A comparison of nonparametric priors in hierarchical mixture modelling of lifetime data

RAFFAELE ARGIENTO\textsuperscript{a}, ALESSANDRA GUGLIELMI\textsuperscript{a,b} and ANTONIO PIEVATOLO\textsuperscript{a}

IMATI-CNR\textsuperscript{a} and Politecnico di Milano\textsuperscript{b}

23 May 2007

Abstract

We will pursue a Bayesian nonparametric approach in the hierarchical mixture modelling of lifetime data in two situations: density estimation, when the distribution is a mixture of parametric densities with a nonparametric mixing measure, and accelerated failure time regression modelling, when the same type of mixture is used for the distribution of the error term. The Dirichlet process is a popular choice for the mixing measure, yielding a Dirichlet process mixture model for the error; as an alternative, we also allow the mixing measure to be equal to a normalized inverse-Gaussian prior, built from normalized inverse-Gaussian finite dimensional distributions, as recently proposed in the literature. Markov chain Monte Carlo techniques will be used to estimate the predictive distribution of the survival time, along with the posterior distribution of the regression parameters. A comparison between the two models will be carried out on the grounds of their predictive power and their ability to identify the number of components in a given mixture density.

Keywords: Accelerated failure time regression models, Bayesian semiparametrics, Mixture models, MCMC algorithms.


1 Introduction

In the survival literature, the accelerated failure time model is usually meant as the multiplicative effect of a fixed \( p \)-vector of covariates \( x = (x_1, \ldots, x_p)' \) on the failure time \( T \), \textit{i.e.} \( \log T = -x'\beta + W \), where \( \beta = (\beta_1, \ldots, \beta_p)' \) is the vector of regression parameters and \( W \) denotes the error. Recently this model has received much attention in the Bayesian community, in particular in papers where the error \( W \) (or \( \exp(W) \)) has been represented hierarchically as a mixture of parametric densities with a Dirichlet process as mixing measure, \textit{i.e.}, the
well-known Dirichlet process mixture (DPM) models, introduced by Lo (1984). In Kuo and Mallick (1997), the authors model exp(W) as a Dirichlet process location mixture of normals (with variance fixed and small). Kottas and Gelfand (2001) and Gelfand and Kottas (2003) propose a flexible semiparametric class of zero median distributions for the error, which essentially consists of a Dirichlet process scale mixture of split 0-mean normals (with skewness handled parametrically). In Ghosh and Goshal (2006) the distribution of exp(W) is given as a scale mixture of Weibull distributions with a Dirichlet process as mixing measure, and the consistency of the posterior distribution of the regression parameters is established. Pointing out that the marginal prior for exp(W) in Kuo and Mallick (1997) gives positive probability to the negative reals, Hanson (2006) proposes a DPM model of gamma densities, mixing over both the scale and the shape of the gammas, for the distribution of exp(W).

In the paper we will consider the model $T = \exp^{-x' \beta} \cdot V$, where $V := e^W$. It is equally reasonable to work on $V$ rather than $W$, with the advantage that the survival time is modelled directly, thus facilitating also prior specification. Here the law of $V$ is assumed to be a nonparametric hierarchical mixture model, i.e. a mixture of some parametric family of densities on the positive real line, mixed by a random distribution function $G$ on $\mathbb{R}^s$ ($s$ is a positive integer). The paper mainly focuses on the performances of two hierarchical mixture models for the error $V$, comparing Bayesian inferences on the predictive survival times. On one hand we will assume that $G$ has a Dirichlet process prior, yielding DPM models for $V$; on the other hand, $G$ will have the normalized inverse-Gaussian prior (N-IG prior), as introduced in Lijoi, Mena and Prünster (2005), thus defining what we call N-IG mixtures for short. The matching between the two priors is achieved centering $G$ at the same distribution function $G_0$, and assuming equal prior means for the number of components in the mixture. We will also compare the models in the related case of mixture density estimation, where we investigate their ability in identifying the actual number of components of a given density. In all cases a sensitivity assessment is performed.

N-IG mixtures of normals have been studied in Lijoi, Mena and Prünster (2006), but, at the best of our knowledge, no papers are available on N-IG mixture with kernel having support on $\mathbb{R}^+$, including a regression component. The N-IG prior preserves the same tractability as the Dirichlet process prior, and, furthermore, it is characterized by a more elaborate clustering property: the weight given to a single observation depends on the whole number of ties in the sample. Moreover, the prior distribution of the number of components for the N-IG mixture is more “spread out” compared to the same prior under the DPM model.

Following Hanson (2006), we will consider hierarchical mixtures of gamma densities, mixed on both the scale and the shape parameters. As far as the centering distribution $G_0$ is concerned, we will consider distributions yielding conjugate and nonconjugate models. In the
first case the resulting marginal prior mean for \( V \) does not exist; however prior information can be expressed by means of the median of \( V \).

Posterior inference on \( \beta \) and \( P(T > t) \) is carried out via the Polya urn Gibbs sampler, that is, an MCMC scheme where the main update step is the sampling from the full conditional distributions of the parameters of the kernel densities. Neal (2000) gave an extensive review of such algorithms, describing also the data augmentations needed for improving the mixing of the algorithm and for dealing with the case where the kernel is non-conjugate with \( G_0 \). We will provide a detailed illustration of the augmented state space occurring in this case. A further data augmentation will be needed for incorporating censoring. The two “competing” models will be tested on real and simulated data. We will use the same notation to denote probability measures and their distribution functions.

The plan of the paper is as follows. Section 2 presents the accelerated failure time model, while in Section 3 we introduce notation and basic definitions about nonparametric hierarchical mixtures. Section 4 gives the structure of the accelerated failure time model and the nonparametric priors under consideration. Computational algorithms are discussed in Section 5 and applications are presented in Section 6. Conclusions and comments are given in Section 7.

2 The AFT model

The accelerated failure time (AFT) model (Kalbfleisch and Prentice, 2002) specifies a log-linear relationship between time-to-event \( T \) and covariates \( x = (x_1, \ldots, x_p)' \) via

\[
\log T = -x' \beta + W,
\]

where \( W \) is an error variable with support in \( \mathbb{R} \), and \( \beta = (\beta_1, \ldots, \beta_p)' \) is the vector of regression parameters. The equivalent formulation

\[
T = \exp(-x' \beta) V,
\]

where \( V := \exp(W) \) is a nonnegative random error, clarifies that the role of covariates \( x \) is to accelerate (or decelerate) the time-to-event. This model gained popularity as an alternative to the Cox proportional hazard model because of its intuitive physical interpretation. It is quite natural to consider a semiparametric approach for it, when \( \beta \) is modeled parametrically and the error \( W \) (or \( V \)) nonparametrically to avoid too restrictive parametric assumption on the error distribution. Semiparametric approaches to the AFT model, in the frequentist realm, date back to the initial work of Buckley and James (1979). More recent work on censored data, based on \( L_1 \) regression, includes Jin, Lin, Wei and Ying (2003) and Jin, Lin
and Ying (2006), but computationally difficulties could probably explain its limited usage. On the other hand, the Bayesian semiparametric approach is especially attractive in this regard, because inference is exact and predictive power may be gained by assuming a centering parametric “baseline” form for the error distribution.

