

# Statistical Analysis of Discontinuous Phenomena with Potts Functionals

DISSERTATION

zur Erlangung des akademischen Grades  
eines Doktors der Naturwissenschaften  
des Fachbereichs Mathematik  
der Ludwig-Maximilians-Universität München

vorgelegt von

Angela Kempe

30. Januar 2004

1. Berichtstatter: Prof. Dr. G. Winkler
  2. Berichtstatter: Prof. Dr. P. L. Davies
- Tag der mündlichen Prüfung: 25. Juni 2004

Für meine Oma



# Preface

This thesis was carried out under the supervision of Prof. Dr. Gerhard Winkler. From 01.01.2000 to 31.12.2002 it was funded by the GSF-National Research Center for Environment and Health, Neuherberg, and from 01.01.03 to 31.01.2004 by the Graduate Program Applied Algorithmic Mathematics of the Deutsche Forschungsgemeinschaft at the Munich University of Technology.

The simulations were realized on the software package ANTSINFIELDS developed by F. FRIEDRICH (2003). It was also used for most of the illustrations.

I thank the members of the GSF-Institute of Biomathematics and Biometry for their support. I am particularly indebted to Gerhard Winkler, Olaf Wittich, Volkmar Liebscher, and Felix Friedrich for their sedulous help and encouragement. Many thanks to my friends and my mother.

Munich, January 2004

Angela Kempe



# Contents

<b>Preface</b>	<b>i</b>
<b>Zusammenfassung</b>	<b>vii</b>
<b>Introduction</b>	<b>1</b>
<b>I Analysis of Potts Functionals and their Minimizers</b>	<b>5</b>
<b>1 Segmentations and Potts Functionals</b>	<b>9</b>
1.1 Potts Functionals . . . . .	9
1.2 Segmentations . . . . .	11
1.3 Relation to Bayesian Approaches . . . . .	15
<b>2 MAP Estimators</b>	<b>19</b>
2.1 Splitting the Minimization . . . . .	19
2.2 Existence . . . . .	22
2.3 Uniqueness . . . . .	24
2.4 Dependence on Hyperparameters . . . . .	28
2.5 Continuity . . . . .	36
2.6 Measurable Section . . . . .	40
<b>3 Exact Optimization</b>	<b>41</b>
3.1 Minimization for Fixed Hyperparameter . . . . .	42
3.2 Simultaneous Minimization in $\gamma$ . . . . .	45
<b>II Choice of Hyperparameters</b>	<b>49</b>
<b>4 Equivariance and Hyperparameters</b>	<b>53</b>
4.1 Equivariance . . . . .	53

4.2	Normalization of Data . . . . .	58
<b>5</b>	<b>Interval Criteria</b>	<b>63</b>
5.1	Invariant Attributes . . . . .	63
5.2	$F$ -Longest Interval Criteria . . . . .	64
5.3	Longest Interval Criterion . . . . .	72
<b>6</b>	<b>Stopping Criteria</b>	<b>81</b>
6.1	Longest Run Criterion . . . . .	81
6.2	Multiresolution Criterion . . . . .	83
<b>7</b>	<b>Model Selection Criteria</b>	<b>85</b>
7.1	The Akaike Information Criterion . . . . .	86
7.2	The Schwarz Information Criterion . . . . .	87
7.3	Equivariant Versions . . . . .	87
<b>8</b>	<b>Further Ideas</b>	<b>89</b>
8.1	Iterative Procedures . . . . .	89
8.2	Constant Estimates . . . . .	91
8.3	Morphological Criteria . . . . .	91
<b>III</b>	<b>Application to Data</b>	<b>93</b>
<b>9</b>	<b>Data Sets from Life Sciences</b>	<b>97</b>
9.1	Functional Magnetic Resonance Imaging . . . . .	97
9.2	Fractionation Experiments . . . . .	103
<b>10</b>	<b>Simulations</b>	<b>111</b>
10.1	Constant Signal . . . . .	113
10.2	One Jump Signal . . . . .	117
10.3	Boxcar Shaped Signal . . . . .	118
<b>IV</b>	<b>Consistency</b>	<b>127</b>
<b>11</b>	<b>Fixed Data Length</b>	<b>131</b>
<b>12</b>	<b>Increasing Data Length</b>	<b>133</b>
12.1	Potts Functionals on $L^2([0, 1])$ . . . . .	135
12.2	Epi-Convergence and Relative Compactness . . . . .	138
12.3	Proof of the Main Theorem . . . . .	150

Discussion and Outlook	153
<b>V Appendix</b>	<b>155</b>
<b>A <math>\gamma</math>-Scanning</b>	<b>159</b>
<b>B Model Selection Criteria</b>	<b>165</b>
B.1 A Family of Regression Models . . . . .	165
B.2 The Akaike Information Criterion . . . . .	168
B.3 The Schwarz Information Criterion . . . . .	171
<b>C Calculations Model Selection Criteria</b>	<b>177</b>
C.1 Proofs for AIC . . . . .	177
C.2 Regularity Conditions Likelihood Function . . . . .	179
C.3 The Hessian Matrix of the Log Likelihood Function . . . . .	182
<b>Symbols</b>	<b>185</b>



# Zusammenfassung

Das Hauptziel der vorliegenden Dissertationsschrift ist eine eingehende und möglichst umfassende Untersuchung der *Potts-Funktionale*

$$\begin{aligned} \bar{H}_\gamma(\cdot, y) : \mathbb{R}^N &\longrightarrow \mathbb{R}, \\ x &\longmapsto \gamma \cdot |\{1 \leq i \leq N - 1 : x_i \neq x_{i+1}\}| + \sum_{i=1}^N (y_i - x_i)^2, \quad \gamma > 0, \end{aligned} \tag{1}$$

für gegebenes  $y \in \mathbb{R}^N$  und ihrer Minimierer. Die Zeitreihen  $y$  werden interpretiert als Daten, und  $x$  ist ein Kandidat für eine geeignete Repräsentation der Daten. Der erste Term des Funktionals bestraft Sprünge von  $x$ , der zweite Term misst seine Datentreue. Die Minimierung des Funktionals (1) in  $x$  bedeutet also, eine Darstellung der Daten zu finden, die diese beiden konkurrierenden Forderungen bestmöglich erfüllt.

Im Potts-Funktional sind die wesentlichen Eigenschaften eines Signals Glattheit und Sprünge. Ein Signal  $x$  hat einen Sprung, wenn die Intensitäten in aufeinanderfolgenden Zeitpunkten verschieden sind. Der erste Term bestraft also Sprünge unabhängig von ihrer Höhe. Der dazu gehörende komplementäre Glattheitsbegriff ist beim Potts-Funktional sehr streng: Nur konstante Signale sind glatt. Der Parameter  $\gamma$  bestimmt den Grad der Glattheit. Ein Maß für die Datentreue ist in (1) die Summe der Abstandsquadrate. Hier sind natürlich auch andere Abstandsbegriffe denkbar, wir beschränken uns aber auf diesen Datenterm, da hier eine ganze Reihe von Aussagen bewiesen werden kann.

Aufgrund ihrer Einfachheit sind Potts-Funktionale in Situationen geeignet, in denen man nichts oder nur wenig über die genaue Erzeugung der Daten oder der zugrundeliegenden Ausgangssignale weiß. Minimierer der Potts-Funktionale sind also ein geeignetes Instrument zur Extraktion von einfachen Features wie Sprüngen und Plateaus. Wir zeigen dies an zwei Datenbeispielen: fMRT-Daten zur Gehirnkartierung und Fraktionierungskurven aus Experimenten für Microarrays. Bei diesen Daten erscheint es sinnvoll, die

Zeitreihen durch im Sinne des Potts-Funktionalen glatte Signale mit möglichst wenigen Sprüngen darzustellen.

Im ersten Teil der Arbeit werden die Potts-Funktionale eingeführt und Eigenschaften ihrer Minimierer  $x^*(\gamma, y)$  untersucht. Wir zeigen die Existenz und Eindeutigkeit von Minimierern für fast alle  $y$ . Außerdem stellt sich heraus, dass es für fast alle Daten  $y$  nur endlich viele verschiedene Minimierer gibt, wenn der Parameter  $\gamma$  den gesamten Bereich von Unendlich bis Null durchläuft. Wir zeigen die gemeinsame Stetigkeit von Minimierern  $x^*(\gamma, y)$  in  $\gamma$  und  $y$ . Schließlich leiten wir noch exakte Algorithmen zur Berechnung der Minimierer her.

Im zweiten Teil beschäftigen wir uns mit der Wahl des Parameters  $\gamma$ , was als Modellwahl aufgefasst werden kann. Wir interpretieren Minimierer der Potts-Funktionale als Schätzer für die Daten  $y$ . Zuerst gehen wir auf eine Minimalforderung an einen vernünftigen Schätzer ein, seine Äquivarianz bezüglich Skalierung des Signals und einer Verschiebung des Nullpunkts. Für festen Parameter  $\gamma$  sind die Minimierer nicht äquivariant. Aus der Skalierungseigenschaft der Minimierer leiten wir eine hinreichende Bedingung für Äquivarianz her. Wir stellen eine neue Klasse datenadaptierter Parameterwahlen vor. Diese Intervallkriterien beruhen darauf, dass der Minimierer für ganze Parameterintervalle derselbe ist, und die Länge dieser Intervalle als ein Maß für Stabilität des zugehörigen Minimierers interpretiert werden kann. Ferner zeigen wir, wie einige bekannte Stoppkriterien und Modellwahlkriterien auch als datenadaptierte Parameterwahl für Schätzer aus dem Potts-Funktional genutzt werden können. Schließlich skizzieren wir weitere Ideen zur Parameterwahl, die vor allem die Probleme der Intervallkriterien beheben sollen. Im Teil III wenden wir die beschriebenen Methoden auf Zeitreihen aus den erwähnten realen Datensätzen und simulierte Daten an.

Der letzte Teil befasst sich mit der Konsistenz der Schätzer. Wir untersuchen das asymptotische Verhalten von Minimierern des Potts-Funktionalen in zwei Situationen. Zunächst nehmen wir an, dass die Daten von einem einfachen Regressionsmodell erzeugt worden sind. Wir zeigen, dass Minimierer gegen eine geglättete Version des Ausgangssignals konvergieren, wenn das Rauschen gegen Null geht. Im zweiten Szenario betrachten wir die Daten als verrauschte Diskretisierung einer Funktion  $f$ . Bei Verfeinerung der Diskretisierung und geeigneter Skalierung des Parameters konvergieren die Minimierer des Potts-Funktionalen, wenn man sie mit Treppenfunktionen identifiziert, gegen die Funktion  $f$ .

Im Anhang findet man exemplarisch für eine Zeitreihe Bilder aller Minimierer bei Variation von  $\gamma$ . Außerdem beinhaltet er eine kurze Zusammenfassung der Modellwahlkriterien und einfache, aber langatmige Rechnungen.

# Introduction

The main aim of the present thesis is a detailed and rigorous analysis of the functional

$$\begin{aligned} \bar{H}_\gamma(\cdot, y) : \mathbb{R}^N &\longrightarrow \mathbb{R}, \\ x &\longmapsto \gamma \cdot |\{1 \leq i \leq N - 1 : x_i \neq x_{i+1}\}| + \sum_{i=1}^N (y_i - x_i)^2, \quad \gamma > 0, \end{aligned} \tag{2}$$

given  $y \in \mathbb{R}^N$ , and its minimizers. The time series  $y$  is interpreted as data, and  $x$  as a candidate for an appropriate ‘interpretation’ or ‘representation’ of data. The first term penalizes roughness of  $x$ , and the second one measures fidelity to data. Minimizing the functionals (2) in  $x$  results in particular in a tradeoff between these two competing terms. For historical reasons, the functionals in (2) will be called *Potts functionals*.

These functionals are a simple instance of variational approaches. The general idea is to design a functional of signals and data which captures and rates the relevant signal features and, simultaneously, fidelity to data. Then - given a special set of data - one selects a signal according to some rule which justifies to consider it as the desired ‘reconstruction’, ‘filter output’, or ‘representation’. The functionals in variational approaches typically consist of a penalty and a data term, and are formally similar to posterior or penalized likelihood functions.

In the present analysis of the Potts functionals the relevant features are smoothness and jumps. The general philosophy, however, applies to many situations where both, a notion of homogeneity and a notion of heterogeneity, are present. There are two basic concepts inherent in the Potts functionals:

- (i) ‘Jump’ or ‘break’: In the Potts functional a jump is present, where the values of the signal  $x$  in two subsequent time points differ from each other. The first term penalizes the number of jumps irrespectively of their size.

The complementary notion of ‘smoothness’: This concerns the behavior of the signal between two subsequent jumps. It is a consequence of the penalty that the notion of smoothness is rather strict in the Potts functional: Only constant signals are smooth. Note that the parameter  $\gamma > 0$  controls the degree of smoothness.

- (ii) Fidelity to data, i.e. some measure of distance between data  $y$  and the signal  $x$ ; in (2) it is the sum of squares.

Although the above general reflections apply with suitable modifications to a broader class of functionals, we restrict ourselves to the analysis of (2).

Due to its simplicity, the Potts functional is appropriate in situations where there is little or no ground truth. Its minimizers are a suitable tool for the extraction of primitive signal features like plateaus and jumps. We will illustrate this by two selected data sets from life sciences: time series’ from fMRI human brain mapping and from fractionation experiments for cDNA microarrays. In view of these data sets, one may doubt about too ‘specific’ methods or too detailed models for their analysis. Estimates rely sensitively on assumptions and therefore are not robust against even slight changes of models. Fitting too many parameters in a specific model introduces more variance despite slight decrease in bias. A way out of this misery is to try a parsimonious approach. The principle of parsimony is a philosophical matter and will not be addressed here; let us sloppily interpret it as reduction to essential features.

In these data examples, it is reasonable to represent time series’ by signals with only a few jumps and smooth in the sense of Potts functionals. Therefore, Potts functionals and their minimizers seem to be appropriate in these situations.

In its original form the *Potts model* was introduced in R.B. POTTS (1952) as a generalization of the well-known Ising model from E. ISING (1925) for binary spin systems to a finite number of states. It is a Gibbs field of the form  $\exp(-K(x))/Z$  on a discrete lattice where  $K(x)$  simply counts neighbor pairs with different values. Since the penalty in (2) corresponds to a Potts prior in a discrete Bayesian model, we call (2) a *Potts functional*. Whereas the original Potts model lives on a finite discrete space, we work in Euclidean spaces.

The Potts functionals can also be interpreted as a degenerate case of other classical functionals. Let us only mention an example which is both, a specialization of the explicit edge model in the seminal paper S. GEMAN and D. GEMAN (1984) on Bayesian image analysis, and a reformulation of

the elasticity model from the article A. BLAKE (1983) and the monograph A. BLAKE and A. ZISSERMAN (1987). For  $y \in \mathbb{R}^N$  this functional is given by

$$BZ(\cdot, y) : \mathbb{R}^N \longrightarrow \mathbb{R}, \quad x \longmapsto \sum_{i=1}^{N-1} \varphi(x_{i+1} - x_i) + \sum_{i=1}^N (y_i - x_i)^2,$$

with the truncated square function

$$\varphi(x) = \varphi_{\mu, \gamma}(x) = \min\left\{\frac{x^2}{\mu^2}, \gamma\right\}, \quad \gamma > 0, \mu > 0.$$

Here a difference  $|x_{k+1} - x_k|$  is considered to be a jump if it is greater than  $\delta = \gamma^{1/2}\mu$ . This functional smoothes  $x$  between two subsequent jump locations in the  $L^2$ -sense. The functional converges pointwise to the Potts functional as  $\delta \rightarrow 0$ .

In the first part of this thesis, we introduce Potts functionals and analyze their minimizers  $x^*(\gamma, y)$ . We prove existence of minimizers and their uniqueness for almost all  $y$ . Moreover, we show that for almost all  $y$  there are only finitely many different minimizers as the hyperparameter  $\gamma$  varies from infinity to zero. We prove joint continuity of minimizers  $x^*(\gamma, y)$  in  $y$  and  $\gamma$ . We show further the existence of a measurable section of the set of all minimizers. Finally, we derive exact optimization algorithms.

Part II is concerned with model choice, which amounts to the choice of the hyperparameter  $\gamma$ . This problem is ubiquitous in nonparametric statistics. In case of the Potts functionals, driven by the results on dependence of minimizers on the hyperparameter and its continuity, there is a chance to treat this question rigorously. Another advantage of this simple functional is that the estimates can be computed *exactly* for all values of  $\gamma$ . We consider equivariance with respect to certain group actions. This is a minimal requirement on estimators. We establish a scaling property of minimizers and derive a sufficient condition for equivariance.

We present a special class of data adapted parameter choices using the  $\gamma$ -intervals on which minimizers of the Potts functionals do not change. We show how some stopping criteria and model selection criteria can be interpreted as data adapted parameter choices for the Potts functional. Finally, we sketch some further ideas for the choice of  $\gamma$ , especially to overcome certain problems of interval criteria.

In Part III, we apply these methods to two data sets from life sciences and to simulated data.

The last part deals with consistency of estimators. We study the asymptotic behavior of minimizers of the Potts functionals in two scenarios. First, we assume that data are generated from linear regression models. We show that minimizers converge to a smoothed version of the signal if noise tends to zero. In the second scenario, data are sampled from some function  $f$  and corrupted by noise. We prove that minimizers - identified with step functions - converge to  $f$  for increasing sampling rate.

The Appendix contains exemplarily plots of minimizers for all hyperparameters for one time series. In addition, we give a brief summary of model selection criteria, and collect straightforward, but tedious, calculations.

**Part I**

**Analysis of Potts Functionals  
and their Minimizers**



This part is concerned with Potts functionals and the analysis of their minimizers  $x^*(\gamma, y)$ . In Chapter 1, we introduce the Potts functionals which will play the role of statistical models. Their minimizers will serve as estimators, and as a tool for the extraction of characteristic features from data. In Chapter 2, we investigate properties of the minimizers of the Potts functionals. We prove their existence and uniqueness. Moreover, we show that for almost all  $y$  there are only finitely many different MAP estimators when the hyperparameter  $\gamma$  is varied from infinity to zero. In particular, the minimizer is the same for all values of the hyperparameter in intervals. This observation plays an important role in the proof of the joint continuity of minimizers  $x^*(\gamma, y)$  in  $\gamma$  and  $y$ . The existence of a measurable section of the set-valued map  $(\gamma, y) \mapsto X^*(\gamma, y)$  is a property of own interest. In addition, it is a crucial property of minimizers for consistency, and a given measurable section provides a unique minimizer. Chapter 3 completes this part with exact algorithms for the computation of the minimizers. They eliminate all the uncertainties of popular Markov Chain Monte Carlo methods like Simulated Annealing. This enables us to study the Potts functional in detail and rigor. Only with exact optimization, we can tell artefacts and effects due to modelling from those caused by suboptimal optimization. This is indispensable for a rigorous validation of methods.



# Chapter 1

## Signals, Segmentations and Potts Functionals

In this first chapter, we introduce Potts functionals. These functionals and their minimizers are the central objects of this thesis. Potts functionals will play the role of statistical models, and their minimizers will serve as estimators, and as a tool for the extraction of characteristic features from data. In Section 1.1, we introduce basic notions and notations. We give the definition of the Potts functional and its ‘maximum posterior estimator’. In Section 1.2, we introduce segmentations. We identify signals with minimal segmentations, and rewrite the one dimensional Potts functional in terms of such segmentations. Finally, we discuss briefly the relationship of the Potts functional to the formally identical posterior log likelihood functions of classical Bayesian models.

### 1.1 Potts Functionals

In this section, we first introduce some notions and notation. In particular, we define signals and data, and then the Potts functionals and the associated MAP estimator.

Let  $S$  be a finite set of design points or *sites*. In the one dimensional case, the sites frequently correspond to *time points*. At the present state we do not distinguish between time series’ and multidimensional data sets like images. We endow  $S$  with a neighborhood structure, induced by a graph, the nodes of which are the sites. Recall that a *simple graph* is a graph without loops in which at most one edge connects any two vertices. Throughout this text the graph will be *undirected*.

**Definition 1.1.1** *Let an undirected simple graph structure with edges in  $S$*

be given. Sites  $s$  and  $t$  in  $S$  are called **neighbors** or **neighbor sites** in  $S$  if they are connected by an edge. This will be indicated by the symbol  $s \sim t$ . We will say that  $S$  is endowed with a **neighbor structure**.

Now we specify signals and data.

**Definition 1.1.2** A **signal**  $x$  is a family  $(x_s)_{s \in S}$  of **intensities**  $x_s \in \mathbb{R}$ . Similarly, we define **data**  $y = (y_s)_{s \in S}$  with single **observations**  $y_s \in \mathbb{R}$ . The space  $\{(x_s)_{s \in S} : x_s \in \mathbb{R}\}$  will be denoted by  $\mathbb{X}$ .

Both, signals  $x$  and data  $y$ , are elements of  $\mathbb{R}^S$ . The notion of a jump in a signal will play an important role. It can be made precise in various ways, cf. G. WINKLER et al. (2004). The following definition is the only one compatible with the Potts functionals. It is particularly simple.

**Definition 1.1.3** A pair of neighbors  $s \sim t$  in  $S$  is called a **jump** of  $x$  if  $x_s \neq x_t$ . The set  $J(x) = \{\{s, t\} \in S \times S : s \sim t, x_s \neq x_t\}$  is called the **jump set** of  $x$ .

For a finite set  $A$ , the symbol  $|A|$  will denote the number of elements of  $A$ . In particular,  $|J(x)|$  denotes the number of jumps of  $x$ .

Now we define the Potts functionals which assign a real value to each pair of a signal  $x$  and data  $y$ . Let us agree that low values of the functional correspond to desired pairs  $(x, y)$ .

**Definition 1.1.4** Let  $D$  be a real functional on  $\mathbb{X} \times \mathbb{X}$ , and  $\gamma > 0$ . The **Potts functional** with **hyperparameter**  $\gamma$  is given by

$$H_\gamma : \mathbb{X} \times \mathbb{X} \longrightarrow \mathbb{R}, \quad (x, y) \longmapsto \gamma \cdot |J(x)| + D(x, y). \quad (1.1)$$

The term

$$\gamma \cdot |J(x)| \quad (1.2)$$

is called the **Potts penalty term** and  $D(x, y)$  is called the **data term**.

If there is no danger of confusion we will skip the subscript  $\gamma$  and write  $H(x, y)$  instead of  $H_\gamma(x, y)$ . The data term  $D$  measures fidelity of a signal  $x$  to data  $y$ . In contrast, the Potts penalty term prefers signals with as few jumps as possible. Obviously, there is a tradeoff between ‘smoothness’ and fidelity to data. A ‘best’ pair  $(x, y)$  is given by a solution of the following minimization problem: given  $y \in \mathbb{X}$ ,

$$\text{minimize } x \longmapsto H_\gamma(x, y) \quad \text{in } x \in \mathbb{X}. \quad (1.3)$$

The minimizers will play the role of statistical estimators.

**Definition 1.1.5** For each data set  $y \in \mathbb{X}$  let

$$X^*(\gamma, y) = \{x^* \in \mathbb{X} : H_\gamma(x^*, y) = \min_{x \in \mathbb{X}} H_\gamma(x, y)\}$$

denote the set of minimizers of  $x \mapsto H_\gamma(x, y)$  with the Potts functional from (1.1). We call elements  $x^* = x^*(\gamma, y)$  of  $X^*(\gamma, y)$  the **maximum posterior (MAP) estimates** for the Potts functional.

In general,  $X^*(\gamma, y)$  is not a singleton. On the other hand, in important special cases MAP estimates are unique, as we shall show in Section 2.3.

The functional  $H$  in (1.1) is formally equal to the log likelihood function of a (Bayesian) posterior distribution of exponential form. For a brief introduction to the Bayesian approach of signal analysis see Section 1.3. A ‘prior’ distribution of the form  $c \cdot \exp(-K)$  will turn out to be improper if  $K$  is invariant under diagonal translations. Although this holds in particular for  $K(x) = \gamma \cdot |J(x)|$ , we adopt nomenclature from Bayesian statistics and speak of ‘maximum posterior’ estimators.

## 1.2 Segmentations

We are interested in locations of abrupt changes in a signal, and in regions of ‘smoothness’ inbetween. This leads in a natural way to the concept of segmentations. They are the natural setting for the Potts functionals. They also will allow to reduce the continuous minimization problem (1.3) to a discrete one.

A segmentation partitions the sites into regions where the signal  $x$  is ‘smooth’; across the boundaries of these regions the signal has abrupt changes or it ‘jumps’. In continuous time the notion of smoothness usually is made precise by the choice of a function space like a Sobolev space  $W^{p,k}$ . In our discrete-time setting, smoothness may be defined in terms of discrete derivatives. The simplest definition - and in fact the only one compatible with the Potts penalty - is that a signal is smooth, where it is constant. Therefore, we define a segmentation of a signal as a partition of the sites into intervals and the intensities on these intervals. Then we identify a signal with a minimal segmentation and rewrite the Potts functional in terms of such segmentations. Let us start with some definitions. A subset  $I$  of sites in  $S$  is called *connected* if for all different sites  $s, t \in I \subseteq S$  there is a sequence  $s = s_0 \sim s_1 \sim \dots \sim s_n = t$  of neighbors in  $I$ . Given a signal  $x = (x_s)_{s \in S}$ , we can decompose  $S$  into connected sets  $I$  on which  $x_s = x_t$  for all  $s, t \in I$ . This provides a decomposition of  $S$  into sets of constant intensity.

**Definition 1.2.1** Let  $S$  be a finite set endowed with a neighbor structure. A connected subset  $I$  of  $S$  will be called an **interval**. A set  $\mathcal{P} = \{I_1, \dots, I_k\}$ ,  $k \in \mathbb{N}$ , of mutually disjoint intervals  $I_1, \dots, I_k \subseteq S$  with  $\cup_{j=1}^k I_j = S$  is called a **partition** of  $S$ . The set of all these partitions will be denoted by  $\mathfrak{P}$ .

Given a partition  $\mathcal{P} \in \mathfrak{P}$  and a site  $s \in S$ , the uniquely determined interval in  $\mathcal{P}$  containing  $s$  will be denoted by  $I(s)$ . The number  $|I|$  will be called the *length* of  $I$ . Two disjoint intervals  $I, J \in \mathcal{P}$  will be called *neighbor intervals* if there are  $s \in I$  and  $t \in J$  with  $s \sim t$ . Furthermore, we will write  $I \sim J$  if  $I$  and  $J$  are neighbors.

**Definition 1.2.2** A pair  $(\mathcal{P}, \mu_{\mathcal{P}})$  of a partition  $\mathcal{P} \in \mathfrak{P}$  and a family  $\mu_{\mathcal{P}} \in \mathbb{R}^{\mathcal{P}}$  of real values will be called a **segmentation**. The set of all segmentations will be denoted by  $\mathfrak{S}$ . A segmentation will be called **minimal** if  $\mu_I \neq \mu_J$  for neighboring intervals  $I \sim J$  in  $\mathcal{P}$ . The space of all minimal segmentations will be denoted by  $\mathfrak{M}$ .

The set of all segmentations can be written as

$$\mathfrak{S} = \bigcup_{\mathcal{P} \in \mathfrak{P}} \{\mathcal{P}\} \times \mathbb{R}^{\mathcal{P}}.$$

Let now a segmentation  $(\mathcal{P}, \mu_{\mathcal{P}}) \in \mathfrak{S}$  be given. It uniquely defines a signal  $x \in \mathbb{X}$  by  $x_s = \mu_{I(s)}$ . Conversely, let  $x \in \mathbb{X}$  be a signal. Taking as intervals *arbitrary* connected sets of constant intensity of  $x$  gives a partition  $\mathcal{P} \in \mathfrak{P}$ . In general, such a partition is not unique, whereas the maximal intervals of constant intensity determine a unique minimal partition  $\mathcal{P}$ .

**Definition 1.2.3** Let  $x \in \mathbb{X}$  be a signal. The partition given by the maximal intervals of constant intensity of  $x$  is called the **partition induced by the signal**  $x$  and will be denoted by  $\mathcal{P}(x)$ . Denote further for each  $I \in \mathcal{P}(x)$  the constant value of  $x$  on  $I$  by  $\mu_I(x)$  and let  $\mu_{\mathcal{P}(x)}(x) = (\mu_I(x))_{I \in \mathcal{P}(x)}$ . The segmentation  $(\mathcal{P}(x), \mu_{\mathcal{P}(x)}(x))$  is called the **segmentation induced by the signal**  $x$ .

By definition, the segmentation  $(\mathcal{P}(x), \mu_{\mathcal{P}(x)}(x))$  induced by  $x$  is minimal. We will write  $\mathcal{P}$  for  $\mathcal{P}(x)$  and  $\mu_{\mathcal{P}}$  for  $\mu_{\mathcal{P}(x)}(x)$  if there is no danger of confusion. In view of the preceding remarks, we can summarize:

**Theorem 1.2.4** *The map*

$$\Sigma : \mathbb{X} \longrightarrow \mathfrak{M}, \quad x \longmapsto (\mathcal{P}(x), \mu_{\mathcal{P}(x)}(x))$$

*is one-to-one and onto.*

The Potts functionals will now be rewritten in terms of segmentations. We start with the data term  $D(x, y)$ . We will restrict ourselves to data terms of the form

$$D(x, y) = \sum_{s \in S} \rho(y_s - x_s)$$

with a function  $\rho : \mathbb{R} \rightarrow \mathbb{R}$ . Throughout the text we require the following natural conditions to hold:

**Hypothesis 1.2.5** *Assume that  $\rho$  fulfills the following conditions:*

(H1)  $\rho$  is symmetric around zero,

(H2)  $\rho(0) = 0$ ,

(H3)  $\rho(u)$  increases in  $u \geq 0$ .

Note, that under these conditions the function  $\rho$  is nonnegative. Standard examples are  $\rho(u) = u^2$  and  $\rho(u) = |u|$ .

**Lemma 1.2.6** *Let  $y \in \mathbb{X}$ . Suppose that  $(\mathcal{P}, \mu_{\mathcal{P}}) \in \mathfrak{M}$  is induced by  $x \in \mathbb{X}$ . Then*

$$D(x, y) = \sum_{s \in S} \rho(y_s - x_s) = \sum_{I \in \mathcal{P}} \sum_{s \in I} \rho(y_s - \mu_I) =: D((\mathcal{P}, \mu_{\mathcal{P}}), y). \quad (1.4)$$

**Proof** Since  $x_s = \mu_{I(s)}$ , a rearrangement of the terms gives

$$\sum_{s \in S} \rho(y_s - x_s) = \sum_{I \in \mathcal{P}} \sum_{s \in I} \rho(y_s - \mu_{I(s)}) = \sum_{I \in \mathcal{P}} \sum_{s \in I} \rho(y_s - \mu_I)$$

which is  $D((\mathcal{P}, \mu_{\mathcal{P}}), y)$  in (1.4).  $\square$

Most parts of the text will deal with one dimensional signals or *time series*'. The set  $S = \{1, \dots, N\}$  will be endowed with the *nearest neighbor structure* defined by  $s \sim s + 1$  for  $s = 1, \dots, N - 1$ . In this case one has  $|J(x)| = |\mathcal{P}(x)| - 1$  and the functionals in (1.1) will be called *one dimensional Potts functionals*. Let us rewrite them in terms of segmentations.

**Proposition 1.2.7** *Let  $S = \{1, \dots, N\}$  be endowed with the nearest neighbor structure. Then the functional*

$$\tilde{H} : \mathfrak{M} \times \mathbb{X} \longrightarrow \mathbb{R}, ((\mathcal{P}, \mu_{\mathcal{P}}), y) \longmapsto \gamma \cdot (|\mathcal{P}| - 1) + \sum_{I \in \mathcal{P}} \sum_{s \in I} \rho(y_s - \mu_I) \quad (1.5)$$

is connected to the one dimensional Potts functional

$$H_\gamma : \mathbb{X} \times \mathbb{X} \longrightarrow \mathbb{R}, \quad (x, y) \longmapsto \gamma \cdot |J(x)| + \sum_{s \in S} \rho(y_s - x_s)$$

by the identity

$$\tilde{H}_\gamma((\mathcal{P}, \mu_{\mathcal{P}}), y) = H_\gamma(x, y) \tag{1.6}$$

where  $x = \Sigma^{-1}(\mathcal{P}, \mu_{\mathcal{P}})$  with  $\Sigma$  from Theorem 1.2.4.

**Proof** By Lemma 1.2.6, and since in one dimension the Potts penalty has the form  $\gamma \cdot (|\mathcal{P}(x)| - 1)$ , each value of the Potts functional can be written as

$$\begin{aligned} H_\gamma(x, y) &= \gamma \cdot |J(x)| + \sum_{s \in S} \rho(y_s - x_s) \\ &= \gamma \cdot (|\mathcal{P}(x)| - 1) + \sum_{I \in \mathcal{P}(x)} \sum_{s \in I} \rho(y_s - \mu_I(x)) \\ &= \tilde{H}_\gamma((\mathcal{P}, \mu_{\mathcal{P}}), y) \end{aligned}$$

and (1.6) is verified. □

Throughout the text we will adopt the following

**Convention** We will identify a signal  $x \in \mathbb{X}$  with the induced segmentation  $(\mathcal{P}, \mu_{\mathcal{P}}) \in \mathfrak{M}$ . In view of Theorem 1.2.4 and Proposition 1.2.7, we will omit the tilde and write  $H_\gamma((\mathcal{P}, \mu_{\mathcal{P}}), y)$  instead of  $\tilde{H}_\gamma(x, y)$ .

In one dimension, a segmentation boils down to the jump locations and the intensities on the intervals between the jumps. In this case, the Potts penalty (1.2) is a function of  $|\mathcal{P}(x)|$  whereas in higher dimensions it is not. We illustrate this by the following example. Moreover, in higher dimensions, jump sets have boundaries with own regularity properties. This is not in the focus of this thesis.

**Example 1.2.8** Consider a  $3 \times 3$  square grid  $S$  with a four neighborhood, where the neighbors of a site are those in the northern, southern, eastern, and western direction. As shown in Figure 1.1, in case of a partition of  $S$  with two intervals there are either two, or three, or four, or six jumps.

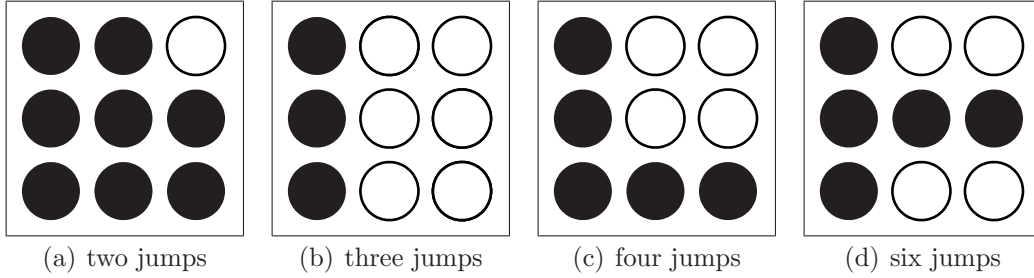


Figure 1.1: Partitions of a  $3 \times 3$ -grid with precisely two intervals

### 1.3 Relation to Bayesian Approaches

In this section, we briefly discuss the relation of the Potts functionals to Bayesian models. First, we recall the Bayesian approach of signal analysis. We will then show that proper prior distributions which are invariant under diagonal translations do not exist on  $\mathbb{R}^n$ , and that, in particular,  $\nu(x) \propto \exp(\gamma \cdot |J(x)|)$  is not a probability distribution.

The Potts functional (1.1) with  $\rho(u) = u^2$  is formally identical to the posterior log likelihood functions of classical Bayesian models for image analysis like those in the seminal paper S. GEMAN and D. GEMAN (1984). From the point of view adopted in this section, the main difference between these models and the variational approach is that in these Bayesian models the functionals take values in finite sets like  $\{0, \dots, 255\}$  whereas we work with real intensities.

Continuous state spaces have several advantages. For example, it is evident that analytical and numerical discussions become extremely unwieldy for discrete spaces.

We focus now on the fact that the Potts penalty term (1.2) depends on intensity differences  $x_s - x_t$  only and hence is invariant under diagonal translations  $x \mapsto x + c\mathbf{1}$ , where  $\mathbf{1}_s = 1$  for every  $s \in S$ . Such a concept does not exist on finite spaces. This could only be circumvented wrapping the intensity range around a torus. But this does not make sense in the present context.

The usual framework for the mentioned Bayesian models is the following one: There is a space  $\mathbf{X}$  of vectors  $(x_1, \dots, x_n)$  the components of which take values in a finite space, and a strictly positive *prior distribution*  $\pi$  on  $\mathbf{X}$ . Such a prior is necessarily of *exponential form*  $\pi(x) \propto \exp(-K(x))$ , see G. WINKLER (2003), p. 21. Clearly,  $K$  corresponds to the penalty term (1.2) of the Potts functional. There is also a - say standard - space  $\mathbf{Y}$  of observations  $y$  and a transition probability  $Q(x, dy)$ .  $Q$  governs the random transition from the ‘true’ configuration  $x$  to data  $y$  which are interpreted as

a degraded version of  $x$ . The *posterior distribution*  $\pi(x | y)$  of  $x$  given  $y$  is then computed from the joint distribution  $\pi(x)Q(x, dy)$  using Bayes' formula and has the form  $\pi(x | y) \propto \exp(-(K(x) - \ln Q(x, y)))$ . Hence the posterior distribution is of exponential form with negative log likelihood

$$H(x, y) = K(x) - \ln Q(x, y). \quad (1.7)$$

The quality of an estimator  $\hat{x}(y) \in \mathbf{X}$  is rated by a *loss function*  $L : \mathbf{X} \times \mathbf{X} \rightarrow \mathbb{R}_+$  with  $L(x, x) = 0$ . The *Bayes risk* of  $\hat{x}$  for a loss function  $L$  is given by the expectation

$$R(\hat{x}, L) = \sum_{x \in \mathbb{X}} \pi(x) \mathbb{E}_{Q(x, dy)} \left( L(\hat{x}(y), x) \right)$$

where  $\mathbb{E}_{\mathbb{P}}(X)$  denotes the expectation of  $X$  under the probability measure  $\mathbb{P}$ . An estimator  $\hat{x}$  is called *Bayes estimator* if it minimizes the Bayes risk. Different Bayes estimators correspond to different loss functions. For the loss function

$$L(x, x') = \begin{cases} 1 & \text{if } x = x', \\ 0 & \text{otherwise.} \end{cases}$$

the Bayes risk  $R(\hat{x}, L)$  is minimized if and only if

$$\hat{x} = x^*(y) = \operatorname{argmax}_{x \in \mathbb{X}} \pi(x|y),$$

see for example G. WINKLER (2003). Such an estimator is called *maximum posterior (MAP) estimator*. The Potts functional  $H_\gamma(x, y)$  from (1.1) is precisely of the form of the negative posterior log likelihood function of an exponential posterior. Therefore, one may wonder whether the Potts functional cannot be studied within the Bayesian framework. In a strict sense, the answer is 'no'. In fact, proper prior distributions with the above invariance property do not exist on  $\mathbb{R}^n$  (and neither on  $\mathbb{Z}^n$ ). The deeper reason behind is that no noncompact topological group admits a finite Haar measure. For Euclidean space this reads as follows.

**Lemma 1.3.1** *Suppose that  $\nu$  is a Borel measure on  $\mathbb{R}^n$  which is invariant under diagonal translations  $(x_1, \dots, x_n) \mapsto (x_1 + c, \dots, x_n + c)$ ,  $c \in \mathbb{R}$ . Then either  $\nu(\mathbb{R}^n) = 0$  or  $\nu(\mathbb{R}^n) = \infty$ .*

For convenience of the reader, we give the standard arguments.

**Proof** The map  $t \mapsto t\mathbf{1}/n^{1/2}$  is an isometric isomorphism between  $\mathbb{R}$  and the diagonal  $D = \{c\mathbf{1} : c \in \mathbb{R}\}$  of  $\mathbb{R}^n$  with its natural norm. Let  $P_D$  denote the

orthogonal projection of  $\mathbb{R}^n$  onto the diagonal and let the image measure  $q$  of  $\nu$  on  $D$  be given by  $q(B) = \nu \circ P_D^{-1}(B)$  for each Borel set  $B$  of  $D$ . As will be proved below,  $q$  is invariant with respect to all translations  $x \mapsto x + c\mathbf{1}$ ,  $x \in D$ ,  $c \in \mathbb{R}$ . Assume now that  $\nu(\mathbb{R}^n) < \infty$ . Then  $q(D) < \infty$  as well and in particular,  $q$  is  $\sigma$ -finite. By a well-known characterization of Lebesgue measure (see for example J. ELSTRODT (1996), Theorem 2.2, p. 89),  $q$  then is a (nonnegative) multiple of Lebesgue measure. Hence  $\nu(\mathbb{R}^n) < \infty$  and consequently  $q(\mathbb{R}^n) < \infty$  would enforce  $q(D) = 0$  which implies the assertion. Let now  $c < c'$  and consider an interval  $[c\mathbf{1}, c'\mathbf{1}]$  on  $D$ . Then

$$\begin{aligned} q([c\mathbf{1}, c'\mathbf{1}] + b\mathbf{1}) &= q([(c+b)\mathbf{1}, (c'+b)\mathbf{1}]) = \nu \circ P_D^{-1}([(c+b)\mathbf{1}, (c'+b)\mathbf{1}]) \\ &= \nu(P_D^{-1}([c\mathbf{1}, c'\mathbf{1}] + b\mathbf{1})) = \nu(P_D^{-1}([c\mathbf{1}, c'\mathbf{1}])) = q([c\mathbf{1}, c'\mathbf{1}]). \end{aligned}$$

Hence  $q$  is translation invariant on  $D$ , and in view of the preceding discussion the proof is complete.  $\square$

The result holds *mutatis mutandis* for  $\mathbb{Z}^n$  with the discrete  $\sigma$ -algebra since under invariance all integer translates of some  $x \in \mathbb{Z}^n$  must have equal mass.

Sometimes a probability measure with density proportional to  $\exp(-(K(x) - \ln Q(x, y)))$  exists even if  $\exp(-K(\cdot))$  is not Lebesgue integrable. In fact, the data term may weight down the leading term sufficiently. Then substitutes for most Bayesian posterior estimators can be defined, in particular ‘posterior means’ etc. *Improper priors* appear in fields like intrinsic auto-regression, cf. H. R. KÜNSCH (1987), or J. BESAG and CH. KOOPERBERG (1995), see also J. O. BERGER (1980, 1985). We will not pursue this aspect in this thesis.



# Chapter 2

## MAP Estimators

In this chapter, we investigate MAP estimators for the Potts functionals. In Section 2.1, we observe that the minimization of  $H((\mathcal{P}, \mu_{\mathcal{P}}), y)$  can be divided into two steps: the minimization in the values on the intervals of a given partition, followed by the minimization in all partitions. If the former is assumed to be known, the continuous minimization problem boils down to a finite discrete minimization problem. This is the crucial observation for the investigation of MAP estimators in the following sections. It allows, for instance, to develop the exact algorithms in Chapter 3. In Section 2.2 we show that under natural conditions MAP estimators exist for all  $y \in \mathbb{X}$ . Moreover, for almost all  $y$ , MAP estimates are unique in important special cases. This is shown in Section 2.3. In Section 2.4, we specify the dependence of the estimator on the hyperparameter. In Section 2.5, we prove joint continuity of MAP estimators in the arguments  $\gamma$  and  $y$ . Finally, in Section 2.6, we show that there is a measurable section of the set-valued map  $(\gamma, y) \mapsto X^*(\gamma, y)$ .

### 2.1 Splitting the Minimization

By Theorem 1.2.4, we can identify a signal  $x \in \mathbb{X}$  with the minimal segmentation  $(\mathcal{P}(x), \mu_{\mathcal{P}(x)}(x)) \in \mathfrak{M}$ . In the one dimensional case, it is natural - and even crucial - to work with  $H((\mathcal{P}, \mu_{\mathcal{P}}), y)$  instead of  $H(x, y)$ . This is justified by Proposition 1.2.7. The minimization problem: given  $y \in \mathbb{X}$ ,

$$\text{minimize } x \longmapsto H(x, y) \quad \text{in } x \in \mathbb{X} \tag{2.1}$$

is equivalent to the problem: given  $y \in \mathbb{X}$ ,

$$\text{minimize } (\mathcal{P}, \mu_{\mathcal{P}}) \longmapsto H((\mathcal{P}, \mu_{\mathcal{P}}), y) \quad \text{in } (\mathcal{P}, \mu_{\mathcal{P}}) \in \mathfrak{M}. \tag{2.2}$$

The most important observation is that the latter minimization problem can be divided into two parts: into the minimization of  $\mu_{\mathcal{P}} \mapsto H((\mathcal{P}, \mu_{\mathcal{P}}), y)$  in  $\mu_{\mathcal{P}} \in \mathbb{R}^{\mathcal{P}}$  for each fixed partition  $\mathcal{P}$ , and into the subsequent minimization over all partitions  $\mathcal{P} \in \mathfrak{P}$ . Since by Hypothesis 1.2.5, the function  $\rho$  is non-negative, the data term  $D((\mathcal{P}, \mu_{\mathcal{P}}), y)$  is minimal in  $\mu_{\mathcal{P}}$  if and only if for each  $I \in \mathcal{P}$  the sum  $\sum_{s \in I} \rho(y_s - \mu_I)$  is minimal in  $\mu_I$ . If these minimizers are known - which is the case in standard situations - then the continuous minimization problem (2.1) is reduced to a *finite* discrete minimization problem. This observation is at the core of the analysis of MAP estimators and of the algorithm for the computation of MAP estimators in one dimension.

Let now  $\mathcal{P} \in \mathfrak{P}$  be fixed. In the following,  $\mu_{\mathcal{P}}^*$  will always denote a family  $(\mu_I^*)_{I \in \mathcal{P}}$  given as a solution of

$$\sum_{s \in I} \rho(y_s - \mu_I^*) = \min_{\mu \in \mathbb{R}} \sum_{s \in I} \rho(y_s - \mu), \quad I \in \mathcal{P}. \quad (2.3)$$

Suppose now that such a  $\mu_{\mathcal{P}}^*$  exists. Define the functional

$$H^*(\cdot, y) : \mathfrak{P} \longrightarrow \mathbb{R}, \quad \mathcal{P} \longmapsto H^*(\mathcal{P}, y) = H((\mathcal{P}, \mu_{\mathcal{P}}^*), y). \quad (2.4)$$

The definition makes sense even if  $\mu_{\mathcal{P}}^*$  is not unique since the value of  $H^*$  is the same for all such minimizers.

**Proposition 2.1.1** *Suppose that Hypothesis 1.2.5 is fulfilled. Then the following holds:*

(1) *The functional*

$$H(\cdot, y) : \mathfrak{S} \longrightarrow \mathbb{R}, \quad (\mathcal{P}, \mu_{\mathcal{P}}) \longmapsto H((\mathcal{P}, \mu_{\mathcal{P}}), y) \quad (2.5)$$

*has a minimum if and only if the functional*

$$H^*(\cdot, y) : \mathfrak{P} \longrightarrow \mathbb{R}, \quad \mathcal{P} \longmapsto H^*(\mathcal{P}, y) \quad (2.6)$$

*has a minimum.*

(2) *The segmentation  $(\mathcal{P}^*, \mu_{\mathcal{P}^*}^*) \in \mathfrak{S}$  is a minimizer of (2.5) if and only if  $\mathcal{P}^*$  minimizes (2.6). Moreover,  $(\mathcal{P}^*, \mu_{\mathcal{P}^*}^*)$  is a minimal segmentation.*

**Proof** (1) Once the partition  $\mathcal{P} \in \mathfrak{P}$  is given, the jump term  $\gamma \cdot (|\mathcal{P}| - 1)$  in  $H((\mathcal{P}, \mu_{\mathcal{P}}), y)$  from (1.5) is fixed, and the minimization of (2.5) boils down to the minimization of the data term. Due to Hypothesis 1.2.5, the latter is minimal if and only if each sum  $\sum_{s \in I} \rho(y_s - \mu_I)$  is minimal. Hence the minimization of  $H((\mathcal{P}, \mu_{\mathcal{P}}), y)$  in  $(\mathcal{P}, \mu_{\mathcal{P}}) \in \mathfrak{S}$  is equivalent to the minimization

of the data term in  $\mu_{\mathcal{P}}$  for each partition  $\mathcal{P}$  and, provided that  $\mu_{\mathcal{P}}^*$  exists for each  $\mathcal{P}$ , the subsequent minimization of  $H^*(\mathcal{P}, y)$  in  $\mathcal{P} \in \mathfrak{P}$ .

(2) Suppose that  $\mu_{\mathcal{P}}^* \in \mathbb{R}^{\mathcal{P}}$  from (2.3) exists. Suppose further that  $\mathcal{P}^* \in \mathfrak{P}$  minimizes (2.6). It remains to prove that  $(\mathcal{P}^*, \mu_{\mathcal{P}^*}^*)$  is a *minimal* segmentation. Assume that it is not minimal. Then there are neighbor intervals  $I \sim J \in \mathcal{P}^*$  such that  $\mu_I^* = \mu_J^*$ . Merging  $I$  and  $J$  gives a partition  $\mathcal{Q} \in \mathfrak{P}$  with the same data term as  $\mathcal{P}^*$  but  $|\mathcal{Q}| = |\mathcal{P}^*| - 1$ . Hence,

$$\gamma(|\mathcal{Q}| - 1) + \sum_{I \in \mathcal{Q}} \sum_{s \in I} \rho(y_s - \mu_I^*) < \gamma(|\mathcal{P}^*| - 1) + \sum_{I \in \mathcal{P}^*} \sum_{s \in I} \rho(y_s - \mu_I^*)$$

in contradiction to the assumption that  $(\mathcal{P}^*, \mu_{\mathcal{P}^*}^*)$  is a minimizer of (2.5).  $\square$

**Theorem 2.1.2** *Suppose that Hypothesis 1.2.5 is fulfilled. If a minimum of (2.5) exists, then*

$$\min_{(\mathcal{P}, \mu_{\mathcal{P}}) \in \mathfrak{S}} H((\mathcal{P}, \mu_{\mathcal{P}}), y) = \min_{(\mathcal{P}, \mu_{\mathcal{P}}) \in \mathfrak{M}} H((\mathcal{P}, \mu_{\mathcal{P}}), y).$$

Moreover,

$$\min_{(\mathcal{P}, \mu_{\mathcal{P}}) \in \mathfrak{M}} H((\mathcal{P}, \mu_{\mathcal{P}}), y) = \min_{\mathcal{P} \in \mathfrak{P}} \left( \gamma \cdot (|\mathcal{P}| - 1) + \sum_{I \in \mathcal{P}} \sum_{s \in I} \rho(y_s - \mu_I^*) \right)$$

where  $(\mu_I^*)_{I \in \mathcal{P}}$  is given by (2.3).

**Proof** Proposition 2.1.1 provides the statement.  $\square$

In case of time series' this reads as follows.

**Corollary 2.1.3** *Let be  $S = \{1, \dots, N\}$ . Under Hypothesis 1.2.5, a MAP estimator  $x^*$  of the one dimensional Potts functional exists if and only if a minimizer  $(\mathcal{P}^*, \mu_{\mathcal{P}^*}^*)$  of (2.5) exists.  $x^*$  minimizes*

$$H(\cdot, y) : \mathbb{X} \longrightarrow \mathbb{R}, \quad x \longmapsto H(x, y)$$

if and only if

$$x_s^* = \mu_{I(s)}^*, \quad I(s) \in \mathcal{P}^*, \quad s \in S. \quad (2.7)$$

In particular, if minima exist, then

$$\min_{x \in \mathbb{X}} H(x, y) = \min_{\mathcal{P} \in \mathfrak{P}} \left( \gamma \cdot (|\mathcal{P}| - 1) + D((\mathcal{P}, \mu_{\mathcal{P}}^*), y) \right)$$

with  $D((\mathcal{P}, \mu_{\mathcal{P}}), y)$  from (1.4).

After these preparations we discuss properties of MAP estimators of the one dimensional Potts functional.

## 2.2 Existence

In this section, we prove existence of MAP estimators. The Potts functionals  $H_\gamma(x, y)$  have minima under natural conditions. In addition to Hypothesis 1.2.5 we will assume:

**Hypothesis 2.2.1** *Let  $\rho$  be such that*

(M) *for every interval  $I \subset S$ , a minimizer  $\mu_I^*$  of*

$$D_I : \mathbb{R} \longrightarrow \mathbb{R}, \quad \mu_I \longmapsto \sum_{s \in I} \rho(y_s - \mu_I)$$

*exists.*

By Hypothesis 1.2.5, the function  $\rho$  is nonnegative, and by (M), for each partition  $\mathcal{P} \in \mathfrak{P}$  there is a minimizer  $\mu_{\mathcal{P}}^*$  given by (2.3).

Hypothesis 1.2.5 does not imply Hypothesis 2.2.1 without further assumptions as the following example shows.

**Example 2.2.2** As a counterexample consider the function

$$\rho(u) = \begin{cases} |u| & \text{for } |u| < 1, \\ 2|u| & \text{for } |u| \geq 1. \end{cases}$$

This function obviously fulfills Conditions (H1)-(H3) from Hypothesis 1.2.5. Let further data  $y_1 = 0$  and  $y_2 = 2$  be given, and define for  $I = \{1, 2\}$  the sum  $\sum_{s \in I} \rho(y_s - \mu_I)$  as the function

$$f(\mu) = \rho(y_1 - \mu) + \rho(y_2 - \mu) = \rho(\mu) + \rho(2 - \mu).$$

Inserting  $\rho$  gives

$$f(\mu) = \begin{cases} |\mu| + 2|2 - \mu| & \text{for } \mu \in ]-1, 1[, \\ 2|\mu| + |2 - \mu| & \text{for } \mu \in (1, 3), \\ 2|\mu| + 2|2 - \mu| & \text{for } \mu \in (-\infty, -1] \cup [3, \infty) \cup \{1\}. \end{cases}$$

The graph of  $f$  is displayed in Figure 2.1. Obviously this function has no minimum. It is upper but not lower semicontinuous.

Under a continuity assumption, condition (M) is automatically fulfilled.

**Lemma 2.2.3** *If  $\rho$  is lower semicontinuous then Hypothesis 1.2.5 implies Hypothesis 2.2.1.*

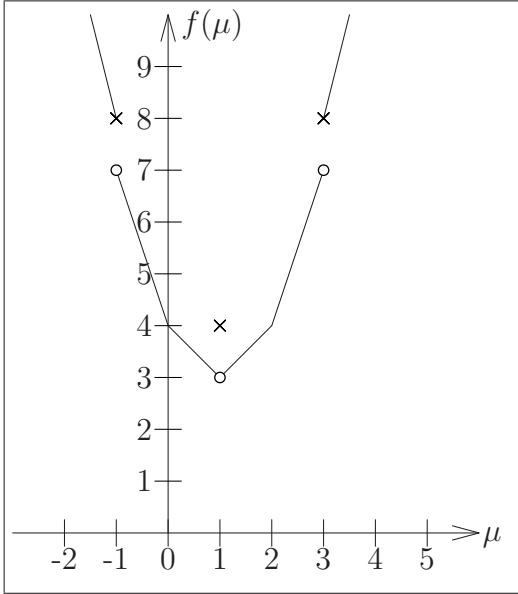


Figure 2.1: The symbol  $\times$  indicates that  $f$  takes this value, and  $\circ$  indicates that it does not.

**Proof** Assume that  $\rho$  fulfills Hypothesis 1.2.5. Consider for any interval  $I \subset S$  the function

$$f(\mu) = \sum_{s \in I} \rho(y_s - \mu).$$

Denote by  $y_{\min}$  and  $y_{\max}$  the smallest and largest, respectively, of the values  $y_s$ ,  $s \in I$ . On  $(-\infty, y_{\min}]$ , the function  $f$  is decreasing. On  $[y_{\max}, \infty)$ , it is increasing. This implies  $f(\mu) \geq f(y_{\min})$  for  $\mu \leq y_{\min}$ , and  $f(\mu) \geq f(y_{\max})$  for  $\mu \geq y_{\max}$ . Therefore,

$$\inf_{\mu \in \mathbb{R}} f(\mu) = \inf_{\mu \in [y_{\min}, y_{\max}]} f(\mu).$$

As a sum of lower semicontinuous functions,  $f$  is lower semicontinuous itself. It has a minimum on the real compact interval  $[y_{\min}, y_{\max}]$ , i. e.

$$\inf_{\mu \in \mathbb{R}} f(\mu) = \min_{\mu \in [y_{\min}, y_{\max}]} f(\mu),$$

and Hypothesis 2.2.1 is fulfilled.  $\square$

**Remark 2.2.4** Hypothesis 1.2.5 implies Hypothesis 2.2.1 for a single  $y \in \mathbb{R}^S$  if  $u \mapsto \rho(u)$  is lower semicontinuous for  $|u| \leq \max_{s,t \in S} |y_s - y_t|$ .

Existence of MAP estimates is now readily proved.

