UNIVERSITÀ COMMERCIALE "LUIGI BOCCONI" – Milano

Facoltà di Economia

Dottorato di Ricerca in Statistica

Ciclo XVII

Bayesian semiparametric inference for accelerated failure time models

Coordinatore: Chiar.mo Prof. Pietro Muliere

Tesi di

Raffaele Argiento

N. Matricola 901645 DT

UNIVERSITÀ COMMERCIALE "LUIGI BOCCONI"

ISTITUTO DI METODI QUANTITAVI

The thesis "**Bayesian semiparametric inference for accelerated failure time models**" by **Raffaele Argiento** is recommended for acceptance by the members of the delegated committee, as stated by the enclosed reports, in partial fulfilment of the requirements for the degree of **Dottore di Ricerca in Statistica**.

Dated: September 2006

Research Supervisor: **Dott. Antonio Pievatolo**

Internal Examiner: **Prof.ssa Sonia Petrone**

External Examiner: **Prof.ssa Alessandra Guglielmi**

UNIVERSITÀ COMMERCIALE “LUIGI BOCCONI”

Date: **September 2006**

Author : **Raffaele Argiento**

Title: **Bayesian semiparametric inference for accelerated failure time models**

Department: **Istituto di Metodi Quantitativi**

Permission is herewith granted to Università Commerciale “Luigi Bocconi” to circulate and to have copied for non-commercial purposes, at its discretion, the above title upon the request of individuals or institutions.

Signature of Author

Acknowledgments

First I would like to thank the Institute of Quantitative Methods of Bocconi University in Milan for giving me the chance to attend a fruitful PhD course, and the Milano Department of CNR-IMATI for supporting me in this last year.

I'm very glad to have worked with Alessandra, Fabrizio and Marco since they gave me precious advice and uncountable teachings, constantly supporting me with infectious excitement and interest. I could not have completed this work without their support.

I'm also very grateful to Prof. P. Muliere and S. Petrone, for the reliable help that they have given to me in these years.

A special thanks to all the people of CNR-IMATI: Alberto, Antonella, Banshee, Bruno, Carla, Elisa, Gabriella, Licia, Luciana, Renata, Sara, Simona and Tommaso. A friendly environment and cakes help in any work!

Finally, I'm in debt with my family and my friends: they always encourage me in the difficult moments of my life.

A Mamma, Papà, Nicola e Imma:

la mia famiglia.

E ancora proteggi la grazia del mio cuore
adesso e per quando tornerà l'incanto.
L'incanto di te...
di te vicino a me.
*(*Vinicio Capossela, *Ovunque Proteggi)*

I saw it written and I saw it say
Pink moon is on its way
And none of you stand so tall
Pink moon gonna get you all
It's a pink moon ...

*(*Nick Drake, *Pink Moon)*

Preface

This thesis is based on one year work at the Institute of Applied Mathematics and Information Technology of the C.N.R. (National Research Council) in Milan, under the supervision of Dr. Antonio Pievatolo, and in collaboration with Prof. Alessandra Guglielmi and Dr. Fabrizio Ruggeri.

It concerns the study of the *accelerated failure time* (AFT) model in the Bayesian nonparametric setting. In the survival literature, the AFT model is usually meant as the multiplicative effect of a fixed p-vector of covariates $\mathbf{x} = (x_1, \ldots, x_p)'$ on the failure time T, i.e.,

$$T = \mathrm{e}^{-\mathbf{x}'\beta} \cdot V, \tag{1}$$

where $\beta = (\beta_1, \ldots, \beta_p)$ is the vector of regression parameters and V denotes the error.

The error V is usually assumed distributed from a parametric family, but often it is hard to justify a specific choice. Therefore, we take a *nonparametric* approach to the distribution of the error term. Recently this model has received much attention in the Bayesian community, in particular in papers where the error, V or $W = \log(V)$, has been represented hierarchically as a mixture of parametric densities with a Dirichlet process as mixing measure (i.e., the well-known DPM models, introduced by Lo, 1984). Moreover, Lijoi, Mena and Pruñster (2006) introduced the N-IG prior, that could represent a valid alternative to the Dirichlet prior in the contest of mixture modelling.

Therefore, we consider the error V in the AFT model as a mixture of some parametric family of densities on the positive reals, mixed by a random distribution function G. The work mainly focuses on the performances of two hierarchical mixture models by comparing Bayesian inferences on the regression parameters and the 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), thus defining what we call *N-IG mixtures* for short.

Our approach to the comparison of such models is a computational one. We match the nonparametric priors in such a way they carry the same prior information, then we measure the performance of the two models on both simulated and real data. We can summarize this approach in two steps. In the first one, we determine the hyperparameters based on the marginal distribution of the error, and we conduct some sensitivity analysis on the posterior estimates. In the second one, we obtain the predictive estimates and measure the predictive power of the two proposed methods. In the simulated data case, we consider the distance in the uniform metric between the predictive and the "target" distribution. In the real dataset case, we use a "cross validation" method, quantifying, in practice, how far the predictions are from the observed data.

Another important feature of nonparametric mixtures is related to the number of components in the mixture. Indeed, such prior specification is a fruitful extension of parametric finite mixture models. In the nonparametric way, the prior number of components is random and its law (in a sample of fixed size n) is determined by the mixing process. The N-IG prior leads to a less informative prior on the number of components with respect to the Dirichlet prior. We analyse the differences in the posterior estimates of this distribution, arising under the two different prior specifications.

Sometimes, in the Bayesian literature the AFT model is rewritten as $\log T = -\mathbf{x}\beta + W$, where $W = \log V$, then a nonparametric hierarchical mixture of parametric densities, with support on the entire real line, is used to model the distribution of W. However if, for example, W is a non-

parametric mixture of normal densities, then V is not a nonparametric mixture of $\log -$normals, so that a one to one correspondence between the additive and the multiplicative model can not be easily obtained. Therefore, it is equally reasonable to work on V rather than on W, with the advantage that the survival time T is modelled directly, thus facilitating also prior specification.

The plan of the Thesis is as follows.

In Chapter 1 we describe the two basic regression models for survival time data: the proportional hazard model and the AFT model. Then, we briefly review the literature on the nonparametric Bayesian approach to inference for AFT models.

In Chapter 2 we discuss the basic Bayesian nonparametric models that will be used as prior on the unknown distribution of V. This development begins with the Dirichlet processes (Ferguson, 1973), the mixture of Dirichlet process (Antoniak, 1974) and the Dirichlet process mixture models (Lo, 1984). Then we will illustrate the N-IG process and the N-IG process mixture.

A Monte Carlo approach to approximating the posterior distribution would involve sampling the infinite dimensional parameter G. Such an approach cannot be implemented without introducing a finite approximation. In Chapter 3 we will illustrate the basic idea of Escobar (1994) who first considered the DPM model obtained after marginalizing the Dirichlet process. Then we will describe some extensions to the Escobar algorithm, and we will adapt these to the N-IG mixture model.

In Chapter 4 the two competing models are tested on real and simulated datasets. We present four examples. In the first one we consider a simulated dataset and we perform density estimation through an AFT model without covariates. The predictive performances are quantified by computing the distance in the uniform metric between the true density and the predictive estimates. In the two subsequent examples we study two well known datasets, one containing censored observation, the other not. The predictive performances of the two models are compared through a *cross-validation* method. In the fourth example we test a non-conjugate hierarchical mixture

model on the simulated data set and we compare the results with those arising from the first example.

Chapter 4 constitutes the original part of this thesis. The AFT model with N-IG process mixture modelling the error has not been considered before. We also examine in some detail the effect of the choice of the prior mean of the N-IG and Dirichlet process on the marginal distribution of V. Finally, while a very large number of proposals have appeared in the literature, there have not been many attempts to compare competing models systematically.

Contents

Chapter 1

Regression models for survival time data

1.1 Introduction

In this chapter our aim is to introduce some basic definitions and models of survival analysis, i.e., the procedures to analyze data arising as the time until an event occurs. Enormous progress has been achieved in this area in the 20th century. The field of survival analysis is very rich, since time-to-event data arises in many fields of study, including medicine, biology, engineering, public health, epidemiology, economics among the others. In such context events are generically referred to as failures (or deaths), since major areas of application are medical studies.

A complexity that frequently arises in trials having time-to-event endpoints is that a fraction of subjects remains without time to failure at the end of the study. For these elements it is only known that the true time-to-event exceeds the recognised time. We refer to such data as right censored. As Flemming and Lin (2000) asserts: "The necessity of obtaining methods of analysis that involve censoring is probably the most important reason for developing specialised models and procedure for failure time data."

The target of a survival study is to look at the dependence between failure time and some explanatory variables. In medical studies for example, in order to enable some evaluation on the benefit or the risk of treatment on the subjects under observations. Another problem is the estimation and the identification of the distribution of the failure times.

In Section 1.2 we describe the probability objects of interest in a statistical survival analysis. In Section 1.3 we report the classical way of accomplishing estimates of parameters in particular for the well-known Cox proportional hazard model and the accelerated failure time model, and we mention some comparison between these two celebrated models.

Since in our work we focus the attention on the Bayesian semiparametric approach to the accelerated failure time model, in section 1.4 we will give a review of the major works in this area

1.2 Survival analysis, basic definition

Let T be a non-negative absolutely continuous random variable on some measure space $(\Omega, \mathcal{F}, \mathbb{P})$, representing the failure time of an individual in a population. Let $f(\cdot)$ denote the probability density function of T with distribution function

$$F(t) := \mathbb{P}(T \leq t) = \int_0^t f(u)du, \;\; t > 0.$$

In survival analysis it is customary to work with the survival function representing the probability that the individual time-to-event is greater than t,

$$S(t) := 1 - F(t) = \mathbb{P}(T > t), \quad t > 0.$$

Of course, $S(\cdot)$ is a monotone decreasing left continuous function with $S(0) = 1$, and $S(+\infty) := \lim_{t\to+\infty} S(t) = 0$. The density function and the survival function are related by

$$f(t) = -\frac{dS}{dt}(t).$$

The hazard function is the instantaneous rate of failure upon the time t and is defined by

$$\lambda(t) := \lim_{\Delta t \downarrow 0} \frac{\mathbb{P}(t \le T < t + \Delta t | T \ge t)}{\Delta t} = \frac{f(t)}{S(t)}.$$

It uniquely specifies the distribution of T, since it obviously holds

$$\lambda(t) = -\frac{d \log S}{dt}(t).$$

Integrating both sides of this equality with the boundary condition $S(0) = 1$, we have

$$S(t) = \exp\left(-\int_0^t \lambda(u)du\right), \quad t > 0.$$

1.3 Parametric regression models

The earliest efforts in the development of the survival methodology were predominantly focused on the estimation of the hazard function $\lambda(\cdot)$, or equivalently on the survival function $S(\cdot)$. The *life table* technique is one of the oldest methods for analysing survival (or failure time) data (e.g., see Berkson and Gage, 1950; Cutler and Ederer, 1958; Gehan, 1969). This table can be thought of as an "enhanced" frequency distribution table. Kaplan and Meier (1958) proposed a famous estimate of $S(\cdot)$ through a nonparametric maximum likelihood approach. The Kaplan-Meier estimate consists of a step survival function, with value reduced by a multiplicative factor at the times of observed events. In practice the two estimators above are descriptive methods for estimating the

distribution of survival times from a sample.

To improve flexibility in estimation it is useful to make some parametric assumptions on the survival distribution. Modelling the survival time T parametrically, a variety of distributions on the positive reals $\mathbb{R}^+$ has been proposed. We mention, among the many, the exponential, Weibull, log-logistic, log-normal and gamma distributions. Under this model Bayesian and maximum likelihood (ML) methods are used for parameters estimation. Lawless (1982) provides a, frequentist, detailed presentation of parametric methods, while Ibrahim, Chen, and Sinha (2001) give a complete panorama on the Bayesian approach.

However, usually failure time may depend on explanatory variables (or covariates). Therefore it becomes of interest to consider generalisations of parametric models to take account of concomitant information on the individuals sampled. Consider a failure time $T > 0$ and suppose a vector $\mathbf{X} = (X_1, \ldots, X_p)' \in \mathbb{R}^p$ of explanatory variables (or covariates) has been observed (note that these covariates $\mathbf{X}$ can take a variety of functional forms, being dichotomous, discrete or continuous). One of the principal problem dealt with in the statistical analysis is that of modelling and determining the relationship between T and $\mathbf{X}$. The covariates can influence the survival time either acting on the hazard function or directly "accelerating" or "decelerating" the failure time.

1.3.1 The proportional hazard model

Regression models proposed for survival distribution generally involve the assumption of proportional hazard functions (Lehemann, 1953). A proportional hazards model possesses the property that different individuals have hazard functions that are proportional to one another, that is, $\lambda(\cdot|\mathbf{X}_1)/\lambda(\cdot|\mathbf{X}_2)$, the ratio of the hazard functions for two individuals with covariates $\mathbf{X}_1$ and $\mathbf{X}_2$, does not vary with time t. In this framework Cox (1972, 1975) introduced his celebrated model

$$\lambda(t|\mathbf{X} = \mathbf{x}) = \lambda_0(t) \exp(\mathbf{x}'\beta), \quad \text{for each } t \geq 0. \tag{1.1}$$

The function $\lambda_0(\cdot)$ can be considered as a baseline hazard function of an individual for which $X = 0$, $\beta = (\beta_1, \ldots, \beta_p)$ is a vector of regression parameters and $\mathbf{x}'$ means the transpose of $\mathbf{x}$.

The proportional hazard regression model (1.1) has been largely studied, mostly because Cox (1975), by the *partial likelihood* approach, provided methods to estimate the regression parameters β without hypothesis on λ_0; the hazard baseline is considered as a infinite-dimensional nuisance function. Furthermore, Andersen and Gill (1982), through martingale theory, provided an elegant asymptotic theory for the partial likelihood estimate, while Efron (1977) and Oakes (1977) studied the efficiency of the Cox estimates.

1.3.2 The accelerated failure time model

The proportional hazard model (1.1) specifies that the effect of the covariates $\mathbf{X}$ is to act multiplicatively on the hazard function: however, in this framework, it is not easy to interpret, for example, the estimates of regression parameters. A different way to specify how the covariates may influence the survival time T is the Accelerated Failure Time (AFT) model (Cox, 1972; Prentice, 1978) which specifies a log-linear relationship between time-to-event and covariates:

$$\log T = -\mathbf{X}'\beta + W, \tag{1.2}$$

where W is an error variable with support in $\mathbb{R}$, independent of $\mathbf{X}$. Exponentiation of (1.2) yields

$$T = \exp(-\mathbf{X}'\beta)V, \tag{1.3}$$

where $V = \exp(W) > 0$. This expression shows that the role of covariates $\mathbf{X}$ is to *accelerate* (*decelerate*) the time to failure.

If $\lambda_0(\cdot)$ is the hazard function of V, then the hazard function of T can be expressed through

$\lambda_0(\cdot)$ as

$$\lambda(t|\mathbf{X}=\mathbf{x}) = \lambda_0(te^{\mathbf{x}'\beta})e^{\mathbf{x}'\beta}. \tag{1.4}$$

The last identity shows the effect of the covariates on the hazard function in the AFT model, which is multiplicative both on t and on the hazard function. Then in such model one assumes the existence of a baseline hazard function and that the effect of the regression variables is to alter the rate at which a subject proceeds along the time axis.

If there is no censored observation at the time of the analysis, the AFT model can be handled as a generalised linear model (GLM) popularised by McCullag and Nelder (1989). Among the various extensions of the traditional linear model, AFT models and the method of least squares to accommodate censored data seems very appealing, simply because the model is well known, widely used, well understood and well tested, as Wei (1992) points out. Considering $\lambda_0(\cdot)$ as an infinite dimensional nuisance and using a U-statistic representation, Koul, Susarla, and Van Ryzin (1981) showed that their estimates of β are consistent and asymptotically normal under some regularity conditions. Following on the simple idea of using "synthetic data", several extensions of the method have appeared in the literature that use more efficient ways to obtain estimated responses (Lai, Ying, and Zheng, 1992; Zhou, 1992). These developments gave rise to a notable interest, but the lack of stability of estimators driven by them made these approaches not as widely used as the proportional hazard model.

1.3.3 Comparison of the regression models

The two classes of models specified by (1.1) and (1.2) are different, and the only overlap arises when $\lambda_0(\cdot)$ is the hazard function of a two parameters Weibull distributed random variable V, i.e

$$\lambda_0(t) = \lambda q(\lambda t)^{q-1}, \quad \lambda > 0,\ q > 0. \tag{1.5}$$

To see that, consider the subset of log-linear models in which the regression variable acts multiplicatively on the hazard function. Using subscripts 1 and 2 for the respective models, if we assume the same hazard function, we have

$$\lambda_1(t|\mathbf{X}=\mathbf{x}) = \lambda_{01}(t)\exp(\mathbf{x}'\beta_1) = \lambda_{02}(t\exp(\mathbf{x}'\beta_2))\exp(\mathbf{x}'\beta_2) = \lambda_2(t|\mathbf{X}=\mathbf{x})$$

for all $t \in \mathbb{R}^+$ and $\mathbf{x} \in \mathbb{R}^p$. Substituting $\mathbf{x} = 0$ we have $\lambda_{01}(\cdot) = \lambda_{02}(\cdot) = \lambda_0(\cdot)$; moreover if β_{11} and β_{21} are respectively the first component of β_1 and β_2 substitution of $\mathbf{x} = (-\log t/\beta_{21}, 0, \ldots, 0)$ gives

$$\lambda_0(t)t^{-\beta_{11}\beta_{21}^{-1}} = \lambda_0(1)t^{-1}.$$

Now if $q = \beta_{11}\beta_{12}^{-1}$ and $\lambda = \{\lambda_0(1)/q\}^{1/q}$ we obtain the Weibull model (1.5).

In any case, both models provide the necessary flexibility to model concrete problems, testified by the fact that they are largely used in classical survival analysis.

We point out that the Cox proportional hazard regression (PH) model and the associated partial likelihood theory of estimation was breakthrough in developing a flexible method of regression for censored data. The huge success of PH models testify to the many needs for this type of semiparametric regression models. However, observe that the structure of PH is quite different from the generalised linear model for regression, in that the link function is not specified via the mean but rather through the hazard function. On the other hand, the proportionality structure is interesting but it may be hard to interpret the regression coefficients. As Sir D. Cox himself once remarked (Reid, 1994): "Of course, another issue is the physical or substantive basis for the proportional hazards model. I think that's one of its weakness, that accelerated life models are in many ways more appealing because of their quite direct physical interpretation, particularly in an engineering context."

Survival models such as (1.1) and (1.2) are usually referred to as *parametric models* when the distribution of the failure time (or equivalently $\lambda_0(\cdot)$) is parametrically specified. The parametric

assumption, however, may be too restrictive in applications. Models in which no parametric hypotheses are assumed on the baseline hazard function are called *semiparametric models* because of the presence of the finite-dimensional vector of parameters β, and an infinite dimensional parameter $\lambda_0(\cdot)$. Analysis of parametric and semi-parametric survival models has been discussed in a frequentist perspective by Kalbfleisch and Prentice (1980), Lawless (1982), Cox and Oakes (1984), Anderson, borgan, Gill, and Keinding (1993). The Bayesian analysis of survival data is examined in Klein and Moeschberger (1997) and in depth by Ibrahim *et al.* (2001).

Nonparametric and semiparametric Bayesian methods have recently become quite popular in survival analysis, due to recent advances in computing technology and the development of efficient computational algorithms for implementing these methods. The literature on nonparametric Bayesian methods is widely large and the enormous number of references can not be listed here (see e.g., the references chapter of Ibrahim *et al.*, 2001). We mention that, for the Cox proportional hazard model, Bayesian modelling involves the specification of nonparametric prior processes for the baseline hazard λ_0 or the cumulative hazard $\int_0^t \lambda_0(u)du$. In particular is worth to mention the work of Dykstra and Laud (1981) that specify a *gamma process* prior on the hazard rate, and the work of Hjort (1990) that introduced the beta process as prior on the space of cumulative hazard functions, recently extended by De Blasi and Hjort (pear) to the case of *regression* models. Finally we cite Walker and Muliere (1997) who introduced the beta-stacy process as a generalization of the the Dirichlet process that is conjugate to the right censored observations. The property of conjugacy to right censored observations is also a feature of beta process; however, with the beta process the statistician is required to consider hazard rates and cumulative hazards when constructing the prior. The beta-stacy process requires only considerations on the distribution of the observations.

1.4 Semiparametric AFT model

We have already noticed that the β parameter in (1.1) and (1.2) explaining the relationship between the survival time and the covariates is usually the main object of inference. The unspecified function $\lambda_0(\cdot)$ can be treated as an nuisance parameter. Semiparametric models constitute an attempt to avoid restrictive parametric assumptions. As Oakes (1977) observes:"A practical motivation for consideration of semiparametric models is to avoid restrictive assumptions about secondary aspects of a problem while preserving a tight formulation for the features of primary concern." In the context of regression modelling, Gelfand (1999) noted that the objective of semiparametric modelling is "to enrich the class of standard parametric models by wandering nonparametrically near, in some sense, the standard class but retaining the linear structure."

Semiparametric approaches to the AFT model, in the frequentist realm, date back to the initial work of Buckley and James (1979). Bickel, Klaassen, Ritov, and Wellner (1993) provide a large-sample theory. Lin and Geyer (1992) develop computational methods using simulated annealing for rank regression procedures often used in semiparametric inference. More recent approaches include those by Ying, Jung, and Wei (1995) and Yang (1999). All these approaches are essentially fitting techniques focusing on the estimates of regression effects. In fact, although the latter two papers include the analysis of failure time data, there are no predictive survival curve or densities nor mention of how one might obtain these very common loci of inference. Moreover, these frequentist approaches are based on a generalisation of the least squares criterion, the least absolute deviation criterion, resulting in what is referred to as L_1 regression (see, e.g., Rosseeuw and Leroy 1987, for a fuller discussion on this topic). The computational difficulties of this method (for example the possibility of a non-unique solution) compared to the simplicity of the least square method may also explain its limited usage as do the inferential limitations with smaller sample size.

On the contrary, the Bayesian nonparametric approach is especially attractive in this regard,

because inference is exact and predictive power may be gained by assuming a centring parametric "baseline" form for the survival curve. The Bayesian literature on nonparametric methods has grown rapidly since the theoretical background for the construction of priors on function spaces was developed. We recall the pionering work of Freedman (1963), who introduced tail free and Dirichlet random measure and Dubins and Freedman (1965), Fabius (1964), Freedman (1965) and Ferguson (1973, 1974) that formalized and explored the notion of the Dirichlet processes. Moreover the development of Markov Chain Monte Carlo (MCMC) algorithms and the enormous progress in computer science provided a powerful tool to deal with non parametric Bayesian estimation; see Robert and Casella (2004) for a survey on this topic.

In a pioneering work on AFT model from the Bayesian view point, Christensen and Johnson (1988) model $V = \exp(W)$ as a random distribution according to a Dirichlet process, $G \in$ Dir(MG_0), where M is a positive real parameter and G_0 is a distribution on the positive reals. The Dirichlet process has the advantage that the parameters M and G_0 have an easy interpretation, indeed G_0 represents the prior belief about the mean of the distribution of V and M indicates the degree of concentration of the distribution of V around G_0. The larger M, the more concentrated G is around G_0. The discrete nature of the Dirichlet process, however, yields intractable computation of the posterior distribution.

To avoid the discreteness of the Dirichlet processes, Walker and Mallick (1999) proposed a Pólya tree distribution (Lavine, 1992; Mauldin, Sudderth, and Williams, 1992) as the prior for the unknown distribution of W in (1.2). Under some sufficient conditions, Pólya tree priors assign probability one to the set of continuous distributions; furthermore the conjugate nature of Pólya trees makes the analysis less complicated. Walker and Mallick (1999) constrained the random Pólya tree distribution to have median zero introducing a *median regression* model for (1.2).

1.4.1 Semiparametric Bayesian AFT mixture model

We observed that the proportional hazard model is ubiquitous in modelling survival data because of its tractability and flexibility. When the proportional hazard approach is untenable, a natural alternative is the AFT model. However, to date applications have been restricted primarily to parametric versions of the AFT model.

Parametric modelling has long dominated both classical and Bayesian inference work, in the AFT context. As mentioned earlier, such modelling is typically developed using generalised linear model within standard exponential families. Such families are limited, being unimodal with implicit mean-variance relationship. In looking beyond standard parametric families, one is naturally led to mixture models. Finite mixture distributions (Titteringon *et al.*, 1985) are flexible and also feasible to implement due to advances in simulation based model-fitting. See, for example, Diebolt and Robert (1994) and Richardson and J. (1997).

Paradoxically, rather than handling the very large number of parameters resulting from a finite mixture models with a large number of mixands, it may be easier to work with an infinite dimensional specification by assuming a random mixing distribution which is not restricted to a specified parametric family. Here, by *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). Besides, NPHM models provides a natural generalisation of existing parametric AFT models, bridging a gap between parametric and semiparametric approaches. If an experimenter has been fitting log-normal, log-logistic, gamma or Weibull AFT specification to their data, then fitting regression models with a corresponding NPHM model should be quite natural. Prior information used in parametric fit of a dataset (e.g, as in the prior specification of Bedrick, Christensen, and Johnson (2000)) may be immediately incorporated into the semiparametric extension.

In the context of NPHM, Ferguson (1983) and Lo (1984) used a Dirichlet process prior on

the mixing distribution, introducing the so-called Dirichlet processes mixture (DPM) model, and obtained expression for Bayesian estimates in density estimation, i.e., AFT without covariates. Escobar and West (1995) developed this idea further and provided an Markov Chain Monte Carlo algorithms for the computation of the posterior distributions of the parameters in a normal mixture model.

Kuo and Mallick (1997) propose a class of DPM models in the AFT setting with non-zero covariate. In what they called "MDPV" the regression error V of (1.3) has been represented hierarchically as a location mixture of normal kernels. They observe that, as a prior on V, DPM distribution smooths the Dirichlet process with a continuous known kernel with unknown mixing weights where prior belief can be incorporated. The smoothing in DPM eliminates the difficulty, due to Dirichlet processes having support on the class of the discrete densities, encountered by Christensen and Johnson (1988). Anyhow in this specification the marginal prior of V gives positive probability to the negative reals, and the authors handled this problem considering kernel variance small enough to avoid, at least computationally, this inconsistency.

In the framework of DPM models, Kottas and Gelfand (2001) and Gelfand and Kottas (2003) propose median regression approaches to the AFT model (1.2). They proposed a DPM of unimodal parametric densities and, also, a DPM of unimodal step-functions as priors on the error variable W. The first model generalises standard parametric families by considering a mixture of scale families and including a parameter for *skewness*. This model seems very useful for estimating regression effects and for survival analysis where it is known a priori that the error distribution is unimodal.

To allow for multi-modality with flexibility in skewness in the median regression AFT model, Hanson and Johnson (2002) introduce a mixture of Pólya Trees, centred about a 0-mean family of normal distributions, as a prior for the error term W . The model accommodates data-driven deviations from the parametric family, and uncertainty in this direction may be modelled a priori.

As more data are collected, they overwhelm the centring baseline family, and features such as multimodality will become apparent. Moreover, Hanson and Johnson provide a comparison of Bayesian semiparametric approaches between their own model and the Kuo-Mallik (1997) and Kottas-Gelfand (2001) ones.

In Ghosh and Ghosal (2006) the distribution of V is given as a scale mixture of Weibull distributions with Dirichlet process as a mixing measure. They not only give a semiparametric formulation of the AFT model, but develop an asymptotic justification of the model. Indeed, in their paper a discussion on the consistency of posterior distribution of the parameters is established.

In a recent paper, pointing out the inconsistency of the prior marginal of V in the "MDPV" model of Kuo and Mallick (1997), Hanson (2006) proposes as mixing measure a mixture of Dirichlet processes (Antoniak, 1974) in which kernels are gamma densities, mixed both over the scale and the shape parameters. Pointing out that any continuous density on $\mathbb{R}^+$ can be approximated arbitrarily closely by a countable weighted sum of gamma densities, Hanson notes that such mixture model can provide a highly flexible baseline, allowing, e.g, for multiple modes.

In our work we will consider the model (1.3), $T = \exp(-\mathbf{X}'\beta)V$, with V distributed according to a NPHM of some parametric family of densities on the positive reals, mixed by a random distribution function G on $\mathbb{R}^s$ (s is a positive integer). In particular we will focus on the performances of two hierarchical mixture models comparing Bayesian inferences on the regression parameters and the 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 normalised inverse-Gaussian prior (N-IG prior), as introduced in Lijoi, Mena, and Prünster (2005), thus defining what we call N-IG *mixture* for short. N-IG mixtures of normals have been studied in Lijoi et al. (2006), but no approach appears to exist that employs a N-IG mixture with kernel having support on $\mathbb{R}^+$ including a regression component. The N-IG prior, compared to the Dirichlet process prior, while preserving almost the same tractability, is characterised by a more elaborate clustering property.

Chapter 2

Nonparametric hierarchical mixture models

2.1 Bayesian Nonparametric Modelling

The term "nonparametric" is somewhat of a misnomer, since it literally connotes the absence of parameters, but is usually used to indicate models in which the goals of a data analysis include making inferences about functionals of an unknown probability measure $\mathbf{P}$, which are themselves parameters, regardless of whether the class of probability measures under consideration is quite broad (e.g., not indexed by parameters). Nonetheless, the spirit of the term "nonparametric" is to be free of restrictive, inappropriate or unrealistic constraints that are implied by particular parametric models. For example, it is often necessary to consider models that allow for unspecified multimodality, asymmetry and nonlinearity. This can be accomplished by considering a broad class of distributions and by making statistical inference within that context.

Bayesian nonparametric models are constructed on "large" space to provide support for more eventualities than are supported by a parametric model. Technically, (to many) the off-putting aspect of Bayesian nonparametric framework is the mathematical apparatus that is required for

specifying distributions on function spaces and for carrying through prior-to-posterior calculation. Nonparametric modelling begins with the specification of a broad class of models for the data at hands. Let $Y_1, Y_2, \ldots$ be a sequence of observations on a sample spaces $\mathcal{Y}$ endowed with its $\sigma-$field $\mathcal{B}$ (in our work we will assume $\mathcal{Y}$ as a Euclidean space with its Borel $\sigma-$field). We think at $Y_1, Y_2, \ldots$ as conditionally independent and identically distributed (i.i.d.) observation from some unknown probability measure P on $(\mathcal{Y}, \mathcal{B})$ (or equivalently some unknown distribution function G) from P itself. In a nonparametric framework, each element of the set of all probability measures on $\mathcal{Y}$ is a candidate to represent the "true law" P. In the Bayesian context, then, the basic modelling concerns on how to define a random element on the class of all probability measures on $\mathcal{Y}$. By a *random probability measure* $\mathbf{P}$ on $(\mathcal{Y}, \mathcal{B})$ we mean a random element on the space $\mathcal{P}(\mathcal{Y})$ (we will skip $\mathcal{Y}$ when it does not generate confusion) of all probability measures on $(\mathcal{Y}, \mathcal{B})$ when $\mathcal{P}(\mathcal{Y})$ is endowed with the $\sigma-$field of the Borel-sets generates from the weak convergence on $\mathcal{P}$ (see Billingsley, 1968 or Parthasarathy, 1967 for a complete exposition on this theory).

In a Bayesian nonparametric framework, the goal is to make inferences about functionals of $\mathbf{P}$, or possibly about the pdf corresponding to $\mathbf{P}$. We denote with π_P the law of $\mathbf{P}$, this is called *prior* distribution and, since it is a measure on a function space, it may be specified by describing a sampling scheme that generate random distributions function with desired properties or by describing the finite dimensional laws of the stochastic process $\mathbf{P}$. This latter approach is more intuitive, but non trivial propositions are needed to establish existence. Let $(\mathcal{Y}, \mathcal{B})$ be an Euclidean space with its Borel $\sigma-$field, and let $\Delta_{n-1} := \{(x_1, \ldots, x_n) \in \mathbb{R}^n : \sum x_i = 1\}$ be the $n-$dimensional unit simplex. Moreover let us denote, as usual, with the symbol $\Rightarrow$ the weak convergence of a sequence of probability measures.

Theorem 2.1.1 (Regazzini, 1996, 2001) *Let* $\Pi = \{P_{A_1,\ldots,A_n} : A_1, \ldots, A_n \in \mathcal{B}\}$ *be a system of finite dimensional distributions, such that* $P_{A_1,\ldots,A_n} : \Delta^{n-1} \to [0, 1]$ *for each* $n \geq 0$*. Suppose the followings hold.*

(a) *For any $n \geq 1$ and any finite permutation ξ of $(1, \ldots, n)$,*

$$P_{A_1,\ldots,A_n}(C) = P_{A_{\xi(1)},\ldots,A_{\xi(n)}}(C_\xi), \quad \textit{for each } \; C \in \Delta_{n-1}$$

where $C_\xi = \{(x_{\xi(1)}, \ldots, x_{\xi(n)}) : (x_1, \ldots, x_n) \in C\}$.

(b) *$P_{\mathcal{Y}} = \delta_1$, where δ_y represents the point mass at y.*

(c) *For any family of sets $\{A_1, \ldots, A_n\}$ in $\mathcal{B}$, let $\{D_1, \ldots, D_h\}$ be a measurable partition of $\mathcal{Y}$ such that it is finer than the partition generated by $\{A_1, \ldots, A_n\}$. Then, for any $C \in \Delta_{n-1}$,*

$$P_{A_1,\ldots,A_n}(C) = P_{D_1,\ldots,D_h}(C')$$

where

$$C' = \left\{ (x_1, \ldots, x_h) \in [0,1]^k : \left(\sum_{(1)} x_i, \ldots, \sum_{(n)} x_i \right) \in C \right\}$$

with $\sum_{(i)}$ meaning the sum over the index j such that $D_j \subset A_i$;

(d) *For any sequence $(A_n)_{n\geq 1}$ of sets in $\mathcal{B}$ such that $A_n \downarrow \emptyset$,*

$$P_{A_n} \Rightarrow \delta_0.$$

Hence, there exists a unique stochastic process (i.e., random probability measure) $\mathbf{P}$ admitting Π as its family of finite-dimensional distribution, i.e.

$$\mathcal{L}\big(\mathbf{P}(A_1), \ldots, \mathbf{P}(A_n)\big) = P_{A_1,\ldots,A_n} \quad \textit{for each } A_1, \ldots, A_n \in \mathcal{B}.$$

As an example of random probability measure we mention the Dirichlet process (Ferguson, 1973, 1974), that is one of the most used nonparametric prior in Bayesian nonparametric statistics, aris-

ing when the finite dimensional distributions are Dirichlet distributions.

Rather than constructing $\mathbf{P}$ directly via its finite dimensional laws, a non parametric prior can be specified, also, via the de Finetti (de Finetti, 1937) representation theorem, constructing a sequence of *exchangeable* variables.

