technische universität
dortmund

fakultät statistik

On Nonparametric Bayesian Analysis under Shape Constraints with Applications in Biostatistics

Dissertation

zur Erlangung des Grades

eines Doktors der Naturwissenschaften

der Technischen Universität Dortmund

Der Fakultät Statistik

der Technischen Universität Dortmund

vorgelegt von

Björn Bornkamp

Dortmund, September 2009

Gutachter:

Prof. Dr. Katja Ickstadt

Prof. Dr. Roland Fried

Tag der mündlichen Prüfung:

26. November 2009

Table of Contents

List of Symbols and Abbreviations

Symbols

y	Dependent values (in $\mathbb{R}$)
$\boldsymbol{x}$	Independent values/covariates (lying in $\mathcal{X} \subset \mathbb{R}^k$)
$\mathcal{X}$	Covariate space (subset of $\mathbb{R}^k$)
n	Number of observations
$\mathcal{P}_{\boldsymbol{\theta}_x}$	Residual probability distribution at the covariate value $\boldsymbol{x}$
$\boldsymbol{\theta}_{\boldsymbol{x}}$	Parameter describing the residual probability distribution at $\boldsymbol{x}$
$\boldsymbol{\Theta}$	Parameter space, *i.e.* $\boldsymbol{\theta}_{\boldsymbol{x}} \in \boldsymbol{\Theta}$ for all $\boldsymbol{x} \in \mathcal{X}$
$\delta_{\tilde{\xi}}(.)$	Probability measure degenerated at $\tilde{\xi}$ (unit point mass at $\tilde{\xi}$)
$\mathbb{1}_B(x)$	Indicator function for the set B *i.e.* $\mathbb{1}_B(x) = 1$ if $x \in B$ and otherwise 0
$\mathbb{R}_+$,$\mathbb{N}_+$	Set of positive real numbers and set of positive integers.
$\mathbb{S}^N$	The probability simplex, *i.e.* $\mathbb{S}^N = \{(\pi_1, \ldots, \pi_N) \in \mathbb{R}^N : \sum_{h=1}^N \pi_h = 1, \pi_h \geq 0\}$
$\boldsymbol{\xi}$	Latent variables (usually parameters of a kernel function with $\boldsymbol{\xi} \subset \boldsymbol{\Xi}$)
$\Pi(d\boldsymbol{\theta})$	Prior distribution for the (possibly infinite dimensional) parameter $\boldsymbol{\theta}$
$\Pi_n^*(d\boldsymbol{\theta})$	Posterior distribution for the (possibly infinite dimensional) parameter $\boldsymbol{\theta}$

Abbreviations

BNP	Bayesian nonparametrics
MCMC	Markov Chain Monte Carlo
cdf	Cumulative distribution function
TSP	Abbreviation for two sided power (distribution)
MED	Minimum effective dose
plcmr	Positive linear combinations of monotonic ridge functions
GAD	Gestational age at delivery

Chapter 1

Introduction

Nothing in nature is random.... A thing appears random only through the incompleteness of our knowledge.
Baruch Benedictus de Spinoza (1632-1677)

This thesis is concerned with a fundamental problem in the analysis of empirical data: Suppose you obtain data points $(y_i, \boldsymbol{x}_i)$, for $i = 1, \ldots, n$, and the task is to explain the response values $y \in \mathbb{R}$ in terms of input values $\boldsymbol{x} \in \mathcal{X} \subset \mathbb{R}^k$, while incorporating geometric or structural information in form of shape constraints. Examples for shape constraints which will be investigated in this thesis are monotonicity, convexity and concavity constraints, when modelling a functional relationship, and a stochastic ordering constraint when modelling probability distributions.

As a practical example, where imposing a shape constraint is adequate, consider the data displayed in Figure 1.1, containing the height of a child over a period of 312 days (see Ramsay (1998) for further information). In this situation it is reasonable to assume that the height of the child increases monotonically with time, and any statistical model should incorporate this type of geometric prior information (see Section 3.5 for an evaluation of this example). As a second example Figures 1.2 (i) and (ii) display histograms of the birth weight of newborns (in grams) for smoking and non-smoking mothers, participating in the US Collaborative Perinatal Project (see Longnecker, Klebanoff, Zhou and Brock (2001) for details). As nicotine is well known to be a toxic substance, it is reasonable to assume that the distribution of birth weight for newborns of non-smoking mothers is (stochastically) not smaller than that of newborns from smoking mothers. Again this information should be incorporated in the statis-

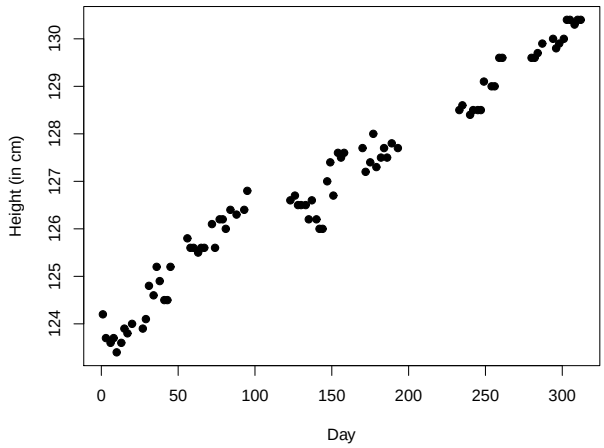

Figure 1.1: Height measurements of a child over 312 days.

tical model, and our analysis in Section 2.1.1 will do so. Another example of shape constrained inference we will encounter in this thesis are pharmaceutical dose-finding studies. Figure 1.3 displays data from a pharmaceutical phase II dose-finding trial for the treatment of the irritable bowel syndrome (see Biesheuvel and Hothorn (2002) for additional information). In these type of trials it is biologically reasonable to assume either a monotonic or a unimodal relationship (due to potential toxicity at larger doses) between the dose administered and the dose effect, and again this should be incorporated in the used statistical model (see Bretz, Hsu, Pinheiro and Liu (2008) for a recent review of pharmaceutical Phase II trials). In Chapter 3 we will investigate this application in more detail.

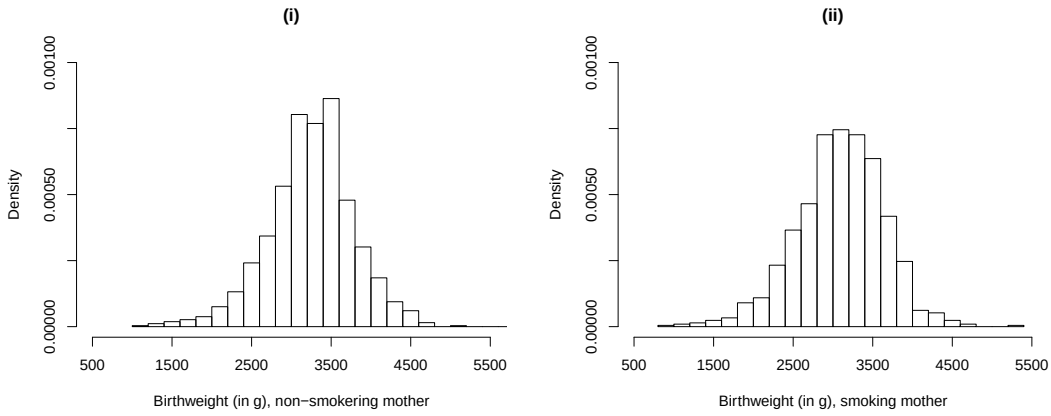

Figure 1.2: Birth weights for newborns with non-smoking and smoking mothers.

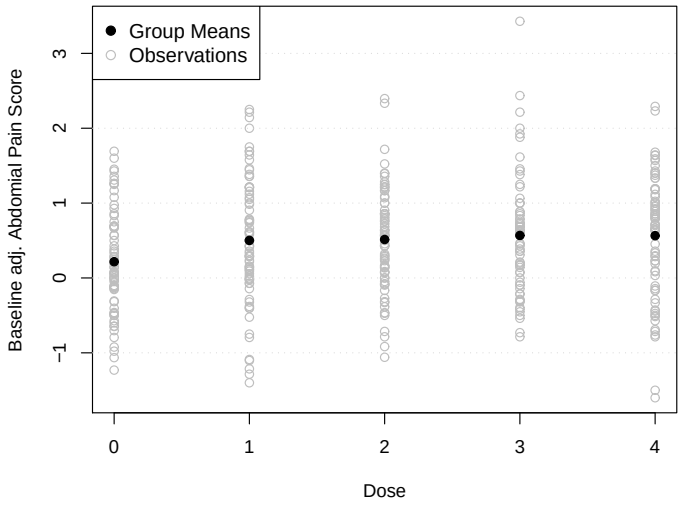

Figure 1.3: Data from a dose-response trial on the irritable bowel syndrome.

In this thesis we will approach the problem of shape constrained modelling from the Bayesian perspective. In this framework, probability is used to model two different kinds of uncertainties. The first use of probability is for modelling the uncertainty in the observations (*i.e.* the residual error), and the second use is for quantifying the uncertainty in the statistical model before any data are observed. Throughout this thesis probability (as in the quote of Spinoza) is hence meant as a quantification of uncertainty (the things we do not know) rather than *true* randomness.

It is, for example, usually not reasonable to assume that observations are made with absolute precision. For the data in Figure 1.1 it is quite likely that there are simple measurement errors, resulting in an uncertainty in the observed values. Additionally in most practical situations it is not reasonable to assume that we have all relevant inputs $\boldsymbol{x}$ available, which influence y (then one could perfectly explain y with the information contained in $\boldsymbol{x}$ and an interpolation model without any error term would be suited). In practice, some inputs might have been forgotten, while others might be too expensive to measure or not measurable at all. In the case of birth weight of newborns, for example, it is hardly imaginable to find a realistic set of predictors, such that one can perfectly predict the birth weight of a newborn. So there is often heterogeneity in the data, which cannot be explained by the inputs $\boldsymbol{x}$. The traditional statistical approach to account for this unobserved heterogeneity (and more generally for uncertainty in

the observed values) is to model y, given the inputs $\boldsymbol{x}$ as the realisation of a probability distribution $\mathcal{P}_{\boldsymbol{\theta}_x}$, depending on parameters $\boldsymbol{\theta}_x$, which vary with $\boldsymbol{x}$. The distribution $\mathcal{P}_{\boldsymbol{\theta}_x}$ for a given $\boldsymbol{\theta}_x$ and $\boldsymbol{x}$ hence represents the uncertainty in y, which *cannot* be explained by the inputs $\boldsymbol{x}$.

An important question is now, how to set up the statistical model for the collection of residual probability distributions $\mathcal{F}_{\mathcal{X}} = \{\mathcal{P}_{\boldsymbol{\theta}_x} \text{ with } \boldsymbol{x} \in \mathcal{X}\}$. There exist two main type of approaches for this task: The *parametric* and the *nonparametric* approach. The parametric approach imposes a strong structural assumption: A probability distribution from a known family is specified for $\mathcal{P}_{\boldsymbol{\theta}_x}$, which depends on a parameter $\boldsymbol{\theta}_x$, lying in a finite dimensional space $\boldsymbol{\Theta}$ for all $\boldsymbol{x} \in \mathcal{X}$. So the functional form of $\mathcal{P}_{\boldsymbol{\theta}_x}$ and $\boldsymbol{\theta}_x$ are assumed to be known and only finitely many unknowns need to be inferred from the data. While the parametric approach is adequate in many modelling situations, there are situations, where it is difficult to make these underlying assumptions. The *nonparametric* approach overcomes this by being less restrictive: Here $\mathcal{P}_{\boldsymbol{\theta}_x}$ and $\boldsymbol{\theta}_x \in \boldsymbol{\Theta}$ (themselves as a probability measure or a function) are treated as unknown, rather than just a set of finitely many parameters describing them and the space $\boldsymbol{\Theta}$ is hence infinite dimensional. Rather than having no parameters (as the name nonparametric might suggest), nonparametric models typically have infinitely many parameters.
We keep the discussion rather general at this point: In some situations, even in the nonparametric setting, a parametric family is assumed for $\mathcal{P}_{\boldsymbol{\theta}_x}$ for a given $\boldsymbol{\theta}_x$ and only the $\boldsymbol{\theta}_x$ is modelled nonparametrically (*i.e.* using an infinite-dimensional object). One example is nonparametric regression with normally distributed errors, here $\mathcal{P}_{\boldsymbol{\theta}_x} = N(\mu(\boldsymbol{x}), \sigma^2)$ (hence $\boldsymbol{\theta}_x = (\mu(\boldsymbol{x}), \sigma^2)'$) and only the conditional mean function $\mu(\boldsymbol{x})$ is modelled nonparametrically. On the other hand, the used notation also includes nonparametric univariate density estimation: Here there is no dependence on any input parameter $\boldsymbol{x}$, and only the residual distribution $\mathcal{P}_{\boldsymbol{\theta}}$ is modelled nonparametrically. This situation occurs, for example, when one uses an infinite mixture of normal distributions for the density of $\mathcal{P}_{\boldsymbol{\theta}}$, *i.e.* $\sum_h \pi_h \phi(y, \mu_h, \sigma_h^2)$, where $\phi(y, \mu, \sigma^2)$ denotes the density of a normal distribution with mean μ and variance σ^2 and $\sum_h \pi_h = 1$. In this situation the parameter $\boldsymbol{\theta}$ is given by $(\pi_1, \mu_1, \sigma_1^2, \pi_2, \mu_2, \sigma_2^2, \ldots)'$, and does not depend on any $\boldsymbol{x}$.

How do we infer reasonable values for $\mathcal{P}_{\boldsymbol{\theta}_x}$ and $\boldsymbol{\theta}_x$ for $\boldsymbol{x} \in \mathcal{X}$, based on the data

$(y_i, \boldsymbol{x}_i)$, $i = 1, \ldots, n$ that have been observed? Here we encounter the second use of probability within the Bayesian framework, namely that for quantifying the uncertainty in the statistical model $\mathcal{P}_{\theta_x}$, $\boldsymbol{x} \in \mathcal{X}$ before any data are observed. The main idea is (roughly) the following: We do not know $\mathcal{P}_{\theta_x}$ for $\boldsymbol{x} \in \mathcal{X}$ and are hence uncertain about their particular value. One can quantify this uncertainty in terms of a probability distribution, the so-called prior distribution for $\mathcal{P}_{\theta_x}, \boldsymbol{x} \in \mathcal{X}$ (see Lindley (2000) for philosophical arguments that probability is an adequate measure for quantifying uncertainty). Once data have been observed, we have the distribution of y_i conditional on $\mathcal{P}_{\theta_{x_i}}$ for $i = 1, \ldots, n$ (the residual or sampling distribution), and the prior distribution for $\mathcal{P}_{\theta_x}$, $\boldsymbol{x} \in \mathcal{X}$. Now one can use the apparatus of probability theory (Bayes theorem) to form the conditional distribution of $\mathcal{P}_{\theta_x}$ given the data (the posterior distribution). This posterior distribution plays a central role in the Bayesian approach: It merges information about $\mathcal{P}_{\theta_x}$ contained in data and prior distribution, and thus forms the basis for subsequent decision making.

In the parametric situation, prior (and posterior) distributions are typically supported on a finite dimensional subset of $\mathbb{R}^k$. In the nonparametric situation the unknown object is a function or a probability measure and hence infinite dimensional, so prior distributions are typically stochastic processes, which in this sense can be seen as extensions of finite dimensional distributions to the infinite case. Setting up a nonparametric prior distribution is hence more challenging than in the parametric situation: There the finite number of parameters often have a meaning in the context of the application and it is typically possible to incorporate prior information (or lack thereof) into the problem. In the nonparametric situation, we need a prior distribution for $\mathcal{F}_{\mathcal{X}} = \{\mathcal{P}_{\theta_x} \text{ with } \boldsymbol{x} \in \mathcal{X}\}$, which is a quite complex object. Intuitively any prior distribution, should ideally fulfill two properties: (i) *Adequacy*: It should reflect the information about the underlying statistical model before any data are observed (which might be, depending on the situation, very scarce). (ii) *Full Support*: It should assign positive probability to all statistical models, which are relevant in the application. This requires that one can approximate any statistical model in a (usually problem specific) mathematical distance measure and the prior needs to assign positive probability to any of these approximations. The full support requirement (ii) is mainly of mathemat-

ical nature, but a relieving property: It ensures that we assign positive probability to all statistical models which are reasonable a-priori, and hence also allow the posterior probability to put its mass there. Certainly the quantification of prior information in (i) is a challenging issue in the nonparametric situation, as prior distributions are defined on abstract spaces such as the space of probability measures on $\mathbb{R}$ or the space of continuous functions on $[0,1]$. However, in many situations it is possible to obtain summaries for the prior distribution such as the first two moments and one can center the priors on a prior guess, for example a parametric model, and associate this prior guess with an appropriately large variability. It is also helpful to employ models, which remain at least in some important aspects interpretable, so that one can match the prior distribution for these interpretable aspects with the potential prior information available.

A very useful tool to reduce the effective complexity of the model are, when adequate, shape constraints (see Appendix A.1 for an overview of different type of shape constraints encountered in this thesis). They allow to incorporate geometric prior information and by this allow to narrow down the class of statistical models achieving that no prior probability mass (and hence also no posterior probability mass) is "wasted" on models, which are implausible a-priori. Statistically this in turn considerably improves the efficiency of inference. For instance in the growth curve example, it makes sense to exclude non-monotonic curves from the prior distribution and inference can focus on the relevant class of monotonic functions. In addition shape constraints are usually very intuitive to interpret and communicate. This thesis is devoted to these types of assumptions in particular in the highly flexible nonparametric situation.

The outline of this thesis is as follows: In the next chapter we will introduce the basics of Bayesian nonparametrics, to equip the reader with the necessary mathematical and statistical background for the main Chapters 3, 4 and 5. In these chapters we develop novel Bayesian nonparametric models for three concrete problems in shape constrained inference: Chapter 3 deals with Bayesian nonparametric regression under a monotonicity assumption, Chapter 4 treats shape constraints on the derivative of the modelled function and Chapter 5 proposes a method for stochastically ordered density regression. Chapter 6 concludes this work.

Chapter 2

Bayesian Nonparametrics

With four parameters I can fit an elephant, and with five I can make him wiggle his trunk.
John von Neumann (1903-1957)

In this chapter we will give a review of Bayesian nonparametric (BNP) methodologies. The literature on nonparametric Bayesian methods has exploded in the last few years and we will not try to review all relevant developments, the focus will lie on ideas, which will be needed as background for Chapters 3, 4 and 5. Other reviews of nonparametric Bayesian methods are given, for example, in Müller and Quintana (2004), O'Hagan and Forster (2004, ch. 13), Dey and Rao (2005, ch. 10-13) or Walker, Damien, Laud and Smith (1999), for an earlier reference. Books about nonparametric Bayesian statistics are, for example, Dey, Sinha and Müller (1998), which describes the use of BNP methods in a variety of practical problems, Ghosh and Ramamoorthi (2003), which concentrates on an asymptotic analysis of Bayesian nonparametrics and the forthcoming book by Hjort, Holmes, Müller and Walker (2010).

Interest in Bayesian nonparametric methodologies started to grow with the publication of the paper by Ferguson (1973), where a nonparametric prior for a probability measure, the Dirichlet process, was introduced in the form it is currently in use, and its conjugacy property was proved. Although numerous generalizations have been proposed in the meantime, the Dirichlet process still provides the building block of many nonparametric models. A second breakthrough was the publication of the papers by Blight and Ott (1975) and by O'Hagan (1978), who use the Gaussian process as a prior for the conditional mean function in nonparametric regression under nor-

mally distributed errors (which is the conjugate prior in this setting), and the paper of Dykstra and Laud (1981), who propose extensions of the gamma process as a prior for intensities in survival analysis (see also Ferguson and Phadia (1979), Doksum (1974), and Hjort (1990) for important contributions to survival analysis). A major practical limitation of these early approaches was the fact that only a (relatively) small and simple class of models could be analyzed. This situation changed when Markov Chain Monte Carlo methodologies such as the Gibbs sampler (Gelfand and Smith 1990) and the Metropolis-Hastings algorithm (Tierney 1994) became popular, which can also be applied in the nonparametric situation with some adaptions. The availability of these algorithms together with the increase in computing power led to an upsurge of interest in nonparametric Bayesian methods in all type of practical applications in the last 20 years, see, for example, the forthcoming review articles of Dunson (2010) and Müller and Quintana (2010), which illustrate the use of nonparametric Bayesian methods in a variety of important and difficult applied problems in biostatistics. On the theoretical side, interest in BNP methodologies has focused in the last years either on probabilistic properties of the underlying stochastic processes used as priors or on asymptotic properties of BNP methodologies.

The outline of the rest of this chapter is as follows: The first section deals with priors for probability measures. Section 2.2 deals with priors for functions and Section 2.3 briefly reviews the asymptotic viewpoint on Bayesian nonparametrics, which is helpful for a mathematical understanding of Bayesian nonparametrics. Although the main examples of shape constrained inference can be found in Chapters 3, 4 and 5, there will be a focus on shape constrained inference already in this chapter.

2.1 Priors for Probability Measures

Let $(\Xi, \mathcal{B})$ be a measurable space. This chapter is concerned with prior distributions for probability measures on $(\Xi, \mathcal{B})$, *i.e.* stochastic processes (also called random probability measures), which generate probability measures P on Ξ. We keep the discussion fairly general in this section, apart from $(\Xi, \mathcal{B})$ being a measurable space no other as-

sumption will be made in most of this section, particularly Ξ remains unspecified, as we will encounter different choices of Ξ in this thesis. First we will concentrate on the Dirichlet process as a prior for a probability measure, following the historic developments and because of its importance. Later we will present a generalization.

The Dirichlet process was popularized in its current form by Ferguson (1973). A Dirichlet process $\mathbb{P}$ on $(\Xi, \mathcal{B})$ has two parameters $(M, P_0)'$, where $M \in \mathbb{R}_+$ is a precision parameter and P_0 a probability measure on $(\Xi, \mathcal{B})$.

Definition 2.1.1 (Dirichlet Process)**.**
P is distributed according to a Dirichlet process, if and only if the joint distribution of $(P(B_1), \ldots, P(B_k))'$ for any finite partition $(B_1, \ldots, B_k)'$ of the sample space Ξ has a $k-1$ dimensional Dirichlet distribution with parameter $(MP_0(B_1), \ldots, MP_0(B_k))'$.

The Dirichlet distribution is a multivariate continuous distribution, generating values on the $k-1$ dimensional probability simplex $\mathbb{S}^k = \{(\pi_1, \ldots, \pi_k) \in \mathbb{R}^k : \sum_{h=1}^k \pi_h = 1, \pi_h \geq 0\}$. From the properties of the beta distribution (the Dirichlet distribution in one dimension) it follows that $\mathrm{E}(P(B)) = P_0(B)$ and $V\mathrm{ar}(P(B)) = \frac{P_0(B)(1-P_0(B))}{M+1}$. Hence the base probability measure P_0 is the prior mean of the Dirichlet process and the parameter M determines the variability in the prior distribution: For larger M, we obtain a smaller variance.

The important property of the Dirichlet process proved by Ferguson (1973) is the conjugacy to an independent and identically distributed sample $y_1, \ldots, y_n$ from a distribution $\mathcal{P}$ on Ξ: Assuming a Dirichlet process prior with parameter M and P_0 for $\mathcal{P}$ the posterior distribution for $\mathcal{P}$ is again a Dirichlet process, with updated parameters $(M+n, P_0^*)$ with $P_0^* = (MP_0 + nF_n)/(M+n)$, where $F_n = \frac{1}{n}\sum_{i=1}^n \delta_{y_i}$ is the empirical probability measure of the observed data, and δ_{y_i} the probability measure degenerated at y_i. The posterior mean is hence given by $\mathrm{E}(P(B)|y_1, \ldots, y_n) = P_0^*(B)$ and the posterior variance by $\mathrm{Var}(P(B)|y_1, \ldots, y_n) = \frac{P_0^*(B)(1-P_0^*(B))}{M+n+1}$ so that the posterior Dirichlet process is a compromise between the prior mean P_0 and the empirical probability measure. When the sample size n increases, one can see from the formula for P_0^* that the data will eventually dominate the posterior mean and the posterior variability of $P(B)$ converges to 0. Additionally a simple application of Chebyshev's inequality and the

triangle inequality in Theorem 2.1.1 shows that $P(B)$ converges in probability towards the true probability $\mathcal{P}(B)$ when $n \to \infty$ (a refined treatment of asymptotics can be found in Section 2.3).

Theorem 2.1.1.
Let P denote a realization from the posterior of a Dirichlet process based on n observations. Then for $B \in \mathcal{B}$

$$P(B) \xrightarrow{P} \mathcal{P}(B) \text{ as } n \to \infty.$$

Proof.

$$\begin{aligned} |P(B) - \mathcal{P}(\mathcal{B})| &\leq |P(B) - P_0^*(B)| + |P_0^*(B) - \mathcal{P}(B)| \\ &\leq |P(B) - P_0^*(B)| + \left| P_0^*(B) - \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}_B(y_i) \right| + \left| \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}_B(y_i) - \mathcal{P}(B) \right| \end{aligned}$$

The first summand converges in probability to zero because of Chebyshev's inequality. The second summand converges deterministically to zero and the last summand converges almost surely towards zero, because of the law of large numbers. □

There are several equivalent ways of defining the Dirichlet process, but the definition given by Sethuraman (1994) turned out to be particularly interesting, because it inspired computational approaches to analyse models based on the Dirichlet process. Additionally it directly shows that the Dirichlet process only generates discrete probability measures, which is not directly obvious from Definition 2.1.1. Sethuraman showed that the law of a Dirichlet process is identical to the law of the random probability measure

$$P(d\boldsymbol{\xi}) = \sum_{h=1}^{\infty} \pi_h \delta_{\boldsymbol{\xi}_h}(d\boldsymbol{\xi}), \text{ with } \boldsymbol{\xi}_h \overset{iid}{\sim} P_0 \ , \tag{2.1}$$

where P_0 is the base probability distribution on Ξ and $\pi_h = V_h \prod_{l<h}(1 - V_l)$ with $V_h \overset{iid}{\sim} \text{Beta}(1, M)$, is the probability mass allocated to $\boldsymbol{\xi}_h$. This representation is often called the stick-breaking representation, because starting with a probability stick of length one, V_1 is the proportion of the stick broken off and allocated to $\boldsymbol{\xi}_1$, V_2 is the proportion of the remaining $1 - V_1$ stick length allocated to $\boldsymbol{\xi}_2$, and so on. The distribution induced for $\pi_1, \pi_2, \ldots$ is also called Griffiths-Engen-McCloskey (GEM) distribution (see Ishwaran and Zarepour (2002)). From this stick-breaking representation it

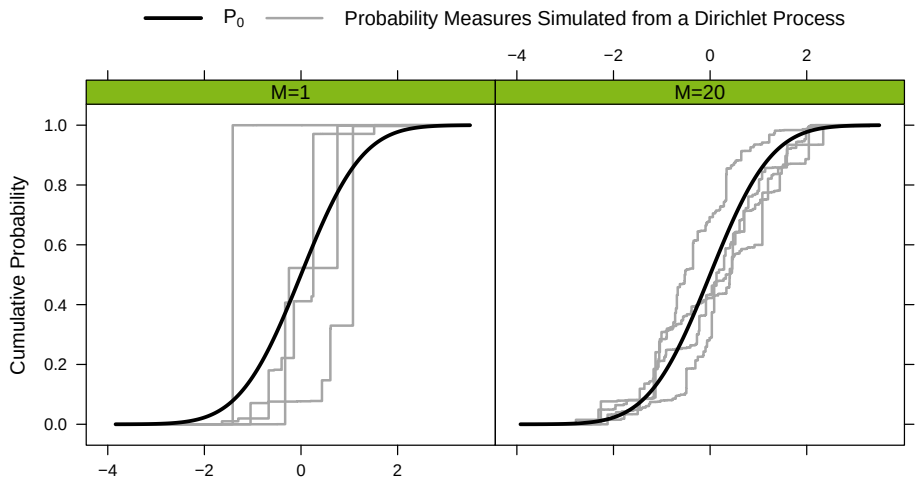

Figure 2.1: Cumulative distribution function of four simulations from a Dirichlet process with a standard normally distributed base measure ($P_0 = N(0,1)$) and different precision parameters M.

becomes obvious that the precision parameter M also determines how the total probability mass is allocated to the support points of the generated probability measure: For small M, most probability mass will be distributed on the first realizations of P_0, for $M \to \infty$ there will be many realisations with a larger probability mass and a specific realization P will be more similar to P_0. This becomes visible also in Figures 2.1 and 2.2, where one can observe the cumulative distribution functions and the probability mass function of realizations from a Dirichlet process for two different values of M.

However the main practical hindrance in using the Dirichlet process is its discreteness. Most commonly continuous phenomena are modelled, and in this case a discrete random probability measure simply does not reflect the prior information. A simple and very versatile approach to overcome this discreteness issue are mixture models. A mixture model can be represented as the following hierarchical model

$$Y_i \sim f(.|\xi_i),\ i = 1,\dots,n, \quad \xi_1,\dots,\xi_n \sim P(d\xi),$$

where $f(.|\xi)$ is the density of a parametric distribution, depending on parameters $\xi \in \Xi$ and $P(d\xi)$ is a discrete mixing distribution, so that some of the ξ_i can be equal in the above formula. Another, equivalent, way of representing this hierarchical model is by

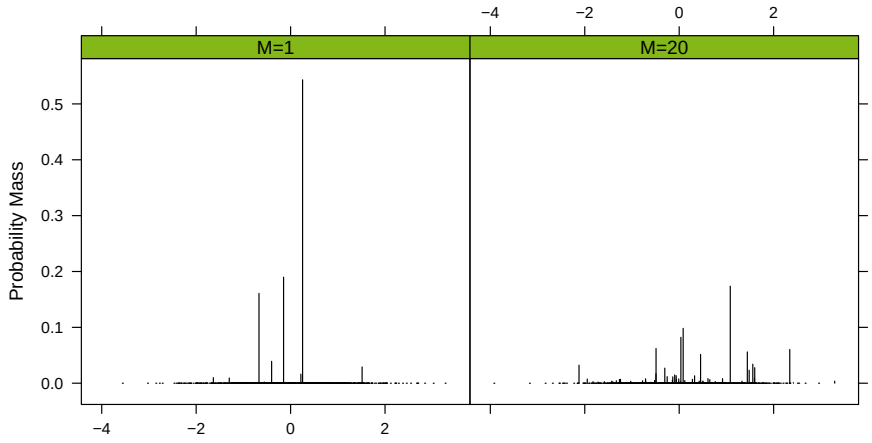

Figure 2.2: Probability mass function for two simulations from a Dirichlet process with $P_0 = N(0,1)$; one with precision parameter $M = 1$ and one with $M = 20$.

integrating out the latent variables $\boldsymbol{\xi}_h$

$$\int f(y, \boldsymbol{\xi}) P(d\boldsymbol{\xi}) = \sum \pi_h f(y, \boldsymbol{\xi}_h).$$

As a prior for the discrete mixing distribution $P(d\boldsymbol{\xi})$ one can now use the Dirichlet process, with parameters M and P_0. So the Dirichlet process is used in a higher level of the hierarchy, as a prior for the latent variables $\boldsymbol{\xi}_h$, instead of modeling the residual distribution directly by a Dirichlet process. The choice of the parametric density $f(., \boldsymbol{\xi})$ depends on the application: When it is desired, for example, to model a distribution on $\mathbb{R}$ one might choose the normal density for $f(., \boldsymbol{\xi})$, so that $\boldsymbol{\xi} = (\mu, \sigma^2)$ and $\Xi = \mathbb{R} \times \mathbb{R}_+$. Then the model is similar to the traditional kernel density estimator with a Gaussian kernel. Similarly one might use the beta density when there are bounds on the distribution, or the gamma distribution when the sample space of interest is $(0, \infty)$. When one is not explicitly interested in modelling a probability distribution, one can even allow an arbitrary integrable function $g(z, \boldsymbol{\xi})$, $g : \mathcal{Z} \times \Xi \mapsto \mathbb{R}$ as the function which is mixed (and this will be done in Chapters 3, 4 and 5). The mixing idea is hence quite general and can be used to model arbitrary functions, not only probability densities.

When the focus is on modelling probability distributions, additional guidance for the choice of $f(.|\boldsymbol{\xi})$ is given by approximation theoretic considerations: It is for example desirable to use a density $f(.|\boldsymbol{\xi})$ generating mixtures, which are rich enough to ap-

proximate (with respect to a specified distance measure) any probability measure on the underlying space (for example, any continuous probability distribution on $\mathbb{R}$). This is a necessary requirement to achieve *full support* in the chosen distance measure for the constructed prior distribution. Actually, this is also one of the major assumptions for consistency in an asymptotic analysis of BNP methods, where the Kullback-Leibler divergence plays a prominent role (see Wu and Ghosal (2008) for a collection of kernels for which the full support property holds with respect to this distance measure). We will discuss the asymptotic aspects more extensively in Section 2.3.

In a mixture modelling framework shape constrained inference problems can often be reduced to unconstrained inference, with a clever choice of the kernel, *i.e.* the density function $f(., \boldsymbol{\xi})$, as Lo (1984) notes (see also Brunner and Lo (1989), Brunner (1992), or Hansen and Lauritzen (2002)). For example, any monotonically decreasing density on $[c, \infty)$ with $c \in \mathbb{R}$ can be represented as a mixture of uniform densities $f(y, \xi) = \frac{1}{\xi - c}\mathbb{1}_{[c,\xi]}(y)$, $\xi > c$. Any unimodal distribution on $\mathbb{R}$ with mode 0 can be represented as a mixture of kernels of the following form $f(y, \xi) = \frac{1}{\xi}(\mathbb{1}_{(0,\xi]}(y) - \mathbb{1}_{[\xi,0)}(y))$ with $\xi \in \mathbb{R}$. Similarly any completely monotone density can be written as a mixture of densities from the exponential distribution, according to a theorem of Bernstein (Lo 1984). Hence with a clever choice of the density function $f(y, \boldsymbol{\xi})$ one can turn a shape constrained inference problem into an unconstrained mixture estimation problem. A variety of further examples exploiting this idea for general constrained inference problems are given in the article of Hoff (2003b). As becomes obvious, this general "discrete mixture" idea is an extremely powerful tool and we will extensively use it in Chapters 3, 4 and 5, also in more abstract settings than the one presented here.

Although the mixture idea was already introduced by Antoniak (1974), shortly after Ferguson's paper, a practical application was not possible because updating of the posterior distribution is computationally very hard. When a base measure P_0 conjugate to the density function $f(., \boldsymbol{\xi})$ is used, analytical updating is possible, but it involves calculation of all possible partitions from the data points, see Lo (1984), and is thus infeasible for realistic sample sizes. The main computational advance was the paper by Escobar and West (1995), where a Gibbs sampler is derived for sampling the posterior distribution in this type of model. Without going into details the basic idea is to inte-

grate out the Dirichlet process via the Blackwell and MacQueen (1973) alternative (so called Pólya urn-) representation of the Dirichlet process. See MacEachern and Müller (1998), Neal (2000) or Ishwaran and James (2001), for reviews of this computational methodology. An alternative versatile method for simulating the posterior distribution is the blocked Gibbs sampler introduced by Ishwaran and James (2001). The main idea of this type of approach is to truncate the Sethuraman representation (2.1) and then use the finite mixture model to approximate the Dirichlet process model. The approach is not limited to the Dirichlet process and it overcomes some of the problems with MCMC mixing encountered with the Pólya urn Gibbs sampler. Although one can control the error made through this finite dimensional approximation (Ishwaran and James (2001) derive error bounds), this approach is not exact. Recently however Papaspiliopoulos and Roberts (2008) and Walker (2007) discuss procedures to turn the blocked Gibbs sampler into an exact simulation approach. We will describe the blocked Gibbs sampler in more detail later in Section 5.2.3.

There exist many discrete random probability measures that can be used as alternatives to the Dirichlet process. Probably the most general alternative discrete random probability measure is described by Ongaro and Cattaneo (2004), which is equivalent to the class of species-sampling models (Pitman (1996), Ishwaran and James (2003)), when one restricts this class to discrete random probability measures. We will describe this random measure here and identify several discrete random probability measures as a special case.

Definition 2.1.2.
A random probability measure $\mathbb{P}$ *belongs to the Ongaro-Cattaneo class when its realizations can be represented as*

$$P(d\boldsymbol{\xi}) = \sum_{h=1}^{N} \pi_h \delta_{\boldsymbol{\xi}_h}(d\boldsymbol{\xi}), \tag{2.2}$$

where $\boldsymbol{\xi}_h$, π_h *and* N *are random variables specified as follows: The* $\boldsymbol{\xi}_h$ *are independent and identically distributed realizations of a nonatomic distribution* P_0 *on* $\boldsymbol{\Xi}$ *(i.e.* $P_0(\{\boldsymbol{\xi}\}) = 0, \forall \boldsymbol{\xi} \in \boldsymbol{\Xi}$*) and are independent from* π_h, $j = 1, \ldots, N$ *and* N*. The weights* $\pi_1, \ldots, \pi_N$ *conditional on* N *have a distribution* Q_N *on the* $N-1$ *dimensional probability simplex* $\mathbb{S}^N = \{(\pi_1, \ldots, \pi_N)' \in \mathbb{R}_+^N : \sum_{j=1}^N \pi_h = 1\}$ *and* N *is a random variable with support* $\{\mathbb{N}_+ \cup \infty\}$.

Most of the work in the following chapters will assume that an Ongaro-Cattaneo random probability measure is used, and results are derived under this assumption. Hence to illustrate the richness of this class, the following remark lists some random probability measures, which arise as a special case of this general model.

Remark 2.1.1.
The following random measures can be identified as special cases of the Ongaro-Cattaneo random measure from Definition 2.1.2.

- **Dirichlet Process.** It directly follows from Equation (2.1) that the Dirichlet process is a member of this class. The parameter $N = \infty$ and the distribution for the weights is the GEM distribution.

- **Prior Process for Finite Mixtures.** In finite mixture models typically N is fixed and a symmetric Dirichlet distribution is used for the weights. In some cases N is treated as unknown and modelled, for example, using a Poisson or a negative binomial distribution shifted to the positive integers. See Frühwirth-Schnatter (2006) for a review of finite mixture models. Note that formally, although typically highly flexible, these type of models are not nonparametric when N is truncated.

- **Stick-Breaking Priors.** Stick-breaking priors are represented exactly as (2.2), with the restriction that N is fixed (either infinite or finite) and a specific prior is assumed for the weights π_h:

$$\pi_h = V_h \prod_{k<h}(1 - V_k),\ h \in \{1, \dots, N\} \text{ with } V_h \overset{iid}{\sim} \text{Beta}(a_h, b_h) \tag{2.3}$$

  where $a_h, b_h > 0$. If N is finite, one sets $V_N = 1$ (to ensure $\sum_{h=1}^{N} \pi_h = 1$). The resulting distribution of the weights π_h is a generalized Dirichlet distribution (Ishwaran and James 2001), which is conjugate to multinomial sampling. This is the basis underlying the blocked Gibbs sampler (and the scope of priors that can be analysed with the blocked Gibbs sampler is the class of stick-breaking priors). When $N = \infty$ one needs additionally $\sum_{h=1}^{\infty} \log(1 + a_h/b_h) = \infty$ as this ensures that $\sum_h \pi_h = 1$ (Ishwaran and James 2001). When choosing $a_h = 1$ and $b_h = M$, $\forall h$, one recovers the Dirichlet process as is obvious from the Sethuraman

representation in Equation (2.1). The two-parameter Poisson-Dirichlet process (sometimes also called Pitman-Yor process) is another famous member of this class. It depends on two parameters a and b and has $N = \infty$, $a_h = 1 - a$ and $b_h = b + hM$ with $a \in [0, 1)$ and $b > -a$.

- **Dirichlet Multinomial Process.** Instead of truncating the stick-breaking representation of the Dirichlet process one can also use an alternative finite mixture to approximate the Dirichlet process (see Ishwaran and Zarepour (2002)). Here one uses a conservative upper bound $N_{\max}$, and a symmetric Dirichlet distribution with parameter $M/N_{\max}$ for the weights π_h. Ishwaran and Zarepour (2002) show that the so obtained finite mixture model converges in distribution to the Dirichlet process infinite mixture model as $N_{\max} \to \infty$.

