NONPARAMETRIC BAYES INFERENCE FOR CONCAVE DISTRIBUTION FUNCTIONS*

MARTIN B. HANSEN† AND
STEFFEN L. LAURITZEN, *Aalborg University*

Abstract

Bayesian inference for concave distribution functions is investigated. This is made by transforming a mixture of Dirichlet processes on the space of distribution functions to the space of concave distribution functions. We give a method for sampling from the posterior distribution using a Pólya urn scheme in combination with a Markov chain Monte Carlo algorithm. The methods are extended to estimation of concave distribution functions for incompletely observed data.

DIRICHLET PROCESS; MARKOV CHAIN MONTE CARLO; CONCAVE DISTRIBUTION FUNCTIONS; DECREASING DENSITY FUNCTION; ORDER RESTRICTED INFERENCE; MULTIPLICATIVE CENSORING; PÓLYA URN SCHEME

AMS 1991 SUBJECT CLASSIFICATION: PRIMARY 62G05
SECONDARY 65U05, 60J05

1 Introduction

In many situations one faces data where there is reason to believe that they arise from a distribution with a decreasing density, or equivalently with a concave distribution function.

An example occurs in renewal processes where the residual lifetime has a concave distribution function, see e.g. Laslett (1982) and Vardi (1989). Another case arises in in spatial statistics, see e.g. Hansen *et al.* (1996), where the so called linear contact distribution (Stoyan *et al.* 1995) is fundamental and has a decreasing density.

Let $X_1, \ldots, X_n$ be a random sample from an unknown (cumulative) distribution function (cdf) F on the positive halfline $[0, \infty)$. When no plausible assumption

*Version of 13 December 2000.

†Address for correspondence: Department of Mathematical Sciences, Aalborg University, Fredrik Bajers Vej 7G, 9220 Aalborg Ø, Denmark.

about the functional form of F is available the empirical distribution function (edf) $F_n(x) = n^{-1}\sum_{i=1}^{n} 1_{(-\infty,x]}(X_i)$ is a natural estimator to use. However, exploiting further information it is possible to find better estimators. Let for instance $X_1, \ldots, X_n$ be a random sample from a concave cdf, F. Grenander (1956) proved that the maximum likelihood estimator of F is the concave majorant of the edf. Various properties of this estimator have been studied in Groeneboom (1985) and Groeneboom and Lopuhaä (1993).

We present a Bayesian framework with prior distributions concentrated on the space of probability measures on the positive real line with concave distribution functions. This involves constructing a probability distribution on the space of cdfs. Ferguson (1973, 1974) and Antoniak (1974) suggested respectively Dirichlet process priors and mixtures of them.

The present paper uses a mixture of Dirichlet processes on the space of probability distributions on $\mathbf{R}_+$. As in Brunner and Lo (1989) and Brunner (1992) we then exploit the unique correspondence between probability distributions on $\mathbf{R}_+$ and concave cumulative distribution functions on $\mathbf{R}_+$ through their representation as mixtures of uniform distributions. Computations are performed using Markov chain Monte Carlo methods and the Pólya urn scheme representation of the Dirichlet distribution given by Blackwell and MacQueen (1973). Our approach resembles that of Brunner (1995) although we use mixtures of Dirichlet process priors, an alternative Monte-Carlo algorithm and a representation which enables straightforward extensions to the case of observations that are censored or incompletely observed as well as easy integration into more complex statistical models.

Both parametric and nonparametric estimators under various censoring models have been studied extensively in the literature. The most common case of censoring is right censoring. For this case the nonparametric maximum likelihood estimator is the well known Kaplan–Meier estimator (Kaplan and Meier 1958). Turnbull (1974, 1976) gives a way of obtaining a nonparametric maximum likelihood estimator of a distribution function F for the general case of interval censoring. Susarla and van Ryzin (1976) and Hjort (1990) and later Damien et al. (1996) and Laud et al. (1996) used a Bayesian approach presenting nonparametric Bayes estimators based on right-censored data. In Muliere and Walker (1997) the idea of placing a Pólya tree prior distribution on the space of probability measures, introduced by Ferguson (1974), is extended to right censored data. The Bayesian approach was also explored by Doss (1994) who used Gibbs sampling to compute with the posterior distribution. For a recent review of Bayesian nonparametric inference see Walker et al. (1999) and the accompanying discussion and for reviews of Bayesian nonparametrics in survival analysis cf. Florens et al. (1999) and Rolin (1998). The present paper uses a flexible approach which easily exploits concavity and still takes censoring into account.

In Section 2 we describe prior distributions on the space of concave distribution functions. In Section 3 we describe Bayesian inference and develop a procedure for Gibbs sampling from the posterior distribution via an urn scheme. In Section 4 the techniques are illustrated with data concerning failure of airconditioning equipment. The methods are, in Section 5, extended to incompletely observed data, and the extension is illustrated with an analysis of data that are right-censored. Finally, we discuss consistency problems and mention some possible issues for future research.

2 Priors for concave distribution functions

2.1 Dirichlet processes and Bayesian updating

As our priors are based upon Dirichlet processes and mixtures of these, we briefly summarize some of their basic properties.

Let μ be a finite measure on $(\mathbf{R}_+, \mathcal{B}(\mathbf{R}_+))$ where $\mathbf{R}_+ = [0, \infty)$ and $\mathcal{B}(\mathbf{R}_+)$ is the Borel σ-field. A stochastic process P indexed by elements B of $\mathcal{B}(\mathbf{R}_+)$, is said to be a *Dirichlet process* on $(\mathbf{R}_+, \mathcal{B}(\mathbf{R}_+))$ with parameter μ if for any measurable partition $(B_1, \ldots, B_k)$ of $\mathbf{R}_+$ the random vector $(P(B_1), \ldots, P(B_k))$ has a Dirichlet distribution with parameter $(\mu(B_1), \ldots, \mu(B_k))$. Hence P may be considered as a random probability measure on $(\mathbf{R}_+, \mathcal{B}(\mathbf{R}_+))$. Often we will write $P \sim \mathcal{D}(\mu)$ or $G \sim \mathcal{D}(\mu)$ if $G(\cdot) = \int_0^{\cdot} dP$ is the cumulative distribution function corresponding to P. Then $G(t), t \in \mathbf{R}_+$ are random variables and if we let $H = \mu/\mu(\mathbf{R}_+)$ we have

1. If $G \sim \mathcal{D}(\mu)$ then $\mathbf{E}\, G(t) = H(t)$
2. If $G \sim \mathcal{D}(\mu)$ and given G, $Y_1, \ldots, Y_n$ is a random sample from G, then

$$G \mid Y_1, \ldots, Y_n \sim \mathcal{D}\left(\mu + \sum_{i=1}^{n} \delta_{Y_i}\right)$$

Property 2. says that the posterior distribution of G given the data is again a Dirichlet process, but with an updated parameter measure. It follows from 1. and 2. that

$$\mathbf{E}\left[G(t) \mid Y_1, \ldots, Y_n\right] = \frac{\mu([0,t]) + \sum_{i=1}^{n} \delta_{Y_i}([0,t])}{\mu(\mathbf{R}_+) + n} \tag{1}$$

so $\mu(\mathbf{R}_+)$ represents our relative belief in the prior or the *prior weight*. Property 1. indicates that H specifies the average behaviour of G, see Ferguson (1973) for these and other properties.

