Politecnico di Milano

Faculty of Systems Engineering
Master of Science in Mathematical Engineering
Department of Mathematics

Bayesian nonparametric multi-response models for clustering of patients in presence of unbalanced in-hospital survival

Advisor: Alessandra Guglielmi
Co-advisor: Fernando A. Quintana

Graduation Thesis of:
Valerio Valvo
Student ID: 771222

Academic Year 2012-2013

Contents

List of Figures

List of Tables

Abstract

In questo lavoro verrà presentato un modello bayesiano non parametrico per l'analisi di dati di sopravvivenza di pazienti colpiti da infarto miocardico acuto con sopraslivellamento del tratto ST e sottoposti ad angioplastica in ospedali della regione Lombardia. I dati sono stati resi disponibili tramite l'Archivio STEMI, un registro di osservazione clinica creato all'interno del Programma Strategico della Regione Lombardia nel 2010, che raccoglie le informazioni cliniche riguardanti i pazienti colpiti da infarto e curati negli ospedali lombardi. Il registro è stato istituito per poter valutare i tempi e l'efficacia delle cure e successivamente selezionare strategie ottimali per la terapia dell'infarto miocardico acuto.

Il modello presentato è multivariato, in quanto la risposta è un vettore a tre componenti, corrispondenti alle tre variabili considerate di interesse in termini di efficienza degli ospedali e di efficacia del percorso di cura: il tempo DB, cioè il tempo intercorso tra l'ingresso in ospedale e l'angioplastica (risposta continua), la sopravvivenza alla dimissione e la sopravvivenza dopo 60 giorni dall'ingresso in ospedale, entrambe risposte binarie.

Lo scopo di questa tesi è stata la costruzione di un algoritmo di tipo MCMC per il calcolo delle inferenze a posteriori. In particolare abbiamo calcolato la distribuzione finale dei parametri del modello, compresa la partizione aleatoria ρ dei pazienti, utilizzando una prior PPMx su ρ. Di conseguenza abbiamo stimato i parametri stessi con le statistiche riassuntive della distribuzione finale. Inoltre, visto che uno degli obiettivi era quello di identificare gruppi di pazienti in qualche modo accomunati da caratteristiche simili, abbiamo fornito una stima della partizione aleatoria, che costituisce il raggruppamento dei pazienti. Infine, abbiamo anche calcolato la distribuzione predittiva di pazienti con determinate covariate.

Abstract

In this work a nonparametric Bayesian model is fitted to study data related to patients admitted to hospitals in Lombardy with ST-elevation myocardial infarction diagnosis and treated with angioplasty. Data are collected in the STEMI dataset, an observational clinical study planned within the Strategic Program of Regione Lombardia 2010, which collect clinical informations on patients with myocardial infarction diagnosis and treated in a hospital situated in Lombardy. The aim of the registry is to evaluate, with statistical analysis of this data, times and effectiveness of treatment.

The introduced model is multivariate where the response is a vector of three components, which are three variables considered important to evaluate the efficiency of the hospitals and the effectiveness of the treatment: Door-to-Balloon time, that is the time between the admission to the hospital and angioplasty (continuous response), in-hospital survival and survival after 60 days from admission, both binary responses.

The aim of this work has been to construct an MCMC algorithm for the computation of posterior inferences. In particular we have computed the final distribution of the parameters of the model, including the random partition ρ of the patients using a PPMx prior on ρ. Consequently, we have estimated the same parameters with the summary statistics of the final distribution. Also, since one of the objectives was to identify groups of patients somehow characterized by similar features, we have provided an estimate of the random partition, which is the clustering of patients. Finally, we have also calculated the predictive distribution of patients with given covariates.

Introduzione

In questo lavoro viene presentato un modello bayesiano non parametrico per l'analisi di dati relativi a pazienti colpiti da infarto miocardico acuto con sopraslivellamento del tratto ST e sottoposti ad angioplastica in ospedali presenti nella regione Lombardia. I dati provengono dall'Archivio STEMI, un registro di osservazione clinica creato all'interno del Programma Strategico della Regione Lombardia nel 2010, istituito per poter valutare i tempi e l'efficacia delle cure e, successivamente, selezionare strategie ottimali per la terapia dell'infarto miocardico acuto. Si dispone di dati anagrafici (età, sesso), clinici (presenza di patologie come l'ipertensione, il diabete, ecc.) e riguardanti l'ospedalizzazione (tempo di insorgenza dei sintomi, mezzo di trasporto utilizzato per raggiungere l'ospedale, tempo del primo elettrocardiogramma, ecc.).

Il modello considerato è caratterizzato da una risposta tridimensionale di tipo mista le cui componenti corrispondono agli *outcome* ritenuti più importanti da un punto di vista medico: il tempo DB intercorso tra l'ingresso in ospedale e l'angioplastica, la sopravvivenza alla dimissione e quella a 60 giorni dall'ingresso in ospedale. Si noti che il tempo DB è una variabile continua mentre le altre due risposte sono binarie.

La struttura della verosimiglianza è della seguente forma: la prima componente è rappresentata da un modello lineare che lega il valore atteso del logaritmo del tempo DB a variabili riguardanti l'ospedalizzazione; la seconda e la terza sono rappresentate tramite un modello lineare generalizzato nel quale le probabilità di sopravvivenza (rispettivamente alla dimissione e a 60 giorni) sono messe in relazione con covariate legate alla situazione clinica e allo stato di salute del paziente.

Il modello prevede inoltre una prior sulla partizione aleatoria delle unità statistiche considerate. In particolare, il numero di clusters k è esso stesso sconosciuto e la presenza della prior sulla partizione implica l'esistenza di una prior su tale numero di clusters. Uno dei modelli presenti in letteratura che assegna una prior alla partizione aleatoria è il Product Partition Model (PPM). Tale prior viene costruita introducendo una funzione di coesione $c(\cdot)$ in grado di misurare quanto siano ben raggruppati tra loro gli elementi di un dato insieme. Inoltre, se la funzione di coesione ha una certa espressione, il PPM coincide esattamente con la prior indotta da un campione i.i.d. da un processo di Dirichlet (DP). Tuttavia, la prior sulla partizione aleatoria nel modello considerato in questa tesi dipenderà anche dalle covariate: a partire dal PPM, viene introdotta un'opportuna funzione $g(\cdot)$, detta funzione di similarità, in grado di formalizzare la similarità tra le covariate. A grandi valori di $g(\cdot)$ corrispondono insiemi di covariate giudicati essere simili. Questa prior, introdotta in Muller, Quintana, and Rosner (2011), costituiscee una prior PPM con covariate e dunque verrà indicata con PPMx. Si distingueranno dunque covariate presenti nella verosimiglianza e covariate presenti nella similarità.

Data la struttura di *clustering* definita, ciascuna delle tre variabili risposta dipenderà, all'interno di ogni cluster, da certi parametri, detti *parametri specifici del cluster*. Per assegnare al vettore dei parametri una prior scambiabile, in modo tale che i di-

versi parametri siano a priori dipendenti e quindi possano scambiarsi informazioni, tale distribuzione viene assegnata come mistura rispetto ad altri parametri, detti *parametri globali*.

L'inferenza bayesiana si basa sulla distribuzione finale, cioè la legge condizionale dei "parametri" (partizione aleatoria, parametri specifici del cluster e parametri comuni). Per il calcolo di tale inferenza è necessario costruire un algoritmo di Markov Chain Monte Carlo (MCMC) che simula una catena markoviana aperiodica e irriducibile, la cui distribuzione limite è la posterior del modello considerato. In particolare, è stato implementato in `C` un algoritmo di tipo *Gibbs sampler* basato sul campionamento di ciascuno dei parametri di interesse θ_i a partire dalle loro distribuzioni *full conditionals*, cioè dalle distribuzioni a posteriori condizionali $\mathcal{L}(\theta_i|\theta_{-i})$ dati tutti gli altri parametri e le osservazioni. Per l'analisi degli output forniti da `C` è stato utilizzato il software `R` (R Development Core Team, 2012).

Ad ogni iterazione del Gibbs sampler viene generata una partizione dei dati, dove l'intera successione delle partizioni simulate rappresenta, al limite, la legge a posteriori della partizione aleatoria. Nella tesi abbiamo utilizzato una stima di tale distribuzione finale per rappresentare il clustering dei pazienti cioè per scoprire gruppi significativi di pazienti che in qualche modo sono accomunati da caratteristiche simili. Infine, attraverso la legge predittiva, abbiamo fatto previsione per nuove unità. A tal fine, in questo lavoro sono stati considerati sia i pazienti presenti nel dataset considerato, sia pazienti non presenti nel dataset ma con una combinazione di covariate di interesse.

Nel Capitolo 1, dopo una breve introduzione all'approccio bayesiano nonparametrico, vengono presentate le principali proprietà del processo di Dirichlet. Successivamente, dopo aver richiamato brevemente la struttura del PPM, si procede con la descrizione del PPMx; vengono forniti dettagli sulla forma generica della funzione di similarità e vengono introdotti i parametri specifici del cluster unitamente ai parametri globali con il fine di completare il modello.

Il Capitolo 2 presenta il problema medico riguardante la patologia in esame e descrive brevemente l'Archivio STEMI. Segue la descrizione del dataset originale e un'analisi statistica preliminare. A partire da tale dataset, vengono poi attuate delle modifiche (standardizzazione o dicotomizzazione) su alcune variabili in modo da ottenere il dataset effettivamente analizzato col modello bayesiano. Abbiamo quindi effettuato un'ulteriore analisi statistica preliminare sul nuovo dataset.

Nel Capitolo 3 viene definito il modello adottato. In primo luogo si definisce la legge generica dell'i-esimo vettore risposta all'interno del j-esimo cluster, per poi in seguito specificare le prior adottate sui parametri specifici del cluster e sui parametri globali. In secondo luogo si procede alla definizione dettagliata della prior adottata sulla partizione, esplicitando dunque la forma della funzione di similarità e tenendo conto delle covariate utilizzate per definire tale prior (in questo lavoro di tipo binario o categorico). Segue la descrizione dell'algoritmo utilizzato per implementare il Gibbs sampler, tenendo conto dell'aggiornamento sia dei parametri specifici del cluster, sia dei parametri comuni.

Nel Capitolo 4 vengono presentati i risultati ottenuti applicando il modello PPMx all'Archivio STEMI. Vengono fornite le stime a posteriori dei parametri globali con relativi intervalli di credibilità. Particolare enfasi viene data alla previsione e al clustering. Infine, viene fatto un confronto con il modello PPM. Dall'analisi effettuata è emerso che all'aumentare del tempo in cui viene effettuato il primo elettrocardiogramma si ha un aumento del tempo DB. Inoltre, variabili quali l'età, la gravità dell'infarto, l'efficienza del

trattamento o la frazione di eiezione (così come l'essere operati in un ospedale a Milano oppure no) hanno un impatto significativo sulla sopravvivenza del paziente.

L'Appendice A richiama brevemente il modello lineare bayesiano con varianza incognita. Tale modello servirà infatti nell'aggiornamento dei parametri specifici del cluster per il primo livello. L'Appendice B riporta le espressioni analitiche delle full-conditionals dei parametri globali per tutti e tre i livelli, distinguendo tra il primo livello e gli altri due. L'Appendice C riporta le espressioni analitiche della media e della varianza a priori dei parametri specifici del cluster.

Questo lavoro costituisce un primo tentativo di costruzione di un modello bayesiano in cui la prior sulla partizione aleatoria dipende dalle covariate, per l'applicazione al dataset sui pazienti affetti da STEMI. Le covariate che sono state incluse nella distribuzione condizionale dei dati, dati i parametri, sembrano quasi tutte significative. Sviluppi futuri potrebbero prevedere un nuovo modello in cui considerare covariate diverse nella verosimiglianza (e tutte le rimanenti nella funzione di similarità), oppure fornire una diversa stima di clustering sotto il modello qui considerato.

Chapter 1

Bayesian Nonparametrics

1.1 Exchangeability assumption

Classical statistics is based on a framework where observations $X_1, X_2, \ldots$ are assumed independent and identical distributed (i.i.d.) from a unknown probability distribution P. We say that we are considering a parametric framework when P belongs to a parametric family, otherwise we are considering a non-parametric framework when P lies in the space of all probability distributions $\mathcal{P}(\mathbb{R})$.

It is possible to distinguish the two cases also in the Bayesian setting. In the parametric case we have a prior π on a finite dimensional space Θ and, given θ, the observations are assumed i.i.d. from P_θ. In the non-parametric case, we have a prior π on the space $\mathcal{P}(\mathbb{R})$ of all probability distributions on $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$ and, given P, the observations are assumed i.i.d. from P.

Under the assumption of exchangeability, de Finetti's Representation Theorem gives a validation of the Bayesian setting.

Let consider an infinite sequence of observations $(X_n)_{n\geq 1}$ defined on some probability space $(\Omega, \mathcal{F}, \mathbb{P})$, with each X_i taking values on $\mathbb{R}$ endowed with the Borel σ-algebra $\mathcal{B}(\mathbb{R})$. This last hypothesis can be relaxed and we could consider observations which take values in a complete metric and separable space $\mathbb{X}$. In this chapter it is enough to consider $\mathbb{X} = \mathbb{R}$.

Definition 1.1 (Finite exchangeability)**.** *The random quantities $y_1, \ldots, y_n$ are finitely exchangeable if their joint probability density is such that*

$$p(y_1, \ldots, y_n) = p(y_{z(1)}, \ldots, y_{z(n)})$$

for all permutations z of the indices of the y_i, $\{1, \ldots, n\}$.

Roughly speaking, $y_1, \ldots, y_n$ are exchangeable if the subscript labels convey no information about the outcomes. There are several types of dependence among a sequence of observations $(X_n)_{n\geq 1}$. Under the exchangeability assumption, the information that the observations X_is provide is independent of the order in which they are collected. For instance, if we sample without replacement from an urn with infinite marbles of different colors, the sequence of colors that we obtain is exchangeable.

Remark. *An infinite sequence of random quantities $y_1, y_2, \ldots$ is infinitely exchangeable if every finite subsequence is finitely exchangeable.*

Definition 1.1 formalizes a symmetry condition between variables: in many cases, the order in which observations are received does not matter. Permuting the order, we would have the same informations.
To enunciate de Finetti's theorem (see for instance Ghosh and Ramamoorthi, 2003, for the proof) we need to define random probability measures. To do this we give formal definitions of the Borel σ-algebra on $\mathcal{P}(\mathbb{R})$ introducing the topology of weak convergence. The space $\mathcal{P}(\mathbb{R})$ is equipped with the topology of the weak convergence which makes it a complete and separable metric space.

Definition 1.2 (Weakly convergence)**.** *A sequence of probability measures $(P_n)_{n\geq 1}$ defined on $\mathcal{P}(\mathbb{R})$ converges weakly to a probability measure P if for all bounded continuous function $f : \mathcal{P} \to \mathbb{R}$*

$$\int_{\mathbb{R}} f\, dP_n \longrightarrow \int_{\mathbb{R}} f\, dP, \ \text{as } n \longrightarrow +\infty.$$

We write $P_n \rightharpoonup P$.
For any P_0 a neighbourhood base consist of sets of the form

$$\bigcap_{i=1}^{n} \Big\{ P : \Big| \int f_i\, dP_0 - \int f_i\, dP \Big| < \epsilon \Big\}$$

where f_i, $i = 1, \ldots, k$, are bounded continuous function on $\mathbb{R}, k \geq 1$ and $\epsilon > 0$. The Borel σ-algebra on $\mathcal{P}(\mathbb{R})$ is the smallest σ-algebra generated by the open sets in the weak topology.

Definition 1.3 (Random probability measure)**.** *A random element defined on $(\Omega, \mathcal{F}, \mathbb{P})$ with values in $\mathcal{P}(\mathbb{R})$ is called random probability measure (r.p.m).*

The two main desirable properties for the class of r.p.m are a large support and a posterior distribution that is analytically tractable. A prior with a large support is an obvious requirement, and a tractable posterior reduces the computational complexity. In fact computational heaviness is still one limitation of Bayesian nonparametrics. The most popular r.p.m.s in literature are Dirichlet Process, Polya Trees and Bernstein Polynomials.

Theorem 1.1 (de Finetti)**.** *The sequence $(X_n)_{n\geq 1}$ is exchangeable if, and only if, there exists a unique probability measure q on $\mathcal{P}(\mathbb{R})$ such that, for any $n \geq 1$ and any Borel sets $B_1, B2, \ldots, B_n$,*

$$\mathbb{P}(X_1 \in B_1, X_2 \in B_2, \ldots, X_n \in B_n) = \int_{\mathcal{P}(\mathbb{R})} \prod_{i=1}^{n} p(B_i) q(dp).$$

Equivalently, the theorem implies that if $(X_n)_{n\geq 1}$ is exchangeable, then

$$\begin{aligned} X_1, \ldots, X_n | P &\overset{i.i.d}{\sim} P \\ P &\sim q(\cdot). \end{aligned}$$

In the parametric case q is concentrated on a parametric family, i.e.

$$\begin{aligned} X_1, \ldots, X_n | \theta &\overset{i.i.d}{\sim} f_\theta(\cdot) \\ \theta &\sim \pi(\cdot), \end{aligned}$$

where $X_i|\theta$ and θ are absolutely continuous (with respect to the Lebesgue measure) or discrete probability distributions, for $i = 1, \ldots, n$, f_θ and $\pi(\cdot)$ are the probability density functions of $X_i|\theta$ and θ respectively. By Bayes' Theorem, in this case the posterior probability distribution of θ, i.e the conditional probability distribution of θ given $X_1, \ldots, X_n$, has probability density function

$$\pi(\theta|X_1 = x_1, \ldots, X_n = x_n) = \frac{\prod_{i=1}^n f_\theta(x_i)\pi(\theta)}{\int_\Theta \prod_{i=1}^n f_\theta(x_i)\pi(\theta)}.$$

The predictive distribution of a new observation $X_{n+1} = x$ has probability density function

$$f(x|X_1 = x_1, \ldots, X_n = x_n) = \int_\Theta f_\theta(x)\pi(d\theta|X_1 = x_1, \ldots, X_n = x_n).$$

Similarly we can make inference and prediction in the nonparametric setting. In this case q is a probability measure on $\mathcal{P}(\mathbb{R})$ and the posterior distribution can be derived by

$$\mathcal{L}(X_{n+1}|X_1, \ldots, X_n) = \int_{\mathcal{P}(\mathbb{R})} \mathcal{L}(X_{n+1}|P)\mathcal{L}(dP|X_1, \ldots, X_n).$$

1.2 The Dirichlet process

The Dirichlet process is a useful family of prior distributions on $\mathcal{P}(\mathbb{R})$ introduced by Ferguson (1973). The Dirichlet prior is easy to elicit, has a manageable posterior and other nice properties. It can be viewed as an infinite-dimensional generalization of the finite-dimensional Dirichlet distribution.

Definition 1.4 (Dirichlet distribution)**.** *Let $\boldsymbol{\alpha} = (\alpha_1, \alpha_2, \ldots, \alpha_k)$, with $\alpha_i > 0$ for $i = 1, 2, \ldots, k$. The random vector $P = (P_1, P_2, \ldots, P_k)$, $\sum_{i=1}^k P_i = 1$, has Dirichlet distribution with parameter $\boldsymbol{\alpha}$ if $(P_1, P_2, \ldots, P_{k-1})$ is absolutely continuous with respect to the Lebesgue measure on $\mathbb{R}^{k-1}$ with density*

$$f(p_1, p_2, \ldots, p_{k-1}) = \frac{\Gamma(\sum_{i=1}^k \alpha_i)}{\Gamma(\alpha_1)\Gamma(\alpha_2)\cdots\Gamma(\alpha_k)} p_1^{\alpha_1-1} p_2^{\alpha_2-1} \cdots p_{k-1}^{\alpha_{k-1}-1} \Big(1 - \sum_{i=1}^{k-1} p_i\Big)^{\alpha_k-1},$$

where $0 \le p_i \le 1\ \forall i,\ 0 \le p_1 + \cdots + p_{k-1} \le 1,\ 0$ otherwise.

We write $P \sim D(\boldsymbol{\alpha})$.

Definition 1.5 (Dirichlet process)**.** *Let α be a finite measure on $\mathbb{R}$, $a = \alpha(\mathbb{R})$; let $\alpha_0(\cdot) = \alpha(\cdot)/a$. A r.p.m. P with values in $\mathcal{P}(\mathbb{R})$ is a Dirichlet process on $\mathbb{R}$ with parameter α if, for a finite measurable partition $B_1, \ldots, B_k$ of $\mathbb{R}$,*

$$(P(B_1), \ldots, P(B_k)) \sim D(\alpha(B_1), \ldots, \alpha(B_k)).$$

We write $P \sim DP(\alpha)$ for short. It can be proved that such process exist (see Ferguson, 1973). If $P \sim DP(\alpha)$, it follows that $\mathbb{E}[P(A)] = \alpha_0(A)$ for any Borel set A, and thus we say that α_0 is the prior expectation of P.

The Dirichlet prior is a conjugate prior on $\mathcal{P}(\mathbb{R})$; in fact, let $(X_1, X_2, \ldots, X_n)$ be a sample from a Dirichlet process P, i.e.

$$X_1, X_2, \ldots, X_n | P \overset{i.i.d}{\sim} P$$
$$P \sim DP(\alpha).$$

Then the posterior distribution of P given $X_1, X_2, \ldots, X_n$, is

$$P|X_1, X_2, \ldots, X_n \sim DP\Big(\alpha + \sum_{i=1}^{n} \delta_{X_i}\Big).$$

In this case the distribution of X_{n+1} can be described as follows:

$$\begin{aligned} X_1 &\sim \alpha_0 \\ X_{n+1}|X_1, \ldots, X_n &\sim \frac{a}{a+n}\alpha_0 + \frac{n}{a+n}\Big(\frac{\sum_{i=1}^{n} \delta_{X_i}}{n}\Big). \end{aligned} \tag{1.1}$$

Notice that the predictive distribution in (1.1), called *Blackwell-MacQueen Urn Scheme*, is a mixture of the baseline measure α_0 and the empirical distribution on the previous observations. This means that there is a positive probability of coincident values for any finite and positive a. Moreover if α_0 is an absolutely continuous probability measure, then X_{n+1} will assume a different distinct value with probability $\frac{a}{a+n}$. Formula (1.1) allows us to sample (marginally) from P without simulating any trajectory of the Dirichlet process.

Let $(X_1, X_2, \ldots, X_n)$ be a sample from P, where $P \sim DP(\alpha)$ and let K_n denote the random variable representing the number of distinct values among $(X_1, X_2, \ldots, X_n)$. Antoniak (1974) proved that the distribution of K_n is the following:

$$\mathbb{P}(K_n = k) = c_n(k) n! a^k \frac{\Gamma(a)}{\Gamma(a+n)},\ k = 1, 2, \ldots, n, \tag{1.2}$$

where $c_n(k)$ is the absolute value of Stirling number of the first kind, which can be tabulated or computed by a software. From (1.2) it is clear that the mass parameter a influences the prior on the number of clusters. Larger a gives rise to a higher prior number of components.

Sethuraman (1994) provided a useful representation of the Dirichlet process. Its construction gives an insight on the structure of the process and provides an easy way to simulate its trajectories. Let consider two independent sequences of random variables $(\theta_i)_{i\geq 1}$ and $(\tau_i)_{i\geq 1}$ such that $\theta_i \overset{i.i.d}{\sim} Beta(1, a)$ and $\tau_i \overset{i.i.d}{\sim} \alpha_0$ defined on some probability space $(\Omega, \mathcal{F}, \mathbb{P})$, and define the following weights:

$$\begin{cases} p_1 = \theta_1 \\ p_n = \theta_n \prod_{i=1}^{n-1}(1-\theta_i), \quad n \geq 2 \end{cases}$$

It is straightforward to see that $0 \leq p_n \leq 1,\ n = 1, 2, \ldots$ and $\sum_{n=1}^{\infty} p_n = 1$ a.s.. This construction is called *stick-breaking*. In fact p_1 represents a piece of a unit-length stick, p_2 represents a piece of the remainder of the stick and so on, where each piece is independently modelled as a $Beta(1, a)$ random variable scaled down to the length of the remainder of the stick. Now we can define a random variable P on $\mathcal{P}(\mathbb{R})$:

$$P(A) = \sum_{n=1}^{\infty} p_n \delta_{\tau_n}(A),\ A \in \mathcal{B}(\mathbb{R}).$$

Sethuraman proved that P has a Dirichlet prior distribution, i.e. P is a Dirichlet process with parameter α. From this construction it is clear that a Dirichlet process has discrete trajectories, i.e. if $P \sim DP(\alpha)$, $\mathbb{P}(\omega : P(\omega) \text{ is discrete}) = 1$.

As mentioned in Section 1.1, $\mathcal{P}(\mathbb{R})$ is a separable metric space, and hence it is possible to define the support of any probability measure π on $\mathcal{P}(\mathbb{R})$ as the smallest closed set of measure 1. Let E be the support of the finite measure α on $\mathbb{R}$. Then it can be shown that $M_\alpha = \{P : \text{ support of } P \subset E\}$ is the weak support of $DP(\alpha)$, i.e. the set of all the probability distributions with support contained in the support of the measure α is the weak support of $DP(\alpha)$ (see Ferguson, 1973).

1.3 The PPMx Model

Recall that if $(X_1, X_2, \ldots, X_n)$ is a sample from a Dirichlet process P, then, by the expression of the posterior distribution (1.1), we deduce that there is a positive probability of having ties in the $\{X_n\}$ sequence for any finite and positive a. This means that this model induce a probability distribution on the space of partitions of a finite set of objects $S = \{1, \ldots, n\}$, say.

As described in Muller, Quintana, and Rosner (2011), we want to develop a probability model for partitioning a set of experimental units, where the probability of any particular partition is allowed to depend on covariates.
Let $i = 1, \ldots, n$ index experimental units and let $\rho_n = \{S_1, \ldots, S_k\}$ denote a partition of the n experimental units into k subsets S_j. Let x_i and y_i denote the covariates and response reported for the ith unit. Let $\boldsymbol{x} = (x_1, \ldots, x_n)$ and $\boldsymbol{y} = (y_1, \ldots, y_n)$ denote the entire set of recorded covariates and response data and let $x_j^* = (x_i,\ i \in S_j)$ and $y_j^* = (y_i,\ i \in S_j)$ denote covariates and response data arranged by clusters. It is convenient to introduce cluster membership indicators $e_i \in \{1, \ldots, k\}$ with $e_i = j$ if $i \in S_j$, and use $(k, e_1, \ldots, e_n)$ to describe the partition. We call probability model $p(\rho_n)$ a clustering model. In particular, the number of cluster k is itself unknown. The clustering model $p(\rho_n)$ implies a prior model $p(k_n)$. Many clustering models include a sampling model $p(\boldsymbol{y}|\boldsymbol{x}, \rho_n)$. We are interested in adding a regression to replace $p(\rho_n)$ with $p(\rho_n|\boldsymbol{x})$.

The PPM (Hartigan, 1990) constructs $p(\rho_n)$ by introducing cohesion functions $c(A) \geq 0$ for $A \subseteq \{1, \ldots, n\}$ that measure how tightly grouped the elements in A are thought to be, and defines a probability model for a partition ρ_n and data $\boldsymbol{y}$ as

$$p(\rho_n) \propto \prod_{j=1}^{k} c(S_j) \qquad \text{and} \qquad p(\boldsymbol{y}|\rho_n) = \prod_{j=1}^{k} p_j\left(y_j^*\right). \tag{1.3}$$

Model (1.3) is conjugate: the posterior $p(\rho_n|\boldsymbol{y})$ is again in the same product form.