**Theorem 2.2.5** *If  $\rho$  satisfies Hypotheses 1.2.5 and 2.2.1, then a minimum of the Potts functional  $x \mapsto H_\gamma(x, y)$  exists.*

**Proof** Hypotheses 1.2.5 and 2.2.1 imply the existence of a minimizer  $(\mu_I^*)_{I \in \mathcal{P}}$  of the data term for each partition  $\mathcal{P}(x) \in \mathfrak{P}$ . Since the penalty term depends only on the partition, the functional

$$H(\cdot, y) : \mathbb{X} \longrightarrow \mathbb{R}, \quad x \longmapsto \gamma \cdot |J(x)| + \sum_{I \in \mathcal{P}(x)} \sum_{s \in I} \rho(y_s - \mu_I(x))$$

has a minimum. Since the number of partitions in  $\mathfrak{P}$  is finite, this implies the existence of a minimum.  $\square$

The assumptions in Theorem 2.2.5 hold in important cases.

**Example 2.2.6** (1) Standard examples are squares and moduli:

- (a) If  $\rho(u) = u^2$  then the minimizer  $\mu_I^*$  is the empirical mean  $\bar{y}_I$ .
- (b) If  $\rho(u) = |u|$  then a minimizer  $\mu_I^*$  is a median of  $\{y_s : s \in I\}$ .

(2) Among others, the following functions from robust statistics satisfy the assumption of Theorem 2.2.5 as well:

- (a) Functions of the form

$$\rho(u) = \begin{cases} \lambda^2 u^2 & : \text{ for } |u| < \delta \\ 2\lambda\sqrt{\alpha}|u| - \alpha & : \text{ for } |u| \geq \delta \end{cases}, \quad \delta = \frac{\sqrt{\alpha}}{\lambda}$$

proposed in P. J. HUBER (1981).

- (b) Cup-shaped functions advocated by F.R. HAMPEL et al. (1986) given by

$$\rho(u) = \begin{cases} \lambda^2 u^2 & : \text{ for } |u| < \delta \\ \alpha & : \text{ for } |u| \geq \delta \end{cases}, \quad \delta = \frac{\sqrt{\alpha}}{\lambda}.$$

We turn now to uniqueness.

## 2.3 Uniqueness

In general, a Potts functional has more than one minimizer. Fortunately, the minimizer is unique in important special cases. We will show that uniqueness holds for Lebesgue almost all  $y \in \mathbb{X}$  if the data term  $D(x, y)$  is a sum of squares.

**Theorem 2.3.1** *For each  $\gamma > 0$  there is a Lebesgue null set  $N_\gamma \subset \mathbb{X}$  such that for each  $y \notin N_\gamma$  the Potts functional*

$$\bar{H}_\gamma(\cdot, y) : \mathbb{X} \longrightarrow \mathbb{R}, \quad x \longmapsto \gamma \cdot |J(x)| + \sum_{s \in S} (y_s - x_s)^2, \quad (2.8)$$

*has a unique minimizer  $x^*(\gamma, y)$ .*

The proof will be given shortly. Let now  $\nu$  be a Borel measure on  $\mathbb{X} = \mathbb{R}^S$  with Lebesgue density. Then each Lebesgue null set is also a  $\nu$ -null set. This implies:

**Corollary 2.3.2** *For each Borel measure on  $\mathbb{X}$  admitting a Lebesgue density and each  $\gamma > 0$  the minimizer  $x^*(\gamma, y)$  of (2.8) is unique for almost all  $y \in \mathbb{X}$ .*

The proof of Theorem 2.3.1 is based on the observation that the existence of two different minimizers imposes a constraint on  $y$ . Recall, that the data term in (2.8) can be rewritten as

$$D((\mathcal{P}(x), \mu_{\mathcal{P}(x)}(x)), y) = \sum_{I \in \mathcal{P}(x)} \sum_{s \in I} (y_s - \mu_I(x))^2 \quad (2.9)$$

and the family of minimizers  $(\mu_I^*)_{I \in \mathcal{P}}$  for a fixed partition  $\mathcal{P} \in \mathfrak{P}$  is given by

$$\bar{y}_I = \frac{1}{|I|} \sum_{t \in I} y_t, \quad I \in \mathcal{P}. \quad (2.10)$$

Note that the function  $\rho(u) = u^2$  fulfills Hypotheses 1.2.5 and 2.2.1. We will show that the set

$$N_\gamma = \{y \in \mathbb{X} : x \mapsto \bar{H}_\gamma(x, y) \text{ has at least two different minimizers}\} \quad (2.11)$$

is a Lebesgue null set.

**Proof of Theorem 2.3.1** (1) Let  $y \in \mathbb{X}$ , and suppose that  $\bar{H}_\gamma(x, y) = \bar{H}_\gamma(x', y)$  for  $x \neq x'$  in  $\mathbb{X}$ . Identifying  $x$  and  $x'$  with their induced segmentations  $(\mathcal{P}(x), \mu_{\mathcal{P}(x)}(x))$  and  $(\mathcal{P}(x'), \nu_{\mathcal{P}(x')}(x'))$  in  $\mathfrak{M}$ , this is equivalent to

$$\begin{aligned} \frac{1}{\gamma} \left( \sum_{I \in \mathcal{P}(x)} \sum_{s \in I} (y_s - \mu_I(x))^2 - \sum_{J \in \mathcal{P}(x')} \sum_{s \in J} (y_s - \mu_J(x'))^2 \right) \\ = |J(x')| - |J(x)|. \end{aligned} \quad (2.12)$$

(2) Let now  $\mathcal{P} \neq \mathcal{Q}$  be partitions in  $\mathfrak{P}$ , and let  $\bar{y}_{\mathcal{P}}(y)$  and  $\bar{y}_{\mathcal{Q}}(y)$  be the associated minimizers of the data terms (2.9), respectively. Define the function

$$f_\gamma^{\mathcal{P}, \mathcal{Q}} : \mathbb{X} \rightarrow \mathbb{R}, \quad y \longmapsto \frac{1}{\gamma} \left( \sum_{I \in \mathcal{P}} \sum_{s \in I} (y_s - \bar{y}_I)^2 - \sum_{J \in \mathcal{Q}} \sum_{s \in J} (y_s - \bar{y}_J)^2 \right) \quad (2.13)$$



**Proof** All empirical means  $\bar{y}_I$  are linear on  $\mathbb{X}$ . Simple calculations then give the assertion.  $\square$

We continue with notation from the proof of Theorem 2.3.1.

**Lemma 2.3.5** *Let be  $\mathcal{P}$  and  $\mathcal{Q}$  be different partitions in  $\mathfrak{B}$ . Then the set  $A_\gamma^{\mathcal{P},\mathcal{Q}}$  defined in (2.14) is a Lebesgue null set.*

**Proof** Application of the binomial formula and Lemma 2.3.4 yield the representation of  $f_\gamma^{\mathcal{P},\mathcal{Q}}$  as the quadratic form

$$f_\gamma^{\mathcal{P},\mathcal{Q}}(y) = \frac{1}{\gamma} \left( \sum_{J \in \mathcal{Q}} |J| \bar{y}_J^2 - \sum_{I \in \mathcal{P}} |I| \bar{y}_I^2 \right) = \frac{1}{\gamma} \left( y^t (B_\mathcal{Q} - B_\mathcal{P}) y \right).$$

Since  $\mathcal{P} \neq \mathcal{Q}$  we have  $B_\mathcal{Q} - B_\mathcal{P} \neq 0$ . Decompose now the set  $A_\gamma^{\mathcal{P},\mathcal{Q}}$  from (2.14) into the two disjoint sets  $\tilde{U}$  and  $\tilde{V}$ , where

$$\tilde{U} := \{y \in A_\gamma^{\mathcal{P},\mathcal{Q}} : \nabla f_\gamma^{\mathcal{P},\mathcal{Q}}(y) = 0\}, \quad \tilde{V} := A_\gamma^{\mathcal{P},\mathcal{Q}} \setminus \tilde{U}$$

and where  $\nabla$  denotes the gradient. Then the following inclusions hold:

$$\begin{aligned} \tilde{U} &\subset \{y \in A_\gamma^{\mathcal{P},\mathcal{Q}} : (B_\mathcal{Q} - B_\mathcal{P})y = 0\} =: U, \\ \tilde{V} &\subset \{y \in A_\gamma^{\mathcal{P},\mathcal{Q}} : (B_\mathcal{Q} - B_\mathcal{P})y \neq 0\} =: V. \end{aligned}$$

Since  $B_\mathcal{Q} - B_\mathcal{P} \neq 0$ , the dimension of the linear subspace  $U$  is strictly smaller than that of  $\mathbb{X} = \mathbb{R}^S$ . Hence, denoting by  $\lambda$  the Lebesgue measure on the Borel- $\sigma$ -field of  $\mathbb{R}^N$ , we have  $\lambda(U) = 0 = \lambda(\tilde{U})$ .

For  $y \in V$  the gradient takes the form

$$\nabla f_\gamma^{\mathcal{P},\mathcal{Q}}(y) = 2(B_\mathcal{Q} - B_\mathcal{P})y \neq 0.$$

Hence, for every  $y \in V$  there is an open neighborhood  $W(y) \subset \mathbb{X}$  such that  $W(y) \cap V$  is a  $\mathcal{C}^\infty$ -submanifold of  $W(y)$ , by M. W. HIRSCH (1976), Theorem 3.2, p. 22. It is  $\dim(W(y) \cap V) < \dim \mathbb{X}$  and hence

$$\lambda(W(y) \cap V) = 0.$$

By B. v. QUERENBURG (1976), Theorem 8.29, p. 91, the set  $V \subset \mathbb{R}^S$ , equipped with the relative topology, enjoys the Lindelöf property, which means that there is a countable subset  $\{u_i : i \in \mathbb{N}\} \subset V$  such that

$$V \subset \bigcup_{i \in \mathbb{N}} W(u_i).$$

Since  $\tilde{V} \subset V$ , we obtain  $\lambda(\tilde{V}) \leq \lambda(V) \leq \sum_{i \in \mathbb{N}} \lambda(W(u_i) \cap V) = 0$ . Therefore we get

$$\lambda\left(A_\gamma^{\mathcal{P}, \mathcal{Q}}\right) \leq \lambda(\tilde{U}) + \lambda(\tilde{V}) = 0$$

which implies the assertion.  $\square$

## 2.4 Dependence on Hyperparameters

In this section, we consider times series  $y$  of length  $N$  with  $S = \{1, \dots, N\}$ . We will have a closer look at the relation between the hyperparameter  $\gamma$  in the one dimensional Potts functionals and the set  $X^*(\gamma, y)$  of corresponding MAP estimates in the case of  $\rho(u) = u^2$ . Hence, we consider the functionals

$$\bar{H}_\gamma(\cdot, y) : \mathfrak{M} \longrightarrow \mathbb{R}, \quad (\mathcal{P}, \mu_{\mathcal{P}}) \longmapsto \gamma \cdot (|\mathcal{P}| - 1) + \sum_{I \in \mathcal{P}} \sum_{s \in I} (y_s - \mu_I)^2. \quad (2.16)$$

The set of all partitions  $\mathcal{P} \in \mathfrak{P}$  with  $|\mathcal{P}| = k$  will be denoted by  $\mathfrak{P}_k$ . With the family of minimizers  $(\bar{y}_I)_{I \in \mathcal{P}}$  from (2.10) we will introduce the minimum data term for partitions  $\mathcal{P} \in \mathfrak{P}_k$  by

$$\tilde{B}_y(k) := \min_{\mathcal{P} \in \mathfrak{P}_k} D((\mathcal{P}, \mu_{\mathcal{P}}^*), y) = \min_{\mathcal{P} \in \mathfrak{P}_k} \sum_{I \in \mathcal{P}} \sum_{s \in I} (y_s - \bar{y}_I)^2. \quad (2.17)$$

The next result is a corollary of Theorem 2.3.1.

**Proposition 2.4.1** *For Lebesgue almost all  $y \in \mathbb{R}^S$  the function  $k \mapsto \tilde{B}_y(k)$  decreases strictly.*

**Proof** First, we show, by way of induction, that for each  $y \in \mathbb{R}^S$  the function  $k \mapsto \tilde{B}_y(k)$  decreases. The set  $\mathfrak{P}_{k+1}$  is related to  $\mathfrak{P}_k$  in the following way:

$$\mathfrak{P}_{k+1} = \left\{ (\mathcal{P} \setminus \{I\}) \cup \{I_1, I_2\} : \mathcal{P} \in \mathfrak{P}_k, I \in \mathcal{P}, I_1 \cup I_2 = I, I_1 \cap I_2 = \emptyset \right\}.$$

Choose  $I \in \mathfrak{P}_k$  and write it as the disjoint union  $I = I_1 \dot{\cup} I_2$  of intervals  $I_1$  and  $I_2$  to get a generic element of  $\mathfrak{P}_{k+1}$ . Then

$$\begin{aligned} \sum_{s \in I} (y_s - \bar{y}_I)^2 &= \sum_{s \in I_1} (y_s - \bar{y}_I)^2 + \sum_{s \in I_2} (y_s - \bar{y}_I)^2 \\ &\geq \sum_{s \in I_1} (y_s - \bar{y}_{I_1})^2 + \sum_{s \in I_2} (y_s - \bar{y}_{I_2})^2 \end{aligned}$$

which implies the assertion.

Hence it is sufficient to show that  $\tilde{B}_y(k) \neq \tilde{B}_y(k')$  for  $k \neq k'$  for Lebesgue almost all  $y \in \mathbb{X}$ . If equality holds then there are two minimal segmentations  $(\mathcal{Q}, \bar{y}_{\mathcal{Q}})$  and  $(\mathcal{R}, \bar{y}_{\mathcal{R}})$  with  $|\mathcal{Q}| = k \neq k' = |\mathcal{R}|$  such that

$$\sum_{I \in \mathcal{Q}} \sum_{s \in I} (y_s - \bar{y}_I)^2 - \sum_{J \in \mathcal{R}} \sum_{s \in J} (y_s - \bar{y}_J)^2 = 0.$$

The set of  $y \in \mathbb{R}^S$  which satisfy this equation is a subset of the Lebesgue null set  $N_\gamma$  from (2.11). Thus, the set of  $y \in \mathbb{R}^S$  for which  $\tilde{B}_y(k)$  is not strictly decreasing has Lebesgue measure zero.  $\square$

With  $\bar{H}_\gamma((\mathcal{P}, \mu_{\mathcal{P}}), y)$  from (2.16), let *minimum functions* be defined as

$$h_y : (0, \infty) \longrightarrow [0, \infty), \quad \gamma \longmapsto \min_{(\mathcal{P}, \mu_{\mathcal{P}}) \in \mathfrak{M}} \bar{H}_\gamma((\mathcal{P}, \mu_{\mathcal{P}}), y), \quad y \in \mathbb{X}.$$

Define further

$$f_y^k : (0, \infty) \longrightarrow \mathbb{R}, \quad \gamma \longmapsto \gamma \cdot (k - 1) + \tilde{B}_y(k), \quad y \in \mathbb{X}, \quad 1 \leq k \leq N. \quad (2.18)$$

By Corollary 2.1.3, we have

$$h_y(\gamma) = \min_{1 \leq k \leq N} f_y^k(\gamma). \quad (2.19)$$

The minimum function has the following properties.

**Proposition 2.4.2** *Choose  $y \in \mathbb{R}^N$ . If  $y$  is a constant signal then the minimum functions coincide with the null function. Otherwise, there are an integer  $m(y) \in \{0, \dots, N - 2\}$  and a set of hyperparameters*

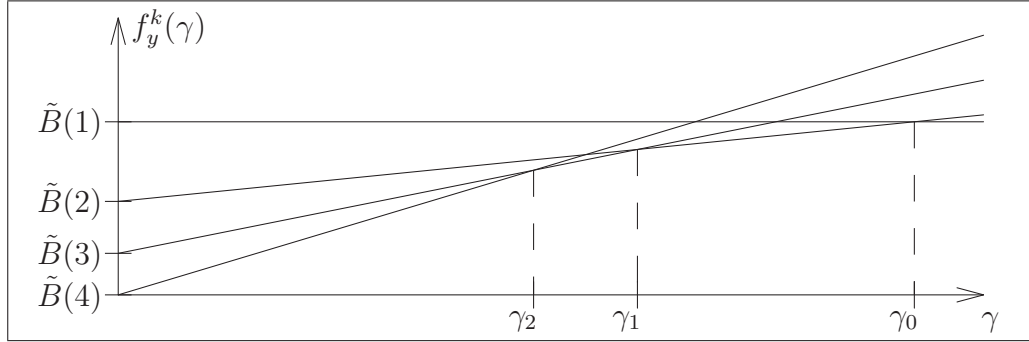
$$\gamma_{m(y)+1} = 0 < \gamma_{m(y)}(y) < \dots < \gamma_0(y) < \infty = \gamma_{-1} \quad (2.20)$$

*such that  $h_y$  is continuous and linear on the intervals  $(\gamma_i(y), \gamma_{i-1}(y))$ ,  $1 \leq i \leq m(y) + 1$ .*

**Proof** We distinguish two cases.

(1)  $|J(y)| = 0$ : For constant  $y$ , the minimum data term  $\tilde{B}_y(k)$  vanishes for all  $k$ . Thus, a segmentation minimizing  $\bar{H}_\gamma$  for constant  $y$  is unique, coincides with  $y$  and the minimum function is the null function.

(2)  $|J(y)| \geq 1$ : First, we observe that there is only a finite number of linear functions  $f_y^k$  from (2.18). Any two functions have exactly one intersection point: In fact, the intercepts of  $f_y^k$  are smaller the larger  $k$  is since the function  $k \mapsto \tilde{B}_y(k)$  is decreasing and the slopes increase in  $k$ . This is illustrated in Figure 2.2 for  $N = 4$ . The finite number of intersection points where the

Figure 2.2: The functions  $f_y^k$ 

minimizing function  $f_y^k$  in (2.19) changes will be denoted by  $m(y)$  and the intersection points will be denoted by  $\gamma_i(y)$ ,  $i = 0, \dots, m(y)$ . Between the points  $\gamma_i(y)$ ,  $i = 0, \dots, m(y)$ , the function  $h_y$  coincides with one of the linear functions  $f_y^k$ .  $\square$

For Lebesgue almost all  $y$  and a given number  $1 \leq k \leq N$  there is precisely one segmentation which minimizes the Potts functional  $\bar{H}_\gamma(\cdot, y)$  from (2.16) on  $\mathfrak{P}_k$ . We define

$$K_y^*(\gamma) = \{k_y^*(\gamma) \in \{1, \dots, N\} : f_y^{k_y^*(\gamma)}(\gamma) = \min_{1 \leq k \leq N} f_y^k(\gamma)\}. \quad (2.21)$$

Note that a minimizer  $k_y^*(\gamma)$  always exists since there is only a finite number of competing functions  $f_y^k(\gamma)$ . The following proposition is a corollary to Theorem 2.3.1.

**Proposition 2.4.3** *Let  $\gamma > 0$ . Then there is a Lebesgue null set  $N \subset \mathbb{X}$ , such that for each  $y \notin N$  and each  $k_y^*(\gamma) \in K_y^*(\gamma)$  there is a unique minimizer of*

$$\bar{H}^*(\cdot, y) |_{\mathfrak{P}_{k_y^*(\gamma)}} : \mathfrak{P}_{k_y^*(\gamma)} \longrightarrow \mathbb{R}, \quad \mathcal{P} \longmapsto \bar{H}((\mathcal{P}, \bar{y}_{\mathcal{P}}), y). \quad (2.22)$$

**Proof** Choose  $y \in \mathbb{X}$ ,  $\gamma > 0$ , and  $k^* \in K_y^*(\gamma)$ . Assume that  $\mathcal{P} \neq \mathcal{Q}$  are two partitions minimizing  $\bar{H}^*$  on  $\mathfrak{P}_{k^*}$ . Then  $\bar{H}^*(\mathcal{P}, y) = \bar{H}^*(\mathcal{Q}, y)$  and, since  $|\mathcal{P}| = |\mathcal{Q}| = k^*$ , we get

$$\sum_{I \in \mathcal{P}} \sum_{s \in I} (y_s - \bar{y}_I)^2 = \sum_{J \in \mathcal{Q}} \sum_{s \in J} (y_s - \bar{y}_J)^2.$$

This is equation (2.12) for  $|\mathcal{P}| = |\mathcal{Q}|$  and the assertion follows from Theorem 2.3.1. In particular, the Lebesgue null set does not depend on  $\gamma$ .  $\square$

**Remark 2.4.4** Suppose now that  $K_y^*(\gamma)$  from (2.21) is a singleton. Then, by Proposition 2.4.3, the number  $k_y^*(\gamma)$  defines a unique minimizer  $x^*(\gamma, y)$ .

We will now show that for almost all  $y$  there are only finitely many different MAP estimators. We will make this more precise in the following theorem.

**Theorem 2.4.5** *Let  $\bar{H}_\gamma(\cdot, y)$  be the one dimensional Potts functional from (2.16) and let be  $x^*(\gamma, y) \in X^*(\gamma, y)$ . For Lebesgue almost all data  $y \in \mathbb{R}^S$  there are an integer  $0 \leq m(y) \leq |S| - 2$  and a set of hyperparameters*

$$\gamma_{m(y)+1} = 0 < \gamma_{m(y)}(y) < \cdots < \gamma_0(y) < \infty = \gamma_{-1}$$

such that

- (1)  $x^*(\gamma, y)$  is unique except for  $\gamma = \gamma_i(y)$ ,  $i = 0, \dots, m(y)$ .
- (2)  $x^*(\gamma, y) = x^*(\gamma', y)$  for all  $\gamma, \gamma' \in (\gamma_i(y), \gamma_{i-1}(y))$ ,  $i = 0, \dots, m(y) + 1$ . The MAP estimator  $x^*(\gamma, y)$  is a constant signal for each  $\gamma > \gamma_0(y)$ , and  $x^*(\gamma, y) = y$  for  $\gamma < \gamma_{m(y)}(y)$ .
- (3) For each  $0 \leq i \leq m(y)$  the functional  $x \mapsto \bar{H}_{\gamma_i(y)}(x, y)$  has precisely the two minimizers belonging to the intervals adjacent to  $\gamma_i(y)$ .

The functions  $i \mapsto |J(x^*(\gamma, y))|$ ,  $\gamma \in (\gamma_i(y), \gamma_{i-1}(y))$ , increase strictly. Theorem 2.4.5 suggests the following definition.

**Definition 2.4.6** *The minimization of the one dimensional Potts functional for all values of the hyperparameter will be called  $\gamma$ -scanning. The intervals  $(\gamma_i(y), \gamma_{i-1}(y))$ ,  $i = 0, \dots, m(y) + 1$ , of the hyperparameter from Theorem 2.4.5 will be called  $\gamma$ -intervals.*

**Proof of Theorem 2.4.5** By Proposition 2.4.3, for fixed hyperparameter  $\gamma$  a minimizer  $k^*(\gamma)$  from (2.21) uniquely defines a minimizing partition. There is a continuum of  $\gamma$ -values which is mapped to the finite set of possible values for  $k^*$ . Now we must guarantee that the partition minimizing  $\bar{H}^*$  from (2.22) on  $\mathfrak{P}_k^*$  is the same for all values of  $\gamma$  which correspond to one number  $k^*$ . Suppose that  $\mathcal{P}$  and  $\mathcal{Q}$  are different partitions in  $\mathfrak{P}_{k^*}$ . Then the linear functions  $\gamma \mapsto \gamma \cdot (k^* - 1) + D((\mathcal{P}, \bar{y}_{\mathcal{P}}), y)$  and  $\gamma \mapsto \gamma \cdot (k^* - 1) + D((\mathcal{Q}, \bar{y}_{\mathcal{Q}}), y)$  have the same slope. Hence, if one is above the other for some  $\gamma$  it will stay above for any  $\gamma$ . There are two alternatives: Either there is exactly one  $k^*(\gamma)$  and then, by Proposition 2.4.3, there is a Lebesgue null set  $N \subset \mathbb{X}$ , independent of  $\gamma$ , such that for all  $y$  in the complement  $N^c = \mathbb{X} \setminus N$  and all  $\gamma$  the number  $k_y^*(\gamma)$  defines a unique minimizer  $x^*(\gamma, y)$ . The other case is

$|K_y^*(\gamma)| \geq 2$ . By Proposition 2.4.2, there is only a finite number of points  $\gamma_i(y)$  where the latter is the case. In Lemma 2.4.7 below, we will show that for Lebesgue almost all  $y$  we have exactly two minimizers  $x^*(\gamma, y)$  for those  $\gamma_i(y)$ .  $\square$

The following lemma completes the proof of Theorem 2.4.5.

**Lemma 2.4.7** *For Lebesgue almost all  $y \in \mathbb{X}$  and each  $\gamma = \gamma_i(y)$ ,  $0 \leq i \leq m(y)$ , from (2.20) there are precisely two MAP estimates  $x^*(\gamma, y)$  belonging to the  $\gamma$ -intervals adjacent to  $\gamma_i(y)$ .*

**Proof** We will show that the set of  $y \in \mathbb{R}^S$  for which there are more than two segmentations minimizing  $\bar{H}_\gamma(\cdot, y)$  from (2.16) for fixed  $\gamma = \gamma(y)$  is contained in a Lebesgue null set. Let  $\mathcal{P}_1$ ,  $\mathcal{P}_2$  and  $\mathcal{P}_3$  be three mutually different partitions in  $\mathfrak{P}$  and assume that they all minimize the functional  $\mathcal{P} \mapsto \bar{H}_\gamma(\mathcal{P}, y)$  from (2.6) with  $\bar{y}_{\mathcal{P}}$  given by (2.10). This implies the following equations

$$\begin{aligned} \gamma \cdot (|\mathcal{P}_1| - 1) + D((\mathcal{P}_1, \bar{y}_{\mathcal{P}_1}), y) &= \gamma \cdot (|\mathcal{P}_2| - 1) + D((\mathcal{P}_2, \bar{y}_{\mathcal{P}_2}), y) \\ \gamma \cdot (|\mathcal{P}_2| - 1) + D((\mathcal{P}_2, \bar{y}_{\mathcal{P}_2}), y) &= \gamma \cdot (|\mathcal{P}_3| - 1) + D((\mathcal{P}_3, \bar{y}_{\mathcal{P}_3}), y). \end{aligned}$$

Application of the binomial formula and straightforward calculations give

$$\begin{aligned} -(|\mathcal{P}_3| - |\mathcal{P}_2|) \sum_{I \in \mathcal{P}_1} |I| \bar{y}_I^2 + & \quad (2.23) \\ + (|\mathcal{P}_3| - |\mathcal{P}_1|) \sum_{I \in \mathcal{P}_2} |I| \bar{y}_I^2 - (|\mathcal{P}_2| - |\mathcal{P}_1|) \sum_{K \in \mathcal{P}_3} |K| \bar{y}_K^2 &= 0. \end{aligned}$$

By Lemma 2.3.4, the left hand side is a quadratic form  $y \mapsto y^t A y$  with

$$A = A_1 + A_2 + A_3$$

where

$$\begin{aligned} A_1 &= -(|\mathcal{P}_3| - |\mathcal{P}_2|) B_{\mathcal{P}_1} \\ A_2 &= (|\mathcal{P}_3| - |\mathcal{P}_1|) B_{\mathcal{P}_2} \\ A_3 &= -(|\mathcal{P}_2| - |\mathcal{P}_1|) B_{\mathcal{P}_3} \end{aligned}$$

with  $B_{\mathcal{P}_i}$  from (2.15). All  $A_i$  have the same block structure as  $B_{\mathcal{P}_i}$ . We will show that the  $A_i$  have blocks of different length which implies that  $A$  does not vanish. Hence the set of  $y$  which fulfill (2.23) is the kernel of a nonzero quadratic form and thus a Lebesgue null subset of  $\mathbb{R}^S$ .

We denote the blocks of  $A_i$  by  $A_{i,n(i)}$ ,  $i = 1, 2, 3$ . Since, by assumption, the partitions are different there is

$$j^* = \min\{j : \text{at least two blocks } A_{i,j} \text{ are of different length}\}.$$

For these blocks there are two cases.

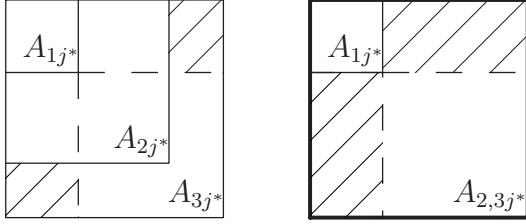


Figure 2.3: Block of matrix  $A$

(a) All blocks are of different size.

(b) Two blocks have the same size.

(a) All  $A_{i,j^*}$  are of different size. Denoting by  $|A|$  the size of a quadratic matrix, we may assume without loss of generality that  $|A_{1,j^*}| < |A_{2,j^*}| < |A_{3,j^*}|$ . The largest block containing the two smaller blocks is symbolically displayed in Figure 2.3(a). In the shaded region of Figure 2.3(a) only the largest block has non-zero entries and hence  $A$  does not vanish.

(b) Assume now that  $|A_{1,j^*}| < |A_{2,j^*}| = |A_{3,j^*}|$ . In Figure 2.3(b) we see the two larger blocks (indicated by the extra thick line) containing the smaller one. In the shaded region of Figure 2.3(b), the matrix  $A_1$  is equal to zero and has no influence on the values of  $A$ . Either, matrix  $A$  is zero in the shaded region, then it is not zero in the smaller block since there  $A_{1,j^*}$  provides a non-zero contribution. Or, if this block of  $A$  is zero, then the entries in the shaded region do not vanish.  $\square$

Let  $y \in \mathbb{R}^N$  outside the exceptional set from Theorem 2.4.5 be given. We will now show in detail how the values  $\gamma_i(y)$ ,  $i = 0, \dots, m(y)$ , from (2.20) are computed explicitly.

The MAP estimator for the Potts functional (2.16) is a constant signal, equal to

$$\bar{y} = \frac{1}{N} \sum_{t \in S} y_t, \quad (2.24)$$

if and only if  $f_y^k(\gamma) > f_y^1(\gamma)$  for all  $1 < k \leq N$ . This condition gives the following equivalent inequalities

$$\gamma \cdot (k-1) + \tilde{B}_y(k) > \tilde{B}_y(1) \quad \text{for all } 1 < k \leq N,$$

$$\begin{aligned}\gamma &> \frac{1}{k-1} \left( \tilde{B}_y(1) - \tilde{B}_y(k) \right) \quad \text{for all } 1 < k \leq N, \\ \gamma &> \gamma_0(y) := \max_{1 < k \leq N} \left( \frac{1}{k-1} \left( \tilde{B}_y(1) - \tilde{B}_y(k) \right) \right).\end{aligned}\quad (2.25)$$

Let  $k_0 = k_0(y)$  be the number  $k$  for which the right hand side of (2.25) is maximal. By Theorem 2.4.5, the one dimensional Potts functional (2.16) has exactly two minimizers for  $\gamma = \gamma_0(y)$ , namely the constant estimate and the estimate with  $k_0 - 1$  jumps. Note, that for  $\gamma > \gamma_0(y)$  the minimum function  $h_y$  is constant. It does not increase, independent of how large  $\gamma$  is.

For  $\gamma < \gamma_0(y)$  the MAP estimate  $x^*(\gamma, y)$  has  $k_0 - 1$  jumps if and only if  $f_y^k(\gamma) > f_y^{k_0}(\gamma)$  for all  $k_0 < k \leq N$ . This is equivalent to

$$\begin{aligned}\gamma \cdot (k-1) + \tilde{B}_y(k) &> \gamma \cdot (k_0-1) + \tilde{B}_y(k_0) \quad \text{for all } k_0 < k \leq N \\ \gamma &> \gamma_1(y) := \max_{k_0 < k \leq N} \left( \frac{1}{k-k_0} \left( \tilde{B}_y(k_0) - \tilde{B}_y(k) \right) \right).\end{aligned}$$

Let  $k_1 = k_1(y)$  be the  $k$  for which the maximum is attained. For  $\gamma = \gamma_1(y)$  the functional  $x \mapsto \bar{H}_\gamma(x, y)$  has exactly two minimizers, one minimizer with  $k_0 - 1$  jumps and one with  $k_1 - 1$  jumps. Hence,  $\gamma_1(y)$  is the next point in the  $\gamma$ -scanning. On the interval  $(\gamma_1(y), \gamma_0(y))$  the minimum function  $h_y$  is linear with slope  $k_0$ . This procedure will be continued until for some  $k_{m(y)} < N$  the number  $k = k(y)$  with  $k_{m(y)} < k \leq N$  which maximizes

$$\frac{1}{k - k_{m(y)}} \left( \tilde{B}_y(k_{m(y)}) - \tilde{B}_y(k) \right)$$

is equal to  $|J(y)| + 1$  which is at most  $N$ . We then have  $\gamma_{m(y)+1}(y) = 0$  and the whole range of the hyperparameter  $\gamma$  is exhausted. Thus, for each  $\gamma \in (0, \gamma_{m(y)}(y))$  the MAP estimator  $x^*(\gamma, y)$  is equal to data  $y$  and has  $|J(y)|$  jumps.

**Remark 2.4.8** The hyperparameter  $\gamma$  controls the number of jumps of the MAP estimator  $x^*(\gamma, y)$ . A large value of  $\gamma$  suppresses jumps. A small value of  $\gamma$  admits more jumps. This means that  $\gamma$  controls the ‘smoothness’ of  $x^*$ . Scanning  $x^*(\gamma, y)$  through the entire range of the hyperparameter reveals the complete potential of the Potts functionals. Plotting all estimates simultaneously is closely related to the *family approach* to the presentation of (parameterized) smoothers or kernel density estimators, proposed in J.S. MARRON and S.S. CHUNG (2001). Exemplarily, Figure 2.4 displays the MAP estimates for dotted data  $y$  for the first four subsequent  $\gamma$ -intervals, starting with  $(\gamma_0, \infty)$ . They are followed by the 7th, 9th, 24th, and the last of the 56  $\gamma$ -intervals. The complete plot of all estimates for this data can be found in Appendix A.

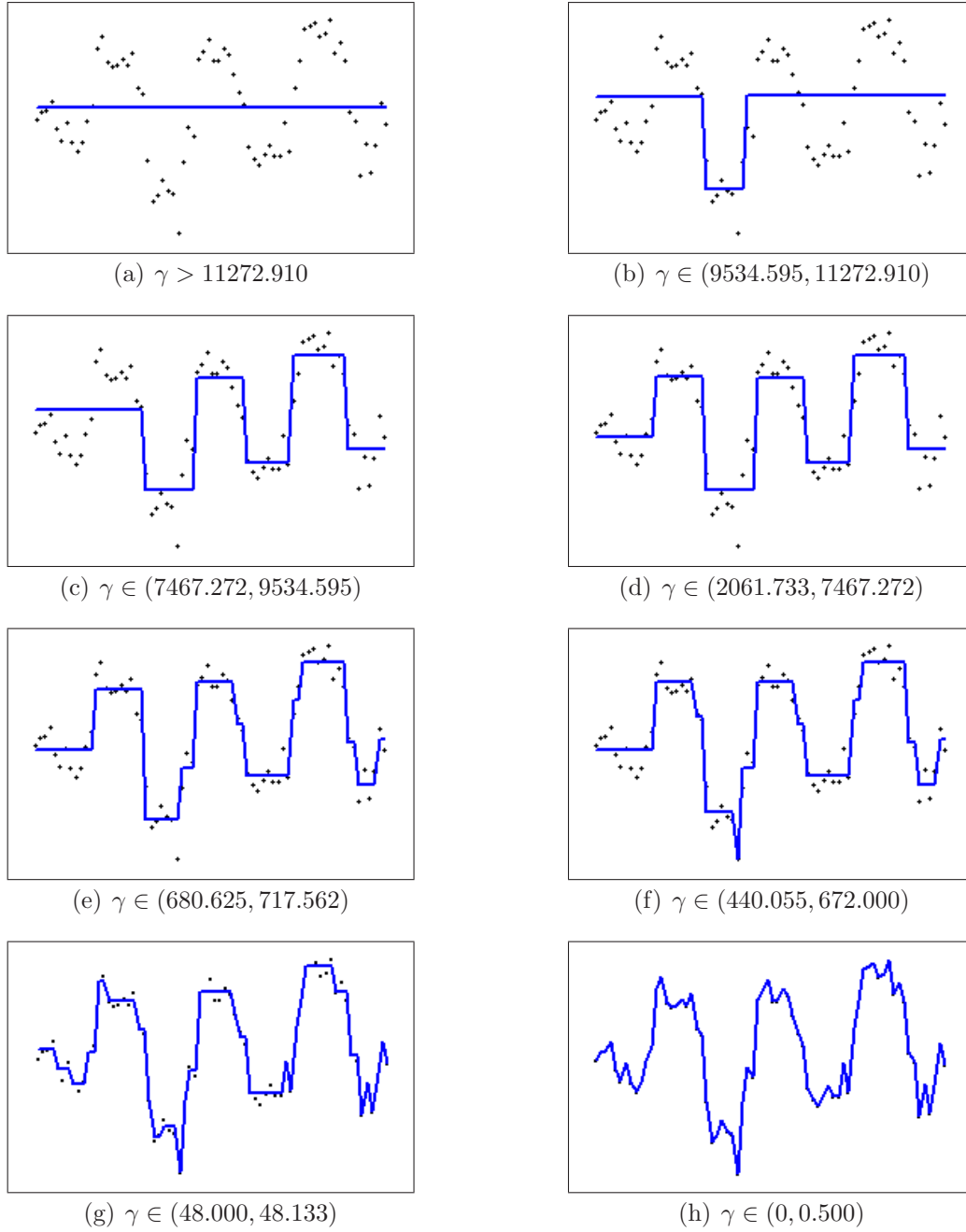


Figure 2.4: Some steps in a  $\gamma$ -scanning. Data are displayed as dots. The frames show MAP estimates of the 1st, 2nd, 3rd, 4th, 7th, 9th, 24th, and the last of the 56  $\gamma$ -intervals, starting with  $(\gamma_0, \infty)$ .

**Remark 2.4.9** Obviously, there is a tradeoff between fidelity to data and smoothness or parsimony in terms of the number of jumps.

**Remark 2.4.10** If  $\gamma$  decreases from infinity to zero the number  $k$  of intervals in the partition induced by the MAP estimator does not necessarily take each value from 1 to  $N - 1$ . This is due to the fact that adding more than one jump may be more profitable than adding only one jump.

## 2.5 Continuity

In this section, we study joint continuity of minimizers  $x^*(\gamma, y)$  in the variables  $(\gamma, y)$ . Recall, that for fixed  $\gamma$  the MAP estimator is unique for almost all  $y \in \mathbb{R}^S$  by Theorem 2.3.1.

Continuity of  $x^*(\gamma, y)$  in  $(\gamma, y)$  implies that for each  $(\gamma', y')$  with unique minimizer there is a neighborhood such that the minimizer is unique for all  $(\gamma, y)$  in this neighborhood as well. In particular, the induced partition is the same and hence the number of jumps of the MAP estimator is constant in a neighborhood of a unique minimizer. For the proof of the main theorem we need some preparatory steps. For each  $\mathcal{P} \in \mathfrak{P}$  consider  $\bar{y}_{\mathcal{P}}$  as an element of  $\mathbb{R}^S$  by the natural identification

$$\bar{y}_{\mathcal{P}} = \underbrace{(\bar{y}_{I_1}, \dots, \bar{y}_{I_1})}_{|I_1|}, \underbrace{(\bar{y}_{I_2}, \dots, \bar{y}_{I_2})}_{|I_2|}, \dots, \underbrace{(\bar{y}_{I_{|\mathcal{P}|}}, \dots, \bar{y}_{I_{|\mathcal{P}|}})}_{|I_{|\mathcal{P}|}})^t.$$

Let

$$\langle y, x \rangle = \sum_{s \in S} y_s x_s$$

be the Euclidian inner product on  $\mathbb{R}^S$  and

$$\|y - x\|^2 = \sum_{s \in S} (y_s - x_s)^2$$

the induced norm. For  $\mathcal{P} \in \mathfrak{P}$  we define the functions

$$f_{\mathcal{P}} : (0, \infty) \times \mathbb{R}^S \longrightarrow \mathbb{R}, \quad (\gamma, y) \longmapsto \gamma(|\mathcal{P}| - 1) + \|y - \bar{y}_{\mathcal{P}}\|^2. \quad (2.26)$$

They are continuous in  $\gamma$  and  $y$  and there are finitely many since  $\mathfrak{P}$  is a finite set. First, we will prove continuity of  $f_{\mathcal{P}}(\gamma, y)$ .

**Lemma 2.5.1** *Choose  $\mathcal{P} \in \mathfrak{P}$ . Then  $f_{\mathcal{P}}$  from (2.26) is continuous with respect to the usual topology on  $(0, \infty) \times \mathbb{R}^S$ .*

**Proof** Let  $(\gamma, y), (\gamma', y') \in (0, \infty) \times \mathbb{R}^S$ , and  $\mathcal{P} \in \mathfrak{P}$ . Then

$$|f_{\mathcal{P}}(\gamma, y) - f_{\mathcal{P}}(\gamma', y')| \leq |\gamma - \gamma'|(|\mathcal{P}| - 1) + \left| \|y - \bar{y}_{\mathcal{P}}\|^2 - \|y' - \bar{y}'_{\mathcal{P}}\|^2 \right|.$$

For the last term we have

$$\begin{aligned} & \left| \|y - \bar{y}_{\mathcal{P}}\|^2 - \|y' - \bar{y}'_{\mathcal{P}}\|^2 \right| \\ &= \left| \langle y - \bar{y}_{\mathcal{P}}, y - \bar{y}_{\mathcal{P}} \rangle - \langle y - \bar{y}_{\mathcal{P}}, y' - \bar{y}'_{\mathcal{P}} \rangle \right. \\ & \quad \left. + \langle y - \bar{y}_{\mathcal{P}}, y' - \bar{y}'_{\mathcal{P}} \rangle - \langle y' - \bar{y}'_{\mathcal{P}}, y' - \bar{y}'_{\mathcal{P}} \rangle \right| \\ &\leq \left| \langle y - \bar{y}_{\mathcal{P}}, y - \bar{y}_{\mathcal{P}} - y' + \bar{y}'_{\mathcal{P}} \rangle \right| \\ & \quad + \left| \langle y - \bar{y}_{\mathcal{P}} - y' + \bar{y}'_{\mathcal{P}}, y' - \bar{y}'_{\mathcal{P}} \rangle \right| \\ &\leq \left( \|y - \bar{y}_{\mathcal{P}}\| + \|y' - \bar{y}'_{\mathcal{P}}\| \right) \cdot \left( \|\bar{y}'_{\mathcal{P}} - \bar{y}_{\mathcal{P}}\| + \|y - y'\| \right). \end{aligned}$$

With  $\bar{y}_{\{S\}} = \bar{y}$  given by  $\bar{y}_s = 1/N \sum_{t \in S} y_t$  for all  $s \in S$  we get it is

$$\|y - \bar{y}_{\mathcal{P}}\| \leq \|y - \bar{y}\|.$$

Since  $y \mapsto \bar{y}_{\mathcal{P}}$  is a projection and hence a contraction, we have

$$\|\bar{y}'_{\mathcal{P}} - \bar{y}_{\mathcal{P}}\| \leq \|y' - y\|.$$

For the sum in the first brackets we then get

$$\begin{aligned} \|y - \bar{y}_{\mathcal{P}}\| + \|y' - \bar{y}'_{\mathcal{P}}\| &\leq \|y - \bar{y}\| + \|y' - \bar{y}'\| \\ &\leq \|y - y'\| + \|y' - \bar{y}'\| + \|\bar{y}' - \bar{y}\| + \|y' - \bar{y}'\| \\ &\leq 2\|y - y'\| + 2\|y' - \bar{y}'\|. \end{aligned}$$

In summary, we arrive at

$$\begin{aligned} \left| \|y - \bar{y}_{\mathcal{P}}\|^2 - \|y' - \bar{y}'_{\mathcal{P}}\|^2 \right| &\leq (2\|y - y'\| + 2\|y' - \bar{y}'\|) \cdot 2\|y' - y\| \\ &= 4\|y - y'\|^2 + 4\|y' - \bar{y}'\| \cdot \|y - y'\|. \end{aligned}$$

Hence, with the constants  $c_1 = |\mathcal{P}| - 1$ ,  $c_2 = 4$ , and  $c_3 = 4\|y' - \bar{y}'\|$  we have the inequality

$$|f_{\mathcal{P}}(\gamma, y) - f_{\mathcal{P}}(\gamma', y')| \leq c_1|\gamma - \gamma'| + c_2\|y - y'\|^2 + c_3\|y - y'\|.$$

Thus, for each  $\varepsilon > 0$  there is  $\delta > 0$  with

$$\delta < \delta(\varepsilon) := \frac{1}{2c_2} \left( \sqrt{(c_1 + c_3)^2 + 4c_2\varepsilon} - (c_1 + c_3) \right)$$

such that

$$|f_{\mathcal{P}}(\gamma, y) - f_{\mathcal{P}}(\gamma', y')| < \varepsilon \quad \text{for all } (\gamma, y) \in U_{\delta}(\gamma', y')$$

where

$$U_{\delta}(\gamma', y') := \{(\gamma, y) \in (0, \infty) \times \mathbb{R}^S : |\gamma' - \gamma| < \delta, \|y' - y\| < \delta\}.$$

This proves the assertion.  $\square$

The main result of this section is the following theorem.

**Theorem 2.5.2** *Suppose that  $(\gamma', y') \in (0, \infty) \times \mathbb{R}^S$  is such that the MAP estimator  $x^*(\gamma', y')$  of the Potts functionals*

$$\bar{H}_{\gamma}(\cdot, y) : \mathbb{X} \longrightarrow \mathbb{R}, \quad x \longmapsto \gamma \cdot |J(x)| + \sum_{s \in S} (y_s - x_s)^2, \quad (2.27)$$

*is unique. Then there is a neighborhood of  $(\gamma', y')$  such that*

$$X^*(\gamma, y) = \{x^*(\gamma', y')\}$$

*for all  $(\gamma, y)$  in this neighborhood. In particular, the mapping*

$$x^* : (0, \infty) \times \mathbb{R}^S \longrightarrow \mathbb{X}, \quad (\gamma, y) \longmapsto x^*(\gamma', y')$$

*is continuous in a neighborhood of  $(\gamma', y')$ .*

**Proof** Let  $\mathcal{P}^*$  be the partition induced by  $x^*(\gamma', y')$ . By Lemma 2.5.1, the function  $f_{\mathcal{P}}$  is continuous for fixed  $\mathcal{P}$ . Hence there is a neighborhood  $U(\mathcal{P})$  of  $(\gamma', y')$  such that

$$f_{\mathcal{P}^*}|_{U(\mathcal{P})} < f_{\mathcal{P}}|_{U(\mathcal{P})},$$

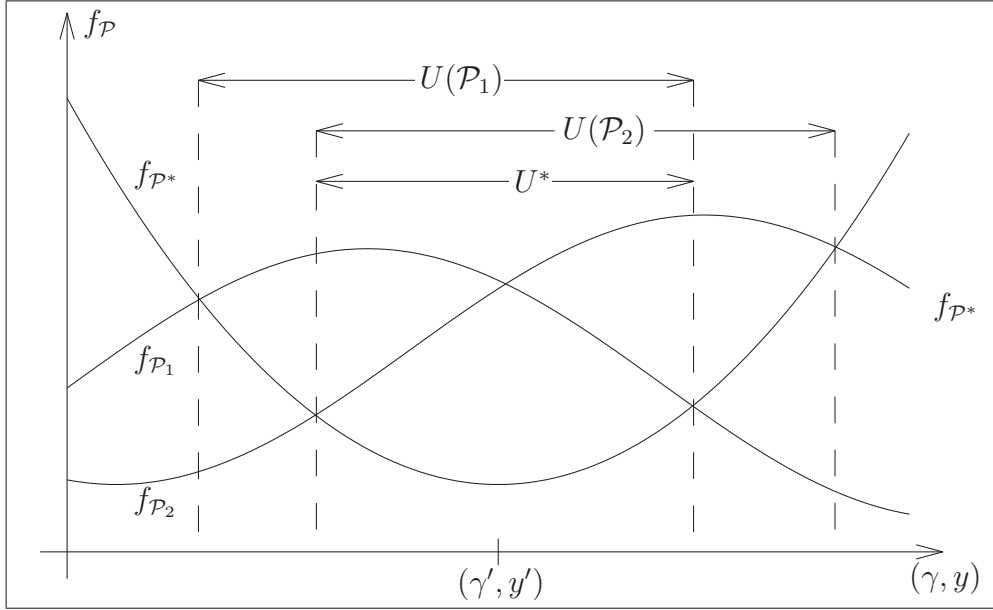
as illustrated in Figure 2.5.

We define

$$U^* := \bigcap_{\mathcal{P} \in \mathfrak{P}, \mathcal{P} \neq \mathcal{P}^*} U(\mathcal{P}).$$

The intersection is finite and hence  $U^*$  is open. On  $U^*$ , it is  $f_{\mathcal{P}^*} < f_{\mathcal{P}}$  for all  $\mathcal{P} \neq \mathcal{P}^*$ . Therefore the minimizing partition is the same in a neighborhood of  $(\gamma', y')$ , namely  $\mathcal{P}^*$ . Obviously also the number  $|J(x^*(\gamma, y))| = |\mathcal{P}^*| - 1$  of jumps is constant for all  $(\gamma, y)$  in  $U^*$ .  $\square$

This proves that the MAP estimator  $x^*(\gamma, y)$  of (2.27) is unique for  $(\gamma, y)$  in a neighborhood of  $(\gamma', y')$  with unique MAP estimator.

Figure 2.5: The neighborhoods  $U(\mathcal{P})$ 

**Corollary 2.5.3** *Let  $(\gamma', y')$  be such that the minimizer  $x^*(\gamma', y')$  of (2.27) is unique. Then there is a neighborhood of  $(\gamma', y')$  such that the minimizer  $x^*(\gamma, y)$  is unique for all  $(\gamma, y)$  in this neighborhood.*

In addition, the proof of Theorem 2.5.2 shows that the partition induced by the MAP estimator  $x^*(\gamma, y)$ , and therefore also the number of jumps, do not change.

**Corollary 2.5.4** *Let  $(\gamma', y')$  be a pair of hyperparameter and data for which the minimizer  $x^*(\gamma', y')$  of (2.27) is unique. Then there is a neighborhood of  $(\gamma', y')$  such that the partition  $\mathcal{P}(x^*(\gamma, y))$  and the number of jumps of  $x^*(\gamma, y)$  are constant in this neighborhood.*

**Remark 2.5.5** The number of jumps is stable - and even constant - under variation of data  $y$  in a neighborhood of  $(\gamma', y')$  for which the minimizer  $x^*(\gamma', y')$  is unique. Hence, the MAP estimator is an adequate instrument to measure the number of jumps.

## 2.6 Measurable Section

Recall that MAP estimators of the Potts functionals

$$\bar{H}_\gamma(\cdot, y) : \mathbb{X} \longrightarrow \mathbb{R}, \quad x \longmapsto \gamma \cdot |J(x)| + \sum_{s \in S} (y_s - x_s)^2 \quad (2.28)$$

are in general not unique. In this section, we will show that there is a measurable section of the set-valued map  $(\gamma, y) \mapsto X^*(\gamma, y)$ .

Let  $(\Omega, \mathfrak{F})$  and  $(X, \mathfrak{A})$  be measurable spaces. Let  $\mathfrak{p}(X)$  denote the power set of  $X$  and let

$$\Phi : \Omega \longrightarrow \mathfrak{p}(X).$$

A *measurable section* of  $\Phi$  is a map

$$\varphi : \Omega \longrightarrow X$$

which is  $\mathfrak{F}$ - $\mathfrak{A}$ -measurable and which fulfills

$$\varphi(\omega) = \Phi(\omega) \quad \text{for every } \omega \in \Omega.$$

**Proposition 2.6.1** *Let  $X^*(\gamma, y)$  be the set of minimizers of (2.28). Then there is a measurable section of the set-valued map  $(\gamma, y) \mapsto X^*(\gamma, y)$ .*

**Proof** Order the set of all partitions of  $S$  for example lexicographically with respect to the length of the intervals to get the ordered set  $\{\mathcal{P}_1, \mathcal{P}_2, \dots, \mathcal{P}_{2^{N-1}}\}$ . Recall that by Theorem 1.2.4 each minimizer  $x^*(\gamma, y)$  of (2.28) can be identified with a minimal segmentation  $(\mathcal{P}(y), \bar{y}_{\mathcal{P}(y)})$ . For a fixed partition  $\mathcal{P}_j$  the corresponding  $\bar{y}_{\mathcal{P}_j}$  is the orthogonal projection of  $y$  to the subspace  $\{z \in \mathbb{R}^S : \mathcal{P}(z) = \mathcal{P}_j\}$ . Define

$$i(\gamma, y) := \min \left\{ j \in \{1, \dots, 2^{N-1}\} : \bar{H}_\gamma(\bar{y}_{\mathcal{P}_j}, y) = \min_{1 \leq k \leq 2^{N-1}} \bar{H}_\gamma(\bar{y}_{\mathcal{P}_k}, y) \right\}.$$

The symbol  $\mathcal{B}(X)$  will denote the  $\sigma$ -field of the Borel subsets of  $X$ . The map

$$i : [0, \infty) \times \mathbb{R}^S \longrightarrow \{1, \dots, 2^{N-1}\}, \quad (\gamma, y) \longmapsto i(\gamma, y)$$

is  $\mathcal{B}([0, \infty)) \otimes \mathcal{B}(\mathbb{R}^S) - \mathfrak{P}(\{1, \dots, 2^{N-1}\})$ -measurable since

$$\begin{aligned} \{i(\gamma, y) = l\} &= \{(\gamma, y) \in [0, \infty) \times \mathbb{R}^S : \bar{H}_\gamma(\bar{y}_{\mathcal{P}_l}, y) < \bar{H}_\gamma(\bar{y}_{\mathcal{P}_k}, y) \text{ for } k < l \\ &\quad \text{and } \bar{H}_\gamma(\bar{y}_{\mathcal{P}_l}, y) \leq \bar{H}_\gamma(\bar{y}_{\mathcal{P}_k}, y) \text{ for } k > l\} \\ &= \bigcup_{k < l} \{\bar{H}_\gamma(\bar{y}_{\mathcal{P}_l}, y) < \bar{H}_\gamma(\bar{y}_{\mathcal{P}_k}, y)\} \cup \bigcup_{k > l} \{\bar{H}_\gamma(\bar{y}_{\mathcal{P}_l}, y) \leq \bar{H}_\gamma(\bar{y}_{\mathcal{P}_k}, y)\} \end{aligned}$$

and the map  $(\gamma, y) \mapsto \bar{H}_\gamma(\bar{y}_{\mathcal{P}_j}, y)$  is  $\mathcal{B}([0, \infty)) \otimes \mathcal{B}(\mathbb{R}^S) - \mathcal{B}(\mathbb{R})$ -measurable for each  $j$ . Hence the map  $(\gamma, y) \mapsto i(\gamma, y) \mapsto (\mathcal{P}_i, \bar{y}_{\mathcal{P}_i})$  is a composition of measurable maps and hence a measurable section of  $(\gamma, y) \mapsto X^*(\gamma, y)$ .  $\square$

# Chapter 3

## Exact Optimization

In Section 2.1, the minimization of the functional  $(\mathcal{P}, \mu_{\mathcal{P}}) \mapsto H((\mathcal{P}, \mu_{\mathcal{P}}), y)$  was split up into two steps. This allows to develop *exact* efficient algorithms for the computation of MAP estimates in one dimension.

These exact algorithms eliminate all the uncertainties of popular Markov Chain Monte Carlo methods like Simulated Annealing. This enables us to study the Potts functional in detail and rigor. Only with exact optimization, we can tell artefacts and effects due to modelling from those caused by suboptimal optimization. This is indispensable for a rigorous validation of methods. For a discussion how ‘incorrect modelling’ along with an ‘incorrect algorithm’ can even produce desired results see G. WINKLER (2003), Chapter 6.2, in particular Figure 6.4.

For fixed hyperparameter, an exact algorithm is derived in Section 3.1. In Section 3.2, it is extended to an algorithm for the computation of estimates for *all* values of the hyperparameter *simultaneously*. Both were sketched in G. WINKLER and V. LIEBSCHER (2002). We will work out the details and give proofs of the recursive formulae.

Assume that  $S = \{1, \dots, N\}$  is endowed with nearest neighbor structure and let data  $y \in \mathbb{X}$  be given. Assume further that the function  $\rho$  in the data term of the Potts functional satisfies the Hypotheses 1.2.5 and 2.2.1.

If  $\mu_{\mathcal{P}}^*$  and  $D((\mathcal{P}, \mu_{\mathcal{P}}^*), y)$  can be computed in  $O(g(N))$ , the algorithm for fixed hyperparameter  $\gamma$  is of complexity  $O(\max\{g(N), N^2\})$  and that one for all  $\gamma$  is of complexity  $O(\max\{g(N), N^3\})$ .