In many situations it is more appropriate to use a mixture of Dirichlet processes, see Antoniak (1974). Then we instead choose a parametric family $H_\theta, \theta \in \Theta \subset \mathbf{R}^k$

of probability measures, constants α_θ, a prior mixing measure ν, and assume that G has the following distribution

$$G \sim \int \mathcal{D}(\alpha_\theta H_\theta)\nu(d\theta),$$

Generally we assume for simplicity that the prior weights do not depend on θ, i.e. $\alpha_\theta = \alpha_0$, so we have

$$G \sim \int \mathcal{D}(\alpha_0 H_\theta)\nu(d\theta), \tag{2}$$

The following reformulation of Lemma 1 in Antoniak (1974) (see also Doss (1994), Theorem 1) gives an explicit expression for the posterior distribution after observing a sample $Y_1, \ldots, Y_n$.

Lemma 1 *Assume that for each $\theta \in \Theta$, H_θ is absolutely continuous, with a density h_θ that is continuous on* $\mathbf{R}$. *If the prior on G is given by (2), then the posterior distribution of θ given $Y = (Y_1, \ldots, Y_n)$ is ν_Y, where*

$$\nu_Y(d\theta) \propto \prod^{*} h_\theta(Y_i)\nu(d\theta). \tag{3}$$

Here "$*$" indicates that the product is extending over distinct values of Y_i only. The conditional distribution of G given θ and Y is

$$G \mid \theta, Y \sim \mathcal{D}\left(\alpha_0 H_\theta + \sum_{i=1}^{n} \delta_{Y_i}\right).$$

2.2 Representing a concave distribution function

The paper relies on the existence of a simple mapping from the space of distribution functions to the space of concave distribution functions. For this we exploit a (well-known) representation theorem for concave distribution functions which can be traced back to A. I. Khincin, see e.g. Feller (1971, p. 158). This can be expressed in the following two alternative forms:

Theorem 1 *Let F be a distribution function on $[0, \infty)$. F is concave if and only if there exists a distribution function G on $[0, \infty)$ such that F admits the representation*

$$F(x) = \int_{[0,\infty)} F_y(x)\, G(dy), \qquad x \in \mathbf{R} \tag{4}$$

where F_y for $y \neq 0$ is the distribution function corresponding to the uniform distribution on $(0, y)$, i.e.

$$F_y(x) = \frac{1}{y}\min\{x, y\}$$

and F_0 is degenerate at 0.

In fact if G is given by (4) then G is uniquely determined by F through the relation

$$G(x) = F(x) - xf(x),$$

where $f = D^+F$ is the derivative of F from the right. This result can alternatively be formulated as:

Theorem 2 *Let F be a distribution function on $[0, \infty)$. Then F is concave if and only if F is the distribution of the product $X = YU$ of two independent random variables with U distributed uniformly on $(0, 1)$ and Y having distribution G for some distribution function G on $[0, \infty)$.*

2.3 Mixture model

The purpose is to make inference about the unknown concave distribution function F given a random sample $X = (X_1, \ldots, X_n)$ from F:

$$X_i \sim F, \qquad i = 1, \ldots, n. \tag{5}$$

In our Bayesian model specification we induce a prior distribution on F by the relation

$$F = \int_0^\infty F_y G(dy) \tag{6}$$

where G is a random cumulative distribution function distributed as a mixture of Dirichlet processes

$$G \sim \int \mathcal{D}(\alpha_0 H_\theta)\nu(d\theta). \tag{7}$$

By Theorem 1, F will be a random concave cdf, and we obtain in this way a prior distribution on the subspace of concave cdfs.

2.4 Multiplicative censoring model

The model can alternatively be expressed in terms of multiplicative censoring (Vardi 1989). Let $Y = (Y_1, \ldots, Y_n)$ be a random sample G

$$Y_i \sim G, \qquad i = 1, \ldots, n \tag{8}$$

where G is random and distributed as

$$G \sim \int \mathcal{D}(\alpha_0 H_\theta)\nu(d\theta). \tag{9}$$

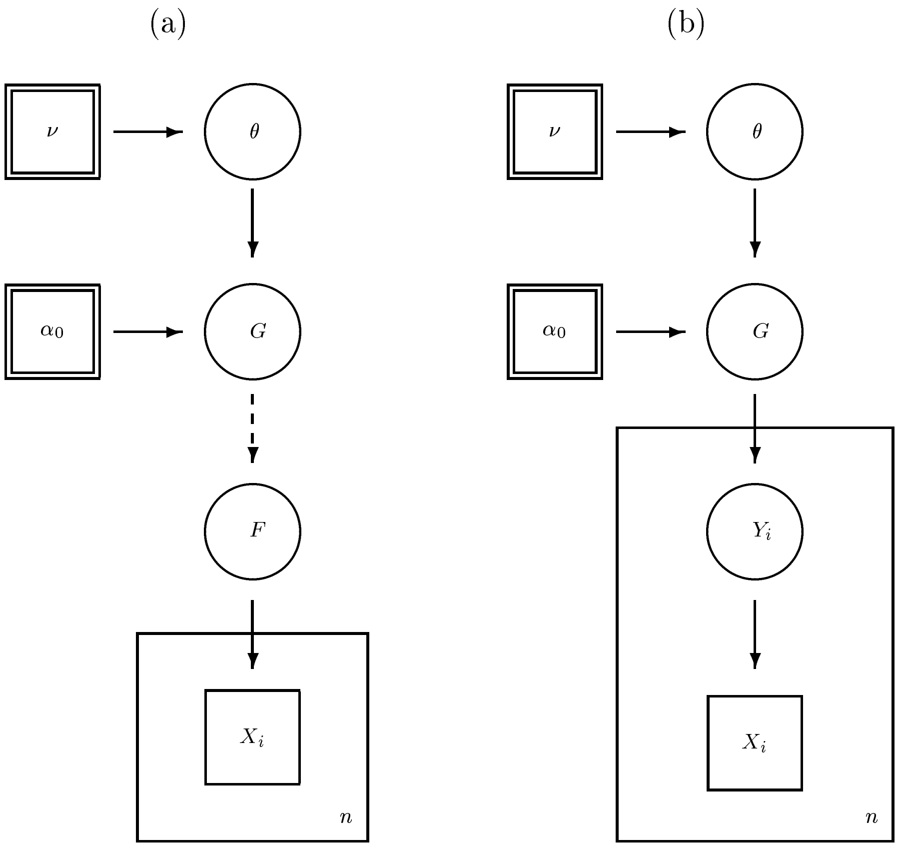

Figure 1: Graphical representations of the models suggested for concave distribution functions: (a) by a mixture; (b) by multiplicative censoring.

