POSTERIOR CONSISTENCY IN SOME BAYESIAN NONPARAMETRIC PROBLEMS

By

Srikanth K. Rajagopalan

A DISSERTATION

Submitted to
Michigan State University
in partial fulfillment of the requirements
for the degree of

DOCTOR OF PHILOSOPHY

Department of Statistics and Probability

1997

ABSTRACT

POSTERIOR CONSISTENCY IN SOME BAYESIAN NONPARAMETRIC PROBLEMS

By

Srikanth K. Rajagopalan

Issues regarding posterior consistency in Bayesian inference are of interest both to frquentists as well as Bayesians. In this dissertation we study different notions of posterior consistency in some Bayesian nonparametric problems, using Dirichlet process and Polya tree process priors.

The first part of the dissertation deals with construction of priors (that yield consistent posteriors) for the class of all distributions symmetric about a point. We consider two natural methods of constructing priors for symmetric distributions, and study the priors obtained by the two methods using Dirichlet processes and Polya tree processes.

The second part deals with the Bayesian analysis of right censored data under a nonparametric formulation. We study different Bayesian approaches to this problem with emphasis on the approaches of Susarla and Van Ryzin (1976) and Tsai (1986), who both use Dirichlet process priors. We establish the posterior consistency for both the approaches and also generalize some of the results to include Polya tree priors as

well.

The Bayesian analysis of interval censored data (again under a nonparametric formulation) is studied in the last part of the dissertation. This portion is rather tentative and we mainly highlight the difficulties in trying to adapt the approaches of Susarla and Van Ryzin (1976), and Tsai (1986) to this problem.

To Amma and Appa; Professor A. M. Goon; Arupda

ACKNOWLEDGMENTS

I would like to express my sincere gratitude to my dissertation advisor, Professor R. V. Ramamoorthi, for his constant help, advice, encouragement, guidance, mentorship and extreme patience. His caring personality, friendly nature, and excellent sense of humour and wit made the whole doctoral experience enjoyable even during trying circumstances.

I would also like to thank, Professors James Hannan, Joseph Gardiner, V. Mandrekar and Habib Salehi for serving on my guidance committee, Professors Hannan and Gardiner for their encouragement, suggestions and many helpful conversations, and Professor V. Mandrekar for useful suggestions on improving the presentation. I would also like to thank Professor J. K. Ghosh for many informal discussions and suggestions.

I cannot thank my parents and sisters enough, for the support and encouragement provided by them during my entire student life. This has been the main motivating force behind all my endeavours and will always be fondly remembered and cherished. I would also like to thank Professor A. M. Goon for the care and interest he showed in my progress as a student, during my undergraduate days, and for the encouragement to pursue graduate studies in Statistics. The help and guidance received from Dr. Arup Kr. Pal during my student days at the Indian Statistical Institute was instrumental in igniting my interest in Probability theory and eventually Mathematical Statistics. Last, but not the least, I am thankful to my Beloved Lord, Bhagavan Sri Sathya Sai Baba for all His Love and Grace, without which, this would not have been possible.

[Major portion of this research was supported by the National Institute of Health Grant 1 RO1 GM49374.]

TABLE OF CONTENTS

CHAPTER 0

An overview

In any statistical experiment, data is collected following a probability model with an unknown parameter θ, lying in a parameter space Θ. The problem of statistical inference deals with drawing meaningful conclusions about θ, given the data. A Bayesian would use a prior probability measure on Θ, representing her/his prior belief/opinion. Given the data, the posterior represents the updated belief/opinion for the Bayesian. Since all Bayes procedures are based on the posterior it is quite natural to require that as more and more data become available, the posterior should concentrate more and more around the true parameter. This idea is formalized as the notion of posterior consistency, which has both Bayesian and frequentist interpretations. Priors that yield consistent posteriors ensure that the data eventually swamps the prior and opinions based on very different priors will merge as the data accumulates. Doob (1948) proved a very general result on consistency, which guarantees that the posterior will be consistent for all θ except on a set of prior measure zero. When Θ is finite dimensional, Freedman (1963), and Schwartz (1965), show that under fairly general conditions, the posterior is consistent at all θ. Freedman (1963) also constructs an example which shows that posterior consistency will not always hold when Θ is the set of all probability measures on the space of positive integers.

Problems of statistical inference with an infinite dimensional parameter space are

of great importance, both theoretically and practically. The Bayesian approach to such nonparametric problems requires the study of (prior and posterior) probability measures on the space of all probability distributions over a set. Freedman's (1963) example shows that posterior consistency may not always hold when Θ is infinite dimensional. Diaconis and Freedman (1986a) and the ensuing discussions highlight the need for a careful study of posterior consistency in nonparametric and semi-parametric problems. In this dissertation we focus on issues concerning different notions of posterior consistency in some nonparametric problems within a Bayesian formulation. Some of the problems that we study are made more complicated because of the fact that we only have censored data.

In Chapter 1, we begin with an introduction to general Bayesian inference and different notions of consistency. We then review and discuss some of the properties of two important families of priors used in Bayesian nonparametrics, namely the Dirichlet processes [Ferguson (1973)], and its generalization, the Polya tree processes [Mauldin et al. (1992), Lavine (1992, 1994)]. We also prove a convergence result for Dirichlet processes that enables us to establish a strong form of consistency for the posterior of a Dirichlet process.

Chapter 2 studies the problem of constructing a family of priors for problems where the parameter set is the space of all distributions symmetric about an arbitrary point on the real line, which we denote by $M^S(\mathbb{R})$. This problem has been studied by Dalal (1979), who constructs a class of priors using Dirichlet process priors, which has been used in the context of the location problem by Diaconis and Freedman (1986). We consider two natural methods of constructing a prior on $M^S(\mathbb{R})$, and study the behaviour of the posterior under the two methods, using both Dirichlet processes and Polya tree processes. We show that using appropriate Dirichlet processes, the two methods yield the same prior on $M^S(\mathbb{R})$, while using appropriate Polya tree processes yield different priors on $M^S(\mathbb{R})$, unless the Polya tree processes being considered

are Dirichlet processes. We also establish the posterior consistency for both the approaches.

In Chapter 3, we consider two different approaches to Bayesian inference with right censored data. Susarla and Van Ryzin (1976), first considered this problem in a Bayesian set-up by considering a Dirichlet process prior for F, the distribution function of interest. They obtain a Bayes estimate and show that this estimate converges to the usual product limit estimate of Kaplan and Meier (1958). Blum and Susarla (1977), complemented this result by proving that the posterior distribution given the right censored data is a mixture of Dirichlet processes. We show that the posterior can be represented as a Polya tree process, a representation which clarifies some of the calculations in Susarla and Van Ryzin (1976). Using this Polya tree representation for the posterior, we then are able to establish the posterior consistency for this approach. Yet another approach to Bayesian inference with right censored data, is to consider priors for the observable random variables as studied by Tsai (1986), who considers a Dirichlet process prior for the distribution of the observable random variables. Under this approach, using a result from Peterson(1977), we are able to establish consistency of the posterior for a wide class of priors.

Chapter 4 is somewhat tentative. Here we consider the Bayesian analysis of the interval censoring problem with single inspection time. We began this study with a goal of obtaining a Bayesian interpretation of the well known Turnbull estimator (1976), which can also be thought of as the nonparametric maximum likelihood estimator (NPMLE). Similar to Chapter 3, here also we look at two different approaches. We highlight the fact that approaches similar to the ones that yield interesting results in the right censoring problem do not yield interesting results in this case. In the first approach we consider a Dirichlet process prior for F, the distribution of interest and study the limiting behaviour of the Bayes estimate. As pointed out in Wang (1993), the NPMLE is not necessarily the limit of the Bayes estimates. We present a set of

examples which show that no obvious relationship connects the limiting Bayes estimate and the NPMLE. We also make an attempt to study consistency properties of the posterior, when we consider priors for the distribution of the observable random variables. Unfortunately, the result that we have in this context, though mathematically nice, is not statistically very useful.

CHAPTER 1

Preliminaries

1.1 General Bayesian inference and posterior consistency

Consider a family of probability measures $\{ Q_\theta : \theta \in \Theta \}$ on a measurable space $(\mathcal{X}, \mathcal{A})$. We view $(\Theta, \mathcal{B})$ as a measurable space such that $Q_\theta(A)$ is $\mathcal{B}$-measurable for every $A \in \mathcal{A}$. We write Q_θ^∞ for the product measure on $\mathcal{X}^\infty$ which makes the coordinate random variables $X_1, X_2, \ldots$, independent with common distribution Q_θ. In general $\mathcal{X}$ and Θ are Borel subsets of complete separable metric spaces. (In this dissertation $\mathcal{X}$ will either be the real line or the positive half line, and Θ will be the set of all probability measures thereon.) Let μ be a prior probability measure on Θ, and let P_μ denote the joint distribution of the parameter and the data:

$$P_\mu(B \times A) = \int_B Q_\theta^\infty(A)\mu(d\theta)$$

for $B \in \mathcal{B}$, and $A \in \mathcal{A}^\infty$. The posterior is the P_μ -distribution of the parameter θ given the data $X_1, X_2, \ldots, X_n$, and is formally defined below. We denote this by $\mu_n(\cdot \mid X_1, X_2, \ldots, X_n)$.

Definition 1.1.1 $\mu_n(\cdot \mid \cdot) : \mathcal{B} \times \mathcal{X}^n \to [0,1]$ *is called a posterior distribution given* $X_1, X_2, \ldots, X_n$ *if,*

1. *For each* $(X_1, X_2, \ldots, X_n) \in \mathcal{X}^n$, $\mu_n(\cdot \mid X_1, X_2, \ldots, X_n)$ *is a probability measure on* $(\Theta, \mathcal{B})$.

2. *For each* $B \in \mathcal{B}$, $\mu_n(B \mid \cdot)$ *is* $\mathcal{A}^n$ *measurable*

3. *For every* $B \in \mathcal{B}, A \in \mathcal{A}^n$

$$P_\mu^n(B \times A) = \int_A \mu_n(B \mid X_1, X_2, \ldots, X_n) dP^n(X_1, X_2, \ldots, X_n),$$

where $P_\mu^n(B \times A) = P_\mu(B \times (A \times \mathcal{X}^\infty))$ *and* $P^n(A) = P_\mu^n(\Theta \times A)$.

The posterior distribution is of course unique only up to P^n null sets. In situations we consider, there is a natural candidate for the posterior and we will generally refer to it as 'the posterior'.

(For the Bayesian) The posterior distribution encapsulates all that is known about θ following the observation of the data $X_1, X_2, \ldots, X_n$, and one would want the posterior to concentrate around the true value of the parameter as more and more data become available. The main topic of study in this dissertation is the consistency property of the posterior sequence $\{\mu_n(\cdot \mid X_1, X_2, \ldots, X_n)\}_{n \geq 1}$ in certain Bayesian nonparametric settings. The sequence of posteriors $\{\mu_n(\cdot \mid X_1, X_2, \ldots, X_n)\}_{n \geq 1}$ is said to be consistent at $\theta_0 \in \Theta$ if, whenever θ_0 is the true value of the parameter θ, as observations accumulate the effect of the prior diminishes and the posterior gets closer and closer to the 'true' prior δ_{θ_0} - the degenerate prior at θ_0. (More formal definitions of posterior consistency will be mentioned later.) Posterior consistency has both Bayesian and frequentist interpretations, and for a detailed discussion of this notion of consistency, especially in a nonparametric set-up, the interested reader is referred to Diaconis and Freedman (1986a) [pages 3, 4, 10-20].

1.2 Probability measures on probability measures

Throughout this dissertation $\mathbb{R}$ will denote the real line, $\mathcal{B}(\mathbb{R})$ will denote the Borel σ-algebra of $\mathbb{R}$, and $M(\mathbb{R})$ will denote the space of all probability measures on $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$. Also, $\mathbb{R}^+$ will denote the positive half line, with $\mathcal{B}(\mathbb{R}^+)$ and $M(\mathbb{R}^+)$ having an analogous interpretation.

On $M(\mathbb{R})$, we consider the smallest σ-algebra that makes the map $P \mapsto P(B)$, measurable for each Borel set $B \in \mathcal{B}(\mathbb{R})$. We denote this σ-algebra by $\mathcal{B}_M$, i.e. $\mathcal{B}_M = \sigma\{P(B) : B \in \mathcal{B}(\mathbb{R})\}$. Since the elements of $M(\mathbb{R})$ are functions on $\mathcal{B}(\mathbb{R})$ taking values in $[0,1]$, $M(\mathbb{R})$ can be viewed as a subset of $[0,1]^{\mathcal{B}(\mathbb{R})}$. If the product space $[0,1]^{\mathcal{B}(\mathbb{R})}$ is equipped with the product σ-algebra (the smallest σ-algebra that makes all the coordinate functions measurable), the restriction of this σ-algebra to $M(\mathbb{R})$ is $\mathcal{B}_M$. However, $M(\mathbb{R})$ is not a measurable subset of $[0,1]^{\mathcal{B}(\mathbb{R})}$. Therefore one needs to be careful in constructing probability measures on $M(\mathbb{R})$. The following two theorems implicit in Ferguson (1973), and mentioned in Ghosh and Ramamoorthi (1996-97), give a way of constructing and defining probability measures on $M(\mathbb{R})$.

Theorem 1.2.1 . *Suppose for each collection $\{B_1, B_2, \ldots, B_k\}$ of subsets of $\mathbb{R}$, a distribution $\mu_{B_1,\ldots,B_k}$ is assigned for $(P(B_1), \ldots, P(B_k))$ such that:*

1. *If $\{A_1, A_2, \ldots, A_l\} \subset \{B_1, B_2, \ldots, B_k\}$, then the marginal distribution of $(P(A_1), \ldots, P(A_l))$ derived from $\mu_{B_1,\ldots,B_k}$ is $\mu_{A_1,\ldots,A_l}$.*

2. *For every partition $\{B_1, B_2, \ldots, B_k\}$ of $\mathbb{R}$, $\mu_{B_1,\ldots,B_k}$ is a probability measure on $S_k = \{(p_1, \ldots, p_k) : p_i \geq 0, \sum p_i = 1\}$ and further if A_i is an union of sets from $\{B_1, B_2, \ldots, B_k\}$, then $\mu_{A_1,\ldots,A_n}$ = distribution of $(\sum_{B_i \subset A_1} P(B_i), \ldots, \sum_{B_i \subset A_n} P(B_i))$*

3. *If $A_n \downarrow \phi$, then $P(A_n) \downarrow 0$ in distribution.*

Then, there exists a probability measure μ on $M(\mathbb{R})$ such that the distribution of $(P(B_1), \ldots, P(B_k))$ under μ is $\mu_{B_1,\ldots,B_k}$.

[The proof is taken from Ghosh and Ramamoorthi (1996-97), and is mentioned here for the sake of completeness.]

Proof: Using 1. and 2., it follows from Klomogorov 's consistency theorem that there exists a probability measure on $[0,1]^{\mathcal{B}}$ with finite dimensional marginals given by $\mu_{B_1,B_2,\cdots,B_k}$. Since $M(\mathbb{R})$ is not a measurable subset of $[0,1]^{\mathcal{B}}$, it is not easy to show that this measure is supported by $M(\mathbb{R})$. So we take an indirect route.

Let $\mathcal{F}$ be the set of all distribution functions on $\mathcal{R}$, and let $\mathcal{F}^*$ be the restriction of functions in $\mathcal{F}$ to a countable dense set Q, say the rationals. Then
$\mathcal{F} = \{F\colon\ F$ is monotone, right continuous, $\lim_{t\to-\infty} F(t) = 0, \lim_{t\to\infty} F(t) = 1\ \}$

and

$\mathcal{F}^* = \{F : F$ is monotone, right continuous on Q, $\lim_{t\to-\infty} F(t) = 0$,
$\lim_{t\to\infty} F(t) = 1\ \}$.

Take any $t_1 < t_2 \cdots < t_k$ in Q. Set the distribution of $(F(t_1), F(t_2), \cdots, F(t_k))$ as the distribution of $(P(-\infty, t_1], P(-\infty, t_2], \cdots, P(-\infty, t_k])$. This assignment gives a consistent specification and hence there exists a probability measure μ on $[0,1]^Q$ with these marginals. We now argue that $\mu(\mathcal{F}^*) = 1$.

It is easy to see that for any fixed $t_1 < t_2, F(t_1) < F(t_2)$ with μ probability 1. Since Q is countable $\mu\{F : F$ is monotone in Q $\}$=1. Condition 3. gives that F is right continuous on Q with probability 1, so that $\mu(\mathcal{F}^*) = 1$.

Let the map $\phi : \mathcal{F} \to \mathcal{F}^*$ be the restriction of F in $\mathcal{F}$ to Q. Since this map is 1-1, onto and measurable, the probability on $\mathcal{F}^*$ can be transferred to a probability measure on $\mathcal{F}$. Under this measure $(P(B_1), P(B_2), \cdots, P(B_k)$ has the marginal distribution $\mu_{B_1,B_2,\cdots,B_k}$ whenever B_i is of the form $(-\infty, t_i], t_i \in Q$.

An usual induction argument shows that the statement holds for all borel sets.

$\Diamond$

Theorem 1.2.2 stated below shows that it is enough to specify $\mu_{B_1,\ldots,B_k}$ for every partition $B_1, B_2, \ldots, B_k$ of $\mathbb{R}$.

Theorem 1.2.2 *Suppose the following two conditions hold:*

1. *For every finite partition $B_1, B_2, \ldots, B_k$ of $\mathbb{R}$, $(P(B_1), \ldots, P(B_k))$ has a distribution $\mu_{B_1,\ldots,B_k}$ on S_k.*

2. *If $B_1, B_2, \ldots, B_k$ and $A_1, A_2, \ldots, A_n$ are two partitions of $\mathbb{R}$, such that each A_i is a union of some B_js, then $\mu_{A_1,\ldots,A_n}$ is the same as the $\mu_{B_1,\ldots,B_k}$ distribution of $(\sum_{B_i \subset A_1} P(B_i), \ldots, \sum_{B_i \subset A_n} P(B_i))$.*

For any collection $A_1, A_2, \ldots, A_n$ of subsets of $\mathbb{R}$, take any partition $B_1, B_2, \ldots, B_k$ of $\mathbb{R}$ such that each A_i is a union of some B_js, and define $\mu_{A_1,\ldots,A_n}$ as the $\mu_{B_1,\ldots,B_k}$ distribution of $(\sum_{B_i \subset A_1} P(B_i), \ldots, \sum_{B_i \subset A_n} P(B_i))$.

Then, $\{\mu_{A_1,\ldots,A_n}$: $A_i \in \mathcal{B}(\mathbb{R})$, $i = 1, 2, \ldots, n$; $n = 1, 2, \ldots\}$ satisfies condition 1 of Theorem 1.2.1.