A sequence of random variables $Y_1, Y_2, ..$ is exchangeable if for any *finite permutation* ξ on the index space $\{1, 2, \dots\}$ we have:

$$\mathcal{L}(Y_1, Y_2, ..) = \mathcal{L}(Y_{\xi(1)}, Y_{\xi(2)}, ..).$$

Theorem 2.1.2 (de Finetti, 1937) *A sequence $Y_1, Y_2, \dots$ of random variables on $(\mathcal{Y}, \mathcal{B})$ is exchangeable if and only if there exist a random probability measure $\mathbf{P}$ on $\mathcal{Y}$ with law denoted by $\pi_P(\cdot)$ such that, for each $n \geq 1$ and for each $A_1, \dots, A_n$ on $\mathcal{B}$, we have*

$$\mathbb{P}\left(Y_1 \in A_1, \dots, Y_n \in A_n\right) = \int \left\{ \prod_{i=1}^{n} P(A_i) \right\} d\pi_P(P).$$

Equivalently, the theorem states that we can obtain the distribution of $Y_1, Y_2, \dots$ choosing first $\mathbf{P} \sim \pi_P$ and then taking $Y_1, Y_2, \dots | \mathbf{P} \sim_{\text{iid}} \mathbf{P}$. In this framework the law π_P is referred to as *de Finetti measure* and given the joint distribution of $Y_1, Y_2, \dots$ this π_P is unique (Hewitt and Savage, 1955). The de Finetti measure of an exchangeable sequence $Y_1, Y_2, \dots$ has the meaning of the prior distribution for the unknown probability measure $\mathbf{P}$. It can be interpreted, in a Bayesian setting as the prior distribution, when the observations Y_i conditioned on some "parameter" $\mathbf{P}$ with prior distribution π_P, are i.i.d.

In the following, by (finite) sample $(Y_1, \dots, Y_n)$ from the random process $\mathbf{P}$ we mean the first n observation of an exchangeable sequence $Y_1, Y_2, \dots$ with de Finetti measure π_P.

There are several reasons why it is often convenient to consider the sequence $Y_1, Y_2, \dots$ directly, marginalizing over $\mathbf{P}$. First, since $\mathbf{P}$ is an infinite dimensional parameter so it can be

advantageous working in a finite dimensional framework, making much of the mathematics simpler. Secondly, interest is often in prediction and the distribution of Y_{n+1} given $Y_1, \ldots, Y_n$ is an immediate consequence. Thirdly, we are "closer" to the data in the sense that we consider the probability distribution for the data explicitly. Also the posterior parameters of π_P (like the posterior mean or variance) can often be determined from the sequence of predictive distributions (consider, for example, the Pòlya urn sequence in section 2.2.1)

2.2 Dirichlet Process

Let $\mathbf{a} > 0$ be a real number and P_0 a probability measure (or equivalently, G_0 a distribution function) on the Euclidean measurable space $(\mathcal{Y}, \mathcal{B})$. A Dirichlet process on $(\mathcal{Y}, \mathcal{B})$ with parameter $(\mathbf{a}P_0)$ is a random probability measure $\mathbf{P}$, such that for each measurable finite partition $\{A_1, \cdots, A_n\}$ of $\mathcal{Y}$, the joint distribution of the vector $(\mathbf{P}(A_1), \ldots, \mathbf{P}(A_n))$ has Dirichlet distribution with parameters $(\mathbf{a}P_0(A_1), \ldots, \mathbf{a}P_0(A_n))$ on the $n-$dimensional unit simplex. From Theorem 2.1.1, under the consistency requirements (a)-(d), the distribution of $\mathbf{P}$ is uniquely defined by its finite dimensional distributions above. We shall denote the distribution of $\mathbf{P}$ by Dir$(\mathbf{a}P_0)$. The parameter $\mathbf{a}$ is called the precision or total mass, P_0 is called the centred measure or the "mean" distribution, and the product $\mathbf{a}P_0$ is called the base measure of the Dirichlet process. To justify this terminology, we observe that, for each $B \in \mathcal{B}$, the random variable $\mathbf{P}(B)$ is distributed as a beta r.v. with parameter $\mathbf{a}P_0(B)$ and $\mathbf{a}\,(1 - P_0(B))$, so that

$$\mathbb{E}\,(\mathbf{P}(B)) = P_0(B)$$

and

$$\text{Var}\,(\mathbf{P}(B)) = \frac{P_0(B)(1 - P_0(B))}{1 + \mathbf{a}}$$

Therefore, if $\mathbf{a}$ is large, $\mathbf{P}$ is tightly concentrated about P_0. However, Sethuraman and Tiwari (1982) pointed out that there is no clear interpretation for the parameter $\mathbf{a}$. Not only it controls the variability of $\mathbf{P}$ around P_0, but it influences the smoothness (or discreetness) of the random distributions. For instance, as $\mathbf{a} \to 0$, $\mathbf{P}$ converges in distribution to a single atomic random measure. We observe also that, from the expression for the variance of $\mathbf{P}(B)$, it is not possible to specify Var($\mathbf{P}(B)$) arbitrarily, and that the shape is determined by P_0.

A key conjugacy result holds for the Dirichlet process. Ferguson (1973) showed that, given a sample from the Dirichlet process $(Y_1, \ldots, Y_n)$, the posterior distribution of $\mathbf{P}$ is $\mathbf{P}|Y_1, \ldots, Y_n \sim$ Dir($\mathbf{a}^* P_0^*$) where $\mathbf{a}^* = \mathbf{a} + n$ and P_0^* is defined as

$$P_0^*(\cdot) = \frac{\mathbf{a}}{\mathbf{a}+n} P_0(\cdot) + \frac{1}{\mathbf{a}+n} \sum_{i=1}^{n} \delta_{Y_i}(\cdot). \tag{2.1}$$

Thus the posterior mean of $\mathbf{P}$ is a linear combination of the prior guess P_0 and the empirical measure $\hat{P}_n = (1/n) \sum_{i=1}^{n} \delta_{Y_j}$.

2.2.1 Pólya Urn and "stick-breaking" representation of Dirichlet Processes

Blackwell and MacQueen (1973), using the de Finetti representation, introduced a very useful construction of the Dirichlet process extending the classical Pólya urn schemes.Let $\mathbf{a}$ be a positive real number and P_0 a probability measure on $(\mathcal{Y}, \mathcal{B})$. A sequence of random variables $Y_1, Y_2, \ldots$ on $\mathcal{Y}$ is called *Pólya sequence* with measure $\mathbf{a}P_0$ when

$$\mathbb{P}(Y_1 \in B) = P_0(B), \quad \text{for each } B \in \mathcal{B}$$

and, for $n \geq 1$, the distribution of Y_{n+1} conditioned on $Y_1, \ldots, Y_n$ is

$$\mathbb{P}(Y_{n+1} \in B | Y_1, \ldots, Y_n) = \frac{\mathbf{a}P_0(B) + \sum_{i=1} \delta_{Y_i}(B)}{\mathbf{a}+n} \tag{2.2}$$

We have,

Theorem 2.2.1 *(Blackwell and MacQueen, 1973) Let $Y_1, Y_2, \ldots$ be a Pólya sequence on $\mathcal{Y}$ with measure $\mathbf{a}P_0(\cdot)$. Then:*

1. *as n grows to infinity, $\mathbb{P}(Y_{n+1} \in \cdot\,|Y_1, \ldots, Y_n) \Rightarrow \mathbf{P}(\cdot)$*

2. *the distribution of $\mathbf{P}$ is a Dirichlet process with parameter $\mathbf{a}P_0$;*

3. *the Pólya sequence is exchangeable and its de Finetti measure is Dir$(\mathbf{a}P_0)$.*

This result gives a simple and concrete procedure for constructing an infinite exchangeable sequence of random variables with a Dirichlet measure as de Finetti measure. The distributions in left side of equations (2.2) are usually called, in the Bayesian context, *predictive* distributions of $\{Y_i\}$. This representation is extremely crucial for Markov Chain Monte Carlo sampling from a Dirichlet process; it also shows that ties are expected in a finite sample $Y_1, \ldots, Y_n$; moreover Blackwell (1973) showed that a random $\mathbf{P}$ following Dir$(\mathbf{a}P_0)$ is a.s. discrete.

Two characteristics are usually indicated as limitations of the Dirichlet process. First, as previously indicated, the support of the Dirichlet process distribution is the set of all discrete distribution. This can be also visualised from the constructive definition of $\mathbf{P}$ given by Sethuraman (1994):

$$\mathbf{P} = \sum_{j=1}^{\infty} \omega_j \delta_{\theta_j},$$

where, with $\kappa_i \sim_{\text{i.i.d}} \text{Beta}(1, \alpha)$, the ω_j's are defined as $\omega_1 = \kappa_1, \ldots, \omega_j = \kappa_j \prod_{r=1}^{j-1}(1 - \kappa_r), \ldots,$ and $\theta_j \sim_{\text{i.i.d}} P_0$. This is often referred to as the "stick-breaking" representation as the weights are defined in a way that the interval $[0, 1]$ (the stick) is successively broken up or partitioned into pieces. The second drawback of the Dirichlet process is that for any disjoint measurable sets B_1

and B_2, the correlation between $\mathbf{P}(B_1)$ and $\mathbf{P}(B_2)$ is negative,

$$\text{Cov}(\mathbf{P}(B_1), \mathbf{P}(B_2)) = -\frac{P_0(B_1)P_0(B_2)}{\mathbf{a}+1}.$$

This for ("small") adjacent sets violates a belief that these two probabilities should be positively correlated.

2.2.2 Mixture of Dirichlet processes

Centring the Dirichlet process on a fixed parametric distribution P_0 may be restrictive for some applied problem. Antoniak (1974) introduced a generalisation of the Dirichlet process that, in some sense, centres the process on a *family* of parametric distributions. The author introduced the so called mixture of Dirichlet processes (MDP), i.e. a random probability $\mathbf{P}$ distributed as a mixture of Dirichlet processes indexed by a parametric family of probabilities $\{P_\theta,\ \theta \in \Theta\}$,

$$\mathbf{P} \sim \int \text{Dir}(\mathbf{a}P_\theta)\pi(d\theta), \tag{2.3}$$

where the mixing distribution $\pi(\cdot)$ is a parametric prior on Θ.

In a Bayesian context, mixture models are essentially hierarchical models that date back to Lindley and Smith (1972) who consider parametric mixtures. In a hierarchical fashion the random variables $Y_1, \ldots, Y_n$ are a sample from the MDP process if

$$\begin{aligned} Y_1, \ldots, Y_n|\mathbf{P} &\sim_{iid} \mathbf{P} \\ \mathbf{P}|\mathbf{a}, \theta &\sim \text{Dir}(\mathbf{a}P_\theta) \\ \theta &\sim \pi(d\theta). \end{aligned}$$

Antoniak (1974) presented theoretical results for the MDP prior and also gave a number of ap-

plications. In particular he showed an important conjugacy result. Let $Y_1, \ldots, Y_n$ a sample from (2.3), then

$$\mathbf{P}|Y_1, \ldots, Y_n \sim \int \text{Dir}(\mathbf{a}^* P_\theta^*)\pi(d\theta|Y_1, \ldots, Y_n)$$

where $\mathbf{a}^*$ and $\mathbf{P}^*$ are defined as in expression (2.1). Moreover, if the family $\{P_\theta,\ \theta \in \Theta\}$ is absolutely continuous with densities $\{f(\cdot|\theta),\ \theta \in \Theta\}$, then the posterior mixing distribution is given by

$$\pi(d\theta|Y_1, \ldots, Y_n) \propto \left(\prod_{j=1}^{k} f(Y_j^*|\theta)\right) \pi(d\theta)$$

were Y_j^* are the distinct observations in the sample $Y_1, \ldots, Y_n$.

Finally we observe that the MDP prior can be specified such that it chooses absolutely continuous probability with probability one, thus overcoming the discreteness problem of the Dirichlet process. This advantage is compensated by the fact that posterior computation becomes very complex. We mention that the complexity arising in the posterior computation has been rescaled by the introduction of simulation procedures first developed by Escobar (1994).

2.3 Normalized Inverse-Gaussian Prior

In a recent paper Lijoi, Mena, and Prünster (2005), following the finite dimensional law specification of Theorem 2.1.1, introduce a new nonparametric prior. As pointed out in Section 2.2, the Dirichlet Process arises when the finite dimensional laws are assumed to be Dirichlet distributions. It is well known (see, e.g. Bilodeau and Brenner 1999) that given n independent gamma random variables $Z_i \sim \text{Gamma}(\bar{a}_i, 1)$, the Dirichlet distribution is defined as the distribution of the vector $(W_1, \ldots, W_n)$, where $W_i = Z_i / \sum_{j=1}^{n} Z_j$ for $i = 1, \ldots, n$. If $\bar{a}_i > 0$ for each i the vector

$(W_1, \ldots, W_n)$ has density on the $n-$dimensional unit simplex Δ_{n-1} given by:

$$f(w_1, \ldots, w_n) = \frac{\Gamma(\sum_{i=1}^n \bar{a}_i)}{\prod_{i=1}^n \Gamma(\bar{a}_i)} \left(\prod_{i=1}^n w_i^{\bar{a}_i - 1} \right)$$

Clearly, if $n = 2$, the latter reduces to beta density with parameter $(\bar{a}_1, \bar{a}_2)$. By substituting the gamma distribution with the inverse-Gaussian we obtain an analogous distribution on the $n-$dimensional simplex.

2.3.1 The normalized Inverse-Gaussian distribution

A positive absolutely continuous random variable Z has inverse-Gaussian distribution with shape parameter $\bar{M} \geq 0$ and scale parameter $\gamma > 0$, which we will denote $Z \sim \text{IG}(\bar{M}, \gamma)$, if its density is given by

$$f(z|\bar{M}, \gamma) = \frac{\bar{M}}{\sqrt{2\pi}} v^{-3/2} \exp\left\{ -\frac{1}{2} \left(\frac{\bar{M}^2}{v} + \gamma^2 v \right) + \gamma \bar{M} \right\}, \quad v \geq 0.$$

An exhaustive account of the inverse Gaussian distribution was provided by Seshadri (1993). Let $Z_1, .., Z_n$ be independent random variables distributed according to a IG$(\bar{M}_i, 1)$ distribution for each $i = 1, \ldots, n$ (γ_i=1 without loss of generality). Lijoi *et al.* (2005) define the normalized inverse-Gaussian (N-IG) distribution with parameter $(\bar{M}_1, \ldots, \bar{M}_1)$, denoted by N-IG$(\bar{M}_1, \ldots, \bar{M}_n)$, as the distribution of the random vector $(W_1, \ldots, W_n)$ where $W_i = Z_i / \sum_{j=1}^n Z_j$ for $i = 1, \ldots, n$. The following proposition provide the density function on Δ_{n-1} of a N-IG random vector.

Proposition 2.3.1 (Lijoi, Mena, and Prünster, 2005) *Suppose that the random vector* $(W_1, \ldots, W_n)$ *is N-IG*$(\bar{M}_1, \ldots, \bar{M}_1)$*, with* $\bar{M}_i > 0$ *for every* $i = 1, \ldots, n$*, then the vector* $(W_1, \ldots, W_n)$ *is absolutely continuous and has a density function on* Δ_{n-1} *given by:*

$$
\begin{aligned}
&f(w_1,\ldots,w_n|\bar{M}_1,\ldots,\bar{M}_n) = \\
&\quad = \frac{e^{\sum_{i=1}^n \bar{M}_i}\prod_{i=1}^n \bar{M}_i}{2^{n/2-1}\pi^{n/2}} \times \mathbf{K}_{-n/2}\left(\sqrt{\mathcal{A}_n(w_1,\ldots,w_n)}\right) \\
&\quad \times \left((w_1\ldots,w_n)^{3/2} \times \left(\mathcal{A}_n(w_1,\ldots,w_n)\right)^{n/4}\right)^{-1}
\end{aligned}
$$

were $\mathcal{A}_n(w_1,\ldots,w_n) = \sum_{i=1}^m \bar{M}_i^2/w_i$ *and* $\mathbf{K}$ *denotes the Bessel function of the third type.*

Let $M = \sum_{j=1}^n \bar{M}_j$ and $p_i = \bar{M}_i/M$ for every $i = 1,\ldots,n$, then

$$
\begin{aligned}
\mathbb{E}(W_i) &= p_i, \\
\text{Var}(W_i) &= p_i(1-p_i)M^2 e^M \Gamma(-2,M),
\end{aligned}
\tag{2.4}
$$

where $\Gamma(\cdot,\cdot)$ denotes the incomplete gamma function, defined for each $M > 0$ and $x \in \mathbb{R}$ as

$$
\Gamma(x,M) := \int_M^\infty t^{x-1}\mathrm{e}^{-t}dt.
$$

We observe that the moments of a N-IG random vector are quite similar to those of a Dirichlet random vector, the structure is the same and they differ just by a multiplicative constant.

2.3.2 The normalized inverse-Gaussian process

Let $M > 0$ be a positive number and P_0 a probability measure on the Euclidean measurable space $(\mathcal{Y},\mathcal{B})$. A N-IG process with parameter MP_0 is a random probability measure such that for each finite and measurable partition $\{A_1,\ldots,A_n\}$ of $\mathcal{Y}$, the joint distribution of the vector $(\mathbf{P}(A_1),\ldots,\mathbf{P}(A_n))$ on the $n-$dimensional unit simplex has a N-IG distribution with parameters $(MP_0(A_1),\ldots,MP_0(A_n))$. The existence of such a process is justified by Theorem 2.1.1.

The moments of an N-IG process with parameter MP_0 follow immediately from (2.4), indeed

for $B \in \mathcal{B}$, $\mathbf{P}$ is distributed according a N-IG distribution with parameters $MP_0(B)$ and $M(1 - P_0(B)$, thus

$$\mathbb{E}(\mathbf{P}(B)) = P_0(B),$$
$$\text{Var}(\mathbf{P}(B)) = P_0(B)(1 - P_0(B))M^2 e^M \Gamma(-2, M).$$

Then we obtain an interpretation of the parameter similar to that of the Dirichlet process. The process $\mathbf{P}$ is "centred" around P_0 and the scalar M is a precision parameter.

2.3.3 Properties of the N-IG process

First, we recall that Ferguson (1973) also proposed an alternative construction of the Dirichlet process as a normalized gamma process. The same can be done in this case by replacing the gamma process with an inverse-Gaussian process, that is, an increasing Lévy process, $\zeta := \{\zeta_t : t \geq 0\}$, which is uniquely characterised by its Lévy measure, $\upsilon(d\nu) = (2\pi\nu^3)^{-1/2} e^{\nu/2} d\nu$. As shown by Regazzini, Lijoi, and Prünster (2003), such a construction holds for any increasing additive process, giving rise to the class of random measures with independent increment (RMI). Through this representation using a result in James (2003), it is possible to show that the N-IG process selects discrete distributions with probability one.

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

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

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. We point out that 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 distribution are known explicitly. What distinguishes the Dirichlet process from the other processes in the class of normalized RMI's 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. It is worth noting, however, that a posterior characterisation of the N-IG process, in term of a latent variable, can be deduced from the work of James (2002).

Let $Y_1^*, \ldots, Y_n^*$ denote the k distinct observations within the sample $(Y_1, \ldots, Y_n)$ with $n_j > 0$ terms being equal to Y_j^*, for $j = 1, \ldots, k$ and $\sum_{j=1}^k n_j = n$. Then the predictive distribution corresponding to a N-IG process is given for each $B \in \mathcal{B}$ by

$$\mathbb{P}(Y_{n+1} \in B | Y_1, \ldots, Y_n) = w_{0,k}^n P_0(B) + w_{1,k}^n \sum_{j=1}^k (n_j - 1/2)\delta_{Y_j^*}(B), \tag{2.5}$$

with

$$w_{0,n}(k) = \frac{\sum_{r=0}^n \binom{n}{r}(-M)^{-r+1}\Gamma(k+1+2r-2n; M)}{2n\sum_{r=0}^{n-1}\binom{n-1}{r}(-M^2)^{-r}\Gamma(k+2+2r-2n; M)} \tag{2.6}$$

and

$$w_{1,n}(k) = \frac{\sum_{r=0}^n \binom{n}{r}(-M)^{-r+1}\Gamma(k+2r-2n; M)}{n\sum_{r=0}^{n-1}\binom{n-1}{r}(-M^2)^{-r}\Gamma(k+2+2r-2n; M)}. \tag{2.7}$$

Thus, similarly to the Dirichlet process, the predictive distributions are linear combinations of the prior guess P_0 and the weighted empirical distributions with explicit expression for the weights. Moreover, from Regazzini (1999) the predictive mechanism (2.5) leads to a generalized Pólya urn scheme for N-IG processes.

A comparison between the predictive distributions of the two models emphasises the distinctive feature of the N-IG process. In the Dirichlet prediction case we have, from (2.2), that Y_{n+1}

is different from previous observations with probability $\mathbf{a}/(\mathbf{a}+n)$ and that it coincides with one of the $(Y_1^*, \ldots, Y_k^*)$ with probability $n/(\mathbf{a}+n)$. Thus the probability allocated to previous observations does not depend on the number k of distinct observations within the sample. Moreover, the weight assigned to each Y_j^* is $n_j/(\mathbf{a}+n)$, and it only depends on the multiplicity of Y_j^*. As pointed out by Ferguson (1973) this is a characterising property of the Dirichlet process that at the same time represents one of its drawbacks. In contrast, the prediction mechanism (2.5) is quite interesting and exploits the available information about k. Given a sample $(Y_1, \ldots, Y_n)$ from a N-IG process, the next observation Y_{n+1} is different from the previous ones with probability $w_{0,n}(k)$ and coincides with an old observation with probability $(n - k/2)w_{1,n}(k)$. As we can see in figure 2.1, for a relatively small value of k (≈ 20 in the figure) the weight that the N-IG process assigns to the prior guess G_0 is smaller than that assigned by the Dirichlet process. An opposite behaviour is shown when k increases. The N-IG prediction 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 that the N-IG assigns to the prior guess.

Also the allocation of probability to each Y_j^* is more elaborate for the N-IG case than for the Dirichlet case. In figure 2.2 we depicted the weights assigned by the two processes to having a tie with a previous observation Y_j^* when n_j $(= 3, 5, 20)$, in a sample of size $n = 100$. Also these quantities, for the N-IG processes, increase with k. Moreover we can see how, for small values of $n_j(= 3)$, the N-IG prior tends to reinforce the observation less than the Dirichlet process, and the opposite behaviour is observed for big value of $n_j(= 20)$. This can be explained as follows: a small value of n_j suggests a weak statistical evidence of Y_j^*, particularly for small values of k. On the opposite way a big value of n_j indicates a strong statistical evidence of Y_j^*.

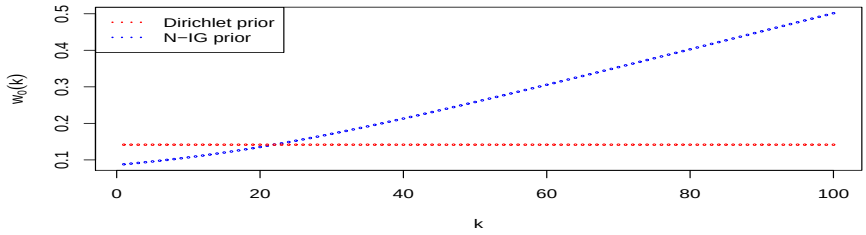

Figure 2.1: Weights, as a function of k, assigned to the prior G_0,appearing in the prediction rules by the Dirichlet process (2.2) and by the N-IG process (2.5) in a sample of size $n = 100$. The parameter are $\mathbf{a} = 14.16$ and $M = 5.39$

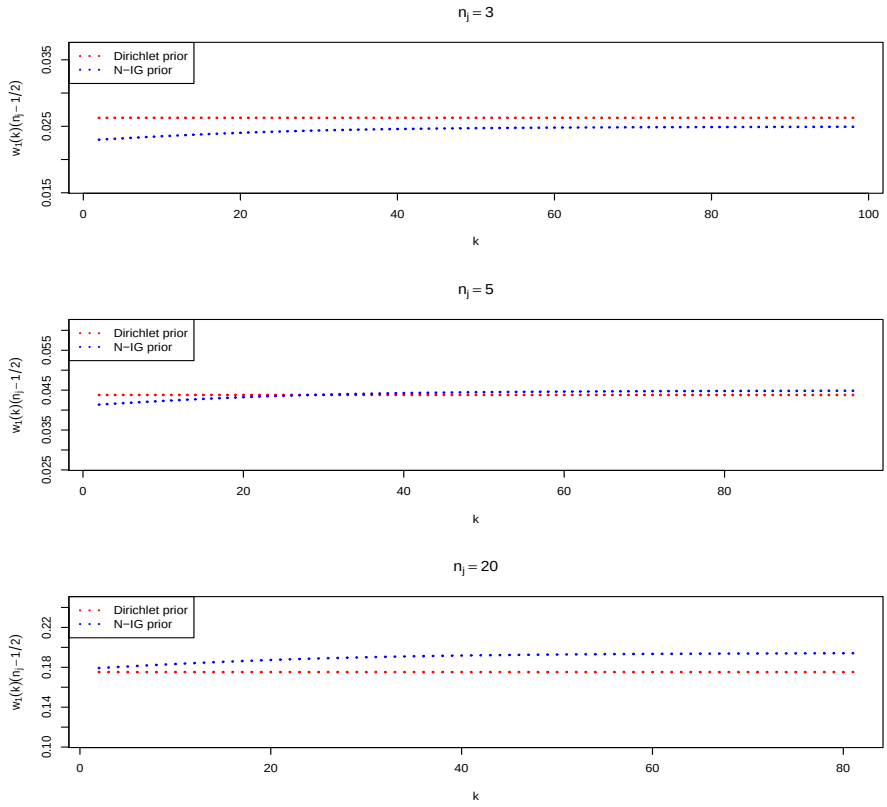

Figure 2.2: Weights, as a function of k, assigned to Y_j^* with multiplicities $n_j = 3, 5, 20$, appearing in the prediction rules by the Dirichlet process (2.2) and by the N-IG process (2.5), in a sample of size $n = 100$. The parameters are $\mathbf{a} = 14.16$ and $M = 5.39$

2.4 Nonparametric hierarchical mixture prior

A standard parametric model that strives to achieve flexibility is the finite parametric mixture model (Titteringon *et al.*, 1985)

$$Y_1, \ldots, Y_n | k, \mathbf{p}, \theta \sim_{iid} \sum_{j=1}^{k} k(\cdot|\theta_j) p_j,$$

where $\{k(\cdot, \theta) : \theta \in \Theta\}$ represents a standard parametric family, $\theta_j \in \Theta$ for $j = 1, \ldots, k$ are assumed to be distinct, so the mixture is comprised of k distinct members of this family, and $\mathbf{p} = (p_1, \ldots, p_k)$ is a fixed unknown discrete probability. Bayesian inference for this model is achieved by placing a prior distribution on $k,\ \mathbf{p} = (p_1, \ldots, p_k),$ and $\{\theta_j, j = 1, \ldots, k\}$. Such a model results in a varying dimensional parameter space and consequently specialized computational techniques, such reversible jump MCMC (Green, 1995), are required.

The nonparametric hierarchical mixture (NPHM) model avoids such concerns as the data are modelled according to an infinite mixture, which is given by

$$f_{\mathbf{P}}(y) = \int k(y; \theta) d\mathbf{P}(\theta) \tag{2.8}$$

where the random probability $\mathbf{P}$ is chosen according to a probability measure (prior) q on the space $\mathcal{P}$. An equivalent (and more used) specification of the nonparametric mixture model is a hierarchical model: the random variables $Y_1, \ldots, Y_n$ are a sample from a NPHM process if

$$\begin{aligned} &Y_i | \theta_i \sim k(\cdot; \theta_i), \quad i = 1, \ldots, n, \\ &\theta_1, \ldots, \theta_n | \mathbf{P} \sim_{iid} \mathbf{P}, \\ &\mathbf{P} \sim q(\cdot) \end{aligned} \tag{2.9}$$

Then, instead of the θ_i's being assumed to be i.i.d. from some parametric distribution (as with

standard Bayesian hierarchical models) greater flexibility is allowed via the introduction of the nonparametric prior $q(\cdot)$ usually centred on a parametric distribution. In our work we will consider q as a Dirichlet prior, obtaining the well known Dirichlet process mixture model, or a N-IG prior, obtaining what we call N-IG *mixture*.

2.4.1 Dirichlet process mixture model

If $\mathbf{P}$ in (2.8) is a Dirichlet process with parameter $\mathbf{a}P_0$, then we obtain the DPM model (Lo, 1984); this model has been largely used in recent Bayesian nonparametric modelling. It reaches a great level of flexibility, and inference can be obtained via those MCMC algorithms that have started to appear in statistical literature since the work of Escobar (1994).

An interesting property of the DPM is its posterior characterisation given by Antoniak (1974). If $Y_1, \ldots, Y_n$ is a sample from the DPM (2.9) with $q(\cdot)$ being a Dir$(\mathbf{a}P_0)$, then the posterior distribution of the mixing process $\mathbf{P}$ is the mixture of Dirichlet processes

$$\mathbf{P}|Y_1, \ldots, Y_n \sim \int \text{Dir}\,(\mathbf{a}^* P_0^*)\, d\pi(\theta_1, \ldots, \theta_n|Y_1, \ldots, Y_n), \tag{2.10}$$

where $\mathbf{a}^* = \mathbf{a} + n$, $P_0^* = \frac{\mathbf{a}}{\mathbf{a}+n} P_0 + \frac{1}{\mathbf{a}+n} \sum_{i=1}^n \delta_{\theta_i}$, and the mixing distribution π is the posterior law of the unobservable parameters $\theta_1, \ldots, \theta_n$, i.e.

$$d\pi(\theta_1, \ldots, \theta_n|Y_1, \ldots, Y_n) \propto \prod_{i=1}^n k(Y_i|\theta_i) \left(\mathbf{a}P_0 + \sum_{j=1}^{i-1} \delta_{\theta_j} \right) (d\theta_i). \tag{2.11}$$

Indeed, by (2.10) and (2.11) in the literature the term *mixture of Dirichlet processes* is often used to indicate both model (2.3) and the DPM model.

Let $\theta_1, \ldots, \theta_n$ be a sample from Dir$(\mathbf{a}G_0)$. The a.s. discreteness of the Dirichlet process ensures that, with positive probability, there may be $k \leq n$ distinct observation $\theta_1^*, \ldots, \theta_n^*$ in the sample. In practice the Dirichlet process, used as mixing distribution in the hierarchical model,

yields a prior specification on the number k of components in the mixture. Antoniak (1974) deduced this distribution conditional on the number of observations,

$$p(k|n) = c_n(k)\mathbf{a}^k \frac{\Gamma(\mathbf{a})}{\Gamma(\mathbf{a}+n)} \quad k \in \{1,\ldots,n\} \tag{2.12}$$

where $c_n(k)$ is the absolute value of a Stirling number of the first kind. From (2.12) it is clear that the real parameter $\mathbf{a}$ influences the prior on the number of components in the mixture specification. Large values of $\mathbf{a}$ give rise to models with a high prior number of components, small values of $\mathbf{a}$ yield priors very concentrated on relatively small values of k.

2.4.2 Normalized inverse-Gaussian mixture model

The most widely used NPHM model is the Dirichlet process mixture just described. A random discrete probability distribution, such as the Dirichlet process, exploited as a mixing measure in the model, is an essential tool for modelling the cluster behaviour. Indeed, the occurrence of ties at higher levels of hierarchy induces a clustering structure within the data. We wonder how a specific choice of the driving random discrete distribution affects the clustering mechanism. In this respect, it is worth mentioning that various new classes of discrete priors generalising the Dirichlet process have been introduced recently. Among them we recall species sampling models (Pitman, 1996), dependent Dirichlet processes (MacEachern, 1999), generalized stick-breaking prior (Ishwaran and James, 2001), normalized random measures with independent increments (Regazzini *et al.*, 2003).

We will focus on the case in which $\mathbf{P}$ is distributed according to a N-IG mixture as in the work of Lijoi *et al.* (2005). The cluster structure induced by this prior specification is quite interesting because of the particular reinforcement mechanism induced by the N-IG process discussed in Section 2.3.3. Let $\theta_1,\ldots,\theta_n$ be a sample from a N-IG(MG_0), then the distribution of the number

of distinct observations k in a sample of size n is given by

$$\begin{aligned} p(k|n) &= \binom{2n-k-1}{n-1} \frac{\mathrm{e}^M(-M^2)^{n-1}}{2^{2n-k-1}\Gamma(k)} \\ &\times \sum_{r=0}^{n-1} \binom{n-1}{r} (-M^2)^{-r} \\ &\times \Gamma(k+2+2r-2n; M) \quad k \in \{1,\ldots,n\}. \end{aligned} \tag{2.13}$$

As the Dirichlet case, a smaller total mass M yields a $p(\cdot|n)$ more concentrated on smaller values of k. This can be explained by the fact that a smaller M gives rise to a smaller $w_{0,n}$ (see (2.5)), that is, the process generates new data with lower probability. However, the $p(\cdot|n)$ induced by the N-IG prior is apparently less informative than that corresponding to the Dirichlet process prior and thus is more robust with respect to a change in the real parameter M (corresponding to $\mathbf{a}$, for the Dirichlet prior). A qualitative illustration is given in Figure 2.3, where the distribution of k given $n = 100$ observation is depicted for the Dirichlet and N-IG processes. The parameters $\mathbf{a}$ and M have been chosen in such a way to match the prior mean of k; we mention that for fixed n the mean of the number of clusters under the NI-G process has a lower bound as shown in Lijoi *et al.* (pear). We observe that the prior distribution $p(\cdot|n)$ under the N-IG prior is more dispersed with respect to the corresponding prior under the Dirichlet process, but notice that as M (or $\mathbf{a}$) grows, the differences between the two priors became less pronounced.

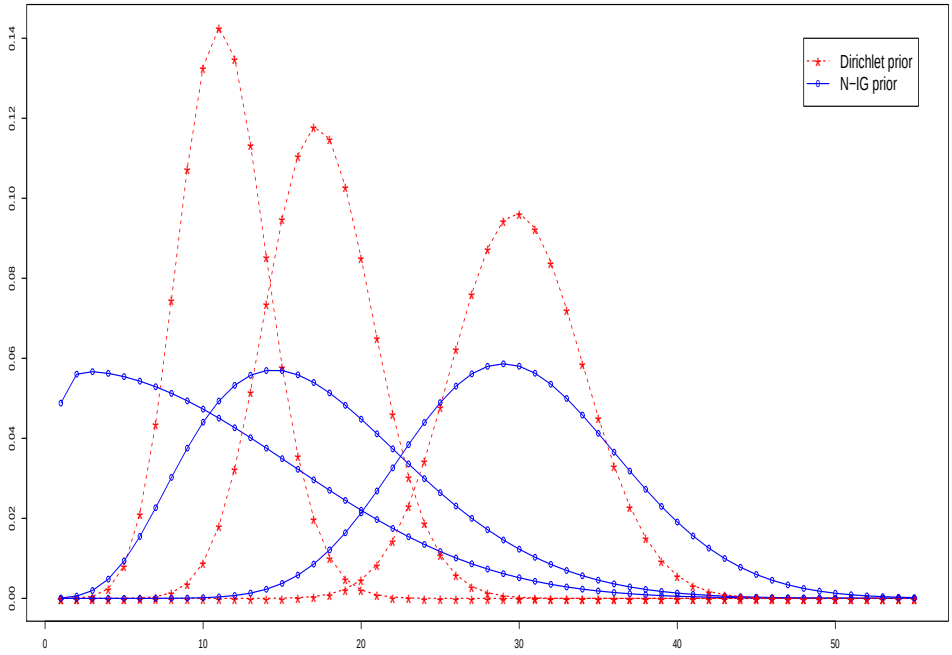

Figure 2.3: Prior probability on the number of different observations k in a sample of $n = 100$ observations, under the N-IG and Dirichlet processes for different choices of the real parameters $\mathbf{a}$ and M. These value have been chosen such that the mean of k is nearly 11 ($\mathbf{a} = 3.10$, $M = 0.01$), 17 ($\mathbf{a} = 5.87$, $M = 1$), and 30 ($\mathbf{a} = 14.16$, $M = 5.39$).

