NONPARAMETRIC BAYES INFERENCE FOR CONCAVE DISTRIBUTION FUNCTIONS

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

Abstract

A way of making Bayesian inference for concave distribution functions is introduced. This is done by uniquely transforming a mixture of Dirichlet processes on the space of distribution functions to the space of concave distribution functions. The approach also gives a way of making Bayesian analysis of multiplicatively censored data. We give a method for sampling from the posterior distribution by use of 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. Finally, consistency issues are touched upon.

DIRICHLET PROCESS; MARKOV CHAIN MONTE CARLO; CONCAVE DISTRIBUTION FUNCTIONS; MONOTONE 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. In our Bayesian framework we use prior distributions concentrated on the space of probability measures on the positive real line with concave distribution functions.

Our motivation for the concavity assumption on distribution functions arose from potential applications in statistical inference for renewal processes and spatial statistics.

*Address for correspondence: Department of Mathematics, Aalborg University, Fredrik Bajers Vej 7E, 9220 Aalborg Ø, Denmark.

Think of k independent stationary renewal processes with lifetime distribution G. Assume that the processes are inspected at an arbitrary time and then each process is surveyed for a fixed period of time T or until the next replacement. The distribution of the residual lifetimes are denoted F. It is possible to prove that the residual lifetimes are multiplicatively censored lifetimes whereby the distribution of F is concave. The application of concavity assumptions on the distribution function and multiplicative censoring with respect to renewal processes was noted already by Laslett (1982) and Vardi (1989).

A similar problem arises in spatial statistics, see e.g. Hansen et al. (1996) and Hansen (1996). Imagine standing in a forest and considering the range of vision in a particular direction. The distribution of this range, usually denoted the linear contact distribution, is of interest in many spatial statistical applications, see e.g. Stoyan et al. (1995). An estimator is usually constructed from a series of sampling points in an observation window, and then using the empirical distribution function of the distance from the sampling points to the random set in a given direction. When the sampling points are placed independently of the random set under study, the distances to the random set may be considered as the residual lifetimes of a renewal process, where the distribution of the renewal time corresponds to the chord length distribution of the random set. So the linear contact distribution is necessarily concave.

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 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, when one introduces prior 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).

The problem of constructing nonparametric Bayes estimators for F 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.

In the present paper we use a mixture of Dirichlet processes on the space of probability distributions on $\mathbb{R}_+$, and then exploit the unique correspondence between probability distributions on $\mathbb{R}_+$ and concave cumulative distribution functions on $\mathbb{R}_+$ through their representation as mixtures of uniform distributions. The posterior distribution seems analytically intractable, so computations are performed using Markov chain Monte Carlo methods.

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 and Meier (1958) estimator. 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 Daimen 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 Polya 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 uses Gibbs sampling to compute with the posterior distribution. The present paper uses an approach similar to Doss (1994) but exploits concavity.

In Section 2 we define Dirichlet process priors and study some of their properties. Representation results for concave distribution functions are given in Section 3. After this we are ready to describe prior distributions on the space of concave distribution functions in Section 4. In Section 5 we propose a way of sampling from the posterior distribution by implementing Gibbs sampling via an urn scheme. Convergence properties of the Gibbs sampler are discussed in Section 6. In Section 7 the techniques are illustrated with data concerning failure of airconditioning equipment. The methods are, in Section 8, 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 Prior distributions on spaces of probability measures

In nonparametric Bayes inference we have to specify a prior distribution on the set of probability distributions on a given sample space. Here we follow the ideas of Ferguson (1973) restricted to the problem of placing the prior distribution on the space of probability measures on the sample space $(\mathbb{R}_+, \mathcal{B}(\mathbb{R}_+))$, where $\mathbb{R}_+ = [0, \infty)$ and $\mathcal{B}(\mathbb{R}_+)$ is the Borel σ-field.

First we let α be a finite measure on $(\mathbb{R}_+, \mathcal{B}(\mathbb{R}_+))$. Then a stochastic process P indexed by elements B of $\mathcal{B}(\mathbb{R}_+)$, is said to be a Dirichlet process on $(\mathbb{R}_+, \mathcal{B}(\mathbb{R}_+))$ with parameter α if for any measurable partition $(B_1, \ldots, B_k)$ of $\mathbb{R}_+$ the random vector $(P(B_1), \ldots, P(B_k))$ has a Dirichlet distribution with parameter $(\alpha(B_1), \ldots, \alpha(B_k))$. Hence P may be considered as a random probability measure on $(\mathbb{R}_+, \mathcal{B}(\mathbb{R}_+))$. Often we will write shortly $P \sim \mathcal{D}(\alpha)$ or $F \sim \mathcal{D}(\alpha)$ if $F(\cdot) = \int_0^{\cdot} dP$ is the cumulative distribution function corresponding to P. For the existence of such probabilities we refer to Ferguson (1973). Clearly the quantity $F(t), t \in \mathbb{R}_+$ is a random variable and if we let $H = \alpha/\alpha(\mathbb{R}_+)$ Ferguson (1973) obtained the following results

1. If $F \sim \mathcal{D}(\alpha)$ then $\mathbb{E}\, F(t) = H(t)$
2. If $F \sim \mathcal{D}(\alpha)$ and given F, $X_1, \ldots, X_n$ is a random sample from F, then
$$F \,|\, X_1, \ldots, X_n \sim \mathcal{D}\left(\alpha + \sum_{i=1}^{n} \delta_{X_i}\right)$$
3. If $F \sim \mathcal{D}(\alpha)$ then F is almost surely a discrete distribution function.

Note that property 2. says that the posterior distribution of F given the data is again a Dirichlet process, but with an updated parameter measure. Moreover it follows from 1. and 2. that

$$\mathbb{E}\,[F(t) \,|\, X_1, \ldots, X_n] = \frac{\alpha([0,t]) + \sum_{i=1}^{n} \delta_{X_i}([0,t])}{\alpha(\mathbb{R}_+) + n} \tag{1}$$

and we see that $\alpha(\mathbb{R}_+)$ represents our relative belief in the prior. Finally we may note that 1. tells us that H specifies the average behaviour of F.

However, in many situations it is not reasonable to use a fixed parameter measure α and we then choose F according to a mixture of Dirichlet distributions, see Antoniak (1974). In this situation we choose a parametric family $H_\theta, \theta \in \Theta \subset \mathbb{R}^k$ of probability measures and a prior mixing measure ν and assume that F has the following distribution

$$F \sim \int \mathcal{D}(\alpha(\mathbb{R}_+)H_\theta)\nu(d\theta).$$

To ease our notation we will use the following expressions

$$X_1, \ldots, X_n \sim F,$$

or

$$X_1, \ldots, X_n \sim \mathcal{L}(F), \qquad n \geq 1,$$

when F is a random distribution function with distribution $\mathcal{L}(F)$ and, given F, $X_1, \ldots, X_n$ is a random sample from F.

3 Representation of concave distribution functions

The idea of the paper relies on the construction of a unique 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. First we note the following lemma which can be traced back to A. I. Khincin, see e.g. Feller (1966, p. 158) for the proof.

Lemma 1 *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 in $(0,1)$ and Y having distribution G for some distribution function G on $[0,\infty)$.*

Using this lemma we can easily show the following integral representation for an arbitrary concave distribution function.

