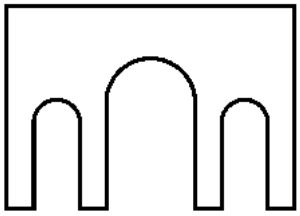


Dissertation an der Fakultät für Mathematik, Informatik
und Statistik der Ludwig-Maximilians-Universität München

Mixed model based inference in structured additive regression

Thomas Kneib

München, 8. November 2005

Erstgutachter:	Prof. Dr. Ludwig Fahrmeir
Zweitgutachter:	Prof. Dr. Helmut Küchenhoff
Drittgutachter:	Prof. Dr. Göran Kauermann
Rigorosum	22. Februar 2006

Für immer die Menschen

Damit sich meine Promotion auch für die vielen hilfreichen Geister gelohnt hat, die mir während ihrer Entstehung zur Seite gestanden haben, sind sie hier alle versammelt, die Helden der Arbeit, die Mitarbeiter der Woche, die Anwohner der Straße der Besten. Für immer die Menschen:

- Ludwig Fahrmeir, der mir eine optimale Förderung zukommen ließ und dessen kollegiale Betreuung sehr zur Schaffung einer stressfreien Arbeitsatmosphäre und damit wiederum zur zügigen Erstellung dieser Arbeit beigetragen hat (auch wenn das für manchen wie ein Widerspruch klingen mag).
- Stefan Lang, von dessen Lob und Kritik ich gleichermaßen profitiert habe und der mir die strukturierte Lebensführung am eigenen Beispiel demonstrierte.
- Susanne Heim, die in unserem gemeinsamen Büro meine täglichen Erfolge und Misserfolge miterleben musste, gegen deren Realismus ich mich immer durch positive Illusionen absetzen konnte, die sich als die kritischste Leserin dieser Arbeit profilierte und die sich bemühte, mir die natürlichen Attraktionen Bayerns näherzubringen.
- Andrea Hennerfeind & Andi Brezger, für die unermüdliche Suche nach Fehlern in meinen Programmteilen, sowie die Hilfe bei diversen kleineren und größeren BayesX-Problemen.
- Günter Rasser, von dessen Forschungsvermeidungsstrategien ich heute noch profitiere und der mich vorbildlich in das Doktorandenleben einwies.
- Susanne Breitner, Michael Höhle & Stefan Krieger, die von Anfang an mit dabei waren, ohne die das tägliche Mittagessen definitv ärmer gewesen wäre und die zu jeder Gelegenheit bereit waren, mit mir über überflüssiges Wissen zu diskutieren.
- Leonhard Held, der stets bemüht war, meinen eingeschränkten Blick auf die Statistik zu erweitern und mir Begeisterung für statistische Sujets zu vermitteln ("Du hast bestimmt noch nie von Skorokhods Repräsentations-Theorem gehört ...").
- Christiane Belitz, Alexander Jerak & Leyre Osuna, die tatkräftig das BayesX-Projekt unterstützten.
- Manuela Hummel, die im Rahmen ihrer Diplomarbeit umfangreiche Simulationen durchführte, deren Ergebnisse die Grundlage für Kapitel 8 dieser Arbeit bilden.
- Conny Oberhauser, die als studentische Hilfskraft die Implementation einiger Erweiterungen für kategoriale Regressionsmodelle übernahm.
- Rudi Eichholz der im Rahmen seiner Diplomarbeit eine Reihe von Analysen der Nigeriadaten durchführte.
- Bärbel und Gottfried Kneib, ohne die ich heute sicher nicht hier wäre.
- Daniela Kropf, für Mut, Aufregung, Ablenkung und Gemeinsamkeiten.

Finanziell gefördert wurde meine Arbeit durch die Deutsche Forschungsgemeinschaft im Rahmen des Sonderforschungsbereichs 386 (Analyse diskreter Strukturen).

Thomas Kneib

Zusammenfassung

Regressionsdaten weisen immer häufiger zusätzlich zu den üblichen Kovariableneffekten räumliche oder räumlich-zeitliche Strukturen auf, so dass adäquate Modellerweiterungen in vielen komplexeren Anwendungen benötigt werden. Ein flexibler Ansatz sollte dabei nicht nur erlauben, räumliche und zeitliche Korrelationen zu berücksichtigem, sondern darüberhinaus die semi- oder nonparametrische Modellierung weiterer Kovariableneffekte zulassen. Da spezifische Regressionsmodelle für verschiedene Klassen von abhängigen Variablen entwickelt wurden, müssen die semiparametrischen Erweiterungen speziell an die jeweilige Situation angepasst werden.

Im Rahmen dieser Arbeit werden zahlreiche Möglichkeiten zur Modellierung komplexer Kovariableninformation wiederholt und im einheitlichen Rahmen von strukturiert additiven Regressionsmodellen zusammengefasst. Insbesondere können nichtlineare Effekte stetiger Kovariablen, zeitlich korrelierte Effekte, räumlich korrelierte Effekte, komplexe Interaktionen oder unbeobachtete Heterogenität berücksichtigt werden. Beginnend mit Regressionsmodellen für abhängige Variablen aus Exponentialfamilien werden Erweiterungen für verschiedene Typen kategorialer Responsevariablen und zur Analyse stetiger Überlebenszeiten beschrieben. Ein neues Inferenz-Konzept, das auf der Verwendung von Methodik für Modelle mit zufälligen Effekten beruht wird eingeführt. Dies erlaubt die Behandlung der verschiedenen Regressionsprobleme in einem einheitlichen Ansatz basierend auf penalisierter Likelihood-Schätzung für die Regressionskoeffizienten und Restricted Maximum Likelihood beziehungsweise marginaler Likelihood Schätzung für die Glättungsparameter. Das neue Schätzverfahren wird in einer Reihe von Anwendungsbeispielen und Simulationsstudien untersucht und erweist sich als vielversprechende Alternative zu konkurrierenden Ansätzen, insbesondere der Schätzung basierend auf Markov Chain Monte Carlo Simulationsverfahren.

Abstract

Due to the increasing availability of spatial or spatio-temporal regression data, models that allow to incorporate the special structure of such data sets in an appropriate way are highly desired in practice. A flexible modeling approach should not only be able to account for spatial and temporal correlations, but also to model further covariate effects in a semi- or nonparametric fashion. In addition, regression models for different types of responses are available and extensions require special attention in each of these cases.

Within this thesis, numerous possibilities to model non-standard covariate effects such as nonlinear effects of continuous covariates, temporal effects, spatial effects, interaction effects or unobserved heterogeneity are reviewed and embedded in the general framework of structured additive regression. Beginning with exponential family regression, extensions to several types of multicategorical responses and the analysis of continuous survival times are described. A new inferential procedure based on mixed model methodology is introduced, allowing for a unified treatment of the different regression problems. Estimation of the regression coefficients is based on penalized likelihood, whereas smoothing parameters are estimated using restricted maximum likelihood or marginal likelihood. In several applications and simulation studies, the new approach turns out to be a promising alternative to competing methodology, especially estimation based on Markov Chain Monte Carlo simulation techniques.

Contents

Part I

Introduction

1 Regression models

One of the main objectives of statistical modeling is to quantify the influence of variables (called covariates) on a measure of interest (the so called dependent variable or the response). A general framework to perform such analyses is provided by regression models which have been developed for a variety of response types. The most prominent regression model is the classical linear model, where the response variable y is assumed to be Gaussian distributed and the covariates $x_1, \ldots, x_p$ act linearly on the response. More specifically, we assume that the following equation holds for the (conditional) expectation of y:

$$E(y|x_1, \ldots, x_p) = \gamma_0 + x_1\gamma_1 + \ldots + x_p\gamma_p = \eta. \tag{1.1}$$

The unknown parameters $\gamma_1, \ldots, \gamma_p$ are called regression coefficients and determine the strength and direction of the corresponding covariate‘s influence. Since (1.1) is linear in the regression coefficients, the sum of the covariate effects η is usually referred to as the linear predictor.

In case of nonnormal responses, a direct connection between the expectation of y and the linear predictor η is not possible, since the domain of $E(y|x_1, \ldots, x_p)$ is no longer the real line. Therefore, the identity link in (1.1) is replaced by a more general transformation h to ensure the correct domain:

$$E(y|x_1, \ldots, x_p) = h(\gamma_0 + x_1\gamma_1 + \ldots + x_p\gamma_p). \tag{1.2}$$

While retaining the assumption of linearity for the influences of the covariates, a nonlinear relationship between η and the expectation of y is introduced if $h \neq id$. For univariate responses with a distribution belonging to an exponential family, models of the form (1.2) are called generalized linear models. Different subtypes are obtained with specific choices for the distribution of y and the response function h. More details on generalized linear models will be presented in Section 4.1.1.

Multivariate versions of generalized linear models allow for regression analyses of multivariate responses and, in particular, of multicategorical responses. In principle, the same structure as in (1.2) is assumed but since the response is multivariate, h is also a multivariate function relating y to a multivariate vector of linear predictors. Such models will be discussed in Section 10.1.

For the analysis of survival or other duration times, special regression models have been developed. The model most commonly used in practice is the Cox model, where the covariates determine the hazard rate of the response instead of the expectation. The Cox model expresses the hazard rate as the product of an unspecified baseline hazard rate $\lambda_0(t)$ not depending on covariates and the exponential of a linear predictor not depending on time, i. e.

$$\lambda(t|x_1, \ldots, x_p) = \lambda_0(t) \exp(x_1\gamma_1 + \ldots + x_p\gamma_p). \tag{1.3}$$

Similarly as in (1.1) and (1.2), all covariates are assumed to affect $\lambda(t|x_1, \ldots, x_p)$ in a linear manner. Note that no intercept is included since it can be absorbed into the baseline hazard rate which is of unspecified functional form.

Obviously, a variety of regression models exist for different types of responses allowing for appropriate modeling of the response distribution. Hence, this thesis will not deal

with new types of regression models in the sense that models for new response distributions are introduced. Instead we aim at a more flexible modeling of different types of covariates. Due to the increasing availability of highly complex regression data, such extensions are clearly needed in practice to obtain valid and realistic statistical models that describe relationships between variables of interest adequately. A specific example which has gained considerable attention in the last years are regression models with spatial or spatio-temporal structures. Such spatio-temporal features can hardly be handled within the linear parametric frameworks discussed above and therefore require suitable forms of covariate modeling. In the next section, we will discuss the necessity of extended regression models and the insufficiency of the classical approaches by means of the applications that will be analyzed in full detail later on.

2 Applications

2.1 Childhood undernutrition in Zambia

Undernutrition is considered one of the most urgent challenges of underdeveloped countries since widespread undernutrition is not only a problem of public health but also causes low labor productivity and, hence, has significant impact on further relevant development outcomes. A framework that allows to identify determinants of undernutrition are regression models with an appropriately defined indicator of undernutrition as the response variable.

Childhood undernutrition is usually determined by assessing the anthropometric status of a child i relative to a reference standard in terms of a Z-score

$$Z_i = \frac{AI_i - \mu_{AI}}{\sigma_{AI}}, \tag{2.1}$$

where AI refers to the child's anthropometric indicator, and μ_{AI} and σ^2_{AI} refer to the median and the standard deviation of the reference population, respectively. Depending on the specific choice for AI, different types of undernutrition can be considered. For example, acute undernutrition is indicated by insufficient weight for height. In this application we will analyze stunting or insufficient height for age indicating chronic undernutrition. In this case, AI is given by height at a certain age.

Our analysis will be based on data collected in Zambia, where stunting rates are generally high with 42% of the children being classified as stunted (Z-score less than minus 2) and 18% as severely stunted (Z-score less than minus 3). The 1992 Zambia Demographic and Health Survey (DHS, Gaisie, Cross & Nsemukila 1993) provides information on a representative sample of $n = 4847$ children including characteristics of the child's parents (education, income, nutritional situation of the parents), characteristics of the child itself (e. g. age), and variables describing the environment in which the child lives (access to clean water, locality of the residence). Table 2.1 contains a description of all variables that will be pursued in the regression analysis in Section 7.

Variable	Description
hazstd	standardized Z-score
bmi	body mass index of the mother
agc	age of the child in months
district	district where the child lives
rcw	mother's employment status with categories "working" (= 1) and "not working" (= −1)
edu1, edu2	mother's educational status with categories "complete primary but incomplete secondary" (edu1=1), "complete secondary or higher" (edu2=1) and "no education or incomplete primary" (edu1=edu2=-1)
tpr	locality of the domicile with categories "urban" (= 1) and "rural" (= −1)
sex	gender of the child with categories "male" (= 1) and "female" (= −1)

Table 2.1: Undernutrition in Zambia: Description of the variables.

To build a regression model for undernutrition, we first have to define a distribution for the response variable. In the present example, it seems reasonable to assume that the Z-score is (at least approximately) Gaussian distributed and, thus, model (1.1) could in principle be applied. However, a thorough investigation of hypothesis about the determinants of undernutrition requires a more flexible modeling of covariate effects. For example, the influence of the body mass index (BMI) of the mother is often expected to be inverse u-shaped. Parents with low BMI values are themselves malnourished and are therefore likely to have undernourished children. At the same time, very high BMI values indicate poor quality of the food and, hence, may also imply malnutrition of the children. Likewise regarding the age of a child, a nonlinear, monotonically decreasing form of the effect is expected since the children are usually born with almost normal anthropometric status. Afterwards, the health status of the children is expected to worsen for a certain time until it stabilizes at a low level. However, the exact shape of the influences is unknown and, hence, no simple model can be established to link the undernutrition score to the covariates.

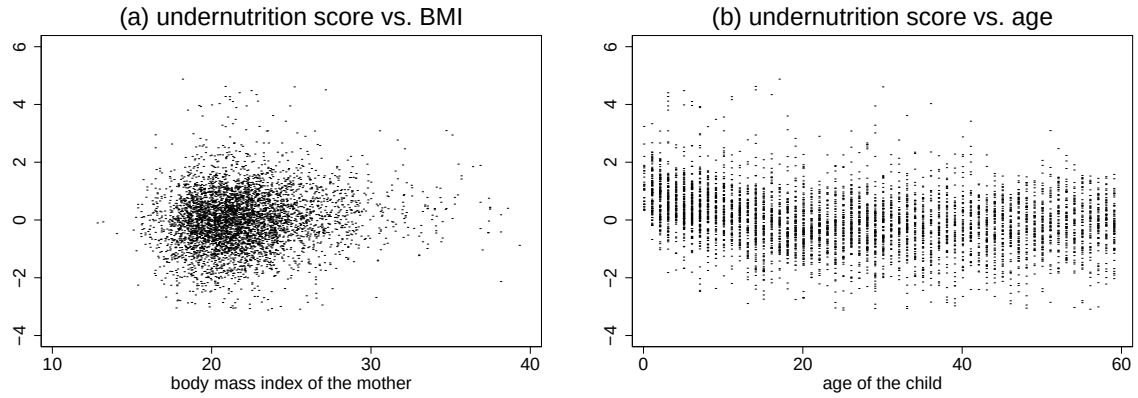


Figure 2.1: Undernutrition in Zambia: Scatterplots of the undernutrition score versus the body mass index of the mother (a) and the age of the child (b).

Figure 2.1 displays scatterplots of the undernutrition score versus the body mass index (BMI) of the mother and the age of the child. For the BMI, the plot does not reveal a clear pattern of the relationship. In contrast, a nonlinear pattern as described above emerges for the age effect. In this case, the reduction to a linear effect may result in false conclusions since nonlinearity is only present in a small part of the age-domain.

As a further problem of the undernutrition data, only a small number of covariates characterizes the environmental conditions of a child. However, these (partly unobserved) covariates may be important influential factors and ignoring them could induce considerable correlations among the observations. This in turn contradicts the basic assumption of independent observations which is routinely made in regression models. Figure 2.2 visualizes the average undernutrition score for the districts within Zambia indicating a clear spatial pattern with better nourished children living in the southern part of Zambia and malnourished children living in the north-eastern part. Note that no exact spatial information on the residence of the children is available and a regression model can only utilize the discrete spatial information in which of the districts a child is living. The dashed regions denote parts of the map where no observations were collected. Therefore,

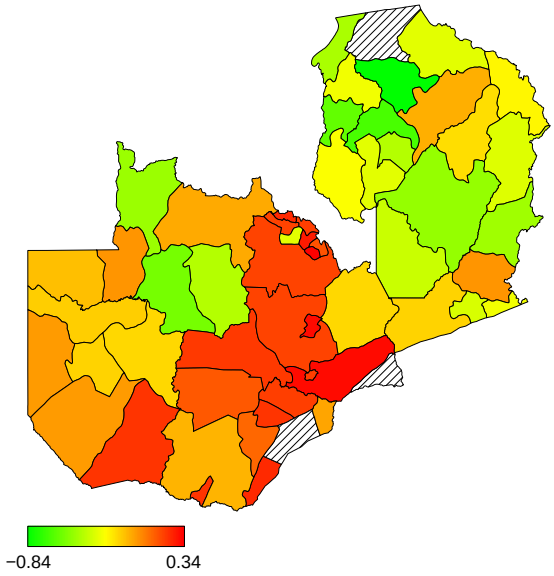


Figure 2.2: Undernutrition in Zambia: Average undernutrition score within the districts of Zambia.

an additional requirement for spatial regression models is the ability to deal with missing data.

A way to overcome both the problem of unknown functional forms of covariate effects and correlations due to unobserved spatially varying covariates, is to introduce a more flexible, geoadditive regression model of the form

$$E(hazstd|\cdot) = f_1(agc) + f_2(bmi) + f_{spat}(district) + \gamma_1 \cdot rcw + \ldots \tag{2.2}$$

The functions f_1 and f_2 should be flexible enough to allow for general functional forms of the covariate effects but, on the other hand, should not overfit the data to retain interpretability. Several modeling strategies fulfilling both requirements will be discussed in Section 4.2.2. All of them will depend on a smoothing parameter that controls the compromise between smoothness of the function estimates and closeness to the data. Hence, an important part of the estimation procedure will be to derive estimates for this smoothing parameters.

The spatial function f_{spat} defined upon the districts of Zambia serves as a surrogate for unobserved, spatially varying covariates and, thus, allows to adjust for possible spatial correlations in the data. Different specifications of spatial effects will be discussed in Section 4.2.3. Similar as for the nonparametric effects, a smoothing parameter will play an important role when estimating the spatial effect.

2.2 Forest health data

Vital forests play an important role in the ecosystem due to their regulatory impact on both the climate and the water cycle. However, a prerequisite for these balancing functions is a sufficiently healthy forest. In this application, we will analyze data on forest health to identify potential factors influencing the health status of trees.

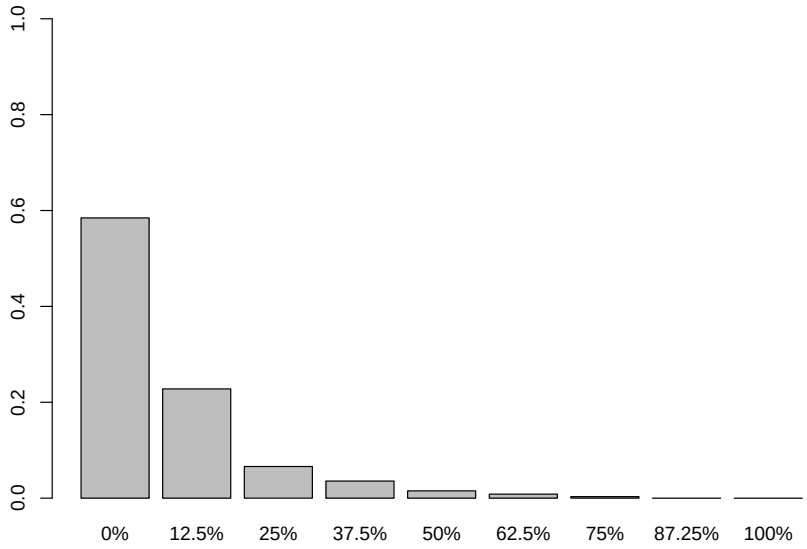


Figure 2.3: Forest health data: Histogram of the nine damage states.

The available data have been collected in annually visual forest inventories between 1983 and 2004 in a northern Bavarian forest district. The health status of 83 beeches within an observation area exhibiting 10 km from south to north and 15 km from west to east was assessed on an ordinal scale. The nine possible categories denote different degrees of defoliation. The scale is divided in 12.5% steps, ranging from healthy trees (0% defoliation) to trees with 100% defoliation. Figure 2.3 shows a histogram of the nine damage states indicating that severe damage is relatively rare. Therefore we aggregated the health status in three categories corresponding to healthy trees (0% defoliation), slightly damaged trees (between 12.5% and 37.5% defoliation), and severely damaged trees (more than 37.5% defoliation).

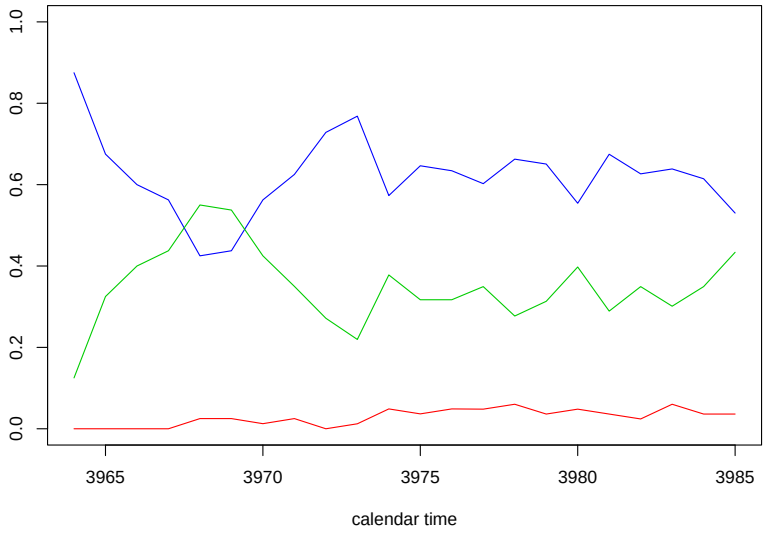


Figure 2.4: Forest health data: Temporal development of the frequency of the three different damage states.

Since the data set has a longitudinal structure consisting of 83 time series of damage states for the 83 trees, temporal correlations have to be considered appropriately. In addition,

the trees are located within a relatively small observation area and spatial correlations are also likely to be encountered. Figure 2.4 shows the temporal development of the frequency of the three different damage states. Obviously, only a small percentage of the trees falls into the class with highest defoliation degree. However, there seems to be a slightly increasing trend for this category. The lowest percentage of healthy trees is observed at the end of the eighties. Afterwards, the population seems to recover, stabilizing at a percentage of approximately 60% of healthy trees in the nineties. All trends are relatively rough due to random variation in the data. Using a (nonparametric) regression model to estimate the trends allows to reduce the disturbance and to obtain more reliable estimates.

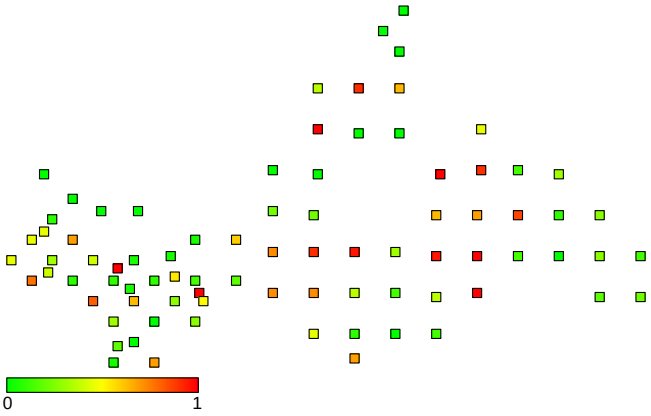


Figure 2.5: Forest health data: Percentage of years for which a tree was classified to be slightly or severely damaged, averaged over the entire observation period.

Figure 2.5 visualizes the distribution of the trees across the observation area. Each box corresponds to a tree and is colored according to the percentage of years for which the tree was classified to be damaged. Accordingly, green boxes correspond to healthy trees whereas red points indicate trees which are classified to be damaged in most years. Around Rothenbuch (the hole in the easterly part of the observation area) there seems to be an increased amount of damaged trees. Employing spatial smoothing techniques will allow us to obtain clearer spatial patterns of healthy and damaged trees. In contrast to the data set on childhood undernutrition, the spatial information in the forest health example consists of exact longitudes and latitudes for the positions of the trees on the lattice map. Hence, different types of spatial smoothing techniques may be required.

In addition to temporal and spatial information, numerous other covariates characterizing the stand and the site of the tree, as well as the soil at the stand are available (see Table 2.2). The set of covariates comprises both categorical and continuous covariates. For the latter, we have to decide whether to include them in a parametric way or, in analogy to the undernutrition example, in a nonparametric way, i. e. as smooth functions $f_j(x_j)$. In Section 12 it will turn out, that some effects of continuous covariates can be adequately approximated by linear effects. This especially concerns effects of tree-specific covariates that do not vary over time.

In principle, temporal correlations can be included in a similar way as the nonparametric effects based on a function $f(t)$ of the calendar time. However, in the present example, the time trend is expected to be distinct for trees of different age due to biological considerations. In Figure 2.6 the temporal development of the percentage of damaged trees

Covariate	Description
age	age of the tree (continuous)
time	calendar time (continuous)
elevation	elevation above sea level (continuous)
inclination	inclination of slope (continuous)
soil	depth of soil layer (continuous)
ph	pH-value in 0-2cm depth (continuous)
canopy	density of forest canopy (continuous)
stand	type of stand (categorical, 1=deciduous forest, -1=mixed forest).
fertilization	fertilization (categorical, 1=yes, -1=no).
humus	thickness of humus layer in 5 categories (ordinal, higher categories represent higher proportions).
moisture	level of soil moisture (categorical, 1=moderately dry, 2=moderately moist, 3=moist or temporary wet).
saturation	base saturation (ordinal, higher categories indicate higher base saturation).

Table 2.2: Forest health data: Description of covariates.

is displayed for young, middle-aged and old trees. Obviously, the three resulting time trends share some common features, e. g. the peak at the end of the eighties, but there is also a clear evidence of a time-varying difference between the trends. To account for this, we might include several time trends for different age groups. A suitable framework is given by varying coefficient terms as presented in Section 4.2.5. However, a drawback of this approach is the arbitrary definition of the age groups. In addition, employing a large number of age classes to obtain a more realistic model may lead to unstable estimates. Therefore, a more sophisticated idea is to add an interaction surface between age and calendar time to the regression model, i. e. a smooth function $f(age, t)$, see Section 4.2.6 for methodological details.

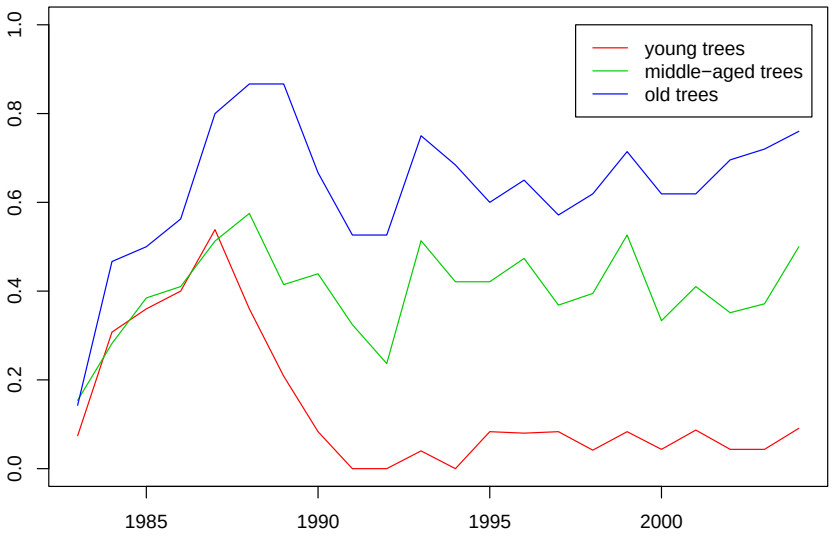


Figure 2.6: Forest health data: Temporal development of the percentage of damaged trees for different ages.

Since the damage states of the trees are not measured on a continuous but a categorical scale with ordered classes, multivariate regression models for categorical responses have to be applied. In such models, predictors of a similar form as in (2.2) are defined separately for each category r. The problems discussed so far suggest a model of the form

$$\eta^{(r)} = \ldots + f_1(t) + f_2(age) + f_3(age, t) + f_{spat}(s) + \ldots, \tag{2.3}$$

see Section 10.1.3 for details on regression models for ordered categories. The smooth functions $f_1(t)$ and $f_2(age)$ account for temporal correlations and nonlinear age effects. According to the above-mentioned considerations, an interaction surface $f_3(age, t)$ between calendar time and age, and a spatial effect $f_{spat}(s)$ are additionally included. Note that in contrast to the childhood undernutrition example, in this case $s = (s_x, s_y)$ denotes exact spatial positions on a lattice map instead of regions.

In Equation (2.3), all effects are assumed to be independent of the category index. Of course this is a rather restrictive assumption, mainly introduced to simplify the regression model. However, the trends visualized in Figure 2.4 indicate that the temporal development obeys significant differences between the damage states. Thus, extended versions of (2.3) allowing for category-specific effects are desired. Conceptually, such models are easily defined by predictors of the form

$$\eta^{(r)} = \ldots + f_1^{(r)}(t) + \ldots$$

but the numerical complexity increases rapidly and additional algorithmic work has to be done. We will discuss such extensions in Section 10.1.3.3.

2.3 Leukemia survival data

As a first example on survival analysis, we will consider a data set on leukemia survival times in Northwest England described by Henderson, Shimakura & Gorst (2002). The data set contains information on all 1,043 cases of acute myeloid leukemia in adults that have been diagnosed between 1982 and 1998 in Northwest England. Continuous covariates include the age of the patient at diagnosis, the white blood cell count (wbc) at diagnosis and the Townsend deprivation index (tpi) which measures the deprivation of the enumeration district of residence. Positive values of this index indicate poorer regions while negative values correspond to wealthier regions. Since the observation area consists of 8,131 enumeration districts, the Townsend index can be considered a subject-specific covariate. The sex of a patient is included in dummy-coding (1=female, 0=male). Spatial information on the residence of a patient is available in form of exact locations in terms of longitude and latitude, but of course we can also aggregate this information to district-level. Figure 2.7 shows the district boundaries together with the exact locations of the observed cases.

The analyses presented Henderson et al. (2002) concentrated on the detection of spatial variation in survival times but retained the assumption of a linear predictor for covariate effects. In Section 16 we will investigate whether this assumption holds by extending the basic hazard rate model (1.3) to

$$\lambda(t) = \lambda_0(t) \exp\left[sex \cdot \gamma_1 + f_1(age) + f_2(wbc) + f_3(tpi) + f_{spat}(s)\right], \tag{2.4}$$

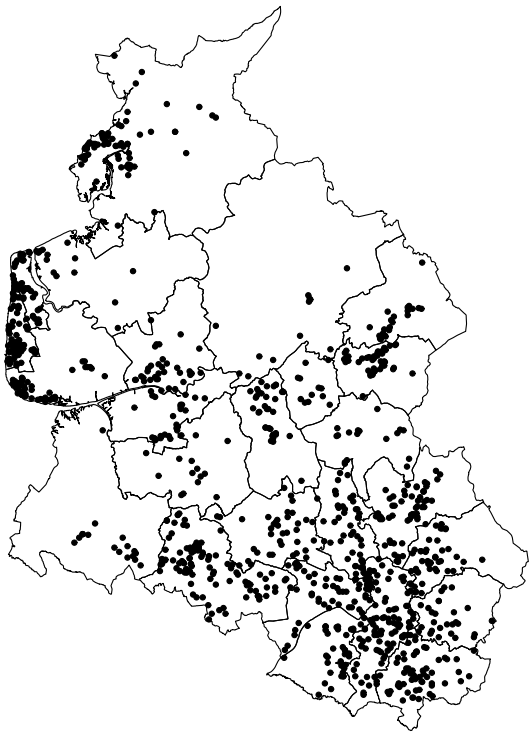


Figure 2.7: Leukemia survival data: Districts of Northwest England and locations of the observations.

where $f_1, \ldots, f_4$ denote smooth nonparametric functions of the corresponding continuous covariates. The spatial effect f_{spat} can be defined upon the districts the individuals are living in as well as upon exact coordinates of the residences.

In addition to these extensions, regression models for hazard rates introduce additional challenges. For example, the baseline hazard rate may be of interest and, thus, is to be estimated simultaneously with the other covariate effects. This is especially important if survival times shall be predicted from the estimated model which is frequently the case in medical applications where expected survival times for the patients have to be predicted. A further difficulty in survival models are time-varying effects of covariates. Such time-varying effects are often expected for effects of medical treatments that are likely to loose their impact over time. In our example we might investigate whether the gender effect varies over time, i. e. whether the survival of males and females follows different nonproportional hazard rates.

The two problems may be tackled by adding additional terms to the predictor (2.4):

$$\lambda(t) = \exp\left[g_0(t) + sex \cdot g_1(t) + f_1(age) + f_2(wbc) + f_3(tpi) + f_{spat}(s)\right]. \tag{2.5}$$

The baseline hazard rate $\lambda_0(t)$ is absorbed into the predictor and represented by the log-baseline $g_0(t)$. The time-varying gender effect is included by replacing the time-constant regression coefficient γ_1 with a time-varying function $g_1(t)$. Based on model (2.5), the baseline hazard rate for males is given by $\exp(g_0(t))$ while for females it is given by $\exp(g_0(t) + g_1(t))$.

The analysis of survival times is not only complicated by the additional model terms that have to be considered, but also by the fact that usually only incomplete data is observed.

The most commonly known phenomenon of this kind is right censoring, i. e. no event occurred for some of the individuals. In this case, we only know that the corresponding individual has survived up to a certain time. In the present example, almost 16% of the observations are right censored. In other words, 16% of the patients did not die from leukemia during the time the study was conducted. Incomplete data requires additional assumptions on the data generating mechanism and also results in more complicated likelihood contributions (see Section 14.2).

2.4 Childhood mortality in Nigeria

As a second example on regression models for survival times, we examine data on childhood mortality in Nigeria. These data have been collected within the 2003 Nigeria Demographic and Health Survey (DHS, National Population Commission (Nigeria) and ORC Macro 2004), a nationally representative survey concerning the health status of women in reproductive age (13–49 years) and their children. The survival times of the children are gathered from retrospective interviews of their mothers and should (in theory) be measured in days. Hence, a continuous time survival model seems to be appropriate.

Numerous covariates are available in addition to the survival times. Besides spatial information on the district the children are living in, we will differentiate between categorical covariates that will be modeled in a parametric way and covariates that will be modeled nonparametrically. The former are described in detail in Table 2.3. The latter comprise the body mass index of the mother (bmi), the age of the mother at birth (age), the number of the child in the birth order ($bord$) and the number of household members ($size$).

To identify covariates that influence the survival of children in underdeveloped countries, we would now like to consider a geoadditive regression model for the hazard rate of the same form as in the previous application on leukemia survival times, i. e.

$$\lambda(t) = \exp\left[g_0(t) + f_1(bmi) + f_2(age) + f_3(bord) + f_4(size) + f_{spat}(s) + u'\gamma\right], \qquad (2.6)$$

where g_0 denotes the log-baseline hazard rate, $f_1, \ldots, f_4$ are flexible nonparametric functions, f_{spat} is a spatial function and u comprises all further categorical covariates. However, direct application of model (2.6) is hindered by the fact that the covariate breastfeeding is time-varying. This variable takes the value one, as long as the child is breastfed, and zero otherwise. Of course, variables that are modeled nonparametrically or the spatial variable may also be time-varying, although we do not observe this problem here. To account for time-varying covariates, additional conceptual work has to be done. In Section 14.2 we will show how piecewise constant time-varying covariates can be included in hazard regression models based on data augmentation.

An additional challenge of the Nigerian survival data are further types of incomplete data that have to be considered. Similar as in the previous section, right censoring is introduced by children that were still alive at the end of the study time. In the Nigeria data set, the population of interest are children up to an age of five and, hence, most observations are right censored since only a small percentage of children dies within the first five years of their life. However, a second type of incomplete data is present for the uncensored survival times. Although these survival times should in theory be given in days, most of

them are actually rounded due to memory effects introduced by the retrospective design of the study. Only survival times within the first two months are observed exactly while all remaining survival times are actually given in months. In contrast, right censoring times are all given in exact days, since these were computed from the date of the interview and the child's birth date.

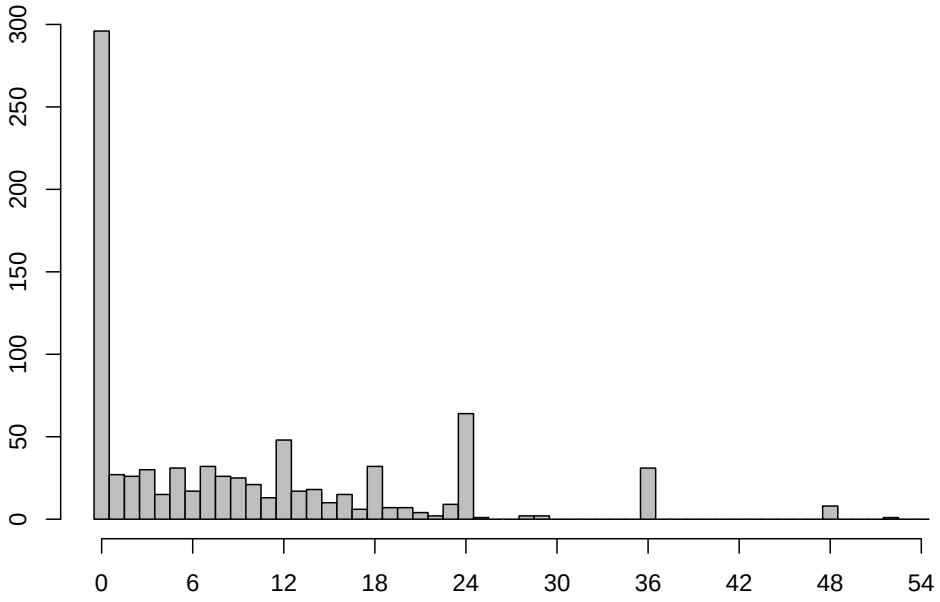


Figure 2.8: Childhood mortality in Nigeria: Frequencies of observed survival times in months.

In addition to the rounding to months, there is a second rounding mechanism illustrated in Figure 2.8, which shows the absolute frequencies of the observed survival times in months. Obviously, a lot of survival times are heaped at the values 12, 18, 24, 36 and 48 while a much smaller number of deaths is recorded between these time points. Such a heaping effect is frequently encountered in retrospective studies on survival times and has to be incorporated appropriately to obtain valid estimates.

From a statistical point of view, both types of rounded survival times can be considered as being interval censored. This means that the event of interest is only known to fall between two time points and no exact event time is observed. In particular, all survival times exceeding two months are treated as interval censored, where the interval is determined by the first and the last day of the corresponding month. For the survival times rounded to 12, 18, 24, 36 and 48 months wider intervals have to be defined. For example, we may assume that all survival times rounded to 36 months had their corresponding event between the time points 30 and 42, i. e. we assume a symmetric interval of 12 months length around this time point.

The introduction of interval censoring into survival models does not change the model definition (2.6) but leads to more complicated likelihood contributions. Right and interval censoring, and further types of incomplete data will be considered in full detail in Section 14.2.

Covariate	Description	Coding
education	educational level of the mother	1=no education, 0=at least primary education
employment	employment status of the mother	1=currently employed, 0=currently unemployed
religion1, religion3	religion of the mother	christian (religion1=1), muslim (religion1=religion3=0), other (religion3=1)
delivery	place of delivery	1=hospital, 0=at home
assistance	assistance at delivery	1=assistance, 0=no assistance
longbirth	long birth (regular contractions last more than 12 hours)	1=yes, 0=no
bleeding	excessive bleeding at birth	1=yes, 0=no
fever	high fever at birth	1=yes, 0=no
convulsion	convulsions not caused by fever	1=yes, 0=no
sex	gender of the child	1=male, 0=female
breastfeeding	time-varying covariate indicating whether the child is currently breastfed	1 = child is breastfed, 0 = child is not breastfed
weight2,...,weight5	birth weight of the child	very small (weight5=1), small (weight4=1), normal (weight3=1), large (weight2=1), very large (weight2=...=weight4=0)
wealth2,...,wealth5	wealth of the household	very rich (wealth5=1), rich (wealth4=1), normal (wealth3=1), poor (wealth2=1), very poor (wealth2=...=wealth5=0)
urban	place of residence	1=urban, 0=rural
water	quality of water supply	1=good, 0=bad
toilet	availability of toilet facility	1=toilet available, 0=no toilet
electricity	availability of electricity	1=electricity available, 0=no electricity
floormaterial	quality of floor material	1=good, 0=bad
initial1, initial2	time when the child was first breastfed	immediately (initial1=1), within 24 hours (initial2=1), later than 24 hours (initial1=initial2=0)

Table 2.3: Childhood mortality in Nigeria: Description of categorical covariates.

3 Outline

Having discussed the problems associated with different types of regression data, the aims of this thesis can be summarized as follows:

- Provide background knowledge on how to account for nonstandard covariate effects, such as nonlinear effects of continuous covariates, temporal effects, spatial effects, interaction effects or unobserved heterogeneity.
- Embed the different types of effects in one unifying framework (structured additive regression).
- Describe how to adapt structured additive regression to different types of regression data, such as univariate responses from exponential families in the context of generalized linear models, categorical responses in multivariate generalized linear models and survival times in Cox-type regression models.
- Introduce an inferential procedure allowing for the estimation of all covariate effects in the different types of regression models in a unified way. This procedure will be based on a mixed model representation of structured additive regression models.
- Investigate the performance of the mixed model approach both in terms of simulation studies and real data applications.

According to the three different types of responses that will be considered, the major part of this work is split into three parts: Part II contains details on regression models for univariate responses from exponential families, Part III presents multicategorical extensions and Part IV deals with the analysis of continuous survival times. Each of the three parts starts with a section describing the respective model in greater detail, beginning with parametric versions and introducing semiparametric extensions afterwards. In Part II this discussion will also comprise an extensive treatment of the different covariate effects included in structured additive regression models. These descriptions will not be repeated in Parts III and IV.

The second section of each of the three parts is dedicated to the presentation of inferential procedures for the respective model. Markov Chain simulation techniques are frequently employed for estimation in semiparametric regression models. We will refer to this as a fully Bayesian estimation strategy since all parameters are treated as random variables and are estimated simultaneously. The focus in this thesis, however, concerns a different estimation technique. The main idea is to reparametrize semiparametric regression models as mixed models and to apply or adapt methodology developed for such mixed models. In particular, smoothing parameters can be estimated based on restricted maximum likelihood or marginal likelihood estimation techniques. This corresponds to a differentiation between parameters of primary interest (the regression parameters) and hyperparameters (the smoothing parameters). While priors are specified for the former, the latter are treated as fixed and estimated in advance from the data. Therefore, this estimation procedure can be considered as an empirical Bayes approach but also has a close connection to penalized likelihood estimation.

As an advantage, problems usually arising in fully Bayesian estimation based on MCMC are not present here. This includes the question on how to determine the burn-in phase and the convergence of the generated Markov chain. In addition, there is no sensitivity with respect to hyperpriors since the hyperparameters are treated as unknown constants and no priors are imposed on them. As a drawback, empirical Bayes methods do not take into account the variability introduced by the estimation of the hyperparameters. In contrast, these are estimated from the data and afterwards inserted into the estimation formulae for the regression coefficients as if they were known. However, in our experience the introduced bias can be neglected even for relatively small numbers of observations. Furthermore, empirical Bayes estimates often result in less variable estimates and, hence, may result in estimates with a lower MSE compared to fully Bayesian estimates. We will investigate this in several simulation studies later on.

Besides providing an alternative estimation concept for structured additive regression, the mixed model representation is also of theoretical value. For example, it allows for deeper insight into the identifiability problems of nonparametric regression models (see Section 5.1). Further benefits will be discussed in Part V.

Mixed model based inference has been considered in additive models with different kinds of univariate responses throughout the last years. We will generalize this approach to further types of covariate effects, such as spatial and interaction effects, and describe how to adapt it to categorical responses and survival times. Fully Bayesian estimation based on MCMC will also briefly be reviewed to enable a comparison between the proposed empirical Bayes approach and its fully Bayesian counterparts.

In addition to theoretical considerations, each part contains applications (the examples discussed in the previous section) and simulation studies giving further insight into the statistical properties of the estimates. Part II also contains a description of the software making all the approaches discussed in this thesis available to researchers and practitioners. Besides the methodological development the implementation in a user-friendly form was a major part of this thesis. The software is available via internet from

http://www.stat.uni-muenchen.de/∼bayesx

This thesis is an extended, modified and reviewed version of the following papers:

- Fahrmeir, Kneib & Lang (2004): Penalized structured additive regression for space-time data: a Bayesian perspective. *Statistica Sinica* **14**, 731–761.
- Kneib & Fahrmeir (2004): A mixed model approach for structured hazard regression. SFB 386 Discussion Paper 400.
- Kneib & Fahrmeir (2005): Structured additive regression for categorical space-time data: A mixed model approach. *Biometrics*, to appear.
- Kneib & Fahrmeir (2005): Supplement to "Structured additive regression for categorical space-time data: A mixed model approach". SFB 386 Discussion Paper 431.
- Kneib (2005): Geoadditive hazard regression for interval censored survival times. SFB 386 Discussion Paper 447.
- Brezger, Kneib & Lang (2005). BayesX: Analyzing Bayesian structured additive regression models. *Journal of Statistical Software*, **14** (11).

Part II

Univariate responses from exponential families

4 Model formulation

4.1 Observation model

4.1.1 Generalized linear models

A common way to build regression models extending the classical linear model for Gaussian responses to more general situations such as binary responses or count data are generalized linear models originally introduced by Nelder & Wedderburn (1972) (for more comprehensive overviews see Fahrmeir & Tutz (2001) or McCullagh & Nelder (1989)). In these models the influence of covariates u on a response variable y is assumed to satisfy the following two assumptions:

Distributional assumption
Conditional on covariates u_i, the responses y_i are independent and the distribution of y_i belongs to a simple exponential family, i. e. its density can be written as

$$f(y_i|\theta_i, \phi, \omega_i) = \exp\left(\frac{y_i\theta_i - b(\theta_i)}{\phi}\omega_i + c(y_i, \phi, \omega_i)\right), \quad i = 1, \ldots, n \tag{4.1}$$

where

θ_i	is the natural parameter of the exponential family (see below),
ϕ	is a scale or dispersion parameter common to all observations,
ω_i	is a weight, and
$b(\cdot)$ and $c(\cdot)$	are functions depending on the specific exponential family.

Structural assumption:
The (conditional) expectation $E(y_i|u_i) = \mu_i$ is linked to the linear predictor

$$\eta_i = u_i'\gamma \tag{4.2}$$

via

$$\mu_i = h(\eta_i) \text{ or } \eta_i = g(\mu_i),$$

where

h	is a smooth, bijective response function,
g	is the inverse of h called the link function and
γ	is a vector of unknown regression coefficients.

Both assumptions are connected by the fact that the mean of y_i is also determined by the distributional assumption and can be shown to be given as

$$\mu_i = b'(\theta_i) = \frac{\partial b(\theta_i)}{\partial \theta}.$$

Therefore, the natural parameter can be expressed as a function of the mean, i. e. $\theta_i = \theta(\mu_i)$. In contrast to the classical linear model, the variance of y_i in general also depends on the linear predictor since

$$\text{var}(y_i|u_i) = \sigma^2(\mu_i) = \frac{\phi v(\mu_i)}{\omega_i}$$

with $v(\mu_i) = b''(\theta_i)$ being the variance function of the underlying exponential family. In Table 4.1 the natural parameter, the expectation, the variance function and the scale parameter are listed for the most commonly used exponential families.

Distribution		$\theta(\mu)$	$b'(\theta)$	$v(\mu)$	ϕ
Normal	$N(\mu, \sigma^2)$	μ	$\mu = \theta$	1	σ^2
Gamma	$Ga(\mu, \nu)$	$-1/\mu$	$\mu = -1/\theta$	μ^2	ν^{-1}
Poisson	$Po(\lambda)$	$\log(\lambda)$	$\lambda = \exp(\theta)$	λ	1
Binomial	$B(n, \pi)$	$\log(\pi/(1-\pi))$	$\pi = \frac{\exp(\theta)}{1+\exp(\theta)}$	$\pi(1-\pi)$	1

Table 4.1: Components of the most commonly used exponential families.

For a given response distribution different response functions are used in practice depending on the specific application. Some examples will be discussed in the following subsections. One particularly important special case is the so called natural (or canonical) link function obtained from

$$\theta_i = \theta(\mu_i) = \eta_i,$$

where the natural parameter is linked directly to the linear predictor. This choice is frequently used as the default link function since it results in simpler estimation equations (although other choices may be more appropriate in some situations as we will see later on).

4.1.1.1 Models for continuous responses

Normal distribution

The classical linear model can be subsumed into the context of generalized linear models by defining $h(\eta) = \eta$, i. e. the response function is simply the identity. For Gaussian distributed responses this also represents the natural link function. The variance function $v(\mu)$ is constant, while the scale parameter equals the variance of the error terms of the linear regression model (see also Table 4.1).

Gamma distribution

If the response values are all nonnegative, the normal distribution in combination with the identity link is often not adequate for an appropriate analysis. Although lognormal models, where the identity link is replaced by the log link, are frequently used in practice, a more natural choice would be a distribution whose support is $\mathbb{R}_+$ by definition. In addition, choosing an appropriate response distribution also allows to account for the fact that usually nonnegative responses follow a skewed and asymmetric distribution. One member of the class of exponential families allowing for both properties is the gamma distribution. Here, the natural response function is given by the negative reciprocal

$$h(\eta) = -\eta^{-1} = \mu.$$

This response function is, however, only rarely used in practice since it does not ensure the nonnegativity of the expectation. Instead the log-link

$$g(\mu) = \log(\mu) = \eta$$

or, equivalently, the exponential response function

$$h(\eta) = \exp(\eta) = \mu$$

are the most common choices when analyzing gamma distributed responses.

4.1.1.2 Models for count data

A regression model for the analysis of count data can be derived under the assumption of Poisson distributed responses. In this case the natural response function is given by the exponential

$$h(\eta) = \exp(\eta) = \mu,$$

and the natural link function is the natural logarithm

$$g(\mu) = \log(\mu) = \eta.$$

Therefore the present model is also referred to as a loglinear model. Note that in contrast to normal and gamma models the scale parameter is fixed at $\phi = 1$ for Poisson data.

4.1.1.3 Models for binary and binomial responses

For binary responses $y_i \in \{0, 1\}$ the expectation is given by the probability $\pi = P(y = 1)$, which requires appropriate response functions to ensure $\pi \in [0, 1]$. Obviously, any cumulative distribution function satisfies this condition and different model formulations are obtained for different choices of the distribution function. In any case, the scale parameter is again fixed at $\phi = 1$.

Logit model

When choosing the natural link function

$$g(\pi) = \log\left(\frac{\pi}{1-\pi}\right) = \eta,$$

the logit model is obtained which corresponds to the logistic distribution function as response function:

$$h(\eta) = \frac{\exp(\eta)}{1 - \exp(\eta)} = \pi.$$

The logistic distribution function is symmetric and has somewhat heavier tails than the standard normal distribution function used in probit models. Due to the intuitive interpretation of the regression coefficients based on odds and odds ratios, the logit model is most commonly used when analyzing binary data, especially in medical applications.

Probit model

In the probit model the logistic distribution function is replaced by the standard normal distribution function. This results in somewhat lighter tails while retaining symmetry. Since for the probit model the evaluation of the likelihood is computationally more demanding and parameter estimates are not interpretable in terms of odds or odds ratios, the logit model is often preferred in practice. An exception are fully Bayesian generalized linear models estimated based on Markov Chain Monte Carlo simulations, where probit models are represented via Gaussian distributed latent variables allowing for simple Gibbs sampling updates (see also the discussion of latent variable approaches in Section 11.5 and Albert & Chib (1993) for a description of fully Bayesian inference in probit models).

Complementary log-log model

An asymmetric binary regression model is obtained with the extreme minimal value distribution function

$$h(\eta) = 1 - \exp(-\exp(\eta)) = \pi.$$

This model is also called the complementary log-log model since it results in the link function

$$g(\pi) = \log(-\log(1 - \pi)) = \eta.$$

Though being less frequently used in the analysis of originally binary data, it is more commonly applied in discrete time survival models since it can be interpreted as a grouped proportional hazards model (see Fahrmeir & Tutz (2001), Ch. 9 and Section 14.4.2).

Binomial responses

To model binomial responses $y_i \sim B(n_i, \pi_i)$, exactly the same models as discussed for binary responses can be used by replacing y_i with $\bar{y}_i = y_i/n_i$ and introducing weights $\omega_i = n_i$, $i = 1, \ldots, n$. In this formulation the expectation is also given by $\pi_i = E(\bar{y}_i)$ and the binomial distribution can easily be subsumed in the exponential family framework.

4.1.2 Structured additive regression

While being flexible in terms of the supported response distributions, generalized linear models obey rather strong assumptions considering the linearity of the influence of covariates and the independence of the observations. In practical regression situations, at least one of the following problems is frequently encountered (compare also the introductory discussions in Section 2):

- For the continuous covariates in the data set, the assumption of a strictly linear effect on the predictor may not be appropriate, i. e. some effects may be of unknown nonlinear form.
- Observations may be spatially correlated.
- Observations may be temporally correlated.

- Heterogeneity among individuals or units may be not sufficiently described by covariates. Hence, unobserved unit- or cluster-specific heterogeneity must be considered appropriately.
- Interactions between covariates may be of complex, nonlinear form.

To overcome these difficulties, we replace the strictly linear predictor in (4.2) by a semiparametric structured additive predictor

$$\eta_i = f_1(\nu_{i1}) + \cdots + f_p(\nu_{ip}) + u_i'\gamma, \tag{4.3}$$

where i is a generic observation index, the ν_j are generic covariates of different type and dimension, and the f_j are (not necessarily smooth) functions of the covariates. These functions comprise nonlinear effects of continuous covariates, time trends and seasonal effects, two-dimensional surfaces, varying coefficient models, i. i. d. random intercepts and slopes, and temporally or spatially correlated effects. Covariates with parametric effects are subsumed in the term $u_i'\gamma$.

At first sight, it may be irritating to use one general notation for nonlinear functions of continuous covariates, i. i. d. random intercepts and slopes, and spatially correlated random effects as in (4.3). However, the unified treatment of the different components in our model has several advantages.

- Since we will adopt a Bayesian perspective, both "fixed effects" and "random effects" are random variables only distinguished by different priors, e. g. diffuse priors for fixed effects and Gaussian priors for i. i. d. random effects (see also the discussion in Hobert & Casella 1996).
- As we will see in Section 4.2, the priors for smooth functions, two-dimensional surfaces, i. i. d., serially and spatially correlated random effects can be cast into one general form.
- The general form of the priors allows for rather general and unified estimation procedures, see Section 5. As a side effect, the implementation and description of these procedures is considerably facilitated.

In order to demonstrate the generality of our approach we point out some special cases of model (4.3) which are well known from the literature.

- **Generalized additive model (GAM) for cross-sectional data:**
  The predictor of a GAM (Hastie & Tibshirani 1990) for observation n, $n = 1, \ldots, N$, is given by
  $$\eta_n = f_1(x_{n1}) + \cdots + f_k(x_{nk}) + u_n'\gamma. \tag{4.4}$$
  In this case, the f_j are smooth functions of continuous covariates x_j which can be modeled by (Bayesian) P-splines, random walks, or Gaussian stochastic process priors, see Section 4.2.2. We obtain a GAM as a special case of (4.3) with $i = n$, $n = 1, \ldots, N$, and $\nu_{ij} = x_{nj}$, $j = 1, \ldots, k$.
- **Generalized additive mixed model (GAMM) for longitudinal data:**
  Consider longitudinal data for individuals $n = 1, \ldots, N$, observed at time points $t \in \{t_1, t_2, \ldots\}$. For notational simplicity we assume the same time points for every

individual, but generalizations to individual-specific time points are obvious. A GAMM extends (4.4) by introducing individual-specific random effects, i. e.

$$\eta_{nt} = f_1(x_{nt1}) + \cdots + f_k(x_{ntk}) + b_{1n}w_{nt1} + \cdots + b_{qn}w_{ntq} + u'_{nt}\gamma, \qquad (4.5)$$

where $\eta_{nt}, x_{nt1}, \ldots, x_{ntk}, w_{nt1}, \ldots, w_{ntq}, u_{nt}$ are predictor and covariate values for individual n at time t and $b_n = (b_{1n}, \ldots, b_{qn})'$ is a vector of q i. i. d. random intercepts (if $w_{ntj} = 1$) or random slopes. While the nonparametric effects f_j are modeled in the same way as for GAMs, the random effects components are assumed to follow i. i. d. Gaussian priors, see Section 4.2.4.

In (4.5), the functions f_j are nonlinear population effects. Individual-specific deviations from these population effects as well as correlations of repeated observations can be modeled through the random effects part of the predictor. As an example, assume that a function $f(t)$ represents the population time trend approximated by a linear combination $f(t) = \sum \beta_j B_j(t)$ of B-spline basis functions $B_j(t)$. Individual-specific deviations can then be expressed as $f_n(t) = \sum b_{jn}B_j(t)$, where the b_{jn} are i. i. d. random effects, and the design variables w_{ntj} are equal to $B_j(t)$. This is in analogy to standard parametric mixed models with, e. g., a linear time trend $\beta_0 + \beta_1 t$ and individual-specific random deviations $b_{0n} + b_{1n}t$ from this trend.

GAMMs can be subsumed into (4.3) by defining $i = (n, t)$, $\nu_{ij} = x_{ntj}$, $j = 1, \ldots, k$, $\nu_{i,k+h} = w_{nth}$, $h = 1, \ldots, q$, and $f_{k+h}(\nu_{i,k+h}) = b_{hn}w_{nth}$. Similarly, GAMMs for clustered data can be written in the general form (4.3).

- **Space-time main effect model - geoadditive model:**
  Suppose we observe longitudinal data with additional geographic information for every observation. A reasonable predictor for such spatio-temporal data (see e. g. Fahrmeir & Lang 2001b) is given by

$$\eta_{nt} = f_1(x_{nt1}) + \cdots + f_k(x_{ntk}) + f_{time}(t) + f_{spat}(s_{nt}) + u'_{nt}\gamma, \qquad (4.6)$$

  where f_{time} is a possibly nonlinear, temporally correlated time trend and f_{spat} is a spatially correlated effect of the location s_{nt} an observation belongs to. Models with a predictor that contains a spatial effect are also called geoadditive models, see Kammann & Wand (2003). The time trend can be modeled in the same way as nonparametric effects of continuous covariates (see Section 4.2.2), and the spatial effect by Markov random fields, stationary Gaussian random fields or two-dimensional P-splines, see Sections 4.2.3 and 4.2.6. Note that observations are marginally correlated after integrating out the temporally or spatially correlated effects f_{time} and f_{spat}. Individual-specific effects can be incorporated as for GAMMs, if appropriate. In the notation of the general model (4.3) we obtain $i = (n, t)$, $\nu_{ij} = x_{ntj}$ for $j = 1, \ldots, k$, $\nu_{i,k+1} = t$ and $\nu_{i,k+2} = s_{nt}$.

- **Varying coefficient model (VCM) - Geographically weighted regression:**
  A VCM as proposed by Hastie & Tibshirani (1993) is given by

$$\eta_n = g_1(x_{n1})z_{n1} + \cdots + g_k(x_{nk})z_{nk},$$

  where the effect modifiers x_{nj} are continuous covariates or time scales and the interacting variables z_{nj} are either continuous or categorical. A VCM can be cast into

the general form (4.3) with $i = n$, $\nu_{ij} = (x_{nj}, z_{nj})$, and by defining the special functions $f_j(\nu_{ij}) = f_j(x_{nj}, z_{nj}) = g_j(x_{nj})z_{nj}$. In structured additive regression models the effect modifiers are not necessarily restricted to be continuous variables as in Hastie & Tibshirani (1993). For example, the geographical location may be used as effect modifier as well, see Fahrmeir, Lang, Wolff & Bender (2003) for an example. VCMs with spatially varying regression coefficients are well known in the geography literature as geographically weighted regression, see e. g. Fotheringham, Brunsdon & Charlton (2002).

- **ANOVA type interaction model:**
  Suppose x_{n1} and x_{n2} are two continuous covariates. Then, the effect of x_{n1} and x_{n2} may be modeled by a predictor of the form

$$\eta_n = f_1(x_{n1}) + f_2(x_{n2}) + f_{1,2}(x_{n1}, x_{n2}) + \ldots,$$

  see e. g. Chen (1993). The functions f_1 and f_2 account for the main effects of the two covariates and $f_{1,2}$ is a two-dimensional interaction surface which can be modeled by two-dimensional P-splines, see Section 4.2.6. The interaction model can be cast into the form (4.3) by defining $i = n$, $\nu_{i1} = x_{n1}$, $\nu_{i2} = x_{n2}$ and $\nu_{i3} = (x_{n1}, x_{n2})$. Similarly, the space-time main effects model (4.6) may be extended to a model incorporating a space-time interaction effect.

4.2 Predictor components and priors

For Bayesian inference, the unknown functions $f_1, \ldots, f_p$ in the structured additive predictor (4.3) or, more exactly, the corresponding vectors of function evaluations and the fixed effects γ are considered as random variables and must be supplemented by appropriate prior assumptions.

Priors for the unknown functions $f_1, \ldots, f_p$ depend on the specific type of the corresponding covariates ν_j and on prior beliefs about the smoothness of f_j. In the following, we express function evaluations $f_j(\nu_{ij})$ as the product of a design vector v_{ij} and a vector of unknown parameters ξ_j, i. e.

$$f_j(\nu_{ij}) = v_{ij}'\xi_j.$$

Therefore, we can rewrite the structured additive predictor (4.3) as

$$\eta_i = v_{i1}'\xi_1 + \ldots + v_{ip}'\xi_p + u_i'\gamma \tag{4.7}$$

or, equivalently, in matrix notation as

$$\eta = V_1\xi_1 + \cdots + V_p\xi_p + U\gamma, \tag{4.8}$$

where the V_j are row-wise stacked matrices composed of the vectors v_{ij}, and U corresponds to the usual design matrix for fixed effects.

A prior for a function f_j is now defined by specifying a suitable design vector v_{ij} and a prior distribution for the vector ξ_j of unknown parameters. In structured additive regression

models as considered in this thesis, the general form of the prior for ξ_j is a multivariate Gaussian distribution with density

$$p(\xi_j|\tau_j^2) \propto \exp\left(-\frac{1}{2\tau_j^2}\xi_j' K_j \xi_j\right), \tag{4.9}$$

where the precision matrix K_j acts as a penalty matrix that shrinks parameters towards zero, or penalizes too abrupt jumps between neighboring parameters. In most cases K_j will be rank deficient, i. e. $k_j = \text{rank}(K_j) < \dim(\xi_j) = d_j$. Hence, the prior for ξ_j is partially improper. The variance parameter τ_j^2 is equivalent to the inverse smoothing parameter in a frequentist approach and controls the trade off between flexibility and smoothness: A small (large) value of τ_j^2 corresponds to an increase (decrease) of the penalty or shrinkage. We will see this at work in several examples discussed throughout the following sections.

Of course, different prior distributions than the multivariate Gaussian prior (4.9) could be considered alternatively. For example, Gamerman (1997) utilizes t-distributions as priors for random effects to obtain more robust estimates and to account for overdispersion. Besag & Kooperberg (1995) employ Laplace priors, where the sum of squares $\Phi(\xi_j) = \xi_j' K_j \xi_j$ in (4.9) is replaced by a sum of absolute values, to obtain improved edge-preserving properties. In the context of image analysis, the functional $\Phi(\xi_j)$ is usually referred to as the potential function and several types of potential functions have been considered in the literature. Hurn, Husby & Rue (2003) give an overview on some potential functions and references to their applications. In this thesis, we will, however, restrict ourselves to prior (4.9) since it provides sufficient generality in most situations and also is a good starting point for more refined analyses. Moreover, Gaussian priors have the advantage that linear transformation of the regression coefficients are also Gaussian distributed. This will be utilized in the reparametrization of structured additive regression models in Section 5.1.

Most of the improper priors used in structured additive regression can be interpreted as intrinsic Gaussian Markov random fields (IGMRFs) which are special cases of Gaussian Markov random fields (GMRFs) with improper distribution (see Rue & Held (2005), especially Ch. 3). Formally, GMRFs and IGMRFs are defined based on labeled graphs and conditional independence properties of the random vector ξ_j but we will not pursue the idea of IGMRFs too much in the following sections. Instead, we briefly discuss one general result which will be useful in Section 5 when estimating structured additive regression models based on mixed model methodology.

This general result states that a vector of regression coefficients ξ_j following an IGMRF prior can always be expressed as the sum of two parts, with the first part lying in the null space spanned by the precision matrix K_j and the second part being orthogonal to this null space (Rue & Held 2005, p. 91). It can then be shown that prior (4.9) is invariant to the addition of any vector belonging to the null space of K_j. This means that the basis of the null space describes the part of a function f_j that is not penalized by (4.9). Furthermore, prior (4.9) can be shown to be proportional to the (proper) distribution of the part of ξ_j belonging to the orthogonal deviation from the null space. In Section 5.1 a special decomposition of ξ_j will be established more formally and further results on the distribution of both parts of the decomposition will be provided. Consequently, we will discuss the dimensions and bases of the null spaces of K_j for different model terms and covariates in the following sections.

4.2.1 Fixed effects

For the parameter vector γ of fixed effects we routinely assume diffuse priors $p(\gamma) \propto const.$ A possible alternative would be to work with a multivariate Gaussian distribution $\gamma \sim N(\gamma_0, \Sigma_{\gamma_0})$. However, since in most cases a noninformative prior is desired, we consider it sufficient to work with diffuse priors. Furthermore, assuming flat priors emphasizes the close connection of our empirical Bayes approach to (penalized) maximum likelihood estimation.

4.2.2 Continuous covariates and time scales

Several alternatives have been recently proposed for modeling effects of continuous covariates or time trends. Most of these approaches can be assigned to one of two model classes using different strategies to ensure smooth and parsimonious estimates. The first class comprises approaches based on adaptive knot selection for splines, see Friedman (1991) or Stone, Hansen, Kooperberg & Truong (1997) for frequentist versions and Denison, Mallick & Smith (1998), Biller (2000), DiMatteo, Genovese & Kass (2001), Biller & Fahrmeir (2001), and Hansen & Kooperberg (2002) for Bayesian variants. In contrast, the second class of models is based on smoothness priors or penalization of the regression coefficients. The following sections will be restricted to penalization approaches, since only these allow for mixed model based inference after an appropriate reparametrization, see Section 5.

4.2.2.1 P-Splines

One increasingly popular idea to estimate smooth effects of continuous covariates are penalized splines or P-splines, introduced by Eilers & Marx (1996). The fundamental assumption of this approach is that the unknown smooth function f_j of a covariate x_j can be approximated by a polynomial spline. For notational simplicity we will drop the index j in the following discussion.

A polynomial spline function (see Dierckx (1993) or de Boor (1978) for mathematically rigorous treatments) is defined based on a set of $M+1$ (not necessarily equally spaced) knots $x_{min} = \kappa_0 < \kappa_1 < \cdots < \kappa_{M-1} < \kappa_M = x_{max}$ within the domain of covariate x. To be more specific, a function $g : [a, b] \to \mathbb{R}$ is called a polynomial spline of degree $l, l \in \mathbb{N}_0$ based on knots $\kappa_0, \ldots, \kappa_M$, if it satisfies the following conditions:

1. $g(x)$ is $(l-1)$ times continuous differentiable and

2. $g(x)$ is a polynomial of degree l for $x \in [\kappa_m, \kappa_{m+1})$, $m = 0, \ldots, M-1$.

The space of polynomial splines can be shown to be a $(M+l)$-dimensional subspace of the space of $(l-1)$ times continuous differentiable functions. Therefore, assuming that $f(x)$ can be approximated by a polynomial spline leads to a representation in terms of a linear combination of $d = M + l$ basis functions B_m, i. e.

$$f(x) = \sum_{m=1}^{d} \xi_m B_m(x),$$

where $\xi = (\xi_1, \ldots, \xi_d)'$ corresponds to the vector of unknown regression coefficients.

One possible set of basis functions is given by the truncated polynomials

$$B_1(x) = 1, \quad B_2(x) = x, \quad \ldots, \quad B_{l+1}(x) = x^l,$$
$$B_{l+2}(x) = (x - \kappa_1)_+^l, \quad \ldots, \quad B_{M+l}(x) = (x - \kappa_{M-1})_+^l,$$

where

$$(x - \kappa_m)_+^l = \begin{cases} (x - \kappa_m)^l & \text{if } x \geq \kappa_m \\ 0 & \text{if } x < \kappa_m. \end{cases}$$

Although frequently used to introduce penalized splines (see e. g. Wand (2003) or Ruppert, Wand & Carroll (2003)), truncated polynomials exhibit a number of drawbacks: Due to their polynomial nature, the evaluation of truncated polynomials may cause numerical problems for large values of the covariate x. More important is the question of how to incorporate a smoothness prior or penalization into the model. With truncated polynomials this is usually achieved by assuming that the coefficients associated with the truncated basis functions $B_{l+2}(x), \ldots, B_{M+l}(x)$ are i. i. d. Gaussian random effects. Hence, increasing the amount of penalization leads (in the limit) to a polynomial of degree l defined by the degree of the basis functions. Applying a different set of basis functions allows to separately control the degree of the limiting polynomial and the degree of the basis functions, thus leading to a richer model class.

This different class of basis functions is called the B(asic)-spline basis, where basis functions of degree l are defined recursively via

$$B_m^l(x) = \frac{x - \kappa_m}{\kappa_{m+l} - \kappa_m} B_m^{l-1}(x) + \frac{\kappa_{m+l+1} - x}{\kappa_{m+l+1} - \kappa_{m+1}} B_{m+1}^{l-1}(x)$$

with

$$B_m^0(x) = \mathbb{1}_{[\kappa_m, \kappa_{m+1})}(x) = \begin{cases} 1 & \kappa_m \leq x < \kappa_{m+1} \\ 0 & \text{else.} \end{cases}$$

Note that additional knots $\kappa_{-l} < \ldots < \kappa_{-1} < a$ and $b < \kappa_{M+1} < \kappa_{M+2} < \ldots < \kappa_{M+l}$ are needed for the recursive construction of the full B-spline basis. In case of equidistant knots, the additional knots are easily defined using the same spacing as for the inner knots. With nonequidistant knots an additional rule has to be defined, e. g. to use the distance between the two leftmost and rightmost inner knots.

Figure 4.1 shows a small set of B-spline basis functions for different degrees l and different knot choices. Obviously, B-splines form a local basis since the basis functions are only positive within an area spanned by $l + 2$ knots. This property is essential for the construction of the smoothness penalty for P-splines. Furthermore, the basis functions are bounded yielding better numerical properties compared to the truncated power series basis. In case of equidistant knots (shown in the left panel) all basis functions are of the same functional form and only shifted along the x-axis. In the contrary case of nonequidistant knots (shown in the right panel), the functional form of the basis functions extremely changes between areas with dense knots and areas where knots are only placed at few positions. As required by the definition, the smoothness of the basis functions increases with increasing degree.

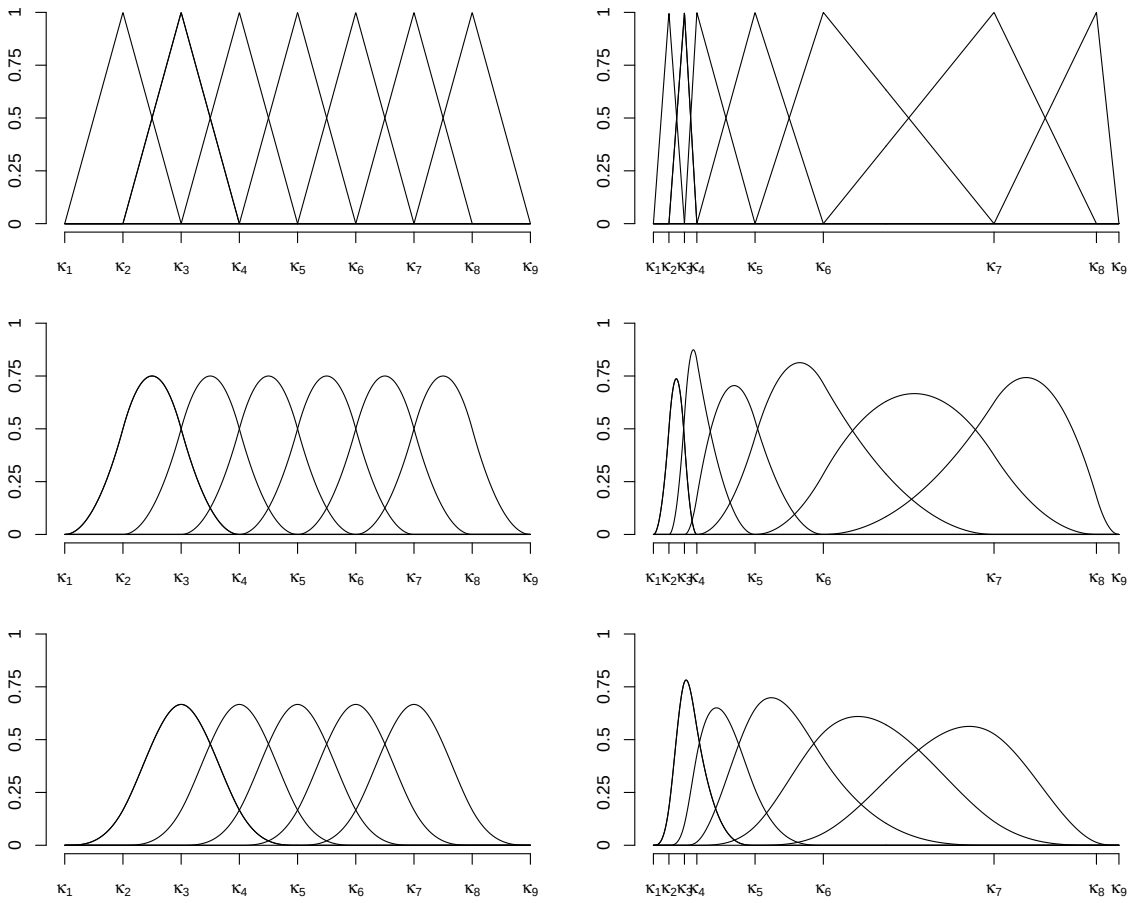


Figure 4.1: B-spline basis functions of degrees $l = 1$ (upper row), $l = 2$ (middle row) and, $l = 3$ (lower row) for equally spaced knots (left panel) and unequally spaced knots (right panel).

Regardless of the specific basis used in an implementation, the d-dimensional design vector v_i consists of the basis functions evaluated at the observations x_i, i. e. $v_i = (B_1(x_i), \ldots, B_d(x_i))'$. The crucial choice is the number of knots: For a small number of knots, the resulting spline may not be flexible enough to capture the variability of the data. For a large number of knots, however, estimated curves tend to overfit the data and, as a result, too rough functions are obtained. Figure 4.2 illustrates the influence of the number of knots when estimating a sinusoidal function. Clearly, 5 or 10 knots lead to a smooth and more or less appropriate fit, while more knots tend to overfit the data.

Whereas adaptive knot selection approaches use specific criteria and strategies to choose an optimal set of basis functions, Eilers & Marx (1996) suggest to use a moderately large number of equally spaced knots (usually between 20 and 40) to ensure enough flexibility, and to define a roughness penalty based on differences of adjacent B-Spline coefficients to guarantee sufficient smoothness of the fitted curves. In their formulation this leads to penalized likelihood estimation with penalty terms

$$pen(\lambda) = \frac{1}{2}\lambda \sum_{m=k+1}^{d} (\Delta^k \xi_m)^2, \tag{4.10}$$

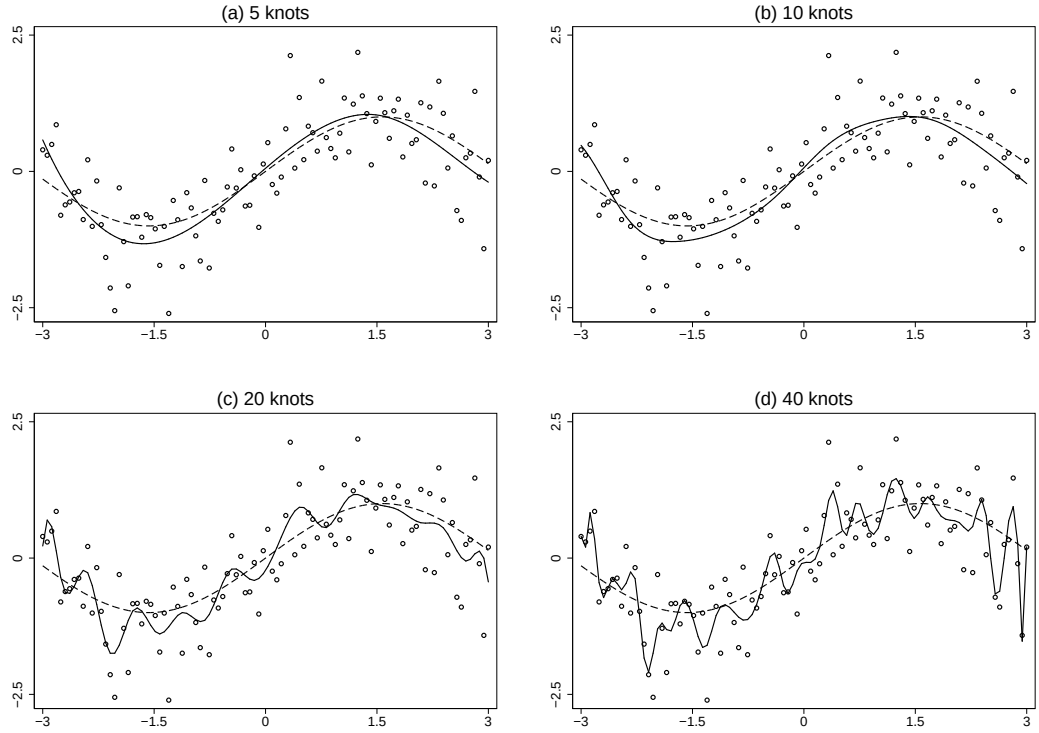


Figure 4.2: Influence of the number of knots. The true function is given by the dashed line, the estimated curve by the solid line.

where λ is a smoothing parameter and the factor $1/2$ is introduced only due to notational convenience. The difference operator Δ^k of order k is defined recursively by

$$\begin{aligned} \Delta^1 \xi_m &= \xi_m - \xi_{m-1}, && (4.11) \\ \Delta^2 \xi_m &= \Delta^1(\Delta^1 \xi_m) = \xi_m - 2\xi_{m-1} + \xi_{m-2}, && (4.12) \\ \Delta^k \xi_m &= \Delta^1(\Delta^{k-1} \xi_m). \end{aligned}$$

While first order differences penalize abrupt jumps between successive parameters, second order differences penalize deviations from the linear trend $2\xi_{m-1} - \xi_{m-2}$. As a consequence, large values of the smoothing parameter λ lead to estimates close to a horizontal line (first order differences) or a linear effect (second order differences). In general, the limiting polynomial when the smoothing parameter goes to infinity is of order $k-1$ and is therefore independent from the degree of the spline basis. This is in contrast to the results for the truncated power series basis discussed before.

In a Bayesian framework, penalized splines are introduced by replacing the difference penalties with their stochastic analogues, i. e., random walks of order k are used as priors for the regression coefficients (Lang & Brezger (2004) and Brezger & Lang (2005)). First and second order random walks for equidistant knots are given by

$$\xi_m = \xi_{m-1} + u_m, \quad m = 2, \ldots, d \tag{4.13}$$

and

$$\xi_m = 2\xi_{m-1} - \xi_{m-2} + u_m, \quad m = 3, \ldots, d \tag{4.14}$$

with Gaussian errors $u_m \sim N(0, \tau^2)$. Diffuse priors $p(\xi_1) \propto const$, and $p(\xi_1)$, $p(\xi_2) \propto const$, are chosen for the initial values. Alternatively, assumptions (4.13) and (4.14) can be replaced by

$$\xi_m|\xi_{m-1} \sim N(\xi_{m-1}, \tau^2)$$

and

$$\xi_m|\xi_{m-1}, \xi_{m-2} \sim N(2\xi_{m-1} - \xi_{m-2}, \tau^2).$$

Figure 4.3 presents the illustration of these assumptions. While a first order random walk induces a constant trend for the conditional expectation of β_m given β_{m-1}, a second order random walk results in a linear trend depending on the two previous values β_{m-1} and β_{m-2}. Figure 4.3 also reveals the impact of the variance parameter τ^2 more clearly. If τ^2 is large, ample deviations from the trend assumed by the random walk prior are possible. In contrast, if τ^2 is small, deviations from this trend will also be small, leading to an almost deterministic behavior of the regression coefficients.

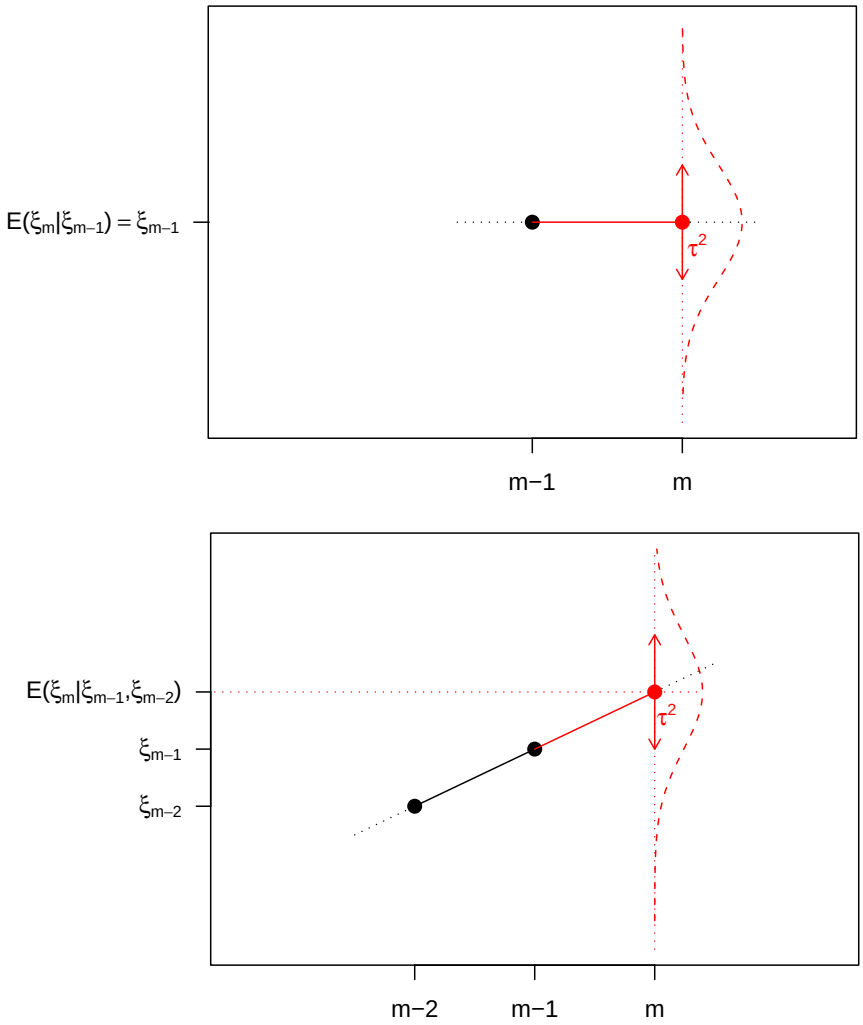


Figure 4.3: Trends induced by first and second order random walk priors.