- **Normalized Random Measures with Independent Increments.** The main idea of this class is to normalize a random measure with independent increments. This class is a special case of the Ongaro and Cattaneo random probability measure, under a certain restriction. James, Lijoi and Prünster (2009) describe this general class of random probability measures and consider its prior to posterior analysis. To introduce this prior we need to introduce the notion of a Poisson random measure $\tilde{N}$ on $\mathbb{R}_+ \times \Xi$ with intensity ν and corresponding random measure $\tilde{\mu}(B) = \int_{\mathbb{R}_+ \times B} s\tilde{N}(ds, d\boldsymbol{\xi})$ (sometimes also called Lévy random measure, see the Appendix A.2 for details). Now define by $T = \tilde{\mu}(\Xi)$ (this is almost surely positive and finite if for $\nu(ds, d\boldsymbol{\xi})$ holds $\int_{\mathbb{R}_+ \times \Xi} \nu(ds, d\boldsymbol{\xi}) = \infty$ and $\int_{\mathbb{R}_+ \times \Xi} [1 - e^{-\lambda s}] \nu(ds, d\boldsymbol{\xi}) < \infty$ for all $\lambda > 0$, see James, Lijoi and Prünster (2009) for details), the normalized random measure is then given by $\tilde{P}(B) = \tilde{\mu}(B)/T$. It is straightforward to see that the so constructed random measure is indeed a random probability measure. Now the main condition to be a member of the Ongaro-Cattaneo class is imposed on the intensity measure: If the intensity measure $\nu(ds, d\boldsymbol{\xi})$ can be factorized as $\nu(ds, d\boldsymbol{\xi}) \propto \rho(ds) P_0(d\boldsymbol{\xi})$, so that the $\tilde{\mu}$ is *homogeneous*, this ensures that the π_h and $\boldsymbol{\xi}_h$ are independent. The Dirichlet process with parameter (M, P_0) can be identified as a special case of this class, when the underlying random measure is a gamma process, *i.e.* the Poisson random measure has the intensity measure: $\nu(ds, d\boldsymbol{\xi}) = M \frac{e^{-s}}{s} ds P_0(d\boldsymbol{\xi})$. Simulation of this

class of random probability measures can be performed by simulating the underlying independent increment process, for example by using the inverse Lévy measure algorithm (Wolpert and Ickstadt 1998b) and then normalizing.

Although the list above is quite long, there are also alternative random probability measures, which do not directly fit into the Ongaro-Cattaneo framework, for example Pólya trees, neutral to the right processes and the logistic Gaussian process (see Walker et al. (1999) and Müller and Quintana (2004)). We will briefly discuss the logistic Gaussian process in the next section. A relatively recent addition to the list of random probability measures is the quantile pyramid as introduced by Hjort and Walker (2009), who directly specify a prior on all quantiles simultaneously to build a prior distribution for a probability measure.

As mentioned in the introduction for setting up prior distributions it is important to be able to calculate prior moments, such as the prior mean and the prior covariance. Then it is possible to adjust the prior mean to center the model at a particular probability measure and associate a variability (*i.e.* uncertainty) statement with it. We will now consider the first two moments of the discrete random probability measure introduced in Definition 2.1.2, hence the results are valid for the classes of probability measures mentioned in Remark 2.1.1.

Theorem 2.1.2.
Let P be distributed according to the Ongaro-Cattaneo random probability measure $\mathbb{P}$ defined in Definition 2.1.2, then for every $B_1, B_2 \in \mathcal{B}$ we have

$$\begin{aligned} \mathrm{E}(P(B_1)) &= P_0(B_1) \\ \mathrm{Cov}(P(B_1), P(B_2)) &= k_0(P_0(B_1 \cap B_2) - P_0(B_1)P_0(B_2)), \end{aligned}$$

where $k_0 = E\left(\sum_{h=1}^{N} \pi_h^2\right)$.

Proof. See the proof of Theorem 2.1.3 and the comments at the end of the proof.

Hence the prior mean and the prior correlation of the discrete random probability measure is determined by the distribution P_0 alone (this follows from the independence of

the weights π_h and the $\boldsymbol{\xi}_h$), while for the covariance also the weights π_h play a role, mainly through the term k_0, being the expected value of the squared weights.

The squared weights always lie in $[0,1]$: The upper bound 1 follows from the fact that $\sum_{h=1}^{N} \pi_h = 1$ and $\pi_h \in [0,1]$, and the lower bound follows from an application of the Cauchy-Schwarz inequality (when $N < \infty$ the lower bound is $1/N$). Hence it is interesting to investigate for which values of π_h one obtains the largest and smallest variability: The upper bound (maximum variability) is achieved when there is only one component in the mixture with non-zero weight, while the lower bound is obtained when there are many components in the mixture and all have the same weight. In this context it is interesting to note that for the Dirichlet process $k_0 = 1/(M+1)$, for a finite mixture model with N components and a symmetric Dirichlet distribution with parameter $\gamma > 0$ for the weights, $k_0 = \frac{\gamma+1}{N\gamma+1}$. The parameter k_0 also has an interesting interpretation, when one uses random probability measures for clustering, where it is the prior probability, that two observations are combined in one cluster, see Fritsch and Ickstadt (2009) for details.

Figures 2.1 and 2.2 also nicely illustrate the role of k_0 for the Dirichlet process: When $M = 1$, *i.e.* $k_0 = 0.5$, there are only a few large weights and many almost zero weights and consequently a larger variability (see the left hand sides of Figures 2.1 and 2.2), while the larger $M = 20$ (hence $k_0 = 1/21$) leads to realizations, where more components have non-negligible probability mass and the prior has a smaller variability (see the right hand sides of Figures 2.1 and 2.2). So the parameter k_0 also determines the number of elements in the mixture, which receive relevant probability mass.

When a mixture model is used, one needs a slight variation of Theorem 2.1.2, to calculate prior mean and prior covariance. Suppose one mixes an integrable function $g(z, \boldsymbol{\xi})$, $g : \mathcal{Z} \times \Xi \mapsto \mathbb{R}$, with respect to a mixing distribution. The following Theorem gives prior mean and prior covariance of $G(z) = \int g(z, \boldsymbol{\xi}) P(d\boldsymbol{\xi})$, when the Ongaro-Cattaneo random probability measure is assumed for the mixing distribution.

Theorem 2.1.3.

The expectation of $G(z)$ and the covariance of $G(z_1)$ and $G(z_2)$ for $z, z_1, z_2 \in \mathcal{Z}$, under the

Ongaro-Cattaneo random probability measure (see Definition 2.1.2) are given by

$$\mathrm{E}(G(z)) = \int_{\Xi} g(z, \boldsymbol{\xi}) P_0(d\boldsymbol{\xi}) \tag{2.4}$$

$$\begin{aligned}\mathrm{Cov}(G(z_1), G(z_2)) = k_0 \Bigg\{ & \int_{\Xi} g(z_1, \boldsymbol{\xi}) g(z_2, \boldsymbol{\xi}) P_0(d\boldsymbol{\xi}) \\ & - \int_{\Xi} g(z_1, \boldsymbol{\xi}) P_0(d\boldsymbol{\xi}) \int_{\Xi} g(z_2, \boldsymbol{\xi}) P_0(d\boldsymbol{\xi}) \Bigg\}\end{aligned} \tag{2.5}$$

where $k_0 = E\left(\sum_{h=1}^{N} \pi_h^2\right) \in [0, 1]$.

Proof:

$$\mathrm{E}(G(z)) = E\left(\sum_{h=1}^{N} \pi_h g(z, \boldsymbol{\xi}_h)\right).$$

Conditional on N and π_h the above expectation would be equal to

$$\sum_{h=1}^{N} \pi_h \int_{\Xi} g(z, \boldsymbol{\xi}) P_0(d\boldsymbol{\xi}) \,,$$

because of the independence of $(N, \pi_1, \pi_2, \ldots)$ and $\boldsymbol{\xi}_1, \boldsymbol{\xi}_2, \ldots$. As $\sum_{h=1}^{N} \pi_h = 1$ regardless of a specific realization, it follows that the above expectation is equal to

$$\int_{\Xi} g(z, \boldsymbol{\xi}) P_0(d\boldsymbol{\xi}).$$

To obtain the covariance between two points z_1 and z_2 one needs to calculate

$$\begin{aligned}\mathrm{E}(G(z_1)G(z_2)) &= E\left(\sum_{h=1}^{N} \pi_h g(z_1, \boldsymbol{\xi}_h) \sum_{h=1}^{N} \pi_h g(z_2, \boldsymbol{\xi}_h)\right) \\ &= E\left(\sum_{h=1}^{N} \pi_h^2 g(z_1, \boldsymbol{\xi}_h) g(z_2, \boldsymbol{\xi}_h) + \sum_{\substack{h=1 \\ h \neq j}}^{N} \sum_{j=1}^{N} \pi_h \pi_j g(z_1, \boldsymbol{\xi}_h) g(z_2, \boldsymbol{\xi}_j)\right).\end{aligned}$$

If again N and the π_h were known, it follows from the independence of $(N, \pi_1, \pi_2, \ldots)$ and $\boldsymbol{\xi}_1, \boldsymbol{\xi}_2, \ldots$ and the independence of $\boldsymbol{\xi}_i$ and $\boldsymbol{\xi}_h$ that the above expression would be equal to

$$k \int_{\Xi} g(z_1, \boldsymbol{\xi}) g(z_2, \boldsymbol{\xi}) P_0(d\boldsymbol{\xi}) + (1 - k) \int_{\Xi} g(z_1, \boldsymbol{\xi}) P_0(d\boldsymbol{\xi}) \int_{\Xi} g(z_2, \boldsymbol{\xi}) P_0(d\boldsymbol{\xi}),$$

where $k = \sum_{h=1}^{N} \pi_h^2$. From this expression the covariance given in Theorem 2.1.3 can easily be calculated. The expressions given in Theorem 2.1.2 can be obtained by using indicator functions, for example $g(z, \boldsymbol{\xi}) = \mathbb{1}_B(\boldsymbol{\xi})$. □

Hence the main conclusions from Theorem 2.1.2 remain valid: The first moment of the mixed distribution $G(z)$ is only affected by the distribution P_0, while the term k_0 plays an important role in the covariance.

As discussed in the introduction an important issue in nonparametric Bayesian analysis is the full support property. The following Theorem proved in Ongaro and Cattaneo (2004), answers the question, when the Ongaro-Cattaneo random probability measure has full support on Ξ (in the chosen distance measure).

Theorem 2.1.4. *(Ongaro and Cattaneo 2004)*
Let $\hat{\mathbb{S}}^m, = \{(\pi_1, \ldots, \pi_m)' \in \mathbb{R}^m | \sum_{h=1}^m \pi_h \leq 1,\ \pi_h \geq 0,\ h = 1, \ldots, m\}$ *and suppose that* N *is an unbounded random variable, i.e.* $\mathbb{P}(N > m^*) > 0$ *for any* $m^* \in \mathbb{N}$, *and that one of the two conditions hold*

(i) for any $l > 1, \exists m(l), l \leq m(l) < \infty$, *such that* $Pr(N = m(l)) > 0$ *and the conditional distribution of* $(\pi_1, \ldots, \pi_{m(l)-1})|N = m(l)$ *has positive Lebesgue density on* $\hat{\mathbb{S}}^{m(l)-1}$.

(ii) $\mathbb{P}(N = \infty) > 0$ *and the conditional distribution of* $(\pi_1, \ldots, \pi_{m^*})|N = \infty$ *admits positive Lebesgue density on the set* $\hat{\mathbb{S}}^{m^*}$ *for any* $m^* \geq 1$.

Let Q *be a probability measure on* Ξ *which is absolutely continuous with respect to* P_0, *i.e.* $P_0(B) = 0$ *implies* $Q(B) = 0$, *for all* $B \in \mathcal{B}$. *Then for any partition* $B_1, \ldots, B_k$ *of* Ξ *we have for any* $\epsilon > 0$

$$Pr(P : |P(B_i) - Q(B_i)| < \epsilon,\ i = 1, \ldots, k) > 0,$$

where "Pr" denotes the probability under the random probability measure from Definition 2.1.2.

This result says that the random probability measure has positive prior probability on all probability distributions Q on Ξ, which are absolutely continuous with respect to P_0. This is important, as it ensures a large support on the prior distribution. A similar kind of result already appears in Ferguson (1973), for the special case of the Dirichlet process. Of course Theorem 2.1.4 also covers this case, as the Dirichlet process is a special case of the general Ongaro-Cattaneo random probability measure.

2.1.1 Example: Modelling Stochastically Ordered Distributions

In this section we will illustrate the use of the Dirichlet process model in a practical example. For this purpose data from the US Perinatal Project will be used, which were already displayed in Figure 1.2. The project was conducted from 1959 to 1966, where the exposure to environmental, social and chemical risk factors was measured for pregnant females to evaluate the potential influence on the health status of the newborn, see Longnecker et al. (2001) or Longnecker, Klebanoff, Brock and Guo (2005).

Suppose one would like to nonparametrically estimate the conditional distribution of the birth weight of newborns corresponding to smoking and non-smoking mothers. It is well known that the smoking status of the mother influences the birth weight of the child (*i.e.*, smoking mothers get babies which weight less). This type of information should be incorporated when estimating both distributions. The desired shape constraint is hence a stochastic ordering (see Appendix A.1). In the notation of the introduction we hence model the residual distribution for smoking and non-smoking mothers and $\boldsymbol{x}$ is simply a categorical variable with values "smoking" and "non-smoking", subsequently abbreviated as "s" and "ns". Nonparametric modelling of stochastically ordered distributions received some attention in recent years see Gelfand and Kottas (2000), Hoff (2003a), Karabatsos and Walker (2007), or Dunson and Peddada (2008). The model we present here, is relatively simple and has been implemented in the `OpenBUGS` software (see `http://mathstat.helsinki.fi/openbugs`) using the `R2WinBUGS` interface (Sturtz, Ligges and Gelman 2005) to the `R` statistical computing language (R Development Core Team 2009).

The main idea is to assume a hierarchical extension of the Dirichlet process: We model the residual density as

$$f_x(y) = \int \phi(y, \mu(x), \sigma^2) P(d\mu(x)) = \sum_{h=1}^{\infty} \pi_h \phi(y, \mu_h(x), \sigma^2)$$

for $x \in \{s, ns\}$ and where $\phi(y, \mu, \sigma^2)$ denotes the normal density with mean μ and variance σ^2. The function μ_h is defined as $\mu_h(x) = \beta_{0h} + \beta_{1h} \mathbb{1}_{\{ns\}}(x)$, with $\beta_{0h} \in \mathbb{R}$ and $\beta_{1h} \in \mathbb{R}_+$. As β_{1h}, is positive this directly assures stochastic ordering of the two residual distributions. For the mixing measure $P(d\mu(x))$ we assume a Dirichlet pro-

cess prior. Probabilistically this makes the residual densities dependent as both residual distributions contain the same normal density components and the same mixing weights, only the means of the normal distributions are shifted. This dependence is desired as it allows for a borrowing of strength effect, while remaining relatively flexible for the two residual distributions (the shape of the two distributions can still be different).

In this particular case the space Ξ on which we build the Dirichlet process consists of the functions of form $\beta_0 + \beta_1 \mathbb{1}_{\{ns\}}(x)$, which are characterized by $(\beta_0, \beta_1)' \in \mathbb{R} \times \mathbb{R}_+$. As the base measure P_0 of the Dirichlet process we choose a normal distribution for β_0 and an independent exponential distribution for β_1, *i.e.* $P_0 = N(m_0, \sigma_0^2) \times Exp(\lambda)$. For the parameters m_0 and σ_0^2 of the normal distribution we use weakly informative hyperprior distributions (although certainly prior information from the literature might be available for this problem): A $N(0, 10^6)$ hyperprior distribution was used for m_0, while for σ_0^{-2} a relatively flat gamma prior distribution was used with parameters $a = 0.01$ and $b = 0.01$, *i.e.* $\sigma_0^{-2} \sim \text{gamma}(0.01, 0.01)$. The parameter λ of the exponential distribution and the inverse of the bandwidth parameter σ^{-2} also receive a $\text{gamma}(0.01, 0.01)$ prior distribution. For the parameter M of the Dirichlet process a gamma prior distribution with parameter $a = 1$ and $b = 1$ was employed. The prior mean and prior variance for M are hence equal to one. The parameter M of the Dirichlet process measures the prior precision as well as how many components receive relevant probability mass (see Theorem 2.1.3 and the discussions thereafter). As we want to use only weakly informative prior distributions in this setting and do not expect exceedingly many components in the mixture we use a prior which favors relatively small values for M.

This model was implemented in `OpenBUGS` using the finite dimensional stick-breaking approximation of the Dirichlet process (using $N = 20$ as a cutoff point). `OpenBUGS` was then run for 20000 iterations after a burn-in of 2000 and using a thinning rate of 2. Summaries of the resulting 10000 iterations can be observed in Figure 2.3. There one can observe the pointwise quantiles of the posterior distribution of the densities, for smokers and non-smokers. The Bayesian nonparametric estimates are relatively smooth with a slightly more pronounced left tail than that of a Gaussian distribu-

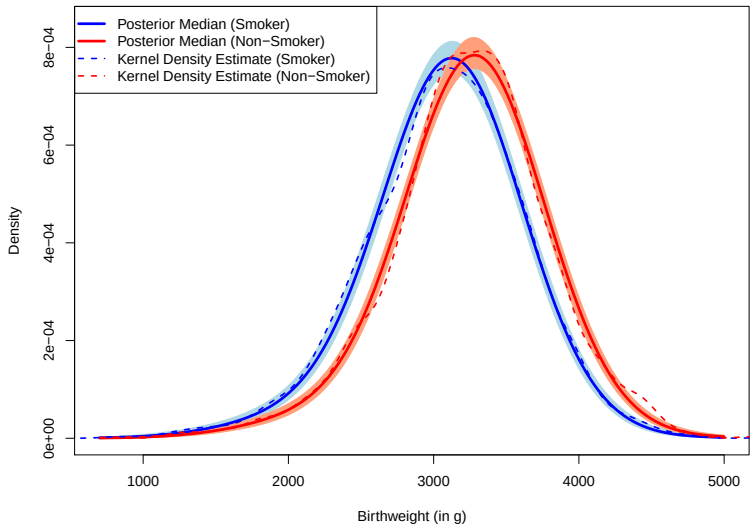

Figure 2.3: Pointwise 0.05, 0.5 and 0.95 quantiles of the posterior distribution of the residual densities for smoking (blue) and non-smoking (red) mothers.

tion. This can be explained by pre-term births, which obviously have a smaller birth weight. The shape of the residual distribution is only slightly different between the two groups as the density is a bit more peaked for non-smokers. Also embedded in the plot are the traditional kernel density estimates based on the `density` function in `R` (using the default bandwidth selection and a Gaussian kernel). It becomes obvious that the kernel density estimate is much more wiggly than the Bayesian nonparametric estimate, and fluctuates around the posterior median. The smoothness of the Bayesian estimate is clearly due to the strong dependence imposed by the hierarchical model, allowing to borrow strength between both estimates of the residual density estimates, while the kernel density estimates are only using data from either the smoking or the non-smoking class.

From the application viewpoint it is interesting to note that there is a remarkable shift of around 100-200g in the birth weight between babies for smoking and non-smoking mothers. A more sophisticated analysis would obviously adjust for possible confounding covariates, such as social status for example, nevertheless this result is quite a strong outcome of the study.

2.2 Priors for Functions

There is a great variety of approaches to build nonparametric prior distributions for functions, but in contrast to the case of probability measures in the last section there is no dominant unifying framework. Hence we will present a selective review of methods, focusing on those, which will turn out to be useful in later chapters. In particular we will not discuss models used in survival analysis (see Ibrahim, Chen and Sinha (2001) for a review of BNP methods in this area), and Bayesian extensions of traditional machine-learning methods, such as trees, see Denison, Holmes, Mallick and Smith (2002) for a review of BNP methods in this area (an interesting machine-learning reference is also Bishop (2006), which treats many machine-learning methods from a Bayesian perspective). In the last section priors for probability measures on a general measurable space were introduced. When building prior distributions for functions, however, one is typically only interested in functions defined on a subset of $\mathbb{R}^k$, rather than functions on arbitrary spaces, which is why we will focus here on this case.

This section is split up into three subsections. First we will discuss the Gaussian process as a prior for (possibly multivariate) functions. Then we give an overview over the diverse area of basis function approaches and more generally approaches based on dictionaries. Here we will first consider the univariate case and later generalize this to the multivariate case, where we will also introduce the class of ridge functions.

2.2.1 Gaussian Processes

The first and most important nonparametric prior distribution for functions is the Gaussian Process, see, for example, O'Hagan and Forster (2004, p. 393–398) and Rasmussen and Williams (2006) for reviews, or Bornkamp (2006), where it is employed for dose-response analysis. So how to build a prior distribution for a function $\mu(.) : \mathcal{X} \mapsto \mathbb{R}$? The main idea underlying the Gaussian process is to model $(\mu(\boldsymbol{x}_1), \ldots, \mu(\boldsymbol{x}_t))'$ for any $t \in \mathbb{N}$ and any $\boldsymbol{x}_i \in \mathcal{X}\ \forall i \in \{1, \ldots, t\}$ jointly as a multivariate normal distribution with mean $(m(\boldsymbol{x}_1), \ldots, m(\boldsymbol{x}_t))'$ determined by a mean function $m(.) : \mathcal{X} \mapsto \mathbb{R}$ and covariance matrix $\boldsymbol{\Gamma}$ with entries $\gamma(\boldsymbol{x}_i, \boldsymbol{x}_j)$ for $i = 1, ..., t;\ j = 1, ..., t$

determined by a covariance function $\gamma(.,.)$. If this holds for any $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_t$ with $t \in \mathbb{N}_+$ the prior distribution for $\mu(.)$ is a Gaussian process. Hence the Gaussian process can be seen as a generalization of the multivariate normal distribution to the infinite case.

The mean and the covariance function determine the mathematical properties of the realizations of a Gaussian process, for example continuity or differentiability (see Adler (1981) for details). Probably the most famous Gaussian process, univariate Brownian motion, for example, has mean function 0 and the non-smooth covariance function $\gamma(x, x') = \min(x, x')$, so that its sample path are also non-smooth. When building a prior distribution for a smooth function the covariance function should be chosen such that the modelled function is also smooth. However, one cannot use any smooth function as a covariance function: The covariance matrix $\boldsymbol{\Gamma}$ with entries $\gamma(\boldsymbol{x}_i, \boldsymbol{x}_j)$ for $i = 1, ..., t$; $j = 1, ..., t$ needs to be positive semi definite for any $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_t$ with $t \in \mathbb{N}$. Functions possessing this property are called positive semi definite (see Cheney and Light (1999)). As $\boldsymbol{\Gamma}$ needs to be symmetric, the covariance function also needs to satisfy $\gamma(\boldsymbol{x}, \boldsymbol{x}') = \gamma(\boldsymbol{x}', \boldsymbol{x})$. Although this is not a requirement for covariance functions per se, $\gamma(.,.)$ also often satisfies $\gamma(\boldsymbol{x}, \boldsymbol{x}') = \gamma(\boldsymbol{x} - \boldsymbol{x}')$ in particular when $\gamma(.,.)$ is used for building a prior distribution for functions. This assumption implies that the underlying Gaussian process is (covariance) stationary. Stationarity is a convenient assumption, because the covariance between the inputs then only depends on the distance between the inputs, allowing for a simple parametrization of the covariance function. The main practical assumption (and restriction) underlying stationarity is that the covariance of $\mu(\boldsymbol{x})$ and $\mu(\boldsymbol{x}')$ does only depend on the difference $\boldsymbol{x} - \boldsymbol{x}'$ (and hence the distance), but it does not differ in different regions of $\mathcal{X}$. The most popular covariance function in practical applications is probably the Gaussian covariance function. For one dimensional inputs it is given by $\gamma(x, x') = \tau^2 \exp(-b(x - x')^2)$ with $b \geq 0$. For multivariate inputs $\boldsymbol{x}$ typically products of functions of this form are used as a covariance function: $\gamma(\boldsymbol{x}, \boldsymbol{x}') = \tau^2 \exp(-\sum_{i=1}^{k} b_i(x_i - x_i')^2)$, where x_i denotes the $i-$th component of $\boldsymbol{x}$. The Gaussian covariance function ensures that the realizations of the Gaussian process are infinitely often differentiable, provided the mean function is also infinitely often differentiable. In most applications the prior mean function is modelled as a linear regression model, *i.e.* $m(\boldsymbol{x}) = \beta_0 + \beta_1 x_1 + \ldots + \beta_k x_k$, sometimes

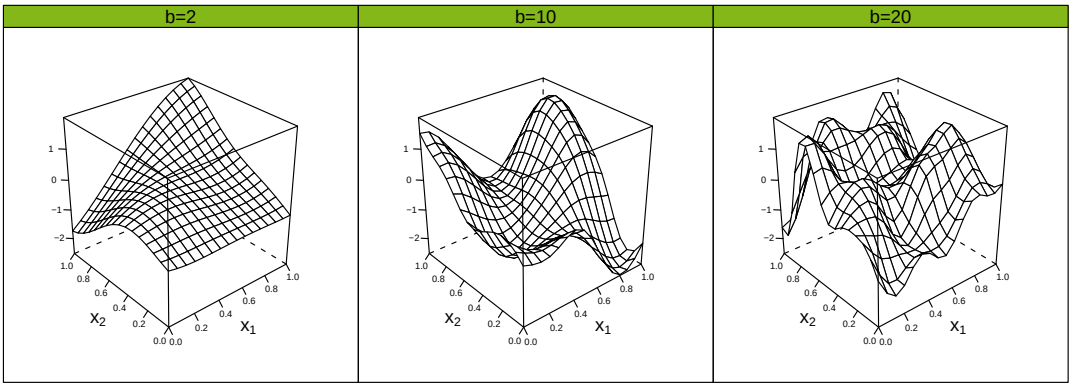

Figure 2.4: Simulation from a Gaussian process prior with mean function 0 and covariance function $\gamma(\boldsymbol{x}, \boldsymbol{x}') = \exp(-\sum_{i=1}^{2} b(x_i - x_i')^2)$ for different values of b.

also with higher order terms and interactions, in some applications however only an intercept term is used, *i.e.* $m(\boldsymbol{x}) = \beta_0$.

In Figure 2.4 one can observe simulated Gaussian processes with a constant mean function $m(\boldsymbol{x}) = 0$ and Gaussian covariance function $\gamma(\boldsymbol{x}, \boldsymbol{x}') = \exp(-\sum_{i=1}^{2} b(x_i - x_i')^2)$ for different values of b. One can see that for small values of b (high correlation within the function) the function is relatively smooth, while for larger values of b (small correlation within the function) the function gets more wiggly. Note that the amount of smoothness stays the same in all of $\mathcal{X}$, which is due to the covariance stationarity of the process.

When $m(\boldsymbol{x})$ and $\gamma(\boldsymbol{x}', \boldsymbol{x})$ do not depend on any unknown hyper-parameters, the Gaussian process prior is conjugate if the residual distribution is a normal distribution, for example, if the observations are independent and $\mathcal{P}_{\boldsymbol{\theta}_x} = N(\mu(\boldsymbol{x}), \sigma^2)$, see O'Hagan and Forster (2004). When a linear regression model is assumed for the mean function (with Gaussian hyperpriors for the linear model coefficients) this conjugacy is still preserved, but if more hyperparameters are treated as unknown, for example in the covariance function (*e.g.* the parameter b in the Gaussian correlation function), MCMC techniques need to be used to analyse the model. Even though the conjugacy properties only hold for normally distributed data, Gaussian processes are also used when the residual distribution is not normal, in particular for Bernoulli distributed data (*i.e.* classification), see Rasmussen and Williams (2006) for details. Another important ap-

plication of Gaussian process priors is multivariate interpolation, for example, in geostatistics or the analysis of deterministic computer experiments (Currin, Mitchell, Morris and Ylvisaker 1991). In these communities this approach is also known under the name kriging or Bayesian kriging. Although developed from entirely different conceptual perspectives the Gaussian process approach has also close connections to neural nets and support vector machines, or when used for interpolation, radial basis function interpolation and interpolation splines, depending on the choice of the covariance function (see Kracker, Bornkamp, Kuhnt, Gather and Ickstadt (2010) for details).

An important issue is to investigate the support properties of Gaussian process priors. Here it turns out, that again the covariance function plays the most important role. Tokdar and Ghosh (2007) investigate the support properties of Gaussian processes with respect to the class of continuous functions on $\mathcal{X}$. They show that when $\mu(.)$ has a Gaussian process prior with a covariance function satisfying a set of certain (relatively weak) conditions that $Pr(\sup_{x \in \mathcal{X}} |\mu(\boldsymbol{x}) - \mu_0(\boldsymbol{x})| < \epsilon) > 0$ for all $\epsilon > 0$ and for any continuous function $\mu_0(\boldsymbol{x})$. Here "Pr" denotes the probability under the prior Gaussian process. The Gaussian covariance function described above satisfies these conditions, but also many other covariance functions commonly used in practice are covered by the result of Tokdar and Ghosh (2007). The Gaussian process hence has full support on the space of continuous functions on $\mathcal{X}$ in sup-norm, provided a suited covariance function is used.

Theoretically the Gaussian process hence has many desirable properties, but there are also practical limitations of the traditional (*i.e.* covariance stationary) Gaussian process prior. The stationarity assumption is too restrictive in quite a few situations. That means that often, for example in situations as simple as dose-response analysis, the underlying function does not have the same smoothness properties throughout all of the input space $\mathcal{X}$. In some parts the function might be quite wiggly and non-smooth, while in others it might be completely flat. This behaviour is relatively difficult to model with a stationary Gaussian process prior. Gramacy and Lee (2008) overcome this issue by using a model that partitions the input space $\mathcal{X}$ according to its smoothness and fits separate Gaussian process priors in each partition. There have also been more direct approaches to overcome the stationarity assumption by using

non-stationary covariance functions; see Xiong, Chen, Apley and Ding (2007) for a recent review.

From the viewpoint of shape constrained inference, Gaussian processes are somewhat less interesting, because Gaussian processes are—by the properties of the normal distribution—not guaranteed to be positive or monotonic. Nevertheless they have also been used for modelling positive functions, for example in financial applications, where one typically models positive quantities, such as stocks. Here one uses a Gaussian process for the logarithm of the quantity (geometric Brownian motion). Another example is the so-called logistic Gaussian process already mentioned in Section 2.1, where a univariate continuous probability density function (*i.e.* a positive function integrating to 1) is modelled as $\frac{\mu(x)}{\int \mu(x)dx}$, and a Gaussian process prior is assumed for $\log(\mu(.))$ (see Müller and Quintana (2004) for further references regarding the logistic Gaussian process).

2.2.2 Basis Function Approaches and Dictionaries

In this section we will introduce a quite general and very versatile alternative approach for building prior distribution for functions, based on basis functions and more generally dictionaries. For ease of exposition, we will focus first on one dimensional functions $\mu(.) : \mathcal{X} \mapsto \mathbb{R}$, where $\mathcal{X}$ is given by an interval $\mathcal{X} = [a, b],\ a, b \in \mathbb{R}$. Later in this section we will generalize this to the multivariate case.

Basis Function Approaches and Dictionaries in One Dimension

The main idea is simple: One models $\mu(.)$ as a linear combination of basis functions $b_1(x), b_2(x), \ldots$, for example basis functions for polynomials, splines, wavelets or any other function basis of interest. The function $\mu(.)$ is then given by $\sum_{j=1}^{J} \beta_j b_j(x)$, where prior distributions are assumed for β_j and J (sometimes J is also fixed at a particular value), see Denison et al. (2002) for an extensive treatment of these type of models. From the theoretical side it is of interest, when such a model is truly nonparametric. That means: When is there a sequence of $\beta_1, \beta_2, \ldots$ and $b_1(x), b_2(x), \ldots$ such that a metric, for example the sup-metric $\sup_{x \in \mathcal{X}} \left| \sum_{j=1}^{J} \beta_j b_j(x) - \mu_0(x) \right|$, converges to zero, when

$J \to \infty$, for any continuous function $\mu_0(.)$. And more importantly: Does the prior distribution have positive prior probability on this approximating sequence of functions? Certainly this depends on the chosen basis and the assumed prior distribution. For many types of commonly used bases, the approximation property holds, see for example Cheney and Light (1999) or Goodman (1995), who explicitly covers polynomials and splines.

When the parameter J is fixed at a particular value (usually a relatively large value) instead of being treated as unknown, this is theoretically a parametric model: It is finite dimensional and typically does not have full support, for example, on the space of continuous functions. Nevertheless if J is chosen large enough, often a rich class of functions can be approximated fairly well, while maintaining the simple computations associated with a parametric model. This advantage can in some situations outweight the gain obtained by using a fully nonparametric model. In fact these type of models have been applied successfully in a wide variety of applications (see, for example, Lang and Brezger (2004)). Before moving on to extensions of the basis function approach, we want to note that also the Gaussian process approach discussed in the last chapter can be seen as a special case of the basis function approach: Due to the Karhunen-Loève expansion one can represent a Gaussian process as an (infinite) sum of orthogonal basis functions (derived from the covariance function of the Gaussian process) with independent normal priors for the coefficients (see Clyde and Wolpert (2007)).

From the discussion above it becomes apparent, that the basis function approach is related to the mixture idea presented in Section 2.1, because one uses a linear combination of functions to model another possibly more complex function. The difference is that in the basis function approach, we only learn the coefficients β_j of the basis functions and possibly their number J. The general mixture approach from the last section, however, learns the coefficients, the number of "basis functions", as well as parameters of the "basis functions" which appear within the linear combination. Hence in a way, the mixture approach is more flexible than the traditional basis function approach and has been extended, for example, by Wolpert and Ickstadt (1998a), Wolpert and Ickstadt (2004) and Clyde and Wolpert (2007) to model general functions (instead

of probability densities). In this approach one hence prespecifies a generating function $b(x, \boldsymbol{\xi}) : \mathcal{X} \times \Xi \mapsto \mathbb{R}$ and then models the function as

$$\int b(x, \boldsymbol{\xi}) dL(d\boldsymbol{\xi}) = \sum \beta_j b(x, \boldsymbol{\xi}_j), \tag{2.6}$$

where $L(d\boldsymbol{\xi})$ is a discrete signed measure on Ξ (see Dudley (2002, p. 178) for details on signed measures). The difference to the mixture model in Section 2.1, is the fact that $L(d\boldsymbol{\xi})$ is not normalized and the β_j may be positive or negative. The main advantage over the traditional basis function approach is that also the "basis functions" themselves are learned from the data (as the parameters $\boldsymbol{\xi}_j$ are treated as unknown). This has the potential to lead to sparse representations, *i.e.* typically only very few generating functions are needed to represent even highly complex functions. This sparseness obviously has advantages in high-dimensional situations but also in terms of interpretation. The parameters $\boldsymbol{\xi}_j$ might, for example, have a particular meaning in the application context, which one would miss, when simply using "many" basis functions. We will call the set of functions obtained from the generating function a dictionary (the name is adopted from the wavelet literature, see Clyde and Wolpert (2007), but there appears to be no clear cut mathematical definition for the word dictionary in this context). Dictionaries typically do not form a basis for a particular function space, but are overcomplete in the sense that one function might be represented in different ways by the same dictionary. This redundancy is not a disadvantage but allows for a sparser representation of the modelled function: When there are several ways to represent a function, one can choose the one with fewest elements in the linear combination.

Note that there are also constructions, which are half-way between the (fixed) basis and the dictionary approach: Some bases have additional parameters, which might be treated as unknowns instead of being known. The knots in a spline basis, for example, can be treated as unknown. This type of approach is more flexible than a fixed basis approach, but does not directly fit in the dictionary framework outlined above.

To illustrate the richness of a dictionary versus a (fixed) basis approach we approximate the function $\mu(x) = 1.4x^3/(x^3 + 0.1^3) + 0.6x^{10}/(x^{10} + 0.8^{10})$ in Figure 2.5 by an 11-dimensional quadratic B-spline basis with equally spaced inner knots (see Dierckx (1993) for details on B-splines) and a two-component dictionary of the form

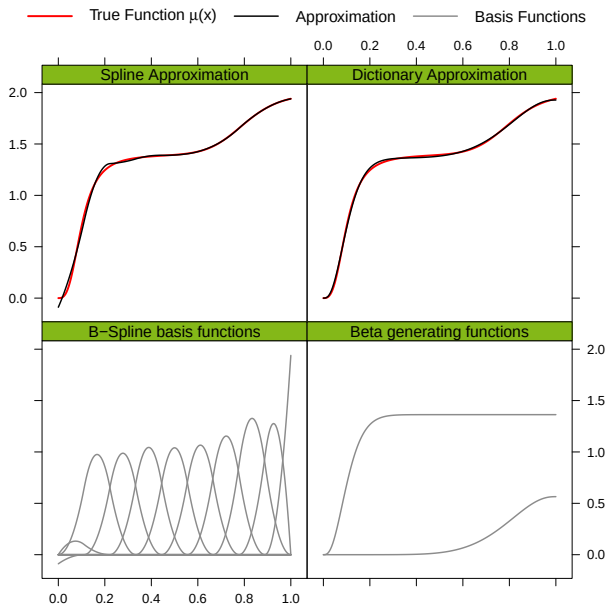

Figure 2.5: Spline and dictionary least squares approximations of the function $\mu(x)$. The lower row shows the basis functions each multiplied with its coefficient β_j.

$\sum_{i=1}^{2} \beta_j B(x, a_j, b_j)$, where the generating function $B(x, a_j, b_j)$ is the cdf of the beta distribution with parameters $a_j, b_j > 0$. In both cases we minimized the least squares distance, evaluated at 201 equally spaced points in $[0, 1]$, between the true function and the approximant. For the spline basis we hence needed to minimize with respect to the 11-dimensional vector $(\beta_1, \ldots, \beta_{11})'$ while for the dictionary we minimized with respect to β_j and a_j,b_j (hence 6 parameters in total). In Figure 2.5 one can observe that both approaches approximate the true function fairly well. The B-spline basis however requires 11 basis functions, while the dictionary needs only 2 generating functions to obtain a similar fit. This situation is typical: Dictionaries can often fit complex functions quite well with a smaller number of elements in the linear combination, because the "basis functions" can be adapted to the problem at hand. This effect becomes even more important, when modelling multivariate functions. On the other hand already in this simple deterministic approximation example it becomes apparent that the dictionary approach is often computationally more complex: The coefficients of the B-spline basis can be determined by solving a linear least squares problem, while for the dictionary approach the involved optimization problem is a nonlinear least squares problem.

Regarding shape constrained inference, the basis function and the dictionary approach turn out to be quite useful: When the basis functions $b_j(x)$ or the generating function $b(x, \boldsymbol{\xi})$, for example, are positive functions (a weak assumption), then one can model a positive function as $\sum_{i=1}^{J} \beta_j b_j(x)$ or $\sum_{i=1}^{J} \beta_j b(x, \boldsymbol{\xi}_j)$, where $\beta_j \geq 0$ for all i. Having a model for positive functions one can straightforwardly build models for monotonic differentiable functions by integrating the basis functions and convex twice differentiable functions by integrating once more. In Chapters 3 and 4 we will use an approach quite similar to these ideas to model monotonic and convex functions, based on a dictionary approach. From the basis function approaches particularly the B-spline basis seems to be interesting for shape constrained inference. Here shape constraints such as monotonicity, convexity or unimodality can directly be reduced to constraints on coefficients of the basis (see Goodman (1995)). It seems that the potential of B-splines in nonparametric Bayesian shape constrained inference has not fully been exploited yet, although Bornkamp and Ickstadt (2009b) use the shape preserving properties of the B-spline basis for elicitation of probability distributions, *i.e.* fitting flexible (possibly shape constrained) distributions to probability statements stated by an expert.