To simplify notation, and since  $y$  is fixed, we will write  $H_{\mathcal{P}}(\mu_{\mathcal{P}})$  and  $D_{\mathcal{P}}(\mu_{\mathcal{P}})$  instead of  $H((\mathcal{P}, \mu_{\mathcal{P}}), y)$  and  $D((\mathcal{P}, \mu_{\mathcal{P}}), y)$ , respectively, throughout this chapter.

### 3.1 Minimization for Fixed Hyperparameter

In this section we assume that the hyperparameter  $\gamma$  is fixed. Let  $\mathfrak{P}^n$  denote the set of all partitions of  $\{1, \dots, n\}$ . Let now  $(\mu_I^*)_{I \in \mathcal{P}}$  for  $\mathcal{P} \in \mathfrak{P}^n$  for any  $1 \leq n \leq N$ , be a family of minimizers given by (2.3). For each  $I \in \mathcal{P}$  define the *interval function*  $H_I$  by

$$H_I(\mu_I^*) = \gamma + \sum_{s \in I} \rho(y_s - \mu_I^*). \quad (3.1)$$

We will use the notation  $[s, t]$  for the interval  $\{s, s+1, \dots, t-1, t\}$ . The symbol  $[r, r]$  denotes the singleton  $\{r\}$ .

**Lemma 3.1.1** *We define for each  $1 \leq n \leq N$  the function*

$$B(n) = \min_{\mathcal{P} \in \mathfrak{P}^n} H_{\mathcal{P}}(\mu_{\mathcal{P}}^*). \quad (3.2)$$

*Let further  $H_I$  be the interval function from (3.1), and set  $B(0) = -\gamma$ ,  $B(1) = 0$ . Then the recursive formula*

$$B(n) = \min_{0 \leq r \leq n-1} \left( B(r) + \min_{\mu \in \mathbb{R}} H_{[r+1, n]}(\mu) \right), \quad 1 < n \leq N \quad (3.3)$$

*holds. A partition minimizes  $H_{\mathcal{P}}(\mu_{\mathcal{P}}^*)$  in  $\mathcal{P} \in \mathfrak{P}^n$  if and only if it is the union of  $[r^* + 1, n]$  and a partition minimizing  $H_{\mathcal{P}}(\mu_{\mathcal{P}}^*)$  in  $\mathcal{P} \in \mathfrak{P}^{r^*}$  where  $r^*$  is a minimizer from (3.3).*

**Proof** Let  $\mathcal{P}_0$  denote the empty set. For  $r \leq n-1$ , the map

$$p : \mathfrak{P}^r \rightarrow \mathfrak{P}^n, \quad \mathcal{P} \mapsto \mathcal{P} \cup [r+1, n]$$

is one-to-one and  $p(\mathfrak{P}^r) = \{\mathcal{P} \in \mathfrak{P}^n : [r+1, n] \in \mathcal{P}\}$ . Then

$$B(n) = \min_{\mathcal{P} \in \mathfrak{P}^n} H_{\mathcal{P}}(\mu_{\mathcal{P}}^*) = \min_{0 \leq r \leq n-1} \left( \min_{\{\mathcal{P} \in \mathfrak{P}^n : [r+1, n] \in \mathcal{P}\}} H_{\mathcal{P}}(\mu_{\mathcal{P}}^*) \right).$$

If  $[r+1, n] \in \mathcal{P}$  the functional  $H_{\mathcal{P}}(\mu_{\mathcal{P}}^*)$  can be split into two terms:

$$\begin{aligned} H_{\mathcal{P}}(\mu_{\mathcal{P}}^*) &= -\gamma + \sum_{I \in \mathcal{P}} H_I(\mu_I^*) \\ &= -\gamma + \sum_{I \in \mathcal{P}, I \neq [r+1, n]} H_I(\mu_I^*) + H_{[r+1, n]}(\mu_{[r+1, n]}^*) \\ &= H_{\mathcal{P} \setminus [r+1, n]}(\mu_{\mathcal{P} \setminus [r+1, n]}^*) + H_{[r+1, n]}(\mu_{[r+1, n]}^*). \end{aligned}$$

Since the second part does not depend on  $\mathcal{P}$ , we get

$$\begin{aligned}
B(n) &= \min_{0 \leq r \leq n-1} \left( H_{[r+1,n]}(\mu_{[r+1,n]}^*) \right. \\
&\quad \left. + \min_{\{\mathcal{P} \in \mathfrak{P}^n : [r+1,n] \in \mathcal{P}\}} \left( H_{\mathcal{P} \setminus [r+1,n]}(\mu_{\mathcal{P} \setminus [r+1,n]}^*) \right) \right) \\
&= \min_{0 \leq r \leq n-1} \left( H_{[r+1,n]}(\mu_{[r+1,n]}^*) + \min_{\mathcal{P} \in \mathfrak{P}^r} H_{\mathcal{P}}(\mu_{\mathcal{P}}^*) \right) \\
&= \min_{0 \leq r \leq n-1} \left( H_{[r+1,n]}(\mu_{[r+1,n]}^*) + B(r) \right)
\end{aligned}$$

which is the desired recursive formula. The assertion for the minimizers is obvious.  $\square$

Note that  $B(N)$  is the minimum of the one dimensional Potts functional for fixed hyperparameter  $\gamma$ .  $B$  is a *Bellman function* in the spirit of R. E. BELLMAN (1957). We adapt the dynamic programming technique to the minimization problem (2.2) for fixed  $\gamma$ .

**Algorithm 3.1.2** The algorithm for the computation of a MAP estimator of the one dimensional Potts functional for fixed  $\gamma$  reads as follows:

- (1) Compute for each interval  $[r, s]$ ,  $1 \leq r < s \leq N$ , a minimizer  $\mu_{[r,s]}^* \in \operatorname{argmin}_{\mu \in \mathbb{R}} H_{[r,s]}(\mu)$ , and the value  $H_{[r,s]}(\mu_{[r,s]}^*)$ . Set  $H_{[r,r]}(\mu_{[r,r]}^*) = \gamma$  for  $1 \leq r \leq N$ .
- (2) Set  $B(0) = -\gamma$  and  $B(1) = 0$ .
- (3) Determine recursively  $B(n)$  for all  $1 < n \leq N$  using (3.3) and store *at least* one  $r_n^* \in \{0, \dots, n-1\}$  for which the minimum in (3.3) is attained.
- (4) Construct a partition of  $\{1, \dots, N\}$  recursively from the right: The minimizer  $r_N^*$  of  $B(N)$  becomes the last point of the last but one interval. The rightmost interval of the partition then is  $[r_N^* + 1, N]$ . Take now a minimizer of  $B(r_N^*)$  to get the next interval to the left of  $[r_N^* + 1, N]$ . Continuing with this procedure gives a partition  $\{I_1, \dots, I_k\}$  of  $\{1, \dots, N\}$  and  $\mu_{\mathcal{P}}^*$  is given by the family of minimizers  $\mu_{I_l}^*$ ,  $l = 1, \dots, k$ .

By this algorithm we obtain a minimizer of the one dimensional Potts functional (1.5).

**Theorem 3.1.3** *The output of Algorithm 3.1.2 is a minimizer  $x^*(\gamma, y)$  of the functional  $x \mapsto H_\gamma(x, y)$ .*

**Proof** In step (4) of Algorithm 3.1.2 we get a partition  $\mathcal{P}^* = \{I_1, \dots, I_k\} \in \mathfrak{P}$  and  $x^* \in \operatorname{argmin}_{x \in \mathbb{X}} H_\gamma(x, y)$  is determined by  $(\mu_I^*)_{I \in \mathcal{P}^*}$ . By definition  $(\mu_I^*)_{I \in \mathcal{P}^*}$  is a minimizer of  $H_{\mathcal{P}^*}(\mu_{\mathcal{P}^*})$  for fixed partition  $\mathcal{P}^*$ . It remains to show that  $\mathcal{P}^*$  minimizes  $H_{\mathcal{P}}(\mu_{\mathcal{P}}^*)$  in  $\mathcal{P} \in \mathfrak{P}$ . We take

$$r_N^* = \operatorname{argmin}_{0 \leq r \leq N-1} \left( B(r) + H_{[r+1, N]}(\mu_{[r+1, N]}^*) \right)$$

to get the rightmost interval  $I_k = [r_N^* + 1, N] \in \mathcal{P}^*$ . By definition of  $r_N^*$  a partition  $(\mathcal{P}^{r_N^*})^* \cup [r_N^* + 1, N]$  minimizes  $H_{\mathcal{P}}(\mu_{\mathcal{P}}^*)$  in  $\mathcal{P} \in \mathfrak{P}$  where  $(\mathcal{P}^r)^* \in \mathfrak{P}^r$  denote a partition which minimizes  $\mathcal{P}^r \mapsto H_{\mathcal{P}^r}(\mu_{\mathcal{P}^r}^*)$  for fixed  $r$ . By Lemma 3.1.1 we obtain the respective last interval of such  $(\mathcal{P}^r)^* \in \mathfrak{P}^r$ . The collection of these intervals is a minimizing partition in  $\mathfrak{P}$ .  $\square$

The function  $f$  which assigns to the length  $N$  of an input the number of basic operations which are necessary to perform the algorithm is called the *complexity* of the algorithm. Basic operations are basic arithmetic operations as summation, subtraction, multiplication, and division, assignments as the fixing of  $\mu_I^*$ , and logical questions as in IF-loops. The expression  $f(N) = O(g(N))$  with the *Landau symbol*  $O$  means that  $|f(N)| \leq c|g(N)|$  for some constant  $c > 0$ .

**Theorem 3.1.4** *If, for each partition  $\mathcal{P} \in \mathfrak{P}$ , a solution  $(\mu_I^*)_{I \in \mathcal{P}}$  of (2.3) and minimum values  $H_I(\mu_I^*)$ ,  $I \in \mathcal{P}$ , can be computed in  $O(g(N))$ , then Algorithm 3.1.2 works in complexity  $O(\max\{g(N), N^2\})$ .*

**Proof** In step (1) of the algorithm we compute minimizers  $\mu_I^*$  and minimum values  $H_I(\mu_I^*)$  which is of complexity  $O(g(N))$  according to the assumption. In step (2) we set two values which has complexity  $O(1)$ . For each  $B(n)$  we compute  $n$  values and perform  $n - 1$  comparisons. The recursive computation of  $B(n)$  for  $1 < n \leq N$  needs  $N(N - 1)/2 - 1$  basic operations which gives complexity  $O(N^2)$  for step (3). The partition is given by the assignment of  $|\mathcal{P}|$  values. Since  $\mathcal{P}$  has at most  $N$  intervals, this results in complexity  $O(N)$  for step (4). Thus, the algorithm has complexity  $O(\max\{g(N), N^2\})$ .  $\square$

**Corollary 3.1.5** *For  $\rho(u) = u^2$  the Algorithm 3.1.2 has complexity  $O(N^2)$ .*

**Proof** We have to show that the computation of minimizers and minimum values of the interval function  $H_I$  for all intervals  $I \subset \{1, \dots, N\}$  is in  $O(N^2)$ . For  $\rho(u) = u^2$ , the minimizer  $\mu_I^*$  of an interval  $I$  is given by the empirical mean

$$\bar{y}_I = \frac{1}{|I|} \sum_{s \in I} y_s.$$

First, set  $\bar{y}_{[r,r]} = y_r$ ,  $r = 1, \dots, N$ . Compute the minimizers of all intervals  $[1, r+1]$ ,  $r = 1, \dots, N-1$ , by

$$\bar{y}_{[1,r+1]} = \frac{1}{r+1} \sum_{s=1}^{r+1} y_s = \frac{r}{r+1} \bar{y}_{[1,r]} + \frac{1}{r+1} y_{r+1}$$

which is in  $O(N)$ . Then the minimizers of the remaining  $O(N^2)$  intervals  $[r, s]$ ,  $2 \leq r < s \leq N$  can be obtained via the formula

$$\bar{y}_{[r,s]} = \frac{1}{s-r+1} \sum_{t=r}^s y_t = \frac{s}{s-r+1} \bar{y}_{[1,s]} - \frac{r-1}{s-r+1} \bar{y}_{[1,r-1]}.$$

The minimum values of the interval function are given by

$$H_I(\bar{y}_{\mathcal{P}}) = \gamma + \sum_{s \in I} (y_s - \bar{y}_I)^2 = \gamma + \sum_{s \in I} y_s^2 - |I| \bar{y}_I^2.$$

We compute for all intervals  $[1, r+1]$ ,  $r = 1, \dots, N-1$ , the sum of squares by

$$\sum_{s=1}^{r+1} y_s^2 = \sum_{s=1}^r y_s^2 + y_{r+1}^2$$

in  $O(N)$  and for all intervals  $[r, s]$ ,  $2 \leq r \leq s \leq N$  we have

$$\sum_{t=r}^s y_t^2 = \sum_{t=1}^s y_t^2 - \sum_{t=1}^{r-1} y_t^2.$$

Thus, the computation of the minimizers and the minimum values of the interval function for all intervals needs  $O(N^2)$  basic operations.  $\square$

## 3.2 Simultaneous Minimization in $\gamma$

The computation of MAP estimates  $x^*(\gamma, y)$  of the one dimensional Potts functional for all values of the hyperparameter  $\gamma$  simultaneously is an extension of Algorithm 3.1.2. Define, similarly to the interval function  $H_I$ , an *interval error function*  $D_I$  by

$$D_I(\mu) = H_I(\mu) - \gamma. \quad (3.4)$$

Note that the partition error function  $D_{\mathcal{P}}$  for a partition  $\mathcal{P} \in \mathfrak{P}$  is the data term from (1.4).

**Lemma 3.2.1** Let  $\mathfrak{P}_k^n$  denote the set of all partitions of  $\{1, \dots, n\}$  with  $|\mathcal{P}| = k$ . We define for  $1 \leq k \leq n \leq N$  the function

$$\tilde{B}(k, n) = \min_{\mathcal{P} \in \mathfrak{P}_k^n} D_{\mathcal{P}}(\mu_{\mathcal{P}}^*).$$

Let further  $D_I$  be the interval error function from (3.4), and set  $\tilde{B}(1, n) = D_{[1, n]}(\mu_{[1, n]}^*)$ ,  $1 \leq n \leq N$ . Then for  $1 < k \leq n \leq N$  the recursive formula

$$\tilde{B}(k, n) = \min_{1 \leq r \leq n-1} \left( \tilde{B}(k-1, r) + \min_{\mu \in \mathbb{R}} D_{[r+1, n]}(\mu) \right) \quad (3.5)$$

holds. A partition with  $k$  intervals minimizes  $H_{\mathcal{P}}(\mu_{\mathcal{P}}^*)$  in  $\mathcal{P} \in \mathfrak{P}^n$  if and only if it is the union of  $[r^* + 1, n]$  and a partition with  $k-1$  intervals minimizing  $H_{\mathcal{P}}(\mu_{\mathcal{P}}^*)$  in  $\mathcal{P} \in \mathfrak{P}^{r^*}$  where  $r^*$  is a minimizer from (3.3).

**Proof** The proof is basically the same as that of Lemma 3.1.1. Note that

$$D_{\mathcal{P}}(\mu_{\mathcal{P}}^*) = D_{\mathcal{P} \setminus [r+1, n]}(\mu_{\mathcal{P} \setminus [r+1, n]}^*) + D_{[r+1, n]}(\mu_{[r+1, n]}^*)$$

if  $[r+1, n] \in \mathcal{P} \in \mathfrak{P}^n$ . □

By Corollary 2.1.3, the minimization problem (2.1) can be rewritten as

$$\begin{aligned} \min_{x \in \mathbb{X}} H(x, y) &= \min_{\mathcal{P} \in \mathfrak{P}} \left( \gamma \cdot (|\mathcal{P}| - 1) + D_{\mathcal{P}}(\mu_{\mathcal{P}}^*) \right) \\ &= \min_{1 \leq k \leq N} \left( \gamma \cdot (k - 1) + \min_{\mathcal{P} \in \mathfrak{P}_k} D_{\mathcal{P}}(\mu_{\mathcal{P}}^*) \right). \end{aligned}$$

The minimum energy function  $h_y$  is defined as

$$h_y(\gamma) = \min_{x \in \mathbb{X}} H_{\gamma}(x, y) = \min_{1 \leq k \leq N} \left( \gamma \cdot (k - 1) + \tilde{B}(k, N) \right). \quad (3.6)$$

The functions  $h_y$  in case of  $\rho(u) = u^2$  were discussed in Section 2.4.

**Algorithm 3.2.2** The algorithm to compute MAP estimates of the one dimensional Potts functional for all values of  $\gamma$  simultaneously reads as follows:

- (1) Compute for all intervals  $[r, s]$ ,  $1 \leq r < s \leq N$ , a minimizer  $\mu_{[r, s]}^* \in \operatorname{argmin}_{\mu \in \mathbb{R}} D_{[r, s]}(\mu)$ , and the value  $D_{[r, s]}(\mu_{[r, s]}^*)$ . Set  $D_{[r, r]}(\mu_{[r, r]}^*) = 0$  for  $1 \leq r \leq N$ .
- (2) Set  $\tilde{B}(1, n) = D_{[1, n]}(\mu_{[1, n]}^*)$ ,  $1 \leq n \leq N$ .
- (3) Determine recursively  $\tilde{B}(k, n)$  by (3.5) for all  $1 < k \leq n \leq N$  and store at least one  $r_{k, n}^*$  for which the minimum of  $\tilde{B}(k, n)$  in (3.5) is attained.

- (4) Construct recursively partitions  $\{I_1^k, \dots, I_k^k\}$  of  $\{1, \dots, N\}$  from minimizers of  $\tilde{B}(k, N)$ ,  $\tilde{B}(k-1, r_{k,N}^*)$ ,  $\dots$  and  $\mu_{\mathcal{P}}^*$  is given by  $\mu_{I_l^k}^*$ ,  $l = 1, \dots, k$ .
- (5) Construct the piecewise linear function  $h_y(\gamma)$  from (3.6).

**Theorem 3.2.3** *Algorithm 3.2.2 computes minimizers  $x^*(\gamma, y)$  of the one dimensional Potts functional  $H_\gamma(x, y)$  for all  $\gamma$  simultaneously.*

**Proof** A fixed number  $k = |\mathcal{P}|$  corresponds to a fixed  $\gamma$ -interval and the assertion follows from Theorem 3.1.3. It was shown in Theorem 2.4.5 that we indeed get MAP estimates for *all* values of  $\gamma$ .  $\square$

**Theorem 3.2.4** *If families of minimizers  $(\mu_I^*)_{I \in \mathcal{P}}$  from (2.3) and minimum values  $H_I(\mu_I^*)$ ,  $I \in \mathcal{P}$ , for all partitions  $\mathcal{P} \in \mathfrak{P}$  can be computed in  $O(g(N))$ , then Algorithm 3.2.2 has complexity  $O(\max\{g(N), N^3\})$ .*

**Proof** The essential difference to Algorithm 3.1.2 lies in step (3). To compute  $\tilde{B}(k, n)$  recursively we perform at most  $N - 1$  comparisons for fixed  $k$  and we have  $N(N - 1)/2$  terms  $\tilde{B}(k, n)$ . This results in complexity  $O(N^3)$  for step (3). To construct the function  $h$  we have to determine the break points  $\gamma_i$ ,  $i = 0, \dots, k$ , which has complexity  $O(N)$ . Hence, the algorithm for the computation of MAP estimates for all values of  $\gamma$  simultaneously has complexity  $O(\max\{g(N), N^3\})$ .  $\square$

With the same arguments as in the case for fixed hyperparameter  $\gamma$  we get

**Corollary 3.2.5** *For  $\rho(u) = u^2$  the Algorithm 3.2.2 has complexity  $O(N^3)$ .*

**Remark 3.2.6** Algorithm 3.1.2 and Algorithm 3.2.2 work irrespectively of uniqueness.



## Part II

# Choice of Hyperparameters



Finally, we want to apply our estimators to time series'. The present variational approach leads to a family of MAP estimators, one for each hyperparameter  $\gamma$ . It remains the natural problem to determine the 'right' or 'best' value of the hyperparameter  $\gamma$ . By piecewise constancy, this is equivalent to decide on a  $\gamma$ -interval. This problem is ubiquitous in nonparametric statistics, examples are the choice of a smoothing parameter, bandwidth selection, and the decision on the prior smoothness or variance.

Concerning variational approaches, this problem was already addressed in the early paper S. GEMAN and D. GEMAN (1984) who resorted to *ad hoc* choices of the hyperparameter in the case of texture classification. Their strategy amounted to supervised learning from probes of known texture. For data like those from brain mapping or gene expression, discussed in detail in Chapter 9, we do not have any ground truth or model to train our methods, only qualitative morphological properties can be characterized with confidence.

One strategy is to compute estimates for all  $\gamma$  and to display them simultaneously in a plot, or as a sequence, and then simply to show them. The 'customer' then may choose an adequate estimate by visual inspection. This is addressed explicitly as the 'family approach' for example in J.S. MARRON and S.S. CHUNG (2001). Others, like P. L. DAVIES and A. KOVAC (2001) vary the hyperparameter in a monotonous way and decide to stop when certain criteria - say for residuals - are fulfilled. Such a proceeding is appropriate if one aims at the reconstruction of an underlying signal, perhaps in a parsimonious fashion. Still others, like J. POLZEHL and V.G. SPOKOINY (2000), estimate variances during an iterative procedure, but they need initial estimates.

For two reasons we are searching for automatic methods. Firstly, decisions should be objective and should not depend on the rating of observers. Secondly, if there are too many time series', as it is the case for the brain or gene data, a non-automatic procedure is not feasible.

The one dimensional Potts functionals are ideally suited to study this question rigorously and in depth. This is nourished by the results on dependence of the MAP estimator on the hyperparameter in Sections 2.4 and 2.5. Another advantage is that the estimates can be computed *exactly* for all values of  $\gamma$ .

Appropriate hyperparameters vary from time series to time series. Hence proper hyperparameters must be chosen in a data adapted way. This is an intricate problem and we are far from a final solution. Below we propose some methods and study basic properties.



# Chapter 4

## Equivariance and Hyperparameters

A reasonable property of estimators is equivariance with respect to certain group actions. In Section 4.1, we establish a scaling property of MAP estimators. It implies that MAP estimators to a fixed value of the hyperparameter are not equivariant with respect to affine linear transformations.

We will derive a sufficient condition for equivariance of estimators resulting from the combination of MAP estimators and a data adapted parameter choice.

In Section 4.2, we discuss normalization of data as an example of a data adapted parameter choice. We will see that equivariance is only a minimal requirement on an estimator.

### 4.1 Equivariance

The term ‘equivariance’ means that an estimator is compatible with certain transformations of data.

**Definition 4.1.1** *Let  $G$  be a group acting on  $\mathbb{X}$ , i. e. there is a map  $\alpha : G \times \mathbb{X} \rightarrow \mathbb{X}$ ,  $\alpha(g, x) =: gx$  with*

$$g_1(g_2x) = (g_1g_2)x \quad \text{and} \quad ex = x, \quad g_1, g_2, e \in G, x \in \mathbb{X}.$$

*A map  $T : \mathbb{X} \rightarrow \mathbb{X}$  is called **equivariant with respect to the group action**  $\alpha$  if*

$$T(gx) = gT(x), \quad g \in G.$$

The choice of transformations is crucial and should follow the needs for a special problem. We will consider the canonical action of the affine linear group

of  $\mathbb{R}$  on  $\mathbb{R}^N$  and ask for equivariant MAP estimators of the one dimensional Potts functionals given by

$$\bar{H}_\gamma(\cdot, y) : \mathbb{X} \longrightarrow \mathbb{R}, \quad x \longmapsto \gamma \cdot |J(x)| + \sum_{s \in S} (y_s - x_s)^2. \quad (4.1)$$

**Convention** Whenever we speak about equivariance, the symbol  $x^*(\gamma, y)$  will denote the measurable section presented in the proof of Lemma 2.6.1. This way, we have uniqueness of MAP estimators.

The smaller the group, the easier it is to find equivariant mappings, and vice versa. As an illustration, we consider constant shifts and find that all MAP estimators are equivariant with respect to this group action. Let  $G_1 = \mathbb{R}$  be the group acting on  $\mathbb{X} \cong \mathbb{R}^N$  by

$$(b, x) \longmapsto x + b\mathbf{1}, \quad b \in G_1.$$

**Proposition 4.1.2** *All MAP estimators  $y \mapsto x^*(\gamma, y)$  of the one dimensional Potts functionals (4.1) are equivariant with respect to  $G_1$ .*

**Proof** We have to show that  $x^*(\gamma, y + b\mathbf{1}) = x^*(\gamma, y) + b\mathbf{1}$  for each  $b \in G_1$ . The shifted  $x^*(\gamma, y) + b\mathbf{1}$  minimizes  $\bar{H}_\gamma(x - b\mathbf{1}, y)$ . Since

$$\begin{aligned} |J(x - b\mathbf{1})| &= |\{s \sim t : x_s - b \neq x_t - b\}| \\ &= |\{s \sim t : x_s \neq x_t\}| = |J(x)| \end{aligned}$$

we have

$$\begin{aligned} \bar{H}_\gamma(x - b\mathbf{1}, y) &= \gamma |J(x - b\mathbf{1})| + \sum_{s \in S} (y_s - (x_s - b))^2 \\ &= \gamma |J(x)| + \sum_{s \in S} ((y_s + b) - x_s)^2 = \bar{H}_\gamma(x, y + b\mathbf{1}). \end{aligned}$$

Thus,  $x^*(\gamma, y) + b\mathbf{1}$  minimizes  $\bar{H}_\gamma(x, y + b\mathbf{1})$  which is the assertion.  $\square$

Now we study equivariance under shifts and scaling.

**Definition 4.1.3** *Let  $\text{Aff}(\mathbb{R}) = \{(b, c) : b \in \mathbb{R}, c \in \mathbb{R} \setminus \{0\}\}$  denote the **affine linear group** of  $\mathbb{R} \setminus \{0\}$  acting on  $\mathbb{X}$  by*

$$t_{b,c} : ((b, c), x) \longmapsto c \cdot x + b\mathbf{1} \quad c \in \mathbb{R}^*, b \in \mathbb{R}. \quad (4.2)$$

*The group actions (4.2) will be called **scale transformations**.*

We consider the equivariance property with respect to this class of transformations as a minimal requirement on estimators. In a large variety of applications, estimators should be equivariant with respect to the choice of the reference point and unit of the  $y$ -axis.

The jump set is an invariant under this action of  $\text{Aff}(\mathbb{R})$ .

**Lemma 4.1.4** *The jump set  $J : \mathbb{X} \rightarrow \mathfrak{p}(S)$ ,  $x \mapsto J(x)$ , where  $\mathfrak{p}(S)$  denotes the power set of  $S$ , is invariant under  $\text{Aff}(\mathbb{R})$ , i. e.*

$$J(t_{b,c}(x)) = J(x), \quad x \in \mathbb{X}.$$

*In particular, the number of jumps  $|J(x)|$  is invariant with respect to scale transformations.*

**Proof** We have

$$J(t_{b,c}(x)) = \{s \sim t : cx_s + b \neq cx_t + b\} = \{s \sim t : x_s \neq x_t\} = J(x)$$

for each scale transformation  $t_{b,c}$  and therefore also  $|J(t_{b,c}(x))| = |J(x)|$ .  $\square$

The scaling property of MAP estimators reads as follows.

**Theorem 4.1.5** *Let  $x^*(\gamma, y)$  be the MAP estimators of the Potts functionals from (4.1). Then*

$$x^*\left(\gamma, t_{b,c}(y)\right) = t_{b,c}\left(x^*\left(\frac{\gamma}{c^2}, y\right)\right) \quad (4.3)$$

*for each scale transformation  $t_{b,c}$ .*

**Proof** With  $t_{b,c}(y) = c \cdot y + b\mathbf{1}$  we have

$$\begin{aligned} \bar{H}_\gamma(x, t_{b,c}(y)) &= \gamma \cdot |J(x)| + \sum_{s \in S} (x_s - (cy_s + b))^2 \\ &= \gamma \cdot |J(x)| + c^2 \cdot \sum_{s \in S} \left(\frac{x_s - b}{c} - y_s\right)^2 \\ &= c^2 \left( \frac{\gamma}{c^2} \cdot |J(x)| + \sum_{s \in S} \left(\frac{x_s - b}{c} - y_s\right)^2 \right). \end{aligned}$$

By Lemma 4.1.4, the number of jumps is invariant under scale transformations, and we get

$$\bar{H}_\gamma(x, t_{b,c}(y)) = c^2 \cdot \bar{H}_{\gamma/c^2}(t_{-b/c, 1/c}(x), y). \quad (4.4)$$

Hence, we have

$$t_{b,c}(\tilde{x}) = t_{b,c}(x^*(\frac{\gamma}{c^2}, y)) = x^*(\gamma, t_{b,c}(y)).$$

This completes the proof.  $\square$

Theorem 4.1.5 implies in particular that MAP estimators, minimizers of the Potts functional (4.1) with fixed  $\gamma$ , are not equivariant.

**Corollary 4.1.6** *Let  $\gamma > 0$  be fixed. Then the map  $y \mapsto x^*(\gamma, y)$  is not equivariant with respect to  $\text{Aff}(\mathbb{R})$ .*

This is only one of the reasons why the choice of the hyperparameter  $\gamma$  should depend on the data at hand. To take  $\gamma$  fixed is even worse if we have to find estimators for several data sets, and hence now less feasible than ever in an automatic procedure. This statement is based on the observation that already for two different time series' there might be no fixed value for  $\gamma$  which yields the desired estimator in both situations. This is illustrated in the following example.

**Example 4.1.7** For  $N = 6$  consider the data displayed in Figure 4.1. For data  $y$  in Figure 4.1(a), we want a constant estimator, whereas the characteristic feature of  $y'$  in Figure 4.1(b) is a jump. To simplify calculation, assume that the respective highest jump has height 1. For  $\gamma > 17/48$  we get the desired constant MAP estimator of  $y$ . If  $\gamma < 1/3$  the MAP estimator of  $y'$  is a signal with one jump. Hence, there is no fixed value for the hyperparameter which works in both situations. This indicates that the hyperparameter should be adapted to data.

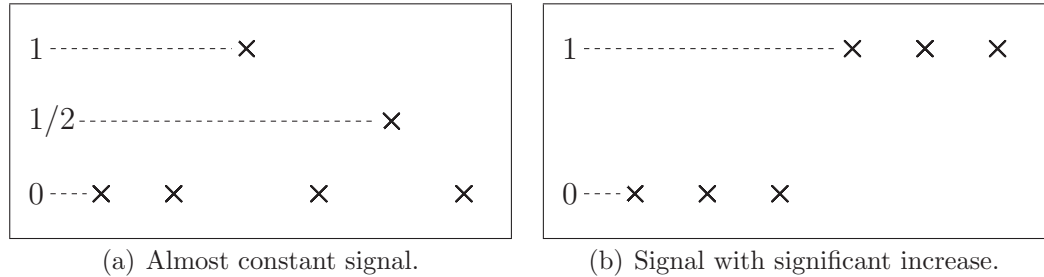


Figure 4.1: Two essentially different signals with identical diameter.

The following definition simply names the procedure of choosing  $\gamma$  according to data.

**Definition 4.1.8** A map

$$\Gamma : \mathbb{X} \longrightarrow (0, \infty), \quad y \longmapsto \Gamma(y)$$

is called **data adapted parameter choice**.

One aim in choosing the parameter in a data adapted way is to derive an equivariant estimator from the MAP estimators of the Potts functionals. We note a sufficient condition for equivariance of these estimators with respect to  $\text{Aff}(\mathbb{R})$ .

**Lemma 4.1.9** Let  $x^*(\gamma, y)$  be the minimizers of the Potts functionals from (4.1) and let  $\Gamma$  be a data adapted parameter choice. If

$$x^*(\Gamma(y), y) = x^*(\Gamma(t_{b,c}(y))/c^2, y) \quad (4.5)$$

then  $y \mapsto x^*(\Gamma(y), y)$  is equivariant with respect to  $\text{Aff}(\mathbb{R})$ .

**Proof** Let  $t_{b,c}$  denote a scale transformation. By Theorem 4.1.5 we have

$$x^*(\Gamma(t_{b,c}(y)), t_{b,c}(y)) = x^*(\Gamma(cy + b\mathbf{1}), (cy + b\mathbf{1})) = c \cdot x^*(\Gamma(cy + b\mathbf{1})/c^2, y) + b\mathbf{1}.$$

Inserting (4.5), we get

$$c \cdot x^*(\Gamma(cy + b\mathbf{1})/c^2, y) + b\mathbf{1} = c \cdot x^*(\Gamma(y), y) + b\mathbf{1} = t_{b,c}(x^*(\Gamma(y), y))$$

which gives the desired identity.  $\square$

In the following chapters, we will present several data adapted parameter choices and check whether they fulfill the sufficient condition in Lemma 4.1.9. The following observation will be useful to characterize the equivariant estimators given by the interval criteria from Section 5.2.

**Proposition 4.1.10** An estimator  $y \mapsto x^*(\Gamma(y), y)$  is equivariant with respect to  $\text{Aff}(\mathbb{R})$  if and only if  $\Gamma(t_{b,c}(y))/c^2$  is in the same  $\gamma$ -interval as  $\Gamma(y)$ .

**Proof** An estimator  $y \mapsto x^*(\Gamma(y), y)$  is equivariant with respect to  $\text{Aff}(\mathbb{R})$  if and only if

$$x^*(\Gamma(cy + b\mathbf{1}), cy + b\mathbf{1}) = c \cdot x^*(\Gamma(y), y) + b\mathbf{1}. \quad (4.6)$$

By the scaling property (4.3), equivariance of  $x^*(\Gamma(y), y)$  is equivalent to

$$c \cdot x^*\left(\frac{\Gamma(cy + b\mathbf{1})}{c^2}, y\right) = c \cdot x^*(\Gamma(y), y) + b\mathbf{1}. \quad (4.7)$$

By Theorem 2.4.5, the MAP estimator is the same on the  $\gamma$ -intervals. Hence, the map  $y \mapsto x^*(\Gamma(y), y)$  is equivariant with respect to  $\text{Aff}(\mathbb{R})$  if and only if  $\Gamma(cy + b\mathbf{1})/c^2$  is in the same  $\gamma$ -interval as  $\Gamma(y)$ .  $\square$

Analogues to Theorem 4.1.5 and its implications are true in a slightly more general frame with only a few modifications.

**Remark 4.1.11** The considerations on equivariance are valid for a more general class of functionals. We will now consider functionals given by

$$F_\gamma(\cdot, y) : \mathbb{X} \longrightarrow \mathbb{R}, \quad x \longmapsto \gamma \cdot P(x) + \sum_{s \in S} (y_s - x_s)^2. \quad (4.8)$$

The term

$$P : \mathbb{X} \longrightarrow \mathbb{R}, \quad x \longmapsto P(x)$$

in the functional  $F_\gamma$  is called  $\alpha$ -homogeneous penalty if

$$P(t_{b,c}(y)) = |c|^\alpha P(x)$$

for some  $\alpha \in \mathbb{R}$  and scale transformations  $t_{b,c}$ . An example for such  $\alpha$ -homogeneous functionals is the functional with the total variation as penalty term,

$$F_\lambda : \mathbb{X} \times \mathbb{X} \longrightarrow \mathbb{R}, \quad (x, y) \longmapsto \lambda \cdot \text{TV}(x) + \sum_{s \in S} (y_s - m_s)^2$$

where the total variation for  $x = (x_s)_{s=1, \dots, N} \in \mathbb{X} = \mathbb{R}^S$  is given by

$$\text{TV}(x) = \sum_{s=1}^{N-1} |x_{s+1} - x_s|.$$

Minimizers of this functional are considered for example in E. MAMMEN and S. VAN DE GEER (1997). The taut string algorithm presented in P. L. DAVIES and A. KOVAC (2001) yields a minimizer of this functional.

## 4.2 Normalization of Data

As an example of data adapted parameter choices, we will now discuss normalization of data, a popular method to prepare them for analysis. First of all we will give a very general definition of normalization.

**Definition 4.2.1** Let be  $A : \mathbb{X} \rightarrow \text{Aff}(\mathbb{R})$ . The map

$$N_A : \mathbb{X} \longrightarrow \mathbb{X}, \quad x \longmapsto A(x)x = c(x)x + b(x)\mathbf{1}$$

is an **affine normalization** if it is constant on the orbits of  $\text{Aff}(\mathbb{R})$  on  $\mathbb{X}$ , i. e. if  $N_A(gx) = N_A(x)$  for  $g \in \text{Aff}(\mathbb{R})$ ,  $x \in \mathbb{X}$ .

Recall that the *orbit* of  $x$  with respect to the group  $G$  is the set  $\{gx : g \in G\}$ . It will turn out that for *any* map  $F : \mathbb{X} \rightarrow \mathbb{X}$  there is an estimator which is equivariant with respect to  $\text{Aff}(\mathbb{R})$ .

**Proposition 4.2.2** Let a map  $F : \mathbb{X} \rightarrow \mathbb{X}$  be given and let  $N_A$  be an affine normalization. Then the estimator

$$F_A : \mathbb{X} \longrightarrow \mathbb{X}, \quad y \longmapsto A(y)^{-1}F(A(y)y)$$

is equivariant with respect to the group action of  $\text{Aff}(\mathbb{R})$ .

**Proof** We have to show that

$$t_{b,c}(F_A(y)) = F_A(t_{b,c}(y)).$$

By definition, an affine normalization is constant on the  $\text{Aff}(\mathbb{R})$ -orbits on  $\mathbb{X}$  and therefore

$$A(cy + b\mathbf{1})(cy + b\mathbf{1}) = A(y)y \quad \text{and} \quad cy + b\mathbf{1} = (A(cy + b\mathbf{1}))^{-1}A(y)y.$$

We conclude

$$\begin{aligned} F_A(t_{b,c}(y)) &= (A(cy + b\mathbf{1}))^{-1}F(A(cy + b\mathbf{1})(cy + b\mathbf{1})) \\ &= (A(cy + b\mathbf{1}))^{-1}F(A(y)y) = (A(cy + b\mathbf{1}))^{-1}A(y)(A(y))^{-1}F(A(y)y) \\ &= (A(cy + b\mathbf{1}))^{-1}A(y)F_A(y) = c \cdot F_A(y) + b\mathbf{1} = t_{b,c}(F_A(y)) \end{aligned}$$

which is the assertion. □

Equivariance does not single out any estimator. Whether an estimator is equivariant essentially depends on the size of the group. In fact, Proposition 4.2.2 tells us that under normalization *any* estimator  $F_A$  is equivariant, irrespective how reasonable it is. Moreover, it is  $F_A \neq F$ . This observation underlines that equivariance is just a minimal requirement.

The following is a standard example for normalization.

**Example 4.2.3** Let be  $y \in \mathbb{R}^N$ . With the empirical mean  $\bar{y}$  and for diameter  $\text{diam}(y) = \max_i y_i - \min_i y_i \neq 0$  we define

$$A(y) = \left( -\frac{2\bar{y}}{\text{diam}(y)}, \frac{2}{\text{diam}(y)} \right) \in \text{Aff}(\mathbb{R}).$$

Then the map

$$N_A : \mathbb{X} \rightarrow \mathbb{X}, \quad y \mapsto \begin{cases} 0 \cdot \mathbf{1} & \text{for } y = a\mathbf{1}, a \in \mathbb{R}, \\ A(y)y = \frac{2(y - \bar{y}\mathbf{1})}{\text{diam}(y)} & \text{otherwise} \end{cases} \quad (4.9)$$

is an affine normalization on  $\mathbb{X}$ . This is obvious for constant  $y$ . Otherwise, we have

$$\overline{cy + b\mathbf{1}} = c\bar{y}\mathbf{1} + b\mathbf{1} \quad \text{and} \quad \text{diam}(cy + b\mathbf{1}) = c \text{diam}(y)$$

and

$$\begin{aligned} N_A(cy + b\mathbf{1}) &= \frac{2(cy + b\mathbf{1} - \overline{(cy + b\mathbf{1})})}{\text{diam}(cy + b\mathbf{1})} = \frac{2(cy + b\mathbf{1} - c\bar{y}\mathbf{1} - b\mathbf{1})}{c \text{diam}(y)} \\ &= \frac{2(y - \bar{y}\mathbf{1})}{\text{diam}(y)} = N_A(y). \end{aligned}$$

We now interpret normalization as special choice of the hyperparameter in the MAP estimator of the Potts functionals.

**Example 4.2.4** Let  $y \mapsto x^*(\gamma, y)$  be the MAP estimator of the Potts functional in (4.1). Let  $N_A(y)$  denote data  $y$  normalized by an affine normalization. Then

$$x^*(\gamma, N_A(y)) = \begin{cases} 0 \cdot \mathbf{1} & \text{for } y = a\mathbf{1}, \\ c(y) \cdot x^*\left(\frac{\gamma}{(c(y))^2}, y\right) + b(y)\mathbf{1} & \text{otherwise} \end{cases}$$

like in the proof of Theorem 4.1.5 with

$$b = b(y) = -\frac{2}{\text{diam}(y)} \cdot \bar{y} \quad \text{and} \quad c = c(y) = \frac{2}{\text{diam}(y)}.$$

Hence,

$$x^*(\gamma, N_A(y)) = \begin{cases} N_A(x^*(\gamma, y)) & \text{for } y = a\mathbf{1}, \\ N_A(x^*\left(\frac{\gamma}{(c(y))^2}, y\right)) & \text{otherwise.} \end{cases}$$

This shows that the MAP estimator for normalized data is the normalization of the MAP estimator for the original data but with scaled hyperparameter. Hence, we do not gain anything by normalization of data in the sense that we are still faced with the problem to choose the hyperparameter  $\gamma$ . Note that the choice of a fixed  $\gamma$ -value for standardized data corresponds to a certain normalization. Thus, a data adapted parameter choice corresponds to the choice of a normalization and a fixed  $\gamma$ .

Moreover, the scaling does only depend on a single number. Thus, the information about  $y$  which enters the data adapted parameter choice is reduced to this single number. Signals with identical diameter may be rather different, see for example those in Figure 4.1. Hence this normalization seems to be too crude and not suitable since the Potts functionals look locally for changes - the neighborhood to detect a jump are only two sites.

The following example shows that in some situations the scaling with the diameter from Example 4.2.3 seems not to be adequate.

**Example 4.2.5** We consider the data from Example 4.1.7. As already mentioned, there is no value for the hyperparameter such that we get the desired MAP estimator in both situations. Consider now the normalization from Example 4.2.3. As shown in Example 4.2.4 the normalization corresponds to a scaling of the  $\gamma$  with a factor depending only on the diameter of data. Since data  $y$  in Figure 4.1(a) and  $y'$  in 4.1(b) have identical diameter, the scaling is the same for both. Thus, normalization does not improve the situation.

Normalization is an extrinsic method to make the MAP estimator equivariant. It does not take into account the special structure of the Potts functional and its MAP estimators. The interval criteria presented in the following chapter are intrinsic. They use the  $\gamma$ -scanning explicitly.



# Chapter 5

## Interval Criteria

In this chapter, we present a special class of data adapted parameter choices. They make use of the  $\gamma$ -intervals in the  $\gamma$ -scanning on which MAP estimators are constant. Choosing the hyperparameter according to these interval criteria leads to equivariant estimators.

In Section 5.1, we observe that there are properties of the  $\gamma$ -intervals which do not change if data are transformed. These invariant attributes are the base for the estimators presented in the subsequent sections as well as for those in the following chapters.

In Section 5.2, we propose criteria based on the invariant attribute ‘to be one of the longest  $\gamma$ -interval’. The idea is that the length of the  $\gamma$ -intervals, after a transformation of the  $\gamma$ -values by a function  $F$ , corresponds to the ‘stability’ of the respective MAP estimator. We characterize those longest interval criteria which lead to equivariant estimators.

In Section 5.3, we focus on the special case of longest interval criteria without a transformation of the  $\gamma$ -axis. It will be discussed in more detail.

### 5.1 Invariant Attributes

The set of the finite positive boundaries of the  $\gamma$ -intervals arising from the minimization of the Potts functional  $H_\gamma(\cdot, y)$  for all values of  $\gamma$  will be denoted by

$$\mathcal{G}(y) = \{\gamma_0(y), \dots, \gamma_{m(y)}(y)\}. \quad (5.1)$$

By Theorem 4.1.5, the scaling property of the  $\gamma$ -values for transformed data  $t_{b,c}(y)$  reads

$$\mathcal{G}(cy + b\mathbf{1}) = \{c^2\gamma_0(y), \dots, c^2\gamma_{m(y)}(y)\} = c^2\mathcal{G}(y).$$

Hence, the lengths of the  $\gamma$ -intervals change but all in equal measure, i. e. for example the longest interval for  $y$  stays the longest interval for  $t_{b,c}(y)$ . The order of the intervals does not change as well as the position in any ranking is retained unchanged. We will call a property of a  $\gamma$ -interval which is maintained for transformed data  $t_{b,c}(y)$  an *invariant attribute*. Choosing a  $\gamma$ -value from the interior of a  $\gamma$ -interval according to such an invariant attribute is a data adapted parameter choice providing a to  $\text{Aff}(\mathbb{R})$  equivariant estimator. Examples are

- (1) the position in the ranking with respect to the length of the intervals, possible after a transformation of the  $\gamma$ -axis. In Definition 5.2.1 we will introduce a criterion based on the attribute to be the longest  $\gamma$ -interval,
- (2) the property to be the  $k$ -th interval, counted from the rightmost interval  $(\gamma_0(y), \infty)$ , for which the corresponding estimator is monotonous. Herefrom arises the last monotone criterion from Section 8.3,
- (3) the property to be the  $k$ -th interval, counted from the rightmost interval  $(\gamma_0(y), \infty)$ , for which the corresponding estimator fulfills some stopping condition like those in Chapter 6, and
- (4) any combination of such properties, such as being the longest interval for which the corresponding estimator is monotonous.

## 5.2 $F$ -Longest Interval Criteria

In this section, we will consider estimators which are obtained as follows: Transform the  $\gamma$ -axis by a strictly increasing continuous function  $F$ , pick the longest of the intervals  $[F(\gamma_{i-1}(y)), F(\gamma_i(y))]$ ,  $i = 1, \dots, m(y)$ , and as estimate the one belonging to some interior point of the original  $\gamma$ -interval. To be definite, we will choose the center point. Denoting by

$$\text{dist}(z, A) = \min_{z' \in A} |z - z'|$$

the distance of a real number  $z$  to a finite set  $A$  of real numbers, this procedure is made more precise in the following definition.

**Definition 5.2.1** *Let  $F : \mathbb{R}^+ \rightarrow \mathbb{R}$  be a continuous and strictly increasing function with  $F(1) = 0$ . If  $\mathcal{G}(y) \neq \emptyset$  let  $z^*(y)$  be given by*

$$z^*(y) = \max\{\tilde{z} \in F([\gamma_{m(y)}(y), \gamma_0(y)]) : \text{dist}(\tilde{z}, F(\mathcal{G}(y))) \geq \text{dist}(F(z), F(\mathcal{G}(y))) \text{ for all } z \in [\gamma_{m(y)}(y), \gamma_0(y)]\}.$$

A data adapted parameter choice of the form

$$\Gamma^F : \mathbb{X} \longrightarrow \mathbb{R}^+, \quad y \longmapsto \begin{cases} 0 & \text{if } \mathcal{G}(y) = \emptyset, \\ F^{-1}(z^*(y)) & \text{otherwise} \end{cases} \quad (5.2)$$

is called  **$F$ -longest interval criterion (FLIC)**. Estimators

$$\mathbb{X} \longrightarrow \mathbb{X}, \quad y \longmapsto x^*(\Gamma^F(y), y) \quad (5.3)$$

where  $y \mapsto x^*(\gamma, y)$  is the MAP estimator of the Potts functional (4.1) are called **FLIC estimators**.

Note that in the definition of the  $F$ -longest interval criterion we only consider the  $\gamma$ -intervals of *finite* length. Since functions  $F$  may map infinity or zero to infinity we do not take into account the estimators corresponding to the leftmost interval  $(0, \gamma_{m(y)}(y))$  and to the rightmost interval  $(\gamma_0(y), \infty)$ .

**Remark 5.2.2** Note that the data adapted parameter choice  $\Gamma^F$  in (5.2) is well-defined.  $\mathcal{G}(y) = \emptyset$  if and only if  $y$  is a constant signal since in this case there is no positive  $\gamma$ -value. Since for constant  $y$  there is only one estimator, namely the constant one, setting  $\gamma = 0$  is reasonable.

In general, we are not interested to get data back, and hence skipping the leftmost  $\gamma$ -interval  $(0, \gamma_{m(y)}(y))$  is not a serious restriction. In contrast, the fact that we will never get the MAP estimator which belongs to the rightmost interval  $(\gamma_0(y), \infty)$  constitutes a real problem. Recall that for  $\gamma > \gamma_0(y)$  the MAP estimator is a constant signal. A reasonable estimator should map a quasi constant time series to a constant one. FLIC estimators do not have this important property, and thus we are forced to watch out for other methods to handle this problem.

Recall that we are interested in equivariant estimators. In case of FLIC estimators, Proposition 4.1.10 can be reduced to the fact that  $\Gamma^F(t_{b,c}(y))/c^2$  and  $\Gamma^F(y)$  coincide.

**Proposition 5.2.3** A FLIC estimator  $y \mapsto x^*(\Gamma^F(y), y)$  is equivariant with respect to  $\text{Aff}(\mathbb{R})$  if and only if for all  $y \in \mathbb{X}$

$$\frac{\Gamma^F(t_{b,c}(y))}{c^2} = \Gamma^F(y).$$

**Proof** We use Proposition 4.1.10. The  $\gamma$ -interval containing  $\Gamma^F(y)$  is an interval  $(\gamma_i(y), \gamma_{i-1}(y))$  for which

$$F(\gamma_{i-1}(y)) - F(\gamma_i(y)) \geq F(\gamma_{j-1}(y)) - F(\gamma_j(y)) \quad \text{for all } j \neq i.$$

Whereas  $\Gamma^F(cy + b\mathbf{1})$  is given as the center point of an interval  $(\gamma_l(y), \gamma_{l-1}(y))$  for which

$$F(c^2\gamma_{l-1}(y)) - F(c^2\gamma_l(y)) \geq F(c^2\gamma_{j-1}(y)) - F(c^2\gamma_j(y)) \quad \text{for all } j \neq l.$$

The value  $\Gamma^F(cy + b\mathbf{1})/c^2$  is in  $(\gamma_i(y), \gamma_{i-1}(y))$  if and only if  $(\gamma_i(y), \gamma_{i-1}(y))$  and  $(\gamma_l(y), \gamma_{l-1}(y))$  coincide. Since the  $\gamma$ -intervals are disjoint, this is the case if and only if both intervals have the same center point. Hence, the estimator  $y \mapsto x^*(\Gamma^F(y), y)$  with the  $F$ -longest interval criterion (5.2) is equivariant with respect to  $\text{Aff}(\mathbb{R})$  if and only if

$$\frac{\Gamma^F(cy + b\mathbf{1})}{c^2} = \Gamma^F(y)$$

which proves the assertion.  $\square$

We will now characterize FLIC estimators which are equivariant with respect to  $\text{Aff}(\mathbb{R})$ .

**Theorem 5.2.4** *Let be  $F(1) = 0$ . The FLIC estimator  $y \mapsto x^*(\Gamma^F(y), y)$  is equivariant with respect to  $\text{Aff}(\mathbb{R})$  if and only if*

$$F(x) = a \frac{x^\beta - 1}{2^\beta - 1}$$

for some  $\beta \in \mathbb{R} \setminus \{0\}$  and  $a > 0$ , or

$$F(x) = a \cdot \ln x$$

for  $\beta = 0$  and some  $a > 0$ .

The proof of Theorem 5.2.4 will be given at the end of this section. It uses the following implications of the equivariance of FLIC estimators.

Recall that a continuous, strictly increasing function  $F$  corresponds one-to-one to a positive measure  $\mu_F$  on the Borel- $\sigma$ -field  $\mathcal{B}(\mathbb{R}^+)$  with

$$\mu_F([a, b)) = F(b) - F(a).$$

A consequence of the equivariance of  $y \mapsto x^*(\Gamma^F(y), y)$  is the following property of the measure  $\mu_F$ .

**Proposition 5.2.5** *Suppose that the FLIC estimator  $y \mapsto x^*(\Gamma^F(y), y)$  to the  $F$ -longest interval criterion  $\Gamma^F$  is equivariant. Then for  $\alpha \geq 0$  and intervals  $I, J \subset \mathbb{R}^+$*

$$\mu_F(I) = \alpha \mu_F(J) \quad \text{implies} \quad \mu_F(CI) = \alpha \mu_F(CJ)$$

for all  $C > 0$ .

The proof is essentially the solution of a functional equation. We will show that the implication in Proposition 5.2.5 is valid first for integer  $\alpha \in \mathbb{N}$  and then for rational numbers. Using a continuity argument we finally arrive at the statement for all positive real numbers.

**Proof of Proposition 5.2.5** (1) Note, that for each set  $A$  of at most  $N - 2$  real positive values we can construct data  $y$  such that  $A = \mathcal{G}(y)$ . Recall from the proof of Proposition 2.4.2 that the  $\gamma$ -values are the intersection points of the functions  $f_y^k(\gamma)$  from (2.18). If  $m \leq N - 2$  values  $\gamma_0, \dots, \gamma_{m-1}$  are given, set  $y_{m+1} = \dots = y_N = 0$ . This fixes that the functions  $f_y^i(\gamma)$  and  $f_y^{i+1}(\gamma)$  intersect at  $\gamma_i$ . We get  $m$  quadratic equations for  $m$  variables  $y_1, \dots, y_m$ . This justifies to consider  $F$  as a function of  $\gamma$  and not of  $y$ .

(2) By Proposition 5.2.3, a FLIC estimator  $y \mapsto x^*(I^F(y), y)$  is equivariant if and only if  $I^F(t_{b,c}(y))/c^2$  is in the same  $\gamma$ -interval as  $I^F(y)$ . Hence, equivariance implies that if the  $F$ -length of the  $\gamma$ -interval  $I$  is maximal then also the  $F$ -length of the transformed interval  $c^2I$  is maximal. Setting  $c^2 = C$  this implies that

$$\mu_F(I) > \mu_F(J) \quad \text{for intervals } I \neq J \subset \mathbb{R}^+$$

implies

$$\mu_F(CI) > \mu_F(CJ) \quad \text{for } C > 0.$$

Let now  $[\gamma_1, \gamma_2]$  and  $[\gamma_3, \gamma_4]$  be intervals with  $\gamma_2 \leq \gamma_3$  or  $\gamma_1 \geq \gamma_4$ . Then the consideration above tells us that

$$F(\gamma_2) - F(\gamma_1) > F(\gamma_4) - F(\gamma_3)$$

implies

$$F(C\gamma_2) - F(C\gamma_1) > F(C\gamma_4) - F(C\gamma_3).$$

Take now  $\gamma_4$  such that

$$F(\gamma_2) - F(\gamma_1) = F(\gamma_4) - F(\gamma_3). \quad (5.4)$$

This is possible due to (1). Since  $F$  increases strictly, we have for  $\gamma_4^n = \gamma_4 - 1/n$  that

$$F(\gamma_2) - F(\gamma_1) > F(\gamma_4^n) - F(\gamma_3) \quad \text{for } n \in \mathbb{N}.$$

According to the assumption, this implies

$$F(C\gamma_2) - F(C\gamma_1) > F(C\gamma_4^n) - F(C\gamma_3) \quad \text{for } n \in \mathbb{N}.$$

Using continuity of  $F$ , we conclude that

$$F(C\gamma_2) - F(C\gamma_1) \geq F(C\gamma_4) - F(C\gamma_3).$$

Since  $F(C\gamma_2) - F(C\gamma_1) > F(C\gamma_4) - F(C\gamma_3)$  contradicts (5.4), we have that (5.4) always implies

$$F(C\gamma_2) - F(C\gamma_1) = F(C\gamma_4) - F(C\gamma_3). \quad (5.5)$$

Thus, we have that (5.4) implies (5.5) for all  $\gamma_1, \gamma_2, \gamma_3, \gamma_4$  with

$$|\{\gamma > 0 : \gamma \in [\gamma_1, \gamma_2] \cap [\gamma_3, \gamma_4]\}| \leq 1. \quad (5.6)$$

(3) Equation (5.4) implies (5.5) for all  $\gamma_1, \gamma_2, \gamma_3, \gamma_4$  without the condition (5.6) on the intervals due to the following reasons: If the interval  $[\gamma_3, \gamma_4]$  lies in  $[\gamma_1, \gamma_2]$  then equality of  $F(\gamma_2) - F(\gamma_1)$  and  $F(\gamma_4) - F(\gamma_3)$  would not be possible since  $F$  is strictly monotone. Consider the case of overlapping intervals and assume without restriction that  $\gamma_1 < \gamma_3 < \gamma_2 < \gamma_4$ . Then (5.4) is equivalent to

$$F(\gamma_3) - F(\gamma_1) = F(\gamma_4) - F(\gamma_2)$$

for the disjoint intervals  $[\gamma_1, \gamma_3]$  and  $[\gamma_2, \gamma_4]$ . This implies

$$F(C\gamma_3) - F(C\gamma_1) = F(C\gamma_4) - F(C\gamma_2).$$

Reordering the terms yields the assumption that (5.4) implies (5.5) for all  $\gamma_1, \gamma_2, \gamma_3, \gamma_4, C \in \mathbb{R}^+$ .