Now, let $U = (U_1, \ldots, U_n)$ be independent and uniformly distributed over the unit interval $(0, 1)$ and let

$$X_i = U_i Y_i, \qquad i = 1, \ldots, n. \tag{10}$$

Then by Theorem 1 and Theorem 2, the vector $X = (X_1, \ldots, X_n)$ will be a random sample from F, specified in (6). The task is here to make inference about the distribution function G given the incomplete data $X = (X_1, \ldots, X_n)$.

2.5 Graphical representation

Both models (5)–(7) and (8)–(10) specify a joint distribution of model parameters and data through a few conditional distributions. This assumption of independence among unspecified submodels is known as the directed Markov assumption (Lauritzen *et al.* 1990). Such directed Markov models are conveniently illustrated via a directed acyclic graph (DAG) as shown in Figure 1. Each quantity in the model appears as a node in the graph and directed edges correspond to direct dependencies. Round nodes denote unobservable quantities in prior distributions, square nodes with a single border denote observed random variables and square nodes with a double border denote fixed quantities. The graph is *directed* because each edge between nodes is directed and *acyclic* because by following the arrows one cannot return to a node after leaving it. The formalism of graphical modelling enables direct identification of consequent conditional independence relations and emphasises modularity in modellling and computation. This in turn paves the way for flexibility in extending the model and incorporating it into more complex systems, as exploited later in Section 5.

General introductions to graphical modelling can e.g. be found in monographs by Whittaker (1990), Lauritzen (1996), or Edwards (2000). DAG models provide a rich class of models which are particularly amenable to analysis by Gibbs sampling, see Gilks *et al.* (1994).

3 Sampling from the posterior distribution

3.1 Gibbs sampling by an urn scheme

A popular way of handling analytically intractable posterior distributions is to construct a Markov chain whose limiting distribution is the required posterior distribution, i.e. in our case $\mathcal{L}(F \mid X)$ or $\mathcal{L}(G \mid X)$. Methods of this type are denoted Markov chain Monte Carlo methods (MCMC). There is a wealth of related but different MCMC algorithms and in this paper we, similar to Doss (1994), use Gibbs sampling, introduced in this form by Geman and Geman (1984) and in this context by Gelfand and Smith (1990). For a general description of such methods, see for example Green (2001).

We are interested in the distribution of the unknown variables in Figure 1 (b) given the observation X denoted by $\mathcal{L}(\theta, G, Y \mid X)$. A Gibbs sampler can then proceed as follows: Pick an arbitrary starting value $Y^{(0)}$ satisfying the constraints of the data, i.e. $Y_i^{(0)} \geq X_i$ for $i = 1, \ldots, n$ and make successive draws from the set of full conditional distributions in the following way:

For $k = 1, 2, \ldots$

(a) Draw $(\theta^{(k)}, G^{(k)}) \sim \mathcal{L}(\theta, G \mid Y^{(k-1)}, X)$

(b) Draw $Y^{(k)} \sim \mathcal{L}(Y \mid \theta^{(k)}, G^{(k)}, X)$

Then $(\theta^{(k)}, G^{(k)}, Y^{(k)})$ forms a Markov chain and $\mathcal{L}(\theta, G, Y \mid X)$ is the stationary distribution of this chain. If one can establish that the chain converges to its stationary distribution then for large k, $(\theta^{(k)}, G^{(k)}, Y^{(k)})$ has a distribution which is approximately equal to $\mathcal{L}(\theta, G, Y \mid X)$.

To start the algorithm we only need a permissible $Y^{(0)}$. We generate θ from ν and subsequently $Y_i^{(0)}$, $i = 1, \ldots, n$ as independent observations from H_θ restricted to the set $[X_i, \infty)$, $i = 1, \ldots, n$. This is done by rejection sampling (see Ripley 1987, pp. 60)

Step (a) above is not carried out quite as it stands. Instead we use Lemma 1, sample $\theta^{(k)}$ from its posterior distribution $\mathcal{L}(\theta \mid Y^{(k-1)}, X) = \mathcal{L}(\theta \mid Y^{(k-1)})$, and just identify $\mathcal{L}(G \mid \theta^{(k)}, Y^{(k-1)}, X)$ i.e. we perform step (a) in the following way

Draw $\theta^{(k)} \sim \nu_{Y^{(k-1)}}$

Draw $G^{(k)} \sim \mathcal{D}\left(\alpha_0 H_{\theta^{(k)}} + \sum_{i=1}^n \delta_{Y_i^{(k-1)}}\right)$,

where the second line is only performed conceptually.

Step (b) is performed in two steps by first generating observations Z from the distribution of $G^{(k)}$ above without actually expressing $G^{(k)}$ explicitly and then using simple rejection to generate observations from $\mathcal{L}(Y \mid G^{(k)}, X)$. Doss (1994) used a constructive representation of the Dirichlet prior due to Sethuraman (1994), but we chose instead to use the urn scheme introduced by Blackwell and MacQueen (1973) to simulate from a Dirichlet process.

Let μ be a finite measure on $(\mathbf{R}_+, \mathcal{B}(\mathbf{R}_+)))$ and $\{Z_l\}_{l=1}^\infty$ be a sequence of random variables in $\mathbf{R}_+$, drawn according to the following scheme:

$$Z_1 \sim \mu/\mu(\mathbf{R}_+)$$

and

$$Z_{l+1} \mid Z_1, \ldots, Z_l \sim \left(\mu + \sum_{i=1}^{l} \delta_{Z_i}\right) / \mu(\mathbf{R}_+), \qquad l = 2, 3, \ldots$$

Blackwell and MacQueen (1973) say that such a sequence $\{Z_l\}_{l=1}^\infty$ is a *Pólya sequence* on $\mathbf{R}_+$ with parameter μ. They show that there exists a random distribution function G such that $G \sim \mathcal{D}(\mu)$ and, conditionally on G, $Z_1, Z_2, \ldots$ are independent with distribution G. Thus letting

$$\mu = \alpha_0 H_{\theta^{(k)}} + \sum_{i=1}^{n} \delta_{Y_i^{(k-1)}}$$

in the Pólya sequence above gives the desired stream of Z-values. Then $Y^{(k)}$ is generated from the Z-stream by rejection sampling.

Expressed differently, G is never sampled, but the procedure alternates between drawing from $\mathcal{L}(\theta \mid Y^{(k-1)})$ and $\mathcal{L}(Y \mid X, \theta^{(k-1)})$.

The algorithm described above is computationally straightforward. Convergence properties of the algorithm can be established as in Doss (1994). The cluster structure of ties among the unobserved variables can be exploited to make the algorithm more efficient along the same lines as in MacEachern (1994) who improves upon the Gibbs sampling algorithm of Escobar (1994). However, we found the present approach to be fully sufficient for the problem at hand.