Multivariate Approaches based on Basis Functions and Dictionaries

The easiest way of generalizing the univariate basis function approach to the multivariate case, are tensor products of univariate bases. If a univariate basis on $[a, b]$ is given by $(b_1(x), \ldots, b_J(x))'$ this can be extended to the bivariate case by taking the Cartesian product of the bases, *i.e.* using $b^*_{ij}(\boldsymbol{x}) = b_i(x_1)b_j(x_2),\ i, j \in \{1, \ldots, J\}$ as a basis for functions on $[a, b] \times [a, b]$, resulting in a basis of J^2 terms. In the general k dimensional setting one hence ends up with J^k terms, so the number of basis functions grows exponentially with the dimension. This is probably the main reason, why this type of approach is seldom used in statistical practice when $k > 2$, despite the fact that the mathematical approximation properties of the so formed basis are typically preserved (for example for polynomials and splines).

One commonly used approach to fight this problem of high dimensionality is to impose (reasonable) additional assumptions on the modelled function. The most commonly used assumption is to impose additivity, *i.e.* modelling a multivariate function as $\mu(\boldsymbol{x}) = \mu_1(x_1) + \ldots + \mu_k(x_k)$. This reduces the problem of estimating one k-variate

function to the problem of estimating k univariate functions, and for each dimension a univariate basis (or a dictionary) can be used. When all univariate functions are modelled with J components the number of basis functions hence only grows as kJ, and thus considerably slower than in the tensor product case. It is, however, obvious that the additivity assumption is restrictive: When looking at an additive function $\mu(.)$ as a function of one variable and varying the other $k-1$ variables, the shape or scale of the function does not change, only its location. The other $k-1$ variables hence only play the role of an intercept and possible interactions cannot be modelled by additive functions. Sometimes interactions of order two (or higher) are included in the function (modelled, for example, by a tensor product of a univariate basis), which makes this type of model more flexible. The formal justification of these type of approaches is based on the, so called, ANOVA decomposition of multivariate functions, see Owen (1998, ch. 3) for details on this topic.

Single-index models are another way to reduce the problem of dimensionality for modelling multivariate functions. Here one models a multivariate function as $\mu(\boldsymbol{x}) = \mu^*(\boldsymbol{a}'\boldsymbol{x})$, for a function $\mu^*(.), \mathbb{R} \rightarrow \mathbb{R}$ and $\boldsymbol{a} \in \tilde{\mathbb{S}}^{k-1}$, with $\tilde{\mathbb{S}}^{k-1}$ the unit sphere, or $\boldsymbol{a} \in \mathbb{R}^k$. Here multivariate regression is reduced to estimation of a univariate function and a linear combination. This functional form allows to model interactions to some extent, but the function is constant on hyperplanes of the form $\boldsymbol{a}'\boldsymbol{x} = c$. For a bivariate function this, for example, implies that its contour lines are straight lines, which is again quite restrictive.

The additive model and the single-index models can be seen as a special case of a more general, unifying approach to model multivariate functions based on linear combinations of ridge functions. A ridge function is a multivariate function defined as $g(\boldsymbol{a}'\boldsymbol{x})$, where $g : \mathbb{R} \rightarrow \mathbb{R}$ is a univariate function and $\boldsymbol{a} \in \mathbb{R}^k$ (or $\boldsymbol{a} \in \tilde{\mathbb{S}}^{k-1}$) is the so called direction vector. The final model for a function is then based on a linear combination of ridge functions $\sum \beta_j g_j(\boldsymbol{a}_j'\boldsymbol{x})$. It can be shown that linear combinations of ridge functions $\sum \beta_j g_j(\boldsymbol{a}_j'\boldsymbol{x})$ can approximate any multivariate continuous function in sup norm, provided linear combinations of the involved univariate functions g_j can approximate any univariate continuous function on $\mathbb{R}$, see Cheney and Light (1999, ch. 22) for a detailed statement of the result. Hence the limitations of the additive and the single-

index model are overcome by this more flexible class of functions. Other interesting results from approximation theory were derived in Barron (1993), who considers the rate of approximation by linear combinations of ridge functions (in terms of J) and shows that, particularly in high dimensions, linear combinations of ridge functions have advantages over fixed basis function approaches (see Theorem 1 and 6 of the paper). Before giving a review of the applications of linear combinations of ridge functions, we would like to note that the dictionary idea introduced earlier is easy to generalize to the multivariate case: Simply use a multivariate generating function $b(\boldsymbol{x}, \boldsymbol{\xi})$ in Equation (2.6). One possible choice for $b(\boldsymbol{x}, \boldsymbol{\xi})$ obviously are ridge functions, so the ridge function approach discussed here, also falls under the umbrella of dictionary approaches treated earlier.

Linear combinations of ridge functions form the mathematical basis of neural networks and projection pursuit regression. One of the simplest type of neural networks (see Lee (2004) for a discussion of neural networks from a Bayesian perspective) is the so-called feed-forward network with one hidden layer. Here a sigmoid function $s(.)$ (such as the logistic function $s(x) = 1/(1 + \exp(-x))$) is used and the function is modelled as $\sum_{j=1}^{J} \beta_j s(a_j^0 + \boldsymbol{a}_j^{1\prime} \boldsymbol{x})$, for parameters $a_j^0 \in \mathbb{R}$ and $\boldsymbol{a}_j^1 \in \mathbb{R}^k$. It is straightforward to see that this type of model fits in the ridge function approach above. Typically J is treated as fixed or one performs model selection or model averaging for a small number of different values for J. For the parameters a_j^0 and $\boldsymbol{a}_j^1$ parametric prior distributions, usually associated via a hyperprior, are used, see Lee (2004) for a review of different prior distributions for neural networks. Practically these type of models are hence parametric, as J is treated as fixed and finite dimensional, but this truncation is often practically irrelevant as long as J can get sufficiently large. Another statistical application of linear combinations of ridge functions is projection pursuit regression, here one estimates the univariate functions g_j by nonparametric regression and simultaneously the linear combinations, but there appears to be no work on adopting this idea in the Bayesian framework. In Chapter 5 we will build a nonparametric prior distribution for multivariate monotonic functions based on linear combinations of ridge functions.

Although this ridge function approach appears rather unrelated to the Gaussian pro-

cess approach discussed earlier, it can be shown that the Gaussian process prior arises from a prior distribution based on a neural network in a limiting case, see Neal (1998) for further references in this regard.

Prior Distributions based on Basis Functions and Dictionaries

When using a dictionary of form (2.6), either univariate or multivariate (for example by using ridge functions), one needs to specify a prior distribution for the discrete measure $L(d\boldsymbol{\xi})$. One idea is to use the class of Lévy random measures as advocated in Tu, Clyde and Wolpert (2008) as a prior for $L(d\boldsymbol{\xi})$. This class of prior distributions implies that the prior distribution of $L(A)$ for a subset $A \subset \Xi$ is infinitely divisible. The notion of a Lévy random measure is also intimately related to Poisson random measures, see Appendix A.2 for details. Analogously to Dirichlet process priors one can exploit certain conjugacy properties, when using particular residual distributions and a particular prior distribution of form (2.6). For example when the residual distribution is a Poisson distribution, a multivariate normal kernel is used and a gamma process is employed as prior for $L(d\boldsymbol{\xi})$, see Wolpert and Ickstadt (1998a) for details. However in its most general form, Lévy random measures do not necessarily form a conjugate class of priors. One possible way to generalize different approaches would be to propose a random (signed) measure in analogy to the Ongaro-Cattaneo construction from Section 2.1. A random discrete (signed) measure belongs to this class, when its realizations are given by

$$L(d\boldsymbol{\xi}) = \sum_{j=1}^{J} \beta_j \delta_{\boldsymbol{\xi}_j}(d\boldsymbol{\xi}), \tag{2.7}$$

where $(\beta_j, \boldsymbol{\xi}_j)$ are independent identically distributed from a non-atomic probability distribution P_0 on $\mathbb{R} \times \Xi$, and J has an independent probability distribution on $\mathbb{N} \cup \infty$. Note that this is quite similar to the Ongaro-Cattaneo construction, here however, the distributions of the "jump-heights" (β_j) and the "jump-locations" $(\boldsymbol{\xi}_j)$ are not assumed to be independent. Lévy random measures can easily be identified as a special case of this general random measure. Theoretical investigations of this type of prior are lacking, although the recent preprint of Pillai and Wolpert (2008), which focuses on Lévy random measures, seems to be an exception. Due to their similarity one might expect that the mathematical properties (*e.g.* support) are similar to the Ongaro-Cattaneo construction but we will not pursue this in more detail as this type of random measure

will not explicitly be considered further in this thesis.

2.3 Asymptotics

As mentioned in the introduction, the central object of interest in Bayesian Statistics is the posterior distribution: From the posterior one can derive the distribution of quantities of interest on which all subsequent decision making will be based. The main idea of asymptotic considerations is to fix one particular true model for the residual probability distributions, and evaluate the behaviour of the posterior distribution under drawing (hypothetical) samples from the true model, with a sample size tending to infinity.
In a practical situation, this is a purely mathematical exercise and seems to be an odd idea, as there is only one particular fixed data set and it is not obvious why one should bother with imaginary, artificial data sets of infinite size. Nevertheless asymptotic considerations can provide an external (*i.e.* non-Bayesian) validation of the used methodologies. A central question in asymptotics is, for example, consistency: Does the posterior distribution converge to a degenerate one point distribution at the true statistical model, if we have perfect information through the data (*i.e.* an infinite sample)? Something seems to be wrong with a posterior distribution, which fails to reflect this perfect information. Asymptotic considerations also provide an interesting viewpoint on BNP methodologies: The full support property, for example, appears as a rigorous requirement for consistency of a posterior distribution, while previously we only required this for intuitive reasons.

While consistency can be established under relatively weak assumptions for parametric Bayesian procedures (see for example Ghosh and Ramamoorthi (2003, ch. 1)), there appeared a disturbing paper by Diaconis and Freedman (1986), which shows in a particular example the inconsistency of a nonparametric posterior distribution in a situation with a seemingly plausible prior distribution. Only at the end of the last century (motivated by path-breaking work of Andrew Barron) the study of consistency started to flourish again, resulting in many papers in the last fifteen years. Reviews of

these developments are given by the book of Ghosh and Ramamoorthi (2003) and the review articles of Ghosal, Ghosh and Ramamoorthi (1999), Walker (2004), Choi and Ramamoorthi (2008) and the forthcoming article of Ghosal (2010). This chapter follows the development in Choi and Ramamoorthi (2008), who base their development mainly on Walker (2003).

When treating asymptotic procedures in the nonparametric setup, it is important how the distance between statistical models is measured and how to define convergence of distributions. In Appendix A.3 we give a review of statistical distance measures. Now we define two central notions needed in this section.

Definition 2.3.1 (Neighbourhood and Support)**.**

(a) The (ϵ-)neighbourhood $N_\epsilon(f)$ of a probability distribution with density f is given by the following set of densities g:

$$N_\epsilon(f) = \{g : d^*(f,g) \leq \epsilon\},$$

where $d^*(.,.)$ is the chosen distance measure and $\epsilon > 0$.

(b) The support of the prior, $\mathcal{S}$, is the set of distributions with density f, for which the prior probability of the set $N_\epsilon(f)$ is larger than zero for all $\epsilon > 0$.

Obviously the support $\mathcal{S}$ of the prior distribution depends on the chosen distance measure d^*. When the weak distance d_W is chosen, we will denote the support as $\mathcal{S}_W$, the support of the prior in Hellinger or Kullback-Leibler neighbourhoods will be denoted as $\mathcal{S}_H$ and $\mathcal{S}_{KL}$. From the results on the interrelationships of the different distance measures and the example on weak and strong neighborhoods in Appendix A.3 it follows that $\mathcal{S}_W \supset \mathcal{S}_H \supset \mathcal{S}_{KL}$. Assuming full support in Kullback-Leibler divergence is hence the strongest assumption. We will see later that this condition is already a sufficient condition for *weak* consistency of a posterior distribution (due to the famous result of Schwartz (1965)). It is, however, not a necessary condition for weak consistency (there exist counter-examples) but is frequently used in the literature despite this fact. Walker (2004), for example, notes that full support in Kullback-Leibler distance is currently accepted as the fundamental property for establishing consistency. Interestingly full support in weak neighbourhoods is not enough for weak consistency:

The prior in the example of Diaconis and Freedman (1986) has full support in weak neighbourhoods, but the posterior distribution fails to be weakly consistent. When interest is in *strong* consistency more than full support in Kullback-Leibler neighborhoods is required. Usually one needs an additional *second* assumption, which somehow restricts the model. This second assumption is typically very model specific, and a variety of approaches exist in the literature. It turns out that this assumption—in some situations—is not necessary or automatically fulfilled, when there are shape constraints involved in the model. At the end of this section we will illustrate this fact by some examples.

We begin by making some assumptions, not made explicitly before in this thesis and introduce some notation. We assume that we observe independent realizations $y_i \in \mathbb{R}$ from a true (residual) probability distribution $\mathcal{P}_{\theta_0}$, which does not depend on any covariates $\boldsymbol{x}$, so that the observations are independent and *identically* distributed. The restriction to independent and identically distributed realizations is mainly for exposition and the main ideas presented here carry over also to the case of non-identically distributed observations (we will consider an example of posterior consistency in the regression case in Chapter 4). Additionally we assume that $\mathcal{P}_{\boldsymbol{\theta}}$ has a density and will denote it by $f_{\boldsymbol{\theta}}$. We denote the prior distribution for the (possibly infinite dimensional) parameter $\boldsymbol{\theta}$ by Π and its posterior after n observations by Π_n^*. Bayes theorem states that the posterior probability for a set $A \subset \boldsymbol{\Theta}$ is given by

$$\Pi_n^*(A) = \frac{\int_A \prod_{i=1}^n f_{\boldsymbol{\theta}}(y_i)\Pi(d\boldsymbol{\theta})}{\int \prod_{i=1}^n f_{\boldsymbol{\theta}}(y_i)\Pi(d\boldsymbol{\theta})}. \tag{2.8}$$

We will say that a posterior distribution is consistent if

$$\Pi_n^*(U|y_1, \ldots, y_n) \to 1$$

almost surely for $n \to \infty$ and every neighborhood U of the true residual probability distribution $\mathcal{P}_{\boldsymbol{\theta}_0}$ with density $f_{\boldsymbol{\theta}_0}$. When U is a weak neighborhood, the posterior is called weakly consistent, when U is a strong neighborhood (*i.e.* a Hellinger or a total variation neighborhood), the posterior will be called strongly consistent (see Appendix A.3 for a practical illustration of the difference between weak and strong neighborhoods).

A standard trick in Bayesian asymptotics is to rewrite the ratio (2.8) as

$$\Pi_n^*(A) = \frac{\int_A \prod_{i=1}^n R(y_i)\Pi(d\boldsymbol{\theta})}{\int \prod_{i=1}^n R(y_i)\Pi(d\boldsymbol{\theta})} = \frac{J_n(A)}{I_n},$$

where $R(y_i) = \frac{f_{\boldsymbol{\theta}}(y_i)}{f_{\boldsymbol{\theta}_0}(y_i)}$ and $\boldsymbol{\theta}_0$ is the true value of the (possibly infinite dimensional) parameter. When speaking about asymptotic considerations we evaluate the behaviour of the posterior distribution under the distribution of the sample $y_1, y_2, \ldots$ distributed according to $\mathcal{P}_{\boldsymbol{\theta}_0}^{\infty} = \mathcal{P}_{\boldsymbol{\theta}_0} \times \mathcal{P}_{\boldsymbol{\theta}_0} \times \ldots$, the product measure of the true residual distributions. So the probability statements in this chapter will be made under repeated sampling from the true statistical model $\mathcal{P}_{\boldsymbol{\theta}_0}$.

The standard approach for establishing consistency now considers the numerator $J_n(A)$ and denominator I_n separately. The relatively easy part consists of showing that the denominator I_n is almost surely larger than $\exp(-c_1 n)$. The part which is typically more difficult is to establish that numerator $J_n(A)$ is almost surely smaller than $\exp(-c_2 n)$. Both results together imply that the posterior is consistent (note that the constants $c_1 > 0$ and $c_2 > 0$ can usually be chosen appropriately to ensure the convergence). We will first consider the denominator, and make use of the Kullback-Leibler support $\mathcal{S}_{KL}$.

Lemma 2.3.1.
If $f_{\boldsymbol{\theta}_0} \in \mathcal{S}_{KL}$ of the prior distribution Π then:

$$\exp(n\beta) I_n \to \infty,$$

almost surely for any $\beta > 0$.

Proof:
The version given here is a condensed version of the proof given in Ghosh and Ramamoorthi (2003, Lemma 4.4.1).

$$\int \prod_{i=1}^n R(y_i)\Pi(d\boldsymbol{\theta}) \geq \int_{K_\epsilon} \exp\left(-\sum_{i=1}^n \log\left(\frac{f_{\boldsymbol{\theta}_0}(y_i)}{f_{\boldsymbol{\theta}}(y_i)}\right)\right)\Pi(d\boldsymbol{\theta}),$$

where K_ϵ is a Kullback-Leibler neighborhood of $f_{\boldsymbol{\theta}_0}$. For each $f_{\boldsymbol{\theta}}$ in K_ϵ we thus have by the law of large numbers that

$$\frac{1}{n}\sum_{i=1}^n \log\left(\frac{f_{\boldsymbol{\theta}_0}(y_i)}{f_{\boldsymbol{\theta}}(y_i)}\right) \to K(f_{\boldsymbol{\theta}_0}, f_{\boldsymbol{\theta}}) < \epsilon$$

Hence equivalently we have for each $f_{\boldsymbol{\theta}}$ in K_ϵ

$$\exp(n2\epsilon)\exp\left(n\left(-\frac{1}{n}\sum_{i=1}^{n}\log\left(\frac{f_{\boldsymbol{\theta}_0}(y_i)}{f_{\boldsymbol{\theta}}(y_i)}\right)\right)\right)\to\infty$$

This ensures that also $\exp(n2\epsilon)I_n$ diverges. □

Definition 2.3.2 (Strong separation)**.**
Let $A\subset\boldsymbol{\Theta}$. For any probability measure ν on A, let q_ν be the prior predictive distribution of y under ν, i.e.

$$q_\nu(y)=\int_A f_{\boldsymbol{\theta}}(y)\nu(d\boldsymbol{\theta}).$$

The set A and the parameter value $\boldsymbol{\theta}_0$ are called strongly separated if for any probability measure ν on A and $\delta>0$

$$\int\sqrt{f_{\boldsymbol{\theta}_0}(y)q_\nu(y)}dy<\delta.$$

The affinity $\int\sqrt{f_{\boldsymbol{\theta}_0}(y)q_\nu(y)}dy$ is directly related to the Hellinger metric (see Appendix A.3). It gets large when the densities are similar and small, when the densities are different. So $\int\sqrt{f_{\boldsymbol{\theta}_0}(y)q_\nu(y)}dy<\delta$ implies $d_H^2(f_{\boldsymbol{\theta}_0},q_\nu)\geq 2(1-\delta)$. The utility of the notion of strong separation will become clear in the proof of Theorem 2.3.1. For an iid sample $y_1,\ldots,y_n$ we have the affinity $\int\sqrt{\prod_{i=1}^n f_{\boldsymbol{\theta}_0}(y_i)\prod_{i=1}^n q_\nu(y_i)}dy_1...dy_n$. Induction (*i.e.* integration with respect to y_{n+1} in the induction step) combined with the definition of strong separation shows that

$$\int\sqrt{\prod_{i=1}^n f_{\boldsymbol{\theta}_0}(y_i)\prod_{i=1}^n q_\nu(y_i)}dy_1...dy_n<\delta^n=:\exp(-n\beta_0),\text{ with }\beta_0:=-\log(\delta),$$

provided A and $\boldsymbol{\theta}_0$ are strongly separated. Now we have all ingredients to state the main step towards consistency as proved by Schwartz (1965).

Theorem 2.3.1 (Schwartz)**.**
If $f_{\boldsymbol{\theta}_0}\in\mathcal{S}_{KL}$

$$\Pi_n^*(A|y_1,\ldots,y_n)\overset{a.s.}{\to}0,$$

for a set $A\subset\boldsymbol{\Theta}$ strongly separated from $\boldsymbol{\theta}_0$.

Proof:
This proof is a slightly more detailed version of the proof given by Choi and Ramamoorthi (2008, Theorem 3.7). It follows from Markov's inequality that

$$P\left(\sqrt{J_n(A)}>e^{-n\gamma}\right)\leq e^{n\gamma}E_{f_{\boldsymbol{\theta}_0}}\left(\sqrt{J_n(A)}\right)$$

Now

$$\begin{aligned} E_{f_{\theta_0}}\left(\sqrt{J_n(A)}\right) &= \int \sqrt{\int_A \frac{\prod_{i=1}^n f_{\boldsymbol{\theta}}(y_i)}{\prod_{i=1}^n f_{\theta_0}(y_i)}\Pi(d\boldsymbol{\theta})}\prod_{i=1}^n f_{\theta_0}(y_i)dy_1...dy_n \\ &= \int \sqrt{\int_A \prod_{i=1}^n f_{\boldsymbol{\theta}}(y_i)\Pi(d\boldsymbol{\theta})}\sqrt{\prod_{i=1}^n f_{\theta_0}(y_i)}dy_1...dy_n \\ &= \sqrt{\Pi(A)}\int \sqrt{\int_A \prod_{i=1}^n f_{\boldsymbol{\theta}}(y_i)\widetilde{\Pi}(d\boldsymbol{\theta})}\sqrt{\prod_{i=1}^n f_{\theta_0}(y_i)}dy_1...dy_n, \quad (2.9) \end{aligned}$$

where $\widetilde{\Pi}$ is the prior distribution Π restricted to the set A *i.e.* $\widetilde{\Pi}(d\boldsymbol{\theta}) = \frac{\Pi(d\boldsymbol{\theta})}{\Pi(A)}$. This means that we can apply the strong separation property to Equation (2.9) with $\widetilde{\Pi}$ playing the role of ν in the definition of strong separation. Hence we have

$$E_{f_{\theta_0}}\left(\sqrt{J_n(A)}\right) \leq \sqrt{\Pi(A)}\exp(-n\beta_0),$$

and consequently

$$P\left(J_n(A) > \exp(-n2\gamma)\right) \leq \exp(n\gamma)\sqrt{\Pi(A)}\exp(-n\beta_0),$$

because $J_n(A)$ is positive and the square root a monotone transformation. This converges to zero for $\gamma < \beta_0$. Hence by the Borel-Cantelli Lemma one has $J_n < \exp(-n2\gamma)$ almost surely for large n, additionally from Lemma 2.3.1 we have that $I_n > \exp(-n\beta)$ for sufficiently large n and any $\beta > 0$, so that in total also the ratio converges to zero (for $\beta < 2\gamma$). □

Note that the original version of Schwartz's theorem was proved using the machinery of uniformly consistent test functions, which still play a dominant role in the study of Bayesian asymptotics. However, as mentioned in Choi and Ramamoorthi (2008) the original formulation of the theorem and the version given here are equivalent. But Theorem 2.3.1 is not enough to establish consistency of the posterior distribution directly. Until now we only proved that the posterior probability of a strongly separated set A converges almost surely to zero. But for consistency we want to know whether the posterior probability of a neighborhood of the true residual probability distribution converges to 1. We will first consider weak convergence and will use Theorem 2.3.1 for this purpose. If U is a weak neighborhood of the true model, *i.e.* $U = \{\boldsymbol{\theta} : d_W(f_{\boldsymbol{\theta}}, f_{\theta_0}) < \epsilon\}$, the main idea is now to write $U^c = \boldsymbol{\Theta}\backslash U$ as a *finite* union of strongly separated sets. To each of these finite sets one can then apply Theorem 2.3.1.

Now it follows from the properties of weak distances that $U^c = \boldsymbol{\Theta} \backslash U$ can be written as a finite union of sets of the form

$$A = \left\{ \boldsymbol{\theta} : \left| \int h(y) f_{\boldsymbol{\theta}}(y) dy - \int h(y) f_{\boldsymbol{\theta}_0}(y) dy \right| > \epsilon \right\}, \tag{2.10}$$

for bounded functions h (see, for example, Choi and Ramamoorthi (2008) and Ghosal (2010)). Additionally the sets of form (2.10) are strongly separated from $f_{\boldsymbol{\theta}_0}$, as the Hellinger distance is stronger than weak distances, see also Appendix A.3. Hence for weak convergence we are already there, because then Theorem 2.3.1 suffices for the result. Hence we have the following

Corollary 2.3.1 (Schwartz)**.**
If $f_{\boldsymbol{\theta}_0} \in \mathcal{S}_{KL}$, then

$$\Pi_n^*(U|y_1, \ldots, y_n) \overset{a.s.}{\to} 1$$

almost surely for a weak neighborhood U of $\mathcal{P}_{\boldsymbol{\theta}_0}$.

For strong neighborhoods however one typically cannot write the complement of the neighborhood as finite union of sets strongly separated from $\boldsymbol{\theta}_0$ (Choi and Ramamoorthi (2008) and Ghosh and Ramamoorthi (2003, p. 58)). Here we need additional assumptions and as mentioned before, the type of assumptions are different for different type of models. Before moving on to a closer study of strong consistency and topics in shape constrained inference we illustrate the key steps for proving consistency using Theorem 2.3.1 and Corollary 2.3.1 in a simple parametric model.

Example 2.3.1 (Illustration on a finite dimensional model)
Suppose one observes independent and identically distributed data $y_i, i = 1, \ldots, n$ generated according to a normal distribution distribution with parameter $\theta \in \mathbb{R}$ and known variance σ_0^2. In addition suppose the information prior to data collection can be expressed by a normal distribution with parameters m_0, s_0^2, and suppose $\theta_0 \in \mathbb{R}$ is the true parameter value. According to Theorem 2.3.1 one needs to check the Kullback-Leibler property as the main requirement for consistency. The Kullback-Leibler divergence between two normal distributions with equal variance is given by $K(f_\theta, f_{\theta_0}) = \frac{1}{2\sigma_0^2}(\theta - \theta_0)^2$. Hence we need to check, whether the prior distribution assigns positive prior probability to $N_\epsilon(\theta_0) = \{\theta | K(f_\theta, f_{\theta_0}) < \epsilon\}$ for any $\theta_0 \in \mathbb{R}$ and any $\epsilon > 0$.

This is the case as the normal distribution has strictly positive density on $\mathbb{R}$. The Kullback-Leibler neighborhood of the true residual density f_{θ_0} hence receives positive probability mass. According to Corollary 2.3.1 this is sufficient for weak consistency of the posterior, so $d_W(f_\theta, f_{\theta_0}) \overset{a.s.}{\to} 0$. Note that $f_{\boldsymbol{\theta}}$ is the normal density here and from the properties of the normal density we know that this can only happen if also $|\theta - \theta_0| \overset{a.s.}{\to} 0$. From this one can also infer that the posterior is strongly consistent, as it follows that $f_\theta(y) \overset{a.s.}{\to} f_{\theta_0}(y)$ for all y. This implies convergence of the posterior in total variation distance. So in this situation (and in most usual parametric settings) weak and strong consistency coincide.

Hence the key step for proving weak convergence of the posterior distribution is to check the Kullback-Leibler property of the prior, see Wu and Ghosal (2008) for a review of priors for which the Kullback-Leibler property holds. If strong consistency is desired, in most nonparametric (*i.e.* infinite dimensional) situations, more work is needed than in the finite dimensional Example 2.3.1, because strong and weak consistency usually do not coincide. We will now briefly outline the most general approach to prove strong consistency. It is based on notion of a sieve, which is defined as a growing finite dimensional approximation to an infinite dimensional model, where the approximation grows with the sample size. For the finite dimensional sieve one then needs to find a finite upper bound on the model complexity in terms of an abstract complexity measure (such as the bracketing entropy or metric entropy, see van der Vaart and Wellner (1996) for a detailed mathematical discussion of these notions) and the non-finite part needs to receive exponentially small probability. If the Kullback-Leibler property additionally holds one can establish strong consistency. Unfortunately finding a sieve with the desired properties is often a difficult task and very model specific.

When shape constraints are involved in the statistical model one can typically exploit properties of the underlying model, which makes a proof of consistency (or strong consistency) easier. In the following we will outline three approaches we found in the literature, where the shape constraints were explicitly exploited in the proof of consistency, one of those methods, will be used in Section 4 to prove consistency for shape constrained regression.

- **Equivalence of Weak and Strong Consistency**
  As seen in Example 2.3.1 establishing strong consistency is simple, when strong and weak neighborhoods are equivalent for the space of residual densities under consideration, as strong consistency then directly follows from Theorem 2.3.1. Walker, Lijoi and Prünster (2005) describe a nonparametric shape constrained situation, where one can use such a result: Estimation of a monotone decreasing density as discussed in Section 2.1. There it was shown that any monotone decreasing density on $[c, \infty)$ can be written as a mixture of kernels of the form $\frac{1}{\xi - c}\mathbb{1}_{[c,\xi]}(y)$ (mixed with respect to ξ). Using this representation Walker, Lijoi and Prünster (2005) show that $d_W(f, g) \to 0$ is equivalent to $d_H(f, g) \to 0$. Hence strong consistency directly follows from Theorem 2.3.1 provided the Kullback-Leibler condition is met.

- **Existence of a Maximum Likelihood Estimate**
  When the classical maximum likelihood estimator (MLE) exists (which is not necessarily the case in nonparametric situations, but quite often in shape constrained nonparametric situations) one can exploit this fact. The MLE exists, for example, for estimation of a monotone density, a convex (or concave) density, a log-concave density, a monotonic regression function under a normal residual distribution and a convex (or concave) regression function under a normal residual distribution. Walker and Hjort (2001) show that in situations when the MLE exists (and satisfies an additional technical assumption), it is straightforward to show that also the posterior is consistent (provided the Kullback-Leibler property holds). An approach for proving consistency in a shape constrained situation, exploiting this fact, can be found in Shively, Sager and Walker (2009), who investigate monotone nonparametric regression. We will apply this idea to convex nonparametric regression in Section 4.

- **Finite Complexity Measures**
  The most general approach to establish strong consistency is based on sieves, which need to have a finite complexity (as measured in bracketing entropy or metric entropy). This is often achieved by truncating the parameter space or similar approaches. When there are shape constraints involved in the model,

these complexity measures are often finite automatically. For example bounded, monotone, or convex (and concave) functions have a finite complexity, see van der Vaart and Wellner (1996). This is, for example, exploited in Example 3.2 of Ghosal, Ghosh and van der Vaart (2000), who study the convergence rate of a nonparametric Bayesian posterior distribution for estimation of a monotonic density.

Hence these three examples illustrate that shape constraints (when adequate) not only improve the efficiency of inference in practice; they can also be exploited in a theoretical asymptotic analysis of the underlying model and are hence of interest both from the practical as well as from the theoretical perspective.

In the next three chapters we will now focus our attention to three concrete modelling situations, where shape constraints are employed.

Chapter 3

Bayesian Monotonic Nonparametric Regression

A Bayesian: one who asks you what you think before a clinical trial in order to tell you what you think afterwards.
Stephen Senn

This chapter deals with monotone nonparametric regression under a normality assumption on the residual density and the paper Bornkamp and Ickstadt (2009a) is based on the material presented here. This chapter relies on the Ongaro-Cattaneo random probability measure introduced in Section 2.1 and the model for monotonic functions is essentially based on a dictionary introduced in Section 2.2. The applications in this chapter are growth curve analysis and pharmaceutical dose-finding trials (we will analyse the examples presented in Figures 1.1 and 1.3), see Bornkamp (2006), Bornkamp et al. (2007) or Bretz et al. (2008) for reviews of dose-finding trials. Recently there has been an increased interest in Bayesian methodologies for pharmaceutical dose-finding studies: Drug development is inherently a sequential and adaptive process and information accrued from earlier phases of development can be used to plan and analyse experiments for subsequent phases. The Bayesian approach hence appears promising for this application.

3.1 Introduction

In numerous applications scientific knowledge suggests that the relationship between an independent and a dependent variable is monotone. Prominent examples include growth curves or dose-response analysis. For the latter task typically non-linear models, such as the four parameter logistic model or the sigmoid Emax are used (see, for example, Pinheiro, Bretz and Branson (2006) or Thomas (2006)). These models have the advantage that their parameters have an application-specific interpretation, which allows the elicitation of an informative prior distribution in a Bayesian framework. Nevertheless they impose a particular functional relationship that may not be justified. One way to overcome this is to use a nonparametric model that can approximate any continuous monotone increasing function. However, nonparametric models suffer from the fact that they are usually difficult to interpret and hence the incorporation of prior knowledge is difficult. In this chapter we propose a flexible nonparametric method for Bayesian monotone regression that is build up analogously to the classical non-linear models mentioned above and thus easily interpretable. The model is based on writing the monotone function as the sum of an intercept parameter (interpretable as the baseline effect) and the product of a scale parameter (interpretable as the maximum effect) with a continuous function that monotonically increases from 0 to 1 (*i.e.* a cumulative distribution function). But instead of assuming a parametric model for the cdf it is modelled nonparametrically as a discrete mixture of parametric distribution functions, where a general random measure is assumed as prior for the mixing distribution.

In the classical statistics literature there have been numerous approaches to nonparametric monotone regression. These are typically based on a two-stage approach of monotonizing and smoothing or vice versa (see, for example, Mukerjee (1988) or more recently Dette, Neumeyer and Pilz (2006)). Other approaches are based on constrained optimization with a spline basis (for example Ramsay (1988) or Wood (1994)) or numerical integration (for example Ramsay (1998)).

In the Bayesian framework the monotonicity assumption can be enforced by constructing a prior distribution on monotone functions. Gelfand and Kuo (1991) and Ram-

gopal, Laud and Smith (1993) proposed nonparametric shape constrained estimates for binary dose-response data. Lavine and Mockus (1995) investigated estimation of a monotone increasing function from data with an unimodal error distribution. They based the prior for the monotone function on a (shifted and scaled) Dirichlet process directly, which results in discontinuous realizations from the prior. Perron and Mengersen (2001) proposed to use mixtures of triangular distribution functions (*i.e.* quadratic splines) as a prior distribution on the space of monotone functions, while Neelon and Dunson (2004) consider monotone regression with piecewise linear functions and an autoregressive prior distribution for the parameters of the basis functions.

Our approach is new in the sense that a general discrete random measure, proposed by Ongaro and Cattaneo (2004), is assumed as prior for the mixing distribution. The parameters of the underlying base distribution functions are hence not treated as known (as is often the case, when a fixed function basis is used, such as polynomials or splines). Instead the base distribution functions in the mixture are themselves learned from the data (as well as their number and weights). This results in a flexible model, and in turn allows for a sparse representation of the underlying curve (see (Clyde and Wolpert 2007) for a discussion of flexibility and sparsity in the context of general nonparametric regression).

We also investigate how to choose the parametric class of the underlying base cdf and find that the recently introduced two-sided power distribution (van Dorp and Kotz 2002) is sufficiently rich from the mathematical perspective and also allows a computationally efficient implementation.

In some aspects our approach is similar to the nonparametric kernel regression technique of Clyde and Wolpert (2007) (see also Section 2.2). These authors model the regression function as a linear combination of kernels and propose to use general pure jump Lévy processes as prior distribution for the mixing measure. However, their approach differs as they do not consider monotone regression and the mixing measure is not normalized (*i.e.* is not a probability distribution).

The outline of this chapter is as follows: In Section 3.2 we will present our model and discuss its statistical aspects as well as issues in selecting the prior distributions. In

Section 3.3 we will first investigate the properties of our method in a simulation study and compare it with two recent proposals for monotone nonparametric regression. Then we will illustrate the suitability of our approach on a real data set from a dose-finding trial in Section 3.4 and on a growth curve example in Section 3.5. We will end this chapter with a summary and notes on possible extensions of the model.

3.2 Monotone Regression

3.2.1 Constructing the Model

We consider a model for continuous, homoscedastic data

$$y_i = \mu(x_i) + \varepsilon_i,\ i = 1, \ldots, n,$$

where $\varepsilon_i \sim N(0, \sigma^2)$ and $\mu(.)$ is a continuous monotone function. The covariate x_i is assumed to come from a bounded region, which we take to be $[0, 1]$. Without loss of generality we can write $\mu(.)$ as

$$\mu(x) = \beta_0 + \beta_1 \mu^0(x), \tag{3.1}$$

where $\mu^0(.)$ is the cdf of a continuous bounded random variable on $[0, 1]$. In this form the intercept β_0 represents the response at 0, while $\beta_0 + \beta_1$ represents the response at 1. The reason for this factorization is that in many applications these parameters have a clear cut interpretation in the application context, for example, in dose-response analysis β_0 represents the placebo and β_1 the maximum effect.

For the cdf $\mu^0(.)$ we propose to model it a-priori as a discrete mixture of parametric distribution functions $F(x, \boldsymbol{\xi})$ of bounded continuous random variables on $[0, 1]$, with parameters $\boldsymbol{\xi} \in \Xi$. We can hence formulate the model as

$$\mu^0(x) = \int_{\Xi} F(x, \boldsymbol{\xi}) P(d\boldsymbol{\xi}),$$

where P is a discrete mixing distribution on Ξ. It is easy to check that $\mu^0(.)$, being a convex combination of distribution functions, is itself a cumulative distribution function. The task is hence to deduce the discrete probability measure $P(d\boldsymbol{\xi})$ from the data

available. In recent years there have been numerous proposals for discrete random probability measures as priors for discrete probability distributions. The most widely used random probability measure is the Dirichlet process, which is mainly due to the fact that it has attractive analytical properties for the purpose of density estimation. We propose to use the general discrete random measure $\mathbb{P}$, introduced by Ongaro and Cattaneo (2004). A discrete random measure belongs to this class if its realisations can be represented as

$$P(d\boldsymbol{\xi}) = \sum_{j=1}^{J} w_j \delta_{\boldsymbol{\xi}_j}(d\boldsymbol{\xi}), \tag{3.2}$$

with $\boldsymbol{\xi}_j, w_j, J$ as given in Definition 2.1.2. Note that this construction contains many discrete random probability measures (for example the Dirichlet process or general stick-breaking processes) as a special case (see Remark 2.1.1).

Assuming the prior $\mathbb{P}$ for $P(d\boldsymbol{\xi})$, $\mu^0(x)$ with $x \in [0,1]$ is a-priori random and given by

$$\mu^0(x) = \int_{\Xi} F(x, \boldsymbol{\xi}) P(d\boldsymbol{\xi}) = \sum_{j=1}^{J} w_j F(x, \boldsymbol{\xi}_j),$$

where $\boldsymbol{\vartheta} := (J, w_1, w_2, \ldots, \boldsymbol{\xi}_1, \boldsymbol{\xi}_2, \ldots)$ has a distribution as specified in the last paragraph.

It has to be noted that the so constructed model for monotone functions is very flexible. We do not impose any structure for locating the $w_j, \boldsymbol{\xi}_j$ in the parameter space, and let these be chosen by the information in data (and prior) according to their posterior density. This allows for a sparse representation of the underlying curve, and in turn for an efficient computer implementation. This is an improvement over existing Bayesian methods for monotone regression as proposed in Perron and Mengersen (2001), who either fix the knot locations or the weights associated with a quadratic spline basis, which potentially results in a higher dimensional model. Also the piecewise linear model of Neelon and Dunson (2004) seems to require a relatively large number of basis functions (in their simulation study a more than hundred dimensional piecewise linear basis is used).

For the purpose of prior elicitation it is important to calculate prior summaries of $\mu^0(.)$ such as the mean or the variance at a certain point on the curve or the correlation

between two points on the curve. The following Lemma shows how such calculations can be done, and is a slight extension of Proposition 1 in Ongaro and Cattaneo (2004).

Lemma 3.2.1. *The expectation of $\mu^0(x)$ and the covariance of $\mu^0(x_1)$ and $\mu^0(x_2)$ for $x, x_1, x_2 \in [0,1]$, with respect to $\mathbb{P}$ are given by*

$$E(\mu^0(x)) = \int_\Xi F(x, \boldsymbol{\xi}) dP_0 \tag{3.3}$$

$$\begin{aligned} Cov(\mu^0(x_1), \mu^0(x_2)) = k_0 \Bigg\{ & \int_\Xi F(x_1, \boldsymbol{\xi}) F(x_2, \boldsymbol{\xi}) dP_0 \\ & - \int_\Xi F(x_1, \boldsymbol{\xi}) dP_0 \int_\Xi F(x_2, \boldsymbol{\xi}) dP_0 \Bigg\} \end{aligned} \tag{3.4}$$

where $k_0 = E\left(\sum_{j=1}^J w_j^2\right) \in [0,1]$.

