L. Bocconi University

Some extensions of the Polya urn scheme with Bayesian applications.

Lorenzo Trippa

Matr.1094861

Advisor: Pietro Muliere

Milan, September 2008

Summary

The reinforcement concept plays a key role in Bayesian statistics. In particular the Polya urn scheme has been extensively studied in Bayesian nonparametrics and adopted in a number of inferential procedures. This intuitive representation of the Dirichlet process underlines the conjugacy property of the Ferguson-Dirichlet prior, which is by large the most known Bayesian nonparametric model. The prior is characterized by its analytic tractability and its large support. These two futures made the Dirichlet process a reference tool for implementing Bayesian inference avoiding restrictive parametric assumptions.

The aim of the thesis is to illustrate some extensions of the Polya urn scheme constructed with specific purposes.

In the first two chapters two extensions for modeling partially exchangeable populations are proposed.

The first extension is finalized to the characterization of a flexible prior for dependent random distributions indexed by covariates. The main properties of the prior are illustrated and its application to survival regression analysis is discussed.

In the second chapter the idea of characterizing partially exchangeable variables by mean of latent colored tessellations is illustrated and its application to binary regression problems is discussed.

In the third chapter an extension of the Bernstein prior is proposed. A novel mixing random measure for Beta kernels which generalizes the Dirichlet process is specified introducing a particular class of reinforced urn processes. It is shown that such extension, if compared with the Bernstein prior, can produce substantially different posterior inference. Some motivations behind these discrepancies are given and it is shown that the extension results more robust than the Bernstein prior, which in some cases suffers from overfitting problems.

In the last chapter an experimental design problem is considered adopting a decisional-theoretic approach. A Bayesian framework for optimizing the tuning parameters of a clinical trial structured in two stages is illustrated. Relevant difficulties in the elicitation of the prior distribution are underlined. A slight

variation of the Bayesian bootstrap reinforcement mechanism is proposed as a practical tool, in order to specify an informative prior, when historical data are available. Adapting the Bayesian bootstrap to non exchangeable variables in such a way to specify an informative prior, which reflect the historical information, allows to construct a quasi-automated procedure to optimize the choice of the tuning parameters.

Acknowledgments

Several people have played a part during my research and I have benefitted from several supports. I acknowledge the essential support of Peter Muller. Second, I sincerely thank Pietro Muliere and Sonia Petrone for the theoretical basis they provided for me. Gary Rosner and Paolo Bulla have been also important for several stimulating conversations.

Contents

Chapter 1

A Dependent Polya Tree Model: Bayesian Nonparametric Survival Regression

Abstract

We propose a probability model for a family of unknown distributions indexed with covariates. The marginal model for each distribution is a Polya tree prior. The proposed model introduces the desired dependence across the marginal Polya tree models by defining dependent random branching probabilities of the unknown distributions. The dependence is on the corresponding branching probabilities across covariate levels.

An important feature of the proposed model is the easy centering of the nonparametric model around any parametric regression model. This is important for the motivating application to the proportional hazards (PH) model. We use the proposed model to implement nonparametric inference for survival regression. The proposed model allows us to center the nonparametric prior around the PH structure. In contrast to many available models that restrict the non-parametric extension of the PH model to the baseline hazard, the proposed model defines

a family of random probability measures that are a priori centered around the PH model but allows any other structure. This includes, for example, crossing hazards, additive hazards, or any other structure as supported by the data.

1.1 Introduction

We propose a new nonparametric Bayesian model for a family $\{\mathcal{P}_x;\, x \in X\}$ of dependent random probability measures $\mathcal{P}_x$ indexed with covariates x. The proposed model is an extension of the Polya tree (PT) model (Lavine (1992,1994)). The main features are: (i) The marginal prior $p(\mathcal{P}_x)$ is a Polya tree model. (ii) The model is a priori centered around any desired family of distributions $\{F_x\}_{x\in X}$. The prior centering model $\{F_x\}_{x\in X}$ can be indexed by hyperparameters. (iii) The availability of posterior Markov chain Monte Carlo (MCMC) simulation schemes to implement inference. The motivating application is to survival regression. We use the proposed model to implement a fully nonparametric extension of the proportional hazards (PH) model.

Without loss of generality we keep the following discussion specific to this application.

Many recent discussions of Bayesian approaches to survival analysis focused on the PH model. A variety of analytically tractable prior distributions with suitably large supports for the baseline hazard function have been proposed. Kalbfleisch (1978) adopted the gamma process for modeling the cumulative hazard, Hjort (1990) introduced the Beta process and Kim and Lee (2003, a) presented a general framework based on the neutral to the right processes.

An important constraint of the PH model is that the time to event distributions associated with different points of the covariate space, for example the survival times of patients with different diagnostic profiles in a clinical study, are stochastically ordered. The prior probability that two random distribution functions cross is zero. The model proposed in this article is motivated by the observation that such a constraint is reasonable as approximate characterization

of the prior information, but should not be strictly enforced by the model. The probability model should allow for violation of the PH structure, as and when indicated by the data. A typical example is the experimental comparison of two competing anticancer therapies in which a priori the trial is expected to indicate a superior treatment but, as illustrated for example in Mantel and Stablein (1988), the survival functions could cross each other.

The objective of the paper is to present a flexible alternative to the PH model for those cases when doubts arise about the PH structure. The adopted approach is hierarchical and it is based on the definition of a class of dependent random probability measures (RPM) centered on parametric distribution functions which satisfy the PH model assumptions.

The outlined modeling approach requires an adequate specification of the joint distribution of the survival functions under different covariates. Consider any two points x_1 and x_2 in the covariate space. The closer x_1 and x_2 are, the more similar should the corresponding distributions be. For infinitesimal distance the differences should vanish.

A joint distribution for RPMs indexed by covariates consistent with the above consideration is defined. I.e., we define a stochastic process $\mathcal{P}$ indexed by the points of the set $\{X \times \mathcal{B}\}$ where X is the covariate space, and $\mathcal{B}$ is the usual Borel σ-field of the real line. It is assumed that $X \subset \mathbb{R}^k$, for a given positive integer k. The salient properties of the proposed process $\mathcal{P}$ are:

1. For any $x \in X$ the marginal process $\{\mathcal{P}_x(B)\}_{B\in\mathcal{B}}$ is a RPM with Polya tree distribution.

2. The process can be a priori centered on a given regression model $\{F_x\}_{x\in X}$. The model $\{F_x\}_{x\in X}$ can include unknown parameters. This is the case, for example, when using a PH model with unknown regression parameters for prior centering.

3. For every $B \in \mathcal{B}$, $x \in X$ and $\{x_i \in X\}_{i\geq 1}$ such that $x_i \to x$, the random variables sequence $\{\mathcal{P}_{x_i}(B)\}_{i\in\mathbb{N}}$ converges in probability to $\mathcal{P}_x(B)$;

4. For every finite subset of the covariate space $\{x_1, \ldots, x_m\}$, the law of the RPMs $(\mathcal{P}_{x_2}, \ldots, \mathcal{P}_{x_m})$ has full support with respect to the weak convergence topology.

The outline of the paper is the following. In section 1.2, a random process called multivariate Beta process (MBP) is defined. To avoid confusion we clarify that it is not related with the Beta process introduced in Hjort (1990). The random model $\mathcal{P}$, introduced in section 3, which we denote the multivariate Polya tree (MPT), is based on the MBP and on the Polya tree scheme. The MBP generates the beta distributed random branching probabilities that are required for the construction of the marginal Polya tree models. In section 4 we introduce a hierarchical prior with unknown parameters in the prior centering distributions. Section 5 briefly describes the main features of posterior MCMC simulation. Section 6 reports some examples. The last section concludes with final remarks and a discussion of open issues.

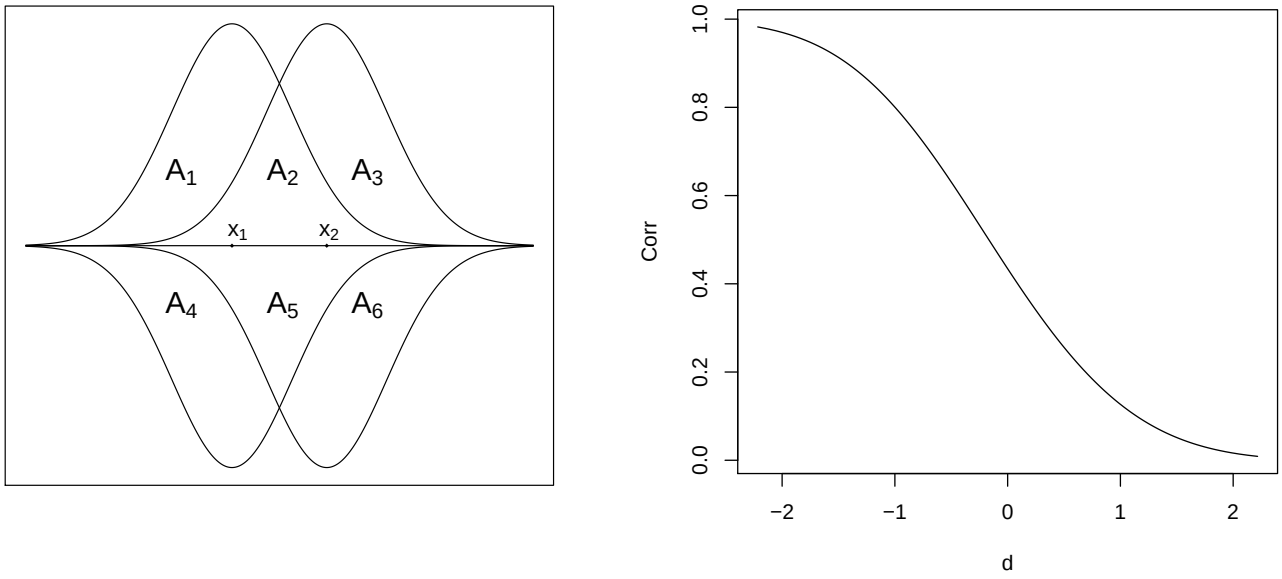
1.2 The Multivariate Beta Process

Before introducing the MPT model we define an instrumental process $\{Y_x\}_{x \in X}$, with Beta marginal distributions, $Y_x \sim Beta(\alpha_0, \alpha_1) \quad \forall x \in X$. The definition of the process is a natural extension of the following characterization of a bivariate random vector (Y_1, Y_2) with beta marginal distributions:

$$(Y_1, Y_2) \stackrel{d}{=} \left(\frac{G_1 + G_2}{G_1 + G_2 + G_4 + G_5}, \frac{G_2 + G_3}{G_2 + G_3 + G_5 + G_6}\right) \tag{1.1}$$

where the equality is in distribution and $G_1, \ldots, G_6$ are independent gamma $r.v.$'s with fixed scale parameter. Representing the shape parameters of $G_1, \ldots, G_6$ as the areas of the sets $A_1, \ldots, A_6$ in Figure 1.1a it is easy to see how the construction can be extended. In Figure 1.1a, the kernels $\alpha_0 q_{x_1}$, $\alpha_0 q_{x_2}$ $-\alpha_1 q_{x_1}$, and $-\alpha_1 q_{x_2}$ centered at x_1 and x_2 specify the distribution of the random vector (Y_1, Y_2). We extend the construction to $\{Y_x\}_{x \in X}$ by expanding from two kernels centered at $\{x_1, x_2\}$ to a family of kernels with location parameters in X. The extension

exploits the fact that a Beta random variable ($r.v.$) can be represented as the ratio of two gamma distributed $r.v.$'s and the infinite divisibility property of the gamma distribution.



(a) (b)

Figure 1.1: Panel (a) shows the kernels $\alpha_0 q_{x_1}(\cdot)$, $\alpha_0 q_{x_2}(\cdot)$, $-\alpha_1 q_{x_1}(\cdot)$ and $-\alpha_1 q_{x_2}(\cdot)$, centered at two covariate values x_1 and x_2. The areas indicated by A_1 through A_6 are used to construct the dependent r.v.'s Y_1 and Y_2 in (1.1). Panel (b) plots the $Corr(Y_x, Y_{x+d})$ as a function of d. Here $X = \mathbb{R}$, $\{Y_x\}_{x\in X} \sim \text{MBP}(\alpha_0, \alpha_1, Q)$, $\alpha_0 = \alpha_1 = 1$, μ is the Lebesgue measure and $\{q_x\}_{x\in\mathbb{R}}$ are gaussian kernels with mean x and variance equal to 1.

Let X be endowed with a σ-field $\mathcal{X}$ and a σ-finite measure μ. Let $\{Q_x\}_{x\in X}$ be a location family of probability measures absolutely continues with respect to μ and derivatives $\{q_x\}_{x\in X}$. Throughout the paper such derivatives will be assumed to be unimodal. Let $\{G(A)\}_{A\in\{\mathcal{X}\times\mathcal{B}\}}$ be a gamma process indexed by the sets of the product σ-field $\mathcal{X}\times\mathcal{B}$. For every $m \in \{1, 2, \ldots\}$ and non overlapping $A_1, \ldots, A_m$, the $r.v.$'s $G(A_1), \ldots, G(A_m)$ are independently gamma distributed

with fixed scale parameter and shape parameters $\nu(A_j)$, where ν is the product measure on $(X \times \mathbb{R})$ of μ and the Lebesgue measure. Finally for every x let $A_x^0 = \{(z, y) \in X \times \mathbb{R} : 0 < y < \alpha_0 q_x(z)\}$ and symmetrically $A_x^1 = \{(z, y) \in X \times \mathbb{R} : -\alpha_1 q_x(z) < y < 0]\}$ denote the areas bounded by the kernels $\alpha_0 q_x(\cdot)$ and $-\alpha_1 q_x(\cdot)$. In Figure 1.1, the areas $A_{x_1}^0$ and $A_{x_1}^1$ are bordered by the two kernels above and below the horizontal axis that are centered in x_1. We define the beta *r.v.*

$$Y_x = \frac{G(A_x^0)}{G(A_x^0) + G(A_x^1)}.$$

We call the constructed process $\{Y_x\}_{x \in X}$ the *multivariate beta process* with parameters (α_0, α_1, Q) and use the notation

$$\{Y_x\}_{x \in X} \sim \text{MBP}(\alpha_0, \alpha_1, Q).$$

The finite dimensional distributions of $\{Y_x\}_{x \in X}$ can be alternatively characterized as follows. Consider a set of covariate values $x_1, \ldots x_m$. Let $S = \{(z, y) \in X \times \mathbb{R} : \min_i[-\alpha_1 q_{x_i}(z)] < y < \max_i[\alpha_0 q_{x_i}(z)]\}$ and $\nu_{x_1,\ldots,x_m}$ denote the ν measure restricted to S. Let $D_{x_1,\ldots,x_m}$ be a Dirichlet random measure: $D_{x_1,\ldots,x_m} \sim DP(\nu_{x_1,\ldots,x_m})$. Then

$$[Y_{x_1}, \ldots, Y_{x_m}] \stackrel{d}{=} [\frac{D_{x_1,\ldots,x_m}(A_{x_1}^0)}{D_{x_1,\ldots,x_m}(A_{x_1}^0 \cup A_{x_1}^1)}, \ldots, \frac{D_{x_1,\ldots,x_m}(A_{x_m}^0)}{D_{x_1,\ldots,x_m}(A_{x_m}^0 \cup A_{x_m}^1)}]. \tag{1.2}$$

The identity between the two definitions of the process $\{Y_x\}_{x \in X}$ follows from the fact that a Dirichlet process can be represented as a normalized gamma process.

For later use we note the following. Exploiting the second representation (1.2) of the MBP it is easy to sample a sequence of partially exchangeable dichotomous variables $\mathbf{Z} = (Z_1, \ldots, Z_m)$ when their joint distribution is specified through a MBP:

$$P(Z_1, \ldots, Z_m \mid Y) = \prod_1^m Y_{x_i}^{Z_i}(1 - Y_{x_i})^{1 - Z_i} \quad \text{and} \quad \{Y_x\}_{x \in X} \sim \text{MBP}(\alpha_0, \alpha_1, Q). \tag{1.3}$$

The urn scheme in Blackwell and MacQueen (1973) can be used to sample $(Z_1, \ldots, Z_m)$. Let $\{(l_i, h_i) \in X \times \mathbb{R}\}_{i \geq 1}$ be an exchangeable sequence with random

distribution $D_{x_1,\ldots,x_m}$. Let $(i_1,\ldots,i_m)$ be a vector of integers recursively defined as $i_j = \inf\{i > i_{j-1} : -\alpha_1 q_{x_j}(l_i) < h_i < \alpha_0 q_{x_j}(l_i)\}$ with $i_0 = 0$. The subsequence i_j, $j = 1,\ldots,m$ selects pairs (l_{i_j}, h_{i_j}) that belong to the areas contoured by the curves $\alpha_0 q_{x_j}$ and $-\alpha_1 q_{x_j}$. The truncated sequence $\{(l_i,h_i)\}_{i\in\{1,\ldots,i_m\}}$ can be generated by the Polya urn. The random vector $\tilde{\mathbf{Z}} = (I(h_{i_1} > 0),\ldots,I(h_{i_m} > 0))$ has the same distribution of $\mathbf{Z}$. The equality in distribution follows from the fact that the random probabilities of $\tilde{\mathbf{Z}}$ are exactly the right hand side of (1.2), and thus have the same distribution as the random probabilities $\{Y_{x_i}\}_{i=1}^m$ that define the probability distribution of $\mathbf{Z}$.

The augmented model with the latent sequence $\{(l_i,h_i)\}_{i\geq 1}$ can be used to implement posterior predictive inference for a future Z_{m+1} with covariate x_{m+1}. We will use this construction for the implementation of posterior simulation in section 1.5.

The correlation function of the process $\{Y_x\}_{x\in X} \sim \text{MBP}(\alpha_0,\alpha_1,Q)$ can be computed through a result in Olkin and Liu (2003). They studied the bivariate beta distribution of a random vector (Y_1,Y_2) characterized by

$$(Y_1,Y_2) \stackrel{d}{=} (\frac{G_1}{G_1+G_3}, \frac{G_2}{G_2+G_3}) \tag{1.4}$$

where G_1, G_2 and G_3 are independent gamma $r.v.$'s with shape parameters a, b and c. They show that

$$E(Y_1Y_2) = h\sum_{j=0}^{\infty} \frac{\Gamma(a+j+1)}{\Gamma(a+b+c+j+1)} \frac{\Gamma(b+j+1)}{\Gamma(a+b+c+j+1)} \frac{1}{j!} \tag{1.5}$$

where

$$h = \frac{ab\Gamma(a+b)\Gamma(b+c)\Gamma(a+b+c+1)}{\Gamma(c)\Gamma(a+1)\Gamma(b+1)} \quad .$$

Consider now $\{Y_x\}_{x\in X} \sim \text{MBP}(\alpha_0,\alpha_1,Q)$. Let, for $i \in \{1,2\}$ and $j \in \{0,1\}$,

$A_i^j = A_{x_i}^j$, $A_i = A_i^0 \cup A_i^1$, $A_{12}^j = A_1^j \cap A_2^j$, $A_{12} = A_{12}^1 \cup A_{12}^0$ and $D = D_{x_1,x_2}$. Then

$$E(Y_{x_1}Y_{x_2}) = E\left(\left[\frac{D(A_{12}^0)}{D(A_1)} + \frac{D(A_1^0 \backslash A_2^0)}{D(A_1)}\right]\left[\frac{D(A_{12}^0)}{D(A_2)} + \frac{D(A_2^0 \backslash A_1^0)}{D(A_2)}\right]\right) =$$
$$E\left(\frac{D(A_{12})}{D(A_1)}\frac{D(A_{12})}{D(A_2)}\right) E\left(\left[\frac{D(A_{12}^0)}{D(A_{12})}\right]^2\right) + E\left(\frac{D(A_{12})}{D(A_1)}\frac{D(A_2 \backslash A_1)}{D(A_2)}\right)\left(\frac{\alpha_0}{\alpha_0+\alpha_1}\right)^2 +$$
$$E\left(\frac{D(A_1 \backslash A_2)}{D(A_1)}\frac{D(A_{12})}{D(A_2)}\right)\left(\frac{\alpha_0}{\alpha_0+\alpha_1}\right)^2 + E\left(\frac{D(A_1 \backslash A_2)}{D(A_1)}\frac{D(A_2 \backslash A_1)}{D(A_2)}\right)\left(\frac{\alpha_0}{\alpha_0+\alpha_1}\right)^2 \tag{1.6}$$

The equality follows from the representation (1.2) and from the tail free property of the Dirichlet process. The right hand side of the equation allows us to compute the quantities $E(Y_{x_1}Y_{x_2})$ using the identity (1.5). This is possible since the law of the random vector

$$\left(\frac{D(A_1 \backslash A_2)}{D(A_1)}, \frac{D(A_2 \backslash A_1)}{D(A_2)}\right)$$

belongs to the distributions family studied in Olkin and Liu (2003). Figure 1.1b illustrates the correlation function of a specific MBP.

The parameters of the MBP have a clear interpretation: α_0 and α_1 characterize the univariate marginal distributions while the correlation function can be flexibly chosen through a suitable specification of Q. A simple example helps to clarify the relationship between Q and the correlation function. Let Q be the location family of the normal distributions on the real line with variance σ^2. The correlation function of the process $\{Y_x\}_{x\in\mathbb{R}}$ depends on the choice of σ^2: the higher the variance the higher the correlations of the bivariate marginal distributions. The triple $\alpha_0, \alpha_1, \nu(A_{x_1}^0 \cap A_{x_2}^0)$ parameterizes the joint distribution of (Y_{x_1}, Y_{x_2}). For fixed values of σ^2, α_0 and α_1, we show in the appendix (proposition 1.5) that the correlation between Y_{x_1} and Y_{x_2} is a decreasing function of the Euclidean distance $|x_1 - x_2|$. It follows that the map $(\sigma, x_1, x_2) \to Corr_{\sigma^2}(Y_{x_1}, Y_{x_2})$ is decreasing with respect to $(|x_1 - x_2|/\sigma)$.

Finally, we state one more result for later use.

Proposition 1.1. Continuity in x. *Given $x \in X$ and a sequence $\{x_i \in X\}_{i\geq 1}$,*

if

$$\sup_{B\in\mathcal{B}} |Q_{x_i}(B) - Q_x(B)| \to 0$$

then the sequence $\{Y_{x_i}\}_{i\geq 1}$ *converges in probability to* Y_x.

(The proof is in the appendix.)

1.3 Multivariate Polya Trees

For later reference we recall the definition of the Polya tree (PT) prior (Lavine, 1992,1994). Let $\Pi = (B \equiv \Omega; B_0, B_1; B_{00}, B_{01}, \ldots)$ be a binary tree of partitions of a separable measurable space Ω such that $(B; B_0, B_1; B_{00}, \ldots)$ generates the measurable sets. Let $\mathcal{A} = (\alpha_0, \alpha_1, \alpha_{00}, \ldots)$ be a sequence of non-negative numbers. Finally, let $E = \bigcup_{m=1}^{\infty}\{0,1\}^m$ denote the index set of Π and $\mathcal{A}$.

Definition 1 (Lavine 1992)**.** *A random probability measure* $\mathcal{P}$ *on* Ω *is said to have a Polya tree (PT) distribution, with parameter* $(\Pi, \mathcal{A})$*, written* $\mathcal{P} \sim PT(\Pi, \mathcal{A})$*, if there exist r.v.'s* $\mathcal{Y} = (Y_\varepsilon, \varepsilon \in E \cup \oslash)$ *such that (i) The r.v.'s in* $\mathcal{Y}$ *are independent; (ii) For every* $\varepsilon \in E \cup \oslash$*,* $Y_\varepsilon \sim Beta(\alpha_{\varepsilon 1}, \alpha_{\varepsilon 0})$ *and* $\mathcal{P}(B_{\varepsilon 1})/\mathcal{P}(B_\varepsilon) = Y_\varepsilon$.

For an extensive discussion of the properties of the Polya tree model we refer also to Mauldin et al. (1992).

1.3.1 The Multivariate Polya Tree Model

We introduce the multivariate Polya tree distribution. The idea formalized in the following definition is to substitute the beta *r.v.*'s that characterize the Polya tree model with random processes indexed by the points of a covariate space.

In the sequel $\{B_0, B_1; B_{00}, \ldots\} = \{(0, 1/2], (1/2, 1]; (0, 1/4], \ldots\}$ is fixed as the standard dyadic tree of partitions of the unit interval. The proposed model is parameterized by $(\mathcal{A}, Q, F)$ defined as follows. Let $\mathcal{A} = (\alpha_\varepsilon, \varepsilon \in E \cup \oslash)$ denote a sequence of positive numbers. For a σ-finite measure μ on $(X, \mathcal{X})$, let $Q = \{Q_x\}_{x\in X}$ denote a location family of probability measures absolutely

continuous with respect to μ. And F is a class of continuous distribution functions $F = \{F_x\}_{x \in X}$.

Definition 2. *A class of random probability measures on the real line $\{\mathcal{P}_x\}_{x \in X}$ has a multivariate Polya tree distribution (MPT), with parameters $(\mathcal{A}, Q, F)$ if there exist random processes $\mathcal{Y} = (\{Y_{\varepsilon,x}\}_{x \in X}, \varepsilon \in E \cup \oslash)$ such that the following hold:*

1. *The random processes in $\mathcal{Y}$ are independent;*
2. *For every $\varepsilon \in E \cup \oslash$, $\{Y_{\varepsilon,x}\}_{x \in X}$ is MBP$(\alpha_\varepsilon, \alpha_\varepsilon, Q)$ distributed.*
3. *For every $m = 1, 2, \ldots$ and every $\varepsilon \in \{0,1\}^m$*

$$\mathcal{P}_x(F_x^{-1}(B_{\varepsilon_1,\ldots\varepsilon_m})) = (\prod_{j=1;\varepsilon_j=1}^{m} Y_{\varepsilon_1,\ldots,\varepsilon_{j-1},x})(\prod_{j=1;\varepsilon_j=0}^{m} (1 - Y_{\varepsilon_1,\ldots,\varepsilon_{j-1},x})) \qquad (1.7)$$

where the first term in the products is interpreted as $Y_{\oslash,x}$ or as $1 - Y_{\oslash,x}$ and, for every $a \in (0,1]$, $F^{-1}(a) = \inf\{b \in \mathbb{R} : F(b) > a\}$.

Note that the indices of α_ε in Definition 2 are shifted by one binary digit compared to Definition 1. This is because the two parameters of the beta r.v.'s in the MPT are equal, whereas in the PT the two parameters could differ.

The following proposition asserts that the defined class of RPMs $\{\mathcal{P}_x\}_{x \in X}$ can be centered on an arbitrary model $\{F_x\}_{x \in X}$. In particular, we will be interested on centering the prior on a PH model.

Proposition 1.2. Prior Mean. *For every $x \in X$ and every Borel subset of the real line B if $\{\mathcal{P}_x\}_{x \in X} \sim MPT(\mathcal{A}, Q, F)$ then the expected value of the r.v. $\mathcal{P}_x(B)$ is equal to $F_x(B)$.*
(The proof is in the appendix)

1.3.2 Choice of the MPT Parameters

The interpretation of the parameter $\{F_x\}_{x \in X}$ is similar to the interpretation of the base measure of a Dirichlet process. The distribution F_x is the subjective prior

guess for the unknown probability model $\mathcal{P}_x$. From a data analysis perspective it is important that the prior can be centered on a prior guess and that the degree of variability of the RPM can be controlled to reflect the strength of the available prior information. The PT prior is a flexible probability model that can achieve both these objectives. See for example Lavine (1992,1994).

In the MPT a third aspect enters the prior elicitation. The investigator can specify the degree of dependence between the jointly modeled RPMs $\{\mathcal{P}_x\}_{x\in X}$. For example, consider a covariate space $X = \{x_1, x_2\}$ constituted of two points. The higher the degree of dependence between the two RPMs, the more borrowing of strength will occur between the two groups defined by the two covariates. In the extreme cases, if the random measures are independent the observations with covariate x_2 are not used to estimate the unknown distribution $\mathcal{P}_{x_1}$, while on the other hand if the random measures are almost surely identical then all data are pooled to carry out inference about the one common RPM.

The degree of dependence between jointly modeled random distributions is easiest characterized by the correlations of the random probabilities. For an exhaustive motivation of this approach we refer to Walker and Muliere (2003). They characterize the joint law of two Dirichlet processes, D_1 and D_2, in such a way that the quantity $Corr(D_1(B), D_2(B))$ is a fixed positive constant independent of the specific Borel subset B. In our case the correlations of interest can be easily evaluated for a rich class of Borel subsets. Given the covariate points x_1 and x_2, for every ordered couple of integers (i_1, i_2), consider

$$Corr\left\{\mathcal{P}_{x_1}\left(F_{x_1}^{-1}\left(\frac{i_1}{2^j}\right), F_{x_1}^{-1}\left(\frac{i_2}{2^j}\right)\right], \mathcal{P}_{x_2}\left(F_{x_2}^{-1}\left(\frac{i_1}{2^j}\right), F_{x_2}^{-1}\left(\frac{i_2}{2^j}\right)\right]\right\}. \quad (1.8)$$

Due to the structure of the MPT model the computation of the correlation coefficients can be reduced to the previously discussed problem of evaluating second order moments of bivariate marginal distributions of independent MBPs. Figure 1.2 illustrates, for a specific MPT parametrization, some of the correlations that can be computed following the outlined procedure.