The Dirichlet Process (DP) model is related with the PPM. In fact, the marginal distribution that a DP induces on partition is also a PPM with cohesions $c(S_j) = \alpha(|S_j| - 1)!$ (Quintana and Iglesias, 2003). Here α denotes the total mass parameter of the DP prior.

The model proposed in Muller, Quintana, and Rosner (2011) is built from the PPM (1.3) modifying the cohesion function $c(S_j)$ with an additional factor that achieves the desired regression. Let $g\left(x_j^*\right)$ denote a nonnegative function of x_j^* that formalize similarity of the x_i with larger values of $g\left(x_j^*\right)$ for sets of covariates that are judged to be similar. The model is defined as

$$p(\rho_n|\boldsymbol{x}) \propto \prod_{j=1}^{k} g\left(x_j^*\right) c\left(S_j\right) \tag{1.4}$$

with the normalization constant $g_n(\boldsymbol{x}) = \sum_{\rho_n} \prod_{j=1}^{k_n} g\left(x_j^*\right) c\left(S_j\right)$. As a default choice we propose to define $g(\cdot)$ as the marginal probability in an auxiliary probability model q, even if x_i are not considered random,

$$g\left(x_j^*\right) = \int \prod_{i \in S_j} q\left(x_i|\xi_j\right) q\left(\xi_j\right) d\xi_j. \tag{1.5}$$

There is no notion of the x_i being random variables. But the use of a probability density $q(\cdot)$ for the construction of $g(\cdot)$ is convenient since it allows for easy calculus. The correlation that is induced by the cluster specific ξ_j in (1.5) leads to higher values of $g\left(x_j^*\right)$ for tightly clustered, similar x_i, as desired. Furthermore, under some minimal assumptions a similarity function $g(\cdot)$ necessarily is of the form (1.5). The function (1.5) satisfies the following two properties that are desirable for a similarity function in (1.4). First, symmetry with respect to permutations of the sample indices j is required. The probability model must not depend on the order of introducing the experimental units. This implies that the similarity function $g(\cdot)$ must be symmetric in its arguments. Second, we require that the similarity function scales across sample size, in the sense that $g(x^*) = \int g(x^*, x)\, dx$. In words, the similarity of any cluster is the average of any augmented cluster.
Under these two requirements (1.5) is the only possible similarity function that satisfies these two constraints (see Muller, Quintana, and Rosner, 2011).

The definition of the similarity function with the auxiliary model $q(\cdot)$ also implies another important property. The random partition model (1.4) is coherent across sample sizes. The model for the first n experimental units follows from the model for $(n+1)$ observations by appropriate marginalization.

The random partition model (1.4) is completed with a sampling model that defines independence across cluster and exchangeability within each cluster. We include cluster specific parameters $\boldsymbol{\theta}_j$ and common hyperparameters $\boldsymbol{\eta}$:

$$p(\boldsymbol{y}|\rho_n, \boldsymbol{\theta}, \boldsymbol{\eta}, \boldsymbol{x}) = \prod_{j=1}^{k} \prod_{i \in S_j} p\left(y_i|x_i, \boldsymbol{\theta}_j, \boldsymbol{\eta}\right) \qquad \text{and} \qquad p(\boldsymbol{\theta}|\boldsymbol{\eta}) = \prod_{j=1}^{k} p\left(\boldsymbol{\theta}_j|\boldsymbol{\eta}\right), \tag{1.6}$$

where $\boldsymbol{\theta} = (\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_k)$. We refer to (1.4) together with (1.6) as a PPM with covariates, and write PPMx for short, as introduced in Muller, Quintana, and Rosner (2011).

Chapter 2

Dataset and medical issue description

2.1 The acute myocardial infarction

As described in Ieva (2013) and Prandoni (2012), the Acute Myocardial Infarction (AMI), one of the Coronary Artery Diseases (CAD), is a pathological condition in which a prolonged ischemia, i.e. a reduced perfusion of the heart muscle leading to a decreased supply of oxygen, causes the death of the heart cells. The most frequent cause of myocardial infarction is the stenosis (constriction) of one or more coronary arteries caused by the occlusion of the blood vessel by a thrombus. Symptoms of this pathology are sudden chest pain, shortness of breath, sweating, nausea, vomit. Via the electrocardiogram (ECG) printout, AMI can be classified into two types: AMI with elevation of the ST section (STEMI) and AMI without elevation of the ST section (see Rugarli, 2010, Chapter 5). The former represents the typical expression of an infarction. Unlike the infarction without elevation of the ST section, STEMI is caused by a serious and prolonged failure of oxygen in the whole depth of a region of the myocardium. In physiological conditions, ECG presents a characteristic pattern of positive and negative waves that is repeated each cardiac cycle. ST segment of the ECG is the section that separates wave S and wave T. It is placed on the base of the trace (Figure 2.1) and it is characterized by the absence of electrical movements. In the case of STEMI this line raises with respect to the physiological level.

AMI diagnosis is mainly based, besides on symptoms (intense and prolonged chest pain with radiation to the left arm), on ECG analysis and on the evaluation of specific markers of myocardial necrosis. The reopening of the occluded coronary vessel could block the necrosis process and preserve at least one part of the myocardium from cellular death, improving the prognosis of the patient. Moreover, the therapeutic intervention of revascularization must be done as early as possible after symptoms. Indeed, therapeutic efficacy is much greater if the reperfusion therapy is implemented within the next two hours after the infarction, it is pretty large within the first 6 and 12 hours and it is lower after 12 hours. For this reason it is very important that patients who suspect to have an infarction go to the hospital as soon as possible, so that the electrocardiographic diagnosis of AMI and the reperfusion therapy are implemented as quickly as possible. Currently, coronary reperfusion can be obtained with a pharmacological or mechanical treatment. The first treatment, called thrombolysis, is based on the assumption of drugs that produce the lysis of the fibrin ties of the thrombus, disintegrating it. The second one is called angioplasty (PTCA - Percutaneous Transluminal Coronary Angioplasty). It is a surgical technique that dilates the stenotic artery by means of a inflatable *balloon*. PTCA

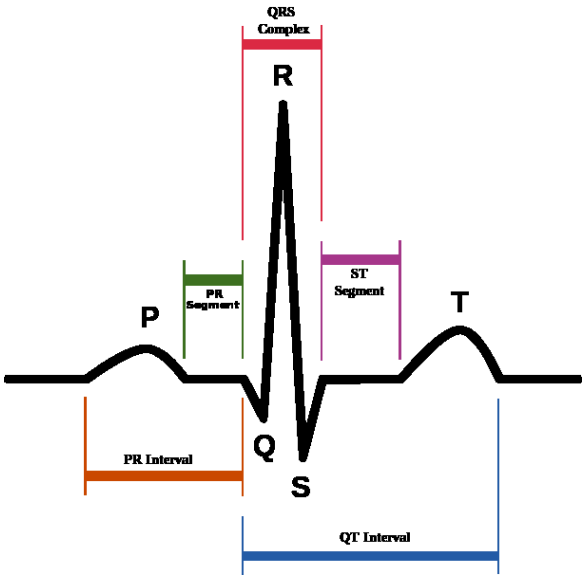


Figure 2.1: Outline of the characteristics wave of the electrocardiogram trace.

very often is associated to the installation of an expandable little tube which constitutes a sort of scaffold capable to maintain dilated the artery and which remain permanently in the vessel. In this way, the cells of the arterial wall grow around the device, fixing it even more. PTCA constitutes nowadays the treatment of choice to restore the coronary reperfusion. It presents some advantages with respect to the thromobolysis: it allows the recanalization in almost all cases, regardless of the elapsed time of onset infarction. Thrombolysis instead achieves this result by a maximum of 75% of the cases and the success probability is reduced with the passage of time. Moreover in PTCA there is no risk of intracranial hemorrhage, while for thrombolysis this risk is 0.5-1%. Disadvantages of such procedure with respect to thrombolysis are that it can be implemented only in centers provided of hemodynamic with expert operators, and it requires a proper organization of the medical and nursing staff to ensure 24 hours availability; finally it requires longer time to be carried out.

2.1.1 The STEMI dataset

Data that will be used come from the STEMI dataset (Ieva, 2013), a register of clinical observation created within the Strategic Program of the Lombardy Region in 2010, which collects clinical informations about patients affected by AMI and treated in hospitals in Lombardy. The purpose is to evaluate timing and effectiveness of the treatment in order to select an optimal path for AMI therapy. In particular, statistical analysis is useful to provide an efficiency framework of the treatment protocols and to find solutions to enhance the quality of the offered services.

Data can be divided into four categories:

- Demographic data: fiscal code, date of birth, gender, age, hospital where the patient was treated.
- Pre-hospital data: presence of diseases like diabetes, high blood pressure, hypercholesterolemia, cardiac pathologies.

- Admission data: time of onset of symptoms and typology, time of first contact of the patient with the medical staff, type of rescue vehicle sent, time of the first ECG, Killip class (i.e. the severity of the infarction), blood pressure and heart rate.

- Therapeutic data: elapsed time between the entrance into hospital and thrombolysis or angioplasty treatment, ejection fraction.

More details are given in the next Section.

In this work we will consider as the outcome of interest the time between the entrance into hospital and angioplasty, the in-hospital survival and the survival after 60 days. The first term is an important indicator of the hospitals efficiency and it is crucial in the success of therapy; the second term is the outcome that indicates the therapy success or failure; the last term is the most significant indicator because treatment efficacy, in terms of survival and quality of life, is evaluated on the mid-term.

2.2 Variables

The dataset contains data on 697 patients undergoing angioplasty after an episode of acute myocardial infarction in 33 hospitals in Lombardia. The response variables of interest are:

- DB (*door balloon time*): it is the time between hospital admission and the arrival time on the operating table where angioplasty can be performed.

- ALIVEIN: survival to discharge from hospital. It is equal to 1 if the patient is alive, 0 otherwise.

- ALIVE60: survival to 60 days from entrance into hospital. It is equal to 1 if the patient is alive, 0 otherwise.

Available covariates are continuous, binary and categorical. The continuous ones are:

- AGE: patient age in years.

- EF: ejection fraction at the entrance into hospital, i.e. volume of blood that the heart ejects from the left ventricle within each heartbeat with respect to the physiological condition.

- ECG: elapsed time, in minutes, between the entrance into hospital and the first electrocardiogram. It can assumes negative values since for some patients it has been possible to carry out the electrocardiogram during the transport in the ambulance thanks to TeleECG.

Binary covariates are:

- GENDER: 1 if the patient is a male, 0 if female.

- STres: it is the efficacy of the treatment, quantified by the reduction (at least 70%) of the gap of the ST section of the electrocardiogram within one hour. It is equal to 1 if there was no efficacy, 0 otherwise.

- COMPLICATION: it assumes the value 1 if, after the operation, there were complications, 0 otherwise.
- WEEKEND: it is equal to 1 if the patient was admitted to the hospital in a non-working day (saturday, sunday or festivity) or between 6:00 pm and 8:00 am, 0 otherwise.
- MILAN: it denotes the geographical location of the hospital, it is equal to 1 if the hospital is in Milan, 0 otherwise.

Other available binary variables are risk factors as DIABETES, SMOKING, HYPERTENSION, CHOLESTEROL, VASCULOPATIA, CKD (Chronic Kidney Disease), preAMI (previous infarction): they assume the value 1 if the patient has them, 0 otherwise.

Finally, the categorical covariates are:

- KILLIP: it is the severity of the infarction. It assumes values from 1 to 4, with increasing severity of the infarction
- AMBULANCE TYPE (MA): it is the modality of access into the hospital. It assumes the value 1 if an advanced ambulance (MSA), i.e. equipped of TeleECG capable of performing the electrocardiogram and send it to the doctor in the operation center, has been used. The value 2 indicates that it has been used a basic ambulance (MSB), i.e. the ambulance without TeleECG, whereas 3 indicates a nursing ambulance (MSI): there is nursing stuff on the ambulance but no doctors and it is not possible to send the electrocardiogram. Finally, the value 4 means that the patient has reached the hospital by his own.
- HOSPITAL: it is an integer index, from 1 to 33, denoting the hospital the patient was admitted to and treated.

2.3 Preliminary analysis of the dataset

The original dataset is larger than the one we have considered for our Bayesian analysis. Some patients from the original dataset were removed. In particular hospitals with a number of patient lower than 5 were not considered. Moreover, 3 patients with a time of discharge greater than 60 days (for whom, then, the variable survival to discharge no longer makes sense) were removed and also 60 patients whose the variable DB is not available, so that the number of patient is 1201. Finally, patients with missing values in the covariates used in the likelihood were also removed.
On the whole, the final dataset contains 697 patients and 33 hospitals: of these, 12 (36.36%) are in Milan (44.48% of treated patients) and 21 (63.64%) outside Milan (55.52% of treated patients). Frequencies of patients in each hospital are shown in Figure 2.2.

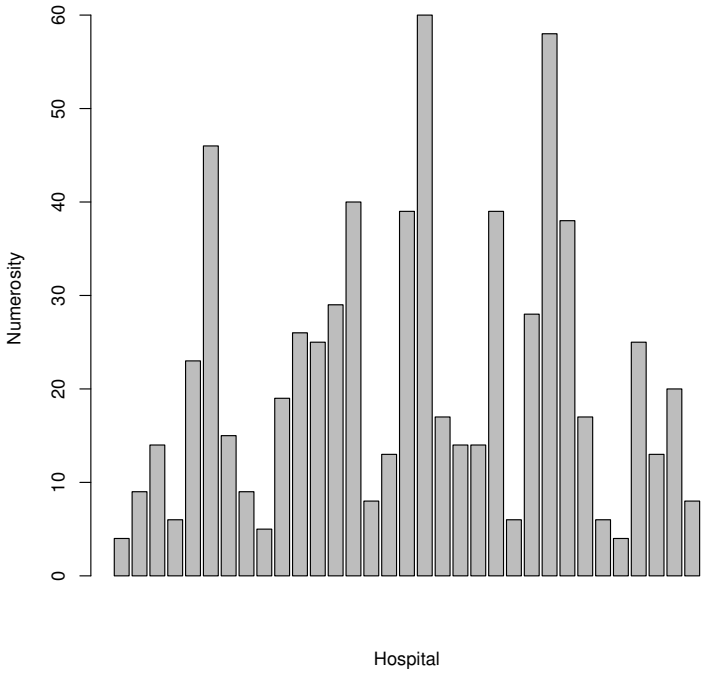


Figure 2.2: Number of patients for hospitals.

The final dataset still contains missing data (NA). The following table shows the percentages of missing data for each variable that presents them.

Table 2.1: Percentages of missing data.

Variable	% NA
DIABETES	0.29
SMOKING	1.00
CHOLESTEROL	11.48
VASCULOPATIA	1.00

The dataset is highly unbalanced, since patients discharged alive from the hospital are 675 (96.8%), while the ones alive at 60 days are 664 (that is, the 98.4% of patients alive at discharge). Table 2.2 shows the percentages of survived patients, distinguishing between those treated in Milan and outside Milan.

Table 2.2: Percentage of survived patients.

(a) discharged alive from hospital

Milan	**outside Milan**
44.15%	55.85%

(b) alive at 60 days

Milan	**outside Milan**
44.13%	55.87%

The age varies from a minimum of 34 years to a maximum of 99 years (the histogram is in figure 2.3). Men are the 77.62% of patients and women 22.38%.

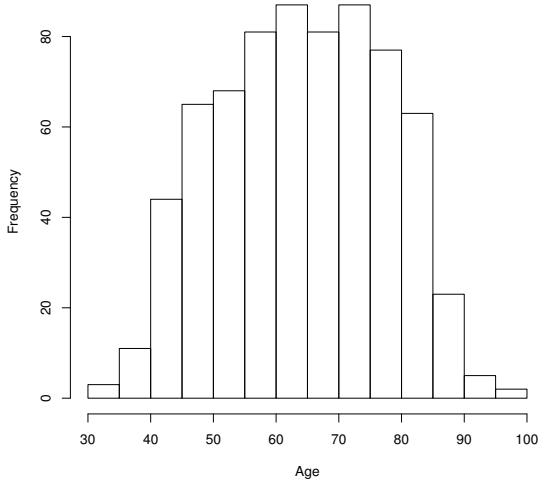


Figure 2.3: Histogram of the ages.

Figure 2.4 shows the boxplots of patients age divided by gender: women are older than men.

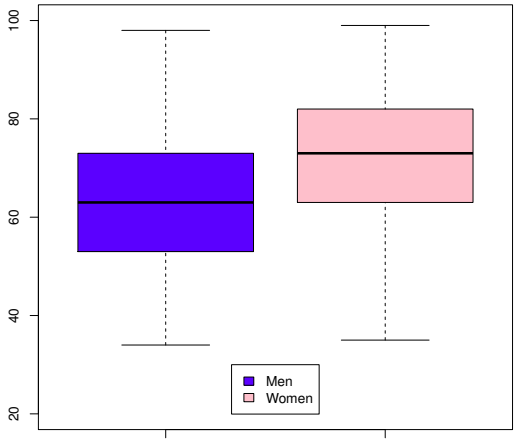


Figure 2.4: Boxplot of the ages with respect to gender.

Previous analysis on the original dataset (Prandoni, 2012) or on a similar dataset (see Guglielmi et al., 2012) have shown that the gender is not a significant covariate, being strongly correlated with age in this type of data.
To understand dependence between gender and age, the Fisher independence test was performed, discretizing data according to Table 2.3: the p-value was $1.05 \cdot 10^{-7}$, showing strong evidence against the null hypothesis of independence.

Table 2.3: Contingency table, age vs gender.

	Men	Women
age≤65	308	51
age>65	233	105

For the variable MA and KILLIP Table 2.4 shows the distribution of patients in the respective categories:

Table 2.4: Percentage of patient respect to ambulance type and killip.

Modality	%
MSA	19.66
MSB	18.36
MSI	2.30
Spontaneous	59.68

Killip	%
1	84.22
2	9.76
3	3.01
4	3.01

As far as the only one continuous response variable, the DB time, is concerned, Figure 2.5 and Figure 2.6 show the boxplots stratified by modality of access (to see if there are some differences depending on the ambulance type used to arrive to the hospital) and by arrival time (working or no working day), respectively:

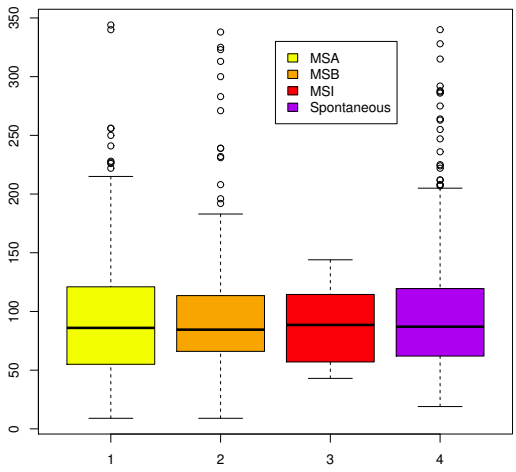


Figure 2.5: Boxplot of DB time respect to the modality of access.

In particular, Figure 2.5 shows no significant difference between DB times related to the different modality of access. We have also performed the nonparametric Kruskal-Wallis test: this is useful when normality assumption can not be assumed, like in this case, and the null hypothesis is that group means are equal. The p-value of the test is equal to 0.5336, showing that there is no evidence against the hypothesis of equality

of all the means. Also comparing the day and the arrival time (Figure 2.6) DB time distributions seem very similar (Kruskal-Wallis test provided a p-value of 0.3907).

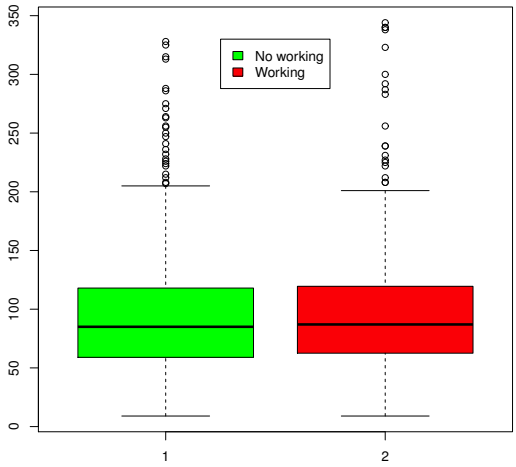


Figure 2.6: Boxplot of DB time with respect to the arrival time: working hours from 8:00 am to 6:00 pm, from Monday to Friday.

2.4 A primary frequentist covariates selection

An analysis for the selection of significant covariates for the responses was performed in Prandoni (2012); we will use it as preliminary analysis, considering covariates that are been evidenced as significant in that work. Moreover, we use now the frequentist analysis of the corresponding univariate models like benchmark for our future choices.
We considered three different univariate regression models, one for every response variable: *door to balloon time* (DB), survival to discharge (ALIVEIN) and survival to 60 days (ALIVE60). For every model a linear regression will be performed, using frequentist statistical techniques.

The first model relates the response variable DB to the covariates ECG, WEEKEND and MA, considered significant thanks to the analysis in Prandoni (2012). For every patient $i = 1, \ldots, n$ it is assumed that

$$log(DB)_i = \beta_0 + \beta_1 ECG_i + \beta_2 WEEKEND_i + \beta_3 MA_i + \epsilon_i, \quad \epsilon_i \overset{i.i.d}{\sim} \mathcal{N}(0, \tau^{-1}),$$

where $\mathcal{N}(0, \tau^{-1})$ denotes the normal distribution with mean 0 and precision τ.
The analysis was carried out using the `lm` function of the MASS package of R for the frequentist approach. Variable MA, which is categorical, has been transformed in a 3 dimensional dummy vector, considering the class "Spontaneo" as reference, so the intercept represents the mean of the logarithmic of DB for those patients who have reached the hospital by their own.

```
Call:
lm(formula = log(DB) ~ ECG + WEEKEND + MSA + MSB + MSI)

Residuals:
     Min       1Q   Median       3Q      Max
-2.20390 -0.33703 -0.02079  0.30915  1.47673

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 4.302971   0.049247  87.376  < 2e-16 ***
ECG         0.007508   0.001116   6.730 3.57e-11 ***
WEEKEND     0.038416   0.040589   0.946    0.344
MSA         0.083135   0.066082   1.258    0.209
MSB         0.006035   0.141062   0.043    0.966
MSI         0.071983   0.052989   1.358    0.175
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.5333 on 691 degrees of freedom
Multiple R-squared: 0.06983,Adjusted R-squared: 0.0631
F-statistic: 10.38 on 5 and 691 DF,  p-value: 1.304e-09
```

The only one covariate that seems to be significant is ECG. The DB time increases if the elapsed time for the first electrocardiogram is high.

Concerning the survival to discharge a logistic regression model has been considered. For every patient $i = 1, \dots, n$, Y_i is the Bernoulli random variable of mean p_i that describes if the patient is alive to discharge. Then we assume

$$Y_i|p_i \overset{i.i.d}{\sim} Be(p_i),$$
$$logit(p_i) = \beta_0 + \beta_1 log(DB)_i + \beta_2 AGE_i + \beta_3 MILAN_i + \beta_4 KILLIP_i$$

As well as MA, also KILLIP is a categorical covariate and it has been transformed onto a 3 dimensional dummy vector, considering the class 1 as reference. The `glm` function of the MASS package of R gives:

```
Call:
glm(formula = ALIVEIN ~ log(DB) + AGE + MILAN + Killip2 + Killip3 +
    Killip4, family = binomial)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-3.1731   0.1004   0.1317   0.1863   1.9069

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) 10.05412    2.50948   4.006 6.16e-05 ***
log(DB)     -0.50340    0.47452  -1.061   0.2887
```

```
AGE         -0.04012    0.02055  -1.953   0.0508 .
MILAN       -0.88339    0.53882  -1.639   0.1011
Killip2     -1.74644    0.69233  -2.523   0.0117 *
Killip3     -1.74396    1.13844  -1.532   0.1256
Killip4     -4.80502    0.65111  -7.380 1.59e-13 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 195.35  on 696  degrees of freedom
Residual deviance: 125.45  on 690  degrees of freedom
AIC: 139.45

Number of Fisher Scoring iterations: 7
```

The analysis shows that the KILLIP variable is very significant if the severity of the infarction is high, a little less significant if the severity is not high. In particular, in both cases the probability of survival to discharge decreases. AGE is quite significant, less than KILLIP variable.

For those patients survived to discharge we have considered the survival to 60 days using one more time the logistic regression model. For every patient $i = 1, \dots, n^{alive}$, Y_i is the Bernoulli random variable of mean r_i that describes if the patient is alive to discharge. Then we assume

$$Y_i|p_i \overset{i.i.d}{\sim} Be(r_i),$$
$$logit(r_i) = \beta_0 + \beta_1 EF_i + \beta_2 STres_i + \beta_3 CKD_i + \beta_4 MILAN_i$$

We got:

```
Call:
glm(formula = ALIVE60 ~ EF + STres + CKD + MILAN, family = binomial)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-3.3223   0.0978   0.1239   0.1912   0.7575

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  0.18657    1.34812   0.138  0.88993
EF           0.09515    0.03025   3.145  0.00166 **
STres       -0.46494    0.66268  -0.702  0.48293
CKD         -0.90928    0.81444  -1.116  0.26423
MILAN       -0.07838    0.62419  -0.126  0.90007
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

```
(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 112.390  on 674  degrees of freedom
Residual deviance:  99.622  on 670  degrees of freedom
AIC: 109.62

Number of Fisher Scoring iterations: 7
```

EF is a significant covariate. In particular, an high value of the ejection fraction gives a positive contribution to the probability of survival to 60 days.

2.5 Standardization and dichotomization of the variables

In this work we will standardize all the continuous variables, i.e. $log(DB)$, AGE, ECG and EF. Furthermore, we dichotomize the variables MA and KILLIP. The reason of this choice is to improve the convergence (in particular the mixing) of the Markov chain MCMC that we will construct. Moreover, in this way it will be possible to make a comparison with the results obtained in Prandoni (2012).

In particular, for the modality of access we distinguish between the class "Spontaneo" ($MA = 1$) and the other classes ($MA = 0$). Similarly, KILLIP will assume the value 0 if the severity of the infarction is medium-low (class 1 and 2), or 1 if the severity of the infarction is most serious (class 3 and 4). Table 2.5 shows the frequencies of patients in the new categories:

Table 2.5: Percentage of patients with respect to modality of access and killip (now dichotomized).

Modality	%
Means	40.32
Spontaneous	59.68

Killip	%
0	93.97
1	6.03

Figure 2.7 shows the boxplot of $log(DB)$ with respect to the new modality of access.

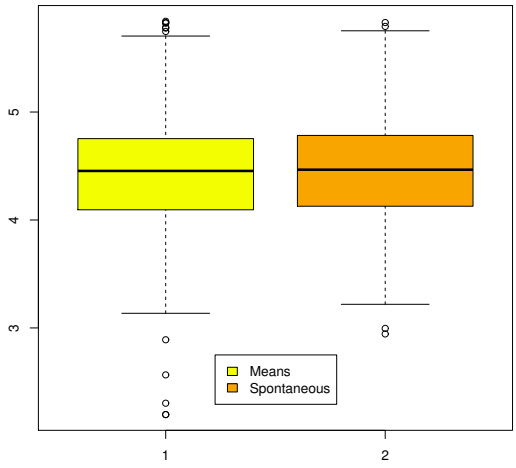


Figure 2.7: Boxplot of $log(DB)$ with respect to the modality of access.

A slight difference in terms of the modality of access can be appreciated. This does not happen if we don't use the logarithm of DB. Using the nonparametric Kruskal-Wallis test as in Section 2.3 we obtain a p-value equal to 0.3499. There is no evidence that the two groups have different means.

2.5.1 Preliminary frequentist analysis of the modified dataset

We are going to repeat the analysis reported in Section 2.4, but now considering the standardized and dichotomized covariates as explained above.

For the first level, since this time $MA = 1$ if the patient reached the hospital by his own means, the intercept represents the mean of the (standardized) logarithmic of the DB time for those patients who have reached the hospital with an ambulance. Using the `lm` function we obtain the following output:

```
Call:
lm(formula = logDB ~ ECG + WEEKEND + MA)

Residuals:
    Min      1Q  Median      3Q     Max
-3.9924 -0.6016 -0.0397  0.5487  2.6341

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -0.12737    0.10217  -1.247    0.213
ECG          0.25265    0.03688   6.851 1.62e-11 ***
WEEKEND      0.07536    0.07345   1.026    0.305
MA           0.03047    0.02943   1.035    0.301
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

```
Residual standard error: 0.9674 on 693 degrees of freedom
Multiple R-squared: 0.06818,Adjusted R-squared: 0.06415
F-statistic:  16.9 on 3 and 693 DF,  p-value: 1.323e-10
```

Again, the only one covariate that seems to be significant is ECG: DB time increases if the elapsed time for the first electrocardiogram increases.