The joint distribution of the regression parameters ξ is easily computed as a product of conditional densities defined by the random walk priors and can be brought into the

general form (4.9) of a multivariate but improper Gaussian distribution. The precision matrix is of the form $K = D_k' D_k$ where D_k is a difference matrix of order k, i. e.

$$
\begin{aligned}
D_1 &= \begin{pmatrix} -1 & 1 & & & \\ & -1 & 1 & & \\ & & \ddots & \ddots & \\ & & & -1 & 1 \end{pmatrix} && (d-1 \times d), \\
D_2 &= \begin{pmatrix} 1 & -2 & 1 & & \\ & 1 & -2 & 1 & \\ & & \ddots & \ddots & \ddots \\ & & & 1 & -2 \quad 1 \end{pmatrix} && (d-2 \times d), \\
D_k &= D_1 D_{k-1} && (d-k \times d).
\end{aligned}
$$

Concerning first order random walks this yields the precision matrix

$$
K = \begin{pmatrix} 1 & -1 & & & \\ -1 & 2 & -1 & & \\ & \ddots & \ddots & \ddots & \\ & & -1 & 2 & -1 \\ & & & -1 & 1 \end{pmatrix} \tag{4.15}
$$

and referring to second order random walks we obtain

$$
K = \begin{pmatrix} 1 & -2 & 1 & & & & \\ -2 & 5 & -4 & 1 & & & \\ 1 & -4 & 6 & -4 & 1 & & \\ & \ddots & \ddots & \ddots & \ddots & \ddots & \\ & & 1 & -4 & 6 & -4 & 1 \\ & & & 1 & -4 & 5 & -2 \\ & & & & 1 & -2 & 1 \end{pmatrix}.
$$

Since the rows of matrix (4.15) sum to zero, the rank deficiency of this matrix equals one. In general, the penalty matrix constructed from a random walk of order k has rank $d-k$.

Instead of conditioning only on preceding values of ξ_m we can also compute the conditional distribution of ξ_m given all other entries of ξ. This distribution then depends on both preceding and succeeding values of ξ_m. From the joint distribution (4.9) these conditional distributions can be easily derived using standard calculations for normal distributions: The weight associated with parameter ξ_r in the conditional distribution of ξ_m is given by $-k_{mr}/k_{mm}$, where k_{mr} and k_{mm} are the corresponding elements of K. The variance of the distribution is given by τ^2/k_{mm} (see for example Rue & Held 2005, p. 22).

Obviously, the weights for a first order random walk equal zero for all but the two nearest neighbors. More specifically we obtain

$$
\xi_m|. \sim \begin{cases} N(\xi_2, \tau^2) & m = 1, \\ N\left(\frac{1}{2}(\xi_{m-1} + \xi_{m+1}), \frac{\tau^2}{2}\right) & m = 2, \ldots, d-1\,, \\ N(\xi_{d-1}, \tau^2) & m = d, \end{cases} \tag{4.16}
$$

i. e. the conditional expectations are given by the local mean of the two next neighbors for all but the boundary coefficients. From a regression perspective, this can also be interpreted as a local linear fit through the two next neighbors as illustrated in Figure 4.4. Note that we could also start directly by assuming a local linear conditional expectation for ξ_m, which would result in the same prior distribution as discussed above. This interpretation will be used in Section 4.2.6 where P-splines are extended to two dimensions.

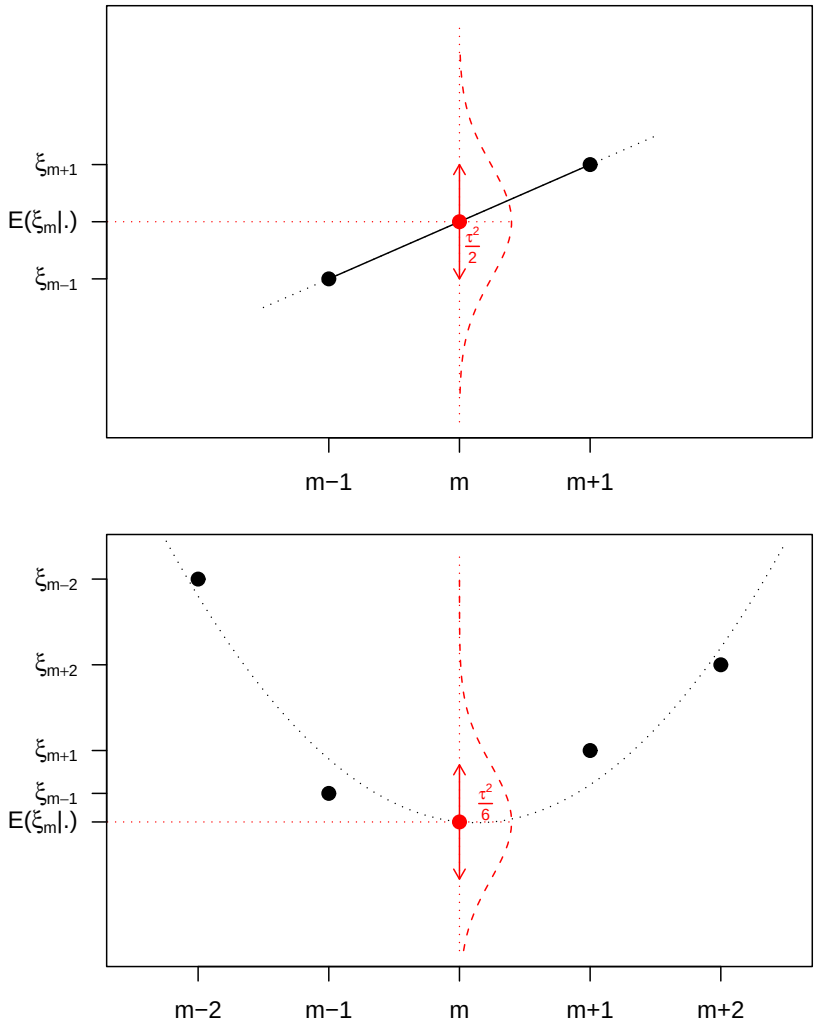


Figure 4.4: Local regression models induced by first and second order random walk priors.

For a second order random walk, similar calculations yield the conditional distributions

$$\xi_m|. \sim \begin{cases} N(2\xi_2 - \xi_3, \tau^2) & m = 1, \\ N\left(\frac{2}{5}\xi_1 + \frac{4}{5}\xi_3 - \frac{1}{5}\xi_4, \frac{\tau^2}{5}\right) & m = 2, \\ N\left(-\frac{1}{6}\xi_{m-2} + \frac{4}{6}\xi_{m-1} + \frac{4}{6}\xi_{m+1} - \frac{1}{6}\xi_{m+2}, \frac{\tau^2}{6}\right) & m = 3, \ldots, d-2, \\ N\left(-\frac{1}{5}\xi_{d-3} + \frac{4}{5}\xi_{d-2} + \frac{2}{5}\xi_d, \frac{\tau^2}{5}\right) & m = d-1, \\ N(-\xi_{d-2} + 2\xi_{d-1}, \tau^2) & m = d. \end{cases} \tag{4.17}$$

Again, these distributions (at least for the coefficients not affected by the boundaries) can be interpreted in a regression setting. Consider a regression of ξ_m on the four nearest

neighbors ξ_{m-2}, ξ_{m-1}, ξ_{m+1} and ξ_{m+2} where we assume a local quadratic influence. This corresponds to a linear regression model with design matrix

$$X = \begin{pmatrix} 1 & m-2 & (m-2)^2 \\ 1 & m-1 & (m-1)^2 \\ 1 & m+1 & (m+1)^2 \\ 1 & m+2 & (m+2)^2 \end{pmatrix}$$

and response vector $y = (\xi_{m-2}, \xi_{m-1}, \xi_{m+1}, \xi_{m+2})'$. If we predict ξ_m from this regression model, we obtain exactly the same coefficients as in (4.17). Figure 4.4 illustrates the interpretation of (4.17) as a local quadratic fit.

Let us now take a closer look at the null space of the precision matrices of random walks. From (4.11) it is obvious that the penalization induced by first order differences remains unchanged when adding a constant to the parameter vector ξ. Equivalently, the prior distribution defined by a first order random walk remains unchanged by adding a constant. Therefore, the one-dimensional null space of precision matrix (4.15) is spanned by a d-dimensional vector of ones. Similarly, taking a closer look at (4.12) reveals that adding a linear trend does neither affect a second order difference penalty nor a second order random walk. Thorough considerations show that, in general, the null space of a P-spline precision matrix is spanned by a polynomial of degree $k-1$ defined by the knots of the P-spline. More specifically, the column vectors of the matrix

$$\begin{pmatrix} 1 & \kappa_1 & \dots & \kappa_1^{k-1} \\ \vdots & \vdots & & \vdots \\ 1 & \kappa_{d_j} & \dots & \kappa_{d_j}^{k-1} \end{pmatrix} \tag{4.18}$$

form a basis of this null space. Equivalently we may use any set of equidistant values instead of the knots. For instance the column vectors of the matrix

$$\begin{pmatrix} 1 & 1 & \dots & 1^{k-1} \\ 1 & 2 & \dots & 2^{k-1} \\ \vdots & \vdots & & \vdots \\ 1 & d & \dots & d^{k-1} \end{pmatrix}$$

span the same null space as (4.18).

Although we restricted the description of P-splines to the equidistant knot setting, they can also be used with nonequidistant knots by defining weighted random walks or weighted difference penalties. However, due to their construction, P-splines should be rather insensitive to the position and number of the knots as long as a sufficiently large number of knots is used. Probably nonequidistant knots may lead to an improved fit if the distribution of covariate x over its domain differs significantly from a uniform distribution. Then it could be useful to define knots on the basis quantiles of x instead of equidistant values. Weighted random walks will be described in the following section.

One reason why P-splines nowadays are one of the most popular smoothing techniques is their low-rank property in the sense of Hastie (1996): P-splines allow to describe a broad class of flexible functions using only a moderate number of parameters. In contrast, smoothing splines need a number of parameters which approximately equals the number

of observations making direct simultaneous estimation of several smoothing splines numerically untractable even for a small number of functions $f_j(x_j)$. Iterative techniques like the backfitting algorithm have to be employed in such a setting. Based on P-splines, a large number of nonparametric effects can be estimated simultaneously. Moreover, routinely used regression diagnostics like the elements of the hat matrix which are not available from iterative procedures can easily be computed (see also Marx & Eilers 1998).

4.2.2.2 Random walks

Instead of assuming a random walk prior for the parameters of a spline representing $f(x)$, a random walk can be imposed on the function evaluations $f(x)$. Such random walk models are frequently used in the analysis of time series (subsumed in the general class of state space models) but can also be applied in additive regression models (see for example Fahrmeir & Lang 2001a). Since the observed values of x are usually not equidistant, appropriate modifications of the random walk priors (4.13) and (4.14) are needed.

Suppose that

$$x_{(1)} < \ldots < x_{(m)} < \ldots < x_{(d)}$$

are the ordered distinct values that are observed for covariate x and define $\xi_m = f(x_{(m)})$. Then $f(x)$ can be written as $f(x) = v'\xi$, where v is a 0/1 incidence vector taking the value one if $x = x_{(m)}$ and zero otherwise, and $\xi = (\xi_1, \ldots, \xi_d)'$ is a vector of regression coefficients. A proper definition of a random walk prior for ξ has to take into account the nonequal distances $\delta_m = x_{(m)} - x_{(m-1)}$ between two adjacent values of x.

First order random walks for nonequidistant observations can be defined via

$$\xi_m = \xi_{m-1} + u_m, \quad m = 2, \ldots, d, \tag{4.19}$$

where the error terms now are Gaussian distributed with variances depending on δ_m:

$$u_m \sim N(0, \delta_m \tau^2). \tag{4.20}$$

A random walk prior with variance (4.20) is motivated from the fact that it can be interpreted as a discretized version of a time-continuous Wiener process. The increments of the Wiener process are stationary Gaussian distributed with mean zero and variance proportional to the length of the interval the increment is defined upon. This property is inherited by the random walk prior which can therefore be interpreted as a realization of a continuous-time process observed only at discrete locations $x_{(m)}$ (see Rue & Held 2005, p. 97).

If we consider the conditional distribution of ξ_m given all other parameters, we obtain a generalized version of (4.16) as

$$\xi_m|. \sim \begin{cases} N\left(\xi_2, \delta_2 \tau^2\right) & m = 1, \\ N\left(\frac{\delta_m}{\delta_m + \delta_{m+1}} \xi_{m-1} + \frac{\delta_{m+1}}{\delta_m + \delta_{m+1}} \xi_{m+1}, \delta_m \frac{\tau^2}{2}\right) & m = 2, \ldots, d-1, \\ N\left(\xi_{d-1}, \delta_d \tau^2\right) & m = d. \end{cases}$$

From these conditional distributions it is easy to verify that the joint distribution of all parameters has precision matrix

$$K = \begin{pmatrix} \delta_2^{-1} & -\delta_2^{-1} & & & & \\ -\delta_2^{-1} & \delta_2^{-1} + \delta_3^{-1} & -\delta_3^{-1} & & & \\ & \ddots & \ddots & \ddots & & \\ & & -\delta_{m-1}^{-1} & \delta_{m-1}^{-1} + \delta_m^{-1} & -\delta_m^{-1} \\ & & & -\delta_m^{-1} & \delta_m^{-1} \end{pmatrix} \tag{4.21}$$

which reduces to the simple form (4.15) for equidistant locations. Note that K can be written as $K = D'WD$ where D is a first order distance matrix for equidistant locations and $W = \text{diag}(\delta_2^{-1}, \ldots, \delta_m^{-1})$ is a diagonal matrix of the weights. This representation of K reveals that again K has rank $d - 1$ and its null space is spanned by a d-dimensional vector of ones.

For second order random walks not only the variance of the error distribution has to be altered, but also the definition of the random walk itself changes. One possible specification is via

$$\xi_m = \left(1 + \frac{\delta_m}{\delta_{m-1}}\right) \xi_{m-1} - \frac{\delta_m}{\delta_{m-1}} \xi_{m-2} + u_m \tag{4.22}$$

with $u_m \sim N(0, w_m \tau^2)$. In accordance with the definition of random walk priors for equidistant locations, (4.22) also implies a linear extrapolation of the two preceding values ξ_{m-1} and ξ_{m-2} for the prediction of ξ_m. Furthermore it can be alternatively derived by replacing the second order differences

$$\Delta^2(\xi_m) = \xi_m - 2\xi_{m-1} + \xi_{m-2} = (\xi_m - \xi_{m-1}) - (\xi_{m-1} - \xi_{m-2})$$

with the weighted version

$$\Delta^2(\xi_m) = \frac{\xi_m - \xi_{m-1}}{\delta_m} - \frac{\xi_{m-1} - \xi_{m-2}}{\delta_{m-1}}.$$

In contrast to the weighted first order random walk, (4.22) can no longer be interpreted as a discretized version of an underlying time-continuous process. Rue & Held (2005, Sec. 3.5) discuss a class of random walk models that preserves this property but is mathematically more demanding than the model presented here.

The weights w_m defining the variances of the random walks can be chosen in different ways. The simplest approach is to work with $w_m = \delta_m$ similar as with first order random walks. A better alternative may be the usage of

$$w_m = \delta_m \left(1 + \frac{\delta_m}{\delta_{m-1}}\right)$$

since it additionally takes into account the distance δ_{m-1} (see Fahrmeir & Lang (2001b) for a derivation of these weights). In either case, the precision matrix of the joint distribution of ξ can again be written as $K = D'WD$, where D is a generalized second order difference matrix defined by (4.22), i. e.

$$D = \begin{pmatrix} \frac{\delta_3}{\delta_2} & -(1 + \frac{\delta_3}{\delta_2}) & 1 & & & \\ & \frac{\delta_4}{\delta_3} & -(1 + \frac{\delta_4}{\delta_3}) & 1 & & \\ & & \ddots & \ddots & \ddots & \\ & & & \frac{\delta_d}{\delta_{d-1}} & -(1 + \frac{\delta_d}{\delta_{d-1}}) & 1 \end{pmatrix}$$

and $W = \text{diag}(w_3, \ldots, w_m)$ contains the variance weights.

In analogy to the discussion for P-splines with second order random walk penalty, the precision matrix for nonequidistant second order random walks also has rank $d-2$ and a basis of the null space is formed by the column vectors of

$$\begin{pmatrix} 1 & x_{(1)} \\ 1 & x_{(2)} \\ \vdots & \vdots \\ 1 & x_{(d)} \end{pmatrix}.$$

Taking a closer look on the discussed random walk priors and P-splines of degree $l = 0$ reveals that both approaches are in fact equivalent if the knots of the P-spline are identified with the ordered observations $x_{(m)}$. Hence, random walk priors can essentially be regarded as a special case of zero degree P-splines but usually require a much larger number of nonequidistant knots.

4.2.2.3 Univariate Gaussian random fields

Inspired by geostatistical models, where the correlation between two locations is modeled explicitly by an intrinsic correlation function, a third way to estimate smooth effects of continuous covariates can be defined. The basic idea is to assume a zero mean stationary Gaussian stochastic process for $\{\xi_x = f(x), x \in \mathbb{R}\}$ and to describe this process via its variance and its correlation function. This will be discussed in greater detail for spatial effects $\{\xi_s, s \in \mathbb{R}^2\}$ in Section 4.2.3.2 but can also be used for univariate smoothing if the Euclidean distance between two points in $\mathbb{R}^2$ is replaced by an appropriate distance measure in $\mathbb{R}$, e. g. the absolute value $|x - x'|$. The distributional assumption of a stochastic process for ξ can then be interpreted as a smoothness prior which enforces the function evaluations $f(x)$ and $f(x')$ for two close points x and x' to be highly correlated. For more details on Gaussian random field priors see Section 4.2.3.2.

4.2.2.4 Seasonal priors

When modeling the effect of a time scale, it may be useful, to split the effect into a trend and a seasonal component

$$f_{time}(t) = f_{trend}(t) + f_{season}(t). \tag{4.23}$$

The trend function will then be modeled by one of the aforementioned approaches, while the seasonal component may be assumed to follow a more general autoregressive prior (compare Fahrmeir & Lang 2001a). A flexible seasonal component $f_{season}(t) = \xi_t$ with period *per* is defined by

$$\xi_t = -\sum_{j=1}^{per-1} \xi_{t-j} + u_t \tag{4.24}$$

with error terms $u_t \sim N(0, \tau^2_{season})$ and diffuse priors for initial values. If $\tau^2_{season} = 0$, the random part in definition (4.24) vanishes and we obtain a fixed seasonal component that

does not vary over time and sums to zero. In general, (4.24) allows for a time-varying seasonal effect, where τ^2_{season} controls the trade-off between a fixed seasonal pattern and flexibility.

Seasonal components can be cast into the general form (4.9) by computing the joint distribution of $\xi_{season} = (\xi_1, \ldots, \xi_T)'$ induced by (4.24), where T denotes the largest value of the time scale. It turns out that the precision matrix of the joint distribution can be written as $K_{season} = D'D$, where D is given by

$$D = \begin{pmatrix} 1 & \ldots & 1 & & & \\ & 1 & \ldots & 1 & & \\ & & \ddots & \ddots & \ddots & \\ & & & \underbrace{1 \quad \ldots \quad 1}_{per} \end{pmatrix} \qquad (T - per + 1) \times T.$$

For a seasonal component of period $per = 4$ for example the precision matrix is given by

$$K_{season} = \begin{pmatrix} 1 & 1 & 1 & 1 & & & & & & \\ 1 & 2 & 2 & 2 & 1 & & & & & \\ 1 & 2 & 3 & 3 & 2 & 1 & & & & \\ 1 & 2 & 3 & 4 & 3 & 2 & 1 & & & \\ & 1 & 2 & 3 & 4 & 3 & 2 & 1 & & \\ & & \ddots & \ddots & \ddots & \ddots & \ddots & \ddots & \ddots & \\ & & & 1 & 2 & 3 & 4 & 3 & 2 & 1 \\ & & & & 1 & 2 & 3 & 3 & 2 & 1 \\ & & & & & 1 & 2 & 2 & 2 & 1 \\ & & & & & & 1 & 1 & 1 & 1 \end{pmatrix}.$$

By construction, this matrix has a $(per-1)$-dimensional null space spanned by the column vectors of

$$\begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \\ -1 & -1 & -1 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \\ -1 & -1 & -1 \\ & \vdots & \end{pmatrix}$$

if $per = 4$. This null space obviously contains all time-constant seasonal patterns. The design vector v_{season} is a T-dimensional 0/1-incidence vector, where the t-th entry equals one if the corresponding observation is measured at time t and zero otherwise.

4.2.3 Spatial covariates

Spatial effects are usually introduced into regression models to account for spatial correlation which is often induced by unobserved, spatially varying covariates. Similarly as

for temporal effects, it is often useful to split a spatial effect into a spatially correlated (structured) part f_{str} and a spatially uncorrelated (unstructured) part f_{unstr}

$$f_{spat}(s) = f_{str}(s) + f_{unstr}(s).$$

A rationale is that a spatial effect is usually a surrogate of many unobserved influential factors, some of them may obey a strong spatial structure and others may be present only locally. By estimating a structured and an unstructured spatial component we aim at distinguishing between the two kinds of influential factors (see Besag York & Mollié 1991). The uncorrelated part f_{unstr} may be estimated based on region-specific i. i. d. Gaussian random effects, see Section 4.2.4. For the specification of the smooth spatial effect f_{str} we can distinguish two different situations: Spatial locations are available exactly in terms of longitude and latitude or observations are clustered in connected geographical regions. Both types of information led to the development of models specifically designed for the estimation of spatial effects in the respective situation. However, the distinction between spatial data given as geographical regions and spatial data given as coordinates is not that clear-cut as we will see later on.

4.2.3.1 Markov random fields

Suppose first that the spatial index $s \in \{1, \ldots, S\}$ represents a location or site in connected geographical regions. For simplicity, we assume that the regions are labeled consecutively. A common way to introduce a spatially correlated effect is to assume that neighboring sites are more alike than two arbitrary sites. Thus, for a valid prior definition a set of neighbors must be defined for each site s. For geographical data two sites s and s' are usually defined to be neighbors if they share a common boundary as illustrated in Figure 4.5 for regular and irregular shaped regions. Neighbors of the black region are indicated in grey.

The simplest (yet most often used) spatial smoothness prior for function evaluations $f_{spat}(s) = \xi_s$ in the present situation is

$$\xi_s | \xi_{s'}, s' \neq s, \tau^2 \sim N\left(\frac{1}{N_s}\sum_{s' \in \partial_s} \xi_{s'}, \frac{\tau^2}{N_s}\right), \tag{4.25}$$

where $s' \in \partial_s$ denotes that site s' is a neighbor of site s and $N_s = |\partial_s|$ is the number of adjacent sites. Thus, the conditional mean of ξ_s is an unweighted average of the function evaluations of neighboring sites. The prior can be considered as a direct generalization of a univariate first order random walk (as discussed in the previous section) to two dimensions and is called an intrinsic Markov random field (IGMRF) or simply a Markov random field (MRF).

Similarly as for the knots of a penalized spline, definition (4.25) seems to be most appropriate if all regions have the same distance in a sense. This assumption is fulfilled if the geographical regions correspond to squares as in the left part of Figure 4.5, but in most applications the geographical regions will be given by irregular shaped regions as in the right part of Figure 4.5. Therefore, more general priors based on weighted rather than

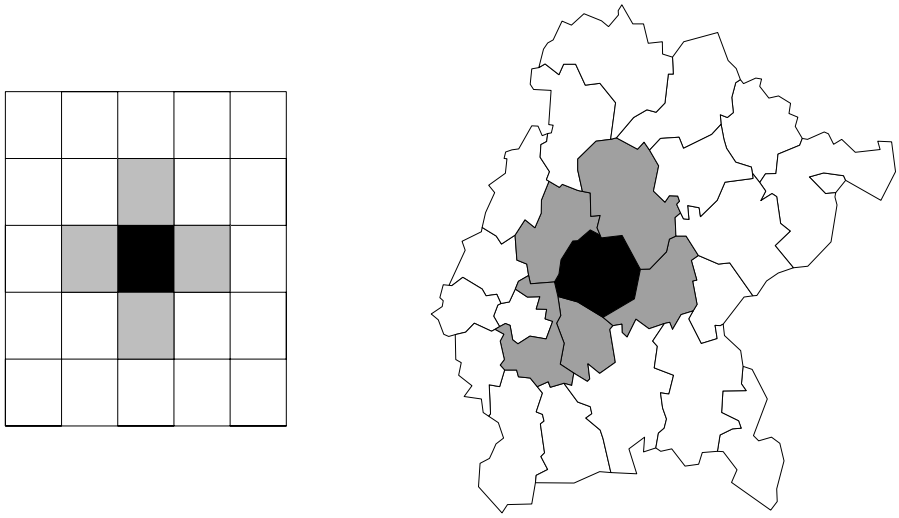


Figure 4.5: Neighborhoods defined by common boundaries for regularly and irregularly shaped regions. Neighbors of the region colored in black are indicated in grey.

unweighted averages as in (4.25) are desirable. In terms of weights $w_{ss'}$, a general spatial prior can be defined as

$$\xi_s|\xi_{s'},\ s' \neq s, \tau^2 \sim N\left(\sum_{s' \in \partial_s} \frac{w_{ss'}}{w_{s+}}\xi_{s'}, \frac{\tau^2}{w_{s+}}\right), \tag{4.26}$$

where $w_{s+} = \sum_{s' \in \partial_s} w_{ss'}$.

Convenient definitions of weights include:

- Equal weights $w_{ss'} = 1$. Here the weighted MRF (4.26) reduces to (4.25).
- Weights inverse proportional to the distance of centroids: $w_{ss'} = c \cdot \exp(-d(s, s'))$, where $d(s, s')$ denotes the Euclidean distance between the centroids of regions s and s', and c is a normalizing constant.
- Weights proportional to the length of the common boundary of regions s and s'.

For both priors (4.25) and (4.26), the S-dimensional design vector $v = (0, \ldots, 1, \ldots, 0)'$ is a 0/1-incidence vector. Its s-th value is 1 if the corresponding observation is located in site or region s, and zero otherwise. Note that it is allowed to include regions s without any corresponding observation, i. e. we may have $v_i[s] = 0$ for all observations $i = 1, \ldots, n$. Due to the smoothness introduced by the prior specification it is possible to estimate spatial effects even for such regions.

The $S \times S$ penalty matrix K can be shown to be given by an adjacency matrix of the form

$$k_{ss} = w_{s+} \tag{4.27}$$

$$k_{ss'} = \begin{cases} -w_{ss'} & \text{if } s \text{ and } s' \text{ are neighbors,} \\ 0 & \text{otherwise.} \end{cases} \tag{4.28}$$

Obviously, the rows and columns of K sum to zero and again the precision matrix is rank deficient. If the geographical map formed by the regions does not fall into several

pieces, the rank deficiency and, thus, the dimension of the null space equal one. The null space is spanned by a S-dimensional vector of ones. If the map can be decomposed in disconnected parts, higher rank deficiencies are obtained but since this leads to problems when estimating the spatial effect, we will assume connectivity in the following. One possibility to obtain connected maps is to introduce additional neighboring rules to fuse the disconnected parts of the map.

Due to Equation (4.28) the matrix K has a sparse structure since most of its entries equal zero. This is illustrated in Figure 4.6 for the map of former West-Germany. This sparse structure may be exploited in some computations, especially when using Markov Chain Monte Carlo algorithms to estimate structured additive regression models or when sampling from Markov random fields (compare Rue 2001). In mixed model based inference, however, this property is of no particular advantage.

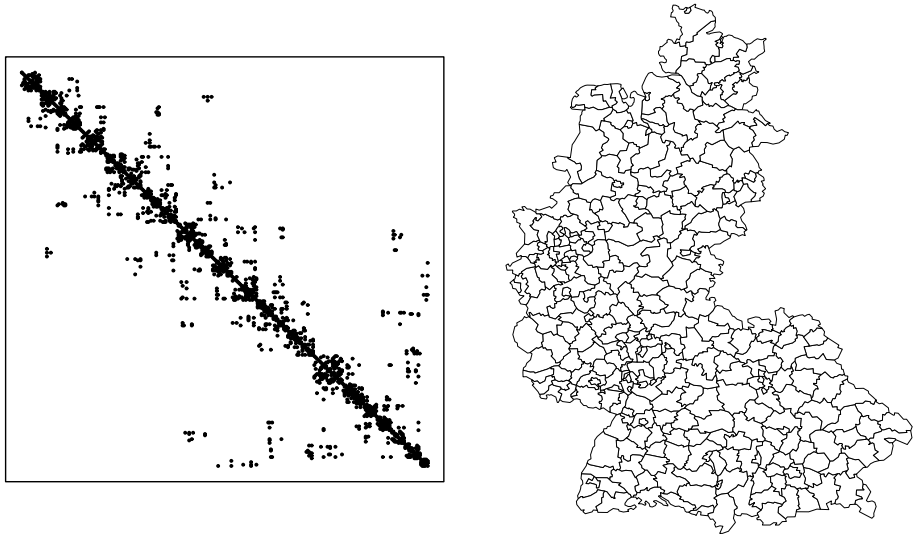


Figure 4.6: Non-zero entries of the precision matrix for a Markov random field defined by the map of West-Germany.

Since geographical information is usually given in terms of irregular shaped regions, extensions of MRFs to higher order two-dimensional random walks are complicated and will not be considered here. More general higher order Markov random fields on regular lattices will be discussed in Section 4.2.6 in the context of penalties for two-dimensional penalized splines.

4.2.3.2 Stationary Gaussian random fields (Kriging)

If exact locations $s = (s_x, s_y)$ are available, Gaussian random field (GRF) priors, originating from the field of geostatistics, can be used to model spatial effects (see for example Diggle, Tawn & Moyeed (1998) or Kammann & Wand (2003) for applications in a spatial regression context). Gaussian random fields have been introduced for spatial interpolation by the South African mining engineer D.G. Krige to map ore grade from drill samples taken at various spatial locations (see Krige 1966). Therefore, the estimation of GRFs is also referred to as Kriging.

In GRFs the spatial component $f_{spat}(s) = \xi_s$ is assumed to follow a zero mean stationary Gaussian random field $\{\xi_s : s \in \mathbb{R}^2\}$ with variance τ^2 and correlation function $\rho(\beta_s, \beta_{s+h})$. In most basic geostatistical models, the correlation function is assumed to be isotropic, i. e. $\rho(\beta_s, \beta_{s+h}) = \rho(||h||)$. This means that correlations between sites that are $||h||$ units apart are the same, regardless of direction and location of the sites. If the assumption of isotropy is questionable, for example when analyzing environmental phenomena like wind speeds, anisotropic correlation functions may be utilized. Yet this considerably increases the complexity of the estimation problem. Some further comments on anisotropic spatial models will be given in Section 4.2.3.5.

Choosing an appropriate correlation function $\rho(r)$, $r = ||h||$, is important since the resulting estimates for the spatial effect inherit properties like continuity and differentiability from the correlation function (Stein 1999, Sec. 2.4). Several proposals have been made in the geostatistics literature, among which the Matérn family is highly recommended due to its flexibility (Stein 1999, p. 31). In its general form, the Matérn correlation function is given by

$$\rho(r; \alpha, \nu) = \frac{1}{2^{\nu-1}\Gamma(\nu)} \left(\frac{r}{\alpha}\right)^{\nu} \mathcal{K}_\nu \left(\frac{r}{\alpha}\right), \quad \alpha > 0, \nu > 0, \tag{4.29}$$

where Γ is the Gamma function and $\mathcal{K}_\nu$ is the modified Bessel function of order ν (see Abramowitz & Stegun 1974, Ch. 9). While evaluation of (4.29) in general requires the numerical evaluation of modified Bessel functions, simple forms $\rho(r; \alpha)$ are obtained for predetermined values $\nu = m + 0.5$, $m = 0, 1, 2, \ldots$, of the smoothness parameter ν, for example

$$\begin{aligned}
\rho(r; \alpha, \nu = 0.5) &= \exp(-|r/\alpha|), \\
\rho(r; \alpha, \nu = 1.5) &= \exp(-|r/\alpha|)(1 + |r/\alpha|), \\
\rho(r; \alpha, \nu = 2.5) &= \exp(-|r/\alpha|)(1 + |r/\alpha| + \tfrac{1}{3}|r/\alpha|^2), \\
\rho(r; \alpha, \nu = 3.5) &= \exp(-|r/\alpha|)(1 + |r/\alpha| + \tfrac{2}{5}|r/\alpha|^2 + \tfrac{1}{15}|r/\alpha|^3).
\end{aligned}$$

Figure 4.7 displays this four correlation functions. With $\nu = 0.5$ we obtain the well known exponential correlation function which is not differentiable in zero and therefore leads to continuous but not differentiable estimates for the spatial effect. Larger values of ν lead to differentiable correlation functions and, hence, to differentiable as well as smoother estimates. In the limit $\nu \to \infty$, we obtain the Gaussian correlation function (Diggle, Ribeiro Jr. & Christensen 2003, Ch. 2.4)

$$\rho(r; \alpha) = \exp(-\tfrac{1}{2}|r/\alpha|^2).$$

The scale parameter α controls how fast correlations die out with increasing distance $r = ||h||$. To simplify the estimation procedure, we determine α in a preprocessing step based on the simple rule

$$\hat{\alpha} = \max_{i,j} ||s_i - s_j||/c. \tag{4.30}$$

The constant $c > 0$ is chosen such that $\rho(c)$ is small (e. g. 0.001), i. e. c specifies the desired effective range of the correlation function. Therefore, the different values of $||s_i - s_j||/\hat{\alpha}$ are spread out over the r-axis of the correlation function and scale invariance of the estimation procedure is ensured. This choice of α has proved to work well in our experience. The

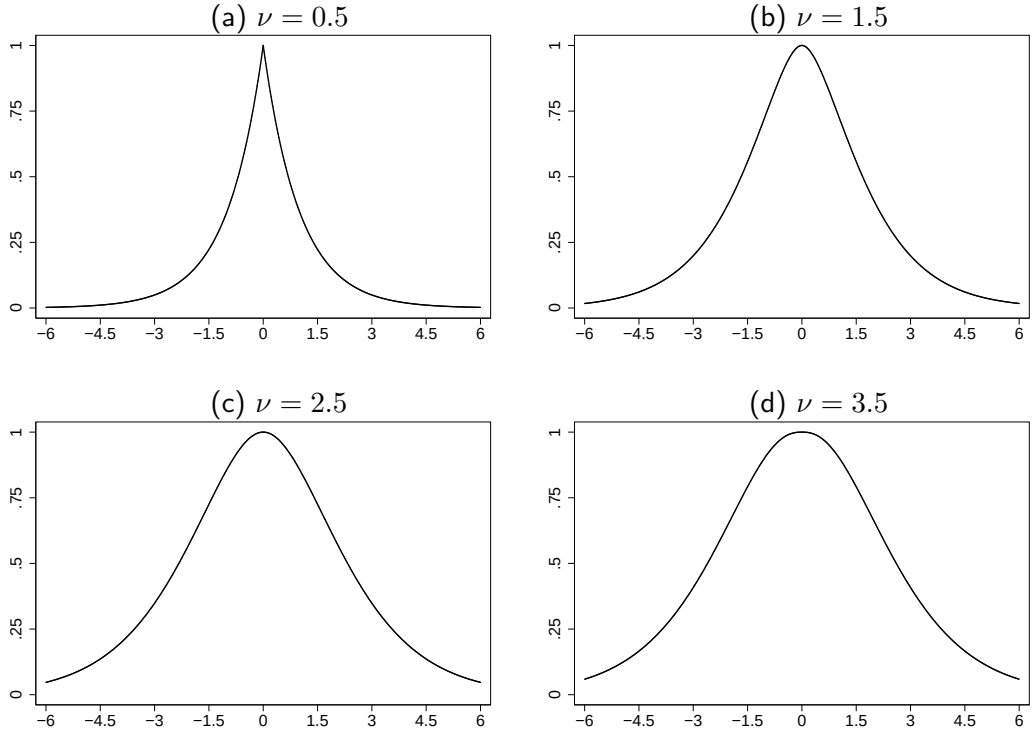


Figure 4.7: Matérn correlation functions for four different choices of the smoothness parameter ν.

limitation to a finite set of values for ν and a pre-chosen value of α based on (4.30) can be justified by the fact that estimation of all parameters of a Matérn correlation function, e. g. based on maximum likelihood, is problematic since different parameter constellations lead to almost indistinguishable correlation functions and hence lead to weakly identified parameters (see Diggle, Ribeiro Jr. & Christensen 2003, p. 66). In addition, Nychka (2000, Sec. 13.3.4) argues that spatial estimates are usually much more sensitive to the choice of the variance τ^2 then to the choice of the other parameters of the Matérn family.

For a finite set of observed spatial locations $s \in \{s_1, \ldots, s_S\}$ the prior for $\xi = (\xi_1, \ldots, \xi_S)'$ can of course be cast into the general form (4.9) since the corresponding random field is assumed to be Gaussian. As a consequence the precision matrix is in fact given by the inverse correlation matrix defined by $\rho(r; \alpha)$, i. e. $K = C^{-1}$ and

$$C[i, j] = \rho(||s_i - s_j||), 1 \leq i, j \leq S.$$

Obviously, this is an example where K is of full rank S. Similar as for Markov random fields, the design vector v for a Gaussian random field is given by a 0/1-incidence vector.

Note that in contrast to the precision matrix of a Markov random field the precision matrix of a Gaussian random field is a full $S \times S$ matrix and has no sparse structure. That is why Gaussian random fields turn out to be computationally demanding in a MCMC analysis based on block updates. However, with mixed model based inference the only relevant property of K is its dimension while the special structure of K is merely not of interest. Strategies for reducing the dimension of the spatial effect will be discussed in the subsequent paragraph on low-rank Kriging.

4.2.3.3 Matérn splines

Gaussian random fields can not only be regarded as priors based on spatial stochastic processes but also have an interpretation as spatial smoothers based on special radial basis functions. Laslett (1994) presents an empirical comparison of splines and GRFs and in a comment on this work, Handcock, Meier & Nychka (1994) present a class of basis functions that contains both thin plate splines and the Matérn correlation functions as special cases. A thorough derivation of the relationship between splines and GRFs can be found in Nychka (2000). Here, we will only give an informal justification.

Consider for the moment a simple spatial regression model

$$y = \beta_0 + f_{spat} + \varepsilon = \beta_0 + V\xi + \varepsilon, \tag{4.31}$$

where $\varepsilon \sim N(0, \sigma^2 I)$ and $\xi \sim N(0, C)$. Then the marginal distribution of y induced by (4.31) is $y \sim N(\beta_0, \Sigma)$, where $\Sigma = V'CV + \sigma^2 I$. Within the classical geostatistical view of the model, C is a covariance matrix obtained from the correlation function $\rho(r)$ and the variance τ^2, i. e. $C[i,j] = \tau^2\rho(||s_i - s_j||)$ and V is a 0/1-incidence matrix formed by the vectors v described in the previous section. However, it is easy to see that different definitions for V and C result in exactly the same marginal model. For example $V[i,j] = \rho(||s_i - s_j||)$ and $C = \tau^2 V^{-1}$ define the same spatial regression model and simply cause a reparametrization of the vector of regression coefficients ξ. In this setting, the design matrix V is no longer an incidence matrix but consists of the correlation function evaluated at the observed locations. Therefore, the functions $\rho(||s_i - s_j||) = \rho_i(s_j)$ can be interpreted as basis functions $\rho_i(s)$ just like the B-spline basis for penalized splines and the distinct locations s_i can be regarded as knots of these basis functions. This results in the so called Matérn splines model. Due to their definition based on the Euclidean distance, the basis functions ρ_i are special radial basis functions. Figure 4.8 shows surface and contour plots of the two-dimensional basis functions obtained with $\nu = 0.5$ and $\nu = 1.5$.

The basis function representation of Kriging estimates also allows to use GRF priors as priors for interaction surfaces of two arbitrary variables. However, problems with this approach may occur if the two interacting variables are not of the same scale since the scale parameter in the correlation function is assumed to be equal for both interacting variables or directions. Therefore, other smoothing techniques developed specifically to model interaction surfaces are preferable, see Section 4.2.6, where the basis functions will be tensor products of one-dimensional B-splines. Interestingly, appropriately normalized B-spline basis functions converge to Gaussian correlation functions when the degree of the B-splines tends to infinity (see Unser, Aldroubi & Eden 1992), which is also the limit of the Matérn correlation function if ν tends to infinity. This establishes another link between the traditional geostatistical view on the estimation of spatial effects and smoothing techniques based on basis functions.

4.2.3.4 Low rank Kriging

For mixed model based inference, the main difference between GRFs and MRFs with respect to their numerical properties, is the dimension of the penalty matrix. For MRFs, the dimension of K equals the number of different regions and is therefore independent

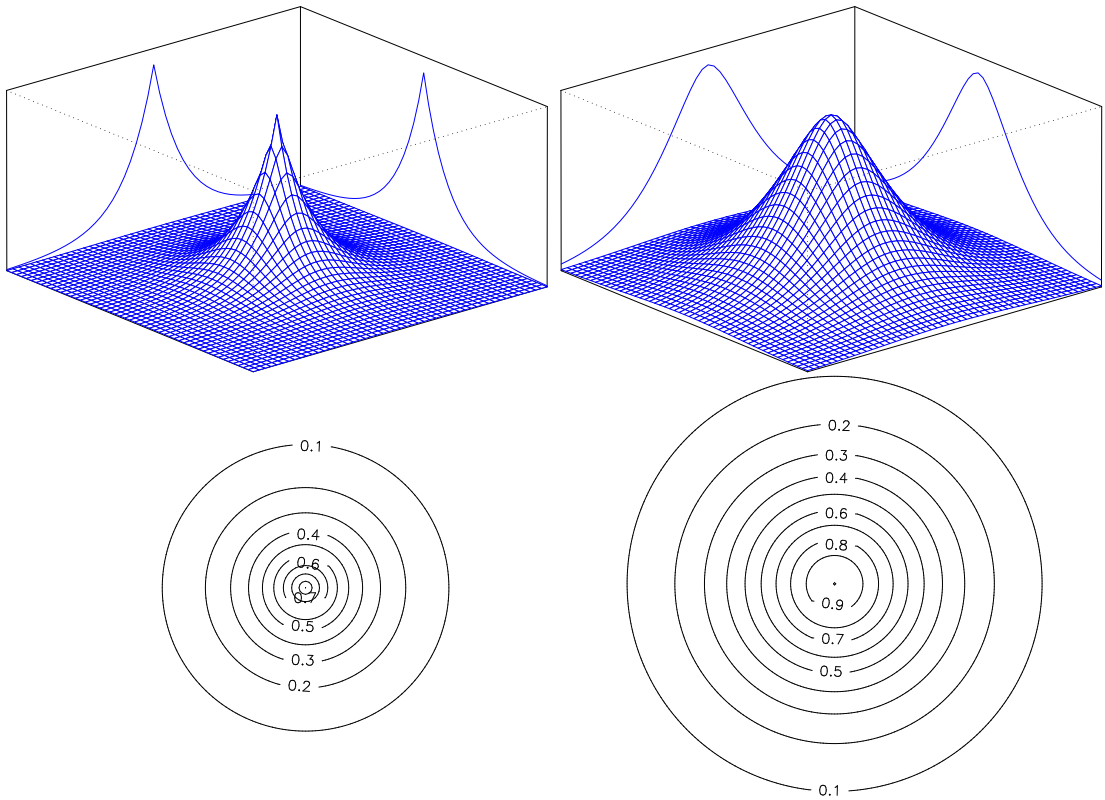


Figure 4.8: Surface and contour plots of Matérn spline basis functions with smoothness parameters $\nu = 0.5$ and $\nu = 1.5$.

from the sample size. On the other hand, for GRFs, the dimension of K is determined by the number of distinct locations which usually is close to or equal to the sample size. So the number of regression coefficients used to describe a MRF is usually much smaller than for a GRF. Accordingly, the estimation of GRFs is computationally much more expensive. To overcome this difficulty, low rank Kriging has been introduced by Nychka, Haaland, O'Connel & Ellner (1998) and formulated in a geoadditive setting by Kammann & Wand (2003). The goal is to approximate Gaussian random fields using only a compact set of parameters.

This goal is achieved by using only a subset $\mathcal{D} = \{\kappa_1, \ldots, \kappa_d\} \subset \mathcal{C} = \{s_1, \ldots, s_S\}$ of the set of all distinct observation points $\mathcal{C}$ as knots of a Matérn spline. These knots can be chosen to be "representative" for the set of distinct locations based on a space filling algorithm (compare Johnson, Moore & Ylvisaker (1990) and Nychka & Saltzman (1998)). Consider the distance measure

$$d(s, \mathcal{D}) = \left(\sum_{\kappa \in \mathcal{D}} ||s - \kappa||^p \right)^{\frac{1}{p}} \tag{4.32}$$

with $p < 0$, between any location $s \in \mathcal{C}$ and a possible set of knots $\mathcal{D}$. Note that $p < 0$ forces the distance measure to be zero for all knots $s = \kappa$. Based on (4.32), an overall coverage criterion for a candidate set of knots is defined via

$$\left(\sum_{s \in \mathcal{C}} d(s, \mathcal{D})^q \right)^{\frac{1}{q}} \tag{4.33}$$

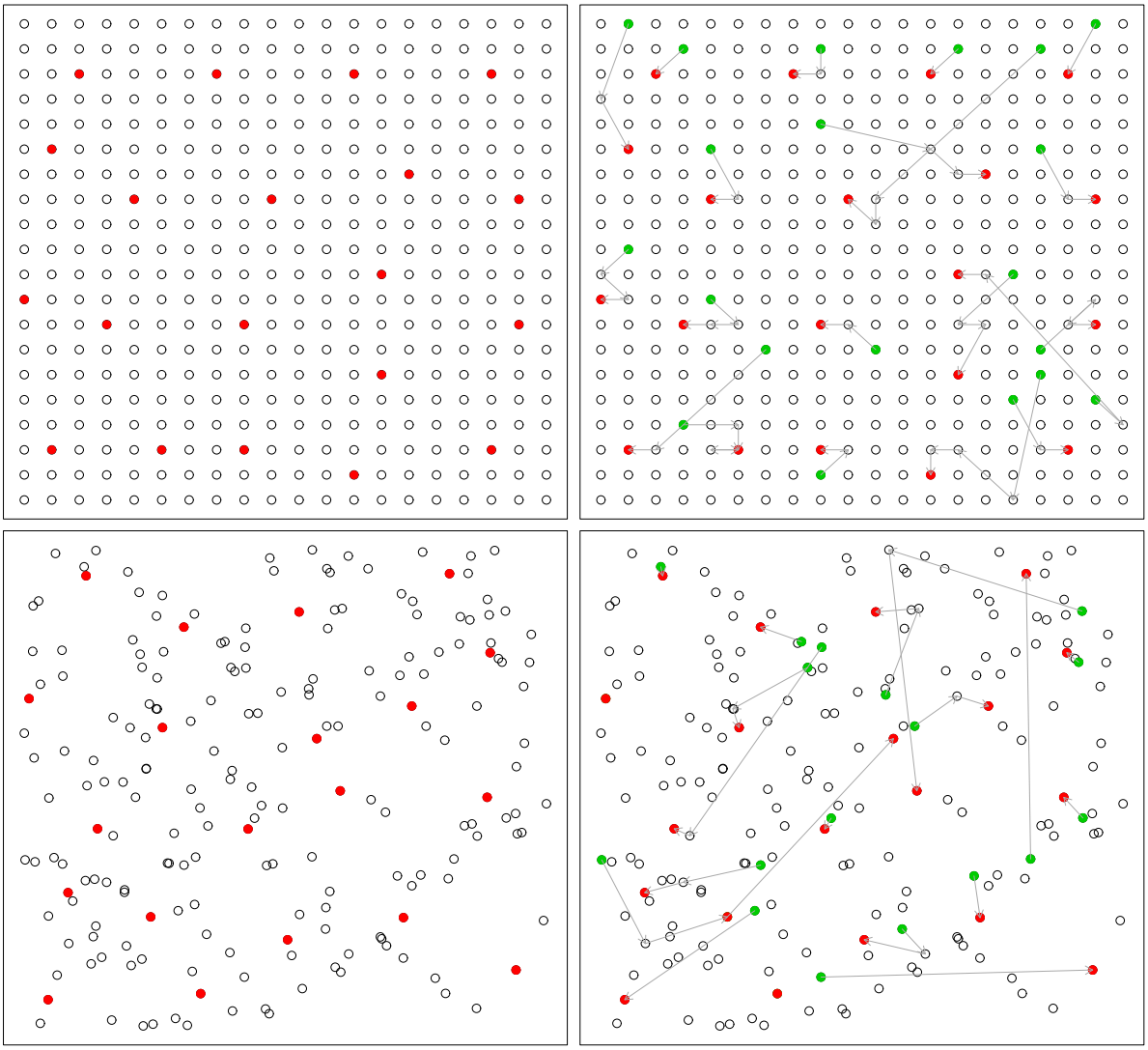


Figure 4.9: Space-filling algorithm: Selection of knots for uniform data (upper panel) and for scattered data (lower panel). The green points indicate the starting design and the red points the optimal set of knots obtained with the space filling algorithm.

with $q > 0$. The parameter q has to be strictly positive in order to obtain large values of $d(s, \mathcal{D})^q$ for points far from $\mathcal{D}$. Standard choices for p and q are $p = -20$ and $q = 20$, respectively. Using a simple swapping algorithm to minimize the coverage criterion yields an optimal set of knots $\mathcal{D}$.

Such a swapping algorithm works as follows:

(i) Randomly choose a start design of candidate knots $\mathcal{D}$ and compute coverage criterion (4.33) for these knots.

(ii) Loop over the set of remaining points $s \in \mathcal{C} \setminus \mathcal{D}$ and the set of candidate knots $\kappa \in \mathcal{C}$. For each combination (s, κ), compute the coverage criterion if κ is replaced by s. Note that this computation can be performed efficiently, since only a small part of the addends in sum (4.32) computed in step (i) is changed.

(iii) Perform the substitution (or swap) that minimizes (4.33) and replace the coverage

criterion with its new value.

(iv) If no further improvement could be achieved, stop the algorithm. Otherwise, return to (ii).

A further reduction of computing time can be achieved if the algorithm does not work with the points themselves but swaps only their indices. An implementation of the swapping algorithm that works in the described manner is available in the R package fields (see Nychka et al. (1998) for a description of its S-Plus predecessor FUNFITS). However, an implementation in a low level programming language like C++ works significantly faster due to the loop-intensive structure of the algorithm.

Figure 4.9 illustrates the knot selection procedure with two examples. The upper panel shows results for a set of regular points on a two-dimensional lattice. Red points indicate the selected, optimal set of knots whereas green points represent the randomly chosen starting design of the swapping algorithm. Arrows are used to show the different design points visited during the minimization procedure. The lower panel of Figure 4.9 shows the same information for an irregular set of points.

Based on knots $\mathcal{C}$, we can now define the approximation $f_{spat}(s) = v'\xi$ with a d-dimensional design vector $v = (\rho(||s - \kappa_1||), \ldots, \rho(||s - \kappa_d||)'$, penalty matrix $K = C$ and $C[i, j] = \rho(||\kappa_i - \kappa_j||)$. The number of knots d allows to control the trade-off between the accuracy of the approximation (d close to the sample size) and the numerical efficiency (d small). In principle, a similar approximation could be attained using a regular grid of knots as for P-splines. However, increasing the number of knots yields the original full Matérn spline for the above definition of the knots, since for $d = S$ we obtain $\mathcal{D} = \mathcal{C}$, while with a regular grid of knots this desirable property is lost.

4.2.3.5 Anisotropic spatial effects

The Kriging approaches discussed so far share the property of isotropy, since the correlation between two points is defined upon the Euclidean distance of the two points. This assumption is likely to be violated in situations where, for example, the spatial effect is used as a surrogate for ecological properties such as the prevailing environmental conditions. A relatively easy way to account for anisotropy is to introduce different distance measures than the Euclidean distance, e. g. replacing $||h|| = \sqrt{h'h}$ with $\sqrt{h'Bh}$ where B is a positive definite matrix. An intuitive interpretation of anisotropy is achieved by choosing

$$B = R(\psi_A)'S(\psi_R)R(\psi_A), \tag{4.34}$$

where $R(\psi_A)$ is a rotation matrix with anisotropy angle $\psi_A \in [0, 2\pi]$ of the form

$$R(\psi_A) = \begin{pmatrix} \cos(\psi_A) & \sin(\psi_A) \\ -\sin(\psi_A) & \cos(\psi_A) \end{pmatrix},$$

and $S(\psi_R)$ is a prolongation matrix with anisotropy ratio $\psi_R \geq 1$ given by

$$S(\psi_R) = \begin{pmatrix} \psi_R^{-1} & 0 \\ 0 & 1 \end{pmatrix}$$

(see Chiles & Delfiner 1999, Ch. 2.5.2). The two parameters ψ_A and ψ_R exhibit an intuitive interpretation. With $\psi_R = 1$, we obtain isotropic correlation function while $\psi_R > 1$ leads to elliptically shaped correlation functions. More precisely, ψ_R is defined by the ratio of the ranges of the correlation function along the two principal axes. The anisotropy angle ψ_A causes a rotation of the correlation function around the origin. This is illustrated in Figure 4.10 for different parameter constellations.

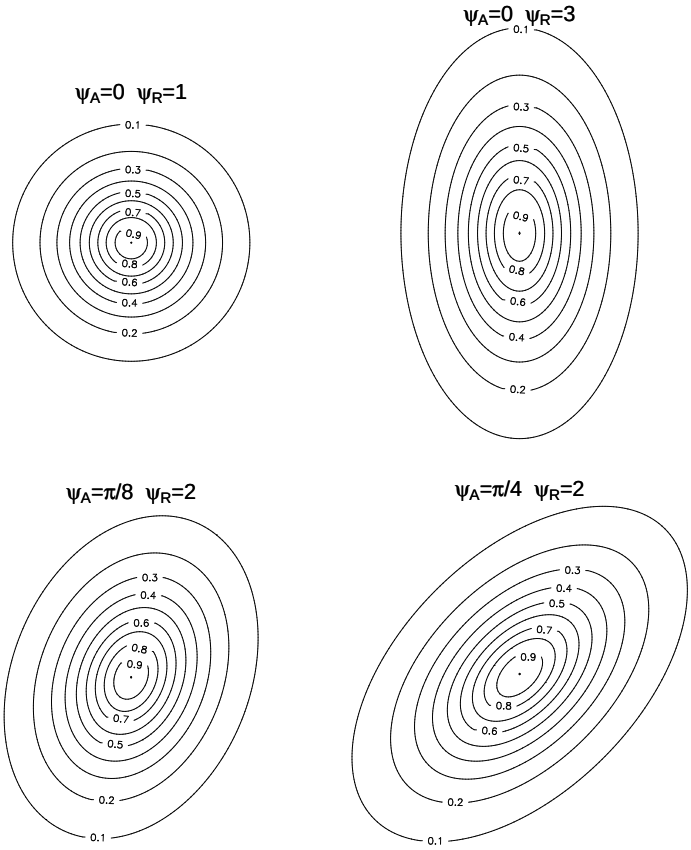


Figure 4.10: Contour plots of correlation functions with different choices for the anisotropy angle ψ_A and the anisotropy ratio ψ_R.

The class of correlation functions resulting from definition (4.34) is called geometrically anisotropic in the geostatistics literature. A broader class of anisotropic correlation functions defined as the product of several isotropic correlation functions allowing for far more complex shaped contours is introduced by Ecker & Gelfand (2003). The reason why anisotropic models are not widely used in statistical models is the introduction of additional parameters that have to be chosen or estimated. Since their influence on the likelihood is of complex, nonlinear form, maximum likelihood estimation becomes considerably more difficult. However, for pre-chosen values of ψ_A and ψ_R models with anisotropic correlation functions are easily implemented.

While the above mentioned approaches merely modify the functional form of the correlation function, further types of anisotropy can be considered. Zimmermann (1993) provides a thorough overview and classification of different kinds of anisotropy and also discusses consequences on the properties of the spatial stochastic process when a special type of anisotropy is assumed. Since most types of anisotropy have undesired side effects

such as nonstationarity of the underlying spatial process, different approaches than the ones discussed above are very rarely realized in practice.

4.2.3.6 Discrete versus continuous spatial modeling

Although described separately, differences between methods for modeling spatial effects based on discrete spatial information (i. e. region data) and effects based on coordinates are not that clear-cut. For example, approaches for exact locations can also be used in the case of connected geographical regions based on the coordinates of the centroids of the regions. Conversely, we can apply MRFs to exact locations if neighborhoods are defined based on a distance measure (see Section 12 for an application) or via discretization of the observation area. However, continuous spatial modeling allows to evaluate a spatial function at arbitrary positions and, hence, interpolations and extrapolations are available (we will make use of this in the example presented in Section 12).

From a theoretical point of view, it can be shown that GRFs may be approximated by (higher order) MRFs (see Rue & Tjelmeland 2002). In addition, MRFs have the advantage, that no stationarity has to be assumed for the spatial effect (see Rue & Held (2005, Ch. 2.6) for a discussion of stationary Gaussian MRFs).

In general, it is not clear which of the presented approaches leads to the "best" fit. For data observed on a discrete lattice, MRFs seem to be most appropriate. If the exact locations are available surface estimators may be more natural, particularly because predictions for unobserved locations are available. However, in some situations surface estimators lead to an improved fit compared to MRFs even for discrete lattices and vice versa. A general approach that can handle both situations is given by Müller, Stadtmüller & Tabnak (1997). We will compare discrete and continuous spatial models in a simulation study in Section 8.

4.2.4 Group indicators, cluster-specific effects and unstructured spatial effects

In many situations, for example longitudinal data, we encounter the problem of unexplained heterogeneity among clusters of observations caused by unobserved covariates. Suppose $c \in \{1, \ldots, C\}$ is a cluster variable indicating the cluster a particular observation belongs to. In case of longitudinal data, clusters are often defined by the repeated measurements on individuals with the cluster index c being the individual index. A common approach to overcome the difficulties of unobserved heterogeneity is to introduce additional cluster-specific Gaussian i. i. d. effects $f(c) = \xi_c$ with

$$\xi_c \sim N(0, \tau^2), \quad c = 1, \ldots, C. \tag{4.35}$$

In this case, the design vector v is a C-dimensional 0/1-incidence vector and the penalty matrix is the identity matrix, i. e. $K = I$. Therefore, the prior of the joint vector of regression coefficients $\xi = (\xi_1, \ldots, \xi_C)'$ is proper and (4.35) is a second example of a prior where no rank deficiency occurs. From a classical perspective, (4.35) defines i. i. d. random effects. However, from a Bayesian point of view, all unknown parameters are assumed to be random and, hence, the notation "random effects" in this context is misleading.

Prior (4.35) may also be used for a more sophisticated modeling of spatial effects if the spatial effect is split into a spatially structured (smooth) part f_{str} and a spatially unstructured (unsmooth) part f_{unstr} as mentioned in Section 4.2.3. Identifying the cluster index c with the spatial index s, the unstructured spatial effects $f_{unstr}(s) = \xi_s$ are assumed to be i. i. d. random effects $\xi_s \sim N(0, \tau^2_{unstr})$.

From a classical perspective, prior (4.35) defines only random intercepts. Models with random slopes are discussed in the following subsection, since they can be seen as varying coefficient models where the cluster indicator acts as effect modifier.

4.2.5 Varying coefficients

The model components considered so far are not appropriate for modeling interactions between covariates. Facing such situations, varying coefficient models have first been introduced by Hastie & Tibshirani (1993) in the context of smoothing splines. Here, the effect of a covariate z is assumed to vary smoothly over the range of a second covariate x, i. e.,

$$f(x, z) = g(x)z. \tag{4.36}$$

In most cases the interacting covariate z is categorical whereas the effect modifier may be either continuous, spatial or an unordered group indicator. Concerning the function g, all priors defined in Sections 4.2.2 for continuous effect modifiers, 4.2.3 for spatial effect modifiers and 4.2.4 for unordered group indicators as effect modifiers can be used. In Hastie & Tibshirani (1993) continuous effect modifiers have been considered exclusively. Models with spatial effect modifiers are presented for example in Fahrmeir et al. (2003) or Gamerman, Moreira & Rue (2003) to model space-time interactions. In the geography literature this type of models is well known as geographically weighted regression (see e. g. Fotheringham et al. 2002). From a classical point of view, models with unordered group indicators as effect modifiers are called models with random slopes.

In matrix notation we obtain $f = \text{diag}(z_1, \ldots, z_n)V^*\xi$ for the vector of function evaluations, where V^* is the design matrix corresponding to the prior for g. Hence, the overall design matrix is given by

$$V = \text{diag}(z_1, \ldots, z_n)V^*.$$

4.2.6 Interaction surfaces

If both interacting covariates are continuous, varying coefficient models are usually too restrictive to model interactions. In this case, a more flexible approach can be based on nonparametric two-dimensional surface fitting. We follow an approach based on bivariate P-splines as described in Lang & Brezger (2004) and Brezger & Lang (2005). In analogy to univariate P-Splines described in Section 4.2.2, we assume that the unknown surface $f(x_1, x_2)$ can be approximated by the tensor product of two univariate B-splines, i. e.

$$f(x_1, x_2) = \sum_{m_1=1}^{d_1} \sum_{m_2=1}^{d_2} \beta_{m_1 m_2} B_{m_1}(x_1) B_{m_2}(x_2). \tag{4.37}$$

Equation (4.37) implicitly defines bivariate B-spline basis functions

$$B_{m_1 m_2}(x_1, x_2) = B_{m_1}(x_1) \cdot B_{m_2}(x_2) \tag{4.38}$$

which are shown in Figure 4.11 for different degrees l. Similarly as with univariate B-splines, an increasing degree l leads to smoother basis functions in terms of continuity and differentiability. More formally, a bivariate polynomial spline is defined as follows (Dierckx 1993):

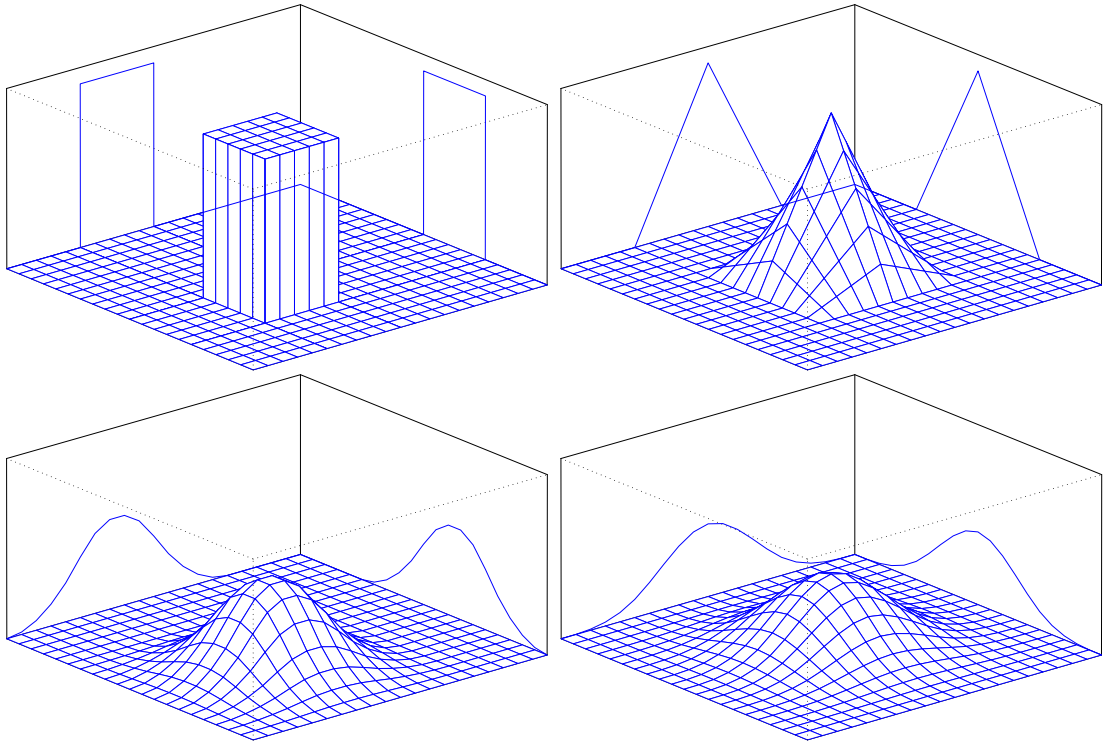


Figure 4.11: Tensor product B-spline basis functions of degree l =0, 1, 2 and 3.

A function $g : [a, b] \times [c, d] \to \mathbb{R}$ is called a bivariate polynomial spline of degree l with knots $x_1^{min} = \kappa_0^{(1)} < \ldots < \kappa_{M_1}^{(1)} = x_1^{max}$ and $x_2^{min} = \kappa_0^{(2)} < \ldots < \kappa_{M_2}^{(2)} = x_2^{max}$ if it satisfies the following conditions:

1. The partial derivatives
$$\frac{\partial^{i+j} g(x_1, x_2)}{\partial^i x_1 \partial^j x_2}, \quad 0 \leq i, j < l,$$
are continuous and

2. $g(x_1, x_2)$ is a polynomial of degree l for $(x_1, x_2) \in [\kappa_{m_1}^{(1)}, \kappa_{m_1+1}^{(1)}) \times [\kappa_{m_2}^{(2)}, \kappa_{m_2+1}^{(2)})$, $m_1 = 0, \ldots, M_1 - 1$, $m_2 = 0, \ldots, M_2 - 1$.

The bivariate polynomial splines then form a $d_1 d_2 = (M_1 + l)(M_2 + l)$-dimensional vector space and the B-splines (4.38) can be shown to define a basis of this function space.

In contrast to the Matérn splines discussed in Section 4.2.3, the tensor product B-spline basis does not define a radial basis. This can be seen most clearly for the basis functions of degree zero and becomes less visible for increasing degree. Figure 4.12 shows contour-plots for tensor product B-splines of degree $l = 1, 2$ and 3. Obviously, the amount of

nonradiality decreases rapidly when the degree of the basis functions is increased. This reflects the fact that in the limit as the degree goes to infinity, appropriately normalized B-spline basis functions converge to Gaussian radial basis functions (Unser et al. 1992).

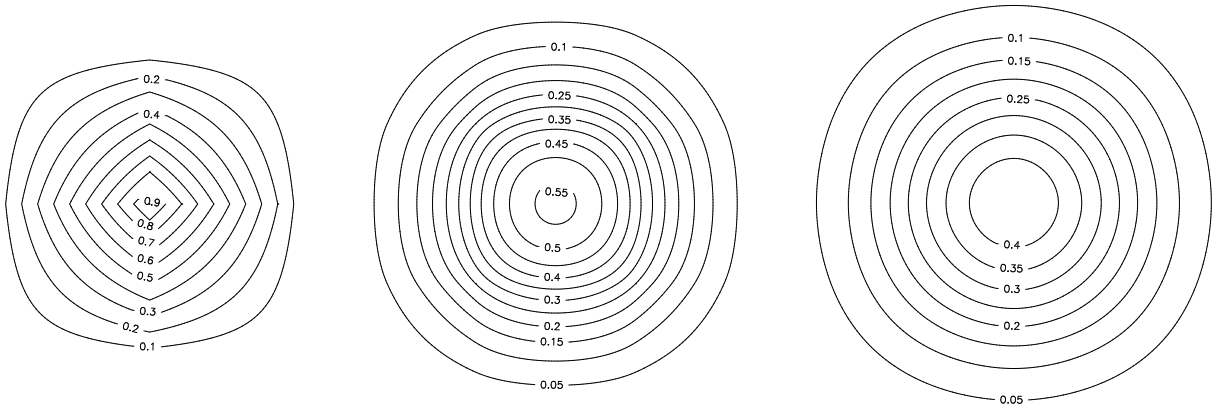


Figure 4.12: Contour plots for tensor product B-spline basis functions of degree l =1, 2 and 3.

In analogy to one-dimensional P-splines, the $d = d_1 d_2$-dimensional design vector v is composed of basis functions evaluated at the respective covariates, i. e.

$$\begin{aligned} v &= (B_{11}(x_1, x_2), \ldots, B_{d_1 1}(x_1, x_2), \ldots, B_{1 d_2}(x_1, x_2), B_{d_1 d_2}(x_1, x_2))' \\ &\quad (B_1(x_1) \cdot B_1(x_2), \ldots, B_{d_1}(x_1) \cdot B_1(x_2), \ldots, B_1(x_1) \cdot B_{d_2}(x_2), \ldots, B_{d_1}(x_1) \cdot B_{d_2}(x_2))'. \end{aligned}$$

The vector of regression coefficients $\xi = (\xi_{11}, \ldots, \xi_{d_1 d_2})'$ can be arranged corresponding to the two-dimensional regular array of basis functions in the (x_1, x_2)-plane. Following the idea of one-dimensional P-splines, we assign two-dimensional random walk priors to enforce smoothness of the estimated surface. Starting with a bivariate first order random walk, we will later discuss several approaches for defining higher order generalizations of Markov random field priors for regular lattices.

4.2.6.1 First order random walk

The easiest way to define a bivariate random walk is to apply a first order random walk in two dimensions, i. e. Markov random field prior (4.25) with neighborhoods defined on a regular lattice. The most commonly used neighborhood structure in this situation is based on the four next neighbors as indicated in Figure 4.13a. However, different schemes are conceivable, e. g. with eight neighbors as shown in Figure 4.13b. For both priors, the penalty matrix can be easily computed from formulae (4.27) and (4.28). It is also obvious from the discussion in Section 4.2.3.1 that the penalty matrix in this case again has a one-dimensional null space spanned by a d-dimensional vector of ones.

Another way to construct the penalty matrix for a bivariate first order random walk with four neighbors is based on the Kronecker sum of the penalty matrices of two one-dimensional first order random walks. If $K_1^{(1)}$ and $K_2^{(1)}$ are penalty matrices for first order random walks in x_1 and x_2 direction, respectively, and I_d denotes a d-dimensional identity

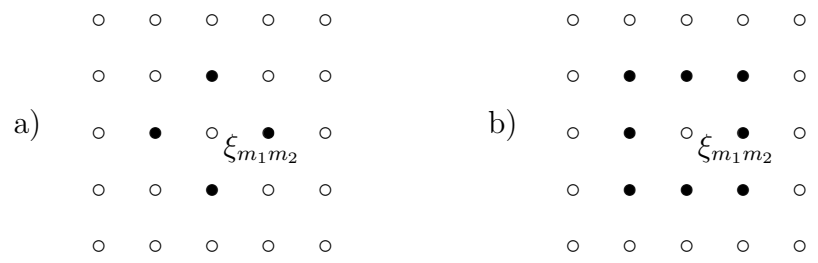


Figure 4.13: Neighborhoods on a regular lattice: The four (eight) next neighbors of $\xi_{m_1m_2}$ are indicated as black points.

matrix, it is easy to verify that

$$K = I_{d_2} \otimes K_1^{(1)} + K_2^{(1)} \otimes I_{d_1} \tag{4.39}$$

is the precision matrix of a bivariate first order random walk. This formulation also allows for the introduction of anisotropic penalizations by the inclusion of direction-specific smoothing parameters (see Eilers & Marx (2003) in the context of two-dimensional penalized splines and Besag & Higdon (1999) in the context of Markov random fields).

A last possibility to define a bivariate first order random walk is to construct it from a local linear fit through the four nearest neighbors ξ_{m_1-1,m_2}, ξ_{m_1,m_2-1}, ξ_{m_1+1,m_2} and ξ_{m_1,m_2+1} in analogy to the discussion of one-dimensional P-splines in Section 4.2.2.

4.2.6.2 Kronecker sum of two second order random walks

A first possibility to define bivariate second order random walks is to replace the penalty matrices $K_j^{(1)}$ of first order random walks in (4.39) with penalty matrices $K_j^{(2)}$ of second order random walks leading to the precision matrix

$$K = I_{d_2} \otimes K_1^{(2)} + K_2^{(2)} \otimes I_{d_1}. \tag{4.40}$$

This induces a dependency structure where the value of $\xi_{m_1m_2}$ depends on the eight next neighbors along the x_1 and x_2 axes (see Figure 4.14). Note that no dependence on neighbors along the diagonals is assumed. In principle, any combination of univariate first and second order random walks and even higher order random walks can be used to construct two-dimensional priors as in (4.40). However, if there is no obvious reason why the two directions should be treated differently, it may be more natural to use either first or second order random walks.

The elements of the precision matrix for the Kronecker sum of two-dimensional random walks can in principle be read from (4.40). Figure 4.15 additionally displays the non-zero entries of K and shows the modifications that are caused by boundary restrictions. This allows to compare the structure of precision matrix (4.40) with other possible definitions of bivariate second order random walks discussed in the following sections.

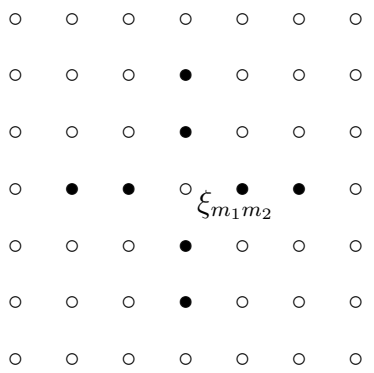


Figure 4.14: Dependence structure of a bivariate second order random walk based on the Kronecker sum of two univariate second order random walks.

The rank of precision matrix (4.40) can be shown to be $d-4$ and the four-dimensional null space is spanned by the column vectors in

$$\begin{pmatrix}
1 & \kappa_1^{(1)} & \kappa_1^{(2)} & \kappa_1^{(1)} \cdot \kappa_1^{(2)} \\
1 & \kappa_2^{(1)} & \kappa_1^{(2)} & \kappa_2^{(1)} \cdot \kappa_1^{(2)} \\
\vdots & \vdots & \vdots & \vdots \\
1 & \kappa_{d_1}^{(1)} & \kappa_1^{(2)} & \kappa_{d_1}^{(1)} \cdot \kappa_1^{(2)} \\
1 & \kappa_1^{(1)} & \kappa_2^{(2)} & \kappa_1^{(1)} \cdot \kappa_2^{(2)} \\
\vdots & \vdots & \vdots & \vdots \\
1 & \kappa_{d_1}^{(1)} & \kappa_2^{(2)} & \kappa_{d_1}^{(1)} \cdot \kappa_2^{(2)} \\
\vdots & \vdots & \vdots & \vdots \\
1 & \kappa_1^{(1)} & \kappa_{d_2}^{(2)} & \kappa_1^{(1)} \cdot \kappa_{d_2}^{(2)} \\
\vdots & \vdots & \vdots & \vdots \\
1 & \kappa_{d_1}^{(1)} & \kappa_{d_2}^{(2)} & \kappa_{d_1}^{(1)} \cdot \kappa_{d_2}^{(2)}
\end{pmatrix},$$

where $\kappa_1^{(1)}, \ldots, \kappa_{d_1}^{(1)}$ and $\kappa_1^{(2)}, \ldots, \kappa_{d_2}^{(2)}$ are the knots in x_1 and x_2 direction, respectively. Interestingly the basis contains an interaction term although no dependencies along the diagonals had been assumed in the model formulation. Therefore, the null space does not only consist of simple planes but also contains more complicated structures with changing slopes in x_1 and x_2 direction.

4.2.6.3 Local quadratic fit

Another possibility to define second order random walks on lattices has been proposed by Besag & Kooperberg (1995). Since the coefficients of the precision matrix for a bivariate first order random walk can be generated from a local linear fit through the four next neighbors, their idea is to perform a local quadratic fit through the next twelve neighbors. To obtain the coefficients associated with $\xi_{m_1m_2}$, we have to compute the prediction for

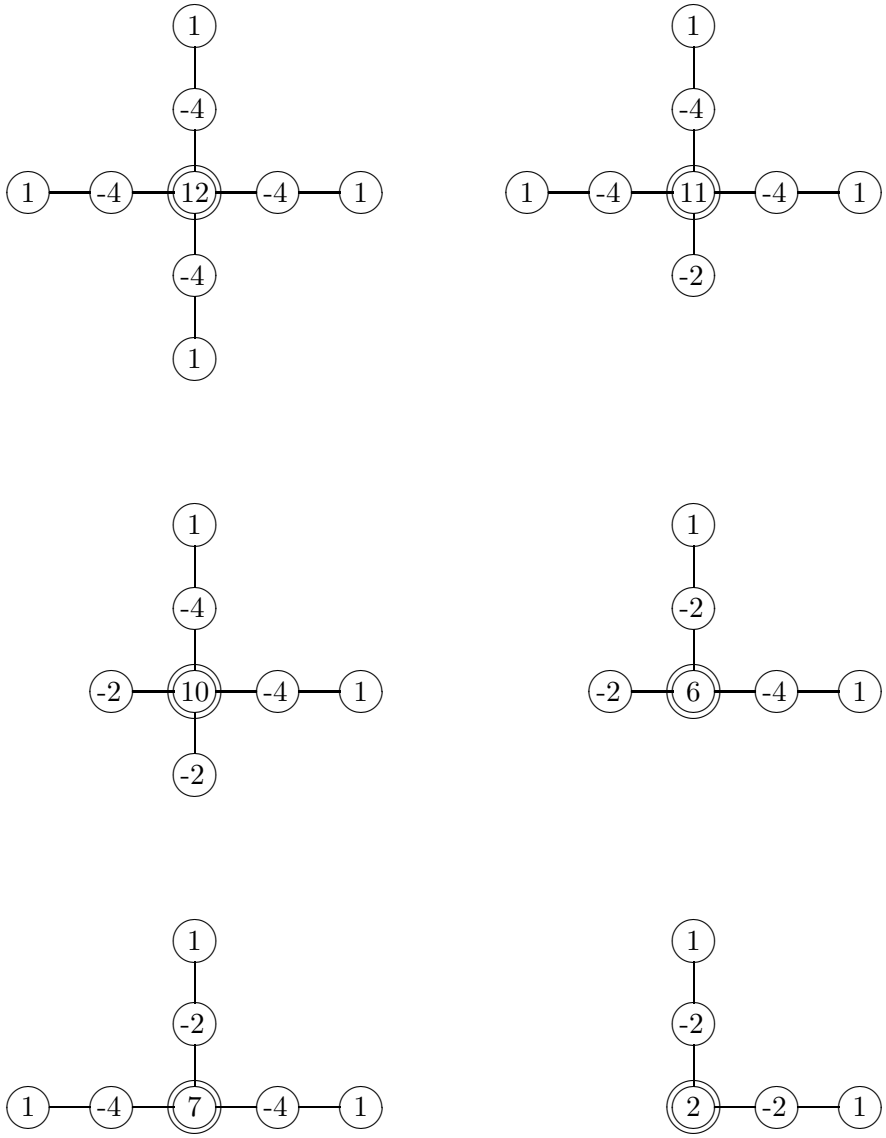


Figure 4.15: Kronecker sum of two univariate second order random walks: Coefficients of the precision matrix.

$\xi_{m_1 m_2}$ from a regression model with design matrix

$$\begin{pmatrix}
1 & m_1-2 & m_2 & (m_1-2)^2 & m_2^2 \\
1 & m_1-1 & m_2-1 & (m_1-1)^2 & (m_2-1)^2 \\
1 & m_1-1 & m_2 & (m_1-1)^2 & m_2^2 \\
1 & m_1-1 & m_2+1 & (m_1-1)^2 & (m_2+1)^2 \\
1 & m_1 & m_2-2 & m_1^2 & (m_2-2)^2 \\
1 & m_1 & m_2-1 & m_1^2 & (m_2-1)^2 \\
1 & m_1 & m_2+1 & m_1^2 & (m_2+1)^2 \\
1 & m_1 & m_2+2 & m_1^2 & (m_2+2)^2 \\
1 & m_1+1 & m_2-1 & (m_1+1)^2 & (m_2-1)^2 \\
1 & m_1+1 & m_2 & (m_1+1)^2 & m_2^2 \\
1 & m_1+1 & m_2+1 & (m_1+1)^2 & (m_2+1)^2 \\
1 & m_1+2 & m_2 & (m_1+2)^2 & m_2^2
\end{pmatrix}$$

and dependent variable

$$\begin{pmatrix} \xi_{m_1-2,m_2} \\ \xi_{m_1-1,m_2-1} \\ \xi_{m_1-1,m_2} \\ \xi_{m_1-1,m_2+1} \\ \xi_{m_1,m_2-2} \\ \xi_{m_1,m_2-1} \\ \xi_{m_1,m_2+1} \\ \xi_{m_1,m_2+2} \\ \xi_{m_1+1,m_2-1} \\ \xi_{m_1+1,m_2} \\ \xi_{m_1+1,m_2+1} \\ \xi_{m_1+2,m_2} \end{pmatrix}.$$

This yields the dependence structure and coefficients shown in Figure 4.16.

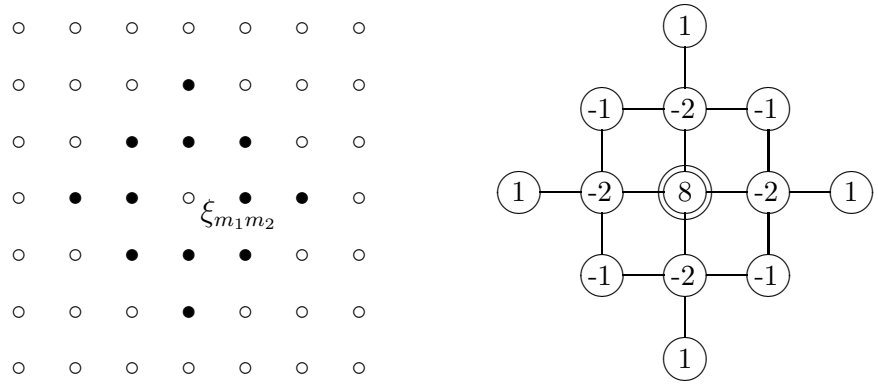


Figure 4.16: Dependence structure and coefficients of the precision matrix for a two-dimensional second order random walk based on a local quadratic fit through the twelve next neighbors.

Although this extension of first order random walks is quite intuitive, it exhibits a number of drawbacks: Most importantly, it is unclear, how to incorporate boundary modifications in this model, since the coefficients given in Figure 4.16 are only valid in the interior of the lattice. Schmid (2004, p. 65) proposes to compute the penalty matrix via

$$K = I_{d_2} \otimes K_1^{(2)} + K_2^{(2)} \otimes I_{d_1} - K_2^{(1)} \otimes K_1^{(1)}, \tag{4.41}$$

where $K_1^{(k)}$ and $K_2^{(k)}$ are penalty matrices for univariate random walks of order k in x_1 and x_2 direction, respectively. In the interior of the lattice, (4.41) leads to the correct coefficients but using it to construct the full precision matrix results in an indefinite matrix, i. e. some of the eigenvalues of K are negative although K should be positive semidefinite by definition.

Other modifications propose to extend the lattice at the boundaries and then to work on a torus (Besag & Higdon 1999), resulting in a theoretically infinite lattice. However, this approach is not very intuitive since observations of extremely high distance become almost neighbors. An alternative is to restrict an infinite lattice to the finite case without

correcting for the boundaries (see Rue & Held 2005, p. 115). While this approach may be used with MCMC simulation techniques, it is not very useful in a mixed model setting since it enlarges the null space of the precision matrix and leads to a noninterpretable basis of this null space.

A further argument against the local quadratic fit is given by Rue & Held (2005, p. 119): The penalty derived from the precision matrix cannot be represented in terms of increments depending only on the same neighborhood of twelve nearest neighbors. In contrast, the increments depend on a much larger neighborhood which is not a desirable property.