Remarks: We will see later how Theorems 1.2.1 and 1.2.2 are used to define the most commonly mentioned prior in Bayesian nonparametrics, called the Dirichlet process. Another way of defining a probability measure on $M(\mathbb{R})$ is via probability measures on the space of all probability measures on sequences, and the Polya tree process discussed later is an example of one such prior.

1.3 Topologies on the space of probability measures

A major focus of this dissertation is on issues related to posterior consistency in nonparametric problems. Thus the parameter space is $M(\mathbb{R})$, and the sequence of posteriors $\{\mu_n(\cdot \mid X_1, X_2, \ldots, X_n)\}_{n\geq 1}$ is a sequence of probability measures on $M(\mathbb{R})$. Since the notion of consistency involves convergence of probability measures on $M(\mathbb{R})$, we next look at some of the commonly considered modes of convergence on $M(\mathbb{R})$, and later present the corresponding notions of convergence on the space of probability measures on $M(\mathbb{R})$.

Weak Topology: The first notion we look at is weak convergence arising from the usual weak convergence on $M(\mathbb{R})$. We recall that on $M(\mathbb{R})$ weak convergence is defined as: Let $\{\{P_n\}_{n\geq 1}, P\} \subset M(\mathbb{R})$. P_n is said to converge weakly (or in the weak topology) to P if $\int f dP_n \to \int f dP$ for all bounded continuous functions f on $\mathbb{R}$.

For any $P_0 \in M(\mathbb{R})$, sets of the form

$$U_{P_0} = \{P : \mid \int f_i dP - \int f_i dP_0 \mid < \delta_i; i = 1, \ldots, k\},$$

where for each i, f_i is a bounded continuous function on $\mathbb{R}$, constitute a base of open neighbourhoods for P_0 under the weak topology.

It is well known that under this topology $M(\mathbb{R})$ is a complete separable metric space, with $\mathcal{B}_{\mathcal{M}}$ as its Borel σ-algebra [Parthasarathy (1967), Chapter II, Section 6].

Kolmogorov Metric: The Kolmogorov metric on $M(\mathbb{R})$ is defined as follows:

$$d_k(P, Q) = sup_{t\in\mathbb{R}} \mid P(-\infty, t] - Q(-\infty, t] \mid .$$

Interest on this metric stems from the 1-1 correspondence between probability

measures on $\mathbb{R}$ and the cumulative distribution functions, and the Glivenko-Cantelli theorem on convergence of empirical distribution functions. Under the metric d_k, $M(\mathbb{R})$ is neither separable not complete.

Total Variation Metric: The total variation metric d_t on $M(\mathbb{R})$ is defined as

$$d_t(P,Q) = sup_{B\in\mathcal{B}(\mathbb{R})} \mid P(B) - Q(B) \mid .$$

This metric is uninteresting in the context of all of $M(\mathbb{R})$. However, when the parameter space is restricted to subsets of $M(\mathbb{R})$ of the form

$$L_1(\mu) = \{\text{all probability measures on} M(\mathbb{R}) \text{ dominated by a } \sigma - \text{ finite measure } \mu\},$$

d_t is extremely useful, and has a nice form. For $\{P,Q\} \subset L_1(\mu)$, $d_t(P,Q) = \frac{1}{2}\int \mid \frac{dP}{d\mu} - \frac{dQ}{d\mu} \mid d\mu$. Further, $L_1(\mu)$ equipped with d_t is a complete separable metric space.

1.4 Convergence of probability measures and posterior consistency

As noted earlier, $M(\mathbb{R})$ when equipped with the weak convergence metric becomes a complete separable metric space with $\mathcal{B}_{\mathcal{M}}$ as its Borel σ-algebra. Thus a natural topology on the space of probability measures on $M(\mathbb{R})$ is the weak topology arising from this metric on $M(\mathbb{R})$. A formal definition is given below:

Definition 1.4.1 *A sequence of probability measures $\{\mu_n\}$ on $M(\mathbb{R})$ is said to converge weakly to a probability measure μ (on $M(\mathbb{R})$) if*

$$\int \phi(P)d\mu_n(P) \to \int \phi(P)d\mu(P),$$

for all bounded continuous functions ϕ on $M(\mathbb{R})$, and we write $\mu_n \rightarrow^w \mu$ or $\mu_n \Rightarrow \mu$.

Under this convergence, the space of probability measures on $M(\mathbb{R})$ also becomes a complete separable metric space [Parthasarathy (1967), Chapter II, Section 6]. A detailed study of weak convergence requires an understanding of the continuous (in the weak topology) functions on $M(\mathbb{R})$. But, we will mainly be interested in the case when $\mu = \delta_{P_0}$, for some $P_0 \in M(\mathbb{R})$. Since convergence in distribution of μ_n to δ_{P_0}, is equivalent to convergence in probability of P_n to P_0, where $P_n \sim \mu_n$, this convergence can be described in terms of the continuous functions on $\mathbb{R}$ rather than those on $M(\mathbb{R})$, as mentioned in the following proposition.

Proposition 1.4.1 *$\mu_n \rightarrow^w \delta_{P_0}$ if $\mu_n(U_{P_0}) \rightarrow 1$ for all U_{P_0} of the form*
$U_{P_0} = \{P :\mid \int f_i dP - \int f_i dP_0 \mid < \delta_i; i = 1, \ldots, k\}$, where for each i f_i is a bounded continuous function on $\mathbb{R}$.

The non separability of $M(\mathbb{R})$ with either the Kolmogorov metric d_k or the total variation metric d_t prevents the induction of a natural topology on $M(M(\mathbb{R}))$, when $M(\mathbb{R})$ is equipped with either d_k or d_t. However Proposition 1.4.1 still enables us to speak of 'convergence' of μ_n to δ_{P_0} in the sense that as $n \rightarrow \infty$, μ_n concentrates more and more around P_0, and these are formally mentioned in the definitions below.

Definition 1.4.2 *A sequence of probability measures $\{\mu_n\}$ on $M(\mathbb{R})$ is said to converge to δ_{P_0} on uniform (total variation) neighbourhoods if*
$\mu_n(P : d_t(P, P_0) < \delta) \rightarrow 1$, for all $\delta > 0$, and we write $\mu_n \rightarrow^t \delta_{P_0}$.

Definition 1.4.3 *A sequence of probability measures $\{\mu_n\}$ on $M(\mathbb{R})$ is said to converge to δ_{P_0} on k-neighbourhoods if $\mu_n(P : d_k(P, P_0) < \delta) \rightarrow 1$, for all $\delta > 0$, and we write $\mu_n \rightarrow^k \delta_{P_0}$.*

Note: The last two notions of convergence provide for a stronger sense of convergence than the weak convergence of Definition 1.4.1 (and Proposition 1.4.1).

Most of our discussion will focus on weak convergence and hence on Proposition 1.4.1. We will on occasion consider convergence in k-neighbourhoods. Convergence on uniform neighbourhoods will in general not be relevant to our discussion.

We now formally define the notion of posterior consistency under the same set-up mentioned in Section 1.1.

Definition 1.4.4 *The sequence of posteriors* $\{\mu_n(\cdot \mid X_1, X_2, \ldots, X_n)\}_{n\geq 1}$ *is said to be*

1. *weakly consistent at* P_0, *if* $\{\mu_n(\cdot \mid X_1, X_2, \ldots, X_n)\} \rightarrow^w \delta_{P_0}$ *a.s.* P_0, *and*

2. *k-consistent at* P_0, *if* $\{\mu_n(\cdot \mid X_1, X_2, \ldots, X_n)\} \rightarrow^k \delta_{P_0}$ *a.s.* P_0, *and*

3. *t-consistent at* P_0, *if* $\{\mu_n(\cdot \mid X_1, X_2, \ldots, X_n)\} \rightarrow^t \delta_{P_0}$ *a.s.* P_0.

We end this section by mentioning a result which is used quite a lot in proving weak consistency of a sequence of posteriors. Throughout this dissertation for any $\mu \in M(M(\mathbb{R}))$, $\bar{\mu} \in M(\mathbb{R})$ will denote the probability measure defined as follows: $\bar{\mu}(A) = E_\mu(P(A))$, for all $A \in \mathcal{B}(\mathbb{R})$.

Proposition 1.4.2 *Let* $\{\mu_n\}_{n\geq 1} \subset M(\mathbb{R})$, *be such that* $\{\bar{\mu}_n\}_{n\geq 1}$ *is tight as a family of probability measures on* $\mathbb{R}$, *then,* $\{\mu_n\}_{n\geq 1}$ *is a tight family of probability measures (with respect to weak convergence) on* $M(\mathbb{R})$.

Proof: The proof is along the same lines as that of Theorem 3.1 of Sethuraman and Tiwari (1982), and is mentioned here for the sake of completeness.

Fix $\delta > 0$. By the tightness of $\{\bar{\mu}_n\}_{n\geq 1}$, for every positive integer d there exists a sequence of compact sets K_d in $\mathbb{R}$, such that $sup_n \bar{\mu}_n(K_d^c) \leq \frac{6\delta}{d^3\pi^2}$.

For $d = 1, 2, \ldots$, let, $M_d = \{P \in M(\mathbb{R}) : P(K_d^c) \leq \frac{1}{d}\}$, and let $M = \cap_d M_d$. Then by its very definition M is a compact subset of $M(\mathbb{R})$, in the weak topology. Further,

by Markov's inequality,

$$\begin{aligned}\mu_n(M_d^c) &\leq dE_{\mu_n}(P(K_d^c)) \\ &= d\bar{\mu}_n(K_d^c) \\ &\leq \frac{6\delta}{d^3\pi^2}\end{aligned}$$

Hence, for any $n = 1, 2, \ldots$, $\mu_n(M) \leq \sum_d \frac{6\delta}{d^3\pi^2} = \delta$.

By Theorem 6.7, on page 47 of Parthasarathy (1967), this proves that $\{\mu_n\}_{n\geq 1}$ is tight. $\diamondsuit$

In the next two sections we introduce the two families of priors that are used in the problems considered in this dissertation, namely the Dirichlet processes and Polya tree processes.

1.5 Dirichlet processes

Dirichlet processes were formally introduced by Ferguson (1973, 1974), who mentions many of its basic properties, and also applies it to a variety of nonparametric problems. In the process a Bayesian interpretation for some of the commonly used nonparametric procedures were provided for the first time. Dirichlet processes arise naturally as an infinite dimensional analogue of the finite dimensional Dirichlet distribution, which itself is the multivariate generalization of the Beta distribution. Here we restrict ourselves to stating its definition and mentioning some of its basic properties. For a detailed account we refer the interested reader to Ferguson (1973, 1974), Schervish (1995), and Ghosh and Ramamoorthi (1997).

Definition 1.5.1 *Let α be a finite non-null measure on $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$. A (prior) probability measure $I\!P$ on $M(\mathbb{R})$ is said to be a Dirichlet process with parameter (or base measure) α if, for every finite measurable partition $\{B_1, B_2, \ldots, B_k\}$*

of $\mathbb{R}$, *the random vector* $(P(B_1), P(B_2), \ldots, P(B_k))$ *has the Dirichlet distribution* $\mathcal{D}(\alpha(B_1), \alpha(B_2), \ldots, \alpha(B_k))$ *under* $I\!P$.

In particular, for any $A \in \mathcal{B}(\mathbb{R})$, $P(A)$ has the Beta distribution $\mathrm{B}(\alpha(A), \alpha(\mathbb{R}) - \alpha(A))$ under $I\!P$. So, $E_{\mathcal{D}(\alpha)}(P(A)) = \frac{\alpha(A)}{\alpha(\mathbb{R})}$– is the 'prior' guess for $P(A)$.

We view the Dirichlet Process as choosing a probability P randomly according to $\mathcal{D}(\alpha)$ and write it as $\mathrm{P} \in \mathcal{D}(\alpha)$.

The existence of the Dirichlet process can be established using Theorems 1.2.1 and 1.2.2 mentioned earlier. A very clever and elegant construction of the Dirichlet process is given by Sethuraman (1994) and is mentioned in the next theorem. This construction gives an insight into some of the peculiarities of the Dirichlet process, and is an extremely useful tool for simulation purposes. We will make use of this construction in Chapter III.

Theorem 1.5.1 *Let* α *be a finite non-null measure on* $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$. *Let* $\{Y_n\}_{n\geq 1}$ *be an i.i.d. sequence of random variables with* $Y_1 \sim \bar{\alpha}$, *and let* $\{\theta_n\}_{n\geq 1}$ *be an i.i.d. sequence of random variables with* $\theta_1 \sim Beta(1, \alpha(\mathbb{R}))$, *and let* $\{Y_n\}_{n\geq 1}$ *and* $\{\theta_n\}_{n\geq 1}$ *be independent. Define* $P_1 = \theta_1$, *and for* $n \geq 2, P_n = \theta_n \prod_1^{n-1}(1 - \theta_i)$. *Then,* $I\!P = \sum_1^\infty P_j \delta_{Y_j}$ *is a Dirichlet process with parameter* α.

Support [Ferguson (1974), Facts 2. and 3.].

1. If $\mathrm{P} \in \mathcal{D}(\alpha)$, then with probability one P is discrete.

2. The topological support (that is, the smallest closed set with probability one) of $\mathcal{D}(\alpha)$, w. r. t. the topology of weak convergence is the set of all distributions whose (topological) support is contained in the (topological) support of α.

Thus, even though the measure theoretic support is 'small', the topological support is fairly large. For example, if the (topological) support of α is $\mathbb{R}$ then the

(topological) support of $\mathcal{D}(\alpha)$ is all of $M(\mathbb{R})$, and $\mathcal{D}(\alpha)$ gives positive mass to every open set in $M(\mathbb{R})$.

Posterior Distribution [Theorem 1. Ferguson (1973)]. Let P $\in \mathcal{D}(\alpha)$. If, given P, $X_1, X_2, \ldots, X_n$ is a sample from P, then the posterior distribution of P given $X_1, X_2, \ldots, X_n$ is $\mathcal{D}(\alpha + \sum_1^n \delta_{X_i})$, where δ_x is the measure giving mass one to x. Thus just like the (finite dimensional) Dirichlet distribution priors for the vector of proportions in a multinomial model, the Dirichlet processes provide a conjugate family of priors for M($\mathbb{R}$).

Predictive Distribution and Bayes Estimates. Let P $\in \mathcal{D}(\alpha)$ and let $\bar{\alpha}(\cdot) = \frac{\alpha(\cdot)}{\alpha(\mathbb{R})}$. The Bayes estimate (w. r. t. squared error loss) of P(A), given a sample $X_1, X_2, \ldots, X_n$ from P, is

$$\hat{F}_n(A) = E(\mathrm{P(A)} \mid X_1, X_2, \ldots, X_n) = p_n \bar{\alpha}(A) + (1 - p_n) F_n(\mathrm{A}),$$

where $F_n(\cdot)$ denotes the sample (empirical) distribution and $p_n = \frac{\alpha(\mathbb{R})}{\alpha(\mathbb{R})+n}$.

The Bayes estimate $\hat{F}_n$ is thus a linear combination of $\bar{\alpha}$ and the sample distribution function F_n. This Bayes estimate can also be looked upon as the 'Predictive distribution' of a future observation given $X_1, X_2, \ldots, X_n$.

Convergence of Dirichlet processes. The Dirichlet process possesses nice continuity properties with respect to the base measure α. In Propositions 1.5.1 and 1.5.2 we mention two such properties, of which Proposition 1.5.1 is well known, while Proposition 1.5.2 is a new result. (Throughout '$\Rightarrow$' will denote weak convergence and all convergences are as 'n goes to ∞'.)

Also we will write $\bar{\alpha}$ to denote the probability measure defined as $\bar{\alpha}(A) = E_{\mathcal{D}(\alpha)}(P(A))$, for all $A \in \mathcal{B}(\mathbb{R})$.

Proposition 1.5.1 *Let α_n, for n = 1, 2, ... be finite non-null measures on $\mathbb{R}$ such that $\bar{\alpha}_n \Rightarrow P_0$, (where $P_0 \in M(\mathbb{R})$) and $\alpha_n(\mathbb{R}) \to \infty$, then $\mathcal{D}(\alpha_n) \Rightarrow \delta_{P_0}$.*

[We mention the proof to illustrate the general principles behind the weak convergence results proved in this dissertation.]

Proof:

By Proposition 1.4.2, since $\{\bar{\alpha}_n\}_{n\geq 1}$ is tight $\{\mathcal{D}(\alpha_n)\}_{n\geq 1}$, is a tight family of probability measures.

Let f be a bounded continuous function on $\mathbb{R}^+$, with compact support. It is enough to show that $\mathcal{D}(\alpha_n)(V_{Q_0}^{\delta}) \to 1$, where

$$V_{Q_0}^{\delta} = \{P : \mid \int f dP - \int f dP_0 \mid < \delta\}.$$

f bounded continuous, with compact support, implies that there exists a simple function $f_\delta = \sum_{i=1}^{k} a_i I_{A_i}$, such that A_i's are P_0 continuity sets, and $sup_x \mid f(x) - f_\delta(x) \mid < \frac{\delta}{3}$. Noting that

$$\mid \int f dP - \int f dP_0 \mid \;\leq\; \mid \int f_\delta dP - \int f_\delta dP_0 \mid + \frac{2\delta}{3}, \text{ and}$$

$$\int f_\delta dP = \sum_{i=1}^{k} a_i P(A_i).$$

Our proof will be complete if we can show that $E_{\mathcal{D}}(\alpha_n)(P(A_i) - P_0(A_i))^2 \to 0$, and this follows from the fact that $\bar{\alpha}_n(A_i) \to P_0(A_i)$, and $(\bar{\alpha}_n(A_i))^2 \to (P_0(A_i))^2$.

$\diamondsuit$

Proposition 1.5.2 . *Let α_n, for $n = 1, 2, \ldots$ be finite non-null measures on $\mathbb{R}$ such that*

1. $\alpha_n(\mathbb{R}) \to \infty$,

2. $sup_{t\in\mathbb{R}} \mid \bar{\alpha}_n(-\infty, t] - P_0(-\infty, t] \mid \to 0$,

   and $sup_{t\in\mathbb{R}} \mid \bar{\alpha}_n(-\infty, t) - P_0(-\infty, t) \mid \to 0$,

then $\mathcal{D}(\alpha_n) \rightarrow^k \delta_{P_0}$.

