours!). Since there exists no functional connection between the estimates resulting from the CRR and the AFT model, with exception of the Weibull model, we use as default a common specification of the regularization priors in both survival model classes. Within the `BayesX` method `regress` we use the default values $h_{1,\lambda} = h_{2,\lambda} = 0.001$ to specify the hyperparameters of the Bayesian lasso and ridge prior. The logarithm of the baseline hazard $f_0(\cdot) := \log\lambda_0(\cdot)$ is modeled, as the nonlinear effects $f_j(\cdot)$, by cubic P-splines placed at $g_j = 20$ knots and the basis function weights are equipped with second-order random walk priors to control the smoothness. The associated hyperparameters of the smoothing variances are also set to the default values $h_{1,\tau_j} = h_{2,\tau_j} = 0.001$, $j = 0,1,...,10$. Due to inferential problems arising with the estimation procedure `regress` under the Bayesian NMIG prior in combination with predictor (12.2), the covariate *age* is modeled linearly in this specific case.

To model the nonlinear covariate effects with the frequentist procedures, we use the `pspline()` term within the formula of the `R`-functions `survreg()` and `coxph()`. The roughness penalty of the P-splines is set to the value `theta=0.8`. In the subsequent sections the main results of the analysis are presented. The abbreviations that denote the models are listed in the Reference Section.

12.3. Results

Results for the linear predictor

Linear effects

We first summarize the results obtained under the purely linear structure of the predictor (12.1). A selection of the point estimates in the CRR model together with the corresponding standard deviations for the regression coefficients are displayed in the upper plot of **Figure 12.1**. The lower plot shows results from the Bayesian methods after applying the hard shrinkage rules HS.STD and HS.IND to select covariates for the final model together with the results from the stepwise procedure and the frequentist lasso. The marked standard deviations for the frequentist lasso are obtained by the approximate method described in Tibshirani (1997), and for the regression coefficients not included in the final model the standard deviations are set to zero.

All presented methods are leading to final models that include the five covariates *age*, *alb*, *bili*, *stage* and *copper* and eliminate *hepmeg*, *platelet*, *spiders*, *trt* and *trig*. The covariate *ascites* is only chosen by the frequentist lasso (CPL.PenL), but the effect of *ascites* is generally very small. Obviously, as designed, the Bayesian NMIG method (CPL.BN, CFL.BN) shrinks small effects to a larger extent than the lasso- or ridge-based methods, so that most of the remaining covariates are excluded, if variable selection is based on the HS.IND criterion. The HS.IND criterion uses the model inclusion probability of each covariate estimated by posterior relative frequency of the NMIG indicator variables $I_j = v_1$, compare **Figure 12.3**. In contrast to the results with the full likelihood, the inclusion probability of the covariate *edrt* under the partial likelihood falls just below the HS.IND selection threshold of 0.5 and does not appear in the resulting final model.

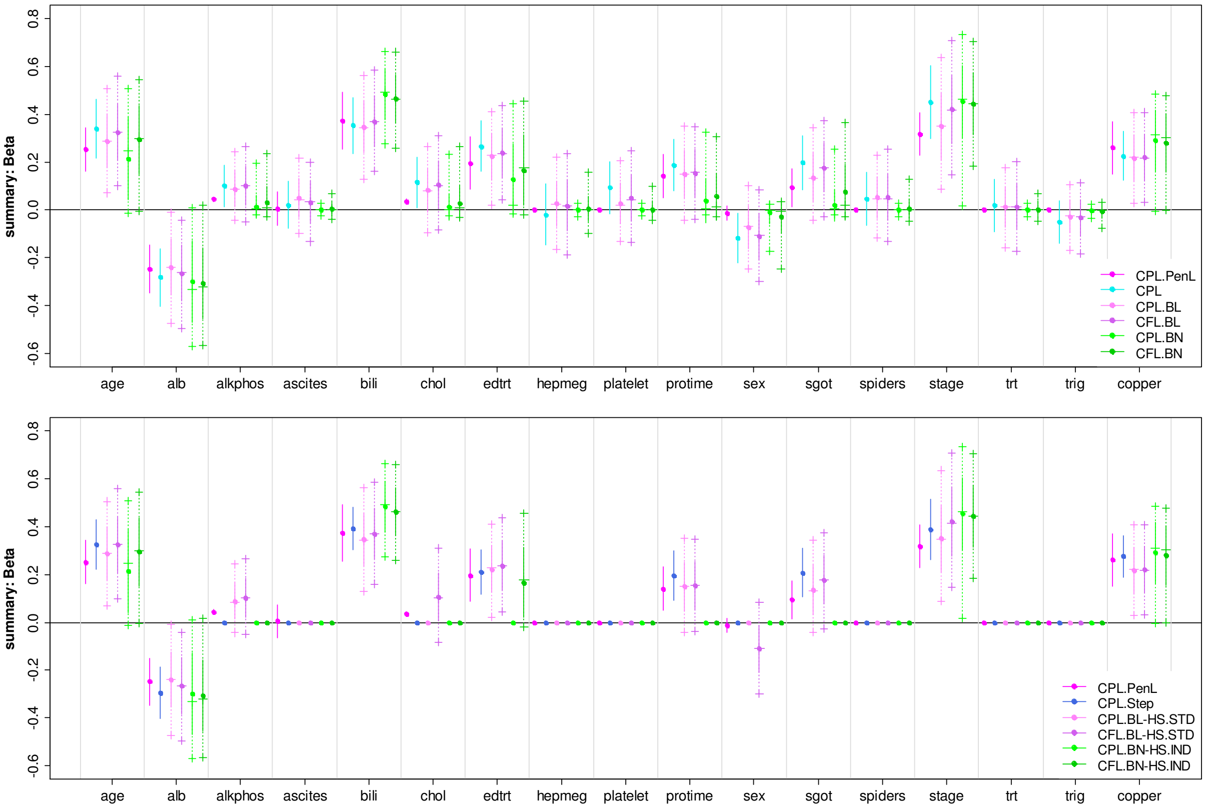


Figure 12.1: Estimated regression coefficients without (upper panel) and with variable selection (lower panel) in the CRR model. The points mark the estimates of the regression coefficients and the solid lines display the corresponding standard errors. For the Bayesian procedures the points mark the mean, the solid lines display the standard errors and the additional dashes mark the median and the 95 % empirical quantiles of the marginal posterior distribution of the regression coefficients.

Figure 12.2 shows the corresponding results from the AFT model with Gaussian and PGM error. The unpenalized estimates show by trend smaller absolute values than those under the CRR assumption with exception of the covariates *ascites* and *spiders*. If we consider the estimates under the NMIG prior, we see, e. g., that the effects of the covariates *bili* and *edtrt* are stronger regularized as the effects of the covariates *protime* and *sgot*, while the unregularized estimates of these covariates are almost of comparable size. In the CRR models we observe a similar behavior for the covariates *age* and *bili* or *edtrt* and *copper*, where respectively also the covariates *age* and *edtrt* are stronger regularized. We note again that the “effective” regularization of the linear effects depends also on the used survival regression model and the modeling of the components within the specific survival model.

Nevertheless, we find for eleven covariates in both survival model classes comparable results with respect to the covariates included in the final models after variable selection. Some differences occur for the three covariates *alkphos* (CPL.Step), *chol* (CFL.BL-HS.STD) and *edtrt* (CPL.BN-HS.IND) and for the remaining three covariates *alb*, *ascites* and *spiders* we observe differences more frequently.

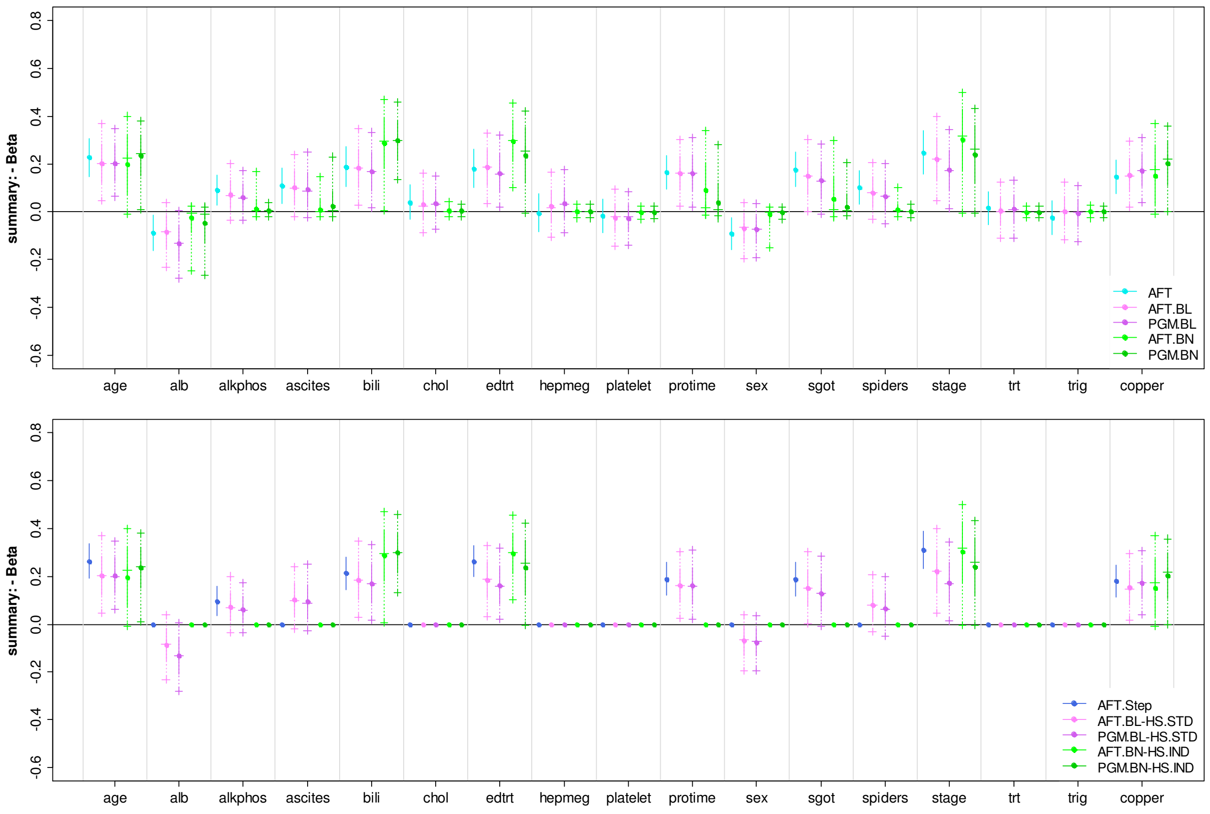


Figure 12.2: Estimated regression coefficients without (upper panel) and with variable selection (lower panel) in the AFT model. The points mark the estimates of the regression coefficients and the lines display the corresponding standard errors. For the Bayesian procedures the points mark the mean, the solid lines display the standard errors and the additional dashes mark the median and 95 % empirical quantiles of the marginal posterior distribution of the regression coefficients.

NMIG indicators

The posterior relative frequencies of the Bayesian NMIG indicator variables $I_j = v_1$ are summarized in **Figure 12.3,** for the CRR model under the full and partial likelihood, and in **Figure 12.4,** for the AFT model with Gaussian and PGM error. The crosses at the bottom of the bars mark the covariates from

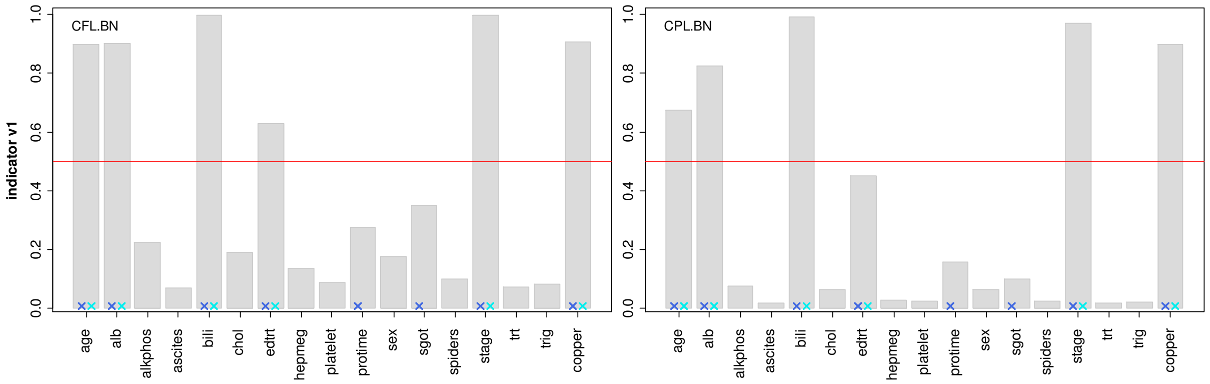


Figure 12.3: Estimated inclusion probabilities based on posterior relative frequencies of the Bayesian NMIG indicator variable value $I_j = v_1$ in the CRR model based on the full likelihood (left side) and the partial likelihood (right side). The crosses at the bottom of the bars mark the covariates from the corresponding frequentist models, which are significant with respect to the p-value 0.05 (cyan) and which are selected by the frequentist stepwise variable selection procedure (dark blue).

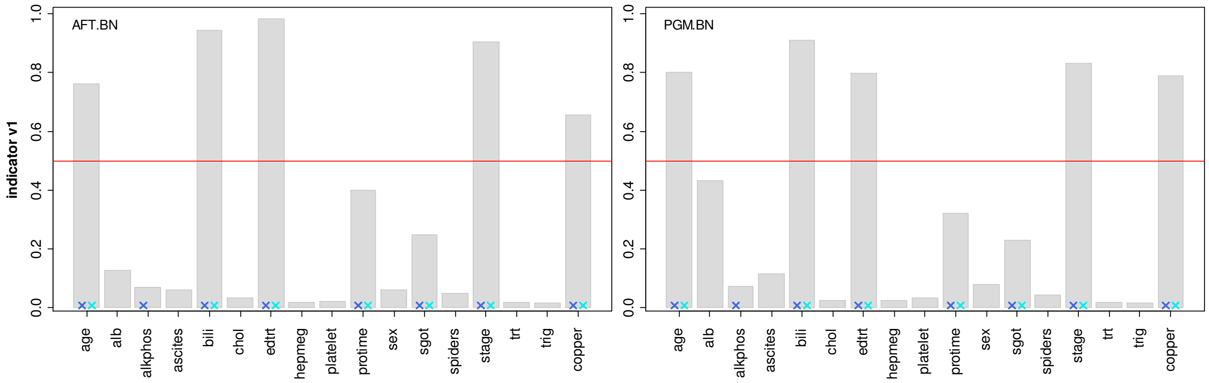


Figure 12.4: Estimated inclusion probabilities based on posterior relative frequencies of the Bayesian NMIG indicator variable value $I_j = v_1$ in the AFT model with Gaussian error distribution (left side) and PGM error distribution (right side). The crosses at the bottom of the bars mark the covariates from the corresponding frequentist models, which are significant with respect to the p-value 0.05 (cyan) and which are selected by the frequentist stepwise variable selection procedure (dark blue).

the corresponding frequentist models, which are significant with respect to the p-value 0.05 (cyan) and which are selected by the frequentist stepwise variable selection procedure (dark blue). With exception of the covariates *alb* in the AFT models and *edtrt* in the CPL model, almost all covariates are across the various models commonly selected resp. deselected by the HS.IND criterion. Nevertheless, we see that the impact of the covariates is more or less pronounced across the survival model classes, but also varies within the survival model class with the model complexity.

Baseline quantities

The estimated log-baseline hazards, $\log\lambda_0(t)+\gamma_0$, obtained from the full likelihood approach in the CRR model with the lasso and NMIG prior, are depicted at the left side of **Figure 12.5**. The corresponding cumulative baseline hazards, obtained by applying the trapezoidal rule for integration, are shown at the right side of **Figure 12.5** together with the Breslow estimate from the partial likelihood based methods. We observe a close conformity of the estimates across the frequentist and

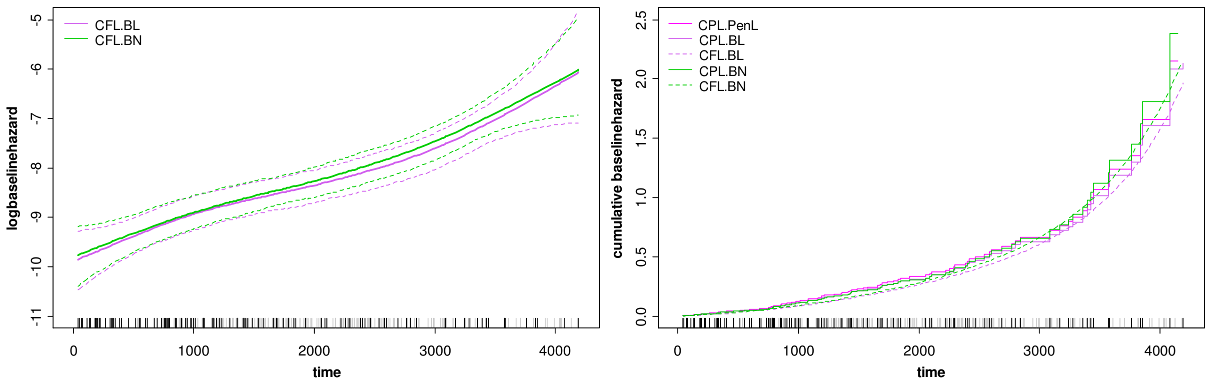


Figure 12.5: Estimation of the log-baseline hazard and cumulative baseline hazard in the CRR model. Left side: Posterior mean estimate of the log-baseline hazard for the Bayesian NMIG and Bayesian lasso regularization based on the full likelihood (solid lines) with 95% pointwise credible bands (dashed lines). Right side: Cumulative baseline hazards obtained as Breslow estimate for the partial likelihood methods and via the trapezoidal rule for the full likelihood methods based on the posterior mean of the involved regression coefficients.

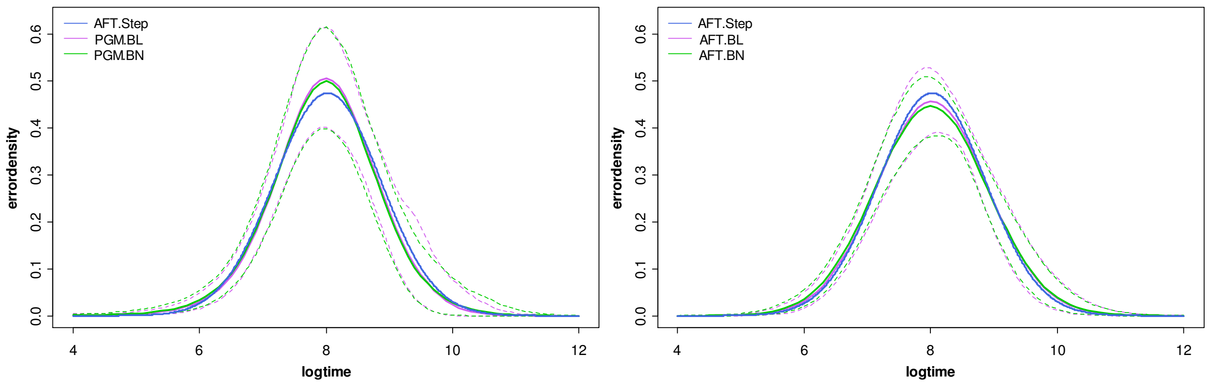


Figure 12.6: Estimation of the baseline error distribution in the AFT model. Left side: Posterior mean estimate of the density (solid lines) when the error is modeled as PGM with 95% pointwise credible bands (dashed lines). Right side: Posterior mean estimate of the density (solid lines) when the error is modeled by a Gaussian distribution with 95 % pointwise credible bands (dashed lines).

Bayesian approaches and across the different types of regularization priors. The same holds for the estimated baseline error densities ($Y_0 = \gamma_0 + \sigma\varepsilon_0$) in the AFT model, as shown in **Figure 12.6**, when a PGM error (left side) or Gaussian error (right side) is assumed. The baseline hazard estimation indicates that the risk to die increases monotonically, closely linear, over the years. In the AFT model there is no eye catching asymmetry for the density estimation observable and the Gaussian error seems to be a good proxy to the baseline error distribution.

Path results for the linear predictor

We compute the paths of the parameter estimates as function of the shrinkage parameter to investigate the evolution of the estimates if the shrinkage parameter is varied. The frequentist lasso estimation procedures provide the regression coefficient paths as add-on to the implemented functions. In the Bayesian procedures the parameter paths are obtained by suppressing the update of the shrinkage parameter in the MCMC process and choosing the starting value at the desired grid points in the range of the penalty-specific shrinkage parameter. The frequentist and Bayesian lasso paths are plotted as a function of the (standardized) complexity parameter t, with $t := \sum|\hat{\beta}_j| / \sum|\hat{\beta}_{j,ML}| \in [0,1]$, where $\hat{\beta}_{j,ML}$ denote the unconstrained maximum likelihood estimates and $\hat{\beta}_j$ correspond to the regularized lasso estimates with $\sum|\hat{\beta}_j| \in [0, \sum|\hat{\beta}_{j,ML}|]$. For the Bayesian NMIG prior we show the results on grid points in the range $\omega \in [0.1,1]$, because for grid points $\omega < 0.1$ we observe very wiggly paths associated to the covariates with larger effects, in particular for the indicator variables. As outlined in Section 4.3.2 larger regression coefficients can have high sampling probabilities for the sate $I_j = v_1$, even if ω is small, so the paths become unstable until $\omega = 0$. This is also confirmed by the paths of the variance parameters (not shown), where the variance parameters of the larger regression coefficients obtain large values nearly over the whole range of the complexity parameter.

Paths of the regression coefficients

The paths of four selected regression coefficients from the Bayesian lasso regularized CRR and AFT models are plotted in **Figure 12.7**. In addition, the associated paths of the pointwise empirical standard deviation (dashed lines) and the pointwise empirical 95% credible region (dotted lines) are shown, which are used to practice variable selection in terms of the HS.STD and HS.CRI criterion. The asterisks on the coefficient paths indicate the grid point t at which the estimates are computed and the black dotted vertical line marks the estimated constraint parameter $\hat{t}$ (from the full Bayesian approach). From the top to the bottom of the figure the paths are computed with the CRR model based on the full and partial likelihood and the AFT model with PGM and Gaussian error. **Figure 12.8** shows the corresponding results under the Bayesian NMIG prior. In both figures the coefficient paths of the frequentist lasso (CPL.PenL) are marked as competitor (black dashed lines).

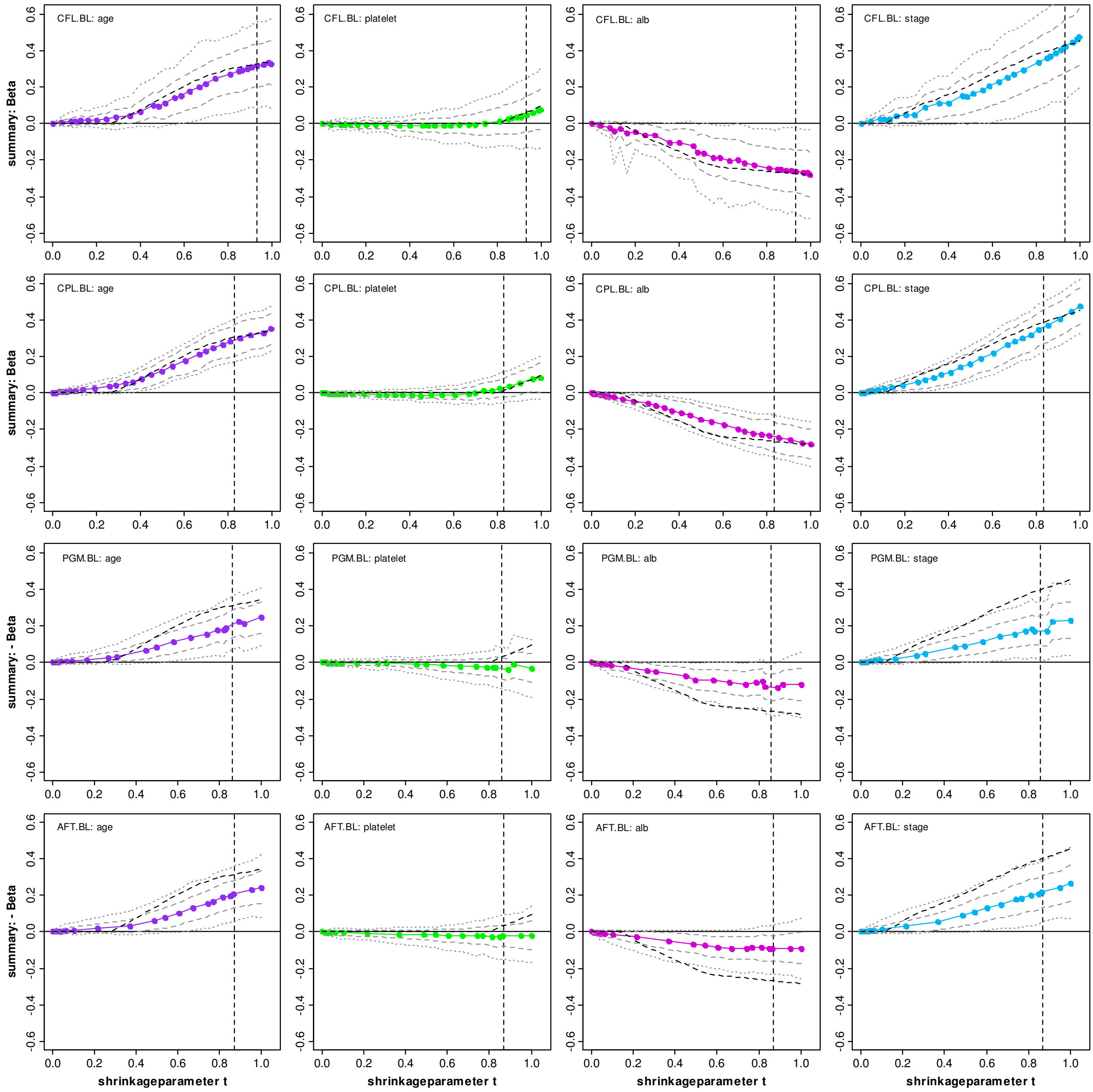


Figure 12.7: Selected coefficient estimates from the Bayesian lasso regularization in the CRR (first two rows) and AFT model (last two rows) as a function of the complexity parameter t. In the first row the estimates are based on the full likelihood and in the second row on the partial likelihood. In the third row the error is modeled by a PGM and in the last row the error is Gaussian. The vertical dashed line marks the corresponding coefficient estimates at the particular (full) Bayesian estimate of the constraint parameter. The gray dotted and dashed lines mark the evolution of the empirical 95% quantiles and standard deviation of the associated marginal posterior distribution. The black dashed paths mark coefficient paths of the penalized lasso procedure.

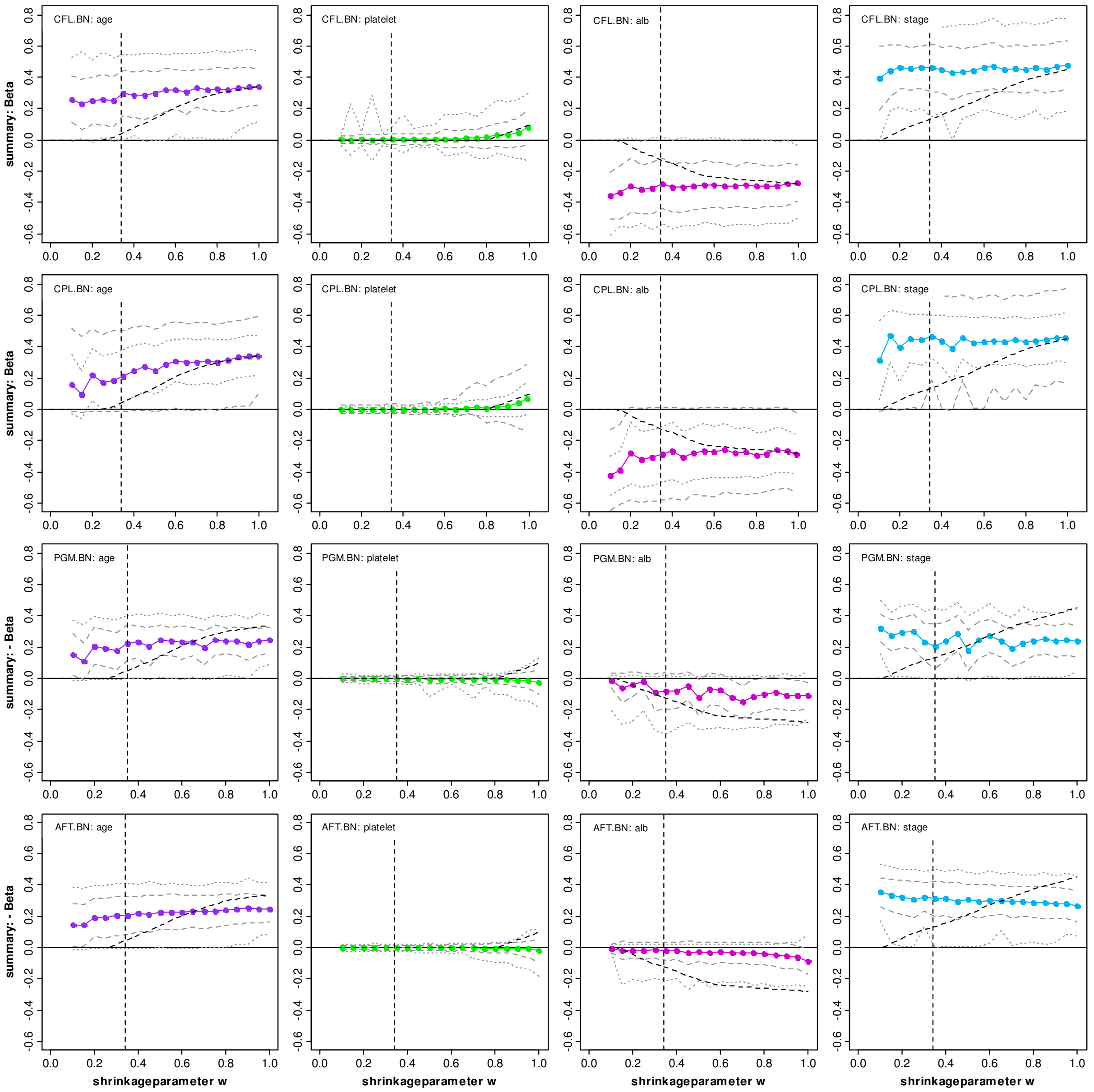


Figure 12.8: Selected coefficient estimates from the Bayesian NMIG regularization in the CRR (first two rows) and AFT model (last two rows) as a function of the complexity parameter t. In the first row the estimates are based on the full likelihood and in the second row on the partial likelihood. In the third row the error is modeled by a PGM and in the last row the error is Gaussian. The vertical dotted line marks the corresponding coefficient estimates at the particular (full) Bayesian estimate of the constraint parameter. The gray dotted and dashed lines mark the evolution of the empirical 95% quantiles and standard deviation of the associated marginal posterior distribution. The black dashed paths mark coefficient paths of the penalized lasso procedure.

In the CRR model the Bayesian and frequentist lasso paths show no strong differences. Due to the sampling based MCMC inference, the Bayesian paths are not piecewise exactly equal to zero as the frequentist lasso paths, where inference and variable selection is carried out simultaneously. For larger regression coefficients the paths of the standard deviation or the credible interval, computed with the partial likelihood (CPL.BL), often cross zero in the region where the frequentist lasso path is set to zero. We remember that e. g. the covariate *platelet* is always excluded from all final models. If we consider the development of the estimator for the covariate *platelet*, there is in all models strong evidence that this covariate has a negligible effect, because the margins of the HS.STD interval always include zero and the size of the effect marginally varies.

Further, the figures highlight the different shrinkage properties of the Bayesian lasso and NMIG prior. The shrinkage of larger regression coefficients, like those of *age* or *stage*, is suppressed by the bimodal structure of the NMIG prior that leaves these larger effects over a wide range of the complexity parameter virtually unpenalized. In comparison to the Bayesian NMIG paths the Bayesian lasso paths indicate a more uniform shrinkage of small and large effects.

Paths of the NMIG posterior inclusion probabilities

The left panel of **Figure 12.9** shows the posterior NMIG inclusion probabilities of the four selected covariates *age*, *alb*, *platelet* and *stage* as a function of the complexity parameter ω. In the right panel we see as competitor the estimated inclusion probabilities from the full Bayesian models, where the shrinkage parameter is jointly estimated, compare **Figure 12.3** and **Figure 12.4**. The vertical lines in the left panel mark the associated estimated value $\hat{\omega}$ from the MCMC sample.

For some of the covariates we observe a similar evolution along increasing values of ω as for the estimated regression coefficients. The inclusion probabilities of the covariates *age* and *stage* increase rapidly to relative high values exceeding the HS.IND selection threshold of 0.5. In contrast, the inclusion probability of the covariate *platelet* only slightly changes over a wide range of the complexity parameter and clearly increases only in the last third of the complexity parameter range. Other covariates, like *alb*, show a different devolution in the CRR or AFT survival model class. In the CRR model the inclusion probabilities of *alb*, obtained with the full and partial likelihood, quickly increase, but in the AFT model with PGM and Gaussian error they stay longer on a lower level below the threshold of 0.5. The inclusion probability values from the paths at the estimate $\hat{\omega}$ are almost comparable with the inclusion probabilities shown in the right panel. Finally, we observe that the paths across the models and methods differ in the unsteadiness, where the wiggliest results are obtained under the AFT model with PGM error, indicating a higher uncertainty in the classification to the component $I_j = v_0$ and $I_j = v_1$.

Paths of the baseline quantities

The following three figures show the impact on the baseline quantities if the shrinkage parameters are varied. We display the impact under the Bayesian lasso and NMIG prior in the CRR model with P-spline baseline hazard, **Figure 12.10**, the AFT model with PGM error, **Figure 12.11**, or Gaussian error, **Figure 12.12**.

In contrast to the Bayesian lasso models (left sides of the figures), there are no rigorous changes in the log-baseline hazard or baseline error density estimates observable under the Bayesian NMIG models (right sides of the figures), if the complexity parameter is varied. Under the lasso prior the log-baseline hazard estimates have a close to linear increase for large values of the complexity parameter t that indicate a weak regularization. With decreasing complexity parameter $t \to 0$, i. e. enhanced regularization, we obtain smaller slopes of the log-baseline hazard in the first 2000 days, and the slope in this interval gets close to zero, if $t \approx 0$, see **Figure 12.10**.

With decreasing complexity parameter $t \to 0$ we observe also the progress of the baseline error density from a symmetric to a heavy skewed shape in the AFT model with PGM error, compare **Figure 12.11**, and with the Gaussian error the density gets more mass in the tails, compare **Figure 12.12**.

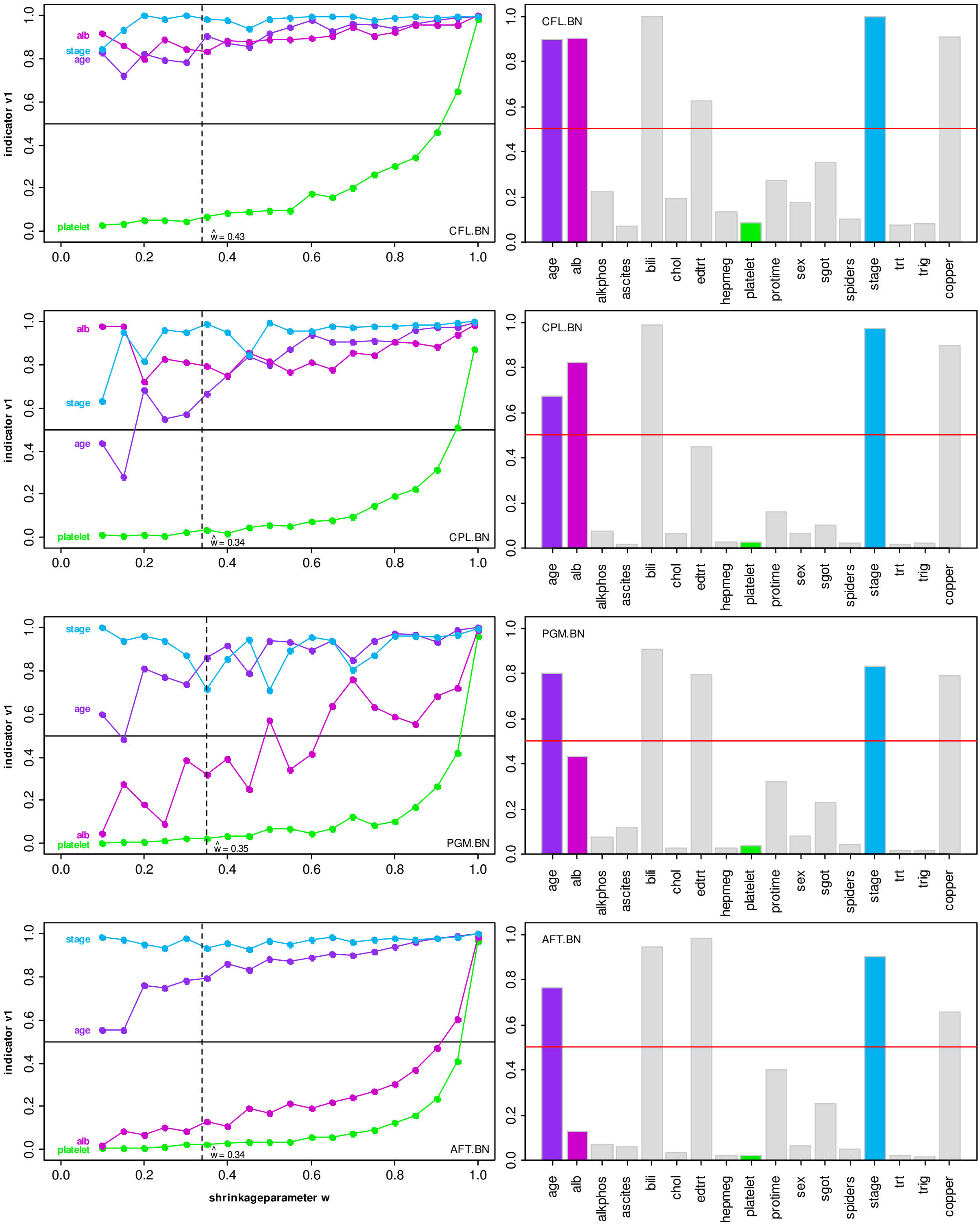


Figure 12.9: Selected posterior relative frequencies of the Bayesian NMIG indicator variables as a function of the complexity parameter ω (left column) and relative posterior frequencies of the Bayesian NMIG indicator variable (right column) under the full Bayesian approach in the CRR and AFT model. The vertical dashed line in the figures of the first column marks the corresponding coefficient estimates at the (full) Bayesian estimate of the complexity parameter. In the first row the estimates are based on the full likelihood and in the second row on the partial likelihood. In the third row the error is modeled by a PGM and in the last row it is Gaussian.

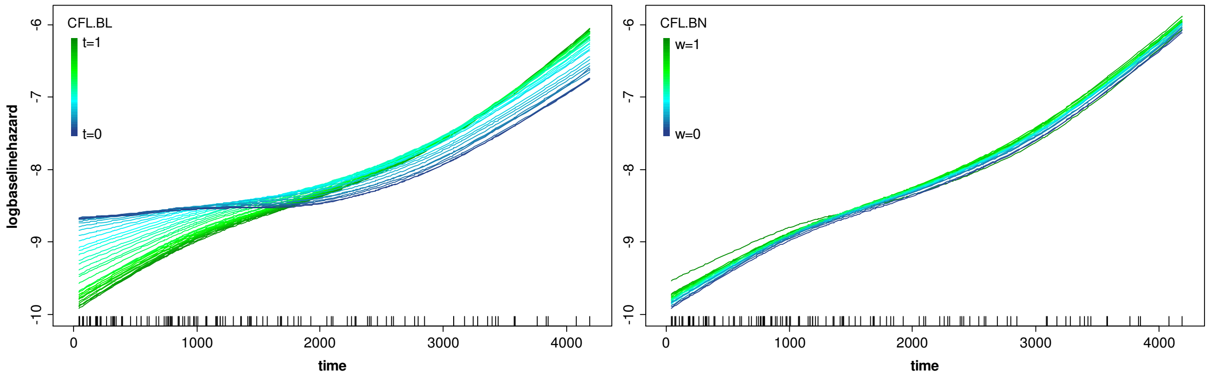


Figure 12.10: Log-baseline hazard estimation for the different values of the shrinkage parameters in the CRR model based on the full likelihood. Posterior mean estimations resulting from the paths of the Bayesian lasso (right side) and the Bayesian NMIG (left side) regularization.

