Bayesian Nonparametric Estimation with Shape Constraints

Pang, L.

DOI
10.4233/uuid:cde75cae-91c8-4f4d-9b65-51bee023cd08

Publication date
2020

Document Version
Final published version

Citation (APA)
Pang, L. (2020). *Bayesian Nonparametric Estimation with Shape Constraints*. [Dissertation (TU Delft), Delft University of Technology]. https://doi.org/10.4233/uuid:cde75cae-91c8-4f4d-9b65-51bee023cd08

Important note
To cite this publication, please use the final published version (if applicable).
Please check the document version above.

BAYESIAN NONPARAMETRIC ESTIMATION WITH SHAPE CONSTRAINTS

Lixue Pang

Lixue Pang
Bayesian nonparametric estimation with shape constraints, Delft, 2020

BAYESIAN NONPARAMETRIC ESTIMATION WITH SHAPE CONSTRAINTS

Dissertation

for the purpose of obtaining the degree of doctor
at Delft University of Technology
by the authority of the Rector Magnificus,
prof. dr. ir. T.H.J.J. van der Hagen,
chair of the Board for Doctorates
to be defended publicly on
Friday 28 August 2020 at 10:00 o'clock

by

Lixue Pang

Master of Applied Probability in Mathematics
University of Science and Technology of China

This dissertation has been approved by the promotors.

Composition of the doctoral committee:

Rector Magnificus,	chairperson
Prof. dr. ir. G. Jongbloed,	Delft University of Technology, promotor
Dr. ir. F.H. van der Meulen,	Delft University of Technology, copromotor
Independent members:	
Prof. A.J. Schmidt-Hieber,	University of Twente
Prof. dr. J.H. van Zanten,	Vrije Universiteit Amsterdam
Prof. dr. ir. A.W. Heemink,	Delft University of Technology
Dr. M.R. Schauer,	Chalmers University of Technology, Sweden
Dr. J. Söhl,	Delft University of Technology
Prof. dr. ir. M. Verlaan,	Delft University of Technology, reserve member

Contents

1. Introduction

In statistical inference, a fundamental problem consists of finding a suitable probability model for a given dataset. Historically, research has focused on parametric solutions, where the probability distribution is specified up to a finite-dimensional parameter. While this simplicity comes with mathematical and computational convenience, the risk of misspecification is considerable. That is, the class of presumed probability distributions may exclude the data-generating distribution. Nonparametric methods aim to alleviate this by allowing for infinite dimensional parameters, thereby enriching the the class of considered probability distributions.

The estimation framework can either be frequentist or Bayesian. In classical frequentist inference, model parameters are considered fixed and unknown. Hence, there is a clear distinction between the data, that are modelled using a probability distribution, and the parameter. Within Bayesian statistics both data and parameters are equipped with a probability distribution. Once specified, the approach is conceptually simple as all inference is to be based on the posterior distribution, which is the distribution of the parameters, conditional on the data. Whereas, this distribution is virtually never tractable in closed form, the past two decades have witnessed a tremendous development of computational tools that enable to sample from the posterior. Using such samples, uncertainty quantification on parameters is relatively straightforward, especially compared to the frequentist approach. Moreover, Bayesian derived point estimates enjoy favourable properties such as admissibility and shrinkage. In the nonparametric setting, maximum likelihood estimators may fail to be consistent. The same can be said about Bayesian estimators if the prior is not chosen carefully.

For nonparametric problems, constructing a prior on the space of all suitable probability density functions is a delicate and difficult matter. Popular priors in Bayesian nonparametric models include Gaussian processes, Dirichlet processes, Pólya Trees and mixtures of these. In most Bayesian nonparametric models, there is no closed form of posterior distribution which increases the practical difficulties to the computations of the posterior distribution. Computational methods need to be developed to sample from the posterior. This is an approximation method for the posterior and in order to apply it, an algorithm is needed. The class of stochastic simulation methods known as Markov Chain Monte Carlo (MCMC) algorithms constitute efficient methods for this purpose. Apart from the computa-

tional aspect, understanding the theoretical properties is also crucial for Bayesian nonparametric inference. Typically, theoretical studies in Bayesian nonparametrics focus on the asymptotic properties of the posterior distribution from a frequentist point of view. The tools which describe the posterior properties lead to the notion of posterior consistency and contraction rate. Posterior consistency means that the posterior probability distribution asymptotically concentrates on any arbitrarily small neighborhood of the true value of the parameter, under the true data generating measure. A stronger property, a (Bayesian) contraction rate, is a lower bound on the radius of balls around the true parameter, while maintaining most of the posterior mass.

In statistical modelling, often there is prior knowledge on the shape of a parameter. For example, it is natural to assume that the expected height of children is nondecreasing with age. Apart from situations in which shape constraints appear naturally, these can also be induced by the inverse nature of a statistical problem. This occurs for example in survival analysis when indirect observations or censored data lead to an inverse problem where the sampling density depends on the distribution of interest in a particular way. Then the monotonicity property of the distribution function of interest induces a shape constraint on the sampling distribution. If available, it is natural to incorporate shape constraints in the statistical inferential method.

In this thesis, we focus on Bayesian nonparametric estimation in the presence of shape constraints. We start with a literature review of Bayesian nonparametric estimation which includes some general results and methods we will use in the following chapters. In the second chapter we deal with the problem of estimating a deceasing density. We derive pointwise contraction rates. In the third and the fourth chapter, we study the distribution of the time until the occurrence of a certain event, where the event time can not be observed directly due to a censoring scheme. The third chapter addresses the mixed case interval censoring problem under the assumption that the distribution function of the event time is concave. The fourth chapter is on estimating a bivariate distribution in the current status continuous mark model.

1.1. Asymptotic properties of the posterior

In this section we restrict our attention to asymptotic properties of the posterior distribution. We begin with a discussion on posterior consistency. Throughout this section, let $\mathcal{P}$ be the parameter set which includes all probability densities p with respect to a dominating measure μ on a sample space $(\Omega, \mathcal{H})$. We assume the set $\mathcal{P}$ is equipped with a suitable topology and $\sigma-$field. Consider estimating

an unknown probability density $p_0 \in \mathcal{P}$ based on observations $X_1, \ldots, X_n \overset{iid}{\sim} p_0$. Write $X^n = (X_1, \ldots, X_n)$. Let Π be the prior distribution on $\mathcal{P}$ and $\Pi(\cdot|X^n)$ be the posterior distribution. An expression for the posterior distribution is given by the *Bayes' formula*,

$$\Pi(A \mid X^n) = \frac{\int_A \prod_{i=1}^n p(X_i)\, \mathrm{d}\Pi(p)}{\int_{\mathcal{P}} \prod_{i=1}^n p(X_i)\, \mathrm{d}\Pi(p)}$$

for any measurable set A.

Definition 1.1.1. (Posterior consistency) The posterior distribution is said to be consistent at p_0 if for every neighbourhood U of p_0, $\Pi(U \mid X^n) \to 1$, p_0-almost surely.

A well known result in posterior consistency for dominated models was derived in Schwartz (1965). This result requires that the prior puts sufficient mass near the true density p_0 and the existence of a uniformly consistent sequence for testing $p = p_0$ versus $p \in U^c$. The first condition is quantified using the Kullback-Leibler divergence. Denote the KL-divergence between p_0 and p as $KL(p_0, p) = \int p_0 \log(p_0/p) d\mu$. We state Schwartz's Theorem as below.

Theorem 1.1.2. *(Schwartz (1965)) Assume $X_1, \ldots, X_n$ are independent and identically distributed with common density p_0. If $\Pi(p : KL(p_0, p) < \varepsilon) > 0$ for all $\varepsilon > 0$ and for every neighborhood U of p_0 there exist test functions $\Phi_n : \Omega^n \to [0, 1]$ such that*

$$\begin{aligned} \mathbb{E}_{p_0}(\Phi_n(X^n)) &\to 0, \\ \sup_{p \in U^c} \mathbb{E}_p(1 - \Phi_n(X^n)) &\to 0, \end{aligned} \tag{1.1}$$

then the posterior distribution is consistent at p_0.

Having posterior consistency, a natural refinement is the quantification of the rate at which the posterior concentrates around the true distribution. To obtain posterior consistency, a neighborhood is defined as a fixed ball of radius ε around p_0. If we let the radius depend on n, so $\varepsilon = \varepsilon_n$. We consider the rate at which we can let $\varepsilon_n \downarrow 0$ while still capturing most of posterior mass.

Definition 1.1.3. (Contraction rate) A sequence ε_n such that for sufficiently large M, $E_{p_0}\Pi(p : d(p, p_0) \geq M\varepsilon_n \mid X^n) \to 0$ is called a contraction rate of the posterior with respect to the semimetric d.

With a similar idea as in Schwartz's theorem, theorem 8.9 of Ghosal & Van der Vaart (2017) gives a general result for deriving contraction rates by using a rate related version of the Kullback-Leibler condition for the prior. Let $N(\varepsilon, \mathcal{P}, d)$ be the minimal number of balls of radius ε needed to cover $\mathcal{P}$.

Theorem 1.1.4. *(Ghosal & Van der Vaart (2017)) Suppose that for two sequences $0 \leq \varepsilon_n \to 0$ and $0 \leq \bar{\varepsilon}_n \leq \varepsilon_n$ with $n\bar{\varepsilon}_n^2 \to \infty$, there exists a constant $c > 0$ and sets $\mathcal{P}_n \subset \mathcal{P}$, such that*

$$\Pi(\mathcal{P}_n^c) \leq \exp(-(c+4)n\bar{\varepsilon}_n^2), \tag{1.2}$$

$$\log N(\varepsilon_n, \mathcal{P}_n, d) \leq n\varepsilon_n^2, \tag{1.3}$$

$$\Pi\left(p\colon \int p_0 \log(p_0/p) \leq \bar{\varepsilon}_n^2, \int p_0(\log(p_0/p))^2 \leq \bar{\varepsilon}_n^2\right) \geq \exp(-cn\bar{\varepsilon}_n^2). \tag{1.4}$$

Then for sufficiently large M, $E_{p_0}\Pi(p\colon d(p,p_0) \geq M\varepsilon_n | X^n) \to 0$ as $n \to \infty$.

Condition (1.4) is similar to the KL-divergence in Schwartz's theorem but with an additional restriction on the expectation of $(\log(p_0/p))^2$. Condition (1.2) indicates that there exists a sequence of sieves $\mathcal{P}_n$ capturing most of the prior mass. Then we only need to consider the model on these smaller sets $\mathcal{P}_n$ in condition (1.3). Condition (1.3) encapsulates that the size of the model should not be too large, it guarantees the existence of a uniformly exponentially consistent test sequence. More precisely, existence of a sequence of test functions Φ_n such that

$$\begin{aligned} \mathbb{E}_{p_0}(\Phi_n) &\leq \exp(-KM^2n\varepsilon_n^2), \\ \sup_{\{p \in \mathcal{P}_n : d(p,p_0) > M\varepsilon_n\}} \mathbb{E}_p(1-\Phi_n) &\leq \exp(-KM^2n\varepsilon_n^2). \end{aligned} \tag{1.5}$$

for some constant $K > 0$. In the argument of the following proof only this weaker condition is used.

Sketch of the proof of theorem 1.1.4. Write the posterior mass on set U as

$$\Pi(U \mid X^n) = D_n^{-1} \int_U \prod_{i=1}^n \frac{p(X_i)}{p_0(X_i)} \, \mathrm{d}\Pi(p),$$

where

$$D_n = \int \prod_{i=1}^n \frac{p(X_i)}{p_0(X_i)} \, \mathrm{d}\Pi(p).$$

By lemma 8.1 in Ghosal, Ghosh & Van der Vaart (2000), condition (1.4) implies that

$$\mathbb{P}_0(D_n \leq \exp(-(c+2)n\bar{\varepsilon}_n^2)) \to 0, \quad \text{as } n \to \infty.$$

Then we can restrict attention to the event $\{D_n \geq \exp(-(c+2)n\bar{\varepsilon}_n^2)\}$. Taking M large enough such that $KM^2 > c+2$ and using (1.5), the posterior distribution on $U_n = \{p\colon d(p, p_0) > M\varepsilon_n\}$ satisfies

$$\begin{aligned}
\mathbb{E}_{p_0}\Pi(U_n \mid X^n) &= \mathbb{E}_{p_0}\Pi(U_n \mid X^n)\Phi_n + \mathbb{E}_{p_0}\Pi(U_n \mid X^n)(1-\Phi_n) \\
&\leq \mathbb{E}_{p_0}\Phi_n + e^{(c+2)n\bar{\varepsilon}_n^2}\left(\Pi(\mathcal{P}_n^c) + \mathbb{E}_{p_0}\int_{U_n\cap\mathcal{P}_n}\prod_{i=1}^n\frac{p(X_i)}{p_0(X_i)}(1-\Phi_n)\,\mathrm{d}\Pi(p)\right) \\
&= \mathbb{E}_{p_0}\Phi_n + e^{(c+2)n\bar{\varepsilon}_n^2}\left(\Pi(\mathcal{P}_n^c) + \int_{U_n\cap\mathcal{P}_n}\mathbb{E}_p(1-\Phi_n)\,\mathrm{d}\Pi(p)\right) \\
&\leq e^{-KM^2n\varepsilon_n^2} + e^{(c+2)n\bar{\varepsilon}_n^2}\left(e^{KM^2n\varepsilon_n^2} + \Pi(\mathcal{P}_n^c)\right) \to 0.
\end{aligned}$$

□

In conclusion, studying consistency and contraction rates of a posterior distribution includes three aspects:

1. computing a metric entropy which is used for controlling the size of the model or checking the existence of uniformly exponentially consistent tests;

2. finding suitable sieves that capture most of the prior mass;

3. ensuring that the prior assigns positive probabilities on neighbourhoods around the true distribution.

There is an extensive literature on Bayesian nonparametrics. Key references include Schwartz (1965), Barron, Schervish & Wasserman (1999), Ghosal, Ghosh & Ramamoorthi (1999), Walker & Hjort (2001), Walker (2004), Ghosal, Ghosh & Van der Vaart (2000), Shen & Wasserman (2001) and Walker, Lijoi & Prunster (2007). There are many specific models the theory has been applied to. For instance, Ghosal & Van der Vaart (2007b) study the rates of convergence of the posterior distribution for estimating smooth densities with Dirichlet mixtures of normal distributions as the prior. A similar prior was considered by Tokdar (2006) and Ghosal & Van der Vaart (2001). For other type of priors, Tokdar & Ghosh (2007) derived the posterior consistency of density estimation using logistic Gaussian process priors. The Bayesian approach also provides a natural way to incorporate shape constraints, like monotonicity in Salomond (2014) and Shively, Sager & Walker (2009), convexity in Hannah & Dunson (2011) and Shively, Walker & Damien (2011), and log-concavity in Mariucci, Ray & Szabó (2017), etc.

1.2. Dirichlet Process (DP)

In this section we introduce a useful and important family of prior distributions known as the Dirichlet process, first proposed by Ferguson (1973). The Dirichlet Process is a probability distribution over probability measures. This process, actually measure, has a large support (with respect to weak topology) and in some cases yields tractability of the posterior distribution. For these reasons it is often used as a prior in Bayesian nonparametrics. We begin with the definition of the Dirichlet process, which arises naturally from the finite-dimensional Dirichlet distribution. Then we address more explanations and important properties, as well as its application.

The Dirichlet process has two parameters: a distribution function G_0 referred to as the base measure and a scalar $\alpha > 0$ known as the concentration parameter. The base measure is the expected value of the process and the concentration parameter specifies how close to G_0 a realisation can be expected.

Definition 1.2.1. (Dirichlet Process) A random measure P on $(\mathbb{R}, \mathcal{B})$ has the Dirichlet process distribution $\mathrm{DP}(G_0, \alpha)$, if for every finite partition $B_1, B_2, \ldots, B_k$ of $\mathbb{R}$,

$$(P(B_1), \ldots, P(B_k)) \sim \mathrm{Dir}(k, \alpha G_0(B_1), \ldots, \alpha G_0(B_k)),$$

where $\mathrm{Dir}(k, a_1, \ldots, a_k)$ denotes Dirichlet distribution of order k with parameters $a_1, \ldots, a_k$ for which the density function is given by

$$f(x_1, \ldots, x_k) = \frac{\Gamma(\sum_{i=1}^k a_i)}{\prod_{i=1}^k \Gamma(a_i)} \prod_{i=1}^k x_i^{a_i - 1}, \qquad \min_{1 \le i \le k} x_i \ge 0 \quad \text{and} \quad \sum_{i=1}^k x_i = 1.$$

1.2.1. Constructions

Multiple ways exist for constructing a realisation from the Dirichlet process. The following theorem shows how a realisation from $\mathrm{DP}(G_0, \alpha)$ can be obtained by a *stick-breaking process*.

Theorem 1.2.2. *(Sethuraman (1994)) Let $Y_1, Y_2, \ldots$ be independent and identically distributed random variables with distribution function G_0. Let $V_1, V_2, \ldots$ be independent $Beta(1, \alpha)$ distributed random variables, independent of the Y_i's. Define $C_1 = V_1$ and for $k \ge 2, C_k = V_k \prod_{j=1}^{k-1}(1 - V_j)$. Then the discrete measure assigning mass C_k to $Y_k (k \ge 1)$, is a realisation from the $DP(G_0, \alpha)$ distribution.*

Note that the mass of V_i shifts towards 1 as $\alpha \downarrow 0$. This implies that the smaller α, the more likely it is that realisations from the $\mathrm{DP}(G_0, \alpha)$ distribution visually show only a few large jumps. On the other hand, when α is relatively large, G will

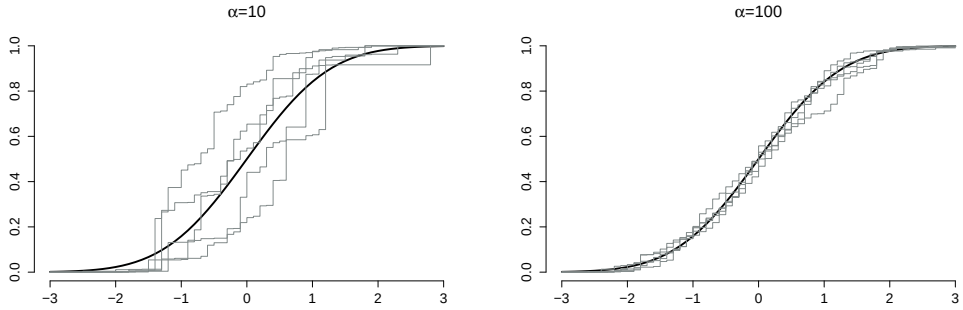

Figure 1.1.: Distribution functions generated from a Dirichlet process with different concentration parameters α. In all cases, the base measure is the standard normal distribution (thick black curve). Each case contains 5 independent realizations (grey dark curves). Note how α controls not only the variability of the realizations around G_0, but also the relative size of the jumps.

show small jumps at many of the Y_i's sampled. Therefore, G is most likely close to G_0 or, put differently, the distribution will be highly concentrated around G_0, see an example in figure 1.1.

There are other ways to generate a sample from the $DP(G_0, \alpha)$.

- *Pólya urn Process* (Blackwell and MacQueen (1973)): Starting from an empty urn, the following steps are taken in order to fill the urn with numbered balls. Set $n = 1$.

  1. Draw from the probability distribution G_0. Put a new ball in the urn labelled by the outcome of the draw.
  2. With probability $\frac{n}{\alpha+n}$, pick a ball already present in the urn and put it back with another ball having the same label. With probability $\frac{\alpha}{\alpha+n}$, go to step 1. Raise n by one.
  3. Repeat step 2 infinitely often.

  The empirical cumulative distribution function based on the numbers on balls in the urn is a draw from $\mathrm{DP}(G_0, \alpha)$. Note also that with this procedure, a large value of α will lead to many new number additions to the urn, so a realisation will likely be close to the base distribution function G_0.

- *The Chinese restaurant process (CRP)* (Aldous (1985)): This process follows the same procedure as the Pólya urn Process but explains it in a different

way. It can be interpreted as follows: Suppose that a Chinese restaurant has an infinite number of tables. The first customer picks a table. The $(n+1)$-th customer chooses one of the already occupied tables with probability $\frac{m}{\alpha+n}$, where m is the number of customers sitting at that table. Otherwise, this customer chooses a new table. After time n, there is a partition of n customers into $k \leq n$ tables. Suppose each table is labelled with a draw from the base measure G_0. With infinitely many customers having entered, the resulting probability distribution over the different tables is a draw from a Dirichlet process with parameters α and G_0.

- *Gamma Process* (see section 4.2.3 in Ghosal & Van der Vaart (2017)): Let $t \mapsto \gamma(t)$ be a Gamma process, i.e. a stochastic process for which $\gamma(t_2) - \gamma(t_1) \sim \text{Gamma}(t_2 - t_1, 1)$ for $t_2 > t_1$. Define a random distribution by $G(t) = \frac{\gamma(\alpha G_0(t))}{\gamma(\alpha)}$. The law of G is the $\text{DP}(\alpha, G_0)$ process.

1.2.2. Properties

We now list some important properties of the Dirichlet Process, more properties can be found in Ghosal & Van der Vaart (2017), chapter 4.

1. *Discreteness.* Realisations from the Dirichlet process are almost surely discrete.

2. *Support.* The support (with respect to the weak topology) of $\text{DP}(G_0, \alpha)$ is given by all probability measures G whose supports are contained in the support of G_0, that is,

$$\text{support}(\text{DP}(G_0, \alpha)) = \{G : \text{support}(G) \subset \text{support}(G_0)\}.$$

3. *Mean.* For every measurable set B, if $G \sim \text{DP}(G_0, \alpha)$, then $E(G(B)) = G_0(B)$ and for any measurable function ψ, $E(\int \psi dG) = \int \psi dG_0$.

4. *Conjugacy.* Let $G \sim \text{DP}(G_0, \alpha)$ and $\Theta_1, \Theta_2, \ldots, \Theta_n$ be independent with common probability measure G. Then the posterior distribution of G given $\Theta_1, \Theta_2, \ldots, \Theta_n$ is $\text{DP}\left(\frac{\alpha G_0 + \sum_{i=1}^n \delta_{\Theta_i}}{\alpha+n}, \alpha + n\right)$ where δ_u denotes the Dirac measure on $\{u\}$.

5. *Distinct values.* Let $G \sim \text{DP}(G_0, \alpha)$ and $\Theta_1, \Theta_2, \ldots, \Theta_n$ be independent with common probability measure G. Then the number of distinct values in the vector $(\Theta_1, \ldots, \Theta_n)$, K_n, satisfies $K_n / \log n \to \alpha$ as $n \to \infty$.

1.2.3. Dirichlet Process Mixture Models (DPM)

Based on the Dirichlet process important 'induced' models, can be obtained, like for example the hierarchical Dirichlet process, a Dirichlet process mixture, a nested Dirichlet process and dependent Dirichlet process. The Dirichlet process mixture model is a general and useful prior in Bayesian nonparametric density estimation. Let $\{\psi(\cdot,\theta) : \theta \in \Theta\}$ be a parameterized class of probability density functions. For a probability measure G, define the mixture density function

$$f_G(x) = \int_\Theta \psi(x,\theta)\,\mathrm{d}G(\theta).$$

By putting a distribution on the mixing measure G, we obtain a distribution on the mixture densities f_G. Dirichlet process mixture models (introduced by Antoniak (1974)) are obtained by endowing G with a Dirichlet process prior. The model can be written as

$$\begin{aligned} f_G(\cdot) &= \int_{\mathbb{R}} \psi(\cdot,\theta)\,\mathrm{d}G(\theta), \\ G &\sim \mathrm{DP}(G_0,\alpha). \end{aligned}$$

Sampling n realizations from f_G can be done according to the following hierarchical scheme:

$$\begin{aligned} X_i \mid \Theta_i &\overset{ind}{\sim} f_{X_i|\Theta_i} = \psi(x_i,\theta_i), \\ \Theta_1,\Theta_2,\ldots,\Theta_n \mid G &\overset{ind}{\sim} G, \\ G &\sim \mathrm{DP}(G_0,\alpha). \end{aligned}$$

Given $X^n = (X_1,\ldots,X_n)$, the posterior expectation of f_G has a simple representation. By conjugacy of the DP, we know $G \mid \Theta_1,\ldots,\Theta_n \sim \mathrm{DP}\left(\frac{\alpha G_0 + \sum_{i=1}^n \delta_{\Theta_i}}{\alpha+n}, \alpha+n\right)$ and the mean of DP is the base measure, then for any measurable function ψ,

$$E\left(\int \psi(x,\theta)\,\mathrm{d}G(\theta)\Big|\Theta_1,\ldots,\Theta_n\right) = \frac{1}{\alpha+n}\left(\alpha\int \psi(x,\theta)\,\mathrm{d}G_0(\theta) + \sum_{i=1}^n \psi(x,\Theta_i)\right).$$

Averaging out with respect to the posterior distribution of Θ, we have an expression for the posterior mean for density f_G:

$$E\left(f_G(x) \mid X^n\right) = \frac{1}{\alpha+n}\left(\alpha f_{G_0}(x) + E\left[\sum_{i=1}^n \psi(x,\Theta_j)\Big|X^n\right]\right).$$

The first part comes from the prior and the second part comes from the observations. From this formula, we see that a large value of the parameter α reflects strong belief in the prior. To approximate the posterior expectation, it is general practice to average out the second part using samples generated from the posterior distribution of $\Theta_1, \ldots, \Theta_n$.

Many algorithms have been developed for drawing from the posterior in the DPM model. For instance, if a conjugate prior is assumed, i.e., G_0 is a conjugate distribution of ψ, Gibbs sampling is straightforward and can easily be implemented (see for instance Bush & MacEachern (1996), MacEachern (1994)). In case of non-conjugate priors, West, Müller & Escobar(1994) first presented the algorithm using a Monte Carlo approximation. MacEachern and Müller (1998) proposed the 'no gaps' and 'complete' algorithms that are based on introducing auxiliary parameters. Moreover, Neal (2000) reviewed the past work and proposed new MCMC algorithms for solving this problem.

1.3. Censoring schemes

Survival analysis is concerned with the analysis of data that correspond to the time until the occurrence of some event of interest. The event can be death, the response to a treatment, or the occurrence of a symptom. However, often the exact survival time is not observed and this is referred to as censoring. The most studied censoring scheme is right censoring, which means we only know the exact event time if it occurred before a particular time (censoring time). Otherwise, the censoring time is observed, with the information that the event has not occurred yet at that time. In a medical study for example this censoring happens when subjects have not yet experienced the event of interest by the end of the study. Another, more general type of censoring is interval censoring. This arises when the event time of interest cannot be directly observed and we only know if it occurred in a specific interval, henceforth leading to observations that are intervals. This situation is encountered in many longitudinal studies where the event of interest, for example the occurrence of a symptom, can only be observed at an examination time. It is clear that right censoring can be viewed as a special case of interval censoring, where the intervals are either of type $[t, t]$ or $[t, \infty)$, but the term interval censoring is often used in situation where intervals of zero length do not occur.

The review book on semiparametric Bayesian models by Ibrahim, Chen & Sinha (2001) presents Bayesian methods for survival analysis and examines several types of parametric and semiparametric models. For nonparametric models, Susarla & Van Ryzin (1976) define a nonparametric Bayesian estimator of the survival function by minimizing the risk under the squared-error loss function when the data

are right censored. They use the class of Dirichlet processes as prior and prove that the Kaplan-Meier estimator (the frequentist maximum likelihood estimator) is a special case of this Bayesian estimator. Under the same model and prior, Ghosh, Ramamoorthi & Srikanth (1999) establish posterior consistency. They also consider the prior (for the underlying distribution) is generated through a prior for the distribution of the observations. and show that a natural extension of their approach to interval censored data is not straightforward. From the computational perspective, Doss (1994) and Doss & Huffer (2003) propose a Gibbs sampling algorithm to deal with censored data from Dirichlet mixture process models. Alternatively, Calle & Gómez (2001) propose an approach by introducing latent variables, only requiring sampling from a Dirichlet distribution.

1.4. Outline

This thesis focusses on Bayesian nonparametric function estimation under shape constraints and/or censoring. For three specific models we

1. derive theoretical properties of the Bayesian procedure (consistency, contraction rates);

2. develop computational methods for obtaining draws from the posterior distribution;

3. apply these methods to real data examples.

We now give a more specific outline of the chapters in this thesis.

In Chapter 2 we deal with nonparametric estimation of a bounded decreasing density function on $\mathbb{R}^+$ with particular emphasis on estimation of the density at zero. Estimating a monotone density constitutes a well studied topic in the literature. The maximum likelihood estimator has been derived in Grenander (1956). It has been pointed out in Woodroofe & Sun (1993) that the MLE is not consistent at 0. This is problematic in a number of inverse problems where estimation crucially depends on the estimate at zero. Some have thus tried to fix this inconsistency with various strategies, such as Kulikov & Lopuhaä (2006),Woodroofe & Sun (1993). It is well known that any decreasing density can be represented as a scale mixture of uniform densities. This suggests that within the Bayesian setting a natural prior distribution on the set of decreasing densities is obtained by endowing the mixing measure with a prior distribution. A prime example of such a prior is the Dirichlet process prior. Indeed, Salomond (2014) considered this model and derived the the posterior contraction rates for the L_1, Hellinger metric and supremum norm, but also pointwisely at any fixed point $x > 0$. For $x = 0$, only posterior consistency is

derived. We explain why the techniques in the proof of Salomond (2014) cannot be used to obtain rates at zero and present an alternative proof (using different arguments). This proof not only yields consistency but also yields a contraction rate of $(\log n/n)^{2/9}$ (up to log factors) that coincides with the case $x > 0$. We argue that with the present method of proof a better rate is not easily obtained. Additionally, we empirically investigate the rate of convergence of the Bayesian procedure for estimating the density at zero when the density of the base measure satisfies $g_0(\theta) \sim e^{-1/\theta}$ or $g_0(\theta) \sim \theta$ for $\theta \downarrow 0$. In a simulation study, we compare the performance of existing frequentist methods and the Bayesian procedure.

Chapter 3 considers estimation of a concave distribution function with mixed interval censored data. This means that for each subject under study, we observe a finite number of inspection times together with information on whether the event has occurred before each of these times. The set of inspection times, including the number of inspections, may be different for each subject. We are interested in estimating the underlying distribution function of the event time, assuming it is concave. Schick & Yu (2000) study the maximum likelihood estimator and show that it si L_1-consistent. Wellner & Zhang (2000) consider a panel count model which includes the mixed case interval censoring model as a special case. This problem has not been addressed before from a theoretical perspective within a Bayesian nonparametric setting. We prove that under weak conditions on the prior the posterior is consistent. The proof relies on Schwartz's method for proving posterior consistency. We also provide computational methods for drawing from the posterior by adapting the algorithms in Calle & Gómez (2001) and Doss & Huffer (2003) and illustrate the performance of the Bayesian method in both a simulation study and two real datasets.

In Chapter 4 we study Bayesian nonparametric estimation for the current status continuous mark model. Here, an event time X is observed under current status censoring (interval censoring case 1). Furthermore, a continuous mark variable Y is only observed in case the event occurred before the censoring time. We are interested in estimating the joint distribution function of (X, Y). This model has applications in the analysis of HIV vaccine trials (see more in Hudgens, Maathuis & Gilbert (2007)). Maathuis & Wellner (2008) show that the nonparametric maximum likelihood estimator for the joint distribution function is inconsistent. Alternative nonparametric estimators, that are consistent, have been proposed in Groeneboom, Jongbloed & Witte (2011) and Groeneboom, Jongbloed & Witte (2012). However, for both estimators no convergence rates have been derived. Within the Bayesian approach, we introduce two histogram type priors for which we derive posterior contraction rates. Using the general theory in Ghosal, Ghosh & Van der Vaart (2000), we derive that this rate is upper bounded by $n^{-1/9}$ under some regularity assumptions on the true distribution function. We propose

computational methods for obtaining draws from the posterior under both priors. For one prior this is a data-augmentation algorithm, whereas for the other one we use probabilistic programming software that is based on Hamiltonian Monte Carlo methods.

2. Bayesian estimation of a decreasing density

Suppose $X_1, \ldots, X_n$ is a random sample from a bounded and decreasing density f_0 on $[0, \infty)$. We are interested in estimating such f_0, with special interest in $f_0(0)$. This problem is encountered in various statistical applications and has gained quite some attention in the statistical literature. It is well known that the maximum likelihood estimator is inconsistent at zero. This has led several authors to propose alternative estimators which are consistent. As any decreasing density can be represented as a scale mixture of uniform densities, a Bayesian estimator is obtained by endowing the mixture distribution with the Dirichlet process prior. Assuming this prior, we derive contraction rates of the posterior density at zero by carefully revising arguments presented in Salomond (2014). Several choices of base measure are numerically evaluated and compared. In a simulation various frequentist methods and a Bayesian estimator are compared. Finally, the Bayesian procedure is applied to current durations data described in Keiding et al. (2012).

2.1. Introduction

2.1.1. Setting

Consider an independent and identically distributed sample $X_1, \ldots, X_n$ from a bounded decreasing density f_0 on $[0, \infty)$. The problem of estimating f_0 based on the sample, only using the information that it is decreasing, has attracted quite some attention in the literature. One of the reasons for this is that the estimation problem arises naturally in several applications.

To set the stage, we discuss a simple idealized example related to the waiting time paradox. Suppose buses arrive at a bus stop at random times, with independent interarrival times sampled from a distribution with distribution function H_0. At some randomly selected time, somebody arrives and has to wait for a certain amount of time until the next bus arrives. A natural question then is: 'what is the distribution of the remaining waiting time until the next bus arrives?' In order to derive this distribution, two observations are important.

The first is, that the time of arrival of the traveller is more likely contained in a long interarrival interval than a short interarrival interval. Under mild assumptions, one can show that actually the length of the whole interarrival interval (so between arrival of the previous and the next bus) containing the time the traveller arrives, can be viewed as a draw from the length biased distribution associated to distribution function H_0. This is the distribution with distribution function

$$\bar{H}_0(y) = \frac{1}{\mu_{H_0}} \int_0^y z \, dH_0(z) \text{ with } \mu_{H_0} = \int_0^\infty z \, dH_0(z). \tag{2.1}$$

It is assumed that $0 < \mu_{H_0} < \infty$.

The second observation is that the remaining waiting time for the traveller is a uniformly distributed fraction of the interarrival time. A residual waiting time X is therefore interpreted as

$$X = UY,$$

where U is uniformly distributed on $(0, 1)$ and, independently of U, Y according to distribution function $\bar{H}_0$ defined in (2.1).

These observations imply that on $[0, \infty)$, X has survival function

$$\begin{aligned} P(X > x) &= P(UY > x) = \int_{y=x}^\infty \int_{u=x/y}^1 du \, d\bar{H}_0(y) = \int_{y=x}^\infty \left(1 - \frac{x}{y}\right) d\bar{H}_0(y) \\ &= \frac{1}{\mu_{H_0}} \int_{y=x}^\infty (y - x) \, dH_0(y) = \frac{1}{\mu_{H_0}} \int_{y=x}^\infty (1 - H_0(y)) \, dy, \end{aligned}$$

using integration by parts in the last step. Differentiating with respect to x, yields the following relation between the sampling density f_0 and distribution function H_0:

$$f_0(x) = \frac{1}{\mu_{H_0}} (1 - H_0(x)), \; x \geq 0. \tag{2.2}$$

In words: the sampling density is proportional to a survival function of the interarrival distribution, which is by definition decreasing. Note that in the classical waiting time paradox, the underlying arrival process is taken to be a homogeneous Poisson process, with exponential interarrival times. In view of (2.2), this leads to the 'paradox' that the distribution of the residual waiting time equals the distribution of the interarrival time itself.

More examples where exactly this model comes into play can for instance be found in the introductory section of Kulikov & Lopuhaä (2006), in Vardi (1989), Watson (1971), Keiding et al. (2012) and references therein. In those examples, the challenge is to estimate the interarrival distribution function H_0 based on a

sample from density f_0. To do this, the 'inverse relation' of (2.2), expressing H_0 in terms of f_0 can be employed:

$$H_0(x) = 1 - \mu_{H_0} f_0(x) = 1 - \frac{f_0(x)}{f_0(0)}, \ x \geq 0. \tag{2.3}$$

Here it is used that $H_0(0) = 0$.

From (2.3) it is clear that in order to estimate H_0 at some specific point $x > 0$, estimating the decreasing sampling density f_0 at zero is of special interest. This value occurs at the right hand side for any choice of $x > 0$.

2.1.2. Literature overview

The most commonly used estimator for f_0 is the maximum likelihood estimator derived in Grenander (1956). This estimator is defined as the maximizer of the log likelihood $\ell(f) = \sum_{i=1}^{n} \log f(X_i)$ over all decreasing density functions on $(0, \infty)$. The solution $\hat{f}_n$ of this maximization problem can be graphically constructed. Starting from the empirical distribution $\mathbb{F}_n$ based on $X_1, \ldots, X_n$, the least concave majorant of $\mathbb{F}_n$ can be constructed. This is a concave distribution function. The left-continuous derivative of this piecewise linear concave function yields the maximum likelihood (or Grenander) estimator for f_0. For more details on the derivation of this estimate, see Section 2.2 in Groeneboom & Jongbloed (2014). As can immediately be inferred from the characterization of the Grenander estimator,

$$\hat{f}_n(0) := \lim_{x \downarrow 0} \hat{f}_n(x) = \max_{1 \leq i \leq n} \frac{\mathbb{F}_n(X_i)}{X_i} \geq \frac{\mathbb{F}_n(X_{(1)})}{X_{(1)}} = \frac{1}{nX_{(1)}},$$

where $X_{(i)}$ denotes the i-th order statistic of the sample. Denoting convergence in distribution by $\xrightarrow{d}$,

$$n f_0(0) X_{(1)} \xrightarrow{d} Y \quad \text{as} \quad n \to \infty$$

where Y has the standard exponential distribution. It is clear that $\hat{f}_n(0)$ does not converge in probability to $f_0(0)$. This inconsistency of $\hat{f}_n(0)$ was first studied in Woodroofe & Sun (1993). There it is also shown that

$$\frac{\hat{f}_n(0)}{f_0(0)} \xrightarrow{d} \sup_{t>0} \frac{N(t)}{t} \stackrel{d}{=} \frac{1}{U} \quad \text{as} \quad n \to \infty,$$

where N is a standard Poisson process on $[0, \infty)$ and U is a standard uniform random variable.

It is clear from (2.3) that this inconsistency is undesirable, as estimating the distribution function of interest, H_0, at any point $x > 0$, requires estimation of

$f_0(0)$. Various approaches have been taken to obtain a consistent estimator of $f_0(0)$. The idea in Kulikov & Lopuhaä (2006) is to estimate $f_0(0)$ by $\hat{f}_n$ evaluated at a small positive (but vanishing) number: $\hat{f}_n(cn^{-1/3})$ for some $c > 0$. There it is shown that the estimator is $n^{1/3}$-consistent, assuming $f_0(0) < \infty$ and $|f_0'(0)| < \infty$.

A likelihood related approach was taken in Woodroofe & Sun (1993). There a penalized log likelihood function is introduced, where the estimator is defined as maximizer of

$$\ell_\alpha(f) = \sum_{i=1}^{n} \log f(X_i) - \alpha n f(0).$$

For fixed $\alpha \geq 0$, this estimator can be computed explicitly by first transforming the data using a data dependent affine transformation and then applying the basic concave majorant algorithm to the empirical distribution function based these transformations data. It is shown (again, assuming $f_0(0) < \infty$ and $|f_0'(0)| < \infty$) that the optimal rate to choose α is $n^{-2/3}$. Then, the maximum penalized estimator $\hat{f}^P_{n,\hat{\alpha}_n}(0)$ is $n^{1/3}$-consistent.

Groeneboom & Jongbloed (2014) proposed to estimate $f_0(0)$ by the histogram estimator $b_n^{-1}\mathbb{F}_n(b_n)$, where $\{b_n\}$ is a sequence of positive numbers with $b_n \to 0$ if $n \to \infty$. The bin widths b_n can e.g. be chosen by estimating the asymptotically Mean Squared Error-optimal choice. Also this estimator is $n^{1/3}$- consistent assuming $f_0(0) < \infty$ and $| f_0'(0) | < \infty$.

2.1.3. Approach

In this paper we take a Bayesian nonparametric approach to the problem. An advantage of the Bayesian setup is the ease of constructing credible regions. To construct frequentist analogues of these, confidence regions, can be quite cumbersome, relying on either bootstrap simulations or asymptotic arguments.

To formulate a Bayesian approach for estimating a decreasing density, note that any decreasing density on $[0, \infty)$ can be represented as a scale mixture of uniform densities (see e.g. Williamson (1956)):

$$f_G(x) = \int_0^\infty \psi_x(\theta) dG(\theta), \text{ where } \psi_x(\theta) = \theta^{-1} 1_{[0,\theta]}(x), \tag{2.4}$$

where G is a distribution function concentrated on the positive half line. Therefore, by endowing the mixing measure with a prior distribution we obtain the posterior distribution of the decreasing density, and in particular of $f_0(0)$. A convenient and well studied prior for distribution functions on the real line is the Dirichlet process (DP) prior (see for instance Ferguson (1973) and Van der Vaart and Ghosal (2017)). This prior contains two parameters: the concentration parameter, usually

denoted by α, and the base probability distribution, which we will denote by G_0. The approach where a prior is obtained by putting a Dirichlet process prior on G in (2.4) was previously considered in Salomond (2014). In that paper, the asymptotic properties of the posterior in a frequentist setup are studied. More specifically, contraction rates are derived to quantify the performance of the Bayesian procedure. This is a rate for which we can shrink balls around the true parameter value, while maintaining most of the posterior mass. More formally, if L is a semimetric on the space of density functions, a contraction rate ε_n is a sequence of positive numbers $\varepsilon_n \downarrow 0$ for which the posterior mass of the balls $\{f : L(f, f_0) \leq \varepsilon_n\}$ converges in probability to 1 as $n \to \infty$, when assuming $X_1, X_2, \ldots$ are independent and identically distributed with density f_0. A general discussion on contraction rates is given in Chapter 8 of Van der Vaart and Ghosal (2017).