Proof: See Theorem 2.1.3.

It is interesting to note that the prior expectation and correlation structure just depend on the distribution P_0, but not on J or Q_J. This is due to the fact that the $\boldsymbol{\xi}_j$ are a-priori independent of J and the w_j. The prior variability is determined by the prior for J and Q_J (via the factor $k_0 \in [0,1]$) as well as by P_0.

In some cases, depending on the choice of $F(., \boldsymbol{\xi})$, P_0, Q_J and the prior for J, it is possible to calculate the resulting integrals in (3.3) and (3.4) analytically, but in general numerical integration or simple Monte Carlo can be used for this purpose.

3.2.2 Choice of F

A typical requirement for the base distribution function $F(., \boldsymbol{\xi})$ would be that any continuous probability distribution function on $[0,1]$ can be approximated by a convex combination of functions $F(., \boldsymbol{\xi}_1), F(., \boldsymbol{\xi}_2), \ldots$. A distribution function possessing this property is the beta distribution function (the regularized incomplete beta function). This choice has—in slightly different and differing contexts—been investigated for example by Diaconis and Ylvisaker (1985), Petrone (1999), and Perron and Mengersen (2001). Due to the close connection with Bernstein polynomials, it is straightforward to show that any continuous distribution function on $[0,1]$ can be approximated arbitrarily well by beta distribution functions (see Diaconis and Ylvisaker (1985)). A severe

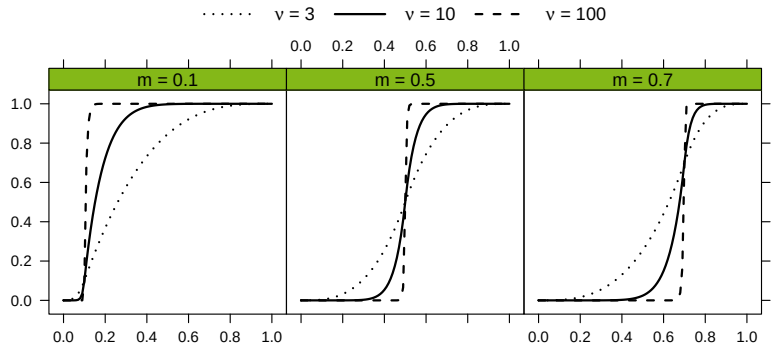

Figure 3.1: TSP distribution functions for different values of m and ν.

drawback of this approach is the fact that the regularized incomplete beta function is not available in closed form and is relatively time consuming to evaluate numerically. In particular when Monte Carlo schemes are employed to analyze the model, this becomes a hindrance as the likelihood has to be evaluated a large number of times in these approaches. In our case F needs to be evaluated nJ times for *one* evaluation of the likelihood.

Here we propose to use the distribution function of the two-sided power (TSP) distribution for F, which was introduced by van Dorp and Kotz (2002) as a viable alternative to the beta distribution. Its distribution function is available in closed form

$$F(x, \boldsymbol{\xi}) = \begin{cases} m \left(\frac{x}{m}\right)^{\nu} & 0 \leq x \leq m \\ 1 - (1-m)\left(\frac{1-x}{1-m}\right)^{\nu} & m \leq x \leq 1 \end{cases}, \tag{3.5}$$

and depends on two parameters $\boldsymbol{\xi} = (m, \nu) \in [0,1] \times \mathbb{R}_+$. If $\nu > 1$ the unique mode is given by m, and ν determines the steepness of the distribution function at m. The uniform and the triangular distribution are special cases of the TSP distribution corresponding to $\nu = 1$ and $\nu = 2$ respectively. Figure 3.1 shows different shapes of the distribution function for different parameter values. In our experience one evaluation of the TSP distribution function in C++ is around 10 to 15 times faster than one evaluation of the beta distribution function as implemented in the GSL library for C++ (version 1.8). In Theorem 3.2.1 below we show that one can also approximate any continuous distribution function on $[0,1]$ arbitrarily close by a convex combination of TSP distribution functions.

Theorem 3.2.1. *Given a continuous cdf $G(.)$ on $[0,1]$ there exist J, $w_1,...,w_J$, $m_1,...,m_J$ and $\nu_1,...,\nu_J$ with $\sum_{j=1}^{J} w_j = 1$, $m_j \in [0,1]$ and $\nu_j > 1$ such that:*

$$\sup_{x\in[0,1]} (|\mu^0(x) - G(x)|) \leq \kappa(J)(1 + 2e^{-1}),$$

where $\kappa(J) := \sup_{k\in\{0,...,J-1\}} \{G(\frac{k+1}{J}) - G(\frac{k}{J})\}$, $\mu^0(x) = \sum_{j=1}^{J} w_j F\{x,(m_j,\nu_j)\}$, and F is the distribution function of a TSP distribution.

Proof: See Appendix B.1.

As $G(.)$ is continuous on the compact interval $[0,1]$ this implies that $\kappa(J) \to 0$ for $J \to \infty$, so the family of mixtures of TSP distribution functions is sufficiently rich for our model.

3.2.3 Prior Distributions

An important aspect in any Bayesian analysis is the choice of prior distributions. In this section we will discuss how to choose the priors for $\boldsymbol{\beta}, \sigma^2$, as well as for the random probability measure $\mathbb{P}$.

Given a fixed $\boldsymbol{\vartheta} = (J, w_1, w_2, \ldots, \boldsymbol{\xi}_1, \boldsymbol{\xi}_2, \ldots)'$, $\mu^0(x)$ is fully determined and we would be back to the linear model context. The ith row of the (hypothetical) design matrix $\boldsymbol{X}$ would just be equal to $(1, \mu^0(x_i))$. So it is convenient and usually sufficient to use the conjugate normal inverse gamma distribution to express the prior information on $\beta_0, \beta_1, \sigma^2$ conditional on $\boldsymbol{\vartheta}$. This distribution depends on parameters a, d, a 2×2 matrix $\boldsymbol{V}$ and a two-dimensional vector $\boldsymbol{m}$. Its density is given by

$$p(\boldsymbol{\beta}, \sigma^2|\boldsymbol{\vartheta}) \propto (\sigma^2)^{-(d+4)/2} \times \\ \times \exp[-\{(\boldsymbol{\beta} - \boldsymbol{m})'\boldsymbol{V}^{-1}(\boldsymbol{\beta} - \boldsymbol{m}) + a\}/(2\sigma^2)],$$

see, for example, O'Hagan and Forster (2004) for details. One possibility to represent weak prior information is to let the prior variances tend to infinity, which is equivalent to letting $\boldsymbol{V}^{-1} \to \boldsymbol{0}$. Setting $a = 0$ and $d = -2$ one obtains the improper non-informative prior distribution

$$p(\boldsymbol{\beta}, \sigma^2|\boldsymbol{\vartheta}) \propto \sigma^{-2}. \tag{3.6}$$

Note that in both cases the prior distribution of $\boldsymbol{\beta}$ and σ^2 does not depend on $\boldsymbol{\vartheta}$, which is valid, because the interpretation of $\boldsymbol{\beta}$ and σ^2 does not change for different values of $\boldsymbol{\vartheta}$.

The likelihood for $\boldsymbol{y} = (y_1, \ldots, y_n)$ is given by

$$f(\boldsymbol{y}|\boldsymbol{\beta}, \sigma^2, \boldsymbol{\vartheta}) \propto (\sigma^2)^{-n/2} \exp\left(-\frac{1}{2\sigma^2}(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})'(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})\right),$$

where the above expression depends on $\boldsymbol{\vartheta}$ through the design matrix $\boldsymbol{X}$. When using the normal-inverse gamma distribution as a prior for $\boldsymbol{\beta}$ and σ^2, these parameters can analytically be integrated out to obtain the marginal posterior distribution for $\boldsymbol{\vartheta}$. In the case of non-singularity of $(\boldsymbol{X}'\boldsymbol{X})^{-1}$ it is proportional to

$$\begin{aligned} p(\boldsymbol{\vartheta}|\boldsymbol{y}) \propto & \det\{(\boldsymbol{V}^{-1} + \boldsymbol{X}'\boldsymbol{X})^{-1}\}^{\frac{1}{2}}(a + (n-2)\hat{\sigma}^2 + \\ & + \{(\boldsymbol{m} - \hat{\boldsymbol{\beta}})'(\boldsymbol{V} + (\boldsymbol{X}'\boldsymbol{X})^{-1})^{-1}(\boldsymbol{m} - \hat{\boldsymbol{\beta}})\})^{-\frac{d+n}{2}} p(\boldsymbol{\vartheta}), \end{aligned}$$

where $\hat{\boldsymbol{\beta}} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{y}$ and $(n-2)\hat{\sigma}^2 = (\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}})'(\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}})$ and $p(\boldsymbol{\vartheta})$ is the prior density of $\boldsymbol{\vartheta}$ (O'Hagan and Forster 2004). When the improper prior (3.6) is used, this reduces to

$$p(\boldsymbol{\vartheta}|\boldsymbol{y}) \propto \det(\boldsymbol{X}'\boldsymbol{X}^{-1})^{\frac{1}{2}}((n-2)\hat{\sigma}^2)^{-\frac{n-2}{2}} p(\boldsymbol{\vartheta}).$$

The random probability measure $\mathbb{P}$ is determined by the distribution of the parameter $\boldsymbol{\vartheta} = (J, w_1, w_2, \ldots, \boldsymbol{\xi}_1, \boldsymbol{\xi}_2, \ldots)'$ which can vary in dimension or might even be infinite dimensional. In a computer implementation necessarily J needs to be finite, which is why we restrict attention here to this case. As Monte Carlo methods are used to obtain a sample of the posterior distribution $p(\boldsymbol{\vartheta}|\boldsymbol{y})$ anyway, $p(\boldsymbol{\vartheta})$ is not restricted to any particular choice. In the following we discuss the choice of the prior distribution $p(\boldsymbol{\vartheta})$ in some detail.

The distribution P_0 of the $\boldsymbol{\xi}_j$ determines the prior mean function and the prior correlation structure of $\mu^0(.)$. Using the results in Lemma 3.2.1 it is possible to calculate prior mean and correlation for a particular selection of P_0, and see whether these match the prior knowledge available. In many cases a good starting point in the case of sparse prior knowledge is to use a uniform distribution on a reasonable finite subset of Ξ. In the application section we will briefly illustrate in an example how to choose P_0.

The prior for J, and Q_J, the distribution of $w_1, \ldots, w_J$, determine the factor $k_0 = E(\sum_{j=1}^{J} w_j^2)$ in Lemma 3.2.1 and hence influence the prior variability. From the theoretical perspective it is appealing to use an unbounded prior distribution for J, and for Q_J (for each J) a distribution with positive Lebesgue density on the $J-1$ dimensional simplex. In this case it can be shown that the random probability measure $\mathbb{P}$ has full support on the set of probability measures absolutely continuous with respect to P_0, see Proposition 3 and Corollary 1 of Ongaro and Cattaneo (2004) (or Theorem 2.1.4) for details. One choice fulfilling these requirements is to use a zero-truncated Poisson distribution (on $1, 2, \ldots$) with rate parameter $\lambda > 0$ as a prior distribution for J, and for Q_J the symmetric Dirichlet distribution with parameter $\gamma > 0$. In this case k_0, the expectation of $\sum_{j=1}^{J} w_j^2$ for given J is $E(\sum_{j=1}^{J} w_j^2 | J) = \frac{\gamma+1}{J\gamma+1}$. So the variance of $\mu^0(.)$ is increasing when J or γ get smaller. In practice we believe the choice of λ and γ should be based on (i) the desired variability in the prior for $\mu^0(.)$ and (ii) the expected number of jumps in the modelled response. Again we will illustrate one particular selection in the next section.

3.2.4 Asymptotics

Before illustrating the methodology on concrete examples we would like to point out that the full support property for $\mathbb{P}$ and the priors for $\boldsymbol{\beta}$ and σ^2 together with the approximation property of TSP distribution functions (Theorem 3.2.1) is already enough to ensure full support in Kullback-Leibler divergence and hence consistency (see Section 2.3). We will discuss this topic in full detail in Chapter 4, where convex regression is considered. The theory presented there also applies to the case of monotonic regression.

3.3 Simulation Study

In this section we will evaluate the performance of the proposed methodology in a simulation study. Two recent classical approaches to monotone regression will be used

to assess the quality of our methodology. One approach is based on local linear regression and the other on penalized regression splines. In total six scenarios will be used, corresponding to three different test functions and two different noise levels. In each case 50 normal random variables are generated with mean $\mu_j(i/49)$, $i = 0, \ldots, 49$, where $\mu_j(.)$, $j = 1, 2, 3$, is one of the three test functions, and variance σ^2, where σ is set to 0.05 or to 0.2, respectively.

The following monotone test functions will be used, see also Figure 3.2

$$\begin{aligned}
\mu_1(x) &= x^7/(x^7 + 0.4^7) \\
\mu_2(x) &= \frac{1}{3}B(x, 1, 1) + \frac{1}{3}B(x, 200, 80) + \frac{1}{3}B(x, 80, 200) \\
\mu_3(x) &= \begin{cases} \frac{10}{6}x & 0 \leq x \leq 0.6 \\ 1 & x > 0.6. \end{cases}
\end{aligned}$$

Here $B(.,.,.)$ denotes the distribution function of the beta distribution. This choice is a compromise, between common shapes in dose-response analysis (μ_1), shapes that could appear in growth curve analysis, where growth often appears in short bursts rather than linearly (μ_2), and a non-smooth shape, which has a sharp corner (*i.e.* a discontinuity in the first derivative) at $x = 0.6$ (μ_3). Hence one might expect that the function $\mu_3(.)$ might be difficult to approximate with a mixture of smooth base distribution functions.

The quality of the methods will be evaluated with the mean absolute estimation error (MAE) over the range $[0, 1]$

$$MAE = \frac{1}{11}\sum_{i=0}^{10} |\mu(i/10) - \hat{\mu}(i/10)|,$$

where $\hat{\mu}(i/10)$ is the point estimate of $\mu(i/10)$. The six simulation scenarios are run with 2000 simulations per scenario.

We compare our approach with the method of Dette, Neumeyer and Pilz (2006), which is implemented in the `monreg` package for `R` (Pilz, Titoff and Dette 2005). The method applies local linear regression and then (if necessary) monotonizes the fit with a technique based on kernel density estimation. Two bandwidths need to be specified, one for the local linear regression and one for the density estimation step. The bandwidth

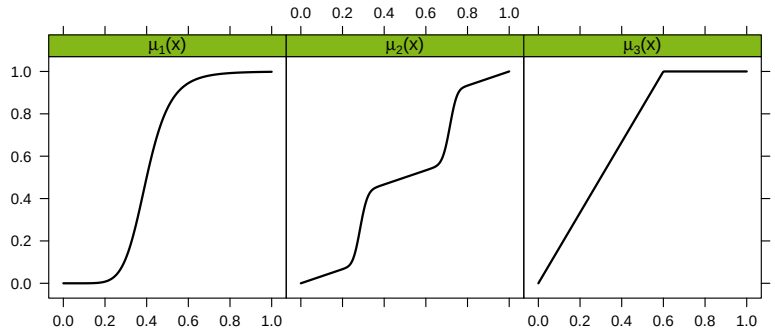

Figure 3.2: Test functions used in simulation study.

h_r for the regression estimator is chosen as recommended in Dette, Neumeyer and Pilz (2006), and the bandwidth for the kernel estimator as h_r^3. The results of this approach in the simulation study can be found in the first column of Table 3.1. The other approach is based on constrained optimization with a spline basis. For this purpose we use the `mgcv` package for `R` (Wood 2007). A penalized cubic regression spline, with ten knots uniformly spread over the range $[0, 1]$ is employed and linear constraints necessary for monotonicity of the spline as described in Wood (1994) are used in a quadratic optimization to find the parameter values for the spline basis functions. The smoothing parameter is set to the value obtained by generalized cross-validation for the unconstrained problem. See the examples for the `pcls` function (in the `mgcv` package) for details. The results of this approach can be found in the middle column of Table 3.1.

For the method proposed in this chapter we use the non-informative prior for $\boldsymbol{\beta}$ and σ^2 from Equation (3.6). The cdf $\mu^0(.)$ is modelled as a nonparametric mixture of TSP distribution functions, where the random measure $\mathbb{P}$ is assumed as a prior on the mixing distribution. For J a zero-truncated Poisson distribution with parameter 1 (resulting in a prior expectation of ≈ 1.58) is used. For the weights a Dirichlet distribution with $\gamma = 1$ (*i.e.* a uniform distribution) on the $J - 1$ dimensional simplex is used for each J. This results in a choice of $k_0 \approx 0.84$, so a relatively large variability. In addition this represents the prior information that the functions are smooth, with not exceedingly many jumps. For the parameter m of the TSP distribution function a uniform distribution on $[0, 1]$ is used, while the parameter ν was given a uniform distribution on $[1, 70]$.

Testfunction	σ	LocLin/Mon	MonRS	MonBayes
$\mu_1(.)$	0.05	0.0216	0.0149	0.0117
	0.2	0.0628	0.0527	0.0475
$\mu_2(.)$	0.05	0.0274	0.0172	0.0188
	0.2	0.0632	0.0582	0.0557
$\mu_3(.)$	0.05	0.0176	0.0161	0.0212
	0.2	0.0520	0.0502	0.0561

Table 3.1: Mean absolute estimation error for compared methods (LocLin/Mon $\hat{=}$ Local linear regression and monotonization, MonRS $\hat{=}$ Monotone regression splines, MonBayes $\hat{=}$ Monotone nonparametric Bayes estimate), based on 2000 simulations.

The reason for bounding ν at 70 is that in this case the TSP distribution function almost corresponds to a step function from 0 to 1, and more extreme shapes might be excluded a-priori. Performing the integration (3.3) it can be seen that this specification approximately corresponds to a linearly increasing prior mean. In addition the variability for $\mu^0(.)$ is reasonably large. For reasons of comparability we refrained from adapting the priors to the simulation scenarios and used the same prior for each scenario. To estimate the function $\mu(.)$ we will use the pointwise medians of its posterior distribution (pointwise posterior means lead to very similar results). Model fitting is done with the reversible jump (RJ) MCMC algorithm described in Appendix C.1, based on 50000 iterations after a burn in of 5000. Every fifth value is used to reduce dependence.

Table 3.1 displays the results for the different scenarios. It can be seen that our method performs best for the first test function and performs equally well to the monotone regression splines for the second test function. For $\mu_3(.)$ the performance is reasonable but as expected slightly worse than the other approaches. Overall the monotone regression splines and the approach proposed in this chapter perform best in the simulation study, while the local linear regression approach followed by a monotonization step shows a slightly worse performance.

3.4 Dose-Response Analysis

Clinical dose-response studies are typically conducted in Phase II of the pharmaceutical development program. Two of the major aims in this phase are (i) to learn about

the shape of the dose-response relationship and (ii) the estimation of a suitable dose to be used in large scale Phase III trials. If prior knowledge on the compound allows to assume monotonicity (an assumption that should critically be checked) our method becomes appropriate.

To illustrate the described method, a real trial data set taken from Biesheuvel and Hothorn (2002) will be used. The data are part of a dose ranging trial on a compound for the treatment of the irritable bowel syndrome with four active doses 1, 2, 3, 4 equally distributed in the dose range $[0, 4]$ and placebo. The primary continuous endpoint was a baseline adjusted abdominal pain score with larger values corresponding to a better treatment effect. In total 369 patients completed the study, with nearly balanced allocation across the doses. This data set was also analyzed by Bretz, Pinheiro and Branson (2004) with a recent classical approach to dose-finding studies that is based on a multiple comparison framework for model selection (see Bretz, Pinheiro and Branson (2005) for a detailed description of this methodology).

Here we illustrate the estimation of the so called minimum effective dose (MED)

$$MED = \min_{x \in (0,4]} \{x : \mu(x) > \mu(0) + \Delta\}, \tag{3.7}$$

where Δ is the threshold from which on a response is regarded as clinically relevant. Note that the MED does not exist, if the function $\mu(.)$ is entirely below $\mu(0) + \Delta$. The MED is often of interest as it can be interpreted as a lower bound on all useful doses. In practice it is desirable not only to obtain a point estimate of the MED, but also a variability statement. As the MED is a functional of $\mu(.)$ this is straightforward to obtain in the Bayesian framework, without relying on asymptotic arguments. For the analysis of the data set we set $\Delta = 0.25$ in accordance with Bretz, Pinheiro and Branson (2004).

No additional information from the clinical team is available for the data, but to illustrate the impact of using informative prior distributions for $\beta_0, \beta_1, \sigma^2$ we will investigate two scenarios: a weakly informative choice and an informative choice of the priors. For the prior $\mathbb{P}$ there is no non-informative choice, and in both settings informative priors with relatively large variance are used.

In the weakly informative setting we use the non-informative prior from Equation (3.6) for $\boldsymbol{\beta}$ and σ^2. For the random probability measure $\mathbb{P}$ the following specification is used: For J, the number of basis functions, a zero-truncated Poisson distribution with parameter 0.5 is used (corresponding to a prior mean of ≈ 1.27). For the distribution Q_J the symmetric Dirichlet distribution with $\gamma = 1$ is employed. This results in $k_0 \approx 0.91$, *i.e.* a relatively large variability. In addition this corresponds to the prior expectation that there will not be many more than 1 or 2 jumps in the response. For the parameters $\boldsymbol{\xi}_j = (m_j, \nu_j)$ of the TSP distribution function again a uniform distribution on $[0,1] \times [1,70]$ is used, as this approximately corresponds to a linear increasing prior mean. For the informative prior we will use a normal-inverse gamma prior for $\boldsymbol{\beta}, \sigma^2$ with parameter $\boldsymbol{m} = (0.21, 0.55)'$, $\boldsymbol{V} = ((0.01, 0)', (0, 0.1)')$, $a = 3.6$ and $d = 4$. The means $\boldsymbol{m}$ are chosen as the empirical means in the placebo group and the group with the highest dose. For β_0 the variance is 0.018 which approximately corresponds to the knowledge that can be obtained from 33 patients (assuming that $\sigma^2 = 0.6$, which is approximately equal to the empirical variance observed in the study), while the variance for β_1 approximately corresponds to the information gained by 3 patients. The prior distribution for σ^2 is chosen, such that its mode is equal to 0.6 and the prior variance of σ^2 is infinite. For the random probability measure $\mathbb{P}$, the priors for J, Q_J and ν are chosen identical to the setting above, while for m a $Beta(1,2)$ distribution on $[0,4]$ was used, corresponding to the believe that $\mu^0(.)$ is concave (the prior mean is approximately equal to the distribution function of a $Beta(1,2)$ distributed variable). Note that in practice the process of eliciting prior distributions should be done with considerable care and *before* the actual data are available, based on data from previous trials or similar compounds. Our selection of informative priors here is just done for illustrative purposes. A useful approach to calibrate the prior distributions (if proper distributions are employed) is to simulate from the prior predictive distribution of the MED and see whether summaries, such as the probability of identifying an MED, the mean or quantiles, are plausible.

To fit the model the RJ-MCMC algorithm described in Appendix C.1 was implemented in C++ and was run in both the weakly informative and informative scenario for 200000 iterations after 5000 iterations burn in. To reduce dependence in the chain only

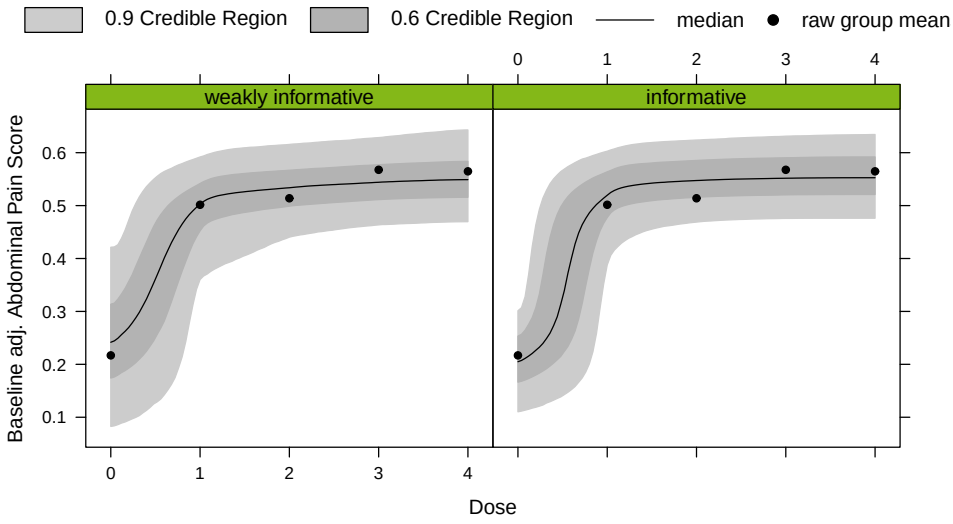

Figure 3.3: Pointwise posterior quantiles of the estimated dose-response curve and observed group means.

every tenth value has been saved and is used for further analysis. We assessed convergence by running parallel chains from very different starting points and confirmed that the results for summaries of the posterior distribution were consistent between the chains. The pointwise posterior distributions for the curve are visualized in Figure 3.3 and it can be seen that both approaches lead to similar results. The most notable difference is that the informative prior leads to smoother posterior quantiles and a reduced uncertainty especially around $\mu(0)$ and in the interval $[0, 1]$. As the prior distribution for β_0 was chosen quite informative this is not surprising. The posterior distribution for the MED is obtained by applying Equation (3.7) to the posterior draws of $\mu(.)$. In the non-informative setting in approximately 30 percent of the posterior draws from the function $\mu(.)$ were entirely below $\mu(0) + \Delta$. This can be interpreted as the probability that the MED does not exist in $[0, 4]$ is 30 percent. For the informative setting the corresponding value is just 8 percent. Summaries of the posterior distribution of the MED conditional on that an MED estimate exists are displayed in Table 3.2. It can be seen that the variance in the posterior distribution drops by more than 50 percent, which is due to the fact that the posterior of the MED becomes concentrated on the interval $[0, 2]$ in the informative setting, while in the weakly-informative setting more probability mass lies in the interval $[2, 4]$. This result is not surprising as one can ex-

	Weakly Informative	Informative
Mean	0.90	0.74
Variance	0.44	0.19
0.025-Quantile	0.16	0.16
0.25-Quantile	0.48	0.44
Median	0.76	0.72
0.75-Quantile	1.00	0.92
0.975-Quantile	3.04	1.96

Table 3.2: Summary statistics for posterior distribution of the MED for the weakly informative and informative choice of priors.

pect that the variability of the MED estimate crucially depends on the variability of the estimate for $\mu(0)$. This variability is reduced in the informative setting, through the incorporation of prior information on the placebo response.

As a point estimate for the MED the posterior mean or median may be chosen. Bretz, Pinheiro and Branson (2004), selecting a hyperbolic $E_{\max}$ model for dose estimation, obtained an MED estimate of 0.74, which is quite similar to the posterior median both in the informative and weakly informative scenario. However, in their approach information about the uncertainty of the MED estimate is not directly available.

It has to be noted that the posteriors of $\mu(.)$ and the MED are not really sensitive with respect to the prior distribution for J. We confirmed this by reanalyzing the data using the prior specifications $\lambda = 5, \gamma = 0.025$ and $\lambda = 10, \gamma = 0.011$ (in both cases the parameters were chosen in a way such that k_0 was approximately equal to 0.91, ensuring equality of the prior variance between the different prior choices). The resulting posteriors for $\mu(.)$ and MED were in both cases essentially equal to the result obtained from the used prior parameters $\lambda = 0.5, \gamma = 1$.

3.5 Growth Curves

Thalange, Foster, Gill, Price and Clayton (1996) report a study of the growth of 5-8 year old children over a 312-days period. It is observed that growth in this period typically occurs in short bursts rather than linearly, which makes our model more appropriate than modelling with standard smooth non-linear parametric models such

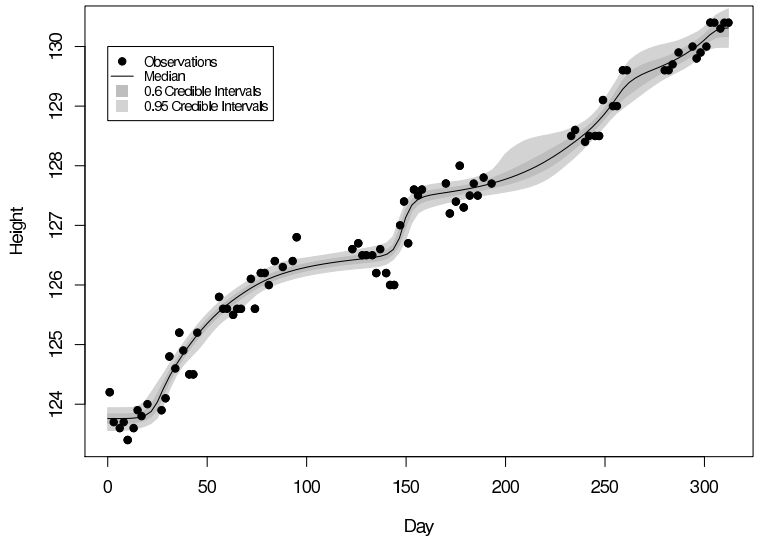

Figure 3.4: Pointwise posterior quantiles of the estimated growth curve and observations.

as the sigmoid Emax and the logistic model. We will analyze the growth data of one child from this study, which were also analyzed by Ramsay (1998). For simplicity we will use the non-informative prior (3.6) for $\beta_0, \beta_1, \sigma^2$. For the location of the modes m of the TSP distribution functions a uniform distribution on $[0, 312]$ is used, while ν was given a uniform distribution on $[1, 70]$. This choice approximately corresponds to a linear increase over the whole period, which is reasonable assuming no specific knowledge on the location of the growth bursts. For the distribution Q_J the uniform distribution on the simplex was used, while for J the number of basis functions a zero-truncated Poisson distribution with parameter 1 was chosen. Again the RJ-MCMC method described in Appendix C.1 is used to obtain a posterior sample of $(\boldsymbol{\beta}, \sigma^2, \boldsymbol{\vartheta})'$. In Figure 3.4 the growth data of the child can be seen in addition to pointwise posterior credibility bands for the growth curve.

The posterior distribution for the parameter J has mean 4.6 and variance 1.3, so there is quite strong evidence of 4 to 5 growth periods within these 312 days. The parameters m and ν of the individual TSP distribution components determine the day and the speed of growth for the corresponding growth-burst, which is quite different between the different growth periods. There is, for example, one quite steep growth step around day 150, while the growth from day 20 to day 95 is somewhat less steep. Note

that this type of direct interpretability of the parameters would be difficult to obtain from a (typically high-dimensional) basis function approach.

An interesting aspect of imposing monotonicity becomes apparent when observing the uncertainty intervals in the periods with no measurements (day 95 – day 123 and also day 193 – day 233). In the first period uncertainty in the function estimate is practically the same as everywhere else, despite the sparse data, but in the second period with no data there is a larger uncertainty. This nicely illustrates the effect of imposing monotonicity: The child does not seem to have grown a lot between day 95 and day 123, so for a continuous monotonic function there are not many ways of connecting the heights at day 95 and 123, and consequently there is less uncertainty in estimating the underlying function. The situation for the second period is quite different; here the child seems to have grown in the period when no measurements have been made and there are various possibilities of connecting both function levels with a continuous monotone function. So the uncertainty for the second period with sparse data is much larger.

3.6 Summary and Outlook

In this chapter a nonparametric method for monotone regression has been presented that relies on representing the monotone function by an intercept parameter, a scale parameter and a cdf. The cdf is modelled as a discrete mixture of known base distribution functions and a general discrete random probability measure is assumed as prior for the mixing distribution. The formed model for the monotone function is very flexible and hence allows for a sparse representation of the curve. The choice of the base distribution function has been discussed and the two-sided power distribution has been proposed as an alternative to the frequently used beta distribution function. The TSP distribution function allows for a more efficient implementation and we showed that convex combinations of TSP distribution functions are rich enough to approximate any continuous distribution function.

Special emphasis in the construction of the model has been laid on interpretability,

such that it is possible to elicitate informative prior distributions for important aspects of the curve. In dose-response analysis, for example, there typically exists prior knowledge about the placebo effect, the standard deviation and sometimes also the maximum effect. Incorporation of prior knowledge for these quantities is straightforward with the presented model. Additionally we discussed and illustrated one choice of priors for the random probability measure $\mathbb{P}$, using a zero-truncated Poisson distribution for the number of base distribution functions in the model and a symmetric Dirichlet distribution for the weights. We believe this is a useful choice of priors although our approach also allows for many other choices, which may turn out to be useful in specific situations.

A simulation study has been performed, which compared our method with two recent classical approaches to nonparametric monotone regression. Our method performs very competitive, but can incorporate possible prior information and provides uncertainty statements for all aspects of the modelling process. The usefulness of the method has also been shown on two real data sets. The first originates from a clinical dose-finding study. We illustrated that inclusion of possible prior knowledge, especially for the placebo response, can greatly reduce the uncertainty inherent in MED estimation. A second example investigates our model for estimating the growth curve of a child. Growth occurs in bursts rather than smoothly, so our model, being able to accommodate multiple growth steps, is quite adequate in this modelling situation.

There are various possibilities to extend the model to binary or categorical responses. The easiest approach would be to specify a (monotone) link function to map the parameter values onto the correct range. Covariates (such as gender, age or center effects typically available in dose-finding trials) can easily be incorporated in the proposed model by replacing β_0 with a additive linear model structure, the same computer implementation could also be used in this scenario.

Convexity or concavity constraints may also be incorporated with the presented methodology and we will pursue this idea further in the next Section 4. The generalization to the case when there are multiple predictors and all are assumed to be in a monotonic relationship with the response variable will be given in Section 5.

Chapter 4

Bayesian Nonparametric Regression under Derivative Constraints

This chapter is an extension of the last chapter in the sense that we treat nonparametric regression under convexity and monotonic convexity constraints (and more generally derivative constraints). For this purpose we impose the model used in the last chapter for the first derivative of the regression function. Hence similar to Chapter 3 this section relies on the Ongaro-Cattaneo random probability measure introduced in Section 2.1 and the approach to shape constrained regression based on dictionaries discussed in Section 2.2. In this chapter we also investigate asymptotic properties of the posterior distribution with methods quite related to those described in Section 2.3. The finite sample performance of the model will then be investigated in a simulation study and the model will be illustrated in a biological application.

4.1 Introduction

It is quite common that mathematical models impose constraints on the derivative in the relationship between variables, such as monotonicity, concavity or convexity. For some examples see Ramgopal, Laud and Smith (1993), Meyer (2008) or Lee, Lim, Kim and Joo (2009). Particularly rich sources of examples are also economics (*e.g.* pro-

duction functions) and finance (*e.g.* option pricing), where the constraints usually can directly be deduced from the economic theory underlying the problem. In this chapter we will introduce a flexible Bayesian model for nonparametric regression under these type of constraints, specifically we will consider monotonicity and positivity constraints on the derivative. The main idea is to build a model for a shape-constrained derivative and then to obtain the original function by integration. Along this way we develop a representation of convex and monotone convex functions, where the convexity and monotone convexity shape-constraint can be imposed by simple finite dimensional inequality constraints.

Nonparametric estimation of functions with shape constraints on derivatives has a long tradition in the area of classical statistics, see, for example, Hildreth (1954), Hanson and Pledger (1976), Mammen (1991) and Groeneboom, Jongbloed and Wellner (2001) for theory on the classical convex least squares estimate (which is a monotonicity constraint on the first derivative). Birke and Dette (2007) recently approach the problem by "convexification" of an unconstrained local polynomial estimate. Another main stream of the literature on derivative constraints is directly motivated by option pricing, see, for example, Aït-Sahalia and Duarte (2003), Yatchew and Härdle (2006), or Birke and Pilz (2009). In addition particularly spline functions have been used to deal with shape constrained derivatives, see, for example, Dierckx (1980) or Schwetlick and Kunert (1993) for references from the numerical analysis literature or Dole (1999) and Turlach (2005).

From the Bayesian perspective, shape constrained nonparametric regression has been considered extensively when it comes to monotonicity constraints, see, for example, Lavine and Mockus (1995), Perron and Mengersen (2001), Neelon and Dunson (2004) and recently Shively, Sager and Walker (2009) or the model described in the last chapter. The case of more general shape constraints on derivatives such as convexity or concavity, however, has not been covered in this extent. An early parametric account is O'Hagan (1973), who considers estimation of a convex quadratic polynomial. More recent references are Chang, Hsiung, Wu and Yang (2005) and Chang, Chien, Hsiung, Wen and Wu (2007), who use the Bernstein polynomial basis for nonparametric shape constrained survival analysis and regression.

The outline of this chapter is as follows: Section 4.2 motivates and describes the model and associated prior distributions in detail and investigates its asymptotic properties. Section 4.3 displays the results of a simulation study and in Section 4.4 we apply our ideas a real data example. Section 4.5 concludes.

4.2 Model

Setting up a prior distribution for nonparametric regression models is challenging. When using a functional basis, the interpretation of the basis coefficients is typically involved, and the elicitation of prior distributions is difficult. Another issue is how to determine the dimension of the underlying basis. Although the model dimension can be treated as unknown (as done for example in Chang et al. (2007)), it is possible that a relatively large number of basis functions is needed to be able to reconstruct the underlying function satisfactorily. A way out of these problems is to use an overcomplete dictionary as introduced in Section 2.2, where not only the basis coefficients are treated as unknown but also the basis functions themselves. The prior distribution developed in this chapter will be based on such a dictionary.

4.2.1 Modelling the Derivative

The main idea is to build the model for a function $\mu(.)$ on $[0,1]$ based on a shape constrained derivative and then obtain the original function to be modelled by integration. We hence first construct a model for a monotonic derivative $\mu'(.)$. This will be done using the approach developed in Chapter 3 for monotonic regression. There a continuous monotonic function $\mu'(.)$ on $[0,1]$ was decomposed as

$$\mu'(x) = \beta_1 + \beta_2 F(x), \tag{4.1}$$

where β_1 and β_2 are parameters and $F(x)$ is the distribution function of a continuous random variable on $[0,1]$. The model for $\mu(.)$ can then be obtained simply by calculat-

ing an integral of $\mu'(.)$

$$\mu(x) = \int_0^x \mu'(t)\, dt = \beta_0 + \beta_1 x + \beta_2 F^*(x), \tag{4.2}$$

where $F^*(x) = \int_0^x F(t)\, dt$. In this decomposition shape constraints for $\mu(.)$ and $\mu'(.)$ reduce to simple finite dimensional constraints for β_1 and β_2. A convexity constraint, for example, corresponds to imposing $\beta_2 \geq 0$, as this results in a monotonic increasing derivative. A monotonic increasing and convexity constraint can be imposed by using $\beta_1 \geq 0$ and $\beta_2 \geq 0$. When $\mu'(.)$ is assumed to be a cdf (as in option pricing) the desired constraints are $\beta_1 = 0$ and $\beta_2 = 1$. Note that a variety of other constraints, such as monotone decrease, and concavity are possible with this basic approach by simple transformations of the data or range (see also Appendix A.1 for details).