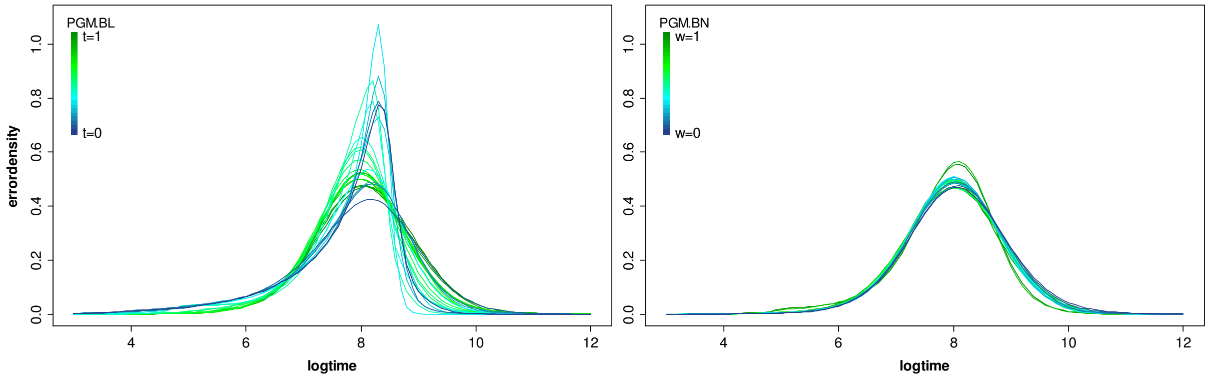


Figure 12.11: Baseline error density estimation for the different values of the shrinkage parameters in the AFT model with Gaussian error. Posterior mean estimations resulting from the paths of the Bayesian lasso (right side) and the Bayesian NMIG (left side) regularization.

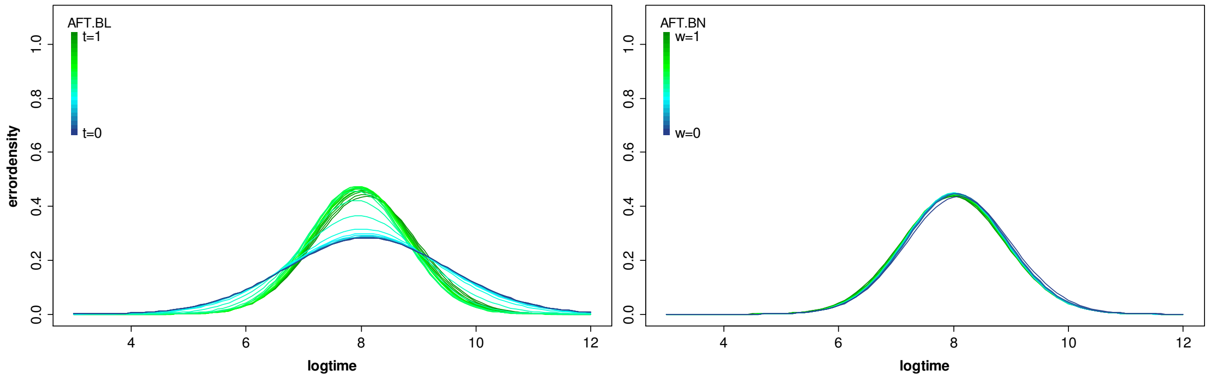


Figure 12.12: Baseline error density estimation for the different values of the shrinkage parameters in the AFT model with PGM error. Posterior mean estimations resulting from the paths of the Bayesian lasso (right side) and the Bayesian NMIG (left side) regularization.

Selected samples for the linear predictor

To highlight again the different regularization structures of the Bayesian lasso and NMIG priors, we consider the generated MCMC samples. The following four figures show the samples of four different covariate effects under the Bayesian lasso and NMIG prior for different fixed values of the complexity parameter t resp. ω. We use the results from the CRR model with P-spline model for the log-baseline

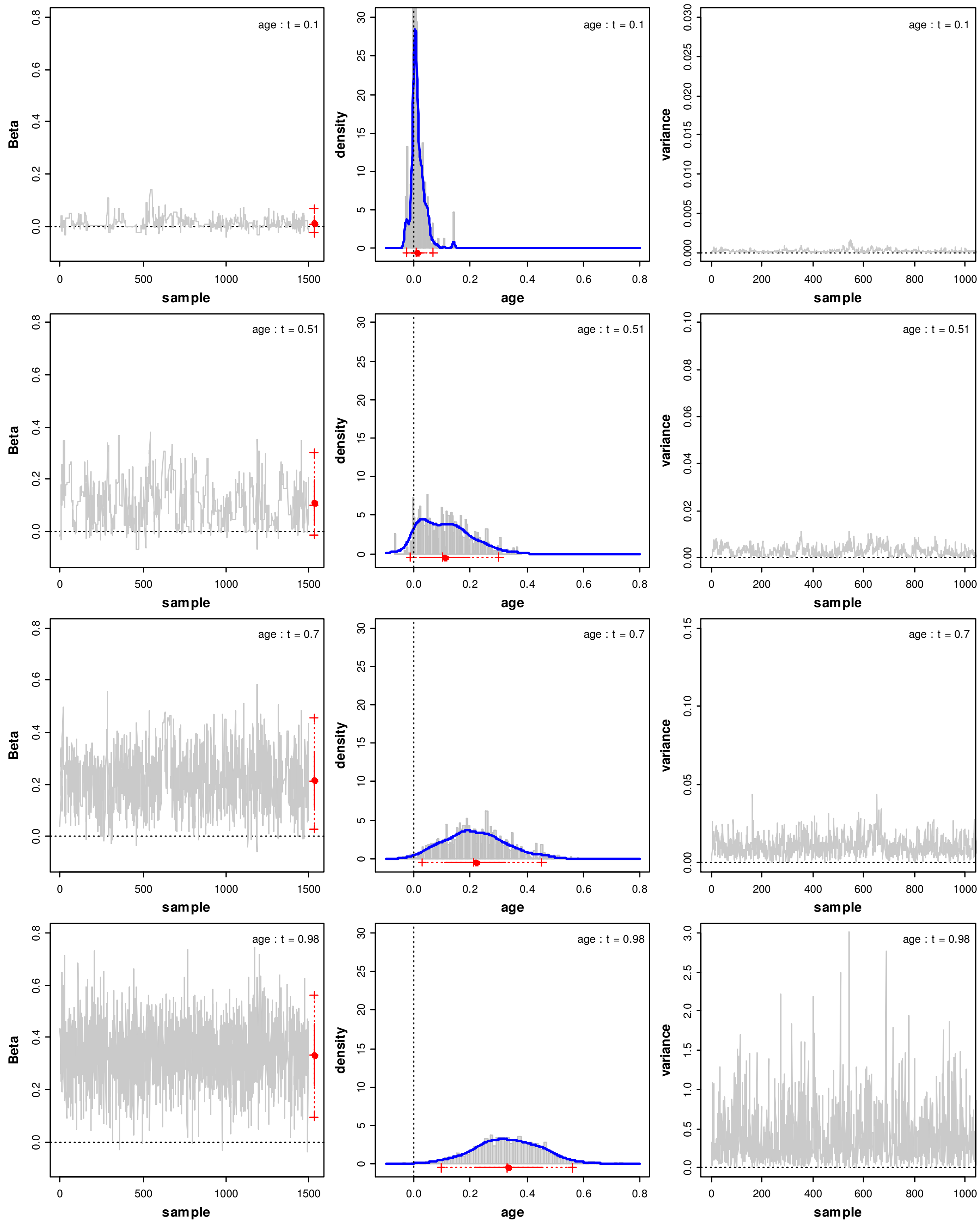


Figure 12.13: MCMC sample of the regression coefficient of covariate *age* and the corresponding variance parameter from the CRR model based on the full likelihood under the Bayesian lasso prior. Left column: Trace plot of the sample of the regression coefficient β_{age} for the fixed values t = 0.10, 0.51, 0.70, 0.98. Middle column: Corresponding kernel density estimates of the marginal posterior density based on Gaussian kernels. Right column: Trace plot of the samples of the related variance parameter τ^2_{age}. The red plots at the border of the first and second column display summary statistics of the marginal posterior distribution. The red points mark the mean, the red solid lines display the standard errors and the red dashes mark the median and 95 % empirical quantiles.

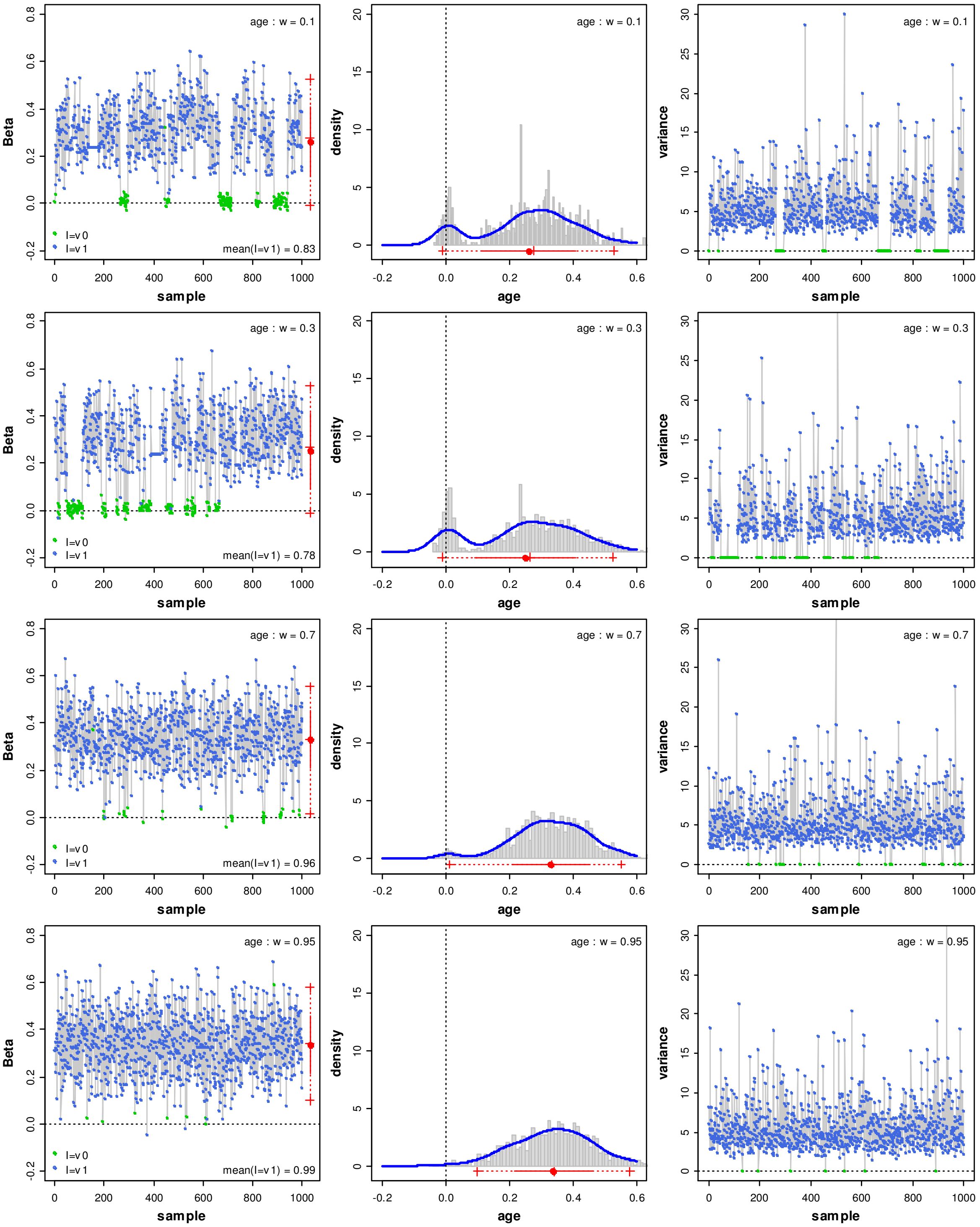


Figure 12.14: MCMC sample of the regression coefficient of covariate *age* and the corresponding variance parameter from the CRR model based on the full likelihood under the Bayesian NMIG prior. Left column: Trace plot of the sample of the regression coefficient β_{age} for the fixed values $\omega = 0.1, 0.30, 0.70, 0.95$. Middle column: Corresponding kernel density estimates of the marginal posterior density based on Gaussian kernels. Right column: Trace plot of the samples of the related variance parameter τ^2_{age}. The red plots at the border of the first and second column display summary statistics of the marginal posterior distribution. The red points mark the mean, the red solid lines display the standard errors and the red dashes mark the median and 95 % empirical quantiles. In the left and right column the green and blue points mark the sampled values, if the sample value of the corresponding indicator equals v_0 or v_1. The posterior mean estimate of the Bayesian NMIG indicator is given at the bottom right side of the figures in the first column.

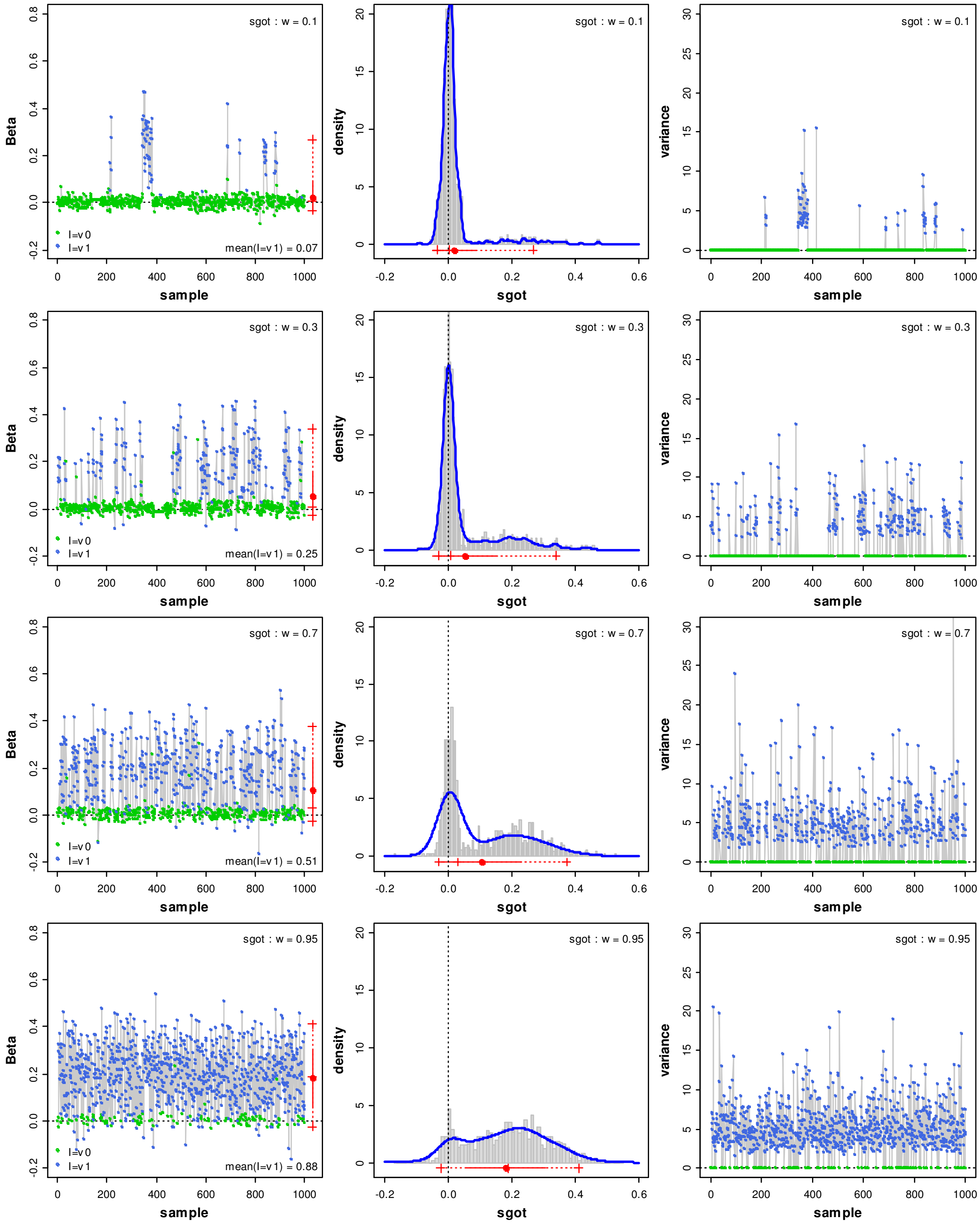


Figure 12.15: MCMC sample of the regression coefficient of covariate *sgot* and the corresponding variance parameter from the CRR model based on the full likelihood under the Bayesian NMIG prior. Left column: Trace plot of the sample of the regression coefficient β_{sgot} for the fixed values $\omega = 0.05, 0.30, 0.70, 0.95$. Middle column: Corresponding kernel density estimates of the marginal posterior density based on Gaussian kernels. Right column: Trace plot of the samples of the related variance parameter τ^2_{sgot}. The red plots at the border of the first and second column display summary statistics of the marginal posterior distribution. The red points mark the mean, the red solid lines display the standard errors and the red dashes mark the median and 95 % empirical quantiles. In the left and right column the green and blue points mark the sampled values when the sample value of the corresponding indicator equals v_0 or v_1. The posterior mean estimate of the Bayesian NMIG indicator is given at the bottom right side of the figures in the first column.

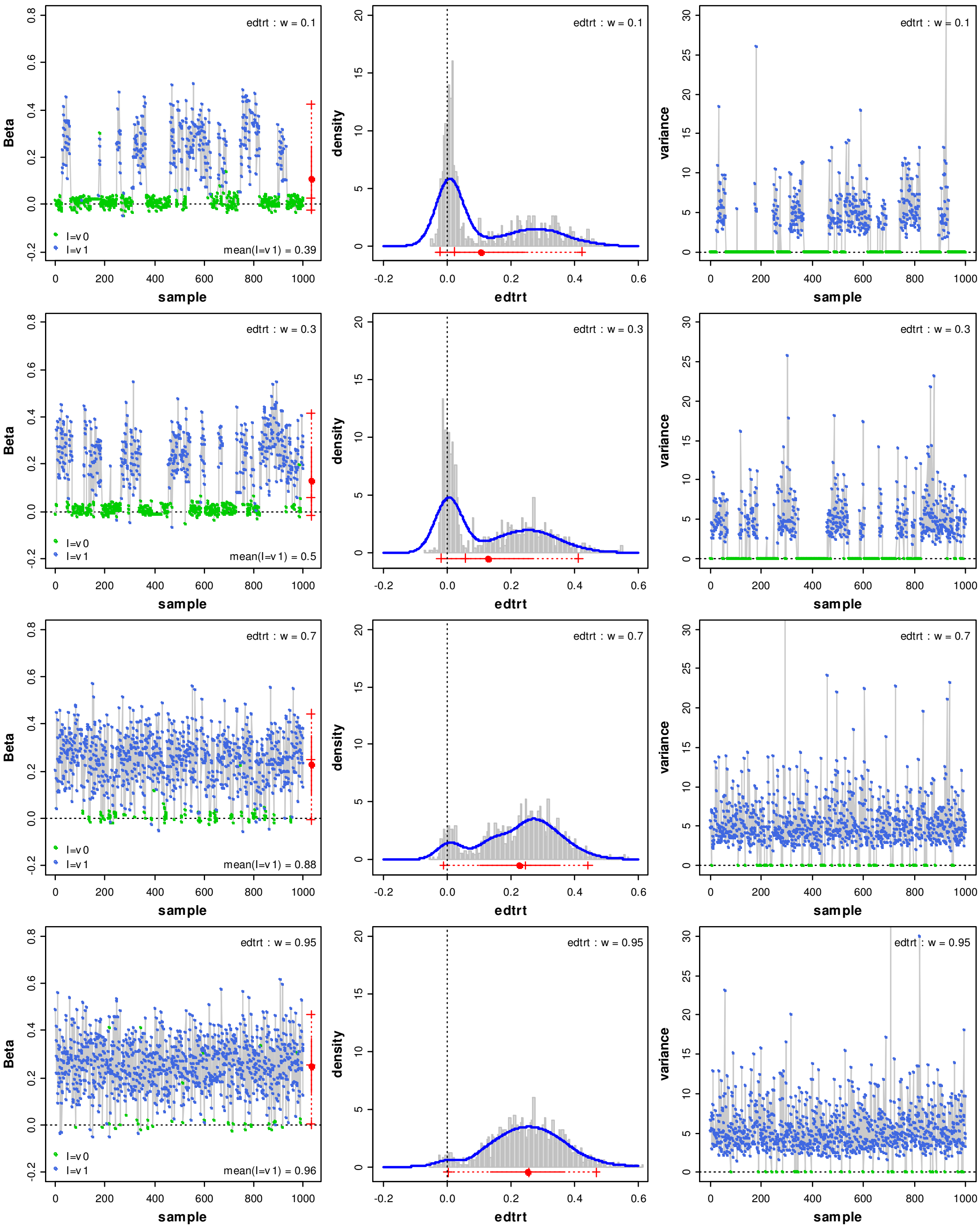


Figure 12.16: MCMC sample of the regression coefficient of covariate *edtrt* and the corresponding variance parameter from the CRR model based on the full likelihood under the Bayesian NMIG prior. Left column: Trace plot of the sample of the regression coefficient β_{edtrt} for the fixed values $\omega = 0.1, 0.30, 0.70, 0.95$. Middle column: Corresponding kernel density estimates of the marginal posterior density based on Gaussian kernels. Right column: Trace plot of the samples of the related variance parameter τ^2_{edtrt} . The red plots at the border of the first and second column display summary statistics of the marginal posterior distribution. The red points mark the mean, the red solid lines display the standard errors and the red dashes mark the median and 95 % empirical quantiles. In the left and right column the green and blue points mark the sampled values when the sample value of the corresponding indicator equals v_0 or v_1. The posterior mean estimate of the Bayesian NMIG indicator is given at the bottom right side of the figures in the first column.

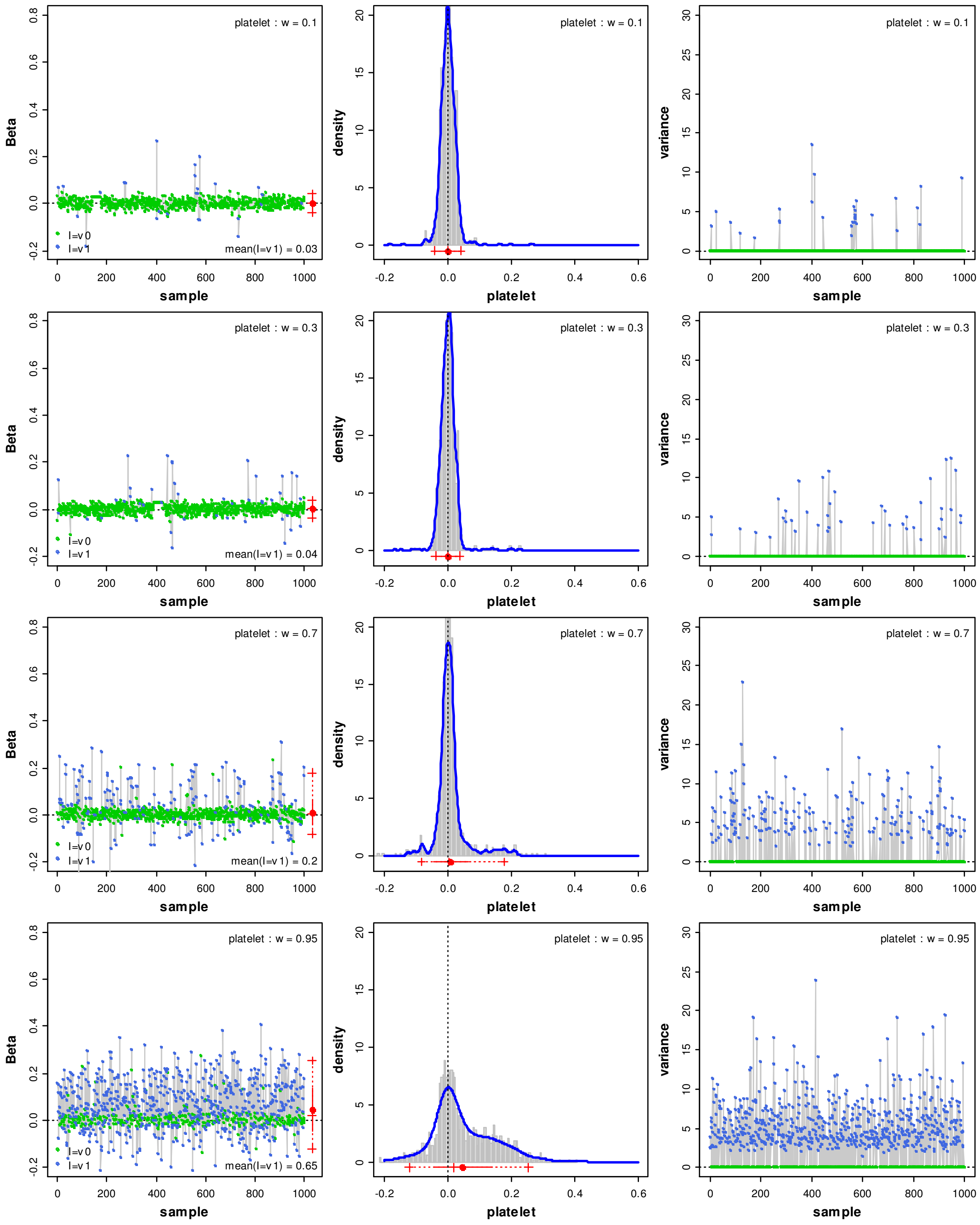


Figure 12.17: MCMC sample of the regression coefficient of covariate *platelet* and the corresponding variance parameter from the CRR model based on the full likelihood under the Bayesian NMIG prior. Left column: Trace plot of the sample of the regression coefficient β_{platelet} for the fixed values $\omega = 0.1, 0.30, 0.70, 0.95$. Middle column: Corresponding kernel density estimates of the marginal posterior density based on Gaussian kernels. Right column: Trace plot of the samples of the related variance parameter τ^2_{platelet}. The red plots at the border of the first and second column display summary statistics of the marginal posterior distribution. The red points mark the mean, the red solid lines display the standard errors and the red dashes mark the median and 95 % empirical quantiles. In the left and right column the green and blue points mark the sampled values when the sample value of the corresponding indicator equals v_0 or v_1. The posterior mean estimate of the Bayesian NMIG indicator is given at the bottom right side of the figures in the first column.

hazard. From the top to the bottom of the figures the complexity parameter is increased, i. e. the regularization is reduced. From the left to the right we see the trace plot of the sample of the regression coefficient (first column), the corresponding marginal posterior density (middle column, computed as kernel density estimate with Gaussian kernel) and the trace plot of the variance parameter sample, which controls the covariate-specific penalization. At the margins of the plots in the first and second column the empirical posterior mean, median and 95% quantiles of the marginal distribution of the regression coefficients are displayed.

In particular **Figure 12.13** shows the associated samples for the covariate *age* under the Bayesian lasso prior, if the complexity parameter t increases from $t = 0.1$ to $t = 0.98$. We clearly observe that the sampled values of the variance $\tau^2_{\beta_j}$ increase with increasing values t, inducing a decrease in the covariate-specific penalty $\tau^{-2}_{\beta_j}$. The impact of the decrease in the penalty is reflected in the sampled values of the regression coefficients and in the associated density estimate, which both become less concentrated around zero.

Figure 12.14 to **Figure 12.17** show the samples of the covariates *age*, *sgot*, *edtrt* and *platelet* under the Bayesian NMIG prior, if the complexity parameter ω increases from $\omega = 0.1$ to $\omega = 0.95$. The green and blue dots in the samples mark the associated values $I_j = v_0$ and $I_j = v_1$ of the indicator variable and the posterior relative frequency of the indicator value $I_j = v_1$ are annotated at the right bottom ($\text{mean}(I = v_1)$). With increasing values of the complexity parameter ω we observe an increase of the number of sampled values $I_j = v_1$ associated with an increase of the sampled variances $\tau^2_{\beta_j}$. The resulting global reduction of penalization is reflected in the samples of the regression coefficients and the corresponding density estimate of the marginal posterior which both are shifted away from zero. Nevertheless, the marginal posteriors of the regression coefficients (and variance parameters) show a more or less pronounced bimodality in each figure, which depends on the effect size and the frequencies of the indicators $I = v_1$. In particular small or large effects (as defined by the setting of the NMIG hyperparameters), like that of *platelet* and *age*, are strongly or weakly regularized over a broad range, as also shown in the coefficient paths, so that a notable bimodality of the effect distribution mainly occurs at the right or left margins of the complexity parameter. In general, the bimodality of the marginal posteriors derogates somehow the empirical mean as appropriate summary statistic in cases when effects of moderate size are present. But, as seen in the Simulation Section 11.5, the associative estimation strategy, using only the subsamples belonging to the nearly unregularized component with $I = v_1$ for effects exceeding a given frequency threshold, has shown no improvement in the predictive performance.

Results for the nonlinear predictor

We finally consider the results achieved for the extended AFT and CRR models given in (12.3) and (12.4), when the predictor (12.2) with nonlinear effects of the continuous covariates is assumed. In summary, we found no strong evidence for the nonlinear form of any effect under both survival models, if the estimates are considered in the regions where most of the observations occur.

Nonlinear effects

Figure 12.18 displays the estimated nonlinear effect of the covariate *bili* as representative. Shown are the results obtained under the Bayesian NMIG and Bayesian lasso prior for the linear effects together with the results from the stepwise variable selection. The left side of **Figure 12.18** shows the results for the estimated CRR model and the right side the corresponding results for the AFT model.

The shape of the estimated functions is similar at both sides of the figure, but under the Bayesian AFT model the pointwise credible bands, indicating the uncertainty, are more concentrated in the regions with most of the observations than those of the Bayesian CRR model and the slope is smaller. The shift in the Bayesian CFL estimates is due to the internal centering of the spline estimates.

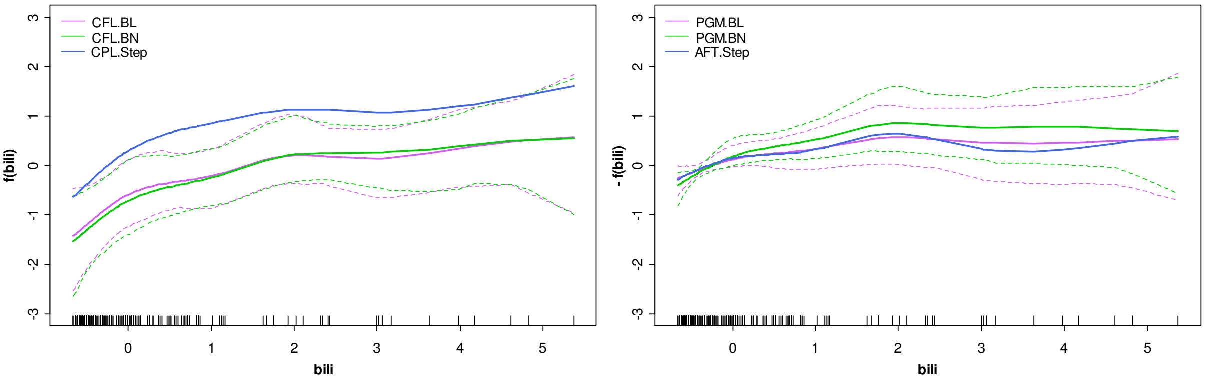


Figure 12.18: Estimation of the nonlinear effect of the covariate *bili* when all continuous covariates are modeled as P-splines in the CRR model (left side) and the AFT model (right side). Both figures display the posterior mean estimates of the nonlinear effect (solid lines) with 95% pointwise credible bands (dashed lines) for the Bayesian lasso and Bayesian NMIG regularization together with the corresponding estimates from the stepwise selection of the frequentist CRR and AFT model with Gaussian error.

Linear effects

The estimates of the remaining seven linearly modeled covariate effects are displayed in **Figure 12.19**, where the upper panel shows the estimates from the CRR model and the lower panel those from the AFT model. The impact of the nonlinear modeling on the linear effects can be viewed by considering the frequentist estimates, e. g. in terms of the covariates *ascites* in the CRR model or *stage* in the AFT model, that differ clearly in size compared to the estimates under strictly linear predictor. Consequently, also the amount of shrinkage is prior-specific adapted and varies. In the CRR model with strictly linear predictor the covariate *ascites* is often excluded from the final models if variable selection (HS.STD, Step) is applied which is still not the case with the nonlinear predictor. The reverse results in the AFT model, e. g. for the covariate *edtrt*. Finally, the effects of the covariates *hepmeg* and *trt* are still close to zero as in the CRR and AFT models with strictly linear predictor. In summary, the nonlinear modeling has a clear impact on the estimates of the linear effects and the variable selection.

NMIG indicators

In terms of the Bayesian NMIG prior this variation is reflected in the associated posterior relative frequencies of the Bayesian NMIG indicator variable $I_j = v_1$. **Figure 12.20** displays the resulting posterior inclusion probabilities in the CRR model with P-spline hazard (left side) and the AFT model with PGM error (right side). In comparison to the previous models with strictly linear predictor, we observe changes in the inclusion probabilities of covariates *stage* and *edtrt*. Only the covariate *stage* exceeds in the CRR model the frequency threshold 0.5 used in the HS.IND selection rule. This covariate is also the only one marked as significant in the frequentist CRR model. The previously higher inclusion probability of the covariate *edtrt* decreases here and falls below the threshold 0.5. Under the AFT model none of both, previously high, inclusion probabilities of *stage* and *edtrt* exceed further the threshold.

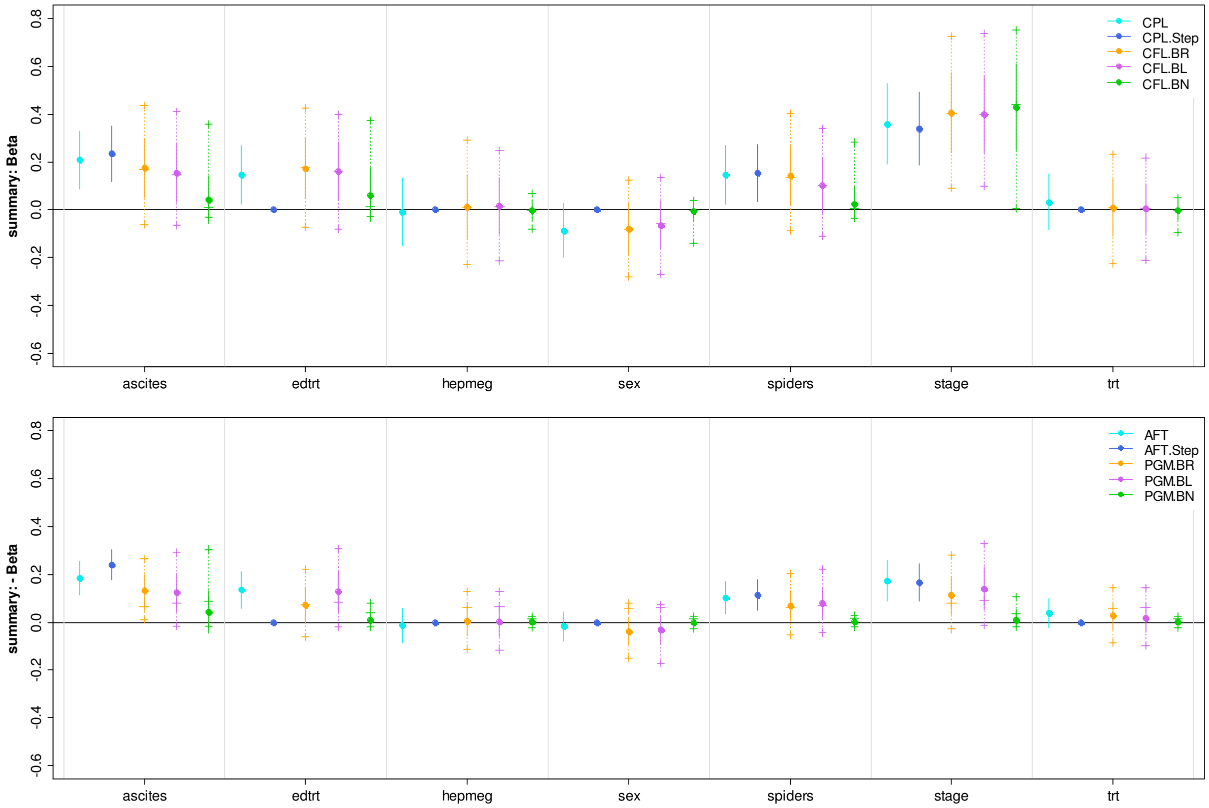


Figure 12.19: Estimated coefficients in the CRR (upper panel) and AFT model (lower panel) when all continuous covariates are modeled as P-splines. The points mark the estimates of the regression coefficients and the lines display the corresponding standard errors. For the Bayesian procedures the points mark the mean, the solid lines display the standard errors and the additional dashes mark the median and 95 % empirical quantiles of the marginal posterior distribution of the regression coefficients.

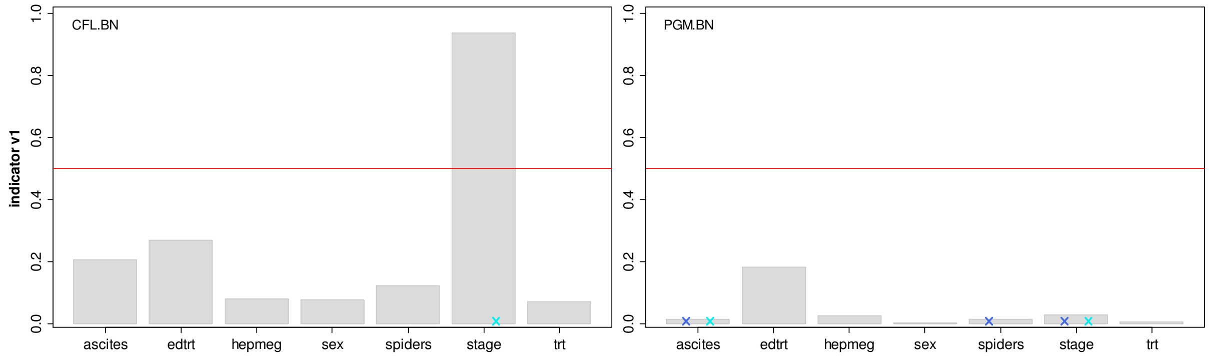


Figure 12.20: Estimated inclusion probabilities based on posterior relative frequencies of the Bayesian NMIG indicator variable value $I_j = v_1$ in the CRR model with P-spline hazard (left side) and the AFT model with PGM error (right side) when continuous covariates are modeled as P-splines. The crosses in the bars mark the covariates from the corresponding frequentist models, which are significant with respect to the p-value 0.05 (cyan) and which are selected by the frequentist stepwise variable selection procedure (dark blue). The red horizontal line marks the cut off value 0.5 of the hard shrinkage selection criterion HS.IND.

Baseline quantities

Finally, **Figure 12.21** shows the estimates of the log-baseline hazard in the CRR model (left side) and the estimates of the baseline error densities in the AFT model (right side). The difference in the vertical location of the log-baseline hazard under the Bayesian NMIG prior results, because the covariate *age* is modeled linear in this specific case. The corresponding estimate under the Bayesian ridge prior, with nonlinear modeled *age*, closely coincides with the displayed estimate of the log-hazard function under the Bayesian lasso prior. The estimated shape of the log-baseline hazard shows no obvious deviations to the shape when all covariate effects are assumed to be linear and the same holds under the Bayesian versions of AFT model. The observed difference in the location to the frequentist approach is due to the internal centering of the spline estimates.