Chapter 3

Markov Chain Monte Carlo methods for NPHM

3.1 The Markov Chain Monte Carlo Methods

Integration plays a fundamental role in Bayesian statistics, and Markov chain Monte Carlo (MCMC) methods are a useful computational tool.

Before Markov chain Monte Carlo became routine and related approximation or numerical methods where developed, Dempster (1980, p. 273) writes: "The application of inference techniques is held back by conceptual factors and computational factors. I believe that Bayesian inference is conceptually much more straightforward than non-Bayesian inference, one reason being that Bayesian inference has a unified methodology for coping with nuisance parameters, whereas non-Bayesian inference has only a multiplicity of ad hoc rules. Hence, I believe that the major barrier to much more widespread application of Bayesian methods is computational.... The development of the field depends heavily on the preparation of effective computer programs."

MCMC methods, which partially resolved the problem delineated by Dempster, originated in the statistical physics literature by Metropolis, Rosenbluth, Teller, and Teller (1953) and sub-

sequently generalized by Hastings (1970), and were firstly used in spatial statistics and image analysis. Tierney (1994) gives a comprehensive theoretical exposition of these algorithms, and Chib and Greenberg (1995) provide a useful tutorial.

For instance, let Y be a random variable in a sample space $\mathcal{Y}$, with density $f(y|\theta)$ depending on a parameter $\theta \in \Theta \subset \mathbb{R}^n$; let $\pi(\theta)$ be the prior distribution of θ. Suppose we are interested in evaluation a quantity like the posterior mean

$$\mathbb{E}[g(\theta|y)] = \int_\theta g(\theta)\pi(\theta|x)d\theta \tag{3.1}$$

were g is some function of the parameter θ. It is clear that, in this framework, it is important to have methods to approximate integrals in some complex space with respect to complicated functions. A powerful tool to address the computational challenges posed by the Bayesian paradigm is the MCMC simulation. The two building blocks of MCMC are, as the name itself suggests, Monte Carlo simulation and Markov chains. Before introducing MCMC, we will recall what we mean for Monte Carlo integration, and we will state some basic definition about Markov chains. We refer to Nummelin (1984), Meyn and Tweedie (1993) and Tierney (1994) for a detailed introduction on this topic.

Firstly suppose we are able to generate a sample of i.i.d. observations, $(\theta_1, \cdots, \theta_n)$, from $\pi(\cdot|y)$. Then we can resort to *Monte Carlo simulation* (McCracken, 1955) and estimate $\mathbb{E}[g(\theta|x)]$ by the sample mean

$$\frac{1}{n}\sum_{i=1}^{n} g(\theta_i).$$

Assuming that g has finite variance under $\pi(\cdot|y)$, the law of large numbers guarantees that $(1/n)\sum_{i=1}^{n} g(\theta_i)$ is a consistent estimator of $\mathbb{E}[g(\theta|x)]$.

Suppose now that $\pi(\cdot|y)$ is a complex distribution such that we are not able to (directly) generate an i.i.d. sample from it, hence we cannot apply the Monte Carlo method. The idea of the

MCMC methods is that of using the Markov Chain (MC) theory to build a sample that is approximately i.i.d. from the *target* distribution $\pi(\cdot|y)$.

A discrete-time Markov chain, with state space $\mathcal{X}$ endowed with a $\sigma-$field $\mathcal{B}(\mathcal{X})$, is a stochastic process $\{X_1, X_2, \cdots\}$ that evolves in time with the property that the future is independent from the past given the present:

$$\mathbb{P}\{X_{m+1} \in A|X_1, \cdots, X_m\} = \mathbb{P}\{X_{m+1} \in A|X_m\},$$

for each $m \geq 0$ and for any A in $\mathcal{B}(\mathcal{X})$.

We identify a MC with the corresponding transition kernel K, defined for any measurable set A and any element x of the state space as:

$$K(x, A) := \mathbb{P}\{X_{m+1} \in A|X_m = x\} \quad \text{for each } m > 0.$$

We are implicitly assuming that the transition probabilities are invariant over time (*time-homogeneity*). A MC has *invariant* (or *stationary*) distribution π if

$$\pi K(A) := \int K(x, A)\pi(dx) = \pi(A),$$

for each measurable A. Not all MC's have an invariant distribution and even when an invariant distribution exists it may not be unique. The basic principle behind MCMC is that certain Markov chains converge to a unique invariant distribution and can thus be used to estimate expectations with respect to this distribution. We refer to Mira (2005) for an introduction on the MCMC methods in Bayesian estimation or, for a complete treatment of this subject, to Robert and Casella (2004) or Gilks, Richardson, and Spiegelhalter (1996).

3.1.1 The Metropolis-Hastings algorithm

A very general procedure to construct a Markov chain $(\theta_1, \theta_2, \dots)$, stationary with respect to a specified distribution, π, is the Metropolis-Hastings algorithm (Metropolis *et al.*, 1953; Hastings, 1970). We can summarize a Metropolis-Hasting's transition probability in the following way: given the current position of the MC, $\theta_m = \theta$, a move to θ' is proposed using the distribution $q(\theta, \theta')$ (such that we are able to sample from it), that may depend on the current position. Such move is accepted with probability

$$\alpha(\theta, \theta') = \min\left\{1, \frac{\pi(\theta')q(\theta', \theta)}{\pi(\theta)q(\theta, \theta')}\right\},$$

where we set $\alpha(\theta, \theta') = 1$ if $\pi(\theta) = 0$. If the move is rejected the current position is retained. We observe that, in the acceptance probability expression, the target distribution enters only as a ratio: $\pi(\theta')/\pi(\theta)$. This means that, the possibly unknown normalizing constant, i.e. the marginal distribution of the data in Bayes formula, cancels and we can thus easily implement the MCMC setting to estimate (3.1), as long as we can evaluate the product $f(x|\theta)\pi(\theta)$ for any given value of θ up to a constant of proportionality.

3.1.2 The Gibbs sampler algorithm

A special case of Metropolis-Hasting is the *Gibbs sampler*. The Gibbs Sampler was given its name by Geman and Geman (1984), who used it for analysing Gibbs distributions. Nevertheless, the work of Geman and Geman (1984) led to the introduction of MCMC into the mainstream statistics via the articles by Gelfand and Smith (1990) and Gelfand, Hills, Racine-Poon, and Smith (1990). To date, most statistical application of MCMC have used Gibbs Sampling. Suppose that for some $n > 1$, the random variable $\underline{\theta}$, with distribution π can be written as $\underline{\theta} = (\theta_1, \dots, \theta_n)$, where the θ_i's are univariate (or multidimensional). Moreover suppose we can simulate from the

corresponding univariate (multidimensional) *full conditional densities*

$$\theta_i|\underline{\theta}^{(-i)} \sim \pi_i(\theta_i|\underline{\theta}^{(-i)})$$

for $i = 1, \ldots, n$, where $\underline{\theta}^{(-i)}$ is the vector $(\theta_1, \ldots, \theta_{i-1}, \theta_{i+1}, \ldots, \theta_n)$.

To produce a Markov sequence $(\underline{\theta}_1, \underline{\theta}_2, \ldots)$ with π as stationary distribution we can use the following procedure: given the current position of the MC, $\underline{\theta}_m = \underline{\theta}$ a move to $\underline{\theta}'$ is completed using

- $\theta'_1 \sim \pi_1(\theta'_1|\theta_2, \ldots, \theta_n)$
- $\theta'_2 \sim \pi_2(\theta'_2|\theta'_1, \theta_3, \ldots, \theta_n)$
  $\vdots \quad \vdots$
- $\theta'_n \sim \pi_n(\theta'_n|\theta'_1, \ldots, \theta'_{n-1})$.

Thus Gibbs sampling consist purely in sampling from the full conditional distributions. A feature of a Gibbs sampler is that even in a high-dimensional problem, all of the simulation may be univariate, which is usually an advantage. This advantage is compensated by the fact that the full conditionals are often not easy to obtain and sampling can be difficult. In this case we could resort to a *Metropolis within Gibbs* algorithm in which the easy to sample full conditionals are used as proposals while the other ones are substituted with different proposals and the corresponding acceptance probability is computed.

3.2 Markov Chain Sampling Methods for NPHM

Modelling a distribution as a mixture of simpler distributions is a useful structure in a wide range of statistical problems. As discussed in Chapter 2, mixtures with a countable infinite number of

components can be reasonably handled, in a Bayesian framework, by employing a nonparametric prior distribution for mixing proportions such as Dirichlet processes or NI-G processes.

Let $(Y_1, \ldots, Y_n)$ a sample from a nonparametric hierarchical mixture prior (2.9), i.e.

$$\begin{aligned} &Y_i|\theta_i \sim k(\cdot|\theta_i), \quad i = 1, \ldots, n, \\ &\theta_1, \ldots, \theta_n|\mathbf{P} \sim_{iid} \mathbf{P}, \\ &\mathbf{P} \sim q(\cdot) \end{aligned}$$

with $k(\cdot|\cdot)$ being a parametric density on the sample space $\mathcal{Y}$ and $q(\cdot)$ a nonparametric prior on the space of distribution $\mathcal{P}(\mathcal{Y})$ (in our study we will consider only the case when $q(\cdot)$ is a Dirichlet prior or a N-IG prior).

Our estimates will be based on the predictive distribution, i.e the distribution of a new observation Y_{n+1} given the sample $(Y_1, \ldots, Y_n)$, with density that can be written as:

$$f(y_{n+1}|Y_1, \ldots, Y_n) = \int f(y_{n+1}|\underline{\theta}) d\pi(\underline{\theta}|Y_1, \ldots, Y_n)$$

where $\pi(\cdot|Y_1, \ldots, Y_n)$ is the posterior distribution of $\underline{\theta} = (\theta_1, \ldots, \theta_n)$. We have

$$\begin{aligned} f(y_{n+1}|\underline{\theta}) &= \int f(y_{n+1}|\underline{\theta}, \theta_{n+1}) L(d\theta_{n+1}|\underline{\theta}) \\ &= \int k(y_{n+1}|\theta_{n+1}) L(d\theta_{n+1}|\underline{\theta}) \end{aligned} \tag{3.2}$$

where $k(\cdot|\theta_{n+1})$ is the known (parametric) kernel distribution and $L(\cdot|\underline{\theta})$ is the predictive distribution of θ_{n+1} given the observation $\underline{\theta} = (\theta_1, \ldots, \theta_n)$. For the two processes under study, $L(\cdot|\underline{\theta})$ is a mixture between the mean distribution of the prior $q(\cdot)$ and the empirical distribution of the observations; see expression (2.2) for the Dirichlet prior and the expression (2.5) for N-IG prior. Finally, the predictive distribution of the nonparametric hierarchical mixture model can be written

as

$$f(y_{n+1}|Y_1,\ldots,Y_n) = \int k(y_{n+1}|\theta_{n+1})L(d\theta_{n+1}|\underline{\theta})d\pi(\underline{\theta}|Y_1,\ldots,Y_n). \tag{3.3}$$

Then NPHM become computationally feasible if we have methods for sampling from the posterior distribution $\pi(\cdot|Y_1,\ldots,Y_n)$ of the unobservable parameters $\theta_1,\ldots,\theta_n$. When the prior $q(\cdot)$ is a Dirichlet process the problem has been largely investigated starting from the seminal papers of Escobar (1994) and MacEachern (1994). We will give a concise review of these works and their generalizations.

3.2.1 Dirichlet Processes

Let $Y_1,\ldots,Y_n$ be a sample from the hierarchical model (2.9), and let

$$q(\cdot) = \text{Dir}(\mathbf{a}G_0),$$

with $\mathbf{a} > 0$ being a real number and G_0 being a distribution on $\mathcal{Y}$. Marginalizing over the process $\mathbf{P}$ we have, see expression (2.11)

$$\pi(d\underline{\theta}|Y_1,\ldots,Y_n) \propto \prod_{i=1}^{n} k(Y_i|\theta_i)\left(\mathbf{a}G_0 + \sum_{j=1}^{i-1}\delta_{\theta_j}\right)(d\theta_i). \tag{3.4}$$

We are interested in a sample from this distribution to evaluate the predictive distribution (3.3). The first work proposing a Monte Carlo strategy to sample from (2.11) dates back to Kuo (1986), where the author proposed an *importance sampling* method (see Robert and Casella, 2004, p. 80-96). However Kuo's algorithm does not sample values conditionally on the data, which can lead to very inefficient estimates (see Escobar, 1992, for a detailed discussion).

The computational difficulties attached to the MDP were solved by Escobar (1994), who in-

troduced a Gibbs sampler algorithm to sample from (2.11) via the full conditional

$$\pi(d\theta_i|\underline{\theta}^{-i}, Y_1, \ldots, Y_n), \quad \text{for each } i = 1, \ldots, n.$$

By Bayes' theorem, observing that θ_i is independent from Y_j with $j \neq i$, we have, for each $i = 1, .., n$,

$$\pi(d\theta_i|\underline{\theta}^{(-i)}, Y_1, \ldots, Y_n) = \frac{k(Y_i|\theta_i)\mathcal{L}(d\theta_i|\underline{\theta}^{(-i)})}{\int k(Y_i|\theta_i)\mathcal{L}(d\theta_i|\underline{\theta}^{(-i)})}. \tag{3.5}$$

The key idea of Escobar's algorithm is the following: we consider the sample $\theta_1, \ldots, \theta_n$ as a part of an exchangeable sequence having Dir($\mathbf{a}G_0$) as de Finetti measure (see Section 2.2.1). Then for each $i = 1, \ldots, n$, we can consider θ_i as the last observation, after $\theta_1, \ldots, \theta_{i-1}, \theta_{i+1}, \ldots, \theta_n$ are observed, so from the prediction rule (2.2)

$$\mathcal{L}(d\theta_i|\underline{\theta}^{-i}) = \frac{\mathbf{a}}{\mathbf{a} + n - 1} G_0(d\theta_i) + \frac{1}{\mathbf{a} + n - 1} \sum_{j \neq i} \delta_{\theta_j}.$$

therefore, (3.5) can be written as

$$\pi(d\theta_i|\underline{\theta}^{(-i)}, Y_i, \ldots, Y_n) = \frac{\mathbf{a}k(Y_i|\theta_i)G_0(d\theta_i) + \sum_{j \neq i} k(Y_i|\theta_j)\delta_{\theta_j}(d\theta_i)}{\mathbf{a}q_0(Y_i) + q_i(Y_i)} \tag{3.6}$$

where $q_0(\cdot)$ is the marginal distribution defined on the sample space by

$$q_0(y) = \int k(y|\theta)G_0(d\theta) \tag{3.7}$$

and

$$q_i(y) = \sum_{j \neq i} k(y|\theta_j). \tag{3.8}$$

Let $(\underline{\theta}^{(1)}, \underline{\theta}^{(2)}, \ldots)$ be a Markov chain generated by a Gibbs sampler procedure with full condi-

tional given by (3.6); then Escobar (1994) showed, when $k(\cdot|\cdot)$ is a normal with fixed variance, that the distribution $\pi(\underline{\theta}|Y_1, \ldots, Y_n)$ is stationary for such $(\underline{\theta}^{(1)}, \underline{\theta}^{(2)}, \ldots)$. The same result holds for the more general case, when $k(\cdot|\theta)$ is an absolutely continuous family of distributions (see Escobar and West, 1995).

The convergence of Escobar's MCMC method may be rather slow, and sampling thereafter may be inefficient. The problem is that there are often groups of observations that with high probability are associated with the same θ. Since the algorithm cannot change the θ_i for more than one observation simultaneously, an update of the common value requires passage through a low-probability intermediate state in which observation in the group do not have the same θ value. A more efficient algorithm can be obtained resorting to MacEachern (1994), where updating θ's is done in clusters. Let fix some notation. In a DPM model with continuous base measure G_0, let $\underline{\theta}^* = \{\theta_1^*, \ldots, \theta_k^*\}$ denote the set of distinct θ_i's, were $k \leq n$ is the number of distinct elements in the vector $\underline{\theta} = (\theta_1, \ldots, \theta_n)$. Let $c = (c_1, \ldots, c_n)$ denote the vector of the configuration indicator defined by $c_i = c_j$ if and only if $\theta_i = \theta_j$. We will use the term *cluster* to refer to the set of all observation Y_i, or just the index i, or the corresponding θ_i's, with identical configuration index c_i. The numerical values of the c_i are arbitrary, as long as they faithfully represent whether or not $c_i = c_j$; that is, the c_i are important only in that they determine what is called the *configuration* in which the data items are grouped in clusters. We will always consider c such that $c_i \in \{1, \ldots, k\}$, for $i = 1, \ldots n$, and we will indicate with n_{c_j} the size of the cluster associated with the value c_j, i.e., $n_{c_j} = \#\{i : c_i = c_j\}$. Note that knowledge of $\underline{\theta}$ is equivalent to knowledge of k, c, and $\underline{\theta}^*$.

We can rewrite the full conditional (3.6) in term of the new parameterisation as

$$\pi(d\theta_i|\underline{\theta}^{(-i)}, Y_i, \ldots, Y_n) = \frac{\mathbf{a}k(Y_i|\theta_i)G_0(d\theta_i) + \sum_{j=1}^{k^{(-i)}} n_j^{(-i)} k(Y_i|\theta_j^*)\delta_{\theta_j^*}(d\theta_i)}{\mathbf{a}q_0(Y_i) + q_i(Y_i)} \tag{3.9}$$

were $k^{(-i)}$ is the multiplicity of the cluster vector $\underline{\theta}^{*(-i)}$, obtained from the vector $\underline{\theta}^{(-i)}$, and $n_j^{(-i)}$ is the multiplicity of θ_j^* in $\underline{\theta}^{*(-i)}$. Simulation of θ_i from the above conditional is straightforward: with conditional probabilities proportional to $n_j^{(-i)} k(Y_i|\theta_j^*)$, the sample is one of the existing clusters, otherwise, the sample θ_i is drawn anew from the marginal $k(Y_i|\theta_i)G_0(d\theta_i)$. Then we can write the full conditional for the configuration vector c as:

$$\begin{aligned}
&\mathbb{P}(c_i = c_j | c^{(-i)}, \underline{\theta}^*, Y_i) = \frac{n_{c_j}^{(-i)} k(Y_i|\theta_{c_j}^*)}{\mathbf{a} q_0(Y_i) + q_i(Y_i)}, \text{ if } j \neq i \text{ and } n_{c_j}^{(-i)} > 0, \\
&\mathbb{P}(c_i \neq c_j \text{ for all } j \neq i | c^{(-i)}, \underline{\theta}^*, Y_i) = \frac{\mathbf{a} q_0(Y_i)}{\mathbf{a} q_0(Y_i) + q_i(Y_i)}.
\end{aligned} \tag{3.10}$$

The algorithm introduced by MacEachern (1994) for a mixture of normals and by Neal (1992) for models of categorical data, uses an analytical integration over θ_i eliminating them from the algorithm. This procedure requires the computation of complex expressions, which therefore relatively limits its applicability in hierarchical models.

The computational difficulties with the MacEachern algorithm are solved when combined with the Escobar algorithm. This improvement is used in Bush and MacEachern (1996) and West, Müller, and Escobar (1994) who construct a Markov chain $\{\underline{\theta}_m^*, c_m\}_{m \geq 1}$ updating the configuration vector via the full conditionals (3.10) and the cluster vector $\underline{\theta}^*$ using the property that the $\theta_{c_j}^*$'s are conditionally independent with posterior densities

$$p(\theta_{c_j}^* | c, Y_1, \ldots, Y_n) = \left(\prod_{i \in I_{c_j}} k(Y_i|\theta_{c_j}^*) \right) G_0(d\theta_{c_j}^*), \tag{3.11}$$

where $I_{c_j} = \{i : c_i = c_j\}$, for $c_j = 1, \ldots, k$.

We have already observed that, when more observations are associated with the same cluster, the algorithm of Escobar (1994) can not perform excellently. In practice this event occurs when for some $\underline{\theta}^*$ the sum $\sum_{j=1}^k k(Y_i, \theta_j^*)$ becomes very large relative to $q_0(Y_i)$ on any iteration (see

expression (3.9)). This occurs when the Markov chain has "stabilized" on a small number of clusters, and it is then unlikely to generate a "new" value of θ^*. The last described algorithm, obtained combining the Escobar and MacEachern procedures, does not suffer of this pathology in that it "shuffles", through the (3.11), the θ_j^*'s after every step, providing more movement in the MCMC sampler, which in turn improves convergence.

We can summarize the Gibbs sampler procedure as follow: let the current state of the Markov chain consist of $c = (c_1, \ldots, c_n)$ and $\underline{\theta}^* = (\theta_1^*, \ldots, \theta_k^*)$. Repeatedly sample as follow:

For $i = 1, \ldots, n$ [Escobar-MacEachern step]

- If the present value of c_i is associated with no other observation (i.e., $n_{c_i}^{(-i)} = 0$), we remove θ_{c_i} from the state.

- We draw a new value for c_i from $c_i | c^{(-i)}, Y_i, \underline{\theta}^*$ as defined by equation (3.10). If the new c_i is not associated with any other observation, we draw a value for $\theta_{c_i}^*$ from $k(Y_i|\theta_i)G_0(d\theta_i)$ and we add it to the state.

For c_i *with* $i = 1, \ldots, n$ [Shuffle step]

- We draw a new value of $\theta_{c_i}^*$ from the posterior based on prior G_0 and all data points currently associated with the cluster c_j defined in (3.11).

All the algorithms we mentioned are based on the Pòlya urn (predictive) representation of the Dirichlet process. For this reason, usually in literature, one refers to them as *Pòlya urn Gibbs samplers*. This methods are been largely used in nonparametric Bayesian statistics using DPM processes as prior distribution, a survey of these models is given in MacEachern and Müller (2000). We mention that on some of these models prior distributions are also introduced on the hyperpa-

rameters $\mathbf{a}$ and G_0 or sets of covariates, but conditional on these additional parameters, the portion of the model involving the DPM has the form given above.

3.2.2 N-IG process

To exploit a N-IG mixture process for inferential purposes, it is essential to derive an appropriate sampling scheme. In such framework, knowledge of predictive distributions (2.5) is crucial. Indeed this formula suggests to build a generalized-urn Gibbs sampler procedure, adapting the methods described in Section 3.2.1. Let $\underline{\theta} = (\theta_1, \ldots, \theta_n)$ be a sample from a N-IG(MG_0) process, were M is a positive real parameter and G_0 is a distribution on $\Theta \subset \mathbb{R}^m$. Then, $(\theta_1, \ldots, \theta_n)$ can be characterized by the following generalized Pòlya urn scheme. Letting $(\varphi_1, \ldots, \varphi_n)$ be an i.i.d. sample from G_0, $\underline{\theta}$ can be generated as follows: set $\theta_1 = \varphi_1$, and for $i = 2, \ldots, n$, generate from

$$(\theta_i|\theta_1, \ldots, \theta_{i-1}) = \begin{cases} \varphi_i \text{ with prob. } w_{0,i-1}(k_i) \\ \theta^*_{i,j} \text{ with prob. } (n_{i,j} - 1/2)w_{1,i-1}(k_i),\ j = 1, \ldots, k_i \end{cases} \tag{3.12}$$

were k_i represents the number of distinct observations, denoted by $\theta^*_{i,1}, \ldots, \theta^*_{i,k_i}$ with multiplicities $(n_{i,1}, \ldots, n_{i,k_i})$, and $w_{0,\cdot}$ and $w_{1,\cdot}$ are given in expressions (2.6) and (2.7).

Now, let $Y_1, \ldots, Y_n$ be a sample on the space $\mathcal{Y}$ from a NPHM model with

$$q(\cdot) = \text{N-IG}(M, G_0).$$

We are interested in the evaluation of the predictive density given in (3.3). We construct a MCMC method to sample from the posterior distribution of $\underline{\theta}$ given the observations $(Y_1, \ldots, Y_n)$. From the generalized urn representation (3.12), using Escobar's algorithm idea and the cluster parametrization in term of $\underline{\theta}^*, k$ and c, we can write the full conditional as

$$\begin{aligned}\pi(\theta_i|\underline{\theta}^{(-i)}, Y_1, \ldots, Y_n) &\propto w_{0,n-1}(k^{(-i)})k(Y_i|\theta_i)G_0(d\theta_i) \\ &+ \sum_{j=1}^{k^{(-i)}}(n_j^{(-i)} - 1/2)k(Y_i|\theta_j^*)\delta_{\theta_j*}(d\theta_i)\end{aligned} \tag{3.13}$$

were the normalizing constant D is given by

$$D = w_{0,n-1}(k^{(-i)})q_0(Y_i) + w_{1,n-1}(k^{(-i)})q_i(Y_i)$$

with

$$q_0(y) = \int k(y|\theta_i)G_0(d\theta_i) \quad \text{and} \quad q_i(y) = \sum_{j=1}^{k^{(-i)}}(n_j^{(-i)} - 1/2)k(y|\theta_j^*).$$

The N-IG mixture has a behaviour similar to the DPM. In fact, the full conditional (3.13) has the same structure of (3.9). The difference between the two models consists in the weights that the processes respectively use to choose the cluster for an observation Y_i.

To construct an efficient MCMC algorithm for sampling from the posterior distribution π of the $\underline{\theta}$ given the data, we can resort to the idea of MacEachern and write the full conditional for the vector of the clusters c, as follows

$$\begin{aligned}&\mathbb{P}(c_i = c_j|c^{(-i)}, \underline{\theta}^*, Y_i) \propto w_{1,n-1}(k^{(-i)})(n_{c_j}^{(-i)} - \frac{1}{2})k(Y_i|\theta_{c_j}^*), \text{ if } j \neq i \text{ and } n_j^{(-i)} > 0, \\ &\mathbb{P}(c_i \neq c_j \text{ for all } j \neq i|c^{(-i)}, \underline{\theta}^*, Y_i) \propto w_{0,n-1}(k^{(-i)})q_0(Y_i).\end{aligned} \tag{3.14}$$

Then we can implement an Escobar-MacEachern "shuffle" algorithm for the N-IG mixture model. We use the full conditional (3.14) to update the configuration vector c, and we use a mixing strategy for the cluster θ_j^* as outlined in the previous section.

Convergence for such algorithm can be obtained considering it as a particular case of the strategy introduced by Ishwaran and James (2001) for stick-breaking priors. In their paper the authors describe an algorithm to handle NPHM's when the mixing prior is in the class of stick-breaking process, as the N-IG does.

Recently, new algorithms have been proposed for dealing with nonparametric mixtures. Ishwaran and James (2003) proposed a generalized weighted Chinese restaurant algorithm that cover species-sampling mixture models. They formally derived the posterior distribution of a species sampling mixture. To draw approximate samples from this distribution, they devised a computational scheme that requires knowledge of the conditional distribution of the species sampling distributed random measure $\mathbf{P}$, given the unobservable parameters $\underline{\theta}$. When feasible, such an algorithm has the merit of reducing the posterior approximation error. In the case of the N-IG process, sampling from the posterior law is not straightforward, because it is characterized in terms of latent variables (see Section 2.3.3). Nieto-Barajas, Prünster, and Walker (2004) proposed a method similar to that of Ishwaran and James (2003) to sample from the posterior of a mixture of normalized random measures driven by increasing additive processes (RMI). Problems of the same type of that described above arise if one is willing to implement this scheme for a N-IG mixture.

3.3 Method for non-Conjugate Models

The Pòlya urn Gibbs samplers described in the last Sections are practicable only if $k(\cdot|\cdot)$ and $G_0(\cdot|\cdot)$ are *conjugate* in θ, allowing analytic evaluation of $q_0 = \mathbb{P}(c_i \neq c_j \text{ for } i \neq j| \dots)$ (see expressions (3.10) and (3.7)). West, Müller, and Escobar (1994) presented the first algorithm designed specifically for use with non-conjugate models. In their algorithm they approximate q_0 taking a random draw from G_0, say θ', and replace $\int k(Y_i|\theta_i)G_0(d\theta_i)$ with $k(Y_i|\theta')$ (one sample Monte Carlo approximation). This approximation is quite inaccurate because it can lead to the

wrong stationary distribution (see, MacEachern and Müller, 1998 or Neal, 2000).

MacEachern and Müller (1998) introduced a sampling plan that entirely avoids the difficult integration to evaluate q_0. In their *no gaps* algorithm the configuration vector is constructed such that $c_i \in \{1, \ldots, k\}$ for each i. Then the cluster vector is $\underline{\theta}^* = (\theta_1^*, \ldots, \theta_k^*)$ and a set of, i.i.d. from G_0, auxiliary variables $(\theta_{k+1}^*, \ldots, \theta_n^*)$ are introduced *augmenting* $\underline{\theta}^*$, these are interpreted as not yet used clusters. In the augmented model the Gibbs sampler is simplified, practically, evaluation of integrals of the type q_0 is performed marginalizing over the augmenting variables.

The work of MacEachern and Müller (1998) was extended by Neal (2000). In this paper the author gives a complete review of the past work and presents two new approaches for handling non-conjugate priors. It suggests a Metropolis-Hasting and a data augmentation procedures that both refine the previous sample schemes to update the configuration parameter c.

Another approach to handling non-conjugate prior was devised by Walker and Damien (1998). Their method avoids the integrals needed for Pòlya Gibbs sampling, but requires instead that the probability under G_0 of the set for which $k(Y_i|\theta) > u$ (were $u > 0$ is a real number) be computable, and that one is able to sample from G_0 on this set. Although these operations are feasible for some models, they will in general be quite difficult, especially when θ is multidimensional.

Finally, Green and Richardson (2001) and Jain and Neal (2004) developed a Markov chain sampling method based on splitting and merging components that is applicable to non-conjugate models. These methods are more complex than those already discussed, since they attempt, using a reversible jump strategy (see Green, 1995), to solve the difficult problem of obtaining a good performance in situations where the other methods tend to become trapped in local modes.

3.3.1 Data Augmentation Methods

The basic idea of data augmentation methods on sampling algorithms is the introduction of some appropriate auxiliary variables that make the sampling procedure much easier. Suppose we wish to

sample from a distribution π_θ for θ, then we can resort to an auxiliary variable τ such that it is easy to sample from the joint distribution $\pi_{\theta\tau}$. We can built a Markov chain with the following strategy. Let the permanent state of the chain be θ, and introduce the auxiliary variable τ temporarily during an update of the following form:

1. Draw a value τ from its conditional distribution given θ, as defined by $\pi_{\theta\tau}$.
2. Perform some update (θ', τ') that leaves $\pi_{\theta\tau}$ invariant.
3. Discard τ', leaving only the value of θ'

It easy to see that this update for θ leaves π_θ invariant as long as π_θ is the marginal distribution of θ under $\pi_{\theta\tau}$.

We are going to report now how use a data augmentation strategy to modify the MacEachern-Escobar Gibbs sampler algorithm, described in the previous sections, in such a way that the analytic integration with respect to G_0 in (3.7), to evaluate the quantities $q_0(Y_i)$, can be avoided.

As shown in Section 3.2 to give an estimation of the predictive distribution (3.3) we need to build a Markov chain having $\pi(d\underline{\theta}|Y_1, \ldots, Y_n)$ as stationary distribution, and this can be done by a Gibbs sampler algorithm via the full conditionals

$$\pi(d\theta_i|\underline{\theta}^{(-i)}, Y_1, \ldots, Y_n) \propto k(Y_i|\theta_i)\mathcal{L}_q(d\theta_i|\underline{\theta}^{(-i)}) \text{ for } i = 1, \ldots, n.$$

where $\mathcal{L}_q$ is the predictive law of the nonparametric prior q. For both the N-IG prior and the Dirichlet prior, the law $\mathcal{L}_q$ can be written, for $i = 1, \ldots, n$, as:

$$\mathcal{L}_q(d\theta_i|\underline{\theta}^{(-i)}) = H_0(n, k^{(-i)})G_0(d\theta_i) + \sum_{j=1}^{k^{(-i)}} H_1(n, n_j^{(-i)}, k^{(-i)})\delta_{\theta_j^*}(d\theta_i), \tag{3.15}$$

where the weights H_0 and H_1 are given in equation (2.2) for the Dirichlet prior and in equation (2.5) for the N-IG prior.

Following the idea of Algorithm 8 in Neal (2000) we can use a data augmentation technique to update θ_i from (3.15) introducing, for each $i = 1, \ldots, n$, a vector of auxiliary variables $\tau = (\tau_1, \ldots, \tau_s)$ iid according to the distribution G_0. In this way the law (3.15) conditionally to τ is

$$\mathcal{L}_q(d\theta_i|\underline{\theta}^{(-i)}, \tau) = H_0(n, k^{(-i)})\frac{1}{s}\sum_{l=1}^{s}\delta_{\tau_l}(d\theta_i) + \sum_{j=1}^{k^{(-i)}} H_1(n, n_j, k^{(-i)})\delta_{\theta^*}(d\theta_i),$$

for $i = 1, \ldots, n$. To complete the algorithm, we have to sample from the conditional law

$$\pi(\tau|\theta_i, \underline{\theta}^{(-i)}) = \frac{\pi(\theta_i|\tau, \underline{\theta}^{(-i)}) \cdot \pi(\tau)}{\pi(\theta_i|\underline{\theta}^{(-i)})}.$$

Some simple algebra gives that

$$\pi(\tau|\theta_i, \underline{\theta}^{(-i)}) \propto \begin{cases} \prod_{l=1}^{s} G_0(\tau_l) & \text{if } \theta_i = \theta_j \text{ for some } j \neq i \\ \\ \delta_{\theta_i}(\tau_{\bar{l}}) \cdot \prod_{l \neq \bar{l}} G_0(\tau_l) & \text{otherwise; with } \bar{l} \in \{1, \ldots, s\}. \end{cases}$$

In practice, to sample from the conditional distribution of the auxiliary parameter given the current value of θ_i and the rest of the state, we will proceed as follows: if $\theta_i = \theta_j$ for some $j \neq i$, the auxiliary parameter has no connection with the rest of the state, and its are drawn independently from G_0. If $\theta_i \neq \theta_j$ for all $j \neq i$, then it must be equal to one of the s auxiliary parameters. Technically, we should randomly select which auxiliary parameter it is associated with, but since it turns out to make no difference, we can just let θ_i be the first of these. Finally, using the usual reparametrization of $\underline{\theta}$ in term of c and $\underline{\theta}^*$ we can summarize the Gibbs sampler algorithm with auxiliary variable. Let the state of the Markov chain consist of $c = (c_1, \ldots, c_n)$ and $\underline{\theta}^* = (\theta_1^*, \ldots, \theta_k^*)$. Repeatedly sample as follows:

For i=1,. . . ,n [Data augmentation step]

- Let $h = k^{(-i)} + s$, and relabel the distinct c_j for $j \neq i$ with values in $\{1, \ldots, k^{(-i)}\}$. If $c_i = c_j$ for some $j \neq i$, draw $(\tau_1, \ldots, \tau_s)$ independently from G_0 and let $\theta^*_{k^{(-i)}+l} = \tau_l$ for $l = 1, \ldots, s$. If $c_i \neq c_j$ for all $j \neq i$, let c_i have the label $k^{(-i)} + 1$, draw $(\tau_1, \ldots, \tau_{s-1})$ independently from G_0 and let $\theta^*_{(k^{(-i)}+1)+l} = \tau_l$ for $l = 1, \ldots, s-1$.

- Draw a new value for c_i from $\{1, \ldots, h\}$ using the following probabilities:

$$\mathbb{P}(c_i = c | c^{(-i)}, Y_i, \theta^*_1, \ldots, \theta^*_h) \propto \begin{cases} H_1(n, n_c^{(-i)}, k^{(-i)}) \cdot k(Y_i|\theta^*_c) \text{ for } 1 \leq c \leq k^{(-i)} \\ \\ \frac{H_0(n,k^{(-i)})}{m} \cdot k(Y_i|\theta^*_c) \qquad \text{for } k^{(-i)} \leq c \leq h. \end{cases}$$

  Finally change the state to contain only those θ^*_c that are now associated with one or more observations.