Because (4.1) holds for any continuous monotonic function, it follows that any continuously differentiable convex function can be represented in form (4.2) (with $\beta_2 \in \mathbb{R}_+$). The same is true for any other continuously differentiable shape constrained space of functions, when the shape constraint can be expressed in terms of β_0, β_1 and β_2. An additional advantage of representations (4.1) and (4.2) is the fact that the parameters $\boldsymbol{\beta} = (\beta_0, \beta_1, \beta_2)'$ are still relatively easy to interpret: β_0 represents the function value at 0, β_1 represents the value of the derivative at 0 and β_2 the amount that the first derivative increases in $[0, 1]$. In a sense parameters β_0, β_1 and β_2 hence constitute the "parametric" part of the function (4.2), and the shape constraints only need to be imposed on this finite dimensional part. The "nonparametric" part consists of the function F, which determines the shape of the first derivative. We propose to model the distribution function $F(.)$ as a convex combination of parametric distribution functions $G(x, \boldsymbol{\xi})$ with parameter $\boldsymbol{\xi} \in \Xi$, *i.e.*

$$F(x) = \sum_{j=1}^{J} w_j G(x, \boldsymbol{\xi}_j) = \int_{\Xi} G(x, \boldsymbol{\xi}) P(d\boldsymbol{\xi}), \tag{4.3}$$

where $J \in \{\mathbb{N}_+ \cup \infty\}$, $\sum_{j=1}^{J} w_j = 1$ and $P(d\boldsymbol{\xi})$ is a discrete probability (mixing) measure. At this point we leave $G(., \boldsymbol{\xi})$ unspecified, but note that a convex combination of distribution functions $G(., \boldsymbol{\xi}_j)$ should be sufficiently flexible to achieve a large support. The function $F^*(.)$ appearing in (4.2) is hence modelled by $F^*(x) = \sum_{j=1}^{J} w_j G^*(x, \boldsymbol{\xi}_j) = \int_{\Xi} G^*(x, \boldsymbol{\xi}) P(d\boldsymbol{\xi})$, where $G^*(., \boldsymbol{\xi}) = \int_0^x G(t, \boldsymbol{\xi})\, dt$.

In summary we will use the following model (restricting ourselves in this chapter to a Gaussian error model):

$$Y_i = \mu(X_i) + \varepsilon_i,\ \varepsilon_i \overset{iid}{\sim} N(0, \sigma^2),\ i = 1, \ldots, n$$
$$\mu \sim \Pi_\mu(d\mu(.)),\ \sigma^2 \sim p(d\sigma^2), X_i \sim Q(dx),$$

where $\mu(x) = \beta_0 + \beta_1 x + \beta_2 F^*(x)$, $F^*(x) = \int_\Xi G^*(x, \boldsymbol{\xi}) P(d\boldsymbol{\xi})$, $p(d\sigma^2)$ is the prior distribution for σ^2, $Q(dx)$ is the probability distribution of the covariates on $[0, 1]$ and $\Pi_\mu(d\mu(.))$ consists of a parametric prior distribution for $\boldsymbol{\beta} = (\beta_0, \beta_1, \beta_2)'$, which enforces the desired shape constraint and a nonparametric prior distribution $\mathbb{P}$ for the mixing distribution $P(d\boldsymbol{\xi})$ and the base distribution function $G(., \boldsymbol{\xi})$. In the next section we will concentrate on how to choose the components of the prior distribution Π_μ.

4.2.2 Prior Distributions

For the parameters β_0, β_1 and β_2 the support of the prior is determined by the type of constraint to be used. If a parameter is positive one might, for example, use a truncated normal or an exponential distribution as prior, while a normal prior might be used in the unrestricted case. When a particular value is assumed for the parameter, degenerate one-point distributions can be used. The nonparametric part of the prior Π_μ is given by the prior for the discrete mixing probability measure $P(d\boldsymbol{\xi})$ and the choice of the base distribution function $G(., \boldsymbol{\xi})$. Several random probability measures $\mathbb{P}$ might be used as a prior for the discrete mixing distribution $P(d\boldsymbol{\xi})$. We will subsequently use the general prior of Ongaro and Cattaneo (2004). A random probability measure $\mathbb{P}$ belongs to this class when its realizations (discrete probability measures with support points in a space Ξ) can be represented as

$$P(d\boldsymbol{\xi}) = \sum_{j=1}^{J} w_j \delta_{\boldsymbol{\xi}_j}(d\boldsymbol{\xi}), \tag{4.4}$$

where $\boldsymbol{\xi}_j, w_j$ and J are random variables as given in Definition 2.1.2. Ongaro and Cattaneo (2004) (see also Theorem 2.1.4) show that the so constructed random probability measure has full support on the space of probability distributions absolutely continuous with respect to P_0, when the prior distribution for J is unbounded and one of the

two support assumptions (i) or (ii) of Theorem 2.1.4 are fulfilled. Several choices fulfill these requirements, for example the shifted Poisson distribution on $1, 2, 3, \ldots$ with parameter $\lambda > 0$ and a symmetric Dirichlet distribution on the probability simplex with parameter $\delta > 0$ for each J. In an actual application, one needs to select one particular prior specification. Hence we need additional guidance on how to choose the priors. One possibility is to calculate prior mean function and prior variability (and covariance) of the a-priori random function $\mu(.)$ and choose the prior parameters, so that they match possible prior information and prior uncertainty. The following result shows how to calculate these quantities and is an extension of Lemma 3.2.1, as we now also investigate the influence of the prior for $\boldsymbol{\beta}$ on the prior moments.

Lemma 4.2.1 (Prior Expectation and Prior Covariance)**.** *Under the assumption that the priors for $\boldsymbol{\beta}$ and $\mathbb{P}$ are independent, the prior expectation with respect to Π_μ conditional on $x \in [0,1]$ is given by*

$$E(\mu(x)|x) = E(\beta_0) + E(\beta_1)x + E(\beta_2) \int_\Xi G^*(x, \boldsymbol{\xi})\, dP_0(d\boldsymbol{\xi}), \tag{4.5}$$

where $G^*(x, \boldsymbol{\xi}) = \int_0^x G(t, \boldsymbol{\xi})\, dt$. The covariance of $\mu(x_1), \mu(x_2) \in [0,1]$ is given by

$$Cov(\mu(x_1), \mu(x_2)|x_1, x_2) = E(a_1' \boldsymbol{B} a_2) + E(\beta_2)^2 Cov(F^*(x_1), F^*(x_2)), \tag{4.6}$$

where $a_i = (1, x_i, F^*(x_i))'$, $\boldsymbol{B} = Cov(\boldsymbol{\beta})$,
$Cov(F^*(x_1), F^*(x_2)) = k_0(\int_\Xi G^*(x_1, \boldsymbol{\xi}) G^*(x_2, \boldsymbol{\xi}) dP_0(d\boldsymbol{\xi}) -$
$\int_\Xi G^*(x_1, \boldsymbol{\xi}) dP_0(d\boldsymbol{\xi}) \int_\Xi G^*(x_2, \boldsymbol{\xi}) dP_0(d\boldsymbol{\xi}))$
and $k_0 = E(\sum_{j=1}^J w_j^2) \in [0,1]$.

Proof. See Appendix B.2.

It is interesting to note that the prior mean function only depends on the prior for $\boldsymbol{\beta}$ and P_0, but not on the distribution of J and $(w_1, \ldots, w_J)|J$. For the covariance however, they play a crucial role. In Equation (4.6) both summands depend on $k_0 \in [0,1]$, in the sense that the variability in the prior distribution gets larger, when k_0 gets larger (for this note that $Cov(F^*(x_1), F^*(x_2))$ and $E(F^*(x_1)F^*(x_2))$ are maximal, when $k_0 = 1$). Hence the priors for J and $w_1, \ldots, w_J|J$ can be chosen based on the desired variability in the prior for $\mu(.)$, as well as on how many base distribution functions

are deemed to be necessary. Note that in general straightforward Monte Carlo (or other numerical techniques) can be used to calculate prior mean and covariance function from Lemma 4.2.1 for the particular model and priors under consideration. In the context of elicitation, it might also be worthwhile to investigate the prior distribution for the derivative: The prior mean function of the derivative is given by $E(\mu'(x)|x) = E(\beta_1) + E(\beta_2) \int_{\Xi} G(x, \boldsymbol{\xi})\ dP_0(d\boldsymbol{\xi})$. A starting point for prior elicitation here would be to select the distribution P_0 such that the prior mean function for the first derivative is (approximately) linearly increasing, so that the prior mean function for $\mu(.)$ is approximately a quadratic polynomial.

In the following we will consider how to choose the base distribution function $G(., \boldsymbol{\xi})$. To assure that an arbitrary differentiable convex function can be represented as (4.2) we need to find a function $G(., \boldsymbol{\xi})$ such that any continuous monotonic function can be approximated in sup-norm by a function of form (4.1) (modelling $F(x)$ by (4.3)). So $G(., \boldsymbol{\xi})$ needs to be chosen such that for any continuous distribution function H on $[0, 1]$, there exists $\boldsymbol{\xi}_1, \boldsymbol{\xi}_2, \ldots, \boldsymbol{\xi}_J, \ldots$ with $\boldsymbol{\xi}_j \in \Xi$ and w_j, subject to $\sum_{j=1}^{J} w_j = 1$ so that

$$\sup_{x \in [0,1]} \left| \sum_{j=1}^{J} w_j G(x, \boldsymbol{\xi}_j) - H(x) \right| \to 0, \text{ for } J \to \infty. \tag{4.7}$$

Several choices of distribution functions have this property. A second important requirement on $G(., \boldsymbol{\xi})$ is that the integral over the $G(., \boldsymbol{\xi})$ is available in a closed form, to be able to calculate $G^*(., \boldsymbol{\xi})$ and $F^*(.)$ efficiently (otherwise one would need to perform numerical integration to evaluate the likelihood function, which would be computationally prohibitive).

A cdf possessing both properties is the distribution function of the two-sided power (TSP) distribution (van Dorp and Kotz 2002) already used in Chapter 3, see Equation (3.5) for a definition of the TSP distribution function. In Chapter 3 it is shown in Theorem 3.2.1 that (4.7) holds for the TSP distribution function. In addition, as desired, the integral over a TSP distribution function is available analytically and given by

$$G^*(x, \boldsymbol{\xi}) = \begin{cases} \frac{m^2}{\nu+1} \left(\frac{x}{m}\right)^{\nu+1} & 0 \leq x \leq m \\ x - m + \frac{(1-m)^2}{(\nu+1)} \left(\frac{1-x}{1-m}\right)^{\nu+1} + \frac{2m-1}{\nu+1} & m \leq x \leq 1 \end{cases}.$$

4.2.3 Asymptotic Considerations

In this section we investigate the asymptotic properties of the posterior distribution for the special case of convex nonparametric regression. Particularly we will investigate, whether the posterior distribution concentrates its probability mass at a one point distribution on the true value, when one has perfect information, *i.e.* an infinite sample size. As discussed in Chapter 2.3 one can explicitly exploit the convexity shape constraint in a consistency proof, as was done by Shively, Sager and Walker (2009) for the case of monotone nonparametric regression.

Let $\zeta = (\mu(.), \sigma^2)$ and let the true value of ζ be $\zeta_0 = (\mu_0(.), \sigma_0^2)$. Here $\mu_0(.)$ is hence a continuously differentiable convex function and $\sigma_0 > 0$. In addition the Kullback-Leibler divergence between two densities (dominated by the Lebesgue measure on $\mathbb{R}$) is denoted by $K(f, g)$ and the Hellinger distance by $d_H(f, g)$ (see Appendix A.3), then we define

$$K_Q(\zeta, \zeta_0) = \int K(\phi(\mu(x), \sigma), \phi(\mu_0(x), \sigma_0))Q(dx), \text{ and}$$

$$H_Q(\zeta, \zeta_0) = \int d_H(\phi(\mu(x), \sigma), \phi(\mu_0(x), \sigma_0))Q(dx),$$

where $\phi(\mu, \sigma)$ denotes the density of a normal distribution with parameters μ and σ. K_Q and H_Q are hence the average Kullback-Leibler divergence and the average Hellinger distance between the normal residual distributions, where the average is taken with respect to the distribution of the covariates. In addition denote the parameters ζ with H_Q distance larger than ϵ from ζ_0 as H_ϵ, *i.e.* $H_\epsilon = \{\zeta | H_Q(\zeta, \zeta_0) > \epsilon\}$. In the following we will list the conditions sufficient to achieve full support in Kullback-Leibler distance K_Q.

(A1) The prior for σ^2 has a strictly positive density on $[0, \infty)$.

(A2) The priors for $\boldsymbol{\beta}$ have full support on $\mathbb{R} \times \mathbb{R} \times \mathbb{R}_+$.

(A3) The base distribution function $G(., \boldsymbol{\xi})$ fulfills property (4.7).

(A4) The random probability measure $\mathbb{P}$ fulfills the support requirements stated in Theorem 2.1.4 and the base probability measure P_0 has strictly positive support on Ξ.

Now let Π_n^* be the posterior distribution of ζ. The following Theorem shows consistency of the posterior distribution in Hellinger distance, using the same method of proof as in Shively, Sager and Walker (2009), *i.e.* exploiting the fact that a consistent maximum likelihood estimator for convex regression exists.

Theorem 4.2.1 (Posterior Consistency)**.** *Under the Assumptions (A1)-(A4), the joint posterior distribution Π_n^* of $\mu(.)$ and σ fulfills for any $\epsilon > 0$*

$$\Pi_n^*(H_\epsilon) \xrightarrow{a.s.} 0 \text{ for } n \to \infty.$$

Proof. See Appendix B.3.

The posterior probability of the complementary event $\overline{H}_\epsilon = \{\zeta | H_Q(\zeta, \zeta_0) \leq \epsilon\}$ hence converges to one and the posterior concentrates on the true values $(\mu_0(.), \sigma_0^2)'$ in the H_Q distance. From this, one can directly conclude almost sure convergence of the Bayes estimate, *i.e.* the posterior mean, using the convexity of Hellinger distance and an application of Jensen's inequality, see Shively, Sager and Walker (2009) for details.

4.3 Simulation Study

In this section we will evaluate the finite sample performance of the approach with respect to estimation of the original function and its derivative in a simulation study and compare it with two alternative approaches. To simulate our data we used three different test functions and two different noise levels. The test functions are given by

$$\begin{aligned}
\mu_1(x) &= \exp(3x - 3) \\
\mu_2(x) &= \begin{cases} 0.1 - x + 2x^2 & x \leq 0.7 \\ 1.8x - 0.88 & x > 0.7 \end{cases} \\
\mu_3(x) &= 0.2 - x + 5 \int_0^x B(t, 100, 15)\, dt + 2 \int_0^x B(t, 5, 10)\, dt,
\end{aligned}$$

where $B(t, \alpha, \beta)$ denotes the distribution function of a beta distribution with parameters α and β. The functions as well as their first derivatives are displayed in Figure 4.1. The first function μ_1 is an exponentially increasing function and has a rather smooth

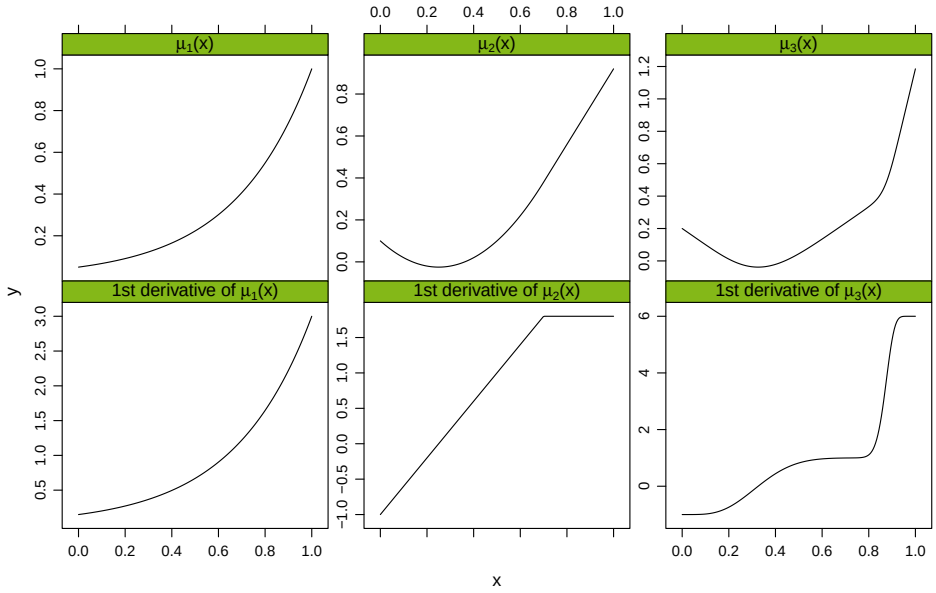

Figure 4.1: Testfunctions used for simulations (upper row) and corresponding first derivatives (lower row).

shape. The second function is included because it has a non-smooth first derivative with a sharp edge at $x = 0.7$, and is constant after that. The last test function μ_3 is included because it has a slightly more complex shape than the other two test functions. For each scenario we simulated 100 uniformly distributed covariates in $[0, 1]$ and conditional on those we simulated normally distributed observations with mean given by one of the three test functions and the standard deviation σ given by either 0.3 or 0.1.

We will use two alternative methods to compare the performance of the proposed methodology. The first one is the classical convex least squares estimate proposed by Hildreth (1954), as implemented in the `conreg` function of the `cobs` R package (Ng and Maechler 2008). Note that the fitted function for this approach is a piecewise linear function, and hence not smooth. As all functions in the simulation study are smooth we additionally used a second approach based on shape constrained penalized B-splines. We used cubic B-splines with 10 inner knots located at the $0.05, 0.15, 0.25, \ldots,$ $0.85, 0.95$ empirical quantile of the covariates. As noted for example by Dierckx (1980), for the B-spline basis there are simple linear constraints, which ensure convexity of a spline. These constraints can be imposed in a quadratic programming approach to obtain a convex estimate. To achieve a smooth estimate we used a difference penalty

as described in Eilers and Marx (1996). The smoothing parameter for the penalty is set to the value obtained by generalized crossvalidation for the unconstrained problem. For both penalized splines and the classical convex least squares estimate we used the derivative of the estimate as an estimate for the first derivative.

The approach we propose in this chapter is applied with the following prior selection: For the parameters $\boldsymbol{\beta}$ we use (possibly positive truncated) improper constant prior distributions (resulting in a convexity shape-constraint). The nonparametric part of the prior (4.3) is specified as follows: For $J-1$ we use a Poisson distribution with parameter $\lambda = 1.5$ (*i.e.* the prior mean for J is 2.5), corresponding to the experience that typically a fairly small number of basis functions is sufficient to model a monotonic function. The distribution of the weights $\boldsymbol{w}|J$ is chosen as a symmetric Dirichlet distribution with parameter $\delta = 0.5$. This specification of the priors for J and $\boldsymbol{w}|J$ leads to a $k_0 \approx 0.71$, so a quite uninformative selection of priors. For $G(.,\boldsymbol{\xi})$ we use the TSP distribution function. The parameters m and ν of the TSP distribution receive a $U(0,1)$ and a $U(1,50)$ distribution. This approximately corresponds to a linear increasing prior mean function for the derivative (and hence approximately to a quadratic polynomial for $\mu(.)$). The approach is implemented using MCMC techniques based on Gibbs sampling and a reversible jump step (Green 1995) for updating the function $F^*(.)$ (see Appendix C.2). For each simulation we used 5000 iterations burn-in, a total of 55000 iterations and a thinning rate of 5, giving a total of 10000 iterations. For estimating the mean function $\mu(.)$ and its derivative in this approach we use the pointwise median as an estimate.

We compare estimation of the function and its derivative through the mean absolute estimation (MAE) error:

$$MAE = \frac{1}{101}\sum_{i=0}^{101} |\mu(i/101) - \hat{\mu}(i/101)|.$$

The results of the simulation study can be found in Table 4.1. There it can be seen that the Bayesian approach, with the chosen weakly informative selection of priors performs quite well compared to the other approaches, particularly for Scenario 2. When it comes to estimation of the first derivative it becomes obvious that the estimation error increases quite dramatically for all methods. Estimation of derivatives seems con-

		Function			1st Derivative		
Testfct.	σ	ConvLS	ConvRS	ConvBayes	ConvLS	ConvRS	ConvBayes
$\mu_1(.)$	0.3	0.0530	0.0441	0.0451	0.6722	0.3722	0.3273
	0.1	0.0201	0.0179	0.0183	0.3441	0.2389	0.1801
$\mu_2(.)$	0.3	0.0538	0.0488	0.0477	0.7154	0.4927	0.3400
	0.1	0.0208	0.0197	0.0186	0.3636	0.2904	0.1852
$\mu_3(.)$	0.3	0.0571	0.0531	0.0575	0.9831	0.7436	0.6442
	0.1	0.0223	0.0233	0.0213	0.5313	0.5084	0.2970

Table 4.1: Mean absolute estimation error for true mean function and first derivative of the mean function (ConvLS $\hat{=}$ Convex least squares, ConvRS $\hat{=}$ Convex regression splines, ConvBayes $\hat{=}$ Convex nonparametric Bayes estimate), based on 1000 simulations.

siderably more difficult. However the penalized splines perform considerably better than the convex maximum likelihood estimator, which is probably due to the smoothness penalty. Additionally the Bayesian approach performs better than the other two approaches, particularly in Scenarios 2 and 3. We believe the good performance is mainly due to two reasons: (i) the Bayes estimate also uses a type of smoothness penalization through the prior distribution for J, and more importantly, the Bayesian estimate for the derivative is based on an averaging over posterior simulations from an MCMC sample (the pointwise median is used as the estimate). As the derivative is quite variable, it is reasonable that this averaging has a positive effect (in the sense of leading to a smoother and possibly also more reliable estimate) compared to taking just one particular estimate.

4.4 Length of Dugongs

In this section we will apply our methodology to a data set taken from Ratkowksy (1983). The data consist of length and age measurements of dugongs captured near Townsville in Queensland, Australia. Dugongs are large marine mammals living at the shores of the Indian Ocean and primarily in Australia. In this situation it is reasonable to assume that the growth of dugongs primarily happens at an early age and the growth rate gets less as the dugong gets older. Mathematically this is hence essentially a concavity restriction. Additionally it is reasonable to assume that the length is

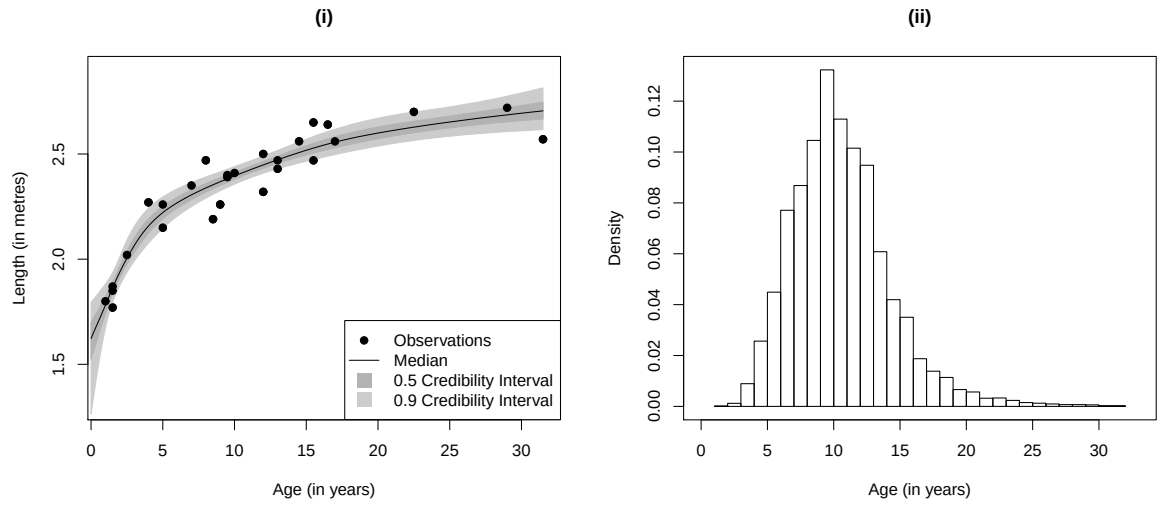

Figure 4.2: (i) Plot of the fitted data set and credibility intervals and (ii) the uncertainty distribution for a dugong of length 2.5.

a monotonic function of age.

Although prior information might be available for this particular application, we will use a weakly informative selection of priors. Hence for the parameters $\boldsymbol{\beta}$ we use constant improper priors and impose the constraints $\beta_1 \geq 0$ and $\beta_2 \geq 0$. The nonparametric part of the prior (4.3) is specified as follows: For $J-1$ we use a Poisson distribution with parameter $\lambda = 1$ (*i.e.* the prior mean for J is 2). The distribution of the weights $\boldsymbol{w}|J$ is chosen as a uniform distribution on the simplex. For this specification $k_0 \approx 0.74$, so a fairly high variability. For $G(.,\boldsymbol{\xi})$ again the TSP distribution is used. The parameters m and ν of the two-sided power distribution receive a $U(0,1)$ and a $U(1,50)$ distribution; note that this approximately corresponds to a linear increasing prior mean function for the derivative. As our code is designed to deal with convex increasing regression we reversed the x axis, negated the observed values and retransformed back to display the results.

We ran the MCMC algorithm described in Chapter C.2 for 200000 iterations after a burnin of 10000 and used a thinning rate of 10 resulting in 20000 simulations from the posterior. Convergence has been checked by running multiple chains from different starting values and observing that the results are consistent between the chains. In

Figure 4.2 (i) one can observe the credibility intervals for the fitted conditional mean function of length versus age. The uncertainty becomes largest in the sparse region of the data, *i.e.* for dugongs of an age higher than around 17 years. Here the monotonicity restriction is certainly beneficial for the estimation process, because the three observations with age larger than 17 point into a downwards direction. An unconstrained nonparametric model would certainly predict a downturn of the length at larger ages due to the lack of other information.

One interesting application of the data set would be to determine the age of the dugong (something usually quite difficult to obtain) from its length (which is easier to obtain). Particular in the Bayesian framework one can report not only a point estimate for the age for a given length but also an uncertainty interval, taking into account the uncertainty in model and data. Suppose for example one finds a dugong of length 2.4 metres and want to infer its age. Hence we need to account for two kinds of uncertainties, first the uncertainty induced from not knowing $\mu(.)$ and second the individual variation of the animal ϵ^*, which is also unknown. Essentially we thus want to find the value x^* for which $2.4 = \mu(x^*) + \epsilon^*$. For $\mu(.)$ we have a posterior sample and ϵ^* is normally distributed with mean 0 and standard deviation σ, where for σ also a posterior distribution is available. So posterior simulations for x^* can be produced from the MCMC output for $\mu(.)$ and σ. In Figure 4.2 (ii) one can observe the corresponding histogram of x^* values. The posterior median is given by 10.3 years and the 0.05 and 0.95 quantile are given by 5.4 years and 17.4 years, respectively, which is a surprisingly large uncertainty. It is interesting to observe that uncertainty is larger in the right tail: This is probably due to the fact that the data are relatively sparse for older dugongs and the fact that the curve flattens out for higher ages.

The traditional way of evaluating these data would be to use parametric non-linear regression models such as $\mu(x) = \alpha - \beta\gamma^x$ with $\alpha, \beta > 1$ and $\gamma \in (0, 1)$ as is done in the examples of the WinBUGS language (version 1.4). A usual concern about these parametric models however is the fact that there is no application specific motivation for using this (or any other) particular parametrization. Hence the uncertainty involved in the modelling process can greatly be underestimated.

4.5 Conclusions

In this chapter we have proposed a method for estimating a continuous differentiable function under derivative constraints. For this purpose we derived a representation of continuously differentiable functions with a monotonic derivative, in which shape constraints such as convexity or monotonic convexity reduce to simple finite dimensional constraints on the parametric part of the model. We think that this representation might also be of interest for other approaches to derivative-constrained inference.

Future work might apply this model, for example, in problems in economics or finance. We believe this model particularly has great potential for example for option pricing, where the first derivative of the modelled function is a probability distribution function, a constraint which our model can easily accommodate by choosing $\beta_1 = 0$ and $\beta_2 = 1$.

A limitation of the representation employed in this chapter is the fact that it allows essentially for monotonicity and positivity constraints on the first derivative, although obviously more general constraints might be desirable. One example of a constraint not covered by our approach is a unimodality restriction on the derivative. This would result in, so-called, ogive curves for the original function to be modelled (for example the non-linear sigmoid Emax model is an example of a parametric ogive model). Also a unimodal function can be modelled using a derivative constraint, here however the derivative constraint is more involved: The derivative of a differentiable unimodal function is first positive, then (possibly) negative, but after that not positive again. This seems difficult to represent with simple constraints.

In the next section we will relax the assumption of a parametric residual distribution. Additionally we will consider multivariate input variables and build a model for a stochastically ordered densities based on multivariate monotone functions (which is again a positivity constraint on the partial derivatives of the function).

Chapter 5

Stochastically Ordered Multiple Regression

This chapter generalizes Chapters 3 and 4 in two directions: (i) A nonparametric model is assumed for the residual density (based on a mixture of normal densities, as described in Section 2.1.1) and (ii) a nonparametric model for multivariate monotonic functions based on ridge functions (see Section 2.2.2) is constructed. In addition again the Ongaro-Cattaneo random measure, introduced in Section 2.1, will be used to build the model, but in this chapter it will be employed on a function space rather than a subset of Euclidean space. The application studied in this section is in epidemiology using data from the US Collaborative Perinatal Project. Specifically we investigate the gestational age at delivery (GAD) of newborns as a function of the risk factors DDE and PCB, while imposing a stochastic ordering constraint in the relationship between GAD and the risk factors. The paper Bornkamp, Ickstadt and Dunson (2009) is based on the material in this chapter.

5.1 Introduction

In many biomedical applications, subject-specific knowledge suggests that the conditional distribution of a response $y \in \mathbb{R}$ given predictors $\boldsymbol{x} \in \mathcal{X} \subset \mathbb{R}^k$ increases

(or decreases) stochastically with increasing $\boldsymbol{x}$. One example arises in epidemiology, where the exposure to toxic substances or environmental risk factors often can be assumed to be related to health risk in a monotonic way. A different example appears in clinical trials, where the effect of a pharmaceutical compound (or a combination of compounds or therapies) is assumed to be increasing with increasing dose level (or intensity of therapy). In these situations, it is natural to model the distribution of the response conditionally on covariates, such as age, as stochastically ordered with increasing value of the exposures. For ease of exposition we focus on the increasing case, but a stochastic decrease can be considered analogously.

Nonparametric modelling of stochastically increasing densities with respect to an ordered *categorical* covariate has recently been discussed quite extensively in a Bayesian framework by Gelfand and Kottas (2000), Hoff (2003a), Karabatsos and Walker (2007) and Dunson and Peddada (2008), among others. The generalization to a multivariate continuous predictor is considerably more difficult. When normality and homoscedasticity are imposed on the residual density, the problem reduces to estimation of an isotonic regression in multiple predictors (*e.g.* Dykstra and Robertson (1982)). Mukarjee and Stern (1994) and Dette and Scheder (2006) proposed to monotonize an unconstrained nonparametric regression fit. To reduce complexity in modeling of the multivariate surface subject to monotonicity constraints, additivity constraints can be imposed as in Bacchetti (1989), Morton-Jones, Diggle, Parker, Dickinson and Binks (2000) and Tutz and Leitensdorfer (2007) or more recently in Shively, Sager and Walker (2009) in a Bayesian framework.

Such methods focus on the mean of the response distribution, while in many applications the distribution tails may be of even greater interest. For example, in epidemiology, subjects in the right or left tail have an adverse health response. In order to assess how the entire conditional response distribution changes with predictors, it is important to avoid restrictive assumptions such as normality and homoscedasticity. Bayesian density regression methods, proposed by Müller, Erkanli and West (1996) and Dunson, Pillai and Park (2007) among others, allow the conditional response density to change flexibly with predictors. To address the curse of dimensionality problem, such methods borrow strongly across different regions of the predictor space.

Efficiency can substantially be improved through imposing stochastic ordering constraints. To our knowledge, Wang and Dunson (2009) proposed the only method to incorporate stochastic ordering over a continuous predictor in nonparametric density regression. Our focus is on generalizing their approach to allow multiple predictors, while incorporating ideas of Chapter 3 for building multivariate monotonic functions.

Section 5.2 describes our model and discusses its properties. Section 5.3 contains a simulation study, Section 5.4 applies the methods to an epidemiology data set, and Section 5.5 concludes.

5.2 Methodology

5.2.1 Model

Although there is a rich literature on multivariate stochastic ordering, the focus has been on multivariate responses. In this chapter we address the problem of nonparametric conditional distribution modeling subject to stochastic ordering in multiple predictors. We refer to the proposed order restriction as SO-$\mathcal{X}$, with $\mathcal{X}$ the (possibly multivariate) input space of the predictors. In particular, letting $F_{\boldsymbol{x}}(y)$ denote the conditional distribution function of y given predictors $\boldsymbol{x} \in \mathcal{X} \subset \mathbb{R}^k$, restriction SO-$\mathcal{X}$ corresponds to

$$F_{\boldsymbol{x}}(y) \geq F_{\boldsymbol{x}'}(y), \text{ for all } y \in \mathbb{R} \text{ and } \boldsymbol{x} \leq \boldsymbol{x}',$$

where $\boldsymbol{x} \leq \boldsymbol{x}'$ if and only if $x_m \leq x'_m$ for all $m = 1, \ldots, k$. This is a generalization of the stochastic ordering constraint for two probability distributions given in Appendix A.1.

Let $\mathcal{F}_{\mathcal{X}} = \{F_{\boldsymbol{x}}, \boldsymbol{x} \in \mathcal{X}\}$ denote an uncountable collection of continuous conditional distribution functions, with each $F_{\boldsymbol{x}}$ in $\mathcal{F}_{\mathcal{X}}$ having support on $\mathbb{R}$ and with $\mathcal{X} \subset \mathbb{R}^k$. We propose a prior for $\mathcal{F}_{\mathcal{X}}$, which will be a distribution over the set of all possible collections $\mathcal{F}_{\mathcal{X}}$ subject to restriction SO-$\mathcal{X}$. To induce such a prior, we propose to characterize each $F_{\boldsymbol{x}}$ as a location-scale mixture of Gaussians, with the variances con-

stant with $\boldsymbol{x}$ while the conditional means vary according to unknown multivariate monotone functions. Such a restriction on the component-specific mean functions is sufficient to ensure SO-$\mathcal{X}$, as is shown formally below.

Letting $f_{\boldsymbol{x}}$ denote the density corresponding to distribution function $F_{\boldsymbol{x}}$, we model the residual densities as

$$f_{\boldsymbol{x}}(y) = \int \phi(y, \mu(\boldsymbol{x}), \sigma^2) P(d\mu, d\sigma^2) = \sum_h \pi_h \phi(y, \mu_h(\boldsymbol{x}), \sigma_h^2), \tag{5.1}$$

where $\phi(y, \mu, \sigma^2)$ is the normal density with mean μ and variance σ^2, the $\mu_h : \mathcal{X} \to \mathbb{R}$ are multivariate monotonic functions satisfying $\mu_h(\boldsymbol{x}) \leq \mu_h(\boldsymbol{x}')$ for all $\boldsymbol{x} \leq \boldsymbol{x}'$, and P is a discrete mixing probability measure with support on $\mathcal{M} \times \mathbb{R}_+$, where $\mathcal{M}$ is the space of multivariate monotonic functions mapping from $\mathcal{X} \to \mathbb{R}$. In the following we take $\mathcal{X} = [0,1]^k$ without loss of generality for bounded predictors. By assuming the mixing measure is almost surely discrete, we hence obtain a countable mixture with π_h a probability weight on the hth component, which has associated mean function μ_h and variance σ_h^2. For each $\boldsymbol{x} \in \mathcal{X}$, the conditional density is expressed as a univariate Gaussian mixture, with the densities stochastically ordered due to the monotonicity of each μ_h. We focus on Gaussian mixtures as they are well-established and computationally tractable, note however, that most of the theory in this paper also applies to other kernels (for example the proof of Theorem 5.2.1, see Appendix B.4).

As a general prior for the discrete mixing measure on $\mathcal{M} \times \mathbb{R}_+$, we focus on the class proposed by Ongaro and Cattaneo (2004) (see also Section 2.1), which includes a broad variety of priors as special cases. A random probability measure belongs to this class when its realizations can be represented as

$$P(\cdot) = \sum_{h=1}^{N} \pi_h \delta_{\boldsymbol{\xi}_h}(\cdot),$$

where $\boldsymbol{\xi}_h, \pi_h, N$ are random variables as specified in Definition 2.1.2. The Dirichlet process with parameter MP_0 is obtained by setting $N = \infty$ and using the GEM distribution with parameter M for the weights π_h (Ishwaran and Zarepour 2002). A truncated Dirichlet process has a fixed N and a generalized Dirichlet distribution for the weights (see Remark 2.1.1 for details).

The following Lemma establishes that mixture model (5.1) induces the SO-$\mathcal{X}$ restriction on the conditional distributions and that any collection of continuous conditional distributions in SO-$\mathcal{X}$ can be approximated using (5.1).

Theorem 5.2.1 (Support)**.**

(i) Under model (5.1) the conditional distributions satisfy

$$F_{\boldsymbol{x}}(y) \geq F_{\boldsymbol{x}'}(y), \text{ for all } y \in \mathbb{R}, (\boldsymbol{x}, \boldsymbol{x}') \in \mathcal{X} \times \mathcal{X}, \boldsymbol{x} \leq \boldsymbol{x}';$$

(ii) Given a set $\tilde{\mathcal{F}}_{\mathcal{X}}$ of continuous distributions satisfying SO-$\mathcal{X}$ order, with conditional distribution functions $\tilde{F}_{\boldsymbol{x}}(y)$ on $\mathbb{R}$, there exist, for an arbitrarily small $\epsilon > 0$, π_h, $\mu_h(\boldsymbol{x})$ and σ_h^2 such that

$$\sup_{\boldsymbol{x} \in [0,1]^k} \left\{ \sup_{y \in \mathbb{R}} \left| \sum_{h=1}^{N} \pi_h \Phi(y, \mu_h(\boldsymbol{x}), \sigma_h^2) - \tilde{F}_{\boldsymbol{x}}(y) \right| \right\} \leq \epsilon + \frac{1}{N},$$

where $\Phi(y, \mu, \sigma^2)$ is the distribution function of a normal distribution with mean μ and variance σ^2.

Proof: See Appendix B.4.

Because the probability of having any observation exactly at a given $\boldsymbol{x}$ is zero for predictors having a continuous density, the ability to estimate $f_{\boldsymbol{x}}(y)$ necessarily relies on borrowing of information across different locations. We cannot simply define separate mixtures of normals for each location. Lemma 5.2.1 shows how the dependence arises through the prior, while also providing an expression for the prior expectation.

Lemma 5.2.1 (Prior Moments)**.** *Marginalizing out the random mixing measure P, the expectation of $F_{\boldsymbol{x}}(y)$ and the covariance of $F_{\boldsymbol{x}}(y)$ and $F_{\boldsymbol{x}'}(y)$ for $\boldsymbol{x}, \boldsymbol{x}' \in \mathcal{X}$ are given by*

$$\begin{aligned} E\{F_{\boldsymbol{x}}(y)\} &= \int \Phi(y, \mu(\boldsymbol{x}), \sigma^2)) dP_0, \\ Cov\{F_{\boldsymbol{x}}(y), F_{\boldsymbol{x}'}(y)\} &= k_0 \left\{ \int \Phi(y, \mu(\boldsymbol{x}), \sigma^2) \Phi(y, \mu(\boldsymbol{x}'), \sigma^2) dP_0 \right. \\ &\qquad \left. - \int \Phi(y, \mu(\boldsymbol{x}), \sigma^2) dP_0 \int \Phi(y, \mu(\boldsymbol{x}'), \sigma^2) dP_0 \right\} \end{aligned}$$

where $\Phi(y, \mu(\boldsymbol{x}'), \sigma^2)$ is the distribution function of a normal distribution with mean μ and variance σ^2, P_0 is a nonatomic probability distribution on $\mathcal{M} \times [0, \infty)$ and $k_0 \in [0, 1]$ is given by $E(\sum_{h=1}^{N} \pi_h^2)$.

Proof: See Theorem 2.1.3.

Hence the prior mean and the prior correlation structure is determined by the base measure P_0 alone, while the parameter k_0 of the random measure, jointly with P_0, determines the variability. In practice we need to specify the base measure P_0 of the nonparametric prior, consisting of a prior distribution H on the monotonic function space $\mathcal{M}$ as well as a prior distribution on $[0, \infty)$ for the variance parameter. Because standard choices can be used for the prior for the variance (e.g., inverse-gamma), we focus in the next section on how to choose H.