As desired, the closer the two covariates, the stronger the dependence between the

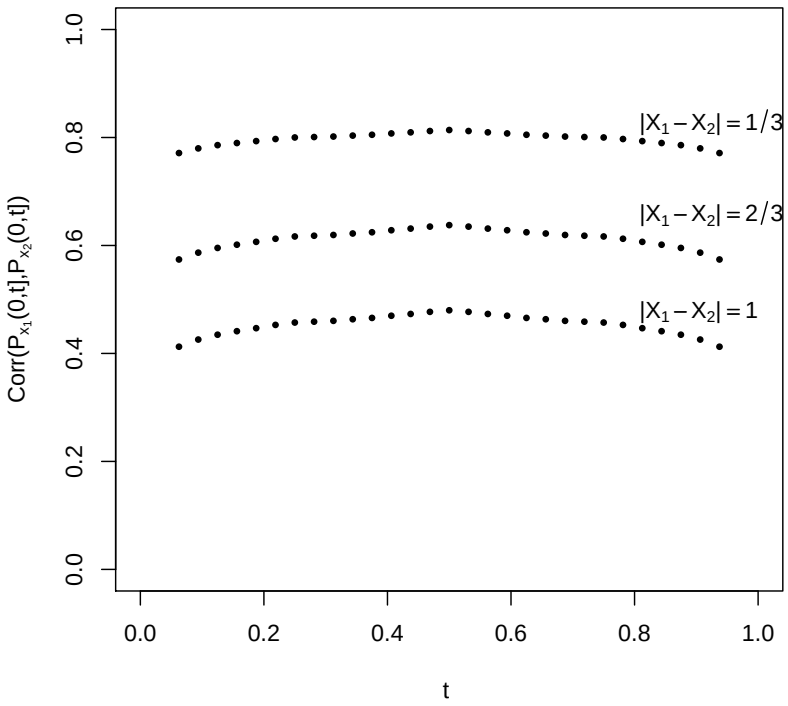


Figure 1.2: Correlation between $\mathcal{P}_{x_1}(0,t]$ and $\mathcal{P}_{x_2}(0,t]$. Here $\{\mathcal{P}_x\}_{x\in X} \sim \text{MPT}(\mathcal{A}, Q, F)$, $X = \mathbb{R}$, $\alpha_\varepsilon = 2$ for every $\alpha_\varepsilon \in \mathcal{A}$, F_x is the uniform distribution function on $[0,1]$ for every x and $\{q_x\}_{x\in X}$ is the family of Normal kernels with variance $\sigma^2 = 2$.

corresponding random measures. The parameters $(\mathcal{A}, F)$ of the MPT characterize the marginal distribution of a single random measure $\mathcal{P}_x$. If $(\mathcal{A}, F)$ are fixed, then the choice of Q allows to flexibly model the strength of dependence between the random measures. Consider, for example, the MPT parametrization adopted in Figure 1.2. In this case the effects of the choice of σ^2 on the prior are clear. Following the same arguments used earlier to discuss the relationship between Q and the correlation function of $\{Y_x\}_{x\in X} \sim \text{MBP}(\alpha_0, \alpha_1, Q)$, it can be shown that the ratio $|x_1 - x_2|/\sigma$ determines the degree of dependence between $\mathcal{P}_{x_1}$ and $\mathcal{P}_{x_2}$.

1.3.3 Properties of the MPT

The next proposition asserts that for a sequence of covariate points $\{x_i\}_{i\geq 1}$ converging to x, the differences between the associated random distributions, $\mathcal{P}_{x_i}$ and $\mathcal{P}_x$, become negligible.

Proposition 1.3. Continuity in x. *Given $\{x_i \in X\}_{i\geq 1}$ and $x \in X$, if*

$$\lim_i [\sup_{B\in\mathcal{B}^k} |Q_{x_i}(B) - Q_x(B)|] \to 0 \quad \textit{and} \quad \lim_i [\sup_{B\in\mathcal{B}} |F_{x_i}(B) - F_x(B)|] \to 0$$

then for every Borel subset B of the real line the sequence $\{\mathcal{P}_{x_i}(B)\}_{i\geq 1}$ converges in probability to $\mathcal{P}_x(B)$.
(The proof is in the appendix)

If, for example, F is a Weibull model and Q is a multivariate normal location family, both assumptions of the above proposition can be verified, for every sequence $\{x_i\}_{i\geq 1}$ converging to x, through dominated convergence arguments

Next we characterize the support of $\mathcal{P}_x$, marginally for each x. The support includes all probability distributions that are absolutely continuous with respect to F_x. If the support of F_x is the real line then the law has full weak support. Moreover, the support remains the same even conditional on restrictions on $\mathcal{P}_{x_i}$ for neighboring x_i. Specifically, let $\{x_1, \ldots, x_m\}$ denote distinct covariate points. The RPM $\mathcal{P}_x$ still has full support, even conditional on $\mathcal{P}_{x_1}, \ldots, \mathcal{P}_{x_m}$ belonging to weak neighborhoods $U_1, \ldots, U_m$, respectively. Similar considerations hold if the support of F_x is the positive real line.

This property distinguishes the MPT from many other Bayesian models for heterogeneous populations. Consider, for example, a Bayesian PH model F, when the covariate space X is a subset of the real line. For any monotone bounded continuous function f the map $x \to \int f dF_x$ is monotone. If $x_1 < x_2 < x_3$ and it is known that the expected values of f associated with x_1 and x_3 belong to the interval (a, b), then also $\int f dF_{x_2}$ belongs to (a, b). In contrast, assume that an MPT is centered on a PH model and it is known that $\int f d\mathcal{P}_{x_1}$ and $\int f d\mathcal{P}_{x_3}$ are in (a, b). The conditional probability that $\int f d\mathcal{P}_{x_2}$ is outside (a, b) still remains strictly positive.

Proposition 1.4. Support of the MPT. *For every strictly positive δ, for every $m = 1, 2, \ldots$, distinct $x_1, \ldots, x_m \in X$, and for any family of probability measures $\mathcal{S}_1, \ldots, \mathcal{S}_m$ such that $\mathcal{S}_1 \ll F_{x_1}, \ldots, \mathcal{S}_m \ll F_{x_m}$, the following holds. If $[V_1, V_2, \ldots, V_k]$ is a partition of the real line into intervals V_j, then the event*

$$\{|\mathcal{P}_{x_i}(V_j) - \mathcal{S}_i(V_j)| < \delta\,; \quad \forall i \in (1, \ldots, m), \quad \forall j \in (1, \ldots, k)\}$$

has strictly positive probability.
(The proof is in the appendix)

The proposition guarantees that for any finite sequence of continuous and bounded real functions $\{f_1, \ldots, f_l\}$ and for every positive ε the event

$$\{\bigcap_{i=1}^{m}\bigcap_{j=1}^{l}\left|\int f_j d\mathcal{P}_{x_i} - \int f_j d\mathcal{S}_i\right| < \varepsilon\}$$

has a priori strictly positive probability.

The next result gives a desirable asymptotic property of the MPT model. Let $\{x_i\}_{i\in\{1,\ldots,n\}}$ be a vector of covariates. Let $p(\mathcal{P}_x \mid \{x_i, W_i\}_{i\in\{1,\ldots,n\}})$ denote the conditional law of the RPM $\mathcal{P}_x$ given conditionally independent *r.v.*'s $\{W_i\}_{i\in\{1,\ldots,n\}}$ with

$$W_i \mid \{\mathcal{P}_x\}_{x\in X} \sim \mathcal{P}_{x_i} \quad \text{and} \quad \{\mathcal{P}_x\}_{x\in X} \sim \mathrm{MPT}(\mathcal{A}, Q, F)\cdot$$

Finally, let $\mathfrak{F}$ denote the set of all the distributions on the real line.

Theorem 1.1. Consistency of the MPT. *Let $X^\star = \{x_1^\star, \ldots, x_m^\star\}$ be distinct points of the covariate space. Let $\{x_i \in X^\star\}_{i\geq 1}$ be a sequence such that*

$$\sum_{i=1}^{n} I(x_i = x_j^\star) \to \infty \qquad \forall j \in \{1, \ldots, m\}\cdot$$

Consider bounded continuous real-valued functions $f_1, \ldots, f_l$, a vector of distributions $\mathcal{S}_{x_1^\star}, \ldots, \mathcal{S}_{x_m^\star}$ and a sequence of independent r.v.'s $\{W_i\}_{i\geq 1}$, $W_i \sim \mathcal{S}_{x_i}$. If the distribution functions $F_{x_1^\star}, \ldots, F_{x_m^\star}$ are strictly monotone then for every strictly positive ε and $j \in \{1, \ldots, m\}$

$$p(\mathcal{P}_{x_j^\star} \in \Delta_{j,\varepsilon} \mid \{x_i, W_i\}_{i\in\{1,\ldots,n\}}) \to 1 \qquad a.s.$$

where $\Delta_{j,\varepsilon} = \{P \in \mathfrak{F} : |\int f_i dP - \int f_i d\mathcal{S}_j| < \varepsilon, \quad \forall i \in (1, \ldots, l)\}$.
(The proof is in the appendix.)

In survival analysis we are interested in distributions $\mathcal{S}_{x_1^\star}, \ldots, \mathcal{S}_{x_m^\star}$ with support on the non-negative numbers. Assume that $F_{x_1^\star}, \ldots, F_{x_m^\star}$ are strictly monotone on $[0, \infty)$. Then it can be verified that the asymptotic property still holds. The theorem guarantees that conditional on an increasing number of observations under each of a finite number of covariate the posterior distribution eventually concentrates around the unknown true distributions. For example $\{x_1^\star, \ldots, x_m^\star\}$ could be different dose levels of an experimental drug administered to an homogeneous population and the MPT could be centered on a Weibull model. Posterior inference about the unknown survival functions is weakly consistent.

1.4 Mixture of MPTs: A Nonparametric PH Model

We introduce a hierarchical extension of the MPT model. The extension is similar to mixtures of PT models. Mixtures of PTs have been studied by several authors, including Hanson (2006), Hanson and Johnson (2002) and Berger and Guglielmi (2001). A mixture of PTs is defined as the law of a RPM $\mathcal{P}$, such that conditional on a random parameter θ, the random measure $\mathcal{P}$ has a PT distribution centered on a parametric distribution F_θ. A number of desirable properties of this nonparametric prior have been discussed in literature. Mixtures of PTs naturally generalize and enrich the support of a parametric prior. For example, Hanson and Johnson (2002) illustrate the advantages that can be achieved by modeling the residual distribution in a linear regression model through a mixture of PTs, and compare the mixture model with the PT prior without mixture. The prior on θ can be chosen to reflect prior knowledge about the unknown distribution. The PT in the second level of the hierarchical model allows us to extend the support of the prior. The parametrization of the PT prior controls the degree

of variability of the RPM $\mathcal{P}$ around F_θ. Moreover the parametrization of the PT model can be chosen to have $\mathcal{P}$ almost surely absolutely continuous.

Another important advantage of the mixture of PT models compared to the PT prior is discussed in Lavine (1994). The predictive model is not strictly dependent on a fixed sequence of nested partitions of the real line. The awkward sensitivity of the inference with respect to the choice of the partitions is mitigated, and posterior predictive distributions are usually smoother and more regular than under a PT prior.

A mixture of MPTs is a natural extension of the mixtures of PT models. We consider a class of RPMs $\{\mathcal{P}_x\}_{x\in X}$ indexed by covariates x such that conditionally on a random parameter θ the process $\{\mathcal{P}_x\}$ is a MPT centered on F_θ,

$$\{\mathcal{P}_x\}_{x\in X} \mid \theta \sim MPT(\mathcal{A}, Q, F_\theta) \quad \text{and} \quad \theta \sim \lambda.$$

The parameter θ indexes a regression model and therefore a whole class of distributions. In the sequel $\{\mathcal{P}_x\}_{x\in X}$ will be centered on a survival regression model. Simple modification allows to adapt the framework to other regression problems. Let $\{W_i\}_{i=1}^n$ denote survival times, possibly censored from the right, of a sample of subjects with covariates $\{x_i\}_{i=1}^n$. Let $T_i = \min(W_i, C_i)$ and $\delta_i = I(W_i \leq C_i)$ denote the observed times and censoring indicators. Here C_i is the i-th censoring time. It is assumed that the censoring times are noninformative: $C_1, \ldots C_n$ are independently distributed with respect to the random distributions of the survival times.

We propose the following hierarchical nonparametric Bayesian model for time to event data. Let $H_{\theta_1}(\cdot)$ be a continuous cumulative hazard function with parameter $\theta_1 \in \Theta_1$, let $L_{\theta_2}(\cdot)$ be a link function with parameter $\theta_2 \in \Theta_2$. Define $F_{x,\theta_1,\theta_2}(\cdot) = 1 - \exp[-H_{\theta_1}(\cdot)L_{\theta_2}(x)]$ and $F_\theta = \{F_{x,\theta_1,\theta_2}(\cdot)\}_{x\in X}$. We define the survival regression model

$$(\theta_1, \theta_2) \sim \lambda\,,$$

$$\{\mathcal{P}_x\} \mid \theta \sim \mathrm{MPT}(\mathcal{A}, Q, F_\theta)\,,$$

and the $r.v.$'s W_i, $i = 1, \ldots, n$, are conditionally independently distributed

$$W_i \mid \{\mathcal{P}_x\} \sim \mathcal{P}_{x_i} \,.$$

Under this model, for every covariate point x the (marginal) law of the RPM $\mathcal{P}_x$ is a mixture of PTs.

In the examples that we propose the first level of the hierarchical prior is a Weibull model: $\theta_1 = (\theta_{11}, \theta_{12})$, $H_{\theta_1}(t) = \theta_{11} t^{\theta_{12}}$, and $L_{\theta_2}(x) = \exp(\theta_2 x)$. A vague prior is adopted for the regression parameters: θ_{12} and θ_2 are independently normally distributed with null mean and large variances while θ_{11} is exponentially distributed with mean equal to 1.

1.5 Posterior Inference

1.5.1 Posterior inference with MPT models

In this section we state an algorithm to implement posterior inference. The interest is in the conditional distribution of the random distributions $\{\mathcal{P}_x\}_{x \in X}$ given the data $\{W_i\}_{i=1}^n$ with covariate values $\{x_i\}_{i=1}^n$:

$$\{\mathcal{P}_x\}_{x \in X} \sim MPT(\alpha_0, \alpha_1, Q) \quad \text{and} \quad W_i | \{\mathcal{P}_x\}_{x \in X} \sim \mathcal{P}_{x_i} \quad i = 1, \ldots, n.$$

The Multivariate Beta processes of the MPT model are both a priori and a posteriori independently distributed. Indeed, given the data $\{W_i\}_{i=1}^n$, a sufficient statistics for updating the distribution of $\{Y_{\varepsilon_1}, \ldots, Y_{\varepsilon_m}\}$, where $\{\varepsilon_1, \ldots, \varepsilon_m\}$ are arbitrarily chosen finite dyadic sequences having maximum length $\tilde{k}$, consists in the array of indicators

$$I_{\tilde{k}} = \left\{ I\left(F_{x_i}(\frac{j}{2^{\tilde{k}}}) < W_i \leq F_{x_i}(\frac{j+1}{2^{\tilde{k}}}) \right) ; i \in (1, \ldots, n), j \in (1, \ldots, 2^{\tilde{k}} - 1) \right\}$$

and the distribution of the array conditionally on $\{Y_{\varepsilon_1}, \ldots, Y_{\varepsilon_m}\}$, can be conveniently factorized simply exploiting the equality (1.7).

From this fact it follows that to compute the quantities

$$E\left(\mathcal{P}_x\Big(F_x^{-1}(\frac{j_1}{2^k}), F_x^{-1}(\frac{j_2}{2^k}) \Big] \,\Big|\, \{W_i\}_{i=1}^n \right)$$

for any ordered couple of integers (j_1, j_2), with the purpose of approximating the predictive distributions, it suffices to quantify the conditional expectations $E(Y_{\varepsilon,x}|\{W_i\}_{i=1}^n)$ for every ε belonging to an adequately selected finite subset of $(\cup_{j=0}^{\infty}\{0,1\}^j)$. It becomes then apparent, writing the likelihood of the array I_k, that the computation of the predictive probabilities reduces to binary regression problems in which the interest is in the a posteriori expectation of the *r.v.* Y_x conditionally on a finite sequence of observations $\{Z_i, \ldots, Z_n\}$ distributed as in (1.3).

From the characterization (1.2) it follows that

$$E(Y_x|Z_1,\ldots,Z_n) = \frac{\int \frac{D(A_x^0)}{D(A_x^0)+D(A_x)} \prod_{i=1}^{n} \frac{D(A_{x_i}^0)^{Z_i} D(A_{x_i}^1)^{1-Z_i}}{D(A_{x_i}^0)+D(A_{x_i}^1)} \; d\mathcal{D}_{x_1,\ldots,x_n,x}}{\int \prod_{i=1}^{n} \frac{D(A_{x_i}^0)^{Z_i} D(A_{x_i}^1)^{1-Z_i}}{D(A_{x_i}^0)+D(A_{x_i}^1)} \; d\mathcal{D}_{x_1,\ldots,x_n,x}} \tag{1.9}$$

where $\mathcal{D}_{x_1,\ldots,x_n,x}$ denotes the law of the Dirichlet process D having parameter $\nu_{x_1,\ldots,x_n,x}$. Both the integrals in (1.9) can be approximated through a Monte Carlo procedure sampling iteratively the Dirichlet process and computing at each iteration the integrand functions. The Dirichlet process can be approximately sampled exploiting the stick-breaking representation given in Sethuraman (1994). For a formal study about the application of the stick-breaking representation to approximate the distribution and the expected value of a functional of a Dirichlet process we refer to Muliere and Tardella (1998); they give theoretical basis for using the representation in Monte Carlo procedures.

In Figure 3 an example in which the conditional expectations $E(Y_x|Z_1,\ldots,Z_n)$ are approximated through the outlined algorithm is illustrated. The dotted line in the same graph is the estimate of the regression curve obtained with an alternative nonparametric prior studied in Coram and Lalley (2006).

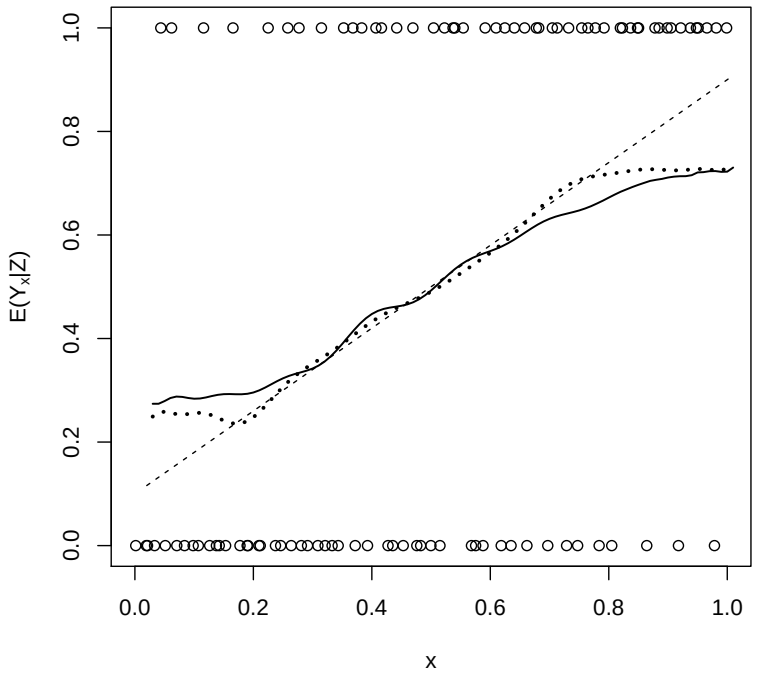


Figure 3: Dashed line: the retrogression curve from which have been generate the binary r.v.'s Solid lines: estimate of the regression curve. Prior parametrization: $\alpha_0 = \alpha_1 = 2$*,* q_x *is the gaussian kernel with mean* x *and variance equal to 1.*

The described procedure results simple and adequately efficient but allows exclusively to approximate the predictive distribution functions. To assess the precision of the estimates it is necessary to approximate a posteriori the marginal distributions of the Multivariate Beta processes of the MPT model. The credible intervals of the random quantities $\mathcal{P}_x\Big(F_x^{-1}(\frac{j_1}{2^k}), F_x^{-1}(\frac{j_2}{2^k})\Big]$ can be a posteriori computed sampling iteratively from the conditional distribution of $\{Y_{\varepsilon_1,x}, \ldots, Y_{\varepsilon_m,x}\}$ given $\{W_i\}_{i=1}^n$, where $\{\varepsilon_1, \ldots, \varepsilon_m\}$ is an adequately selected set of finite binary sequences.

The sampling technique introduced in the previous section to generate a binary vector $[Z_1, \ldots, Z_n]$ distributed as in (1.3) suggests the structure of a Gibbs sampling algorithm to generate *r.v.*'s from the conditional distribution of Y_x given $\{Z_1, \ldots, Z_n\}$. We underline again that $\{Y_{\varepsilon_1,x}, \ldots, Y_{\varepsilon_m,x}\}$ are a posteriori inde-

pendent $r.v.$'s and that the problem of sampling $Y_{\varepsilon_j,x}$ conditionally on $\{W_i\}_{i=1}^n$ is identical to that of sampling Y_x given $\{Z_1,\ldots,Z_n\}$.

The procedure based on the Polya urn scheme that allows to sample the partially exchangeable binary $r.v.$s $\{Z_1,\ldots,Z_n,Z_{n+1}\}$ associated with the covariate points $\{x_1,\ldots,x_n,x\}$ splits the Polya sequence $\{h_i,l_i\}_{i\geq 1}$ in subsequences $(L_1=\{h_i,l_i\}_{i=1}^{i_1},\ldots,L_{n+1}=\{h_i,l_i\}_{i=i_n+1}^{i_n})$, where $\{i_1,\ldots,i_{n+1}\}$ are the random integers recursively defined in Section 1. This scheme clarify that the posterior distribution of Y_x is a mixture of beta distributions, indeed the limit composition of the Polya urn, conditionally on the latent subsequences $(L_1,\ldots,L_n)$ is still Dirichlet distributed. More formally

$$P(Y_x\in\cdot|\{Z_i\}_{i=1}^n)=P(\frac{D(A_x^0)}{D(A_x^0)+D(A_x^1)}\in\cdot|\{Z_i\}_{i=1}^n)=$$

$$\int\int_{\cdot}\beta\Big(s,\alpha_0(x,L_1,\ldots,L_n),\alpha_1(x,L_1,\ldots,L_n)\Big)\,ds\,dP(L_1,\ldots,L_n|\{Z_i\}_{i=1}^n)$$

where the first integral is over the possible configurations of $(L_1,\ldots,L_n)$, $\beta(\cdot,\alpha_0,\alpha_1)$ denotes the beta density and

$$\alpha_j(x,L_1,\ldots,L_n)=\alpha_j+\sum_{i=1}^{i_n}I(-j\alpha_1q_x(l_i)<h_i<(1-j)\alpha_0q_x(l_i))\quad\text{for}\quad j\in\{0,1\}\cdot$$

The equality is obtained exploiting the fact that the Dirichlet process is conjugate.

To sample $\{L_1,\ldots,L_n\}$ conditionally on $\{Z_i\}_{i=1}^n$ we sample iteratively from the full conditional of L_j given $\{L_1,\ldots,L_{j-1},Lj+1,\ldots,L_n\}$ and $\{Z_i\}_{i=1}^n$; the iterations simulates a Markov chain with the desired invariant distribution. The expression

$$P(L_j\in\cdot|L_1,\ldots,L_{j-1},L_{j+1},\ldots,L_n,\{Z_i\}_{i=1}^n)\propto$$

$$P(L_j\in\cdot|L_1,\ldots,L_{j-1},L_{j+1},\ldots,L_n)\;I\Big(-(1-Z_j)\alpha_1q_{x_j}(l_{i_j})<h_{i_j}<Z_j\alpha_0q_{x_j}(l_{i_j})\Big)$$

clarifies that to sample from the full conditional it suffices to sample from the Polya urn updated by $(L_1,\ldots,L_{j-1},L_{j+1},\ldots,L_n)$, till a ball with color belonging to $(A_{x_j}^0\cup A_{x_j}^1)$ is drawn, and to substitute, if necessary, only the last ball: if the last ball belongs to the color set $\{A_{x_j}^0\}$ and $Z_j=0$ than it is substituted with a

ball randomly selected from the updated composition of the Polya urn restricted to $\{A^0_{x_j}\}$ and vice versa.

Assuming that the parametrization of the MPT $\{\mathcal{P}_x\}_{x\in X}$ is such that $\mathcal{P}_x$ is $a.s.$ absolutely continuous for every $x \in X$, the Polya urn scheme can be adapted to generate from the partially exchangeable law of the random vector $(W_1, \ldots, W_n)$. Indeed we could sample the subsequences $(L^0_1, \ldots, L^0_n)$ in such a way that

$$P(I(h_{i^0_1} > 0), \ldots, I(h_{i^0_n} > 0)) = P(I(F^{-1}_{x_1}(\frac{1}{2}) < W_1), \ldots, I(F^{-1}_{x_n}(\frac{1}{2}) < W_n)),$$

continue sampling from the urns in the inferior levels of the MPT till level $\tilde{k}$, where

$$\tilde{k} = \inf\{k \in \mathbb{N} : \sum_i I(F^{-1}_{x_i}(\frac{j}{2^k}) < W_i \leq F^{-1}_{x_i}(\frac{j+1}{2^k})) \leq 1 \quad \forall j \in (0, \ldots, 2^k - 1)\} \tag{1.10}$$

and finally sample each of the $r.v.$s in $(W_1, \ldots, W_n)$ from the adequately restricted distributions functions $(F_{x_1}, \ldots, F_{x_n})$. An alternative way to generate the random vector consists in generating a sequence of processes $\{D^0, D^{00}, D^{01}, \ldots\}$ characterized as in (1.2), in this case it is necessary to generate only a finite number of Dirichlet processes: it suffices to sample at the various levels of the MPT the indicators I_k and stop the procedure at level $\tilde{k}$. Note that after having generated the Dirichlet processes we would be able to conditionally sample from the Polya urns; for example after D^0 has been generated it is trivial to conditionally sample $(L^0_1, \ldots, L^0_n)$.

1.5.2 Posterior inference with mixtures of MPT models

The considerations in the above paragraph suggest how to generate a Markov chain in order to approximate a posteriori the distribution of θ when

$$\theta \sim \lambda, \qquad \mathcal{P}|\theta \sim MPT(\mathcal{A}, Q, F_\theta)$$

and the $r.v.$'s $\{W_i\}_{i=1}^n$ are conditionally independent, $W_j|\theta, \mathcal{P} \sim \mathcal{P}_{x_j}$. In the following paragraphs we consider the joint law of θ, $\{W_i\}_{i=1}^n$, of the processes

$(D^0, D^{01}, \dots)$ and $(L_0^1, L_0^2, \dots)$ and of a nuisance parameter θ^* that will be better specified. The objective is to construct conditionally on $\{W_i\}_{i=1}^n$ a Markov chain whose invariant distribution is the conditional joint distribution of θ and of a subset of the subsequences $(L_0^1, L_0^2, \dots)$ in such a way to allow to sample from the predictive distributions. The iterative procedure consists of three steps.

1)Assume that $\tilde{\theta}$ is sampled from the conditional distribution $\theta|\{W_i\}_{i=1}^n$, while $\tilde{\theta}^*$ is sampled from a proposal $p(\theta, \theta^*)$ and that $\{D^\varepsilon\}_{\varepsilon \in E(\tilde{\theta}, \tilde{\theta}^*)}$ are the limit compositions of a subset of the Polya urns of the MPT model; $E(\tilde{\theta}, \tilde{\theta}^*)$ identifies the finite dyadic sequences of length inferior to $\max(\tilde{k}_{\tilde{\theta}}, \tilde{k}_{\tilde{\theta}^*})$ where $\tilde{k}_{\tilde{\theta}}$ and $\tilde{k}_{\tilde{\theta}^*}$ are defined as in (1.10).