Proof: For any $P \in M(\mathbb{R})$, let $P(t) = P(-\infty, t]$, and let $P(t-) = P(-\infty, t)$. We need to show that for any $\epsilon > 0$,

$$\mathcal{D}(\alpha_n)(P : \sup_{t \in \mathbb{R}} \mid P(t) - P_0(t) \mid \geq \epsilon) \rightarrow 0.$$

Let m be a fixed positive integer. Let $\phi(u) := inf\{x : P_0(x) \geq u\}$, and let $x_{m,k} := \phi(k/m)$, for $k = 1, 2, \ldots, m$. We observe that $P_0(\phi(u)-) \leq u \leq P_0(\phi(u))$, and hence, $P_0(x_{m,1}-) \leq 1/m$, $P_0(x_{m,m-1}) \geq (1 - 1/m)$, and for $2 \leq k \leq m$, $(P_0(x_{m,k}-) - P_0(x_{m,k-1})) \leq 1/m$.

Let $1 \leq k \leq m-1$. For $x_{m,k-1} \leq t < x_{m,k}$,

$$\mid P(t) - P_0(t) \mid \leq \mid P(x_{m,k}-) - P_0(x_{m,k-1}) \mid \vee \mid P(x_{m,k-1}) - P_0(x_{m,k}-) \mid,$$

and for $t \geq x_{m,m-1}$,

$$\mid P(t) - P_0(t) \mid \leq (1 - P_0(x_{m,m-1})) \vee (1 - P(x_{m,m-1}).$$

Therefore $sup_{t \in \mathbb{R}} \mid P(t) - P_0(t) \mid \leq B_m$, where

$$B_m = max_k\{B_{m,k}\} \vee (1 - P_0(x_{m,m-1})) \vee (1 - P(x_{m,m-1})), and$$

$$B_{m,k} = \{\mid P(x_{m,k}-) - P_0(x_{m,k-1}) \mid \vee \mid P(x_{m,k-1}) - P_0(x_{m,k}-) \mid\}$$

Hence,

$$\mathcal{D}(\alpha_n)(P : sup_{t \in \mathbb{R}} \mid P(t) - P_0(t) \mid \geq \epsilon) \leq \mathcal{D}(\alpha_n)(P : B_m \geq \epsilon).$$

Let $\delta > 0$, and let N_m be such that, for all $n \geq N_m$,

$$\begin{aligned} \sup_{t\in\mathbb{R}} \mid \bar{\alpha_n}(t-) - P_0(t-) \mid &< \frac{\delta}{2m}, \\ \sup_{t\in\mathbb{R}} \mid \bar{\alpha_n}(t) - P_0(t) \mid &< \frac{\delta}{2m}, \text{ and} \\ \alpha_n(\mathbb{R}) &> \frac{2m}{\delta}. \end{aligned}$$

By Markov's inequality, and our choice of $x_{m,k}s$,

$$\begin{aligned} & \mathcal{D}(\alpha_n)(P :\mid P(x_{m,k}-) - P_0(x_{m,k-1}) \mid \geq \epsilon) \\ \leq\ & E_{\mathcal{D}(\alpha_n)} \frac{(P(x_{m,k}-) - P_0(x_{m,k-1}))^2}{\epsilon^2} \\ \leq\ & \frac{1}{\epsilon^2}[(P_0(x_{m,k}-) - P_0(x_{m,k-1}))^2 + \frac{6\delta}{2m}], \text{for all } n \geq N_m \\ \leq\ & \frac{1}{\epsilon^2}[\frac{1}{m^2} + \frac{6\delta}{2m}] \end{aligned}$$

The second inequality above follows from our choice of N_m and from the fact that for any finite non-null measure α on $\mathbb{R}$, and $t \in \mathbb{R}$,

$$E_{\mathcal{D}(\alpha)}(P(t-))^2 = \bar{\alpha}(t-) \times \frac{\bar{\alpha}(t-) + \frac{1}{\alpha(\mathbb{R})}}{1 + \frac{1}{\alpha(\mathbb{R})}},$$

and

$$E_{\mathcal{D}(\alpha)}(P(t-)) = \bar{\alpha}(t-).$$

Hence,

$$\begin{aligned} & \mathcal{D}(\alpha_n)(P : sup_{t\in\mathbb{R}} \mid P(t) - P_0(t) \mid \geq \epsilon) \\ & \qquad \leq\ \mathcal{D}(\alpha_n)(P : B_m \geq \epsilon) \\ & \qquad \leq\ \frac{1}{\epsilon^2}[\frac{2}{m} + 6\delta], \text{ for all } n \geq N_m. \end{aligned}$$

Since $\delta > 0$ is arbitrary and the last mentioned inequality holds for all m,

$\mathcal{D}(\alpha_n)(P : sup_{t\in\mathbb{R}} \mid P(t) - P_0(t) \mid \geq \epsilon) \to 0.$

$\Diamond$

Posterior consistency: It is well known that a $\mathcal{D}(\alpha)$ prior leads to a posterior that is weakly consistent at all $P_0 \in M(\mathbb{R})$. (This fact follows on observing that the posterior for $\mathcal{D}(\alpha)$ given $X_1, X_2, \ldots, X_n$ is $\mathcal{D}(\alpha + \sum_{i=1}^n \delta_{X_i})$, and then taking $\alpha_n = \mathcal{D}(\alpha + \sum_{i=1}^n \delta_{X_i})$ in Proposition 1.5.1.) The next theorem mentions that in this case, the posterior has the stronger k-consistency property.

Theorem 1.5.2 *Let $P \in \mathcal{D}(\alpha)$, and given P, let $X_1, X_2, \ldots, X_n$ be a sample from P. Then $\mathcal{D}(\alpha)(U_{P_0} \mid X_1, X_2, \ldots, X_n) \to 1$ a.s. P_0, for all k-neighbourhoods U_{P_0} of P_0.*

Proof: We observe that the posterior for $\mathcal{D}(\alpha)$ given $X_1, X_2, \ldots, X_n$ is $\mathcal{D}(\alpha + \sum_{i=1}^n \delta_{X_i})$. The proof now follows from Proposition 1.5.2 by taking $\alpha_n = \alpha + \sum_{i=1}^n \delta_{X_i}$, and a simple application of the Glivenko-Cantelli theorem. $\Diamond$

Critics of the Dirichlet process point to the fact that with probability one, the Dirichlet process selects a discrete distribution, as its major shortcoming. Polya tree processes discussed in the next section, is a family of workable priors which overcome this drawback of the Dirichlet process.

1.6 Polya tree processes

Polya tree priors (or Polya tree processes) are a generalization of Dirichlet processes, and share many of the properties of the Dirichlet processes. These processes are described through a large number of parameters and a suitable choice of these parameters allows the statistician to overcome some of the shortcomings of the Dirichlet

processes. Here also we mention only some of its basic properties, and for a detailed account, the interested reader is referred to Lavine (1992, 1994), Mauldin et.al. (1992), Schervish (1995), and Ghosh and Ramamoorthi (1997).

Let $\pi_0 = \mathbb{R}$ and $\mathbf{\Pi} = \{\pi_m; m = 0, 1, ...\}$, where $\pi_0, \pi_1, \ldots$, is sequence of partitions of $\mathbb{R}$ such that $\mathcal{B} = \sigma(\cup_0^\infty \pi_m)$ and such that every $B \in \pi_{m+1}$ is an interval and is obtained by splitting some $B' \in \pi_m$ into two pieces. Let $B_\phi = \mathbb{R}$ and let $\pi_m = \{B_{\epsilon_1,\ldots,\epsilon_m} : \epsilon_j = 0$ or 1 for j = 1, ..., m$\}$ and let $B_{\epsilon_1,\ldots,\epsilon_m 0} \in \pi_{m+1}$ and $B_{\epsilon_1,\ldots,\epsilon_m 1} \in \pi_{m+1}$ be the two pieces into which $B_{\epsilon_1,\ldots,\epsilon_m}$ is split.

Definition 1.6.1 *A random probability measure P on ($\mathbb{R}$, $\mathcal{B}(\mathbb{R})$) is said to have a Polya tree distribution or a Polya tree prior with parameters $(\mathbf{\Pi}, \boldsymbol{\alpha})$ and we write P $\in$ PT $(\mathbf{\Pi}, \boldsymbol{\alpha})$, if there exists a collection of non-negative numbers*

$$\boldsymbol{\alpha} = \{\alpha_{\epsilon_1,\ldots,\epsilon_m} : \epsilon_j = 0 \text{ or } 1, \text{ for } j = 1, \ldots, m; m = 1, 2, \ldots\}$$

such that the following hold:

1. $\{P(B_{\epsilon_1,\ldots,\epsilon_m 0} \mid B_{\epsilon_1,\ldots,\epsilon_m}) : \epsilon_j = 0 \text{ or } 1 \text{ for } j = 1, \ldots, m; m = 1, 2, \ldots\}$ *are independent random variables.*

2. $P(B_{\epsilon_1,\ldots,\epsilon_m 0} \mid B_{\epsilon_1,\ldots,\epsilon_m})$ *has the beta distribution* $B(\alpha_{\epsilon_1,\ldots,\epsilon_m 0}, \alpha_{\epsilon_1,\ldots,\epsilon_m 1})$.

Polya tree priors seem to have their origin in Blackwell (1973), and Ferguson (1974, page 620) (even though both Blackwell and Ferguson do not use the phrase 'Polya tree priors'), and recently Lavine (1992, 1994) and Mauldin et al.(1992) investigate some of their interesting properties and set the course for their use in Bayesian analysis.

Existence. The existence of Polya tree processes can be shown by first realizing it as a prior on the space of probabilities on the sequence space $\{0, 1\}^{\mathbf{N}}$ and then transfering it to M($\mathbb{R}$). A more elegant way is to use de Finnetti's theorem. We refer the reader to Mauldin et al.(1992) for a discussion of these issues.

Support. The support of a Polya tree process is controlled by the choice of the parameters $\boldsymbol{\alpha}$ and of course the partitions $\boldsymbol{\Pi}$. Mauldin et al. (1992) give sufficient conditions for the Polya tree prior to give mass one to the space of all continuous probability distributions. If for simplicity we consider the Polya tree prior for (0, 1] with $\pi_m = \{ ((\frac{(i-1)}{2^m}), (\frac{i}{2^m})] : i = 1, \ldots, 2^m\}$ - the set of all dyadic intervals of length $\frac{1}{2^m}$, and take $\alpha_{\epsilon_1,\ldots,\epsilon_m} = m^2$, the resulting Polya tree will be absolutely continuous with probability one. This feature of Polya tree priors make it more attractive as a prior especially in the context of density estimation problems. Lavine (1992, 1994) have a discussion on the implications and interpretations of various choices of the partitions $\boldsymbol{\Pi}$ and the non negative numbers $\boldsymbol{\alpha}$.

From now on to avoid cumbersome notation we will write $B_{\boldsymbol{\epsilon}}$ for $B_{\epsilon_1,\ldots,\epsilon_m}$ and $\alpha_{\boldsymbol{\epsilon}}$ for $\alpha_{\epsilon_1,\ldots,\epsilon_m}$ unless it is very important to write otherwise.

Connection to Dirichlet process [Lavine (1994), Fact 2.]. The Polya tree prior is a generalization of the Dirichlet process in the sense explained below.

a) A Dirichlet Process $\mathcal{D}(\alpha)$ is a Polya tree w.r.t. any sequence of partitions $\boldsymbol{\Pi}$ with $\alpha_{\boldsymbol{\epsilon}} = \alpha(B_{\boldsymbol{\epsilon}})$, for all $B_{\boldsymbol{\epsilon}} \in \boldsymbol{\Pi}$.

b) A Polya tree PT($\boldsymbol{\Pi}$, $\boldsymbol{\alpha}$)is a Dirichlet process if $\alpha_{\boldsymbol{\epsilon}} = \alpha_{\boldsymbol{\epsilon}0} + \alpha_{\boldsymbol{\epsilon}1}$, for all possible values of $\boldsymbol{\epsilon}$. The parameter α of the associated Dirichlet process is specified as $\alpha(B_{\boldsymbol{\epsilon}}) = \alpha_{\boldsymbol{\epsilon}}$.

Posterior Distribution [Mauldin et. al.(1992), Theorem 4.3]. Let P $\in$ PT($\boldsymbol{\Pi}$, $\boldsymbol{\alpha}$) and, given P, let $X_1, \ldots, X_n$ be a sample from P. Then the posterior distribution of P given $X_1, \ldots, X_n$ is PT($\boldsymbol{\Pi}$, $\boldsymbol{\alpha}_{X_1,\ldots,X_n}$) where $\alpha_{\boldsymbol{\epsilon}}$ in $\boldsymbol{\alpha}$ is replaced by $(\alpha_{\boldsymbol{\epsilon}} + \sum_1^n I[X_i \in B_{\boldsymbol{\epsilon}}])$ in $\boldsymbol{\alpha}_{X_1,\ldots,X_n}$. Thus the Polya tree priors form a conjugate family of priors.

Posterior distribution given incomplete/partial observations [Lavine (1994), page 1223]. One feature of interest to us is the fact that a Polya tree process permits easy posterior updating even in the presence of partial information. More precisely, let

P $\in$ PT(Π, α) and given P, let $X_1, \ldots, X_n$ be a sample from P. Then the posterior distribution of P given $\{ X_1 \in B_{\epsilon^1}, \ldots, X_n \in B_{\epsilon^n}\}$ is again a Polya tree with respect to Π, with α_ϵ changing to $\alpha_\epsilon + \sum_1^n I\{B_{\epsilon^i} \subset B_\epsilon\}$.

[Remark: In the case where we have some observations fully specified and some partially specified, the updating for the posterior is first done for the fully specified observations and then for the partially specified observations in an obvious way.]

Bayes Estimates. Let P $\in$ PT(Π, α) then the Bayes estimate (w.r.t. squared error loss) of $P(B_{\epsilon_1,\ldots,\epsilon_m})$ given a sample $X_1, X_2, \ldots, X_n$ from P is,

$$E(P(B_{\epsilon_1,\ldots,\epsilon_m})) = \prod_1^m \frac{\alpha_{\epsilon_1,\ldots,\epsilon_i} + \sum_1^n I[X_j \in B_{\epsilon_1,\ldots,\epsilon_i}]}{\alpha_{\epsilon_1,\ldots,\epsilon_{i-1}0} + \alpha_{\epsilon_1,\ldots,\epsilon_{i-1}1} + \sum_1^n I[X_j \in B_{\epsilon_1,\ldots,\epsilon_{i-1}}]}.$$

As with the Dirichlet process, here also, if the α_ϵ's are small (compared to n), the Bayes estimate is close to the sample distribution function. This expression for the Bayes estimate also describes the predictive distribution of a future observation given $X_1, X_2, \ldots, X_n$. Details of this can be found in Mauldin et al. (1992).

Posterior consistency. Calculations similar to that carried out for the Dirichlet process, also shows that the Polya tree priors lead to posteriors that are weakly consistent at all $P_0 \in M(\mathbb{R})$. But, unlike the Dirichlet process, the stronger k-consistency need not hold for the Polya tree priors.

Remarks: The properties of Dirichlet process and Polya tree processes on $M(\mathbb{R})$ that have been mentioned in the last two sections have obvious extension to $M(\mathbb{R}^+)$ and $M(\mathbb{R}^+ \times \{0,1\})$, the two other spaces discussed in this dissertation.

CHAPTER 2

Polya Tree Priors for Symmetric Distributions

2.1 Introduction and Summary

In many semi-parametric inference problems, within a Bayesian formulation, identifiability conditions requires the Bayesian to consider priors on the class of distributions symmetric around an arbitrary point on the real line. A typical example is the location problem. Diaconis and Freedman (1986a, 1986b) consider the location problem, using symmetrized Dirichlet process priors. The first paper and the subsequent discussions provide a good summary of the basic issues involved in such problems, and also elaborate on the need for families of rich priors on the class of symmetric distributions. More recently, Ghosal et al. (1996), consider the same problem using using symmetrized Polya tree priors on distributions with symmetric densities.

Dalal (1979) constructs a class of priors which are invariant under a finite group of transformations, using the Dirichlet process priors and calls them Dirichlet Invariant processes. In this chapter, we study priors on the class of symmetric distributions, using the Polya tree processes. We consider two natural methods (discussed below) of constructing priors on the class of symmetric distributions and compare the two

methods using Dirichlet processes and Polya tree processes.

A prior $\mathbb{P}$ on the class of all symmetric distributions on $\mathbb{R}$, denoted by $M^S(\mathbb{R})$, can be constructed in two natural ways.

Method 1. For any prior (say λ_1) on $M(\mathbb{R})$, the map $P \mapsto P_1^*$, defined by

$$P_1^*(A) = \frac{P(A) + P(-A)}{2}, \text{ for } A \in \mathcal{B}(\mathbb{R})$$

induces a prior on $\mathrm{M}^S(\mathbb{R})$.

Method 2. For any prior (say λ_2)on $M(\mathbb{R}^+)$, the map $P \mapsto P_2^*$, defined by

$$P_2^*(A) = \frac{P(\mathbb{R}^+ \cap A)}{2} + \frac{P(\mathbb{R}^- \cap A)}{2}, \text{for } A \in \mathcal{B}(\mathbb{R}^+),$$

induces a prior on $M^S(\mathbb{R})$.

Dalal (1979), looks at symmetrization using Method 1, with $\lambda_1 = \mathcal{D}(\alpha)$, where α is a symmetric measure on $\mathbb{R}$. Using the transformation invariance property of Dirichlet processes, it can be verified [Hannum and Hollander (1983), Theorem 2.1] that with $\lambda_2 = \mathcal{D}(2\alpha^+)$, the Method 2 symmetrization is equivalent to the (Method 1) symmetrization considered by Dalal (1979).

In the next section we look at the two methods of symmetrization using analogous Polya tree priors and show that the two methods yield the same prior, iff the Polya tree processes being considered are Dirichlet processes. In the last section we consider (a version of the) posterior distributions under the two methods and establish the weak consistency for the sequence of posterior distributions.

2.2 Symmetrization using Polya tree processes

In this section we study the two methods of symmetrizations using Polya trees that can be considered analogous to $\mathcal{D}(\alpha)$ and $\mathcal{D}(2\alpha^+)$. With this in mind, we now

introduce notation that is crucial to our construction and results. Let,

$$\pi_m^+ = \{B^+_{\epsilon_1,\ldots\epsilon_m} : \epsilon_j = 0 \text{ or } 1 \text{ for j} = 1,\ldots,m\},$$

where $\{B^+_{\epsilon_1,\ldots\epsilon_m} : \epsilon_j = 0$ or 1 for j $= 1,\ldots,m\}$ is a partition of $\mathbb{R}^+$. Let $B^-_{\epsilon_1,\ldots\epsilon_m} = -B^+_{\epsilon_1,\ldots\epsilon_m}$, and let

$$\pi_m^- = \{B^-_{\epsilon_1,\ldots\epsilon_m} : \epsilon_j = 0 \text{ or } 1 \text{ for j} = 1,\ldots,\text{m}\}.$$

Let $\mathbf{\Pi}^+ = \{\pi_m^+ : m = 1, 2, \ldots\}$, and $\mathbf{\Pi}^- = \{\pi_m^- : m = 1, 2, \ldots\}$and let $\mathbf{\Pi} = \mathbf{\Pi}^+ \cup \mathbf{\Pi}^-$.