Corollary 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*

$$F(x) = \int_{[0,\infty)} F_y(x)\, G(dy), \qquad x \in \mathbb{R} \tag{2}$$

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

$$F_y(x) = \begin{cases} 0 & x < 0 \\ x/y & 0 \leq x < y \\ 1 & x \geq y. \end{cases}$$

Proof If F has the representation (2) it is clearly concave. Assume that F is concave. Define G, Y and U as in Lemma 1. It then follows that

$$F(x) = \mathbb{P}(UY \leq x) = G(0) + \int_{(0,\infty)} \mathbb{P}(U \leq x/y) G(dy) = \int_{[0,\infty)} F_y(x) G(dy).$$

which was to be shown. □

The result above could also be phrased in terms of the general theory of Choquet, see e.g. Phelps (1966), and this was indeed done by Johansen (1967): The set of concave distribution functions is convex, F_y are the extreme points of this set, and the integral representation in (2) is expressing an arbitrary element of the convex set as a mixture of its extreme points. The convex set is in fact a simplex, i.e. the mixing measure G is uniquely determined by F. The uniqueness can be established directly:

Theorem 1 *If F is concave and represented by G as in (2) 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.

Proof Assume that G satisfies (2). A direct argument gives that

$$f(x) = D^+F(x) = \lim_{h\to 0} \int_{[0,\infty)} \frac{(F_y(x+h) - F_y(x))}{h} G(dy) = \int_{(x,\infty)} \frac{1}{y} G(dy).$$

If we use the expression for the function F_y we get from (2) that

$$F(x) = \int_{[0,\infty)} F_y(x)\, G(dy) = \int_{[0,x]} G(dy) + \int_{(x,\infty)} \frac{x}{y} G(dy) = G(x) + xf(x)$$

and the result follows. □

4 Prior distributions on concave distribution functions and Bayesian inference

4.1 Concave distribution functions

Let $X = (X_1, \dots, X_n)$ be a random sample from an unknown concave distribution function F

$$X_i \sim F, \qquad i = 1, \dots, n. \tag{3}$$

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

$$F = \int_0^\infty F_y G(dy) \tag{4}$$

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

$$G \sim \int \mathcal{D}(\alpha(\mathbb{R}_+)H_\theta)\nu(d\theta). \tag{5}$$

By Corollary 1, F will be a random concave cdf, and we obtain in this way a prior distribution on the subspace of concave cdfs. The purpose is here to make inference about the concave distribution function F given data $X = (X_1, \dots, X_n)$.

4.2 Multiplicative censoring model

There is another way of expressing the model given by (3), (4), and (5), namely by the multiplicative censoring model of Vardi (1989). Let $Y = (Y_1, \dots, Y_n)$ be a random sample from the distribution function G

$$Y_i \sim G, \qquad i = 1, \dots, n \tag{6}$$

where G is a random distribution function assumed to be distributed as

$$G \sim \int \mathcal{D}(\alpha(\mathbb{R}_+)H_\theta)\nu(d\theta). \tag{7}$$

Now, let $U = (U_1, \ldots, U_n)$ be a random sample of uniformly distributed random variables over the unit interval $(0, 1)$. Finally let

$$X_i = U_i Y_i, \qquad i = 1, \ldots, n. \tag{8}$$

Then by Lemma 1 and Theorem 1, the vector $X = (X_1, \ldots, X_n)$ will be a random sample from F, specified in (4). Notice, that G is again a mixture of Dirichlet distributions, see equation (5). The task is here to make inference about the distribution function G given the incomplete data $X = (X_1, \ldots, X_n)$.

4.3 Graphical representation

Both models (3)–(5) and (6)–(8) specifies a joint distribution of model parameters and data by a few conditional distributions. This implicit assumption of independence among unspecified submodels is called the directed Markov assumption (Lauritzen et al., 1990).

A convenient way of illustrating directed Markov models (Gilks et al., 1994) and thereby the models (3)–(5) and (6)–(8) is 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.

General introductions to graphical modelling can be found in the monographs by Whittaker (1990) and Lauritzen (1996). DAG models provide a rich class of models which are particularly amenable to analysis by Gibbs sampling, see Section 5.

4.4 Bayesian inference

Next we wish to obtain the posterior distribution of F (or G) given the data X. The posterior distribution appears analytically intractable. Therefore we develop an algorithm for generating $Z^{(k)} = (Z_1^{(k)}, \ldots, Z_M^{(k)}) \sim G^{(k)}$, $k = 1, \ldots, K$ from a sample of random distribution functions, $G^{(k)}$, $k = 1, \ldots, K$, with common distribution $\mathcal{L}(G \mid X)$. This enables us to approximate the implicitely given distribution functions $G^{(k)}$ by the edf for the random sample $Z^{(k)}$. Hereby, again for a reasonably large

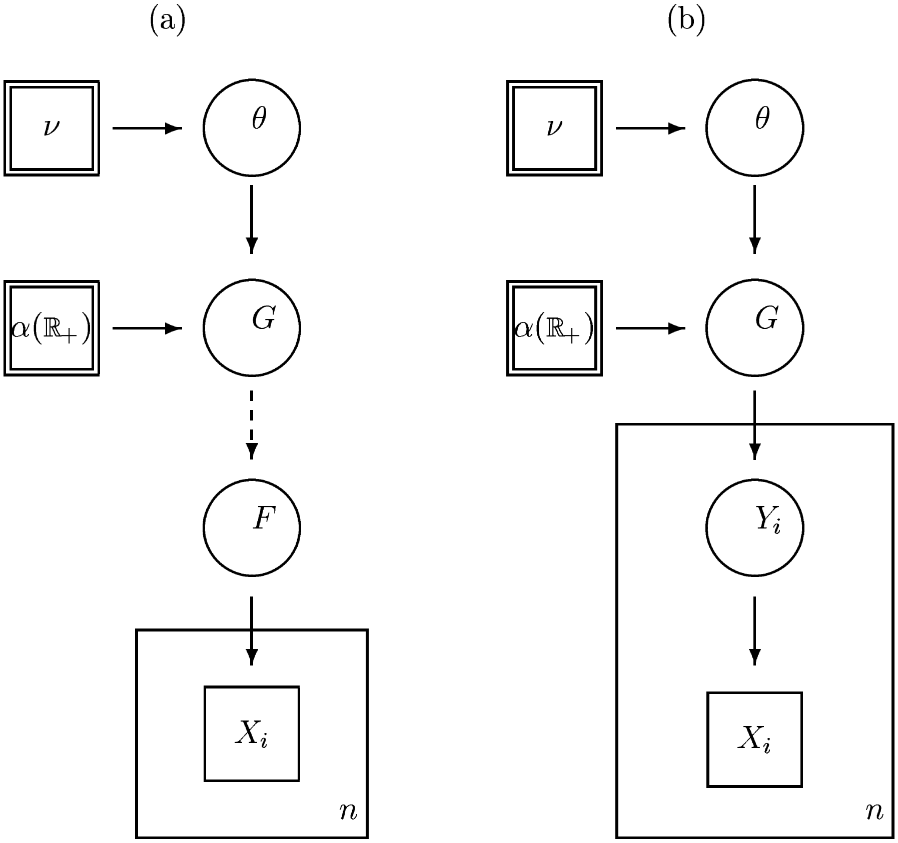

Figure 1: Graphical representations of the models suggested for concave distribution functions: (a) by a mixture; (b) by multiplicative censoring.

sample size M, it would be possible to obtain observations from distributions such as $\mathcal{L}(\text{mean}(G) \,|\, X)$ or $\mathcal{L}(\text{median}(G) \,|\, X)$ and thereby give us a method for making Bayesian inference for the multiplicative censoring model. Furthermore, by noticing that if

$$\widetilde{G} \sim \mathcal{L}(G \,|\, X) \tag{9}$$

then

$$\widetilde{F} = \int F_y \widetilde{G}(dy) \sim \mathcal{L}(F \,|\, X), \tag{10}$$

we are also able to generate a random distribution function from $\mathcal{L}(F \,|\, X)$. This makes it possible, as before, to obtain observations from distributions such as $\mathcal{L}(\text{mean}(F) \,|\, X)$, $\mathcal{L}(\text{median}(F) \,|\, X)$ or make an estimate of $\mathbb{E}\,[F \,|\, X]$.