More formally, $\theta \sim \lambda$, $\theta^*|\theta \sim p_\theta$, $\{D^\varepsilon\}_{\varepsilon \in E}$ are the random limits of the Polya urns of the MPT model having laws independent with respect to (θ, θ^*) and the $r.v.$s $\{W_i\}_{i=1}^n$ are conditionally independently distributed with

$$P(W_i \in dw|\theta, \theta^*, \{D^\varepsilon\}_{\varepsilon \in E}) =$$

$$\lim_k \left(\prod_{j=0, \varepsilon_j \in \varepsilon^+}^{k} 2\frac{D^{\varepsilon_j}(A^0_{x_i})}{D^{\varepsilon_j}(A^0_{x_i}) + D^{\varepsilon_j}(A^1_{x_i})} \prod_{j=0, \varepsilon_j \in \varepsilon^-}^{k} 2\frac{D^{\varepsilon_j}(A^0_{x_i})}{D^{\varepsilon_j}(A^0_{x_i}) + D^{\varepsilon_j}(A^1_{x_i})} \right) dF_{x_i, \theta}(w)$$

where $(\varepsilon^+, \varepsilon^-, \varepsilon_0, \varepsilon_1, \dots)$ allows to track the dyadic expansion of $F_{x_i, \theta}(s)$. Note that the random sequence $\{D^\varepsilon\}_{\varepsilon \in E(\theta, \theta^*)}$ is well defined, it is a function of θ, θ^*, $\{W_i\}_{i=1}^n$ and $\{D^\varepsilon\}_{\varepsilon \in E})$. To evaluate the ratio

$$\frac{P(\theta \in d\tilde{\theta}, \theta^* \in d\tilde{\theta}^*|\{W_i\}_{i=1}^n, \{D^\varepsilon\}_{\varepsilon \in E(\theta, \theta^*)})}{P(\theta \in d\tilde{\theta}^*, \theta^* \in d\tilde{\theta}|\{W_i\}_{i=1}^n, \{D^\varepsilon\}_{\varepsilon \in E(\theta, \theta^*)})} \tag{1.11}$$

we can exploit the following expression

$$P(\theta \in d\tilde{\theta}, \theta^* \in d\tilde{\theta}^*|\{W_i\}_{i=1}^n, \{D^\varepsilon\}_{\varepsilon \in E(\theta, \theta^*)}) \propto$$

$$h\lambda(\tilde{\theta}) p_{\tilde{\theta}}(\tilde{\theta}^*) \prod_i dF_{x_i, \tilde{\theta}}(W_i) \prod_{j=1}^{\max(\tilde{k}_{\tilde{\theta}}, \tilde{k}_{\tilde{\theta}^*})} \frac{D^{\varepsilon_{ji}}(A^0_{x_i}) I_(\varepsilon_{ji} \in \varepsilon_i^+) + D^{\varepsilon_{ji}}(A^1_{x_i}) I(\varepsilon_{ji} \in \varepsilon_i^-)}{D^{\varepsilon_{ji}}(A^0_{x_i}) + D^{\varepsilon_{ji}}(A^1_{x_i})}$$

where h is equal to 1 if the event $\{\theta = \tilde{\theta}, \theta^* = \tilde{\theta}^*\}$ is compatible with the conditioning event and null otherwise. It is therefore possible to evaluate the ratio,

draw a uniform $r.v.$ U and accordingly with the standard rule of the Metropolis Hastings algorithm switch $\tilde{\theta}$ and $\tilde{\theta}^*$ or not .

2) The second step starts from a realization from the conditional distribution $\theta, \{D^{\varepsilon}\}_{E(\theta,\theta^*)} | \{W_i\}_1^n$ and the subsequences $(L_1^0, \ldots, L_n^0, \ldots)$ are sampled till the level $\tilde{k}_\theta$ conditionally on $(\theta,\ \{D^{\varepsilon}\}_{E(\theta,\theta^*)}$ and $\{W_i\}_1^n)$. We underline again that when θ and the limit compositions of the Polya urns are known it becomes simple to conditionally generates the latent sequences.

3)Finally the third step starts from a realization of θ and $(L_1^0, \ldots, L_n^{\tilde{k}_\theta})$ conditionally on $\{W_i\}_{i=1}^n$, a new proposal θ^* is generated conditionally on θ and the sequence $\{D^{\varepsilon}\}_{\varepsilon \in E(\theta,\theta^*)}$ is generated conditionally on θ, θ^*, $(L_1^0, \ldots, L_n^{\tilde{k}\theta})$ and $\{W_i\}_{i=1}^n$. We underline again that once the sequences $(L_1^0, \ldots, L_n^{\tilde{k}\theta})$ are known the processes till the level $\tilde{k}_\theta$ can be simulated exploiting the fact that the Dirichlet prior is conjugate. If $\tilde{k}_\theta < \tilde{k}_{\theta^*}$ the remaining subsequences till $L_n^{\tilde{k}_{\theta^*}}$ are sampled conditionally on $(\theta, \{W_i\}_{i=1}^n)$ taking a considerable advantage from the fact that from each urn need to be drown at most one subsequence before generating the Dirichlet processes.

Throughout the description of the procedure we have assumed to be able to generate the Dirichlet processes; the stick breaking representation allows to approximately sample the processes. The fact that it is not possible to generates all the countable atoms of the processes is not an issue, indeed the approximations allows to exactly perform the Metropolis Hastings step and to exactly generate the sequences $(L_1^0, \ldots, L_n^{\tilde{k}_\theta})$. The error term of the computation of the ratio (1.11) based on the truncated stick breaking series can always be bounded and reduced adding further atoms and random weights to the series till conditionally on the generated uniform $r.v.$ U it become evident the output of the first step of the algorithm. Similarly the sequences $(L_1^0, \ldots, L_n^{\tilde{k}_\theta})$ can be exactly sampled extending the truncated stick breaking series when necessary.

At each iteration of the procedure a realization of θ and of a subset of the sequences $(L_1^0, L_2^0, \dots)$ conditionally on the data is obtained; note that the subset can be arbitrarily extended. Once we are able to a posteriori generate both the parameter θ and the compositions of the Polya urns it become straightforward to sample from the predictive distribution and to sample arbitrarily precisely the random distribution $\mathcal{P}_x$.

A slight modification of the algorithm allows to deal with censured time to event data, the idea is to sample at each iteration the latent survival times. Let $\{W_i\}_{i=1}^n$ denotes a finite sequence of possibly censored from the right times to event. We adopt the notation $T_i = \min(W_i, C_i)$ and $\delta_i = I(W_i \leq C_i)$, where C_i is the i-th censoring time. It is assumed that the censoring times are noninformative: $C_1, \dots C_n$ are independently distributed with respect to the random distributions of the survival times.

Assume to start the first step of the procedure from a realization from

$$\theta, \theta^*, \{D^\varepsilon\}_{\varepsilon \in E(\theta,\theta^*)}, \{W_i\}_{i=1}^n | \{T_i, \delta_i\}\cdot$$

At the second step if $\delta_n = 0$, after having obtained a realization of $(L_1^0, \dots, L_{n-1}^0, \dots, L_{n-1}^{\tilde{k}_\theta})$, in this case

$$\tilde{k}_\theta = \inf\{k \in \mathbb{N} : \sum_{i=1}^{n-1} I(F_{x_i,\theta}^{-1}(\frac{j}{2^k}) < W_i \leq F_{x_i,\theta}^{-1}(\frac{j+1}{2^k})) \leq 1 \quad \forall j \in (0, \dots, 2^k - 1)\}$$

a random variable from the conditional distribution

$$W_n | \theta, \{W_i\}_{i=1}^{n-1}, W_n \geq T_n, (L_1^0, \dots, L_{n-1}^0, \dots, L_{n-1}^{\tilde{k}_\theta})$$

can be generated. To obtain a realization from

$$W_n | \theta, \{W_i\}_{i=1}^{n-1}, (L_1^0, \dots, L_{n-1}^0, \dots, L_{n-1}^{\tilde{k}_\theta})$$

it suffices to sample from the updated Polya urns till the level $\tilde{k}_\theta$, then it can happen that there exists j such that $L_j^{\tilde{k}_\theta}$ and $L_n^{\tilde{k}_\theta}$ come from the same urn or not. In the last case we need exclusively to generate W_n from the adequately restricts distribution $F_{x_n,\theta}$. In the first case a realization of $L_j^{\tilde{k}_\theta+1}$ coherent with

W_j is obtained and $L_n^{\tilde{k}_\theta+1}$ is sampled from the updated Polya urn; finally at the minimum level $\tilde{k}_\theta + m$ such that $L_j^{\tilde{k}_\theta+1}$ and $L_n^{\tilde{k}_\theta+1}$ come from two distinct urns W_n is generated from the adequately restricted distribution $F_{x_n,\theta}$. The described strategy can be adopted to conditionally sample all the censured survival times.

A natural simplification of the described algorithm consists in truncating the number of levels of the Polya Tree structure simply assuming a priori that for every ε of length greater than a fixed integer $Y_{\varepsilon,x} = \frac{1}{2} \quad \forall x \in X$.

1.6 Examples.

1.6.1 Example 1: A non-PH Survival Model

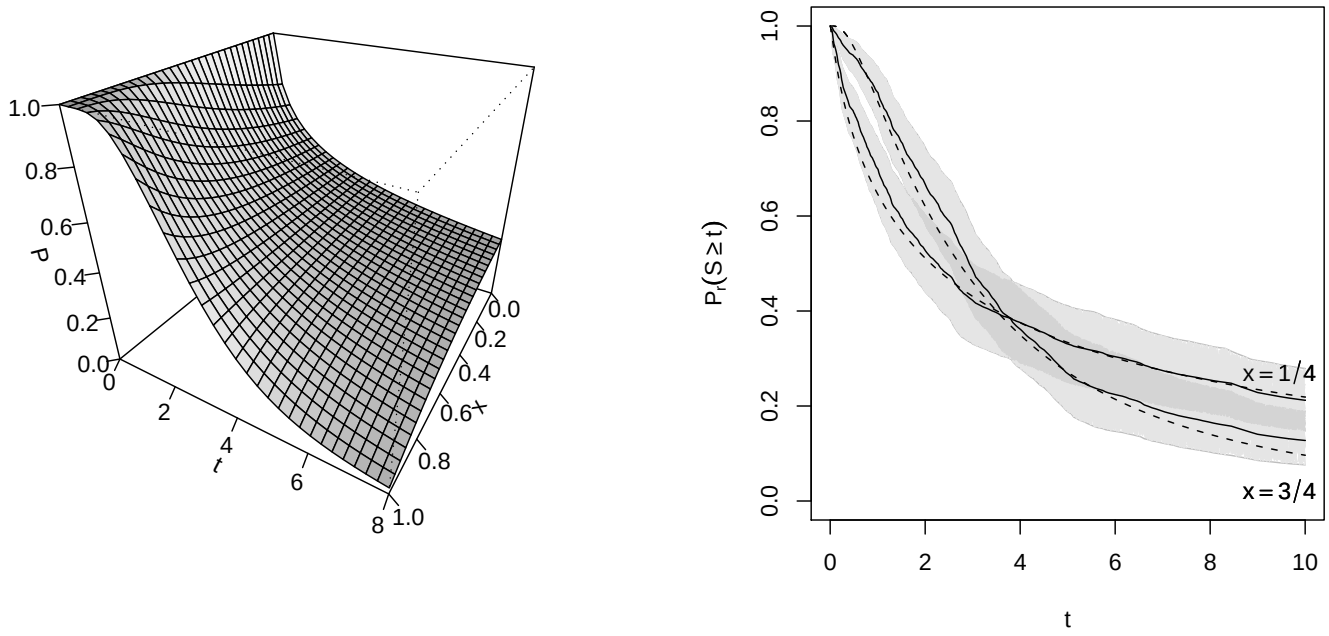
The first example illustrates the flexibility of the MPT model. We consider a simulated data set generated from the following model:

$$X = [0,1], \quad W_i|x_i \sim Lognormal(\mu_{x_i}, \sigma_{x_i}), \quad \mu_{x_i} = \frac{1}{2}x_i + \frac{5}{8}, \quad \sigma_{x_i} = \frac{5}{2} - 2x_i$$

and $C_i \sim Lognormal(2.4, 1)$. The implied probability of a single observation being censored is $p(C_i < W_i) \approx \frac{1}{5}$. The covariates x_i are uniformly sampled from the unit interval. As shown in Figure 1.3 the survival functions can not be adequately approximated by traditional survival models like the PH model, the accelerated failure time model or the proportional odds model. The survival curves are not stochastically ordered. A sample of 250 observations have been generated and fitted through the MPT model. The prior parameterization is the following: $F_x(t) = 1 - \exp(4t) \quad \forall x \in X$, $\alpha_\varepsilon = 4 \quad \forall \varepsilon \in E$ and $\{q_x\}_{x \in X}$ are normal kernels with variance equal to 1 and mean x.

We evaluated posterior inference using the previously outlined MCMC scheme. Some results are shown in Figure 1.3.

Posterior inference achieves a good balance between borrowing strength across covariate levels versus reporting meaningful covariate effects. These two goals are in conflict with each other. Excessive prior dependence between the random survival functions diminishes the ability to report meaningful covariate effects.



(a) (b)

Figure 1.3: Example 1: The left panel shows the simulation truth. The figure plots the survival functions $S_x(w) \equiv P(W > w \mid x)$ against w and x. The right panel shows inference for $x = 1/4$ and $x = 3/4$. The solid curves are the estimated survival functions $E(S_x(w) \mid data)$ under the proposed model. The dashed lines show the simulation truth. The shaded bands show central 90% credible intervals.

On the other, in the absence of prior dependence we would not borrow any strength across x. The relevant sample size for inference on the unknown survival function S_x would be restricted to only observations at that covariate level, and inference for S_x, $x \notin \{x_i;\, i = 1, \ldots, n\}$, would be impossible.

The reported inference also highlights that the model avoids the restrictions of previously mentioned semi-parametric models for survival regression. At the same time the model includes enough structure and smoothing to still provide meaningful estimates of the unknown survival functions.

1.6.2 Example 2: A PH Model

The second example demonstrates that, non surprisingly, simpler ad hoc inferential procedures can outperform the MPT model when the assumptions of the PH model hold. However, slight violations can suffice to lead to comparable performances of the two approaches.

We consider two scenarios. Under the first scenario the survival regression satisfies the PH assumption.

$$P_r(W > w) = \left(\frac{4}{4+w}\right)^{4+x}.$$

The second one introduces a minor violation of the PH assumption:

$$P_r(W > w) = \left(1 - \frac{1}{10}x\right)\left(\frac{4}{4+w}\right)^{4+x} + \left(\frac{1}{10}x\right)\exp\left(-\frac{1}{10}\left(\frac{5}{4}w\right)^3\right).$$

In both scenarios a binary covariate space is assumed, $x \in X \equiv \{0, 1\}$.

We compare inference under the MPT mixture model versus the Bayesian bootstrap for PH models studied in Kim and Lee (2003, b). Table 1.1 reports mean squared error (MSE) for estimating $S_x(w)$ under the two models, as well as Kaplan-Meier estimates. Estimates are based on $n = 50$ observations on each of the two arms, $x = 0$ and $x = 1$. MSEs are reported for $w \in \{0.5, 1, 1.5, 2\}$ and $x = 1$.

The MPT mixture is centered on the Weibull model and the MPT structure is parameterized as follows: $\alpha_\varepsilon = 4, \forall \varepsilon \in E$ and q_x are normal kernels with standard

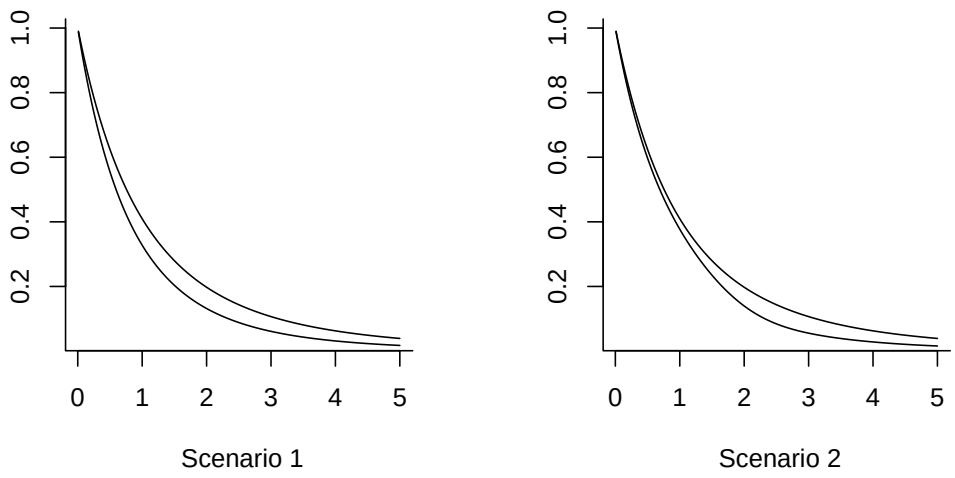


Figure 1.4: Example 2: $S_x(w)$, $x = 0, 1$. Simulation truth under the two scenarios.

deviation equal to 3.

In both scenarios the regression models perform better than the simple Kaplan-Meier estimator, the Bayesian bootstrap is the optimal choice in the first scenario, while in the second scenario the MSEs under the mixture of MPT model is comparable to the Bayesian bootstrap.

1.6.3 Example 3: A Lung Cancer Trial

We apply the proposed hierarchical model to the analysis of survival data from a clinical trial performed by the Lung Cancer Study Group (Lad et. al., 1988). The objective of the trial was to determine the potential benefit of adjuvant chemotherapy for patients with incompletely resected non-small-cell lung cancer. The patients enrolled in the study were randomly treated with a radiotherapy, which was the standard of care, or with the radiotherapy plus a chemotherapy. The trial proved the beneficial effect of the chemotherapy as an adjuvant to the radiotherapy. The survival data are published in Piantadosi (1997); 164 patients were enrolled in the trial and 28 were alive at the end of the follow up period.

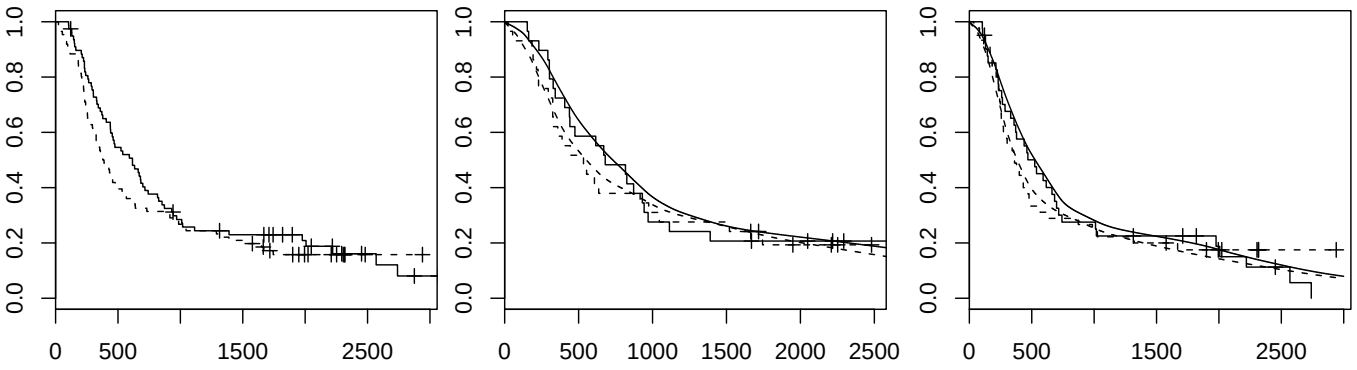
Table 1.1: Example 2: MSE for point estimates of $S_x(w)$ for four values of w and $x = 1$. MSEs for the Bayesian bootstrap (BB) and the mixture of MPT models (MMPT) are evaluated by Monte Carlo integration using $N = 100$ simluations of the inferential procedures.

MSE of $\hat{P}_r(W > w \mid x = 1)$					
		$w = 0.5$	$w = 1.0$	$w = 1.5$	$w = 2.0$
Scenario 1	MMPT	0.0039	0.0036	0.0028	0.0023
	BB	0.0035	0.0033	0.0028	0.0021
	KM	0.0049	0.0044	0.0032	0.0023
Scenario 2	MMPT	0.0037	0.0035	0.0031	0.0022
	BB	0.0038	0.0037	0.0029	0.0022
	KM	0.0048	0.0046	0.0035	0.0024

The most relevant prognostic factors are cancer histology (coded as $X_1 = 0$ for squamous and $X_1 = 1$ for non squamous); performance status at the beginning of the trial ($X_2 = 0$ for Kernofsky indicator ≥ 7, and $X_2 = 1$ otherwise); and treatment assignment ($X_0 = 1$ for treatment and $X_0 = 0$ for control).

The data are shown in Figure 1.5a. The Kaplan-Meier survival functions cross around $w = 1000$ days. For a while after the crossing the differences are negligible. At the end of the third year the survival function for the treatment group is below the curve for control. The details of the prior are as in the previous example, $\alpha_\varepsilon = 4 \ \forall \varepsilon \in E$. The kernels q_x in this case are multivariate normal densities with diagonal covariance matrix and the three standard deviations equal to 3.

Figure 1.5 summarizes posterior inference. The estimates correctly identify the crossing survival functions. This is the case despite prior centering around the PH structure which would not allow such peculiarities. However, in contrast to the PH model the proposed MPT model allows posterior inference to violate the structure of the prior centering measure if the data so requires.



(a) KM curves (b) Non-squamous and good performance status (c) Squamous

Figure 1.5: Example 3: Panel (a) shows the data. Panels (b) and (c) show the estimated survival curves for treatment (solid) and control (dotted lines). For comparison the Kaplan-Meier curves are included.

1.7 Discussion

We introduced a new prior for dependent RPMs. The proposed MPT model allows us to relax the assumptions of a Bayesian parametric model by extending the support of the unknown distribution functions.

The construction of probability models for dependent random measures has been studied by several authors in recent years.

Among these MacEachern (1999) and De Iorio et. al. (2004) explore the idea of defining dependent RPMs by means of an extension of the Dirichlet process mixture model while Dunson and Park (2008) discuss an extension of the stick-breaking representation.

The definition of the MPT model is based on the construction of dependent Beta $r.v.$'s through overlapping gamma processes. The MBP characterization can easily be modified to obtain dependent Dirichlet random vectors or dependent Dirichlet processes. Let G be an homogeneous gamma process on $\mathbb{R} \times (0, 1)$ and a a positive constant. Normalizing the random measurers

$$G_x(A) \equiv G([x, x + a] \times A) \qquad \forall A \in \mathcal{B},\, x \in \mathbb{R}$$

we can obtain dependent Dirichlet processes indexed by $x \in \mathbb{R}$.

The two main guiding principles in the proposed construction of dependent RPMs are the need of specifying prior distributions with a clear interpretation and the aim of constructing a probability model with large support.

A natural continuation to the above line of research is the study of the asymptotic properties of Bayesian inferential procedures based on the proposed class of dependent RPMs. We hope that the strict similarities of the MPT model with the tail free prior distributions could result helpful in future research for exploring its asymptotic behavior.

1.8 Appendix

Proposition 1.5.

Let be α_0, α_1 two fixed strictly positive real values and $r_2^0 = r_1^0 = \alpha_0(1-a), \quad r_c^0 = \alpha_0 a, \quad r_2^1 = r_1^1 = \alpha_1(1-a), \quad r_c^1 = \alpha_1 a$, *where a is a positive constant in $(0,1)$. If $G_1^0, G_2^0, G_1^1, G_2^1, G_c^0, G_c^1$ are independent gamma r.v. with same scale parameter and shape parameters $r_1^0, r_2^0, r_1^1, r_2^1, r_c^0, r_c^1$ the function*

$$\rho(a) = Correlation(\frac{G_1^0 + G_c^0}{G_1^0 + G_c^0 + G_1^1 + G_c^1}, \frac{G_2^0 + G_c^0}{G_2^0 + G_c^0 + G_2^1 + G_c^1})$$

is decreasing.

Proof. It is know that the vector

$$g = [\frac{G_1^0}{G_1^0 + G_2^0 + G_1^1 + G_2^1 + G_c^0 + G_c^1}, \ldots, \frac{G_c^1}{G_1^0 + G_2^0 + G_1^1 + G_2^1 + G_c^0 + G_c^1}]$$

is Dirichlet distributed with parameter $(r_1^0, r_2^0, r_1^1, r_2^1, r_c^0, r_c^1)$.

Let $\{l_i\}_{i\geq 1}$ be a Blackwell-McQueen random sequence from an urn with initial weights $(r_1^0, r_2^0, r_1^1, r_2^1, r_c^0, r_c^1)$ and colors $(C_1^0, C_2^0, C_1^1, C_2^1, C_c^0, C_c^1)$. Let S be the first drawn ball from the Bleckwel-McQeen urn whose color belongs to the colors set $(C_1^0, C_1^1, C_c^0, C_c^1)$, Finally, let V be the first drawn ball, after that S has been extracted, whose color belong to the colors set $(C_2^0, C_2^1, C_c^0, C_c^1)$.

$$E(\frac{G_1^0 + G_c^0}{G_1^0 + G_c^0 + G_1^1 + G_c^1}\frac{G_2^0 + G_c^0}{G_2^0 + G_c^0 + G_2^1 + G_c^1}) = P(S \in \{C_1^0, C_c^0\}, V \in \{C_2^0, C_c^0\}) =$$
$$= \sum_{m=1}^{\infty} P(\bigcap_{k=1}^{m-1} l_k \in \{C_2^0, C_2^1\}, l_m \in \{C_1^0, C_c^0\}) \times$$
$$\left[P(V \in \{C_2^0, C_c^0\} | \bigcap_{k=1}^{m-1} l_k \in \{C_2^0, C_2^1\}, l_m \in \{C_1^0, C_c^0\}) \right] \quad (1.12)$$

For each component of the sum the first product term is equal to

$$p'(a,m) = \frac{\alpha_0}{(2-a)(\alpha_0+\alpha_1) + m - 1} \prod_{k=1}^{m-1} \frac{(1-a)(\alpha_0+\alpha_1) + k - 1}{(2-a)(\alpha_0+\alpha_1) + k - 1}$$

while the second is equal to

$$p''(a,m) = (\frac{\alpha_0}{\alpha_0+\alpha_1})(1 - a\frac{a(\alpha_0+\alpha_1)+1}{(\alpha_0+\alpha_1)+m}) + (a\frac{a\alpha_0+1}{(\alpha_0+\alpha_1)+m})$$

The following inequalities can be verified with simple algebra. If a^* and a^{**} belong to the interval $(0,1)$ and $a^{**} > a^*$, then $\sum_{m=j}^{\infty} p'(a^{**},m) \leq \sum_{m=j}^{\infty} p'(a^*,m)$ $\forall j \in \mathbb{N}$, $p''(a^{**},m) > p''(a^*,m)$ $\forall m \in \mathbb{N}$, and $p''(a,m) > p''(a,m+1)$ $\forall m \in \mathbb{N}$, $\forall a \in (0,1)$. From the three inequalities and the equality $\sum_{m=1}^{\infty} p'(a^{**},m) = \sum_{m=1}^{\infty} p'(a^*,m)$ it follows that

$$\sum_{m=1}^{\infty} p'(a^{**},m)p''(a^{**},m) > \sum_{m=1}^{\infty} p'(a^*,m)p''(a^*,m) \quad \cdot$$

The last inequality completes the proof, indeed the two terms are equivalent to the the expected value in (1.12) under two different parameterizations (i.e. a^* and a^{**}) of the random vector G. □

Proof of Proposition 1.1

Define for every i

$G^0_{x,x_i} = G((z,y) \in X \times \mathbb{R} : 0 < y < \alpha_0 \min(q_{x_i}(z), q_x(z)))$

$G^0_{x\setminus x_i} = G((z,y) \in X \times \mathbb{R} : 0 < y < \alpha_0 q_x(z), y \geq \alpha_0 q_{x_i}(z))$

$G^0_{x_i\setminus x} = G((z,y) \in X \times \mathbb{R} : 0 < y < \alpha_0 q_{x_i}(z), y \geq \alpha_0 q_x(z))\cdot$

We observe that $G(A^0_x) = G^0_{x,x_i} + G^0_{x\setminus x_i}$, $\quad G(A^0_{x_i}) = G^0_{x,x_i} + G^0_{x_i\setminus x}$ and

$E((A^0_x - A^0_{x_i})^2) = E((G^0_{x\setminus x_i} - G^0_{x_i\setminus x})^2) = Var(G^0_{x\setminus x_i}) + Var(G^0_{x_i\setminus x}) = 2Var(G^0_{x_i\setminus x}).$

From the hypothesis: $\nu((z,y) \in X \times \mathbb{R} : 0 < y < q_{x_i}(z), q_x(z) > y) \to 0\cdot$

It follows that $Var(G^0_{x_i\setminus x}) \to 0\cdot$

We can similarly define $G^1_{x,x_i}, G^1_{x\setminus x_i}, G^1_{x_i\setminus x}$ and show $E((A^1_x - A^1_{x_i})^2) \to 0$.