2.1.4. Contributions

In Theorem 4 in (Salomond (2014)) the rate $(\log n/n)^{2/9}$ is derived for pointwise loss at any $x > 0$. For $x = 0$, only posterior consistency is derived, essentially under the assumption that the base measure admits a density g_0 for which there exists $1 < a_1 \leq a_2$ such that $e^{-a_1/\theta} \lesssim g_0(\theta) \lesssim e^{-a_2/\theta}$ when θ is sufficiently small (theorem 4). These are interesting results, though one would hope to prove the rate $n^{-1/3}$ for all $x \geq 0$. Under specific conditions on the underlying density, this rate is attained by estimators to be discussed in section 2.4. We explain why the techniques in the proof of (Salomond (2014)) cannot be used to obtain rates at zero and present an alternative proof (using different arguments). This proof not only reveals consistency, but also yields a contraction rate equal to $n^{-2/9}$ (up to log factors) that coincides with the case $x > 0$. We argue that with the present method of proof a better rate is not easily obtained. Many results from Salomond (2014) are important ingredients to the proof we present. The first key contribution of this paper is to derive the claimed contraction rate, combining some of Salomond's results with new arguments.

We also address computational aspects of the problem and show how draws from the posterior can be obtained using the algorithm presented in Neal (2000). Using this algorithm we conduct four studies.

- For a fixed dataset, we compare the performance of the posterior mean under various choices of base measure for the Dirichlet process.
- We investigate empirically the rate of convergence of the Bayesian procedure for estimating the density at zero when $g_0(\theta) \sim e^{-1/\theta}$ or $g_0(\theta) \sim \theta$ for $\theta \downarrow 0$. The simulation results suggest that for both choices of base measure the rate is $n^{-1/3}$. If $g_0(\theta) \sim e^{-1/\theta}$ this implies that the derived rate $n^{-2/9}$ (up

to log factors) is indeed suboptimal, as anticipated by (Salomond (2014)). If $g_0(\theta) \sim \theta$ the rate $n^{-1/3}$ is interesting, as it contradicts the belief that "due to the similarity to the maximum likelihood estimator, the posterior distribution is in this case not consistent" (page 1386 in (Salomond (2014))).

- We compare the behaviour of various proposed frequentist methods and the Bayesian method for estimating $f_0(0)$. Here we vary the sample sizes and consider both the Exponential and half-Normal distribution as true data generating distributions.
- Pointwise credible sets can be approximated in a direct way from MCMC-output, which is much more straightforward than the construction of frequentist confidence intervals based on large-sample limiting results.

2.1.5. Outline

In section 2.2 we derive pointwise contraction rates for the density evaluated at x, for any $x \geq 0$. In section 2.3 a Markov Chain Monte Carlo method for obtaining draws from the posterior is given, based on the results of Neal (2000). This is followed by a review of some existing methods to consistently estimate f_0 at zero. Section 2.5 contains numerical illustrations. The appendix contains some technical results.

2.1.6. Frequently used notation

For two sequences $\{a_n\}$ and $\{b_n\}$ of positive real numbers, the notation $a_n \lesssim b_n$ (or $b_n \gtrsim a_n$) means that there exists a constant $C > 0$ that is independent of n and such that $a_n \leq C b_n$. We write $a_n \asymp b_n$ if both $a_n \lesssim b_n$ and $a_n \gtrsim b_n$ hold. We denote by F and F_0 the cumulative distribution functions corresponding to the probability densities f and f_0 respectively. We denote the L_1-distance between two density functions f and g by $L_1(f, g)$, i.e. $L_1(f, g) = \int |f(x) - g(x)| \, \mathrm{d}x$. The Kullback-Leibler divergence 'from f to f_0' is denoted by $KL(f, f_0) = \int f(x) \log \frac{f(x)}{f_0(x)} dx$.

2.2. Point-wise posterior contraction rates

Let $\mathcal{F}$ denote the collection of all bounded decreasing densities on $[0, \infty)$ and recall that $X_1, X_2, \ldots$ are i.i.d. with density $f \in \mathcal{F}$. Denote the distribution of $X^n = (X_1, \ldots, X_n)$ under f by $\mathbb{P}_f$ and expectation under $\mathbb{P}_f$ by $\mathbb{E}_f$. In this section we are interested in the asymptotic behaviour of the posterior distribution of $f(x)$ in a frequentist setup. This entails that we study the behaviour of the posterior

distribution on $\mathcal{F}$ while assuming a true underlying density f_0. Set $\mathbb{P}_0 = \mathbb{P}_{f_0}$ and $\mathbb{E}_0 = \mathbb{E}_{f_0}$. Denote the prior measure on $\mathcal{F}$ by Π and the posterior measure by $\Pi(\cdot \mid X^n)$.

Given a loss function L on $\mathcal{F}$, we say that the posterior is consistent with respect to L if for any $\varepsilon > 0$, $\mathbb{E}_0\Pi(L(f, f_0) > \varepsilon \mid X^n) \to 0$ when $n \to \infty$. If $\{\varepsilon_n\}$ is a sequence that tends to zero, then we say that the posterior contracts at rate ε_n (with respect to L) if $\mathbb{E}_0\Pi(L(f, f_0) > \varepsilon_n \mid X^n) \to 0$ when $n \to \infty$. The rate $\{\varepsilon_n\}$ is called a contraction rate.

Salomond (2014) derived contraction rates based on the Dirichlet process prior for the L^1-, Hellinger- and point-wise loss function.

In the following theorem we derive sufficient conditions for posterior contraction in terms of the behaviour of the density of the base measure near zero. In that, we closely follow the line of proof in Salomond (2014). Although the argument in Salomond (2014) for proving posterior contraction rate ϵ_n for $f_0(x)$ with $x > 0$ is correct, we prove the theorem below for $x \geq 0$ rather than only for $x = 0$. The reason for this is twofold: *(i)* many steps in the proof for $x > 0$ are also used in the proof for $x = 0$; *(ii)* we obtain one theorem covering pointwise contraction rates for all $x \geq 0$. For the base measure we have the following assumption.

Assumption 2.2.1. The base distribution function of prior, G_0, has a strictly positive Lebesgue density g_0 on $(0, \infty)$. There exists positive numbers $\theta_0, \underline{a}, \underline{k}, \overline{a}$ such that

$$\underline{k}e^{-\underline{a}/\theta} \leq g_0(\theta) \leq \theta^{\overline{a}} \quad \text{for all} \quad \theta \in (0, \theta_0). \tag{2.5}$$

For the data generating density we assume

Assumption 2.2.2. The data generating density $f_0 \in \mathcal{F}$ and

- there exists an $x_0 > 0$ such that $\sup_{x\in[0,x_0]}|f_0'(x)| < \infty$;
- the exist positive constants β and τ such that $f_0(x) \leq e^{-\beta x^\tau}$ for x sufficiently large.

Theorem 2 in (Salomond (2014)) asserts the existence of a positive constant C such that

$$\Pi\left(f \in \mathcal{F}\colon L_1(f, f_0) \geq C\left(\frac{\log n}{n}\right)^{1/3}(\log n)^{1/\tau} \mid X^n\right) \to 0,$$

$\mathbb{P}_0 -$ almost surely$(n \to \infty)$. This result will be used in the proof for deriving an upper bound on the pointwise contraction rate of the posterior at zero.

Define a sequence of subsets of $\mathcal{F}$ by

$$\mathcal{F}_n = \{f \in \mathcal{F} : f(0) - f(x) \leq M_n x, \text{ for all } x \in [0, \xi_n]\}, \tag{2.6}$$

where $\xi_n \asymp n^{-2/9}$ and $M_n \asymp (\log n)^\beta$.

Theorem 2.2.3. *Let $X_1, X_2, \ldots$ be independent random variables, each with density f_0 satisfying assumption 2.2.2. Let Π_n be the prior distribution on $\mathcal{F}_n$ that is obtained via (2.4), where $G \sim DP(G_0, \alpha)$ and G_0 satisfies assumption 2.2.1. Assume $\beta > 1/3$ (in the behaviour of the sequence $\{M_n\}$). For any $x \in [0, \infty)$ with $f_0'(x) < 0$ there exists a constant $C > 0$ such that,*

$$\mathbb{E}_0 \Pi \left(f \in \mathcal{F}_n : |f(x) - f_0(x)| > C n^{-2/9} (\log n)^\beta \,\middle|\, X^n \right) \to 0.$$

for $n \to \infty$.

In the proof we will use the following lemma (see appendix B and lemma 8 of (Salomond (2014))).

Lemma 2.2.4. *Let $\epsilon_n = (\log n/n)^{1/3}$ and f_0 satisfy assumption 2.2.2. Define*

$$D_n = \int \prod_{i=1}^n \frac{f(X_i)}{f_0(X_i)} \, \mathrm{d}\Pi(f). \tag{2.7}$$

There exist strictly positive constants c_1 and c_2 such that

$$\mathbb{P}_0 \left(D_n < c_1 e^{-c_2 n \epsilon_n^2} \right) = o(1) \quad \text{as} \quad n \to \infty. \tag{2.8}$$

We now give the proof of Theorem 2.2.3.

Proof of Theorem 2.2.3. The posterior measure of a measurable set $\mathcal{E} \subset \mathcal{F}$ is given by

$$\Pi(\mathcal{E}|X^n) = D_n^{-1} \int_{\mathcal{E}} \prod_{i=1}^n \frac{f(X_i)}{f_0(X_i)} \, \mathrm{d}\Pi(f),$$

where D_n is as defined in (2.7). By lemma 2.2.4 there exist positive constants c_1 and c_2 such that $\mathbb{P}_0(\mathcal{D}_n) = o(1)$, where $\mathcal{D}_n = \{D_n < c_1 e^{-c_2 n \epsilon_n^2}\}$. Let $C > 0$. Define $\eta_n = n^{-2/9} (\log n)^\beta$, $B_n(x) = \{f \in \mathcal{F}_n : |f(x) - f_0(x)| > C\eta_n\}$ and consider (test-)

functions $\Phi_n : \mathbb{R} \to [0,1]$. We bound

$$\begin{aligned}
&\mathbb{E}_0 \Pi(B_n(x)|X^n) \\
&\quad = \mathbb{E}_0 \Pi(B_n(x)|X^n) 1_{\mathcal{D}_n} + \mathbb{E}_0 \Pi(B_n(x)|X^n) 1_{\mathcal{D}_n^c} \Phi_n(x) \\
&\qquad\qquad + \mathbb{E}_0 \Pi(B_n(x)|X^n) 1_{\mathcal{D}_n^c} (1 - \Phi_n(x)) \\
&\quad \leq \mathbb{E}_0 \left[1_{\mathcal{D}_n}\right] + \mathbb{E}_0(\Phi_n(x)) + \mathbb{E}_0 \left[D_n^{-1} \int_{B_n(x)} \prod_{i=1}^n \frac{f(X_i)}{f_0(X_i)} (1 - \Phi_n(x)) \, \mathrm{d}\Pi(f) 1_{\mathcal{D}_n^c} \right] \\
&\quad \leq \mathbb{P}_0 \left(\mathcal{D}_n\right) + \mathbb{E}_0(\Phi_n(x)) + c_1^{-1} e^{c_2 n \epsilon_n^2} \mathbb{E}_0 \int_{B_n(x)} \prod_{i=1}^n \frac{f(X_i)}{f_0(X_i)} (1 - \Phi_n(x)) \, \mathrm{d}\Pi(f) \\
&\quad = o(1) + \mathbb{E}_0(\Phi_n(x)) + c_1^{-1} e^{c_2 n \epsilon_n^2} \int_{B_n(x)} \mathbb{E}_f (1 - \Phi_n(x)) \, \mathrm{d}\Pi(f). \qquad (2.9)
\end{aligned}$$

To construct the specific test functions $\Phi_n(x)$, we distinguish between $x > 0$ and $x = 0$. For case $x > 0$, it follows from the proofs of theorems 3 and 5 in Salomond (2014) that there exists a sequence test functions such that

$$\begin{aligned}
\mathbb{E}_0 \, \Phi_n(x) &= o(1) \\
\sup_{f \in B_n(x)} \mathbb{E}_f (1 - \Phi_n(x)) &\leq e^{-C' n (C \eta_n)^3} = e^{-C' C^3 n \epsilon_n^2}.
\end{aligned}$$

for some constant $C' > 0$. Substituting these bounds into (2.9) and choosing $C > (c_2/C')^{1/3}$ shows that $\mathbb{E}_0 \Pi(B_n(x)|X^n) \to 0$ as $n \to \infty$. This finishes the proof for $x > 0$.

We now consider the case $x = 0$. Define subsets

$$\begin{aligned}
B_n^+(0) &= \{f \in \mathcal{F}_n \colon f(0) - f_0(0) > C\eta_n\} \\
B_n^-(0) &= \{f \in \mathcal{F}_n \colon f(0) - f_0(0) < -C\eta_n\}.
\end{aligned}$$

As $B_n(0) = B_n^+(0) \cup B_n^-(0)$, $\Pi(B_n(0)|X^n) \leq \Pi(B_n^+(0)|X^n) + \Pi(B_n^-(0)|X^n)$. For bounding $\mathbb{E}_0 \Pi(B_n^-(0)|X^n)$, use the same test function defined in Salomond (2014). Then it follows from the inequalities in (2.9), applied with $B_n^-(0)$ instead of $B_n(x)$, that $\mathbb{E}_0 \Pi(B_n^-(0)|X^n) = o(1)$ as $n \to \infty$.

For bounding $\mathbb{E}_0 \Pi(B_n^+(0)|X^n)$, we also use the inequalities in (2.9), applied with $B_n^+(0)$ instead of $B_n(x)$. However, we also intersect with the event

$$A_n = \{f \colon L_1(f, f_0) \leq C \varepsilon_n (\log n)^{1/\tau}\}$$

to obtain

$$\mathbb{E}_0 \Pi(B_n^+(0)|X^n) \leq o(1) + \mathbb{E}_0(\Phi_n(0)) + c_1^{-1} e^{c_2 n \epsilon_n^2} \int_{B_n^+(0) \cap A_n} \mathbb{E}_f (1 - \Phi_n(0)) \, \mathrm{d}\Pi(f).$$

This holds true since theorem 2 in (Salomond (2014)) gives $\Pi(A_n^c|X^n) \to 0$, $\mathbb{P}_0$-almost surely.

Now define

$$\Phi_n^+(0) = 1\left\{ n^{-1}\sum_{i=1}^{n} 1_{[0,\xi_n]}(X_i) - \int_0^{\xi_n} f_0(t)\,\mathrm{d}t > \tilde{c}_n \right\},$$

where

$$\xi_n \asymp n^{-2/9} \qquad \text{and} \qquad \tilde{c}_n = C\xi_n\eta_n/3 \asymp n^{-4/9}(\log n)^{\beta}. \tag{2.10}$$

By Bernstein's inequality (Van der Vaart (1998), lemma 19.32),

$$\mathbb{E}_0\,\Phi_n^+(0) \le 2\exp\left(-\frac{1}{4}\frac{n\tilde{c}_n^2}{M\xi_n + \tilde{c}_n}\right) = o(1).$$

Here we bound the second moment of $1_{[0,\xi_n]}(X_i)$ under $\mathbb{P}_0$ by $f_0(0)\xi_n$ and use that $f_0(0) \le M$.

It remains to bound

$$I := e^{c_2 n\epsilon_n^2}\int_{B_{n2}^+(0)\cap A_n} \mathbb{E}_f(1-\Phi_n^+(0))\,\mathrm{d}\Pi(f).$$

Since both f and f_0 are nonincreasing we have

$$\int_0^{\xi_n}(f(t)-f_0(t))\,\mathrm{d}t \ge (f(\xi_n)-f_0(0))\xi_n.$$

Hence

$$\begin{aligned}\int_0^{\xi_n} f_0(t)\,\mathrm{d}t &\le \int_0^{\xi_n} f(t)\,\mathrm{d}t + (f_0(0)-f(\xi_n))\xi_n \\ &\le \int_0^{\xi_n} f(t)\,\mathrm{d}t + \xi_n(f_0(0)-f(0)+M_n\xi_n),\end{aligned}$$

the final inequality being a consequence of $f \in \mathcal{F}_n$. Since for $f \in B_n^+(0)$ we have $f_0(0)-f(0) \le -C\eta_n$ we get

$$\int_0^{\xi_n} f_0(t) \le \int_0^{\xi_n} f(t)\,\mathrm{d}t + \xi_n(M_n\xi_n - C\eta_n).$$

Using the derived bound we see that

$$I_2 \le e^{c_2 n\epsilon_n^2}\int_{B_n^+(0)\cap A_n} \mathbb{P}_f\left(\sqrt{n}\left(\frac{1}{n}\sum_{i=1}^{n}\mathbf{1}_{[0,\xi_n]}(X_i) - \int_0^{\xi_n} f(t)\,\mathrm{d}t\right) \le -v_n\right)\mathrm{d}\Pi(f),$$

where

$$v_n = -\sqrt{n}\left(\tilde{c}_n + \xi_n(M_n\xi_n - C\eta_n)\right). \tag{2.11}$$

Note that $M_n\xi_n \asymp \eta_n$, by choice of M_n, ξ_n. Taking C big enough such that $M_n\xi_n \leq C\eta_n/3$ we have $v_n \geq C\sqrt{n}\eta_n\xi_n/3$ is positive (recall that $\tilde{c}_n$ is defined in (2.10)). Using that f is nonincreasing and that $f \in A_n$ we get

$$\begin{aligned}\mathbb{E}_f \mathbf{1}_{[0,\xi_n]}(X_1) = \int_0^{\xi_n} f(t)\,\mathrm{d}t &\leq \|f_0 - f\|_1 + \xi_n f_0(0) \\ &\leq C\epsilon_n(\log n)^{1/\tau} + M\xi_n \leq 2M\xi_n.\end{aligned}$$

Bernstein's inequality gives

$$I \leq 2e^{c_2 n\epsilon_n^2} \exp\left(-\frac{1}{4}\frac{v_n^2}{2M\xi_n + v_n/\sqrt{n}}\right).$$

If we take $\eta_n = n^{-2/9}(\log n)^{\beta}$, then

$$\frac{v_n^2}{2M\xi_n + v_n/\sqrt{n}} \gtrsim n^{1/3}(\log n)^{2\beta}.$$

This tends to infinity faster than $n\varepsilon_n^2 = n^{1/3}(\log n)^{2/3}$ whenever $2\beta > 2/3$, i.e. when $\beta > 1/3$. □

Remark 2.2.5. The derived rate is not the optimal but cannot be easily improved upon with the present type of proof. At first sight, one may wonder whether the tests $\Phi_n^+(0)$ can be improved upon by choosing different sequences $\{\tilde{c}_n\}$ and M_n, ξ_n. Unfortunately, the choice of ξ_n and M_n cannot be much improved upon. To see this, for bounding I with Bernstein's inequality we need that v_n in (2.11) is positive. Assume $\xi_n = n^{-\beta_1}$ and $\eta_n = n^{-\beta_2}$ (up to $\log n$ factors), we must have $\beta_1 \geq \beta_2$. Hence this restriction leads to $v_n \asymp -\sqrt{n}(\tilde{c}_n + \xi_n\eta_n)$.

Define $b_n = \max(\varepsilon_n(\log n)^{1/\tau}, \xi_n)$. Then $\mathbb{E}_f \mathbf{1}_{[0,\xi_n]}(X_1) \lesssim b_n$, we can bound I by

$$2\exp\left(c_2 n^{1/3}(\log n)^{2/3} - \frac{1}{4}\frac{v_n^2}{b_n + v_n/\sqrt{n}}\right).$$

We have two cases according to sequence b_n.

1. $b_n = \xi_n$, implies $\beta_1 \leq 1/3$. We have $\frac{v_n^2}{b_n + v_n/\sqrt{n}} \asymp n\xi_n\eta_n^2 = n^{1-\beta_1-2\beta_2}$ should tend to infinity faster than $n^{1/3}$, hence $\beta_1 + 2\beta_2 \leq 2/3$. By combine all restrictions, we derive that β_2 necessarily has to satisfy $1/6 \leq \beta_2 \leq 2/9$.

2. $b_n = \varepsilon_n (\log n)^{1/\tau}$, implies $\beta_1 > 1/3$. Then $\frac{v_n^2}{b_n + v_n/\sqrt{n}} \asymp n^{4/3}(\xi_n \eta_n)^2 = n^{4/3-2\beta_1-2\beta_2} \geq n^{1/3}$ gives $\beta_1 + \beta_2 \leq 1/2$. Hence $\beta_2 < 1/6$.

Therefore, η_n can not go to zero faster than $n^{-2/9}(\log n)^{\beta}$.

Remark 2.2.6. As point-wise consistency is proved in Theorem 2.2.3, Theorem 4 in Salomond (2014) implies that the posterior median is a consistent estimator at any fixed point. Moreover, the posterior median has the same converge rate $n^{2/9}(\log n)^{\beta}$. The consistency of the posterior mean is not clear now. However, the posterior mean of f is a decreasing density function, which provides a convenient way for estimation. We use either mean or median estimator according to different purpose in the simulation study.

2.2.1. A difficulty in the proof of theorem 4 in Salomond

The construction of the tests $\{\Phi_n^+(0)\}$ in the proof of theorem 2.2.3 is new. In Salomond (2014) a different argument is used, which we now shortly review (it is given in section 3.3 of that paper). First we give a lemma for the following discussion.

Lemma 2.2.7. *Let Π be the prior distribution on $\mathcal{F}$ that is obtained via (2.4), where $G \sim DP(G_0, \alpha)$ and G_0 satisfies there exists positive numbers $\theta_0, \overline{a}, \overline{k}$ such that*

$$g_0(\theta) \leq \overline{k} e^{-\overline{a}/\theta} \quad \text{for all} \quad \theta \in (0, \theta_0).$$

Then for any x (possibly sequence) in $(0, \theta_0)$,

$$\Pi\left(\{f : f(0) - f(x) \geq A\}\right) \leq \frac{\overline{k}}{\overline{a}A} x e^{-\overline{a}/x} \quad \text{for every} \quad A > 0.$$

Proof. By the mixture representation of decreasing function f, (2.4), and Markov's inequality we have

$$\Pi\left(\{f : f(0) - f(x) \geq A\}\right) = \Pi\left(\int_0^x \theta^{-1} \, \mathrm{d}G(\theta) \geq A\right) \leq A^{-1} \int_0^x \theta^{-1} g_0(\theta) \, \mathrm{d}\theta.$$

By assumption 2.2.1 this is bounded by

$$\begin{aligned} \overline{k} A^{-1} \int_0^x \theta^{-1} e^{-\overline{a}/\theta} \, \mathrm{d}\theta &= \overline{k} A^{-1} \int_{1/x}^{\infty} u^{-1} e^{-\overline{a}u} \, \mathrm{d}u \\ &\leq \overline{k} A^{-1} x \int_{1/x}^{\infty} e^{-\overline{a}u} \, \mathrm{d}u = \overline{k} (\overline{a}A)^{-1} x e^{-\overline{a}/x}. \end{aligned}$$

□

Let $\{h_n\}$ be a sequence of positive numbers. Trivially, we have

$$f(0) - f_0(0) = f(0) - f(h_n) + f(h_n) - f_0(0).$$

Since both f and f_0 are nonincreasing, $f(h_n) \leq f(x)$ and $f_0(0) \geq f_0(x)$, for all $x \in [0, h_n]$. Hence,

$$f(0) - f_0(0) \leq f(0) - f(h_n) + f(x) - f_0(x), \qquad \text{for all} \quad x \in [0, h_n].$$

This implies

$$f(0) - f_0(0) \leq f(0) - f(h_n) + h_n^{-1} L_1(f, f_0).$$

Using this bound and define a new sequence $\tilde{\eta}_n$, we get

$$\begin{aligned} \mathbb{E}_0 \Pi \left(f(0) - f_0(0) > C\tilde{\eta}_n | X^n \right) \leq{}& \mathbb{E}_0 \Pi \left(f(0) - f(h_n) > C\tilde{\eta}_n/2 | X^n \right) \\ &+ \mathbb{E}_0 \Pi \left(L_1(f, f_0) > C\tilde{\eta}_n h_n/2 | X^n \right). \end{aligned} \tag{2.12}$$

Choose $\tilde{\eta}_n$ and h_n such that $\tilde{\eta}_n h_n = 2\varepsilon_n$. Theorem 1 in Salomond (2014) implies that the second term on the right-hand-side tends to zero. We aim to choose $\tilde{\eta}_n$ such that the first term on the right-hand-side in (2.12) also tends to zero. This term can be dealt with using lemma 2.2.4:

$$\begin{aligned} \mathbb{E}_0 \Pi \left(f(0) - f(h_n) > C\tilde{\eta}_n/2 | X^n \right) &\leq \mathbb{P}_0(\mathcal{D}_n) + c_1^{-1} e^{c_2 n \varepsilon_n^2} \Pi \left(f(0) - f(h_n) > C\tilde{\eta}_n/2 \right) \\ &= o(1) + c_1^{-1} e^{c_2 n \varepsilon_n^2} \Pi \left(f(0) - f(h_n) > C\tilde{\eta}_n/2 \right). \end{aligned}$$

Using lemma 2.2.7, the second term on the right-hand-side can be bounded by

$$\frac{2\overline{k}}{\overline{a} c_1 C} \frac{h_n}{\tilde{\eta}_n} e^{c_2 n \varepsilon_n^2 - \overline{a} h_n^{-1}} \asymp \frac{h_n^2}{\varepsilon_n} e^{c_2 n \varepsilon_n^2 - \overline{a} h_n^{-1}}$$

Since $n\varepsilon_n^2 = n^{1/3} (\log n)^{2/3}$, the right-hand-side in the preceding display tends to zero ($n \to \infty$) upon choosing $h_n^{-1} \asymp n^{1/3} (\log n)^{\beta}$ and $\beta > 2/3$. This yields

$$\tilde{\eta}_n \asymp \varepsilon_n h_n^{-1} \asymp (\log n)^{\beta + 1/3},$$

which unfortunately does not tend to zero. Hence, we do not see how the presented argument can yield point-wise consistency of the posterior at zero.

2.2.2. Attempt to fix the proof by adjusting the condition on the base measure

A natural attempt to fix the argument consists of changing the condition on the base measure. If the assumption on g_0 would be replaced with

$$\underline{k} e^{-\underline{a}/\theta^\gamma} \leq g_0(\theta) \leq \overline{k} e^{-\overline{a}/\theta^\gamma} \quad \text{for all} \quad \theta \in (0, \theta_0), \tag{2.13}$$

then lemma 2.2.7 would give the bound

$$\Pi\left(\{f : f(0) - f(x) \geq A\}\right) \leq \frac{\overline{k}}{\overline{a}A} x e^{-\overline{a}/x^{\gamma}}.$$

Now we can repeat the argument and check whether it is possible to choose γ and $\{h_n\}$ such that both $\tilde{\eta}_n \to 0$ and

$$\frac{h_n^2}{\varepsilon_n} e^{cn\varepsilon_n^2 - \overline{a}h_n^{-\gamma}} = o(1) \tag{2.14}$$

hold true simultaneously. The requirement $\tilde{\eta}_n \to 0$ leads to taking $h_n = n^{-1/3}(\log n)^{\tilde{\beta}}$, with $\tilde{\beta} > 1/3$. With this choice for h_n, equation (2.14) can only be satisfied if $\gamma > 1$. Now if we assume (2.13) with $\gamma > 1$, then we need to check whether lemma 2.2.4 is still valid. This is a delicate issue as we need to trace back in which steps of its proof the assumption on the base measure is used. In appendix B of (Salomond (2014)) it is shown that the result in lemma 2.2.4 follows upon proving that

$$\Pi(\mathcal{S}_n) \geq \exp\left(-c_1 n\varepsilon_n^2\right), \tag{2.15}$$

with $\varepsilon_n = (\log n/n)^{1/3}$ (as in the statement of the lemma). Here, the set $\mathcal{S}_n$ is defined as

$$\mathcal{S}_n = \left\{ f : KL(f_{0,n}, f_n) \leq \epsilon_n^2, \int f_{0,n}(x) \left(\log \frac{f(x)}{f_0(x)}\right)^2 \mathrm{d}x \leq \epsilon_n^2, \int_0^{\theta_n} f(x)\,\mathrm{d}x \geq 1 - \epsilon_n^2 \right\},$$

where

$$\theta_n = F_0^{-1}(1 - \epsilon_n/(2n)), \qquad f_n(\cdot) = \frac{f(\cdot)I_{[0,\theta_n]}(\cdot)}{F(\theta_n)}, \qquad f_{0,n}(\cdot) = \frac{f_0(\cdot)I_{[0,\theta_n]}(\cdot)}{F_0(\theta_n)}.$$

In lemma 8 of (Salomond (2014)) it is proved that $\Pi(\mathcal{S}_n) \gtrsim \exp\left(-C_1\varepsilon_n^{-1}\log\varepsilon_n\right)$ for some constant $C_1 > 0$, which implies the specific rate ε_n. The proof of this lemma is rather complicated, the key being to establish the existence of a set $\mathcal{N}_n \subset \mathcal{S}_n$ for which $\Pi(\mathcal{N}_n) \gtrsim \exp\left(-C_1\varepsilon_n^{-1}\log\varepsilon_n\right)$. Next, upon tracking down at which place the prior mass condition is used for that result (see appendix A.1), we find that it needs to be such that

$$\sum_{i=1}^{m_n} \log G_0(U_i) \gtrsim \varepsilon_n^{-1}\log\varepsilon_n \tag{2.16}$$

where $m_n \asymp \varepsilon_n^{-1}$ and $U_i = (i\varepsilon_n, (i+1)\varepsilon_n]$ (see in particular inequality (A.1) in the appendix). Now assume (2.13), then

$$G_0(U_i) \geq \underline{k}\int_{U_i} e^{-\underline{a}/\theta^{\gamma}}\,\mathrm{d}\theta \geq \underline{k}\varepsilon_n \exp\left(-\underline{a}(i\varepsilon_n)^{-\gamma}\right)$$

Hence

$$\sum_{i=1}^{m_n} \log G_0(U_i) \gtrsim \log \underline{k} + \varepsilon_n^{-1} \log \varepsilon_n - \underline{a} \sum_{i=1}^{n} (i\varepsilon_n)^{-\gamma}$$
$$\gtrsim \log \underline{k} + \varepsilon_n^{-1} \log \varepsilon_n - \varepsilon_n^{-\gamma},$$

if $\gamma > 1$ (which we need to assume for (2.14) to hold). From this inequality we see that (2.16) can only be satisfied if $\gamma \in (0, 1]$. We conclude that with the line of proof in (Salomond (2014)) the outlined problem in the proof of consistency near zero cannot be fixed by adjusting the prior to (2.13): one inequality requires $\gamma > 1$, while another inequality requires $\gamma \in (0, 1]$ and these inequalities need to hold true jointly.

2.3. Gibbs Sampling in the DPM model

Since a decreasing density can be represented as a scale mixture of uniform densities (see (2.4)) and the mixing measure is chosen according to a Dirichlet process, the model is a special instance of a so-called Dirichlet Process Mixture (DPM) Model. Algorithms for drawing from the posterior in such models have been studied in many papers over the past two decades, a key reference being Neal (2000). Here we shortly discuss the algorithm coined "algorithm 2" in that paper. We assume G_0 has a density g_0 with respect to Lebesgue measure.

Let $\#(x)$ denote the number of distinct values in the vector x and let x_{-i} denote the vector obtained by removing the i-th element of x. Denote by $\vee(x)$ and $\wedge(x)$ the maximum and minimum of all elements in the vector x respectively.

The starting point for the algorithm is a construction to sample from the DPM model:

$$\begin{aligned} Z := (Z_1, \ldots, Z_n) &\sim \mathrm{CRP}(\alpha) \\ \Theta_1, \ldots, \Theta_{\#(Z)} &\overset{iid}{\sim} G_0 \\ X_1, \ldots, X_n \mid \Theta_1, \ldots, \Theta_{\#(Z)}, Z_1, \ldots, Z_n &\overset{ind}{\sim} \mathrm{Unif}(0, \Theta_{Z_i}). \end{aligned} \tag{2.17}$$

Here $\mathrm{CRP}(\alpha)$ denotes the "Chinese Restaurant Process" prior, which is a distribution on the set of partitions of the integers $\{1, 2, \ldots, n\}$. This distribution is most easily described in a recursive way. Initialize by setting $Z_1 = 1$. Next, given $Z_1, \ldots, Z_i$, let $L_i = \#(Z_1, \ldots, Z_i)$ and set

$$Z_{i+1} = \begin{cases} L_i + 1 & \text{with probability } \alpha/(i+\alpha) \\ k & \text{with probability } N_k/(i+\alpha). \end{cases}$$

where k varies over $\{1, \ldots, L_i\}$ and $N_k = \sum_{j=1}^{i} \mathbf{1}\{Z_j = k\}$ is the number of current Z_j's equal to k. In principle this process can be continued indefinitely, but for our purposes it ends after n steps. One can interpret the vector Z as a partitioning of the index set $\{1, \ldots, n\}$ (and hence the data $X = (X_1, \ldots, X_n)$) into $\#(Z)$ disjoint sets (sometimes called "clusters"). For ease of notation, write $\Theta = (\Theta_1, \ldots, \Theta_{\#(Z)})$.

An algorithm for drawing from the posterior of (Z, Θ) is obtained by successive substitution sampling (also known as Gibbs sampling), where the following two steps are iterated:

1. sample $\Theta \mid (X, Z)$;
2. sample $Z \mid (X, \Theta)$.

The first step entails sampling from the posterior within each cluster. For the $k-$th component of Θ, Θ_k, this means sampling from

$$f_{\Theta_k|X,Z}(\theta_k|x,z) \propto f_{\Theta_k}(\theta_k) \prod_{j:z_j=k} f_{X_j|\Theta_k}(x_j|\theta_k) = g_0(\theta_k) \prod_{j:z_j=k} \psi(x_j|\theta_k). \tag{2.18}$$

Sampling $Z \mid (X, \Theta)$ is done by cycling over all Z_i $(1 \leq i \leq n)$ iteratively. For $i \in \{1, \ldots, n\}$ and $k \in \{1, \ldots, 1 + \vee(Z)\}$ we have

$$\begin{aligned} f_{Z_i|Z_{-i},X,\Theta}(k \mid z_{-i}, x, \theta) &\propto f_{X_i|Z_i,Z_{-i},\Theta}(x_i \mid k, z_{-i}, \theta) f_{Z_i|Z_{-i},\Theta}(k \mid z_{-i}, \theta) \\ &= f_{X_i|\Theta_{Z_i}}(x_i \mid \theta_k) f_{Z_i|Z_{-i}}(k \mid z_{-i}) \end{aligned} \tag{2.19}$$

The right-hand-side of this display equals

$$\begin{cases} \dfrac{N_{k,-i}}{n-1+\alpha} \psi(x_i \mid \theta_k) & \text{if} \quad 1 \leq k \leq \vee(Z), \\ \dfrac{\alpha}{n-1+\alpha} \displaystyle\int \psi(x_i \mid \theta) dG_0(\theta) & \text{if} \quad k = 1 + \vee(Z), \end{cases} \tag{2.20}$$

where $N_{k,-i} = \sum_{j \in \{1,\ldots,n\} \setminus \{i\}} \mathbf{1}\{Z_j = k\}$. The expression for $k = 1 + \vee(Z)$ follows since in that case sampling from $X_i \mid \Theta_k$ boils down to sampling from the marginal distribution of X_i. Summarising, we have the algorithm 1.

It may happen that over subsequent iterations of the Gibbs sampler certain clusters disappear. Then $\#(Z)$ and $\vee(Z)$ will not be the same. If this happens, the Θ_j corresponding to the disappearing cluster is understood to be removed from the vector Θ (because the cluster becomes "empty", the prior and posterior distribution of such a Θ_j are equal). The precise labels do not have a specific meaning and are only used to specify the partitioning into clusters.

In this step we need to evaluate $\int \psi(x_i|\theta) \, \mathrm{d}G_0(\theta)$. One option is to numerically evaluate this quantity for $i = 1, \ldots, n$ (it only needs to be evaluated once). Alternatively, the "no-gaps" algorithm of MacEachern and Müller (1998) or "algorithm 8" of Neal (2000) can be used and refer for further details to these papers.

Algorithm 1 Gibbs Sampling in DPM model

1: Initialise Z, Θ.
2: **for** each iteration **do**
3: **for** $i = 1, 2, \ldots, n$ **do**
4: Update Z_i according to (2.19),
5: That is, set Z_i equal to k with probabilities proportional to those given in (2.20).
6: **end for**
7: **for** $k = 1, \ldots, \#(Z)$ **do**
8: Update Θ_k by sampling from the density in (2.18).
9: **end for**
10: **end for**

2.4. Review of existing methods for estimating the decreasing density at zero

In this section we review some consistent estimators for a decreasing density f_0 at zero that have appeared in the literature. These will be compared with the Bayesian method of this paper using a simulation study in section 2.5.

2.4.1. Maximum penalized likelihood

In Woodroofe & Sun (1993), the maximum penalized likelihood estimator is defined as the maximizer of the following penalized log likelihood function:

$$\ell_\alpha(f) = \sum_{i=1}^{n} \log f(X_i) - \alpha n f(0).$$

Here $\alpha \geq 0$ is a (small) penalty parameter. This estimator has the same form as the maximum likelihood estimator (MLE), being piece-wise constant with at most n discontinuities. For fixed $\alpha \geq 0$, for ease of notation here let $x_1 < \cdots < x_n < \infty$ denote the ordered observed values and

$$w_0 = 0 \quad \text{and} \quad w_k = \alpha + \gamma x_k, \quad k = 1, \ldots, n$$

where γ is the unique solution of the equation

$$\gamma = \min_{1 \leq s \leq n} \left\{ 1 - \frac{\alpha s/n}{\alpha + \gamma x_s} \right\}.$$

Denote by $f^P(\alpha, \cdot)$ the penalized estimator with penalty parameter α. Taking $\alpha < x_n$, $f^P(\alpha, \cdot)$ is a step function with

$$f^P(\alpha, x) = f^P(\alpha, x_k), \quad \forall x_{k-1} < x \leq x_k, \quad \forall k = 1, \ldots, n.$$

At zero it is defined by right continuity and for $x \notin [0, x_n]$ as $f^P(\alpha, x) = 0$. Here

$$f^P(\alpha, x_k) = \min_{0 \leq i < k} \max_{k \leq j \leq n} \frac{(j-i)/n}{w_j - w_i}.$$

Geometrically, for $k = 1, 2, \ldots, n$, $f^P(\alpha, x_k)$ is the left derivative of the least concave majorant of the empirical distribution function of the transformed data $w_i, i = 1, \ldots, n$ evaluated at w_k. Note that an alternative expression for $f^P(\alpha, 0)$ is $(1-\gamma)/\alpha$ which can be easily calculated.

Theorem **4** in Woodroofe & Sun (1993) states that

$$n^{1/3}\{f^P(\alpha_n, 0) - f_0(0)\} \Rightarrow^d \sup_{t>0} \frac{W(t) - (c + \beta t^2)}{t}$$

where $\alpha_n = cn^{-2/3}$, $\beta = -f_0(0)f_0'(0)/2$ and $W(t)$ denotes the standard Brownian motion. In Woodroofe & Sun (1993), the theoretically optimal constant c is determined by minimizing the expected absolute value of the limiting distribution f^P, resulting in $c = 0.649 \cdot \beta^{-1/3}$.

2.4.2. Simple and 'adaptive' estimators

In Kulikov & Lopuhaä (2006), $f_0(0)$ is estimated by the maximum likelihood estimator $\hat{f}_n$ evaluated at a small positive (but vanishing) number: $\hat{f}_n(cn^{-1/3})$ for some $c > 0$. Of course, the estimator depends on the choice of the parameter c.

In Kulikov & Lopuhaä (2006), Theorem 3.1, it is shown that

$$A_{21}\left\{n^{1/3}(\hat{f}_n(cB_{21}n^{-1/3}) - f_0(cB_{21}n^{-1/3})) + cB_{21}f_0'(0)\right\}$$

converges in distribution to $D_R[W(t) - t^2](c)$ when $n \to \infty$. Here $D_R[Z(t)](c)$ is the right derivative of the least concave majorant on $[0, \infty)$ of the process $Z(t)$, evaluated at c. Furthermore, $B_{21} = 4^{1/3}f_0(0)^{1/3}|f_0'(0)|^{-2/3}$ and $A_{21} = \sqrt{B_{21}/f_0(0)}$.

Based on this asymptotic result, two estimators are proposed, denoted as f^S and f^A ('S' for simple, 'A' for adaptive). The first is a simple one with $cB_{21} = 1$, then $f^S(0) = \hat{f}_n(n^{-1/3})$. The second is $f^A(0) = \hat{f}_n(c^*B_{21}n^{-1/3})$, where $c^* \approx 0.345$ is taken such that the the second moment of the limiting distribution is minimized. Of course, to really turn this into an estimator, B_{21} has to be estimated. Details on this are presented in section 2.5.5.

2.4.3. Histogram estimator

In chapter 2 of Groeneboom & Jongbloed (2014) a natural and simple histogram-type estimator for $f_0(0)$ is proposed. Let $\{b_n\}$ be a vanishing sequence of positive numbers and consider the estimator $f^H(0) = b_n^{-1}\mathbb{F}_n(b_n)$, where $\mathbb{F}_n$ is the empirical distribution of $X_1, \ldots, X_n$. It can be shown that $\mathrm{E}f^H(0) - f_0(0)$ behaves like $b_n f_0'(0)/2$ and the variance of $f^H(0)$ behaves like $f_0(0)/(b_n)$ as $n \to \infty$. Then the asymptotic mean square error (MSE) optimal choice for b_n is $(2f_0(0)/f_0'(0)^2)^{1/3}n^{-1/3} = 2^{-1/3}B_{21}n^{-1/3}$, where B_{21} is as defined in the Section 2.4.2.