5.2.2 Prior for Multivariate Monotone Functions

Placing a prior on the space of multivariate monotonic functions is challenging. The use of multivariate basis expansions or tensor products of univariate bases quickly becomes infeasible as the dimension increases, because more and more basis functions are needed to obtain an adequate approximation (Barron 1993). Another challenging issue is how to impose monotonicity on the multivariate basis. A common strategy is to impose additional constraints to simplify the problem, with two such possibilities corresponding to additive models (where $\mu(x_1, \ldots, x_k) = \mu_1(x_1) + \ldots + \mu_k(x_k)$) or single index models (where $\mu(\boldsymbol{x}) = \mu^*(\boldsymbol{a}'\boldsymbol{x})$, with $\mu^* : \mathbb{R} \to \mathbb{R}$ and $\boldsymbol{a} \in \mathbb{R}^k$), see Section 2.2.2 for details. For additive models, monotonicity is imposed through restricting each univariate function to be monotonic, while for single index models, one can let $\boldsymbol{a} \in \mathbb{R}^k_+$ and μ^* be monotonic. Unfortunately, additive models do not allow interactions, and the single index model is constant on hyperplanes of the form $\boldsymbol{a}'\boldsymbol{x} = const$.

We propose to base our prior on linear combinations of ridge functions, $\sum c_j g_j(\boldsymbol{a}_j'\boldsymbol{x})$, where the $g_j : \mathbb{R} \to \mathbb{R}$ are univariate continuous functions and the $\boldsymbol{a}_j \in \mathbb{R}^k$ are direction vectors. As discussed in Chapter 2.2, linear combinations of sufficiently-flexible ridge functions can approximate any multivariate continuous function in sup norm, and are ideally suited for multivariate cases in requiring only a few ridge functions to characterize fairly complex relationships (Barron 1993). As a sufficient but not necessary condition to ensure monotonicity, we assume $c_j \in \mathbb{R}_+$, the $g_j(.)$ to be monotonic

and $\boldsymbol{a}_j \in \mathbb{R}^k_+$. We refer to the resulting class of functions as positive linear combinations of monotonic ridge (plcmr) functions. As it is not straightforward to find simple, and hence computationally-tractable, necessary restrictions for monotonicity and we find the plcmr class to be highly-flexible, we restrict consideration to this class. It is straightforward to show that all plcmr functions are multivariate monotone, with additive and single-index models arising as special cases.

We will carefully specify our prior on the space of plcmr functions on $[0,1]^k$ to facilitate interpretation and computation expressing the function $\mu(\boldsymbol{x})$ as

$$\mu(\boldsymbol{x}) = \beta_0 + \beta_1 \mu^0(\boldsymbol{x}), \tag{5.2}$$

with $\beta_0 \in \mathbb{R}$ the value at $\boldsymbol{x} = (0,\ldots,0)'$, $\beta_1 \in \mathbb{R}_+$ the maximum change between $(0,\ldots,0)'$ and $(1,\ldots,1)'$, $\mu^0(\boldsymbol{x}) = \int G(\boldsymbol{\alpha}'\boldsymbol{x},\boldsymbol{\xi})Q(d\boldsymbol{\alpha},d\boldsymbol{\xi}) = \sum_{j=1}^J w_j G(\boldsymbol{\alpha}_j'\boldsymbol{x},\boldsymbol{\xi}_j)$, Q a discrete mixing measure, $\boldsymbol{w} \in \mathbb{S}^J$, $\boldsymbol{\alpha}_j \in \mathbb{S}^k$ and G a univariate cdf on $[0,1]$ depending on parameters $\boldsymbol{\xi} \in \Xi$. Restricting $\boldsymbol{\alpha}_j$ to fall on the probability simplex has the advantage that automatically $\boldsymbol{\alpha}_j'\boldsymbol{x} \in [0,1]$ for any $\boldsymbol{\alpha} \in \mathbb{S}^k$ and any $\boldsymbol{x} \in [0,1]^k$. Hence $\boldsymbol{\alpha}$ measures the proportions of the total increase in the function $\mu^0(.)$ attributable to the different covariates. Lemma 5.2.2 provides a condition on the base distribution G under which a plcmr function can be approximated using (5.2).

Lemma 5.2.2. *Any plcmr function $\sum c_j g_j(\boldsymbol{a}_j'\boldsymbol{x})$ on $[0,1]^k \to \mathbb{R}$ can be approximated arbitrarily well in supremum norm by a function of form (5.2), provided*

$$\sup_{x\in[0,1]} \left| \sum_{j=1}^J w_j G(x,\boldsymbol{\xi}_j) - \tilde{G}(x) \right|$$

can be made arbitrarily small, for $\boldsymbol{w} \in \mathbb{S}^J$, $\boldsymbol{\xi}_j \in \Xi$ and any distribution function $\tilde{G}$ on $[0,1]$.

Proof: See Appendix B.5

In order to induce smoothness in the collection of conditional distributions over the predictor space, it is appealing to focus on continuous multivariate monotonic functions. In this case, the prior is dense in the space of continuous plcmr functions when the base distribution function G can approximate any continuous cdf on $[0,1]$ arbitrarily well. Several choices fulfill this property. One example is the distribution function

of the standard two-sided power (TSP) distribution of van Dorp and Kotz (2002), as used in Chapter 3, see Equation (3.5). The TSP cdf is sufficiently flexible (see Theorem 3.2.1), numerically straightforward to evaluate and available in a closed form (unlike for example the beta cdf).

Assuming the functions follow (5.2), a specification of the prior H is completed with parametric priors for β_0 and β_1 and a nonparametric prior for the mixing distribution Q based on Ongaro and Cattaneo (2004). A typical choice is to use $J - 1 \sim Poi(\rho)$, while the components $(m, \nu, \boldsymbol{\alpha})'$ of the base measure Q_0 are chosen to match prior information and prior uncertainty. A useful tool is to simulate the prior distribution and assess whether the resulting simulations lead to a-priori plausible results. A default choice in this setting are uniform distributions on reasonable subsets of the parameter space.

5.2.3 Implementation

In this section we describe the implementation and specific priors used. Assume we observe independently distributed data $(y_i, \boldsymbol{x}_i, \boldsymbol{z}_i)$, $i = 1, \ldots, n$, where y_i is a univariate response, $\boldsymbol{x}_i \in [0,1]^k$ are the covariates which are in a multivariate monotonic relationship with respect to y_i and $\boldsymbol{z}_i \in \mathbb{R}^p$ are additional unconstrained covariates we would like to adjust for in the analysis.

For the mixing measure P (from Equation (5.1)) we use the truncated Dirichlet process with parameter MP_0, which provides an accurate approximation to the Dirichlet process, while facilitating an efficient implementation via a blocked Gibbs sampler (Ishwaran and James 2001). We choose the truncation level $N = 20$, which provides a conservative upper bound on the number of mixture components occupied by individuals in the sample. The resulting model for the data is

$$P \sim DP_N(MP_0),\ P = \sum_{h=1}^{N} \pi_h \delta_{(\mu_h(\boldsymbol{x}), \sigma_h^{-2})}$$

$$y_i | \boldsymbol{x}_i, \boldsymbol{z}_i, P \overset{iid}{\sim} \sum_{h=1}^{N} \pi_h \phi(\mu_h(\boldsymbol{x}_i) + \boldsymbol{\gamma}' \boldsymbol{z}_i, \sigma_h^2)$$

where $DP_N(MP_0)$ denotes the truncated Dirichlet process with parameter MP_0 and N

components. The weights π_h have the truncated stick-breaking representation $\pi_h = V_h \prod_{l<h}(1 - V_h)$ with $V_h \overset{iid}{\sim} Beta(1, M)$ and $\pi_N = 1 - \sum_{h=1}^{N-1} \pi_h$. The atoms in the mixture (μ_h, σ_h^{-2}) are iid realizations of the base measure P_0 with $P_0 = H \times Exp(\omega)$, H is the prior on the space of plcmr functions and the $\mu_h(\boldsymbol{x})$ are given by $\mu_h(\boldsymbol{x}) = \beta_{0h} + \beta_{1h} \sum_{j=1}^{J_h} w_{hj} G(\boldsymbol{\alpha}'_{hj}\boldsymbol{x}, m_{hj}, \nu_{hj})$. We adjust for possible additional predictors $\boldsymbol{z}_i$ linearly.

For β_{0h} a normal distribution with parameter m_0 and variance ν_0^{-1} will be used. The parameter m_0 in turn has a normal prior with mean w_0 and variance τ_0, while ν_0 has a gamma prior with parameters a_{ν_0}, b_{ν_0}. As a common focus is in assessing whether the predictors have any effect on the response distribution, it is important to allow a completely flat relationship. This can be accomplished through using a mixture of a point mass at 0 and an exponential distribution with parameter λ as the prior for β_{1h}. The mixing probability π_0 is given a Beta(a_{π_0}, b_{π_0}) hyperprior, while λ is given a gamma(a_λ, b_λ) hyperprior. These hyperpriors induce a heavier-tailed and hence a more robust specification.

In specifying the prior for the mixing distribution Q in Equation (5.2), we assign the number of components J_h a Poisson(ρ) distribution shifted by one. The hierarchical prior for ρ is given a gamma(a_ρ, b_ρ) hyperprior. The weights w_j in the mixture follow a uniform distribution on $\mathbb{S}^J$ for each J. For the base measure Q_0 we use the following distribution $U(0,1) \times U(1,20) \times D(\mathbf{1})$ for the parameter $(m, \nu, \boldsymbol{\alpha})$, where $D(\mathbf{1})$ is the $(k-1)$-dimensional Dirichlet distribution with parameter $(1, \ldots, 1)'$ *i.e.* a uniform distribution on the simplex. This corresponds to the prior assumption that all variables are equally important a-priori and ensures an approximately linearly increasing prior mean function for the univariate function μ, with a reasonable variability.

For the precisions σ_h^{-2} an exponential prior with parameter ω is used, where ω has a gamma hyperprior with parameters a_ω, b_ω. The precision parameter M of the truncated Dirichlet process is also treated as unknown and receives a conjugate gamma hyperprior with parameters a_M, b_M. As a prior for the additional covariates $\boldsymbol{\gamma}$ a multivariate normal prior is used with mean $\boldsymbol{\mu}_\gamma$ and covariance matrix $\boldsymbol{\Sigma}_\gamma$.

To analyze the model, MCMC techniques based on the blocked Gibbs sampler will be used. This algorithm introduces a latent class membership variable K_i with categories

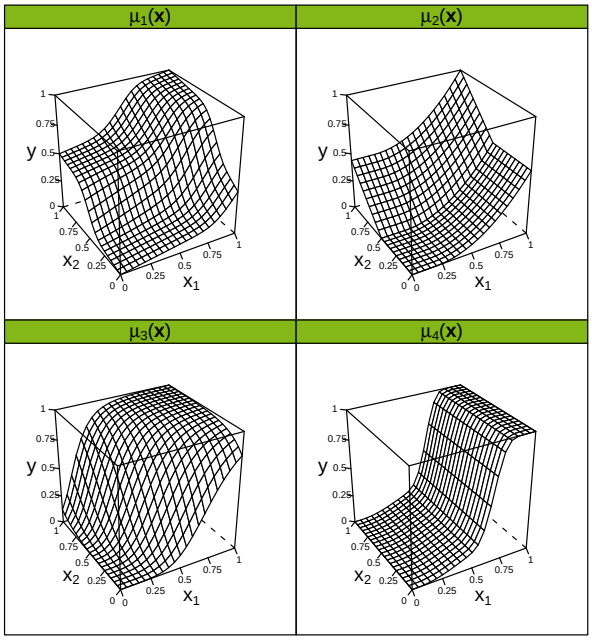

Figure 5.1: Test functions used in the simulation study.

$1, \ldots, N$ for each observation and iterates between updating the class memberships variables and the class specific parameters. Most of the class specific parameters can be updated by Gibbs steps, while an RJ-MCMC step is used to update the functions $\mu_h^0(.)$. Additionally the hyperparameters are updated in Gibbs steps, which is possible because conjugate hyperpriors were used. Appendix C.3 contains a detailed description of the MCMC algorithm.

5.3 Simulation Study

In this section we will evaluate the performance of the methodology with respect to estimation of the conditional response density and the conditional mean function. For this purpose we consider four simulation scenarios. For each scenario 250 uniform random covariates $\boldsymbol{x}$ in $[0,1]^2$ are generated and the response values were then simulated according to the following four densities

1. $f(y|\boldsymbol{x}) = \phi(0, 0.15^2)$

2. $f(y|\boldsymbol{x}) = \phi(\mu_1(\boldsymbol{x}), 0.05^2)$

3. $f(y|\boldsymbol{x}) = 0.6\phi(\mu_2(\boldsymbol{x}), 0.09^2) + 0.4\phi(0, 0.05^2)$

4. $f(y|\boldsymbol{x}) = 0.7\phi(\mu_3(\boldsymbol{x}), 0.06^2) + 0.3\phi(\mu_4(\boldsymbol{x}), 0.08^2)$.

In Scenario 1 a null model with no effect of the covariates $\boldsymbol{x}$ is used, while in Scenario 2 there is only one mixture component. In Scenario 3 only part of the population shows an effect with increasing covariates, while in Scenario 4 there are two groups within the population reacting differently to the predictors $\boldsymbol{x}$. The multivariate functions used for the scenarios are given by (see also Figure 5.1)

$$\begin{aligned}
\mu_1(\boldsymbol{x}) &= 1/3B(0.5x_1 + 0.5x_2, 1, 1) + 1/3B(0.7x_1 + 0.3x_2, 20, 10) + \\
&\quad +1/3B(0.15x_1 + 0.85x_2, 15, 20) \\
\mu_2(\boldsymbol{x}) &= \max(x_1 - 0.2, 0)^2/0.64 + \max(x_2 - 0.5, 0)/0.5 \\
\mu_3(\boldsymbol{x}) &= \frac{z^5}{z^5 + 0.3^5}, \text{ where } z = (x_1 + 0.5x_2 + x_1x_2)/2.5 \\
\mu_4(\boldsymbol{x}) &= 0.8B(x_1, 30, 15) + 0.2B(x_1, 3, 4),
\end{aligned}$$

where $B(x, \alpha, \beta)$ is the distribution function of the beta distribution with parameters α and β. The functions represent a selection of rather smooth functions with a subtle interaction structure (μ_1, μ_3), a non-smooth additive function (μ_2) and a function in which only one of the two predictors has an increasing effect (μ_4), with a rather sudden change from baseline to maximum response.

We will apply our methodology with the following weakly informative setting of the hyperpriors: $w_0 = 0, \tau_0 = 10, a_{\nu_0} = 0.5, b_{\nu_0} = 0.5, a_\lambda = 1, b_\lambda = 1, a_\rho = 1, b_\rho = 1.5, a_{\pi_0} = 1, b_{\pi_0} = 1, a_M = 1, b_M = N, a_\omega = 0.5, b_\omega = 0.5$. The components of the prior for the mixing distribution Q (used in the prior for $\mu^0(.)$) are chosen exactly as specified in Section 5.2.3. To analyze the model we use the algorithm described in the last section for 5000 iterations after a burn-in of 1000 iterations for each simulated data set. Every simulation scenario has been repeated for 500 simulations.

In Figure 5.2 the mean of the pointwise 0.05, 0.5 and 0.95 quantiles of the conditional density at the locations $(1/3, 1/3)'$, $(1/3, 2/3)'$, $(2/3, 1/3)'$ and $(2/3, 2/3)'$ can be observed (averaged over the 500 simulations). It can be seen that the methodology in

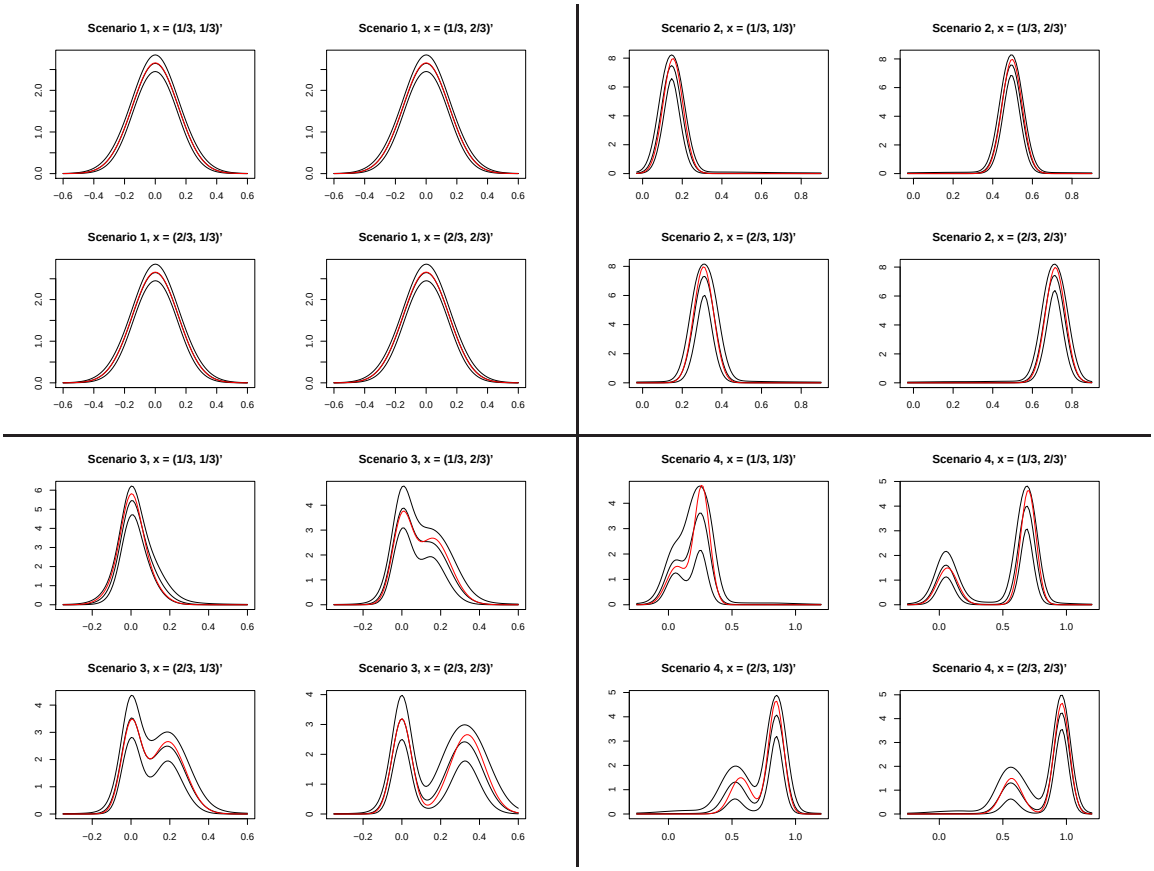

Figure 5.2: Mean 0.05,0.5 and 0.95 posterior quantile of the conditional density over 500 simulations (solid grey lines) and true conditional density (red line) at four locations in the input space for each of the four scenarios.

almost all situations nicely recovers the shape of the distribution. As one might expect, the performance in the more complex Scenarios 3 and 4 is slightly worse than in Scenarios 1 and 2.

In addition we evaluate the performance for estimating the conditional mean function. For this purpose we used the posterior mean of the conditional mean function as an estimate in each simulation scenario and evaluated the absolute estimation error with respect to the true conditional mean function at 55 uniformly distributed locations in the predictor space. In Figure 5.3 one can observe the mean absolute estimation error at these 55 locations. There it can be seen that for Scenarios 1 and 2 the methodology works well for all of the input space. In Scenario 3 and 4 the largest estimation errors occur at the boundary of the input space, when there is a steep increase (which is probably due to the fact that the data are sparse in these regions). Additionally

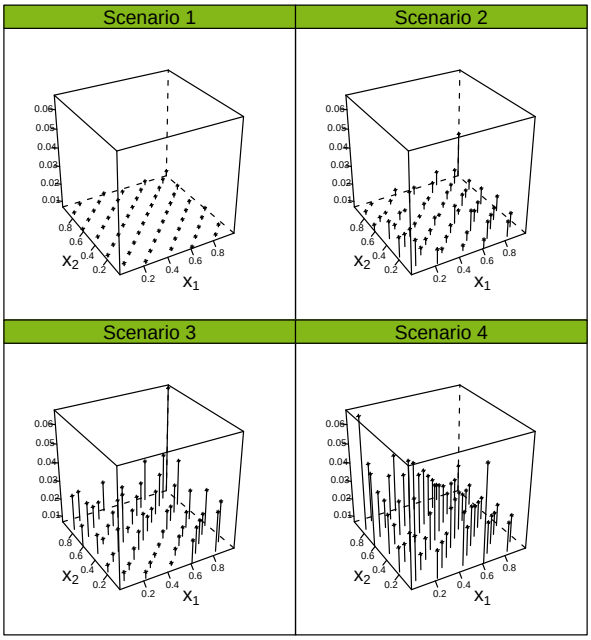

Figure 5.3: Mean absolute error for estimating the conditional mean function in $[0,1]^2$.

we compared the results obtained by monotonic density regression with other standard approaches for multivariate regression, which are typically designed for estimating the mean. The results of our methodology can be found in Table 5.1 under the name SOMR. We report results here only for quadratic local polynomial regression (LP) and a bivariate spline fitting approach (BS), which are the two methodologies that performed best of the classical approaches we investigated (we also investigated Projection Pursuit and MARS). For local quadratic regression we used the `locfit` package (Loader 2007) in `R`. A nearest-neighbour bandwidth selected by generalized cross-validation was chosen, since it lead to the best performance. For the bivariate spline, we used the `mgcv` package (Wood 2009) with the default setting for smoothing parameter selection. We employed the monotonic rearrangement described by Chernozhukov, Fernández-Val and Galichon (2007) to monotonize the LP and BS estimated functions. Univariate rearrangement was applied in both directions, and to eliminate order dependence we take the average over both possibilities. The results of the monotonized local polynomials and monotonized bivariate spline can be found in Table 5.1 under the names LP-Mon and BS-Mon. All approaches were applied for 500 simulation runs.

Method	LP	LP-Mon	BS	BS-Mon	SOMR
Scenario 1	0.0194	0.0196	0.0147	0.0147	0.0081
Scenario 2	0.0152	0.0143	0.0137	0.0135	0.0121
Scenario 3	0.0272	0.0249	0.0271	0.0258	0.0188
Scenario 4	0.0386	0.0366	0.0373	0.0361	0.0271

Table 5.1: Mean absolute estimation error for the conditional mean function.

The results of the simulation study can be observed in Table 5.1. There it can be seen that density regression performs best compared to the other approaches. The performance of the standard approaches gets close to SOMR in Scenario 2, probably because this is the scenario for which these type of approaches are typically designed for. Among the classical approaches both the rearranged bivariate spline and rearranged local polynomial estimate work quite well, with some slight advantage for the rearranged spline. The good performance of SOMR is not entirely surprising as the model used for data generation is quite similar to the Bayesian model we used for evaluation of the simulations. However, the main purpose of these simulations for us was (i) to see whether our methodology recovers the truth in realistic simulations before applying it to real data and (ii) to ensure that the results are at least as good as alternative approaches in cases in which the assumed ordering holds.

5.4 Application to Epidemiologic Data

In this section we apply our methodology to data from the US Collaborative Perinatal Project, which was conducted from 1959 to 1966. In the 1990s a random sample of blood sera of the participants were reanalyzed for potential toxic substances, see Longnecker et al. (2001) or Longnecker et al. (2005). We focus on the relationship between DDE (a metabolite of DDT) and PCB in the blood serum of the mother and the gestational age of the newborn at delivery (GAD). Dichloro-Diphenyl-Trichloroethane (DDT) is a pesticide which was primarily used as an agricultural insecticide and has been mostly been banned in the 1970s, although it is still in use in some developing countries. Polychlorinated biphenyls (PCB) are organic compounds that were primarily used in electrical equipment, and have been associated with a wide range of ad-

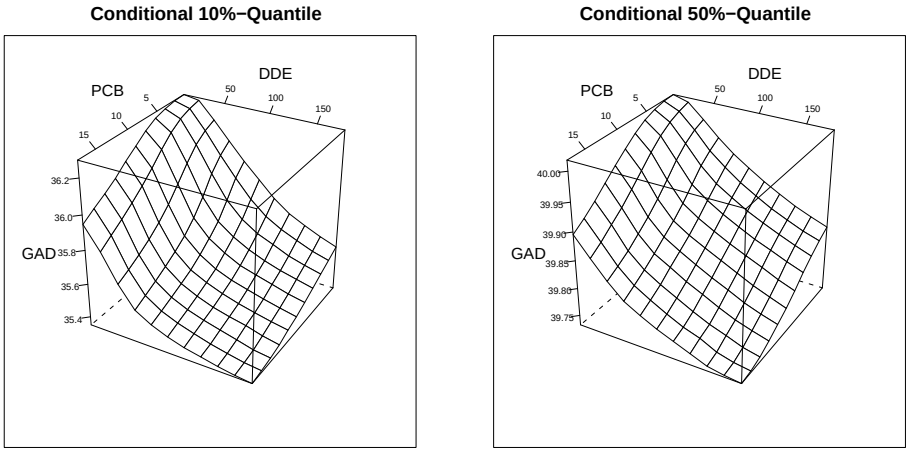

Figure 5.4: Posterior median of the conditional 10% and 50% quantile.

verse health effects. Note that both toxic substances were still in use in the United States when the data were collected.

Here we focus on GAD (in weeks) in relationship to DDE (in μg/L) and the total serum PCB (in μg/L). For model fitting we reversed and scaled these two predictors into the interval $[0, 1]$ and for the results transformed back. As additional unconstrained covariates we include the serum triglycerides (in μg/L) and the binary inputs smoking habit (1 = smoking) and race (1 = black). Triglycerides were standardized before model fitting. In the analysis we excluded all values with length of gestation longer than 45 weeks (approx. 10 months) for plausibility reasons and 68 cases with missing values, leaving a total sample size of 2312 for analysis.

For the priors we chose $w_0 = 30$, $\tau_0 = 10000$, $a_\lambda = 0.01$, $b_\lambda = 0.01$, $a_\omega = 0.01$, $b_\omega = 0.01$, $a_{\nu_0} = 0.1$, $b_{\nu_0} = 0.1$, to adapt the prior parameters to the right scale. All other parameters received the same prior distributions as in the simulation example. The prior for γ was chosen as a multivariate normal with mean vector $\mathbf{0}$ and diagonal covariance matrix $6.7\mathbf{I}_{3\times 3}$, where 6.7 is an estimate of the approximate variance in the observations. The prior for γ hence approximately reflects the information obtained in one observation.

We ran three independent chains of the MCMC sampling algorithm of Section 2 for 110000 iterations after using a burn-in of 10000 iterations and a thinning of 10, leaving

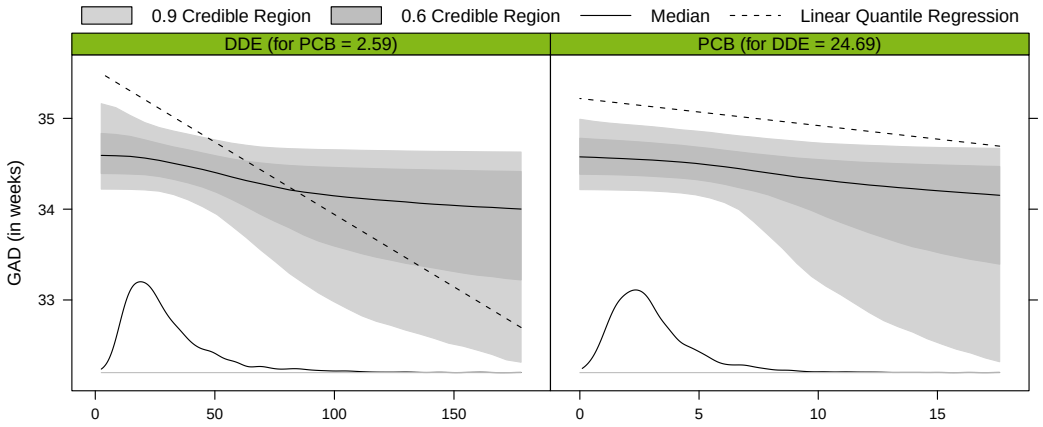

Figure 5.5: Posterior of the conditional 5% quantile, (scaled and shifted) kernel density estimates of the covariate distribution and a linear 5% quantile regression fit.

a total of 10000 iterations per chain. The results between the chains were consistent hence the presented analysis is based on the last 3500 iterations per chain resulting in a total of 10500 simulations.

Figure 5.4 plots the bivariate posterior median of the 50% and the 10% quantile of the conditional distribution against DDE and PCB, when the additional covariates are set to 0. There it can be seen that both substances seem to affect the gestational age at delivery only slightly, with a steeper decrease in the direction of DDE for both the 10% and the 50% quantile. Comparing the 10% and the 50% conditional quantile it becomes obvious that the 10% conditional quantile is affected slightly stronger by an increasing DDE and PCB, as the posterior median is decreasing steeper and stronger in overall effect for the conditional 10% quantile (in particular in the DDE direction).

Figure 5.5 shows the posterior distribution of the conditional 5% quantile for DDE (holding PCB fixed at its median) and PCB (holding DDE at its median value), and all other covariates are set to 0. It can be seen that uncertainty in the estimate is quite large, in particular for DDE values larger than 50 and PCB values larger than 5. This can be attributed to the fact that most of the participants in the study had rather small PCB and DDE values, which is illustrated in the figure by including the (scaled and shifted) kernel density estimates of the covariates DDE and PCB. Primarily for DDE there seems to be an effect for persons with high exposure (*i.e.* larger than 40), but

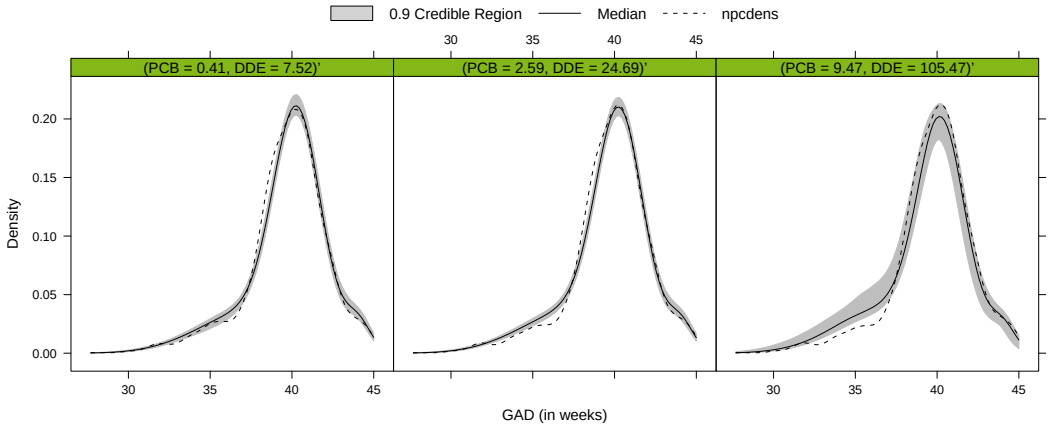

Figure 5.6: Posterior distribution of the conditional densities at three locations in the input space, together with the estimate of the `npcdens` function.

this effect cannot be estimated with high precision, as data are relatively sparse in this region. Using the `rq` function in the `quantreg` R package (Koenker 2008), we also fitted a parametric quantile regression model to the data (using linear effects for DDE and PCB and the same additional covariates) and the results are superimposed in Figure 5.5. Even though quantile regression is based on quite a different statistical model, results of quantile regression roughly agree with our results, but due to the linearity the quantile regression fails to give a detailed information on the shape of the relationship. We also fitted more complex effects (for example quadratic or cubic) for DDE and PCB, but it turned out to be difficult to get stable results, as the results for quantile regression are quite variable for the 5% quantile. This is probably due to the fact that mostly information about the 5% quantile is used for quantile regression, as opposed to explicitly modelling the whole conditional residual distribution, as in our methodology. Figure 5.6 shows the conditional densities at different locations in the predictor space. For this purpose we are looking at the conditional distribution, when both DDE and PCB are at their median value and at two extreme quantiles (the 1% and the 99% quantile). There it can be seen that the shape of the residual distribution looks relatively non-normal, with a more pronounced left tail. In the simulations typically two to five components were occupied (with modal value three). It is interesting to see that the shape of the residual density largely remains identical throughout the predictor space, only the uncertainty intervals are larger in parts, where the data are sparser

Covariate	0.05 Quantile	Median	0.95 Quantile
Triglycerides	-0.29	-0.22	-0.14
Smoking habit	-0.27	-0.12	0.02
Race	-0.78	-0.62	-0.47

Table 5.2: Posterior summaries of additional covariates.

(see also Figure 5.5). It also seems that there is a tendency that the left tail gets slightly more pronounced, in particular at the extreme quantiles of the predictor space. This is in accordance with the results in Figure 5.4, where we observed that the 10% conditional quantile is more effected by DDE and PCB than the conditional 50% quantile. Superimposed one can find the conditional density estimate using the `npcdens` function in the `np` R package (Hayfield and Racine 2008), which implements an unconstrained non-Bayesian method for kernel estimation of a set of conditional densities (see also Hall, Racine and Li (2004)). The fixed bandwidth was selected by maximum likelihood cross-validation. Both methods obtain rather similar results, with the main difference being in the left tail. Here the Bayesian approach is less wiggly, which is at least partially due to the implicit averaging over the posterior simulations in the Bayesian approach (rather than using one particular point estimate), additionally the conditional density is considerably smaller in the left tail for larger values of the input. This is most likely due to the fact that stochastic ordering is imposed in our methodology, while the alternative approach is unconstrained.

It is also interesting to compare the results with those obtained by Wang and Dunson (2009), who modelled the conditional density of GAD versus DDE with univariate monotonic density regression. The posterior medians for the conditional densities are quite similar between the approaches, while the variability intervals for the conditional densities are wider in Wang and Dunson (2009). This is probably due to the fact that the bivariate shape constraint employed here restricts the conditional density considerably more than in the one-dimensional case and hence reduces uncertainty in estimation. Table 5.2 contains the credibility intervals for the parameter estimates γ corresponding to the additional covariates. There it can be seen that both race and triglycerides have an impact on GAD, while for smoking habit there seems to be a less pronounced negative effect, as its credibility interval contains zero.

5.5 Conclusions

In this chapter we presented a model for estimating conditional densities under the SO-$\mathcal{X}$ stochastic order, *i.e.*, the stochastic ordering is assumed with respect to multivariate continuous predictors. The model relies on representing the conditional distributions as a location-scale mixture of normal distributions and the stochastic ordering constraint is imposed by assuming that the means of the components in the mixture are multivariate monotonically increasing functions. This type of model is extremely flexible, in particular we show that any collection of conditional densities under SO-$\mathcal{X}$ stochastic order can be approximated arbitrarily well by the proposed model. The model relies on a prior distribution for multivariate monotonic functions and we used positive linear combinations of monotonic ridge functions for this purpose. This class is quite flexible (compared to additive or single-index models for example) and seems well suited for sparse representation of multivariate functions.

Typical regression models focused on characterizing predictor effects on the center of the response distribution are insufficient in some applications. This is particularly the case when the tails of the response distribution are of primary interest. For example, in many applications, the greatest interest is in the extremes corresponding to unusual health responses, pollution levels, financial events or weather conditions. In such settings, most of the literature has focused on either using quantile regression models that focus on a single quantile (e.g., 95th) or models for extremes that effectively discard all information below a certain quantile. By using density regression methods, one simultaneously models all quantiles and hence allows inferences on differing predictor effects on the center and extreme quantiles, while using all the available data. A concern in density regression is the curse of dimensionality, as it is challenging to allow the response distribution to change flexibly over the predictor space. The incorporation of stochastic ordering constraints in multiple predictors is a highly effective strategy for reducing the effective dimensionality of the problem.

Chapter 6

Summary and Outlook

A little uncertainty is good for everyone.
Henry Kissinger

In this thesis we have developed methods for nonparametric Bayesian analysis under shape constraints. Chapter 2 reviews Bayesian nonparametric methodologies for probability distributions and general functions, and briefly discusses the asymptotic behaviour of BNP methods. Then in Chapters 3, 4 and 5 three novel nonparametric Bayesian models for shape constrained inference have been developed. The first model in Chapter 3 is built for nonparametric monotone regression under a parametric normality assumption on the residual distribution. For this purpose we derive a representation of monotonic functions, which is fairly interpretable and rich enough to ensure full support for monotone continuous functions in sup-norm. While a monotonicity constraint can be adequate in a variety of modelling situations, we focused on pharmaceutical dose-finding trials and growth curves to illustrate the underlying methodology. Chapter 4 develops an approach to model convex (and possibly monotone) functions. For this purpose we extend the representation derived in Chapter 3 to the case of convex functions, and consider the asymptotic behaviour of the underlying nonparametric posterior distribution. The particular modelling example we consider in this section are body length data in biology. Chapter 5 finally extends the other two chapters in two directions: (i) We relax the assumption of a parametric residual density, and model it as a mixture of normal distributions and (ii) we consider multivariate inputs, requiring the development of a nonparametric prior distribution for

multivariate monotone functions. The method regresses the whole residual density (rather than just the conditional mean) against the inputs x and employs a stochastic ordering constraint on the residual distribution. In this chapter we analyse data from the US perinatal project, where the distribution of the gestational age of newborns is studied as a function of two toxic substances, DDE and PCB.

A notion, appearing repeatedly in this thesis has been the question of full support of Bayesian nonparametric models. While for parametric models one is typically interested in full support on a finite dimensional space (for example $\mathbb{R}^k$), in the nonparametric situation one wants to achieve full support on spaces of functions (for example the space of continuous monotone or differentiable convex functions in Chapters 3 and 4 or the space of stochastically ordered continuous densities on $\mathbb{R}$ in Chapter 5). Practically full support means that the BNP model puts positive prior probability to all relevant residual models under consideration, so that the posterior is able to concentrate its mass on the true residual model. This is a fairly intuitive practical requirement, but it also appears as a rigorous requirement for consistency in an asymptotic study of BNP methods (see Chapters 2.3 and 4).

The shape constraints considered in this thesis can, in essence, all be reduced to a positivity constraint on a function or a (partial) derivative. As discussed in Chapter 2.2, having a model for a positive function one can achieve monotonicity and convexity (or concavity) constraints simply by integration (also log-concavity can be covered along this route). A shape constraint, which cannot be treated with this approach is unimodality. The derivative of a differentiable unimodal function is first positive, then (possibly) negative, but after that not positive again. This type of derivative constraint is relatively challenging to translate into simple constraints. This might be one of the reasons, why the literature on unimodality assumptions is relatively sparse, although it is adequate in a variety of situations. In phase II dose-finding trials, for example, there might be a downturn of the dose effect at larger doses, due to potential toxicity of the compound. Future work might concentrate on the unimodality restriction both for densities and functions.

Bayesian nonparametrics in general is currently still a rapidly evolving field, which

shows great promise for a *realistic* modelling of complex dependences and data structures. In the following we would like to discuss two points related to BNP, which we find important, but have not received much attention in this thesis (and usually do not receive much attention in the literature).

This thesis deals with nonparametric models and those are typically less prone to violating modeling assumptions than parametric models. Nevertheless we think that model checking is important, particularly for complex applications and models. A simple way to check the adequacy of a Bayesian model, with respect to the data set under consideration, is to simulate data sets from the posterior predictive distribution of the model and observe, whether the data are "similar" to the actually observed data, as measured by features of the underlying data (these feature will usually be problem specific). Another option is to apply cross-validation methods, *i.e.* splitting the data set in training and testing part and see how well the model predicts data from the testing set.

While Bayesian nonparametrics is a field that often convinces through its innovative solutions in difficult applied problems, a main hindrance for routine practical application is the lack of easy-to-use software (usually one needs to code models by oneself). The `BUGS` language can successfully handle many (and complex) parametric problems fairly well, and is an irreplaceable tool to implement Bayesian inference in practice (in academia *and* industry). It is however not designed for nonparametric problems and can handle only a few of them. A nonparametric Bayesian analogue of a general purpose software is currently missing. A software package that comes close to this, is certainly the `DPpackage` (Jara 2009) for `R`, which offers a huge variety of specific nonparametric Bayesian models and priors, but the user cannot specify an (almost) arbitrary model as in `BUGS`.

