Similarity Analysis in Species Sampling Mixture Models

Carlos A. Navarrete

Abstract. Species Sampling Mixture Models (SSMMs) rely on modeling the data on top of a structure that considers the inherent clustering of the observations under the hypothesis of prior exchangeability. This work proposes a method to study the information given by the posterior clustering behaviour of SSMMs, called Similarity Analysis. It is based fundamentally on decomposing the similarity matrix obtained from a sample of the partitions in an intrinsic and an extrinsic part. This gives valuable information about the individual characteristics that explain the clustering, specially in the presence of covariates. A new approach to the representation of partitions and their interpretation is also given. Applications in Bayesian density estimation, linear regression models and multivariate regression models for binary response are included.

Keywords: Species Sampling Mixture Models, Dirichlet process, cluster analysis, Bayesian density estimation, multivariate binary regression, Gibbs sampling.

Contents

CHAPTER 1

Introduction

The main goal of this work is to study the relationship between individuals, their covariates and the posterior probability of grouping the observations, that is, a partition for the data, in the context of Species Sampling Mixture Models (SSMMs). Models defined from mixtures of Species Sampling Models (SSMs) present a great innovation, compared with the *full parametric* scheme: they consider a prior probability distribution for the latent inherent clustering structure of the data. This is done by means of specifying a random probability measure (RPM) for individual parameters, which in turn defines a random partition structure, allowing the model to accommodate the posterior distribution of the parameters in a way that represents the observations guided by the model specification, but not constrained to a unique form of it: there are individual *variations* in the specific way such model is followed. For individuals sharing the same subject-specific random effects, the difference between them is explained by sampling variability in the likelihood. This characteristic makes SSMM a very flexible tool. In this work we focus on the posterior partition structures and their interpretation, based on individual covariates.

For data $y_1, \ldots, y_n$, we consider the following hierarchical model:

$$\begin{aligned} y_i|\theta_i, x_i, \nu &\sim F(\phi(\theta_i, x_i), \nu) \\ \theta_i|G &\sim G \\ G &\sim SSM(G_0, p) \end{aligned} \tag{1}$$

The individual parameters $(\theta_1, \ldots, \theta_n)$ come from a SSM centered in a baseline measure G_0 and defined by a predictive rule p, to be defined below. $(x_1, \ldots, x_n)$ are optional vectors of covariates and ϕ is a function relating θ_i and x_i in F, for

$i = 1, 2, \ldots, n$, like, for example, the traditional linear regression form

$$\phi(\theta_i, x_i) = E(y_i|\theta_i, x_i) = \theta_{0i} + \theta_{1i}x_{1i} + \ldots + \theta_{(q-1)i}x_{(q-1)i}.$$

The parameters θ_i need not be always functionally related to the covariates x_i. For instance, one could also define $\theta_{qi} = Var(y_i|\theta_i, x_i)$ in the context of a Normal specification for F. ν represents optional additional parameters, which may be fixed or given an hyperprior distribution $\pi|\eta$. θ_i, $i = 1, \ldots, n$ are, in general, vectors in $\mathbb{R}^m$, for some known positive integer m. The framework provided by SSMMs assume that the observations come from a mixture of distributions, and this mixture is defined by the SSM. Equivalently, (1) can be written as

$$y_i|x_i, \theta_i, \nu \sim \int F(\phi(\theta_i, x_i), \nu) dG(\theta_i) \tag{2}$$

with $G(\theta_i)$ coming in turn from a SSM. Model (1) can be extended by traditional Bayesian hierarchical specifications. A common choice for F is a normal kernel $N(\mu, V)$. Many applications consider $\theta_i = \mu_i$ and a common variance $\nu = V$, which leads to a mixture of normals. This scheme can be extended to the more general case $\theta_i = (\mu_i, V_i)$. A typical base distribution for the first case is Normal, and for the latter Normal/Inverse Gamma or Multivariate Normal/Inverse Wishart, both chosen for conjugacy.

SSMs are RPMs introduced by Pitman (1996), which include, as a particular case, the Dirichlet Process (DP) (Ferguson 1973, Blackwell and MacQueen, 1973). The name Species Sampling comes from the idea that $\theta_1, \theta_2, \ldots$ come from a big population formed by different *species*, and each value θ_i is a *tag* associated to a new species found. The following definitions, all due to Pitman (1996) introduce SSMs.

DEFINITION 1. *Given a baseline distribution G_0, (θ_n) is a species sampling sequence iff it is a sample from a random distribution G of the form*

$$G = \sum_i \omega_i \delta_{(\hat{\theta}_i)}(\cdot) + (1 - \sum_i \omega_i) G_0 \tag{3}$$

for some sequence of random variables (ω_i) such that $\omega_i >= 0$ for every i and $\sum_i \omega_i \leq 1$ a.s. Marginally, $\{\hat{\theta}_i, i = 1, 2, \ldots\}$ is a random sample from G_0 and, given G, $\theta_1, \theta_2, \ldots$ are independent and identically distributed according to G.

DEFINITION 2. *A Species Sampling Model (SSM) is a random distribution G of the form (3) with a random sample $\theta_1, \ldots, \theta_n$ from G.*

SSMs belong to a class of Bayesian models which are often called *non-parametric*, although they are defined, in fact, in terms of an infinite (but countable) number of parameters. We will refer to them as the non-parametric part of the model. We will be interested in *proper* SSMs, in which $\sum_i \omega_i = 1$ a.s., which determine discrete distributions with probability one. Moreover, when the SSM is proper, its construction determines a partition in the sampled values of $(\theta_1, \ldots, \theta_n)$. In other words, proper SSMs are based, fundamentally, on a probability distribution over all possible partitions of the vector of parameters $(\theta_1, \ldots, \theta_n)$, which are assumed to be exchangeable, plus the *locations* of the clusters, distributed depending on both the partition and the baseline measure G_0. The RPM (3) is then simplified to

$$G(\cdot) = \sum_i \omega_i \delta_{(\hat{\theta}_i)}(\cdot) \tag{4}$$

The connection between proper SSMs and the underlying partition structure that characterizes them becomes clear when introducing *predictive rules*.

DEFINITION 3. *A predictive rule is a rule specifying the distribution of θ_1 and the conditional distribution of θ_{n+1} given $\theta_1, \ldots, \theta_n$ for any $n = 1, 2, \ldots$.*

Pitman (1996) shows that, given a continuous distribution G_0, a sequence (θ_n) is a species sampling sequence if it is exchangeable and subject to a predictive rule of the form

$$\begin{aligned} P(\theta_1 \in \cdot) &= G_o(\cdot) \\ P(\theta_{n+1} \in \cdot | \theta_1, \ldots, \theta_n) &= \sum_{j=1}^{k} p_j(N_n) \mathrm{I}(\theta_j^* \in \cdot) + p_{k+1}(N_n) G_0(\cdot). \end{aligned} \tag{5}$$

Given n, the construction (5) defines the probability distribution of $(\theta_1, \ldots, \theta_n)$ in a recursive way. This constitutes, in fact, a generalization of the Polya urn scheme that characterizes the Dirichlet Process (Blackwell and MacQueen, 1973). The first value, θ_1, is directly sampled from the baseline distribution G_0. The rest of the parameters θ_i, conditional on the previously sampled values, can either *copy* one sampled value, or generate a new one from G_0. It can be seen in (5) that the probability of obtaining *ties* in the sampled values θ_i is positive. k represents the

number of *unique* values $(\theta_1^*, \ldots, \theta_k^*)$ in $(\theta_1, \ldots, \theta_n)$, that is, the number of *clusters*, and N_n is the vector of cluster sizes n_j in the implicit partition of $(\theta_1, \ldots, \theta_n)$, for $j = 1, 2, \ldots, k$. Exchangeability imposes the constraint that the probabilities of copying a previous value or sampling a new one depend on the already sampled values only on N_n. Connecting (5) with (4), Pitman (1996) shows that (4) is the limit in total variation norm of (5) when $n \to \infty$, and thus the weigths ω_i can be interpreted as limit proportions of the recorded tags. For any SSM, marginalizing over the RPM G leads to a joint distribution $p(\theta_1, \ldots, \theta_n)$ that can be expressed as the product of conditional distributions as in (5). From the point of view of partitions, this determines an *exchangeable partition probability function* (EPPF). The EPPF is defined as (Pitman 1996)

$$P\left(\bigcap_{j=1}^{k}(\theta_i = \theta_j^* \text{ for all } i \in C_j)\right) = p(n_1, \ldots, n_k) \tag{6}$$

for some *symmetric* function p of k-tuples of non-negative integers with sum n. Allowing n to vary, let $[n]_k = (n_1, \ldots, n_k)$ with $\sum_{j=1}^{k} n_j = n$. This represents a partition with k clusters, and each cluster j has n_j elements. Pitman (1996) shows that an EPPF must satisfy, for any sequence $[n]_k$ and for any k

$$\begin{aligned} p([1]) &= 1 \text{ and} \\ p([n]_k) &= \sum_{j=1}^{k+1} p([n^{j+}]_k) \end{aligned} \tag{7}$$

where $[n^{j+}]_k$ is defined from $[n]_k$ incrementing n_j by one (so a partition with k or $k+1$ clusters may be obtained, the latter case when $j = k+1$). Conversely, any symmetric function p satisfying (7) is an EPPF. EPPFs are important because SSMs can be alternatively defined by an EPPF plus the baseline measure G_0. In this case, the predictive probability functions in (5) are easily shown to be given by

$$p_j(N_n) = \frac{p([n^{j+}]_k)}{p([n]_k)}, 1 \leq j \leq k+1 \tag{8}$$

It follows that the choice of predictive probabilities are constrained by (7) and (8), and so they are not arbitrary. Two available forms of SSM will be considered in this work: the Dirichlet Process and the Pitman-Yor process, also called Poisson-Dirichlet process.

Dirichlet Process. The Dirichlet Process (DP) (Ferguson 1973, Blackwell and MacQueen 1973) is a well known particular case of SSM, with (5) determined by $p_j = n_j/(n+M)$ for $j \leq k$ and $p_{k+1} = M/(n+M)$, where n_j is the size of cluster j, and M is a positive *mass* parameter, in what is referred to as Blackwell & MacQueen Polya urn representation. The EPPF for the DP is

$$p(n_1, \ldots, n_k) = \frac{M^{k-1} \prod_{j=1}^{k} (n_j - 1)!}{[M+1]_{n-1}} \tag{9}$$

with $[x]_m = \prod_{j=1}^{m} (x + j - 1)$. Sethuraman (1994) showed that the DP corresponds to (4) with $\hat{\theta}_1, \hat{\theta}_2, \ldots$ being i.i.d. from a baseline distribution G_0 and with weights defined as

$$\begin{aligned} &\omega_1 = 1 \text{ and} \\ &\omega_h = \prod_{j=1}^{h-1} (1 - V_j) V_h \text{ for } h > 1, \end{aligned} \tag{10}$$

where $V_1, V_2, \ldots$ are i.i.d. samples from a $Beta(1, M)$ distribution. This construction is often referred to as *Stick Breaking representation.*

Pitman-Yor process. Another example of SSM is the Pitman-Yor process, in which case we have $p_j = (n_j - \alpha)/(n+M)$ for $j \leq k$ and $p_{k+1} = (M + k\alpha)/(n+M)$. This process can be seen as a generalization of the Dirichlet process. The process can be defined in two alternative ways:

$$\alpha = -M/m \text{ for some } m = 2, 3, \ldots \text{ and } M > 0, \text{ or} \tag{11}$$

$$0 \leq \alpha \leq 1 \text{ and } M > -\alpha. \tag{12}$$

Note that the process defined by (11) determines partitions with a number of clusters less or equal to m with probability 1. The EPPF for this model is given by (Pitman 1996)

$$p(n_1, \ldots, n_k) = \frac{\left(\prod_{j=1}^{k-1} (M + j\alpha)\right) \left(\prod_{j=1}^{k} [1 - \alpha]_{n_j - 1}\right)}{[M+1]_{n-1}} \tag{13}$$

When $\alpha = 0$ and $M > 0$, the PY model reduces to the DP case.

The class of SSM models admits several other important special cases. These include the Dirichlet-multinomial process (Muliere and Secchi 1995), the beta-two process (Ishwaran and Zarepour 2000) and the stick-breaking priors (Ishwaran and

James 2001, 2003b). Additional properties of SSMs can be found in Pitman (1996). For a related class of RPMs see Lijoi, Mena and Prünster (2007).

The specific topic of relating covariates with the nonparametric part of the model has been studied mainly by means of introducing dependence on the covariates in the stick-breaking representation of the Dirichlet Process. MacEachern (1999) introduces the Dependent Dirichlet Process (DDP), allowing explicit dependence on covariates on random distributions coming from a DP. DDP models allow various types of dependence. One type comes from replacing the random values $\hat{\theta}_i$ in (10) by stochastic processes $\theta_i\hat{(x)}$. Additionaly, the base measure may be allowed to depend on x, too. Other source of dependence is obtained from replacing the random variables V_j in (10) with stochastic processes $V_j(x)$, which also introduces dependence on the mass parameter, setting $M = M(x)$. De Iorio, Müller, Rosner and MacEachern (2004) study ANOVA-type structures based on the DDP. They impose the structure on the locations $\hat{\theta}_i := \hat{\theta}_i(x) = m_i + A_{\nu i} + B_{\omega i}$, where $x_i = (\nu_i, \omega_i)$ in an ANOVA fashion. They show that this model can be written as a DP mixture model of the form

$$\begin{aligned} (y_i|x_i) &\sim H_{x_i}(y_i) \\ H_{x_i}(y) &= \int N(y|\alpha d_i, S) dF(\alpha), \\ F &\sim DP(M, p^0) \end{aligned} \tag{14}$$

where d_i denote a design vector to select the appropiate ANOVA effects corresponding to x_i. They show that this scheme can be generalized by further hierarchical specifications, which include the possibility to represent a categorical response by means of specifying additional latent parameters. Their conclusions extend naturally to mixtures of proper SSMs, since they are valid for the general representation (4). In a related approach, Griffin and Steel (2006) make the weights in the Sethuraman representation dependent on the covariates in what is called Order-Based DDP (πDDP). They induce an ordering π in the random variables which define the weights in (10) at each covariate value, such that distributions for similar covariates values are associated with similar orderings, and so they are are close. They apply mixtures of πDDP processes to time series and spatial data. Dunson, Pillai and Park (2007) propose a model with a formulation similar to (14) in the context of

Bayesian density regression. But their approach replaces the stick-breaking prior (10) by weighted mixtures of DP priors

$$G_x = \sum_{j=1}^{n} b_j(x) G^*_{x_j},$$
$$G^*_{x_j} \sim DP(M, G_0)$$

independently for j=1,2,...,n.

In this work, we focus on the relation of posterior partitions obtained from (1) and individual covariates, without specifying, explicitly, dependence on the covariates in the SSM that generates the individual parameters θ_i. Instead, we study the indirect influence of the covariates, through the likelihood, in the posterior clustering process. By exchangeability, the prior probability for any particular observation to join a cluster (actually, for a sample of the individual parameter, coming from the SSM) depends only on the cluster sizes in the partition. *A posteriori*, this is modified by the likelihood of the observations, and the influence of the covariates in the process is determined by their influence in the likelihood. To identify individual influence, and then associate it to covariate values, it is necessary to know how the clustering process and the likelihood interact. MacEachern and Müller (1998, 2000) note that the individual parameters $(\theta_1, \ldots, \theta_n)$ can be reparameterized as pairs $(c_i, \theta^*_{c_i})$, where c_i is the cluster corresponding to the i-th element, and $\theta^*_{c_i}$ represents its location. Then, conditional on the random partition determined by the SSM, the model becomes

$$y_i | c_i, \theta^*_{c_i}, x_i, \nu \sim F(\phi(\theta^*_{c_i}, x_i), \nu) \; i = 1, \ldots, n,$$

considering a partition for $\theta_1, \ldots, \theta_n$, with $k \leq n$ possible different values for $\theta^*_{c_i}$. Note that the partition of $\theta_1, \ldots, \theta_n$ is in $1:1$ relation with a partition of $y_1, \ldots, y_n$, unless the optional additional parameters ν determine an additional clustering process, a case not treated here. From (5) and (2), the posterior distribution for the

cluster membership of element θ_n, given $(\theta_1, \ldots, \theta_{n-1})$ is determined by

$$\begin{aligned} & p(c_n|\theta_1, \ldots, \theta_{n-1}, y_n, x_n, \nu) \\ & \propto \sum_{j=1}^{k_{n-1}} p(y_n|\phi(\theta_j^*, x_n), \nu) p_j(N_{n-1}) I(c_n = j) \\ & + p_{(k_{n-1}+1)}(N_{n-1}) \left(\int p(y_n|\phi(\theta, x_n), \nu) dG_0(\theta) \right) I(c_n = k_{n-1} + 1) \end{aligned} \tag{15}$$

where k_{n-1} is the number of clusters in the partition of $(\theta_1, \ldots, \theta_{n-1})$. In turn, the posterior location for cluster c_n, given the elements that form it, is given by

$$p(\theta_{c_n}^*|y, x, \nu) \propto \left(\prod_{y_i \in C_n} p(y_i|\phi(\theta_{c_n}^*, x_i), \nu) \right) G_0(\theta_n^*). \tag{16}$$