2.5. Numerical illustrations

In this section we use the algorithm described in Section 2.3 to sample from the posterior distribution. We consider two data generating settings for the true density function: the standard Exponential distribution and the half-Normal distribution. Both densities are bounded, decreasing and satisfy assumption 2.2.2. Suppose in the j-th iteration of the Gibbs sampler (possibly after discarding "burn in" samples) we have obtained $\left(\Theta_{Z_1}^{(j)}, \ldots, \Theta_{Z_n}^{(j)}\right)$. At iteration j, if the stationary region of the mcmc sampler has been reached, a sample from the posterior distribution is given by

$$\hat{f}^{(j)}(x) := \frac{1}{n}\sum_{i=1}^{n} \psi_x(\Theta_{Z_i}^{(j)}). \tag{2.21}$$

Two natural derived Bayesian point estimators are the posterior mean and the median. Assuming J iterations, a Rao-Blackwellized estimator for the posterior mean is obtained by computing $J^{-1}\sum_{j=1}^{J}\hat{f}^{(j)}(x)$ and an estimator for the posterior median at x is the median value in $\{\hat{f}^{(j)}(x),\ j = 1, \ldots, J\}$. We implemented our procedures in **Julia**, see Bezanson et al.(2017). The computer code and datasets for replication of our examples forms part of the BayesianDecreasingDensity repository (https://github.com/fmeulen/BayesianDecreasingDensity). For plotting we used functionalities of the **ggplot2** package (see Wickham (2016)) in **R**. The computations were performed on a MacBook Pro, with a 2.7GHz Intel Core i5 with 8 GB RAM.

2.5.1. Base measures

To assess the influence of the base measure in the Dirichlet-process prior, we consider the following choices for the base measure:

(A) The density of the base measure vanishes exponentially fast near zero, as the lower bound of Assumption 2.2.1 requires:

$$g_0(\theta) \propto e^{-\theta-\theta^{-1}}\mathbf{1}_{[0,\infty)}(\theta). \tag{2.22}$$

(B) The density of the Gamma$(2,1)$ distribution

$$g_0(\theta) = \theta e^{-\theta}\mathbf{1}_{[0,\infty)}(\theta).$$

(C) The density of the Pareto$(\bar{\alpha},\tau)$ distribution. That is

$$g_0(\theta) = \bar{\alpha}\tau^{\bar{\alpha}}\theta^{-\bar{\alpha}-1}\mathbf{1}_{[\tau,\infty)}(\theta).$$

Here, we consider various choices for the threshold parameter τ.

(D) The density is obtained as a mixture of the Pareto$(\bar{\alpha},\tau)$ density, where the mixing measure on τ has the Gamma(λ,β) distribution. This implies that $g_0(\theta) \asymp \theta^{\lambda-1}$ for $\theta \downarrow 0$. The parameter $\bar{\alpha}$ is fixed here, but could be equipped with with a "hyper" prior without adding much additional computational complexity.

Note that cases (A), (B), (D)(when $\lambda > 1$) satisfy Assumption 2.2.1 and case (C) does not. In cases (A) and (B) the update on the "cluster centra" θ does not boil down to sampling from a "standard" distribution. In this case either rejection sampling or a Metropolis-Hastings step can be used, the details of which are given in section A.2 in the appendix. In case (C) we have partial conjugacy, which in this case means that the θ's can be sampled from a Pareto distribution. Finally, case (D) can be dealt with by Gibbs sampling. More precisely, conditional on the current value of τ, the θ's can be sampled from the Pareto distribution just as in case (C). Next, τ is sampled conditional on $(\theta_1,\theta_{\#z})$ from the density

$$p(\tau \mid \theta_1,\theta_{\#z}) \propto p(\theta_1,\theta_{\#z} \mid \tau)p(\tau) \propto \tau^{\lambda+(\#z)\bar{\alpha}-1}e^{-\beta\tau}\mathbf{1}\{\tau \leq \min(\theta_1,\ldots,\theta_{\#z})\}$$

(where we use "Bayesian notation", to simplify the expressions). Hence, this boils down to sampling from a truncated Gamma distribution.

2.5.2. Estimates of the density for two simulated datasets

We obtained datasets of size 100 by sampling independently from both the standard Exponential distribution and the halfNormal distribution. In the prior specification, the concentration parameter α was fixed to 1 in all simulations, while the base measure was varied over cases (A), (B), (C) with $\bar{\alpha} = 1$, $\tau \in \{0.005, 0.05, 0.5\}$

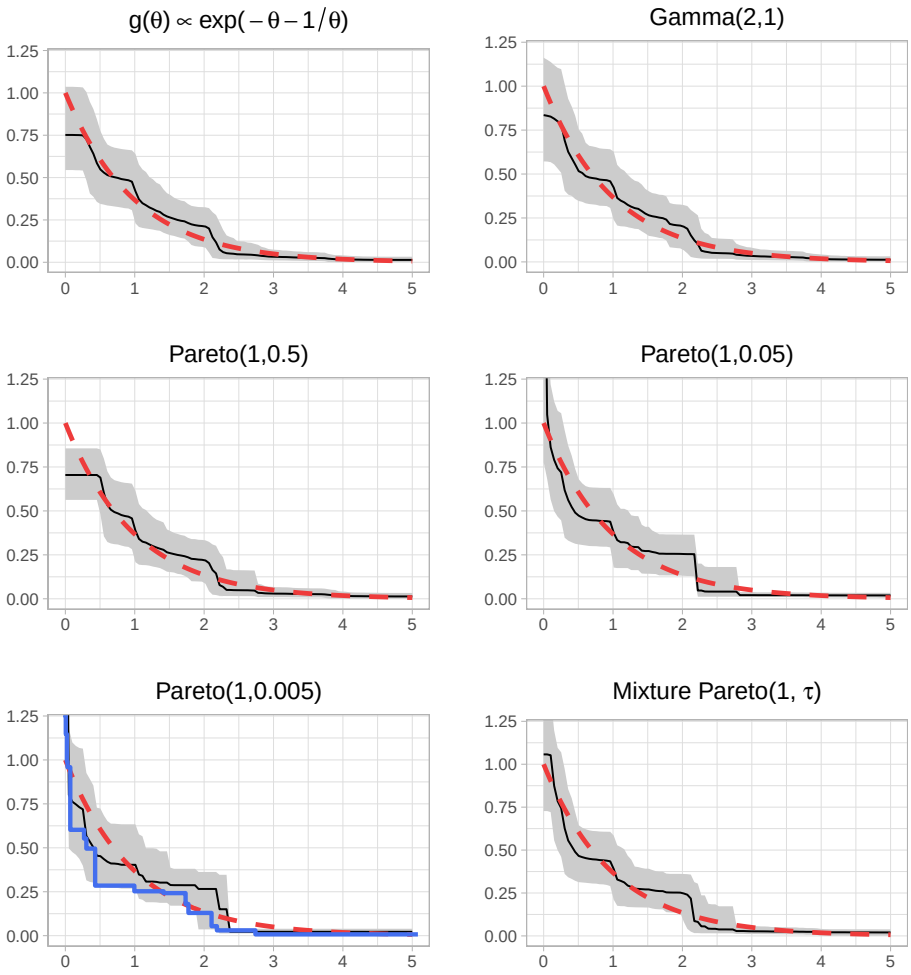

Figure 2.1.: In each panel the same dataset was used, which is a sample of size 100 from the standard Exponential distribution. The black curve is the posterior mean and the shaded grey area depicts pointwise 95% credible intervals. The dashed red curve is the true density. The title in each of the figures refers to the base measure. In the mixture Pareto case, the mixing measure on τ was taken to be the Gamma$(2, 1)$ distribution. In the lower left figure, the solid blue step-function is the maximum likelihood estimate. The inconsistency of this estimator at zero is clearly visible. Moreover, the figure suggests also inconsistency of the posterior mean when the base measure is taken to be the Pareto$(1, 0.005)$ distribution.

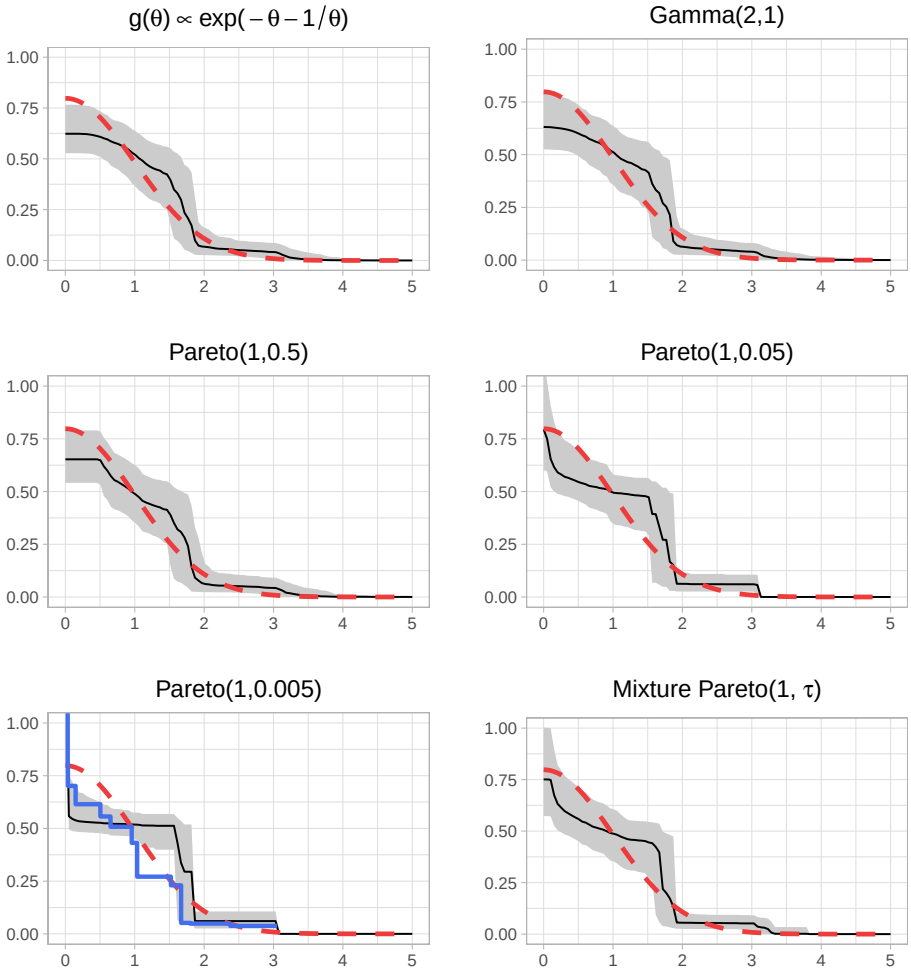

Figure 2.2.: Same experiment as in Figure 2.1, this time with a sample of size 100 from the halfNormal distribution.

and (D) with $\bar{\alpha} = 1$, $\lambda = 2$ and $\beta = 1$. The algorithm was run for 50.000 iterations and the first half of the iterates were discarded as burn in. The computing time was approximately 2 minutes. In case Metropolis-Hastings steps were used for updating θ's, the acceptance rates of the random-walk updates was approximately 0.35, both in case (A) and (B). The results are displayed in figures 2.1 and 2.2. From the top figures we see that the posterior mean and pointwise credible bands visually look similar for the choices of base-measure under (A) and (B). If the base measure is chosen according to (C), the middle and bottom-left figures show the

effect of the parameter τ. Choosing τ too small (here: 0.005) the posterior mean appears inconsistent at zero, similar as the Grenander estimator which is added to the figure for comparison. For somewhat larger values of τ (middle-left figure), the estimate near zero is like a histogram estimator. Finally, the bottom-right figure shows the posterior mean under the base measure specification (D). Here, the posterior mean looks comparable as obtained under (A) and (B), suggesting that we are able to learn the parameter τ from the data. In fact, whereas the prior mean of τ equals 2, the average of the non burn in samples of τ equals 0.66. We have repeated the whole experiment with sample size 1000. The results are in Appendix A.3.

2.5.3. Distribution of the posterior mean for $f(0)$ under various bases measures

In this section we compare base measures (A), (B) and (D) for estimating f at zero. In the experiment, we considered samples of sizes either 50 or 250. We computed the posterior mean for $f(0)$ for each sample based on $10,000$ MCMC-iterations, discarding the first half as burnin. The Monte-Carlo sample size was taken equal to 500. Figure 2.3 summarises the results. While the density for base measure (D) is slightly more spread, contrary to base measures (A) and (B), it concentrates on correct values for both the Exponential and HalfNormal distribution.

2.5.4. Empirical assessment of the rate of contraction

We also performed a large scale experiment to empirically assess the rate of contraction of the posterior median at zero, under either choices (A), (B) or (D) for the base measure. Our proof for deriving the contraction rate really requires a base-measure as under (A) and now the underlying idea is to see in a simulation study whether $g_0(\theta) \sim \theta$ for θ near 0 is suitable or not. In the experiment, we first fixed a sample size n and generated n independent realisations from the standard Exponential distribution. We then ran the MCMC sampler for 20.000 iterations, and kept the final iterate for initialisation of all chains ran for that particular sample size. Next, we repeated 50 times

1. sample a dataset of size n from the standard Exponential distribution;

2. run the MCMC algorithm for 2500 iterations;

3. compute the median value at zero obtained in those samples.

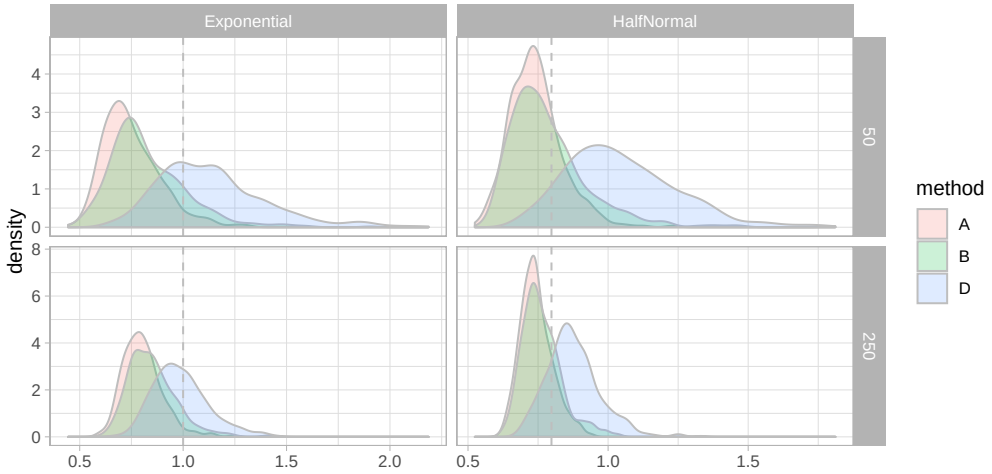

Figure 2.3.: Posterior mean estimator for $f(0)$ for sample sizes $50, 250$ in case the true data-generating distribution is either standard Exponential or halfNormal. The posterior mean is computed by taking $10,000$ MCMC-samples and discarding the first $5,000$ as burnin samples. The Monte-Carlo sample size was taken equal to 500. For the considered sample sizes, only method (D) concentrates around the correct values.

The Metropolis-Hastings proposals for updating the θ's were tuned such that the acceptance rate was about 20% in all cases. If the averages are denoted by $y_1, \ldots, y_{100}$, we finally computed the Root Mean Squared Error, defined by $\sqrt{0.02 \sum_{i=1}^{50} (y_i - 1)^2}$. By repeating this experiment for all three choices of base measure and various values of n, we obtained figure 2.4. The contraction rate is an asymptotic property, and hence there is definitely uncertainty on which values of n correspond to that. The computed slopes do not give a conclusive answer to the actual rate of contraction. For the halfNormal distribution, it is conceivable that methods (A) and (B) yield rate $n^{-1/3}$, whereas method (D) gives a rate almost $n^{-1/2}$. The latter can intuitively be explained by the fact that the slope of the density of the halfNormal is zero at zero which coincides with realisations from the prior. For the Exponential distribution, methods (A) and (B) support rate $n^{-2/9}$, whereas method (D) has worse rates. For completeness, we tabulated the computed slopes in Table 2.1. The difficulty with rate-assessment by finite samples for Dirichlet mixture priors has been noted recently in Wehrhahn, Jara & Barrientos (2019) as well.

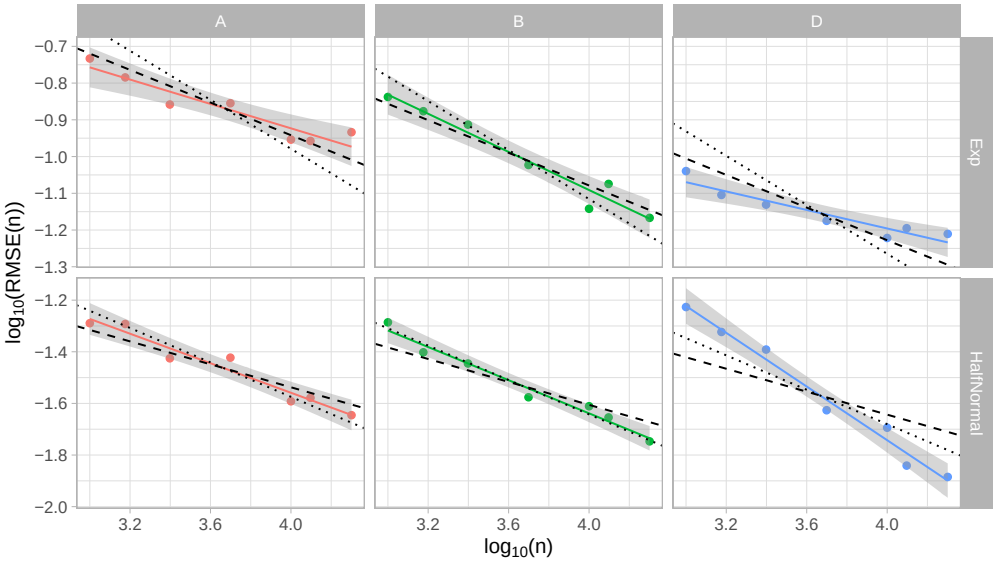

Figure 2.4.: The base10-log of the RMSE versus the base10-log of the sample size under 3 different base measures (method A: $g(\theta) \propto \exp(-\theta - 1/\theta)$, method B: $g(\theta) \propto \theta \exp(-\theta)$, method D: mixture of Pareto). Each dot corresponds to the average of the posterior means using Monte-Carlo size 50. In each panel a least-squares fit is added along with a 95%-confidence interval. The dashed and dotted lines are best least squares fits with slopes $-2/9$ and $-1/3$ respectively.

	Method		
	A	B	D
Exp	−0.166	−0.260	−0.126
halfNormal	−0.286	−0.321	−0.520

Table 2.1.: Slopes of fitted lines in Figure 2.4.

2.5.5. Comparing between Bayesian and various frequentist methods for estimating f_0 at 0

In this section we present a simulation study comparing our Bayesian estimator (posterior median) with various frequentist estimators available for $f_0(0)$ discussed in section 2.4. We simulated 50 samples of sizes $n = 50, 200, 10000$ from the standard exponential distribution and halfNormal distribution. For each sample, the following estimators are calculated: the posterior median estimator f^B, the penalized NPMLE f^P, the two estimators f^S and f^A and the histogram type

estimator f^H. All these estimators require choosing some input parameters.

1. The posterior median estimator $f^B(0)$ is computed using the DPM prior with concentration parameter $\alpha = 1$ and base measure in (2.22). The total number of MCMC iterations was chosen to be 30000, with 15000 burn-in iterations. The posterior median was computed as median value of samples for $\hat{f}(0)$ in equation (2.21).

2. For the penalized estimator $f^P(\alpha_n, 0)$ the parameter $\alpha_n = 0.649\hat{\beta}_n^{-1/3}n^{-2/3}$ was taken with

$$\hat{\beta}_n = \max\left\{ f^P(\alpha_0, 0)\frac{f^P(\alpha_0, 0) - f^P(\alpha_0, x_m)}{2x_m}, n^{-1/3} \right\}.$$

   Here x_m is the second point of jump of $f^P(\alpha_0, \cdot)$ and $\alpha_0 = 0.0516, 0.0205$ for $n = 50, 200$ (listed in Woodroofe & Sun (1993)).

3. For $f^S(0) = \hat{f}_n(n^{-1/3})$ no tuning is needed. For the other estimator we take $f^A(0) = \hat{f}_n(0.345\hat{B}_{21}n^{-1/3})$, where

$$\hat{B}_{21} = 4^{1/3}f^S(0)^{1/3}|\hat{f}_n'(0)|^{-2/3}, \tag{2.23}$$

   a consistent estimator of B_{21} where

$$\hat{f}_n'(0) = \min\{n^{1/6}(\hat{f}_n(n^{-1/6}) - \hat{f}_n(n^{-1/3})), -n^{-1/3}\}.$$

4. For the histogram estimator $f^H(0) = \mathbb{F}_n(\hat{b}_n)/\hat{b}_n$, $\hat{b}_n = 2^{-1/3}\hat{B}_{21}n^{-1/3}$ was chosen with $\hat{B}_{21}$ as in (2.23).

Figure 2.5 shows, for each combination of sample size and estimation method described, the boxplots of the 50 realized values based on samples from the standard exponential distribution. Figure 2.6 shows these boxplots for the samples from the halfNormal distribution.

In table 2.2 we compare the bias, variance and mean squared error of these consistent estimators based on data from the standard exponential distribution. For the standard exponential data, the penalized estimator $f^P(0)$ performs best in the MSE sense. The Bayesian estimator f^B has smallest variance, but big bias when the sample size is large ($n = 10000$). This might be explained by the small contraction rate $n^{-1/6}$ at zero, but also by the fact that the Bayesian method is not specifically aimed at only estimating the density at zero, but instead the full density.

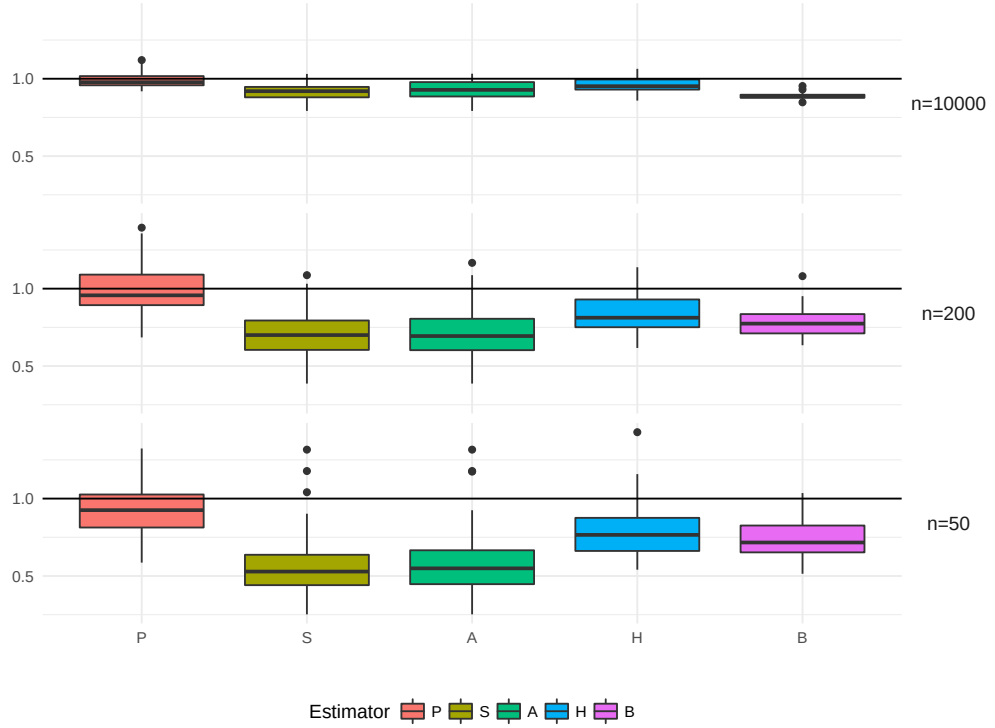

Figure 2.5.: Boxplots based on 50 replications, where a sample of size n is drawn from the standard exponential distribution. Here P,S,A,H,B correspond to the penalized maximum likelihood-, simple-, adaptive-, histogram- and posterior median- estimator respectively. The horizontal lines indicate the true value of $f_0(0) = 1$.

Table 2.3 lists the bias, variance and MSE values of the estimators with observations sampled from the halfNormal distribution. For the halfNormal data, the histogram estimator f^H behaves best in the bias and MSE sense. This can probably be explained by the behaviour of f_0 near zero, note that $f_0'(0) = 0$ in the halfNormal case. The estimator for $f_0'(0)$, $\hat{f}_n'(0)$, probably quite unstable which leads to big value for $\hat{B}_{21}$ resulting in a big bandwidth $\hat{b}_n$. As the behaviour of the underlying density is "flat" near zero, the MSE-optimal choice of bandwidth is of the slower order $n^{-1/5}$. The posterior mean again has smallest variance.

2.5.6. Application to fertility data

In Keiding et al. (2012) data concerning the fertility of a population are analysed. The aim is to estimate the distribution of the duration for women to become preg-

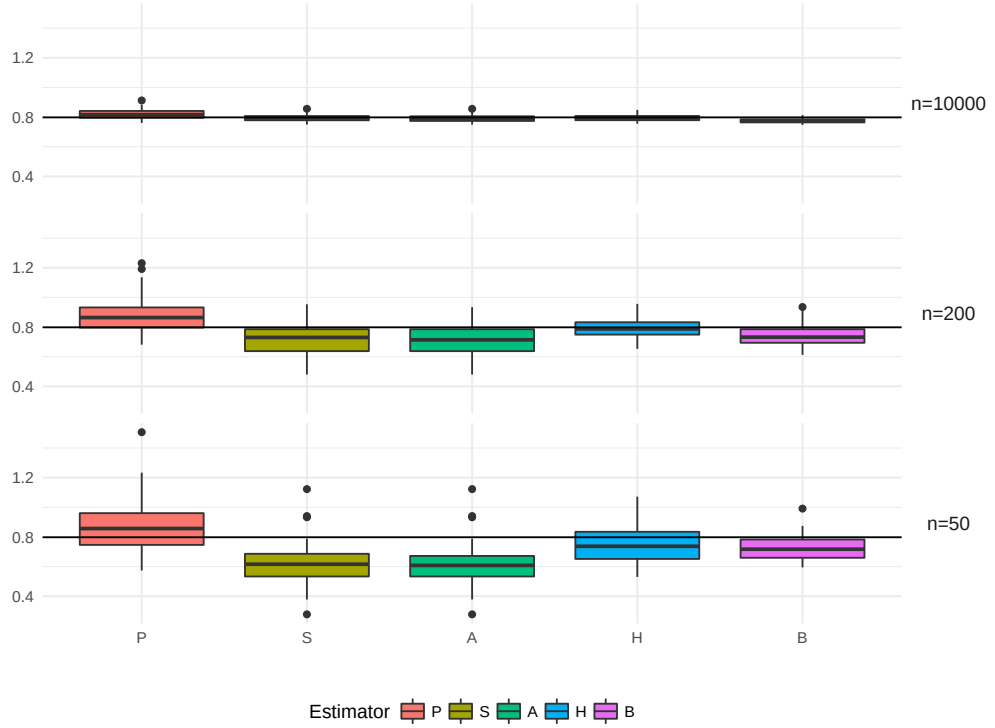

Figure 2.6.: Boxplots based on 50 replications, where a sample of size n is drawn from the halfNormal distribution. The rows correspond to the sample sizes $n = 50$, 200 and 10000. Here P,S,A,H,B correspond to the penalized maximum likelihood-, simple-, adaptive-, histogram- and posterior median- estimator respectively. The horizontal lines indicate the true value of $f_0(0) = \sqrt{2/\pi}$.

nant from when they start attempting, based on data from so-called current durations. These current durations can be modeled as described in the introduction. Indeed, the true durations are modeled as sample from an unknown distribution function H_0. According to length-biased sampling, individuals are selected and then the time since the start of attempting to become pregnant is administered. This is called the current duration, and can be seen as a uniform random fraction of the true duration of the selected individual. This current duration then has bounded decreasing probability density f_0 as given in (2.2). The distribution function of the durations H_0, can be expressed in terms of f_0 as in Equation(2.3). For more information on the design of this study we refer to Keiding et al. (2012). For illustration purpose we only used the $n = 618$ measured current durations that do not exceed 36 months. Figure 2.7 shows the histogram of 618 raw data, modeled

n		f^P	f^S	f^A	f^H	f^B
50	Bias	-0.067	-0.423	-0.402	-0.214	-0.266
	Var	0.033	0.042	0.049	0.030	0.013
	MSE	0.037	0.222	0.210	0.076	0.084
200	Bias	-0.001	-0.286	-0.271	-0.158	-0.221
	Var	0.029	0.020	0.027	0.015	0.007
	MSE	0.029	0.101	0.100	0.040	0.056
10000	Bias	-0.011	-0.084	-0.072	-0.041	-0.112
	Var	0.002	0.002	0.003	0.002	0.0004
	MSE	0.002	0.010	0.009	0.004	0.013

Table 2.2.: Simulated bias, variance and mean squared error for the five estimators from standard exponential distribution.

n		f^P	f^S	f^A	f^H	f^B
50	Bias	0.063	-0.182	-0.185	-0.043	-0.073
	Var	0.029	0.022	0.022	0.016	0.007
	MSE	0.033	0.055	0.056	0.018	0.012
200	Bias	0.080	-0.086	-0.088	-0.011	-0.051
	Var	0.014	0.012	0.012	0.004	0.005
	MSE	0.020	0.019	0.020	0.004	0.008
10000	Bias	0.0216	-0.0022	-0.0060	-0.0019	-0.0239
	Var	0.0010	0.0005	0.0006	0.0005	0.0002
	MSE	0.0015	0.0005	0.0006	0.0005	0.0008

Table 2.3.: Simulated bias, variance and mean squared error for the five estimators based on samples from the standard halfNormal distribution.

as sample from the decreasing density f_0.

In this section we estimate the density f_0 using base measure choice (A) which satisfies assumption 2.2.1 and (D) which does not satisfy assumption 2.2.1 with concentration parameter $\alpha = 1$. Then each MCMC iterate of the posterior mean can be converted to an iterate for H_0 using the relation (2.3). In Groeneboom & Jongbloed (2015) chapter 9, pointwise confidence bands for f_0 and H_0 are constructed based on the smoothed maximum likelihood estimator. Having derived the estimators, producing such confidence bands needs quite some fine tuning. In this section, we construct the Bayesian counterpart of the confidence bands, credible regions for H_0. Contrary to the frequentist approach, having the machinery

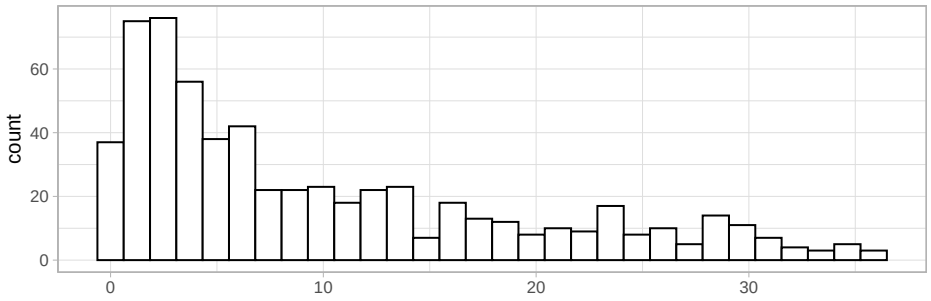

Figure 2.7.: Histogram of the current durations fertility data that do not exceed 36 months.

available for computing the posterior mean, the pointwise credible sets can be obtained directly from the MCMC output. The results for the fertility data are shown in Figures 2.8 using base measures (A) and (D) respectively.

2.6. Discussion

In this paper we have used Bayesian analysis to nonparametrically estimate a decreasing density based on a random sample. Particular emphasis is given to estimation of the density at zero and sufficient criteria on the base measure of the prior are derived to obtain contraction rate $n^{-2/9}$. Besides a base measure attaining this rate, we have investigated the relative performance of other base measures by means of a Monte Carlo study. This study was extended to compare multiple frequentist estimators for estimating the density at zero to a Bayesian derived point estimator.

It remains an open question whether for a given density function f there exists a base measure such that the contraction rate for estimation of $f(0)$ is $n^{-1/3}$. From the simulation study it appears that taking a mixture of Pareto densities as base measure empirically yields satisfactory performance and henceforth we recommend taking base measure (D) from Section 2.5.1.

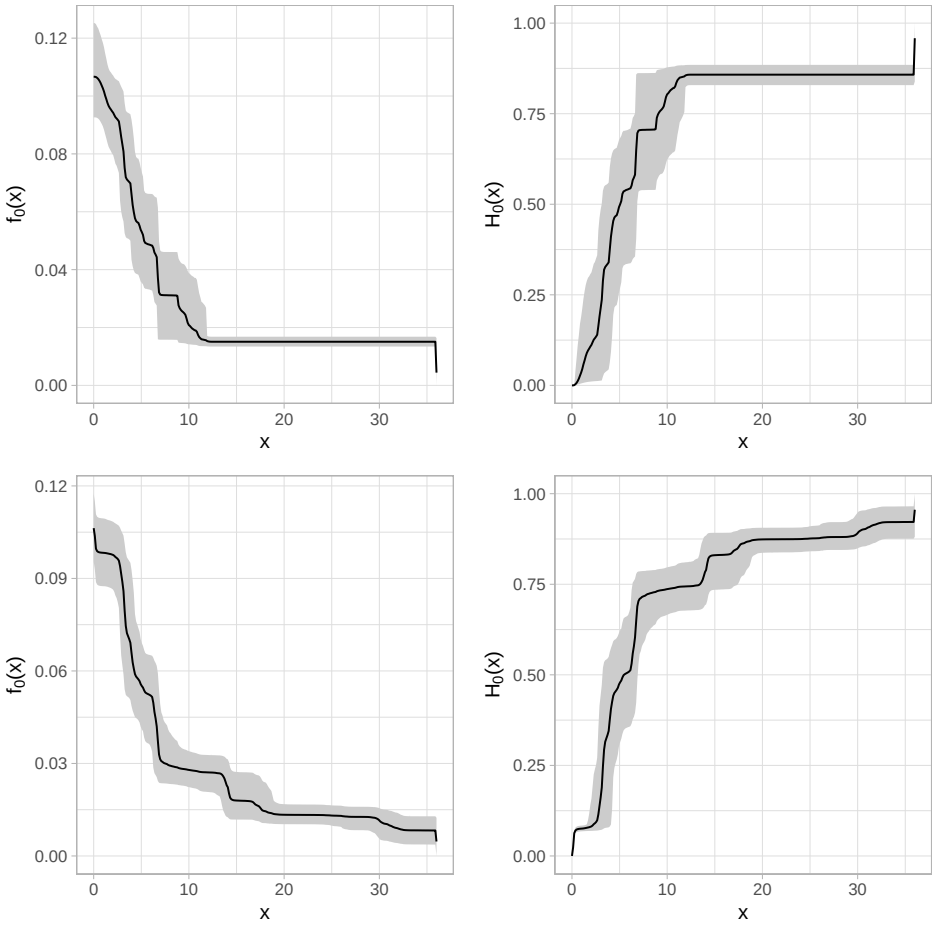

Figure 2.8.: Fertility data. Top: results for base measure (A) and $\alpha = 1$. Bottom: results for base measure (D) and $\alpha = 1$. Left: posterior mean and 95% pointwise credible sets for probability density function f_0. Right: corresponding estimate and pointwise credible sets for the distribution function $H_0(x) = 1 - f_0(x)/f_0(0)$.

3. Bayesian nonparametric estimation of a concave distribution function with mixed interval censored data

Assume we observe a finite number of inspection times together with information on whether a specific event has occurred before each of these times. Suppose replicated measurements are available on multiple event times. The set of inspection times, including the number of inspections, may be different for each event. This is known as mixed case interval censored data. We consider Bayesian estimation of the distribution function of the event time while assuming it is concave. We provide sufficient conditions on the prior such that the resulting procedure is consistent from the Bayesian point of view. We also provide computational methods for drawing from the posterior and illustrate the performance of the Bayesian method in both a simulation study and two real data sets.

3.1. Introduction

In survival analysis, one is interested in the time a certain event occurs. For example, the event may be the onset of a disease. A well known complication often encountered in practice is censoring, where the precise time at which an event occurs is unknown, but partial information on it is available. In right censoring for example, one only observes the event if it occurs before a certain censoring time, otherwise one observes the censoring time accompanied by the information that the event occurred after this time. In interval censoring, one never sees the exact event time. Only an interval of positive length (possibly infinite) is observed which contains the event time of interest.

Suppose X models the actual event time for one subject. Instead of observing X directly, we observe a finite number of $i$nspection times $0 < t_1 < t_2 < \cdots < t_k < \infty$, together with the information which of the intervals $(t_{j-1}, t_j]$ contains X. We will assume a setting in which we obtain data that are modelled as independent and

identically distributed realisations of $X_1, \ldots, X_n$, each of which is distributed as X. For each subject, the set of inspection times, as well as the number of inspections, may be different. This type of data is known as *mixed-case interval censored data*. Our model includes both the interval censring case 1 model (also known as current status model) and interval censoring case 2 model for which $k = 1$ and $k = 2$ respectively. In many statistical models, there are reasons to impose specific assumptions on functional parameters, for example shape constraints. Incorporating such constraints into the estimation procedure often improves the accuracy of the resulting estimator. In this paper, we consider the problem of estimating the distribution function F of X, assuming that F is concave.

3.1.1. Related literature

In Groeneboom & Wellner (1992), the point-wise asymptotic distribution of the maximum likelihood estimator (mle) of the distribution function in the interval censoring case 1 model is derived. For interval censoring case 2, the asymptotic point-wise distribution of the mle is still not known. In the mixed case interval censoring model, the mle has been studied by Schick & Yu (2000) where it is shown to be L_1-consistent. In Wellner & Zhang (2000) a panel count model is considered, which includes the mixed case interval censoring model as a special case, namely when the counting process has only one jump. For this panel count model, Wellner & Zhang (2000) study two estimators. In case the counting process has only one jump and there is one inspection time, their estimators coincide with the mle for current status data ($k = 1$). If $k > 1$, this is not the case. Dümbgen, Freitag & Jongbloed (2004) consider the current status model with the additional constraint that the underlying distribution function F_0 is concave. It is shown that the supremum distance between the nonparametric least squares estimator and the underlying distribution function F_0 is of order $(\log n/n)^{2/5}$. For mixed case interval censoring, the MLE is shown to be asymptotically consistent under the assumption that F_0 is concave or convex-concave in Dümbgen, Freitag & Jongbloed (2006). In addition, an algorithm for computing the mle is proposed there.

From the Bayesian perspective, Susarla & Van Ryzin (1976) derived a nonparametric Bayesian estimator for the event time distribution function based on right-censored data, using the Dirichlet process prior. A special feature in this right-censoring model is that the posterior mean estimator can be constructed explicitly. For interval censored data, this explicit construction is not available. Calle & Gómez (2001) propose a nonparametric Bayesian approach in the interval censoring model and use a Markov Chain Monte Carlo algorithm to obtain estimators for the posterior mean. Doss & Huffer (2003) consider the Dirichlet Process prior

in the interval censoring model. They develop and compare various Monte Carlo based algorithms for computing Bayesian estimators. A host of closely related Bayesian nonparametric models have been implemented in the DP-package in the R-language, Cf. Jara et al. (2011).

3.1.2. Contribution

In this paper, we define and study a Bayesian estimator of the event time distribution based on *m*ixed-case interval censored data under the additional assumption that the distribution function is concave. An advantage of the Bayesian setup is the ease of constructing credible regions. To construct frequentist analogues of these, confidence regions, can be quite cumbersome, relying on either bootstrap simulations or asymptotic arguments. We address this problem from a theoretical perspective and provide conditions on the prior such that the resulting procedure is consistent. That is, assuming data are generated from a "true" distribution, we show that the posterior asymptotically (as the sample size increases) converges to this distribution. The proof relies on Schwartz' method for proving posterior consistency (Cf. Section 6.4 in Ghosal & Van der Vaart (2017)). In addition, we provide computational methods for drawing from the posterior and illustrate its performance in a simulation study. Finally, we apply the Bayesian procedure on two real data sets and construct pointwise credible sets.

3.1.3. Outline

Section 3.2 sets off with introducing notation and formally describing the model. In section 3.3 we derive posterior consistency under a weak assumption on the prior distribution on the class of concave distribution functions. A Markov Chain Monte Carlo algorithm for obtaining draws from the posterior using the Dirichlet Mixture Process prior is detailed in section 3.4. In section 3.5 we perform a simulation study to illustrate the behaviour of the proposed Bayesian method. Furthermore, we apply it to two data sets in section 3.6, one concerned with Rubella and the other with breast cancer. The appendix contains proofs of some technical results.

3.2. Model, likelihood and prior

3.2.1. Model and likelihood

Suppose X is a random variable in $[0, \infty)$ with concave distribution function F_0. Instead of observing X, we observe the random vector (K, T, Δ) that is constructed

as follows. First, K is sampled from a discrete distribution with probability mass function p_K on $\{1, 2 \dots\}$, representing the number of inspection times. Given $K = k$, $T \in \mathbb{R}^k$ is sampled from a density g_k supported on the set $\{t = (t_1, \dots, t_k) \in (0, L]^k : 0 < t_1 < \cdots < t_k < \infty\}$ for some constant L. This random vector contains the (ordered) inspection times. Finally, $\Delta \in \{0, 1\}^{k+1}$ is the vector indicating in which of the $k + 1$ intervals generated by T the event actually happened. Thus, it is defined as the vector with j-th component

$$\Delta_j = 1_{(T_{j-1}, T_j]}(X) \text{ for } 1 \leq j \leq k + 1$$

where $T_0 = 0$ and $T_{k+1} = \infty$ by convention.