4.2.6.4 Approximation of the biharmonic differential operator

To overcome the difficulties discussed in the previous paragraph, Rue & Held (2005) propose to define a two-dimensional second order random walk based on a penalty obtained from an approximation of the biharmonic differential operator

$$\left(\frac{\partial^2}{\partial^2 x_1} + \frac{\partial^2}{\partial^2 x_2}\right)^2 = \frac{\partial^4}{\partial^4 x_1} + 2\frac{\partial^4}{\partial^2 x_1 \partial^2 x_2} + \frac{\partial^4}{\partial^4 x_2}. \tag{4.42}$$

The biharmonic differential operator is a two-dimensional extension of the squared second derivative, which is routinely used for penalization in nonparametric regression in combination with smoothing splines. Seeking the maximum solution of the penalized log-likelihood criterion

$$l(\xi) - \frac{1}{2}\lambda \int\int \left(\frac{\partial^2}{\partial^2 x_1} + \frac{\partial^2}{\partial^2 x_2}\right)^2 f(x_1, x_2) dx_1 dx_2$$

leads to the thin-plate spline which is the two-dimensional analogon to the natural cubic smoothing spline. Approximating the derivatives in (4.42) by difference operators based on the twelve nearest neighbors results in a precision matrix with non-zero elements defined by

$$-\left(\Delta^2_{(1,0)} + \Delta^2_{(0,1)}\right)^2 = -\left(\Delta^4_{(1,0)} + 2\Delta^2_{(1,0)}\Delta^2_{(0,1)} + \Delta^4_{(0,1)}\right), \tag{4.43}$$

where $\Delta^2_{(1,0)}$ and $\Delta^2_{(0,1)}$ denote second order difference operators along the x_1 and x_2 direction, respectively. This yields the coefficients given in Figure 4.17. Restrictions at the boundaries can be incorporated via careful modifications of the biharmonic differential operator and are also included in Figure 4.17 (compare Terzopoulos (1988) for details). Note that the penalty defined in (4.43) corresponds to squares of increments

$$(\xi_{m_1+1,m_2} + \xi_{m_1-1,m_2} + \xi_{m_1,m_2+1} + \xi_{m_1,m_2-1}) - 4\xi_{m_1 m_2}$$

which can be regarded as extensions of the increments defined for univariate second order random walks. Rue & Held (2005) moreover discuss refined ways to approximate the biharmonic differential operator using larger neighborhoods or by inclusion of differences along the diagonals in approximation (4.43).

The null space of the bivariate random walk based on approximating the biharmonic

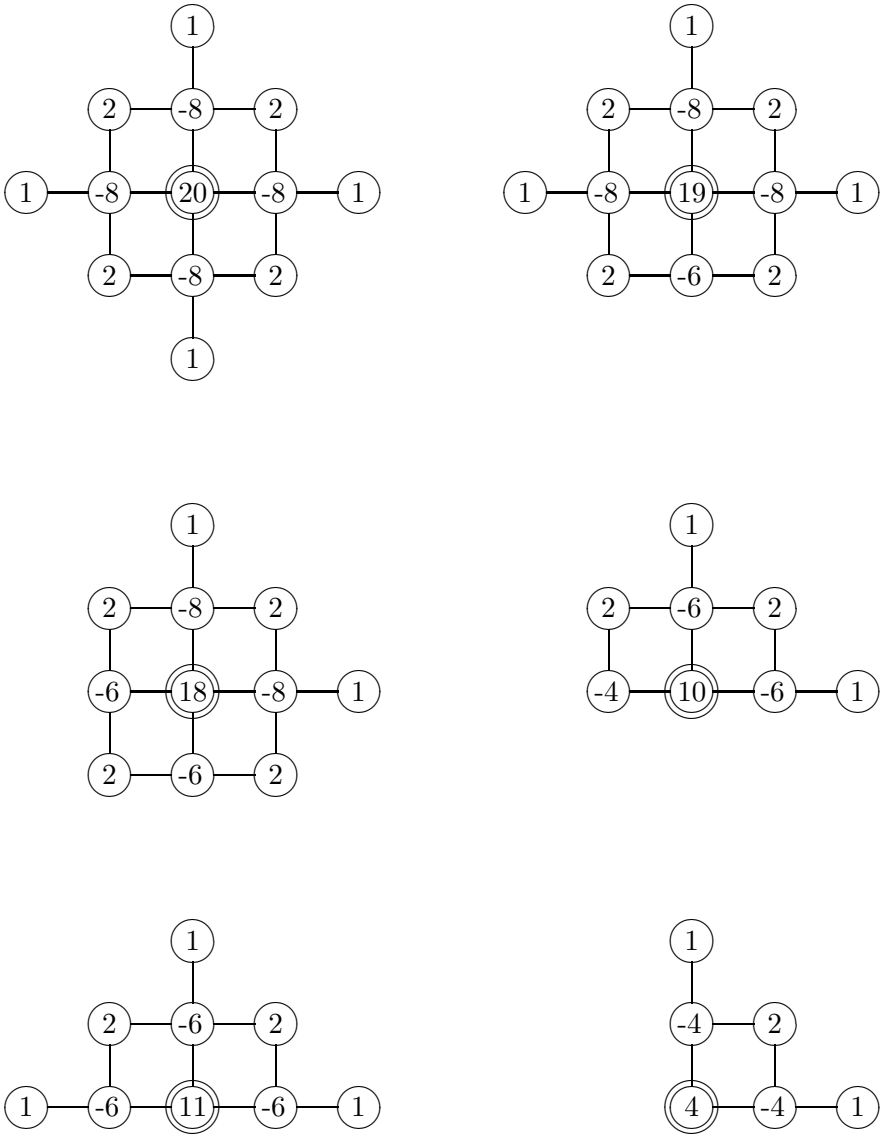


Figure 4.17: Approximation to the biharmonic differential operator: Coefficients of the precision matrix.

differential operator is of dimension three and is spanned by the column vectors of

$$\begin{pmatrix} 1 & \kappa_1^{(1)} & \kappa_1^{(2)} \\ 1 & \kappa_2^{(1)} & \kappa_1^{(2)} \\ \vdots & \vdots & \vdots \\ 1 & \kappa_{d_1}^{(1)} & \kappa_1^{(2)} \\ 1 & \kappa_1^{(1)} & \kappa_2^{(2)} \\ \vdots & \vdots & \vdots \\ 1 & \kappa_{d_1}^{(1)} & \kappa_2^{(2)} \\ \vdots & \vdots & \vdots \\ 1 & \kappa_1^{(1)} & \kappa_{d_2}^{(2)} \\ \vdots & \vdots & \vdots \\ 1 & \kappa_{d_1}^{(1)} & \kappa_{d_2}^{(2)} \end{pmatrix}.$$

In contrast to the null space for the Kronecker sum of two second order random walks, the basis does not contain an interaction. Consequently, it consists of simple planes in the (x_1, x_2)-space.

4.2.6.5 Kronecker product of two random walks

A penalty based on eight or 24 neighbors, as indicated in Figure 4.18, can be defined by the Kronecker product of two first or second order random walk penalty matrices (see e. g. Besag & Kooperberg (1995) or Besag & Higdon (1999)), i. e.

$$K = K_2^{(k)} \otimes K_1^{(k)}. \tag{4.44}$$

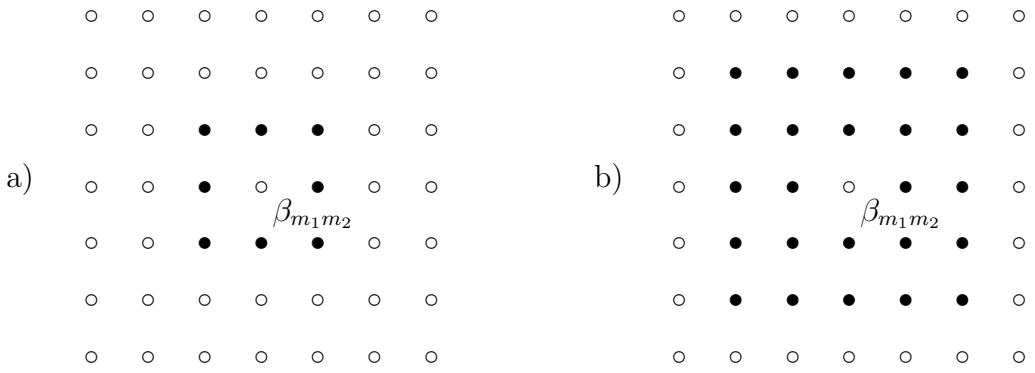


Figure 4.18: Neighborhoods consisting of eight and 24 neighbors.

For first order random walks, this corresponds to a penalization of differences of differences as increments

$$\Delta_{(1,0)}\Delta_{(0,1)}\xi_{m_1m_2} = \xi_{m_1m_2} - \xi_{m_1-1,m_2} - \xi_{m_1,m_2-1} + \xi_{m_1-1,m_2-1}$$

and results in a penalty matrix with weights given in Figure 4.19 for the interior of the lattice.

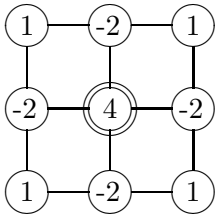


Figure 4.19: Weights in the interior of a precision matrix defined by the Kronecker product of two first order random walks.

Second order versions can be interpreted in a similar manner but will not be pursued further since usually Kronecker product priors are supposed to be unsuitable as priors for

two-dimensional surfaces. This can already be seen from the null space of penalty matrix (4.44). Since the rank of a Kronecker product is given by the product of the ranks of the factors, (4.44) has a $(d_1 + d_2 - 1)$ or $(2d_1 + 2d_2 - 4)$-dimensional null space for first and second order random walks, respectively. Therefore, the null space usually is high-dimensional and contains arbitrary row and column effects (Besag & Higdon 1999) but no constant effect. A useful application of Kronecker product priors may be the modeling of interaction effects, e. g. time-space interactions based on the Kronecker product of a spatial and a temporal penalty matrix. In this case, the invariance with respect to the addition of arbitrary time and space effects has a reasonable interpretation, if temporal and spatial main effects are included.

4.2.6.6 Comparison

When comparing the different bivariate random walks discussed in the previous paragraphs, Kronecker sums of univariate random walks and the approximation to the biharmonic differential operator are the most promising alternatives to define priors for the parameters of tensor product splines. Both the local quadratic fit and the Kronecker product of penalty matrices lead to problems which make them unsuitable in the present setting. An empirical comparison of bivariate random walks will be performed in a simulation study in Section 8.

Considering differences between first and second order random walks, the latter have larger null spaces including different types of planes and are therefore able to better approximate effects that are in fact planes. Generally, second order random walks seem to be more appropriate when estimating smooth surfaces while first order random walks may be more suitable for wigglier functions. This hypothesis will also be investigated further in the simulation study in Section 8.

5 Inference

When performing Bayesian inference, all inferential conclusions are based on the posterior of the model. In an empirical Bayes approach to structured additive regression, no hyperpriors are assigned to the hyperparameters, i. e. the variances τ_j^2 are treated as fixed. In this case, the specific form of the posterior only depends on the parameterization of the regression terms in the model. If we choose the original parameterization in terms of the ξ_j as discussed in the previous sections, the posterior is given by

$$p(\xi_1, \ldots, \xi_p, \gamma, |y) \propto L(y, \xi_1, \ldots, \xi_p, \gamma) \prod_{j=1}^{p} p(\xi_j|\tau_j^2), \tag{5.1}$$

where $L(\cdot)$ denotes the likelihood which is the product of individual likelihood contributions defined by the exponential family density in (4.1) and $p(\xi_j|\tau_j^2)$ is the prior for regression coefficients ξ_j as given in (4.9). Note that posterior (5.1) has the form of a penalized likelihood, where the penalty terms equal the prior distributions of the regression coefficients. Hence, empirical Bayes estimates derived by maximizing (5.1) can also be considered penalized likelihood estimates.

Since the variances shall also be estimated from the data, an appropriate estimation procedure has to be provided. In principle, methodology developed for mixed models could be used, since from a frequentist perspective, prior (4.9) merely defines ξ_j to be a random effect with a correlated random effects distribution. However, the fact that some of the priors are improper prevents direct usage of mixed model estimation techniques. An increasingly popular idea to solve this problem is to reparametrize models with penalties as mixed models with i. i. d. random effects. This approach goes back to Green (1987) for smoothing splines and has been used in a variety of settings throughout the last years. Lin & Zhang (1999) consider models for longitudinal data consisting of nonparametric effects modeled by smoothing splines and random effects to account for correlations caused by the longitudinal data structure. Zhang (2004) further extends this model and includes varying coefficient terms with smoothing splines as effect modifiers. Mixed model approaches to penalized splines based on truncated power series representations have been described in Wand (2003) and Ruppert et al. (2003). Kammann & Wand (2003) added spatial effects to these models using (low rank) Kriging terms. Penalized splines with B-spline bases were considered in Currie & Durbán (2002).

In either case, the mixed model representation yields a variance components model, and techniques for estimating the variance parameters are yet available or can be easily adapted. Probably the most common approach is to employ restricted maximum likelihood, also termed marginal likelihood in the literature. Kauermann (2004) provides some theoretical results on (restricted) maximum likelihood estimates for smoothing parameters in penalized spline models.

Before addressing the estimation of regression coefficients and variance parameters in a mixed model in detail, we will first show how to rewrite structured additive regression models as variance components models.

5.1 Mixed model representation

In order to reformulate structured additive regression models as mixed models, we first take a closer look on the general prior (4.9) for the regression coefficients ξ_j. This prior specifies a multivariate Gaussian distribution. However, in most cases the precision matrix K_j is rank deficient rendering (4.9) an improper distribution. Assuming that K_j is known and does not depend on further parameters to be estimated, we can reexpress ξ_j via a one-to-one transformation in terms of a parameter vector β_j with flat prior and a parameter vector b_j with i. i. d. Gaussian prior. While β_j captures the part of a function f_j that is not penalized by K_j, b_j captures the orthogonal deviation from this unpenalized part. The dimensions of both vectors depend on the rank of the penalty matrix K_j. If K_j has full rank, the unpenalized part vanishes completely and by choosing $b_j = K_j^{1/2}\xi_j$ we immediately obtain $b_j \sim N(0, \tau_j^2 I)$.

For the general case of rank deficient K_j, things are somewhat more complicated. If we assume that the j-th parameter vector has dimension d_j and the corresponding penalty matrix has rank k_j the decomposition of ξ_j into a penalized and an unpenalized part is of the form

$$\xi_j = \tilde{X}_j\beta_j + \tilde{Z}_j b_j \tag{5.2}$$

with a $d_j \times (d_j - k_j)$ matrix $\tilde{X}_j$ and a $d_j \times k_j$ matrix $\tilde{Z}_j$.

Requirements for decomposition (5.2) are:

(i) The composed matrix $(\tilde{X}_j\ \tilde{Z}_j)$ has full rank to make the transformation in (5.2) a one-to-one transformation. This also implies that both $\tilde{X}_j$ and $\tilde{Z}_j$ have full column rank.

(ii) $\tilde{X}_j$ and $\tilde{Z}_j$ are orthogonal, i. e. $\tilde{X}_j'\tilde{Z}_j = 0$.

(iii) $\tilde{X}_j'K_j\tilde{X}_j = 0$, resulting in β_j being unpenalized by K_j.

(iv) $\tilde{Z}_j'K_j\tilde{Z}_j = I_{k_j}$, resulting in an i. i. d. Gaussian prior for b_j.

In general the matrices defining (5.2) can be set up as follows: Establish $\tilde{X}_j$ as a $(d_j - k_j)$-dimensional basis of the null space of K_j. As a consequence, requirement (iii) is automatically fulfilled. The matrix $\tilde{Z}_j$ can be constructed as $\tilde{Z}_j = L_j(L_j'L_j)^{-1}$, where the full column rank $d_j \times k_j$ matrix L_j is determined by the factorization of the penalty matrix K_j into $K_j = L_jL_j'$. This ensures requirements (i) and (iv). If we furthermore choose L_j such that $L_j'\tilde{X}_j = 0$ and $\tilde{X}_jL_j' = 0$ hold, requirement (ii) is satisfied too. The factorization of the penalty matrix can be based on the spectral decomposition $K_j = \Gamma_j\Omega_j\Gamma_j'$, where the $k_j \times k_j$ diagonal matrix Ω_j contains the positive eigenvalues ω_{jm}, $m = 1, \ldots, k_j$, of K_j in descending order, i. e. $\Omega_j = \text{diag}(\omega_{j1}, \ldots, \omega_{jk_j})$ and the $d_j \times k_j$ orthogonal matrix Γ_j is formed of the corresponding eigenvectors. From the spectral decomposition we can choose $L_j = \Gamma_j\Omega_j^{1/2}$.

Note that the factor L_j is not unique and in many cases numerically superior factorizations exist. For instance, for P-splines a more favorable choice for L_j is given by $L_j = D'$, where D is the difference matrix used to construct the penalty matrix. For random walks a weighted version of the difference matrix can be used, i. e. $L_j = D'W^{\frac{1}{2}}$, and for seasonal

effects again the factor $L_j = D'$, where D is defined in Section 4.2.2.4, is an alternative choice.

From the general prior (4.9) for ξ_j and decomposition (5.2), it follows that $p(\beta_j) \propto const$ and

$$b_j \sim N(0, \tau_j^2 I_{k_j}), \tag{5.3}$$

since

$$\frac{1}{\tau_j^2}\xi_j' K_j \xi_j = \frac{1}{\tau_j^2} b_j' b_j.$$

This is a special version of the general decomposition result discussed for IGMRFs in Section 4.2.

Defining the vectors $x_{ij}' = v_{ij}'\tilde{X}_j$ and $z_{ij}' = v_{ij}'\tilde{Z}_j$ allows us to rewrite the predictor (4.7) as

$$\begin{aligned}
\eta_i &= \sum_{j=1}^{p} v_{ij}'\xi_j + u_i'\gamma \\
&= \sum_{j=1}^{p} (v_{ij}'\tilde{X}_j\beta_j + v_{ij}'\tilde{Z}_j b_j) + u_i'\gamma \\
&= \sum_{j=1}^{p} (x_{ij}'\beta_j + z_{ij}' b_j) + u_i'\gamma \\
&= x_i'\beta + z_i' b.
\end{aligned}$$

The design vector z_i and the vector b are composed of the vectors z_{ij} and the vectors b_j, respectively. More specifically, we obtain $z_i = (z_{i1}', z_{i2}', \ldots, z_{ip}')'$ and the stacked vector $b = (b_1', \ldots, b_p')'$. Similarly, the vector x_i and the vector β are given by $x_i = (x_{i1}', x_{i2}', \ldots, x_{ip}', u_i')'$ and $\beta = (\beta_1', \ldots, \beta_p', \gamma')'$. In matrix notation, proceeding in a similar way yields

$$\begin{aligned}
\eta &= \sum_{j=1}^{p} V_j\xi_j + U\gamma \\
&= \sum_{j=1}^{p} (X_j\beta_j + Z_j b_j) + U\gamma \\
&= X\beta + Zb,
\end{aligned} \tag{5.4}$$

where the matrices X and Z are composed in complete analogy as in the vector-based presentation.

Finally, we obtain a GLMM with fixed effects β and random effects $b \sim N(0, Q)$ where $Q = \text{blockdiag}(\tau_1^2 I_{k_1}, \ldots, \tau_p^2 I_{k_p})$. Hence, we can utilize GLMM methodology for simultaneous estimation of the functions f_j and the variance parameters τ_j^2, see the subsequent sections. Due to the flat prior of β, posterior (5.1) transforms to

$$p(\beta, b|y) \propto L(y, \beta, b) \exp\left(-\frac{1}{2} b' Q^{-1} b\right) \tag{5.5}$$

and the log-posterior is given by

$$l_p(\beta, b|y) = l(y, \beta, b) - \sum_{j=1}^{p} \frac{1}{2\tau_j^2} b_j' b_j, \tag{5.6}$$

where $l(y, \beta, b)$ denotes the log-likelihood corresponding to $L(y, \beta, b)$.

The decomposition of ξ_j also leads to a similar decomposition for $f_j(\nu_{ij})$ into a penalized and an unpenalized part:

$$\begin{aligned} f_j(\nu_{ij}) &= v_{ij}' \tilde{X}_j \beta_j + v_{ij}' \tilde{Z}_j b_j \\ &= x_{ij}' \beta_j + z_{ij}' b_j. \end{aligned} \tag{5.7}$$

This decomposition will form the basis for the construction of pointwise credible intervals for $f_j(\nu_{ij})$ (see Section 5.2.1).

The mixed model representation furthermore allows for a different perspective on the identification problem inherent to nonparametric regression models. For each of the model components with improper prior (except varying coefficient terms), the matrix X_j representing the deterministic part of f_j contains a column of ones corresponding to the mean level of the respective function. Provided that there is at least one such term and that we have an intercept included in the model, linear dependencies in the design matrix X of fixed effects occur. To get around this, we delete all the vectors of ones except for the intercept which has a similar effect as centering the functions $f_j(\nu_{ij})$.

A similar problem occurs in models with varying coefficient terms, if a covariate, u say, is included both in a parametric way and as interaction variable of a varying coefficient term, i. e. if a model with a predictor of the form

$$\eta = \ldots + u \cdot \gamma + u \cdot f(w) + \ldots$$

shall be estimated. In this case, γ is not identifiable since the design matrix for the unpenalized part resulting from the decomposition of $f(w)$ contains a column of ones. This column of ones is multiplied by u and, thus, u is represented twice in the overall design matrix of the fixed effects. This in turn results in a singular coefficient matrix when estimating the regression coefficients. As a consequence, covariates have to be considered either as parametric effects or as interaction variables and not as both.

For models with interaction surfaces, there are also modeling restrictions that have to be respected to ensure identifiability. If we consider a regression model of the form

$$\eta = \ldots + f_1(u) + f_2(w) + f_{1,2}(u, w) + \ldots,$$

where f_1 and f_2 shall be modeled as univariate penalized splines and $f_{1,2}$ is to be included as a bivariate P-spline, we already know from the above considerations, that the mean levels of these functions are not identifiable. Moreover, if second order random walks are employed as priors for both the univariate terms and the interaction term, further linear dependencies in X are introduced. Either with a bivariate second order random walk based on a Kronecker sum or an approximation to the biharmonic differential operator (see Sections 4.2.6.2 and 4.2.6.4), the design matrix X contains columns which are

multiples of the original covariate vectors u and w. The same observation holds for the reparametrizations of the univariate terms. Hence, linear dependencies caused by these vectors are observed and the resulting model is not identifiable. As a consequence, care has to be taken to choose a suitable formulation for the estimation problem at hand. However, the mixed model approach has the advantage, that identifiability problems can be diagnosed from the rank of the design matrix X, while in fully Bayesian inference based on MCMC, they are not necessarily recognizable from the output of the estimation procedure.

5.2 Estimation of regression coefficients

To derive empirical Bayes or posterior mode estimates of the regression coefficients, the posterior (5.5) or equivalently the log-posterior (5.6) have to be maximized with respect to β and b given the variances. In analogy to the estimation in generalized linear models, this maximization can be carried out utilizing a Fisher-Scoring algorithm requiring the score function and the expected Fisher-information matrix. Since (5.6) has the form of a penalized log-likelihood whose penalty depends only on the random effects b, it is useful to look at the derivatives for β and b separately (Fahrmeir & Tutz 2001, pp. 298/99). For the score function, we obtain

$$s(\beta, b) = \frac{\partial l_p(\beta, b|y)}{\partial(\beta, b)} = \begin{pmatrix} s_\beta(\beta, b) \\ s_b(\beta, b) \end{pmatrix}$$

with

$$s_\beta(\beta, b) = \frac{\partial l_p(\beta, b|y)}{\partial \beta} = X'DS^{-1}(y - \mu) \tag{5.8}$$

and

$$s_b(\beta, b) = \frac{\partial l_p(\beta, b|y)}{\partial b} = Z'DS^{-1}(y - \mu) - Q^{-1}b. \tag{5.9}$$

Just like in ordinary generalized linear models, D and S are given by the derivative of the response function with respect to the predictor and the variance of the response, respectively:

$$D = \text{diag}(D_i) = \text{diag}\left(\frac{\partial h(\eta_i)}{\partial \eta}\right) \tag{5.10}$$

and

$$S = \text{var}(y|\beta, b) = \text{diag}(\sigma_i^2) = \text{diag}(\phi v(\mu_i)/\omega_i). \tag{5.11}$$

Here, $v(\mu)$ denotes the variance function that is determined by the exponential family the response variable belongs to, ϕ is the scale parameter of the exponential family and the ω_i are positive weights (compare Section 4.1.1).

In a similar way, the expected Fisher information is decomposed as follows:

$$F(\beta, b) = \begin{pmatrix} F_{\beta\beta}(\beta, b) & F_{\beta b}(\beta, b) \\ F_{b\beta}(\beta, b) & F_{bb}(\beta, b) \end{pmatrix},$$

where

$$\begin{aligned} F_{\beta\beta}(\beta, b) &= X'DS^{-1}DX, \\ F_{\beta b}(\beta, b) &= F_{b\beta}(\beta, b)' = X'DS^{-1}DZ, \\ F_{bb}(\beta, b) &= Z'DS^{-1}DZ + Q^{-1}. \end{aligned}$$

Now, the regression coefficients can be estimated by iterating

$$\begin{pmatrix}\hat{\beta}^{(k+1)}\\ \hat{b}^{(k+1)}\end{pmatrix} = \begin{pmatrix}\hat{\beta}^{(k)}\\ \hat{b}^{(k)}\end{pmatrix} + (F^{(k)})^{-1}s^{(k)}, \tag{5.12}$$

beginning with some starting values $(\beta^{(0)'}, b^{(0)'})'$.

Defining the working observations

$$\tilde{y} = X\hat{\beta}^{(k)} + Z\hat{b}^{(k)} + D^{-1}(y-\mu) \tag{5.13}$$

leads to the equivalent estimation process of iteratively solving the linear system of equations

$$\begin{pmatrix}X'WX & X'WZ\\ Z'WX & Z'WZ + Q^{-1}\end{pmatrix}\begin{pmatrix}\hat{\beta}^{(k+1)}\\ \hat{b}^{(k+1)}\end{pmatrix} = \begin{pmatrix}X'W\tilde{y}\\ Z'W\tilde{y}\end{pmatrix} \tag{5.14}$$

with working weights

$$W = \mathrm{diag}(w_i) = DS^{-1}D. \tag{5.15}$$

The latter procedure is called iteratively weighted least squares since it has the form of weighted least squares estimation except that the weights change in every iteration. For the natural link function discussed in Section 4.1.1 the formula for the weights simplifies to $W = D = S$.

Note that iteratively solving (5.14) is equivalent to approximating the likelihood $L(y, \beta, b)$ with the likelihood of a multivariate Gaussian distribution having an iteratively reweighted covariance matrix W^{-1}. This approximation will also be used to obtain estimates of the variance parameters in Section 5.3.

5.2.1 Construction of credible intervals and credible bands

Formula (5.14) also serves as a basis for constructing credible intervals of the function estimates $\hat{f}_j$ (see Lin & Zhang 1999). If we denote the coefficient matrix on the left hand side of (5.14) by H, the approximate covariance matrix of the regression coefficients is given by H^{-1}. Since $\hat{f}_j = X_j\hat{\beta}_j + Z_j\hat{b}_j$ is a linear combination of the regression coefficients, the covariance matrix of $\hat{f}_j$ is given by

$$\mathrm{cov}(\hat{f}_j) = (X_j, Z_j)\,\mathrm{cov}\left(\hat{\beta}_j', \hat{b}_j'\right)(X_j, Z_j)', \tag{5.16}$$

where $\mathrm{cov}\left(\hat{\beta}_j', \hat{b}_j'\right)$ can be taken from the corresponding blocks in H^{-1}. Assuming approximate normality of the estimated regression coefficients allows to construct pointwise credible intervals based on the diagonal elements in (5.16). Suggestions on the construction of simultaneous confidence bands, either based on simulation techniques or analytic approximations can be found in Ruppert et al. (2003, Ch. 6.5).

5.3 Marginal likelihood for variance components

In this section, we turn to the estimation of the variance parameters $\tau^2 = (\tau_1^2, \ldots, \tau_p^2)'$ (and the dispersion parameter ϕ if needed). To motivate the estimation via (approximate)

restricted maximum likelihood we first explain the estimation of variance parameters for Gaussian response (see also Verbeke & Molenberghs (2000, Ch. 5) or McCulloch & Searle (2001, Ch. 6)). Then we generalize the results to nonnormal data. In the beginning we will more or less adopt the frequentist perspective on mixed models since methodology for mixed models is usually derived in this context. However, restricted maximum likelihood estimation also offers a nice Bayesian interpretation.

5.3.1 Maximum likelihood estimation

In linear mixed models, estimation of variance parameters is often based on maximum likelihood (ML). These ML estimates are usually obtained from the marginal distribution of y after having integrated out the random effects b:

$$y \sim N(X\beta, \Sigma), \tag{5.17}$$

where

$$\Sigma = \sigma^2 I + ZQZ'$$

is the marginal covariance matrix of y. Note that the consideration of (5.17) as a marginal distribution depends on the frequentist perspective that only b is a random vector while β is assumed to be deterministic. From a Bayesian perspective, this distinction is not reasonable, since both vectors of regression coefficients are assumed to be random but follow different types of priors. In the following we will see how these different views of the model influence the estimation of the variance parameters. The marginal formulation is preferred since distribution (5.17) and its density explicitly depend on Q and therefore allow the determination of derivatives with respect to τ^2. This does not hold for the conditional distribution of $y|b$ which is given by

$$y|b \sim N(X\beta + Zb, \sigma^2 I).$$

For Gaussian response, the maximum likelihood estimator of β can be derived analytically from the marginal distribution (5.17) yielding

$$\hat{\beta} = (X'\Sigma^{-1}X)^{-1}X'\Sigma^{-1}y.$$

Plugging this expression into the likelihood (the density of the marginal distribution of y) yields the profile likelihood for τ^2 and σ^2 which may be maximized numerically using the EM-algorithm or Fisher-Scoring (compare Harville (1977) for details). However, maximum likelihood estimation does not take into account the loss of degrees of freedom due to the estimation of β and, as a consequence, the obtained estimators are usually biased towards zero. A way to overcome or at least to mitigate this bias is the usage of restricted maximum likelihood (REML) introduced by Patterson & Thompson (1971).

5.3.2 Restricted maximum likelihood estimation

In this case, estimation is based on the likelihood of some error contrasts $u = A'y$ rather than on the likelihood of y and also fits better in the Bayesian model formulation that

was adopted in the previous sections. An error contrast is defined as a linear combination $a'y$ with expectation zero, causing the distribution of $a'y$ not to depend on β. An example for a set of such error contrasts are the residuals $\hat{\varepsilon}_i = y_i - x_i'\hat{\beta}$ where the vector a consists of the ith row of the residual matrix $R = I - X(X'X)^{-1}X'$. Consequently, the vector of all residuals $\hat{\varepsilon}$ has the distribution

$$\hat{\varepsilon} \sim N(0, R\sigma^2)$$

with the desired property $E(\hat{\varepsilon}) = 0$. However, since R is inherently singular, the distribution is partially improper and therefore the usage of the residuals is not advisable in general.

Since only $n - \dim(\beta)$ linear independent error contrasts exist, REML estimation is commonly based on error contrasts obtained from the decomposition

$$AA' = X(X'X)^{-1}X' \text{ with } A'A = I, \tag{5.18}$$

where A is an $n \times (n - \dim(\beta))$ matrix with full column rank. It is easy to show, that the resulting error contrasts $u = A'y$ fulfill $E(u) = 0$.

Now, the marginal density of u is given by (Harville 1974)

$$p(u) = \left(\frac{1}{2\pi}\right)^{\frac{n-\dim(\beta)}{2}} |X'X|^{\frac{1}{2}}|\Sigma|^{-\frac{1}{2}}|X'\Sigma^{-1}X|^{-\frac{1}{2}} \exp\left[-\frac{1}{2}(y - X\hat{\beta})'\Sigma^{-1}(y - X\hat{\beta})\right] \tag{5.19}$$

and restricted maximum likelihood estimators of τ^2 and σ^2 are obtained by maximizing

$$l^*(\tau^2, \sigma^2) = -\frac{1}{2}\log(|\Sigma|) - \frac{1}{2}\log(|X'\Sigma^{-1}X|) - \frac{1}{2}(y - X\hat{\beta})'\Sigma^{-1}(y - X\hat{\beta})$$

using some numerical technique. Note that the restricted log-likelihood does not depend on the special choice of A as long as $n - \dim(\beta)$ linear independent error contrasts are used (Verbeke & Molenberghs 2000, Sec. 5.3).

Generalization of the restricted maximum likelihood technique to nonnormal data is not straightforward, since the definition of error contrasts is not possible for more general responses due to the nonlinear dependency of y on β in generalized linear mixed models. Therefore, we present an alternative approach by Harville (1974) to the estimation of variances in Gaussian mixed models that leads to exactly the same restricted log-likelihood for τ^2 and σ^2 and that can additionally be extended to more general responses.

Recall that in the Bayesian formulation of mixed models not only b is assumed to be a random variable but also β. While the prior distribution of b is proper, the distribution of β is flat, i. e.

$$p(\beta) \propto const.$$

From the Bayesian perspective it seems reasonable, to integrate both b and β out of the distribution of y. The resulting marginal distribution of y (as regards to b and β) now has to be maximized with respect to the variance parameters. Harville (1974) showed that proceeding in this way leads to exactly the same likelihood as above. For this reason, REML estimation is also sometimes referred to as marginal likelihood estimation in the literature.

Replacing ML variance estimates with their REML counterparts allows for a further interpretation: The ML estimators are obtained by jointly maximizing the posterior with respect to the regression coefficients β and the variances τ^2. Then the ML estimators correspond to the variance components of the posterior mode. In contrast, REML estimates are given by the mode of the marginal posterior for the variances. The latter strategy coincides with the usual strategy in an empirical Bayes approach, where hyperparameters are treated as fixed constants which have to be estimated from their marginal posterior.

To derive REML estimates for GLMMs, we approximate the logarithm of the likelihood with the Pearson χ^2-statistic, yielding

$$\begin{aligned} l(y, \beta, b) &\approx \sum_{i=1}^{n} \frac{(y_i - \mu_i)^2}{\omega_i v(\mu_i)/\phi} \\ &= (y - \mu)' S^{-1} (y - \mu). \end{aligned}$$

This is in fact equivalent to the Laplace approximation of $l(y, \beta, b)$ with a quadratic function (compare Tierney & Kadane 1986). Using the definition of the working observations $\tilde{y}$ in (5.13) gives

$$(y - \mu) = D(\tilde{y} - X\beta - Zb)$$

and therefore we have

$$\begin{aligned} l(y, \beta, b) &\approx (\tilde{y} - X\beta - Zb)' D' S^{-1} D (\tilde{y} - X\beta - Zb) \\ &= (\tilde{y} - X\beta - Zb)' W (\tilde{y} - X\beta - Zb). \end{aligned}$$

Ignoring the dependence of W on the variance parameters (Breslow & Clayton 1993) gives rise to the fact that the likelihood can be approximated by the log-likelihood of a linear mixed model for the working observations $\tilde{y}$. To be more specific, we assume

$$\tilde{y} | \beta, b \overset{a}{\sim} N(X\beta + Zb, W^{-1}).$$

The determination of the marginal distribution of $\tilde{y}$ (as regards to b and β) yields an approximate restricted log-likelihood for the generalized linear mixed model:

$$l^*(\tau^2, \phi) = -\frac{1}{2} \log(|\Sigma|) - \frac{1}{2} \log(|X' \Sigma^{-1} X|) - \frac{1}{2} (\tilde{y} - X\hat{\beta})' \Sigma^{-1} (\tilde{y} - X\hat{\beta}) \tag{5.20}$$

with $\Sigma = W^{-1} + ZQZ'$ being an approximation to the marginal covariance matrix of $\tilde{y}$.

To finally obtain REML-estimates of the variance parameters, we have to maximize (5.20) with respect to τ^2 (and ϕ if necessary). One suitable optimization procedure is the Newton-Raphson algorithm which is based on the first and second derivative of $l^*(\tau^2, \phi)$ with respect to the variance parameters. A modification of the Newton-Raphson algorithm is given by Fisher-Scoring, where the second derivative is replaced by its expectation. Since this leads to simplified estimation equations, we will focus on Fisher-Scoring.

Note that in several models the derivatives with respect to the dispersion parameter ϕ are not needed, e. g. for Poisson or Binomial data where the dispersion parameter is fixed. In these cases the corresponding derivatives have to be eliminated from the formulae for the score-function and the expected Fisher-information presented in the following sections.

5.3.3 Numerical details: Score function

The score-function $s^*(\tau^2, \phi)$ is a $(p+1)$-dimensional vector containing the derivatives of (5.20) with respect to τ^2 and ϕ:

$$\begin{aligned} s^*(\tau^2, \phi) &= (s_1^*, \ldots, s_{p+1}^*)' \\ &= \left(\frac{\partial l^*(\tau^2, \phi)}{\partial \tau_1^2}, \ldots, \frac{\partial l^*(\tau^2, \phi)}{\partial \tau_p^2}, \frac{\partial l^*(\tau^2, \phi)}{\partial \phi} \right)'. \end{aligned}$$

The first p elements of this score-function are given by

$$s_j^* = -\frac{1}{2} \operatorname{tr} \left(P \frac{\partial \Sigma}{\partial \tau_j^2} \right) + \frac{1}{2} (\tilde{y} - X\hat{\beta})' \Sigma^{-1} \frac{\partial \Sigma}{\partial \tau_j^2} \Sigma^{-1} (\tilde{y} - X\hat{\beta}) \tag{5.21}$$

with

$$P = \Sigma^{-1} - \Sigma^{-1} X (X' \Sigma^{-1} X)^{-1} X' \Sigma^{-1}, \tag{5.22}$$

see Harville (1977).

The crucial point is, that formulae (5.21) and (5.22) are inapplicable for data sets with more than about $n = 3000$ observations, since they involve the computation and manipulation of several $n \times n$ matrices including P and Σ. In particular, the determination of Σ^{-1}, which requires $\mathcal{O}(n^3)$ computations, is almost impractical if n is large. In addition, n^2 storing locations are needed for each $n \times n$ matrix, resulting in an enormous amount of total memory that is required to compute the score function.

The inversion of Σ may be avoided using some matrix identities derived by Lin & Zhang (1999), yielding

$$s_j^* = -\frac{1}{2} \operatorname{tr} \left(P \frac{\partial \Sigma}{\partial \tau_j^2} \right) + \frac{1}{2} (\tilde{y} - X\hat{\beta} - Z\hat{b})' W \frac{\partial \Sigma}{\partial \tau_j^2} W (\tilde{y} - X\hat{\beta} - Z\hat{b}) \tag{5.23}$$

and

$$P = W - W(X \; Z) H^{-1} (X \; Z)' W \tag{5.24}$$

with

$$H = \begin{pmatrix} X'WX & X'WZ \\ Z'WX & Z'WZ + Q^{-1} \end{pmatrix}.$$

Given (5.23), the involved derivatives are of the form

$$\frac{\partial \Sigma}{\partial \tau_j^2} = Z \frac{\partial Q}{\partial \tau_j^2} Z' \tag{5.25}$$

which reduces to

$$\frac{\partial \Sigma}{\partial \tau_j^2} = Z \frac{\partial Q}{\partial \tau_j^2} Z' = Z_j Z_j'$$

in a variance components model as in (5.4) with $Q = \text{blockdiag}(\tau_1^2 I, \ldots, \tau_p^2 I)$.

Plugging derivative (5.25) into formula (5.23) yields

$$s_j^* = -\frac{1}{2}\operatorname{tr}\left(PZ\frac{\partial Q}{\partial \tau_j^2}Z'\right) + \frac{1}{2}(\tilde{y} - X\hat{\beta} - Z\hat{b})'WZ\frac{\partial Q}{\partial \tau_j^2}Z'W(\tilde{y} - X\hat{\beta} - Z\hat{b}). \tag{5.26}$$

Though avoiding the inversion of Σ, the computation of (5.26) still involves the determination and multiplication of the $n \times n$ matrix P, and so the problems described above remain essentially unchanged. Using an elementary property of the trace in combination with the alternative definition of P in (5.24) we can further simplify the score function:

$$\begin{aligned} -\frac{1}{2}\operatorname{tr}\left(PZ\frac{\partial Q}{\partial \tau_j^2}Z'\right) &= -\frac{1}{2}\operatorname{tr}\left(WZ\frac{\partial Q}{\partial \tau_j^2}Z'\right) + \frac{1}{2}\operatorname{tr}\left(W(X\ Z)H^{-1}(X\ Z)'WZ\frac{\partial Q}{\partial \tau_j^2}Z'\right) \\ &= -\frac{1}{2}\operatorname{tr}\left(Z'WZ\frac{\partial Q}{\partial \tau_j^2}\right) + \frac{1}{2}\operatorname{tr}\left(Z'W(X\ Z)H^{-1}(X\ Z)'WZ\frac{\partial Q}{\partial \tau_j^2}\right) \end{aligned} \tag{5.27}$$

Now the matrices inside the traces are no longer of dimension $n \times n$ but are reduced to dimension $\dim(b) \times \dim(b)$. Moreover, most of the matrix products do not have to be evaluated explicitly since they can be derived at low computational cost from the matrix H which is involved in the determination of $\hat{\beta}$ and $\hat{b}$ (compare Equation (5.14)). The matrix of sums of squares an crossproducts (SSCP)

$$\begin{pmatrix} X'WX & X'WZ \\ Z'WX & Z'WZ \end{pmatrix} \tag{5.28}$$

can be computed from H by simply subtracting Q^{-1} from the lower right corner. Then most of the matrices needed to evaluate the traces in (5.27) turn out to be submatrices of the SSCP-matrix (5.28). Only H^{-1} and $\partial Q/\partial \tau_j$ have to be computed in addition. In this way the main computational burden is shifted from the computation and manipulation of $n \times n$ matrices to the inversion of the matrix H. Accordingly, the storage requirements are reduced from order n^2 to $(\dim(\beta) + \dim(b))^2$.

Assuming the special structure of a variance components model yields even simpler formulae. In this case (5.27) reduces to

$$-\frac{1}{2}\operatorname{tr}\left(PZ\frac{\partial Q}{\partial \tau_j^2}Z'\right) = -\frac{1}{2}\operatorname{tr}\left(Z_j'WZ_j\right) + \frac{1}{2}\operatorname{tr}\left(Z_j'W(X\ Z)H^{-1}(X\ Z)'WZ_j\right)$$

and the second part of (5.26) becomes

$$\frac{1}{2}(\tilde{y} - X\hat{\beta} - Z\hat{b})'WZ_jZ_j'W(\tilde{y} - X\hat{\beta} - Z\hat{b}).$$

It is important not to determine Z_jZ_j' explicitly as this matrix product has dimension $n \times n$. Instead it is advantageous (and computationally favorable) to compute the squared norm of the vector $(\tilde{y} - X\hat{\beta} - Z\hat{b})'WZ_j$.

If an additional dispersion parameter is present, the last entry in the score vector is given by

$$s_{p+1}^* = -\frac{1}{2}\operatorname{tr}\left(P\frac{\partial \Sigma}{\partial \phi}\right) + \frac{1}{2}(\tilde{y} - X\hat{\beta})'\Sigma^{-1}\frac{\partial \Sigma}{\partial \phi}\Sigma^{-1}(\tilde{y} - X\hat{\beta}). \tag{5.29}$$

Concerning the derivative of Σ with respect to the dispersion parameter ϕ, we get

$$\begin{aligned}\frac{\partial \Sigma}{\partial \phi} &= \frac{\partial W^{-1}}{\partial \phi} \\ &= D^{-1}\frac{\partial S}{\partial \phi}D^{-1} \\ &= D^{-1}\,\mathrm{diag}\left(\frac{\partial \phi v(\mu_i)/\omega_i}{\partial \phi}\right)D^{-1} \\ &= D^{-1}\,\mathrm{diag}\left(v(\mu_i)/\omega_i\right)D^{-1} \\ &= \frac{1}{\phi}W^{-1}. \end{aligned} \tag{5.30}$$

Plugging this derivative into (5.29) and using the matrix identities as described above yields

$$s^*_{p+1} = -\frac{1}{2\phi}\,\mathrm{tr}\left(PW^{-1}\right) + \frac{1}{2\phi}(\tilde{y} - X\hat{\beta} - Z\hat{b})'WW^{-1}W(\tilde{y} - X\hat{\beta} - Z\hat{b}). \tag{5.31}$$

The first part of (5.31) may again be simplified by inserting the expression for P in (5.24) giving

$$\begin{aligned}-\frac{1}{2\phi}\,\mathrm{tr}\left(PW^{-1}\right) &= -\frac{1}{2\phi}\,\mathrm{tr}\left(WW^{-1}\right) + \frac{1}{2\phi}\,\mathrm{tr}\left(W(X\ Z)H^{-1}(X\ Z)'WW^{-1}\right) \\ &= -\frac{n}{2\phi} + \frac{1}{2\phi}\,\mathrm{tr}\left((X\ Z)'W(X\ Z)H^{-1}\right).\end{aligned}$$

Note again, that besides of H^{-1} no new matrices have to be computed to evaluate these expressions since $(X\ Z)'W(X\ Z)$ equals the SSCP-matrix (5.28).

The second part of (5.31) reduces to

$$\frac{1}{\phi}(\tilde{y} - X\hat{\beta} - Z\hat{b})'W(\tilde{y} - X\hat{\beta} - Z\hat{b}),$$

which differs only by the multiplication with the inverse scale parameter from the weighted sum of squared residuals.

5.3.4 Numerical details: Expected Fisher-information

The expected Fisher-information $F^*(\tau^2, \phi) = (F^*_{jk})$, $j, k = 1, \ldots, p+1$, is a $(p+1)\times(p+1)$-dimensional matrix with upper left $p \times p$-corner consisting of the negative expectations of the second derivatives $-E\left(\frac{\partial^2 l^*(\tau^2,\phi)}{\partial \tau_j^2 \partial \tau_k^2}\right)$, $j, k = 1, \ldots, p$. These are given by (Lin & Zhang 1999)

$$F^*_{jk} = -E\left(\frac{\partial^2 l^*(\tau^2, \phi)}{\partial \tau_j^2 \partial \tau_k^2}\right) = \frac{1}{2}\,\mathrm{tr}\left(P\frac{\partial \Sigma}{\partial \tau_j^2}P\frac{\partial \Sigma}{\partial \tau_k^2}\right). \tag{5.32}$$

Applying (5.24) and (5.25) gives rise to simpler formulae:

$$\begin{aligned} F^*_{jk} &= \frac{1}{2}\operatorname{tr}\left(WZ\frac{\partial Q}{\partial \tau_j^2}Z'WZ\frac{\partial Q}{\partial \tau_k^2}Z'\right) - \frac{1}{2}\operatorname{tr}\left(W(X\ Z)H^{-1}(X\ Z)'WZ\frac{\partial Q}{\partial \tau_j^2}Z'WZ\frac{\partial Q}{\partial \tau_k^2}Z'\right) \\ &\quad -\frac{1}{2}\operatorname{tr}\left(WZ\frac{\partial Q}{\partial \tau_j^2}Z'W(X\ Z)H^{-1}(X\ Z)'WZ\frac{\partial Q}{\partial \tau_k^2}Z'\right) \\ &\quad +\frac{1}{2}\operatorname{tr}\left(W(X\ Z)H^{-1}(X\ Z)'WZ\frac{\partial Q}{\partial \tau_j^2}Z'W(X\ Z)H^{-1}(X\ Z)'WZ\frac{\partial Q}{\partial \tau_k^2}Z'\right). \end{aligned} \tag{5.33}$$

Due to equality, only the second or the third term in (5.33) have to be evaluated. Shifting some matrices inside the traces allows to reexpress the term in sole dependence on H^{-1}, the derivatives of Q and submatrices of the SSCP-matrix (5.28):

$$\begin{aligned} F^*_{jk} &= \frac{1}{2}\operatorname{tr}\left(Z'WZ\frac{\partial Q}{\partial \tau_j^2}Z'WZ\frac{\partial Q}{\partial \tau_k^2}\right) - \operatorname{tr}\left(Z'W(X\ Z)H^{-1}(X\ Z)'WZ\frac{\partial Q}{\partial \tau_j^2}Z'WZ\frac{\partial Q}{\partial \tau_k^2}\right) \\ &\quad +\frac{1}{2}\operatorname{tr}\left(Z'W(X\ Z)H^{-1}(X\ Z)'WZ\frac{\partial Q}{\partial \tau_j^2}Z'W(X\ Z)H^{-1}(X\ Z)'WZ\frac{\partial Q}{\partial \tau_k^2}\right). \end{aligned}$$

Again, the largest matrix involved in the computation of F^*_{jk} is H^{-1}. Further efficacy can be achieved for the variance components model (5.4). Here the above expression reduces to

$$\begin{aligned} F^*_{jk}(\vartheta) &= \frac{1}{2}\operatorname{tr}(Z_k'WZ_jZ_j'WZ_k) - \operatorname{tr}(Z_k'W(X\ Z)H^{-1}(X\ Z)'WZ_jZ_j'WZ_k) \\ &\quad +\frac{1}{2}\operatorname{tr}(Z_k'W(X\ Z)H^{-1}(X\ Z)'WZ_jZ_j'W(X\ Z)H^{-1}(X\ Z)'WZ_k). \end{aligned}$$

Concerning the derivatives with respect to the dispersion parameter, the general form of is given by

$$F^*_{p+1,j} = F^*_{j,p+1} = \frac{1}{2}\operatorname{tr}\left(P\frac{\partial \Sigma}{\partial \tau_j^2}P\frac{\partial \Sigma}{\partial \phi}\right), \quad j = 1, \ldots, p, \tag{5.34}$$

and

$$F^*_{p+1,p+1} = \frac{1}{2}\operatorname{tr}\left(P\frac{\partial \Sigma}{\partial \phi}P\frac{\partial \Sigma}{\partial \phi}\right). \tag{5.35}$$

Using the derivatives in (5.25) and (5.30) together with the definition of P in (5.24), we

can restate (5.34) as

$$\begin{aligned}
F^*_{j,p+1} &= \frac{1}{2}\operatorname{tr}\left(PZ\frac{\partial Q}{\partial \tau_j^2}Z'P\frac{1}{\phi}W^{-1}\right) \\
&= \frac{1}{2\phi}\operatorname{tr}\left(WZ\frac{\partial Q}{\partial \tau_j^2}Z'WW^{-1}\right) - \frac{1}{2\phi}\operatorname{tr}\left(WZ\frac{\partial Q}{\partial \tau_j^2}Z'W(X\ Z)H^{-1}(X\ Z)'WW^{-1}\right) \\
&\quad -\frac{1}{2\phi}\operatorname{tr}\left(W(X\ Z)H^{-1}(X\ Z)'WZ\frac{\partial Q}{\partial \tau_j^2}Z'WW^{-1}\right) \\
&\quad +\frac{1}{2\phi}\operatorname{tr}\left(W(X\ Z)H^{-1}(X\ Z)'WZ\frac{\partial Q}{\partial \tau_j^2}Z'W(X\ Z)H^{-1}(X\ Z)'WW^{-1}\right) \\
&= \frac{1}{2\phi}\operatorname{tr}\left(Z'WZ\frac{\partial Q}{\partial \tau_j^2}\right) - \frac{1}{\phi}\operatorname{tr}\left((X\ Z)'WZ\frac{\partial Q}{\partial \tau_j^2}Z'W(X\ Z)H^{-1}\right) \\
&\quad +\frac{1}{2\phi}\operatorname{tr}\left((X\ Z)'W(X\ Z)H^{-1}(X\ Z)'WZ\frac{\partial Q}{\partial \tau_j^2}Z'W(X\ Z)\right).
\end{aligned}$$

In case of a variance components model we obtain

$$\begin{aligned}
F^*_{j,p+1} &= \frac{1}{2\phi}\operatorname{tr}\left(Z_j'WZ_j\right) - \frac{1}{\phi}\operatorname{tr}\left((X\ Z)'WZ_jZ_j'W(X\ Z)H^{-1}\right) \\
&\quad +\frac{1}{2\phi}\operatorname{tr}\left((X\ Z)'W(X\ Z)H^{-1}(X\ Z)'WZ_jZ_j'W(X\ Z)\right).
\end{aligned}$$

The expression in (5.35) can be rewritten as

$$\begin{aligned}
F^*_{p+1,p+1} &= \frac{1}{2}\operatorname{tr}\left(P\frac{\partial \Sigma}{\partial \phi}P\frac{\partial \Sigma}{\partial \phi}\right) \\
&= \frac{1}{2\phi^2}\operatorname{tr}\left(PW^{-1}PW^{-1}\right) \\
&= \frac{1}{2\phi^2}\operatorname{tr}\left(WW^{-1}WW^{-1}\right) - \frac{1}{2\phi^2}\operatorname{tr}\left(W(X\ Z)H^{-1}(X\ Z)'WW^{-1}WW^{-1}\right) \\
&\quad -\frac{1}{2\phi^2}\operatorname{tr}\left(WW^{-1}W(X\ Z)H^{-1}(X\ Z)'WW^{-1}\right) \\
&\quad +\frac{1}{2\phi^2}\operatorname{tr}\left(W(X\ Z)H^{-1}(X\ Z)'WW^{-1}W(X\ Z)H^{-1}(X\ Z)'WW^{-1}\right) \\
&= \frac{n}{2\phi^2} - \frac{1}{\phi^2}\operatorname{tr}\left((X\ Z)'W(X\ Z)H^{-1}\right) \\
&\quad +\frac{1}{2\phi^2}\operatorname{tr}\left((X\ Z)'W(X\ Z)H^{-1}(X\ Z)'W(X\ Z)H^{-1}\right).
\end{aligned}$$

5.4 Mixed model based inference in STAR

Having the estimation procedure as well as numerical improvements at hand, estimation of structured additive regression models based on mixed model methodology can be summarized in the following two steps:

1. Obtain updated estimates $\hat{\beta}$ and $\hat{b}$ given the current variance parameters as solutions of the system of equations

$$\begin{pmatrix} X'WX & X'WZ \\ Z'WX & Z'WZ + Q^{-1} \end{pmatrix} \begin{pmatrix} \beta \\ b \end{pmatrix} = \begin{pmatrix} X'W\tilde{y} \\ Z'W\tilde{y} \end{pmatrix}.$$

2. Estimates for the variance parameters $\vartheta = (\tau^2, \phi)$ are updated by

$$\vartheta^{(k+1)} = \vartheta^{(k)} + F^*(\vartheta^{(k)})^{-1} s^*(\vartheta^{(k)}).$$

The two estimation steps are performed in turn and iterated until convergence.

5.5 Inference based on MCMC

Since we will compare mixed model based empirical Bayes estimates with their fully Bayesian counterparts in several simulation studies in the following, we give a short overview over MCMC inference in structured additive regression. In this case, no reparametrization is needed and inference can be performed directly for the parameters $\gamma, \xi_1, \ldots, \xi_p$. A nice introductory text about MCMC inference is for example given in Green (2001). More details on MCMC inference in structured additive regression models can be found in Fahrmeir & Lang (2001a), Lang & Brezger (2004) and Brezger & Lang (2005).

In a fully Bayesian approach, parameter estimates are generated by drawing random samples from the posterior (5.1) via MCMC simulation techniques. The variance parameters τ_j^2 can be estimated simultaneously with the regression coefficients ξ_j by assigning additional hyperpriors to them. The most common assumption is, that the τ_j^2 are independently inverse gamma distributed, i. e. $\tau_j^2 \sim IG(a_j, b_j)$, with hyperparameters a_j and b_j specified a priori. A standard choice is to use $a_j = b_j = 0.001$. In some data situations (e. g. for small sample sizes), the estimated nonlinear functions f_j may depend considerably on the particular choice of hyperparameters. It is therefore good practice to estimate all models under consideration using a (small) number of different choices for a_j and b_j to assess the dependence of results on minor changes in the prior assumptions.

Suppose first that the distribution of the response variable is Gaussian, i. e. $y_i|\eta_i, \sigma^2 \sim N(\eta_i, \sigma^2/\omega_i)$, $i = 1, \ldots, n$ or $y|\eta, \sigma^2 \sim N(\eta, \sigma^2\Omega^{-1})$ where $\Omega = diag(\omega_1, \ldots, \omega_n)$ is a known weight matrix. In this case an additional hyperprior for the scale parameter σ^2 has to be specified. Similarly as for the variances of the regression coefficients, an inverse Gamma distribution $\sigma^2 \sim IG(a_0, b_0)$ is a convenient choice.

For Gaussian responses the full conditionals for fixed effects as well as nonlinear functions f_j are multivariate Gaussian. Thus, a Gibbs sampler can be used where posterior samples are drawn directly from the multivariate Gaussian distributions. The full conditional $\gamma|\cdot$ for fixed effects with diffuse priors is Gaussian with mean

$$E(\gamma|\cdot) = (U'\Omega U)^{-1} U'\Omega(y - \tilde{\eta}) \tag{5.36}$$

and covariance matrix

$$Cov(\gamma|\cdot) = \sigma^2 (U'\Omega U)^{-1}, \tag{5.37}$$

where U is the design matrix of fixed effects and $\tilde{\eta} = \eta - U\gamma$ is the part of the additive predictor associated with the other effects in the model. Similarly, the full conditional for the regression coefficients ξ_j of a function f_j is Gaussian with mean

$$m_j = E(\xi_j|\cdot) = \left(\frac{1}{\sigma^2}V_j'\Omega V_j + \frac{1}{\tau_j^2}K_j\right)^{-1}\frac{1}{\sigma^2}V_j'\Omega(y - \eta_{-j}) \tag{5.38}$$

and covariance matrix

$$P_j^{-1} = Cov(\xi_j|\cdot) = \left(\frac{1}{\sigma^2}V_j'\Omega V_j + \frac{1}{\tau_j^2}K_j\right)^{-1}, \tag{5.39}$$

where $\eta_{-j} = \eta - V_j\xi_j$. Although the full conditional is Gaussian, drawing random samples in an efficient way is not trivial, since linear equation systems with a high-dimensional precision matrix P_j must be solved in every iteration of the MCMC scheme. Following Rue (2001), drawing random numbers from $p(\xi_j|\cdot)$ can be conducted as follows: First the Cholesky decomposition $P_j = LL'$ is computed. Proceeding by solving $L'\xi_j = z$, where z is a vector of independent standard normal distributed random variables, yields $\xi_j \sim N(0, P_j^{-1})$. Afterwards, the mean m_j is computed by solving $P_j m_j = \frac{1}{\sigma^2}V_j'\Omega(y-\eta_{-j})$. This is achieved by first solving $L\nu = \frac{1}{\sigma^2}V_j'\Omega(y - \eta_{-j})$ by forward substitution followed by backward substitution $L'm_j = \nu$. Finally, adding m_j to the previously simulated ξ_j yields $\xi_j \sim N(m_j, P_j^{-1})$.

For all effects considered in Section 4.2 except GRFS, the posterior precision matrices P_j can be transferred into a band matrix like structure with bandsize depending on the prior. If f_j corresponds to a Markov random field, the posterior precision matrix is usually a sparse matrix but not a band matrix (compare Section 4.2.3.1). In this case the regions of a geographical map must be reordered, for example using the reverse Cuthill-McKee algorithm, to obtain a band matrix like precision matrix. Random samples from the full conditional can now be drawn in a very efficient way using Cholesky decompositions for band matrices or band matrix like matrices (see for example the envelope method for matrices with local bandwidths described in George & Liu 1981).

The full conditionals for the variance parameters τ_j^2, $j = 1, \ldots, p$, and σ^2 are all inverse Gamma distributions with parameters

$$a_j' = a_j + \frac{\text{rank}(K_j)}{2} \quad \text{and} \quad b_j' = b_j + \frac{1}{2}\xi_j'K_j\xi_j \tag{5.40}$$

for τ_j^2. For σ^2, we obtain

$$a_0' = a_0 + \frac{n}{2} \quad \text{and} \quad b_0' = b_0 + \frac{1}{2}(y-\eta)'(y-\eta). \tag{5.41}$$

If more general responses from an exponential family are given, a Metropolis-Hastings-algorithm based on iteratively weighted least squares (IWLS) proposals can be used. The idea of IWLS updates has been introduced by Gamerman (1997) for the estimation of generalized linear mixed models and adapted to the present situation of structured additive regression in Brezger & Lang (2005).

Suppose we want to update the regression coefficients ξ_j of the j-th function f_j with current value ξ_j^c of the chain. Then, according to IWLS, a new value ξ_j^p is proposed by drawing a random number from the multivariate Gaussian proposal distribution $q(\xi_j^c, \xi_j^p)$ with precision matrix and mean

$$P_j = V_j' W(\xi_j^c) V_j + \frac{1}{\tau_j^2} K_j \quad \text{and} \quad m_j = P_j^{-1} V_j' W(\xi_j^c)(\tilde{y} - \eta_{-j}). \tag{5.42}$$

The working weights and working observations W and $\tilde{y}$ are defined in complete analogy to the discussion in Section 5.2 and the vector $\eta_{-j} = \eta - V_j \beta_j$ is the part of the predictor associated with all remaining effects in the model. The proposed vector ξ_j^p is accepted as the new state of the chain with probability

$$\alpha(\xi_j^c, \xi_j^p) = \min\left(1, \frac{p(\xi_j^p|\cdot) q(\xi_j^p, \xi_j^c)}{p(\xi_j^c|\cdot) q(\xi_j^c, \xi_j^p)}\right)$$

where $p(\xi_j|\cdot)$ is the full conditional for ξ_j (i. e. the conditional distribution of ξ_j given all other parameters and the data y).

A fast implementation requires efficient sampling from the Gaussian proposal distributions. Using the same algorithms as for Gaussian responses implies that the number of calculations required to draw random numbers from the proposal distribution is linear in the number of parameters and observations. Also the computation of the acceptance probabilities is linear in the number of observations.

The full conditionals for the variance parameters τ_j^2 remain inverse gamma with parameters

$$a_j' = a_j + \frac{\text{rank}(K_j)}{2} \quad \text{and} \quad b_j' = b_j + \frac{1}{2} \xi_j' K_j \xi_j$$

and updating can be done by simple Gibbs steps, drawing random numbers directly from the inverse gamma densities.

Convergence of the Markov chains to their stationary distributions can be assessed by inspecting sampling paths and autocorrelation functions of sampled parameters. In the majority of cases, however, the IWLS updating scheme has excellent mixing properties and convergence problems do not occur.

From a theoretical point of view, fully Bayesian inference allows for inference in structured additive regression models avoiding the need of the Laplace approximation involved in mixed model based inference. Furthermore a frequently claimed advantage of full over empirical Bayes inference is that the variability caused by the estimation of hyperparameters is only considered appropriately by the former. However, in our experience this effect can be neglected in most situations, at least if the data contain enough information.

On the other hand, MCMC inference introduces hyperpriors for the variance parameters and no generally applicable rule for the choice of the hyperparameters is available. While the influence of these hyperparameters is usually small, it may become relevant in sparse data situations. Furthermore, mixing and convergence of several Markov chains have to be monitored when performing MCMC inference to achieve accurate sampling based approximations to the quantities of interest. Mostly this is achieved by a visual inspection

of sampling paths since no generally accepted measure is available to determine the actual convergence.

Although a theoretical comparison of empirical and fully Bayesian inference reveals several differences between both inferential concepts, it should be noted that in many applications and also in most of our simulations, differences where comparably small. If larger differences are observed, this may be an indicator for a weakly identified model or other problems caused by the model formulation. Therefore it may be advisable, to compare the results of both approaches, if possible.

6 BayesX

The mixed model methodology presented in the previous sections and all extensions that will be discussed in the following parts have been implemented in the public domain software package BayesX (Brezger, Kneib & Lang 2005a) as a part of this thesis. The program does not only comprise tools for performing empirical Bayes inference but also allows for full Bayesian inference based on MCMC and for the estimation of Gaussian and nongaussian directed acyclic graphs. Functions for handling and manipulating data sets and geographical maps, as well as for visualizing results are added for convenient use.

In this section, we mainly give an overview about the general usage of BayesX (Section 6.1) and describe the different types of objects existing in BayesX and their specific methods (Section 6.2). Instructions for downloading the program are given in the concluding Section 6.3. A complex example on childhood undernutrition in Zambia will be used for a tutorial-like introduction in Section 7.

6.1 Usage of BayesX

After having started BayesX, a main window divided into four sub-windows appears on the screen. These sub-windows are a command window for entering and executing commands, an output window for displaying results, a review window for easy access to past commands, and an object-browser that displays all objects currently available.

BayesX is object-oriented although the concept is limited, that means inheritance and other concepts of object-oriented languages like C++ or S-Plus are not supported. For every object type a number of object-specific methods may be applied to a particular object. To be able to estimate Bayesian regression models we need a dataset object to incorporate, handle and manipulate data, a remlreg object to estimate semiparametric regression models, and a graph object to visualize estimation results. If spatial effects are to be estimated, we additionally need map objects. Map objects mainly serve as auxiliary objects for remlreg objects and are used to read the boundary information of geographical maps and to compute the neighborhood matrix and weights associated with the neighbors. The syntax for generating a new object in BayesX is

> *objecttype objectname*

where *objecttype* is the type of the object, e. g. `dataset`, and *objectname* is the arbitrarily chosen name of the new object. In the following section, we give an overview on the most important methods of the object types required to estimate Bayesian structured additive regression models.

6.2 Object types in BayesX

6.2.1 Dataset objects

Data (in form of external ASCII files) are read into BayesX with the `infile` command. The general syntax is:

> *objectname*.`infile` [*varlist*] [, *options*] `using` *filename*

While *varlist* denotes a list of variable names separated by blanks (or tabs), *filename* specifies the name (including full path) of the external ASCII file storing the data. The variable list may be omitted if the first line of the file contains the variable names. BayesX assumes that each variable is stored in an extra column. Two options may be passed: the `missing` option to indicate missing values and the `maxobs` option for reading in large data sets. Specifying, for example, `missing = M` defines the letter 'M' as an indicator for a missing value. The default values are a period '.' or 'NA' (which remain valid indicators for missing values even if an additional indicator is defined).

The `maxobs` option can be used to speed up the import of large data sets, since it allows BayesX to allocate enough memory in advance to store the whole data set. Otherwise, the data reading process has to be restarted several times when the internal memory limit is reached leading to very long execution times. The usage of `maxobs` is strongly recommended if the number of observations exceeds 10,000. For instance, `maxobs=100000` indicates that the data set has 100,000 or less observations. Note that `maxobs=100000` does not mean that reading the data is stopped after 100,000 observations. Regardless of the specific value of `maxobs`, BayesX will read the complete data set, but new memory must be allocated when the number of observations of your data set exceeds the limit specified in `maxobs`.

Having read in the data, the data set can be inspected by double-clicking on the respective object in the object-browser.

Apart from the `infile` command, many more methods for handling and manipulating data are available, e. g. the `generate` command to create new variables, the `drop` command to drop observations and variables or the `descriptive` command to obtain summary statistics for the variables, see Chapter 4 of the BayesX reference manual (Brezger, Kneib & Lang 2005b).

6.2.2 Map objects

The boundary information of a geographical map is read into BayesX using the `infile` command of map objects. The current version supports two file formats, boundary files and graph files. A boundary file stores the boundaries of every region in form of closed polygons. Having read a boundary file, BayesX automatically computes the neighbors and associated weights of each region. By double-clicking on the respective object in the object-browser the map may be inspected visually.

The syntax for reading boundary files is

> *objectname*.`infile` [, `weightdef=` *wd*] `using` *filename*

where option `weightdef` specifies how the weights associated with each pair of neighbors are generated. Currently, there are three weight specifications available, `weightdef=adjacency` (the default), `weightdef=centroid`, and `weightdef=combnd`. If `weightdef=adjacency` is specified, the weights for each pair of neighbors are set equal to one. Setting

`weightdef=centroid` results in weights inverse proportional to the distance of the centroids of neighboring regions whereas `weightdef=combnd` produces weights proportional to the length of the common boundary (compare the discussion in Section 4.2.3).

The file following the keyword `using` is assumed to contain the boundaries in form of closed polygons. To give an example, we print a small part of the boundary file of Zambia that will be used in the following section. The map corresponding to this part of the boundary file can be found in Figure 6.1.

```
⋮

"52",48
28.080507,-12.537530
28.083376,-12.546980
28.109501,-12.548961
28.134972,-12.566787
28.154797,-12.585320
28.165771,-12.593912
28.165771,-12.593912
28.160769,-12.609917
28.152800,-12.633824
28.144831,-12.657733
28.132877,-12.677656
28.120922,-12.701565
28.120922,-12.717505
28.120922,-12.741411
28.116938,-12.761335
28.108969,-12.777274
28.100998,-12.793213
28.089045,-12.817122
28.085060,-12.837045
28.081076,-12.856968
28.081076,-12.876892
28.080862,-12.884153
28.080862,-12.884153
28.076630,-12.879521
28.031454,-12.881046
27.974281,-12.884675
27.910725,-12.878692
27.686228,-12.880120
27.665676,-12.854732
27.653563,-12.818301
27.639263,-12.759848
27.648254,-12.699927
27.662464,-12.680613
27.662464,-12.680613
27.666534,-12.675080
27.703260,-12.679779
27.752020,-12.695455
27.797932,-12.702188
27.836775,-12.707567
27.867813,-12.699892
27.902308,-12.667418
27.922668,-12.630853
27.943035,-12.596350
27.963434,-12.571486
27.983179,-12.563844
28.016331,-12.554779
28.070650,-12.542199
28.080507,-12.537530
⋮
```

For each region of the map, the boundary file must contain the identifying name of the region (in quotation marks) and the number of lines the polygon consists of, separated by a comma. Afterwards follow the polygons that form the boundary of the region. Note that the first and the last point must be identical (see the example above) to obtain a closed polygon. Compare Chapter 5 of the reference manual (Brezger, Kneib & Lang 2005b) for a detailed description of some special cases, e. g. regions divided into subregions.

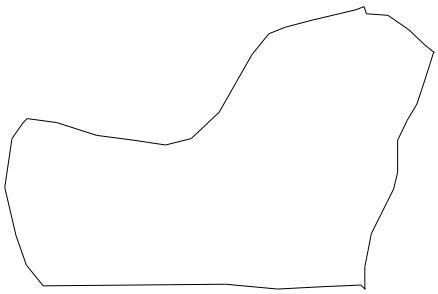


Figure 6.1: Part of the map of Zambia corresponding to the boundary information given above.

Reading the boundary information from an external file and assessing the neighborhood matrix may be a computationally intensive task if the map contains a huge number of regions or if the polygons are given in great detail. To avoid doing these computations every time BayesX is restarted, the generated neighborhood information can be directly

stored in a graph file. A graph file simply contains the nodes N and edges E of a graph $G = (N, E)$, which is a convenient way of representing the neighborhood structure of a geographical map. While the nodes of the graph correspond to the region codes, the neighborhood structure is represented by the edges of the graph. Weights associated with the edges may be given in a graph file as well.

We now describe the structure of a graph file as it is expected by BayesX. The first line of a graph file must contain the total number of nodes of the graph. The remaining lines provide the nodes of the graph together with their edges and associated weights. One node corresponds to three consecutive lines. The first of the three lines determines the name of the node which may simply be the name of a geographical region. In the second line, the number of edges of that particular node is given. The third line contains the corresponding edges of the node, where an edge is declared by the index of a neighboring node. The index starts with zero. For example, if the fourth and the seventh node/region in the graph file are connected/neighbours, the edge index for the fourth node/region is 3 and for the seventh node/region 6.

We illustrate the structure of a graph file with an example. The following few lines are the beginning of the graph file corresponding to the map of Zambia:

```
57
11
4
1 2 4 5
12
4
0 3 5 6
13
2
0 4
⋮
```

The first line specifies the total number of nodes, in the present example 57 nodes. The subsequent three lines correspond to the node with name '11', which is the first region in the map of Zambia. Region '11' has 4 neighbors, namely the second, third, fifth and sixth node appearing in the graph file. Lines 5 to 7 in the example correspond to node '12' and its four neighbors and lines 8 to 10 correspond to node '13'.

In a graph file, it is also possible to specify weights associated with the edges of the nodes. Since no weights are stated in the preceding example, all weights are automatically assumed to be equal to one. Nonequal weights are specified in the graph file by adding them following the edges of a particular node. An example of the beginning of a graph file with weights is given below:

```
57
11
4
1 2 4 5 0.461261 1.74605 1.13411 1.17406
12
4
0 3 5 6 0.461261 0.388206 0.537407 0.756847
13
```

2
0 4 1.74605 1.1692
⋮

Here, the edges of the first node '11' have weights 0.461261, 1.74605, 1.13411 and 1.17406.

A graph file is read into BayesX by using the `infile` command, but now the additional keyword `graph` has to be specified as an option:

> *objectname*`.infile, graph using` *filename*

A variety of boundary and graph files for different countries and regions are available at the BayesX homepage, see Section 6.3 for the internet address.

6.2.3 Remlreg objects

In BayesX, structured additive regression models can be estimated based on mixed model methodology using the `regress` command of remlreg objects. The general syntax is

> *objectname*`.regress` *model* [`weight` *weightvar*] [`if` *expression*] [`,` *options*] `using` *dataset*

Executing this command estimates the regression model specified in *model* using the data specified in *dataset*, where *dataset* is the name of a dataset object created previously. An `if` statement may be included to analyze only parts of the data and a weight variable *weightvar* to estimate weighted regression models. Options may be passed to specify the response distribution, details of estimation procedure (for example the termination criterion), etc. The syntax of models is:

$depvar = term_1 + term_2 + \cdots + term_r$

Here, *depvar* specifies the dependent variable in the model and $term_1$,...,$term_r$ define the way the covariates influence the response variable. The different terms must be separated by '+' signs. For clarification, some examples are given in the following. An overview about the possibilities for univariate model terms and interactions supported by BayesX is given in Table 6.2 More details can be found in Chapter 8 of the BayesX reference manual (Brezger, Kneib & Lang 2005b).

Suppose we aim at modeling the effect of three covariates X1, X2 and X3 on the response variable Y. Traditionally, a strictly linear predictor is assumed which can be estimated in BayesX by the model specification

```
Y = X1 + X2 + X3
```

Note that an intercept is automatically included into the models and must not be added explicitly. If we assume nonlinear effects of the continuous variables X1 and X2, for instance quadratic P-splines with second order random walk smoothness priors, the according model is declared by:

```
Y = X1(psplinerw2,degree=2) + X2(psplinerw2,degree=2) + X3
```

The second argument in the model formula above is optional. If omitted, a cubic spline will be estimated by default. Moreover, some more optional arguments may be passed,

e. g. to define the number of knots. For details we refer to the BayesX methodology manual.

Suppose now that an additional variable L is observed which provides information about the geographical location an observation belongs to. A spatial effect based on a Markov random field prior is added by:

```
Y = X1(psplinerw2,degree=2) + X2(psplinerw2,degree=2) + X3 +
    L(spatial,map=m)
```

The option `map` specifies the map object that contains the boundaries of the regions and the neighborhood information required to estimate a spatial effect.

The distribution of the response is specified by adding the option `family` to the options list. For instance, `family=gaussian` defines the responses to be Gaussian. Other valid specifications are found in Table 6.1. Note that response distributions for categorical responses and continuous survival time analysis are given too, although the corresponding methodology has not been described yet.

Family	**Response distribution**	**Link**
`family=gaussian`	Gaussian	identity
`family=binomial`	binomial	logit
`family=binomialprobit`	binomial	probit
`family=binomialcomploglog`	binomial	complementary log-log
`family=poisson`	Poisson	log
`family=gamma`	gamma	log
`family=multinomial`	unordered multinomial	logit
`family=cumprobit`	cumulative multinomial	probit
`family=cumlogit`	cumulative multinomial	logit
`family=seqprobit`	sequential multinomial	probit
`family=seqlogit`	sequential multinomial	logit
`family=cox`	continuous-time survival data	

Table 6.1: Summary of response distributions supported by BayesX.

6.2.4 Graph objects

Graph objects serve as a kind of graphics device and allow for the visualization of data and estimation results obtained from other objects in BayesX. Currently, graph objects may be used to draw scatter plots between variables (method `plot`), or to color geographical maps stored in map objects (method `drawmap`). For illustration purposes, method `drawmap` is presented to color the regions of a map according to some numerical characteristics. The syntax is:

> *objectname*`.drawmap` *plotvar regionvar* [`if` *expression*]`, map=`*mapname* [*options*] `using` *dataset*

Prior/Effect	**Syntax example**	**Description**
Linear effect	X1	Linear effect of X1.
First or second order random walk	X1(rw1) X1(rw2)	Nonlinear effect of X1.
P-spline	X1(psplinerw1) X1(psplinerw2)	Nonlinear effect of X1.
Seasonal prior	X1(season,period=12)	Time varying seasonal effect of X1 with period 12.
Markov random field	X1(spatial,map=m)	Spatial effect of X1 where X1 indicates the region an observation belongs to. The boundary information and the neighborhood structure is stored in the map object m.
Two-dimensional P-spline	X1(geosplinerw1,map=m) X1(geosplinerw2,map=m) X1(geosplinebiharmonic,map=m)	Spatial effect of X1. Estimates a two-dimensional P-spline based on the centroids of the regions. The centroids are stored in the map object m.
Stationary Gaussian random field	X1(geokriging,map=m)	Spatial effect of X1. Estimates a stationary Gaussian random field based on the centroids of the regions. The centroids are stored in the map object m.
Random intercept	X1(random)	I. i. d. Gaussian (random) effect of the group indicator X1, e. g. X1 may be an individuum indicator when analyzing longitudinal data.
Baseline in Cox models	X1(baseline)	Nonlinear shape of the baseline effect $\lambda_0(X1)$ of a Cox model (see Part IV). The log-baseline $\log(\lambda_0(X1))$ is modeled by a P-spline with second order penalty.

Type of interaction	**Syntax example**	**Description**
Varying coefficient term	X2*X1(rw1) X2*X1(rw2) X2*X1(psplinerw1) X2*X1(psplinerw2)	Effect of X2 varies smoothly over the range of the continuous covariate X1.
Random slope	X2*X1(random)	The regression coefficient of X2 varies with respect to the unit or cluster index X1.
Geographically weighted regression	X2*X1(spatial,map=m)	Effect of X2 varies geographically. Covariate X1 indicates the region an observation belongs to.
Two-dimensional surface	X2*X1(pspline2dimrw1) X2*X1(pspline2dimrw2) X2*X1(pspline2dimbiharmonic)	Two-dimensional surface for the continuous covariates X1 and X2.
Stationary Gaussian random field	X1*X2(kriging)	Stationary Gaussian random field for coordinates X1 and X2.
Time-varying effect in Cox models	X2*X1(baseline)	Effect of X2 varies over time, where the time-axis is given by X1

Table 6.2: Univariate and interaction terms in BayesX.

Method `drawmap` generates the map stored in the map object *mapname* and displays it either on the screen or stores it as a postscript file (if option `outfile` is specified). The regions with region code *regionvar* are colored according to the values of the variable *plotvar*. The variables *plotvar* and *regionvar* are supposed to be stored in the dataset object *dataset*. Several options are available for customizing the graph, e. g. for changing from grey scale to color scale or for storing the map as a postscript file (see Chapter 6 of the BayesX reference manual and Section*7.5).

6.3 Download

The latest version of BayesX can be downloaded from

`http://www.stat.uni-muenchen.de/~bayesx/`

The BayesX homepage also contains two tutorial based on the Zambia data set to make new users familiar with the functionality of BayesX. One of these tutorials is reproduced in the next section. Furthermore, the manuals can be downloaded, several boundary and graph files are available, and new features of the latest releases are announced.

7 Childhood undernutrition in Zambia

In this section we will demonstrate the flexibility of structured additive regression models in nonstandard regression situations based on the Zambian childhood undernutrition data introduced in Section 2.1. Our analysis will follow the discussion in Kandala, Lang, Klasen & Fahrmeir (2001), where a fully Bayesian model for the same data is presented. In addition, this section contains a tutorial-like introduction to the usage of BayesX for estimating structured additive regression models based on mixed model methodology. The data set and the map of Zambia are available on the BayesX homepage. Thus, the reader is enabled to reproduce the results discussed in this section on his or her own.

7.1 Reading data set information

In a first step, the available data set (as described in Table 2.1 on page 5) is read into BayesX by creating a dataset object named `d`

```
> dataset d
```

and using the method `infile`:

```
> d.infile, maxobs=5000 using c:\data\zambia.raw
```

Note that we assume the data to be provided in the external file `c:\data\zambia.raw`. The first few lines of this file look like this:

```
hazstd bmi agc district rcw edu1 edu2 tpr sex
0.0791769   21.83   4   81   -1   1   0   1   -1
-0.2541965   21.83   26   81   -1   1   0   1   -1
-0.1599823   20.43   56   81   1   -1   -1   1   1
0.1733911   22.27   6   81   -1   0   1   1   1
```

In our example the file contains the variable names in the first line. Hence, it is not necessary to pass a list of variable names to the `infile` command. Option `maxobs` is used to speed up the execution time of the `infile` command, compare the discussion in Section 6.2.1. However, this does only play an important role for larger data sets with more than 10,000 observations and could therefore be omitted in the present example.

After having read the data set, we can inspect the data visually. The execution of

```
> d.describe
```

opens an Object-Viewer window containing the data in form of a spreadsheet (see Figure 7.1). This can also be achieved by double-clicking on the respective dataset object in the object-browser.

Further methods allow to examine characteristics of the variables in the dataset object. Given a categorial variable, e. g. *sex*, the `tabulate` command

```
> d.tabulate sex
```

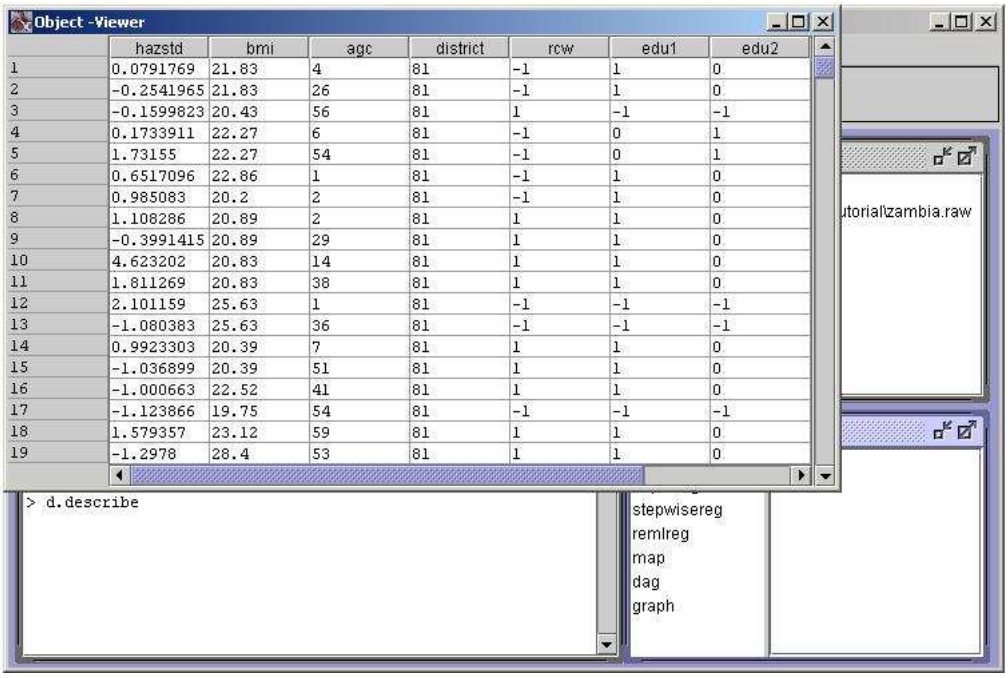


Figure 7.1: An Object-Viewer containing the dataset.

may be used to produce the frequency table

```
Variable: sex

          Value        Obs          Freq           Cum
             -1       2451        0.5057        0.5057
              1       2396        0.4943             1
```

in the output window. For continuous variables the `descriptive` command prints several characteristics of the variable in the output window. For example, executing

```
> d.descriptive bmi
```

leads to

```
  Variable     Obs        Mean     Median        Std       Min       Max
       bmi    4847   21.944349       21.4   3.2879659      12.8     39.29
```

7.2 Compute neighborhood information

In the following, we aim at estimating a spatially correlated effect of the district a child lives in. Hence, we need the boundaries of the districts in Zambia to assess the neighborhood information of the corresponding map. Initializing a map object via

```
> map m
```

and reading the boundaries using the `infile` command

```
> m.infile using c:\data\zambia.bnd
```

causes BayesX to automatically compute the neighborhood structure of the map.