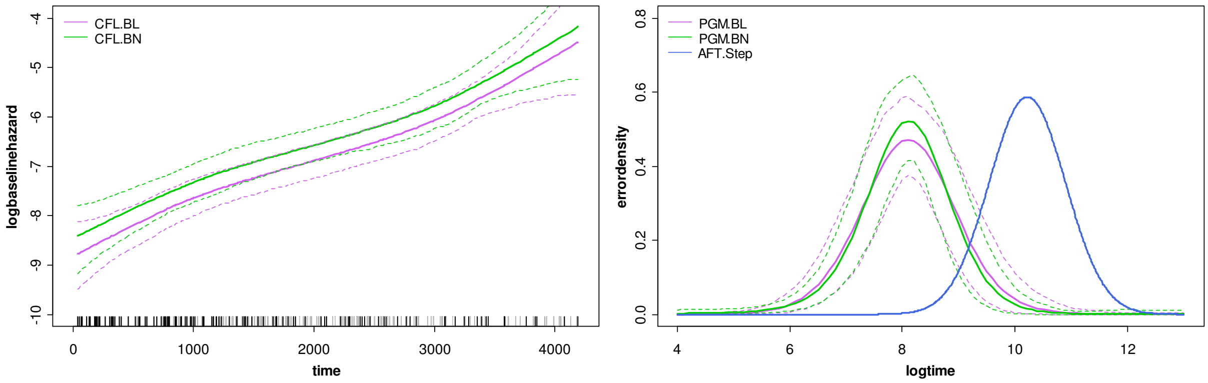


Figure 12.21: Estimation of the log-baseline hazard in the CRR model P-spline hazard (left side) and estimation of the baseline error density in the AFT model with PGM error (right side) when continuous covariates are modeled as P-splines. Left side: Posterior mean estimate of the log-baseline hazard function (solid lines) with 95% pointwise credible bands (dashed lines) for the Bayesian lasso and Bayesian NMIG regularization. Right side: Posterior mean estimate of the baseline error density together (solid lines) with 95% pointwise credible bands (dashed lines) for the Bayesian lasso and Bayesian NMIG regularization together with the frequentist estimates of the AFT model with Gaussian error distribution and the corresponding result after stepwise selection.

Final remark

In summary, the nonlinear modeling of continuous covariates has shown an impact of the remaining linear effect estimates. Although none of the estimated nonlinear effects show a clear nonlinearity, the size of some estimated linear covariate effects clearly changes compared to the models with strictly linear predictor. The differences under the various regularization models are induced mainly by the diversification of the linear effects, as shown in terms of the unregularized estimates. The shape of the Bayesian estimates of the baseline quantities seems to be only marginally affected by the increased model complexity introduced with the nonlinear modeling of some covariate effects. Sleeper and Harrington (1990) found evidence for the nonlinear form of the covariate *age*, but they use a reduced set of covariates in combination with the log-transformation of the continuous covariates *alkphos* and *bili* in order to reduce the influence of outliers on the spline fits.

13. Adult myeloid leukemia in northwest England

13.1. Data

The data set used in this section contains information of adult myeloid leukemia patients in northwest England who have been diagnosed between 1982 and 1998. Previous analyses can be found in Henderson et al. (2002), where the detection of spatial variation in survival times is based on strictly linear covariate effects, while more flexible forms are considered in Kneib and Fahrmeir (2007) in the context of geoadditive hazard regression models. The leukemia data was originally provided by Leonhard Held (University of Zurich, UZH).

In the data we have $n = 1043$ observations, where 15.8 % of the observations are right censored. The mean survival time of a patient is 533 days with median 185 days and a range from 1 to 4977 days. **Table 12.1** displays the available variables of the dataset. For the Townsend index, which measures the deprivation in the given 24 districts of residence, higher values indicate poorer regions while smaller values correspond to wealthier regions. In the data the values of the Townsend index range from −6.09 to 9.55. The 24 administrative districts of northwest England are shown in **Figure 13.1**. For the analysis we use effect coded districts with reference to district 24. We restrict our analysis to the level of the districts to enable the application of the shrinkage priors, while geostatistical models are possible, if the available exact locations of the patient's residences are used, compare Henderson et al. (2002) and Kneib and Fahrmeir (2007).

time	number of days between registration and death
cens	status at endpoint, 0 = censored, 1= death
age	age of the patient in years
sex	sex of a patient (−1 = female, 1 = male)
wbc	white blood cell count at diagnosis
tpi	Townsend deprivation index, which measures the deprivation for the enumeration district of residence. Higher values indicate less affluent areas.
district	24 districts of patient's residence. The enumeration of the districts is displayed in **Figure 13.1**
xcoord, ycoord	Exact location coordinates (latitude, longitude) of the patient's residence

Table 13.1: List of available covariates used in the analysis of the leukemia data.

13.2. Analysis

Kneib and Fahrmeir (2007) analyzed the data in the framework of geoadditive hazard regression models with a generalized mixed model based approach for inference. They consider, besides the flexible shape of the baseline hazard, also nonlinear covariate effects and utilize the available spatial information by means of a district level analysis, since the observations are clustered by the districts in connected geographical regions. We apply the extended AFT and CRR model and specify, with respect to this previous analysis, the effects of the three continuous covariates *age*, *wbc* and *tpi* throughout nonlinear to make the results comparable, even though there was no strong evidence found for a nonlinear influence of the covariates *age* and *wbc*.

The structure of the predictor is given by

$$\eta_i = \gamma_0 + \gamma_1 \text{sex}_i + f_1(\text{age}_i) + f_2(\text{wbc}_i) + f_3(\text{tpi}_i) + \sum\nolimits_{j=1}^{23} \beta_j d_{ij}\,, \tag{13.1}$$

where $f_1(\cdot), f_2(\cdot)$ and $f_3(\cdot)$ are smooth functions of the three continuous covariates which are modeled by cubic P-splines. The linear effect of the covariate *sex* is kept unregularized and the district effects β_j, $j=1,...,23$, are equipped with the Bayesian shrinkage priors. Also in the frequentist stepwise procedure only the district effects β_j are considered for the variable selection.

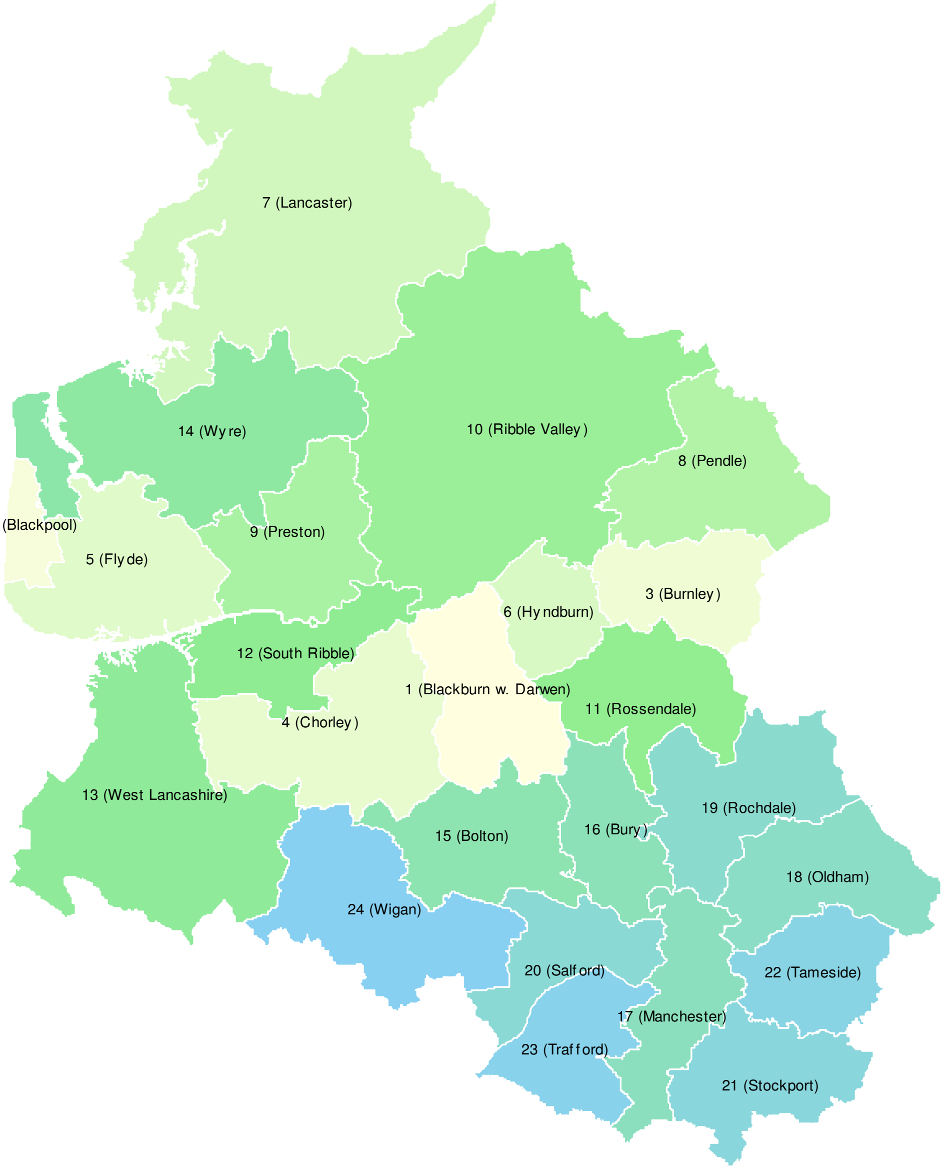


Figure 13.1: Administrative districts of the ceremonial counties Lancashire (1-14) and Greater Manchester (15-24) in North West England (Source: http://en.wikipedia.org/wiki/Subdivisions_of_England).

In contrast, Kneib and Fahrmeir (2007) model the districts with a spatial effect $f_{spat}(\cdot)$, where the spatial neighborhood structure is utilized in the inferential procedure. In particular they assume $f_{spat}(j) = \beta_j = N_j^{-1} \sum_{j' \in \delta_j} \beta_{j'} + u_j$, $j=1,...,24$, with Gaussian error $u_j \sim N(0, \tau_{spat}^2 N_j^{-1})$, where $j' \in \delta_j$ denotes that district j' is a neighbor of district j, in the sense that they share a common boundary, and N_j is the number of neighbor districts. In summary, the effect of a district j is assumed to be conditionally Gaussian, with the mean of the effects of neighbor districts as expectation and a variance that is inverse proportional to the number of its neighbors. We utilize the `spatial` term within the BayesX-method `regress`, to reproduce the spatial results as competitor (CFL.BS).

To fit the AFT and CRR models, we use the same specification of the model components and priors as for the analysis of the PBC data, compare Section 12.2. With the described basic prior specification for the regularized linear effects we obtain in the AFT model with PGM error small, close to zero estimates as district effects. Therefore we fit in addition the Bayesian AFT model with adjusted prior versions to reduce the shrinkage and support a stronger influence of the districts in the final models. The hyperparameters of the shrinkage parameters are set to $h_{2,\lambda} = 1.1$ for the Bayesian lasso and ridge prior and $h_{1,\omega} = 32$, $h_{2,\omega} = 64$ for the Bayesian NMIG prior. For the MCMC runs, we use 30000 iterations with a burnin of 15000 and thin the chain by 15, which results in a MCMC posterior sample of size 1000.

To model the nonlinear effect of the covariates *age*, *wbc* and *tpi*, we use the default settings in the `pspline()` term within the formula of the `R`-functions `survreg()` and `coxph()`. The `penalized` procedure is not applicable, because it does not support the combined estimation of nonlinear effects.

13.3. Results

Linear effects

The frequentist and Bayesian estimates of the district effects are summarized in **Figure 13.2** for the CRR model and in **Figure 13.3** for the AFT model. Besides the point estimates the one-standard-deviation region is marked by the solid lines around the point estimate and in addition for the Bayesian estimates also the 95 % credible intervals are given by the dashed lines together with the posterior median estimate. The standard deviation and credible regions are utilized to drive the Bayesian hard shrinkage variable selection as formulated for the HS.STD and HS.CRI criterion together with the Bayesian NMIG indicator based criterion HS.IND. In the lower panel of the figures we find some of the results of the variable selection, in particular those from the stepwise selection, the Bayesian lasso and ridge prior in combination with the HS.STD criterion and the Bayesian NMIG prior with HS.IND criterion. For the spatial results (CFL.BS) we use the 80% pointwise credible interval as in Kneib and Fahrmeir (2007) to select the districts.

Figure 13.2 shows the results for the CRR model obtained with the full likelihood under the P-spline model for the log-baseline hazard and the partial likelihood based frequentist estimates. With the frequentist analysis (CPL, CPL.Step) and Bayesian analysis under the different shrinkage priors (CFL.BL, CFL.BR, CFL.BN) we find commonly a clear increased risk to die in district 6 and a clear decreased risk in districts 9 and 11. Under the Bayesian ridge prior, with the uniform proportion of shrinkage for all regression coefficients, we observe a weaker regularization of the estimates compared to the lasso or NMIG prior.

The absolute values of the spatial district estimates (CFL.BS) tend in general to smaller values due to the considered neighborhood structure. For example, the estimate of district 6 is affected by the surrounding neighbor districts 1, 3, 10 and 11 and the spatial prior structure cause an adaption of this estimate to the neighborhood mean that is by trend smaller. As a consequence, the spatial estimates of the three particular districts 6, 9 and 11 have the largest differences to the estimates from the other

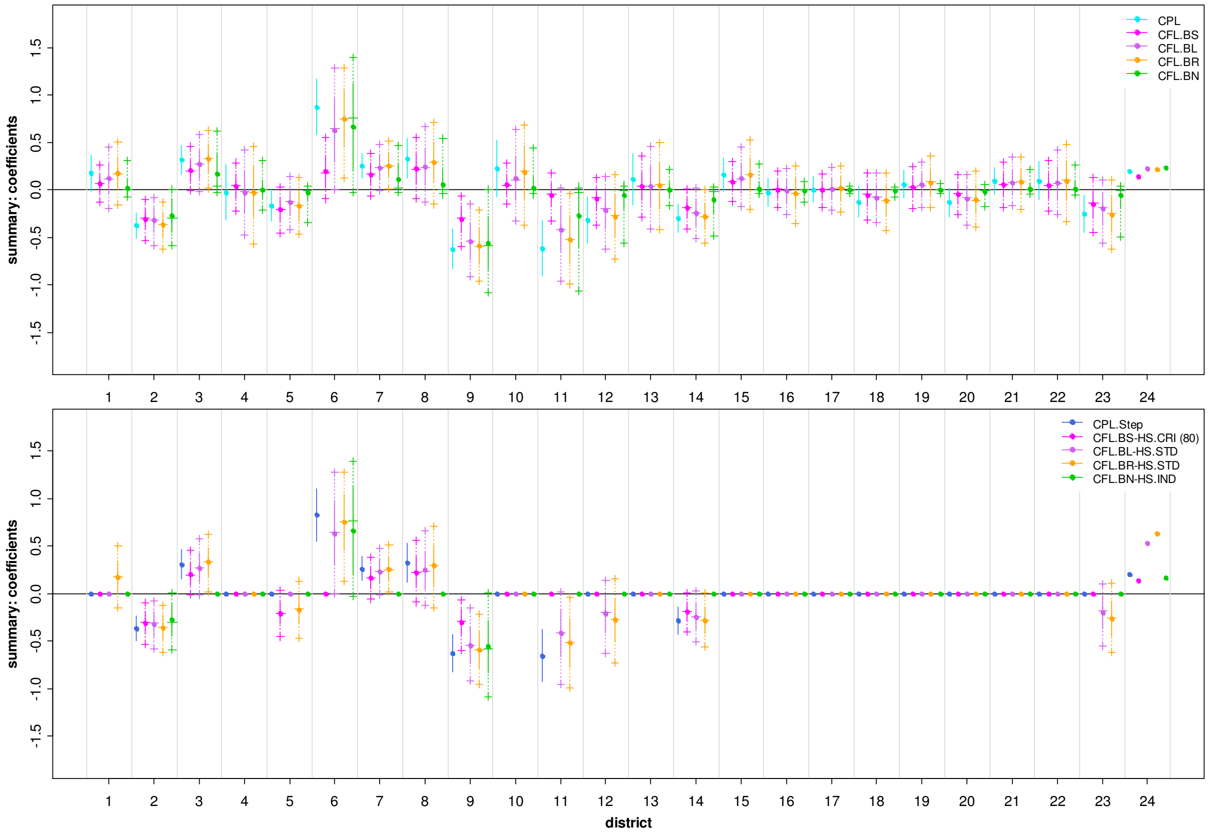


Figure 13.2: Estimated district coefficients without (upper panel) and with variable selection (lower panel) in the CRR model. The points mark the estimates of the regression coefficients and the lines display the corresponding standard errors. For the Bayesian procedures the points mark the mean, the solid lines display the standard errors and the additional dashes mark the median and 95 % empirical quantiles of the marginal posterior distribution of the regression coefficients. The selection of the spatial district effects (CFL.BS) is based on hard shrinkage with the empirical 80% quantiles.

remaining applied methods. If variable selection is applied only a few districts stay in the final models, compare **Figure 13.5** for a spatial visualization.

Figure 13.3 shows the results for the AFT model based on the modified hyperparameter setting. In concordance to the CRR model the effects of districts 6 and 11 induce a clear decrease and increase of the survival time under all methods. In contrast to the CRR model the influence of district 9 is negligible, since the estimates are close to zero under all estimation procedures and this district does not appear in the final models after variable selection.

In the analysis of the PBC data the absolute values of the unpenalized estimated effects obtained with the AFT model were often smaller as those from the CRR model. In this section we observe the opposite. This highlights again the aspect that the estimated effects from both model classes are not directly comparable. However, besides the differences in the effect sizes, in summary we observe at least the same direction in the risk / survival time affection for all estimates in the CRR and AFT model and we find a range of districts (4, 10, 13, 15-22) that does not appear in the sparse final models of both survival model classes.

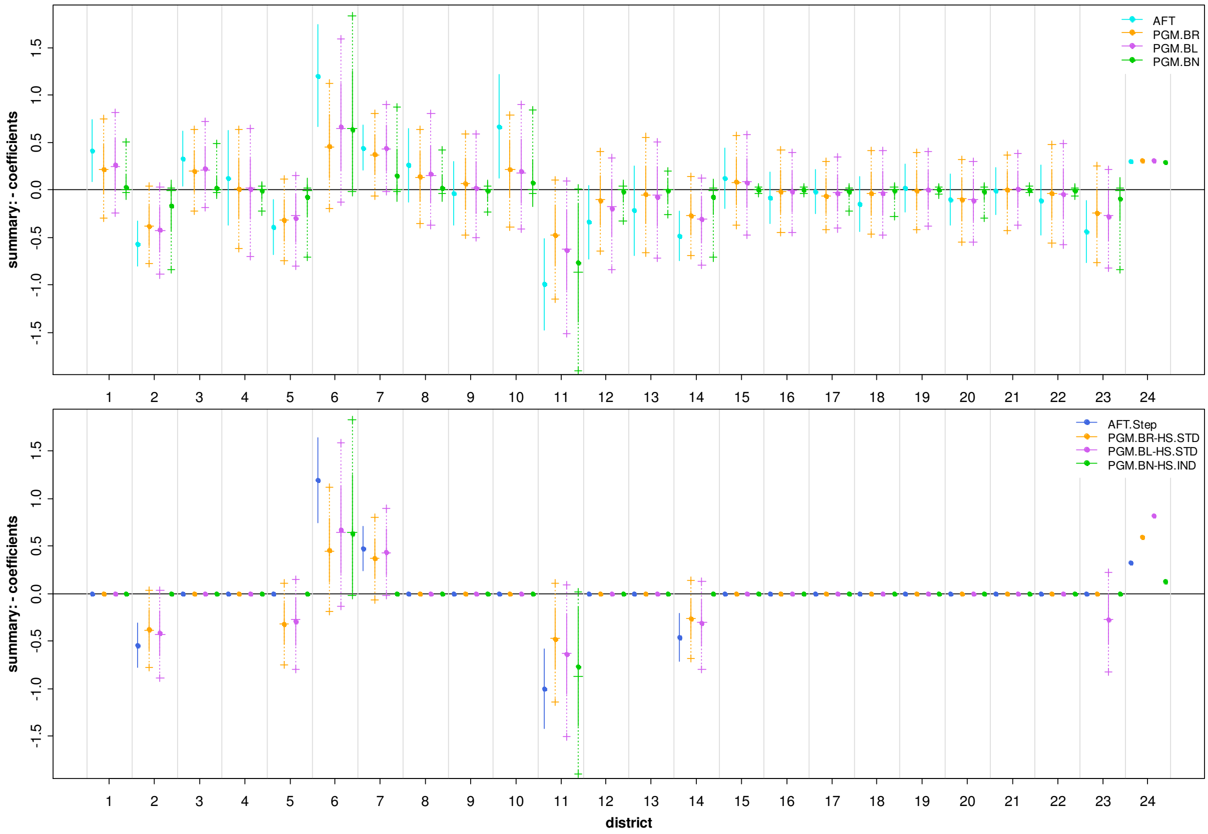
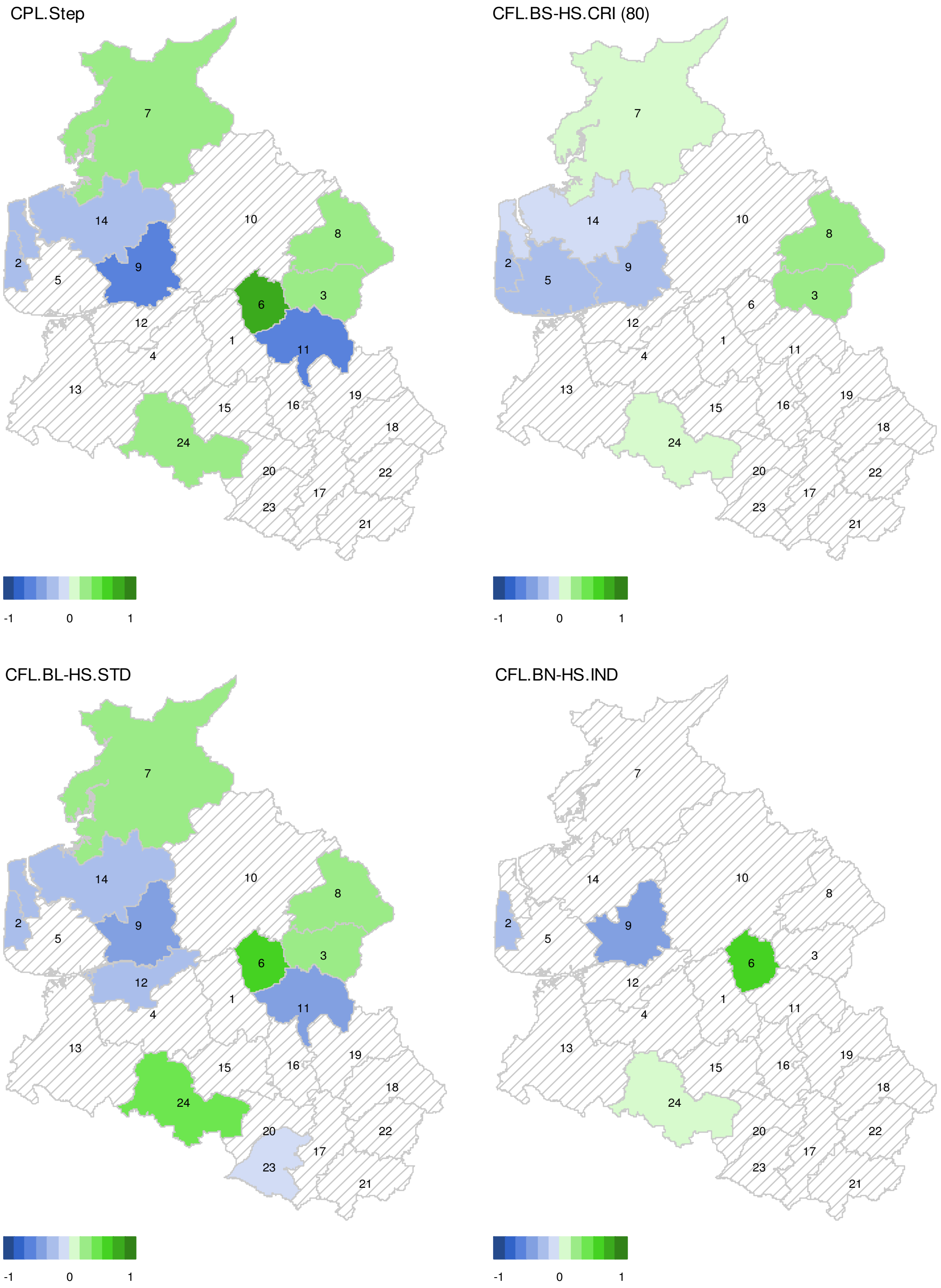


Figure 13.3: Estimated coefficients without (upper panel) and with variable selection (lower panel) in the AFT model. In the frequentist model the error is modeled by a Gaussian distribution and in the Bayesian models by a PGM. The points mark the estimates of the regression coefficients and the lines display the corresponding standard errors. For the Bayesian procedures the points mark the mean, the solid lines display the standard errors and the additional dashes mark the median and 95 % empirical quantiles of the marginal posterior distribution of the regression coefficients.

Spatial visualization

The district importance structures resulting from the frequentist and Bayesian variable selection are visualized for the CRR model in **Figure 13.4** and for the AFT model in **Figure 13.5**. The results correspond to the estimates in the lower panels of **Figure 13.2** and **Figure 13.3**. The green shaded districts have effects that increase the risk to die in CRR model or decrease the survival time in the AFT model, and in contrast the blue shaded districts have effects that decrease the risk to die in CRR model or increase the survival time in the AFT model. The shades change for effect sizes in the range from −1.2 to 1.2 with difference 0.2.

Regions with an increased risk resp. shorter survival times are located in the northern and eastern part of the map, while the regions with a decreased risk resp. longer survival times are located in the north-western part. In the southern part of the map there are also some districts with enhanced or less pronounced effects, but most of the districts there show no influence on the patient's survival.


CPL.Step

CFL.BS-HS.CRI (80)

CFL.BL-HS.STD

CFL.BN-HS.IND

Figure 13.4: Estimation of the district coefficients with variable selection in the CRR model. Upper left: Frequentist CRR model based on the partial likelihood and stepwise variable selection. Upper right to lower right: Bayesian CRR model with spatial, lasso and NMIG regularization. The selection of the spatial district effects is based on the hard shrinkage with the empirical 80 % quantiles. The colored legend ranges from -1.2 to 1.2, where the color changes at the distance 0.2.

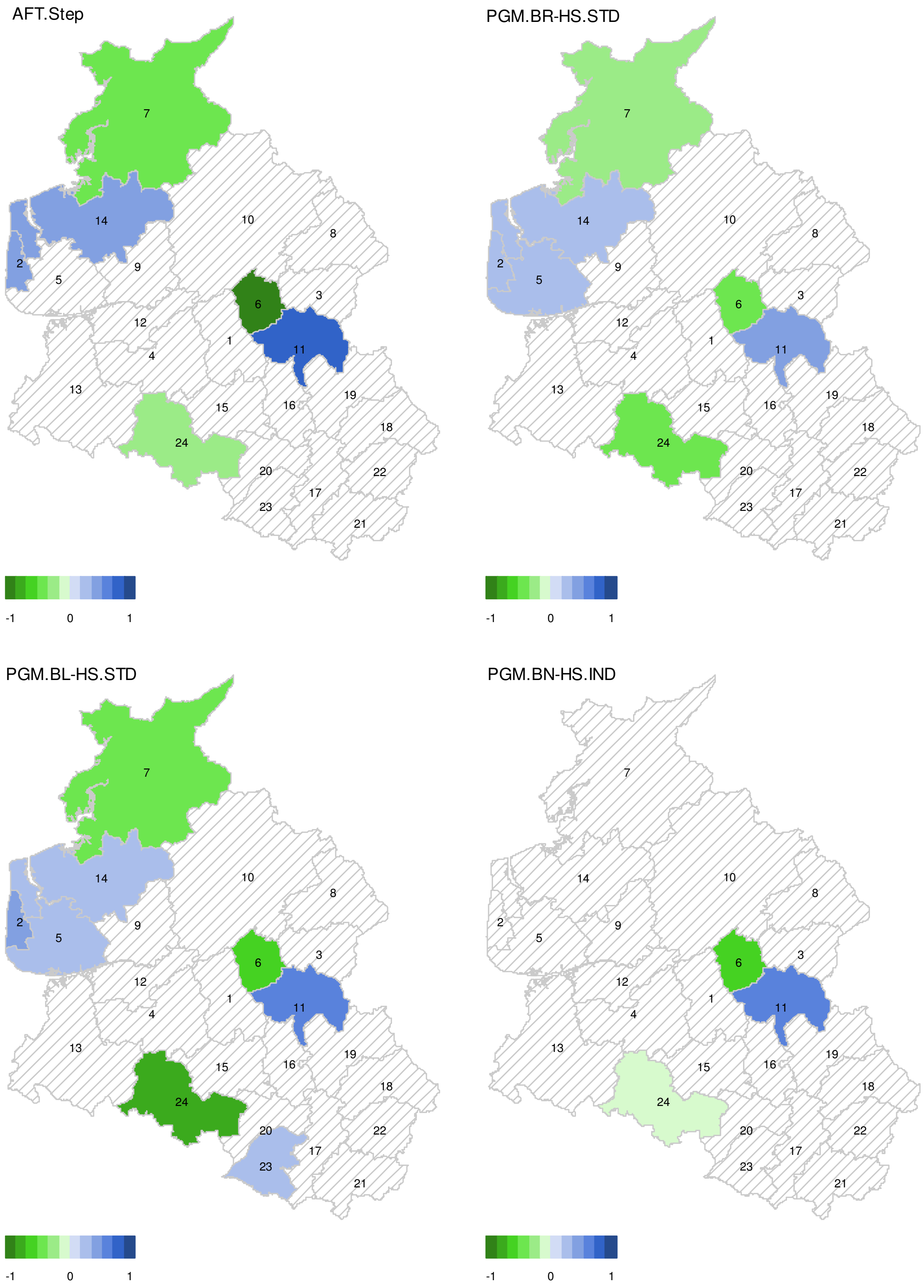


Figure 13.5: Estimation of the district coefficients with variable selection in the AFT model. Upper left: Frequentist AFT model with Gaussian error and stepwise variable selection. Upper right to lower right: Bayesian AFT model with PGM error and ridge, lasso and NMIG regularization. The colored legend ranges from -1.2 to 1.2, where the color changes at the distance 0.2.

NMIG indicators

The posterior relative frequencies of the Bayesian NMIG indicator variables $I_j = v_1$ utilized in the HS.IND selection criterion are presented in **Figure 13.6**. The left panel shows the results from the CRR model with the full likelihood (CFL.BN) the right panel the results from the AFT model with PGM error (AFT.PGM). With the basic hyperparameter setting used so far, we achieve only with the CRR model estimated inclusion probabilities that exceed the HS.IND threshold of 0.5. With exception of district 8 the highest inclusion probabilities are given for those districts that are also selected by the stepwise procedure (CPL.Step) marked at the bottom of the bars, i. e. for districts 2, 3, 6, 7, 9, 11 and 14.

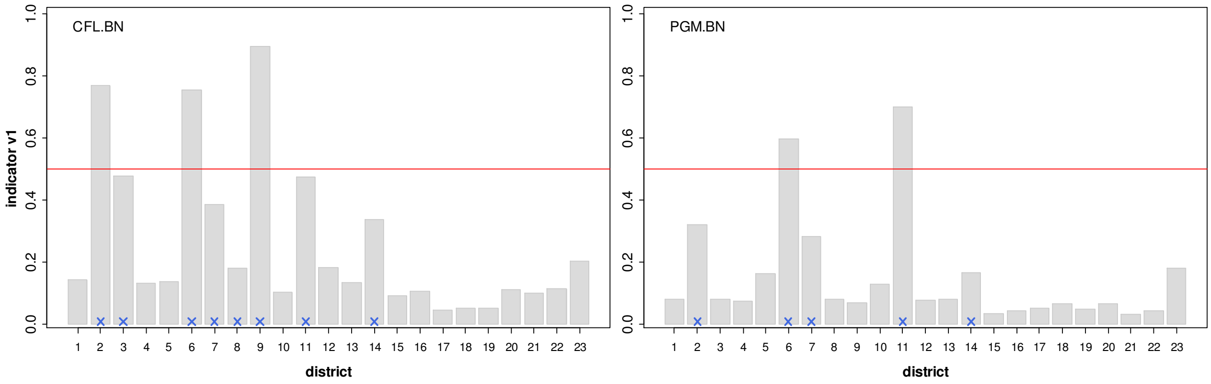


Figure 13.6: Estimated inclusion probabilities based on posterior relative frequencies of the Bayesian NMIG indicator variable value $I_j = v_1$ in the CRR model (left side) and in the AFT model with PGM error (right side). The blue crosses mark the covariates which are selected by the frequentist stepwise variable selection procedure in the CRR and AFT model. The red horizontal line marks the cut off value 0.5 of the hard shrinkage selection criterion HS.IND.

In the AFT model the adjusted hyperparameters $h_{1,\omega} = 32$, $h_{2,\omega} = 64$ force an increase in the Bayesian NMIG complexity parameter ω that reduces the shrinkage of the covariate effects. They are leading to an estimate of the complexity parameter ($\hat{\omega} \approx 0.28$) in the same range as obtained with the CFL.BN model. Besides a general increase of all relative frequencies, compared to the results with the basic hyperparameter setting $h_{1,\omega} = h_{2,\omega} = 1$, we obtain a clear increase in the relative frequencies of the districts 2, 6, 7, 11, 14, 23, which also have enhanced inclusion probabilities in the CFL model. However, only two districts (6, 11) exceed the HS.IND-threshold 0.5. If we rank the districts according to the sizes of the inclusion probabilities, or adjust the HS.IND-threshold, we could also include districts 2 and 7. In Simulation Section 11.5 we have seen that the adaption of the threshold often improves the predictive performance of the associated sparse final model.

Nonlinear effects and baseline quantities

Finally, we consider the estimates of the nonlinear effects in the predictor and the baseline quantities, i. e. the log-baseline hazard function in the CRR model and the baseline error density in the AFT model. As previously observed in the analysis of the PBC data, the estimated shapes of the nonlinear effects are again only marginal affected by the specific regularization (or spatial) prior assumed for the district effects. Since the same holds for the frequentist spline estimates, with and without stepwise variable selection, we present in the following the results by means of the frequentist stepwise selection (Step) and the Bayesian NMIG (BN) prior. The differences in the location of the estimated nonlinear effects or the location parameter in the AFT model are due to the centering of the nonlinear effects in the Bayesian procedures.

Figure 13.7 illustrates the estimated nonlinear effects of the covariates *tpi*, *age* and *wbc* in the CRR model together with the estimates of the log-baseline hazard function. The estimate of the log-baseline hazard (CFL.BN) is compared to the resulting Bayesian estimate when the spatial neighborhood information is used to model the district effects (CFL.BS). Neglecting the differences in the location, the nonlinear effects from the Bayesian or frequentist models show a similar shape. Again the almost linear influence of the covariates *age* and *wbc* is approved and both effects decrease the survival time with increasing values. In the range of the interval with most of the observations the log-baseline hazard decreases in general, but we observe some intervals, e. g. in the second year, where the log-baseline increases (CFL.BS) or the slope is reduced (CFL.BN).

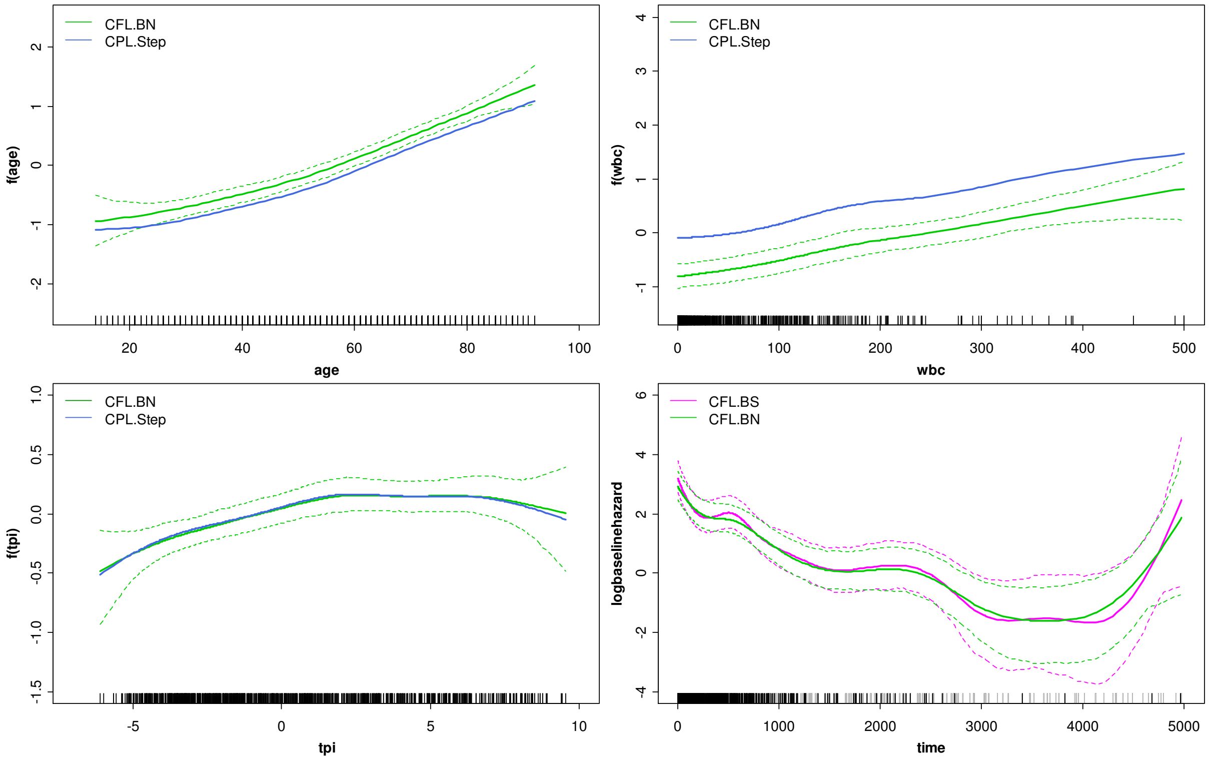


Figure 13.7: Estimates of the nonlinear effect of the covariates *age* (upper left panel), *wbc* (upper right panel) and *tpi* (lower left panel) and the log-baseline hazard (lower right panel) in the CRR model. Displayed are the estimates under the Bayesian NMIG regularization prior in the CRR model based on the full likelihood (CFL.BN) together with those from the frequentist stepwise variable selection (CPL.Step). The log-baseline hazard estimate is compared with the Bayesian estimate utilizing the spatial neighborhood information (CFL.BS). For the Bayesian models the solid lines show the posterior mean estimates and the dotted lines mark the corresponding 95 % pointwise credible bands. The stripes at the x-axis mark the observed values in the data.

Figure 13.8 shows the results for the AFT model (PGM.BN, AFT.Step). Neglecting also the differences in the location, we achieve with the PGM error model a comparable shape of the error density as in the frequentist counterpart, when the error is assumed to be Gaussian. The estimates of the covariates *age* and *wbc* are almost linear as in the CRR model and the effect of *tpi* is rather nonlinear, but the nonlinear effect show a larger slope in the AFT model.

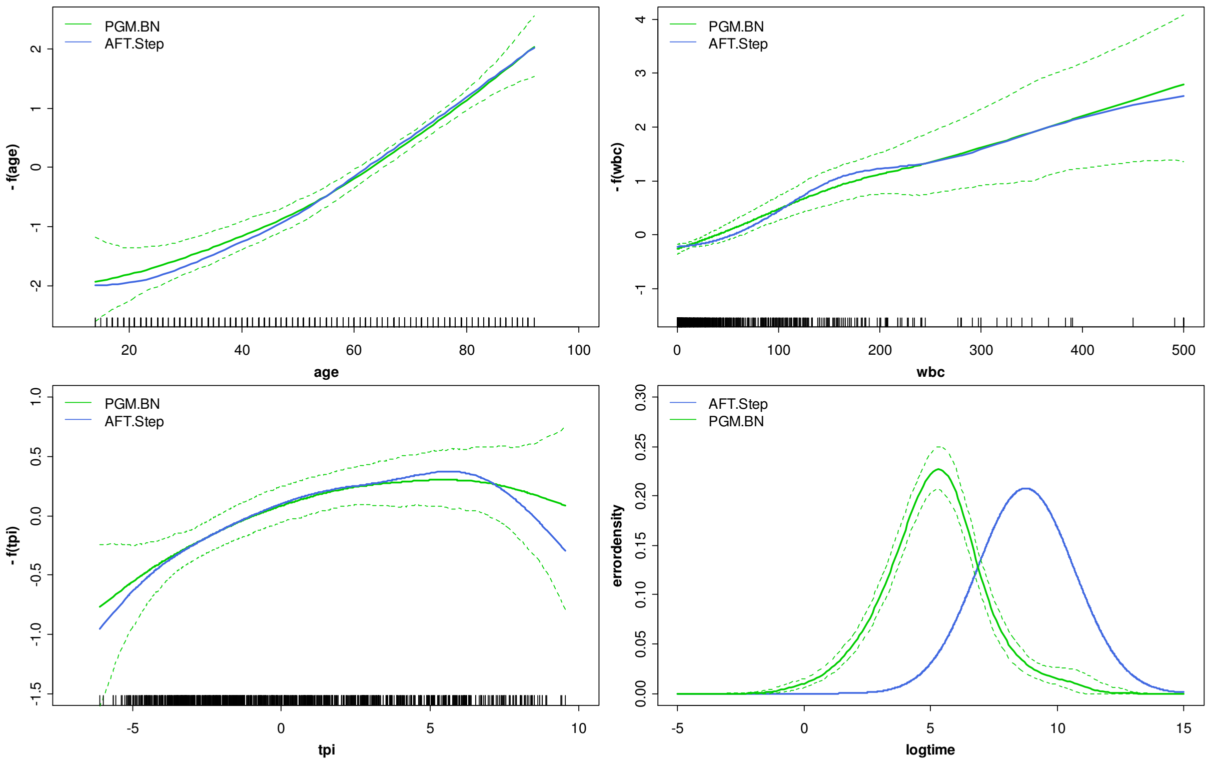


Figure 13.8: Estimates of the nonlinear effect of the covariates *age* (upper left panel), *wbc* (upper right panel) and *tpi* (lower left panel) and the log-baseline hazard (lower right panel) in the AFT model. Displayed are the estimates under the Bayesian NMIG regularization prior in the AFT model with PGM error (AFT.BN) together with those from the frequentist model under stepwise variable selection (AFT.Step). For the Bayesian models the solid lines show the posterior mean estimates and the dotted lines mark the corresponding 95 % pointwise credible bands. The stripes at the x-axis mark the observed values in the data.