This procedure is repeated independently, so for sample size n the data is a realisation of

$$\mathcal{D}_n := \{(K_i, T^i, \Delta^i) = (K_i, T_{i,1}, \dots, T_{i,K_i}, \Delta_{i,1}, \dots, \Delta_{i,K_i+1}),\ i = 1, \dots, n\}.$$

Define the sets

$$\mathcal{C}_k = \{t \in (0, L]^k : 0 < t_1 < \cdots < t_k < \infty\} \tag{3.1}$$

and $\mathcal{H}_k = \{\delta \in \{0, 1\}^{k+1} : \sum_{j=1}^{k+1} \delta_j = 1\}$, $k = 1, 2, \dots$. Then $\mathcal{D}_n \in (\bigcup_{k=1}^{\infty} \{k\} \times \mathcal{C}_k \times \mathcal{H}_k)^n$.

Upon conditioning on the observed inspection times, we can define the likelihood of the distribution function F by

$$L(F) = \prod_{i=1}^{n} \Big(p_K(K_i) g_{K_i}(T^i) \prod_{j=1}^{K_i+1} (F(T_{i,j}) - F(T_{i,j-1}))^{\Delta_{i,j}} \Big). \tag{3.2}$$

We denote the joint distribution of $\{(K_i, T^i),\ 1 \leq i \leq n\}$ by $\mathbb{P}_{K,T}$. Given these $(K_i, T^i)s$ the vectors Δ^i have multinomial distributions with probabilities depending on F_0. The distribution of $\mathcal{D}_n$ will be denoted by $\mathbb{P}_0$. Expectation with respect to measures will be denoted by $\mathbb{E}$, supplemented by a subscript referring to the measure.

3.2.2. Prior specification

In order to estimate the underlying concave distribution function in a Bayesian way, we construct a prior distribution on the set of all concave distribution functions. For $\theta > 0$, denote the uniform density function on $[0, \theta]$ by $\varphi(\cdot \mid \theta)$ and its distribution function by $\Psi(\cdot \mid \theta)$, i.e.

$$\varphi(x, \theta) = \frac{1}{\theta} 1\{x \leq \theta\} \text{ and } \Psi(x, \theta) = \frac{\min(x, \theta)}{\theta} \text{ respectively}, \quad x \geq 0. \tag{3.3}$$

It is well known that any concave distribution function F on $[0,\infty)$ allows the mixture representation (see Feller (1966))

$$F(x) = \int \Psi(x,\theta)\,\mathrm{d}G(\theta), \tag{3.4}$$

where G is a distribution function on $[0,\infty)$. In what follows, we sometimes stress this representation and denote the concave distribution function by F_G. In order to put a prior measure Π on the set

$$\mathcal{F} = \Big\{F : F \text{ is a concave distribution on } [0,\infty)\Big\},$$

we use (3.4) together with a prior distribution Π^* on the set of all mixing distribution functions G on $(0,\infty)$ (denote as $\mathcal{M}$). Having chosen such a prior measure, we denote the resulting posterior measure on $\mathcal{F}$ by $\Pi(\cdot|\mathcal{D}_n)$.

3.3. Posterior consistency

In this section we establish consistency of the posterior distribution $\Pi(\cdot|\mathcal{D}_n)$ under a weak condition on the prior measure Π. Generally, the posterior is said to be consistent at F_0 (with respect to a semimetric d) if for any $\varepsilon > 0$, $\mathbb{E}_0\Pi(d(F,F_0) > \varepsilon \mid \mathcal{D}_n) \to 0$ when $n \to \infty$.

For any distribution function G, denote $G_{i,j} = G(T_{i,j}) - G(T_{i,j-1})$. Given the inspection times $\{T^i, 1 \le i \le n\}$, we say that distribution functions G and F belong to the same equivalence class if the increments between the adjacent times are the same: $G_{i,j} = F_{i,j}$ for all $i = 1,\dots,n$, $j = 1,\dots,K_i+1$. Then given data $\mathcal{D}_n$, we define a distance d between two (equivalence classes of) distribution functions G and F by

$$d_n(G,F) = \frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{K_i+1} |G_{i,j} - F_{i,j}|\,. \tag{3.5}$$

Recall that Π^* is a prior on the set $\mathcal{M}$, then G is in the weak support of Π^* if every weak neighborhood of G has positive measure.

Theorem 3.3.1. *Fix $F_0 \in \mathcal{F}$ and $x \in [0,\infty)$. Consider the mixed-case interval censoring model described in section 3.1. Assume F_0 has a continuous density function f_0 on $(0,\infty)$ with $f_0(0) \le M < \infty$ and that the weak support of the prior distribution Π^* is $\mathcal{M}$. If $\mathbb{E}K^r < \infty$, for some $r > 1/2$, then for any $\epsilon > 0$, we have $\mathbb{P}_0$-almost surely that*

$$\Pi(F \in \mathcal{F} : d_n(F,F_0) > \epsilon |\, \mathcal{D}_n) \to 0 \text{ as } n \to \infty.$$

Note that d_n in Theorem 3.3.1 is a random semidistance since it depends on the inspection times $\{K_i, T^i, i = 1, \ldots, n\}$, also depending on n. Define the measure μ on the Borel $\sigma-$field $\mathcal{B}$ on $[0, \infty)$ that measures the "expected proportion of inspection times contained in a Borel set $B \in \mathcal{B}$" by

$$\mu(B) = \sum_{k=1}^{\infty} p_K(k) k^{-1} \int g_k(t) \sum_{j=1}^{k} \mathbf{1}_B(t_j) \, \mathrm{d}t.$$

As a special case, assume that given k, $S_1, \ldots, S_k$ are independent and identically distributed with density function ξ on $[0, \infty)$ and $\{T_1 < T_2 < \cdots < T_k\}$ are the ordered S_j's. Then when $k = 1$,

$$\mu(B) = \int g_1(t_1) \mathbf{1}_B(t_1) \, \mathrm{d}t_1 = \int_B \xi(x) \, \mathrm{d}x.$$

When $k = 2$, for any $a \in [0, \infty)$

$$\begin{aligned}\mu((0, a]) =& \frac{1}{2} \int g_2(t) (\mathbf{1}\{t_1 \leq a\} + \mathbf{1}\{t_2 \leq a\}) \, \mathrm{d}t = \frac{1}{2} (\mathbb{P}(t_1 \leq a) + \mathbb{P}(t_2 \leq a)) \\ =& \frac{1}{2} \left(1 - \left(1 - \int_0^a \xi(x) \, \mathrm{d}x \right)^2 + \left(\int_0^a \xi(x) \, \mathrm{d}x \right)^2 \right) = \int_0^a \xi(x) \, \mathrm{d}x\end{aligned}$$

Hence, measure μ has density ξ in interval case 1 and 2.

The following result establishes posterior consistency with respect to $L_1(\mu)$ loss.

Theorem 3.3.2. *Let F_0, Π and K satisfy the conditions of Theorem 3.3.1. Then for any $\epsilon > 0$, we have*

$$\mathbb{E}_0 \, \Pi \left(F \in \mathcal{F} : \int |F - F_0| \, d\mu > \epsilon | \mathcal{D}_n \right) \to 0 \text{ as } n \to \infty.$$

3.3.1. Proofs

For proving Theorem 3.3.1 we use the Schwartz's approach to derive posterior consistency. In the proof of this theorem, Lemma 3.3.3 is used to control the prior mass of a neighbourhood of the true distribution. Lemma 3.3.4 provides appropriate test functions. Both lemmas are stated below; the proofs are in appendix B.

Lemma 3.3.3. *Let F_0 and Π^* satisfy the conditions of Theorem 3.3.1. Define, for $F_1, F_2 \in \mathcal{F}$, $k = 1, 2, \ldots$ and $t \in \mathcal{C}_k$ as defined in (3.1):*

$$h_{k, F_1, F_2}(t) = \sum_{j=1}^{k+1} (F_0(t_j) - F_0(t_{j-1})) \log \frac{F_1(t_j) - F_1(t_{j-1})}{F_2(t_j) - F_2(t_{j-1})} \tag{3.6}$$

(where $t_0 = 0$ and $t_{k+1} = \infty$ by convention). If we define,

$$S(\eta) = \left\{ F \in \mathcal{F} \colon \sum_{k=1}^{\infty} p_K(k) \int g_k(t) h_{k,F_0,F}(t)\, \mathrm{d}t < \eta \right\}. \tag{3.7}$$

then for all $\eta > 0$, $\Pi(S(\eta)) > 0$.

Note that that for the specific choice $F_1 = F_0$, by Jensen's inequality, $h_{k,F_0,F} \geq 0$ for all $F \in \mathcal{F}$.

Lemma 3.3.4. *For $\epsilon > 0$, define $U_\epsilon := \{F \in \mathcal{F} \colon d_n(F, F_0) > \epsilon\}$. Then there exists a sequence of test functions Φ_n such that for all $n \geq 1$,*

$$\begin{aligned} \mathbb{E}_0(\Phi_n) &\leq Ce^{-nc} \\ \mathbb{E}_{(K,T)} \left\{ \sup_{F \in U_\epsilon} \mathbb{E}_F[1 - \Phi_n | K, T] \right\} &\leq Ce^{-nc} \end{aligned} \tag{3.8}$$

for some positive constants c and C.

Proof of Theorem 3.3.1. Choose $\epsilon > 0$ and define the set U_ϵ as in Lemma 3.3.4. Define

$$Z_{i,j} = \frac{F(T_{i,j}) - F(T_{i,j-1})}{F_0(T_{i,j}) - F_0(T_{i,j-1})}.$$

Using expression (3.2) of the likelihood, the posterior mass of the set U_ϵ can be written as

$$\Pi(U_\epsilon | \mathcal{D}_n) = D_n^{-1} \int_{U_\epsilon} \prod_{i=1}^{n} \prod_{j=1}^{K_i+1} Z_{i,j}^{\Delta_{i,j}}\, \mathrm{d}\Pi(F),$$

where

$$D_n = \int \prod_{i=1}^{n} \prod_{j=1}^{K_i+1} Z_{i,j}^{\Delta_{i,j}}\, \mathrm{d}\Pi(F).$$

Fix $0 < \eta < c/2$, where c is as it appears in Lemma 3.3.4. Also fix $F \in S(\eta)$.

We first show that Lemma 3.3.3 implies for any $\eta' > \eta$ we have $\mathbb{P}_0$-a.s. that

$$D_n \geq \exp(-n\eta')\Pi(S(\eta))$$

for all n sufficiently large. By Lemma 3.3.3, we have $\Pi(S(\eta)) > 0$. Let $\Pi_{S(\eta)}$ be Π restricted to $S(\eta)$ and normalised to a probability measure. For $i \geq 1$ define

$$Y_{i,j} = -\int \Delta_{i,j} \log Z_{i,j}\, \mathrm{d}\Pi_{S(\eta)}(F) \mathbf{1}_{\{1,2,\ldots,K_i+1\}}(j).$$

Note that,

$$\begin{aligned}\mathbb{E}_0\left[\sum_{j=1}^{K_1+1} Y_{1,j}\right] &= \mathbb{E}_{K_1,T_1}\left[\mathbb{E}_{F_0}\left[\sum_{j=1}^{K_1+1} Y_{1,j}|T^{K_1}, K_1\right]\right] \\ &= \mathbb{E}_{K_1,T_1}\left[\sum_{j=1}^{K_1+1}\int -(F_0(T_{1,j}) - F_0(T_{1,j-1}))\log Z_{i,j}\,\mathrm{d}\Pi_{S(\eta)}(F)\right] \\ &= \sum_{k=1}^{\infty} p_K(k)\int\int g_k(t)h_{k,F_0,F}(t)\,\mathrm{d}t\,\mathrm{d}\Pi_{S(\eta)}(F) \le \eta < \infty.\end{aligned}$$

Therefore, the law of large numbers yields

$$\frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{K_i+1} Y_{i,j} \to \mathbb{E}_0\left[\sum_{j=1}^{K_1+1} Y_{1,j}\right] \le \eta, \qquad \mathbb{P}_0 - a.s.$$

Hence, $\mathbb{P}_0$-a.s. for any $\eta' > \eta$,

$$\begin{aligned}D_n &\ge \int_{S(\eta)} \prod_{i=1}^{n}\prod_{j=1}^{K_i+1} Z_{i,j}^{\Delta_{i,j}}\,\mathrm{d}\Pi(F) = \Pi(S(\eta))\int \prod_{i=1}^{n}\prod_{j=1}^{K_i+1} Z_{i,j}^{\Delta_{i,j}}\,\mathrm{d}\Pi_{S(\eta)}(F) \\ &= \Pi(S(\eta))\int \exp\left(\sum_{i=1}^{n}\sum_{j=1}^{K_i+1}\Delta_{i,j}\log Z_{i,j}\right)\,\mathrm{d}\Pi_{S(\eta)}(F) \\ &\ge \Pi(S(\eta))\exp\left(-n\cdot\frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{K_i+1} Y_{i,j}\right) \\ &\ge \exp(-n\eta')\,\Pi(S(\eta)) \qquad (3.9)\end{aligned}$$

for n sufficiently large, where we used Jensen's inequality in the second inequality.

Now we can finish the proof by combining this result with the test functions Φ_n satisfying (3.8) (by Lemma 3.3.4).

By inequality (3.9), we can bound $\mathbb{E}_0\Pi(U_\epsilon|\mathcal{D}_n)$ as follows,

$$\begin{aligned}\mathbb{E}_0\Pi(U_\epsilon|\mathcal{D}_n) &= \mathbb{E}_0\Pi(U_\epsilon|\mathcal{D}_n)\Phi_n + \mathbb{E}_0\Pi(U_\epsilon|\mathcal{D}_n)(1-\Phi_n) \\ &\le \mathbb{E}_0\Phi_n + \Pi(S(\eta))^{-1}e^{n\eta'}\mathbb{E}_0\int_{U_\epsilon}\prod_{i=1}^{n}\prod_{j=1}^{K_i+1} Z_{i,j}^{\Delta_{i,j}}(1-\Phi_n)\,\mathrm{d}\Pi(F) \\ &= \mathbb{E}_0\Phi_n + \Pi(S(\eta))^{-1}e^{n\eta'}\mathbb{E}_{(K,T)}\int_{U_\epsilon}\mathbb{E}_F(1-\Phi_n)\,\mathrm{d}\Pi(F) \\ &\le Ce^{-cn} + \Pi(S(\eta))^{-1}\cdot Ce^{-(c-\eta')n} = o(1) \qquad \text{as} \quad n\to\infty.\end{aligned}$$

The final step follows by choosing $\eta' < c$. Since $\sum_{n=1}^{\infty} e^{-bn} < \infty$ for any constant b, almost sure convergence follows by the Borel-Cantelli lemma. □

Proof of Theorem 3.3.2. First note that the proof of (16) in Dümbgen, Freitag & Jongbloed (2006) shows for all distribution functions $F, F_0 \in \mathcal{F}$,

$$d'_n(F, F_0) = \frac{1}{n} \sum_{i=1}^{n} K_i^{-1} \sum_{j=1}^{K_i} |F(T_{i,j}) - F_0(T_{i,j})| \leq d_n(F, F_0). \tag{3.10}$$

For any $\epsilon > 0$, denote set

$$A_n = \left\{ \sup_{F \in \mathcal{F}} \left| d'_n(F, F_0) - \int |F - F_0| d\mu \right| > \epsilon/2 \right\},$$

Now we prove that $\mathbb{P}_{(K,T)}(A_n) \to 0$ as $n \to \infty$. Fix $F_0 \in \mathcal{F}$ and denote

$$\psi_i(F) = n^{-1} K_i^{-1} \sum_{j=1}^{K_i} |F(T_{i,j}) - F_0(T_{i,j})|.$$

Then $d'_n(F, F_0) = \sum_{i=1}^{n} \psi_i(F)$. Note that $\mathbb{E}_{(K,T)} d'_n(F, F_0) = \int |F - F_0| \, \mathrm{d}\mu$. It is sufficient to show that

$$\mathbb{E}_{(K,T)} \sup_{F \in \mathcal{F}} |d'_n(F, F_0) - \mathbb{E}_{(K,T)} d'_n(F, F_0)| \to 0. \tag{3.11}$$

By theorem B.1.2, it is implied by the existence of a sequence $\delta_n \to 0$ such that

$$\mathbb{E}_{(K,T)} \sum_{i=1}^{n} \sup_{F \in \mathcal{F}} |\psi_i(F)| = O(1), \tag{3.12}$$

$$\mathbb{E}_{(K,T)} \sum_{i=1}^{n} \mathbf{1}\{\sup_{F \in \mathcal{F}} |\psi_i(F)| > \delta_n\} \sup_{F \in \mathcal{F}} |\psi_i(F)| = o(1), \tag{3.13}$$

$$\text{for any } u > 0, \quad \log \mathcal{N}(u, \mathcal{F}, \rho_n) = c(u). \tag{3.14}$$

Here

$$\mathcal{N}(u, \mathcal{F}, \rho_n) = \min \left\{ \#\mathcal{G} : \mathcal{G} \subset \mathcal{F}, \inf_{G \in \mathcal{G}} \rho_n(F, G) \leq u \text{ for all } F \in \mathcal{F} \right\},$$

and

$$\rho_n(F, F') = \sum_{i=1}^{n} |\psi_i(F) - \psi_i(F')|.$$

For (3.12) and (3.13), note that $\sup_{F\in\mathcal{F}}|\psi_i(F)|\leq n^{-1}$, hence $\mathbb{E}_{(K,T)}\sum_{i=1}^n \sup_{F\in\mathcal{F}}|\psi_i(F)|\leq 1$. By taking $n\delta_n\to\infty$, e.g. $\delta=\frac{1}{\sqrt{n}}$,

$$\mathbb{E}_{(K,T)}\sum_{i=1}^n 1\{\sup_{F\in\mathcal{F}}|\psi_i(F)|>\delta_n\}\sup_{F\in\mathcal{F}}|\psi_i(F)|\leq n^{-1}\mathbb{E}\sum_{i=1}^n 1\{n^{-1}>\delta_n\}=1\{n^{-1}>\delta_n\}\to 0$$

For (3.14), note that

$$\begin{aligned}\rho_n(F,F')=\sum_{i=1}^n|\psi_i(F)-\psi_i(F')|&\leq n^{-1}\sum_{i=1}^n K_i^{-1}\sum_{j=1}^{K_i}||F(T_{i,j})-F_0(T_{i,j})|-|F'(T_{i,j})-F_0(T_{i,j})||\\&\leq n^{-1}\sum_{i=1}^n K_i^{-1}\sum_{j=1}^{K_i}|F(T_{i,j})-F'(T_{i,j})|=\int|F-F'|\,\mathrm{d}\upsilon\leq\left(\int|F-F'|^2\,\mathrm{d}\upsilon\right)^{1/2}\end{aligned}$$

where the measure υ is defined by $\upsilon(\cdot)=n^{-1}\sum_{i=1}^n K_i^{-1}\sum_{j=1}^{K_i}\delta_{T_{i,j}}(\cdot)$. In the final step, we use Hölder's inequality and that υ has total mass 1. Further, using Lemma 2.1 and equation (2.5) in van de Geer (2000) we obtain

$$\log\mathcal{N}(u,\mathcal{F},\rho_n)\leq\log\mathcal{N}(u,\mathcal{F},L_2(\upsilon))\leq Cu^{-1}$$

for some constant C and any $u>0$.

Therefore, denote $B_\epsilon=\{F\in\mathcal{F}:\int|F-F_0|\,\mathrm{d}\mu>\epsilon\}$, by $\mathbb{P}_{(K,T)}(A_n)\to 0$ as $n\to\infty$ and inequality (3.10), we have

$$\begin{aligned}\mathbb{E}_0\Pi(B_\epsilon|\mathcal{D}_n)&=\mathbb{E}_0\Pi(B_\epsilon|\mathcal{D}_n)\mathbf{1}_{A_n}+\mathbb{E}_0\Pi(B_\epsilon|\mathcal{D}_n)\mathbf{1}_{A_n^c}\\&\leq\mathbb{E}_0(\mathbf{1}_{A_n})+\mathbb{E}_0\Pi(F\in\mathcal{F}\colon d_n'(F,F_0)>\epsilon/2|\mathcal{D}_n)\\&\leq\mathbb{P}_{(K,T)}(A_n)+\mathbb{E}_0\Pi(F\in\mathcal{F}\colon d(F,F_0)>\epsilon/2|\mathcal{D}_n)\to 0\end{aligned}$$

as $n\to\infty$. □

3.4. Computational methods

Assume the mixing measure G is a Dirichlet process with base measure G_0 (with density g_0) and concentration rate α. The prior distribution this induces on $\mathcal{F}$ through (3.4) is called a Dirichlet Mixture Process (DMP). Denoting by $\#(x)$ the number of distinct values in a vector x, a sample $X_1,\ldots,X_n$ from the DMP can be generated using the following steps:

$$\begin{aligned}Z:=(Z_1,\ldots,Z_n)&\sim\mathrm{CRP}(\alpha)\\\Theta_1,\ldots,\Theta_{\#(Z)}&\overset{iid}{\sim}G_0\\X_1,\ldots,X_n\mid\Theta_1,\ldots,\Theta_{\#(Z)},Z_1,\ldots,Z_n&\overset{ind}{\sim}\mathrm{Unif}(0,\Theta_{Z_i}).\end{aligned}\tag{3.15}$$

Here CRP(α) denotes the "Chinese Restaurant Process" that can be viewed as follows. Assume in a Chinese restaurant, the first customer sits at the first table. Then, given a number of occupied tables, the next customer joins one of these tables with a probability proportional to the number of customers already there, or starts a new table with probability proportional to α. Interpreting Z_i as the number of customers sitting at table i after n customer arrivals, this leads to a distribution on the space of partitions of the integers $\{1, 2, \ldots, n\}$.

In the interval censoring model, we do not observe the X_i's, but for each i the interval $(L_i, R_i] = (T_{i,J_i-1}, T_{i,J_i}]$ that contains X_i. We are then interested in the conditional distribution of (Z, Θ) given the data $\mathcal{D}_n$. In case we would have complete observations $X_1, \ldots, X_n$, there are algorithms to sample from this conditional distribution (see Neal (2000)). Having only the interval censored data, we can adapt such algorithms, treating the unobserved event times X_i as latent variables in the same fashion as this is done in the case of right censoring by Hansen & Lauritzen (2002). Given the exact values X_i, we can use existing algorithms to generate samples from the posterior. Subsequently, we update the X_i's in each iteration by sampling conditionally on the time intervals $(L_i, R_i]$ where the event happened.

We initialise a Gibbs sampler by specifying values of (Z, Θ, X) that satisfy the constraints in the model. This means that $\Theta_{Z_i} \geq X_i$ and $X_i \in (L_i, R_i]$ for $i = 1, \ldots, n$. For ease of notation let $\Theta = (\Theta_1, \ldots, \Theta_{\#(Z)})$ and $X = (X_1, \ldots, X_n)$ for $i = 1, \ldots, n$. Then the following steps are iterated:

1. sample $Z \mid (X, \Theta, \mathcal{D}_n)$;

2. sample $\Theta \mid (X, Z, \mathcal{D}_n)$;

3. sample $X \mid (\mathcal{D}_n, \Theta, Z)$.

Given X, $\mathcal{D}_n$ does not play any role when sampling Z and Θ. Hence the first two steps are the same as in the case of precise observations. More details on this step in that setting can be found in section 2.3 and Neal (2000). The final step is to sample the latent variables X given $\mathcal{D}_n$, Z and Θ. For this, note that

$$f_{X_i|\mathcal{D}_n,\Theta,Z}(x|\mathcal{D}_n, \theta, z) \propto f(x|\theta_{z_i})\mathbf{1}_{(L_i,R_i]}(x) = \varphi(x|\theta_{z_i})\mathbf{1}_{(L_i,R_i]}(x).$$

This is the density of the uniform distribution on interval $(L_i, R_i] \cap [0, \Theta_{Z_i}]$. Note that with the initialisation described above, $(L_i, R_i] \cap [0, \Theta_{Z_i}]$ is non-empty.

Using the conjugacy property of Dirichlet process (see e.g. Ferguson (1973)), the conditional expectation of the posterior of F is given by

$$\mathbb{E}\left[\int \Psi(x,\theta)dG(\theta)\,|\,\Theta, Z, \mathcal{D}_n\right] = \frac{1}{\alpha+n}\left(\alpha\int \Psi(x,\theta)dG_0(\theta) + \sum_{i=1}^{n}\Psi(x,\Theta_{Z_i})\right).$$

Hence, the posterior mean of F can be obtained using a Markov Chain Monte Carlo approximation of the posterior of (Θ, Z) given $\mathcal{D}_n$. Having the algorithms to generate from the distribution of $(\Theta, Z)|\mathcal{D}_n$, assume in the j-th iteration we obtained $\left(\Theta^{(j)}_{Z_1^{(j)}}, \ldots, \Theta^{(j)}_{Z_n^{(j)}}\right)$. At iteration j, a sample from the posterior is given by

$$\hat{F}^{(j)}(x) := \frac{\alpha}{\alpha+n}\int \Psi(x,\theta)dG_0(\theta) + \frac{1}{\alpha+n}\sum_{i=1}^{n}\Psi(x,(\Theta^{(j)}_{Z_i^{(j)}})). \tag{3.16}$$

After J iterations, an estimator for the posterior mean is given by $J^{-1}\sum_{j=1}^{J}\hat{F}^{(j)}(x)$.

Remark 3.4.1. In case the Dirichlet process is truncated, the target density is of fixed dimension. One of the referees raised the question whether probabilistic programming languages such as JAGS, BUGS, Stan or Turing can be used. First of all, we do not consider truncation here as, strictly speaking, it is not necessary. However, we fully agree that from a practical point of view the proposed approach may be implemented using one of the suggested Bayesian computational packages in case of truncation. What might be tricky here is that the workehorse algorithm in for example Stan (Hamiltonian Monte Carlo) uses automatic differentiation for computing gradients. However the density of the uniform distribution on $[0,\theta]$, viewed as a function of θ is not differentiable.

A host of related Bayesian nonparametric models have been implemented in the DP-package (Cf. Jara et al. (2011)).

3.5. Simulation results

In this section, we first study the posterior mean estimators of a concave distribution function based on simulated interval censored data. Next, we compare the Bayesian and the frequentist methods in this setting.

We simulate data by repeating independently n times the following scheme:

1. sample K from the discrete uniform distribution on the integers $\{1,\cdots,20\}$;
2. sample K inspection times $T_1 < \cdots < T_K$ by sorting K independent and identically distributed random variables (we choose the Gamma distribution with shape parameter equal to 2 and rate parameter equal to 1);

3. sample X from the standard Exponential distribution;

4. set $L := \sup_j\{T_j : T_j < X\}$ and $R := \inf_j\{T_j : T_j \geq X\}$ (where $T_0 = 0$, $T_{k+1} = \infty$).

This leads to the dataset $\mathcal{D}_n$ containing the observation intervals $(L_i, R_i]$ for $1 \leq i \leq n$.

The prior is specified by a Dirichtlet Process for the mixture measure. As seen in the formula (3.16), the concentration parameter α expresses our confidence on the prior. In the following, we take a "small" value $\alpha = 1$. Write $Y \sim Par(s, \xi)$ with $s > 0$ and $\xi > 0$ if $f_Y(y) = \xi s^\xi y^{-\xi-1}\mathbf{1}\{y \geq s\}$. We choose the base measure to be a mixture of $Par(s, \xi)$ distributions, where $\xi = 1$ is fixed and the parameter s is drawn from the $Gamma(2, 1)$-distribution. This hierarchical specification leads to partial conjugacy in the Gibbs sampler with one extra step in which S is updated. For both algorithms, S can be sampled from

$$\begin{aligned} f_{S|\Theta,Z}(s|\theta, z) &\propto f_{\Theta|S,Z}(\theta|s, z) f_S(s) \\ &= f_S(s) \prod_{i=1}^{\#(z)} f_{\Theta_i|S}(\theta_i|s) \propto f_S(s) s^{\xi \#(z)} \mathbf{1}\{s \leq \wedge(\theta)\}, \end{aligned}$$

the product being taken over all (distinct) values in the vector θ.

We take sample size $n = 100$. To show the algorithm's performance, we show a traceplot and autocorrelation function of $\hat{F}^j(1)$ over 30.000 iterations in Figure 3.1.

We compute the posterior mean estimator for the function F_0 using equation (3.16) for two samples from the standard exponential distribution: one with sample size 50 and the other with sample size 500. Figure 3.2 shows the results. The total number of MCMC iterations was chosen to be $30,000$, with $15,000$ burn-in iterations.

We now compare different estimation methods:

- the posterior mean for a concave distribution function;
- the maximum likelihood estimator under concavity;
- the maximum likelihood estimator without shape constraints.

We took $n = 500$ and considered $K_i = 1$, $K_i = 2$ for $i = 1, \ldots, n$ (interval censoring case 1 and 2) and K_i independently sampled from the discrete uniform distribution on the integers $\{1, 2, \ldots, 20\}$, which we denote by $K \sim \text{Unif}(1, 20)$.

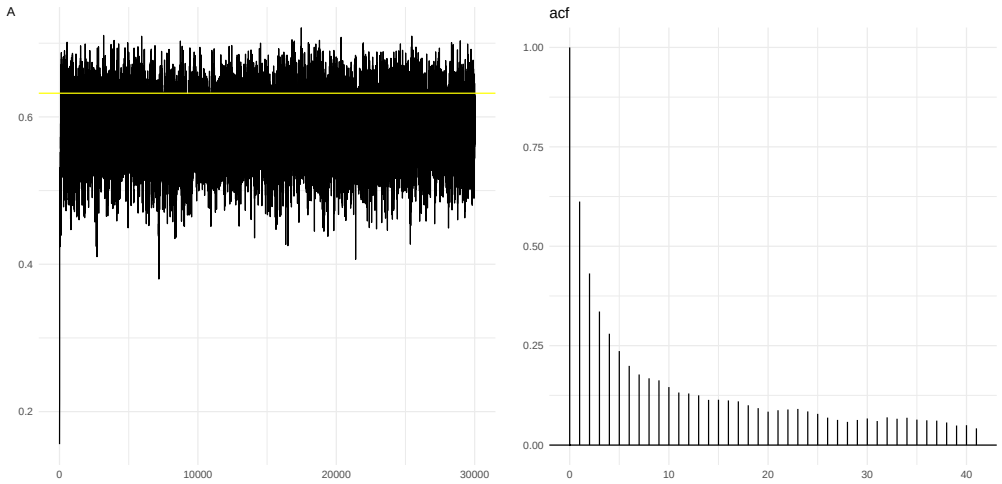

Figure 3.1.: Traceplot (left) and autocorrelation plot (right) for the posterior distribution function evaluated at 1 using the algorithm detailed in Section 3.4. The horizontal line in the left-hand figure depicts the true value $F_0(1) = 1 - e^{-1}$.

We use the same prior specification as before. Figure 3.3 depicts the estimators $\hat{F}$ (here we have three estimators: the NPMLE using the algorithm in Wellner & Zhan (1997), the concave MLE studied in Dümbgen, Freitag & Jongbloed (2006) and the Bayesian posterior mean estimator) and error curves $\hat{F} - F_0$, where F_0 is the true underlying distribution function. As the true distribution is smooth it is not surprising that NPMLE performs worst, as it is a step function. With an increasing number of inspection times, the procedure of generating the inspection time and event time gives a narrow inspection interval for each event. Although the NPMLE does not consider the concavity assumption on F_0, it suggests a concave shape. As can be seen in all cases, the concave MLE and the posterior mean estimator behave similarly.

Using the setting of mixed interval censoring ($K \sim \text{Unif}(1, 20)$), we generated 50 data sets of sizes $n = 50, 100, 200, 400, 800$ from the standard exponential and half-normal distribution and computed the NPMLE, the concave MLE, the posterior mean for each of the cases. Fix grid points $t_j = j/100, j = 1, \ldots, m$, where we took $m = 800$. Figures 3.4 and 3.5 show the log of the mean square error of $\hat{F}$ evaluated at $t = t_j, j = 1, \ldots, m$ for each sample size n, that is

$$\log R(\hat{F}, F)(t) = \log \frac{1}{50} \sum_{k=1}^{50} (\hat{F}^{(k)}(t) - F(t))^2$$

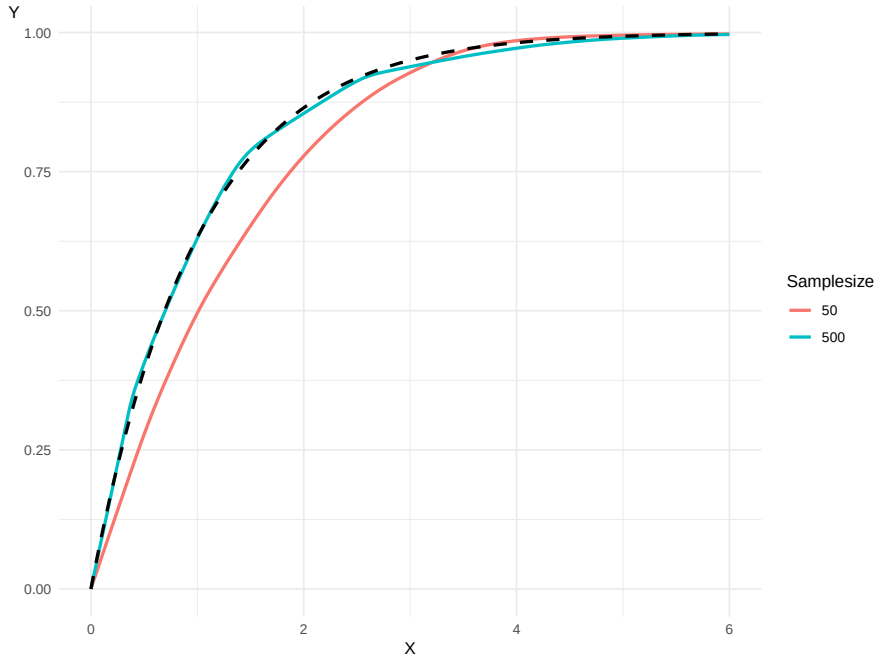

Figure 3.2.: Posterior mean in case the data are sampled from the standard exponential distribution. Two solid lines depict the posterior mean and the dashed line is true distribution function (standard exponential). The concentration parameter is $\alpha = 1$ and the base measure is a mixture of Pareto distributions. The total number of MCMC iterations was chosen to be 30.000, with 15.000 burn-in iterations.

where $\hat{F}^{(k)}$ represent estimator based on the $k-$th data set. We see that all three estimators give small error. As seen from figure 3.3, it can be explained by the setting of how to generate mixed interval censoring data. We see that the posterior mean gives smallest error when t is small, whereas all three estimators are comparable when $t \in [1, 4]$ of case $n = 800$. Finally, the NPMLE performs best when t is big based on the data sets sample from the half-normal distribution.

We also consider a global value, the integrated square errors:

$$\mathrm{ISE}^{(k)}(\hat{F}, F) = \frac{1}{m}\sum_{j=1}^{m}(\hat{F}^{(k)}(t_j) - F(t_j))^2$$

for each sample size n, where $\hat{F}^{(k)}$ represent estimator based on the $k-$th data set, $k = 1, \ldots, 50$. Figure 3.6 shows the mean of integrated square errors. In most of the cases, we see that the concave MLE has the smallest mean integrated square

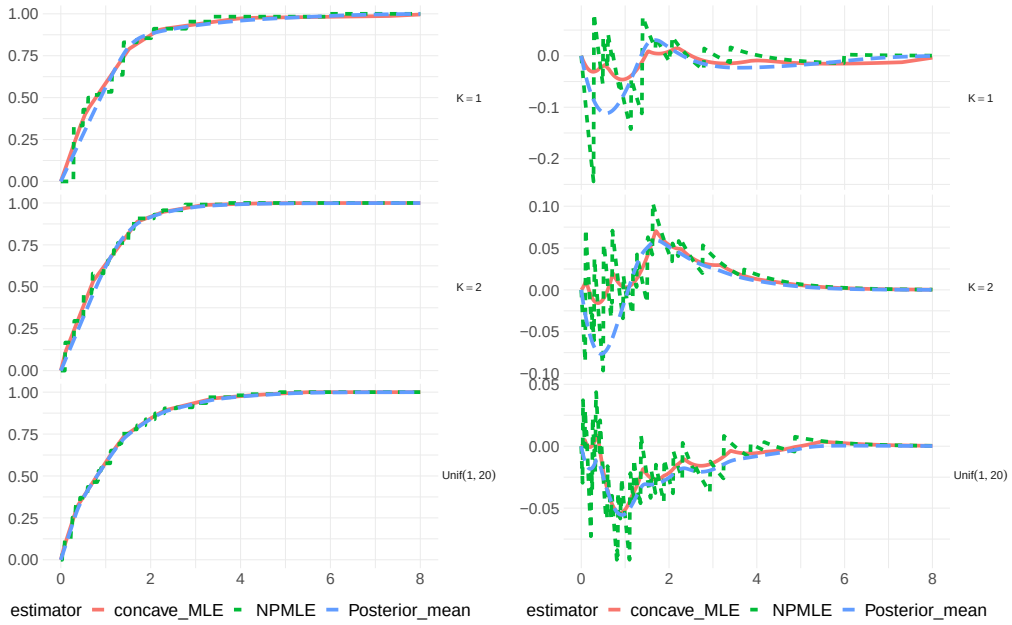

Figure 3.3.: Three cumulative distribution function estimators of F_0 (left) and error curves (right) $\hat{F} - F_0$, $\hat{F}$: posterior mean, NPMLE and concave MLE in case $n = 500$ sample size data are sampled from the standard exponential distribution. From top to bottom corresponding different inspection time $K = 1, 2$ and $K \sim \text{Unif}(1, 20)$.

error, The posterior mean laying between NPMLE and the concave MLE and close to the concave MLE in case of half-normal distribution.

3.6. Case study

In this section we illustrate the applicability of our method in real data examples. Using a nonparametric frequentist approach, producing confidence bands for the underlying distribution usually needs quite some fine tuning (see e.g. Groeneboom & Jongbloed (2014)). Contrary to the frequentist approach, within the Bayesian approach it is simple to construct pointwise credible regions from MCMC output. We applied the Bayesian approach and two frequentist estimators to the Rubella data and Breast cancer data sets.

Example 3.6.1. Rubella is a highly contagious childhood disease. The Rubella data concerns the prevalence of rubella in $n = 230$ Austrian males (see for more information Keiding et al. (1996)). The male individuals included in the data set

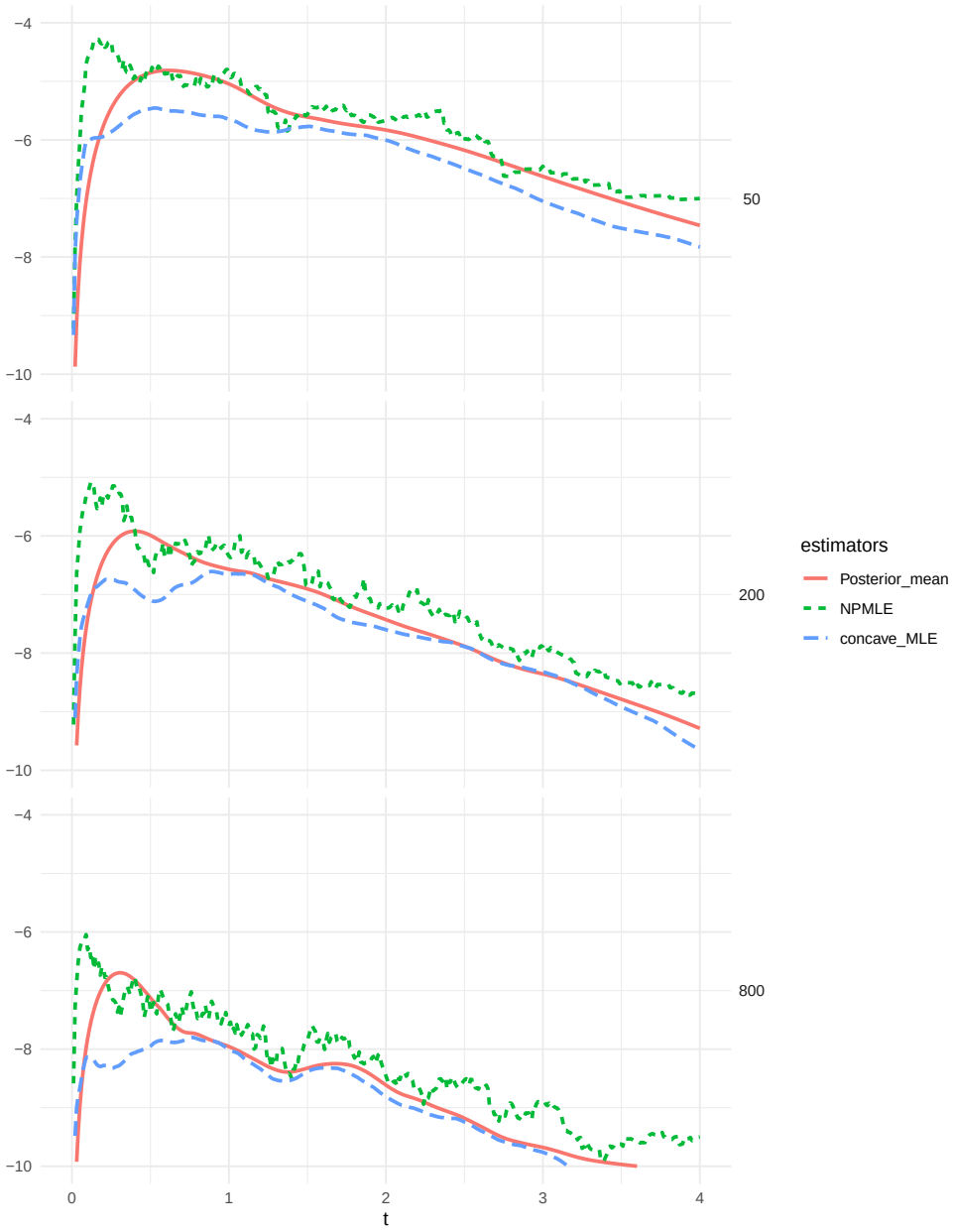

Figure 3.4.: The log mean square error $(\log R(\hat{F}, F))$ evaluated at grid points $\{0.01, 0.02, \ldots, 8.0\}$ for NPMLE, the concave MLE and posterior mean in 50 data sets of different sample size $n = 50, 200, 800$ case sampled from the standard exponential distribution.