For the second level, using the `glm` function we obtain the following output:

```
Call:
glm(formula = ALIVEIN ~ logDB + AGE + MILAN + KILLIP, family = binomial)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-3.1718   0.0978   0.1284   0.1825   1.7944

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)   6.8295     0.7015   9.735  < 2e-16 ***
logDB        -0.2230     0.2604  -0.856   0.3919
AGE          -0.5671     0.2577  -2.201   0.0277 *
MILAN        -0.9357     0.5255  -1.781   0.0750 .
KILLIP       -1.5399     0.2143  -7.185 6.74e-13 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 195.35  on 696  degrees of freedom
Residual deviance: 128.11  on 692  degrees of freedom
AIC: 138.11

Number of Fisher Scoring iterations: 7
```

The KILLIP variable results strongly significant again, confirming as before that if the severity of the infarction is high the probability of survival to discharge is low. Also AGE and MILAN are significant and both of them contribute negatively to the survival to discharge.

Finally, for the third level the result of the analysis is the following:

```
Call:
glm(formula = ALIVE60 ~ EF + STres + CKD + MILAN, family = binomial)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-3.3223   0.0978   0.1239   0.1912   0.7575
```

```
Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  4.74007    0.53962   8.784  < 2e-16 ***
EF           0.91937    0.29233   3.145  0.00166 **
STres       -0.46494    0.66268  -0.702  0.48293
CKD         -0.90928    0.81444  -1.116  0.26423
MILAN       -0.07838    0.62419  -0.126  0.90007
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 112.390  on 674  degrees of freedom
Residual deviance:  99.622  on 670  degrees of freedom
AIC: 109.62

Number of Fisher Scoring iterations: 7
```

Also in this case it seems that an high value of the ejection fraction gives a positive contribution to the probability of survival to 60 days.

Chapter 3

The Bayesian model

We have to construct a Bayesian model where the response vector has three components: the DB time (in logarithmic scale), the survival to discharge and the survival to 60 days. Also, we want to consider the clustering structure induced by the Dirichlet process, as described in Chapter 1.

For every patient $i = 1, \dots, n$, we consider the vector

$$\boldsymbol{Y}_i := (Y_{i1}, Y_{i2}, Y_{i3}) := (log(DB)_i, VIVOIN_i, VIVO60_i).$$

Keeping the same notation as in Section 1.3, let $\rho_n = \{S_1, \dots, S_k\}$ denote a partition of the n experimental units into k subsets S_j and let e_i indicate which "latent class" is associated with observation y_i, i.e. which group in the partition is associated with observation y_i with $e_i = j$ if $i \in S_j$. Recall that $\boldsymbol{\theta}_j$, $j = 1, \dots, k$, and $\boldsymbol{\eta}$ indicate respectively *clusters specific parameters* and common *hyperparameters*, respectively (where k is the number of clusters).
Moreover, in the sequel we will refer to data $y_1, \dots, y_n$ as part of an indefinite exchangeable sequence. For each class, e, the parameters ϕ_e determine the distribution of observations from that class. α and G_0 are, respectively, the concentration parameter and the base distribution of a Dirichlet process (i.e., with base measure αG_0). Finally, if y_i belongs to the class e, the likelihood will be denoted by $F(y_i, \phi_e)$.

Conditionally to the partition, the data are independent between each cluster and, within clusters, depend of cluster specific parameters and covariates; see (1.6) for the general expression of the conditional distribution of the data. Moreover, for each $i = 1, \dots, n$ we have

$$\mathcal{L}(\boldsymbol{Y}_i|\boldsymbol{\theta}_j, e_i = j, \boldsymbol{x}_i) = \mathcal{L}(Y_{i1}|\boldsymbol{\theta}_{j1}, e_i = j, \boldsymbol{x}_{i1})\mathcal{L}(Y_{i2}|\boldsymbol{\theta}_{j2}, Y_{i1}, \boldsymbol{x}_{i2})\mathcal{L}(Y_{i3}|\boldsymbol{\theta}_{j3}, Y_{i2}, \boldsymbol{x}_{i3}), \quad (3.1)$$

where $\boldsymbol{x}_{il}$ denotes the set of covariates associated to the ith patient relatively to the level l ($l = 1, 2, 3$). In particular, the covariates are

$$\begin{aligned}
\boldsymbol{x}_{i1} &= \{ECG_i, WEEKEND_i, MA_i\},\\
\boldsymbol{x}_{i2} &= \{AGE_i, MILAN_i, KILLIP_i\},\\
\boldsymbol{x}_{i3} &= \{EF_i, STres_i, CKD_i, MILAN_i\}.
\end{aligned}$$

3.1 Construction of the model

3.1.1 Likelihood and prior

The conditional distributions of each variable Y_i is defined in (3.1). Each block there is so defined:

$$\begin{aligned} Y_{i1}|\mu_{ij},\tau_j &\overset{ind}{\sim} \mathcal{N}(\mu_{ij},\tau_j^{-1}),\\ \mu_{ij} &= \beta_{0j}^1 + \beta_{1j}^1 ECG_i + \beta_{2j}^1 WEEKEND_i + \beta_{3j}^1 MA_i, \end{aligned} \tag{3.2}$$

$$\begin{aligned} &Y_{i2}|p_{ij}Y_{i1} \overset{ind}{\sim} Be(p_{ij}),\\ &logit(p_{ij}) = \beta_{0j}^2 + \beta_{1j}^2 log(DB)_i + \beta_{2j}^2 AGE_i + \beta_{3j}^2 MILAN_i + \beta_{3j}^2 KILLIP_i, \end{aligned} \tag{3.3}$$

$$\begin{aligned} &Y_{i3}|r_{ij}Y_{i2} \overset{ind}{\sim} Be(r_{ij}I_{\{1\}}Y_{i2}),\\ &logit(r_{ij}) = \beta_{0j}^3 + \beta_{1j}^3 EF_i + \beta_{2j}^3 STres_i + \beta_{3j}^3 CKD_i + \beta_{4j}^3 MILAN_i. \end{aligned} \tag{3.4}$$

so that the law of the ith response vector in the jth cluster is

$$\mathcal{L}(\boldsymbol{Y}_i|\boldsymbol{\beta}_j^1,\boldsymbol{\beta}_j^2,\boldsymbol{\beta}_j^3,\tau_j,\boldsymbol{x}_i) = \mathcal{L}(Y_{i1}|\boldsymbol{\beta}_j^1,\tau_j,\boldsymbol{x}_{i1})\mathcal{L}(Y_{i2}|Y_{i1},\boldsymbol{\beta}_j^2,\boldsymbol{x}_{i2})\mathcal{L}(Y_{i3}|Y_{i2},\boldsymbol{\beta}_j^3,\boldsymbol{x}_{i3}), \tag{3.5}$$

where $\boldsymbol{\beta}_j^1 = (\beta_{j0}^1,\dots,\beta_{j3}^1)$, $\boldsymbol{\beta}_j^2 = (\beta_{j0}^2,\dots,\beta_{4j}^2)$, $\boldsymbol{\beta}_j^3 = (\beta_{j0}^3,\dots,\beta_{j4}^3)$ denote the regression coefficients for the cluster j, whereas τ_j^{-1} represents the variance of Y_{i1}. Here $(\boldsymbol{\beta}_j^1,\tau_j)$ constitutes the first set of cluster specific parameters, $\boldsymbol{\beta}_j^2$ the second and $\boldsymbol{\beta}_j^3$ the third. In particular, $\boldsymbol{\theta}_j = (\boldsymbol{\beta}_j^1,\tau_j,\boldsymbol{\beta}_j^2,\boldsymbol{\beta}_j^3)$ is the parameters vector of dimension 15.

As far as the prior distribution $p(\boldsymbol{\theta}_1,\dots,\boldsymbol{\theta}_k|\boldsymbol{\eta})$ in (1.6) is concerned, we assign it as follows: each $p(\boldsymbol{\theta}_j|\boldsymbol{\eta})$, where $\boldsymbol{\eta} = (b,\boldsymbol{\beta}_0^1,\boldsymbol{\beta}_0^2,\boldsymbol{\beta}_0^3)$, is given as the product of three blocks:

$$p(\boldsymbol{\beta}_j^1,\tau_j,\boldsymbol{\beta}_j^2,\boldsymbol{\beta}_j^3) = p(\boldsymbol{\beta}_j^1|\tau_j)p(\tau_j)p(\boldsymbol{\beta}_j^2)p(\boldsymbol{\beta}_j^3) \tag{3.6}$$

where

$$\boldsymbol{\beta}_j^1|\tau_j \sim \mathcal{N}_4\left(\boldsymbol{\beta}_0^1, \frac{1}{k_0^1\tau_j}I\right),\ \tau_j \sim \mathcal{G}(a,b), \tag{3.7}$$

$$\boldsymbol{\beta}_j^2 \sim \mathcal{N}_5\left(\boldsymbol{\beta}_0^2, \frac{1}{k_0^2}I\right), \tag{3.8}$$

$$\boldsymbol{\beta}_j^3 \sim \mathcal{N}_5\left(\boldsymbol{\beta}_0^3, \frac{1}{k_0^3}I\right). \tag{3.9}$$

We are going to complete the prior specification assigning a prior distribution for the common parameters $\boldsymbol{\eta} = (b,\boldsymbol{\beta}_0^1,\boldsymbol{\beta}_0^2,\boldsymbol{\beta}_0^3)$, so that the prior is specified as follows:

$$\begin{aligned} \boldsymbol{\theta}_j|\boldsymbol{\eta} &\overset{i.i.d}{\sim} \pi(\boldsymbol{\theta}_j|\boldsymbol{\eta}), \\ \boldsymbol{\eta} &\sim \pi(\boldsymbol{\eta}). \end{aligned}$$

In particular, for any $j = 1, \ldots, k$, we assume:

$$\boldsymbol{\beta}_0^1 \sim \mathcal{N}_4(\boldsymbol{m}_1, B_1), b \sim \mathcal{G}(a_1, a_2), \tag{3.10}$$

$$\boldsymbol{\beta}_0^2 \sim \mathcal{N}_5(\boldsymbol{m}_2, B_2), \tag{3.11}$$

$$\boldsymbol{\beta}_0^3 \sim \mathcal{N}_5(\boldsymbol{m}_3, B_3), \tag{3.12}$$

where, for $l = 1, 2, 3$, $\boldsymbol{m}_l$, are fixed and $B_l = g_l(X_l^T X_l)^{-1}$, being X_l the design matrix of the level l. That is, we put a Zellner g-prior on the three common means.
An important result related to the structure of the Gibbs sampler that we will implement is that the upgrade of the common parameters does not depends directly on the data but only on the clusters specific parameters .
Therefore, the law of $\boldsymbol{\theta}_j$ can be factorized as follows:

$$\begin{aligned} \mathcal{L}\left(\boldsymbol{\theta}_j|\boldsymbol{\eta}\right) &= \mathcal{L}\left(\boldsymbol{\beta}_j^1, \tau_j, \boldsymbol{\beta}_j^2, \boldsymbol{\beta}_j^3|b, \boldsymbol{\beta}_0^1, \boldsymbol{\beta}_0^2, \boldsymbol{\beta}_0^3\right) \\ &= \mathcal{L}\left(\boldsymbol{\beta}_j^1, \tau_j|\boldsymbol{\beta}_0^1, b\right) \mathcal{L}\left(\boldsymbol{\beta}_j^2|\boldsymbol{\beta}_0^2\right) \mathcal{L}\left(\boldsymbol{\beta}_j^3|\boldsymbol{\beta}_0^3\right) \end{aligned} \tag{3.13}$$

In conclusion our model has the following structure:

$$\begin{aligned} \boldsymbol{Y}_i|\boldsymbol{\theta}_{e_i} &\overset{ind}{\sim} p(y_i|\boldsymbol{\theta}_{e_i}) \qquad i = 1, \ldots, k \\ \boldsymbol{\theta}_i|\boldsymbol{\eta} &\sim \pi(\boldsymbol{\theta}_i|\boldsymbol{\eta}) \qquad i = 1, \ldots, k \\ \boldsymbol{\eta} &\sim \pi(\boldsymbol{\eta}) \\ (e_1, \ldots, e_n) &\leftrightarrow \rho \sim \pi(\rho|\boldsymbol{x}) \propto \prod_{j=1}^{k} c(S_j) g(x_j^*). \end{aligned} \tag{3.14}$$

We have just described the first 3 blocks, the only left is $\pi(\rho)$.

3.2 Similarity functions in the PPMx model

We have to describe the prior $\pi(\rho|\boldsymbol{x}^S)$ as in (1.4). Let's denote with $\boldsymbol{x}^L$ and $\boldsymbol{x}^S$ the covariates in the likelihood and similarity, respectively. Remember that, by definition (1.5), $g(\cdot)$ is the marginal probability in an auxiliary model q:

$$g(x_j^*) = \int \prod_{i \in S_j} q\left(x_i|\xi_j\right) q\left(\xi_j\right) d\xi_j,$$

where $x_j^* = (x_i, i \in S_j)$. Observing that the argument of the integral is the product between a likelihood and a prior, we obtain the general expression of the similarity using the Bayes' theorem in the following way:

$$\int q(x|\lambda)q(\lambda)d\lambda = q(x|\theta)\frac{q(\theta)}{q(\theta|x)} = likelihood \times \frac{prior}{posterior}, \tag{3.15}$$

and this expression is independent of θ. As pointed out in Muller, Quintana, and Rosner (2011), an appropriate choice of the auxiliary model q implies computational simplicity.

In this work, covariates considered in the similarity functions are of categorical and binary type. The following table summarizes which covariates are in the likelihood and which in the similarity.

Table 3.1: Variables in the likelihood and similarity, distinguishing between continuous, categorical and binary case.

	$\boldsymbol{x}^L$	$\boldsymbol{x}^S$
Continuous	ECG AGE EF	
Binary	WEEKEND MA MILAN KILLIP STres CKD	CHOLESTEROL COMPLICATIONS DIABETES SMOKING HYPERTENSION preAMI GENDER VASCULOPATIA
Categorical		HOSPITAL

3.2.1 Similarity, categorical covariates

When constructing a similarity function for categorical covariates, a default choice is based on Dirichlet prior. Assume x_i is a categorical covariate, $x_i \in \{1, \dots, C\}$, with $\boldsymbol{x} = (x_1, \dots, x_C)$. Let $X = (X_1, \dots, X_C)$ be the vector of the observed numerosities for every category and let n_j be the size of the cluster. Then, $X|p \sim Multinomial(n_j, p)$:

$$\pi(\boldsymbol{x}|p) = \frac{n_j!}{x_1! \cdots x_C!} p_1^{x_1} \cdots p_C^{x_C}, \quad \sum_{i=1}^{C} x_i = n_j,$$

A priori $p \sim Dirichlet(\boldsymbol{\gamma})$:

$$\pi(p) = \frac{1}{B(\boldsymbol{\gamma})} p_1^{\gamma_1 - 1} \cdots p_C^{\gamma_C - 1}, \quad \sum_{i=1}^{M} p_i = 1,$$

where $\frac{1}{B(\boldsymbol{\gamma})} = \frac{\Gamma(\gamma_1 + \cdots + \gamma_C)}{\Gamma(\gamma_1) \cdots \Gamma(\gamma_C)}$.

Given the data, it is straightforward to see that $p|\boldsymbol{x} \sim Dirichlet(\boldsymbol{\gamma} + \boldsymbol{x})$. Therefore, using (3.15), the similarity in the categorical case is

$$g(x_j^*) = \pi(\boldsymbol{x}|p) \times \frac{\pi(p)}{\pi(p|\boldsymbol{x})} = \frac{n_j!}{x_1! \cdots x_C!} \frac{B(\boldsymbol{\gamma} + \boldsymbol{x})}{B(\boldsymbol{\gamma})}. \tag{3.16}$$

This is a Dirichlet-Multinomial model without the multinomial coefficient. For binary covariates the similarity function becomes a Beta-Binomial probability without the Binomial coefficient. We choose $\boldsymbol{\gamma} = (\gamma_1, \ldots, \gamma_C)$ such that $\gamma_i = t \ \forall i \in \{1, \ldots, C\}$. It is recommended to choose $\gamma_i < 1$. This is to facilitate the formation of clusters that are characterized by the categorical covariates. For example, for $C = 2$, the bimodal nature of a Beta distribution with such parameters assigns high probability to binomial success probabilities close to 0 or 1. Similarly, the Dirichlet distribution with parameters $\gamma_i < 1$ favours clusters corresponding to specific levels of covariates (see Muller, Quintana, and Rosner, 2011).
Also, if $\gamma_i = 1$ it results

$$g(x_j^*) = \frac{n_j!}{x_1! \cdots x_C!}\Gamma(C)\frac{\Gamma(x_1+1)\cdots\Gamma(x_C+1)}{\Gamma(n_j+C)} = n_j!\frac{\Gamma(C)}{\Gamma(n_j+C)},$$

and so, in this case, the numerosities x_i would not be considered.

3.3 Neal's Algorithm

Use of Dirichlet process mixture models has become computationally feasible with the development of Markov chain methods for sampling from the posterior distribution of the parameters of the component distributions. Methods based on Gibbs sampling can easily be implemented for model based on conjugate prior distributions, but when non-conjugated priors are used, as is appropriate in many contexts, straightforward Gibbs sampling requires that an often difficult numerical integration be performed.

As resumed in Neal (2000), West, Muller, and Escobar (1994), used a Monte Carlo approximation to this integral, but the error from using such an approximation is likely to be large in many contexts. MacEachern and Muller (1998) devised an exact approach for handling non-conjugate priors that uses a mapping from a set of auxiliary parameters to the set of parameters currently in use. Their algorithms based on this approach are widely applicable, but somehow inefficient.

Neal (2000) reviews this past work and presents a new approach to Markov chain sampling. A wide class of methods for handling non-conjugated priors uses Gibbs sampling in a space with auxiliary parameters. This approach yields an algorithm that resembles use of a Monte Carlo approximation to the necessary integrals, but which does not suffer from any approximation error. The basic idea of auxiliary variable methods is that we can sample from a distribution π_x for x by sampling from some distribution π_{xy} for (x, y), with respect to which the marginal distribution of x is π_x. The permanent state of the Markov chain will be x, but a variable y will be introduced temporarily during an update of the following form:

1. Draw a value for y from its conditional distribution given x, as defined by the joint distribution π_{xy}.

2. Perform some update of (x, y) that leaves π_{xy} invariant.

3. Discard y, leaving only the value of x.

It is easy to see that this update for x will leave π_x invariant as long as π_x is the marginal distribution of x under π_{xy}.

We can use this technique to update the e_i for a Dirichlet process mixture model without having to integrate with respect to G_0. The permanent state of the Markov chain will consist of the e_i and the ϕ_e and when e_i is updated we will introduce temporary auxiliary variables that represent possible values for the parameters of components that are not associated with any other observations. We than update e_i by Gibbs sampling with respect to the distribution that includes these auxiliary parameters.

Since observations y_i are exchangeable and the component labels e_i are arbitrary, we can assume that we are updating e_i for the last observation, and that e_j for other observations have value in the set $\{1, \ldots, k^-\}$, where k^- is the number of distinct e_j for $j \neq i$. We can now visualize the conditional prior distribution for e_i given the other e_j in terms of m auxiliary components and their associated parameters. The probability of e_i being equal to a e in $\{1, \ldots, k^-\}$ will be $n_{-i,e}/(n-1+\alpha)$, where $n_{-i,e}$ is the number of times e occurs among the e_j for $j \neq i$. The probability of e_i having some other value will be $\alpha/(n-1+\alpha)$, which we will split equally among the m auxiliary components we have introduced (see Neal, 2000 for details). In our case it will be $m = 1$. The algorithm can be summarized as follows:

Algorithm 8. Let the state of the Markov chain consist of $e = (e_1, \ldots, e_n)$ and $\phi = (\phi_e : e \in \{e_1, \ldots, e_n\})$. Repeatedly sample as follow:

- For $i = 1, \ldots, n$: Let k^- be the number of distinct e_j for $j \neq i$, and let $h = k^- + m$. Labels these e_j with values in $\{1, \ldots, k^-\}$. If $e_i = e_j$ for some $j \neq i$, draw values independently from G_0 for those ϕ_e for which $k^- < e \leq h$. If $e_i \neq e_j$ for all $j \neq i$, let e_i have the label $k^- + 1$, and draw values independently from G_0 for those ϕ_e for which $k^- + 1 < e \leq h$. Draw a new value for e_i from $\{1, \ldots, h\}$ using the following probabilities:

$$P(e_i = e \,|\, e_{-i}, y_i, \phi_1, \ldots, \phi_h) \propto \begin{cases} n_{-i,e} F(y_i, \phi_e) & \text{for } 1 \leq e \leq k^- \\ \alpha/m F(y_i, \phi_e) & \text{for } k^- < e \leq h \end{cases} \tag{3.17}$$

  where $n_{-i,e}$ is the number of cej for $j \neq i$ that are equal to e. Change the state to contain only those ϕ_e that are now associated with one or more observations.

- For all $e \in \{e_1, \ldots, e_n\}$: Draw a new value from $\phi_e | y_i$ such that $e_i = e$, or perform some other update to ϕ_e that leaves this distribution invariant.

- Perform the update of all common parameters.

Note that the relabellings of the e_j above are conceptual; they may or may not require any actual computation, depending on the data structures used.

First step of Neal's algorithm concerns the formation of the clusters. For any data, the basic idea is to see how likely the data itself can be added to an existing cluster or, otherwise, form a new cluster between m possible ones. We fix $m = 1$. A colourful description of this partition structure is given by the *Chinese restaurant process* (Arratia, Barbour, and Tavaré, 1992). Imagine that $n-1$ customers are seated in k^- tables. The nth customer (that we label with i) will either choose an empty table with probability proportional to $\alpha F(y_i, \phi_e)$ for some $\alpha > 0$, or an occupied table with probability proportional to the number of occupants at the given table.

In our case $y_i = (y_{i1}, y_{i2}, y_{i3})$, where, for $i = 1, \ldots, n$, y_{i1}, y_{i2} and y_{i3} are independently drawn respectively from a Normal and two Bernoulli distributions. Remembering the expression of the law of the response vector (3.5), for a fixed class c, $F(y_i, \phi_e)$, has the following expression:

$$\begin{aligned} F(y_i, \phi_e) &= \Big(\frac{\tau_j}{2\pi}\Big)^{1/2} exp\big\{ -\frac{\tau_j}{2}(y_{i1} - \mu_{ij})^2 \big\} \\ &\times \; p_{ij}^{y_{i2}}(1 - p_{ij})^{1-y_{i2}} \\ &\times \; r_{ij}^{y_{i3}}(1 - r_{ij})^{1-y_{i3}} \mathbb{I}_{\{y_{i2}=1\}}, \end{aligned}$$

where μ_{ij}, τ_j, $logit(p_{ij})$ and $logit(r_{ij})$ are defined in Section 3.1.

As said in Section 1.3, the marginal distribution that a DP induces on partitions is also a PPM with cohesion $c(S_j) = \alpha(|S_j| - 1)!$. Indeed, an equivalent way to write (3.17) is the following:

$$P(e_i = e |\, e_{-i}, y_i, \phi_1, \ldots, \phi_h) \propto \begin{cases} \frac{c(S_j \cup \{y_i\})}{c(S_j)} F(y_i, \phi_e) & \text{for } 1 \le e \le k^- \\ c(\{y_i\})/m F(y_i, \phi_e) & \text{for } k^- < e \le h \end{cases}$$

since $c(S_j \cup \{y_i\})/c(S_j) = |S_j| = n_{-i,c}$, and $c(\{y_i\}) = \alpha$.
To implement the PPMx we have to consider the similarity function in addiction to cohesion, modifying the probabilities in the following way:

$$P(e_i = e |\, e_{-i}, y_i, \phi_1, \ldots, \phi_h) \propto \begin{cases} \frac{c(S_j \cup \{y_i\})}{c(S_j)} \frac{g(x_j^* \cup \{x_i\})}{g(\{x_j^*\})} F(y_i, \phi_e) & \text{for } 1 \le e \le k^- \\ c(\{y_i\}) g(\{x_i\})/m F(y_i, \phi_e) & \text{for } k^- < e \le h \end{cases} \tag{3.18}$$

where $x_j^* = (x_i,\ i \in S_j)$ represents the set of covariates in the cluster S_j as introduced in Section 1.3.

3.3.1 Updates for cluster specific parameters

The second step of the algorithm is related to parameters actualization that appear in the likelihood, i.e. $\boldsymbol{\beta}_j^1$, τ_j, $\boldsymbol{\beta}_j^2$ and $\boldsymbol{\beta}_j^3$, for $j = 1, \ldots, k$. The law of these parameters is factorized as in (3.6).

It is clear that such update concerns every cluster created in the current iteration. For this reason and to simplify notation we will write $\boldsymbol{\beta}$, τ, $\boldsymbol{\beta}_0$ and k_0 instead of $\boldsymbol{\beta}_j^l$, τ_j, $\boldsymbol{\beta}_0^l$, k_0^l for $j = 1, \ldots, k$ and $l = 1, 2, 3$, depending on the level, and consider only those statistical units i belonging to cluster j; that is, the set $\{i : c(i) = j\}$ whose cardinality will be denoted with n_j.