In the present paper we only consider the case with $M = 1$ and construct an estimate of $\mathbb{E}[G \,|\, X]$ by using the edf of the sample $(Z^{(1)}, \ldots, Z^{(K)})$ and letting

$$\widehat{G}_K(t) = \frac{1}{K} \sum_{k=1}^{K} 1_{(-\infty,t]}(Z^{(k)}). \tag{11}$$

Then a natural estimate of $\mathbb{E}[F \,|\, X]$ is obtained from the edf $\widehat{G}_K$ as

$$\widehat{F}_K(t) = \int F_y(t) d\widehat{G}_K(y) = \frac{1}{K} \sum_{k=1}^{K} F_{Z^{(k)}}(t) \tag{12}$$

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

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

5 An urn scheme for sampling from the posterior distribution

In Bayesian statistics one often faces analytically intractable posterior distributions. A popular way of handling this problem is to construct a Markov chain whose limiting distribution is the required posterior distribution, i.e. in our case $\mathcal{L}(F \,|\, X)$ or $\mathcal{L}(G \,|\, X)$. 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). For a general description and survey, see for example the papers by Gelfand and Smith (1990), Smith and Roberts (1993), and Besag and Green (1993).

In the present context we are interested in the distribution of the unknown variables in Figure 1 given the observation X denoted by $\mathcal{L}(\theta, G, Y \,|\, X)$, whence a Gibbs sampler can 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 $(\theta^{(k)}, G^{(k)}) \sim \mathcal{L}(\theta, G \,|\, Y^{(k-1)}, X)$
    (b) Draw $Y^{(k)} \sim \mathcal{L}(Y \,|\, \theta^{(k)}, G^{(k)}, X)$
```

It is then seen that $(\theta^{(k)}, G^{(k)}, Y^{(k)})$ forms a Markov chain and that $\mathcal{L}(\theta, G, Y \,|\, X)$ is a stationary distribution of the chain, for more details cf. e.g. Tierney (1994, Section

2.2). If one can establish that the chain converges to its stationary distribution (see Section 6), then for large k, $(\theta^{(k)}, G^{(k)}, Y^{(k)})$ has a distribution which is approximately equal to $\mathcal{L}(\theta, G, Y \,|\, X)$. Thus by repeating the algorithm a large number K of times, one obtains $(\theta^{(k)}, G^{(k)}, Y^{(k)})$, $k = 1, \ldots, K$.

In fact we will not carry out step (a) directly; instead we will sample $\theta^{(k)}$ from the distribution $\mathcal{L}(\theta \,|\, Y^{(k-1)}, X)$, and use a simple urn scheme to obtain a direct sample $Z^{(k)}$ from a random distribution G drawn from the conditional distribution $\mathcal{L}(G \,|\, \theta^{(k)}, Y^{(k-1)}, X)$ which is then used to get $Y^{(k)}$ by a rejection step. Thus we do not explicitly represent G itself. See Sections 5.2 and 5.3 below for details.

This procedure allows for both step (b) to be performed but also to use the samples $Z^{(1)}, \ldots, Z^{(K)}$ from $G^{(1)}, \ldots, G^{(K)}$ to obtain an estimator as descibed in Section 4.4. Note that the Markov property of the DAG in Figure 1(b) gives that

$$\mathcal{L}(\theta, G \,|\, Y^{(k-1)}, X) = \mathcal{L}(\theta, G \,|\, Y^{(k-1)}) \text{ and } \mathcal{L}(Y \,|\, \theta^{(k)}, G^{(k)}, X) = \mathcal{L}(Y \,|\, G^{(k)}, X).$$

5.1 Initialization

To start the algorithm we only need $Y^{(0)}$ and it is desirable to generate it from a distribution which is close to the stationary distribution in order to make convergence happen quickly. It seems reasonable to let $\theta \sim \nu$ and to generate $Y_i^{(0)}$, $i = 1, \ldots, n$ 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)) in the following way

```
Draw $\theta \sim \nu$

For $i = 1, \ldots, n$

    Repeat
        Draw $U \sim U(0,1)$
        Draw $Y_i^{(0)} \sim H_\theta$
    Until $Y_i^{(0)} \geq X_i$ and $UY_i^{(0)} \leq X_i$
```

5.2 Sampling from $\mathcal{L}(\theta, G \,|\, Y^{(k-1)}, X)$

As mentioned above this step is not carried out as it stands. Instead we sample $\theta^{(k)}$ from the distribution $\mathcal{L}(\theta \,|\, Y^{(k-1)}, X) = \mathcal{L}(\theta \,|\, Y^{(k-1)})$ and identify $\mathcal{L}(G \,|\, \theta^{(k)}, Y^{(k-1)}, X) = \mathcal{L}(G \,|\, \theta^{(k)}, Y^{(k-1)})$, which allows us to apply the Polya sequence in Section 5.3 below. We need the following reformulation of Lemma 1 in Antoniak (1974), (see also Doss (1994, Theorem 1)).

Lemma 2 *Assume that for each $\theta \in \Theta$, H_θ is absolutely continuous, with a density h_θ that is continuous on $\mathbb{R}$. If the prior on G is given by (7), then the posterior distribution of θ given $Y = (Y_1, \ldots, Y_n)$ is given by the measure ν_Y which is absolutely continuous with respect to ν and is defined by*

$$\nu_Y(d\theta) = C(Y) \left(\prod^{*} h_\theta(Y_i) \right) \frac{(\alpha_\theta(\mathbb{R}_+))^{\#(Y)} \Gamma(\alpha_\theta(\mathbb{R}_+))}{\Gamma(\alpha_\theta(\mathbb{R}_+) + n)} \nu(d\theta)$$

where the "$*$" in the product indicates that the product is taken over distinct values only, "#" the number of distinct values, Γ is the gamma function, and $C(Y)$ is a normalizing constant. Moreover the conditional distribution of G given θ and Y is given by

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

The lemma enables us in principle to perform the step by the following

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

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

5.3 Sampling from $\mathcal{L}(Y \,|\, G^{(k)}, X)$

We will perform this in two steps by first considering a way to generate observations from $G^{(k)} \sim \mathcal{D}\left(\alpha_{\theta^{(k)}} + \sum_{i=1}^{n} \delta_{Y_i^{(k-1)}} \right)$ without actually expressing $G^{(k)}$ explicitly and then use a simple rejection method to generate observations from $\mathcal{L}(Y \,|\, G^{(k)}, X)$. 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 $(\mathbb{R}, \mathcal{B}(\mathbb{R}_+)))$ and $\{Z_l\}_{l=1}^{\infty}$ be a sequence of random variables in $\mathbb{R}_+$, drawn according to the following scheme:

$$Z_1 \sim \mathcal{D}(\mu)$$

and

$$Z_{l+1} \,|\, Z_1, \ldots, Z_l \sim \mathcal{D}(\mu_l), \qquad l = 2, 3, \ldots$$

where $\mu_l = \mu + \sum_{i=1}^{l} \delta_{Z_i}$. Using the terminology of Blackwell and MacQueen (1973), $\{Z_l\}_{l=1}^{\infty}$ is a Pólya sequence on $\mathbb{R}_+$ with parameter μ. Blackwell and MacQueen (1973) show that

$$Z_1, Z_2, \ldots \sim \mathcal{D}(\mu).$$

As $\mathcal{L}(G \,|\, \theta^{(k)}, Y^{(k-1)})$ has a $\mathcal{D}(\alpha_{\theta^{(k)}} + \sum_{i=1}^{n} \delta_{Y_i^{(k-1)}})$-distribution it is easy to sample variables by using a Pólya sequence. As we actually want sample values from $\mathcal{L}(Y \,|\, G^{(k)}, X)$ we use the following rejection scheme