This thesis illustrates through both a broad range of applications considered and their appealing theoretical properties that shape constraints (when appropriate) are an extremely useful tool to narrow down the effective complexity of a nonparametric model. The prior (and hence also the posterior) can then concentrate on the relevant part of the parameter space, leading to more efficient inference. The main argument for shape constraints from a practical perspective is the fact that they are usually directly moti-

vated by the basic science underlying the considered application, while other type of modeling assumptions (such as, for example, parametric assumptions) often lack such a motivation. From a more theoretical perspective shape constraints are appealing as they usually can directly be exploited in an asymptotic analysis of the nonparametric model, as illustrated in Chapter 2.3.

Appendix A

Complementary Material

The more constraints one imposes, the more one frees one's self of the chains that shackle the spirit. The arbitrariness of the constraint only serves to obtain precision of execution.
Igor Stravinsky (1882-1971)

In this section we provide some additional material, which is too long to be presented in the text directly, but is needed for a better understanding for parts of this thesis.

A.1 Shape Constraints

Here we give a review of different shape constraints used in this thesis.

Monotonic Functions
A function $\mu : \mathbb{R}^k \to \mathbb{R}$ is monotone increasing, whenever

$$\mu(\boldsymbol{x}) \leq \mu(\boldsymbol{x}') \text{ for } \boldsymbol{x} \leq \boldsymbol{x}', \tag{A.1}$$

where $\boldsymbol{x} \leq \boldsymbol{x}'$ if and only if $x_m \leq x'_m$ for all $m = 1, \ldots, k$. A function is monotone decreasing, when (A.1) holds with the first "$\leq$" replaced by "$\geq$". When the function μ is monotonic increasing and differentiable it follows that all partial derivatives are positive, $\frac{\partial}{\partial x_m}\mu(\boldsymbol{x}) \geq 0$.

Convex Functions
A function $\mu : \mathcal{C} \to \mathbb{R}$ defined on a convex set $\mathcal{C} \subset \mathbb{R}^k$ is convex, when

$$\mu(t\boldsymbol{x} + (1-t)\boldsymbol{x}') \leq t\mu(\boldsymbol{x}) + (1-t)\mu(\boldsymbol{x}') \text{ for any } t \in [0,1] \text{ and any } \boldsymbol{x}, \boldsymbol{x}' \in \mathcal{C}.$$

In the one dimensional case ($k = 1$) this means that a convex function μ between points $x_0 < x_1$ from $\mathcal{C}$ always lies below the function linearly interpolating x_0 and x_1. If the convex function is differentiable on $\mathcal{C}$ and $k = 1$ the first derivative of μ is a monotone increasing function on $\mathcal{C}$ and the second derivative a positive function.

A function $\mu : \mathcal{C} \to \mathbb{R}$ is concave, when $-\mu$ is a convex function.

Unimodal Functions
A function $\mu : \mathbb{R} \to \mathbb{R}$ is unimodal with mode m, when μ is monotone increasing on $(-\infty, m)$ and monotone decreasing on (m, ∞).

Similarly a probability measure is unimodal, when its probability density function (or probability mass function) is unimodal.

Stochastic Ordering
A probability distributions $\mathcal{P}_1$ on $\mathbb{R}$ is stochastically larger than another probability distribution $\mathcal{P}_2$ on $\mathbb{R}$, whenever

$$\mathcal{P}_1((-\infty, x]) \leq \mathcal{P}_2((-\infty, x]),\ \forall x \in \mathbb{R}.$$

This means that for each $x \in \mathbb{R}$, $\mathcal{P}_2$ has more probability mass on values smaller than x than $\mathcal{P}_1$. This is the usual stochastic order in one dimension.

A.2 Poisson Random Measure

In this section we will introduce the notion of a Poisson random measure. A normalized Poisson random measure is used in Section 2.1 to build a prior for a discrete mixing probability measure, and later in Section 2.2 for building a prior distribution for functions.

Poisson Random Measure Let $(\Xi, \mathcal{B})$ be a measurable space. $\tilde{N}$ is a Poisson random measure on $\mathbb{R}_+ \times \Xi$ with intensity measure ν, if

(i) For any $B \in \mathcal{B}$ the distribution of $\tilde{N}(B)$ is Poisson$(\nu(B))$.

(ii) For any collection of pairwise disjoint sets: $B_1, \ldots, B_k \in \mathcal{B}$ the random variables $\tilde{N}(B_1), \ldots, \tilde{N}(B_k)$ are mutually independent.

The measure ν needs to fulfill $\int_\Xi \int_{(0,1)} s\nu(ds, d\xi) < \infty$ and $\int_\Xi \int_{[1,\infty)} \nu(ds, d\xi) < \infty$.

Now by $\tilde{\mu}(B) = \int_B \int_{\mathbb{R}_+} s\tilde{N}(ds, d\xi)$ one can define a random measure, which is normalized in Section 2.1 to define a random probability measure, and can be used as a prior for a discrete measure (as suggested at the end of Section 2.2). Note that $\tilde{\mu}(.)$ as defined above is sometimes also called a Lévy random measure, with ν the associated Lévy measure.

A.3 Statistical Distance Measures and Convergence Concepts

Here we will introduce some distance measures, which turn out to be useful in particular for the asymptotic study of BNP methods. In all cases we consider the "distance" between two densities f and g dominated by the Lebesgue measure on $\mathbb{R}$, other cases can be defined analogously. A useful review of different statistical distance measures is given by Gibbs and Su (2002).

Weak Distances

A sequence of probability measures with densities f_n converges weakly against a distribution with density f if $\int h(y)f_n(y)dy \to \int h(y)f(y)dy$ for all continuous bounded functions h (weak convergence also implies, for example, that the cdfs $F_n(y)$ converge against $F(y)$ for all $y \in \mathbb{R}$, if $F(y)$ is continuous). There are several equivalent metrics that metrize weak convergence, for example the Lévy or the Prohorov metric (see Dudley (2002, ch. 11)). As we are not interested in the specific form of the metric, we will denote by $d_W(f, g)$ any metric, which metrizes weak convergence. See Ghosh and Ramamoorthi (2003, p. 12-13, p. 60) for further material and references regarding weak distances and neighborhoods.

Total Variation Distance

The total variation distance (sometimes also called L_1 distance) is given by

$$d_{TV}(f,g) = \frac{1}{2}\int_{\mathbb{R}} |f(y) - g(y)|dy.$$

Hellinger Distance

The Hellinger distance is given by

$$d_H(f,g) = \left\{ \int_{\mathbb{R}} \left(\sqrt{f(y)} - \sqrt{g(y)} \right)^2 dy \right\}^{\frac{1}{2}}.$$

This can be re-written as $d_H(f,g) = \left\{ 2\left(1 - \int_{\mathbb{R}} \sqrt{f(y)g(y)}dy\right)\right\}^{\frac{1}{2}}$. Note that the term $\int_{\mathbb{R}} \sqrt{f(y)g(y)}dy$, a monotonic transformation of the Hellinger distance, is also often called *affinity*. Particularly $d_H(f,g) < \epsilon$ implies $\int_{\mathbb{R}} \sqrt{f(y)g(y)}dy > 1 - \epsilon^2/2$.

Kullback-Leibler Divergence

The Kullback-Leibler divergence is given by

$$K(f,g) = \int_{\mathbb{R}} \log\{f(y)/g(y)\}f(y)dy.$$

Note that the Kullback-Leibler distance is not symmetric and hence not a metric, *i.e.* $K(f,g) \neq K(g,f)$. It holds that $K(f,g) \geq 0$ with equality only if $f = g$.

Interrelationships

There are several interrelationships between the different distance measures (see Gibbs and Su (2002)), some important ones are

(i) $\frac{1}{2}d_H^2(f,g) \leq d_{TV}(f,g) \leq d_H(f,g) \leq \sqrt{2}$
This shows that the total variation distance and the Hellinger distance are equivalent, in the sense that they induce the same type of convergence and consistency (strong consistency).

(ii) $d_H(f,g) \leq \sqrt{K(f,g)}$
This shows that the Kullback-Leibler divergence is stronger than the other two distance measures.

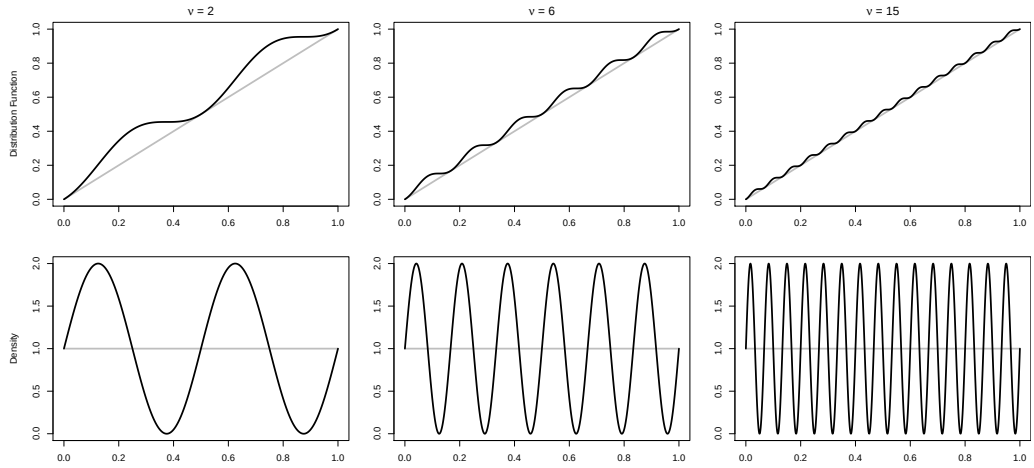

Figure A.1: $F_\nu(y)$ and density $f_\nu(y)$ for different ν. In grey: Probability distribution function and density of the uniform distribution.

(iii) $K(f,g) \leq 2d_H^2(f,g)\left(1+\log\left(\sup_{y\in\mathbb{R}}\left|\frac{f(y)}{g(y)}\right|\right)\right)$.
This result is taken from Ghosal, Ghosh and van der Vaart (2000, p. 525).

(iv) Here we show that the weak convergence and distance is less strong than total variation convergence and distance. For $\left|\int_{\mathbb{R}} h(y)f(y)dy - \int_{\mathbb{R}} h(y)g(y)dy\right|$ (with bounded and continuous h) holds

$$\left|\int_{\mathbb{R}} h(y)(f(y)-g(y))dy\right| \leq \left|\int_{\mathbb{R}} h(y)\,|f(y)-g(y)|\,dy\right| \leq 2Kd_{TV}(f,g), \quad \text{(A.2)}$$

where $K = \sup_{y\in\mathbb{R}}|h(y)|$. So whenever d_{TV} (or equivalently d_H) goes to zero also d_W needs to go to zero. On the other hand d_{TV} does not necessarily go to zero, when d_W does, see the discussion below for an example.

Discussion Weak and Strong Consistency

It is not straightforward to understand the *practical impact* of the difference between weak and strong (*i.e.* total variation and Hellinger) neighborhoods and convergence. We will illustrate this difference by a practical example.

Consider the density $f_\nu(y) = 1+\sin(2\pi\nu y)$ on $[0,1]$ with $\nu \in \mathbb{N}$. The associated probability distribution function is given by $F_\nu(y) = y + \frac{1-\cos(2\pi\nu y)}{2\nu\pi}$. Now the distribution function F_ν converges to the uniform distribution on $[0,1]$, when $\nu \to \infty$, see Figure A.1

(it is easy to see from the formula for $F_\nu(y)$ that $F_\nu(y) \to y, \forall y \in [0,1]$, when $\nu \to \infty$), but the density $f_\nu(y)$ does not converge against the density of the uniform distribution for any y (its oscillating behaviour even gets stronger with ν getting larger, see Figure A.1). This situation is hence an example, where the limiting probability distribution is in a weak neighborhood of the uniform distribution, but not in a strong neighborhood of the uniform distribution.

Hence when the main focus is to estimate concrete values of the residual density, strong consistency is what one should aim for. But in many practical situations, weak consistency seems already be enough, for example when one is only interested in posterior probabilities or other terms defined as integrals.

Appendix B

Proofs

B.1 Proof of Theorem 3.2.1

To proof Theorem 3.2.1, we will construct a TSP approximation, similar to the Bernstein polynomial approximation. Note that this TSP approximation is used for proving Theorem 3.2.1 only, although it might turn out to be useful also in other applications. The first part of the proof is along the lines of the proof for the modified Bernstein polynomials of (Perron and Mengersen 2001).

Denote by $G(.)$ an arbitrary continuous probability distribution function on $[0,1]$ and by $\mu^0(.)$ its approximation. $F(x,m,\nu)$ will be denoted as the distribution function of a two-sided power distribution function. We define the TSP approximation of $G(.)$ to be:

$$\mu^0(x) = \sum_{k=0}^{J-1} \left\{ G\left(\frac{k+1}{J}\right) - G\left(\frac{k}{J}\right) \right\} F\left(x, \frac{k}{J-1}, J+1\right).$$

We now proceed by bounding separately $\mu^0(x) - G(x)$ (part (a)) and $G(x) - \mu^0(x)$ (part (b)) from above, which gives an upper bound for $|\mu^0(x) - G(x)|$. As $G(0) = \mu^0(0) = 0, G(1) = \mu^0(1) = 1$ we just need to consider $x \in (0,1)$.

a) Considering $\mu^0(x) - G(x)$, we write

$$G(x) = \sum_{k=0}^{k^*-1} \left\{ G\left(\frac{k+1}{J}\right) - G\left(\frac{k}{J}\right) \right\} + \left\{ G(x) - G\left(\frac{k^*}{J}\right) \right\},$$

where $k^* = \lfloor Jx \rfloor$.

Then

$$\begin{aligned}
\mu^0(x) - G(x) &= \sum_{k=0}^{k^*-1} \left\{ G\left(\frac{k+1}{J}\right) - G\left(\frac{k}{J}\right) \right\} \left\{ F(x, \frac{k}{J-1}, J+1) - 1 \right\} \\
&+ \left\{ G\left(\frac{k^*+1}{J}\right) - G\left(\frac{k^*}{J}\right) \right\} F(x, \frac{k^*}{J-1}, J+1) \\
&+ \sum_{k=k^*+1}^{J-1} \left\{ G\left(\frac{k+1}{J}\right) - G\left(\frac{k}{J}\right) \right\} F(x, \frac{k}{J-1}, J+1) \\
&- G(x) + G\left(\frac{k^*}{J}\right) \\
&= -\sum_{k=0}^{k^*-1} \left\{ G\left(\frac{k+1}{J}\right) - G\left(\frac{k}{J}\right) \right\} \left\{ 1 - F(x, \frac{k}{J-1}, J+1) \right\} \\
&- \left\{ G(x) - G\left(\frac{k^*}{J}\right) \right\} \left\{ 1 - F(x, \frac{k^*}{J-1}, J+1) \right\} \\
&+ \left\{ G\left(\frac{k^*+1}{J}\right) - G(x) \right\} F(x, \frac{k^*}{J-1}, J+1) \\
&+ \sum_{k=k^*+1}^{J-1} \left\{ G\left(\frac{k+1}{J}\right) - G\left(\frac{k}{J}\right) \right\} F(x, \frac{k}{J-1}, J+1) \\
&\leq \kappa(J) \left\{ F(x, \frac{k^*}{J-1}, J+1) + \sum_{k=k^*+1}^{J-1} F(x, \frac{k}{J-1}, J+1) \right\}.
\end{aligned}$$

Here, $\kappa(J) = \sup_{0,\ldots,J-1} \{G(\frac{k+1}{J}) - G(\frac{k}{J})\}$. Bounding the two terms on the right leads to the desired result. First, trivially

$$F\left(x, \tfrac{k^*}{J-1}, J+1\right) \leq 1.$$

For the sum we first note that $x < \frac{k^*+1}{J-1}$, so that the upper branch in the definition of the TSP distribution applies (see Equation (3.5)) for all summands.

The sum can then be bounded as follows

$$\begin{aligned}
\sum_{k=k^*+1}^{J-1} F(x, \frac{k}{J-1}, J+1) &= \sum_{k=k^*+1}^{J-1} \frac{k}{J-1} \left\{ \frac{(J-1)x}{k} \right\}^{J+1} \\
&= (J-1)^J x^{J+1} \sum_{k=k^*+1}^{J-1} k^{-J} \\
&\leq (J-1)^J x^{J+1} \left\{ (k^*+1)^{-J} + \int_{k^*+1}^{J-1} y^{-J} dy \right\} \\
&= \frac{\{(J-1)x\}^J}{(k^*+1)^J} x + \frac{\{(J-1)x\}^{J-1}}{(k^*+1)^{J-1}} x^2 - x^{J+1} \\
&\leq \left(\frac{J-1}{J} \right)^J x + \left(\frac{J-1}{J} \right)^{J-1} x^2 - x^{J+1} \\
&\leq 2e^{-1}.
\end{aligned}$$

Hence we have $\mu^0(x) - G(x) \leq \kappa(J)(1 + 2e^{-1})$.

b) For $G(x) - \mu^0(x)$ different calculations but the same ideas as in part a) lead to

$$G(x) - \mu^0(x) \leq \kappa(J) \left\{ 1 - F(x, \frac{k^*}{J-1}, J+1) + \sum_{k=0}^{k^*-1} \{1 - F(x, \frac{k}{J-1}, J+1)\} \right\}.$$

The first term can again be bounded from above by one. For the sum one obtains

$$\begin{aligned}
\sum_{k=0}^{k^*-1} \{1 - F(x, \frac{k}{J-1}, J+1)\} &= (J-1)^J (1-x)^{J+1} \sum_{k=0}^{k^*-1} \left(\frac{1}{J-1-k} \right)^J \\
&= (J-1)^J (1-x)^{J+1} \sum_{k=J-k^*}^{J-1} k^{-J} \\
&\leq (J-1)^J (1-x)^{J+1} \left\{ (J-k^*)^{-J} + \int_{J-k^*}^{J-1} y^{-J} dy \right\} \\
&= \left\{ \frac{(J-1)(1-x)}{(J-k^*)} \right\}^J (1-x) + \left\{ \frac{(J-1)(1-x)}{(J-k^*)} \right\}^{J-1} (1-x)^2 \\
&- (1-x)^{J+1} \\
&\leq \left(\frac{J-1}{J} \right)^J (1-x) + \left(\frac{J-1}{J} \right)^{J-1} (1-x)^2 - (1-x)^{J+1} \\
&\leq 2e^{-1}.
\end{aligned}$$

Parts (a) and (b) together yield

$$|\mu^0(x) - G(x)| \leq \kappa(J)(1 + 2e^{-1}) \ \forall x \in [0, 1].$$

□

B.2 Proof of Lemma 4.2.1

$$\begin{aligned}
Cov(\mu(x_1), \mu(x_2)|x_1, x_2) &= Cov(\beta_0 + \beta_1 x_1 + \beta_2 F^*(x_1), \beta_0 + \beta_1 x_2 + \beta_2 F^*(x_2)) \\
&\overset{(1)}{=} E(Cov(\beta_0 + \beta_1 x_1 + \beta_2 F^*(x_1), \beta_0 + \beta_1 x_2 + \beta_2 F^*(x_2)|F^*(.))) \\
&+ \; Cov(E(\beta_0) + E(\beta_1)x_1 + E(\beta_2)F^*(x_1), E(\beta_0) + E(\beta_1)x_2 + E(\beta_2)F^*(x_2)) \\
&= E(a_1' \boldsymbol{B} a_2) + E(\beta_2)^2 Cov(F^*(x_1), F^*(x_2)),
\end{aligned}$$

with $a_i = (1, x_i, F^*(x_i))'$, $\boldsymbol{B} = Cov(\boldsymbol{\beta})$. In (1) the law of total covariance is used, and the fact that $E(\beta_0 + \beta_1 x + \beta_2 F^*(x)|F^*(.)) = E(\beta_0) + E(\beta_1)x + E(\beta_2)F^*(x)$. The formula for the covariance of $F^*(x_1)$ and $F^*(x_2)$ can be obtained from Theorem 2.1.3. □

B.3 Proof of Theorem 4.2.1

We will use the exactly the same method of proof as in Shively, Sager and Walker (2009), and will proceed in three steps (i)-(iii) (as a short reminder note that $\zeta = (\mu(.), \sigma^2)'$). The first two steps are preliminary steps needed in third step: In (i) we will show full support in Kullback-Leibler distance then in (ii) we will show that the log-likelihood ratio $\frac{1}{n} \log L_n \to 0$, where $L_n = \left(\frac{\prod_{i=1}^n \phi(y_i, \hat{\mu}_n(x_i), \hat{\sigma}_n^2)}{\prod_{i=1}^n \phi(y_i, \mu_0(x_i), \sigma_0^2)} \right)$ and $(\hat{\mu}_h(.), \hat{\sigma}_n^2)$ is the maximum likelihood estimator of ζ_0. In (iii) finally we will show that $\Pi_n^*(H_\epsilon) \to 0$, where $H_\epsilon = \{\zeta | H_Q(\zeta, \zeta_0) > \epsilon\}$ contains the ζ with H_Q-distance at least ϵ from ζ_0.

(i) Full support in Kullback-Leibler divergence

The average Kullback Leibler divergence K_Q can explicitly be calculated:

$$K_Q(\zeta, \zeta_0) = \log(\sigma_0/\sigma) - 0.5 \left(\frac{\sigma_0^2}{\sigma^2} - 1 \right) + \frac{1}{2\sigma^2} \int_{\mathbb{R}} (\mu_0(x) - \mu(x))^2 Q(dx).$$

For full support in Kullback Leibler divergence the prior Π needs to assign positive prior probability to $B_\epsilon = \{\zeta | K_Q(\zeta, \zeta_0) < \epsilon\}$ for any $\zeta_0 = (\mu_0(.), \sigma_0^2)'$. It is straightforward to see that this is fulfilled for our prior, because any convex function can be represented in from (4.2), and the assumptions stated in (A1)-(A4) are assumed to hold.

(ii) Convergence of the log-likelihood ratio

From Groeneboom, Jongbloed and Wellner (2001) (see also Hanson and Pledger (1976)) we have that the convex least squares estimator $\hat{\mu}_n$ exists, but no explicit analytic form is available for the estimator (it needs to be calculated iteratively via quadratic programming); the maximum likelihood estimator for σ_0^2 is $\hat{\sigma}_n^2 = \frac{1}{n}\sum_{i=1}^n (y_i - \hat{\mu}_n(x_1))^2$. Now $\hat{\mu}_n$ fulfills (Groeneboom, Jongbloed and Wellner (2001), Hanson and Pledger (1976)):

$$\sup_{x\in(0,1)} |\hat{\mu}_n(x) - \mu_0(x)| \to 0 \text{ a.s.} \tag{B.1}$$

and the log-likelihood ratio is given by

$$\begin{aligned}
\frac{1}{n}\log L_n &= \frac{1}{n}\left(\log(\sigma_0/\hat{\sigma}_n) - \frac{1}{2\hat{\sigma}_n^2}\sum_{i=1}^n (y_i - \hat{\mu}_n(x_i))^2 + \frac{1}{2\sigma_0^2}\sum_{i=1}^n (y_i - \mu_0(x_i))^2\right) \\
&= \frac{1}{n}\left(\log(\sigma_0/\hat{\sigma}_n) - \frac{n}{2} + \frac{n}{2\sigma_0^2}\frac{1}{n}\sum_{i=1}^n (y_i - \mu_0(x_i))^2\right).
\end{aligned}$$

As $\frac{1}{n}\sum_{i=1}^n (y_i - \mu_0(x_i))^2$ almost surely converges to σ_0^2 and in addition

$$\sum_{i=1}^n (y_i - \hat{\mu}_n(x_i))^2 \leq \sum_{i=1}^n (y_i - \mu_0(x_i))^2 + \sum_{i=1}^n (\hat{\mu}_n(x_i) - \mu_0(x_i))^2.$$

From (B.1), $\hat{\sigma}_n^2$ converges to σ_0^2, and putting all results together the log likelihood ratio converges to 0.

(iii) Convergence of posterior

Similar as in Section 2.3 we separately consider the numerator and denominator of

$$\frac{\int_{H_\epsilon}\prod_{i=1}^n \phi(y_i,\mu(x_i),\sigma^2)\Pi(d\zeta)}{\int\prod_{i=1}^n \phi(y_i,\mu(x_i),\sigma^2)\Pi(d\zeta)} = \frac{\int_{H_\epsilon}\prod_{i=1}^n R(y_i,x_i)\Pi(d\zeta)}{\int\prod_{i=1}^n R(y_i,x_i)\Pi(d\zeta)} = \frac{J_n(H_\epsilon)}{I_n},$$

with $R(y_i,x_i) = \frac{\phi(y_i,\mu(x_i),\sigma^2)}{\phi(y_i,\mu_0(x_i),\sigma_0^2)}$. The almost sure lower bound for I_n can be established exactly as in Lemma 2.3.1 using (i). For the numerator we have

$$\begin{aligned}
J_n(H_\epsilon) &= \int_{H_\epsilon}\frac{\prod_{i=1}^n \phi(y_i,\mu(x_i),\sigma^2)}{\prod_{i=1}^n \phi(y_i,\mu_0(x_i),\sigma_0^2)}\Pi(d\zeta) \\
&\leq L_n^{1/2}\int_{H_\epsilon}\left(\frac{\prod_{i=1}^n \phi(y_i,\mu(x_i),\sigma^2)}{\prod_{i=1}^n \phi(y_i,\mu_0(x_i),\sigma_0^2)}\right)^{1/2}\Pi(d\zeta).
\end{aligned}$$

From (ii) we have that $L_n^{1/2} \leq \exp(nd)$ almost surely for any $d > 0$ and sufficiently large n. Hence we consider the remaining part $K_n(H_\epsilon) = \int_{H_\epsilon} \left(\frac{\prod_{i=1}^n \phi(y_i, \mu(x_i), \sigma^2)}{\prod_{i=1}^n \phi(y_i, \mu_0(x_i), \sigma_0^2)} \right)^{1/2} \Pi(d\zeta)$. Note that we proceed quite similar as in Theorem 2.3.1, but we do not need the notion of strong separation. Again, however Markov's inequality will be used.

$$\begin{aligned} E(K_n(H_\epsilon)) &= \prod_{i=1}^n \int_{\mathbb{R}} \int_{[0,1]} \int_{H_\epsilon} \left(\frac{\prod_{i=1}^n \phi(y_i, \mu(x_i), \sigma^2)}{\prod_{i=1}^n \phi(y_i, \mu_0(x_i), \sigma_0^2)} \right)^{1/2} \Pi(d\zeta) \phi(y_i, \mu_0(x_i), \sigma_0^2) Q(dx_i) dy_i \\ &= \int_{H_\epsilon} \prod_{i=1}^n \int_{\mathbb{R}} \int_{[0,1]} \sqrt{\phi(y_i, \mu(x_i), \sigma^2) \phi(y_i, \mu_0(x_i), \sigma_0^2)} Q(dx_i) dy_i \Pi(d\zeta). \end{aligned}$$

Here E denotes the expectation with respect to $(x_1, y_1)', (x_2, y_2)', \ldots$. Now from the direct relationship of the affinity and Hellinger distance (see Appendix A.3) and the properties of the set H_ϵ we can conclude

$$E(K_n(H_\epsilon)) \leq \int_{H_\epsilon} (1 - \epsilon^2/2)^n \Pi(d\zeta) \leq (1 - \epsilon^2/2)^n.$$

From this one can apply Markov's inequality to yield $P(K_n(H_\epsilon) > \exp(-n\eta)) < \exp(n\eta) \exp(-n(-\log(1 - \tilde{\epsilon})))$, with $\eta > 0$ and $\tilde{\epsilon} = \epsilon^2/2$. Hence by the Borel-Cantelli Lemma $K_n(H_\epsilon) < \exp(-n\eta)$ almost surely for $\eta < -\log(1 - \tilde{\epsilon})$ (here note that $\log(1 - \tilde{\epsilon})$ is negative). Hence $J_n(H_\epsilon) \leq \exp(-n(\eta - d))$, and one can choose $d < \eta$. Combined with (i) this yields the desired convergence result. □

B.4 Proof of Theorem 5.2.1

Part (i)

First define $\delta_h := \mu_h(\boldsymbol{x}') - \mu_h(\boldsymbol{x})$.

The conditional distribution functions at $\boldsymbol{x}$ and $\boldsymbol{x}'$ are given by $F_{\boldsymbol{x}}(y) = \sum \pi_h \Phi\left(\frac{y - \mu_h(\boldsymbol{x})}{\sigma_h}\right)$ and $F_{\boldsymbol{x}'}(y) = \sum \pi_h \Phi\left(\frac{y - \mu_h(\boldsymbol{x}) - \delta_h}{\sigma_h}\right)$, Stochastic ordering now follows because of the monotonicity of $\Phi(.)$ and from the fact that $\delta_h \geq 0 \, \forall h$.

Part (ii)

To show: For $\epsilon > 0$ there exist π_h, σ_h^2 and $\mu_h(\boldsymbol{x})$ such that

$$\sup_{\boldsymbol{x} \in [0,1]^k} \left\{ \sup_{y \in \mathbb{R}} \left| \sum_{h=1}^N \pi_h \Phi(y, \mu_h(\boldsymbol{x}), \sigma_h^2) - \tilde{F}_{\boldsymbol{x}}(y) \right| \right\} \leq \epsilon + \frac{1}{N}.$$

We will explicitly construct an approximation, which has the desired error bound. For this purpose we set $\pi_h = 1/N$ and $\mu_h(\boldsymbol{x}) = q_{\frac{2h-1}{2N}}(\boldsymbol{x})$, where $q_\alpha(\boldsymbol{x})$ is the α quantile of the conditional distribution at $\boldsymbol{x}$. Note that $q_\alpha(\boldsymbol{x})$, as a function of $\boldsymbol{x}$, is a multivariate monotonic function for any choice of $\alpha \in (0,1)$; this directly follows from the definition of the SO-$\mathcal{X}$ stochastic order. Finally we set $\sigma_h^2 = c$, with $c > 0$.

Then we consider (with $\mu_h(\boldsymbol{x})$ as defined above):

$$\begin{aligned}\left|\frac{1}{N}\sum_{h=1}^{N}\Phi(y,\mu_h(\boldsymbol{x}),c) - \tilde{F}_{\boldsymbol{x}}(y)\right| &\leq \left|\frac{1}{N}\sum_{h=1}^{N}\Phi(y,\mu_h(\boldsymbol{x}),c) - \frac{1}{N}\sum_{h=1}^{N}1_{(y\geq\mu_h(\boldsymbol{x}))}\right| + \\ &+ \left|\frac{1}{N}\sum_{h=1}^{N}1_{(y\geq\mu_h(\boldsymbol{x}))} - \tilde{F}_{\boldsymbol{x}}(y)\right| = (*) + (**)\end{aligned}$$

Now we consider the two summands separately.

First consider $(**)$ (with $\mu_h(\boldsymbol{x})$ as defined above):

Because we use the conditional quantiles, it follows from (Fang and Wang 1994, Theorem 4.1) that

$$\sup_{y\in\mathbb{R}}\left|\frac{1}{N}\sum_{h=1}^{N}1_{(y\geq\mu_h(\boldsymbol{x}))} - \tilde{F}_{\boldsymbol{x}}(y)\right| = \frac{1}{2N}$$

For $(*)$ first choose $c = c_{\epsilon,\delta}$ in a way such that for each $h \in \{1,\ldots,N\}$:

$$\left|\Phi(y,\mu_h(\boldsymbol{x}),c_{\epsilon,\delta}) - 1_{(y\geq\mu_h(\boldsymbol{x}))}\right| < \epsilon, \forall y \in (-\infty,\mu_h(\boldsymbol{x})-\delta]\cup[\mu_h(\boldsymbol{x})+\delta,\infty),$$

where δ is a small positive number. In addition it is straightforward to see that

$$\left|\Phi(y,\mu_h(\boldsymbol{x}),c_{\epsilon,\delta}) - 1_{(y\geq\mu_h(\boldsymbol{x}))}\right| \leq 0.5$$

for $y \in (\mu_h(\boldsymbol{x})-\delta,\mu_h(\boldsymbol{x})+\delta)$. Note that this inequality is sharp with the maximum achieved at $y = \mu_h(\boldsymbol{x})$. Hence for a particular $y \in \mathbb{R}$ and a sufficiently small δ we have

$$\begin{aligned}\left|\frac{1}{N}\sum_{h=1}^{N}\left(\Phi(y,\mu_h(\boldsymbol{x}),c_{\epsilon,\delta}) - 1_{(y\geq\mu_h(\boldsymbol{x}))}\right)\right| &\leq \frac{1}{N}\sum_{h=1}^{N}\left|\left(\Phi(y,\mu_h(\boldsymbol{x}),c_{\epsilon,\delta}) - 1_{(y\geq\mu_h(\boldsymbol{x}))}\right)\right| \\ &\leq \frac{(N-1)\epsilon+0.5}{N} \leq \epsilon + \frac{1}{2N},\end{aligned}$$

because for sufficiently small δ, y can at most be in one of the intervals $(\mu_h(\boldsymbol{x})-\delta,\mu_h(\boldsymbol{x})+\delta)$, $h \in \{1,\ldots,N\}$. Hence we obtain the desired result that $(*)+(**) \leq$

$\epsilon + \frac{1}{N}$. The upper bound holds for all y and $\boldsymbol{x}$. Hence it is also valid, even when taking the supremum over $y \in \mathbb{R}$ and subsequently the supremum over $\boldsymbol{x} \in [0,1]^k$. Note that properties of the normal distribution have not explicitly been used in the proof. In fact the same proof goes through for any other unimodal distribution that includes parameters for the mode and steepness at the mode, so that it can converge towards a step function with increasing steepness at the mode (e.g. t-distributions, logistic distribution, Laplace distribution, etc).

B.5 Proof of Lemma 5.2.2

This is merely a matter of rewriting and redefining the function: First write $\sum c_j g_j(\boldsymbol{a}_j'\boldsymbol{x})$ as $\sum c_j g_j'(\boldsymbol{\alpha}_j'\boldsymbol{x})$ with $g_j'(\boldsymbol{\alpha}_j'\boldsymbol{x}) = g_j(\boldsymbol{a}_j'\mathbf{1}\boldsymbol{\alpha}_j'\boldsymbol{x})$, where $\boldsymbol{\alpha}_j = \boldsymbol{a}_j'/\boldsymbol{a}_j'\mathbf{1}$, then one can write $\sum c_j g_j'(\boldsymbol{\alpha}_j'\boldsymbol{x})$ as $\beta_0 + \beta_1 \sum w_j g_j^*(\boldsymbol{\alpha}_j'\boldsymbol{x})$, where $\beta_0 = \sum c_j g_j'(\boldsymbol{\alpha}_j'\mathbf{0})$, $\beta_1 = \sum c_j(g_j'(\boldsymbol{\alpha}_j'\mathbf{1}) - g_j'(\boldsymbol{\alpha}_j'\mathbf{0}))$, $g_j^*(\boldsymbol{\alpha}_j'\boldsymbol{x}) = \frac{g_j'(\boldsymbol{\alpha}_j'\boldsymbol{x}) - g_j'(\boldsymbol{\alpha}_j'\mathbf{0})}{g_j'(\boldsymbol{\alpha}_j'\mathbf{1}) - g_j'(\boldsymbol{\alpha}_j'\mathbf{0})}$ and $w_j = \frac{c_j(g_j'(\boldsymbol{\alpha}_j'\mathbf{1}) - g_j'(\boldsymbol{\alpha}_j'\mathbf{0}))}{\sum c_j(g_j'(\boldsymbol{\alpha}_j'\mathbf{1}) - g_j'(\boldsymbol{\alpha}_j'\mathbf{0}))}$. It is then straightforward to see that we have a form $\beta_0 + \beta_1 \sum w_j g_j^*(\boldsymbol{\alpha}_j'\boldsymbol{x})$ and the $g_j^*(x)$ are univariate distribution functions. Hence the sup norm convergence occurs when all individual $g_j^*(x)$ can be approximated in sup norm.

Appendix C

Computer Algorithms

The practicing Bayesian is well advised to become friends with as many numerical analysts as possible.
James O. Berger

C.1 Implementation for Section 3

As J is not fixed and the parameter $\boldsymbol{\vartheta}$ has a varying dimensional distribution we use the reversible jump methodology of (Green 1995) for the implementation of the approach. A move set is defined and a move selected at random at each iteration. Let the number of iterations for the burn-in period be b, the number of total iterations be T and the thinning rate be ϕ.

In the following the three moves **UPDATE**, **ADD** and **REMOVE** will be described in more detail. For this purpose let $l(.)$ denote the marginal density function for $\boldsymbol{\vartheta}$, $p(.)$ the joint prior distribution for $\boldsymbol{\vartheta}$ and $q(.)$ the proposal distribution. Suppose the chain is currently in state $\boldsymbol{\vartheta}^t = (J^t, w_1^t, \ldots, w_J^t, \boldsymbol{\xi}_1^t, \ldots, \boldsymbol{\xi}_{J^t}^t)'$. Let $p_U(J), p_A(J)$ and $p_R(J)$ denote the probabilities for choosing the three different move types **UPDATE**, **ADD** and **REMOVE**, given that there are currently J summands in the model. As J needs to be a positive integer one needs to select $p_R(1) = 0$. In the calculations done for the simulation study and the example we used equal probabilities for the different possible move types. Figure C.1 gives an overview of the algorithm. We now define the moves to explore the posterior distribution of $\boldsymbol{\vartheta}$ in more detail. Our C++ implementation of the algorithm is on the log-scale for computational reasons, the description here is not.

RJ-MCMC ALGORITHM

while($t < T$**)**{
  Draw u from uniform distribution
  if($u < p_U(J^t)$**)** obtain $\boldsymbol{\vartheta}^{t+1}$ with UPDATE step
  if($u < p_U(J^t) + p_A(J^t)$**)** obtain $\boldsymbol{\vartheta}^{t+1}$ with ADD step
  else obtain $\boldsymbol{\vartheta}^{t+1}$ with REMOVE step
  if($t \; mod \; \phi = 0$ **and** $t > b$**)**{
    Save $\boldsymbol{\vartheta}^{t+1}$
    Save a draw of $p(\boldsymbol{\beta}, \sigma | \boldsymbol{y}, \boldsymbol{\vartheta}^{t+1})$
  }
}

Figure C.1: Algorithm to obtain an MCMC sample from the posterior of $\boldsymbol{\beta}, \sigma^2, \boldsymbol{\vartheta}$.

UPDATE:

Fixed dimensional update of the parameters $w_1, \ldots, w_{J^t}$ and $\boldsymbol{\xi}_1, \ldots, \boldsymbol{\xi}_{J^t}$ with Metropolis-Hastings moves.

ADD:

1. Sample $(w^*, \boldsymbol{\xi}^*)$ from a proposal distribution $q(w, \boldsymbol{\xi})$.
   Sample position r uniformly from $1, ..., J^t + 1$.

2. Set the proposal
   $\boldsymbol{\vartheta}^* = (J^t + 1, w_1^t(1 - w^*), \ldots, w_{r-1}^t(1 - w^*), w^*, w_r^t(1 - w^*), \ldots, w_{J^t}^t(1 - w^*),$
   $\boldsymbol{\xi}_1^t, \ldots, \boldsymbol{\xi}_{r-1}^t, \boldsymbol{\xi}^*, \boldsymbol{\xi}_r^t, \ldots \boldsymbol{\xi}_{J^t}^t)'$

3. Calculate $l(\boldsymbol{\vartheta}^*)$ and $p(\boldsymbol{\vartheta}^*)$ and $q(w^*, \boldsymbol{\xi}^*)$ and the Jacobian $\mathcal{J} = (1 - w^*)^{J^t - 1}$.