First level

Notice that the first factor in (3.5) is the final law of a Bayesian linear model with unknown variance (as described in Appendix A). Indeed, let's consider the variables $Y_{i1} = log(DB)_i$ and remember the expression of the likelihood in (3.2). Define $w_i = log(DB)_i$ and let $\boldsymbol{w} = (w_1, \ldots, w_{n_j})$. Then, it results:

$$\boldsymbol{w} \sim \mathcal{N}(X\boldsymbol{\beta}, \tau^{-1} I_{n_j}),$$

where X is the design matrix and I_{n_j} represents the identity matrix of dimension n_j. The prior on the $\boldsymbol{\beta}$s and τ is written in (3.10):

$$\boldsymbol{\beta}|\tau, \boldsymbol{\beta}_0 \sim \mathcal{N}_4\left(\boldsymbol{\beta}_0, \frac{1}{\tau}B_0\right), \tau|b \sim \mathcal{G}(a, b),$$

with $B_0 = I_{n_j}/k_0$.
We label this distribution *Normal-Gamma* and write $\boldsymbol{\beta}, \tau \sim \mathcal{NG}(\boldsymbol{\beta}_0, B_0, a, b)$ for short. It can be shown (see Appendix A) that the posterior distribution is still Normal-Gamma:

$$\boldsymbol{\beta}, \tau|\boldsymbol{w} \sim \mathcal{NG}\left(\boldsymbol{\beta}^*, B^*, a^*, b^*\right),$$

where

$$\begin{aligned} B^* &= \left(B_0^{-1} I_{n_j} + X^T X\right)^{-1} \\ \boldsymbol{\beta}^* &= B^*\left(B_0^{-1}\boldsymbol{\beta}_0 + X^T \boldsymbol{w}\right) \\ a^* &= a + n_j/2 \\ b^* &= b + \frac{1}{2}\left(\boldsymbol{\beta}_0^T B_0^{-1}\boldsymbol{\beta}_0 + \boldsymbol{w}^{\boldsymbol{T}}\boldsymbol{w} - \left(\boldsymbol{\beta}^*\right)^T \left(B^*\right)^{-1}\left(\boldsymbol{\beta}^*\right)\right). \end{aligned}$$

Second and third level

Let's consider the binary variables $Y_{i2} = ALIVEIN_i$ and $Y_{i3} = ALIVE60_i$.

The *logit* appears in the expression of the likelihoods defined in (3.3) and (3.4). Hence, in this case, unlike the previous one, it is impossible to find a conjugated prior. For this reason, a Random Walk Metropolis Hastings algorithm will be performed to sampling from the unknown posterior distribution (see for instance Chib and Greenberg, 1995). Let $\boldsymbol{z} \sim \mathcal{N}(\mathbf{0}, \Sigma)$ and $\boldsymbol{\beta}^{prop} = \boldsymbol{\beta}^{old} + \boldsymbol{z}$. We accept to upgrade the value of $\boldsymbol{\beta}$ from $\boldsymbol{\beta}^{old}$ to $\boldsymbol{\beta}^{prop}$ with probability $\alpha = min\left\{1, \pi\left(\boldsymbol{\beta}^{prop}\right)/\pi\left(\boldsymbol{\beta}^{old}\right)\right\}$, where $\pi(\boldsymbol{\beta})$ is the *target* distribution proportional to the product between the likelihood and the prior:

$$\pi(\boldsymbol{\beta}) \propto \prod_{i=1}^{n_j}\left\{\left(\frac{e^{\boldsymbol{x}_i^T\boldsymbol{\beta}}}{1+e^{\boldsymbol{x}_i^T\boldsymbol{\beta}}}\right)^{y_i}\left(\frac{1}{1+e^{\boldsymbol{x}_i^T\boldsymbol{\beta}}}\right)^{1-y_i}\right\} exp\left\{-\frac{k_0}{2}\left(\boldsymbol{\beta}-\boldsymbol{\beta}_0\right)^T\left(\boldsymbol{\beta}-\boldsymbol{\beta}_0\right)\right\},$$

where $\boldsymbol{x}_i$ is the ith row of the appropriate design matrix. We try to obtain an acceptance rate between 30% and 40% fixing Σ proportional to the identity matrix: $\Sigma = sI$. The choice of s is quite important to achieve the desired acceptance rate. The value of this parameter may change if other hyperparameters change.

3.3.2 Updates for common parameters

We want now calculate the full conditionals distributions of the hyperparameters $\boldsymbol{\eta} = (b, \boldsymbol{\beta}_0^1, \boldsymbol{\beta}_0^2, \boldsymbol{\beta}_0^3)$ introduced in Section 3.1. As already mentioned, the upgrade of these common hyperparameters does not depends directly on the data but only on the cluster specific parameters. That is, once $\boldsymbol{\theta}_j$ have been updated for any $j = 1, \ldots, k$, being k the number of clusters in the current iteration, it is possible to proceed to such upgrade. Let $\boldsymbol{\tau}$ denote the entire vector of τ_j: $\boldsymbol{\tau} = (\tau_1, \ldots, \tau_k)$. It results

$$\begin{aligned}\pi\left(\boldsymbol{\eta}|\boldsymbol{\theta}_1,\ldots,\boldsymbol{\theta}_k\right) &= \pi\left(b,\boldsymbol{\beta}_0^1,\boldsymbol{\beta}_0^2,\boldsymbol{\beta}_0^3|\boldsymbol{\beta}_1^1,\ldots,\boldsymbol{\beta}_k^1,\boldsymbol{\tau},\boldsymbol{\beta}_1^2,\ldots,\boldsymbol{\beta}_k^2,\boldsymbol{\beta}_1^3,\ldots,\boldsymbol{\beta}_k^3\right) \\ &= \pi\left(\boldsymbol{\beta}_0^1,b|\boldsymbol{\beta}_1^1,\ldots,\boldsymbol{\beta}_k^1,\boldsymbol{\tau}\right)\pi\left(\boldsymbol{\beta}_0^2|\boldsymbol{\beta}_1^2,\ldots,\boldsymbol{\beta}_k^2\right)\pi\left(\boldsymbol{\beta}_0^3|\boldsymbol{\beta}_1^3,\ldots,\boldsymbol{\beta}_k^3\right)\end{aligned} \tag{3.19}$$

We distinguish between the first, second and third level. Indeed, these last two levels have the same functional form. Also, the index l that identifies the level will be understood by the context.

First level

Remember that $\pi\left(\boldsymbol{\beta}_j,\tau_j|\boldsymbol{\beta}_0,b\right)=\pi\left(\boldsymbol{\beta}_j|\tau_j\boldsymbol{\beta}_0\right)\pi\left(\tau_j|b\right)$. We have to upgrade $\boldsymbol{\beta}_0$ and b given $\boldsymbol{\beta}_j$ and τ_j, $j\in\{1,\ldots,k\}$. As described before we have

$$\begin{aligned}\pi\left(\boldsymbol{\beta}_1,\ldots,\boldsymbol{\beta}_k,\boldsymbol{\tau}|\boldsymbol{\beta}_0,b\right) &\propto \pi\left(\boldsymbol{\beta}_1,\ldots,\boldsymbol{\beta}_k|\boldsymbol{\beta}_0,\boldsymbol{\tau}\right)\pi\left(\boldsymbol{\tau}|b\right) \\ &= \prod_{j=1}^{k}\pi\left(\boldsymbol{\beta}_j|\boldsymbol{\beta}_0,\tau_j\right)\pi\left(\tau_j|b\right).\end{aligned} \tag{3.20}$$

Hyperparameters upgrade is given by the Bayes' rule:

$$\pi\left(\boldsymbol{\beta}_0,b|\boldsymbol{\beta}_1,\ldots,\boldsymbol{\beta}_k,\boldsymbol{\tau}\right)\propto\pi\left(\boldsymbol{\beta}_1,\ldots,\boldsymbol{\beta}_k,\boldsymbol{\tau}|\boldsymbol{\beta}_0,b\right)\pi\left(\boldsymbol{\beta}_0,b\right), \tag{3.21}$$

being $\pi\left(\boldsymbol{\beta}_0,b\right)=\pi\left(\boldsymbol{\beta}_0\right)\pi\left(b\right)$, since there are no $\boldsymbol{\beta}_0$ and b in the likelihood.
After some calculations (see Appendix B) we obtain

$$\begin{aligned}b|\tau_1,\ldots,\tau_k &\sim \mathcal{G}\left(a_1+ak,a_2+\sum_{j=1}^{k}\tau_j\right), \\ \boldsymbol{\beta}_0|\boldsymbol{\beta}_1,\ldots,\boldsymbol{\beta}_k,\tau_1,\ldots,\tau_k &\sim \mathcal{N}_4\left(\boldsymbol{m}_k,B_k\right),\end{aligned}$$

where

$$\begin{aligned}\boldsymbol{m}_k &= \left(\frac{1}{g}\left(X^TX\right)+k_0\sum_{j=1}^{k}\tau_jI\right)^{-1}\left(\frac{1}{g}\left(X^TX\right)\boldsymbol{m}+k_0\sum_{j=1}^{k}\tau_j\boldsymbol{\beta}_j\right), \\ B_k &= \left(\frac{1}{g}\left(X^TX\right)+k_0\sum_{j=1}^{k}\tau_jI\right)^{-1}.\end{aligned}$$

Second and third level

In this case the structure of the full conditionals is simpler than before because of the absence of τ_j in the covariance. We only have a prior on the mean vector $\boldsymbol{\beta}_0$, that is:

$$\pi\left(\boldsymbol{\beta}_1,\ldots,\boldsymbol{\beta}_k|\boldsymbol{\beta}_0\right) \propto \prod_{j=1}^{k}\pi\left(\boldsymbol{\beta}_j|\boldsymbol{\beta}_0\right). \tag{3.22}$$

Using Bayes' rule, the upgrade of $\boldsymbol{\beta}_0$ is given by the following expression:

$$\pi\left(\boldsymbol{\beta}_0|\boldsymbol{\beta}_1,\ldots,\boldsymbol{\beta}_k\right) \propto \pi\left(\boldsymbol{\beta}_1,\ldots,\boldsymbol{\beta}_k|\boldsymbol{\beta}_0\right)\pi\left(\boldsymbol{\beta}_0\right). \tag{3.23}$$

After some calculations (see Appendix B) we obtain:

$$\boldsymbol{\beta}_0|\boldsymbol{\beta}_1,\ldots,\boldsymbol{\beta}_k \sim \mathcal{N}_5\left(\boldsymbol{m}_k, B_k\right),$$

where

$$\begin{aligned} \boldsymbol{m}_k &= \left(\frac{1}{g}\left(X^TX\right)+k_0kI\right)^{-1}\left(\frac{1}{g}\left(X^TX\right)\boldsymbol{m}+k_0\sum_{j=1}^{k}\boldsymbol{\beta}_j\right), \\ B_k &= \left(\frac{1}{g}\left(X^TX\right)+k_0kI\right)^{-1}. \end{aligned}$$

This results holds for $\boldsymbol{\beta}_i = \boldsymbol{\beta}_i^l$, $l = 2, 3$.

3.4 Summary of the hyperparameters

In order to give greater clarity, we briefly summarize the hyperparameters of the model described so far:

- α denotes the total mass parameter of the DP prior.
- k_0^l are constants proportional to the inverse of the covariance of $\boldsymbol{\beta}_j^l$, $l = 1, 2, 3$.
- a is the shape parameter of the Gamma distribution that models the precision τ_j.
- a_1 is the shape parameter of the Gamma distribution on b.
- a_2 is the rate of the Gamma distribution on b.
- $\boldsymbol{m}_l$ are the means of the Normal distributions on $\boldsymbol{\beta}_0^l$, $l = 1, 2, 3$.
- g_l are the constants in the Zellner g-priors on $\boldsymbol{\beta}_0^l$, $l = 1, 2, 3$.

Chapter 4

Application to the STEMI dataset

In this chapter we analyze the Bayesian inference of the model described in Chapter 3 to the data in Chapter 2. A posterior estimates relate the number of clusters K_n and the global parameters $\boldsymbol{\eta} = (b, \boldsymbol{\beta}_0^1, \boldsymbol{\beta}_0^2, \boldsymbol{\beta}_0^3)$.

We have to fix the value of all the hyperparameters summarized in Section 3.4. Remember that α denotes the total mass parameter of the DP prior. It is known that the number of clusters K_n depends on it. Specifically, from Antoniak (1974), when the prior is the PPM with no covariates, a priori we have:

$$\mathbb{E}\left[K_n\right] = \sum_{i=1}^{n} \frac{\alpha}{\alpha + i - 1}.$$

Here $n = 697$; Table 4.1 shows $\mathbb{E}\left[K_n\right]$ as a function of some values of α.

Table 4.1: A priori values of the mean of K_n for some value of α.

α	0.1	1	10
$\mathbb{E}\left[K_n\right]$	1.70	7.13	43.09

Moreover, we have observed that, under the PPMx model, the number of clusters is smaller than that under the PPM model. For this reason we fixed $\alpha = 10$.
A priori the expression of the mean and variance of cluster specific parameters $\boldsymbol{\theta}_j$, for $j = 1, \ldots, k$, is

$$\mathbb{E}\left[\tau_j\right] = \frac{a a_2}{a_1 - 1} \qquad \text{and} \qquad Var(\tau_j) = \frac{a a_2^2 (a + a_1 - 1)}{(a_1 - 1)^2 (a_1 - 2)}.$$

Note that (see Appendix C), from (3.7), it results that $b = \mathbb{E}(\tau_j | b)/Var(\tau_j | b)$.
Moreover, it is straightforward to see that the prior marginal mean of $\boldsymbol{\beta}_j^l$ is equal to $\boldsymbol{m}_l$, for $l = 1, 2, 3$, whereas the expression of the prior marginal covariance matrix changes between the first level and the other two (see Appendix C):

$$Cov\left[\boldsymbol{\beta}_j^1\right] = \frac{a_1}{a_2(a-1)} \frac{1}{k_0^1} I + g_1 \left(X_1^T X_1\right)^{-1}, \qquad Cov(\boldsymbol{\beta}_j^l) = \frac{1}{k_0^l} I + g_l \left(X_l^T X_l\right)^{-1}, \; l = 2, 3.$$

As a default choice, we have fixed $k_0^l = 1/10$ and $g_l = 10$ in order to have, a priori, an "average" value for the variances within the clusters. Finally, we set $a = 2$, $a_1 = 2$,

$a_2 = 1$, so that the precision τ_j has finite mean equal to 1 and infinite variance.
Concerning the values of $\boldsymbol{m}_l$, we set them equal to the frequentist estimates obtained in Section 2.5.1.

In conclusion, we consider the model (3.14) under the following "benchmark" prior:

$$\boldsymbol{\beta}_j^1|\tau_j \sim \mathcal{N}_4\left(\boldsymbol{\beta}_0^1, \frac{10}{\tau_j} I\right),\ \tau_j \sim \mathcal{G}(2,b),\ \boldsymbol{\beta}_0^1 \sim \mathcal{N}_4\left(\boldsymbol{m}_1, 10\left(X_1^T X_1\right)^{-1}\right),\ b \sim \mathcal{G}(2,1), \quad (4.1)$$

$$\boldsymbol{\beta}_j^2 \sim \mathcal{N}_5\left(\boldsymbol{\beta}_0^2, 10I\right),\ \boldsymbol{\beta}_0^2 \sim \mathcal{N}_5\left(\boldsymbol{m}_2, 10\left(X_2^T X_2\right)^{-1}\right), \quad (4.2)$$

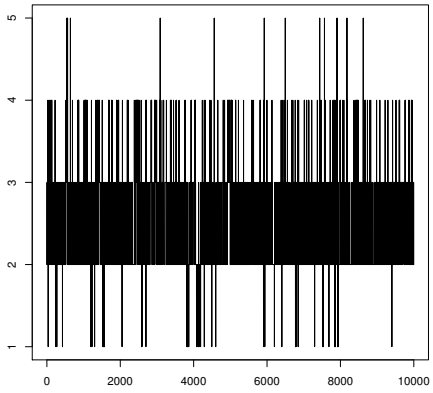
$$\boldsymbol{\beta}_j^3 \sim \mathcal{N}_5\left(\boldsymbol{\beta}_0^3, 10I\right),\ \boldsymbol{\beta}_0^3 \sim \mathcal{N}_5\left(\boldsymbol{m}_3, 10\left(X_3^T X_3\right)^{-1}\right). \quad (4.3)$$

Posterior and predictive estimates have been obtained implementing in C the Gibbs sampler algorithm described in Section 3.3. We run the algorithm for 1,050,000 iterations with a burn-in of 50,000 iterations and a thinning of 100 to reduce the autocorrelation of the Markov chain (in particular of K_n, that seems to be highly correlated). The final sample size is 10,000.

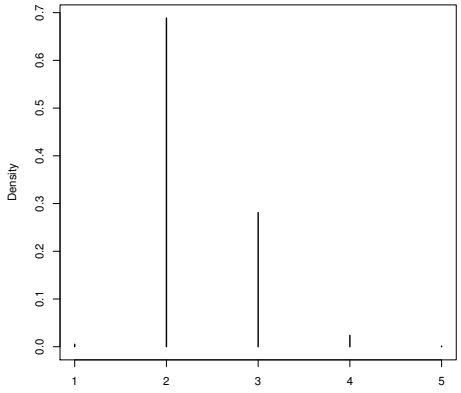
4.1 The PPMx model

4.1.1 Posterior estimates of global parameters

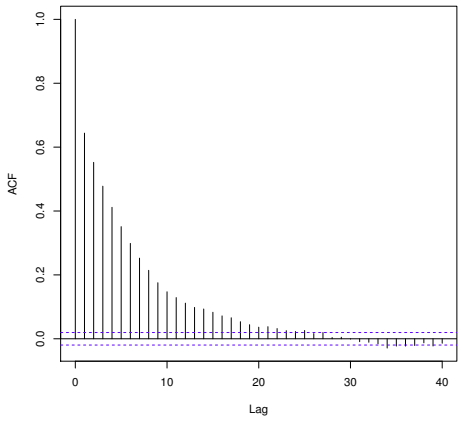
Concerning the number of clusters K_n, Figure 4.1 shows some details about the convergence of the Markov chain approximating the marginal posterior distribution.



(a) Traces of K_n



(b) Posterior distribution of K_n

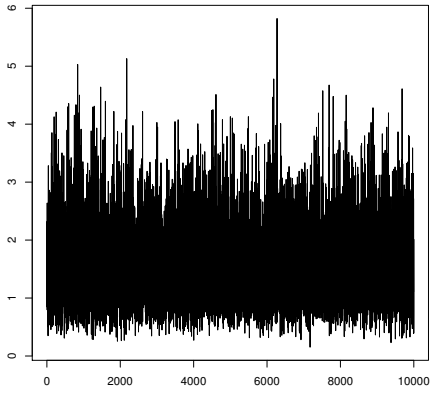


(c) ACF of K_n

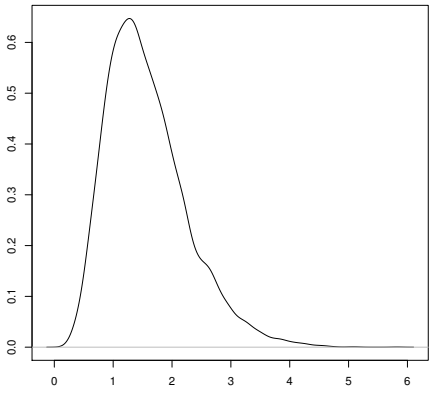
Figure 4.1: Markov chain sample of the number of clusters K_n.

We get $\mathbb{E}[K_n] = 2.33$ and $Var(K_n) = 0.29$. From Figure 4.1c it is clear that there is convergence but that it is quite slow. Actually, this behaviour is related only to the variable K_n. All the other global parameters converge more quickly, as we are going to show now.

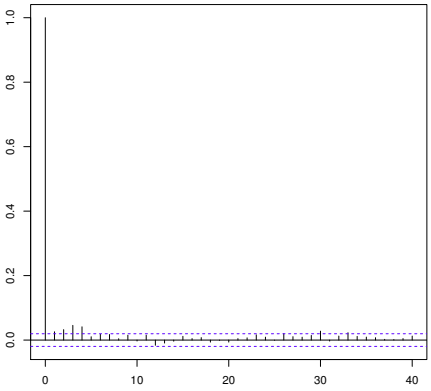
Figure 4.2 shows the trace plot, the density estimation and the auto correlation function of the component of the MC posterior of the parameter b.



(a) Traces of b



(b) Kernel density estimation of the posterior draws of b



(c) ACF of b

Figure 4.2: Markov chain sample of b.

The posterior mean and variance of b are $\mathbb{E}[b] = 1.58$ and $Var(b) = 0.47$, respectively.

Concerning the first component, the logarithm of DB, Figure 4.3 and 4.4 show the trace plots and the density estimation of the $\boldsymbol{\beta}_0^1$ components, whereas the values of the posterior means and 95% credible regions are shown in Table 4.2.

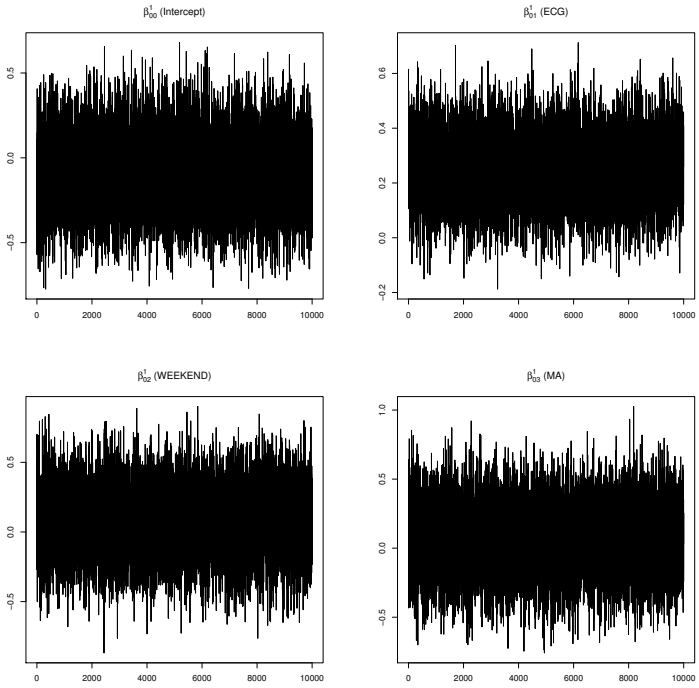


Figure 4.3: Traces of $\boldsymbol{\beta}_0^1$ components.

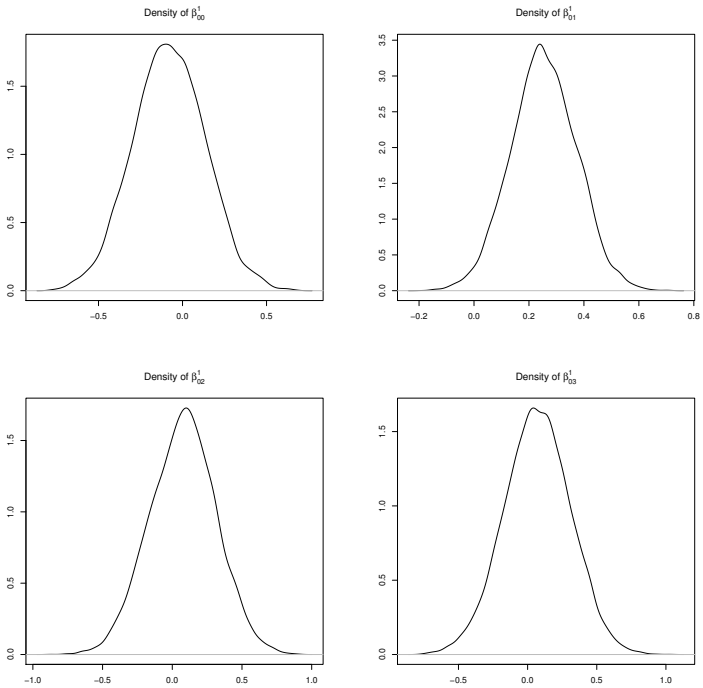


Figure 4.4: Posterior kernel density estimation of $\boldsymbol{\beta}_0^1$ components.

Table 4.2: Posterior means and credible regions for $\boldsymbol{\beta}_0^1$ components.

variable	parameter	mean	2.5%	50%	97.5%
Intercept	β_{00}^1	-0.08	-0.50	-0.08	0.34
ECG	β_{01}^1	0.25	0.02	0.25	0.49
WEEKEND	β_{02}^1	0.08	-0.39	0.08	0.54
MA	β_{03}^1	0.07	-0.41	0.07	0.53

From these results we note that the values of the parameter corresponding to ECG are centered on positive numbers. We deduce that the DB time increases if the elapsed time for the first electrocardiogram is high. WEEKEND and MA variable do not seem to be particularly significant, since they are centered around zero.

Concerning the second component, the the survival to discharge, Figure 4.5 and 4.6 shows the trace plots and the density estimation of the $\boldsymbol{\beta}_0^2$ components, whereas the values of the posterior means and 95% credible regions are shown in Table 4.3.

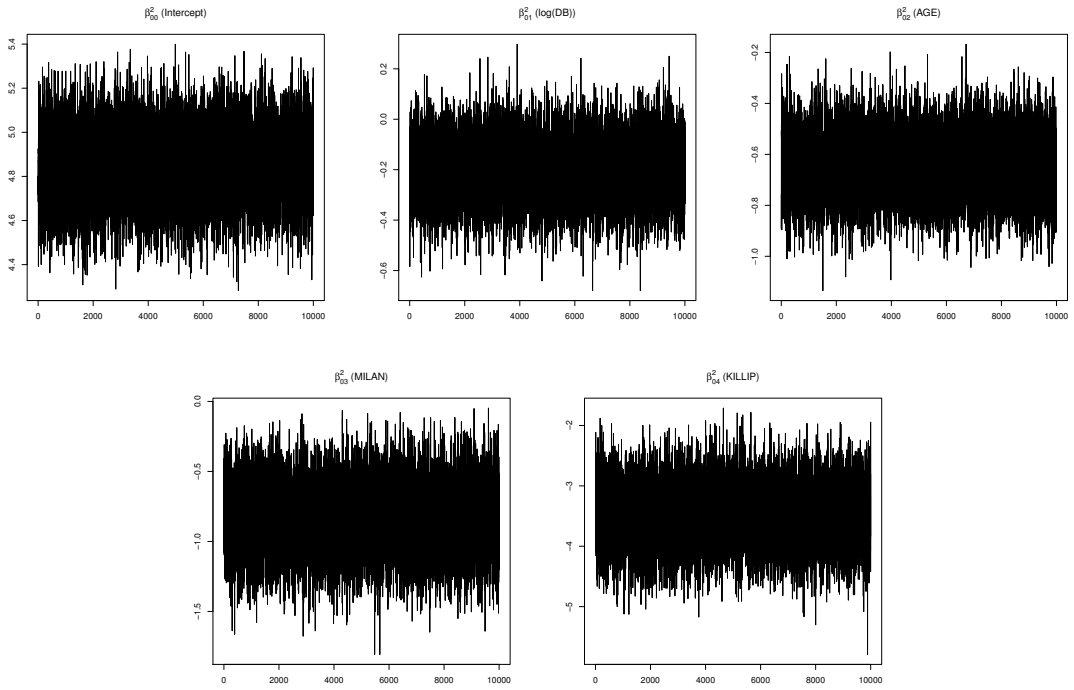


Figure 4.5: Traces of $\boldsymbol{\beta}_0^2$ components.