  For c_i with $i = 1, \ldots, n$ [Updating $\underline{\theta}^*$]

- Draw a new value of $\theta^*_{c_i}$ from the posterior based on prior G_0 and all data points currently associated with the cluster c_j as defined in (3.11).

This approach is similar to that of MacEachern and Müller (1998), the difference being that in the MacEachern-Müller approach the auxiliary parameter is introduced on the space of the vector $\underline{\theta}^*$ and then a Gibbs sampler procedure is built to update the augmented $\underline{\theta}^*$. On the contrary in the Neal (2000) approach just described, first we use a Gibbs sampler strategy, then for each $i = 1, \ldots, n$ the parameter θ_i is augmented by the τ vector, in this algorithm the auxiliary parameter is regarded as existing only temporarily, during the update of θ_i.

We observe also that, as $s \to \infty$, this algorithm approaches the behaviour of the Escobar-MacEachern algorithm described in Section 3.2.1, since the m (or $m-1$) values for θ^*_c drawn from G_0 effectively produce a Monte Carlo approximation to the quantities $q_0(Y_i),\ i = 1, \ldots, n$. However, the equilibrium distribution of the Markov chain defined by the data augmentation pro-

cedure is correct for any value of s, unlike the simulation when a Monte Carlo approximation is used to implement the Escobar-MacEachern algorithm (see, West *et al.*, 1994 and MacEachern and Müller, 1998).

Chapter 4

A comparison of two NPHMs in regression for survival time data

4.1 Introduction

In this chapter we will study a semiparametric accelerated failure time (AFT) survival model (see Section 1.4) analyzing the performances of two nonparametric prior specifications for the error variable.

Let $T_1, \ldots, T_n$ be survival times of n subjects. In the AFT model, the covariates act multiplicatively on the survival time. We assume for each $i = 1, \ldots, n$

$$T_i = \mathrm{e}^{-\mathbf{x}_i'\beta} V_i,$$

or equivalently, letting $W_i = \mathrm{e}^{V_i}$,

$$\log(T_i) = -\mathbf{x}_i'\beta + W_i,$$

where $\mathbf{x}_i = (x_{i1} \ldots, x_{ip})'$ is a known vector of covariates for the ith patient, and β is an unknown column vector of p regression coefficients. In this thesis we assume the error terms $V_i,\ i = 1 \ldots, n$ as a sample from a NPHM model, i.e. we assume that $V_1, \ldots, V_n$, given a random distribution $\mathbf{G}$ on the Euclidean space Θ, are independent and identically distributed from the following density

$$f(\cdot|\mathbf{G}) = \int k(\cdot|\theta) d\mathbf{G}(\theta)$$

with the unknown $\mathbf{G}$ chosen according to a nonparametric law $q(\cdot)$. In particular we direct our the attention at two particular choices for $\mathbf{G}$: $\mathbf{G}$ is a Dirichlet process yielding a DPM model (see Section 2.4.1), or $\mathbf{G}$ is a N-IG mixture model (see Section 2.4.2). Our primary interest is a comparison between the two models specification. As already pointed out in the previous Chapters, the N-IG mixture model represents an interesting alternative to the DPM model. Indeed, as the Dirichlet process, the N-IG process selects discrete distributions with probability one, and it preserves almost the same tractability; nevertheless it is characterised by a more elaborate clustering structure that makes use of all the information contained in the data (see Section (2.3.3)). The matching between the two priors is achieved centering $\mathbf{G}$ at the same distribution function G_0, and letting the prior means of the number of components in the mixture coincide.

We will consider hierarchical mixtures of gamma densities, mixed on both the scale and the shape. The centering distribution G_0 we choose following Hanson yields an infinite mean marginal prior for V; then we resort to a *median* regression model, and prior information will be expressed by means of the median of V. Posterior inference on regression parameters β and on the survival function $S(t) := P(T > t)$ is carried out via Gibbs sampling, incorporating censoring when necessary. Of course, density estimation can be performed within this model ignoring the regression aspect (i.e. simply assuming a null vector of covariates). The two "competing" models were tested on real and simulated data. We will use the same notation to denote distribution functions and the

corresponding probability measures.

4.2 Quantile Regression Models

The theory of linear models is essentially a theory for models of conditional expectations. In many applications, however, it is fruitful to go beyond these models. Quantile regression is gradually emerging as a comprehensive approach to the statistical analysis of linear and nonlinear response models. Employing the standard additive regression formulation, the $p-$th quantile linear regression model (the special case $p = 1/2$ is called *median* regression model) for response observation Y_i, with associated covariate vector $\mathbf{X}_i$, $i = 1, \ldots, n$, can be written as

$$Y_i = \mathbf{X}_i'\beta + W_i, \tag{4.1}$$

where the W_i are assumed conditionally independent from an error distribution with $p-$th quantile equal to zero, i.e.,

$$\int_{-\infty}^{0} f_W(w)dw = p, \tag{4.2}$$

with the function $f_W(\cdot)$ denoting the error density. There is a fairly extensive literature on classical estimation for model (4.1), we refer, for example, to the review papers by Buckinsky (1998) and Yu, Lu, and Stander (2003), and to the work of Ying *et al.* (1995) for the model with censored observations. In this literature no likelihood specification for the response distribution are made (a part for the quantile constrain (4.2)), and point estimation for β proceeds by optimization of some *loss* function. Any inference beyond point estimation is based on asymptotic arguments or resampling methods and thus relies on the availability of large samples.

The Bayesian approach to these models enables exact inference as opposed to the asymptotic inference of the classical approach, moreover Bayesian inference deals in a better way with pa-

rameters uncertainty. The relative ease with which MCMC methods may be used for obtaining the posterior distribution, even in complex situations, has made Bayesian inference much more accessible and attractive. MCMC methods make the entire posterior distribution of parameter β of interest available.

As mentioned in Section 1.4, the special case of median regression has been considered in the Bayesian literature (see, e.g, Walker and Mallick, 1999, Kottas and Gelfand, 2001, Hanson and Johnson, 2002) and little works exists for general quantile regression modelling. See, e.g., Yu and Moyeed (2001) for a parametric approach based on the asymmetric Laplace distribution for the error, Dunson and Taylor (2005) for an approximate method based on the substitution likelihood for quantiles or the recent work of Hjort and Petrone (2006) for nonparametric inference. We mention also the work of Kottas and Krnjajič (2005) who propose a Bayesian nonparametric methodology for quantile regression modelling, developing some MDP models for the error distribution in additive quantile regression formulation.

In our work we consider a multiplication regression model

$$T = \mathrm{e}^{-\mathbf{X}'\beta} \cdot V,$$

without the intercept parameter β_0. Given β, this specification leads to a proportionality relation between the quantile function of the error variable V and the time variable T:

$$Q_T(p) = Q_V(p) \cdot \mathrm{e}^{-\mathbf{X}'\beta}, \quad p \in (0,1). \tag{4.3}$$

If we fix $p = 1/2$ then $m := Q_V(1/2)$ is the prior median of V that represents the baseline median of T, i.e. the prior median with no effect of covariates ($\mathbf{X} = 0$). In our model specification we will choose $f_V(\cdot)$ such that

$$\int_0^m f_V(v)dv = 1/2 \tag{4.4}$$

thus resulting in a semiparametric $m-$*median* regression model.

4.3 The model

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

$$
\begin{aligned}
&T_i = \mathrm{e}^{-\mathbf{x}_i'\beta} \cdot V_i,\ i = 1, \ldots, n, \\
&V_i|\theta_i \overset{ind}{\sim} k(\cdot|\theta_i), \\
&\theta_i|\mathbf{G} \overset{iid}{\sim} \mathbf{G}, \\
&\mathbf{G} \sim q, \quad G_0(A) := \mathbb{E}_q(\mathbf{G}(A)),\ A \in \mathcal{B}(\Theta) \\
&\beta \perp \mathbf{G},\ \beta \sim \pi(\beta),
\end{aligned}
\tag{4.5}
$$

were $k(\cdot|\theta_i)$ is a family of densities on $\mathbb{R}^+$, depending on a vector of parameters θ_i belonging to a Borel subset Θ of $\mathbb{R}^r$, and q is the prior distribution of the random distribution function $\mathbf{G}$, G_0 being a distribution function on Θ, expressing the "mean" of $\mathbf{G}$. Here we will assume that q is a Dirichlet prior or a N-IG prior. The Bayesian model specification is completed assuming G_0 depends on a vector of s hyperparameters $\gamma = (\gamma_1, .., \gamma_s)$ (possibly random and distributed according to $\pi(\gamma)$). In ours model specification, the nonparametric prior $q(\cdot)$ is chosen such that it is centred on the same "mean" distribution G_0 for both the Dirichlet and the N-IG specification. The two priors depend also on two positive parameters expressing the "total mass", which we denoted by $\mathbf{a}$ for the Dirichlet process, and by M for the N-IG process. This parameter can be interpreted as the confidence we have on the choice of G_0 as center measure (see the expressions (2.2) and

(2.4)), but it influences also the induced priors on the number of components in the mixture model. Indeed, as pointed out in Section (2.4), the distribution assigned to the random distribution function $\mathbf{G}$ induces a prior distribution on the number of components k on a sample $(\theta_1, \ldots, \theta_n)$ of dimension n from $\mathbf{G}$. Expressions for this distribution, which we called $p(k|n),\ k = 1, \ldots, n$, are given in (2.12) for the Dirichlet process and in (2.13) for the N-IG process. Hence, the matching between the two non parametric priors is completed choosing $\mathbf{a}$ and M such that

$$\mathbb{E}_{\text{Dir}}(k|n) = \mathbb{E}_{\text{N-IG}}(k|n),$$

i.e., we will assume that the prior means of the number of components in both mixture models coincide.

4.4 Hyperparameters

In the hierarchical model (4.5), we are assuming the error variables $V_i,\ i = 1, \ldots, n$ as a sample from a NPHM model, i.e. given $\mathbf{G}$, as an i.i.d. sample from the density:

$$f(v|\mathbf{G}) = \int k(v|\theta) d\mathbf{G}(\theta), \tag{4.6}$$

were the unknown $\mathbf{G}$ is chosen according the law $q(\cdot)$. Lo (1984), in the context of DPM models, discusses various choices of the family of kernel densities $\{k(v|\theta), \theta \in \Theta\}$ that include histogram models, uniform densities over $(0, \theta)$, exponential densities with parameter θ, and normal densities with $\theta = (\mu, \sigma^2)$. In the context of AFT, Kuo and Mallick (1997) used normal densities with a fixed variance, achieving a prior from V that gives positive mass to non positive values. Recently DPM with kernel having support on $\mathbb{R}^+$ are been studied by Ghosh and Ghosal (2006) using a family of Weibull densities or by Hanson (2006) using gamma densities (see Section 1.4.1).

Following Hanson (2006) we considered hierarchical mixtures of gamma densities $k(\cdot;\theta)$, $\theta = (\vartheta_1, \vartheta_2)$, with mean ϑ_1/ϑ_2. We observe that in the no-sample problem, the Bayes estimate of $f(v|\mathbf{G})$ is the marginal distribution

$$q_0(v) = f_V(v) := \int k(v|\theta) dG_0(\theta).$$

Therefore, we should choose $G_0(\cdot)$ so that $q_0(v)$ represents our prior belief about the distribution of the error variable V, in the NPHM model. However, as pointed out in Section 3.2, assuming a $G_0(\cdot)$ conjugate in θ with the kernels family $\{k(\cdot|\theta), \theta \in \Theta\}$ leads to a computationally convenient *conjugate* hierarchical mixture model.

For this reason we choose the centering distribution G_0 on $\mathbb{R}^+ \times \mathbb{R}^+$ as the product of two exponential distributions, i.e. ϑ_1 and ϑ_2, under G_0, are independent, exponentially distributed with parameters γ_1 and γ_2, respectively. Indeed, for model (4.5), under both Dirichlet and N-IG priors, the marginal prior density of V is, for $v > 0$,

$$\begin{aligned} f_V(v) &= \int_0^{+\infty} d\vartheta_1 \int_0^{+\infty} d\vartheta_2 \frac{\vartheta_2^{\vartheta_1}}{\Gamma(\vartheta_1)} v^{\vartheta_1 - 1} \mathrm{e}^{-\vartheta_2 v} \gamma_1 \mathrm{e}^{-\gamma_1 \vartheta_1} \gamma_2 \mathrm{e}^{-\gamma_2 \vartheta_2} \\ &= \frac{\gamma_1 \gamma_2}{v(v+\gamma_2)(\gamma_1 + \log(\frac{v+\gamma_2}{v}))^2}, \end{aligned} \tag{4.7}$$

with distribution function

$$F_V(v) = \frac{\gamma_1}{\left(\gamma_1 + \log(1 + \frac{\gamma_2}{v})\right)}, \;\; v > 0. \tag{4.8}$$

and quantile function

$$Q_V(p) = \frac{\gamma_2}{\mathrm{e}^{\gamma_1(1/p-1)} - 1}, \;\; p \in (0,1).$$

This distribution has infinite mean, but information about the hyperparameters γ_1 and γ_2 will

be derived through (4.4) fixing the median m of V. Therefore, we have

$$\gamma_1 = \log(1 + \gamma_2/m). \tag{4.9}$$

On the other hand, the "remaining" hyperparameter γ_2 controls the dispersion of V: the interquartile range of the marginal prior, as a function of m and γ_2, is

$$\gamma_2\Big(\big((1+\gamma_2/m)^{1/3}-1\big)^{-1} - \big(\big(1+\gamma_2/m\big)^3-1\big)^{-1}\Big),$$

which, for fixed m, increases with increasing γ_2. The 90% prior probability interval for V is

$$\Big[\frac{\gamma_2}{\big(1+\frac{\gamma_2}{m}\big)^{19}-1}, \frac{\gamma_2}{\big(1+\frac{\gamma_2}{m}\big)^{1/19}-1}\Big]. \tag{4.10}$$

The length, L, of the interval (4.10) has a strictly positive lower bound, for each choice of the median m, given by:

$$L = \frac{2\cdot 0.9 - 1}{0.9\cdot(1-0.9)} m = 8.\overline{8}\cdot m,$$

and for fixed γ_2 the quantity L is an increasing function of m.

We point out that this choice for G_0 is helpful in the algorithm implementation, but is not extremely flexible. Indeed, as shown in figure 4.1, $f_V(\cdot)$ is decreasing with an asymptote at 0 for any γ_1, γ_2 (even if $\gamma_1 \neq \log(1+\gamma_2/m)$).

The hyperparameters (γ_1, γ_2) affect also the mean and the variance of the gamma components of the mixture model. Indeed, let $V|\theta \sim \Gamma(\vartheta_1, \vartheta_2)$ indicate a gamma distributed random variable with mean $\mu = \vartheta_1/\vartheta_2$ and variance $\sigma^2 = \vartheta_1/\vartheta_2^2$, with $\theta \sim \text{Exp}(\gamma_1) \times \text{Exp}(\gamma_2)$. Then, given γ_1 and γ_2 the marginal prior on μ is

$$f(\mu|\gamma_1,\gamma_2) = \frac{\gamma_1\gamma_2}{(\gamma_2+\gamma_1\mu)^2}, \quad \mu > 0,$$

and given μ, γ_1, and γ_2 the precision σ^{-2} is distributed as $\Gamma(2, \gamma_1\mu^2 + \gamma_2\mu)$. We observe that the induced density for μ is monotone decreasing and can be very diffuse on the positive reals. Moreover the larger the γ_2 parameter (and consequently the parameter $\gamma_1 = \log(1 + \gamma_2/m)$), the smaller the precision of the component is expected to be.

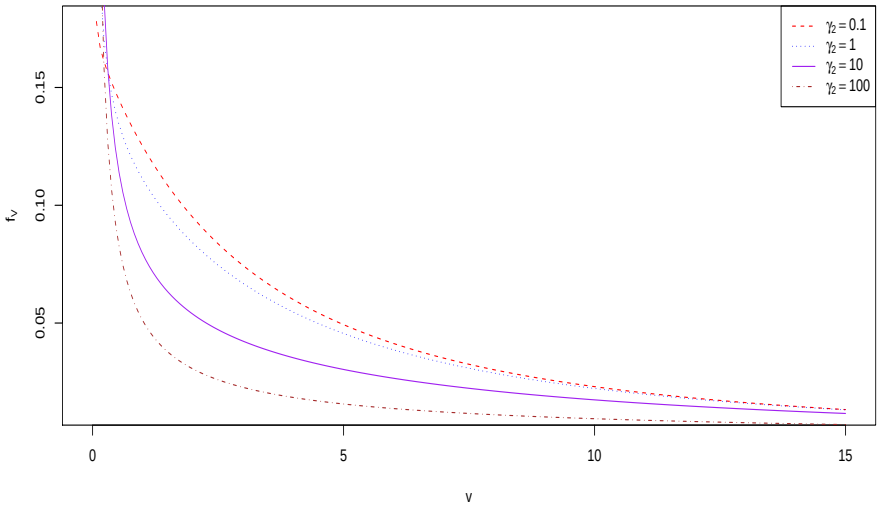

Figure 4.1: Graphics of the marginal prior distribution (4.7) of the error variable V for some choice of the hyperparameter γ_2. The median $m = 5.67$ and the gamma parameters satisfy the relation (4.9)

4.5 The regression coefficient β

To avoid identification problems with the shape parameters of the gamma kernel density ϑ_2 in the AFT model, we do not consider an intercept parameter β_0. Indeed, a prior π_{β_0} does lead to an identification problem. To see this, consider for simplicity, the parametric hierarchical model

$$T_i = \mathrm{e}^{-\beta_0}\mathrm{e}^{-\mathbf{x}_i'\beta}V_i, \quad i = 1, \ldots, n,$$

with $V \sim \Gamma(\vartheta_1, \vartheta_2)$.Then, given β_0 and β, we have

$$T \sim \Gamma\big(\vartheta_1, (\vartheta_2 \mathrm{e}^{\beta_0})\mathrm{e}^{\mathbf{x}_i'\beta}\big)$$

and the product $\vartheta_2 \mathrm{e}^{\beta_0}$ generates an identification inconsistency in the hierarchical model specification.

A flat prior distribution is often imposed upon the β regression parameters (Kuo and Mallick, 1997; Hanson, 2006), regardless of the form of the kernel density $k(\cdot;\theta)$. We introduced instead the reparametrization $\alpha_j = e^{\beta_j}$, $j = 1, \ldots, p$, and assigned independent gamma priors to the α_j's. In this way, with gamma kernel densities, the full conditional posterior distributions of the α_j's associated to binary covariates are still gamma (see Section 4.6.1).

4.6 The algorithm

The Bayesian estimate of the distribution of the survival time is the predictive distribution of a new observation T_{n+1} with covariate $\mathbf{x}_{n+1}$, given the sample $T_1, \ldots, T_n$, from the hierarchical model (4.5), i.e.

$$f_{T_{n+1}}(t|T_1, \ldots, T_n, \mathbf{x}_{n+1}) = \int f_{T_{n+1}}(t|\beta, \underline{\theta}, \mathbf{x}_{n+1}) d\pi(\beta, \underline{\theta}|T_1, \ldots, T_n). \tag{4.11}$$

were $\pi(\cdot, \cdot|T_1, \ldots, T_n)$ is the joint posterior distribution of β and $\underline{\theta} = (\theta_1, \ldots, \theta_n)$. Observe now that, given the vector β and the covariate $\mathbf{x}_{n+1}$, the conditional predictive survival density is given by

$$f_{T_{n+1}}(t|\beta, \underline{\theta}, \mathbf{x}_{n+1}) = f_{V_{n+1}}(t\mathrm{e}^{\mathbf{x}_{n+1}'\beta}|\underline{\theta})\mathrm{e}^{\mathbf{x}_{n+1}'\beta},$$

where $f_{V_{n+1}}(\cdot|\underline{\theta})$ has the following closed form expression

$$f_{V_{n+1}}(v|\underline{\theta}) = \frac{1}{\mathbf{a}+n}\left[\int k(v|\theta_{n+1})dG_0(\theta_{n+1}) + \sum_{i=1}^{n} k(v|\theta_i)\right] \tag{4.12}$$

for the DPM model (see Expression (2.2) and Expression (3.2)), and

$$f_{V_{n+1}}(v\mid,\underline{\theta}) = w_{0,n}(k)\int k(v;\theta_{n+1})G_0(d\theta_{n+1}) + w_{1,n}(k)\sum_{j=1}^{k}\left(n_j - \frac{1}{2}\right)k(v;\theta_j^*) \tag{4.13}$$

for the N-IG mixture model (see Expression (2.5) and Expression (3.2)). We observe that the integral in (4.12) and (4.13) is the marginal prior density of V in (4.7).

To estimate the predictive distribution, we resorted to an MCMC procedure to produce a Markov sequence $\{\beta^{(j)}, \underline{\theta}^{(j)}\}_{j\geq 1}$ having the posterior $\pi(\cdot,\cdot|T_1,\ldots,T_n)$ as stationary distribution. In this way, an estimate of the posterior predictive density given a covariate vector $\mathbf{x}_{n+1}$ is given by

$$\hat{f}_{T_{n+1}}(t|T_1,\ldots,T_n,\mathbf{x}_{n+1}) = \frac{1}{J}\sum_{j=1}^{J} f_{V_{n+1}}(t\mathrm{e}^{\mathbf{x}'_{n+1}\beta^{(j)}}|\theta^{(j)})\mathrm{e}^{\mathbf{x}'_{n+1}\beta^{(j)}} \tag{4.14}$$

The desired Markov sequence $\{\beta^{(j)}, \underline{\theta}^{(j)}\}_{j\geq 1}$ is constructed using a Gibbs sample procedure updating β from the full conditional $\pi(\beta|\underline{\theta}, T_1,\ldots,T_n)$, and $\underline{\theta}$ from the full conditional $\pi(\underline{\theta}|\beta, T_1,\ldots,T_n)$. Moreover in the survival analysis context, the predictive distribution of an individuals with covariate value $\mathbf{x}_{n+1}$ is usually presented as the predictive survival function. Therefore, we also computed

$$\hat{S}_{T_{n+1}}(t|T_1,\ldots,T_n,\mathbf{x}_{n+1}) = \frac{1}{J}\sum_{j=1}^{J} S_{V_{n+1}}(t\mathrm{e}^{\mathbf{x}'_{n+1}\beta^{(j)}}|\underline{\theta}^{(j)}), \tag{4.15}$$

where $S_{V_{n+1}}(t\mathrm{e}^{\mathbf{x}'_{n+1}\beta^{(j)}}|\underline{\theta}^{(j)})$ can be easily found by integrating (4.12) or (4.13).

4.6.1 Updating β

Observing that $T_i = \mathrm{e}^{-\mathbf{x}_i'\beta}V_i,\ i = 1, \ldots, n$, from expression (4.6) we derived the density of T_i,

$$f_{T_i}(t|\mathbf{G}, \beta) = f_{V_i}(v|\mathbf{G})\left|\frac{dv}{dt}\right| = \int \mathrm{e}^{\mathbf{x}_i'\beta} k(t\mathrm{e}^{\mathbf{x}_i'\beta}|\theta_i) d\mathbf{G}(\theta_i).$$

Then, the conditional density of $\beta = (\beta_1, \ldots, \beta_p)$ given $\underline{\theta}$ and $T_1, \ldots, T_n$ is given, up to a proportionality constant, by

$$\pi(\beta|\underline{\theta}, T_1, \ldots, T_n) \propto \pi(\beta) \prod_{i=1}^{n} \mathrm{e}^{\mathbf{x}_i'\beta} k(T_i\mathrm{e}^{\mathbf{x}_i'\beta}|\theta_i). \tag{4.16}$$

Introducing the reparametrization $\alpha_j = \mathrm{e}^{\beta_j},\ j = 1, \ldots, p$, the conditional distribution (4.16) can be written as

$$\pi(\alpha|\underline{\theta}, T_1, \ldots, T_n) \propto \pi(\alpha) \left(\prod_{j=1}^{p} \alpha_j^{\sum_{i=1}^{n} \vartheta_{1,i}\mathbf{x}_{i,j}} \right) \mathrm{e}^{-\sum_{i=1}^{n} T_i\vartheta_{2,i}\left(\prod_{j=1}^{p} \alpha_j^{\mathbf{x}_{i,j}}\right)}.$$

Now, if β_j is a coefficient corresponding to a binary covariate $\mathbf{x}_{i,j}$, then assuming independent prior for the α's, the full conditional of the corresponding α_j is given by

$$\pi(\alpha_j|\alpha^{(-j)}, \underline{\theta}, T_1, \ldots, T_n) \propto \pi(\alpha_j)\alpha_j^{\sum_{\{\mathbf{x}_{i,j}\neq 0\}} \vartheta_{1,i}\mathbf{x}_{i,j}} \mathrm{e}^{-\left\{\sum_{\{\mathbf{x}_{i,j}\neq 0\}} T_i\vartheta_{2,i}\left(\prod_{l\neq j} \alpha_l^{\mathbf{x}_{i,l}}\right)\right\}\alpha_j}, \tag{4.17}$$

where, as usual, $\alpha^{(-j)}$ means the vector α without the element α_j.

If, as prior for α_j, we consider a gamma distribution with mean g_j/h_j, we obtain a *conjugate* model. Therefore, with this choice, the full conditional distribution of α_j is still a gamma, and

$$\alpha_j|\alpha^{(-j)}, \underline{\theta}, T_1, \ldots, T_n \sim \Gamma\left(g_j + \textstyle\sum_{\{\mathbf{x}_{i,j}\neq 0\}}\ ;\ h_j + \sum_{\{\mathbf{x}_{i,j}\neq 0\}} T_i\vartheta_{2,i} \prod_{l\neq j} \alpha_l^{\mathbf{x}_{i,l}}\right).$$

Unluckily, the same conjugacy property does not hold when β_j is the coefficient of a continu-

ous covariate; in this case the full conditional is a non-standard density given by

$$\pi(\alpha_j|\underline{\theta}, T_1, \ldots, T_n) \propto \pi(\alpha_j)\alpha_j^{\sum_{i=1}^n \vartheta_{1,i}\mathbf{x}_{i,j}} \mathrm{e}^{\sum_{i=1}^n T_i\vartheta_{2,i}\left(\prod_{j=1}^p \alpha_j^{\mathbf{x}_{i,j}}\right)}. \tag{4.18}$$

To sample from the density proportional to (4.18), we used a Metropolis-Hasting step in the Gibbs-sampler algorithm. If α_j is the current state of the chain we proposed a new value α_j' from a Log-Normal(μ, σ) distribution, with mean in the current state of the chain, $\mu = \alpha_j$, and standard deviation

$$\frac{1}{\sum_{i=1}^n T_i\vartheta_{2,i}\alpha_j^{\mathbf{x}_{i,j}}\mathbf{x}_{i,j}^2 + h_j\alpha_j}$$

This expression is an approximation of the dispersion of the density in (4.18), arising from a second order Taylor expansion of this density around the current state of the chain.

4.6.2 Updating θ's

Conditionally on the vector β, the model (4.5) is equivalent to a NPHM model described in Section 2.9; indeed, for $i = 1, \ldots, n$, the observation T_i are deterministically related, through the relation $V_i = \mathrm{e}^{\mathbf{x}_i'\beta}T_i$, to a sample, $V_1, \ldots, V_n$, from a NPHM; therefore, instead of β and $T_1, \ldots, T_n$, we can use β and $V_1, \ldots, V_n$ as conditioning variables in the expressions that follow.

To update the non-observable vector $\underline{\theta}$ we resorted to a Pòlya urn Gibbs sampler scheme such as in Section 3.2. In particular, introducing the cluster reparametrization in term of $\underline{\theta}^*, k$ and c, we used an Escobar-MacEachern "shuffle" procedure.

The vector c is updates through the full conditional

$$\mathbb{P}(c_i = c_j|c^{(-i)}, \underline{\theta}^*, V_i, \beta) = \frac{n_{c_j}^{(-i)}k(V_i|\theta_{c_j}^*)}{\mathbf{a}q_0(V_i) + q_i(V_i)}, \text{ if } j \neq i \text{ and } n_{c_j}^{(-i)} > 0,$$
$$\mathbb{P}(c_i \neq c_j \text{ for all } j \neq i|c^{(-i)}, \underline{\theta}^*, V_i, \beta) = \frac{\mathbf{a}q_0(V_i)}{\mathbf{a}q_0(V_i) + q_i(V_i)},$$

for the Dirichlet prior, and through

$$\mathbb{P}(c_i = c_j | c^{(-i)}, \underline{\theta}^*, Y_i) \propto w_{1,n-1}(k^{(-i)})(n_{c_j}^{(-i)} - \frac{1}{2})k(Y_i|\theta^*_{c_j}), \text{ if } j \neq i \text{ and } n_j^{(-i)} > 0,$$

$$\mathbb{P}(c_i \neq c_j \text{ for all } j \neq i | c^{(-i)}, \underline{\theta}^*, Y_i) \propto w_{0,n-1}(k^{(-i)})q_0(Y_i).$$

for the N-IG prior; with normalizing constant

$$D = w_{0,n-1}(k^{(-i)})q_0(Y_i) + w_{1,n-1}(k^{(-i)})q_i(Y_i).$$

The function $q_0(\cdot)$ coincides to the prior marginal (4.7), and $q_i(\cdot)$ are given, for each i, in (3.8) and (3.14). Then, independently from the nonparametric law $q(\cdot)$, the updating of $\theta^*_{c_j}$, for each $j = 1, .., k$, is performed trough the posterior distribution

$$f(\theta^*_{c_j} | c, V_1, \ldots, V_n, \beta) \propto \left(\prod_{i \in I_{c_j}} k(V_i|\theta^*_{c_j}) \right) G_0(\theta^*_{c_j}) \tag{4.19}$$

were $I_{c_j} = \{i : c_i = c_j\}$. We already pointed out that the choice of the family of kernel densities $\{k(\cdot|\theta), \theta \in \Theta\}$ and the "centering" distribution $G_0(\cdot)$, is such that the model is conjugate in θ, so that the computation of the integral $q_0(\cdot)$, to update the configuration vector c, becomes particularly handy. Moreover the conjugacy is helpful also in updating $\underline{\theta}^*$. Indeed, if c_j is a "one-observation" cluster, i.e. $n_{c_j} = 1$, then we shall simulate $\theta^*_{c_j} = (\vartheta^*_{1,c_j}, \vartheta^*_{2,c_j})$ from the following density (we omit the subscript c_j and the superscript $*$, to simplify the notation)

$$\begin{aligned} f(\theta | c, V_1, \ldots, V_n, \beta) &\propto k(V_j|\theta)G_0(\theta) \qquad (4.20) \\ &= \frac{\vartheta_2^{\vartheta_1}}{\Gamma(\vartheta_1)} V_j^{\vartheta_1 - 1} \mathrm{e}^{-\vartheta_2 V_j} \gamma_1 \mathrm{e}^{-\gamma_1 \vartheta_1} \gamma_2 \mathrm{e}^{-\gamma_2 \vartheta_2}. \end{aligned}$$

From the last expression it clear that

$$f(\vartheta_2|\vartheta_1, c, V_1, \ldots, V_n, \beta) \propto \vartheta_2^{\vartheta_1} \mathrm{e}^{-\{V_j+\gamma_2\}\vartheta_2},$$

and then

$$\begin{aligned} f(\vartheta_1|c, V_1, \ldots, V_n, \beta) &= \frac{f((\vartheta_1, \vartheta_2)|c, V_1, \ldots, V_n, \beta)}{f(\vartheta_2|\vartheta_1, c, V_1, \ldots, V_n, \beta)} \\ &\propto \vartheta_1 \mathrm{e}^{-\{\gamma_1+\ln(1+\frac{\gamma_2}{V_j})\}\vartheta_1}. \end{aligned}$$

Hence, to sample an observation from the density (4.20), we first sampled $\vartheta_1 \sim \Gamma\big(2; \gamma_1 + \ln(1 + \gamma_2/V_j)\big)$, then $\vartheta_2 \sim \Gamma\big(\vartheta_1 + 1; V_j + \gamma_2\big)$.

The same sample scheme cannot be applied in the case the cluster c_j contains more that one observation, i.e. $n_{c_j} > 1$. In this case the density (4.19) becomes

$$f(\theta|c, V_1, \ldots, V_n, \beta) \propto \frac{\vartheta_2^{n_{c_j}\vartheta_1}}{\Gamma(\vartheta_1)^{n_{c_j}}} \Big(\prod_{i\in I_{c_j}} V_i\Big)^{n_{c_j}} \vartheta_1 \mathrm{e}^{-\{\sum_{i\in I_{c_j}} V_i\}\vartheta_2} \gamma_1 \mathrm{e}^{-\gamma_1\vartheta_1} \gamma_2 \mathrm{e}^{-\gamma_2\vartheta_2}. \quad (4.21)$$

This is a non standard density, but we can observe that

$$f(\vartheta_2|\vartheta_1, c, V_1, \ldots, V_n, \beta) \propto \vartheta_2^{n_{c_j}\vartheta_1} \mathrm{e}^{-\{\sum_{i\in I_{c_j}} V_i+\gamma_2\}\vartheta_2},$$

and then the marginal of density of ϑ_1 is given by

$$f(\vartheta_1|c, V_1, \ldots, V_n, \beta) \propto \frac{\Gamma(n_{c_j} \cdot \vartheta_1)}{\Gamma(\vartheta_1)^{n_{c_j}}} \vartheta_1 \exp - \left\{ \left(\gamma_1 + \ln \frac{(\sum_{i\in I_{c_j}} V_i + \gamma_2)^{n_{c_j}}}{\prod_{i\in I_{c_j}} V_i} \right) \vartheta_1 \right\}. \quad (4.22)$$

Sampling from (4.21) is achieved by first updating ϑ_1 from (4.22) via a Metropolis-Hasting step. If ϑ_1 is accepted, then we sample ϑ_2 from $\Gamma\big(n_{c_j}\vartheta_1 + 1; \sum_{i\in I_{c_j}} V_i + \gamma_2\big)$, otherwise both are left at their previous values. The Metropolis-Hasting step to sample from density (4.22) was performed

proposing a new observation ϑ_1' from normal distribution $N(\mu, \sigma)$, with mean in the current state of the chain ϑ_1 and standard deviation

$$\sigma = \frac{1}{\sqrt{2}} \frac{\sqrt{n_{c_j} + 3}}{\gamma_1 + \ln \frac{\left(\sum_{i \in I_{c_j}} V_i + \gamma_2\right)^{n_{c_j}}}{\prod_{i \in I_{c_j}} V_i} - n_{c_j} \ln n_{c_j}}.$$

First we observe that the denominator in the last expression is greater than zero, since the following inequality holds

$$n_{c_j}^{n_{c_j}} \leq \frac{\left(\sum_{i \in I_{c_j}} V_i + \gamma_2\right)^{n_{c_j}}}{\prod_{i \in I_{c_j}} V_i}$$

because $\gamma_2 > 0$ and by Jensen's inequality

$$\left(\prod_{i \in I_{c_j}} V_i \right)^{1/n_{c_j}} \leq \frac{\sum_{i \in I_{c_j}} V_i}{n_{c_j}}.$$

Then, we obtained the estimate σ of the dispersion of (4.22), by observing that

$$f(\vartheta_1 | c, V_1, \ldots, V_n, \beta) \leq K \cdot \Gamma \left(\vartheta_1 \Big| \frac{n_{c_j} + 3}{2} \; ; \; \gamma_1 + \ln \frac{\left(\sum_{i \in I_{c_j}} V_i + \gamma_2\right)^{n_{c_j}}}{\prod_{i \in I_{c_j}} V_i} - n_{c_j} \ln n_{c_j} \right) \tag{4.23}$$

where K is a constant and $\Gamma(\cdot|\mathrm{s}, \mathrm{r})$ is the density of a gamma distributed random variable with shape parameter s and rate parameter r.