In Method 1 we take $\lambda_1 = \text{PT}(\mathbf{\Pi}, \boldsymbol{\alpha})$, where $\boldsymbol{\alpha}$ is a symmetric collection, i.e. under PT($\mathbf{\Pi}$, $\boldsymbol{\alpha}$), $P(B^+_{\boldsymbol{\epsilon}0} \mid B^+_{\boldsymbol{\epsilon}})$ and $P(B^-_{\boldsymbol{\epsilon}0} \mid B^-_{\boldsymbol{\epsilon}})$ both have the $Beta(\alpha_{\boldsymbol{\epsilon}0}, \alpha_{\boldsymbol{\epsilon}1})$ distribution.

In Method 2 we take $\lambda_2 = \text{PT}(\mathbf{\Pi}^+, 2\boldsymbol{\alpha}^+)$, where $\boldsymbol{\alpha}^+$ has an obvious interpretation.

In the remainder of this chapter λ_1 will always represent PT($\mathbf{\Pi}$, $\boldsymbol{\alpha}$), and λ_2 will always represent PT($\mathbf{\Pi}^+, 2\boldsymbol{\alpha}^+$).

Theorem 2.2.1 *The priors induced on $M^S(\mathbb{R})$ by Method 1 (using $\lambda_1 = PT(\mathbf{\Pi}, \boldsymbol{\alpha})$) and Method 2 (using $\lambda_2 = PT(\mathbf{\Pi}^+, 2\boldsymbol{\alpha}^+)$) are the same if and only if*

$$\alpha_{\epsilon_1,\ldots,\epsilon_m} = \alpha_{\epsilon_1,\ldots,\epsilon_m 0} + \alpha_{\epsilon_1,\ldots,\epsilon_m 1},$$
$$\textit{for all } \{\epsilon_1,\ldots,\epsilon_m\} \in \{0,1\}^m\textit{; for } m = 1, 2, \ldots$$

and hence for Polya tree processes the two methods yield the same prior if and only if it is a Dirichlet process.

Proof: The proof is by the principle of mathematical induction and uses elementary properties of the Beta distribution. We recall that if $X \sim Beta(a, b)$, then $E(X) = \frac{a}{a+b}$, $E(X)^2 = \frac{a(a+1)}{(a+b)(a+b+1)}$.

If part: If the condition holds , then $\lambda_1 = \mathcal{D}(\alpha)$, with the measure α given by $\alpha(B_\epsilon) = \alpha_\epsilon$ for all $B_\epsilon \in \pi$, and $\lambda_2 = \mathcal{D}(2\alpha^+)$. The result now follows from the remarks made in the last section about symmetrization using Dirichlet process priors.

Only if part: If the priors induced by the two methods on $M^S(\mathbb{R})$ are the same, then,

$$E_{\lambda_1}[P(B^+_{\epsilon_1,\epsilon_2,\ldots,\epsilon_n}) + P(B^-_{\epsilon_1,\epsilon_2,\ldots,\epsilon_n})]^2 = E_{\lambda_2}[P(B^+_{\epsilon_1,\epsilon_2,\ldots,\epsilon_n})]^2$$

for all $\epsilon_1, \epsilon_2, \ldots, \epsilon_n \in \{0,1\}^m$; for $m = 1, 2, \ldots$

To avoid trivialities, we will assume that

$\alpha_{\epsilon_1,\epsilon_2,\ldots,\epsilon_n} > 0$, for all $\epsilon_1, \epsilon_2, \ldots, \epsilon_n \in \{0,1\}^m$. Then,

$$\begin{aligned}
& E_{\lambda_1}[P(B_0^+) + P(B_0^-)]^2 \\
= \; & E(X_1Y_1 + (1-X_1)Y_2)^2, \\
& \text{where } X_1 \sim Beta(\alpha,\alpha), Y_i \sim Beta(\alpha_0,\alpha_1)(i=1,2), \\
& \text{and } \{X-1, Y_1, Y_2\} \text{ are independent.} \\
= \; & 2[\frac{\alpha(\alpha+1)}{2\alpha(2\alpha+1)} \times \frac{\alpha_0(\alpha_0+1)}{(\alpha_0+\alpha_1)(\alpha_0+\alpha_1+1)}] + 2[\frac{\alpha\alpha}{2\alpha(2\alpha+1)} \times (\frac{\alpha_0}{\alpha_0+\alpha_1})^2] \\
= \; & \frac{1}{2\alpha+1}\frac{\alpha_0}{\alpha_0+\alpha_1}[\frac{(\alpha+1)(\alpha_0+1)}{\alpha_0+\alpha_1+1} + \frac{\alpha\alpha_0}{\alpha_0+\alpha_1}]
\end{aligned}$$

Similarly,

$$E_{\lambda_2}[P(B_0^+)]^2 = \frac{\alpha_0(2\alpha_0+1)}{(\alpha_0+\alpha_1)(2\alpha_0+2\alpha_1+1)}$$

Therefore, $E_{\lambda_1}[P(B_0^+) + P(B_0^-)]^2 = E_{\lambda_2}[P(B_0^+)]^2$, iff

$$[\frac{(\alpha+1)(\alpha_0+1)}{(\alpha_0+\alpha_1+1)} + \frac{\alpha\alpha_0}{\alpha_0+\alpha_1}] = \frac{(2\alpha_0+1)(2\alpha+1)}{2\alpha_0+2\alpha_1+1},$$

which in turn holds iff $\alpha = \alpha_0 + \alpha_1$.

Let,

$$\alpha_{\epsilon_1,\epsilon_2,\ldots,\epsilon_j} = \alpha_{\epsilon_1,\epsilon_2,\ldots,\epsilon_j 0} + \alpha_{\epsilon_1,\epsilon_2,\ldots,\epsilon_j 1},$$

$$\text{for } \epsilon_1,\epsilon_2,\ldots,\epsilon_j \in \{0,1\}^j; j = 1,2,\ldots,(n-1).$$

Then we will show that

$\alpha_{\epsilon_1,\epsilon_2,\ldots,\epsilon_n} = \alpha_{\epsilon_1,\epsilon_2,\ldots,\epsilon_n 0} + \alpha_{\epsilon_1,\epsilon_2,\ldots,\epsilon_n 1}$, for $\epsilon_1,\epsilon_2,\ldots,\epsilon_n \in \{0,1\}^n$,
by equating $E_{\lambda_1}[P(B^+_{\epsilon_1,\epsilon_2,\ldots,\epsilon_n 0}) + P(B^-_{\epsilon_1,\epsilon_2,\ldots,\epsilon_n 0})]^2$ and $E_{\lambda_2}[P(B^+_{\epsilon_1,\epsilon_2,\ldots,\epsilon_n 0})]^2$.

$$\begin{aligned}
& E_{\lambda_1}[P(B^+_{\epsilon_1,\epsilon_2,\ldots,\epsilon_n 0}) + P(B^-_{\epsilon_1,\epsilon_2,\ldots,\epsilon_n 0})]^2 \\
= \; & 2[\frac{\alpha(\alpha+1)}{2\alpha(2\alpha+1)} \prod_{j=1}^{n} \frac{\alpha_{\epsilon_1,\epsilon_2,\ldots,\epsilon_j}(\alpha_{\epsilon_1,\epsilon_2,\ldots,\epsilon_j}+1)}{(\alpha_{\epsilon_1,\epsilon_2,\ldots,\epsilon_{j-1}0} + \alpha_{\epsilon_1,\epsilon_2,\ldots,\epsilon_{j-1}1})(\alpha_{\epsilon_1,\epsilon_2,\ldots,\epsilon_{j-1}0} + \alpha_{\epsilon_1,\epsilon_2,\ldots,\epsilon_{j-1}1}+1)} \\
& \times \frac{\alpha_{\epsilon_1,\epsilon_2,\ldots,\epsilon_n 0}(\alpha_{\epsilon_1,\epsilon_2,\ldots,\epsilon_n 0}+1)}{(\alpha_{\epsilon_1,\epsilon_2,\ldots,\epsilon_n 0} + \alpha_{\epsilon_1,\epsilon_2,\ldots,\epsilon_n 1})(\alpha_{\epsilon_1,\epsilon_2,\ldots,\epsilon_n 0} + \alpha_{\epsilon_1,\epsilon_2,\ldots,\epsilon_n 1}+1)}] \\
& + 2[\frac{\alpha(\alpha+1)}{2\alpha(2\alpha+1)} \prod_{j=1}^{n} (\frac{\alpha_{\epsilon_1,\epsilon_2,\ldots,\epsilon_j}}{(\alpha_{\epsilon_1,\epsilon_2,\ldots,\epsilon_{j-1}0} + \alpha_{\epsilon_1,\epsilon_2,\ldots,\epsilon_{j-1}1})})^2 (\frac{\alpha_{\epsilon_1,\epsilon_2,\ldots,\epsilon_n 0}}{(\alpha_{\epsilon_1,\epsilon_2,\ldots,\epsilon_n 0} + \alpha_{\epsilon_1,\epsilon_2,\ldots,\epsilon_n 1})})^2 \\
= \; & 2[\frac{1}{2\alpha(2\alpha+1)} \frac{\alpha_{\epsilon_1,\epsilon_2,\ldots,\epsilon_n}(\alpha_{\epsilon_1,\epsilon_2,\ldots,\epsilon_n}+1)\alpha_{\epsilon_1,\epsilon_2,\ldots,\epsilon_n 0}(\alpha_{\epsilon_1,\epsilon_2,\ldots,\epsilon_n 0}+1)}{(\alpha_{\epsilon_1,\epsilon_2,\ldots,\epsilon_n 0} + \alpha_{\epsilon_1,\epsilon_2,\ldots,\epsilon_n 1})(\alpha_{\epsilon_1,\epsilon_2,\ldots,\epsilon_n 0} + \alpha_{\epsilon_1,\epsilon_2,\ldots,\epsilon_n 1}+1)}] \\
& + 2[\frac{1}{2\alpha(2\alpha+1)} (\frac{\alpha_{\epsilon_1,\epsilon_2,\ldots,\epsilon_n}\alpha_{\epsilon_1,\epsilon_2,\ldots,\epsilon_n 0}}{\alpha_{\epsilon_1,\epsilon_2,\ldots,\epsilon_n 0} + \alpha_{\epsilon_1,\epsilon_2,\ldots,\epsilon_n 1}})^2]
\end{aligned}$$

(The second equality follows from the above hypothesis.)

Similarly,

$$\begin{aligned}
E_{\lambda_2}[P(B^+_{\epsilon_1,\epsilon_2,\ldots,\epsilon_n 0})]^2 = \; & \frac{(2\alpha_{\epsilon_1,\epsilon_2,\ldots,\epsilon_n})(2\alpha_{\epsilon_1,\epsilon_2,\ldots,\epsilon_n}+1)}{(2\alpha_0+2\alpha_1)(2\alpha_0+2\alpha_1+1)} \\
& \times \frac{2\alpha_{\epsilon_1,\epsilon_2,\ldots,\epsilon_n 0}(2\alpha_{\epsilon_1,\epsilon_2,\ldots,\epsilon_n 0}+1)}{(2\alpha_{\epsilon_1,\epsilon_2,\ldots,\epsilon_n 0} + 2\alpha_{\epsilon_1,\epsilon_2,\ldots,\epsilon_n 1})(2\alpha_{\epsilon_1,\epsilon_2,\ldots,\epsilon_n 0} + 2\alpha_{\epsilon_1,\epsilon_2,\ldots,\epsilon_n 1}+1)}
\end{aligned}$$

Using the above hypothesis again, we can conclude that,

$$E_{\lambda_1}[P(B^+_{\epsilon_1,\epsilon_2,\ldots,\epsilon_n 0}) + P(B^-_{\epsilon_1,\epsilon_2,\ldots,\epsilon_n 0}]^2 = E_{\lambda_2}[P(B^+_{\epsilon_1,\epsilon_2,\ldots,\epsilon_n 0})]^2, \text{ iff}$$
$$\frac{(\alpha_{\epsilon_1,\epsilon_2,\ldots,\epsilon_n}+1)(\alpha_{\epsilon_1,\epsilon_2,\ldots,\epsilon_n 0}+1)}{\alpha_{\epsilon_1,\epsilon_2,\ldots,\epsilon_n 0}+\alpha_{\epsilon_1,\epsilon_2,\ldots,\epsilon_n 1}+1} + \frac{\alpha_{\epsilon_1,\epsilon_2,\ldots,\epsilon_n}\alpha_{\epsilon_1,\epsilon_2,\ldots,\epsilon_n 0}}{\alpha_{\epsilon_1,\epsilon_2,\ldots,\epsilon_n 0}+\alpha_{\epsilon_1,\epsilon_2,\ldots,\epsilon_n 1}}$$
$$= \frac{(2\alpha_{\epsilon_1,\epsilon_2,\ldots,\epsilon_n}+1)(2\alpha_{\epsilon_1,\epsilon_2,\ldots,\epsilon_n 0}+1)}{2\alpha_{\epsilon_1,\epsilon_2,\ldots,\epsilon_n 0}+2\alpha_{\epsilon_1,\epsilon_2,\ldots,\epsilon_n 1}+1}$$

which in turn holds iff

$$\alpha_{\epsilon_1,\epsilon_2,\ldots,\epsilon_n} = \alpha_{\epsilon_1,\epsilon_2,\ldots,\epsilon_n 0} + \alpha_{\epsilon_1,\epsilon_2,\ldots,\epsilon_n 1}$$

$\Diamond$

2.3 The posterior distribution and its consistency

We observe that there is a 1-1 correspondence between $M^S(\mathbb{R})$ and $M(\mathbb{R}^+)$, and will make use of this correspondence in the remainder of this section. With this in mind we briefly review the properties of this correspondence, that are relevant to our discussion.

Let,

$$\phi : M^S(\mathbb{R}) \mapsto M(\mathbb{R}^+),$$

be defined as $\phi(P)(A) = 2P(A)$, for $A \in M(\mathbb{R}^+)$. ϕ is 1-1 and onto. We will on occasion write P^+ for $\phi(P)$ in the remainder of this section.

Let μ be a (prior) probability measure on $M^S(\mathbb{R})$, then the map ϕ induces the prior probability measure $\mu\phi^{-1}$ on $M(\mathbb{R}^+)$. The following propositions summarize the important consequences of using the map ϕ and consider the following set-up:

Let $P \sim \mu$, and given P, let $\{X_1, X_2, \ldots, X_n\}$ be i.i.d P.

Proposition 2.3.1 *The posterior distribution of P given $\{X_1, X_2, \ldots, X_n\}$, is the same as the conditional distribution of μ given $\{\mid X_1 \mid, \mid X_2 \mid, \ldots, \mid X_n \mid\}$, i. e.*

$$\mu(\cdot \mid X_1, X_2, \ldots, X_n) = \mu(\cdot \mid \mid X_1 \mid, \mid X_2 \mid, \ldots, \mid X_n \mid).$$

[This follows from the fact that $\{\mid X_1 \mid, \mid X_2 \mid, \ldots, \mid X_n \mid\}$ is a sufficient statistic for symmetric distributions.]

Proposition 2.3.2 $\mu(\cdot \mid \mid X_1 \mid, \mid X_2 \mid, \ldots, \mid X_n \mid) =$
$\mu\phi^{-1}(\phi(\cdot) \mid \mid X_1 \mid, \mid X_2 \mid, \ldots, \mid X_n \mid)$

Proof: For notational convenience we will consider the case $n = 1$. We need to show that for any $B \in \mathcal{B}(\mathbb{R}^+)$, the measures μ_1, and μ_2 on $M^S(\mathbb{R})$, defined as

$$\begin{aligned} \mu_1(C) &= \int_B \mu\phi^{-1}(\phi(C) \mid\mid X \mid\mid)\bar{\mu}(dx) \\ \mu_2(C) &= \int_C 2P(B)\mu(dP) \end{aligned}$$

are the same where $C \subset M^S(\mathbb{R})$, and $\bar{\mu}(A) = E_\mu(P(\mid X \mid \in A))$.

$$\begin{aligned} \mu_2(C) &= \int_C 2P(B)\mu(dP) \\ &= \int_C (\phi P)(B)\mu(dP) \\ &= \int_{\phi(C)} P(B)\mu\phi^{-1}(dP) \\ &= \int_B \mu\phi^{-1}(\phi(C) \mid\mid X \mid\mid)\bar{\mu}(dx) \\ &= \mu_1(C) \end{aligned}$$

(The third equality above follows from the change of variable theorem, and the fourth equality follows from the definition of conditional distributions.) $\diamondsuit$

Proposition 2.3.3 *Let $P_0 \in M^S(\mathbb{R})$, and let $\{\mu_n\}_{n\geq 1}$ be probability measures on $M^S(\mathbb{R})$, then $\mu_n \Rightarrow \delta_{P_0}$ iff $\mu_n\phi^{-1} \Rightarrow \delta_{P_0^+}$.*

Proof: For any f bounded continuous function on $\mathbb{R}$, and any $P \in M^S(\mathbb{R})$, we observe that

$$\begin{aligned}
\int_{\mathbb{R}} f(x)dP(x) &= \int_{\mathbb{R}} \frac{(f(x)+f(-x))}{2}dP(x) \\
&= 2\int_{\mathbb{R}^+} \frac{(f(x)+f(-x))}{2}dP(x) \\
&= \int_{\mathbb{R}^+} f^+(x)d(\phi P)(x), \text{ where } f^+(x) = \frac{(f(x)+f(-x))}{2} \\
&= \int_{\mathbb{R}^+} f^+(x)dP^+(x).
\end{aligned}$$

Similarly we can show that for any f^+ bounded continuous function on $\mathbb{R}^+$, there exist an f bounded continuous on $\mathbb{R}$ such that $\int_{\mathbb{R}^+} f^+(x)dQ(x) = \int_{\mathbb{R}} f(x)d(\phi^{-1}Q)(x)$.

Therefore with $\{\{P_n\}_{n\geq 1}, P\} \subset M^S(\mathbb{R})$, $P_n \Rightarrow P$ iff $\phi(P_n) \Rightarrow \phi(P)$, and hence $\mu_n \Rightarrow \delta_{P_0}$ iff $\mu_n\phi^{-1} \Rightarrow \delta_{P_0^+}$.