Finally from the convergence in mean square of $\{G(A^0_{x_i})\}_{i\geq 1}$ and $\{G(A^1_{x_i})\}_{i\geq 1}$ it follows the convergence in probability of $\{Y_{x_i}\}_{i\geq 1}$ to Y_x.

Proof of proposition 1.2

$E\left(\mathcal{P}_x(F_x^{-1}(B_\varepsilon))\right) = \lambda(B_\varepsilon) \quad \forall \varepsilon \in \cup_{i=1}^{\infty}\{0,1\}^i$, where λ denotes the Lebesgue measure. The same equality holds for the elements of the algebra of finite unions

of B_ε sets. The class $\{B \subseteq (0,1] : E\left(\mathcal{P}_x(F_x^{-1}(B))\right) = \lambda(B)\}$ is monotone. It follows from the monotone class theorem that for every Borell subset B of $(0,1]$ $E\left(\mathcal{P}_x(F_x^{-1}(B))\right) = \lambda(B)$. Finally if a distribution Q satisfy the equality $Q\left(F_x^{-1}(B)\right) = \lambda(B)$ for every Borell subset B of $(0,1]$ then $Q = F_x$.

Proof of Proposition 1.3

From Proposition 1 follows that: $\mathcal{P}_{x_i}(F_{x_i}^{-1}(B_\varepsilon)) \xrightarrow{p} \mathcal{P}_x(F_x^{-1}(B_\varepsilon)) \quad \forall \varepsilon \in \cup_{i=1}^{\infty}\{0,1\}^i$. This holds also for the elements of the algebra of the finite union of B_ε sets.

Let $B_1, B_2, \ldots$ be an increasing sequence with limit $B \subset [0,1]$ such that:
$\mathcal{P}_{x_i}(F_{x_i}^{-1}(B_j)) \xrightarrow{p} \mathcal{P}_x(F_x^{-1}(B_j)) \quad \forall j \in (1,2,\ldots)$.

$\mathcal{P}_{x_i}(F_{x_i}^{-1}(B)) - \mathcal{P}_{x_i}(F_{x_i}^{-1}(B_j)) \geq 0 \quad a.s. \quad \forall i,j$ and,
$E(\mathcal{P}_{x_i}(F_{x_i}^{-1}(B)) - \mathcal{P}_{x_i}(F_{x_i}^{-1}(B_j))) = \lambda(B \setminus B_j)$.

It follows from the Markov inequality that: $\forall \delta > 0, \forall \varepsilon > 0\ \exists m$ such that
$P(|\mathcal{P}_{x_i}(F_{x_i}^{-1}(B_m)) - \mathcal{P}_{x_i}(F_{x_i}^{-1}(B))| > \frac{\varepsilon}{3}) < \frac{\delta}{3} \quad \forall i$ and
$P(|\mathcal{P}_x(F_x^{-1}(B_m)) - \mathcal{P}_x(F_x^{-1}(B))| > \frac{\varepsilon}{3}) < \frac{\delta}{3}$.

Moreover there exists an integer k such that
$P(|\mathcal{P}_{x_j}(F_{x_j}^{-1}(B_m)) - \mathcal{P}_x(F_x^{-1}(B_m))| > \frac{\varepsilon}{3}) < \frac{\delta}{3} \quad \forall j > k$,
thus $P(|\mathcal{P}_{x_j}(F_{x_j}^{-1}(B)) - \mathcal{P}_x(F_x^{-1}(B))| > \varepsilon) < \delta \quad \forall j > k$.

Similarly if $B_1, B_2, \ldots$ decrease to B and
$\mathcal{P}_{x_i}(F_{x_i}^{-1}(B_j)) \xrightarrow{p} \mathcal{P}_x(F_x^{-1}(B_j)) \quad \forall j \in (1,2,\ldots)$,
then $\mathcal{P}_{x_i}(F_{x_i}^{-1}(B)) \xrightarrow{p} \mathcal{P}_x(F_x^{-1}(B))$.

From the monotone class theorem it follows that for every Borell subset $B \subset [0,1]$,

$\mathcal{P}_{x_i}(F_{x_i}^{-1}(B)) \xrightarrow{p} \mathcal{P}_x(F_x^{-1}(B))$.

Let be $a \in \mathbb{R}$ and $c = F_x(a)$, F_{x_i} and F_x are continuous thus

$$|\mathcal{P}_{x_i}((-\infty, F_{x_i}^{-1}(c)]) - \mathcal{P}_x((-\infty, a])| \stackrel{a.s.}{=} |\mathcal{P}_{x_i}(F_{x_i}^{-1}(0, c]) - \mathcal{P}_x(F_x^{-1}(0, c])| \xrightarrow{p} 0$$

$$E(|\mathcal{P}_{x_i}((-\infty, F_{x_i}^{-1}(c)]) - \mathcal{P}_{x_i}((-\infty, a])|) = |c - F_{x_i}(a)| \longrightarrow 0$$

It follows that: $|\mathcal{P}_{x_i}((-\infty, F_{x_i}^{-1}(c)]) - \mathcal{P}_{x_i}((-\infty, a])| \xrightarrow{p} 0$ and

$$|\mathcal{P}_{x_i}((-\infty, a]) - \mathcal{P}_x((-\infty, a])| \xrightarrow{p} 0 \quad .$$

It follows, through monotone class arguments based on the convergence in total variation of the sequence $\{F_{x_i}\}_{i \geq 1}$ to F_x, that for every Borel subset of the real line B,
$\mathcal{P}_{x_i}(B) \xrightarrow{p} \mathcal{P}_x(B)$.

Proof of Proposition 1.4

It is first proved a preliminary result (A) and then the proposition (B).

A)If $\{Y_x\}_{x \in X} \sim$ *MBP*(α_0, α_1, Q), *for every strictly positive* ε, *for every* $m = 1, 2, \ldots$, *for every distinct* $x_1, \ldots, x_m \in X$ *and* $a_1, \ldots, a_m \in [0, 1]$ *the event*

$$\{|Y_{x_i} - a_i| < \varepsilon \quad \forall i \in 1, \ldots m\}$$

has strictly positive probability.

There exist strictly positive numbers (g_1^0, g_1^1) such that $|\frac{g_1^0}{g_1^0 + g_1^1} - a_1| < \frac{\varepsilon}{2}$.
Given $G(A_{x_1}^0) = g_1^0$ and $G(A_{x_1}^0) = g_1^0$ there exist strictly positive numbers (g_2^0, g_2^1) such that if $G(A_{x_2}^0 \setminus A_{x_1}^0) = g_2^0$ and $G(A_{x_2}^1 \setminus A_{x_1}^1) = g_2^1$ then $|\frac{G(A_{x_2}^1)}{G(A_{x_2}^0) + G(A_{x_2}^1)} - a_2| < \frac{\varepsilon}{2}$.

For induction: given the equalities

$G(A^0_{x_1}) = g^0_1, \quad G(A^1_{x_1}) = g^1_1,$
$G(A^0_{x_2} \setminus A^0_{x_1}) = g^0_2, \quad G(A^1_{x_2} \setminus A^1_{x_1}) = g^1_2,$
$G(A^0_{x_3} \setminus (A^0_{x_1} \cup A^0_{x_2})) = g^0_3, \quad G(A^1_{x_3} \setminus (A^1_{x_1} \cup A^1_{x_2})) = g^1_3,$
$\ldots,$
$G(A^0_{x_i} \setminus (A^0_{x_1} \cup A^0_{x_2} \cup \ldots \cup A^0_{x_{i-1}})) = g^0_i, G(A^1_{x_i} \setminus (A^1_{x_1} \cup A^1_{x_2} \cup \ldots \cup A^1_{x_{i-1}})) = g^1_i,$

exist positive numbers (g^0_{i+1}, g^1_{i+1}) such that if the above equalities hold,
$G(A^0_{x_{i+1}} \setminus (A^0_{x_1} \cup \ldots \cup A^0_{x_i})) = g^0_{i+1}$ and $G(A^1_{x_{i+1}} \setminus (A^1_{x_1} \cup \ldots \cup A^1_{x_i})) = g^1_{i+1}$
then $|\frac{G(A^1_{x_{i+1}})}{G(A^0_{x_{i+1}}) + G(A^1_{x_{i+1}})} - a_{i+1}| < \frac{\varepsilon}{2}$.

The quantities $\nu(A^0_{x_1}), \nu(A^1_{x_1}), \nu(A^0_{x_2} \setminus A^0_{x_1}), \ldots, \nu(A^1_{x_m} \setminus (A^1_{x_1} \cup \ldots \cup A^1_{x_{m-1}}))$ are strictly positive due to the fact that the kernels $\{q_x\}_{x \in X}$ constitute a location family of unimodal densities. Hence the density function of

$$[G(A^0_{x_1}), G(A^1_{x_1}), G(A^0_{x_2} \setminus A^0_{x_1}), \ldots, G(A^1_{x_m} \setminus (A^1_{x_1} \cup \ldots \cup A^1_{x_{m-1}}))]$$

in an adequate neighborhood of $(g^0_1, g^1_1, \ldots, g^1_m)$ is strictly positive, this proves the assertion.

B) From the definition of the MPT and the above reported result it follows that for every positive ε^* and every integer j the event

$$\{|\mathcal{S}_i(F^{-1}_{x_i}(\frac{l}{2^j}, \frac{l+1}{2^j}]) - \mathcal{P}_{x_i}(F^{-1}_{x_i}(\frac{l}{2^j}, \frac{l+1}{2^j}])| < \varepsilon^*, \forall l \leq (2^j - 1), \forall i \in (1, \ldots, m)\} \tag{1.13}$$

has strictly positive probability. Finally for any given positive δ and partition $[V_1, \ldots, V_k]$ constituted of intervals, choosing adequately j and ε^* the event (1.13) becomes a subset of the event of interest,

$$\{|\mathcal{S}_i(V_l) - \mathcal{P}_{x_i}(V_l))| < \delta, \forall l \in (1, \ldots, k), \forall i \in (1, \ldots, m)\} \quad .$$

Proof of Theorem 1.1

If $0 \leq a < b \leq 1 \ \exists \quad \delta(a,b) > 0$ s.t. $P(Y_{x^\star{}_{1,\oslash}} \in (a,b) | Y_{x^\star{}_{2,\oslash}}, \ldots, Y_{x^\star{}_{m,\oslash}}) > \delta(a,b)$, where $\delta(a,b)$ does not depend from $(Y_{x^\star{}_{2,\oslash}}, \ldots, Y_{x^\star{}_{m,\oslash}})$. This fact is a direct consequence of the representation (1.2) and of the tail free property of the Dirichlet process: the random vectors

$$[\frac{D_{x^\star{}_1,\ldots,x^\star{}_m}(A^0_{x^\star{}_2})}{D_{x^\star{}_1,\ldots,x^\star{}_m}(A^0_{x^\star{}_2} \cup A^1_{x^\star{}_2})}, \ldots, \frac{D_{x^\star{}_1,\ldots,x^\star{}_m}(A^0_{x^\star{}_m})}{D_{x^\star{}_1,\ldots,x^\star{}_m}(A^0_{x^\star{}_m} \cup A^1_{x^\star{}_m})}]$$

and

$$[D_{x^\star{}_1,\ldots,x^\star{}_m}(A^0_{x^\star{}_1} \backslash (\cup_{i=2}^m A^0_{x^\star{}_i})), D_{x^\star{}_1,\ldots,x^\star{}_m}(A^1_{x^\star{}_1} \backslash (\cup_{i=2}^m A^1_{x^\star{}_i}))]$$

are independent.

From the inequality it follows that, for every strictly positive ξ the conditional probabilities

$$P\Big(Y_{x^\star{}_{1,\oslash}} \in (1-\mathcal{S}_{x^\star{}_1}(F^{-1}_{x^\star{}_1}\left(\frac{1}{2}\right))-\xi, 1-\mathcal{S}_{x^\star{}_1}(F^{-1}_{x^\star{}_1}\left(\frac{1}{2}\right))+\xi)\Big| \{W_i\}_{i=1}^n, Y_{x^\star{}_{2,\oslash}}, \ldots, Y_{x^\star{}_{m,\oslash}}\Big) =$$

$$P\Big(Y_{x^\star{}_{1,\oslash}} \in (1-\mathcal{S}_{x^\star{}_1}(F^{-1}_{x^\star{}_1}\left(\frac{1}{2}\right))-\xi, 1-\mathcal{S}_{x^\star{}_1}(F^{-1}_{x^\star{}_1}\left(\frac{1}{2}\right))+\xi)\Big| \Big\{I(W_i > F^{-1}_{x^\star{}_1}\Big(\frac{1}{2}\Big))\Big\}_{\{i\leq n: x_i = x^\star{}_1\}}, Y_{x^\star{}_{2,\oslash}}, \ldots, Y_{x^\star{}_{m,\oslash}}\Big)$$

almost surely (w.r.t.$\{W_i\}_{i=1}^n$) converge uniformly to 1: this is a direct consequence of the law of large numbers and of the fact that the second derivative of the loglikelihood of the *r.v.*'s $\{I(W_i > F^{-1}_{x^\star}(\frac{1}{2})\}_{\{i\leq n: x_i = x^\star{}_1\}}$ converges to $(-\infty)$.

From the equality

$$P(Y_{x^\star{}_{1,\oslash}} \in \cdot | \{W_i\}_{i=1}^n) =$$

$$\int P(Y_{x^\star{}_{1,\oslash}} \in \cdot | \{W_i\}_{i=1}^n, Y_{x^\star{}_{2,\oslash}}, \ldots, Y_{x^\star{}_{m,\oslash}}) dP_{Y_{x^\star{}_{2,\oslash}},\ldots,Y_{x^\star{}_{m,\oslash}} | \{W_i\}_{i=1}^n}$$

it follows that

$$P\Big(Y_{x^\star{}_{1,\oslash}} \in (1-\mathcal{S}_{x^\star{}_1}(F^{-1}_{x^\star{}_1}\Big(\frac{1}{2}\Big))-\xi, 1-\mathcal{S}_{x^\star{}_1}(F^{-1}_{x^\star{}_1}\Big(\frac{1}{2}\Big))+\xi)\Big| \{W_i\}_{i=1}^n\Big) \to 1 \qquad a.s.$$

Similarly if $\mathcal{S}_{x^\star{}_1}(F^{-1}_{x_1^\star}(\frac{j}{2^k}), F^{-1}_{x_1^\star-1}(\frac{j+1}{2^k})] > 0$ and $B_\varepsilon = (\frac{j}{2^k}, \frac{j+1}{2^k}]$ the posterior distribution of $Y_{x^\star{}_1,\varepsilon}$ is consistent.

From the monotonicity of $F_{x^\star{}_1}$ it follows that given a continuity point v of $\mathcal{S}_{x^\star{}_1}$ for every $\xi > 0$

$$P(\mathcal{P}_{x^\star{}_1}(-\infty, v) \in (\mathcal{S}_1(-\infty, v)-\xi, \mathcal{S}_1(-\infty, v)+\xi)|\{W_i\}_{i=1}^n) \to 1 \qquad a.s.$$

This fact completes the proof.

Bibliography

Berger,J.O.,Guglielmi, A.(2001) Bayesian testing of a parametric model versus nonparametric alternatives *Journal of the American Statistical Association*, Vol. 96, 174-184

Blackwell,D.,MacQueen J.B.(1973) Ferguson distributions via Polya urn schemes *The Annals of Statistics*, Vol. 1, No. 2, 353-355

Coram,M.,Lalley,S.P.(2006) Consistency of Bayes estimators of a binary regression function *The Annals of Statistics*Vol. 34, No. 3,1233-1269

Dunson,D.B.,Park,B.K.(2008) Kernel stick-breaking processes *Biometrika*, Vol. 95, 307-323

Hanson,T.,Johnson, W.(2002) Modeling regression error with a mixture of Polya trees *Journal of the American Statistical Association*, Vol. 97, 1020-1033

Hanson(2006) Inference for mixtures of finite Polya tree models *Journal of the American Statistical Association*, Vol. 101, 1548-1565

Hjort,N.,L.(1990) Nonparametric Bayes estimators based on Beta processes in models for life history data *The Annals of Statistics*Vol. 18, No. 3,1259-1294

Kalbfleisch,J.D. (1978) Non-Parametric bayesian analysis of survival time data *Journal of the Royal Statistical Society. Series B* , Vol. 40, No. 2, 214-221.

Kim,Y., Lee,J.(2003, a) Bayesian analysis of proportional hazard models *The Annals of Statistics*, Vol. 31, No. 2, 493-511

Kim,Y., Lee,J.(2003, b) Bayesian Bootstrap for proportional hazards models *The Annals of Statistics*, Vol. 31, No. 6, 10905-1922

De Iorio,M., Mller,P.,Rosner,G.L.,MacEachern,S.(2004) An ANOVA Model for Dependent Random Measures *Journal of the American Statistical Association*, Vol. 99, 205-215

Lad T.,Rubinstein L.,Sadeghi,A.(1988) The benefit of adjuvant treatment for resected locally advanced non-small-cell lung cancer *Journal of Clinical Oncology*,6,9-17

Lavine,M.(1992) Some aspects of Polya tree distributions for statistical modelling *The Annals of Statistics*, Vol. 20, No. 3, 1222-1235

Lavine,M.(1994) More aspects of Polya tree distributions for statistical modelling *The Annals of Statistics*, Vol. 22, No. 3, 1161-1176

MacEachern,S.(1999), Dependent Nonparametric Processes *ASA Proceedings of the Section on Bayesian Statistical Science* Alexandria, VA: American Statistical Association.

Mantel,N.,Stablein,M.D.(1988) The crossing hazard function problem *The Statistician*, Vol. 37, No. 1, 59-64

Mauldin,R.D.,Sudderth,W.D.,Williams,S.C.(1992) Polya trees and random distributions *The Annals of Statistics*, Vol. 20, No. 3, pp. 1203-1221

Muliere,P.,Tardella,L. (1998) Approximating distributions of random functionals of Ferguson-Dirichlet priors *The Canadian Journal of Statistics*, 26, 283-297

Olkin,I.,Liu,R.(2003) A bivariate beta distribution *Statistics and Probability Letters*, Vol. 62, 407-412

Piantadosi,S.(1997) *Clinical trials: a methodologic perspective*, New York: Wiley

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

Walker,S.,Muliere,P.(2003) A bivariate Dirichlet process *Statistics and Probability Letters*, Vol. 64, 1-7

Chapter 2

Reinforced random tessellations for Bayesian nonparametric binary regression.

Abstract

A Bayesian nonparametric model for binary random variables is introduced. The characterization of the probability model is based on the Polya urn scheme and on a random tessellation model. These two probabilistic structures are combined in order to adapt, under the hypothesis of partial exchangeability, the reinforcement mechanism of the Polya urn. A Gibbs sampling algorithm for implementing predictive inference is illustrated and an application of the inferential procedure is discussed.

2.1 Introduction

Consider an heterogeneous population of subjects with covariates. A dichotomous random variable (r.v.) is associated to each subject with response probability depending on the covariates: $P(Y_i = 1|X_i) = f(X_i)$, i indexes the individual and X_i represents its profile. We propose a Bayesian nonparametric procedure

to estimate f.

The most widely used parametric models are defined by means of a parametric function l_θ, which maps the covariates space on the real line, and a cumulative density function (cdf) H; such models assume that $f \in \{H \circ l_\theta : \theta \in \Theta\}$: in the probit model, for example, l_θ is linear and H is the standard Gaussian distribution function. This framework is exploited in Newton et al. (1996) to define a semiparametric Bayesian binary regression model, therein l_θ is linear and a Normal prior distribution on the coefficients is combined with a Dirichlet-Ferguson prior (Ferguson (1973)) for the cdf H. Wood and Kohn (1998) proposed an alternative Bayesian semiparametric model, the cdf H is fixed and a flexible prior for the function l, which is assumed to be a sum of functions of single covariates, is defined.

More recently Choudhuria et al. (2007) studied a nonparametric prior distribution on the response probability functions which is not fully concentrated on the monotone functions nor on the additive ones: the typical assumptions of the semiparametric models are removed. Therein the link function H is the standard Normal cdf and l is modeled as a Gaussian process.

In this chapter a Multivariate-Beta prior, whose characterization is based on a random tessellation distribution, is proposed to model the dependency between the response variables. A tessellation is a mosaic of non overlapping adjacent polytopes which cover a subset of $\mathbb{R}^d$. The prior is defined through an instrumental Dirichlet process (DP) on a space of tessellations of the covariates space having cells associated with the values 0 or 1: the parameter of the process is a rescaled probability distribution of a random tessellation with black and white cells.

Random tessellation theory has received much attention during the last decade and find applications in diverse areas as geostatistics and stereology; prominent classes studied in these fields are the Poisson-Voronoi and the Poisson-hyperplane tessellations, for a comprehensive overview on this subject we refer to Okabe et al. (1992).

The regression model is structured as the response variables $\{Y_i\}_{i\geq 1}$ would be functions of a latent sequence of random tessellations drawn from an unknown distribution: if the cell of the i-th tessellation where is located X_i is associated with the value 1 then Y_i will be equal to 1 and 0 otherwise. The prior is characterized by an easily interpretable dependency structure. If the unknown tessellations probability law is expected to concentrate on few tessellations with few cells, intuitively, Y_1 is strongly predictive of the response variable Y_2, especially if X_1 is near to X_2.

The chapter is organized in 4 sections. In Section 2 the the Poisson-hyperplane tessellation is briefly described. We consider this random tessellation model to specify the parameter of the DP on the tessellations space. In Section 3 the binary regression model is stated, an MCMC algorithm to estimate the response probability function f is given and an application of the proposed inferential procedure is illustrated. Some final remarks are given in the last section.

2.2 Poisson-hyperplane tessellations

Random tessellations are stochastic partitions of multidimensional spaces into polytopes. Let $\mathcal{C}$ be the class of compact subsets of $\mathbb{R}^d$, $\mathcal{P}$ denotes the class of polytopes in $\mathbb{R}^d$ and $\mathfrak{B}$ the Borel σ-field of $\mathcal{P}$ with respect to the Hausdorff metric. Indicate the interior of the polytope p with $int(p)$. T is the class of the counting measures on $\mathcal{P}$ such that $\forall t \in T$:

- $t(p) \in \{0, 1\} \qquad \forall p \in \mathcal{P}$,
- $\forall p_1, p_2 \in \mathcal{P}$ if $t(p_1) = t(p_2) = 1$ then $int(p_1) \bigcap int(p_2) = \emptyset$,
- $\forall x \in \mathbb{R}^d \; \exists \; p \in \mathcal{P}$ such that $t(p) = 1$ and $x \in p$,
- $\forall c \in \mathcal{C}$ the set $\{p \in \mathcal{P} : t(p) = 1, p \cap c \neq \emptyset\}$ has finite cardinality.

$\mathcal{T}$ indicates the smallest σ-field that guarantees, for every set B in $\mathfrak{B}$, the measurability of the function $t \longrightarrow t(B)$. The measurable space $(T, \mathcal{T})$ endowed with a probability measure P defines a random tessellation.

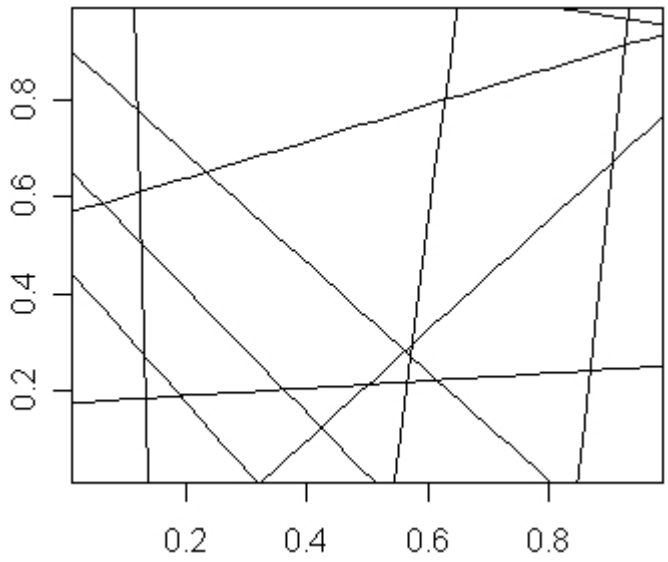


Figure 1: Bidimensional hyperplane tessellation.

In what follows $S^{d-1} = \{x \in \mathbb{R}^d : \|x\| = 1\}$ indicates the unit sphere and , for every $s \in S^{d-1}$ and $r \geq 0$, $H_{s,r}$ denotes the hyperplane $\{x \in \mathbb{R}^d : \langle x, s\rangle = r\}$. Q is an homogeneous Poisson process on the hyperplanes space $\mathbb{H}^d$ with intensity λ if for every measurable set $A \subset S^{d-1} \times [0, \infty)$, $Q(H_{s,r} \in \mathbb{H}^d : (s, r) \in A)$ is a Poisson r.v. with parameter $\lambda\gamma(A)$, where γ is the product between the Lebesgue measures on the real line and the uniform distribution on the unit sphere. If $A_1, \ldots, A_J$ are disjoint then $Q(H_{s,r} \in \mathbb{H}^d : (s, r) \in A_1), \ldots, Q(H_{s,r} \in \mathbb{H}^d : (s, r) \in A_J)$ are independently distributed.

The atoms of the Poisson process Q create a tessellation of $\mathbb{R}^d$: the class of nonempty intersections of the closed halfspeces bounded by the Q atoms constitutes a Poisson-hyperplane tessellation. A single cell can be represented as follows:

$$C = \left\{ \bigcap_{j=1}^{l} (x \in \mathbb{R}^d : \langle s_j, x\rangle \geq r_j) \right\} \bigcap \left\{ \bigcap_{j=l+1}^{m} (x \in \mathbb{R}^d : \langle s_j, x\rangle \leq r_j) \right\};$$

$Q(H_{r_j,s_j}) = 1 \; \forall j \in \{1, \ldots, m\}$ and $Q(H \in \mathbb{H}^d : H \cap int(C) \neq \emptyset) = 0$. The Poisson process Q, and consequently also the tessellation, is isotropic and stationary: the laws of both processes are invariant with respect to translations and rotations.

The outlined random tessellation, has received much attention in the literature, for an overview of the probabilistic properties of this model we refer to Stoyan et al. (1995). We will exploit the following proposition.

Proposition 2.1. *Given a Poisson-Hyperplane tessellation with intensity λ, the probability that a segment $\overline{ab}$ is entirely contained in one cell is*

$$P_r\left(Q(H \in \mathbb{H}^d : H \cap \overline{ab} \neq \emptyset) = 0\right) = \exp\left(-\lambda \frac{\|\overline{ab}\|}{2} \frac{\Gamma(\frac{d}{2})}{\Gamma(\frac{d+1}{2})\Gamma(\frac{1}{2})}\right). \tag{2.1}$$

Proof. If the random vector $(u_1, \dots u_d)$ is uniformly distributed on the unit sphere, then $(u_1^2, \dots u_{d-1}^2)$ is Dirichlet distributed with parameter $(\frac{1}{2}, \dots, \frac{1}{2})$, see for example (Eaton (1981)). It follows that

$$\gamma\left((s,r) \in S^{d-1} \times \mathbb{R}^+ : H_{s,r} \cap \overline{ab} \neq \emptyset\right) = \frac{\|\overline{ab}\|\Gamma(\frac{d}{2})}{2\pi^{d/2}} \int_{S^{d-1}} z_1^{1/2} \prod_{i=1}^{d} z_i^{-1/2} dz =$$

$$= \frac{\|\overline{ab}\|}{2} \frac{\Gamma(\frac{d}{2})}{\Gamma(\frac{d+1}{2})\Gamma(\frac{1}{2})}$$

□

2.3 Partial exchangeability via random tessellations

The distribution of a partially exchangeable sequence of random variables can be defined through an auxiliary latent exchangeable sequence of colored tessellations $\{t_i\}_{i\geq 1}$; the intuitive idea is that the i-th random response Y_i from an individual with covariates X_i is the color of the cell containing X_i of the i-th tessellation.