Map objects may be visualized using method `describe`:

```
> m.describe
```

results in the graph shown in Figure 7.2. Additionally, `describe` prints further information about the map object in the output window including the name of the object, the number of regions, the minimum and maximum number of neighbors and the bandwidth of the corresponding adjacency or neighborhood matrix:

```
MAP m
Number of regions: 54
Minimum number of neighbors: 1
Maximum number of neighbors: 9
Bandsize of corresponding adjacency matrix: 24
```

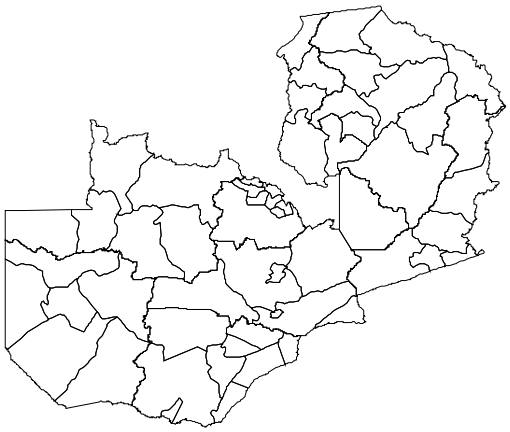


Figure 7.2: The districts of Zambia.

In order to see to which extend the available file format affects the computation time of the `infile` command, we create a second map object and read the information from the graph file of Zambia. In this case, we have to specify the additional keyword `graph`:

```
> map m1
> m1.infile, graph using c:\data\zambia.gra
```

Obviously, reading geographical information from a graph file is much faster than reading from a boundary file. However, using graph files also has the drawback of loosing the full information on the polygons forming the map. As a consequence, we can not visualize a map object originating from a graph file. Therefore, typing

```
> m1.describe
```

raises an error message and visualizing estimation results of spatial effects can only be based on map objects created from boundary files, although estimation can be carried out using graph files. Since we continue to work with the map object `m`, we delete `m1`:

```
> drop m1
```

7.3 Analysis based on structured additive regression

To estimate a regression model based on mixed model techniques, we first initialize a remlreg object:

```
> remlreg r
```

By default, estimation results are written to the subdirectory `output` of the installation directory. In this case, the default filenames are composed of the name of the remlreg object and the type of the specific file. Usually, it is more convenient to store the results in a user-specified directory which can be defined by the `outfile` command of remlreg objects:

```
> r.outfile = c:\data\r
```

Note that `outfile` does not only determine an output directory but also a base filename (the character 'r' in our example). Therefore executing the command above leads to storage of the results in the directory '`c:\data`' and all filenames will start with the character 'r'. Of course, the base filename may be different from the name of the remlreg object.

Beyond parameter estimates BayesX also produces some further information on the estimation process. In contrast to parameter estimates, this information is not stored automatically but is printed in the output window. Hence, saving the contents of the output window is advisable. This can be achieved automatically by opening a log file using the `logopen` command

```
> logopen, replace using c:\data\logzambia.txt
```

Every information written to the output window is then additionally duplicated in the log file. Option `replace` allows BayesX to overwrite an existing file with the same name as the specified log file. Without `replace` results are appended to an existing file.

The regression model presented in Kandala et al. (2001) which will also be applied here is given by the following semiparametric predictor:

$$\begin{aligned}\eta &= \gamma_0 + \gamma_1 rcw + \gamma_2 edu1 + \gamma_3 edu2 + \gamma_4 tpr + \gamma_5 sex \\ &\quad + f_1(bmi) + f_2(agc) + f_{str}(district) + f_{unstr}(district)\end{aligned}$$

As the two continuous covariates *bmi* and *agc* are assumed to expose a possibly nonlinear effect on the Z-score, they are modeled nonparametrically as cubic P-splines with second order random walk prior in the present example. The spatial effect of the district is split into a spatially correlated part $f_{str}(district)$ and an uncorrelated part $f_{unstr}(district)$, see Section 4.2.3. The correlated part is modeled by Markov random field prior (4.25), where the neighborhood matrix is obtained from the map object `m`. The uncorrelated part is modeled by an i.i.d Gaussian effect.

For estimation of the model method `regress` of remlreg objects is called:

```
> r.regress hazstd = rcw + edu1 + edu2 + tpr + sex + bmi(psplinerw2)
  + agc(psplinerw2) + district(spatial,map=m) + district(random),
  family=gaussian lowerlim=0.01 eps=0.0005 using d
```

Options `lowerlim` and `eps` serve the control of the estimation process. Since small variances are near to the boundary of their parameter space, the usual Fisher-scoring algorithm needs to be modified: If the fraction of the penalized part of an effect relative to the total effect is less than `lowerlim`, the estimation of the corresponding variance is stopped and the estimate is defined to be the current value of the variance. The option `eps` defines the termination criterion for the estimation process. The default value for `lowerlim` is 0.001, the default value for `eps` is 0.00001. However, since our analysis is mainly for explanatory purpose, we choose somewhat weaker conditions allowing for faster 'convergence' of the algorithm. On a 2.4 GHz PC estimation takes about 2 minutes and 30 seconds with the above specifications of `lowerlim` and `eps`.

A further option of method `regress` concerns the maximum number of iterations (`maxit`) that should be performed in the estimation. Even if no convergence could be achieved within `maxit` iterations, BayesX produces results based on the current values of all parameters although a warning message will be printed in the output window.

For the fixed effects we obtain the following results (in the output window):

Variable	Post. Mode	Std. Dev.	p-value	95% Confidence	Interval
const	0.0610357	0.0341574	0.0367456	-0.00592622	0.127998
rcw	0.00767158	0.0136564	0.286931	-0.0191003	0.0344434
edu1	-0.0605105	0.0261369	0.0103181	-0.111749	-0.00927192
edu2	0.234917	0.0459925	4.41249e-06	0.144754	0.325081
tpr	0.0904093	0.0218891	6.17648e-05	0.047498	0.133321
sex	-0.0585716	0.0129304	2.04243e-05	-0.0839203	-0.0332229

For interpretation, note the following: According to the definition of the Z-score in Equation (2.1) on page 5, a large value of the Z-score indicates better nourished children while negative values indicate malnutritioned children. In general, the findings are as expected: Children living in urban areas are better nourished just as children of highly educated mothers. Male children are more likely to be undernourished and children of currently working mothers also show a slightly negative effect.

For the nonparametric effects, some information (e. g. on the variance of the corresponding effect) is also printed in the output window. For example, for the effect of *bmi* we obtain

```
f_bmi_pspline

Estimated variance:  1.14816e-05
Inverse variance:    87095.9
Smoothing parameter: 69863.5
(Smoothing parameter = scale / variance)
NOTE: Estimation of the variance was stopped after iteration 6 because the
      corresponding penalized part was small relative to the linear predictor.

Variance and smoothing parameter are stored in file
c:\data\r_f_bmi_pspline_var.res

Results are stored in file
c:\data\r_f_bmi_pspline.res
Postscript file is stored in file
c:\data\r_f_bmi_pspline.ps

Results may be visualized using method 'plotnonp'
Type for example: objectname.plotnonp 1
```

Since results for a nonparametric effect usually consist of a lot of parameters, these are not printed on the screen but written to external ASCII files. The names of these files are indicated in the output window (see the example above). For the different regression terms of the model the files comprise the posterior mode, the 80% and 95% credible interval, the standard deviations and the corresponding 80% and 95% posterior probabilities of the estimated effects. For example, the beginning of the file `c:\data\r_f_bmi_pspline.res` for the effect of *bmi* looks like this:

```
intnr   bmi   pmode   ci95lower   ci80lower   std   ci80upper   ci95upper   pcat95   pcat80
1   12.8   -0.22305   -0.304661   -0.276409   0.0416301   -0.169692   -0.141439   -1   -1
2   13.15   -0.215246   -0.292828   -0.26597   0.0395749   -0.164522   -0.137663   -1   -1
3   14.01   -0.19607   -0.264173   -0.240597   0.0347394   -0.151544   -0.127968   -1   -1
```

The levels of the credible intervals and posterior probabilities may be changed by the user using the options `level1` and `level2`. For example, specifying `level1=99` and `level2=70` in the option list of the `regress` command leads to the computation of 99% and 70% credible intervals and posterior probabilities. The defaults are `level1=95` and `level2=80`.

Nonparametric and spatial effects are visualized automatically by BayesX and the resulting graphs are stored in postscript format. For example, the effect of *bmi* is visualized in the file `c:\data\r_f_bmi_pspline.ps`. The names of the postscript files are supplied in the output window, see the example above. In addition, a file containing BayesX commands that would produce the automatically generated graphics is stored in the output window. The advantage is that the user can look up these commands and specify additional options to customize the graphs (see the following two sections for details). In our example, the name of this file is `c:\data\r_graphics.prg`.

Moreover a file with ending `.tex` is created in the output directory. This file contains a summary of the estimation results and may be compiled using LaTeX.

Having completed the estimation, the log file can be closed by typing

```
> logclose
```

Note that the log file is closed automatically when you exit BayesX.

7.4 Visualizing estimation results

Nonparametric and spatial effects are most intuitively presented by visualization. BayesX provides three options to display estimation results:

- As mentioned in the previous section, BayesX automatically stores most of the results in postscript files.
- Post-estimation commands of remlreg objects allow to display results immediately after having executed a `regress` command (as long as the corresponding remlreg object is available).

- Graph objects can be used to visualize the content of dataset objects. Hence, creating a dataset object from one of the ASCII files containing the estimation results allows to generate graphics of these results.

In this section we describe the application of the post-estimation commands as well as the usage of graph objects to enable the user to reproduce the automatically generated plots directly in BayesX. Section 7.5 describes how to customize plots.

7.4.1 Post-estimation commands

After having estimated a regression model plots for nonparametric effects of metrical covariates can be generated by the post-estimation command `plotnonp`.

```
> r.plotnonp 1
```

and

```
> r.plotnonp 2
```

produce the graphs shown in Figure 7.3 in an Object-Viewer window. Each effect matches a predetermined number obtained from numbering the terms in the regression model which is supplied in the output window (compare the example on the effect of *bmi* above). Note that `plotnonp` can only be applied as long as the corresponding regression object is available, and hence `plotnonp` is called a post-estimation command.

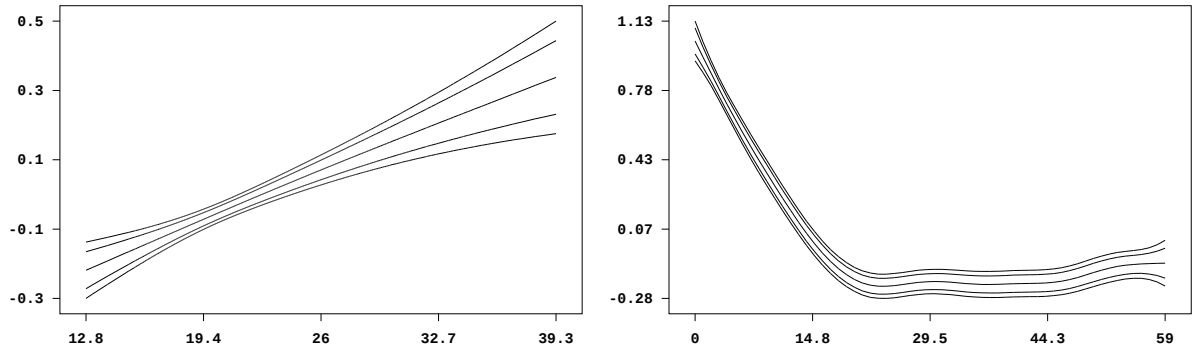


Figure 7.3: Effect of the body mass index of the child's mother and of the age of the child together with pointwise 80% and 95% credible intervals.

The effect of the mother's body mass is almost linearly increasing indicating a lower risk for undernutrition for better nourished mothers. The effect of the age of the child is obviously nonlinear and strongly decreasing between birth and an age of about 20 months. This continuous worsening of the nutritional status may be caused by the fact that most of the children obtain liquids other than breast milk already shortly after birth. After 20 months a relatively stable, low level is reached. The slight increase of the effect after 24 months is introduced by a change of the reference standard at that point.

By default the plots produced by `plotnonp` contain the posterior mode and pointwise credible intervals according to the levels specified in the `regress` command. Hence, by default Figure 7.3 includes pointwise 80% and 95% credible intervals.

A plot may be stored in postscript format using the `outfile` option. Executing

```
> r.plotnonp 1, replace outfile = c:\data\f_bmi.ps
```

stores the plot for the estimated effect of *bmi* in the file `c:\data\f_bmi.ps`. Again, specifying `replace` allows BayesX to overwrite an existing file (otherwise an error message would be raised, if the file is already existing). Note that BayesX does not display the graph on the screen if option `outfile` is passed.

Estimation results of spatial effects are best visualized by drawing the respective map and coloring the regions of the map according to some characteristic of the posterior, e. g. the posterior mode. In case of the structured spatial effect, the respective Figure 7.4 can be achieved using the post-estimation command `drawmap`:

```
> r.drawmap 3
```

The map shows pronounced undernutrition in the northern part of Zambia and better nutrition in the southern part. This is in agreement with findings on the spatial segregation of poverty and deprivation within Zambia (see Kandala et al. (2001) for a more detailed discussion of estimation results).

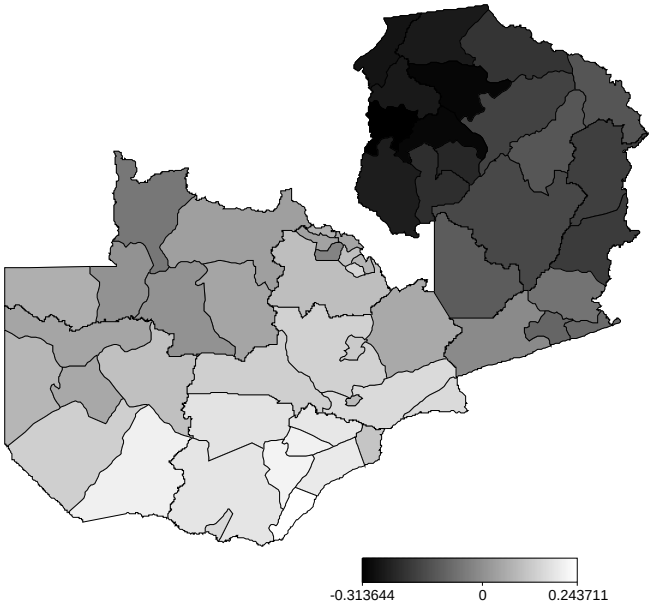


Figure 7.4: Posterior mode estimates of the structured spatial effect.

7.4.2 Graph objects

The commands presented in the previous subsection work only as long as the corresponding regression object is available in the current BayesX session. However, it may also be desirable to visually inspect results of former analyses which can be achieved using graph objects. Note that graph objects are also used in the batch file of the commands to reproduce the automatically generated graphics. Therefore, the purpose of this subsection is also to get the content of this batch file across.

First, the estimation results have to be stored in a dataset object, e. g. by running the commands

```
> dataset res
> res.infile using c:\data\r_f_bmi_pspline.res
```

to obtain the results for the effect of *bmi*. Now, the estimation results (or any content of a dataset object) may be visualized using a graph object which we create by typing

```
> graph g
```

Executing the `plot` command of graph objects

```
> g.plot bmi pmode ci95lower ci80lower ci80upper ci95upper using res
```

reproduces the graph in the left part of Figure 7.3.

In analogy, spatial effects can be displayed using method `drawmap` of graph objects:

```
> res.infile using c:\data\r_f_district_spatial.res
> g.drawmap pmode district, map=m using res
```

Since – in contrast to a remlreg object – no map object is associated with a graph object, we explicitly have to specify the desired map in the option list.

Moreover, graph objects allow to plot other characteristics of the posterior than the posterior mode. For instance, the posterior 95% probabilities may be visualized by

```
> g.drawmap pcat95 district, map=m using res
```

The result is shown in Figure 7.5.

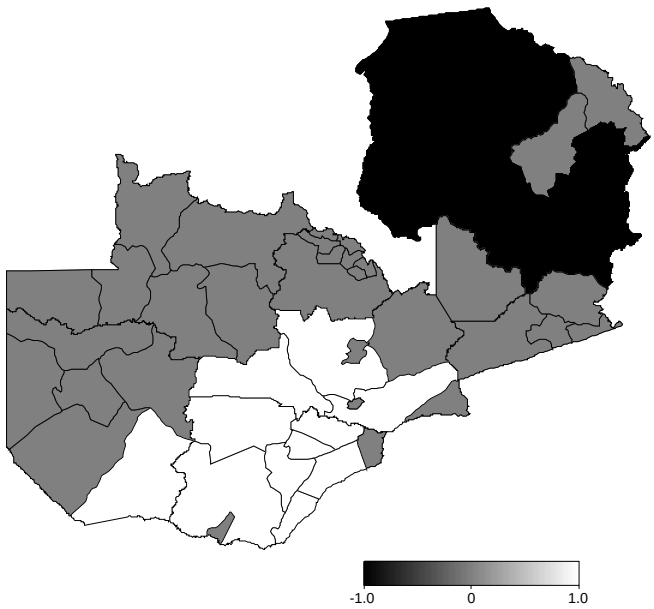


Figure 7.5: Posterior 95% probabilities of the structured spatial effect. Black (white) denotes strictly negative (positive) credible intervals

As a further advantage, visualization of the estimation results of uncorrelated spatial effects is enabled. Since these are modeled as unstructured random effects, BayesX is

unable to recognize them as spatial effects. However, proceeding as follows gives us the possibility to plot the unstructured spatial effect shown in Figure 7.6:

```
> res.infile using c:\data\r_f_district_random.res
> g.drawmap pmode district, map=m color swapcolors using res
```

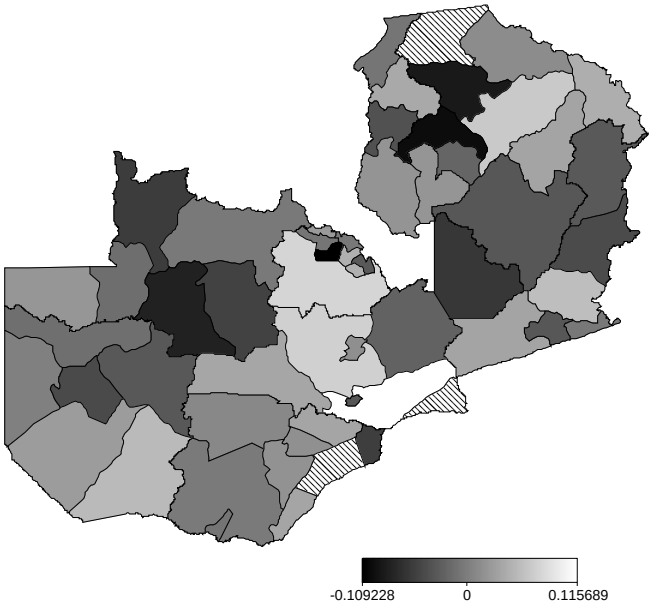


Figure 7.6: Posterior mode estimates of the unstructured spatial effect.

7.5 Customizing graphics

Several options for customizing graphics are available in BayesX. In the following, all options are described for the usage with the post-estimation commands but may be used in combination with graph objects as well.

Nonparametric effects may be preferentially presented with one single credible interval. The number of included intervals can be determined by the `levels` option. Possible values of this option are `1` and `2`, corresponding to the levels specified in the `regress` command (compare Section 7.3). If the default values of `level1` and `level2` have been used, specifying `levels=2` in the `plotnonp` command causes BayesX to plot the 80% credible interval only (Figure 7.7):

```
> r.plotnonp 1, levels=2
```

The additional introduction of axis labels and titles helps to distinguish more clearly between different covariates. Both text fields are supported by BayesX as demonstrated in the following examples (compare Figure 7.8 for the resulting plots):

```
> r.plotnonp 1, title="Mother body mass index"
> r.plotnonp 1, xlab="bmi" ylab="f_bmi" title="Mother body mass index"
```

By default, BayesX displays x- and y-axis with five equidistant ticks according to the range of the data that is to be visualized. These defaults may be overwritten using the options `xlimbottom`, `xlimtop` and `xstep` for the x-axis and `ylimbottom`, `ylimtop` and `ystep` for

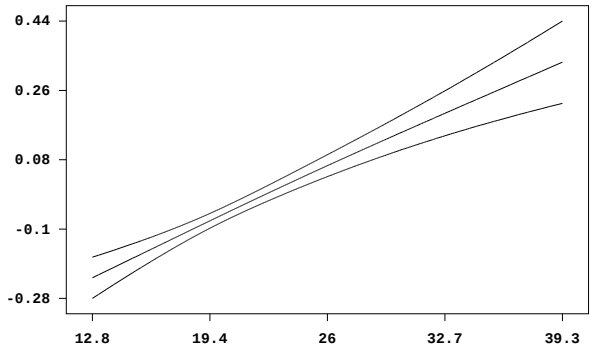


Figure 7.7: Effect of the body mass index of the child's mother with pointwise 80% credible intervals only.

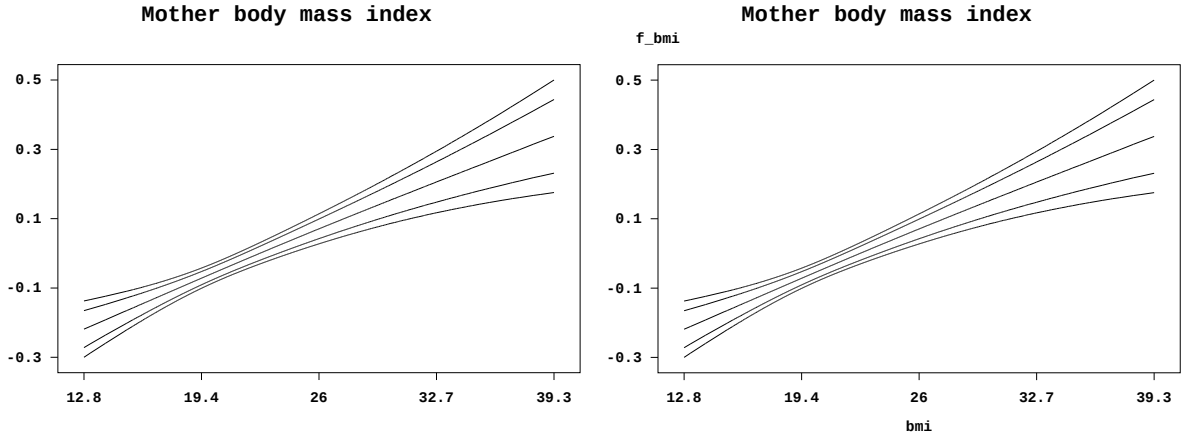


Figure 7.8: Inclusion of title and axis labels.

the y-axis, respectively. The meaning of these options is more or less self-explanatory. A demonstration with the following commands leads to the graph shown in Figure 7.9:

```
> r.plotnonp 1, xlab="bmi" ylab="f_bmi" title="Mother body mass index"
  ylimbottom=-0.8 ylimtop=0.6 ystep=0.2 xlimbottom=12 xlimtop=40
```

Figure 7.9 also includes a customized graph for the effect of the age of the child created by

```
> r.plotnonp 2, xlab="age" ylab="f_age" title="Age of the child in months"
  ylimbottom=-0.3 ystep=0.3 xlimbottom=0 xlimtop=60 xstep=10
```

Now we turn to the options referring to method `drawmap`. By default, `drawmap` represents different values of the posterior mode on a grey scale. Using option `color` forces BayesX to switch to color scales instead. By default, higher values are indicated by greenish colors and smaller values by reddish colors. Specifying `swapcolors` reverses this definition. Therefore, the command

```
> r.drawmap 3, color swapcolors
```

leads to the graph shown in Figure 7.10 with higher values being represented by reddish colors and smaller values by greenish colors.

Similar options as for the visualization of nonparametric effects exist for method `drawmap`.

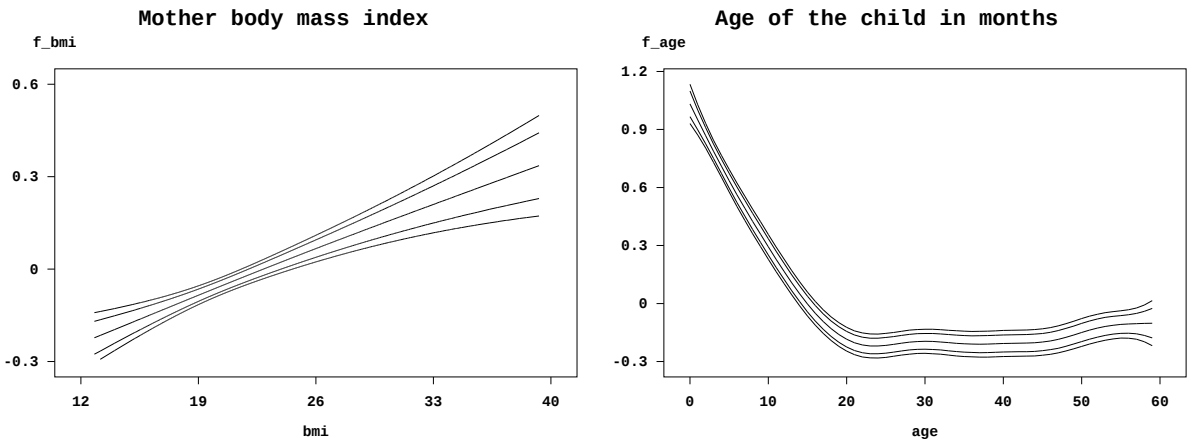


Figure 7.9: Redefining x- and y-axis.

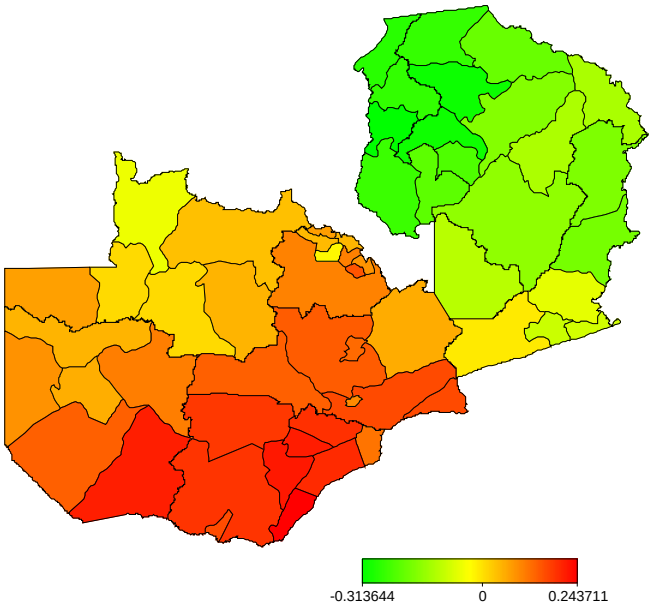


Figure 7.10: Posterior mode of the structured spatial effect in color.

For example, a title may be included by specifying the option `title`

```
> r.drawmap 3, color swapcolors title="Structured spatial effect"
```

or the range of values to be displayed may be defined using the options `lowerlimit` and `upperlimit`:

```
> r.drawmap 3, color swapcolors title="Structured spatial effect"
  lowerlimit=-0.3 upperlimit=0.3
```

The graph produced by the second command is shown in Figure 7.11.

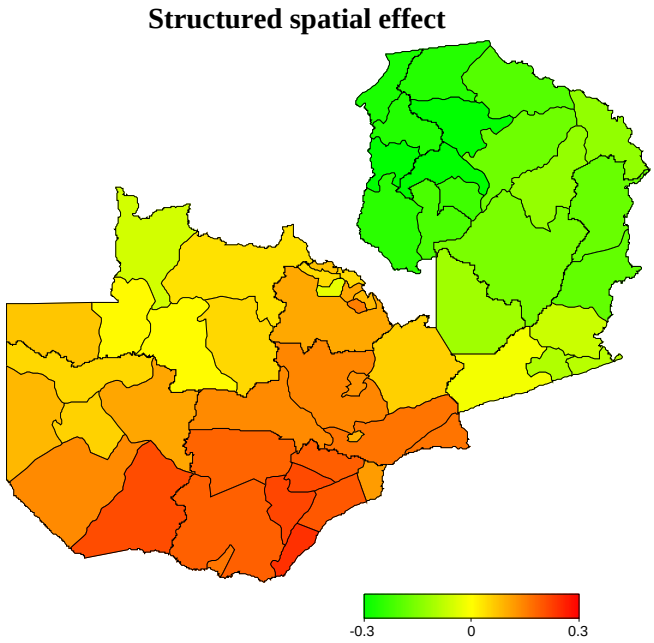


Figure 7.11: Specifying a title and the range of the plot for spatial effects.

8 A simulation study in spatial smoothing techniques

In order to compare the different spatial smoothing techniques discussed in Section 4.2 and to investigate the statistical properties of the mixed model approach compared to its fully Bayesian counterpart, we conducted a number of simulation studies with different spatial setups. In the first part of this section, we consider discrete spatial information, i. e. spatial information is given in terms of regions the individual observations belong to. This part is further divided into two sections dealing with a smooth spatial function (Section 8.1.1) and a more wiggly spatial function (Section 8.1.2) as underlying effects. In the second part of the simulation study, continuous spatial information is considered, i. e. spatial information is given in terms of longitude and latitude of the corresponding observation. This simulation study was conducted as a part of her diploma thesis by Manuela Hummel (2005).

8.1 Discrete spatial information

8.1.1 Smooth spatial function

In this section, we examine simple spatial regression models with predictors

$$\eta_i = \beta_0 + f_{spat}(s_i),$$

where f_{spat} is a smooth spatial function defined upon the centroids of the 54 districts s within Zambia (compare also the application in the previous section). As indicated in Figure 8.1, the spatial function is given by a trend increasing linearly from west to east.

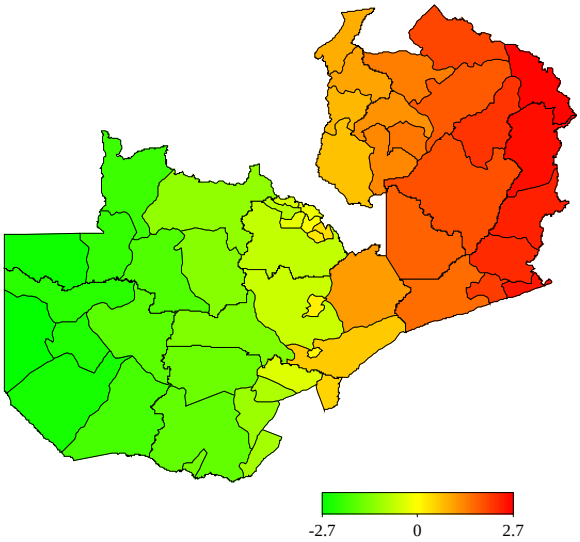


Figure 8.1: Smooth regional data: True spatial function.

We simulated $R = 100$ data sets, each consisting of $n = 500$ observations, based on four different distributional assumptions:

- A binary logit model, i. e. $y_i \sim B(1, \pi_i)$, where $\pi_i = \frac{\exp(\eta_i)}{1+\exp(\eta_i)}$,

- a Binomial logit model with $m = 10$ repeated binary observations, i. e. $y_i \sim B(10, \pi_i)$, where $\pi_i = \frac{\exp(\eta_i)}{1+\exp(\eta_i)}$,
- a loglinear Poisson model, i. e. $y_i \sim Po(\lambda_i)$, where $\lambda_i = \exp(\eta_i)$,
- and a linear model for Gaussian responses, i. e. $y_i \sim N(\mu_i, \sigma^2)$, where $\mu_i = \eta_i$ and $\sigma^2 = 0.25$.

In each case, the 54 regions were randomly assigned to the observations.

The resulting data sets were analyzed based on four strategies:

- Assume Markov random field prior (4.25) for f_{spat} and estimate the model based on mixed model methodology.
- Assume a Gaussian random field with $\nu = 1.5$ based on the centroids of the regions for f_{spat} and estimate the model using mixed model methodology.
- Assume Markov random field prior (4.25) for f_{spat} and estimate the model based on MCMC with hyperparameters $a = b = 0.001$ for the variance τ^2_{spat}.
- Assume a Gaussian random field with $\nu = 1.5$ based on the centroids of the regions for f_{spat} and estimate the model using MCMC with hyperparameters $a = b = 0.001$ for the variance τ^2_{spat}.

To compare the results of the different approaches average estimates

$$\bar{f}_{spat}(s) = \frac{1}{R}\sum_{r=1}^{R} \hat{f}^{(r)}_{spat}(s),$$

where $\hat{f}^{(r)}_{spat}(s)$ denotes estimates from the r-th replication, and the empirical bias

$$\text{bias}(\hat{f}_{spat}(s)) = \bar{f}_{spat}(s) - f_{spat}(s)$$

were investigated. In addition, we computed empirical mean squared errors (MSE)

$$MSE(\hat{f}^{(r)}_{spat}) = \frac{1}{54}\sum_{s=1}^{54}\left[f_{spat}(s) - \hat{f}^{(r)}_{spat}(s)\right]^2, \quad r = 1, \ldots, R$$

and empirical coverage probabilities, i. e. relative frequencies indicating how often the true function $f_{spat}(s)$ was covered by the credible interval of the estimate. For empirical Bayes estimates, credible intervals are computed as described in Section 5.2. In the fully Bayesian approach pointwise credible intervals are simply obtained by computing the respective empirical quantiles of sampled function values.

From a closer inspection of these quantities the following conclusions can be drawn:

- In general, differences between the empirical Bayes (EB) and the fully Bayesian (FB) approach are small in terms of bias (Figures 8.2 and 8.3 show results for Poisson and Bernoulli distributed responses).
- For Gaussian and Binomial distributed responses, all four estimation techniques result in highly reliable, almost unbiased estimates (results not shown).

- For Poisson distributed responses, the Kriging approaches in general lead to less biased estimates than the MRF approaches but have higher bias at the westerly boundary (Figure 8.2).
- The most pronounced bias is found for Bernoulli distributed responses (Figure 8.3). Interestingly, the highest bias for the MRF approaches is not observed at the boundaries of the observation area but in one of the middle countries.
- Consider the bias relative to the true function (as in the diagonal plots in Figures 8.2 and 8.3), reveals that the bias can in general be neglected.
- In terms of average coverage probabilities, differences between EB and FB approaches become more distinct. While the FB approach is close to the nominal level for Binomial, Poisson and Gaussian distributed responses, the EB approach only meets the nominal level in combination with MRFs but is too conservative in combination with GRFs (Table 8.1).
- For Bernoulli distributed responses, all approaches lead to relatively wide credible intervals, but the empirical Bayes approach in combination with GRFs still shows the most conservative behavior.
- Considering empirical MSEs, the EB approach results in somewhat better estimates for Bernoulli distributed responses, while differences are generally quite small for the remaining response distributions (Figure 8.4). Note that the MSEs are not displayed on the same scale for all distributions to emphasize the comparison between the smoothing techniques.
- For Poisson distributed responses, an outlier with very high MSE is observed for the empirical Bayes approach.
- For Binomial and Gaussian distributed responses, GRFs lead to somewhat preferable estimates compared to MRFs. For Bernoulli and Poisson distributed responses differences are small.

		Bernoulli	Binomial	Poisson	Gaussian
MRF EB	80%	0.876	0.835	0.819	0.818
	95%	0.978	0.962	0.958	0.956
MRF FB	80%	0.886	0.838	0.836	0.818
	95%	0.980	0.965	0.966	0.957
Kriging EB	80%	0.953	0.989	0.977	0.997
	95%	0.995	1.000	0.998	1.000
Kriging FB	80%	0.866	0.837	0.845	0.820
	95%	0.977	0.967	0.972	0.960

Table 8.1: Smooth regional data: Average empirical coverage probabilities. Values that are more than 2.5% below (above) the nominal level are indicated in green (red).

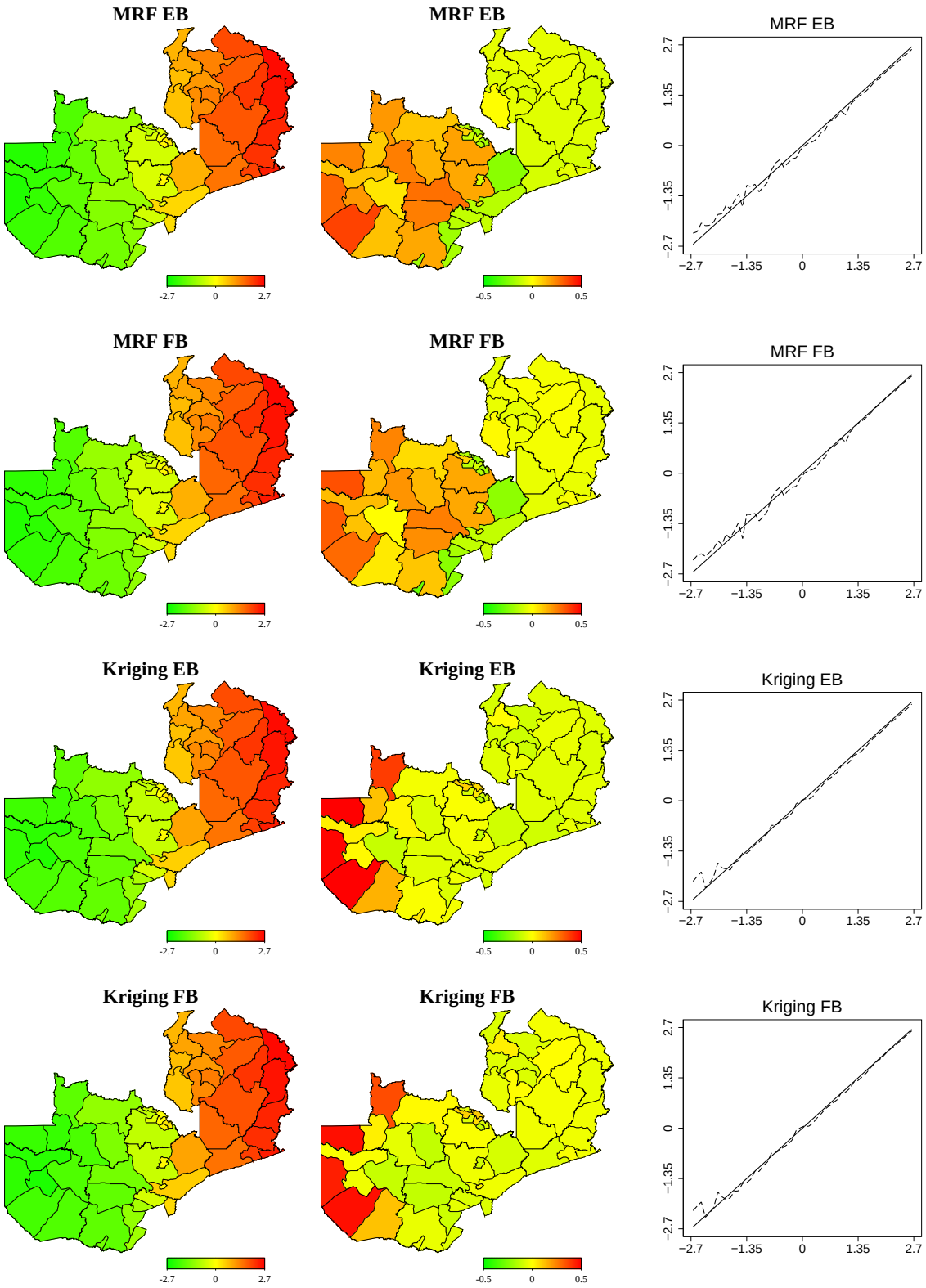


Figure 8.2: Smooth regional data: Average estimates (left panel) and estimated bias (middle panel) for Poisson distributed responses. In the diagonal plots (right panel) average estimates are plotted against the ordered true values (dashed line). The reference curve obtained from plotting true values against true values (solid line) is included for comparison.

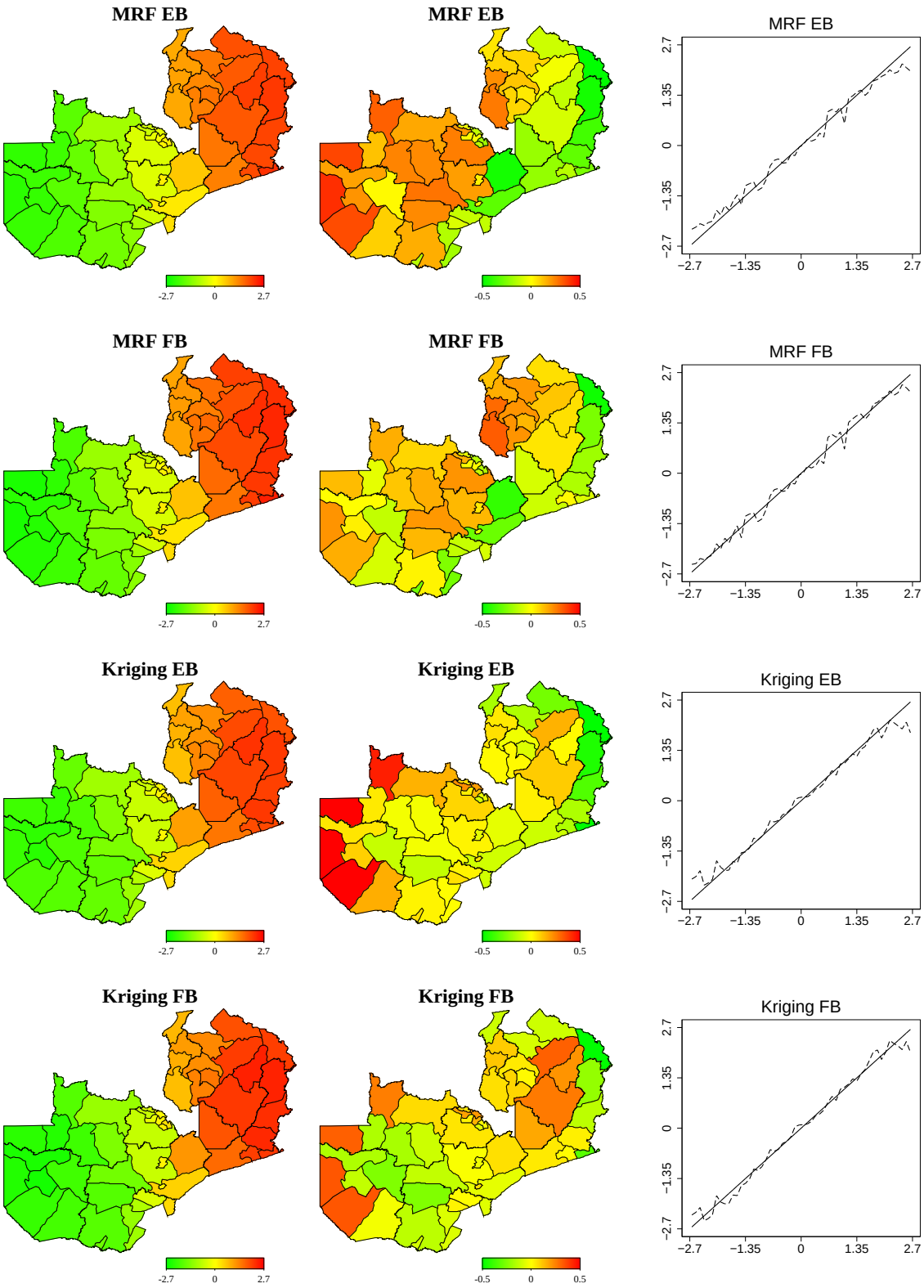


Figure 8.3: Smooth regional data: Average estimates (left panel) and estimated bias (middle panel) for Bernoulli distributed responses. In the diagonal plots (right panel) average estimates are plotted against the ordered true values (dashed line). The reference curve obtained from plotting true values against true values (solid line) is included for comparison.

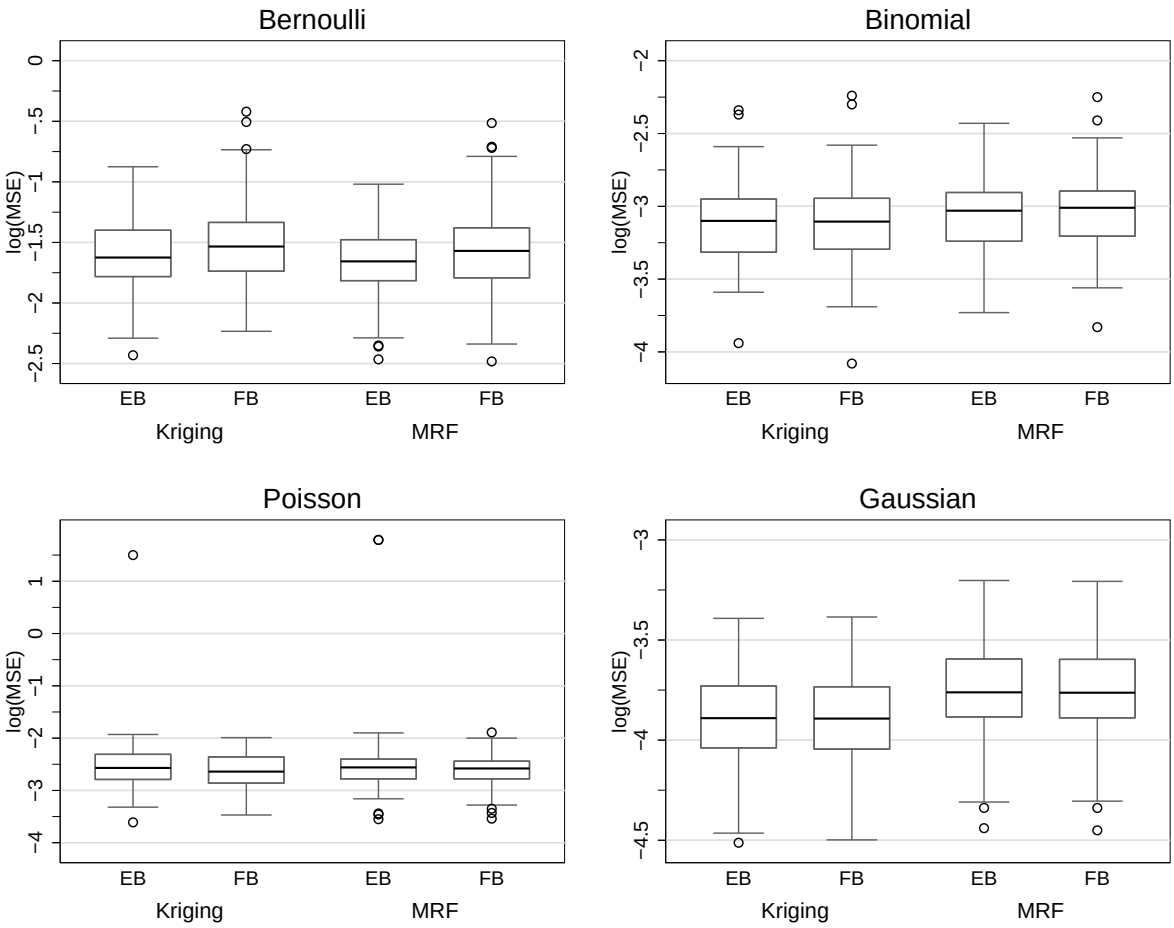


Figure 8.4: Smooth regional data: Empirical $\log(MSE)$ for the different response distributions, inferential procedures, and spatial priors.

8.1.2 Wiggly spatial function

In this section, the smooth spatial function is replaced with a more wiggly function computed as a multiple of the region-specific average of the undernutrition score analyzed in Section 7 (see Figure 8.5). The simulation setup and the examined smoothing techniques are the same as in the previous section.

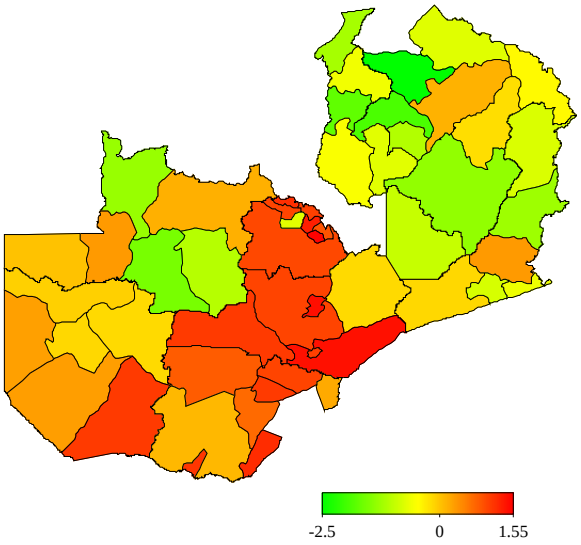


Figure 8.5: Wiggly regional data: True spatial function.

Again results are compared in terms of average estimates and empirical bias, average empirical coverage probabilities, and empirical MSEs. The conclusions drawn from these comparisons are

- Differences between EB and FB estimates are generally small in terms of bias (Figures 8.6 and 8.7 display results for Binomial and Poisson distributed responses).
- For Binomial and Gaussian responses, the spatial function is recovered almost unbiased with both MRFs and GRFs (Figure 8.6 shows results for Binomial distributed responses).
- With GRFs some small regions with higher bias are found, where the GRF approach is not flexible enough to capture the high variability of the spatial function in this part of the map. With MRFs the problem is not encountered.
- For Bernoulli and Poisson distributed responses, the bias becomes more visible and the map is considerably oversmoothed (Figure 8.7 shows results for Poisson distributed responses).
- Considering average coverage probabilities, both the EB and the FB approach are somewhat below, but relatively close to the nominal levels when using MRFs for Binomial and Gaussian distributed responses (Table 8.2). Too narrow credible intervals are observed for Bernoulli and Poisson distributed responses.
- For results obtained with GRFs, differences become more visible: While the EB approach is generally too conservative (except for Bernoulli distributed responses), the FB approach usually leads to credible intervals which are too narrow.

- For Bernoulli distributed responses, all average coverage probabilities are below the nominal level.
- In terms of MSEs results obtained with MRFs are clearly preferable to those produced by GRFs for all response distributions (Figure 8.8).
- Differences between the EB and the FB approach are almost invisible in terms of MSEs.

		Bernoulli	Binomial	Poisson	Gaussian
MRF EB	80%	0.721	0.793	0.728	0.791
	95%	0.884	0.947	0.909	0.948
MRF FB	80%	0.752	0.797	0.753	0.792
	95%	0.920	0.948	0.928	0.946
Kriging EB	80%	0.778	0.996	0.940	1.000
	95%	0.924	1.000	0.972	1.000
Kriging FB	80%	0.671	0.743	0.712	0.751
	95%	0.855	0.903	0.875	0.915

Table 8.2: Wiggly regional data: Average empirical coverage probabilities. Values that are more than 2.5% below (above) the nominal level are indicated in green (red)

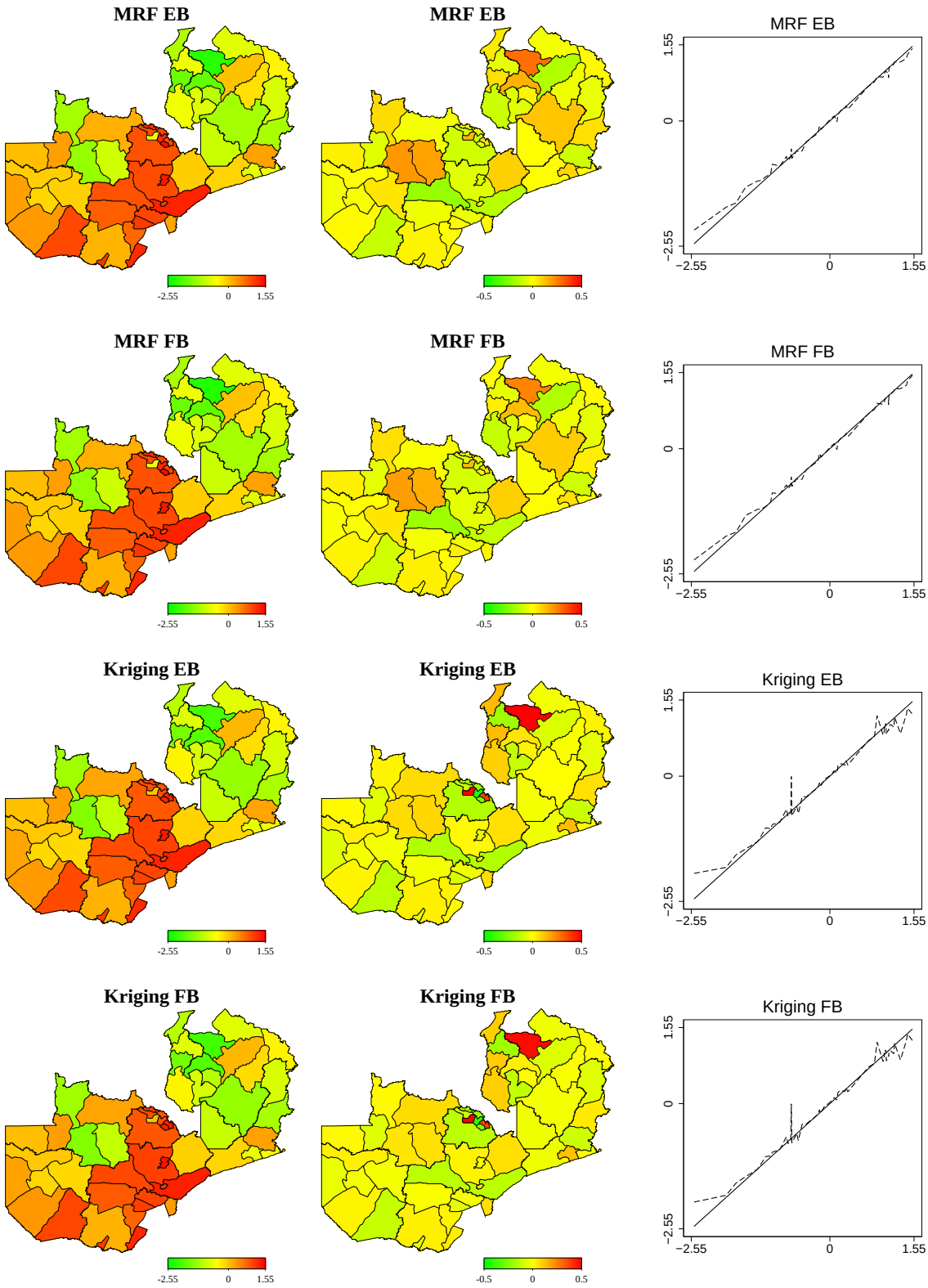


Figure 8.6: Wiggly regional data: Average estimates (left panel) and estimated bias (middle panel) for Binomial distributed responses. In the diagonal plots (right panel) average estimates are plotted against the ordered true values (dashed line). The reference curve obtained from plotting true values against true values (solid line) is included for comparison.

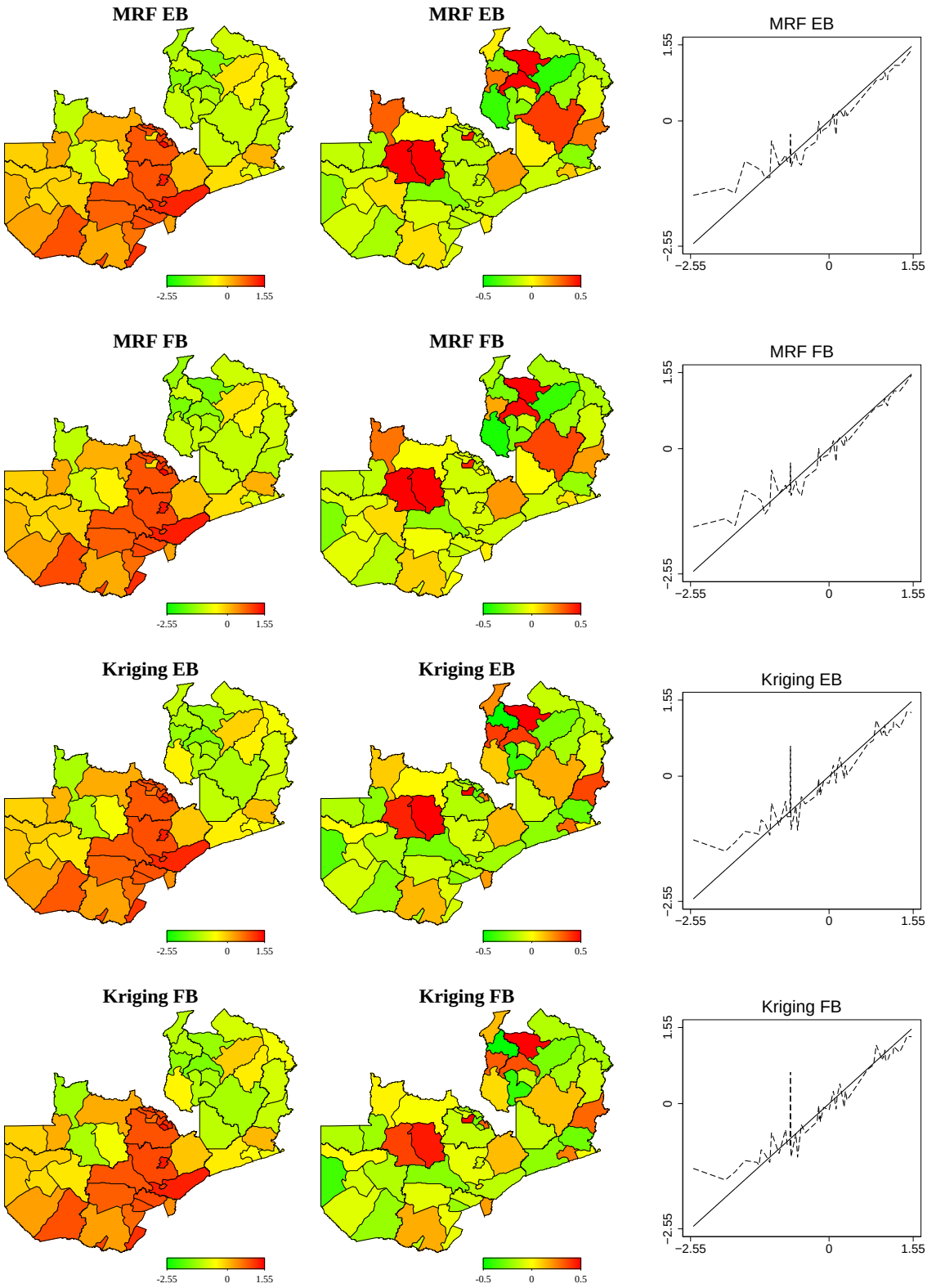


Figure 8.7: Wiggly regional data: Average estimates (left panel) and estimated bias (middle panel) for Poisson distributed responses. In the diagonal plots (right panel) average estimates are plotted against the ordered true values (dashed line). The reference curve obtained from plotting true values against true values (solid line) is included for comparison.

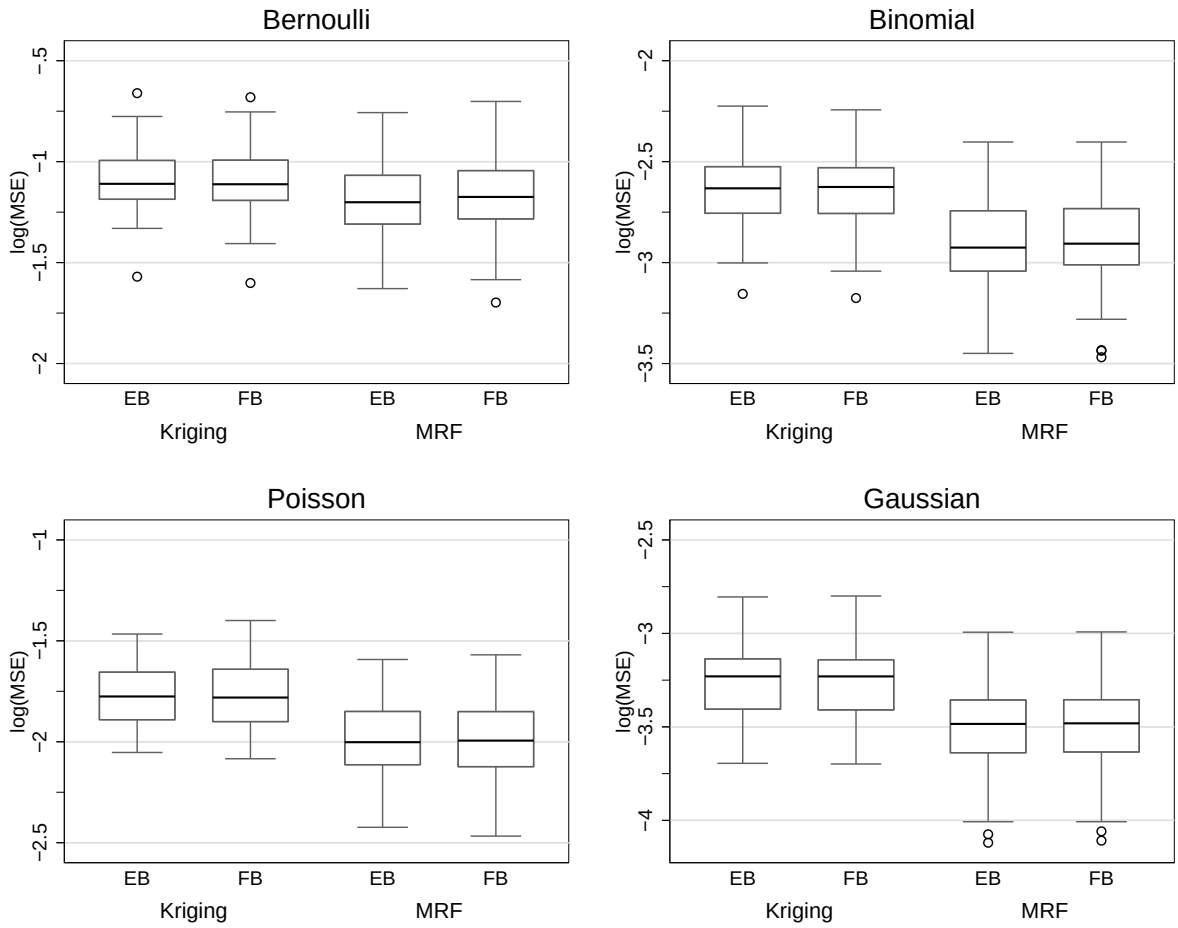


Figure 8.8: Wiggly regional data: Empirical $\log(MSE)$ for the different response distributions, inferential procedures, and spatial priors.

8.1.3 Extensions

Due to the findings in the two previous sections, some additional approaches and modifications have been explored.

8.1.3.1 Two-dimensional P-splines

For Binomial responses, bivariate P-splines defined on the centroids of the regions with first and second order random walk penalties based on Kronecker sums (see Section 4.2.6) were considered as competitors to MRFs and GRFs. Figure 8.9 displays the resulting MSEs both for the smooth and the wiggly function. For the smooth function, bivariate P-splines further improve the fit obtained with GRFs. The best results are achieved with a second order penalty, complementing findings for univariate P-splines, where an increased order of the penalty leads to smoother estimates. In contrast, for the wiggly function, results obtained with bivariate P-splines are dissatisfying due to a considerable amount of oversmoothing. Again second order P-splines lead to the smoothest results and, hence, exhibit the largest MSEs.

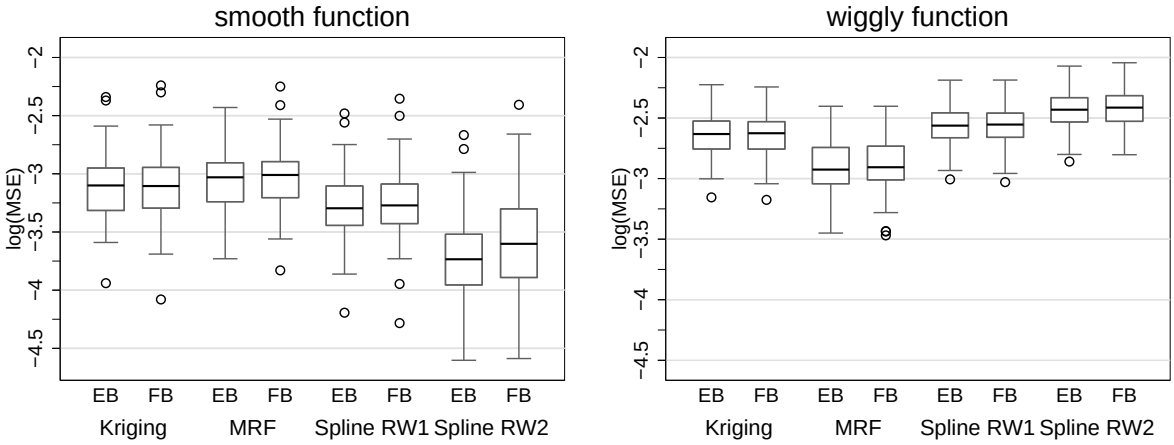


Figure 8.9: Regional data: Empirical $log(MSE)$ for Binomial distributed responses.

8.1.3.2 Nondifferentiable GRFs

In general, spatial smoothers resulting in differentiable surfaces seem to be inappropriate when analyzing data with a wiggly underlying function. This can be seen in the right part of Figure 8.9, where both GRFs and bivariate P-splines are outperformed by MRFs which naturally lead to discontinuous surface estimates. To investigate whether GRFs allowing for more variable, nondifferentiable estimates improve the fit in such situations, the analyses for Binomial distributed responses were rerun with a smaller value of the smoothness parameter of the GRF, i. e. $\nu = 0.5$. This corresponds to a GRF with exponential correlation function and results in continuous but not differentiable estimates. As indicated in Figure 8.10 such modified GRFs perform comparable to MRFs for the wiggly spatial function.

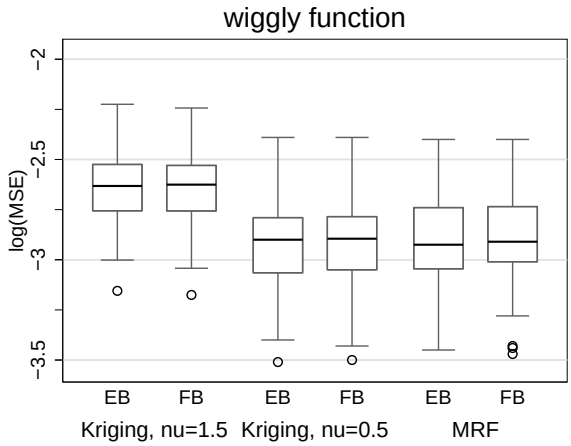


Figure 8.10: Wiggly regional data: Empirical $log(MSE)$ for Binomial distributed responses.

8.1.3.3 Weighted MRFs

As a last extension, we considered Markov random fields with weights inverse proportional to the distance of the centroids as defined in Section 4.2.3.1 instead of unweighted MRFs. The idea is that additional information is included resulting in an improved fit for the smooth spatial function. Figure 8.11 shows the resulting MSEs for Binomial and Gaussian distributed responses. Obviously, the MSEs decrease with weighted MRFs resulting in estimates of comparable quality as with GRFs.

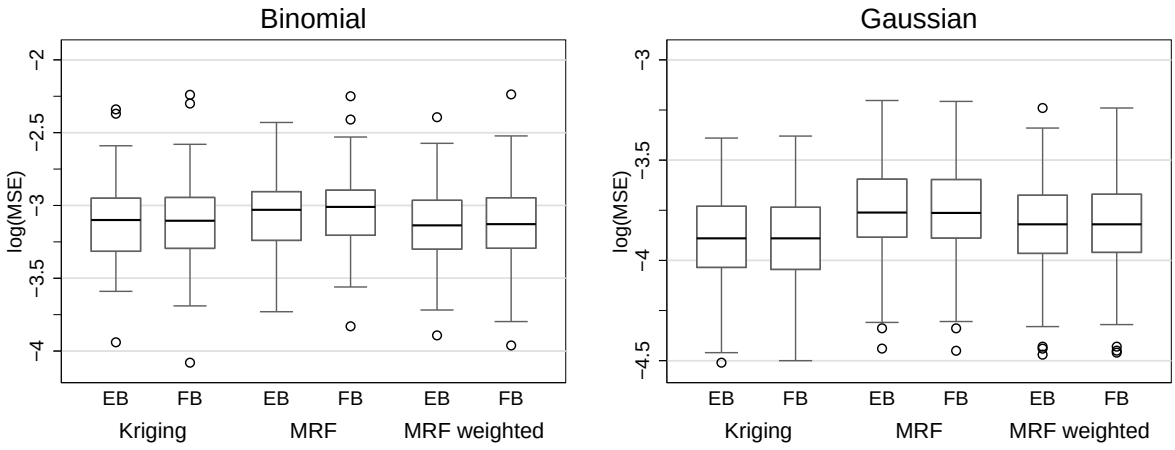


Figure 8.11: Smooth regional data: Empirical $log(MSE)$ for Binomial and Gaussian distributed responses.

8.2 Continuous spatial information

In a second simulation study, continuous spatial information given in terms of longitude and latitude of the corresponding observations was considered. Here, the regression model is given by a spatial predictor of the form

$$\eta_i = \beta_0 + f_{spat}(s_i),$$

where $s = (s_x, s_y)$ indicates the coordinates of the respective observation. Based on the coordinates, the spatial function (see Figure 8.12) was defined as

$$f(s_x, s_y) = 1.9 \cdot \left[1.35 + \exp(s_x) \cdot \sin(13 \cdot (s_x - 0.6)^2) \cdot \exp(-s_y) \cdot \sin(7 \cdot y)\right] - 3.5.$$

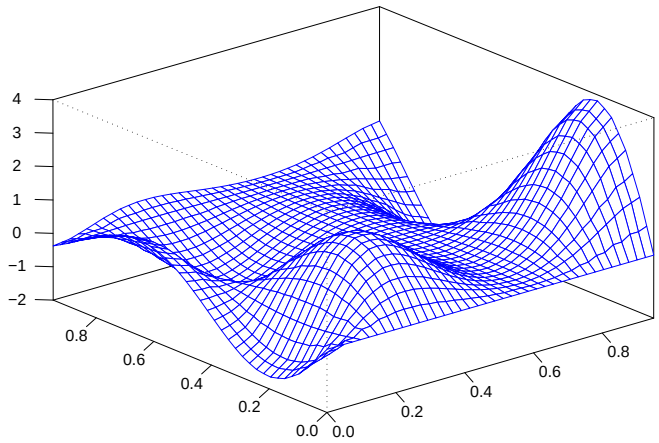


Figure 8.12: Lattice data: True spatial function.

We simulated $R = 100$ data sets consisting of $n = 400$ observations for each of the four response distributions (Bernoulli, Binomial, Poisson, and Gaussian). For Poisson distributed responses, the spatial function had to be scaled with a factor of 0.25 since otherwise too extreme observations were encountered. The x and y coordinates of the observations are given by a 20×20 grid of equidistant points between 0 and 0.95, so each point appeared exactly one time in the data set.

We compared the following spatial models:

- Assume a low rank Kriging term with 100 knots and $\nu = 1.5$ for the spatial effect and estimate the model based on mixed model methodology.
- Discretize the observation area into clusters formed by four grid points. Assume a Markov random field prior for the clusters, where the four next clusters are used as neighbors and estimate the model based on mixed model methodology. The discretization had to be used since a MRF based on the grid points themselves was found to be numerically to extensive in combination with mixed model methodology.
- Assume a two-dimensional P-spline with first order random walk prior (see Section 4.2.6.1) and estimate the model based on mixed model methodology.
- Assume a two-dimensional P-spline with a Kronecker sum second order random walk prior (see Section 4.2.6.2) and estimate the model based on mixed model methodology.

- Assume a low rank Kriging term with 100 knots and $\nu = 1.5$ for the spatial effect and estimate the model based on MCMC with hyperparameters $a = b = 0.001$ for the variance τ^2_{spat}.
- Assume a Markov random field prior defined upon the grid points, where the four next points on the grid are used as neighbors and estimate the model based on MCMC with hyperparameters $a = b = 0.001$ for the variance τ^2_{spat}.
- Assume a two-dimensional P-spline with first order random walk prior and estimate the model based on MCMC with hyperparameters $a = b = 0.001$ for the variance τ^2_{spat}.
- Assume a two-dimensional P-spline with a Kronecker sum second order random walk prior and estimate the model based on MCMC with hyperparameters $a = b = 0.001$ for the variance τ^2_{spat}.

Some further extensions and modifications including two-dimensional P-splines with priors based on approximating the biharmonic differential operator (Section 4.2.6.4) will be considered in Section 8.2.1.

In analogy to the measures defined in Section 8.1.1 we computed average estimates and empirical bias, average empirical coverage probabilities, and empirical mean squared errors to compare the estimates. The results can be summarized as follows:

- Almost unbiased estimates are obtained with Gaussian responses and all estimation approaches (results not shown).
- For Binomial distributed responses the bias is generally small but visible for observations with extreme values of the spatial function. Two-dimensional P-splines with second order random walk penalty lead to the estimates with the smallest bias. Figure 8.13 shows results obtained with MRFs and P-splines with RW2 penalty.
- In case of Poisson and Bernoulli distributed responses the surface is recovered only dissatisfactory due to a considerable amount of oversmoothing. For Bernoulli distributed responses, the best results are obtained with MRFs estimated by MCMC. Figure 8.14 displays results for Bernoulli distributed responses obtained with GRFs and MRFs.
- FB estimates for Markov random fields are generally more wiggly than EB estimates, since they are defined upon the grid points themselves and not on a discretized version.
- Otherwise differences between EB and FB approaches are small in terms of bias.
- Considering empirical coverage probabilities, no general conclusion holding for all response distributions can be drawn (Table 8.3).
- For Bernoulli distributed responses, average coverage probabilities are usually relatively close to the nominal values except for FB Markov random fields, where the credible intervals are too conservative and two-dimensional P-splines with second order random walk prior, where the credible intervals are too narrow.
- Binomial distributed responses lead to average coverage probabilities above the nominal level for all approaches. FB MRFs and EB GRFs result in credible intervals

that are far too conservative.

- For Poisson distributed responses, problems are observed for the two-dimensional P-splines. Due to the oversmoothed estimates, average coverage probabilities are considerably below the nominal levels both with EB and FB estimates.
- In case of Gaussian responses, most average coverage probabilities are above but relatively close to the nominal levels. Exceptions are FB MRFs which are too narrow and EB GRFs, where the credible intervals are too conservative.
- In terms of MSE, differences between the empirical Bayes and the fully Bayesian approach are generally small (Figure 8.15). Exceptions are the results obtained with Markov random fields, which are most likely caused by the different definitions for EB and FB inference (see also Section 8.2.1), and two-dimensional P-splines with second order random walk prior.
- The fully Bayesian Markov random field approach leads to very wiggly estimates and therefore yields rather high MSEs although having a relatively small bias. Further investigation revealed that using a discretized version of the MRF as with the empirical Bayes approach also leads to improved estimates in terms of MSE (see also Section 8.2.1).
- If there is enough information in the data, two-dimensional P-splines with second order random walk prior lead to the best estimates (Binomial and Gaussian distributed responses). However, if the signal to noise ratio is low, as it is especially for Poisson distributed responses, a considerable amount of oversmoothing is observed and, hence, the resulting estimates perform poorly in terms of MSE. For Bernoulli distributed responses the empirical Bayes approach still leads to satisfactory results while the fully Bayesian approach is far off when using two-dimensional P-splines with second order random walk prior.
- The low rank Kriging estimates shows almost no differences for EB and FB estimates. They perform preferable to estimates obtained with MRFs and first order random walk P-splines.

		Bernoulli	Binomial	Poisson	Gaussian
MRF EB	80%	0.785	0.813	0.819	0.858
	95%	0.912	0.951	0.921	0.975
MRF FB	80%	0.920	0.903	0.848	0.312
	95%	0.990	0.987	0.950	0.534
Kriging EB	80%	0.853	0.968	0.851	0.99
	95%	0.949	0.994	0.934	0.999
Kriging FB	80%	0.798	0.837	0.804	0.822
	95%	0.939	0.962	0.937	0.958
Spline RW1 EB	80%	0.804	0.88	0.808	0.850
	95%	0.926	0.976	0.907	0.972
Spline RW1 FB	80%	0.838	0.885	0.720	0.849
	95%	0.961	0.982	0.872	0.972
Spline RW2 EB	80%	0.764	0.867	0.640	0.855
	95%	0.912	0.975	0.797	0.973
Spline RW2 FB	80%	0.638	0.867	0.542	0.852
	95%	0.821	0.976	0.729	0.972

Table 8.3: Lattice data: Average empirical coverage probabilities. Values that are more than 2.5% below (above) the nominal level are indicated in green (red)

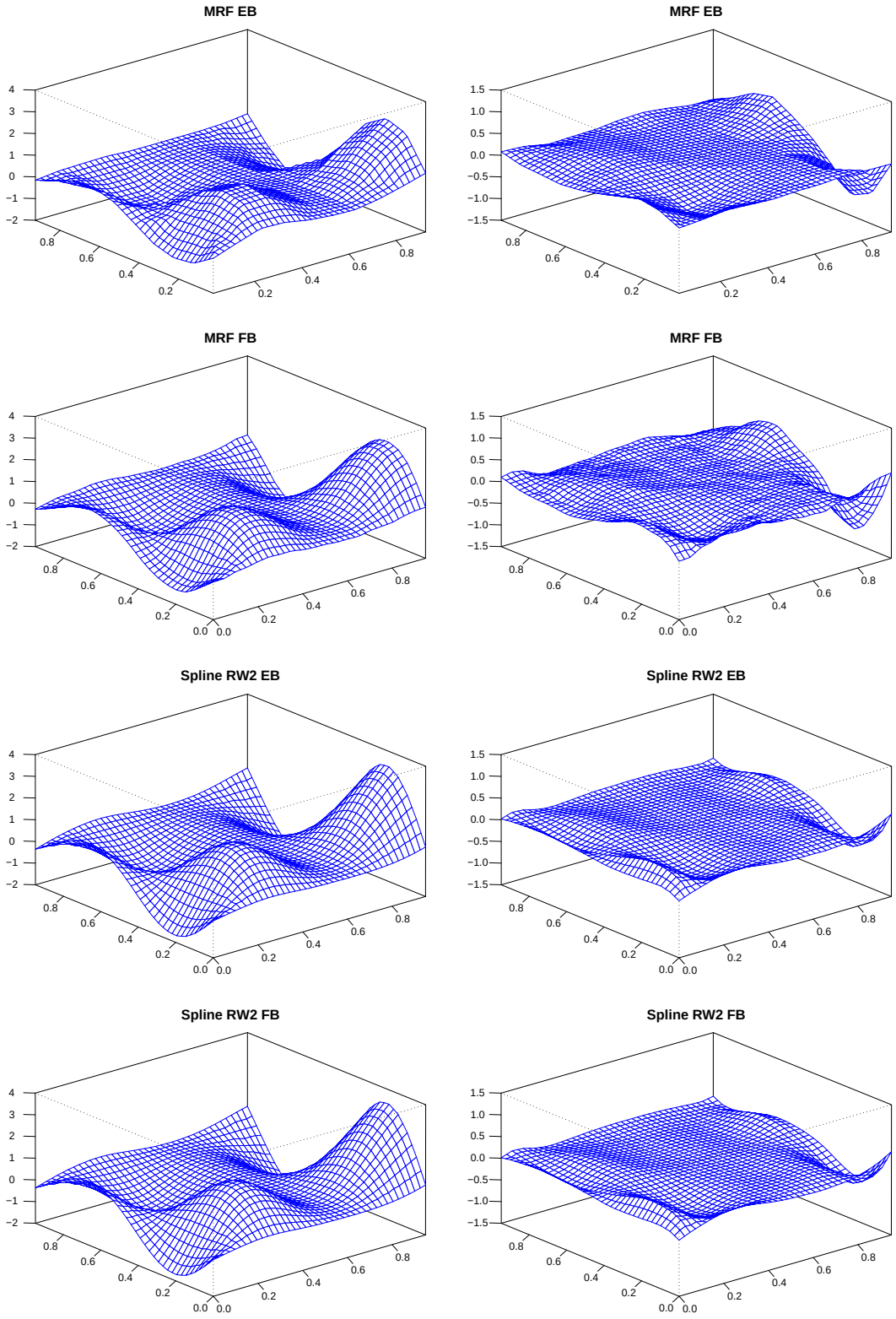


Figure 8.13: Lattice data: Average estimates (left panel) and estimated bias (right panel) for Binomial distributed responses.

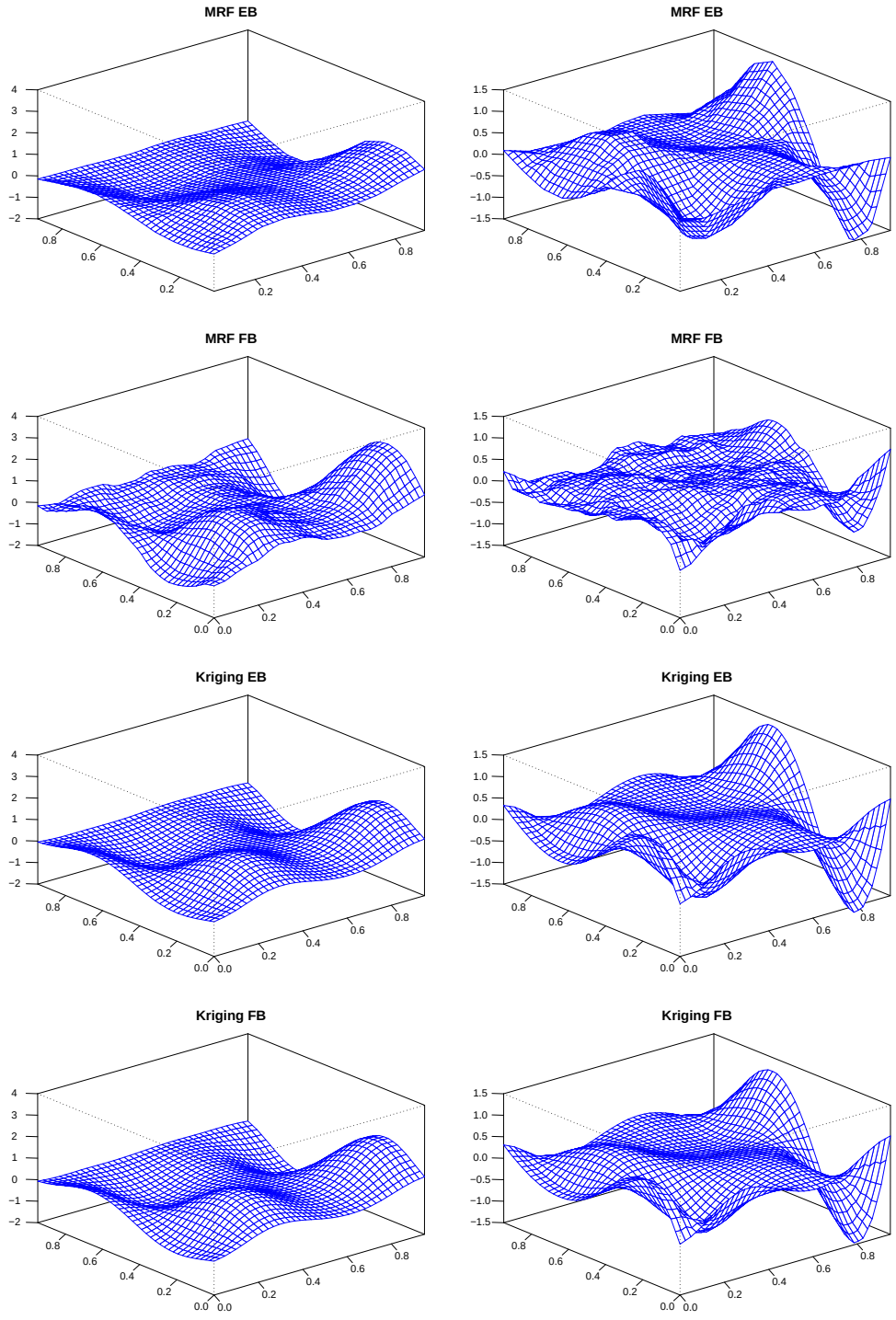


Figure 8.14: Lattice data: Average estimates (left panel) and estimated bias (right panel) for Bernoulli distributed responses.