```
$l = 1$

for $i = 1, \ldots, n$

    Repeat
        Draw $Z_l \sim (\alpha_{\theta^{(k-1)}} + \sum_{i=1}^{n} \delta_{Y_i^{(k-1)}} + \sum_{i=1}^{l-1} \delta_{Z_i}) / (\alpha_{\theta^{(k-1)}}(\mathbb{R}_+) + n + (l-1))$
        $l = l + 1$
        Draw $U \sim \mathrm{U}(0,1)$
    Until $Z_l U \leq X_i$ and $Z_l \geq X_i$
    $Y_i^{(k)} = Z_l$
```

Note that, the sequence $\{Z_l\}$ represents successive draws from an urn with a continuum of colours ($\mathbb{R}_+$) with different 'chances' of being drawn attached to them. Initially the urn has n balls of colours $Y_1^{(k-1)}, \ldots, Y_n^{(k-1)}$ each with probability $1/(\alpha_{\theta^{(k-1)}} + n)$ of being drawn and a continuum of colours $\mathbb{R}_+$ with a chance of being drawn given by the density $\alpha_{\theta^{(k-1)}}/(\alpha_{\theta^{(k-1)}} + n)$. After each draw the ball is replaced and another ball of its same colour is added to the urn.

5.4 Final algorithm

The algorithm can now altogether be formulated as

Algorithm

```
Draw $\theta \sim \nu$

for $i = 1, \ldots, n$

    Repeat
        Draw $U \sim U(0,1)$
        Draw $Y_i^{(0)} \sim H_\theta$
    Until $Y_i^{(0)} \geq X_i$ and $UY_i^{(0)} \leq X_i$

for $k = 1, \ldots, K$
```

```
    Draw $\theta^{(k)} \sim \nu_Y^{(k-1)}$
    $l = 1$
    for $i = 1, \dots, n$
        Repeat
            Draw $Z_l \sim (\alpha_{\theta^{(k)}} + \sum_{i=1}^{n} \delta_{Y_i^{(k-1)}} + \sum_{i=1}^{l-1} \delta_{Z_i}) / (\alpha_{\theta^{(k)}}(\mathbb{R}_+) + n + (l-1))$
            $l = l + 1$
            Draw $U \sim \mathrm{U}(0,1)$
        until $Z_l U \leq X_i$ and $Z_l \geq X_i$
        $Y_i^{(k)} = Z_l$
```

The algorithm is computationally straightforward and the sampling process results in approximate draws from a random distribution function G with distribution $\mathcal{L}(G \mid X)$. We briefly discuss the convergence properties of the algorithm below.

6 Convergence of the algorithm

To start the algorithm we first choose $\theta \sim \nu$ and then generate $Y_i^{(0)} \sim H_{\theta,i}$ independently for $i = 1, \dots, n$, where $H_{\theta,i}$ is H_θ constrained to the data. Assume that for each $\theta \in \Theta$, H_θ is absolutely continuous, with a density h_θ. Then $H_{\theta,i}$ is absolutely continuous with a density $h_{\theta,i}$ proportional to

$$h_{\theta,i}(y) \propto \frac{h_\theta(y)}{y} 1_{(x_i,\infty)}(y).$$

Let τ denote the distribution of $Y^{(0)} = (Y_1^{(0)}, \dots, Y_n^{(0)})$ i.e. τ is the product measure $\tau = H_{\theta,1} \times \cdots \times H_{\theta,n}$. Recall that the Markov chain constructed in the algorithm is specified by

```
Draw $Y^{(0)} \sim \tau$

For $k = 1, 2, \dots$

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

The validity of the algorithm requires a proof that $\mathcal{L}(Y^{(k-1)}, \theta^{(k)}, G^{(k)} \mid Y^{(0)})$ converges to the stationary distribution of the Markov chain. The result below can be established in complete analogy with Doss (1994). We omit the details.

Let $\mathcal{B}_\Theta$ be the Borel σ-field on Θ, let $\mathcal{B}_{\mathbb{R}^k}$ be the Borel σ-field on $\mathbb{R}^k$ and let $\mathcal{B}_\mathcal{P}$ be the Borel σ-field on $\mathcal{P}$ defined by the smallest σ-field on $\mathcal{P}$ such that the function $P \to \mathbb{P}(A)$ is measurable for each Borel set A.

Theorem 2 *Assume there exist a set $E_0 \subset \Theta$ with $\nu(E_0) > 0$, a $\delta > 0$ and for $i = 1, \ldots, n$ disjoint sets $E_i \subset (X_i, \infty)$ with positive finite Lebesgue measure such that*

(A1) *$\nu_Y(C_0) \geq \delta\nu(C_0)$ for all $Y \in E_1 \times \cdots \times E_n$ and all $C_0 \subset E_0$.*

(A2) *$H_\theta(C_i) \geq \delta\lambda(C_i)$ for all $\theta \in E_0$ and $C_i \subset E_i$, $i = 1, \ldots, n$.*

(A3) *$\nu_Y(E_0) > 0$ whenever $Y_i \in (X_i, \infty)$.*

(A4) *$H_\theta(E_i) > 0$ for all $i = 1, \ldots, n$, whenever $\theta \in E_0$.*

(A5) *There exists $\eta > 0$ such that*

$$\frac{(\alpha_\theta(\mathbb{R}_+))^n\Gamma(2\alpha_\theta(\mathbb{R}_+) + n)}{\Gamma(\alpha_\theta(\mathbb{R}_+) + 2n)} > \eta \text{ for all } \theta \in E_0.$$

(A6) *For every Y such that $Y_i \in (X_i, \infty)$, $i = 1, \ldots, n$ we have*

$$\prod_{i=1}^n h_\theta(Y_i) > 0 \text{ for } \nu \text{ a.e. } \theta.$$

Then

$$\sup_{B \in \mathcal{B}_{\mathbb{R}^k} \times \mathcal{B}_\Theta \times \mathcal{B}_\mathcal{P}} |P\{(Y^{(k-1)}, \theta^{(k)}, G^{(k)}) \in B \,|\, Y^{(0)}\} - P\{(Y, \theta, G) \in B \,|\, X\}| \to 0$$

for τ-almost all $Y^{(0)}$.

This, in particular implies the following

Corollary 2 *Assume conditions (A1)–(A6) of Theorem 2 are fulfilled. Then*

$$\sup_{B \in \mathcal{B}_\mathcal{P}} |P\{G^{(k)} \in B | Y^{(0)}\} - P\{G \in B \,|\, X\}| \to 0$$

for τ-almost all $Y^{(0)}$.