$\diamondsuit$

In view of Propositions 2.3.1, 2.3.2, and 2.3.3, (and ease of notation), we will study the priors (and posteriors) induced by Method 1 and Method 2 on $M(\mathbb{R}^+)$, rather than $M^S(\mathbb{R})$. We will see that even though the two methods are not always equivalent, the posteriors still are weakly consistent for both the methods. We recall that in Method 1, the prior on $M(\mathbb{R}^+)$ is $\lambda_1 \circ g^{-1}$ where $(g \circ P)(A) = P(A) + P(-A)$, and in Method 2 the prior on $M(\mathbb{R}^+)$ is λ_2, with $\lambda_1 = PT(\boldsymbol{\Pi}, \boldsymbol{\alpha})$ and $\lambda_2 = PT(\boldsymbol{\Pi}^+, 2\boldsymbol{\alpha}^+)$.

The next proposition follows from the properties of Polya tree priors mentioned in Chapter 1.

Proposition 2.3.4 *Let $P \sim \lambda_2$, and given P, let $\{T_1, T_2, \ldots, T_n\}$ be a random sample from P. The posterior distribution of P given $\{T_1, T_2, \ldots, T_n\}$ is $PT(\boldsymbol{\pi}^+, 2\boldsymbol{\alpha}^+_{T_1,T_2,\ldots,T_n})$, and further the posterior is consistent at all $Q_0 \in M(\mathbb{R}^+)$.*

Thus, in view of the comments made earlier, the Method 2 symmetrization using Polya tree priors yields (weakly) consistent posteriors. The more interesting result is that the Method 1 symmetrization also yields a (weakly) consistent sequence of posteriors. To establish this we first need to consider (a version of) the posterior distribution under Method 1.

Theorem 2.3.1 *Let $P \sim \lambda_1 \circ g^{-1}$, and given P, let $T_1, T_2, \ldots, T_n$ be a random sample from P. The posterior distribution of P given $\{T_1, T_2, \ldots, T_n\}$, is given as*

$$\begin{aligned} & \lambda_1 \circ g^{-1}(\cdot \mid T_1, T_2, \ldots, T_n) \\ = & \sum_{\{x_1,x_2,\ldots,x_n : |x_i| = T_i\}} PT(\boldsymbol{\pi}, \boldsymbol{\alpha}_{x_1,x_2,\ldots,x_n}) \circ g^{-1}(\cdot) f_{T_1,T_2,\ldots,T_n}(x_1, x_2, \ldots, x_n), \end{aligned}$$

where, $f_{T_1,T_2,\ldots,T_n}(\mp T_1, \ldots, \mp T_n) = lim_k \frac{Pr(X_1 \in B^{\mp}_{\epsilon_1^k}, \ldots, X_n \in B^{\mp}_{\epsilon_n^k})}{Pr(T_1 \in B^{+}_{\epsilon_1^k}, \ldots, T_n \in B^{+}_{\epsilon_n^k})}$, *and* $B^+_{\epsilon_i^k} \downarrow T_i$ *for* $i = 1, \ldots, n$

[Remark: In an intuitive sense, $f_{T_1,T_2,\ldots,T_n}(x_1, x_2, \ldots, x_n) =$
$Pr(X_1 = x_1, \ldots, X_n = x_n | \mid X_1 \mid = T_1, \ldots, \mid X_n \mid = T_n).$]

Proof: For any measurable $C \subset M(\mathbb{R}^+)$, and $B \in B(\mathbb{R}^{+n})$, let

$$\begin{aligned} \mu(C) &:= \int_B \lambda_1 \circ g^{-1}(C \mid T_1, T_2, \ldots, T_n) \bar{\alpha}^n \circ g^{-1}(d\boldsymbol{T}), \\ \nu(C) &:= \int_C P^n(B) PT(\boldsymbol{\pi}, \boldsymbol{\alpha}) \circ g^{-1}(dP), \end{aligned}$$

where $\bar{\alpha}^n \circ g^{-1}(B) = E_{PT(\boldsymbol{\pi},\boldsymbol{\alpha}) \circ g^{-1}} P^n(B)$, and $P^n(B) = P(\boldsymbol{T} \in B)$, with $\boldsymbol{T} = \{T_1, T_2, \ldots, T_n\}$, and $T_i \sim^{i.i.d.} P$.

To verify that $\lambda_1 \circ g^{-1}(\cdot \mid T_1, T_2, \ldots, T_n)$ is indeed a version of the posterior , we need to verify that for every $B \in B(\mathbb{R}^{+n})$, the measures μ, and ν on $M(\mathbb{R}^+)$, defined above are the same. In fact, it is enough to verify the same for B of the form $B = B_1 \times \cdots \times B_n$, where $B_i \in M(\mathbb{R}^+)$.

We observe that if $B = B_1 \times \cdots \times B_n$, with $B_i \in M(\mathbb{R}^+)$, then

$$\begin{aligned} \bar{\alpha}^n \circ g^{-1}(B) &= E_{PT(\boldsymbol{\pi},\boldsymbol{\alpha})\circ g^{-1}}(P(B_1) \times \cdots \times P(B_n)) \\ &= E_{PT(\boldsymbol{\pi},\boldsymbol{\alpha})}[((P(B_1^+) + P(B_1^-)) \times \cdots \times (P(B_n^+) + P(B_n^-))]. \end{aligned}$$

Let $\bar{\alpha}_n((B_1^+ \cup B_1^-) \times \cdots \times (B_n^+ \cup B_n^-)) :=$
$E_{PT(\boldsymbol{\pi},\boldsymbol{\alpha})}[((P(B_1^+) + P(B_1^-)) \times \cdots \times (P(B_n^+) + P(B_n^-))]$.

It suffices to check that the moments of $\{P(B_{\boldsymbol{\epsilon}^i}^+) : \boldsymbol{\epsilon}^i \in \{0,1\}^m\}$, under the two measures $\mu(\cdot)$, and $\nu(\cdot)$ are the same when $B = B_1 \times \cdots \times B_n$, i.e., we want to verify that for positive integers r_i,

$$E_\mu[\prod_{i=1}^{2^m} P(B_{\boldsymbol{\epsilon}^i}^+)^{r_i}] = E_\nu[\prod_{i=1}^{2^m} P(B_{\boldsymbol{\epsilon}^i}^+)^{r_i}].$$

$$\begin{aligned} & E_\nu[\prod_{i=1}^{2^m} P(B_{\boldsymbol{\epsilon}^i}^+)^{r_i}] \\ = & \int (\prod_{j=1}^{n} P(B_j^+)) \prod_{i=1}^{2^m} P(B_{\boldsymbol{\epsilon}^i}^+)^{r_i} PT(\boldsymbol{\pi},\boldsymbol{\alpha}) \circ g^{-1}(dP) \\ = & \int \prod_{j=1}^{n} (P(B_j^+) + P(B_j^-)) \prod_{i=1}^{2^m} (P(B_{\boldsymbol{\epsilon}^i}^+) + P(B_{\boldsymbol{\epsilon}^i}^-))^{r_i} PT(\boldsymbol{\pi},\boldsymbol{\alpha})(dP) \end{aligned}$$

Also,

$$
\begin{aligned}
& E_\mu[\prod_{i=1}^{2^m} P(B^+_{\epsilon^i})^{r_i}] \\
= & \int_{B_1^+ \times \cdots \times B_n^+} [\int (\prod_{i=1}^{2^m} (P(B^+_{\epsilon^i}))^{r_i}) \sum PT(\boldsymbol{\pi}, \boldsymbol{\alpha}_{x_1,x_2,\ldots,x_n}) \circ g^{-1}(dP) \\
& f_{T_1,T_2,\ldots,T_n}(x_1, x_2, \ldots, x_n)] \bar{\alpha}^n \circ g^{-1}(d\boldsymbol{T}) \\
= & \sum lim_k \frac{Pr(X_1 \in B^{\mp}_{\epsilon_1^k}, \ldots, X_n \in B^{\mp}_{\epsilon_n^k})}{Pr(T_1 \in B^{+}_{\epsilon_1^k}, \ldots, T_n \in B^{+}_{\epsilon_n^k})} \int_{(B_1^+ \cup B_1^-) \times \cdots \times (B_n^+ \cup B_n^-)} [\int \prod_{i=1}^{2^m} (P(B^+_{\epsilon^i}) + P(B^-_{\epsilon^i}))^{r_i} \\
& PT(\boldsymbol{\pi}, \boldsymbol{\alpha}_{x_1,x_2,\ldots,x_n})(dP)] \bar{\alpha}_n(d\boldsymbol{T}) \\
= & lim_k \sum \frac{Pr(X_1 \in B^{\mp}_{\epsilon_1^k}, \ldots, X_n \in B^{\mp}_{\epsilon_n^k})}{Pr(T_1 \in B^{+}_{\epsilon_1^k}, \ldots, T_n \in B^{+}_{\epsilon_n^k})} \int_{(B_1^+ \cup B_1^-) \times \cdots \times (B_n^+ \cup B_n^-)} [\int \prod_{i=1}^{2^m} (P(B^+_{\epsilon^i}) + P(B^-_{\epsilon^i}))^{r_i} \\
& PT(\boldsymbol{\pi}, \boldsymbol{\alpha}_{x_1,x_2,\ldots,x_n})(dP)] \bar{\alpha}_n(d\boldsymbol{x}) \\
= & lim_k \sum \frac{Pr(X_1 \in B^{\mp}_{\epsilon_1^k}, \ldots, X_n \in B^{\mp}_{\epsilon_n^k})}{Pr(T_1 \in B^{+}_{\epsilon_1^k}, \ldots, T_n \in B^{+}_{\epsilon_n^k})} \int \prod_{j=1}^{n} (P(B_i^+) + P(B_i^-)) \prod_{i=1}^{2^m} (P(B^+_{\epsilon^i}) + P(B^-_{\epsilon^i}))^{r_i} \\
& PT(\boldsymbol{\pi}, \boldsymbol{\alpha})(dP) \\
= & lim_k \sum \frac{Pr(X_1 \in B^{\mp}_{\epsilon_1^k}, \ldots, X_n \in B^{\mp}_{\epsilon_n^k})}{Pr(T_1 \in B^{+}_{\epsilon_1^k}, \ldots, T_n \in B^{+}_{\epsilon_n^k})} E_\nu[\prod_{i=1}^{2^m} P(B^+_{\epsilon^i})^{r_i}] \\
= & E_\nu[\prod_{i=1}^{2^m} P(B^+_{\epsilon^i})^{r_i}]
\end{aligned}
$$

[The second and the third equality above follows from the dominated convergence theorem; the fourth equality uses the fact that $PT(\boldsymbol{\pi}, \boldsymbol{\alpha}_{x_1,x_2,\ldots,x_n})$ is the posterior of $PT(\boldsymbol{\pi}, \boldsymbol{\alpha})$ given $\{x_1, x_2, \ldots, x_n\}$.] $\Diamond$

Theorem 2.3.2 *Let* $\bar{\alpha}_{x_1,x_2,\ldots,x_n}(\cdot) = E_{PT(\boldsymbol{\pi},\boldsymbol{\alpha}_{x_1,x_2,\ldots,x_n}) \circ g^{-1}}(P(\cdot))$. *Then* $\{\bar{\alpha}_{x_1,x_2,\ldots,x_n}(\cdot) :| x_i |= T_i, i = 1, 2, \ldots, n\}_{n \geq 1}$ *is a tight family of probability measures on* $(\mathbb{R}^+, \mathcal{B}(\mathbb{R}^+))$, *whenever there exists a* $Q_0 \in M(\mathbb{R}^+)$, *such that* $| \frac{1}{n} \sum_1^n I(T_i \in B^+_{\epsilon}) - Q_0(B^+_{\epsilon}) | \to 0$, *for all* $\epsilon \in \cup_m \{0,1\}^m$.

Proof: For any $(\epsilon_1, \epsilon_2, \ldots, \epsilon_m) \in \{0,1\}^m$, and any $\{x_n :| x_n |= T_n\}_{n \geq 1}$ we will

show that $\mid \bar{\alpha}_{x_1,x_2,\ldots,x_n}(B^+_{\epsilon_1,\epsilon_2,\ldots,\epsilon_m}) - Q_0(B_{\epsilon_1,\epsilon_2,\ldots,\epsilon_m}) \mid \to 0$. This in turn will imply that $\bar{\alpha}_{x_1,x_2,\ldots,x_n}(\cdot) \Rightarrow Q_0$, and tightness follows.

We observe that for any $\{x_n :\mid x_n \mid = T_n\}_{n\geq 1}$,

$$\frac{1}{n}\sum_1^n I(T_i \in B^+_{\boldsymbol{\epsilon}}) = \frac{1}{n}\sum_1^n I(x_i \in B^+_{\boldsymbol{\epsilon}}) + \frac{1}{n}\sum_1^n I(x_i \in B^-_{\boldsymbol{\epsilon}}),$$

and hence

$$\begin{aligned} & \mid \bar{\alpha}_{x_1,x_2,\ldots,x_n}(B^+_{\epsilon_1,\epsilon_2,\ldots,\epsilon_m}) - Q_0(B_{\epsilon_1,\epsilon_2,\ldots,\epsilon_m}) \mid \\ \leq\ & \mid \bar{\alpha}_{x_1,x_2,\ldots,x_n}(B^+_{\epsilon_1,\epsilon_2,\ldots,\epsilon_m}) - \frac{1}{n}\sum_1^n I(x_i \in B^+_{\boldsymbol{\epsilon}}) + \frac{1}{n}\sum_1^n I(x_i \in B^-_{\boldsymbol{\epsilon}}) \mid \\ & + \mid \frac{1}{n}\sum_1^n I(T_i \in B^+_{\boldsymbol{\epsilon}}) - Q_0(B^+_{\boldsymbol{\epsilon}}) \mid, \end{aligned}$$

and hence it is enough to verify that

$$\mid \bar{\alpha}_{x_1,x_2,\ldots,x_n}(B^+_{\epsilon_1,\epsilon_2,\ldots,\epsilon_m}) - \frac{1}{n}\sum_1^n I(x_i \in B^+_{\boldsymbol{\epsilon}}) + \frac{1}{n}\sum_1^n I(x_i \in B^-_{\boldsymbol{\epsilon}}) \mid \to 0.$$

The last mentioned convergence follows by observing that,

$$\begin{aligned} \bar{\alpha}_{x_1,x_2,\ldots,x_n}(B^+_{\epsilon_1,\epsilon_2,\ldots,\epsilon_m}) \ =\ & \prod_{j=1}^m \frac{\alpha_{\epsilon_1,\epsilon_2,\ldots,\epsilon_j} + \sum I(x_i \in B^+_{\epsilon_1,\epsilon_2,\ldots,\epsilon_j})}{\alpha_{\epsilon_1,\epsilon_2,\ldots,\epsilon_{j-1}0} + \alpha_{\epsilon_1,\epsilon_2,\ldots,\epsilon_{j-1}1} + \sum I(x_i \in B^+_{\epsilon_1,\epsilon_2,\ldots,\epsilon_{j-1}})} \\ & + \prod_{j=1}^m \frac{\alpha_{\epsilon_1,\epsilon_2,\ldots,\epsilon_j} + \sum I(x_i \in B^-_{\epsilon_1,\epsilon_2,\ldots,\epsilon_j})}{\alpha_{\epsilon_1,\epsilon_2,\ldots,\epsilon_{j-1}0} + \alpha_{\epsilon_1,\epsilon_2,\ldots,\epsilon_{j-1}1} + \sum I(x_i \in B^-_{\epsilon_1,\epsilon_2,\ldots,\epsilon_{j-1}})} \end{aligned}$$

$\Diamond$

Corollary 2.3.3 $\{\lambda_1 \circ g^{-1}(\cdot \mid T_1, T_2, \ldots, T_n)\}_{n\geq 1}$ *is a tight family of probability measures on* $M(\mathbb{R}^+)$, *whenever there exists a* $Q_0 \in M(\mathbb{R}^+)$, *such that* $\mid \frac{1}{n}\sum_1^n I(T_i \in B^+_{\boldsymbol{\epsilon}}) - Q_0(B^+_{\boldsymbol{\epsilon}}) \mid \to 0$, *for all* $\boldsymbol{\epsilon} \in \cup_m \{0,1\}^m$.

Proof: The tightness of the family of probability measures

$\{\bar{\alpha}_{x_1,x_2,\ldots,x_n}(\cdot) :| x_i |= T_i, i = 1,2,\ldots,n\}_{n\geq 1}$ implies (according to Proposition 1.4.2), that $\{PT(\pi, \alpha_{x_1,x_2,\ldots,x_n})\}_{n\geq 1}$ is a tight family of probability measures on $M(\mathbb{R}^+)$. Therefore for every $\delta > 0$ there exists a compact $M^*_\delta \subset M(\mathbb{R}^+)$, such that $PT(\pi, \alpha_{x_1,x_2,\ldots,x_n})(M^*_\delta) > 1-\delta$, for all $\{x_n\}_{n\geq 1}$ such that $| x_n |= T_n$. This in turn implies that $\lambda_1 \circ g^{-1}(M^*_\delta \mid T_1, T_2, \ldots, T_n) > 1-\delta$, for all n. $\Diamond$

We are now in a position to establish the posterior (weak) consistency for the Method 1 symmetrization using Polya tree priors.

Theorem 2.3.4 *Let $Q_0 \in M(\mathbb{R}^+)$, be such that*
$| \frac{1}{n}\sum_1^n I(T_i \in B^+_\epsilon) - Q_0(B^+_\epsilon) | \to 0$, *for all* $\epsilon \in \cup_m\{0,1\}^m$,
then $\lambda_1 \circ g^{-1}(\cdot \mid T_1, T_2, \ldots, T_n) \Rightarrow \delta_{Q_0}$.

Proof: Let f be a bounded continuous function on $\mathbb{R}^+$, with compact support. It is enough to show that $\lambda_1 \circ g^{-1}(V^\delta_{Q_0} \mid T_1, T_2, \ldots, T_n) \to 1$, where

$$V^\delta_{Q_0} = \{Q :| \int f dQ - \int f dQ_0 |< \delta\}.$$

f, bounded continuous with compact support implies that there exists a simple function of the form

$$f_\delta(\cdot) = \sum_{i=1}^k a_i I_{B_{\epsilon^i}}(\cdot),$$

with $B_{\epsilon^i} \in \cup_m\{0,1\}^m$, such that $sup_{x\in\mathbb{R}^+} \mid f(x) - f_\delta(x) |< \frac{\delta}{3}$.