This updating scheme controls the formation of clusters, and therefore its understanding is very important for the interpretation of the resulting partitions. The predictive probabilities $p_j(N_{n-1})$, $j = 1, \ldots, (k_{n-1} + 1)$ are the most obvious way to control the clustering process. One example is the DP, where the probability of repeating a previously sampled value is distributed *a priori* equally for each value (considering every location individually, no matter how many times it is repeated), and $p_{k^- +1}(N_n^-) \propto M$, the so-called *mass parameter*, which controls the probability of forming new clusters. When the prior distribution of the partitions is ruled by the PY process, the probability of forming new clusters is, additionally, dependent on the number of clusters. That comes from the prior specification. In the posterior form, the construction tends to join *similar* individuals in terms of their likelihood. The process also depends on the specification of G_0, since it rules the prior distribution for the locations. For the probability of creating new clusters in (15), we have

$$\int p(y_n|\phi(\theta, x_n), \nu) dG_0(\theta) = E_{G_0(\theta)} \left[p(y_n|\phi(\theta, x_n), \nu) \right] \tag{17}$$

So, as a general rule, an observation will tend to form a new cluster when the already sampled values are less likely to represent it, compared with the *expected likelihood* defined by the base distribution. One could think of this process as sampling from a spectrum of colors. Depending on some degree of tolerance, different colors can be sampled and then classified in more basic categories, like *blue*, *yellow* or *red*. In (15), basic color classification (clusters) group individuals which are likely to be represented by the same *version* of the model. The degree of tolerance determines,

for instance, if some tone of *purple* is classified in *blue* or *red*, or rather forces the process to create a new category called, say, *purple*. Tolerance is in turn determined by the specification of the likelihood in (2), the base measure G_0 and the clustering mechanism (5). From a data analysis perspective it is interesting to note that model (1) allows for two extreme cases: all the parameters θ_i are equal, and all of them are distinct, reducing inference to parametric models. But more generally, a discrete RPM prior for the unknown distribution of the individual parameters represents an intermediate choice between models with all parameters equal or different. By adequately choosing the predictive probabilities $\{p_j\}$, the analyst can favor different partition structures. In the DP case, for instance, a large value of M implies many clusters, while small values of M favor partitions with a a reduced number of clusters. The prior expectation and variance of the number of clusters are given by Liu (1996)

$$\sum_{i=1}^{n} \frac{M}{M+i-1} \qquad \text{and} \qquad \sum_{i=1}^{n} \frac{M(i-1)}{(M+i-1)^2}.$$

The extreme cases mentioned earlier follow by letting $M \to 0$ and $M \to \infty$, respectively. Some authors (e.g. Escobar & West, 1995) treat M as an unknown parameter itself, choosing a prior distribution (usually Gamma) to reflect uncertainty. The above expressions for prior mean and variance can be used for prior elicitation purposes (Kottas et al. 2005). In contrast, the PY process has more flexible partition structures, taking also in consideration the number of clusters in the clustering process.

In the last years, efficient algorithms have been developed in the context of Gibbs Sampling for Dirichlet Process Mixture Models (DPMMs), which extend naturally to SSMMs. Among others, we can mention the works of Escobar (1994), Escobar & West (1995), Bush & MacEachern (1996), MacEachern & Müller (1998, 2000) and Walker (2007). Many Gibbs Sampling algorithms are based on noting that, as a consequence of prior exchangeability in $\theta_1, \ldots, \theta_n$, given the number of observations n, the prediction rule (5) implies that

$$P(\theta_i \in \cdot | \theta_{-i}) = \sum_{j=1}^{k^{-i}} p_j^{-i}(N_{n-1}) \mathrm{I}(\theta_j^* \in \cdot) + p_{k^{-i}+1}(N_{n-1}) G_0(\cdot) \tag{18}$$

where $\theta_{-i} = \{\theta_1, \ldots, \theta_{i-1}, \theta_{i+1}, \ldots, \theta_n\}$ and k^{-i} is the number of clusters without considering observation i. A posteriori, this fact leads to the modified posterior probability for cluster configurations

$$\begin{aligned} & p(c_i|\theta_{-i}, y_i, x_i, \nu) \\ & \propto \sum_{j=1}^{k_{-i}} p(y_i|\phi(\theta_j^*, x_i), \nu) p_j(N_{n-1}) I(c_i = j) \\ & + p_{(k_{-i}+1)}(N_{n-1}) \left(\int p(y_i|\phi(\theta, x_i), \nu) dG_0(\theta) \right) I(c_i = k_{-i} + 1) \end{aligned} \tag{19}$$

where k_{-i} is the number of clusters in the partition of without θ_i. The posterior location for cluster c_i is similar to (16). This allows direct sampling for θ_1, ..., θ_n conditional on the rest of the parameters (Escobar, 1994, Escobar and West, 1995). MacEachern and Müller (1998) showed that it is more efficient to sample from this distribution in two steps. First, sample cluster memberships c_i, $i = 1, 2, \ldots, n$, and then update the locations θ_j, $j = 1, 2, \ldots, k$. This individual allocation scheme is used in most applications presented in this work, and is detailed in Appendix A. Walker (2007) proposes a new alternative, based on the introduction of latent variables which reduce the infiniteness of the RPM representation to a finite case. Several other algorithms are available now, like Neal (1998), Neal and Jain (2000) and Dahl (2005). Most of them rely on accept/reject methods. The only other algorithm used in this work is the one proposed by Dahl, which made possible simulations for the application shown in Chapter 5.

For the purpose of identifying individual influence in the clustering, we propose a new way of representing partitions. Partitions are represented by Partition Matrices (PMs), defined by $\rho_{ij} = 1$ if individuals i and j belong to the same cluster, 0 in other case, for every pair (i, j) in the sample. This representation requires taking in consideration that partitions are uniquely defined by equivalence relations, and these, in turn, are characterized by *reflexiveness, symmetry* and *transitivity*. In general, the clustering process is characterized, beyond sampled partitions, by probabilities of joining pairs of individuals. This is represented by Similarity Matrices (SMs), defined by the probabilities of pairs (i, j) to be in the same cluster. They constitute the expectation of PMs, and make it possible the definition of loss functions associated to the decision of selecting one partition for the data, as is

shown in Lau and Green (2007). The proposed representation allows, additionally, identification of individual influence by means of decomposing the PM or SM in two parts: one *intrinsic* and one *extrinsic*, the former relating two individuals directly, and the latter relating individuals by application of a transitivity rule. The decomposition is based, in turn, in identifying the *first cluster representative* (FCR) of each cluster, defined as the representative with lowest label of each cluster. The partitions themselves, considering only the individuals forming each cluster, are invariant to the order of the observations. But PMs and SMs, and their decompositions, represent partitions in relation with the order of the observations in the sample, and this can be changed at will by rearranging the rows and columns of the PM or SM. This method, called Similarity Analysis, provides valuable information concerning the role of the covariates in the clustering process, since it identifies which individuals "lead" the groups, and this is the same as knowing which ones are more likely to represent the rest. Graphical ways to represent this information are proposed, along with a guide to their interpretation, based on the covariates. As a side effect of the proposed partition representation, some new algorithms to find candidate partitions for the data are also proposed.

Mixtures of Dirichlet processes are currently being used in different contexts. De La Cruz, Quintana and Müller (2007) develop a semiparametric hierarchical model for classification based on logitudinal markers. Their model is an extension of the DDP, with an additional probability model for group classification. They also investigate the effect of the dependence introduced by the DDP compared with a model with independent DP priors. Jara, García-Zattera and Lesaffre (2007) propose an extension to multivariate probit models based on a DP mixture. The mixture is with respect to both location and covariance of a normal kernel. They study different parametrization alternatives for the covariance matrix, addressing identification constraints and tractability of computations. To show SSMMs and Similarity Analysis in practice, three applications are presented in this work. The first one (Chapter 3) reviews the Bayesian density estimation model proposed by Escobar and West (1995), extended to SSMMs. As Escobar and West do, we also apply their model to Galaxy data from Roeder (1990), and show how to find candidate partitions based on the algorithms proposed in this work. The Galaxy

dataset has been analyzed by a number of authors, including Escobar & West (1995), Richardson & Green (1997), Stephens (2000), Ishwaran & James (2003), Quintana (2006) and Navarrete, Quintana and Müller (2008). The second application presents a linear regression model based on (1) and explore the posterior clustering behavior of simulated data, based on covariates and Similarity Analysis. The third application shows a multivariate binary response model applied to real data, modelling the relative risk of Atrial Fibrilation events at 30 days and 1 year of follow-up. The statistical model presented is similar to the application in Jara et al (2007), but it considers a full specification of the covariance structure, dealing with a high dimensional parameter space.

The rest of this work is organized as follows. Chapter 2 introduces PMs, SMs, their graphical representation and properties and their decomposition in intrinsic and extrinsic parts. Some methods to find partitions for the data based on the information given by the SM are also discussed. Chapter 3 shows the Bayesian Density application. Chapter 4 shows the application related to linear regression models, based on simulated data. Chapter 5 presents the multivariate regression model. A final discussion is given in Chapter 6, and a short introduction to computational algorithms used in this work is given in Appendix A.

CHAPTER 2

Similarity analysis

As seen in the previous chapter, SSMMs rely on an inherent mechanism of clustering the observations. In fact, any set of data can be seen to be partitioned, at least in trivial ways, like one cluster per individual, or one cluster for all observations. In this section, a novel way of representing the prior and posterior partition structure will be introduced, which will be seen to reveal valuable information from data modelled by a SSMM. For the purpose of this work, we need to establish a link between posterior partitions and individual information. The first step consists in representing partitions based on individuals and the relations between them. This is the concept behind *partition matrices*. Then, we need to summarize the information provided by the clustering mechanism, considering that partitions are sampled following a stochastic process which depends both on prior cluster configuration and individual information. This will be done by means of *similarity matrices*, and their decomposition in *intrinsic* and *extrinsic* parts, the base of Similarity Analysis. The link with individual information is based on the concept of *first cluster representatives* and the interpretation of the decomposition of the similarity.

1. Partitions and Partition Matrices

1.1. Equivalence relations and partitions. Let us consider a set of n elements $[n] = \{1, 2, \ldots, n\}$. A relation r between the elements of $[n]$ is defined as a subset of the product $[n] \times [n] = \{(i, j) : i \in [n], j \in [n]\}$. Two elements, i and j, of $[n]$ are said to be related iff the pair (i, j) belongs to r. From now on, we will be interested in *equivalence relations* (ERs), which satisfy three well known properties: *reflexiveness* (every element is related with itself), *symmetry* (if a is related with b, then b is related with a), and *transitivity* (if a is related with b, and b is related with c, then a is related with c). Every ER "$\sim$" in $[n]$ defines a partition $\{C_1, \ldots, C_k\}$, which means $[n] = \bigcup_{j=1}^{k} C_i$ and $C_i \cap C_j = \emptyset$ for every $i, j = 1, 2, \ldots, k$. Every

element $i \in [n]$ will belong to an equivalence class, which we will call *cluster* C_i, $1 \leq C_i \leq k$, calling k the number of such clusters, for some $1 \leq k \leq n$. Equivalently, the cluster label of i, c_i will be j iff $i \in C_j$. In order to label the clusters in a unique way, we will follow the convention of labeling them *in order of appearence*:

(1) $c_1 = 1$

(2) $\min\{i : c_i = j_1\} < \min\{i : c_i = j_2\} \Leftrightarrow 1 \leq j_1 < j_2 \leq k$

In other words, the first element must belong to C_1, and the first element of $[n] - (C_1 \cup \ldots \cup C_{j-1})$ must belong to C_j, for $2 \leq j \leq k$. This is equivalent to the following recursive formula for cluster labels.

PROPOSITION 1. *Let $\sim$ be an ER in $[n]$, and let $\rho_{ij} = 1$ if $i \sim j$, 0 in other case. Labeling clusters in the partition in order of appearence is equivalent to defining the cluster label for element i, $i = 1, 2, \ldots, n$ as*

$$\begin{aligned} c_1 &= 1 \\ c_i &= \max_{l<i}\{c_l \rho_{l,i}\} + (1 - \max_{l<i}\{\rho_{l,i}\}) \max_{l<i}\{c_l + 1\} \end{aligned} \tag{20}$$

Proof. Let $\sim$ be an ER, and let i be an element in $[n]$. The case $i = 1$ is obvious. For $i > 1$, if element i joins a previous cluster, then ρ_{il} will be one for any element l belonging to the same cluster. As one element can not belong to more than one cluster at a time, then we have $c_i = \max_{l<i}\{c_l \rho_{l,i}\}$. Otherwise, if element i forms a new cluster, then the label of the new cluster must be the maximum label already defined, plus one, which is equivalent to the second term of the right side of (20).

1.2. Partition matrices. From now on,we will represent a partition by an $n \times n$ matrix R formed by the elements $\rho_{ij},\ i, j \in 1, 2, \ldots, n$. Reflexiveness implies that the main diagonal of R is formed by ones. Symmetry imposes R to be symmetric. Transitivity is fulfilled when for every triplet of elements $\{i, j, k\} \in [n]$, the number of ones in $\{\rho_{ij}, \rho_{ik}, \rho_{jk}\}$ is not two. There are several equivalent ways to define this contition algebraically. One is saying that transitivity is equivalent to the following condition:

PROPOSITION 2. *An $n \times n$ symmetric matrix $R = (\rho_{ij})$ with main diagonal 1 represents a transitive relation if and only if*

$$1 - (\rho_{ij} + \rho_{ik} + \rho_{jk}) + 2\rho_{ij}\rho_{ik}\rho_{jk} \geq 0 \tag{21}$$

for every triplet $\{i, j, k\}$ *in* $[n]$.

Proof. The latter condition sums zero or one, except in the case when there are two ones in the triplet $\{\rho_{ij}, \rho_{ik}, \rho_{jk}\}$, when the sum is -1. The reason to highlight this specific result will be seen later on.

Definition 4. *A partition matrix (PM) of size* n*,* R*, is a square, symmetric matrix with elements* $\rho_{ij} \in \{0, 1\}, i, j = 1, \ldots, n$*, with ones in the main diagonal and satisfying* (21).

For a fixed number of elements n, any partition of such elements can be represented by a partition matrix, and conversely, any matrix satisfying the preceding definition represents a unique partition of n elements, defined by (20). Any particular order of the elements in $[n]$ defines a different PM, but given a specific order of the elements in $[n]$, the PM is unique. Any row or column represents a cluster, and the sum of a row or column is equal to the size of the respective cluster.

Proposition 3. *If the rows and columns of a PM are sorted by cluster label, a block-diagonal matrix is obtained. On the other hand, every PM is a permutation of the rows and columns of a block diagonal matrix of zeroes and ones.*

Example. Suppose we have 5 elements, and the partition is $\{\{1, 3\}, \{2\}, \{4, 5\}\}$. Sorting the rows and columns of the PM by cluster label is equivalent to renaming element 3 as 2, and viceversa. The partition is now $\{\{1, 2\}, \{3\}, \{4, 5\}\}$, and it is represented by a block diagonal matrix.

Definition 5. *The first cluster representative (FCR) of an element* i *is* $d_i = \min\{l \in [n] : c_l = c_i\}$*. The first representative of a cluster* C *is given by* $\min\{i \in [n] : c_i = c\}$.

FCRs provide a link from partitions to individual influence. FCRs are defined in relation to the specific order of the sample, and the information they provide is intrepretable when the order of the observations is meaningful, based on individual characteristics, which are defined by individual covariates in the scope of this work.

Proposition 4. *Any PM* R *can be written as* $R = Q^T Q$*, where* Q *is a* $(n \times n)$ *matrix, with row* $l, l = 1, 2, \ldots, n$ *representing a cluster whose FCR is element* l*,*

and each column representing an element of $[n]$*. The elements* q_{li} *take value* 1 *if element* i *belongs to cluster* c_{d_l}*, and* 0 *otherwise. The matrix* Q *defined this way is upper triangular and unique.*

Proof. For any PM R, its elements r_{ij} can be written as $r_{ij} = \sum_{l=1}^{i} q_{li}q_{lj}$ with q as defined above, since the only way that elements i and j can be related is when they both belong to the same cluster, and each cluster is uniquely identified by its first representative. The number of possible FCRs for element i is at most i, because i can either join a previous cluster, represented by a preceding element in the sample, or form a new one. An element i can, actually, be classified with elements represented by an observation with higher label, but then, due to the partition representation rule (20), i becomes the FCR. There can not be more elements than clusters, so the matrix is upper triangular. Since clusters are represented uniquely by their FCR, and elements can belong to only one cluster, the matrix Q is unique.

1.3. Example. Suppose we have 5 elements, partitioned in three clusters: $\{\{1,2\},\{3\},\{4,5\}\}$. In relation notation we have $r = \{\{1,1\},\{2,2\},\{3,3\},\{4,4\},\{5,5\},\{1,2\},\{2,1\},\{4,5\},\{5,4\}\}$. The PM is then

$$R = \begin{pmatrix} 1 & 1 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & 1 & 1 \end{pmatrix}$$

and the partition decomposition is given by $R = Q^T Q$ with

$$Q = \begin{pmatrix} 1 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & 0 & 0 \end{pmatrix}$$

Two special partitions can be highlighted. One is the *minimal* partition, which has as many clusters as elements, and it is represented by the identity matrix. The other is the *maximal* partition, where all elements are grouped in one cluster, and it is represented by a square matrix of ones. If we represent ERs in $[n]$ as subsets

of $[n] \times [n]$, the *maximal* partition $\{(i,j) : i = 1,\dots,n;\ j = 1,\dots,n\}$ is, indeed, maximal, in the sense that it covers the ER corresponding to any other partition. In the same way, the *minimal* partition's ER $\{(i,i) : i = 1,\dots,n\}$ is a subset of any other partition's ER. The $\subseteq$ relation in ERs corresponds to the $\leq$ relation element by element in the partition matrices. The intersection of two PMs is obtained by taking the logical *AND* element by element (or the product), obtaining a new ER. This is not the case for the union, in general, and the result must be *completed* to satisfy transitivity. For PMs, the union is equivalent to the logical *OR* element by element (or the sum defining $1 + 1 = 1$).