Inequality (4.23) follows from the multiplication theorem of the gamma function (see Gradshtey and Ryzhik, 1994, p.946).

4.6.3 Censored observations

Suppose that the first n_1 observations in the sample $T_1, \ldots, T_n$ are known only up to the censoring intervals

$$T_i \in [a_i, b_i), \;\; i = 1, \ldots, n_1,$$

where $b_i = \infty$ for right-censored data. In the MCMC algorithm, we used a data augmentation strategy sampling at each iteration the latent T_i's from the conditional distribution of $[T_i \big| T_i \in [a_i, b_i), \beta, \theta_i]$ for $i = 1, \ldots, n_1$. Indeed, we note that T_i is conditionally independent from $T_j,\ j \neq i$, given $T_i \in [a_i, b_i), \beta$ and θ_i. The updating of the censored T_i was performed at each iteration of the MCMC algorithm using that $T_i \in [a_i, b_i)$ implies $V_i \in \left[a_i \mathrm{e}^{\mathbf{x}_i'\beta}, b_i \mathrm{e}^{\mathbf{x}_i'\beta}\right)$, and then $\left[V_i \big| V_i \in \left[a_i \mathrm{e}^{\mathbf{x}_i'\beta}, b_i \mathrm{e}^{\mathbf{x}_i'\beta}\right), \theta_i\right]$ is distributed as a gamma random variable, with shape ϑ_1 and rate ϑ_2, restricted to the interval $\left[a_i \mathrm{e}^{\mathbf{x}_i'\beta}, b_i \mathrm{e}^{\mathbf{x}_i'\beta}\right)$. We thus sampled $[T_i \big| T_i \in [a_i, b_i), \beta, \theta_i],\ \ i = 1, \ldots, n_1$ using the inverse cumulative distribution function method, i.e. first sampling $U \sim \text{Unif}\big(K(a_i \mathrm{e}^{\mathbf{x}_i'\beta} | \theta_i), K(b_i \mathrm{e}^{\mathbf{x}_i\beta} | \theta_i)\big)$ and then taking $V_i = K^{-1}(U | \theta_i)$, where $K(\cdot|\theta)$ is the cumulative distribution function of a gamma random variable. Finally we set

$$T_i = V_i \mathrm{e}^{-\mathbf{x}_i'\beta},\ \ i = 1, \ldots, n_1.$$

4.7 Data Illustration

4.7.1 Simulated data for density estimation

We studied, first, an AFT model with no effect of covariate, i.e. $\mathbf{x}_i = 0$, for each $i = 1, \ldots, n$. In practice we performed a nonparametric density estimate of a random sample $T_1, \ldots, T_n$ from a NPHM. We considered a simulated data set from a mixture of 3 gamma densities. We generated a sample $t_1, \ldots, t_n$ of size $n = 100$ from the density

$$f(t) = 0.2 \cdot \Gamma(t|40, 20) + 0.6 \cdot \Gamma(t|6, 1) + 0.2 \cdot \Gamma(t|200, 20), \tag{4.24}$$

(with mean 6 and variance 10.12) and computed the posterior density estimates from the Dirichlet and the N-IG mixtures of gammas.

We assumed $M = 0.01$ and $M = 5.39$, which corresponds to a prior mean of the num-

ber of components under the N-IG prior equal to the (actual) minimum value 11.3700 and 30, respectively. The matching with the Dirichlet process prior is achieved when $\mathbf{a} = 3.0981$ and $\mathbf{a} = 14.1614$, respectively. We set the median m of V equal to 5.6702 (i.e. the true median), 56.7016 and 0.5670, and the hyperparameter γ_2 in G_0 equal to 0.01, 1 and 10, corresponding to values for the IQR and 90% prior probability interval listed in Table 4.1.

For each choice of hyperparameters γ_2, m, M and $\mathbf{a}$, we run several chains. We observed that convergence was relatively fast, essentially occurring after $2,000$ iterations. We finally run a long chain for each model, discarding the first $10,000$ iteration (*burn-in*) and then keeping the values every 50 iterations (*thinning*) to reduce autocorrelation. These choices are rather conservative and, indeed, smaller burn-in and thinning would also be adequate.

As a measure of the performances of our estimates, for every choice of the hyperparameters, we computed the error, in the uniform metric, between the true distribution function and the predictive distributions (under both priors). Let $\hat{F}_T(\cdot|t_1, \ldots, t_n)$ represents the predictive Bayesian estimation of the "true" cumulative distribution function $F_T(\cdot)$ based on the data $t_1, \ldots, t_n$ $\big(\hat{F}_T(\cdot|t_1, \ldots, t_n)$ is obtained integrating (4.14)$\big)$; then the error in the uniform metric (EUM) is defined as

$$\text{EUM}(t_1, \ldots, t_n) = \sup_t \left|\hat{F}_T(t|t_1, \ldots, t_n) - F_T(t)\right| \tag{4.25}$$

.

Figures 4.2, 4.3 and 4.4 display the true density, the histogram from the simulated data, and the density estimates under the Dirichlet process prior and the N-IG prior for each value of the hyperparameters; in each graph the two estimates are indistinguishable. Moreover, Table 4.2 presents the observed errors for some choices of the hyperparameters.

We provided also an estimate of the posterior number of clusters in the sample $t_1, \ldots, t_n$. Figures 4.5, 4.6 and 4.7 show the posterior estimates for the each choice of the hyperparameters.

The posterior clusters distributions under the N-IG prior seem to be more robust than the corresponding distributions under the Dirichlet prior. The two models produce almost indistinguishable predictive density estimates, but the posterior clusters configurations are quite different. We argue that the elaborate predictive form of the predictive distribution of a N-IG process makes it quite suited, as prior distribution, in the analysis of the cluster structure of data.

We observe that, under both nonparametric laws $q(\cdot)$, the hyperparameter γ_2 works as a smoothing parameter, since it controls the prior variance of the gamma component of the mixture model. It is interesting to note that (relatively) large γ_2's empirically fit the observed data worst, underestimating the number of modes of the generating distribution (the opposite pathology can arise using small values of γ_2).

In the no covariate NPHM setting, the prior marginal $f(\cdot|G_0) = \int k(\cdot|\theta)dG_0(\theta)$ represents the prior belief on the distribution of T. We already pointed out that our choice of G_0 and $\{k(\cdot|\theta), \Theta\}$ is such that $f(\cdot|G_0)$ is easily computable, but it suffers of lack of flexibility. Indeed, our posterior density estimates are an average of mixtures between the prior marginal and some gamma components (see (4.14),(4.12) and (4.13)). This leads to posterior estimates with an asymptote in zero (as $f(\cdot|G_0)$) also when the data do not indicate this kind of trend, as in the case under observation. This discrepancy becomes more evident when the parameters M and $\mathbf{a}$ become bigger (stronger confidence in the prior), or when the prior median m is set at relatively small values (informative prior).

Finally to quantify the difference in estimating the density (4.24) independently from the observed sample, we computed the mean error between the exact cumulative distribution function and the estimates:

$$\mathbb{E}_{T_1,\ldots,T_n}\big(\text{EUM}(T_1,\ldots,T_n)\big). \tag{4.26}$$

We performed a Monte Carlo estimate of (4.26) for a smaller sample size, generating $J = 200$

samples of size $n = 30$, $\{T_1^{(j)}, \ldots, T_n^{(t)}\}_{j=1}^{J}$, and computing

$$\hat{\mathbb{E}}(EUM(T_1, \ldots, T_n) = \frac{1}{J} \sum_{j=1}^{J} EUM\big(T_1^{(j)}, \ldots, T_n^{(j)}\big).$$

The mean errors (and the corresponding standard deviations) of the estimates for some choices of the hyperparameters are presented in Table 4.3. The value we obtained seem to confirm that the two prior specifications are equivalent in density estimation.

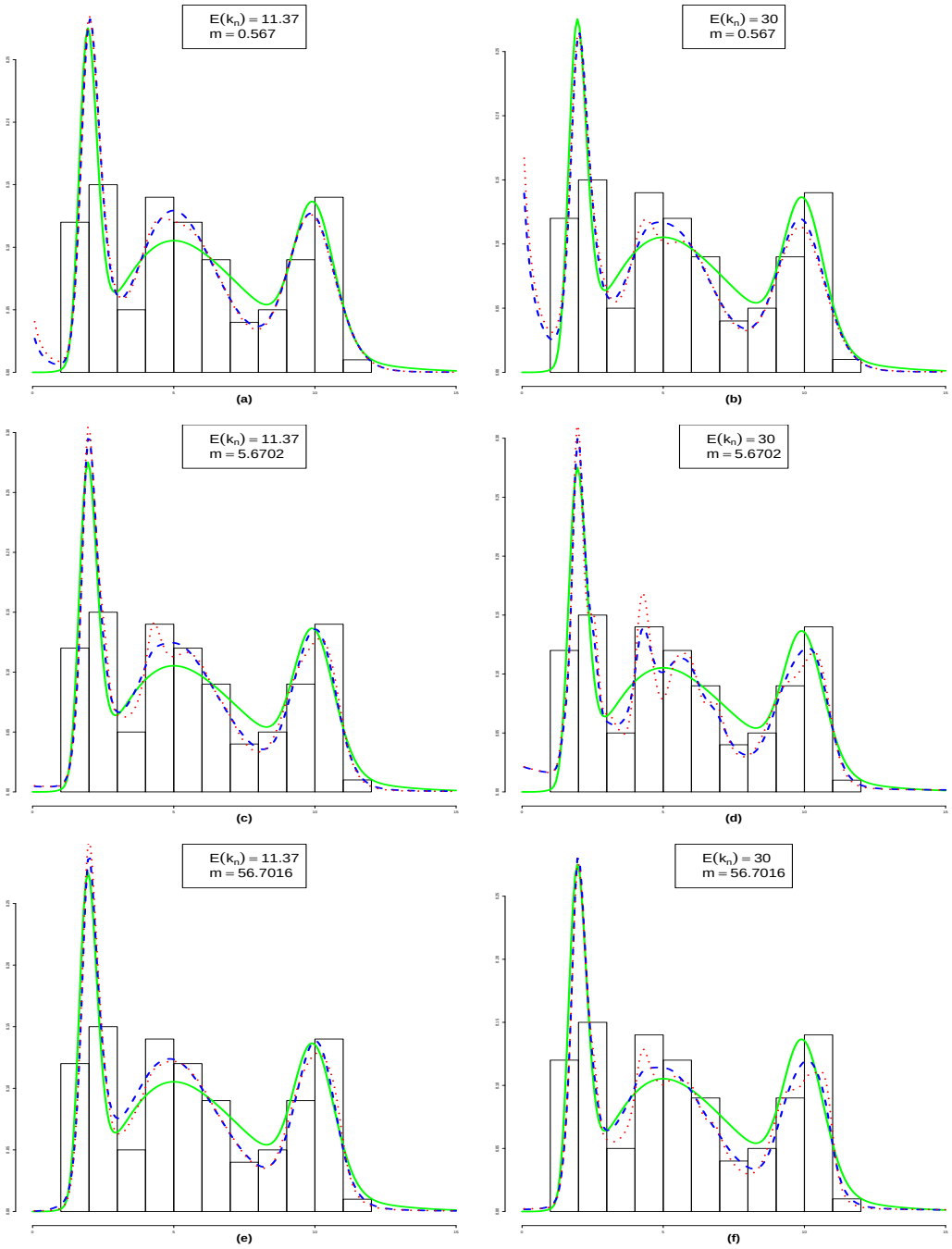

Figure 4.2: Histogram from the simulated data and density estimates under the Dirichlet process prior (dotted red) and the N-IG prior (dashed blue) when $\gamma_2 = 0.01$. $\mathbb{E}(k_n)$ indicates the prior number of component. In each graph the solid (green) line denotes the true density.

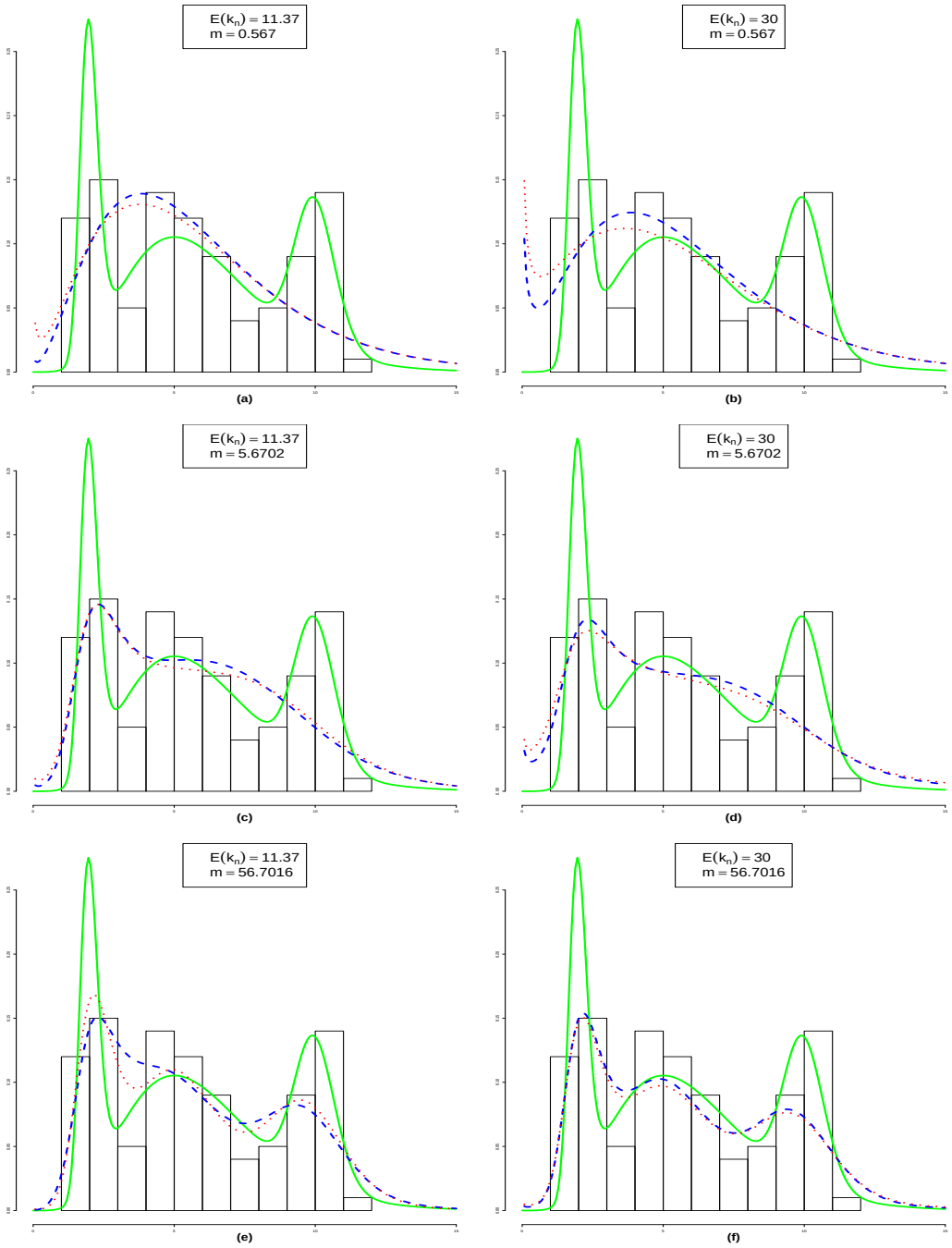

Figure 4.3: Histogram from the simulated data and density estimates under the Dirichlet process prior (dotted red) and the N-IG prior (dashed blue) when $\gamma_2 = 1$. $\mathbb{E}(k_n)$ indicates the prior number of component. In each graph the solid (green) line denotes the true density.

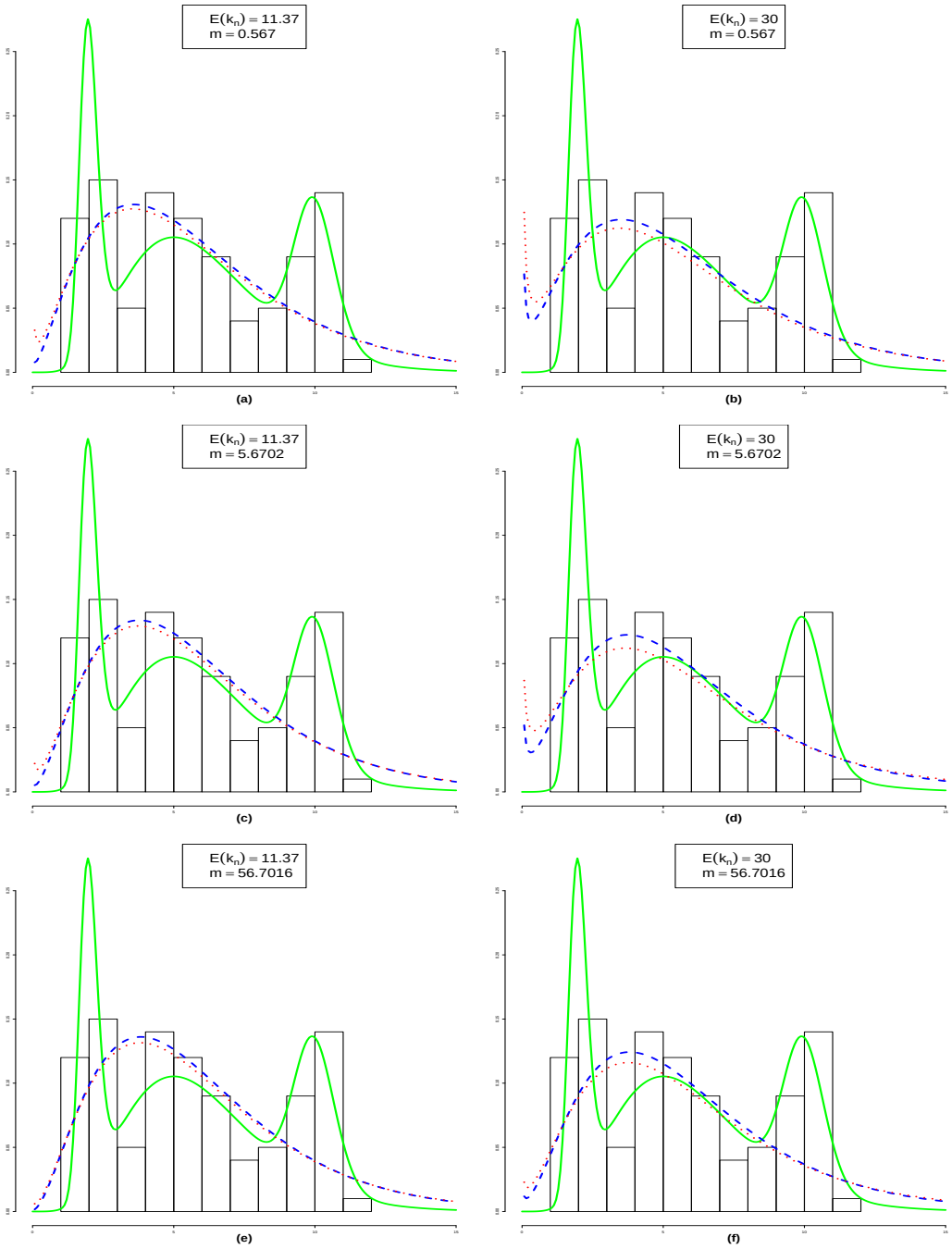

Figure 4.4: Histogram from the simulated data and density estimates under the Dirichlet process prior (dotted red) and the N-IG prior (dashed blue) when $\gamma_2 = 10$. $\mathbb{E}(k_n)$ indicates the prior number of component. In each graph the solid (green) line denotes the true density.

		m		
		0.5670	5.6702	56.7016
γ_2	0.01	1.5255 [0.0254,10.8627]	15.1339 [0.2937,107.8237]	151.2172 [2.9796,1077.4701]
	1	2.4299 $[4 \cdot 10^{-9}, 18.1950]$	16.3819 [0.0479,116.4774]	152.5298 [2.5382,1086.3024]
	10	6.0543 $[7 \cdot 10^{-24}, 60.0826]$	24.2965 [$4 \cdot 10^{-8}$,181.9538]	163.8171 [0.4787,1164.7669]

Table 4.1: IQR and 90% probability interval for the marginal prior of V.

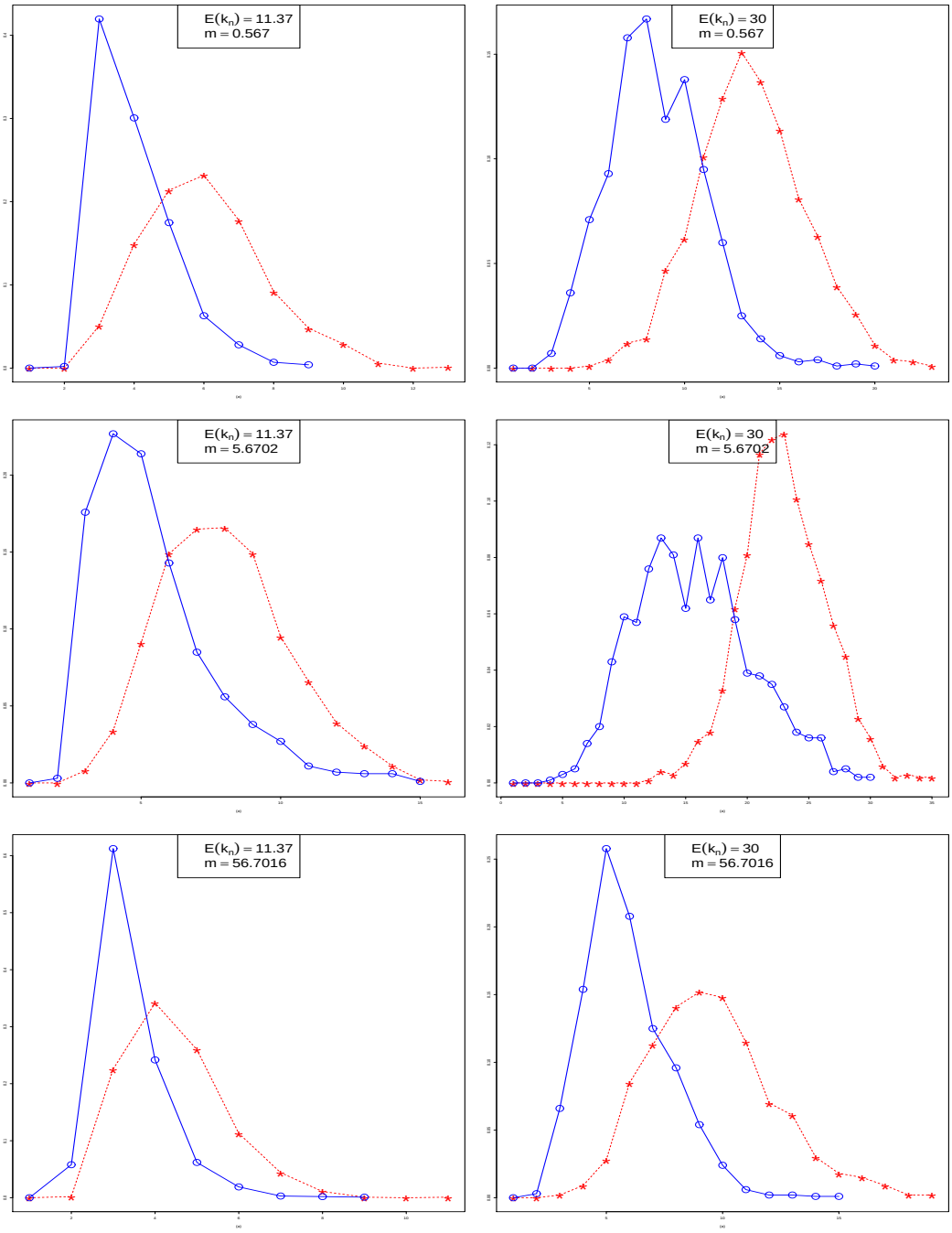

Figure 4.5: Posterior distribution estimates of the number of clusters under the Dirichlet case (dotted red) and N-IG case (continuous blue) when $\gamma_2 = 0.01$.

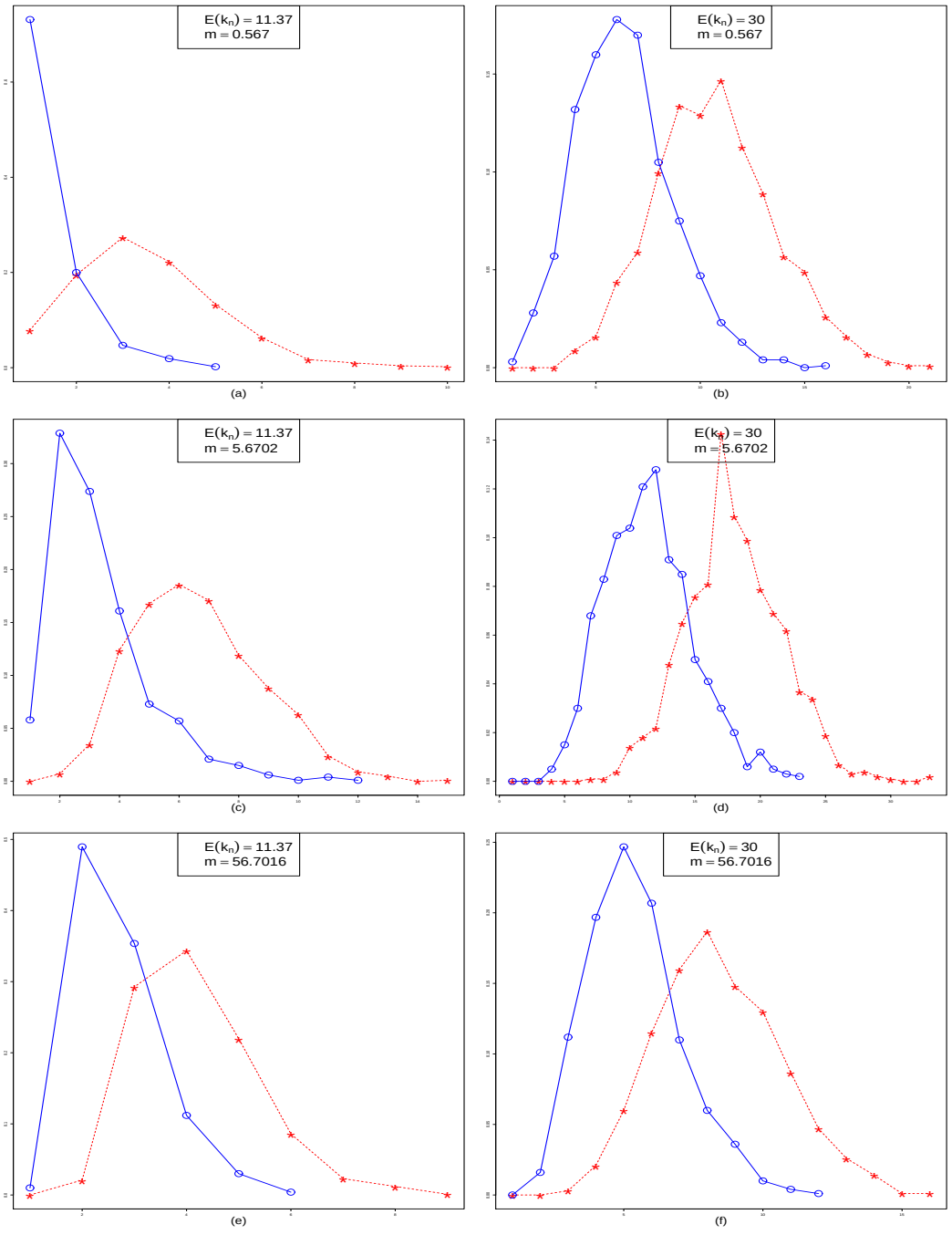

Figure 4.6: Posterior distribution estimates of the number of cluster under the Dirichlet case (dotted red) and N-IG case (continuous blue) when $\gamma_2 = 1$.

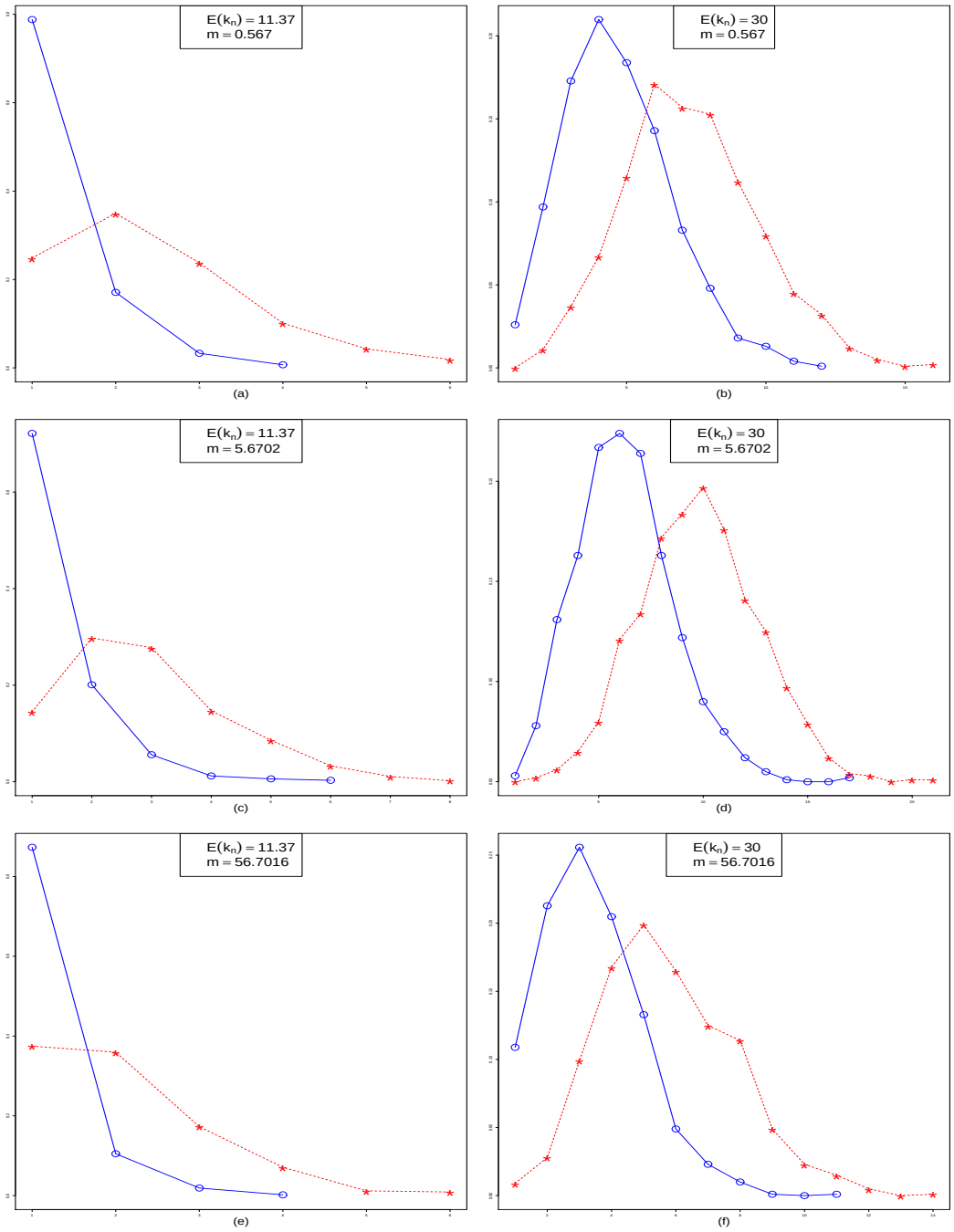

Figure 4.7: Posterior distribution estimates of the number of cluster under the Dirichlet case (dotted red) and N-IG case (continuous blue) when $\gamma_2 = 10$.

		M		**a**	
		0.01	5.39	3.0981	14.1614
m	5.6702	0.0481	0.0386	0.0466	0.0339
	56.7016	0.0334	0.0975	0.0448	0.0768

Table 4.2: Errors in the uniform metric for the simulated dataset of size 100 between the true and estimated distribution functions.

		$\mathbb{E}(k\|n) = 6.2$		$\mathbb{E}(k\|n) = 14.5$	
	$\gamma 2$	N-IG	DIR	N-IG	DIR
	0.01	0.1263 (0.046)	0.1112 (0.044)	0.1281 (0.040)	0.1129 (0.030)
$m = 5.6706$	1	0.1086 (0.032)	0.1028 (0.031)	0.1325 (0.021)	0.1470 (0.017)
	10	0.1144 (0.023)	0.1380 (0.022)	0.1774 (0.018)	0.2216 (0.016)
	0.01	0.1281 (0.042)	0.1254 (0.043)	0.2037 (0.035)	0.2289 (0.025)
$m = 56.7061$	1	0.1198 (0.040)	0.1147 (0.032)	0.2057 (0.020)	0.2504 (0.019)
	10	0.1120 (0.024)	0.1445 (0.023)	0.2370 (0.019)	0.3011 (0.021)
	0.01	0.1177 (0.048)	0.1179 (0.049)	0.1825 (0.032)	0.2055 (0.030)
$m = 0.5671$	1	0.1054 (0.029)	0.1177 (0.028)	0.1914 (0.022)	0.2525 (0.023)
	10	0.1326 (0.020)	0.1518 (0.017)	0.2156 (0.022)	0.2843 (0.026)

Table 4.3: Mean errors and corresponding standard deviations (in brackets) between the estimates and the true distribution for samples of size 30. The error is the distance, in the uniform metric, between distribution functions.

4.7.2 Dataset not involving censoring

We studied 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 a measurement error in the survival time of the 17th patient (AG positive); indeed, it is atypically high (65 week) related to the elevated number of white blood cell recovered ($\text{WBC}_{17} = 10^5$). Anyway we decided to use this dataset as a test for comparing our models.

Patient	AG factor	WBC	Survival Time	Patient	AG factor	WBC	Survival Time
1	1	2300	65	18	0	4400	56
2	1	750	156	19	0	3000	65
3	1	4300	100	20	0	4000	17
4	1	2600	134	21	0	1500	7
5	1	6000	16	22	0	9000	16
6	1	10500	108	23	0	5300	22
7	1	10000	121	24	0	10000	3
8	1	17000	4	25	0	19000	4
9	1	5400	39	26	0	27000	2
10	1	7000	143	27	0	28000	3
11	1	9400	56	28	0	31000	8
12	1	32000	26	29	0	26000	4
13	1	35000	22	30	0	21000	3
14	1	10^5	1	31	0	79000	30
15	1	10^5	1	32	0	10^5	4
16	1	52000	5	33	0	10^5	43
17	1	10^5	65				

Table 4.4: Feigl-Zelen dataset

We considered the bivariate vectors of covariates $\mathbf{x}_i = (x_{i,1}, x_{i,2})'$ such that $x_{i1} = 1$ if AG positive and 0 if AG negative and $x_{i2} \in [0, 1]$. Indeed, in order to maintain numerical stability, we normalized the continuous covariate by defining:

$$x_{i,2} = \frac{\text{WBC}_i - \min_i(\text{WBC}_i)}{\max_i(\text{WBC}_i) - \min_i(\text{WBC}_i)} \quad i = 1, \ldots, 33.$$

We assumed $M = 0.01$ ($a = 2.1478$) and $M = 10$ ($a = 16.3400$), which corresponds to prior means of the number of components in the mixture equal to 6.5108 and 18.3966, respectively.