3.2 Bayesian inference

By repeating the algorithm a large number K of times, one obtains $(\theta^{(k)}, G^{(k)}, Y^{(k)})$, $k = 1, \ldots, K$.

Next we wish to represent the posterior distribution of F (or G) given the data X. We then approximate the implicitly given distribution functions $G^{(k)}$ by the edf for the random samples $Z^{(k)}$. Noticing that if

$$\widetilde{G} \sim \mathcal{L}(G \mid X) \tag{11}$$

then

$$\widetilde{F} = \int F_y \widetilde{G}(dy) \sim \mathcal{L}(F \mid X), \tag{12}$$

and we are thus also able to generate a random distribution function from $\mathcal{L}(F \mid X)$. Next we construct an estimate of $\mathbf{E}\,[G \mid X]$ by

$$\widehat{G}_K(t) = \frac{1}{KL} \sum_{k=1}^{K} \sum_{l=1}^{L} 1_{[0,t]}(Z_l^{(k)}) \tag{13}$$

and a natural estimate of $\mathbf{E}\,[F \mid X]$ is then obtained from $\widehat{G}_K$ as

$$\widehat{F}_K(t) = \int F_y(t) d\widehat{G}_K(y) = \frac{1}{KL} \sum_{k=1}^{K} \sum_{l=1}^{L} F_{Z_l^{(k)}}(t) \tag{14}$$

leading to an estimate of the decreasing density of $\mathbf{E}\,[F \mid X]$ by

$$\hat{f}_K(t) = \frac{1}{KL} \sum_{k=1}^{K} \sum_{l=1}^{L} \frac{1}{Z_l^{(k)}} 1_{\{Z_l^{(k)} > t\}}.$$

The choice of L and K is mainly determined on the type of inference. If the sampling variation of $G \mid X$ is of interest, one should make repeated experiments where L is large and if the posterior mean of $G \mid X$ is of interest, L can in principle be chosen as small as one likes, as long as K is sufficiently large.

The Gibbs sampling procedure may provide an alternative estimate of $\mathbf{E}[G|X]$ with a potentially lower variance than (13), namely by noticing that

$$\mathbf{E}[G \mid \theta, Y, X] \sim \frac{\alpha_0 H_\theta + \sum_{i=1}^n \delta_{Y_i}}{\alpha_0 + n}.$$

whereby a Rao-Blackwellized estimate (Gelfand and Smith 1990) of $\mathbf{E}[G|X]$ can be obtained by

$$\widehat{G}(t) = \frac{1}{K} \sum_{k=1}^{K} \frac{\alpha_0 H_{\theta^{(k)}}([0,t]) + \sum_{i=1}^{K} \delta_{Y_i^{(k)}}([0,t])}{\alpha_0 + n}.$$

4 Failure of airconditioning equipment

We now discuss an example to illustrate the techniques. The stepfunctions in Figure 2 give the edfs of the intervals in operating hours between successive failures of air-conditioning equipment in 13 Boeing 720 aircrafts. These data set were presented in Proschan (1963) and further analyzed by Cox and Lewis (1966).

This is an example where we have several series rather than a single one. It is of interest to find the distribution of the failure-time, for instance when one considers preventive maintenance.

Cox and Lewis (1966) argued that the combined data follow a distribution which is nearly exponential. Therefore it seems sensible to specify a prior distribution for F centered around an exponential distribution with density

$$f_\theta(x) = \theta e^{-\theta x}.$$

Assuming a common exponential distribution yields a maximum likelihood estimate of the reciprocal mean parameter $\theta_0 = 0.0107$. However, it seems more reasonable to assume individual survival functions for each aircraft, see Cox and Lewis (1966).

We therefore analyzed the data using a prior on F given by model (5)–(7), where H_θ has density

$$h_\theta(y) = \theta y e^{-\theta x} \tag{15}$$

so that

$$\int_y F_y(x) h_\theta(y)\, dy = \theta e^{-\theta x}.$$

We chose to consider two cases $\alpha_0 = 1$ and $\alpha_0 = 100$, representing essentially a situation of a diffuse prior and a situation where we are rather confident about our