Let $(T \otimes \mathbb{R}^{\mathbb{N}}, \mathcal{T} \otimes \mathcal{B}(\mathbb{R}^{\mathbb{N}}))$ be the measurable space of colored tessellations: each tessellation t is constituted by the cells $C_1, C_2, \dots$ ordered with respect to an arbitrary criterion and the i-th real coordinate of the product space indicates the color of the i-th cell. In what follows C_X^t denotes the cell with the minimum index containing X and $\mathcal{Y}(C_X^t)$ the cell color. Consider a random probability measure $\mathcal{D}$ on $(T \otimes \mathbb{R}^{\mathbb{N}}, \mathcal{T} \otimes \mathcal{B}(\mathbb{R}^{\mathbb{N}}))$ having Dirichlet distribution with parameter $M(P_\lambda \times$

F^{∞}), where M is a positive constant, P_{λ} denotes the distribution of a Poisson-Hyperplane tessellation and F is a distribution on the possible cells colors, if the responses are dichotomous F concentrates on two colors. The elements of the partially exchangeable sequence $\{Y_i\}_{i\geq 1}$ are, conditionally on $\mathcal{D}$, independently distributed with

$$P_r(Y_i \in B|\mathcal{D}, X_i) = \mathcal{D}(t \in T \otimes \mathbb{R}^{\mathbb{N}} : \mathcal{Y}(C^t_{X_i}) \in B) \qquad \forall B \in \mathcal{B}(\mathbb{R}).$$

The definition of the model implies that the random probability measure $\mathcal{D}_X$ associated to a single covariate point X is Dirichlet distributed with weight parameter M and base measure F. The equality (2.1) allows to evaluate the dependency between the random probability measures $\mathcal{D}_{X_1}$ and $\mathcal{D}_{X_2}$ for any couple of covariates X_1 and X_2. For every measurable subset B of the support of F:

$$\begin{aligned} &E(\mathcal{D}_{X_1}(B)\mathcal{D}_{X_2}(B)) = P_r(Y_1 \in B, Y_2 \in B|X_1, X_2) = \\ &= F^2(B) + (F(B) - F^2(B))\frac{1}{M+1}\exp\left(-\lambda\|X_1 - X_2\|\frac{\Gamma(\frac{d}{2})}{2\Gamma(\frac{d+1}{2})\Gamma(\frac{1}{2})}\right) \end{aligned}$$

and the correlation between $\mathcal{D}_{X_1}(B)$ and $\mathcal{D}_{X_2}(B)$ is

$$Corr(\mathcal{D}_{X_1}(B), \mathcal{D}_{X_2}(B)) = \exp\left(-\lambda\|X_1 - X_2\|\frac{\Gamma(\frac{d}{2})}{2\Gamma(\frac{d+1}{2})\Gamma(\frac{1}{2})}\right).$$

If the response variables are dichotomous the parameters of the defined model can be clearly interpreted. The random probability $f(x)$ associated to a point x in the covariates space is a priori Beta distributed with parameters $(MF(1), MF(0))$, while the correlation between $f(x_1)$ and $f(x_2)$ depends on the intensity parameter λ and on the distance between x_1 and x_2.

Remark 1. The correlation function (2.2) depends exclusively on the random tessellation distribution P_{λ} that parameterizes the DP $\mathcal{D}$. It can be easily verified that this peculiarity is preserved if the Poisson-Hyperplane tessellation model is substituted, in the construction of the prior, with alternative random tessellation models. Consider, for example, the dead leaves tessellation model (Matheron (1968)), with spherical leaves having random radius: $P_r(R > r) = \exp(-\frac{r^2}{2/a^2})$. An elementary application of the results in (Bordenave et al. (2006)), allows to

obtain the correlation function of this slightly modified version of the proposed model: for every $B \in \mathcal{B}(\mathbb{R})$

$$Corr(\mathcal{D}_{X_1}(B), \mathcal{D}_{X_2}(B)) = \frac{\Phi(-a\|X_1 - X_2\|)}{\Phi(a\|X_1 - X_2\|)}, \tag{2.2}$$

where Φ denotes the standard Gaussian cumulative distribution function.

Remark 2. An alternative characterization of dependent DPs, whose dependency relationships allow a representation similar to the expression (2.2), is discussed in Walker and Muliere (2003) and Muliere et. al. (2005). A relevant difference between such characterization and the introduced prior consists in the fact that, with the proposed model, the quantities $Corr(\mathcal{D}_{X_1}(B), \mathcal{D}_{X_2}(B))$ gradually decrease as the distances $\|X_1 - X_2\|$ become larger.

2.3.1 Predictive inference.

The predictive probabilities,

$$\hat{f}(x) = P_r(Y_{n+1} = 1 | Y_1, \ldots, Y_n, X_1, \ldots, X_n, X_{n+1} = x)$$

can be computed through a Gibbs sampling algorithm based on the Blackwell MacQueen representation of an exchangeable sequence of variables with Dirichlet random distribution (Blackwell and MacQueen (1973)), in our case the elements of the sequence will be the latent tessellations. The algorithm is initialized by a finite sequence of colored tessellations, one for every observation. In each iteration the tessellation corresponding to a single observation is sampled conditionally on the other tessellations and on the data. The Blackwell MacQueen scheme suggests how to sample from the conditional distribution of the colored tessellation t_j given $\{t_1, \ldots, t_{j-1}, t_{j+1}, \ldots, t_n\}$ and $\{Y_1, \ldots, Y_n\}$. For every $l \in \{1, \ldots, j-1, j+1, \ldots, n\}$

$$P_r(t_j = t_l | t_1, \ldots, t_{j-1}, t_{j+1}, \ldots, t_n, \{Y_i\}_{i=1}^n) = \frac{I(\mathcal{Y}(C_{X_j}^{t_l}) = Y_j)}{MF(Y_j) + \sum_{h \neq j} I(\mathcal{Y}(C_{X_j}^{t_h}) = Y_j)}$$

and to sample from the conditional distribution

$$t_j | t_1, \ldots, t_{j-1}, t_{j+1}, \ldots, t_n, \{Y_i\}_{i=1}^n, t_j \neq t_1, \ldots t_j \neq t_n$$

it suffices to generate a P_λ distributed tessellation with $\mathcal{Y}(C^t_{X_j}) = Y_j$ and the remaining cells randomly colored accordingly with F.

The Gibbs sampling algorithm, exploiting the fact that the Blackwell MacQueen scheme indicates the conditional probabilities

$$P(Y_{n+1} = 1|t_1, \ldots, t_n, Y_1, \ldots, Y_n) = \frac{MF(1) + \sum_{i=1}^{n} \mathcal{Y}(C^{t_i}_{X_{n+1}})}{M + n}$$

allows to compute $\hat{f}$ through Monte Carlo iterations. Note that the algorithm can be adopted to estimate the response probability function also if the partially exchangeable probability model is parameterized through a random tessellation distribution alternative to the Poisson Hyperplane model. It suffices that the adopted tessellation can be exactly simulated. The dead leaves model and the Poisson-Voronoi tessellation are two examples of tessellation processes which can be easily generated.

2.3.2 Simulation example

The outlined algorithm has been applied to the simulated data set represented in *Figure 2.b*: 100 of points have been randomly selected in $[0, 1]^2$ and for each one a Bernoulli r.v. has been generated accordingly with the probabilities of success illustrated in *Figure 2.a*. The surface in the graph is the rescaled density of a mixture of two truncated bivariate Gaussian distributions. The prior distribution has been parameterized with $M = 2$, $F(0) = 1/2$ and $\lambda = 10$, from which follows that the response probabilities are a priori marginally uniformly distributed and their distribution is characterized by the correlation function $Corr(f(x_1), f(x_2)) = \exp(-\frac{10}{\Pi}\|x_1 - x_2\|)$. In *Figure 2.c* is represented an approximation of the predictive function $\hat{f}(x)$ which has been obtained computing through the Gibbs sampling algorithm the values of $\hat{f}$ on a grid of 2500 points on the unit square. The Bayesian estimate $\hat{f}$ captures the bimodal shape of the response probability function adopted to generate the data and seems to well reflect the data set structure suggested by the scatter plot.

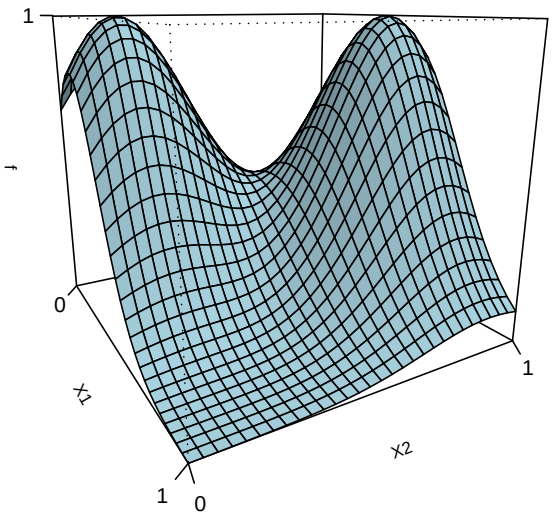


Figure 2.a: Response probability function.

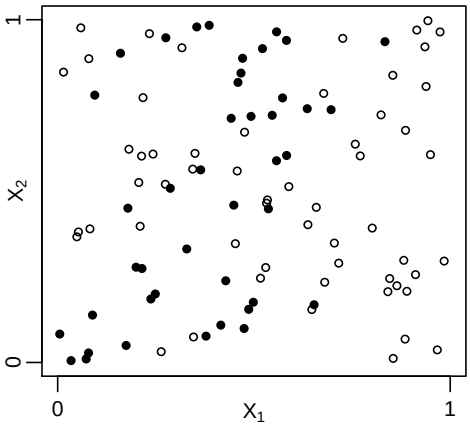


Figure 2.b: Simulated data set, $\circ = 0$ and $\bullet = 1$.

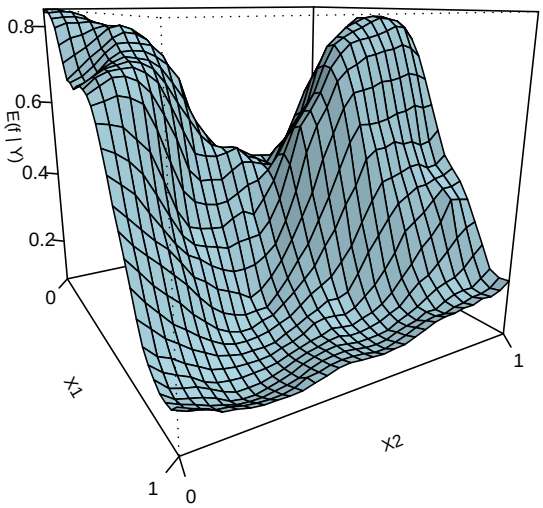


Figure 2.c: The predictive response probability function.

2.3.3 Example: Spatial variation in risk of disease

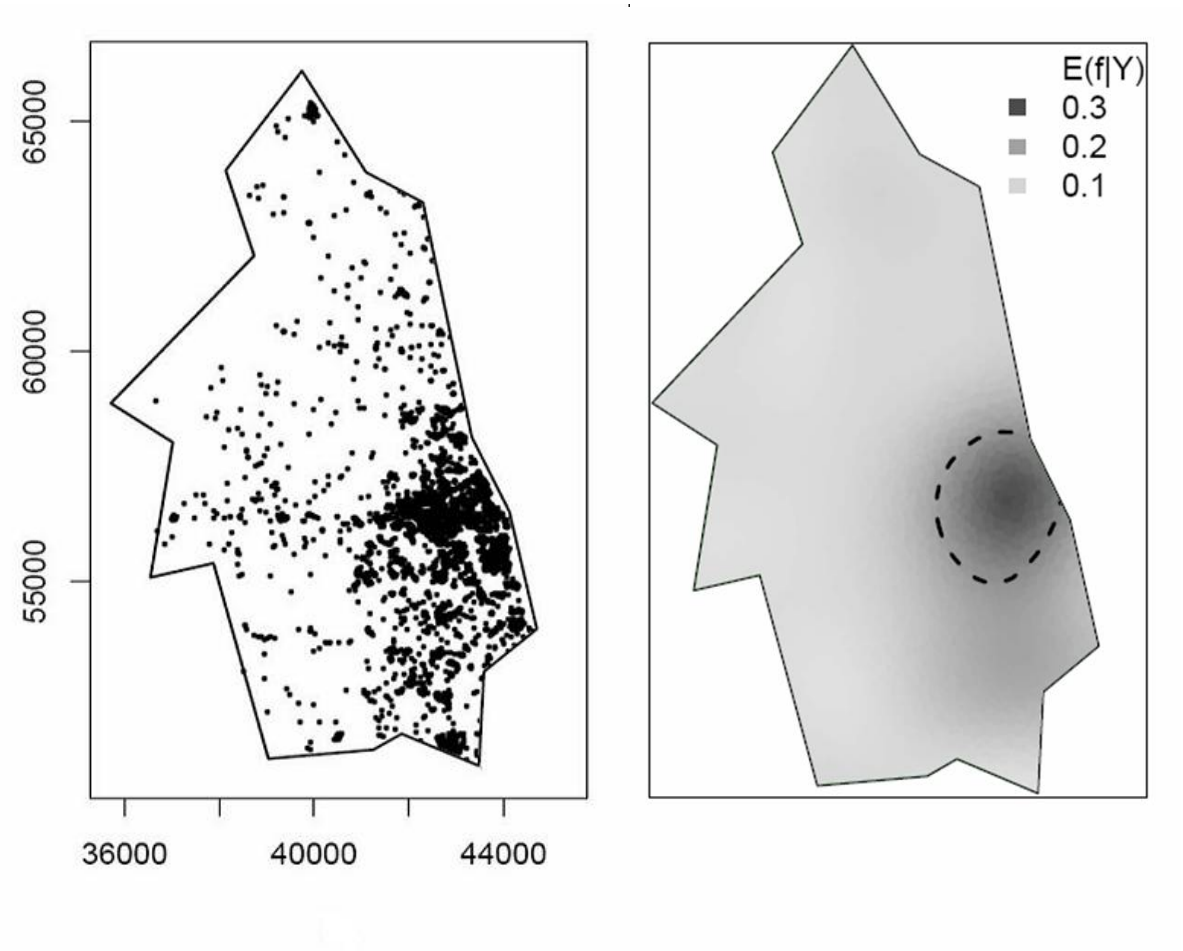
We describe the results of the analysis of an epidemiological data set through the proposed inferential procedure. The data set reports the postcodes of patients, in Northeast England, affected with primary biliary cirrhosis (PBC), and of a control group of residents of the region. For a detailed description of the data set we refer to Prince et al. (2001). One of the main findings of the cited study was that the risk of disease varies considerably across the region. Identifying the high-risk areas can substantially contribute to ascertain the environmental risk factors. Controls, as described in Prince et al. (2001), are randomly selected from the region population and concentrate mostly in the urban areas.

As discussed in Kelsall and Diggle (1998), the problem of identifying the variations of the risk of disease across a specific area can be formalized as a binary regression problem. Such approach considers both cases and controls as realizations of non homogeneous Poissont processes with intensities $\eta_1(x)$ (cases) and $\eta_0(x)$ (controls). We are interested in estimating the regression function

$$f(x) = \frac{\eta_1(x)}{\eta_1(x) + \eta_0(x)}.$$

A flat regression function corresponds to a constant risk across the area while possible variations could be determined by environmental factors which vary across the region. Figure 3.a represents the residence locations of the 761 individuals with PBC and of the 3044 controls. Figure 3.b represents the Bayesian estimate of the regression function f. The graph suggests a relevant difference in the incidence of disease between the urban areas of Newcastle and Gateshead and the surrounding areas.

The proposed inferential procedure identifies the area with the highest risk of disease and, the posterior confidence bands of f allows to assess the strength of evidence against the hypothesis of a constant regression function. Confidence bands can be computed approximately sampling, from the posterior distribution, the unknown response probability function f through the outlined Monte Carlo sampling scheme. Conditionally on the latent tessellations $\{t_1, \ldots, t_n\}$, the con-



(a) (b)

Figure 3: Panel (a): The polygonal approximation of the study area and the locations of cases of PBC and controls. Panel (b): Estimated response probability function.

jugacy property of the DP allows to sample with arbitrary accuracy the random process f. Muliere and Tardella (1998) discuss a computational procedure in order to generate with arbitrary accuracy a DP. The dashed line in Figure 3.b identifies the areas which, with posterior probability higher than 95%, present a lower risk of disease than the point x^* associated with the highest estimated response probability.

2.4 Final remarks.

In this chapter a class of dependent Beta random variables indexed by the points of a covariate space is characterized introducing a Dirichlet random distribution on a space of colored tessellations. The aim of the adopted modeling approach is to extend the appealing reinforcement mechanism proper of the Polya urn model to regression problems. The idea of constructing a partially exchangeable model merging the Polya urn scheme and the Poisson-Hyperplane tessellation can be generalized adopting alternative exchangeable random partition models, see for example Pitman (1995), as well as alternative tessellation models.

Bibliography

Blackwell, D. and MacQueen, J. B. (1973) Ferguson distributions via Pólya-urn schemes. *The Annals of Statistics*, Vol. 1, No. 2, 353–355.

Bordenave, C.,Gousseau, Y.,Roueff, F.(2006) The dead leaves model: a general tessellation modeling occlusion. *Adv. in Appl. Probab.*, 38, 31–46.

Choudhuria, N., Ghosalb, S. and Roy,A.(2007) Nonparametric binary regression using a Gaussian process prior. *Statistical Methodology* , 4, 227-243

Eaton,M.L. (1981) On the projections of isotropic distributions. *The Annals of Statistics*, 9, 391–400.

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

Matheron, G. (1968). Modle squentiel de partition alatoire. *Tech. Rep., Centre de Morphologie Mathmatique,* Fontainebleau

Muliere, P.,Secchi, P., Walker, S.(2005) Partially exchangeable processes indexed by the vertices of a k-tree constructed via reinforcement. *Stochastic processes and their applications*, 115, 661-677

Muliere, P.,Tardella, L. (1998) Approximating distributions of random functionals of Ferguson-Dirichlet priors. *The Canadian Journal of Statistics*, 26, 283-297

Newton, M.A., Czado, C. and Chappell, R. (1996) Bayesian inference for semiparametric binary regression. *Journal of the American Statistical Association*, 91, 142-153

Kelsall, J.E., Diggle, P.J.(1998) Spatial variation in risk of disease : a nonparametric binary regression approach. *Applied statistics*, 47, 559-573

Okabe, A., Boots,B., Sugihara,K., Nok Chiu,S. (1992) Spatial Tessellations: Concepts and Applications of Voronoi Diagrams. *John Wiley*

Pitman, J. (1995) Exchangeable and partially exchangeable random partitions. *Probab. Theory Related Fields* ,102, 145-158.

Prince,I.M. et al.(2001) The geographical distribution of primary biliary cirrhosis in a well-defined cohort. *Hepatology* , 34, 1083-1088

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

Walker, S., Muliere, P. (2003) A bivariate Dirichlet process . *Statistics & Probability Letters*, 64, 1-7

Wood, S., Kohn, R. (1998) A Bayesian approach to robust binary nonparametric regression. *Journal of the American Statistical Association*, 93, 203-213

Chapter 3

Extended Bernstein Prior via Reinforced Urn Processes

Abstract

A Reinforced Urn Process which induces a prior on the space of mixtures of Bernstein distributions is introduced. A nonparametric Bayesian model based on this prior is presented: the elicitation is treated and some connections with Dirichlet mixtures are given. In the last part of the chapter an MCMC algorithm to compute the predictive distribution is discussed.

3.1 Introduction

The Bernstein polynomial of degree k associated to a bounded function F on $[0,1]$ is defined as

$$B(x;k,F) = \sum_{j=0}^{k} F\left(\frac{j}{k}\right)\binom{k}{j} x^j (1-x)^{k-j}, \quad x \in [0,1]. \tag{3.1}$$

It is known that Bernstein polynomials well approximate F under general assumptions. If x is a continuity point of F: $\lim_{k\to\infty} B(x;k,F) = F(x)$. When F is a distribution function (d.f.) the Bernstein approximation is also a d.f. on $[0,1]$ and, if $F(0) = 0$, the polynomial $B(x;k,F)$ can be expressed as a mixture

of beta distributions. Under these hypotheses the derivative of the Bernstein polynomial is $B'(\cdot\,; k, F) = \sum_{j=1}^{k} w_{jk}\beta(\cdot, j, k-j+1)$, where $\beta(\cdot, a, b)$ denotes the beta density with parameters (a, b) and $w_{jk} = F\left(\frac{j}{k}\right) - F\left(\frac{j-1}{k}\right)$.

Petrone (1999) applied Bernstein polynomials in Bayesian inference: a non-parametric prior on the set of absolutely continuous distributions on the unit interval is expressed by specifying a probability measure on the set of d.f.'s which belong to the Bernstein polynomials' space. Therein is studied the random d.f. $B(x; K, F) = I_{(1,\infty)}(x) + \sum_{j=0}^{K} F\left(\frac{j}{K}\right)\binom{K}{j} x^j (1-x)^{K-j} I_{[0,1]}(x)$, where F is a Dirichlet process (Ferguson (1973)) on the unit interval parameterized with a finite measure α and K has an independent distribution p on the integers. Provided that $\alpha(\{0\}) = 0$, $F(0) = 0$ *a.s.*, and the random Bernstein polynomial is *a.s.* a mixture of beta distributions. In practice, the d.f. F_0 obtained by the normalization of the measure α expresses the initial guess of an unknown d.f., although $B(x)$ is centered on $\sum_k B(x; k, F_0)p(k)$, which is generally different from F_0. Computational procedures to estimate an unknown d.f. through the described Bayesian model have been proposed by Petrone (1999) and Petrone and Wasserman (2002); however, they assume that the random variable (r.v.) K has finite support, that is, the order of the random polynomial is a priori bounded.

We propose an extension of the Bernstein prior that can be easily centered on whichever continuous d.f. on a real interval. Posterior inference can be fairly simply approximated through simulation techniques; in particular, computing the predictive distribution does not require to truncate the order of the polynomial.

The first part of the chapter is devoted to defining a prior probability measure on the beta mixtures space

$$\mathbf{B} = \left\{ \sum_{\tilde{S}} w_{jk}\beta(j, k-j+1) : \sum_{\tilde{S}} w_{jk} = 1;\ w_{jk} \geq 0,\ \forall (j,k) \in \tilde{S} \right\}, \qquad (3.2)$$

where $\tilde{S} = \{(j,k) : j, k \in \mathbb{N}^+; j \leq k\}$. To this aim, we first construct a prior for the mixing weights $W = \{W_{j,k}\}_{(j,k)\in\tilde{S}}$; note that W is a random probability measure on $\tilde{S}$. Then, the prior on the beta mixtures space is obtained by mapping W to $\mathbf{B}$: $W \longrightarrow \sum_{\tilde{S}} W_{j,k}\beta(j, k-j+1)$. The peculiarity of our construction is

to define the random mixing distribution W by means of an auxiliary Reinforced Urn Process (RUP). Muliere et al. (2000) give a detailed account of RUP's probabilistic properties. We show that a random probability measure W on $\tilde{S}$, which generalizes the Dirichlet process, can be characterized through a RUP $\{X_n\}_{n\geq 0}$. Indeed, the RUP generates an exchangeable sequence of r.v.'s $\{Y_n\}_{n\geq 1}$ with values in $\tilde{S}$, whose de Finetti measure gives the required probability law of the random measure W.

The above construction defines a prior with large support on the space of absolutely continuous d.f.'s on $[0,1]$. In the second part of the chapter, we extend the Bernstein model to construct a prior for a random d.f. on a general real interval. The main features of the extension are (i) the possibility of easily centering the prior on any a priori guess F_0, (ii) the availability of simulation techniques for implementing posterior inference and (iii) the ability of the model of combining mixture components with remarkably heterogeneous variances. The latter property differentiates the proposed extension from the Bernstein model discussed in Petrone (1999); we illustrate that such peculiarity can determine relevant differences between the predictive distributions corresponding to the two models. Such differences have an intuitive explanation. The beta components of a Bernstein density B' act as kernels, with k having the role of a smoothing parameter. Intuitively, one can think of B' as a smoothing of a histogram with bins of equal length $1/k$, where the rectangular kernels of the histogram are replaced by the beta kernels $\beta(j, k-j+1)$. For a better local behavior, it is natural to think of histograms with bins of different length, smoothed through the extended Bernstein mixtures in $\mathbf{B}$, where the mixture is taken with respect to the *joint* distribution W of j and k.

The outline of the chapter is the following. In Section 3.2 some definitions and results about RUPs from Muliere et al. (2000) are reviewed. In Section 3.3, a class of RUPs $\{X_n\}_{n\geq 0}$ which generates a random probability measure W on $\tilde{S}$ is introduced. The extended Bernstein model is presented in Section 3.4. In Section 3.5, an MCMC algorithm for posterior inference and some applications

are illustrated. Final remarks and open issues are briefly discussed in the last section.

3.2 Reinforced Urn Processes

RUPs may be defined very briefly as reinforced random walks on a state space of urns. Indeed they blend two basic probabilistic models: a Pólya urn scheme and a random walk in such a way as to describe the attitude of a random mover to repeat transitions already occurred in the past. Each RUP consists of four basic elements.

Definition 3. *Let*

1. *S be a countable state space;*
2. *$C = \{c_1, \ldots, c_k\}$ be a finite set of colors of cardinality $k \geq 1$;*
3. *$U(s) = (a_s(c_1), \ldots, a_s(c_k))$ be an urn composition function which maps S into the set of k-tuples of nonnegative numbers whose sum is strictly positive;*
4. *$q : S \times C \to S$ be a law of motion such that for every $x, y \in S$, there is at most one color $c(x, y) \in C$ such that $q(x, c(x, y)) = y$.*

Fixed $X_0 = x_0 \in S$, for $n \geq 1$ if $X_{n-1} = x \in S$, a ball is sampled from the urn associated with x and, given its color c, we set

$$X_n = q(x, c).$$

Finally the ball is replaced in the urn along with one of the same color.

The sequence $X = \{X_n,\, n \geq 0\}$ is said to be a Reinforced Urn Process with initial state x_0 and parameters S, C, U, q.

Muliere et al. (2000) show that partial exchangeability is a momentous feature enjoyed by the RUPs. Therefore, if a RUP $\{X_n\}_{n\geq 0}$ is also recurrent, a

representation theorem of Diaconis and Freedman (1980) states that it is a mixture of Markov chains. More formally let $R^{(0)} = \{x_0\}$ and for all $x \in S$, let $R_x = \{y \in S : a_x(c(x,y)) > 0\}$ be the set of all the states attainable from the state x in one step. Then define, for $n \geq 1$, $R^{(n)} = \bigcup_{x \in R^{(n-1)}} R_x$ and $R = \bigcup_{n=0}^{\infty} R^{(n)}$, the set of states visited with strictly positive probability by the RUP. The above cited result assures the existence of a probability measure μ on the set $\mathcal{P}$ of $R \times R$ transition matrices such that for every $n \geq 1$ and $(x_1, \cdots, x_n) \in R^n$,

$$P[X_0 = x_0, X_1 = x_1 \cdots, X_n = x_n] = \int_{\mathcal{P}} \prod_{j=0}^{n-1} \pi(x_j, x_{j+1}) \mu(d\pi). \tag{3.3}$$

Consider the random matrix Π with distribution μ. Let $\Pi(x)$ be the x-th row of Π and $\alpha(x)$ the measure on R which assigns mass $a_x(c)$ to $q(x,c)$ for each $c \in C$ such that $a_x(c) > 0$ and null mass to all the other elements of R. The following theorem describes the measure μ making the point about the properties of the random matrix Π.

Theorem 3.1. [*Muliere et al. (2000)*] *If the RUP* $\{X_n\}_{n \geq 0}$ *is recurrent, the rows of* Π *are mutually independent random probability distributions on* R *and, for all* $x \in R$*, the law of* $\Pi(x)$ *is that of a Dirichlet process with parameter* $\alpha(x)$*.*

Following Diaconis and Freedman (1980), for a process $\{X_n\}_{n \geq 0}$ on S, a x_0-block is defined to be a finite sequence of states which begins with x_0 and contains no further x_0. Let S^* be the countable space of all finite sequences of elements of S. Under the recurrence hypothesis of a RUP $\{X_n\}_{n \geq 0}$, the sequence of the successive x_0-blocks, which retraces the trajectory of the process, say $B_1, B_2, \ldots$ with $B_n \in S^*$ for every $n \geq 1$, is well defined. A simple example helps to clarify that the trajectory of a recurrent process can be decomposed in x_0-blocks. Let $S = \{0, 1, 2\}$ and $x_0 = 0$; a specific trajectory, say $\{0, 2, 1, 0, 0, 1, 0, \ldots\}$, is constituted by x_0-blocks: $\{B_1 = [0, 2, 1],\ B_2 = [0], B_3 = [0, 1], \ldots\}$. The blocks sequence $\{B_n\}_{n \geq 1}$ is well defined because of the recurrence hypothesis: the event $\{\cup_{m=1}^{\infty} \cap_{n \geq m} (X_n \neq x_0)\}$ has null probability.

As illustrated in the following paragraph, the sequence $\{B_n\}_{n \geq 1}$ is exchangeable and, for every measurable function φ from S^* to another space S', the same

property characterizes the sequence $\{\varphi(B_n)\}_{n\geq 1}$. In the next section, this property is exploited to define a random probability measure on the countable space $\tilde{S} = \{(i,k) : i,k \in \mathbb{N}^+; i \leq k\}$.