DEFINITION 6. *Let R be a reflexive and symmetric relation in $[n]$. The transitive closure R^+ is the intersection of all ERs in $[n]$ that cover R.*

DEFINITION 7. *Let R be a reflexive and symmetric matrix with elements $\rho_{ij} \in \{0,1\}$. We will define the completion operator for R as $R^+ = \{\rho^+_{ij}\}$ with*

$$\rho^+_{ij} = \rho_{ij} + (1 - \rho_{ij}) \max_{k=1,2,\dots,n} \{\rho_{ik}\rho_{kj}\}$$

Completing R several times, until no change is done, results in the transitive closure of R. What the previous operation does is *connecting* the elements that should be related for the relation to be transitive, and it results in a *bigger* relation, in the sense that it includes the original one. Algorithms to find the transitive closure have been developed in Graph Theory. For further information, see Warshall (1962), Yoeli (1961) and Nuutila (1995).

2. Similarity Matrices

DEFINITION 8. *A similarity matrix (SM) of size n, S, is a square, symmetric and semi positive definite matrix with elements s_{ij} in the $[0,1]$ interval, such that $s_{ii} = 1$ for all $i = 1, 2, \dots, n$.*

A SM is intended to represent the marginal probabilities for every pair (i,j) to be related. It is a generalization of a PM, since PMs are particular cases of SMs with probabilities zero and one. It can also be seen as the expectation of a random sample of PMs coming from the same process. Then, by the Law of Large Numbers, it can be estimated by $\hat{S} = \frac{1}{N}\sum_{r=1}^{N} R_r$, based on a sample of PMs $R_1, \dots, R_N$. But this is not enough. Just like PMs must satisfy (21) in order to be transitive

and so represent an actual partition, a similarity matrix must be somehow *coherent* in order to give information about the clustering structure of the data, so that the probability of joining two observations takes into account the transitivity property on the underlying partitions. The starting point comes out noting that, just like PMs can be decomposed as the cross-product of a matrix of cluster memberships, it would be desireable that a SM could also share this property, and for that purpose, since symmetry is guaranteed, SMs should be semi positive definite.

2.1. Diagonalization of a SM. From Linear Algebra, real symmetric matrices are semi positive definite iff any diagonal representation (under congruence) has only positive diagonal elements, or equivalently, if all eigenvalues are positive (or zero). We will make use of this fact to explore what being semi positive definite means in terms of the similarities s_{ij}. The diagonalization can be made by elementary row and column operations, in order to make zeroes from down and from the right of the diagonal to the extreme of the matrix. That is, for $i = 1, 2, \ldots, n-1$ substract to row $(i+1)$ row i multiplied by the i-th diagonal element, substract to column $(i+1)$ column i multiplied by the same i-th diagonal element, and repeat the operation for $i+2$ to n. At the end of the process, a diagonal matrix congruent with S is obtained, and it must be checked that every element in the diagonal is positive.

For $n = 2$, it is easy to verify that $S = (s_{ij})$ is semi positive definite iff $1 - s_{12} \geq 0$, which is always true. The case $s_{12} = 1$ defines a PM.

For $n = 3$, the diagonal elements obtained are $v_1 = 1$, $v_2 = 1 - s_{12}$ and $v_3 = (1 - (s_{12}^2 + s_{13}^2 + s_{23}^2) + 2s_{12}s_{13}s_{23})/(1 - s_{12}^2)$. Since v_1 and v_2 are always positive, it can be concluded that S is semi positive definite if and only if

$$1 - (s_{12}^2 + s_{13}^2 + s_{23}^2) + 2s_{12}s_{13}s_{23} \geq 0. \tag{22}$$

This is equivalent to (21) for PMs, with the obvious detail that $r_{ij}^2 = r_{ij}$ for the elements of a PM. This condition constitutes, in fact, a generalized definition of transitivity, which we will call *coherence*. Triplets close to intransitivity for PMs are discarded, like, for example, $(s_{ij}, s_{ik}, s_{jk}) = (0.1, 0.9, 0.9)$. The limiting region

is shown in figure (1), and the transitive triplets are located in the interior, and border, of the region.

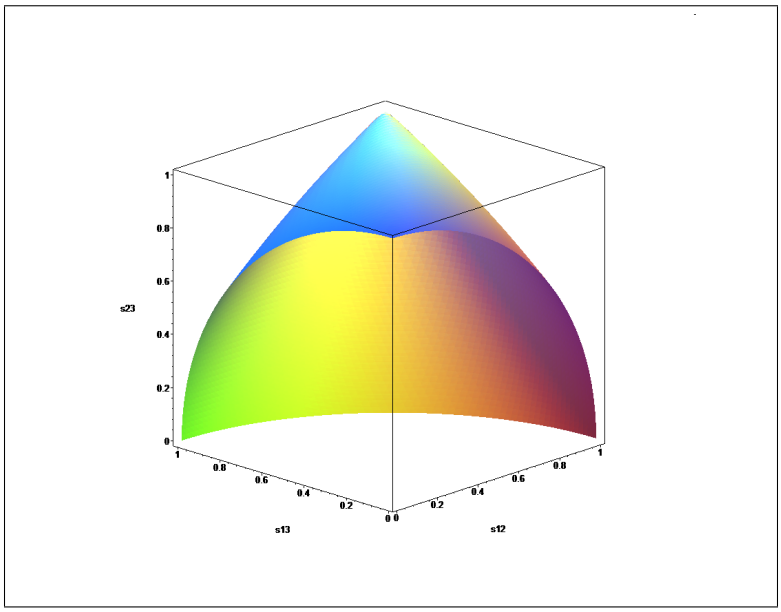

FIGURE 1. Coherence region for n=3. Transitive triplets are located in the interior of the region.

What $v_i \geq 0$ actually shows is the overall coherence of row (column) i with the previous rows (columns).

2.1.1. *Example.* Let S be the following 5×5 matrix.

	1	2	3	4	5
1	1.00	0.30	0.60	0.70	0.70
2	0.30	1.00	0.60	0.90	0.85
3	0.60	0.60	1.00	0.75	0.20
4	0.70	0.90	0.75	1.00	0.90
5	0.70	0.85	0.20	0.90	1.00

The diagonal elements of S, after diagonalization, are, approximately, $v_1 = 1$, $v_2 = 0.91$, $v_3 = 0.45$, $v_4 = -0.01$, $v_5 = -0.25$. There are problems then, since elements 4 and 5 seem to contradict the similarity of the rest. For instance, $s_{24} = 0.9$, $s_{14} = 0.7$, but $s_{12} = 0.3$. Also, $s_{35} = 0.2$, but $s_{15} = 0.7$ and $s_{13} = 0.6$. We are going to change some values, arbitrarily, in order to obtain more coherence. To choose

better values, let us examine the formula for v_4:

$$v_4 = 1 - s_{14}^2 - \frac{s_{24} - s_{12}s_{14}}{1 - s_{12}^2} - \frac{s_{34} - s_{13}s_{14} - \frac{(s_{24} - s_{12}s_{14})(s_{23} - s_{12}s_{13})}{1 - s_{12}^2}}{1 - s_{13}^2 - \frac{(s_{23} - s_{12}s_{13})^2}{1 - s_{12}^2}}$$

What we need is to put the values in an acceptable range in order to turn v_4 in a non-negative value. If we set, for instance, $s_{24} = s_{42} = 0.3$, we obtain $v_4 = 0.31$ and $v_5 = -2$ (notice that the previous diagonal values remain unchanged). We have fixed one incoherence. The reasoning for v_5 goes along the same way. We do not show its formula here, for space considerations, but it is calculated straightforwardly from the row and column operations already mentioned. We see that element 5 "pretends" to be close to 1 and 2, but 1 and 2 are distant from each other. So 5 must "decide" if joining 1 or 2. The same problem with 4 and 2. Also, 5 is distant from 3, which is close to 4, a "close friend" to 5. In order to make the system more coherent, we have to distance 5 from 2, and improve the relations with 3. Setting $s_{25} = s52 = 0.2$ and $s35 = s53 = 0.5$ we get $v_5 = 0.09$, which passes the test. The new matrix, which is indeed a SM, is

	1	2	3	4	5
1	1.00	0.30	0.60	0.70	0.70
2	0.30	1.00	0.60	0.30	0.20
3	0.60	0.60	1.00	0.75	0.50
4	0.70	0.30	0.75	1.00	0.90
5	0.70	0.20	0.50	0.90	1.00

Of course, this was shown as a fabricated example. With real data, one can not just change the SM. An incoherent SM reveals inner contradictions on the clustering process, demanding further research.

2.1.2. *Decomposition of the similarity.* In general, the diagonal elements are built up based on the recursive formulas

$$\begin{aligned} q_{ij} &= \left(s_{ij} - \sum_{l=1}^{i-1} q_{li} q_{lj} \right) / q_{ii} \\ v_i^2 = q_{ii}^2 &= 1 - \sum_{l=1}^{i-1} q_{li}^2 \end{aligned} \tag{23}$$

The condition that the diagonal values q_{ii} should be positive is equivalent to checking the coherence of row and column i with the previous ones, as can be seen in

the expansion of q_{ii}^2:

$$\begin{aligned} q_{ii}^2 &= 1 - \sum_{l=1}^{i-1} q_{li}^2 \\ &= 1 - s_{1i}^2 - (s_{2i} - s_{12}s_{1i})^2 \\ &\quad -((s_{3i} - s_{13}s_{1i}) - (s_{23} - s_{12}s_{13})(s_{2i} - s_{12}s_{1i}))^2 - \dots \end{aligned}$$

The above formula takes in consideration and compensates every possible source of incoherence in the related triplets in the rows and columns, evaluating recursively every row and column with its predecessors. For every triplet $\{i, j, k\}, i < j < k$ the factors $(s_{ik} - s_{ij}s_{jk})$ constitute a measure of influence of the transitive rule in the triplet. As the row and/or column number increases, the previous expansion evaluates recursively every possible combination of triplets with index lower or equal to i, that is, the elements that i can *join* when they are considered FCRs of their clusters. Of course, here we do not have actual clusters, but a tendency to group, which differs between the observations, and it is relative to the order in which the observations are sampled. But in the total similarity, the order of the observations is irrelevant, except for the order of the rows and columns, and the values s_{ij} remain the same when they correspond to the same individuals. (23) corresponds to the recursive formula for the elements of Q in the Cholesky decomposition $S = Q^T Q$. From this point of view, we have for any $1 \leq i \leq j \leq n$

$$\begin{aligned} S_{ij} &= I_{ij} + E_{ij} \text{ with} \\ I_{ij} &= q_{ii}q_{ij} \\ E_{ij} &= \sum_{l=1}^{i-1} q_{li}q_{lj} \end{aligned} \tag{24}$$

This decomposition is quite informative, since it explains the likeness of elements i and j in terms of one part that depends exclusively on i and j (*intrinsic* similarity), and a remaining part that depends on how their *common* relations influence their tendency to join (*extrinsic* similarity). Starting recursively from $n = 2$, it can be observed that I_{ij} is a measure of closeness between i and j, beyond transitivity:

For $n = 2$,

$$\begin{aligned} I_{12} &= S_{12} \\ E_{12} &= 0 \end{aligned}$$

For $n = 3$,

$$\begin{aligned} I_{12} &= S_{12} \\ I_{13} &= S_{13} \\ I_{23} &= S_{23} - S_{12}S_{13} \\ E_{12} &= 0 \\ E_{13} &= 0 \\ E_{23} &= S_{12}S_{13} \end{aligned}$$

Both I_{ij} and E_{ij} have absolute value lower than one since, by construction of the Cholesky decomposition, we have $S_{ii} = 1 = \sum_{l=1}^{i} q_{li}^2$. Numerically, I_{ij} and E_{ij} can take negative values, and if I_{ij} is negative, E_{ij} must take higher values in order to compensate that, for S_{ij} to be positive, and viceversa. In practice, however, negative values are rare, and close to zero. Now, if we take $j = i$, we have that $I_{ii} = 1 - \sum_{l=1}^{i-1} q_{li}^2$ represents the degree of *autonomy* of element i from the previous elements. This process can be compared to a group of n persons, who meet each other depending on both the agreement with one specific individual, and the similarity to the people that relate to him. An individual with high autonomy is like a person with strong opinions, who rarely follows others, but others may follow him, like a leader. Autonomy can be understood as the *idiosyncratic* part of an individual's ideas, not depending on the opinions of the other individuals that relate to him. In our case, due to the representation of partitions (20), the observations that an element i can possibly relate to, in a constructive view of the partition, are $1, 2, \ldots, i-1$. Highly autonomous individuals represent a generalization of *cluster representatives*. Here, autonomy represents to what degree and individual can be considered the *first representative* of its kind. We will see that these quantities give essential information for similarity analysis in SSMMs, justifying a formal definition.

DEFINITION 9. *Let $S = (s_{ij})$ be a SM of size n, and $S = Q^T Q$ its Cholesky decomposition. Let q_{ij} be the elements of the upper triangular matrix Q. We will call decomposition of the similarity s_{ij} to the sum $s_{ij} = I_{ij} + E_{ij}$, with*

- *Intrinsic similarity of elements i and j to $I_{ij} = q_{ii}q_{ij}$,*
- *Extrinsic similarity of elements i and j to $E_{ij} = \sum_{l=1}^{i-1} q_{li}q_{lj}$ and*

- *Autonomy of element i to* $A_i = I_{ii} = q_{ii}^2$.

2.1.3. *Autonomy versus probability of creating new clusters.* It is tempting to interpret the previously defined quantities in terms of probabilities, specially with Autonomy, in which case we have

$$\sum_{l=1}^{i} q_{li}^2 = 1$$

One could associate q_{li}^2 to the probability of element i joining element l. But, for instance, $q_{12}^2 = s_{12}^2$, which is different to the probability of element 2 to join 1, which is s_{12}. Then we also have $A_2 = q_{22}^2 = 1 - s_{12}^2$, which is greater than the probability of element 2 forming a new cluster, which is $1 - s_{12}$. The quantities in Definition 9 constitute a decomposition of probabilities, and should be treated like that. Nevertheless, they are quite informative. One reason to consider these quantities instead of the empirical probability of creating new clusters is that the latter distinguish only highly autonomous individuals when the statistical model favours partitions with few clusters, which is a very common case. Instead, the mentioned quantities will be shown in the applications to be much more *sensitive* to the information provided by an individual, which is to say its covariates. Another reason is related with the unique representation of partitions (20) and the recursive nature of the Polya urn cluster allocating process. The sampled partitions follow these rules, so, for instance, the first element in the sample will generate a new cluster with probability one. But this particular element may not be that special, it is just the first one in the list. A fundamental hypothesis of Species Sampling models is the exchangeability of $\theta_1, \ldots, \theta_n$ given the nonparametric distribution G. But in the sample, when an element forms a new cluster, cluster labels are changed in order to follow the representation rule for partitions. In the clustering process itself, every element has a priori the same probability to form new clusters, and any element can be the *first one*. So the information provided by autonomy can not be fully extracted from the empirical probability of generating new clusters, in general, because the latter is associated with the order of appearence of the elements in the sample, when the model was fit. Instead, decomposed similarity and its derivatives follow directly from the similarity matrix, which does not consider any particular order of the observations in terms of the *values* of the matrix, but only in the

order of its rows and columns. If the rows and columns of the SM are rearranged in any order, the similarity decomposition, based on the Cholesky decomposition of the rearranged SM, automatically accommodates to the permutation, and the new first element will have autonomy one, and highly autonomous elements will have the *first cluster representative* interpretation relative to the new order of the elements. Looking at the decomposition in PMs (Proposition 4) can be clarifying. Two elements i and j, $i < j$ can be related for two reasons. One case is when both belong to a cluster represented by a FCR l, for some $l < i$. We can consider this an *extrinsic* relation, and it is equivalent to $E^R_{ij} = \sum_{l=1}^{i} q_{li}q_{lj}$ with $q_{li} = I(d_i = l)$, $q_{lj} = I(d_j = l)$. The other case is when i is itself a FCR, and then the relation can be considered *intrinsic*. It is equivalent to $I_{ij} = I(d_j = i)$. It is implied that $d_i = i$, since i must be a FCR for d_j to take i as its value. With these considerations, the decomposition in Definition 9 is the same for PMs, although for any pair (i, j), either the intrinsic or the extrinsic part will be zero. As the SM S is an approximation to the expectation of the PMs R, its decomposition is based on the same principle, and it can be interpreted in a similar way.

It is important to consider, in the context of SSMMs, which motivate this work, the connection of the discussion in this section and predictive rules. (5) is equivalent to (18), the latter being a representation of the clustering process adapted for Gibbs Sampling. But in (18), an element i could be reallocated in a cluster represented by a FCR with higher label than i, say $j > i$. There is no contradiction between this and similarity decomposition, since, after the reallocation, to respect the partition representation, cluster labels are changed, and element i in this case will become the FCR of its cluster, even if its reallocation did not originate a new cluster. So, the probability that $d_i = i$ is not equal to the probability that element i forms a new cluster.