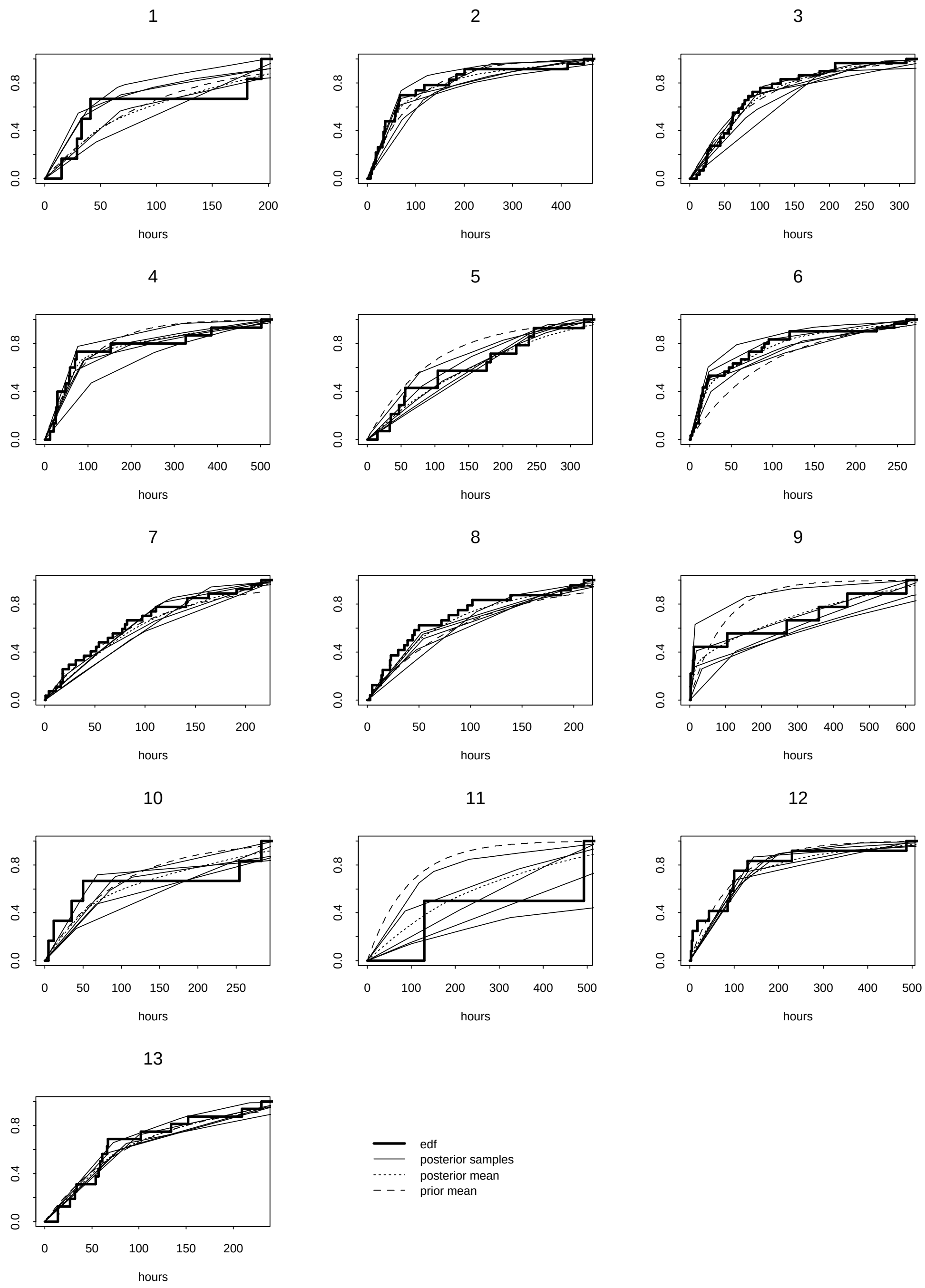

Figure 2: Edf of failure times of airconditioning equipment of 13 aircrafts (thick line). The thinner lines show samples from the posterior distributions. The dotted line represents the posterior mean and the dashed line is the prior mean.

prior. Only results for $\alpha_0 = 1$ are reported here. Furthermore we took ν to be a gamma distribution $\Gamma(a, b)$ with $a = 1000$ and $b = 1000\theta_0$, i.e.

$$\nu(d\theta) \propto \theta^{a-1} e^{-b\theta}\, d\theta.$$

Hence the prior for F is concentrated around the exponential distribution corresponding to the maximum likelihood estimate.

The reason for taking a gamma prior on θ is that it is a conjugate family for the exponential distribution. If ν is $\Gamma(a, b)$, then Lemma 1 in combination with (15) gives that the posterior ν_Y is $\Gamma(a + 2n^*, b + \sum^* y_i)$, where n^* is the number of distinct observations in Y and $\sum^* y_i$ is the sum of distinct y_i's.

Trace plots of empirical means of Y indicate almost immediate convergence to equilibrium of the Gibbs sampling algorithm introduced in Section 4.

Figure 2 shows estimates of the distribution function F for each aircraft. These were obtained by running the algorithm for a burn-in of 100 iterations and then generating a random sample $Z_1^{(100)}, \ldots, Z_{1000}^{(100)}$ from $G^{(100)}$. This was repeated independently $K = 1000$ times. Samples from the posterior distribution (obtained by the estimate $\widehat{F}_K(t)$ for the first 5 repetitions, see (14)), are shown by fulldrawn thinner lines. The dotted line is the posterior mean (obtained by averaging all 1000 samples of $\widehat{F}_K(t)$), whereas the dashed line is the prior distribution.

5 Extensions to further censoring

The graphical model formulation as displayed in Figure 1 in combination with Gibbs sampling provides a powerful tool for Bayesian inference in structured model formulations. Below we capitalize on this fact by introducing a straightforward extension to situations with further censoring.

Assume that our model setup is as specified by (5)–(7) or alternatively by (8)–(10). Moreover we suppose that the X's are not directly observed; instead we observe

$$(B_i, \delta_i) = \begin{cases} (\{X_i\}, 1), & X_i \notin A_i \\ (A_i, 0) & X_i \in A_i \end{cases}, \; i = 1, \ldots, n, \tag{16}$$

where the A_i's are subsets of $\mathbf{R}_+$. In case of right-censored data $A_i = (c_i, \infty)$ and X_i is censored to the right by c_i. We wish to obtain the posterior distribution of F given the incomplete data. Graphical representations of the suggested model are shown in Figure 3.

By inspecting the graphical representation in Figure 3 (b) and reading off the conditional independences we can for $(B, \delta) = ((B_1, \delta_1), \ldots, (B_n, \delta_n))$ formulate the following general Gibbs sampling scheme

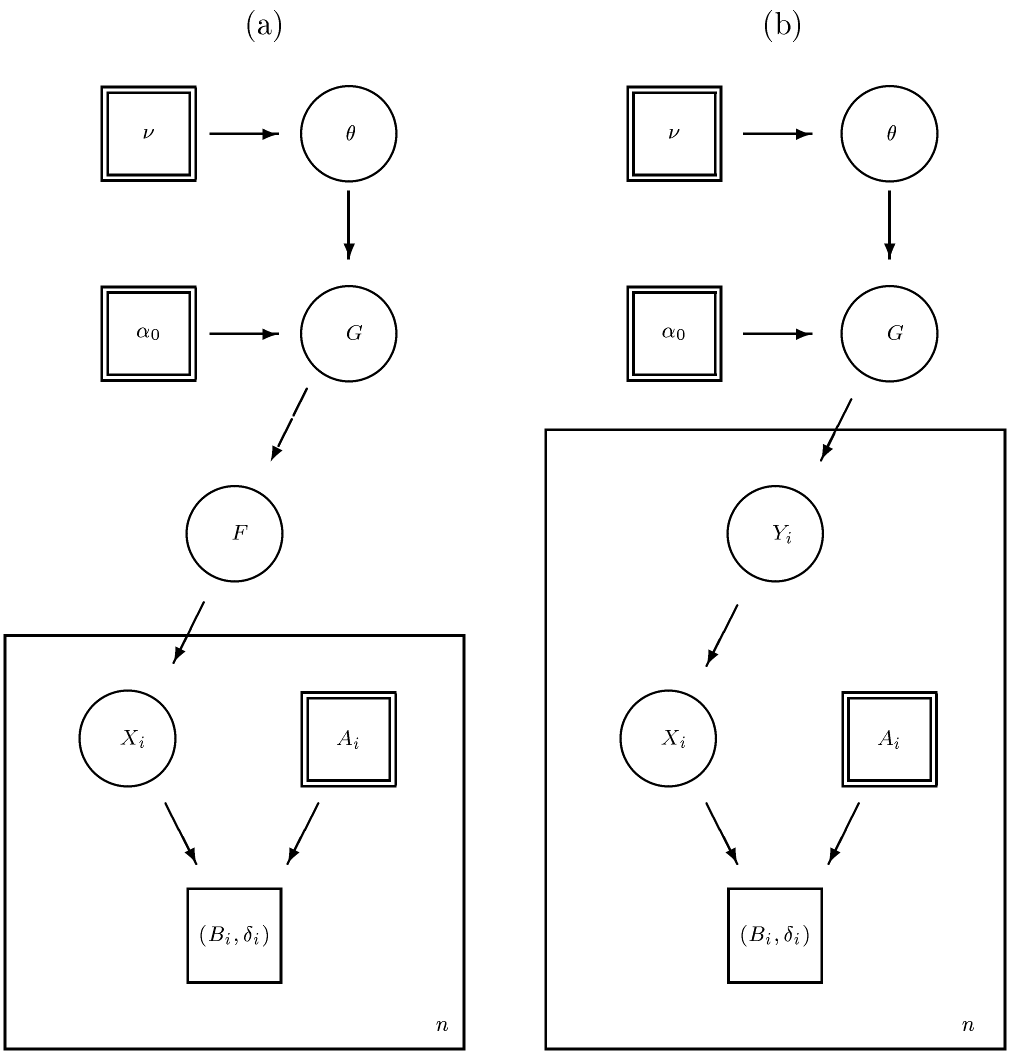

Figure 3: Graphical representations of the models suggested for Bayesian inference in the case of incompletely observed data from concave distribution functions represented (a) by mixtures; (b) by multiplicative censoring.