3 Nonparametric hierarchical mixture models

By a nonparametric hierarchical mixture (NPHM) model we mean a mixture of parametric distributions (usually absolutely continuous) with a random mixing distribution (i.e., a random probability measure). The model can be expressed via a latent variable $\theta$ as follows

$$Z|\theta \sim k(\cdot; \theta),$$

$$\theta \sim G,$$

$$G \sim q,$$

where $k(\cdot; \theta)$ denotes a parametric (absolutely continuous) density for any $\theta$, and $G$ is a random probability measure (or a random distribution function) with prior $q$. NPHM models provides a natural generalization of existing parametric AFT models. In the context of NPHM, Ferguson (1983) and Lo (1984) used a Dirichlet process prior on the mixing distribution, introducing the widely-known Dirichlet processes mixture (DPM) models, popularized by Escobar and West (1995) who provided a first Markov Chain Monte Carlo algorithm for the computation of the Bayesian estimates in the context of density estimation. Here, two NPHM models for the error $V$ will be compared, namely when $G$ is a Dirichlet process, or an N-IG process. To explain the notation, we introduce an inverse-Gaussian random variable $Z$ with shape parameter $\bar{M} \geq 0$ and scale parameter $\gamma > 0$, which we will denote by $Z \sim IG(\bar{M}, \gamma)$, i.e. a positive absolutely continuous random variable $Z$ with density

$$f(z; \bar{M}, \gamma) = \frac{\bar{M}}{\sqrt{2\pi}} z^{-3/2} \exp\left\{ -\frac{1}{2} \left( \frac{\bar{M}^2}{z} + \gamma^2 z \right) + \gamma \bar{M} \right\}, \quad z \geq 0.$$
the N-IG prior and $G \sim DP(aG_0(\cdot))$ to denote the Dirichlet prior with parameter measure $aG_0(\cdot)$.

Recall that, for both the Dirichlet process and N-IG priors, $E(G(A)) = G_0(A)$, whereas $\text{Var}(G(A)) = G_0(A)(1 - G_0(A))M^2e^M\Gamma(-2, M)$ for the N-IG prior ($\Gamma(\cdot, \cdot)$ denotes the incomplete gamma function), and, on the other hand, $\text{Var}(G(A)) = (G_0(A)(1 - G_0(A)))/(a + 1)$ under the Dirichlet prior. It is worth noting that a N-IG process $G$ on $\Theta$ belongs to the class of random measures with independent increments (equivalently, $G$ is a completely random measure, that is, if $A_i$ and $A_j$ are disjoint and measurable, then $G(A_i)$ and $G(A_j)$ are independent random variables); see Regazzini, Lijoi and Prünster (2003) for an introduction to random measures with independent increments in Bayesian nonparametrics. The N-IG process is also, when its parameter measure is non-atomic, a special case of species sampling model. This class of probability measures, due to Pitman (1996), is defined as

$$P = \sum_{i \geq 1} P_i \delta_{Y_i} + \left(1 - \sum_{i \geq 1} P_i\right)H$$

where $0 < P_i < 1$ are random weights such that $\sum_{i \geq 1} P_i \leq 1$, independent of the locations $Y_i$, which are iid with some non-atomic distribution $H$ (on a general Polish space). Moreover, the N-IG process $G$ is an example of Poisson-Kingman models, where $G$ has representation (1) with $\sum P_i = 1$ a.s. and the weights $(P_i)_i$ are normalized ranked points of a Poisson process on the positive reals (see Pitman 2003). Therefore the N-IG process selects discrete distributions with probability one. The peculiarities of the N-IG and Dirichlet processes compared with other members of these classes (and, indeed, within all random probability measures) is represented by the fact that their finite-dimensional distributions are known explicitly. What distinguishes the Dirichlet process from the other processes in the class of normalized random measures with independent increments and species sampling models (and thus also from the N-IG process) is its conjugacy, as shown in James, Lijoi and Prünster (2006). Anyhow, this is no longer a problem, given the availability of suitable sampling schemes. However, a posterior characterization of the N-IG process, in term of a latent variable, can be deduced from the work of James (2002).

### 4 Nonparametric hierarchical mixtures in the AFT model

Let $T_1, \ldots, T_n$ be the survival times of $n$ subjects, and let $x_i = (x_{i1}, \ldots, x_{ip})'$ be the covariate vector associated with (observed or censored) $t_i$, $i = 1, \ldots, n$. The model we consider can be hierarchically expressed as follows

$$T_i = e^{-x'_i \beta} \cdot V_i, \ i = 1, \ldots, n,$$
where \( k(\cdot; \theta_i) \) is a family of densities on \( \mathbb{R}^+ \), depending on a vector of parameters \( \theta_i \) belonging to a Borel subset \( \Theta \) of \( \mathbb{R}^s \), and \( q \) is the prior distribution on the random d.f. \( G \); \( G_0 \) is a d.f. on \( \Theta \), expressing the “mean” of \( G \). Moreover, we assume that

\[
\beta \perp G, \quad \beta \sim \pi(\beta).
\]

The Bayesian model specification is completed assuming that \( G_0 \) depends on \( s \) hyperparameters \( \gamma_1, \ldots, \gamma_s \) (possibly random and distributed according to \( \pi(\gamma_1, \ldots, \gamma_s) \)). In the paper, we will assume that \( G \) has a Dirichlet process prior, or an N-IG process prior, more precisely \( G \sim DP(aG_0(\cdot)) \) or \( G \sim N-IG(M, G_0(\cdot)) \), \( a \) and \( M \) being positive parameters.

Observe that, if \( G \) is an N-IG process with diffuse mean \( G_0 \), then the predictive distributions of \( \theta_{n+1} \), given \( \theta_1, \ldots, \theta_n \), resulting from (3) are

\[
P(\theta_{n+1} \in A | \theta_1, \ldots, \theta_n) = \frac{\sum_{j=1}^{k} (n_j - \frac{1}{2}) \delta_{\theta_j}(A)}{\sum_{j=1}^{k} (n_j - \frac{1}{2})}, \quad A \in \mathcal{B}(\Theta),
\]

where \( \theta_1^*, \ldots, \theta_k^* \) (\( n_1^*, \ldots, n_k^* \), \( \sum_j n_j = n \)) denote the \( k \) distinct observations (multiplicities) within \( \theta_1, \ldots, \theta_n \) and

\[
w_{0,n}(k) = \frac{\sum_{r=0}^{n} \binom{n}{r} (-M^2)^{-r+1} \Gamma(k + 1 + 2r - 2n; M)}{2n \sum_{r=0}^{n-1} \binom{n-1}{r} (-M^2)^{-r} \Gamma(k + 2 + 2r - 2n; M)},
\]

\[
w_{1,n}(k) = \frac{\sum_{r=0}^{n-1} \binom{n-1}{r} (-M^2)^{-r+1} \Gamma(k + 2 + 2r - 2n; M)}{n \sum_{r=0}^{n-1} \binom{n-1}{r} (-M^2)^{-r} \Gamma(k + 2 + 2r - 2n; M)}.
\]

The N-IG process “prediction mechanism” is quite interesting and exploits the available information about \( k \); indeed, given a sample \( (\theta_1, \ldots, \theta_n) \) from a N-IG process, the next observation \( \theta_{n+1} \) is different from the previous ones with probability \( w_{0,n}(k) \) and coincides with an old observation with probability \( (n - k/2)w_{1,n}(k) \). A comparison with the predictive distributions from a Dirichlet process, where \( P(\theta_{n+1} \in A | \theta_1, \ldots, \theta_n) = a/(a + n)G_0(A) + 1/(a + n) \sum_{j=1}^{k} n_j \delta_{\theta_j}(A) \), reveals that the probability allocated to a new observation and to the old observations by the N-IG process depend on the number of different observations in the sample. For instance, Figure 1 shows that, for a relatively small value of \( k \) (between 5 and 10 for “small” total masses and \( \approx 20 \) for “big” total masses) the weight the N-IG process assigns to the prior guess \( G_0 \) is smaller than that assigned by the Dirichlet process with a sample size equal to 100; an opposite behaviour is shown when \( k \) increases. The N-IG prediction
rule takes the number of distinct observation \( k \) into account; since \( w_{0,n}(k) \) is an increasing function of \( k \), the more distinct observations are present in the sample (i.e. not many ties), the higher the weight the N-IG assigns to the prior guess \( G_0 \) is. Moreover, the distribution of the number of distinct observations \( k \) in a sample of size \( n \) from the N-IG process is

\[
p_{N-IG}(k|n) = \binom{2n-k-1}{n-1} \frac{e^M(-M^2)^{n-1}}{2^{2n-k-1} \Gamma(k)} \sum_{r=0}^{n-1} \binom{n-1}{r} \frac{\Gamma(k + 2 + 2r - 2n; M)}{(-M^2)^r}.
\]

It should be pointed out that, for \( n \) fixed, the mean of distribution (7) has a lower bound, given by the limit as \( M \) goes to 0.