2.1.4. *Similarity analysis and covariates.* The posterior probability for the partition of $(\theta_1, \ldots, \theta_n)$ given the observations $(y_1, x_1), \ldots, (y_n, x_n)$ is given by

$$\prod_{i=1}^{n} p(c_i|\theta_1, \ldots, \theta_{i-1}, y_i, x_i)$$

and the conditional probabilities in the product are defined in (15). From the point of view of FCRs, given $d_1, \ldots, d_{i-1}$, (15) is equivalent to setting recursively

$$d_i = l < i \quad \text{with prob.} \quad \propto p_{c_{d_l}}(N_{i-1}) p(y_i|\phi(\theta_l, x_i)) q_{ll} \tag{25}$$

$$d_i = i \quad \text{with prob.} \quad \propto p_{k_{i-1}+1}(N_{i-1}) \left(\int p(y_i|\phi(\theta_i, x_i)) dG_0(\theta_i) \right) \tag{26}$$

As seen before (Chapter 1), the expresion

$$\int p(y_i|\phi(\theta_i, x_i)) dG_0(\theta_i)$$

corresponds to the expectation of $p(y_i|\phi(\theta_i, x_i))$ defined by G_0, so we will refer to it as *expected likelihood*. The construction of $(d_1, \ldots, d_n)$ determines the decomposition $\{q_{ij}, 1 \leq i \leq j \leq n\}$ for every partition, since $q_{li} = I(d_i = l)$, and this is relative to the order of the observations. From this construction, it can be seen that individuals, based on their covariates, tend to join a previous one if its parameters represent it with high likelihood, or to form a new cluster if its expected likelihood, based on G_0, is relatively high for a new sample of the parameters from G_0, compared with the parameter values of the clusters represented by the previous FCRs.

Suppose there is an important association between the response y_i and the covariates x_i, and the rows and columns of the SM are sorted according to x. This sorting means arranging the rows and columns of the SM in an order that depends on the values of x. For this task, it does not matter if x is continuous or discrete. The important thing is that the order considered is interpretable in terms of the covariate. If the association between y and x is correctly specified in the model, then a combination of parameters which is highly likely for the first observation (y_1, x_1) should give high likelihood to the rest of the observations too. If the relation somehow changes for different values of the covariates, for instance if there is an interaction not considered in the model, or the functional specification ϕ of the relation is not the most adequate in (2), then the sampled values for the first observation will be likely for the rest of the observations up to a certain value of the sorting covariate. After this value, it will be more likely that the observations come from a newly sampled value of the parameters that represent them better. Let us suppose now that the covariates relate poorly with the response. Then the clustering process will behave erratically when depending on the order of x, and one

could see a variable number of clusters, depending on the specification of the rest of the covariates, but clusters will tend to mix, making it difficult to distinguish which elements form them, and which are the FCRs. Other clearly identifiable case is the presence of observations which differ notably from the rest in their behavior. We could call them *outliers*, understanding that the term does not necessarily refer to observations with *extreme* covariate values, but observations which seem to relate with their covariates in a different way. In that case, we will see a tendency to cluster all the observations together, except for the outliers, which will appear as FCRs of their own size-one cluster.

2.1.5. *Example.* Suppose the data come from

$$y_i \sim N(\alpha_0 + \alpha_1 x_i, \sigma^2) \text{ for } i = 1, 2, \ldots, n \text{ and } x_i \in \{0, 1\}$$

and we propose

$$\begin{aligned} y_i &\sim N(\alpha_i, \sigma^2) \\ \alpha_i | F &\sim F \\ F &\sim SSM(G_0, p) \end{aligned}$$

Given α_i, the likelihood for y_i is proportional to

$$\exp(-\frac{1}{2\sigma^2}(y_i - \alpha_i)^2).$$

The real mean of the observations is α_0 for every y_i such that $x_i = 0$, and $(\alpha_0 + \alpha_1)$ for the other group. Suppose the observations are sorted by x_i, although this covariate is not considered in the proposed model. Then $x_1 = 0$, and an estimation α_1, highly likely for y_1, will be likely for any observation of group 0, but it will not fit as well for group 1. So the first observation (y_l, x_l) of group 1 will define a new estimation α_l with high probability, an estimation which will be highly likely for the rest of the group. This process depends also on other parameters. In this case, if the variance σ^2 is overestimated, it will be difficult for the clustering process to detect important differences on the means. On the contrary, if the variance is subestimated, clustering will tend to compensate this by assigning more individual effects than needed. The specification of the baseline measure G_0 is crucial, and should support the entire range of possible values for the parameters α. If G_0 is poorly specified, the effect will be a confussion in the clustering process, and

there will be no clear clustering tendency, making the similarity analysis hard to interpret. If now we fit a better specified model $y_i \sim N(\alpha_{0i} + \alpha_{1i}x_i, \sigma^2)$, with the nonparametric part defined accordingly, then an estimation $(\alpha_{01}, \alpha_{11})$ corresponding to the observation (y_1, x_1) will be likely for the rest of the observations too, regardless of covariate values.■

SSMMs are so flexible in part because of their ability to adequate to differences in the observations which are not considered explicitly in the topmost part of the hierarchy. If the SM shows a tendency to a particular clustering of the observations, and this clustering coincides with the presence of a factor, it is a strong sign that the mentioned factor could be included in an explicit way in the model. This is also applicable to continuous covariates, since, when one important covariate or interaction between covariates is missing, the nonparametric part of the model will tend to correct the lack of fit proposing different parameter values for the observations, according to the missing covariate(s). These covariates may be available, or may be considered latent, in which case a reformulation of the model could be considered. The intrinsic similarity I, in particular, shows the clustering tendency in a more precise way, beyond the clustering explained by simple application of the transitivity rule, highlighting the specific observations that rule the partitioning, which are the FCRs. Consider a SM S, decomposed as $S = I + E$. Sorting the rows and columns of S by any particular covariate, and then decomposing it, gives important information about the relevance of the covariate in the model. Let π be a permutation of $\{1, 2, \ldots, n\}$, and let S^π be the SM S with its rows and columns rearranged by π. S^π can be decomposed as $S^\pi = I^\pi + E^\pi$. For any pair of individuals (i, j), the similarity values remain the same, that is, $S_{ij} = S^\pi_{\pi(i)\pi(j)}$. But the decomposition changes, and now we have new intrinsic and extrinsic parts, which depend on a different set of elements, determined by the permutation π. If we define π as an order depending on a specific covariate, or a set of covariates, then the decomposition can be *explained* in terms of intrinsic and extrinsic similarity *due* to the covariate values associated with i and j, whose positions are now $\pi(i)$ and $\pi(j)$. This analysis is made based on a graphic representation of I. In a complimentary way, one can combine scatter plots of the covariates with the information given by

I, obtaining a good approximation to the kind of grouping obtained, depending on the sorting of the rows and columns of S.

2.2. Similarity Decomposition Graphs (SDG). At this stage, we need a practical graphical representation for the decomposition of the SM, one that could extract the relevant information from the clustering process. Here we propose the following. Construct a square grid for all the elements in the sample. The rows and columns of the grid represent each individual, sorted accordingly to the SM. Then fill the points in every coordinate with RGB colors, defined as Red, Green and Blue channels in a $[0, 1]$ scale. In the Red channel, put the intrinsic part of the similarity. In the Blue channel, put the extrinsic part, and in the Green channel put 1 minus the total similarity. Following this scheme, for high similarity areas, the intrinsic part will be colored in red, and the extrinsic in blue. For low similarity areas, the intrinsic part will be colored in orange or brown, and the extrinsic part in green. There are specific patterns that can be recognized and interpreted. These are, mainly:

- Unordered pattern. This pattern appears when the order of the rows and columns in the SM is not directly associated with the clustering, or it is not evident. It is convenient to rearrange the rows and columns of the SM to make the clustering suitable for interpretation.
- Blue blocks A blue block represents a set of observations that mix from one cluster to another, so in average they seem to share a common cluster, but there is no FCR defined.
- Blue blocks with red delimeter. This is the representation of a well formed cluster. The blue block indicates the extent of the cluster, starting with well defined vertical and horizontal red bars, originated by the FCR. In the diagonal, the FCR is represented by a red point, and the bars show that the other members of the cluster relate to it intrinsically. The rest of the block is made up of extrinsic relations, following the transitivity rule.
- Isolated red points in the main diagonal. These represent individuals which are FCR of their own singleton cluster. This means the model is treating them as separate entities, so they can be considered as different of

the rest of the sample, at least in the dimension explained by the covariate that defines the order of the rows and columns of the SM.

- Green areas. Green areas represent pairs of elements which rarely relate to each other, and separate one cluster from other. Evaluating which individuals (specifically their covariates) separate the clusters is the key to determining which covariate may explain the clustering.

All the previous description constitutes a guide, which must be evaluated considering which covariate determines the order of the observations, and wether this covariate is included in the model, or not, and how. A partition with two or more clusters, determined by a covariate that is not being considered in the main specification could advise on considering the mentioned covariate, or the information it represents. If, on the contrary, the covariate is actually included, it may be a sign that the functional form of the covariate is not the most adequate, and one could try, for instance, a quadratic form instead of a linear form.

In the next chapters we will explore these relations in more detail with concrete applications. To conclude this part, let us discuss one more aspect relating SMs and PMs.

3. Choosing a partition for the data

Sometimes it is necessary to choose one specific partition for the data, which is a decision problem. As Bernardo and Smith (1994) state, the elements of a decision problem in the inference context are:

(1) $a \in A$, a set of available "answers" to the inference problem.
(2) $\omega \in \Omega$, a set of unknown states of the world.
(3) $u : A \times \Omega \to \mathbb{R}$, a function attaching utilities to each consequence of a decision to summarise inference in the form of an "answer" a, and an ensuing state of the world, ω.
(4) $p(\omega)$, a specification, in the form of a probability distribution, of current beliefs about the possible states of the world.

In our context, element (4) is given by the SSMM specification, and (2) is the set of all possible partitions in $[n]$. Lau and Green (2007) propose an answer to (3) in terms of a loss function defined on pairwise coincidences, discussed by Binder

(1978), which adequates to our aim to make inference on partitions based on a SM. This loss function considers pairs of items and the cost of clustering together items that should be separate, and also the cost of setting apart items that should be together. For item (1), the huge size of the set of possible partitions for the data makes insufficient any practical sample size of posterior partitions. Here we propose some simple methods to find good candidate partitions based on the expected loss.

3.1. Expected loss and distance.

DEFINITION 10. *(Binder, 1978, Lau and Green 2007) The expected loss function $EL_{a:b}(\cdot)$ for a PM $R = (\rho_{ij})$ given a SM $S = (s_{ij})$, both of size n, and positive weights a and b is defined as*

$$\begin{aligned} EL_{a:b}(R|S) &= \sum_{i,j=1,2,\ldots,n} \{a\ I(c_i \neq c_j)p(c_i = c_j) + b\ I(c_i = c_j)p(c_i \neq c_j)\} \\ &= \sum_{i,j=1,2,\ldots,n} \{a\ (1-\rho_{ij})s_{ij} + b\ \rho_{ij}(1-s_{ij})\} \end{aligned} \tag{27}$$

where a and b represent the relative cost of separating two elements that should be joined, and joining two elements that should be separated, respectively.

The Euclidean distance between a SM $S = (s_{ij})$ and a PM $R = (\rho_{ij})$ is

$$D(R,S) = \sqrt{\sum_{i,j=1,2,\ldots,n} (s_{ij} - \rho_{ij})^2}$$

Noting that $\rho_{ij}(1-\rho_{ij}) = 0$ for any pair (i,j) in $[n] \times [n]$, we have

$$EL_{1:1}(R|S) = D(R,S)^2 + \sum_{i,j=1,2,\ldots,n} s_{ij}(1-s_{ij})$$

so finding the PM R that minimizes $EL_{1:1}(R|S)$ is the same as minimizing $D(R,S)$.

3.2. Sampling partitions. The number of possible partitions for $[n]$ is huge, but always finite, and it is called the *Bell* number, B_n (see Klaska (1997), Rota (1964)). Recursively, Bell numbers satisfy

$$B_{n+1} = \sum_{k=0}^{n} \left(\begin{array}{c} n \\ k \end{array} \right) B_k,$$

and we have, for instance, that the number of partitions for 100 elements approximates to 4.8×10^{115}. Anyway, we can always define a probability distribution for the partitions, by means of assigning a probability value to every possible partition

$p(R_i) = p_i \geq 0,\ i = 1, 2, \ldots, B_n$ subject to $\sum_{i=1}^{B_n} p_i = 1$. Sampling partitions from a huge discrete space has the potencial problem that many possible partitions cannot even get a chance to be sampled, specially when n is big, as B_n can be far greater than any practical sample size for simulations. Nevertheless, we can actually obtain samples of any size, and select from the sampled partitions, for instance, the mode. For this task, taking the distance from sampled partitions to a reference PM or SM can be useful to identify the partitions efficiently. From the Bayesian perspective, selecting the sampled partition with minimum expected loss (Lau &Green, 2007) based on a similarity matrix is the preferred choice.

3.2.1. *Methods to find PMs based on a SM.* SM's do not represent partitions, but, in typical problems, the SM is our approximation to the expected PM, so we may want to find a list of PM's close, in some sense, to S, to choose then the partition with minimum EL.

Let $S = \{s_{ij}\}$ be a SM.

- *Direct method*
  Define a cutpoint $c \in [0, 1]$, for instance $c = 0.5$, and set $\rho_{ij} = \mathrm{I}(s_{ij} >= c)$. A quick method to find a PM from S comes straightforward from (20). It is guaranteed to obtain an actual PM, good enough to get a preliminary result, although it may not be the optimal answer in terms of a chosen expected loss.
- *Cut & Complete*
  Define a cutpoint as before, and complete $R = \{\rho_{ij}\}$ to obtain the transitive closure. This can be used itself as a method to find a candidate partition, or as a step in the next algorithm.
- *Succesive Completion*
  Define a sequence of cutpoints, which can be the list of unique values of the SM. Define a list of partitions based on cutting and completing for every value in the list, and then choose the PM with minimum EL.
- *Minimize & Complete*
  From the definition of the expected loss function, it is easy to see that,

given a SM $S = \{s_{ij}\}$, a relation $R = \{\rho_{ij}\}$ that minimizes the EL can be constructed, defining

$$\rho_{ij} = \mathrm{I}(a(1 - s_{ij}) - bs_{ij} < 0).$$

Then, completing R gives us the result. In the case $a = b$, the EL is minimized when $\rho_{ij} = I(s_{ij} > 0.5)$.

- *Optimizing a PM*

  Sometimes it is possible to optimize a PM, that is, given a starting PM R_0, find a PM R_1 with lower EL. The contribution of each pair (i, j) to the overall EL is $ar_{ij}(1 - s_{ij}) + b(1 - r_{ij})s_{ij}$, where r_{ij} and s_{ij} are the corresponding elements of the PM and the SM, respectively, and a and b are the constants which define the EL. So, one can select a pair (i, j) with a high (maybe the highest) contribution to the EL, and change it, in a merge-split fashion.

  - If $r_{ij} = 1$, setting $r_{ij} = 0$ implies splitting the cluster where i and j belong. Put i and j in new singleton clusters, and for every element k that was related with i, put it in j's cluster if $a(1 - s_{ki}) > bs_{ki}$ and in i's cluster in other case.
  - If $r_{ij} = 0$, set $r_{ij} = 1$ and merge clusters corresponding to elements i and j.

  If the new partition obtained has a lower EL, it can be used as a better choice. If not, try changing another pair. It is recommended to work with a relatively short list of pairs, ordered decreasingly by contribution to the EL, and stop if no better solution is found.

3.2.2. *Example.* Let's continue with the SM from 2.1.1. We saw that the first matrix was not positive definite, and then corrected it to

	1	2	3	4	5
1	1.00	0.30	0.60	0.70	0.70
2	0.30	1.00	0.60	0.30	0.20
3	0.60	0.60	1.00	0.75	0.50
4	0.70	0.30	0.75	1.00	0.90
5	0.70	0.20	0.50	0.90	1.00

To form a PM from that, let's start defining $R = (r_{ij})$ by $r_{ij} = I(s_{ij} > 0.5)$. We obtain

	1	2	3	4	5
1	1	0	1	1	1
2	0	1	1	0	0
3	1	1	1	1	0
4	1	0	1	1	1
5	1	0	0	1	1

We see that this matrix does not represent a real partition, since there are related triplets of the PM which sum 2.

i	j	k	r_{ij}	r_{jk}	r_{ik}	$r_{ij} + r_{ik} + r_{jk}$
1	2	3	0	1	1	2
		4	0	0	1	1
		5	0	0	1	1
	3	4	1	1	1	3
		5	1	0	1	2
	4	5	1	1	1	3
2	3	4	1	1	0	2
		5	1	0	0	1
	4	5	0	1	0	1
3	4	5	1	1	0	2

These triplets have to be completed to satisfy transitivity, and that means changing the matrix so that all mentioned triplets that sum 2 sum 3. The new matrix is

	1	2	3	4	5
1	1	1	1	1	1
2	1	1	1	1	0
3	1	1	1	1	1
4	1	1	1	1	1
5	1	0	1	1	1

This matrix has to be completed again, since the previous operation created more incoherencies:

i	j	k	r_{ij}	r_{jk}	r_{ik}	$r_{ij}+r_{ik}+r_{jk}$
1	2	3	1	1	1	3
		4	1	1	1	3
		5	1	0	1	2
	3	4	1	1	1	3
		5	1	1	1	3
	4	5	1	1	1	3
2	3	4	1	1	1	3
		5	1	1	0	2
	4	5	1	1	1	3
3	4	5	1	1	1	3

Then, completing again, we obtain the maximal partition, with $EL_{1:1} = 8.9$. Let us check now the minimal partition, represented by the identity matrix. Its expected loss is 11.1. The list of possible cuts determined by S is {0.9, 0.75, 0.7, 0.6, 0.5, 0.3, 0.2}. After cutting at 0.9 we obtain a matrix which does not need to be completed, with $EL_{1:1} = 9.5$.