Following the idea of MacEachern and Müller (2000) that viewed an NPHM model as a robust

extension of a parametric model, we performed a preliminary analysis of the parametric AFT model

$$
\begin{aligned}
&T_i = \mathrm{e}^{-\mathbf{x}_i'\beta} V_i \\
&V_i \sim_{iid} \Gamma(\vartheta_1, \vartheta_2) \\
&\theta \sim \pi(\theta);\ \beta \sim \pi(\beta),\ \ i = 1, \ldots, n,
\end{aligned}
$$

as described in Ibrahim *et al.* (2001, p.40). Then, we obtained the Bayesian estimates of the parameters that we called $\hat{\theta}_{pre}$, $\hat{\beta}_{pre}$ and $\hat{\alpha}_{pre}$ (where $\hat{\alpha}_{j,pre}$ is the estimate of $\alpha_j = \mathrm{e}^{\beta_j},\ \ j = 1, \ldots, p$). We used these estimates as prior information in the nonparametric framework. We assumed the prior median m of the error variable V such that

$$
m = K^{-1}\left(\frac{1}{2}|\hat{\theta}_{pre}\right) = 14.8484,
$$

where $K(\cdot|\theta)$ is a cumulative distribution function of a gamma random variable with mean ϑ_1/ϑ_2. Moreover, we assigned independent gamma priors $\Gamma(g_j, h_j)$ to the regressors parameter $\alpha_j,\ j = 1, \ldots, p$, such that

$$
\mathbb{E}(\alpha_j) = \frac{g_j}{k_j} = \hat{\alpha}_{j,pre} \quad j = 1, \ldots, p
$$

with a non-informative variance $\mathrm{Var}(\alpha_j) = g_j/h_j^2 = 1000$ (in particular $\hat{\alpha}_{pre} = (0.5213, 7.372)$)

Figures 4.8, 4.9 and 4.10 display the estimates of the survival functions for 2 "new" patients (corresponding to covariates $(1, 0.5)$ and $(0, 0.5)$ respectively) when $\gamma_2 = 1$, 10 and 100. The predictive survival function are practically indistinguishable, when $M = 0.01$ ($\mathbf{a} = 2.1478$). A different behaviour in the tails of the predictive survival function arises when the parameter M ($\mathbf{a}$) increases: the survival predictive functions under the N-IG specification have heavier tails indicating a more robust estimate under this prior specification.

Each survival function estimate was obtained through a sample from a Markov chain $\{\underline{\theta}^{(j)}, \beta^{(j)}\}_{j=1}^{J}$ built as described in Section 4.6. We run several independent chains and we observed an high autocorrelation in the chain, maybe due to the high number of parameters in the algorithm; nevertheless convergence was quite fast and achieved after nearly $2,000$ iterations. We finally run a long chain with a burn-in period of $10,000$ iterations, and a thinning of 100 iterations to reduce autocorrelation, obtaining a final samples size $J = 1,000$ from the posterior joint distribution. The Bayesian estimates of α_1 and α_2, together with the 90% credible intervals, are presented in Table 4.5 for $m = 14.8480$ and the different values of γ_2. Figures from 4.11 to 4.13 show the tipical traces, the autocorrelation functions and the scatter plots of the chains $\{\alpha^{(j)} = (\alpha_1^{(j)}, \alpha_2^{(j)})\}_{j=1}^{J}$, for one choice of the hyperparameter.

We observe how the hyperparameter γ_2 works as a smoothing parameter in this example too. Indeed, assuming $\gamma_2 = 1$ (small) leads to a wave trend in the estimated survival function, that indicates a multimodality in the relative density (see Figure 4.8). In particular in this case, also the posterior joint density of $\alpha = (\alpha_1, \alpha_2)$ seems to have two modes. This can be imputed to the presence of influential points, as the 17th observation, in the data.

To evaluate the performance of the models considered we should check how well the model predicts. We consider a cross-validation method (see Gelfand, Dey, and Chang, 1992). Let $T_1, \ldots, T_n$ be a sample from the NPHM model (4.5), and let $t_1, \ldots, t_n$ be the observed values from the sample. In a cross-validation approach we want to check the observed non-censored survival time t_i against the predictive distribution, $f(\cdot|t_{(i)}, \mathbf{x}_i)$ arising from the model, all the observations $t_j,\ j \neq i$, and the covariates value of the ith patient $\mathbf{x}_i$. Actually, if the model holds, t_i could be viewed as a random observation from $f_{T_i}(\cdot|t^{(i)})$. To do this we considered $g(T_i; t_i) := T_i - t_i$ whose median under $f_{T_i}(\cdot|t^{(i)})$ has been calculated and denoted by $r_i := \text{Med}(T_i|t^{(i)}) - t_i$. We used the set of $\{r_i, i = 1, \ldots, n\}$, called *generalized residuals*, for model assessment (various possible choices of $g(\cdot;\cdot)$, called *checking functions*, are discussed in Gelfand *et al.*, 1992). We

considered

$$s_j := \text{Med}\left(\text{abs}\big(T_i - \text{Med}(T_i|t^{(i)}, \mathbf{x}_i)\big)\big|t^{(i)}, \mathbf{x}_i\right) \tag{4.27}$$

to standardize the generalized residuals by letting $r'_i := \frac{r_i}{s_i}$, for $i = 1, \ldots, n$. Then the quantity

$$\text{I} := \sum_{i=1}^{n} |r'_i| \tag{4.28}$$

can be considered as as index of the model fit.

This approach can be viewed as the Bayesian analogue to the well known frequentist strategy of examining the studentized residuals. We point out that usually in the Bayesian framework the residuals are computed through the conditional mean of the checking function $g(T_j; t_j)$. Since in our model the predictive distribution $f_{T_j}(\cdot|t^{(j)})$ does not admit mean, here we decided to use a median estimate.

We computed the predictive fit index (4.28) when $m = 14.8480$, $\gamma_2 = 1$ and $\mathbb{E}(k|n) = 6.5108$, obtaining $I_{NIG} = 108.97$ and $I_{DIR} = 103.06$, indicating a slightly better fit of the DPM model. The Figures 4.14, 4.15 and 4.16 show the plots of the standardaized residuals r'_i upon the continuous covariate $x_{i,2}$, $i = 1, \ldots, n$.

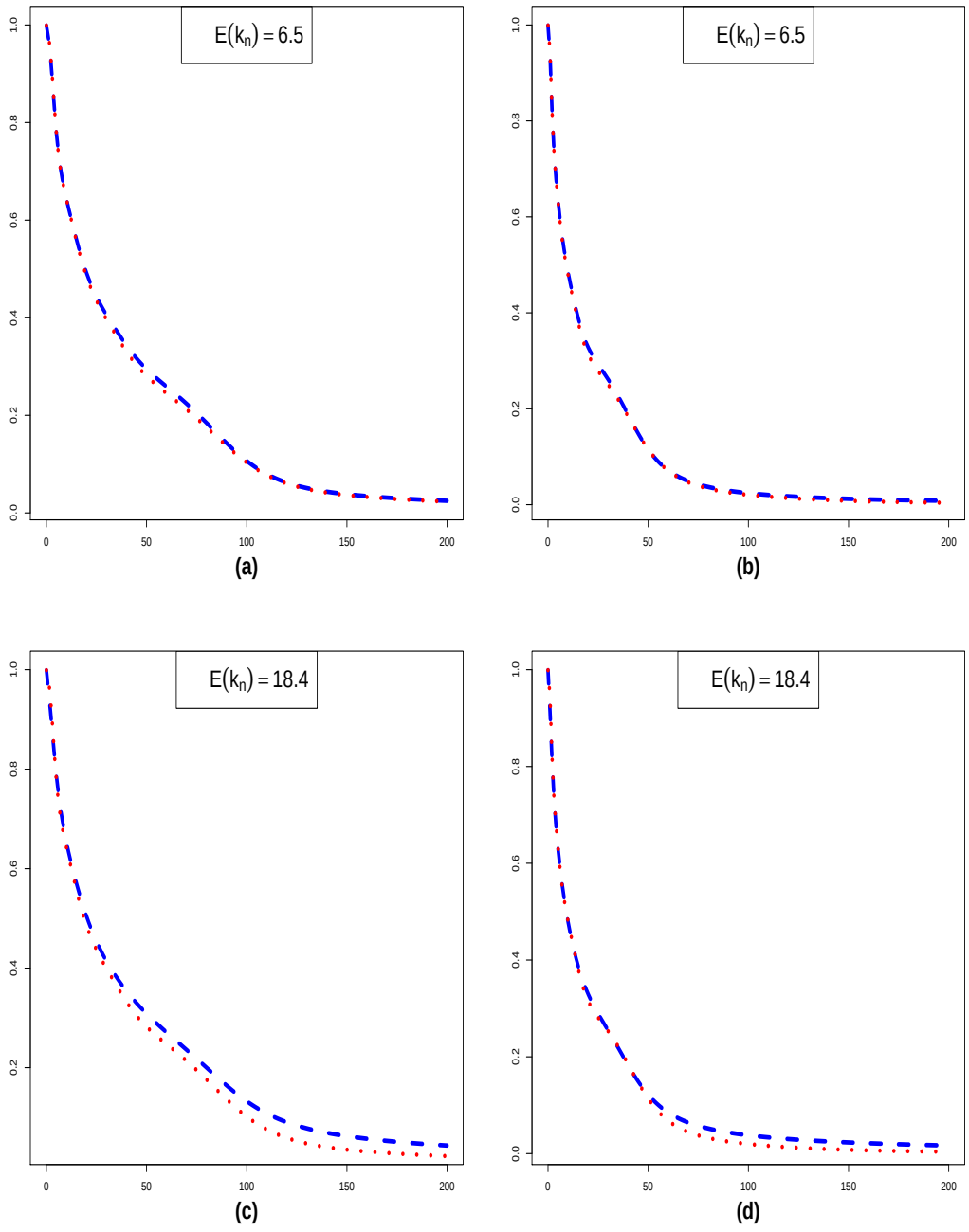

Figure 4.8: Estimated survival functions under the Dirichlet process prior (dotted red) and the N-IG prior (dashed blue) for 2 patients (covariate $(1, 0.5)$ in the left column and covariate $(0, 0.5)$ in the right column) from Example 2 when $\gamma_2 = 1$ and $m = 14.85$.

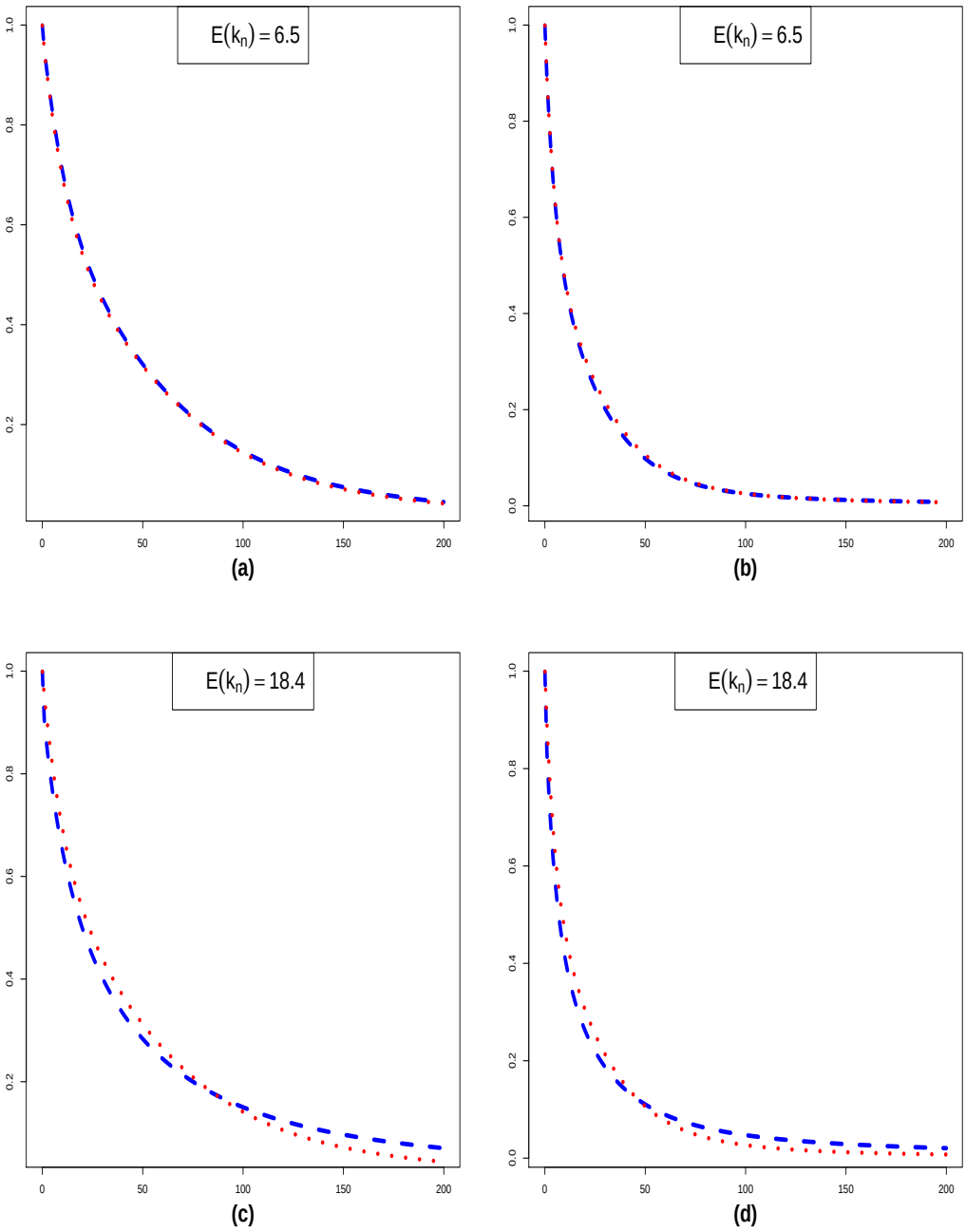

Figure 4.9: Estimated survival functions under the Dirichlet process prior (dotted red) and the N-IG prior (dashed blue) for 2 patients (covariate $(1, 0.5)$ in the left column and covariate $(0, 0.5)$ in the right column) from Example 2 when $\gamma_2 = 10$ and $m = 14.85$.

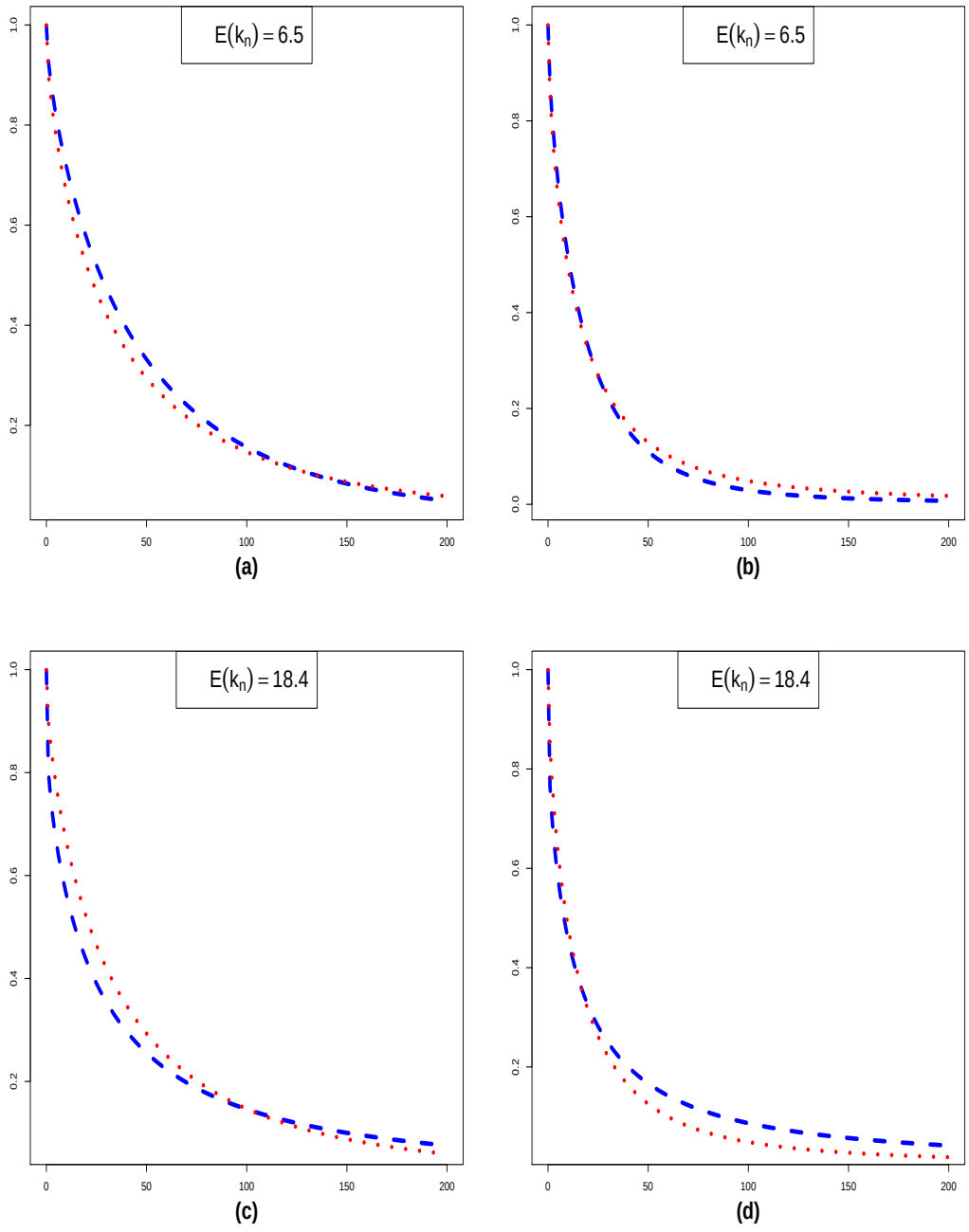

Figure 4.10: Estimated survival functions under the Dirichlet process prior (dotted red) and the N-IG prior (dashed blue) for 2 patients (covariate $(1, 0.5)$ in the left column and covariate $(0, 0.5)$ in the right column) from Example 2 when $\gamma_2 = 100$ and $m = 14.85$.

		Nig-mixture prior	
		M=0.01	**M**=10
γ_2	1	$\hat{\alpha}_1 = 0.5156\ (0.3318; 0.8661)$ $\hat{\alpha}_2 = 4.8386\ (1.4532; 15.7634)$	$\hat{\alpha}_1 = 0.4859\ (0.2497; 0.8056)$ $\hat{\alpha}_2 = 4.6024\ (1.3554; 16.0212)$
	10	$\hat{\alpha}_1 = 0.4052\ (0.1549; 0.7794)$ $\hat{\alpha}_2 = 8.6771\ (0.9026; 38.1015)$	$\hat{\alpha}_1 = 0.3984\ (0.1489; 0.7818)$ $\hat{\alpha}_2 = 10.2115\ (1.1436; 34.2550)$
	100	$\hat{\alpha}_1 = 0.4305\ (0.1967; 0.7867)$ $\hat{\alpha}_2 = 4.3092\ (0.9009; 10.3039)$	$\hat{\alpha}_1 = 0.6581\ (0.2342; 1.3701)$ $\hat{\alpha}_2 = 13.7599\ (1.6032; 46.6534)$

(a)

		MDP prior	
		a=2.1478	**a=** 16.3390
γ_2	1	$\hat{\alpha}_1 = 0.5107\ (0.3337; 0.8258)$ $\hat{\alpha}_2 = 5.3749\ (1.4407; 16.4946)$	$\hat{\alpha}_1 = 0.5070\ (0.3434; 0.7914)$ $\hat{\alpha}_2 = 5.3749\ (1.4403; 16.0518)$
	10	$\hat{\alpha}_1 = 0.4282\ (0.1557; 0.7722)$ $\hat{\alpha}_2 = 7.8721\ (1.0260; 21.4419)$	$\hat{\alpha}_1 = 0.4338\ (0.1555; 0.8062)$ $\hat{\alpha}_2 = 8.3606\ (1.0105; 30.8108)$
	100	$\hat{\alpha}_1 = 0.4980\ (0.2053; 0.9722)$ $\hat{\alpha}_2 = 10.5538\ (1.1902; 44.7378)$	$\hat{\alpha}_1 = 0.4851\ (0.1955; 0.9497)$ $\hat{\alpha}_2 = 10.1354\ (1.1850; 36.3112)$

(b)

Table 4.5: Estimates of α_1 and α_2, with 90% probability credible intervals, for the Feigl & Zelen dataset under the N-IG mixture (a) and DPM (b) priors.

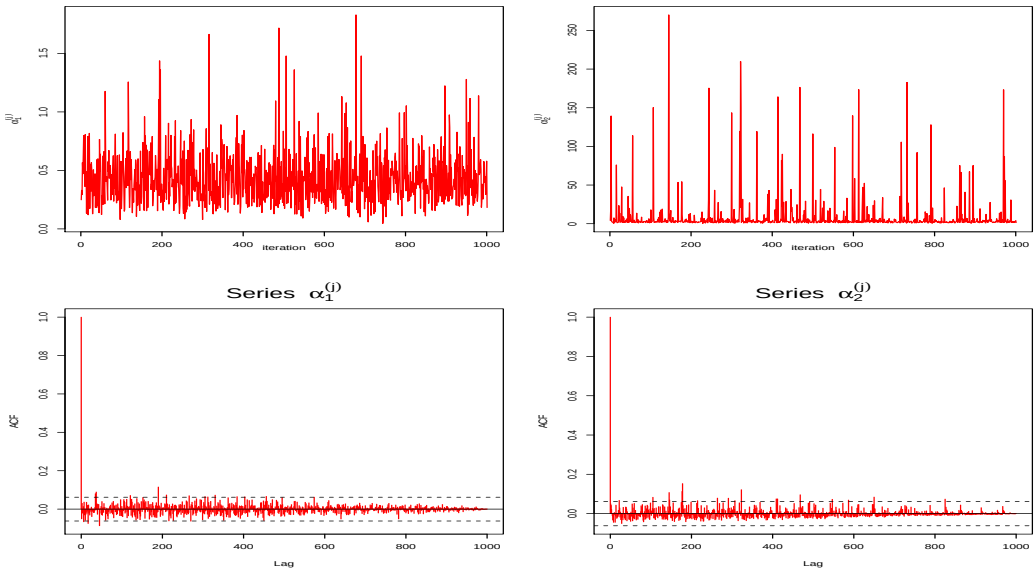

Figure 4.11: Traces and estimated autocorrelation functions of the Markov chain sample $\{\alpha_1^{(j)}, \alpha_2^{(j)}\}$ under the Dirichlet prior when $\mathbb{E}(k|n) = 18.4$, m=14.85 and $\gamma 2 = 10$

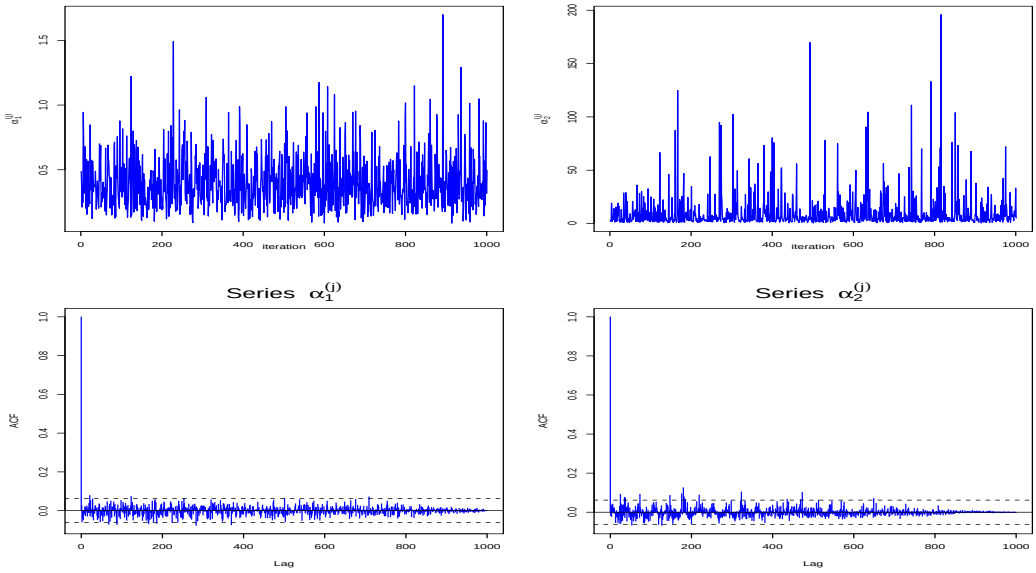

Figure 4.12: Traces and estimated autocorrelation functions of the Markov chain sample $\{\alpha_1^{(j)}, \alpha_2^{(j)}\}$ under the N-IG prior when $\mathbb{E}(k|n) = 18.4$, m=14.85 and $\gamma 2 = 10$

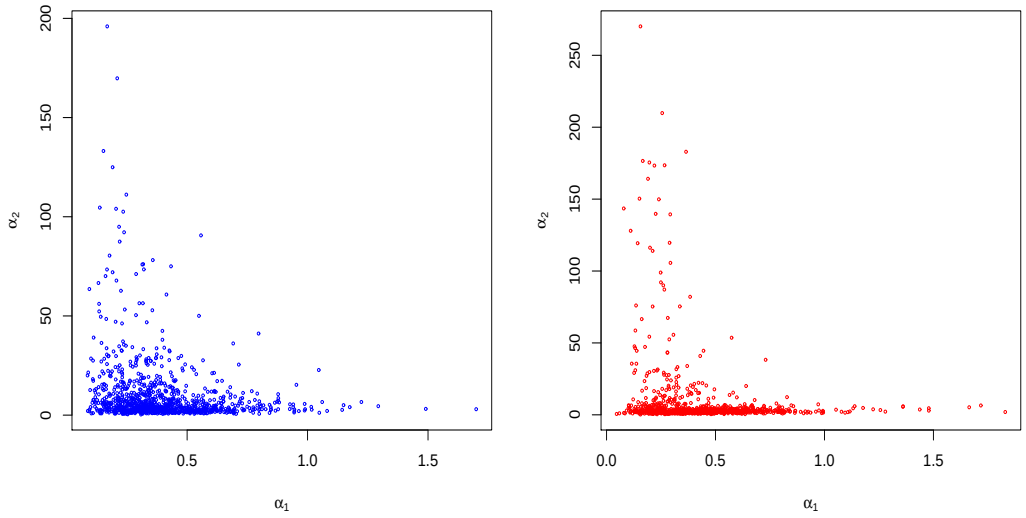

Figure 4.13: Scatter plots of the series $\{\alpha_1^{(j)}, \alpha_2^{(j)}\}_j$ under the N-IG process prior (left column blue) and Dirichlet process prior (right column red), when $\gamma_2 = 10$ and $m = 14, 84$. $\mathbb{E}(k|n) = 6,5$ in graph (a-b) and $\mathbb{E}(k|n) = 18.4$ in (c-d).

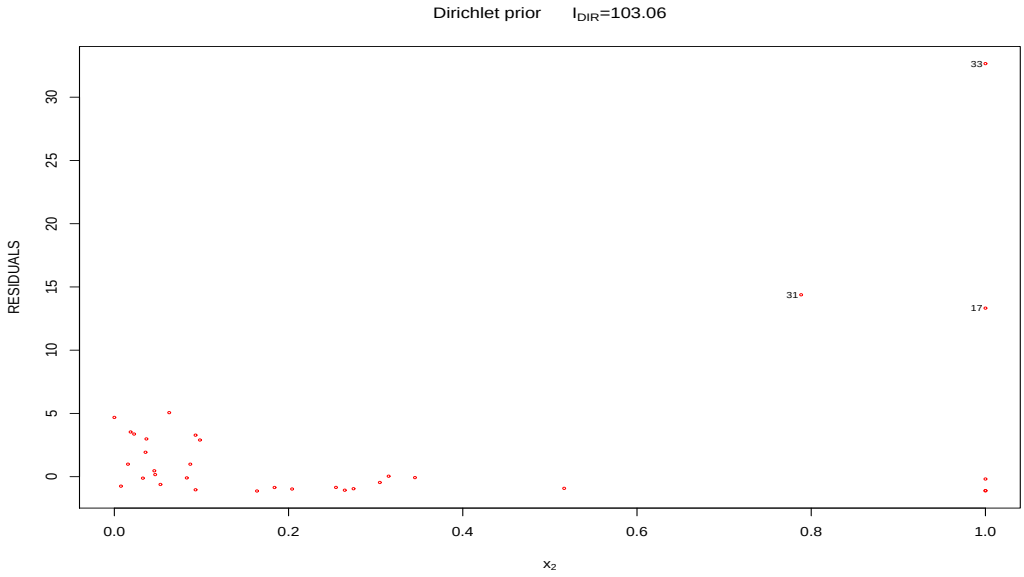

Figure 4.14: Standardized residuals, on the Feigl & Zelen data set, plotted respect to the continuous covariate $x_{\cdot 2}$, under the Dirichlet process prior. The Influent points are labelled with the identification number.$\gamma_2 = 1,\ m = 14.84$ and $\mathbb{E}(k|n) = 6.5$

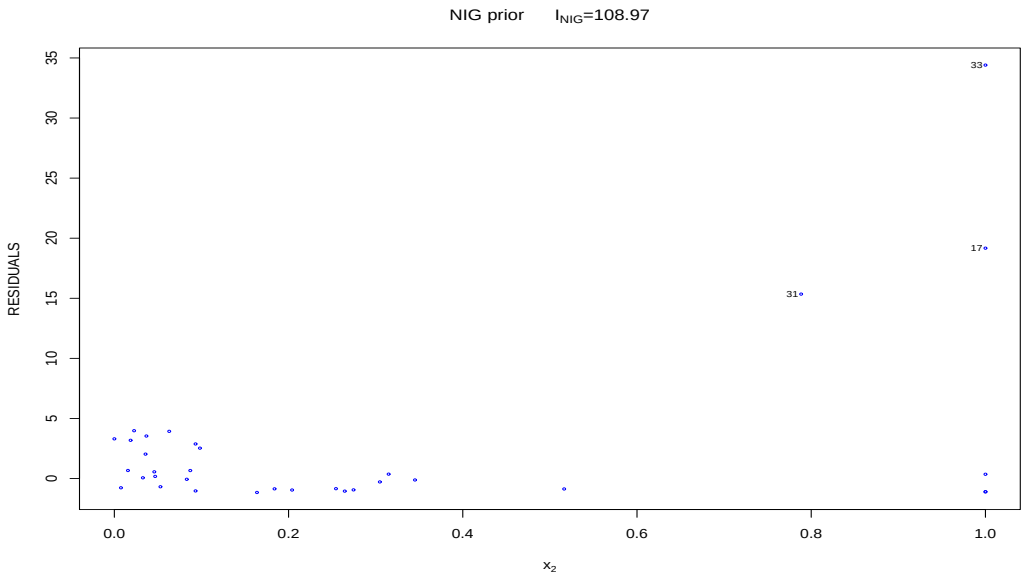

Figure 4.15: Standardized residuals, on the Feigl & Zelen data set, plotted respect to the continuous covariate $x_{\cdot 2}$, under the N-IG process prior. The Influent points are labelled with the identification number. $\gamma_2 = 1,\ m = 14.84$ and $\mathbb{E}(k|n) = 6.5$

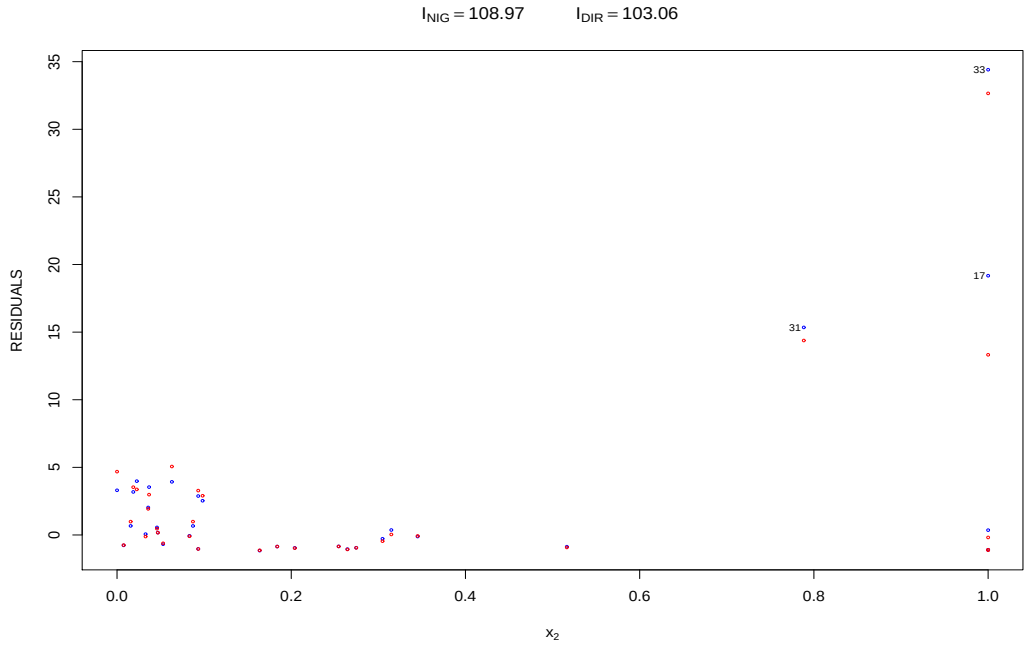

Figure 4.16: Standardized residuals against the continuous covariate under both the N-IG (blue) and Dirichlet (red) priors. $\gamma_2 = 1,\ m = 14.84$ and $\mathbb{E}(k|n) = 6.5$

4.7.3 Dataset involving censoring

As a third example, we considered 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, see Table 4.6, consist of $n = 121$ survival times in days 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 $\mathbf{x}_i = (x_{i1}, x_{i,2})'$, with $x_{i1} = 0$ if patient i was assigned to Treatment A, and $x_{i,2}$ denoting the patient's entry age.

As in the previous example, an explorative parametric analysis was performed, and the computed estimates were used in the nonparametric model specification. The prior median of the error variable V was assumed to be equal to the parametric estimate of the median $m = 2.4356$, and the prior on $\alpha = (\alpha_1, \alpha_2)$ was assumed to be a product of independent gamma, $\Gamma(g_j, h_j)$, $j = 1, 2$, such that $\mathbb{E}(\alpha_1) = 1.559$ and $\mathbb{E}(\alpha_2) = 1.113$, with an elevate variability, $\text{Var}(\alpha_j) = 1.000$, $j = 1, 2$. Figures 4.17, 4.18 and 4.19 display the estimated survival distributions, under the two semiparametric Bayesian models considered here, for 2 patients with covariates $(1, 36)$ and $(0, 36)$, for different values of the hyperparameters. The estimated survival functions are practically indistinguishable for small values of M (or $\mathbf{a}$), see figure 4.17. As before, a different behaviour on the tails arises when M $(\mathbf{a})$ increases, but unlike the previous example the predictive survival function under the Dirichlet prior have heavier tails than those under the N-IG prior. We argue that the N-IG process is particularly sensible to the presence of clusters in the data. In the Feigl-Zelen data set the presence of influential points with high survival times (see Figure 4.16) generates a cluster

that returns heavy tails in the survival estimates. This behaviour is accentuate when the N-IG is used as prior.