7 Failure of airconditioning equipment

We now discuss an example to illustrate the techniques. The stepfunctions in Figure 3 give the edfs of the intervals in operating hours between successive failures of airconditioning equipment in 13 Boeing 720 aircrafts. This data set was presented in Proschan (1963) and further analyzed by Cox and Lewis (1966).

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

Cox and Lewis (1966) argued that the combined data follow a distribution which is nearly exponential. Therefore it seems sensible to use a parametric family of exponential distributions as a model for the combined data. If this approach is used one obtains a maximum likelihood estimate of the reciprocal mean parameter $\theta_0 = 0.0107$. However, it seems reasonable as a starting point to assume an individual survival function for each aircraft, see Cox and Lewis (1966). An alternative to using an exponential distribution is just to assume a decreasing density.

We analyzed the data set using a prior on F given by model (3)–(5), where H_θ is an exponential distribution with reciprocal mean parameter θ. With $\alpha_\theta(\mathbb{R}_+)$ being constant we chose to consider two cases $\alpha_\theta(\mathbb{R}_+) = 1$ and $\alpha_\theta(\mathbb{R}_+) = 100$, representing essentially a situation of a diffuse prior and a situation where we are rather confident about our prior. Furthermore we took ν to be a gamma distribution with position parameter 1000 and shape parameter $1000{\cdot}\theta_0$, hence F is centered around the $\mathrm{Exp}(\theta_0)$-distribution. The reason for taking a gamma prior on θ is that it is a conjugate family for the exponential distribution. If ν is the $\mathrm{Gamma}(a, b)$-distribution, then (see Lemma 2) ν_Y is the $\mathrm{Gamma}(a + 2n^*, b + \sum^* y_i)$-distribution where n^* is the number of distinct observations in Y and $\sum^* y_i$ is the sum of distinct y_i's.

In order to study convergence properties of the algorithm we have chosen to present a diagnostics plot in Figure 2. For each run of the algorithm we obtain a random sample $Y^{(k)} = (Y_1^{(k)}, \dots, Y_n^{(k)})$ from $\mathcal{L}(Y \mid G^{(k)}, X)$ which for large k should be a random sample from $\mathcal{L}(Y \mid X)$. We study convergence properties of the algorithm by plotting $\mathrm{mean}(Y^{(k)}) = n^{-1} \sum_{i=1}^n Y_i^{(k)}$ versus k for three independent starts of the algorithm. From the diagnostics plot we see that for all aircrafts the trace of the sampled values for each run quickly settled down to a stable level, which we take as an indication that the stationary regime has been achieved almost from the beginning of the sampling procedure. Figure 3 shows a plot of different methods for estimating the distribution function F for each aircraft. It shows a plot of the distribution function of a future observation for the cases where $\alpha_\theta(\mathbb{R}_+) = 1$ and $\alpha_\theta(\mathbb{R}_+) = 100$. These were obtained by running the algorithm for a burn-in of 100 iterations and first obtaining a G from $\mathcal{L}(G \mid X)$, and then generating a random variable Y from

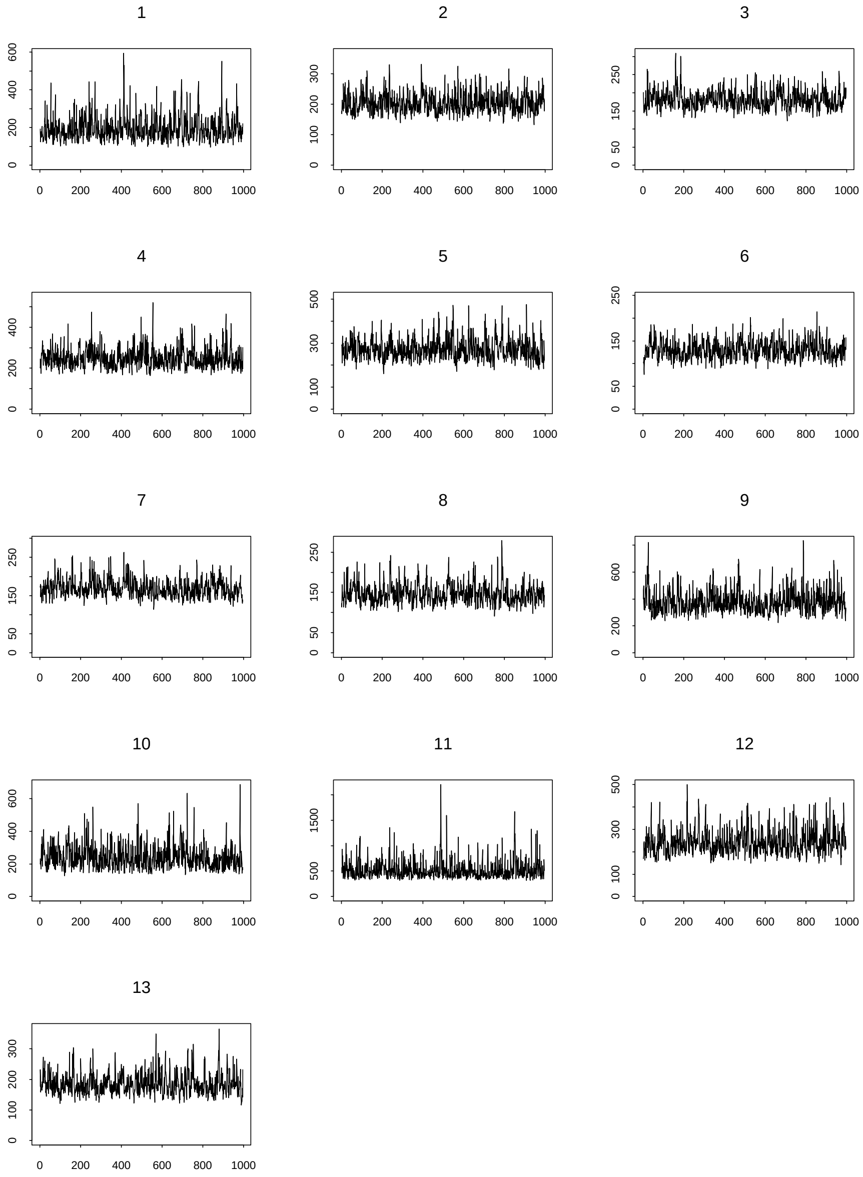

Figure 2: Airconditioning equipment. Sampled values of mean($Y^{(k)}$) versus k for Algorithm 1.