14. Cytogenetically normal acute myeloid leukemia

14.1. Data

In this section we analyze data for patients diseased with cytogenetically normal acute myeloid leukemia (CN-AML). AML is a cancer of the myeloid line of blood cells which is characterized by the rapid growth of abnormal white blood cells that accumulate in the bone marrow and interfere with the production of normal blood cells. Gene expression profiling can be used to develop a gene signature that predicts the overall survival time of patients in combination with prognostic factors like molecular markers and patient characteristics. The CN-AML data was provided by U. Mansmann (IBE, Munich) and is analyzed, e. g., in Benner et al. (2010) and Metzeler et al. (2008).

The data comprises two independent cohorts of patients used as training and test cohort, where the available second test data enables a further investigation in the predictive accuracy for the applied methods. The training cohort stems from the multicenter AMLCG-199 trial of the German AML Cooperative Group between 1999 and 2003 and consists of $n = 163$ adult patients with CN-AML, where 35.0 % of the observed survival times are censored. In the training data the median survival time is 280 days with range from 0 to 2399 days. The independent test cohort consists of $n = 80$ patients who were diagnosed with CN-AML in 2004. In the test data we have a median survival time of 247.5 days with range from 1 to 837 days and 57.5 % of censored survival times. In both cohorts, survival time is defined as time from study entry until death from any cause. The original data consists of $p_x = 44757$ covariates, i. e., 44754 microarray probe sets for each individual and additional (known highly) prognostic covariates, like the *age* of the patient and the two molecular markers *FLT3* (tandem duplications of the fms-like tyrosine kinase 3) and *NPM1* (mutations in the nucleophosmin 1), are recorded. To avoid manual tuning of the regularization priors, the continuous covariates in the training and test data were standardized to have zero mean and unit variance.

14.2. Analysis

As in Metzeler et al. (2008), univariate Cox scores, measuring the correlation between each of the probe sets and the survival time in the training cohort, are used to rank and to reduce the number of probe sets. We present results based on the preselected probe sets with the 50 and 200 highest ranks of the Cox score. As additional prognostic covariates, we include the *age* of the patient and the two molecular markers *FLT3*, *NPM1* into the models, where the effect of *age* is modeled either as linear or nonlinear, utilizing P-splines. Further we fit models, where these three covariates are omitted to consider only the impact of the $p_x = 50$ and $p_x = 200$ probe sets on the patients survival time. In summary, we use the predictors

$$\eta_i = \gamma_0 + \gamma_1 \mathrm{FLT3}_i + \gamma_2 \mathrm{NPM1}_i + \gamma_3 \mathrm{age}_i + \sum\nolimits_{j=1}^{p_x} \beta_j \mathrm{probeset}_{i,j} \,, \tag{14.1}$$

$$\eta_i = \gamma_0 + \gamma_1 \mathrm{FLT3}_i + \gamma_2 \mathrm{NPM1}_i + f_1(\mathrm{age}_i) + \sum\nolimits_{j=1}^{p_x} \beta_j \mathrm{probeset}_{i,j} \,, \tag{14.2}$$

$$\eta_i = \gamma_0 + \sum\nolimits_{j=1}^{p_x} \beta_j \mathrm{probeset}_{i,j} \,, \tag{14.3}$$

$$\eta_i = \gamma_0 + \gamma_1 \mathrm{FLT3}_i + \gamma_2 \mathrm{NPM1}_i + \gamma_3 \mathrm{age}_i \,. \tag{14.4}$$

To fit the AFT and CRR models, we use the same specification of the model components and priors as for the analysis of the PBC data, compare Section 12.2. In the frequentist framework the nonlinear effect of the covariate *age* is modeled with the default settings for the `pspline()` term within the formula of the `R`-functions `survreg()` and `coxph()`. In the CRR model with the strictly linear predictors (14.1) and (14.3) we compute also the frequentist lasso and ridge estimates with the `penalized()` function.

Performance

To measure predictive accuracy, we use the time-dependent empirical Brier score $\mathrm{BS}(t)$, as proposed by Graf et al. (1999), which is defined as the time-dependent mean square error between the observed survival status and the predicted survival probability. Under random censoring the empirical Brier score is given as

$$\mathrm{BS}(t) = \frac{1}{n}\sum_{i=1}^{n}\left[\frac{\hat{S}(t \mid x_i)^2 I(t_i \le t, d_i = 1)}{\hat{G}(t_i)} + \frac{(1-\hat{S}(t \mid x_i))^2 I(t_i > t)}{\hat{G}(t)}\right],$$

where $\hat{S}(t \mid x_i)$ is the estimated survival probability of the i-th individual at time t, $I(\cdot)$ denotes the indicator function and $\hat{G}(\cdot)$ is the Kaplan-Meier estimate of the censoring distributions survival function based on the observations $(t_i, 1-d_i)$, $i = 1,...,n$. The average over time for a fixed time point $t^* > 0$ is given by integrated version of the Brier score

$$\mathrm{IBS} = \frac{1}{t^*}\int_0^{t^*} \mathrm{BS}(s)ds\,.$$

The integrated Brier score (IBS) can finally be viewed as a performance measure of the predicted survival functions within the interval $[0, t^*]$, where lower values indicate a better performance. Additionally, the IBS of a proposed prediction model can be used to derive a measure of explained variation $R^2_{\mathrm{IBS}} := 1 - \mathrm{IBS}/\mathrm{IBS}_0$, with IBS_0 defined as the integrated Brier score corresponding to the Kaplan-Meier estimate of the survival function $\hat{S}(t)$. In the CRR model, when the partial likelihood is used for inference, the estimate of the survival function is based on the Breslow estimator $\hat{\Lambda}_0^{BR}(t)$ for the cumulative baseline hazard, i. e. $\hat{S}(t\,|\,\mathbf{x}_i) = \exp(-\hat{\Lambda}(t\,|\,\mathbf{x}_i))$, with $\hat{\Lambda}(t\,|\,\mathbf{x}_i) = \hat{\Lambda}_0^{BR}(t)\exp(\hat{\eta}_i)$, where $\hat{\eta}_i$ denotes the estimated linear predictor. In the case of the full likelihood we utilize the trapezoidal rule to compute the cumulative baseline hazard from the estimate of the log-baseline hazard function.

14.3. Results

For a selection of the estimated models the IBSs in the training data ($t^* = 2399$ days) and test data ($t^* = 837$ days) are visualized in the following figures. All figures show the results achieved with the full predictor and with the reduced predictor, if variable selection is applied. The number of probe sets included in the final models are displayed at the bottom of the bars. As competitors the IBSs of the Kaplan-Meier estimate are marked (KM: $\mathrm{IBS}_{\mathrm{train}} = 0.212$ and $\mathrm{IBS}_{\mathrm{test}} = 0.205$), together with the resulting IBSs for the unregularized frequentist models using the predictor (14.4) which includes only the three pheno-covariates (CPL3: $\mathrm{IBS}_{\mathrm{train}} = 0.171$ and $\mathrm{IBS}_{\mathrm{test}} = 0.190$, AFT3 with Gaussian error: $\mathrm{IBS}_{\mathrm{train}} = 0.178$ and $\mathrm{IBS}_{\mathrm{test}} = 0.192$). The penalized lasso (CPL.PenL) applied to the complete number of probe sets, $p_x = 44754$, using the purely linear predictor (14.1), is leading to a final CRR model with 15 selected probe sets and to the integrated Brier scores $\mathrm{IBS}_{\mathrm{train}} = 0.138$ and $\mathrm{IBS}_{\mathrm{test}} = 0.182$, both marked as further benchmarks by black bars at the y-axis in the figures.

Integrated Brier score with 50 probe sets

The IBS-results, based on the reduced data with $p_x = 50$ preselected probe sets and predictor (14.1), where the pheno-covariate *age* is modeled linear, are shown in **Figure 14.1**. **Figure 14.2** shows the corresponding results using predictor (14.2), where the covariate *age* is modeled as P-spline and finally **Figure 14.3** shows in terms of the CRR model the IBS-results with predictor (14.3) without the pheno-variables.

We consider at first the results with the purely linear predictor in **Figure 14.1**. Compared to the models from partial likelihood approach, the flexible modeling of the log-baseline hazard leads to models with better performance in the trainings data, but the higher adaptiveness to the trainings data causes a loss of performance in the test data. In the test data the IBSs of the full likelihood based models (CFL) are larger than those of the associated partial likelihood based models (CPL) and frequently exceed the reference IBS of the CPL3 model. Also the unregularized models (CPL, CPL.Step, CPL.B, CFL.B) achieve comparatively large IBS values. The best predictive performances are obtained with the CPL models in combination with the ridge penalty (CPL.PenR, CPL.BR), closely followed by the models with the lasso penalty (CPL.PenL, CPL.BL). The Bayesian approaches achieve marginally smaller IBSs than their frequentist counterparts and the lowest value is obtained under the Bayesian ridge prior (CPL.BR: $\text{IBS}_{\text{test}} = 0.168$). Variable selection causes a loss of performance as indicated by the results from the stepwise procedure, the frequentist lasso and the hard shrinkage selection rules. Hard shrinkage via the standard deviation criterion is leading to sparse models, including 6 probe sets in the final Bayesian ridge model (CPL.BR-HS.STD: $\text{IBS}_{\text{test}} = 0.172$) and 5 probe sets in the final Bayesian lasso model (CPL.BL-HS.STD: $\text{IBS}_{\text{test}} = 0.178$), with only a marginal loss in the predictive performance. Both values are smaller than the IBS of the sparse penalized lasso model, $\text{IBS}_{\text{test}} = 0.182$, that selects the 15 covariates from the complete probe set $p_x = 44757$. Using the AFT model for inference, right panel of **Figure 14.1**, the best performances are achieved under the assumption of a Gaussian error distribution, but the differences to the results with the PGM error are less pronounced as differences between the CPL and CFL model. Comparing the estimates for the baseline error under the Gaussian and PGM error assumption, see **Figure 14.9,** shows also a high similarity. As in the CRR model the best performances are obtained with the ridge prior (AFT.BR: $\text{IBS}_{\text{test}} = 0.175$) and the lasso prior (AFT.BL: $\text{IBS}_{\text{test}} = 0.176$), if all probe sets are included in the predictor. Variable selection increases the IBS and we get $\text{IBS}_{\text{test}} = 0.184$ if the HS.STD criterion is applied in the Bayesian ridge model (AFT.BR-HS.SDT). The associated final model includes, amongst others, the probe sets with Cox score ranks 11, 12 and 21 and these probe sets occur in almost all predictors of the sparse CRR and AFT models.

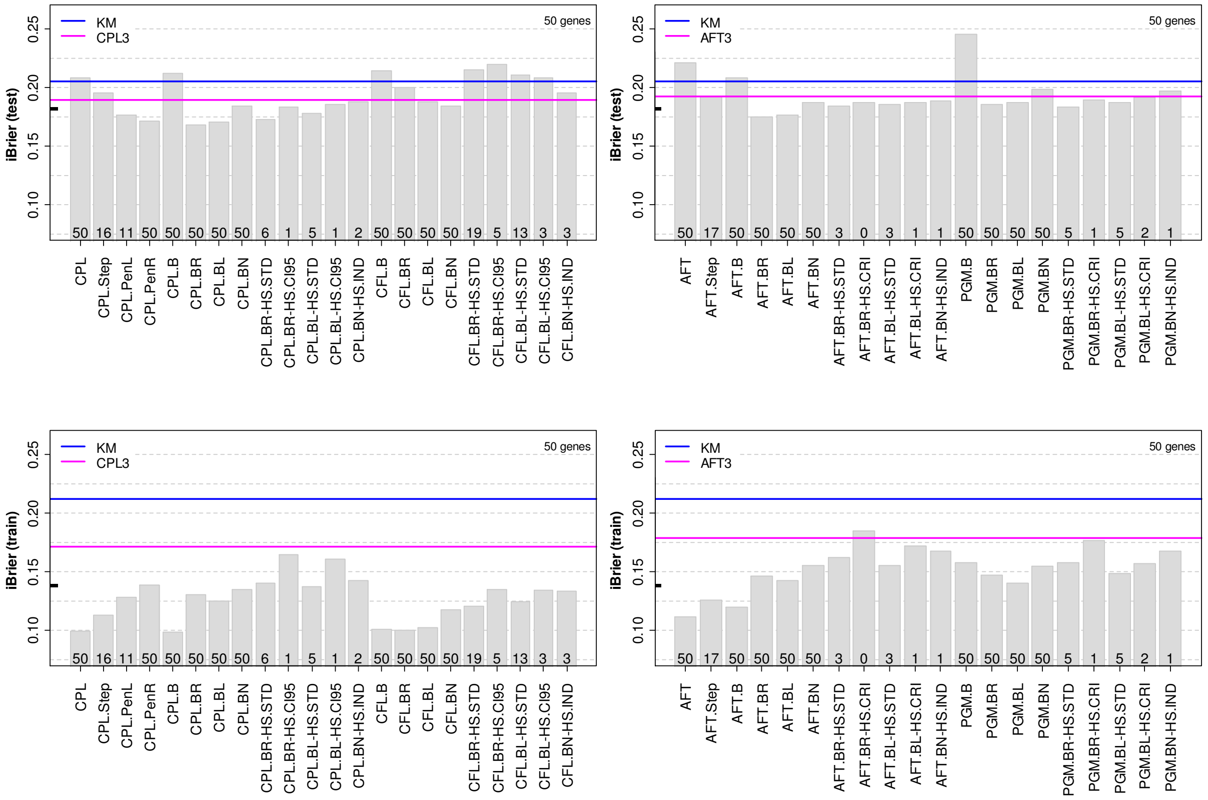


Figure 14.1: Integrated Brier scores in the test data (upper panel) and the trainings data (lower panel) for the CRR model (left panel) and the AFT model (right panel) with predictor (14.1) using $p_x = 50$ preselected probe sets. The blue horizontal line marks the IBS of the Kaplan-Meier-Estimate ($\text{IBS}_{\text{train}} = 0.212$ and $\text{IBS}_{\text{test}} = 0.205$). The magenta horizontal line marks the IBS of the frequentist CRR model ($\text{IBS}_{\text{train}} = 0.171$ and $\text{IBS}_{\text{test}} = 0.190$) and Gaussian AFT model ($\text{IBS}_{\text{train}} = 0.178$ and $\text{IBS}_{\text{test}} = 0.192$) with predictor (14.4). The black bar at the y-axis marks the IBS from the frequentist lasso using $p_x = 44754$ probe sets ($\text{IBS}_{\text{train}} = 0.138$ and $\text{IBS}_{\text{test}} = 0.182$). The associated numbers of covariates included in the estimated predictor are displayed at the bottom of the bars.

We found further that the nonlinear modeling of the covariate *age* does not clearly improve the IBS performance in the test data, compare **Figure 14.2,** and also the visual inspection of the nonlinear estimate, **Figure 14.8**, shows no strong evidence to model this covariate effect as nonlinear. For example with the ridge prior we the get the values $\text{IBS}_{\text{test}} = 0.169$ (CPL.BR) and $\text{IBS}_{\text{test}} = 0.174$ (AFT.BR), the application of the HS.STD selection criterion is leading to $\text{IBS}_{\text{test}} = 0.176$ (CPL.BR-HS.SDT) and $\text{IBS}_{\text{test}} = 0.183$ (AFT.BR-HS.SDT). These values are almost comparable to the IBS values obtained with the strictly linear predictor. In addition, with the nonlinear modeling the IBSs of the full likelihood based CRR models and the AFT models with PGM error are further increased.

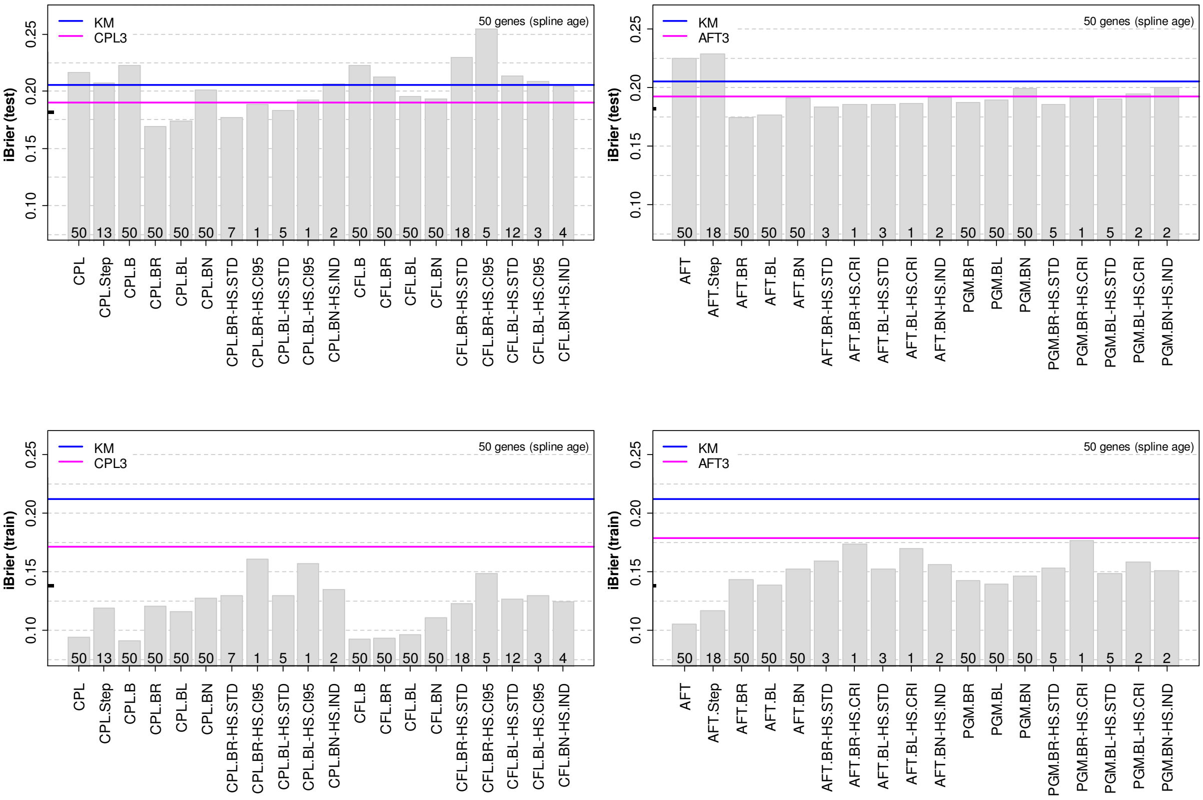


Figure 14.2: Integrated Brier scores in the test data (upper panel) and the trainings data (lower panel) for the CRR model (left panel) and the AFT model (right panel) with predictor (14.2) using $p_x = 50$ preselected probe sets. The blue horizontal line marks the IBS of the Kaplan-Meier-Estimate ($\text{IBS}_{\text{train}} = 0.212$ and $\text{IBS}_{\text{test}} = 0.205$). The magenta horizontal line marks the IBS of the frequentist CRR model ($\text{IBS}_{\text{train}} = 0.171$ and $\text{IBS}_{\text{test}} = 0.190$) and Gaussian AFT model ($\text{IBS}_{\text{train}} = 0.178$ and $\text{IBS}_{\text{test}} = 0.192$) with predictor (14.4). The black bar at the y-axis marks the IBS from the frequentist lasso using $p_x = 44754$ probe sets ($\text{IBS}_{\text{train}} = 0.138$ and $\text{IBS}_{\text{test}} = 0.182$). The associated numbers of covariates included in the estimated predictor are displayed at the bottom of the bars.

In summary both covariate groups, the clinical covariates as well as the microarray features, separately influence the predictive performance. The models including only the three unregularized pheno-covariates (magenta lines) have a decreased IBS compared to the IBS of the Kaplan-Meier estimate (blue lines), while the microarray features need to be regularized to enhance the predictive performance. **Figure 14.3** shows the IBSs of the CRR models with predictor (14.3) that includes only the probe sets. Therein the regularized estimates provide models with an improved performance compared to the model with only the three pheno-variables (CPL3), but the best predictive performances are obtained if both covariate groups are commonly included in combination with a ridge or lasso type shrinkage of the probe sets. With a reduced number of $p_x = 50$ preselected microarray features we are able to find models with a similar or improved predictive performance compared to the frequentist lasso that searches the final model within all microarray features. The increased flexibility resp. model complexity introduced by the P-spline model for the covariate *age* or the P-spline model for the baseline quantity does not enhance the predictive performance. And finally with respect to the results in the trainings data, where the CRR model indicates a higher adaptivity to the data as the AFT model, we would rather use the CRR model to reflect the impact of the covariates to the patient's survival.

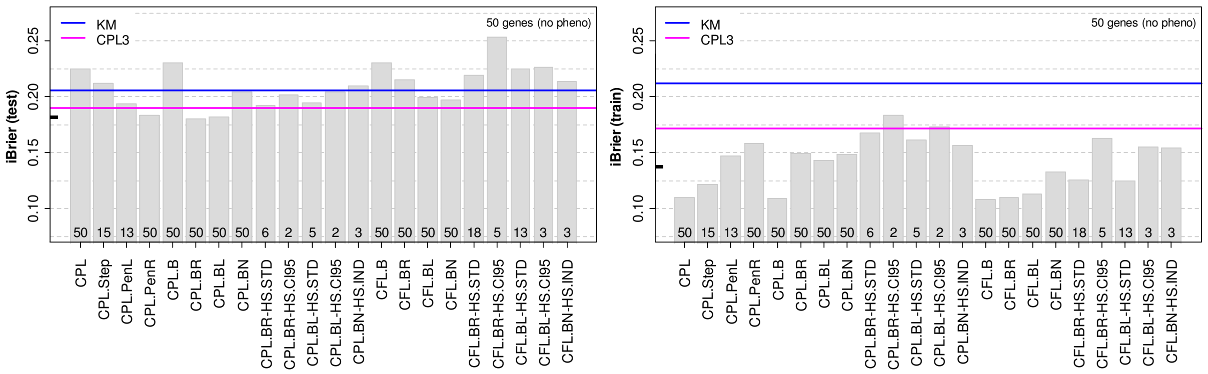


Figure 14.3: Integrated Brier scores in the test data (left side) and the trainings data (right) for the CRR model with predictor (14.3) and $p_x = 50$ preselected probe sets. The blue horizontal line marks the IBS of the Kaplan-Meier-Estimate ($IBS_{train} = 0.212$ and $IBS_{test} = 0.205$). The magenta horizontal line marks the IBS of the frequentist CRR model ($IBS_{train} = 0.171$ and $IBS_{test} = 0.190$). The associated numbers of covariates included in the estimated predictor are displayed at the bottom of the bars.

Integrated Brier score with 200 probe sets

Figure 14.4 summarizes the results when inference is based on predictor (14.1) with *age* modeled as linear and the increased number of $p_x = 200$ preselected probe sets. The left and right panel show respectively the results for the CRR model with the partial likelihood and the AFT model with Gaussian error. Due to the poor performance of the models with higher complexity, i. e. the CRR model with P-spline log-baseline hazard and the AFT model with PGM error, the results are omitted. Compared to the models with $p_x = 50$ probe sets we obtain an improvement in the predictive performance for the models CPL.PenR, AFT.BR, AFT.BR-HS.STD and AFT.BL, where the best performance results for the frequentist ridge model CPL.PenR ($IBS_{test} = 0.159$) with all 200 probe sets included in the final predictor.

In the CRR model the predictive performance of the Bayesian models with the lasso and ridge prior, separately or combined with the HS.STD selection criterion, is almost comparable to the associated models with $p_x = 50$ preselected probe sets. The IBS from the NMIG prior exceeds now the IBS of the Kaplan-Meier estimate. We can also report a loss in the predictive performance for the final model from the frequentist lasso procedure, the IBS of the model CPL.PenL is in the range of the model CPL3 that merely includes the three pheno-covariates in the predictor. In the AFT model the predictive performance of the models AFT.BR ($IBS_{test} = 0.164$), AFT.BL ($IBS_{test} = 0.164$) is clearly improved and now almost comparable to the model CPL.BR ($IBS_{test} = 0.162$). The final predictor of the model AFT.BR-HS.STD shares eight probe sets with the predictor from CPL.BR-HS.STD and contains also the three probe sets (Cox score ranks 11, 12, 21) from the associated model based on

$p_x = 50$ preselected probe sets. The models CPL.BR-HS.STD with $p_x = 50$ and $p_x = 200$ share two probe sets (Cox score ranks 11, 12).

In summary, with an increased number of probe sets we obtain for some models an improvement in the predictive performance. The models with the best performances include again all of the $p_x = 200$ preselected probe sets. With the hard shrinkage selection rule HS.STD we find sparse models with comparable performance to the models with $p_x = 50$ preselected probe sets, but the included probe sets in the final models differ and expert knowledge is required for the interpretation. The probe sets with ranks 11 and 12 are also almost always included in the final spares models.

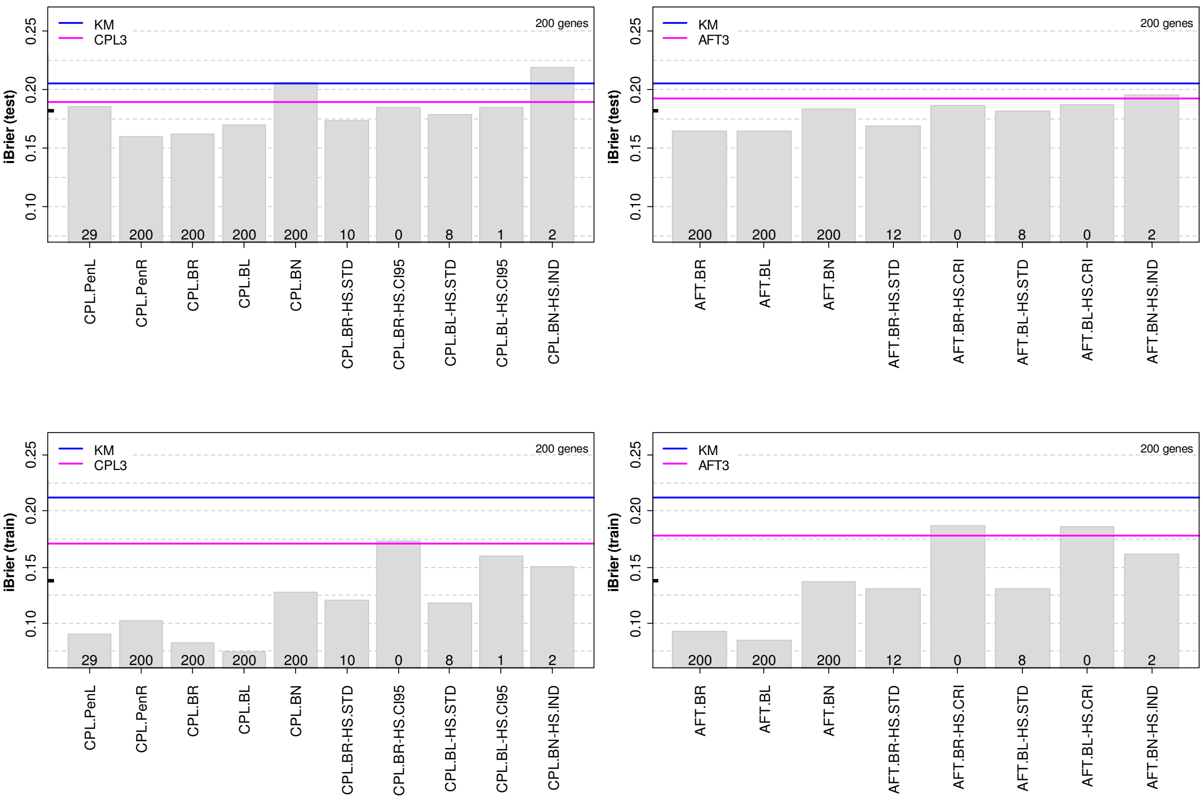


Figure 14.4: Integrated Brier scores (IBS) in the test data (upper panel) and the trainings data (lower panel) for the CRR model (left panel) and the AFT model (right panel) with predictor (14.1) using $p_x = 200$ preselected probe sets. The blue horizontal line marks the IBS of the Kaplan-Meier-Estimate ($\text{IBS}_{\text{train}} = 0.212$ and $\text{IBS}_{\text{test}} = 0.205$). The magenta horizontal line marks the IBS of the frequentist CRR model ($\text{IBS}_{\text{train}} = 0.171$ and $\text{IBS}_{\text{test}} = 0.190$) and Gaussian AFT model ($\text{IBS}_{\text{train}} = 0.178$ and $\text{IBS}_{\text{test}} = 0.192$) with predictor (14.4). The black bar at the y-axis marks the IBS from the frequentist lasso using $p_x = 44754$ probe sets ($\text{IBS}_{\text{train}} = 0.138$ and $\text{IBS}_{\text{test}} = 0.182$). The associated numbers of covariates included in the estimated predictor are displayed at the bottom of the bars.

Paths of the integrated Brier score

As further illustration, the upper panel **Figure 14.5** displays the impact of the regularization parameter on the IBS in the test data by means of the CRR model with predictor (14.1) using $p_x = 50$ preselected probe sets. Shown are the IBS-paths under the Bayesian and frequentist lasso and ridge regularization (upper left and right panel) and the Bayesian NMIG regularization (lower left panel). As further competitor the results from the frequentist CPL model and the stepwise selection are shown for increasing number of covariates in the predictor (lower right panel). The horizontal lines mark the

$\text{IBS}_{\text{test}} = 0.190$ of the frequentist Cox model with predictor (14.4) including only the three pheno-covariates (CPL3).

The first observation, that we can make from these figures is, that again the partial likelihood based models with the ridge regularization of the probe sets leads to the smallest IBS values, and the IBS path is also fairly insensitive with respect to the regularization parameter over a wide range of the shrinkage parameter. In addition, the behavior of the frequentist ridge IBS-path (CPL.PenR) is quite close to the Bayesian version (CPL.BR). With increased value of the shrinkage parameter λ the penalization of the probe sets is increased, which obviously improves the performance of the resulting models. The strongest impact on the increase of the performance is observable in the range $0 < \lambda < 20$, while larger values $\lambda > 20$ increase the performance only marginally. Also the sparse models obtained with the HS.STD criterion show the same development of the IBS-path, but on a higher level of the IBS. Nevertheless, the path is clearly below the CPL3 benchmark for higher values of the shrinkage parameter. For ridge regression in combination with the full likelihood (CFL.BR), there seems to be some instability in estimation for $\lambda > 5$ that yields abrupt changes in the IBS even for small variations of the regularization parameter.

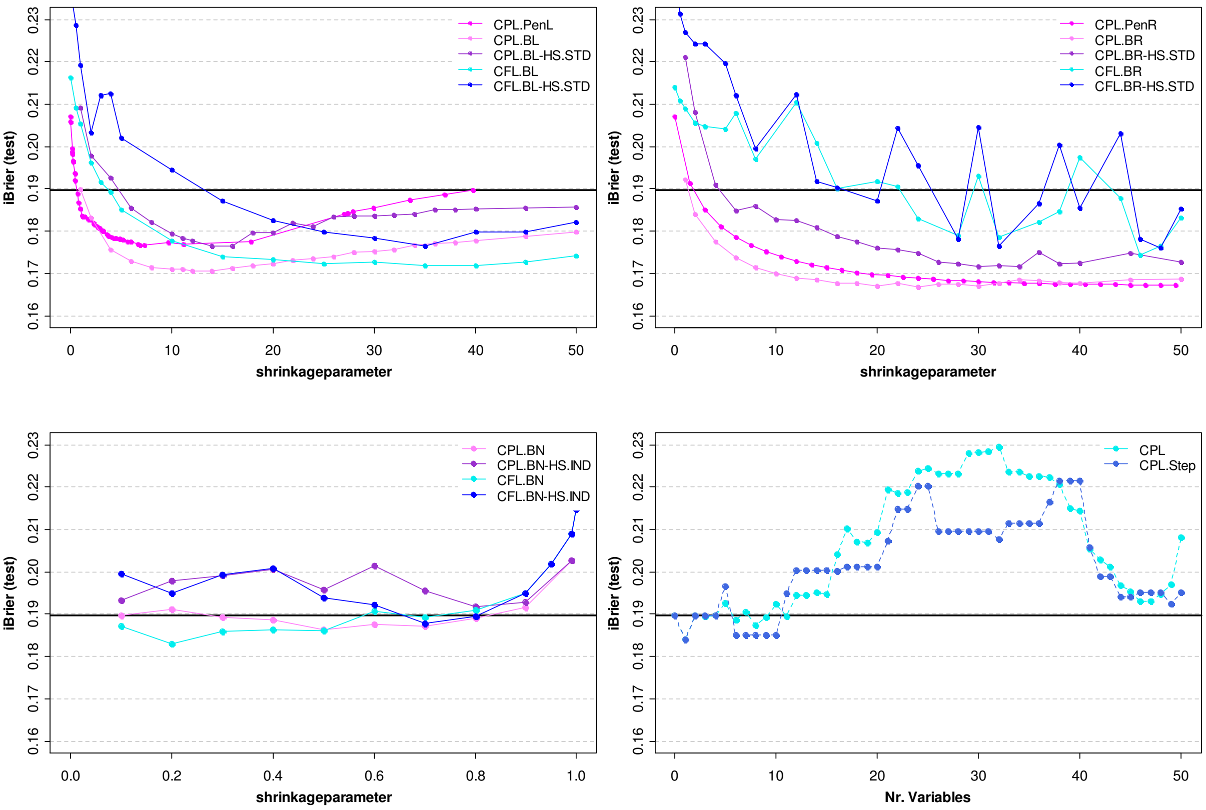


Figure 14.5: Paths of the integrated Brier scores for varying shrinkage parameter in the CRR model with predictor (14.1) using $p_x = 50$ preselected probe sets. Upper panel: Lasso type regularization (left side) and ridge type regularization (right side). Lower panel: Bayesian NMIG regularization (left side) and frequentist CRR model with and without stepwise selection (right side) for increasing number of covariates in the predictor. The black solid horizontal line marks the $\text{IBS}_{\text{test}} = 0.190$ of the frequentist CRR model with predictor (14.4) including only the 3 pheno-covariates.

In general, the CFL and CPL lasso variants are in closer agreement and the CPL lasso-path does not show such an irregular behavior as under the ridge prior. Again, the Bayesian lasso estimates based on the partial likelihood (CPL.BL) performs remarkably well, but the full likelihood based estimates (CPL.BL) are close and yield an improved performance for large regularization parameters λ. The IBSs achieved with the frequentist lasso, which always selects a subset from the 50 probe sets, do not reach the low values possible with the Bayesian lasso or the (Bayesian and frequentist) ridge estimates that include all 50 preselected probe sets in the predictor. Variable selection via the hard shrinkage criterion HS.STD does not globally improve the predictive performance, but provides, dependent on the particular value of the shrinkage parameter, sparse models with comparably good performance. In contrast to the ridge-paths the predictive performance of the lasso-paths starts to decrease for larger values of the shrinkage parameter. For the frequentist lasso this is easy to understand, because with increased shrinkage parameter at least only the three pheno-covariates are included in the final model, and the IBS-path must converge to the CPL3 benchmark. The Bayesian lasso-paths indicate a similar behavior, but the decrease of the performance is less pronounced in the plotted range of the shrinkage parameter. But, at the limit $\lambda \to \infty$ or $\omega \to 0$ all regularized IBS-paths converge to CPL3 benchmark.

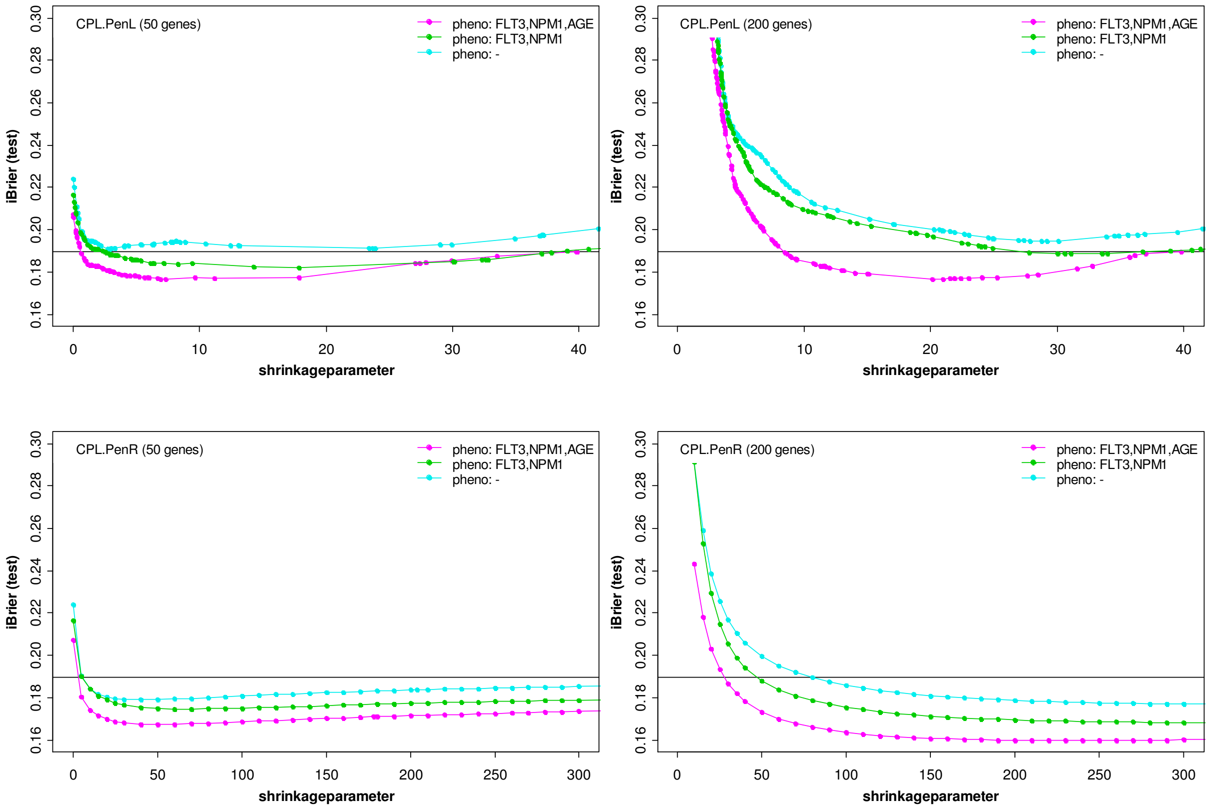


Figure 14.6: Paths of the integrated Brier scores for varying shrinkage parameter under the frequentist lasso (upper panel) and ridge (lower panel) regularized model using $p_x = 50$ (left panel) and $p_x = 200$ (right panel) preselected probe sets. The upper right legend identifies the additional unregularized pheno-variables in the predictor. The black solid horizontal line marks the $IBS_{test} = 0.190$ of the frequentist CRR model with predictor (14.4) including only the 3 pheno-covariates. The paths of the penalized lasso are evaluated at the values of the shrinkage parameter, where a covariate is removed from the predictor.