	1	2	3	4	5
1	1	0	0	0	0
2	0	1	0	0	0
3	0	0	1	0	0
4	0	0	0	1	1
5	0	0	0	1	1

After cutting at 0.75, the following matrix is obtained

	1	2	3	4	5
1	1	0	0	0	0
2	0	1	0	0	0
3	0	0	1	1	0
4	0	0	1	1	1
5	0	0	0	1	1

This matrix must be completed, since for transitivity we must have $r_{35} = r_{53} = 1$

	1	2	3	4	5
1	1	0	0	0	0
2	0	1	0	0	0
3	0	0	1	1	1
4	0	0	1	1	1
5	0	0	1	1	1

The EL for this PM is 8.5. After cutting at 0.7 we obtain the same matrix, and after cutting at 0.6 we obtain

	1	2	3	4	5
1	1	0	1	1	1
2	0	1	1	0	0
3	1	1	1	1	1
4	1	0	1	1	1
5	1	0	1	1	1

The completion of this matrix is the maximal PM, whose EL is 8.9. So the choice from this method is the previous PM. Let's see if it can be optimized. The partial contributions to the overall EL is listed now.

	1	2	3	4	5
1	0.0	0.3	0.6	0.7	0.7
2	0.3	0.0	0.6	0.3	0.2
3	0.6	0.6	0.0	0.25	0.5
4	0.7	0.3	0.25	0.0	0.1
5	0.7	0.2	0.5	0.1	0.0

Changing the pairs which add 0.7 to the EL results in merging element 1's cluster with the cluster formed by 3, 4 and 5, resulting in

	1	2	3	4	5
1	1	0	1	1	1
2	0	1	0	0	0
3	1	0	1	1	1
4	1	0	1	1	1
5	1	0	1	1	1

TABLE 1. List of all possible partitions for five elements

Element clusters					$EL_{1:1}$	Element clusters					$EL_{1:1}$	Element clusters					$EL_{1:1}$
1	1	1	1	1	8.9	1	1	1	1	2	10.1	1	1	1	2	1	11.5
1	1	1	2	2	9.5	1	1	1	2	3	11.1	1	1	2	1	1	10.7
1	1	2	1	2	11.9	1	1	2	1	3	11.9	1	1	2	2	1	11.3
1	1	2	2	2	9.3	1	1	2	2	3	10.9	1	1	2	3	1	12.3
1	1	2	3	2	11.9	1	1	2	3	3	10.3	1	1	2	3	4	11.9
1	2	1	1	1	6.5	1	2	1	1	2	10.1	1	2	1	1	3	8.9
1	2	1	2	1	10.7	1	2	1	2	2	11.1	1	2	1	2	3	11.5
1	2	1	3	1	9.9	1	2	1	3	2	11.9	1	2	1	3	3	9.1
1	2	1	3	4	10.7	1	2	2	1	1	7.5	1	2	2	1	2	11.1
1	2	2	1	3	9.9	1	2	2	2	1	9.7	1	2	2	2	2	10.1
1	2	2	2	3	10.5	1	2	2	3	1	9.9	1	2	2	3	2	11.9
1	2	2	3	3	9.1	1	2	2	3	4	10.7	1	2	3	1	1	7.9
1	2	3	1	2	11.5	1	2	3	1	3	10.3	1	2	3	1	4	10.3
1	2	3	2	1	11.1	1	2	3	2	2	11.5	1	2	3	2	3	11.9
1	2	3	2	4	11.9	1	2	3	3	1	9.3	1	2	3	3	2	11.3
1	2	3	3	3	8.5	1	2	3	3	4	10.1	1	2	3	4	1	10.3
1	2	3	4	2	12.3	1	2	3	4	3	11.1	1	2	3	4	4	9.5
1	2	3	4	5	11.1												

This PM has $EL_{1:1} = 6.5$, and is in fact the PM with minimum EL. The list of all possible partitions for 5 elements and their expected loss is shown in table 1.

CHAPTER 3

Application: Bayesian Density Estimation Model

This chapter reviews a well known Bayesian non-parametric model with no covariates. It is adequate to see the prior and posterior clustering mechanism derived from the SSMM specification. The application presented is already a classical example of cluster analysis, representing a good opportunity to see the information derived from Similarity Analysis. We will study the clustering behavior of the data and propose partitions based on the SM. The interpretation of SDGs will be discussed here, too.

1. Bayesian density estimation

Bayesian density estimation comes from the posterior predictive distribution $p(y_{n+1}|y_1,\ldots,y_n)$. In the context of the general model (1) without considering covariates, letting $y=(y_1,\ldots,y_n)$ represent the observations, and $\theta=(\theta_1,\ldots,\theta_n)$ their corresponding parameters coming from a SSM, we have

$$p(y_{n+1}|y)=\int p(y_{n+1}|\theta_{n+1})p(\theta_{n+1}|\theta,y)p(\theta|y)d\theta d\theta_{n+1} \tag{28}$$

From a Gibbs Sampling point of view, $(\theta_{n+1}|\theta,y)$ comes from the predictive rule

$$p(\theta_{n+1}|\theta,y)\propto\sum_{j=1}^{k_n}p_j(N_n)\delta_{\theta_j^*}(\theta_{n+1})+\ p_{(k_n+1)}(N_n)G_0(\theta_{n+1}) \tag{29}$$

where k_n is the currently imputed number of clusters for $(\theta_1,\ldots,\theta_n)$ and θ_j^*, $j=1,\ldots,k_n$ are the currently imputed locations for each cluster, which in turn come from (16). p_j are the weights of the predictive rule defined in (5), for $j=1,\ldots,k_n+1$. Having a sampled value for θ_{n+1}, $(y_{n+1}|\theta_{n+1},y)$ is sampled from (30), by conditional independence. The formulas for sampling the parameters are intended for an individual allocation Gibbs sampling scheme. For details, see section A.

1.0.3. *Escobar & West (1995) density estimation model.* Here we have $\theta_i = (\mu_i, V_i)$. We consider

$$\begin{aligned} y_i &\sim \mathrm{N}(\mu_i, V_i) \\ (\mu_i, V_i)|G &\sim G \\ G &\sim \mathrm{DP}(M, G_0) \\ G_0(V_i) &\equiv \mathrm{IG}(\nu_0, \nu_1) \\ G_0(\mu_i|V_i) &\equiv \mathrm{N}(m, \tau V_i) \\ m &\sim \mathrm{N}(a, A) \\ \tau &\sim \mathrm{IG}(\lambda_0, \lambda_1) \end{aligned} \tag{30}$$

This model differs form our basic model (1) in the absense of covariates, and the additional hyperpriors for the parameters in the baseline measure G_0, in the part that concerns the means μ_i, $i = 1, \ldots, n$. For SSMMs in general, parameters can be sampled based on the following scheme: first, update the partition of $(\theta_1, \ldots, \theta_n)$ from (19), for $i = 1, 2, \ldots, n$, and then sample the locations of the clusters from (16). In this particular case, the posterior predictive rule for cluster configuration is based on

$$\begin{aligned} P(c_i = j|c_{-i}, \mu_j^*, V_j^*) &\propto N_j (2\pi V_j^*)^{-1/2} \exp\left[-\frac{1}{2V_j^*}(y_i - \mu_j^*)^2\right] \quad \text{for } j = 1, \ldots, k_{-i} \\ P(c_i = k_{-i} + 1|c_{-i}) &\propto M(2\pi(\tau+1))^{-1/2} \frac{\nu_1^{\nu_0}}{\Gamma(\nu_0)} \Gamma(\nu_0 + 1/2) \\ &\quad \left[\frac{y_i^2 + m^2 + 2(\tau+1)\nu_1 - 2my_i}{2(\tau+1)}\right]^{-(\nu_0+1/2).} \end{aligned} \tag{31}$$

For cluster locations update we have

$$V_j^*|\mu_j^*, c \sim \mathrm{IG}\left(\nu_0 + N_j/2, \frac{1}{2}\sum_{c_i=j}(y_i - \mu_j^*)^2 + \nu_1\right) \tag{32}$$

$$\mu_j^*|V_j^*, c \sim \mathrm{N}\left(\frac{\tau\sum_{c_i=j} y_i + m}{\tau N_j + 1}, \frac{\tau V_j^*}{\tau N_j + 1}\right) \tag{33}$$

Secondly, the rest of the parameters is updated. For m and τ we have

$$m|\mu^*, V^* \sim \mathrm{N}\left(\left(\frac{1}{\tau}\sum_{j=1}^{k}\frac{1}{V_j^*}+\frac{1}{A}\right)^{-1}\left(\frac{1}{\tau}\sum_{j=1}^{k}\frac{\mu_j^*}{V_j^*}+\frac{a}{A}\right), \left(\frac{1}{\tau}\sum_{j=1}^{k}\frac{1}{V_j^*}+\frac{1}{A}\right)^{-1}\right) \tag{34}$$

$$\tau|\mu^*, V^* \sim \mathrm{IG}\left(k/2+\lambda_0, \lambda_1+\frac{1}{2}\sum_{j=1}^{k}\frac{(\mu_j^*-m)^2}{V_j^*}\right) \tag{35}$$

Model (30) bases the distribution of the data entirely in the nonparametric part. The *tuning* of the model specification comes from the baseline measure G_0, which in turn depends on the hyperparameters m and τ, and the mass parameter M. So, it is very important to specify the hyperparameters a, A, λ_0 and λ_1 based on the available information. Escobar and West (1995) also propose a prior specification $M \sim \Gamma(a_0, b_0)$, a Gamma prior with shape $a_0 > 0$ and scale $b_0 > 0$. Learning about M is important due to the relation of this parameter with the mechanism for creating clusters. Their scheme for updating this parameter is shown in Appendix A. For further details, refer to Escobar and West (1995).

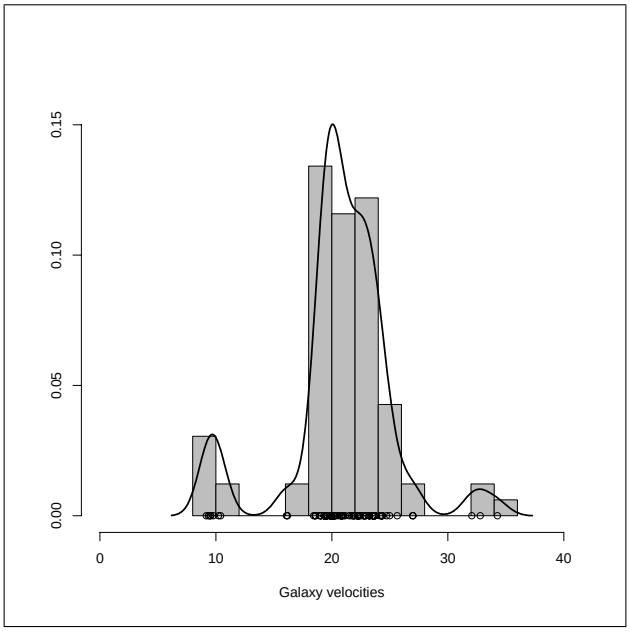

FIGURE 1. Galaxy data

TABLE 1. Galaxy dataset

Observation	1	2	3	4	5	6	7
Value	9.172	9.35	9.483	9.558	9.775	10.227	10.406
Observation	8	9	10	11	12	13	14
Value	16.084	16.17	18.419	18.552	18.6	18.927	19.052
Observation	15	16	17	18	19	20	21
Value	19.07	19.33	19.343	19.349	19.44	19.473	19.529
Observation	22	23	24	25	26	27	28
Value	19.541	19.547	19.663	19.846	19.856	19.863	19.914
Observation	29	30	31	32	33	34	35
Value	19.918	19.973	19.989	20.166	20.175	20.179	20.196
Observation	36	37	38	39	40	41	42
Value	20.215	20.221	20.415	20.629	20.795	20.821	20.846
Observation	43	44	45	46	47	48	49
Value	20.875	20.986	21.137	21.492	21.701	21.814	21.921
Observation	50	51	52	53	54	55	56
Value	21.96	22.185	22.209	22.242	22.249	22.314	22.374
Observation	57	58	59	60	61	62	63
Value	22.495	22.746	22.747	22.888	22.914	23.206	23.241
Observation	64	65	66	67	68	69	70
Value	23.263	23.484	23.538	23.542	23.666	23.706	23.711
Observation	71	72	73	74	75	76	77
Value	24.129	24.285	24.289	24.366	24.717	24.99	25.633
Observation	78	79	80	81	82		
Value	26.96	26.995	32.065	32.789	34.279		

2. Galaxy data

The data consists on $n = 82$ measured velocities (in 10^3 km/s) of galaxies from six well-separated conic sections in space, relative to our own galaxy. A histogram with kernel-smoothing density estimation for the data is shown in figure (1). At first glance, the data shows a multimodal distribution, leading the observer to suppose the data may be grouped in three or four, or up to six clusters. The interest focuses, then, in estimating the probability density and make inference about the clustering, from a Bayesian point of view, based on the observed data and available information about it. The complete dataset is shown in table 1, to make the cluster allocation easier to understand.

3. Model specification

The following values are assigned to the hyperparameters in (30), following Escobar & West (1995): $\nu_0 = 2$, $\nu_1 = 1$, $a = 20$, $A = 1000$, $\lambda_0 = 0.5$, $\lambda_1 = 50$. Two prior distributions were considered for M. First model ($DP1$) considered $a_0 = 2$,

$b_0 = 4$, which puts a prior expected value for M of 0.5. A second model ($DP2$) specified $a_0 = 10$, $b_0 = 2$, with prior expectation of 5, intended to favour partitions with a higher number of clusters. For both models, MCMC sample size was 10000, after *burning* 2000 initial samples and thinning 150 samples each time.

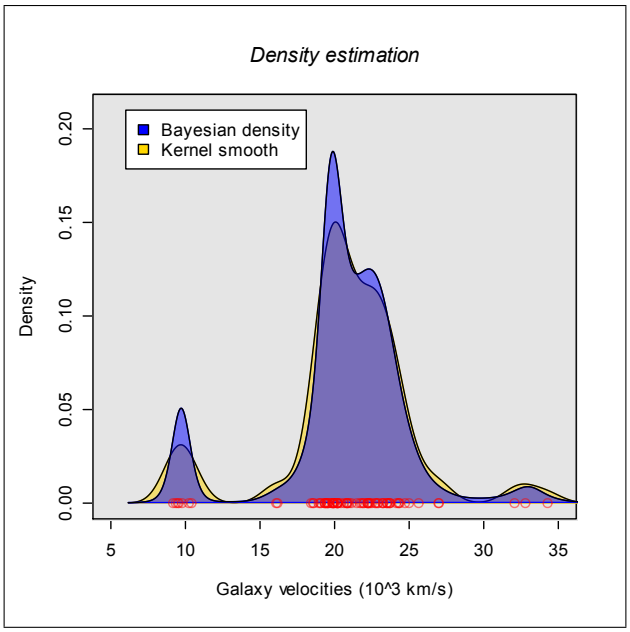

FIGURE 2. Density estimation from model DP1

4. Partitions

The formulation of Model (30) is not focused explicitly in making inference about the partitions. Nevertheless, the SSMM structure is based on considering the latent partition structure of the data, and posterior partitions are obtained. In Chapter 2, various methods to obtain partitions for the data based on an approximation of their expectation are given. Table 2 summarizes partitions obtained for Model $DP1$ by four criteria: mode of sampled partitions, sampled partition with minimum expected loss $EL_{1:1}$, the partition obtained from the similarity matrix by application of the *Succesive Completion (SC)* method, and the partition obtained from the *Minimize & Complete (MC)* method. Two extra partitions are presented ($Opt1$ and $Opt2$), obtained from *optimizing* the second partition. For the decision problem implied in the choice of partitions, we suggest setting default values $a = 1$

TABLE 2. Partitions from Model DP1

PM method	$EL_{1:1}$	Clusters					
		1	2	3	4	5	6
Mode	1880.176	1-7	8-79	80-82			
Best sampled	1872.610	1-7	8	9	10-42	43-79	80-82
SC	1876.258	1-7	8-9	10-79	80-82		
MC	1880.176	1-7	8-79	80-82			
Opt1	1871.48	1-7	8-9	10-42	43-79	80-82	
Opt2	1859.476	1-7	8-9,43-79	10-42	80-82		

and $b = 1$ in (27), unless one has specific reasons to do otherwise. A conservative guess from visual appreciation is consistent with the mode of the sampled partitions. Its sampling frequency was 1879/10000. This partition was also obtained by the MC method. The sampled partition with minimum EL was sampled with a frequency of 1/10000, so the possibility that it could not have been sampled can not be discarded. The partition obtained with the SC method was actually sampled, with a frequency of 13/10000. Neither of the optimized partitions were sampled. Although $Opt2$ has the best EL, it is not plausible for the data, so the best partition found is $Opt1$. The posterior mean number of clusters was 5.08 (s.d. 1.8). For Model DP2, no Mode partition could be selected, since the highest frequency for sampled partitions was 2, a value obtained by almost 30 configurations. The sampled partition with minimum EL obtained consisted of 13 clusters, with an $EL_{1:1}$ of 608.99. The partition obtained by the SC method had an $EL_{1:1}$ of 594.12 and it consisted of 17 clusters. This method was able to find 13 non-sampled partitions with lower EL than the sampled ones, with number of clusters varying from 12 to 26. No optimization was done for this model's partitions. The posterior number of clusters was 11.49 (s.d. 3.19). Density estimation is shown on figures (2) and (3). Further information for mass parameter and posterior number of clusters is given in figures (4) and (5).