(4) Suppose now that

$$m(F(\gamma_2) - F(\gamma_1)) = F(\gamma_4) - F(\gamma_3) \quad (5.7)$$

for  $m \in \mathbb{N}$  and for all  $\gamma_1, \gamma_2, \gamma_3, \gamma_4 \in \mathbb{R}^+$ . We divide the interval  $[\gamma_3, \gamma_4]$  into  $m$  pieces of the same  $F$ -length and insert division points  $\gamma_3 = \gamma_0^m < \gamma_1^m < \dots < \gamma_m^m = \gamma_4$  with

$$F(\gamma_2) - F(\gamma_1) = F(\gamma_i^m) - F(\gamma_{i-1}^m) \quad \text{for all } i = 1, \dots, m$$

which is possible since  $F$  is continuously increasing. We conclude that

$$F(C\gamma_2) - F(C\gamma_1) = F(C\gamma_i^m) - F(C\gamma_{i-1}^m) \quad \text{for all } i = 1, \dots, m.$$

Summing up these  $m$  equations gives

$$m(F(C\gamma_2) - F(C\gamma_1)) = F(C\gamma_4) - F(C\gamma_3). \quad (5.8)$$

Hence, we have that (5.7) implies (5.8) for  $m \in \mathbb{N}$  and therefore also for a factor  $\tilde{m} = 1/m$ .

(5) Assume now that

$$\frac{1}{q}(F(\gamma_2) - F(\gamma_1)) = \frac{1}{p}(F(\gamma_4) - F(\gamma_3)) \quad \text{for } p, q \in \mathbb{N}. \quad (5.9)$$

Due to the continuity of  $F$  we can find  $\gamma_5 < \gamma_6$  such that

$$\frac{1}{q}(F(\gamma_2) - F(\gamma_1)) = F(\gamma_6) - F(\gamma_5) = \frac{1}{p}(F(\gamma_4) - F(\gamma_3))$$

which implies

$$\frac{1}{q}(F(C\gamma_2) - F(C\gamma_1)) = F(C\gamma_6) - F(C\gamma_5) = \frac{1}{p}(F(C\gamma_4) - F(C\gamma_3)).$$

Thus, (5.9) implies

$$\frac{p}{q}(F(C\gamma_2) - F(C\gamma_1)) = F(C\gamma_4) - F(C\gamma_3)$$

for  $p/q \in \mathbb{Q}^+$ .

(6) Assume now that

$$\alpha(F(\gamma_2) - F(\gamma_1)) = F(\gamma_4) - F(\gamma_3) \quad \text{for } \alpha \in \mathbb{R}^+. \quad (5.10)$$

Take  $\alpha_n \in \mathbb{Q}^+$ ,  $\alpha_n < \alpha$  for all  $n \in \mathbb{N}$ , such that  $\alpha_n \rightarrow \alpha$ . Since  $F$  increases strictly, there are  $\gamma_4^n < \gamma_4$  with

$$\alpha_n(F(\gamma_2) - F(\gamma_1)) = F(\gamma_4^n) - F(\gamma_3).$$

Due to the considerations in (5), this implies

$$\alpha_n(F(C\gamma_2) - F(C\gamma_1)) = F(C\gamma_4^n) - F(C\gamma_3).$$

Continuity of  $F$  and its increase imply in the limit  $n \rightarrow \infty$  that  $\gamma_4^n \rightarrow \gamma_4$  and

$$\alpha(F(C\gamma_2) - F(C\gamma_1)) = F(C\gamma_4) - F(C\gamma_3). \quad (5.11)$$

Hence, (5.10) implies (5.11) for all  $\alpha \in \mathbb{R}^+$ ,  $\gamma_1, \gamma_2, \gamma_3, \gamma_4, C \in \mathbb{R}^+$ .  $\square$

**Lemma 5.2.6** *The equivariance of the FLIC estimator  $y \mapsto x^*(\Gamma^F(y), y)$  implies that there is a function  $f : \mathbb{R}^+ \rightarrow \mathbb{R}^+$  such that for all  $C > 0$  and intervals  $I \subset (0, \infty)$*

$$\mu_F(CI) = f(C)\mu_F(I). \quad (5.12)$$

*The function  $f$  is a **real character** of the multiplicative group  $\mathbb{R}^+$ , i. e.*

$$f(1) = 1 \quad \text{and} \quad f(C_1 \cdot C_2) = f(C_1) \cdot f(C_2).$$

**Proof** By Proposition 5.2.5, we have that  $F(b) - F(a) = \alpha(F(d) - F(c))$  for  $\alpha > 0$  implies that  $F(Cb) - F(Ca) = \alpha(F(Cd) - F(Cc))$  for all  $C > 0$ . Set now  $a = 1$ ,  $b = 2$ ,  $c = a$ , and  $d = b$  as well as

$$\alpha = \frac{F(b) - F(a)}{F(2) - F(1)}.$$

Since (5.10) implies (5.11), the equation

$$\alpha(F(2) - F(1)) = F(b) - F(a)$$

then implies

$$\frac{F(b) - F(a)}{F(2) - F(1)} \cdot (F(2C) - F(C)) = F(Cb) - F(Ca).$$

Reordering the terms on left hand side yields

$$\frac{F(2C) - F(C)}{F(2) - F(1)} (F(b) - F(a)) = F(Cb) - F(Ca).$$

Choose

$$f(C) = \frac{F(2C) - F(C)}{F(2) - F(1)}.$$

This proves that there is  $f : \mathbb{R}^+ \rightarrow \mathbb{R}^+$  such that for all  $C > 0$  and intervals  $I \subset (0, \infty)$

$$\mu_F(CI) = f(C)\mu_F(I).$$

By (5.12) we have that

$$f(C_1C_2)\mu_F(I) = \mu_F(C_1C_2I) = f(C_1)\mu_F(C_2I) = f(C_1)f(C_2)\mu_F(I).$$

The function  $f$  is continuous since  $F$  is continuous and it fulfills the functional equation  $f(C_1 \cdot C_2) = f(C_1) \cdot f(C_2)$ . Clearly,  $f(1) = 1$ .  $\square$

The next lemma gives us the explicit form of functions  $F$  for which FLIC estimators are equivariant.

**Lemma 5.2.7** *Assume  $F(1) = 0$ . If  $y \mapsto x^*(\Gamma^F(y), y)$  is equivariant with respect to  $\text{Aff}(\mathbb{R})$  then*

$$F(x) = a \frac{x^\beta - 1}{2^\beta - 1} \tag{5.13}$$

for some  $\beta \in \mathbb{R} \setminus \{0\}$  and  $a > 0$ , or

$$F(x) = a \cdot \ln x \tag{5.14}$$

for  $\beta = 0$  and some  $a > 0$ .

**Proof** By Lemma 5.2.6 there is a function  $f : \mathbb{R}^+ \rightarrow \mathbb{R}^+$  such that (5.12) is fulfilled and  $f$  is a real character. The function  $f$  is continuous since  $F$  is continuous, and we conclude that  $f$  is of the form  $f(C) = C^\beta$ ,  $\beta \in \mathbb{R}$ . Thus, we have that

$$F(Cb) - F(Ca) = C^\beta (F(b) - F(a)), \quad \beta \in \mathbb{R}. \quad (5.15)$$

We now have to distinguish two cases.

(1) For  $\beta \neq 0$  we insert  $C = x$ ,  $a = 1$ , and  $b = 2$  in (5.15) and get

$$F(2x) - F(x) = x^\beta (F(2) - F(1)).$$

The choice  $C = 2$ ,  $a = 1$ , and  $b = x$  gives

$$F(2x) - F(2) = 2^\beta (F(x) - F(1)).$$

Subtracting these two equations and elementary calculations yield

$$F(x) = \frac{F(2) - F(1)}{2^\beta - 1} \cdot x^\beta + \frac{2^\beta F(1) - F(2)}{2^\beta - 1}.$$

Taking into account that  $F(1) = 0$  and setting  $a := F(2) > 0$  we arrive at

$$F(x) = \frac{a}{2^\beta - 1} \cdot x^\beta - \frac{a}{2^\beta - 1}$$

for  $a > 0$ .

(2) For  $\beta = 0$ , we have  $F(Cb) - F(Ca) = F(b) - F(a)$ . Inserting  $C = y$ ,  $a = 1$ , and  $b = x$  we obtain

$$F(x \cdot y) = F(x) + F(y) - F(1).$$

Now  $F(1) = 0$  gives the functional equation of the natural logarithm, and hence we get

$$F(x) = a \cdot \ln x$$

for some  $a > 0$ . □

Now we are in the position to give the proof of the main result of this section.

**Proof of Theorem 5.2.4** It is easy to see that a FLIC estimator  $y \mapsto x^*(\Gamma^F(y), y)$  is equivariant if  $F$  has the stated form. In case of (5.13) we have that

$$\text{dist}(F(c^2 z), F(c^2 \mathcal{G}(y))) = \min_{z' \in \mathcal{G}(y)} |F(c^2 z) - F(c^2 z')|$$

$$= \min_{z' \in \mathcal{G}(y)} (c^2)^\beta |F(z) - F(z')| = c^{2\beta} \text{dist}(F(z), F(\mathcal{G}(y))).$$

If  $F$  has the form (5.14), then

$$\begin{aligned} \text{dist}(F(c^2 z), F(c^2 \mathcal{G}(y))) &= \min_{z' \in \mathcal{G}(y)} |a \ln(c^2 z) - a \ln(c^2 z')| \\ &= \min_{z' \in \mathcal{G}(y)} |a \ln(z) - a \ln(z')| = \text{dist}(F(z), F(\mathcal{G}(y))). \end{aligned}$$

Lemma 5.2.7 completes the proof.  $\square$

Examples of  $F$ -longest interval criteria which lead to equivariant estimators for the Potts functional (4.1) are the following ones.

- (1) For  $F(x) = x$  the  $F$ -longest interval criterion reads: Take the estimator corresponding to the  $\gamma$ -interval  $(\gamma_j(y), \gamma_{j-1}(y))$  for which the difference  $\gamma_{j-1}(y) - \gamma_j(y)$  is maximal. It will be discussed in more detail in the next section.
- (2) For  $F(x) = \ln x$  the  $F$ -longest interval criterion suggests to take the estimator corresponding to the  $\gamma$ -interval  $(\gamma_j(y), \gamma_{j-1}(y))$  for which the ratio  $\gamma_{j-1}(y)/\gamma_j(y)$  is maximal.

**Remark 5.2.8** It is plausible that by taking the logarithm of the  $\gamma$ -axis enlarges the intervals close to zero which were originally rather short: If two successive  $\gamma$ -values have the distance  $l$  we get

$$\frac{\gamma_{i-1}}{\gamma_i} = \frac{\gamma_i + l}{\gamma_i} = 1 + \frac{l}{\gamma_i}.$$

Hence the length  $l$  of the interval has definitely more weight for smaller values  $\gamma_i$  than for larger ones. The choice of the  $\gamma$ -interval for which this ratio is maximal corresponds then to the preference of intervals close to zero.

### 5.3 Longest Interval Criterion

We consider a special case of the interval criteria introduced in the preceding section. In this section, the transformation is the identity such that the  $F$ -longest  $\gamma$ -interval is indeed the longest of the finite intervals  $(\gamma_{i-1}(y), \gamma_i(y))$ ,  $i = 1, \dots, m(y)$ . We show that this provides an almost unique data adapted parameter choice. Finally, we discuss the application of the longest interval criterion to data with trend. It will turn out that in most of these cases this criterion suggests the estimator with exactly one jump. The iterative procedure suggested in Section 8.1 is a first approach to deal with this problem.

**Definition 5.3.1** For  $F(x) = x - 1$  the FLIC estimator will be simply called the *longest interval criterion (LIC) estimator*.

The FLIC estimator is made unique by the measurable choice of  $x^*(\gamma, y)$ . In contrast thereto, for almost all  $y$  the LIC estimator is unique without this restriction.

**Proposition 5.3.2** For Lebesgue almost all  $y \in \mathbb{X}$ , the LIC estimate is unique.

**Proof** (1) We will first show that for Lebesgue almost all  $y \in \mathbb{X}$  the lengths of any two different  $\gamma$ -intervals are different.

For each  $y \in \mathbb{X}$ , the number of  $\gamma$ -intervals is equal to  $m(y) + 1$  with  $m(y) \in \{0, \dots, |S| - 2\}$  from Theorem 2.4.5.

We will denote the interval  $(\gamma_i(y), \gamma_{i-1}(y))$  as the  $i$ -th  $\gamma$ -interval. Its length is given by

$$\text{length}(i\text{-th } \gamma\text{-interval}) = \gamma_{i-1}(y) - \gamma_i(y).$$

At the end of Section 2.4 we gave explicit formulas for the computation of the  $\gamma$ -values. Using the notation from there, for given  $y \in \mathbb{X}$ , the length of the first  $\gamma$ -interval is given by

$$\begin{aligned} \gamma_0(y) - \gamma_1(y) &= \frac{\tilde{B}_y(1) - \tilde{B}_y(k_0(y))}{k_0(y) - 1} - \frac{\tilde{B}_y(k_0(y)) - \tilde{B}_y(k_1(y))}{k_1(y) - k_0(y)} \\ &= \frac{\tilde{B}_y(1)}{k_0 - 1} - \tilde{B}_y(k_0) \frac{k_1 - 1}{(k_0 - 1)(k_1 - k_0)} + \frac{\tilde{B}_y(k_1)}{k_1 - k_0} \end{aligned}$$

The length of the subsequent  $\gamma$ -intervals is of the same form. The set of  $y \in \mathbb{X}$  for which the lengths of at least two  $\gamma$ -intervals coincide is enclosed in the solutions of quadratic equations. By analogous arguments as in the proofs of Theorem 2.3.1 and Lemma 2.4.7 we see that those  $y$  are contained in a Lebesgue null set.

(2) For all  $y \in \mathbb{X}$ , there is a finite number of  $\gamma$ -intervals.  $\square$

It was already mentioned that the longest interval criterion suffers from the disadvantage that the  $\gamma$ -interval corresponding to the constant estimator has to be excluded. Another problem represent data with ‘trend’ where this term has to be explained. Our interpretation is that the longest interval criterion catches the essential feature - here it is the trend - but not the features of the signal on a smaller scale. At the present state, there are no rigorous results, we will consider several examples of idealized data. In these cases, there is only a small number of  $\gamma$ -intervals and the computation is done quickly.

In Example 5.3.3, we consider a signal which has, loosely spoken, no trend. In this case, the only finite  $\gamma$ -interval is the correct one and its right boundary  $\gamma_0(y)$  increases in  $N$ . Hence, the correct  $\gamma$ -interval enlarges and finally takes the whole  $\gamma$ -range if the length  $N$  of data goes to infinity.

In Example 5.3.4, we consider data with trend where the jumps are all of the same height. We observe that the  $\gamma$ -interval with one jump - catching the trend - is always the longest one, independently of how large  $N$  is.

We now want to extract how different jump heights affect the longest interval in the  $\gamma$ -scanning. In Example 5.3.5, the length of the correct interval does not depend on the factor  $k$  by which one jump is higher, whereas the length of the interval with one jump increases in  $k$ . We conclude that in such cases the longest  $\gamma$ -interval is always the one of the trend-catching estimator with one jump at the position of the highest jump. The higher the jump, the longer the corresponding  $\gamma$ -interval is.

Finally, we consider data where the jumps are of the same height but the lengths of the intervals in the partition are different. From Example 5.3.6 we conclude that in such a symmetric setting the longest interval criterion yields the correct estimator if the length of the plateau separating the two jumps is large enough.

These examples demonstrate the difficulties of the longest interval criterion with data with trend.

The rest of the section contains the announced examples for idealized data. In the first example we consider a signal without trend.

**Example 5.3.3 (no trend)** We consider data as in Figure 5.1 with four plateaus of the same length and three jumps of the same height. The direction of the jumps is alternating such that we would say that this signal has no trend. We will denote by  $E(l \text{ jumps at } \dots)$  the value of the Potts functional

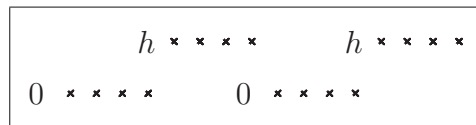


Figure 5.1: No trend, 4 plateaus of same length  $N/4$ , jumps of the same height  $h$

$x \mapsto H_\gamma(x, y) = \gamma \cdot |J(x)| + \sum_{i=1}^N (y_i - x_i)^2$  for an estimate  $x$  with  $l$  jumps at the locations  $\dots$ . It is

$$\bar{y} = \frac{h}{2} \quad \text{and} \quad \sum_{i=1}^N y_i^2 = \frac{1}{2}Nh^2.$$

In the case of idealized data, a minimum value is only achieved if the jump set of the estimate is contained in the jump set of data  $y$ . Hence, the values

to be considered are reduced to the following few.

$$\begin{aligned}
E(\text{no jump}) &= \sum_{i=1}^N (y_i^2 - \bar{y})^2 = \sum_{i=1}^N y_i^2 - N\bar{y}^2 = \frac{1}{4}Nh^2 \\
E(1 \text{ jump at } \frac{N}{4}) &= \gamma + \frac{1}{6}Nh^2 = E(1 \text{ jump at } \frac{3N}{4}) \text{ (symmetry)} \\
E(2 \text{ jumps at } \frac{N}{4}, \frac{N}{2}) &= 2\gamma + \frac{1}{8}Nh^2 \\
&= E(2 \text{ jumps at } \frac{N}{2}, \frac{3N}{4}) \text{ (symmetry)} \\
&= E(2 \text{ jumps at } \frac{N}{4}, \frac{3N}{4}) \text{ (symmetry)} \\
E(3 \text{ jumps at } \frac{N}{4}, \frac{N}{2}, \frac{3N}{4}) &= 3\gamma
\end{aligned}$$

The values  $\gamma_i(y)$  from Theorem 2.4.5 where the number of jumps changes are then obtained by

$$\begin{aligned}
E(0) = E(1) &\Leftrightarrow \gamma = \frac{1}{12}Nh^2, \\
E(0) = E(2) &\Leftrightarrow \gamma = \frac{1}{16}Nh^2, \\
E(0) = E(3) &\Leftrightarrow \gamma = \frac{1}{12}Nh^2 =: \gamma_{0 \rightarrow 3}.
\end{aligned}$$

Thus, in the scanning there are only two  $\gamma$ -intervals, namely the one corresponding to the constant estimator (no jumps) and the one corresponding to data (three jumps). The only finite  $\gamma$ -interval therefore is the correct one with three jumps. The boundary  $\gamma_0(y) = \gamma_{0 \rightarrow 3}$  of the correct interval increases in  $N$ . Hence, the correct  $\gamma$ -interval enlarges and finally takes the whole  $\gamma$ -range if the number  $N$  of data goes to infinity.

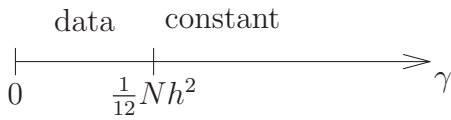


Figure 5.2:  $\gamma$ -scanning

Now we consider data with trend.

**Example 5.3.4 (trend)** The data in Figure 5.3 have also four plateaus of the same length and three jumps of the same height. In contrast to the previous example, the direction of the jumps is always the same such that we

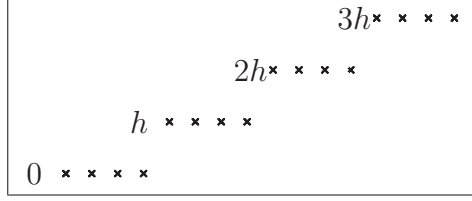


Figure 5.3: Trend, 4 plateaus of the same length  $N/4$ , jumps of the same height  $h$

will say that the signal has a trend. The mean value and the sum of squares are

$$\bar{y} = \frac{3}{2}h \quad \text{and} \quad \sum_{i=1}^N y_i^2 = \frac{7}{2}Nh^2.$$

The values of the Potts functional for the possible scenarios of jump sets are

$$\begin{aligned} E(\text{no jump}) &= \frac{5}{4}Nh^2, \\ E(1 \text{ jump at } \frac{N}{4}) &= \gamma + \frac{1}{2}Nh^2 = E(1 \text{ jump at } \frac{3N}{4}) \text{ (symmetry),} \\ E(1 \text{ jump at } \frac{N}{2}) &= \gamma + \frac{1}{4}Nh^2, \\ E(2 \text{ jumps at } \frac{N}{4}, \frac{N}{2}) &= 2\gamma + \frac{1}{8}Nh^2, \\ &= E(2 \text{ jumps at } \frac{N}{2}, \frac{3N}{4}) \text{ (symmetry),} \\ &= E(2 \text{ jumps at } \frac{N}{4}, \frac{3N}{4}) \text{ (symmetry),} \\ E(3 \text{ jumps at } \frac{N}{4}, \frac{N}{2}, \frac{3N}{4}) &= 3\gamma. \end{aligned}$$

The  $\gamma$ -values are obtained by

$$\begin{aligned} E(0) = E(1) &\Leftrightarrow \gamma = Nh^2, \\ E(0) = E(2) &\Leftrightarrow \gamma = \frac{9}{16}Nh^2, \\ E(0) = E(3) &\Leftrightarrow \gamma = \frac{5}{12}Nh^2, \\ E(1) = E(2) &\Leftrightarrow \gamma = \frac{1}{8}Nh^2, \\ E(1) = E(3) &\Leftrightarrow \gamma = \frac{1}{8}Nh^2, \\ E(2) = E(3) &\Leftrightarrow \gamma = \frac{1}{8}Nh^2. \end{aligned}$$

Hence, there are three  $\gamma$ -intervals, corresponding to the constant estimator,



Figure 5.4:  $\gamma$ -scanning

to the estimator with one jump, and to data. The length  $l$  of the two finite  $\gamma$ -intervals is

$$l(\text{1 jump}) = Nh^2 - \frac{1}{8}Nh^2 = \frac{7}{8}Nh^2,$$

$$l(\text{3 jumps}) = \frac{1}{8}Nh^2 - 0 = \frac{1}{8}Nh^2.$$

The longest finite one is thus always the one corresponding to the estimator with one jump.

In the next example we vary the height of the jumps.

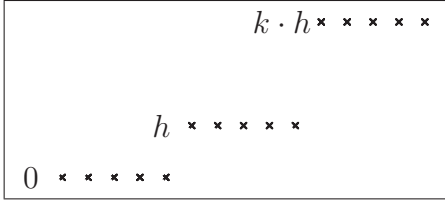


Figure 5.5: Trend, 3 plateaus of the same length  $N/3$ , jumps of different height,  $k > 2$

**Example 5.3.5 (trend plus variable jump height)** The data displayed in Figure 5.5 have three plateaus of the same length and two jumps of different height. Without restriction, we assume that the second jump is higher, i. e. we consider only the case  $k > 2$ . We get

$$\bar{y} = \frac{k+1}{3}h \quad \text{and} \quad \sum_{i=1}^N y_i^2 = \frac{k^2+1}{3}Nh^2.$$

Further, we have

$$E(\text{no jump}) = \frac{2(k^2 - k + 1)}{9}Nh^2,$$

$$E(\text{1 jump at } \frac{N}{3}) = \gamma + \frac{(k-1)^2}{6}Nh^2,$$

$$E(\text{1 jump at } \frac{2N}{3}) = \gamma + \frac{1}{6}Nh^2,$$

$$E(\text{2 jumps at } \frac{N}{3}, \frac{2N}{3}) = 2\gamma.$$

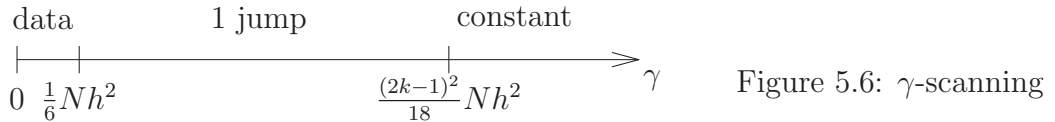
For  $k > 2$  we always have

$$E(1 \text{ jump at } \frac{N}{3}) > E(1 \text{ jump at } \frac{2N}{3}).$$

This is intuitively clear since one would set a single jump at the position  $2N/3$  of the higher jump of data. The  $\gamma$ -values are given by

$$\begin{aligned} E(0) = E(1) &\Leftrightarrow \gamma = \frac{(2k-1)^2}{18}Nh^2 =: \gamma_{0 \rightarrow 1}, \\ E(0) = E(2) &\Leftrightarrow \gamma = \frac{k^2 - k + 1}{9}Nh^2 =: \gamma_{0 \rightarrow 2}, \\ E(1) = E(2) &\Leftrightarrow \gamma = \frac{1}{6}Nh^2 =: \gamma_{1 \rightarrow 2}. \end{aligned}$$

For  $k > 2$  we have always  $\gamma_{0 \rightarrow 1} > \gamma_{0 \rightarrow 2}$ . Hence, the  $\gamma$ -interval left to the one of the constant estimator is the one of the estimator with one jump. Again,



there are three  $\gamma$ -intervals. The lengths of the finite intervals are

$$\begin{aligned} l(1 \text{ jump}) &= \gamma_{0 \rightarrow 1} - \gamma_{1 \rightarrow 2} = \frac{(2k-1)^2}{18}Nh^2 - \frac{1}{6}Nh^2 = \frac{(2k-1)^2 - 3}{18}Nh^2, \\ l(2 \text{ jumps}) &= \gamma_{1 \rightarrow 2} - 0 = \frac{1}{6}Nh^2. \end{aligned}$$

Note that the length of the correct interval with two jumps does not depend on  $k$  whereas the length of the interval with one jump increases in  $k$ . For  $k > 2$ , the  $\gamma$ -interval corresponding to the estimator with one jump is always the longest.

In the following example, the jumps are of the same height but the length of the plateaus varies.

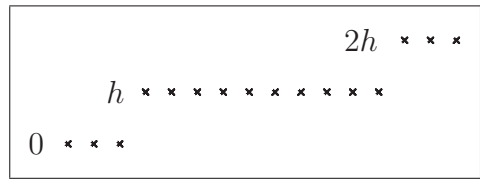


Figure 5.7: Trend, different lengths ( $3N/16$  resp.  $5N/8$  resp.  $3N/16$ ) of the 3 plateaus, same height  $h$  of jumps

**Example 5.3.6 (trend plus different plateau length)** We consider the data from Figure 5.7. Assume that the length of the middle plateau is  $N/k$ ,  $k > 1$ . Hence, the two others are each of length  $(k-1)N/2k$ . We get

$$\bar{y} = h \quad \text{and} \quad \sum_{i=1}^N y_i^2 = \frac{2k-1}{k}Nh^2.$$

The minimum values are

$$\begin{aligned} E(\text{no jump}) &= \frac{k-1}{k}Nh^2, \\ E(1 \text{ jump at } \frac{k-1}{2k}N) &= \gamma + \frac{k-1}{k(k+1)}Nh^2, \\ &= E(1 \text{ jump at } \frac{k+1}{2k}N) \text{ (symmetry),} \\ E(2 \text{ jumps at } \frac{k-1}{2k}N, \frac{k+1}{2k}N) &= 2\gamma. \end{aligned}$$

For the  $\gamma$ -values in the scanning we get

$$\begin{aligned} E(0) = E(1) \Leftrightarrow \gamma &= \frac{k-1}{k+1}Nh^2 =: \gamma_{0 \rightarrow 1}, \\ E(0) = E(2) \Leftrightarrow \gamma &= \frac{k-1}{2k}Nh^2 =: \gamma_{0 \rightarrow 2}, \\ E(1) = E(2) \Leftrightarrow \gamma &= \frac{k-1}{k(k+1)}Nh^2 =: \gamma_{1 \rightarrow 2}. \end{aligned}$$

Under the assumption  $k > 1$  it is always  $\gamma_{0 \rightarrow 1} > \gamma_{0 \rightarrow 2}$ . The lengths of the finite  $\gamma$ -intervals are

$$\begin{aligned} l(1 \text{ jump}) &= \gamma_{0 \rightarrow 1} - \gamma_{1 \rightarrow 2} = \frac{k-1}{k+1}Nh^2 - \frac{k-1}{k(k+1)}Nh^2 = \frac{(k-1)^2}{k(k+1)}Nh^2, \\ l(2 \text{ jumps}) &= \gamma_{1 \rightarrow 2} - 0 = \frac{k-1}{k(k+1)}Nh^2. \end{aligned}$$

We get

$$l(1 \text{ jump}) < l(2 \text{ jumps}) \Leftrightarrow 1 < k < 2.$$

Thus the longest interval is the correct one if the plateau separating the two jumps is long enough.



# Chapter 6

## Stopping Criteria

P. L. DAVIES and A. KOVAC (2001) suggest the criteria presented in the subsequent sections to stop their iterative ‘taut string’ algorithm. In this chapter, we show that these stopping criteria may provide data adapted parameter choices. There are at least two classes of stopping criteria. Those of this chapter impose requirements on the residuals. In contrast to this, the criteria presented in Section 8.3 rely on properties of the estimators itself.

The conditions considered here decide whether the residuals of an estimator can be classified as noise. By the derived criteria we can stop for example an iterative procedure like the one presented in Section 8.1. They can also be used to terminate the scanning through the range of the hyperparameter starting from infinity and hence are a method to choose  $\gamma$ .

The *residuals*  $(r_s)_{s \in S}$  of an estimate  $\hat{y}$  are given by  $r_s = \hat{y}_s - y_s$ . We will consider the residuals of a minimizer  $x^*(\gamma, y)$  of  $\bar{H}_\gamma(\cdot, y)$  from (4.1) given by

$$r(\gamma, y) = x^*(\gamma, y) - y. \quad (6.1)$$

### 6.1 Longest Run Criterion

The criterion presented in this section looks for the longest run of signs of the residuals.

**Definition 6.1.1** *The **longest run condition** is given by*

$$\max\{|I| : I \in \mathcal{P}(\text{sgn}(r))\} \leq R$$

where  $\mathcal{P}(\text{sgn}(r))$  denotes the partition induced by  $\text{sgn}(r) = (\text{sgn}(r_s))_{s \in S}$  of the signs of the residuals and  $R$  is some given number.

The value for  $R$  we use is

$$R = \lceil \log_2 N - 1.47 \rceil. \quad (6.2)$$

This is a suggestion from P. L. DAVIES and A. KOVAC (2001) and a justification for this value can be found there.

Now we will embed this condition in the concept of data adapted parameter choices.

**Definition 6.1.2** *The longest run criterion is given by*

$$\Gamma^{LR}(y) = \max \left\{ \gamma \in \mathcal{G}(y) : \max \left\{ |I| : I \in \mathcal{P}(\text{sgn}(r(\gamma, y))) \right\} \leq R \right\}$$

for some given  $R$ .

We will see that it leads to an estimator which is equivariant with respect to  $\text{Aff}(\mathbb{R})$ . We will first investigate how the residuals are transformed when  $\text{Aff}(\mathbb{R})$  acts on  $\mathbb{X}$ .

**Lemma 6.1.3** *Let  $x^*(\gamma, y)$  denote a minimizer of the Potts functional (4.1). Let further  $r(\gamma, y)$  be the residuals of  $x^*(\gamma, y)$ . Then*

$$r(\gamma, t_{b,c}(y)) = c \cdot r\left(\frac{\gamma}{c^2}, y\right)$$

for each scale transformation  $t_{b,c}$ .

**Proof** Using the scaling property (4.3) of minimizers of  $\bar{H}_\gamma(\cdot, y)$  from Theorem 4.1.5 we get

$$\begin{aligned} r(\gamma, t_{b,c}(y)) &= x^*(\gamma, t_{b,c}(y)) - t_{b,c}(y) = t_{b,c}(x^*(\frac{\gamma}{c^2}, y)) - t_{b,c}(y) \\ &= c \cdot x^*(\frac{\gamma}{c^2}, y) + b\mathbf{1} - (cy + b\mathbf{1}) = c \cdot r(\frac{\gamma}{c^2}, y) \end{aligned}$$

which is the assertion.  $\square$

We immediately get that the longest run criterion leads to equivariant estimators.

**Theorem 6.1.4** *The longest run criterion fulfills*

$$\Gamma^{LR}(t_{b,c}(y)) = \frac{\Gamma^{LR}(y)}{c^2}.$$

*In particular, the estimator  $y \mapsto x^*(\Gamma^{LR}(y), y)$  is equivariant with respect to  $\text{Aff}(\mathbb{R})$ .*

**Proof** The sign of the residuals of  $x^*(\gamma, t_{c,b}(y))$  is  $\text{sgn}(c) \cdot \text{sgn}(r(\gamma/c^2, y))$ , hence they all change simultaneously or stay the same. By Lemma 6.1.3 we thus have that

$$\max \left\{ |I| : I \in \mathcal{P} \left( \text{sgn}(r(\gamma, t_{b,c}(y))) \right) \right\} \leq R$$

if and only if

$$\max \left\{ |I| : I \in \mathcal{P} \left( \text{sgn}(r(\gamma/c^2, y)) \right) \right\} \leq R$$

and transformation of  $\gamma$  yields the stated equality.  $\square$

## 6.2 Multiresolution Criterion

The criterion presented in this section is based on a multiresolution analysis of the residuals.

**Definition 6.2.1** *The **multiresolution coefficients** are defined as*

$$w_{i,j} = \frac{1}{\sqrt{j-i+1}} \sum_{k=i}^j (y_k - \hat{y}_k) \quad \text{for all } 1 \leq i \leq j \leq N.$$

*The **multiresolution condition** then is given by*

$$|w_{i,j}| \leq \sqrt{\tau \ln(N)} \cdot \hat{\sigma}$$

*where  $\tau$  is some positive number, and for the standard deviation  $\sigma$  of data  $y$  the estimator*

$$\hat{\sigma} = \frac{1.4862}{\sqrt{2}} \cdot \text{median}(|y_s - y_{s-1}|, s = 2, \dots, N) \quad (6.3)$$

*is used.*

P. L. DAVIES and A. KOVAC (2001) suggest to choose the parameter  $\tau$  between 2 and 2.5.

Choosing the hyperparameter such that the multiresolution condition is fulfilled can be formulated as data adapted parameter choice.

**Definition 6.2.2** *The **multiresolution criterion** is given by*

$$\Gamma^{MR}(y) = \max \left\{ \gamma \in \mathcal{G}(y) : \right. \\ \left. |w_{i,j}(\gamma, y)| \leq \sqrt{\tau \ln(N)} \cdot \hat{\sigma}(y) \text{ for all } 1 \leq i \leq j \leq N \right\}$$

*for some given  $\tau$ .*

Minimizers of  $\bar{H}_\gamma(\cdot, y)$  with the hyperparameter chosen by the multiresolution criterion also lead to equivariant estimators.

**Theorem 6.2.3** *The multiresolution criterion fulfills*

$$\Gamma^{MR}(t_{b,c}(y)) = \frac{\Gamma^{MR}(y)}{c^2}.$$

*In particular, the estimator  $y \mapsto x^*(\Gamma^{MR}(y), y)$  is equivariant with respect to  $\text{Aff}(\mathbb{R})$ .*

**Proof** By Lemma 6.1.3, we get for the multiresolution coefficients

$$|w_{i,j}(\gamma, t_{b,c}(y))| = |c| \cdot |w_{i,j}(\gamma/c^2, y)|,$$

and for the estimated standard deviation we arrive at

$$\hat{\sigma}(t_{b,c}(y)) = \frac{1.4862}{\sqrt{2}} \cdot \text{Median}(|cy_s + b - cy_{s-1} - b|, s = 2, \dots, N) = |c| \cdot \hat{\sigma}(y).$$

The requirement

$$|w_{i,j}(\gamma, t_{b,c}(y))| \leq \hat{\sigma}(t_{b,c}(y)) \cdot \sqrt{\tau \cdot \ln N}$$

is then equivalent to

$$|c| \cdot |w_{i,j}(\gamma/c^2, y)| \leq |c| \cdot \hat{\sigma}(y) \cdot \sqrt{\tau \cdot \ln N}$$

which proves the assertion.  $\square$

These criteria may also be used to stop iterative procedures like the one presented in Section 8.1.

# Chapter 7

## Model Selection Criteria

In this chapter, we connect classical model selection criteria and the concept of data adapted parameter choices. We will identify the estimators obtained by the application of the Akaike and the Schwarz information criteria with MAP estimators for the Potts functional (4.1) for a special choice of  $\gamma$ .

We assume that the true deterministic signal  $x$  is corrupted by additive Gaussian white noise, i. e.

$$y_s = x_s + \xi_s(\omega), \quad s = 1, \dots, N, \quad (7.1)$$

where  $\xi_s$ ,  $s = 1, \dots, N$ , are independent and identically distributed normal random variables with mean zero and variance  $\sigma^2$ .

The choice of a  $\gamma$ -interval is equivalent to the determination of the number of intervals in the partition of the segmentation induced by  $x^*(\gamma, y)$ . This number can be interpreted as the dimension of the parameter in the family of simplest regression models given by the log likelihood functions

$$\ln L(\theta^k|Y) = -\frac{N}{2} \ln(2\pi\sigma_k^2) - \frac{1}{2\sigma_k^2} \|y - \mu^k\|^2. \quad (7.2)$$

This family is not only simple but will turn out to be the proper class of models for the Potts functionals with sum of squares in the data term. We will denote by  $\mathcal{P}_k^* = \{I_1, \dots, I_k\}$  a partition in  $\mathfrak{P}_k$  which minimizes the term  $\sum_{I \in \mathcal{P}} \sum_{s \in I} (y_s - \bar{y}_I)^2$  in  $\mathcal{P} \in \mathfrak{P}_k$ . The maximum likelihood estimators of the parameter vector  $\theta^k = (\mu^k, \sigma_k^2)$  are then given by

$$\hat{\mu}^k(y) = \left( \underbrace{\bar{y}_{I_1}, \dots, \bar{y}_{I_1}}_{|I_1|}, \dots, \underbrace{\bar{y}_{I_k}, \dots, \bar{y}_{I_k}}_{|I_k|} \right) \quad (7.3)$$

and

$$\hat{\sigma}_k^2(y) = \frac{1}{N} \|y - \hat{\mu}^k(y)\|^2 = \frac{1}{N} \sum_{I \in \mathcal{P}_k^*} \sum_{s \in I} (y_s - \bar{y}_I)^2. \quad (7.4)$$

There are several criteria to select and reduce the parameter dimension. We consider two classical model selection criteria. Let  $L(\hat{\theta}^k|Y)$  denote the likelihood function of the model with parameter  $\theta$  evaluated at the maximum likelihood estimator  $\hat{\theta}^k = (\hat{\theta}_1^k, \dots, \hat{\theta}_k^k)$  in the subspace  $\Theta^k$  of the parameter space  $\Theta$ . Let  $k$  be the number of parameters to be estimated.

## 7.1 The Akaike Information Criterion

H. AKAIKE (1973, 1974) suggested an information criterion (AIC) of the following form: Maximize the log likelihood function separately for the competing models and choose the model for which

$$\text{AIC}(k) = \ln L(\hat{\theta}^k|Y) - k \quad (7.5)$$

is largest. This has become known as the *Akaike information criterion*. It is based on the minimization of the Kullback-Leibler information. Some more details and a derivation of a corrected version can be found in Appendix B.2. The relation to MAP estimators is the following. Suppose that the variance  $\sigma^2$  is known. Then, in case of the set of candidate models given by the family of densities (7.2), the Akaike information criterion reads: Maximize

$$\begin{aligned} \text{AIC}(k) &= \ln L(\hat{\mu}^k|Y) - k \\ &= -\frac{N}{2} \ln(2\pi\sigma^2) - \frac{1}{2\sigma^2} \sum_{I \in \mathcal{P}_k^*} \sum_{s \in I} (y_s - \bar{y}_I)^2 - k. \end{aligned}$$

The first term can be ignored since it is the same for all competing models. Hence, with known variance  $\sigma^2$ , the Akaike information criterion amounts to the maximization of

$$k \longmapsto -\frac{1}{2\sigma^2} \sum_{I \in \mathcal{P}_k^*} \sum_{s \in I} (y_s - \bar{y}_I)^2 - k \quad (7.6)$$

for  $1 \leq k \leq N$ . We realize the following connection of AIC to MAP estimation.

**Theorem 7.1.1** *Assume that the variance  $\sigma^2 > 0$  is known. Let  $k^*$  be a maximizer of (7.6). Then the corresponding maximum likelihood estimator  $\hat{\mu}^{k^*}$  minimizes the Potts functional  $\bar{H}(\cdot, y)$  in (4.1) for  $\gamma = 2\sigma^2$  and vice versa.*

**Proof** If the number  $k$  of intervals in a partition  $\mathcal{P}$  is fixed, a minimizer of the Potts functional is given by a minimizer of the data term. Since the data

term is a sum of squares, a minimizer maximizes the log likelihood function (7.2) on the space of signals with exactly  $k$  intervals in the induced partition. An estimator obtained by the maximization of (7.6) thus minimizes the Potts functional (4.1) for the special choice of  $\gamma = 2\sigma^2$ .  $\square$

## 7.2 The Schwarz Information Criterion

Further, we will consider the *Schwarz information criterion* (SIC) introduced by G. SCHWARZ (1978). He suggested to choose the model for which

$$\text{SIC}(k) = \ln L(\hat{\theta}^k|Y) - \frac{1}{2} k \ln N \quad (7.7)$$

is maximal. Some more details and a derivation of a corrected version can be found in Appendix B.3.

Similarly, we see that in case of known variance  $\sigma^2$  the maximization of the Schwarz information criterion (7.7) is equivalent to the maximization of

$$k \longmapsto -\frac{1}{2\sigma^2} \sum_{I \in \mathcal{P}_k^*} \sum_{s \in I} (y_s - \bar{y}_I)^2 - \frac{k}{2} \ln N. \quad (7.8)$$

Hence, for the Schwarz information criterion we obtain the following identification.

**Theorem 7.2.1** *Assume that the variance  $\sigma^2 > 0$  is known. Let  $k^*$  be a maximizer of (7.8). Then the corresponding maximum likelihood estimator  $\hat{\mu}^{k^*}$  minimizes the Potts functional (4.1) for  $\gamma = \sigma^2 \ln N$  and vice versa.*

**Proof** As in Theorem 7.1.1 for the Akaike information criterion, an estimator resulting from the maximization of (7.8) minimizes the Potts functional for  $\gamma = \sigma^2 \ln N$ .  $\square$

## 7.3 Equivariant Versions

Hence, these model selection criteria give a suggestion for the choice of  $\gamma$ . From Theorem 4.1.5 we know that this will not yield equivariant estimators. The idea is now to interpret them as data adapted parameter choices by inserting a suitable estimator  $\hat{\sigma}^2(y)$  for the variance. We then define

$$\Gamma^{AIC}(y) = 2\hat{\sigma}^2(y)$$

and

$$\Gamma^{SIC}(y) = \hat{\sigma}^2(y) \ln N.$$

We conclude that the estimators  $y \mapsto x^*(\Gamma^{AIC}(y), y)$  and  $y \mapsto x^*(\Gamma^{SIC}(y), y)$  are equivariant with respect to  $\text{Aff}(\mathbb{R})$  if the estimator  $\hat{\sigma}^2(y)$  fulfills

$$\hat{\sigma}^2(t_{b,c}(y)) = c^2 \hat{\sigma}^2(y).$$

Examples for such an estimator are the maximum likelihood estimator  $\hat{\sigma}_k^2$  from (7.4) or the estimator of the standard deviation from (6.3).

**Remark 7.3.1** These model selection criteria can be considered as a first step towards interval criteria. The application of these criteria corresponds to the reduction of the whole range of the hyperparameter to a finite discrete set of  $\gamma$ -values, and then to choose  $\gamma$  from this set by a certain rule. The  $k$ -th  $\gamma$ -value corresponds to the penalty term for the  $k$ -th model. The rule is the maximization of a functional representing AIC and SIC, respectively. In contrast to the interval criteria, the model selection criteria only choose between a finite set of  $\gamma$ -values whereas the interval criteria incorporate all values of  $\gamma$  between  $\gamma_{m(y)}(y)$  and  $\gamma_0(y)$ .

We will see that the classical model selection criteria, the Akaike information criterion as well as the Schwarz information criterion, are not appropriate for data sets like brain data from functional magnetic resonance imaging (see Section 9.1) or fractionation curves (presented in Section 9.2). These criteria more or less return data. One reason is that the model (7.1) and the set of candidate models given by the log likelihood functions in (7.2) is clearly not the adequate class of models for this kind of data. Then we might expect better results with other classes of models, but here we would have difficulties to derive the corrected versions of these criteria which decisively depend on the model assumptions, see Appendices B and C. On the other hand, these corrections are necessary since the original criteria rely on the asymptotics as the length  $N$  of data tends to infinity, and the time series' from fMRI experiments have only a length of 70 time points, respectively, we have only 29 time points for the fractionation curves. Moreover, our aim was not to study the performance of these criteria but to use them for the determination of  $\gamma$  in the Potts functional. And therefore, the family (7.2) is the adequate one.

# Chapter 8

## Further Ideas

This chapter contains further approaches to the choice of the hyperparameter. At the moment, they are simply ideas, there is no rigorous treatment. The iterative procedure presented in Section 8.1 is a first approach to overcome the problem that the longest interval criterion cannot deal with data with trend. Section 8.2 is concerned with the problem that the interval criteria are not constructed to give a constant estimate. With the morphological criteria from Section 8.3 we have a general frame for a criterion which was originally designed for the application of MAP estimators to the gene expression data from Section 9.2.

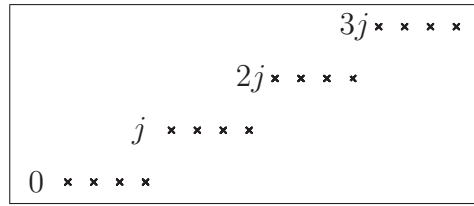
### 8.1 Iterative Procedures

In Section 5.3 we discussed the application of the longest interval criterion to data with trend. It turned out that in most of these cases this criterion suggests the estimator with exactly one jump. The iterative procedure suggested in the following is a first approach to deal with this problem. The idea is to decompose the signal into the essential features on different scales caught by the respectively longest  $\gamma$ -interval. Repeated application of the longest interval criterion from Section 5.3 gives a decomposition of the signal into components beginning with the most striking feature (such as a trend) to the details. The algorithm of this iterative procedure reads as follows:

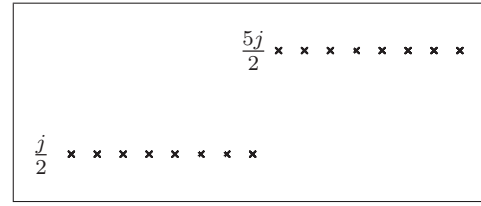
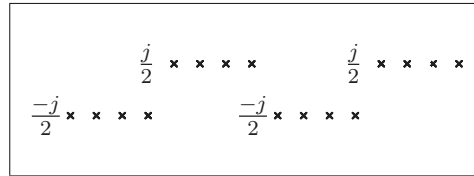
**Algorithm 8.1.1** Start with data  $y = y^{(0)}$ .

While some stopping condition is not fulfilled repeat the following steps.

- (1) Apply the longest interval criterion to  $y^{(i)}$  to get the estimate  $\hat{y}^{(i)}$ . Keep it as the  $i$ -th component of the signal.



(a) original data

(b) first component: estimate obtained by the longest interval in  $\gamma$ -scanning

(c) first residuals: difference of first component and data

Figure 8.1: Steps in the iterative procedure of Algorithm 8.1.1

- (2) Compute the residuals  $r_s^{(i)} = \hat{y}_s^{(i)} - y_s^{(i)}$ .
- (3) Check the stopping condition for the residuals  $r^{(i)}$ . If it is not fulfilled, treat the residuals as new data  $y^{(i+1)}$ .

Sum up the single components  $\hat{y}^{(i)}$  to get an estimate for the original data  $y$ .

An open problem is the mentioned stopping condition. A first approach is to use the stopping criteria from Chapter 6 but they are external criteria. It would be nicer to have an intrinsic criterion derived from the  $\gamma$ -scanning. The idea behind the iterative procedure is to continue until the number of jumps of the recomposed final estimate does not change any more. This can only be the case when the location of the jumps for two subsequent estimates is the same or when the subsequent estimate is a constant. For idealized data, in Example 8.1.2, the iterative procedure is stopped when the residuals are identically zero. This is only the case for data in a Lebesgue null subset of  $\mathbb{R}^N$ . Thus, we are faced with the problem to get the constant estimate which will be discussed in Section 8.2.

For the idealized data from Example 5.3.4, we perform the iterative procedure from Algorithm 8.1.1.

**Example 8.1.2** The data shown in Figure 8.1(a) have a trend, all plateaus are of the same length, and all jumps have the same height. We saw in Example 5.3.4 that the longest  $\gamma$ -interval for such a signal is that one with one jump at  $N/2$  displayed in Figure 8.1(b). Substraction of this estimate

from the original signal gives the data shown in Figure 8.1(c). This new signal has the same form as that one in Example 5.3.3. There is only one finite interval in the  $\gamma$ -scanning. Thus, the number of jumps immediately changes from zero to three, i. e. data is recovered. Here the procedure naturally stops.

We add the estimates of the longest (in the second case the only) finite interval from Figure 8.1(b) and Figure 8.1(c) of these two steps. This indeed gives back the original time series from Figure 8.1(a).

## 8.2 Constant Estimates

A reasonable estimator should map a nearly constant time series to a constant one. Our FLIC estimators do not have this important property. The  $\gamma$ -interval corresponding to the constant estimator extends to infinity and therefore its length cannot be compared to the length of the other  $\gamma$ -intervals. To handle this problem, one approach could be the following: Apply well-known statistical tests to exclude that data  $y$  arise from a constant underlying true signal or are only noise.

An intrinsic approach is to make use of the  $\gamma$ -scanning again: We investigate properties of the distribution of the random variable  $\gamma_0(Y)$  when  $Y$  is considered to a random vector. Further, we could try to approximate the density of  $\gamma_0(Y)$  in this case, and construct a test to decide whether the value of  $\gamma_0(y)$  at hand is a realization of  $\gamma_0(Y)$ .

## 8.3 Morphological Criteria

Sometimes we have morphological information about the signal. This leads to another class of stopping criteria. They decide whether the estimators, and not the residuals as in Chapter 6, have certain properties.

An important example is monotony. Hence, we impose the additional restriction that  $x^*(\gamma, y)$  is monotone.

**Example 8.3.1** Tracking  $x^*(\gamma, y)$  we find  $\gamma$ -intervals on which  $x^*(\gamma, y)$  is monotone and intervals where it is not. If we want to have a monotone signal, we take the hyperparameter  $\gamma$  as some point of the last interval (starting from the rightmost interval  $(\gamma_0(y), \infty)$ ) on which  $x^*(\gamma, y)$  is monotone. This gives an estimator  $y \mapsto x^*(\Gamma(y), y)$  with a data adapted parameter choice  $\Gamma(y)$ . Since each scale transformation  $t_{b,c}$  for  $c > 0$  preserves monotony such an estimator is equivariant with respect to the subgroup of  $\text{Aff}(\mathbb{R})$  with  $c > 0$ .



**Part III**  
**Application to Data**



Due to its simplicity, the Potts functionals are appropriate in situations where there is no or only little ground truth. Their minimizers are a suitable tool for the extraction of primitive signal features like plateaus and jumps. In Chapter 9, we will illustrate this by two data sets from life sciences. In view of these data sets one may doubt about too ‘specific’ methods or too detailed models for their analysis. In our data examples we expect that the observation period can be partitioned into intervals where the underlying signal can be represented by a constant. Therefore, it is reasonable to fit minimizers of the Potts functionals to this kind of data.

Chapter 10 contains a brief statistical survey of the different methods applied to simulated data.



# Chapter 9

## Data Sets from Life Sciences

In this chapter, we present two kinds of data from life sciences. We will apply the methods from the preceding chapters to exemplary data from functional magnetic resonance imaging and from fractionation experiments for cDNA experiments. Characteristic for both types of experiments is the lack of reliable information about the detailed shape of the curves as well as a suitable noise model. Hence, MAP estimators of the Potts functionals with suitable parameter choice seem to be adequate to extract the essential features from these data.

### 9.1 Functional Magnetic Resonance Imaging

The main target of human brain mapping is the non-invasive localization of functional areas in the human brain where certain outer stimuli are processed. This is done by identifying regions of increased activity in the human brain in response to outer stimuli. Typically such stimuli are boxcar shaped as indicated in Figure 9.1. They may represent ‘light or sound on and off’, i.e. visual or acoustic stimuli, or tactile ones like finger tipping on a desk. Functional magnetic resonance imaging (fMRI) exploits the blood oxygenation level dependent (BOLD) effect. This is basically a change of paramagnetic properties caused by an increase of blood flow in response to the

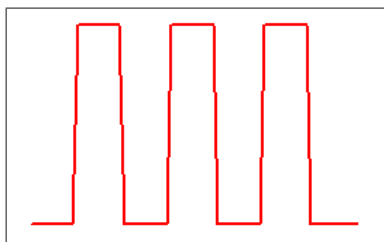


Figure 9.1: A boxcar shaped signal representing ‘on-off’ stimuli in fMRI brain mapping.

demand of activated neurons for more oxygen. The degradation mechanism along the path ‘(complex) eye - (highly complex) brain - (complicated) measuring device’ is only partially known. Moreover, measurement is indirect, since the recorded BOLD effect is a physiological quantity related to a local increase of blood flow and not a direct function of cortical activation.

The data at hand represent the time series of the intensities in one voxel along the duration of the experiment. Hence, one would call a voxel ‘active’ if the corresponding time series imitates the outer stimulus. This is thought to be boxcar shaped, as shown in Figure 9.1. Thus, the approach with the Potts functionals to get significant plateaus should be appropriate.

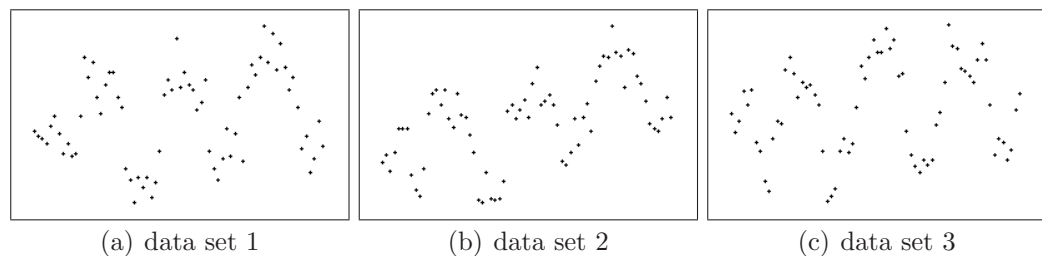


Figure 9.2: The three fMRI data sets.

Data displayed in Figure 9.2 are from an fMRI experiment performed by D. AUER, Max-Planck-Institute of Psychiatry, Munich. All three time series show the three characteristic bumps as in the representation of the outer stimulus in Figure 9.1.

Figures 9.3, 9.4, and 9.5 show the results of the application of the criteria introduced in Chapters 5, 6, 7, and 8, to these exemplary time series’.

First, we will briefly recapitulate the methods we used for the generation of the pictures. LIC estimator denotes the longest interval estimator from Definition 5.3.1, discussed in Section 5.3. This estimator is the one corresponding to the  $\gamma$ -interval for which the difference  $\gamma_{i-1}(y) - \gamma_i(y)$  is maximal. The log-longest interval estimator is the FLIC estimator from Definition 5.2.1 with  $F(x) = \ln x$ . Hence it is the one corresponding to the  $\gamma$ -interval for which the ratio  $\gamma_{i-1}(y)/\gamma_i(y)$  is maximal. For the model selection criteria we assumed the Gaussian model from (7.1). The corrected versions of the Akaike information criterion (AIC) and the Schwarz information criterion (SIC) are modification of the original criteria introduced in Section 7.1 and 7.2 for short time series’. Their derivation for this special class of models and the exact formulae can be found in Appendix B.2 and B.3, respectively.

The fMRI data have a length of 70 time points. Hence, the maximum of the allowed run length of the signs of the residuals for the longest run criterion

from Section 6.1, according to (6.2), is

$$R = \lceil \log_2 70 - 1.47 \rceil = 5.$$

For the multiresolution criterion from Section 6.2 we use the factor  $\tau = 2$ . The iterative procedure is the repeated application of the longest interval criterion as described in Algorithm 8.1.1. As stopping criterion we use the multiresolution criterion with factor  $\tau = 2$ .

We consider the results of these estimators for the fMRI data set 1 displayed in Figure 9.2(a). The longest interval estimate shown in Figure 9.3(b) yields the expected estimate with the characteristic bumps. It coincides with the estimate from the longest run criterion, see Figure 9.3(f), and the one from the multiresolution criterion, in Figure 9.3(g). Also the iterative procedure works well. Since already in the first iteration the residuals fulfill the multiresolution criterion the estimate in Figure 9.3(h) coincides with the LIC estimate. The log-longest interval criterion, see Figure 9.3(c), and the model selection criteria, see Figure 9.3(d) and Figure 9.3(e), basically return data.