A simple example shows that the exchangeability of the x_0-blocks follows from the representation (3.3). From the definition of x_0-block it follows that $\{B_1 = [0,2], B_2 = [0]\}$ and $\{X_1 = 0, X_2 = 2, X_3 = 0, X_4 = 0\}$ are two representations of the same event; as well, permuting B_1 and B_2, the events $\{B_1 = [0], B_2 = [0,2]\}$ and $\{X_1 = 0, X_2 = 0, X_3 = 2, X_4 = 0\}$ are equivalent. Moreover, the count of the transitions between states in the two cases is identical: there is 1 transition from 0 to 0, 1 from 0 to 2 and 1 transition from 2 to 0. Thus, $Pr(B_1 = [0,2], B_2 = [0]) = Pr(B_1 = [0], B_2 = [0,2])$, since both are obtained from (3.3) and the integrand functions are the same. The same arguments can be adopted to verify that all the permutations of a finite sequence of x_0-blocks $B_1, \ldots, B_n$ are equally probable.

3.3 Probability measures on the beta mixtures space

The class of RUPs that we consider to define probability measures on the set of mixtures of beta distributions is parameterized as follows:

1. the state space: $\tilde{S} = \{(i,k) : i,k \in \mathbb{N}^+, i \leq k\}$,
2. the initial state of the process: $\tilde{x}_0 = (1,1)$,
3. the color space: $\tilde{C} = \{\tilde{c}_1, \tilde{c}_2, \tilde{c}_3\}$,
4. the urn composition function $\tilde{U}(s) = (a_s(\tilde{c}_1), a_s(\tilde{c}_2), a_s(\tilde{c}_3)) \quad \forall s \in \tilde{S}$,
5. the law of motion:

$$\tilde{q}((i,k),c) = \begin{cases} (i,k+1) & c = \tilde{c}_1 \\ (i+1,k+1) & c = \tilde{c}_2 \\ (1,1) & c = \tilde{c}_3 \end{cases} \qquad \forall (i,k) \in \tilde{S}.$$

The tilde symbol is used to distinguish the above described class from the whole RUPs family.

As anticipated, if a RUP $\{\tilde{X}_n\}_{n\geq 0}$ is recurrent, a sequence of exchangeable r.v.'s $\{Y_n\}_{n\geq 1}$ can be easily defined by mapping the x_0-blocks $B_1, B_2, \ldots$ to a measurable space. We consider the map:

$$\varphi : [B_i = (\tilde{x}_0, \tilde{x}_1, \ldots, \tilde{x}_m)] \to [Y_i = \tilde{x}_m], \quad i \geq 1. \tag{3.4}$$

Let $(\Omega, \mathfrak{F}, \mathbf{P})$ be the probability space on which the process $\{\tilde{X}_n\}_{n\geq 0}$ and the random sequence $\{Y_n = \varphi(B_n)\}_{n\geq 1}$ are defined. By de Finetti's representation theorem, the limit of the empirical distributions of the latter sequence, $W = \{W_s = \lim \frac{1}{N} \sum_1^N I_{(s)}(Y_n)\}_{s\in\tilde{S}}$, is well defined on the same space; W is a random probability measure on $\tilde{S}$, with probability law given by the de Finetti measure of the exchangeable sequence $\{Y_n = \varphi(B_n)\}_{n\geq 1}$.

Let us now denote the class of probability measures on the unit interval with $\triangle$, the Borel σ-field generated by the topology of the weak convergence with $\mathcal{H}$ and its restriction to $\mathbf{B}$ with $\mathfrak{B}$. Given the RUP $\{\tilde{X}_n\}_{n\geq 0}$, and the associated random probability measure W, a probability measure on $(\mathbf{B}, \mathfrak{B})$ can be induced in a natural way via a measurable map from Ω to $\mathbf{B}$:

$$\omega \to \sum W_{i,k}(\omega)\beta(i, k-i+1)\cdot$$

We will still denote such probability measure with $\mathbf{P}$.

From the properties of Bernstein polynomials, it follows easily that the prior $\mathbf{P}$ has full weak support Δ under mild assumptions.

Proposition 3.1. *If the urn composition function of the RUP $\{\tilde{X}_n\}_{n\geq 0}$ is such that $Pr(\varphi(B_1) = (i,k)) > 0 \quad \forall\, (i,k) \in \tilde{S}$, then the probability measure $\mathbf{P}$ defined above has full weak support Δ.*

The proofs of the results in this section are provided in the Appendix.

The following results show that the urn composition function $\tilde{U}(s)$ can be chosen in such a way that the random probability measure $W = \{W_s\}_{s\in\tilde{S}}$ is a

Dirichlet process with parameter G, where G is an arbitrary finite measure on $\tilde{S}$. The following theorem holds for a general RUP.

Theorem 3.2. *Let $\{X_n\}_{n\geq 0}$ be a recurrent RUP with parameter (S, U, C, q) and initial state x_0. Let φ be the map from $\cup_{m\geq 1} S^m$ to S such that*

$$\varphi : (s_1, s_2, \ldots, s_m) \to s_m \qquad \forall\, (s_1, s_2, \ldots, s_m) \in S^m,\ \forall\, m \geq 1.$$

If

$$\sum_{c\in C} a_{s^*}(c) = \sum_{s\in S} a_s(c(s, s^*)) \qquad \forall\, s^* \in S, \tag{3.5}$$

then the sequence $\{Y_n = \varphi(B_n)\}_{n\geq 1}$ is exchangeable and its associated de Finetti measure is a Dirichlet process with parameter α, where α is the finite measure on S with $\alpha(s) = a_s(c(s, x_0))$, $\forall s \in S$.

An elementary example of RUP that satisfies the condition (3.5) is the following. Let $S = \{0, 1, 2\}$ and $x_0 = 0$. Consider the urn composition such that the urn associated to the state s contains $(2 - s)$ balls of color c_1 and one of color c_2. If $q(s, c_1) = s + 1$ and $q(s, c_2) = 0$, then the described RUP parametrization implies that the de Finetti measure associated to $\{Y_n = \varphi(B_n)\}_{n\geq 1}$ is a Dirichlet distribution with parameters $(1, 1, 1)$.

The next proposition states how the urn compositions of the RUP $\{\tilde{X}_n\}_{n\geq 0}$ can be initialized to get a Dirichlet process prior for the sequence of the last states of the x_0-blocks.

Proposition 3.2. *Let G be a finite measure on $\tilde{S}$. Let $\{\mu_h\}_{h\geq 1}$ be a sequence of measures on $(0, 1]$ with $\mu_h((0, 1] \setminus \mathbb{Q}) = 0$ and*

$$\mu_h(r) = G((i, k) \in \tilde{S} : k \geq h, \frac{i}{k} = r),\ \forall r \in \mathbb{Q} \quad .$$

Let $\{\tilde{X}_n\}_{n\geq 0}$ be the RUP with initial state $x_0 = (1, 1)$ and parameters $(\tilde{S}, \tilde{U}, \tilde{C}, \tilde{q})$ where $\tilde{U}(i, k) = (\mu_{k+1}(\frac{i-1}{k}, \frac{i}{k+1}], \mu_{k+1}(\frac{i}{k+1}, \frac{i}{k}], G(i, k))$. Then, the sequence $\{Y_n = \varphi(B_n)\}_{n\geq 1}$ is exchangeable and its de Finetti measure is a Dirichlet process with parameter G.

3.4 Extended Bernstein prior

In this section, a Bayesian nonparametric model whose prior distribution can be easily centered is specified through the probability measure $\mathbf{P}$ defined on the mixtures' space $\mathbf{B}$.

We consider an exchangeable sequence $\{Z_i\}_{i\geq 1}$ of continuous real valued r.v.'s; equivalently, $Z_i \mid F$ are i.i.d. according to F, where F is an absolutely continuous random d.f.. Let F_0 be an arbitrarily chosen continuous d.f. which represents the initial guess on the unknown d.f. F of Z_i. To assign a nonparametric prior on F, centered on F_0, we transform the Z_i's into $[0,1]$ by letting $\zeta_i = F_0(Z_i)$, and assume that

$$\zeta_i \mid B \overset{i.i.d.}{\sim} B, \qquad B \sim \mathbf{P}.$$

Therefore, the probability law of $\{Z_i\}_{i\geq 1}$ is uniquely defined by the sequence of d.f.'s

$$P(Z_1 \leq z_1, \ldots, Z_n \leq z_n) = \int_{\mathbf{B}} \prod_{j=1}^{n} B(F_0(z_i)) d\mathbf{P}(B), \qquad n \geq 1.$$

In particular

$$P(Z_i \leq z \mid B) = \int_0^z b(F_0(t))\, F_0'(t) dt, \tag{3.6}$$

where b denotes the beta mixture density of B. Thus, the prior models the unknown density as F_0' times a "distortion factor" $b(F_0(\cdot))$.

The law of the sequence $\{Z_i\}_{i\geq 1}$ can be alternatively represented conditionally on the RUP $\tilde{X} = \{\tilde{X}_i\}_{n\geq 0}$ used in the construction of $\mathbf{P}$. In this case:

$$P(Z_1 \leq z_1, \ldots, Z_n \leq z_n \mid \tilde{X}) = \prod_{i=1}^{n} P(Z_i \leq z_i \mid \varphi(B_i)) \tag{3.7}$$

$$P(Z_i \leq z \mid \varphi(B_i)) = \int_0^{F_0(z)} \beta(x, \varphi^{(1)}(B_i), \varphi^{(2)}(B_i) - \varphi^{(1)}(B_i) + 1) dx$$

where $\varphi^{(1)}$ and $\varphi^{(2)}$ are the two components of the φ function. In this perspective, every observation Z_i has an associated latent r.v. $\varphi(B_i)$.

The introduced prior is centered on F_0 if and only if $E(B(x)) = x$ for every x in $[0,1]$, so that $P(Z_1 \leq z) = F_0(z)$. The simplest strategy to achieve this

equality is constraining the parameter of the RUP $\tilde{X}$ as follows:

$$\frac{a_{i,k}(\tilde{c}_2)}{a_{i,k}(\tilde{c}_1) + a_{i,k}(\tilde{c}_2)} = \frac{i}{k+1} \qquad \forall (i,k) \in \tilde{S}. \tag{3.8}$$

If these conditions are satisfied, the first x_0-block evolves exactly as a Pólya urn with initially two balls of two different colors; more formally $\forall\,(i,k) \in \tilde{S}$:

$$P\left(\tilde{X}_k = (i+1, k+1) | \tilde{X}_{k-1} = (i,k), \tilde{X}_k \neq (1,1)\right) = \frac{i}{k+1}$$
$$P\left(\tilde{X}_k = (i, k+1) | \tilde{X}_{k-1} = (i,k), \tilde{X}_k \neq (1,1)\right) = \frac{k-i+1}{k+1}.$$

These are the transition probabilities of a Pólya urn containing i white balls and $k-i+1$ black. On the other hand, conditionally on $\{\tilde{X}_{k-1} = (i,k)\}$ and $\{\tilde{X}_k = (1,1)\}$, from (3.7) we have that $F_0(Z_1)$ has a beta distribution with parameters $(i, k-i+1)$, which can be viewed as the random limit proportion of white balls in a Pólya urn initially containing i white balls and $k-i+1$ black. In a unified perspective, the law of $F_0(Z_1)$ can be represented as the random limit proportion of white balls in a Pólya urn initially containing 1 white and 1 black, which is well known to be uniformly distributed.

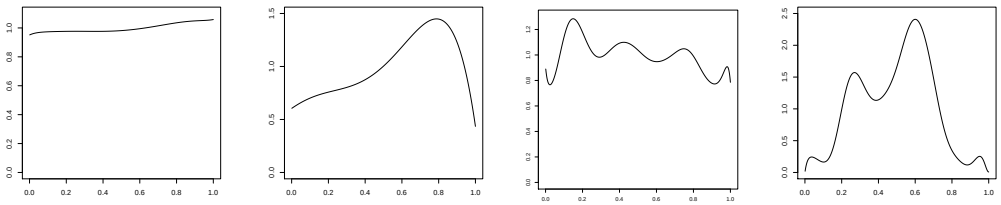
It can be easily shown, through these arguments, that for example if the RUP is parameterized as in Proposition 3.2, with

$$G(i,k) = M\frac{e^{-\lambda}\lambda^{k-1}}{k!} \quad \forall\,(i,k) \in \tilde{S}, \quad M > 0, \lambda > 0,$$

then the random probability measure $B\,(F_0(\cdot))$ is centered on F_0. More generally, if $P(\varphi(B_1) = (i,k))$ is a function of k, the equality $P(Z_1 \leq z) = F_0(z)$ is achieved. If the support of F_0 is a closed interval $[\gamma, \eta]$, it follows from proposition (3.1) that the support of the random distribution $B\,(F_0(\cdot))$ is the set of the probability measures on the interval. In this example, the expected values of the random weights $W_{i,k}$ are functions of λ: the higher is λ, the larger are the weights of the random mixture B which correspond to beta components characterized by small variances.

The constant λ can be interpreted as a smoothing parameter. For $\lambda \approx \infty$, given a finite partition $\{A_1, \ldots A_m\}$ of $[0,1]$, the random vector $[B(A_1), \ldots B(A_m)]$

will be approximately Dirichlet distributed, while if $\lambda \approx 0$ it will be approximately equal to the Lebesgue measure of the partition sets. For fixed values of λ, the parameter M is decisive to regulate the variances of the r.v.'s $W_{i,k}$: the lower it is, the higher is the uncertainty on the random probability measure $B(F_0(\cdot))$. Figure **??** represents the density functions of four sampled probability measures B; the comparison of the four graphs emphasizes the interpretation of the parameters M and λ.



$\lambda = 5, M = 100$ $\lambda = 5, M = 3$ $\lambda = 50, M = 100$ $\lambda = 50, M = 3$

Figure 3.1: Samples of density functions from differently parameterized prior distributions.

3.5 Inference

The proposed extension of the Bernstein prior allows to define a Gibbs sampling algorithm in order to approximate the predictive distribution of Z_{n+1} conditionally on $Z_1, Z_2, \ldots, Z_n$

$$P(Z_{n+1} \leq z | Z_1, \ldots, Z_n) = \sum_{\tilde{S}} E(W_{i,k} | Z_1, \ldots, Z_n) \int_0^{F_0(z)} \beta(x, i, k-i+1) dx. \tag{3.9}$$

The algorithm is based on the updating structure proper of the Pólya scheme. If the prior is suitably parameterized, the computational procedure does not require to truncate the number of components of the random mixture $B(F_0(\cdot))$. This is

an advantage with respect to the algorithms in literature for the Bernstein model, which bound the order of the random polynomial, see for example Petrone (1999).

An essential description of the algorithm is given as follows. Since, using (3.7), we have

$$E(W_{i,k} \mid Z_1, \ldots, Z_n) = \int E(W_{i,k} \mid B_1, \ldots, B_n) dP(B_1, \ldots, B_n \mid Z_1, \ldots, Z_n),$$

to compute the predictive distribution (3.9) it suffices to sample from the posterior distribution of $(B_1, \ldots, B_n \mid Z_1, \ldots, Z_n)$. To this aim, fix a sequence of x_0-blocks $B_1^1, \ldots B_n^1$ such that $P([B_1, \ldots, B_n] = [B_1^1, \ldots, B_n^1]) > 0$ and sample iteratively B_l^{j+1}, from the conditional distribution of B_l given $(Z_1, \ldots, Z_n)$ and $(B_1 = B_1^{j+1}, \ldots, B_{l-1} = B_{l-1}^{j+1}, B_{l+1} = B_{l+1}^j, \ldots, B_n = B_n^j)$. The iterations simulate a Markov chain $\{B_1^j, \ldots, B_n^j\}_{j \geq 1}$ whose stationary distribution is the conditional law of $(B_1, \ldots, B_n)$ given $(Z_1, \ldots, Z_n)$. So, for every couple (i, k) in $\tilde{S}$, $E(W_{i,k}|Z_1, Z_2, \ldots, Z_n)$ can be approximated through

$$\frac{1}{N} \sum_{j=1}^{N} E(W_{i,k}|B_1 = B_1^j, \ldots, B_n = B_n^j) \cdot$$

These conditional expectations can be computed by means of the following equalities:

$$E(W_{i,k}|B_1, B_2, \ldots, B_n) = \sum_{\varphi(B)=(i,k)} P(B_{n+1} = B|B_1, B_2, \ldots, B_n) \quad (3.10)$$

$$P(\tilde{X}_{m+1} = \tilde{x}_{m+1}, \ldots, \tilde{X}_{m+k} = \tilde{x}_{m+k}|\tilde{X}_0 = \tilde{x}_0, \ldots, \tilde{X}_m = \tilde{x}_m) =$$

$$= \prod_{j=0}^{k-1} \frac{a_{\tilde{x}_{m+j}}(c(\tilde{x}_{m+j}, \tilde{x}_{m+j+1})) + \sum_{i=0}^{m+j-1} I_{(\tilde{x}_i, \tilde{x}_{i+1})}(\tilde{x}_{m+j}, \tilde{x}_{m+j+1})}{a_{\tilde{x}_{m+j}}(\tilde{c}_1) + a_{\tilde{x}_{m+j}}(\tilde{c}_2) + a_{\tilde{x}_{m+j}}(\tilde{c}_3) + \sum_{i=0}^{m+j-1} I_{(\tilde{x}_i)}(\tilde{x}_{m+j})}$$

In two relevant cases it is easy to sample (B_l) conditionally on $(Z_1, \ldots, Z_n)$ and $(B_1, \ldots, B_{l-1}, B_{l+1}, \ldots, B_n)$: when the condition (3.8) is satisfied and when the random probability measure associated to $\{\varphi(B_n)\}_{n \geq 1}$ is a Dirichlet process.

In the first case, let $\tilde{k} = \max\left(\varphi^{(2)}(B_j), j \in \{1, \ldots, l-1, l+1, \ldots, n\}\right)$; the same trick adopted to illustrate that condition (3.8) allows to center the prior

distribution on F_0 points out the equality:

$$\sum_{\tilde{\mathcal{S}}} E(W_{i,k}|B_1,\ldots,B_{l-1},B_{l+1},\ldots,B_n)\beta(\cdot,i,k-i+1) = \tag{3.11}$$
$$= \sum_{k<\tilde{k}}\sum_{i=1}^{k} P(\varphi(B_l)=(i,k)|B_1,\ldots,B_{l-1},B_{l+1},\ldots,B_n)\beta(\cdot,i,k-i+1) +$$
$$+ \sum_{i=1}^{\tilde{k}} P(\varphi(B_l^{\tilde{k}})=(i,\tilde{k})|B_1,\ldots,B_{l-1},B_{l+1},\ldots,B_n)\beta(\cdot,i,\tilde{k}-i+1)$$

where $B_l^{\tilde{k}}$ denotes the x_0-block truncated at the $\tilde{k}$-th component. The equality (3.11) allows to sample from the conditional distributions of $\varphi(B_l)$ and B_l given $(Z_1,\ldots,Z_n)$ and $(B_1,\ldots,B_{l-1},B_{l+1},\ldots,B_n)$:

$$P(\varphi(B_l)=(i^*,k^*)|Z_1,\ldots,Z_n,B_1,\ldots,B_{l-1},B_{l+1},\ldots,B_n) = \tag{3.12}$$
$$= \frac{E(W_{i^*,k^*}|B_1,\ldots,B_{l-1},B_{l+1},\ldots,B_n)\beta(F_0(Z_l),i^*,k^*-i^*+1)}{\sum_{\mathcal{S}} E(W_{i,k}|B_1,\ldots,B_{l-1},B_{l+1},\ldots,B_n)\beta(F_0(Z_l),i,k-i+1)}$$

$$P(B_l=B^*|\varphi(B_l)=(i^*,k^*),Z_1,\ldots,Z_n,B_1,\ldots,B_{l-1},B_{l+1},\ldots,B_n) =$$
$$= \frac{P(B_l=B^*|B_1,B_2,\ldots,B_{l-1},B_{l+1},\ldots,B_n)I_{(i^*,k^*)}(\varphi(B^*))}{\sum_{\varphi(B)=(i^*,k^*)} P(B_l=B|B_1,B_2,\ldots,B_{l-1},B_{l+1},\ldots,B_n)}$$

Finally, adapting the expression (3.11), the density function of Z_{n+1} conditionally on $(Z_1,\ldots,Z_n)$ and $(B_1,\ldots,B_n)$ in z

$$F_0'(z)\sum_{\tilde{S}} E(W_{i,k}|B_1,B_2,\ldots,B_n)\beta(F_0(z),i,k-i+1)$$

can be exactly computed.

The sampling procedure in the second case is much the same as in the first; the only difference consists in the computation of the left member of the equality (3.11) which, exploiting the fact that the Dirichlet prior is conjugate, is equal to:

$$\frac{G(\tilde{S})f(\cdot)}{G(\tilde{S})+n-1} + \frac{\sum_{j\neq l}\beta(\cdot,\varphi^{(1)}(B_j),\varphi^{(2)}(B_j)-\varphi^{(1)}(B_j)+1)}{G(\tilde{S})+n-1}$$

where f is the density function of the r.v. $F_0(Z_1)$ and G is the parameter of the Dirichlet process.

If the specification of the prior distribution is different from those in the two discussed cases the Gibbs sampling exploits the approximations

$$E(W_{i,k}|Z_1,\ldots,Z_n) \approx E(W_{i,k}|Z_1,\ldots,Z_n,\max\{\varphi^{(2)}(B_1),\ldots,\varphi^{(2)}(B_n)\} < \tilde{k})$$

where $\tilde{k}$ is a fixed large integer, and sample from the conditional distribution of B_l given $(Z_1,\ldots,Z_n)$, $(B_1,\ldots,B_{l-1},B_{l+1},\ldots,B_n)$ and $\varphi^{(2)}(B_l) < \tilde{k}$.

We have applied the described algorithm to the Buffalo snowfall data. This dataset consists of 63 observations and it has been extensively used in the density estimation literature, see for example Silverman (1986). We recall that the predictive d.f. is the optimal Bayesian estimate of the unknown d.f. under a quadratic loss function.

Four different prior distributions are considered. F_0 is set equal to a truncated Normal distribution with mean $\theta = 70$ and standard deviation $\sigma = 100$ having support $[0, 140]$. The RUP $\tilde{X}$ is parameterized as in Proposition 1 with $G(i,k) = M\frac{e^{-\lambda}\lambda^{k-1}}{k!}$ and the predictive distribution is computed for four different values of the couple (M,λ). The predictive densities suggest a trimodal distribution. The different shapes of the four estimates agree with the interpretation of the parameters M and λ given in section (3.4): the higher is the value of M the more the predictive distribution is similar to the initial guess F_0, while different values of λ correspond to predictive densities with different degrees of smoothness.

In some cases, the proposed model outperforms the Dirichlet-Bernstein prior, as illustrated in the following example. The random distributions in the Dirichlet-Bernstein model are mixtures of a random number K of beta distributions with Dirichlet distributed random weights. The posterior distribution from a Bernstein-Dirichlet prior, when the unknown density is very picked, typically concentrates on distributions that overfit the data on the tails: the densities generated from the posterior have peaks in correspondence to single observations on the tails. This behavior of the Dirichlet-Bernstein model has an intuitive explanation. If a relevant portion of observations lies in a small subinterval of the support, the

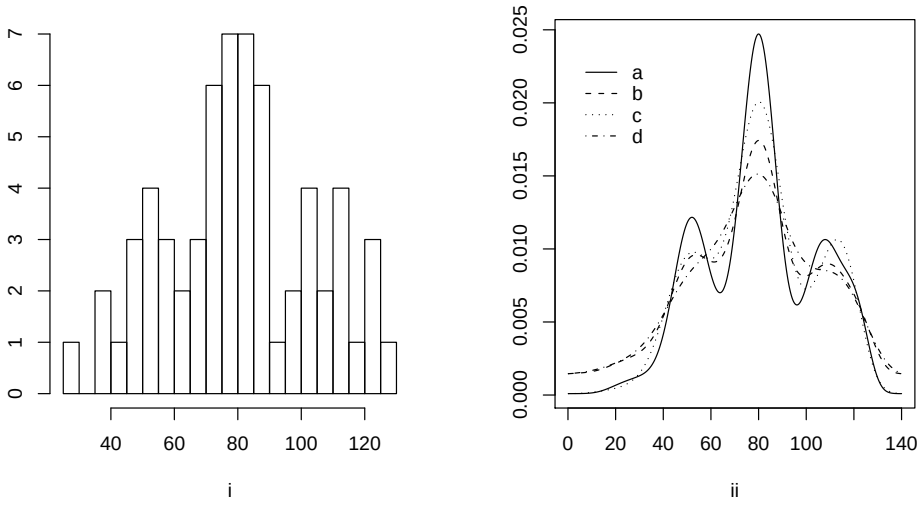


Figure 3.2: i) Histogram of the Buffalo snowfall data. ii) Predictive density functions; a, b, c and d correspond to the parametrizations ($\lambda = 100, M = 1$), ($\lambda = 100, M = 20$), ($\lambda = 50, M = 1$) and ($\lambda = 50, M = 20$).

posterior concentrates on mixtures with a high number of components and, conditionally on a large value of K, the Dirichlet-Bernstein model inherits the peculiarities of the Dirichlet process; in particular, the predictive distribution closely follows that of the Dirichlet process which has point masses on the observations.

The extended Bernstein prior solves the issue: in the Dirichlet-Bernstein model K determines how the mixture components concentrate around their means, while the proposed prior is more flexible in combining components with different shapes. The underlined difference is similar to the improvement that can result if a data-set is fitted through a location-scale mixture of Gaussians (see for example Escobar and West (1995)) rather than by a location mixture.

Figure 3 represents the density estimates obtained fitting a sample generated from a mixture of a truncated Normal density and a Uniform distribution, through the Bernstein model and the proposed extension. We have kept the parameterizations of the two priors as similar as possible. In the first case, $P(K = k) \propto 0.9^k$ and the distribution function F in (3.1) is modeled through a

Dirichlet process centered on the uniform distribution $U[0,1]$, in the second case, the prior is centered on the Uniform distribution and the expected values of the mixing weights $\{W_{j,k}\}_{(j,k)\in\tilde{S}}$ depend only on k: $E(W_{j,k}) = \frac{0.9^k}{9k}$.

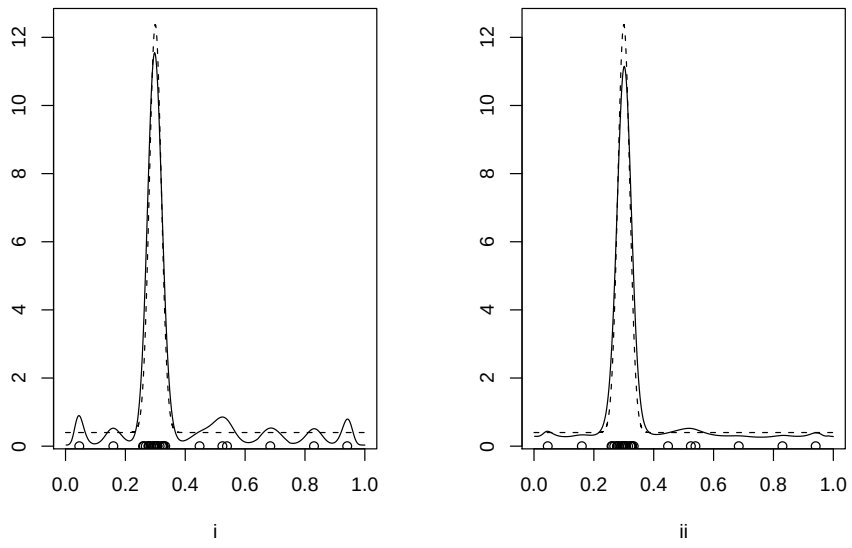


Figure 3.3: i)Bernstein prior. ii)Extended Bernstein prior. Solid lines: predictive densities. Dotted line: the density from which the data have been generated. Sample size=30.

The comparison between the predictive densities in Figure 3 emphasizes the previously mentioned difference. In both cases, the mode of the density from which the data have been generated is well approximated. The predictive densities approach in a similar manner the peak of the unknown density and concurrently appear rather different in correspondence of the two tails. The predictive density obtained updating the Bernstein prior is characterized by several peaks while in the second case the predictive density has approximately flat tails.

3.6 Final remarks

In this chapter, a random mixing measure for mixtures of beta kernels, which is a generalization of the Dirichlet process, is characterized through a RUP and some advantages with respect to the Bernstein prior are illustrated.

The theoretical properties of the Bernstein prior and of other Bayesian mixture models based on the Dirichlet process have been investigated by several authors. Among these, Ghosal (2001), Ghosal et. al. (2008) and Kruijer and van der Vaart (2008) give results on the asymptotic properties of the Bernstein model. The general results of Ghosal et. al. (2008) can be applied to the Bernstein model, for studying the effect of the a priori distribution of the unknown polynomial degree K on the convergence rates of the posterior distribution. The recent work of Kruijer and van der Vaart (2008) underlines the relevance of investigating random mixing measures alternative to the Dirichlet process; they show that a slight modification of the random mixing measure of the Bernstein model, under specific assumptions, improves the posterior rate of convergence.