As far as the regression parameters are concerned, a flat prior distribution is often imposed upon the \( \beta \) regression parameters (Kuo and Mallick 1997, Hanson 2006), regardless of the form of the kernel density \( k(\cdot; \theta) \). We introduce instead the reparameterization \( \alpha_j = e^{\beta_j} \), \( j = 1, \ldots, p \), and assign independent gamma priors to the \( \alpha_j \)'s. In this way, with gamma kernel densities, the full conditional posterior distributions of the \( \alpha_j \)'s associated to binary covariates are still gamma.

### 4.1 A conjugate prior

As mentioned in the Introduction, we are going to consider hierarchical mixtures of gamma densities \( k(\cdot; \theta) \), \( \theta = (\vartheta_1, \vartheta_2) \), with mean \( \vartheta_1/\vartheta_2 \), and to choose the centering distribution \( G_0 \) on \( \mathbb{R}^+ \times \mathbb{R}^+ \) as the product of two exponential distributions, i.e. \( \vartheta_1 \) and \( \vartheta_2 \), under \( G_0 \), are independent, exponentially distributed with parameters \( \gamma_1 \) and \( \gamma_2 \), respectively. This choice for \( G_0 \) is computationally convenient, since, in this case, a closed-form expression for the marginal prior of \( V \) is available, which is needed in the implementation of the Gibbs sampling procedure (a conjugate hierarchical mixture model is obtained). Indeed, for model (2)-(4) under both Dirichlet process and N-IG priors, the marginal prior density of \( V \) is, for \( v > 0 \),

\[
f_V(v) = \int_{\Theta} k(v; \theta)G_0(d \theta) = \frac{\gamma_1 \gamma_2}{v(v + \gamma_2)(\gamma_1 + \log(\frac{v + \gamma_2}{v}))^2},
\]

with distribution function \( F_V(v) = \gamma_1 / \left( \gamma_1 + \log(1 + \frac{\gamma_2}{v}) \right) \), \( v > 0 \). This distribution has infinite mean, but information about the hyperparameters \( \gamma_1 \) and \( \gamma_2 \) can be derived fixing the median \( m \) of \( V \), thus resulting into a semiparametric \( m \)-median regression model, analogously as in Kottas and Gelfand (2001). Therefore, in the paper, we will assume \( \gamma_1 = \log(1 + \gamma_2/m) \). On the other hand, the hyperparameter \( \gamma_2 \) controls the dispersion of \( V \): the interquartile range of the marginal prior, as a function of \( m \) and \( \gamma_2 \), is

\[
\gamma_2 \left( (1 + \gamma_2/m)^{1/3} - 1 \right)^{-1} - ((1 + \gamma_2/m)^{3} - 1)^{-1}
\]
which, for fixed \( m \), increases with increasing \( \gamma_2 \). Observe that the 90% prior probability interval for \( V \) is
\[
\left[ \frac{\gamma_2}{(1 + \frac{2s}{m})^{19}} - 1, \frac{\gamma_2}{(1 + \frac{2s}{m})^{1/19}} - 1 \right].
\]

4.2 More flexible priors

It should be pointed out that assuming \( G_0 \) as the product of two exponential distributions is not an extremely flexible choice, since \( f_V \) in (8) is decreasing with an asymptote at 0 for any \( \gamma_1, \gamma_2 \), in such a way that it always puts an appreciable amount of mass near zero (even if \( \gamma_1 \neq \log(1 + \gamma_2/m) \)). Therefore, we let \( G_0 \) be the product of two independent gamma distributions, i.e. \( \vartheta_1 \) and \( \vartheta_2 \) under \( G_0 \) are independent gamma distributed with parameter \( (\omega_1, \gamma_1) \) and \( (\omega_2, \gamma_2) \) respectively. We could think of further generalizations (such as \( \vartheta_1 \) and \( \vartheta_2 \) not independent under \( G_0 \)), but we limited ourselves to a neighborhood of the error distribution (8), by modifying some features only (by which the existence of some moments can be guaranteed, for example). The new marginal prior for the variable \( V \) is, for \( v > 0 \),
\[
f_V(v) = \int_0^{+\infty} d\vartheta_1 \int_0^{+\infty} d\vartheta_2 \frac{\vartheta_1^{\omega_1} \vartheta_2^{\omega_2}}{\Gamma(\vartheta_1) \Gamma(\vartheta_2)} \frac{\vartheta_1^{\omega_1 - 1} e^{-\vartheta_1 v} \vartheta_2^{\omega_2} \Gamma(\omega_1) \Gamma(\omega_2) \vartheta_2^{\omega_2 - 1} e^{-\vartheta_2 v}}{\vartheta_1^{\omega_1} e^{-\vartheta_1 v} \vartheta_2^{\omega_2} \Gamma(\omega_1) \Gamma(\omega_2)} \vartheta_1^{\omega_1 - 1} e^{-\vartheta_1 (\gamma_1 + \log(1 + \gamma_2/v))} d\vartheta_1
\]
\[(9)\]
\[
= \frac{\gamma_1^\omega_1 \gamma_2^\omega_2}{\Gamma(\omega_1) \Gamma(\omega_2)} \frac{1}{v(\gamma_1 + \gamma_2)} \int_0^{+\infty} \frac{\Gamma(\vartheta_1 + \omega_2)}{\Gamma(\vartheta_1)} \vartheta_1^{\omega_1 - 1} e^{-\vartheta_1 (\gamma_1 + \log(1 + \gamma_2/v))} d\vartheta_1 \int_0^{+\infty} \frac{\Gamma(\vartheta_1 + \omega_2)}{\Gamma(\vartheta_1)} \vartheta_1^{\omega_1} e^{-\vartheta_1 (\gamma_1 + \log(1 + \gamma_2/v))} d\vartheta_1.
\]
where \( \Gamma(\cdot,s,r) \) is the density of a gamma distributed random variable with shape parameter \( s \) and rate parameter \( r \), and \( r(v) = \gamma_1 + \log(1 + \gamma_2/v) \). Of course, (8) can be recovered from (9) assuming \( \omega_1 = \omega_2 = 1 \). Anyhow, distribution (9) is more flexible than the error distribution in the conjugate case, it has an asymptote in zero, but admits a mode for \( \omega_2 > 1 \). In Figure 2 the graphics of \( f_V(\cdot) \) for some choices of the hyperparameters are depicted. Moreover, \( f_V \) also admits \( j \)-th moment for \( \omega_2 > j \). In this case, \( E_q(V^j) = E_q(E_q(V^j|\vartheta_1, \vartheta_2)) = E_q(\vartheta_1^j)E_q(1/\vartheta_2^j) \), and \( E_q(1/\vartheta_2^j) \) exists if and only if \( \omega_2 > j \). In particular
\[
E_q(V) = \frac{\omega_1 \gamma_2}{(\omega_2 - 1) \gamma_1}, \quad \omega_2 > 1.
\]

5 The algorithm

Bayesian inference is done through MCMC simulation. The algorithm for the AFT model is built upon one for the hierarchical mixture model, by adding a module for the updating of the regression parameters. The desired output of the algorithm for the hierarchical mixture model is an estimate of the predictive density of a new observation \( V_{n+1} \), given the data
This can be obtained as the ergodic average of the conditional densities of $V_{n+1}$, given $\theta^{(j)} = (\theta_1^{(j)}, \ldots, \theta_n^{(j)})$ and the data, where $\{\theta^{(j)}\}_{j \geq 1}$ is the output of a Markov chain having $f(\theta \mid data)$ as its equilibrium distribution:

\begin{equation}
\hat{f}_{V_{n+1}}(v \mid data) = \frac{1}{J} \sum_{j=1}^{J} f_{V_{n+1}}(v \mid data, \theta^{(j)}).
\end{equation}

The density $f_{V_{n+1}}(v \mid data, \theta)$ can be found by conditioning on $G$ first, and then integrating it out with respect to its distribution given $\theta$ and the data, which leaves us with the following closed form expressions:

\begin{equation}
f_{V_{n+1}}(v \mid data, \theta) = \frac{a}{a+n} f_V(v) + \frac{1}{a+n} \sum_{i=1}^{n} k(v; \theta_i)
\end{equation}

for the DPM model, and

\begin{equation}
f_{V_{n+1}}(v \mid data, \theta) = w_{0,n}(k) f_V(v) + w_{1,n}(k) \sum_{j=1}^{k} \left( n_j - \frac{1}{2} \right) k(v; \theta_j^*)
\end{equation}

for the N-IG mixture, $f_V(v)$ in (12) and (13) as given in (8) or in (9). In this latter case, numerical integration is needed.