Data set 2 displayed in Figure 9.2(b) has a slight trend. Here, by trend we will denote the fact that the ground levels of the bumps increase. This becomes evident in Figure 9.4(b) showing the LIC estimator. It gives the estimator with exactly one jump which catches this trend. This aligns with the theoretical considerations in Section 5.3 for data with trend. The estimators from the log-longest interval criterion, see Figure 9.4(c), and from the model selection criteria, in Figure 9.4(d) and Figure 9.4(e), return data. The longest run criterion, see Figure 9.4(f), and the multiresolution criterion, as shown in Figure 9.4(g), both provide sensible estimates. They have no problems with the trend. Thereby, the longest run criterion seems to be more restrictive than the multiresolution criterion, or, in other words, it resolves finer details. In the estimate displayed in Figure 9.4(h), obtained by repeated application of the longest interval criterion, one could find the rough structure of the data, namely three bumps, irrespectively of the trend. This is an indication that iteration helps to overcome the problem with data with trend. There are visible more details than for the multiresolution criterion which is used as stopping criterion. This is due to the fact that the final estimate does not appear in the scanning but is a sum of the curve in Figure 9.4(b) and further LIC estimates for the residuals.

Data set 3 shown in Figure 9.2(c) also has a slight trend. The longest interval estimate, see Figure 9.5(b), has no problem with the slight increase of the base line and provides a reasonable estimate. Again, the log-longest interval criterion and the model selection criteria, see Figure 9.5(d), Figure 9.5(e),

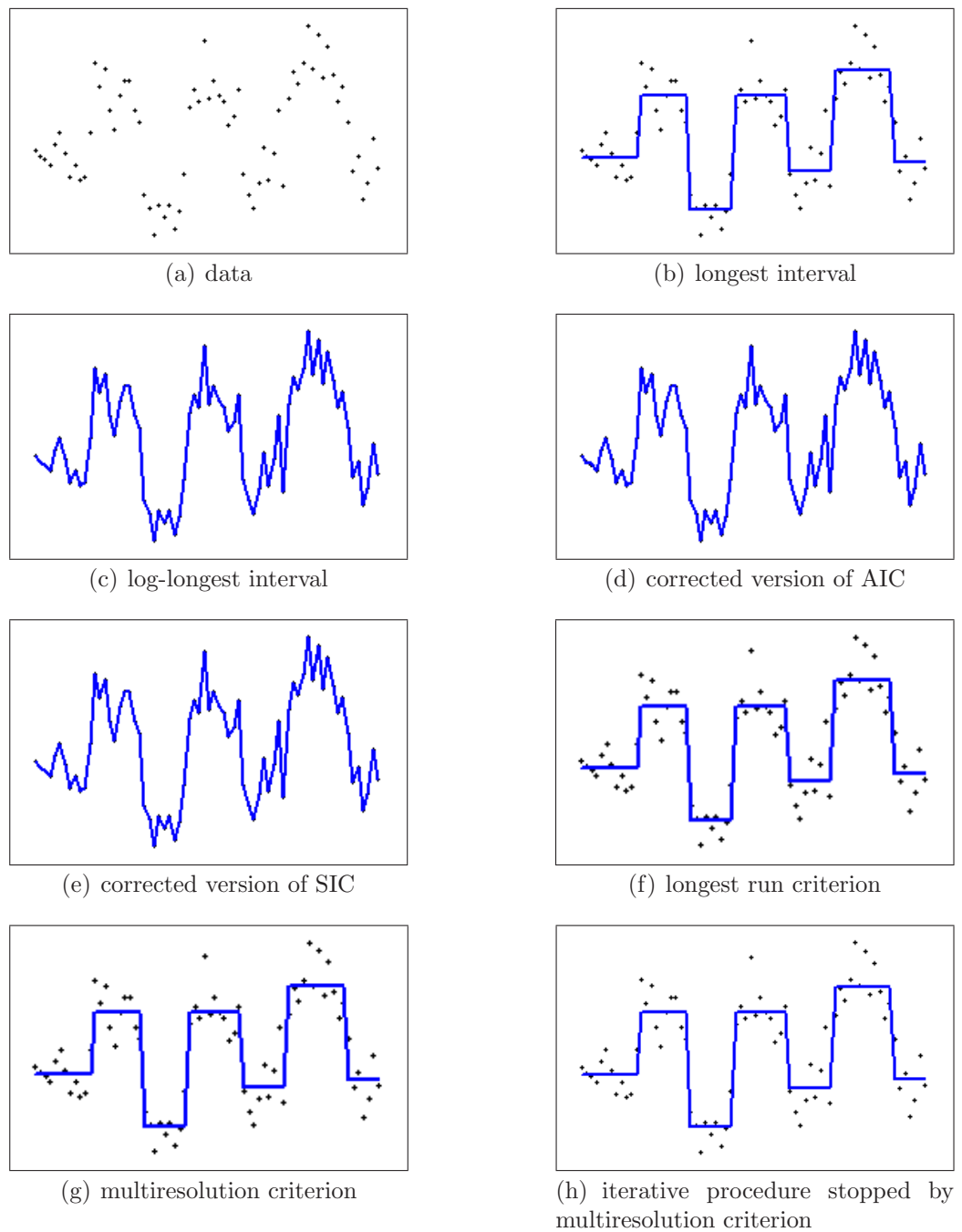


Figure 9.3: Application of different criteria to fMRI data set 1.

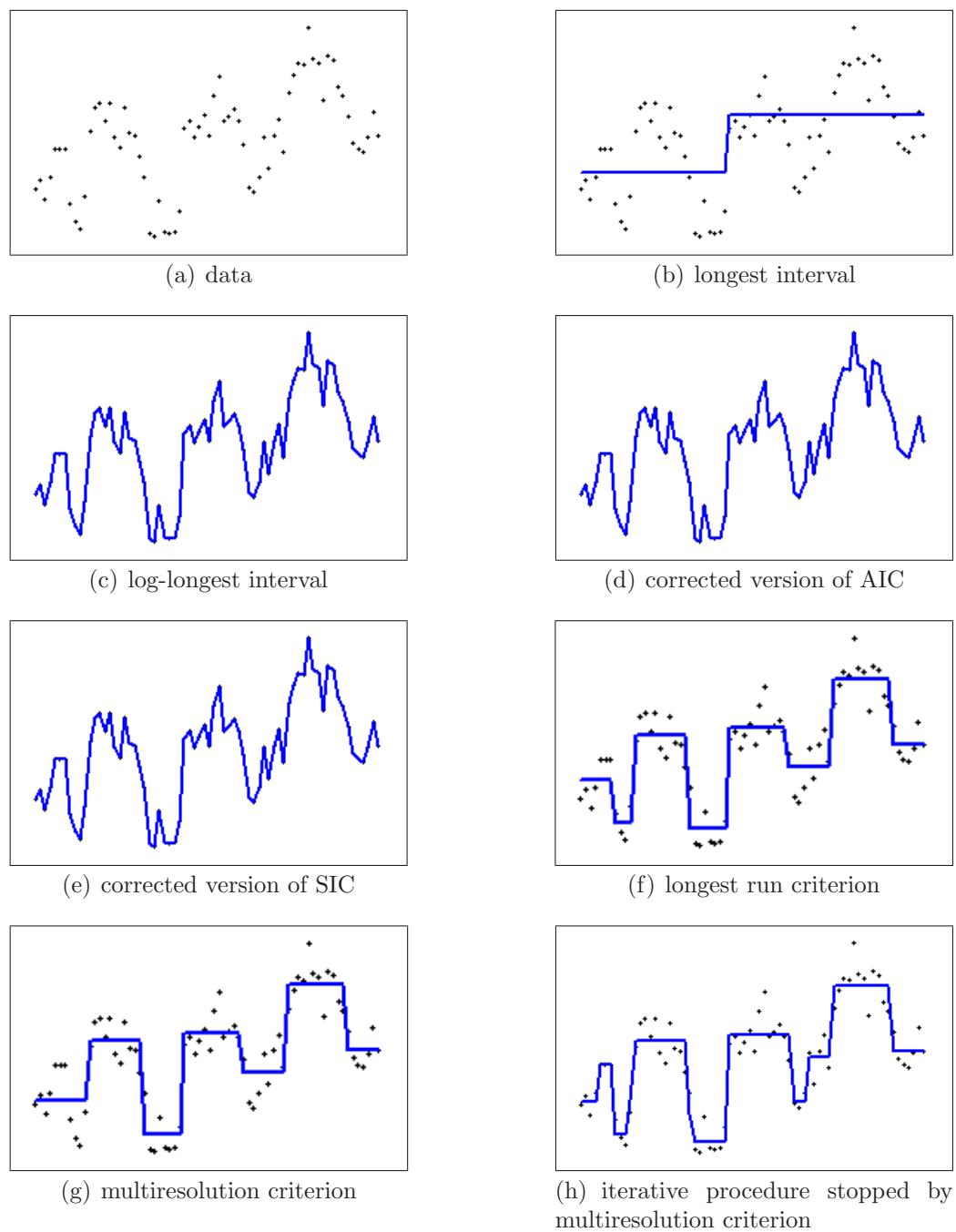


Figure 9.4: Application of different criteria to fMRI data set 2

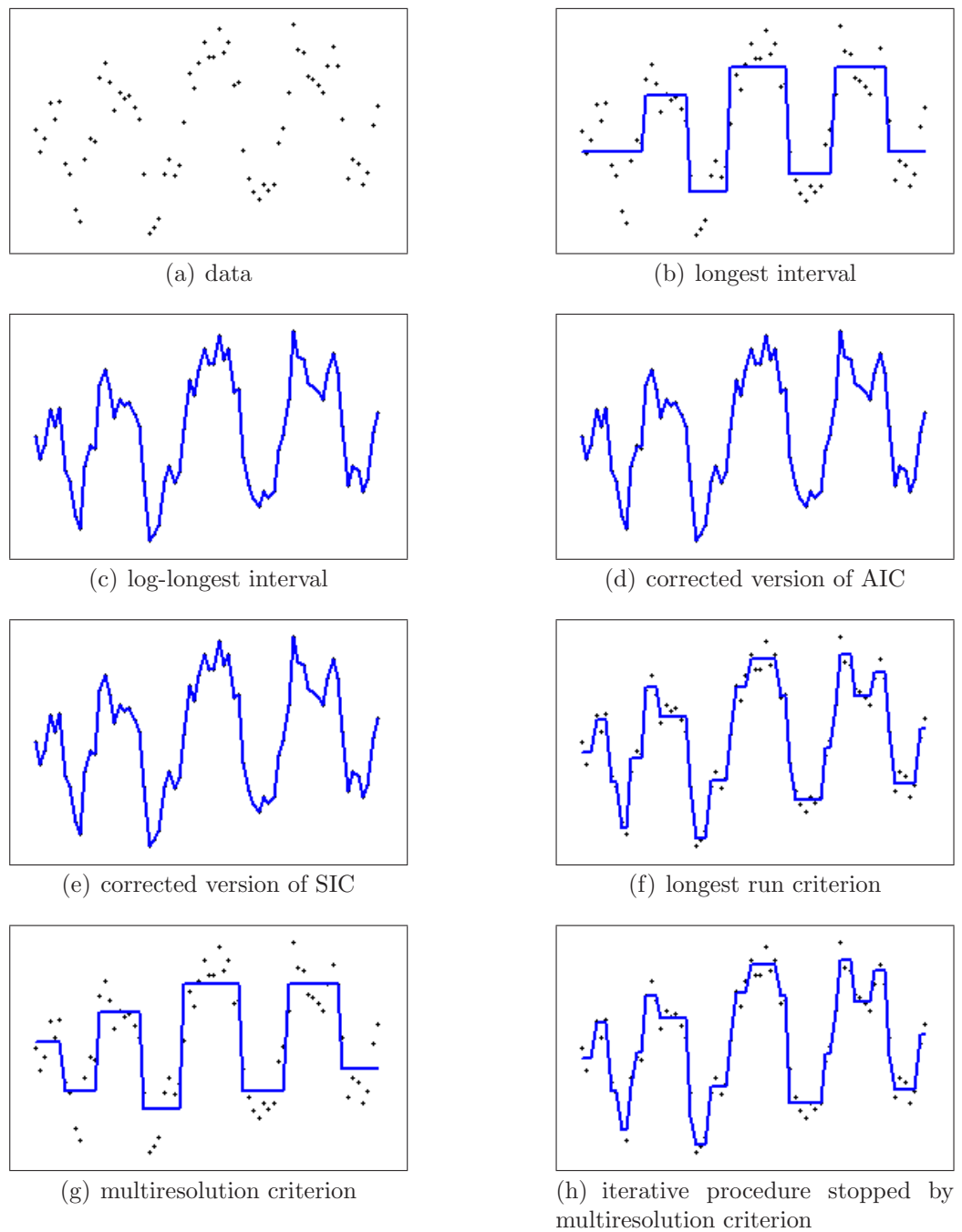


Figure 9.5: Application of different criteria to fMRI data set 3

and Figure 9.5(c), return data. The estimator of the longest run criterion incorporates quite a lot of details as shown in Figure 9.5(f). In Figure 9.5(b) we see that rather long sequences of data points stay above or under the estimate. Precisely these long runs (recall that the maximum of the allowed run length was  $R = 5$ ) are broken by the insertion of further intervals. In contrast, the estimator suggested by the multiresolution criterion, displayed in Figure 9.5(g), yields a sensible estimate. The beginning of a bump on the left should not be counted as a proper bump. The iterative longest interval performs similarly as the longest run criterion. Note that the resulting estimator in Figure 9.5(h) and the longest run estimate in Figure 9.5(f) are very similar. This is surprising since we used the multiresolution criterion to stop the iteration and not the longest run criterion.

In summary, for all three data sets we find two groups of estimators: the model selection criteria and the log-longest interval criterion returning data on the one hand, and the longest interval criterion and the iterative procedure together with the stopping criteria on the other hand. For data without trend, the methods of the latter group give the same reasonable estimate. As expected, in case of trend in data the longest interval criterion fails but the iterative procedure serves its purpose to overcome this problem. Considering further data sets, we have the impression that the longest interval criterion has tendency to insert rather less jumps than the others. This results in a more distinct representation of the outer stimulus if present.

## 9.2 Fractionation Experiments

The final aim of cDNA microarray experiments is to explore the structure of unknown genes. To this end, single stranded sections of *known* cDNA which are called *targets* are put on spots of microchips. A microchip typically consists of about 20.000 spots. Each section is a finite sequence of four nucleic acids, which are coded by the letters A(denin), C(ytosin), G(uanin), and T(hymin). If nucleic acids are added then they tend to bind to the targets where T binds to A, and G binds to C. Hence sections of single stranded *unknown* cDNA tend to pair with complementary DNA targets. The binding energy is maximal for perfect matches and such a perfect match means high stability. With perfect match the unknown sequence could be identified perfectly. A main problem is *cross-hybridization*, which means that DNA sections pair with DNA of similar, but not precisely equal, structure to the complementary sequence.

A new and innovative experiment provides data which hopefully will allow to identify mismatch dissociation. It is called ‘Specificity Assessment From

Fractionation Experiments’, or in short-hand notation, ‘SAFE’, see A.L. DROBYSHEV et al. (2003). It is plausible that the ‘melting temperature’ of double stranded DNA depends on length and contents of specific sequences. Further, increasing washing stringencies with *formamide* solutions has similar effects as increasing temperature since both decrease the binding energies. In the experiment, the chips are washed repeatedly (29 times) with formamide solutions of increasing concentration, and time series of intensities are recorded, called *fractionation curves*. Since the binding energy of cross-hybridizing cDNA is lower, it is washed away at lower concentrations. It is plausible that there is a critical concentration where a special kind of cDNA is abruptly washed away from the spot. Therefore, a statistical analysis should aim at the identification of locations and heights of abrupt decreases.

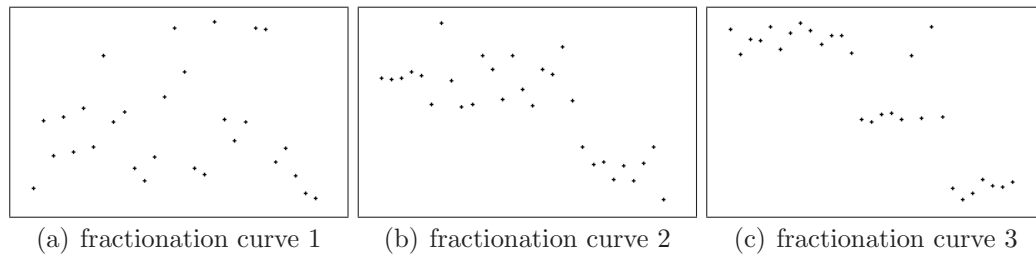


Figure 9.6: Three typical fractionation curves.

We will consider estimates for the three typical fractionation curves of single spots from such an experiment shown in Figure 9.6.

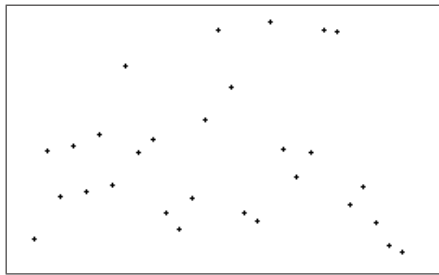
Figures 9.7, 9.8, and 9.9 show results of the application of the criteria introduced in Chapters 5, 6, 7, and 8, to these fractionation curves.

We use the same methods as for the fMRI data from Section 9.1. The length of the fractionation curves is 29 time points. For the longest run criterion, note that here the maximum of the allowed run length of the signs of the residuals, given by (6.2), is

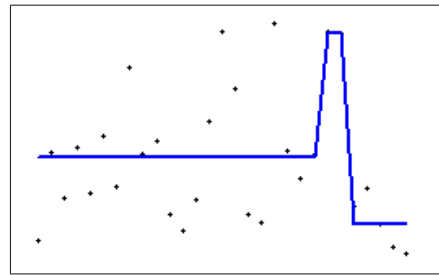
$$R = \lceil \log_2 29 - 1.47 \rceil = 4.$$

For the fractionation curves we have the additional restriction that by physical reasons the ‘true’ signal should be decreasing. Therefore, it is reasonable to apply here in addition the last monotone criterion from Example 8.3.1. It was designed especially for these data.

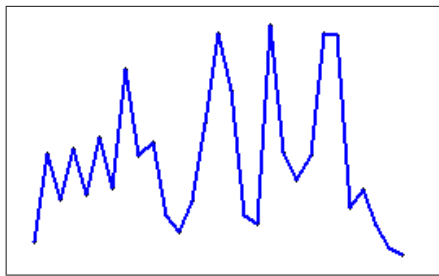
The data in Figure 9.6(a) do not show a striking decrease and by visual inspection one may classify them as noise. This means that under the special conditions there is almost no specific hybridization, the corresponding gene



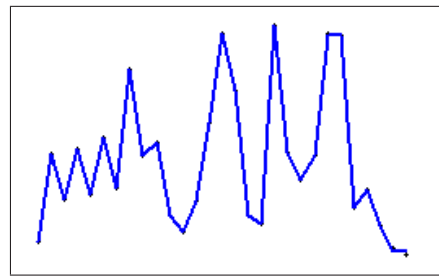
(a) data



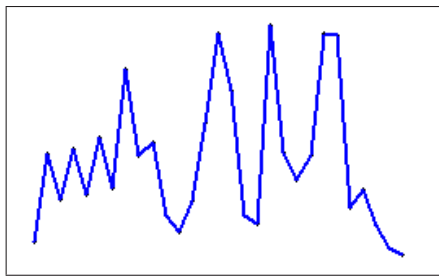
(b) longest interval



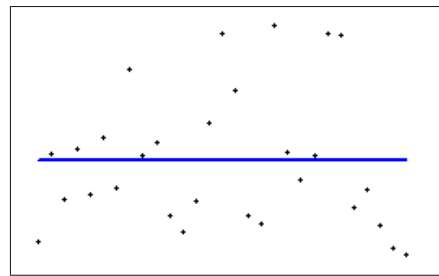
(c) log-longest interval



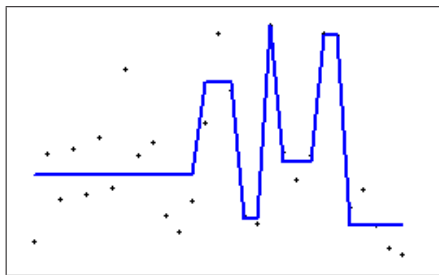
(d) corrected version of AIC



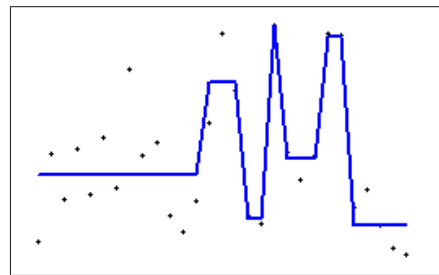
(e) corrected version of SIC



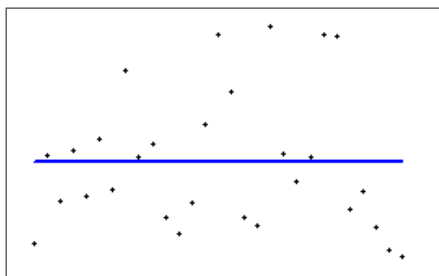
(f) longest run criterion



(g) multiresolution criterion



(h) iterative procedure stopped by multiresolution criterion



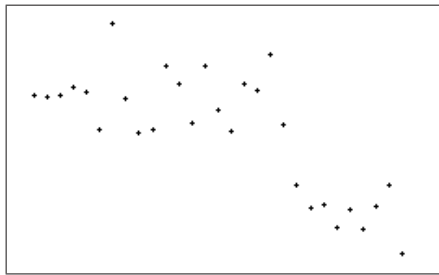
(i) last monotone criterion

Figure 9.7: Application of different criteria to the fractionation curve of a spot with no specific hybridization

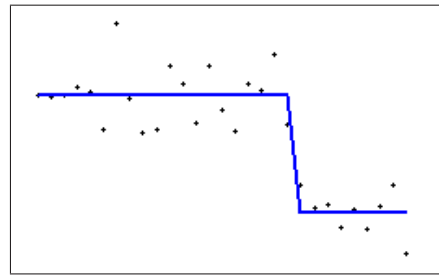
is not expressed. The longest interval criterion, which by default excludes the constant estimate, suggests an almost constant estimate with a narrow bump, see Figure 9.7(b). It seems to be rather a suggestion of embarrassment since a human inspector would not necessarily extract this bump as the essential feature of the signal. The log-longest interval, for the estimate see Figure 9.7(c), and the model selection criteria, see Figure 9.7(d) and Figure 9.7(e), return data. The reasons are the same as in case of the brain data. The longest run criterion yields a constant estimate, shown in Figure 9.7(f), which is a reasonable estimate for this spot. The data points seem to lie uniformly above and under the estimate. The estimate from the multiresolution criterion picks some details, see Figure 9.7(g). In the same way the estimate resulting from the repeated application of the longest interval criterion behaves, shown in Figure 9.7(h). It does not coincide with the estimate in Figure 9.7(g) - which should not be the case anyway - but has the same shape. The last monotone criterion yields the sensible constant estimate, see Figure 9.7(i).

The fractionation curve 2 in Figure 9.6(b) shows a quite clear cut and seems to jump down to another level. It corresponds to a fairly good spot where the right complementary cDNA bound. The location of the sharp decrease indicates that at this concentration the binding energy was amortized by the repeated washing. The longest interval criterion detects this jump exactly. The estimate from Figure 9.8(b) is the desired one and corresponds to what a trained observer would have suggested. The log-longest interval and the model selection criteria cannot provide anything else than data, see Figures 9.8(c), 9.8(d), and 9.8(e). As shown in Figure 9.8(f), the longest run criterion yields the desired estimator. Also the multiresolution criterion suggests the estimate which exactly one jump, see Figure 9.8(g). The iterative procedure stops at the first application of the longest interval criterion since the multiresolution criterion for the residuals is immediately fulfilled. Thus the estimator, see Figure 9.8(h), coincides with the LIC estimator. The last monotone estimate in the  $\gamma$ -scanning, displayed in Figure 9.8(i), is also the one with one jump. The estimate following in the  $\gamma$ -scanning inserts a peak which breaks the monotony.

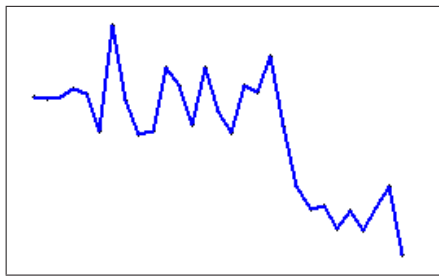
In the fractionation curve 3, shown in Figure 9.6(c), we clearly recognize three levels with two abrupt local breaks. This indicates that one part is washed away at one concentration level and another one at a second level. Hence, the time series corresponds to a spot with cross-hybridization. As we know from the examples from Section 5.3, the longest interval criterion cannot detect the gradual decrease, it provides the estimate with exactly one jump catching the main feature trend, see Figure 9.9(b). As we meanwhile would guess, the



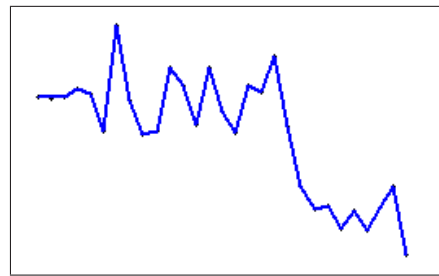
(a) data



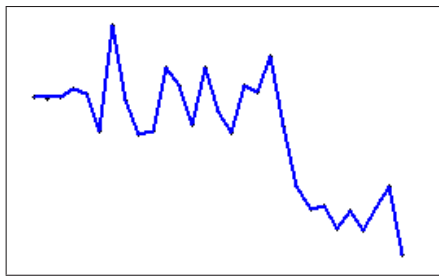
(b) longest interval



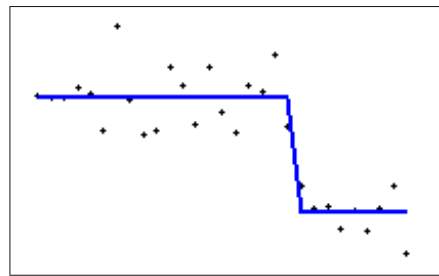
(c) log-longest interval



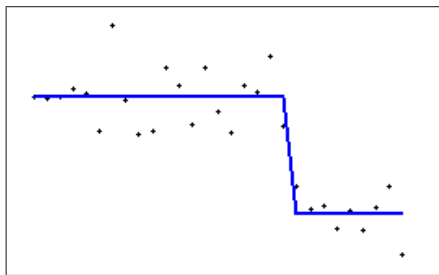
(d) corrected version of AIC



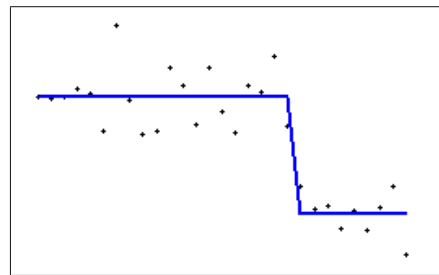
(e) corrected version of SIC



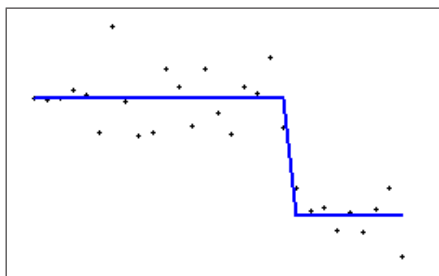
(f) longest run criterion



(g) multiresolution criterion

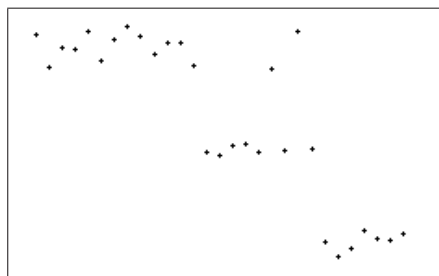


(h) iterative procedure stopped by multiresolution criterion

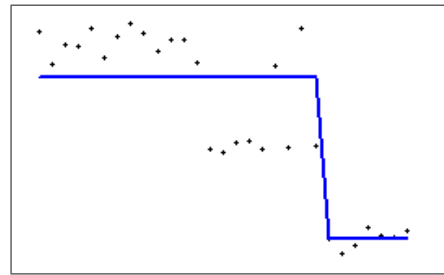


(i) last monotone criterion

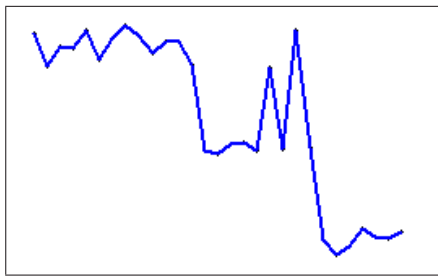
Figure 9.8: Application of different criteria to the fractionation curve of a fairly good spot where the right complementary cDNA bound



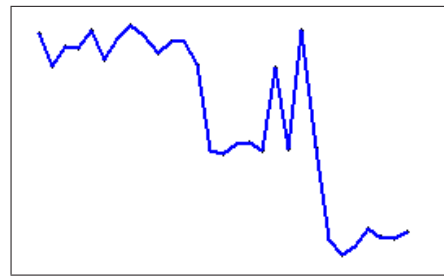
(a) data



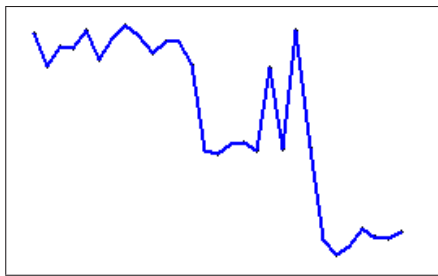
(b) longest interval



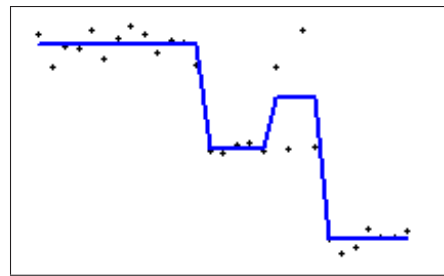
(c) log-longest interval



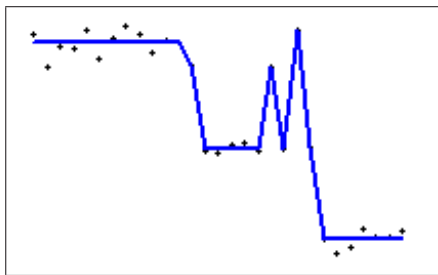
(d) corrected version of AIC



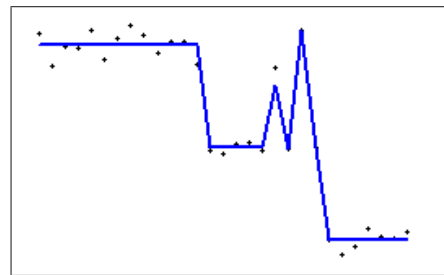
(e) corrected version of SIC



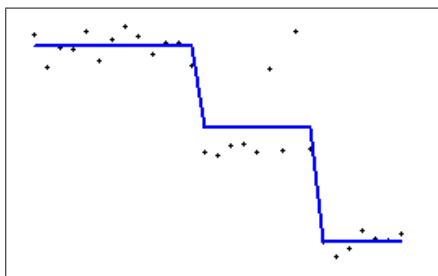
(f) longest run criterion



(g) multiresolution criterion



(h) iterative procedure stopped by multiresolution criterion



(i) last monotone criterion

Figure 9.9: Application of different criteria to the fractionation curve of a spot with cross-hybridization

log-longest interval criterion and the model selection criteria return data, see Figures 9.9(c), 9.9(d), and 9.9(e). The estimate resulting from the longest run criterion has jumps at suitable locations between the three plateaus. In addition, as we can see in Figure 9.9(f), it has to catch the ‘outliers’ of the center plateau since otherwise the run of the signs would be too long, see Figure 9.9(i). The estimate suggested by the multiresolution criterion, displayed in Figure 9.9(g), inserts even additional intervals to handle the outliers. We have the same situation for the estimate in Figure 9.9(h), arising from the iterative procedure. The reason is that it was stopped by the latter. The last monotone estimator yields exactly the desired estimate. It covers the three plateaus without additional peaks, see Figure 9.9(i).

In summary, the last monotone criterion outperforms all others with respect to the specific demand of the data.



# Chapter 10

## Simulations

A systematic study for the various criteria to choose the hyperparameter is desirable. Unfortunately this is presently beyond our time horizon. Nevertheless, we will apply the methods described in Part II to simulated data and then simply show the results. This will at least support our intuition gained from theoretical considerations. It also hopefully gives some hints which aspects should be pursued in future work.

We will consider the following estimators using the italic terms in the plots.

- *AICC*: the corrected version of the Akaike information criterion. It is a modification for short time series of the original criterion introduced in Section 7.1 . The exact formula can be found in Appendix B.2 ;
- *interval*: the longest interval estimator from Definition 5.3.1;
- *iterative*: the estimator provided by the repeated application of the longest interval criterion as described in Algorithm 8.1.1, and stopped by the multiresolution criterion;
- *monotone*: the last monotone estimator from Example 8.3.1;
- *multi*: the estimator from the multiresolution criterion described in Section 6.2 with factor  $\tau = 2$  ;
- *run*: the longest run estimator with maximal allowed run length of signs of residuals, given by (6.2).

Depending on the given signal, not all methods are reasonable, and we will exclude not well-suited estimators from our considerations.

We generate data  $y$  by regression models of the form

$$y_s = u_s + \xi_s, \quad s \in \{1, \dots, N\}, \quad (10.1)$$

where  $u$  is a given signal and  $\xi_1, \dots, \xi_N$  are independent and identically distributed Gaussian random variables with mean zero and variance  $\sigma^2 > 0$ . It is reasonable to start with signals  $u$  which are simple and smooth in the sense of the Potts functionals. In Section 10.1, we take a constant signal. In Section 10.2, we consider a signal with one jump down, corresponding to the fractionation curves from Chapter 9.2. In Section 10.3, the boxcar shaped signal is an analogon to the outer stimulus in fMRI human brain mapping from Section 9.1. The different signals  $u$  are displayed in Figure 10.1.

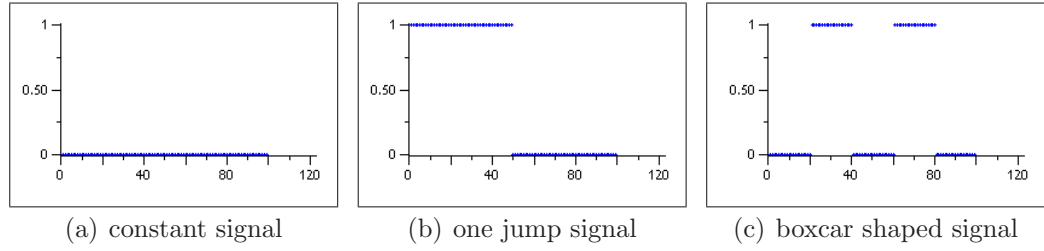


Figure 10.1: The underlying signals  $u$ .

We want to compare the different estimators in two respects, the number of jumps and the fidelity to the underlying signal  $u$ . The distance of the estimates  $x^*$  to the underlying signal  $u$  is measured by

- the sum of the absolute deviations

$$d_1(x^*, u) = \sum_{s=1}^N |x_s^* - u_s|,$$

- the sum of squares

$$d_2(x^*, u) = \sum_{s=1}^N (x_s^* - u_s)^2,$$

and

- the supremum of the absolute deviations

$$d_\infty(x^*, u) = \sup_{s=1, \dots, N} |x_s^* - u_s|.$$

In the model (10.1), we will fix the following parameters. The jumps in the underlying signals  $u$  are all of the same height  $h = 1$ . The length of data

is  $N = 100$ . We generate data for three different noise variances, taking  $\sigma^2 = 0.5h$ ,  $\sigma^2 = h$ , and  $\sigma^2 = 2h$ . We generate 200 time series' for each setting.

The number of jumps and the distances for the different methods will be displayed in box-and-whisker plots generated with `SPLUS 2000`. The white line in the light blue box marks the median of the data set. The box is bounded by the lower and the upper quartiles, the height of the box is the interquartile range (IQR). The whiskers represent the smallest and the largest value, respectively, or have a length of 1.5 IQR, if there are values which differ from the quartiles more than this amount. These values are considered as outliers and will be marked by black bars. If the dark blue regions of two boxes do not overlap then the medians are significantly different at a rough 5% level.

Our study is far away from a systematic simulation study. It is a very first attempt to investigate and compare the different methods for very specific signals  $u$ . Next steps would be to take other kinds of noise, especially from distributions with heavy tails, to vary the length of data, and to consider further types of signals.

## 10.1 Constant Signal

In this section, the underlying signal  $u$  is the constant null signal shown in Figure 10.1(a). We will exclude the longest interval criterion since it can never give a constant signal. Repeated application has the same disadvantage and thus, the iterative procedure is excluded as well.

We will first consider the number of jumps of the estimate. Figure 10.2(a) displays box-plots of this number for all estimates for noise variance  $\sigma^2 = 0.5$ . The number of jumps of the AICC estimates is considerably higher than for the others. We skip AICC in Figure 10.2(b) in order to get a more detailed picture for the last monotone and the stop criteria. For the last monotone criterion there is almost no variation in the number of jumps, it is either zero or one. The median for the multiresolution criterion is significantly smaller than the one for the longest run criterion. Both estimators yield some outliers with a fairly high number of jumps.

AICC seems to be not suited in this situation, it will not be discussed further. To check whether and how the number of jumps for the other methods depends on the noise variance, we plotted them for the three noise variances  $\sigma^2 = \text{var05} = 0.5$ ,  $\sigma^2 = \text{var1} = 1$ , and  $\sigma^2 = \text{var2} = 2$  in Figure 10.3. The number of jumps for the last monotone estimator (a) seems to change not all. We can also not say that the one for the multiresolution criterion (b)

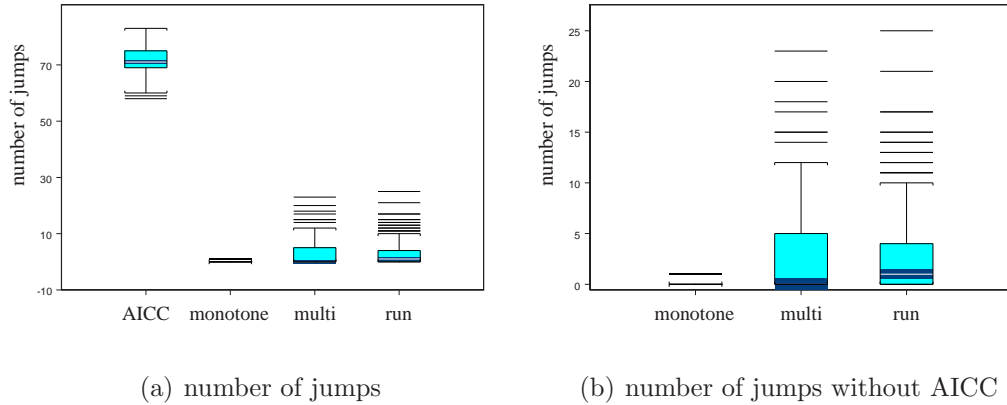


Figure 10.2: box-plots for the number of jumps of the estimates for the underlying constant signal with noise variance  $\sigma^2 = 0.5$

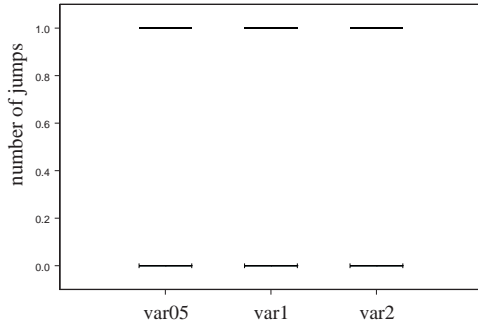
changes significantly. The behavior of the longest run criterion in (c) seems to be interesting: the number of jumps at variance 0.5 is significantly higher than the one at variance 2.

We will now consider the distance of the estimates from the true signal. Since the data term in the Potts functionals is a sum of squares, we will focus on the distance  $d_2$ , the sum of the squared deviations of the estimate from  $u$ . Again, as already indicated by the number of jumps, AICC is not comparable to the others, see Figure 10.4(a), and will be omitted in the following discussion. For noise variance  $\sigma^2 = 0.5$ , the distance of the last monotone estimates to the underlying constant is significantly smaller than the one for the others, see Figure 10.4(b).

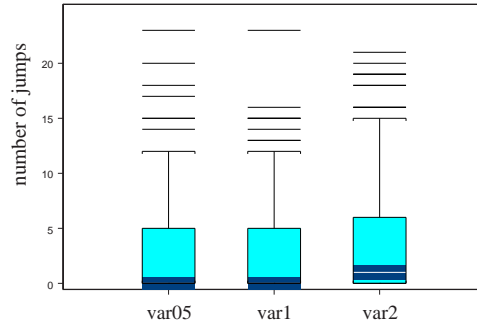
As expected, the median of the distance of the last monotone estimate increases significantly for increasing noise variance, see Figure 10.5(a). The box-plots (b) and (c) show a similar behavior for the two stopping criteria.

For sake of completeness, Figure 10.6 displays the box-plots of the sum and the supremum of the absolute deviations.

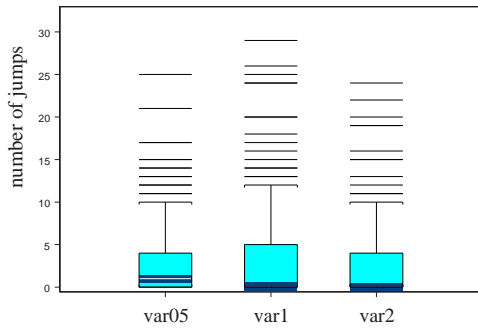
The qualitative statements on the number of jumps and the distance measures are the same for all noise variance. This is not astonishing for a constant underlying signal. A little bit surprising is the following observation. Although the stopping criteria were designed to decide whether data can be considered as white noise, the last monotone criterion discovers the underlying constant signal better in the simulations we performed here. In summary, for constant signals, the last monotone estimator outperforms the others.



(a) last monotone criterion

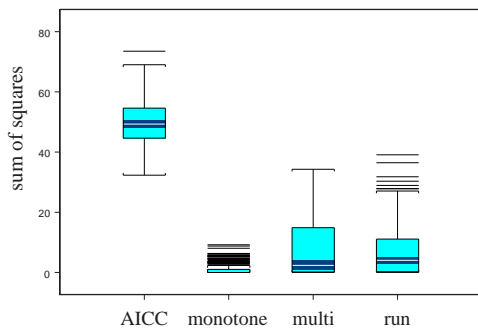


(b) multiresolution criterion

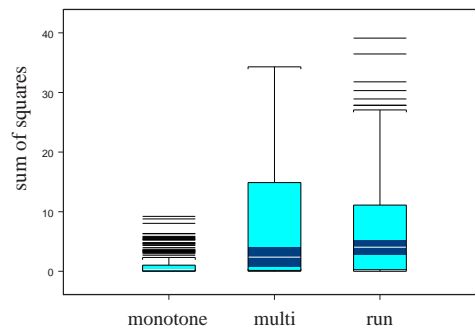


(c) longest run criterion

Figure 10.3: box-plots of the number of jumps for different estimators for an underlying constant at different noise variances  $\text{var05}=0.5$ ,  $\text{var1}=1$ , and  $\text{var2}=2$ .

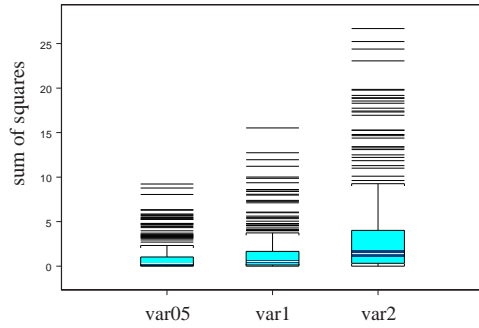


(a) sum of squares

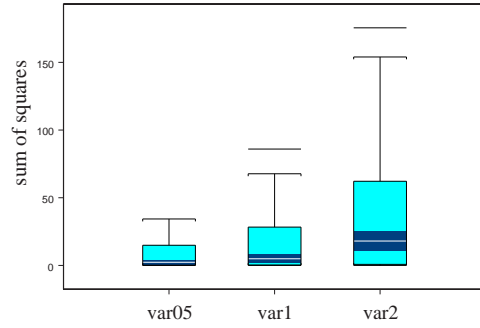


(b) sum of squares without AICC

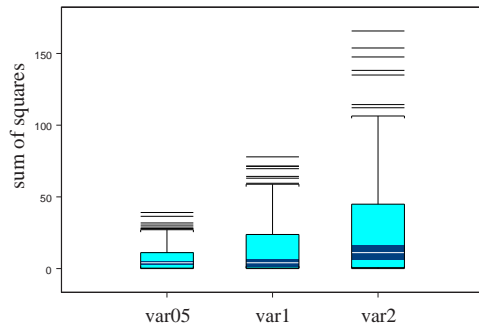
Figure 10.4: box-plots for the sum of squares for underlying constant signal with noise variance  $\sigma^2 = 0.5$



(a) last monotone criterion

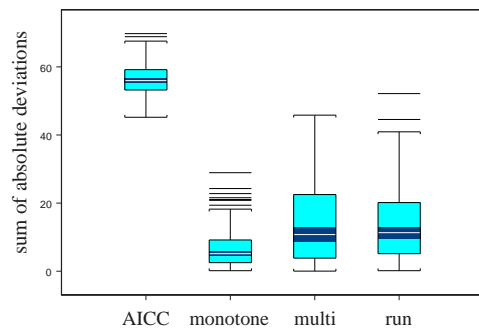


(b) multiresolution criterion

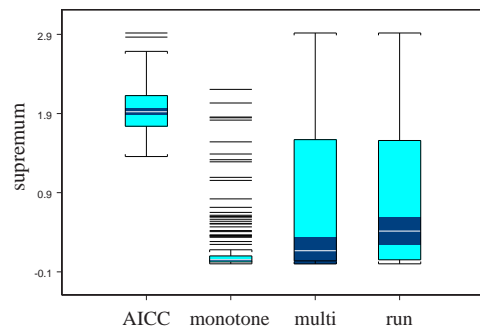


(c) longest run criterion

Figure 10.5: box-plots of the sum of squares for different estimators for an underlying constant for different noise variances  $\text{var05}=0.5$ ,  $\text{var1}=1$ , and  $\text{var2}=2$ .

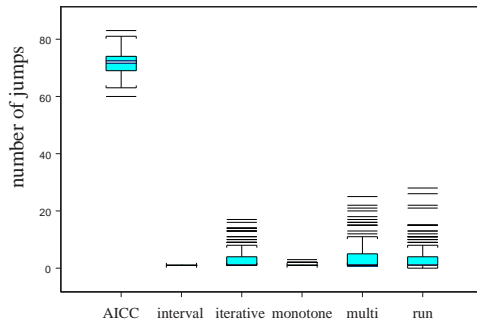


(a) sum of absolute deviations

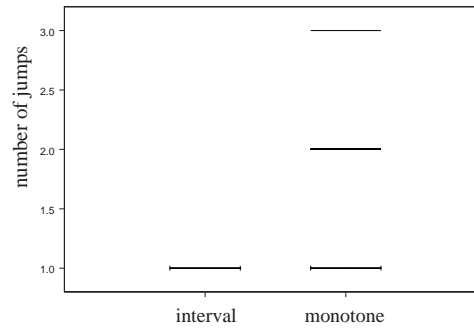


(b) supremum of absolute deviations

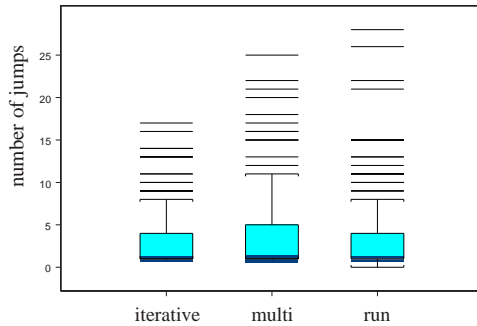
Figure 10.6: box-plots of different measures for the deviation of the estimates from the underlying constant signal with noise variance  $\sigma^2 = 0.5$



(a) number of jumps



(b) number of jumps for first group



(c) number of jumps for second group

Figure 10.7: box-plots for the number of jumps of the estimates for underlying one jump signal with noise variance  $\sigma^2 = 0.5$

## 10.2 One Jump Signal

In this section, we consider the mirrored heaviside function with jump height  $h = 1$  shown in Figure 10.1(b).

First of all, we ask whether the estimates have the ‘correct’ number of jumps equal to one. Figure 10.7(a) displays box-plots for the number of jumps of all estimates for noise variance  $\sigma^2 = 0.5$ . This number for the AICC estimates is always considerably higher than for the others. Omitting AICC, we find two ‘groups’ of estimators with respect to the number of jumps: the last monotone and the stop criteria on the one hand, and the longest interval and the last monotone criterion on the other hand. The longest interval criterion always detects one jump, for the last monotone criterion there are ‘outliers’ of two and three jumps, see Figure 10.7(b). As shown in Figure 10.7(c), the other methods result in a broader spectrum of number of jumps, but their medians are not significantly different from the one of the first group.

The qualitative statements about the number of jumps are the same for noise variance  $\sigma^2 = 1$  and  $\sigma^2 = 2$ , as we can see in Figure 10.8.

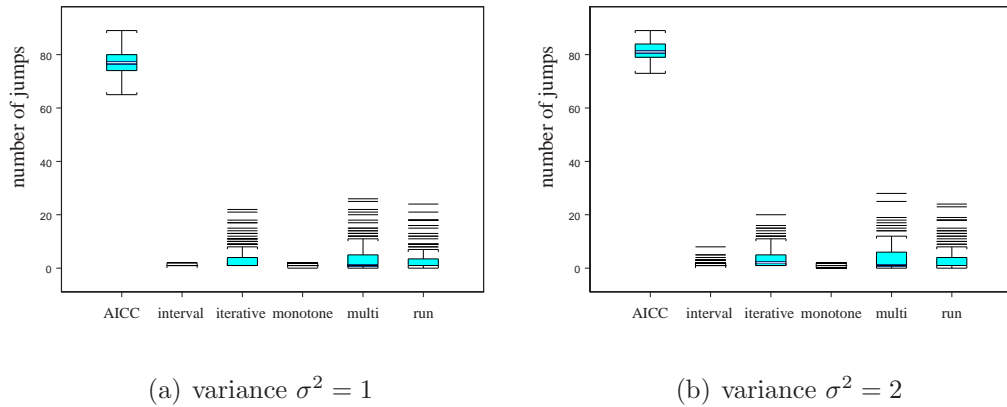


Figure 10.8: box-plots for the number of jumps of the estimates for underlying one jump signal at different noise variances

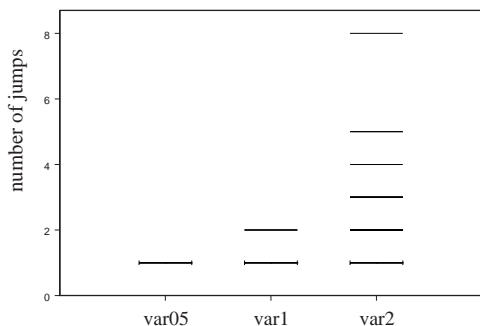
We will now consider the number of jumps separately for each estimator for varying noise. The median of the number of jumps for the iterative procedure stopped by the multiresolution criterion is significantly higher for  $\sigma^2 = 2$  than for the lower variances, see Figure 10.9(b). For all other methods the medians do not differ significantly for different noise variances.

We will now address the distance of the estimates from the true signal. For all distance measures, there is clearly the same partition of the estimators in groups as for the number of jumps, see Figure 10.10 for the sum of squares, and Figure 10.11 for the sum and the supremum of the absolute deviations. For all methods, the median of the sum of squares increases significantly for increasing noise, see Figure 10.12.

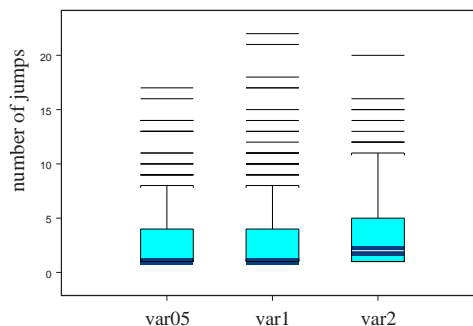
The simulations for the other noise variances for all the distances do not provide new insight. One would expect that if the noise variance is equal to or even twice the jump height of the underlying signal that then there is a significant decrease in the reliability of the estimators. This is not the case. The last monotone and the longest interval criterion find the correct number of jumps absolutely reliable, independent of the noise variance. As a consequence, the resulting estimates are significantly closer to the true underlying signal than the others. We conclude that for a one jump signal the last monotone criterion and the longest interval criterion perform best.

### 10.3 Boxcar Shaped Signal

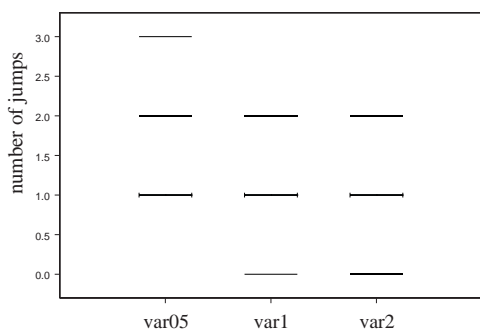
In case of the boxcar shaped signal displayed in Figure 10.1(c), the application of the last monotone criterion makes no sense. Concerning AICC, the



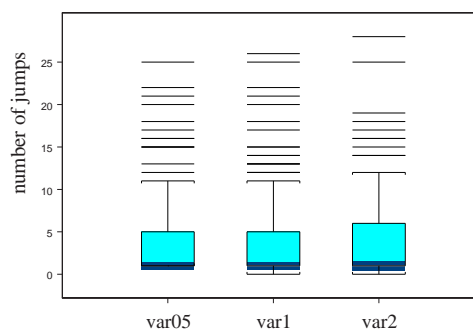
(a) longest interval criterion



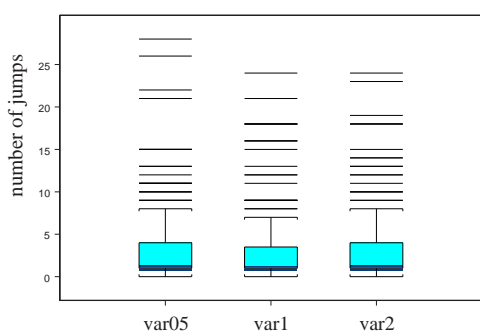
(b) iterative procedure



(c) last monotone criterion

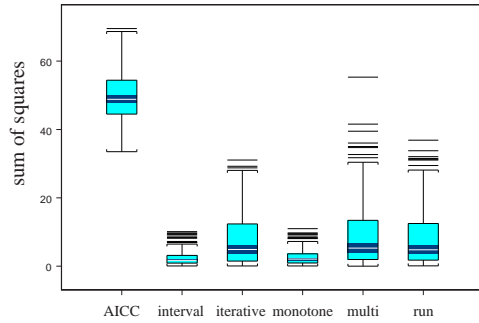


(d) multiresolution criterion

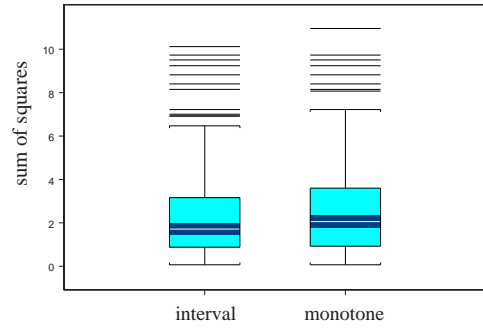


(e) longest run criterion

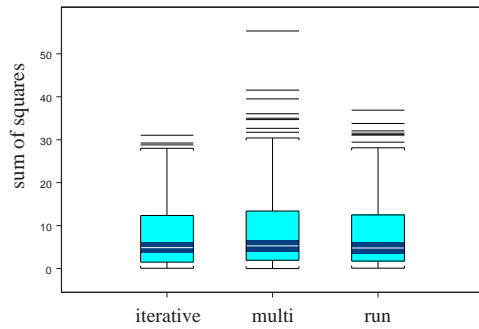
Figure 10.9: box-plots of the number of jumps for different estimators for an underlying one jump signal at different noise variances  $\text{var05} = 0.5$ ,  $\text{var1} = 1$ , and  $\text{var2} = 2$ .