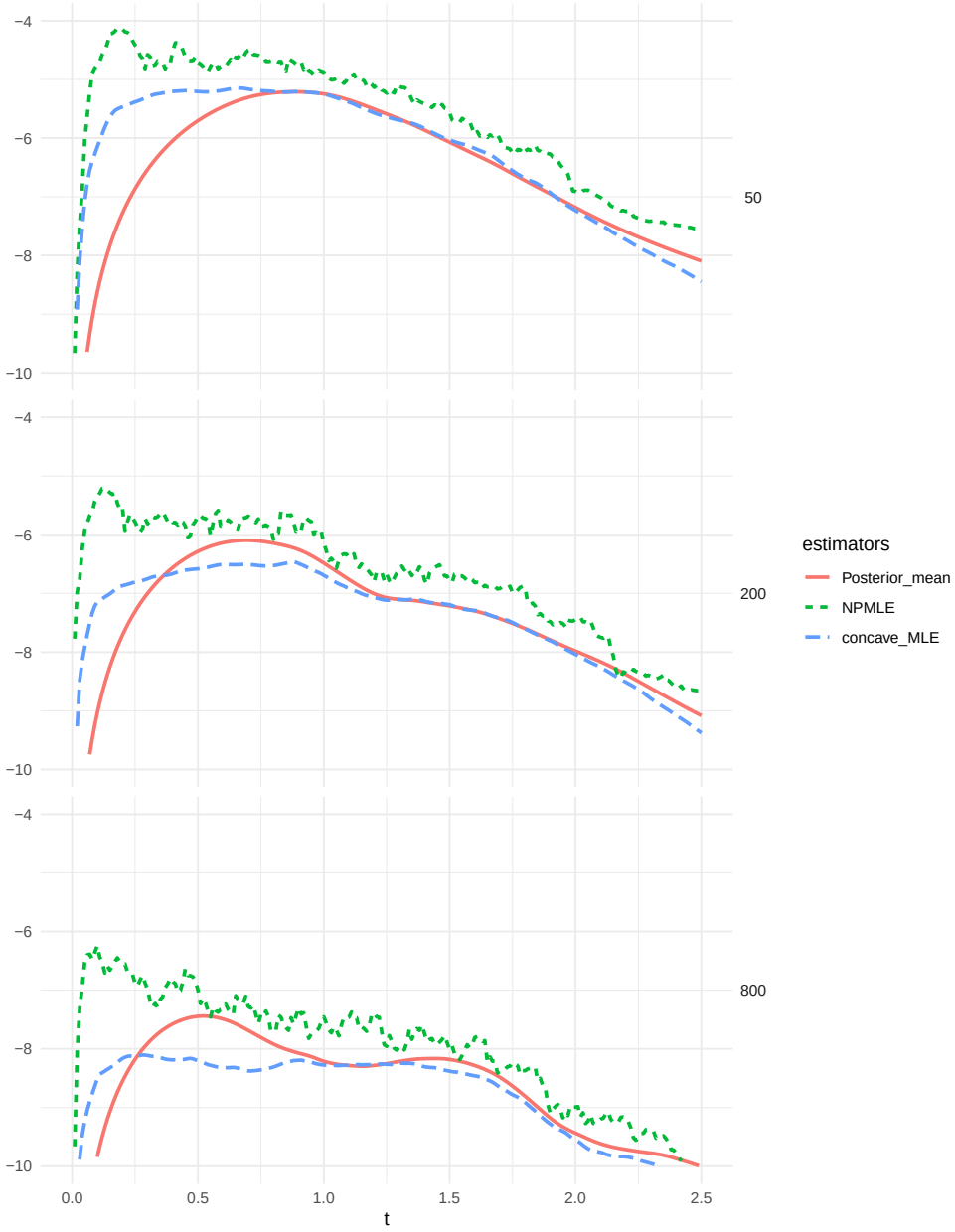

Figure 3.5.: The log mean square error ($\log R(\hat{F}, F)$) evaluated at grid points $\{0.01, 0.02, \ldots, 8.0\}$ for NPMLE, the concave MLE and posterior mean in 50 data sets of different sample size $n = 50, 200, 800$ case sampled from the standard halfnormal distribution.

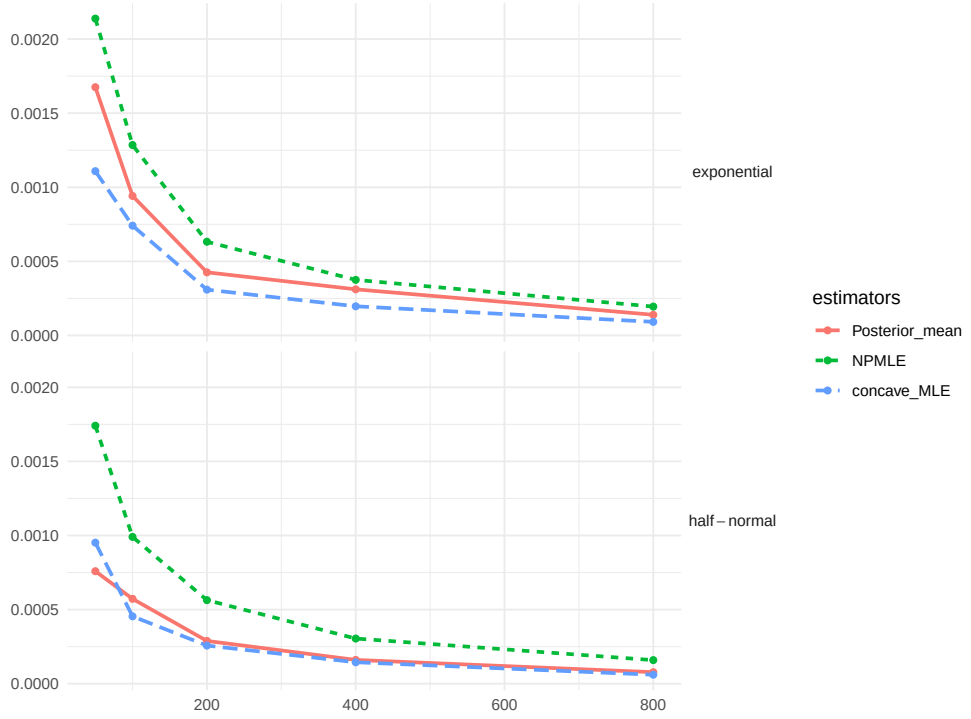

Figure 3.6.: Mean of $ISE^{(k)}(\hat{F}, F)$ for NPMLE, the concave MLE and posterior mean in 50 data sets of different sample size $n = 50, 100, 200, 400, 800$ case sampled from the standard exponential and halfnorm distribution.

represent an unvaccinated population. The data records whether a person got infected or not before a certain time. Here the upper limit of a persons's life span is set equal to 100. Because there is only one inspection time per person, the data are actually case 1 interval censored. Figure 3.7 visualises the data, showing that the time intervals either start at 0 or end at 100.

The settings for computing the posterior mean are as described in the previous section (DP as the prior, with concentration parameter $\alpha = 1$ and the mixture of Pareto as the base measure. The total number of iterations was set to 30.000 where the initial 15.000 iterations have been treated as burn. Figure 3.8 shows the three estimators and 95% pointwise credible sets for the underlying distribution function. The mle (assuming the distribution function to be concave) is comparable with the posterior mean. However, the posterior mean provides a smoother estimator as it is obtained by averaging and not as a maximizer of a likelihood (both the mle and mle under concavity assumption only change slope at censoring times).

Example 3.6.2. In the Breast cancer study discussed in Finkelstein & Wolfe

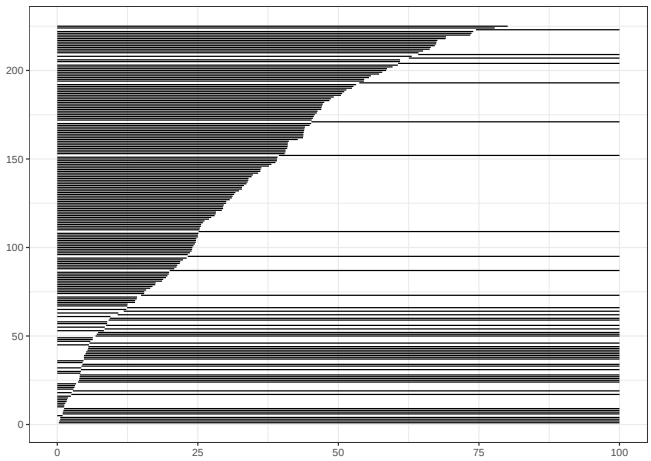

Figure 3.7.: Visualisation of Rubella data. The x-axis is the range of event time. The horizontal lines displays the time intervals.

(1986), 94 early breast cancer patients were given radiation therapy with (RCT, 48) or without (RT, 46) adjuvant chemotherapy between 1976 and 1980. They were supposed to be seen at clinic visits every 4 to 6 months. However, actual visit times differ from patient to patient, and times between visits also vary. In each visit, physicians evaluated the appearance breast retraction. The data contain information about the time to breast retraction, hence, interval censored. Figure 3.9 visualises the data, we use the right end point 100 for the right censoring case.

The settings for computing the posterior mean are as in example 3.6.1. Figure 3.10 shows the three estimators under two treatments (RT and RCT) and 95% credible sets for the underlying survival function.

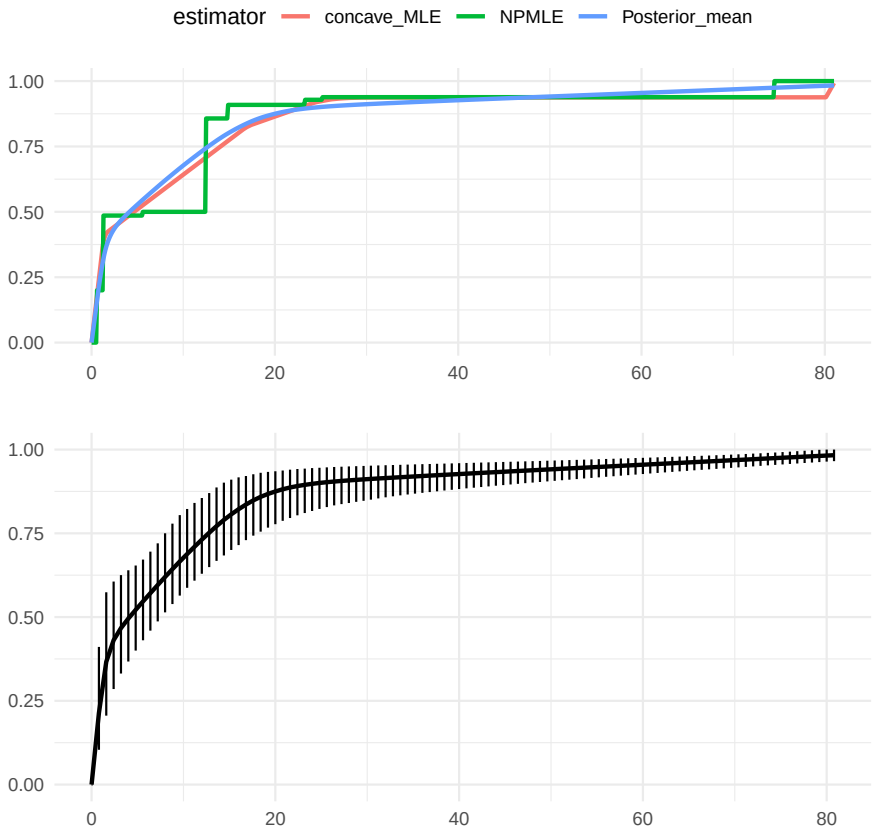

Figure 3.8.: NPMLE, concave MLE and posterior mean estimators (A) and 95% point-wise credible sets of the estimated posterior mean (B) for the underlying distribution function based on the Rubella data.

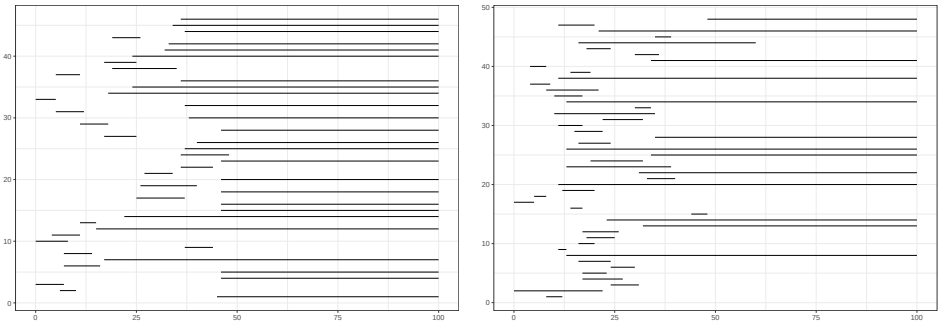

Figure 3.9.: Visualisation of the Breast cancer data (left: RT, right: RCT). The x-axis is the range of event times. The horizontal lines display the time intervals.

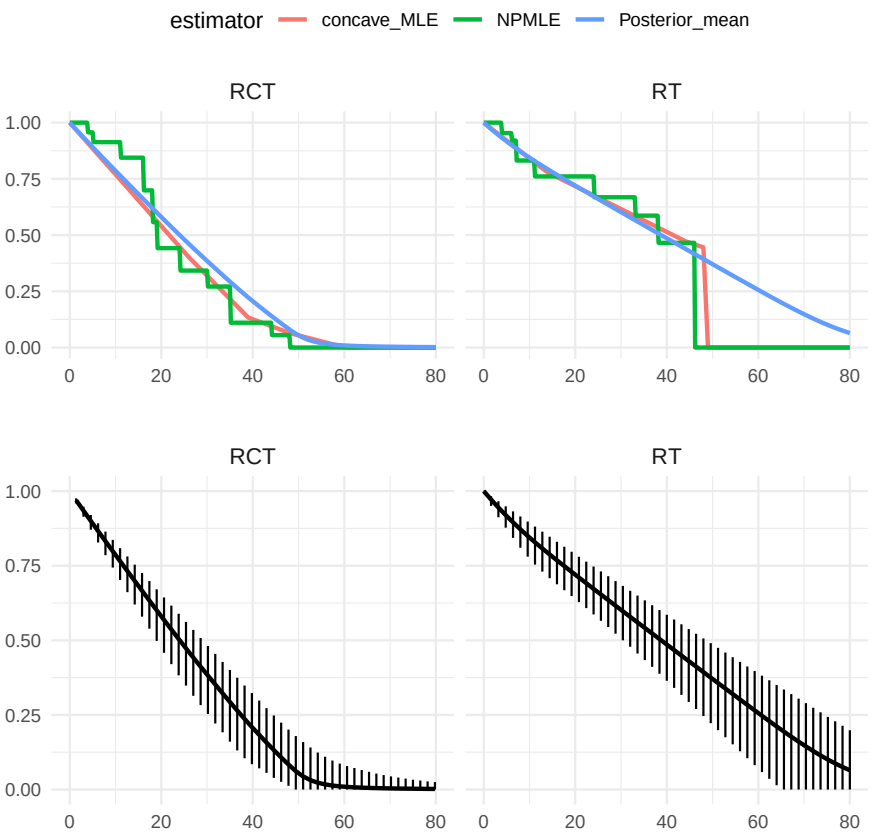

Figure 3.10.: NPMLE, concave MLE and posterior mean estimators (top) and 95% point-wise credible sets of the estimated posterior mean (bottom) for the underlying survival function $1 - F(x)$ ($F(x)$ denotes the distribution function) based on the Breast cancer data.

4. Bayesian nonparametric estimation for current status continuous mark model

This chapter consider the current status continuous mark model where, if the event takes place before an inspection time T a "continuous mark" variable is observed as well. A Bayesian nonparametric method is introduced for estimating the distribution function of the joint distribution of the event time (X) and mark (Y). We consider a prior that is obtained by assigning a distribution on heights of cells, where cells are obtained from a partition of the support of the density of (X, Y). As distribution on cell heights we consider both a Dirichlet prior and a prior based on the graph-Laplacian on the specified partition. Our main result shows that under appropriate conditions, the posterior distribution function contracts pointwisely at rate $(n/\log n)^{-\rho/3(\rho+2)}$, where ρ is the Hölder smoothness of the true density. In addition to the theoretical results, computational methods for drawing from the posterior using probabilistic programming are provided. The performance of the computational methods is illustrated in two examples.

4.1. Introduction

4.1.1. Problem formulation

Survival analysis is concerned with statistical modelling of the time until a particular event occurs. The event may for example be the onset of a disease or failure of equipment. Rather than observing the time of event exactly, censoring is common in practice. If the event time is only observed when it occurs prior to a specific (censoring) time, one speaks of right censoring. In case it is only known whether the event took place before a censoring time or not, one speaks of current status censoring. The resulting data are then called current status data.

In this chapter we consider the current status continuous mark model where, if the event takes place before an inspection time T, a "continuous mark" variable is observed as well. More specifically, denote the event time by X and the mark

by Y. Independent of (X, Y), there is an inspection time T with density function g on $[0, \infty)$. Instead of observing each (X, Y) directly, we observe inspection time T together with the information whether the event occurred before time T or not. If it did so, the additional mark random variable Y is also observed, for which we assume $P(Y = 0) = 0$. Hence, an observation of this experiment can be denoted by $W = (T, Z) = (T, \Delta \cdot Y)$ where $\Delta = \mathbf{1}_{\{X \leq T\}}$ (note that, equivalently, $\Delta = \mathbf{1}_{\{Z>0\}}$). We will assume this experiment is repeated n times independently, leading to the observation set $\mathcal{D}_n = \{W_i, i = 1, \ldots, n\}$. We are interested in estimating the joint distribution function F_0 of (X, Y) nonparametrically, based on $\mathcal{D}_n$.

An application of this model is the HIV vaccine trial studied by Hudgens, Maathuis & Gilbert (2007). Here, the mark is a specifically defined viral distance that is only observed if a participant to the trial got HIV infected before the moment of inspection.

4.1.2. Related literature

In this section we review earlier research efforts on models closely related to that considered here.

Survival analysis with a continuous mark can be viewed as the continuous version of the classical competing risks model. In the latter model, failure is due to either of K competing risks (with K fixed) leading to a mark value that is of categorical type. As the mark variable encodes the cause of failure it is only observed if failure has occurred before inspection. These "cause events" are known as competing risks. Groeneboom, Maathuis & Wellner (2008) study nonparametric estimation for current status data with competing risks. In that paper, they show that the nonparametric maximum likelihood estimator (NPMLE) is consistent and converges globally and locally at rate $n^{1/3}$.

Huang & Louis (1998) consider the continuous mark model under right-censoring, which is more informative compared to the current-status case because the exact event time is observed for noncensored data. For the nonparametric maximum likelihood estimator of the joint distribution function of (X, Y) at a fixed point, asymptotic normality is shown.

Hudgens, Maathuis & Gilbert (2007) consider interval censoring case k, $k = 1$ being the specific setting of current-status data considered here. In this paper the authors show that both the NPMLE and a newly introduced estimator termed "midpoint imputation MLE" are inconsistent. However, coarsening the mark variable (i.e. making it discrete, turning the setting to that of the competing risks model), leads to a consistent NPMLE. This is in agreement with the results in Maathuis & Wellner (2008).

Groeneboom, Jongbloed & Witte (2011) and Groeneboom, Jongbloed & Witte

(2012) consider the exact setting of this paper using frequentist estimation methods. In Groeneboom, Jongbloed & Witte (2011) two plug-in inverse estimators are proposed. They prove that these estimators are consistent and derive the pointwise asymptotic distribution of both estimators. Groeneboom, Jongbloed & Witte (2012) define a nonparametric estimator for the distribution function at a fixed point by finding the maximiser of a smoothed version of the log-likelihood. Pointwise consistency of the estimator is established. In both papers numerical illustrations are included.

4.1.3. Contribution

In this chapter, we consider Bayesian nonparametric estimation of the bivariate distribution function F_0 in the current status continuous mark model. This approach has not been adopted before, neither from a theoretical nor computational perspective (within the Bayesian setting). Whereas consistent nonparametric estimators exist within frequentist inference, convergence rates are unknown. We prove consistency and derive Bayesian contraction rates for the bivariate distribution function of (X, Y) using a prior on the joint density f of (X, Y) that is piecewise constant. For the values on the bins we consider two different prior specifications. Our main result shows that under appropriate conditions, the posterior distribution function contracts pointwisely at rate $(n/\log n)^{-\rho/3(\rho+2)}$, where ρ is the Hölder smoothness of the true density.

The proof is based on general results from Ghosal & Van der Vaart (2017) for obtaining Bayesian contraction rates. Essentially, it requires the derivation of suitable test functions and proving that the prior puts sufficient mass in a neighbourhood of the "true" bivariate distribution. The latter is proved by exploiting the specific structure of our prior. In addition to our theoretical results, we provide computational methods for drawing from the posterior using probabilistic programming in the Turing Language under Julia (see Bezanson et al.(2017), Ge, Xu & Ghahramani (2018)). The performance of our computational methods is illustrated in two examples.

4.1.4. Outline

The outline of this chapter is as follows. In section 4.2 we introduce further notation for the current status continuous mark model and detail the two priors considered. Subsequently, we derive posterior contraction rates under some assumptions on the underlying bivariate distribution in section 4.3. The proof is given in section 4.4. Section 4.5 contains numerical illustrations.

4.1.5. Notation

For two sequences $\{a_n\}$ and $\{b_n\}$ of positive real numbers, the notation $a_n \lesssim b_n$ (or $b_n \gtrsim a_n$) means that there exists a constant $C > 0$, independent of n, such that $a_n \leq C b_n$. We write $a_n \asymp b_n$ if both $a_n \lesssim b_n$ and $a_n \gtrsim b_n$ hold. We denote by F and F_0 the cumulative distribution functions corresponding to the probability densities f and f_0 respectively. The Hellinger distance between two densities f, g is written as $h^2(f,g) = \frac{1}{2}\int (f^{1/2} - g^{1/2})^2$. The Kullback-Leibler divergence of f and g and the L_2-norm of $\log(f/g)$ (under f) by

$$KL(f,g) = \int f \log \frac{f}{g}, \qquad V(f,g) = \int f \left(\log \frac{f}{g}\right)^2.$$

4.2. Likelihood and prior specification

4.2.1. Likelihood

In this section we derive the likelihood for the joint density f based on data $\mathcal{D}_n$. As $W_1, \ldots, W_n$ are independent and identically distributed, it suffices to derive the joint density of $W_1 = (T_1, Z_1)$ (with respect to an appropriate dominating measure). Recall that f denotes the density of (X, Y). Let F denote the corresponding distribution function of (X, Y). The marginal distribution function of X is given by $F_X(t) = \int_0^t \int_0^\infty f(u,v)\,\mathrm{d}v\,\mathrm{d}u$. Define the measure μ on $[0,\infty)^2$ by

$$\mu(B) = \mu_2(B) + \mu_1(\{x \in [0,\infty) : (x,0) \in B\}), \quad B \in \mathcal{B}$$

where $\mathcal{B}$ is the Borel $\sigma-$algebra on $[0,\infty)^2$ and μ_i is Lebesgue measure on $\mathbb{R}^i$. The density of the law of W_1 with respect to μ is then given by

$$s_f(t,z) = g(t)\left(\mathbf{1}_{\{z>0\}}\partial_2 F(t,z) + \mathbf{1}_{\{z=0\}}(1 - F_X(t))\right), \tag{4.1}$$

where $\partial_2 F(t,z) = \frac{\partial}{\partial z} F(t,z) = \int_0^t f(u,z)\,\mathrm{d}u$. By independence the likelihood of f based on $\mathcal{D}_n$ is given by $l(f) = \prod_{i=1}^n s_f(T_i, Z_i)$.

4.2.2. Prior

In this section, we define a prior on the class of all bivariate density functions on $\mathbb{R}^2$, denote as

$$\mathcal{F} = \left\{ f : \mathbb{R}^2 \to [0,\infty) : \int_{\mathbb{R}^2} f(x,y)\,\mathrm{d}x\,\mathrm{d}y = 1 \right\}.$$

For any $f \in \mathcal{F}$, if S denotes the support of f and $\cup_j C_j, j = 1, \ldots, p_n$ is a partition of S, we define a prior on $\mathcal{F}$ by constructing

$$f_{\boldsymbol{\theta}}(x, y) = \sum_j \frac{\theta_j}{|C_j|} \mathbf{1}_{C_j}(x, y), \qquad (x, y) \in \mathbb{R}^2,$$

where $|C| = \mu_2(C)$ is the Lebesgue measure of the set C. Let $\boldsymbol{\theta}$ denote the vector $\boldsymbol{\theta} = (\theta_1, , \ldots, \theta_{p_n})$. We require that all θ_j are nonnegative and that $\boldsymbol{\theta}$ satisfies $\sum_j \theta_j = 1$. We consider two types of prior on $\boldsymbol{\theta}$.

1. *D*irichlet. For a fixed parameter $\alpha = (\alpha_1, \ldots, \alpha_{p_n})$ consider $\boldsymbol{\theta} \sim \text{Dirichlet}(\alpha)$. This prior is attractive as draws from the posterior distribution can be obtained using a straightforward data-augmentation algorithm (Cf. Section 4.5).

2. *N*ormal with graph Laplacian covariance matrix. For a positive-definite matrix Υ, let $\boldsymbol{H} \sim N_{p_n}(0, \tau^{-1}\Upsilon^{-1})$, conditionally on τ. Each element of $\boldsymbol{H}$ corresponds to one value of $\boldsymbol{\theta}$. Next, set

$$\theta_j = \frac{\psi(H_j)}{\sum_j \psi(H_j)}, \quad \text{where} \quad \psi(x) = e^x/(1 + e^x). \tag{4.2}$$

   The matrix Υ is chosen as follows. The partition of S induces a graph structure on the bins, where each bin corresponds to a node in the graph, and nodes are connected when bins are adjacent (meaning that they are either horizontal or vertical "neighbours"). Let L denote the graph Laplacian of the graph obtained in this way. This is the $p_n \times p_n$ matrix given by

$$L_{i,i'} = \begin{cases} \text{degree node } i & \text{if } i = i' \\ -1 & \text{if } i \neq i' \text{ and nodes } i \text{ and } i' \text{ are connected} \\ 0 & \text{otherwise.} \end{cases} . \tag{4.3}$$

   Now we take

$$\Upsilon = L + p_n^{-2} I.$$

Remark 4.2.1. A property of the Dirichlet prior is that values of θ_j in adjacent bins are a negatively correlated, preventing the density to capture smoothness. See more in numerical study section 4.5. The idea of the graph-Laplacian prior is to induce positive correlation on adjacent bins and thereby specify a prior that produces draws is smoother on the graph corresponding to the partition. As we will see, this comes at the cost of increased computational complexity.

Remark 4.2.2. One can argue whether the presented prior specifications are truly nonparametric. It is not if one adopts as definition that the size of the parameter should be learned by the data. For that, a solution could be to put a prior on p_n as well. While possible, this would severely complicate drawing from the posterior. As an alternative, one can take large values of p_n (so that the model is high-dimensional), and let the data determine the amount of smoothing by incorporating flexibility in the prior. As the Dirichlet prior lacks smoothness properties, fixing large values of p_n will lead to overparametrisation, resulting in high variance estimates (under smoothing). On the contrary, as we will show in the numerical examples, for the graph Laplacian prior, this overparametrisation can be substantially balanced/regularised by equipping the parameter τ with a prior distribution. The idea of histogram type priors with positively correlated adjacent bins has recently been used successfully in other settings as well, see for instance Gugushvili et al. (2018), Gugushvili et al. (2019).

4.3. Posterior contraction

In this section we derive a contraction rate for the posterior distribution of F_0. Denote as $\Pi_n(\cdot \mid \mathcal{D}_n)$ under the prior measure Π_n described in section 4.2.2.

Assumption 4.3.1. The underlying joint density of the event time and mark, f_0, has compact support given by $\mathcal{M} = [0, M_1] \times [0, M_2]$ and is ρ-Hölder continuous on $\mathcal{M}$ ($\rho \in (0, 1]$). That is, there exists a positive constant L such that for any (x_1, y_1) and (x_2, y_2) in $\mathcal{M}$,

$$|f_0(x_1, y_1) - f_0(x_2, y_2)| \leq L \, \|(x_1, y_1) - (x_2, y_2)\|^{\rho}. \tag{4.4}$$

In addition, there exist positive constants $\underline{M}$ and $\overline{M}$ such that

$$\underline{M}(\min(x, y)^{\rho}) \leq f_0(x, y) \leq \overline{M}, \quad \text{for all } (x, y) \in \mathcal{M}. \tag{4.5}$$

Assumption 4.3.2. The censoring density g is bounded away from 0 and infinity on $(0, M_1)$. That is, there exist positive constants and $\overline{K}$ such that $0 < \underline{K} \leq g(t) \leq \overline{K} < \infty$ for all $t \in (0, M_1)$.

Assumption 4.3.3. Conditions for prior:

1. For the Dirichlet prior, parameter $\alpha = (\alpha_1, \ldots, \alpha_{p_n})$ satisfies $a p_n^{-1} \leq \alpha_l \leq 1$ for all $l = 1, \ldots, p_n$ and some constant $a \in \mathbb{R}^+$.

2. For the graph-Laplacian prior, the prior specification is completed by specifying a prior distribution for τ supported on the positive halfline. For computational convenience, we assign the Gamma(β, γ) distribution prior for τ with density function $f_\tau(\tau) \propto \tau^{\beta-1}e^{-\gamma\tau}$ which is a conjugate prior of normal distribution.

Theorem 4.3.4. *Fix $(x, y) \in [0, M_1] \times (0, M_2]$. Consider either of the priors defined in section 4.2.2 and hyper-parameters satisfy assumption 4.3.3. Define $\eta_n = (n/\log n)^{-\frac{\rho}{3(\rho+2)}}$ where ρ denotes the Hölder parameter in assumption 4.3.1. If f_0 and g satisfy assumptions 4.3.1 and 4.3.2 respectively, then for sufficiently large C*

$$\mathbb{E}_0\Pi_n(f \in \mathcal{F}\colon |F(x, y) - F_0(x, y)| > C\eta_n \mid \mathcal{D}_n) \to 0, \quad as \quad n \to \infty.$$

Before we give a proof of theorem 4.3.4, we state two lemmas which are sufficient to give the contraction rate in the theorem. Define $\varepsilon_n \asymp (n/\log n)^{-\frac{\rho}{2(\rho+2)}}$ (more specifically (C.9)), note that $\varepsilon_n \leq \eta_n$ and $n\varepsilon_n^2 \to \infty$ as $n \to \infty$.

Lemma 4.3.5. *Fix f_0 and g satisfying the conditions in assumption 4.3.1 and 4.3.2. Define*

$$S_n = \left\{f \in \mathcal{F}\colon KL(s_{f_0}, s_f) \leq \varepsilon_n^2, V(s_{f_0}, s_f) \leq \varepsilon_n^2\right\}. \tag{4.6}$$

Then we have $\Pi_n(S_n) \geq e^{-cn\varepsilon_n^2}$ for some constant $c > 0$.

Lemma 4.3.6. *Fix $(x, y) \in [0, M_1] \times (0, M_2]$. Define $U_n(x, y) := \{f \in \mathcal{F}\colon |F(x, y) - F_0(x, y)| > C\eta_n\}$. There exists a sequence of test functions Φ_n such that*

$$\begin{aligned} \mathbb{E}_0(\Phi_n) &= o(1), \\ \sup_{f \in U_n(t,z)} \mathbb{E}_f(1 - \Phi_n) &\leq c_1 e^{-c_2 C^2 n\varepsilon_n^2}, \end{aligned} \tag{4.7}$$

for some positive constants c_1, c_2 and C appeared in theorem 4.3.4.

Proof of Theorem 4.3.4. The proof follows from the general idea in Ghosal, Ghosh & Van der Vaart (2000). Fix $(x, y) \in [0, M_1] \times (0, M_2]$, define $U_n(x, y) := \{f \in \mathcal{F}\colon |F(x, y) - F_0(x, y)| > C\eta_n\}$. Write the posterior mass on the set $U_n(x, y)$ as

$$\Pi_n(U_n(x, y) \mid \mathcal{D}_n) = D_n^{-1} \int_U \prod_{i=1}^n \frac{s_f(W_i)}{s_{f_0}(W_i)} \, \mathrm{d}\Pi_n(f),$$

where

$$D_n = \int \prod_{i=1}^n \frac{s_f(W_i)}{s_{f_0}(W_i)} \, \mathrm{d}\Pi_n(f).$$

Lemma 4.3.5 implies that (see lemma 8.1 in Ghosal, Ghosh & Van der Vaart (2000))

$$\mathbb{P}_0(D_n \leq \exp(-(c+1)n\varepsilon_n^2)) \to 0, \quad \text{as } n \to \infty.$$

Then we can only consider on event $\{D_n \geq \exp(-(c+1)n\varepsilon_n^2)\}$. Using the test sequence in lemma 4.3.6 the posterior mass of $U_n(x,y)$ satisfies

$$\begin{aligned}
\mathbb{E}_0\Pi_n(U_n(x,y) \mid \mathcal{D}_n) &= \mathbb{E}_0\Pi_n(U_n(x,y) \mid \mathcal{D}_n)\Phi_n + \mathbb{E}_0\Pi_n(U_n(x,y) \mid \mathcal{D}_n)(1-\Phi_n) \\
&\leq \mathbb{E}_0\Phi_n + e^{(c+1)n\varepsilon_n^2}\mathbb{E}_0 \int_{U_n(x,y)} \prod_{i=1}^{n} \frac{s_f(W_i)}{s_{f_0}(W_i)}(1-\Phi_n)\,\mathrm{d}\Pi_n(f) \\
&= \mathbb{E}_0\Phi_n + e^{(c+1)n\varepsilon_n^2} \int_{U_n(x,y)} \mathbb{E}_f(1-\Phi_n)\,\mathrm{d}\Pi_n(f) \\
&\leq o(1) + c_1 e^{(c+1)n\varepsilon_n^2} e^{-c_2 C^2 n\varepsilon_n^2} \to 0.
\end{aligned}$$

The final step follows by taking C (appeared in theorem 4.3.4) large enough such that $c_2C^2 > c+1$.

□

4.4. Proof of Lemmas

4.4.1. Proof of lemma 4.3.5

Proof. To give a lower bound for $\Pi_n(S_n)$, we construct a subset Ω_n of S_n and derive a lower bound of $\Pi_n(\Omega_n)$ for both priors considered in section 4.2.2.

We first give a sequence of approximations for f_0. Let $\delta_n = (n/\log n)^{-\frac{1}{\rho+2}}$. Denote $A_{n,j} = ((j-1)\delta_n, j\delta_n]$, $B_{n,k} = ((k-1)\delta_n, k\delta_n]$ for $j = 1, 2, \ldots, J_n - 1, k = 1, 2, \ldots, K_n - 1$ and $A_{n,J_n} = ((J_n-1)\delta_n, M_1], B_{n,K_n} = ((K_n-1)\delta_n, M_2]$, $J_n = \lfloor M_1\delta_n^{-1} \rfloor, K_n = \lfloor M_2\delta_n^{-1} \rfloor$. Then $\cup_{j,k}(A_{n,j} \times B_{n,k})$ is a regular partition on $\mathcal{M}$. Let $f_{0,n}$ be the piecewise constant density function defined by

$$f_{0,n}(t,z) = \sum_{j=1}^{J_n}\sum_{k=1}^{K_n} \frac{w_{0,j,k}}{|A_{n,j} \times B_{n,k}|} 1_{A_{n,j}\times B_{n,k}}(t,z), \tag{4.8}$$

where $w_{0,j,k} = \int_{A_{n,j}}\int_{B_{n,k}} f_0(u,v)\,\mathrm{d}v\,\mathrm{d}u$. That is, we approximate f_0 by averaging it on each bin. Note that $f_{0,n}$ has support $\mathcal{M}$. Define the set

$$\Omega_n := \left\{ f \in \mathcal{F} : ||f - f_{0,n}||_\infty \leq \frac{1}{6}\underline{M}\delta_n^\rho,\ \mathrm{supp}(f) \supseteq \mathcal{M} \right\}. \tag{4.9}$$

By Lemma C.1.1 in appendix, we know that $\Omega_n \subseteq S_n$. Now we give a lower bound for $\Pi_n(\Omega_n)$, for the two type of priors.

Let $p_n = J_n K_n$ denote the total number of bins. According to the prior specifications in section 4.2.2, for any $f \in \mathcal{F}$, we parameterize

$$f_{\boldsymbol{\theta}}(x,y) = \sum_{j,k} \frac{\theta_{j,k}}{|A_{n,j} \times B_{n,k}|} \mathbf{1}_{A_{n,j} \times B_{n,k}}(x,y), \qquad (x,y) \in \mathbb{R}^2,$$

where $\boldsymbol{\theta}$ denotes the vector obtained by stacking all coefficients $\{\theta_{j,k}, j = 1, \ldots, J_n, k = 1, \ldots, K_n\}$. Recall that $f_{0,n}$ is defined by the local averages $\{w_{0,j,k}, j, k \geq 1\}$. For any $(t,z) \in A_{n,j} \times B_{n,k}$, $j, k \geq 1$, we have

$$|f_{\boldsymbol{\theta}}(t,z) - f_{0,n}(t,z)| = |A_{n,j} \times B_{n,k}|^{-1} |\theta_{j,k} - w_{0,j,k}| \leq \delta_n^{-2} \max_{j,k} |\theta_{j,k} - w_{0,j,k}|.$$

In the second step we use $|A_{n,j} \times B_{n,k}| \geq \delta_n^2$ for all j, k. Hence

$$\left\{ f_{\boldsymbol{\theta}} \in \mathcal{F} : \max_{j,k} |\theta_{j,k} - w_{0,j,k}| \leq \frac{1}{6} \underline{M} \delta_n^{\rho+2} \right\} \subseteq \Omega_n. \tag{4.10}$$

Consider the two type of priors defined in section 4.2.2.

• Endowing prior $\boldsymbol{\theta} \sim \text{Dirichlet}(\alpha)$, fixed $\alpha = (\alpha_1, \ldots, \alpha_{p_n})$, $a p_n^{-1} \leq \alpha_l \leq 1$ for all $l = 1, \ldots, p_n$.

By Lemma 6.1 in Ghosal, Ghosh & Van der Vaart (2000), we have

$$\begin{aligned}
\Pi_n(\Omega_n) &\geq \Pi_n \left(\max_{j,k} |\theta_{j,k} - w_{0,j,k}| \leq \frac{1}{6} \underline{M} \delta_n^{\rho+2} \right) \\
&\geq \Gamma\left(\sum_{l=1}^{p_n} \alpha_l \right) \left(\frac{1}{6} \underline{M} \delta_n^{\rho+2} \right)^{p_n} \prod_{l=1}^{p_n} \alpha_l \\
&\geq \exp\left(\log \Gamma(a) + p_n \log\left(\frac{1}{6} \underline{M} \delta_n^{\rho+2} \right) + p_n \log(a p_n^{-1}) \right) \\
&\gtrsim \exp(-C_1 \delta_n^{-2} \log n) = \exp(-cn\varepsilon_n^2)
\end{aligned}$$

for some constant $C_1, c > 0$. This finishes the proof for the Dirichlet prior.

• Let $\theta_{j,k} = \frac{\psi(H_{j,k})}{\sum_{j,k} \psi(H_{j,k})}$ as defined in (4.2) and let $\tau \sim \text{Gamma}(\beta, \gamma)$, $\boldsymbol{H} \mid \tau \sim N_{p_n}(0, \tau^{-1}\Sigma^{-1})$, where

$$\Sigma = L + p_n^{-2} I$$

and each element of the vector $\boldsymbol{H}$ has exactly same order with $\boldsymbol{\theta}$).

For the fixed values $w_{0,j,k}, 1 \leq j \leq J_n, 1 \leq k \leq K_n$, there exists a matrix $\boldsymbol{H}_0$ such that

$$w_{0,j,k} = \frac{\psi(H_{0,j,k})}{\sum_{j,k} \psi(H_{0,j,k})}.$$

For the ease of exposition, we choose $\boldsymbol{H}_0$ such that it satisfies $\sum_{j,k}\psi(H_{0,j,k})=1$, then $w_{0,j,k}=\psi(H_{0,j,k})$ for all j,k.