The **lower left panel in Figure 14.5** shows the results achieved with the NMIG prior structure and the lower right panel shows the IBS path under the frequentist CRR model and combined stepwise selection, when the number of probe sets p_x included in the predictor increases according to the ranks of the Cox scores. The reduced shrinkage of some influential regression coefficients under the

Bayesian NMIG prior, compare **Figure 14.7,** leads to models with similar poor performance values as the models obtained with the frequentist CPL and CPL.Step procedure. None of these methods enhance the performance to the IBS-levels obtained under the lasso or ridge regularization.

Finally, **Figure 14.6** shows the movement of the frequentist lasso (CPL.PenL) and ridge (CPL.PenR) IBS-paths for increasing values of the shrinkage parameter if the number of pheno-variables in the predictor are reduced. The left and right panels display the results with $p_x = 50$ and $p_x = 200$ probe sets in the regularized predictor component. We obtain from this figure, that the inclusion of all three pheno-covariates is important to improve the predictive performance of the estimated models. From the lower left panel, where the CPL.PenR path is plotted over a wider range of the shrinkage parameter as in **Figure 14.5,** we see, that similar to the lasso regularization, also the performance under the ridge regularization decreases for larger values of the shrinkage parameter, where the regularized estimates become closer to zero.

Paths of the regression coefficients

Figure 14.7 illustrates the associated paths of the estimated regression coefficients as a function of the regularization parameter for the four selected probe sets with Cox score ranks 1, 11, 12, and 21. We show the results from the CRR model with ridge, lasso and NMIG regularization of the $p_x = 50$ covariates in predictor (14.4). The vertical lines in the figure mark the estimated regression coefficients at the estimated value of the complexity parameter.

It turns out that especially the probe sets associated with Cox score ranks 11 and 12 yield overall larger estimates and are therefore deemed to be important over wide ranges of the complexity parameter values. As previously mentioned, both probe sets are almost always included in the final models resulting from any used selection method applied to the data with $p_x = 50$ and $p_x = 200$ preselected probe sets. Especially, in the case of the Bayesian and frequentist ridge model (upper panel), the estimated effects at the estimated shrinkage parameter $\hat{\lambda}$ have smaller values compared the Bayesian and frequentist lasso (middle panel) and Bayesian NMIG estimates (lower left panel) and in particular the effects of the important probe sets with Cox score ranks 11 and 12 are clearly smaller. With respect to the previous IBS results, particularly the uniform shrinkage of all effects, as under the ridge penalty, seem to improve predictive performance.

The lower right panel of **Figure 14.7** shows the posterior relative frequencies of the binary NMIG indicator $I_j = v_1$ as a function of the complexity parameter. Here, the inclusion probabilities of the two variables with Cox score ranks 1 and 21 rapidly decrease, if the complexity parameter moves towards zero in direction with reduced model complexity. In contrast, the estimated inclusion probabilities for the probe sets with Cox score ranks 11 and 12 do not vary very much in the plotted range of the complexity parameter and always yield the conclusion that these probe sets should be contained in the final model. But, higher inclusion probabilities cause a reduction of the shrinkage, compare lower left panel of **Figure 14.7**, and this may be the main reason why the NMIG models have such a poor performance. Due to the results from the ridge regression, the best performances are obtained if all covariates, also the important ones, are uniformly shrunken with the same proportion, and this is definitely not the case under the NMIG prior (as designed).

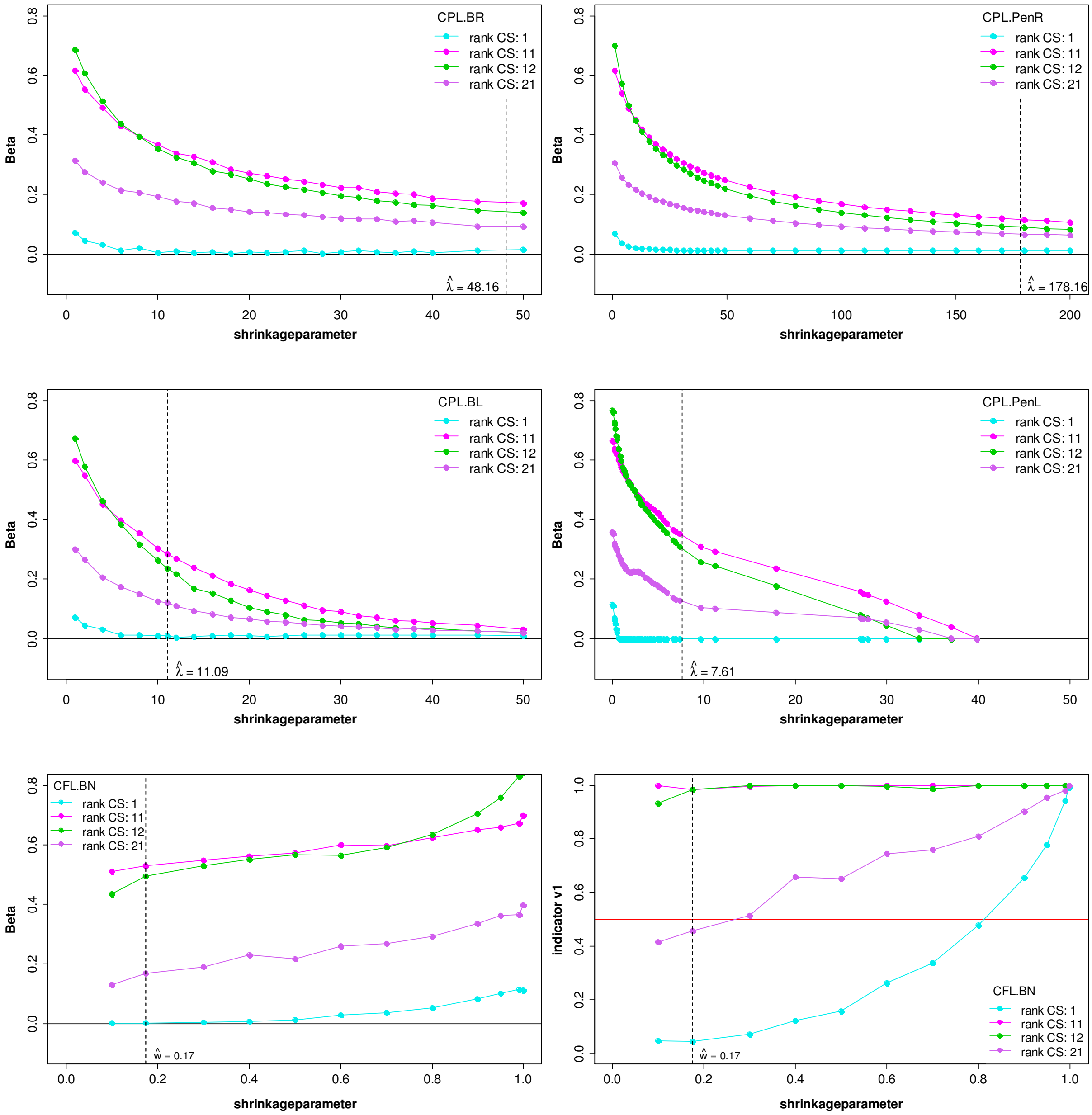


Figure 14.7: Shrinkage parameter dependent paths of four selected estimated regression coefficients and NMIG inclusion probabilities in the CRR model based on the full and partial likelihood with predictor (14.1) using $p_x = 50$ preselected probe sets. The four probe sets are identified by the rank of the Cox score (rank CS) in the legend of the last figure. Upper panel: Frequentist ridge (left side) and Bayesian ridge (right side) regularization. Middle panel: Frequentist lasso (left side) and Bayesian lasso (right side) regularization. Lower panel: Bayesian NMIG regularization (left side) and posterior inclusion probabilities based on the relative frequencies of Bayesian NMIG indicator variable value $I_j = v_1$ (right side), where the horizontal red line marks the 0.5 cut off value for variable selection. The vertical dashed lines in the figures mark the estimated values at the estimated shrinkage parameter.

Nonlinear effects

Figure 14.8 presents the estimated nonlinear effect of the covariate *age* with predictor (14.2) under the CRR and AFT model. While the spline estimates in both survival model classes do show some nonlinearity, the associated credible intervals all cover the linear effect so that there is only weak evidence for the necessity of a nonlinear modeling. We have also seen from **Figure 14.2** that the spline modeling of the *age* leads indeed to a better fit in the training data, but there is no remarkable benefit for the predictive performance in the test data.

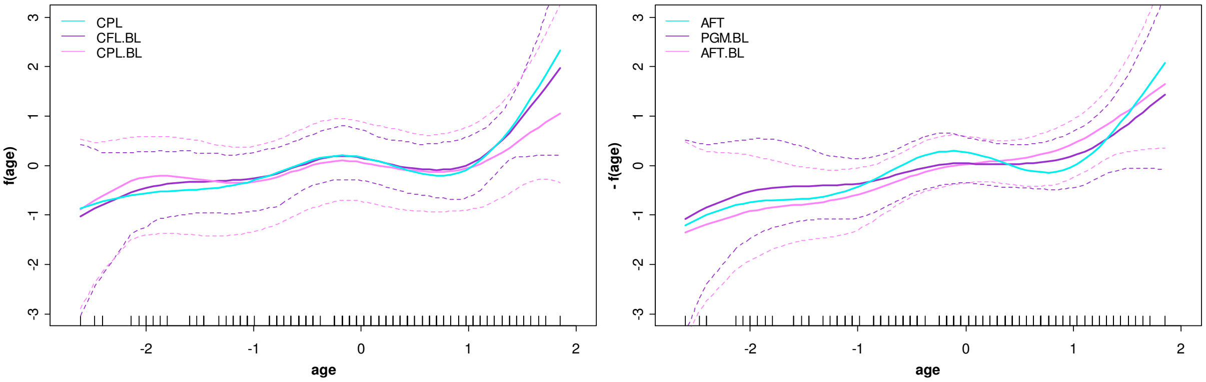


Figure 14.8: Estimation of the nonlinear *age* effect the CRR model (left side) and the AFT model (right side) with predictor (14.2) using $p_x = 50$ preselected probe sets. Left side: Estimations from the frequentist CRR model and Bayesian lasso regularized model based on the full and partial likelihood. Right side: Estimations from the frequentist Gaussian AFT model and the Bayesian Lasso regularized AFT model with Gaussian and PGM error. For the Bayesian models the solid lines show the posterior mean estimates and the dotted lines mark the corresponding 95 % pointwise credible bands. The black stripes at the x-axis mark the observed values.

Baseline quantities

In the following we shortly summarize the estimated components of the CRR and AFT survival model with strictly linear predictor (14.1) using the $p_x = 50$ preselected probe sets under the various approaches. The upper left side of **Figure 14.9** displays the estimated log-baseline hazard rate in the CRR model obtained with the full likelihood. After a short period of constant or moderately increasing (or bathtub shaped) log-baseline hazard in the first 300 days, the hazard rate shows afterwards an almost linear decrease, so there seems to be an enhanced risk to die in the last quarter of the first year. The corresponding cumulative baseline hazards in comparison to the Breslow estimators based on the partial likelihood are shown at the upper right side of **Figure 14.9**. In the first year period, that contains most of the observations, the estimates from the full and partial likelihood closely coincide.

In the simulations we have seen that with low sample sizes it is hard to detect variations in the shape of the baseline error density. Comparing the estimated baseline error densities achieved with the AFT model assuming a Gaussian and a PGM error, lower panel of **Figure 14.9,** we observe a minor asymmetry with the PGM estimate and in summary a Gaussian error seems to be a good proxy for the underlying baseline error distribution also with respect to the enhanced predictive performance and the low sample size.

Linear effects

In **Figure 14.10** a selection of the estimated probe set effects in the CRR and AFT model are displayed. With exception of the probe sets with Cox score ranks 1 and 5 we show the results for those probe sets that frequently appear in the final models after variable selection. As previously observed in the simulation section, with comparable prior tuning the estimates in the CRR model based on the full likelihood are less regularized as those under the partial likelihood. This also explains the fall off in the predictive performance compared to the partial likelihood estimates, because with stronger regularization the performance of the full likelihood estimates increases, compare **Figure 14.5**.

NMIG indicators

Under the adaptive Bayesian NMIG prior structure most of the $p_x = 50$ probe sets in the predictor (14.1) are assigned to the close to zero component of the bimodal prior and are therefore strongly regularized. **Figure 14.11** shows the posterior relative frequencies of the binary NMIG indicator value $I_j = v_1$ in the CRR and AFT model. In the figures the probe sets with Cox score ranks 11, 12, 21, 41 and 46 stand out with larger posterior inclusion probabilities, but only the probe set with Cox score rank 11 achieves commonly an estimated inclusion probability clearly exceeding the threshold of 0.5. As a consequence the additional hard shrinkage variable selection based on the HS.IND-threshold 0.5 is leading to very sparse models with an IBS in the test data close to the IBS of the models AFT3 and CPL3, that include only the three pheno-covariates in the predictor. Due to the reduced shrinkage of the important probe set with rank 11, the adaptation of the threshold, to force more probe sets to the final model, should not improve the predictive performance.

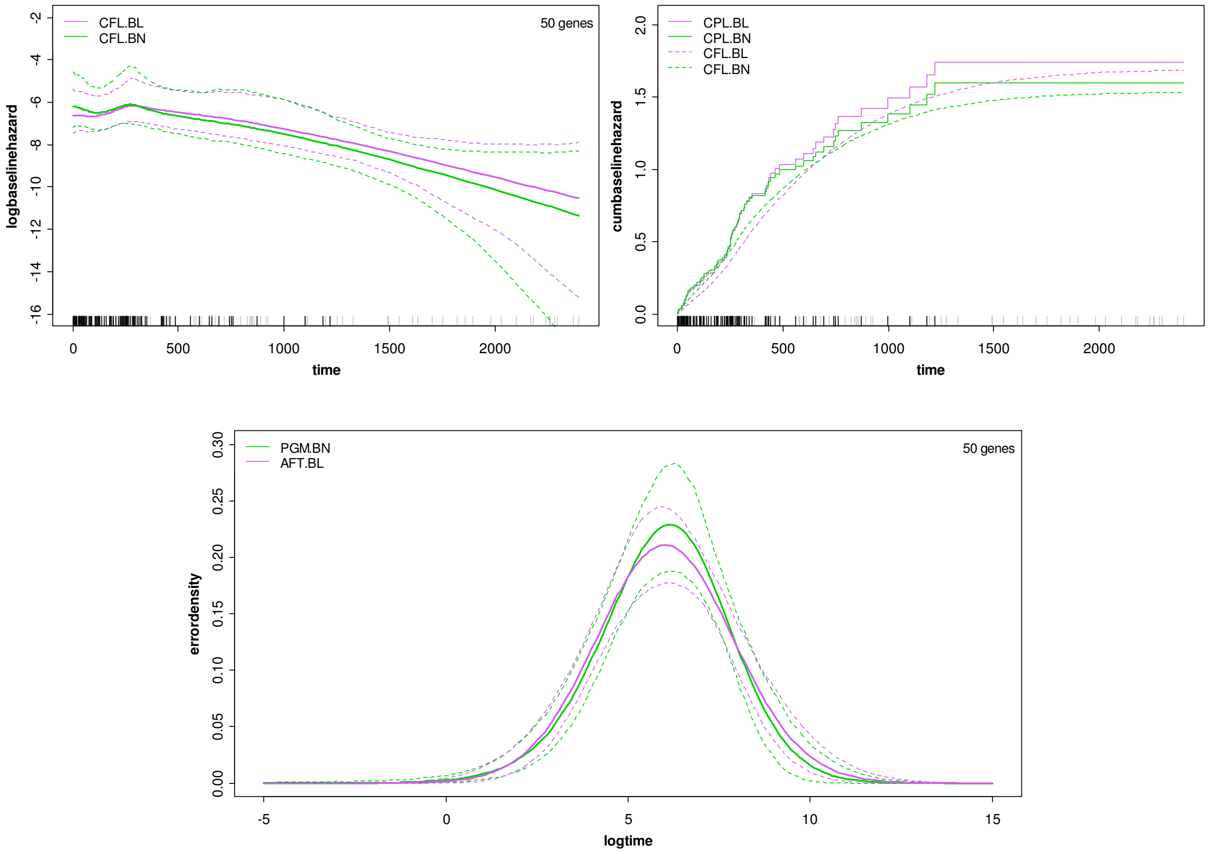


Figure 14.9: Estimation of the baseline quantities in the CRR model (upper panel) and the AFT model (lower panel) with predictor (14.1) using $p_x = 50$ preselected probe sets under the Bayesian lasso and NMIG regularization of the linear effects. Upper left side: Estimates of the log-baseline hazard from the CRR model based on the full likelihood. Upper right side: Estimation of the cumulative baseline hazard in the CRR model. The dashed lines display the estimate from the full likelihood when the cumulative baseline hazard is computed via numerical integration of the baseline hazard using the trapezoidal rule. Lower panel: Estimates from the AFT model with lasso regularization and Gaussian error and Bayesian NMIG regularization with PGM error. Commonly the solid lines show the posterior mean estimates and the dotted lines mark the corresponding 95 % pointwise credible bands.

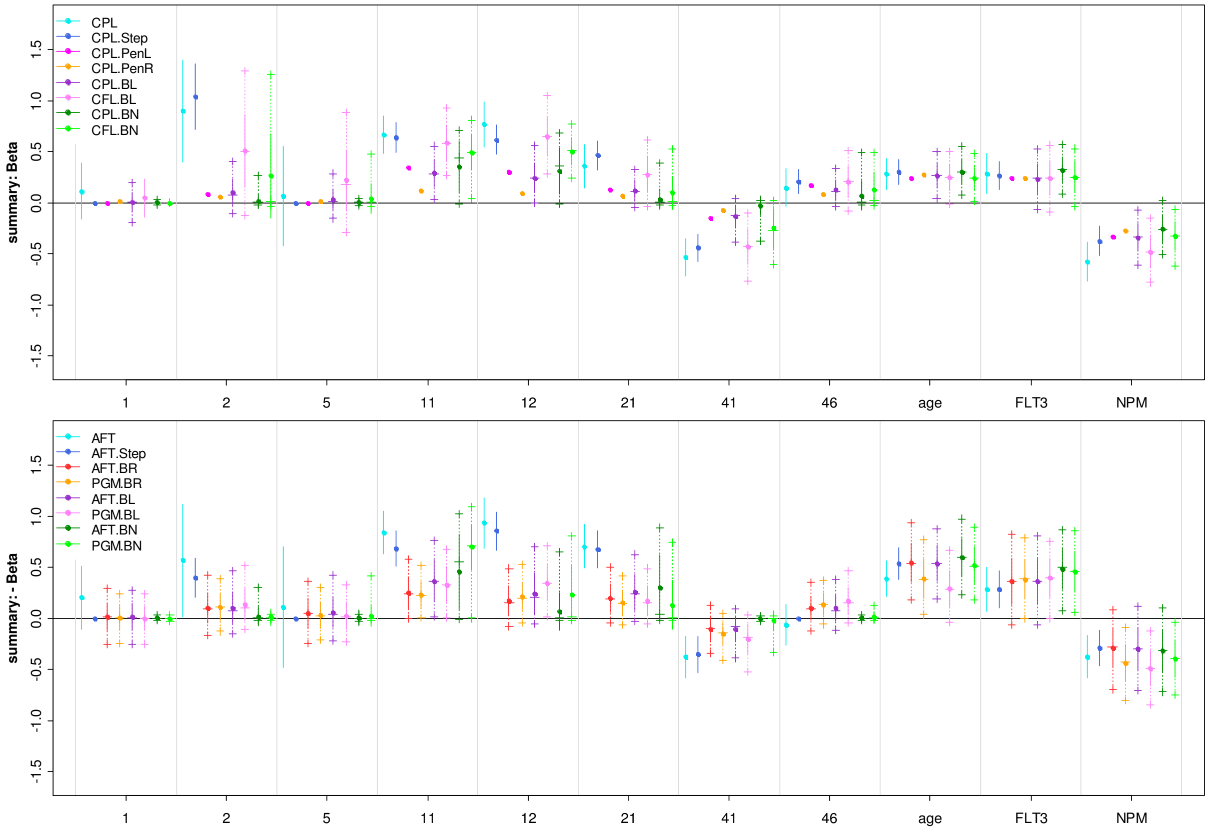


Figure 14.10: Selected estimated regression coefficients in the CRR model (upper panel) and AFT model (lower panel) with predictor (14.1) using $p_x = 50$ preselected probe sets. The probe sets are sorted according to the rank of the Cox scores that are displayed at the x-axis. The points mark the estimates of the regression coefficients and the lines display the corresponding standard errors. Additional for the Bayesian procedures the dashes mark the median and the 95 % quantiles of the marginal posterior distribution of the regression coefficients.

Final remark

In summary it becomes apparent that with increasing model complexity, formed by using flexible baseline quantities or nonlinear effects, the adaptation in the training data increases, but the estimated models loose their predictive performance in the test data. Besides a compromise regarding the model complexity, the IBSs seem to suggest that the best strategy to achieve precise predictions is to include all covariates without variable selection in the predictor, but to apply stronger regularization to the regression coefficient vector. This claim is further supported by the results from the NMIG prior structure, with the sophisticated selection-like shrinkage of small and large effects, which leads to somewhat deteriorated IBSs, and the IBS results for the sparse models under the frequentist lasso compared to the Bayesian counterpart that includes all covariates in the predictor. Nevertheless, also variable selection for the Bayesian lasso and ridge regularized estimates, based on the standard deviation (HS.STD) criterion, is leading to final models with comparably good predictive performance, if one is willing to accept a little loss in the predictive performance for the benefit of a sparse predictive model.

In our analyses we found strong evidence for the importance of at least two probe sets associated with Cox score ranks 11 and 12. Finally, our flexible model classes allow us to validate the assumption of linearity of pheno-covariates that are often available in addition to genetic information, but we did not

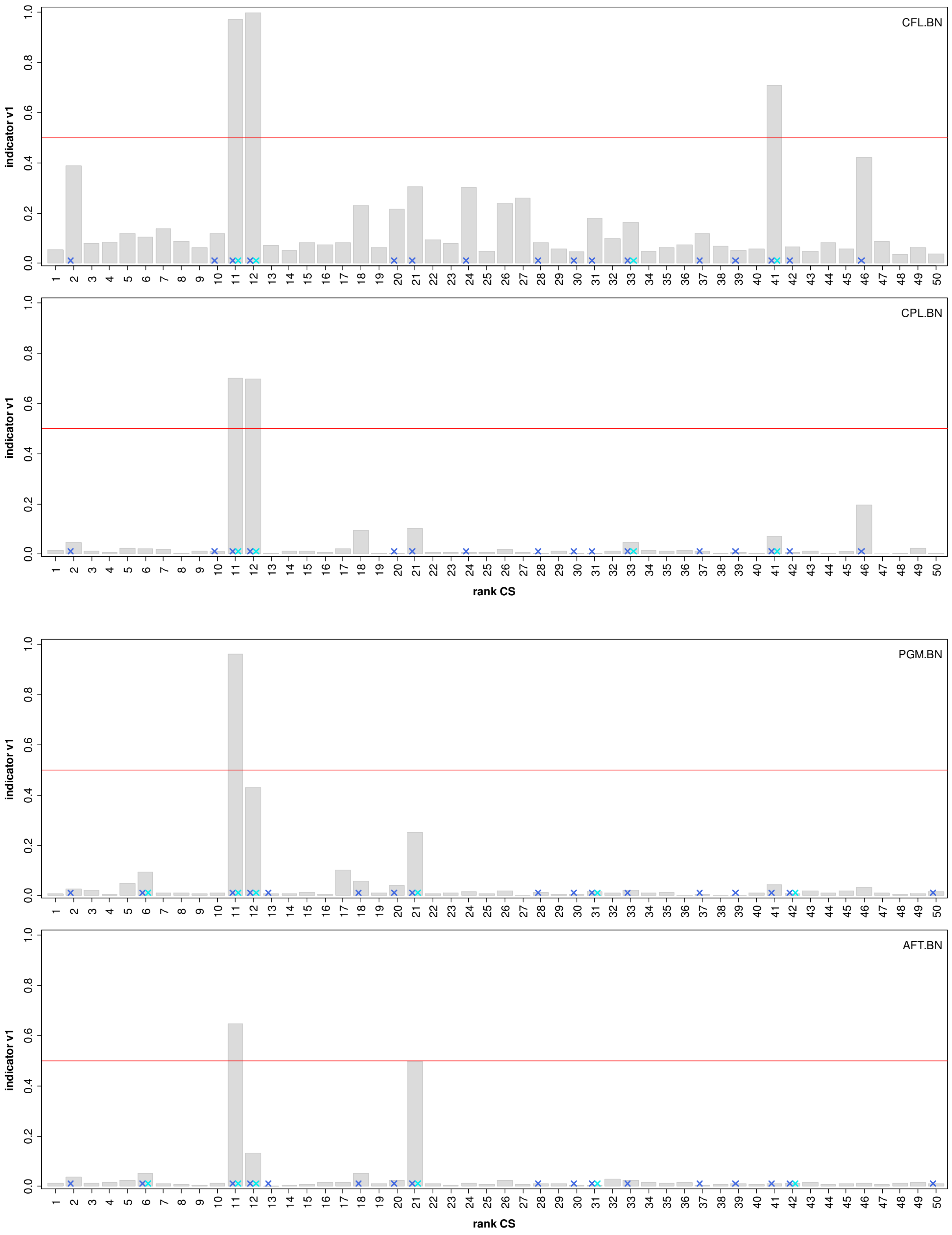


Figure 14.11: Estimated inclusion probabilities based on posterior relative frequencies of the Bayesian NMIG indicator variable value $I_j = v_1$ in the models with predictor (14.1) using $p_x = 50$ probe sets. First and second row: CRR model based on the full and partial likelihood. Third and fourth row: AFT model with PGM and Gaussian error. The probe sets are sorted according to the rank of the Cox scores that are displayed at the x-axis. The crosses at the bottom of the bars mark the covariates from the corresponding frequentist models, which are significant with respect to the p-value 0.05 (cyan) and which are selected by the frequentist stepwise variable selection procedure (dark blue). The red horizontal line marks the cut off value 0.5 of the hard shrinkage selection criterion HS.IND.

find such evidence for a nonlinear effect in case of the clinical covariate *age*. Also the inspection of the shape of the baseline quantities provides information for the need of the flexible modeling, and in particular the Gaussian error in the AFT model seems to be appropriate to represent the baseline survival time of the population.

CONCLUSION

15. Results

We have developed different types of regularization priors for flexible accelerated failure time and hazard regression models that allow the combined modeling of complex predictor structures together with the regularization of linear effects of possibly high-dimensional covariate vectors. We considered random-walk smoothing priors for the model components that are represented by linear combinations of basis functions, like the baseline survival quantities or the nonlinear effects in the predictor. For the regularization of the linear covariate effects we examined three different priors, the Bayesian ridge and lasso prior and a normal mixture of inverse gamma distributions (NMIG) as prior that supplements regularization with a natural possibility for variable selection based on latent indicator variables. The developed methods are implemented in `R`-functions and the `BayesX` software.

The provided Bayesian approach is of practical relevance, e. g. in the context of gene expression data, since the flexible modeling of clinical covariates can be combined with the regularization of high-dimensional microarray features and pre-specification (and validation) of parametric assumptions about the underlying baseline survival time is redundant. The combined flexible modeling can be viewed as improvement over a purely parametric approach, since it enables a visual inspection of the linearity of effects or parametric shapes of baseline survival time and provides more flexible functional shapes when needed. The flexible formulation of model components increases the model complexity, which limits the scope of application for the extended modeling with respect to reliable inference and also with respect to the predictive performance in cases, where only a few number of observations are available. In such situations (simpler) parametric structures, e. g of the baseline quantities, can be specified that are also included in our approaches as special cases.

The restriction that posterior mean estimates in regularized regression models do not directly provide access to the variable selection property can be overcome by the application of hard shrinkage selection rules for the regularized estimates of the regression coefficients or for the posterior inclusion probabilities provided by the NMIG prior structure. But in our simulations and applications we found some evidence that posterior mean models without variable selection are beneficial, even if the sparsity assumption is fulfilled by the data under consideration or when considering the prediction from regularized regression models. In most of our simulations we have seen that the predictive performance of sparse final models rarely achieves the predictive performance of the models with the full predictor. Similar results are obtained from the frequentist perspective, where the performance of the frequentist ridge models with full predictor was often comparable or higher with respect to the sparse models obtained with the frequentist lasso. We found also evidence from the practical perspective, Section 14, that the predictive performance of an estimated model is enhanced, if the covariate effects are regularized and all covariates are included in the model. In particular the reduced shrinkage of larger effects with NMIG prior caused a clear loss of the predictive performance and the best results are obtained with the ridge regularization. In some of the considered simulations the HS.STD criterion provided sparse final models with a comparable predictive performance as the full model, also in combination with the NMIG prior.

In summary, our analyses allow to conclude that, depending on the specific purpose of the analysis, different variants of Bayesian regularization seem to be more or less suitable. The NMIG prior structure showed excellent results in the simulations if the underlying model is sparse and the effects are well separated in clearly small or large effects. The estimated models have a high predictive performance and (hard shrinkage) variable selection can lead in such cases to sparse final models with a similar predictive performance as the models with the full predictor. In the higher-dimensional cases the NMIG-specific reduced shrinkage of larger effects leads to an increased performance compared to the ridge or lasso prior, but in models with moderate to small effects the estimates obtained with the ridge or lasso prior outperform the estimates with the NMIG prior. Since we do not know in reality, if the underlying model is really sparse or something about the effect sizes, it is hard to find recommendations for the general use of a specific regularization prior or the unrestricted application of variable selection.

16. Outlook

16.1. Prior tuning

A major topic of the future research is the further investigation of the shrinkage and selection prior tuning. We provided in Section 4 some crude guidelines for the prior tuning in terms of the standardized constraint parameter or the intersection points of the mixture components of marginal NMIG prior for the regression coefficients. These guidelines consider the prior tuning from a “local” univariate perspective for single regression coefficients and do take not into account the “global” problem, where the regularization depends also on the number of regularized covariates, the correlations between the covariates the sample size and the used regression model.

With respect to the results obtained with the higher-dimensional linear predictors there is a need for a systematic investigation of the prior tuning to counterbalance (to a certain degree) the strong regularization of larger effects. In particular in the case of the NMIG prior, modifications are required to stabilize the estimation of inclusion probabilities according to the impact of the covariates and the separation of moderately large from zero effects. We have seen that the effect of the sole modification of the model complexity is limited, since it increases at last the inclusion probabilities of the zero effects and does not promote the separation of the moderate from the zero effects. So the adaption must also take place on the level of the variance parameters and we have to clarify which of the four hyperparameters need to be modified and how they are to modify. In the case of the lasso or ridge prior possibly the adaptive versions with covariate-specific shrinkage parameter (and common hyperparameters) may help to reduce the strong shrinkage of the larger effects and we have to determine how the hyperparameters change dependent on the dimension of the considered regularized covariates. The investigation in models with adaptive versions of the proposed regularization priors will also be of interest with respect to the front-up scaling of the covariates. Covariate-specific shrinkage parameters, with individual hyperparameters, allow a covariate-specific tuning of the marginal prior of the regression coefficients, which leads to further flexibility and to the option to adapt the prior to the scale of covariates. A systematic connection between the prior tuning and the covariate scale could help to overcome the necessity to standardize covariates up-front, since the priors are allowed to adapt to the varying covariate scales.

A strategy for tuning the priors becomes also important in the lower-dimensional cases, if we will consider the group-prior versions. In particular with the NMIG prior, the inclusion probability increases for larger groups of associated regression coefficients (with comparable sizes) and we have to take into account the group size to avoid e. g. that large groups with small effects achieve a higher or similar inclusion probability than small groups with large effects. So far we have not regarded the correlation structure of the covariates with respect to the prior tuning, and this is also a topic for our future research. Correlations of the covariates can be considered e. g. by the incorporation of the empirical correlation matrix of the covariates into the Gaussian prior of the regression coefficients, compare e. g. George and McCulloch (1993). In addition, the investigation of asymptotic properties of the estimates under the NMIG prior, in analogy to the results presented in Ishwaran and Rao (2005b) for Gaussian regression models, is of interest and in this case it might be also necessary to modify the priors to achieve a non-vanishing impact of the regularization priors even for large sample sizes.

At last, the prior tuning must be verified in the specific (survival) regression model and possibly adapted to the specific regression context, in particular the tuning of the NMIG prior. The NMIG prior, with hyperparameter constellation used in the CRR simulations and applications, was tuned in the CRR model. From the Section 13 we have seen that this basic tuning of NMIG prior leads to very different shrinkage behavior in the AFT and CRR model. While the unregularized estimated effects under the CRR or AFT model differ only moderately (some effects are larger in the AFT model) the basic NMIG tuning leads to close to zero estimates for all inclusion probabilities in the AFT model. So we have to verify the performance of this hyperparameter constellation in the AFT model or search for alternatives.

16.2. Regularization priors

On the predictor side, in particular with respect to the regularization of linear covariate effects, there are a lot of self-evident generalizations. Since the scale mixture of normals class is quite large, as demonstrated for example in Griffin and Brown (2005), other types of regularization priors for linear predictors, that support such a hierarchical representation, can be considered in the same unified Bayesian framework. Recently Li and Lin (2010) represented the Bayesian elastic net prior as mixture of normals with truncated gamma mixing distribution for the variance function which fits also in our framework and can be utilized for correlated predictors.

In addition, the common regularization for associated groups of regression coefficients, arising e. g. from categorical covariates, can be considered. We mentioned this expansion already in the Sections 4.1.3, 4.2.3 and 4.3.3, and there are only marginal modifications in the present implementation of the methods necessary to enable the group-versions of the Bayesian ridge, lasso and NMIG regularization (also the adaptive group-versions) for subsets of covariates in the extended predictor.

We can also think about a mixture of gdP-distributions as marginal prior of the regression as an alternative to the mixture of Student t-distributions induced by the NMIG prior structure. This would lead to a more beneficial behavior of the associated penalty at the origin, due the non-continuous first derivate of the marginal prior of the regression coefficients, and should lead to similar theoretical properties as already derived for the gdP prior.

16.3. Generalized accelerated hazard model

Further it would be interesting to consider the NMIG prior in a kind of regression model selection framework. Etezadi-Amoli and Ciampi (1987) formulate an extended model for the hazard rate function that includes several survival model types as special cases. They assume that covariates $\mathbf{x} = (x_1, ... x_p)'$ change the hazard function according to

$$\lambda(t \mid \mathbf{x}) = \lambda_0(t \cdot \exp(\mathbf{x}'\boldsymbol{\beta}_1)) \cdot \exp(\mathbf{x}'\boldsymbol{\beta}_2),$$

where two different covariate effects $\boldsymbol{\beta}_1$ and $\boldsymbol{\beta}_2$ modify the baseline hazard function $\lambda_0(\cdot)$ and the hazard function $\lambda(\cdot)$ separately. The authors use regression splines to model the baseline hazard and the ML approach for inference. This generalized accelerated hazard (GAH) model provides a natural generalization of the CRR and the AFT model, because the CRR model results for $\boldsymbol{\beta}_1 = \mathbf{0}$ and the AFT model results for $\boldsymbol{\beta}_1 = \boldsymbol{\beta}_2$. Also the accelerated hazard (AH) model of Chen and Wang (2000) is included as third special case if $\boldsymbol{\beta}_2 = \mathbf{0}$. The GAH model allows the simultaneous treatment of different assumptions about the covariate impact in one unified regression model and provides the possibility to discriminate covariates with respect to the fashion how they influence survival time. This weakens the reliance on specific assumptions about the impact of the whole set of covariates and provides a very flexible model class, where e. g. subsets of covariates can act in a AFT-, CRR- or AH-fashioned way on the survival time. The extension of our methodology to the GAH model, in particular the usage of the selection (or shrinkage) priors in the GAH models is appealing to uncover either the underlying particular regression model class for the whole set of covariates (at least CRR with $\boldsymbol{\beta}_1 = \mathbf{0}$ or AH with $\boldsymbol{\beta}_2 = \mathbf{0}$) or to classify single covariates due to their specific form of influence ($\beta_{1,j} = 0$ or $\beta_{2,j} = 0$). To detect subsets of covariates with AFT like impact ($\boldsymbol{\beta}_1 = \boldsymbol{\beta}_2$), a modification of the shrinkage behavior of the priors is necessary. One option is to allow the shrinkage of coefficients toward multiple prior means (including the zero mean), where the grouping of coefficients around specific values of a common grid of prior means for $\boldsymbol{\beta}_1$ and $\boldsymbol{\beta}_2$ may guide the classification. E. g. MacLehose and Dunson (2010) use a Dirichlet process prior to induce a clustering of the regression coefficients into groups in the context of the Bayesian lasso prior. For realization we can also think about the expansion of the NMIG prior by introducing a finite mixture (with more than two components) for the variance parameter, where the latent component labels (indicator variables) guide the classification. At least also an extended structured additive predictor could be considered in this general model class.

16.4. Time dependent covariates

In general, the AFT model would benefit from the further extension of the predictor to take into account other covariate effects like random effects, covariate interactions or spatial effects and the implementation of alternative censoring schemes to the right censoring. Another potential for the generalization of our approach is the extension of the AFT (or GAH) model to take into account effects of time-dependent covariates $x(t)$, which seems to be rather a practical and numerical than a theoretical challenge. E. g. Cox and Oakes (1984) propose the generalization

$$t_0 = f(t,\beta) = \int_0^t \exp(x(s)\beta)ds = \int_0^t \exp(\eta(s))ds\,, \qquad (*)$$

where t_0 is the unobservable baseline survival time and t the observable survival time. Inference is carried out with the hazard rate formulation of the (right-censored) likelihood and the likelihood contributions are based on the components $\lambda(t) = \lambda_0(\int_0^t \exp\big(\eta(s)\big)ds)\cdot\exp(\eta(s))$ and $\Lambda_0(\int_0^t \exp\big(\eta(s)\big)ds)$, where $\lambda_0(\cdot)$ denotes the hazard function and $\Lambda_0(\cdot)$ the cumulative hazard function of the baseline survival time. To obtain a full likelihood function, Tseng et al. (2005) use (in the context with longitudinal data) a piecewise constant hazard function to approximate the baseline hazard, but flexible AFT models can also be obtained in a similar fashion like in the CRR model by representing the log-baseline hazard as P-spline. It seems to be straightforward to derive a related hierarchical model structure for the AFT model with time-dependent covariates as for the CRR model in Sections 7 and 8, if we treat the unobservable baseline survival times t_0 as latent model variables that are imputed according to $(*)$. The further use of the IWLS proposals may be a point to think about, since also first and second order derivates of the baseline hazard are involved in the construction. In particular, the use of the conditional prior proposals, Knorr-Held (1999), may be advantageous in this context, since this kind of proposal requires only the evaluation of the log-likelihood and not the derivates. Due to the integral formulation $(*)$, the computational effort for the numerical evaluation of the integrals (including $\Lambda_0(\cdot)$) during the samplers increases. A similar representation in terms of the log-linear AFT model with PGM error may be possible, but the imputation of the exact event times is no longer feasible, since no measurements of the time-dependent covariates are available after the observed event or censoring time to evaluate $(*)$.

16.5. Software

The current versions of the functions `baftpgm()` and `bcoxpl()` enable the fitting of models with predictors that describe linear and smooth effects of time independent covariates. Both functions will be further developed to consider extensions of the predictor to take into account other effect types, to incorporate alternative regularization priors that fit in our hierarchical model structure, to consider the common regularization of associated groups of covariates, to incorporate alternative censoring schemes to the right-censoring and to accelerate the samplers by outsourcing the basic routines e. g. to C++.

Within the `BayesX` software the predictor can capture a greater variety of effect types in combination with the regularization of the linear effects utilizing the adaptive and non-adaptive versions on the Bayesian ridge, lasso and NMIG prior. Besides the continuous time hazard regression for right censored observations the methods are available for other censoring schemes like interval censoring or a broader class of response distributions like right censored discrete time hazard regression, those from the exponential family or categorical responses. Also the routine `regress` can be extended with low expense to take into account the (adaptive) group-versions of the implemented regularization priors.

APPENDIX

A Extended AFT model

A.1. Penalized Gaussian mixture

Let X follow a Gaussian mixture distribution, i. e., $X \sim w_1 N(m_1, s_1^2) + ... + w_g N(m_g, s_g^2)$, where w_j, $j = 1,...,n$, denote the mixture weights with $w_j > 0$ and $w_1 + w_2 + ... + w_g = 1$, and m_j, s_j^2 represent the mean and variance of the associated Gaussian basis densities $\varphi(\cdot)$.