### 5.1 MCMC for the conjugate model

To produce the Markov sequence $\{\theta^{(j)}\}_{j \geq 1}$ with $f(\theta \mid data)$ as its stationary distribution, we use a Polya urn Gibbs sampler scheme such as Algorithm 2 in Neal (2000), obtained from the combination of Escobar’s (1994) and McEachern’s (1994) algorithm. We describe the algorithm in some detail in order to understand clearly what the state space is and how the Gibbs updates allow its exploration.

Consider an augmentation of the state space from the space of $\theta$ vectors to that of $(c, \theta)$ vectors, where $c = (c_1, \ldots, c_n)$ is a vector of indexes such that $c_i = c_j$ if $\theta_i = \theta_j$. Thus $c_i \in \{1, \ldots, k\}$ (or any set of $k$ different labels) for all $i$’s, with any permutation of the labels consistent with $\theta$: for example, if $n = 3$, and $\theta_1 = \theta_2 \neq \theta_3$, then $c$ may be either $(1, 1, 2)$ or $(2, 2, 1)$. Given $j \in \{1, \ldots, k\}$, $\theta_j^*$ takes the value of any element in $\theta$ found in one of the positions occupied by $j$ within $c$.

Let us denote by $n_j^{(-i)}$ the multiplicity of $j$ within $c$ after removing $\theta_i$ from $\theta$, by $\theta_{-i}$ the vector $\theta$ without $\theta_i$, and by $c_{-i}$ the vector of indexes remaining in vector $c$ after removing $c_i$. The kernel of the Markov chain is obtained as the combination of two separate Gibbs kernels: 1) update pairs $(c_i, \theta_i)$, for $i = 1, \ldots, n$, drawing a sample from the full conditional of $(c_i, \theta_i)$, starting with $c_i$ given $c_{-i}$, $\theta_{-i}$ and the data and then sampling from the full conditional of
\(\theta_i; 2)\) block update the sub-vectors of \(\theta\) formed by elements with the same value (which is equivalent to updating the \(\theta_j^*\) values).

Then, for the DPM model, the conditional probability that a new \(c_i'\) takes the value \(j\) is

\[
\Pr(c_i' = j \mid c_{-i}, \theta_{-i}, \text{data}) = \begin{cases} \frac{k(v_i|\theta_j^*) n_j^{(-i)}}{\sum_j k(v_i|\theta_j^*) n_j^{(-i)} + a_F(v_i)} & \text{for } j \in c_{-i} \\ \frac{a_F(v_i)}{\sum_j k(v_i|\theta_j^*) n_j^{(-i)} + a_F(v_i)} & \text{for } j = k_{\text{new}} \end{cases}
\]

where \(k_{\text{new}}\) indicates that the value of \(\theta\) associated to the \(i\)-th observation will be a new one. A single step of the first Gibbs kernel is completed with the update of \(\theta_i\) to \(\theta_i'\) given all the other random quantities:

\[
\Pr(d\theta_i' \mid c_i', c_{-i}, \theta_{-i}, \text{data}) = \begin{cases} \delta_{\theta_i'}(d\theta_i') & \text{if } c_i' \in c_{-i} \\ b k(v_i \mid \theta_i') g_0(\theta_i') d\theta_i' & \text{if } c_i' = k_{\text{new}} \end{cases}
\]

where \(\delta_{\theta_i'}\) is the point mass at \(\theta_i'^*\), \(g_0\) is the density function of \(G_0\) and \(b\) is a normalizing constant. Hanson (2006) gives a method for the exact sampling from the density \(b k(v_i \mid \theta_i') g_0(\theta_i')\).

After this step, it is convenient to re-assign values from 1 to \(k'\) to the labels in \(c\), where \(k'\) may be equal to \(k - 1\), \(k\), or \(k + 1\), according to the value of \(n_{c_i}^{(-i)}\) and to the sampled value of \(c_i'\) (for example, if \(n_{c_i}^{(-i)} = 0\) and \(c_i' = k_{\text{new}}\), then \(k' = k\)).

The block-update of the second Gibbs kernel can be described as the update of the \(\theta_j^*\) values with \(j\) ranging from 1 to \(k'\), with full conditional proportional to

\[
g_0(\theta_j^*) \prod_{i : c_i = j} k(v_i; \theta_j^*),
\]

for which a Metropolis step is required in case \(n_{c_i} > 1\).

The sampling scheme remains the same for the N-IG mixture model, where only (13) changes to adjust for the different weights in the predictive distribution (5):

\[
\Pr(c_i' = j \mid c_{-i}, \theta_{-i}, \text{data}) = \begin{cases} \frac{w_{1,n-1}(k^{(-i)}) (n_j^{(-i)} - \frac{1}{2}) + k(v_i|\theta_j^*)}{w_{1,n-1}(k^{(-i)}) b_i + w_{0,n-1}(k^{(-i)}) f_{V}(v_i)} & \text{for } j \in c_{-i} \\ \frac{w_{0,n-1}(k^{(-i)}) f_{V}(v_i)}{w_{1,n-1}(k^{(-i)}) b_i + w_{0,n-1}(k^{(-i)}) f_{V}(v_i)} & \text{for } j = k_{\text{new}} \end{cases}
\]

As the weights depends on \(k\) in this case, we have denoted by \(k^{(-i)}\) the number of distinct labels in \(c_{-i}\), and

\[
b_i := \sum_{j \in c_{-i}} \left( n_j^{(-i)} - \frac{1}{2} \right) k(v_i; \theta_j^*)
\]

where the positive part removes the \(j\)-th term from the sum when \(n_j^{(-i)} = 0\).
The extension of the Gibbs sampler to the AFT model is obtained by letting
\[ v_i = e^{x_i' \beta} t_i = \left( \prod_{j=1}^{p} \alpha_j^{x_{ij}} \right) t_i \]
(with the reparameterization introduced in Section 3) and by updating \((c, \theta)\) as just described. The \(\alpha_j\) parameters corresponding to binary covariates have a full conditional gamma distribution, whereas a Metropolis step is required for the remaining regression parameters. The actual state of the Markov chain is thus \((c, \theta, \alpha)\).

In the survival analysis context, the predictive distribution of an individual with covariate value \(x\) is usually presented as the predictive survival function. Therefore, with data \(t_1, \ldots, t_n\), instead of (11) one calculates
\[ \hat{S}_{T_{n+1}}(t \mid x, data) = \frac{1}{J} \sum_{j=1}^{J} S_{V_{n+1}}(e^{x' \beta(j)} t \mid data, \theta^{(j)}). \]
The survival function \(S_{V_{n+1}}(\cdot \mid data, \theta)\) can be easily found by integrating (12) or (13).

When censored survival times are present, one can proceed by augmenting the unknown parameter space with the unobserved survival times, and enlarging \(\theta\) and \(c\) consequently. The unobserved survival times are sampled one at a time from their full conditional distributions truncated at their respective censoring points.