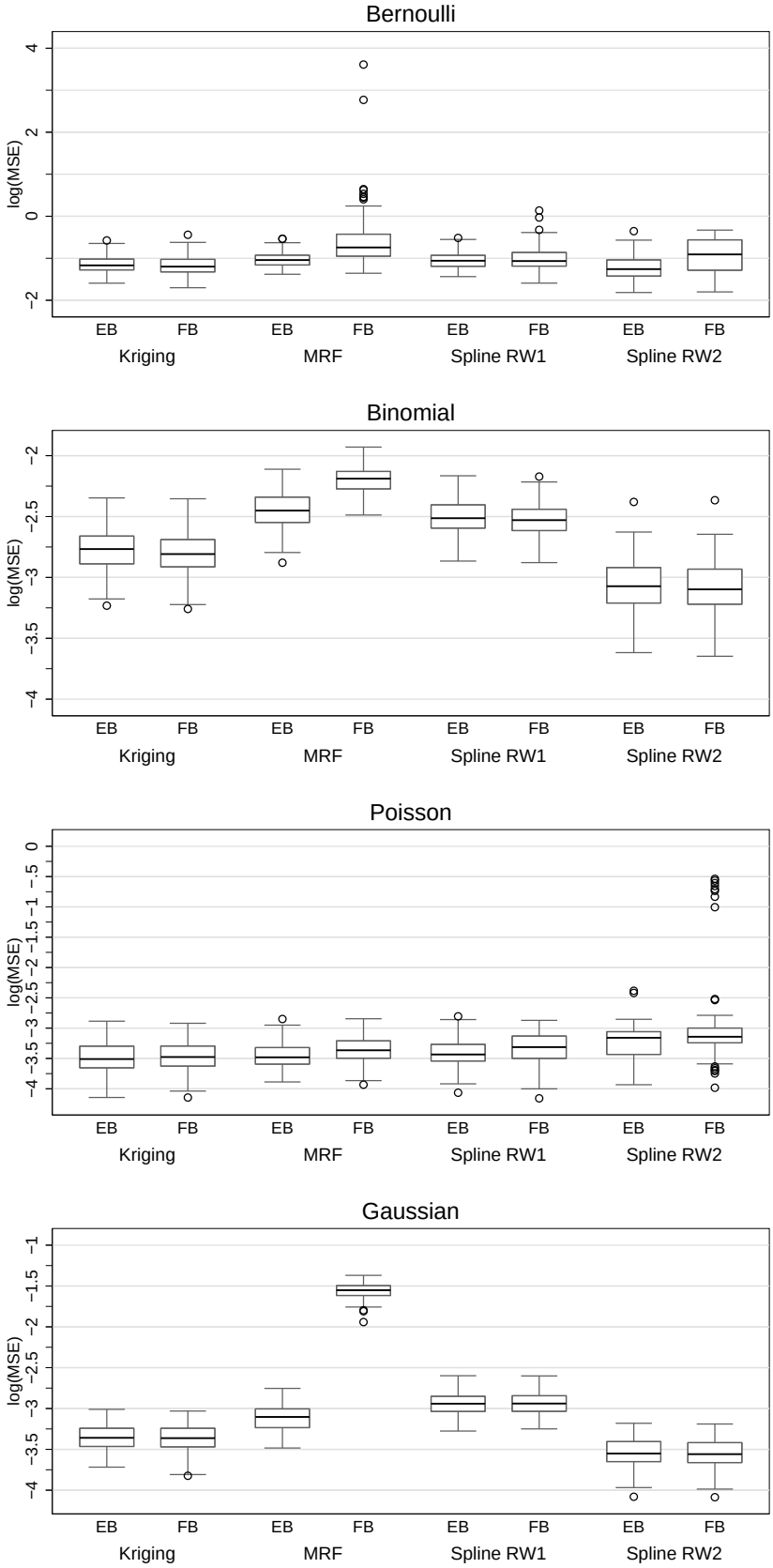


Figure 8.15: Lattice data: Empirical $\log(MSE)$ for the different response distributions, inferential procedures, and spatial priors.

8.2.1 Extensions

8.2.1.1 Markov random fields

As already mentioned in the comparison in the previous section, results obtained with Markov random fields were completely different for EB and FB inference. Here, we want to investigate if these differences are caused by the inferential techniques or by the distinct definitions of the MRFs. Therefore, we considered two further definitions for MRFs in combination with fully Bayesian inference and Gaussian distributed responses. These were a discretized version, where the MRF is defined upon clusters formed by four grid points (as in the definition for EB inference in the previous section), and a version, where eight instead of four next neighbors are employed.

In Figure 8.16, the corresponding MSEs are visualized. Obviously, using the discretized version of the Markov random field leads to smoother, more reasonable estimates both in an empirical and a fully Bayesian analysis. In contrast, increasing the number of neighbors did not have the expected impact. Results obtained with this type of MRFs are even worse than those using only four neighbors.

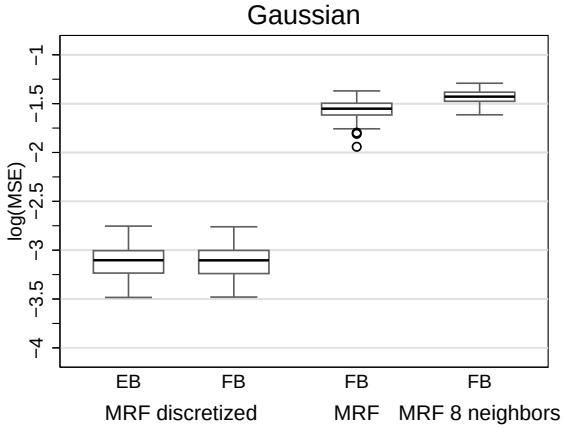


Figure 8.16: Lattice data: Empirical log(MSE) for Gaussian distributed responses obtained with different types of Markov random fields.

8.2.1.2 Approximation of the biharmonic differentiable operator

An alternative to the definition of penalties for bivariate P-splines based on Kronecker sums was introduced in Section 4.2.6.4 using an approximation to the biharmonic differential operator. Applying this kind of penalty to Bernoulli and Gaussian responses yields the MSEs shown in Figure 8.17. In both cases, the P-spline with biharmonic penalty performes comparable to the two other versions of bivariate penalized splines.

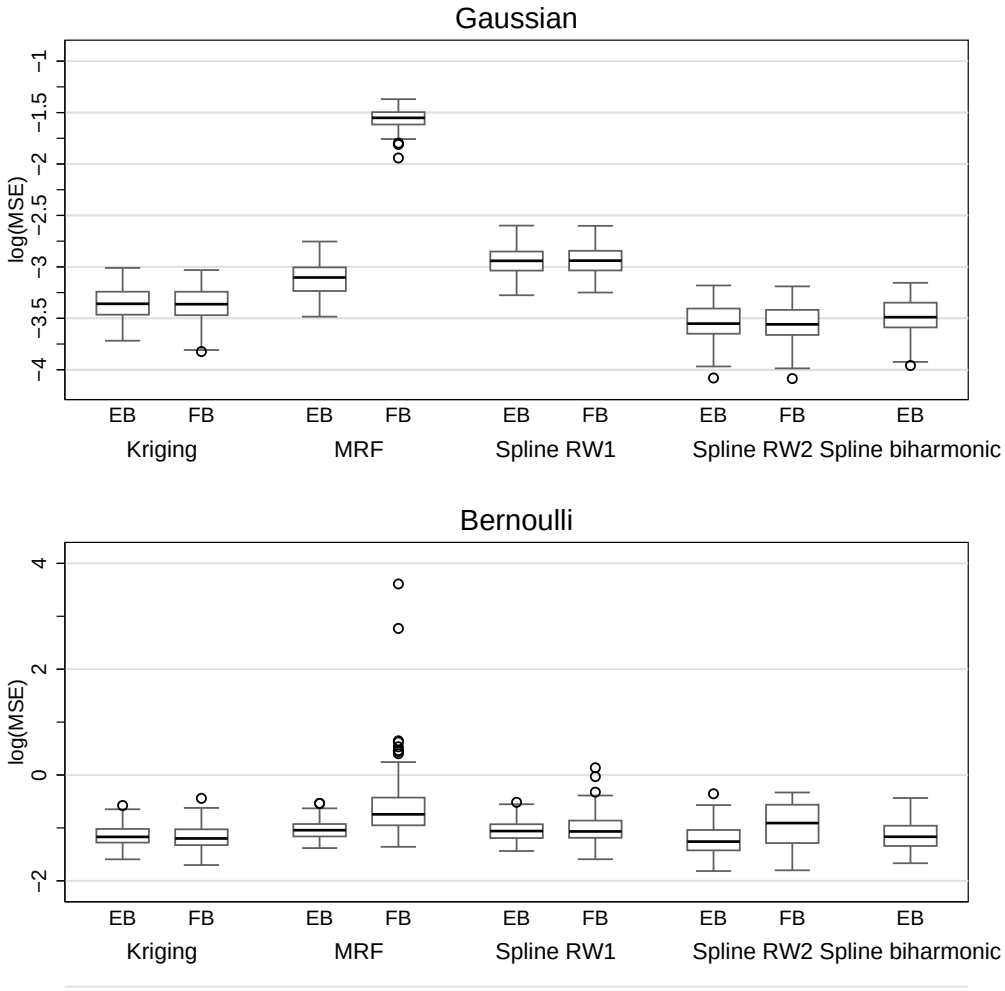


Figure 8.17: Lattice data: Empirical $log(MSE)$ for Gaussian and Bernoulli distributed responses.

9 A simulation study in spatio-temporal longitudinal data

9.1 Simulation setup

The present simulation study aims at imitating typical spatio-temporal longitudinal data. To assess the impact of information contained in different types of responses, the following study is based on binary, binomial (with three repeated binary observations), Poisson, and Gaussian regression models. In each case, data were generated from logit, loglinear and additive models using the structured additive predictor

$$\eta_{it} = f_1(x_{it1}) + f_{spat}(s_{it}) + b_{i1} + b_{i2}x_{it2} + b_{i3}x_{it3} + \gamma_1 x_{it2} + \gamma_2 x_{it3}$$

for $i = 1, \ldots, 24$ individuals and $t = 1, \ldots, 31$ repeated measurements, resulting in 744 observations per simulation run. The function f_1 is a sine function, and f_{spat} is a spatial function with linearly increasing trend defined upon the $s = 1, \ldots, 124$ districts of Bavaria and Baden-Württemberg, the two southern states of Germany (see Figure 9.1). The parameters b_{i1}, b_{i2} and b_{i3} are i.i.d individual-specific Gaussian random effects. From a classical perspective b_{i1} is a random intercept, b_{i2} and b_{i3} represent random slopes. The effects γ_1, γ_2 are usual fixed effects.

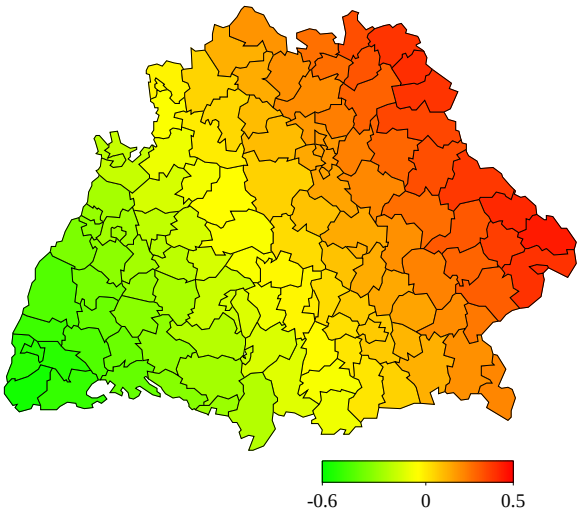


Figure 9.1: True spatial function $f_{spat}(s)$.

For the covariate x_1, values were randomly drawn from 186 equidistant gridpoints between -3 and +3. Each gridpoint was randomly assigned four times. Similarly, values for the covariates x_2 and x_3 were drawn from 186 equidistant gridpoints between -1 and +1. The function f_{spat} has 124 different values; each value was randomly assigned 6 times. The i.i.d Gaussian (random) effects were obtained as drawings

$$\begin{aligned} b_{i1} &\sim N(0; 0.25), \\ b_{i2} &\sim N(0; 0.25), \quad i = 1, \ldots, 24, \\ b_{i3} &\sim N(0; 0.36). \end{aligned}$$

Keeping the resulting 744 predictor values η_{it}, $i = 1, \ldots, 24$, $t = 1, \ldots, 31$, fixed, binary, binomial, Poisson, and Gaussian responses were generated using logit, loglinear Poisson

and additive Gaussian models, respectively. For each model, the simulation was repeated over 250 such simulation runs, producing responses $y_{it}^{(r)}$, $r = 1, \ldots, 250$, for the predictor. For additive Gaussian models, the errors are i. i. d. drawings from $N(0; 0.25)$.

Using these artificial data, we compared performance in terms of bias, MSE and average coverage probabilities. For f_1 we assumed a cubic P-spline with second order random walk prior, and for the spatial effect f_{spat} the MRF prior (4.25). The models were estimated based on either mixed model methodology, yielding empirical Bayes (EB) estimates or using Markov Chain Monte Carlo simulation techniques resulting in full Bayesian (FB) estimates. For MCMC inference, we compared two different choices for the hyperpriors of the variance parameters, namely $a = 1, b = 0.005$ and $a = b = 0.001$.

9.2 Results

A general, but not surprising conclusion is that bias and MSE tend to decrease with increasing information contained in the responses, i. e., when moving from binary responses to Poisson or Gaussian responses. A further observation is that the REML estimate has convergence problems in about 25% of the analyzed models. In the case of no convergence, usually only one of the variance components switched between two values which were close to each other, while iterations converged for the remaining variance components. A closer inspection of estimates with and without convergence showed that differences in terms of MSE can be neglected and the arbitrary choice of one of the two switching values leads to reasonable estimates. Therefore, it is justified to use the final values after the maximum number of iterations (400) to compute empirical MSEs, bias, average coverage probabilities, etc.

The true sine curve f_1 and the average estimates obtained from all 250 posterior estimates, $r = 1, \ldots, 250$, are visually very close for all four observation models. As examples we show average estimates of f_1 for binary and Gaussian responses in Figure 9.2, since these two response distributions result in the worst and best fit, respectively. Since both choices for the hyperpriors led to almost indistinguishable results, we only show average estimates obtained with $a = b = 0.001$ for FB estimates. Obviously, the FB estimates are somewhat closer to the true curve for Binary responses but these differences vanish rapidly when switching to response distributions containing more information.

For binary responses, averages of posterior estimates and empirical bias for f_{spat}, are displayed in Figure 9.3. We conclude the following: At least for binary observations, the often recommended standard choice $a = 1, b = 0.005$ for hyperparameters of inverse Gamma priors for smoothing parameters results in oversmoothing of the spatial effect, whereas FB inference with $a = b = 0.001$ and EB inference perform considerably better and with comparable bias.

For Poisson responses (lower panel of Figure 9.4), the bias becomes smaller and the true surface is recovered satisfactorily both with full and empirical Bayes estimation. Also, differences for the different choices of hyperparameters are no longer present and we therefore excluded results with $a = 1$ and $b = 0.005$. Estimation properties for binomial observations (upper panel of Figure 9.4) are between results for binary and Poisson models. For Gaussian observations (not shown), we obtain the best results, and EB and FB results

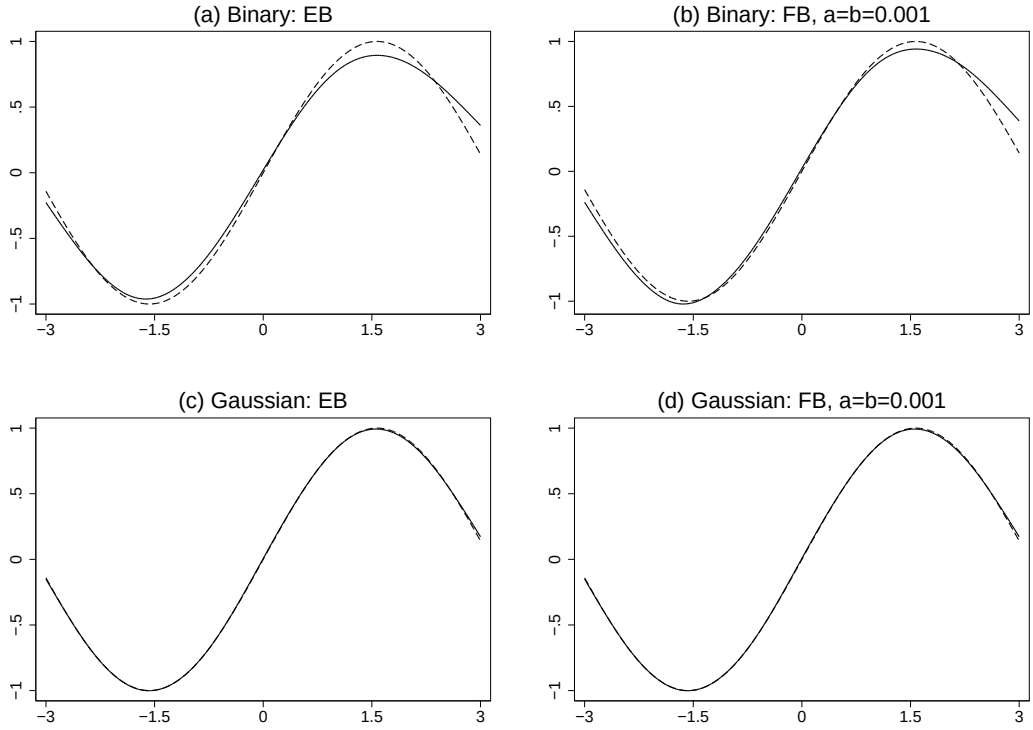


Figure 9.2: Average estimates (solid line) and true values (dashed line) for the nonparametric effect $f_1(x_1)$.

almost coincide.

In principle, results for the estimated random effects b_1, b_2 and b_3 indicate similar conclusions as for the spatial effect. For Gaussian distributed responses, all three approaches yield almost unbiased estimates (results not shown). Binomial and Poisson responses introduce some bias, especially when considering the random slopes b_2 and b_3, but still results are almost the same regardless of the utilized method (see Figure 9.5 for results obtained with EB inference and FB inference with hyperparameters $a = b = 0.001$). For binary responses, considerably biased estimates are obtained. Again the hyperparameter choice $a = 1$, $b = 0.005$ lead to the worst estimates due to oversmoothing, while the two remaining methods perform comparable (results not shown).

Figures 9.6 and 9.7 show empirical MSEs for the sine curve f_1 and the spatial effect f_2, averaged over all covariate values, and for the random effects averaged over the individuals $i = 1, \ldots, 24$. From these figures we see that generally EB estimation behaves remarkably well in terms of MSEs when compared to FB inference. While differences are relatively small for Poisson and Gaussian responses, they become more visible for binary and Binomial responses. For binary responses, the hyperparameter choice $a = 1, b = 0.005$ implies highest MSE for the spatial effect, and also for the random intercept. Considering the random slopes b_{i2} and b_{i3}, comparable estimates are obtained with EB inference and FB inference with hyperparameters $a = b = 0.001$. The second set of hyperparameters leads to estimates which have considerably higher MSE, especially for binary responses.

Average coverage probabilities of pointwise credible intervals for a nominal level of 95%

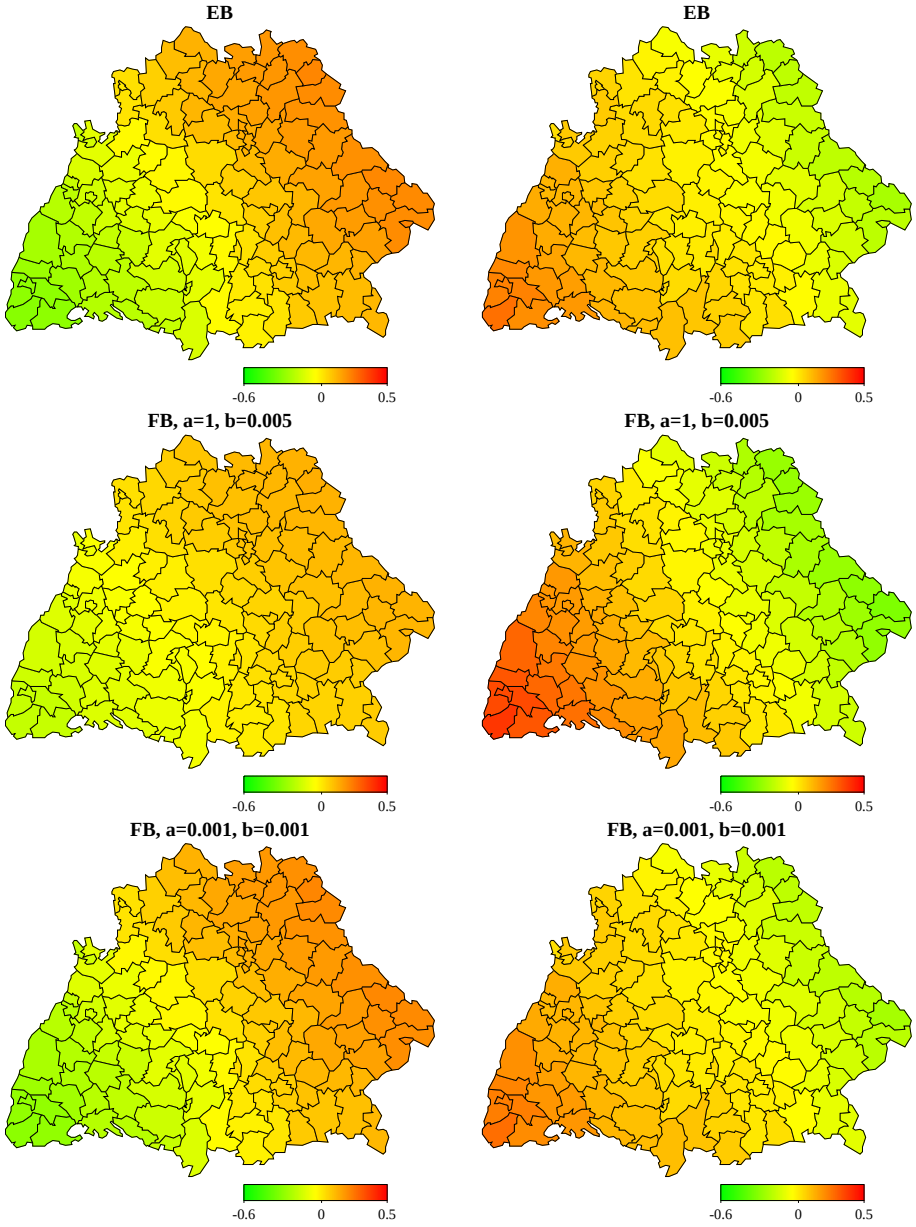


Figure 9.3: Binary responses: Average estimates (left panel) and empirical bias (right panel) for $f_{spat}(s)$.

are shown in Table 9.1 for the different effects. The results provide some evidence for the following: All three Bayesian approaches have comparable coverage properties for Gaussian and Poisson responses. For binary responses, some differences can be seen. While the average coverage probabilities are still quite acceptable for the nonparametric function f_1, they are partly considerably below the nominal level of 95% for the spatial effect f_{spat} and the i.i.d effects b_1, b_2 and b_3. Only FB inference with $a = b = 0.001$ gives satisfactory results. For Binomial responses still some difficulties are observed, but average coverage probabilities are relatively close to the nominal level for all three approaches. For the spatial effect, credible intervals are mostly too conservative, irrespective of the specific inferential procedure.

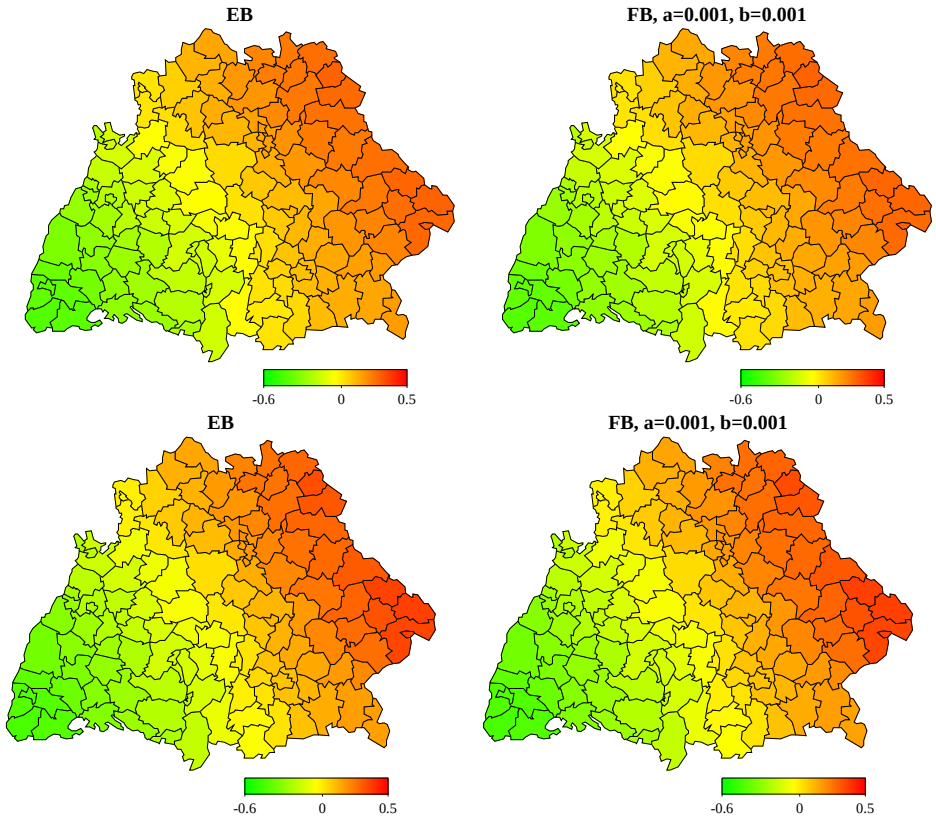


Figure 9.4: Binomial (upper panel) and Poisson (lower panel) responses: Average estimates for $f_{spat}(s)$.

The final comparison concerns estimation of variance components of the random effects b_{i1}, b_{i2} and b_{i3}. For each type of response, Table 9.2 compares averages of estimates with the "empirical" variances, obtained from the 24 i. i. d. drawings from the corresponding normals. A comparison with these empirical variances is fairer than with "true" values (given in brackets). For Gaussian responses, FB estimates with $a = b = 0.001$ have larger bias than EB and FB estimates with $a = 1$, $b = 0.005$. For binary responses, on the

	distribution	f_1	f_{spat}	b_1	b_2	b_3
EB	Gaussian	0.980	0.993	0.993	0.976	0.986
	Bernoulli	0.967	0.900	0.915	0.723	0.854
	Binomial	0.975	0.99	0.963	0.914	0.947
	Poisson	0.980	0.998	0.971	0.949	0.970
FB ($a = 1, b = 0.005$)	Gaussian	0.971	0.996	0.993	0.975	0.985
	Bernoulli	0.958	0.884	0.856	0.568	0.67
	Binomial	0.970	0.984	0.962	0.861	0.932
	Poisson	0.974	0.998	0.973	0.946	0.969
FB ($a = b = 0.001$)	Gaussian	0.973	0.996	0.995	0.978	0.989
	Bernoulli	0.971	0.985	0.935	0.883	0.909
	Binomial	0.971	0.995	0.969	0.926	0.959
	Poisson	0.973	0.998	0.976	0.955	0.973

Table 9.1: Average coverage probabilities for the different effects based on a nominal level of 95%. Values that are more than 2.5% below (above) the nominal level are indicated in green (red).

		emp. value (true value)	bias			
			Gaussian	Bernoulli	Binomial	Poisson
EB	b_{i1}	0.196 (0.25)	0.010	-0.014	0.003	-0.005
	b_{i2}	0.226 (0.25)	0.006	-0.047	-0.014	-0.006
	b_{i3}	0.329 (0.36)	0.017	-0.029	-0.003	0.007
FB	b_{i1}	0.196 (0.25)	0.009	-0.066	0.002	-0.001
$(a = 1, b = 0.005)$	b_{i2}	0.226 (0.25)	0.001	-0.177	-0.070	-0.019
	b_{i3}	0.329 (0.36)	0.013	-0.215	-0.032	-0.004
FB	b_{i1}	0.196 (0.25)	0.030	0.024	0.039	0.026
$(a = b = 0.001)$	b_{i2}	0.226 (0.25)	0.028	-0.019	0.014	0.020
	b_{i3}	0.329 (0.36)	0.051	0.024	0.057	0.048

Table 9.2: Average bias of the variance components.

other side, FB estimates with $a = 1$, $b = 0.005$ have considerable bias. For binomial and Poisson responses, differences between the two FB versions are less distinct, but EB estimates are mostly better. A conclusion emerging from these results is that REML estimates of variance components are preferable in terms of bias.

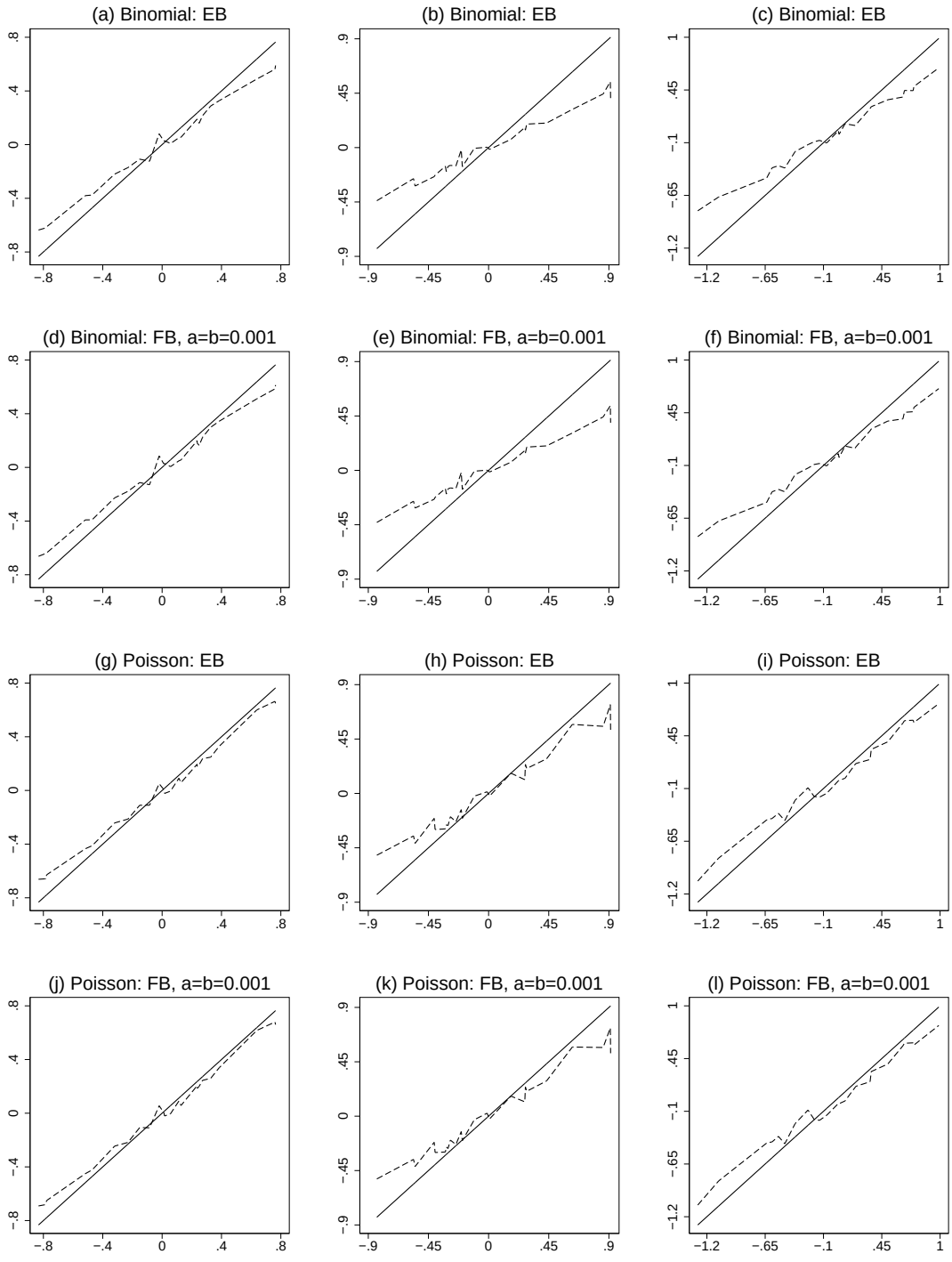


Figure 9.5: Average estimates for the random intercept b_1 (left panel) and the random slopes b_2 (middle panel) and b_3 (right panel). Average estimates are plotted against the ordered true values (dashed line). The reference curve obtained from plotting true values against true values (solid line) is included for comparison

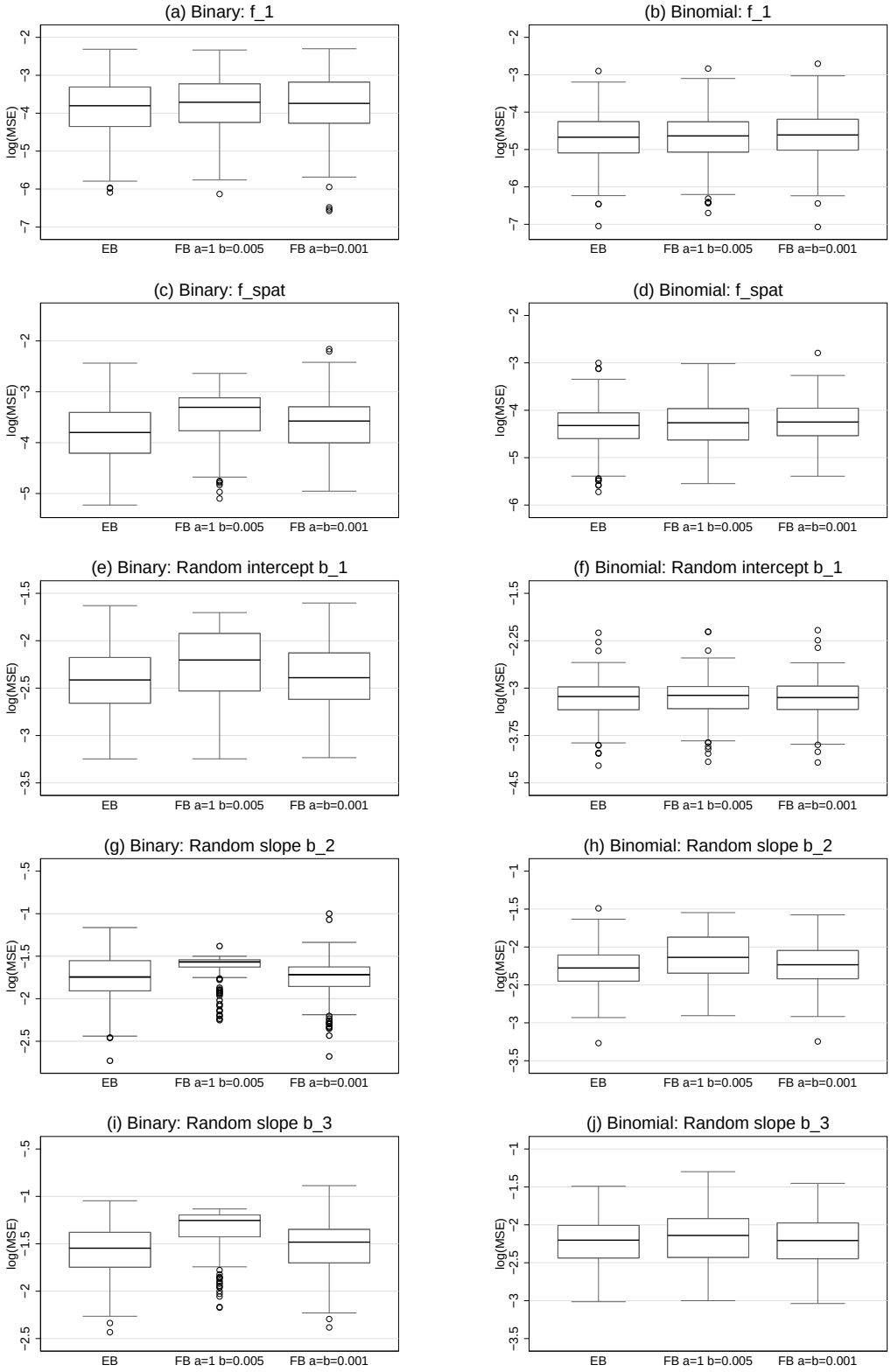


Figure 9.6: Binary (left panel) and Binomial (right panel) responses: Boxplots for log(MSE).

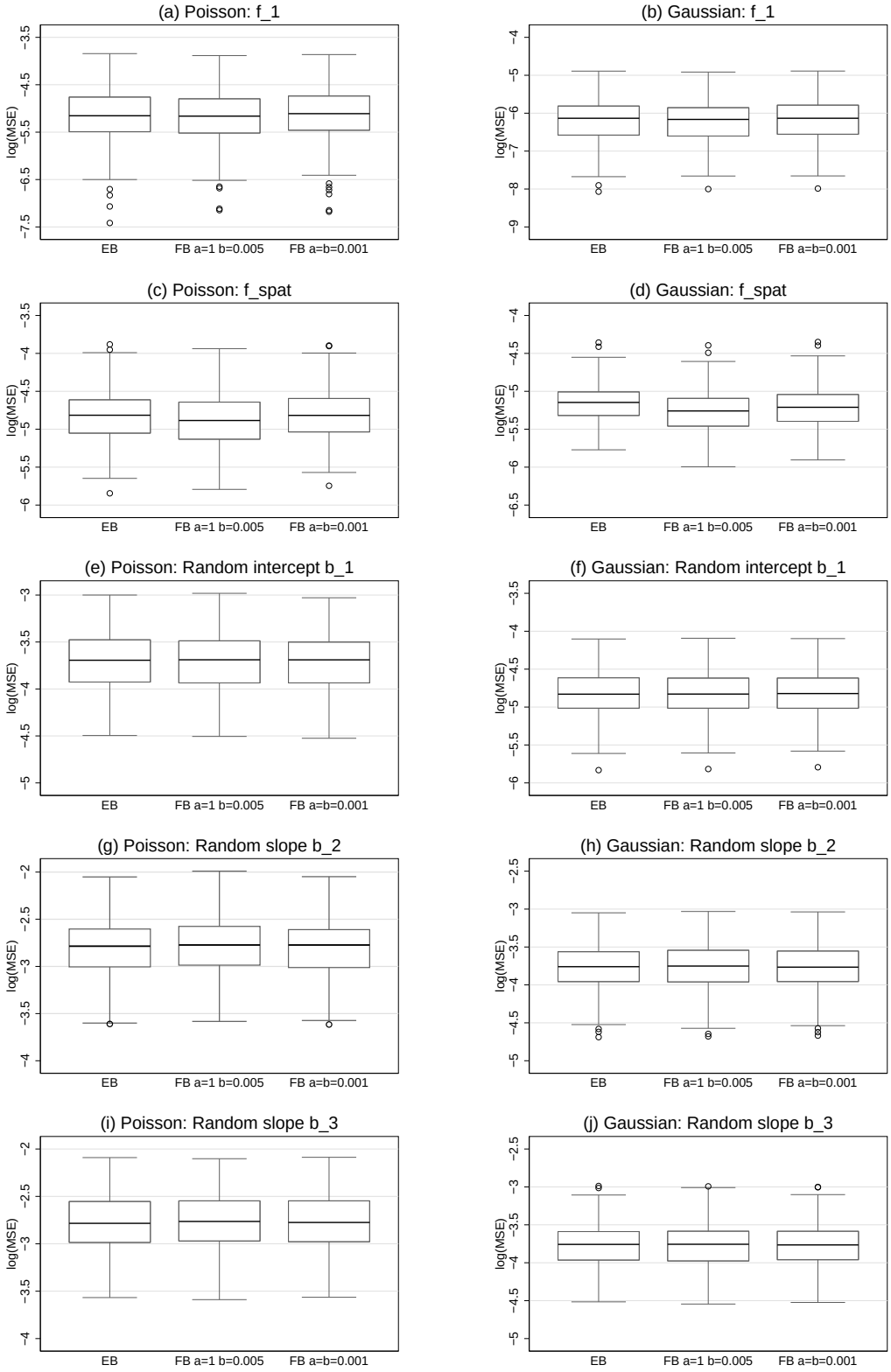


Figure 9.7: Poisson (left panel) and Gaussian (right panel) responses: Boxplots for log(MSE).

Part III

Multicategorical responses

10 Model formulation

10.1 Observation model

The models discussed in Part II are only appropriate for the analysis of univariate responses, such as Bernoulli, binomial, Poisson or Gaussian distributed responses. A more complicated situation arises when multivariate extensions of regression models are considered. In the following, we will focus on one important special case of multivariate regression, where the response variable $Y \in \{1, \ldots, k\}$ is categorical. In fact, the multivariate regression problem is established by the representation of Y in terms of k indicator variables $y^{(r)}$ with

$$y^{(r)} = \begin{cases} 1 & Y = r, \\ 0 & \text{else.} \end{cases} \tag{10.1}$$

Regression models for categorical responses are then formulated based on this multivariate representation of the data. Expressing categorical data in this form is particularly suited for considering covariates that change over the categories $\{1, \ldots, k\}$ or when modeling effects that affect the probability of a certain category in a category-specific way. Here, different regression models are imposed on each of the indicator variables instead of applying one overall model to the categorical response Y.

As noticeable from (10.1), one of the indicator variables is redundant since its value can be inferred from the remaining $q = k - 1$ variables. Therefore, category k is usually considered as the reference category and not further taken into account. Collecting all the remaining indicator variables in a vector $y = (y^{(1)}, \ldots, y^{(q)})'$, leads to the multinomial distribution, i. e. $y \sim M(1, \pi)$, with $\pi = (\pi_i^{(1)}, \ldots, \pi_i^{(q)})'$ being the q-dimensional vector of probabilities

$$\pi^{(r)} = P(Y = r) = P(y^{(r)} = 1), \quad r = 1, \ldots, q.$$

Note that the indicator variables are not independent but negatively correlated, since $y^{(r)} = 1$ requires all other indicators to be zero. Therefore, we truly are in a multivariate framework when analyzing categorical responses.

The rest of this section is organized as follows: In Section 10.1.1 we will briefly introduce multivariate generalized linear models as a general framework for multivariate regression and point out how categorical response models can be cast into this framework. Sections 10.1.2 to 10.1.4 discuss different types of categorical regression models in greater detail and also contain extensions to structured additive regression models similarly to the univariate case. Models for nominal responses with unordered categories will be discussed in Section 10.1.2. Cumulative and sequential models for ordered categorical responses are described in Sections 10.1.3 and 10.1.4.

10.1.1 Multivariate generalized linear models

To set up regression models for categorical responses, we will first introduce a multivariate class of models extending univariate generalized linear models. Following Fahrmeir & Tutz

(2001, Ch. 3), a multivariate generalized linear model for a q-dimensional response vector y_i is defined by the following assumptions:

Distributional assumption

Given covariates u_i, the q-dimensional response vectors y_i are (conditionally) independent and have a distribution that belongs to an exponential family, i. e. the density of y_i can be written as

$$f(y_i|\theta_i, \phi, \omega_i) = \exp\left(\frac{y_i'\theta_i - b(\theta_i)}{\phi}\omega_i + c(y_i, \phi, \omega_i)\right), \tag{10.2}$$

where, in analogy to univariate GLMs, ϕ is a scale parameter common to all observations, ω_i is a weight, and θ_i is the q-dimensional natural parameter of the exponential family.

Structural assumption

The expectation $\mu_i = E(y_i|u_i)$ is determined by a q-dimensional vector of linear predictors

$$\eta_i = U_i\gamma \tag{10.3}$$

and a vector-valued response function $h : \mathbb{R}^q \to \mathbb{R}^q$ via

$$\mu_i = h(\eta_i) = h(U_i\gamma), \tag{10.4}$$

where U_i is a $(q \times p)$ design matrix formed of the covariates u_i and γ is a p-dimensional vector of regression coefficients.

The models for categorical responses that we will discuss next are special cases of multivariate generalized linear models, where y_i follows a multinomial distribution

$$y_i \sim M(m_i, \pi_i) \quad \text{and} \quad \pi_i = (\pi_i^{(1)}, \ldots, \pi_i^{(q)}).$$

The density of such a multinomial distribution is given by

$$\begin{aligned} f(y_i) &= \frac{m_i!}{y_i^{(1)}! \cdot \ldots \cdot y_i^{(q)}!(m_i - y_i^{(1)} - \ldots - y_i^{(q)})!} \cdot \\ &\quad (\pi_i^{(1)})^{y_i^{(1)}} \cdot \ldots \cdot (\pi_i^{(q)})^{y_i^{(q)}}(1 - \pi_i^{(1)} - \ldots - \pi_i^{(q)})^{m_i - y_i^{(1)} - \ldots - y_i^{(q)}} \end{aligned} \tag{10.5}$$

and can be shown to be of the general form (10.2) of an exponential family density by defining

$$\begin{aligned} \theta_i &= \left(\log\left(\frac{\pi_i^{(1)}}{1 - \pi_i^{(1)} - \ldots - \pi_i^{(q)}}\right), \quad \ldots, \quad \log\left(\frac{\pi_i^{(q)}}{1 - \pi_i^{(1)} - \ldots - \pi_i^{(q)}}\right)\right)' \\ b(\theta_i) &= -\log(1 - \pi_i^{(1)} - \ldots - \pi_i^{(q)}) \\ c(y_i, \phi, \omega_i) &= \log\left(\frac{m_i!}{y_i^{(1)}! \cdot \ldots \cdot y_i^{(q)}!(m_i - y_i^{(1)} - \ldots - y_i^{(q)})!}\right) \\ \omega_i &= m_i \\ \phi &= 1. \end{aligned}$$

Specific types of regression models are now derived by determining appropriate response functions h and design matrices U with respect to the potential arrangement of the data. Examples will be presented in the following sections.

10.1.2 Models for nominal responses

In case of a nominal response $Y \in \{1, \ldots, k\}$ with unordered categories the most commonly used model is the multinomial logit model (see e. g. Fahrmeir & Tutz (2001, Ch. 3.2) or Agresti (2002, Ch. 7)), which can be regarded as a direct generalization of the univariate logit model. Here, the probability of category r is specified as

$$P(Y = r) = \pi^{(r)} = h^{(r)}\left(\eta^{(1)}, \ldots, \eta^{(q)}\right) = \frac{\exp(\eta^{(r)})}{1 + \sum_{s=1}^{q} \exp(\eta^{(s)})}. \tag{10.6}$$

where $\eta^{(r)}$ is a category-specific linear predictor depending on covariates and regression coefficients. The particular form of this predictor will be discussed later on in this section. Equivalently to the response function defined in (10.6), we can consider the link function, i. e. the inverse response function

$$g^{(r)}(\pi^{(1)}, \ldots, \pi^{(q)}) = \eta^{(r)} = \log\left(\frac{\pi^{(r)}}{1 - \sum_{s=1}^{q} \pi^{(s)}}\right).$$

Like most models for categorical responses, the multinomial logit model (10.6) can be motivated by considering latent variables and specific assumptions connecting these latent variables with the categorical response Y. For the multinomial logit model, this connecting mechanism is the principle of maximum utility. Latent variable representations also allow for additional insight in the properties of categorical response models, especially to formulate identifiability restrictions for the regression coefficients.

10.1.2.1 The principle of maximum random utility

In general, the principle of maximum random utility considers latent utilities

$$L^{(r)} = l^{(r)} + \varepsilon^{(r)}, \quad r = 1, \ldots, k,$$

where $l^{(r)}$ is deterministic and $\varepsilon^{(1)}, \ldots, \varepsilon^{(k)}$ are i. i. d. random variables with some continuous cumulative distribution function F. Once these latent variables are realized, the categorical response Y is determined by

$$Y = r \quad \Leftrightarrow \quad L^{(r)} = \max_{s=1,\ldots,k} L^{(s)}. \tag{10.7}$$

In the context of decision theory, $L^{(r)}$ describes the randomly disturbed profit a person has if it chooses alternative r. The principle of maximum random utility merely states that one always chooses the alternative that maximizes the profit.

When constituting the systematic part of utility $L^{(r)}$, the simplest form is given by

$$l^{(r)} = u'\alpha^{(r)}, \tag{10.8}$$

where u is a vector of covariates not depending on the specific category and $\alpha^{(r)}$ is a category-specific vector of regression coefficients. If category-specific covariates $w^{(r)}$ are available, (10.8) can be extended to

$$l^{(r)} = u'\alpha^{(r)} + w^{(r)\prime}\delta, \tag{10.9}$$

where the coefficient vector δ is postulated to be identical for all categories.

Finally, the probability for a specific decision can be computed in terms of the distribution of the error variables $\varepsilon^{(r)}$:

$$\begin{aligned} P(Y=r) &= P(L^{(r)} - L^{(1)} \geq 0, \ldots, L^{(r)} - L^{(k)} \geq 0) \\ &= P\left(\varepsilon^{(1)} \leq l^{(r)} - l^{(1)} + \varepsilon^{(r)}, \ldots, \varepsilon^{(k)} \leq l^{(r)} - l^{(k)} + \varepsilon^{(r)}\right) \\ &= \int_{-\infty}^{\infty} \prod_{s \neq r} F(l^{(r)} - l^{(s)} + \varepsilon) f(\varepsilon) d\varepsilon, \end{aligned} \tag{10.10}$$

where f denotes the density function corresponding to F. Different choices for F (or f) lead to specific models for nominal responses.

10.1.2.2 Multinomial logit model

Taking the extreme value distribution as error distribution with cumulative density function

$$F(\varepsilon) = \exp(-\exp(-\varepsilon))$$

and density

$$f(\varepsilon) = \exp(-\exp(-\varepsilon))\exp(-\varepsilon)$$

results in the multinomial logit model (10.6), since (10.10) can be rewritten in the following way:

$$\begin{aligned} P(Y=r) &= \int_{-\infty}^{\infty} \prod_{s \neq r} F(l^{(r)} - l^{(s)} + \varepsilon) f(\varepsilon) d\varepsilon \\ &= \int_{-\infty}^{\infty} \prod_{s \neq r} \exp(-\exp(-l^{(r)} + l^{(s)} - \varepsilon)) \exp(-\exp(-\varepsilon)) \exp(-\varepsilon) d\varepsilon \\ &= \int_{-\infty}^{\infty} \exp\left(-\sum_{s=1}^{k} \exp(-l^{(r)} + l^{(s)}) \exp(-\varepsilon)\right) \exp(-\varepsilon) d\varepsilon \\ &= \int_{0}^{\infty} \exp\left(-\sum_{s=1}^{k} \exp(-l^{(r)} + l^{(s)}) t\right) dt \\ &= \frac{1}{\sum_{s=1}^{k} \exp(-l^{(r)} + l^{(s)})} \\ &= \frac{\exp(l^{(r)})}{\sum_{s=1}^{k} \exp(l^{(s)})} \end{aligned} \tag{10.11}$$

It is clear from (10.10) that only the $q = k - 1$ differences of the latent variables are identifiable. Hence, one of the $\alpha^{(r)}$ has to be restricted. Choosing k as reference category and setting $\alpha^{(k)} = 0$ finally leads to the multinomial logit model (10.6) with linear predictors

$$\eta^{(r)} = u'\alpha^{(r)} + (w^{(r)} - w^{(k)})'\delta = u'\alpha^{(r)} + \bar{w}^{(r)\prime}\delta, \tag{10.12}$$

where $\bar{w}^{(r)} = w^{(r)} - w^{(k)}$. These predictors can be summarized in the multivariate form (10.3) by defining

$$U_i = \begin{pmatrix} u_i' & & (w_i^{(1)} - w_i^{(k)})' \\ & \ddots & \vdots \\ & u_i' & (w_i^{(q)} - w_i^{(k)})' \end{pmatrix}$$

and

$$\gamma = (\alpha^{(1)\prime}, \ldots, \alpha^{(q)\prime}, \delta')'.$$

Choosing standard normal distributed errors in (10.10) results in the multinomial probit model which can be further generalized by allowing for correlated random errors $\varepsilon^{(r)}$. However, the application of multinomial probit models faces numerical problems since the analytic evaluation of probability (10.10) is no longer feasible. Simulation based methods are available, either using simulated likelihood methods (e. g. Keane (1994) or Ziegler & Eymann (2001)) or MCMC simulation techniques, where the Gaussian latent variables are augmented by sampling from appropriate normal distributions (e. g. Chib & Greenberg (1998) or Fahrmeir & Lang (2001b)). In contrast, the empirical Bayes approach discussed in the following cannot be applied directly.

10.1.2.3 Structured additive regression for nominal responses

Similar to the problems discussed in Section 4.1.2 for univariate responses, simple regression models for categorical responses face restricted applicability in real data situations due to their purely parametric nature. However, the multivariate situation is somewhat more complicated, since we have to distinguish between covariates with effects varying over the categories and category-specific covariates with effects fixed over the categories. Replacing the strictly linear predictor in (10.9) by a structured additive predictor which combines both types of effects yields the latent utilities

$$l^{(r)} = u'\alpha^{(r)} + w^{(r)\prime}\delta + f_1^{(r)}(\nu_1) + \ldots + f_l^{(r)}(\nu_l) + f_{l+1}(\nu_{l+1}^{(r)}) + \ldots + f_p(\nu_p^{(r)}), \qquad (10.13)$$

where $u'\alpha^{(r)}$ and $w^{(r)\prime}\delta$ model parametric effects of covariates with linear influence as in (10.9), $f_1^{(r)}(\nu_1), \ldots, f_l^{(r)}(\nu_l)$ are nonlinear functions of covariates fixed for all categories and $f_{l+1}(\nu_{l+1}^{(r)}), \ldots, f_p(\nu_p^{(r)})$ are nonlinear effects of category-specific covariates. According to the specifications discussed in Section 4.2, nonlinear effects of continuous covariates, spatial effects, interaction effects based on varying coefficients or interaction surfaces, and random effects are all comprised in this framework. Correspondingly, the generic covariates $\nu_1, \ldots, \nu_l, \nu_{l+1}^{(r)}, \ldots, \nu_p^{(r)}$ denote covariates of different types and dimension.

Proceeding as in the purely parametric model reveals that the predictors of a structured additive multinomial logit model are given by

$$\eta^{(r)} = u'\alpha^{(r)} + \bar{w}^{(r)\prime}\delta + f_1^{(r)}(\nu_1) + \ldots + f_l^{(r)}(\nu_l) + \bar{f}_{l+1}(\nu_{l+1}^{(r)}) + \ldots + \bar{f}_p(\nu_p^{(r)}), \quad r = 1, \ldots, q,$$

where

$$\bar{f}_j(\nu_j^{(r)}) = f_j(\nu_j^{(r)}) - f_j(\nu_j^{(k)}).$$

Again, only differences of effects enter the predictors for category-specific covariates.

10.1.2.4 Special Cases

In order to demonstrate the flexibility of categorical structured additive regression, we briefly describe some special cases of (10.13) which have been previously introduced in the literature.

Multinomial models in the spirit of generalized additive models have been proposed by Kooperberg, Bose & Stone (1997) and Yau, Kohn & Wood (2003). In both cases, the latent utilities are given by

$$l^{(r)} = \sum_{k=1}^{K} f_k^{(r)}(x_k) + \sum_{k=1}^{K} \sum_{l=k+1}^{K} f_{kl}^{(r)}(x_k, x_l) + \varepsilon^{(r)}. \tag{10.14}$$

Thus, effects of K continuous covariates $x_1, \ldots, x_K$ are modeled in terms of main effects $f_k^{(r)}$ and interactions $f_{kl}^{(r)}$. Such a model can be subsumed in a structured additive regression model (10.13) with a total number of $p = K + \frac{K(K-1)}{2}$ model terms by defining the generic covariates $\nu_1 = x_1, \ldots, \nu_K = x_K, \nu_{K+1} = (x_1, x_2), \ldots, \nu_p = (x_{K-1}, x_K)$ and functions $f_1^{(r)}(\nu_1) = f_1^{(r)}(x_1), \ldots, f_K^{(r)}(\nu_K) = f_K^{(r)}(x_K), f_{K+1}^{(r)}(\nu_{K+1}) = f_{1,2}^{(r)}(x_1, x_2), \ldots, f_p^{(r)}(\nu_p) = f_{K-1,K}^{(r)}(x_{K-1}, x_K)$. In structured additive regression, the nonparametric main effects are modeled using penalized splines or any of the smoothing techniques discussed in Section 4.2.2. Interactions can be estimated based on two-dimensional tensor product P-splines as presented in Section 4.2.6. Note that in (10.14) all covariates are assumed to be global and, therefore, effects of category-specific covariates are not included.

To estimate models with latent utilities (10.14), Kooperberg et al. (1997) extend the methodology of multivariate adaptive regression splines (MARS, Friedman 1991) to multinomial logit models. Nonparametric and interaction effects are modeled using linear splines and their tensor products, respectively. Smoothness of the estimated curves is not achieved by penalization but via stepwise inclusion and deletion of basis functions based on an information criterion, e. g. AIC. The approach is implemented in an R-routine called polyclass and will be comprised as a competing method in the simulation study on categorical structured additive regression in Section 13.

Yau et al. (2003) assume Gaussian errors in the latent utilities resulting in multinomial probit models. Nonparametric and interaction effects are modeled via radial or thin plate spline basis functions. Smoothness and parsimony of the estimated model are approached by some Bayesian variable selection technique employing Markov Chain Monte Carlo techniques. In particular, an algorithm based on data augmentation that involves sampling of the latent utilities is considered, since in this case all full conditionals are of simple form and Gibbs sampling steps can be performed.

A comparable model that allows for category-specific covariates is presented in Tutz & Scholz (2004). They consider semiparametric additive latent variables of the form

$$l^{(r)} = u'\alpha^{(r)} + w^{(r)\prime}\delta + \sum_{j=1}^{l} f_j^{(r)}(x_j) + \sum_{j=l+1}^{p} f_j(x_j^{(r)}) + \varepsilon^{(r)}. \tag{10.15}$$

In contrast to (10.14), no interaction effects are included but parametric as well as nonparametric effects of category-specific covariates are considered. Of course, model (10.15)

is of the form (10.13) if all generic covariates ν_j and $\nu_j^{(r)}$ are in fact continuous covariates, i. e. $\nu_j = x_j$ and $\nu_j^{(r)} = x_j^{(r)}$.

Tutz & Scholz (2004) model the nonparametric effects in (10.15) using penalized splines as described in Section 4.2.2.1. They propose to choose the smoothing parameters according to minimal AIC but actually perform this minimization based on a grid search which leads to intractable computational effort already for a small number of nonparametric model terms. In contrast, the mixed model based approach presented in the following section allows for the routine determination of a large number of nonparametric effects.

Fahrmeir & Lang (2001b) introduce semiparametric regression models for the analysis of spatio-temporal categorical data within a Bayesian framework. They consider geoadditive latent utilities

$$l^{(r)} = u'\alpha + f_{time}^{(r)}(t) + f_{spat}^{(r)}(s) + \sum_{k=1}^{K} f_k^{(r)}(x_k) + \varepsilon^{(r)}, \tag{10.16}$$

where f_{time} is a nonlinear function of time which may be decomposed further into a trend and a seasonal component as described in Section 4.2.2.4. Analogously, the spatial effect f_{spat} might be split up into a spatially structured and a spatially unstructured part as discussed in Section 4.2.3. In addition, (10.16) contains nonparametric effects f_k of continuous covariates and a further set of covariates u whose effects are modeled parametrically. While Fahrmeir & Lang (2001b) use random walks to model nonparametric and temporal effects, Markov random field priors for the spatial effect and seasonal priors for the seasonal effect, Brezger & Lang (2005) describe extensions where nonparametric effects are modeled by penalized splines, and interaction effects can be included based on two-dimensional tensor product P-splines. In either case, estimation is based on Markov Chain Monte Carlo simulation techniques, allowing for both multinomial logit and probit models.

Clearly, model (10.16) is a submodel of (10.13), since all effects are of the general form discussed in Section 4.2. Moreover, the general structured additive regression model (10.13) allows for category-specific covariates while all covariates are assumed to be global in (10.16).

As a last submodel of (10.13), we consider the mixed logit model which is particularly popular in econometrics (see for example Train 2003, Ch. 6). Here, the only nonstandard effects are random slopes with category-specific interaction variables, i. e.

$$l^{(r)} = u'\alpha^{(r)} + w^{(r)\prime}\delta + w^{(r)\prime}b + \varepsilon^{(r)},$$

where b is a cluster-specific random effect not depending on the category. This model is incorporated into (10.13) by defining functions $f_j(\nu_j^{(r)}) = w_j^{(r)\prime}b_j$.

Mixed logit models were originally introduced to overcome the restrictive implications of multinomial logit models which exhibit the independence from irrelevant alternatives property (see Chapter 3.3 in Train (2003) for a detailed discussion of this property). Basically, the idea is to allow for correlations between the latent utilities by introducing random effects. In a model without random effects, the latent utilities $l^{(1)}, \ldots, l^{(k)}$ of a particular observation are independent, since the error terms $\varepsilon^{(1)}, \ldots, \varepsilon^{(k)}$ are assumed to be independent. Introducing random effects b induces correlations between the latent

utilities. Of course, the random effects do not have to be individual-specific but can also be defined upon clusters of observations as long as they are global, i. e., do not vary over the categories. Note also that no random intercepts can be addressed since the interaction variable has to be category-specific.

From a Bayesian perspective, correlated latent utilities are not only achieved with classical random effects but also with other effects of category-specific covariates. For example, including a P-spline $f(x^{(r)})$ of a continuous covariate $x^{(r)}$ likewise leads to correlated latent utilities since in a Bayesian formulation, a P-spline is also a correlated random effect.

Classical approaches to the estimation of mixed logit models typically applied in econometrics involve simulation based methods (Keane 1994). If formulated as a structured additive regression model, the parametric mixed logit model is not only easily extended to a semiparametric version but is also estimable based on mixed model methodology or MCMC.

10.1.3 Cumulative models for ordinal responses

10.1.3.1 Parametric cumulative models

If the categories of the response can be ordered, i. e. Y is a variable measured on ordinal scale, adequate models can be developed by switching to a different type of latent variable mechanism (see Fahrmeir & Tutz (2001, Ch. 3.3) or Agresti (2002, Ch. 7) for more rigorous treatments). In the models investigated in this section, $Y \in \{1, \ldots, k\}$ is connected with a latent variable

$$L = l + \varepsilon$$

via

$$Y = r \quad \Leftrightarrow \quad \theta^{(r-1)} < L \leq \theta^{(r)}, \tag{10.17}$$

where $-\infty = \theta^{(0)} < \theta^{(1)} < \ldots < \theta^{(k)} = \infty$ are ordered thresholds. In contrast to the models for nominal responses, there is only one latent variable whose deterministic part is once again determined in a regression way. In the simplest form it is given by $l = u'\alpha$. From (10.17), the cumulative distribution function of Y is then easily obtained as

$$P(Y \leq r) = F(\theta^{(r)} - u'\alpha), \tag{10.18}$$

where F denotes the cumulative distribution function of the error term ε. In contrast to the multinomial logit model, both the covariates u and the regression coefficients α are assumed to be fixed for all categories in this basic model. Since the model defines cumulative probabilities, it is usually called a cumulative regression model. In Section 10.1.4 we will discuss a second approach for the analysis of ordinal responses which is based on a sequential mechanism.

From (10.18), the response function can be derived to be

$$\begin{aligned} P(Y = 1) &= \pi^{(1)} = h^{(1)}\left(\eta^{(1)}, \ldots, \eta^{(q)}\right) = F(\eta^{(1)}) \\ P(Y = r) &= \pi^{(r)} = h^{(r)}\left(\eta^{(1)}, \ldots, \eta^{(q)}\right) = F(\eta^{(r)}) - F(\eta^{(r-1)}), \quad r > 1, \end{aligned}$$

with linear predictors

$$\eta^{(r)} = \theta^{(r)} - u'\alpha.$$

10.1.3.2 Interpretation of covariate effects

Figure 10.1 gives an intuitive interpretation of cumulative response models. Suppose that $f = F'$ is the density of ε. Then the distribution of L follows the same density shifted by $u'\alpha$ and the regression coefficients α determine the direction and strength of the shift caused by a specific covariate combination u. This is illustrated for two different values of $u'\alpha$ in Figure 10.1. The probability for a specific category r is now obtained by splitting the density at the thresholds. As an example, the probability for category $r = 2$ is shaded in grey. Obviously, increasing the shift reduces the probability for a small category and conversely increases the probability for higher categories.

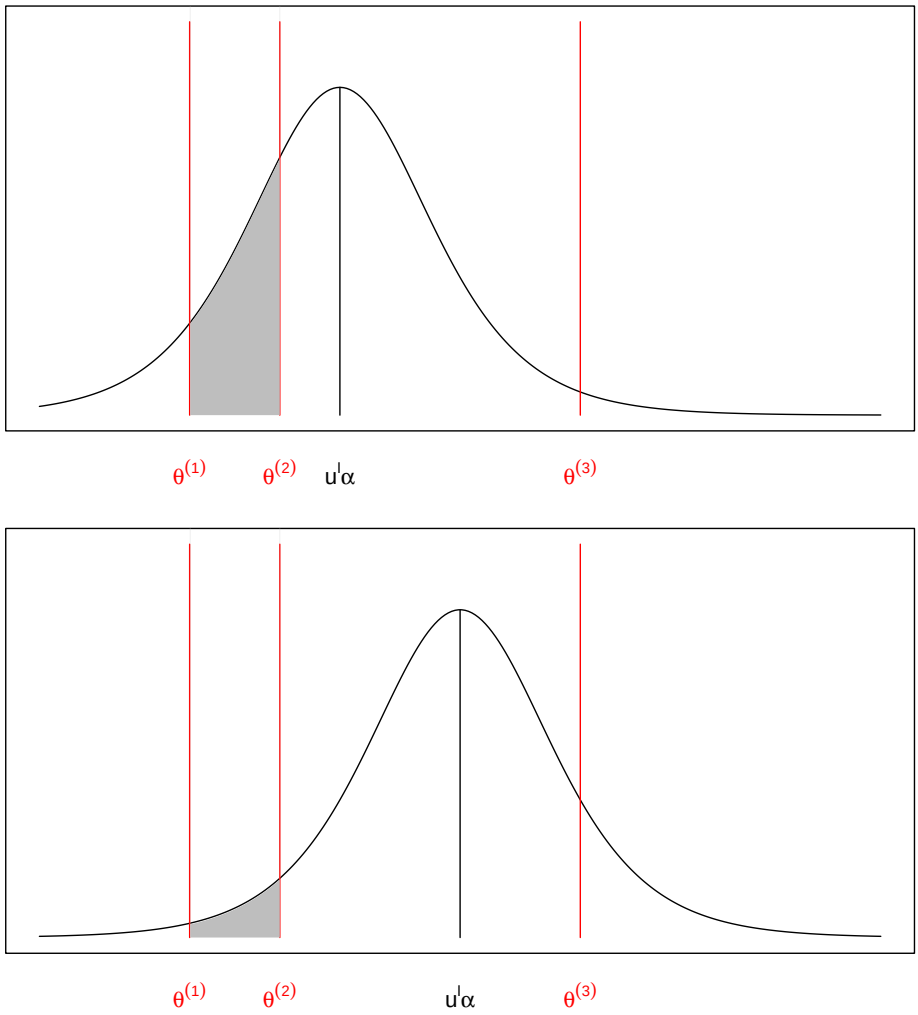


Figure 10.1: Interpretation of cumulative regression models: The linear predictor $u'\alpha$ shifts the density of L. The shaded region denotes the probability $P(Y = 2)$.

Choosing a specific distribution for ε completes the model formulation. Popular variants are the logistic distribution or the standard normal distribution resulting in the cumulative logit and the cumulative probit model, respectively. In the former case, (10.17) yields

$$P(Y \leq r) = \frac{\exp(\theta^{(r)} - u'\alpha)}{1 + \exp(\theta^{(r)} - u'\alpha)}$$

which can be equivalently written in terms of log odds:

$$\log\left(\frac{P(Y \leq r)}{P(Y > r)}\right) = \theta^{(r)} - u'\alpha.$$

In other words, the cumulative logit model parameterizes the log odds of the cumulative probabilities $P(Y \leq r)$.

10.1.3.3 Extended cumulative models

In the basic cumulative regression model (10.17), the covariates have no impact on the values of the thresholds. However, in some applications it is likely that the thresholds themselves are covariate-dependent or, equivalently, that some of the covariate effects are category-specific. This leads to extended cumulative models where the linear predictor is given by

$$\eta^{(r)} = \theta^{(r)} - u'\alpha - w'\delta^{(r)}. \tag{10.19}$$

In such a model, the covariate-dependent thresholds are given by $\tilde{\theta}^{(r)}(w) = \theta^{(r)} - w'\delta^{(r)}$. Though being easily defined, extended cumulative models are of considerably increased complexity, since the ordering restrictions $\theta^{(1)} < \ldots < \theta^{(q)}$ have to hold for the covariate-dependent thresholds, i. e. the inequalities

$$\theta^{(1)} - w'\delta^{(1)} < \ldots < \theta^{(q)} - w'\delta^{(q)}$$

have to be fulfilled for all possible values of the covariates w. Especially in sparse data situations this may cause numerical problems in the estimation process.

Cumulative regression models can be cast in the general form of multivariate generalized linear models by defining the design matrix

$$U_i = \begin{pmatrix} 1 & -w_i' & & & & & -u_i' \\ & & 1 & -w_i' & & & -u_i' \\ & & & & \ddots & & \vdots \\ & & & & & 1 & -w_i' & -u_i' \end{pmatrix}$$

and the regression coefficients

$$\gamma = (\theta^{(1)}, \delta^{(1)\prime}, \ldots, \theta^{(q)}, \delta^{(q)\prime}, \alpha')'.$$

10.1.3.4 Cumulative structured additive regression models

To account for possible deviations from the parametric form of the predictor supposed in (10.19), we replace the parametric predictor by a structured additive predictor, thereby differentiating between category-specific effects and effects fixed for all categories. This yields the cumulative structured additive regression model

$$\eta^{(r)} = \theta^{(r)} - u'\alpha - w'\delta^{(r)} - f_1(\nu_1) - \ldots - f_l(\nu_l) - f_{l+1}^{(r)}(\nu_{l+1}) - \ldots - f_p^{(r)}(\nu_p), \tag{10.20}$$

where $f_1(\nu_1), \ldots, f_l(\nu_l)$ are nonlinear functions fixed for all categories and $f_{l+1}^{(r)}(\nu_{l+1}), \ldots, f_p^{(r)}(\nu_p)$ are nonlinear category-specific effects. Again, the functions $f_1, \ldots, f_l$, $f_{l+1}^{(r)}, \ldots, f_p^{(r)}$ are general functions of generic covariates accounting for different types of effects such as nonlinear effects of continuous covariates, temporal effects, spatial effects, interaction effects or random effects. To explain the generality of structured additive regression models for ordinal responses, we will examine selected special cases of (10.20) recently described in the literature.

Tutz (2003) introduces semiparametric ordinal models, where effects of continuous covariates are modeled nonparametrically based on penalized splines. Both global effects fixed for all categories and category-specific effects are taken into account and extensions based on varying coefficient terms are discussed. To ensure identifiability of models with category-specific nonparametric effects, penalization across categories is added to the usual penalties for the different effects. Semiparametric ordinal models are special ordinal structured additive regression models, where all covariates ν_j are in fact continuous covariates x_j, i. e. we have $f_j(\nu_j) = f_j(x_j)$, $j = 1, \ldots, l$ and $f_j^{(r)}(\nu_j) = f_j^{(r)}(x_j)$, $j = l+1, \ldots, p$. If varying coefficient terms are included, some of the generic covariates are bivariate, that is $\nu_j = (x_{j_1}, x_{j_2})$ and $f_j(\nu_j) = x_{j_1} g_j(x_{j_2})$.

Penalized likelihood estimation with smoothing parameters chosen according to AIC based on a grid search is proposed in Tutz (2003). This approach becomes quite computationally intensive in high-dimensional problems with a large number of nonparametric effects. In contrast, the mixed model based estimation approach described in Section 11 allows for the routine computation of complex regression models for ordinal responses including the estimation of several smoothing parameters. As a potential drawback, mixed model based estimation does not directly allow to include penalties across categories and, therefore, identifiability problems may be observed more frequently.

Fully Bayesian spatio-temporal models for ordinal responses based on MCMC simulation techniques are described in Fahrmeir & Lang (2001b), where predictors of the form

$$\eta^{(r)} = \theta^{(r)} - u'\alpha - f_{time}(t) - f_{spat}(s) - f_1(x_1) - \ldots - f_l(x_l) \tag{10.21}$$

are considered. In (10.21), all effects are assumed to be global, i. e. no category-specific effects are included. Fahrmeir & Lang (2001b) use a Markov random field prior for the spatial effect and random walk or seasonal priors for the temporal and the nonparametric effects. Extensions based on penalized splines and their tensor products can be found in Brezger & Lang (2005). The general ordinal structured additive regression model (10.20) extends (10.21) by allowing for GRF priors for the spatial effect and the inclusion of category-specific effects.

10.1.4 Sequential models for ordinal responses

While the cumulative models discussed in the previous section are in many situations appropriate for the analysis of responses with ordered categories, a different model may be preferred if the ordering is caused by a sequential mechanism. This may be the case if the categories can only be achieved successively, for example when analyzing different stages of a disease. Here, sequential models seem to be more natural, since they model

transitions between the categories in a successive manner. The model bases on latent variables $L^{(1)}, \ldots, L^{(q)}$ with

$$L^{(r)} = l + \varepsilon^{(r)}$$

and i. i. d. random errors $\varepsilon^{(r)}$ (see Fahrmeir & Tutz 2001, Ch. 3.3). In this case, the random part of the latent variable is category-specific, whereas the deterministic part is not. The probability for achieving category r is modeled conditional on the achievement of the previous category $r-1$ using thresholds $\theta^{(1)}, \ldots, \theta^{(r)}$:

$$Y = r | Y \geq r \quad \Leftrightarrow \quad L^{(r)} \leq \theta^{(r)}. \tag{10.22}$$

Note that, in contrast to cumulative models, no ordering restrictions have to be fulfilled by the thresholds since each threshold is associated with exactly one specific transition.

Setting the deterministic part of $L^{(r)}$ to $l = u'\alpha$ yields the conditional probabilities

$$P(Y = r | Y \geq r) = F(\theta^{(r)} - u'\alpha),$$

where again F denotes the cumulative distribution function of the random errors. From this expression, the response function is obtained as

$$P(Y = r) = \pi^{(r)} = h^{(r)}(\eta^{(1)}, \ldots, \eta^{(q)}) = F(\eta^{(r)}) \prod_{s=1}^{r-1} [1 - F(\eta^{(s)})]$$

with linear predictors

$$\eta^{(r)} = \theta^{(r)} - u'\alpha.$$

Different choices for the cumulative distribution function F result in different models, e. g. sequential logit and sequential probit models with the logistic or the standard normal distribution, respectively.

In complete analogy to cumulative models, sequential models can be extended to contain covariate-dependent thresholds or, equivalently, category-specific effects by introducing extended linear predictors

$$\eta^{(r)} = \theta^{(r)} - u'\alpha - w'\delta^{(r)}.$$

The advantage of sequential models is that no ordering restrictions are needed for the thresholds and, hence, the numerical problems mentioned for cumulative models do not arise. Both the basic and the extended form of sequential models can be cast in the general form of multivariate generalized linear models by defining the design matrix

$$U_i = \begin{pmatrix} 1 & -w_i' & & & & & -u_i' \\ & & 1 & -w_i' & & & -u_i' \\ & & & & \ddots & & \vdots \\ & & & & & 1 & -w_i' & -u_i' \end{pmatrix}$$

and regression coefficients

$$\gamma = (\theta^{(1)}, \delta^{(1)\prime}, \ldots, \theta^{(q)}, \delta^{(q)\prime}, \alpha')'.$$

Sequential structured additive regression models can be introduced similarly as with cumulative models. This results in structured additive predictors

$$\eta^{(r)} = \theta^{(r)} - u'\alpha - w'\delta^{(r)} - f_1(\nu_1) - \ldots - f_l(\nu_l) - f_{l+1}^{(r)}(\nu_{l+1}) - \ldots - f_p^{(r)}(\nu_p), \tag{10.23}$$

where $f_1(\nu_1), \ldots, f_l(\nu_l)$ are nonlinear functions fixed for all categories and $f_{l+1}^{(r)}(\nu_{l+1}), \ldots, f_p^{(r)}(\nu_p)$ are nonlinear category-specific effects accounting for nonlinear effects of covariates, interactions, spatial correlations, and so on.

As an attractive feature, sequential regression models offer an interpretation as discrete time survival models. In this case, the transition from category $r-1$ to category r implies that the corresponding individual survived the $(r-1)$-th time point, while stopping at category $r-1$ means death at the $(r-1)$-th time point. In Section 14.4.2, we will describe how (right) censoring can be included in sequential models and how such modified sequential models can be estimated based on data augmentation and binary regression. This equivalent expression in terms of binary regression models may also be the reason why sequential models are less commonly considered in the literature (see Tutz (2003) for an extension describing semiparametric regression models for sequential responses). However, a direct treatment of sequential categorical responses exhibits the advantage of avoiding unnecessary data augmentation.

10.2 Likelihood and priors

To complete the Bayesian formulation of categorical structured additive regression models, we have to specify a likelihood and priors for the different types of effects. Similarly as in the discussion for univariate responses in Part II, all functions $f_j(\nu_j)$, $f_j^{(r)}(\nu_j)$ or $f_j(\nu_j^{(r)})$ contained in (10.13), (10.20) and (10.23), can be written in terms of a design vector and a vector of regression coefficients, i. e.

$$f_j(\nu_j) = v_j'\xi_j, \quad f_j^{(r)}(\nu_j) = v_j'\xi_j^{(r)} \quad \text{and} \quad f_j(\nu_j^{(r)}) = v_j^{(r)\prime}\xi_j. \tag{10.24}$$

Assuming conditional independence of the observations allows to compute the likelihood of the vector ξ combining all regression coefficients as

$$L(\xi) = \prod_{i=1}^{n} f(y_i|\xi),$$

where $f(y_i)$ is a multinomial density of the form (10.5). The precise form of the likelihood contributions $f(y_i)$ depends on the model and is derived from the predictors $\eta_i^{(r)}$ by computing the probability vector $\pi = (\pi_i^{(1)}, \ldots, \pi_i^{(q)})$ via (10.4) and applying the response function for that specific model.

Dropping possible category indices from the regression coefficients, the priors for these coefficients are of the well known form

$$p(\xi_j|\tau_j^2) \propto \exp\left(-\frac{1}{2\tau_j^2}\xi_j' K_j \xi_j\right) \tag{10.25}$$

of a multivariate but generally improper Gaussian distribution. Specific types of priors for the effects of continuous covariates, spatial effects, random effects, varying coefficients and interaction effects were investigated in Section 4.2 and can immediately be transferred to the effects in categorical regression models.

In summary, the posterior of ξ is given by

$$p(\xi|Y) \propto L(\xi) \prod_j p(\xi_j|\tau_j^2), \tag{10.26}$$

and posterior mode estimates for ξ are obtained by maximizing the right hand side of (10.26), or, taking logarithms, the penalized log-likelihood

$$l_p(\xi|Y) = l(\xi) - \frac{1}{2} \sum_j \frac{1}{\tau_j^2} \xi_j' K_j \xi_j.$$

Although this optimization could be performed based on a Fisher-scoring-type algorithm, inference in Section 11 will be based on a different parameterization, since there is no obvious rule on how to determine the variance parameters in the original formulation. Hence, we apply the same idea as in Part II to generate a categorical mixed model with proper priors and extend estimation techniques for mixed models to the multivariate case.

11 Inference

Inference in categorical structured additive regression (STAR) models consists of three parts: In Section 11.1, we will describe how to reparametrize categorical STAR models as mixed models in order to obtain proper prior distributions. Section 11.2 presents a penalized likelihood estimation procedure for the estimation of regression coefficients when variance parameters are given. Section 11.3 refers to the estimation of these variances and Section 11.4 summarizes mixed model based estimation of categorical STAR models in form of an algorithm. The final Section 11.5 provides some brief comments on the estimation of categorical STAR models based on MCMC simulation techniques.

11.1 Mixed model representation

In order to rewrite categorical structured additive regression models as mixed models, we proceed in a similar way as in the univariate case. The vectors of regression coefficients are decomposed into a penalized and an unpenalized part as in Equation (5.2) on page 64, yielding different types of decompositions for the ingredients of the predictors discussed in Section 10.1. The products of design vectors and regression coefficients given in (10.24) are consequently decomposed as

$$v_{ij}'\xi_j = v_{ij}'(\tilde{X}_j\beta_j + \tilde{Z}_j b_j) \quad = x_{ij}'\beta_j + z_{ij}'b_j, \tag{11.1}$$

$$v_{ij}'\xi_j^{(r)} = v_{ij}'(\tilde{X}_j\beta_j^{(r)} + \tilde{Z}_j b_j^{(r)}) = x_{ij}'\beta_j^{(r)} + z_{ij}'b_j^{(r)} \quad \text{and} \tag{11.2}$$

$$v_{ij}^{(r)\prime}\xi_j = v_{ij}^{(r)\prime}(\tilde{X}_j\beta_j + \tilde{Z}_j b_j) \quad = x_{ij}^{(r)\prime}\beta_j + z_{ij}^{(r)\prime}b_j, \tag{11.3}$$

respectively.