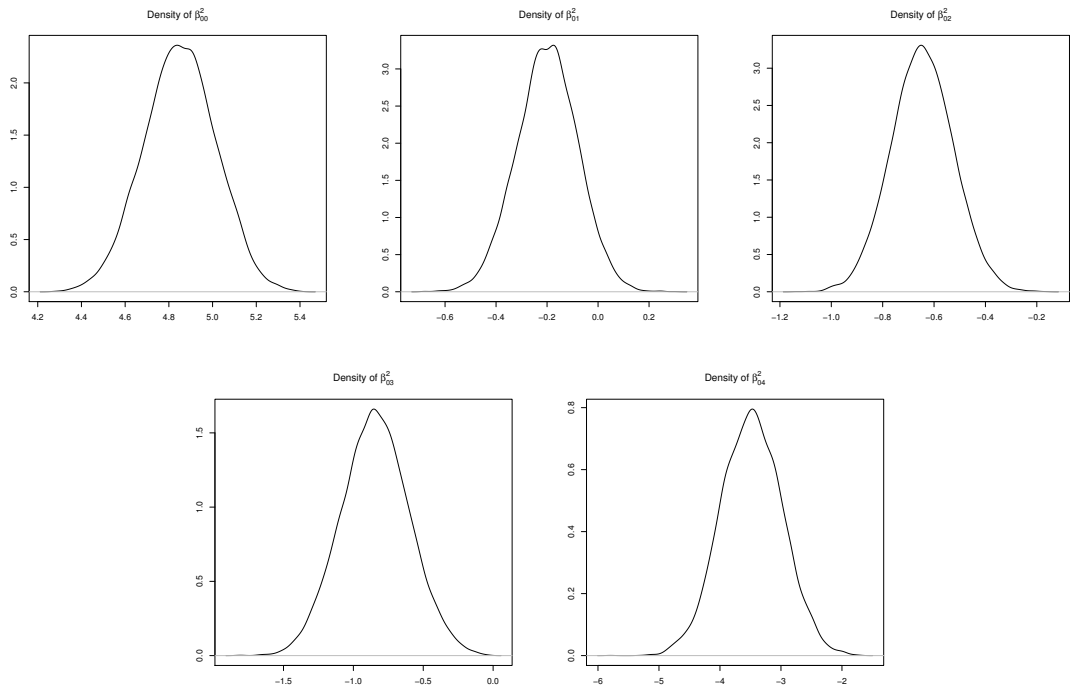


Figure 4.6: Posterior kernel density estimation of $\boldsymbol{\beta}_0^2$ components.

Table 4.3: Posterior means and credible regions for $\boldsymbol{\beta}_0^1$ components.

variable	parameter	mean	2.5%	50%	97.5%
Intercept	β_{00}^2	4.85	4.53	4.85	5.17
$log(DB)$	β_{01}^2	-0.20	-0.44	-0.20	0.03
AGE	β_{02}^2	-0.65	-0.88	-0.65	-0.41
MILAN	β_{03}^2	-0.84	-1.32	-0.84	-0.36
KILLIP	β_{04}^2	-3.48	-4.46	-3.48	-2.49

All the variables we have considered are significant. From Figure 4.6 we have that all parameters are centered on negative values. The DB time gives a negative contribution to the survival to discharge. Moreover, with increasing age, the probability of survival to discharge decreases. Also the severity of the infarction has a strong impact on the survival to discharge: if the severity is high the probability of survival to discharge is low. Finally, it seems that if the patient is treated in an hospital in Milan the survival probability to discharge is lower.

Concerning the third component, the survival to 60 days, Figure 4.7 and 4.8 shows the trace plots and the density estimation of the $\boldsymbol{\beta}_0^3$ components, whereas the values of the posterior means and 95% credible regions are shown in Table 4.4.

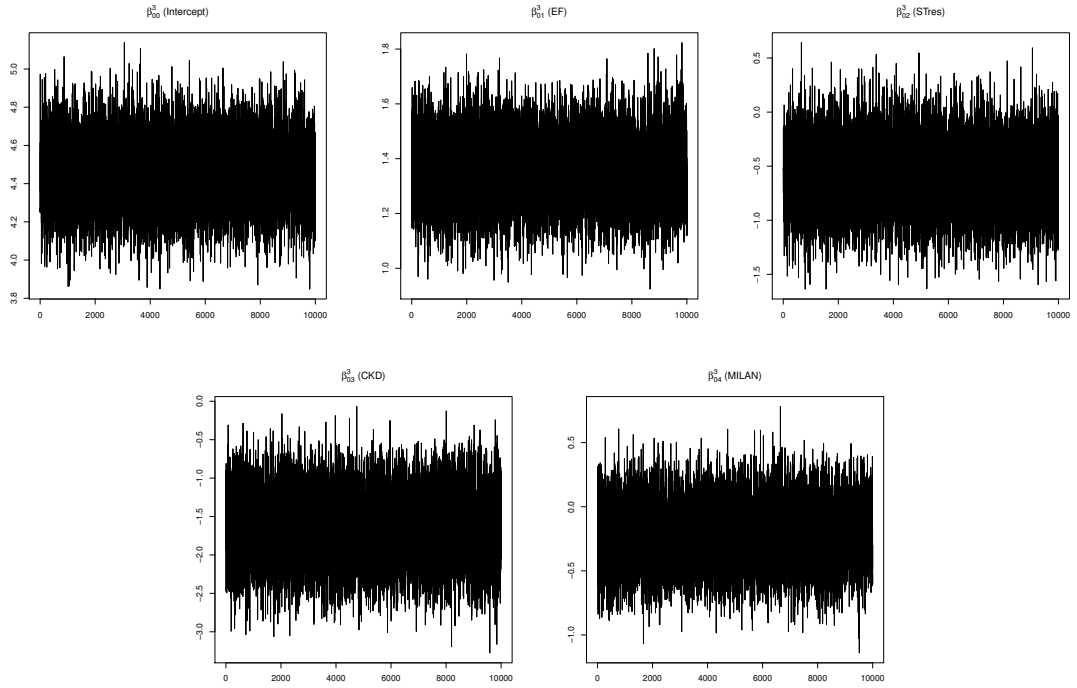


Figure 4.7: Traces of $\boldsymbol{\beta}_0^3$ components.

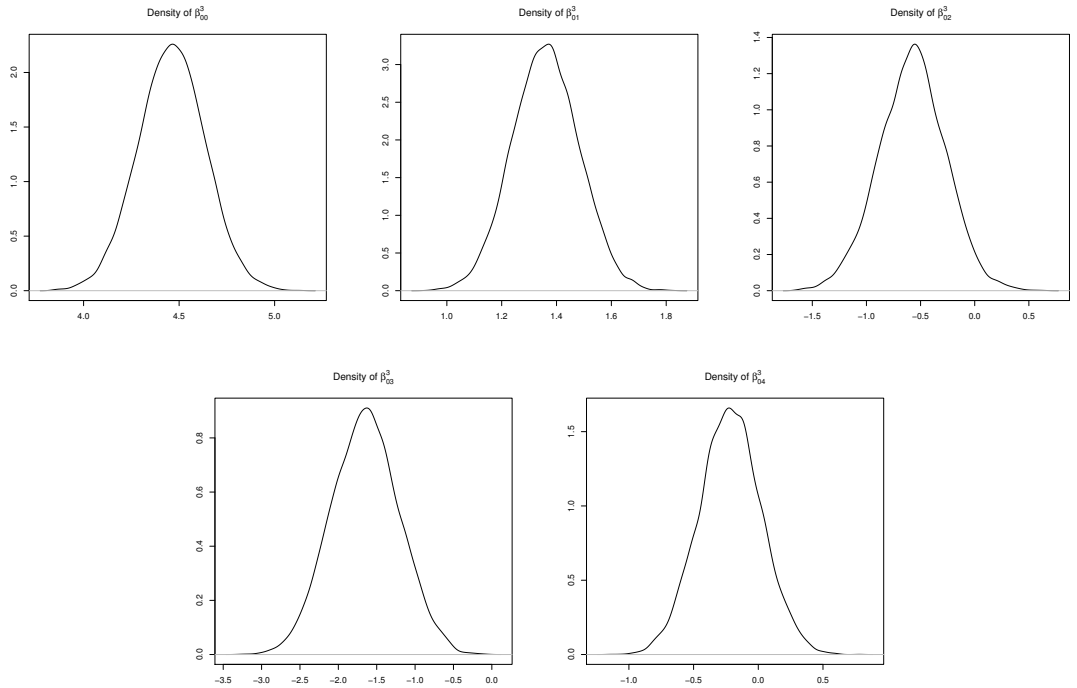


Figure 4.8: Posterior kernel density estimation of $\boldsymbol{\beta}_0^3$ components.

Table 4.4: Posterior means and credible regions for $\boldsymbol{\beta}_0^1$ components.

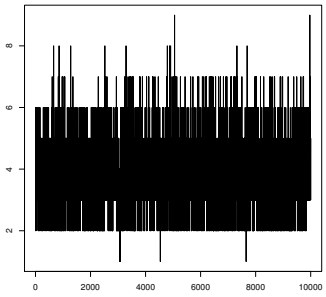
variable	parameter	mean	2.5%	50%	97.5%
Intercept	β_{00}^3	4.46	4.11	4.46	4.81
EF	β_{01}^3	1.36	1.12	1.36	1.60
STres	β_{02}^3	-0.58	-1.20	-0.57	0.02
CKD	β_{03}^3	-1.65	-2.50	-1.65	-0.78
MILAN	β_{04}^3	-0.22	-0.69	-0.22	0.26

Also in this case all the variables in the likelihood are significant. The parameter corresponding to EF is the only one centered on positive numbers. This means that there are more probabilities of survival to 60 days if the value of the ejection fraction increases. STres variable contributes negatively to the survival to 60 days. This is reasonable: if the treatment is not efficient there are less probabilities that the patient survives. Also CKD gives a negative contribution. Finally, observe that, as for the second component, the variable MILAN seems to give a negative contribution to the survival at 60 days, even if the evidence is stronger for the survival to discharge.

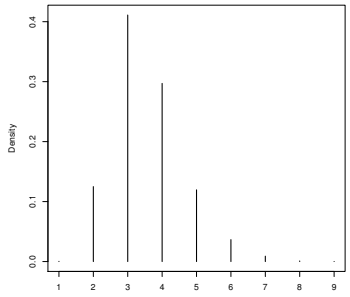
This effect has already been underlined in Guglielmi et al. (2013): epidemiologists confirmed that epidemiology is different between Milan and his neighbourhoods.

4.1.2 Robustness analysis

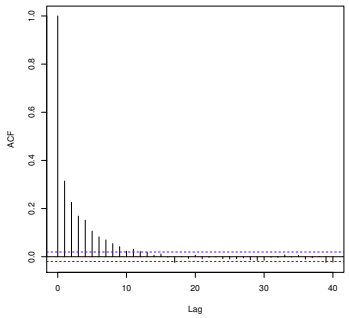
We have performed a sensitivity analysis with respect to the hyperparameters.



(a) Traces of K_n



(b) Posterior number of clusters



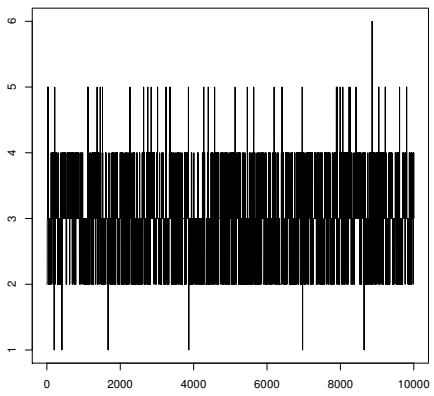
(c) ACF of K_n

Figure 4.9: Markov chain sample of the number of clusters K_n ($\alpha = 100$).

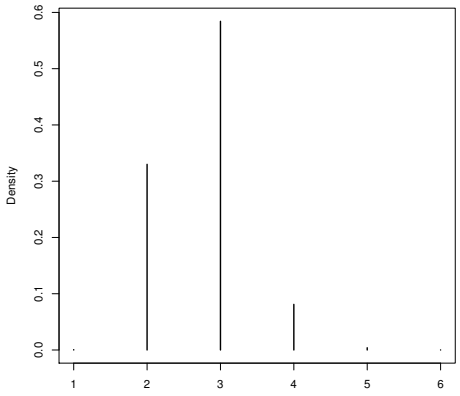
We run several iterations changing the values of the constants k_0^l, the total mass parameter α and the parameters of the Gammas a, a_1 and a_2. All the posterior estimates are very robust, since the estimated values are very similar. Some changes appear for the only K_n. Specifically, if $\alpha = 100$ it seems that K_n converges more quickly, as shown in Figure 4.9. In this case we get $\mathbb{E}(K_n) = 3.56$ and $Var(K_n) = 1.09$. As expected, both mean and variance of K_n are larger than before. A possible reason for which we have faster convergence is that the chain explores the space of the states by varying more frequently, since the variance is larger.

In this Subsection we assume that $k_0^2 = k_0^3 = 1/2$, keeping $k_0^1 = 1/10$ as before. In particular, for the $\boldsymbol{\beta}_0^l$ components we report the trace plots and the tables of the credible regions.

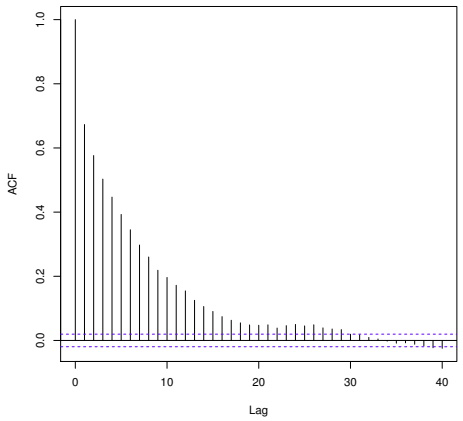
Concerning the number of clusters K_n, Figure 4.10 shows some detail about the posterior estimate and convergence of the Markov chain.



(a) Traces of K_n



(b) Posterior number of clusters

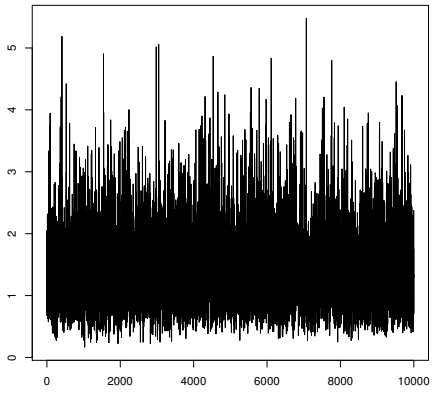


(c) ACF of K_n

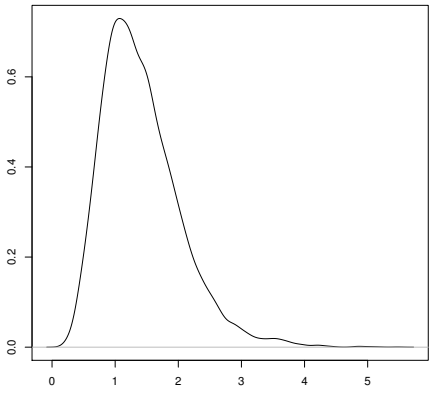
Figure 4.10: Markov chain sample of the number of clusters K_n ($k_0^2 = k_0^3 = 1/2$).

In this case it is $\mathbb{E}[K_n] = 2.76$ and $Var(K_n) = 0.37$.

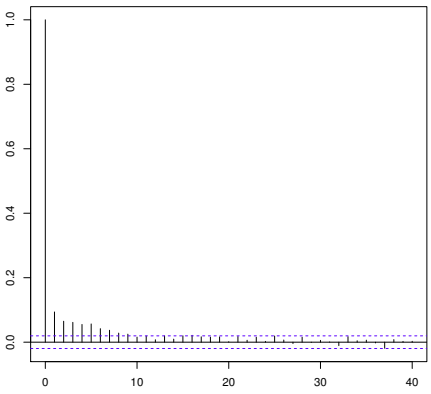
Figure 4.11 shows the trace plot, the density estimation and the auto correlation function of the parameter b.



(a) Traces of b



(b) Posterior kernel density estimation of b



(c) ACF of b

Figure 4.11: Markov chain sample of b ($k_0^2 = k_0 3 = 1/2$).

The posterior mean and variance of b are $\mathbb{E}[b] = 1.42$ and $Var(b) = 0.39$

Concerning the first component, the logarithm of DB, Figure 4.12 shows the trace plots of the $\boldsymbol{\beta}_0^1$ components.

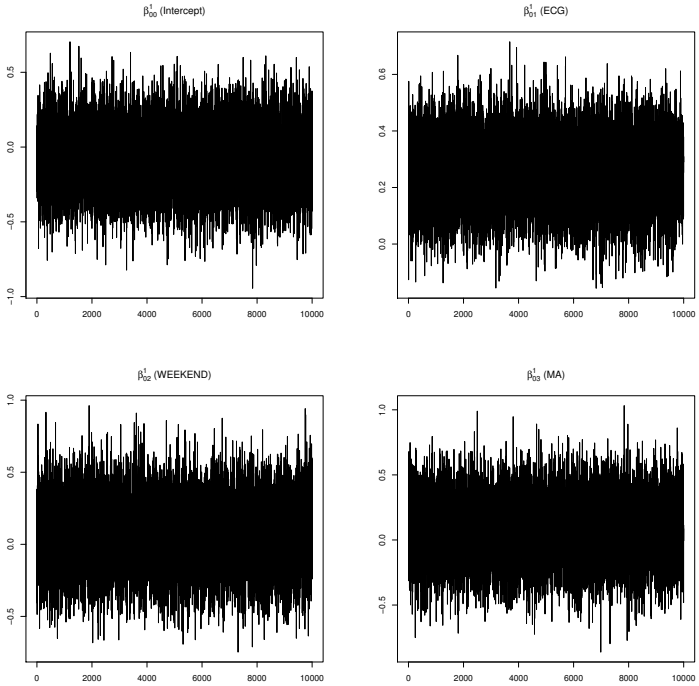


Figure 4.12: Traces of $\boldsymbol{\beta}_0^1$ components ($k_0^2 = k_0^3 = 1/2$).

Table 4.5 shows the values of the a posterior means and 95% credible regions of $\boldsymbol{\beta}_0^1$ components.

Table 4.5: Posterior means and credible regions for $\boldsymbol{\beta}_0^1$ components ($k_0^2 = k_0 3 = 1/2$).

variable	parameter	mean	2.5%	50%	97.5%
Intercept	β_{00}^1	-0.08	-0.49	-0.08	0.34
ECG	β_{01}^1	0.25	0.02	0.25	0.49
WEEKEND	β_{02}^1	0.07	-0.40	0.07	0.53
MA	β_{03}^1	0.07	-0.39	0.07	0.54

Concerning the second component, the survival to discharge, Figure 4.13 shows the trace plots of the $\boldsymbol{\beta}_0^2$ components.

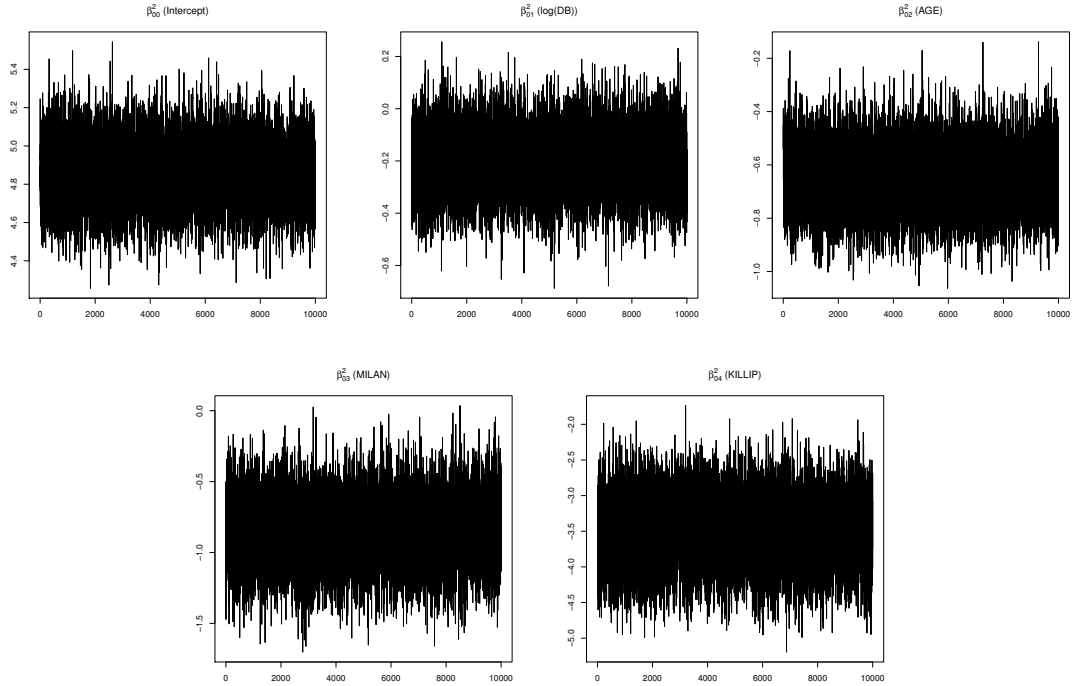


Figure 4.13: Traces of $\boldsymbol{\beta}_0^2$ components ($k_0^2 = k_0^3 = 1/2$).

Table 4.6 shows the values of the a posterior means and 95% credible regions of $\boldsymbol{\beta}_0^2$ components.

Table 4.6: Posterior means and credible regions for $\boldsymbol{\beta}_0^1$ components ($k_0^2 = k_0 3 = 1/2$).

variable	parameter	mean	2.5%	50%	97.5%
Intercept	β_{00}^2	4.86	4.54	4.86	5.18
$log(DB)$	β_{01}^2	-0.20	-0.43	-0.20	0.03
AGE	β_{02}^2	-0.64	-0.88	-0.64	-0.41
MILAN	β_{03}^2	-0.85	-1.32	-0.85	-0.39
KILLIP	β_{04}^2	-3.51	-4.44	-3.52	-2.60

Concerning the third component, the survival to 60 days, Figure 4.14 shows the trace plots of the $\boldsymbol{\beta}_0^3$ components.

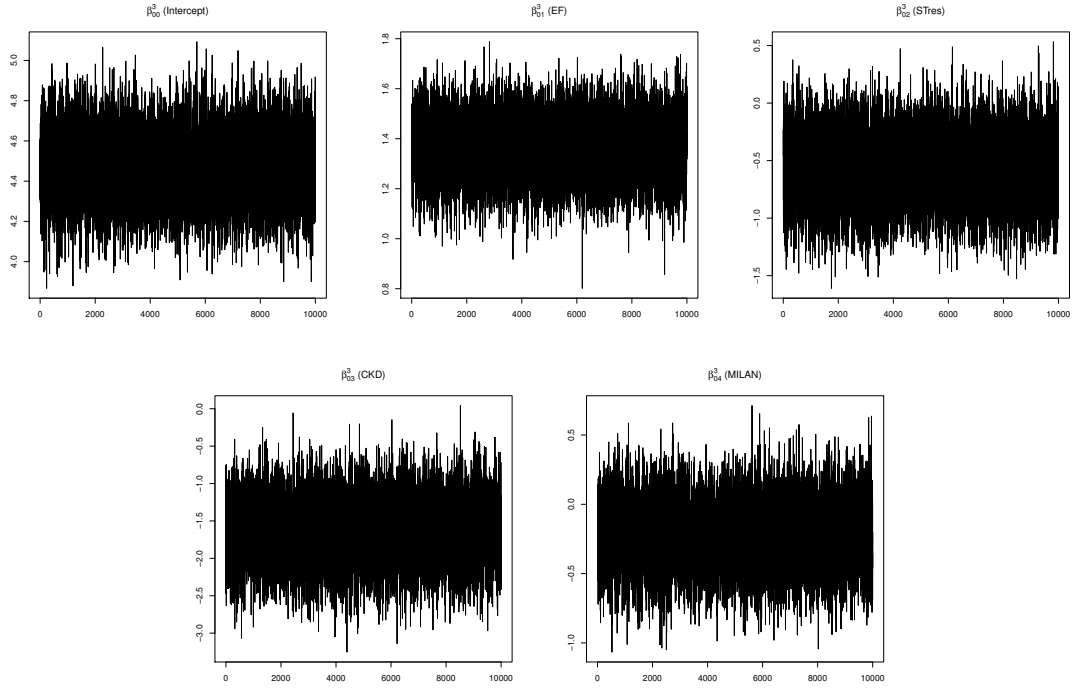


Figure 4.14: Traces of $\boldsymbol{\beta}_0^3$ components ($k_0^2 = k_0 3 = 1/2$).

Table 4.7 shows the values of the a posterior means and 95% credible regions of $\boldsymbol{\beta}_0^3$ components.

Table 4.7: Posterior means and credible regions for $\boldsymbol{\beta}_0^1$ components ($k_0^2 = k_0 3 = 1/2$).

variable	parameter	mean	2.5%	50%	97.5%
Intercept	β_{00}^3	4.47	4.14	4.47	4.81
EF	β_{01}^3	1.36	1.13	1.36	1.59
STres	β_{02}^3	-0.57	-1.15	-0.58	0.01
CKD	β_{03}^3	-1.64	-2.46	-1.64	-0.83
MILAN	β_{04}^3	-0.22	-0.69	-0.22	0.24

As mentioned before, we notice that all the posterior estimates are very robust. The only different thing that can be observed is the behaviour of K_n. Mean and variance are slightly higher than before and the mode is equal to 3 instead of 2.

4.1.3 Prediction

With reference to the hyperparameters fixed in the "benchmark" prior at the beginning of the Chapter, we are interested in making prediction for the response of a new unit that presents some combination of covariates of interest. This can be done in the same Gibbs sampler and the idea is very similar to what we do in the first step of Neal's algorithm (see equation (3.18)): let x^{n+1} denote the vector of covariate values where we want prediction (could be an observed or unobserved combination). Keeping the same notations of Chapter 3, proceed to allocate this new individual to a new cluster or in an existing one with the following probabilities:

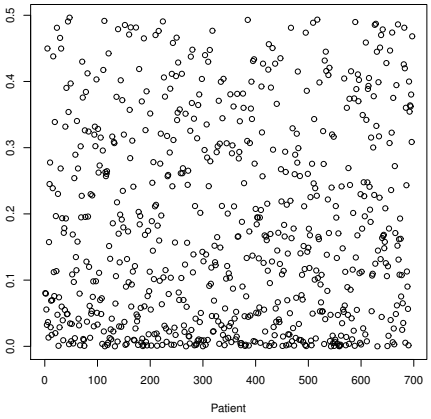
$$P(e_{n+1} = j | \, \phi_1, \ldots, \phi_h) \propto \begin{cases} \frac{c(S_j \cup \{n+1\})}{c(S_j)} \frac{g(x_j^* \cup \{x_{n+1}\})}{g(\{x_j^*\})} & \text{for } 1 \leq j \leq k^- \\ c(\{n+1\}) g(\{x_{n+1}\}) & \text{for } k^- < c \leq h \end{cases} \tag{4.4}$$

Note that only the cohesion and similarity functions $c(\cdot)$ and $g(\cdot)$ contribute in giving a label to the new unit. Once assigned the new individual to the cluster j, we associate to him cluster specific parameters $\boldsymbol{\theta}_j$. We repeat this for all the iterations. Therefore, it is possible generate the three dimensional response $\boldsymbol{Y}_{n+1}$ from a Normal and two Bernoulli distributions with appropriate parameters (as described in Section 3.1.1).