### 5.2 MCMC for the non-conjugate model

When the model is non-conjugate, we have the problem of calculating values \(f_V(v_i)\) in (16) and (13). In addition, exact sampling from the full conditional distribution of \(\theta_i'\) in (14) is not possible anymore. We should also consider that the \(v_i\) values change at each update of the \(\beta\) parameters in the Gibbs sampler for the AFT model.

A way round to this problem is to substitute \(G_0\) with the empirical distribution of a random sample of size \(m\) from \(G_0\) itself. We may justify this procedure formally as a data augmentation if we re-express the mechanism leading to the generation of \(\theta\) in the hierarchical model (3) by integrating \(G\) out and by introducing the new augmentation variables in the following way, in the case of the DPM model: for \(i = 1, \ldots, n\), let \(\tau^{(i)} = (\tau_1^{(i)}, \ldots, \tau_m^{(i)})\) be an iid sample of size \(m\) from \(G_0\) and let
\[ \theta_1 \mid \tau^{(1)} \sim \frac{1}{m} \sum_j \delta_{\tau_j^{(1)}}, \]
for \(i\) in 2, \ldots, \(n\),
\[ \theta_i \mid \theta_1, \ldots, \theta_{i-1}, \tau^{(i)} \sim \frac{a/m}{a+i-1} \sum_j \delta_{\tau_j^{(i)}} + \frac{i-1}{a+i-1} \sum_{j=1}^{i-1} \delta_{\theta_j}. \]
This second expression indicates that, conditionally on the values of $\theta_j$ for $j < i$, the distribution of $\theta_i$ does not depend on $\tau(i)$. This is just the Polya urn scheme where we have replaced a single value sampled from $G_0$ by a value chosen at random from a sample of size $m$ of $G_0$, which is the same. Therefore, the new augmented state space for the Gibbs sampler is now that of vectors $(\underline{\theta}, c, \tau^{(1)}, \ldots, \tau^{(n)})$.

The two Gibbs kernels are now organized as follows: 1) for every $i = 1, \ldots, n$, sample from the full conditionals of $(c_i, \theta_i)$ and of $\tau(i)$ sequentially; 2) block-update $\underline{\theta}$ through the update of the $\theta_j^*$ values. This is equivalent to Algorithm 8 of Neal (2000), with an explicit description of the state space. It can be shown that the sampling distributions defining the first kernel take the following form:

\begin{equation}
\text{Pr}(c' = j \mid \underline{c} - i, \underline{\theta} - i, \tau^{(1)}, \ldots, \tau^{(n)}, \text{data}) \propto \begin{cases} k(v_i \mid \theta^*_i)n^{(-i)}_j & \text{for } j \in \underline{c} - i \\ \frac{1}{m} \sum_{l=1}^{m} k(v_i \mid \tau^{(i)}_l) & \text{for } j = k_{\text{new}} \end{cases}
\end{equation}

\[
\text{Pr}(d\theta' \mid \text{rest}) \propto \begin{cases} \delta_{\theta'_i}(d\theta'_i) & \text{if } c'_i \in \underline{c} - i \\ \sum_{l=1}^{m} k(v_i \mid \tau^{(i)}_l)\delta_{\theta'_i}(d\theta'_i) & \text{if } c'_i = k_{\text{new}} \end{cases}
\]

\[
\text{Pr}(d\tau^{(i)} \mid \text{rest}) = \begin{cases} \prod_{l=1}^{m} G_0(d\tau^{(i)}_l) & \text{if } c'_i \in \underline{c} - i \\ \delta_{\tau^{(i)}_i}(d\tau^{(i)}_i) \prod_{l \neq i} G_0(d\tau^{(i)}_l) & \text{if } c'_i = k_{\text{new}} \end{cases}
\]

where $\tau^{(i)}_l$ is the element of $\tau^{(i)}$ that was assigned to $\theta'_i$. The block-update kernel coincides with the one given in equation (15) for the conjugate case, with the difference that a Metropolis step is always required.

As before, for the N-IG process mixture, only the step for the label update in equation (17) changes, with

\[w_{1,n-1}(k^{(-i)}) \left( n^{(-i)}_j - \frac{1}{2} \right)^+\]

replacing $n^{(-i)}_j$ and $w_{0,n-1}(k^{(-i)})$ replacing $a$.

6 Data illustrations

6.1 Simulated data for density estimation

We consider simulated data from a mixture of 3 gamma densities to perform density estimation. We generated a random sample of size 100 from the density

\[0.2 \cdot \text{gamma}(40, 20) + 0.6 \cdot \text{gamma}(6, 1) + 0.2 \cdot \text{gamma}(200, 20),\]
(the mean is 6 and the variance is 10.12) and computed the posterior density estimates from the Dirichlet and the N-IG mixtures of gammas (the no-covariate AFT model) in the conjugate case when $G_0$ is the product of two exponential distributions. We assumed $M = 0.01$ and $M = 5.39$, which correspond to prior means of the number of components under the N-IG prior equal to 11.37 and 30, respectively. The matching with the Dirichlet process prior is achieved when $a = 3.10$ and $a = 14.16$, respectively. We set the median $m$ of $V$ equal to 5.67 (i.e. the true median), 56.7 and 0.57, and the hyperparameter $\gamma_2$ in $G_0$ equal to 0.01, 1 and 10, corresponding to the values of the IQRs and the 90% prior probability intervals listed in Table 1. For each choice of the hyperparameters we ran several MCMC chains and observed that they stabilize after around 2000 iterations. Then we ran the chains for 50000 iterations keeping the values every 50 iterations to reduce autocorrelation to compute density estimates.

As a measure of the performances of the models, we computed the error, in the uniform metric (EUM), between the true distribution function and the predictive distributions (under both priors). Figure 3 displays the true density, the histogram from the simulated data, and the density estimates under the Dirichlet and N-IG process priors for some values of the hyperparameters; in each graph the two estimates are very similar. Table 2 presents the errors for the values of the hyperparameters as in Figure 3, confirming the impression gained from the graphs. However, according to Figure 4, the posterior mode of the number of clusters under the N-IG mixture (corresponding to the prior (7)) is closer than the posterior mode under the DPM model to the actual number of components for every choice of the hyperparameters.

To quantify the difference in estimating the true density for repeated samples and at a smaller sample size, we computed the mean error between the exact d.f. and the estimates. If $\hat{F}_T(\cdot|T_1, \ldots, T_n)$ denotes the Bayesian estimate of the true $F_T(\cdot)$, based on a data sample $(T_1, \ldots, T_n)$, and $\text{EUM}(T_1, \ldots, T_n) := \sup_t |\hat{F}_T(t|T_1, \ldots, T_n) - F_T(t)|$, we computed $E_{T_1, \ldots, T_n}(\text{EUM}(T_1, \ldots, T_n))$ generating 200 samples of size $n = 30$. The mean errors (and the corresponding standard deviations) of the estimates for some choices of the hyperparameters are presented in Table 3. The values obtained seem to confirm a substantial equivalence of the two nonparametric priors in density estimation.