Denote $\boldsymbol{x}=(x_1,\ldots,x_m)$, $\boldsymbol{x}\in\mathbb{R}^m$ for some $m\in\mathbb{N}$. Define the function $\zeta(\boldsymbol{x})=\frac{\psi(x_1)}{\sum_{j=1}^m\psi(x_j)}$. Using the inequality $\frac{ab}{(a+b)^2}\le\frac{1}{4}$ for $a,b\ge 0$, the partial derivatives of ζ satisfy

$$\left|\frac{\partial\zeta(\boldsymbol{x})}{\partial x_1}\right|=\frac{\psi(x_1)(\sum_{j\ge 2}\psi(x_j))}{(\sum_{j=1}^m\psi(x_j))^2(1+e^{x_1})}\le\frac{1}{4},$$
$$\left|\frac{\partial\zeta(\boldsymbol{x})}{\partial x_l}\right|=\frac{\psi(x_1)\psi(x_l)}{(\sum_{j=1}^m\psi(x_j))^2(1+e^{x_l})}\le\frac{1}{4},\quad l=2,\ldots,m.$$

Then we have for any $\boldsymbol{x}=(x_1,\ldots,x_m)$ and $\boldsymbol{x}_l^0=(x_1,\ldots,x_l^0,\ldots,x_m)\in\mathbb{R}^m$, $l=1,\ldots,m$

$$|\zeta(\boldsymbol{x})-\zeta(\boldsymbol{x}_l^0)|\le\frac{1}{4}|x_l-x_l^0|.$$

Hence for any $\boldsymbol{x},\boldsymbol{x}^0=(x_1^0,\ldots,x_m^0)\in\mathbb{R}^m$,

$$\begin{aligned}|\zeta(\boldsymbol{x})-\zeta(\boldsymbol{x}^0)|&\le|\zeta(\boldsymbol{x})-\zeta(x_1^0,x_2,\ldots,x_m)|+|\zeta(x_1^0,x_2,\ldots,x_m)-\zeta(x_1^0,x_2^0,\ldots,x_m)|\\&\quad+\cdots+|\zeta(x_1^0,\ldots,x_{m-1}^0,x_m)-\zeta(x_1^0,\ldots,x_{m-1}^0,x_m^0)|\\&\le\frac{1}{4}\sum_{l=1}^m|x_l-x_l^0|.\end{aligned}$$

Let $m=p_n$ and $\boldsymbol{x}$ correspond to the vector $\boldsymbol{H}$. Then we have for $j,k\ge 1$,

$$|\theta_{j,k}-w_{0,j,k}|\le\frac{1}{4}\sum_{j,k}|H_{j,k}-H_{0,j,k}|.$$

Combining this with (4.10), we have

$$\Pi_n(\Omega_n)\ge\Pi_n(\{f_{\boldsymbol{H}}\in\mathcal{F}:\boldsymbol{H}\in B_n\}),$$

where

$$B_n=\left\{\boldsymbol{H}:|H_{j,k}-H_{0,j,k}|\le\frac{2}{3}\underline{M}\delta_n^{\rho+2}p_n^{-1},\text{for all}\,j,k\right\}.$$

It is therefore sufficient to give a lower bound for the prior probability on $\{f_{\boldsymbol{H}}:\boldsymbol{H}\in B_n\}$. Note that

$$\begin{aligned}&\Pi_n(\{f_{\boldsymbol{H}}:\boldsymbol{H}\in B_n\})\\&\quad=\frac{\gamma^\beta}{\Gamma(\alpha)}\int_0^\infty\tau^{\beta-1}e^{-\gamma\tau}(2\pi)^{-\frac{p_n}{2}}\tau^{p_n/2}|\Sigma^{-1}|^{1/2}\int_{B_n}\exp\left(-\frac{1}{2}\tau\boldsymbol{H}^T\Sigma\boldsymbol{H}\right)\mathrm{d}\boldsymbol{H}\,\mathrm{d}\tau.\end{aligned}$$

In order to calculate the integral $\int_{B_n} \exp(-\frac{1}{2}\tau \boldsymbol{H}^T \Sigma \boldsymbol{H})\, \mathrm{d}\boldsymbol{H}$ at the right hand side for τ fixed, we first note the following facts. Denote the eigenvalues of Σ by $0 < \lambda_1 < \cdots < \lambda_{p_n}$. Then Σ has the following properties:

$$|\Sigma| = \lambda_1 \cdots \lambda_{p_n} \leq (\lambda_{p_n})^{p_n}, \tag{4.11}$$

$$\mathrm{tr}(\Sigma) = \sum_{l=1}^{p_n} \lambda_l = \sum_{l=1}^{p_n} (L_{l,l} + p_n^{-2}) = p_n^{-1} + \sum_{l=1}^{p_n} L_{l,l}, \tag{4.12}$$

$$\boldsymbol{x}^T \Sigma \boldsymbol{x} \leq \lambda_{p_n} \boldsymbol{x}^T \boldsymbol{x}, \quad \text{for any } p_n \text{ dim vector } \boldsymbol{x}, \tag{4.13}$$

where $|\Sigma|$ denotes the determinant of Σ. By definition of the Laplacian matrix L, (4.3), we know

$$\sum_{l=1}^{p_n} L_{l,l} = 2 \cdot 4 + 3(2(J_n - 2) + 2(K_n - 2)) + 4(J_n - 2)(K_n - 2) < 4p_n,$$

the first term denotes we have 4 nodes of 2 connections (corners), the second item denotes $2(J_n - 2) + 2(K_n - 2)$ of 3 connections (edges) and the final term counts $(J_n - 2)(K_n - 2)$ of full 4 connections (inside). Using (4.12), we know

$$\lambda_{p_n} \leq \sum_{l=1}^{p_n} \lambda_l = p_n^{-1} + \sum_{l=1}^{p_n} L_{l,l} \leq 4p_n + p_n^{-1} \tag{4.14}$$

Using (4.13), we have

$$\int_{B_n} \exp\left(-\frac{1}{2}\tau \boldsymbol{H}^T \Sigma \boldsymbol{H} \right) \mathrm{d}\boldsymbol{H} \geq \int_{B_n} \exp\left(-\frac{1}{2}\tau \lambda_{p_n} \boldsymbol{H}^T \boldsymbol{H} \right) \mathrm{d}\boldsymbol{H}.$$

We give an upper bound for $\boldsymbol{H}^T \boldsymbol{H}$. By assumption (4.5) again, we have $\underline{M}\delta_n^{\rho+2} \leq w_{0,j,k} = \psi(H_{0,j,k}) \leq 4\overline{M}\delta_n^2$, then we can bound $H_{0,j,k}$ by

$$\log(\underline{M}\delta_n^{\rho+2}) \leq \log\left(\frac{\underline{M}\delta_n^{\rho+2}}{1 - \underline{M}\delta_n^{\rho+2}} \right) \leq H_{0,j,k} \leq \log\left(\frac{4\overline{M}\delta_n^2}{1 - 4\overline{M}\delta_n^2} \right) \leq \log(8\overline{M}\delta_n^2) < 0.$$

Then for any $\boldsymbol{H} \in B_n$,

$$H_{j,k} \leq H_{0,j,k} + 2\underline{M}\delta_n^{\rho+2} p_n^{-1} \leq \log(8\overline{M}\delta_n^2) + \underline{M}\delta_n^{\rho+2} p_n^{-1} < 0.$$

Hence

$$H_{j,k}^2 \leq (H_{0,j,k} - 2\underline{M}\delta_n^{\rho+2} p_n^{-1})^2 \leq (\log(\underline{M}\delta_n^{\rho+2}) - 2\underline{M}\delta_n^{\rho+2} p_n^{-1})^2 \leq C_2 (\log n)^2$$

for some constant $C_2 > 0$. Then we have

$$\boldsymbol{H}^T\boldsymbol{H} \leq C_2 p_n (\log n)^2.$$

Using this and the fact that B_n is a hyper-rectangle in $\mathbb{R}^{p_n}$,

$$\begin{aligned}\int_{B_n} e^{-\frac{1}{2}\tau \boldsymbol{H}^T\Sigma\boldsymbol{H}}\,\mathrm{d}\boldsymbol{H} &\geq \exp\left(-\frac{1}{2}C_2\tau\lambda_{p_n}p_n(\log n)^2\right)\int_{B_n} 1\cdot \mathrm{d}\boldsymbol{H}\\ &= \left(\frac{4}{3}\underline{M}\delta_n^{\rho+2}p_n^{-1}\right)^{p_n}\exp\left(-\frac{1}{2}C_2\tau\lambda_{p_n}p_n(\log n)^2\right).\end{aligned}$$

Hence we have

$$\begin{aligned}\Pi_n(\{f_{\boldsymbol{H}} : \boldsymbol{H}\in B_n\}) &\geq \frac{\gamma^\beta}{\Gamma(\alpha)}(2\pi)^{-\frac{p_n}{2}}3^{-p_n}(4\underline{M}\delta_n^{\rho+2}p_n^{-1})^{p_n}|\Sigma^{-1}|^{1/2}\\ &\quad\times\int_0^\infty \tau^{\beta+\frac{p_n}{2}-1}\exp\left(-\left(\frac{1}{2}C_2\lambda_{p_n}p_n(\log n)^2+\gamma\right)\tau\right)\mathrm{d}\tau\\ &= \frac{\gamma^\beta}{\Gamma(\alpha)}(4\underline{M}(12\pi)^{-\frac{1}{2}}\delta_n^{\rho+2}p_n^{-1})^{p_n}|\Sigma^{-1}|^{1/2}\frac{\Gamma(\beta+p_n/2)}{(\frac{1}{2}C_2\lambda_{p_n}p_n(\log n)^2+\gamma)^{\beta+\frac{p_n}{2}}}\\ &\geq \frac{\gamma^\beta}{\Gamma(\alpha)}(4\underline{M}(12\pi)^{-\frac{1}{2}}\lambda_{p_n}^{-\frac{1}{2}}\delta_n^{\rho+2}p_n^{-1})^{p_n}\left(\frac{\beta+p_n/2}{(\frac{1}{2}eC_2\lambda_{p_n}p_n(\log n)^2+\gamma)}\right)^{\beta+\frac{p_n}{2}}(\beta+p_n/2)^{-1/2}.\end{aligned}$$

In the final step we use (4.11), $|\Sigma^{-1}|^{1/2} = |\Sigma|^{-\frac{1}{2}} \geq (\lambda_{p_n})^{-\frac{1}{2}p_n}$ and $\Gamma(x) \asymp (x/e)^x x^{-1/2}$ when x is large enough. By the inequality (4.14), we further have

$$\begin{aligned}\Pi_n(\{f_{\boldsymbol{H}} : \boldsymbol{H}\in B_n\}) &\gtrsim \exp\Bigg(p_n\log(4\underline{M}(12\pi)^{-\frac{1}{2}}\delta_n^{\rho+2}p_n^{-1}(4p_n+p_n^{-1})^{-\frac{1}{2}})\\ &\qquad+\left(\beta+\frac{p_n}{2}\right)\log(C_3p_n^{-1}(\log n)^{-2})-\frac{1}{2}\log(\beta+p_n/2)\Bigg)\\ &\gtrsim \exp(-C_4\delta_n^{-2}\log n) = \exp(-cn\varepsilon_n^2)\end{aligned}$$

for some positive constants C_3, C_4, c. In the last step we use (C.9).

For both types of prior, we derived $\Pi(\Omega_n) \gtrsim \exp(-cn\varepsilon_n^2)$, finishing the proof. □

4.4.2. Proof of lemma 4.3.6

Proof. Recall that $\eta_n = (n/\log n)^{-\frac{\rho}{3(\rho+2)}}$. Note that $\eta_n \asymp \varepsilon_n^{2/3}$. For $(t,z)\in[0,M_1]\times(0,M_2]$, define sets

$$\begin{aligned}U_{n,1}(t,z) &= \{f : F(t,z) > F_0(t,z)+C\eta_n\},\\ U_{n,2}(t,z) &= \{f : F(t,z) < F_0(t,z)-C\eta_n\}.\end{aligned}$$

Then $U_n(t,z) = U_{n,1}(t,z) \cup U_{n,2}(t,z)$. We consider different test functions in different regimes of t: $t \in (0, M_1)$ and $t \in \{0, M_1\}$.

Now fix $(t,z) \in (0,M_1) \times (0,M_2]$. Define test sequences

$$\Phi_n^+(t,z) = \mathbf{1}\left\{ \frac{1}{n}\sum_{i=1}^n \kappa_n^+(t,z;T_i,Z_i) - \int_t^{t+h_n} g(x)F_0(x,z)\,\mathrm{d}x > e_n/2 \right\},$$
$$\Phi_n^-(t,z) = \mathbf{1}\left\{ \frac{1}{n}\sum_{i=1}^n \kappa_n^-(t,z;T_i,Z_i) - \int_{t-h_n}^{t} g(x)F_0(x,z)\,\mathrm{d}x < -e_n/2 \right\},$$

where

$$\kappa_n^+(t,z;T,Z) = \mathbf{1}_{[t,t+h_n]}(T)\mathbf{1}_{(0,z]}(Z),$$
$$\kappa_n^-(t,z;T,Z) = \mathbf{1}_{[t-h_n,t]}(T)\mathbf{1}_{(0,z]}(Z),$$

and let

$$h_n = (2M_2)^{-1}C\eta_n \min(1, \overline{M}^{-1}) \quad \text{and} \quad e_n = \frac{1}{2}C\underline{K}\eta_n h_n$$

be two sequences tending to zero. Recall that C is defined in theorem 4.3.4. By assumption 4.3.2, we have $\underline{K} \le g \le \overline{K}$. Then for any bivariate density function f,

$$\begin{aligned}
\mathbb{E}_f(\kappa_n^+(t,z;T,Z)) &= \int \mathbf{1}_{[t,t+h_n]}(x)\mathbf{1}_{(0,z]}(u)s_f(x,u)\,\mathrm{d}\mu(x,u) \\
&= \int_t^{t+h_n}\int_0^z g(x)\partial_2 F(x,u)\,\mathrm{d}\mu_2(x,u) \\
&= \int_t^{t+h_n} g(x)F(x,z)\,\mathrm{d}x \\
&\le \int_t^{t+h_n} g(x)\,\mathrm{d}x \le \overline{K}h_n
\end{aligned}$$

where s_f is the density function of (T,Z) defined in (4.1). The same upper bound holds for $\mathbb{E}_f(\kappa_n^-(t,z;T,Z))$. By Bernstein's inequality (Van der Vaart (1998), lemma 19.32),

$$\mathbb{E}_0(\max(\Phi_n^+(t,z), \Phi_n^-(t,z))) \le 2\exp\left(-\frac{1}{16}\frac{ne_n^2}{\overline{K}h_n + e_n/2}\right) = o(1).$$

When $f \in U_{n,1}(t,z)$, for any $x \in [t, t+h_n]$, by the monotonicity of F and $f_0 \le \overline{M}$, we have

$$\begin{aligned}
F(x,z) - F_0(x,z) &\ge F(t,z) - F_0(t,z) - (F_0(x,z) - F_0(t,z)) \\
&\ge C\eta_n - \overline{M}M_2h_n \ge C\eta_n/2.
\end{aligned}$$

Then it follows

$$\int_t^{t+h_n} g(x)(F(x,z)-F_0(x,z))\,\mathrm{d}x \geq \frac{C\eta_n}{2}\int_t^{t+h_n} g(x)\,\mathrm{d}x \geq \frac{C\underline{K}}{2}\eta_n h_n = e_n.$$

Hence, for $f \in U_{n,1}$ we have

$$\begin{aligned}\mathbb{E}_f(1-\Phi_n^+(t,z)) &= \mathbb{P}_f\left(\frac{1}{n}\sum_{i=1}^n \kappa_n^+(t,z|T_i,Z_i) - \int_t^{t+h_n} g(x)F_0(x,z)\,\mathrm{d}x < e_n/2\right)\\ &\leq \mathbb{P}_f\left(\frac{1}{n}\sum_{i=1}^n \kappa_n^+(t,z|T_i,Z_i) - \int_t^{t+h_n} g(x)F(x,z)\,\mathrm{d}x \leq -e_n/2\right).\end{aligned}$$

Further, Bernstein's inequality gives

$$\mathbb{E}_f(1-\Phi_n^+(t,z)) \leq 2\exp\left(-\frac{1}{16}\frac{ne_n^2}{\overline{K}h_n + e_n/2}\right) \leq c_1 e^{-c_2C^2n\varepsilon_n^2}$$

for some constants $c_1, c_2 > 0$.

When $f \in U_{n,2}(t,z)$, $x \in [t-h_n, t]$, we have

$$\begin{aligned}F(x,z)-F_0(x,z) &\leq F(t,z)-F_0(t,z)+F_0(t,z)-F_0(x,z)\\ &\leq -C\eta_n + \overline{M}M_2h_n \leq -C\eta_n/2\end{aligned}$$

and

$$\int_{t-h_n}^t g(x)(F(x,z)-F_0(x,z))\,\mathrm{d}x \leq -\frac{C\underline{K}}{2}\eta_n h_n = -e_n.$$

Hence for $f \in U_{n,2}$, the type II error satisfies

$$\mathbb{E}_f(1-\Phi_n^-(t,z)) \leq \mathbb{P}_f\left(\frac{1}{n}\sum_{i=1}^n \kappa_n^-(t,z|T_i,Z_i) - \int_{t-h_n}^t g(x)F(x,z)\,\mathrm{d}x \geq e_n/2\right).$$

Using Bernstein's inequality again, we have

$$\mathbb{E}_f(1-\Phi_n^-(t,z)) \leq c_1 e^{-c_2C^2n\varepsilon_n^2}, \quad \text{for some } c_1, c_2 > 0.$$

For the boundary case $(t,z) \in \{0, M_1\} \times (0, M_2]$. With the similar idea, in order to give non-zero test sequences, we use κ_n^+ define $\Phi_n^+(0,z), \Phi_n^-(0,z)$ and κ_n^- define $\Phi_n^+(M_1,z), \Phi_n^-(M_1,z)$. When $f \in U_{n,1}(0,z)$, using the tests sequence $\Phi_n^+(0,z)$ defined in case $t \in (0, M_1)$, we have

$$\sup_{f\in U_{n,1}(0,z)} \mathbb{E}_f(1-\Phi_n^+(0,z)) \leq c_1 e^{-c_2C^2n\varepsilon_n^2}.$$

When $f \in U_{n,2}(M_1, z)$, using the tests sequence $\Phi_n^-(M_1, z)$ defined in case $t \in (0, M_1)$, we have

$$\sup_{f \in U_{n,2}(M_1,z)} \mathbb{E}_f(1 - \Phi_n^-(M_1, z)) \leq c_1 e^{-c_2 C^2 n\varepsilon_n^2}.$$

Note that for any $f \sim \Pi_n$ and $t \in A_{n,j}$, $j = 1, \ldots, J_n$,

$$\int_0^{M_2} f(t, v)\, \mathrm{d}v = |A_{n,j}|^{-1} \sum_{k=1}^{K_n} \theta_{j,k} \leq \delta_n^{-1} K_n = M_2. \tag{4.15}$$

Here we use $\theta_{j,k} \leq 1$ and $|A_{n,j}| \geq \delta_n$. When $f \in U_{n,2}(0, z)$, for any $x \in [0, h_n]$, using (4.15) we have

$$\begin{aligned} F(x, z) - F_0(x, z) &\leq F(x, z) - F(0, z) + F(0, z) - F_0(0, z) \\ &\leq \int_0^x \int_0^z f(u, v)\, \mathrm{d}v\, \mathrm{d}u - C\eta_n \\ &\leq M_2 h_n - C\eta_n \leq -C\eta_n/2 \end{aligned}$$

and

$$\int_0^{h_n} g(x)(F(x, z) - F_0(x, z))\, \mathrm{d}x \leq -e_n.$$

Define tests sequence

$$\Phi_n^-(0, z) = \mathbf{1}\left\{ \frac{1}{n} \sum_{i=1}^n \kappa_n^+(0, z|T_i, Z_i) - \int_0^{h_n} g(x) F_0(x, z)\, \mathrm{d}x < -e_n/2 \right\}.$$

Hence by the Bernstein's inequality,

$$\mathbb{E}_f(1 - \Phi_n^-(0, z)) \leq c_1 e^{-c_2 C^2 n\varepsilon_n^2}.$$

By the similar arguments as above, when $f \in U_{n,1}(M_1, z)$, for any $x \in [M_1 - h_n, M_1]$, using (4.15) we have

$$\begin{aligned} F(x, z) - F_0(x, z) &\geq F(x, z) - F(M_1, z) + F(M_1, z) - F_0(M_1, z) \\ &\geq C\eta_n - \int_{M_1 - h_n}^{M_1} \int_0^z f(u, v)\, \mathrm{d}v\, \mathrm{d}u \\ &\geq C\eta_n - M_2 h_n \geq C\eta_n/2 \end{aligned}$$

and

$$\int_0^{h_n} g(x)(F(x, z) - F_0(x, z))\, \mathrm{d}x \geq -e_n.$$

Define tests sequence

$$\Phi_n^+(M_1, z) = \mathbf{1}\left\{\frac{1}{n}\sum_{i=1}^{n}\kappa_n^-(M_1, z|T_i, Z_i) - \int_{M_1-h_n}^{M_1} g(x)F_0(x, z)\,\mathrm{d}x > e_n/2\right\},$$

hence,

$$\mathbb{E}_f(1 - \Phi_n^+(M_1, z)) \leq c_1 e^{-c_2 C^2 n\varepsilon_n^2}.$$

To conclude, take $\Phi_n(t, z) = \max(\Phi_n^+(t, z), \Phi_n^-(t, z))$, we derived

$$\mathbb{E}_0\Phi_n(t, z) = o(1),$$
$$\sup_{f\in U_n(t,z)} \mathbb{E}_f(1 - \Phi_n(t, z)) \leq c_1 e^{-c_2 C^2 n\varepsilon_n^2}.$$

□

4.5. Computational study

In this section we present algorithms for drawing from the posterior distribution for both priors described in section 4.2.2.

4.5.1. Dirichlet prior

First, we consider the case where $\{(X_i, Y_i),\ i = 1, \ldots, n\}$ is a sequence of independent random vectors, with common density f_0 that is piecewise constant on $A_{n,j} \times B_{n,k}$ and compactly supported. This "no-censoring" model has likelihood

$$l(\boldsymbol{\theta}) = \prod_{j,k} \theta_{j,k}^{C_{j,k}},$$

where $C_{j,k} = \sum_i \mathbf{1}\{(X_i, Y_i) \in A_{n,j} \times B_{n,k}\}$ denotes the number of observations that fall in bin $A_{n,i} \times B_{n,k}$. Clearly, the Dirichlet prior is conjugate for the likelihood, resulting in the posterior being of Dirichlet type as well and known in closed form. In case of censoring, draws from the posterior for the Dirichlet prior can be obtained by data-augmentation, where the following two steps are alternated

1. Given $\boldsymbol{\theta}$ and censored data, simulate the "full data". This is tractable since the censoring scheme tells us in which collection of bins the actual observation can be located. Then one can renormalise the density f conditional on these bins and select a specific bin accordingly and generate the "full data". Cf. Figure 4.1.

2. Given the "full data", draw samples for $\boldsymbol{\theta}$ from the posterior according to a Dirichlet distribution.

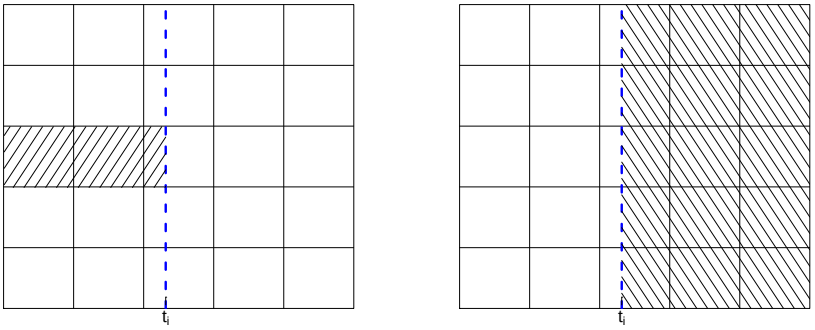

Figure 4.1.: Left: if $x_i \leq t_i$ the mark is observed. Right: if $x_i > t_i$ the mark is not observed.

4.5.2. Graph Laplacian prior

For the graph-Laplacian prior, one could opt for a data-augmentation scheme as well, but its attractiveness is lost, since step (2) is not anymore of simple form. Therefore, we propose to bypass data-augmentation in this case and use a probabilistic programming language to draw from the posterior. In such a language, only the hierarchical scheme and sampling method need to be specified. From this, the likelihood and prior are computed. Subsequently generic implementations of sampling methods are called. An example of such a language is STAN, where Hamiltonian Monte Carlo (HMC), or more specifically the No U-Turn Sampler (NUTS) (see for instance Robert et al. (2018), Van de Meent et al. (2018), Betancourt (2018)), is the sampler used. More recently, an implementation in the Julia language (see Bezanson et al.(2017)) has been provided in the Turing package (see Ge, Xu & Ghahramani (2018)). In this section we will use this package.

Unfortunately, there is presently no easy way to specify models with censored observations within the Turing-language. However, the model with censoring can easily be tweaked into a more familiar form that specifies the likelihood correctly. The only essential for a probabilistic programming language are the likelihood and hierarchical model specification. Specification of the prior is completely straightforward, while the likelihood can be specified by assuming a model with (conditionally independent) Bernoulli distributed random variables $Z_1, \ldots, Z_n$. For the i-th observation, let $\mathcal{I}_i$ denote the set of indices corresponding to the shaded areas as in either left- or right-hand-side panel of Figure 4.1. Hence, the union of all

boxes with indices in $\mathcal{I}_i$ specifies the area where the i-th observation is located. The success probability of Z_i is then given by inner product of the vector of shaded areas with the vector of corresponding probabilities $\theta_{j,k}$. Viewed in this way, the observation vector is simply a vector of length n consisting of ones, corresponding to observations $z_1 = z_2 = \ldots = z_n = 1$. The actual amount of programming is modest (Cf. appendix C.2).

4.5.3. Numerical examples

In the following simulations, we use the DynamicNUTS sampler from Hoffman & Gelman (2014). For the Dirichlet prior we took $5,000$ iterations of which the first half was discarded as burn-in. For the graph-Laplacian prior we took $2,000$ iterations or which the first 100 iterations were discarded as burn-in.

We will consider the following data generating settings for the joint distribution of (X, Y):

1. $f(x, y) = (x + y)\mathbf{1}_{[0,1]\times[0,1]}(x, y)$ (similar to example in Groeneboom, Jongbloed & Witte (2012));

2. the density of a Gaussian copula with correlation equal to -0.7.

In all cases we assume that $T \sim \sqrt{U}$ where U is uniformly distributed on $[0, 1]$. This implies that the density of T is given by $t \mapsto 2t\mathbf{1}_{[0,1]}(t)$. For the graph-Laplacian prior we took $\Sigma = L + 0.01I$ where L is defined in (4.3).

Experiment 1

Here we take density (1), sample size 100 and $J_n = K_n = 5$. In Figure 4.2 we show traceplots for the DynamicNUTS sampler. In the top row of Figure 4.3 we show for both priors a plot where each bin is coloured according to the deviation of the estimated posterior mean bin probability from the true bin probability. Clearly, the graph-Laplacian gives a much better fit. Moreover, the deviations visually appear to be smoother, in the sense that adjacent blocks tend to have similar colours.

Next, we repeat the experiment, though with a much finer grid specified by $J_n = K_n = 10$. Traceplots and a plot of the errors made are in figures 4.4 and the bottom panel of 4.3 respectively. Clearly, the errors are much smaller compared to $J_n = K_n = 5$. Moreover, the smoothing effect induced by the graph Laplacian prior is clearly visible. The sampler seems to have mixed after iteration 500 and for this reason the initial 500 samples were discarded as burnin samples. To compare the performance under both priors, we calculated the square root of the summed squared errors ($\sqrt{SSE}$). The results are as follows:

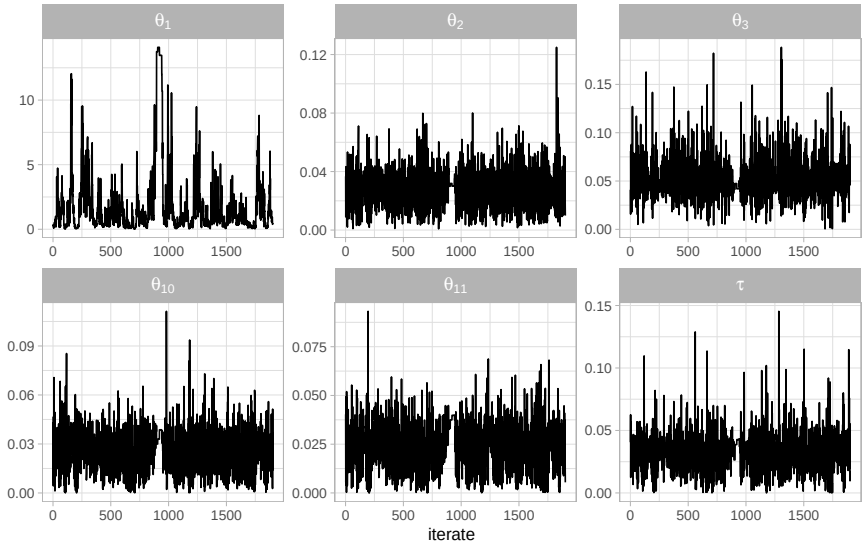

Figure 4.2.: Trace plots for a selected group of parameters in experiment 1, $J_n = K_n = 5$.

Resolution / prior	Dirichlet	graph-Laplacian
$J_n = K_n = 5$	0.201	0.035
$J_n = K_n = 10$	0.070	0.018

This confirms the superior performance of the graph-Laplacian prior for this example. As the true density is smooth, the latter is as expected.

Experiment 2

Here, we take the Gaussian copula, again with sample size $n = 100$. The setup of the experiment is the same as that of experiments 1. The results are displayed in figure 4.5. Again, we computed the square root of the summed squared errors ($\sqrt{SSE}$). The results are as follows:

Resolution / prior	Dirichlet	graph-Laplacian
$J_n = K_n = 5$	0.225	0.147
$J_n = K_n = 10$	0.160	0.080

As expected, the performance of the graph-Laplacian outperforms that of the Dirichlet.

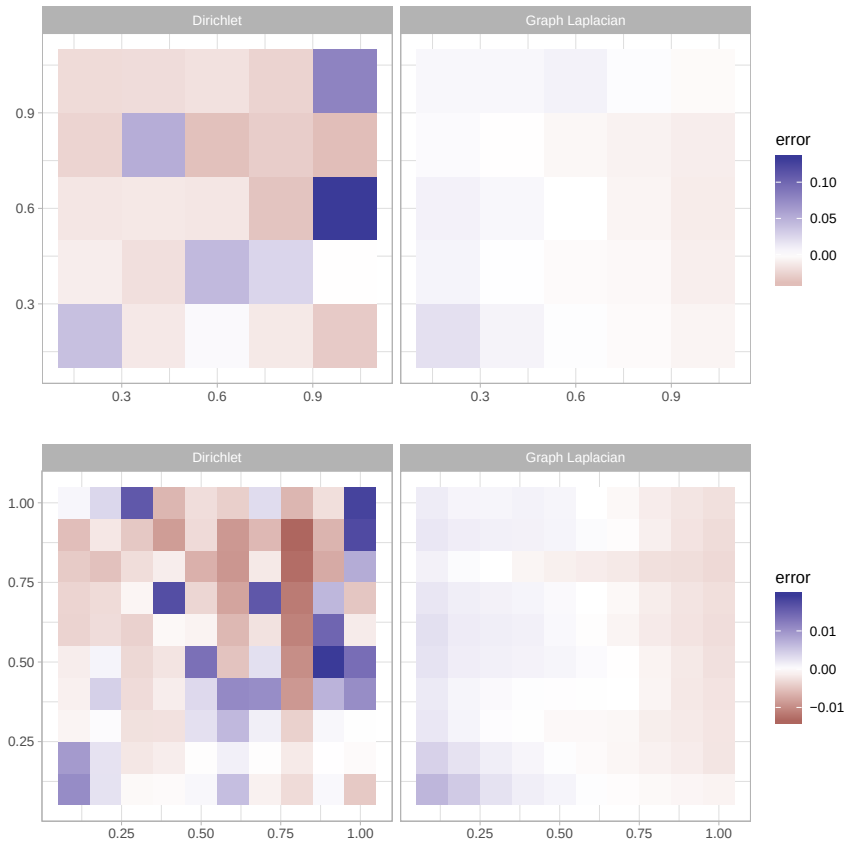

Figure 4.3.: Experiment 1: each bin is coloured according to the error within the bin, which is the estimated posterior mean of the bin probability minus the true bin probability. Left: Dirichlet prior. Right: graph-Laplacian prior. Note that the scale of colouring is the same in both figures. Top: $J_n = K_n = 5$. Bottom $J_n = K_n = 10$.

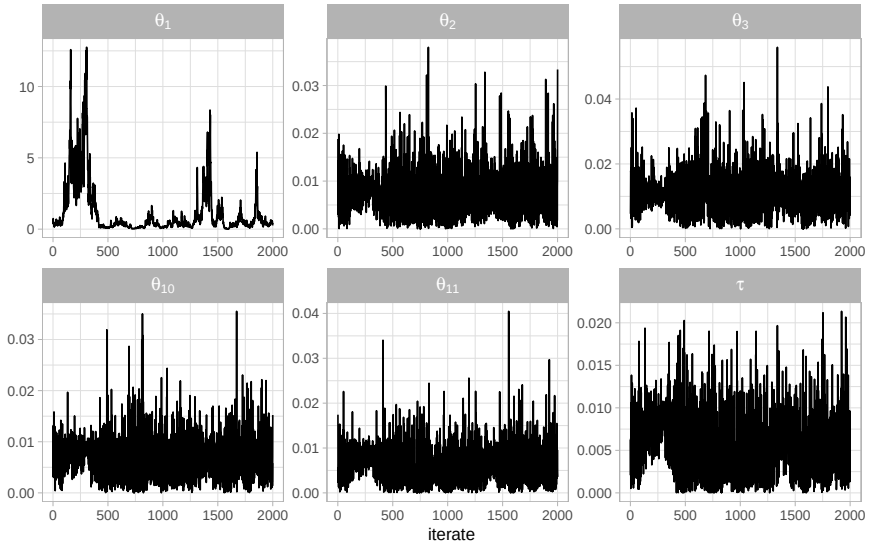

Figure 4.4.: Trace plots for a selected group of parameters in experiment 1, $J_n = K_n = 10$.

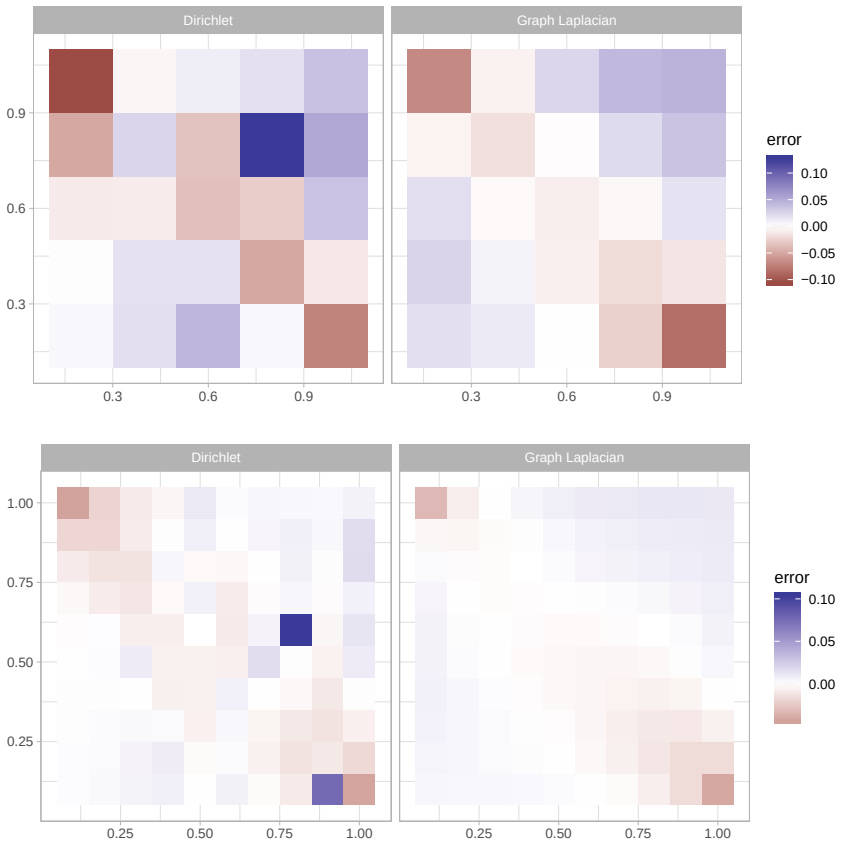

Figure 4.5.: Experiment 2: each bin is coloured according to the error within the bin, which is the estimated posterior mean of the bin probability minus the true bin probability. Left: Dirichlet prior. Right: graph-Laplacian prior. Note that the scale of colouring is the same in both figures. Top: $J_n = K_n = 5$. Bottom $J_n = K_n = 10$.

A. Supplement to Chapter 2

A.1. Review and supplementary proof of inequality (2.15)

In this section we point out a technical issue arising in the proof of inequality (2.15). As mentioned in section 2.2.2, it suffices to lower bound the prior mass of a certain subset $\mathcal{N}_n$ of $\mathcal{S}_n$, for which lower bounding $\Pi(\mathcal{N}_n)$ is tractable. To construct this set, we first need some approximation results.

Lemma A.1.1. *For any $\theta_0 > 0$ there exists a discrete measure $\tilde{P} = \sum_{i=1}^{\tilde{N}} \tilde{p}_i \delta_{y_i}$, with $y_i \in [\theta_0, \infty)$, $p_i \in [0,1]$, $\tilde{N} \lesssim 1/\varepsilon_n$ and $\sum_{i=1}^{\tilde{N}} p_i = \int_{\theta_0}^{\infty} f_0(x)\,\mathrm{d}x$ such that*

$$\int_{\theta_0}^{\infty} \left(\sqrt{f_0(x)} - \sqrt{f_{\tilde{P}}(x)} \right)^2 \mathrm{d}x \lesssim \varepsilon_n^2.$$

Moreover, the sequence $\{y_i\}$ can be taken such that $|y_i - y_j| \geq 2\varepsilon_n^2$ for all $i, j \leq \tilde{N}$.

Proof. Without the claimed separation property, existence of the discrete measure follows from lemma 11 in Salomond (2014). Denote this measure by $P = \sum_{i=1}^{N} p_i \delta_{z_i}$ and note that $N \lesssim 1/\varepsilon_n$. The set $y_1, \ldots, y_{\tilde{N}}$ is obtained from $\{z_1, \ldots, z_N\}$ by removing points from the latter set which are not $2\varepsilon_n^2$-separated. Clearly, $\tilde{N} \leq N \lesssim 1/\varepsilon_n$. The mass p_i of any removed point z_i is subsequently added to the point y_j $(1 \leq j \leq \tilde{N})$ that is closest to z_i. Denote the mass of y_j, obtained in this way, by $\tilde{p}_j$. Hence, we can written $\tilde{P} = \sum_{j=1}^{\tilde{N}} \tilde{p}_j \delta_{y_j} = \sum_{i=1}^{N} p_i \delta_{y_{k(i)}}$, where $k(i) = j$ if p_i assigned to $\tilde{p}_j$. Furthermore,

$$\begin{aligned}
L_1\left(f_P, f_{\tilde{P}}\right) &= \int \left| \sum_{i=1}^{N} p_i \psi_x(z_i) - \sum_{j=1}^{\tilde{N}} \tilde{p}_j \psi_x(y_j) \right| \mathrm{d}x \\
&= \int \left| \sum_{i=1}^{N} p_i (\psi_x(z_i) - \psi_x(y_{k(i)})) \right| \mathrm{d}x \\
&= \int \left| \sum_{i : z_i \neq y_{k(i)}} p_i (\psi_x(z_i) - \psi_x(y_{k(i)})) \right| \mathrm{d}x
\end{aligned}$$

Since for any $\theta_0 < \theta_1 < \theta_2$,

$$\begin{aligned}\int |\psi_x(\theta_1) - \psi_x(\theta_2)| \, \mathrm{d}x &= \int_{x \le \theta_1} + \int_{\theta_1 < x \le \theta_2} + \int_{x > \theta_2} |\psi_x(\theta_1) - \psi_x(\theta_2)| \, \mathrm{d}x \\ &= 2(\theta_2 - \theta_1)/\theta_2 \lesssim \theta_2 - \theta_1.\end{aligned}$$

This implies that

$$\begin{aligned}L_1\left(f_P, f_{\tilde{P}}\right) &\le \sum_{i : z_i \neq y_{k(i)}} p_i \int |\psi_x(z_i) - \psi_x(y_{k(i)})| \, \mathrm{d}x \\ &\le \sum_{i : z_i \neq y_{k(i)}} p_i \varepsilon_n^2 \lesssim \varepsilon_n^2\end{aligned}$$

The claimed result now follows from the triangle inequality and that the squared Hellinger distance is bounded by the L_1-distance. □

Lemma A.1.2. *Assume f_0 satisfies assumption 2.2.2. There exists a discrete probability measure $\tilde{P}$, supported on $\{i\varepsilon_n, 1 \le i \le N'\} \cup \{y_j, 1 \le j \le \tilde{N}\}$, with $N' = \lfloor x_0/\varepsilon_n \rfloor$ such that*

$$\int_0^\infty \left(\sqrt{f_0(x)} - \sqrt{f_{\tilde{P}}(x)}\right)^2 \mathrm{d}x \lesssim \varepsilon_n^2.$$

Proof. By lemma A.1.1 applied with $\theta_0 = x_0$ it suffices to prove $\int_0^{x_0} (\sqrt{f_0(x)} - \sqrt{f_{\tilde{P}(x)}})^2 \, \mathrm{d}x \lesssim \varepsilon_n^2$. Define the measure $\tilde{P} = \sum_{i=1}^{N'} p_i' \delta_{i\varepsilon_n} + \sum_{j=1}^{\tilde{N}} \tilde{p}_j \delta_{y_j}$, where $\tilde{p}_j$ is as defined in lemma A.1.1 and

$$p_i' = \begin{cases} (f_0((i-1)\varepsilon_n) - f_0(i\varepsilon_n))i\varepsilon_n & \text{if} \quad i < N' \\ (f_0((N'-1)\varepsilon_n) - a)N'\varepsilon_n & \text{if} \quad i = N' \end{cases}$$

with $a = \sum_{j=1}^{\tilde{N}} \tilde{p}_j / y_j$. Then for $x \in ((i-1)\varepsilon_n, i\varepsilon_n]$,

$$\begin{aligned}f_{\tilde{P}}(x) &= \sum_{k=i}^{N'} p_k' \psi_x(k\varepsilon_n) + \sum_{j=1}^{\tilde{N}} \tilde{p}_j \psi_x(y_j) = \sum_{k=i}^{N'} \frac{p_k'}{k\varepsilon_n} + a \\ &= \sum_{k=i}^{N'-1} k\varepsilon_n \frac{f_0((k-1)\varepsilon_n) - f_0(k\varepsilon_n)}{k\varepsilon_n} + \frac{f_0((N'-1)\varepsilon_n) - a}{N'\varepsilon_n} N'\varepsilon_n + a \\ &= f_0((i-1)\varepsilon_n)\end{aligned}$$

By the mean value theorem, it follows that

$$\begin{aligned}\int_0^{x_0}\left(\sqrt{f_0(x)}-\sqrt{f_{\tilde{P}}(x)}\right)^2\,\mathrm{d}x &= \sum_{i=1}^{N'}\int_{(i-1)\varepsilon_n}^{i\varepsilon_n}\left(\sqrt{f_0(x)}-\sqrt{f_0((i-1)\varepsilon_n)}\right)^2\,\mathrm{d}x\\ &\le \sum_{i=1}^{N'}\int_{(i-1)\varepsilon_n}^{i\varepsilon_n}\left(\frac{f_0'(\zeta_i)}{2\sqrt{f_0(\zeta_i)}}(x-(i-1)\varepsilon_n)\right)^2\,\mathrm{d}x\\ &\le \frac{(\sup_{x\in[0,x_0]}|f_0'(x)|)^2}{4f_0(\theta_0)}\sum_{i=1}^{N'}\int_{(i-1)\varepsilon_n}^{i\varepsilon_n}(x-(i-1)\varepsilon_n)^2\,\mathrm{d}x\\ &= \frac{(\sup_{x\in[0,x_0]}|f_0'(x)|)^2}{12f_0(\theta_0)}\sum_{i=1}^{N'}\varepsilon_n^3\lesssim\varepsilon_n^2\end{aligned}$$

where $\zeta_i\in((i-1)\varepsilon_n,i\varepsilon_n)$. □

By lemmas A.1.1 and A.1.2 we have.