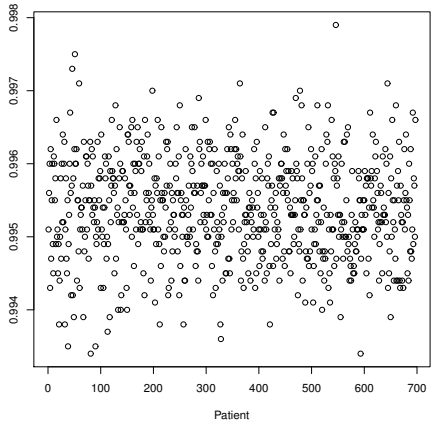
In this way, the probabilities of survival to discharge and at 60 days of the $(n+1)$th unit (denoted with p_{n+1} and r_{n+1}), can be estimated by the ergodic means.
Concerning the first response variable, an alternative way of displaying the prediction is to compute the predictive density of the first response. That is, we can compute, on an appropriate grid, the values that the Normal distribution with actualized parameters assumes.

Prediction for patients in the dataset

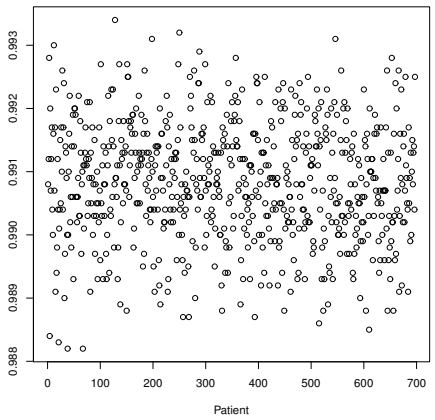
We first consider the 697 units in the dataset (i.e. we used the data twice). In this case, concerning the variable $log(DB)$ we calculated the Bayesian p-values for every unit. This is possible since the true values of the variable Y_{i1} are available for every i. Later we will do prediction for unobserved units. In this case the values of Y_{i1} are not available, so that we will proceed calculating the predictive density for every unit of interest. Figure 4.15 shows the Bayesian p-values together with the predictive probabilities of survival to discharge and at 60 days for every patient in the dataset.



(a) Bayesian p-values



(b) Predictive probabilities of survival to discharge



(c) Predictive probabilities of survival at 60 days

Figure 4.15: Bayesian p-values for Y_{i1} and predictive survival probabilities for Y_{i2} and Y_{i3}.

The number of units with Bayesian p-value lesser than 0.05 is 204 (29.3%). We observe that the predictive probabilities of survival for every patient are all high and near to 1.

Prediction for patients with a combination of covariates of interest

We want now to predict the behaviour of unobserved units. In order to do this, we have to create a dataset with units presenting covariates of interest. In particular, we choose to fix some covariate in the likelihood, assuming as missing all the other ones in the similarity. We select the most significant covariates of interest based on the results that we get in the previous section, i.e. all continuous (and standardized) covariates (ECG, AGE and EF) and some binary one (KILLIP, MILAN, CKD and STres). All selected binary covariates assume the value 0 or 1. We let ECG assume values in $\{-4, 0, 4\}$, whereas AGE and EF assume values in $\{-2, 0, 2\}$. So the number of (unobserved) units in the new dataset generated is equal to $3^3 \times 4^2 = 432$.

Concerning the $log(DB)$, as mentioned above, we are going to calculate the predictive densities for every unit $i = 1, \ldots, 432$ on an appropriate grid of values. The problem that

we have to solve is to find a common range of values for which all the densities are considerably larger than 0. A good choice seems to be the interval $[-4, 4]$. Figure 4.16 shows the predictive densities, distinguishing by the three different values of ECG variable established before.

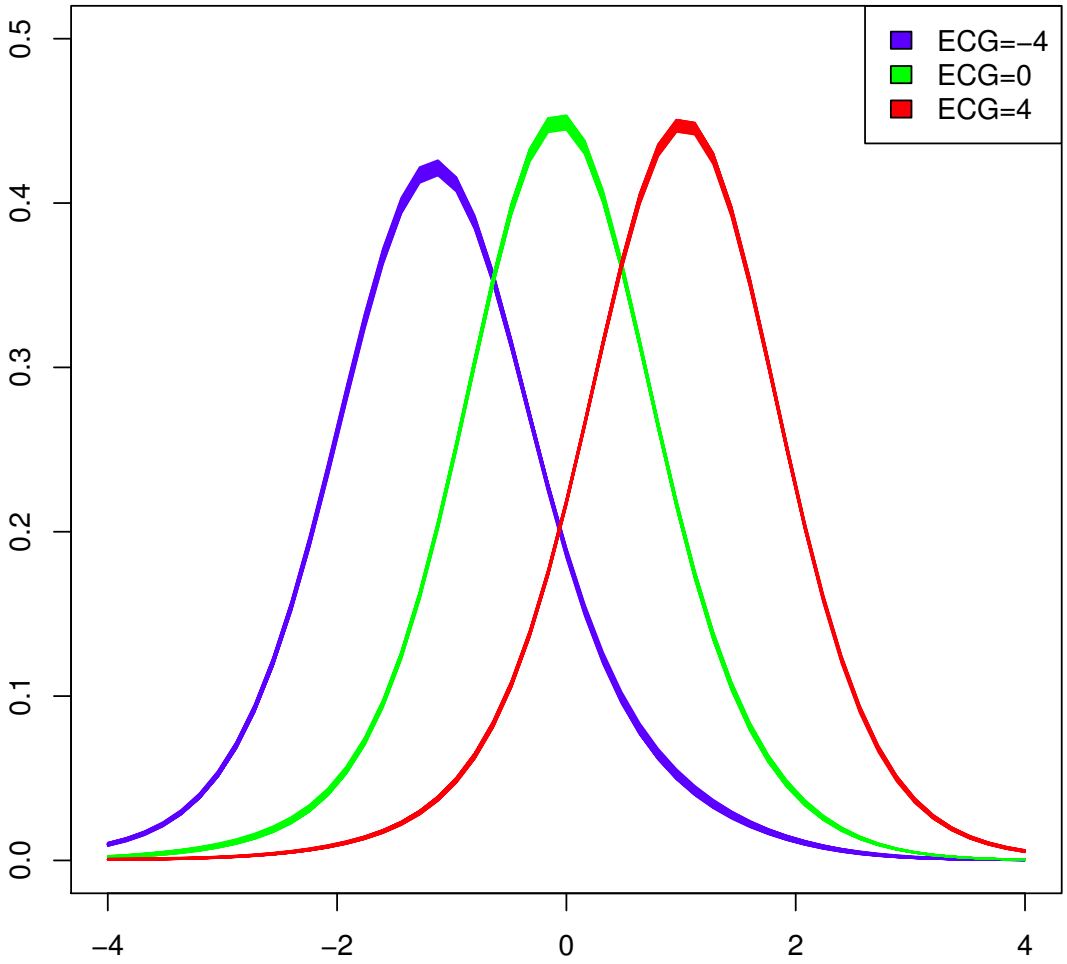
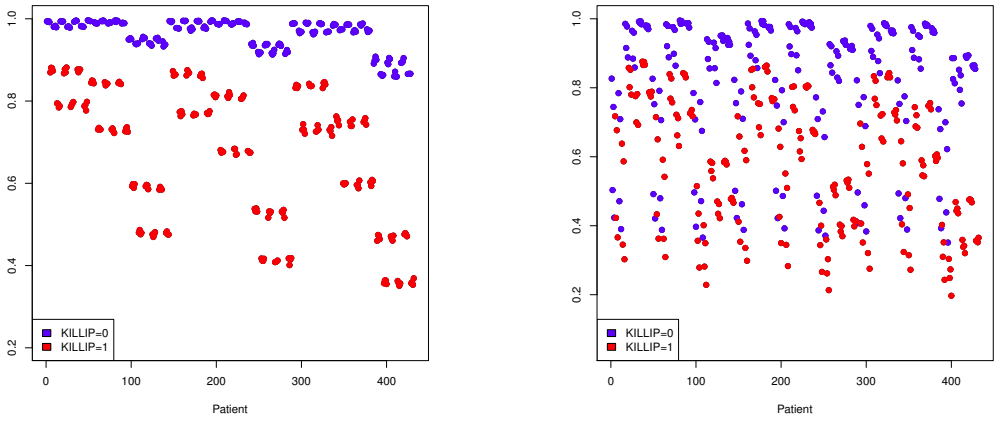


Figure 4.16: Predictive posterior densities of $log(DB)$ for the 432 units generated.

It is clear that the variable ECG explains perfectly the three groups in evidence in the above figure. Indeed, the numerosities of each group are exactly 432/3=144.

Concerning the survival to discharge and at 60 days, Figure 4.17 shows the predictive probabilities, distinguishing by patients for which the severity of the infarction is high (KILLIP=1) or low (KILLIP=0).



(a) Predictive probabilities of survival to discharge.

(b) Predictive probabilities of survival at 60 days.

Figure 4.17: Predictive probabilities of survival for the 432 units generated (divided by KILLIP).

From Figure 4.17a we can clearly identify separated groups of units with higher or lower probability of survival to discharge. The same holds for Figure 4.17b relatively to the probability of survival at 60 days, even if with less evidence. As expected, the variable KILLIP explains correctly these groups. In practice, the strong significance of the KILLIP variable with respect to the survival to discharge is confirmed. We observe that the units with the highest probabilities of survival to discharge are those for which the severity of the infarction is not high (Figure 4.17a).

4.1.4 Clustering

Each iteration of the Markov chain yields a clustering of the patients. We want now to find a method by which we can get a cluster estimate of the data. Let's denote with $\boldsymbol{e}_1, \ldots, \boldsymbol{e}_B$ the posterior clustering distribution obtained using MCMC, where B is the number of sampled clusterings (i.e. B is the final sample size). Every $\boldsymbol{e}_i$ is a vector of labels of dimension n. Usually, for every clustering $\boldsymbol{e} \in \{\boldsymbol{e}_1, \ldots, \boldsymbol{e}_B\}$, an association matrix $\delta(\boldsymbol{e})$ of dimension $n \times n$ can be formed whose (i, j) element is $\delta_{i,j}(\boldsymbol{e})$, an indicator of whether element i is clustered with element j. Element-wise averaging of these associations matrices yields the pairwise probability matrix of clustering, denoted with $\hat{\pi}$.

In the literature one can find various approaches. As summarized in Dahl (2006), Medvedovic and Sivaganesan (2002) suggest forming a clustering estimate by using the pairwise probability matrix $\hat{\pi}$ as a distance matrix in hierarchical agglomerative clustering. It seems counterintuitive, however, to apply an ad hoc clustering method on top of a model which itself produces clustering.
Perhaps the simplest method is to select the observed clustering that maximizes the density of the posterior clustering distribution. This is known as the maximum a posteriori

(MAP) clustering. Unfortunately, the MAP clustering may only be slightly more probable than the next best alternative, yet represent a very different allocation of observations.

Dahl (2006) introduces the *least-square model-based clustering*, a new method for estimating the clustering of observations using draws from a posterior clustering distributions. The method is based on the pairwise probability matrix $\hat{\pi}$ that units are clustered together, but, unlike in Medvedovic and Sivaganesan (2002), here it selects one of the observed clusterings in the Markov chain as the point estimate. Specifically, the last square clustering $\boldsymbol{e}_{ls}$ is the observed clustering $\boldsymbol{e}$ which minimizes the sum of squared deviations of its association matrix $\delta(\boldsymbol{e})$ from the pairwise probability matrix $\hat{\pi}$:

$$\boldsymbol{e}_{LS} = \underset{\boldsymbol{e}\in\{\boldsymbol{e}_1,\ldots,\boldsymbol{e}_B\}}{\arg\min} \sum_{i=1}^{n}\sum_{j=1}^{n} \left(\delta_{i,j}(\boldsymbol{e}) - \hat{\pi}_{i,j}\right)^2 . \tag{4.5}$$

The least square clustering has the advantage that it uses information from all the clusterings (via the pairwise probability matrix) and is intuitively appealing because it selects the "average" clustering, instead of forming a clustering via an external, ad hoc clustering algorithm.

We computed the least-square model-based cluster estimate as in (4.5): we obtained 3 groups as reported in Table 4.8. In order to display the estimated clusters, we plot the points in the plane (the same color means units in the same cluster) defined by two continuous covariates. Available continuous variable are ECG, AGE, EF and DB (in logarithmic scale, even if it is not a covariate), so that all possible combinations of plots are $\binom{4}{2} = 6$. However we consider only the 3 plots where the variable DB appears, since it seems that we get a good clustering. Figure 4.18 shows this three plots (with appropriate scales), distinguishing between patients deceased to discharge, at 60 days and the remaining survived, whereas Table 4.8 shows the frequencies in the groups.

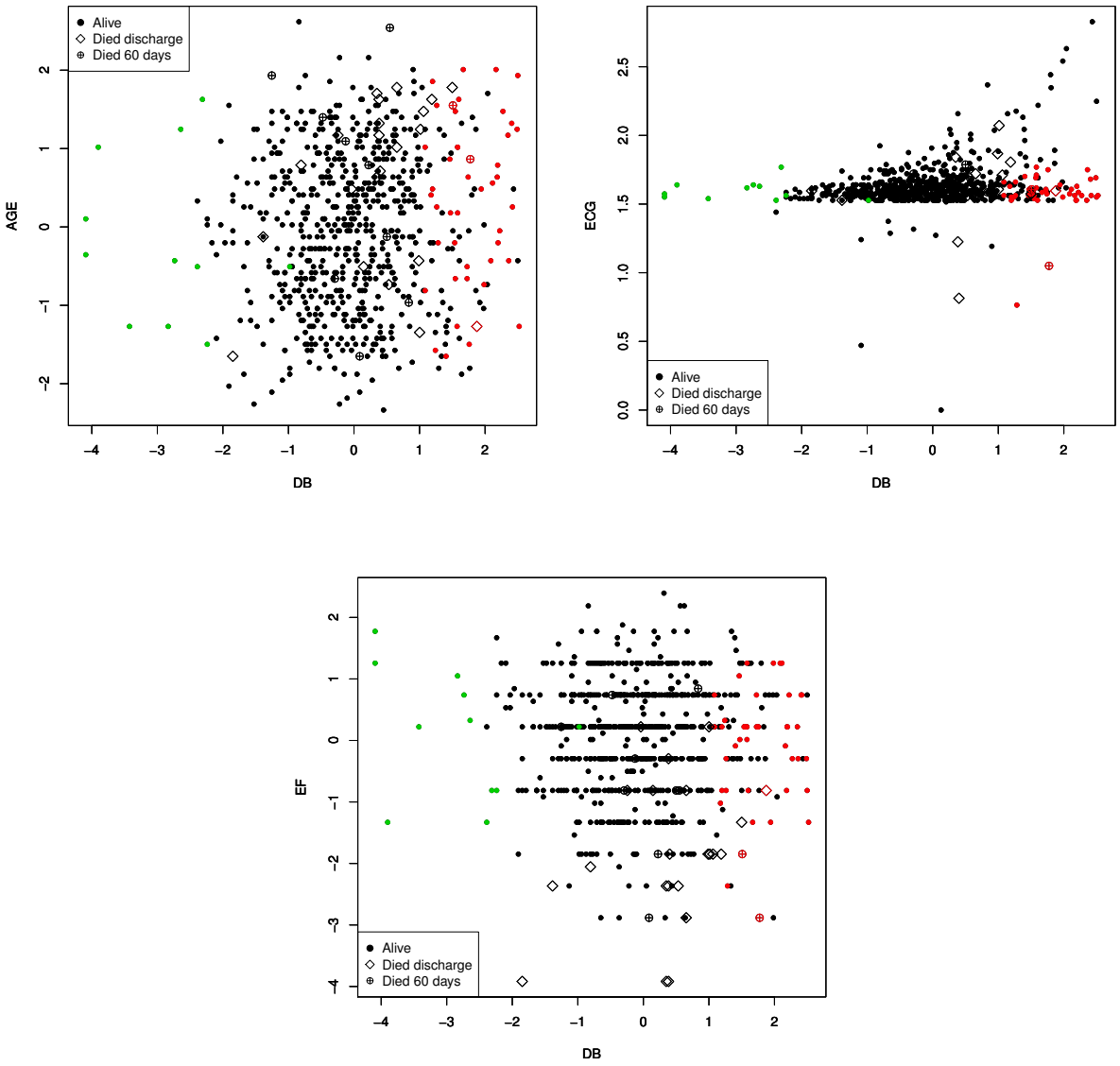


Figure 4.18: Plots of AGE, ECG and EF with respect to $log(DB)$ highlighting the clusters.

Table 4.8: Numerosities of the groups.

Labels	1	2	3
Numerosities	641	45	11
Color	Black	Red	Green

It is clear that it is the response DB that drives the clusters. Moreover, it does not seems to be a relationship between the clustering and patients survival.

In order to make some comparison, we are now going to show the same plot but for different values of the hyperparameters. Figure 4.19 and 4.20 show the clustering obtained when $k_0^2 = k_0^3 = 1/2$ and $\alpha = 100$, respectively, whereas Tables 4.9 and 4.10 show the respective groups numerosities.

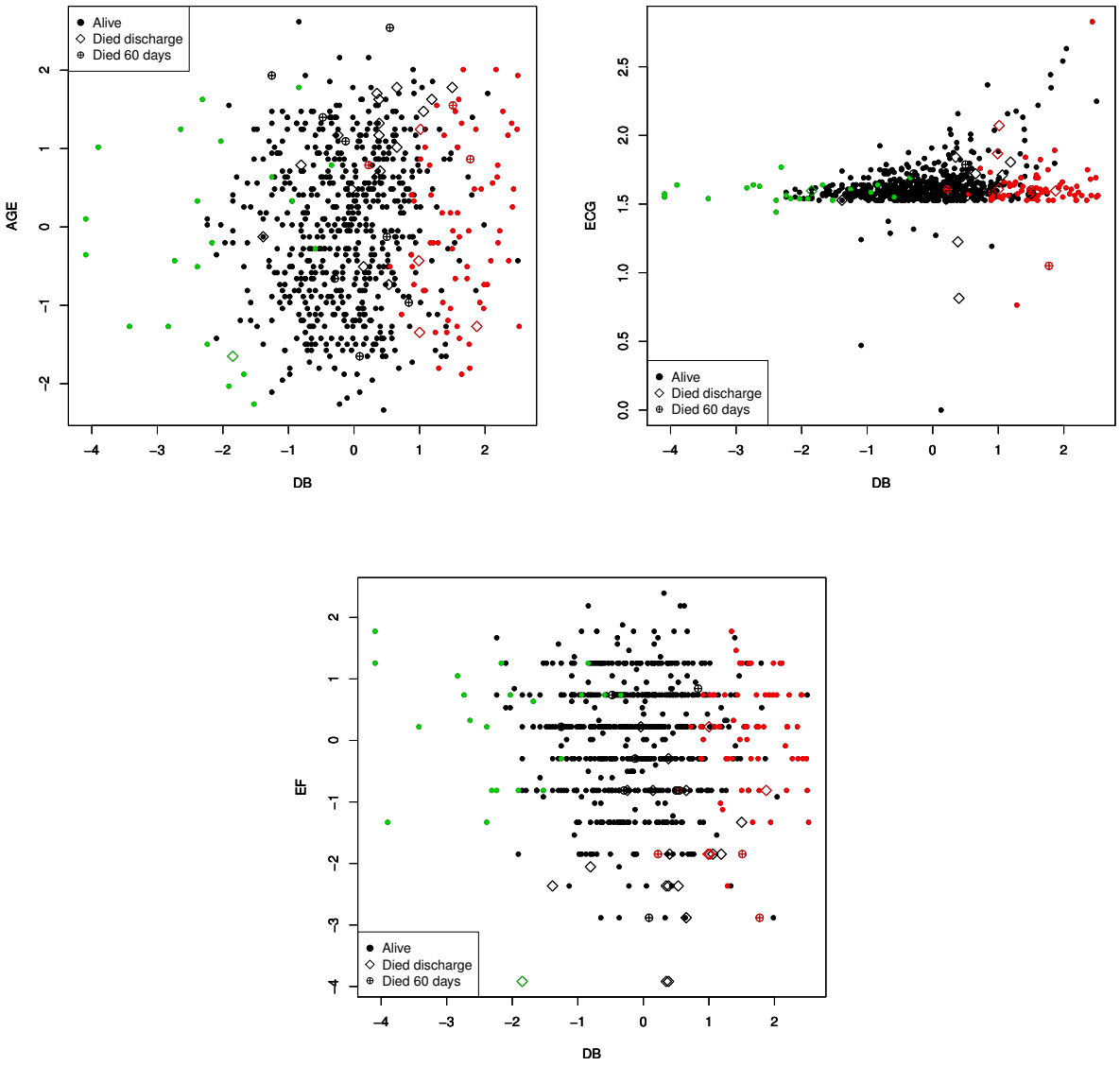


Figure 4.19: Plots of AGE, ECG and EF with respect to $log(DB)$ highlighting the clusters ($k_0^2 = k_0^3 = 1/2$).

Table 4.9: Numerosities of the groups ($k_0^2 = k_0^3 = 1/2$).

Labels	1	2	3
Numerosities	600	75	22
Color	Black	Red	Green

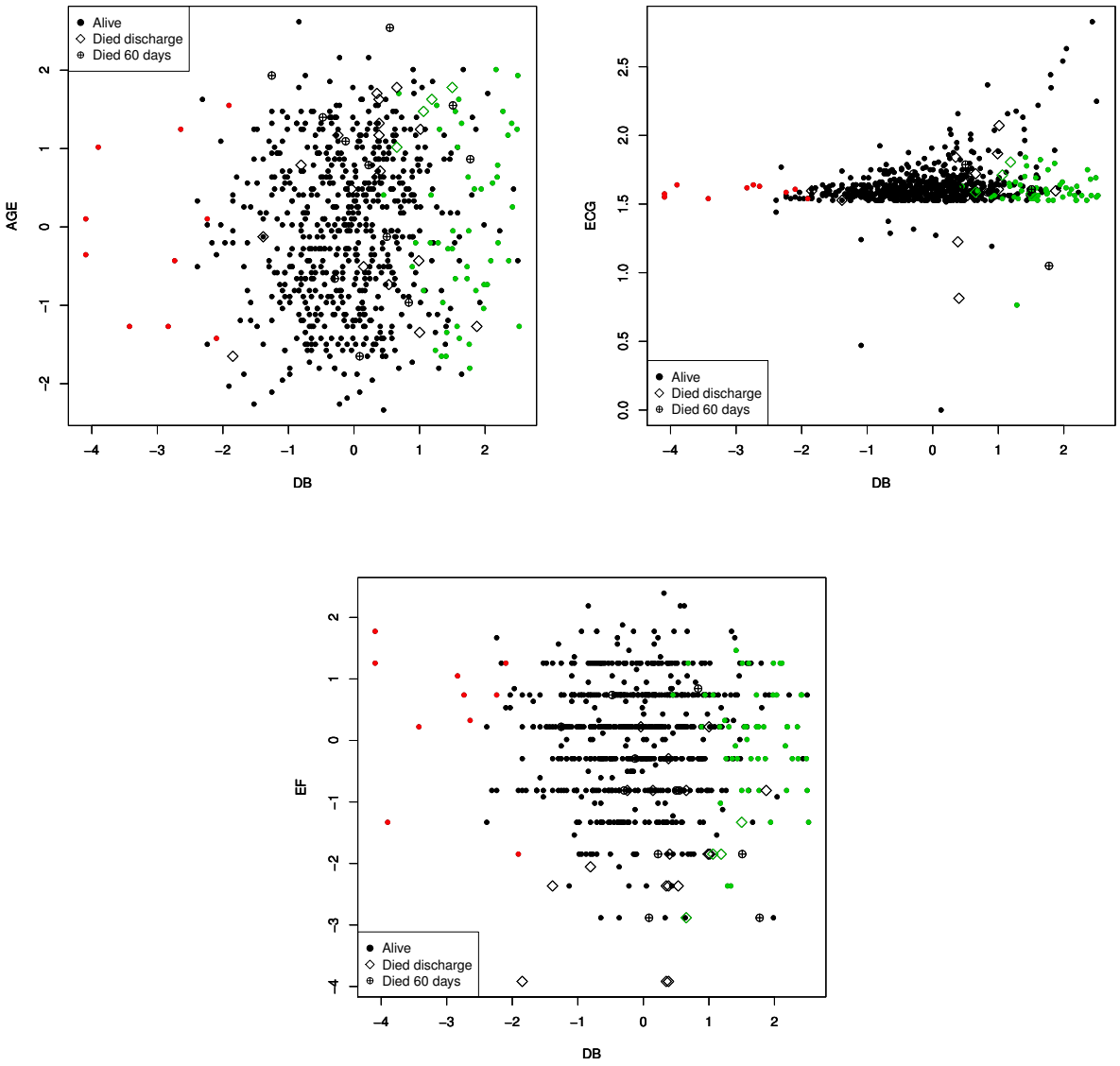


Figure 4.20: Plots of AGE, ECG and EF with respect to $log(DB)$ highlighting the clusters ($\alpha = 100$).

Table 4.10: Numerosities of the groups ($\alpha = 100$).

Labels	1	2	3
Numerosities	632	10	55
Color	Black	Red	Green

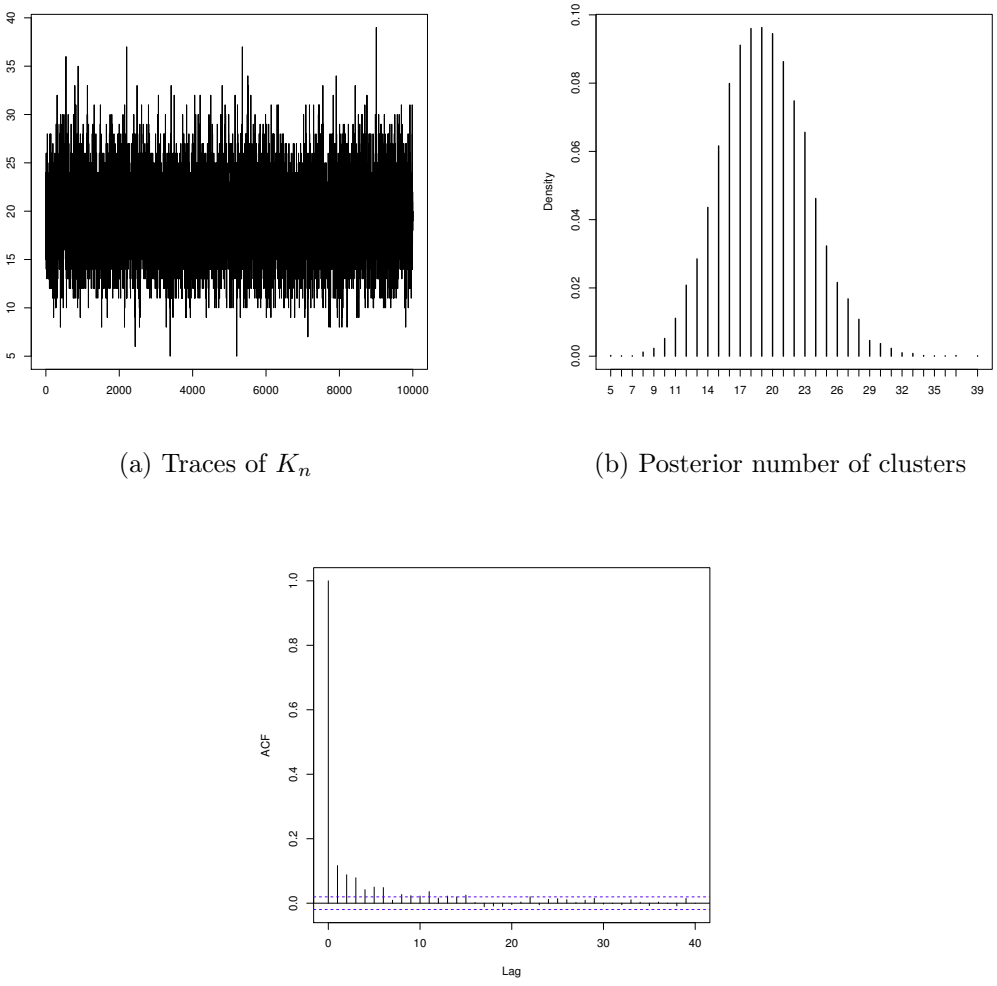
It seems that increasing the values of k_0^2 and k_0^3 (i.e. decreasing the a priori variance of cluster specific parameters $\boldsymbol{\beta}_j^2$ and $\boldsymbol{\beta}_j^3$) the numerosities of the detected groups by Dahl's algorithm increase too. The same result holds with increasing the value of the total mass parameter α: increasing α seems to create larger clusters, but not more of them. Apparently, the choice of this parameter is not strongly relevant.