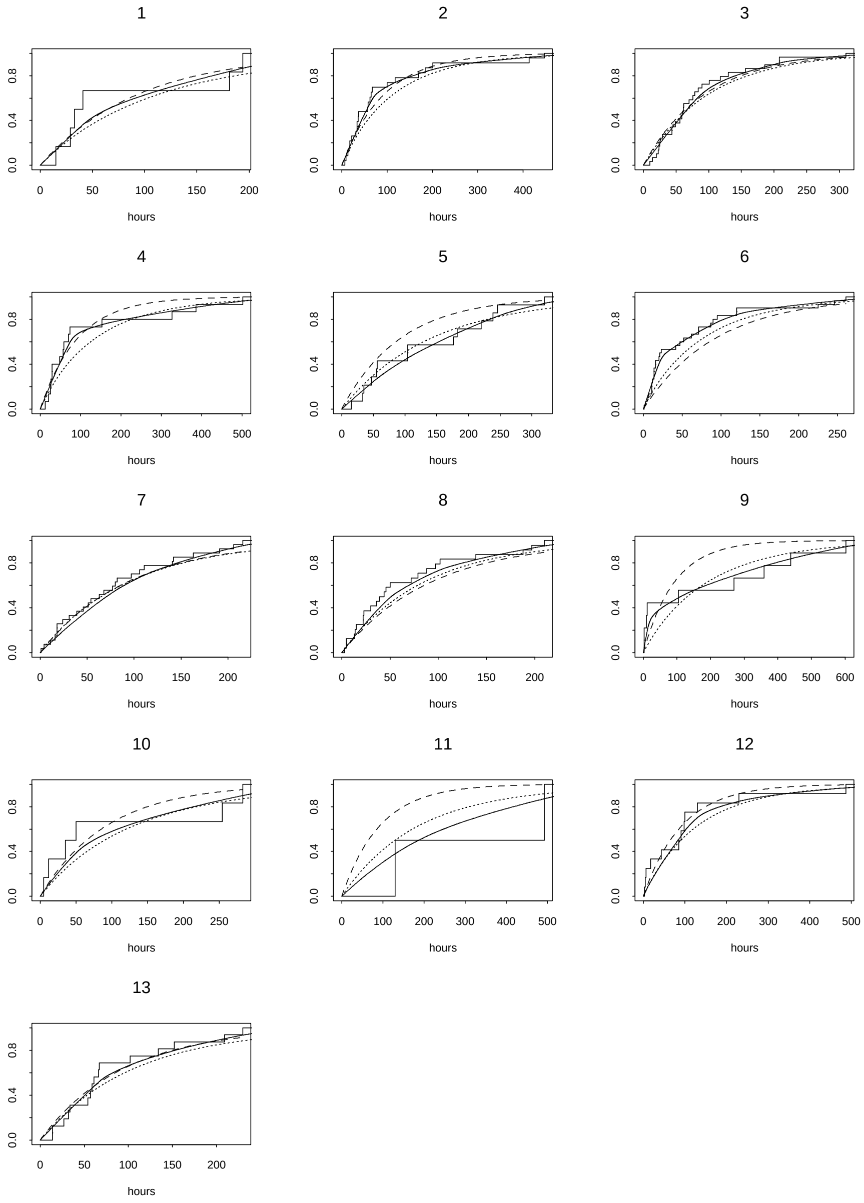

Figure 3: Estimated distribution function for aircraft data. The empirical distribution function is illustrated by a stepfunction. The fulldrawn and dotted lines corresponds to weight 1 and 100, respectively. The mean of the prior distribution is illustrated by a dashed line.

this G. This was repeated independently $K = 3000$ times, yielding a random sample $Y_1, \ldots, Y_K$. An estimate of $\mathbb{E}[G \mid X]$ and $\mathbb{E}[F \mid X]$ was obtained by using the empirical distribution function of the random samples, as shown in (11) and (12).

The prior distribution is also displayed. From Figure 3 we see that the Bayes estimate is concave and provides a smoothed estimate of the distribution function. For small weights we note that the estimate follows the data quite closely. However, for larger weights the estimate has a tendency to have a shape more closely resembling the shape of the prior.

It is straightforward to check theoretical convergence of the algorithm in the present example by verifying assumption (A1)–(A6) of Section 6.

8 Extensions to further censoring

The graphical model formulation as provided in Figure 1 in combination with Gibbs sampling provides a powerful tool for Bayesian inference in structured model formulations. In the present section we capitalize on this fact to introduce a straightforward extension to situations with further censoring.

Assume that our model setup is as specified by (3)–(5) or alternatively by (6)–(8). Moreover we suppose that the X's are not directly observed; instead we observe

$$(B_i, \delta_i) = \begin{cases} (\{X_i\}, 1), & X_i \notin A_i \\ (A_i, 0) & X_i \in A_i \end{cases}, \; i = 1, \ldots, n, \tag{13}$$

where the A_i's are subsets of $\mathbb{R}_+$, with positive Lebesgue measure. In case of right-censored data $A_i = (c_i, \infty)$ and X_i is censored on the right by c_i. We wish to obtain the posterior distribution of F given the incomplete data. The model can be interpreted as incomplete data from a decreasing density or alternatively as multiplicatively censored data which are incompletely observed. Graphical representations of the suggested model are shown in Figure 4.

By inspecting the graphical representation in Figure 4 (b) and reading off the conditional independences we can for $B = ((B_1, \delta_1), \ldots, (B_n, \delta_n))$ formulate the following general Gibbs sampling scheme

```
For $k = 1, 2, \ldots$
    (a) Draw $(\theta^{(k)}, G^{(k)}) \sim \mathcal{L}(\theta, G \mid Y^{(k-1)}, X^{(k-1)}, B) = \mathcal{L}(\theta, G \mid Y^{(k-1)})$
    (b) Draw $Y^{(k)} \sim \mathcal{L}(Y \mid \theta^{(k)}, G^{(k)}, X^{(k-1)}, B) = \mathcal{L}(Y \mid G^{(k)}, X^{(k-1)})$
    (c) Draw $X^{(k)} \sim \mathcal{L}(X \mid \theta^{(k)}, Y^{(k)}, B) = \mathcal{L}(X \mid Y^{(k)}, B)$
```