For $k = 1, 2, \ldots$

(a) Draw $(\theta^{(k)}, G^{(k)}) \sim \mathcal{L}(\theta, G \mid Y^{(k-1)}, X^{(k-1)}, B, \delta) = \mathcal{L}(\theta, G \mid Y^{(k-1)})$

(b) Draw $Y^{(k)} \sim \mathcal{L}(Y \mid \theta^{(k)}, G^{(k)}, X^{(k-1)}, B, \delta) = \mathcal{L}(Y \mid G^{(k)}, X^{(k-1)})$

(c) Draw $X^{(k)} \sim \mathcal{L}(X \mid \theta^{(k)}, Y^{(k)}, B, \delta) = \mathcal{L}(X \mid Y^{(k)}, B)$,

where as before, G is only generated implicitly.

To illustrate a situation with incomplete data, we have chosen to use the data given in Kaplan and Meier (1958).

We analyzed these data using a prior on F given by the model (5)–(7) and (16). So in our notation $(B_1, \delta_1) = (\{0.8\}, 1)$, $(B_2, \delta_2) = (\{3.1\}, 1)$, $(B_3, \delta_3) = (\{5.4\}, 1)$,

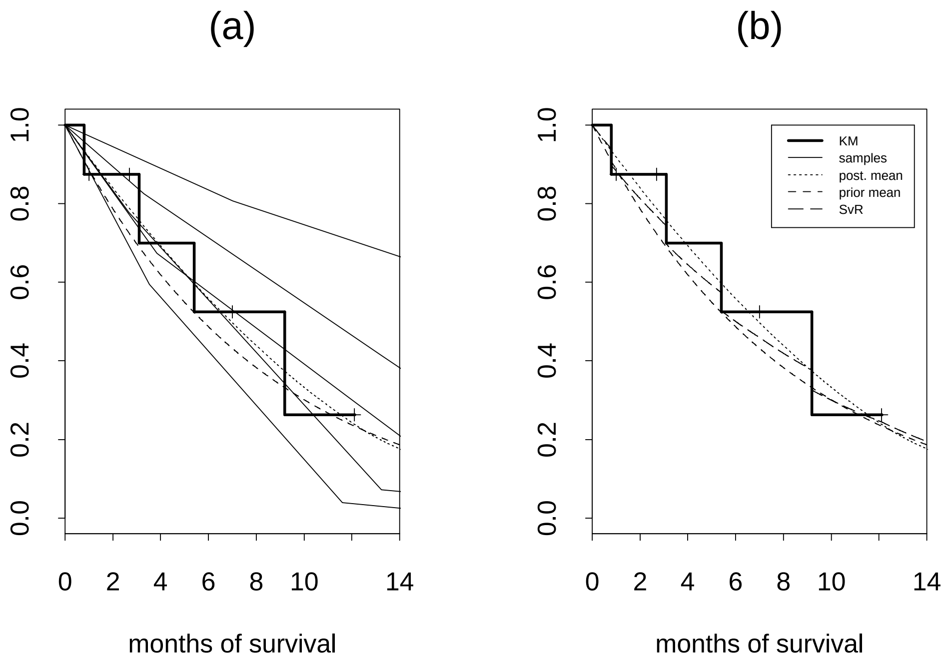

Figure 4: (a) The thinner lines are samples from the posterior distribution. The dotted line represent the posterior mean and the short-dashed line the prior mean (b) The long-dashed line is the Susarla and van Ryzin (1976) estimator. (In both plots crosses represent censoring times)

$(B_4, \delta) = (\{9.2\}, 1)$, $(B_5, \delta_5) = (\{(1.0, \infty)\}, 0)$, $(B_6, \delta_6) = (\{(2.7, \infty)\}, 0)$, $(B_7, \delta_7) = (\{(7.0, \infty)\}, 0)$, $(B_8, \delta_8) = (\{(12.1, \infty)\}, 0)$. Furthermore we chose H_θ using the same rationale as in the previous example to be $\Gamma(2, 0.12)$. As Susarla and van Ryzin (1976) we chose $\alpha_0 = 16$, which represents a rather high degree of smoothing, and took ν to be $\Gamma(100, 12)$. Figure 4 (a) displays corresponding estimates. These were obtained by running the algorithm for a burn in of 100 iterations and then generating a random sample $Z_1^{(100)}, \ldots, Z_{1000}^{(100)}$ from $G^{(100)}$. This was repeated independently $K = 1000$ times. Samples from the posterior distribution (obtained by the estimate $1 - \widehat{F}_K(t)$ for the first 5 repetitions, see (14)), is shown as the fulldrawn thinner lines. The dotted lines are the posterior mean (obtained by averaging all 1000 samples of $1 - \widehat{F}_K(t)$), whereas the thick line represents the KM-estimator based on the data and the short-dashed line is the posterior mean.

In Figure 4(b) the estimates from Figure 4 (a) are supplemented with the Susarla and van Ryzin (1976) estimator with weight $\beta = 16$ (long-dashed line).

6 Discussion

6.1 Consistency

In complex problems such as those dealt with in the present paper, one must often choose prior distributions for the unknown parameters in a pragmatic and rather *ad hoc* fashion. Consequently it is of interest to know whether a sufficiently large sample will eventually overrule any reasonable choice of prior distribution. In our context, this amounts to the following question. Let ν, H_θ and α_0 represent a given choice of prior distributions for the unknown (concave) distribution function, and let the $X_1, \ldots, X_n$ denote a sample from a concave distribution function F^*. Is it true that the posterior distribution for F converges (F^*-almost surely in the weak-star topology) to the measure degenerate at F^*? Or, slightly weaker, does the posterior expectation of F converge to F^* under similar conditions?

In the uncensored case, without the concavity assumption, the answer to these questions is positive which follows from the explicit representation of the posterior distribution, see for example equation (1).