The mixture distribution density is defined as

$$f_X(x) = \sum_{j=1}^{g} w_j \varphi\left(x \mid m_j, s_j^2\right).$$

A.1.1. Mean and variance

The mean and variance of the Gaussian mixture are given as

$$\mu_X = \mathbb{E}(X) = \sum_{j=1}^{g} w_j m_j, \quad \sigma_X^2 = \mathbb{V}\mathrm{ar}(X) = \sum_{j=1}^{g} w_j \left(m_j^2 + s_j^2\right) - \mu_X^2.$$

This can be shown as follows: The expectation of a function $h(\cdot)$ with respect to the mixture distribution density $f_X(\cdot)$ is given by

$$\mathbb{E}\big(h(X)\big) = \int h(x) f_X(x) dx = \int h(x) \sum_{j=1}^{g} w_j \varphi(x \mid m_j, s_j^2) dx = \sum_{j=1}^{g} w_j \int h(x) \varphi(x \mid m_j, s_j^2) dx = \sum_{j=1}^{g} w_j \mathbb{E}_j\big(h(X)\big),$$

where $\mathbb{E}_j(h(X))$ is the expectation of $h(X)$ with respect to the j-th basis density $\varphi(\cdot \mid m_j, s_j^2)$.

The mean μ_X and the variance σ_X^2 of the mixture distribution are then obtained as special cases with the specifications $h(X) = X$ and $h(X) = (X - \mu_X)^2$, i. e.

$$\mu_X = \mathbb{E}(X) = \sum\nolimits_{j=1}^{g} w_j \mathbb{E}_j(X) = \sum\nolimits_{j=1}^{g} w_j m_j,$$

$$\begin{aligned}\sigma_X^2 = \mathbb{V}\mathrm{ar}(X) &= \sum\nolimits_{j=1}^{g} w_j \mathbb{E}_j (X - \mu_X)^2 = \sum\nolimits_{j=1}^{g} w_j \mathbb{E}_j(X^2) - 2\mu_X \sum\nolimits_{j=1}^{g} w_j m_j + \mu_X^2 \sum\nolimits_{j=1}^{g} w_j \\ &= \sum\nolimits_{j=1}^{g} w_j \left(\mathbb{V}\mathrm{ar}_j(X^2) + m_j^2\right) - \mu_X^2 = \sum\nolimits_{j=1}^{g} w_j \left(s_j^2 + m_j^2\right) - \mu_X^2,\end{aligned}$$

where $m_j = \mathbb{E}_j(X)$ and $s_j^2 = \mathbb{V}\mathrm{ar}_j(X)$ denote the expectation and the variance of X with respect to the j-th basis density. In the equations for the variance we used the condition $\sum w_j = 1$ and the variance partition $\mathbb{V}\mathrm{ar}(X) = \mathbb{E}_j(X^2) - \mathbb{E}_j^2(X) = \mathbb{E}_j(X^2) - m_j^2$.

A.1.2. Standardization

To standardize the mixture distribution, $\mu_X = 0$ and $\sigma_X^2 = 1$, we use the unconstrained coefficients $\boldsymbol{\alpha} = (\alpha_1, ..., \alpha_g)'$ as defined in Section 2.3 with the identifiability constraint $\alpha_g = 0$ and the connection $w_k(\boldsymbol{\alpha}) = \exp(\alpha_k) \big/ \sum_{j=1}^{g} \exp(\alpha_j)$, $k = 1, ..., g$. To shorten the notation, we write $\exp(\alpha_j) = e^{\alpha_j}$.

For a standardized mixture distribution we obtain the following condition for the mean

$$\begin{aligned} \mu_X = 0 \quad &\Leftrightarrow \quad \frac{\sum_{j=1}^{g} e^{\alpha_j} m_j}{\sum_{j=1}^{g} e^{\alpha_j}} = 0 \\ &\Leftrightarrow \quad \sum_{j=1}^{g} e^{\alpha_j} m_j = 0 \\ &\Leftrightarrow \quad \sum_{j=1}^{g-3} e^{\alpha_j} m_j + e^{\alpha_{g-2}} m_{g-2} + e^{\alpha_{g-1}} m_{g-1} + m_g = 0 \end{aligned}$$

and for the variance

$$\begin{aligned} \sigma_X^2 = 1 \quad &\Leftrightarrow \quad \frac{\sum_{j=1}^{g} e^{\alpha_j} - \sum_{j=1}^{g} e^{\alpha_j}(s_j^2 + m_j^2)}{\sum_{j=1}^{g} e^{\alpha_j}} = 0 \\ &\Leftrightarrow \quad \sum_{j=1}^{g} e^{\alpha_j}(1 - s_j^2 - m_j^2) = 0 \\ &\Leftrightarrow \quad \sum_{j=1}^{g-3} e^{\alpha_j}(1 - s_j^2 - m_j^2) + e^{\alpha_{g-2}}(1 - s_{g-2}^2 - m_{g-2}^2) + e^{\alpha_{g-1}}(1 - s_{g-1}^2 - m_{g-1}^2) + 1 - s_g^2 - m_g^2 = 0. \end{aligned}$$

The specification of the coefficients $\alpha_{g-1}, \alpha_{g-2}$ from the mean condition is

$$\begin{aligned} &\text{(I)} \quad e^{\alpha_{g-1}} = -\sum_{j=1}^{g-3} e^{\alpha_j} \frac{m_j}{m_{g-1}} - e^{\alpha_{g-2}} \frac{m_{g-2}}{m_{g-1}} - \frac{m_g}{m_{g-1}}, \\ &\text{(II)} \quad e^{\alpha_{g-2}} = -\sum_{j=1}^{g-3} e^{\alpha_j} \frac{m_j}{m_{g-2}} - e^{\alpha_{g-1}} \frac{m_{g-1}}{m_{g-2}} - \frac{m_g}{m_{g-2}}. \end{aligned}$$

The specifications of the coefficients $\alpha_{g-1}, \alpha_{g-2}$ from the variance condition is

$$\begin{aligned} &\text{(III)} \quad e^{\alpha_{g-1}} = -\sum_{j=1}^{g-3} e^{\alpha_j} \frac{1 - s_j^2 - m_j^2}{1 - s_{g-1}^2 - m_{g-1}^2} - e^{\alpha_{g-2}} \frac{1 - s_{g-2}^2 - m_{g-2}^2}{1 - s_{g-1}^2 - m_{g-1}^2} - \frac{1 - s_g^2 - m_g^2}{1 - s_{g-1}^2 - m_{g-1}^2}, \\ &\text{(V)} \quad e^{\alpha_{g-2}} = -\sum_{j=1}^{g-3} e^{\alpha_j} \frac{1 - s_j^2 - m_j^2}{1 - s_{g-2}^2 - m_{g-2}^2} - e^{\alpha_{g-1}} \frac{1 - s_{g-1}^2 - m_{g-1}^2}{1 - s_{g-2}^2 - m_{g-2}^2} - \frac{1 - s_g^2 - m_g^2}{1 - s_{g-2}^2 - m_{g-2}^2}. \end{aligned}$$

Inserting (V) in (I) we get

$$e^{\alpha_{g-1}} = -\sum_{j=1}^{g-3} e^{\alpha_j} \frac{m_j}{m_{g-1}} + \left(\sum_{j=1}^{g-3} e^{\alpha_j} \frac{1 - s_j^2 - m_j^2}{1 - s_{g-2}^2 - m_{g-2}^2} + e^{\alpha_{g-1}} \frac{1 - s_{g-1}^2 - m_{g-1}^2}{1 - s_{g-2}^2 - m_{g-2}^2} + \frac{1 - s_g^2 - m_g^2}{1 - s_{g-2}^2 - m_{g-2}^2} \right) \frac{m_{g-2}}{m_{g-1}} - \frac{m_g}{m_{g-1}}$$

$$e^{\alpha_{g-1}} \left(1 - \frac{1 - s_{g-1}^2 - m_{g-1}^2}{1 - s_{g-2}^2 - m_{g-2}^2} \frac{m_{g-2}}{m_{g-1}} \right) = -\sum_{j=1}^{g-3} e^{\alpha_j} \left(\frac{m_j}{m_{g-1}} - \frac{1 - s_j^2 - m_j^2}{1 - s_{g-2}^2 - m_{g-2}^2} \frac{m_{g-2}}{m_{g-1}} \right) + \frac{1 - s_g^2 - m_g^2}{1 - s_{g-2}^2 - m_{g-2}^2} \frac{m_{g-2}}{m_{g-1}} - \frac{m_g}{m_{g-1}}$$

$$e^{\alpha_{g-1}} = \frac{-\sum_{j=1}^{g-3} e^{\alpha_j} \left((1 - s_{g-2}^2 - m_{g-2}^2) m_j - (1 - s_j^2 - m_j^2) m_{g-2} \right) + (1 - s_g^2 - m_g^2) m_{g-2} - (1 - s_{g-2}^2 - m_{g-2}^2) m_g}{(1 - s_{g-2}^2 - m_{g-2}^2) m_{g-1} - (1 - s_{g-1}^2 - m_{g-1}^2) m_{g-2}}.$$

Inserting (III) in (II) we get accordingly

$$e^{\alpha_{g-2}} = -\sum_{j=1}^{g-3} e^{\alpha_j} \frac{m_j}{m_{g-2}} + \left(\sum_{j=1}^{g-3} e^{\alpha_j} \frac{1-s_j^2-m_j^2}{1-s_{g-1}^2-m_{g-1}^2} + e^{\alpha_{g-2}} \frac{1-s_{g-2}^2-m_{g-2}^2}{1-s_{g-1}^2-m_{g-1}^2} + \frac{1-s_g^2-m_g^2}{1-s_{g-1}^2-m_{g-1}^2} \right) \frac{m_{g-1}}{m_{g-2}} - \frac{m_g}{m_{g-2}}$$

$$e^{\alpha_{g-2}} \left(1 - \frac{1-s_{g-2}^2-m_{g-2}^2}{1-s_{g-1}^2-m_{g-1}^2} \frac{m_{g-1}}{m_{g-2}} \right) = -\sum_{j=1}^{g-3} e^{\alpha_j} \left(\frac{m_j}{m_{g-2}} - \frac{1-s_j^2-m_j^2}{1-s_{g-1}^2-m_{g-1}^2} \frac{m_{g-1}}{m_{g-2}} \right) + \frac{1-s_g^2-m_g^2}{1-s_{g-1}^2-m_{g-1}^2} \frac{m_{g-1}}{m_{g-2}} - \frac{m_g}{m_{g-2}}$$

$$e^{\alpha_{g-2}} = \frac{-\sum_{j=1}^{g-3} e^{\alpha_j} \left((1-s_{g-1}^2-m_{g-1}^2)m_j - (1-s_j^2-m_j^2)m_{g-1} \right) + (1-s_g^2-m_g^2)m_{g-1} - (1-s_{g-1}^2-m_{g-1}^2)m_g}{(1-s_{g-1}^2-m_{g-1}^2)m_{g-2} - (1-s_{g-2}^2-m_{g-2}^2)m_{g-1}}.$$

With the definitions

$$c_{j,g-1} := -\frac{(1-s_{g-2}^2-m_{g-2}^2)m_j - (1-s_j^2-m_j^2)m_{g-2}}{(1-s_{g-2}^2-m_{g-2}^2)m_{g-1} - (1-s_{g-1}^2-m_{g-1}^2)m_{g-2}}, \quad j=1,\ldots,g-3,g,$$

$$c_{j,g-2} := -\frac{(1-s_{g-1}^2-m_{g-1}^2)m_j - (1-s_j^2-m_j^2)m_{g-1}}{(1-s_{g-1}^2-m_{g-1}^2)m_{g-2} - (1-s_{g-2}^2-m_{g-2}^2)m_{g-1}}, \quad j=1,\ldots,g-3,g,$$

and taking the logarithm we get, in addition to the identifying constraint $\alpha_g = 0$, the conditions

$$\alpha_{g-1} = \log\left(\sum_{j=1}^{g-3} e^{\alpha_j} c_{j,g-1} + c_{g,g-1} \right), \quad \alpha_{g-2} = \log\left(\sum_{j=1}^{g-3} e^{\alpha_j} c_{j,g-2} + c_{g,g-2} \right),$$

to ensure, that the mixture density is standardized. Using equal basis variances $s_j^2 = s^2$, $j=1,\ldots,g$, we obtain

$$c_{j,g-1} = -\frac{m_j - m_{g-2}}{m_{g-1} - m_{g-2}} \frac{1-s^2-m_j m_{g-1}}{1-s^2-m_{g-1}m_{g-2}}, \quad j=1,\ldots,g-3,g,$$

$$c_{j,g-2} = -\frac{m_j - m_{g-1}}{m_{g-2} - m_{g-1}} \frac{(1-s^2-m_j m_{g-2})}{1-s^2-m_{g-1}m_{g-2}}, \quad j=1,\ldots,g-3,g.$$

A.1.3. Linear transformation

The linear transformation $Y = \mu + \sigma X$ of the Gaussian mixture distribution X is also a Gaussian mixture distribution with density

$$f_Y(y) = \frac{1}{\sigma} f_X\left(\frac{y-\mu}{\sigma} \right) = \sum_{j=1}^{g} w_j \frac{1}{\sigma} \varphi\left(\frac{y-\mu}{\sigma} \,|\, m_j, s_j^2 \right) = \sum_{j=1}^{g} \frac{w_j}{\sqrt{2\pi}\sigma s_j} \exp\left(-\frac{1}{2\sigma^2 s_j^2} (y_i - \eta_i - \sigma m_j)^2 \right)$$
$$= \sum_{j=1}^{g} w_j \varphi\left(y \,|\, \mu + \sigma m_j, \sigma^2 s_j^2 \right),$$

where $m_{Y,j} := \mu + \sigma m_j$ denote the knots and $s_{Y,j}^2 := \sigma^2 s_j^2$ the variances of the associated Gaussian basis densities $\varphi(\cdot)$. The mean and variance of Y are given as $\mu_Y = \mu + \sigma \mu_X$ and $\sigma_Y^2 = \sigma^2 \sigma_X^2$.

A.2. Full conditional of the scale parameter

With the definitions from Section 6.1.4

$$A_\sigma := \frac{n}{2} + h_{\sigma,1} + 1, \quad B_\sigma := \frac{1}{2}(\mathbf{y}-\boldsymbol{\eta})' \mathbf{S}_r^{-1} (\mathbf{y}-\boldsymbol{\eta}) + h_{\sigma,2}, \quad C_\sigma := (\mathbf{y}-\boldsymbol{\eta})' \mathbf{S}_r^{-1} \mathbf{m}_r$$

the full conditional of the scale parameter is built as

$$p\left(\sigma^2 \,|\, \cdot\right) \propto \left(\frac{1}{\sigma^2}\right)^{A_\sigma} \exp\left(-\frac{1}{\sigma^2}B_\sigma + \frac{1}{\sigma}C_\sigma\right).$$

Unimodality of the full conditional

To see the unimodality of the full conditional, we consider the logarithm of the full conditional and the first derivate with respect to σ^2:

$$\log p\left(\sigma^2 \,|\, \cdot\right) = A_\sigma \log\left(\frac{1}{\sigma^2}\right) - \frac{1}{\sigma^2}B_\sigma + \frac{1}{\sigma}C_\sigma = -A_\sigma \log\sigma^2 - \frac{B_\sigma}{\sigma^2} + \frac{C_\sigma}{\sqrt{\sigma^2}},$$

$$\frac{d}{d\sigma^2}\log p(\sigma^2 \,|\, \cdot) = -\frac{A_\sigma}{\sigma^2} + \frac{B_\sigma}{\sigma^4} - \frac{C_\sigma}{2\sigma^3} = -\frac{1}{\sigma^4}\left(A_\sigma\sigma^2 + \frac{1}{2}C_\sigma\sigma - B_\sigma\right). \qquad \text{(A.1)}$$

The possible two roots σ_+ and σ_- of the expression

$$-\frac{1}{2\sigma^4}\left(2A_\sigma\sigma^2 + C_\sigma\sigma - 2B_\sigma\right) = 0$$

from the first derivate (A.1) are given as

$$\sigma_{+/-} = \frac{1}{4A_\sigma}\left(-C_\sigma \pm \sqrt{C_\sigma^2 + 16A_\sigma B_\sigma}\right) = \frac{1}{4A_\sigma}\left(-C_\sigma \pm \left|C_\sigma\right|\sqrt{1 + \frac{16A_\sigma B_\sigma}{C_\sigma^2}}\right).$$

Since $A_\sigma > 0$, $B_\sigma > 0$ and since the expression $\sqrt{1 + 16A_\sigma B_\sigma / C_\sigma^2} > 1$, we see that for each $C_\sigma \neq 0$ there is only one of the roots positive, i. e. $\sigma_+ = (-C_\sigma + \sqrt{C_\sigma^2 + 16A_\sigma B_\sigma})/4A_\sigma$, which must then be the single mode of the full conditional distribution.

B Marginal distributions of regularization prior components

In the following derivations we require some properties of the gamma function. The gamma function is defined via an improper integral $\Gamma(x) := \int_0^\infty t^{x-1}\exp(-t)dt$, $x > 0$ and satisfies

$$\Gamma\left(x+1\right) = x\Gamma\left(x\right). \qquad \text{(B.1)}$$

Based on the initial definition of the gamma function, we derive with $x = a + 1$ and the substitution $x = t/b$ the representation

$$\int_0^\infty x^a \exp\left(-bx\right)dx = \frac{\Gamma\left(a+1\right)}{b^a}. \qquad \text{(B.2)}$$

B.1. Bayesian ridge prior

Marginal distribution of the regression coefficients

Version (A): Using the hierarchy of the Bayesian ridge prior, version (A), compare Section 4.1.1, we obtain with the inverse gamma prior of the variance parameters $\tau_{\beta_j}^2 \,|\, h_{1,\lambda}, h_{2,\lambda} \sim \text{IGamma}\left(h_{1,\lambda}, 0.5h_{2,\lambda}\right)$ and with the Gaussian prior of the regression coefficients $\beta_j \,|\, \tau_{\beta_j}^2 \sim N(0, \tau_{\beta_j}^2)$ the marginal prior

$$
\begin{aligned}
p\left(\beta_j \mid h_{1,\lambda}, h_{2,\lambda}\right) &= \int N\left(\beta_j \mid 0, \tau_{\beta_j}^2\right) \cdot \mathrm{InvGamma}\left(\tau_{\beta_j}^2 \mid h_{1,\lambda}, \tfrac{1}{2} h_{2,\lambda}\right) d\tau_{\beta_j}^2 \\
&= \frac{(\frac{1}{2} h_{2,\lambda})^{h_{1,\lambda}}}{\sqrt{2\pi}\Gamma(h_{1,\lambda})} \int \left(\tau_{\beta_j}^{-2}\right)^{(h_{1,\lambda}+\frac{1}{2})+1} \exp\left(-(\tfrac{1}{2}\beta_j^2 + \tfrac{1}{2} h_{2,\lambda})\tau_{\beta_j}^{-2}\right) d\tau_{\beta_j}^2 \\
&= \frac{(\frac{1}{2} h_{2,\lambda})^{h_{1,\lambda}} \Gamma\left(h_{1,\lambda} + \frac{1}{2}\right)}{\sqrt{2\pi}\Gamma(h_{1,\lambda})} \left(\tfrac{1}{2}\beta_j^2 + \tfrac{1}{2} h_{2,\lambda}\right)^{-\left(h_{1,\lambda}+\frac{1}{2}\right)} = \frac{\Gamma\left(h_{1,\lambda} + \frac{1}{2}\right)}{\sqrt{h_{2,\lambda}\pi}\Gamma(h_{1,\lambda})} \left(\frac{\beta_j^2}{h_{2,\lambda}} + 1\right)^{-\left(h_{1,\lambda}+\frac{1}{2}\right)} \\
&= \frac{\Gamma\left(\frac{2h_{1,\lambda}+1}{2}\right)}{\Gamma\left(\frac{2h_{1,\lambda}}{2}\right)\sqrt{2h_{1,\lambda}\frac{h_{2,\lambda}}{2h_{1,\lambda}}\pi}} \left(1 + \frac{\beta_j^2}{2h_{1,\lambda}\frac{h_{2,\lambda}}{2h_{1,\lambda}}}\right)^{-\frac{2h_{1,\lambda}+1}{2}} .
\end{aligned}
$$

In the third conversion we use the connection (B.2). Thus the resulting distribution is a scaled Student t-distribution with $d = 2h_{1,\lambda}$ degrees of freedom and scale parameter $s = \sqrt{h_{2,\lambda}/2h_{1,\lambda}}$

$$\beta_j \mid h_{1,\lambda}, h_{2,\lambda} \sim t\left(df = 2h_{1,\lambda}, s = \sqrt{h_{2,\lambda}/2h_{1,\lambda}}\right).$$

Version (B): Using the hierarchy of the Bayesian ridge prior, version (B), compare Section 4.1.1, we obtain with the inverse gamma prior of the variance parameters $\tau_\beta^2 \mid h_{1,\lambda}, h_{2,\lambda} \sim \mathrm{IGamma}\left(h_{1,\lambda}, 0.5h_{2,\lambda}\right)$ and with the multivariate Gaussian prior of the regression coefficients $\boldsymbol{\beta} \mid \tau_\beta^2 \sim N(\mathbf{0}, \tau_\beta^2 \mathbf{I})$ the marginal prior

$$
\begin{aligned}
p\left(\boldsymbol{\beta} \mid h_{1,\lambda}, h_{2,\lambda}\right) &= \int N\left(\boldsymbol{\beta} \mid \mathbf{0}, \tau_\beta^2\right) \cdot \mathrm{InvGamma}\left(\tau_\beta^2 \mid h_{1,\lambda}, \tfrac{1}{2} h_{2,\lambda}\right) d\tau_\beta^2 \\
&= \frac{(\frac{1}{2} h_{2,\lambda})^{h_{1,\lambda}}}{\sqrt{2\pi}^{p_x}\Gamma(h_{1,\lambda})} \int \left(\tau_\beta^{-2}\right)^{(h_{1,\lambda}+\frac{p_x}{2})+1} \exp\left(-(\tfrac{1}{2}\boldsymbol{\beta}'\boldsymbol{\beta} + \tfrac{1}{2} h_{2,\lambda})\tau_\beta^{-2}\right) d\tau_\beta^2 \\
&= \frac{(\frac{1}{2} h_{2,\lambda})^{h_{1,\lambda}} \Gamma\left(h_{1,\lambda} + \frac{p_x}{2}\right)}{\sqrt{2\pi}^{p_x}\Gamma(h_{1,\lambda})} \left(\tfrac{1}{2}\boldsymbol{\beta}'\boldsymbol{\beta} + \tfrac{1}{2} h_{2,\lambda}\right)^{-\left(h_{1,\lambda}+\frac{p_x}{2}\right)} = \frac{\Gamma\left(h_{1,\lambda} + \frac{p_x}{2}\right)}{\sqrt{\pi h_{2,\lambda}}^{p_x}\Gamma(h_{1,\lambda})} \left(\frac{\boldsymbol{\beta}'\boldsymbol{\beta}}{h_{2,\lambda}} + 1\right)^{-\left(h_{1,\lambda}+\frac{1}{2}\right)} \\
&= \frac{\Gamma\left(\frac{2h_{1,\lambda}+p_x}{2}\right)}{\Gamma\left(\frac{2h_{1,\lambda}}{2}\right)\sqrt{2h_{1,\lambda}\frac{h_{2,\lambda}}{2h_{1,\lambda}}\pi}^{p_x}} \left(1 + \frac{\boldsymbol{\beta}'\boldsymbol{\beta}}{2h_{1,\lambda}\frac{h_{2,\lambda}}{2h_{1,\lambda}}}\right)^{-\frac{2h_{1,\lambda}+1}{2}} .
\end{aligned}
$$

In the third conversion we use again the connection (B.2). Thus the resulting distribution is a multivariate, p_x -dimensional, scaled Student t-distribution with $d = 2h_{1,\lambda}$ degrees of freedom and scale matrix $\boldsymbol{\Sigma}^{\frac{1}{2}} = \sqrt{h_{2,\lambda}/2h_{1,\lambda}}\mathbf{I}$

$$\boldsymbol{\beta} \mid h_{1,\lambda}, h_{2,\lambda} \sim t\left(df = 2h_{1,\lambda}, \boldsymbol{\Sigma}^{\frac{1}{2}} = \sqrt{h_{2,\lambda}/2h_{1,\lambda}}\mathbf{I}\right).$$

B.2. Bayesian lasso prior

Marginal distribution of the variance parameters

Using the hierarchy of the Bayesian lasso prior, Section 4.2.1, we obtain with the gamma prior of the shrinkage parameter $\lambda^2 \mid h_{1,\lambda}, h_{2,\lambda} \sim \mathrm{Gamma}\left(h_{1,\lambda}, h_{2,\lambda}\right)$ and with the exponential prior of the variances $\tau_{\beta_j}^2 \mid \lambda^2 \sim \mathrm{Exp}(\tfrac{1}{2}\lambda^2)$ the marginal prior of the variance parameter as

$$
\begin{aligned}
p\left(\tau_{\beta_j}^2 \mid h_{1,\lambda}, h_{2,\lambda}\right) &= \int \operatorname{Exp}\left(\tau_{\beta_j}^2 \mid \tfrac{1}{2}\lambda^2\right) \operatorname{Gamma}\left(\lambda^2 \mid h_{1,\lambda}, h_{2,\lambda}\right) d\lambda^2 \\
&= \frac{h_{2,\lambda}^{h_{1,\lambda}}}{2\Gamma(h_{1,\lambda})} \int \left(\lambda^2\right)^{(h_{1,\lambda}+1)-1} \exp\left(-\tfrac{1}{2}\lambda^2(\tau_{\beta_j}^2 + 2h_{2,\lambda})\right) d\lambda^2 \\
&= \frac{h_{2,\lambda}^{h_{1,\lambda}}}{2\Gamma(h_{1,\lambda})} \frac{\Gamma(h_{1,\lambda}+1)2^{h_{1,\lambda}+1}}{(\tau_{\beta_j}^2 + 2h_{2,\lambda})^{h_{1,\lambda}+1}} = \frac{h_{1,\lambda}(2h_{2,\lambda})^{h_{1,\lambda}}}{(\tau_{\beta_j}^2 + 2h_{2,\lambda})^{h_{1,\lambda}+1}} \\
&= \frac{h_{1,\lambda}(2h_{2,\lambda})^{h_{1,\lambda}}}{(2h_{2,\lambda})^{h_{1,\lambda}+1}(\frac{1}{2h_{2,\lambda}}\tau_{\beta_j}^2 + 1)^{h_{1,\lambda}+1}} = \frac{h_{1,\lambda}}{2h_{2,\lambda}} \left[\frac{\tau_{\beta_j}^2}{2h_{2,\lambda}} + 1\right]^{-(h_{1,\lambda}+1)},
\end{aligned}
$$

which is the density of a generalized Pareto distribution

$$
\tau_{\beta_j}^2 \mid h_{1,\lambda}, h_{2,\lambda} \sim \text{gPareto}(\text{shape} = h_{1,\lambda}, \text{scale} = 2h_{2,\lambda}/h_{1,\lambda}) .
$$

In the third conversion we use the connections (B.1) and (B.2) to solve the integral.

Marginal distribution of the regression coefficients

To derive the marginal distribution of the regression coefficients, we require the following integral-representations of the parabolic cylinder function $D_{-2\nu}(\cdot)$ of order -2ν and $D_{-2\nu+1}(\cdot)$ of order $-2\nu+1$

$$
\int_0^\infty x^{\nu-1} \exp(-mx)(x+y)^{-\nu-1/2} dx = \frac{2^\nu \Gamma(\nu)}{y^{1/2}} \exp\left(\frac{ym}{2}\right) D_{-2\nu}\left(\sqrt{2ym}\right), \tag{B.3}
$$

$$
\int_0^\infty x^{\nu-1} \exp(-mx)(x+y)^{-\nu+1/2} dx = \frac{2^{\nu-1/2} \Gamma(\nu)}{m^{1/2}} \exp\left(\frac{ym}{2}\right) D_{-2\nu+1}\left(\sqrt{2ym}\right), \tag{B.4}
$$

compare e. g. Griffin and Brown (2005).

With the Gaussian prior of the regression coefficients $\beta_j \mid \tau_{\beta_j}^2 \sim N(0, \tau_{\beta_j}^2)$ and the generalized Pareto prior of the variance parameters the marginal prior of the regression coefficients is obtained as the integral

$$
\begin{aligned}
\pi\left(\beta_j \mid h_{1,\lambda}, h_{2,\lambda}\right) &= \int_0^\infty N\left(\beta_j \mid 0, \tau_{\beta_j}^2\right) \text{gPareto}\left(\tau_{\beta_j}^2 \mid h_{1,\lambda}, h_{2,\lambda}\right) d\tau_{\beta_j}^2 \\
&= \frac{h_{1,\lambda}}{2h_{2,\lambda}\sqrt{2\pi}} \int_0^\infty \tau_{\beta_j}^{-1} \exp\left(-\frac{\beta_j^2}{2\tau_{\beta_j}^2}\right) \left[\frac{\tau_{\beta_j}^2}{2h_{2,\lambda}} + 1\right]^{-(h_{1,\lambda}+1)} d\tau_{\beta_j}^2 .
\end{aligned}
$$

With the substitution $x_j = 1/\tau_{\beta_j}^2$ we get

$$
\begin{aligned}
p\left(\beta_j \mid h_{1,\lambda}, h_{2,\lambda}\right) &= \frac{h_{1,\lambda}}{2h_{2,\lambda}\sqrt{2\pi}} \int_0^\infty \left(x_j\right)^{-\frac{3}{2}} \exp\left(-\frac{\beta_j^2}{2} x_j\right) \left[\frac{1}{2h_{2,\lambda} x_j} + 1\right]^{-(h_{1,\lambda}+1)} dx_j \\
&= \frac{h_{1,\lambda}}{2h_{2,\lambda}\sqrt{2\pi}} \int_0^\infty \left(x_j\right)^{h_{1,\lambda}-\frac{1}{2}} \exp\left(-\frac{\beta_j^2}{2} x_j\right) \left[x_j + \frac{1}{2h_{2,\lambda}}\right]^{-(h_{1,\lambda}+1)} dx_j .
\end{aligned} \tag{B.5}
$$

If we use the representation (B.3) of the parabolic cylinder function $D_{-2\nu}(\cdot)$ with the parameters $m = \beta_j^2/2$, $\nu = h_{1,\lambda} + 1/2$ and $y = 1/2h_{2,\lambda}$, we obtain for the marginal densities of the regression coefficients

$$
p\left(\beta_j \mid h_{1,\lambda}, h_{2,\lambda}\right) = \frac{h_{1,\lambda}}{\sqrt{\pi}} \frac{2^{h_{1,\lambda}}}{\sqrt{2h_{2,\lambda}}} \Gamma\left(h_{1,\lambda} + \tfrac{1}{2}\right) \exp\left(\frac{1}{4} \frac{\beta_j^2}{2h_{2,\lambda}}\right) D_{-2(h_{1,\lambda}+1/2)}\left(\frac{|\beta_j|}{\sqrt{2h_{2,\lambda}}}\right).
$$

Derivate of marginal log-prior of the regression coefficients

To obtain the derivate

$$\frac{d}{d\beta_j}\log p\left(\beta_j \mid h_{1,\lambda},h_{2,\lambda}\right) = \frac{\frac{d}{d\beta_j}p\left(\beta_j \mid h_{1,\lambda},h_{2,\lambda}\right)}{p\left(\beta_j \mid h_{1,\lambda},h_{2,\lambda}\right)},$$

we have to evaluate the derivate in the nominator. Using the expression in (B.5) we obtain

$$\frac{d}{d\beta_j}p\left(\beta_j \mid h_{1,\lambda},h_{2,\lambda}\right) = -\frac{h_{1,\lambda}\beta_j}{2h_{2,\lambda}\sqrt{2\pi}}\int_0^\infty \left(x_j\right)^{h_{1,\lambda}+\frac{1}{2}}\exp\left(-\frac{\beta_j^2}{2}x_j\right)\left[x_j+\frac{1}{2h_{2,\lambda}}\right]^{-(h_{1,\lambda}+1)}dx_j.$$

If we use the representation (B.4) of the parabolic cylinder function $D_{-2v+1}(\cdot)$ with the parameters $m=\beta_j^2/2$, $\nu=h_{1,\lambda}+3/2$ and $y=1/2h_{2,\lambda}$, we obtain

$$\frac{d}{d\beta_j}p\left(\beta_j \mid h_{1,\lambda},h_{2,\lambda}\right) = -\frac{\beta_j}{|\beta_j|}\frac{1}{\sqrt{\pi}}\frac{h_{1,\lambda}}{2h_{2,\lambda}}2^{h_{1,\lambda}+1}\Gamma\left(h_{1,\lambda}+\tfrac{3}{2}\right)\exp\left(\frac{1}{4}\frac{\beta_j^2}{2h_{2,\lambda}}\right)D_{-2h_{1,\lambda}-2}\left(\frac{|\beta_j|}{\sqrt{2h_{2,\lambda}}}\right).$$

Finally, we obtain with $\Gamma\left(h_{1,\lambda}+\frac{3}{2}\right)=(h_{1,\lambda}+\frac{1}{2})\Gamma\left(h_{1,\lambda}+\frac{1}{2}\right)$ from (B.1) the result

$$\frac{d}{d\beta_j}\log p\left(\beta_j \mid h_{1,\lambda},h_{2,\lambda}\right) = -\frac{\beta_j}{|\beta_j|}\frac{2h_{1,\lambda}+1}{\sqrt{2h_{2,\lambda}}}\frac{D_{-2h_{1,\lambda}-2}\left(\dfrac{|\beta_j|}{\sqrt{2h_{2,\lambda}}}\right)}{D_{-2h_{1,\lambda}-1}\left(\dfrac{|\beta_j|}{\sqrt{2h_{2,\lambda}}}\right)}$$

and the derivate of penalty function reads

$$\text{pen}'\left(|\beta_j|;h_{1,\lambda},h_{2,\lambda}\right) = \frac{\left(2h_{1,\lambda}+1\right)}{\sqrt{2h_{2,\lambda}}}\frac{D_{-2\left(h_{1,\lambda}+1\right)}\left(\left|\beta_j\right|/\sqrt{2h_{2,\lambda}}\right)}{D_{-2\left(h_{1,\lambda}+1/2\right)}\left(\left|\beta_j\right|/\sqrt{2h_{2,\lambda}}\right)}.$$

Using the connection $D_\nu(0)=2^{\frac{\nu}{2}}\pi^{\frac{1}{2}}\Gamma^{-1}(\frac{1-\nu}{2})$ we obtain at the origin the penalty

$$\text{pen}'\left(0 \mid h_{1,\lambda},h_{2,\lambda}\right) = \frac{2h_{1,\lambda}+1}{\sqrt{2h_{2,\lambda}}}\frac{D_{-2h_{1,\lambda}-2}(0)}{D_{-2h_{1,\lambda}-1}(0)} = \frac{1}{\sqrt{h_{2,\lambda}}}\frac{\Gamma(h_{1,\lambda}+1)}{\Gamma(h_{1,\lambda}+\frac{1}{2})}.$$

B.3. Bayesian NMIG prior

Conditional distribution of the variance parameter

With the hierarchy of the Bayesian MNIG prior from Section 4.3.1 we obtain for the variance parameters $\tau_{\beta_j}^2=I_j\psi_j^2$, as the product of a Bernoulli distributed indicator I_j and an inverse gamma distributed variance ψ_j^2, the densities

$$\begin{aligned}
p\left(I_j\psi_j^2 \mid v_0,v_1,\omega,h_{1,\psi},h_{2,\psi}\right) &= \int\frac{1}{|x|}\left[(1-\omega)^{\delta_{v_0}(x)}\omega^{\delta_{v_1}(x)}\right]\text{IGamma}\left(\frac{\tau_j^2}{x} \mid h_{1,\psi},h_{2,\psi}\right)dx \\
&= \frac{1-\omega}{v_0}\text{IGamma}\left(\frac{\tau_j^2}{v_0} \mid h_{1,\psi},h_{2,\psi}\right)+\frac{\omega}{v_1}\text{IGamma}\left(\frac{\tau_j^2}{v_1} \mid h_{1,\psi},h_{2,\psi}\right) \\
&= \left(1-\omega\right)\cdot\text{IGamma}\left(\tau_j^2 \mid h_{1,\psi},v_0h_{2,\psi}\right)+\omega\cdot\text{IGamma}\left(\tau_j^2 \mid h_{1,\psi},v_1h_{2,\psi}\right),
\end{aligned}$$

which are mixtures of scaled inverse gamma distributions with common shape parameter $h_{1,\psi}$ and the scale parameters $v_0 h_{2,\psi}$ and $v_1 h_{2,\psi}$.