As illustrated in the previous section, the proposed mixture model may provide a better small sample behavior than the Bernstein model. An open problem, in the above direction of research, is to investigate if it can also improve the convergence rates in density estimation.

Appendix

Proposition 3.1. Under the assumptions, it can be easily verified that, given a finite subset $\{(i_1, k_1), (i_2, k_2), \ldots, (i_m, k_m)\}$ of $\tilde{S}$ and a vector of positive numbers $(w_1, w_2, \ldots, w_m)$ such that $\sum w_j \leq 1$, we have

$$P\left(\bigcap_{j=1}^{m}[w_j - \delta \leq W_{i_j,k_j} \leq w_j + \delta]\right) > 0, \qquad \forall \delta > 0.$$

Then, given a finite set $\{g_1, g_2, \ldots, g_l\}$ of continuous functions and a probability measure Q on $[0,1]$, for every strictly positive ϵ

$$\begin{aligned}\mathbf{P}\left[B \in \mathbf{B} : \max_{j=1,\ldots,l}\left[\left|\int g_j db - \int g_j dQ\right|\right] < \epsilon\right] \\ \geq \mathbf{P}\left[B \in \mathbf{B} : \max_{j=1,\ldots,l}\left[\left|\int g_j db - \int g_j dB_k^Q\right|\right] < \tfrac{\epsilon}{2}\right] > 0\end{aligned} \tag{3.13}$$

where B_k^Q is a Bernstein distribution satisfying the inequalities

$$\left|\int g_j dB_k^Q - \int g_j dQ\right| < \frac{\epsilon}{2} \qquad \forall j = 1, 2, \ldots, l,$$

whose existence is ensured by the fact that Bernstein-distributions are dense in $\triangle$. □

Theorem 3.2. Without loss of generality consider $S = \mathbb{N}$ and $x_0 = 0$. The finite-dimensional laws of the process $\{X_n\}_{n\geq 0}$ are described in (3.3) where μ is the distribution of the random transition matrix Π such that the rows are independent and the i-th row $\Pi(i)$ is a Dirichlet processes with parameter $\beta_i = \{\beta_{i0}, \beta_{i1}, \ldots\}$, where $\beta_{ij} = a_i(c(i,j))$.

Equality (3.5) implies $\sum_j \beta_{ij} = \sum_j \beta_{ji}$. As the probability measure μ is conjugate, conditionally on $[X_0 = x_0, X_1 = x_1, \cdots, X_m = x_m]$, the rows $\Pi(i)$ are independent Dirichlet processes an the parameter of the i-th raw is

$$\left\{\beta_{i0} + \sum_{l=0}^{m-1} I_{(i,0)}(x_l, x_{l+1}), \quad \beta_{i1} + \sum_{l=0}^{m-1} I_{(i,1)}(x_l, x_{l+1}), \ldots\right\}.$$

If $x_m = x_0$,

$$\sum_j \left(\beta_{ij} + \sum_{l=0}^{m-1} I_{(i,j)}(x_l, x_{l+1})\right) = \sum_j \left(\beta_{ji} + \sum_{l=0}^{m-1} I_{(j,i)}(x_l, x_{l+1})\right) \tag{3.14}$$

and the condition (3.5) still holds.

Let us denote $A^{[b]} = \prod_{i=1}^{b}(A + i - 1)$, $\forall\, b \in \mathbb{N}^+$, $\forall A \in \mathbb{R}^+$ and $A^{[0]} = 1$.

If $P(\varphi(B_1) = i) = \dfrac{\beta_{i0}}{\sum_l \beta_{l0}}$ for every $i \in \{0, 1, \ldots\}$, then

$$P(Y_1 = y_1, Y_2 = y_2, \ldots, Y_m = y_m) = \frac{\prod_i \beta_{i0}^{[n_i]}}{\left(\sum_i \beta_{i0}\right)^{[m]}}, \tag{3.15}$$

where $n_i = \sum_{l=1}^{m} I_{(i)}(y_l)$, and the de Finetti measure associated with the sequence $\{Y_n = \varphi(B_n)\}_{n\geq 1}$ is a Dirichlet process with parameter α.

Consider the process $\{X_n^*\}_{n\geq 0}$ with the same state space of $\{X_n\}_{n\geq 0}$ such that $P[X_0^* = x_0, \cdots, X_n^* = x_n] = \int_{\mathcal{P}} \prod_{j=0}^{n-1} \pi(x_j, x_{j+1}) \mu^*(d\pi)$, where μ^* is the distribution of a random transition matrix Π^* such that the rows are independent random measure and the row $\Pi^*(i)$ is a Dirichlet process with parameter $\beta_i^* = (\beta_{0i}, \beta_{1i}, \beta_{2i}, \ldots)$.

Observe that

$$
\begin{aligned}
&P(X_0 = x_0, X_1 = x_1, X_2 = x_2, \cdots, X_{m-1} = x_{m-1}, X_m = x_0) \\
&= \prod_l \left(\frac{\prod_j \beta_{lj}^{[n_{lj}]}}{\left(\sum_j \beta_{lj}\right)^{[\sum_j n_{lj}]}} \right) \\
&\quad = P(X_0^* = x_0, X_1^* = x_{m-1}, X_2^* = x_{m-2} \cdots, X_{m-1}^* = x_1, X_m^* = x_0) \qquad (3.16)
\end{aligned}
$$

where $n_{lj} = \sum_{i=0}^{m-1} I_{(l,j)}(x_i, x_{i+1})$, thus $P(\varphi(B_1) = i) = P(X_1^* = i) = \dfrac{\beta_{i0}}{\sum_l \beta_{l0}}$ and equality (3.15) is verified. □

Proposition 3.2. The proposition is a direct application of Theorem 3.2. The first condition (3.5) is straightforwardly verified. Thus we need to verify the recurrence of the process.

The identity

$$
P\left(\tilde{X}_{k-1} = (i,k)\right) = \frac{\mu_k\left(\left[\frac{i-1}{k}, \frac{i}{k}\right]\right)}{G(\tilde{S})} \quad \forall\, (i,k) \in \tilde{S}. \qquad (3.17)
$$

holds by definition for $k = 1$ and can be recursively verified for every $k \in \mathbb{N}^+$ exploiting the following equality:

$$
\begin{aligned}
P\left(\tilde{X}_k = (i,k+1)\right) &= P\left(\tilde{X}_k = (i,k+1) | \tilde{X}_{k-1} = (i,k)\right) P\left(\tilde{X}_{k-1} = (i,k)\right) + \\
&+ P\left(\tilde{X}_k = (i,k+1) | \tilde{X}_{k-1} = (i-1,k)\right) P\left(\tilde{X}_{k-1} = (i-1,k)\right). \qquad (3.18)
\end{aligned}
$$

Indeed, the first term of the sum in (3.18) is equal to $\dfrac{\mu_{k+1}(\left[\frac{i-1}{k}, \frac{i}{k+1}\right])}{\mu_k(\left[\frac{i-1}{k}, \frac{i}{k}\right])} \dfrac{\mu_k(\left[\frac{i-1}{k}, \frac{i}{k}\right])}{G(\tilde{S})}$

if $k \geq i$ and 0 otherwise, while the second is equal to $\dfrac{\mu_{k+1}(\frac{i-1}{k+1}, \frac{i-1}{k}]}{\mu_k(\frac{i-2}{k}, \frac{i-1}{k}]} \dfrac{\mu_k(\frac{i-2}{k}, \frac{i-1}{k}]}{G(\tilde{S})}$ if $i \geq 2$ and 0 otherwise. It follows that

$P(\varphi(B_1) = (i,k)) = \dfrac{G(i,k)}{G(\tilde{S})}, \qquad \lim\limits_{k\to\infty} P(\varphi^{(2)}(B_1) > k) = 0 \qquad$ and

$\lim\limits_{k\to\infty} P(\varphi^{(2)}(B_1) > k | \tilde{X}_{h-1} = (i,h)) = 0 \qquad \forall (i,h) \in \tilde{S}\cdot$

Then the equality

$$\lim_{k\to\infty} P\left(\varphi^{(2)}(B_i) > k \middle| \varphi^{(2)}(B_j) = k_j; j \in \{1,\dots,i-1\}\right) = \tag{3.19}$$

$$= \lim_{k\to\infty} \sum_i \left[P\left(\tilde{X}_{\hat{k}+\sum_j k_j} = (i, \hat{k}+1) \middle| \varphi^{(2)}(B_j) = k_j; j \in \{1,\dots,i-1\} \right) \right.$$

$$\left. P\left(\varphi^{(2)}(B_1) > k \middle| \tilde{X}_{\hat{k}} = (i, \hat{k}+1) \right) \right] = 0$$

where $\hat{k} = \max_{j\in\{1,\dots,i-1\}}\{k_j\}$, proves the recurrence of the process. □

Bibliography

Diaconis, P. and Freedman, D. (1980) de Finetti's theorem for Markov chains. *The Annals of Probability*, Vol. 8, No. 1, 155–130.

Escobar, M.D., West, M (1995) Bayesian density estimation and inference using mixtures. *Journal of the American Statistical Association*, Vol. 90, No. 1, 577–588.

Ferguson, T. S. (1973) A Bayesian analysis of some nonparametric. problems. *Ann. Statist.*, Vol. 1, No. 2, 209–230.

Ghosal, S. (2001)Convergence rates for density estimation with Bernstein polynomials. *Ann. Statist.*, Vol. 29, No. 5, 1264–1280

Ghosal, S., Lember, J., van der Vaart, A. (2008)Nonparametric Bayesian model selection and averaging. *Electronic journal of statistics*, Vol. 2, 63–89

Kruijer, W., van der Vaart, A. (2008)Posterior convergence rates for Dirichlet mixtures of beta densities. *Journal of statistical planning and inference*, Vol. 138, 1981–1992.

Muliere, P., Secchi, P., Walker, S. G. (2000) Urn schemes and reinforced random walks. *Stochastic Processes and their Applications*, 88, 59–78.

Petrone, S. (1999) Random Bernstein polynomials. *Scandinavian Journal of Statistics*, 26, 373–393.

Petrone, S. and Wasserman, L. (2002) Consistency of Bernstein Polynomial Posteriors. *Journal of the Royal Statistical Society. Series B* , Vol. 64, No. 1, pp. 79-100.

Silverman, B.W. (1986) Density estimation for statistics and data analysis. Chapman and Hall, London.

Chapter 4

A Bayesian approach to randomized discontinuation trials design

Abstract

During the last years the randomized discontinuation design has been successfully applied in many clinical trials. Most applications are to oncology phase II trials for cytostatic agents. The design consists of two stages, a first preliminary open stage and a subsequent phase during which a subgroup of patients are randomly treated with the investigated agent or with a control therapy. The design is characterized by the following tuning parameters: the duration of the preliminary stage, the number of patients in the trial, and the selection criterium for the second stage. We discuss an optimal choice of the tuning parameters based on a Bayesian decision theoretic framework. We define a probability model for putative cytostatic agents and specify a suitable utility function. A computational procedure to select the optimal decision is illustrated and the efficacy of the prosed approach is evaluated through a simulation study.

4.1 Introduction

During the last decade several non-cytotoxic agents have been studied in attempts to establish their efficacy for inhibiting cancer cell growth. The clinical development of cytostatic agents, as recommended in literature, usually involves phase I,II, and III trial. It has been recognized that phase II trials in such process play a key role in that they screen between the multitude of cytostatic agents actually studied and focus the amount of resources required for phase III trials on the most promising; this peculiarity continues to stimulate interest on their design finalized to improve their efficiency.

The randomized discontinuation design (RDD) has been applied in oncological phase II clinical studies (Stadler (2007)) for assessing the cytostatic properties of new agents. It consists of two stages (Rosner et al (2002)): in the first one all patients enrolled in the clinical trial are treated with the new agent, at the end of this period the progression of the disease of each patient is evaluated and those which have attained a state of stable disease participate to the second stage during which they are randomly treated with the putative agent or the standard of care.

The structure of the RDD, if compared with other phase II designs for anticancer therapies, is motivated by cytostatic agents characteristics which differentiate these from the cytotoxic therapies. Most part of the clinical trials in oncology evaluate if a putative anticancer drug may shrink a tumor mass while the appreciable outcome of a cytostatic agent is a state of stable disease. Another peculiarity of many cytostatic agents due to the biological mechanisms of action is that they can be supposed to have a greatly heterogeneous effectiveness on a court of patients with similar prognostic characteristics. Millar and Linch (2003) underline that, due to these characteristics, in much cases, useful clinical designs to evaluate cytotoxic therapies are inappropriate for cytostatic agents.

The first stage of the RDD selects an homogeneous group of patients while the second is finalized to evaluate if the experimental outcomes evidence a cytostatic

activity of the studied agent. Producing an homogeneous group is useful to evaluate the agent because the patients population participating to the first stage is usually characterized by greatly heterogeneous tumor growth rates, moreover the selected patients are the most likely to substantially benefit from the treatment. It follows that, if preclinical studies guarantee that the agent doesn't affect the tumor growth after the treatment is stopped, the RDD, if compared with a two arm randomized trial, can significantly improve the probability the trial detects the biologic activity of the agent (Fedorov and Liu (2005)).

The RDD implementation requires the choice of some specific features of the design. We discuss the choice of the number of patients participating to the first stage of the clinical trial and of the durations of the two stages, while the eligibility criteria to participate in the second stage will be assumed settled through ethical and medical considerations. The choice of the three tuning parameters is critical to achieve a suitable balance between the resources employed in the trial and its capability to detect the potential cytostatic activity of the investigated agent. Stadler (2007) point out their relevance and illustrates that a too short prerandomization phase can be insufficient to distinguish the subpopulation of those patients that substantially benefit from the treatment while, on the other hand, a too long open phase requires an excessive number of enrolled patients, because of the high risk of a small portion of patients eligible for randomization.

The choice of the features of the RDD is formalized adopting a Bayesian decision-theoretic approach (DeGroot (2004)). In this setting the decision maker considers different scenarios, which could correspond to various degree of efficacy of the putative agent, and specifies a probability measure on the scenarios'set which reflects its a priori beliefs. The choice problem, in this case the design to be adopted, is then solved through a utility function representative of the benefit the decision maker associates to each combination of scenario, decision and experimental observations: the optimal choice maximizes the expected utility.

The a priori knowledge of the institution that plans the clinical trial, involved in the decisional framework and necessary to specify a probability measure on a

scenarios'set, can be based on many factors as experimental data from preclinical studies and from the phase I trial or the historical experience of the progression of the disease when treated with the standard of care. Korn et al. (2001) gives a detailed account of the clinical and preclinical basis to be considered in designing phase II trials for cytostatic agents.

The utility function synthesizes the most relevant aspects of the study that will depend on the adopted design, the unknown efficacy of the putative agent and the experimental outcomes that will be observed. The costs and benefits for the patients participating in the study, as well as for the future patients, will depend on all these three factors. The final decision wether to perform a phase III trial or to consider the agent ineffective will be a function of the experimental observations. The necessary resources to conduct the phase II trial, if early stopping rules are considered, will be determined by the chosen clinical design and by the observed outcomes. Further components as the scientific knowledge that will be acquired during the trial and the commercial potential of the new agent can be simultaneously captured by the utility function. Even if a plurality of interests are involved in the clinical trial design; a decision-theoretic approach and in particular specifying a utility function clarifies which guidelines and principles are adopted for planning the trial.

Clinical trials simulations are widely recommended in literature (Nestorov, et al. (2001))and are usually conducted to compare competing designs. Simulations are helpful devices to evaluate the experimental designs' properties under different scenarios. A simulation study is based on two major components: a set of alternative designs and a set of probabilistic models. The purpose is to identify the alternative with the better overall performance. This approach implies, also if not explicitly, an a priori guess on the possible scenarios that could be verified during the clinical trial and the ability of valuing each combination of design, scenario and experimental observations. These two guidelines are reflected both by the choice of the probability models set of the simulation study and by the informal way of formulating an overall assessment of each considered design which

summarizes its performances under different scenarios and its associated costs. The two mentioned implicit basis typical of a simulations study are formalized by the Bayesian decision-theoretic approach by mean of the utility function and of the prior distribution on a probabilistic models' set. The utility function explicates the adopted criterions to evaluate alternative designs, such clarification is particularly useful when different experimental schemes are evaluated. As an example consider a comparison between a RDD and a two arm randomized trial with same duration and number of involved patients, if different costs are associated with the randomization of the patients at the beginning of the trial and with the randomization of a subgroup after an interim analysis, then the evaluation cannot be based exclusively on the designs'power to detect the biologic activity of the agent but should also reflect such difference; an adequate utility function can easily integrate these two perspectives.

The decision-theoretic approach to the experimental design usually assume that the decision maker maximize the expected utility choosing simultaneously the design of the experiment that will be conducted and how its future actions will depend on the experimental observations, in our case decides which outcomes will suggest the biological activity of the agent. In the clinical trials setting it can be opportune to constrain the relationship between the experimental observations and the future actions and to consider that the authorities could refuse to validate the experimental evidence following a Bayesian framework and adopting the decision maker prior distribution due to a lack of symmetry in the a priori knowledge or to prudential concerns. The choice of the clinical trial characteristics can be formalized through a decision-theoretic framework which consider these constraints: the prior distribution does not depend on the presence of any constraints while the utility function has to adapt to them. As an example, consider a set of experimental data which, from the perspective of the decision maker, could suggests the biological activity of the agent but would be insufficient for the authority to prove it; the benefit the decision maker associates to these experimental observations, if combined with a costly design and an effective agent,

depends on the constraints established by the authority. Also in absence of any authorities a constrained modeling of the decisional problem can be appropriate to account for the decision maker prudential concerns. Further constraints, as a maximum number of patients involved in the trial or a maximum period of treatment, due to medical or financial arguments, can be formalized in the decision-theoretic framework.

4.2 Tumor growth model and prior specification

Many tumor growth models for specific patients populations both stochastic and deterministic have been proposed in literature. In what follows we model the tumor growth through a Gompertzian diffusion process (Ferrante et al. (2000)), nevertheless the decisional framework we propose can be straightforwardly generalized to whichever alternative model.

The cell cycle is constituted by the proliferative phase, characterized by the cell growth and the DNA reproduction followed by the cell split into two distinct cells, and by the quiescent phase. These two phases are regulated by proliferative and anti-proliferative signals from which depend the cell transitions from one phase to the other. At the beginning of the tumor mass growth process a large proportion of cells are in the proliferative phase, resulting, initially, in an exponential dynamic followed by a decreasing growth rate which can be related with a lack of oxygen and nutrient in the tumoral tissue. The Gompertz curve model often well approximate tumor growth data (Ribba et al. (2006)) reproducing the initial exponential dynamic and the subsequent growth rate decrease. The Gompertz function solves the following differential equation :

$$dX_t/dt = a \cdot X_t - b \cdot X_t \cdot \log(X_t) \qquad X_0 = x_0.$$

Where a and b are two constants and X_t represents the tumor volume at time t.

We assume the individual tumor proliferation is characterized by the same relationship between its local growth rate and its volume, expressed by the

above equation, but randomly perturbed by contingent factors which are modeled through a Brownian motion. The resulting random process being the solution of the following stochastic differential equation (sde):

$$dX_t^i = [(a_i X_t^i - b_i X_t^i \log(X_t^i)]dt + \sigma_i X_t^i dW_t^i \qquad X_0^i = x_0^i \quad t \in [0,T]$$

where i indexes the i-th patient participating in the trial and $\{W_t^i\}_{i\in I}$ are independent Brownian motions. As well as in the deterministic equation if b is null the growth is exponential in the stochastic one, under the same condition, the process becomes a geometric Brownian motion. The stochastic model describes the tumor growth if the individual patient is treated with the same therapy during the period $[0,T]$, while if two therapeutic regimens are involved the resulting process will solve the same sde with a_i and b_i substituted by two picewise constant functions: $a_i(t) = a_i^0 I(t \in T_i^0) + a_i^1 I(t \in T_i^1)$ and $b_i(t) = b_i^0 I(t \in T_i^0) + b_i^1 I(t \in T_i^1)$ where T_i^0 and T_i^1 are the time intervals during which the therapies are administered. In the former case the solution of the sde (Ferrante et al. (2000)) is:

$$X_t^i = \exp\left[\frac{a_i - \sigma_i^2/2}{b_i} + \left(\log(x_0^i) - \frac{a_i - \sigma_i^2/2}{b_i}\right)e^{-b_i t} + \sigma_i \int_0^t e^{-b_i(t-s)} dW_s^i\right]. \tag{4.1}$$

In the latter the markovian property of the process allows to easily compute the finite-marginal distributions which are multivariate lognormal, the transition probability density of the process (4.1) being:

$$\begin{aligned} f_{X_{t+s}|X_t}(x_{t+s}) = \frac{1}{x_{t+s}}\left[2\pi\sigma^2\frac{1-e^{-2bs}}{2b}\right]^{-1/2} \cdot \\ \cdot \exp\left[-\frac{[\log(x_{t+s}) - e^{-b_s}\log(X_t) - \frac{a-\sigma^2/2}{b}(1-e^{-bs})]^2}{\sigma^2\frac{1-e^{-2bs}}{b}}\right] \end{aligned} \tag{4.2}$$

Under the hypothesis that the new treatment, labeled 1, is superior with respect to the standard of care the inequality

$$E(X_{t+s}^i | X_t^i = x_t, a_i^0, b_i^0, \sigma_i) > E(X_{t+s}^i | X_t^i = x_t, a_i^1, b_i^1, \sigma_i) \tag{4.3}$$

holds for every i, x_t and s: the treatment permanently improve the patients' condition. Considering the derivative of the conditional expectations with respect

s when s is null it can be observed that the above inequality is permanently verified if and only if, for every positive x_t, $(-b_i^0 \log(x_t) + a_i^0) > (-b_i^1 \log(x_t) + a_i^0)$ or equivalently if $a_i^1 < a_i^0$ and $b_i^1 = b_i^0$.

$$E(X_{t+s}|X_t = x_t, a, b) =$$
$$= \exp\left[e^{-bs}\log(x_t) + \frac{a - \sigma^2/2}{b}(1 - e^{-bs}) + \frac{\sigma^2}{4b}(1 - e^{-2bs})\right]$$
$$\frac{d\left(\log E[X_{t+s}|X_t = x_t, a, b]\right)}{ds} = \left(-b\log(x_t) + a - \sigma^2/2\right)e^{-bs} + \frac{\sigma^2}{2}e^{-2bs}$$

The following theorem underlines also that if $a^1 < a^0$ and $b^1 = b^0$ the conditional distribution $P(X_{t+s} \in \cdot|X_t = x, a^1, b^1, \sigma)$ is stochastically dominated by $P(X_{t+s} \in \cdot|X_t = x, a^0, b^0, \sigma)$.

Theorem 4.1. [*Levy (1973)*] *Let F and G be two alternative log-normal distributions. Let X_F and X_G two r.v. with distributions F and G. F stochastically dominate G if and only if $E(\log(X_F)) > E(\log(X_G))$ and $Var(\log(X_F)) = Var(\log(X_G))$.*

In our case the equalities $Var(\log(X_{t+s})|X_t = x, a, b, \sigma) = \frac{\sigma^2}{2b}(1 - e^{-2bs})$ and $E(\log(X_{t+s})|X_t = x, a, b, \sigma) = \frac{a-\sigma^2/2}{b} + (\log(x) - \frac{a-\sigma^2/2}{b})e^{(-bs)}$ guarantee the stochastic inequality.

Two classes of possible scenarios are considered in the choice problem. The first one corresponds to the hypothesis that the treatment doesn't effect the tumor growth: in this case the individual patients tumoral processes are entirely characterized by the parameters $a_i^0 = a_i^1, b_i$ and σ_i which vary across the heterogeneous population accordingly with an unknown distribution. The second class corresponds to the hypothesis that the new agent inhibits or accelerate the tumor growth of the individual patients; in this case a_i^0, b_i, σ_i and a_i^1 parameterize the laws of the processes. In the following paragraph a prior distribution of the future experimental observations, that includes these scenarios, is specified.

A parametric distribution F_θ is chosen for three functionals of the Gompertzian diffusions

$$\varphi_{1i} = \frac{a_i^0 - \sigma_i^2/2}{b_i}, \qquad \varphi_{2i} = \frac{\sigma_i^2}{2b_i}, \qquad \varphi_{3i} = b_i$$

which regulate, if the patients do not assume the new agent, the tumors growth processes after these had reached a minimal threshold ε. φ_{1i}, φ_{2i} and φ_{3i} have a clear interpretation. The lognormal conditional distribution (4.2) is parameterized by $\mu = e^{-b_i s}\log(X_t) + \frac{a_i^0 - \sigma_i^2/2}{b_i}(1 - e^{-b_i s})$ and $\lambda^2 = \frac{\sigma_i^2}{2b_i}(1 - e^{-2b_i s})$. μ is a convex linear combination between φ_{1i} and $\log(X_t)$, φ_{2i} is a multiplicative constant from which depends λ^2 and, for fixed values of φ_{1i} and φ_{2i}, the conditional distribution is entirely characterized by the product $\varphi_{3i}s$: if φ_{3i} and s are respectively multiplied and divided by a positive constant the transition density doesn't change.

The probabilities π_1 and π_{-1} are specified for the events $\{E = 1\}$ and $\{E = -1\}$ that the new agent inhibits or accelerates the tumor growth, a parametric distribution G_θ is chosen for the differences $\psi_i = (\frac{a_i^0 - \sigma_i^2/2}{b_i} - \frac{a_i^1 - \sigma_i^2/2}{b_i})$, and a prior distribution is specified on the parameters space Θ.

$\theta_{j1}, \theta_{j2}|E$ for $j = 1, 2, 3, 4, 5$, are independently $normalgamma(\mu_j, \tau_j, \nu_j, \rho_j)$ distributed.

$\varphi_{ji}|E, \theta$ for $i \in I$ and $j = 1, 2, 3$ are independently $lognormal(\theta_{j1}, \theta_{j2}^{-1})$ distributed.

$\psi_i|\theta, E = 1$ for $i \in I$ are independently $lognormal(\theta_{41}, \theta_{42}^{-1})$ distributed, symmetrically $-\psi_i|\theta, E = -1$ are $lognormal(\theta_{51}, \theta_{52}^{-1})$ distributed, while if the agent doesn't affect the tumor growth $\{E = 0\}$ the quantities ψ_i are null.

Finally a density function g is specified for the times $\{t_i\}_{(i\in I)}$ from when the tumor reach the threshold ε to the beginning of the trial: $g(t_i) = \frac{1}{B-A}I(t_i \in [A, B])$. The sequence $\{t_i\}_{(i\in I)}$ is assumed independent with respect to $\{\varphi_{1i}, \varphi_{2i}, \varphi_{3i}, \psi_i\}_{(i\in I)}$.

The outlined probability model for the tumoral processes combines the a

priori distributions for the quantities $\varphi_1, \varphi_2, \varphi_3$ and ψ, whose interpretations are directly suggested by the transition probabilities of the Gompertzian process. These quantities are assumed to vary across the enterogenous population; the normal-gamma prior allows to formulate the initial guess about their distributions taking into account the strength of the a priori knowledge. Frequently in Bayesian statistics the elicitation procedures of informative prior, postulate a priori the existence of imaginary data sets representative of the initially available knowledge and update a non informative prior through the unobserved data. This strategy can effectively guide the elicitation of the adopted normal-gamma distributions and clarify which parameterizations of the prior adequately represents the inial guess, indeed when the Jeffreys' prior for log-normal distributed data is updated a normal-gamma distribution is obtained (Padgett and Wei (1977)).

In the hypothesis that historical data from patients treated with the standard of care or with the novel agent have been collected, the elicitation of the prior distribution can be driven from these observations; if the tumoral growths have been periodically measured, a practical elicitation strategy consists in the plug in of the estimates (Gutierretz et al. (2006)) of the parameters of the individual growth processes in the Jeffreys' distribution to obtain an informative prior.

A possible alternative to the described continuous prior distribution could be the elicitation of a discrete one concentrated on a finite number of points, indeed such a specification could facilitate the interpretation of the a priori information underlying the decision-making process when it need to be communicated to an authority or to a plurality of subjects. The application of a discrete prior, as well as a simulations study of the frequentist properties of alternative designs, clarify the set of scenarios considered by the decision maker; of course the decision-theoretic framework allows to weigh the scenarios and to integrate the possible performances of a single design by mean of the utility function.