In the censored case, with or without the concavity assumption, the problem is difficult and the answer clearly depends on the censoring mechanism. If certain parts of the state space are never observed, the estimate cannot be consistent in the above sense. And in general the situation could be bad even in the uncensored case with the concavity assumption. The risk is that the posterior expectation represents an oversmoothing. Brunner and Lo (1989) use Doob's theorem (Doob 1949) on the present problem and state that the Bayes estimates are consistent for almost all parameter values. The problem using this result is that "almost all" refers to the prior distribution which is concentrated on the (small) set of discrete distributions, whereas we really want consistency w.r.t. true distributions F^* in the support of the prior distribution. See for example Diaconis and Freedman (1986) who show how unpleasant things could potentially be.

We have not been able to resolve this question in general, but we do believe that the Bayes estimate suggested in this paper will be consistent in cases with reasonable censoring patterns. This conjecture is supported by a small simulation experiment, the results of which are displayed in Figure 5. Observations X_i were simulated from F^* being uniform on the interval $(0,1)$, and the Bayes estimate of F calculated using a prior distribution similar to that used in Section 4. Thus H_θ is $\Gamma(2,\theta)$ ν is $\Gamma(1000, 1000\theta_0)$, where $\theta_0 = .75$. Different weights were attached to the prior by considering the two cases $\alpha_0 = 1$ and $\alpha_0 = 100$. In a sense the uniform distribution represents the most difficult case when the main risk is oversmoothing, as the uniform distributions are extreme points of the set of concave distribution functions.

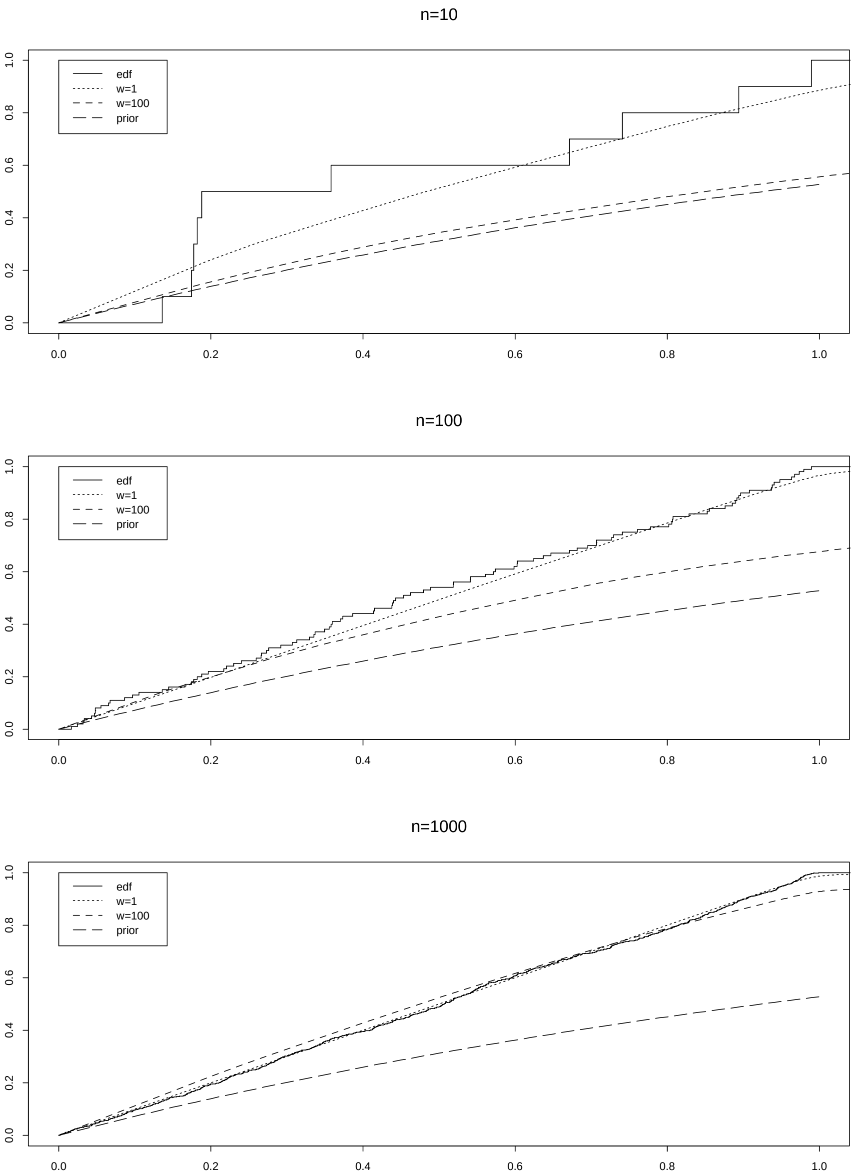

Figure 5: Bayes estimates of a concave distribution function based on 10, 100, and 1000 observations from a uniform distribution. The empirical distribution function and the prior mean are displayed for comparison.

The posterior expectation of F was calculated after 10, 100, and 1000 observations. It appears from the figure that this posterior expectation gets closer to F^* as the sample size increases.

6.2 Perspectives

The basic element of the method described is the fact that the class of distributions in the non-parametric model have a mixture representation in such a way that that the mixing measure itself can be used to parameterize the distributions. Thus the results in this paper are much in the same spirit as work of Lo (1984), Florens *et al.* (1992), Escobar and West (1995), Florens *et al.* (1999), and others.

There is a considerable potential for exploiting this in other cases with a similar structure. In particular statistical inverse problems springs to mind. Assume we have a random sample $Y_1, \ldots, Y_n$ from a distribution function F, which is related to an underlying distribution G via the following integral equation

$$F(x) = \int k(x, y) dG(y),$$

where $k(\cdot, \cdot)$ is an integral kernel. This is a familiar problem in statistics and arises in e.g. convolution and monotone density problems. If G is the distribution of interest this is an instance of a statistical inverse problem. One can attack this type of problem by assuming a nonparametric prior on the distribution G, and then perform the analysis under this assumption. This seems to lead to complicated posterior calculations as we are dealing with indirect observations from the prior distribution. Nevertheless, the problem can be resolved in many cases by MCMC-techniques, as done in the present paper.

Also it could be of interest to extend the methods to other prior distributions than Dirichlet-processes, such as for example Pólya-tree distributions.

Finally, we should emphasize the flexibility of the present approach for incorporating Bayesian nonparametrics in graphical models. Once modules for sampling $\mathcal{L}(Y \mid G)$ and $\mathcal{L}(G \mid \theta, Y)$ has been implemented, data can be analyzed by standard packages for Bayesian inference like BUGS (Gilks *et al.* 1994).

Acknowledgement

This work was (partially) supported by MaPhySto — Centre for Mathematical Physics and Stochastics, funded by a grant from the Danish National Research Foundation.

References

Antoniak, C. E. (1974). Mixtures of Dirichlet processes with applications to Bayesian nonparametric problems. *Annals of Statistics*, **2**, 1152–74.

Blackwell, D. and MacQueen, J. B. (1973). Ferguson distributions via Pólya urn schemes. *Annals of Statistics*, **1**, 353–5.

Brunner, L. J. (1992). Bayesian nonparametric methods for data from a unimodal density. *Statistics and Probability Letters*, **14**, 195–9.