Marginal distribution of the variance parameters

The marginal densities of the variance parameters are obtained by marginalization of the complexity parameter which is equipped with a beta prior, i. e. $\omega \sim \text{Beta}(h_{1,\omega}, h_{2,\omega})$:

$$\begin{aligned}
p(\tau_j^2 \mid \ldots) &= \frac{\Gamma(h_{1,\omega}+h_{2,\omega})}{\Gamma(h_{1,\omega})\Gamma(h_{s,\omega})} \int p\left(I_j \psi_j^2 \mid v_0, v_1, \omega, h_{1,\psi}, h_{2,\psi}\right) \omega^{h_{1,\omega}-1}(1-\omega)^{h_{2,\omega}-1} 1_{[0,1]}(\omega) d\omega \\
&= \frac{\Gamma(h_{1,\omega}+h_{2,\omega})}{\Gamma(h_{1,\omega})\Gamma(h_{s,\omega})} \text{IGamma}\left(\tau_j^2 \mid h_{1,\psi}, v_0 h_{2,\psi}\right) \int \omega^{h_{1,\omega}-1}(1-\omega)^{(h_{2,\omega}+1)-1} 1_{[0,1]}(\omega) d\omega \\
&\quad + \frac{\Gamma(h_{1,\omega}+h_{2,\omega})}{\Gamma(h_{1,\omega})\Gamma(h_{2,\omega})} \text{IGamma}\left(\tau_j^2 \mid h_{1,\psi}, v_1 h_{2,\psi}\right) \int \omega^{(h_{1,\omega}+1)-1}(1-\omega)^{h_{2,\omega}-1} 1_{[0,1]}(\omega) d\omega \\
&= \frac{\Gamma(h_{1,\omega}+h_{2,\omega})}{\Gamma(h_{1,\omega})\Gamma(h_{2,\omega})} \frac{\Gamma(h_{1,\omega})\Gamma(h_{2,\omega}+1)}{\Gamma(h_{1,\omega}+h_{2,\omega}+1)} \text{IGamma}\left(\tau_j^2 \mid h_{1,\psi}, v_0 h_{2,\psi}\right) \\
&\quad + \frac{\Gamma(h_{1,\omega}+h_{2,\omega})}{\Gamma(h_{1,\omega})\Gamma(h_{2,\omega})} \frac{\Gamma(h_{1,\omega}+1)\Gamma(h_{2,\omega})}{\Gamma(h_{1,\omega}+h_{2,\omega}+1)} \text{IGamma}\left(\tau_j^2 \mid h_{1,\psi}, v_1 h_{2,\psi}\right) \\
&= \frac{h_{2,\omega}}{h_{1,\omega}+h_{2,\omega}} \text{IGamma}\left(\tau_j^2 \mid h_{1,\psi}, v_0 h_{2,\psi}\right) + \frac{h_{1,\omega}}{h_{1,\omega}+h_{2,\omega}} \text{IGamma}\left(\tau_j^2 \mid h_{1,\psi}, v_1 h_{2,\psi}\right).
\end{aligned}$$

In the last conversion we use the connection (B.1). In summary, the marginal distribution is a mixture of inverse gamma distributions with common shape parameter $h_{1,\psi}$ and scale parameters $v_0 h_{2,\psi}$ and $v_1 h_{2,\psi}$

$$\tau_{\beta_j}^2 \mid \ldots \sim \frac{h_{2,\omega}}{h_{1,\omega}+h_{2,\omega}} \cdot \text{IGamma}(h_{1,\psi}, v_0 h_{2,\psi}) + \frac{h_{1,\omega}}{h_{1,\omega}+h_{2,\omega}} \cdot \text{IGamma}(h_{1,\psi}, v_1 h_{2,\psi}) \,. \tag{B.6}$$

Marginal distribution of the regression coefficients

With the Gaussian prior of the regression coefficients $\beta_j \mid \tau_{\beta_j}^2 \sim N(0, \tau_{\beta_j}^2)$ and with the marginal mixture prior of the variance parameters (B.6) we obtain with similar conversions as in Subsection B.1, the marginal prior of the regression coefficients as mixture of two scaled Student t-distributions with $d = 2h_{1,\lambda}$ degrees of freedom and scale parameters $s_0 = \sqrt{v_0 h_{2,\psi}/h_{1,\psi}}$ and $s_1 = \sqrt{v_1 h_{2,\psi}/h_{1,\psi}}$

$$\beta_j \mid \cdot \sim \frac{h_{2,\omega}}{h_{1,\omega}+h_{2,\omega}} t\left(d = 2h_{1,\psi}, s = \sqrt{\frac{v_0 h_{2,\psi}}{h_{1,\psi}}} \right) + \frac{h_{1,\omega}}{h_{1,\omega}+h_{2,\omega}} t\left(d = 2h_{1,\psi}, s = \sqrt{\frac{v_1 h_{2,\psi}}{h_{1,\psi}}} \right).$$

C Taylor expansion of second order

General approach

Let $f : \mathbb{R}^p \to \mathbb{R}$, $f(\boldsymbol{\theta}) = f(\theta_1, \ldots, \theta_p)$ denote a real valued, two times continuous differentiable function and let $\boldsymbol{\theta}^{(c)} = (\theta_1^{(c)}, \ldots, \theta_p^{(c)})'$ denote the current state of the Markov chain. The quadratic approximation $\hat{f}(\cdot)$ to the function $f(\cdot)$ at the current state is obtained by second order Taylor expansion of the function $f(\cdot)$ with respect to the current state of the chain $\boldsymbol{\theta}^{(c)}$, which is given as

$$\hat{\mathrm{f}}(\boldsymbol{\theta}) \approx \mathrm{f}\left(\boldsymbol{\theta}^{(c)}\right)+\left(\boldsymbol{\theta}-\boldsymbol{\theta}^{(c)}\right)'\mathbf{s}_{\theta}\left(\boldsymbol{\theta}^{(c)}\right)+\frac{1}{2}\left(\boldsymbol{\theta}-\boldsymbol{\theta}^{(c)}\right)'\mathbf{H}_{\theta}\left(\boldsymbol{\theta}^{(c)}\right)\left(\boldsymbol{\theta}-\boldsymbol{\theta}^{(c)}\right) \qquad \text{(C.1)}$$

with the score vector and hessian matrix defined as the derivates

$$\mathbf{s}_{\theta}\left(\boldsymbol{\theta}^{(c)}\right):=\frac{\partial \mathrm{f}\left(\boldsymbol{\theta}^{(c)}\right)}{\partial\boldsymbol{\theta}}, \quad \mathbf{H}_{\theta}\left(\boldsymbol{\theta}^{(c)}\right):=\frac{\partial^2 \mathrm{f}\left(\boldsymbol{\theta}^{(c)}\right)}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}'}.$$

If the components of the approximation $\hat{\mathrm{f}}(\boldsymbol{\theta})$ that do not depend on $\boldsymbol{\theta}$ are omitted, the exponential function of the approximation

$$\begin{aligned}\exp\left(\hat{\mathrm{f}}(\boldsymbol{\theta})\right) &\propto \exp\left(\boldsymbol{\theta}'\mathbf{s}_{\theta}\left(\boldsymbol{\theta}^{(c)}\right)+\frac{1}{2}\boldsymbol{\theta}'\mathbf{H}_{\theta}\left(\boldsymbol{\theta}^{(c)}\right)\boldsymbol{\theta}-\boldsymbol{\theta}'\mathbf{H}_{\theta}\left(\boldsymbol{\theta}^{(c)}\right)\boldsymbol{\theta}^{(c)}\right)\\ &=\exp\left(-\frac{1}{2}\boldsymbol{\theta}'\left(-\mathbf{H}_{\theta}\left(\boldsymbol{\theta}^{(c)}\right)\right)\boldsymbol{\theta}+\boldsymbol{\theta}'\left(\mathbf{s}_{\theta}\left(\boldsymbol{\theta}^{(c)}\right)-\mathbf{H}_{\theta}\left(\boldsymbol{\theta}^{(c)}\right)\boldsymbol{\theta}^{(c)}\right)\right)\end{aligned}$$

is proportional to the density of a multivariate Gaussian distribution with mean vector $\mathbb{E}(\boldsymbol{\theta}\,|\,\boldsymbol{\theta}^{(c)})=\hat{\boldsymbol{\mu}}_{\boldsymbol{\theta}}^{(c)}$ and precision matrix $\mathrm{Prec}(\boldsymbol{\theta}\,|\,\boldsymbol{\theta}^{(c)})=\hat{\boldsymbol{\Pi}}_{\boldsymbol{\theta}}^{(c)}$ given by

$$\hat{\boldsymbol{\mu}}_{\boldsymbol{\theta}}=\left(\boldsymbol{\Pi}_{\boldsymbol{\theta}}^{(c)}\right)^{-1}\left(\mathbf{s}_{\theta}\left(\boldsymbol{\theta}^{(c)}\right)-\mathbf{H}_{\theta}\left(\boldsymbol{\theta}^{(c)}\right)\boldsymbol{\theta}^{(c)}\right), \quad \hat{\boldsymbol{\Pi}}_{\boldsymbol{\theta}}^{(c)}=-\mathrm{H}_{\theta}\left(\boldsymbol{\theta}^{(c)}\right).$$

In the case of an improper Gaussian distribution, i. e. the Hessian matrix is not of full rank, the expression $(\boldsymbol{\Pi}_{\boldsymbol{\theta}}^{(c)})^{-1}$ denotes a generalized inverse of the precision matrix. In terms of the covariance matrix $\mathbb{C}\mathrm{ov}(\boldsymbol{\theta}\,|\,\boldsymbol{\theta}^{(c)})=\hat{\boldsymbol{\Sigma}}_{\boldsymbol{\theta}}^{(c)}=(\hat{\boldsymbol{\Pi}}_{\boldsymbol{\theta}}^{(c)})^{-1}$ we write

$$\hat{\boldsymbol{\mu}}_{\boldsymbol{\theta}}^{(c)}=\boldsymbol{\theta}^{(c)}+\hat{\boldsymbol{\Sigma}}_{\boldsymbol{\theta}}^{(c)}\mathbf{s}_{\theta}\left(\boldsymbol{\theta}^{(c)}\right), \quad \hat{\boldsymbol{\Sigma}}_{\boldsymbol{\theta}}^{(c)}=-\mathbf{H}_{\theta}^{-1}\left(\boldsymbol{\theta}^{(c)}\right) \qquad \text{(C.2)}$$

and $\mathrm{direc}(\hat{\boldsymbol{\mu}}_{\boldsymbol{\theta}}^{(c)},\boldsymbol{\theta}^{(c)}):=-\mathbf{H}_{\theta}^{-1}(\boldsymbol{\theta}^{(c)})\mathbf{s}_{\theta}(\boldsymbol{\theta}^{(c)})=\hat{\boldsymbol{\mu}}_{\boldsymbol{\theta}}^{(c)}-\boldsymbol{\theta}^{(c)}$ denotes the difference vector between the current state and the approximated mean vector.

Connection to Fisher-scoring

From the representation of the mean vector in (C.2) we can see the close connection to the Fisher-scoring algorithm. If the function $\mathrm{f}(\boldsymbol{\theta})=\log(\mathrm{L}(\boldsymbol{\theta}))$ denotes the log-likelihood function of the parameter $\boldsymbol{\theta}$ and we want to maximize the function $\mathrm{f}(\cdot)$, then we try to find the root of the score function, i. e. $\mathbf{s}_{\theta}(\hat{\boldsymbol{\theta}})=0$. With the first order Taylor expansion to the score function $\hat{\mathrm{s}}_{\theta}(\boldsymbol{\theta})\approx\mathrm{s}_{\theta}(\boldsymbol{\theta}^{(c)})-\mathbf{H}_{\theta}(\boldsymbol{\theta}^{(c)})(\boldsymbol{\theta}-\boldsymbol{\theta}^{(c)})$ the problem reads $\hat{\boldsymbol{\theta}}\approx\boldsymbol{\theta}^{(c)}-\mathbf{H}_{\theta}^{-1}(\boldsymbol{\theta}^{(c)})\mathrm{s}_{\theta}(\boldsymbol{\theta}^{(c)})$. Starting with an appropriate value $\boldsymbol{\theta}^{(0)}$, one iteratively compute the values

$$\boldsymbol{\theta}^{(c+1)}=\boldsymbol{\theta}^{(c)}-\mathbf{H}_{\theta}^{-1}\left(\boldsymbol{\theta}^{(c)}\right)\mathrm{s}_{\theta}\left(\boldsymbol{\theta}^{(c)}\right)$$

until the algorithm converges to the desired solution $\hat{\boldsymbol{\theta}}$, which also maximizes the likelihood $\mathrm{L}(\boldsymbol{\theta})=\exp(\mathrm{f}(\boldsymbol{\theta}))$. Thus the mean vector $\hat{\boldsymbol{\mu}}_{\boldsymbol{\theta}}^{(c)}$ of (C.2) can be interpreted as a one step Fisher-scoring approximation to the mode (i. e. the maximum) of the function $\exp(\mathrm{f}(\boldsymbol{\theta}))$ in the direction $\mathrm{direc}(\hat{\boldsymbol{\mu}}_{\boldsymbol{\theta}}^{(c)},\boldsymbol{\theta}^{(c)})=\hat{\boldsymbol{\mu}}_{\boldsymbol{\theta}}^{(c)}-\boldsymbol{\theta}^{(c)}$.

Additional quadratic penalty term

In the presence of an additional quadratic penalty term from a zero mean multivariate Gaussian distribution, i. e. $\mathrm{f}_{\mathrm{pen}}(\boldsymbol{\theta})=\mathrm{f}(\boldsymbol{\theta})+0.5\boldsymbol{\theta}'\boldsymbol{\Sigma}_{\theta}^{-1}\boldsymbol{\theta}$ we get accordingly using the Taylor expansion of $\mathrm{f}(\cdot)$ from (C.1)

$$\begin{aligned}\exp\left(\hat{\mathrm{f}}_{\mathrm{pen}}(\boldsymbol{\theta})\right) &\propto \exp\left(\boldsymbol{\theta}'\mathbf{s}_{\theta}\left(\boldsymbol{\theta}^{(c)}\right)+\frac{1}{2}\boldsymbol{\theta}'\mathbf{H}_{\theta}\left(\boldsymbol{\theta}^{(c)}\right)\boldsymbol{\theta}-\boldsymbol{\theta}'\mathbf{H}_{\theta}\left(\boldsymbol{\theta}^{(c)}\right)\boldsymbol{\theta}^{(c)}-\frac{1}{2}\boldsymbol{\theta}'\boldsymbol{\Sigma}_{\theta}^{-1}\boldsymbol{\theta}\right)\\ &=\exp\left(-\frac{1}{2}\boldsymbol{\theta}'\left(-\mathbf{H}_{\theta}\left(\boldsymbol{\theta}^{(c)}\right)+\boldsymbol{\Sigma}_{\theta}^{-1}\right)\boldsymbol{\theta}+\boldsymbol{\theta}'\left(\mathbf{s}_{\theta}\left(\boldsymbol{\theta}^{(c)}\right)-\mathbf{H}_{\theta}\left(\boldsymbol{\theta}^{(c)}\right)\boldsymbol{\theta}^{(c)}\right)\right).\end{aligned}$$

This is also a multivariate Gaussian density with mean vector $\mathbb{E}_{\mathrm{pen}}(\boldsymbol{\theta}\,|\,\boldsymbol{\theta}^{(c)})=\hat{\boldsymbol{\mu}}^{(c)}_{\boldsymbol{\theta},\mathrm{pen}}$ and covariance matrix $\mathbb{C}\mathrm{ov}_{\mathrm{pen}}(\boldsymbol{\theta}\,|\,\boldsymbol{\theta}^{(c)})=\hat{\boldsymbol{\Sigma}}^{(c)}_{\boldsymbol{\theta},\mathrm{pen}}$ given as

$$\hat{\boldsymbol{\mu}}^{(c)}_{\boldsymbol{\theta},\mathrm{pen}}=\hat{\boldsymbol{\Sigma}}^{(c)}_{\boldsymbol{\theta},\mathrm{pen}}\left(\mathbf{s}_{\theta}\left(\boldsymbol{\theta}^{(c)}\right)-\mathbf{H}_{\theta}\left(\boldsymbol{\theta}^{(c)}\right)\boldsymbol{\theta}^{(c)}\right),\quad \hat{\boldsymbol{\Sigma}}^{(c)}_{\boldsymbol{\theta},\mathrm{pen}}=\left(-\mathbf{H}_{\theta}\left(\boldsymbol{\theta}^{(c)}\right)+\boldsymbol{\Sigma}_{\theta}^{-1}\right)^{-1}. \qquad \text{(C.3)}$$

Approach with penalized score vector and Hessian matrix

The same result as in (C.3) is achieved, if the quadratic penalty is included in penalized score vector and the penalized Hessian matrix while accomplishing the Taylor expansion to $\mathrm{f}_{\mathrm{pen}}(\boldsymbol{\theta})$. The penalized score vector and penalized Hessian matrix are then given as

$$\mathbf{s}_{\theta,\mathrm{pen}}\left(\boldsymbol{\theta}^{(c)}\right):=\frac{\partial \mathrm{f}_{\mathrm{pen}}\left(\boldsymbol{\theta}^{(c)}\right)}{\partial\boldsymbol{\theta}}=\mathbf{s}_{\theta}\left(\boldsymbol{\theta}^{(c)}\right)-\boldsymbol{\Sigma}_{\theta}^{-1}\boldsymbol{\theta}^{(c)},\quad \mathbf{H}_{\theta,\mathrm{pen}}\left(\boldsymbol{\theta}^{(c)}\right):=\frac{\partial^{2}\mathrm{f}_{\mathrm{pen}}\left(\boldsymbol{\theta}^{(c)}\right)}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}'}=\mathbf{H}_{\theta}\left(\boldsymbol{\theta}^{(c)}\right)-\boldsymbol{\Sigma}_{\theta}^{-1},$$

and with

$$\exp\left(\hat{\mathrm{f}}_{\mathrm{pen}}(\boldsymbol{\theta})\right)\propto\exp\left(-\frac{1}{2}\boldsymbol{\theta}'\left(-\mathbf{H}_{\theta,\mathrm{pen}}\left(\boldsymbol{\theta}^{(c)}\right)\right)\boldsymbol{\theta}+\boldsymbol{\theta}'\left(\mathbf{s}_{\theta,\mathrm{pen}}\left(\boldsymbol{\theta}^{(c)}\right)-\mathbf{H}_{\theta,\mathrm{pen}}\left(\boldsymbol{\theta}^{(c)}\right)\boldsymbol{\theta}^{(c)}\right)\right)$$

we get the mean vector and the covariance matrix of the corresponding multivariate Gaussian distribution as

$$\hat{\boldsymbol{\mu}}^{(c)}_{\boldsymbol{\theta},\mathrm{pen}}=\hat{\boldsymbol{\Sigma}}^{(c)}_{\boldsymbol{\theta},\mathrm{pen}}\left(\mathbf{s}_{\theta,\mathrm{pen}}\left(\boldsymbol{\theta}^{(c)}\right)-\mathbf{H}_{\theta,\mathrm{pen}}\left(\boldsymbol{\theta}^{(c)}\right)\boldsymbol{\theta}^{(c)}\right),\quad \hat{\boldsymbol{\Sigma}}^{(c)}_{\boldsymbol{\theta},\mathrm{pen}}=-\mathbf{H}^{-1}_{\theta,\mathrm{pen}}\left(\boldsymbol{\theta}^{(c)}\right), \qquad \text{(C.4)}$$

which coincides with the representations in (C.3). We can also write the mean as $\hat{\boldsymbol{\mu}}^{(c)}_{\boldsymbol{\theta},\mathrm{pen}}=\boldsymbol{\theta}^{(c)}+\hat{\boldsymbol{\Sigma}}^{(c)}_{\boldsymbol{\theta},\mathrm{pen}}\mathbf{s}_{\theta,\mathrm{pen}}\left(\boldsymbol{\theta}^{(c)}\right)$, with the difference vector between the current state and the approximated mean vector $\mathrm{direc}(\hat{\boldsymbol{\mu}}^{(c)}_{\boldsymbol{\theta},\mathrm{pen}},\boldsymbol{\theta}^{(c)}):=\hat{\boldsymbol{\Sigma}}^{(c)}_{\boldsymbol{\theta},\mathrm{pen}}\mathbf{s}_{\theta,\mathrm{pen}}\left(\boldsymbol{\theta}^{(c)}\right)=\hat{\boldsymbol{\mu}}^{(c)}_{\boldsymbol{\theta},\mathrm{pen}}-\boldsymbol{\theta}^{(c)}$.

D BayesX methods and R functions

Simulation studies and data analysis is carried out with the free software `R` and `BayesX`. The sources of the software and references, to obtain methodological or implementational details of the used procedures, are listed in the Reference Section.

D.1. BayesX methods

regress: We use the method `regress` implemented in the software tool `BayesX` (Belitz, C., Brezger, A., Kneib, T., Lang, S. and Umlauf, N.) to fit the regularized CRR-type regression models based on the full likelihood. The implemented MCMC simulation techniques are described in Section 9.1. In general `BayesX` supports the estimation of structured additive regression models like generalized additive models, generalized additive mixed models, generalized geoadditive mixed models, dynamic models, varying coefficient models, as well as the regression for categorical responses, hazard regression for continuous survival times and continuous time multi-state models within a unifying framework. The method `regress` is extended to consider the shrinkage priors of Section 4.1 to Section 4.3. For details of the methodological background we refer to the `BayesX` homepage, where a complete list of references is available to download. The provided `BayesX` reference manual contains a detailed description of the `BayesX` commands, in particular the subsection *"Shrinkage of fixed effects"* covers the syntax for the Bayesian lasso, ridge and NMIG penalty. Below we shortly summarize the syntax for the regularization of the linear effects for the `BayesX 2.1` version.

D.2. R functions

penalized: The `penalized()` (Goeman, J. J.) function of the `R`-package `{penalized}` is used to fit frequentist penalized Cox regression models based on the partial likelihood with the lasso or ridge penalty for the linear effects. The implemented algorithm for maximizing the penalized partial likelihood follows the full gradient of the likelihood from a given starting value of the regression coefficients at each step of the maximization, and switches to faster Newton-Raphson steps when it gets close to the optimum. The tuning (resp. shrinkage) parameter λ is determined by Cross-validation as implemented in the package functions `optL1()` and `profL1()`. In particular for the Cox model the cross-validated partial likelihood of Verweij and van Houwelingen (1993) builds the base for choosing the tuning parameter. As recommended by the author of the package, we use in our simulations and applications the function `optL1()` in combination with the function `profL1()` to validate if `optL1()` has converged to the desired optimum.

bayessurv2: As Bayesian competitor to the extended AFT model we used the function `bayessurv2()` of the `R`-package `{bayesSurv}` (Komárek, A.). In this function the error distribution is also expressed as a penalized univariate Gaussian mixture with a finite fixed number of components. The function supports the estimation of unregularized linear effects or random effects. The results for *intercept1* und *scale1* from the generated file `gspline.sim` represent the samples of the location component γ_0 and the scale component σ of the AFT model with respect to the unstandardized PGM. No identification constraints for the location and the scale parameter are implemented and therefore, the trace-plots of the parameters *intercept1* and *scale1* do not show stationarity anyhow. But the file `mixmoment.sim`, which contains the mean and variance of the baseline error density $f_{Y_0}(y)$, $Y_0 := \gamma_0 + \sigma\varepsilon$, can be used to check the stationarity via the stationarity of the moment estimates. A recomputation of the weights to show the stationarity is not supported.

pendensity: In the simulation section we also use the function `pendensity()` of the `R`-package `{pendensity}` (Schellhase, C) for frequentist estimation the error distribution density in the AFT model. This function is designed for the estimation of penalized densities using P-splines, with Gaussian or B-spline basis functions, and also allows for the inclusion factor covariates. We only applied the function to uncensored data without covariates to provide a frequentist competitor for the results from the Bayesian methods. For details of the specific `R`-package functions we refer to the corresponding help files of the package documentation.

baftpgm and bcoxpl : The described Bayesian approach to fit the Bayesian AFT models with extended predictor and flexible error distribution is implemented in `R`-function `baftpgm()`. The corresponding function for estimation of the Bayesian CRR model with extended predictor based on the partial likelihood is implemented in `R`-function `bcoxpl()`. The following `R`-functions are used within the functions `baftpgm()` or `bcoxpl()`.

`uni.slice()`, `ars{ars}` for an optional single update of the transformed error weights of the PGM error density in the AFT model using slice sampling (`uni.slice`) or adaptive rejection sampling (`ars`).

`tnorm{msm}` to generate truncated normal random variables to impute the latent exact survival times in the AFT model with PGM error.

`rinvGauss{SuppDists}` to generate inverse Gaussian random variables to update the variance parameter $\tau^2_{\beta_j}$ of the Bayesian lasso regularized linear effects.

`rinvgamma{MCMCpack}` to generate inverse gamma distributed random variables to update the variance component ψ^2_j of the Bayesian NMIG regularized linear effects.

`rmvnorm{mvtnorm}` to generate multivariate normal random variables of the regularized linear effects β_j. The procedures to determine the matrix root are based on the eigenvalue decomposition (default), the singular value decomposition and the Cholesky decomposition of the covariance matrix.

`rdiric{VGAM}` to generate Dirichlet random variables.

D.3. Arguments of the BayesX method regress

The `BayesX` syntax of the method `regress` is extended to consider linear effects that are equipped with the lasso, ridge or NMIG shrinkage-prior. For the $p \geq 1$ regularized linear effects γ_j, $j = 1,...,p$, of the corresponding covariates $X1,...,Xp$ the linear predictor components is given as

$$\eta = ... + \gamma_1 X1 + ... + \gamma_p Xp + ...$$

The specific `BayesX 2.1` syntax of the individual model terms for the linear covariate effects has the general form

Ridge-prior: `X1(ridge[,options])+...+ Xp(ridge[,options])`

Lasso-prior: `X1(lasso[,options]) +...+ Xp(lasso[,options])`

NMIG-prior: `X1(nigmix[,options]) +...+ Xp(nigmix[,options])`

with the following shrinkage-prior specific options:

Optional arguments for lasso and ridge terms

`a, b` (*)

Non-negative, real values, to specify the hyperparameters $\mathtt{a} \mathrel{\hat{=}} h_{1,\lambda} \geq 0$ and $\mathtt{b} \mathrel{\hat{=}} h_{2,\lambda} \geq 0$ of the inverse gamma prior of the shrinkage parameter λ. This option is specified in the first lasso/ridge model term of the predictor.

Default value: `a=0.001, b=0.001`

`adaptive`

Logical value, that specifies the adaptive version of the shrinkage priors, i. e. an individual shrinkage parameter λ_j for each covariate effect is estimated. This option is specified in the first shrinkage model term.

Default value: `false`

`effect`

Real value, that enables to specify the starting value of the linear effect γ_j. As default, the starting values are set to the posterior mode, which are initially computed via backfitting within Fisher scoring steps, compare the BayesX methodology manual for details. If a large penalty for the regularized effects is used, e. g. when the paths of the regression coefficients are computed, the implemented computation of the starting values fails sometimes and the sampler does not start. In such a situation the external specification of small starting values for the regularized effects overcomes this problem.

Default value: `-`

`shrinkage` (*)

Non-negative real value, that specifies the starting value of the shrinkage parameter λ. This option is specified in the first `lasso/ridge` model term of the predictor.

Default value: `shrinkage=1`

`shrinkagefix`

Logical value, that specifies, if the shrinkage parameter λ should be fixed at the value specified in option `shrinkage`. The shrinkage parameter is treated as fixed, if this option is set in the first `lasso/ridge` model term.

Default value: `false`

`tau2`

Positive real value, that specifies the starting value of the variance parameter $\tau^2_{\beta_j}$. Values have to be set in each `lasso/ridge` model term if the default starting value should be modified.

Default value: `tau2=0.1`

Optional arguments for the nigmix terms

`a, b` (*)

Non-negative real values, to specify the hyperparameters $\mathtt{a} \mathrel{\hat{=}} h_{1,\psi} \geq 0$ and $\mathtt{b} \mathrel{\hat{=}} h_{2,\psi} \geq 0$ of the inverse gamma prior of the variance component parameter ψ^2_j. This option is specified in the first `nigmix` model term of the predictor.

Default value: `a=5, b=50`

`aw, bw` (*)

Non-negative real values, to specify the hyperparameters $\mathtt{aw} \mathrel{\hat{=}} h_{1,\omega} \geq 0$ and $\mathtt{bw} \mathrel{\hat{=}} h_{2,\omega} \geq 0$ of the beta prior for the complexity parameter ω. This option is specified in the first `nigmix` model term of the predictor.

Default value: `aw=1, bw=1`

`adaptive`

Logical value, that specifies the adaptive version of the shrinkage priors, i. e. an individual shrinkage parameter ω_j for each covariate effect is estimated. This option is specified in the first shrinkage model term.

Default value: `false`

`effect`

Real value, that enables to specify the starting value of the linear effect γ_j. As default, the starting values are set to the posterior mode, which are initially computed via backfitting within Fisher scoring steps, compare the BayesX methodology manual for details.

Default value: -

`I`

Sets the starting value of the indicator variable I_j. Values have to be 0 or 1 to set the indicator variable point mass at the values $v_0 > 0$ or $v_1 > 0$ and have to be set in each `nigmix` predictor term.

Default value: `I=1`

`t2`

Provides a starting value for the variance parameter ψ_j^2. Values have to be positive and have to be set in each `nigmix` predictor term.

Default value: `t2=11`

`v0`, `v1` (*)

Non-negative real values, to specify the point mass of the indicator variables at the values `v0`$\triangleq v_0 > 0$ or `v1`$\triangleq v_1 > 0$.

Default value: `v0=0.005`, `v1=1`

`w` (*)

Specifies the starting value of the complexity parameter ω. Values have to be in the interval (0,1).

Default value: `w=0.5`

`wfix`

Logical value, that specifies, if the complexity parameter ω should be fixed at the value specified in option `w`. The shrinkage parameter is treated as fixed, if this option is not omitted in the first `nigmix` model term.

Default value: `false`

The options marked with (*) can be specified in each shrinkage term if the adaptive versions of the penalties should be used.

As an example we consider two covariates `X1` and `X2`. If the predictor is written as `X2(lasso,shrinkagepar=2,shrinkagefix)+X1(lasso,shrinkagepar=1.5)` the procedure uses the options of the first lasso term given by `X2(lasso,...)`, i. e. the shrinkage parameter is fixed at the value 2. The option `shrinkagepar=1.5` of the second term is ignored. Since the remaining possible lasso options are not modified the default settings are used.

By default the shrinkage parameter is estimated as well as all other model parameters. It is also possible to fix the shrinkage parameter through the iterations in order to use a prespecified amount of shrinkage or to compute the parameter paths as function of the shrinkage parameter. We used this option to compute the Bayesian versions of the lasso and NMIG coefficient paths in the application section.

Resulting objects for lasso and ridge terms

The following listed files contain the associated results to the lasso regularization of the linear effects and are additionally generated by the use of the method `regress`. The prefix `*` denotes the replacement character for the user specified base-name prefix. If the covariates are regularized by the ridge prior, the filenames contain the string `ridge` instead of `lasso` with the results from the ridge regression in the corresponding files.

`*_f_lasso_hyperpar_startdata.raw`: Contains the used values of the options `a`, `b`, `shrinkagefix` and `adaptive` as explicit or implicit specified in the model terms. The values 0/1 of `shrinkagefix` and `adaptive` correspond to the logical values false/true. The values coincide per row in the non adaptive case.

`*_f_lasso_shrinkage.res`: Contains summary statistics of the marginal empirical posterior distribution of the `shrinkage` parameter λ like the posterior mean, standard deviation and quantiles.

`*_f_lasso_shrinkage_sample.raw`: Contains the sampled values of the `shrinkage` parameter λ. The columns of the file coincide in the non adaptive case.

`*_lasso_shrinkage_startdata.raw`: Contains the starting values of the `shrinkage` parameter λ as set in the options. In the non adaptive case the values for each variable coincide.

`*_f_lasso_var.res`: Contains summary statistics of the marginal empirical posterior distribution of the covariate specific variance parameters `tau2` like the posterior mean, standard deviation and quantiles.

`*_f_lasso_variance_sample.raw`: Contains the sampled values of the covariate specific variance parameters `tau2`.

`*_f_lasso_variance_startdata.raw`: Contains the starting values of the `tau2` parameters for each penalized covariate effect.

`*_lasso_Effects.res`: Contains summary statistics of the marginal empirical posterior distribution of the covariate effects β_j like the posterior mean, standard deviation and quantiles. If the number of regularized covariates is larger than the `blocksize` parameter (default value `blocksize=20`, compare the BayesX manual), the results are partitioned in different files, where each file contains the results of the covariates corresponding to one block with the size given in `blocksize`. The file names run from `*_lasso_Effects1.res`, `*_lasso_Effects2.res`, ... to the number of the resulting blocks.

`*_lasso_Effects_sample.raw`: Contains the sampled values of the covariate effects β_j. Files are partitioned in blocks as described in `*_lasso_Effects.res`.

`*_lasso_Effects_startdata.raw`: Contains the starting values of the covariate effects β_j if specified in the `effect` option. Files are partitioned in blocks as described in `*_lasso_Effects.res`.

Resulting objects for the nigmix terms

Using the NMIG prior the resulting additional files are:

`*_f_nigmix_hyperpar_startdata.raw`: Contains the used values of the options `v0`, `v1`, `a`, `b`, `aw`, `bw`, `wfix` and `adaptive` as explicit or implicit specified in the model terms. The values 0/1 of `wfix` and `adaptive` correspond to the logical values false/true. The values coincide per row in the non adaptive case.

`*_f_nigmix_shrinkage.res`: Contains summary statistics of the marginal empirical posterior distribution of the complexity parameter ω like the posterior mean, standard deviation and quantiles.

`*_f_nigmix_shrinkage_sample.raw`: Contains the sampled values of the complexity parameter ω. The columns of the file coincide in the non adaptive case.

`*f_nigmix_shrinkage_startdata.raw`: Contains the starting values of the complexity parameter ω as set in the options. In the non adaptive case the values for each variable coincide.

`*_f_nigmix_indicator.res`: Contains relative frequencies of the indicator variable value v_1.

`*_f_nigmix_indicator_sample.raw`: Contains the sampled values of the covariate specific variance parameter component I_j. The values 0 or 1 indicate point mass at the values $v_0 > 0$ or $v_1 > 0$.

`*_f_nigmix_indicator_startdata.raw`: Contains the starting values of the variance parameter component I_j for each penalized covariate effect.

`*_f_nigmix_t2.res`: Contains summary statistics of the marginal empirical posterior distribution of the covariate specific variance parameter component ψ_j^2 like the posterior mean, standard deviation and quantiles.

`*_f_nigmix_t2_sample.raw`: Contains the sampled values of the covariate specific variance parameter component ψ_j^2.

`*_f_nigmix_t2_startdata.raw`: Contains the starting values of the variance parameter component ψ_j^2 for each penalized covariate effect.

`*_f_nigmix_var.res`: Contains summary statistics of the marginal empirical posterior distribution of the covariate specific variance parameters $\tau_{\beta_j}^2 = I_j\psi_j^2$ like the posterior mean, standard deviation and quantiles.

`*_f_nigmix_variance_sample.raw`: Contains the sampled values of the covariate specific variance parameters $\tau_{\beta_j}^2 = I_j\psi_j^2$.

`*_f_nigmix_variance_startdata.raw`: Contains the starting values of the variance parameters $\tau_{\beta_j}^2 = I_j\psi_j^2$ for each penalized covariate effect.

`*_nigmix_Effects.res`: Contains summary statistics of the marginal empirical posterior distribution of the covariate effects β_j like the posterior mean, standard deviation and quantiles. Files are partitioned in blocks as described in `*_lasso_Effects.res`.

`*_nigmix_Effects_sample.raw`: Contains the sampled values of the covariate effects β_j. Files are partitioned in blocks as described in `*_lasso_Effects.res`.

`*_nigmix_Effects_startdata.raw`: Contains the starting values of the covariate effects β_j if specified in the `effect` option. Files are partitioned in blocks as described in `*_lasso_Effects.res`.

D.4. Arguments of the R-function bcoxpl

Usage

```
bcoxpl(dataset,unpenpri,penpri,splinepri,
        simpar=list(niter=10000,nthin=1,nburn=0,nwrite=10000,catniter=100),
        dir=list(outdir=getwd(),outnam="bcpl",overwrite=F),dirfunctions)
```

Arguments

`dataset`: a list, containing the data with the following components:

`data`

matrix containing, the variables in the model, i.e. the right censored survival times, the censoring indicator and the covariates that enter the predictor.

`time`

character string, specifying the column of the data frame specified in the argument `data` that is interpreted as the observed survival time.

`delta`

character string, specifying the column of the data frame specified in the argument `data` that is interpreted as the censoring indicator with values 0=alive and 1=death.

`unpenpri`: an optional list, to specify the parameters for the unpenalized linear effects:

`names`

a character vector, containing the column names of the `data` that identifies the covariates corresponding to the unregularized linear effects of the predictor.

`start.effect`

a numeric vector (with the same length as `names`), containing the initial values of the unpenalized effects.

`blocksize`

an integer, to define the size of the blocks for a simultaneous update of the corresponding unpenalized effects partitioned into these blocks. The value has to be less than or equal to the length of the `names` vector. Consider the situation with covariates $x_1, x_2, ..., x_{20}$ and the specification `blocksize=10`. Then the effects of the covariate blocks $x_1, x_2, ..., x_{10}$ and $x_{11}, x_{12}, ..., x_{20}$ are simultaneous updated. In each block-update the components of the full conditional corresponding to the blocks, which are not updated at this time, are discarded. If not specified, the blocksize is set to the minimum of the length of the `names` vector and `20`.

`randomblocks`

logical value, that indicates if the covariates are randomly assigned to the blocks for each iteration. If not specified, the value is set to `FALSE`.

`penpri`: an optional list, to specify the parameters of the regularized linear effects:

`type`

a character string, that is assumed to name an element from `"nigmix"`, `"ridge"`, `"lasso"`, `"adnigmix"`, `"adridge"` or `"adlasso"` to specify the NMIG, lasso ridge or the corresponding adaptive priors.

`names`

character vector, containing the column names of the `data` that identify the covariates corresponding to the effects that are regularized by the prior defined in `type`.

`blocksize`

the same explanation as in the `unpenpri` list.

`randomblocks`

the same explanation as in the `unpenpri` list.

`start.effect`

a numeric vector with the same length as `names`, containing the initial values of the regularized effects.

`v0, v1`

non-negative, numeric values, to specify the point mass of the NMIG-prior (`type=″nigmix″`) indicator variables I_j at the values `v0`$\hat{=} v_0 > 0$ or `v1`$\hat{=} v_1 > 0$. The default values are `v0=0.005` and `v1=1`.

`h1.t2, h2.t2`

both non-negative, numeric values, to specify the hyperparameters `h1.t2`$\hat{=} h_{1,\psi} \geq 0$ and `h2.t2`$\hat{=} h_{2,\psi} \geq 0$ of the inverse gamma prior for the variance component parameter ψ_j^2 if the NMIG-prior (`type=″nigmix″`) is selected. The default values are `h1.t2=5` and `h2.t2=5`.

`start.t2`

non-negative numeric vector with the same length as `names`, that provides the initial values for the variance component parameter ψ_j^2.

`start.I`

a positive numeric vector with the same length as `names`, that gives the initial value of the indicator variable I_j for the NMIG-prior (`type=″nigmix″`). Values have to be 0 or 1 to set the indicator variable point mass at the values $v_0 > 0$ or $v_1 > 0$.

`start.tau2`

a positive numeric vector with the same length as `names`, that specifies the initial values of the variance parameter $\tau_{\beta_j}^2$ for the lasso- or ridge prior (`type=″lasso″` or `type=″ridge″`).

`h1.shrink`, `h2.shrink`

each is a single positive numeric (or a positive numeric vector with the same length as `names` if the adaptive prior versions are specified), to specify the hyperparameters `h1.shrink`$\hat{=} h_{1,\lambda} \geq 0$ and `h2.shrink`$\hat{=} h_{2,\lambda} \geq 0$ of the inverse gamma priors of the shrinkage parameter λ (`type=″lasso″` or `type=″ridge″`) or the hyperparameters `h1.shrink`$\hat{=} h_{1,\omega} \geq 0$ and `h2.shrink`$\hat{=} h_{2,\omega} \geq 0$ of the beta prior for ω (`type=″nigmix″`). The default values are `h1.shrink=0.001`, `h2.shrink1=0.001`, if `type=″ridge″` or `type=″lasso″`, and `h1.shrink=1`, `h2.shrink1=1`, if `type=″nigmix″`.

`start.shrink`

a single numeric (or a numeric vector with the same length as `names` if the adaptive prior versions are specified), interpreted as the initial value of the shrinkage parameter λ (`type=″lasso″` or `type=″ridge″`) or ω (`type=″nigmix″`). The default value is 0.5.

`fix.shrink`

Logical value, that specifies, if the shrinkage parameter λ (`type=″lasso″` or `type=″ridge″`) or ω (`type=″nigmix″`) should be fixed at the value given in option `start.shrink`. If not specified, the default value is set to `FALSE` so that the shrinkage parameter is estimated.

`splinepri`: an optional list, to specify the parameters of the regularized smooth effects:

`names`

a character vector, containing the column names of the `data` that identify the covariates corresponding to the non linear effects of the predictor.

`blocksize`

the same explanation as in the `unpenpri` list, but a vector with the same length as `names`. As default, the blocksize of each spline corresponds to the number of used basis functions.

`randomblocks`

the same explanation as in the `unpenpri` list, but a vector with the same length as `names`.

`degree`

an integer vector with the same length as `names`, containing the degrees of the splines. The components of degree vector are set per default to `degree=3`.

`nbasis`

an integer vector with the same length as `names`, containing the number of B-spline basis functions to model the nonlinear effects.

`difforder`

an integer vector with the same length as `names`, containing the difference order of the smoothing penalty.

`h1.tau2, h2.tau2`

both non-negative numeric vectors, with the same length as `names` containing the hyperparameter values `h1.tau2`$\hat{=} h_{1,\tau_j} > 0$ and `h2.tau2`$\hat{=} h_{2,\tau_j} > 0$ of the inverse gamma priors for the smoothing variances $\tau^2_{\alpha_j}$. The default values are `h1.tau2=0.001` and `h2.tau2=0.001`.

`start.tau2`

non-negative numeric vector, with the same length as `names` containing the initial values of the smoothing variances $\tau^2_{\alpha_j}$. The components of starting vector are set per default to 1.

`start.effect`

a list, with the same length as `names`. Each component of the list is a numeric vector that contains the initial values of the basis function weights.

`simpar`: a list, giving the parameters of the MCMC simulation:

`niter`

an integer, giving the number of iterations for the sampler.

`nthin`

an integer, giving the thinning parameter of the chain to compute the characteristics of the parameter specific marginal empirical posterior distribution like the mean, standard deviation and quantiles. The sequence from `nburn` to `niter` by `nthin` is used for printing these results on the screen. In the output files all sampled values given by `niter` are stored.

`nburn`

an integer, that sets the number of initial sampled values treated as burn-in values.

`nwrite`

an interval, with which the sampled values are written to the output files.

`seed`

an optional single value, interpreted as an integer to define the seed parameter of the implied function `set.seed()`.

catniter

an interval, at which information about the number of performed iterations is printed on the screen.

dir: a list that specifies the directory information to store the sampled values with components:

outdir

a character string, that specify a directory where the output files should be stored. All output files will be named with "outnam_" as prefix.

outnam

a character string, used as prefix of the generated output files.

overwrite

a logical value, which enables to overwrite existing files with the same outnam in the outdir directory.

dirfunctions: a character string, specifying the directory where collection of implemented function are stored.

Value

A character vector, that contains the storing paths of each generated file to load the results into R.

Files created

The * prefix denotes the replacement character for the user specified base name as defined in outnam.

*_mcmc_call.RData: File with the arguments of the function call.

*_mcmc_design.RData: File that contains the specification of all parameters of the function.

*_mcmc_result.RData: File that contains the storing paths of each generated file.

*_sim_unpen_gamma.RData: Optional file that contains the samples of the unpenalized effects.

*_sim_unpen_accepted.RData: Optional file that contains the acceptance status of the unpenalized effects in each iteration. 0=rejected, 1=accepted.

*_sim_pen_beta.RData: Optional file that contains the samples of the penalized effects.