Finally, we tested the two nonparametric priors in the nonconjugate case when $G_0$ is the product of two gamma distributions with parameters $(\omega_1, \gamma_1)$ and $(\omega_2, \gamma_2)$ respectively. We assumed $\omega_2 = 2$ which yields a prior marginal mean $E_q(V) = \omega_1 \gamma_2/\gamma_1$ for the error. In this case, we set the mean of $V$, instead of the median, equal to 6 (the true mean), or 0.6, 60, and fixed $\omega_1$ equal to 1, 3, 10 and 30, and $\gamma_2$ equal to 0.01, 0.1 and 1 (with $\gamma_1$ calculated from (10)). We ran several MCMC chains: the convergence was as fast as in the conjugate
algorithm. For the estimation we ran a long chain for each set of hyperparameters with a burn-in of 10000 and a thinning of 50 iterations, so that the predictive densities were based on an MCMC sample of size 1000. Figure 5 shows the true density, the histogram from the simulated data, and the density estimates under the Dirichlet and N-IG process priors for some values of the hyperparameters; the two estimates are still equivalent and very close to those in the conjugate case. This is confirmed by Table 4, where the EUMs are displayed for $\gamma_2 = 0.1$. However, the posterior distributions of the number of clusters under the N-IG process prior assign more probability to smaller values (see Figure 6), emphasizing, as for the conjugate case, its strength in finding clustering structures in the data. On the other hand, the flexibility of the nonconjugate “mean” prior $G_0$ described in Section 4.2 is manifest in Figure 7, where the asymptote of $f_V$ in 0 disappears in both density estimates as $\omega_1$ increases.

### 6.2 Dataset not involving censoring

We study a famous dataset, in Feigl and Zelen (1965), where the survival times (in weeks) after diagnosis of 33 patients suffering from leukemia are presented. For each patient, two covariates were recorded, the white blood cell (WBC) count and the test result on the AG factor (positive and negative) at the time of diagnosis. As pointed out in Cook and Weisberg (1982) for instance, this dataset is controversial probably for the presence of measurement errors for at least one patient, but we decided to use it as a test for comparing our models. We normalized the bivariate vectors of covariates, $x_i = (x_{i1}, x_{i2})'$, $x_{i1} = 1$ if AG positive and 0 if AG negative, $x_{i2} \in [0, 1]$ as far as WBC is concerned. We assumed $M = 0.01 (a = 2.1478)$ and $M = 10 (a = 16.3400)$, which correspond to prior means of the number of components in the mixture equal to 6.5108 and 18.3966, respectively. Figure 8 displays the estimates of the survival functions for 2 “new” patients (corresponding to covariates (1,0.5) and (0,0.5) respectively) when $m = 14.8480$ (a parametric estimate of the “true” median), and $\gamma_2 = 1$ and 100; as panels (a) and (b) show, there are no significant differences in the estimates when the prior mean $E(K_n)$ is small (approximately 6.5), while, on the other hand, when $E(K_n)$ is bigger ($\simeq 18.4$), the estimates differ more sensibly. The Bayesian estimates of $\alpha_1$ and $\alpha_2$, together with the 90% credible intervals, are presented in Table 5 for $m = 14.8480$ and different values of $\gamma_2$; for comparison, the maximum likelihood estimates for a parametric AFT model with gamma distributed error are $\alpha_1 = 0.3288 (0.128)$ and $\alpha_2 = 1.9702 (0.989)$, with standard errors between parentheses.

To establish a predictive fit measure for the model, we adopt a cross-validation approach via the predictive densities of each survival time, given all the others; see Gelfand, Dey, Chang (1992). If $\tilde{y}_j, j = 1, \ldots, n$, is the estimated (from the Gibbs sampling procedure) median of the distribution of $T_j \mid T^{(-j)}$, and $\tilde{s}_j$ is the estimated predictive median of $|T_j - \tilde{y}_j|$
(given $T^{(-i)}$), then $\frac{t_j - \tilde{y}_j}{\tilde{s}_j}$ represents a standardized residual. In this case, a predictive fit index can be defined as

$$I := \sum_{j=1}^{n} \left| \frac{t_j - \tilde{y}_j}{\tilde{s}_j} \right|.$$  

We find that, when $m = 14.8480$, $\gamma_2 = 1$ and $E(K_n) = 6.5108$, then $I_{NIG} = 108.97$ and $I_{DIR} = 103.06$, indicating a slightly better fit of the DPM model. Figure 9 shows the standardized residual plot, where cases no. 17, 31 and 33 seem outlying and jointly influential points (as noticed by Dunn and Smyth, 1996). Their removal produces in fact essentially equal predictive fit indexes. At any rate, since the number of observations is not large, the predictive powers of the DPM model and the N-IG mixture are comparable, as in the density estimation experiment.

### 6.3 Dataset involving censoring

As a second regression example, we consider survival times in thousands of days of small-cell lung cancer data patients with right censoring from Ying, Jung and Wei (1995), and studied also in Walker and Mallick (1999), Yang (1999), Kottas and Gelfand (2001), Hanson (2006). The standard therapy is to use a combination of etoposide (E) and cisplatin (P); however the optimal sequencing and administration schedule was not defined. The original dataset and a more complex study were originally presented in Maksymiuk et al. (1994). The data here consist of $n = 121$ survival times of patients with limited-stage small-cell lung cancer who were randomly assigned to two different regimens (Treatment $A$: P followed by E, administered to 62 patients, and Treatment $B$: E followed by P, administered to 59 patients); moreover 23 patients were administratively right-censored. In this case the covariates are $x_i = (x_{i1}, x_{i2})'$, with $x_{i1} = 0$ if patient $i$ was assigned to Treatment $A$, and $x_{i2}$ denoting the patient’s entry age. In Figure 10 the estimated survival distributions, under the two semiparametric Bayesian models considered here, for 2 patients, are shown, for some values of the hyperparameters. Observe that, for high values of $M$ and $a$ (i.e. stronger confidence in the prior than for small values of the “total masses”), both estimates of the survival function show a bump near the origin. This is imputable to the choice of the marginal prior of $V$, presenting a “heavy” asymptote in zero for each value of the hyperparameters. The Bayesian estimates of $\alpha_1$ and $\alpha_2$, together with the 90% credible intervals - for $m = 2.436$ and different values of $\gamma_2$ - are presented in Table 6; the two models provide very similar results, that actually agree with Bayesian estimates provided by the literature. We ran the MCMC chains for 100000 iterations, with a thinning of 100, and a burn-in period of 10000.

As before, we computed a predictive fit index adopting a cross-validation approach via
the predictive densities of each non-censored survival time, given all the others. In this case, keeping the same notation as before, a predictive fit index is

\[ I := \sum_{j \in S^*} \frac{|t_j - \tilde{y}_j|}{\tilde{s}_j}, \]

where \( S^* \) denotes the index set of non-censored data in the sample. For \( m = 2.436, \gamma_2 = 1 \) and \( E(K_n) = 12.41 \), the indexes of predictive fit are \( I_{N-IG} = 94.42 \) and \( I_{DIR} = 99.99 \), indicating a slightly better fit for the N-IG mixture; see Figure 11. Again the difference in predictive power between the two nonparametric alternatives does not seem decisive.

Finally, we adopted the “nonconjugate” baseline measure \( G_0 \) for the two priors to remove the artificial trend near the origin of the estimated survival functions. Figure 12 shows the estimates for a patient with covariate \((0, 36)\) for different values of \( \omega_1 \) (1, 3 and 10) and \( \omega_2 = 2, \gamma_2 = 1, E(K_n) = 41.9 \). In this way, the “a priori” information near the origin change along with \( \omega_1 \); we also remark that the difference between the two estimates seems to decrease as \( \omega_1 \) increases. On the other hand, we checked that the adoption of such more flexible priors has no substantial effect on the estimates of \( \alpha_1 \) and \( \alpha_2 \).