(a) sum of squares

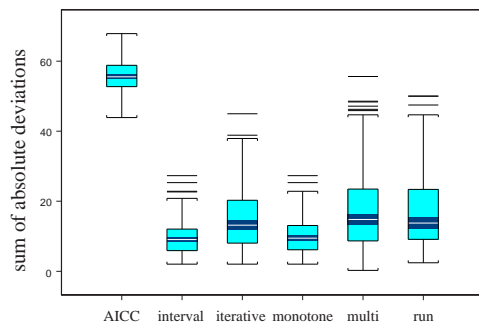


(b) first group

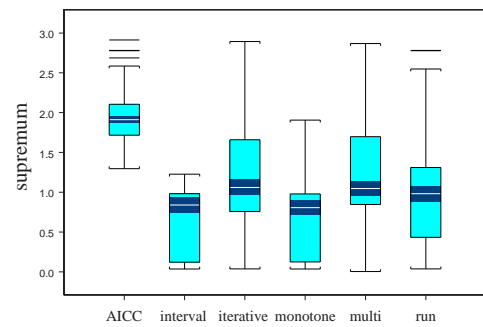


(c) second group

Figure 10.10: box-plots for the sum of squares for underlying one jump signal with noise variance  $\sigma^2 = 0.5$

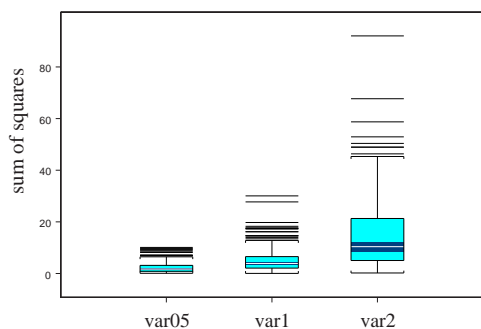


(a) sum of absolute deviations

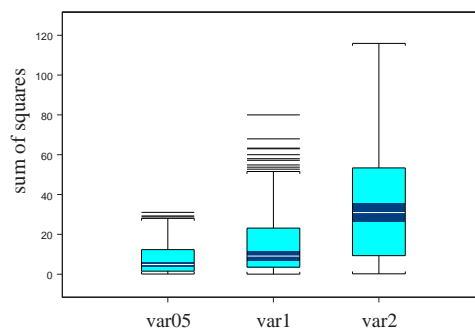


(b) supremum of absolute deviations

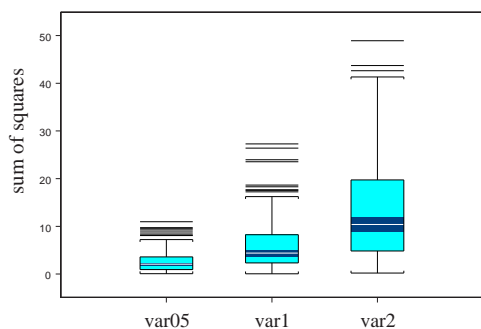
Figure 10.11: box-plot of the sum and supremum of absolute deviations of the estimates from the underlying one jump signal for  $\sigma^2 = 0.5$ .



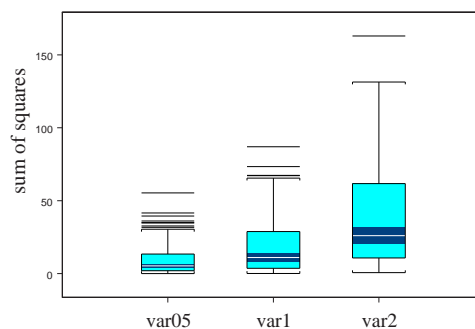
(a) longest interval criterion



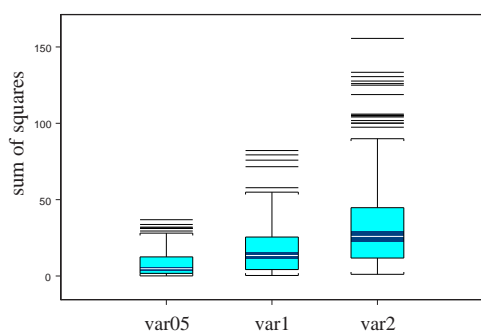
(b) iterative procedure



(c) last monotone criterion

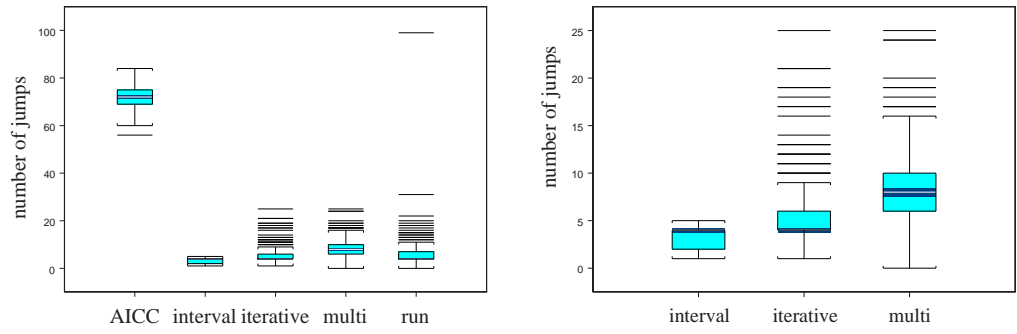


(d) multiresolution criterion



(e) longest run criterion

Figure 10.12: box-plots of the sum of squares for different estimators for an underlying one jump signal for different noise variances  $\text{var05} = 0.5$ ,  $\text{var1} = 1$ , and  $\text{var2} = 2$ .



(a) number of jumps, all methods

(b) number of jumps without AICC and longest run

Figure 10.13: box-plot of the number of jumps of the estimate for underlying boxcar shaped signal at noise variance  $\sigma^2 = 0.5$ .

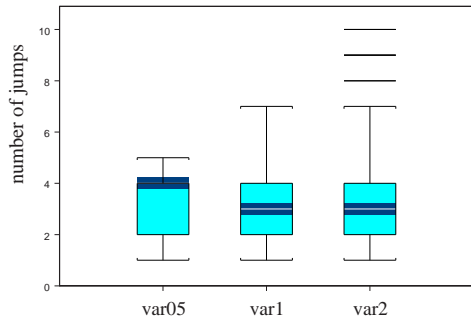
situation is the same as in the preceding sections, it is in any category out of the range of the other criteria.

We will first consider the number of jumps of the estimates where the correct one is four. In the present case, only the longest run criterion has an outlier in the number of jumps, see Figure 10.13(a). Comparing the other in Figure 10.13(b), we see that the median of the multiresolution criterion is significantly higher than the medians of the others.

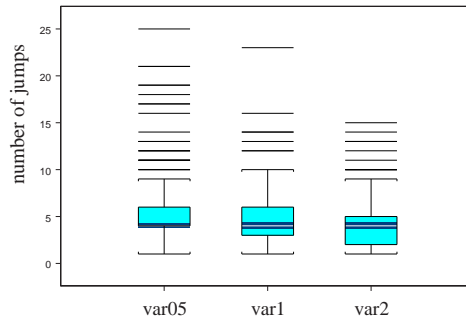
The median of the number of jumps for the longest run criterion is only for noise variance  $\sigma^2 = 0.5$  equal to the correct number of jumps. For higher noise variances, the number of jumps has the tendency to be too small, see Figure 10.14(a). This leads to the assertion that the longest interval underestimates the number of jumps with increasing noise. The median for the iterative procedure does not differ significantly for different noise variances, see Figure 10.14(b). We observe a slighter tendency to underestimate the number of jumps for higher variance in the noise. As shown in Figure 10.14(d), the median for the longest run criterion for noise variance  $\sigma^2 = 2$  is significantly smaller than for  $\sigma^2 = 1$ . The median of the number of jumps for the multiresolution criterion becomes significantly smaller the larger the variance is. At the highest variance in our comparisons, the correct number of jumps is detected more often than for lower variances, see Figure 10.14(e).

For the distance measures, we will restrict ourselves to the discussion of the sum of squares. For it, we have a box-plot (see Figure 10.15) for noise variance  $\sigma^2 = 0.5$  similar to the one for the number of jumps. The median for the multiresolution criterion is significantly higher than the others.

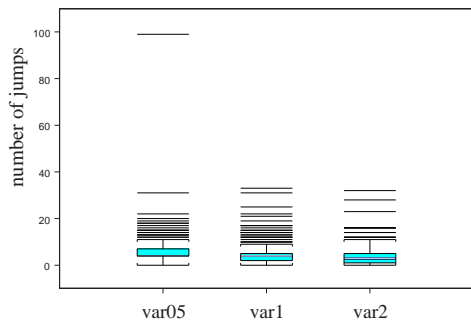
All criteria show a significant increase of the median of the sum of squares



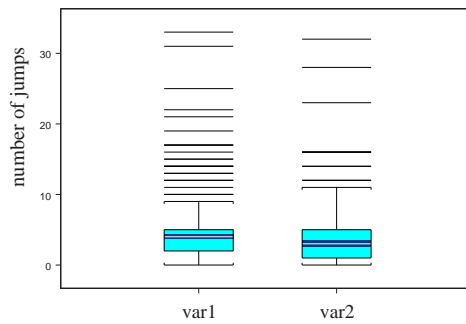
(a) longest interval criterion



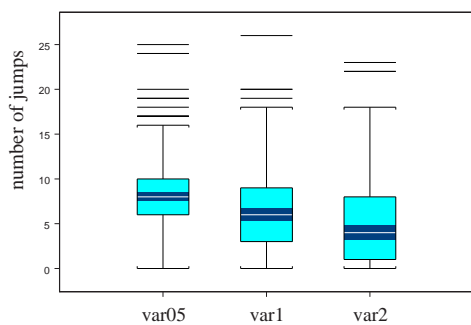
(b) iterative procedure



(c) longest run criterion



(d) longest run criterion for higher variances



(e) multiresolution criterion

Figure 10.14: box-plots of the number of jumps for different estimators for an underlying boxcar shaped signal at different noise variances  $\text{var05} = 0.5$ ,  $\text{var1} = 1$ , and  $\text{var2} = 2$ .

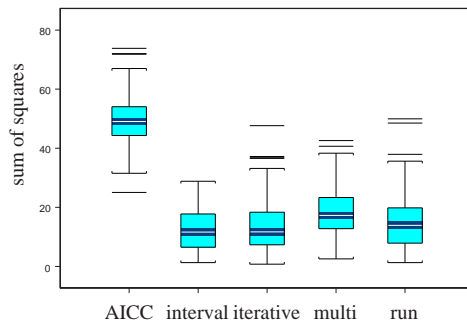


Figure 10.15: box-plot of the sum of squares of the estimate for the underlying boxcar shaped signal at noise variance  $\sigma^2 = 0.5$ .

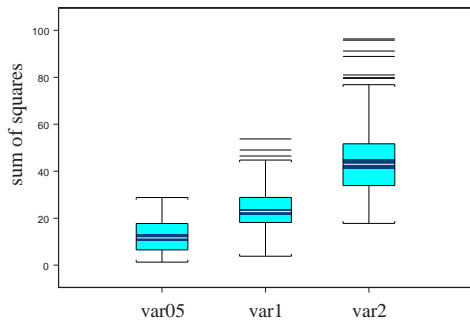
(a) sum of squares

for increasing noise variance, see Figure 10.16.

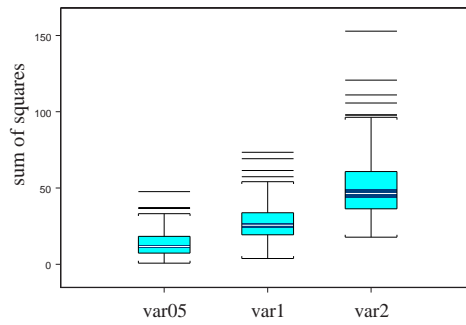
We obtain similar results for the sum and the supremum of the absolute deviations, see Figure 10.17.

The comparison of the different methods for noise variance  $\sigma^2 = 1$  does not provide further insight since the results are qualitatively the same. Only for  $\sigma^2 = 2$ , the number of jumps and the distance to the underlying signal for the multiresolution criterion differs no more from the other criteria.

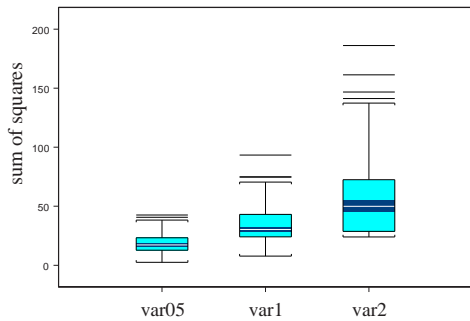
The longest interval criterion is the one with the smallest variation in the number of jumps. For  $\sigma^2 = 0.5$ , the estimates have in most cases the correct number of jumps. For higher variances, it has a tendency to insert rather three than four jumps. Here, the other methods perform better, but have also a larger range of number of jumps including outliers. The longest interval criterion outperforms the other methods with respect to the number of jumps, especially if one is interested in a parsimonious representation of data. In addition, the distance  $d_2$  of the estimates is not larger than for the others.



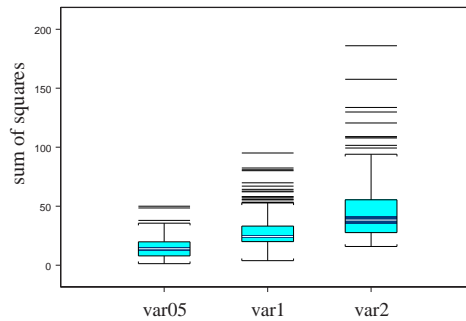
(a) longest interval criterion



(b) iterative procedure

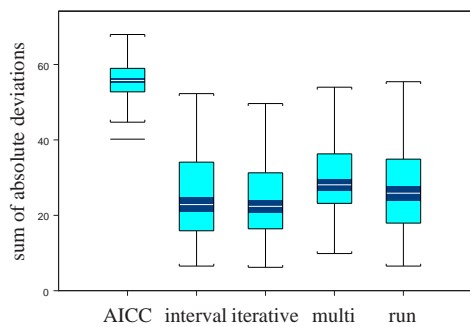


(c) multiresolution criterion

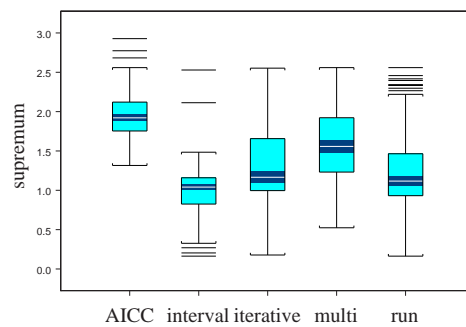


(d) longest run criterion

Figure 10.16: box-plot of the sum of squares of the different estimates for the underlying boxcar shaped signal at the noise variances  $\text{var05} = 0.5$ ,  $\text{var1} = 1$ , and  $\text{var02} = 2$ .



(a) sum of absolute deviations



(b) supremum of absolute deviations

Figure 10.17: box-plot of the sum and the supremum of the absolute deviations of the estimate for underlying boxcar shaped jump signal at noise variance  $\sigma^2 = 0.5$ .



**Part IV**  
**Consistency**



In this part, we study the asymptotic behavior of minimizers  $x^*(\gamma, Y^n)$  of the Potts functionals for data generated by regression models of the form

$$Y_s^n = u_s^{k_n} + \xi_{n,s}, \quad s \in S_{k_n} = \{1, \dots, k_n\}.$$

More precisely,

$$\xi_{n,s} : (\Omega, \mathcal{F}, \mathbb{P}) \longrightarrow \mathbb{R}, \quad \omega \longmapsto \xi_{n,s}(\omega), \quad n \in \mathbb{N}, s \in S_{k_n} = \{1, \dots, k_n\}$$

are random variables and  $u^{k_n} \in \mathbb{R}^{S_{k_n}}$  is deterministic for two different scenarios for  $\xi$ ,  $u$ , and  $(k_n)_{n \in \mathbb{N}}$ .

We interpret minimizers of the Potts functionals as estimators of a true signal or a smoothed version, depending whether  $\gamma$  is drawn to zero or fixed. This amounts to consistency of MAP estimators.

In Chapter 11, we assume  $k_n = N$ . Data  $Y^n$  are generated by adding a random vector  $(\xi_{n,s})_s$  to the true deterministic signal  $u$  of fixed length  $N$ . We show the convergence of estimators  $x^*(\gamma_n, Y^n)$  towards  $x^*(\gamma^*, u)$  if  $\xi_{n,s} \rightarrow 0$  for all  $s$  and  $\gamma_n \rightarrow \gamma^*$ .

In contrast to the fixed data length there, Chapter 12 deals the case of increasing data length  $k_n = n$ . The deterministic signal  $u^n$  is the mean value of a square integrable function  $f$  over an appropriate interval. Data  $Y^n$  are then generated by adding independent and identically distributed centered Gaussian random variables  $(\xi_{n,s})_s$  to  $u^n$ . Using the concept of epi-convergence, we show that an embedded sequence of minimizers  $x^*(n\gamma_n, Y^n)$  of the Potts functionals with scaled hyperparameter converges to  $f$  if the sequence  $(\gamma_n)_n$  converges in an appropriate way.



# Chapter 11

## Fixed Data Length

In this chapter, we fix the dimension  $k_n = N$  of data, and hence  $S_{k_n} = \{1, \dots, N\}$ . Let a signal  $u \in \mathbb{R}^N$  be given. We consider random variables as data given by

$$Y_s^n = u_s + \xi_{n,s}, \quad s \in S = \{1, \dots, N\}, \quad (11.1)$$

where  $\xi_{n,s}$ ,  $s = 1, \dots, N$ , are real random variables on some probability space  $(\Omega, \mathcal{F}, \mathbb{P})$ .

We consider the Potts functionals

$$\bar{H}_\gamma(\cdot, y) : \mathbb{X} \longrightarrow \mathbb{R}, \quad x \longmapsto \gamma \cdot |J(x)| + \sum_{s \in S} (y_s - x_s)^2. \quad (11.2)$$

Due to the existence of a measurable section of the set-valued map  $(\gamma, y) \mapsto X^*(\gamma, y)$  statements of the form ‘ $x^*(\gamma, Y^n)$  converges  $\mathbb{P}$ -almost surely’ make sense.

Let  $\gamma_n$ ,  $n \in \mathbb{N}$ , be random variables, also defined on the probability space  $(\Omega, \mathcal{F}, \mathbb{P})$ . We will prove stochastic continuity of MAP estimators if the sequence  $(\xi_{n,\cdot})_{n \in \mathbb{N}}$  of random variables tends to zero and  $(\gamma_n)$  tends to  $\gamma'$ .

**Theorem 11.0.1** *Let  $(\gamma, y) \mapsto x^*(\gamma, y)$  be a measurable section of MAP estimators  $(\gamma, y) \mapsto X^*(\gamma, y)$  for (11.2). Let  $u \in \mathbb{R}^N$  and  $\gamma' \geq 0$  be given. Suppose that data  $Y^n$  are given by (11.1) and*

$$\xi_{n,s} \longrightarrow 0 \quad \mathbb{P}\text{-almost surely} \quad \text{as } n \rightarrow \infty \quad \text{for all } s \in S.$$

Let further  $(\gamma_n)_{n \in \mathbb{N}} \subset (0, \infty)$  be a random sequence with

$$\gamma_n \longrightarrow \gamma' \quad \mathbb{P}\text{-almost surely} \quad \text{as } n \rightarrow \infty.$$

Then

$$x^*(\gamma_n, Y^n) \longrightarrow x^*(\gamma', u) \quad \mathbb{P}\text{-almost surely} \quad \text{as } n \rightarrow \infty.$$

**Proof** Let  $\omega \in \Omega$  be such that  $\gamma_n(\omega) \rightarrow \gamma'$  and  $\xi_{n,s}(\omega) \rightarrow 0$  for  $n \rightarrow \infty$  and all  $s \in S$ . The latter implies that  $Y^n(\omega) \rightarrow u$  for  $n \rightarrow \infty$ . A measurable section  $(\gamma, y) \mapsto x^*(\gamma, y)$  provides a unique  $x^*(\gamma', u)$ . By Theorem 2.5.2, there is a neighborhood of  $(\gamma', u)$  such that  $(\gamma, y) \mapsto x^*(\gamma, y)$  is continuous in this neighborhood. We have that  $x^*(\gamma_n(\omega), Y^n(\omega)) \rightarrow x^*(\gamma', u)$  for  $n \rightarrow \infty$ . Thus,

$$\begin{aligned} & \{\omega \in \Omega : \gamma_n(\omega) \rightarrow \gamma'\} \cap \{\omega \in \Omega : \xi_{n,s}(\omega) \rightarrow 0, s = 1, \dots, N\} \\ & \subset \{\omega \in \Omega : x^*(\gamma_n(\omega), Y^n(\omega)) \rightarrow x^*(\gamma', u)\} \end{aligned}$$

which proves the assertion.  $\square$

The case  $\gamma' = 0$  in the preceding theorems yields consistency to recover the true signal  $u$ .

**Corollary 11.0.2** *Let  $(\gamma, y) \mapsto x^*(\gamma, y)$  be a measurable section of the MAP estimators for (11.2). Suppose that data  $Y^n$  are given by (11.1) and*

$$\xi_{n,s} \rightarrow 0 \quad \mathbb{P}\text{-almost surely} \quad \text{as } n \rightarrow \infty \quad \text{for all } s \in S.$$

*Let further  $(\gamma_n)_{n \in \mathbb{N}} \subset (0, \infty)$  be a random sequence with*

$$\gamma_n \rightarrow 0 \quad \mathbb{P}\text{-almost surely} \quad \text{as } n \rightarrow \infty.$$

*Then*

$$x^*(\gamma_n, Y^n) \rightarrow u \quad \mathbb{P}\text{-almost surely} \quad \text{as } n \rightarrow \infty.$$

**Proof** For  $\gamma = 0$  the unique minimizer of  $\bar{H}_\gamma(\cdot, y)$  is  $y$  and therefore we have  $x^*(0, u) = u$  in Theorem 11.0.1.  $\square$

# Chapter 12

## Increasing Data Length

In this chapter, the set of sites on which signals and data are defined increases in  $n$ . For sake of simplicity, let  $k_n = n$ , and thus the set of sites is

$$S_n = \{1, \dots, n\}.$$

Signals  $x$  and data  $y$  are elements of  $\mathbb{R}^{S_n}$ .

Let  $\lambda$  denote the Lebesgue measure on the Borel- $\sigma$ -field  $\mathcal{B}([0, 1])$ . In the following,  $L^2([0, 1])$  will denote the Hilbert space  $L^2([0, 1], \mathcal{B}([0, 1]), \lambda)$  of equivalence classes of square-integrable functions which coincide almost everywhere. It is equipped with the inner product

$$\langle f, g \rangle = \int_{[0,1]} f \cdot g \, d\lambda,$$

and the  $L^2$ -norm is given by

$$\|f\| = \left( \int_{[0,1]} |f|^2 \, d\lambda \right)^{1/2}.$$

In the sequel, we adopt the following point of view: We think of data as arising from a discretization of a ‘true’ signal  $f \in L^2([0, 1])$  corrupted by noise. We consider the following *discretization* of  $f$  given by

$$\Delta_n : L^2([0, 1]) \longrightarrow \mathbb{R}^{S_n}, \quad f \longmapsto \bar{f}^n$$

where  $\bar{f}^n = (\bar{f}_s^n)_{s \in S_n}$  is defined as

$$\bar{f}_s^n = n \int_{\frac{s-1}{n}}^{\frac{s}{n}} f \, d\lambda, \quad s \in S_n = \{1, \dots, n\}. \quad (12.1)$$

Random data are given by

$$Y_s^n = \bar{f}_s^n + \xi_{n,s}, \quad s \in S_n = \{1, \dots, n\} \quad (12.2)$$

where  $\xi_{n,s}$ ,  $s \in S_n$ , are random variables on some probability space  $(\Omega, \mathcal{F}, \mathbb{P})$ . We require the following condition to hold:

**Hypothesis 12.0.3** *We assume that for all  $n \in \mathbb{N}$  the triangular array of random variables  $\xi_{n,s}$ ,  $s \in S_n$ , are independent and identically distributed Gaussian random variables with mean zero and variance  $\sigma^2 > 0$ . We will use the short-hand notation  $\xi_{n,s} \sim \mathcal{N}(0, \sigma^2)$ .*

The results of this section are still true if the random variables  $(\xi_{n,s})_{n \in \mathbb{N}, s \in S_n}$  are uniformly sub-Gaussian random variables. We refer to the forthcoming article V. LIEBSCHER et al. (2004).

Increasing  $n$  thus corresponds to refining the discretization. Consistency means that the finer the discretization the higher is the accuracy of the estimators.

We will consider minimizers of the Potts functionals

$$\bar{H}_\gamma^n(\cdot, y) : \mathbb{R}^{S_n} \longrightarrow \mathbb{R}, \quad x \longmapsto \gamma \cdot |J(x)| + \sum_{s \in S_n} (y_s - x_s)^2 \quad (12.3)$$

for  $y \in \mathbb{R}^{S_n}$  as estimators.

To get consistency, we have to compare estimators of different discretization levels. To this end, we will introduce counterparts of the Potts functionals defined on the common Hilbert space  $L^2([0, 1])$ . Discrete data  $y^n$  are identified with the (blurred) conditional expectations of the true signal with respect to the equidistant partition of  $[0, 1]$  into  $n$  intervals. The corresponding ‘continuous’ Potts functionals are defined such that their minimizers are step functions which correspond to minimizers of (12.3). It turns out that under a suitable rescaling of the hyperparameter, the minimizing step functions converge to the function  $f$ . This is the main result formulated in Theorem 12.0.5.

First of all, we have to make precise which step functions correspond to signals and data in  $\mathbb{R}^{S_n}$ . It is not clear which value the step function should take at a point of discontinuity. The determination is arbitrary, and without restriction we decide to take a right-continuous step function. The space of all right-continuous step functions on  $[0, 1]$  given by  $\text{span}\{\mathbf{1}_{[a,b)} : 0 \leq a < b \leq 1\}$  will be denoted by  $\mathcal{T}([0, 1])$ .

**Definition 12.0.4** Signals  $x \in \mathbb{R}^{S_n}$  correspond to right-continuous step functions via the *embedding*

$$j^n : \mathbb{R}^{S_n} \longrightarrow \mathcal{T}([0, 1]), \quad x \longmapsto \mathfrak{x}^n = \sum_{s=1}^n x_s \cdot \mathbf{1}_{[\frac{s-1}{n}, \frac{s}{n})}. \quad (12.4)$$

**Convention** In the sequel, we will frequently identify a right-continuous step function with its equivalence class in  $L^2([0, 1])$ . This is not dangerous since for right-continuous step functions  $\tau \in \mathcal{T}([0, 1])$  the map  $\tau \mapsto [\tau]$  is one-to-one. In particular, we use the right-continuous representant for the definition of the jump set, see Definition 12.1.1 below.

The main result of this chapter is the following theorem.

**Theorem 12.0.5** Let be  $f \in L^2([0, 1])$  and Hypothesis 12.0.3 be fulfilled. Assume that  $Y^n$  is determined by the model (12.2). Then, for any measurable section  $(\gamma, y) \mapsto x^*(\gamma, y)$  of minimizers of the Potts functional (12.3), and for any random sequence  $(\gamma_n)_{n \in \mathbb{N}} \subset (0, \infty)$  with

$$\gamma_n \longrightarrow 0 \quad \text{and} \quad \gamma_n \frac{n}{\log n} \longrightarrow \infty \quad \mathbb{P}\text{-almost surely} \quad \text{as} \quad n \rightarrow \infty,$$

we have

$$j^n(x^*(n\gamma_n, Y^n)) \longrightarrow f \quad \text{in} \quad L^2([0, 1]) \quad \text{as} \quad n \rightarrow \infty.$$

**Proof** The proof will be given in Section 12.3. □

## 12.1 Potts Functionals on $L^2([0, 1])$

We will now construct counterparts to the Potts functionals from (12.3) defined on  $L^2([0, 1])$ , and show the relation of their minimizers to MAP estimators.

**Definition 12.1.1** The jump set  $\mathcal{J}(\tau)$  for (an  $L^2$ -equivalence class of) a right-continuous step function  $\tau \in \mathcal{T}([0, 1])$  will be defined as the set of discontinuities of its right-continuous representant.

For functions  $f \in L^2([0, 1]) \setminus \mathcal{T}([0, 1])$  we set  $|\mathcal{J}(f)| = \infty$ .

The image of  $\mathbb{R}^{S_n}$  under  $j^n$ , the step functions identified with their  $L^2$ -equivalence classes, will be denoted by  $\mathfrak{X}^n$  and the  $\sigma$ -field generated by the system  $\{[\frac{s-1}{n}, \frac{s}{n}) : s \in S_n\}$  of intervals by  $\mathfrak{B}^n$ .

We can now define the following functionals on  $L^2([0, 1])$ .

**Definition 12.1.2** For  $f \in L^2([0, 1])$  the functionals

$$\begin{aligned} \tilde{H}_\gamma(\cdot, f) &: L^2([0, 1]) \longrightarrow \bar{\mathbb{R}}, \\ g &\longmapsto \begin{cases} \gamma \cdot |\mathcal{J}(g)| + \|f - g\|^2 & \text{if } g \in \mathfrak{X}^n, \\ +\infty & \text{otherwise.} \end{cases} \end{aligned} \quad (12.5)$$

will be called **continuous Potts functionals**.

The term ‘continuous’ should indicate that it is, in contrast to the Potts functionals from (12.3) which are defined for intensities on the discrete set  $S_n$ , defined for functions on the real interval  $[0, 1)$ .

The space  $\mathfrak{X}^n$  is a closed linear subspace of  $L^2([0, 1])$ , and hence, we have the *orthogonal projection of  $L^2([0, 1])$  onto  $\mathfrak{X}^n$*  given by

$$P_{\mathfrak{X}^n} : L^2([0, 1]) \longrightarrow \mathfrak{X}^n, \quad f \longmapsto \sum_{s \in S_n} \bar{f}_s^n \cdot \mathbf{1}_{[\frac{s-1}{n}, \frac{s}{n})} \quad (12.6)$$

where  $\bar{f}^n$  is given by (12.1).

Note that  $\bar{f}^n = (\bar{f}_s^n)_{s \in S_n}$  is an element of  $\mathbb{R}^{S_n}$  and  $P_{\mathfrak{X}^n} f = j^n(\bar{f}^n)$ .

We will rewrite the data term  $\|f - g\|^2$  of the continuous Potts functional.

**Lemma 12.1.3** Let be  $f \in L^2([0, 1])$ . Then the following holds.

(1) For  $\mathfrak{r}^n = j^n(x)$ ,  $x \in \mathbb{R}^{S_n}$ , we have

$$\|f - \mathfrak{r}^n\|^2 = \frac{1}{n} \sum_{s \in S_n} (\bar{f}_s^n - x_s)^2 + \|(\text{id} - P_{\mathfrak{X}^n})f\|^2.$$

(2) For  $\mathfrak{h}^n = j^n(y)$  and  $\mathfrak{r}^n = j^n(x)$ ,  $x, y \in \mathbb{R}^{S_n}$ , we can write

$$\|\mathfrak{h}^n - \mathfrak{r}^n\|^2 = \frac{1}{n} \sum_{s \in S_n} (y_s - x_s)^2.$$

**Proof** (1) For  $\mathfrak{r}^n = j^n(x)$ ,  $x \in \mathbb{R}^{S_n}$ , we can write

$$\begin{aligned} \|f - \mathfrak{r}^n\|^2 &= \|(\text{id} - P_{\mathfrak{X}^n})f\|^2 + \|P_{\mathfrak{X}^n} f - \mathfrak{r}^n\|^2 \\ &= \|(\text{id} - P_{\mathfrak{X}^n})f\|^2 + \int_0^1 (P_{\mathfrak{X}^n} f - \mathfrak{r}^n)^2 d\lambda \\ &= \|(\text{id} - P_{\mathfrak{X}^n})f\|^2 + \int_0^1 \left( \sum_{s \in S_n} \bar{f}_s^n \mathbf{1}_{[\frac{s-1}{n}, \frac{s}{n})} - \sum_{s \in S_n} x_s \mathbf{1}_{[\frac{s-1}{n}, \frac{s}{n})} \right)^2 d\lambda \end{aligned}$$

$$\begin{aligned}
&= \|(\text{id} - P_{\mathfrak{X}^n})f\|^2 + \int_0^1 \sum_{s \in S_n} (\bar{f}_s^n - x_s)^2 \mathbf{1}_{[\frac{s-1}{n}, \frac{s}{n})} d\lambda \\
&= \|(\text{id} - P_{\mathfrak{X}^n})f\|^2 + \sum_{s \in S_n} (\bar{f}_s^n - x_s)^2 \lambda([\frac{s-1}{n}, \frac{s}{n})) \\
&= \|(\text{id} - P_{\mathfrak{X}^n})f\|^2 + \frac{1}{n} \sum_{s \in S_n} (\bar{f}_s^n - x_s)^2
\end{aligned}$$

which is the assertion.

(2) The assertion follows directly from (1) since  $j^n(y) = P_{\mathfrak{X}^n}j^n(y)$ .  $\square$

The following lemma shows how MAP estimators of the Potts functionals are connected to minimizers of their counterparts embedded into  $L^2([0, 1])$ .

**Lemma 12.1.4** *Let  $\bar{H}_\gamma^n$  be the Potts functionals from (12.3) and  $\tilde{H}_\gamma^n$  the continuous ones given by (12.5). Then*

(1) *For  $x \in \mathbb{R}^{S_n}$ ,  $\mathfrak{x}^n = j^n(x)$ , and  $f \in L^2([0, 1])$  the following statements are equivalent*

- (a)  $\mathfrak{x}^n$  minimizes  $\tilde{H}_\gamma^n(\cdot, f)$ .
- (b)  $\mathfrak{x}^n$  minimizes  $\tilde{H}_\gamma^n(\cdot, P_{\mathfrak{X}^n}f)$ .
- (c)  $x$  minimizes  $\bar{H}_{n\gamma}^n(\cdot, \bar{f}^n)$ .

(2) *For  $x, y \in \mathbb{R}^{S_n}$ ,  $\mathfrak{x}^n = j^n(x)$ , and  $\mathfrak{y}^n = j^n(y)$ , the following statements are equivalent*

- (a)  $x$  minimizes  $\bar{H}_{n\gamma}^n(\cdot, y)$ .
- (b)  $\mathfrak{x}^n$  minimizes  $\tilde{H}_\gamma^n(\cdot, \mathfrak{y}^n)$ .

**Proof** Note that  $j^n : \mathbb{R}^{S_n} \rightarrow \mathfrak{X}^n \subset L^2([0, 1])$  is one-to-one and onto. Hence, we can identify  $j^n(x) \in \mathfrak{X}^n$  with  $x \in \mathbb{R}^{S_n}$ .

(1) By Lemma 12.1.3 we get the identity

$$\tilde{H}_\gamma^n(\mathfrak{x}^n, f) = \tilde{H}_\gamma^n(\mathfrak{x}^n, P_{\mathfrak{X}^n}f) + \|(\text{id} - P_{\mathfrak{X}^n})f\|^2.$$

Since the functionals  $\tilde{H}_\gamma^n(\cdot, f)$  and  $\tilde{H}_\gamma^n(\cdot, P_{\mathfrak{X}^n}f)$  differ only by the constant  $\|(\text{id} - P_{\mathfrak{X}^n})f\|^2$  they have the same minimizers. This is the equivalence of (a) and (b).

Again from Lemma 12.1.3 we derive the following relation

$$\tilde{H}_\gamma^n(\mathfrak{x}^n, f) = \gamma \cdot |J(x)| + \frac{1}{n} \sum_{s \in S_n} (\bar{f}_s^n - x_s)^2 + \|(\text{id} - P_{\mathfrak{X}^n})f\|^2$$

$$\begin{aligned}
&= \frac{1}{n} (n\gamma \cdot |J(x)| + \sum_{s \in S_n} (\bar{f}_s^n - x_s)^2) + \|(\text{id} - P_{\mathfrak{X}^n})f\|^2 \\
&= \frac{1}{n} \bar{H}_{n\gamma}^n(x, \bar{f}^n) + \|(\text{id} - P_{\mathfrak{X}^n})f\|^2
\end{aligned}$$

for  $x = (j^n)^{-1}(\mathfrak{x}^n)$ . The last term on the right hand side is constant and does not have influence on minimizers of  $\tilde{H}_\gamma^n(\cdot, f)$ . Note that a minimizer of  $\tilde{H}_\gamma^n(\cdot, f)$  is necessarily in  $\mathfrak{X}^n$ . Hence,  $x \in \mathbb{R}^{S_n}$  minimizes  $\bar{H}_{n\gamma}^n(\cdot, \bar{f}^n)$  if and only if  $j^n(x)$  minimizes  $\tilde{H}_\gamma^n(\cdot, f)$  which is the equivalence of (b) and (c).

(2) Using the reformulation of the data term from Lemma 12.1.3 and the coincidence of the number of jumps for  $x$  and  $\mathfrak{x}^n$  we get

$$\tilde{H}_\gamma^n(\mathfrak{x}^n, \mathfrak{y}^n) = \gamma \cdot |J(x)| + \frac{1}{n} \sum_{s \in S_n} (y_s - x_s)^2 = \frac{1}{n} \bar{H}_{n\gamma}^n(x, y)$$

for  $\mathfrak{x}^n = j^n(x)$  and  $\mathfrak{y}^n = j^n(y)$ . Hence,  $x \in \mathbb{R}^{S_n}$  minimizes  $\bar{H}_{n\gamma}^n(\cdot, y)$  if and only if  $\mathfrak{x}^n$  minimizes  $\tilde{H}_\gamma^n(\cdot, \mathfrak{y}^n)$ .  $\square$

## 12.2 Epi-Convergence and Relative Compactness

For the proof of Theorem 12.0.5 we need several preparatory steps. Note that  $(j^n(x^*(n\gamma_n, Y^n)))_{n \in \mathbb{N}}$  is a sequence of minimizers of the continuous Potts functionals  $\tilde{H}_{\gamma_n}^n(\cdot, j^n(Y^n))$ . Denoting  $\xi^n = j^n((\xi_{n,1}, \dots, \xi_{n,n}))$ , we obtain

$$\begin{aligned}
j^n(Y^n) &= \sum_{s \in S_n} Y_s^n \cdot \mathbf{1}_{[\frac{s-1}{n}, \frac{s}{n})} = \sum_{s \in S_n} (\bar{f}_s^n + \xi_{n,s}) \cdot \mathbf{1}_{[\frac{s-1}{n}, \frac{s}{n})} \\
&= j^n(\bar{f}^n) + \xi^n.
\end{aligned}$$

By Lemma 12.1.4,  $j^n(x^*(n\gamma_n, Y^n))$  also minimizes  $\tilde{H}_{\gamma_n}^n(\cdot, f + \xi^n)$ . Note further that  $f \in L^2([0, 1])$  is the unique minimizer of the functional

$$H_0^\infty(\cdot, f) : L^2([0, 1]) \longrightarrow \mathbb{R}, \quad g \longmapsto \|f - g\|^2. \quad (12.7)$$

Hence, the assertion of Theorem 12.0.5 is that a sequence of minimizers of the continuous Potts functionals  $\tilde{H}_{\gamma_n}^n(\cdot, j^n(Y^n))$  converges to the unique minimizer of  $H_0^\infty(\cdot, f)$ . This leads to the concept of epi-convergence, see for example A. BRAIDES (2002).

**Definition 12.2.1** Let  $(\Theta, d)$  be a metric space and  $F_n : \Theta \rightarrow \mathbb{R} \cup \{\pm\infty\}$ ,  $n \in \mathbb{N}$ , be numerical functions. The sequence  $(F_n)_{n \in \mathbb{N}}$  is **epi-convergent** to  $F_\infty : \Theta \rightarrow \mathbb{R} \cup \{\pm\infty\}$  if

- (1) for all  $\theta \in \Theta$  and for each sequence  $(\theta_n)_{n \in \mathbb{N}}$  with  $\theta_n \rightarrow \theta$  for  $n \rightarrow \infty$  the **lim inf-inequality**

$$F_\infty(\theta) \leq \liminf_{n \rightarrow \infty} F_n(\theta_n) \quad (12.8)$$

holds, and

- (2) for all  $\theta \in \Theta$  there is a sequence  $(\theta_n)_{n \in \mathbb{N}}$  with  $\theta_n \rightarrow \theta$  for  $n \rightarrow \infty$  such that the **lim sup-inequality**

$$F_\infty(\theta) \geq \limsup_{n \rightarrow \infty} F_n(\theta_n) \quad (12.9)$$

is fulfilled.

We will write  $F_n \xrightarrow{\text{epi}} F_\infty$  as  $n \rightarrow \infty$ .

The following theorem from G. BEER (1993) summarizes the main conclusions from epi-convergence.

**Theorem 12.2.2** ([5], Theorem 5.3.6) Let be  $(\Theta, d)$  a metric space and  $F_n, F_\infty : \Theta \rightarrow \mathbb{R} \cup \{\pm\infty\}$ ,  $n \in \mathbb{N}$ . Suppose  $(F_n)_{n \in \mathbb{N}} \xrightarrow{\text{epi}} F_\infty$  as  $n \rightarrow \infty$ .

- (1) Let  $(\theta_n)_{n \in \mathbb{N}}$  be a converging sequence of minimizers of  $F_n$ . Then  $\lim_{n \rightarrow \infty} \theta_n$  is a minimizer of  $F_\infty$ .
- (2) Let  $(\theta_n)_{n \in \mathbb{N}}$  be a sequence of minimizers of  $F_n$ . If there is a compact set  $K \subset \Theta$  such that  $\{\theta^* \in \Theta : F_n(\theta^*) = \min_{\theta \in \Theta} F_n(\theta)\} \subset K$  for large enough  $n$  then a minimizer of  $F_\infty$  exists and

$$d(\theta_n, \{\theta^* \in \Theta : F_\infty(\theta^*) = \min_{\theta \in \Theta} F_\infty(\theta)\}) \rightarrow 0$$

as  $n \rightarrow \infty$ .

- (3) Let  $(\theta_n)_{n \in \mathbb{N}}$  be a sequence of minimizers of  $F_n$ . If, in addition to the assumptions in (2), the functional  $F_\infty$  has a unique minimizer  $\theta^*$  then

$$\theta_n \rightarrow \theta^*$$

as  $n \rightarrow \infty$ .

To use the conclusions of Theorem 12.2.2 for our purposes we have to show the epi-convergence of  $\tilde{H}_{\gamma_n}^n(\cdot, f + \xi^n)$  and the relative compactness of the set of minimizers.

Define

$$\tilde{H}_0^\infty(\cdot, f) : L^2([0, 1]) \longrightarrow \mathbb{R}, \quad (g, f) \longmapsto \|f - g\|^2 + \sigma^2 \quad (12.10)$$

for  $f \in L^2([0, 1])$ . Clearly, differing only by a constant from the original functional  $H_0^\infty(\cdot, f)$  in (12.7), it has the same minimizers. The following proposition shows that  $\tilde{H}_0^\infty$  is the limit functional of  $(\tilde{H}_{\gamma_n}^n(\cdot, f + \xi^n))_{n \in \mathbb{N}}$ .

**Proposition 12.2.3** *Almost surely, for all sequences  $(\gamma_n)_{n \in \mathbb{N}} \subset (0, \infty)$  with  $\gamma_n \rightarrow 0$  and  $\gamma_n \frac{n}{\log n} \rightarrow \infty$  holds*

$$\tilde{H}_{\gamma_n}^n(\cdot, f + \xi^n) \xrightarrow{\text{epi}} \tilde{H}_0^\infty(\cdot, f) \quad \text{as } n \rightarrow \infty.$$

**Proof** The proof will be given later (page 146). □

The proof of this epi-convergence result needs some technical preparations. Let  $Z$  be a  $\mathcal{N}(0, \sigma^2)$ -distributed random variable. Then

$$\mathbb{P}(Z > z) < (2\pi)^{-1/2} \frac{\sigma}{z} e^{-\frac{z^2}{2\sigma^2}}, \quad z > 0, \quad (12.11)$$

see for example H. BAUER (1990), Lemma 4.1, p. 30. We derive the following

**Lemma 12.2.4** *The random variable*

$$X := \sup_{n \in \mathbb{N}} \frac{1}{\log n} \max_{1 \leq s \leq t \leq n} \frac{(\xi_{n,s} + \cdots + \xi_{n,t})^2}{t - s + 1}$$

is  $\mathbb{P}$ -almost surely finite, i. e.

$$\mathbb{P}(X < \infty) = 1.$$

**Proof** The prove uses the Borel-Cantelli-Lemma. We consider for  $z > 0$  the sets

$$A_{s,t}^n := \left\{ \omega \in \Omega : \frac{(\xi_{n,s}(\omega) + \cdots + \xi_{n,t}(\omega))^2}{t - s + 1} > z^2 \log n \right\}, \quad 1 \leq s \leq t \leq n.$$

We get

$$\mathbb{P}\left(\frac{(\xi_{n,s} + \cdots + \xi_{n,t})^2}{t - s + 1} > z^2 \log n\right) = 2 \cdot \mathbb{P}\left(\frac{\xi_{n,s} + \cdots + \xi_{n,t}}{\sqrt{t - s + 1}} > z\sqrt{\log n}\right).$$

Since  $(\xi_{n,s})_{s \in S_n}$  are independent centered Gaussian random variables with variance  $\sigma^2 > 0$  we have

$$\frac{\xi_{n,s} + \cdots + \xi_{n,t}}{\sqrt{t-s+1}} \sim \mathcal{N}(0, \sigma^2).$$

Inserting the inequality (12.11), we arrive at

$$\mathbb{P}\left(\frac{(\xi_{n,s} + \cdots + \xi_{n,t})^2}{t-s+1} > z^2 \log n\right) < \sqrt{\frac{2}{\pi}} \cdot \frac{\sigma}{z\sqrt{\log n}} \cdot \exp\left(-\frac{z^2 \log n}{2\sigma^2}\right).$$

Thus, for the sum of the probabilities of the sets  $A_{s,t}^n$ ,  $1 \leq s \leq t \leq n$ , we get

$$\begin{aligned} & \sum_{n \in \mathbb{N}} \sum_{1 \leq s \leq t \leq n} \mathbb{P}\left(\frac{(\xi_{n,s} + \cdots + \xi_{n,t})^2}{t-s+1} > z^2 \log n\right) \\ & < \sum_{n \in \mathbb{N}} \sum_{1 \leq s \leq t \leq n} \sqrt{\frac{2}{\pi}} \frac{\sigma}{z\sqrt{\log n}} e^{-\frac{z^2 \log n}{2\sigma^2}} = \sqrt{\frac{2}{\pi}} \sum_{n \in \mathbb{N}} \frac{\sigma}{z\sqrt{\log n}} e^{-\frac{z^2 \log n}{2\sigma^2}} \cdot \frac{n(n+1)}{2} \\ & < \sqrt{\frac{2}{\pi}} \frac{\sigma}{z} \sum_{n \in \mathbb{N}} n^2 \cdot e^{-\frac{z^2 \log n}{2\sigma^2}} = \sqrt{\frac{2}{\pi}} \frac{\sigma}{z} \sum_{n \in \mathbb{N}} n^{-\frac{z^2 - 4\sigma^2}{2\sigma^2}} < \infty. \end{aligned}$$

The Borel-Cantelli-Lemma yields that  $\mathbb{P}$ -almost surely

$$\frac{1}{\log n} \max_{1 \leq s \leq t \leq n} \frac{(\xi_{n,s}(\omega) + \cdots + \xi_{n,t}(\omega))^2}{t-s+1} > z^2$$

only for finitely many  $n \in \mathbb{N}$ . This implies that  $\mathbb{P}$ -almost surely

$$\sup_{n \in \mathbb{N}} \frac{1}{\log n} \max_{1 \leq s \leq t \leq n} \frac{(\xi_{n,s}(\omega) + \cdots + \xi_{n,t}(\omega))^2}{t-s+1} < \infty$$

which is the assertion.  $\square$

For any finite set  $J \subset (0, 1)$  we will define the  $\sigma$ -field

$$\mathcal{B}_J = \sigma(\{[a, b) : a \in J \cup \{0\}, b \in J \cup \{1\}\})$$

and the partition  $\mathcal{P}_J$  of  $[0, 1)$  induced by the atoms of the  $\sigma$ -field  $\mathcal{B}_J$ . For the set  $J = \{j_1, \dots, j_{|J|}\}$  the induced partition  $\mathcal{P}_J$  of  $[0, 1)$  provides a partition  $\mathcal{P}_J^n = \{I_1^n, \dots, I_{|J|+1}^n\}$  of  $S_n$  with

$$I_l^n = \bigcup_{s \in S_n} \{s \in S_n : nj_{l-1} \leq s < nj_l\}, \quad 1 \leq l \leq |J| + 1,$$

where  $j_0 := 0$  and  $j_{|J|+1} := 1$ .

Minimizers of the continuous Potts functional  $\tilde{H}_\gamma^n(\cdot, f)$  are, as in the discrete case, determined by their jumps set.

**Lemma 12.2.5** *Minimizers  $g^*$  of  $\tilde{H}_\gamma^n(\cdot, f)$  from (12.5) are in  $\mathfrak{X}^n$  and determined by their jump set  $\mathcal{J}(g^*)$ . If  $\mathcal{J}(g^*)$  is fixed then  $g^*$  is given as*

$$g^* = \sum_{I \in \mathcal{P}_{\mathcal{J}(g^*)}} \frac{\int_I f d\lambda}{\lambda(I)} \mathbf{1}_I.$$

**Proof** This lemma is the analogue to Proposition 2.1.1. Once a jump set  $J$  in (12.5) is fixed, the minimization of  $\tilde{H}_\gamma^n(\cdot, f)$  boils down to the minimization of  $g \mapsto \|f - g\|^2$  on the subspace

$$\mathfrak{X}_J^n = \{g \in \mathfrak{X}^n : \mathcal{J}(g) = J\}.$$

The data term  $\|f - g\|^2$  is minimal on  $\mathfrak{X}_J^n$  if and only if  $g$  is the orthogonal projection of  $f$  to  $\mathfrak{X}_J^n$  given by

$$P_{\mathfrak{X}_J^n} f = \sum_{I \in \mathcal{P}_J} \frac{\int_I f d\lambda}{\lambda(I)} \mathbf{1}_I$$

which is the stated form. □

If the cardinality of the jump sets  $(J_n)_{n \in \mathbb{N}}$  does not increase too fast, the projection of the random variable  $\xi^n$  to  $\mathfrak{X}_{J_n}^n$  tends to zero.

**Lemma 12.2.6** *There is a set of  $\mathbb{P}$ -probability 1 on which for all sequences  $(J_n)_{n \in \mathbb{N}}$  of finite sets in  $(0, 1)$ ,  $J_n \subset \{1/n, 2/n, \dots, (n-1)/n\}$  the relation*

$$\lim_{n \rightarrow \infty} \frac{\log n}{n} |J_n| = 0 \tag{12.12}$$

*implies*

$$P_{\mathfrak{X}_{J_n}^n} \xi^n \longrightarrow 0 \quad \text{in } L^2([0, 1)) \quad \text{as } n \rightarrow \infty.$$

**Proof** The orthogonal projection of  $\xi^n$  to  $\mathfrak{X}_{J_n}^n$  is given by

$$\begin{aligned} P_{\mathfrak{X}_{J_n}^n} \xi^n &= \sum_{I \in \mathcal{P}_{J_n}} \frac{\int_I \xi^n d\lambda}{\lambda(I)} \mathbf{1}_I \\ &= \sum_{I \in \mathcal{P}_{J_n}} \frac{\int_I \left( \sum_{s \in S_n} \xi_{n,s} \mathbf{1}_{[\frac{s-1}{n}, \frac{s}{n})} \right) d\lambda}{\lambda(I)} \mathbf{1}_I \\ &= \sum_{I \in \mathcal{P}_{J_n}} \frac{1}{\lambda(I)} \left( \sum_{s \in S_n} \xi_{n,s} \int_I \mathbf{1}_{[\frac{s-1}{n}, \frac{s}{n})} d\lambda \right) \mathbf{1}_I. \end{aligned}$$

The integral is zero if  $[(s - 1/n), s/n) \cap I = \emptyset$  and equal to  $1/n$  otherwise. Thus, we get

$$P_{\mathfrak{X}_{J_n}^n} \xi^n = \sum_{I \in \mathcal{P}_{J_n}} \frac{1}{n\lambda(I)} \left( \sum_{\{s \in S_n: [\frac{s-1}{n}, \frac{s}{n}) \subset I\}} \xi_{n,s} \right) \mathbf{1}_I.$$

Let  $\mathcal{P}_{J_n}^n$  denote the partition of  $S_n$  corresponding to  $\mathcal{P}_{J_n}$ . Then

$$\begin{aligned} \|P_{\mathfrak{X}_{J_n}^n} \xi^n\|^2 &= \int_0^1 (P_{\mathfrak{X}_{J_n}^n} \xi^n)^2 d\lambda = \sum_{I \in \mathcal{P}_{J_n}} \left( \frac{1}{n\lambda(I)} \sum_{s \in I^n} \xi_{n,s} \right)^2 \lambda(I) \\ &= \frac{1}{n} \sum_{I \in \mathcal{P}_{J_n}^n} \frac{1}{|I|} \left( \sum_{s \in I} \xi_{n,s} \right)^2 \\ &\leq \frac{1}{n} (|J_n| + 1) \cdot \sup_{I \in \mathcal{P}_{J_n}^n} \frac{(\sum_{s \in I} \xi_{n,s})^2}{|I|} \\ &\leq \frac{\log n}{n} (|J_n| + 1) X \longrightarrow 0 \quad \mathbb{P}\text{-almost surely} \end{aligned} \quad (12.13)$$

for  $X$  from Lemma 12.2.4. Condition (12.12) then gives the assertion.  $\square$

We consider now the data term of  $\tilde{H}_\gamma^n$ .

**Lemma 12.2.7** *Almost surely, for any sequence of sets  $(J_n)_{n \in \mathbb{N}}$  with (12.12) and  $(g_n)_{n \in \mathbb{N}} \subset L^2([0, 1])$  where  $g_n$  is  $\mathcal{B}_{J_n}$ -measurable and  $g_n \rightarrow g$  for  $n \rightarrow \infty$ , we have*

$$\|f + \xi^n - g_n\|^2 \longrightarrow \|f - g\|^2 + \sigma^2 \quad \text{as } n \rightarrow \infty.$$

**Proof** First observe that

$$\begin{aligned} \|f + \xi^n - g_n\|^2 &= \|f\|^2 + 2\langle f, \xi^n \rangle + \|\xi^n\|^2 - 2\langle f, g_n \rangle - 2\langle \xi^n, g_n \rangle + \|g_n\|^2 \\ &= \|f - g_n\|^2 + 2\langle f, \xi^n \rangle + \|\xi^n\|^2 - 2\langle \xi^n, g_n \rangle. \end{aligned}$$

We consider the single terms.

(1) We have that  $\|f - g_n\|^2 \rightarrow \|f - g\|^2$  for  $n \rightarrow \infty$  since  $g_n$  converges to  $g$  for  $n \rightarrow \infty$ .

(2) Since  $\xi^n \in \mathfrak{X}^n$  the inner product of  $f$  and  $\xi^n$  can be computed by

$$\langle f, \xi^n \rangle = \langle P_{\mathfrak{X}^n} f, \xi^n \rangle = \int_0^1 \sum_{s \in S_n} (\bar{f}_s^n \cdot \xi_{n,s}) \mathbf{1}_{[\frac{s-1}{n}, \frac{s}{n})} d\lambda = \frac{1}{n} \sum_{s \in S_n} \bar{f}_s^n \cdot \xi_{n,s}.$$

By Hypothesis 12.0.3 on  $(\xi_{n,s})_{s \in S_n}$ , the expression  $\langle f, \xi^n \rangle$  is as a sum of independent Gaussian random variables also a centered Gaussian random variable with variance

$$\begin{aligned} \text{Var}(\langle f, \xi^n \rangle) &= \sum_{s \in S_n} \left( \frac{\bar{f}_s^n}{n} \right)^2 \text{Var}(\xi_{n,s}) = \frac{\sigma^2}{n} \cdot \frac{1}{n} \sum_{s \in S_n} (\bar{f}_s^n)^2 \\ &= \frac{\sigma^2}{n} \cdot \|P_{\mathfrak{X}^n} f\|^2 \leq \frac{\sigma^2}{n} \cdot \|f\|^2. \end{aligned}$$

To show that  $\langle f, \xi^n \rangle$  tends  $\mathbb{P}$ -almost surely to zero we will use the following theorem, see D. WILLIAMS (1991), Theorem 12.2, p. 112: Let  $(X_n)_{n \in \mathbb{N}}$  be a sequence of independent random variables with  $\mathbb{E}(X_n) = 0$  and  $\text{Var}(X_n) < \infty$  for every  $n$ . Then  $\sum_n \text{Var}(X_n) < \infty$  implies that  $\sum_n X_n$  converges almost surely.