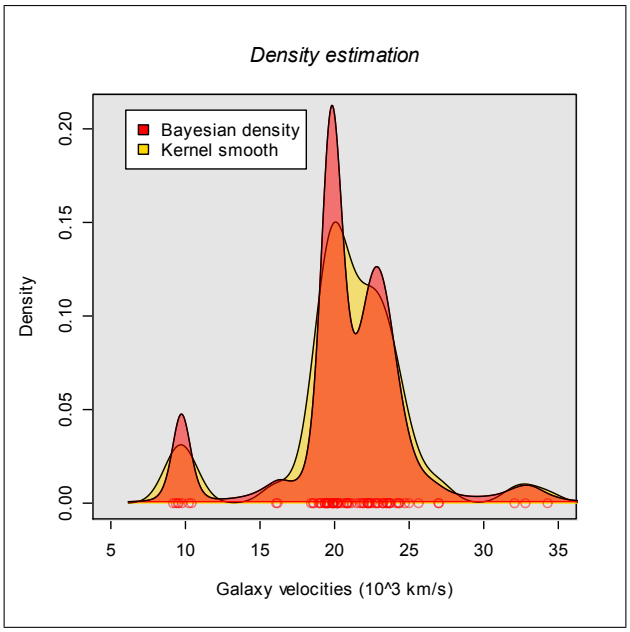

FIGURE 3. Density estimation from model DP2

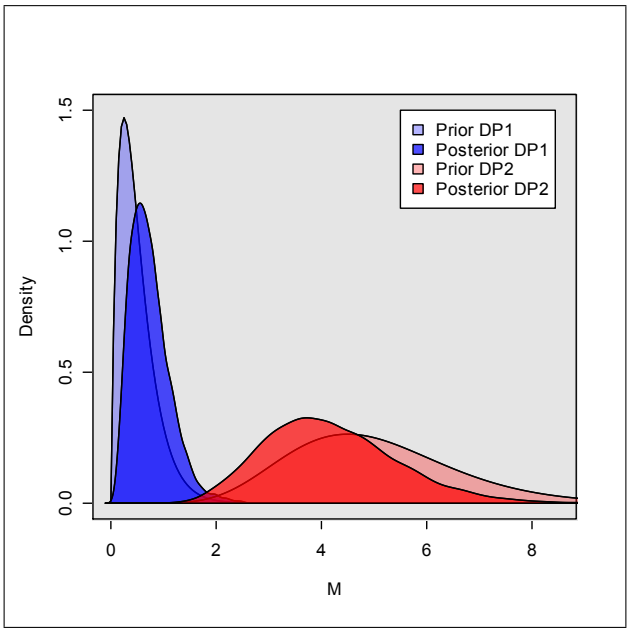

FIGURE 4. Prior and posterior densities for mass parameter M

5. Similarity analysis

From the similarity matrices (figures 6 and 7) it is possible to appreciate the clustering behavior of the observations. High similarity areas are colored in white and yellow, low similarity in red. In model DP1, three well defined clusters can be distinguished, which correspond to partitions Mode and MC in table (2). The SM for model DP1 indicates that the second big cluster, corresponding to observations 8 to 79, can be split in two, three or four sub clusters, with less certainty than the first mentioned three clusters. One likely form for this splitting is detailed in partition $Opt1$ from table (2).

It is interesting to see how the choice of a partition based on the mean of the sampled partitions (the SM) can differ from what is seen from the sampled partitions themselves. The forming of a new cluster is in direct relation with the autonomy of its first element, due to the nature of the Polya urn mechanism (see (5) and (15)). In figures (8) and (11), the autonomy and an empirical estimation of the probability of creating new clusters, respectively, can be appreciated for model DP1. It must be noted that such probability is conditional on both, current cluster configurations and sampling order. This is important to provide an adequate interpretation. Both figures show the same information in this case, since the order of the observations is the same as the order of the cluster labels, although the autonomy is more sensitive. We can see the strong trend of observations 8, 10 and 80 to form new clusters (the first observation always forms a new cluster, obviously), as seen in the partitions from table (2). But, when the big central cluster is split, all of the mentioned candidate partitions consider the split beginning in observation 43. The selection is based on expected loss, that is to say, on the SM. But observing the autonomy and new cluster probability plots, it so happens that observation 43 is in fact the locally less autonomous observation. Instead, the observations with higher posterior probability of leading clusters are 40 and 46. The reason for this becomes clear when looking at the graphical representation of PMs. Let us consider as example an extreme case. Suppose we have 25 elements, and one half of the sample consist on one partition, and the other half on a different one, as shown in Figure 9. The first conclusion from seeing the SM points to a partition of three clusters, although we know the sample concentrates in two partitions of two clusters

each, shown on the left and right side of the graph. When two frequent clusters overlap, the SM may reflect relations of pairs of individuals laying at or close to the "boundaries" on both clusters. These relations, in the SM, are extrinsic, following different FCRs, corresponding to the overlapping clusters. This information is considered in the decomposition of the SM, and the "real" clusters can be better identified by their FCRs. Figure 10 shows the decomposition of the SM in the example, and an autonomy plot. It can be seen there that the supposed central cluster is less important than the ones in the extremes, based on the definition of the FCRs and the autonomy of elements 1 and 15. So, in the Galaxy data, based on Similarity Analysis, it can be concluded that, instead of having a central cluster splitted at observation 43, it is more plausible that observations 41 to 45 belong alternatively to a cluster represented by observation 40, or to a cluster at the left of another cluster, in turn represented by observation 46. Let us consider partition $Opt1$, and put observations 40, 41, and 42 in cluster 4. This partition obtains an EL of 1880.201. If we put observations 43, 44 and 46 on cluster 3, the obtained EL is 1880.655. So, these partitions are more distant from the SM than $Opt1$ or, equivalently, their $EL_{1:1}$ is higher. If the decision is based strictly on EL, then the best choice found is $Opt1$. For model DP2, the autonomy and new cluster probability are shown in figures (12) and (13). This model encourages the observations to form new clusters in a higher way that DP1. The change is readily captured in the autonomy plot. The empirical probabilities for creating new clusters, on the other side, are also higher than their counterparts for model DP1. A comparison between autonomy and empirical probability of creating new clusters is shown in figure (14). The similarity decomposition for both models is reflected in figures (15) and (16). It can be appreciated that most of the clustering seen in the SM is due to coherency coming from the transitivity rule, and the intrinsic part reveals which observations actually *lead* the process. In figure (15), a clear and homogeneus cluster for observations 1 to 7 can be observed. It is lead, naturally, by element 1, marked in red, followed by a blue block that completes the cluster. From 8 to 80, one can appreciate a second big cluster, subdivided by three or four less important subclusters. From 8 to 9, and from 10 to a not well defined extent, two clusters lead by observations 8 and 10 can be seen. Somewhere around

element 43, a new cluster is started, but there is no recognizable FCR. This means that, although these observations tend to group, the cluster memberships mix in the MCMC sample, so it is not an homogeneus cluster. A similar situation can be observed for observations 79 and 80. Elements 81, 82 and 83 clearly form a well defined cluster. Note that in SDGs the vertical axis follow the traditional order for plots, and not the matrix order, as in plots for SMs. In the SDG for model $DP2$ (16), it can be seen that the model favors partitions with more clusters, but this is reflected in a greater tendency of the observations to mix, and this can be seen in the greener color of the central clusters, plus the fact that no FCR can be distinguished in their zone.

6. Density estimation model extended to Pitman-Yor process

Pitman-Yor (PY) models are another case of SSM, and include the DP as a particular case ($\alpha = 0$, $M = 1$). See (11) for definition. These models extend the DP to much wider possibilities. Here, as illustrative examples, we will fit two models based on PY priors for the clustering process to the Galaxy data. The first model, called $PY3$, is defined considering $\alpha = -1$ and $M = 3$. When α is negative, the prior probability of creating new clusters is $(M + k\alpha)/(n + M)$, with α equal to $-M/m$, for some positive integer m. The mentioned probability is then $M(1 - k/m)/(n + M)$, which is null for $k = m$. So the clustering process can only sample partitions with a maximum of m clusters. In our case, that number is 3. The second model, $PY4$, considers $\alpha = 0.9$ and $M = -0.1$. The prior probability of forming new clusters is $(0.9k-0.1)/(n-0.1)$. A priori, this specification does not restrict the number of clusters as the previous one, but, if a relatively big number of clusters is reached, it tends to maintain that situation. For model $PY3$, the partitions consist in an almost complete separation of the observations in the three main clusters discussed before (figure 18), commanded by observations 1, 8 and 81 (figure 19). The rigid clustering prior determines a very clear partition in the data, as can be seen in the SDG (20). The number of clusters in the sample was 3, almost constantly. Model $PY4$, in contrast, is much more flexible. The mean of the number of clusters in the sampled partitions was 20.7 with a standard deviation of 10.2. In the previous models we observed a tendency to group the observations in three main clusters. The second cluster, in turn, showed a tendency to split in two or

three, when model specification allowed it. In model $PY4$, given the resulting highly partitioned cluster configurations, one could expect an important mixing in clusters. The Bayesian density estimation (figure 21) shows, in contrast, a multimodality that does not differ too much from the models based on DP specifications. Figure (23) shows the SDG. There, we can see that the big number of sampled clusters traduces in the presence of a high number of autonomous individuals. Elements 8, 9, 10 and the last 5 or 6 observations have a clearly independent behavior, as seen in the corresponding red spots in the diagonal of the SDG, with no clusters surrounding them. The autonomy plot (figure 22) compliments this information. It can be seen that, in summary, the posterior density estimation of figure (21) is strongly influenced by individual contributions.

7. Discussion

Model $DP1$ resembles the specifications of Escobar and West (1994), with similar results. Similarity Analysis represents a contribution towards a better understanding and interpretation of the clustering mechanism underlying in SSMMs. We studied three additional models and compared their clustering behaviors. One model ($DP2$) was based on the DP, like $DP1$, but specifying a prior for the mass parameter that favours a high number of clusters. This resulted in an larger number of clusters, and an increased probability for some elements to be relocated in different clusters. Nevertheless, Similarity Analysis pointed to similar conclusions compared with the first model, proving to be robust to variability in the sampled partitions. The MCMC sample size of 10000 observations was clearly insufficient to obtain a fair representation of all plausible partitions under this specification. The best partitions obtained, in terms of EL, were sampled only once or twice in the whole sample. Nevertheless, the information provided by the sampled partitions was enough to get conclusions from Similarity Analysis. The last two models considered were based on the PY process. Mixing on PY models brings the possibility of considering the number of clusters in the clustering mechanism. We used this feature in two ways. First, we explored a model in which the number of clusters was bounded above by three with probability one. Under this specification, the model showed an almost constant partition of three well defined clusters, consistent with the information provided by the previous models. Secondly, we specified a model

TABLE 3. Summary statistics for some posterior parameters in Bayesian Density Estimation models

	Mean	S.Dev.	P05	P10	P50	P90	P95
DP1							
Pred. μ	20.77	4.28	9.76	19.21	21.43	23.14	24.42
Pred. V	4.52	57.32	0.28	0.36	3.29	5.78	6.74
m	17.41	6.48	7.61	10.27	17.37	24.52	27.46
τ	211.41	280.39	35.18	44.30	122.52	456.03	721.15
N. Clusters	5.09	1.80	3.00	3.00	5.00	7.00	8.00
M	0.74	0.38	0.24	0.31	0.68	1.25	1.45
α	0.00						
DP2							
Pred. μ	20.79	4.89	9.80	15.78	20.66	24.05	27.10
Pred. V	3.45	38.70	0.25	0.30	0.69	3.20	5.79
m	19.34	3.00	14.26	15.60	19.43	22.89	24.01
τ	85.43	74.48	27.28	32.50	66.33	155.06	200.32
N. Clusters	11.49	3.19	7.00	8.00	11.00	16.00	17.00
M	4.18	1.26	2.36	2.69	4.05	5.84	6.45
α	0.00						
PY3							
Pred. μ	20.86	4.36	9.74	20.75	21.38	21.80	22.13
Pred. V	4.23	1.67	0.35	0.67	4.56	5.80	6.16
m	17.13	9.76	2.06	6.15	16.91	28.44	32.60
τ	396.97	439.93	64.09	82.88	244.43	1037.26	1669.22
N. Clusters	3.00	0.00	3.00	3.00	3.00	3.00	3.00
M	3.00						
α	-1.00						
PY4							
Pred. μ	20.46	4.82	9.79	14.31	21.05	23.46	25.91
Pred. V	1.96	7.65	0.26	0.31	0.78	3.81	4.76
m	19.72	3.30	13.95	15.63	20.10	23.20	24.29
τ	56.55	62.53	14.52	18.48	42.38	102.03	138.12
N. Clusters	20.70	10.18	8.00	10.00	19.00	34.00	41.00
M	-0.10						
α	0.90						

that allowed for extra flexibility in the choice of number of clusters based, additionally, on the current number of clusters. This resulted in a different clustering behavior than in model $DP2$. This time, the higher number of clusters expressed in the formation of various individual clusters, decreasing the mixing, as shown in Similarity Analysis. General statistics for selected posterior parameter estimations for all models are shown in table 3.

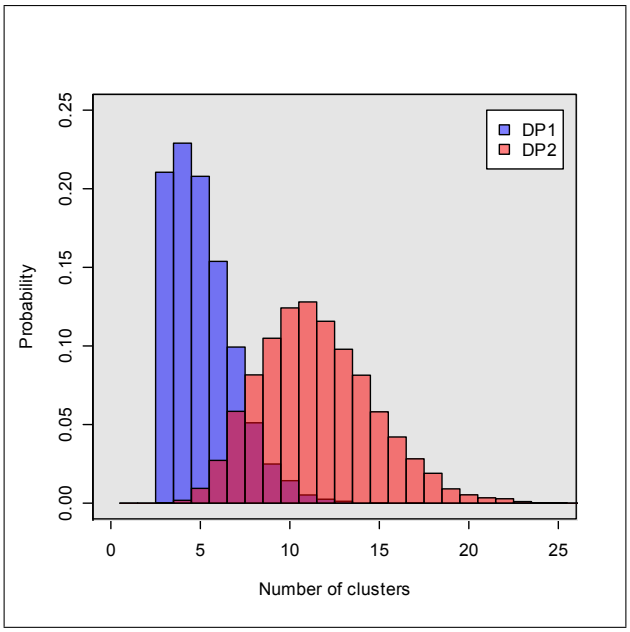

FIGURE 5. Posterior number of clusters for models DP1 and DP2

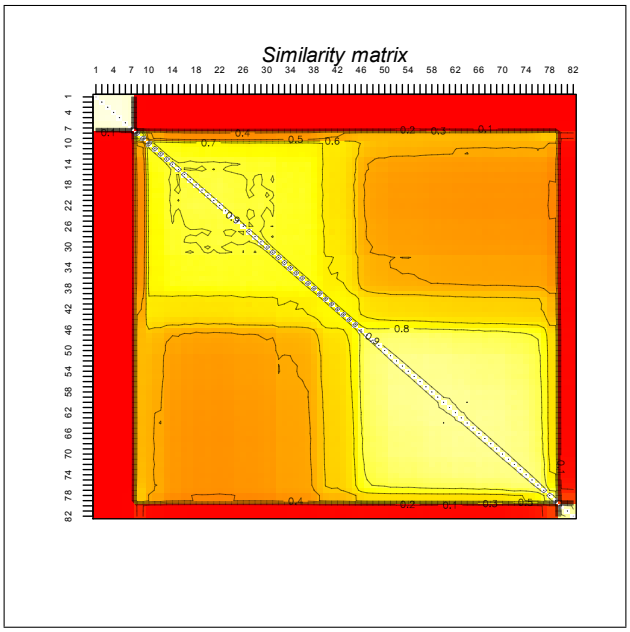

FIGURE 6. Similarity matrix for model DP1.