7 Comments

From a predictive point of view, the illustrative examples show that there is not a substantial difference between the DPM and N-IG process priors. However, in the density estimation examples, some peculiarities arise in the posterior distributions of the number of clusters the two priors use to recover the density. The N-IG process prior seems to be more effective in the estimation of the number of components in the mixture for most values of the hyperparameters. Indeed, in Argiento (2007) extensive experiments showed that the posterior of \( K_n \) under the N-IG prior is more robust with respect to \( M \), and more sensitive to the choice of the hyperparameters in some regions. On the other hand, the posterior under the Dirichlet prior is sensitive to the choice of the total mass. This is the reason why, in the DPM literature, this model is usually presented with one more level in the hierarchy, \( i.e. \) randomizing the total mass \( a \), thus assuming \( G \) to be a mixture of Dirichlet processes. Instead of adding more levels in the model hierarchy, it seems worthy of investigation to consider a parametric mixture with a (normalized) generalized gamma process as mixing measure, encompassing both Dirichlet and N-IG mixtures; see Lijoi, Mena, Prünster (2006). Generalized gamma processes depend on two real parameters, the “total mass” \( M \), considered in the N-IG process as well, and \( \sigma \in (0, 1) \) which controls the reinforcement in the predictive distributions, in addition to \( G_0 \). The evaluation of the effect of this new prior on Bayesian inferences for model (2)-(3) will be considered in a future investigation.
The calculation of the weights (6) requires multiple precision arithmetics, because of the presence of the sum of several incomplete gamma functions. Therefore we did all the computations and the MCMC simulations using R (R Development Core Team, 2006), but we used Maple for setting up a table with the necessary weights (which do not change during simulations). An alternative to Maple is the PARI C library (The PARI Group, 2006), which can be used both at initialization and at run time, because C subroutines can be loaded into R. Given the availability of these multiple precision computational tools, the calculation of the weights is not a serious concern.

8 Acknowledgments

We are indebted to Ramses Mena, who provided some computer code and precious suggestions. We would also like to thank Antonio Lijoi, Igor Prünster and Fabrizio Ruggeri for helpful discussions.

References


Figure 1: Weights, as a function of $k$, assigned to the prior $G_0$, appearing in the predictive distributions of a sample of size $n = 100$ from a Dirichlet process (red) and from a N-IG process (see (6)); here the “total masses” are $a = 3.10$, $M = 0.01$ (left) and $a = 14.16$, $M = 5.39$ (right).

Table 1: IQR and 90% probability interval for the marginal prior (8) of $V$.

<table>
<thead>
<tr>
<th></th>
<th>m</th>
<th>0.57</th>
<th>5.67</th>
<th>56.7</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.01</td>
<td>1.53</td>
<td>0.03,10.86</td>
<td>0.29,107.82</td>
<td>2.98,1077.47</td>
</tr>
<tr>
<td>1</td>
<td>2.43</td>
<td>4·10$^{-9}$, 18.20</td>
<td>0.05,116.48</td>
<td>2.54,1086.30</td>
</tr>
<tr>
<td>10</td>
<td>6.05</td>
<td>7·10$^{-24}$, 60.08</td>
<td>4·10$^{-8}$,181.95</td>
<td>0.48,1164.77</td>
</tr>
</tbody>
</table>

Table 2: Errors in the uniform metric for the simulated dataset of size 100 between the true and estimated distribution functions for the values of the hyperparameters as in Figure 3.

<table>
<thead>
<tr>
<th></th>
<th>M</th>
<th>a</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>0.01</td>
<td>5.39</td>
<td>3.10</td>
<td>14.16</td>
</tr>
<tr>
<td>5.67</td>
<td>0.048</td>
<td>0.047</td>
<td>0.034</td>
</tr>
<tr>
<td>56.7</td>
<td>0.033</td>
<td>0.045</td>
<td>0.077</td>
</tr>
</tbody>
</table>
Figure 2: Graphics of the error marginal prior (9) for some choices of the hyperparameters. (a): $\omega_1 = 1$, $\omega_2 = 1$, $\lambda_1 = 0.002$, $\lambda_2 = 0.01$; (b): $\omega_1 = 3$, $\omega_2 = 2$, $\lambda_1 = 1$, $\lambda_2 = 4$; (c): $\omega_1 = 4$, $\omega_2 = 4$, $\lambda_1 = 0.007$, $\lambda_2 = 0.04$; (d): $\omega_1 = 149$, $\omega_2 = 4$, $\lambda_1 = 1$, $\lambda_2 = 0.2$.

Table 3: Mean errors and corresponding standard deviations (in brackets) between the estimates and the true distribution for samples of size 30 from Example 1. The error is the distance, in the uniform metric, between distribution functions.

<table>
<thead>
<tr>
<th>$\gamma_2$</th>
<th>(E(K_n) = 6.2)</th>
<th>(E(K_n) = 14.5)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>N-IG</td>
<td>DIR</td>
</tr>
<tr>
<td>$m = 5.67$</td>
<td>0.01</td>
<td>0.126 (0.046)</td>
</tr>
<tr>
<td></td>
<td>1</td>
<td>0.109 (0.032)</td>
</tr>
<tr>
<td></td>
<td>10</td>
<td>0.114 (0.023)</td>
</tr>
<tr>
<td>$m = 56.7$</td>
<td>0.01</td>
<td>0.128 (0.042)</td>
</tr>
<tr>
<td></td>
<td>1</td>
<td>0.120 (0.040)</td>
</tr>
<tr>
<td></td>
<td>10</td>
<td>0.112 (0.024)</td>
</tr>
<tr>
<td>$m = 0.57$</td>
<td>0.01</td>
<td>0.118 (0.048)</td>
</tr>
<tr>
<td></td>
<td>1</td>
<td>0.105 (0.029)</td>
</tr>
<tr>
<td></td>
<td>10</td>
<td>0.133 (0.020)</td>
</tr>
</tbody>
</table>
Figure 3: Histogram from the simulated data and density estimates under the Dirichlet process prior (dotted red) and the N-IG prior (dashed blue) when $\gamma_2 = 0.01$. Panel (a): $a = 3.10$, $M = 0.01$, $m = 5.67$, Panel (b): $a = 14.16$, $M = 5.39$, $m = 5.67$, Panel (c): $a = 3.10$, $M = 0.01$, $m = 56.7$, Panel (d): $a = 14.16$, $M = 5.39$, $m = 56.7$. In each graph the solid (green) line denotes the true density.

Figure 4: Posterior distributions of the number of the clusters under the N-IG mixtures (solid blue) and the DPM model (dotted red) under the “conjugate” marginal error distribution (8) when $\gamma_2 = 0.1$. The values of the parameters are those used in Figure 3.
Figure 5: Histogram from the simulated data and density estimates under the Dirichlet process prior (dotted red) and the N-IG prior (dashed blue) under the “nonconjugate” marginal error distribution (9) when $\gamma_2 = 0.1$. Panel (a): $\omega_1 = 3$, $a = 3.10$, $M = 0.01$, $E(V) = 6$, Panel (b): $\omega_1 = 3$, $a = 14.16$, $M = 5.39$, $E(V) = 6$, Panel (c): $\omega_1 = 1$, $a = 3.10$, $M = 0.01$, $E(V) = 60$, Panel (d): $\omega_1 = 1$, $a = 14.16$, $M = 5.39$, $E(V) = 60$. In each graph the solid (green) line denotes the true density.
Figure 6: Posterior distributions of the number of the clusters under the N-IG mixtures (solid blue) and the DPM model (dotted red) under the “nonconjugate” marginal error distribution (9) when $\gamma_2 = 0.1$. The values of the parameters are those used in Figure 5.

Figure 7: Density estimates under the Dirichlet process prior (dotted red) and the N-IG prior (dashed blue) for the “nonconjugate” marginal error distribution (9) for different values of $\omega_1$ (1, 3 and 30). The values of remaining hyperparameters are $\gamma_2 = 0.1$, $a = 14.16$, $M = 5.39$, $E(V) = 6$. 
Table 4: Errors in the uniform metric for the simulated dataset of size 100 between the true and estimated distribution functions for the “nonconjugate” $G_0$ when $\gamma_2 = 0.1$.