*_sim_pen_accepted.RData: Optional file that contains the acceptance status of the penalized effects in each iteration. 0=rejected, 1=accepted.

*_sim_pen_I.RData: Optional file that contains the samples of the indicator variables if the NMIG-prior is used.

*_sim_pen_t2.RData: Optional file that contains the samples of the variance components ψ_j^2 if the NMIG-prior is used.

*_sim_pen_shrink.RData: Optional file that contains the samples of the shrinkage parameter.

*_sim_pen_tau2.RData: Optional file that contains the samples of the variance parameters.

*_sim_spline_beta_xx.RData: Optional file that contains the samples of the basis function weights. xx in the filename denotes the covariate name corresponding to the smooth effect.

*_sim_spline_accepted_xx.RData: Optional file that contains the acceptance status of the smooth effects in each iteration. 0=rejected, 1=accepted. xx in the filename denotes the covariate name corresponding to the smooth effect.

`*_sim_spline_tau2_xx.RData`: Optional file that contains the samples of the smoothing variance spline estimation. `xx` in the filename denotes the covariate name corresponding to the smooth effect.

Example

```
# Using the veteran data from the R-package {survival}
my.veteran <- as.matrix(veteran[,-2])
bcoxpl(dataset= list(data=my.veteran,time="time",delta="status"),
       penpri=list(type="lasso",names=c("karno","age"),
       start.effect=rep(0.01,2),start.tau2 = rep(1/10,2),
       h1.shrink = 0.01, h2.shrink = 0.01, start.shrink = 1),
       simpar= list(niter=10000,catniter=100),
       dirfunctions=file.path(".","RWD","FUNCTIONS"))
```

D.5. Arguments of the R-function baftpgm

The function is described in the version that was used for the simulations.

Usage

```
baftpgm(dataset,
     errorpar=list(method.alpha="mhcond",order.alpha="fix1",
                   method.rlabel="gibbs", djust.alpha="no",
                   scalebasis=FALSE,scalebasis.type="s"),
     errorpri,unpenpri,penpri,splinepri,
     simpar=list(niter=10000,nthin=1,nburn=0,nwrite=10000,catniter =100),
     dir=list(outdir=getwd(),outnam="bpgm", overwrite=FALSE),
     dirfunctions,errorplot)
```

Arguments

`dataset`: a list, containing the data with the following components:

`data`

a matrix, containing the variables in the model, i.e. the right censored survival times, the censoring indicator and the covariates that enters the predictor.

`logT`

a character string, specifying the column of the data frame specified in the argument `data` that is interpreted as the logarithm of the observed survival time.

`delta`

a character string, specifying the column of the data frame specified in the argument `data` that is interpreted as the censoring indicator with values 0=alive and 1=death.

`errorpar`: a list, giving the method to update the error weights, compare Section 6.1.3:

`method.alpha`

a character string, that is assumed to name an element from `"ars"`, `"slice"`, `"dirichlet"`, `"mhmarg"`, `"mhcond"`, `"mcondstep"` or `"mcondblok"` to specify the update method of the error weights.

`order.alpha`

a character string, that is assumed to name an element from `"fix1"`, `"fix2"`, `"random1"` or `"random2"` to specify the order of the updated error weights. This argument is only used in combination with `method.alpha=″mcondstep″`, `method.alpha=″ars″` or `method.alpha=″slice″`.

`scalebasis`

logical value. If specified as `TRUE`, in each iteration the basis knots and basis variances are recomputed so that the error distribution has zero mean and unit variance.

`errorpri`: a list, giving the parameters of the error density:

`type`

a character string, which is assumed to name an element from `"gaussian"` or `"pgm"` to specify, if the error distribution is assumed to be Gaussian or a penalized Gaussian mixture (PGM).

`difforder`

an integer, giving the difference order of the smoothing penalty for the error distribution if `type=″pgm″`.

`start.muknots`

a numeric vector, specifying the position of the Gaussian basis function means m_j. As default $g_0 = 31$ knots building a sequence from $m_1 = -4.5$ to $m_{g_0} = -4.5$ with differences 0.3 are used.

`start.s2knots`

a numeric value or numeric vector, with the same length as `start.muknots` specifying the variances s_j^2 of the Gaussian basis functions. As default, all variances are set to the value $s_j^2 = 0.2^2$.

`zero.alpha`

an integer, giving the index of the reference knot. As default, the middle knot is used.

`start.weight`

a positive numeric vector of the same length as `start.muknots`, with the starting values of the error weights w_j.

`start.alpha`

an optional numeric vector of the same length as `start.muknots`, with the starting values of the transformed error weights. If not specified, each transformed weight, except the reference weight `zero.alpha`, is set to 0.01.

`start.intercept`

a numeric value, giving the initial value of the shift γ_0 of the error distribution.

`start.rlabel`

a vector, that specifies the initial labels r_i of the mixture components, into which the residuals are intrinsically assigned. The label have the values from $\{1,...,g_0\}$.

`h1.sigma2, h2.sigma2`

both non-negative numeric values, to specify the hyperparameters `h1.sigma2`$\hat{=} h_{1,\sigma} \geq 0$ and `h2.sigma2`$\hat{=} h_{2,\sigma} \geq 0$ of the inverse gamma prior for the scale parameter σ^2. Default values are `h1.sigma2=0.001` and `h2.sigma2=0.001`.

`start.sigma2`

a positive numeric, giving the initial value of the scale parameter σ^2. The default value is `start.sigma2=1`.

`slice.sigma2.m`, `slice.sigma2.w`, `slice.sigma2.lower`, `slice.sigma2.upper`

arguments correspond the arguments `m`, `w`, `lower` and `upper` of the `R`-function `uni.slice()`.

`h1.tau2`, `h2.tau2`

both non-negative numeric values, containing the hyperparameter values `h1.tau2`$\hat{=} h_{1,\tau_0} \geq 0$ and `h2.tau2`$\hat{=} h_{2,\tau_0} \geq 0$ of the inverse gamma prior for the smoothing variance $\tau^2_{\alpha_0}$ for the error density. The default values are `h1.tau2=0.001` and `h2.tau2=0.001`.

`start.tau2`

non-negative numeric, giving the initial value of the smoothing variances $\tau^2_{\alpha_0}$. The default value is `start.tau2=0.001`.

`scaledpri`

option, to specify the scale-dependent prior versions, compare Section 6.2.4 The default is `scaledpri=FALSE`.

`unpenpri`: an optional list, to specify the parameters for the unpenalized linear effects:

For details compare the description of the function `bcoxpl()`.

`penpri`: an optional list, to specify the parameters of the regularized linear effects:

For details compare the description of the function `bcoxpl()`.

`splinepri`: an optional list, to specify the parameters of the regularized smooth effects:

For details compare the description of the function `bcoxpl()`.

`simpar`: a list, giving the parameters of the MCMC simulation:

For details compare the description of the function `bcoxpl()`.

`dir`: a list, that specifies a directory information to store the sampled values with components:

For details compare the description of the function `bcoxpl()`.

`dirfunctions`: a character string, specifying the directory with the implemented function:

For details compare the description of the function `bcoxpl()`.

`errorplot`: an optional list, to plot the estimated error density through the iterations in a postscript file:

`plotiter`

an integer, to specify an interval at which the error density is printed in the output file.

`rn.grid`

a vector, that specifies the grid points at which the error density is evaluated and plotted.

Value

A character vector, that contains the storing paths of each generated file to load the results into `R`.

Files created

The `*` prefix denotes the replacement character for the user specified base name as defined in `outnam`.

`*_mcmc_call.RData`: File with the arguments of the function call.

`*_mcmc_design.RData`: File that contains the specification of all parameters of the function.

`*_mcmc_result.RData`: File that contains the storing paths of each generated file.

`*_sim_unpen_gamma.RData`: Optional file that contains the samples of the unpenalized effects.

`*_sim_pen_beta.RData`: Optional file that contains the samples of the penalized effects.

`*_sim_pen_I.RData`: Optional file that contains the samples of the indicator variables if the NMIG-prior is used.

`*_sim_pen_t2.RData`: Optional file that contains the samples of the variance components ψ_j^2 if the NMIG-prior is used.

`*_sim_pen_shrink.RData`: Optional file that contains the samples of the shrinkage parameter.

`*_sim_pen_tau2.RData`: Optional file that contains the samples of the variance parameters.

`*_sim_spline_beta_xx.RData`: Optional file that contains the samples of the basis function weights. `xx` in the filename denotes the covariate name corresponding to the smooth effect.

`*_sim_spline_tau2_xx.RData`: Optional file that contains the samples of the smoothing variance spline estimation. `xx` in the filename denotes the covariate name corresponding to the smooth effect.

In addition

`*_sim_error_alpha.RData`: Optional file that contains the samples of the transformed error weights $\alpha_{0,j}$.

`*_sim_error_accepted.RData`: Optional file that contains the acceptance status of the transformed error weights if `method` is set to `"mhmarg"`, `"mhcond"`, `"mcondstep"` or `"mcondblock"` in each iteration. 0=rejected, 1=accepted.

`*_sim_error_sigma2.RData`: File that contains the samples of the scale parameter σ^2.

`*_sim_error_tau2.RData`: File that contains the samples of the smoothing parameter $\tau^2_{\alpha_0}$.

`*_sim_error_muknots.RData`: Optional file, if `scalebasis=TRUE`, that contains the corrected positions of the basis function knots m_j of each iteration.

`*_sim_error_s2knots.RData`: Optional file, if `scalebasis=TRUE`, that contains the corrected positions of the basis function variances s_j^2 of each iteration.

`*_sim_error_rlabel.RData`: samples labels r_i of the mixture components into which the residuals are assigned.

`*_sim_errorvideo.ps`: File created if the option `errorplot` is specified.

Example

```
load(file.path(...,"pbcliver.RData"))
bpgm <- baftpgm(dataset=list(data=pbcliver,logT="logtime",delta="delta"),
        errorpar=list(method.alpha="slice"),
        errorpri=list(type="pgm",start.muknots=seq(-4.5,4.5,length.out=21),
                      start.s2knots=0.25^2,start.intercept=1,
             h1.sigma2=0.001,h2.sigma2=0.001,start.sigma2=1^2,
             h1.tau =1,h2.tau2=0.001,start.tau2=0.001),
penpri=list(type="lasso",names=c("chol","age"),
            start.effect=rep(0.01,2),start.tau2=rep(1/10,2),
            h1.shrink=0.01,h2.shrink=0.01,start.shrink=1),
simpar=list(niter=10000,catniter=100),
dirfunctions=file.path(".","RWD","FUNCTIONS"),
errorplot=list(plotiter=100,rn.grid=seq(1,15,by=0.1)))
```

REFERENCES

Abbreviations

We use the following abbreviations

AIC	Akaike-Information-Criterion
DIC	Deviance-Information-Criterion
gP	generalized Pareto
gdP	generalized double Pareto
IQR	interquartile range
IBS	integrated Brier score
MAP	maximum a posteriori
MCMC	Markov-Chain-Monte-Carlo
MH	Metropolis-Hastings
ML	maximum Likelihood
MSE	mean squared error
PGM	penalized Gaussian mixture
P-spline	penalized spline
p.d.f.	probability density function
c.d.f.	cumulative distribution function
i.i.d.	independent and identically distributed

For the description of the simulation and application results, the following abbreviations are used to reduce the writing.

In the case of the accelerated failure time model of Section 10 we use the abbreviations

PGM: if the baseline error distribution is modeled by a penalized Gaussian mixture,

AFT: if the baseline error distribution is Gaussian.

For the Cox type hazard rate models of Section 11 we use

CPL: if inference is based on the partial likelihood,

CFL: if inference is based on the full likelihood P-spline baseline hazard,

WB: for the special case of the full likelihood corresponding to the Weibull model.

When results are achieved via Bayesian inference the previous abbreviations are combined with

B: to denote models without regularization of the linear effects,

BL: to denote models with Bayesian lasso regularization of the linear effects,

BN: to denote models with Bayesian NMIG regularization of the linear effects,

BR: to denote models with Bayesian ridge regularization of the linear effects,

BT: to denote models where the predictor contains only the “true” nonzero effects.

The frequentist models are combined with

Step: to denote the backward stepwise selection based on AIC,

PenL: to denote a penalized partial likelihood based CRR model with lasso penalty,

PenR: to denote a penalized partial likelihood based CRR model with ridge penalty,

T: denote models where the predictor contains only the “true” nonzero effects.

For the Bayesian approaches, the hard shrinkage methods described in Section 4.4 are additionally assigned with

HS.CRI: if hard shrinkage is done via the 95% credible region,

HS.STD: if hard shrinkage is done via the one standard error region,

HS.IND: if hard shrinkage is done via the NMIG indicator variables.

For example, *WB.BN-HS.IND* denotes the Bayesian Weibull model under NMIG penalty, if the covariate specific indicators are used to select the covariates for the final model, and *CPL.PenL* is the short cut for the frequentist lasso penalty applied to the linear predictors, if inference is carried out with the partial likelihood.

The notation for different update schemes of the transformed error weights in the AFT model with PGM error are introduced in Section 6.1.3. They are combined with the following suffixes to indicate the specification of some options of the function `baftpgm()`, compare Appendix D.5:

FK: indicates the specification `scalebasis=FALSE`,

otherwise (if `scalebasis=TRUE`) the knots are transformed to standardize the error density estimation in each iteration loop of the sampler. In particular the two update schemes *“slice”* and *“mcondstep”* enable to vary order of the update of the transformed error weights:

R0: indicates the specification `order.alpha=″fix2″`,

R1: indicates the specification `order.alpha=″random1″`,

R2: indicates the specification `order.alpha=″random2″`.

For example, *“sliceR1FK”* denotes that the update schemes *“slice”* is used to update the transformed error weights (`method.alpha=″slice″`) with the options `order.alpha=″random1″` and `scalebasis=FALSE`.

Bibliography

Andersen, Borgan, Gill and Keiding (1993). *Statistical Models Based on Counting Processes*. Springer Series in Statistics. Springer, New York.

Andersen, P. K. and Gill, R. D. (1982). Cox's Regression Model for Counting Processes: A Large Sample Study. *The Annals of Statistics* , ***10(4)***, 1100-1120.

Armagan, A. and Zaretzki, R. L. (2010). Model selection via adaptive shrinkage with t priors. *Computational Statistics*, ***25(3)***, 441-461.

Armagan, A., Dunson, D. B. and Lee, J. (2013). Generalized double Pareto shrinkage. *Statistica Sinica*, ***23(1)***, 119-143.

Bender, R., Augustin, T. and Blettner, M. (2005). Simulating survival times for Cox regression models. *Statistics in Medicine*, ***24***, 1713-1723.

Benner, A., Zucknick, M., Hielscher, T., Ittrich, C., Mansmann, U. (2010). High-dimensional Cox Models: The Choice of Penalty as Part of the Model Building Process. *Biometrical Journal*, ***52(1)***, 50-69.

Breslow N. E. (1972). Discussion of the paper by Cox, D. R. (1972). *Journal of the Royal Statistical Society, Series B*, ***34(2)***, 216-217.

Breslow, N. E. (1974). Covariance Analysis of Censored Survival Data. *Biometrics*, ***30(1)***, 89-99.

Brezger, A. and Lang, S. (2006). Generalized structured additive regression based on Bayesian P-splines. *Computational Statistics and Data Analysis*, ***50(4)***, 967-991.

Buckley, J. and James, I. (1979). Linear regression with censored data. *Biometrika*, ***66(3)***, 429-436.

Cai, T., Huang, J. and Tian, L. (2009). Regularized Estimation for the Accelerated Failure Time Model. *Biometrics*, ***65(2)***, 394-404.

Carvalho, C. M., Polson, N. G. and J.G. Scott (2010). The horseshoe estimator for sparse signals. *Biometrika*, ***97(2)***, 465-480.

Chen, Y. Q. and Wang, M.-C. (2000). Analysis of Accelerated Hazards Models. *Journal of the American Statistical Association*, ***95(450)***, 608-618.

Clyde, M. and George, E. (2000). Flexible empirical Bayes estimation for wavelets. *Journal of the Royal Statistical Society, Series B, **62(4)***, 681-698.

Cox, D. R. (1972). Regression Models and Life Tables (with Discussion). *Journal of the Royal Statistical. Society, Series B, **34(2)***, 187-220.

Cox, D. R. (1975). Partial Likelihood. *Biometrika, **62***, 269-276.

Cox, D. R. and Oakes, D. (1984). *Analysis of Survival Data.* Monographs on Statistics & Applied Probability. London-New York: Chapman & Hall/CRC.

Datta, S., Le-Rademacher, J. and Datta, S. (2007). Predicting Patient Survival from Microarray Data by Accelerated Failure Time Modeling Using Partial Least Squares and LASSO. *Biometrics*, ***63(1)***, 259-271.

De Boor, C. (2001). *A Practical Guide to Splines*. Applied Mathematical Sciences. Springer, New York.

Efron, B., Hastie, T., Johnstone, I. M. and Tibshirani, R. (2004). Least angle regression. *The Annals of Statistics, **32(2)***, 407-499.

Eilers, P. H. C. and Marx, B. D. (1996). Flexible smoothing using B-splines and penalties (with comments and rejoinder). *Statistical Science*, ***11(2)***, 89-121.

Eltoft, T., Kim, T. and Lee, T.-W. (2006). Multivariate Scale Mixture of Gaussian Modeling. *Lecture Notes in Computer Science*, ***3889***, 799-806.

Engler, D. and Li, Y. (2009). Survival Analysis with High-Dimensional Covariates: An Application in Microarray Studies. *Statistical Applications in Genetics and Molecular Biology*, ***8(1)***, Article 14.

Etezadi-Amoli, J. and Ciampi, A. (1987). Extended Hazard Regression for Censored Survival Data with Covariates: A SplineApproximation for the Baseline Hazard Function. *Biometrics*, ***43(1)***, 181-192.

Fahrmeir, L. and Kneib, T. (2009). Propriety of posteriors in structured additive regression models: theory and empirical evidence. *Journal of Statistical Planning and Inference*, ***139(3)***, 843-859.

Fahrmeir, L. and Kneib, T. (2011). *Bayesian Smoothing and Regression for Longitudinal, Spatial and Event History Data*. Oxford University Press.

Fahrmeir, L., Kneib, T. and Lang, S. (2004). Penalized structured additive regression for space-time data: a Bayesian perspective. *Statistica Sinica*, ***14***, 731-761.

Fahrmeir, L., Kneib, T., Konrath, S. (2010). Bayesian regularization in structured additive regression: a unifying perspective on shrinkage, smoothing and predictor selection. *Statistics and Computing*, ***20(2)***, 203-219.

Fan, J. and Li, R. (2001). Variable Selection via Nonconcave Penalized Likelihood and its Oracle Properties. *Journal of the American Statistical Association*, ***96(456)***, 1348-1360.

Fan, J. and Li, R. (2002). Variable Selection for Cox Proportional Hazards Model and Frailty Model. *The Annals of Statistics*, ***30(1)***, 74-99.

Frühwirth-Schnatter, S. (2006). *Finite Mixture and Markov Switching Models*. Springer Series in Statistics. Springer, New York, Berlin, Heidelberg.

Fu, W. J. (1998). Penalized Regressions: The Bridge Versus the Lasso. *Journal of Computational and Graphical Statistics,* ***7(3)***, 397-416.

Gamerman, D. (1997). Efficient sampling from the posterior distribution in generalized linear models. *Statistics and Computing*, ***7***, 57-68.

Gelman, A. (2006). Prior distributions for variance parameters in hierarchical models. *Bayesian Analysis*, ***1(3)***, 515-533.

Gelman, A. (2008). Scaling regression inputs by dividing by two standard deviations. *Statistics in Medicine*, ***27(15)***, 2865-2873.

Gelman, A., Carlin, J. B., Stern, H. S. and Rubin, D. B. (2004). *Bayesian Data Analysis (2nd Edition)*. Texts in Statistical Science. Chapman & Hall/CRC.

George, E. I. and McCulloch, R. E. (1993). Variable Selection via Gibbs sampling. *Journal of the American Statistical Association*, ***88(423)***, 881-889.

George, E. I. and McCulloch, R. E. (1997). Approaches for Bayesian Variable Selection. *Statistica Sinica,* ***7***, 339-373.

Geweke, J. (1996). *Variable Selection and Model Comparison in Regression*. In Bernardo et al. (eds.): Bayesian Statistics 5, Oxford University Press.

Gilks, W. R. and Wild, P. (1992). Adaptive Rejection Sampling for Gibbs Sampling. *Journal of the Royal Statistical Society, Series C*, ***41(2)***, 337-348.

Gilks, W. R., Richardson, S. and Spiegelhalter, D. J. (1996). *Markov Chain Monte Carlo in Practice*. Interdisciplinary Statistics. Chapman & Hall/CRC.

Goeman, J. J. (2010). L1 penalized estimation in the Cox proportional hazards model. *Biometrical Journal*, ***52(1)***, 70-84.

Graf, E., Schmoor, C., Sauerbrei, W., Schumacher, M. (1999). Assessment and comparison of prognostic classification schemes for survival data. *Statistics in Medicine*, ***18,*** 2529-2545.

Gray, R. J. (1992). Flexible Methods for Analyzing Survival Data Using Splines, With Applications to Breast Cancer Prognosis. *Journal of the American Statistical Association*, ***87(420)***, 942-951.

Griffin, J. E. and Brown, P. J. (2005). *Alternative prior distributions for variable selection with very many more variables than observations*. Technical report, Department of Statistics, University of Warwick.

Griffin, J. E. and Brown, P. J. (2007). *Bayesian adaptive lassos with non-convex penalization*. Technical report, Department of Statistics, University of Warwick.

Gui, J. and Li, H. (2005). Penalized Cox regression analysis in the high-dimensional and low-sample size settings, with applications to microarray gene expression data. *Bioinformatics,* ***21(13)***, 3001-3008.

Hans, C. (2009). Bayesian Lasso regression. *Biometrika*, ***96(4)***, 835-845.

Hans. C. (2011). Elastic Net Regression Modeling with the Orthant Normal Prior. *Journal of the American Statistical Association*, ***106(496)***, 1383-1393.

Henderson, R., Shimakura, S. and Gorst, D. (2002). Modeling Spatial Variation in Leukemia Survival Data, *Journal of the American Statistical Association*, ***97***, 965-972.

Hennerfeind, A., Brezger, A. and Fahrmeir, L. (2006). Geoadditive survival models. *Journal of the American Statistical Association*, ***101(475)***, 1065-1075.

Hoerl, A. E. and Kennard, R. W. (1970). Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, ***12(1)***, 55-67.

Hoeting, J. A., Madigan, D., Raftery, A. E. and Volinsky, C. T. (1999). Bayesian Model Averaging: A Tutorial. *Statistical Science*, ***14(4)***, 382-417.

Huang, J. and Ma, S. (2010). Variable selection in the accelerated failure time model via the bridge method. *Lifetime Data Analysis*, ***16***, 176-195.

Huang, J., Ma, S. and Xie, H. (2006). Regularized estimation in the accelerated failure time model with high-dimensional covariates. *Biometrics*, ***62(3)***, 813-820.

Ibrahim, J. G., Chen, M.-H. and Sinha D. (2001). *Bayesian Survival Analysis*. Springer Series in Statistics. Springer, New York.

Ishwaran, H. and Rao, S. J. (2003). Detecting Differentially Expressed Genes in Microarrays Using Bayesian Model Selection. *Journal of the American Statistical Association,* ***98(462)***, 438-455.

Ishwaran, H. and Rao, S. J. (2005a). Spike and Slab Gene Selection for Multigroup Microarray Data. *Journal of the American Statistical Association, **100(471)***, 764-780.

Ishwaran, H. and Rao, S. J. (2005b). Spike and Slab Variable Selection: Frequentist and Bayesian Strategies. *The Annals of Statistics, **33(2)***, 730–773.

Johnson, B. A. (2008). Variable selection in semiparametric linear regression with censored data. Technical Report, Department of Biostatistics, Emory University Atlanta.

Johnson, B. A. (2009). On lasso for censored data. *Electronic Journal of Statistics*, ***3***, 485-506.

Johnson, B. A., Lin, D. Y. and Zeng, D. (2008). Penalized Estimating Functions and Variable Selection in Semiparametric Regression Models. *Journal of the American Statistical Association*, ***103(482)***, 672-680.

Kaderali, L. (2006). A Hierarchical Bayesian Approach to Regression and its Application to Predicting Survival Times in Cancer. Dissertation, Universität zu Köln. Köln.

Kaderali, L., Zander, T., Faigle, U., Wolf, J., Schultze, J. L. and Schrader, R. (2006). CASPAR: a hierarchical Bayesian approach to predict survival times in cancer from gene expression data. *Bioinformatics, **22(12)***, 1495-1502.

Kalbfleisch J. D. (1978). Non-Parametric Bayesian Analysis of Survival Time Data. *Journal of the Royal Statistical Society. Series B*, ***40(2)***, 214-221.

Kalbfleisch, J. D., and Prentice, R. L. (2002). *The Statistical Analysis of Failure Time Data (2nd ed.)*. Wiley Series in Probability and Statistics. New York: John Wiley & Sons.

Kim, Y. and Kim, D. (2009). Bayesian partial likelihood approach for tied observations. *Journal of Statistical Planning and Inference*, ***139***, 469-477.

Klein, J. P. and Moeschberger, M. L. (2003). *Survival Analysis: Techniques for Censored and Truncated Data*. (2nd ed.). Springer, New York.

Kneib T. and Fahrmeir L. (2007). A Mixed Model Approach for Geoadditive Hazard Regression. *Scandinavian Journal of Statistics*, ***34(1)***, 207-228.

Kneib, T., Konrath, S., Fahrmeir, L. (2011). High-dimensional Structured Additive Regression Models: Bayesian Regularization, Smoothing and Predictive Performance. *Journal of the Royal Statistical Society, Series C*, ***60(1)***, 51-70.

Knight, K. and Fu, W. J. (2000). Asymptotics for Lasso-type estimators. *The Annals of Statistics, **28(5)***, 1356-1378.

Knorr-Held, L. (1999). Conditional Prior Proposals in Dynamic Models. *Scandinavian Journal of Statistics*, ***26***, 129-144.

Komárek, A. and Lesaffre, E. (2008a). Bayesian Accelerated Failure Time Model With Multivariate Doubly Interval-Censored Data and Flexible Distributional Assumptions. *Journal of the American Statistical Association*, ***103(482)***, 523-533.

Komárek, A., Lesaffre, E. (2008b). Generalized linear mixed model with a penalized Gaussian mixture as a random-effects distribution. *Computational Statistics and Data Analysis*, ***52(7)***, 3441–3458.

Komárek, A., Lesaffre, E. and Hilton, J. F. (2005). Accelerated Failure Time Model for Arbitrarily Censored Data With Smoothed Error Distribution. *Journal of Computational and Graphical Statistics*, ***14***, 726–745.

Komárek, A., Lesaffre, E. and Legrand, C. (2007). Baseline and treatment effect heterogeneity for survival times between centers using a random effects accelerated failure time model with flexible error distribution. *Statistics in Medicine*, ***26***, 5457-5472.

Konrath, S. (2007). *Bayesianische Regularisierung mit Anwendungen*. Masterthesis, LMU München.

Konrath, S., Kneib, T. and Fahrmeir, L. (2013). Bayesian Smoothing, Shrinkage and Variable Selection in Hazard Regression. To appear in Becker, C., Fried, R. and Kuhnt, S. (eds.). *Robustness and Complex Data Structures*. Springer, Berlin, Heidelberg.

Kyung, M. Gill, J., Ghosh, M. and Casella, G. (2010). Penalized Regression, Standard Errors, and Bayesian Lassos. *Bayesian Analysis*, ***5(2)***, 369-412.

Lambert, P. (2013). Nonparametric additive location-scale models for interval censored data. *Statistics and Computing*, ***23(1)***, 75-90.

Lang, S. and Brezger, A. (2004). Bayesian P-Splines. *Journal of Computational and Graphical Statistics*, ***13***, 183-212.

Lee, A., Caron, F., Doucet, A. and Holmes, C. (2012). Bayesian Sparsity-Path-Analysis of Genetic Association Signal using Generalized t Priors. *Statistical Applications in Genetics and Molecular Biology,* ***11(2***), pp. 5.

Lee, K. H., Chakraborty, S. and Sun, J. (2011). Bayesian Variable Selection in Semiparametric Proportional Hazards Model for High Dimensional Survival Data. *The International Journal of Biostatistics*, ***7(1)***, Article 21.

Lesaffre, E., Komárek, A. and Declerck, D. (2005). An overview of methods for interval-censored data with an emphasis on applications in dentistry. *Statistical Methods in Medical Research*, ***14(6)***, 539-552.

Li, F. and Zhang, N. R. (2010). Bayesian Variable Selection in Structured High-Dimensional Covariate Spaces With Applications in Genomics. *Journal of the American Statistical Association*, ***105(491)***, 1202-1214.

Li, Q. and Lin, N. (2010). The Bayesian Elastic Net. *Bayesian Analysis*, ***5(1)***, 847-866.

Lin, D. Y. (2007). On the Breslow estimator. *Lifetime Data Analysis,* ***13(4)***, 471-480.

Lindley, D. V. and Smith, A. F. M. (1972). Bayes Estimates for the Linear Model. *Journal of the Royal Statistical Society, Series B*, ***34(1)***, 1-41.

Lu, W. and Zhang, H. (2010). On Estimation of Partially Linear Transformation Models. *Journal of the American Statistical Association*, ***105(490)***, 683-691.

MacLehose, R. F. and Dunson, D. B. (2010). Bayesian Semiparametric Multiple Shrinkage. *Biometrics*, ***66(2)***, 455-462.

Metzeler, K. H., Hummel, M., Bloomfield, C. D., Spiekermann, K., Braess, J., et al. (2008). An 86-probe-set gene-expression signature predicts survival in cytogenetically normal acute myeloid leukemia. *Blood*, ***112***, 4193-4201.

Murphy, S. A. and van der Vaart, A. W. (2000). On profile Likelihood. *Journal of the American Statistical Association*, ***95(450)***, 449-465.

Neal, R. M. (2003). Slice Sampling. *The Annals of Statistics*, ***31(3)***, 705-761.

Oakes, D. (1972). Discussion of the paper by Cox, D. R. (1972). *Journal of the Royal Statistical Society, Series B*, ***34(2)***, 208.

Park, M. Y. and Hastie, T. (2007). L1-regularization path algorithm for generalized linear models. *Journal of the Royal Statistical Society, Series B*, ***69(4)***, 659-677.

Park, T. and Casella, G. (2008). The Bayesian Lasso. *Journal of the American Statistical Association*, ***103(482)***, 681-686.

Peto, R. (1972). Discussion of the paper by Cox, D. R. (1972). *Journal of the Royal Statistical Society, Series B*, ***34(2)***, 205-207.

Polson, N. G. and Scott, J. G. (2010). *Shrink Globally, Act Locally: Sparse Bayesian Regularization and Prediction*. In J. Bernardo, M. Bayarri, J. O. Berger, A. Dawid, D. Heckerman, A. Smith, and M. West, editors, Proceedings of the 9th Valencia World Meeting on Bayesian Statistics. Oxford University Press.

Polson, N. G. and Scott, J. G. (2011). *The Bayesian Bridge*. Technical Report, University of Texas at Austin. http://arxiv.org/abs/1109.2279, 2011b.

Polson, N. G. and Scott, J. G. (2012). Local shrinkage rules, Lévy processes and regularized regression. *Journal of the Royal Statistical Society, Series B*, ***74(2)***, 287-311.

Robert (1995). Simulation of truncated normal variables. *Statistics and Computing*, **5**, 121-125.

Rockova, V., Lesaffre, E., Luime, J. and Löwenberg, B. (2012). Hierarchical Bayesian formulations for selecting variables in regression models. *Statistics in Medicine*, ***31***, 1221-1237.

Rue, H. (2001). Fast sampling of Gaussian Markov random fields. *Journal of the Royal Statistical Society, Series B*, ***63(2)***, 325-338.

Rue, H. and Held, L. (2005). *Gaussian Markov Random Fields: Theories and Applications*. Chapman & Hall/CRC.

Ruppert, D., Wand, M. P. and Carroll, R. J. (2003). *Semiparametric Regression*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, Cambridge.

Scheipl, F. (2011). *Bayesian Regularization and Model Choice in Structured Additive Regression*. Dissertation, LMU München.

Schellhase, C. (2010). R-package pendensity - Density estimation with a Penalized Mixture Approach. R package version 0.2.2.

Schellhase, C. and Kauermann, G. (2011). Density estimation and comparison with a penalized mixture approach. *Computational Statistics*, ***27(4)***, 757-777.

Sha, N., Tadesse, M. G. and Vannucci, M. (2006). Bayesian variable selection for the analysis of microarray data with censored outcomes. *Bioinformatics*, ***22(18)***, 2262-2268.

Shyur, H.-J., Elsayed, E. A. and Luxhøj, J. T. (1999). A General Model for Accelerated Life Testing with Time-Dependent Covariates. *Naval Research Logistics*, ***46(3)***, 303-321.

Sinha, D. (1993). Semiparametric Bayesian Analysis of Multiple Event Time Data. *Journal of the American Statistical Association*, ***88(423)***, 979-983.

Sinha, D., Ibrahim, J. G. and Chen, M.-H. (2003). A Bayesian justification of Cox's partial likelihood. *Biometrika,* ***90(3)***, 629-641.

Sleeper, L. A. and Harrington, D. P. (1990). Regression Splines in the Cox Model with Application to Covariate Effects in Liver Disease. *Journal of the American Statistical Association*, ***85(412)***, 941-949.

Smith, M. und Kohn, R. (1996). Nonparametric regression using Bayesian variable selection, *Journal of Econometrics,* ***75***, 317-343.

Spiegelhalter, D. J., Best, N. G., Carlin, B. P. and van der Linde, A. (2002). Bayesian measures of model complexity and fit (with discussion). *Journal of the Royal Statistical Society, Series B*, ***64(4)***, 583-639.

Tachmazidou, I., Johnson, M. R. and De Iorio, M. (2010). Bayesian Variable Selection for Survival Regression in Genetics. *Genetic Epidemiology*, ***34***, 689–701.

Tanner, M. A. and Wong, W. H. (1987). The Calculation of Posterior Distributions by Data Augmentation. *Journal of the American Statistical Association*. ***82(398)***, 528-540.

Therneau, T. M. and Grambsch, P. M. (2000). *Modeling survival data: Extending the Cox model.* Springer, New York.

Tibshirani, R. (1996). Regression Shrinkage and Selection via the Lasso. *Journal of the Royal Statistical Society, Series B,* ***58(1)***, 267-288.

Tibshirani, R. (1997). The LASSO Method for Variable Selection in the Cox Model. *Statistics in Medicine*, ***16(4)***, 385-395.

Tseng, Y.-K., Hsieh, F. and Wang, J.-L. (2005). Joint modeling of accelerated failure time and longitudinal data. *Biometrika*, ***92(3)***, 587-603.

van Houwelingen, H. C., Bruinsma, T., Hart, A. A. M., van't Veer, L. J. and Wessels, L. F. A. (2006). Cross-validated Cox regression on microarray gene expression data. *Statistics in Medicine*, ***25(18)***, 3201-3216.

Verweij, P. J. M. and van Houwelingen, H. C. (1993). Cross-Validation in Survival Analysis. *Statistics in Medicine*, ***12(24)***, 2305-2314.

Verweij, P. J. M. and van Houwelingen, H. C. (1994). Penalized Likelihood in Cox Regression. *Statistics in Medicine*, ***13(23-24)***, 2427-2436.

Verweij, P. J. M. and van Houwelingen, H. C. (1995). Time-Dependent Effects of Fixed Covariates in Cox Regression. *Biometrics*, ***51(4)***, 1550-1556.

Volinsky, C. T., Madigan, D., Raftery, A. E., Kronmal, R. A. (1997). Bayesian Model Averaging in Proportional Hazard Models: Assessing the Risk of a Stroke. *Journal of the Royal Statistical Society, Series C*, ***46(4)***, 433-448.

Wang, S., Nan, B., Zhu, J. and Beer, D. G. (2008). Doubly penalized Buckley-James method for survival data with high-dimensional covariates. *Biometrics*, ***64***, 132-140.

Wei, L. J. (1992). The Accelerated Failure Time Model: A useful alternative to the Cox Regression Model in Survival Analysis. *Statistics in Medicine*, ***11***, 1871-1879.

West, M. (1987). On Scale Mixtures of Normal Distributions. *Biometrika*, ***74(3)***, 646-648.

Wood, S. N. (2006). *Generalized Additive Models*. Chapman & Hall/CRC.

Yuan, M. and Lin, Y. (2006). Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society, Series B,* ***68(1)***, 49-67.

Zhang, H. H. and Lu, W. (2007). Adaptive Lasso for Cox's Proportional Hazards Model. *Biometrika*, ***94(3)***, 691-703.

Zou, H. (2006). The Adaptive Lasso and Its Oracle Properties. *Journal of the American Statistical Association,* ***101(476)***, 1418-1429.

Zou, H. and Hastie, T. (2005). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society, Series B,* ***67(2)***, 301-320.

Datasets

CN-AML: The CN-AML data used in Section 14 with gene expressions from patients with cytogenetically normal acute myeloid leukemia (CN-AML) was provided by U. Mansmann (IBE, Munich). The data is analyzed e. g. in Benner et al. (2010) and Metzeler et al. (2008).

Leukemia: The leukemia data of adult myeloid leukemia patients in northwest England used in Section 13 was originally provided by Leonhard Held (University of Zurich, UZU). The data is also analyzed e. g. in Henderson et al. (2002) and Kneib and Fahrmeir (2007).

PBC liver: The primary biliary cirrhosis of the liver (PBC liver) data used in Section 12 is provided and described e.g. in the `R`-package `survival` (`pbc{survival}`) or on the book-homepage of Therneau and Grambsch (2000). The data is also analyzed e. g. in Tibshirani (1997), Hoeting et al. (1999), Therneau and Grambsch (2000), Sleeper and Harrington (1990), Johnson (2008), Johnson (2009) and Fahrmeir and Kneib (2011).

Software and packages

Belitz, C., Brezger, A., Kneib, T., Lang, S. and Umlauf, N. `BayesX`: *Bayesian Inference in Structured Additive Regression Models*. URL http://www.stat.uni-muenchen.de/~bayesx. For details, compare e. g. http://www.stat.uni-muenchen.de/~bayesx/manual/reference_manual.pdf and http://www.stat.uni-muenchen.de/~bayesx/manual/methodology_manual.pdf.

R Development Core Team. `R`: *A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL http://www.R-project.org/.

Komárek, A. `R-package {bayesSurv}`*: Bayesian Survival Regression with Flexible Error and Random Effects Distributions.* URL http://CRAN.R-project.org/package=bayesSurv. Details are available e. g. in Komárek et al. (2007).

Goeman, J., J. `R-package {penalized}`: *L1 (lasso and fused lasso) and L2 (ridge) penalized estimation in Generalized Linear Models and in the Cox model.* URL http://CRAN.R-project.org/package=penalized. Details are available in Goeman (2010).

Park, M. Y. and Hastie, T. `R-package {glmpath}`*: L1 Regularization Path for Generalized Linear Models and Cox Proportional Hazards Model.* URL http://CRAN.R-project.org/package=glmpath. Details are available in Park and Hastie (2007).

Schellhase, C. `R-package {pendensity}`: *Density Estimation with a Penalized Mixture Approach*. URL http://CRAN.R-project.org/package=pendensity. Details are available in Schellhase and Kauermann (2011).

Neal, R. M. `R-function uni.slice()`: Implementation of the simple slice sampler, URL http://www.cs.toronto.edu/~radford/ftp/slice-R-prog). Details are available in Neal (2003).

Perez Rodriguez, P. `R-package {ars}`: *Adaptive Rejection Sampling*. Original C++ code from A. Komárek based on `ars.f` written by P. Wild and W. R. Gilks. URL http://CRAN.R-project.org/package=ars. Details are available in Gilks and Wilde (1992).

Jackson, C. `R-package {msm}`: Multi-state Markov and hidden Markov models in continuous time. http://CRAN.R-project.org/package=msm. The included function `tnorm()` is used for simulation of truncated normal variables. Details concerning this function are available in Robert (1995).

EIDESSTATTLICHE VERSICHERUNG

(Siehe Promotionsordnung vom 12.07.11, § 8, Abs. 2 Pkt. .5.)

Hiermit erkläre ich an Eidesstatt, dass die Dissertation von mir selbstständig, ohne unerlaubte Beihilfe angefertigt ist.

Konrath, Susanne
Name, Vorname

Mering, den 16.04.2013
Ort, Datum

Unterschrift Doktorandin