We will consider the independent random variables

$$X_n := \langle f, \xi^n \rangle^2 - \mathbb{E}(\langle f, \xi^n \rangle^2) = \langle f, \xi^n \rangle^2 - \frac{\sigma^2}{n} \cdot \|P_{\mathfrak{X}^n} f\|^2.$$

They have zero mean and the variance is given by

$$\begin{aligned} \text{Var}(X_n) &= \mathbb{E}(X_n^2) = \mathbb{E}(\langle f, \xi^n \rangle^4) - \frac{2\sigma^2}{n} \|P_{\mathfrak{X}^n} f\|^2 \cdot \mathbb{E}(\langle f, \xi^n \rangle^2) + \frac{\sigma^4}{n^2} \|P_{\mathfrak{X}^n} f\|^4 \\ &= \mathbb{E}(\langle f, \xi^n \rangle^4) - \frac{\sigma^4}{n^2} \|P_{\mathfrak{X}^n} f\|^4. \end{aligned}$$

Using the formula for the moments of the normal distribution (see for example J. SCHMETTERER (1966), p. 71) we get for the fourth moment of  $\langle f, \xi^n \rangle$

$$\mathbb{E}(|\langle f, \xi^n \rangle|^4) = 3 \cdot (\text{Var}(\langle f, \xi^n \rangle))^2 = \frac{3\sigma^4}{n^2} \cdot \|P_{\mathfrak{X}^n} f\|^4.$$

This yields

$$\text{Var}(X_n) = \frac{2\sigma^4}{n^2} \|P_{\mathfrak{X}^n} f\|^4 \leq \frac{2\sigma^4}{n^2} \|f\|^4,$$

and hence, the sum of variances of  $X_n$  converges. By the cited theorem above, this implies that  $\sum_n X_n$  converges  $\mathbb{P}$ -almost surely which implies that

$$\langle f, \xi^n \rangle^2 - \frac{\sigma^2}{n} \cdot \|P_{\mathfrak{X}^n} f\|^2 \longrightarrow 0 \quad \mathbb{P}\text{-almost surely} \quad \text{for } n \rightarrow \infty.$$

Since

$$\langle f, \xi^n \rangle^2 \leq |\langle f, \xi^n \rangle^2 - \frac{\sigma^2}{n} \cdot \|P_{\mathfrak{X}^n} f\|^2| + |\frac{\sigma^2}{n} \cdot \|P_{\mathfrak{X}^n} f\|^2|$$

we have that

$$\langle f, \xi^n \rangle^2 \longrightarrow 0 \quad \mathbb{P}\text{-almost surely} \quad \text{for } n \rightarrow \infty$$

which implies that  $\langle f, \xi^n \rangle$  tends to zero  $\mathbb{P}$ -almost surely.

(3) The third term is

$$\|\xi^n\|^2 = \int_0^1 \left( \sum_{s \in S_n} \xi_{n,s} \mathbf{1}_{[\frac{s-1}{n}, \frac{s}{n})} \right)^2 d\lambda = \int_0^1 \sum_{s \in S_n} \xi_{n,s}^2 \mathbf{1}_{[\frac{s-1}{n}, \frac{s}{n})} d\lambda = \frac{1}{n} \sum_{s \in S_n} \xi_{n,s}^2.$$

The  $(\xi_{n,s}^2)$  are independent and identically distributed random variables with finite expectation  $\sigma^2$ . Hence,

$$\|\xi^n\|^2 \longrightarrow \sigma^2 \quad \mathbb{P}\text{-almost surely} \quad \text{for } n \rightarrow \infty$$

by the strong law of large numbers.

(4) For the remaining term we obtain

$$\langle \xi^n, g_n \rangle = \langle \xi^n, P_{\mathfrak{X}_{J_n}^n} g_n \rangle = \langle P_{\mathfrak{X}_{J_n}^n} \xi^n, g_n \rangle$$

since  $g_n$  is  $\mathcal{B}_{J_n}$ -measurable. By the Cauchy-Schwarz-inequality we have

$$|\langle P_{\mathfrak{X}_{J_n}^n} \xi^n, g_n \rangle| \leq \|P_{\mathfrak{X}_{J_n}^n} \xi^n\| \cdot \|g_n\|.$$

By assumption,  $(J_n)_{n \in \mathbb{N}}$  fulfills condition (12.12) such that  $\mathbb{P}$ -almost surely

$$P_{\mathfrak{X}_{J_n}^n} \xi^n \longrightarrow 0 \quad \text{in } L^2([0, 1]) \quad \text{as } n \rightarrow \infty$$

by Lemma 12.2.6. Since  $(g_n)_{n \in \mathbb{N}}$  converges in  $L^2([0, 1])$  the sequence  $(\|g_n\|)_{n \in \mathbb{N}}$  is bounded. Thus, we have

$$\langle P_{\mathfrak{X}_{J_n}^n} \xi^n, g_n \rangle \longrightarrow 0 \quad \mathbb{P}\text{-almost surely} \quad \text{as } n \rightarrow \infty.$$

In summary, the items (1)-(4) give the assertion. □

We will need that the set of functions in  $L^2([0, 1])$  with a limited number of jumps is a closed subset which corresponds to lower semicontinuity of the number of jumps.

**Definition 12.2.8** *A functional  $F : X \rightarrow \mathbb{R} \cup \{\pm\infty\}$  is called **lower semi-continuous** if for all  $u \in X$  and for all sequences  $(u_n)_n$  tending to  $u$  we have*

$$F(u) \leq \liminf_{n \rightarrow \infty} F(u_n).$$

The last ingredient to prove the epi-convergence of the continuous Potts functionals is the lower semicontinuity of the number of jumps.

**Lemma 12.2.9** *The map*

$$L^2([0, 1]) \longrightarrow \mathbb{N}_0 \cup \{\infty\}, \quad g \longmapsto \begin{cases} |\mathcal{J}(g)| & \text{for } g \in \mathcal{T}([0, 1]), \\ +\infty & \text{otherwise} \end{cases}$$

*is lower semicontinuous.*

**Proof** We will show that the set  $\{g \in L^2([0, 1]) : |\mathcal{J}(g)| > M\}$  is open for all  $M \in \mathbb{N}_0$ . Assume that  $(g_n)_{n \in \mathbb{N}}$  is a sequence in  $\mathcal{T}([0, 1])$  converging to  $g \in \mathcal{T}([0, 1])$  with  $|\mathcal{J}(g_n)| \leq M < |\mathcal{J}(g)|$ . The sequence  $(\mathcal{J}(g_n))_{n \in \mathbb{N}}$  is a sequence of bounded closed sets. Let  $\mathcal{C}(X)$  be the set of all closed sets in the compact metric space  $(X, d)$ . For  $(A, B) \in \mathcal{C}(X) \times \mathcal{C}(X)$  define  $h(A, B) = \sup\{d(x, B) : x \in A\}$ . Equipped with the Hausdorff metric  $d_{\mathcal{H}}$  given by

$$d_{\mathcal{H}}(A, B) = \max\{h(A, B), h(B, A)\}$$

the space  $(\mathcal{C}(X), d_{\mathcal{H}})$  is a compact metric space. Since in a metric space compactness and sequentially compactness is the same, we can extract a subsequence, again denoted by  $(\mathcal{J}(g_n))_{n \in \mathbb{N}}$ , with

$$\mathcal{J}(g_n) \cup \{0, 1\} \xrightarrow{d_{\mathcal{H}}} J \cup \{0, 1\} \quad \text{as } n \rightarrow \infty$$

for some closed set  $J \subset [0, 1]$ . The cardinality is lower semicontinuous with respect to  $d_{\mathcal{H}}$  and thus,  $J$  is finite. Let  $(s, t) \subset [0, 1)$  be such that  $(s, t) \cap J = \emptyset$  and let  $\varepsilon > 0$ . Then  $(s + \varepsilon, t - \varepsilon) \cap \mathcal{J}(g_n) = \emptyset$  and hence  $g_n$  is constant on  $(s + \varepsilon, t - \varepsilon)$ . We observe that  $g_n \mathbf{1}_{(s+\varepsilon, t-\varepsilon)}$  converges in  $L^2([0, 1])$  to  $g \mathbf{1}_{(s+\varepsilon, t-\varepsilon)}$  such that  $g$  is constant on  $(s + \varepsilon, t - \varepsilon)$  as well. Since  $\varepsilon$  was arbitrary we conclude that  $g$  is constant on  $(s, t)$ . Hence,  $g \in \mathcal{T}([0, 1])$  and  $\mathcal{J}(g) \subseteq J$ . Lower semicontinuity of the cardinality on the space of closed subsets of  $[0, 1]$  yields

$$|\mathcal{J}(g)| \leq |J| \leq \liminf_{n \rightarrow \infty} |\mathcal{J}(g_n)| \leq \limsup_{n \rightarrow \infty} |\mathcal{J}(g_n)|.$$

Since  $|\mathcal{J}(g_n)| \leq M$  for all  $n \in \mathbb{N}$ , we have that

$$|\mathcal{J}(g)| \leq \limsup_{n \rightarrow \infty} |\mathcal{J}(g_n)| \leq M$$

which contradicts the assumption  $|\mathcal{J}(g)| > M$ .  $\square$

We will now prove the epi-convergence of the continuous Potts functionals.

**Proof of Proposition 12.2.3** We have to show that for a set of  $\mathbb{P}$ -measure one the following two inequalities hold:

(i) If  $g_n \rightarrow g$  for  $n \rightarrow \infty$  then the liminf-inequality (12.8) is fulfilled, i. e.

$$\liminf_{n \rightarrow \infty} \tilde{H}_{\gamma_n}^n(g_n, f + \xi^n) \geq \tilde{H}_0^\infty(g, f).$$

(ii) For all  $g \in L^2([0, 1])$  there is a sequence  $(g_n)_{n \in \mathbb{N}} \subset L^2([0, 1])$  which converges to  $g$  in  $L^2([0, 1])$  such that the limsup-inequality (12.9) is fulfilled, i. e.  $\limsup_{n \rightarrow \infty} \tilde{H}_{\gamma_n}^n(g_n, f + \xi^n) \leq \tilde{H}_0^\infty(g, f)$ .

(i) Let  $g_n \rightarrow g$  for  $n \rightarrow \infty$ . We want to investigate the limes inferior of  $\tilde{H}_{\gamma_n}^n(\cdot, f + \xi^n)$  meaning that there is a subsequence  $(g_{n_k})_k$  of  $(g_n)_n$  such that

$$\liminf_{n \rightarrow \infty} \tilde{H}_{\gamma_n}^n(g_n, f + \xi^n) = \lim_{k \rightarrow \infty} \tilde{H}_{\gamma_{n_k}}^{n_k}(g_{n_k}, f + \xi^{n_k})$$

Taking this subsequence if necessary, we may assume without loss of generality that  $\tilde{H}_{\gamma_n}^n(g_n, f + \xi^n)$  converges in  $\mathbb{R} \cup \{\pm\infty\}$ . We distinguish the following three cases:

(a) Suppose that  $g_n \notin \mathfrak{X}^n$  for infinitely many  $n \in \mathbb{N}$ . Then

$$\liminf_{n \rightarrow \infty} \tilde{H}_{\gamma_n}^n(g_n, f + \xi^n) = \limsup_{n \rightarrow \infty} \tilde{H}_{\gamma_n}^n(g_n, f + \xi^n) = +\infty$$

and the liminf-inequality is trivially fulfilled.

(b) Suppose that

$$|\mathcal{J}(g_n)| > \frac{\tilde{H}_0^\infty(g, f)}{\gamma_n} \quad \text{for infinitely many } n \in \mathbb{N}.$$

Then

$$\tilde{H}_{\gamma_n}^n(g_n, f + \xi^n) > \tilde{H}_0^\infty(g, f) + \|f + \xi^n - g_n\|^2 \geq \tilde{H}_0^\infty(g, f)$$

for infinitely many  $n \in \mathbb{N}$ . Hence, we have

$$\liminf_{n \rightarrow \infty} \tilde{H}_{\gamma_n}^n(g_n, f + \xi^n) = \limsup_{n \rightarrow \infty} \tilde{H}_{\gamma_n}^n(g_n, f + \xi^n) \geq \tilde{H}_0^\infty(g, f),$$

and the liminf-inequality is fulfilled.

The cases (a) and (b) do not exclude each other but are complementary to

(c) We have that  $g_n \in \mathfrak{X}^n$  for finitely many  $n \in \mathbb{N}$  and

$$|\mathcal{J}(g_n)| > \frac{\tilde{H}_0^\infty(g, f)}{\gamma_n} \quad \text{for finitely many } n \in \mathbb{N}.$$

We will show that in this case  $(\mathcal{J}(g_n))_{n \in \mathbb{N}}$  fulfills condition (12.12). Since  $\log n/n$  and  $|\mathcal{J}(g_n)|$  are both nonnegative, we have  $\liminf_{n \rightarrow \infty} \frac{\log n}{n} |\mathcal{J}(g_n)| \geq 0$  as well. For the limes superior we get

$$\begin{aligned} 0 \leq \liminf_{n \rightarrow \infty} \frac{\log n}{n} |\mathcal{J}(g_n)| &\leq \limsup_{n \rightarrow \infty} \frac{\log n}{n} |\mathcal{J}(g_n)| \\ &\leq \limsup_{n \rightarrow \infty} \frac{\log n}{n \cdot \gamma_n} \tilde{H}_0^\infty(g, f) = 0 \end{aligned}$$

since by assumption  $\lim_{n \rightarrow \infty} \gamma_n \frac{n}{\log n} = \infty$ . Hence, we have

$$\liminf_{n \rightarrow \infty} \frac{\log n}{n} |\mathcal{J}(g_n)| = \limsup_{n \rightarrow \infty} \frac{\log n}{n} |\mathcal{J}(g_n)| = \lim_{n \rightarrow \infty} \frac{\log n}{n} |\mathcal{J}(g_n)| = 0$$

which is (12.12). Application of Lemma 12.1.3 then yields

$$\tilde{H}_{\gamma_n}^n(g_n, f + \xi^n) = \gamma_n |\mathcal{J}(g_n)| + \|f + \xi^n - g_n\|^2 \geq \|f + \xi^n - g_n\|^2$$

for infinitely many  $n \in \mathbb{N}$ . Hence,

$$\begin{aligned} \liminf_{n \rightarrow \infty} \tilde{H}_{\gamma_n}^n(g_n, f + \xi^n) &\geq \liminf_{n \rightarrow \infty} \|f + \xi^n - g_n\|^2 \\ &= \lim_{n \rightarrow \infty} \|f + \xi^n - g_n\|^2 = \|f - g\|^2 + \sigma^2 = \tilde{H}_0^\infty(g, f) \end{aligned}$$

$\mathbb{P}$ -almost surely.

(ii) Choose  $g_n$  as some best approximation of  $g$  in  $\mathfrak{X}^n$  with at most  $1/\sqrt{\gamma_n}$  jumps. Then  $g_n \rightarrow g$  for  $n \rightarrow \infty$ , and the condition (12.12) is fulfilled since  $\frac{\log n}{n} |\mathcal{J}(g_n)| \geq 0$  and

$$\limsup_{n \rightarrow \infty} \frac{\log n}{n} |\mathcal{J}(g_n)| \leq \limsup_{n \rightarrow \infty} \frac{\log n}{n} \frac{1}{\sqrt{\gamma_n}} = \limsup_{n \rightarrow \infty} \frac{\log n}{n \cdot \gamma_n} \cdot \sqrt{\gamma_n} = 0.$$

By Lemma 12.1.3, we have  $\mathbb{P}$ -almost surely

$$\begin{aligned} \limsup_{n \rightarrow \infty} \tilde{H}_{\gamma_n}^n(g_n, f + \xi^n) &= \limsup_{n \rightarrow \infty} (\gamma_n \cdot |\mathcal{J}(g_n)| + \|f + \xi^n - g_n\|^2) \\ &\leq \limsup_{n \rightarrow \infty} \sqrt{\gamma_n} + \limsup_{n \rightarrow \infty} \|f + \xi^n - g_n\|^2 \\ &= 0 + \|f - g\|^2 + \sigma^2 = \tilde{H}_0^\infty(g, f) \end{aligned}$$

by the assumption  $\gamma_n \rightarrow 0$ . □

To draw conclusions from epi-convergence as in Theorem 12.2.2, we need the following compactness result.

**Lemma 12.2.10** *For any  $f \in L^2([0, 1])$  the set  $\{P_{\mathfrak{X}_J^n} f : J \subset (0, 1)\}$  is relatively compact.*

**Proof** The proof is done in several steps.

(1) For  $0 \leq s < t \leq 1$  the map  $(s, t) \mapsto \mathbf{1}_{[s,t]}$  is continuous since

$$\begin{aligned} \|\mathbf{1}_{[s,t]} - \mathbf{1}_{[s',t']}\|^2 &= \int_0^1 (\mathbf{1}_{[s,t]} - \mathbf{1}_{[s',t']})^2 d\lambda \\ &= \int_0^1 (\mathbf{1}_{[s,t]})^2 d\lambda - 2 \int_0^1 (\mathbf{1}_{[s,t]} \cdot \mathbf{1}_{[s',t']}) d\lambda + \int_0^1 (\mathbf{1}_{[s',t']})^2 d\lambda \\ &= t - s - 2 \cdot \max\left(\min(t, t') - \max(s, s'), 0\right) + t' - s'. \end{aligned}$$

Hence, the sets  $\{\sum_{i=1}^m \alpha_i \mathbf{1}_{I_i} : |\alpha_i| \leq C, I_i \subset [0, 1) \text{ right half-open intervals}\}$  are compact for all  $m \in \mathbb{N}$  and  $C > 0$  as they are continuous images of compact sets.

(2) If  $f = \mathbf{1}_I$  for some right half-open interval  $I \subset [0, 1)$  then, for any  $J \subset (0, 1)$ , the projection  $P_{\mathfrak{X}_J^n} f$  is a linear combination of at most three different indicator functions. Namely,

$$\begin{aligned} P_{\mathfrak{X}_J^n} f &= \sum_{I' \in \mathcal{P}_J} \frac{1}{\lambda(I')} \int_{I'} f d\lambda \mathbf{1}_{I'} = \sum_{I' \in \mathcal{P}_J} \frac{1}{\lambda(I')} \int_{I'} \mathbf{1}_I d\lambda \mathbf{1}_{I'} \\ &= \sum_{I' \in \mathcal{P}_J} \frac{1}{\lambda(I')} \lambda(I' \cap I) \mathbf{1}_{I'}. \end{aligned}$$

For those  $I' \in \mathcal{P}_J$  with  $I' \subset I$  it is  $\lambda(I' \cap I) = \lambda(I')$  and  $\sum_{I' \subset I} \mathbf{1}_{I'} = \mathbf{1}_I$ . Further contributions yield only the - at most two - intervals in  $\mathcal{P}_J$  which have nonempty intersection with  $I$ .

(3) If  $f = \sum_{i=1}^m \alpha_i \mathbf{1}_{I_i}$  is a right-continuous step function and  $J$  arbitrary then

$$P_{\mathfrak{X}_J^n} f = \sum_{I' \in \mathcal{P}_J} \frac{1}{\lambda(I')} \int_{I'} \left(\sum_{i=1}^m \alpha_i \mathbf{1}_{I_i}\right) d\lambda \mathbf{1}_{I'} = \sum_{I' \in \mathcal{P}_J} \frac{1}{\lambda(I')} \int_{I'} \sum_{i=1}^m \alpha_i \lambda(I' \cap I_i) \mathbf{1}_{I'}$$

and by the same arguments as in (2) we have  $P_{\mathfrak{X}_J^n} f = \sum_{j=1}^{m'} \beta_j \mathbf{1}_{J_j}$  for some  $m' \leq 3$ . It is

$$\begin{aligned} |\beta_j| &= \left| \frac{1}{\lambda(J_j)} \int_{J_j} f d\lambda \right| \leq \frac{1}{\lambda(J_j)} \int_{J_j} |f| d\lambda \leq \frac{1}{\lambda(J_j)} \cdot \lambda(J_j) \cdot \sup_{t \in J_j} |f(t)| \\ &\leq \|f\|_\infty < \infty \end{aligned}$$

since  $f$  is a step function. Hence, by (1), we get that  $\{P_{\mathfrak{X}_J^n} f : J \subset (0, 1)\}$  is relatively compact for right-continuous step functions  $f \in \mathcal{T}([0, 1])$ .

(4) Suppose now that  $f \in L^2([0, 1])$  is arbitrary and let  $\varepsilon > 0$ . We will show that we can cover  $\{P_{\mathfrak{x}_J^n} f : J \subset (0, 1)\}$  by finitely many  $\varepsilon$ -balls. Since step functions are dense in  $L^2([0, 1])$  we can fix a function  $g \in \mathcal{T}([0, 1])$  such that  $\|f - g\| < \varepsilon/2$ . By contractivity of projections, we get that  $\|P_{\mathfrak{x}_J^n} f - P_{\mathfrak{x}_J^n} g\| < \varepsilon/2$  for all finite  $J \subset (0, 1)$ . By (3), every covering of  $\{P_{\mathfrak{x}_J^n} g : J \subset (0, 1)\}$  contains a finite covering. Thus, there are finitely many  $J_1, \dots, J_p \subset (0, 1)$  such that

$$\max_{1 \leq l \leq p} \|P_{\mathfrak{x}_{J_l}^n} g - P_{\mathfrak{x}_{J_l}^n} g\| < \frac{\varepsilon}{2} \quad \text{for all finite } J \subset (0, 1).$$

This implies

$$\max_{1 \leq l \leq p} \|P_{\mathfrak{x}_J^n} f - P_{\mathfrak{x}_{J_l}^n} g\| \leq \|P_{\mathfrak{x}_J^n} f - P_{\mathfrak{x}_J^n} g\| + \min_{1 \leq l \leq p} \|P_{\mathfrak{x}_J^n} g - P_{\mathfrak{x}_{J_l}^n} g\| < \frac{\varepsilon}{2} + \frac{\varepsilon}{2} = \varepsilon,$$

and hence,

$$\{P_{\mathfrak{x}_J^n} f : J \subset (0, 1)\} \subset \bigcup_{i=1}^p B(P_{\mathfrak{x}_{J_i}^n} g, \varepsilon)$$

which is the pre-compactness of  $\{P_{\mathfrak{x}_J^n} f : J \subset (0, 1)\}$ . By Lemma 3.9, p. 18, in F. HIRZEBRUCH and W. SCHARLAU (1971), a subspace of a metric space is relatively compact exactly if it is pre-compact and its closure is complete. Its closure is complete as it is a closed subset of the complete space  $L^2([0, 1])$ . Thus, the proof is complete.  $\square$

## 12.3 Proof of the Main Theorem

At last, we complete the proof of the main theorem of this section.

**Proof of Theorem 12.0.5** By the fundamental Theorem 12.2.2 and the reformulation in Lemma 12.1.4, it is enough to prove that almost surely there is a compact set containing

$$\bigcup_{n \in \mathbb{N}} \{g_n \in L^2([0, 1]) : g_n \text{ minimizes } \tilde{H}_{\gamma_n}^n(\cdot, f + \xi^n)\}.$$

First note that all minimizers  $g_n$  of  $\tilde{H}_{\gamma_n}^n(\cdot, f + \xi^n)$  are projections  $P_{\mathfrak{x}_{J_n}^n}(f + \xi^n)$  for some random sets  $J_n \subset \{1/n, \dots, (n-1)/n\}$ . We compare the value of the minimizers with the value for the constant zero and get

$$\begin{aligned} \tilde{H}_{\gamma_n}^n(P_{\mathfrak{x}_{J_n}^n}(f + \xi^n), f + \xi^n) &= \gamma_n \cdot |J_n| + \|f + \xi^n - P_{\mathfrak{x}_{J_n}^n}(f + \xi^n)\|^2 \\ &\leq \tilde{H}_{\gamma_n}^n(0, f + \xi^n) = \|f + \xi^n\|^2 \end{aligned}$$

which implies

$$\gamma_n \cdot |J_n| \leq \|P_{\mathfrak{x}_{J_n}^n}(f + \xi^n)\|^2 \leq \|P_{\mathfrak{x}_{J_n}^n}(f + \xi^n)\|^2 + \|P_{\mathfrak{x}_{J_n}^n}(f - \xi^n)\|^2.$$

Application of the parallelogram inequality and the contraction property of projections then yield

$$\gamma_n \cdot |J_n| \leq 2\|P_{\mathfrak{x}_{J_n}^n} f\|^2 + 2\|P_{\mathfrak{x}_{J_n}^n} \xi^n\|^2 \leq 2\|f\|^2 + 2\|P_{\mathfrak{x}_{J_n}^n} \xi^n\|^2.$$

By inequality (12.13) in the proof of Lemma 12.2.6, we then get

$$\gamma_n \cdot |J_n| \leq 2\|f\|^2 + 2X \frac{\log n}{n} (|J_n| + 1)$$

for all  $n \in \mathbb{N}$  and  $X$  finite  $\mathbb{P}$ -almost surely. This is equivalent to

$$\left(\gamma_n \frac{n}{\log n} - 2X\right) \frac{\log n}{n} |J_n| \leq 2\|f\|^2 + 2X \frac{\log n}{n}$$

for all  $n \in \mathbb{N}$ . Since by assumption  $\lim_{n \rightarrow \infty} \gamma_n \frac{n}{\log n} = \infty$  and  $\lim_{n \rightarrow \infty} \frac{\log n}{n} = 0$  we have that

$$\lim_{n \rightarrow \infty} \frac{\log n}{n} |J_n| = 0 \quad \mathbb{P}\text{-almost surely}$$

which is condition (12.12). From Lemma 12.2.6 it then follows that  $\mathbb{P}$ -almost surely  $P_{\mathfrak{x}_{J_n}^n} \xi^n \rightarrow 0$ . Since, by Lemma 12.2.10, the set  $\{P_{\mathfrak{x}_{J_n}^n} f : J_n \subset (0, 1)\}$  is relatively compact in  $L^2([0, 1])$ , relative compactness of

$$\begin{aligned} & \bigcup_{n \in \mathbb{N}} \{g_n \in L^2([0, 1]) : g_n \text{ minimizes } \tilde{H}_{\gamma_n}^n(\cdot, f + \xi^n)\} \\ & \subset \bigcup_{n \in \mathbb{N}} \{P_{\mathfrak{x}_{J_n}^n} f + P_{\mathfrak{x}_{J_n}^n} \xi^n : J_n \subset (0, 1)\} \end{aligned}$$

follows by continuity of addition. This completes the proof.  $\square$



# Discussion and Outlook

We studied the Potts functionals

$$\bar{H}_\gamma : \mathbb{X} \longrightarrow \mathbb{R}, \quad x \longmapsto \gamma \cdot |J(x)| + \sum_{s \in S} (y_s - x_s)^2, \quad \gamma > 0 \quad (1)$$

for  $y \in \mathbb{R}^S$ . There are functionals with modified and more complicated penalties and variants in the data term. A more general class of functionals with  $\alpha$ -homogenous penalties was briefly mentioned in Remark 4.1.11. For the minimizers of these functionals we can prove equivariance properties similar to those for the MAP estimators of the Potts functionals. Other statements like existence, uniqueness, and continuity of minimizers are not straightforward to generalize, or even do not hold. For functionals with total variation penalty term, minimizers are considered for example in E. MAMMEN and S. VAN DE GEER (1997); an algorithm for the approximation of a minimizer can be found in P. L. DAVIES and A. KOVAC (2001). We are not aware of rigorous results for the dependence on the hyperparameter or continuity. In contrast, in Chapter 3 we presented exact and fast algorithms, at least in the case of time series<sup>7</sup>.

In summary, even moderate change of the penalty causes problems to carry out the ‘program’ outlined in this thesis. Another way to modify the Potts functionals is to allow other data terms. Sum of squares could for example be replaced by sum of absolute deviations. Further examples for data terms which guarantee at least the existence of minimizers were given in Example 2.2.6.

Statements on uniqueness and dependence on hyperparameters of minimizers are the programme for future work. The present thesis may serve as a kind of outline of what we wish to prove for more general functionals. This was also the reason to study just the functionals in (1). There we could achieve in some sense the maximum of rigorous results.

Another field of future work will be the choice of hyperparameters. Still missing is an intrinsic criterion which includes constant estimators. The

application of the last monotone criterion presented in Section 8.3 is a step into this direction but has the restriction that it needs additional information or assumptions on the morphology. One approach for other criteria is the analysis of the distribution of  $\gamma_0(y)$  in case of data  $y = Y^n$  arising from a constant signal which could be described by the model

$$Y_s^n = c + \xi_{n,s}, \quad s \in S = \{1, \dots, N\}, \quad (2)$$

where  $c$  is some constant and  $(\xi_{n,s})_{s \in S}$  are random variables from a certain distribution. The aim will be to prove a central limit theorem. This would allow to construct tests in order to exclude data arising from a constant signal with a certain error probability.

Central limit theorems are also in the focus of further consistency investigations. We think of a (possible random) sampled function which provides data replacing the constant  $c$  in the model (2) by the collected function values. The aim is then to prove convergence of suitably embedded minimizers of an appropriately scaled Potts functionals with fixed  $\gamma$  towards the original function.

**Part V**  
**Appendix**



In order to keep the text streamlined, we relieve it from too many figures and lengthy calculations, and collect them in this appendix. We also give a brief summary for the Akaike and the Schwarz information criteria for convenience of the reader.

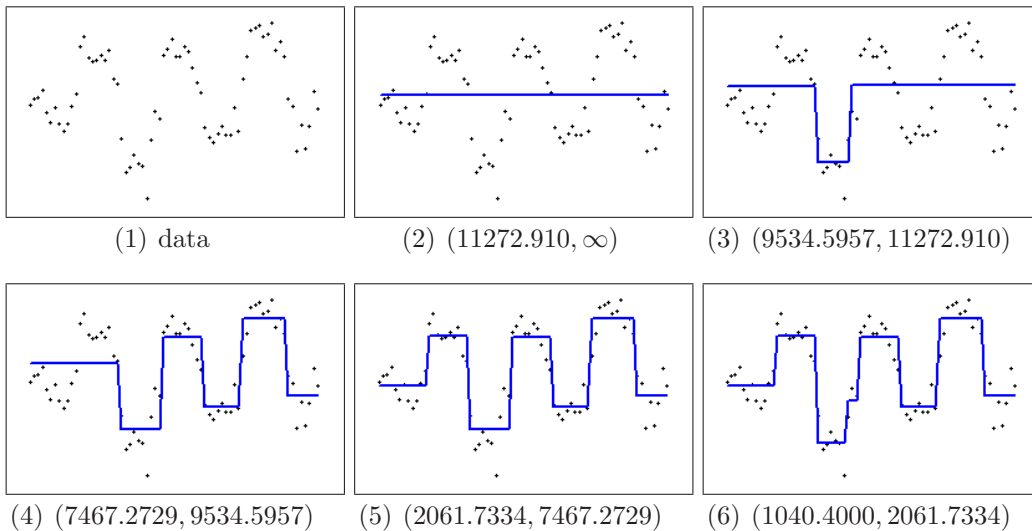
In Appendix A, exemplarily for a time series from the fMRI experiment described in Section 9.1, a full  $\gamma$ -scanning is displayed to illustrate Theorem 2.4.5. The subsequent two chapters are devoted to the model selection criteria. In Appendix B, we carry out in more detail what was only sketched in Chapter 7. We justify the choice of the family of regression models and summarize the well-known justifications and derivations of the Akaike and Schwarz information criteria. In addition, we then give a complete derivation of variants of these criteria in the special case of the considered family of regression models. They correct the original (asymptotic) criteria in the case of shorter time series' and seemed to be required for the fMRI-data (70 time points) and the fractionation curves (29 time points) from Chapter 9. The proofs of necessary lemmata, elementary but lengthy calculations, are contained in Appendix C.

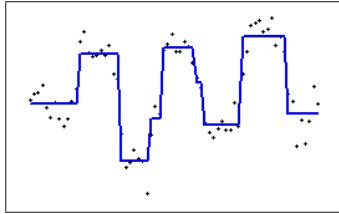


# Appendix A

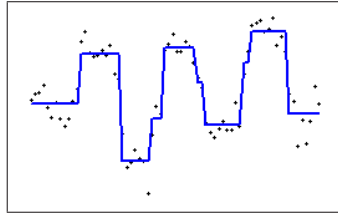
## $\gamma$ -Scanning

This is a supplement to Section 2.4. By Theorem 2.4.5 we know that for almost all data  $y$  there are only finitely many different MAP estimators  $x^*(\gamma, y)$  which are the same on  $\gamma$ -intervals. The number of jumps increases in these intervals, to be more precise, the functions  $i \mapsto |J(x^*(\gamma, y))|$ ,  $\gamma \in (\gamma_i(y), \gamma_{i-1}(y))$ , increase strictly. Figure 2.4 was a sample of MAP estimates for a certain time series, for sake of completeness, we display here the full  $\gamma$ -scanning. Figure A.1 displays the MAP estimates (blue line) for dotted data for all  $\gamma$ -intervals, starting with the constant estimate of the rightmost interval  $(\gamma_0, \infty)$ , and ending with data.

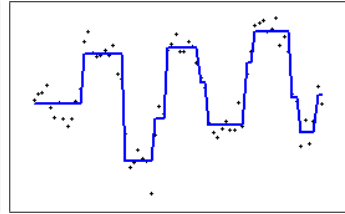




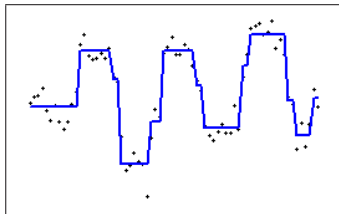
(7) (882.44446, 1040.4000)



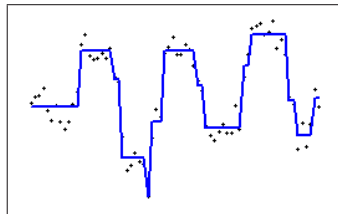
(8) (717.56250, 882.44446)



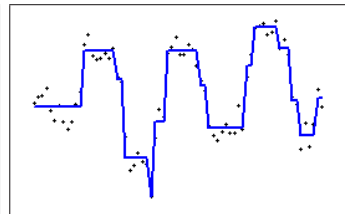
(9) (680.62500, 717.56250)



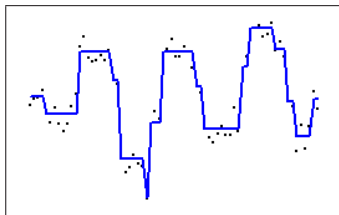
(10) (672.00000, 680.62500)



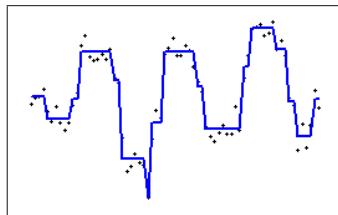
(11) (440.05554, 672.00000)



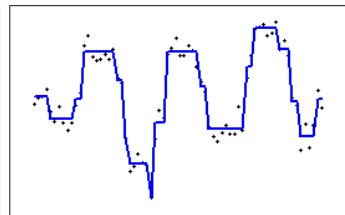
(12) (433.50000, 440.05554)



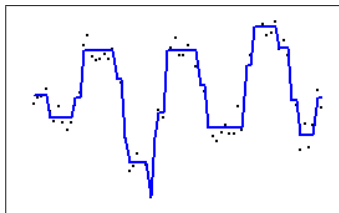
(13) (337.50000, 433.50000)



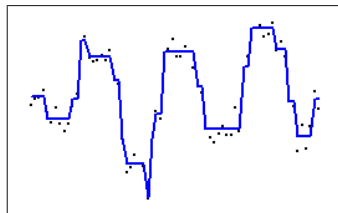
(14) (307.20001, 337.50000)



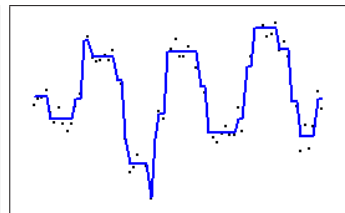
(15) (228.16667, 307.20001)



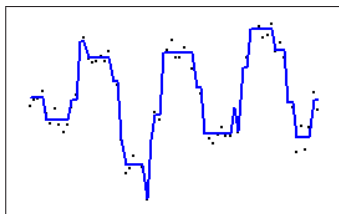
(16) (210.04167, 228.16667)



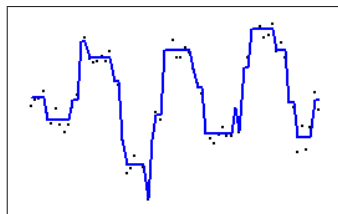
(17) (169.17461, 210.04167)



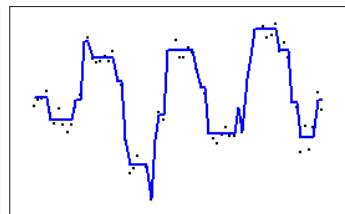
(18) (162.00000, 169.17461)



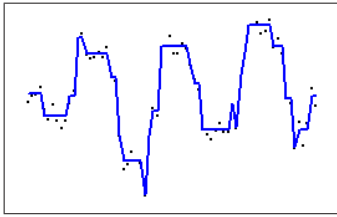
(19) (138.28572, 162.00000)



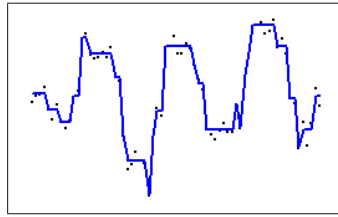
(20) (128.00000, 138.28572)



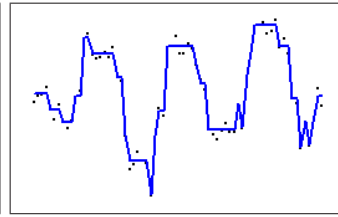
(21) (126.75000, 128.00000)



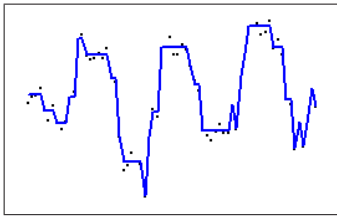
(22) (121.50000, 126.75000)



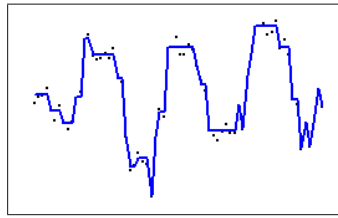
(23) (91.000000, 121.50000)



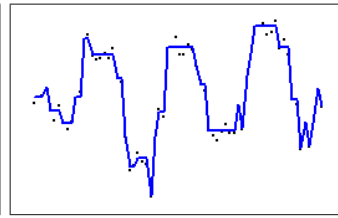
(24) (84.500000, 91.000000)



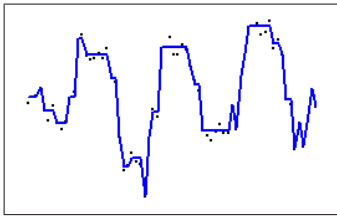
(25) (48.133335, 84.500000)



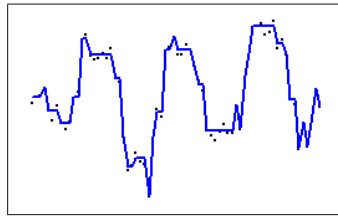
(26) (48.000000, 48.133335)



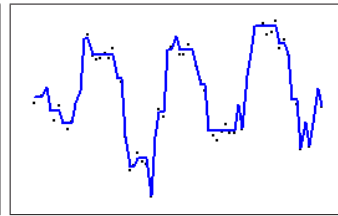
(27) (42.666668, 48.000000)



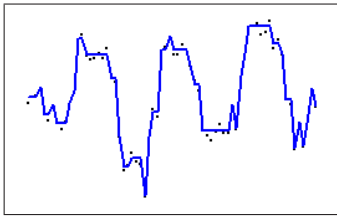
(28) (41.607143, 42.666668)



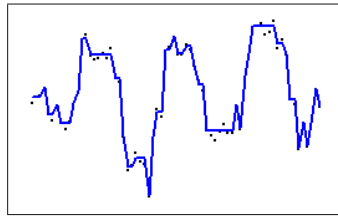
(29) (40.500000, 41.607143)



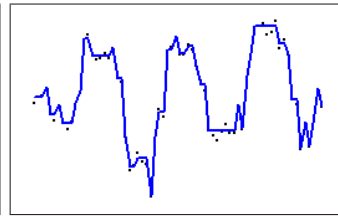
(30) (37.500000, 40.500000)



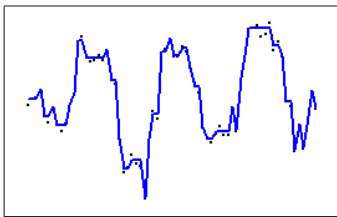
(31) (36.000000, 37.500000)



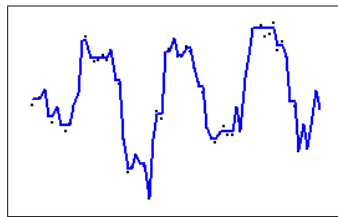
(32) (34.133335, 36.000000)



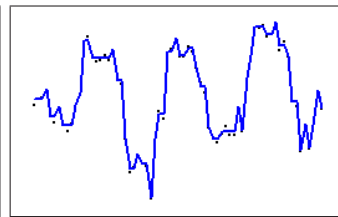
(33) (33.482143, 34.133335)



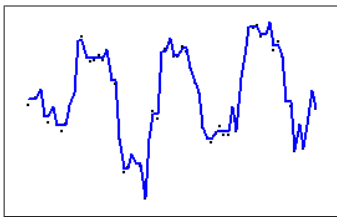
(34) (32.666668, 33.482143)



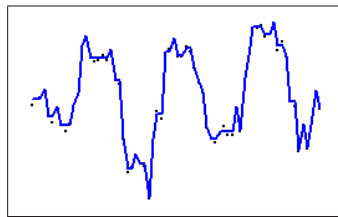
(35) (31.083334, 32.666668)



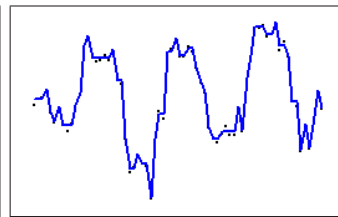
(36) (24.500000, 31.083334)



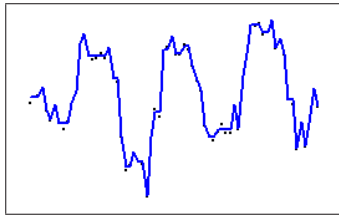
(37) (24.500000, 24.500000)



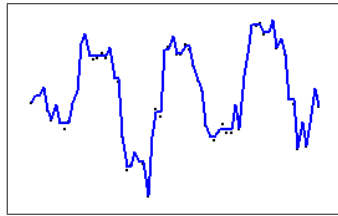
(38) (24.500000, 24.500000)



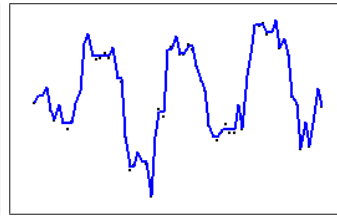
(39) (18.000000, 24.500000)



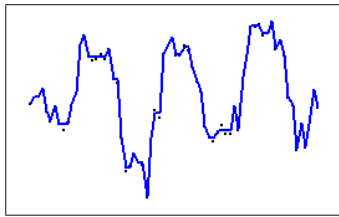
(40) (13.500000, 18.000000)



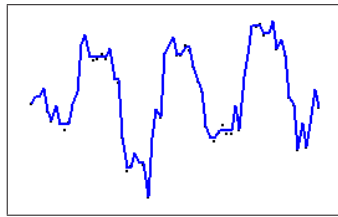
(41) (12.500000, 13.500000)



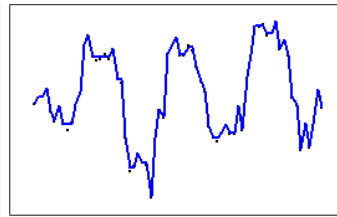
(42) (12.500000, 12.500000)



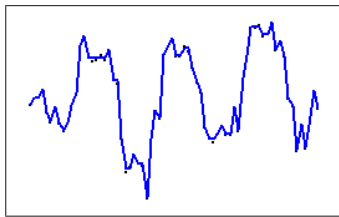
(43) (12.500000, 12.500000)



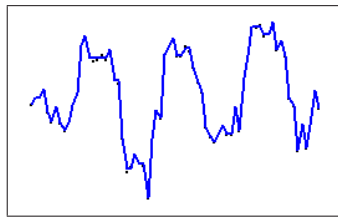
(44) (12.375000, 12.500000)



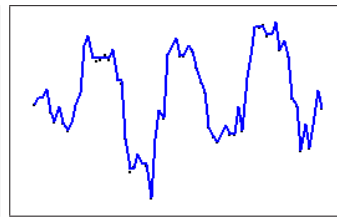
(45) (12.000000, 12.375000)



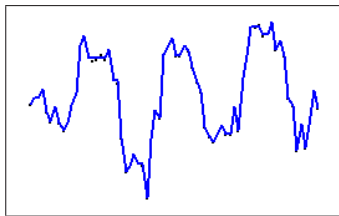
(46) (8.000000, 12.000000)



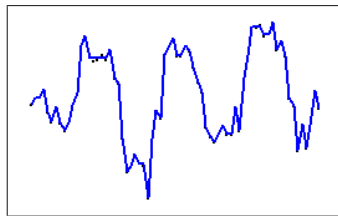
(47) (8.000000, 8.000000)



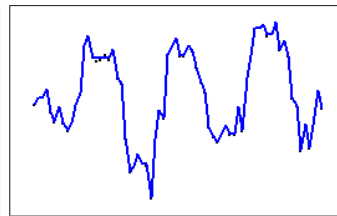
(48) (8.000000, 8.000000)



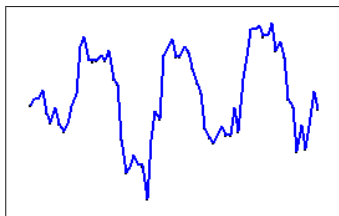
(49) (8.000000, 8.000000)



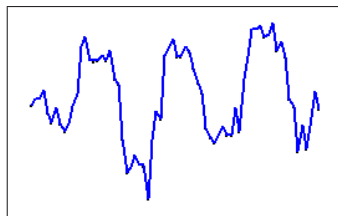
(50) (4.1666665, 8.000000)



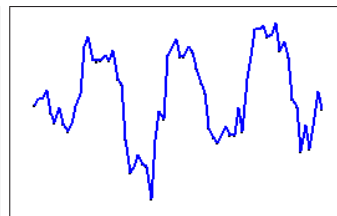
(51) (3.599999, 4.1666665)



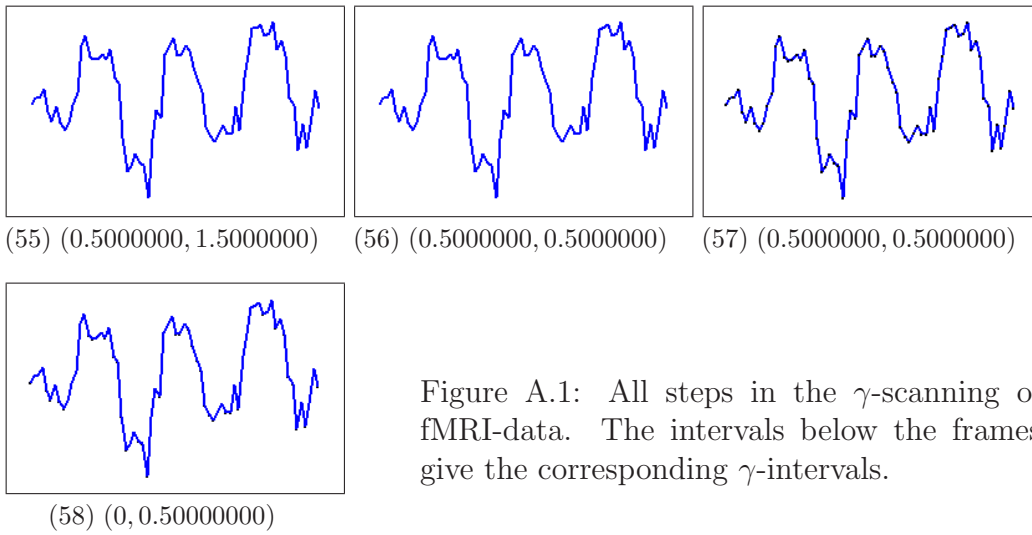
(52) (2.000000, 3.599999)



(53) (2.000000, 2.000000)



(54) (1.500000, 2.000000)





# Appendix B

## Model Selection Criteria

In Chapter 7 we used classical model selection criteria to choose the hyperparameter. We observed a close relation of the resulting estimators to MAP estimators for the Potts functional.

In Section B.1, we will now first specify a parametric family of regression models and give the corresponding maximum likelihood estimators. These models are piecewise constant approximations of data. The number of constant pieces is the dimension of the parameter (eventually plus one), and thus, it will be interpreted as the number of intervals in the partition induced by a MAP estimator. Hence, this family is the suitable one for the Potts functionals. There are several criteria to select and reduce the parameter dimension. We consult two classical model selection criteria. In Section B.2, we will briefly recall the well-known justification and derivation of the Akaike information criterion from H. AKAIKE (1974) in general. Finally, we will derive a variant of it in the setting of the special model class, and arrive at the well-known corrected version, known as AICC. It corrects the original criterion for shorter time series'. The analogous procedure for the Schwarz information criterion from G. SCHWARZ (1978) is carried out in Section B.3. Concerning the corrected version we will derive here, we are not aware of that this special form was already treated in the literature.

### B.1 A Simple Family of Regression Models

In Theorem 2.4.5 we established the connection between the  $\gamma$ -intervals and MAP estimators  $x^*(\gamma, y)$ . Recall that a MAP estimator was identified with a minimal segmentation. We further know that for almost all  $y \in \mathbb{X}$ , a MAP estimator is unique on the  $\gamma$ -intervals and that its number of jumps is constant on these  $\gamma$ -intervals. Hence, the choice of a  $\gamma$ -interval is equivalent to

the determination of the number of intervals in the partition of the (unique) segmentation corresponding to  $x^*(\gamma, y)$  for  $\gamma$  in this interval. This number is interpreted as the dimension of the parameter of the family of the simplest regression models which will be introduced below. This is not only a simple family but also the parametric family of models *suitable for the Potts functional*. We will restrict ourselves to the Potts functionals

$$\bar{H}_\gamma : \mathbb{R}^S \times \mathbb{R}^S \longrightarrow \mathbb{R}, \quad (x, y) \longmapsto \gamma \cdot |J(x)| + \sum_{s \in S} (y_s - x_s)^2.$$

We assume that the true deterministic signal  $x$  is corrupted by additive Gaussian white noise, i. e.

$$y_s = x_s + \varepsilon_s(\omega), \quad s = 1, \dots, N, \quad (\text{B.1})$$

where  $\varepsilon_s$ ,  $s = 1, \dots, N$ , are independent and identically distributed normal random variables with mean zero and variance  $\sigma^2$ . By Theorem 1.2.4 we can identify a signal  $x \in \mathbb{X}$  with the induced minimal segmentation  $(\mathcal{P}(x), \mu_{\mathcal{P}(x)}(x))$ . Suppose that  $\mathcal{P}(x) = \{J_1, \dots, J_k\}$ . Hence, a signal  $x \in \mathbb{X}$  corresponds to the parameter vector

$$\mu = \left( \underbrace{\mu_{J_1}, \dots, \mu_{J_1}}_{|J_1|}, \underbrace{\mu_{J_2}, \dots, \mu_{J_2}}_{|J_2|}, \dots, \underbrace{\mu_{J_k}, \dots, \mu_{J_k}}_{|J_k|} \right)^t \in \mathbb{R}^N.$$

We will now recall the definition of likelihood functions and of maximum likelihood estimators.

**Definition B.1.1** *Let  $y$  be a realization of the random variable  $Y$ . Given a family*

$$\Pi = \{f(\cdot, \theta); \theta \in \Theta\}$$

*of densities and a set of parameters  $\Theta \subset \mathbb{R}^N$ , the function*

$$L(\cdot | Y) : \Theta \longrightarrow \mathbb{R}, \quad \theta \longmapsto f(y, \theta)$$

*is called the **likelihood function** of  $y$ .*

*An estimator  $\hat{\theta}(y)$  which maximizes the likelihood function  $\theta \mapsto L(\theta | Y)$*

$$\hat{\theta} = \operatorname{argmax}_{\theta \in \Theta} L(\theta | Y)$$

*is called a **maximum likelihood estimator**.*



To avoid confusion with too many indices the log likelihood function of the  $k$ -th model will be written as

$$\ln L(\theta^k|Y) = -\frac{N}{2} \ln(2\pi\sigma_k^2) - \frac{1}{2\sigma_k^2} \sum_{I \in \mathcal{P}} \sum_{s \in I} (y_s - \mu_I^k)^2, \quad (\text{B.6})$$

having in mind that  $\mu_I^k$  and  $\mathcal{P}$  are the respective minimizers.

## B.2 The Akaike Information Criterion

Let  $L(\hat{\theta}^k|Y)$  denote the likelihood function of the distribution with parameter  $\theta$  evaluated at the maximum likelihood estimator  $\hat{\theta}^k = (\hat{\theta}_1^k, \dots, \hat{\theta}_k^k)$  in the subspace  $\Theta^k$  of the parameter space  $\Theta$ . Let  $k$  be the number of parameters to be estimated. H. AKAIKE (1973, 1974) suggested an information criterion (AIC) of the following form: Maximize the log likelihood function separately for the competing distributions and choose that distribution for which

$$\text{AIC}(k) = \ln L(\hat{\theta}^k|Y) - k \quad (\text{B.7})$$

is largest. This has become known as the *Akaike information criterion*. It is based on the minimization of the Kullback-Leibler information.

**Definition B.2.1** *Let be given a family of distributions with densities  $L(\theta|Y)$ . Let  $y$  be a sample of the distribution with parameter  $\theta_0$ . Denoting by  $\mathbb{E}_{\theta_0}$  the expectation with respect to  $\theta_0$ , the **Kullback-Leibler information** of a distribution with likelihood  $L(\theta|Y)$  with respect to the distribution with likelihood  $L(\theta_0|Y)$  is defined as*

$$KL(\theta_0, \theta) = \mathbb{E}_{\theta_0} \left( \ln \frac{L(\theta_0|Y)}{L(\theta|Y)} \right).$$

The Kullback-Leibler information has the following properties

- (1)  $KL(\theta_0, \theta) = \mathbb{E}_{\theta_0}(\ln L(\theta_0|Y)) - \mathbb{E}_{\theta_0}(\ln L(\theta|Y))$
- (2)  $KL(\theta_0, \theta) \geq 0$  and  $KL(\theta_0, \theta) = 0$  if and only if  $\theta = \theta_0$ .

We consider  $L(\theta_0|y)$  as the likelihood function of the data generating distribution. If one wants to select between distributions with parameters  $\theta^k$  from different subspaces  $\Theta^k$  of  $\Theta$  one approach is to minimize the Kullback-Leibler information in  $\theta$ . Since the first term  $\mathbb{E}_{\theta_0}(\ln L(\theta_0|Y))$  in (1) is the

same for all competing distributions this is equivalent to the maximization of the *discrepancy*

$$d(\theta_0, \theta) = \mathbb{E}_{\theta_0}(\ln L(\theta|Y)). \tag{B.8}$$

of distributions with parameter  $\theta$  with respect to the true distribution with parameter  $\theta_0$ . Once a subspace  $\Theta^k$  is fixed, the Kullback-Leibler information is minimal for the maximum likelihood estimator  $\hat{\theta}^k(y)$  in this subspace.

Therefore, when the parameter  $\theta$  must be estimated, the evaluation of the discrepancy at the maximum likelihood estimator  $\hat{\theta}^k(y)$

$$\delta(\Theta^k, \theta_0) = d(\theta_0, \theta)|_{\theta=\hat{\theta}^k(y)} \tag{B.9}$$

is a useful measure of the discrepancy (B.8) of the distribution with parameter in the subspace  $\Theta^k$ . As  $\theta_0$  is unknown, (B.9) cannot be computed exactly.

H. AKAIKE (1974) used the maximized log likelihood function as an estimator for (B.9). He derived that under certain regularity conditions its bias can be estimated by  $k$  and arrived at (B.7) as an unbiased estimator for (B.9).

### A Corrected Version of AIC

It is well known that AIC is applicable in a very general framework but has a large bias when the sample size is small, see for example C. M. HURVICH and C. L. TSAI (1989). The corrected version AICC, presented for example in C. M. HURVICH and C. L. TSAI (1989), corrects for this bias but its special form depends on the form of the candidate models. J.E. CAVANAUGH (1997) derives AIC and AICC for linear regression models which connects both. In this section we construct a corrected version of the Akaike information criterion for the family (B.6) of densities where the number  $k$  of intervals in the partition  $\mathcal{P}$  varies. The derivation follows the lines in J.E. CAVANAUGH (1997).