Observing that

$$\begin{aligned} | \int f dQ - \int f dQ_0 | \leq 2 \ &< \ \frac{\delta}{3} + | \int f_\delta dQ - \int f_\delta dQ_0 |, \text{ and,} \\ \int f_\delta dQ - \int f_\delta dQ_0 \ &= \ \sum_{i=1}^k a_i (Q(B_{\epsilon^i}) - Q_0(B_{\epsilon^i}), \end{aligned}$$

we can conclude that there exists a γ sufficiently small such that

$U^{\gamma}_{Q_0} \subset V^{\delta}_{Q_0}$, where

$$U^{\gamma}_{Q_0} = \cup_1^k \{Q :\mid Q(B_{\epsilon^i}) - Q_0(B_{\epsilon^i}) \mid < \gamma\}.$$

To complete our proof, it is enough to to prove that $\lambda_1 \circ g^{-1}(U^{\gamma}_{Q_0} \mid T_1, T_2, \ldots, T_n) \rightarrow 1$. This follows by observing that

$$\begin{aligned} & \lambda_1 \circ g^{-1}(\{Q :\mid Q(B_{\epsilon^i}) - Q_0(B_{\epsilon^i}) \mid > \gamma\} \mid T_1, T_2, \ldots, T_n) \\ \leq \; & E_{\lambda_1 \circ g^{-1}(\cdot \mid T_1, T_2, \ldots, T_n)} \frac{[\mid Q(B_{\epsilon^i}) - Q_0(B_{\epsilon^i}) \mid]^2}{\gamma^2} \\ \rightarrow \; & 0. \end{aligned}$$

[The last mentioned convergence follows by an argument similar to the one used in the proof of Theorem 2.3.2]

$\Diamond$

Corollary 2.3.5 *The sequence of posteriors* $\lambda_1 \circ g^{-1}(\cdot \mid T_1, T_2, \ldots, T_n)$ *is (weakly) consistent at all* $Q_0 \in M(re^+)$.

Proof: We need to show that $\lambda_1 \circ g^{-1}(\cdot \mid T_1, T_2, \ldots, T_n) \Rightarrow \delta_{Q_0}$ a.s. Q_0, for all $Q_0 \in M(\mathbb{R}^+)$. By the SLLN, $\mid \frac{1}{n} \sum_1^n I(T_i \in B^+_{\epsilon}) - Q_0(B^+_{\epsilon}) \mid \rightarrow 0$, for all $\epsilon \in \cup_m \{0,1\}^m$ a.s. Q_0. By Theorem 2.3.4, the result follows. $\Diamond$

CHAPTER 3

Nonparametric Bayesian inference with right censored observations

3.1 Introduction and summary

In several clinical, epidemiological, biomedical, and reliability studies the outcomes of interest are response times, for example, time to death, time to relapse, time to failure etc. These endpoints may however, not be observed in all subjects. Subjects may be lost to follow-up due to withdrawal, or due to the occurrence of an end point unrelated to the outcome of interest in the study. These data are termed right censored. The typical situation where right censorship occurs is when subjects enter a study at different (random) time points and are followed until a specific endpoint is observed or until the termination of the study. The time of occurrence of the desired end point will be right censored if by the time of termination of the study the event of interest has not taken place.

Formally the right censoring problem is defined as: Let $X_1, X_2, \ldots, X_n$ and $Y_1, Y_2, \ldots, Y_n$ be non-negative i.i.d. random variables with distributions F and G respectively. We view the Xs as lifetimes and the Ys as censoring times and also assume that the Xs and Ys are independent. The observed data are:

$\{ (Z_1, \delta_1), \ldots, (Z_n, \delta_n) \}$, where $Z_i = X_i \wedge Y_i$, and $\delta_i = I[X_i \leq Y_i]$, and the goal is to make inferences on F. This problem was first considered in a non-Bayesian framework by Kaplan and Meier (1958), who introduced the product limit estimator and interpreted it as a "maximum likelihood estimator" of F. In this chapter our main focus is on issues related to posterior consistency in the Bayesian inference with right censored data.

In the censored data context, priors can be constructed in many ways. One way is to consider a prior for F directly as done by Susarla and Van Ryzin (1976) who used a Dirichlet process prior for F. This approach was further studied by Blum and Susarla (1978) who realized the posterior as a mixture of Dirichlet processes. In Section 2, we pursue the approach of the former paper and show that the posterior can be represented as a Polya tree process. This representation clarifies and simplifies many of the calculations of both papers and also enables us to establish the posterior consistency.

Priors can also be constructed via the distribution of the observables (Z, δ), and identifiability conditions can then be used to transfer to a prior for (F, G). This is the method adapted by Tsai (1986) who considers a Dirichlet process prior for the distribution of (Z, δ). In Section 3, we briefly discuss issues related to posterior consistency with priors of this kind.

Yet another approach is to construct priors for F via a prior for the cumulative hazard function. Examples of such construction are Ferguson and Phadia (1979), and Hjort(1990). This approach is not discussed in this dissertation.

3.2 Dirichlet process priors for F

The set-up that we consider can be described as follows: X and Y are non-negative random variables corresponding to life time and censoring time, with distribution

F $\in M(\mathbb{R}^+)$ and G $\in M(\mathbb{R}^+)$. Also {F, $X_1, X_2, \ldots, X_n$}, and {G, $Y_1, Y_2, \ldots, Y_n$} are independent. The observed data is $\{(Z_1, \delta_1), \ldots, (Z_n, \delta_n)\}$. With priors of the form $\mu = \mu_1 \times \mu_2$ for (F, G), our interest is in the marginal posterior distribution $\mu_X(\cdot \mid (Z_1, \delta_1), \ldots, (Z_n, \delta_n))$ (of F) on $M(\mathbb{R}^+)$ given $\{(Z_1, \delta_1), \ldots, (Z_n, \delta_n)\}$. It is clear that if $Z_i = z$ and $\delta_i = 0$, then with regard to X_i, the only information we have is $X_i > z$ and thus $\mu_X(\cdot \mid (Z_i, \delta_i))$ should be given by $\mu_1(\cdot \mid X_i > z)$. In other words, in view of the independence of F and G under the prior, the marginal posterior distribution of F does not depend on μ_2 and hence the prior on G plays no role in the analysis that we are interested in. The next proposition is a formal statement of this fact.

Proposition 3.2.1 *Let $\mu = \mu_1 \times \mu_2$ be a prior on $M(\mathbb{R}^+) \times M(\mathbb{R}^+)$ and let $(Z_1, \delta_1), \ldots, (Z_n, \delta_n)$ be a sequence of observations realized as described above. Assume WLOG that $\delta_1 = \cdots = \delta_m = 1$, and $\delta_{m+1} = \cdots = \delta_n = 0$. Then (a version of) the posterior distribution $\mu_X(\cdot \mid (Z_1, \delta_1), \ldots, (Z_n, \delta_n))$ is given by $\mu_1(\cdot \mid X_1 = Z_1, \ldots, X_m = Z_m, X_{m+1} > Z_{m+1}, \ldots, X_n > Z_n)$.*

As mentioned earlier, μ_2 does not play any role in the marginal posterior, and hence in the results that follow. For the remainder of this section we focus our attention on priors of the form $\mathcal{D}(\alpha) \times \delta_{G_0}$ on $M(\mathbb{R}^+) \times M(\mathbb{R}^+)$. This model was first investigated by Susarla and Van Ryzin (1976), who obtained a Bayes estimate for F and showed that this Bayes estimate converges to the Kaplan Meier estimate as $\alpha(\mathbb{R}^+) \to 0$. Blum and Susarla (1978) complemented this result by showing that the posterior distribution is a mixture of Dirichlet processes. The mixture representation is somewhat cumbersome and we feel that the Polya tree approach is more natural for the censored data framework and makes the computations transparent. We next introduce notation that will be used in the remainder of this dissertation.

Let $\boldsymbol{Z} = (Z_1, Z_2, \ldots, Z_n)$, where $Z_1 < \cdots < Z_n$. Consider the sequence of nested partitions $\{\pi_m(\boldsymbol{Z})\}_{m\geq 1}$ given by:

$$\begin{aligned}
\pi_1(\boldsymbol{Z}) &: B_0 = (0, Z_1], B_1 = (Z_1, \infty) \\
\pi_2(\boldsymbol{Z}) &: B_{00}, B_{01}, B_{10} = (Z_1, Z_2], B_{11} = (Z_2, \infty), \text{ and for } l \leq (n-1), \text{ let} \\
\pi_{l+1}(\boldsymbol{Z}) &: B_{0_l 0}, B_{0_l 1}, \ldots, B_{1_l,0} = (Z_l, Z_{l+1}], B_{1_l 1} = (Z_{l+1}, \infty),
\end{aligned}$$

where 1_l is a string of l 1s, and 0_l is a string of l 0s. The remaining B_{ϵ}s are arbitrarily partitioned into two intervals such that $\{\pi_m(\boldsymbol{Z})\}_{m\geq 1}$ forms a sequence of nested partitions which generates $\mathcal{B}(\mathbb{R}^+)$.

Let $\alpha_{\epsilon_1,\ldots,\epsilon_l} = \alpha(B_{\epsilon_1,\ldots,\epsilon_l})$, and $C^n_{\epsilon_1,\ldots,\epsilon_l} = \sum_{\delta_i=0} I[(Z_i, \infty) \subset B_{\epsilon_1,\ldots,\epsilon_l}]$

For any $\{(Z_1, \delta_1), \ldots, (Z_n, \delta_n)\}$, let $\boldsymbol{Z^*}$ denote the vector of distinct values of the observations for which the corresponding $\delta = 0$ (i.e. the corresponding observation is a censored observation) arranged in an increasing order.

Theorem 3.2.1 *Let $\mu = \mathcal{D}(\alpha)\times\delta_{G_0}$ be the prior on $M(\mathbb{R}^+) \times M(\mathbb{R}^+)$. Then the posterior distribution $\mu_1(\cdot \mid (Z_1, \delta_1), \ldots, (Z_n, \delta_n))$ is a Polya tree process with parameters $\boldsymbol{\pi}_n^{(\boldsymbol{Z},\boldsymbol{\delta})} = \{\pi^*_n(\boldsymbol{Z^*})\}_{n\geq 1}$ and $\alpha_n^{(\boldsymbol{Z},\boldsymbol{\delta})} = \{\acute{\alpha}_{\epsilon_1,\ldots,\epsilon_l}\}$, where $\acute{\alpha}_{\epsilon_1,\ldots,\epsilon_l} = \alpha_{\epsilon_1,\ldots,\epsilon_l} + \sum_{\delta_i=1} I[Z_i \in B_{\epsilon_1,\ldots,\epsilon_l}] + C^n_{\epsilon_1,\ldots,\epsilon_l}$*

[Remark: Note that if $B_{\epsilon_1,\ldots,\epsilon_l} = (Z_k, \infty)$ then $\acute{\alpha}_{\epsilon_1,\ldots,\epsilon_l} = \alpha(B_{\epsilon_1,\ldots,\epsilon_l})$ + no. of individuals surviving at time Z_k, and for every other $B_{\epsilon_1,\ldots,\epsilon_l}$ $\acute{\alpha}_{\epsilon_1,\ldots,\epsilon_l} = \alpha(B_{\epsilon_1,\ldots,\epsilon_l})$ + no. of uncensored observations in $B_{\epsilon_1,\ldots,\epsilon_l}$.]

Proof: Since $\mathcal{D}(\alpha)$ is a Polya tree process with respect to any sequence of partitions, it is a Polya tree with respect to the partitions $\boldsymbol{\pi}_n^{(\boldsymbol{Z},\boldsymbol{\delta})}$, with parameter $\boldsymbol{\alpha}$ $= \alpha(B_{\epsilon_1,\ldots,\epsilon_l})$. The proof now follows from Proposition 3.2.1, and the results on the posterior of a Polya tree process given exact and incomplete observations mentioned in Chapter 1. $\diamondsuit$

We will denote this posterior by $\mathrm{PT}(\boldsymbol{\pi}_n^{(\boldsymbol{Z},\boldsymbol{\delta})}, \alpha_n^{(\boldsymbol{Z},\boldsymbol{\delta})})$.

Note: Throughout this dissertation for any distribution function F, $\bar{F}(t) := 1 - F(t)$, for $t \in \mathbb{R}$.

Proposition 3.2.2 *Let $\delta_j = 0$, for some $1 \leq j \leq n$, then the Bayes estimate of $\bar{F}(z_j)$ given the observables $\{(z_1, \delta_1), \ldots, (z_n, \delta_n)\}$ is given by*
$\bar{F}_n(z_j) = \prod_{\{i: z^*_{(i)} \leq z_j\}} [\frac{\alpha(z^*_{(i)}, \infty) + n_i}{\alpha(z^*_{(i-1)}, \infty) + n_i + \lambda_i}]$,
where $n_i = \#\{z_k \geq z^*_{(i)}\}$, and $\lambda_i = \#\{z_k \in (z^*_{(i-1)}, z^*_{(i)} : \delta_k = 1\}$

Proof: The Bayes estimate of $\bar{F}(z_j)$ is given as $\bar{F}_n(z_j) = \int \bar{F}(z_j) d\mu_1(F \mid (z_1, \delta_1), \ldots, (z_n, \delta_n))$. Using the Polya tree representation for the posterior $\mu_1(\cdot \mid (z_1, \delta_1), \ldots, (z_n, \delta_n))$, the result follows from the properties of Polya tree processes mentioned in Chapter 1. $\Diamond$

Let $M_0 \subset M(\mathbb{R}^+)$ be the class of all distribution functions F, such that

1. F and G_0 have no points of discontinuity in common, and

2. Support$(F) \subset$ Support(G_0)

Let $F_0 \in M_0$, and consider the set V_{F_0} of all sequences $\{(z_n, \delta_n)\}_{n \geq 1}$ such that

1. For any (z_j, δ_j) with $\delta_j = 0$, a) $\frac{1}{n} \sum_{i=0}^{n} I\{z_i \geq z_j\} \to \bar{F}_0(z_j)\bar{G}_0(z_j-)$, and b) $\bar{\hat{G}}_n(z_j-) \to \bar{G}_0(z_j-)$, where $\hat{G}_n$ is the Kaplan Meier estimate of G,

2. $\{z_j : \delta_j = 0\}$ is a dense subset in the support of F_0.

Remark: It follows from the SLLN that , $\frac{1}{n} \sum_{i=0}^{n} I\{Z \geq z\} \to \bar{F}_0(z)\bar{G}_0(z-)$, a.s. (F_0, G_0), whenever z is a point of discontinuity of G_0. Also the SLLN for censored data [Stute and Wang (1994)], implies that $\bar{\hat{G}}_n(z-) \to \bar{G}_0(z-)$.
Therefore $P^{\infty}_{F_0,G_0}(V_{F_0}) = 1$. To prove any consistency result for the posterior at (F_0, G_0), it is enough to prove the result for $\{(z_n, \delta_n)\}_{n \geq 1} \in V_{F_0}$.

Theorem 3.2.2 *Let* $\{(z_n, \delta_n)\}_{n\geq 1} \in V_{F_0}$, *then the Bayes estimate of* F,

$$\hat{F}_n(\cdot \mid (z_1,\delta_1),\ldots,(z_n,\delta_n)) = \int F(\cdot)PT((\pi_n^{(\boldsymbol{z},\boldsymbol{\delta})}, \alpha_n^{(\boldsymbol{z},\boldsymbol{\delta})})(dF),$$

converges weakly to $F_0(\cdot)$.

Proof: Let us consider a fixed sequence $\{(z_j,\delta_j)\}_{n\geq 1} \in V_{F_0}$, and let (z_j,δ_j) be a coordinate such that $\delta_j = 0$. By our assumptions on V_{F_0}, it is enough to show that $\bar{F}_n(z_j) \to \bar{F}_0(z_j)$. For simplicity of notation we will assume that $\delta_1 = 0$, and we will show that $\bar{F}_n(z_1) \to \bar{F}_0(z_1)$.

Let $0 = z_{(0)} < z_{(1)} < \ldots < z_{(n(1))} = z_1$, be the zs among $\{(z_1,\delta_1),\ldots,(z_n,\delta_n)\}$ for which the corresponding δs are 0, and are $\leq z_1$. By proposition 2,

$$\bar{F}_n(z_1) = \prod_{j=1}^{n(1)} \frac{\alpha(z_{(j)},\infty)+n_j}{\alpha(z_{(j-1)},\infty)+n_j+\lambda_j},$$

where $n_j = \sum_k I\{z_k \geq z_{(j)}\}$, and $\lambda_j = \sum_{k:\delta_k=1} I\{z_k \in (z_{(j-1)}, z_{(j)})\}$

Rewriting the expression on the right hand side of the last equation, we get,

$$\begin{aligned}\bar{F}_n(z_1) &= \frac{\alpha(z_1,\infty)+\sum I\{z_i \geq z_1\}}{\alpha(\mathbb{R}^+)+n} \times \prod_{j=1}^{n(1)-1} \frac{\alpha(z_{(j)},\infty)+n_j}{\alpha(z_{(j)},\infty)+n_{j+1}+\lambda_{j+1}} \\ &= A_n(z_1) \times B_n(z_1),\end{aligned}$$

where $A_n(z_1) = \frac{\alpha(z_1,\infty)+\sum I\{z_i\geq z_1\}}{\alpha(\mathbb{R}^+)+n}$, and $B_n(z_1) = \prod_{j=1}^{n(1)-1} \frac{\alpha(z_{(j)},\infty)+n_j}{\alpha(z_{(j)},\infty)+n_{j+1}+\lambda_{j+1}}$.

Since $\{(z_j,\delta_j)\}_{n\geq 1} \in V_{F_0}$ and $\delta_1 = 0$, $A_n(z_1) \to \bar{G}_0(z_1-)\bar{F}_0(z_1)$.

Let $d_j = \#\{z_i : \delta_i = 0, and z_i = z_{(j)}\}$, then $n_{i+1} + \lambda_{i+1} = n_i - d_i$, and hence $B_n(Z_1) = \prod_{j=1}^{n(1)-1} \frac{\alpha(z_{(j)},\infty)+n_j}{\alpha(z_{(j)},\infty)+n_j-d_j}$. Therefore

$$\prod_{j=1}^{n(1)-1} \frac{n_j-d_j}{n_j} \leq (B_n(z_1))^{-1} \leq \prod_{j=1}^{n(1)-1} \frac{\alpha(\mathbb{R}^+)+n_j-d_j}{\alpha(\mathbb{R}^+)+n_j}$$

Observing that $\prod_{j=1}^{n(1)-1} \frac{n_j - d_j}{n_j}$ is the Kaplan Meier estimate for $\bar{G}(z_1-)$, and hence converges to $\bar{G}_0(z_1-)$, we can conclude that $B_n(z_1) \to (G_0(z_1-))^{-1}$. Therefore $\bar{F}_n(z_1) = A_n(z_1) \times B_n(z_1) \to \bar{F}_0(z_1)$.

$\Diamond$

Combining the above theorem with the fact that $P^{\infty}_{F_0,G_0}(V_{F_0}) = 1$, we get the following corollary.