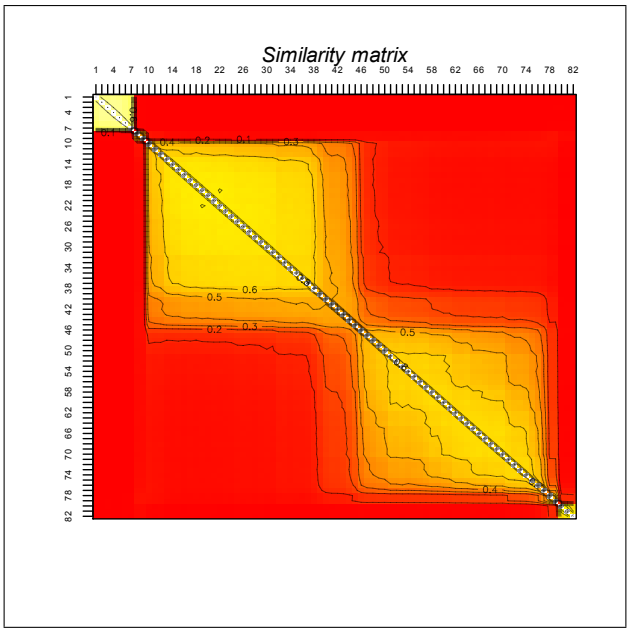

FIGURE 7. Similarity matrix for model DP2

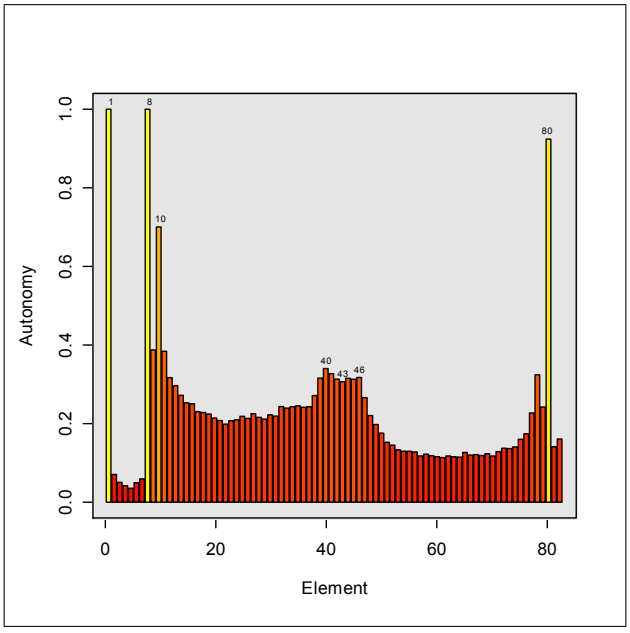

FIGURE 8. Autonomy plot for model DP1

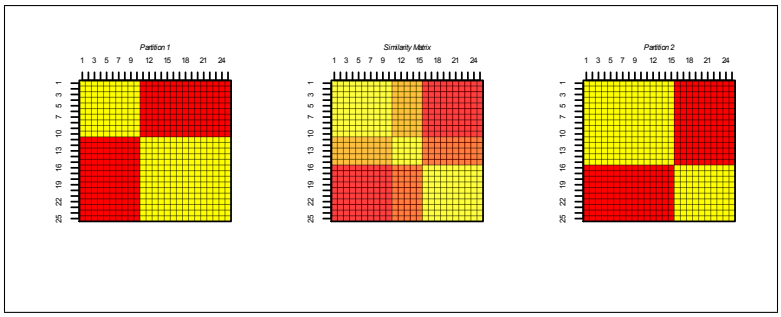

FIGURE 9. Example of two overlapping partitions and the SM.

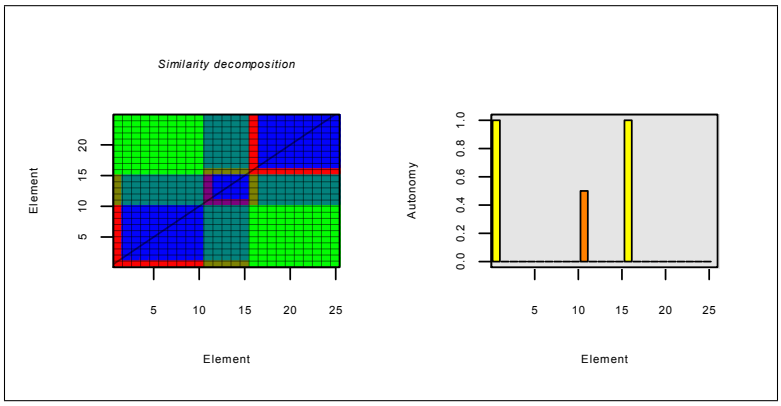

FIGURE 10. SDG and autonomy plot for example of overlapping partitions.