The elicitation of the prior distribution, in particular when the informative basis consist exclusively in the knowledge of a plurality of experts, is a critical momentum of the decisional process, also if considerable research has been

conducted to improve the elicitation procedures (see for example Garthwaite el al (2005) and references therein), the involved difficulties when is necessary to deal, as in our case, with a considerable number of uncertain quantities are still recognized. Such difficulties in the trial design problem can be partially soften. The RDD has been applied to evaluate the disease-stabilizing activity of new anticancer therapies adopting dichotomous primary end points; it follows that the frequentist properties of the trial depends uniquely by three unknown parameters: the probability of a patient to be eligible for randomization and, conditionally on the eligibility, the response probabilities under the treatment and control regimens. The application of the decisional theoretic framework can be therefore validated computing the first and second error rates induced by the optimal design the for hypothetical values of these probabilities. The validation of the optimal decision is particularly suitable if difficulties arise eliciting the experts knowledge about the tumoral growth processes across the heterogeneous population but, at the same time, the experts can explicit their beliefs about the three pivotal probabilities.

4.3 Decisional problem

The utility gain from the trial $U(d, \phi, X)$ is a function of the adopted design $d \in D$, the parameters which characterize the investigated phenomena $\phi \in \Phi$, in our case $\phi = (E, \theta)$, and of the experimental observations $x \in X$. Its a priori expectation is

$$E(U(d)) = \int_{\Phi} \int_{X} U(d, \phi, X) P(dX|\phi, d) Q(d\phi) \tag{4.4}$$

where Q is the prior distribution on Φ. The choice problem of selecting a design from the action space D coincides with the maximization of the expected utility: $E(U(d^*)) = \max_{d \in D} E(U(d))$.

The action space is constituted of alternative designs which differ in the overall number of patients participating in the trial N and in the durations of the two

phases T_1, T_2. The utility associated to each combination (d, ϕ, X), is a balance between the costs of the trial and the benefits which derive from recommending a phase III trial for an effective agent: $U(d, \phi, X) = B(d, \phi, X) - C(d, \phi, X)$. The institution recommends to continue the study if the experimental observations give, from a frequentist point of view, a significant evidence of the superiority of the novel agent with respect to the standard of care: $R_d(X) \in S_d$, where R_d is the adopted test static and S_d is the rejection set of the null hypothesis that the novel treatment is not superior to the control. The costs of the DDS depend on the number of patients participating to the two phases (N, n) and on their durations:

$$C(d, \phi, X) = c_1 N + c_2 n + c_3(T_1 N + T_2 n)$$

c_1 and c_2 are the costs associated to the individual patients participating in the trial and to those that after having reached a state of stable disease could discontinue the therapy, while c_3 expresses a cost for each unit of time the patients are involved in the trial. The benefits from (d, ϕ, X), when the experimental data suggests a significant cytostatic activity of the agent are proportional to the degree of efficacy of the therapy:

$$B(d, \phi, X) = I(R_d(X) \in S_d) E(\log(1 + \psi^+)|\phi);$$

the more the treatment inhibits the tumor growth processes the more valuable is conducting further studies.

The eligibility criteria of the RDD avoids that those patients for which, at the end of the first stage, a conspicuous growth or a consistent shrinkage of the tumor mass is observed, participate to the second stage of the trial. The i-th patient participates to the second phase if

$$\delta_1(X_{0i}, T_1) \geq X_{T_1 i} \geq \delta_2(X_{0i}, T_1);$$

the two bounds δ_1 and δ_2 depends on the tumor volume at the beginning of the trial X_{0i} and on the length of the first stage T_1. The above indicated criteria is motivated by the need to not discontinue the treatment to those patients who

seams to have appreciably benefited from the treatment and by the enrichment strategy of the design intended to select an homogenous group of patients likely to give evidence of the cytostatic activity of the putative agent.

The optimal strategy d^* can be computed through a Monte Carlo optimization algorithm. The intuitive idea on which is based the algorithm is that the expected utility (4.4) associated to each point d of the designs space D can be approximated through a Monte Carlo procedure, simulating iteratively, accordingly with the prior distribution, the population parameters ϕ as well as the experimental outcomes X and computing at each iteration the utility function $U(X, \phi, d)$. The empirical mean of the iteratively sampled quantity $U(X, \phi, d)$ converges to the expectation of interest. We can reasonably assume that the designs space is finite: the number of days of the two stages of the trial as well as the number of involved patients can vary from 0 to a fixed maximum. To improve the efficiency of the algorithm, instead of simulating a large amount of samples for each point of the design space, the $U(d)$ function can be computed fitting a smooth surface $\tilde{U}(d)$ to the Monte Carlo samples set; the approach has been considered in details in Muller and Parmigiani (1995) and allows to approximate the expected utility of a design d exploiting not only the sampled random quantities associated to d but also borrowing strength from those associated to the neighboring designs. The structure of the outlined algorithm is the following:

1. Step 1. Iteratively $\mathcal{N}$ points $\{d_i \in D, i = 1, \ldots, \mathcal{N}\}$ are selected and at each iteration a random quantity U_i equal in distribution to $U(X, \phi, d_i)$ is generated.

2. Step 2. The surface $U(d)$ is approximated fitting a standard least squares multivariate polynomial regression function $\tilde{U}(d)$ of degree $\mathcal{M}$ to the Monte Carlo samples set.

3. Step 3. Finally d^* is computed maximizing $\tilde{U}(d)$.

The computation of the expected utilities surface is reduced to a standard statistical problem. Many parsimonious model selection procedure proposed in literature can be applied to choose from the sequence of nested models which

vary from the polynomial regression of first degree to to the saturated model (Bursham and Anderson (1998)); most part of the selection criteria guarantee the almost sure convergence of the estimates $\{\tilde{U}(d)\}_{d\in D}$ to the expected utilities of interest as the Monte Carlo sample size diverges.

For illustrative purpose we apply the proposed optimization procedure to an hypothetical decisional process finalized to design the trial for evaluating a novel therapy. Figure 4.1 gives a representation of the a priori distribution of the tumoral growth process and underlines the variability components embedded in the adopted probabilistic model. The a priori probabilities assigned to the event that the novel agent slow the tumoral growth $\{E = 1\}$ is 0.7 while $P(E = -1) = 0.1$. The randomization in the second phase is balanced, so that half of the patients ($i.e. \frac{n}{2} \leq n_1 < \frac{n}{2} + 1$) continue to receive the novel treatment. The single patients are supposed to be enrolled after a time varying from 10 to 14 months after the tumoral mass has reached the threshold $\epsilon = 0.02$. The eligibility criteria that determines the participation to the blinded phase, as in the trials reported in (Stadler et al. (2005)) and (Ratain et al. (2006)), consists of two thresholds of the relative variation of the tumor mass during the open phase: $\delta_1(X_0, T_1) = 1.3X_0$ and $\delta_2(X_0, T_1) = 0.7X_0$. The primary end point, as in the two cited trials, is dichotomous: the disease is considered stable if the tumor mass increase less than 30% during the second phase. To assess if the novel treatment slow the tumor growth the exact Fisher test with significance level $\alpha = 0.95$ is adopted. The utility function is parameterized as follows: $c_1 = c_2 = 0.004$ and $c_3 = 0.00003$. The action space D is constituted by the designs set $\{[N, T_1, T_2] \in (0, 1, \ldots, 300)^3\}$; the alternative of not performing the trial (*i.e.* $N = 0$), as well as the alternative of performing a traditional two-arm randomized trial ($i.e. T_1 = 0$) are thus contemplated.

The optimal strategy has been computed through a Monte Carlo sample of 10^7 simulated trials with tuning parameters randomly spread across the action space. Figure 4.2 illustrates the orthogonal sections of the fitted surface $\tilde{U}(d)$ which intersect at $d^* = (N = 75, T_1 = 64, T_2 = 90)$.

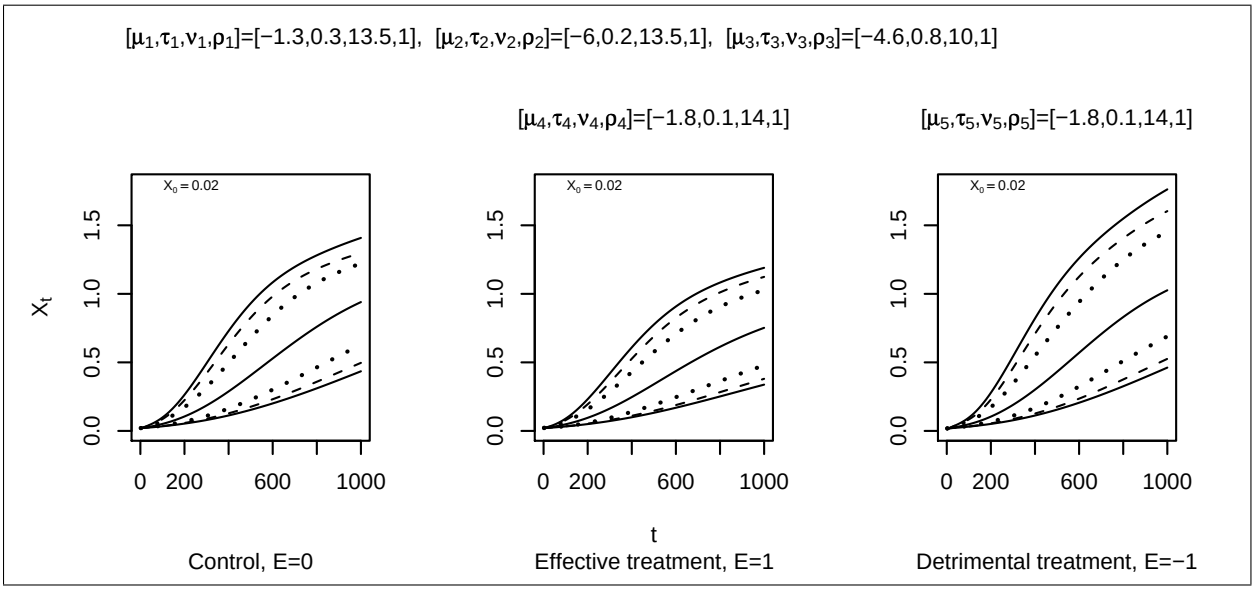


Figure 4.1 Solid line: the a priori median function and the 80% confidence band of the growth process X_t^1. Dashed line: the 80% confidence band of the median function of the conditional distributions $X_t^1|\varphi_i, \psi_i$. Dotted line: the 80% confidence band of the median function of the conditional distributions $X_t^1|\theta$.

The plots display also the approximations of the expected utilities $U(d)$ associated to alternative designs obtained through a standard Monte Carlo procedure confirming the ability of the smoothing algorithm to well approximate the utility surface. The rational decision maker favors the RDD: the optimal discontinuation strategy if compared with the standard two arms trial designs belonging to the action space guarantees a superior expected utility gain. The example underlines also that the choice of the durations of the two RDD phases strongly impacts on the utility of the trial; the illustrated losses of deviating from the optimal selection of T_1 and T_2 are mainly related to the expected loss of power of the trial, indeed the adjunct costs of extending the two phases contribute minimally, i.e. for less than $c_3(\Delta T_1 + \Delta T_2)N$, to the represented variations.

As anticipated the frequentist properties of the RDD, if a dichotomous final end point is adopted, depends exclusively by the probability that a single patient would be eligible for the blinded phase p_e and, conditionally on its eligibility, on the response probabilities under the treatment p_1 and control p_0 regimens.

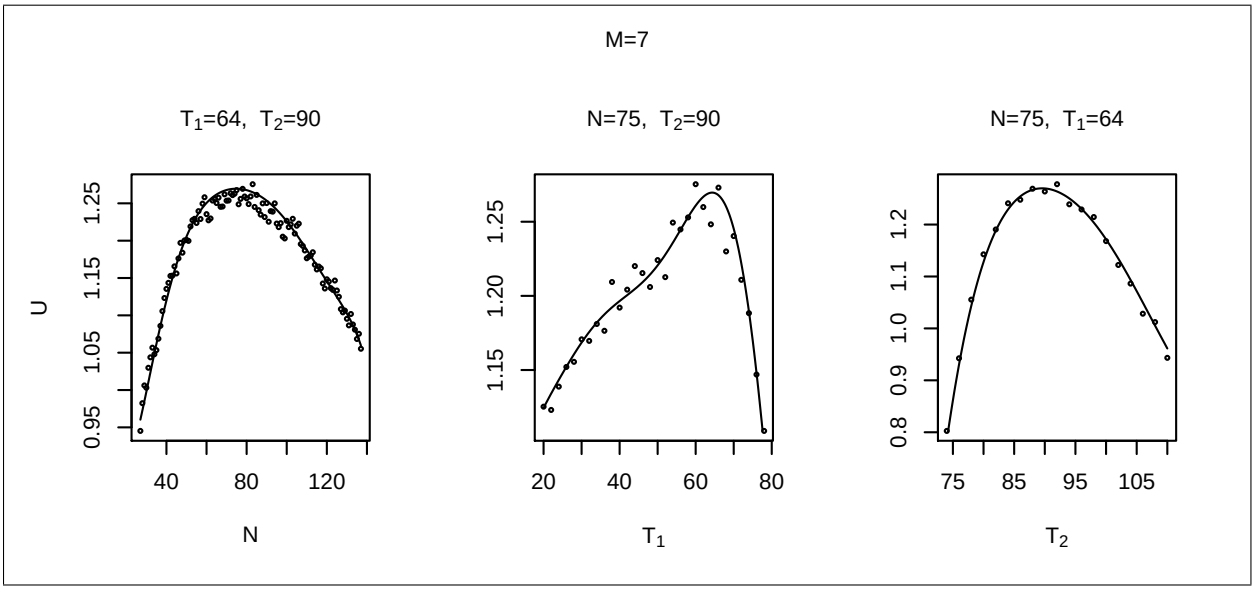


Figure 4.2. Solid line: Fitted expected utility surface $\tilde{U}(d)$. Dots: Monte Carlo approximation of $U(d)$ obtained through 10^4 simulated trials for each design.

Table 1 illustrates the power of the optimal design for hypothetical values of the unknown parameters p_e, p_0 and p_1, as expected the probabilities of rejecting the null hypothesis result satisfactory only for those scenarios in which p_0 is widely superior to p_1. The frequentist operating characteristics reflect the a priori beliefs expressed by the prior distribution, indeed conditionally on on the event $\{p_0 < p_1\}$ it concentrates on large values of their differences.

4.4 A default decisional procedure.

In some cases the institution that plans the trial has access to a considerable amount of longitudinal data inherent the tumor growth processes relative to populations similar to the trial population. The typical instance is the one in which during trials in which have been enrolled similar patients populations tumor growth data have been collected. In such cases it can be desirable to rely as much as possible on these information source, to base the RDD tuning parameters choice on the historical data and to minimize the necessity of elicitating the experts knowledge.

p_e	0.2				0.4				0.6				0.8			
p_1	0.2	0.4	0.6	0.8	0.2	0.4	0.6	0.8	0.2	0.4	0.6	0.8	0.2	0.4	0.6	0.8
p_0																
0.2	0.02	0.08	0.28	0.60	0.02	0.20	0.62	0.93	0.02	0.32	0.81	0.99	0.03	0.42	0.91	1.00
0.4		0.02	0.08	0.28		0.03	0.19	0.61		0.03	0.28	0.81		0.03	0.37	0.92
0.6			0.02	0.08			0.03	0.20			0.03	0.32			0.03	0.42
0.8				0.02				0.02				0.02				0.03

Table 1. Probabilities of rejecting the null hypothesis of ineffectiveness of the novel treatment for hypothetical values of p_e, p_0 and p_1.

In the following paragraphs we describe an ad hoc strategy for specifying an informative prior in the outlined circumstances, as will be evident it is somehow related with the bootstrap techniques. The guidelines that are followed consist in the ideas of restricting the subjective component of the prior specification to the unknown distribution of the treatment effects and of constructing a prior that track as much as possible the data.

We assume to have observed in the historical experience at discrete times the tumor processes of N patients under the control regimen:

$$\{\tilde{X}^0_{1,t_0}, \tilde{X}^0_{1,t_1}, \ldots, \tilde{X}^0_{1,t_k}\}, \{\tilde{X}^0_{2,t_0}, \ldots, \tilde{X}^0_{2,t_k}\}, \ldots, \{\tilde{X}^0_{N,t_0}, \ldots, \tilde{X}^0_{N,t_k}\}$$

the tilde symbol is adopted to distinguish the historical data from the future experimental observations. As will be apparent the procedure that is proposed can be easily adapted to the cases in which the historical patients have heterogeneous follow up periods as well as to those in which the tumor masses have not been measured at fixed scheduled times. To specify the prior distribution for the tumoral trajectories under the control regimen it is stated that the law of the tumoral process of each patient that will be enrolled is a mixture of conditional laws of Gompertz diffusion processes which reflects the partially known historical

trajectories; we adopt the intuitive notation

$$\{X^0_{i,t}\}_{t\geq 0} \sim \frac{1}{N}\sum_{j=1}^{N} \mathbf{GP}(a_j, b_j, \sigma_j | X^0_{i,t_0} = \tilde{X}^0_{j,t_0}, \ldots, X^0_{i,t_k} = \tilde{X}^0_{j,t_k}) \tag{4.5}$$

That is the law of the process is a mixture of N components and each component reflects one of the historical trajectories. The parameters (a_j, b_j, σ_j) in (4.5) are estimated with the standard maximum likelihood technique (Gutierretz et al. (2006)). An appealing slight modification to the scheme, when few measurements for each patient are available, consists in adequately stratifying the historical patients population (see for example Schnatter and Kaufmann (2008)) and in the adoption of the assumption that the three parameters in each stratum are constant. Note that due to the fact that the Gompertz diffusion is a lognormal Markov process it is straightforward to sample at discrete times a process with the defined law.

To characterize the future tumoral trajectories under the treatment regimen we associates to each historical patient the law of hypothetical longitudinal observations, such distributions are representative of the a priori guess about the treatment effects that would have been observed on the historical population if treated with the novel agent. These finite dimensional laws depend exclusively on the historical data and on the a priori distribution of the parameters ψ_j which can be interpreted as the treatment effects and are modeled exactly as in the previous sections. It is exploited the linear relation

$$\log(\tilde{X}^0_{i,t_{j+1}}) =$$

$$\gamma(b_i, t_j - t_{j+1}) \log(\tilde{X}^0_{i,t_j}) + \beta(a_i, b_i, \sigma_i, t_j - t_{j+1}) + \delta(b_i, \sigma_i, t_j - t_{j+1}) Z_{ij} \tag{4.6}$$

of the Gompertz model, where the array $\{Z_{ij}; i \in (1, \ldots N), j \in (0, \ldots k-1)\}$ is constituted of independent standard gaussian r.v.s. Solving the above equations the realizations of the gaussian r.v.s Z_{ij} are computed and plugged in the stochastic linear expressions

$$\log(\tilde{X}^1_{i,t_{j+1}}) =$$

$$\gamma(b_i, t_j - t_{j+1}) \log(\tilde{X}^1_{i,t_j}) + \beta(a_i - \psi_i b_i, b_i, \sigma_i, t_j - t_{j+1}) + \delta(b_i, \sigma_i, t_j - t_{j+1}) Z_{ij} \tag{4.7}$$

which are the analog of the relations in (4.6) for the treatment regimen; the only random components in (4.7) are the treatment effects ψ_j. The hypothetical observations $(\tilde{X}^1_{1,t_0}, \ldots, \tilde{X}^1_{1,t_k}, \ldots, \tilde{X}^1_{N,t_k})$ are defined as the random solutions of the equations in (4.7) assuming that for every historical patient $\tilde{X}^0_{i,t_0} = \tilde{X}^1_{i,t_0}$. Finally conditionally on $\tilde{X}^1 = (\tilde{X}^1_{1,t_0}, \ldots, \tilde{X}^1_{1,t_k}, \ldots, \tilde{X}^N_{1,t_k})$ and on the random vector $\psi = (\psi_1, \ldots, \psi_N)$ the future tumoral trajectories under the treatment regimen are characterized as follows:

$$\{X^1_{i,t}\}_{t \geq 0} | \tilde{X}^1, \psi \sim \frac{1}{N} \sum_j \mathbf{GP}(a_j - \psi_j b_j, b_j, \sigma_j | X^1_{i,t_0} = \tilde{X}^1_{j,t_0}, \ldots, X_{i,t_k} = \tilde{X}^1_{j,t_k}) \tag{4.8}$$

This simple plug in strategy can be easily extended to characterize the law of the tumoral process of a patient that will be under the treatment regimen till time $T1$ and subsequently under the control regimen. Also in this case the process law can be represented as a mixture of N components, one for each historical patient and each component consists in the conditional law of a Gompertz diffusion process whose transition densities switch parametrization at time T_1. The parameters (a_j, b_j, σ_j) as in (4.8) are estimated through the standard maximum likelihood technique and the random quantities ψ are adequately modeled, it is thus exclusively necessary to suitably substitute the hypothetical observations. It suffices to compute for each historical patient the distribution of $\tilde{X}^0_{i,T_1}$ conditionally on $\{\tilde{X}^0_{i,t_0}, \tilde{X}^0_{i,t_2}, \ldots, \tilde{X}^0_{i,t_k}\}$, to slightly modify the equations in (4.6) and (4.7) adding to the vector $(t_0, t_1, \ldots, t_k)$ the instant T_1 and substituting in (4.7) $\beta(a_i - \psi_i b_i, b_i, \sigma_i, t_j - t_{j+1})$ with

$$\beta(a_i - \psi_i b_i, b_i, \sigma_i, t_j - t_{j+1}) I(t_{j+1} \leq T_1) + \beta(a_i, b_i, \sigma_i, t_j - t_{j+1}) I(t_{j+1} > T_1)$$

The hypothetical observations are then defined as the solutions of the slightly modified random equations, in this case the equations random quantities consist in the independently distributed random vectors $(\tilde{X}^0_{1,T_1}, \ldots, \tilde{X}^0_{N,T_1})$ and $(\psi_1, \ldots, \psi_N)$.

We have applied the proposed approach to a data-set reporting the tumoral growths of a subgroup of 61 patients enrolled in a recent multicenter trial; it is therefore implicitly assumed the interest in choosing an adequate parametrization for a randomized discontinuation trial in which the tumoral dynamics under the control regimen are expected to be similar to those in the historical data-set. The adopted utility function is similar to that of the previous example; the parameters $c_1 = 0.002$, $c_2 = 0.002$ and $c_3 = 0.0001$ have a clear interpretation, indeed the optimal choice achieve the equilibrium between the marginal variations of the costs and of the expected benefits. The a priori probability given to the event that the novel agent has a cytostatic activity is 0.7 and the prior on the treatment effects ψ_i is parameterized in such a way to have, conditionally on $\{E = 1\}$, the mean growth of the tumor mass after 4 months from the baseline measurement equal to 12% under the treatment regimen versus the 24% under the control regimen. The primary end point is still binary, the disease is considered stable if the tumor mass increase less than 20% during the second trial phase, and the eligibility criteria to participate to the randomized stage are the following: $\delta_1(X_0; T1) = 1.2X_0$, $\delta_2(X_0; T1) = 0.8X_0$. The action space D is constituted by the alternative designs in $\{[N, T_1, T_2] \in (0, 1, \ldots, 300)^3\}$. The optimal choice $d^* = (T_1 = 72, T_2 = 145, N = 221)$ that maximize the expected utility has been computed by mean of the previously outlined Monte Carlo procedure.

The opportunity of tailoring a prior distribution on a historical data set results advantageous and reduces considerably the complexity of the decisional process. The uncertainties that are not subjectively modeled are limited to the treatment effects, while the full elicitation of the a priori knowledge, as it has been underlined in the previous sections, requires a formal representation of the available information inherent a plurality of aspects as the degree of variability of the baseline measurements that will be observed and the variability of the tumoral progressions. The procedure moreover allows the analysis of the optimal design d^* operative characteristics following a predictive approach; indeed it is straightforward to compute for a plurality of possible degree of efficacy of

the putative agent the predictive probability that the optimal design detects the biologic activity of the agent via Monte Carlo method.

4.5 Discussion

Recent applications of the RDD have confirmed that it can be an appealing alternative to the standard two arm design for assessing the cytostatic activity of novel agents. Between the main features of the RDD in literature have been highlighted the patients' propensity to accept the trial (Stadler (2007)) and the ability of the design to focus the investigation on an homogeneous cohort of patients suitable to give empirical evidence of the treatment effectiveness.

In this chapter the criticality of the tuning parameters of the RDD is underlined and the application of the decision-theoretic framework finalized to their optimal choice is illustrated. The decisional procedure is discussed assuming that the results of the clinical trial, typically the decision to perform a phase III trial or to consider the novel treatment ineffective, have to be based exclusively on the experimental data. On the other hand it is recognized the necessity to take into account the initial knowledge on the treatment and control regimens to suitably design the trial; it doesn't seem practicable to base the choice of T_1 and T_2 other than on the available information. The Bayesian modeling of the tumoral growth processes is a natural candidate to embed the a priori knowledge in the decisional process.

A relevant advantage of the application of the decisional-theoretic paradigm is that it allows to compare the RDD with the two arm design taking into account the differences in the necessary resources to implement the designs and the patients perception of the trials. The underlined difficulty posed by the decisional process of an adequate elicitation of the experts knowledge, in the hypothesis that the decision maker cannot be supported by historical data, follows from the need of a probability model able to capture a plurality of relevant aspects as the connection between the tumor mass and its growth rate, the degree of predictability

of the future tumoral trajectory once it has bee observed for a time interval, the degree of heterogeneity among the patients population and the strength of the experts knowledge about such factors. Even if the simplification of the probabilistic structure at the base of the decisional framework can soften this difficulty a more definitive solution consists in the final validation of the optimal strategy evaluating the optimal design frequentist operating characteristics.

Bibliography

Ratain, M. et al. (2006) Phase II Placebo-Controlled Randomized Discontinuation Trial of Sorafenib in Patients With Metastatic Renal Cell Carcinoma *Journal of Clinical Oncology*, Vol 24, No. 16, 2505-2512.

Rosner GL, Stadler W, Ratain MJ (2002) Randomized discontinuation design: Application to cytostatic antineoplastic agents. *Journal of Clinical Oncology* Vol. 20, No. 22 ,4478-4484

Millar, A. W. Lynch, K. P. (2003) Rethinking clinical trials for cytostatic drugs. *Nature Reviews Cancer*, Vol. 3; No. 7, 540-545

Fedorov, F., Liu, T.(2005) Randomized Discontinuation Trials: Design and Efficiency. *Glaxo Smith Kline Technical Report* .

DeGroot,M.(2004) *Optimal Statistical Decisions.* New York:Wiley-Interscience.

Korn,E.L. et al. (2001) Clinical Trial Designs for Cytostatic Agents: Are New Approaches Needed? *Journal of Clinical Oncology* Vol. 19, No. 1, 265-272

Nestorov, I. et al.(2001) Modeling and Simulation for Clinical Trial Design Involving a Categorical Response: A Phase II Case Study with Naratriptan. *Pharmaceutical Research* Vol.18, No.8, 1210-1219

Ferrante,L.,Bompadre,S.,Possati,L. , Leone, L.(2000) Parameter estimation in a gompertzian stochastic model for tumor growth. *Biometrics*

Gutirrez,R., Gutirrez-Sncheza,R., Nafidia,A., Ramosa,E.(2006)A new stochastic Gompertz diffusion process with threshold parameter: Computational aspects and applications *Applied Mathematics and Computation*, 183, 738-747

Ribbaa, B., Sautb, O., Colinb ,T., Breschc, D., Grenierd, E., Boissela,J.P. (2006) A multiscale mathematical model of avascular tumor growth to investigate the therapeutic benefit of anti-invasive agents *Journal of Theoretical Biology*, Vol. 243, No.4, 532-541

Schnatter,S., Kaufmann,S.(2008) Model-Based Clustering of Multiple Time Series *Journal of Business & Economic Statistics*, Vol. 26, No.1, 78-89

Levy, H.(1973)Stochastic dominance among Log-Normal prospects. *International Economic Review*, Vol. 14, No. 3, 601-614

Padgett, W.,Wei,L.J. (1977) Bayes estimation of reliability for the two-parameter lognormal distribution. *Communications in Statistics*, Vol. 6,No. 5, 443 - 457

Muller,P., Parmigiani,G. (1995) Optimal Design Via Curve Fitting of Monte Carlo Experiments. *Journal of the American Statistical Association*, Vol. 90, No 432, 1322-1330

Stadler,W.M. (2007) The randomized discontinuation trial: a phase II design to assess growth-inhibitory agents. *Molecular Cancer Therapeutics*, Vol 6, 1180-1185

Garthwaite,P.H., Kadane,J.B., O'Hagan,A. (2005) Statistical Methods for Eliciting Probability Distributions. *Journal of the American Statistical Association*, Vol.100, No 470, 680-701

Stadler,W.M., Rosner, G.,Small, E. et al (2005) Successful Implementation of the Randomized Discontinuation Trial Design: An Application to the Study of the Putative Antiangiogenic Agent Carboxyaminoimidazole in Renal Cell Carcinoma −CALGB 69901 *Journal of Clinical Oncology*, Vol.23, No 16, 3726-3732

Bursham,K.P., Anderson,D.R.(1998) *Model Selection and Inference.* Springer-Verlag, New York.