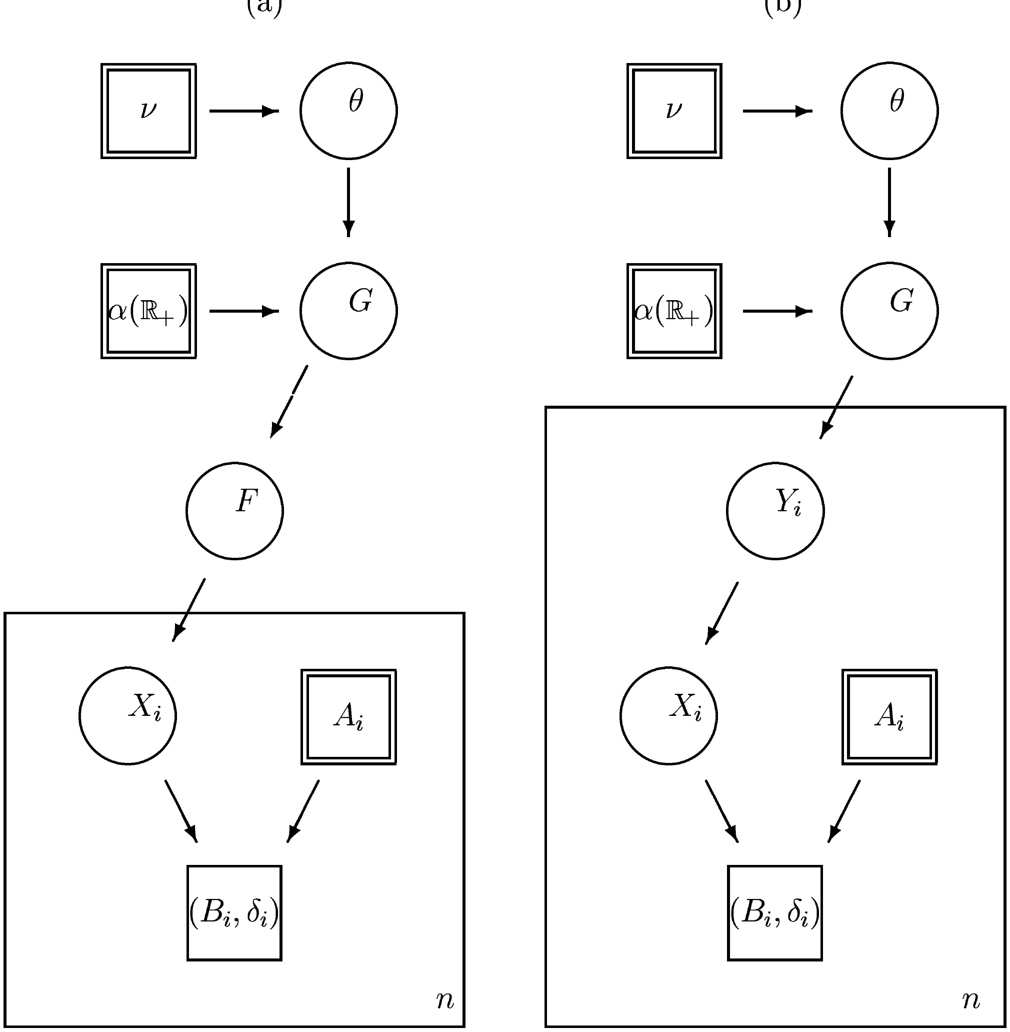

Figure 4: Graphical representations of the models suggested for Bayesian inference in the case of incompletely observed data from concave distribution functions represented (a) by mixtures; (b) by multiplicative censoring.

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

We analyzed the data set using a prior on F given by the model (3)–(5) and (13). So in our notation $(B_1, \delta_1) = (\{0.8\}, 1)$, $(B_2, \delta_2) = (\{3.1\}, 1)$, $(B_3, \delta_3) = (\{5.4\}, 1)$, $(B_4, \delta) = (\{9.2\}, 1)$, $(B_5, \delta_5) = (\{(1.0, \infty)\}, 0)$, $(B_6, \delta_6) = (\{(2.7, \infty)\}, 0)$, $(B_7, \delta_7) = (\{(7.0, \infty)\}, 0)$, $(B_8, \delta_8) = (\{(12.1, \infty)\}, 0)$. Furthermore we let H_θ be the gamma distribution with shape parameter 2 and position parameter $\theta = 0.12$, (which is the maximum likelihood estimator for the reciprocal mean, if data is assumed to follow an exponential distribution). We chose $\alpha_\theta(\mathbb{R}_+) = 16$ and took ν to be a gamma distribution with position parameter 12 and shape parameter 100. Figure 5 shows an estimate of the survival function in the case $\alpha_\theta(\mathbb{R}_+) = 16$. This was obtained

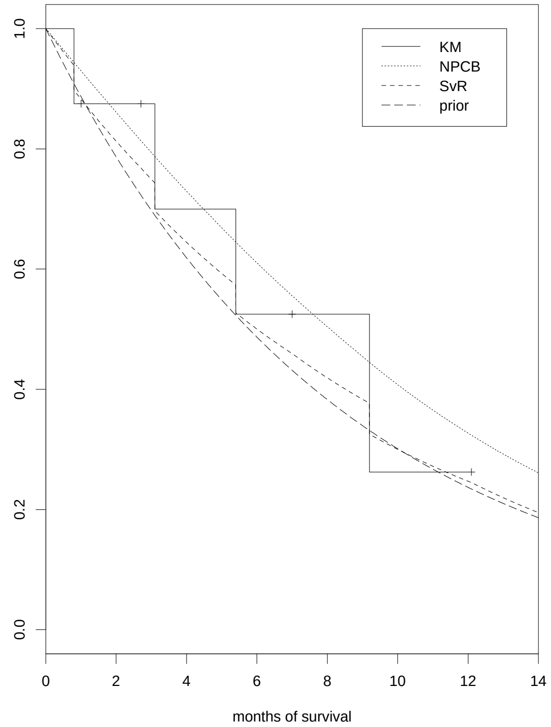

Figure 5: Estimated survival function for the Kaplan-Meier data. The Kaplan-Meier estimator is illustrated by a stepfunction (KM). The dotted line corresponds to the nonparametric concave Bayes estimator (NPCB). The two dashed lines correspond to the prior mean (prior) and Susarla and van Ryzin's (1978) nonparametric estimator (SvR).

by running the algorithm for a burn in of 100 iterations and first obtaining a G from $\mathcal{L}(G \mid Z)$, and then generating a random variable Y from this G. This was repeated independently $K = 3000$ times. The Bayes estimate of the survival function $\mathbb{E}\,[1 - F \mid Z]$ was estimated by $1 - \widehat{F}_K$, where $\widehat{F}_K$ is constructed analogously to (12). In Figure 5 we have also plotted the Kaplan-Meier estimator of the survival function. Moreover, the nonparametric Bayes estimator (NPB) as introduced in Susarla and van Ryzin (1976) with weight $\beta = 16$ is plotted. From Figure 5 we see that taking the concave assumption on the distribution function yields a smooth survival function compared to the nonparametric approach without concavity assumption. Moreover we see that the weight 16 seems to yield a rather diffuse prior in the case of concavity compared to the Bayes approach without concavity assumption.

9 Discussion

9.1 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 $\alpha(\mathbb{R}_+)$ 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_0. Is it true that the posterior distribution for F converges (F_0-almost surely in the weak-star topology) to the measure degenerate at F_0? Or, slightly weaker, does the posterior expectation of F converge to F_0 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. See for example Diaconis and Freedman (1986) who show how bad things could potentially be.

We have not been able to resolve this question in general, but we do believe that in the case without censoring, or with reasonable censoring patterns, the Bayes estimate suggested in this paper will be consistent. This conjecture is supported by a small simulation experiment, the results of which are displayed in Figure 6. Observations X_i were simulated from F_0 being uniform on the interval $(0,1)$, and the Bayes estimate of F calculated using a prior distribution similar to that used in Section 7. Thus H_θ is exponential with parameter θ and ν a gamma distribution with position parameter 1000 and shape parameter $1000 \cdot \theta_0$, where $\theta_0 = .75$. Different weights were attatched to the prior by considering the two cases $\alpha_\theta(\mathbb{R}_+) = 1$ and $\alpha_\theta(\mathbb{R}_+) = 100$. 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.

The posterior expectation of F was calculated after 10, 100, and 1000 observations. It appears from the figure that this posterior expectation gets closer and closer to F_0 as the sample size increases.