Choosing appropriate matrices $\tilde{X}_j$ and $\tilde{Z}_j$ according to the considerations in Section 5.1 leads to a flat prior for β_j and $\beta_j^{(r)}$, i. e. these regression coefficients are considered as fixed effects in the resulting mixed model. The penalized parts b_j and $b_j^{(r)}$ transform to i. i. d. random effects with variances τ_j^2 and $(\tau_j^{(r)})^2$, i. e.

$$b_j \sim N(0, \tau_j^2 I) \quad \text{and} \quad b_j^{(r)} \sim N(0, (\tau_j^{(r)})^2 I).$$

To restate categorical STAR models in the compact matrix notation of mixed models, we define overall design matrices for fixed effects and random effects. These design matrices are formed in analogy to the design matrices U_i discussed in Section 10.1. For nominal responses, we obtain

$$X_i = \begin{pmatrix} u_i' & & & \bar{w}_i^{(1)\prime} & x_{i1}' & & & \dots & x_{il}' & & & \bar{x}_{i,l+1}^{(1)\prime} & \dots & \bar{x}_{ip}^{(1)\prime} \\ & \ddots & & \vdots & & \ddots & & & & \ddots & & \vdots & & \vdots \\ & & u_i' & \bar{w}_i^{(q)\prime} & & & x_{i1}' & \dots & & & x_{il}' & \bar{x}_{i,l+1}^{(q)\prime} & \dots & \bar{x}_{ip}^{(q)\prime} \end{pmatrix}$$

and

$$Z_i = \begin{pmatrix} z_{i1}' & & & \dots & z_{il}' & & & \bar{z}_{i,l+1}^{(1)\prime} & \dots & \bar{z}_{ip}^{(1)\prime} \\ & \ddots & & & & \ddots & & \vdots & & \vdots \\ & & z_{i1}' & \dots & & & z_{il}' & \bar{z}_{i,l+1}^{(q)\prime} & \dots & \bar{z}_{ip}^{(q)\prime} \end{pmatrix},$$

where $\bar{x}_{ij}^{(r)} = x_{ij}^{(r)} - x_{ij}^{(k)}$ and $\bar{z}_{ij}^{(r)} = z_{ij}^{(r)} - z_{ij}^{(k)}$. Accordingly, the regression coefficients are combined into the vectors

$$\beta = (\alpha^{(1)\prime}, \ldots, \alpha^{(q)\prime}, \delta', \beta_1^{(1)\prime}, \ldots, \beta_1^{(q)\prime}, \ldots, \beta_l^{(1)\prime}, \ldots, \beta_l^{(q)\prime}, \beta_{l+1}', \ldots, \beta_p')'$$

and

$$b = (b_1^{(1)\prime}, \ldots, b_1^{(q)\prime}, \ldots, b_l^{(1)\prime}, \ldots, b_l^{(q)\prime}, b_{l+1}', \ldots, b_p')',$$

corresponding to fixed and random effects, respectively. For cumulative and sequential models the design matrices are given by

$$X_i = \begin{pmatrix} 1 & -w_i' & & & -u_i' & -x_{i1} & \ldots & -x_{il} & & & \\ & & \ddots & & \vdots & \vdots & & \vdots & \cdots & & \\ & & & 1 & -w_i' & -u_i' & -x_{i1} & \ldots & -x_{il} & & \\ & & & & -x_{i,l+1}' & & & \ldots & -x_{ip}' & & \\ & & \cdots & & & \ddots & & & & \ddots & \\ & & & & & & -x_{i,l+1}' & \ldots & & & -x_{ip}' \end{pmatrix}$$

and

$$Z_i = \begin{pmatrix} -z_{i1} & \ldots & -z_{il} & -z_{i,l+1}' & & & \ldots & -z_{ip}' & & \\ \vdots & & \vdots & & \ddots & & & & \ddots & \\ -z_{i1} & \ldots & -z_{il} & & & -z_{i,l+1}' & \ldots & & & -z_{ip}' \end{pmatrix}$$

with regression coefficients

$$\beta = (\theta^{(1)}, \delta^{(1)\prime}, \ldots, \theta^{(q)}, \delta^{(q)\prime}, \alpha', \beta_1', \ldots, \beta_l', \beta_{l+1}^{(1)'}, \ldots, \beta_{l+1}^{(q)'}, \ldots, \beta_p^{(1)'}, \ldots, \beta_p^{(q)'})'$$

and

$$b = (b_1', \ldots, b_l', b_{l+1}^{(1)'}, \ldots, b_{l+1}^{(q)'}, \ldots, b_p^{(1)'}, \ldots, b_p^{(q)'})'.$$

Finally, defining the stacked vector of predictors $\eta = (\eta_i)$ and the stacked matrices $X = (X_i)$ and $Z = (Z_i)$, all categorical structured additive regression models can be written in form of the predictor

$$\eta = X\beta + Zb$$

with priors

$$p(\beta) \propto const$$

and

$$b \sim N(0, Q). \tag{11.4}$$

Hence, categorical STAR models can be represented as categorical mixed models with a proper prior for the random effects. The covariance matrix Q of the random effects is of blockdiagonal form and has to be arranged in agreement with the sequence of the random effects in b. For nominal models, Q is given by

$$Q = \text{blockdiag}\left((\tau_1^{(1)})^2 I, \ldots, (\tau_1^{(q)})^2 I, \ldots, (\tau_l^{(1)})^2 I, \ldots, (\tau_l^{(q)})^2 I, \tau_{l+1}^2 I, \ldots, \tau_p^2 I\right),$$

and for cumulative and sequential models we obtain

$$Q = \text{blockdiag}\left(\tau_1^2 I, \ldots, \tau_l^2 I, (\tau_{l+1}^{(1)})^2 I, \ldots, (\tau_{l+1}^{(q)})^2 I, \ldots, (\tau_p^{(1)})^2 I, \ldots, (\tau_p^{(q)})^2 I\right).$$

Since estimation of categorical STAR models is now transformed to the estimation of a categorical mixed model, we can apply the same estimation strategy as for univariate responses. Therefore, we separately consider the estimation of regression coefficients and variance components in the next two sections and summarize the estimation scheme in Section 11.4.

11.2 Estimation of regression coefficients

To obtain posterior mode estimates for the regression coefficients, we have to maximize posterior (10.26) or, equivalently, in mixed model formulation

$$p(\beta, b|Y) \propto L(\beta, b)p(b),$$

where the likelihood $L(\beta, b)$ equals the likelihood defined in terms of the original parameterization, i. e. $L(\xi) = L(\beta, b)$, and the prior of the random effects $p(b)$ is given in (11.4). The log-posterior has the form of a penalized likelihood

$$l_p(\beta, b) = l(\beta, b) - \frac{1}{2} b' Q^{-1} b$$

and maximization can be carried out using a similar Fisher-Scoring algorithm as in the univariate case. Rewriting this Fisher-Scoring algorithm as an iteratively weighted least squares (IWLS) scheme yields the following system of equations:

$$\begin{pmatrix} X'WX & X'WZ \\ Z'WX & Z'WZ + Q^{-1} \end{pmatrix} \begin{pmatrix} \beta \\ b \end{pmatrix} = \begin{pmatrix} X'W\tilde{y} \\ Z'W\tilde{y} \end{pmatrix}, \tag{11.5}$$

which has to be solved iteratively to obtain posterior mode estimates. The matrix of working weights $W = DS^{-1}D$ has a blockdiagonal structure established by the blockdiagonal matrices $D = \text{blockdiag}(D_1 \ \dots \ D_n)$ and $S = \text{blockdiag}(S_1 \ \dots \ S_n)$, with $q \times q$ matrices D_i and S_i given by

$$D_i = \frac{\partial h(\eta_i)}{\partial \eta} = \begin{pmatrix} \frac{\partial h^{(1)}(\eta_i)}{\partial \eta^{(1)}} & \dots & \frac{\partial h^{(q)}(\eta_i)}{\partial \eta^{(1)}} \\ \vdots & \ddots & \vdots \\ \frac{\partial h^{(1)}(\eta_i)}{\partial \eta^{(q)}} & \dots & \frac{\partial h^{(q)}(\eta_i)}{\partial \eta^{(q)}} \end{pmatrix}$$

and

$$S_i = \text{cov}(y_i) = \begin{pmatrix} \pi_i^{(1)}(1 - \pi_i^{(1)}) & -\pi_i^{(1)}\pi_i^{(2)} & \dots & -\pi_i^{(1)}\pi_i^{(q)} \\ -\pi_i^{(1)}\pi_i^{(2)} & \ddots & & \vdots \\ \vdots & & \ddots & -\pi_i^{(q-1)}\pi_i^{(q)} \\ -\pi_i^{(1)}\pi_i^{(q)} & \dots & -\pi_i^{(q-1)}\pi_i^{(q)} & \pi_i^{(q)}(1 - \pi_i^{(q)}) \end{pmatrix}.$$

The working observations $\tilde{y}$ are defined by

$$\tilde{y} = \eta + (D^{-1})'(y - \pi).$$

The system of equations (11.5) defines a general way of estimating regression coefficients in categorical STAR models. The different models for ordered and unordered responses

only vary in the specific structure of the design matrices and the derivatives of the response function with respect to the linear predictor in the matrices D_i. For the multinomial logit model, these derivatives can be shown to be given by

$$\frac{\partial h^{(r)}(\eta_i)}{\partial \eta^{(j)}} = \begin{cases} \dfrac{\exp(\eta^{(r)})\left(1+\sum_{s=1,s\neq r}^{q}\exp(\eta^{(s)})\right)}{\left(1+\sum_{s=1}^{q}\exp(\eta^{(s)})\right)^2} = \pi^{(r)}(1-\pi^{(r)}) & r=j \\ -\dfrac{\exp(\eta^{(r)})\exp(\eta^{(j)})}{\left(1+\sum_{s=1}^{q}\exp(\eta^{(s)})\right)^2} = -\pi^{(r)}\pi^{(j)} & r\neq j. \end{cases}$$

Therefore, we have $D_i = S_i$ which leads to the simplified working weights $W_i = D_i = S_i$.

In case of cumulative and sequential models, the derivatives of h depend on the density f of the latent variables. For cumulative models, we obtain

$$\frac{\partial h^{(r)}(\eta_i)}{\partial \eta^{(j)}} = \begin{cases} f(\eta^{(j)}) & j=r, \\ -f(\eta^{(j)}) & j=r-1, \\ 0 & \text{else}, \end{cases}$$

while for sequential models the derivatives are given by

$$\frac{\partial h^{(r)}(\eta_i)}{\partial \eta^{(j)}} = \begin{cases} f(\eta^{(r)})\prod_{s=1}^{r-1}(1-F(\eta^{(s)})) & j=r, \\ -F(\eta^{(r)})f(\eta^{(j)})\prod_{s=1,s\neq j}^{r-1}(1-F(\eta^{(s)})) & j<r, \\ 0 & j>r. \end{cases}$$

11.3 Marginal likelihood for variance components

The aim of estimating the variance components will be approached on the basis of the marginal likelihood discussed in Section 5.3.2 for univariate responses. This means, we want to maximize the marginal likelihood

$$L^*(Q) = \int L(\beta, b, Q)d\beta db \tag{11.6}$$

with respect to the variance parameters in Q.

Since direct evaluation of the integral in (11.6) is not possible in general, we use a Laplace approximation to $L(\beta, b, Q)$ which, in fact, is equivalent to the approximation made in IWLS in the last section. This results in the restricted log-likelihood or marginal log-likelihood

$$l^*(Q) \approx -\frac{1}{2}\log(|\Sigma|) - \frac{1}{2}\log(|X'\Sigma^{-1}X|) - \frac{1}{2}(\tilde{y}-X\hat{\beta})'\Sigma^{-1}(\tilde{y}-X\hat{\beta}), \tag{11.7}$$

where $\Sigma = W^{-1} + Z'QZ$ is an approximation to the marginal covariance of $\tilde{y}|b$. Maximization of (11.7) can now be conducted by Newton Raphson or Fisher Scoring, compare

the formulae presented in Section 5.3.2. There, we also derived numerically feasible expressions for the derivatives of (11.7) allowing for the computation of REML estimates even for fairly large data sets. Although these expressions were derived for univariate responses, they can also be used within a multicategorical setting. One should, however, keep in mind that the weight matrix W is no longer diagonal but blockdiagonal.

11.4 Mixed model based inference in categorical STAR

Estimation of multicategorical STAR models can now be summarized in the following algorithm:

(i.) Compute the design matrices of the mixed model representation of the categorical STAR model.

(ii.) Update $\hat{\beta}$ and $\hat{b}$ given the current variances using IWLS, i. e. solve the system of equations (11.5).

(iii.) Update the variances in Q using marginal likelihood estimation.

(iv.) Compute relative changes of the estimates and return to (ii.) if they are above a certain threshold. Otherwise stop the estimation process.

Based on the final estimates $\hat{\beta}$ and $\hat{b}$, estimates for the function evaluations can be defined by inserting $\hat{\beta}$ and $\hat{b}$ into (11.1) to (11.3), e. g.

$$\hat{f}_j^{(r)}(\nu_{ij}) = x'_{ij}\hat{\beta}_j^{(r)} + z'_{ij}\hat{b}_j^{(r)}$$

or

$$f_j(\nu_j^{(r)}) = x_{ij}^{(r)\prime}\hat{\beta}_j + z_{ij}^{(r)\prime}\hat{b}_j.$$

These expressions also form the basis for the construction of credible intervals. The inverse of the coefficient matrix in (11.5) can be shown to be the approximate covariance matrix of $\hat{\beta}$ and $\hat{b}$ and standard deviations for linear combinations of the estimated coefficients can be easily calculated (see also the comments in Section 5.2.1).

11.5 MCMC inference based on latent variables

Similarly to the discussion in Section 5.5, we will now give a brief overview over fully Bayesian inference in multicategorical structured additive regression based on MCMC.

For most models with categorical responses, efficient sampling schemes based on latent utility representations can be developed. The seminal paper by Albert & Chib (1993) develops algorithms for probit models with ordered categorical responses. The case of probit models with unordered multicategorical responses is dealt with e. g. in Fahrmeir & Lang (2001b). Recently another important data augmentation approach for categorical logit models has been presented by Holmes & Held (2006). The adaption of these sampling schemes to structured additive regression models is more or less straightforward.

We briefly illustrate the concept for ordinal data in a cumulative probit model as discussed in Section 10.1.3. Recall that the model was formulated based on latent utilities

$$L = \eta + \varepsilon$$

as

$$Y = r \quad \Leftrightarrow \quad \theta^{(r-1)} < L \leq \theta^{(r)}.$$

In a probit model, the error term ε is standard Gaussian and, hence, the latent utility is also Gaussian distributed given the predictor η. Augmenting the latent variables L_i as additional parameters, an additional sampling step for updating the L_i is required. Fortunately, sampling the L_i's is relatively easy and fast because in a probit model, the full conditionals are truncated normal distributions. More specifically $L_i|\cdot$ follows a normal distribution $N(\eta_i, 1)$ truncated at the left by $\theta^{(r-1)}$ and truncated at the right by $\theta^{(r)}$ if $y_i = r$. The advantage of defining a probit model through the latent variables L_i is that the full conditionals for the regression parameters ξ_j (and γ) are Gaussian with precision matrices and means given by similar expressions as for Gaussian responses in Section 5.5 but with y_i replaced by L_i. Hence, the efficient and fast sampling schemes for Gaussian responses can be used with slight modifications. Similar updating schemes may be developed for multinomial probit models with unordered categories (see Fahrmeir & Lang (2001b) for details) or sequential probit models.

For categorical logit models updating schemes based on latent utilities can be based on ideas presented in Holmes & Held (2006) but here additional hyperparameters have to be introduced to obtain Gaussian full conditionals for the regression coefficients. Alternatively, updating schemes based on IWLS proposals constructed in analogy to the description for univariate responses can be used (see Section 5.5 and Brezger & Lang (2005)).

12 A space-time study in forest health

As an application of categorical structured additive regression models, we will now analyze the spatio-temporal data set on forest health described in the introductory Section 2.2. The intention is to find factors influencing the health status of beeches in a northern Bavarian forest district.

In addition to temporal and spatial information, numerous covariates characterizing the stand and the site of the tree, as well as the soil at the stand are given (see Table 2.2 on page 10). Both continuous and categorical covariates are available and have to modeled appropriately. In a first exploratory analysis, all continuous covariates were included as penalized splines but it turned out that a reduced model leads to a comparable fit with a much smaller number of parameters. Especially effects of tree-specific covariates that do not vary over time were estimated to be approximately linear and, hence, could be included in the parametric part of the predictor. This led to the cumulative probit model

$$P(Y_{it} \leq r) = \Phi(\theta^{(r)} - [f_1(t) + f_2(a_{it}) + f_3(t, a_{it}) + f_{spat}(s_i) + u'_{it}\gamma]), \tag{12.1}$$

where Φ is the standard normal cumulative distribution function, $Y_{it} \in \{1, 2, 3\}$ denotes the damage state of tree i, $i = 1, \ldots, 83$ at time t, $t = 1983, \ldots, 2004$, $f_1(t)$ is a flexible time trend, $f_2(a_{it})$ is a nonparametric age effect, $f_3(t, a_{it})$ is an interaction surface, $f_{spat}(s_i)$ is a spatial effect and $u'_{it}\gamma$ comprises all further covariates with parametric effects. Effect coding was used for categorical covariates and the category with the highest frequency was chosen as the reference category. The nonparametric effects f_1 and f_2 were modeled using cubic P-splines with 20 inner knots and second order random walk prior. For the interaction effect we assumed a cubic bivariate P-spline with 20 inner knots for each of the interacting variables and a first order random walk prior based on the four next neighbors. For the spatial effect, various parameterizations were examined, compare the next section. Note that in model (12.1) effects have to be interpreted in the following way: Higher (lower) values of covariate effects correspond to worse (healthier) state of the trees (compare the illustration in Figure 10.1 on page 145).

12.1 Comparison of spatial smoothing techniques

As discussed in Section 4.2, several alternatives are available to model spatial effects in the present application. In this section, we will compare the following four approaches:

- Model f_{spat} by a Markov random field with neighborhoods based on a distance measure. We used a radius of 1.2 kilometers around each tree to define neighborhoods.
- Model f_{spat} by a full Gaussian random fields. No dimension reduction is needed since the number of trees is comparably small. For the correlation function we chose a Matérn function with $\nu = 1.5$ and scale parameter α obtained in a pre-processing step according to the rule specified in (4.30) on page 44.
- Model f_{spat} by a cubic bivariate penalized spline with first order random walk penalty and 12 inner knots for each direction.
- Neglect spatial correlations, i. e. do not include any spatial effect at all.

Within BayesX (see Section 6 for a general introduction), cumulative probit models are estimated using a series of commands like the following:

```
> dataset d
> d.infile using c:\data\beeches.dat
> remlreg r
> r.regress bu3 = x*y(kriging, full) + age(psplinerw2)
  + time(psplinerw2) + age*time(pspline2dimrw1) + elevation
  + ... + saturation4, family=cumprobit using d
```

First, we create a dataset object and store the available variables in this object. Afterwards, we can estimate the regression model, where we have to specify the option `family=cumprobit` to obtain a cumulative probit model. Cumulative logit models are requested by `family=cumlogit`. Similar as in Section 7, univariate P-splines for the main effects of age and calendar time are specified by the terms `age(psplinerw2)` and `time(psplinerw2)`. In addition, we consider a bivariate P-spline with first order random walk prior in the term `age*time(pspline2dimrw1)` to account for interactions between age and calendar time. Different types of priors can be requested by `pspline2dimrw2` (second order random walk based on the Kronecker sum of two univariate second order random walks) or `pspline2dimbiharmonic` (second order random walk based on an approximation of the biharmonic differential operator), see Section 4.2.6 for theoretical details.

In the above example, the spatial effect is estimated by the term `x*y(kriging, full)`. The option `full` specifies that all observation points shall be used for the Kriging estimate, i. e. no low-rank approximation is to be employed. Alternatively, the Kriging term may be replaced by a bivariate P-spline (`x*y(pspline2dimrw1)`) or a Markov random field (`tree(spatial,map=m)`). In the latter case, a map object containing the adjacency information has to be created in addition and the corresponding tree indices have to be specified as the covariate.

	no spatial effect	MRF	GRF	2d P-spline
-2*log-likelihood	1641.02	1122.76	1128.94	1164.62
degrees of freedom	71.68	118.82	117.92	114.01
AIC	1784.38	1360.39	1364.78	1392.63
BIC	2178.04	2012.89	2012.35	2018.73
GCV	0.844	0.546	0.550	0.570

Table 12.1: Forest health data: Information criteria and generalized cross-validation for models with different spatial effects.

Table 12.1 presents characteristics of the model fit obtained with the four different spatial model specifications. Obviously, including a spatial effect leads to an improved fit, regardless of the chosen parametrization. Differences between the spatial models are smaller, but the Markov random field model performs best in terms of Akaikes information criterion and the generalized cross validation statistic. When considering the Bayesian information criterion, the GRF model is only slightly better and therefore we will focus on the MRF model in the following sections.

variable	$\hat{\gamma}_j$	std. dev.	p-value	95% ci	
$\theta^{(1)}$	-1.602	1.740	0.358	-5.012	1.809
$\theta^{(2)}$	1.943	1.740	0.265	-1.468	5.353
elevation	-0.002	0.003	0.578	-0.008	0.004
inclination	0.019	0.014	0.199	-0.010	0.047
soil	-0.001	0.013	0.920	-0.028	0.025
ph	-0.054	0.210	0.797	-0.466	0.358
canopy	-2.319	0.364	<0.0001	-3.033	-1.606
stand	0.239	0.125	0.055	-0.006	0.485
fertilization	-0.397	0.243	0.102	-0.872	0.079
humus0	-0.244	0.107	0.022	-0.453	-0.035
humus1	-0.125				
humus2	0.141	0.086	0.100	-0.027	0.308
humus3	0.124	0.101	0.221	-0.074	0.322
humus4	0.104	0.141	0.462	-0.172	0.380
moisture1	-0.647	0.290	0.026	-1.216	-0.078
moisture2	0.292				
moisture3	0.355	0.199	0.074	-0.035	0.744
saturation1	0.183	0.295	0.533	-0.394	0.761
saturation2	-0.517				
saturation3	-0.300	0.304	0.325	-0.896	0.297
saturation4	0.634	0.397	0.110	-0.145	1.413

Table 12.2: Forest health data: Posterior mode estimates for fixed effects when the spatial effect is modeled by a Markov random field.

Table 12.2 summarizes estimation results for the parametric effects in (12.1) when using a Markov random field prior for the spatial effect. The only significant effect (at the 1% level) is the effect of the forest canopy density. An increased density leads to higher probabilities for lower categories and, hence, for a healthy state of the tree. Note that this conclusion is specific for beeches and does not necessarily hold for other tree species. Borderline significant effects are obtained for the type of stand and fertilization. While deciduous forest has higher probabilities of being damaged, fertilization has a positive influence on the health status (corresponding to a negative regression coefficient). Interestingly, the effect of the pH-value is not significant, although it is usually assumed to be an important determinant of forest health. However, in our example most of the trees share a similar level of the pH-value since they are all located within a relatively small observation area. Therefore, the insignificant effect is likely to be caused by the fact that there is only marginal variation in the pH-value between the trees.

Regarding the categorical covariates humus, moisture and saturation, there is usually one category with almost significant effect (note that no standard errors are available for the effects of the reference category). An increasing thickness of the humus layer seems to have a negative effect on the health status of the trees (corresponding to increasing regression coefficients). Concerning the level of soil moisture, the effects are very close for the second and the third category. Apparently, dry soil has a positive effect on the health status of the trees. For the effect of the base saturation, no clear conclusions can

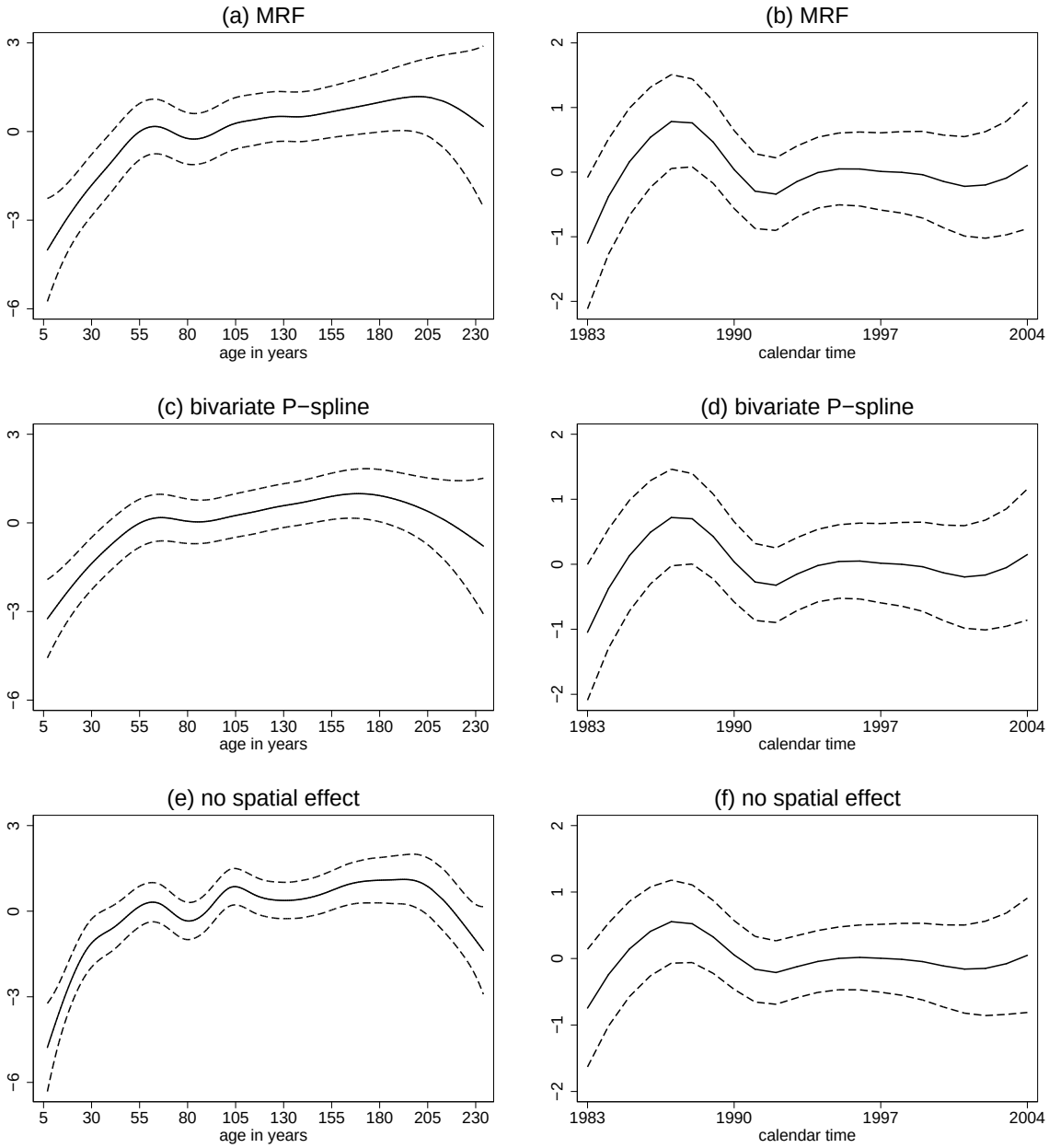


Figure 12.1: Forest health data: Posterior mode estimates for the nonparametric effects (solid line) and 95% credible intervals (dashed lines).

be drawn. While a small and a very high base saturation increase the probability for higher damage states, moderate values lead to healthier trees. Therefore, extreme values seem to have a negative influence.

Figure 12.1 displays the estimates of the time trend and the nonparametric age effect obtained with a MRF and a bivariate P-spline for the spatial effect, and with a model that neglects spatial correlations. While both spatial models lead to similar estimates for the time trend, that recover the trend for slightly damaged trees shown in Figure 2.4, the age effect is somewhat different. With a MRF, the age effect is rapidly increasing for young trees, reaching a maximum at an age of 55 years. Afterwards, it is decreasing for a

short period followed by a slight increase up to an age of about 210 years. With bivariate P-splines a much smoother, almost inverse u-shaped estimate is obtained, resulting in smaller effective degrees of freedom (as indicated in Table 12.1) but also a declined model fit. In contrast, excluding the spatial effect results in a wigglier estimate for the age effect with an additional peak around 105 years. Obviously, the age effect absorbs some of the effects which are otherwise covered by the spatial component. The time trend has a similar functional form but is less expressed when excluding the spatial component.

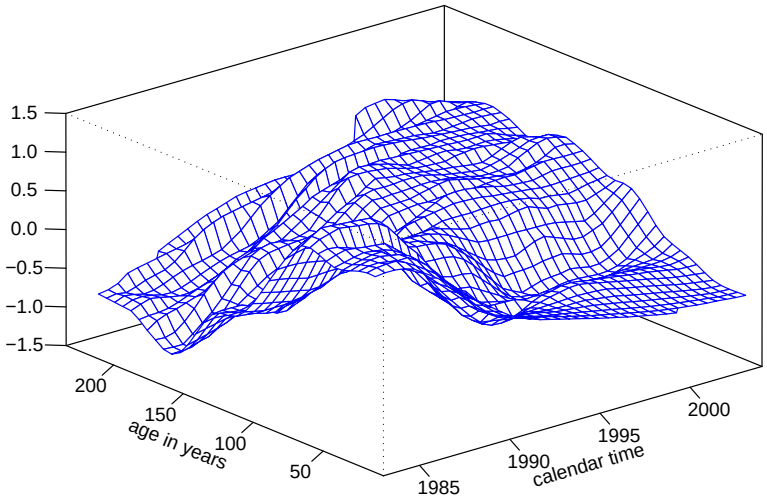


Figure 12.2: Forest health data: Posterior mode estimates for the interaction effect when the spatial effect is modeled by a Markov random field.

The estimated interaction effect between calendar time and age is visualized in Figure 12.2. Apparently, young trees were in poorer health state in the eighties but recovered in the nineties, unlike older trees which showed the contrary behavior. A possible interpretation is that it takes longer until older trees are affected by harmful environmental circumstances while younger trees are affected nearly at once but manage to accommodate when growing older.

Estimates for spatial effects obtained with a Markov random field and a bivariate penalized spline are presented in Figure 12.3. The posterior modes at the observation points seem to show a similar spatial structure for both models, although the range of the spatial effect is somewhat smaller for the MRF. Surprisingly, this does not imply that the spline-based estimates lead to more significant effects in terms of pointwise posterior probabilities (middle row of Figure 12.3). Obviously, Markov random fields produce more precise estimates in the sense that the respective credible intervals are narrower. An advantage of penalized spline estimates is that they can be evaluated at arbitrary points and, therefore, naturally allow for interpolation as well as extrapolation. For Markov random fields, in contrast, estimates are only defined for the observation points themselves. Of course, we can perform linear interpolation within the convex hull formed by the observation points, but this induces a strong assumption about the structure of the underlying spatial effect. The last row of Figure 12.3 displays both the linear interpolated Markov random field and the penalized spline evaluated on a regular grid covering the observation area. From this presentation, differences between the two estimates can be seen much more clearly than from the presentation at the observation points only. In addition, at least the penalized

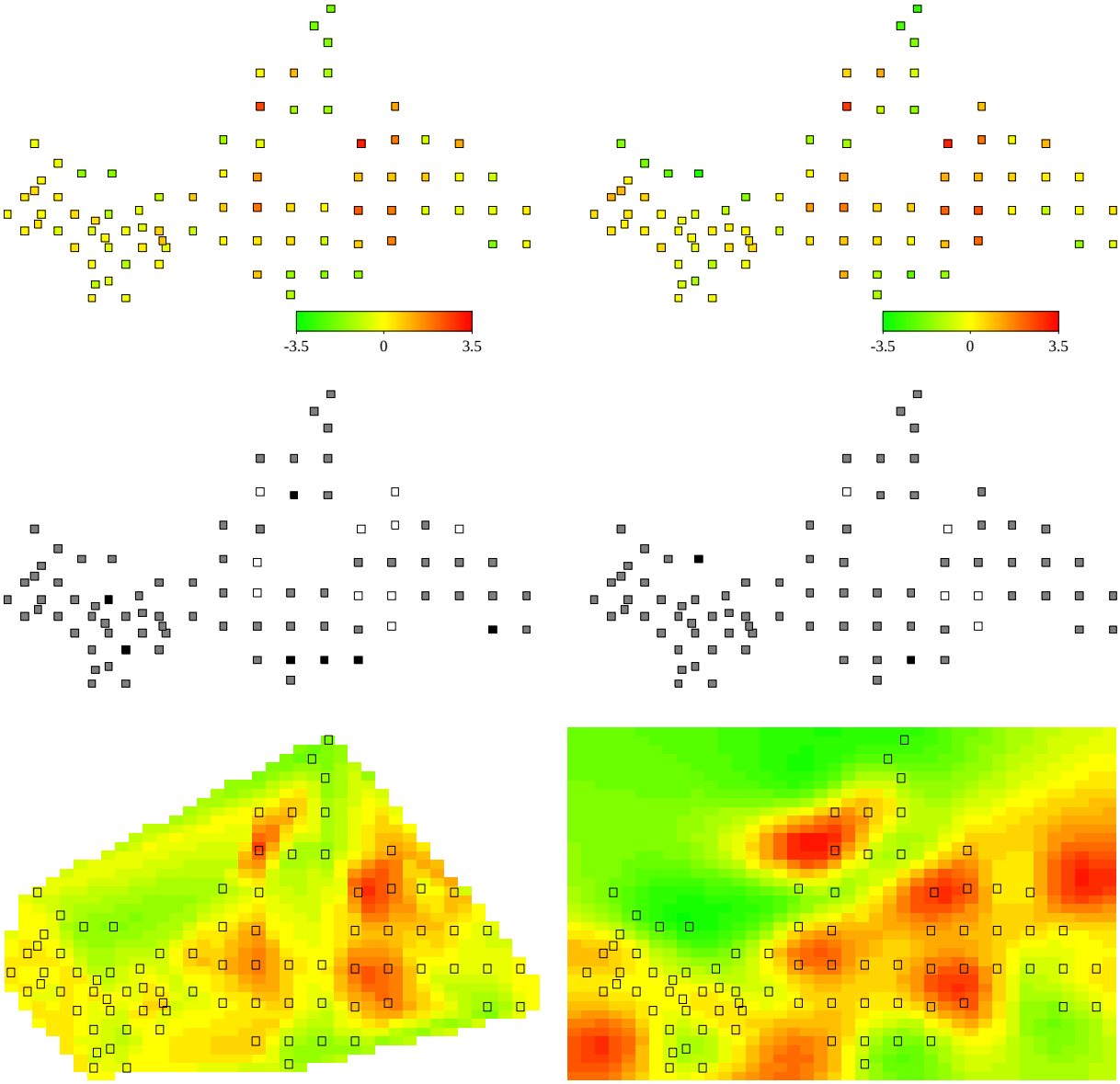


Figure 12.3: Forest health data: Spatial effects obtained with a Markov random field (left panel) and a bivariate P-spline (right panel). The upper row displays posterior mode estimates evaluated at the observation points, the middle row displays 95% posterior probabilities, where black (white) points denote trees with strictly negative (positive) credible intervals. The lower row displays a linear interpolation of the posterior modes for the MRF and the estimated posterior mode surface for the bivariate spline.

spline estimates allow to identify larger regions with increased or decreased risk. For the Markov random field, such an interpretation seems to be difficult due to the linear interpolation. In fact, the linear interpolation produces much more yellow points, where the estimated effect is approximately zero, while the penalized spline reveals a spatial pattern also in these parts of the observation area.

12.2 Comparison with fully Bayesian estimates

In addition to the mixed model based empirical Bayes approach, BayesX also supports fully Bayesian estimation of structured additive regression models with ordered responses. In this section, we will re-estimate model (12.1) using MCMC in combination with a Markov random field prior for the spatial effect and compare the results to those from the previous section.

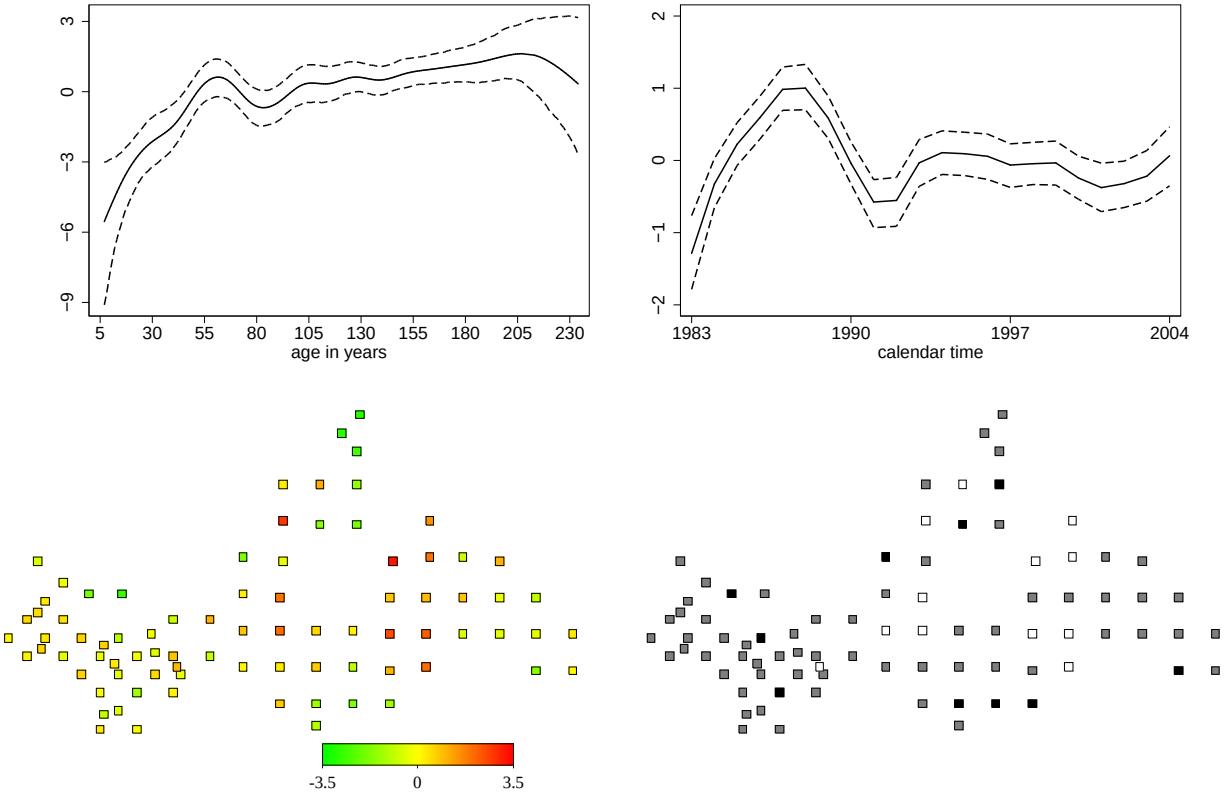


Figure 12.4: Forest health data: Fully Bayesian posterior mean estimates of nonparametric and spatial effects obtained with a Markov random field. The lower right figure indicates 95% posterior probabilities, where black (white) points denote trees with strictly negative (positive) credible intervals.

The resulting nonparametric and spatial effects are visualized in Figure 12.4. While the estimated spatial effect is very close to its empirical Bayesian counterpart, both in terms of point estimates and posterior probabilities, there are differences for the nonparametric effects. Although showing roughly the same functional form as in Figure 12.1, the posterior mean estimates are more wiggly and are, therefore, less intuitively interpretable. The credible intervals for the posterior means are more narrow than those obtained for the empirical Bayes posterior mode estimates. Considering the interaction effect (shown in Figure 12.5), differences between empirical Bayes estimates and fully Bayesian estimates are very small.

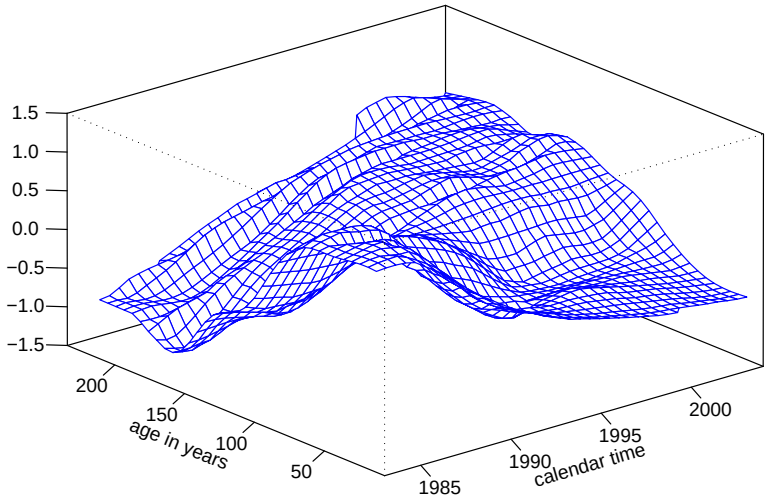


Figure 12.5: Forest health data: Fully Bayesian estimates of the interaction effect obtained with a Markov random field prior for the spatial effect.

12.3 Category-specific trends

The empirical time trends shown in Figure 2.4 on page 8 indicated very dissimilar trends for the three categories. In contrast, the regression model (12.1) assumes a common time trend and, in fact, the estimates shown in Figure 12.1 seem to be dominated by the trend for slightly damaged trees. To relax the assumption of a common time trend, we replaced the predictor in (12.1) by

$$\eta_{it}^{(r)} = \theta^{(r)} - \left[f_1^{(r)}(t) + f_2(a_{it}) + f_{spat}(s_i) + u_{it}'\gamma \right], \tag{12.2}$$

where $f_1^{(r)}(t)$ is a category-specific time trend. In contrast to (12.1), this model does not contain an interaction effect, since the category-specific time trend would also imply a category-specific interaction making model (12.2) hardly identifiable from the present data. We experimented with discrete interactions based on a categorized age effect but, due to the small number of cases in category '3', these models turned out to be either not

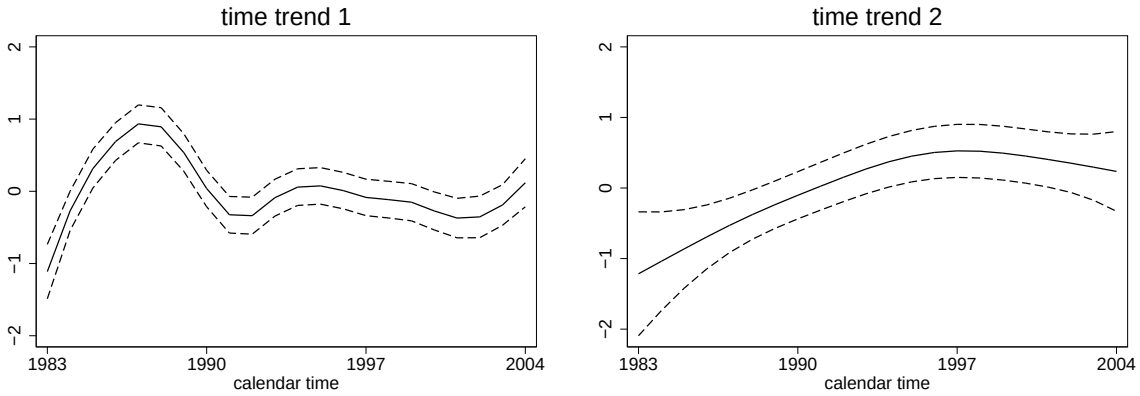


Figure 12.6: Forest health data: Posterior mode estimates of category-specific time trends when the spatial effect is modeled by a Markov random field.

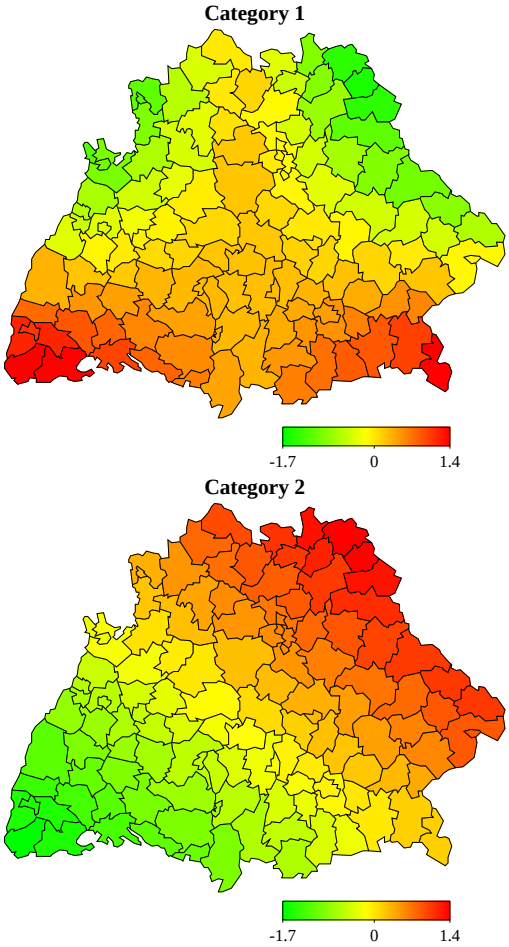
identifiable or not interpretable. Therefore, we decided to exclude the interaction term and to estimate the main effects model (12.2) for illustration purposes. The spatial effect is again assumed to follow a Markov random field prior.

Within BayesX, category-specific effects are requested by adding the additional option `catspecific` to the respective effect. In our example, the specification `time(psplinerw2)` has to replaced by `time(psplinerw2,catspecific)`. Note that parametric effects can also be specified to be category-specific, i. e. the specification of `x(catspecific)` corresponding to effects $x\gamma^{(r)}$ is also allowed.

Figure 12.6 visualizes the two estimated trend functions. The trend $f_1^{(1)}(t)$, which influences the first threshold, almost equals the overall trend obtained in model (12.1) except for the narrower credible intervals. In contrast, the second trend $f_1^{(2)}(t)$ exhibits a different shape. Starting from a negative value, it increases over most of the observation period before showing a slight decrease at the end of the nineties. This corresponds very well to the empirical trend of category '3' which shows a similar pattern (see Figure 2.4 on page 8).

13 Simulation studies for multicategorical responses

To investigate the performance of the presented mixed model based approach to categorical structured additive regression, we conducted several simulation studies based on a multinomial logit model and a cumulative probit model with three categories. In either model the predictors were defined to be the sum of a nonparametric effect and a spatial effect, see Figures 13.1 and 13.2 for detailed descriptions of the simulation design.



- Predictor:

$$\eta_i^{(r)} = f_1^{(r)}(x_i) + f_2^{(r)}(s_i)$$

- Category 1:

$$\begin{aligned} f_1^{(1)}(x) &= \sin[\pi(2x-1)] \\ f_2^{(1)}(s) &= -0.75|s_x|(0.5+s_y) \end{aligned}$$

- Category 2:

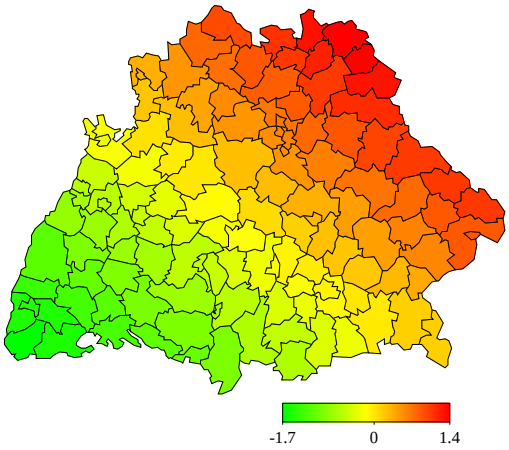
$$\begin{aligned} f_1^{(2)}(x) &= \sin[2\pi(2x-1)] \\ f_2^{(2)}(s) &= 0.5(s_x+s_y) \end{aligned}$$

- x is chosen from an equidistant grid of 100 values between -1 and 1.
- (s_x, s_y) are the centroids of the 124 districts s of the two southern states of Germany (see Figures).

Figure 13.1: Simulation design for the multinomial logit model.

13.1 Comparison of different modeling approaches

The first aim of this simulation study was the comparison of different parameterizations of the spatial effect and of different approaches to the estimation of categorical STAR models. In total, 250 simulation runs with $n = 500$ observations were generated from the multinomial logit model described in Figure 13.1. We used cubic P-splines with second order random walk penalty and 20 knots to estimate effects of the continuous covariate. The spatial effect was estimated either by a Markov random field, a (full) Gaussian random field or a two-dimensional P-spline (based on 10×10 inner knots). Concerning the competing fully Bayesian approach by Fahrmeir & Lang (2001b) and Brezger & Lang



- Predictor:

$$\eta_i^{(r)} = \theta^{(r)} - f_1(x_i) - f_2(s_i)$$

- Functions:

$$\begin{aligned} f_1(x) &= \sin[\pi(2x-1)] \\ f_2(s) &= 0.5(s_x + s_y) \end{aligned}$$

- x is chosen from an equidistant grid of 100 values between -1 and 1.
- (s_x, s_y) are the centroids of the 124 districts s of the two southern states of Germany (see Figure).

Figure 13.2: Simulation design for the cumulative probit model.

(2005), inverse Gamma priors $IG(a, b)$ with $a = b = 0.001$ were assigned to the variances. In the fully Bayesian framework, the GRF approach was computationally to demanding since it requires the inversion of a full precision matrix for the spatial effect in each iteration. Therefore, we excluded the fully Bayesian GRF approach from the comparison. As a further competitor, we utilized the R-implementation of the procedure polyclass described in Kooperberg et al. (1997), where nonparametric effects and interaction surfaces are modeled by linear splines and their tensor products. Smoothness of the estimated curves is not achieved by penalization but via stepwise inclusion and deletion of basis functions based on AIC.

The results of the simulation study can be summarized as follows:

- Generally, REML estimates have somewhat smaller median MSE than their fully Bayesian counterparts, with larger differences for spatial effects (see Figures 13.3a to 13.3d).
- Estimates for the effects of the continuous covariate are rather insensitive with respect to the model choice for the spatial effect (Figures 13.3a and 13.3b).
- Two-dimensional P-splines lead to the best fit for the spatial effect although data are provided with discrete spatial information (Figures 13.3c and 13.3d).
- Polyclass is outperformed by both the empirical and the fully Bayesian approach and, therefore, results are separately displayed in Figure 13.3e together with REML estimates based on two-dimensional P-splines. Presumably, the poor performance of polyclass is mainly caused by the special choice of linear splines, resulting in rather peaked estimates. Smoother basis functions, e. g. truncated cubic polynomials might improve the fit substantially but are not available in the implementation.
- Empirical and fully Bayesian estimates lead to comparable bias for the nonparametric effects. Results obtained with polyclass are less biased but show some spikes caused by the modeling with linear splines. Therefore, we can conclude that the poor performance of polyclass in terms of MSE is mainly introduced by additional

variance compared to empirical and fully Bayesian estimates (Figure 13.4).

- For spatial effects, both empirical and fully Bayesian estimates tend to oversmooth the data, i. e. estimates are too small for high values of the spatial functions and vice versa. In contrast, polyclass leads to estimates which are too wiggly and, therefore, overestimate spatial effects (Figures 13.5 and 13.6).
- For some simulation runs with spatial effects modeled by MRFs, no convergence of the REML algorithm could be achieved. This also happened when the spatial effect was modeled by a two-dimensional P-spline but in a much smaller number of cases. Obviously, the same convergence problems as described in Section 9 appear in a categorical setting. However, the arguments given there still hold and, hence, estimates obtained from the final (100-th) iteration are considered in these cases.

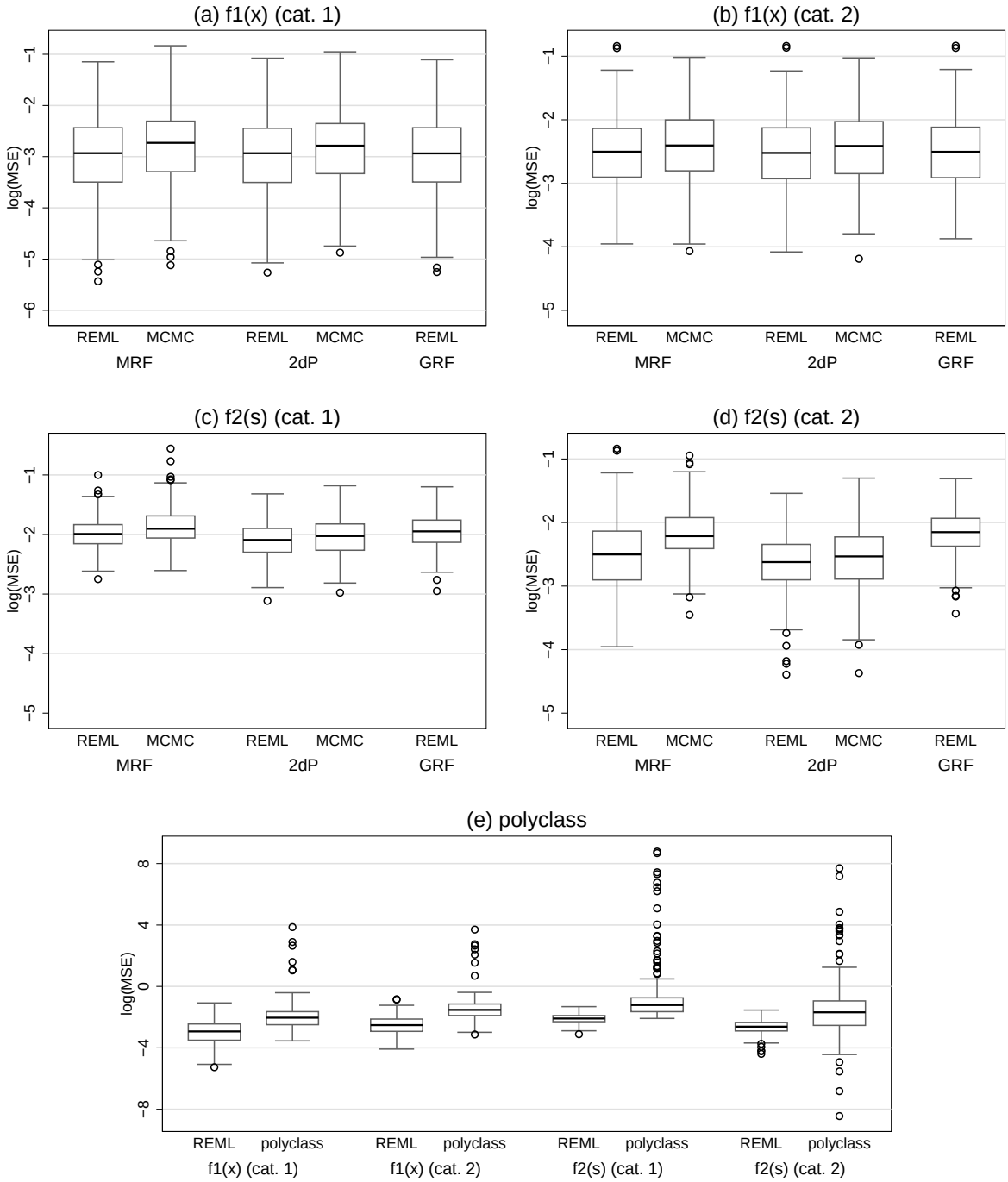


Figure 13.3: Comparison of different modeling approaches: Boxplots of log(MSE).

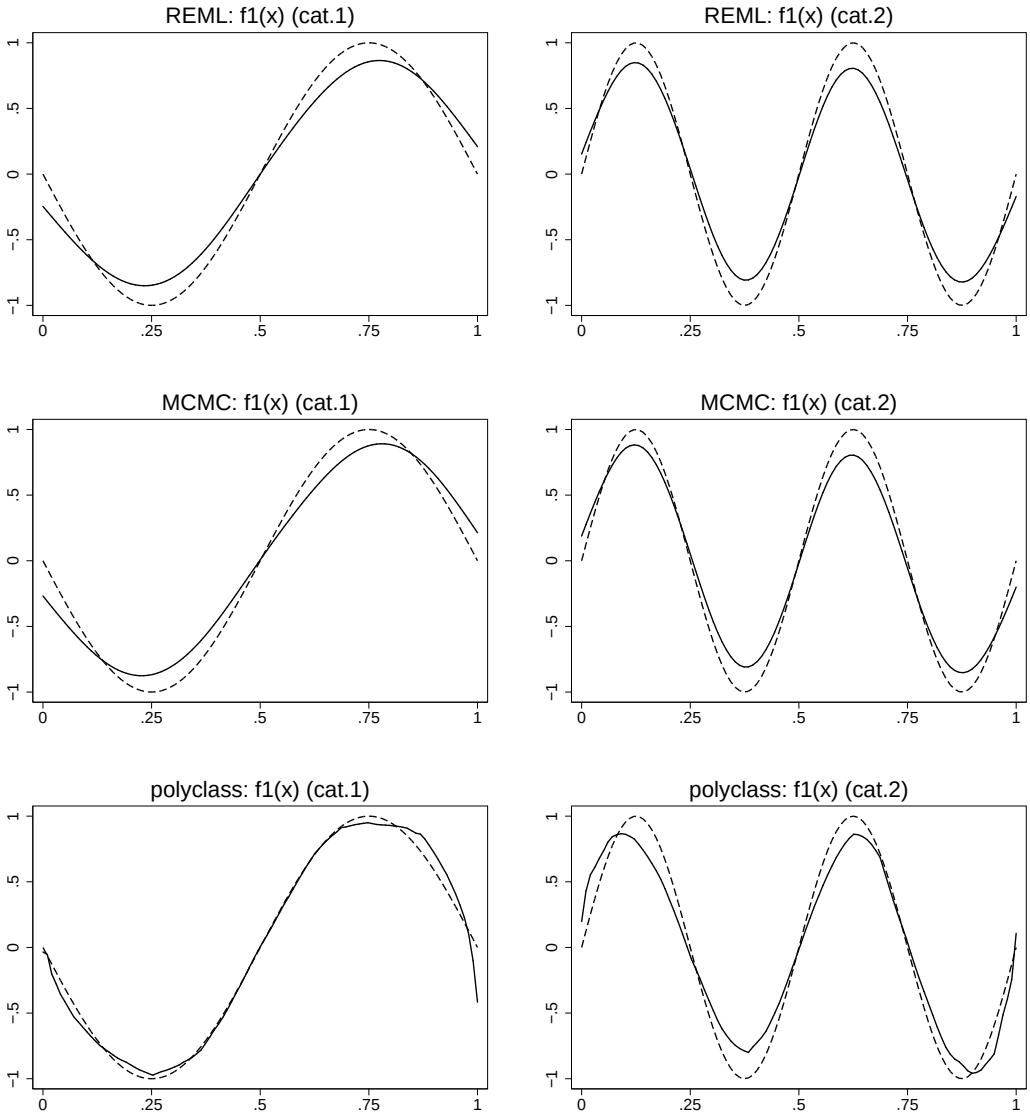


Figure 13.4: Comparison of different modeling approaches: Bias of nonparametric estimates. Estimates are displayed as solid lines, the true functions are included as dashed lines.

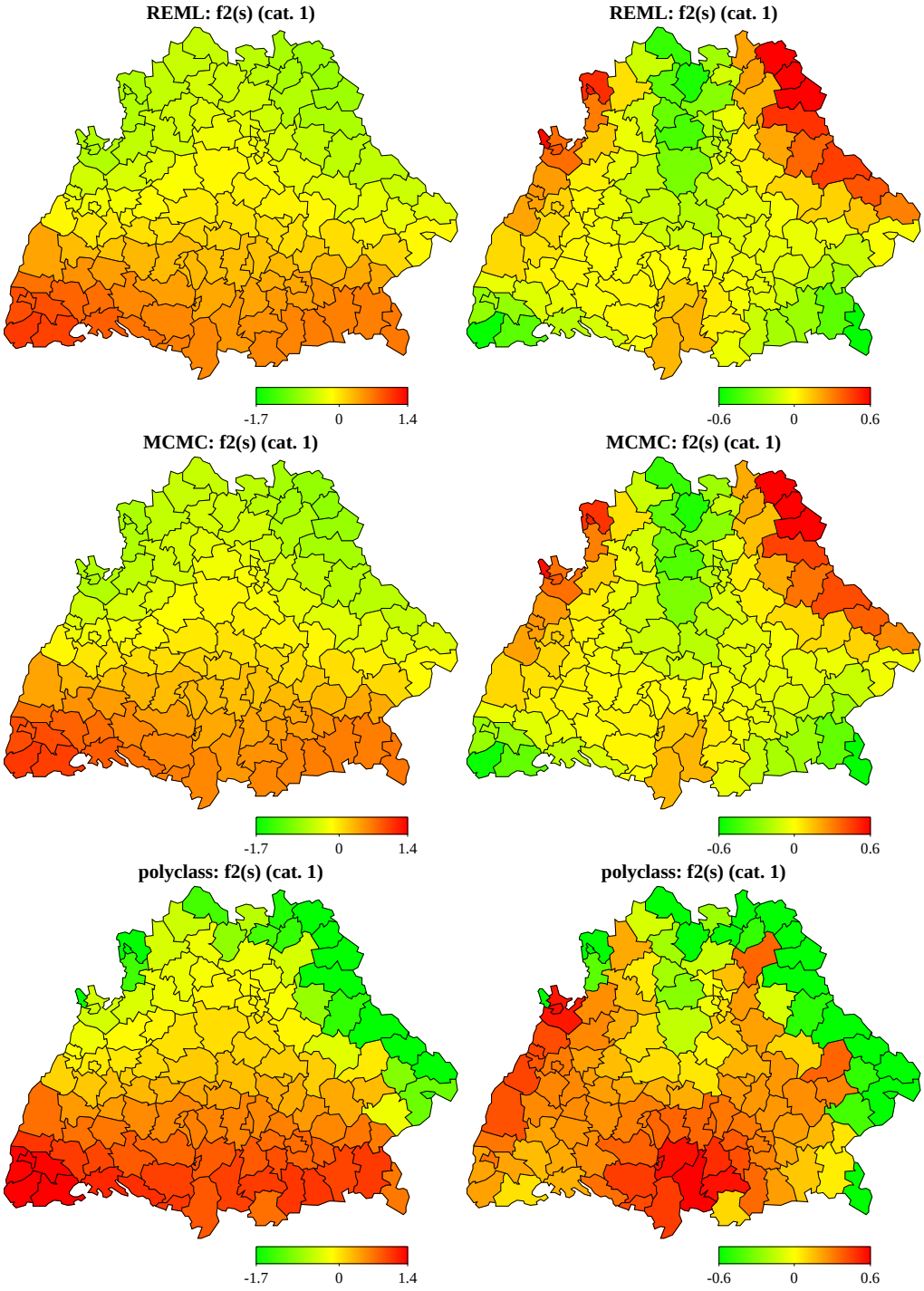


Figure 13.5: Comparison of different modeling approaches: Average estimates (left panel) and empirical bias (right panel) of estimates for $f_2^{(1)}(s)$.

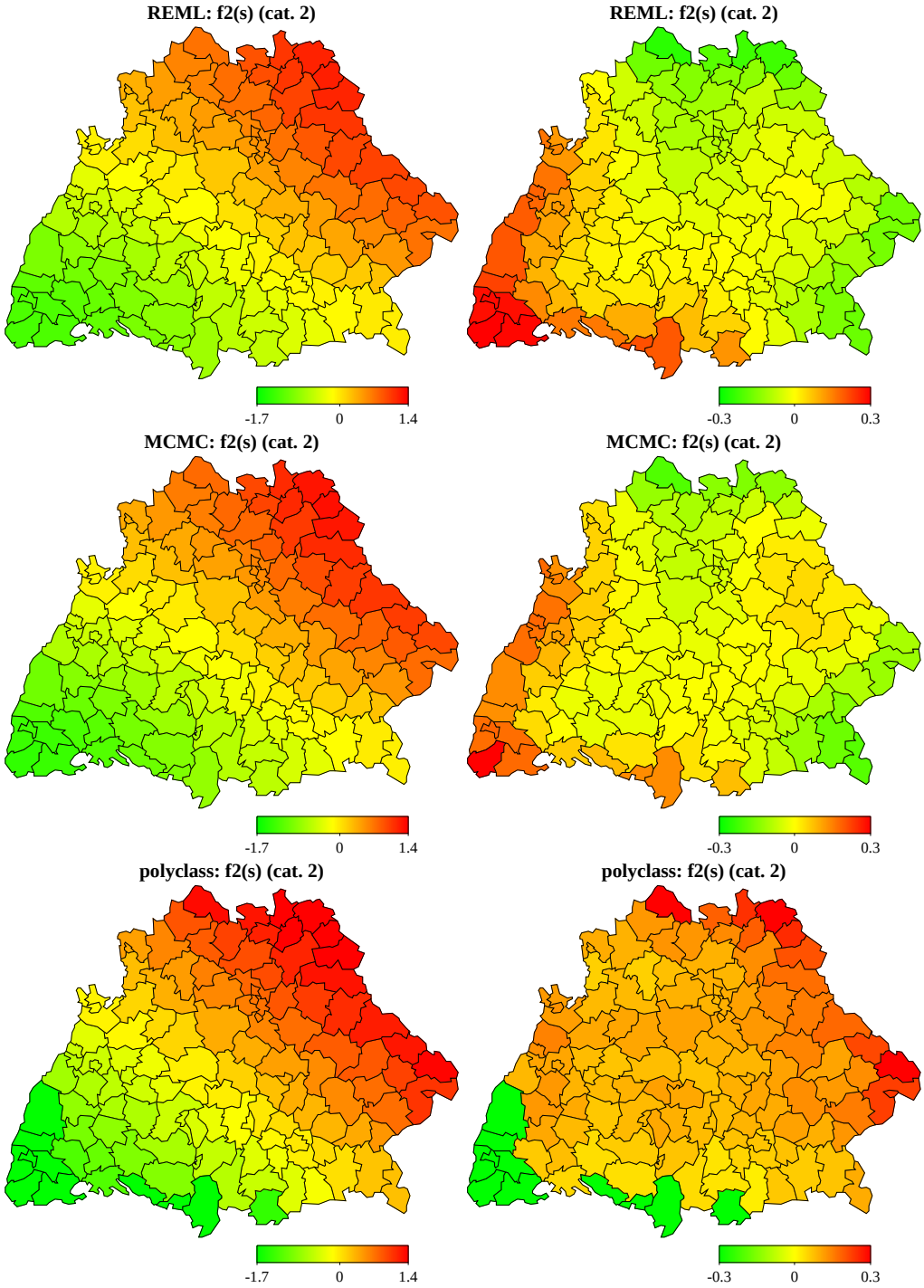


Figure 13.6: Comparison of different modeling approaches: Average estimates (left panel) and empirical bias (right panel) of estimates for $f_2^{(2)}(s)$.

13.2 Bias of REML estimates

It is frequently argued that results from REML estimation procedures in GLMMs tend to be biased due to the Laplace approximation involved, especially in sparse data situations (compare e. g. Lin & Breslow 1996). Therefore, as a second aim, we investigated whether this observation holds in a categorical setting in a second simulation study. Based on the models described in Figures 13.1 and 13.2, data sets with different sample sizes, namely $n = 500$, $n = 1000$ and $n = 2000$, were generated. Results from the REML estimation procedure were compared to their fully Bayesian counterparts which do not employ any approximations but work with the exact posterior. For both approaches, the spatial effect was estimated by a MRF while nonparametric effects were again modeled by cubic P-splines with second order random walk penalty and 20 inner knots.

The results of the simulation lead to the following conclusions:

- In general, the bias is smaller for MCMC estimates, most noticeably for more wiggly functions. For increasing sample sizes, differences almost vanish and both approaches give nearly unbiased estimates (Figures 13.7 to 13.12).
- REML estimates perform superior to MCMC estimates in terms of MSE for all sample sizes (Figures 13.13 and 13.14). Hence, although additional bias is introduced by considering mixed model based estimates, these estimates are in general preferable since they have a much smaller variation and, therefore, lead to smaller MSEs.

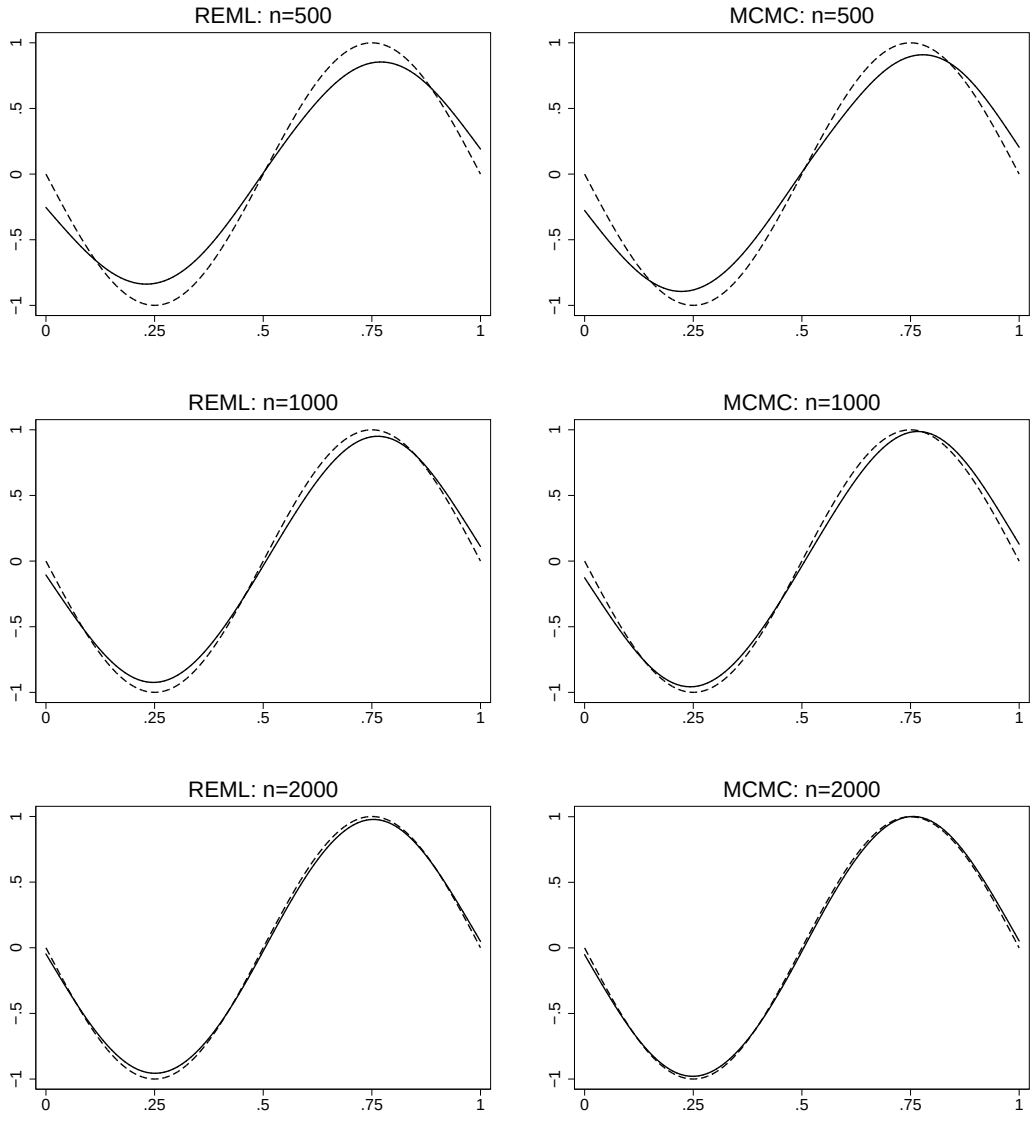


Figure 13.7: Multinomial logit model: Bias of nonparametric estimates for $f_1^{(1)}(x)$. Estimates are displayed as solid lines, the true functions are included as dashed lines.

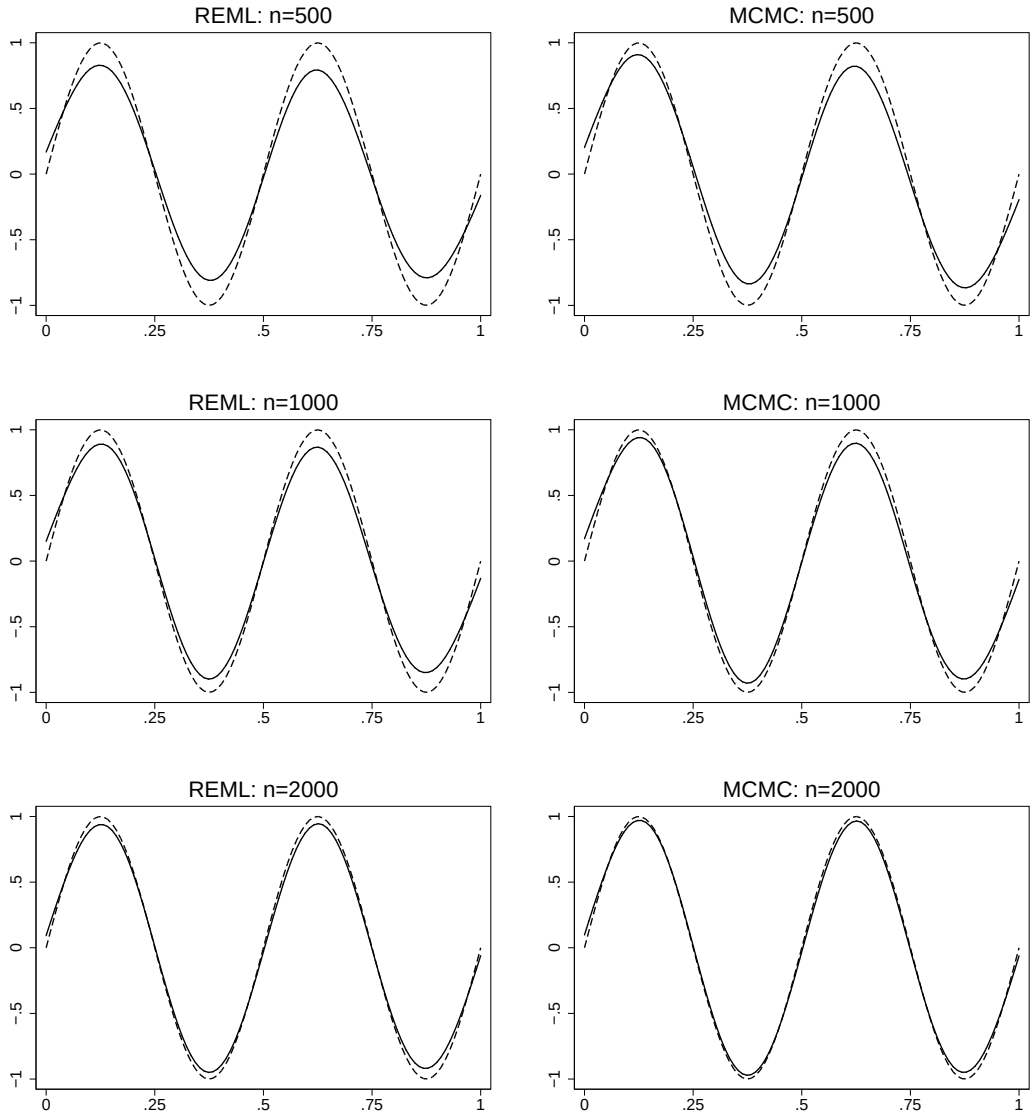


Figure 13.8: Multinomial logit model: Bias of nonparametric estimates for $f_1^{(2)}(x)$. Estimates are displayed as solid lines, the true functions are included as dashed lines.

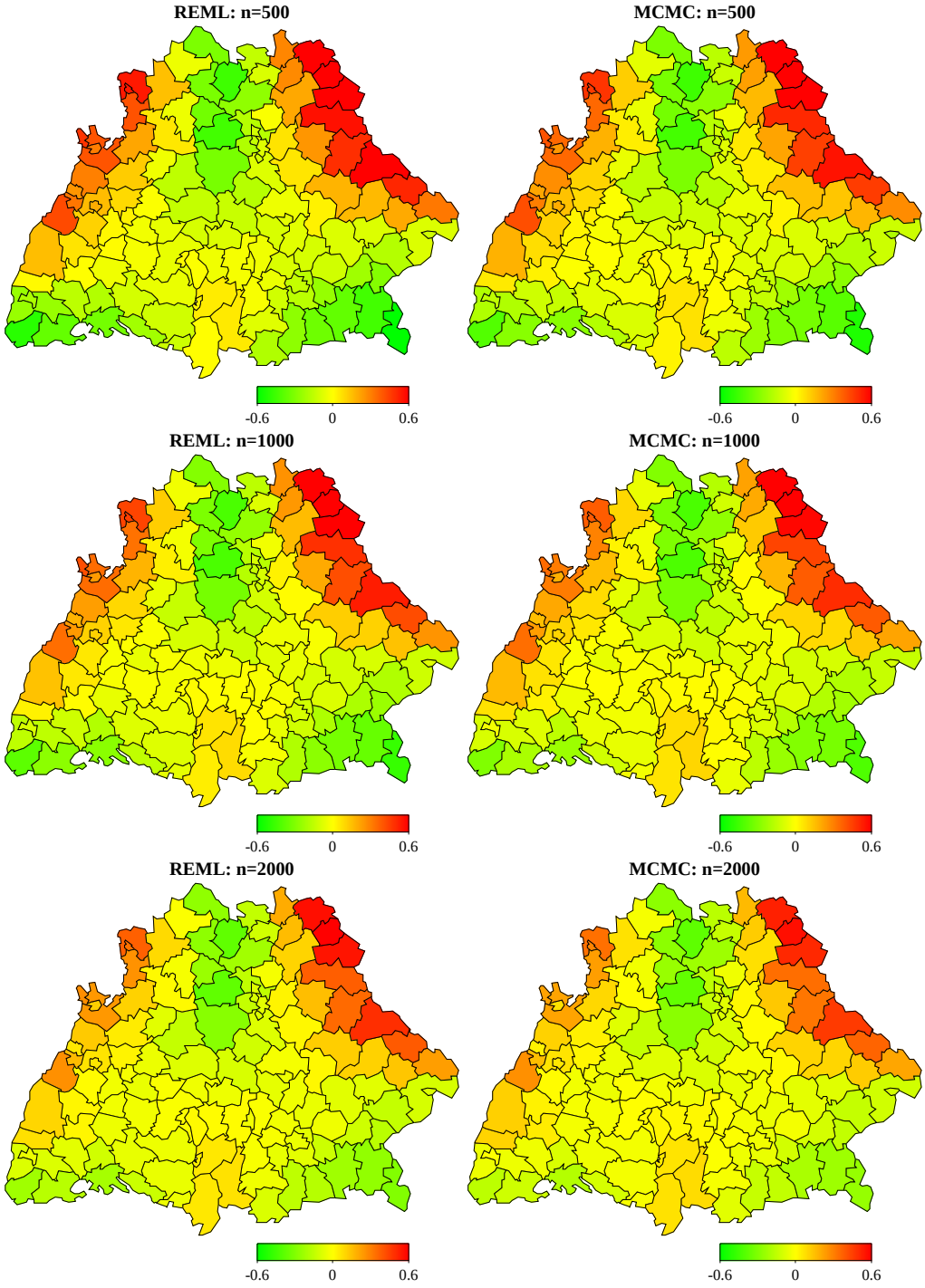


Figure 13.9: Multinomial logit model: Bias of spatial estimates for $f_2^{(1)}(s)$.

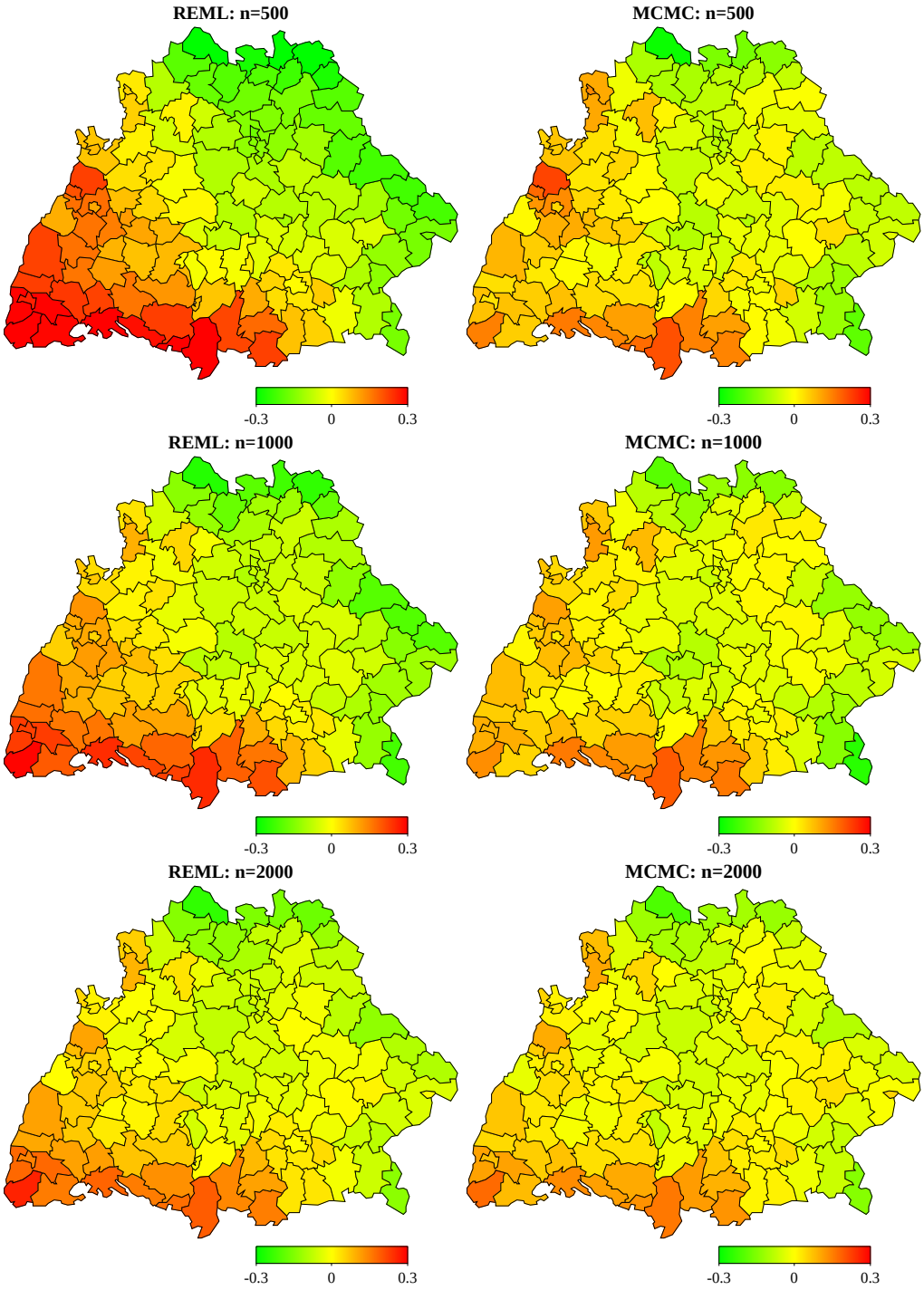


Figure 13.10: Multinomial logit model: Bias of spatial estimates for $f_2^{(2)}(s)$.

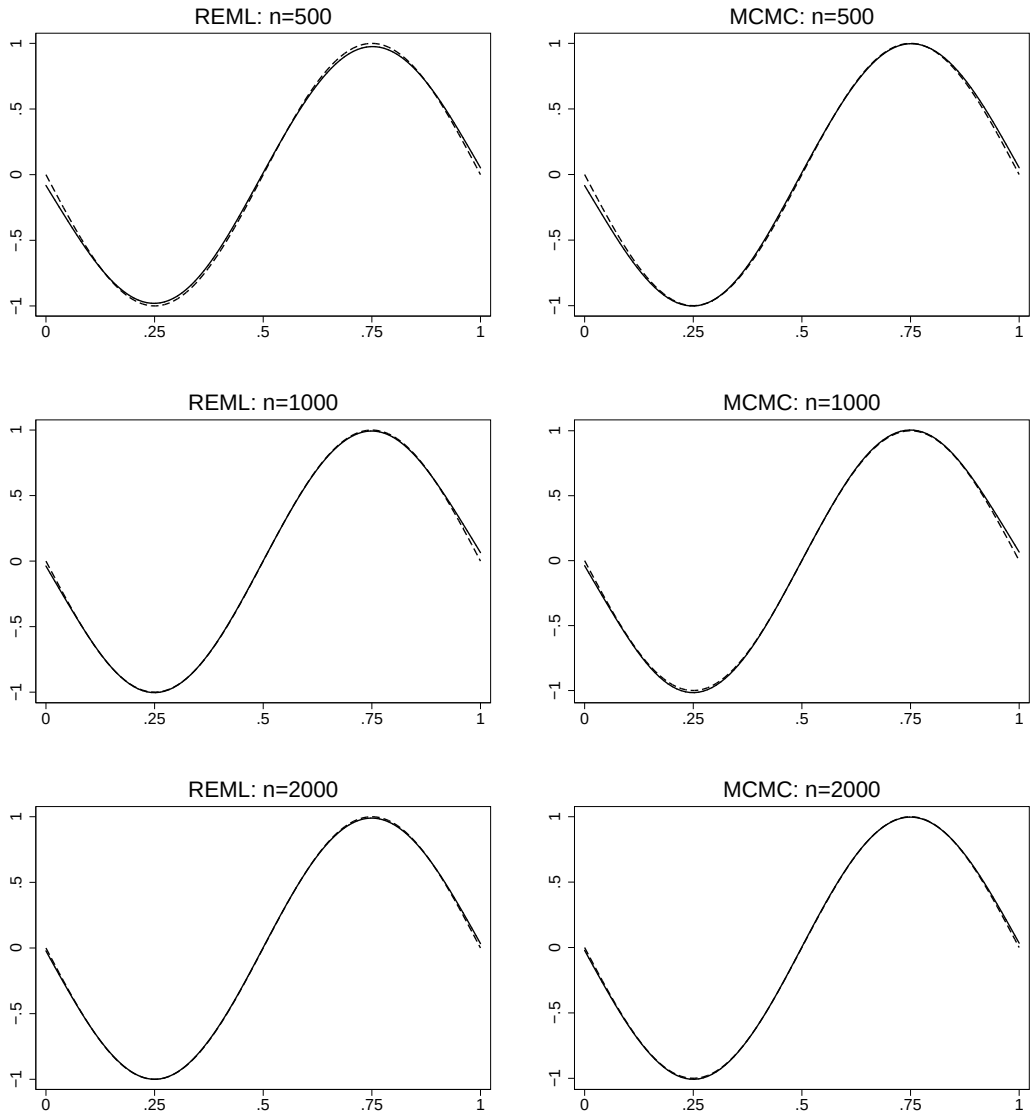


Figure 13.11: Cumulative probit model: Bias of nonparametric estimates for $f_1(x)$.