Brunner, L. J. (1995). Using the Gibbs sampler to simulate from a Bayes estimate of a decreasing density. *Communications in Statistics. Theory and Methods*, **24**, 215–26.

Brunner, L. J. and Lo, A. Y. (1989). Bayes methods for a symmetric unimodal density and its mode. *Annals of Statistics*, **17**, 1550–66.

Cox, D. R. and Lewis, P. A. W. (1966). *The statistical analysis of series of events.* Methuen, New York.

Damien, D., Laud, P. W., and Smith, A. F. M. (1996). Implementation of Bayesian non-parametric inference based on beta processes. *Scandinavian Journal of Statistics*, **23**, 27–36.

Diaconis, P. and Freedman, D. (1986). On the consistency of Bayes estimates. *Annals of Statistics*, **14**, 1–26.

Doob, J. L. (1949). Application of the theory of martingales. *Colloques Internationaux du Centre National de la Recherche Scientifique, Paris*, **13**, 23–7.

Doss, H. (1994). Bayesian nonparametric estimation for incomplete data via successive substitution sampling. *Annals of Statistics*, **22**, 1763–86.

Edwards, D. (2000). *Introduction to graphical modelling*, (2nd edn). Springer-Verlag, New York.

Escobar, M. D. (1994). Estimating normal means with a dirichlet process prior. *Journal of the American Statistical Association*, **89**, 268–77.

Escobar, M. D. and West, M. (1995). Bayesian density estimation and inference using mixtures. *Journal of the American Statistical Association*, **90**, 577–88.

Feller, W. (1971). *An introduction to probability theory and its applications*, (2nd edn). John Wiley and Sons, New York.

Ferguson, T. S. (1973). A Bayesian analysis of some nonparametric problems. *Annals of Statistics*, **1**, 209–30.

Ferguson, T. S. (1974). Prior distributions on spaces of probability measures. *Annals of Statistics*, **2**, 615–29.

Florens, J.-P., Mouchart, M., and Rolin, J.-M. (1992). Bayesian analysis of mixtures: Some results on exact estimability and identification. In *Bayesian statistics 4*, (ed. J. M. Bernardo, J. O. Berger, A. P. Dawid, and A. F. M. Smith), pp. 127–45. Oxford University Press, Oxford.

Florens, J.-P., Mouchart, M., and Rolin, J.-M. (1999). Semi- and non-parametric Bayesian analysis of duration models with Dirichlet priors: A survey. *International Statistical Review*, **67**, 187–210.

Gelfand, A. E. and Smith, A. F. M. (1990). Sampling-based approaches to calculating marginal densities. *Journal of the American Statistical Association*, **85**, 398–409.

Geman, S. and Geman, D. (1984). Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **6**, 721–41.

Gilks, W. R., Thomas, A., and Spiegelhalter, D. J. (1994). A language and program for complex Bayesian modelling. *The Statistician*, **43**, 169–78.

Green, P. (2001). A primer on Markov chain Monte Carlo. In *Complex stochastic systems*, Monographs on statistics and applied probability, Vol. 97, (ed. O. E. Barndorff-Nielsen, D. R. Cox, and C. Klüppelberg), pp. 1–62. Chapman and Hall/CRC Press, London/Boca Raton.

Grenander, U. (1956). On the theory of mortality measurement, part II. *Skandinavisk Aktuarietidskrift*, **39**, 125–53.

Groeneboom, P. (1985). Estimating a monotone density. In *Proceedings of the Berkeley conference in honor of Jerzy Neyman and Jack Kiefer, vol. II*, (ed. L. L. Cam and R. Olshen), pp. 539–55. Wadsworth, Monterey.

Groeneboom, P. and Lopuhaä, H. (1993). Isotonic estimators of monotone densities and distribution functions. *Statistica Neerlandica*, **47**, 175–84.

Hansen, M. B., Gill, R. D., and Baddeley, A. (1996). Kaplan-Meier type estimators for linear contact distributions. *Scandinavian Journal of Statistics*, **23**, 129–55.

Hjort, N. L. (1990). Nonparametric Bayes estimators based on beta processes in models for life history data. *Annals of Statistics*, **18**, 1259–94.

Kaplan, E. L. and Meier, P. (1958). Nonparametric estimation from incomplete observations. *Journal of the American Statistical Association*, **53**, 457–81.

Laslett, G. M. (1982). The survival curve under monotone density constraints with applications to two-dimensional line segment processes. *Biometrika*, **69**, 153–60.

Laud, P. W., Smith, A. F. M., and Damien, P. (1996). Monte Carlo methods for approximating a posterior hazard rate process. *Statistics and Computing*, **6**, 77–84.

Lauritzen, S. L. (1996). *Graphical models*. Oxford University Press.

Lauritzen, S. L., Dawid, A. P., Larsen, B. N., and Leimer, H.-G. (1990). Independence properties of directed Markov fields. *Networks*, **20**, 491–505.

Lo, A. Y. (1984). On a class of Bayesian nonparametric estimates: I. Density estimation. *Annals of Statistics*, **12**, 351–7.

MacEachern, S. N. (1994). Estimating normal means with a conjugate style Dirichlet process prior. *Communications in Statistics. Simulation and Computation*, **23**, 727–41.

Muliere, P. and Walker, S. (1997). A Bayesian non-parametric approach to survival analysis using Polya trees. *Scandinavian Journal of Statistics*, **24**, 331–40.

Proschan, F. (1963). Theoretical explanation of observed decreasing failure rate. *Technometrics*, **5**, 375–83.

Ripley, B. D. (1987). *Stochastic simulation.* John Wiley and Sons, New York.

Rolin, J.-M. (1998). Bayesian survival analysis. In *Encyclopedia of biostatistics*, pp. 271–86. Wiley, New York.

Sethuraman, J. (1994). A constructive definition of Dirichlet priors. *Statistica Sinica*, **4**, 639–50.

Stoyan, D., Kendall, W. S., and Mecke, J. (1995). *Stochastic geometry and its applications*, (2nd edn). John Wiley and Sons, Chichester.

Susarla, V. and van Ryzin, J. (1976). Nonparametric Bayesian estimation of survival curves from incomplete observations. *Journal of the American Statistical Association*, **71**, 897–902.

Turnbull, B. W. (1974). Nonparametric estimation of a survivorship function with doubly censored data. *Journal of the American Statistical Association*, **69**, 169–73.

Turnbull, B. W. (1976). The empirical distribution function with arbitrarily grouped, censored, and truncated data. *Journal of the Royal Statistical Society, Series B*, **38**, 290–5.

Vardi, Y. (1989). Multiplicative censoring, renewal processes, deconvolution and decreasing density: Nonparametric estimation. *Biometrika*, **76**, 751–61.

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

Whittaker, J. (1990). *Graphical models in applied multivariate analysis.* John Wiley and Sons, Chichester.