We observe also that, for high values of M or $\mathbf{a}$ (strong confidence in the prior), the estimated survival functions have an undesirable trend near zero (see for example Figure 4.19 (e) or (f)). This is imputable to the choice of the marginal prior of the error variable V, which has an asymptote in zero for each choice of the hyperparameters. Clearly this behaviour is emphasized when the confidence on the prior, quantified by the parameter M ($\mathbf{a}$), is high.

The survival estimates we just described, are based on a Markov sample $\{(\underline{\theta}^{(j)}, \alpha^{(j)}\}_{j=1}^{J}$ from the posterior joint distribution of $(\underline{\theta}, \alpha)$, obtained with the MCMC procedure in Section 4.6. As usual, we run several independent chains for each choice of hyperparameters. We noted that the autocorrelation in each sample was quite high, likely due to the presence of censored observations, that increases the number of parameters in the algorithm. Convergence, however, was sufficiently fast. We run long Markov chains with a burn-in period of $10,000$ iteration and a thinning of 100 observation. The final sample size was $J = 1,000$. Figures from 4.20 to 4.22 show the tipical traces, the autocorrelation functions and the scatter plots for the chains $\{\alpha^{(j)} = (\alpha_1^{(j)}, \alpha_2^{(j)})\}_{j=1}^{J}$ for one choice of the hyperparameters. The Bayesian estimates of α_1 and α_2, together with the 90% credible intervals, are presented in Table 4.7 for $m = 2.4356$ and different values of γ_2.

As before a cross-validation method to evaluate the predictive performances of the two models under study was set. Let $g(T_i, t_i) = T_i - t_i$ the checking function, the generalized residuals

$$r_i := \text{Med}(g(T_i, t_i)|t^{(i)}, \mathbf{x}_i),$$

can be computed only for the non-censored observation; therefore if S^* denotes the index set of non-censored data in the sample, we quantified the goodness in prevision of the models by the

Treatment A				Treatment B			
Entry age	Survival time (thousand days)	Entry age	Survival time (thousand days)	Entry age	Survival time (thousand days)	Entry age	Survival time (thousand days)
56	0.73	52	0.998	72	1.225	60	0.511
70	1.98^{+}	52	0.311	55	0.556	44	0.372
56	0.26	51	1.843^{+}	68	0.17	60	1.82^{+}
54	1.883^{+}	68	0.455	60	0.174	68	0.728
74	1.194	59	0.315	58	0.219	70	0.613
65	1.624^{+}	50	0.624	62	0.241	36	0.352
60	0.967	69	0.473	72	0.394	51	0.343
66	1.779^{+}	71	0.354	64	0.731	57	1.232
74	0.643	55	0.893	72	0.395	65	0.232
63	1.645^{+}	64	0.577	58	0.687	68	0.428
39	0.749	55	0.441	67	0.23	42	1.573^{+}
64	0.882	69	0.478	75	0.209	68	1.457^{+}
65	0.164	57	1.433^{+}	55	0.703	65	0.398
71	1.221	64	1.043	72	0.799	70	0.166
47	0.523	47	0.465	58	1.315	56	0.364
75	0.201	68	0.524	72	0.265	72	0.789
66	0.288	55	0.529	60	0.199	63	0.083
57	1.123^{+}	53	0.49	62	0.426	45	0.757
67	0.442	62	0.755	59	0.34	69	0.329
56	1.133^{+}	64	1.008	68	0.488	56	1.12^{+}
57	1.204^{+}	62	0.525	66	0.292	61	0.181
49	0.429	59	0.22	59	0.426	67	0.49
74	0.47	65	0.464	54	0.305	72	0.285
65	0.667	58	1.102^{+}	68	1.005^{+}	59	1.043^{+}
62	1.11^{+}	72	0.938	63	0.382	65	0.435
66	0.622	63	0.597	77	0.325	58	0.897^{+}
68	0.98^{+}	53	0.476	52	0.916^{+}	79	0.44
57	0.935	69	0.251	73	0.172	71	0.251
79	0.152	71	0.539	78	0.339	63	0.254
55	0.552	52	0.746^{+}	65	0.371		
52	0.256	54	0.835^{+}				

Table 4.6: Small cell lung cancer data. The superscript "+" indicates censoring

index

$$\mathrm{I}^{+} := \sum_{j \in S^*} \frac{|r_j|}{\tilde{s}_j}$$

where s_j is the index of dispersion introduced in (4.27). We computed the index I^{+} for both models for $m = 2.4356$, $\gamma_2 = 1$ and $\mathbb{E}(K|n) = 12.4107$, and obtained $I^{+}_{NIG} = 94.42$ and $I^{+}_{DIR} = 99.99$, indicating a slightly better fit of the N-IG mixture model for this dataset. Plots of the standardized generalized residual are shown in Figure 4.23, 4.24, and 4.25; the graphics indicate a good fit of both models, and no sign of curvature or heteroscedascity are present.

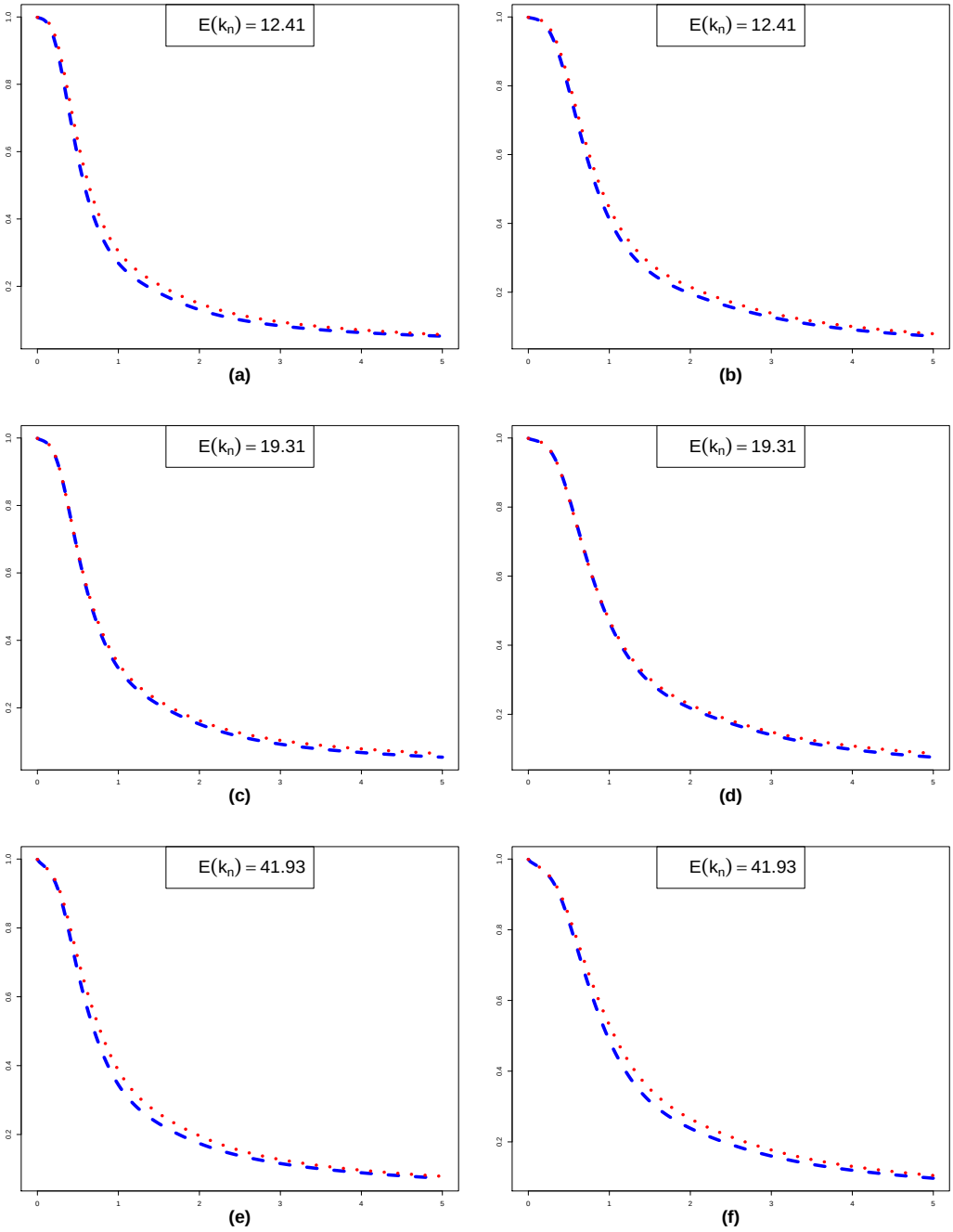

Figure 4.17: Estimated survival functions under the Dirichlet process prior (dotted red) and the N-IG prior (dashed blue) for 2 patients (covariate $(1, 36)$ in the left column and covariate $(0, 36)$ in the right column) from Example 3 when $m = 2.4356$ and $\gamma_2 = 0.1$.

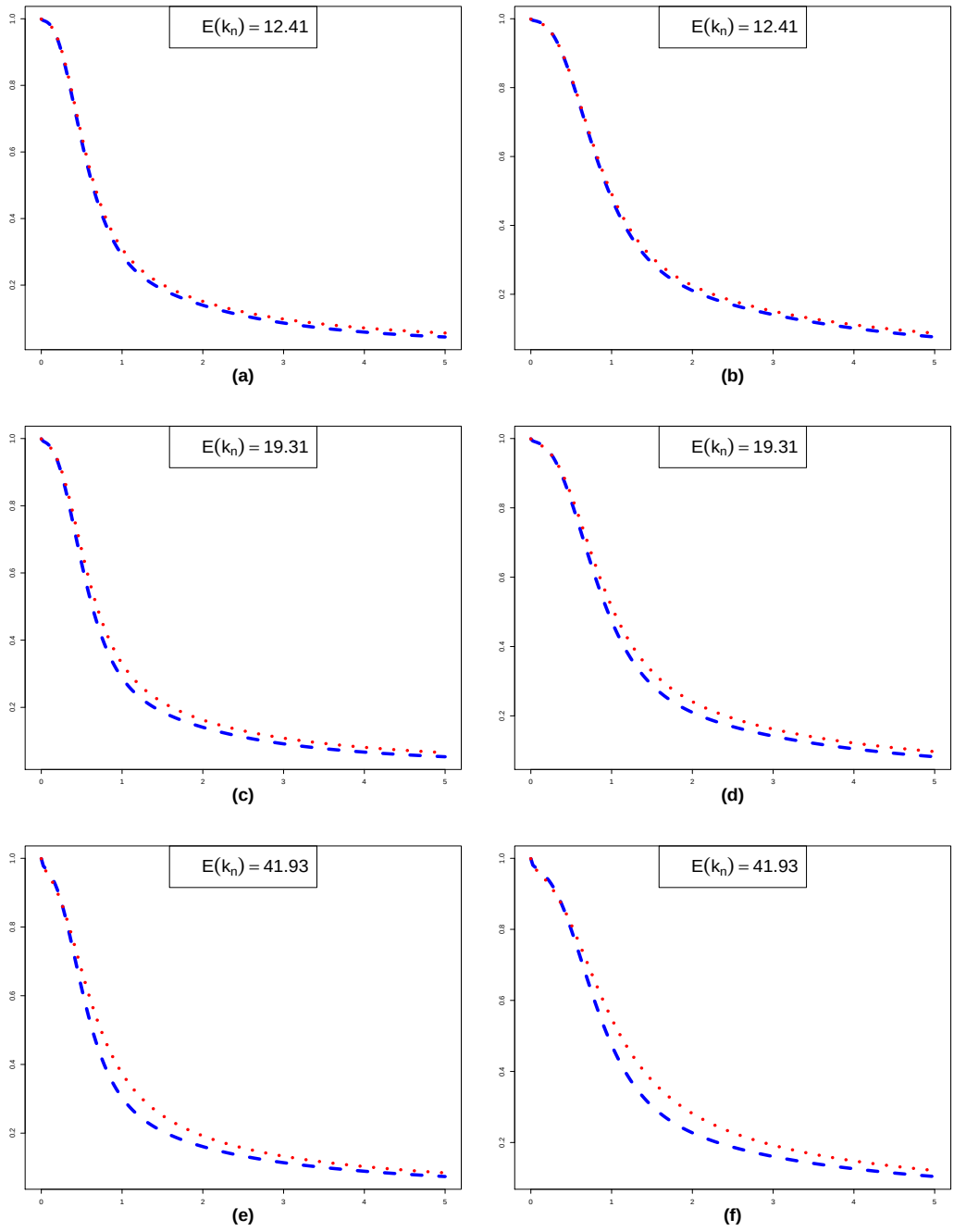

Figure 4.18: Estimated survival functions under the Dirichlet process prior (dotted red) and the N-IG prior (dashed blue) for 2 patients (covariate $(1, 36)$ in the left column and covariate $(0, 36)$ in the right column) from Example 3 when $m = 2.4356$ and $\gamma_2 = 1$

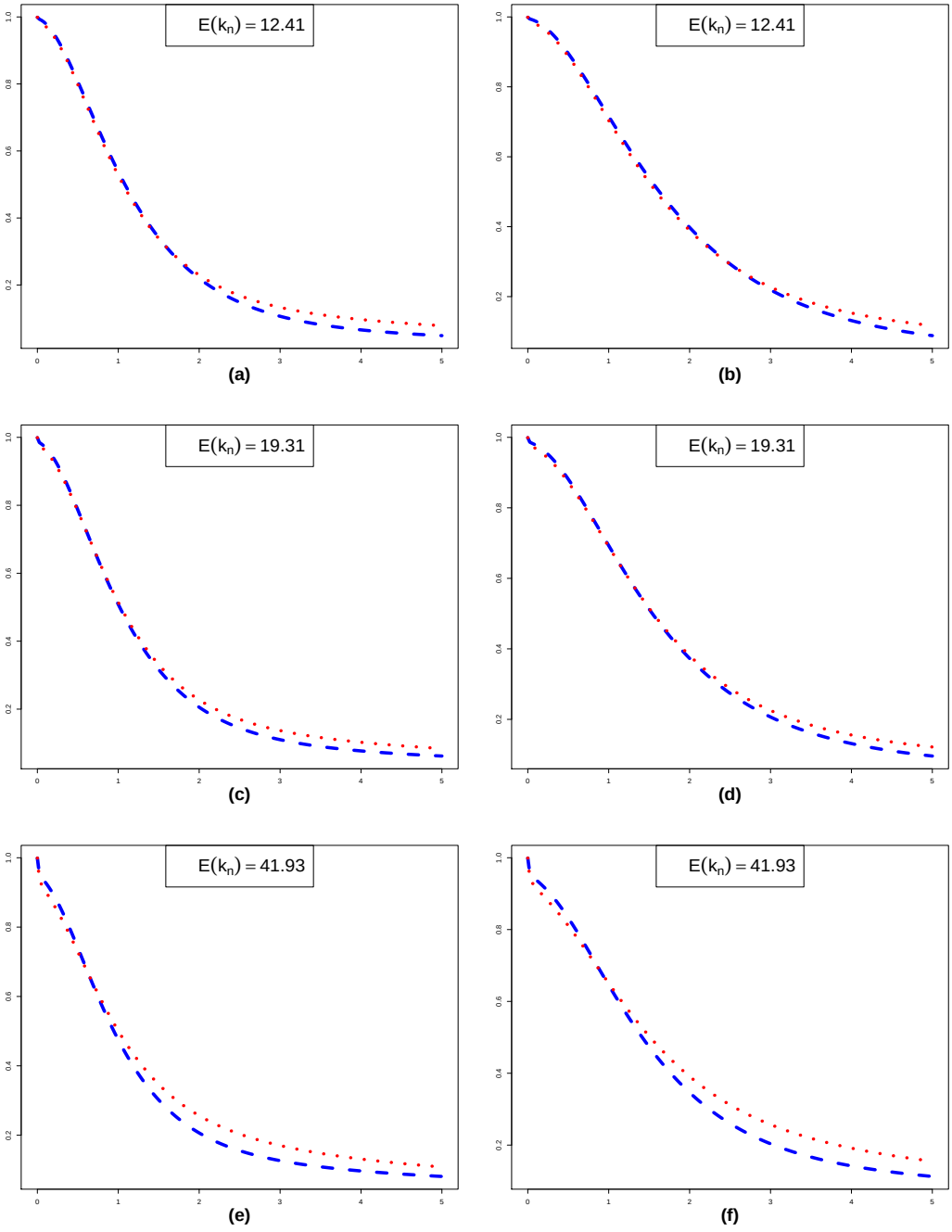

Figure 4.19: Estimated survival functions under the Dirichlet process prior (dotted red) and the N-IG prior (dashed blue) for 2 patients (covariate $(1, 36)$ in the left column and covariate $(0, 36)$ in the right column) from Example 3 when $m = 2.4356$ and $\gamma_2 = 10$

		N-IG mixture prior		
		M=0.001	**M**=1	**M**=10
γ_2	0.1	$\hat{\alpha_1} = 1.1492; (1.1814; 1.8193)$ $\hat{\alpha_2} = 1.0113; (0.9998; 1.0234)$	$\hat{\alpha_1} = 1.4209; (1.1238; 1.7623)$ $\hat{\alpha_2} = 1.0165; (1.0061; 1.0280)$	$\hat{\alpha_1} = 1.4087; (1.1145; 1.7820)$ $\hat{\alpha_2} = 1.0180; (1.0071; 1.0284)$
	1	$\hat{\alpha_1} = 1.5330; (1.2028; 1.9031)$ $\hat{\alpha_2} = 1.0150; (1.0047; 1.0248)$	$\hat{\alpha_1} = 1.5289; (1.2010; 1.8953)$ $\hat{\alpha_2} = 1.0149; (1.0051; 1.0248)$	$\hat{\alpha_1} = 1.5222; (1.1813; 1.9260)$ $\hat{\alpha_2} = 1.0152; (1.0060; 1.0239)$
	10	$\hat{\alpha_1} = 1.5210; (1.0221; 1.0418)$ $\hat{\alpha_2} = 1.0318; (1.0221; 1.0418)$	$\hat{\alpha_1} = 1.5311; (1.1676; 1.9406)$ $\hat{\alpha_2} = 1.0307; (1.0207; 1.0413)$	$\hat{\alpha_1} = 1.5155; (1.1584; 1.9667)$ $\hat{\alpha_2} = 1.0290; (1.0206; 1.0387)$

(a)

		MDP prior		
		a=3.2713	**a**= 6.2467	**a**=22.3172
γ_2	0.1	$\hat{\alpha_1} = 1.4365; (1.1492; 1.7719)$ $\hat{\alpha_2} = 1.0141; (1.0032; 1.0251)$	$\hat{\alpha_1} = 1.4086; (1.1229; 1.7532)$ $\hat{\alpha_2} = 1.0163; (1.0049; 1.0279)$	$\hat{\alpha_1} = 1.3867; (1.0961; 1.7613)$ $\hat{\alpha_2} = 1.0211; (1.0126; 1.0288)$
	1	$\hat{\alpha_1} = 1.5149; (1.1741; 1.8865)$ $\hat{\alpha_2} = 1.0160; (1.0070; 1.0253)$	$\hat{\alpha_1} = 1.5123; (1.1922; 1.8694)$ $\hat{\alpha_2} = 1.0176; (1.0090; 1.0267)$	$\hat{\alpha_1} = 1.5295; (1.1564; 1.9838)$ $\hat{\alpha_2} = 1.0199; (1.0122; 1.0280)$
	10	$\hat{\alpha_1} = 1.5178; (1.1499; 1.9471)$ $\hat{\alpha_2} = 1.0324; (1.0221; 1.0430)$	$\hat{\alpha_1} = 1.5214; (1.1277; 1.9425)$ $\hat{\alpha_2} = 1.0318; (1.0227; 1.0413)$	$\hat{\alpha_1} = 1.5369; (1.1226; 2.0492)$ $\hat{\alpha_2} = 1.0331; (1.0245; 1.0423)$

(b)

Table 4.7: Estimates of α_1 and α_2, with 90% probability credible intervals, for the small-cell lung cancer dataset under the N-IG mixture (a) and MDP (b) priors, when $m = 2.4356$.

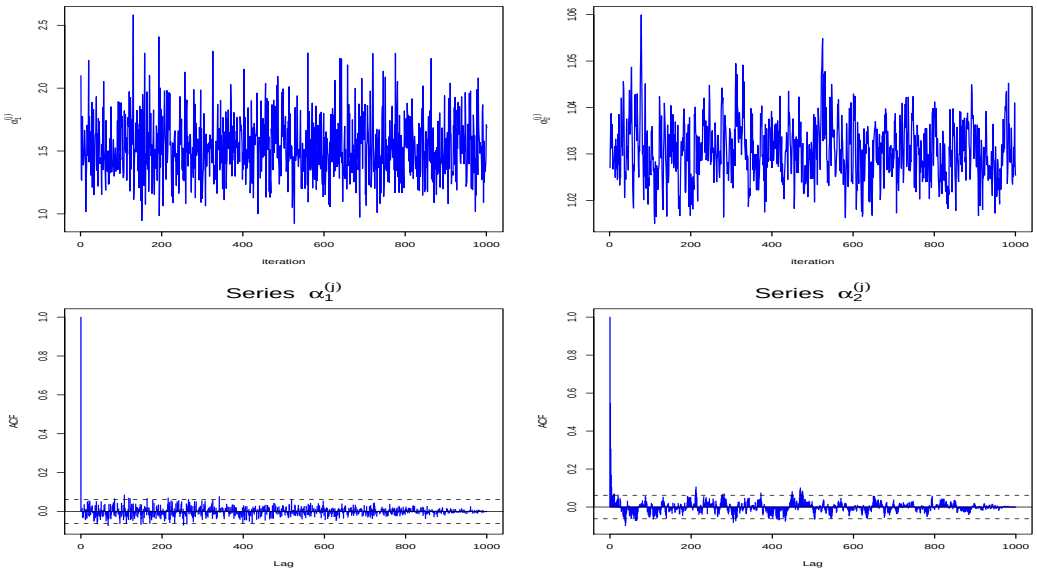

Figure 4.20: Traces and estimated autocorrelation functions of the Markov chain sample $\{\alpha_1^{(j)}, \alpha_2^{(j)}\}$ under the N-IG process prior when $\mathbb{E}(k|n) = 19.31$, m=2.44 and $\gamma 2 = 10$.

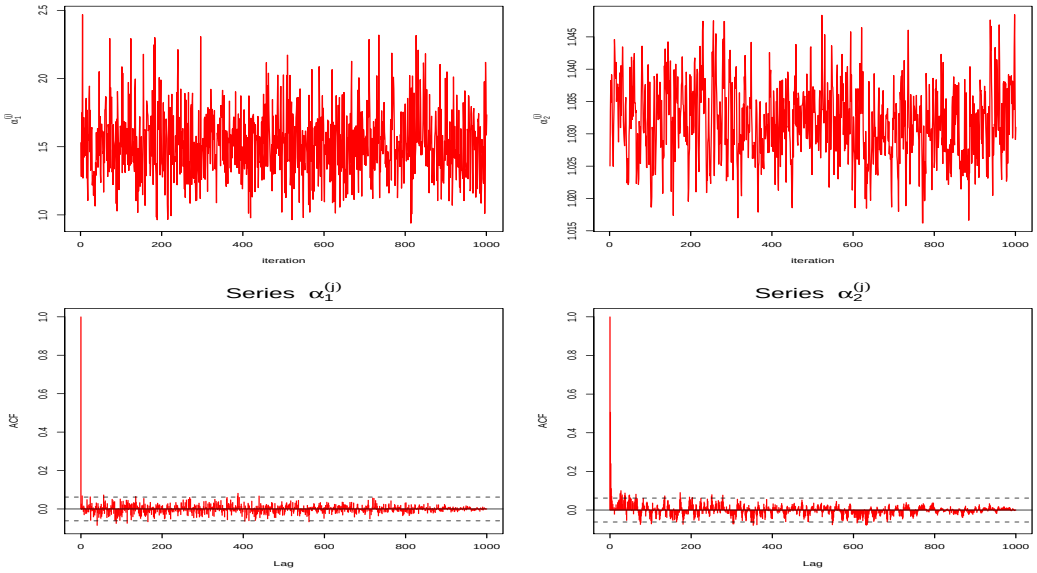

Figure 4.21: Traces and estimated autocorrelation functions of the Markov chain sample $\{\alpha_1^{(j)}, \alpha_2^{(j)}\}$ under the Dirichlet process prior when $\mathbb{E}(k|n) = 19.31$, m=2.44 and $\gamma 2 = 10$.

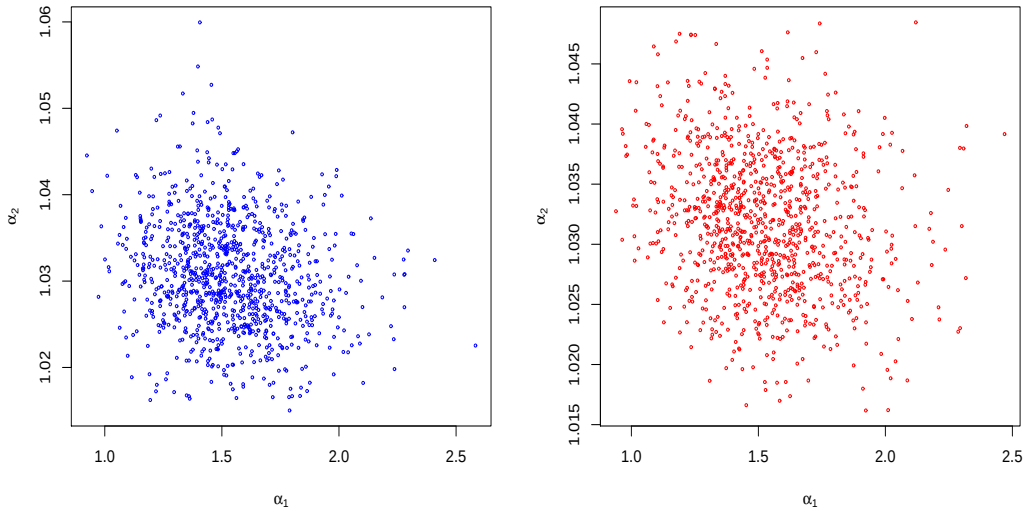

Figure 4.22: Scatter plots of the series $\{\alpha_1^{(j)}, \alpha_2^{(j)}\}_j$ under the N-IG process prior (left column blue) and Dirichlet process prior (right column red). $\gamma_2 = 10,\ m = 2.44$ and $\mathbb{E}(k|n) = 19.31$

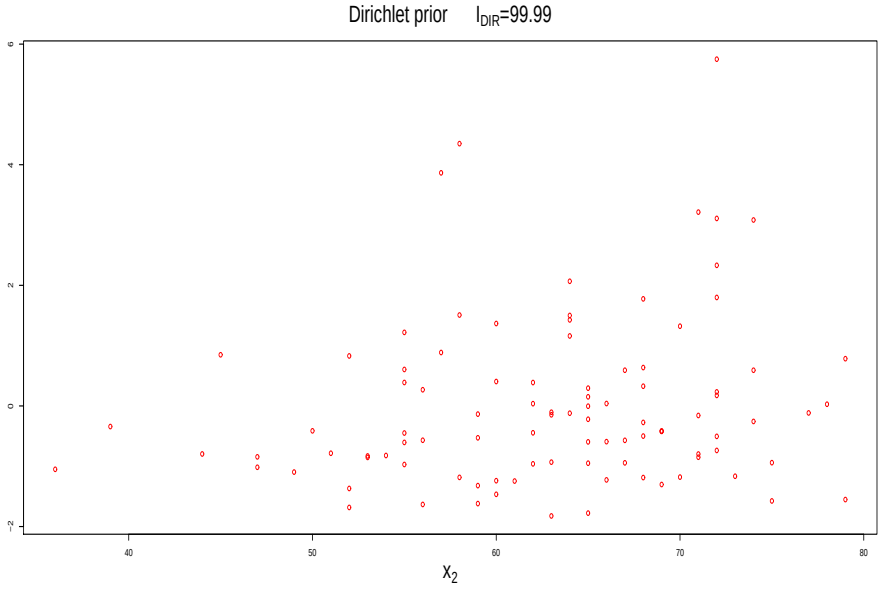

Figure 4.23: Standardized residuals, from the third Example, plotted respect to the continuous covariate $x_{\cdot 2}$, under the Dirichlet process prior. $\gamma_2 = 1,\ m = 2.43$ and $\mathbb{E}(k|n) = 12.41$

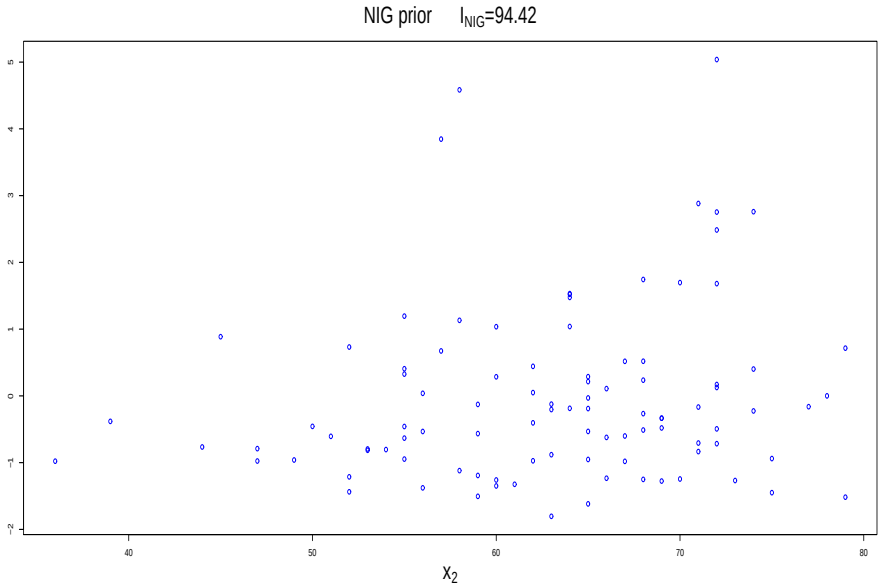

Figure 4.24: Standardized residuals, from the third Example, plotted respect to the continuous covariate $x_{\cdot 2}$, under the N-IG process prior. $\gamma_2 = 1,\ m = 2.43$ and $\mathbb{E}(k|n) = 12.41$

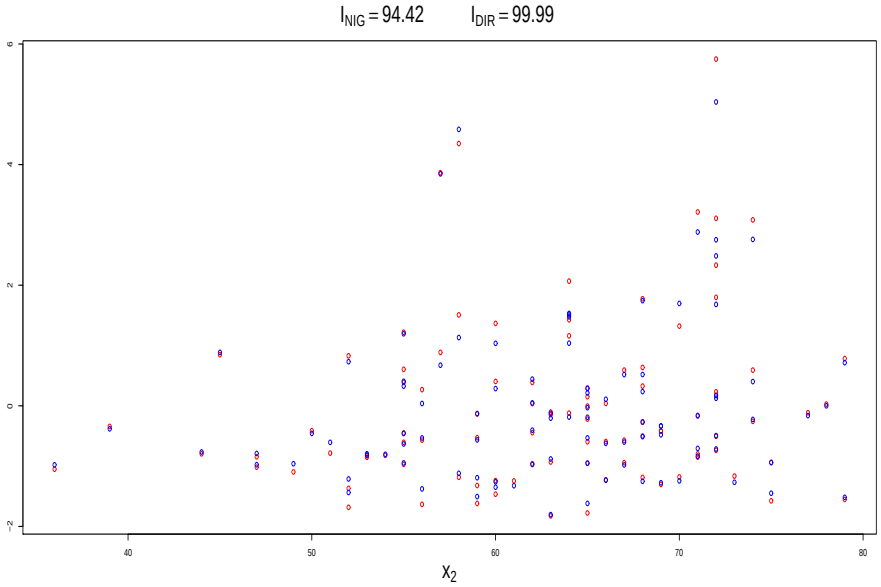

Figure 4.25: Standardized residuals against the continuous covariate under both the N-IG (blue) and Dirichlet (red) priors. $\gamma_2 = 1,\ m = 2.44$ and $\mathbb{E}(k|n) = 12.41$

4.8 A non conjugate model

In the previous sections we observed that the marginal prior of the error term in the AFT regression model, given by

$$\begin{aligned} f_V(v) &= \int_0^{+\infty} d\vartheta_1 \int_0^{+\infty} d\vartheta_2 \frac{\vartheta_2^{\vartheta_1}}{\Gamma(\vartheta_1)} v^{\vartheta_1 - 1} \mathrm{e}^{-\vartheta_2 v} \gamma_1 \mathrm{e}^{-\gamma_1 \vartheta_1} \gamma_2 \mathrm{e}^{-\gamma_2 \vartheta_2} \\ &= \frac{\gamma_1 \gamma_2}{v(v+\gamma_2)(\gamma_1 + \log(\frac{v+\gamma_2}{v}))^2}, \end{aligned}$$

is not flexible enough. Indeed, it is monotone with an asymptote in zero for each choice of the hyperparameters. In practice, in spite of the knowledge we have on lifetimes, we are not able to introduce all prior informations in to the model. As seen in section 4.7.1, this can produce an undesirable behaviour in posterior estimates (see figure 4.5). To obtain a more flexible prior, we propose to change the mean distribution $G_0(\cdot)$ of the non parametric prior, both in the N-IG and the Dirichlet cases. The hierarchical model for the variable $V_1, \ldots, V_n$ is:

$$\begin{aligned} & V_i|\theta_i \overset{ind}{\sim} k(\cdot|\theta_i), \\ & \theta_i|\mathbf{G} \overset{iid}{\sim} \mathbf{G}, \\ & \mathbf{G} \sim q, \quad G_0(A) := \mathbb{E}_q(\mathbf{G}(A)),\ A \in \mathcal{B}(\Theta) \end{aligned} \tag{4.29}$$

where $\{k(\cdot|\theta), \theta \in \Theta\}$ is a family of gamma kernels with $\theta = (\vartheta_1, \vartheta_2)$ and mean ϑ_1/ϑ_2. We let G_0 be the product of two independent gamma distributions, i.e., in the hierarchical specification, we choose $\theta = (\vartheta_1, \vartheta_2)$ such that, ϑ_1 and ϑ_2 under G_0 are independent gamma distributed with parameter (ω_1, γ_1) and (ω_2, γ_2) respectively. The new marginal prior for the variable V is:

$$\begin{aligned} f_V(v) =& \int_0^{+\infty} d\vartheta_1 \int_0^{+\infty} d\vartheta_2 \frac{\vartheta_2^{\vartheta_1}}{\Gamma(\vartheta_1)} v^{\vartheta_1 - 1} \mathrm{e}^{-\vartheta_2 v} \frac{\gamma_1^{\omega_1}}{\Gamma(\omega_1)} \vartheta_1^{\omega_1} \mathrm{e}^{-\gamma_1 \vartheta_1} \frac{\gamma_2^{\omega_2}}{\Gamma(\omega_2)} \vartheta_2^{\omega_2} \mathrm{e}^{-\gamma_2 \vartheta_2} \\ =& \frac{\gamma_1^{\omega_1} \gamma_2^{\omega_2}}{v(v+\gamma_2)^{\omega_2}(\gamma_1 + \log(\frac{v+\gamma_2}{v}))^{\omega_1}} \cdot \int_0^{+\infty} d\vartheta_1 \frac{\Gamma(\vartheta_1 \cdot \omega_2)}{\Gamma(\vartheta_1)} \Gamma(\vartheta_1|s, r(v)), \end{aligned} \tag{4.30}$$

where $\Gamma(\cdot|s,r)$ is the density of a gamma distributed random variable with shape parameter s and rate parameter r, and $r(v) = \gamma_1 + \log(1+\gamma_1/v)$. Distribution (4.30) is more flexible than distribution (4.7), it has an asymptote in zero, but for $\omega_2 > 1$ it admits a mode. In figure 4.26 the graphics of $f_V(\cdot)$ for some choices of the hyperparameters are depicted. We can see that for ω_1 and ω_2 big enough the asymptote of $f_V(\cdot)$ becomes negligible. The new marginal prior also admits $k-$th moment for $\omega_2 > k$. Indeed, if $V \sim f_V(\cdot)$ then $\mathbb{E}(V^k) = \mathbb{E}(\mathbb{E}(V^k|\vartheta_1,\vartheta_2)) = \mathbb{E}(\vartheta_1^k)\mathbb{E}(1/\vartheta_2^k)$, and the second term of the last product exist if and only if $\omega_2 > k$. In particular

$$\mathbb{E}(V) = \frac{\omega_1\gamma_2}{(\omega_2-1)\gamma_1}, \;\; \omega_2 > 1. \tag{4.31}$$

We derive the induced mean and variance of the gamma components in the mixture: let $V|(\vartheta_1,\vartheta_2) \sim \Gamma(\vartheta_1,\vartheta_2)$ indicate a gamma distributed random variable with mean $\mu = \vartheta_1/\vartheta_2$ and variance $\sigma^2 = \vartheta_1/\vartheta_2^2$. With $(\vartheta_1,\vartheta_2) \sim \Gamma(\omega_1,\gamma_1) \times \Gamma(\omega_2,\gamma_2)$, we have that $\mathbb{E}(\mu) = \mathbb{E}(V)$, and

$$\mathbb{E}(\sigma^2) = \frac{\gamma_2}{\omega_2-2}\mathbb{E}(V), \;\; \omega_2 > 2. \tag{4.32}$$

Finally we compute

$$\text{Var}(V) = \left(\frac{\omega_2+\omega_1-1}{(\omega_2-1)\gamma_1}+1\right)\mathbb{E}(\sigma^2), \;\; \omega_2 > 2. \tag{4.33}$$

In this way we can use equations (4.31), (4.32) and (4.33) to express the prior information on the problem at hand.