4.2 Comparison with the PPM model

Finally, we are interested in comparing the PPMx model with the PPM model, i.e. to understand what inference we get if we do not assume a prior depending on covariates. As mentioned in Section 1.3, the PPM coincides exactly with a DP in the sense that the marginal distribution that a DP induces on partition is also a PPM with cohesions $c(S_j) = \alpha(|S_j|-1)!$, as in our case. Roughly speaking, we do not consider the similarity functions in equation (3.18). In such a way we consider only the covariates in the likelihood, and the same number $n = 697$ of patients.
To make comparison we keep the same values of hyperparameters fixed at the begin of the chapter in the "benchmark" prior.

4.2.1 Posterior estimates of global parameters

Concerning the number of clusters K_n, Figure 4.21 shows some detail about the posterior estimate and convergence of the Markov chain.



(a) Traces of K_n

(b) Posterior number of clusters

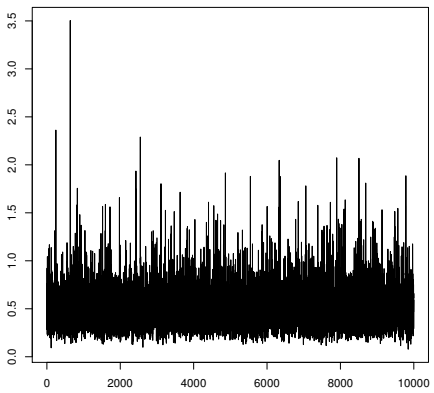
(c) ACF of K_n

Figure 4.21: Markov chain sample of the number of clusters K_n in the PPM.

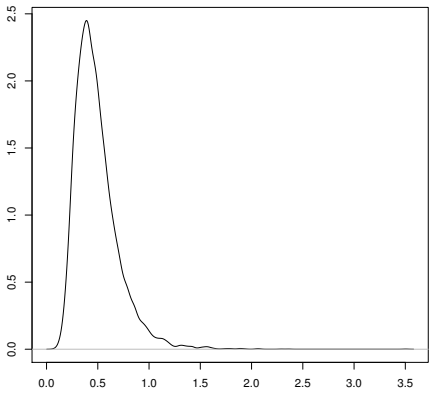
We get $\mathbb{E}[K_n] = 19.24$ and $Var(K_n) = 16.08$. It is clear that the Markov chain

converges. If we get rid of the covariates in the prior for the partition parameter, the convergence is faster. Also, posterior mean and variance are pretty different and larger. This is related to the value of the total mass parameter α that, as we will see later, here seems to be more relevant.

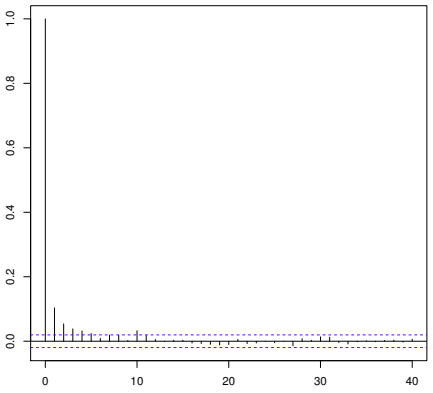
Figure 4.22 shows the trace plot, the density estimation and the auto correlation function of the parameter b.



(a) Traces of b



(b) Posterior kernel density estimation of b



(c) ACF of b

Figure 4.22: Markov chain sample of b in the PPM.

The posterior mean and variance of b are $\mathbb{E}[b] = 0.49$ and $Var(b) = 0.05$. These values are quite different from those obtained in the PPMx model.

Concerning the first component, the logarithm of DB, Figure 4.23 and 4.24 shows the trace plots and the density estimation of the $\boldsymbol{\beta}_0^1$ components, whereas Table 4.11 shows the values of the posterior means and 95% credible regions distinguishing between the PPM and the PPMx model.

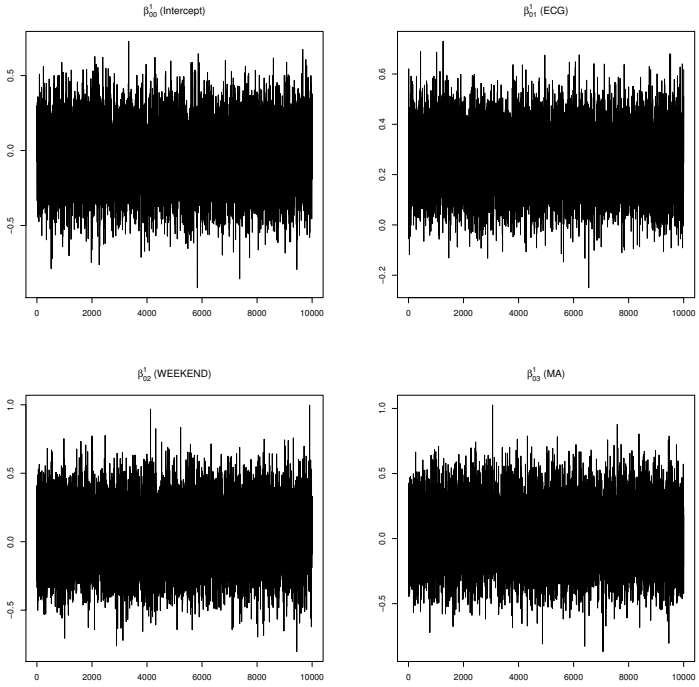


Figure 4.23: Traces of $\boldsymbol{\beta}_0^1$ components in the PPM.

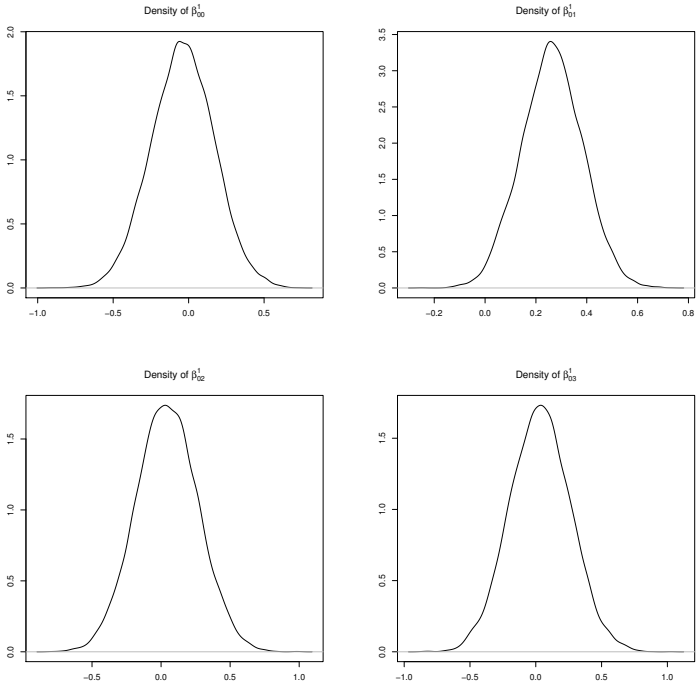


Figure 4.24: Posterior kernel density estimation of $\boldsymbol{\beta}_0^1$ components in the PPM.

Table 4.11: Posterior means and credible regions for $\boldsymbol{\beta}_0^1$ components, comparison between PPMx and PPM.

		PPMx				PPM			
variable	parameter	mean	2.5%	50%	97.5%	mean	2.5%	50%	97.5%
Intercept	β_{00}^1	-0.08	-0.50	-0.08	0.34	-0.04	-0.45	-0.04	0.37
ECG	β_{01}^1	0.25	0.02	0.25	0.49	0.26	0.03	0.26	0.50
WEEKEND	β_{02}^1	0.08	-0.39	0.08	0.54	0.04	-0.39	0.04	0.49
MA	β_{03}^1	0.07	-0.41	0.07	0.53	0.03	-0.41	0.03	0.49

Concerning the second component, the the survival to discharge, Figure 4.25 and 4.26 shows the trace plots and the density estimation of the $\boldsymbol{\beta}_0^2$ components, whereas Table 4.12 shows the values of the posterior means and 95% credible regions distinguishing between the PPM and the PPMx model.

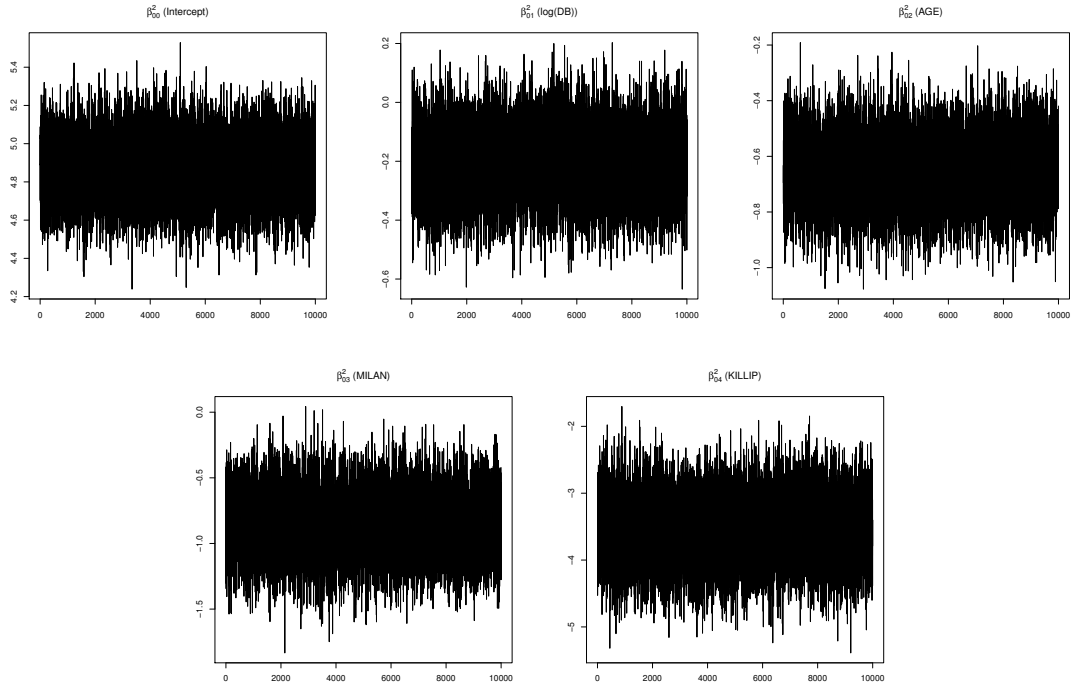


Figure 4.25: Traces of $\boldsymbol{\beta}_0^2$ components in the PPM.

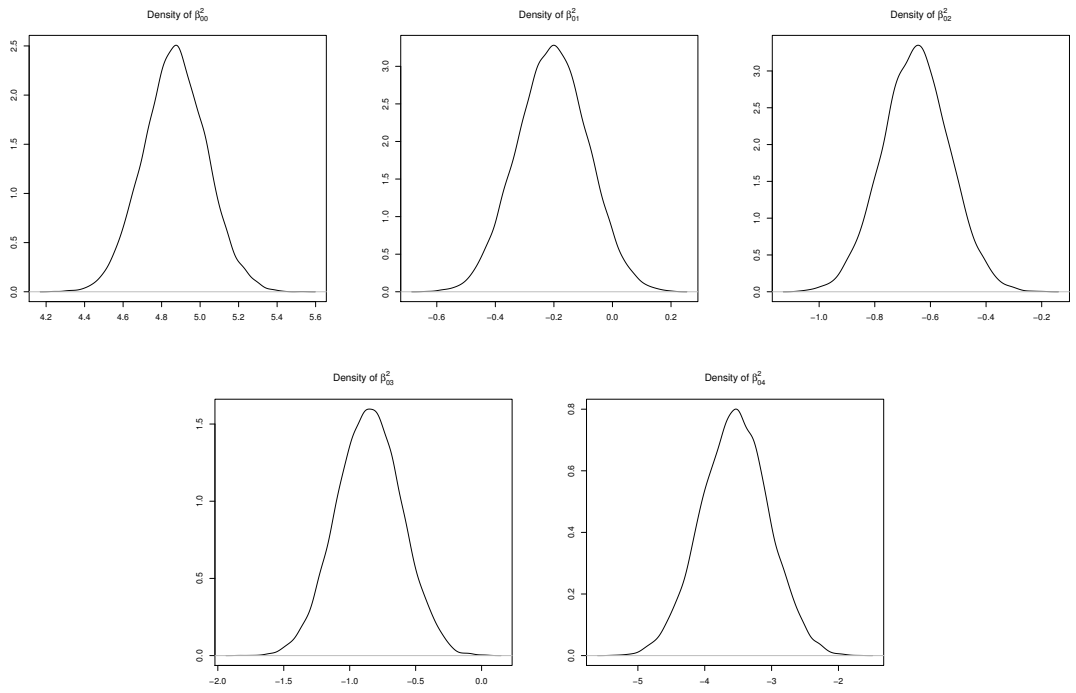


Figure 4.26: Posterior kernel density estimation of $\boldsymbol{\beta}_0^2$ components in the PPM.

Table 4.12: Posterior means and credible regions for $\boldsymbol{\beta}_0^2$ components, comparison between PPMx and PPM.

		PPMx				PPM			
variable	parameter	mean	2.5%	50%	97.5%	mean	2.5%	50%	97.5%
Intercept	β_{00}^2	4.85	4.53	4.85	5.17	4.87	4.55	4.87	5.19
$log(DB)$	β_{01}^2	-0.20	-0.44	-0.20	0.03	-0.20	-0.44	-0.20	0.03
AGE	β_{02}^2	-0.65	-0.88	-0.65	-0.41	-0.65	-0.89	-0.65	-0.41
MILAN	β_{03}^2	-0.84	-1.32	-0.84	-0.36	-0.85	-1.33	-0.85	-0.37
KILLIP	β_{04}^2	-3.48	-4.46	-3.48	-2.49	-3.55	-4.53	-3.55	-2.57

Concerning the third component, the the survival to 60 days, Figure 4.27 and 4.28 shows the trace plots and the density estimation of the $\boldsymbol{\beta}_0^3$ components, whereas Table 4.13 shows the values of the posterior means and 95% credible regions distinguishing between the PPM and the PPMx model.

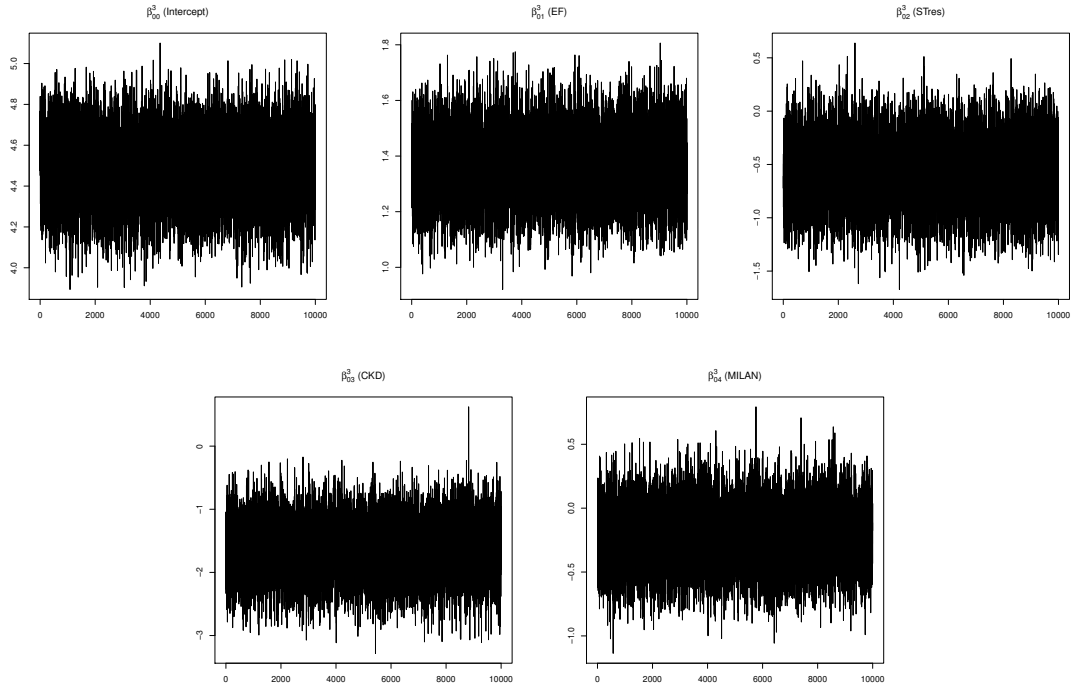


Figure 4.27: Traces of $\boldsymbol{\beta}_0^3$ components in the PPM.

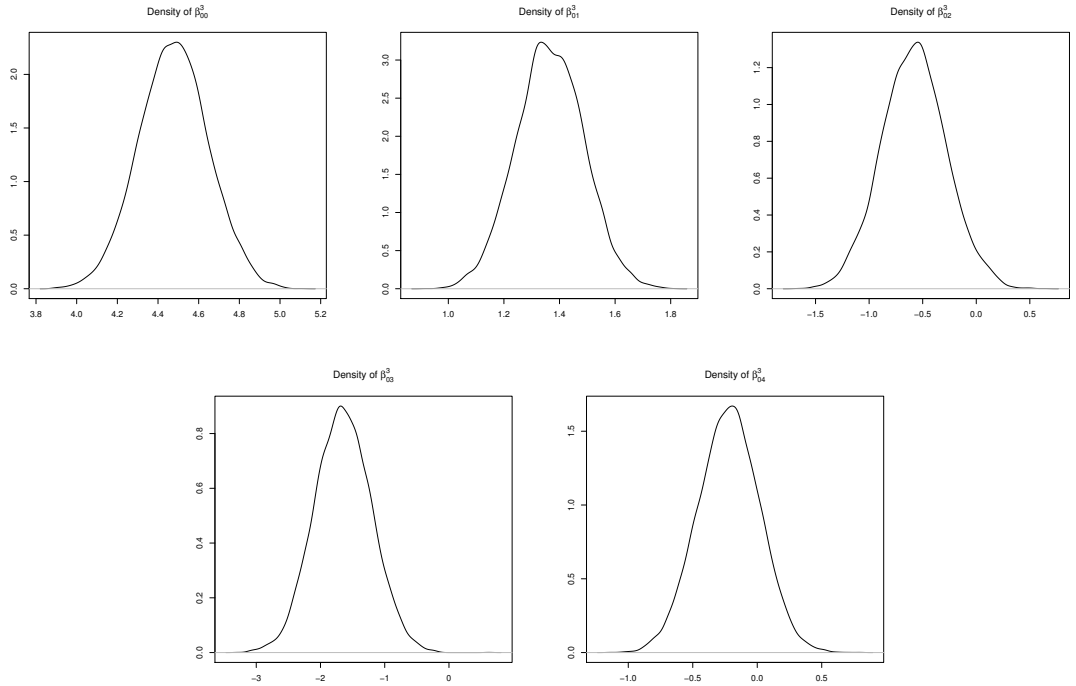


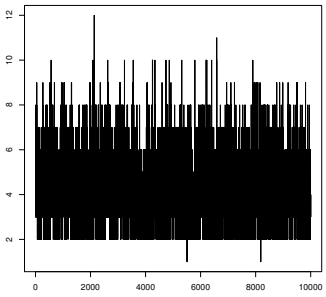
Figure 4.28: Posterior kernel density estimation of $\boldsymbol{\beta}_0^3$ components in the PPM.

Table 4.13: Posterior means and credible regions for $\boldsymbol{\beta}_0^3$ components, comparison between PPMx and PPM.

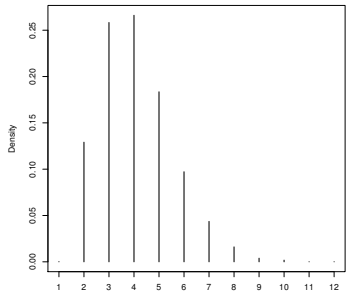
		PPMx				PPM			
variable	parameter	mean	2.5%	50%	97.5%	mean	2.5%	50%	97.5%
Intercept	β_{00}^2	4.46	4.11	4.46	4.81	4.48	4.14	4.48	4.81
EF	β_{01}^2	1.36	1.12	1.36	1.60	1.37	1.13	1.37	1.61
STres	β_{02}^2	-0.58	-1.20	-0.57	0.02	-0.58	-1.17	-0.58	0.03
CKD	β_{03}^2	-1.65	-2.50	-1.65	-0.78	-1.65	-2.50	-1.65	-0.77
MILAN	β_{04}^2	-0.22	-0.69	-0.22	0.26	-0.22	-0.69	-0.22	0.24

From Tables 4.11, 4.12 and 4.13, we get that posterior estimates of the global parameters obtained under the PPM are very similar to those obtained under the PPMx. In particular, concerning the posterior means and quantiles, for each variable the signs are the same in both models and the values are strongly similar.

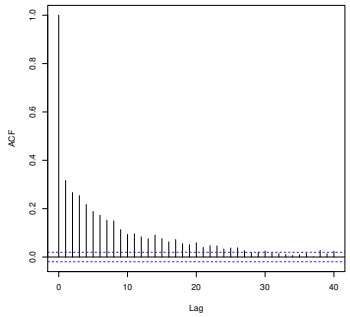
As mentioned before, the PPM model is less robust wit respect to the value of α, the total mass parameter. Indeed the posterior distribution of K_n can be rather different. Let $\alpha = 1$. Figure 4.29 shows the posterior estimate of the number of clusters and the convergence of the Markov chain.



(a) Traces of K_n



(b) Posterior number of clusters



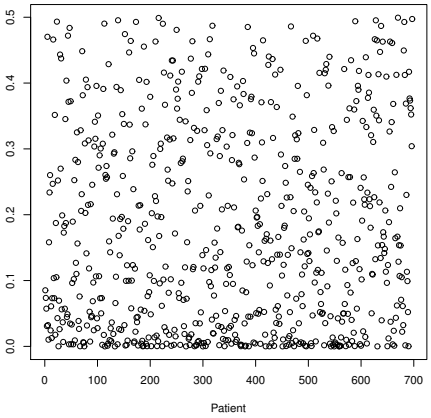
(c) ACF of K_n

Figure 4.29: Markov chain sample of the number of clusters K_n in the PPM ($\alpha = 1$).

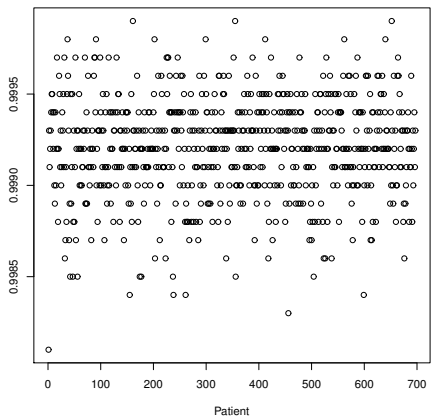
In this case it is $\mathbb{E}[K_n] = 4.09$ and $Var(K_n) = 2.17$ and we observe that K_n is very sensitive to changes of α.

4.2.2 Prediction

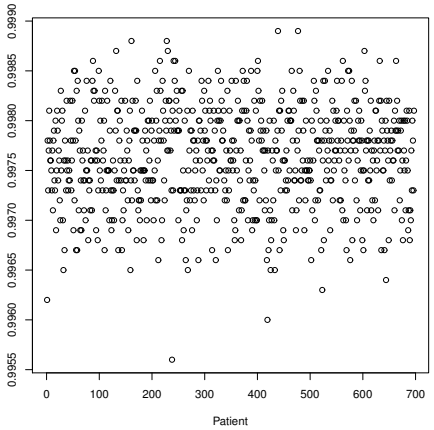
Concerning the prediction, in order to make comparison between the two models we proceed like in Section 4.1.3 but considering only patients in the dataset. In this way, for every patient we obtained the predictive probabilities of survival to discharge and at 60 days and the Bayesian p-values, as showed in Figure 4.30



(a) Bayesian p-values in the PPM



(b) Predictive probabilities of survival to discharge in the PPM



(c) Predictive probabilities of survival at 60 days in the PPM

Figure 4.30: Bayesian p-values for Y_{i1} and predictive survival probabilities for Y_{i2} and Y_{i3}.

Also in this the number of units with Bayesian p-value lesser than 0.05 is quite high (208 units, corresponding to 29.8%) and the predictive probabilities of survival for every patient are all high and near to 1.

4.2.3 Clustering

In order to see what change in clustering units implementing the PPM, we computed the least-square cluster estimate. It is clear that the number of clusters of the optimal partition will be larger than before; as can be noticed observing the density of the posterior number of clusters in Figure 4.21c, the minimum is 5.

Figure 4.31 shows the same three plots of Section 4.1.4, distinguishing between patients deceased to discharge, at 60 days and the remaining survived, whereas Table 4.14 shows the groups numerosities.