<table>
<thead>
<tr>
<th>m</th>
<th></th>
<th>M</th>
<th>0.01</th>
<th>5.39</th>
<th>3.0981</th>
<th>14.1614</th>
</tr>
</thead>
<tbody>
<tr>
<td>6</td>
<td></td>
<td>0.048</td>
<td>0.049</td>
<td>0.051</td>
<td>0.062</td>
<td></td>
</tr>
<tr>
<td>60</td>
<td></td>
<td>0.049</td>
<td>0.065</td>
<td>0.041</td>
<td>0.074</td>
<td></td>
</tr>
</tbody>
</table>

Figure 8: Estimated survival functions under the Dirichlet process prior (dotted red) and the NIG prior (dashed blue) for 2 patients (covariate (1, 0.5) in the left column and covariate (0, 0.5) in the right column) from the leukemia dataset when $m = 14.8480$. Moreover $a = 2.1478$, $M = 0.01$, $\gamma_2 = 1$ in (a) and (b) and $a = 16.34$, $M = 10$, $\gamma_2 = 100$ in (c) and (d).
Table 5: Estimates of $\alpha_1$ and $\alpha_2$, with 90% probability credible intervals, for the leukemia dataset under the NIG mixture and DPM priors.

<table>
<thead>
<tr>
<th>$\gamma_2$</th>
<th>NIG-mixture prior</th>
<th>DPM prior</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$M=0.01$</td>
<td>$M=10$</td>
</tr>
<tr>
<td>1</td>
<td>$\hat{\alpha}_1 = 0.5156$ (0.3318; 0.8661)</td>
<td>$\hat{\alpha}_1 = 0.4859$ (0.2497; 0.8056)</td>
</tr>
<tr>
<td></td>
<td>$\hat{\alpha}_2 = 4.8386$ (1.4532; 15.7634)</td>
<td>$\hat{\alpha}_2 = 4.6024$ (1.3554; 16.0212)</td>
</tr>
<tr>
<td>10</td>
<td>$\hat{\alpha}_1 = 0.4052$ (0.1549; 0.7794)</td>
<td>$\hat{\alpha}_1 = 0.3984$ (0.1489; 0.7818)</td>
</tr>
<tr>
<td></td>
<td>$\hat{\alpha}_2 = 8.6771$ (0.9026; 38.1015)</td>
<td>$\hat{\alpha}_2 = 10.2115$ (1.1436; 34.2550)</td>
</tr>
<tr>
<td>100</td>
<td>$\hat{\alpha}_1 = 0.4305$ (0.1967; 0.7867)</td>
<td>$\hat{\alpha}_1 = 0.6581$ (0.2342; 1.3701)</td>
</tr>
<tr>
<td></td>
<td>$\hat{\alpha}_2 = 4.3092$ (0.9009; 10.3039)</td>
<td>$\hat{\alpha}_2 = 13.7599$ (1.6032; 46.6534)</td>
</tr>
</tbody>
</table>

Figure 9: Standardized residuals against the continuous covariate for the leukemia dataset under the N-IG mixture (blue) and DPM (red) priors, when $m = 14.8480$, $\gamma_2 = 1$ and $E(K_n) = 6.5108$. 
Figure 10: Estimated survival functions under the Dirichlet process prior (dotted red) and the N-IG prior (dashed blue) for 2 patients (covariate (1, 36) in the left column and covariate (0, 36) in the right column) from the small-cell lung cancer dataset when $m = 2.436$, $\gamma_2 = 1$ and $a = 3.271$, $M = 0.001$ (first row), $a = 22.317$, $M = 10$ (second row).
Table 6: Estimates of $\alpha_1$ and $\alpha_2$, with 90% probability credible intervals, for the small-cell lung cancer dataset under the N-IG mixture and DPM priors, when $m = 2.436$. The estimates in the columns correspond to a prior mean $E(K_n)$ of the number of clusters equal to 12.41, 19.31 and 41.93, respectively.

<table>
<thead>
<tr>
<th>$\gamma_2$</th>
<th>M=0.001</th>
<th>M=1</th>
<th>M=10</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.1</td>
<td>$\hat{\alpha}_1 = 1.149; (1.181;1.819)$</td>
<td>$\hat{\alpha}_1 = 1.421; (1.124;1.762)$</td>
<td>$\hat{\alpha}_1 = 1.409; (1.115;1.782)$</td>
</tr>
<tr>
<td></td>
<td>$\hat{\alpha}_2 = 1.011; (1.000;1.023)$</td>
<td>$\hat{\alpha}_2 = 1.017; (1.006;1.028)$</td>
<td>$\hat{\alpha}_2 = 1.018; (1.007;1.028)$</td>
</tr>
<tr>
<td></td>
<td>$\hat{\alpha}_1 = 1.533; (1.203;1.903)$</td>
<td>$\hat{\alpha}_1 = 1.529; (1.201;1.895)$</td>
<td>$\hat{\alpha}_1 = 1.522; (1.181;1.926)$</td>
</tr>
<tr>
<td></td>
<td>$\hat{\alpha}_2 = 1.015; (1.005;1.025)$</td>
<td>$\hat{\alpha}_2 = 1.015; (1.005;1.025)$</td>
<td>$\hat{\alpha}_2 = 1.015; (1.006;1.024)$</td>
</tr>
<tr>
<td>10</td>
<td>$\hat{\alpha}_1 = 1.521; (1.022;1.042)$</td>
<td>$\hat{\alpha}_1 = 1.531; (1.168;1.941)$</td>
<td>$\hat{\alpha}_1 = 1.516; (1.158;1.967)$</td>
</tr>
<tr>
<td></td>
<td>$\hat{\alpha}_2 = 1.032; (1.022;1.042)$</td>
<td>$\hat{\alpha}_2 = 1.031; (1.021;1.041)$</td>
<td>$\hat{\alpha}_2 = 1.029; (1.021;1.039)$</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>$\gamma_2$</th>
<th>M=0.001</th>
<th>M=1</th>
<th>M=10</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.1</td>
<td>$\hat{\alpha}_1 = 1.437; (1.149;1.772)$</td>
<td>$\hat{\alpha}_1 = 1.409; (1.123;1.753)$</td>
<td>$\hat{\alpha}_1 = 1.387; (1.096;1.761)$</td>
</tr>
<tr>
<td></td>
<td>$\hat{\alpha}_2 = 1.014; (1.003;1.025)$</td>
<td>$\hat{\alpha}_2 = 1.016; (1.005;1.028)$</td>
<td>$\hat{\alpha}_2 = 1.021; (1.013;1.029)$</td>
</tr>
<tr>
<td></td>
<td>$\hat{\alpha}_1 = 1.515; (1.174;1.887)$</td>
<td>$\hat{\alpha}_1 = 1.512; (1.192;1.869)$</td>
<td>$\hat{\alpha}_1 = 1.530; (1.156;1.984)$</td>
</tr>
<tr>
<td></td>
<td>$\hat{\alpha}_2 = 1.016; (1.007;1.025)$</td>
<td>$\hat{\alpha}_2 = 1.018; (1.009;1.027)$</td>
<td>$\hat{\alpha}_2 = 1.020; (1.012;1.028)$</td>
</tr>
<tr>
<td>10</td>
<td>$\hat{\alpha}_1 = 1.518; (1.150;1.947)$</td>
<td>$\hat{\alpha}_1 = 1.521; (1.128;1.943)$</td>
<td>$\hat{\alpha}_1 = 1.537; (1.123;2.049)$</td>
</tr>
<tr>
<td></td>
<td>$\hat{\alpha}_2 = 1.032; (1.022;1.043)$</td>
<td>$\hat{\alpha}_2 = 1.032; (1.023;1.041)$</td>
<td>$\hat{\alpha}_2 = 1.033; (1.025;1.042)$</td>
</tr>
</tbody>
</table>
Figure 11: Standardized residuals against the continuous covariate for the small-cell lung cancer dataset under the N-IG mixture (blue) and DPM (red) priors, when $m = 2.436$.

Figure 12: Estimated survival functions under the Dirichlet process prior (dotted red) and the N-IG prior (dashed blue) of a patient with covariate $(0, 36)$ when the marginal error distribution is as in (9) for different values of $\omega_1$ (1, 3 and 10 from the left to the right) and $\omega_2 = 2$, $\gamma_2 = 1$, $E(K_n) = 41.9$. 