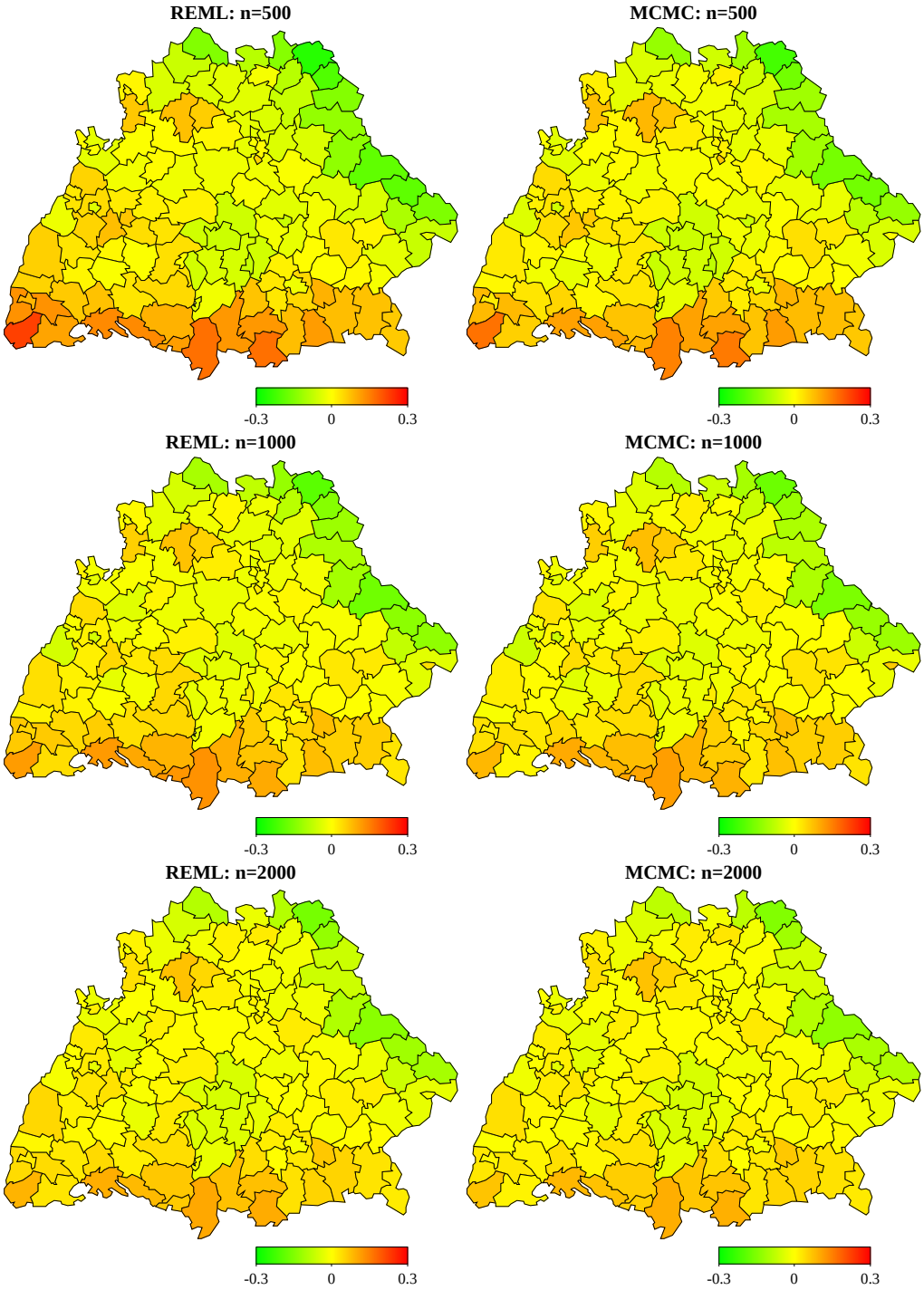


Figure 13.12: Cumulative probit model: Bias of spatial estimates for $f_2(s)$.

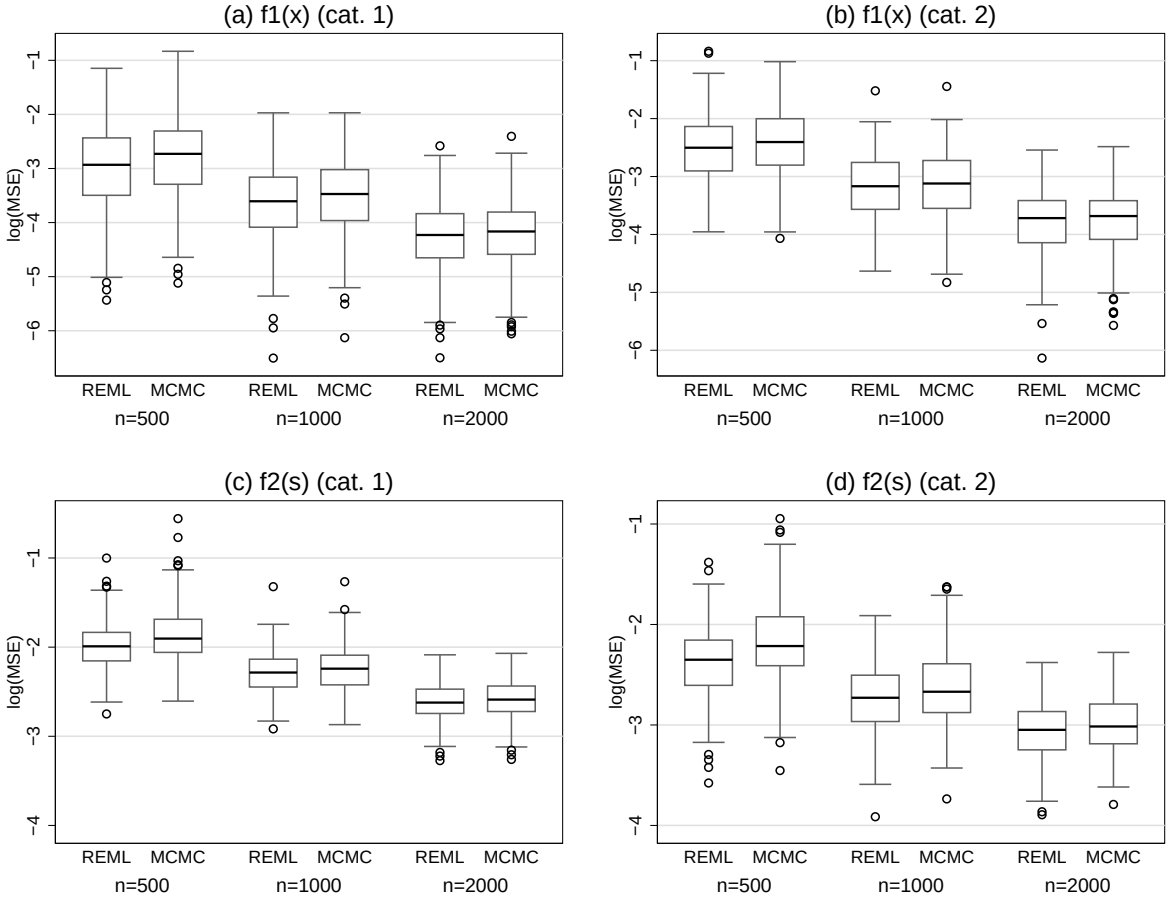


Figure 13.13: Multinomial logit model: Boxplots of log(MSE) for different sample sizes.

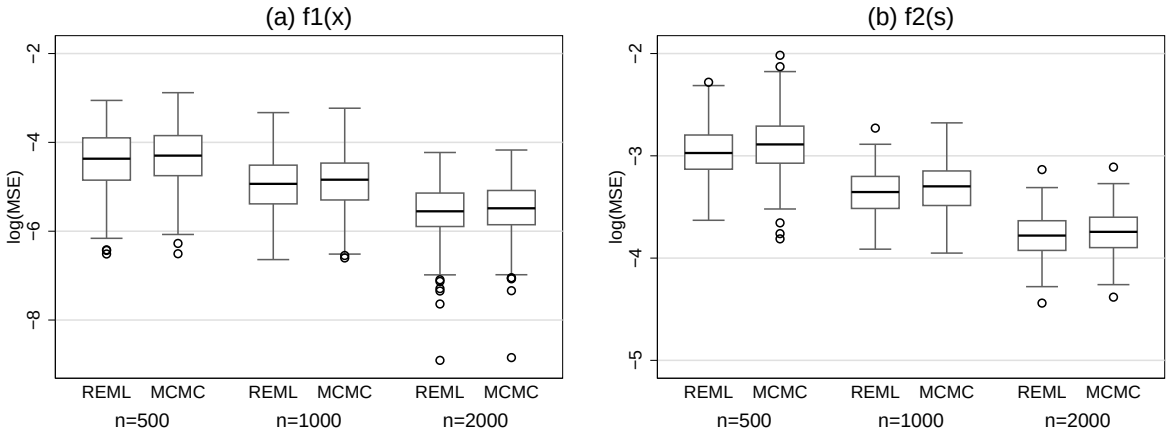


Figure 13.14: Cumulative probit model: Boxplots of log(MSE) for different sample sizes.

Part IV

Continuous survival times

14 Model formulation

14.1 Observation model

A specific type of regression models which receive considerable attention, especially in medical applications, is associated with the analysis of survival times. The quantity of interest in such models is, for example, the time between diagnosis and death of a certain disease. Of course, similar data structures frequently appear in other fields like the analysis of times to failure of machines in engineering or the analysis of the time between wedding and divorce in social sciences. In general, the influence of covariates on the duration time up to a certain event is of interest. Therefore, the terms duration time and duration time analysis might be preferred to survival times and survival analysis but we will retain the latter in agreement with the convention in statistics.

In principle, survival times could be analyzed with models introduced in Part II, especially those accounting for the nonnegativity of survival times. However, special models have been developed to incorporate typical features of survival data. The most important challenge is that survival times are often not completely observed. Different types of incomplete data have been considered in the literature and led to the development of appropriate estimates for quantities characterizing the distribution of survival times. Famous examples are the Nelson-Aalen estimator for the cumulative hazard rate or the Kaplan-Meier estimator for the survivor function (see for example Klein & Moeschberger 2003). In the following, we will present regression models for survival times that allow for accommodation to different types of covariate effects and to incorporate several types of incomplete survival data. The rest of this section discusses these models in more detail. Section 14.2 describes the likelihood construction for different censoring mechanisms, i. e. different types of incomplete data. Some special cases of the present model and competing approaches are provided in Section 14.4 and Section 14.5. Section 15 describes how the empirical Bayes approach of the previous Parts II and III can be extended to estimate regression models for survival times.

14.1.1 The Cox model

Since the publication of the seminal paper of Cox (1972), influences of covariates u on survival times T are commonly described by a regression model for the hazard rate

$$\lambda(t|u) = \lim_{\Delta t \to 0} \frac{1}{\Delta t} P(t \leq T < t + \Delta t | T \geq t, u).$$

The hazard rate can be interpreted as the instantaneous rate of an event in the interval $[t, t + \Delta t]$ given survival up to time t. In the Cox proportional hazards model a multiplicative structure of the form

$$\lambda(t|u) = \lambda_0(t) \exp(u'\gamma) \tag{14.1}$$

is assumed for the hazard rate, where $\lambda_0(t)$ is an unspecified smooth baseline hazard rate and $u'\gamma$ is a linear predictor formed of (time-constant) covariates u and regression coefficients γ. In a conventional analysis, the baseline hazard rate $\lambda_0(t)$ remains unspecified

and estimation of the regression coefficients is based on the partial likelihood. In a second step, the baseline hazard can be approximated by a step function using Breslow's estimate (see for example Therneau & Grambsch (2000) or Klein & Moeschberger (2003) for more in-depth treatments of the Cox model and general discussions about the analysis of survival times).

The Cox model (14.1) (with time-constant covariates) is called proportional hazards model, since the ratio of the hazard rates for two individuals with covariate vectors u_1 and u_2 is given by

$$\frac{\lambda(t|u_1)}{\lambda(t|u_2)} = \exp((u_1 - u_2)'\gamma) \tag{14.2}$$

i. e. the hazard rates for the individuals are proportional because the ratio in (14.2) does not depend on t. This is a rather restrictive assumption which is likely to fail in realistically complex data examples. One possibility to relax the proportional hazards assumption is the inclusion of time-varying effects $\gamma(t)$ or time-varying covariates $u(t)$. However, both approaches lead to models that can no longer be estimated based on the partial likelihood and require more elaborate estimation techniques.

In addition to these issues, similar problems as discussed in the two previous parts may occur:

- Concerning the continuous covariates in the data set, the assumption of a strictly linear effect on the predictor may not be appropriate, i. e. some effects may be of unknown nonlinear form.
- Survival times may be spatially correlated.
- Heterogeneity among individuals or units may not be sufficiently described by covariates. Hence, unobserved unit or cluster-specific heterogeneity must be considered appropriately.
- Interactions between covariates may be of complex, nonlinear form.

14.1.2 Structured hazard regression

In order to account for the possibility of nonproportional hazards, nonstandard covariate effects and spatial correlations, we extend the classical Cox model to a semiparametric structured hazard rate model

$$\lambda_i(t) = \exp(\eta_i(t)), \quad i = 1, \ldots, n, \tag{14.3}$$

with structured additive predictor

$$\eta_i(t) = u_i(t)'\gamma + g_0(t) + \sum_{k=1}^{K} g_k(t) w_{ik}(t) + \sum_{j=1}^{J} f_j(\nu_{ij}(t)), \tag{14.4}$$

where $g_0(t) = \log(\lambda_0(t))$ is the log-baseline hazard, $g_k(t)$ represent time-varying effects of covariates $w_{ik}(t)$, $f_j(\nu_{ij}(t))$ are nonlinear effects of different types of generic covariates and $u_i(t)'\gamma$ corresponds to effects of covariates that are modeled in the usual parametric way. Note that all covariates in (14.4) are allowed to be time-varying. We will describe

at the end of Section 14.2 how time-varying covariates can be handled based on data augmentation. In analogy to Part II, the functions f_j comprise nonlinear effects of continuous covariates, spatial effects, i. i. d. random effects, interaction surfaces and varying coefficient terms. Time-varying effects $g_k(t)w_{ik}(t)$ can also be cast into the framework of varying coefficients if the survival time itself is considered to be the effect modifier. In fact, survival models with time-varying effects have already been treated in the original paper on varying coefficients by Hastie & Tibshirani (1993), where estimation was based on an adjusted partial likelihood.

To ease the description of inferential details in Section 15 and to obtain a compact formulation of structured hazard regression models, we introduce some matrix notation. All different effects in (14.4) can be cast into one general form and, similar as in Parts II and III, each vector of function evaluations can be written as the product of a design vector $v_{ij}(t)$ and a possibly high-dimensional vector of regression coefficients ξ_j. This also applies to the time-varying effects and, hence, after appropriate reindexing and suppressing the time index, the predictor (14.4) can be rewritten as

$$\eta_i = u_i'\gamma + v_{i1}'\xi_1 + \ldots + v_{ip}'\xi_p, \tag{14.5}$$

where $u_i'\gamma$ represents parametric effects while each of the terms $v_{ij}'\xi_j$ represents a nonparametric effect. Defining stacked vectors and matrices $\eta = (\eta_i)$, $U = (u_i)$, $V_j = (v_{ij})$, finally yields the expression

$$\eta = U\gamma + V_1\xi_1 + \ldots + V_p\xi_p.$$

14.2 Likelihood contributions for different censoring mechanisms

Usually, the Cox model and semiparametric extensions are developed for right censored observations which are one specific type of incomplete survival data. More formally spoken, if the true survival time is given by T and C is a censoring time, only $\tilde{T} = \min(T, C)$ is observed along with the censoring indicator $\delta = \mathbb{1}_{(T \leq C)}$. Many applications, however, confront the analyst with more complicated types of incomplete data involving more general censoring schemes. For example, interval censored survival times T are not observed exactly but are only known to fall into an interval $[T_{lo}, T_{up}]$. If $T_{lo} = 0$, these survival times are also referred to as being left censored. Such data structures frequently occur in medical applications, where the quantity of interest (e. g. the development of a certain disease) can only be diagnosed at discrete time points when the patient visits a doctor. In this case, the interval $[T_{lo}, T_{up}]$ is defined by the last time at which the patient was healthy and the time the disease is diagnosed. Interval censoring can also be used to account for rounded survival times which are often recorded in retrospective studies, see Section 17 for an application.

In addition to the censoring mechanisms discussed so far, each of the censoring schemes may appear in combination with left truncation of the corresponding observation. This means, that the survival time is only observed if it exceeds the truncation time T_{tr}. Accordingly, some survival times are not observable and the likelihood has to be adjusted appropriately. Figure 14.1 illustrates the different censoring schemes we will consider in the following: The true survival time is denoted by T which is observed for individuals

1 and 2. While individual 1 is not truncated, individual 2 is left truncated at time T_{tr}. Similarly, individuals 3 and 4 are right censored at time C and individuals 5 and 6 are interval censored with interval $[T_{lo}, T_{up}]$ and the same pattern of left truncation.

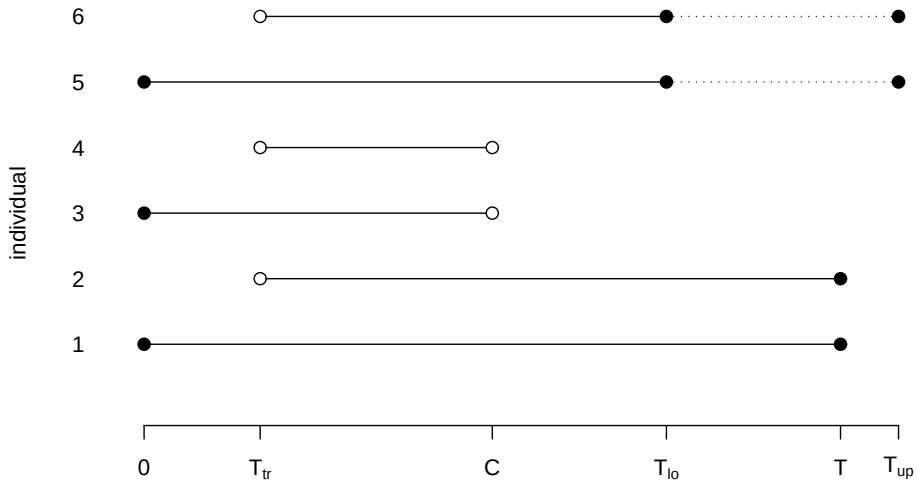


Figure 14.1: Illustration of different censoring schemes: For individuals 1 and 2, the true survival time T is observed, individuals 3 and 4 are right censored at time C, and individuals 5 and 6 are interval censored, where the interval is given by $[T_{lo}, T_{up}]$. Individuals 2, 4 and 6 are left truncated at time T_{tr}.

In a general framework combining all the aforementioned issues, an observed survival time can be completely described by the quadruple $(T_{tr}, T_{lo}, T_{up}, \delta)$, with

$$\begin{array}{ll} T_{lo} = T_{up} = T,\ \delta = 1 & \text{if the observation is uncensored,} \\ T_{lo} = T_{up} = C,\ \delta = 0 & \text{if the observation is right censored,} \\ T_{lo} < T_{up},\ \delta = 0 & \text{if the observation is interval censored.} \end{array}$$

For left truncated observations, we have $T_{tr} > 0$ and $T_{tr} = 0$ for observations which are not truncated.

Based on these definitions, we can construct the likelihood contributions for the different censoring schemes in terms of the hazard rate (14.3) and the survivor function

$$S(t) = P(T > t) = \exp\left(-\int_0^t \lambda(u)du\right).$$

Under the common assumption of noninformative censoring and conditional independence, the likelihood for $\xi = (\gamma', \xi_1', \ldots, \xi_p')'$ is given by the product of individual likelihood contributions

$$L(\xi) = \prod_{i=1}^{n} L_i(\xi), \tag{14.6}$$

where for an observation i with survival data $(T_{tr}, T_{lo}, T_{up}, \delta)$

$$\begin{aligned} L_i(\xi) &= \lambda(T_{up})\frac{S(T_{up})}{S(T_{tr})} \\ &= \lambda(T_{up})\exp\left(-\int_{T_{tr}}^{T_{up}} \lambda(t)dt\right) \end{aligned}$$

if the corresponding observation is uncensored,

$$\begin{aligned} L_i(\xi) &= \frac{S(T_{up})}{S(T_{tr})} \\ &= \exp\left(-\int_{T_{tr}}^{T_{up}} \lambda(t)dt\right) \end{aligned}$$

if the corresponding observation is right censored, and

$$\begin{aligned} L_i(\xi) &= \frac{S(T_{lo}) - S(T_{up})}{S(T_{tr})} \\ &= \exp\left(-\int_{T_{tr}}^{T_{lo}} \lambda(t)dt\right) - \exp\left(-\int_{T_{tr}}^{T_{up}} \lambda(t)dt\right) \\ &= \exp\left(-\int_{T_{tr}}^{T_{lo}} \lambda(t)dt\right)\left[1 - \exp\left(-\int_{T_{lo}}^{T_{up}} \lambda(t)dt\right)\right] \end{aligned} \tag{14.7}$$

if the corresponding observation is interval censored (see Klein & Moeschberger (2003, Ch. 3) for the derivation of these likelihood expressions). Note that some numerical integration technique has to be employed for an explicit evaluation of the likelihood contributions, since none of the integrals can, in general, be solved analytically. In our implementation we used the trapezoidal rule, see Section 15.2 for more details.

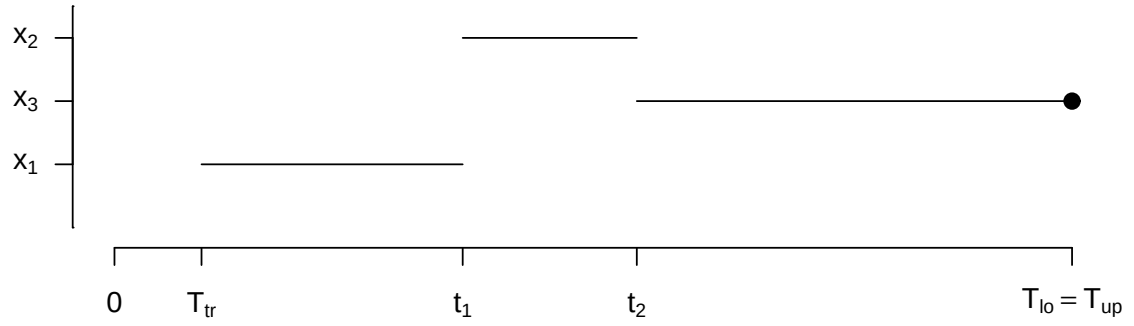


Figure 14.2: Illustration of time-varying covariates: Covariate $x(t)$ takes the three different values x_1, x_2 and x_3 on the subsequent intervals $[T_{tr}, t_1], [t_1, t_2]$ and $[t_2, T_{up}]$.

The quadruple notation presented above also allows for the easy inclusion of piecewise constant, time-varying covariates via data augmentation. For a decomposition $T_{tr} < t_1 < \ldots < t_q < T$ of the time axis we can equivalently decompose an integral of the hazard rate as

$$\int_{T_{tr}}^{T} \lambda(t)dt = \int_{T_{tr}}^{t_1} \lambda(t)dt + \int_{t_1}^{t_2} \lambda(t)dt + \ldots + \int_{t_{p-1}}^{t_p} \lambda(t)dt + \int_{t_p}^{T} \lambda(t)dt.$$

Therefore, an observation $(T_{tr}, T_{lo}, T_{up}, \delta)$ can be replaced by a set of observations $(T_{tr}$, t_1, t_1, $0)$, $(t_1, t_2, t_2, 0)$, $\ldots$, $(t_{p-1}, t_p, t_p, 0)$, $(t_p, T_{lo}, T_{up}, \delta)$ without changing the likelihood. Therefore, observations with time-varying covariates can be split into several observations, where the values $t_1 < \ldots < t_p$ are defined by the changepoints of the covariate and the covariate is now time-constant on each of the intervals. In theory, more general paths for a covariate $x(t)$ can also be included if $x(t)$ is known for $T_{tr} \leq t \leq T_{lo}$. In this case, the

likelihood (14.6) can also be evaluated numerically but a general path $x(t)$ is, of course, difficult to store in a data matrix.

Figure 14.2 illustrates the data augmentation step for a left truncated, uncensored observation and a covariate $x(t)$ that takes the three different values x_1, x_2 and x_3 on the subsequent intervals $[T_{tr}, t_1], [t_1, t_2]$ and $[t_2, T_{up}]$. In this example, the original observation $(T_{tr}, T_{up}, T_{up}, 1)$ has to be replaced by $(T_{tr}, t_1, t_1, 0)$, $(t_1, t_2, t_2, 0)$ and $(t_2, T_{up}, T_{up}, 1)$.

14.3 Priors

In principle, all priors discussed in Section 4.2 can also be used to model the different effects in the predictor (14.4) when analyzing survival times. Therefore, nonparametric effects of continuous covariates can be estimated on the basis of penalized splines, random walks or univariate Gaussian random fields priors (see Section 4.2.2), spatial effects can be modeled via Markov random fields or stationary Gaussian random fields (Section 4.2.3), cluster or individual-specific random effects can be included (Section 4.2.4), varying coefficient terms with continuous or spatial effect modifiers can be considered (Section 4.2.5), and interaction surfaces can be introduced based on two-dimensional penalized splines (Section 4.2.6). Note that in the context of regression models for survival times random effects are often referred to as frailties (see for example Therneau & Grambsch 2000, Ch. 9). This terminology originates from the fact that random effects were introduced in survival models to account for the frailty of patients with respect to a certain disease.

The only additional terms in predictor (14.4) not considered in Section 4.2 are the log-baseline $g_0(t)$ and the time-varying effects $g_k(t)$, $k = 1, \ldots, K$. From a conceptual point of view, the time scale t does not differ from a continuous covariate which renders all modeling possibilities presented in Section 4.2.2 applicable. Given that we are usually interested in estimating a smooth baseline hazard rate and smooth time-varying effects, we will only consider penalized splines. Moreover, as shown in Section 4.2.2.2 random walks are in fact special cases of penalized splines with degree $l = 0$. As exclusive advantage, random walk models allow for an exact computation of the integrals involved in the likelihood formulae, because all time-dependent functions are piecewise constant and, hence, the integrals reduce to sums (see also Section 14.4.1 on piecewise exponential models). In fact, the trapezoidal rule mimics this approximation. Therefore, the numerical benefit of random walk models is comparably small while P-spline models usually lead to smoother estimates.

14.4 Special cases

In order to clarify how structured hazard rate models correspond to other regression models for survival times discussed in the literature, we will take a closer look at two exemplary approaches which turn out to be special cases of the Cox model. Their advantage is that they can be estimated using standard software for generalized linear models or extended, semiparametric models for univariate responses from exponential families. However, some additional assumptions have to imposed on the data, so that the extended Cox model is generally preferable.

14.4.1 Piecewise exponential model

In the piecewise exponential model (see for example Fahrmeir & Tutz 2001, Ch. 9.1) a special form is assumed for the time-varying quantities, i. e. the baseline hazard rate $\lambda_0(t)$, the time-varying effects $g_k(t)$, $k = 1, \ldots, K$, and the time varying covariates $u(t)$, $w_k(t)$ and $\nu_j(t)$. Suppose that the time axis $[a_0, \infty)$ is divided into successive intervals

$$[a_0, a_1), [a_1, a_2), \ldots, [a_{q-1}, a_q), [a_q, \infty).$$

Then, assuming that all time-varying quantities are piecewise constant on the intervals, we can rewrite the predictor (14.4) as

$$\eta_i(t) = \eta_{is} = u_{is}'\gamma + \alpha_{0s} + \sum_{k=1}^{K} \alpha_{ks} w_{iks} + \sum_{j=1}^{J} f_j(\nu_{ijs}), \quad t \in [a_{s-1}, a_s), \quad s = 1, \ldots, q,$$

where for example $\alpha_{0s} = g_0(t)$, $t \in [a_{s-1}, a_s)$, corresponds to the value of the log-baseline in interval $[a_{s-1}, a_s)$ and similar definitions hold for the other time-varying quantities.

If we only consider right censored observations, the log-likelihood in a piecewise exponential model can be written in terms of an event indicator

$$y_{is} = \begin{cases} 1, & \text{observation } i \text{ has an event in } [a_{s-1}, a_s), \\ 0, & \text{observation } i \text{ survives or is censored in } [a_{s-1}, a_s), \end{cases}$$

as

$$\sum_{i=1}^{n} \delta_i \eta_i(t) - \sum_{i=1}^{n} \Lambda(t_i) = \sum_{s=1}^{q} \sum_{i \in R_s} (y_{is}\eta_{is} - \Delta_{is} \exp(\eta_{is})),$$

where R_s denotes the risk set for interval $[a_{s-1}, a_s)$, i. e. R_s contains the indices of the observations that had no event and were not censored prior to a_{s-1}. The Δ_{is} are offsets defined by

$$\Delta_{is} = \max\{0, \min\{a_s - a_{s-1}, t_i - a_{s-1}\}\}.$$

Taking a closer look on the likelihood of the piecewise exponential model reveals that, in fact, the likelihood is equivalent to the likelihood of a loglinear Poisson model for the indicator variables y_{is} with predictors η_{is} and offsets Δ_{is}. Therefore, the estimation techniques discussed for univariate responses in Part II are immediately applicable after some data augmentation. For example, if survival data

t_i	δ_i	x_{i1}	x_{i2}
0.25	1	0	3
0.12	0	1	5
$\vdots$	$\vdots$	$\vdots$	$\vdots$

are given, the observations have to be modified to

y_{is}	i	a_{is}	δ_i	Δ_{is}	x_{i1}	x_{i2}
0	1	0.1	1	0.1	0	3
0	1	0.2	1	0.1	0	3
1	1	0.3	1	0.05	0	3
0	2	0.1	0	0.1	1	5
0	2	0.2	0	0.02	1	5
⋮	⋮	⋮	⋮	⋮	⋮	⋮

To prevent too rough estimates for the baseline hazard and the time-varying effects, suitable prior assumptions can be imposed on the coefficients α_{0s} and α_{ks}. For example, choosing a random walk prior retains the assumption of a step function while circumventing overfitting. Of course, penalized splines of higher degree can also be used to fit smooth effects.

Although piecewise exponential models allow for easy fitting of extended survival models based on a Poisson regression, the structured hazard regression model has several advantages. In the piecewise exponential model, estimation of the baseline and time-varying effects highly depends on the grid the time axis is divided into. In case of too large intervals the approximation to the baseline will be poor which, in turn, may also affect the estimation of the remaining parameters. On the other hand, using small intervals results in expanded data augmentation and, thus, strongly increases the effective number of observations. Especially for large data sets, this may lead to long execution times and numerical problems due to the large matrices involved. Within the structured hazard regression model such problems do not arise.

From a conceptual perspective, the piecewise exponential model has the drawback that it does not allow the likelihood to be transformed into a Poisson likelihood for more general but the right censoring scheme. In contrast, interval and left censoring can be easily included in the extended Cox model, since the likelihood can be calculated using the same numerical integration techniques as for right censored observations.

14.4.2 Discrete time models

Discrete time survival models assume a similar structure as the piecewise exponential model, where the time axis is divided into q intervals $[a_{s-1}, a_s)$, $s = 1, \ldots, q$, but do no longer exploit the continuous time information. Instead, only the discrete information about the interval the event has happened in is considered. More formally, a discrete random variable T_d is introduced with

$$T_d = s \quad \Leftrightarrow \quad T \in [a_{s-1}, a_s),$$

i. e. the discrete time index s is identified with the time interval $[a_{s-1}, a_s)$. This may be a natural assumption if observations can only be collected at equidistant discrete time points due to the sampling mechanism, or if the event of interest actually only happens at discrete times. An example for the latter are durations of unemployment, because employments usually start at the beginning of a month and, hence, the durations are really given in months.

For discrete time models, a discrete hazard function can be defined by

$$\lambda_s = P(T_d = s | T_d \geq s), \quad s = 1, \ldots, q. \tag{14.8}$$

Here, the continuous hazard function transforms to a conditional probability which can be related to covariates using sequential models for ordinal responses as discussed in Section 10.1.4. Each survival time corresponds to a sequence of binary decisions indicating whether the corresponding individual survived the s-th interval or not.

Although sequential models exactly fit the needs of a discrete time survival model without censoring, it is easier to model the binary decisions directly if right censoring is to be included. Therefore, binary indicators y_{is}, $s = 1, \ldots, t_i$ are introduced with

$$y_{is} = \begin{cases} 1, & \text{if } t_i = s \text{ and } \delta_i = 1, \\ 0, & \text{else.} \end{cases}$$

Hence, for a right censored observation all binary indicators equal zero while for an uncensored observation the last indicator equals one and all other indicators are zero. Corresponding to the data augmentation of the response variable, all covariates have to be replicated as illustrated in the following example. Suppose that discrete time survival data

t_i	δ_i	x_{i1}	x_{i2}
4	0	0	2
3	1	1	0
⋮	⋮	⋮	⋮

are given. The first observation is censored in the fourth interval. The second observation is uncensored and has an event in the third interval. Augmenting the data as described above yields

y_{is}	i	s	δ_i	x_{i1}	x_{i2}
0	1	1	0	0	2
0	1	2	0	0	2
0	1	3	0	0	2
0	1	4	0	0	2
0	2	1	1	1	0
0	2	2	1	1	0
1	2	3	1	1	0
⋮	⋮	⋮	⋮	⋮	⋮

Note that the data augmentation allows for the routine inclusion of time-varying covariates, as x_{i1} and x_{i2} do not necessarily have to be constant over time and can adopt different values for each interval.

To build a regression model for the augmented data, we use the fact that the models for binary responses discussed in Section 4.1.1 allow to equivalently write the discrete hazard function (14.8) as

$$\lambda_s = P(y_{is} = 1) = h(\eta_{is}),$$

where h is the response function corresponding to the cumulative distribution function in the sequential model. Thus, the two model formulations are exactly equivalent. The predictor η_{is} can be specified as a structured additive predictor, for example by

$$\eta_{is} = u'_{is}\gamma + \alpha_{0s} + \sum_{k=1}^{K} \alpha_{ks} w_{iks} + \sum_{j=1}^{J} f_j(\nu_{ijs}), \quad s = 1, \ldots, q. \tag{14.9}$$

Both random walks and penalized splines can be used as priors to derive smooth estimates for the parameters representing the baseline hazard α_{0s} and the time-varying effects α_{ks}. Note that this is not possible in the sequential model, giving a further argument in favor of binary models.

One particularly interesting example showing the close connection between extended Cox models and discrete time models is the complementary log-log model. Consider for the moment a continuous time model with hazard rate

$$\lambda(s) = \lambda_0(s) \exp(\tilde{\eta}).$$

Introducing parameters α_{0s} which represent the logarithm of the integral over the baseline hazard

$$\alpha_{0s} = \log \left(\int_{a_{s-1}}^{a_s} \lambda_0(u) du \right),$$

allows us to express the continuous hazard rate in terms of the discrete hazard rate of a complementary log-log model (see Kalbfleisch & Prentice 1980, Ch. 2.4.2):

$$\lambda_s = 1 - \exp(-\exp(\alpha_{0s} + \tilde{\eta})).$$

Note that the predictor $\tilde{\eta}$ remains unchanged when switching between the continuous time and the discrete time view. Therefore, estimation of the covariate effects can either be based on the extended Cox model or the binary complementary log-log model. Correspondingly, the complementary log-log model can be regarded a discretized or grouped version of a time-continuous model. Fleming & Harrington (1990)[pp. 126] show a similar relationship between logistic regression models for discrete survival times and the Cox model.

Comparing discrete time models with structured hazard regression models reveals similar drawbacks as for the piecewise exponential model. Especially the data augmentation substantially increases the amount of data to be handled if the number of discrete time points is large. However, a discrete time model may be more appropriate if a lot of tied observations are observed and the assumption of continuous time is not satisfied, or if the data generating process directly leads to discrete times (as in the example on unemployment durations discussed above).

As another disadvantage, discrete time models do not enable the inclusion of left or interval censored observations. Only if interval censoring is introduced by a specific mechanism, discrete time models can be a competing analyzing strategy. If, for example, the time until the appearance of a certain disease is investigated and the presence of this disease is controlled at equidistant time points, the resulting duration times may either be analyzed based on an interval censoring approach for continuous times or on a discrete time model. We will contrast both possibilities in a simulation study in Section 19.

14.5 Related approaches

In order to demonstrate the generality and flexibility of structured hazard regression models, we will now discuss several related models introduced in the literature. First, we will concentrate on approaches allowing only for right censored survival times, since most efforts on nonparametric and spatial extensions of the Cox model are spent in this area.

Fully Bayesian models for jointly estimating the baseline hazard rate and possibly time-varying covariate effects, are described in Ibrahim, Chen & Sinha (2001). A comparable class of models based on penalized splines is presented in Kauermann (2005), where a data augmentation scheme similar as in piecewise exponential models allows to transform the likelihood to a Poisson likelihood and mixed model methodology for generalized linear mixed models can be employed.

Survival models which add a spatial component to the linear predictor in (14.1) have been developed recently. Li & Ryan (2002) model the spatial component by a stationary Gaussian random field (GRF). The baseline hazard rate, however, is treated as a nuisance parameter, and no procedure for estimating the spatial effects is provided. Henderson et al. (2002) propose a Cox model with gamma frailties, where the frailty means follow either a Markov random field (MRF) or a stationary GRF Kriging model. They use a kind of hybrid MCMC scheme, plugging in the Breslow estimator for the baseline hazard at each updating step. Banerjee, Wall & Carlin (2003) assume a parametric Weibull baseline hazard rate and MRF or GRF priors for the spatial component. Carlin & Banerjee (2002) and Banerjee & Carlin (2003) extend this work by including nonparametric estimation of the baseline hazard rate. Effects of continuous covariates are still assumed to be of linear parametric form as in (14.1). A semiparametric fully Bayesian approach to structured hazard regression models of the form (14.4) has been developed by Hennerfeind, Brezger & Fahrmeir (2005).

The literature dealing with extended censoring schemes, especially interval censoring, is much more limited. Cai & Betensky (2003) extend the mixed model based approach by Cai, Hyndman & Wand (2002) to the estimation of the baseline hazard rate in the presence of interval censoring based on penalized splines. Their model also allows for the inclusion of parametric covariate effects. A class of hazard regression models for interval censored survival times extending the work by Kooperberg, Stone & Truong (1995) is described in Kooperberg & Clarkson (1997). The baseline hazard rate, covariate effects and time-varying effects are approximated by linear splines. Tensor product splines are proposed to model interaction surfaces. Smoothness of the estimated curves and surfaces is not ensured via penalization but through a variable selection procedure based on information criteria. A Bayesian approach to correlated interval censored survival times is presented in Komárek, Lesaffre, Härkänen, Declerck & Virtanen (2005). While interval censoring is modeled via data augmentation, frailties are used to incorporate correlations.

A completely different class of models for the analysis of survival times are transformation or accelerated failure time (AFT) models, where the survival times are modeled in a linear regression fashion. To account for the nonnegativity, a transformation is applied. The most common models are based on a logarithmic transformation yielding

$$\ln(T) = u'\gamma + \sigma\varepsilon. \tag{14.10}$$

The transformed survival time is directly related to the linear predictor composed of covariates u and regression coefficients γ plus an error term ε multiplied by the scale parameter σ. Different models are characterized by the distribution of the error term. For example, extreme value distributed errors correspond to Weibull distributed survival times whereas Gaussian distributed errors correspond to lognormal survival times. In either case, the regression coefficients have a nice interpretation. Exponentiating (14.10) indicates that the covariates act multiplicatively on the survival time. For instance, for a simple model with only one binary variable u we have

$$T = \exp(\gamma_0)\exp(\gamma_1 u)\exp(\sigma\varepsilon).$$

Comparing observations with $u = 0$ and $u = 1$ respectively reveals that the latter are moving towards the event with accelerated (decelerated) speed if $\gamma_1 > 0$ ($\gamma_1 < 0$).

In case of right censored survival times, the likelihood of an accelerated failure time model is easily calculated (see for example Fahrmeir, Hamerle & Tutz 1996, Ch. 7), and the parameters can be estimated with a Newton-Raphson algorithm. Extended transformation models for interval censored survival times in combination with a generalized estimating equations approach to account for correlations are described in Bogaerts et al. (2002).

The definition of extended transformation models with structured additive predictors is, in principle, straightforward. However, we concentrate on extended Cox models since AFT models do not allow for the inclusion of time-varying covariates or time-varying effects. Furthermore, a parametric distribution has to be assumed for the survival times and, hence, accelerated failure times are less flexible than the Cox model.

A possibility to circumvent at least the latter problem is presented in Komárek, Lesaffre & Hilton (2005). They employ a mixture of normals for the error distribution, where the mixture weights are forced to vary smoothly over the mixture components resulting in a smooth, flexible error distribution. Since the mixing distributions are Gaussian, likelihood contributions for right and interval censored observations are easily calculated in terms of Gaussian cumulative distribution functions. A combination of structured additive predictors and flexible error distributions seems to be a promising alternative to extended Cox models, since AFT models naturally lead to nonproportional hazards without including time-varying effects or covariates. Furthermore, mixed model methodology should allow the estimation of the smoothing parameter associated with the penalization of the mixing weights. In contrast, in Komárek et al. (2005) the smoothing parameter is chosen according to AIC using a grid search algorithm which would be computationally intractable in more complex models.

15 Inference

Similarly as in the two previous parts, mixed model based inference in structured hazard regression models can be performed in three steps: Section 15.1 describes the reparametrization of structured hazard regression models in terms of a simple mixed model with proper priors. The following two sections describe the estimation of regression coefficients for given variances and marginal likelihood estimation of the variances. Especially the estimation of the variance parameters requires additional attention since the estimation of mixed models for survival data is less developed than for responses from exponential families. The resulting estimates, again, have an interpretation as posterior mode or empirical Bayes estimates.

The basic quantity for Bayesian inference is the posterior which – in the original model parameterization – is obtained by combining prior information and the likelihood contributions given in Section 14.2 to

$$p(\xi|data) \propto L(\xi) \prod_{j=1}^{p} p(\xi_j|\tau_j^2). \tag{15.1}$$

Equation (15.1) has to be maximized to obtain posterior mode estimates. Equivalently, we may consider the log-posterior

$$l_p(\xi|data) = l(\xi) + \sum_{j=1}^{p} \frac{1}{\tau_j^2} \xi_j' K_j \xi_j,$$

which equals a penalized likelihood with penalty terms $\frac{1}{\tau_j^2} \xi_j' K_j \xi_j$.

15.1 Mixed model representation

In structured hazard regression models, the reparametrization of the ξ_j, which is required to establish a variance components mixed model, is of exactly the same form as for univariate responses discussed in Section 5.1. Each vector of regression coefficients ξ_j is decomposed into an unpenalized and a penalized part yielding

$$\xi_j = \tilde{X}_j \beta_j + \tilde{Z}_j b_j,$$

where the design matrices $\tilde{X}_j$ and $\tilde{Z}_j$ are defined in complete analogy to the discussion in Section 5.1. This in turn allows to rewrite the predictor (14.4) as

$$\eta_i = \sum_{j=1}^{p} v_{ij}' \xi_j + u_i' \gamma = x_i' \beta + z_i' b, \tag{15.2}$$

where $p(\beta) \propto const$, $b \sim N(0, Q)$ and $Q = \text{blockdiag}(\tau_1^2 I, \ldots, \tau_p^2 I)$. Accordingly, the posterior transforms to

$$p(\beta, b|data) \propto L(\beta, b) \exp\left(-\frac{1}{2} b' Q^{-1} b\right) \tag{15.3}$$

and the log-posterior is given by

$$l_p(\beta, b|data) = l(\beta, b) - \sum_{j=1}^{p} \frac{1}{2\tau_j^2} b_j' b_j. \tag{15.4}$$

Note that the (log-)likelihoods in (15.3) and (15.4) are completely equivalent to those discussed in Section 14.2 when the additive predictor is replaced by the expression in (15.2).

15.2 Regression coefficients

To construct a Newton-Raphson update step for the regression coefficients, we need first and second derivatives of (15.4) with respect to β and b. To ease notation, consider for the moment a hazard rate of the form

$$\lambda(t) = \exp(x(t)'\xi)$$

which essentially reflects the structure of a structured hazard regression model. Defining

$$D_j(t) = -\frac{\partial}{\partial \xi_j} \int_0^t \lambda(u)du = -\int_0^t x_j(u)\lambda(u)du$$

and

$$E_{jk}(t) = -\frac{\partial^2}{\partial \xi_j \partial \xi_k} \int_0^t \lambda(u)du = -\int_0^t x_j(u)x_k(u)\lambda(u)du,$$

first and second derivatives of the log-likelihood contributions for uncensored and right censored observations can easily be shown to be given by

$$\frac{\partial l_i(\xi)}{\partial \xi_j} = \delta \cdot x_j(T_{up}) + D_j(T_{up}) - D_j(T_{tr})$$

and

$$\frac{\partial^2 l_i(\xi)}{\partial \xi_j \partial \xi_k} = E_{jk}(T_{up}) - E_{jk}(T_{tr}).$$

In case of interval censored survival times formulae become more complicated. Considering the first derivative, we obtain from (14.7)

$$\begin{aligned}
\frac{\partial l_i(\xi)}{\partial \xi_j} &= \frac{\partial}{\partial \xi_j}\left[-\int_{T_{tr}}^{T_{lo}} \lambda(t)dt + \log\left(1 - \exp\left(-\int_{T_{lo}}^{T_{up}} \lambda(t)dt\right)\right)\right] \\
&= D_j(T_{lo}) - D_j(T_{tr}) + \frac{\dfrac{\partial}{\partial \xi_j}\left(1 - \exp\left(-\displaystyle\int_{T_{lo}}^{T_{up}} \lambda(t)dt\right)\right)}{1 - \exp\left[\Lambda(T_{lo}) - \Lambda(T_{up})\right]} \\
&= D_j(T_{lo}) - D_j(T_{tr}) - \frac{\exp\left[\Lambda(T_{lo}) - \Lambda(T_{up})\right]\left[D_j(T_{lo}) - D_j(T_{up})\right]}{1 - \exp\left[\Lambda(T_{lo}) - \Lambda(T_{up})\right]}
\end{aligned}$$

Proceeding in the same way with the second derivatives yields

$$\begin{aligned}
\frac{\partial^2 l_i(\xi)}{\partial \xi_j \partial \xi_k} &= E_{jk}(T_{lo}) - E_{jk}(T_{tr}) - \frac{\exp\left[\Lambda(T_{lo}) - \Lambda(T_{up})\right]^2 \left[D_j(T_{up}) - D_j(T_{lo})\right]\left[D_k(T_{up}) - D_k(T_{lo})\right]}{\left\{1 - \exp\left[\Lambda(T_{lo}) - \Lambda(T_{up})\right]\right\}^2} \\
&\quad - \frac{\exp\left[\Lambda(T_{lo}) - \Lambda(T_{up})\right]\left\{\left[D_k(T_{up}) - D_k(T_{lo})\right]\left[D_j(T_{up}) - D_j(T_{lo})\right] - \left[E_{jk}(T_{up}) - E_{jk}(T_{lo})\right]\right\}}{1 - \exp\left[\Lambda(T_{lo}) - \Lambda(T_{up})\right]}.
\end{aligned}$$

Note that for $T_{tr} = 0$ these results are equivalent to those presented in Kooperberg & Clarkson (1997).

In order to evaluate the derivatives $D_j(t)$ and $E_{jk}(t)$, and the cumulative hazard rate $\Lambda(t)$, some numerical integration technique is needed. Due to its simplicity, we used the trapezoidal rule in the implementation, where an integral of the form

$$\int_{T_1}^{T_2} g(t)dt$$

is approximated by the sum

$$\sum_{k=1}^{K}(t_k - t_{k-1})\frac{g(t_k) + g(t_{k-1})}{2}$$

based on knots $T_1 = t_0 < \ldots < t_K = T_2$. This corresponds to approximating g with a piecewise linear function as illustrated in Figure 15.1. In order to achieve a sufficient approximation, the number of knots K should not be too small. In particular, it has to be larger than the number of parameters characterizing g. In our implementation we use $K = 300$ as default value. Concerning the position of the knots, two different choices are available: Equidistant knots and knots based on quantiles of the observed survival times. In our experience, the former works more stable, especially if larger holes without data are observed on the time axis. However, improved estimates for the covariate effects may be obtained with a quantile based grid (see also the simulation study in Section 18).

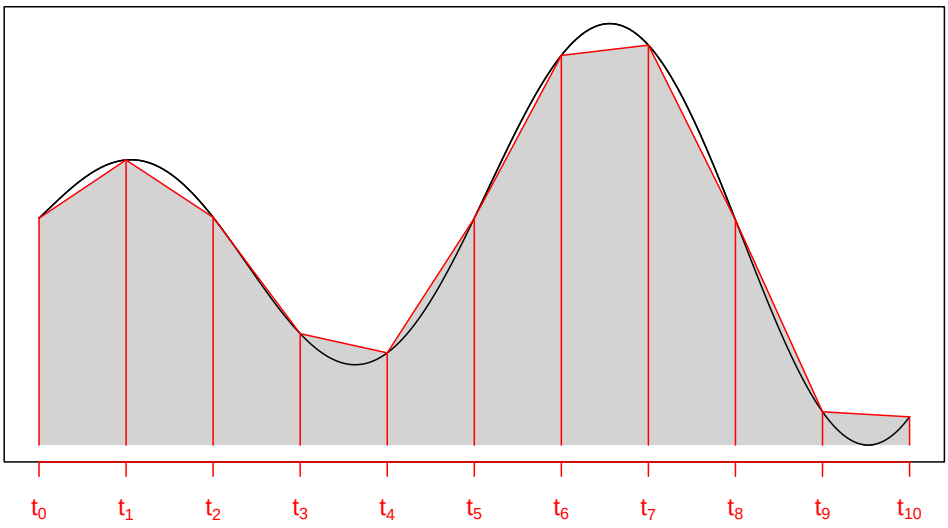


Figure 15.1: Trapezoidal rule: $g(t)$ is approximated by a piecewise linear function through the points $(t_k, g(t_k))$. The integral $\int_{T_1}^{T_2} g(t)dt$ can then be replaced by a sum over the corresponding trapezoids resulting in the shaded region.

If some covariates are constant over time, i. e. $x_{ij}(t) \equiv x_{ij}$, the evaluation of $D_j(t)$ can be significantly simplified. In this case, $D_j(t)$ reduces to $-x_{ij}\Lambda(t)$, and $\Lambda(t)$ has to be computed only once. Similar simplifications hold for the computation of $E_{jk}(t)$ if both $x_{ij}(t)$ and $x_{ik}(t)$ do not depend on t.

In a mixed model, the score-vector can be partitioned according to the derivatives with respect to the unpenalized and the penalized vector of regression coefficients, i. e.

$$s = \begin{pmatrix} s_\beta \\ s_b \end{pmatrix} = \begin{pmatrix} \dfrac{\partial l_p(\beta, b)}{\partial \beta} \\ \dfrac{\partial l_p(\beta, b)}{\partial b} \end{pmatrix}.$$

In analogy, the observed Fisher-information is partitioned into four blocks:

$$F = \begin{pmatrix} F_{\beta\beta} & F_{\beta b} \\ F_{b\beta} & F_{bb} \end{pmatrix} = - \begin{pmatrix} \dfrac{\partial^2 l_p(\beta, b)}{\partial \beta \partial \beta'} & \dfrac{\partial^2 l_p(\beta, b)}{\partial \beta \partial b'} \\ \dfrac{\partial^2 l_p(\beta, b)}{\partial b \partial \beta'} & \dfrac{\partial^2 l_p(\beta, b)}{\partial b \partial b'} \end{pmatrix}.$$

On the basis of the expressions presented above, both quantities are easy to calculate and allow to update estimates for the regression coefficients given the variances via a Newton-Raphson step:

$$\begin{pmatrix} \hat{\beta}^{(k+1)} \\ \hat{b}^{(k+1)} \end{pmatrix} = \begin{pmatrix} \hat{\beta}^{(k)} \\ \hat{b}^{(k)} \end{pmatrix} + (F^{(k)})^{-1} s^{(k)}.$$

15.3 Marginal likelihood for variance components

In Gaussian linear mixed models, a well established method for the estimation of variance components is restricted maximum likelihood (REML), which - in contrast to ordinary maximum likelihood - takes into account the loss of degrees of freedom due to the estimation of the regression coefficients. As Harville (1974) showed, REML estimation is equivalent to maximizing the marginal likelihood for the variance components

$$L^{marg}(Q) = \int L_{pen}(\beta, b, Q) d\beta db. \tag{15.5}$$

Consequently, REML estimation can be extended to more general situations including regression models for survival times. Up to now, marginal likelihood estimation has mostly been applied in the context of subject-specific frailty models based on the partial likelihood (compare e. g. Therneau & Grambsch (2000) or Ripatti & Palmgren (2000)). Cai et al. (2002) use marginal likelihood estimates for the smoothing parameter of the baseline hazard but do not provide estimation equations. Instead, they maximize the marginal likelihood numerically which may become quite computerintensive if the model includes more than one variance component as for a structured additive predictor.

In the following, we describe a possibility to estimate variances in a structured hazard regression model based on the full marginal likelihood (not the partial marginal likelihood). Two approximation steps allow to use a Fisher-Scoring algorithm for the maximization of (15.5), yielding estimation equations which are numerically simple to evaluate. First, applying a Laplace approximation to the marginal log-likelihood results in

$$l^{marg}(Q) \approx l(\hat{\beta}, \hat{b}) - \frac{1}{2} \log |Q| - \frac{1}{2} \hat{b}' Q^{-1} \hat{b} - \frac{1}{2} \log |F|.$$

Assuming that both $l(\hat{\beta}, \hat{b})$ and $\hat{b}$ vary only slowly when changing the variance components, we can further reduce the marginal log-likelihood to

$$l^{marg}(Q) \approx -\frac{1}{2}\log|Q| - \frac{1}{2}\log|F| - \frac{1}{2}b'Q^{-1}b, \tag{15.6}$$

where b denotes a fixed value not depending directly on the variances, e. g. a current estimate. The second approximation seems to be reasonable, since, at least in generalized additive models, it is well known that small changes of the smoothing parameters do not affect the estimates of the regression coefficients very much. A similar argument is given by Breslow & Clayton (1993) to simplify the marginal likelihood for variance components of generalized linear mixed models. Although the approximation steps may look rather crude at first sight, they proved to work well in simulations as well as in real data applications.

First and second derivatives of (15.6) can easily be calculated, based on differentiation rules for matrices (see for example McCulloch & Searle (2001, appendix M) or Harville (1997, Ch. 15)). Since expressions for general covariance matrices Q become quite lengthy, we make use of the blockdiagonal structure of Q in variance components models to obtain simpler formulae. For the score function this yields

$$\begin{aligned} s_j^* &= \frac{\partial l^{marg}(Q)}{\tau_j^2} \\ &= -\frac{1}{2}\operatorname{tr}\left(Q^{-1}\frac{\partial Q}{\partial \tau_j^2}\right) - \frac{1}{2}\operatorname{tr}\left(F^{-1}\frac{\partial F}{\partial \tau_j^2}\right) + \frac{1}{2}b'Q^{-1}\frac{\partial Q}{\partial \tau_j^2}Q^{-1}b \\ &= -\frac{k_j}{2\tau_j^2} + \frac{1}{2\tau_j^4}\operatorname{tr}\left(G_{b_jb_j}\right) + \frac{1}{2\tau_j^4}{b_j}'b_j, \end{aligned}$$

where $k_j = \operatorname{rank}(K_j)$, $G = F^{-1}$ denotes the inverse Fisher information (for the regression coefficients) and $G_{b_jb_j}$ is the diagonal block of F^{-1} corresponding to b_j. The expected Fisher-information can be shown to be given by

$$\begin{aligned} F_{jk}^* &= E\left(-\frac{\partial^2 l^{marg}(Q)}{\partial \tau_j^2 \partial \tau_k^2}\right) \\ &= \frac{1}{2\tau_j^4\tau_k^4}\operatorname{tr}\left(G_{b_jb_k}G_{b_kb_j}\right), \end{aligned}$$

where $G_{b_jb_k}$ denotes the off-diagonal block of F^{-1} corresponding to b_j and b_k. Both expressions are numerically simple to evaluate since F^{-1} and b are direct byproducts from the estimation of the regression coefficients. Based on the score-function and the Fisher-information we can compute updated variances $\tau^2 = (\tau_1^2, \ldots, \tau_p^2)'$ via a Fisher-scoring step:

$$(\tau^2)^{(k+1)} = (\tau^2)^{(k)} + (F^*)^{-1}s^*.$$

16 Leukemia survival data

As a first example for the applicability of structured hazard regression, we consider the data set on leukemia survival times described in Section 2.3. This data set has already been analyzed in Henderson et al. (2002), where the main focus was on the detection of spatial variation in the data while the assumption of a linear predictor for covariate effects was retained. Modeling such covariates as penalized splines allows to check whether the assumption of linearity is appropriate or whether a more flexible modeling improves the fit. Furthermore, our analysis allows for the joint estimation of the baseline hazard and, therefore, offers deeper insight into the temporal development of the risk to die from leukemia.

Since spatial information on the residence of a patient is available both in form of exact locations in terms of longitude and latitude, as well as aggregated to district-level, we can employ discrete and continuous modeling of the spatial effect. Comparing results from district-level and individual-level analyses allows to judge the loss of information caused by the aggregation of observations within districts.

In either situation, the hazard rate for observation i is given by

$$\lambda_i(t) = \exp(\eta_i(t))$$

with a structured additive predictor

$$\eta_i(t) = \gamma_0 + \gamma_1 sex_i + g_0(t) + f_1(age_i) + f_2(wbc_i) + f_3(tpi_i) + f_{spat}(s_i),$$

where g_0 is the (centered) log-baseline, f_1, f_2 and f_3 are smooth functions of the continuous covariates and f_{spat} is a spatial effect. Both g_0 and the f_j will be modeled as cubic P-splines with second order difference penalty and 20 inner knots. In an individual-level analysis, $s_i = (s_{xi}, s_{yi})$ is the exact location of the patient's residence, while in a district-level analysis s_i denotes the district the patient lives in.

16.1 District-level analysis

First, we conducted a district level analysis. In this case, a natural choice to model the spatial effect is Markov random field prior (4.25). In BayesX, such a district-level model is estimated using the following commands:

```
> dataset d
> d.infile using c:\data\leukemia.dat
> map m
> m.infile using c:\data\nwengland.bnd
> remlreg r
> r.regress delta = time(baseline) + age(psplinerw2) + wbc(psplinerw2) +
  tpi(psplinerw2) + district(spatial,map=m) + sex, family=cox using d
```

We start with the usual steps to create a dataset object containing the variables and a map object containing the spatial information. Afterwards, a structured hazard regression model is estimated by specifying `family=cox` as response distribution. The censoring

indicator `delta` is defined as the response variable and the survival times act as a covariate with option `baseline` indicating that the corresponding time defines the log-baseline. By default, BayesX uses equidistant knots for the numerical integration to evaluate derivatives of the log-likelihood, compare Section 15.2. A quantile based grid can be requested by adding the option `gridchoice=quantiles` to the baseline specification.

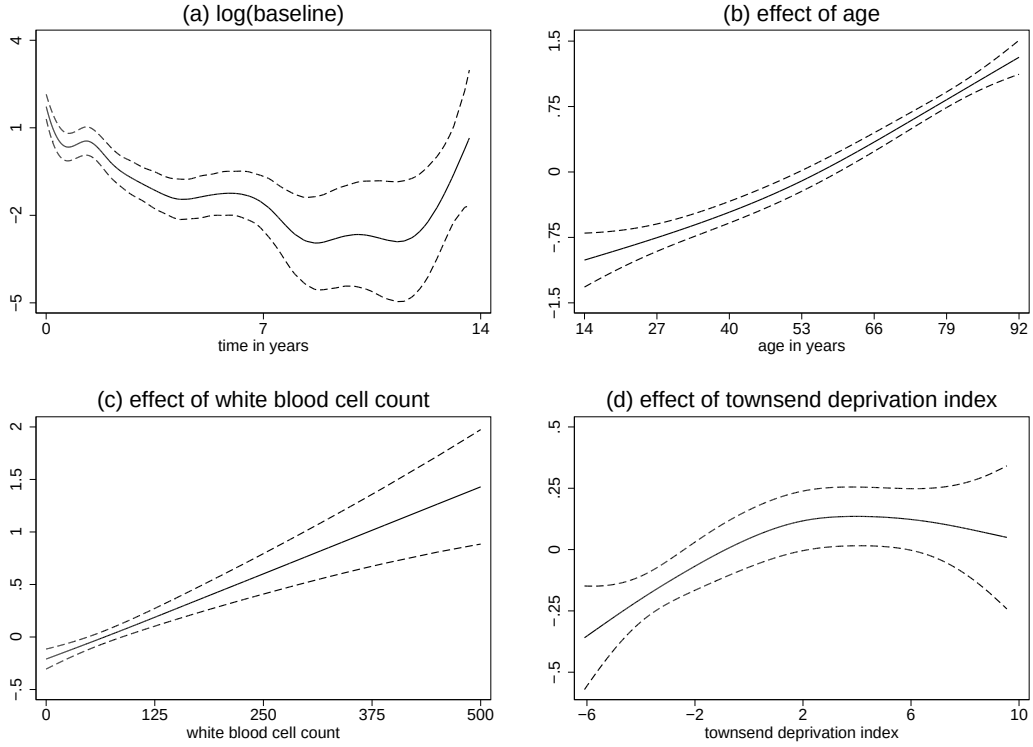


Figure 16.1: District-level analysis: Posterior mode estimates of the log-baseline, the effects of age, white blood cell count, and the Townsend deprivation index together with pointwise 95% *credible intervals.*

Employing a Markov random field for the spatial effect leads to the estimates for the log-baseline g_0 and the nonparametric effects f_j shown in Figure 16.1. The log-baseline decreases monotonically over nearly the whole observation period, alternating between relatively steep decreasing periods and almost flat periods. At the end of the observation period, there is a strong increase in g_0. However, only 26 individuals survived more than 10 years and, therefore, this increase should not be over-interpreted.

Obviously, the effects f_1 and f_2 of age and white blood cell count are almost linear and, hence, a reduced model probably leads to a comparable fit. We will check this possibility in the next section with individual-specific spatial effects. Both effects are quite similar to those found by Henderson et al. (2002), as is the effect of sex ($\hat{\gamma}_1 = 0.076$). Note that Henderson et al. modeled sex in effect-coding.

In contrast, the effect of the deprivation index is clearly nonlinear with lowest values for the developed enumeration districts. Moving to the right on the tpi-scale first increases the risk to die from leukemia but remains almost constant when approaching zero. Although both effects of age and wbc are nearly linear, the flexible modeling is a clear improvement

over a purely parametric approach, since it allows to check for the linearity of some effects but also allows more flexible functional forms, where needed.

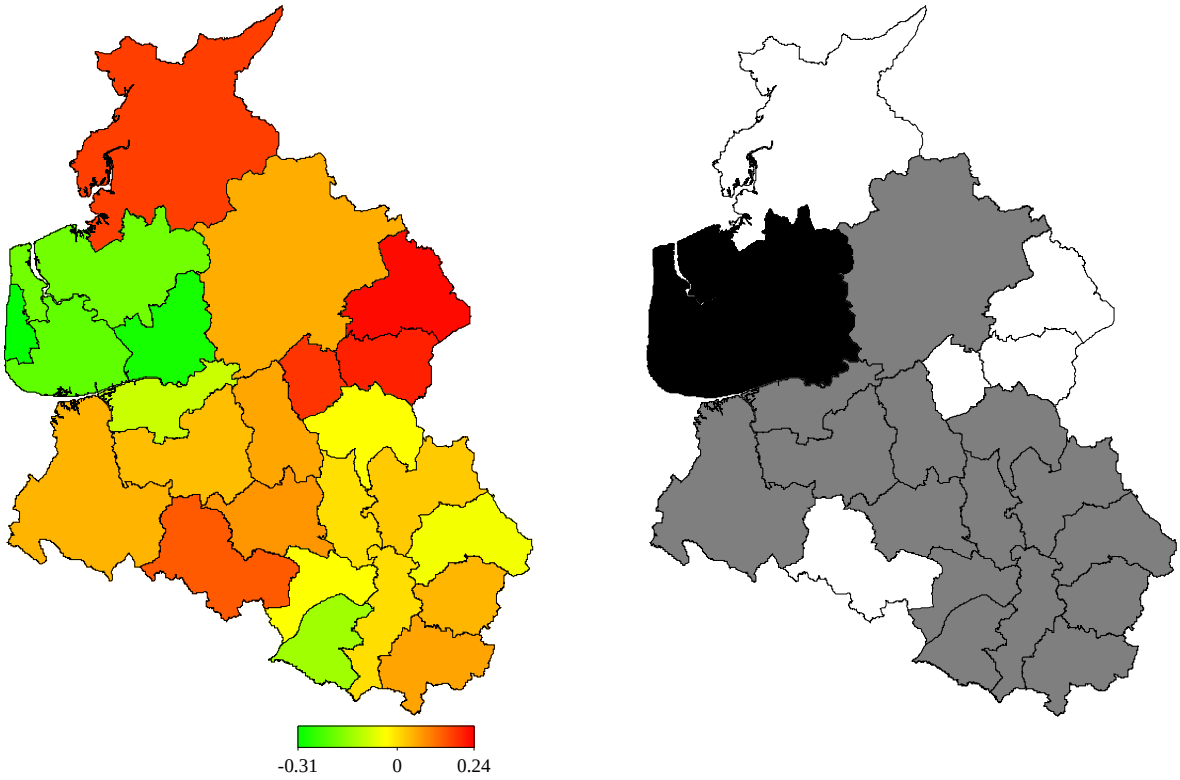


Figure 16.2: District-level analysis: Estimated spatial effect (left part) and pointwise 80% significance map (right part). Black denotes districts with strictly negative credible intervals, whereas white denotes districts with strictly positive credible intervals.

Looking at the estimated spatial effect in the left part of Figure 16.2, we find several districts with low risk in the western part of the map, surrounded by districts of increased risk. In the southern part of the map, there are also some districts with lower risk but the spatial effect is less pronounced here. This structure is confirmed by the significance map in the right part of Figure 16.2, where black denotes districts with strictly negative credible intervals and white denotes districts with strictly positive credible intervals.

16.2 Individual-level analysis

Of course, performing a district-level analysis is questionable when more detailed information is available. Therefore, we replaced the MRF with a stationary GRF based on the exact locations of the residences. The usage of a complete Kriging term would require the computation and inversion of an approximately 1,100 times 1,100 matrix, due to the total number of about 1,100 regression parameters in this model. To increase the efficiency in terms of computation time and memory requirements, the low-rank Kriging approach described in Section 4.2.3.4 is preferable. The choice of 50, 100, and 200 knots led to essentially the same results indicating that the approach is rather insensitive to the number of knots (see also rows 2–4 on the corresponding model fits in Table 16.1).

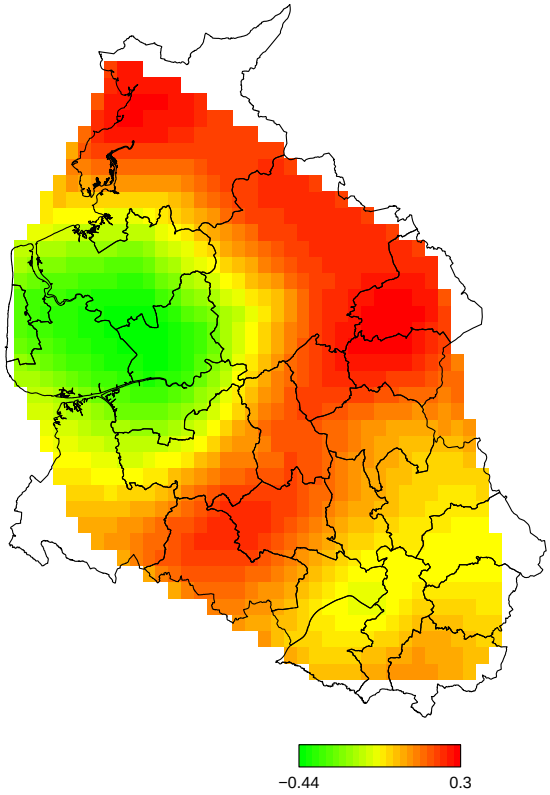


Figure 16.3: Individual-level analysis: Estimated spatial effect obtained with a (low-rank) Kriging term and 50 knots.

Effects for continuous covariates are not distinguishable by eye from the estimates in the district-level model and are therefore not presented again. Figure 16.3 shows the spatial effect for a low-rank Kriging term with 50 knots. In general, results are comparable to those from the district-level analysis but the Kriging approach reveals a more detailed spatial pattern and also finds a somewhat larger spatial variation. In particular, there is considerable variation of the spatial effect within most of the districts. When performing a district-level analysis, such information is lost, since a constant risk level is assumed in each district. This assumption may be problematic due to the fact that district boundaries are political constructs and usually do not reflect factors relevant for the risk of patients. In summary, the computationally feasible low-rank Kriging approach seems to be preferable to the Markov random field approach whenever individual-level information is available. This is also confirmed by comparing the models in terms of AIC, where the GRF approach leads to a reduced AIC regardless of the number of knots (see rows 1–4 in Table 16.1).

Information criteria can also be used to check, whether nonparametric modeling of covariate effects leads to an improved fit compared to purely parametric or semiparametric models. We compared the fully nonparametric model with three semiparametric models:

- A model where the effects of age, wbc and tpi are modeled linearly (row 5 in Table 16.1),
- a model where the effects of age, wbc and tpi are modeled linearly and a quadratic effect for tpi is included in addition (row 6 in Table 16.1), and

f_{spat}	knots	age	wbc	tpi	sex	-2*log-like.	df	AIC
MRF	-	nonp.	nonp.	nonp.	constant	11893.8	28.7	11951.2
GRF	50	nonp.	nonp.	nonp.	constant	11883.5	30.1	11943.7
GRF	100	nonp.	nonp.	nonp.	constant	11882.2	30.9	11944.0
GRF	200	nonp.	nonp.	nonp.	constant	11882.0	31.0	11944.1
GRF	50	linear	linear	linear	constant	11894.5	27.2	11948.9
GRF	50	linear	linear	quad.	constant	11890.2	27.9	11946.0
GRF	50	linear	linear	nonp.	constant	11890.4	28.1	11946.7
GRF	50	nonp.	nonp.	nonp.	time-var.	11884.1	31.1	11946.2

Table 16.1: Log-likelihood, degrees of freedom and AIC for different model specifications.

- a model where the effects of age and wbc are modeled linearly and tpi is included nonparametrically (row 7 in Table 16.1).

From the results presented in Table 16.1 we can conclude that the fully nonparametric model leads to the best fit in terms of AIC although the nonparametric effects of age and wbc were visually very close to straight lines.

16.3 Inclusion of time-varying effects

To check the proportional hazards assumption for males and females, we finally included a time-varying effect of sex in the fully nonparametric individual-level model. Although the estimated effect is somewhat increasing over time (see Figure 16.4), it almost equals a horizontal line and has rather wide credible intervals including such a horizontal line. Hence, we may conclude that the proportional hazard assumption is valid for the subpopulations of males and females. This observation is also supported by a comparison based on AIC, were a model with time-varying effect of sex yields a higher value than a model with time-constant effect (rows 2 and 8 in Table 16.1).

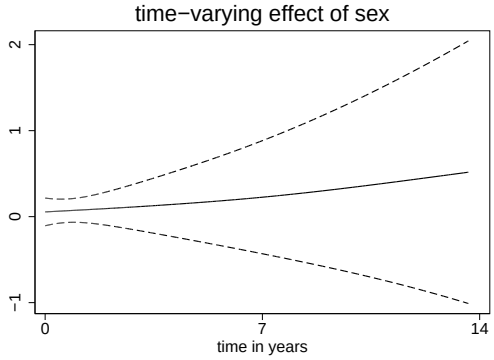


Figure 16.4: Individual-level analysis: Estimate for the time-varying effect of sex.

To include time-varying effects in structured hazard regression models, BayesX uses a similar syntax as for the definition of interactions. In the present example, a time-varying effect is obtained by replacing the covariate `sex` by the term `sex*time(baseline)`. This

syntax reflects the fact that time-varying effects can be considered as varying coefficient terms, where the survival time denotes the effect modifier.

17 Childhood mortality in Nigeria

As a second example on structured hazard regression, we examine the Nigerian childhood mortality data introduced in Section 2.4. In this application, interval censoring of the survival times has to be incorporated in addition to the nonparametric and spatial effects considered in the previous example. To be more specific, all uncensored survival times exceeding two months are treated as interval censored, where the interval is determined by the first and the last day of the corresponding month. For the survival times rounded to 12, 18, 24, 36 and 48 months (see Figure 2.8 on page 14) wider intervals have to be defined. We assigned symmetric intervals of 6 or 12 months length to these survival times.

For the hazard rate $\lambda(t) = \exp(\eta(t))$ we choose the geoadditive predictor

$$\eta(t) = g_0(t) + f_1(bmi) + f_2(age) + f_3(bord) + f_4(size) + f_{spat}(s) + u(t)'\gamma, \qquad (17.1)$$

where $g_0(t)$ denotes the log-baseline hazard rate, $f_1, \ldots, f_4$ are functions of the continuous covariates 'body mass index of the mother' (bmi), 'age of the mother at birth' (age), 'number of the child in the birth order' (bord) and 'number of household members' (size). Function f_{spat} models a spatial effect based on the district s the mother lives in. Fixed effects of numerous categorical covariates describing the economic situation of the family, circumstances at birth, and the breastfeeding behavior of the mother are collected in $u(t)$, see Table 2.3 in Section 2.4 for a detailed description. While most of these categorical covariates are time invariant, the duration of breastfeeding is described by a time-varying covariate which takes the value one as long as the child is breastfed and zero otherwise. Using the findings from Section 14.2, the breastfeeding information can be easily included in the present model by data augmentation.

Both the log-baseline and nonparametric effects are modeled by cubic P-splines with second order random walk penalty and 20 inner knots while the spatial effect is assumed to follow Markov random field prior (4.25). Due to missing values, the final number of observations accounts to $n = 5323$. 117 children die within the first two months and are, thus, treated as uncensored. The 474 children that die within the remaining study time are treated as interval censored as explained above.

Within BayesX, the presented model can be specified using the following commands:

```
> r.regress delta = intervalright(baseline) + district(spatial,map=m) +
  ... + initial2, family=cox leftint=intervalleft lefttrunc=truncleft
  using d
```

where we assume that a suitable dataset object `d` and a map object `m` have been created previously. In the present example three time variables have to be specified: The left and the right interval boundaries for the interval censored observations and the left truncation time. The latter is introduced by the data augmentation steps that are required to incorporate the time-varying covariate breastfeeding. The right interval boundary is specified within the model formula to define the log-baseline. The two remaining time variables are supplied as global options `leftint` and `lefttrunc`. The exact definitions of the corresponding time variables are given in Section 14.2.

Table 17.1 contains the estimates, standard deviations, p-values and 95% credible intervals of the fixed effects of model (17.1). Though most of the covariates show no significant

variable	$\hat{\gamma}_j$	std.	p-value	95% ci	
intercept	-8.266	0.824	<0.001	-9.882	-6.649
breastfeeding	-4.266	0.128	<0.001	-4.519	-4.014
initial1	-0.086	0.115	0.453	-0.312	0.139
initial2	-0.233	0.106	0.028	-0.441	-0.024
sex	-0.013	0.083	0.873	-0.176	0.149
employment	-0.074	0.090	0.406	-0.251	0.101
education	-0.269	0.116	0.020	-0.497	-0.041
place of delivery	-0.543	0.129	<0.001	-0.797	-0.289
assistance	-0.184	0.107	0.087	-0.395	0.027
longbirth	0.280	0.098	0.004	0.088	0.472
bleeding	0.032	0.109	0.768	-0.183	0.247
fever	0.207	0.127	0.102	-0.042	0.457
convulsion	0.111	0.210	0.595	-0.301	0.524
toilet	-0.151	0.108	0.162	-0.364	0.061
floormaterial	-0.072	0.117	0.535	-0.303	0.157
electricity	-0.119	0.139	0.388	-0.392	0.152
urban	-0.083	0.113	0.460	-0.307	0.139
religion1	-0.521	0.125	<0.001	-0.768	-0.274
religion3	-0.271	0.255	0.287	-0.772	0.228
weight2	0.036	0.130	0.777	-0.219	0.293
weight3	0.012	0.119	0.919	-0.221	0.246
weight4	-0.047	0.180	0.791	-0.401	0.306
weight5	0.370	0.166	0.025	0.045	0.696
wealth2	0.157	0.129	0.224	-0.096	0.410
wealth3	0.215	0.183	0.238	-0.143	0.574
wealth4	-0.084	0.245	0.731	-0.564	0.396
wealth5	-0.499	0.309	0.105	-1.106	0.106
watersource	-0.054	0.054	0.324	-0.161	0.053

Table 17.1: Childhood mortality in Nigeria: Estimates, standard deviations, p-values and 95% credible intervals of fixed effects.

effects, some interesting conclusions can be drawn. For example, the large negative value of the intercept reflects the relatively low overall risk of childhood mortality. The covariate breastfeeding causes the highest decrease in the mortality risk when modeled time-varying, whereas it is of less impact when included in a categorized version. The beneficial effect is expected, since breastfeeding supplies the child with important antibodies. Surprisingly, the estimated effect of the time when the child was first breastfed (represented by the categorical variables initial1 and initial2) is counterintuitive.

Further significant or borderline significant effects comprise education of the mother (reduced risk for higher educated mothers), the place of delivery (reduced risk for births at hospital), assistance (reduced risk if assistance was present at birth), long birth (increased risk), religion of the mother (reduced risk for christian mothers), and weight (increased risk for very small children).

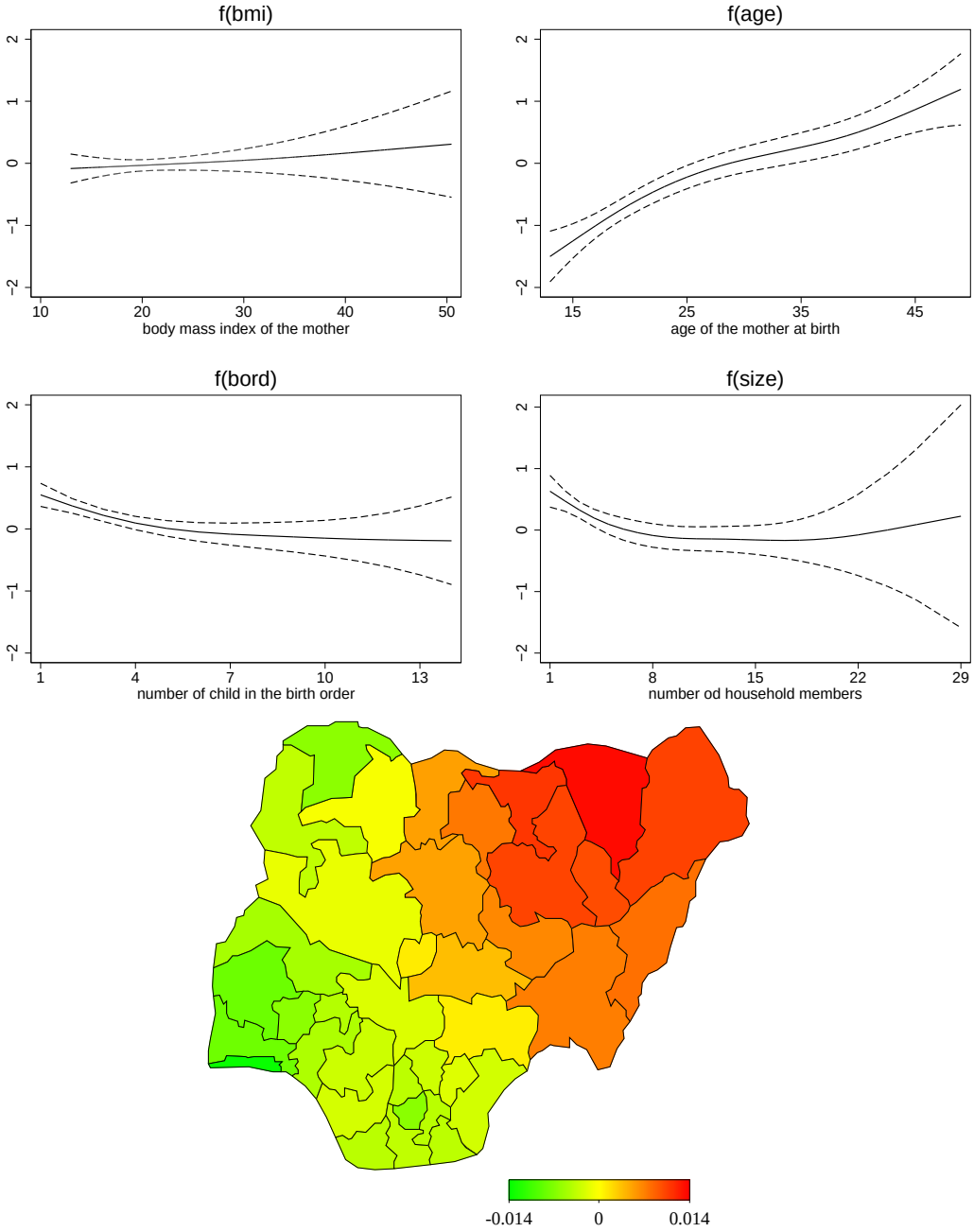


Figure 17.1: Childhood mortality in Nigeria: Estimates of nonparametric effects (with 95% credible intervals) and of the spatial effect.

Estimates of nonparametric and spatial effects are displayed in Figure 17.1. The effect of the maternal body mass index is almost constant with a slight increase for higher values. However, since the credible intervals include zero, the influence of the body mass index can be neglected. The remaining three nonparametric effects are of nonlinear though almost monotone functional form. While a higher age of the mother induced an increased risk, this does not hold for very young mothers although a u-shaped effect might have been expected here. Both a higher number of the child in the birth order and a higher number of household members lead to decreased risk. While the former effect may be explained by an increased knowledge about childcare of the mother, the latter

may reflect the fact that well-endowed households attract additional household members. The range of the estimated spatial effect is very small and a pointwise significance map shows no districts with effects different from zero. It should, however, be noted that in an analysis which exclusively contains a spatial effect, a highly significant spatial pattern emerges. Therefore, observations are clearly spatially correlated, but the spatial variation is completely explained by the covariates.