4.8.1 A numerical example

We performed a nonparametric density estimate of the same sample $t_1,\ldots,t_n$ of size $n = 100$ from the mixture density (4.24), used in section 4.7.1. We centred the marginal prior at the popula-

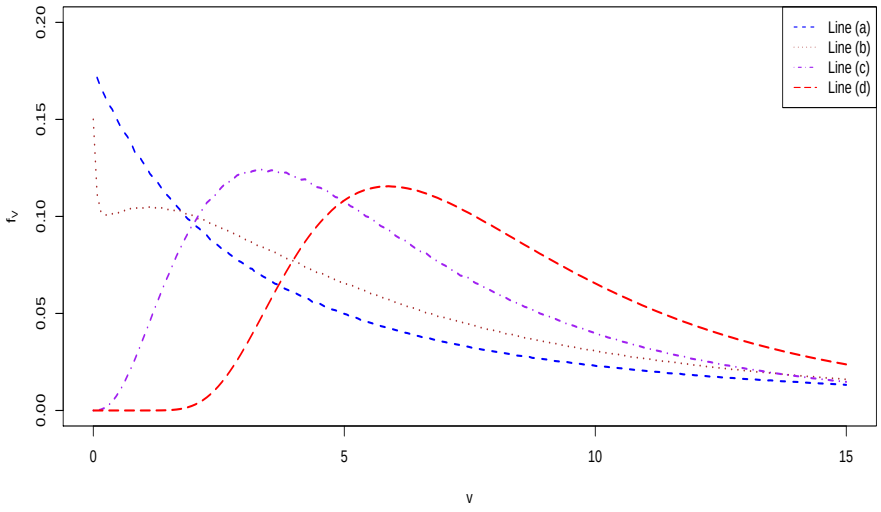

Figure 4.26: Graphics of the marginal prior (4.30) of the variable V of model (4.29) for some choices of the hyperparameters. (a) $\omega_1 = 1,\ \omega_2 = 1,\ \lambda_1 = 0.002,\ \lambda_2 = 0.01$; (b) $\omega_1 = 3,\ \omega_2 = 2,\ \lambda_1 = 1,\ \lambda_2 = 4$; (c) $\omega_1 = 4,\ \omega_2 = 4,\ \lambda_1 = 0.007,\ \lambda_2 = 0.04$; (d) $\omega_1 = 149,\ \omega_2 = 4,\ \lambda_1 = 1,\ \lambda_2 = 0.2$;

tion mean, choosing the hyperparameters so that $\mathbb{E}(V)$=6. To have a benchmark with the estimate obtained under the conjugate model we fixed the dispersion of V through the width L_α of the $(\alpha\cdot)100\%$ prior probability interval determined in Section 4.4, using the approximation arising from the Gaussian distribution:

$$\sqrt{\text{Var}(V)} \approx \frac{L_\alpha}{2 \cdot z_{1-\alpha/2}}$$

where $z_{1-\alpha/2}$ is the $(1-\alpha/2)$ percentile of the Gaussian distribution. Moreover we used $\mathbb{E}(\sigma^2) = 5, 10$ and 50 as a bandwidth parameter, controlling the dispersion of the gamma-components of the nonparametric mixture. We gave a strong confidence to the marginal prior (i.e, we chose $M = 5.39$ and $\mathbf{a} = 14.1614$) because the undesirable behaviour near the origin of the predictive densities under the conjugate model (clearly) was more pronounced in this case.

For each choice of the hyperparameters we ran several chains, using a data augmentation Gibbs sampler algorithm as described in section 3.3, with an auxiliary vector of size 3. The convergence

was relatively fast, and we did not observe a significant worsening with respect to the conjugate algorithm. Then we run a long chain for each model, with a burn-in of 10,000 iteration and a thinning of 50 iterations.

As in the conjugate case the predictive performances under the N-IG and the Dirichlet prior are equivalent. In the left column of Figure 4.27 we depicted the predictive distributions for each choice of hyperparameters, and in Table 4.8 we present the distance in the uniform metric between the "true" distribution and the predictive ones. Under the non conjugate model (with an appropriate choice of hyperparameter) the posterior densities have the desired trend near the origin. Although produce indistinguishable predictive densities, the two models make use of a different clustering of the elements $\underline{\theta}$. The right column of figure 4.27 shows the posterior estimates of the distribution of the number of clusters k. We can see that when $\mathbb{E}(\sigma^2)$ takes the smallest value, the estimate under the N-IG process tends to use a greater number of components with respect to those used in the estimate under the Dirichlet process prior. The opposite trend is observed when the prior mean of the variance of the mixture component is set at the largest value. The N-IG process seems to mitigate the influence of the choice of the total mass given to the "mean" measure, but it is more susceptible to change in the prior information given by the hyperparameters ω and γ.

$\mathbb{E}(k\|n) = 30$	$\mathbb{E}(\sigma^2)$		
	2	10	50
N-IG	0.0508	0.0485	0.0642
Dir	0.0682	0.0647	0.0754

Table 4.8: Errors in the uniform metric for the simulated dataset of size 100 between the true and estimated distribution functions, under the non conjugate model.

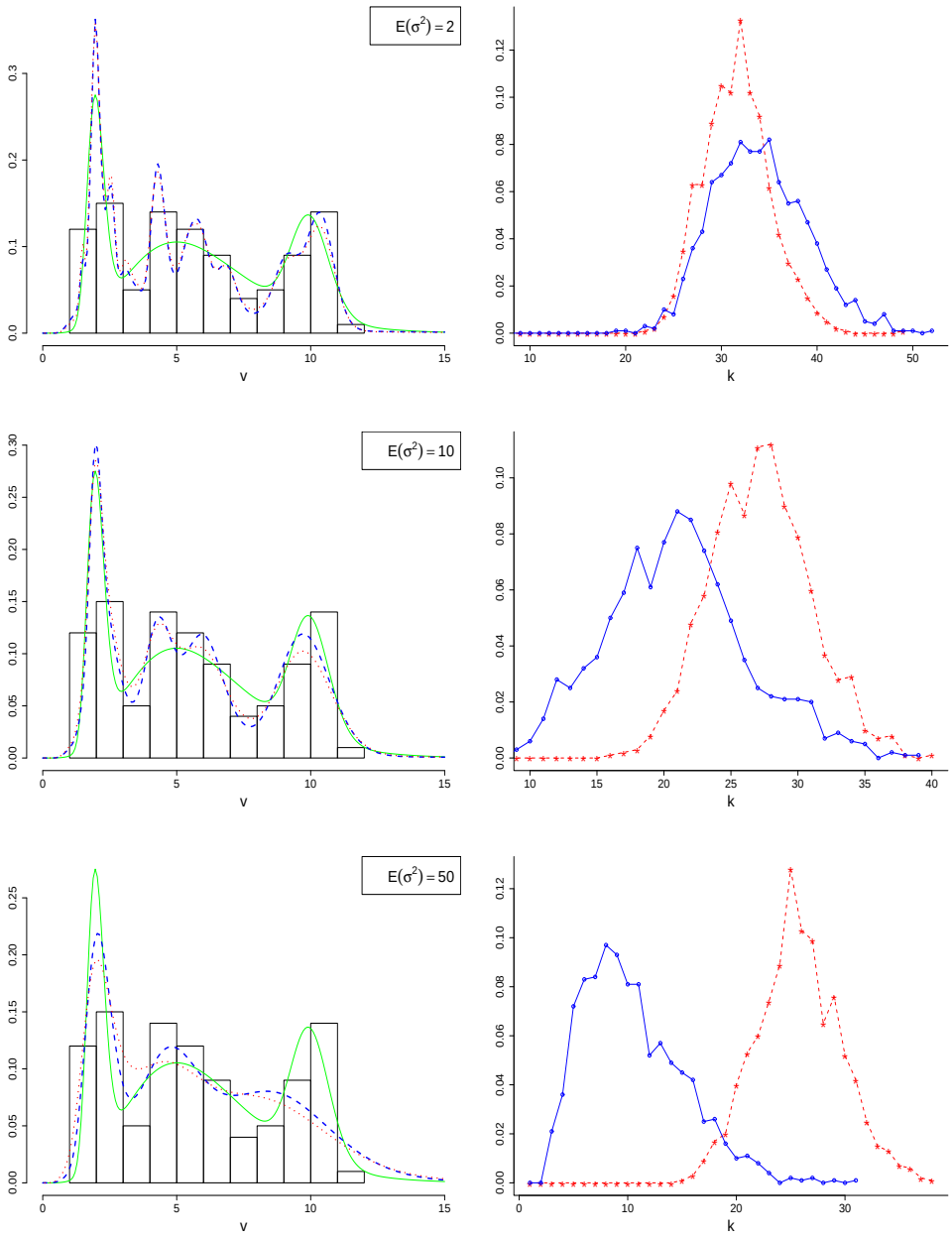

Figure 4.27: Histogram for the simulated data and density estimates under the Dirichlet process prior (dotted red) and the N-IG prior (dashed blue) in the left column. Posterior estimates of the distribution of the number of cluster under the Dirichlet prior (dotted red) and N-IG prior (solid blue) in in the right column. The hyperparameters are such that $\mathbb{E}(V) = 6$, $\text{Var}(V) = 32.78^2$ and $\gamma_2 = 2$. The green solid line denotes the true density.

4.9 Conclusions

In this Thesis we studied the accelerated failure time models in the context of Bayesian semiparametric statistics. In the first two chapters we reviewed the main contributions in this area, and we surveyed on nonparametric Bayesian modelling. In particular we focused the attention on nonparametric hierarchical mixture models, with the goal of comparing the performance of the well known DPM model with N-IG process mixture model, recently introduced in literature. The two competing models have been tested looking at the predictive estimates they produce. The comparison between is carried out in the following way: first we fixed the prior hyperparameters, in such a way they carry as similar a prior information as possible, then we quantified the predictive performances. With the simulated data sets we measured the distance between the predictive and the "true" distribution. With the real data sets we used a *cross validation* procedure to quantify the predictive power of each model.

The estimation was carried out by MCMC simulation methods. In the third chapter we reviewed the mains strategies to handle DPM models. In particular we focused on algorithms that rest on the predictive structure of the Dirichlet process prior. Then we described how to extend this algorithm to the N-IG mixture models, for both the conjugate and the non-conjugate case.

The fourth chapter contains the main results of the work. We described in depth the hierarchical models adopted, and we matched the two nonparametric priors by centering the mixing distributions at the same "mean" G_0 and assuming an equal prior mean of the number of components in the mixture. Then the prior information on the survival times was passed to the model through functionals of the marginal prior.

We presented three examples using a conjugate hierarchical model: the first one on density estimation (AFT with a null vector of covariates) with simulated data; the second one on regression for uncensored survival times; the third one on regression for right-censored data. Finally a non-conjugate hierarchical model is tested on density estimation with simulated data.

A close inspection of the marginal error distribution leads to to think that some artifacts near the origin of the predictive density are due to the choice of the kernels and the "mean" G_0. Indeed the analysis of a non conjugate model, with a more general G_0 eliminates this drawback.

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 in mixture density estimation. However some differences arise in the posterior distribution of the number of clusters the two models use to "built" the predictive estimates: the more elaborate prior reinforcement structure of the N-IG process prior leads to estimates of the number of clusters in the mixture that are difficult to interpret. Indeed this posterior, for both the models, not only depends on the prior total mass of the mixing distribution, it also depends on the hyperparameters within the mean distribution of the non parametric prior processes. In our experiment, with an appropriate choice of these hyperparameters, the N-IG process prior seems to be better than the DPM prior in finding clustering structures in the data. Nevertheless we believe that some more though is needed to give a good statistical interpretation of this characteristic.

In the first regression example (see Section 4.7.2) the model with the Dirichlet process prior error fits the data a little better. However, the second experiment, on the Feigl and Zelen (1965) leukemia dataset (see Section 4.7.3), gives the opposite result. In both cases we notice that the difference in the predictive fit indexes is not dramatic, and it is difficult to choose between the two process priors based only on this. This uncertainty remains even if we consider that the result on the Feigl and Zelen dataset is influenced by a few abnormal observations, whose removal produces essentially equal predictive fit indexes.

A similar conclusion worth looking at the concerns posterior distribution of the regression parameter α. We can see from Table 4.5 and Table 4.7 as, the point and the interval estimates obtained under the two competing priors do not differ in meaningful way.

The Dirichlet process prior is, maybe, the most studied nonparametric prior in Bayesian statis-

tics. One of the reasons of this popularity is its "relative" simplicity. The Pólya urn and the "stick-breaking" representation (see Section 2.2.1) constitute, for example, a very useful tools to work with. Furthermore, in the recent years the Dirichlet process prior has experienced a great success in the context of Bayesian mixture modelling. The idea of overcoming the discreteness of its realizations by exploiting it in hierarchical models, combined with the development of suitable sampling techniques, constitutes one of the reasons of its popularity.

The NIG prior represents a valid alternative to the Dirichlet prior in the NPHM models. Indeed, it preserves almost the same tractability and has an interesting clustering property that makes use of all the information contained in the data,in a predictive sense.

Future extensions of the work will focus on models with one more level in the hierarchical structure, introducing some distributions for the total mass parameters a and M, which determine the prior distribution of the number of components in the mixture, so as to obtain a refined estimate of this distribution. Other extensions can look at the use of a more general nonparametric prior, like the generalized gamma process (see Lijoi *et al.*, pear) that includes both the Dirichlet and the N-IG prior as particular cases.

We mention also that the use of the N-IG prior requires a greater computational effort. Indeed, the computation of the weights (2.6) and(2.7) needs 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 initialisation 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.

Bibliography

Andersen, P. K. and Gill, R. D. (1982). Cox's regression model for counting processes. *The Annals of Statistics*, **10**, 1110–1120.

Anderson, P. K., borgan, O., Gill, R. D., and Keinding, N. (1993). *Statistical Models Based on Counting Process.* Springer-Verlag, New York.

Antoniak, C. E. (1974). Mixtures of Dirichlet processes with applications to Bayesian nonparametric problems. *The Annals of Statistics*, **2**, 1152–1174.

Bedrick, E. J., Christensen, R., and Johnson, W. O. (2000). Bayesian accelerated failure time analysis with application to veterinary epidemioligy. *Statistics in Medicine*, **19**, 221–237.

Berger, J. O. and Guglielmi, A. (2001). Bayesian testing of parametric model versus nonparametric alternative. *Journal of the American Statistical Association*, **96**, 174–184.

Berkson, J. and Gage, R. P. (1950). Calculation of survival rates for cancer. *Proceedings of Staff Meetings of the Mayo Clinic*, **25**, 270–286.

Bickel, P. J., Klaassen, C. A. J., Ritov, Y., and Wellner, J. A. (1993). *Efficient and Adaptive Estimation for Semiparametric Models.* The Johns Hopkins Univ. Press, Baltimore.

Billingsley, P. (1968). *Convergence of Probability Measures.* Wiley, New York.

Bilodeau, M. and Brenner, D. (1999). *Theory of Multivariate Statistics,*. Springer Verlag, New York.

Blackwell, D. (1973). The discreteness of Ferguson selections. *The Annals of Statistics*, **1**, 356–358.

Blackwell, D. and MacQueen, J. B. (1973). Ferguson distributions via Pólya-urn schemes. *The Annals of Statistics*, **1**, 353–355.

Buckinsky, M. (1998). Recent advances in quantile regression models: A pratical guideline for empirical research. *The Journal of Human Resurces*, **33**, 88–126.

Buckley, J. and James, I. (1979). Linear regression with censored data. *Biometrics*, **8**, 907–925.

Bush, C. A. and MacEachern, S. N. (1996). A semi-parametric Bayesian model for randomized block designs. *Biometrika*, **83**, 1013–1021.

Chib, S. and Greenberg, E. (1995). Understanding the Metropolis-Hastings algorithm. *The American Statistician*, **49**, 327–335.

Christensen, R. and Johnson, W. (1988). Modeling accelerated failure time with a Dirichlet process. *Biometrika*, **75**, 693–704.

Cook, R. D. and Weisberg, S. (1982). *Residuals and Influence in Regression*. Chapman & Hall, London.

Cox, D. R. (1972). Regression models and life tables (with discussion). *Journal of the Royal Statistical Society, Series B*, **34**, 187–220.

Cox, D. R. (1975). Partial likelihood. *Biometrika*, **62**, 269–276.

Cox, D. R. and Oakes, D. (1984). *Analisys of Survival Data.* Chapman and Hall, London.

Cutler, S. J. and Ederer, F. (1958). Maximum utilisation of the life table method in analysing. survival. *Journal of Chronic Diseases*, **8**, 699–712.

De Blasi, P. and Hjort, N. L. (to appear). Bayesian survival analysis in proportional hazard models with logistic relative risk. *Scand. J. Statist.*

de Finetti, B. (1937). La prévision: ses lois logiques, ses sources subjectives. *Annales de l'Institut Henri Poincaré*, **7**, 1–68.

Dempster, A. P. (1980). Bayesian inference in applied statistics. In J. M. Bernardo, M. H. De Groot, D. V. Lindley, and A. F. M. Smith, editors, *Bayesian Statistics*, pages 266–291. Valencia University Press, Valencia.

Diebolt, J. and Robert, C. P. (1994). Estimation of finite mixture distributions through Bayesian sampling. *Journal of the Royal Statistical Society, Series B*, **56**, 363–375.

Dubins, L. and Freedman, D. (1965). Random distributions function. *Bulletin of the American Mathematical Society*, **69**, 548–551.

Dunson, D. B. and Taylor, J. (2005). Approximate Bayesian inference for quantiles. *Journal of Nonparametric Statistics*, **17**, 385 – 400.

Dykstra, R. L. and Laud, P. W. (1981). A bayesian nonparametric approach to reliability. *The annals of statistics*, **9**, 356–367.

Efron, B. (1977). The efficiency of Cox's likelihood function for censored data. *Journal of the American Statistical Association*, **72**, 557–567.

Escobar, M. D. (1992). Invited comment on "Bayesian analysis of mixture: Some results on exact estimability and identification," by J.P. Florens, M. Mouchart, and J.M. Rolin. In A. P. D. J. M. Bernardo, J. O. Berger and A. F. M. Smith, editors, *Bayesian Statistic*, volume 4, pages 142–144. Oxford University Press, Oxford.

Escobar, M. D. (1994). Estimating normal means with a Dirichlet process prior. *Journal of the American Statistical Association*, **89**, 268–267.

Escobar, M. D. and West, M. (1995). Bayesian density estimation and inference using mixture. *Journal of the American Statistical Association*, **90**, 577–588.

Fabius, J. (1964). Asymptotic behaviour of Bayes estimates. *The Annals of Mathematical Statistics*, **35**, 846–856.

Feigl, P. and Zelen, M. (1965). Estimation of exponential survival probabilities with concomitant information. *Biometrics*, **21**, 826–838.

Ferguson, T. S. (1973). A Bayesian analisys of some nonparametric problems. *Annals of Statistics*, **1**, 209–230.

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

Ferguson, T. S. (1983). Bayesian density estimation by mixture of normal distributions. In H. Rizvi and Rustagi, editors, *Recent Advances in Statistics*, pages 287–302. Accademic Press, New York.

Flemming, T. R. and Lin, D. Y. (2000). Survival analysis in clinical trials: Past developments and future directions. *Biometrics*, **56**, 971–983.

Freedman, D. A. (1963). On the asymptotic behaviour of Bayes estimates in the discrete case I. *The Annals of Mathematical Statistics*, **34**, 1386–1403.

Freedman, D. A. (1965). On the asymptotic behaviour of Bayes estimates in the discrete case II. *The Annals of Mathematical Statistics*, **36**, 454–456.

Gehan, E. A. (1969). Estimating survival functions from the life table. *Journal of Chronic Diseases*, **13**, 629–644.

Gelfand, A. E. (1999). Approaches for semiparametric Bayesian regression. In S. Ghosh, editor, *Asymptotic, Nonparametrics and Time series*, pages 615–638. Marcel Bekker, New York.

Gelfand, A. E. and Kottas, A. (2003). Bayesian semiparametric regression for median residual life. *Scandinavian Journal of Statistics*, **30**, 651–665.

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

Gelfand, A. E., Hills, S. E., Racine-Poon, A., and Smith, A. F. M. (1990). Illustration of Bayesian inference in normal data modes using Gibbs sampling. *Journal of the American Statistical Association*, **85**, 972–985.

Gelfand, A. E., Dey, D. K., and Chang, H. (1992). Model determination using predictive distributions with implementation via sampling-based methods. In A. P. D. J. M Bernardo, J. O. Berger and A. F. M. Smith, editors, *Bayesian Statistics*, volume 4, pages 147–167. Oxford University Press.

Geman, S. and Geman, D. (1984). Stochastic relaxation, Gibbs distributions and the Bayesian restoration of immages. *IEEE Trans. Pattn. Anal. Mach. Intel.*, **6**, 721–741.

Ghosh, S. K. and Ghosal, S. (2006). Semiparametric accelerated failure time models for censored data. *Bayesian Statistics and its Applications (S. K. Upadhyay et al., eds.), Anamaya Publishers, New Delhi*, pages 213–219.

Gilks, W. R., Richardson, S., and Spiegelhalter, D. J. (1996). *Markov Chain Monte Carlo in Practice*. Chapman & Hall/CRC.

Gradshtey, I. and Ryzhik, I. (1994). *Table of Integrals, Series, and Products (Fifth Edition)*. Academic Press.

Green, P. J. (1995). Reversible jump Markov Chain monte Carlo computation and Bayesian model determination. *Biometrika*, **82**, 711–732.

Green, P. J. and Richardson, S. (2001). Modelling hetergeneity with and without the Dirichlet process. *Scandinavian Journal of Statistics*, **28**, 355–375.

Hanson, T. E. (2006). Modeling censored lifetime data using a mixture of gammas baseline. *Bayesian Analysis*, **1**, 575–594.

Hanson, T. E. and Johnson, O. W. (2002). Modeling regression error with a mixture of Pólya trees. *Journal of the American Statistical Association*, **97**, 1020–1033.

Hastings, W. K. (1970). Monte Carlo sampling methods using Markov chains and their applications. *Biometrika*, **57**, 97–109.

Hewitt, E. and Savage, L. J. (1955). Symmetric measures on Cartesian products. *Transacrions of the American Mathematical Society*, **80**, 470–501.

Hjort, N. L. (1990). Nonparametric bayes estimator based on beta processes in models for life history data. *The annals of Statistics*, **18**, 1259–1294.

Hjort, N. L. and Petrone, S. (2006). Nonparametric quantile inference using Dirichlet proces. *Studi Statistici n.94, Istituto di Metodi Quantitativi, Universitá L.Bocconi, Milano*.

Ibrahim, J. G., Chen, M. H., and Sinha, D. (2001). *Bayesian Survival Analysis*. Springer-Verlag.

Ishwaran, H. and James, L. F. (2001). Gibbs sampling methods fo stick-breaking priors. *Journal of the American Statistical Association*, **96**, 161–173.

Ishwaran, H. and James, L. F. (2003). Generalized weigthed chinese restaurant processes for species sampling mixture models. *Statistica Sinica*, **13**, 1211–1235.

Jain, S. and Neal, M. (2004). A split-merge Markov Chain Monte Carlo procedure for the Dirichlet process mixture models. *Journal of Computational and Graphical Statistics*, **13**, 158–182.

James, L. F. (2002). Poisson process partition calculus with application to exchangeable models and bayesyan nonparametrics,. *Mathematics ArXiv*, **math. PR/0205093**.

James, L. F. (2003). A simple proof of the almost sure discreteness of a class of random measures,. *Statistics & probability letters*, **65**, 363–368.

James, L. F., Lijoi, A., and Prünster, I. (2006). Conjugacy as distinctive feature of the Dirichlet process. *Scandinavian journal of Statistics*, **33**, 105–120.

Kalbfleisch, J. D. and Prentice, R. L. (1980). *The Statistical Analysis of Failure Time Data*. Wiley, New York.

Kaplan, E. L. and Meier, P. (1958). Nonparametric estimation from incomplete observations. *Journal of the American Statistical Association*, **53**, 457–448.

Klein, J. P. and Moeschberger, M. L. (1997). *Survival Analysis. Tecniques for Censored and Truncated Data*. Springer-Verlag, New York.

Kottas, A. and Gelfand, A. E. (2001). Bayesian semiparametric median regression modeling. *Journal of the American Statistical Association*, **96**, 1458–1468.

Kottas, A. and Krnjajič, M. (2005). Bayesian nonparametric modeling in quantile regression. *AMS Technical Report-06*.

Koul, H., Susarla, V., and Van Ryzin, J. (1981). Regression analysis with randomly right censored data. *The Annals of Statistics*, **9**, 1276–1288.

Kuo, L. (1986). Computation of mixture of Dirichlet process. *SIAM Journa of Science and Statistical Computing*, **7**, 60–71.

Kuo, L. and Mallick, B. (1997). Bayesian semiparametric inference for the accelerated failure-time model. *The Canadian Journal of Statistics*, **25**, 457–472.

Lai, T. L., Ying, Z., and Zheng, Z. (1992). Asymptotic normality of a class of adaptive statistics with applications to synthetic data methods for censored regression. *Journal of Multivariate Analysis*, **52**, 159–179.

Lavine, M. (1992). Some aspects of Pólya tree distributions for statistical modelling. *The Annals of Statistics*, **20**, 1222–1235.

Lawless, J. F. (1982). *Statistical Models and Methods for Life Time Data*. Wiley, New York.

Lijoi, A., Mena, R. H., and Prünster, I. (2005). Hierarchical mixture modeling with normalized inverse-Gaussian prior. *Journal of the American Statistical Association*, **100**, 1278–1291.

Lijoi, A., Mena, R. H., and Pruenster, I. (to appear). Controlling the reinforcement in Bayesian nonparametric mixture models.

Lin, D. Y. and Geyer, C. J. (1992). Computational methods for semiparametric linear regression with censored data. *Comp. Graph. Statist.*, **1**, 77–90.

Lindley, D. V. and Smith, A. F. M. (1972). Bayesian estimates for the linear model (with discussion). *Journal of the Royal Statistical Society*, **34**, 1–42.

Lo, A. Y. (1984). On a class of Bayesian nonparametric estimates: I. density estimates. *The Annals of Statistics*, **12**, 351–357.

MacEachern, S. N. (1994). Estimating normal means with a conjugate style Dirichlet process prior. *Communication in Statisistics: Simulation and Computation*, **23**, 727–741.

MacEachern, S. N. (1999). Dependent nonparametric processes. In *ASA Proceedings of the the Section on Bayesian Statistical Science*. American Statistical Association, Alexandria, VA.

MacEachern, S. N. and Müller, P. (1998). Estimating mixture of Dirichlet process models. *Journal of Computational and Graphical Statistics*, **7**, 727–741.

MacEachern, S. N. and Müller, P. (2000). Efficient MCMC schemes for robust models extensions using encompassing Dirichlet process mixture models. In D. Rios-Insua and F. Ruggeri, editors, *In Robust Bayesian Analysis*, pages 295–316. Springer, New York.

Maksymiuk, A. W., Jett, J. R., Earle, J. D., Su, J. Q., Diegert, F. A., Maillard, J. A., Kardinal, C. G., Krook, J. E., Veeder, M. H., Wiesenfeld, M. Tschetter, L. K., and Levitt, R. (1994). Sequencing and schedule effects of cisplatin plus etoposide in small cell lung cancer results of a north central cancer treatment group randomized clinical trial. *Journal of clinical Oncology*, **12**, 70–76.

Mauldin, R. D., Sudderth, W. D., and Williams, S. C. (1992). Pólya trees and random distributions. *The Annals of Statistics*, **20**, 1203–1221.

McCracken, D. D. (1955). The monte Carlo method. *Scientific American*, **192(5)**, 90–96.

McCullag, P. and Nelder, J. (1989). *Generalized Liner Models, 2nd edition*. Chapman and Hall, London.

Metropolis, N., Rosenbluth, A. W., Teller, A. H., and Teller, E. (1953). Equations of state calculation by fast computing machines. *The Journal of Chemical Physics*, **21**, 1087–1092.

Meyn, S. P. and Tweedie, R. L. (1993). *Markov Chain and Stocastic Stability*. Springer, New York.

Mira, A. (2005). MCMC methods to estimate Bayesian parametric models. In D. K. Dey and C. R. Rao, editors, *Handbook of Statistics*, volume 25, pages 415–436. Elsevier, Nort Holland.

Neal, R. M. (1992). Bayesian mixture modeling. In S. G. E. C, R and P. Neudorfer, editors, *Maximum Entropy and Bayesian Methods: Proceedings of the 11th International Workshop on Maximum Entropy abd Bayesian Methods of Statistical Analysis, Seattle, 1991*, pages 197–211. Kluwer Academic, Dordrecht.

Neal, R. M. (2000). Markov Chain sampling methods for Dirichlet process mixture models. *Journal of Computational and Graphical Statistics*, **9**, 249–265.

Nieto-Barajas, L. E., Prünster, I., and Walker, S. G. (2004). Normalized random measure driven by increasing additive process. *The Annals of Statistics*, **32**, 2343–2360.

Nummelin, E. (1984). *General Irreducible Markov Chain and Non-Negative Operators*. Cambridge University Press, Cambridge.

Oakes, D. (1977). The asymptotic information in censored survival data. *Biometrika*, **64**, 441–448.

Parthasarathy, K. R. (1967). *Probability Measures on Metric Spaces.* Academic Press, New York.

Pitman, J. (1996). Some developments of the blackwell-macqueen urn scheme. In T. F. et al., editor, *Statistic, Probability and Game theory, Papers in Honor of David Blackwell*, pages 245–267. Institute of Mathematica Statistics, Hayward.

Prentice, R. L. (1978). Linear rank test with right censored data. *Biometrika*, **65**, 167–179.

Regazzini, E. (1996). *Impostazione non Parametrica di Problemi d'Inferenza statistica Bayesiana,*. Quaderni CNR-Imati.

Regazzini, E. (1999). Old and recent results on the relationship between predictive inference and statistical modelling either in nonparametric or parametric form. In A. D. A. S. J.M. Bernardo, J. Berger, editor, *Bayesian Statistics 6*. Oxford University Press, Oxford.

Regazzini, E. (2001). *Foundation of Bayeian Statistics and Some Theory of Bayesian Nonparametric Methods*. Lecture note, Stanford University.

Regazzini, E., Lijoi, A., and Prünster, I. (2003). Distributional results for means of random measures with independent increment. *The Annals of Statistics*, **31**, 560–585.

Reid, N. (1994). A conversation with Sir David Cox. *Statistical Science*, **9**, 439–455.

Richardson, S. and J., G. P. (1997). On Bayesian analysis of mixtures with an unknown numbers of components. *Journal of Royal Statistical Society Series B*, **59**, 731–792.

Robert, C. P. and Casella, G. (2004). *Monte Carlo Statistical Methods (second edition).* Springer-Verlag, New York.

Rosseeuw, P. J. and Leroy, A. M. (1987). *Robust Regression and outlier Detection*. New York, John Wiley.

Seshadri, V. (1993). *The inverse Gaussian Distribution,*. Oxford University Press., New York.

Sethuraman, J. (1994). A constructive definition of Dirichlet prior. *Statistica Sinica*, **4**, 639–650.

Sethuraman, J. and Tiwari, R. (1982). Convergence of Dirichlet measures and the interpretation of their parameter. In S. Gupta and J. Berger, editors, *Proc. 3rd Purdue Symp. Statistical Decision Theory and Relate Topics*. Accademic Press, New York.

Tierney, L. (1994). Markov Chain for exploring posterior distributions. *The Annals of Statistics*, **22**, 1701–1728.

Titteringon, D. M., Smith, A. F. M., and Makov, U. E. (1985). *Statistical Analysis of Finite Mixture Distributions*. Chichester, Wiley.

Walker, S. and Damien, P. (1998). Sampling methods for Bayesian nonparametric inference involving stochastic processes. In D. S. D. Dey, P. Muller, editor, *Pratical Nonparametric and Semiparametric Bayesian Statistics*, pages 243–254. Springer-Verlag,New York.

Walker, S. and Mallick, B. (1999). A Bayesian semiparametric accelerated failure time model. *Biometrics*, **55**, 477–483.

Walker, S. and Muliere, P. (1997). Beta-Stacy processes and a generalization of the Pòlya-urn scheme. *The annals of Statistics*, **25**, 1762–1780.

Wei, L. J. (1992). The accelerated failure time model: A useful alternative to Cox regression model in survival analysis. *Statistics in Medicine*, **11**, 1871–1879.

West, M., Müller, P., and Escobar, M. D. (1994). Hierarchical priors and mixture models, with applications in regression and density estimation. In A. Smith and P. Freeman, editors, *Aspects of Uncertanity: A Tribute to D. V. Lindley*, pages 363–368. john Wiley and Sons, London.

Yang, S. (1999). Censored median regression using weighted empirical survival and hazard functions. *Journal of the American Statistical Association*, **94**, 137–145.

Ying, Z., Jung, S. H., and Wei, L. J. (1995). Survival analysis with median regression models. *Journal of the American Statistical Association*, **90**, 178–184.

Yu, K. and Moyeed, R. A. (2001). Bayesian quantile regression. *Statistic & Probability Letters*, **54**, 437–447.

Yu, K., Lu, Z., and Stander, J. (2003). Quantile regression: applications and current research areas. *The statistician*, **52**, 331–350.

Zhou, M. (1992). Asymptotic normality of the 'synthetic data' regression estimation for censored survival data. *The Annals of Statistics*, **20**, 1002–1021.