Corollary 3.2.3 *Suppose $\mathcal{D}(\alpha)$ gives mass one to M_0, then the Bayes estimate of F*

$$\hat{F}_n(\cdot \mid (Z_1, \delta_1), \ldots, (Z_n, \delta_n)) = \int F(\cdot) PT((\boldsymbol{\pi}_n^{(\boldsymbol{Z},\boldsymbol{\delta})}, \alpha_n^{(\boldsymbol{Z},\boldsymbol{\delta})})(dF),$$

converges weakly to $F_0(\cdot)$ a.s. P_{F_0,G_0}, for $F_0 \in M_0$.

Remark: It can be seen via Sethuraman's construction (Theorem 1.5.2) of the Dirichlet process that it is possible to choose $\mathcal{D}(\alpha)$ such that $\mathcal{D}(\alpha)(M_0) = 1$. This is achieved by choosing α such that:
a) Support(α) $\subset$ Support(G_0), and b) α and G_0 have no discontinuity points in common.

Theorem 3.2.4 *Let $\mu = \mathcal{D}(\alpha) \times \delta_{G_0}$ be the prior on $M(\mathbb{R}^+) \times M(\mathbb{R}^+)$, where $\mathcal{D}(\alpha)(M_0) = 1$. Let $F_0 \in M_0$, then the marginal posterior on $M(\mathbb{R}^+)$ is weakly consistent at F_0.*

Proof: We will show that the marginal posterior $PT((\boldsymbol{\pi}_n^{(\boldsymbol{z},\boldsymbol{\delta})}, \alpha_n^{(\boldsymbol{z},\boldsymbol{\delta})}) \Rightarrow \delta_{F_0}$, for all $\{(z_j, \delta_j)\}_{n \geq 1} \in V_{F_0}$. By Proposition 1.4.2, and Theorem 3.2.3, the posterior sequence $PT((\boldsymbol{\pi}_n^{(\boldsymbol{z},\boldsymbol{\delta})}), \alpha_n^{(\boldsymbol{z},\boldsymbol{\delta})})$ is a tight family of probability measures on $M(\mathbb{R}^+)$. To complete the proof, it is enough to show that for any f continuous function on $\mathbb{R}^+$, with a compact support and any $\delta > 0$,

$PT(\boldsymbol{\pi}_n^{(\boldsymbol{z},\boldsymbol{\delta})}, \alpha_n^{(\boldsymbol{z},\boldsymbol{\delta})})(U_{F_0}^{\delta}) \to 1$, where

$$U_{F_0}^{\delta} = \{F :\mid \int f dF - \int f dF_0 \mid < \delta\}.$$

Let us fix the sequence $\{(z_j, \delta_j)\}_{n\geq 1} \in V_{F_0}$. For the remaining portion of the proof, we will write $\boldsymbol{\pi}_n$ for $\boldsymbol{\pi}_n^{(\boldsymbol{z},\boldsymbol{\delta})}$, and α_n for $\alpha_n^{(\boldsymbol{z},\boldsymbol{\delta})}$. Also, let $D = \{z_j : \delta_j = 0\}$.

Let f have support $[0, k]$, and let γ be such that $\mid x-y \mid < \gamma$ implies $\mid f(x)-f(y) \mid < \delta/3$.

Let $0 = a_1 < a_2 < \cdots < a_{l+1} = k$, be such that $\mid a_{i+1} - a_i \mid < \gamma/2$, and let $z_{(1)} < z_{(2)} < \cdots < z_{(l)}$, with $z_{(i)} \in D$, and $z_{(i)} \in (a_i, a_{i+1})$.

Let $f_\delta(z) := \sum_1^{l+1} f(z_{(i)}) I\{z \in (z_{(i-1)}, z_{(i)}]\}$, with $z_{(0)} = 0$, and $z_{(l+1)} = k$. Then $\parallel f - f_\delta \parallel < \frac{\delta}{3}$, where $\parallel g \parallel = sup_x \mid g(x) \mid$.

Further,

$$\mid \int f dF - \int f dF_0 \mid \leq \frac{2\delta}{3} + \mid \int f_\delta dF - \int f_\delta dF_0 \mid .$$

Let $U_\delta^1 = \{F :\mid \int f_\delta dF - \int f_\delta dF_0 \mid < \frac{\delta}{3}\}$, then $U_\delta^1 \subset U_{F_0}^{\delta}$.

For any F, $\int f_\delta dF = \sum f(z_{(i)})[\bar{F}(z_{(i-1)}) - \bar{F}(z_{(i)})]$, and hence

$$\mid \int f_\delta dF - \int f_\delta dF_0 \mid \leq 2 \parallel f \parallel \sum \mid \bar{F}_0(z_{(i)}) - \bar{F}(z_{(i)}) \mid,$$

Observing that $PT(\boldsymbol{\pi}_n, \alpha_n)(U_{F_0}^{\delta}) \geq PT(\boldsymbol{\pi}_n, \alpha_n)(U_\delta^1)$, to complete our proof, it is enough to show that $PT(\boldsymbol{\pi}_n, \alpha_n)(F :\mid \bar{F}(z_{(i)} - \bar{F}_0(z_{(i)} \mid > \eta) \to 0$, for every $\eta > 0$. This follows by an application of Markov's inequality, if we are able to show that

$E(\bar{F}(t) \mid (z_1, \delta_1), \ldots, (z_n, \delta_n)) \to \bar{F}_0(t)$, and

$E((\bar{F}(t))^2 \mid (z_1, \delta_1), \ldots, (z_n, \delta_n)) \to (\bar{F}_0(t))^2$ for all $t \in D$.

In the proof of Theorem 3.2.3, we have already seen that

$E(\bar{F}(t) \mid (z_1, \delta_1), \ldots, (z_n, \delta_n)) \to \bar{F}_0(t)$, for all $t \in D$.

We will now show that

$E((\bar{F}(t))^2 \mid (z_1,\delta_1),\ldots,(z_n,\delta_n)) \to (\bar{F}_0(t))^2$, for all $t \in D$.

For simplicity of notation, let us assume that $z_1 \in D$, and we will use the same notation as in the proof of Theorem 3.2.3. Using the properties of Polya tree processes and the Polya tree representation for the posterior, we have

$E((\bar{F}(z_1))^2 \mid (z_1,\delta_1),\ldots,(z_n,\delta_n)) = A_n^*(z_1) \times B_n^*(z_1)$,

where $A_n^*(z_1) = \frac{\alpha(z_1,\infty)+\sum I\{z_i \geq z_1\}}{\alpha(\mathbb{R}^+)+n} \times \frac{\alpha(z_1,\infty)+\sum I\{z_i \geq z_1\}+1}{\alpha(\mathbb{R}^+)+n+1}$ and

$B_n^*(z_1) = \prod_{j=1}^{n(1)-1} \frac{\alpha(z_{(j)},\infty)+n_j}{\alpha(z_{(j)},\infty)+n_{j+1}+\lambda_{j+1}} \times \prod_{j=1}^{n(1)-1} \frac{\alpha(z_{(j)},\infty)+n_j+1}{\alpha(z_{(j)},\infty)+n_{j+1}+\lambda_{j+1}+1}$.

An argument similar to the one used in the proof of Theorem 3.2.3, now yields

$A_n^*(z_1) \to (\bar{G}_0(z_1-)\bar{F}_0(z_1))^2$, and $B_n^*(z_1) \to (\bar{G}_0(z_1))^{-2}$,

and thus $E((\bar{F}(z_1))^2 \mid (z_1,\delta_1),\ldots,(z_n,\delta_n)) \to (\bar{F}_0(z_1))^2$. Similarly

$E((\bar{F}(t))^2 \mid (z_1,\delta_1),\ldots,(z_n,\delta_n)) \to (\bar{F}_0(t))^2$ for all $t \in D$. $\diamondsuit$

3.3 Priors on the distribution of the observables

As in Section 3.2, here also we work under the following set up: X and Y are non-negative random variables corresponding to life time and censoring time, with distribution $\mathrm{F} \in M(\mathbb{R}^+)$ and $\mathrm{G} \in M(\mathbb{R}^+)$. Also $\{\mathrm{F}, X_1, X_2, \ldots, X_n\}$, and $\{\mathrm{G}, Y_1, Y_2, \ldots, Y_n\}$ are independent. The observed data is $\{(Z_1,\delta_1),\ldots,(Z_n,\delta_n)\}$. In this section we consider priors for the distribution of the observable random variables names $\{Z,\delta\}$, and study the posterior given $\{(Z_1,\delta_1),\ldots,(Z_n,\delta_n)\}$.

Let $M^{X,Y} \subset M(\mathbb{R}^+) \times M(\mathbb{R}^+)$ denote the collection of all pairs of distribution functions (F,G) such that

1. F and G have the same support

2. F and G do not have any discontinuity points in common.

We equip $M^{X,Y}$ with the two dimensional Kolmogorov metric d_k^2, defined as $d_k^2((F_1,G_1),(F_2,G_2)) = d_k(F_1,F_2) + d_k(G_1,G_2)$.

We write $M^*(\mathbb{R}^+)$ for $M(\mathbb{R}^+ \times \{0,1\})$ and observe that any $H \in M^*(\mathbb{R}^+)$ is identified by the pair of sub-survival functions (H_0, H_1), where $H_0(t) = H((t,\infty) \times 0)$, and $H_1(t) = H((t,\infty) \times 1)$. We view $M^*(\mathbb{R}^+)$ as the space of sub-survival functions (H_0, H_1), where H_0 and H_1 are right continuous and non-increasing satisfying:
$H_0(0) + H_1(0) = 1$,
and $lim_{t\to\infty}\{H_0(t) + H_1(t)\} = 0$.

Let $M^{Z,\delta} \subset M^*(\mathbb{R}^+)$, be the collection of all $(H_0, H_1) \in M^*(\mathbb{R}^+)$ such that

1. H_0 and H_1 have the same support.

2. H_0 and H_1 do not have any discontinuity points in common.

On $M^{Z,\delta}$, we consider the appropriate 'Kolmogorov' metric d_k^* defined as
$d_k^*((H_0, H_1), (H_0^*, H_0^*)) = sup_t \mid H_0(t) - H_0^*(t) \mid + sup_t \mid H_1(t) - H_1^*(t) \mid$

In this section we will restrict ourselves to $M^{X,Y}$ and $M^{Z,\delta}$ as our parameter spaces of interest.

$$\begin{aligned} \text{Let } T : M^{X,Y} &\mapsto M^{Z,\delta} \\ (F,G) &\mapsto \text{Distribution of } (Z,\delta) \text{ when} X \sim F,\ Y \sim G; \\ &\quad X \text{ and } Y \text{ are independent} \end{aligned}$$

i.e. $T(F,G) = (H_0, H_1)$ such that

$$\begin{aligned} H_0(t) &= \int_t^\infty \bar{F}(s)dG(s) \\ H_1(t) &= \int_t^\infty \bar{G}(s)dF(s) \end{aligned}$$

That the map T is 1-1 is a consequence of the identifiability property of (Z,δ).

Peterson (1977) defines a map

$$\Phi : M^{Z,\delta} \mapsto M^{X,Y},$$

using which one can conclude that T is also onto, and $T^{-1} = \Phi$. The following facts summarize the properties of T that we are interested in.

Fact 1. T is a continuous map from $(M^{X,Y}, d_k^2)$ to $(M^{Z,\delta}, d_k^*)$. (Follows easily from the representation for $T(F,G)$ given above.)

Fact 2.(Peterson (1977)). $T^{-1}(= \Phi)$ is a continuous map from $(M^{Z,\delta}, d_k^*)$ to $(M^{X,Y}, d_k^2)$.

These observations lead to the following theorem.

Theorem 3.3.1 *Let μ be a prior on $M^{Z,\delta}$, and let $\tilde{\mu}$ denote the induced prior on $M^{X,Y}$ via the map Φ, i. e. $\tilde{\mu} = \mu \circ T$. If $\mu(\cdot \mid (Z_1,\delta_1),\ldots,(Z_n,\delta_n))$ is k-consistent at (H_0,H_1), then $\tilde{\mu}(\cdot \mid (Z_1,\delta_1),\ldots,(Z_n,\delta_n))$ is k-consistent at $T^{-1}(H_0,H_1)$.*

A natural class of priors on $M^{Z,\delta}$ is the Dirichlet process priors. These priors were first explored in this set-up by Tsai (1986), who used them to construct a class of self consistent estimates. In this context, using Theorem 1.5.1 (rather Proposition 1.5.2), we can conclude that the posterior given $\{(Z_1,\delta_1),\ldots,(Z_n,\delta_n)\}$ on $M^{Z,\delta}$ is k-consistent at all $(H_0,H_1) \in M^{Z,\delta}$. Theorem 3.3.1 now ensures that we infact have k-consistency for the sequence of posteriors for the induced prior $\mathcal{D}(\alpha) \circ T$ on $M^{X,Y}$ given $\{(Z_1,\delta_1),\ldots,(Z_n,\delta_n)\}$. A consistency result of this sort was proved by Ghosh and Ramamoorthi (1995). However they had mistakenly assumed that $\mathcal{D}(\alpha_1 \times \alpha_2) = \mathcal{D}(\alpha_1) \times \mathcal{D}(\alpha_2)$ and thus believed that they had a version of Theorem 3.2.4. A careful look at their proof shows that it is a special case of Theorem 3.3.1.

Another class of priors on $M^{Z,\delta}$ would be Polya tree priors. Since Polya tree priors can be made to sit on densities, this method gives us (induced) priors on

$M^{X,Y}$ sitting on densities which yield consistent posteriors. The consistency is an immediate consequence of Theorem 3.3.1, since if (F_0, G_0) are continuous, the weak neighbourhoods and k-neighbourhoods coincide.

Theorem 3.3.1 thus provides a wide class of priors for (F, G) which ensures posterior consistency in the context of right censored data. However, there do arise interpretational difficulties. Manageable priors on $M^{Z,\delta}$, like the Dirichlet processes or Polya tree processes, when transferred to $M^{X,Y}$ gives rise to a prior which is not of the form $\mu_1 \times \mu_2$, and hence under the predictive distribution X and Y will not be independent.

CHAPTER 4

Nonparametric Bayesian inference with interval censored observations

4.1 Introduction and summary

In some medical follow up studies and epidemiologic investigations, continuous monitoring for outcome variable of interest is impractical, and assessment of study subjects can only take place at deterministic or random time points. The precise time at which the outcome occurred is not observed, but it is either known to have taken place within a specified time interval determined from two consecutive examination times, or not to have occurred by the last available assessment. The time of occurrence T of the event is said to be interval censored, and one of the important statistical issues is the problem of inference on the distribution of the time of occurrence from a sample of interval censored observations. The problem that we consider in this chapter is the simplest form of interval censoring with a single examination time called Case 1 interval censoring and is described below.

Let $X_1, X_2, \ldots, X_n$ and $Y_1, Y_2, \ldots, Y_n$ be non-negative i.i.d. random variables with distributions F and G respectively. We view the Xs as lifetimes and the Ys as inspection times and also assume that the Xs and Ys are independent. The observed

data are $\{ (Y_1, \delta_1), \ldots, (Y_n, \delta_n) : i = 1, \ldots, n\}$, where $\delta_i = I[X_i \leq Y_i]$, and the goal is to make inferences on F based on $\{(Y_1, \delta_1), \ldots, (Y_n, \delta_n)\}$. This problem has been studied in a non Bayesian framework by among others, Turnbull (1976), Groeneboom and Wellner (1992), and Wang and Gardiner (1996). Groeneboom and Wellner (1992) study in good detail the asymptotic properties of the 'Turnbull estimator' - the non parametric maximum likelihood estimator (NPMLE) of F. Our research in the interval censoring problem was driven by a desire to obtain good Bayesian estimates for F, and also to give a Bayesian interpretation to the Turnbull estimator. The material that is presented in this chapter is essentially a description of the complications associated in using the approaches that work 'well' in the right censoring problem, described in Chapter 3. Unfortunately, our attempts have not been very successful.

Similar to the discussion in Chapter 3, for Bayesian inference in a interval censored data context, priors can be constructed in two natural ways. One way is to consider a prior for F directly. Wang (1993) makes an attempt in this direction by considering a Dirichlet process prior for F. This approach is discussed in Section 2. Another method is to consider a prior for the distribution of the observables $\{(Y_1, \delta_1), \ldots, (Y_n, \delta_n)\}$, and then use identifiability conditions to transfer to a prior for (F, G). We discuss this approach in Section 3.

4.2 Dirichlet process priors for F

The set-up that we consider can be described as follows: X and Y are non-negative random variables corresponding to life time and inspection time, with distribution $\mathrm{F} \in M(\mathbb{R}^+)$ and $\mathrm{G} \in M(\mathbb{R}^+)$. Also $\{\mathrm{F}, X_1, X_2, \ldots, X_n\}$, and $\{\mathrm{G}, Y_1, Y_2, \ldots, Y_n\}$ are independent. The observed data is $\{(Y_1, \delta_1), \ldots, (Y_n, \delta_n)\}$. We consider priors of the form $\mathcal{D}(\alpha) \times \delta_{G_0}$ for (F, G) Our interest is in the Bayes estimate of F and the (marginal) posterior distribution

$\mathcal{D}(\alpha)(\cdot \mid (Y_1, \delta_1), \ldots, (Y_n, \delta_n))$ (of F) on $M(\mathbb{R}^+)$ given $\{(Y_1, \delta_1), \ldots, (Y_n, \delta_n)\}$.

The next proposition follows from Theorem 2 of Blum and Susarla (1977).

Proposition 4.2.1 *Let $\mu = \mathcal{D}(\alpha) \times \delta_{G_0}$ be the prior (for (F, G)) on $M(\mathbb{R}^+) \times M(\mathbb{R}^+)$. Then the posterior distribution (of F) given $\{(Y_1, \delta_1), \ldots, (Y_n, \delta_n)\}$ is a mixture of Dirichlet processes.*

[Remark: As Theorem 2 of Blum and Susarla (1977) suggests, this mixture representation could be pretty complicated.]

Wang (1993) suggests a way of calculating the Bayes estimate of $\bar{F}$ (and hence F) given $\{(Y_1, \delta_1), \ldots, (Y_n, \delta_n)\}$, and highlights the computational difficulties that arise in this case. Also, unlike the right censoring set-up, the limit of the Bayes estimate as $\alpha(\mathbb{R}^+) \to 0$, does not always equal the NPMLE (the Turnbull estimator). Wang(1993, pages 41-42) has an example where the limit of the Bayes estimates as $\alpha(\mathbb{R}^+) \to 0$ is not the NPMLE.