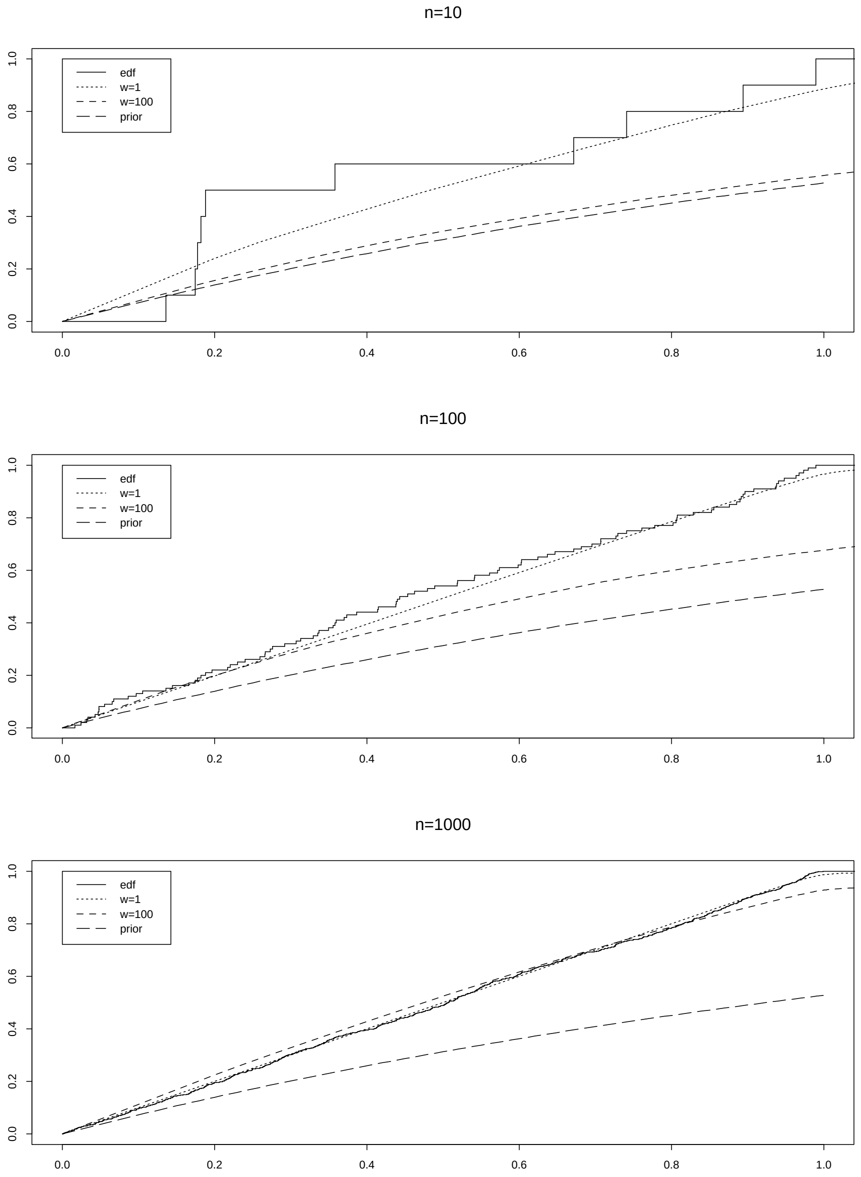

Figure 6: Bayes estimates of a concave distribution function based on 10, 100, and 1000 observations from a uniform distribution. The empirical distribution function and the prior mean are displayed for comparison.

9.2 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 parametrize the distributions. Thus there is a considerable potential for extension to other cases with a similar structure. Also it could be of interest to extend the methods to other prior distributions than Dirichlet-processes, such as for example Polya-tree distributions.

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. *Ann. Statist.*, 2:1152–1174.

Besag, J. and Green, P. J. (1993). Spatial statistics and Bayesian computation. *J. Roy. Statist. Soc. Ser. B*, 55:25–37.

Blackwell, D. and MacQueen, J. B. (1973). Ferguson distributions via Pólya urn schemes. *Ann. Statist.*, 1:353–355.

Cox, D. R. and Lewis, P. A. W. (1966). *The Statistical Analysis of Series of Events.* Methuen, New York.

Daimen, D., Laud, P. W., and Smith, A. F. M. (1996). Implementation of Bayesian non-parametric inference based on beta processes. *Scand. J. Statist.*, 23:27–36.

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

Doss, H. (1994). Bayesian nonparametric estimation for incomplete data via successive substitiution sampling. *Ann. Statist.*, 22:1763–1786.

Feller, W. (1966). *An Introduction to Probability Theory and Its Applications*, volume II. John Wiley and Sons, New York.

Ferguson, T. S. (1973). A Bayesian analysis of some nonparametric problems. *Ann. Statist.*, 1:209–230.

Ferguson, T. S. (1974). Prior distributions on spaces of probability measures. *Ann. Statist.*, 2:615–629.

Gelfand, A. E. and Smith, A. F. M. (1990). Sampling-based approaches to calculating marginal densities. *J. Amer. Statist. Assoc.*, 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*, 6:721–741.

Gilks, W. R., Thomas, A., and Spiegelhalter, D. J. (1994). A language and program for complex Bayesian modelling. *The Statistician*, 43:169–178.

Grenander, U. (1956). On the theory of mortality measurement, part II. *Skand. Aktuarietidskr.*, 39:125–153.

Groeneboom, P. (1985). Estimating a monotone density. In Cam, L. L. and Olshen, R., editors, *Proceedings of the Berkeley Conference in Honor of Jerzy Neyman and Jack Kiefer, Vol. II*, pages 539–555, Monterey. Wadsworth.

Groeneboom, P. and Lopuhaä, H. (1993). Isotonic estimators of monotone densities and distribution functions. *Statist. Neerlandica*, 47:175–184.

Hansen, M. B. (1996). Kaplan-Meier type estimators for first contact distribution functions. In *The XVIIIth International Biometric Conference, Amsterdam, Invited papers*, pages 201–211.

Hansen, M. B., Gill, R. D., and Baddeley, A. (1996). Kaplan-Meier type estimators for linear contact distributions. *Scand. J. Statist.*, 23:129–155.

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

Johansen, S. (1967). Anvendelse af ekstremalpunktsmetoder i sandsynlighedsregningen. University of Copenhagen. In Danish.

Kaplan, E. L. and Meier, P. (1958). Nonparametric estimation from incomplete observations. *J. Amer. Statist. Assoc.*, 53:457–481.

Laslett, G. M. (1982). The survival curve under monotone density constraints with applications to two-dimensional line segment processes. *Biometrika*, 69:153–160.

Laud, P. W., Smith, A. F. M., and Damien, P. (1996). Monte Carlo methods for approximating a posterior hazard rate process. *Statist. Comput.*, 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.

Muliere, P. and Walker, S. (1997). A Bayesian non-parametric approach to survival analysis using Polya trees. *Scand. J. Statist.*, 24:331–340.

Phelps, R. R. (1966). *Lectures on Choquet's Theorem.* Van Nostrand, New York.

Proschan, F. (1963). Theoretical explanation of observed decreasing failure rate. *Technometrics*, 5:375–383.

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

Sethuraman, J. (1994). A constructive definition of Dirichlet priors. *Statist. Sinica*, 4:639–650.

Smith, A. F. M. and Roberts, G. O. (1993). Bayesian computation via the Gibbs sampler and related Markov chain Monte Carlo methods. *J. Roy. Statist. Soc. Ser. B*, 55:3–23.

Stoyan, D., Kendall, W. S., and Mecke, J. (1995). *Stochastic Geometry and Its Applications.* John Wiley and Sons, Chichester, 2nd edition.

Susarla, V. and van Ryzin, J. (1976). Nonparametric Bayesian estimation of survival curves from incomplete observations. *J. Amer. Statist. Assoc.*, 71:897–902.

Tierney, L. (1994). Markov chains for exploring posterior distributions. *Ann. Statist.*, 22:1701–1762.

Turnbull, B. W. (1974). Nonparametric estimation of a survivorship function with doubly censored data. *J. Amer. Statist. Assoc.*, 69:169–173.

Turnbull, B. W. (1976). The empirical distribution function with arbitrarily grouped, censored, and truncated data. *J. Roy. Statist. Soc. Ser. B*, 38:290–295.

Vardi, Y. (1989). Multiplicative censoring, renewal processes, deconvolution and decreasing density: Nonparametric estimation. *Biometrika*, 76:751–761.

Whittaker, J. (1990). *Graphical Models in Applied Multivariate Analysis.* John Wiley and Sons, Chichester.