4. Calculate the Metropolis-Hastings-ratio

$$MH(\boldsymbol{\vartheta}^*, \boldsymbol{\vartheta}^t) = \frac{l(\boldsymbol{\vartheta}^*)}{l(\boldsymbol{\vartheta}^t)} \frac{p(\boldsymbol{\vartheta}^*)}{p(\boldsymbol{\vartheta}^t)} \frac{p_R(J^t + 1)}{p_A(J^t) q(w^*, \boldsymbol{\xi}^*)} \mathcal{J}$$

5. Sample random variate u from uniform distribution on $[0, 1]$
   If $u < MH(\boldsymbol{\vartheta}^*, \boldsymbol{\vartheta}^t)$ set $\boldsymbol{\vartheta}^{t+1} = \boldsymbol{\vartheta}^*$ else set $\boldsymbol{\vartheta}^{t+1} = \boldsymbol{\vartheta}^t$.

REMOVE:

1. Sample position r uniformly from $1, ..., J^t$.

2. Set the proposal
   $\boldsymbol{\vartheta}^* = (J^t - 1, w_1^t/(1-w_r^t), \ldots, w_{r-1}^t/(1-w_r^t), w_{r+1}^t/(1-w_r^t), \ldots, w_J^t/(1-w_r^t),$
   $\boldsymbol{\xi}_1^t, \ldots, \boldsymbol{\xi}_{r-1}^t, \boldsymbol{\xi}_{r+1}^t, \ldots \boldsymbol{\xi}_{J^t}^t)'$

3. Calculate $l(\boldsymbol{\vartheta}^*)$ and $p(\boldsymbol{\vartheta}^*)$ and $q(w_r^t, \boldsymbol{\xi}_r^t)$ and the Jacobian $\mathcal{J} = (1-w_r^t)^{2-J^t}$.

4. Calculate the Metropolis-Hastings-ratio
$$MH(\boldsymbol{\vartheta}^*, \boldsymbol{\vartheta}^t) = \frac{l(\boldsymbol{\vartheta}^*)}{l(\boldsymbol{\vartheta}^t)} \frac{p(\boldsymbol{\vartheta}^*)}{p(\boldsymbol{\vartheta}^t)} \frac{p_A(J^t-1)q(w_r^t, \boldsymbol{\xi}_r^t)}{p_R(J^t)} \mathcal{J}$$

5. Sample random variate u from uniform distribution on $[0,1]$
   If $u < MH(\boldsymbol{\vartheta}^*, \boldsymbol{\vartheta}^t)$ set $\boldsymbol{\vartheta}^{t+1} = \boldsymbol{\vartheta}^*$ else set $\boldsymbol{\vartheta}^{t+1} = \boldsymbol{\vartheta}^t$.

C.2 Implementation for Section 4

In this Section we describe how to obtain an approximate sample from the posterior distribution of $\mu(.)$ and σ^2. We assume that for β_i, $i = 1, 2, 3$ (truncated) normal distributions are used with mean m_{β_i} and variance $\tau_{\beta_i}^{-1}$, which allows that these parameters can be updated in a Gibbs step. Note that a normal distribution with infinite variance can be obtained (*i.e.* an improper constant prior), when $\tau_{\beta_i} = 0$. For σ^{-2} we use a gamma distribution with parameters $a_{\sigma^{-2}}$ and $b_{\sigma^{-2}}$.

The MCMC algorithm cycles through sampling the conditional distributions of $F^*(.)$, $\boldsymbol{\beta}$ and σ in the following way:

1. Sample $F^*(.)$ using reversible jump MCMC (see below for details on this step)

2. Sample β_0, β_1 and β_2 from the full conditional distributions
   - $\beta_0 \sim N\left(\frac{m_{\beta_0}\tau_{\beta_0} + \sum_{i=1}^n y_i^* \sigma^{-2}}{\tau_{\beta_0} + n\sigma^{-2}}, (\tau_{\beta_0} + n\sigma^{-2})^{-1}\right)$,
     where $y_i^* = y_i - \beta_1 x_i - F^*(x_i)$

- $\beta_1 \sim N\left(\frac{m_{\beta_1}\tau_{\beta_1}+\sum_{i=1}^n x_i y_i^* \sigma^{-2}}{\tau_{\beta_1}+\sum_{i=1}^n x_i^2\sigma^{-2}}, (\tau_{\beta_1}+\sum_{i=1}^n x_i^2\sigma^{-2})^{-1}\right)$, where $y_i^* = y_i - \beta_0 - F^*(x_i)$
- $\beta_2 \sim N\left(\frac{m_{\beta_2}\tau_{\beta_2}+\sum_{i=1}^n F^*(x_i) y_i^* \sigma^{-2}}{\tau_{\beta_2}+\sum_{i=1}^n F^*(x_i)^2\sigma^{-2}}, (\tau_{\beta_2}+\sum_{i=1}^n F^*(x_i)^2\sigma^{-2})^{-1}\right)$, where $y_i^* = y_i - \beta_0 - \beta_1 x_i$

When a parameter is truncated, sampling is done from the corresponding truncated normal distribution.

3. Sample σ from the full conditional distribution for σ^{-2}, *i.e.*

$$\sigma^{-2} \sim \text{gamma}(a_{\sigma^{-2}} + n/2, b_{\sigma^{-2}} + 0.5\sum_{i=1}^n (y_i - \mu(x_i))^2).$$

Details of reversible jump MCMC step:

To update F^* we use an extension of the algorithm described in Section C.1. As J is not fixed and the parameter $\mathbf{\Psi} = \{J, w_1, \ldots, w_J, \xi_1, \ldots, \xi_J\}$ has a varying dimensional distribution we again use the reversible jump methodology (Green 1995) for implementation. Suppose the chain is currently in state $\mathbf{\Psi}^t = (J^t, w_1^t, \ldots, w_J^t, \xi_1^t, \ldots, \xi_{J^t}^t)'$. One MCMC step then consists of two substeps:

(1) First a fixed dimensional update of the parameters $w_1, \ldots, w_{J^t}$ and $\xi_1, \ldots, \xi_{J^t}$ using Metropolis-Hastings.

(2) Then as in the previous section either an **ADD** or a **REMOVE** step is performed as defined in detail in Section C.1.

Note that here $l(.)$ is here the likelihood function for $\mathbf{\Psi}$ with the current realizations of $\boldsymbol{\beta}$ and σ^2 held fixed (in Section C.1 these additional parameters had been integrated out) and $p(.)$ the joint prior distribution for $\mathbf{\Psi}$ and $\boldsymbol{\beta}$, σ^2 with $\boldsymbol{\beta}$ and σ^2 held fixed.

C.3 Implementation for Section 5

The MCMC algorithm iterates through the following steps:

1) Simulate γ from $N(\boldsymbol{m}_\gamma^*, \boldsymbol{\Sigma}_\gamma^*)$, where

$$\boldsymbol{\Sigma}_\gamma^* = (\boldsymbol{Z}'\boldsymbol{\Sigma}^{-1}\boldsymbol{Z} + \boldsymbol{\Sigma}_\gamma^{-1})^{-1} \text{ and } \boldsymbol{m}_\gamma^* = \boldsymbol{\Sigma}_\gamma^*(\boldsymbol{Z}'\boldsymbol{\Sigma}^{-1}(\boldsymbol{y} - \boldsymbol{\mu}) + \boldsymbol{\Sigma}_\gamma^{-1}\boldsymbol{m}_\gamma),$$

where $\boldsymbol{y} = (y_1, \ldots, y_n)'$, $\boldsymbol{\mu} = (\mu_{K_1}(\boldsymbol{x}_1), \ldots, \mu_{K_n}(\boldsymbol{x}_n)) \in \mathbb{R}^n$, the rows of $\boldsymbol{Z} \in \mathbb{R}^{n \times p}$ are given by $\boldsymbol{z}_i$ and $\boldsymbol{\Sigma} = \text{diag}(\sigma_{K_1}^2, \ldots, \sigma_{K_n}^2) \in \mathbb{R}^{n \times n}$.

2) The latent variables K_i, $\forall i = 1, \ldots, n$ are updated with a multinomial distribution, where the probabilities of the N categories are proportional to $\pi_h \phi(y_i, \mu_h(\boldsymbol{x}_i) + \boldsymbol{z}_i'\gamma, \sigma_h), h = 1, \ldots, N$.

3) The class specific mean functions $\mu_h(\boldsymbol{x})$ and bandwidth parameters σ_h^{-2} are updated for $h \in 1, \ldots, N$. For this determine m_h, the number of observations currently allocated to cluster h.

- If $m_h = 0$: Update from prior distributions

$$\sigma_h^{-2} \sim \text{Exp}(\omega)$$

$$\beta_{0h} \sim N(m_0, 1/\nu_0)$$

$$\beta_{1h} \sim \pi_0 \delta_0 + (1 - \pi_0)\text{Exp}(\lambda)$$

$$J_h - 1 \sim \text{Poi}(\rho)$$

$$(m_{hj}, \nu_{hj}, \boldsymbol{\alpha}_{hj})' \sim U(0,1) \times U(1,20) \times D(\boldsymbol{1})$$

- If $m_h > 0$

$$\sigma_h^{-2} \sim \text{gamma}\left(1 + m_h/2, \omega + 0.5 \sum_{i:K_i=h} (y_i - \mu_h(\boldsymbol{x}_i) - \boldsymbol{z}_i'\gamma)^2\right)$$

$$\beta_{0h} \sim N((r\sigma_h^{-2} + \nu_0 m_0)/(m_h\sigma_h^{-2} + \nu_0), (m_h\sigma_h^{-2} + \nu_0)^{-1}),$$

where $r = \sum_{i:K_i=h}(y_i - \beta_{1h}\mu_h^0 - \boldsymbol{\gamma}'\boldsymbol{z}_i)$,

$$\beta_{1h} \sim \pi_h^* \delta_0 + (1 - \pi_h^*) N_+(m^*, s^{*2}),$$

where $\pi_h^* = \left(1 + \frac{(1-\pi_0)\lambda(1-\Phi(0,m^*,s^{*2}))}{\pi_0 \phi(0,m^*,s^{*2})}\right)^{-1}$,
$m^* = \frac{\sum_{i:K_i=h} \mu_h^0(\boldsymbol{x}_i)(y_i - \beta_{0h} - \boldsymbol{\gamma}'\boldsymbol{z}_i) - \lambda\sigma_h^2}{\sum_{i:k_i=h} \mu_h^0(\boldsymbol{x}_i)^2}$ and $s^{*2} = (\sigma_h^{-2} \sum_{i:K_i=h} \mu_h^0(\boldsymbol{x}_i)^2)^{-1}$.

Here $N_+(m^*, s^{*2})$ denotes the distribution obtained by truncating a normal distribution with mean m^* and variance s^{*2} to $[0, \infty)$.

The functions $\mu^0(.)$ are updated using a RJ-MCMC algorithm. The algorithm performs the following two steps:

- UPDATE: Update $m_{jh}, \nu_{jh}, \boldsymbol{\alpha}_{jh}$ using Metropolis-Hasting updates
- Choose either an ADD or a REMOVE step with equal probability. If ADD is chosen try to add a basis function to $\mu_h^0(.)$, if REMOVE is chosen try to remove a basis function from $\mu_h^0(.)$.

The procedure follows the algorithm described in Section C.1.

4) The stick-breaking variables are updated as $v_N = 1$ and $v_h \sim Beta(1 + m_h, M + \sum_{i=h+1}^{N} m_i)$ for $h = 1, \ldots, N-1$, where m_h is the number of observations allocated to cluster number h for $h = 1, \ldots, N$ (see Ishwaran and James (2001) for details).

5) Update the parameters in the prior distributions M, ω, λ, m_0, ν_0, ρ, π_0 from their corresponding full conditionals.

$$M \sim \text{gamma}(a_M + N - 1, b_M - \textstyle\sum_{h=1}^{N-1} \log(1 - v_h))$$

$$\omega \sim \text{gamma}(a_\omega + N, b_\omega + \textstyle\sum_{h=1}^{N} \sigma_h^{-2})$$

$$\lambda \sim \text{gamma}(a_\lambda + \textstyle\sum_{h=1}^{N} 1_{\beta_{1h}>0}, b_\lambda + \textstyle\sum_{h=1}^{N} \beta_{1h})$$

$$m_0 \sim N\left(\frac{w_0/\tau_0 + \nu_0 \sum_{h=1}^{N} \beta_{0h}}{\tau_0^{-1} + N\nu_0}, \frac{1}{\tau_0^{-1} + N\nu_0}\right)$$

$$\nu_0 \sim \text{gamma}(a_{\nu_0} + N/2, b_{\nu_0} + 0.5 \textstyle\sum_{h=1}^{N} (\beta_{0h} - m_0)^2)$$

$$\rho \sim \text{gamma}(a_\rho + \textstyle\sum J_h - N, b_\rho + N)$$

$$\pi_0 \sim \text{Beta}(a_\pi + \textstyle\sum_{h=1}^{N} 1_{\beta_{1h}=0}, b_\pi + \textstyle\sum_{h=1}^{N} 1_{\beta_{1h}>0}).$$

Bibliography

ADLER, R. J. (1981) *The Geometry of Random Fields*, Wiley, New York.

AÏT-SAHALIA, Y. AND DUARTE, J. (2003) Nonparametric option pricing under shape restrictions, *Journal of Econometrics* **116**, 9–47.

ANTONIAK, C. (1974) Mixtures of Dirichlet processes with applications to Bayesian nonparametric problems, *Annals of Statistics* **2**, 1152–1174.

BACCHETTI, P. (1989) Additive isotonic models, *Journal of the American Statistical Association* **84**, 289–294.

BARRON, A. R. (1993) Universal approximation bounds for superpositions of a sigmoidal function, *IEEE Transactions on Information Theory* **39**, 930–944.

BIESHEUVEL, E. AND HOTHORN, L. A. (2002) Many-to-one comparisons in stratified designs, *Biometrical Journal* **44**, 101–116.

BIRKE, M. AND DETTE, H. (2007) Estimating a convex function in nonparametric regression, *Scandinavian Journal of Statistics* **34**, 384–404.

BIRKE, M. AND PILZ, K. (2009) Nonparametric option pricing with no-arbitrage constraints, *Journal of Financial Econometrics* **7**, 53–76.

BISHOP, C. M. (2006) *Pattern Recognition and Machine Learning*, Springer, New York.

BLACKWELL, D. AND MACQUEEN, J. B. (1973) Ferguson distributions via Pólya urn schemes, *Annals of Statistics* **1**, 353–355.

BLIGHT, B. J. N. AND OTT, L. (1975) A Bayesian approach to model inadequacy for polynomial regression, *Biometrika* **62**, 79–88.

BORNKAMP, B. (2006) Comparison of model-based and model-free approaches for the analysis of dose-response studies, Diplomarbeit, Fakultät Statistik, Technische Universität Dortmund, www.statistik.tu-dortmund.de/~bornkamp/diplom.pdf.

BORNKAMP, B., BRETZ, F., DMITRIENKO, A., ENAS, G., GAYDOS, B., HSU, C.-H., KÖNIG, F., KRAMS, M., LIU, Q., NEUENSCHWANDER, B., PARKE, T., PINHEIRO, J. C., ROY, A., SAX, R. AND SHEN, F. (2007) Innovative approaches for designing and analyzing adaptive dose-ranging trials, *Journal of Biopharmaceutical Statistics* **17**, 965–995.

BORNKAMP, B., FRITSCH, A., KUSS, O. AND ICKSTADT, K. (2009) Penalty specialists among goalkeepers: A nonparametric Bayesian analysis of 44 years of German Bundesliga, *in* B. Schipp and W. Krämer (eds.), *Statistical Inference, Econometric Analysis and Matrix Algebra: Festschrift in Honour of Götz Trenkler*, Physica Verlag, pp. 63–76.

BORNKAMP, B. AND ICKSTADT, K. (2009a) Bayesian nonparametric estimation of continuous monotone functions with applications to dose-response analysis, *Biometrics* **65**, 198–205.

BORNKAMP, B. AND ICKSTADT, K. (2009b) A note on B-splines for semiparametric elicitation, *The American Statistician* **63**, 373–377.

BORNKAMP, B., ICKSTADT, K. AND DUNSON, D. B. (2009) Stochastically ordered multiple regression, *tentatively accepted in Biostatistics* .

BORNKAMP, B., PINHEIRO, J. C. AND BRETZ, F. (2009) MCPMod: An R package for the design and analysis of dose-finding studies, *Journal of Statistical Software* **29**(7), 1–23.

BRETZ, F., HSU, J. C., PINHEIRO, J. C. AND LIU, Y. (2008) Dose finding - a challenge in statistics, *Biometrical Journal* **50**, 480–504.

BRETZ, F., PINHEIRO, J. C. AND BRANSON, M. (2004) On a hybrid method in dose-finding studies, *Methods of Information in Medicine* **43**, 457–460.

BRETZ, F., PINHEIRO, J. C. AND BRANSON, M. (2005) Combining multiple comparisons and modeling techniques in dose-response studies, *Biometrics* **61**, 738–748.

BRUNNER, L. (1992) Bayesian nonparametric methods for data from a unimodal density, *Statistics and Probability Letters* **14**, 195–199.

BRUNNER, L. AND LO, A. (1989) Bayes methods for a symmetric unimodal density and its mode, *Annals of Statistics* **17**, 1550–1566.

CHANG, I.-S., CHIEN, L.-C., HSIUNG, C. A., WEN, C.-C. AND WU, Y.-J. (2007) Shape restricted regression with random Bernstein polynomials, *in* R. Liu, W. Strawderman and C.-H. Zhang (eds.), *Complex Datasets and Inverse Problems: Tomography, Networks and Beyond*, Lecture Notes - Monograph Series, Vol. 54, Institute of Mathematical Statistics, pp. 187–202.

CHANG, I.-S., HSIUNG, C. A., WU, Y.-J. AND YANG, C.-C. (2005) Bayesian survival analysis using Bernstein polynomials, *Scandinavian Journal of Statistics* **32**, 447–466.

CHENEY, W. AND LIGHT, L. (1999) *A Course in Approximation Theory*, Brooks and Cole, Boston.

CHERNOZHUKOV, V., FERNÁNDEZ-VAL, I. AND GALICHON, A. (2007) Improving estimates of monotone functions by rearrangement, Technical report, arXiv:0806.4730v2.

CHOI, T. AND RAMAMOORTHI, R. (2008) Remarks on consistency of posterior distributions, *in* B. Clarke and S. Ghosal (eds.), *Pushing the Limits of Contemporary Statistics: Contributions in Honor of Jayanta K. Ghosh*, Institute of Mathematical Statistics Collections, Vol. 3, pp. 170–186.

CLYDE, M. A. AND WOLPERT, R. L. (2007) Nonparametric function estimation using overcomplete dictionaries, *in* J. M. Bernardo, M. J. Bayarri, J. O. Berger, A. P.

Dawid, D. Heckerman, A. F. M. Smith and M. West (eds.), *Bayesian Statistics 8*, Oxford University Press, pp. 91–114.

CURRIN, C., MITCHELL, T. J., MORRIS, M. AND YLVISAKER, D. (1991) Bayesian prediction of deterministic functions with applications to the design and analysis of computer experiments, *Journal of the American Statistical Association* **86**, 953–963.

DENISON, D. G. T., HOLMES, C. C., MALLICK, B. K. AND SMITH, A. F. M. (2002) *Bayesian Methods for Nonlinear Classification and Regression*, Wiley, Chichester.

DETTE, H., NEUMEYER, N. AND PILZ, K. F. (2006) A simple nonparametric estimator of a strictly monotone regression function, *Bernoulli* **12**, 469–490.

DETTE, H. AND SCHEDER, R. (2006) Strictly monotone and smooth nonparametric regression for two or more variables, *The Canadian Journal of Statistics* **34**, 535–561.

DEY, D. AND RAO, C. R. (2005) *Handbook of Statistics, Volume 25: Bayesian Thinking, Modeling and Computation*, Elsevier B.V., Amsterdam.

DEY, D., SINHA, D. AND MÜLLER, P. (1998) *Practical Nonparametric and Semiparametric Bayesian Statistics*, Springer, Berlin.

DIACONIS, P. AND FREEDMAN, D. (1986) On the consistency of Bayes estimates, *Annals of Statistics* **14**, 1–26.

DIACONIS, P. AND YLVISAKER, D. (1985) Quantifying prior opinion, *in* J. M. Bernardo, M. H. DeGroot, D. V. Lindley and A. F. M. Smith (eds.), *Bayesian Statistics 2*, Elsevier Science Publishers B.V., pp. 133–156.

DIERCKX, P. (1980) An algorithm for cubic spline fitting with convexity constraints, *Computing* **24**, 349–371.

DIERCKX, P. (1993) *Curve and Surface Fitting with Splines*, Clarendon Press.

DOKSUM, K. A. (1974) Tailfree and neutral random probabilities and their posterior distributions, *Annals of Probability* **2**, 183–201.

DOLE, D. (1999) Cosmo: A constrained scatterplot smoother for estimating convex, monotonic transformations, *Journal of Business and Economic Statistics* **17**, 444–455.

VAN DORP, J. AND KOTZ, S. (2002) The standard two-sided power distribution and its properties: With applications in financial engineering, *The American Statistician* **56**, 90–99.

DUDLEY, R. M. (2002) *Real Analysis and Probability*, Cambridge University Press, New York.

DUNSON, D. B. (2010) Nonparametric Bayes applications to biostatistics, *in* N. L. Hjort, C. Holmes, P. Müller and S. G. Walker (eds.), *Bayesian Nonparametrics*, Cambridge University Press, Cambridge, to appear.

DUNSON, D. B. AND PEDDADA, S. (2008) Bayesian nonparametric inference on stochastic ordering, *Biometrika* **95**, 859–874.

DUNSON, D. B., PILLAI, N. AND PARK, J. (2007) Bayesian density regression, *Journal of the Royal Statistical Society B* **69**, 163–183.

DYKSTRA, R. L. AND LAUD, P. (1981) A Bayesian nonparametric approach to reliability, *Annals of Statistics* **9**, 356–367.

DYKSTRA, R. L. AND ROBERTSON, T. (1982) An algorithm for isotonic regression for two or more independent variables, *Annals of Statistics* **10**, 708–716.

EILERS, P. H. C. AND MARX, B. D. (1996) Flexible smoothing with *B*-splines and penalties, *Statistical Science* **11**, 89–102.

ESCOBAR, M. D. AND WEST, M. (1995) Bayesian density estimation using mixtures, *Journal of the American Statistical Association* **90**, 577–588.

FANG, K.-T. AND WANG, Y. (1994) *Number-theoretic Methods in Statistics*, Chapmann and Hall, London.

FERGUSON, T. S. (1973) A Bayesian analysis of some nonparametric problems, *Annals of Statistics* **1**, 209–230.

FERGUSON, T. S. AND PHADIA, E. G. (1979) Bayesian nonparametric estimation based on censored data, *Annals of Statistics* **7**, 163–186.

FRITSCH, A. AND ICKSTADT, K. (2009) Improved criteria for clustering based on the posterior similarity matrix, *Bayesian Analysis* **4**, 367–392.

FRÜHWIRTH-SCHNATTER, S. (2006) *Finite Mixture and Markov Switching Models*, Springer, Berlin.

GELFAND, A. E. AND KOTTAS, A. (2000) Nonparametric Bayesian modeling for stochastic order, *Annals of the Institute of Statistical Mathematics* **53**, 865–876.

GELFAND, A. E. AND KUO, L. (1991) Nonparametric Bayesian bioassay including ordered polytomous response, *Biometrika* **78**, 657–666.

GELFAND, A. E. AND SMITH, A. F. M. (1990) Sampling-based approaches for calculating marginal densities, *Journal of the American Statistical Association* **85**, 398–409.

GHOSAL, S. (2010) Dirichlet process, related priors and posterior asymptotics, *in* N. L. Hjort, C. Holmes, P. Müller and S. G. Walker (eds.), *Bayesian Nonparametrics*, Cambridge University Press, Cambridge, to appear.

GHOSAL, S., GHOSH, J. AND VAN DER VAART, A. W. (2000) Convergence rates of posterior distributions, *Annals of Statistics* **28**, 500–531.

GHOSAL, S., GHOSH, J. K. AND RAMAMOORTHI, R. V. (1999) Consistency issues in Bayesian nonparametrics, *in Asymptotics, Nonparametrics and Time Series*, Dekker, New York, pp. 639–667.

GHOSH, J. K. AND RAMAMOORTHI, R. V. (2003) *Bayesian Nonparametrics*, Springer, New York.

GIBBS, A. AND SU, F. E. (2002) On choosing and bounding probability metrics, *International Statistical Review* **70**, 419–435.

GOODMAN, T. N. T. (1995) Bernstein-Schoenberg operators, *in* M. Dæhlen, T. Lyche and L. L. Schumaker (eds.), *Mathematical Methods for Curves and Surfaces*, Vanderbilt University Press, pp. 161–175.

GRAMACY, R. B. AND LEE, H. K. H. (2008) Bayesian treed Gaussian process models with an application to computer modeling, *Journal of the American Statistical Association* **103**, 1119–1130.

GREEN, P. J. (1995) Reversible jump Markov chain Monte Carlo computation and Bayesian model determination, *Biometrika* **82**, 711–732.

GROENEBOOM, P., JONGBLOED, G. AND WELLNER, J. A. (2001) Estimation of a convex function: Characterizations and asymptotic theory, *Annals of Statistics* **29**, 1653–1698.

HALL, P., RACINE, J. S. AND LI, Q. (2004) Cross-validation and the estimation of conditional probability densities, *Journal of the American Statistical Association* **99**, 1015–1026.

HANSEN, M. B. AND LAURITZEN, S. L. (2002) Nonparametric Bayes inference for concave distribution functions, *Statistica Neerlandica* **56**, 110–127.

HANSON, D. AND PLEDGER, G. (1976) Consistency in concave regression, *Annals of Statistics* **4**, 1038–1050.

HAYFIELD, T. AND RACINE, J. S. (2008) Nonparametric econometrics: The np package, *Journal of Statistical Software* **27**(5), 1–32.

HILDRETH, P. (1954) Point estimates of ordinates of concave regressions, *Journal of the American Statistical Association* **49**, 598–619.

HJORT, N. L. (1990) Nonparametric Bayes estimators based on beta processes in models for life history data, *Annals of Statistics* **18**, 1259–1294.

HJORT, N. L., HOLMES, C., MÜLLER, P. AND WALKER, S. G. (eds.) (2010) *Bayesian Nonparametrics*, Cambridge University Press, Cambridge, to appear.

HJORT, N. L. AND WALKER, S. G. (2009) Quantile pyramids for Bayesian nonparametrics, *Annals of Statistics* **37**, 105–131.

HOFF, P. (2003a) Bayesian methods for partial stochastic ordering, *Biometrika* **90**, 303–317.

HOFF, P. (2003b) Nonparametric estimation of convex models via mixtures, *Annals of Statistics* **31**, 174–200.

IBRAHIM, J. G., CHEN, M.-H. AND SINHA, D. (2001) *Bayesian Survival Analysis*, Springer, New York.

ISHWARAN, H. AND JAMES, L. F. (2001) Gibbs sampling methods for stick-breaking priors, *Journal of the American Statistical Association* **96**, 161–173.

ISHWARAN, H. AND JAMES, L. F. (2003) Generalized weighted Chinese restaurant processes for species sampling mixture models, *Statistica Sinica* **13**, 1211–1235.

ISHWARAN, H. AND ZAREPOUR, M. (2002) Exact and approximate sum representations for the Dirichlet process, *Canadian Journal of Statistics* **30**, 269–283.

JAMES, L. F., LIJOI, A. AND PRÜNSTER, I. (2009) Posterior analysis for normalized random measures with independent increments, *Scandinavian Journal of Statistics* **36**, 76–97.

JARA, A. (2009) *DPpackage: Bayesian Nonparametric and Semiparametric Analysis*, R package version 1.0-7, with contributions from Timothy Hanson, Fernando A. Quintana, Peter Mueller and Gary L. Rosner.

KARABATSOS, G. AND WALKER, S. G. (2007) Bayesian nonparametric inference of stochastically ordered distributions, with Polya trees and Bernstein polynomials, *Statistics and Probability Letters* **77**, 907–913.

KOENKER, R. (2008) *quantreg: Quantile Regression*, R package version 4.26.

KRACKER, H., BORNKAMP, B., KUHNT, S., GATHER, U. AND ICKSTADT, K. (2010) Uncertainty in Gaussian Process Interpolation, *in Recent Developments in Applied Probability and Statistics*, Springer, to appear.

LANG, S. AND BREZGER, A. (2004) Bayesian P-splines, *Journal of Computational and Graphical Statistics* **13**, 183–212.

LAVINE, M. AND MOCKUS, A. (1995) A nonparametric Bayes method for isotonic regression, *Journal of Statistical Planning and Inference* **46**, 235–248.

LEE, H. K. H. (2004) *Bayesian Nonparametrics Via Neural Networks*, SIAM, Philadelphia.

LEE, S., LIM, J., KIM, S.-J. AND JOO, Y. (2009) Estimating monotone convex functions via sequential shape modification, *Journal of Statistical Computation and Simulation* **79**, 989–1000.

LINDLEY, D. V. (2000) The philosophy of statistics, *The Statistician* **49**, 293–337.

LO, A. (1984) On a class of Bayesian nonparametric estimates: I. density estimates, *Annals of Statistics* **12**, 351–357.

LOADER, C. (2007) *locfit: Local Regression, Likelihood and Density Estimation.*, R package version 1.5-4.

LONGNECKER, M., KLEBANOFF, M., BROCK, J. AND GUO, X. (2005) Maternal levels of polychlorinated biphenyls in relation to preterm and small-for-gestational-age birth, *Epidemiology* **16**, 641–647.

LONGNECKER, M. P., KLEBANOFF, M. A., ZHOU, H. AND BROCK, J. W. (2001) Association between maternal serum concentration of the DDT metabolite DDE and preterm and small-for-gestational-age babies at birth, *Lancet* **358**, 110–114.

MACEACHERN, S. N. AND MÜLLER, P. (1998) Estimating mixture of Dirichlet process models, *Journal of Computational and Graphical Statistics* **7**, 223–238.

MAMMEN, E. (1991) Nonparametric regression under qualitative smoothness assumptions, *Annals of Statistics* **19**, 741–759.

MEYER, M. C. (2008) Inference using shape-restricted regression splines, *Annals of Applied Statistics* **2**, 1013–1033.

MORTON-JONES, T., DIGGLE, P., PARKER, L., DICKINSON, H. O. AND BINKS, K. (2000) Additive isotonic regression models in epidemiology, *Statistics in Medicine* **19**, 849–859.

MUKARJEE, H. AND STERN, S. (1994) Feasible nonparametric estimation of multiargument monotone functions, *Journal of the American Statistical Association* **89**, 77–80.

MUKERJEE, H. (1988) Monotone nonparametric regression, *Annals of Statistics* **16**, 741–750.

MÜLLER, P., ERKANLI, A. AND WEST, M. (1996) Bayesian curve fitting using multivariate normal mixtures, *Biometrika* **83**, 67–79.

MÜLLER, P. AND QUINTANA, F. (2010) More on nonparametric Bayesian models for biostatistics, *in* N. L. Hjort, C. Holmes, P. Müller and S. G. Walker (eds.), *Bayesian Nonparametrics*, Cambridge University Press, Cambridge, to appear.

MÜLLER, P. AND QUINTANA, F. A. (2004) Nonparametric Bayesian data analysis, *Statistical Science* **19**, 95–100.

NEAL, R. (1998) Regression and classification using Gaussian process priors, *in* J. M. Bernardo, J. O. Berger, A. P. Dawid and A. F. M. Smith (eds.), *Bayesian Statistics 6*, Oxford University Press, pp. 475–501.

NEAL, R. (2000) Markov chain sampling methods for Dirichlet process mixture models, *Journal of Computational and Graphical Statistics* **9**, 249–265.

NEELON, B. AND DUNSON, D. B. (2004) Bayesian isotonic regression and trend analysis, *Biometrics* **60**, 398–406.

NG, P. T. AND MAECHLER, M. (2008) *cobs: COBS – Constrained B-splines (Sparse matrix based)*, R package version 1.1-5.

O'HAGAN, A. (1973) Bayes estimation of a convex quadratic, *Biometrika* **60**, 565–571.

O'HAGAN, A. (1978) Curve fitting and optimal design for prediction (with discussion), *Journal of the Royal Statistical Society B* **40**, 1–42.

O'HAGAN, A. AND FORSTER, J. (2004) *Kendall's Advanced Theory of Statistics, Volume 2B: Bayesian Inference*, 2nd edition, Arnold, London.

ONGARO, A. AND CATTANEO, C. (2004) Discrete random probability measures: A general framework for nonparametric Bayesian inference, *Statistics and Probability Letters* **67**, 33–45.

OWEN, A. (1998) Latin hypercube sampling for very high-dimensional simulations, *ACM Transactions on Modeling and Computer Simulation* **8**, 71–102.

PAPASPILIOPOULOS, O. AND ROBERTS, G. O. (2008) Retrospective Markov chain Monte Carlo methods for Dirichlet process hierarchical models, *Biometrika* **95**, 169–186.

PERRON, F. AND MENGERSEN, K. (2001) Bayesian nonparametric modeling using mixtures of triangular distributions, *Biometrics* **57**, 518–528.

PETRONE, S. (1999) Random Bernstein polynomials, *Scandinavian Journal of Statistics* **26**, 373–393.

PILLAI, N. S. AND WOLPERT, R. L. (2008) Posterior consistency of Bayesian nonparametric models using Lévy random field priors, Technical report, 2008-08, Department of Statistical Science, Duke University, Durham, NC, USA.

PILZ, K., TITOFF, S. AND DETTE, H. (2005) *monreg: Nonparametric monotone regression*, R package version 0.1.

PINHEIRO, J. C., BRETZ, F. AND BRANSON, M. (2006) Analysis of dose-response studies – modeling approaches, *in* N. Ting (ed.), *Dose Finding in Drug Development*, Springer, New York, pp. 146–171.

PITMAN, J. (1996) Some developments of the Blackwell-MacQueen urn scheme, *in* T. S. Ferguson, L. S. Shapley and J. B. MacQueen (eds.), *Statistics, Probability and Game Theory: Papers in Honor of David Blackwell*, Institute of Mathematical Statistics, pp. 245–268.

R DEVELOPMENT CORE TEAM (2009) *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria, ISBN 3-900051-07-0.

RAMGOPAL, P., LAUD, P. W. AND SMITH, A. F. M. (1993) Nonparametric Bayesian bioassay with prior constraints on the shape of the potency curve, *Biometrika* **80**, 489–498.

RAMSAY, J. O. (1988) Monotone regression splines in action, *Statistical Science* **3**, 425–441.

RAMSAY, J. O. (1998) Estimating smooth monotone functions, *Journal of the Royal Statistical Society B* **60**, 365–375.

RASMUSSEN, C. E. AND WILLIAMS, C. K. I. (2006) *Gaussian Processes for Machine Learning*, MIT press, Cambridge, Massachusetts.

RATKOWKSY, D. A. (1983) *Nonlinear Regression Modeling*, Dekker, New York.

SCHWARTZ, L. (1965) On Bayes procedures, *Zeitschrift für Wahrscheinlichkeitstheorie und verwandte Gebiete* **4**, 10–26.

SCHWETLICK, H. AND KUNERT, V. (1993) Spline smoothing under constraints on derivatives, *BIT Numerical Mathematics* **33**, 512–528.

SETHURAMAN, J. (1994) A constructive definition of Dirichlet priors, *Statistica Sinica* **4**, 639–650.

SHIVELY, T. S., SAGER, T. W. AND WALKER, S. G. (2009) A Bayesian approach to non-parametric monotone function estimation, *Journal of the Royal Statistical Society B* **71**, 159–175.

STURTZ, S., LIGGES, U. AND GELMAN, A. (2005) R2WinBUGS: A package for running WinBUGS from R, *Journal of Statistical Software* **12**, 1–16.

THALANGE, N., FOSTER, P., GILL, M., PRICE, M. AND CLAYTON, P. (1996) Model of normal prepubertal growth, *Archives of Disease in Childhood* **75**, 1–5.

THOMAS, N. (2006) Hypothesis testing and Bayesian estimation using a sigmoid Emax model applied to sparse dose designs, *Journal of Biopharmaceutical Statistics* **16**, 657–677.

TIERNEY, L. (1994) Markov chains for exploring posterior distributions, *Annals of Statistics* **22**, 1701–1762.

TOKDAR, S. T. AND GHOSH, J. K. (2007) Posterior consistency of logistic Gaussian process priors in density estimation, *Journal of Statistical Planning and Inference* **137**, 34–42.

TU, C., CLYDE, M. AND WOLPERT, R. L. (2008) Lévy adaptive regression kernels, Technical report, 2006-08, Department of Statistical Science, Duke University, Durham.

TURLACH, B. (2005) Shape constrained smoothing using smoothing splines, *Computational Statistics* **20**, 81–104.

TUTZ, G. AND LEITENSDORFER, F. (2007) Generalized smooth monotonic regression in additive modeling, *Journal of Computational and Graphical Statistics* **16**, 165–188.

VAN DER VAART, A. W. AND WELLNER, J. A. (1996) *Weak Convergence and Empirical Processes*, Springer, New York.

WALKER, S. G. (2003) On sufficient conditions for Bayesian consistency, *Biometrika* **90**, 482–488.

WALKER, S. G. (2004) Modern Bayesian asymptotics, *Statistical Science* **19**, 111–117.

WALKER, S. G. (2007) Sampling the Dirichlet mixture model with slices, *Communications in Statistics - Simulation and Computation* **36**, 45–54.

WALKER, S. G., DAMIEN, P., LAUD, P. W. AND SMITH, A. F. M. (1999) Bayesian nonparametric inference for random distributions and related functions (with discussion), *Journal of the Royal Statistical Society Series B* **61**, 485–527.

WALKER, S. G. AND HJORT, N. L. (2001) On Bayesian consistency, *Journal of the Royal Statistical Society, Series B* **63**, 811–821.

WALKER, S. G., LIJOI, A. AND PRÜNSTER, I. (2005) Data tracking and the understanding of Bayesian consistency, *Biometrika* **92**, 765–778.

WANG, L. AND DUNSON, D. B. (2009) Bayesian isotonic density regression, Technical report, under revision at Biometrika.

WOLPERT, R. L. AND ICKSTADT, K. (1998a) Poisson/gamma random field models for spatial statistics, *Biometrika* **85**, 251–267.

WOLPERT, R. L. AND ICKSTADT, K. (1998b) Simulation of Lévy random fields, *in Practical Nonparametric and Semiparametric Bayesian Statistics*, Springer, Berlin, pp. 227–242.

WOLPERT, R. L. AND ICKSTADT, K. (2004) Reflecting uncertainty in inverse problems: A Bayesian solution using Lévy processes, *Inverse Problems* **20**, 1759–1771.

WOOD, S. (1994) Monotonic smoothing splines fitted by cross validation, *SIAM Journal on Scientific Computing* **15**, 1126–1133.

WOOD, S. (2007) *mgcv*, R package version 1.3-29.

WOOD, S. (2009) *mgcv*, R package version 1.5-5.

WU, Y. AND GHOSAL, S. (2008) Kullback Leibler property of kernel mixture priors in Bayesian density estimation, *Electronic Journal of Statistics* **2**, 298–331.

XIONG, Y., CHEN, W., APLEY, D. AND DING, X. (2007) A non-stationary covariance based kriging method for meta-modeling in engineering design, *International Journal for Numerical Methods in Engineering* **71**, 733–756.

YATCHEW, A. AND HÄRDLE, W. (2006) Nonparametric state price density estimation using constrained least squares and the bootstrap, *Journal of Econometrics* **133**, 579–599.

References for Quotes (ordered as they appear in the thesis)

Spinoza	Ethics I, as cited in Niederreiter (1978), *Bulletin of the American Mathematical Society* **84**(6), p. 957
John v. Neumann	Dyson (2004) *Nature* **427**, p. 297
Stephen Senn	as cited in Spiegelhalter, Abrams and Myles (2004) *Bayesian Approaches to Clinical Trials and Health Care Evaluation*, p.181
Igor Stravinsky	Stravinsky (1974) *Poetics of Music in the Form of Six Lessons*
Henry Kissinger	Observer, Sayings of the Week, 12th December 1976
James O. Berger	Berger (1993) *Statistical Decision Theory and Bayesian Analysis*, p. 262