We mention below four examples that illustrates how the limit of Bayes estimates behaves and its relationship to the NPMLE. In all the four examples the Bayes estimate of F is derived under the $\mathcal{D}(\alpha) \times \delta_{G_0}$ prior for (F, G). Also, $\hat{F}$ will denote the NPMLE while $\tilde{F}$ will denote the limit of the Bayes estimates as $\alpha(\mathbb{R}^+) \to 0$. Since it is clear that if $Y_i = y$ and $\delta_i = 1$, then with regard to X_i, the only information we have is $X_i \leq y$, and if $Y_i = y$ and $\delta_i = 0$, then with regard to X_i, the only information we have is $X_i > y$, the oberved data set will be presented in terms of the intervals that contains X_i. For the remaining portion of this chapter, for any distribution function F, we will write F(a, b] to represent F(b) - F(a).

Example 1.: This example illustrates that the limit of Bayes estimates could be supported on a much bigger set than the NPMLE. The observed data consists of the four intervals (1, ∞), (2, ∞), (0, 3], (4, ∞).

The limit of Bayes estimates in this case turns out to be,

$\tilde{F}(0,1] = \frac{1}{22}$,

$\tilde{F}(1,2] = \frac{2}{22}$,

$\tilde{F}(2,3] = \frac{6}{22}$,

$\tilde{F}(4,\infty] = \frac{13}{22}$,

while the NPMLE is given by,

$\hat{F}(2,3] = \frac{1}{2}$,

$\hat{F}(4,\infty] = \frac{1}{2}$.

Example 2.: This example illustrates that the limit of Bayes estimates could be supported on the same set as the NPMLE, but they may still be different. The observed data consists of the intervals (0, 1], (2, ∞), (3, ∞), (0, 4].

The limit of Bayes estimates in this case turns out to be,

$\tilde{F}(0,1] = \frac{3}{8}$,

$\tilde{F}(3,4] = \frac{5}{8}$,

while the NPMLE is given by,

$\hat{F}(0,1] = \frac{1}{3}$,

$\hat{F}(3,4] = \frac{2}{3}$.

Example 3.: This is an example where the limit of Bayes estimates and the NPMLE are the same. The observed data consists of the intervals (0, 1], (2, ∞), (0, 3], (4, ∞).

The limit of Bayes estimates in this case turns out to be,

$\tilde{F}(0,1] = \frac{1}{2}$,

$\tilde{F}(4,\infty) = \frac{1}{2}$,

while the NPMLE is given by,

$\hat{F}(0,1] = \frac{1}{2}$,

$\hat{F}(4,\infty) = \frac{1}{2}$.

Example 4.: This example illustrates that the limit of Bayes estimates could be supported on a smaller set than the NPMLE. The observed data consists of the intervals (0, 1], (2, ∞), (0, 3], (0, 4], (5, ∞).

The limit of Bayes estimates in this case turns out to be,

$\tilde{F}(0,1] = \frac{3}{5}$,

$\tilde{F}(5,\infty) = \frac{2}{5}$,

while the NPMLE is given by,

$\hat{F}(0,1] = \frac{1}{2}$,

$\hat{F}(2,3] = \frac{1}{6}$,

$\hat{F}(5,\infty) = \frac{1}{3}$.

To gain further insight into the behaviour of the limit of Bayes estimates, we introduce the notion of 'allocation', and 'minimal allocation' for interval censored data.

An 'allocation' based on the observed data (intervals) consists of a) the smallest sub-intervals formed by using the left end points and right end points of the observed intervals, and which can account for all the (unobserved) X_i values, (i.e. the unobserved X_is can all be classified as belonging to one [and only one] of the sub-intervals;), and b) the numbers representing the no. of X_is that can be assigned to each sub-interval.

For instance in Example 1. above, an allocation could consist of the intervals (0, 1], (2, 3], and (4, ∞), with the corresponding number of X_is assigned to the above intervals possibly being 1, 1, and 2 respectively. As is obvious, in any particular example there could be many different possible allocations.

A 'minimal allocation' is an allocation consisting of fewest number of sub-intervals among all allocations.

In Example 1. above, a minimal allocation will consist of only 2 sub-intervals.
(i) (0,1], and (4, ∞), with the corresponding numbers of X_is in the sub-intervals being 1 and 3 respectively, represents a minimal allocation.
(ii) (2, 3] and (4, ∞) with the corresponding numbers of X_is in the sub-intervals being 1 and 3 respectively represents another minimal allocation.
(iii) (2, 3] and (4, ∞) with the corresponding numbers of X_is in the sub-intervals being 2 and 2 respectively represents yet another minimal allocation.

A minimal allocation will be called a 'unique minimal allocation' if there is no other minimal allocation (taking into account both the no. of subintervals included in the allocation and the no. of ways in which the (unobserved) X_is can be assigned to the sub-intervals).

For instance, in Example 2. above, the allocation including the intervals (0, 1] and (3, 4] is a minimal allocation, but not a unique minimal allocation, but in example 3. above, the allocation consisting of the intervals (0, 1], and (4, ∞), with the corresponding numbers of X_i's in the sub-intervals being 2 and 2 respectively, represents a unique minimal allocation.

Based on the above examples and some elementary analysis we have the following conjectures about the behaviour of the limit of Bayes estimates and its relationship with the NPMLE.

Conjecture 1. If there is a unique minimal allocation and the intervals included in the allocation are exactly the same as the intervals which are assigned positive mass by the NPMLE, then, the limit of Bayes estimates (as $\alpha(\mathbb{R}^+) \to 0$) is the NPMLE.

Conjecture 2. The limit of Bayes estimates (as $\alpha(\mathbb{R}^+) \to 0$) assigns positive mass to only those intervals which appear in at least one 'minimal allocation, and vice versa.

4.3 Priors on the distribution of the observables

As in Section 4.2, here also we are working under the following set up: X and Y are non-negative random variables corresponding to life time and inspection time, with distribution F $\in M(\mathbb{R}^+)$ and G $\in M(\mathbb{R}^+)$. Also
{F, $X_1, X_2, \ldots, X_n$}, and {G, $Y_1, Y_2, \ldots, Y_n$} are independent. The observed data is $\{(Y_1, \delta_1), \ldots, (Y_n, \delta_n)\}$.

Let $M^{X,Y} \subset M(\mathbb{R}^+) \times M(\mathbb{R}^+)$ denote the collection of all pairs of distribution functions (F, G) such that

1. $0 < F(x) < 1$ for $x \in \mathbb{R}^+$, and F is continuous on $\mathbb{R}^+$.

2. G is absolutely continuous (with respect to Lebesgue measure) having a density g which is continuous and positive on $\mathbb{R}^+$.

We equip $M^{X,Y}$ with the metric $d_{k,t}$, defined as
$d_{k,t}((F_1, G_1), (F_2, G_2)) = d_k(F_1, F_2) + d_t(G_1, G_2)$, where
$d_t(G_1, G_2) = \frac{1}{2} \int \mid g_1 - g_2 \mid d\lambda$, with λ denoting the lebesgue measure on $\mathbb{R}^+$.

As in Chapter 3, we write $M^*(\mathbb{R}^+)$ for $M(\mathbb{R}^+ \times \{0, 1\})$ and identify any $H \in M^*(\mathbb{R}^+)$ by the pair of sub-distribution functions (H_0, H_1), where $H_0(t) = H((0, t] \times 0)$, and $H_1(t) = H((0, t] \times 1)$. We view $M^*(\mathbb{R}^+)$ as the space of sub-distribution functions (H_0, H_1), where H_0 and H_1 are right continuous and non-decreasing satisfying: $H_0(0) + H_1(0) = 0$, and $lim_{t \to \infty}\{H_0(t) + H_1(t)\} = 1$.

Let $M^{Y,\delta} \subset M^*(\mathbb{R}^+)$, be the collection of all $(H_0, H_1) \in M^*(\mathbb{R}^+)$ such that

1. $H_1(x) = \int_0^x F(u)g(u)du$,

2. $H_0(x) = \int_0^x \bar{F}(u)g(u)du$,

for some $(F, G) \in M^{X,Y}$.

On $M^{Y,\delta}$, we consider the appropriate 'total variation metric' d_t^* defined as

$d_t^*((H_0, H_1), (H_0^*, H_1^*)) =$

$\frac{1}{2}\int \mid \bar{F}(u)g(u) - \bar{F}^*(u)g^*(u) \mid du + \frac{1}{2}\int \mid F(u)g(u) - F^*(u)g^*(u) \mid du.$

In this section we will restrict ourselves to $M^{X,Y}$, and $M^{Y,\delta}$, as our parameter spaces of interest.

$$\begin{aligned} \text{Let } T : M^{X,Y} &\mapsto M^{Y,\delta} \\ (F, G) &\mapsto \text{Distribution of } (Y, \delta) \text{ when} X \sim F,\ Y \sim G; \\ &\quad\ X \text{ and } Y \text{ are independent} \end{aligned}$$

i.e. $T(F, G) = (H_0, H_1)$ such that

$$\begin{aligned} H_0(x) &= \int_0^x F(u)g(u)du, \\ H_1(x) &= \int_0^x \bar{F}(u)g(u)du. \end{aligned}$$

The following proposition summarizes the properties of the map T, that are of interest to us.

Proposition 4.3.1 *The map T from $(M^{X,Y}, d_{k,t})$ to $(M^{Y,\delta}, d_t^*)$, is 1-1, onto, and T and T^{-1} are both continuous.*

Proof: That the map T is onto follows from its very definition. From Theorem 1, of Wang et al. (1994), it follows that T is 1-1.

Let $\{\{(F_n, G_n)\}_{n\geq 1}, (F, G)\} \subset M^{X,Y}$ be such that $d_{k,t}((F_n, G_n), (F, G)) \to 0$. To

prove that T is continuous, we need to show that $d_t^*((H_{n0}, H_{n1}), (H_0, H_1)) \to 0$, with

$$\begin{aligned} H_{n0}(x) &= \int_0^x F_n(u) g_n(u) du, \\ H_{n1}(x) &= \int_0^x \bar{F}_n(u) g_n(u) du, \\ H_0(x) &= \int_0^x F(u) g(u) du, \\ H_1(x) &= \int_0^x \bar{F}(u) g(u) du, \end{aligned}$$

where g_n is a density of G_n, and g is a density of G, both with respect to Lebesgue measure. In fact, it is enough to show that
$\int \mid F_n(u) g_n(u) du - F(u) g(u) \mid du \to 0$. By the triangle inequality,

$$\begin{aligned} & \int \mid F_n(u) g_n(u) du - F(u) g(u) \mid du \\ \leq & \int \mid F_n(u) g_n(u) - F(u) g_n(u) \mid du + \int \mid F(u) g_n(u) - F(u) g(u) \mid du \\ \leq & \mid\mid F_n - F \mid\mid + \int \mid g_n(u) - g(u) \mid du, \end{aligned}$$

which converges to zero by our hypothesis.

Let $\{\{(H_{n0}, H_{n1})\}_{n \geq 1}, (H_0, H_1)\} \subset M^{Y,\delta}$ be such that
$d_t^*((H_{n0}, H_{n1}), (H_0, H_1)) \to 0$, where H_{n0}, H_{n1}, H_0, and H_1, have the same representation as above. To prove that T^{-1} is continuous, we need to show that
$d_{k,t}((F_n, G_n), (F, G)) \to 0$.

Observing that $H_1(x) + H_0(x) = G(x)$, and $H_1^n(x) + H_0^n(x) = G_n(x)$, it follows that $d_t(G_n, G) \to 0$. To complete our proof, we need to show that $d_k(F_n, F) \to 0$. This is proved by showing that every subsequence of $\{F_n\}_{n \geq 1}$ converges to F in the d_k metric. Since the F_ns are distribution functions on $\mathbb{R}^+$, every sequence has a further subsequence that converges to a sub-distribution function F^* at all continuity points

of F^*. Since F is continuous, our proof will be complete, if we can show that $F = F^*$, a.e. Lebesgue measure. By the dominated convergence theorem, $\int \mid F_n(u)g(u) - F^*(u)g(u) \mid du \to 0$. On the other hand, noting that

$$\begin{aligned} & \mid F_n(u)g(u) - F(u)g(u) \mid \\ \leq & \mid F_n(u)g(u) - F_n(u)g_n(u) \mid + \mid F_n(u)g_n(u) - F(u)g(u) \mid, \end{aligned}$$

we can conclude that $\int \mid F_n(u)g(u) - F(u)g(u) \mid du \to 0$. Since $g(u) > 0$, for all $u \in \mathbb{R}^+$, we can now conclude that $F(u) = F^*(u)$ a. e. Lebesgue.

$\Diamond$

This leads us to the following theorem.

Theorem 4.3.1 *Let μ be a prior on $M^{Y,\delta}$, and let $\tilde{\mu}$ denote the induced prior on $M^{X,Y}$ via the map T^{-1}, i. e. $\tilde{\mu} = \mu \circ T$. If $\mu(\cdot \mid (Y_1, \delta_1), \ldots, (Y_n, \delta_n))$ is t-consistent at (H_0, H_1), then $\tilde{\mu}(\cdot \mid (Y_1, \delta_1), \ldots, (Y_n, \delta_n))$ is (k,t)-consistent at $T^{-1}(H_0, H_1)$.*

Since Polya tree priors can be made to sit on densities, it is very tempting to think that Polya tree priors would be a natural class of priors on $M^{Y,\delta}$, and then conclude that this method gives us (induced) priors on $M^{X,Y}$ which yield consistent posteriors. But, unfortunately we do not have a way of constructing Polya tree priors that will sit on $M^{Y,\delta}$ and even more importantly, we not know whether Polya tree priors will yield posteriors that are t-consistent. Also, we do not know of any family of priors that will give mass one to $M^{Y,\delta}$, leave alone yield t-consistent posteriors, and hence Theorem 4.3.1 is not of much use.

Bibliography

[1] Blackwell, D. (1973). Discreteness of Ferguson selections. *Ann. Statist.* **1**, 356-358.

[2] Blum, J. and Susarla, V. (1977). On the posterior distribution of a Dirichlet process given randomly right censored observations. *Stoch. Processes Appl.* **5**, 207-211.

[3] Dalal, S. R. (1979). Dirichlet invariant processes and applications to nonparametric estimation of symmetric distribution functions. *Stoch. Proc. App.* **9**, 99-107.

[4] Diaconis, P. and Freedman, D. (1986a). On the consistency of Bayesestimates (with discussion). *Ann. Statist.* **14**, 1-67.

[5] Diaconis, P. and Freedman, D. (1986b). On inconsistent Bayes estimates of location. *Ann. Statist.* **14**, 68-87.

[6] Doob, J. L. (1948) Application to the theory of martingales. *Coll. Int. du C. N. R. S. Paris*, 22-28.

[7] Ferguson, T. S. (1973). A Bayesian analysis of some nonparametric problems. *Ann. Statist.* **1**, 209-230.

[8] Ferguson, T. S. (1974). Prior distributions on spaces of probability measures. *Ann. Statist.* **2**, 615-629.

[9] Ferguson, T. S., Phadia, E. G. and Tiwari, R. C. (1992). Bayesian nonparametric inference. *Current Issues in Statistical Inference: Essays in Honor of D. Basu* **17**, 127-150. (Institute of Mathematical Statistics Lecture Notes - Monograph Series).

[10] Ferguson, T. S., and Phadia, E. G. (1979). Bayesian nonparametric estimation based on censored data. *Ann. Statist.* **7**, 163-186

[11] Freedman, D. A. (1963). On the asymptotic behavior of Bayes estimates in the discrete case. *Ann. Math. Statist.* **34**, 1386-1403.

[12] Ghosal, S., Ghosh, J. K. and Ramamoorthi, R. V. (1996). Consistent semiparametric Bayesian inference about a location parameter. (To be Published).

[13] Ghosh, J. K., and Ramamoorthi, R. V. (1995). Consistency of Bayesian inference for survival analysis with or without censoring. *Analysis of censored data* **27**, 95-103. (Institute of Mathematical Statistics Lecture Notes - Monograph Series).

[14] Ghosh, J. K., and Ramamoorthi, R. V. (1996-97) Unpublished manuscript on Bayesian nonparametrics.

[15] Groeneboom, P. and Wellner, J. A. (1992) Information bounds and nonparametric maximum likelihood estimation. *Birkhauser Verlag, Basel.*

[16] Hannum, R. and Hollander, M. (1983). Robustness of Ferguson's Bayes estimator of a distribution function. *Ann. Statist.* **11**, 632-639.

[17] Hjort, N. L. (1990). Nonparametric Bayes estimators based on beta processes in models for life history data. *Ann. Statist.* **18**, 1259-1294.

[18] Lavine, M. (1992). Some aspects of Polya tree distributions for statistical modelling. *Ann. Statist.* **20**, 1222-1235.

[19] Lavine, M. (1994). More aspects of Polya tree distributions for statistical modelling. *Ann. Statist.* **22**, 1161-1176.

[20] Mauldin, R. D., Sudderth, W. D. and Williams, S C. (1992). Polya trees and random distributions. *Ann. Statist.* **20**, 1203-1221.

[21] Parthasarathy, K., R. (1967). Probability measures on metric spaces. *Academic Press, New York.*

[22] Peterson, A. V. (1977). Expressing the Kaplan-Meier estimator as a function of empirical subsurvival functions. *J. Amer. Statist. Assoc.* **72**, 854-858.

[23] Schervish, (1995). Theory of Statistics. *Springer-Verlag, New York*

[24] Schwartz, L. (1965) On Bayes' procedures. *Z. Wahrsch. verw. Gebiete* **4**, 10-26.

[25] Sethuraman, J. (1994). A constructive definition of Dirichlet priors. *Statistica Sinica* **4**, 639-650.

[26] Sethuraman, J. and Tiwari, R. C. (1982). Convergence of Dirichlet measures and the interpretation of their parameter. *Statistical Decision Theory and Related Topics III.* **2**, 305-315. (Edited by Shanti S. Gupta and James O. Berger; Published by Academic Press).

[27] Susarla, V. and Van Ryzin, J. (1976). Nonparametric Bayesian estimation of survival curves from incomplete observations, *J. Amer. Statist. Assoc.* **71**, 897-902.

[28] Stute, W. and Wang, J.-L. (1993). The strong law under random censorship. *Ann. Statist.* **21**, 1591-1607.

[29] Tsai, W. Y. (1986). Estmation of survival curves from dependent censorship models via a generalized self-consistency property with nonparametric Bayesian approach. *Ann. Statist.* **14**, 238-249.

[30] Turnbull, B. W. (1976). The empirical distribution function from arbitrarily grouped, censored, and truncated data. *J. Royal Statist. Soc. Ser. B.* **38**, 290-295.

[31] Wang, Z. (1993). Estimation in interval censorship models. *Ph. D. Dissertation, Michigan State University.*

[32] Wang, Z., and Gardiner, J. C. (1996). A class of estimators of the survival function from interval censored data. *Ann. Statist.* **24**, 647-658.

[33] Wang, Z., Gardiner, J. C., and Ramamoorthi, R. V. (1994). Identifiability in interval censorship models. *Stat. and Prob. Letters* **21**, 215-221.