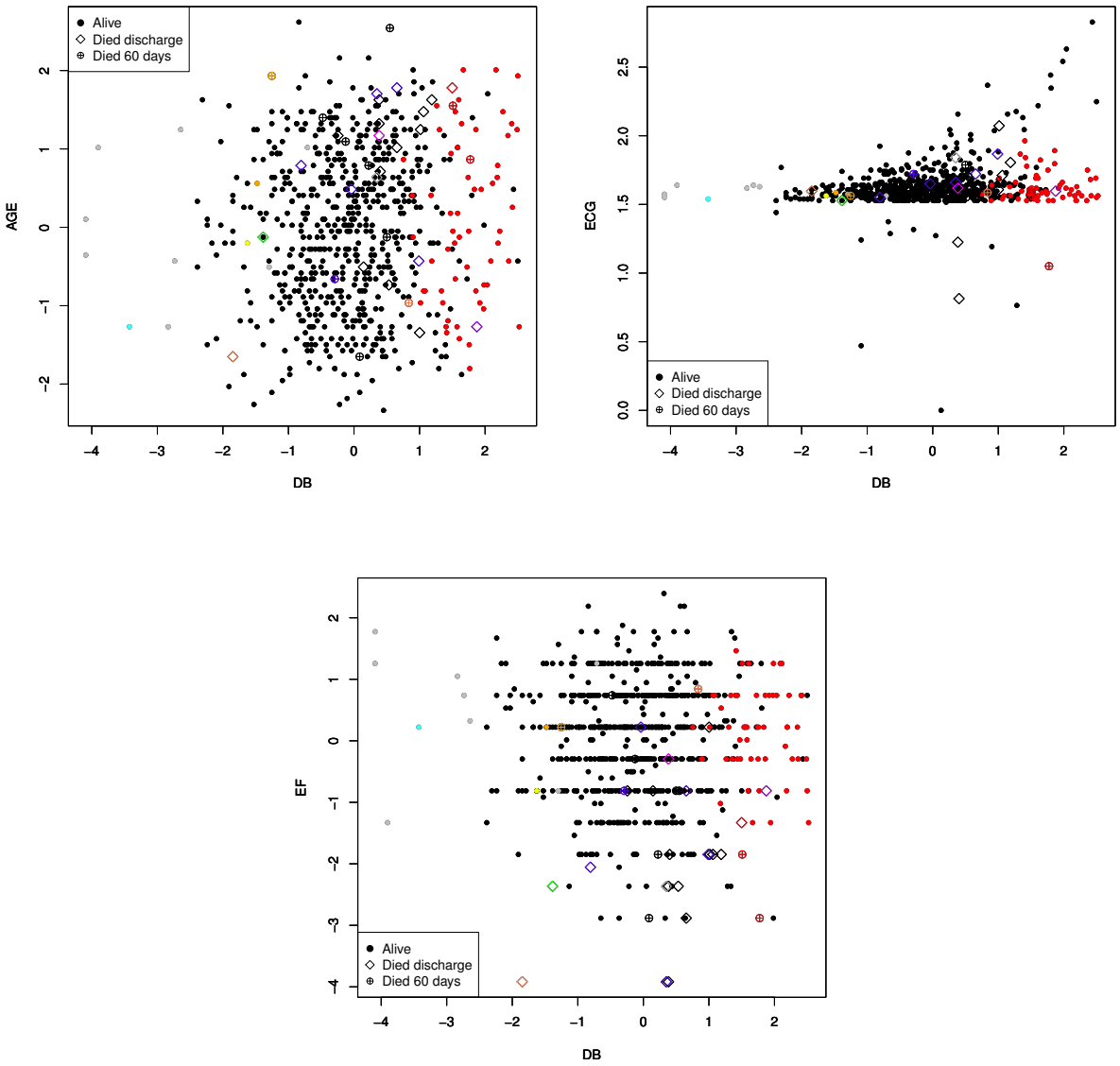


Figure 4.31: Plots of AGE, ECG and EF with respect to $log(DB)$ highlighting the clusters in the the PPM.

Table 4.14: Numerosities of the groups.

Labels	Numerosities	Color
1	612	Black
2	61	Red
3	1	Green
4	6	Blue
5	1	Yellow
6	9	Gray
7	1	Magenta
8	2	Orange
9	1	Purple
10	1	Cyan
11	2	Coral

As expected, the number of groups identified by Dahl's algorithm is larger than before. However, there is always a group containing a large percentage of units (those with the label 1, whose numerosity is 612). In general the clustering seems to make sense. In particular we can observe that all the 6 units labelled with 4 (and plotted with the color Blue) correspond to deceased patients.

4.3 Conclusion and further work

This work constitutes a first attempt of construction of a Bayesian model where the prior on the random partition depends on covariates, for the application to the dataset on patients affected by STEMI described in Chapter 2. Almost all included covariates on the conditional distribution of the data, given the parameters, seem to be significant. Conversely, no covariate seems to be able to explain the computed cluster estimates, which instead are well interpreted by the outcome response DB.

Further work could include assuming a new model where different covariates are set in the likelihood (while all the remaining ones go into the similarity function), or to provide a different cluster estimate under the model we have considered so far.

Appendix A

The Bayesian linear model with unknown variance

Let $\boldsymbol{y} = (y_1, \ldots, y_n)$ be a vector of n response variables and

$$X = \begin{pmatrix} 1 & x_{12} & \cdots & x_{1p} \\ \vdots & \vdots & & \vdots \\ 1 & x_{n_2} & \cdots & x_{n_p} \end{pmatrix},$$

the design matrix of $n \times p$ covariates. A common structure for the model of the relationship between the response variable $\boldsymbol{y}$ and the covariates X is the standard linear regression model given by

$$\boldsymbol{y} = X\boldsymbol{\beta} + \boldsymbol{\epsilon},$$

where $\boldsymbol{\beta}$ is a vector of unknown coefficients and $\boldsymbol{\epsilon}$ is a vector of disturbance term with mean zero. Equally the model can be written as

$$\begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix} = \begin{bmatrix} 1 & x_{12} & \cdots & x_{1p} \\ \vdots & \vdots & & \vdots \\ 1 & x_{n2} & \cdots & x_{np} \end{bmatrix} \begin{bmatrix} \beta_1 \\ \vdots \\ \beta_p \end{bmatrix} \begin{bmatrix} \epsilon_1 \\ \vdots \\ \epsilon_n \end{bmatrix}.$$

The Bayesian approach to fitting the linear model consists of three steps:

1. assign priors to all unknown parameters.
2. write down the likelihood of the data given the parameters;
3. determine the posterior distribution of the parameters given the data using Bayes' theorem.

Step 1: Prior selection

The unknown parameters in the model are the vector of coefficients $\boldsymbol{\beta}$ and the regression variance τ^{-1}. On both priors are applied to represent the knowledge about the distribution of the parameters. It is reasonable to choose the prior distribution in a way that the posterior distribution belongs to a know family. In particular, we choose the Normal-Gamma distribution that is a conjugate prior for the Bayes linear model:

$$\boldsymbol{\beta}, \tau \sim \mathcal{NG}(\boldsymbol{\beta}_0, B_0, a, b).$$

Specifically, $\boldsymbol{\beta}_0$ corresponds to the location parameter and B_0 to the covariance matrix of the normal distribution and a, b denote the parameters of the Gamma distribution. The density function of the prior is then given by:

$$\begin{aligned} f(\boldsymbol{\beta}, \tau) &= f(\boldsymbol{\beta}|\tau) f(\tau) \\ &= \mathcal{N}(\boldsymbol{\beta}|\boldsymbol{\beta}_0, \tau^{-1} B_0) \mathcal{G}(\tau|a, b) \\ &\propto \tau^{a+p/2-1} exp\left\{-\frac{\tau}{2}[(\boldsymbol{\beta}-\boldsymbol{\beta}_0)^T B_0^{-1}(\boldsymbol{\beta}-\boldsymbol{\beta}_0) + 2b]\right\}. \end{aligned} \tag{A.1}$$

Step 2: Likelihood

Given the normal distribution of the error the likelihood function of the model is

$$\begin{aligned} f(\boldsymbol{Y}|\boldsymbol{\beta}, \tau) &= \mathcal{N}(\boldsymbol{Y}|X\boldsymbol{\beta}, \tau^{-1} I) \\ &= \left(\frac{\tau}{2\pi}\right)^{\frac{n}{2}} exp\left\{-\frac{(\boldsymbol{Y}-X\boldsymbol{\beta})^T(\tau^{-1}I)^{-1}(\boldsymbol{Y}-X\boldsymbol{\beta})}{2}\right\}. \end{aligned} \tag{A.2}$$

That is, $\boldsymbol{Y}$ follows a Normal distribution with mean $X\boldsymbol{\beta}$ and precision τI.

Step 3: Posterior distribution

Finally, the posterior distribution is derived by multiplying (A.1) with (A.2):

$$\begin{aligned} &f(\boldsymbol{\beta}, \tau|\boldsymbol{y}) \propto f(\boldsymbol{\beta}, \tau) f(\boldsymbol{y}|\boldsymbol{\beta}, \tau) \\ &\propto \tau^{a+p/2-1} \times exp\left\{-\frac{\tau}{2}[(\boldsymbol{Y}-X\boldsymbol{\beta})^T(\boldsymbol{Y}-X\boldsymbol{\beta}) + (\boldsymbol{\beta}-\boldsymbol{\beta}_0)^T B_0^{-1}(\boldsymbol{\beta}-\boldsymbol{\beta}_0) + 2b]\right\} \end{aligned} \tag{A.3}$$

Applying the identity

$$\begin{aligned} &(\boldsymbol{Y}-X\boldsymbol{\beta})^T(\boldsymbol{Y}-X\boldsymbol{\beta}) + (\boldsymbol{\beta}-\boldsymbol{\beta}_0)^T B_0^{-1}(\boldsymbol{\beta}-\boldsymbol{\beta}_0) + 2b \\ &\qquad = (\boldsymbol{\beta}-\boldsymbol{\beta}^*)^T (B^*)^{-1}(\boldsymbol{\beta}-\boldsymbol{\beta}^*) + 2b^* \end{aligned}$$

to (A.3) yields

$$f(\boldsymbol{\beta}, \tau|\boldsymbol{y}) \propto \tau^{a^*+p/2-1} exp\left\{-\frac{\tau}{2}[(\boldsymbol{\beta}-\boldsymbol{\beta}^*)^T (B^*)^{-1}(\boldsymbol{\beta}-\boldsymbol{\beta}^*) + 2b^*]\right\}, \tag{A.4}$$

where

$$\begin{aligned} B^* &= \left(B_0^{-1} I_n + X^T X\right)^{-1} \\ \boldsymbol{\beta}^* &= B^*\left(B_0^{-1}\boldsymbol{\beta}_0 + X^T\boldsymbol{y}\right) \\ a^* &= a + n/2 \\ b^* &= b + \frac{1}{2}\left(\boldsymbol{\beta}_0^T B_0^{-1}\boldsymbol{\beta}_0 + \boldsymbol{y}^T\boldsymbol{y} - (\boldsymbol{\beta}^*)^T (B^*)^{-1} (\boldsymbol{\beta}^*)\right). \end{aligned}$$

Appendix B

Full-conditionals of the global parameters

Here we compute the full-conditionals of the hyperparameters $\boldsymbol{\eta} = (b, \boldsymbol{\beta}_0^1, \boldsymbol{\beta}_0^2, \boldsymbol{\beta}_0^3)$ in the Gibbs-sampler; we distinguish between the first level and the other two.

It will be useful remember that if $\psi(\boldsymbol{x}) = \boldsymbol{x}^T A \boldsymbol{x} - 2\boldsymbol{p}^T \boldsymbol{x} + c$ is a quadratic form, then

$$\begin{aligned} \psi(\boldsymbol{x}) &= \left(\boldsymbol{x} - A^{-1}\boldsymbol{p}\right)^T A \left(\boldsymbol{x} - A^{-1}\boldsymbol{p}\right) + c \\ &= \left\|\boldsymbol{x} - A^{-1}\boldsymbol{p}\right\|_A^2 + c' \end{aligned} \tag{B.1}$$

B.1 First level

From equation (3.10), for $j = 1, \ldots, k$, we have:

$$\begin{aligned} \boldsymbol{\beta}_j | \tau_j, \boldsymbol{\beta}_0 \sim \mathcal{N}\left(\boldsymbol{\beta}_0, \frac{1}{k_0 \tau_j} I\right), \tau_j | b \sim \mathcal{G}(a, b), \\ \boldsymbol{\beta}_0 \sim \mathcal{N}(\boldsymbol{m}, B), b \sim \mathcal{G}(a_1, a_2), \end{aligned}$$

being $B = g\left(X^T X\right)^{-1}$. Using equation (3.20) and (3.21) we obtain:

$$\pi\left(\boldsymbol{\beta}_0, b | \boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_k, \boldsymbol{\tau}\right) \propto \pi\left(\boldsymbol{\beta}_0\right) \pi(b) \prod_{j=1}^{k} \pi\left(\boldsymbol{\beta}_j | \boldsymbol{\beta}_0, \tau_j\right) \pi\left(\tau_j | b\right) \tag{B.2}$$

$$\propto e^{-\frac{1}{2}(\boldsymbol{\beta_0} - \boldsymbol{m})^T B^{-1} (\boldsymbol{\beta_0} - \boldsymbol{m})} \times b^{a_1 - 1} e^{-a_2 b} \prod_{j=1}^{k} e^{-\frac{k_0 \tau_j}{2} (\boldsymbol{\beta}_j - \boldsymbol{\beta}_0)^T (\boldsymbol{\beta}_j - \boldsymbol{\beta}_0)} \tau_j^{a-1} b^a e^{-b\tau_j}$$

$$= e^{-\frac{1}{2}(\boldsymbol{\beta_0} - \boldsymbol{m})^T B^{-1} (\boldsymbol{\beta_0} - \boldsymbol{m})} \times b^{a_1 - 1} e^{-a_2 b} e^{-\frac{k_0}{2} \sum_{j=1}^{k} \tau_j (\boldsymbol{\beta}_j - \boldsymbol{\beta}_0)^T (\boldsymbol{\beta}_j - \boldsymbol{\beta}_0)} \left(\prod_{j=1}^{k} \tau_j\right)^{a-1} b^{ak} e^{-b \sum_{j=1}^{k} \tau_j}$$

$$\propto b^{a_1 + ak - 1} e^{-\left(a_2 + \sum_{j=1}^{k} \tau_j\right) b} \times e^{-\frac{1}{2} C(\boldsymbol{\beta}_0)}.$$

Using the identity (B.1) we obtain:

$$
\begin{aligned}
C(\boldsymbol{\beta}_0) &= (\boldsymbol{\beta_0} - \boldsymbol{m})^T B^{-1} (\boldsymbol{\beta_0} - \boldsymbol{m}) + k_0 \sum_{j=1}^{k} \tau_j (\boldsymbol{\beta}_0 - \boldsymbol{\beta}_j)^T (\boldsymbol{\beta}_0 - \boldsymbol{\beta}_j) \\
&= \boldsymbol{\beta}_0^T B^{-1} \boldsymbol{\beta}_0 - 2\boldsymbol{m}^T B^{-1} \boldsymbol{\beta}_0 + k_0 \sum_{j=1}^{k} \tau_j \boldsymbol{\beta}_0^T \boldsymbol{\beta}_0 - 2k_0 \left(\sum_{j=1}^{k} \tau_j \boldsymbol{\beta}_j \right)^T \boldsymbol{\beta}_0 + const \\
&= \boldsymbol{\beta}_0^T \left(B^{-1} + k_0 \sum_{j=1}^{k} \tau_j I \right) \boldsymbol{\beta}_0 - 2 \left(B^{-1} \boldsymbol{m} + k_0 \sum_{j=1}^{k} \tau_j \boldsymbol{\beta}_j \right)^T \boldsymbol{\beta}_0 + const' \\
&= \left\| \boldsymbol{\beta}_0 - \left(B^{-1} + k_0 \sum_{j=1}^{k} \tau_j I \right)^{-1} \left(B^{-1} \boldsymbol{m} + k_0 \sum_{j=1}^{k} \tau_j \boldsymbol{\beta}_j \right) \right\|^2_{\left(B^{-1} + k_0 \sum_{j=1}^{k} \tau_j I \right)} + const' \\
&= \| \boldsymbol{\beta}_0 - \boldsymbol{m}_k \|^2_{B_k^{-1}} + const',
\end{aligned}
$$

being $\boldsymbol{m}_k$ and B_k the posterior mean and covariance matrix of $\boldsymbol{\beta}_0$, respectively. In (B.2) we recognize the kernel of a Normal distribution, so that

$$
\begin{aligned}
b|\tau_1, \ldots, \tau_k &\sim \mathcal{G}\left(a_1 + ak, a_2 + \sum_{j=1}^{k} \tau_j \right), \\
\boldsymbol{\beta}_0|\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_k, \tau_1, \ldots, \tau_k &\sim \mathcal{N}_4 (\boldsymbol{m}_k, B_k).
\end{aligned}
$$

B.2 Second and third level

From equations (3.11) and (3.12), for $j = 1, \ldots, k$, we have:

$$
\boldsymbol{\beta}_j|\boldsymbol{\beta}_0 \sim \mathcal{N}\left(\boldsymbol{\beta}_0, \frac{1}{k_0} I \right), \boldsymbol{\beta}_0 \sim \mathcal{N}(\boldsymbol{m}, B),
$$

being $B = g\left(X^T X\right)^{-1}$. Using equation (3.22) and (3.23), we obtain:

$$
\begin{aligned}
\pi(\boldsymbol{\beta}_0|\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_k) &\propto \pi(\boldsymbol{\beta}_0) \prod_{j=1}^{k} \pi(\boldsymbol{\beta}_j|\boldsymbol{\beta}_0) \\
&\propto e^{-\frac{1}{2}(\boldsymbol{\beta_0} - \boldsymbol{m})^T B^{-1} (\boldsymbol{\beta_0} - \boldsymbol{m})} \prod_{j=1}^{k} e^{-\frac{k_0}{2}(\boldsymbol{\beta}_j - \boldsymbol{\beta}_0)^T (\boldsymbol{\beta}_j - \boldsymbol{\beta}_0)} \\
&= e^{-\frac{1}{2}\mathcal{C}(\boldsymbol{\beta}_0)}.
\end{aligned} \tag{B.3}
$$

Using the identity (B.1) we get:

$$
\begin{aligned}
\mathcal{C}(\boldsymbol{\beta}_0) &= (\boldsymbol{\beta_0} - \boldsymbol{m})^T B^{-1} (\boldsymbol{\beta_0} - \boldsymbol{m}) + k_0 \sum_{j=1}^{k} (\boldsymbol{\beta}_0 - \boldsymbol{\beta}_j)^T (\boldsymbol{\beta}_0 - \boldsymbol{\beta}_j) \\
&= \boldsymbol{\beta}_0^T B^{-1} \boldsymbol{\beta}_0 - 2\boldsymbol{m}^T B^{-1} \boldsymbol{\beta}_0 + k_0 k \boldsymbol{\beta}_0^T \boldsymbol{\beta}_0 - 2k_0 \left(\sum_{j=1}^{k} \boldsymbol{\beta}_j \right)^T \boldsymbol{\beta}_0 + const \\
&= \boldsymbol{\beta}_0^T \left(B^{-1} + k_0 k I \right) \boldsymbol{\beta}_0 - 2 \left(B^{-1} \boldsymbol{m} + k_0 \sum_{j=1}^{k} \boldsymbol{\beta}_j \right)^T \boldsymbol{\beta}_0 + const' \\
&= \left\| \boldsymbol{\beta}_0 - \left(B^{-1} + k_0 k I \right)^{-1} \left(B^{-1} \boldsymbol{m} + k_0 \sum_{j=1}^{k} \boldsymbol{\beta}_j \right) \right\|^2_{(B^{-1} + k_0 k I)} + const' \\
&= \| \boldsymbol{\beta}_0 - \boldsymbol{m}_k \|^2_{B_k^{-1}} + const',
\end{aligned}
$$

being $\boldsymbol{m}_k$ and B_k the posterior mean and covariance matrix of $\boldsymbol{\beta}_0$, respectively. As before, in (B.3) we recognize the kernel of a Normal distribution, so that

$$
\boldsymbol{\beta}_0 | \boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_k \sim \mathcal{N}_5 (\boldsymbol{m}_k, B_k) .
$$

Note that the two full-conditionals just computed have the same functional form. The only difference is given by the presence of a prior in the covariance structure (on the τ_js in this case). Indeed, the expression $\sum_{j=1}^{k} \tau_j$ that appears both in the posterior mean and in the posterior covariance in the first level is replaced in the other two levels by k, the number of clusters.

Appendix C

A priori mean and variance of $\boldsymbol{\theta}_j$s

We say that X follows a *Gamma* distribution of shape parameter α and rate parameter β (and write $X \sim Gamma(\alpha, \beta)$) if

$$f_X(x) = \frac{\beta^\alpha}{\Gamma(\alpha)} x^{\alpha-1} e^{-\beta x},$$

and it results

$$\mathbb{E}[X] = \frac{\alpha}{\beta} \qquad \text{and} \qquad Var(X) = \frac{\alpha}{\beta^2}.$$

Moreover, if $X \sim Gamma(\alpha, \beta)$, then $\frac{1}{X}$ follows an *Inverse-Gamma* distribution ($\frac{1}{X} \sim Inv-Gamma(\alpha, \beta)$). It results

$$\mathbb{E}\left[\frac{1}{X}\right] = \frac{\beta}{\alpha - 1} \qquad \text{and} \qquad Var\left(\frac{1}{X}\right) = \frac{\beta^2}{(\alpha-1)^2(\alpha-2)}.$$

A priori mean and variance for the first level

Suppose that

$$\boldsymbol{\beta}|\boldsymbol{\beta}_0, \tau \sim \mathcal{N}\left(\boldsymbol{\beta}_0, \frac{1}{k_0\tau} I\right),\ \tau|b \sim \mathcal{G}(a, b),\ \boldsymbol{\beta}_0 \sim \mathcal{N}(\boldsymbol{m}, g\left(X^T X\right)^{-1},\ b \sim \mathcal{G}(a_1, a_2).$$

Then the mean of $\boldsymbol{\beta}$ is

$$\mathbb{E}[\boldsymbol{\beta}] = \mathbb{E}[\mathbb{E}[\boldsymbol{\beta}|\boldsymbol{\beta}_0, \tau]] = \mathbb{E}[\boldsymbol{\beta}_0] = \boldsymbol{m},$$

and the variance is

$$\begin{aligned}
Var(\boldsymbol{\beta}) &= \mathbb{E}[Var(\boldsymbol{\beta}|\boldsymbol{\beta}_0, \tau)] + Var(\mathbb{E}[\boldsymbol{\beta}|\boldsymbol{\beta}_0, \tau]) \\
&= \mathbb{E}[\frac{1}{k_0\tau} I] + Var(\boldsymbol{\beta}_0) \\
&= \frac{I}{k_0}\mathbb{E}[\mathbb{E}[\frac{1}{\tau}|b]] + g(X^T X)^{-1} \\
&= \frac{I}{k_0}\mathbb{E}[\frac{b}{a-1}] + g(X^T X)^{-1} \\
&= \frac{a_1}{a_2(a-1)}\frac{1}{k_0} I + g(X^T X)^{-1}.
\end{aligned}$$

The mean of τ is

$$\mathbb{E}[\tau] = \mathbb{E}[\mathbb{E}[\tau|b]] = \mathbb{E}[\frac{a}{b}] = \frac{aa_2}{a_1 - 1},$$

and the variance is

$$\begin{aligned} Var(\tau) &= \mathbb{E}[Var(\tau|b)] + Var(\mathbb{E}[\tau|b]) \\ &= \mathbb{E}[\frac{a}{b^2}] + Var(\frac{a}{b}) \\ &= a(Var(\frac{1}{b}) + \mathbb{E}[\frac{1}{b}]^2) + a^2 Var(\frac{1}{b}) \\ &= a(a+1)Var(\frac{1}{b}) + a\mathbb{E}[\frac{1}{b}]^2 \\ &= a(a+1)\frac{a_2^2}{(a_1-1)^2(a_1-2)} + a\frac{a_2^2}{(a_1-1)^2} \\ &= \frac{a_2^2(a+a_1-1)}{(a_1-1)^2(a_1-2)}. \end{aligned}$$

A priori mean and variance for the second and third level

Suppose that

$$\boldsymbol{\beta}|\boldsymbol{\beta}_0 \sim \mathcal{N}\left(\boldsymbol{\beta}_0, \frac{1}{k_0} I\right),\ \boldsymbol{\beta}_0 \sim \mathcal{N}(\boldsymbol{m}, g\left(X^T X\right)^{-1}.$$

Then the mean of $\boldsymbol{\beta}$ is

$$\mathbb{E}[\boldsymbol{\beta}] = \mathbb{E}[\mathbb{E}[\boldsymbol{\beta}|\boldsymbol{\beta}_0]] = \mathbb{E}[\boldsymbol{\beta}_0] = \boldsymbol{m},$$

and the variance is

$$\begin{aligned} Var(\boldsymbol{\beta}) &= \mathbb{E}[Var(\boldsymbol{\beta}|\boldsymbol{\beta}_0)] + Var(\mathbb{E}[\boldsymbol{\beta}|\boldsymbol{\beta}_0,]) \\ &= \mathbb{E}[\frac{1}{k_0} I] + Var(\boldsymbol{\beta}_0) \\ &= \frac{I}{k_0} + g(X^T X)^{-1}. \end{aligned}$$

Bibliography

Antoniak, C. E. (1974). Mixtures of Dirichlet Processes with Applications to Bayesian Nonparametric Problems. *The Annals of Statistics,* **2**, 1152–1174.

Arratia, R., Barbour, A. D., and Tavaré, S. (1992). Poisson process approximations for the Ewens sampling formula. *The Annals of Applied probability* **2**, 519–535.

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

Dahl, D. B. (2006). Model-Based Clustering for Expression Data via a Dirichlet Process Mixture Model. *Bayesian Inference for Gene Expression and Proteomics* (chap 10).

Ferguson, T. S. (1973). A Bayesian Analysis of Some Nonparametric Problems. *The Annals of Statistics,* **1**, 209–230.

Ghosh, J. K. and Ramamoorthi, R. V. (2003). *Bayesian Nonparametric*. Springer, New York.

Guglielmi, A., Ieva, F., Paganoni, A. M., and Ruggeri, F. (2012). A Bayesian random-effects Model for survival probabilities after Acute Myocardical Infarcion. *Chilean Journal of Statistics*.

Guglielmi, A., Ieva, F., Paganoni, A. M., Ruggeri, F., and Soriano, J. (2013). Semi-parametric Bayesian Models for clustering and classification in presence of unbalanced in-hospital survival. *Journal of the Royal Statistical Society, C.* **3**, 15–29.

Hartigan, J. A. (1990). Partition models. *Communications in Statistics, Part A - Theory and Methods* **19**, 2745–2756.

Ieva, F. (2013). Designing and mining a multicenter observational clinical registry concerning patients with Acute Coronary Syndromes. *New Diagnostic, Therapeutic and Organizational Strategies for Acute Coronary Syndromes Patients,* ed. Grieco, N., Marzegalli, M., and Paganoni A. M., Springer.

MacEachern, S. N. and Muller, P. (1998). Estimating Mixture of Dirichlet Process Models. *Journal of Computational and Graphical Statistics* **7**, 223–238.

Medvedovic, M. and Sivaganesan, S. (2002). Bayesian Mixture Model-Based Clustering of Gene Expression Profiles. *Bioinformarics* **18**, 1194–1206.

Muller, P., Quintana, F. A., and Rosner, G. L. (2011). A Product Partition Model with Regression on Covariates. *Journal of Computational and Graphical Statistics,* **20**, 260–278.

Neal, R. M. (2000). Markov Chain Sampling Methods for Dirichlet Process Mixture Models. *Journal of Computational and Graphical Statistics* **9**, 249–265.

Prandoni, E. (2012). Modelli Bayesiani Semiparametrici per le probabilità di sopravvivenza in seguito ad Infarto Miocardico Acuto. Master Thesis. Politecnico di Milano.

Quintana, F. A. and Iglesias, P. L. (2003). Bayesian Clustering and Product Partition Models. *Journal of the Royal Statistical Society* **65**, 557–574.

Rugarli, C. (2010). *Medicina interna sistematica.* Elsevier.

Sethuraman, J. (1994). A constructing Definition of Dirichlet Process Prior. *Statistica Sinica,* **2**, 639–650.

West, M., Muller, P., and Escobar, M. D. (1994). Hierarchical Priors and Mixture Models, with Application in Regression and Density Estimation. *Aspects of Uncertainty*, 363–386.