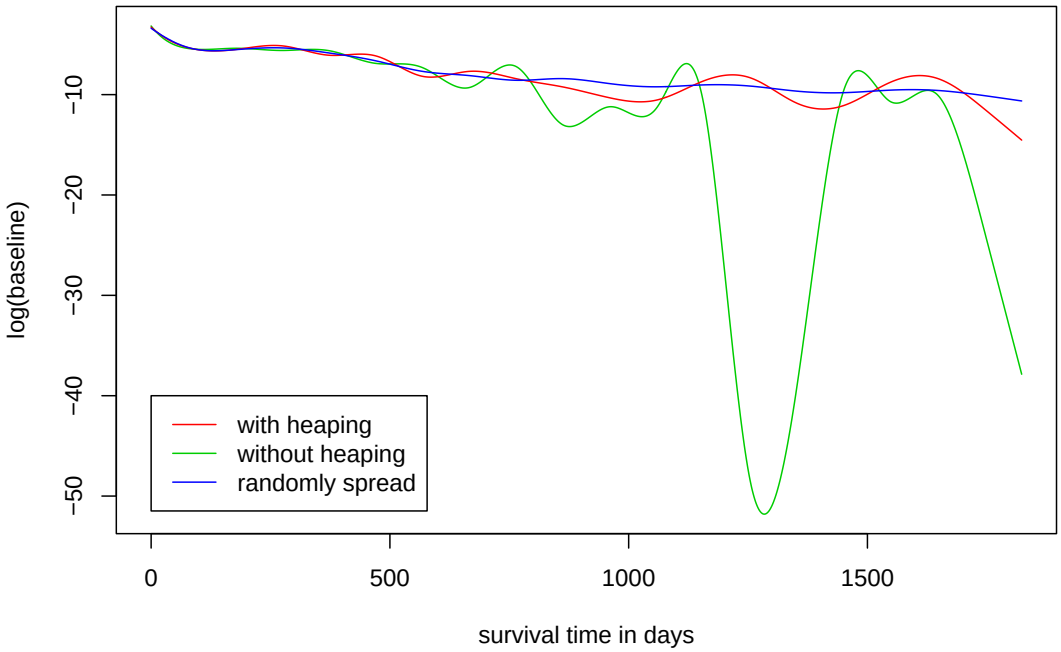


Figure 17.2: Childhood mortality in Nigeria: Estimated log-baselines based on interval censoring with heaping, interval censoring without heaping and randomly spread uncensored observations.

Figure 17.2 shows the estimated log-baseline hazard rate obtained with three different models:

- A model based on the interval censoring approach, where all observed death times beyond two months are treated as interval censored and heaping effects are incorporated, i. e. both roundings to months and (half) years are considered (red line).
- A model based on the the interval censoring approach, where all observed death times beyond two months are treated as interval censored but the heaping effect is neglected, i. e. only rounding to months is considered (green line).
- A model based on right censoring that mimics the first model by randomly spreading the death times across the corresponding intervals. Hence, this model also accounts for both types of roundings.

Obviously, ignoring the heaping effect leads to highly implausible results, with risk estimates approximating zero, where no deaths are recorded. Incorporating the heaping effect significantly reduces this phenomenon but still leaves some fluctuations in the estimate which are not expected to reflect the true temporal development of the hazard rate. Surprisingly, model 3 leads to the most plausible, smooth estimate for the log-baseline. Probably, this outcome is caused by the additional information assumed in this model.

Since all observed death times are treated as exactly observed, the model contains more information than the corresponding model based on interval censoring which is, therefore, more susceptible to produce artificial behavior.

Surprisingly, all three approaches lead to almost identical results for the covariate effects. However, as we will see in the simulation study in Section 19,L the interval censoring approach is expected to produce more accurate estimates even if the baseline is estimated with some error.

18 A simulation study comparing different amounts of right censoring

18.1 Simulation setup

To gain deeper insight in the statistical properties of mixed model based estimates for survival data with different percentages of right censoring, we performed a simulation study. More general censoring schemes will be considered in a second simulation study in Section 19. As a competing method, we used the fully Bayesian approach to structured hazard regression proposed in Hennerfeind et al. (2005). For the empirical Bayes approach, both a quantile based grid and an equidistant grid with 300 knots were utilized in the numerical integration algorithm (compare Section 15.2).

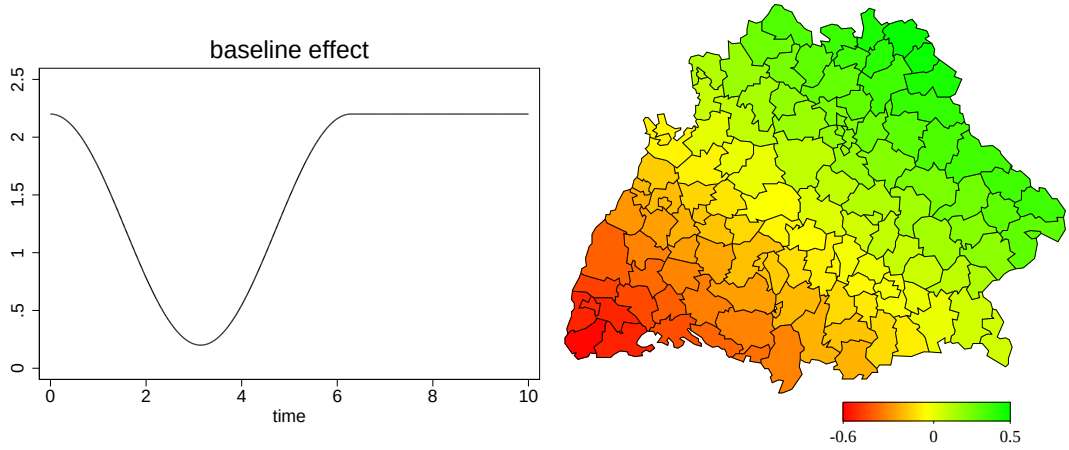


Figure 18.1: Baseline hazard (left) and spatial effect (right) employed in the simulation study.

We generated $R = 250$ data sets, each with $n = 750$ observations based on the following structured additive predictor:

$$\eta_i(t) = g_0(t) + f(x_i) + f_{spat}(s_i). \tag{18.1}$$

The baseline hazard rate $\lambda_0(t) = \exp(g_0(t))$ (shown in the left part of Figure 18.1) is chosen to reflect a situation, where the risk for an event is initially high, decreasing for some time and rising again at the end of the observation period. Such bathtub-shaped hazard rates are quite common in studies on survival times, as we have already seen in Section 16, but can hardly be handled within a regression approach assuming a parametric form of the baseline, e. g. a Weibull distributed baseline. The nonparametric effect $f(x)$ is given by a sine curve

$$f(x) = 0.6 \cdot \sin(\pi(2x - 1)),$$

where x is chosen randomly from an equidistant grid of 75 values within the interval $[0, 1]$. The spatial function f_{spat} is defined based on the centroids of the 124 districts of the two southern states of Germany (Bavaria and Baden-Württemberg) and is shown in the right part of Figure 18.1. Again, the value s is randomly assigned to the observations. To obtain censored observations, we generated independent, exponentially distributed

censoring times C_i and defined the observed survival time to be $t_i = \min(T_i, C_i)$, where T_i is generated according to the hazard rate $\lambda_i(t) = \exp(\eta_i(t))$. Three different amounts of right censoring were considered:

- no censoring at all,
- moderate censoring ($C_i \sim \text{Exp}(0.2)$, corresponding to 10-15% censored observations), and
- high censoring ($C_i \sim \text{Exp}(0.6)$, corresponding to 20-25% censored observations).

In general, it is not clear how to simulate survival times from a Cox-type model with hazard rate $\lambda_i(t)$, since this hazard rate does not correspond to a commonly known distribution. We used a technique based on the inversion principle described in Bender, Augustin & Blettner (2005). This principle allows the simulation of Cox models with arbitrary baseline hazard as long as the cumulative baseline hazard $\Lambda_0(t)$ and its inverse $\Lambda_0^{-1}(t)$ are available (at least for numerical evaluation). In this case uncensored survival times T_i can be simulated via

$$T_i = \Lambda_0^{-1}[-\log(U_i)\exp(-f(x_i) - f_{spat}(s_i))],$$

where U_i is a random variable uniformly distributed on $[0, 1]$, i. e. $U_i \sim U[0, 1]$.

18.2 Results

The results of the mixed model approach were compared to the fully Bayesian MCMC approach by Hennerfeind et al. (2005) based on empirical MSEs, bias and average coverage probabilities. The results can be summarized as follows:

- In terms of MSE, none of the approaches can be considered superior in all situations (Figure 18.2).
- Covariate effects $f(x)$ and $f_{spat}(s)$ are estimated with approximately the same precision regardless of the amount of censoring. For the log-baseline, the median MSE remains roughly the same but the variability of the MSE increases with and increasing percentage of censoring.
- In case of no or medium censoring, REML estimates based on an equidistant grid and the MCMC approach yield estimates of comparable quality for the baseline hazard, while the quantile based grid results in higher MSEs. For a high amount of censoring, the equidistant and the quantile based grid interchange. This change is due to the fact that with no censoring or only a small amount of censoring outliers with large survival times are more likely to be encountered. With a high amount of censoring, these observations are usually censored at a smaller time point. Obviously, an equidistant grid can easier accommodate for holes without data on the time axis.
- For the covariate effects, differences in terms of MSE are generally very small.
- For the nonparametric effect $f(x)$, the quantile based grid yields lowest MSEs with very small differences in case of high censoring.

- For the spatial effect, a minor superiority of the mixed model based estimates can be observed.
- Average estimates for $f(x)$ and $f(s)$ are almost the same for MCMC estimates and estimates obtained with a quantile based grid for all censoring mechanisms (Figures 18.3 and 18.4 show the empirical bias for the cases with no and high censoring). In contrast, using an equidistant grid for the integration introduces considerably more bias.
- In terms of average coverage probabilities, the MCMC approach produces more conservative credible intervals than both mixed model approaches (Table 18.1). When comparing the different integration techniques for the empirical Bayes approach, differences are small for the baseline while more conservative credible intervals are obtained with the quantile based grid for the covariate effects.
- All approaches meet the nominal levels but are mostly to conservative.

		$g_0(t)$		$f_1(x)$		$f_{spat}(s)$	
		80%	95%	80%	95%	80%	95%
REML equidistant grid	no censoring	0.918	0.967	0.784	0.949	0.899	0.984
	medium censoring	0.921	0.962	0.810	0.954	0.904	0.984
	high censoring	0.859	0.952	0.816	0.953	0.905	0.983
REML quantile based grid	no censoring	0.870	0.94	0.852	0.968	0.939	0.993
	medium censoring	0.896	0.945	0.854	0.964	0.940	0.993
	high censoring	0.900	0.969	0.844	0.965	0.933	0.991
MCMC	no censoring	0.931	0.976	0.842	0.967	0.943	0.995
	medium censoring	0.944	0.977	0.841	0.965	0.943	0.994
	high censoring	0.974	0.986	0.836	0.966	0.934	0.993

Table 18.1: Average coverage probabilities. Values that are more than 2.5% below (above) the nominal level are indicated in green (red)

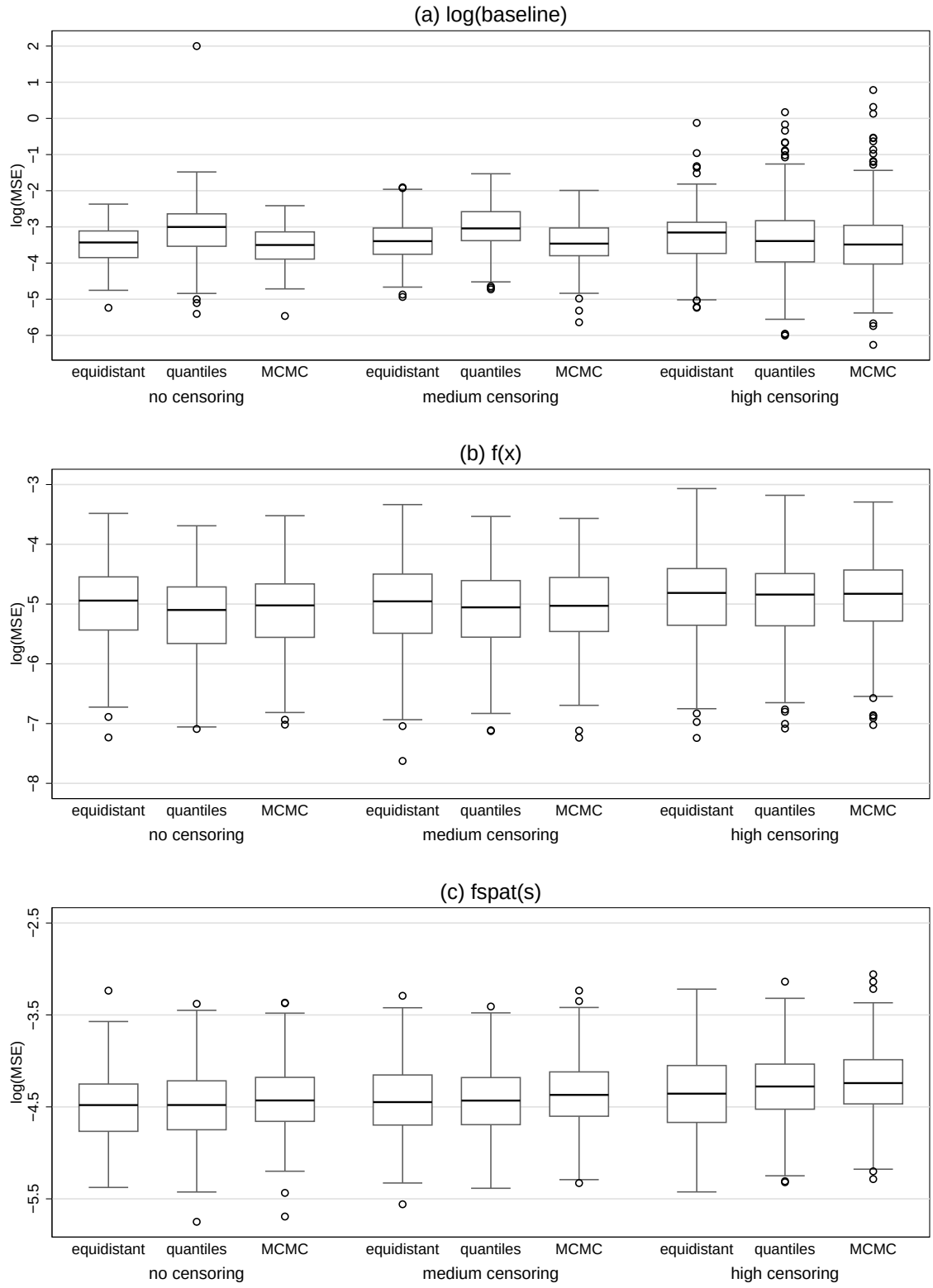


Figure 18.2: Boxplots of log(MSE).

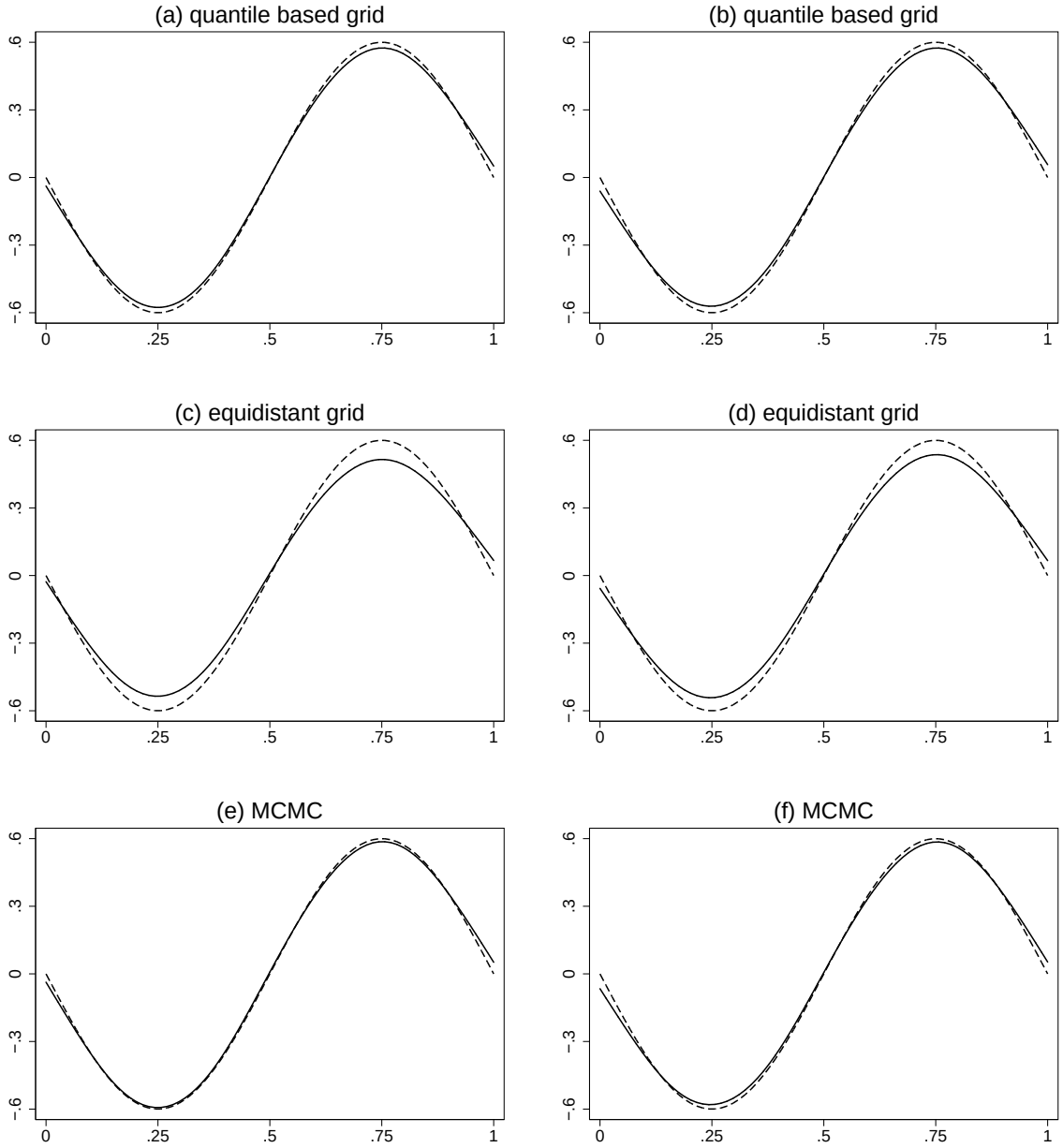


Figure 18.3: Bias for $f(x)$. Results obtained with no censoring are displayed in the left panel, results obtained with high censoring in the right panel. The true function is given as dashed line, average estimates as solid line.

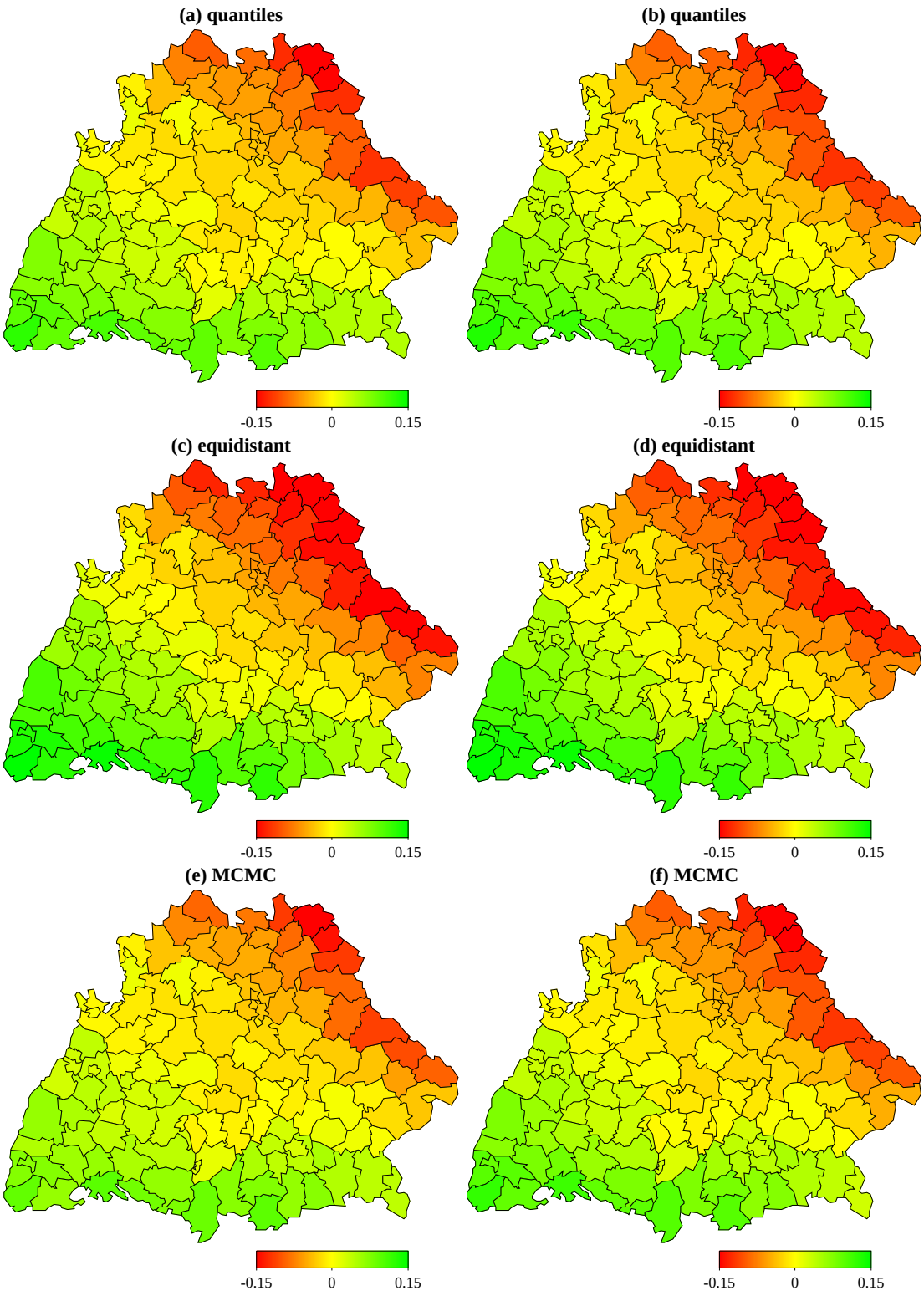


Figure 18.4: Bias for $f_{spat}(s)$. Results obtained with no censoring are displayed in the left panel, results obtained with high censoring in the right panel.

19 Ignoring interval censoring: A simulation study

19.1 Simulation setup

In order to investigate the impact of ignoring interval censoring when analyzing survival data, we conducted a simulation study that mimics a situation frequently found in clinical studies: The survival status of a patient is assessed at fixed dates until the end of the study. Exact survival times were generated from a geoadditive model with hazard rate

$$\lambda(t; x, s) = \exp(g_0(t) + f(x) + f_{spat}(s)),$$

where $g_0(t)$ is the log-baseline hazard rate, $f(x)$ is a function of the continuous covariate x with sinusoidal form, and $f_{spat}(s)$ is a spatial function defined by the density of a mixture of two two-dimensional normal distributions. Two different baseline hazard rates were applied: A bathtub-shaped one with strong variation over the whole time-domain and a relatively flat one. All survival times exceeding 8 were treated as right censored at $C = 8$. The remaining interval $[0, 8]$ was divided into l equidistant intervals and each observation was assigned to the interval the corresponding survival time belonged to. To evaluate the impact of interval censoring, we compared three different values for l, namely $l = 8$, $l = 16$ and $l = 32$, corresponding to intervals with length 1, 0.5 and 0.25. The simulation design is summarized in more detail in Figure 19.1.

The resulting data sets were analyzed based on three different strategies:

- Use the correct censoring mechanisms, i. e. treat all observations with survival times less than 8 as interval censored and all other observations as right censored (IC).
- Use a binary discrete time survival model with complementary log-log link. Such a model can be seen as a grouped Cox model (compare Section 14.4.2 and Fahrmeir & Tutz (2001, Ch. 9)) (CLL).
- Treat all observations with survival times less than 8 as uncensored and all other observations as right censored. To account for interval censoring, uncensored observations are spread randomly across the corresponding intervals (UC).

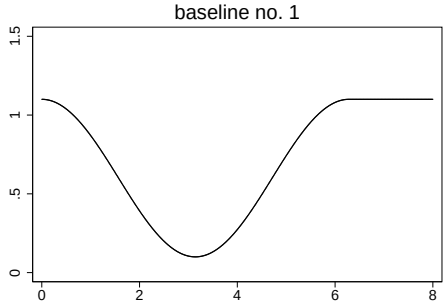
Note that we also tried to treat all survival times less than 8 as uncensored without spreading the observations across the intervals. However, due to numerical problems this strategy could not be routinely applied and is, therefore, not included in the comparison.

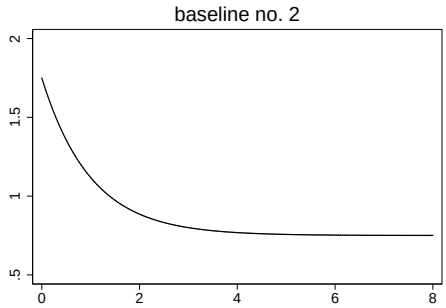
Both the log-baseline and the effect of x are modeled by cubic P-splines with second order random walk penalty and 20 inner knots. The spatial effect is estimated using Markov random field prior (4.25).

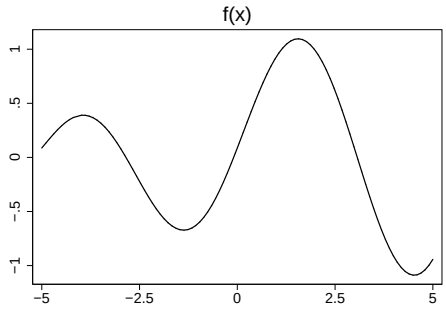
19.2 Results

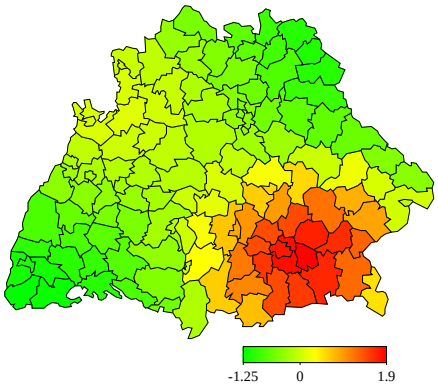
The results of the simulation study can be summarized as follows:

- In case of the bathtub-shaped baseline, the interval censoring approach leads to the best estimates for the baseline hazard rate. While the discrete time model performs









- Hazard rate:

$$\lambda(t; x, s) = \exp(g_0(t) + f(x) + f_{spat}(s))$$

- Baseline no. 1:

$$\exp(g_0(t)) = \begin{cases} 0.5 \cdot [\cos(t) + 1.2], & t \leq 2\pi \\ 0.5 \cdot [1 + 1.2], & t > 2\pi \end{cases}$$

- Baseline no. 2:

$$\exp(g_0(t)) = \exp(-t) + 0.75$$

- $f(x) = \sin(1.05x) \cdot \log(x + 6)$
- x is chosen randomly from an equidistant grid of 100 values between -5 and 5.
- $f_{spat}(s) = N(\mu_1, \Sigma_1, s_x, s_y) + N(\mu_2, \Sigma_2, s_x, s_y) - 1.4$ with

$$\mu_1 = \begin{pmatrix} 0.75 \\ 0.25 \end{pmatrix}, \quad \Sigma_1 = \begin{pmatrix} 0.05 & 0.01 \\ 0.01 & 0.05 \end{pmatrix},$$

$$\mu_2 = \begin{pmatrix} 0.25 \\ 0.75 \end{pmatrix}, \quad \Sigma_2 = \begin{pmatrix} 0.1 & 0.01 \\ 0.01 & 0.1 \end{pmatrix}.$$

- (s_x, s_y) are the centroids of the 124 districts s of the two southern states of Germany.
- Survival times exceeding 8 are considered as right censored.
- The interval $[0, 8]$ is divided in $l = 8, 16$ or 32 equidistant parts for interval censoring.
- Number of observations per replication: $n = 500$.
- Number of simulation runs: $R = 100$.

Figure 19.1: Simulation design.

comparably well for a sufficient large number of intervals, the uncensored approach remains dissatisfying (Figure 19.2a).

- In contrast, in case of the flat baseline, the discrete time model leads to the best estimates for the baseline for a small number of intervals. For a larger number of intervals, both the interval censoring approach and the discrete time model give comparable results and, again, outperform the uncensored approach. (Figure 19.3a).

Considering covariate effects, both types of baseline hazard rates lead to similar conclusions. Hence, we will focus on results for the bathtub-shaped baseline.

- For a sufficiently large number of intervals, all strategies lead to a comparable fit for the nonparametric effect $f(x)$ in terms of MSE. For a smaller number of intervals, the interval censoring approach and the discrete time model give preferable results compared to the uncensored approach (Figure 19.2b and Figure 19.3b).
- Considering the spatial effect, the discrete time model leads to the best fit for a small and a medium number of intervals. The quality of both the interval censoring and the uncensored approach increases with an increasing number of intervals, but only the interval censoring approach reaches results comparable to those of the discrete time model (Figure 19.2c and Figure 19.3c).
- Figures 19.4 and 19.5 show similar results based on average estimates for the spatial function and the nonparametric effect, respectively. While the uncensored approach introduces noticeably more bias for a small number of intervals, the discrete time model and the interval censoring approach lead to comparable estimates. When increasing the number of intervals, differences between the three strategies become smaller but are still present.
- Considering average coverage probabilities (see Table 19.1 for the results obtained with the bathtub-shaped baseline), the interval censoring approach leads to the best results.
- In case of the baseline hazard rate, too narrow credible intervals are obtained with all approaches if only a small number of intervals is given. However, the interval censoring approach still is closest to the nominal level. For an increasing number of intervals, the discrete time model comes closer to the nominal level while the uncensored approach does not improve. The interval censoring approach meets the nominal level already with a medium number of intervals.
- Both the interval censoring approach and the discrete time model produce comparable and satisfying average coverage probabilities for the covariate effects. In contrast, the uncensored approach produces too narrow credible intervals for small and medium numbers of intervals.

Based on these results, we come to the conclusion that the impact of interval censoring depends on the structure of the underlying model, especially on the baseline hazard rate. While details of the model may be lost by ignoring interval censoring for highly fluctuating baselines and a relatively small number of intervals, this effect decreases for an increasing number of intervals. When the baseline is relatively flat, interval censoring does not per se lead to improved estimates but, in any case, performs better than an approach based on randomly spreading the observations across the intervals. Note also, that the discrete time model is only applicable when all intervals have the same width. When considering average coverage probabilities for the baseline, the discrete time approach as well as treating survival times as uncensored, lead to credible intervals which are too narrow.

		$g_0(t)$		$f_1(x)$		$f_{spat}(s)$	
		80%	95%	80%	95%	80%	95%
8 intervals	IC	0.681	0.882	0.828	0.963	0.790	0.925
	CLL	0.388	0.537	0.845	0.967	0.837	0.962
	UC	0.416	0.680	0.576	0.770	0.551	0.703
16 intervals	IC	0.813	0.947	0.838	0.962	0.848	0.970
	CLL	0.576	0.757	0.845	0.963	0.871	0.980
	UC	0.416	0.680	0.702	0.894	0.742	0.886
32 intervals	IC	0.813	0.942	0.850	0.963	0.876	0.981
	CLL	0.661	0.844	0.852	0.961	0.883	0.983
	UC	0.281	0.472	0.786	0.937	0.808	0.939

Table 19.1: Bathtub-shaped baseline: Average coverage probabilities. Values that are more than 2.5% below (above) the nominal level are indicated in green (red)

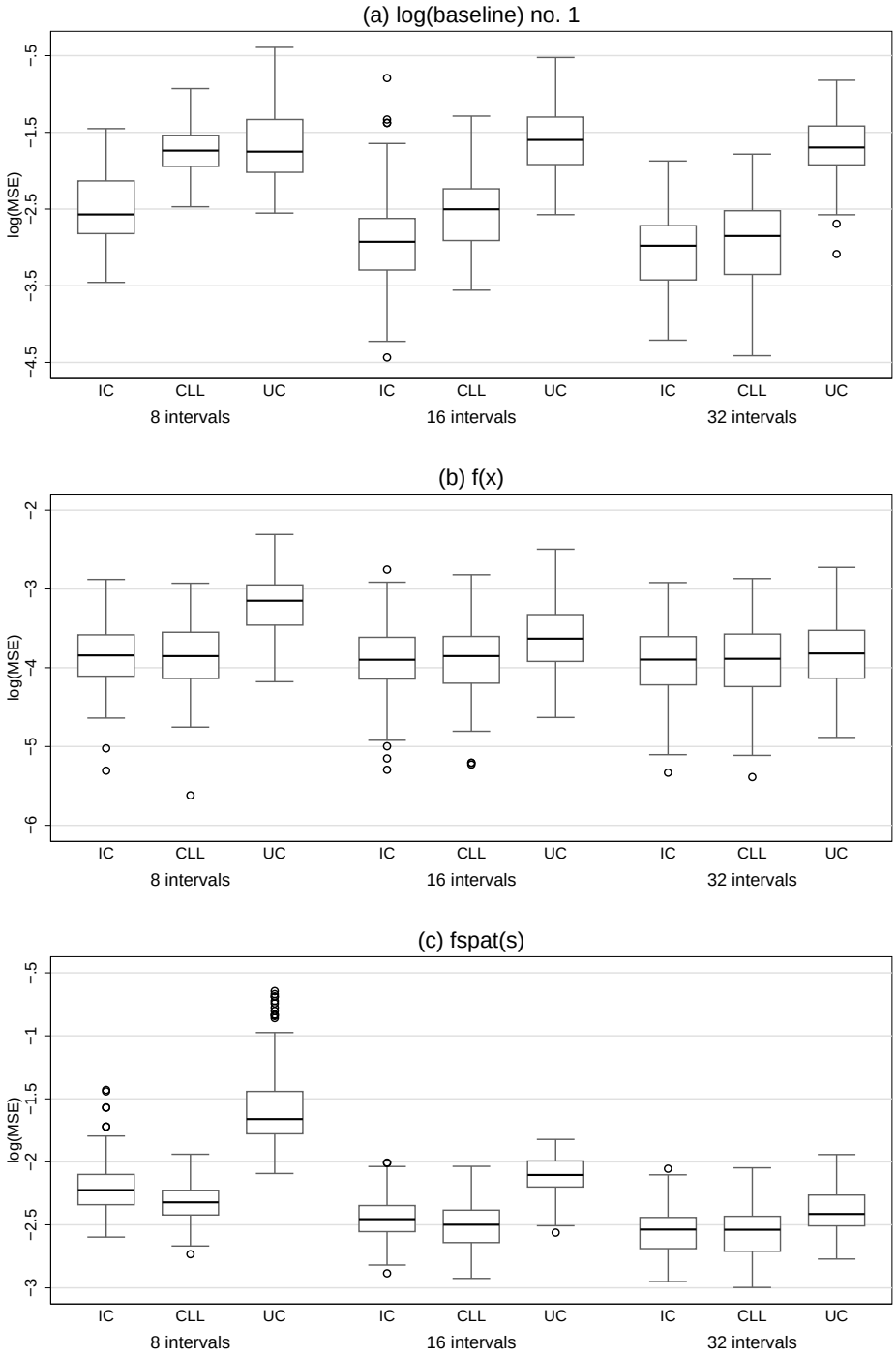


Figure 19.2: Bathtub-shaped baseline: Boxplots of log(MSE) for the baseline hazard rate, the nonparametric effect and the spatial effect. IC denotes results from treating survival times as interval censored, CLL denotes results from the discrete time complementary log-log model and UC denotes results from treating the survival times as uncensored.

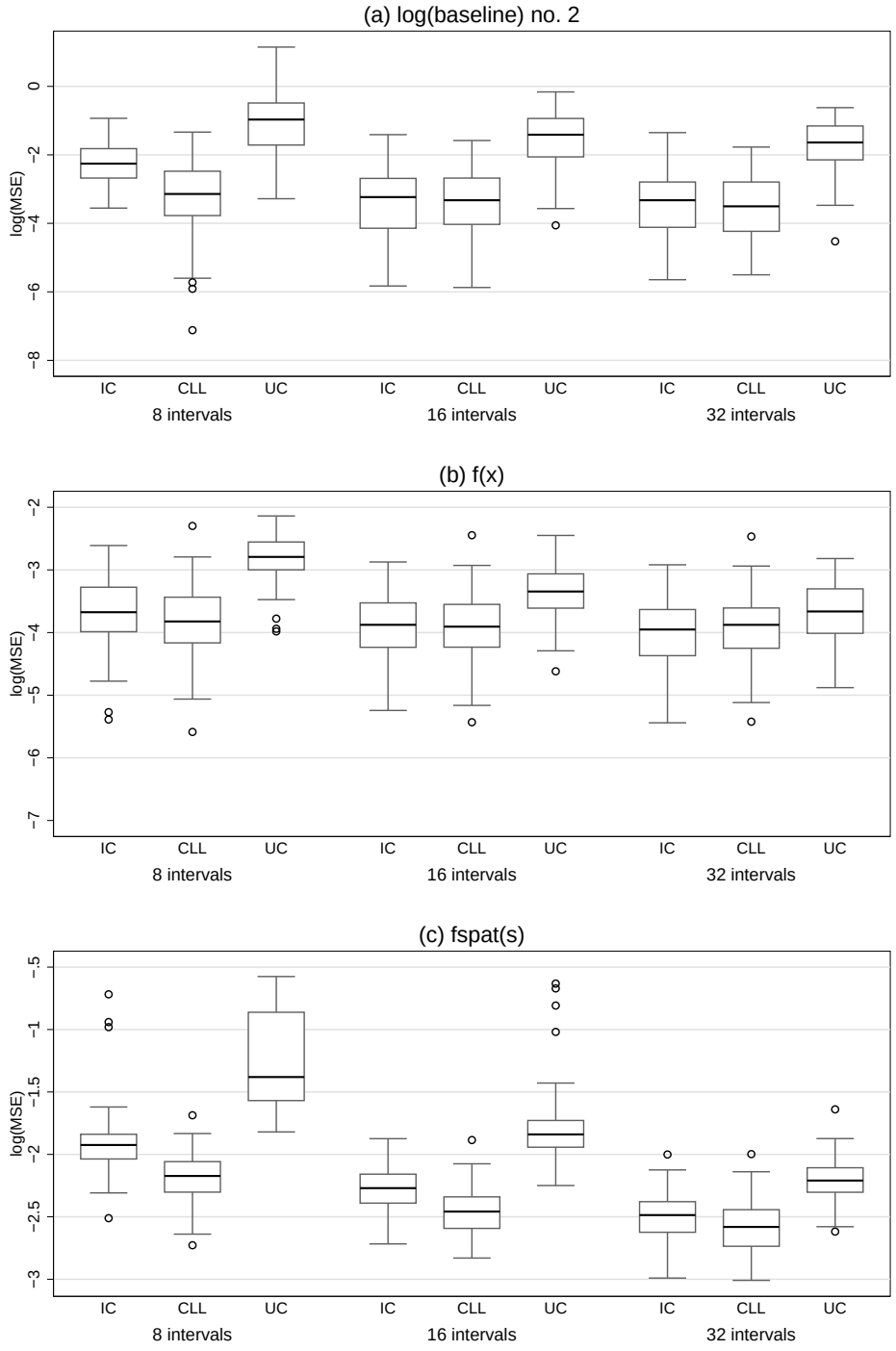


Figure 19.3: Smooth baseline: Boxplots of log(MSE) for the baseline hazard rate, the nonparametric effect and the spatial effect. IC denotes results from treating survival times as interval censored, CLL denotes results from the discrete time complementary log-log model and UC denotes results from treating the survival times as uncensored.

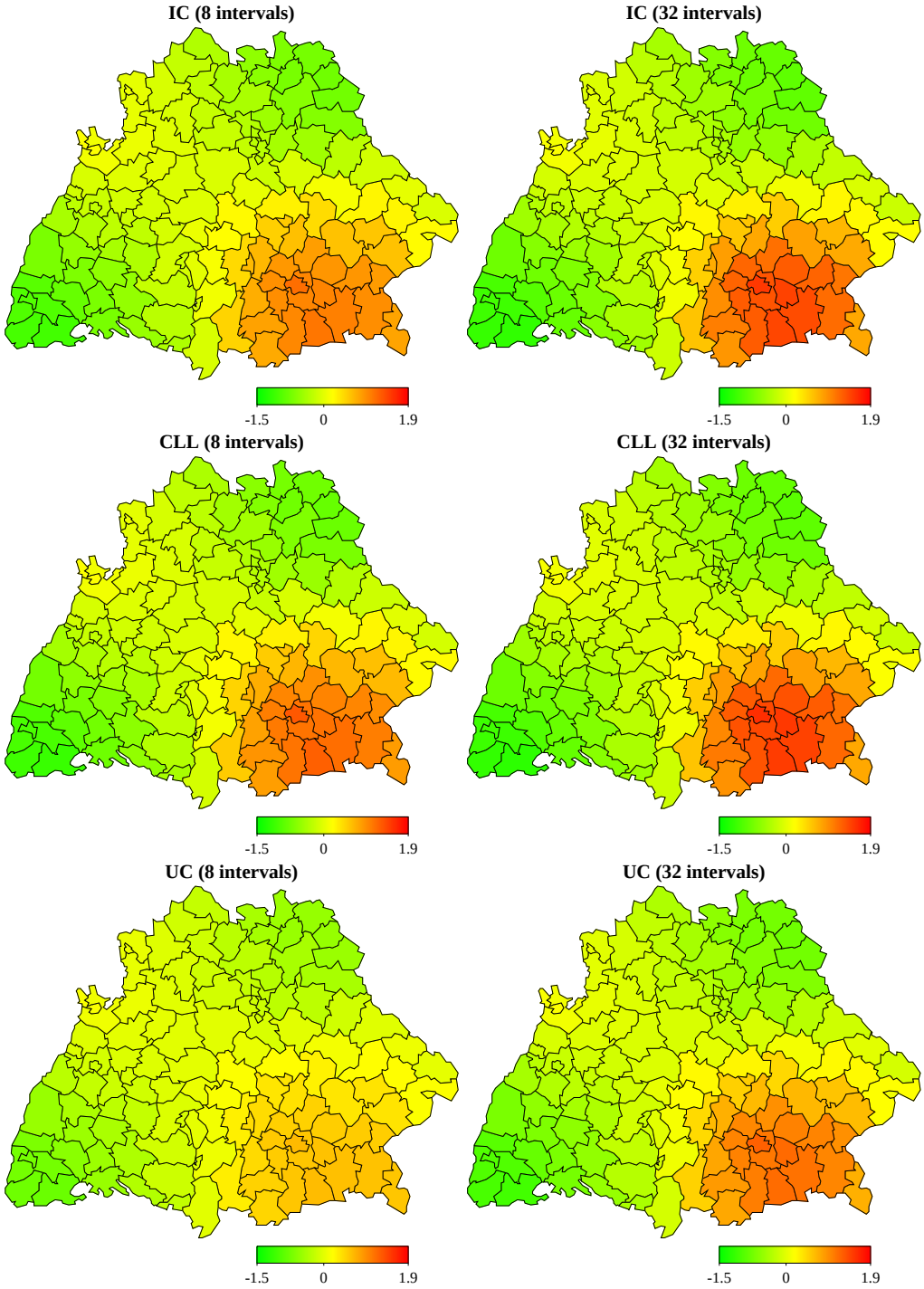


Figure 19.4: Bathtub-shaped baseline: Average estimates for the spatial effect $f_{spat}(s)$. IC denotes results from treating survival times as interval censored, CLL denotes results from the discrete time complementary log-log model and UC denotes results from treating the survival times as uncensored.

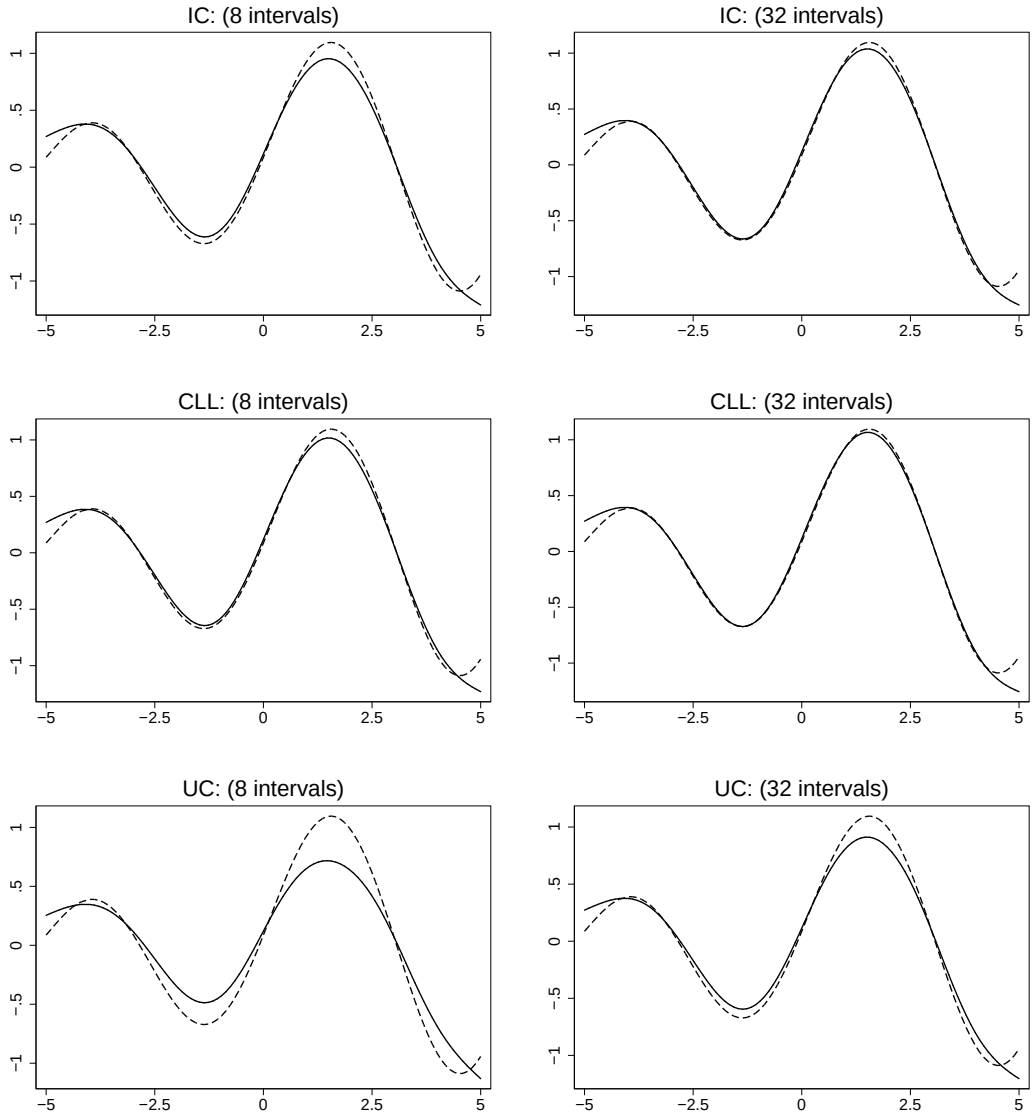


Figure 19.5: Bathtub-shaped baseline: Average estimates (solid lines) and true values (dashed lines) for the nonparametric effect $f(x)$. IC denotes results from treating survival times as interval censored, CLL denotes results from the discrete time complementary log-log model and UC denotes results from treating the survival times as uncensored.

Part V

Summary and outlook

Due to the increasing availability of highly complex spatio-temporal regression data, appropriate statistical methodology is clearly needed in practice. Theoretical background for spatio-temporal regression models and extensions is provided by structured additive regression models, originally introduced within a fully Bayesian framework with estimation based on Markov chain Monte Carlo simulation techniques. Structured additive regression comprises numerous types of effects, including nonparametric effects of continuous covariates, spatial effects, time-varying seasonal effects, unobserved heterogeneity or complex interaction effects, and allows for inference in a unified framework. Within this thesis, an alternative to MCMC inference for the estimation of structured additive regression models was presented and adapted to different types of responses. This inferential procedure bases on the representation of structured additive regression models as variance components models and, therefore, allows to apply mixed model methodology to estimate all model components. Especially, the estimation of the smoothing parameters corresponding to the nonparametric terms of the model is included.

In the first part of this thesis, structured additive regression models for univariate responses and the different model terms were reviewed in detail, and theoretical properties related to the mixed model representation were derived. In the next step, the reparametrization in terms of a variance components model was introduced and a numerical algorithm for the estimation of all model components was provided. This included the derivation of numerically tractable formulae that allow for fast optimization even with a large number of observations and model terms. Within the following parts, structured additive regression models and mixed model based inference were transferred to the more general situations of multivariate models for categorical responses and regression models for continuous survival times. In the latter case, several censoring mechanisms were considered, comprising left, right, and interval censoring as well as left truncation.

In addition to the theoretical development of mixed model based inference, simulation studies and real data applications were conducted to assess the performance of the new approach compared to competing methodology, especially fully Bayesian inference based on MCMC. These comparisons revealed that mixed model methodology provides a promising alternative estimation strategy that in most situations performs equally well or even better than its counterparts. Moreover, specific problems encountered with MCMC inference such as the determination of burn-in or questions concerning the mixing and the convergence of the generated Markov chain are not present, since inference is based on the optimization of likelihood based criteria. Furthermore, the empirical Bayes approach does not rely on specific choices of priors for the variance components. Hence, sensitivity with respect to prior assumptions is not an issue.

Of course, mixed model based inference also exhibits a few drawbacks. For example, the complexity of the estimation problem rapidly increases with the number of regression coefficients. This may be the case in models, where both a structured and an unstructured spatial effect of a large number of regions are considered. In MCMC inference, such estimation problems can be split into several smaller parts, since the separate terms are updated one by one based on their full conditionals. The backfitting algorithm provides a similar idea within the mixed model formulation, but has the disadvantage that credible intervals for the estimated effects are no longer available. Moreover, with today's computing resources approximately 2,000–3,000 regression parameters can be estimated based on the formulae presented in Section 5.3 which should be sufficient in most applications. Of

course, the corresponding computations will be computerintensive and may take a lot of time in models of this size.

An open question in mixed model based inference is in which situations the estimation is guaranteed to converge and in which situations convergence problems will occur. In the simulation studies in Sections 9 and 13, convergence problems were encountered in a number of replications, but it could also be shown that after a large number of iterations results showed no differences to those from converged models. In these cases, only one variance component switched between two values relatively close to each other and, thus, differences in the corresponding estimates could be neglected. Convergence properties of the posterior mode estimates obtained with mixed model methodology could be elucidated by investigating the relationship between posterior means and posterior modes in structured additive regression models. In most of the simulation studies and applications, both approaches yielded estimates which are relatively close to each other. Conditions characterizing situations where posterior means and modes are similar would be of interest, since in these cases convergence problems are less likely to occur.

The mixed model representation of structured additive regression models does not only provide an alternative estimation technique, but also allows for additional theoretical insights. For example, the identifiability problems of nonparametric regression models could easily be stated and interpreted in terms of the mixed model representation (compare Section 5.1). Hopefully, further theoretical results will be derived using the mixed model representation as a building block, e. g. to judge whether the posterior is proper in a fully Bayesian approach to structured additive regression models. Impropriety of the posterior can not be diagnosed from the MCMC output and fitting a model with improper posterior may lead to false conclusions. It is well known from models with (proper) random effects that certain choices for the hyperparameters of the inverse gamma priors assigned to the variance parameters lead to models with improper posteriors (Hobert & Casella 1996). Sufficient and necessary conditions for the existence of the posterior in mixed models are stated in Sun, Tsutakawa & He (2001). In principle, it should be possible to adapt these conditions to structured additive regression models based on the mixed model representation.

In the future, application of the mixed model approach in further complex examples will be of great importance. Especially in connection with more detailed case studies, questions concerning model choice and model validation will require additional attention. Currently, information criteria like AIC and BIC or the generalized cross validation criterion are used to select an adequate model (see for example Section 12 and Section 16). However, diagnostic tools should supplement such overall criteria to investigate whether the different covariates are modeled appropriately. Besides graphical assessments, formal tests on the functional form and the presence of covariate effects would be desirable. Throughout the last years, likelihood ratio and score tests on variance components in mixed models have been developed. Hence, the mixed model representation of structured additive regression models in principle enables their application in nonparametric regression settings. However, these tests either rely on restrictive assumptions concerning the longitudinal arrangement of the data (Stram & Lee (1994, 1995), Verbeke & Molenberghs (2003)), consider only one variance component (Craniceanu & Ruppert 2004a), or employ a simulated distribution of the test statistic under the null hypothesis (Craniceanu & Ruppert 2004a, 2004b), rendering their routine application in complex examples com-

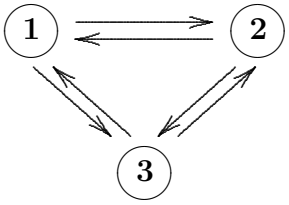
plicated. Within a fully Bayesian framework, tests on the presence of covariate effects are often performed by introducing additional dummy variables that determine whether a model term is to be included or excluded (e. g. Yau et al. 2003). Kauermann & Eilers (2004) present a similar idea within a mixed model based framework for the selection of expressed genes. An extension of their concepts to either testing model components in a structured additive regression model or their variance parameters would be a valuable enhancement of mixed model based inference.

Further developments could aim at the generalization of structured additive regression to parametric regression models other than those considered in this work. For example, Osuna (2004) presents extended regression models for count data with overdispersion or zero inflation and according inferential schemes based on MCMC. In principle, these models could also be estimated based on mixed model methodology. In case of categorical responses, additional parametric regression models such as two-step models (compare Fahrmeir & Tutz 2001, Ch. 3.3.6) may be straightforwardly adapted to the semiparametric structured additive regression setting by defining appropriate link functions and design matrices.

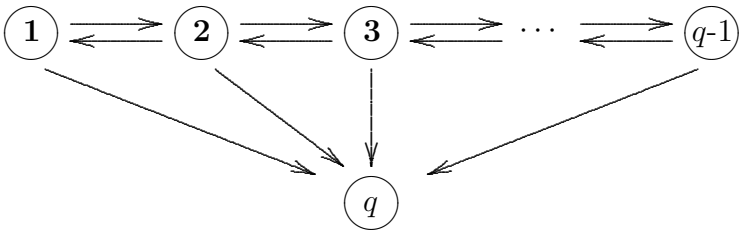
A more challenging task is the development of multinomial probit models, especially when allowing for correlated latent utilities. In the econometric literature, correlated probit models are frequently proposed as an alternative to the multinomial logit model, since they allow for less restrictive substitution patterns and do not exhibit the independence from irrelevant alternatives property (see e. g. Train 2003, Ch. 5). In this context, parametric multinomial probit models are usually fitted based on simulated maximum likelihood techniques. With semiparametric extensions, the mixed model approach would be a useful alternative, since it provides a theoretical framework for both the determination of the variance-covariance parameters of the latent variables and the smoothing parameters. However, the application is not straightforward, since computation of the likelihood requires the evaluation of choice probabilities (compare Equation (10.10) on page 140) on the basis of specialized numerical methods. Within a fully Bayesian approach, correlated multinomial probit models are easily estimated by augmenting the latent factors as additional parameters (see e. g. Chib & Greenberg (1998) for a parametric example). The fully Bayesian estimation scheme is straightforwardly extendable to more general, semiparametric predictors.

Based on the work for survival times presented in Part IV, more general multi state models could be considered, where a stochastic process with finite state space is observed in continuous time (see Andersen & Keiding (2002) for an introduction to multi state models). Note that discrete time multi state models can already be analyzed with categorical regression models via data augmentation (similar as in discrete time survival models, compare Section 14.4.2). Multi state models are most easily described by transition graphs or Markov graphs. Figure 19.6 displays Markov graphs for three different types of multi state models. In the first case, the state space is formed by three states and all states are mutually accessible. This implies that all the three states are recurrent and, hence, this model is also called a model for recurrent events. An example for this type of multi state data are durations of (un)employment, where the three states are given by full time employment, part time employment and unemployment. Since none of the three states is absorbing, i. e. all states are left with positive probability, they can be modeled as recurrent events.

(a) Recurrent events



(b) Disease progression



(c) Competing risks

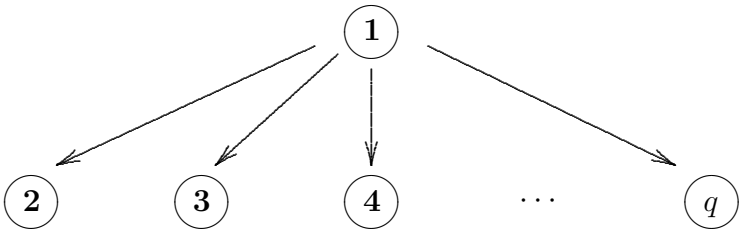


Figure 19.6: Markov graphs for three different types of multi state models.

Figure 19.6b shows a different type of multi state model called the disease progression model that describes the temporal development of a certain disease. If the severity of this disease can be grouped into $q-1$ ordered stages of increasing severity, a reasonable model might look like this: Starting from disease state 'j', an individual can only move to contiguous states, i. e. either the disease gets worse and the individual moves to state '$j+1$', or the disease attenuates and the individual moves to state '$j-1$'. Obvious modifications have to be made for state '1' and state '$q-1$', since no further improvements or deterioration is possible at this points. In addition to the disease states, death can be included as a further state 'q'. In Figure 19.6b it is assumed that individuals can die at any disease stage.

A further multi state model is the model of competing risks. In this case, different types of absorbing states are modeled simultaneously. For example, we might consider different types of cancer and analyze the survival time of patients up to death from one of these cancer types. Then state '1' corresponds to healthy individuals, while the remaining states '2',...,'q' correspond to the different cancer types.

Multi state models can fully be described by transition-specific intensity processes. From these intensities, transition probabilities in terms of cumulative distribution functions or survivor functions can be derived. In case of survival data, only one transition is possible from the state 'alive' to the state 'death' and, hence, only one intensity or hazard rate is considered. The simplest multi state model is a homogenous Markov process, where the transition times between changes of the state are independent and the transition intensities are constant over time. This corresponds to exponential distributed durations in the different states and transition probabilities determined by an embedded Markov chain. In semiparametric multi state models, the assumption of a constant hazard rate for the transition times is replaced with a flexible hazard rate modeled in analogy to the Cox model. This flexible hazard rate then depends on covariates and the nonparametric extensions considered in Part IV can also be employed. More difficult models arise, if some effects are assumed to be the same for several transition intensities, or if some covariates only influence specific transitions.

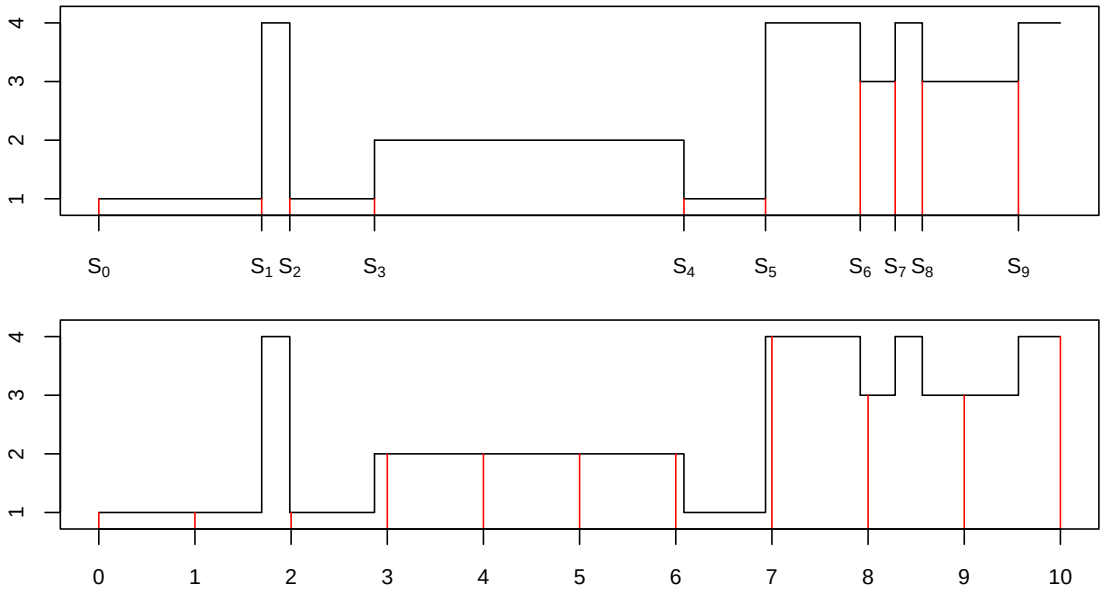


Figure 19.7: Realization of a multi state model with exact transition times (top) and interval censored transition times observed at equidistant time points (bottom).

If exact transition times are observed, the likelihood of a multi state model can be calculated in terms of the transition intensities and mixed model methodology should, in principle, be applicable with only minor accommodations. Yet, as with interval censored survival times, observations of multi state models are frequently not collected at the exact transition times but at some prespecified time points. This data structure is illustrated in Figure 19.7. In the upper part of the Figure, the exact transition times $S_1, \ldots, S_9$ are observed along with the corresponding states. In contrast, if we can only observe the state of the process at the discrete time points $0, 1, \ldots, 10$, we arrive at the situation displayed in the lower part of the Figure. Due to the coarsening mechanism, some transitions may be lost (e. g. between the time points 6 and 7) while for other duration times the same state is observed several times (e. g. between S_3 and S_4). In this case, the likelihood construction becomes significantly more complicated and additional numerical efforts have to be made (see Commenges (2002) for a review of recent work on interval censored multi state models).

A completely different model for the analysis of duration times is the accelerated failure time model discussed in Section 14.5. It has the advantage that proportionality of the hazard rates is not assumed. However, the Cox model is more routinely applied in practice, since it does not specify a parametric distribution for the survival times and, hence, yields more general estimates in this respect. Moreover, the inclusion of time-varying covariates and effects is possible in the Cox model, allowing both for more realistic models and relaxation of the assumption of proportional hazards. Still, accelerated failure time models would provide an attractive alternative to the Cox model if the duration time distribution could be estimated in a more flexible way. Komárek et al. (2005) present an approach based on mixtures of Gaussian densities with penalized mixing coefficients that provides this extension. However, in Komárek et al. (2005) the smoothing parameter associated with the mixing coefficients has to be chosen according to AIC based on a grid search. In contrast, a combination with mixed model methodology would allow both for the inclusion of more general covariate effects and for the estimation of this smoothing parameter in a unified approach.

References

Abramowitz, M. & Stegun, I. A. (1974). *Handbook of Mathematical Functions*. Dover, New York.

Agresti, A. (2002). *Categorical Data Analysis*. Wiley, New York.

Albert, J. H. & Chib, S. (1993). Bayesian analysis of binary and polychotomous response data. *Journal of the American Statistical Association* **88**, 669–679.

Andersen, P. K. & Keiding, N. (2002). Multi-state models for event history analysis. *Statistical Methods in Medical Research* **11**, 91–115.

Banerjee, S., Wall, M. M., Carlin, B. P. (2003). Frailty modelling for spatially correlated survival data, with application to infant mortality in Minnesota. *Biostatistics* **4**, 123–142.

Banerjee, S. & Carlin, B. P. (2003). Semiparametric spatio-temporal frailty modeling. *Environmetrics* **14**, 523–535.

Bender, R., Augustin, T. & Blettner, M. (2005). Generating survival times to simulate Cox proportional hazards models. *Statistics in Medicine* **24**, 1713–1723.

Besag, J., York, J. & Mollié, A. (1991). Bayesian image restoration with two applications in spatial statistics (with discussion). *Annals of the Institute of Statistical Mathematics* **43**, 1–59.

Besag, J. & Higdon, D. (1999). Bayesian analysis of agricultural field experiments (with discussion). *Journal of the Royal Statistical Society B* **61**, 691–746.

Besag, J. & Kooperberg, C. (1995). On conditional and intrinsic autoregressions. *Biometrika* **82**, 733–746.

Biller, C. (2000). Adaptive Bayesian regression splines in semiparametric generalized linear models. *Journal of Computational and Graphical Statistics* **9**, 122–140.

Biller, C. & Fahrmeir, L. (2001). Bayesian varying-coefficient models using adaptive regression splines. *Statistical Modeling* **1**, 195–211.

Bogaerts, K., Leroy, R., Lesaffre, E. & Declerck, D. (2002) Modelling tooth emergence data based on multivariate interval-censored data. *Statistics in medicine* **21**, 3775–3787.

Breslow, N. E. & Clayton, D. G. (1993). Approximate inference in generalized linear mixed models. *Journal of the American Statistical Association* **88**, 9–25.

Brezger, A., Kneib, T. & Lang, S. (2005). BayesX: Analyzing Bayesian structured additive regression models. *Journal of Statistical Software*, **14** (11).

Brezger, A., Kneib, T. & Lang, S. (2005). BayesX Manuals. Available from http//www.stat.uni-muenchen.de/∼bayesx.

Brezger, A. & Lang, S. (2005). Generalized additive regression based on Bayesian P-splines. *Computational Statistics and Data Analysis*, to appear.

Cai, T., Hyndman, R. J. & Wand, M. P. (2002). Mixed model-based hazard estimation. *Journal of Computational and Graphical Statistics* **11**, 784–798.

Cai, T. & Betensky, R. A. (2003). Hazard regression for interval-censored data with penalized spline. *Biometrics* **59**, 570–579.

Carlin, B. P. & Banerjee, S. (2002). Hierarchical multivariate CAR models for spatio-temporally correlated survival data. In: Bernardo et al. (eds.): *Bayesian Statistics 7*. University Press, Oxford.

Chen, Z. (1993). Fitting multivariate regression functions by interaction spline models. *Journal of the Royal Statistical Society B* **55**, 473–491.

Chib, S. & Greenberg, E. (1998). Analysis of multivariate probit models. *Biometrika* **85**, 347–361.

Chiles, J.-P. & Delfiner, P. (1999). *Geostatistics. Modeling Spatial Uncertainty*. Wiley, New York.

Commenges, D. (2002). Inference for multi-state models from interval-censored data. *Statistical Methods in Medical Research* **11**, 167–182.

Cox, D. R. (1972). Regression models and life tables (with discussion). *Journal of the Royal Statistical Society B* **34**, 187–220.

Craniceanu, C. M. & Ruppert, D. (2004a). Restricted likelihood ratio tests in nonparametric longitudinal models. *Statistica Sinica* **14**, 713–729.

Craniceanu, C. M. & Ruppert, D. (2004b). Likelihood ratio tests in linear mixed models with one variance component. *Journal of the Royal Statistical Society B* **66**, 165–185.

Currie, I. & Durbán, M. (2002). Flexible smoothing with P-splines: A unified approach. *Statistical Modelling* **2**, 333–349.

de Boor, C. (1978). *A Practical Guide to Splines*. Springer, Berlin.

Denison, D. G. T., Mallick, B. K. & Smith, A. F. M. (1998). Automatic Bayesian curve fitting. *Journal of the Royal Statistical Society B* **60**, 333–350.

Dierckx, P. (1993). *Curve and Surface Fitting with Splines*. Clarendon Press, Oxford.

Diggle, P. J., Ribeiro Jr., P. J. & Christensen O. F. (2003). An introduction to model based geostatistics. In: Møller, J. (ed.): *Spatial Statistics and Computational Methods*. Springer, New York.

Diggle, P. J., Tawn, J. & Moyeed R. (1998). Model-based geostatistics (with discussion). *Journal of the Royal Statistical Society C* **47**, 299–350.

DiMatteo, I., Genovese, C. R. & Kass, R. E. (2001). Bayesian curve-fitting with free-knot splines. *Biometrika* **88**, 1055–1071.

Ecker, M. D. & Gelfand, A. E. (2003). Spatial modelling and prediction under stationary non-geometric range anisotropy. *Environmental and Ecological Statistics* **10**, 165–178.

Eilers, P. H. C. & Marx, B. D. (1996). Flexible smoothing using B-splines and penalties (with comments and rejoinder). *Statistical Science* **11**, 89–121.

Eilers, P. H. C. & Marx, B. D. (2003). Multivariate calibration with temperature interaction using two-dimensional penalized signal regression. *Chemometrics and Intelligent Laboratory Systems* **66**, 159–174.

Fahrmeir, L., Hamerle, A. & Tutz, G. (1996). *Multivariate statistische Verfahren.* De Gruyter, Berlin.

Fahrmeir, L., Kneib, T. & Lang, S. (2004). Penalized structured additive regression for space-time data: a Bayesian perspective. *Statistica Sinica* **14**, 731–761.

Fahrmeir, L. & Knorr-Held, L. (2000). Dynamic and semiparametric models. In: M. Schimek (ed.): *Smoothing and Regression: Approaches, Computation and Application.* Wiley, New York.

Fahrmeir, L. & Lang, S. (2001a). Bayesian inference for generalized additive mixed models based on Markov random field priors. *Journal of the Royal Statistical Society C* **50**, 201–220.

Fahrmeir, L. & Lang, S. (2001b). Bayesian semiparametric regression analysis of multicategorical time-space data. *Annals of the Institute of Statistical Mathematics* **53**, 11–30

Fahrmeir, L., Lang, S.,Wolff, J. & Bender, S. (2003). Semiparametric Bayesian time-space analysis of unemployment duration. *Journal of the German Statistical Society* **87**, 281–307.

Fahrmeir, L. & Tutz, G. (2001). *Multivariate Statistical Modeling Based on Generalized Linear Models*, Springer, New York.

Fleming, T. R. & Harrington, D. P. (1990). *Counting Processes and Survival Analysis*, Wiley, New York.

Fotheringham, A. S., Brunsdon, C., & Charlton, M. (2002). *Geographically Weighted Regression: The Analysis of Spatially Varying Relationships.* Wiley, Chichester.

Friedman, J. H. (1991). Multivariate adaptive regression splines (with discussion). *The Annals of Statistics* **19**, 1–141.

Gaisie, K., Cross, A. R. & Nsemukila, G. (1993). *Zambia Demographic and Health Survey 1992,* University of Zambia, Lusaka.

Gamerman, D. (1997). Sampling from the posterior distribution in generalized linear models. *Statistics and Computing* **7**, 57–68.

Gamerman, D., Moreira, A. R. B. & Rue, H. (2003). Space-varying regression models: specifications and simulation. *Computational Statistics and Data Analysis* **42**, 513–533.

George, A. & Liu, J. W.-H. (1981). *Computer Solution of Large Sparse Positive Definite Systems.* Prentice-Hall, Englewood Cliffs.

Green, P. J. (1987). Penalized likelihood for general semiparametric regression models. *International Statistical Review* **55**, 245–259.

Green, P. J. (2001). A primer on Markov chain Monte Carlo. In: Barndorff-Nielsen, O. E., Cox, D. R. and Klüppelberg, C. (eds.): *Complex Stochastic Systems*. Chapman & Hall, London.

Handcock, M. S., Meier, K. & Nychka, D. (1994). Comment on 'Kriging and splines: An empirical comparison of their predictive performance in some applications'. *Journal of the American Statistical Association* **89**, 401–403.

Hansen, M. H. & Kooperberg, C. (2002). Spline adaptation in extended linear models (with discussion and rejoinder). *Statistical Science* **17**, 2–51.

Harville, D. A. (1974). Bayesian inference for variance components using only error contrasts. *Biometrika* **61**, 383–85.

Harville, D. A. (1977). Maximum likelihood approaches to variance component estimation and to related problems. *Journal of the American Statistical Association* **72**, 320–338.

Harville, D. A. (1997). *Matrix Algebra From a Statistician's Perspective.* Springer, New York.

Hastie, T. (1996). Pseudosplines. *Journal of the Royal Statistical Society B* **58**, 379–396.

Hastie, T. & Tibshirani, R. (1990). *Generalized Additive Models.* Chapman and Hall, London.

Hastie, T. & Tibshirani, R. (1993). Varying-coefficient models. *Journal of the Royal Statistical Society B* **55**, 757–796.

Hastie, T. & Tibshirani, R. (2000). Bayesian backfitting (with comments and rejoinder). *Statistical Science* **15**, 196–223.

Henderson, R., Shimakura, S. & Gorst, D. (2002). Modeling spatial variation in leukemia survival data. *Journal of the American Statistical Association* **97**, 965–972.

Hennerfeind, A., Brezger, A. & Fahrmeir, L. (2005). Geoadditive survival models. Under revision for *Journal of the American Statistical Association*.

Hobert, J. P. & Casella, G. (1996). The effect of improper priors on Gibbs sampling in hierarchical linear mixed models. *Journal of the American Statistical Association* **91**, 1461–1473.

Holmes, C. C., & Held, L. (2006). Bayesian auxiliary variable models for binary and multinomial regression. *Bayesian Analysis* **1**, 145–168.

Hummel, M. (2005). *Bayes-Inferenz in Regressionsmodellen mit räumlicher Komponente.* Diploma thesis, University of Munich.

Hurn, M. A., Husby, O. K. & Rue, H. (2003). A tutorial on image analysis. In: J. Møller (ed.): *Spatial Statistics and Computational Methods*. Springer, New York.

Ibrahim, J. G., Chen, M.-H. & Sinha, D. (2001). *Bayesian Survival Analysis*. Springer, New York.

Johnson, M. E., Moore, L. M. & Ylvisaker, D. (1990). Minimax and maximin distance designs. *Journal of Statistical Planning and Inference* **26**, 131–148.

Kalbfleisch, J. D. & Prentice, R. L. (1980). *The Statistical Analysis of Failure Time Data*. Wiley, New York.

Kammann, E. E. & Wand, M. P. (2003). Geoadditive models. *Journal of the Royal Statistical Society C* **52**, 1–18.

Kandala, N. B., Lang, S., Klasen, S. & Fahrmeir, L. (2001). Semiparametric analysis of the socio-demographic and spatial determinants of undernutrition in two African countries. *Research in Official Statistics* **1**, 81–100.

Kauermann, G. (2004). A note on smoothing parameter selection for penalised spline smoothing. *Journal of Statistical Planning and Inference* **127**, 53–69.

Kauermann, G. (2005) Penalised spline fitting in multivariable survival models with varying coefficients. *Computational Statistics and Data Analysis* **49**, 169–186.

Kauermann, G. & Eilers, P. H. C. (2004) Modeling microarray data using a threshold mixture model. *Biometrics* **60**, 376–387.

Keane, M. P. (1994). A computationally practical simulation estimator for panel data. *Econometrica* **62**, 95–116.

Klein, J. P. & Moeschberger, M. L. (2003). *Survival Analysis*. Springer, New York.

Kneib, T. (2005). Geoadditive hazard regression for interval censored survival times. SFB 386 Discussion Paper 447.

Kneib, T. & Fahrmeir, L. (2004). A mixed model approach for structured hazard regression. SFB 386 Discussion Paper 400.

Kneib, T. & Fahrmeir, L. (2005). Structured additive regression for categorical space-time data: A mixed model approach. *Biometrics*, to appear.

Knorr-Held, L. (2004). Simultaneous posterior probability statements from Monte Carlo output. *Journal of Computational and Graphical Statistics* **13**, 20–35.

Komárek, A., Lesaffre, E. & Hilton, J. F. (2005). Accelerated failure time model for arbitrarily censored data with smoothed error distribution. *Journal of Computational and Graphical Statistics* **45**, 726–745.

Komárek, A., Lesaffre, E., Härkänen, T., Declerck, D. and Virtanen, J. I. (2005). A Bayesian analysis of multivariate doubly-interval-censored dental data. *Biostatistics* **6**, 145–155.

Kooperberg, C., Bose, S. & Stone, C. J. (1997). Polychotomous regression. *Journal of the American Statistical Association* **92**, 117–127.

Kooperberg, C., Stone, C. J. & Truong, Y. K. (1995). Hazard regression. *Journal of the American Statistical Association* **90**, 78–94.

Kooperberg, C. & Clarkson, D. B. (1997) Hazard regression with interval-censored data. *Biometrics* **53**, 1485–1494.

Krige, D. G. (1966) Two-dimensional weighted moving average trend surfaces for ore-evaluation. *Journal of the South African Institute of Mining and Metallurgy* **66**, 13–38.

Lang, S. & Brezger, A. (2004). Bayesian P-splines. *Journal of Computational and Graphical Statistics* **13**, 183–212.

Laslett, G. (1994). Kriging and splines: An empirical comparison of their predictive performance in some applications. *Journal of the American Statistical Association* **89**, 391–400.

Li, Y. & Ryan, L. (2002). Modeling spatial survival data using semiparametric frailty models. *Biometrics* **58**, 287–297.

Lin, X. & Breslow, N. E. (1996). Bias correction in generalized linear mixed models with multiple components of dispersion. *Journal of the American Statistical Association* **91**, 1007–1016.

Lin, X. & Zhang, D. (1999). Inference in generalized additive mixed models by using smoothing splines. *Journal of the Royal Statistical Society B* **61**, 381–400.

Marx, B. & Eilers, P. H. C. (1998). Direct generalized additive modeling with penalized likelihood. *Computational Statistics and Data Analysis* **28**, 193–209.

McCullagh, P. & Nelder, J. A. (1989). *Generalized Linear Models*. Chapman & Hall, London.

McCulloch, C. E. & Searle, S. R. (2001). *Generalized, Linear, and Mixed Models*. Wiley, New York.

Müller, H. G., Stadtmüller, U. & Tabnak, F. (1997). Spatial smoothing of geographically aggregated data, with applications to the construction of incidence maps. *Journal of the American Statistical Association* **92**, 61–71.

National Population Commission (Nigeria) and ORC Macro (2004). *Nigeria Demographic and Health Survey 2003*. Calverton, Maryland.

Nelder, J. A. & Wedderburn R. W. M. (1972). Generalized linear models. *Journal of the Royal Statistical Society Series A* **135**, 370–384.

Nychka, D. (2000). Spatial-process estimates as smoothers. In: M. Schimek (ed.): *Smoothing and Regression: Approaches, Computation and Application*. Wiley, New York.

Nychka, D. & Saltzman, N. (1998). *Design of air-quality monitoring networks*. In: Nychka, D., Piegorsch, W. W., Cox, L. H. (eds.): *Case Studies in Environmental Statistics*. Springer, New York.

Nychka, D., Haaland, M., O'Connel, M. & Ellner, S. (1998). *Funfits, data analysis and statistical tools for estimating functions*, In: Nychka, D., Piegorsch, W. W., Cox, L. H. (eds.): *Case Studies in Environmental Statistics*. Springer, New York.

Osuna, L. (2004). *Semiparametric Bayesian Count Data Models*. Dissertation, Dr. Hut Verlag, Munich.

Patterson, H. D. & Thompson, R. (1971). Recovery of inter-block information when block sizes are unequal. *Biometrika* **58**, 545–554.

Ripatti, S. & Palmgren, J. (2000). Estimation of multivariate frailty models using penalized likelihood. *Biometrics* **56**, 1016–1022.

Rue, H. (2001). Fast sampling of Gaussian Markov random fields with applications. *Journal of the Royal Statistical Society B* **63**, 325–338.

Rue, H. & Held, L. (2005). *Gaussian Markov Random Fields. Theory and Applications*. CRC / Chapman & Hall, London.

Rue, H. & Tjelmeland, H. (2002). Fitting Gaussian Markov random fields to Gaussian fields. *Scandinavian Journal of Statistics* **29**, 31–49.

Ruppert, D., Wand, M. P. & Carroll, R. J. (2003). *Semiparametric Regression*. University Press, Cambridge.

Schmid, V. (2004). *Bayesianische Raum-Zeit-Modellierung in der Epidemiologie*. Dissertation, Dr. Hut Verlag, Munich.

Stein, M. L. (1999). *Interpolation of Spatial Data. Some Theory for Kriging*. Springer, New York.

Stone, C. J., Hansen, M. H., Kooperberg, C. & Truong, Y. K. (1997). Polynomial splines and their tensor products in extended linear modeling (with discussion). *Annals of Statistics* **25**, 1371–1470.

Stram, D. O. & Lee, J. W. (1994). Variance components testing in the longitudinal mixed effects model. *Biometrics* **50**, 1171–1177.

Stram, D. O. & Lee, J. W. (1995). Correction to "Variance components testing in the longitudinal mixed effects model". *Biometrics* **51**, 1196.

Sun, D., Tsutakawa, R. K. & He, Z. (2001). Propriety of posteriors with improper priors in hierarchical linear mixed models. *Statistica Sinica* **11**, 77–95.

Terzopoulos, D. (1988). The computation of visible-surface representations. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **10**, 417–438.

Therneau, T. M. & Grambsch, P. M. (2000). *Modeling Survival Data: Extending the Cox Model*. Springer, New York.

Tierney, L. & Kadane, J. B. (1986). Accurate approximations for posterior moments and marginal densities. *Journal of the American Statistical Association* **81**, 82–86.

Train, K. E. (2003). *Discrete Choice Methods With Simulations*. University Press, Cambridge.

Tutz, G. (2003). Generalized semiparametrically structured ordinal models. *Biometrics* **59**, 263–273.

Tutz, G. & Scholz, T. (2004). Semiparametric modelling of multicategorical data. *Journal of Statistical Computation and Simulation* **74**, 183–200.

Unser, M., Aldroubi, A. & Eden, M. (1992). On the asymptotic convergence of B-spline wavelets to Gabor functions. *IEEE Transactions on Information Theory* **38**, 864–872.

Verbeke, G. & Molenberghs, G. (2000). *Linear Mixed Models for Longitudinal Data*. Springer, New York.

Verbeke, G. & Molenberghs, G. (2003). The use of score tests for inference on variance components. *Biometrics* **59**, 254–262.

Wahba, G. (1978). Improper prior, spline smoothing and the problem of guarding against model errors in regression. *Journal of the Royal Statistical Society B* **40**, 364–372.

Wand, M. P. (2003) Smoothing and mixed models. *Computational Statistics* **18**, 223–249.

Yau, P., Kohn, R. & Wood, S. (2003). Bayesian variable selection and model averaging in high dimensional multinomial nonparametric regression. *Journal of Computational and Graphical Statistics* **12**, 23–54.

Zhang, D. (2004). Generalized linear mixed models with varying coefficients for longitudinal data. *Biometrics* **60**, 8–15.

Ziegler, A. & Eymann, A. (2001). Zur Simulated Maximum-Likelihood-Schätzung von Mehrperioden-Mehralternativen-Probitmodellen. *Allgemeines Statistisches Archiv* **85**, 319–342.

Zimmermann, D. (1993). Another look at anisotropy in geostatistics. *Mathematical Geology* **25**, 453–470.

Lebenslauf

Thomas Kneib

geboren am 4.12.1976 in Bad Sobernheim

Schulbildung:

- 1983 – 1987 Grundschule Bad Sobernheim
- 1987 – 1996 Emanuel-Felke-Gymnasium Bad Sobernheim
- Juni 1996 Abitur

Zivildienst:

- September 1996 – September 1997
- Krankenhaus der Anhaltischen Diakonissenanstalt Dessau
  Urologische und plastisch-chirurgische Station

Praktikum:

- Oktober 1997 – Februar 1998
- Statistisches Landesamt Sachsen-Anhalt in Halle (Saale)

Studium:

- Sommersemester 1998 – Wintersemester 2002/03
- Diplom-Studiengang Statistik an der Ludwig-Maximilians-Universität München
- Vordiplom: Sommersemester 2000
- Anwendungsgebiete: Politische Wissenschaften, Soziologie
- Vertiefungsgebiet im Hauptstudium: Mathematische Stochastik
- Diplomarbeit: Bayes-Inferenz in generalisierten geoadditiven gemischten Modellen

Studienbegleitende Tätigkeiten:

- Januar bis März 2000 studentische Hilfskraft im Statistischen Beratungslabor der Ludwig-Maximilians-Universität München
- Februar 2001 bis Februar 2003 studentische Hilfskraft am Sonderforschungsbereich 386 (Statistische Analyse diskreter Strukturen)

Beruflicher Werdegang:

- März bis Mai 2003 wissenschaftliche Hilfskraft am Sonderforschungsbereich 386
- Juni bis Juli 2003 wissenschaftlicher Mitarbeiter am Institut für Statistik der LMU (Krankheitsvertretung für Volker Schmid)
- August bis Oktober 2003 wissenschaftliche Hilfskraft am Sonderforschungsbereich 386
- November bis Dezember 2003 Mitarbeiter im Statistischen Beratungslabor
- Seit Januar 2004 wissenschaftlicher Mitarbeiter am Institut für Statistik und dem Sonderforschungsbereich 386