We assume not that the true variance  $\sigma_0^2$  in the model (B.1) is known. Hence, the variance is an additional parameter to be estimated. The expected value of the discrepancy (B.9) with respect to the true distribution corresponding to the parameter  $\theta_0$  can be written as

$$\begin{aligned} \mathbb{E}_{\theta_0}(d(\theta_0, \theta)|_{\theta=\hat{\theta}^k(y)}) &= \mathbb{E}_{\theta_0}(\mathbb{E}_{\theta_0}(\ln L(\theta|Y))|_{\theta=\hat{\theta}^k}) \\ &= \mathbb{E}_{\theta_0}(\ln L(\hat{\theta}^k|Y)) \\ &+ \mathbb{E}_{\theta_0}(\ln L(\theta_0|Y)) - \mathbb{E}_{\theta_0}(\ln L(\hat{\theta}^k|Y)) \end{aligned} \tag{B.10}$$

$$+ \mathbb{E}_{\theta_0}(\mathbb{E}_{\theta_0}(\ln L(\theta|Y))|_{\theta=\hat{\theta}^k}) - \mathbb{E}_{\theta_0}(\ln L(\theta_0|Y)). \tag{B.11}$$

The computations for the single parts (B.10) and (B.11) will be given in the following lemmata. For the difference (B.10) we get

**Lemma B.2.2**

$$\mathbb{E}_{\theta_0}(\ln L(\theta_0|Y)) - \mathbb{E}_{\theta_0}(\ln L(\hat{\theta}|Y)) = -N \ln(\sigma_0 \sqrt{2\pi}) + \mathbb{E}_{\theta_0}(N \ln(\hat{\sigma} \sqrt{2\pi})).$$

**Proof** The proof is given in Appendix C.1.  $\square$

For part (B.11) we have

**Lemma B.2.3**

$$\begin{aligned} & \mathbb{E}_{\theta_0}(\mathbb{E}_{\theta_0}(\ln L(\theta|Y))_{|\theta=\hat{\theta}^k}) - \mathbb{E}_{\theta_0}(\ln L(\theta_0|Y)) \\ &= \mathbb{E}_{\theta_0}(-N \ln(\hat{\sigma}_k \sqrt{2\pi})) - \frac{N(N+k)}{2(N-k-2)} + N \ln(\sigma_0 \sqrt{2\pi}) + \frac{N}{2}. \end{aligned}$$

**Proof** The proof is given in Appendix C.1.  $\square$

With the maximum likelihood estimators  $\hat{\mu}^k$  from (B.4) and  $\hat{\sigma}_k^2$  from (B.5) for  $\theta_0 = (\mu_0, \sigma_0^2)$  we arrive at the following corrected version of the Akaike information criterion (AICC). It is of the same form as for example in the setting of estimating the order of autoregressive time series, see for example P. J. BROCKWELL and R. A. DAVIS (1991).

**Theorem B.2.4** *Let  $y$  be a sample from the distribution with parameter  $(\mu_0, \sigma_0)$  from model (B.1). A corrected version of AIC from (7.5) including second order terms reads: Choose the distribution with that number  $k^*$  intervals in the corresponding minimizing partition for which*

$$AICC(k) = -\frac{N}{2} \ln(\hat{\sigma}_k^2) - \frac{N(N+k)}{2(N-k-2)}$$

*is maximal.*

**Proof** Using Lemma B.2.2 and Lemma B.2.3 we get for the term  $\mathbb{E}_{\theta_0}(\mathbb{E}_{\theta_0}(\ln L(\theta|Y))_{|\theta=\hat{\theta}^k})$  the expression

$$\begin{aligned} \mathbb{E}_{\theta_0}(\mathbb{E}_{\theta_0}(\ln L(\theta|Y))_{|\theta=\hat{\theta}^k}) &= \mathbb{E}_{\theta_0}(\ln L(\hat{\theta}^k|Y)) - \frac{N(N+k)}{2(N-k-2)} + \frac{N}{2} \\ &= \mathbb{E}_{\theta_0}(-N \ln(\hat{\sigma}_k \sqrt{2\pi})) - \frac{N(N+k)}{2(N-k-2)}. \end{aligned}$$

The first term is estimated by  $-N \ln(\hat{\sigma}_k \sqrt{2\pi}) = -N \ln(\sqrt{2\pi}) - \frac{N}{2} \ln(\hat{\sigma}_k^2)$ . Since there the first part is the same for all competing distributions, the

maximization of the expectation of the discrepancy (B.9) is equivalent to the maximization of

$$k \mapsto -\frac{N}{2} \ln(\hat{\sigma}_k^2) - \frac{N(N+k)}{2(N-k-2)}.$$

□

For the question how AICC improves the approximations for AIC, we refer to J.E. CAVANAUGH (1997).

### B.3 The Schwarz Information Criterion

Let  $y$  be data of length  $N$ . Let  $\{M_1, \dots, M_L\}$  denote the set of candidate models which are not necessarily nested. Assume that the model  $M_k$  can uniquely be parameterized by a parameter vector  $\theta^k$  in the parameter space  $\Theta^k$ . Let  $L(\hat{\theta}^k|Y)$  denote the likelihood function based on the model  $M_k$  with parameter  $\theta^k$  evaluated at the maximum likelihood estimator  $\hat{\theta}^k(y) = \operatorname{argmax}_{\theta^k \in \Theta^k} L(\theta^k|Y)$ . The number of parameters to be estimated in the model  $M_k$  will be called the dimension  $D_k$  of the  $k$ -th model. For the case of independent and identically distributed observations  $y_s, s = 1, \dots, N$ , and linear models and under the assumption that the likelihood functions belong to the regular exponential family G. SCHWARZ (1978) suggested the following model selection criterion: Choose that model for which

$$\text{SIC}(k) = \ln L(\hat{\theta}^k|Y) - \frac{1}{2} D_k \ln N \tag{B.12}$$

is maximal. The derivation of the Schwarz information criterion is based on a Bayesian approach. The expression (B.12) is an approximation of a transformation of the posterior probability of the  $k$ -th candidate model. The model which maximizes SIC in (B.12) should for large  $N$  correspond to the model with maximal posterior probability. Assume that we have

- (1) a prior probability  $\pi(M_k)$  of the  $k$ -th model being true,
- (2) a prior distribution  $g(\theta^k|M_k)$  for the parameter  $\theta^k$  given the  $k$ -th model,
- (3) a family of distributions given by the family of their likelihood functions  $L(\theta^k|Y)$ .

Denoting by  $h(Y)$  the marginal density of the data and using Bayes' formula the joint posterior density of the model  $M_k$  and the parameter  $\theta^k$  given data  $Y$  is

$$f((M_k, \theta^k)|Y) = \frac{\pi(M_k)g(\theta^k|M_k)L(\theta^k|Y)}{h(Y)}.$$

We get the distribution of the model  $M_k$  given data  $Y$  by integration of  $f((M_k, \theta^k)|Y)$  over  $\Theta^k$ ,

$$P(M_k|Y) = \frac{\pi(M_k)}{h(Y)} \int_{\Theta^k} L(\theta^k|Y)g(\theta^k|M_k) d\theta^k. \quad (\text{B.13})$$

The maximization of the posterior probability (B.13) is equivalent to the maximization of its transformation

$$\ln P(M_k|Y) + \ln h(Y) = \ln \pi(M_k) + \ln \left( \int_{\Theta^k} L(\theta^k|Y)g(\theta^k|M_k) d\theta^k \right). \quad (\text{B.14})$$

J.E. CAVANAUGH and A.A. NEATH (1999) show in a general setting that under certain regularity conditions (B.12) provides a large-sample approximation to  $\ln P(M_k|Y) + \ln h(Y)$ .

### A Variant of SIC

It is obvious that SIC has the tendency to choose more parsimonious models than AIC. Since the Schwarz information criterion is also based on asymptotic approximations, its performance can improved by adding second order terms in small sample settings. As in the case of the Akaike information criterion, corrections for the bias depend on the set of candidate models. We will derive a corrected version of SIC in the special setting of the family of log likelihood functions in (B.6) following the lines in J.E. CAVANAUGH and A.A. NEATH (1999). In the rest of this section we will always assume that  $\ln L(\theta^k|Y)$  from (B.6) fulfills the regularity conditions in Appendix C.2.

We will first consider the integral

$$\int_{\Theta^k} L(\theta^k|Y)g(\theta^k|M_k) d\theta^k \quad (\text{B.15})$$

which appears on the right hand side of (B.14).

**Lemma B.3.1** *An approximation to the integral (B.15) is given by*

$$\begin{aligned} & \int_{\Theta^k} L(\theta^k|Y)g(\theta^k|M_k) d\theta^k \\ & \approx L(\hat{\theta}^k|Y) \cdot \int_{\Theta^k} e^{\frac{1}{2}(\theta^k - \hat{\theta}^k)^t M(\ln L(\hat{\theta}^k|Y))(\theta^k - \hat{\theta}^k)} g(\theta^k|M_k) d\theta^k. \end{aligned}$$

**Proof** Let  $M(\ln L(\hat{\theta}^k|Y))$  be the Hessian matrix of  $\ln L(\theta^k|Y)$  evaluated at the maximum likelihood estimator  $\hat{\theta}^k$ . If  $\theta^k$  is close to  $\hat{\theta}^k$  we get

$$\ln L(\theta^k|Y) \approx \ln L(\hat{\theta}^k|Y) + \frac{1}{2}(\theta^k - \hat{\theta}^k)^t M(\ln L(\hat{\theta}^k|Y))(\theta^k - \hat{\theta}^k). \quad (\text{B.16})$$

Using this Taylor expansion the integral (B.15) can be written in the form

$$\begin{aligned} \int_{\Theta^k} L(\theta^k|Y)g(\theta^k|M_k) d\theta^k &= \int_{\Theta^k} e^{\ln L(\theta^k|Y)}g(\theta^k|M_k) d\theta^k \\ &\approx \int_{\Theta^k} e^{\ln L(\hat{\theta}^k|Y)+\frac{1}{2}(\theta^k-\hat{\theta}^k)^t M(\ln L(\hat{\theta}^k|Y))(\theta^k-\hat{\theta}^k)}g(\theta^k|M_k) d\theta^k \\ &= L(\hat{\theta}^k|Y) \cdot \int_{\Theta^k} e^{\frac{1}{2}(\theta^k-\hat{\theta}^k)^t M(\ln L(\hat{\theta}^k|Y))(\theta^k-\hat{\theta}^k)}g(\theta^k|M_k) d\theta^k \end{aligned}$$

which is the stated formula.  $\square$

The *observed Fisher information matrix* is given by

$$F_N(\hat{\theta}^k|Y) = -\frac{1}{N}M(\ln L(\hat{\theta}^k|Y)). \quad (\text{B.17})$$

Provided the regularity conditions from Appendix C.2 hold the maximum likelihood estimator  $\hat{\theta}^k$  converges almost surely to some  $\theta_*^k$  and  $F_N(\theta^k|Y)$  converges almost surely uniformly in  $\theta^k$  to a matrix which is positive definite in a neighborhood of  $\theta_*^k$ . As a consequence, for large  $N$  it is possible to find positive constants independent of  $N$  such that  $\det(F_N(\hat{\theta}^k|Y))$  is bounded in between, see J.E. CAVANAUGH and A.A. NEATH (1999). In the further derivation we set  $g(\theta^k|M_k) = 1$  which is interpreted as to use a non-informative prior.

**Lemma B.3.2** *With  $g(\theta^k|M_k) = 1$  we have for the integral (B.15)*

$$\begin{aligned} \int_{\Theta^k} L(\theta^k|Y)g(\theta^k|M_k) d\theta^k &= \int_{\Theta^k} L(\theta^k|Y) d\theta^k \\ &\approx L(\hat{\theta}^k|Y) \cdot N^{-D_k/2} \cdot (2\pi)^{D_k/2} \cdot \left(\det(F_N(\hat{\theta}^k|Y))\right)^{-1/2}. \end{aligned}$$

**Proof** Application of Lemma B.3.1 to (B.15) leads to

$$\int_{\Theta^k} L(\theta^k|Y) d\theta^k \approx L(\hat{\theta}^k|Y) \cdot \int_{\Theta^k} e^{\frac{1}{2}(\theta^k-\hat{\theta}^k)^t M(\ln L(\hat{\theta}^k|Y))(\theta^k-\hat{\theta}^k)} d\theta^k.$$

With  $F_N(\hat{\theta}^k|Y)$  defined in (B.17) we can rewrite the remaining integral as

$$\int_{\Theta^k} e^{\frac{1}{2}(\theta^k-\hat{\theta}^k)^t M(\ln L(\hat{\theta}^k|Y))(\theta^k-\hat{\theta}^k)} d\theta^k = \int_{\Theta^k} e^{-\frac{N}{2}(\theta^k-\hat{\theta}^k)^t F_N(\hat{\theta}^k|Y)(\theta^k-\hat{\theta}^k)} d\theta^k.$$

Since

$$\int_{\mathbb{R}^k} e^{-\frac{1}{2}x^t Ax} dx = \sqrt{\frac{(2\pi)^k}{\det A}}$$

we get

$$\int_{\Theta^k} L(\theta^k|Y) d\theta^k \approx L(\hat{\theta}^k|Y) \cdot N^{-D_k/2} \cdot \sqrt{\frac{(2\pi)^{D_k}}{\det(F_N(\hat{\theta}^k|Y))}}.$$

□

The considerations above are summarized in the following theorem.

**Theorem B.3.3** *Let data  $y \in \mathbb{X}$  be given. Assume that  $g(\theta^k|M_k) = 1$ . Then we get the following approximation for (B.14),*

$$\begin{aligned} \ln P(M_k|Y) + \ln h(Y) & \quad (B.18) \\ \approx \ln \pi(M_k) + \ln L(\hat{\theta}^k|Y) - \frac{D_k}{2} \ln N + \frac{D_k}{2} \ln(2\pi) - \frac{1}{2} \ln \left( \det(F_N(\hat{\theta}^k|Y)) \right). \end{aligned}$$

*Ignoring terms which are bounded as  $N$  tends to infinity, the maximization of the posterior probability (B.13) leads to the Schwarz information criterion:*

$$\text{maximize } k \longmapsto SIC(k) = \ln L(\hat{\theta}^k|Y) - \frac{D_k}{2} \ln N. \quad (B.19)$$

**Proof** The approximation (B.18) follows from Lemma B.3.1 and Lemma B.3.2. As  $N$  tends to infinity the term  $D_k/2 \cdot \ln(2\pi)$  is obviously bounded. It is shown in J.E. CAVANAUGH and A.A. NEATH (1999) that the term containing the determinant of  $F_N(\hat{\theta}^k|Y)$  is asymptotically bounded due to the assumed regularity conditions. □

We specialize now to the model (B.1). In Appendix C.2 we show that the regularity conditions are fulfilled in case of the family of likelihood functions  $L(\theta^k|Y)$  from (B.6). Here is  $\Theta^k = \mathbb{R}^{D_k}$ .

We will derive a variant of the Schwarz information criterion (B.19) for the family of likelihood functions in (B.6) in the case when the variance  $\sigma^2$  has to be estimated. The parameter vector  $\theta^k = (\mu^k, \sigma_k)$  is an element of  $\Theta = \mathbb{R}^{k+1}$ . First, we compute the Hessian matrix  $M(\ln L(\theta^k|Y))$  of  $\ln L(\theta^k|Y)$ .

**Proposition B.3.4** *Let  $\hat{\theta}^k = (\hat{\mu}^k, \hat{\sigma}_k)$  be the maximum likelihood estimator of  $\ln L(\theta^k|Y)$  from (B.6) and let  $\mathcal{P} = \{I_1, \dots, I_k\}$  be the partition induced by  $\hat{\mu}^k$ . Evaluation of the Hessian matrix  $M$  of  $\ln L(\theta^k|Y)$  at the maximum likelihood estimator  $\hat{\theta}^k$  gives*

$$M(\ln L(\hat{\theta}^k|Y)) = -\frac{1}{\hat{\sigma}_k^2} \begin{pmatrix} |I_1| & & & \\ & \ddots & & 0 \\ & & |I_k| & \\ 0 & & & 2N \end{pmatrix} \quad (B.20)$$

**Proof** The proof is in Appendix C.3.  $\square$

The derivatives of the log likelihood function of higher order do not vanish in the case of estimated variance. Hence, the Taylor expansion (B.16) of  $\ln L(\theta^k|Y)$  around the maximum likelihood estimator  $\hat{\theta}^k = (\hat{\mu}^k, \hat{\sigma}_k^2)$  is only an approximation. From Theorem B.3.3 we can derive the following variant of the Schwarz information criterion presents a corrected version of  $\text{SIC}(k)$  from (B.19) in case of smaller length  $N$  of data  $y$ .

**Corollary B.3.5** *Let data  $y \in \mathbb{X}$  be given. We consider the model (B.1) and the family of log likelihood functions  $\ln L((\mu^k, \sigma_k)|Y)$  from (B.6) with the maximum likelihood estimators  $\hat{\mu}^k$  from (B.4) and  $\hat{\sigma}_k^2$  from (B.5). Let  $\mathcal{P}$  be the partition induced by  $\hat{\mu}^k$ . Suppose further that  $g(\theta^k|Y) = 1$  and that the prior  $\pi(M_k)$  is the same for all models. Then the maximization of the posterior probability (B.13) over all candidate models  $M_k$  is approximately equivalent to the maximization of*

$$\text{SICC}(k) = \frac{k+1-N}{2} \ln(2\pi\hat{\sigma}_k^2) - \frac{1}{2} \sum_{I \in \mathcal{P}} \ln |I|. \quad (\text{B.21})$$

for  $k = 1, \dots, L$ .

**Proof** By Theorem B.3.3 we get for the transformation (B.14) of the posterior distribution

$$\begin{aligned} \ln P(M_k|Y) + \ln h(Y) &= \ln \pi(M_k) \\ &\quad + \ln L(\hat{\theta}^k|Y) - \frac{k+1}{2} \ln N \\ &\quad + \frac{k+1}{2} \ln(2\pi) \\ &\quad - \frac{1}{2} \ln \left( \det(F_N(\hat{\theta}^k|Y)) \right). \end{aligned} \quad (\text{B.22})$$

From the computations in Appendix C.3 we get the observed Fisher information matrix  $F_N(\hat{\theta}^k|Y)$  as

$$F_N((\hat{\mu}^k, \hat{\sigma}_k)|Y) = \frac{1}{N\hat{\sigma}_k^2} \begin{pmatrix} |I_1| & & & \\ & \ddots & & 0 \\ & & |I_k| & \\ & 0 & & 2 \end{pmatrix}.$$

with determinant

$$\det(F_N(\hat{\theta}^k|Y)) = \left( \frac{1}{N\hat{\sigma}_k^2} \right)^{k+1} \cdot 2 \cdot \prod_{I \in \mathcal{P}} |I|.$$

Hence we get for (B.22)

$$-\frac{1}{2} \ln \left( \det(F_N(\hat{\theta}^k|Y)) \right) = \frac{k+1}{2} \ln N + \frac{k+1}{2} \ln(\hat{\sigma}_k^2) - \frac{1}{2} \ln 2 - \frac{1}{2} \sum_{I \in \mathcal{P}} \ln |I|.$$

The log likelihood function evaluated at the maximum likelihood estimator gives

$$\begin{aligned} \ln L(\hat{\theta}^k|Y) &= -\frac{N}{2} \ln(2\pi\hat{\sigma}_k^2) - \frac{1}{2\hat{\sigma}_k^2} \sum_{I \in \mathcal{P}} \sum_{s \in I} (y_s - \hat{\mu}_I)^2 \\ &= -\frac{N}{2} \ln(2\pi\hat{\sigma}_k^2) - \frac{N}{2} \end{aligned}$$

Together we get for the transformation (B.14) of the posterior distribution

$$\begin{aligned} \ln P(M_k|Y) + \ln h(Y) &\approx \ln \pi(M_k) - \frac{N}{2} \ln(2\pi\hat{\sigma}_k^2) - \frac{N}{2} - \frac{k+1}{2} \ln N \\ &\quad + \frac{k+1}{2} \ln(2\pi) \\ &\quad + \frac{k+1}{2} \ln N + \frac{k+1}{2} \ln(\hat{\sigma}_k^2) - \frac{1}{2} \ln 2 - \frac{1}{2} \sum_{I \in \mathcal{P}} \ln |I| \\ &= \frac{k+1-N}{2} \ln(2\pi\hat{\sigma}_k^2) - \frac{1}{2} \sum_{I \in \mathcal{P}} \ln |I| \end{aligned} \quad (\text{B.23})$$

$$+ \ln \pi(M_k) - \frac{N + \ln(2N)}{2}. \quad (\text{B.24})$$

Ignoring the terms in (B.24) which do not depend on  $k$  and assuming the same prior for all models we arrive at the following corrected version of the Schwarz information criterion: If we want to maximize the posterior distribution we have to choose that number  $k$  of intervals in the partition  $\mathcal{P}$  which maximizes (B.23).  $\square$

# Appendix C

## Calculations for the Model Selection Criteria

In this chapter, we collect proofs and calculations for the model selection criteria from Chapter B.

### C.1 Proofs Concerning the Akaike Information Criterion

In this section we prove the two Lemmata from Section B.2 used in the proof of Theorem B.2.4.

**Proof of Lemma B.2.2** With the maximum likelihood estimators  $\hat{\mu}^k$  and  $\hat{\sigma}_k^2$  for  $\theta_0 = (\mu_0, \sigma_0^2)$  the single parts in part (B.10) give the following. We have

$$\begin{aligned}\mathbb{E}_{\theta_0}(\ln L(\theta_0|Y)) &= \mathbb{E}_{\theta_0}\left(-N \ln(\sigma_0 \sqrt{2\pi}) - \frac{1}{2\sigma_0^2} \|y - \mu_0\|^2\right) \\ &= -N \ln(\sigma_0 \sqrt{2\pi}) - \frac{N}{2}\end{aligned}$$

and

$$\begin{aligned}\mathbb{E}_{\theta_0}(\ln L(\hat{\theta}|Y)) &= \mathbb{E}_{\theta_0}\left(-N \ln(\hat{\sigma} \sqrt{2\pi}) - \frac{1}{2\hat{\sigma}^2} \|y - \hat{\mu}\|^2\right) \\ &= \mathbb{E}_{\theta_0}\left(-N \ln(\hat{\sigma} \sqrt{2\pi})\right) - \frac{N}{2}.\end{aligned}$$

The difference (B.10) is then equal to

$$\mathbb{E}_{\theta_0}(\ln L(\theta_0|Y)) - \mathbb{E}_{\theta_0}(\ln L(\hat{\theta}|Y)) = -N \ln(\sigma_0 \sqrt{2\pi}) + \mathbb{E}_{\theta_0}(N \ln(\hat{\sigma} \sqrt{2\pi})). \quad \square$$

**Proof of Lemma B.2.3** We have

$$\begin{aligned}\mathbb{E}_{\theta_0}(\ln L(\theta|Y)) &= \mathbb{E}_{\theta_0}\left(-N \ln(\sigma\sqrt{2\pi}) - \frac{1}{2\sigma^2}\|y - \mu\|^2\right) \\ &= -N \ln(\sigma\sqrt{2\pi}) - \frac{N}{2} \cdot \frac{\sigma_0^2}{\sigma^2} - \frac{1}{2\sigma^2}\|\mu_0 - \mu\|^2.\end{aligned}$$

Evaluation at the maximum likelihood estimator  $\hat{\theta}^k = (\hat{\mu}^k, \hat{\sigma}_k^2)$  results in

$$\mathbb{E}_{\theta_0}(\ln L(\theta|Y))_{|\theta=\hat{\theta}^k} = -N \ln(\hat{\sigma}_k\sqrt{2\pi}) - \frac{N}{2} \cdot \frac{\sigma_0^2}{\hat{\sigma}_k^2} - \frac{1}{2\hat{\sigma}_k^2}\|\mu_0 - \hat{\mu}^k\|^2.$$

Taking the expectation above gives

$$\begin{aligned}\mathbb{E}_{\theta_0}(\mathbb{E}_{\theta_0}(\ln L(\theta|Y))_{|\theta=\hat{\theta}^k}) & \tag{C.1} \\ &= \mathbb{E}_{\theta_0}\left(-N \ln(\hat{\sigma}_k\sqrt{2\pi})\right) - \frac{N^2}{2}\mathbb{E}_{\theta_0}\left(\frac{\sigma_0^2}{\|y - P_k y\|^2}\right) - \frac{N}{2}\mathbb{E}_{\theta_0}\left(\frac{\|\mu_0 - \hat{\mu}^k\|^2}{\|y - P_k y\|^2}\right).\end{aligned}$$

The random variable  $\|y - P_k y\|^2/\sigma_0^2$  is  $\chi^2$ -distributed with  $N - k$  degrees of freedom, see for example in H. WITTING (1985). The expected value of the inverse of a  $\chi_{\text{df}}^2$ -distributed random variable is equal to  $1/(\text{df} - 2)$ . Hence

$$\frac{N^2}{2}\mathbb{E}_{\theta_0}\left(\frac{\sigma_0^2}{\|y - P_k y\|^2}\right) = \frac{N^2}{2} \frac{1}{N - k - 2}.$$

The third term in (C.1) can be written as

$$\begin{aligned}\frac{\|\mu_0 - \hat{\mu}^k\|^2}{\|y - P_k y\|^2} &= \frac{\frac{1}{\sigma_0^2}\|\mu_0 - \hat{\mu}^k\|^2}{\frac{1}{\sigma_0^2}\|y - P_k y\|^2} \\ &= \sum_{l=1}^k \sum_{s \in I_l} \frac{1}{|I_l|} \frac{(\sqrt{|I_l|} \frac{1}{\sigma_0^2}(\mu_{0s} - \bar{y}_{I_l}))^2}{\frac{1}{\sigma_0^2}\|y - P_k y\|^2}.\end{aligned}$$

The numerator is standard normal distributed and the denominator is  $\chi^2$ -distributed with  $(N - k)$  degrees of freedom. Then the ratio

$$\frac{(\sqrt{|I_l|} \frac{1}{\sigma_0^2}(\mu_{0s} - \bar{y}_{I_l}))^2}{\frac{1}{N-k} \frac{1}{\sigma_0^2} \|y - P_k y\|^2}$$

is  $F(1, N - k)$ -distributed with expected value  $\frac{N-k}{N-k-2}$ . Together we have

$$\mathbb{E}_{\theta_0}\left(\frac{1}{2\hat{\sigma}_k^2}\|\mu_0 - \hat{\mu}^k\|^2\right) = \frac{N}{2} \sum_{l=1}^k \sum_{s \in I_l} \frac{1}{|I_l|} \cdot \frac{1}{N - k} \cdot \frac{N - k}{N - k - 2}$$

$$\begin{aligned}
&= \frac{N}{2} \cdot \frac{1}{N-k-2} \sum_{l=1}^k |I_l| \frac{1}{|I_l|} \\
&= \frac{Nk}{2(N-k-2)}.
\end{aligned}$$

For (C.1) we then get

$$\begin{aligned}
&\mathbb{E}_{\theta_0}(\mathbb{E}_{\theta_0}(\ln L(\theta|Y))_{|\theta=\hat{\theta}^k}) \\
&= \mathbb{E}_{\theta_0}(-N \ln(\hat{\sigma}_k \sqrt{2\pi})) - \frac{N^2}{2(N-k-2)} - \frac{Nk}{2(N-k-2)} \\
&= \mathbb{E}_{\theta_0}(-N \ln(\hat{\sigma}_k \sqrt{2\pi})) - \frac{N(N+k)}{2(N-k-2)}.
\end{aligned}$$

The difference (B.11) is then

$$\begin{aligned}
&\mathbb{E}_{\theta_0}(\mathbb{E}_{\theta_0}(\ln L(\theta|Y))_{|\theta=\hat{\theta}^k}) - \mathbb{E}_{\theta_0}(\ln L(\theta_0|Y)) \\
&= \mathbb{E}_{\theta_0}(-N \ln(\hat{\sigma}_k \sqrt{2\pi})) - \frac{N(N+k)}{2(N-k-2)} + N \ln(\sigma_0 \sqrt{2\pi}) + \frac{N}{2}.
\end{aligned}$$

□

## C.2 Regularity Conditions on the Likelihood Function

In this section, we check the regularity conditions, mentioned in Section B.3, for the derivation of the Schwarz Information Criterion. Let data  $y$  be a sample from the model (B.1) with parameter  $\theta^0 = ((\mu_J)_{J \in \mathcal{P}^0}, \sigma_0)$ . We will show that the likelihood function  $L(\theta^k|Y_N)$  from (B.6) of the  $k$ -th model fulfills the regularity conditions from J.E. CAVANAUGH and A.A. NEATH (1999). We will use their notation and define

$$V_N(\theta^k) = -\frac{1}{N} L(\theta^k|Y_N).$$

- (1)  $V_N(\theta^k)$  has first- and second order derivatives which are continuous over  $\Theta^k$  since  $L(\theta^k|Y_N)$  has.
- (2)  $V_N(\theta^k)$  has a unique global minimum at  $\hat{\theta}^k = (\hat{\mu}^k, \hat{\sigma}_k)$  with  $\hat{\mu}^k$  from (B.4) and  $\hat{\sigma}_k$  from (B.5).

- (3) We compute the expectation of  $V_N(\theta^k)$  with respect to the distribution with parameter  $\theta^0$  and define

$$W_N(\theta) = \mathbb{E}_{\theta^0}(V_N(\theta^k)) = \mathbb{E}_{\theta^0}\left(-\frac{1}{N} \ln L(\theta^k | Y_N)\right).$$

We decompose the data term

$$\begin{aligned} \sum_{I \in \mathcal{P}} \sum_{s \in I} (y_s - \mu_I^k)^2 &= \sum_{I \in \mathcal{P}} \sum_{s \in I} (y_s - \mu_{J(s)}^0 + \mu_{J(s)}^0 - \mu_I^k)^2 \\ &= \sum_{J \in \mathcal{P}^0} \sum_{s \in J} (y_s - \mu_J^0)^2 + \sum_{I \in \mathcal{P}} \sum_{J \in \mathcal{P}^0} |I \cap J| (\mu_J^0 - \mu_I^k)^2 \\ &\quad + 2 \sum_{I \in \mathcal{P}} \sum_{s \in I} (y_s - \mu_{J(s)}^0) (\mu_{J(s)}^0 - \mu_I^k) \end{aligned} \quad (\text{C.2})$$

and get

$$\begin{aligned} W_N(\theta^k) &= \frac{1}{2} \ln(2\pi\sigma_k^2) + \frac{1}{2N\sigma_k^2} \mathbb{E}_{\theta^0} \left( \sum_{I \in \mathcal{P}} \sum_{s \in I} (y_s - \mu_I^k)^2 \right) \\ &= \frac{1}{2} \ln(2\pi\sigma_k^2) + \frac{\sigma_0^2}{2\sigma_k^2} + \frac{1}{2N\sigma_k^2} \sum_{I \in \mathcal{P}} \sum_{J \in \mathcal{P}^0} |I \cap J| (\mu_J^0 - \mu_I^k)^2. \end{aligned}$$

An increase of  $N$  in the frame of the Schwarz information criterion is interpreted as a refinement of the sampling of a function which is defined on a real interval of  $\mathbb{R}$ . That is, the number  $|I|$  of sample points in the discrete interval  $I$  depends on  $N$ . The ratio  $|I \cap J|/N$  represents the fraction of sample points in  $I \cap J$ . As  $N$  tends to infinity this fraction tends to the length of the continuous interval  $I \cap J$ , also denoted by  $|I \cap J|$ . Thus, we have for  $N \rightarrow \infty$

$$W_N(\theta^k) \rightarrow W(\theta^k) := \frac{\ln(2\pi\sigma_k^2)}{2} + \frac{\sigma_0^2}{2\sigma_k^2} + \frac{1}{2\sigma_k^2} \sum_{I \in \mathcal{P}} \sum_{J \in \mathcal{P}^0} |I \cap J| (\mu_J^0 - \mu_I^k)^2$$

uniformly in  $\theta^k$ .  $W(\theta^k)$  has first- and second order derivatives which are continuous over  $\Theta^k$ . Using the calculations in the proof of Proposition B.3.4 we get

$$\begin{aligned} \frac{\partial}{\partial \mu_I} W(\theta^k) &= \frac{1}{\sigma_k^2} \left( \sum_{J \in \mathcal{P}^0} |I \cap J| \mu_J^0 - |I| \mu_I \right) \\ \frac{\partial^2}{\partial \mu_I \partial \mu_{I'}} W(\theta^k) &= \begin{cases} 0 & I \neq I' \\ \frac{|I|}{\sigma_k^2} & I = I' \end{cases} \end{aligned}$$

$$\begin{aligned} \frac{\partial}{\partial \sigma_k} W(\theta^k) &= \frac{1}{\sigma_k} - \frac{\sigma_0^2}{\sigma_k^3} - \frac{1}{\sigma_k^3} \sum_{I \in \mathcal{P}} \sum_{J \in \mathcal{P}^0} |I \cap J| (\mu_J^0 - \mu_I)^2 \\ \frac{\partial^2}{\partial \sigma_k^2} W(\theta^k) &= -\frac{1}{\sigma_k^2} + \frac{3\sigma_0^2}{\sigma_k^4} + \frac{3}{\sigma_k^4} \sum_{I \in \mathcal{P}} \sum_{J \in \mathcal{P}^0} |I \cap J| (\mu_J^0 - \mu_I)^2 \\ \frac{\partial^2}{\partial \mu_I \partial \sigma_k} W(\theta^k) &= \frac{2}{\sigma_k^2} \left( \sum_{J \in \mathcal{P}^0} |I \cap J| \mu_J^0 - |I| \mu_I \right) \end{aligned}$$

- (4) Necessary conditions for  $W(\theta^k)$  to have a minimum at  $\theta_*^k = ((\mu_I^*)_{I \in \mathcal{P}}, \sigma_k^*)$  are

$$\begin{aligned} \frac{\partial}{\partial \mu_I} W(\theta^k) = 0 &\Leftrightarrow \mu_I^* = \frac{1}{|I|} \sum_{J \in \mathcal{P}^0} |I \cap J| \mu_J^0 \\ \frac{\partial}{\partial \sigma_k} W(\theta^k) = 0 &\Leftrightarrow (\sigma_k^*)^2 = \sigma_0^2 + \sum_{I \in \mathcal{P}} \sum_{J \in \mathcal{P}^0} |I \cap J| (\mu_J^0 - \mu_I^*)^2 \end{aligned}$$

It is indeed a minimum since

$$\begin{aligned} &\frac{\partial^2}{\partial \sigma_k^2} W(\theta^k) |_{\theta^k = \theta_*^k} \\ &= -\frac{1}{(\sigma_k^*)^4} \left( (\sigma_k^*)^2 - 3(\sigma_0^2 + \sum_{I \in \mathcal{P}} \sum_{J \in \mathcal{P}^0} |I \cap J| (\mu_J^0 - \mu_I^*)^2) \right) = \frac{2}{(\sigma_k^*)^2} \\ &\frac{\partial^2}{\partial \mu_I \partial \mu_{I'}} W(\theta^k) |_{\theta^k = \theta_*^k} = \begin{cases} 0 & I \neq I' \\ \frac{|I|}{(\sigma_k^*)^2} & I = I' \end{cases} \\ &\frac{\partial^2}{\partial \mu_I \partial \sigma_k} W(\theta^k) |_{\theta^k = \theta_*^k} = \frac{2}{\sigma_k^2} \left( \sum_{J \in \mathcal{P}^0} |I \cap J| \mu_J^0 - |I| \mu_I^* \right) = 0. \end{aligned}$$

- (5) With the decomposition from (C.2) we can write  $V_N(\theta^k)$  as

$$\begin{aligned} V_N(\theta^k) &= -1/N \ln L(\theta^k | Y_N) \\ &= \frac{1}{N} \sum_{J \in \mathcal{P}^0} \sum_{s \in J} (y_s - \mu_J^0)^2 \end{aligned} \tag{C.3}$$

$$+ 2 \cdot \frac{1}{N} \sum_{I \in \mathcal{P}} \sum_{s \in I} (y_s - \mu_{J(s)}^0) (\mu_{J(s)}^0 - \mu_I^k) \tag{C.4}$$

$$+ \frac{1}{N} \sum_{I \in \mathcal{P}} \sum_{J \in \mathcal{P}^0} |I \cap J| (\mu_J^0 - \mu_I^k)^2. \tag{C.5}$$

We consider the single parts. For part (C.3) we have

$$\frac{1}{N} \sum_{J \in \mathcal{P}^0} \sum_{s \in J} (y_s - \mu_J^0)^2 \longrightarrow \mathbb{E}_{\theta^0}((y_s - \mu_J^0)^2) = \sigma_0^2 \quad \theta^0\text{-almost surely.}$$

The second part (C.4) tends to zero since

$$\frac{1}{N} \sum_{I \in \mathcal{P}} \sum_{s \in I} (y_s - \mu_{J(s)}^0) \longrightarrow \mathbb{E}_{\theta^0}(y_s - \mu_{J(s)}^0) = 0 \quad \theta^0\text{-almost surely.}$$

In the last part (C.5) there are no random variables included and we have convergence of the discrete interval length to the continuous one. Thus have

$$-\frac{1}{N} \ln L(\theta^k | Y_N) \longrightarrow W(\theta^k) \quad \theta^0\text{-almost surely}$$

as  $N \rightarrow \infty$  and uniformly in  $\theta^k$ .

- (6) The uniform almost sure convergence of the second derivatives of  $V_N(\theta^k)$  to those of  $W(\theta^k)$  follows from the calculations in Section C.3.
- (7) The Hessian matrix of  $W(\theta^k)$  is positive definite in a neighborhood of  $\theta_*^k$ ,

$$\begin{aligned} \frac{\partial^2}{\partial \sigma_k^2} W(\theta^k) > 0 &\Leftrightarrow \sigma_k^2 < 3(\sigma_0^2 + \sum_{I \in \mathcal{P}} \sum_{J \in \mathcal{P}^0} |I \cap J| (\mu_J^0 - \mu_I)^2) \\ &\Leftrightarrow \sigma_k^2 < 3(\sigma_k^*)^2. \end{aligned}$$

In this neighborhood, its eigenvalues are bounded and bounded away from zero.

### C.3 The Hessian Matrix of the Log Likelihood Function

In this section, we will compute the Hessian matrix of the log likelihood function in question. We will compute the second derivatives of  $\ln L(\theta^k | Y)$  from (B.6). In case of  $\theta^k = \mu^k = (\mu_I^k)_{I \in \mathcal{P}}$  we have

$$\frac{\partial}{\partial \mu_I} \ln L(\theta^k | Y) = \frac{1}{\sigma^2} \left( \sum_{s \in I} y_s - |I| \mu_I \right)$$

$$\frac{\partial^2}{\partial \mu_I \mu_J} \ln L(\theta^k | Y) = \begin{cases} 0 & I \neq J \\ -\frac{|I|}{\sigma^2} & I = J \end{cases}$$

For  $\theta^k = (\mu^k, \sigma_k)$  we replace there  $\sigma$  by  $\sigma_k$  and get in addition

$$\begin{aligned} \frac{\partial}{\partial \sigma_k} \ln L(\theta^k | Y) &= -\frac{N}{2} \frac{1}{2\pi\sigma_k^2} \cdot 2\pi \cdot 2\sigma_k - \left(\frac{-3}{\sigma_k^3}\right) \cdot \frac{1}{2} \sum_{I \in \mathcal{P}} \sum_{s \in I} (y_s - \mu_I)^2 \\ &= -\frac{N}{\sigma_k} + \frac{1}{\sigma_k^3} \sum_{I \in \mathcal{P}} \sum_{s \in I} (y_s - \mu_I)^2 \\ \frac{\partial^2}{\partial \mu_I \sigma_k} \ln L(\theta^k | Y) &= -\frac{2}{\sigma_k^3} \left( \sum_{s \in I} y_s - |I| \mu_I \right) \\ \frac{\partial^2}{\partial \sigma_k^2} \ln L(\theta^k | Y) &= \frac{N}{\sigma_k^2} + \left(\frac{-3}{\sigma_k^4}\right) \sum_{I \in \mathcal{P}} \sum_{s \in I} (y_s - \mu_I)^2 \\ &= \frac{1}{\sigma_k^2} \left( N - \frac{3}{\sigma_k^2} \sum_{I \in \mathcal{P}} \sum_{s \in I} (y_s - \mu_I)^2 \right) \end{aligned}$$

Evaluation at the maximum likelihood estimator  $\hat{\theta}^k = \hat{\mu}^k$  from (B.4) gives

$$\frac{\partial^2}{\partial \mu_I \mu_J} \ln L(\mu^k | Y) |_{\mu^k = \hat{\mu}^k} = \begin{cases} 0 & I \neq J, \\ -\frac{|I|}{\sigma^2} & I = J, \end{cases}$$

respectively, for  $\hat{\theta}^k = (\hat{\mu}^k, \hat{\sigma}_k)$  with  $\hat{\mu}^k$  from (B.4) and  $\hat{\sigma}_k$  from (B.5)

$$\begin{aligned} \frac{\partial^2}{\partial \mu_I \mu_J} \ln L(\theta^k | Y) |_{\theta^k = (\hat{\mu}^k, \hat{\sigma}_k)} &= \begin{cases} 0 & I \neq J \\ -\frac{|I|}{\hat{\sigma}_k^2} & I = J \end{cases} \\ \frac{\partial^2}{\partial \mu_I \sigma_k} \ln L(\theta^k | Y) |_{\theta^k = (\hat{\mu}^k, \hat{\sigma}_k)} &= -\frac{2}{\hat{\sigma}_k^3} \left( \sum_{s \in I} y_s - |I| \hat{\mu}_I^k \right) \\ &= -\frac{2}{\hat{\sigma}_k^3} \cdot 0 = 0 \\ \frac{\partial^2}{\partial \sigma_k^2} \ln L(\theta^k | Y) |_{\theta^k = (\hat{\mu}^k, \hat{\sigma}_k)} &= \frac{1}{\hat{\sigma}_k^2} \left( N - \frac{3}{\hat{\sigma}_k^2} \sum_{I \in \mathcal{P}} \sum_{s \in I} (y_s - \hat{\mu}_I^k)^2 \right) \\ &= \frac{1}{\hat{\sigma}_k^2} \left( N - \frac{3}{\hat{\sigma}_k^2} \cdot N \hat{\sigma}_k^2 \right) = -\frac{2N}{\hat{\sigma}_k^2}. \end{aligned}$$



# Symbols

$S$	set of sites	9
$s \sim t$	neighbors	10
$\mathbb{X}$	space of families $(x_s)_{s \in S}$ , $x_s \in \mathbb{R}$	10
$J(x)$	jump set of $x \in \mathbb{X}$	10
$ A $	number of elements of $A$	10
$ J(x) $	number of jumps of $x$	10
$\gamma$	hyperparameter	10
$H_\gamma$	Potts functional on $\mathbb{X} \text{ times } \mathbb{X}$	10
$X^*(\gamma, y)$	set of minimizers of the Potts functional	11
$x^*(\gamma, y)$	minimizer of the Potts functional	11
$\mathcal{P}$	partition of $S$	12
$\mathfrak{P}$	set of all partitions of $S$	12
$I \sim J$	neighboring intervals	12
$(\mathcal{P}, \mu_{\mathcal{P}})$	segmentation	12
$\mathfrak{S}$	set of segmentations	12
$\mathfrak{M}$	space of minimal segmentations	12
$\mathcal{P}(x)$	partition of $S$ induced by $x$	12
$\mu_I(x)$	constant value of $x$ on $I \in \mathcal{P}(x)$	12
$\mathbf{1}$	$(1, \dots, 1) \in \mathbb{R}^S$	15
$\mathbb{E}_{\mathbb{P}}(X)$	expectation of $X$ under the probability measure $\mathbb{P}$	16
$\mu_I^*$	minimizer of $\mu \mapsto \sum_{s \in I} \rho(y_s - \mu)$	20
$H^*(\cdot, y)$	functional on $\mathfrak{P}$	20
$\bar{H}_\gamma(\cdot, y)$	Potts functional on $\mathbb{X}$ with sum of squares	25
$\bar{y}_I$	empirical mean of the $y_s$ , $s \in I$	25
$\nabla$	gradient	27
$\mathfrak{P}_k$	set of all partitions $\mathcal{P} \in \mathfrak{P}$ with $ \mathcal{P}  = k$	28
$N^c$	complement of $N$	31
$\bar{y}$	empirical mean of the $y_s$ , $s \in S$	33
$\langle y, x \rangle$	Euclidian scalar product $\sum_{s \in S} y_s x_s$	36
$\ y - x\ ^2$	Euclidian norm on $\mathbb{R}^S$	36
$\mathcal{B}(X)$	$\sigma$ -field of the Borel subsets of $X$	40
$\mathfrak{P}^n$	set of all partitions of $\{1, \dots, n\}$	42

$H_I$	interval function	42
$[s, t]$	interval $\{s, s + 1, \dots, t - 1, t\}$	42
$O$	Landau symbol	44
$D_I$	interval error function	45
$\mathfrak{P}_k^n$	set of all partitions of $\{1, \dots, n\}$ with $ \mathcal{P}  = k$	46
$\mathbb{R}^*$	$\mathbb{R} \setminus \{0\}$	54
$\text{Aff}(\mathbb{R})$	group of the affine linear transformations of $\mathbb{R}$	54
$t_{b,c}$	scale transformation $x \mapsto c \cdot x + b\mathbf{1}$	54
$\mathfrak{p}(A)$	power set of $A$	55
$\Gamma$	data adapted parameter choice	57
$N_A$	affine normalization	59
$\mathcal{G}(y)$	set of finite positive $\gamma$ -values in the scanning	63
$\xi_{n,s}$	random variables on probability space $(\Omega, \mathcal{F}, \mathbb{P})$	129
$\lambda$	Lebesgue measure on the Borel- $\sigma$ -field $\mathcal{B}([0, 1])$	133
$\Delta_n f$	discretization of $f$	133
$\mathcal{N}(\mu, \sigma^2)$	normal distribution with mean $\mu$ and variance $\sigma^2$	134
$\bar{H}_\gamma^n(\cdot, y)$	Potts functional on $\mathbb{R}^{S_n}$ with sum of squares	134
$\mathcal{T}([0, 1])$	space of right-continuous step functions on $[0, 1]$	134
$j^n$	embedding of $x \in \mathbb{R}^{S_n}$ into $\mathcal{T}([0, 1])$	135
$\mathcal{J}(\tau)$	jump set (for an $L^2$ -equivalence class) of a right-continuous step function $\tau$	135
$\mathfrak{X}^n$	image of $\mathbb{R}^{S_n}$ under embedding $j^n$	135
$\mathfrak{B}^n$	$\sigma$ -field induced by $\{[\frac{s-1}{n}, \frac{s}{n}) : s \in S_n\}$	135
$\tilde{H}_\gamma^n(\cdot, f)$	continuous Potts functional on $L^2([0, 1])$	136
$P_{\mathfrak{X}^n}$	orthogonal projection of $L^2([0, 1])$ onto $\mathfrak{X}^n$	136
$\xi^n$	embedding $j^n((\xi_{n,1}, \dots, \xi_{n,n}))$	138
$H_0^\infty(g, f)$	$L^2$ -norm of $f - g$	138
$\tilde{H}_0^\infty(\cdot, f)$	functional on $L^2([0, 1])$ equal to $H_0^\infty(\cdot, f) + \sigma^2$	140
$\mathcal{B}_J$	$\sigma$ -field $\sigma(\{[a, b) : a, b \in J \cup \{0, 1\}\})$ for $J \subset (0, 1)$	141
$\mathcal{P}_J$	partition of $[0, 1]$ induced by the set $J$	141
$\mathcal{P}_J^n$	partition of $S_n$ induced by the partition $\mathcal{P}_J$	141
$\mathfrak{X}_J^n$	subspace of $g \in \mathfrak{X}^n$ with $\mathcal{J}(g) = J$	142
$d_{\mathcal{H}}$	Hausdorff metric	146
$L(\theta Y)$	likelihood function	167
$M(\ln L(\hat{\theta}^k Y))$	Hessian matrix of $\ln L(\theta^k Y)$ evaluated at the maximum likelihood estimator	172

# Index

- $\Gamma$ -convergence, 139
- $\alpha$ -homogeneous penalty, 58
- $\gamma$ -interval, 31
- $\gamma$ -scanning, 31
- affine linear group, 54
- AIC, 86, 168
- AICC, 170
- Akaike information criterion, 86, 168
- Bayes estimator, 16
- Bayes risk, 16
- BOLD effect, 97
- complexity of an algorithm, 44
- continuous Potts functional, 136
- data, 10
- data adapted parameter choice, 57
- data term, 10
- diameter, 60
- discrepancy, 169
- discretization, 133
- embedding, 135
- epi-convergence, 139
- equivariant, 53
- exponential form of a probability measure, 15
- F-longest interval criterion, 65
- family approach, 34
- FLIC, 65
  - estimator, 65
- fMRI, 97
- fractionation curves, 104
- hyperparameter, 10
- induced partition, 12
- induced segmentation, 12
- intensities, 10
- interval, 12
- interval error function, 45
- interval function, 42
- invariant attribute, 64
- jump, 10
- jump set, 10
- jump set for right-continuous step functions, 135
- Kullback-Leibler information, 168
- Landau symbol, 44
- length of an interval, 12
- LIC, 73
- likelihood function, 166
- log likelihood function, 167
- longest interval criterion
  - estimator, 73
- longest run condition, 81
- loss function, 16
- lower semicontinuous, 145
- MAP, 11
  - estimator, 11, 16
- maximum likelihood estimator, 166
- maximum posterior estimator, 11, 16

- measurable section, 40
- minimal segmentation, 12
- minimum function, 29
- multiresolution condition, 83
  
- nearest neighbor, 13
- neighbor
  - sites, 10
  - structure, 10
- neighboring intervals, 12
- neighbors, 10
  
- observations, 10
- observed Fisher information matrix, 173
- orbit, 59
- orthogonal projection, 136
  
- partition, 12
- posterior distribution, 16
- Potts
  - functional, 10
  - one dimensional, 13
  - penalty term, 10
- Potts model, 2
- prior
  - distribution, 15
  - improper, 17
  
- real character, 69
- residuals, 81
- right-continuous step functions, 134
  
- scale transformation, 54
- Schwarz information criterion, 87
- segmentation, 12
- SIC, 87, 171
- SICC, 175
- signal, 10
- simple graph, 9
- site, 9
- step functions, 134
  
- taut string algorithm, 58
- time point, 9
- time series, 13
  
- undirected graph, 9

# Bibliography

- [1] H. Akaike. Information theory and an extension of the maximum likelihood principle. *Proceedings of Second International Symposium on Information Theory*, pages 267–281, 1973.
- [2] H. Akaike. A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, 19(6):716 – 723, 1974.
- [3] D. Auer. Data. Max-Planck-Institute of Psychiatry, Munich.
- [4] H. Bauer. *Wahrscheinlichkeitstheorie*. deGruyter, 1990.
- [5] G. Beer. *Topologies on closed and closed convex sets*, volume 268 of *Mathematics and its Applications*. Kluwer Academic Publishers, Dordrecht, 1993.
- [6] R. E. Bellman. *Dynamic Programming*. Princeton University Press, Princeton, NJ, 1957.
- [7] J. O. Berger. *Statistical Decision Theory and Bayesian Analysis*. Springer Verlag, 2nd edition, 1980, 1985.
- [8] J. Besag and Ch. Kooperberg. On conditional and intrinsic autoregressions. *Biometrika*, 82(4):733–746, 1995.
- [9] A. Blake. The least disturbance principle and weak constraints. *Pattern Recognition Lett.*, 1:393–399, 1983.
- [10] A. Blake and A. Zisserman. *Visual Reconstruction*. The MIT Press Series in Artificial Intelligence. MIT Press, Massachusetts, USA, 1987.
- [11] A. Braides.  *$\Gamma$ -Convergence for Beginners*. Oxford University Press, Oxford, 2002.
- [12] P. J. Brockwell and R. A. Davis. *Time Series: Theory and Methods*. Springer Series in Statistics. Springer Verlag, 2nd edition, 1991.

- [13] J.E. Cavanaugh. Unifying the derivations for the Akaike and corrected Akaike information criteria. *Statistics and Probability Letters*, 33:201–208, 1997.
- [14] J.E. Cavanaugh and A.A. Neath. Generalizing the Derivation of the Schwarz Information Criterion. *Communications in Statistics - Theory and Methods*, 28:49–66, 1999.
- [15] P. L. Davies and A. Kovac. Local extremes, runs, strings and multiresolution. *The Annals of Statistics*, pages 1–65, 2001.
- [16] A.L. Drobyshev, C. Machka, M. Horsch, M. Seltmann, V. Liebscher, M. Hrabé de Angelis, and J. Beckers. Specificity assessment from fractionation experiments (SAFE): a novel method to evaluate microarray probe specificity based on hybridisation stringencies. *Nucleic Acids Research*, 31(2):1–10, 2003.
- [17] J. Elstrodt. *Maß- und Integrationstheorie*. Springer Verlag, 1996.
- [18] F. Friedrich. *ANTSINFIELDS: A Software Package for Random Field Models and Imaging*. Institute of Biomathematics and Biometry, National Research Center for Environment and Health, Neuherberg/München, Germany, 2003.
- [19] S. Geman and D. Geman. Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images. *IEEE Trans. PAMI*, 6:721–741, 1984.
- [20] F.R. Hampel, E.M. Ronchetti, P.J. Rousseeuw, and W.A. Stahel. *Robust Statistics*. Wiley Series in Probability and Mathematical Statistics. Wiley & Sons, New York, 1986.
- [21] M. W. Hirsch. *Differential Topology*. Springer Verlag, 1976.
- [22] F. Hirzebruch and W. Scharlau. *Einführung in die Funktionalanalysis*. Spektrum Akademischer Verlag, 1971.
- [23] P. J. Huber. *Robust statistics*. Wiley Series in Probability and Mathematical Statistics. John Wiley & Sons, New York etc., 1981.
- [24] C. M. Hurvich and C. L. Tsai. Regression and time series model selection in small samples. *Biometrika*, 76(2):297–307, 1989.
- [25] E. Ising. Beitrag zur Theorie des Ferromagnetismus. *Z. Physik*, 31:253, 1925.

- [26] H. R. Künsch. Intrinsic autoregressions and related models on the two-dimensional lattice. *Biometrika*, 74:517–524, 1987.
- [27] V. Liebscher, A. Kempe, and O. Wittich. Consistency of Potts Smoothers. to appear, 2004.
- [28] E. Mammen and S. van de Geer. Locally adaptive regression splines. *The Annals of Statistics*, 25(1):387–413, 1997.
- [29] J.S. Marron and S.S. Chung. Presentation of smoothers: the family approach. *Comput. Stat.*, 16(1):195–207, 2001.
- [30] J. Polzehl and V.G. Spokoiny. Adaptive weights smoothing with applications to image restoration. *J. R. Statist. Soc., Ser. B*, **62**(2):335–354, 2000.
- [31] R.B. Potts. Some generalized order-disorder transitions. *Proc. Camb. Phil. Soc.*, 48:106–109, 1952.
- [32] J. Schmetterer. *Introduction to Mathematical Statistics*. Springer Verlag, 1966.
- [33] G. Schwarz. Estimating the dimension of a model. *The Annals of Statistics*, 6(2):461–464, 1978.
- [34] B. v. Querenburg. *Mengentheoretische Topologie*. Springer Verlag, 1976.
- [35] D. Williams. *Probability with Martingals*. Cambridge Mathematical Textbooks, 1991.
- [36] G. Winkler. *Image Analysis, Random Fields and Markov Chain Monte Carlo Methods. A Mathematical Introduction*, volume 27 of *Applications of Mathematics*. Springer Verlag, Berlin, Heidelberg, New York, second edition, 2003. Completely rewritten and revised.
- [37] G. Winkler and V. Liebscher. Smoothers for discontinuous signals. *J. Nonpar. Statist.*, 14(1-2):203–222, 2002.
- [38] G. Winkler, V. Liebscher, O. Wittich, and A. Kempe. Don’t Shed Tears over Breaks. to appear, 2004.
- [39] H. Witting. *Mathematische Statistik 1*. Teubner Verlag, 1985.



# Lebenslauf

Angela Kempe

geboren am 23. Oktober 1973  
in München

## Schulausbildung

1980 - 1984 Grundschule an der Grafinger Straße in München  
1984 - 1993 Gymnasium Max Josef-Stift in München  
Abschluß Abitur

## Studium

10/93 - 05/99 Studium der Mathematik mit Nebenfach Physik  
an der Technischen Universität München

Diplomarbeit zum Thema  
'Risikoabsicherung durch Quantil-Hedging'  
Betreuerin: Prof. Dr. C. Klüppelberg  
Abschluß Diplom-Mathematikerin

## berufliche Tätigkeit

09/99 - 10/99 wissenschaftliche Hilfskraft am Institut für Biomathematik  
und Biometrie, GSF-Forschungszentrum für Umwelt und  
Gesundheit, Neuherberg  
11/99 - 12/99 wissenschaftliche Hilfskraft im SFB 386 (Statistische Analyse  
diskreter Strukturen, Teilbereich A5 Räumliche Statistik)

## Promotion

01/00 - 12/02 Doktorandin am Institut für Biomathematik und Biometrie,  
GSF-Forschungszentrum für Umwelt und Gesundheit,  
Neuherberg  
01/03 - 01/04 Stipendiatin im Graduiertenkolleg  
'Angewandte Algorithmische Mathematik'  
an der Technischen Universität München