Corollary A.1.3. *Assume f_0 satisfies assumption 2.2.2. There exists a discrete probability measure $\tilde{P}$, supported on $\{i\varepsilon_n,\,1\le i\le N'\}\cup\{y_j,\,1\le j\le\tilde{N}\}$, with $\min_{1\le j\le\tilde{N}}y_j\ge x_0$, $N'=\lfloor x_0/\varepsilon_n\rfloor$ and $\tilde{N}\lesssim 1/\varepsilon_n$ such that*

$$\int_0^{\infty}\left(\sqrt{f_0(x)}-\sqrt{f_{\tilde{P}}(x)}\right)^2\,\mathrm{d}x\lesssim\varepsilon_n^2.$$

Moreover, the sequence $\{y_i\}$ can be taken such that $|y_i-y_j|\ge 2\varepsilon_n^2$ for all $i,j\le\tilde{N}$.

For easy reference, we redefine the weights $\tilde{p}_j$ of the measure $\tilde{P}$ from this corollary so that we can write $\tilde{P}=\sum_{j=1}^{N'}\tilde{p}_j\delta_{j\varepsilon_N}+\sum_{j=1}^{\tilde{N}}\tilde{p}_{N'+j}\delta_{y_j}$.

Next, we use the support points and masses of the constructed measure $\tilde{P}$. To this end, define

$$\begin{aligned}U_i &= (i\varepsilon_n,(i+1)\varepsilon_n] \quad \text{for} \quad i=1,\dots,N'\\ U_{N'+i} &= [\theta_0\vee(y_i-\varepsilon_n^2),y_i+\epsilon_n^2] \quad \text{for} \quad i=1,\dots,\tilde{N}\\ U_0 &= [0,\infty)\cap(\cup_{i=1}^{\tilde{N}+N'}U_i)^c,\end{aligned}$$

such that $U_0,U_1,\dots,U_{N'+\tilde{N}}$ is a partition of $[0,\infty)$. Now define the following set of decreasing densities

$$\mathcal{N}_n=\{f_{P'}\,:\,P'([0,\infty))=1,\,|P'(U_i)-\tilde{p}_i|\le\epsilon_n^2/\tilde{N},\,1\le i\le\tilde{N}+N'\}$$

To prove that $\mathcal{N}_n$ is a subset of $\mathcal{S}_n$ a key property is that the measure $\tilde{P}$ is constructed such that $\int_0^\infty \left(\sqrt{f_0} - \sqrt{f_{\tilde{P}}}\right)^2 \lesssim \varepsilon_n^2$ (see the proof of lemma 8 in Salomond (2014)). Moreover, the prior mass of $\mathcal{N}_n$ is tractable because $U_0, U_1, \ldots, U_{N'+\tilde{N}}$ is a partition of $[0, \infty)$.

Remark A.1.4. If the set $\mathcal{N}_n$ is defined with the masses $p_1, \ldots, p_N$ from lemma A.1.1 (as is done in Salomond (2014), then the resulting sets $\{U_i\}$ do not form a partition. This results in intractable expressions for $\Pi(\mathcal{N}_n)$. For that reason, we defined another discrete measure $\tilde{P}$ such that the support points are $2\varepsilon_n^2$ separated thereby fixing the issue.

The arguments for lower bounding $\Pi(\mathcal{N}_n)$ can now be finished as outlined in Salomond (2014). Without loss of generality, for n sufficiently large we can assume $\alpha G_0(U_i) < 1$, for $i = 0, 1, \ldots, N' + \tilde{N}$. Similar to Lemma 6.1 in Ghosal, Ghosh & Van der Vaart (2000), we have

$$\begin{aligned}\Pi(\mathcal{N}_n) &\geq \mathrm{Dir}(P'(U_i) \in [\tilde{p}_i \pm \epsilon_n^2/\tilde{N}],\ 1 \leq i \leq \tilde{N} + N') \\ &\geq \Gamma(\alpha) \prod_{i=1}^{N'+\tilde{N}} \frac{1}{\Gamma(\alpha G_0(U_i))} \int_{0 \vee (\tilde{p}_i - \epsilon_n^2/\tilde{N})}^{\tilde{p}_i + \epsilon_n^2/\tilde{N}} x_i^{\alpha G_0(U_i) - 1} \, \mathrm{d}x_i.\end{aligned}$$

Here we use $(P'(U_0))^{\alpha G_0(U_0)-1} \geq 1$. As $x_i^{\alpha G_0(U_i)-1} \geq 1$ we have

$$\int_{0 \vee (p_i - \epsilon_n^2/\tilde{N})}^{p_i + \epsilon_n^2/\tilde{N}} x_i^{\alpha G_0(U_i) - 1} \, \mathrm{d}x_i \geq 2\epsilon_n^2 \tilde{N}^{-1}.$$

Substituting this bound into the lower bound on $\Pi(\mathcal{N}_n)$, combined with the inequalities $\beta\Gamma(\beta) = \Gamma(\beta + 1) \leq 1$ for $0 < \beta \leq 1$ and $\tilde{N} \lesssim \epsilon_n^{-1}$, we obtain

$$\Pi(\mathcal{N}_n) \gtrsim \epsilon_n^{3(N'+\tilde{N})} \prod_{i=1}^{N'+\tilde{N}} G_0(U_i) = \exp\left(3(N' + \tilde{N}) \log \epsilon_n + \sum_{i=1}^{N'+\tilde{N}} \log G_0(U_i)\right).$$

When $N' < i \leq N' + \tilde{N}$ it is trivial that $G_0(U_i) \gtrsim \epsilon_n^2$ and therefore

$$\sum_{i=N'+1}^{N'+\tilde{N}} \log G_0(U_i) \gtrsim \tilde{N} \log \epsilon_n.$$

For bounding $G_0(U_i)$ when $i \leq N'$, we use the property of g_0 in (2.5): $g_0(\theta) \geq \underline{k} e^{-\underline{a}/\theta}$. In this case we have

$$G_0(U_i) \geq \underline{k} \int_{U_i} e^{-\underline{a}/\theta} \, \mathrm{d}\theta \geq \underline{k} \epsilon_n \exp(-\underline{a}/(i\epsilon_n)).$$

Implying

$$\sum_{i=1}^{N'} \log G_0(U_i) \geq N' \log(\underline{k}\epsilon_n) - \underline{a}\epsilon_n^{-1} \sum_{i=1}^{N'} i^{-1}.$$

Since $\sum_{i=1}^{N'} i^{-1} \asymp \log(N') \asymp \log \epsilon_n^{-1}$, we therefore have

$$\sum_{i=1}^{N'} \log G_0(U_i) \gtrsim \epsilon_n^{-1} \log \epsilon_n. \tag{A.1}$$

Therefore, we obtain

$$\Pi(\mathcal{S}_n) \geq \Pi(\mathcal{N}_n) \gtrsim e^{C_1 \epsilon_n^{-1} \log \epsilon_n} \gtrsim e^{-C_1 n \epsilon_n^2}$$

for some $C_1 > 0$. This is exactly as is required.

A.2. Some details on the simulation in section 2.5

In this section we provide some computational details for updating the θ-values in the MCMC-sampler. Given the initialisation of (X, Z, Θ), we numerically evaluate $\int \psi(x_i|\theta) dG_0(\theta)$ for $i = 1, \ldots, n$. If g_0 is not conjugate to the uniform distribution, we use the random walk type Metropolis-Hastings method sampling from $f_{\Theta_k|X,Z}$ using the normal distribution. For update each Z_i, if $N_{Z_i,-i} = 0$, we first remove Θ_{Z_i}. If we draw a new "cluster" for Z_i, $1 + \vee(Z)$, then we also draw a new sample for Θ_{Z_i} according to (2.18). In this case, the product $\prod_{j:z_j=k} \psi(x_j|\theta_k)$ only has one item, that is $f_{\Theta|X,Z}(\theta|x, z) \propto g_0(\theta)\psi(x_i|\theta)$. Sampling a value for θ is done as follows:

1. If the base density g_0 is as in (2.22), then we use rejection sampling. To that end, if we set $Y = 1/\Theta$, then

$$f_{Y|X,Z}(y|x, z) = \frac{1}{y^2} f_{\Theta|X,Z}\left(\frac{1}{y}|x, z\right) = C\frac{1}{y} e^{-y-1/y} \mathbf{1}_{[0,1/x_i]}(y),$$

   where C is a constant such that $\int_0^\infty f_{Y|X,Z}(y|x, z)\, \mathrm{d}y = 1$. For reject sampling, we choose the proposal density $g(y)$ to be uniform on $[0, 1/x_i]$. Since $\frac{1}{y} e^{-y-1/y} \leq 0.18$ for any $y > 0$, an upper bound for $\frac{f_Y(y)}{g(y)}$ is given by $M = \frac{0.18 \cdot C}{x_i}$. Hence, we sample from $f_{Y|X,Z}$ as follows:

   a) sample $y \sim g(y)$, $u \sim \mathrm{Unif}(0, 1)$;

b) if

$$u \leq \frac{f(y)}{Mg(y)} = \frac{Ce^{-y-1/y}}{Myx_i} = \frac{e^{-y-1/y}}{0.18y},$$

then accept and set $\theta_{z_i} = 1/y$; else return to step (a).

2. If the base density g_0 is Gamma$(2, 1)$, then

$$f_{\Theta|X,Z}(\theta|x, z) = Ce^{-\theta}\mathbf{1}_{[x_i,\infty)}(\theta),$$

where $C = 1/\int_{x_i}^{\infty} e^{-\theta}\,\mathrm{d}\theta = e^{x_i}$. Hence the cumulative distribution function F_Θ satisfies $F_\Theta(\theta) = \int_{x_i}^{\theta} Ce^{-t}\,\mathrm{d}t = 1 - e^{x_i-\theta}$, when $\theta \geq x_i$. By the inverse cdf method, θ can be sampled by first sampling $u \sim$ Unif$(0, 1)$ and next computing $x_i - \log(u)$.

A.3. Results for the simulation experiment of Section 2.5.2 with sample size $n = 1000$

The results with $n = 1000$ are shown in figures A.1 and A.2.

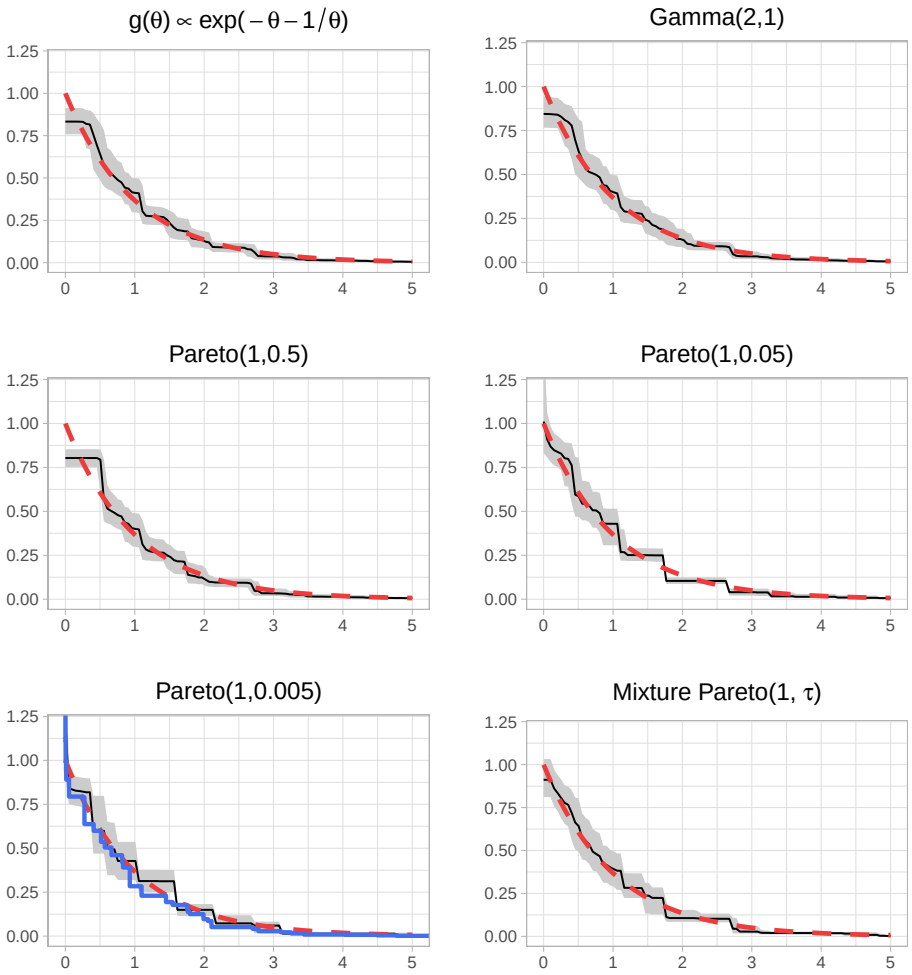

Figure A.1.: Same experiment as in Figure 2.1, this time with a sample of size 1000 from the standard Exponential distribution.

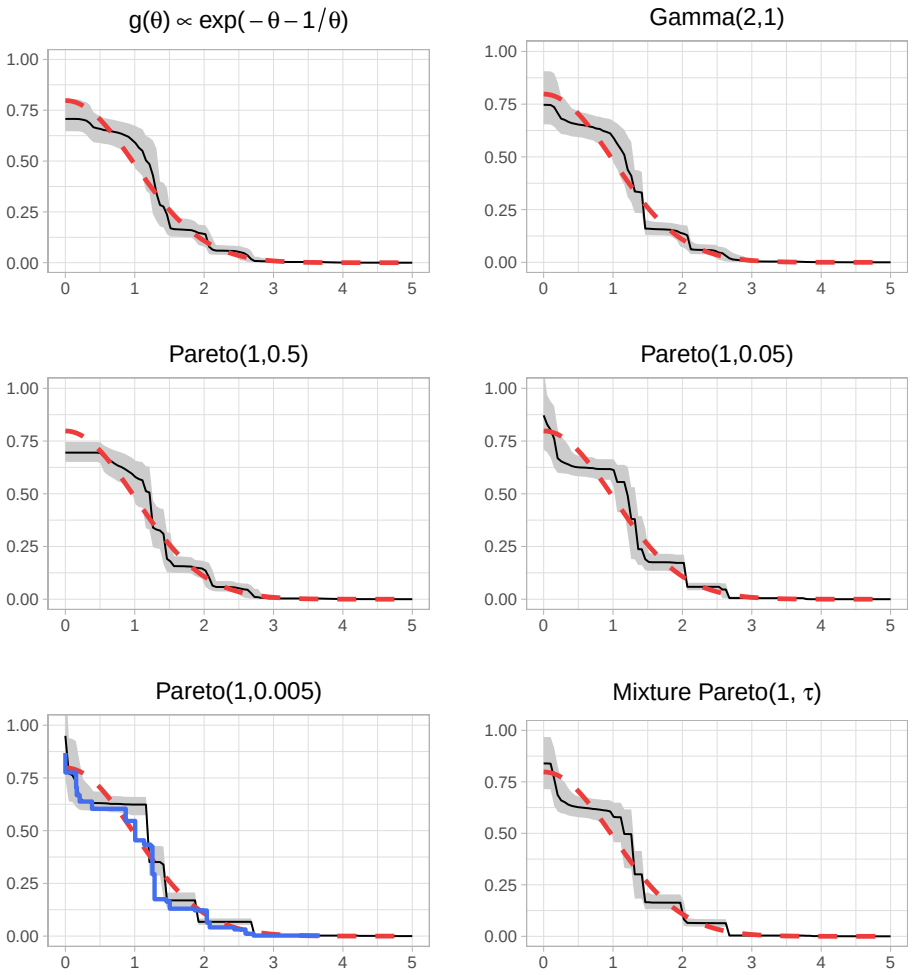

Figure A.2.: Same experiment as in Figure 2.2, this time with a sample of size 1000 from the halfNormal distribution.

B. Supplement to Chapter 3

B.1. Proofs of technical results

In the proof of lemma 3.3.3, we use the following lemma, it constructs a sequence of approximations for F_0.

Lemma B.1.1. *Let F_0 satisfy the conditions stated in theorem 3.3.1. Then there exists a sequence of piece-wise linear concave distribution functions (F_m) such that*

$$\sum_{k=1}^{\infty} p_K(k) \int g_k(t) h_{k,F_0,F_m}(t)\,\mathrm{d}t \to 0 \qquad as \quad m \to \infty.$$

Proof. Since F_0 is a concave distribution function, its density f_0 is decreasing on $[0,\infty)$. We start off with the construction of functions f_m that approximate f_0 (Cf. Theorem 18 in Wu & Ghosal (2008)). Choose $m \in \mathbb{N}$ and let $\tilde{f}_{0,m} = \frac{f_0 \mathbf{1}_{[0,m]}}{F_0(m)}$, then $\tilde{f}_{0,m} \to f_0$ point-wise as $m \to \infty$. Let a_1 and a_2 be real numbers such that $f_0(0) > a_1 > a_2 > 0$. By the continuity of f_0, there exists $x_2 > x_1$ satisfying $\tilde{f}_{0,m}(x_1) = a_1$ and $\tilde{f}_{0,m}(x_2) = a_2$. See also Figure B.1. Let $m_1 \in \mathbb{N}$ and $m_2 \in \mathbb{N}$ satisfy $\frac{m_1}{m} < x_1 \leq \frac{m_1+1}{m}$ and $\frac{m_2}{m} < x_2 \leq \frac{m_2+1}{m}$. Then define

$$\tilde{f}_m(x) = \begin{cases} \tilde{f}_{0,m}(\frac{i}{m}), & \frac{i-1}{m} < x \leq \frac{i}{m}, 1 \leq i \leq m_1 \\ a_1, & \frac{m_1}{m} < x \leq \frac{m_1+1}{m} \\ \tilde{f}_{0,m}(\frac{i-1}{m}), & \frac{i-1}{m} < x \leq \frac{i}{m}, m_1+1 < i \leq m^2. \end{cases}$$

and $\tilde{f}_m(0) = \tilde{f}_{0,m}(m^{-1})$. Because f_0 is continuous on $[0,m]$, $\tilde{f}_m$ converges point-wise to f_0 as $m \to \infty$. Note $\tilde{f}_m$ is not a probability density function, as it will not integrate to one. We now normalize $\tilde{f}_m$ to a density function f_m. First we can rewrite $\tilde{f}_m$ as

$$\tilde{f}_m(x) = \sum_{i=1}^{m^2} \tilde{w}_i \varphi(x, i/m),$$

where φ is defined as (3.3) and

$$\tilde{w}_i = \begin{cases} \frac{i}{m}(\tilde{f}_0(\frac{i}{m}) - \tilde{f}_0(\frac{i+1}{m})), & 1 \le i < m_1 \\ \frac{m_1}{m}(\tilde{f}_0(\frac{m_1}{m}) - a_1), & i = m_1 \\ \frac{m_1+1}{m}(a_1 - \tilde{f}_0(\frac{m_1+1}{m})), & i = m_1 + 1 \\ \frac{i}{m}(\tilde{f}_0(\frac{i-1}{m}) - \tilde{f}_0(\frac{i}{m})), & m_1 + 1 < i < m^2 \\ m\tilde{f}_0(\frac{m^2-1}{m}), & i = m^2. \end{cases}$$

Let

$$w_i = \begin{cases} \tilde{w}_i \frac{1-\sum_{j=1}^{m_1-1}\tilde{w}_j - \sum_{j=m_2+1}^{m^2}\tilde{w}_j}{\sum_{j=m_1}^{m_2}\tilde{w}_j}, & m_1 \le i \le m_2, \\ \tilde{w}_i, & \text{otherwise.} \end{cases}$$

Then $\sum_{i=1}^{m^2} w_i = 1$ and $w_i \ge 0$ (for m sufficiently large). Finally, define a sequence of probability density functions

$$f_m(x) = \sum_{i=1}^{m^2} w_i \varphi(x, i/m). \tag{B.1}$$

Note that for $x \ge x_2$, $f_m(x) = \tilde{f}_m(x)$. For each $x \in [0, m]$,

$$\begin{aligned} |f_m(x) - \tilde{f}_m(x)| &= \left| \sum_{i=1}^{m^2} w_i \varphi(x, \frac{i}{m}) - \sum_{i=1}^{m^2} \tilde{w}_i \varphi(x, i/m) \right| \\ &= \left| \sum_{i=m_1}^{m_2} (w_i - \tilde{w}_i) \varphi(x, i/m) \right| \\ &= \left| \left(\frac{1 - \sum_{j=2}^{m_1-1}\tilde{w}_j - \sum_{j=m_2+1}^{m^2}\tilde{w}_j}{\sum_{j=m_1}^{m_2}\tilde{w}_j} - 1 \right) \sum_{i=m_1}^{m_2} \tilde{w}_i \varphi(x, i/m) \right| \\ &\le \left| \frac{1 - \sum_{j=2}^{m_1-1}\tilde{w}_j - \sum_{j=m_2+1}^{m^2}\tilde{w}_j}{\sum_{j=m_1}^{m_2}\tilde{w}_j} - 1 \right| \left(\sum_{i=m_1}^{m_2} \tilde{w}_i \right) \frac{m}{m_1} \\ &= \left| 1 - \sum_{i=2}^{m_1-1} \tilde{w}_i - \sum_{i=m_2+1}^{m^2} \tilde{w}_i - \sum_{i=m_1}^{m_2} \tilde{w}_i \right| \frac{m}{m_1} \\ &= \left| 1 - \frac{1}{m} \sum_{i=2}^{m^2} \tilde{f}_{0,m}(i/m) - \frac{a_1}{m} \right| \frac{m}{m_1} \to 0 \end{aligned}$$

Here we use that $m/m_1 \to x_1^{-1}$ and that the expression within the modular signs converges to 0 as difference between $\int_0^m \tilde{f}_{0,m}(x)\,\mathrm{d}x$ and its Riemann sum approximate. Then we have $|f_m - \tilde{f}_{0,m}| \to 0$ point-wise and $\tilde{f}_{0,m} \to f_0$ point-wise. Hence

f_m is a decreasing density and converges to f_0 point-wise. See an example in figure B.1 for visualize $f_0, \tilde{f}_m$ and f_m.

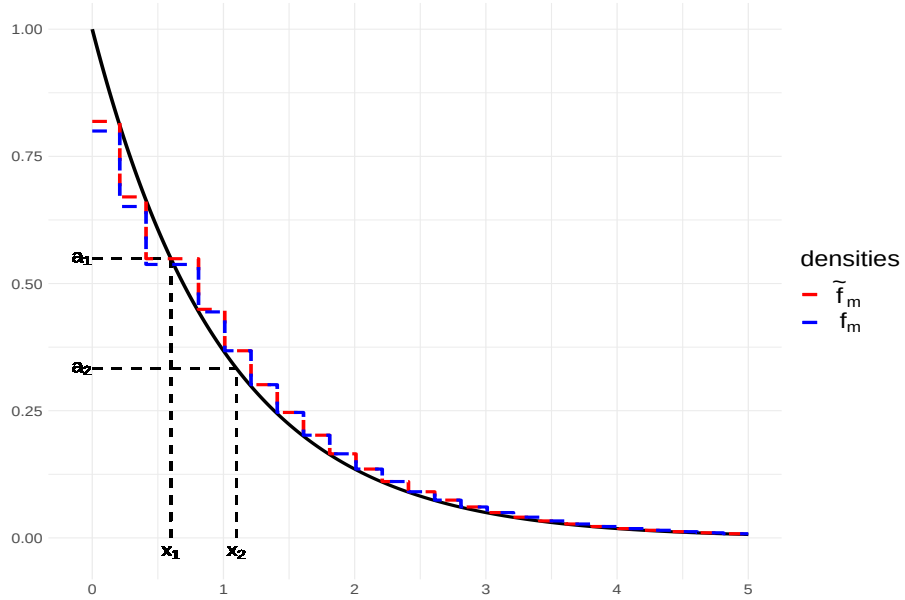

Figure B.1.: Approximation for a decreasing function f_0. First we construct a step function $\tilde{f}_m$, then we normalize the weights $\tilde{w}_i$ to w_i such that f_m defined by (B.1) is a decreasing density function.

Define $F_m(x) = \int_0^x f_m(t)\,\mathrm{d}t$, then using dominated convergence, we have $F_m \to F_0$ point-wise. As $m \to \infty$ $(m > L)$, then it follows that for all k and t

$$h_{k,F_0,F_m}(t) = \sum_{j=1}^{k+1}(F_0(t_j) - F_0(t_{j-1}))\log\frac{F_0(t_j) - F_0(t_{j-1})}{F_m(t_j) - F_m(t_{j-1})} \to 0.$$

The next step is to find an integrable upper bound for $|h_{k,F_0,F_m}|$. Denote $p_j = F_0(t_j) - F_0(t_{j-1})$ for $j = 1, \ldots, k+1$ and note that $\sum_{j=1}^{k+1} p_j = 1$. Then

$$|h_{k,F_0,F_m}(t)| \leq \sum_{j=1}^{k+1} p_j|\log p_j| + \sum_{j=1}^{k+1} p_j|\log(F_m(t_j) - F_m(t_{j-1}))|.$$

Using Lagrange multipliers, the first sum achieves its maximal value over all probability vectors when all p_j's would be equal. Hence it can be bounded by $\log(k+1)$. For the second sum, by the construction of f_m we know that when $x < x_2$, $f_m(x) \geq f_m(x_2) = \tilde{f}_m(x_2) = \tilde{f}_{0,m}(m_2/m) \geq a_2$; when $x_2 \leq x \leq m$,

$f_m(x) = \tilde{f}_m(x) \geq \tilde{f}_{0,m}(x) = \frac{f_0(x)}{F_0(m)} \geq f_0(x)$. Since there exists $j_0 \in \{1, \ldots, k+1\}$ such that $t_{j_0-1} < x_2 \leq t_{j_0}$, the second sum can be bounded by $I_1 + I_2 + I_3$, where

$$
\begin{aligned}
I_1 &= -\sum_{j=1}^{j_0-1} p_j \log(a_2(t_j - t_{j-1})) \\
I_2 &= -p_{j_0} \log(a_2(x_2 - t_{j_0-1}) + F_0(t_{j_0} \wedge m) - F_0(x_2)) \\
I_3 &= -\sum_{j=j_0+1}^{k} p_j \log p_j + p_{k+1} |\log(F_0(m) - F_0(t_k))|
\end{aligned}
$$

Again using the Lagrange multipliers, we have

$$
\begin{aligned}
I_1 &= -\sum_{j=1}^{j_0-1} \frac{p_j}{a_2(t_j - t_{j-1})} (a_2(t_j - t_{j-1})) \log(a_2(t_j - t_{j-1})) \\
&\leq -\frac{M}{a_2} \sum_{j=1}^{j_0-1} (a_2(t_j - t_{j-1})) \log(a_2(t_j - t_{j-1})) \leq \frac{M}{a_2} \log k
\end{aligned}
$$

In the second step we use $p_j \leq M(t_j - t_{j-1})$. In the final step, we use that $\sum_{j=1}^{j_0-1} a_2(t_j - t_{j-1}) \leq 1$. To bound I_2, we know that $-x \log x \leq \frac{1}{e}$ when $x \in (0, 1]$. Splitting I_2 into two parts, we have

$$
\begin{aligned}
I_2 &\leq -(F_0(t_{j_0}) - F_0(x_2)) \log(F_0(t_{j_0} \wedge m) \\
&\quad - F_0(x_2)) - (F_0(x_2) - F_0(t_{j_0-1})) \log(a_2(x_2 - t_{j_0-1})) \\
&\leq \frac{1}{e} \left(\frac{F_0(t_{j_0}) - F_0(x_2)}{F_0(t_{j_0} \wedge m) - F_0(x_2)} + \frac{F_0(x_2) - F_0(t_{j_0-1})}{a_2(x_2 - t_{j_0-1})} \right) \\
&\leq \frac{1}{e} \left(\frac{F_0(t_{j_0})}{F_0(t_{j_0} \wedge m)} + \frac{M}{a_2} \right) \leq \frac{1}{e} \left(\frac{1}{F_0(L)} + \frac{M}{a_2} \right)
\end{aligned}
$$

In the last step, we used that $\frac{F_0(t_{j_0})}{F_0(t_{j_0} \wedge m))} \leq \max(1, 1/F_0(m)) \leq 1/F_0(L)$. Similarly, we can bound I_3 by

$$
\begin{aligned}
I_3 &\leq \log k + p_{k+1} |\log(F_0(m) - F_0(t_k))| \\
&\leq \log k + \frac{1}{e} \frac{1 - F_0(t_k)}{F_0(m) - F_0(t_k)} \leq \log k + \frac{1}{e} \frac{1}{F_0(m)} \leq \log k + \frac{1}{e} \frac{1}{F_0(L)}
\end{aligned}
$$

Therefore, having these bounds we obtain

$$
| \, h_{k,F_0,F_m}(t) \, | \leq \left(\frac{M}{a_2} + 2 \right) \log(k+1) + \frac{1}{e} \left(\frac{2}{F_0(L)} + \frac{M}{a_2} \right).
$$

By the assumption in theorem 3.3.1, we have $\mathbb{E}\log(K+1) \leq C(r)K^r\infty < \infty$ for some constant $C(r)$ depend on r, hence

$$\sum_{k=1}^{\infty} p_K(k) \int g_k(t)|h_{k,F_0,F_m}(t)|\,\mathrm{d}t < \infty.$$

Therefore, by the dominated convergence theorem,

$$\sum_{k=1}^{\infty} p_K(k) \int g_k(t)h_{k,F_0,F_m}(t)\,\mathrm{d}t \to 0.$$

□

B.1.1. Proof of lemma 3.3.3

Proof. By lemma B.1.1, for any $\eta > 0$ there exists a sequence of piece-wise linear concave distribution functions (F_m) such that

$$\sum_{k=1}^{\infty} p_K(k) \int g_k(t)h_{k,F_0,F_m}(t)\,\mathrm{d}t < \eta/2 \tag{B.2}$$

for all m big enough. Recall definition (B.1),

$$f_m(x) = \sum_{i=1}^{m^2} w_i\varphi(x, \frac{i}{m}) = \int \varphi(x,\theta)\,\mathrm{d}P_m(\theta)$$

where $P_m(\cdot) = \sum_{i=1}^{m^2} w_i\delta_{i/m}(\cdot)$. Without loss of generality, assume $w_i > 0$ for all $i = 1, \ldots, m^2$. Given m fixed, for some $0 < \epsilon < \min(1, e^{\eta/4} - 1)$, define a discrete probability measure $P'_{m,\epsilon}(\cdot) = \sum_{i=1}^{m^2} w_i\delta_{(i+\epsilon/2)/m}(\cdot)$. Moreover, define the bounded Lipschitz distance on the set of probability measure on $[0,\infty)$ by

$$d_{BL}(P,Q) = \sup_{\psi\in\mathcal{C}_1} \left| \int \psi\,\mathrm{d}P - \int \psi\,\mathrm{d}Q \right|,$$

where $\mathcal{C}_1$ denotes the set of Lipschitz continuous functions on $[0,\infty)$ with Lipschitz constant 1. Then d_{BL} induces the weak topology (See Appendix A.2 in Ghosal & Van der Vaart (2017)). Choose $0 < \delta \leq \frac{\epsilon}{4m}(1 - e^{-\eta/4}) \min_{1\leq i\leq m^2} w_i$ and define the open set

$$\Omega_m = \left\{P \in \mathcal{M} : d_{BL}(P, P'_{m,\epsilon}) < \delta\right\}.$$

Choose Lipschitz continuous functions $\psi_j, j = 1, \ldots, m$ with compact support $[\frac{j}{m}, \frac{j+\epsilon}{m}]$, satisfying $\psi_j(\theta) = \frac{\epsilon}{4m}$ if $\theta \in (\frac{j+\frac{1}{4}\epsilon}{m}, \frac{j+\frac{3}{4}\epsilon}{m})$ and $0 \le \psi_j \le \frac{\epsilon}{4m}$. Denote $U_j = [\frac{j}{m}, \frac{j+\epsilon}{m}]$, $j = 1, \ldots, m^2$. Then for any $P \in \Omega_m$, $j = 1, \ldots, m^2$, we have

$$\left| \int \psi_j \, \mathrm{d}P - \int \psi_j \, \mathrm{d}P'_{m,\epsilon} \right| \le d_{BL}(P, P'_{m,\epsilon}) < \delta.$$

It also follows that for $j = 1, \ldots, m^2$,

$$\begin{aligned} \frac{\epsilon}{4m} P(U_j) &\ge \int \psi_j \, \mathrm{d}P \ge \int \psi_j \, \mathrm{d}P'_{m,\epsilon} - \delta \\ &\ge \frac{\epsilon}{4m} \int_{(j+\frac{1}{4}\epsilon)/m}^{(j+\frac{3}{4}\epsilon)/m} 1 \, \mathrm{d}P'_{m,\epsilon} - \delta = \frac{\epsilon}{4m} w_j - \delta \ge \frac{\epsilon}{4m} e^{-\eta/4} w_j. \end{aligned}$$

That is $P(U_j) \ge e^{-\eta/4} w_j$, for $j = 1, \ldots, m^2$. Using this lower bound and the mixture representation (3.4), we have for any $x \ge 0$, $P \in \Omega_m$,

$$\frac{f_m(x)}{f_P(x)} \le \frac{\sum_{i=1}^{m^2} w_i \varphi(x, \frac{i}{m})}{\sum_{i=1}^{m^2} \int_{U_i} \varphi(x, \theta) \, \mathrm{d}P(\theta)} \le \frac{\sum_{i=1}^{m^2} w_i \frac{m}{i} 1_{\{x \le \frac{i}{m}\}}}{\sum_{i=1}^{m^2} \frac{m}{i+\epsilon} 1_{\{x \le \frac{i}{m}\}} P(U_j)} \le (1+\epsilon) e^{\eta/4} \le e^{\eta/2}.$$

As this implies

$$F_m(t_j) - F_m(t_{j-1}) = \int_{t_{j-1}}^{t_j} f_m(x) \, \mathrm{d}x \le e^{\eta/2} \int_{t_{j-1}}^{t_j} f_P(x) \, \mathrm{d}x = e^{\eta/2} (F_P(t_j) - F_P(t_{j-1})),$$

we have that

$$\begin{aligned} h_{k,F_m,F_P}(t) &= \sum_{j=1}^{k+1} (F_0(t_j) - F_0(t_{j-1})) \log \frac{F_m(t_j) - F_m(t_{j-1})}{F_P(t_j) - F_P(t_{j-1})} \\ &\le \frac{\eta}{2} \sum_{j=1}^{k+1} (F_0(t_j) - F_0(t_{j-1})) \le \eta/2. \end{aligned} \tag{B.3}$$

Note that $h_{k,F_0,F_P}(t) = h_{k,F_0,F_m}(t) + h_{k,F_m,F_P}(t)$. Combining inequalities (B.2) and (B.3), we have

$$\sum_{k=1}^{\infty} p_K(k) \int g_k(t) h_{k,F_0,F_P}(t) \, \mathrm{d}t < \eta.$$

That means $\{F_P \in \mathcal{F} : P \in \Omega_m\} \subset S(\eta)$. Since Ω_m is an open weak neighborhood of P'_m in the neighborhood a and support$(\Pi^*) = \mathcal{M}$, we have $\Pi^*(\Omega_m) > 0$.

Recall that the prior Π on $\mathcal{F}$ is induced by the prior Π^* on $\mathcal{M}$ and the mixture representation (3.4), therefore $\Pi(S(\eta)) \ge \Pi^*(\Omega_m) > 0$.

□

B.1.2. Proof of lemma 3.3.4

Proof. We construct a test function depending on data $\mathcal{D}_n$. For any $\epsilon > 0$, define the event $A_n = \{d_n(\hat{F}_n, F_0) \geq \epsilon/2\}$, where $\hat{F}_n$ is the maximum likelihood estimator of the underlying distribution based on observations $\mathcal{D}_n$ (see Theorem 3 in Dümbgen, Freitag & Jongbloed (2006)) and d_n is defined as (3.5). Define $\Phi_n = \mathbf{1}\{A_n\}$, then as $n \to \infty$,

$$\begin{aligned} \mathbb{E}_0 \Phi_n &= \mathbb{E}_{K,T}\{\mathbb{E}_{F_0}[\Phi_n | K, T]\} \\ &= \mathbb{E}_{K,T}\{\mathbb{P}_{F_0}[d_n(\hat{F}_n, F_0) \geq \epsilon/2 | K, T]\} \to 0 \end{aligned} \tag{B.4}$$

The final step holds because the consistency of $\hat{F}_n$, $\mathbb{P}_{F_0}[d_n(\hat{F}_n, F_0) \geq \epsilon/2 | K, T] \to 0$ and this probability is bounded by 1. Similarly, given (K, T), for all $F \in U_\epsilon$

$$\begin{aligned} \mathbb{E}_F[1 - \Phi_n | (K, T)] &= \mathbb{P}_F[\{d_n(\hat{F}_n, F_0) \leq \epsilon/2\} \cap \{d_n(F, F_0) > \epsilon\} | (K, T)] \\ &\leq \mathbb{P}_F[d_n(F_0, F) - d_n(\hat{F}_n, F_0) \geq \epsilon/2 | (K, T)] \\ &\leq \mathbb{P}_F[d_n(F, \hat{F}_n) \geq \epsilon/2 | (K, T)] \end{aligned}$$

Then it is sufficient to prove for any $\epsilon > 0$,

$$\mathbb{E}_{(K,T)} \left\{ \sup_{F \in U_\epsilon} \mathbb{P}_F[d_n(F, \hat{F}_n) > \epsilon | (K, T)] \right\} \to 0.$$

We state that

$$\sup_{F \in U_\epsilon} \mathbb{P}_F[d_n(F, \hat{F}_n) > \epsilon | (K, T)] \to 0. \tag{B.5}$$

Then (B.4) and (B.5) are equivalent to the existence of a uniformly exponentially consistent test for testing $H_0 \colon F = F_0$ versus $H_1 \colon F \in U_\epsilon$ (see Proposition 4.4.1 in Ghosh & Ramamoorthi (2003)).

Now we show the inequality (B.5) holds. For a fixed $F \in \mathcal{F}$, the consistency result in Dümbgen, Freitag & Jongbloed (2006) claims that $d_n(F, \hat{F}_n) \to_p 0$, Actually, they proved that $\mathbb{P}_F[d_n(F, \hat{F}_n) > \epsilon] \to 0$ given the censoring times (K, T). We checking all steps of the proof in Dümbgen, Freitag & Jongbloed (2006), the consistency is follows from the finite expectation of K and the bound $F \leq 1$. Define

$$H^2(F, G) = (2n)^{-1} \sum_{i,j} (F_{i,j} - G_{i,j})^2.$$

The consistency result is follows from the following steps:

1. $d_n(F, \hat{F}_n) \leq 8^{1/2} H(F, \hat{F}_n)$;

2. $H(F,\hat{F}_n)^2 \le n^{-1}\sum_{i,j}(\Delta_{i,j}-F_{i,j})(\hat{F}_{n,i,j}/F_{i,j})^{1/2}$;

3. $n^{-1}\sum_{i,j}(\Delta_{i,j}-F_{i,j})(\hat{F}_{n,i,j}/F_{i,j})^{1/2} \le \sup_{G\in\mathcal{F}}|\sum_i(\psi_i(G)-\mathbb{E}_F\psi_i(G))|$;

where $\psi_i(G) = n^{-1}\sum_j \Delta_{i,j}(G_{i,j}/F_{i,j})^{1/2}$. Hence, it is sufficient to show

$$\mathbb{P}_F\left\{\sup_{G\in\mathcal{F}}\left|\sum_i(\psi_i(G)-\mathbb{E}_F\psi_i(G))\right| > \epsilon\right\} \to 0.$$

By theorem B.1.2, this is a consequence of the following conditions: for some sequences $\delta_n \to 0, b_n \to 0$,

$$\mathbb{E}_F\sum_{i=1}^n \sup_{G\in\mathcal{F}}|\psi_i(G)| = O(1), \tag{B.6}$$

$$\mathbb{E}_F\sum_{i=1}^n \mathbf{1}\{\sup_{G\in\mathcal{F}}|\psi_i(G)| > \delta_n\}\sup_{G\in\mathcal{F}}|\psi_i(G)| = b_n, \tag{B.7}$$

$$\text{for any } u>0, \quad \log\mathcal{N}(u,\mathcal{F},\rho_n) \le c(u). \tag{B.8}$$

where

$$\mathcal{N}(u,\mathcal{F},\rho_n) = \min\left\{\#\mathcal{G}\colon \mathcal{G}\subset\mathcal{F}, \inf_{G'\in\mathcal{G}}\rho_n(G,G') \le u \text{ for all } G\in\mathcal{F}\right\},$$

and

$$\rho_n(G,G') = \sum_{i=1}^n|\psi_i(G)-\psi_i(G')|.$$

We first give the main inequalities to derive these conditions. For (B.6),

$$\mathbb{E}_F\sum_{i=1}^n \sup_{G\in\mathcal{F}}|\psi_i(G)| \le n^{-1}\sum_i(K_i+1)^{1/2}.$$

For (B.7),

$$\mathbb{E}_F\sum_{i=1}^n 1\{\sup_{G\in\mathcal{F}}|\psi_i(G)| > \delta_n\}\sup_{G\in\mathcal{F}}|\psi_i(G)| \le n^{-1}\sum_i(K_i+1)^r(n\delta_n)^{-2\kappa} \to 0,$$

where $\kappa\in(0,\frac{1}{2})$, recall that $EK^r<\infty$ and choosing $n\delta_n\to\infty$. As for (B.8), ρ_n can be bounded by a finite measure, hence

$$\log\mathcal{N}(u,\mathcal{F},\rho_n) \le Cu^{-1}$$

for some constant C. (For more details see the proof of Theorem 3 in Dümbgen, Freitag & Jongbloed (2006)). Hence,

$$b_n = n^{-1}\sum_i (K_i+1)^r (n\delta_n)^{-2\kappa}, c(u) = Cu^{-1}.$$

By equation (B.12), we have

$$\mathbb{P}_F\left\{\sup_{G\in\mathcal{F}}\left|\sum_i(\psi_i(G)-\mathbb{E}_F\psi_i(G))\right|>\epsilon\right\} \le 4\epsilon^{-1}b_n + 128C\epsilon^{-1}\epsilon^{-1}\exp\left(-\frac{\epsilon^2}{512n\delta_n^2}\right)$$

Note that the right side do not depend on F, hence the inequality (B.5) holds. □

B.1.3. A technical result for proving uniform convergence

The following theorem follows from theorem 8.2 in Pollard (1990).

Theorem B.1.2. *Let $f_1(w,t), f_2(w,t), \ldots, f_n(w,t)$ be independent processes with integrable envelopes $F_1(w), F_2(w), \ldots, F_n(w)$. If for each $\epsilon > 0$,*

1. *there is a sequence $\delta_n \to 0$ such that*

$$\frac{1}{n}\sum_{i=1}^n \mathbb{E}F_i\mathbf{1}\{F_i>\delta_n\} < \epsilon, \quad \text{for all} \quad n,$$

2. $\log N(u, \mathcal{F}_{nw}, \rho_n) = c(u)$,

then

$$\sup_t|\sum_{i=1}^n (f_i(w,t)-\mathbb{E}f_i(w,t))|\to 0 \quad \text{in probability.}$$

Here $\mathcal{N}(u, \mathcal{F}_{nw}, \rho_n)$ is the covering number of $\mathcal{F}_{nw}$ with distance

$$\rho_n = \rho_n(t,t') = \sum_{i=1}^n |f_i(w,t)-f_i(w,t')|.$$

Proof. Define event $A_{n,i} := \{F_i > \delta_n\}$, then we split the expectation into two parts:

$$\begin{aligned}\sup_t\left|\sum_i(f_i(w,t)-\mathbb{E}f_i(w,t))\right| \le& \sup_t\left|\sum_i(f_i(w,t)\mathbf{1}\{A_{n,i}\}-\mathbb{E}f_i(w,t)\mathbf{1}\{A_{n,i}\})\right| \\ &+\sup_t\left|\sum_i(f_i(w,t)\mathbf{1}\{A_{n,i}^c\}-\mathbb{E}f_i(w,t)\mathbf{1}\{A_{n,i}^c\})\right|\end{aligned}$$

For the first item in the right side, by the condition 1, we have

$$\begin{aligned}
&\mathbb{P}\left\{\sup_t \Big|\sum_i (f_i(w,t)\mathbf{1}\{A_{n,i}\} - \mathbb{E}f_i(w,t)\mathbf{1}\{A_{n,i}\})\Big| > \epsilon/2\right\} \\
&\qquad \le 2\epsilon^{-1}\mathbb{E}\left\{\sup_t \Big|\sum_i (f_i(w,t)\mathbf{1}\{A_{n,i}\} - \mathbb{E}f_i(w,t)\mathbf{1}\{A_{n,i}\})\Big|\right\} \qquad \text{(B.9)} \\
&\qquad \le 4\epsilon^{-1}\mathbb{E}\left\{\sum_i \sup_t f_i(w,t)\mathbf{1}\{A_{n,i}\}\right\} = 4\epsilon^{-1}b_n
\end{aligned}$$

For the second item, denote $f_i^* = f_i\mathbf{1}\{A_{n,i}^c\}$. Using symmetrization, we have

$$\mathbb{P}\left\{\sup_t \Big|\sum_i (f_i^*(w,t) - \mathbb{E}f_i^*(w,t))\Big| > \epsilon/2\right\} \le 4\mathbb{E}_\sigma\mathbb{P}\left\{\sup_t \Big|\sum_i \sigma_i f_i^*(w,t)\Big| > \epsilon/8\right\},$$

where $\sigma_i = 1$ or -1 with probability $1/2$ independently. By the definition of covering number $\mathcal{N}(\epsilon/16, \mathcal{F}_{nw}, \rho_n)$, given w, for each t in $\mathcal{F}_{nw}$, there exists t' such that the distance $\rho_n(t,t') \le \epsilon/16$. Then we have

$$\begin{aligned}
\mathbb{P}\left\{\sup_t \Big|\sum_i \sigma_i f_i^*(w,t)\Big| > \epsilon/8\right\} &\le \mathbb{P}\left\{\max_{t'} \Big|\sum_i \sigma_i f_i^*(w,t')\Big| + \rho_n(t,t') > \epsilon/8\right\} \\
&\le \mathbb{P}\left\{\max_{t'} \Big[\sum_i \sigma_i f_i^*(w,t')\Big] > \epsilon/16\right\} \\
&\le \mathcal{N}(\epsilon/16, \mathcal{F}_{nw}, \rho_n) \max_{t'} \mathbb{P}\left\{\Big|\sum_i \sigma_i f_i^*(w,t')\Big| > \epsilon/16\right\} \qquad \text{(B.10)}
\end{aligned}$$

By the Hoeffding's inequality and $f_i^*(w,t') \le \delta_n$, we further have

$$\mathbb{P}\left\{\Big|\sum_i \sigma_i f_i^*(w,t')\Big| > \epsilon/16\right\} \le 2\exp\left(-\frac{2(\epsilon/16)^2}{\sum_i (2f_i^*(w,t'))^2}\right) \le 2\exp\left(-\frac{\epsilon^2}{512n\delta_n^2}\right) \qquad \text{(B.11)}$$

Therefore, combining inequalities (B.9), (B.10) and (B.11), we have

$$\mathbb{P}\left\{\sup_t \Big|\sum_i (f_i(w,t) - \mathbb{E}f_i(w,t))\Big| > \epsilon\right\} \le 4\epsilon^{-1}b_n + 8c(\epsilon/16)\epsilon^{-1}\exp\left(-\frac{\epsilon^2}{512n\delta_n^2}\right). \qquad \text{(B.12)}$$

By choosing $n\delta_n^2 \to 0$, we have the right side tend to 0.

□

C. Supplement to Chapter 4

C.1. Technical proof

Lemma C.1.1. *Define set*

$$\Omega_n := \left\{ f \in \mathcal{F} : ||f - f_{0,n}||_\infty \leq \frac{1}{6}\underline{M}\delta_n^\rho,\ supp(f) \supseteq \mathcal{M} \right\},$$

where $f_{0,n}$ is defined in (4.8). Then Ω_n is a subset of S_n (which is defined in (4.6)).

Proof. By the definition of $f_{0,n}$ in (4.8), for any $(t,z) \in A_{n,j} \times B_{n,k}$,

$$\begin{aligned}
|f_{0,n}(t,z) - f_0(t,z)| &= \left| |A_{n,j} \times B_{n,k}|^{-1} \int_{A_{n,j}} \int_{B_{n,k}} f_0(u,v)\,\mathrm{d}v\,\mathrm{d}u - f_0(t,z) \right| \\
&\leq |A_{n,j} \times B_{n,k}|^{-1} \int_{A_{n,j}} \int_{B_{n,k}} |f_0(u,v) - f_0(t,z)|\,\mathrm{d}v\,\mathrm{d}u \\
&\leq \max_{(u,v)\in A_{n,j}\times B_{n,k}} |f_0(u,v) - f_0(t,z)|.
\end{aligned}$$

By assumption (4.4) on f_0, we have

$$\max_{(u,v)\in A_{n,j}\times B_{n,k}} |f_0(u,v) - f_0(t,z)| \leq L \max_{(u,v)\in A_{n,j}\times B_{n,k}} ||(u,v) - (t,z)||^\rho \leq L(2\sqrt{2}\delta_n)^\rho.$$

Hence

$$\begin{aligned}
||f_{0,n} - f_0||_\infty &= \max_{j,k} \Big| \max_{(t,z)\in A_{n,j}\times B_{n,k}} |f_{0,n}(t,z) - f_0(t,z)| \Big| \\
&\leq \max_{j,k} \Big| \max_{(t,z)\in A_{n,j}\times B_{n,k}} L(2\sqrt{2}\delta_n)^\rho \Big| = L(2\sqrt{2}\delta_n)^\rho. \qquad \text{(C.1)}
\end{aligned}$$

Note that for any $(t,z) \in \mathcal{M}$ and $f_1, f_2 \in \mathcal{F}$,

$$\begin{aligned}
|s_{f_1}(t,z) - s_{f_2}(t,z)| &= \Bigg| g(t) \Bigg(\mathbf{1}_{\{z>0\}} \int_0^t (f_1(u,z) - f_2(u,z))\,\mathrm{d}u \\
&\qquad + \mathbf{1}_{\{z=0\}} \int_t^{M_1} \int_0^{M_2} (f_1(u,v) - f_2(u,v))\,\mathrm{d}v\,\mathrm{d}u \Bigg) \Bigg| \\
&\leq g(t)(M_1||f_1 - f_2||_\infty + M_1 M_2 ||f_1 - f_2||_\infty) \\
&\leq M_1(1 + M_2) g(t) ||f_1 - f_2||_\infty.
\end{aligned}$$

Further, we have

$$\|s_{f_1} - s_{f_2}\|_1 = \int_{\mathcal{M}} |s_{f_1} - s_{f_2}| \, \mathrm{d}\mu \leq \overline{K} M_1^2 M_2 (1 + M_2) \|f_1 - f_2\|_\infty. \tag{C.2}$$

By Lemma 8 of Ghosal & Van der Vaart (2007b), we know

$$\begin{aligned} KL(s_{f_0}, s_f) &\lesssim h^2(s_{f_0}, s_f) \left(1 + \log \|s_{f_0}/s_f\|_\infty\right), \\ V(s_{f_0}, s_f) &\lesssim h^2(s_{f_0}, s_f) \left(1 + \log \|s_{f_0}/s_f\|_\infty\right)^2. \end{aligned} \tag{C.3}$$

For any $f \in \Omega_n$, we give upper bounds of $h^2(s_{f_0}, s_f)$ and $\|s_{f_0}/s_f\|_\infty$. By (C.1) and (C.2), we know

$$\|s_{f_0} - s_f\|_1 \leq \overline{K} M_1^2 M_2 (1 + M_2)(\|f_0 - f_{0,n}\|_\infty + \|f_{0,n} - f\|_\infty) \lesssim \delta_n^\rho.$$

Using the inequality $h^2(f_1, f_2) \leq \frac{1}{2}\|f_1 - f_2\|_1$, we then have

$$h^2(s_{f_0}, s_f) \leq \frac{1}{2}\|s_{f_0} - s_f\|_1 \lesssim \delta_n^\rho. \tag{C.4}$$

We now give an upper bound on $\|s_{f_0}/s_f\|_\infty$, note that

$$\left\| \frac{s_{f_0}}{s_f} \right\|_\infty \leq \max \left\{ \left\| \frac{\partial_2 F_0}{\partial_2 F} \right\|_\infty, \left\| \frac{1 - F_{0,X}}{1 - F_X} \right\|_\infty \right\} \leq \left\| \frac{f_0}{f} \right\|_\infty \leq \left\| \frac{f_0}{f_{0,n}} \right\|_\infty \cdot \left\| \frac{f_{0,n}}{f} \right\|_\infty. \tag{C.5}$$

By the lower bound in inequality (4.5), we have for any $(t, z) \in A_{n,j} \times B_{n,k}$,

$$f_{0,n}(t, z) = |A_{n,j} \times B_{n,k}|^{-1} w_{0,j,k} \geq \underline{M} |A_{n,j} \times B_{n,k}|^{-1} \int_{A_{n,j} \times B_{n,k}} (\min(u, v))^\rho \, \mathrm{d}v \, \mathrm{d}u.$$

When $\min(j, k) > 1$,

$$\int_{A_{n,j} \times B_{n,k}} (\min(u, v))^\rho \, \mathrm{d}v \, \mathrm{d}u \geq \delta_n^\rho |A_{n,j} \times B_{n,k}|.$$

When $\min(j, k) = 1$ and $j \neq k$,

$$\int_{A_{n,j} \times B_{n,k}} (\min(u, v))^\rho \, \mathrm{d}v \, \mathrm{d}u = \frac{1}{\rho + 1} \delta_n^\rho |A_{n,j} \times B_{n,k}|.$$

When $j = k = 1$,

$$\int_{A_{n,j} \times B_{n,k}} (\min(u, v))^\rho \, \mathrm{d}v \, \mathrm{d}u = 2 \int_0^{\delta_n} \mathrm{d}v \int_0^v u^\rho \, \mathrm{d}u = \frac{2}{(\rho + 1)(\rho + 2)} \delta_n^{\rho + 2}.$$

Hence, in any of the cases, using $\rho \leq 1$, we obtain

$$\int_{A_{n,j} \times B_{n,k}} (\min(u, v))^\rho \, \mathrm{d}v \, \mathrm{d}u \geq \frac{1}{3} \delta_n^\rho |A_{n,j} \times B_{n,k}|.$$

Then it follows that

$$f_{0,n}(t, z) \geq \frac{M}{3} \delta_n^\rho. \tag{C.6}$$

Combining with (C.1),

$$\left\| \frac{f_0}{f_{0,n}} \right\|_\infty \leq 1 + \left\| \frac{f_0 - f_{0,n}}{f_{0,n}} \right\|_\infty \leq 1 + \frac{2^\rho 3L}{\underline{M}}. \tag{C.7}$$

Further, using (C.6) again, by definition of Ω_n, for $f \in \Omega_n$,

$$f(t, z) \geq f_{0,n}(t, z) - \frac{1}{6} \underline{M} \delta_n^\rho \geq \frac{1}{2} f_{0,n}(t, z).$$

Note that this implies if $f = 0$, then we have $f_{0,n} = 0$. Hence,

$$\left\| \frac{f_{0,n}}{f} \right\|_\infty \leq 2. \tag{C.8}$$

Substituting (C.7) and (C.8) into (C.5) gives that

$$\left\| \frac{s_{f_0}}{s_f} \right\|_\infty \leq 2 \left(1 + \frac{2^\rho 3L}{\underline{M}} \right).$$

Substituting this bound and (C.4) into (C.3) implies that there exists a $C_1 > 0$ such that

$$KL(s_{f_0}, s_f) \leq C_1 \delta_n^\rho, \quad V(s_{f_0}, s_f) \leq C_1 \delta_n^\rho.$$

Define

$$\varepsilon_n = \sqrt{C_1} (n / \log n)^{-\frac{\rho}{2(\rho+2)}} = \sqrt{C_1} \delta_n^{\frac{\rho}{2}}, \tag{C.9}$$

then we have $\Omega_n \subseteq S_n$. □

C.2. Programming details in the Turing language

For each observation index $i \in \{1, \ldots, n\}$ the indices $\mathcal{I}_i$ need to be computed and stored. Say that information is in the object **ci** (censoring information). Say that we define a function **bernpar** that takes the full parameter vector **theta** and **ci** and outputs the corresponding success probability. Finally, if **z** denotes the observation vector (taken to be a vector of length n containing solely ones) and **L** is the graph-Laplacian, then the model is specified as follows:

```
@model GraphLaplacianModel(z,ci,L) = begin
    tau ~ InverseGamma(.1,.1)
    H ~ MvNormalCanon(L*tau)
    theta = invlogit(H)
    for k in eachindex(z)
        z[k] ~ Bernoulli(bernpar(theta,ci[k]))
    end
end
```

Here, **invlogit** refers to the function ψ in (4.2).

References

Aldous, D.J. (1985). Exchangeability and related topics. In *École d'Été de Probabilités de Saint-Flour XIII — 1983*, p. 1—198.

Antoniak, C. (1974). Mixtures of Dirichlet processes with applications to Bayesian nonparametric problems. *The Annals of Statistics*. **2**, p. 1152—1174.

Balabdaoui, F., Jankowski, H., Pavlides, M., Seregin, A. and Wellner, J.A. (2011). On the Grenander estimator at zero. *Statist. Sinica* **2**1, p. 873–899.

Barron, A., Schervish, M.J. and Wasserman, L. (1999). The consistency of posterior distributions in nonparametric problems. *The Annals of Statistics*. **2**7, p. 536—561.

Bertoin J. (1998). *Lévy Processes.* Cambridge University Press.

Betancourt, M. (2018). A conceptual introduction to Hamiltonian Monte Carlo. arXiv:1701.02434.

Bezanson, J. and Edelman, A. and Karpinski, S. and Shah, V. (2017). Julia: A Fresh Approach to Numerical Computing. *SIAM Review* **5**9, p. 65-98.

Blackwell, D. and MacQueen J.B. (1973). Ferguson distributions via Polya urn schemes. *Ann. Statist.* **1**, p. 353-355.

Bush, C.A. and MacEachern, S.N. (1996). A Semiparametric Bayesian Model for Randomised Block Designs. *Biometrika*. **8**3, p. 275–285.

Calle, M.L. and Gómez, G. (2001). Nonparametric Bayesian estimation from interval-censored data using Monte Carlo methods. *Journal of Statistical Planning and Inference* **9**8, p. 73–87.

Chen, D., Sun, J. and Peace, K. E. (2013). *Interval-Censored Time-to-Event Data: Methods and Applications*. Chapman and Hall/CRC Biostatistics Series.

Choudhuri, N., and Ghosal, S. and Roy, A. (2007). Nonparametric binary regression using a Gaussian process prior. *Statistical Methodology*. **4**, p. 227–243.

Doss, H. (1994). Bayesian nonparametric estimation for incomplete data via successive substitution sampling. *The Annals of Statistics*. **2**2, p. 1763–1786.

Doss, H., and Huffer F.W. (2003). Monte Carlo Methods for Bayesian Analysis of Survival Data Using Mixtures of Dirichlet Process Priors. *Journal of Computational and Graphical Statistics* **1**2, p. 282—307.

Doss, H. and Sellke T. (1982). The Tails of Probabilities Chosen From A Dirichlet Prior. *Ann. Statist.* **1**0, p. 1302–1355.

Dümbgen, L., Freitag, S., and Jongbloed, G. (2004). Consistency of Concave Regression, With an Application to Current Status Data. *Mathematical Methods of Statistics* **1**3, p. 69—81.

Dümbgen, L., Freitag, S., and Jongbloed, G. (2006). Estimating a Unimodal Distribution From Interval-Censored Data. *Journal of the American Statistical Association* **1**01 , p. 1094–1106.

Feller, W. (1966). An Introduction to Probability Theory and Its Applications. *V*ol. II, John Wiley and Sons, New York.

Ferguson, T.S. (1973). A Bayesian Analysis of Some Nonparametric Problem. *Ann. Statist.* **1**, p. 209–230.

Ferguson, T.S. and Phadia, E.G. (1979). Bayesian Nonparametric Estimation Based on Censored Data. *The Annals of Statistics*. **7**, p. 163–186.

Finkelstein, D. M. and Wolfe, R. A. (1986). Isotonic Regression for interval censored survival data using an E-M algorithm. *Communications in Statistics: Theory and Methods*. **1**5, p. 2493–2505.

Ge, H., Xu, K., and Ghahramani, Z. (2018). Turing: A Language for Flexible Probabilistic Inference. *Proceedings of the Twenty-First International Conference on Artificial Intelligence and Statistics*, PMLR: **8**4, p. 1682—1690.

Ghosal, S., Ghosh, J.K. and Ramamoorthi, R.V. (1999). Posterior consistency of Dirichlet mixtures in density estimation. *The Annals of Statistics*. **2**7, p. 143–158.

Ghosal, S., Ghosh, J.K. and Van der Vaart, A.W. (2000). Convergence rates of posterior distributions. *The Annals of Statistics*. **2**8, p. 500–531.

Ghosh, J.K. and Ramamoorthi, R.V. (2003). Bayesian Nonparametrics. *Springer Series in Statistics*.

Ghosh, J.K., Ramamoorthi, R.V. and Srikanth, K.R. (1999). Bayesian analysis of censored data. *Statistics and Probability Letters*. **4**1, p. 255—265.

Ghosal, S. and Van der Vaart, A.W. (2007). Entropies and rates of convergence for maximum likelihood and Bayes estimation for mixtures of normal densities. *The Annals of Statistics*. **2**9, p. 1233–1236.

Ghosal, S. and Van der Vaart, A.W. (2007b). Posterior Convergence Rates of Dirichlet Mixtures at Smooth Densities. *The Annals of Statistics*. **3**5, p. 697–723.

Ghosal, S. and Van der Vaart, A.W. (2017). *Fundamentals of Nonparametric Bayesian Inference.* Cambridge University Press.

Gómez, G., Calle, M.L., and Oller, R. (2004). Frequentist and Bayesian approaches for interval-censored data. *Statistical Papers*. **4**5, p. 139–173.

Gómez, G., Calle, M.L., Oller, R. and Langohr, K. (2009). Tutorial on methods for interval-censored data and their implementation in R. *Statistical Modelling*.**9**, p. 259—297.

Grenander, U. (1956). On the theory of mortality measurement. II. *Skand. Aktuarietidskr.*, **3**9, p. 125—153.

Groeneboom, P., Maathuis, M. H. and Wellner, J. A. (2008). Current status data with competing risks: Consistency and rates of convergence of the MLE. *Ann. Statist.* **3**6, p. 1031–1063.

Groeneboom, P. and Jongbloed, G. (2014). *Nonparametric estimation under shape constraints*. Cambridge University Press.

Groeneboom, P. and Jongbloed, G. (2015). Nonparametric confidence intervals for monotone functions. *Ann. Statist.* **4**3, p.2019–2054.

Groeneboom, P., Jongbloed, G. and Wellner, J.A. (2001). Estimation of a convex function: characterizations and asymptotic theory. *Ann. Statist.* **2**9, p. 1653–1698.

Groeneboom, P., Jongbloed, G. and Witte, B. I. (2011). Smooth Plug-in Inverse Estimators in the Current Status Continuous Mark Model. *Scandinavian Journal of Statistics*. **3**9, p. 15–33.

Groeneboom, P., Jongbloed, G. and Witte, B. I. (2012). A maximum smoothed likelihood estimator in thecurrent status continuous mark model. *J. Nonparametr. Stat.* **2**4, p. 85–101.

Groeneboom, P., and Wellner, J.A. (1992). Information Bounds and Nonparametric Maximum Likelihood Estimation. Basel: Birkhäuser.

Gugushvili, S., van der Meulen, F. H., Schauer, M. R. and Spreij, P. (2019). Nonparametric Bayesian volatility estimation. (eds) *2017 MATRIX Annals, MATRIX Book Series*. **2**, p. 279–302.

Gugushvili, S., van der Meulen, F. H., Schauer, M. R. and Spreij, P. (2018). Bayesian wavelet de-noising with the Caravan prior. arXiv:1810.07668, to appear in ESAIM.

Hannah, L.A. and Dunson, D.B. (2011). Bayesian nonparametric multivariate convex regression. arXiv:1109.0322.

Hansen, M.B., and Lauritzen, S.L. (2002). Nonparametric Bayes inference for concave distribution functions. *Statistica Neerlandica.***5**6, p. 110—127.

Hartog J. and van Zanten, H. (2017). Nonparametric Bayesian label prediction on a graph. arXiv:1612.01930 [stat.CO]

Hoffmann, M., Rousseau, J., and Schmidt-Hieber, J. (2015). On adaptive posterior concentration rates. *Ann. Statist.* **4**3, p. 2259—2295.

Hoffman, M. D. and Gelman, A. (2014). The No-U-turn Sampler: Adaptively Setting Path Lengths in Hamiltonian Monte Carlo. *J. Mach. Learn. Res.* **1**5, p. 1593–1623.

Huang, Y. and Louis, T. A. (1998). Nonparametric estimation of the joint distribution of survival time and mark variables. *Biometrika.* **8**5, p. 7856—7984.

Hudgens, M. G., Maathuis, M. H. and Gilbert, P. B. (2007). Nonparametric estimation of the joint dis-tribution of a survival time subject to interval censoring and a continuous mark variable. *Biometrics* **6**3, p. 372—380.

Ibrahim, J.G., Chen, M and Sinha, D. (2001) *Bayesian Survival Analysis.* New York: Springer-Verlag.

Alejandro Jara and Timothy Hanson and Fernando Quintana and Peter Müller and Gary Rosner (2011). DPpackage: Bayesian Semi- and Nonparametric Modeling in R. *J*ournal of Statistical Software, Articles **4**0(5), p. 1–30.

Keiding, N., Begtrup, K., Scheike, T. H. and Hasibeder, G. (1996). Estimation from Current Status Data in Continuous Time. *Lifetime Data Anal.* **2**, p. 119–129.

Slama, R. and Højbjerg Hansen, O.K. and Ducot, B. and Bohet, A. and Sorensen, D. and Allemand, L. and Eijkemans, M.J. and Rosetta, L. and Thalabard, J.C. and Keiding, N. and others. (2012). Estimation of the frequency of involuntary infertility on a nation-wide basis. *Human Reproduction* **2**7, p. 1489–1498.

Krachey, E.C. (2009) Variations on the Accelerated Failure Time Model: Mixture Distributions, Cure Rates, and Different Censoring Scenarios. PhD in Statistics, North Carolina State University.

Kulikov, V.N. and Lopuhaä, H.P. (2006). The behavior of the NPMLE of a decreasing density near the boundaries of the support. *Ann. Statist.* **3**4, p. 742—768.

Maathuis, M. H. and Wellner, J. A. (2008). Inconsistency of the MLE for the joint distribution of intervalcensored survival times and continuous marks. *Scand. J. Statist.* **3**5, p. 83—103.

MacEachern, S.N. (1994). Estimating Normal Means With a Conjugate Style Dirichlet Process Prior. *Communications in Statistics: Simulation and Computation*. **2**3, p. 727–741.

MacEachern, S.N. and Müller, P. (1998). Estimating Mixture of Dirichlet Process Models. *J. Comput. Graph. Statist.* **7**(2), 223–238.

Mariucci, E., Ray, K. and Szabó, B. (2017). A Bayesian nonparametric approach to log concave density estimation. arXiv:1703.09531.

Meyer, M.C. and Woodroofe, M. (2004). Consistent maximum likelihood estimation of a unimodal density using shape restrictions. *Can. J. Statist.* **3**2, p. 55—100.

Moala, F. A. and O'Hagan, A. (2010). Elicitation of multivariate prior distributions: A nonparametric Bayesian approach. *Journal of Statistical Planning and Inference.* **1**40, p. 1635–3758.

Neal, R.M. (2000). Markov Chain Sampling Methods for Dirichlet Process Mixture Models. *Journal of Comp. and Graph. Statist.* **9**, p. 249–265.

Orbanz, P. (2014). *Notes on Bayesian Nonparametrics*. Version: May 16, 2014.

Pollard, D. (1990). *Empirical Processes: Theory and Applications.* Haywood, CA: IMS.

Robert C. P., Elvira V., Tawn N., and Wu C. (2018). Accelerating MCMC algorithms *Wiley Interdiscip. Rev. Comput. Stat.* **1**0, e1435.

Salomond, J.B. (2014). Concentration rate and Consistency of the posterior distribution for selected priors under monotonicity constraints. *Electron. J. Statist.* **8**, p. 1380–1404.

Schick, A., and Yu, Q. (2000). Consistency of the GMLE With Mixed Case Interval-Censored Data. *Journal of Statistics* **2**7, p. 45—55.

Schwartz, L. (1965). On Bayes procedures. Z. *Wahrsch. verw. Gebiete*. **4**, p. 10–26.

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

Shen, X. and Wasserman, L. (2001). Rates of convergence of posterior distributions. *The Annals of Statistics*. **2**9, p. 687–714.

Shively, T.S., Sager, T.W. and Walker, S.G. (2009). A Bayesian approach to nonparametric monotone function estimation. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **7**1, p. 159—175.

Shively, T.S., Walker, S.G. and Damien, P. (2011). Nonparametric function estimation subject to monotonicity, convexity and other shape constraints. *J. Econometrics*. **1**61, p. 166—181.

Susarla, V. and Van Ryzin, J. (1976). Nonparametric Bayesian estimation of survival curves from incomplete observations. *Journal of the American Statistical Association*. **7**1, p. 897–902.

Tokdar, S.T. (2006). Posterior consistency of Dirichlet location-scale mixture of normals in density estimation and regression. *The Indian Journal of Statistics*. **1**37, p. 90–110.

Tokdar, S.T. and Ghosh, J. (2007). Posterior consistency of logistic Gaussian process priors in density estimation. *Journal of statistical planning and inference*. **1**37, p. 34–42.

van de Meent, J. W., Paige, B., Yang, H. and Wood, F. (2018). An Introduction to Probabilistic Programming. arXiv:1809.10756.

van de Geer, S. (2000) *Empirical processes in M-estimation*. Cambridge University Press.

Van der Vaart, A.W. and Ghosal, S. (2017) *Fundamentals of Nonparametric Bayesian Inference*. Cambridge Series in Statistical and Probabilistic Mathematics.

Van der Vaart, A.W. (1998). *Asymptotic Statistics*. Cambridge University Press.

Vardi, Y. (1989). Multiplicative censoring, renewal processes, deconvolution and decreasing density: nonparametric estimation. *Biometrika* **7**6, p.751–761.

Walker, S. (2004). New approaches to Bayesian consistency. *The Annals of Statistics*. **3**2, p. 2028–2043.

Walker, S. and Hjort, N.L. (2001). On Bayesian consistency. *Journal of the Royal Statistical Society*. **6**3, p. 811–821.

Walker, S., Lijoi, A. and Prunster, I. (2007). On rates of convergence for posterior distributions in infinite-dimensional models. *The Annals of Statistics.* **3**5, p. 738–746.

Wasserman, L. (1998). Asymptotic properties of nonparametric Bayesian procedures. *Practical Nonparametric and Semiparametric Bayesian Statistics.* p. 293–304.

Watson, G.S. (1971). Estimating functionals of particle size distributions. *Biometrika* **5**8, p. 483–490.

Wellner, J.A., and Zhan, Y. (1997). A hybrid algorithm for computation of the nonparametric maximum likelihood estimator from censored data. *Journal of the American Statistical Association.* **9**2,p. 945–959.

Wellner, J.A., and Zhang, Y. (2000). Two Estimators of the Mean of a Counting Process With Panel Count Data. *The Annals of Statistics.* **2**8, p. 779—814.

Wehrhahn, C., Jara, A., and Barrientos A.F. (2019). On the small sample behavior of dirichlet process mixture models for data supported on compact intervals.*C*ommunications in Statistics-Simulation and Computation. p. 1–25.

West, M., Müller, P. and Escobar, M.D. (1994). Hierarchical Priors and Mixture Models, With Application in Regression and Density Estimation. *Aspects of Uncertainfy.* New York: Wiley, p. 363–386.

Wickham, H. (2016). ggplot2: Elegant graphics for data analysis, 2nd ediition, Springer-Verlag, New York.

Williamson, R. E. (1956). Multiply monotone functions and their Laplace transforms. *Duke Math. J.* **2**3, p. 189–207.

Woodroofe, M. and Sun, J. (1993). A penalized maximum likelihood estimate of $f(0+)$ when f is non-increasing. *Statist. Sinica* **3**, p. 501–515.

Wu, Y. and Ghosal, S. (2008). Kullback Leibler property of kernel mixture priors in Bayesian density estimation. *Electron. J. Statist.* **2**8, p. 298–331.

Summary

This thesis deals with a number of statistical problems where either censoring or shape-constraints play a role. These problems have mostly been treated from a frequentist statistical perspective. Over the past decades, the Bayesian approach to statistics has gained popularity and this is the approach that is adopted in this thesis. We consider nonparametric statistical models, i.e. models indexed by a parameter that is not of finite dimension. For three different models we investigate the asymptotic properties of the posterior distribution under a frequentist setup. We derive either posterior consistency or posterior contraction rates. Such results are relevant, as these provides a frequentist justification of using point estimators derived from the posterior. Besides theoretical results, we develop computational methods for obtaining draws from the posterior. Overall, this work is at the intersection of the research areas "estimation under shape constraints and censoring", "Bayesian nonparametrics" and "Bayesian computation".

In Chapter 2 we deal with nonparametric estimation of a bounded decreasing density function on $\mathbb{R}^+$ and in particular on estimation the density at zero. It is well know that the maximum likelihood estimator in this model is inconsistent at zero. Any decreasing density can be represented as a scale mixture of uniform densities and hence a prior on the set of decreasing densities can be obtained by endowing the mixing measure with a Dirichlet process prior distribution. For $x > 0$, the rate $(\log n/n)^{2/9}$ is derived for point-wise loss in a recent work by Salomond. For $x = 0$, Salomond's arguments do not show consistency. Under some assumption on the base measure of the Dirichlet process prior, we derive a contraction rate equal to $(\log n/n)^{2/9}$ (up to log factors) that coincides with the case $x > 0$. Besides, we investigate empirically the rate of convergence of the Bayesian procedure for estimating the density at zero. This investigation suggests that under specific conditions on both the underlying density and the base measure, it is conceivable that the optimal rate $n^{-1/3}$ is attained by the posterior mean. In a simulation study, we compare the performance of previously introduced frequentist methods and our Bayesian procedure.

In Chapter 3, we study a Bayesian estimation of the event time distribution based on mixed-case interval censored data. It is additionally assumed that the distribution function is concave. We address this problem from a theoretical per-

spective and provide weak conditions on the prior such that the resulting posterior is consistent. The proof relies on Schwartz's method for proving posterior consistency. We also provide computational methods for drawing from the posterior and illustrate the performance of the Bayesian method in both a simulation study and two real datasets.

Finally, in Chapter 4 we consider the the current status continuous mark model where we aim to estimate the joint distribution function of event time and mark variable. For this model, the mle is inconsistent. Within the Bayesian approach, we introduce two histogram type priors for which we derive posterior contraction rates. Using the general theory introduced in chapter 1, we derive that this rate is upper bounded by $n^{-1/9}$ under Hölder smoothness assumptions of the true distribution function. We propose computational methods for obtaining draws from the posterior under both priors. For one prior this is a data-augmentation algorithm, whereas for the other one we use probabilistic programming software that is based on Hamiltonian Monte Carlo methods.

Samenvatting

Dit proefschrift behandeld een aantal statistische problemen waar censuring en vormrestricties een rol spelen. Deze problemen zijn tot op heden voornamelijk beschouwd vanuit de frequentistische statistiek. Wij behandelen deze problemen vanuit de Bayesiaanse statistiek, een aanpak die gedurende de afgelopen 20 jaar aan populariteit heeft gewonnen. We gebruiken niet-parametrische statistische modellen, dat wil zeggen, modellen die niet geïndexeerd zijn door een eindig-dimensionale parameter. Voor drie verschillende modellen onderzoeken we de aymptotische eigenschappen van de aposteriori verdeling onder frequentistische aannamen. We bewijzen consistentie en leiden convergentiesnelheden af. Zulke resultaten zijn relevant omdat ze frequentistische validatie geven van puntschatters die gebaseerd zijn op de aposteriori verdeling. Afgezien van theoretische resultaten ontwikkelen we ook computationale methoden om trekkingen uit de aposteriori verdeling te genereren. In zijn geheel ligt dit proefschrift op het grensvlak van de onderzoeksvelden "schatten onder vormrestricties en censurering", "niet parameterische Bayesiaanse methoden" en "Bayesiaanse computationele methoden".

In hoofdstuk 2 beschouwen we het niet-parametrisch schatten van een begrensde dalend dichtheid op $\mathbb{R}^+$ en in het bijzonder het schatten van de dichtheid in nul. Het is bekend dat de meest aannemelijke schatter in dit model inconsistent is in nul. Iedere dalende dichtheid kan gerepresenteerd worden door een schaal-mengsel van uniforme dichtheden en daarom kan een apriori verdeling op de verzameling van dalende dichtheden verkregen worden door de maat op het schaal-mengsel van een "Dirichlet process prior" te voorzien. Indien $x > 0$, dan is recent door Salomond aangetoond dat voor puntsgewijze verliesfunctie de convergentiesnelheid $(\log n/n)^{2/9}$ is. Als $x = 0$, dan kan consistentie niet geconcludeerd worden op grond van Salomond's argumenten. Onder een zekere conditie op de "base measure" van het Dirichlet proces leiden we convergentiesnelheid $(\log n/n)^{2/9}$ af (afgezien van log factoren), wat overeenkomt met het geval $x > 0$. Bovendien onderzoeken we empirisch de convergentiesnelheid onder de Bayesiaanse aanpak voor het schatten van de dichtheid in nul. Dit onderzoek suggereert dat onder specifieke condities op zowel de onderliggende dichtheid als ook de "base measure" het aannemelijk is dat de optimale snelheid $n^{-1/3}$ behaald wordt door de aposteriori verwachting. In een simulatiestudie vergelijken we de kwaliteit van een aantal frequentistische meth-

oden uit de literatuur en de voorgestelde Bayesiaanse aanpak, onder een aantal keuzen voor de base measure.

In hoofdstuk 3 bestuderen we het Bayesiaans schatten van een tijdsduur gebaseerd op zogenaamde "mixed-case interval gecensureerde data". Een extra aanname is dat de verdelingsfunctie van deze tijdsduur concaaf wordt verondersteld. We beschouwen dit probleem vanuit theoretisch perspectief en leiden condities af op de apriori verdeling zodat de aposteriori verdeling consistent is. Het bewijs is gebaseerd op Schwartz's methode voor het bewijzen van consistentie. We geven ook computationele methoden om uit de aposteriori verdeling te trekken en illustreren de kwaliteit van de Bayesiaanse methode in zowel een simulatie studie als ook door toepassing op twee dataverzamelingen.

Ten slotte beschouwen we in hoofdstuk 4 het "current status continuous mark" model waar we beogen de de gezamenlijke kansverdeling van de tijdsduur en "mark"-variabele te schatten. In dit model is de meest aannemelijke schatter inconsistent. We introduceren twee histogram-priors en leiden voor beide aposteriori convergentiesnelheden af. We laten zien dat de convergentiesnelheid begrensd wordt door orde $n^{-1/9}$ onder een Hölder gladheidsaanname op de echte verdelingsfunctie. Voor beide apriori verdelingen geven we computationele methoden om uit de apposteriori verdeling te kunnen trekken. Voor één van de apriori verdelingen is dit een "data-augmentation" algoritme, voor het de andere gebruiken we "probabilistic programming" welke gebaseerd is op Hamiltoniaanse Monte Carlo methoden.

Acknowledgements

Throughout the my doctoral career I have received a great deal of support and assistance.

I would first like to thank my promotor Geurt Jongbloed and my supervisor Frank van der Meulen, for providing me the opportunity to join TU Delft. Thank you Geurt for your inspiring guidance and enthusiastic support in this research. Thank you Frank for your valuable expertise in the formulating of the research topic and methodology in particular. Words are powerless to express my gratitude. Your broad knowledge, your availability for in-depth discussion, and your constant encouragement throughout this work have all been indispensable for my future life. The years that we worked together in Delft are treasured memories of mine.

I would also like to thank the members of my committee for your attendance and support. I feel very grateful for the friendship that I made while in Netherlands. Thanks to the friends from the Probability and Statistics group at TU Delft for your wonderful collaboration. You supported me greatly and were always willing to help me.

I would like to thank my family: my parents and my brothers for supporting me throughout my life in general. You are always my spiritual support. Lastly, a special thank to my boyfriend, Shaoxiong Hu. Thank you for your kind temper and patience, especially you always encouraging me and giving me confidence .

Curriculum Vitae

Lixue Pang

14 August 1994 Born in Anhui, China.

Education

2009 – 2013 BSc in Applied Mathematics, Department of Mathematics, Anhui University, China.

2013 – 2015 MSc in Applied Probability, Department of Mathematics, University of Science and Technology of China (USTC), China.
Thesis: Stochastic Analysis on Manifolds with Application in Radially Symmetric Manifolds
Supervisor: Prof. E.P. Hsu

2015 – 2020 PhD in Statistics, Department of Applied Mathematics Delft University of Technology (TUD).
Thesis: Bayesian nonparametric estimation with shape constraint
Promotor: Prof. dr. ir. G. Jongbloed
Supervisor: Dr. ir. F.H van der Meulen