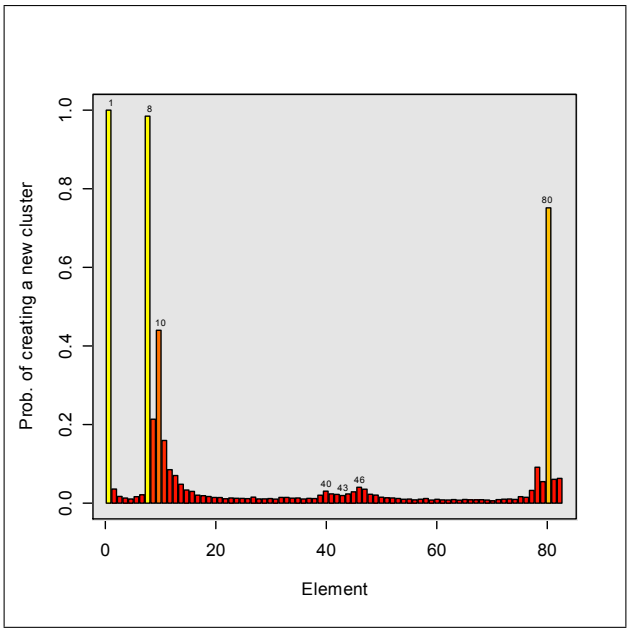

FIGURE 11. Empirical probability of creating a new cluster for model DP1

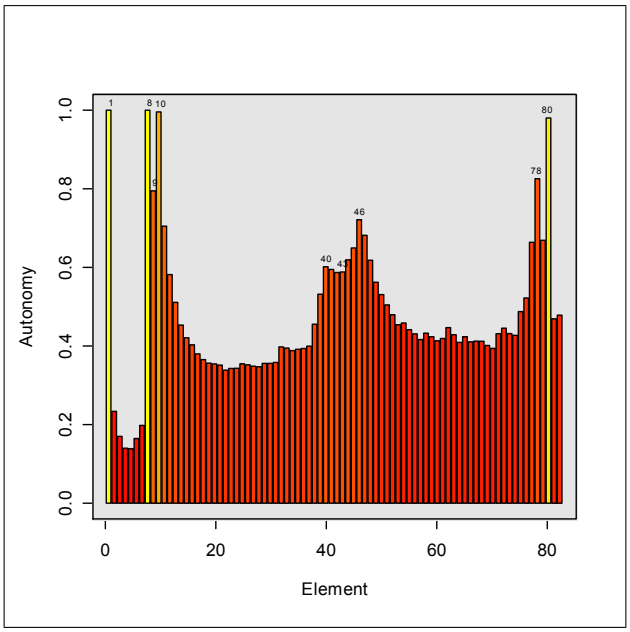

FIGURE 12. Autonomy plot for model DP2

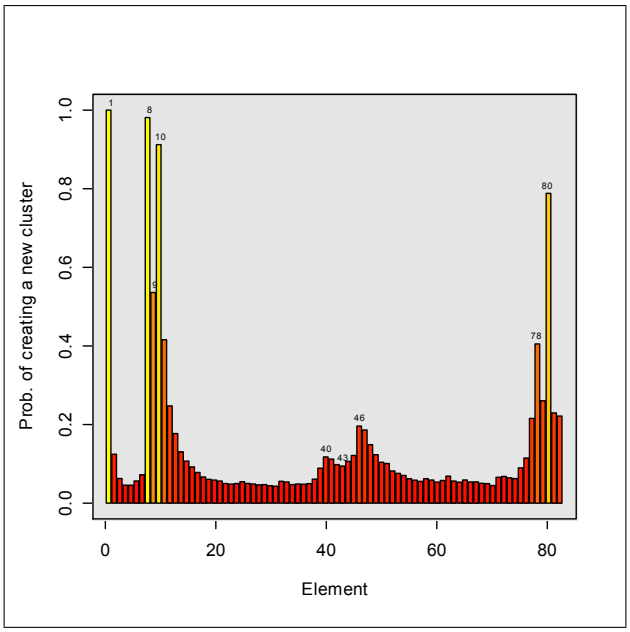

FIGURE 13. Empirical probability of creating a new cluster for model DP2

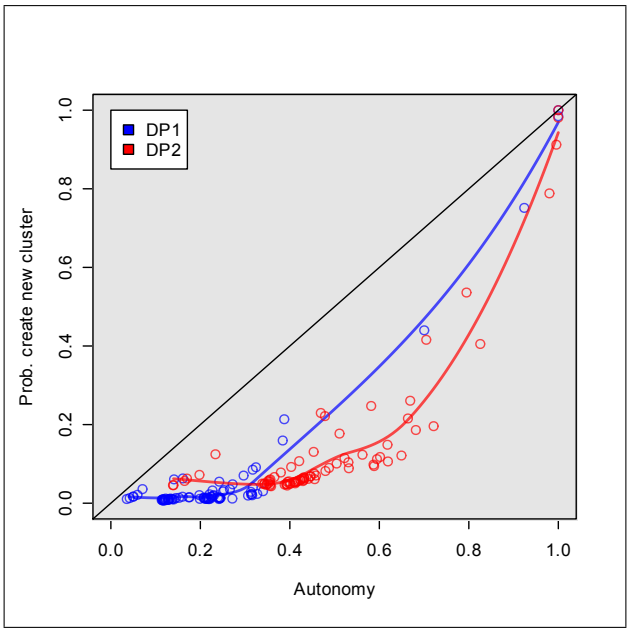

FIGURE 14. Empirical probability of creating new clusters vs autonomy

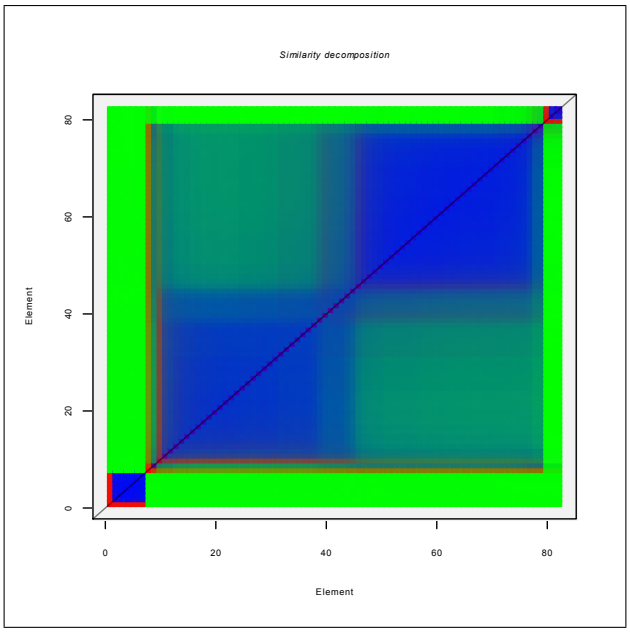

FIGURE 15. Similarity Decomposition Graph for model DP1

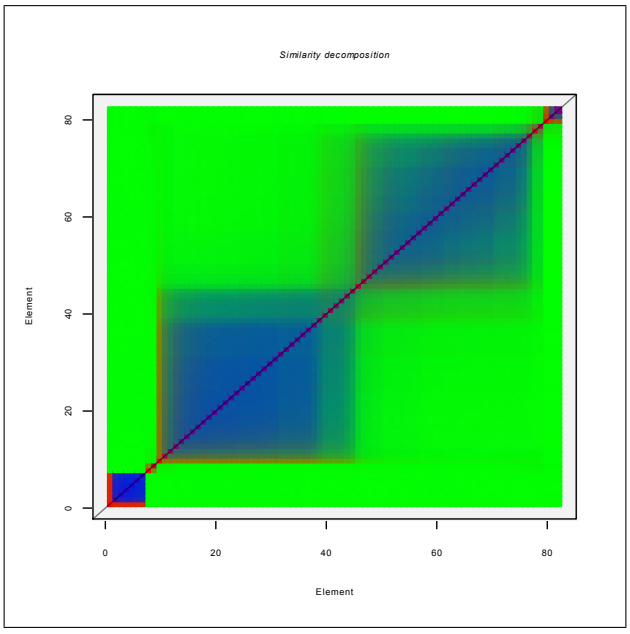

FIGURE 16. Similarity Decomposition Graph for model DP2

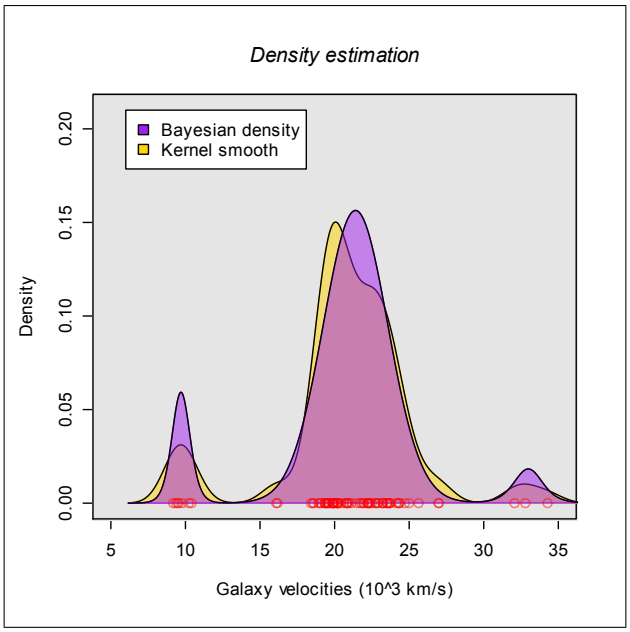

FIGURE 17. Density estimation for model PY3

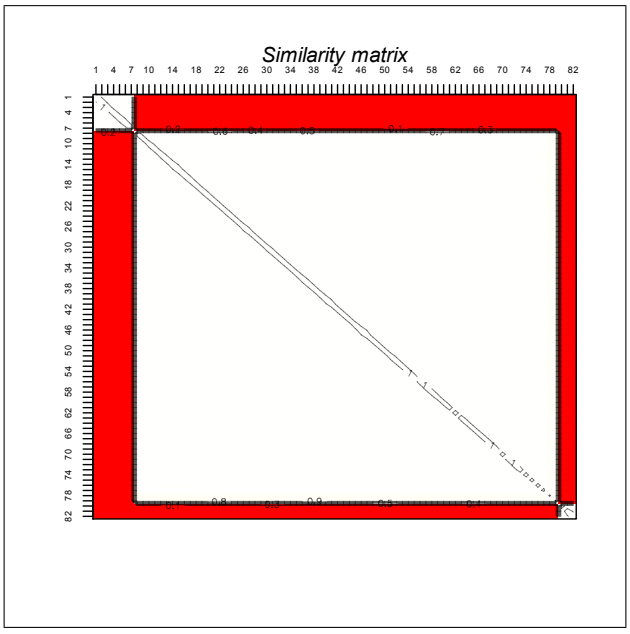

FIGURE 18. Similarity matrix for model PY3

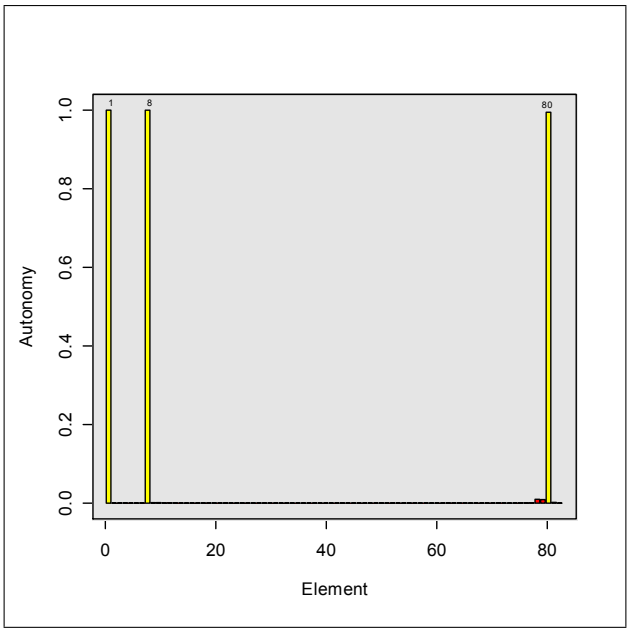

FIGURE 19. Autonomy plot for model PY3

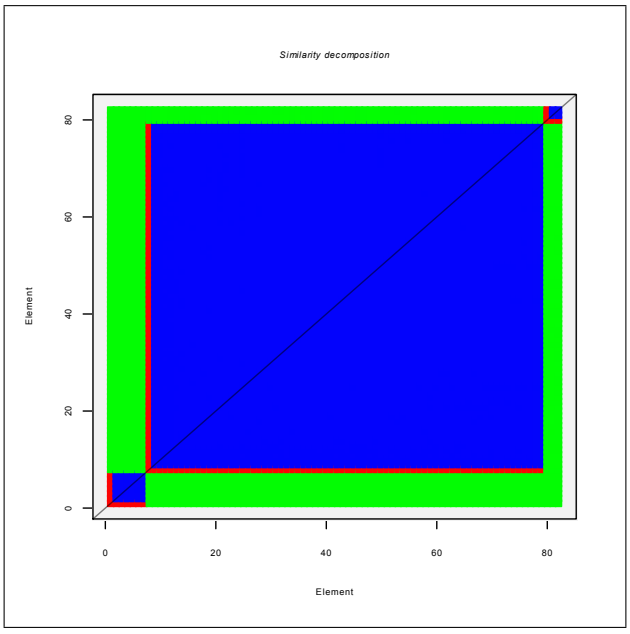

FIGURE 20. Similarity Decomposition Graph for model PY3

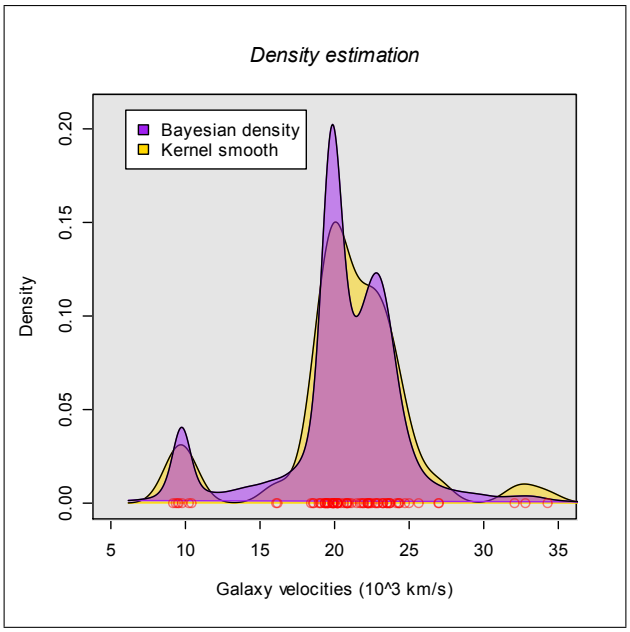

FIGURE 21. Density estimation for model PY4

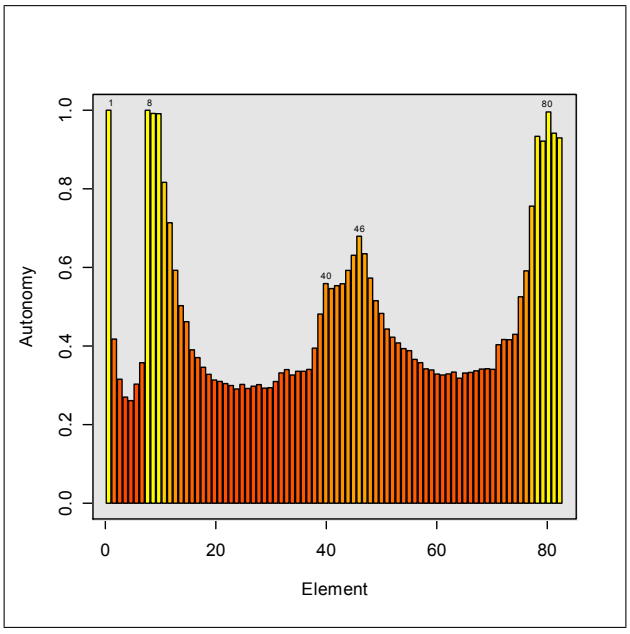

FIGURE 22. Autonomy plot for model PY4

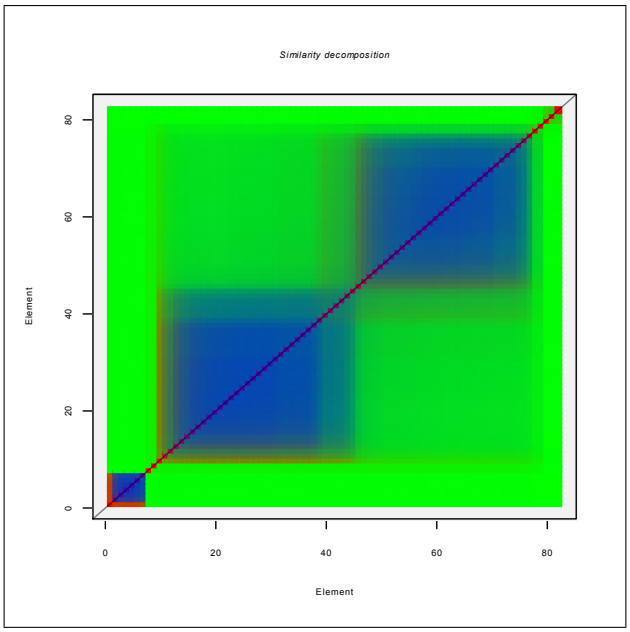

FIGURE 23. Similarity Decomposition Graph for model PY4

CHAPTER 4

Application: Linear Regression Model

In the previous chapter, we explored in detail the clustering mechanism in model (1) in the absense of covariates. Here we will study the behavior of the clustering process when covariates are considered. For that purpose, a multiple linear regression model is specified. The model was conceived keeping in mind the ANOVA-DDP model used by De Iorio et al (2004) as a starting point. However, we do not restrict to categorical covariates representing levels of a factor. We add a continuous covariate and study its effect in the configuration of the partitions, based on FCRs. We use simulated data, to know in advance the *true* relation between response and covariates. We will study posterior cluster configurations under different combinations of covariates specified in the model, and compare information provided by Similarity Analysis in each case. We also present an application to outlier detection to the Forbes data (1857) using Similarity Analysis.

1. Statistical model

We consider the linear regression specification

$$y_i \sim \mathrm{N}(\alpha_i^T d_i, V_i). \tag{36}$$

The mean for each individual is defined as a linear combination of a vector

$$\alpha_i = (\alpha_{i0}\ \alpha_{i1}\ \cdots\ \alpha_{i(q-1)})^T$$

of individual parameters, and a vector

$$d_i = (1\ x_{i1}\ \cdots\ x_{i(q-1)})^T$$

based on individual covariates. We assume the pairs (α_i, V_i) come from a Dirichlet process centered in a base distribution $G = G_1 G_2$, with

$$G_1(\alpha_i) \equiv \mathrm{N}_q(\mu, \Sigma) \tag{37}$$

$$G_2(V_i) \equiv \mathrm{IG}(\nu_0, \nu_1) \tag{38}$$

We need to determine the effect of covariates in the clustering process. For that purpose, all sources of variability in the posterior partitions have to be taken in consideration in the model specification. To introduce additional flexibility in the model, we specify a prior distribution for the parameters in the baseline measure considering

$$(\mu, \Sigma) \sim \mathrm{N}_q\mathrm{IW}_{\eta_0}(\mu_0, \Sigma/\kappa_0; \eta_0, \Lambda_0).$$

To consider a flexible prior for the number of clusters, we also specify a prior distribution over the mass parameter $M \sim \mathrm{Gamma}(a, b)$, as proposed by Escobar and West (1994).

2. Simulated data

We considered a set of 100 simulated observations $y_1, \ldots, y_n$ with two covariates: a group indicator $g_i \in \{0, 1\}$ and a continuous covariate x_i. 60 observations were assigned to group $g = 0$ and 40 to group $g = 1$. Covariates were generated uniformly $U(0, 10)$. For group 0, the simulated observations came from $y_i = 2x_i + 1 + e_i$, e_i i.i.d. $N(0, 1)$, $i = 1, \ldots, 60$. For group 1, $y_i = 0.5x_i - 2 + e_i$, e_i i.i.d. $N(0, 0.5)$, $i = 61, \ldots, 100$. That is, we specified two groups, with different intercepts and slopes. The individual variability is also slightly different between both groups, but constant within each group. Both groups are clearly differentiated, and the effect of the continuous covariate x is evident (Figure 1). In terms of partitions, we know in advance that the individuals are grouped based on their intercepts and their relation with the covariate, expressed by the different slopes. This fact relates clustering with with covariates g_i and x_i, respectively. Individual variability represents an additional possibility of classification, which does not depend on the covariates. The differentiation of the groups becomes more clear at higher values of x.

Four specifications were considered in (36). In Model 1, covariates included are intercept and group effect, with base prior defined by $\mu = (0, 0)^T$. Model 2

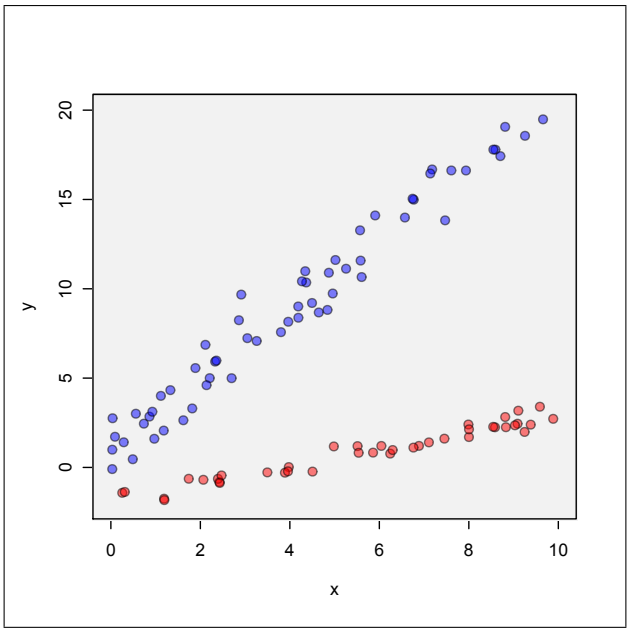

FIGURE 1. Simulated data.

includes intercept and x, with base prior $\mu = (0, 1)^T$. Model 3 includes both group effect and x as covariates, $\mu = (0, 0, 1)^T$. Model 4 includes an additional interaction between group and x, with base mean $\mu = (0, 0, 1, 1)^T$. For all models, the rest of the prior specification was completed setting $\eta_0 = q + 2$, $\Lambda_0 = I$, $\mu_0 = 0$, $\kappa_0 = 1$, $\nu_0 = 1$, $\nu_1 = 1$, understanding every vector of parameters with the corresponding dimensions. Mass parameter prior distribution was based on $a = 1$, $b = 1$.

3. Gibbs Sampling details

Model (36) is non-conjugate, so there is no explicit formulation in (19) for the probability of creating new clusters. To update partitions, *no-gaps* algorithm (MacEacher and Müller, 1998) was used. See Appendix A for details. Cluster

locations are sampled from

$$\begin{aligned}\alpha_j^*|V_j^* &\sim \mathrm{N}_q\left(\left(\frac{1}{V_j^*}D_j^TD_j+\Sigma^{-1}\right)^{-1}\left(\frac{\tilde{y}_j}{V_j^*}D_j+\Sigma^{-1}\mu\right),\right.\\ &\qquad\left.\left(\frac{1}{V_j^*}D_j^TD_j+\Sigma^{-1}\right)^{-1}\right) \qquad (39)\\ V_j^*|\alpha_j^* &\sim \mathrm{IG}\left(\frac{n_j}{2}+\nu_0,\frac{1}{2}\sum_{s_i=j}(y_i-\alpha_j^{*T}d_i)^2+\nu_1\right)\end{aligned}$$

with D_j the design matrix for the observations in cluster j, that is, the rows of D_j are d_i for each observation i in cluster j. (μ,Σ) have a joint distribution $\mathrm{N}_q\mathrm{IW}_{\eta_0}(\mu_0,\Lambda_0/\kappa_0;\eta_0,\Lambda_0)$, with p.d.f.

$$p(\mu,\Sigma)\propto|\Sigma|^{-((\eta_0+q)/2+1)}\exp\left\{-\frac{1}{2}\mathrm{tr}(\Lambda_0\Sigma^{-1})-\frac{\kappa_0}{2}(\mu-\mu_0)^T\Sigma^{-1}(\mu-\mu_0)\right\}$$

which leads to

$$p(\mu,\Sigma|\alpha_1^*,\ldots,\alpha_k^*)\sim\mathrm{N}_q\mathrm{IW}_{\eta_n}(\mu_n,\Lambda_n/\kappa_n;\eta_n,\Lambda_n) \qquad (40)$$

with

$$\begin{aligned}\mu_n &= \frac{\kappa_0}{\kappa_0+k}\mu_0+\frac{\kappa}{\kappa_0+k}\bar{\alpha^*}\\ \kappa_n &= \kappa_0+k\\ \eta_n &= \eta_0+k\\ \Lambda_n &= \Lambda_0+V^*+\frac{\kappa_0 k}{\kappa_0+k}(\bar{\alpha^*}-\mu_0)(\bar{\alpha^*}-\mu_0)^T\\ V^* &= \sum_{j=1}^k(\alpha_j^*-\bar{\alpha^*})(\alpha_j^*-\bar{\alpha^*})^T\\ \bar{\alpha^*} &= (1/k)\sum_{j=1}^k\alpha_j^*\end{aligned}$$

To sample from this distribution, we proceed as follows:

- First draw $z_1,\ldots,z_{\eta_n}$ i.i.d. $\mathrm{N}_q(0,\Lambda_n)$. Then

$$\Sigma=\sum_{i=1}^{\eta_n}z_iz_i^T\sim\mathrm{IW}_{\eta_n}(\Lambda_n).$$

- Draw $\mu|\Sigma,\alpha^*\sim\mathrm{N}_q(\mu_n,\Sigma/\kappa_n)$

4. Results

The posterior predictive distribution for the regression parameters in the four models can be seen in figures (2) to (5). The posterior predictions from all models can be compared with the observed data in figure (6). For Model 1, it can be seen how the non-parametric part of the model attempts to compensate a poor specification. The regression coefficients α_0 and α_1 show trimodal posterior predictive distribution (Figure 2), which in turn is reflected in the clustered posterior predictions in Figure 6. The principal explanation of the clustering comes from the absense of x in the specification of Model 1. The observations tend to group at different levels of x. Visually, we can distinguish three levels for Group 0, and two for Group 1, in concordance with the observed clustering in the posterior predictive distribution of the parameters. Note that the partitions affect all regression parameters α and the variance V jointly. In Model 2, where only intercept and x are included in the regression, the clustering mechanism clearly identifies two groups in exact concordance with the real group memberships. The difference in the variance is also captured, as can be seen in the posterior predictive distribution for V (Figure 3). Parameters in Model 3 show essentially one mode for Intercept and variance, and a bimodal distribution for Group and x effects (Figure 3). The posterior predictive draws (Figure 6) show how the unspecified interaction between Group and x is captured. However, as clusters are defined in terms of all regression parameters *and* the variance, the model fails to detect that each group defined by g has a different slope. Instead, for each group the model requires fitting a pair of slope and intercept. Model 4 is close to the *real* specification for the parameters, although it does not take into explicit account the different variances in the groups. Although for α_0 and α_1 some bimodality can still be appreciated, the estimations in general show a tendency to represent every observation by the same model specification, and only a single mixture component is required. Also, the effect of the relatively vague prior specification can be seen with more clarity in models 3 and 4, where predictions can be seen as conservative in terms of the differences between the groups.

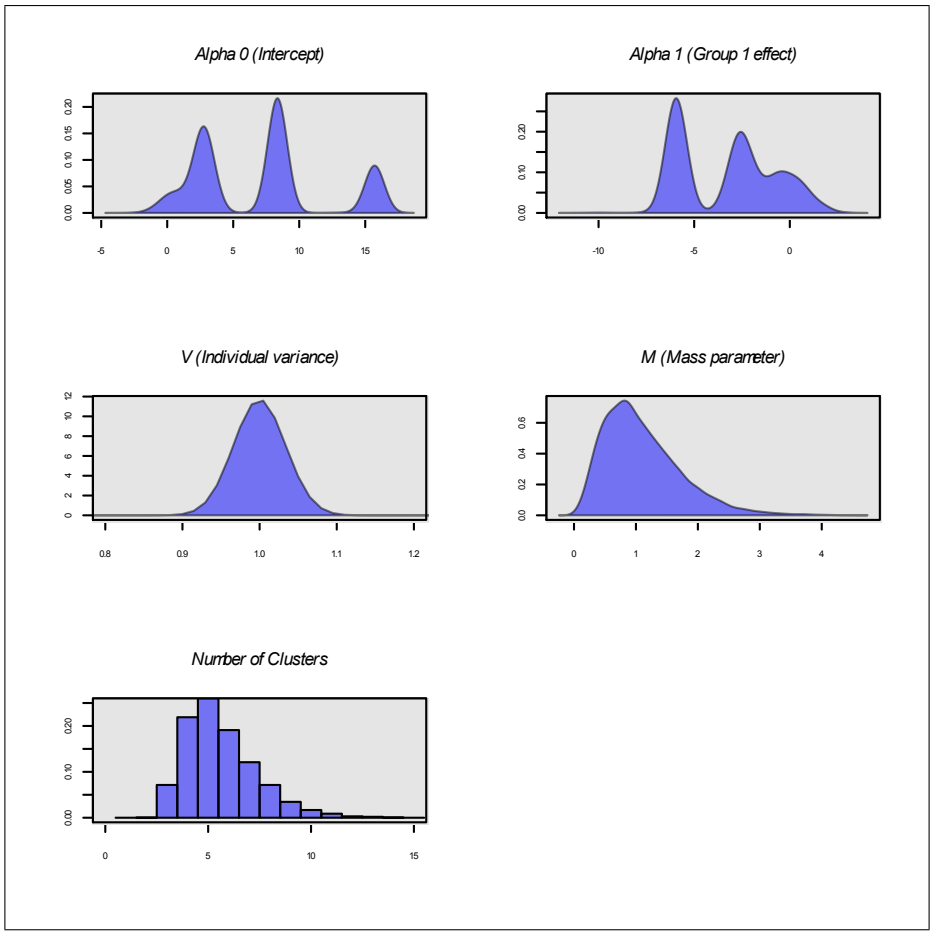

FIGURE 2. Posterior predictive distribution of selected parameters for Model 1.

5. Similarity analysis

Of course, looking at the plot of the data (see figure 6), models 1, 2, and even 3 seem inadequate. They are presented here to illustrate how the clustering process allows SSMM models to compensate a specification which does not fit the data in the best possible way by means of proposing different parameters to distinct groups of individuals. Certainly, the most poorly specified one is model 1. The clustering process correctly captures the lack of a continuous covariate. The intrinsic similarity, based on the SM sorted by group and covariate x, shows how observations are clustered at different levels of x, and also in different groups in concordance with covariate g. This follows from observing the formation of various adjacent clusters

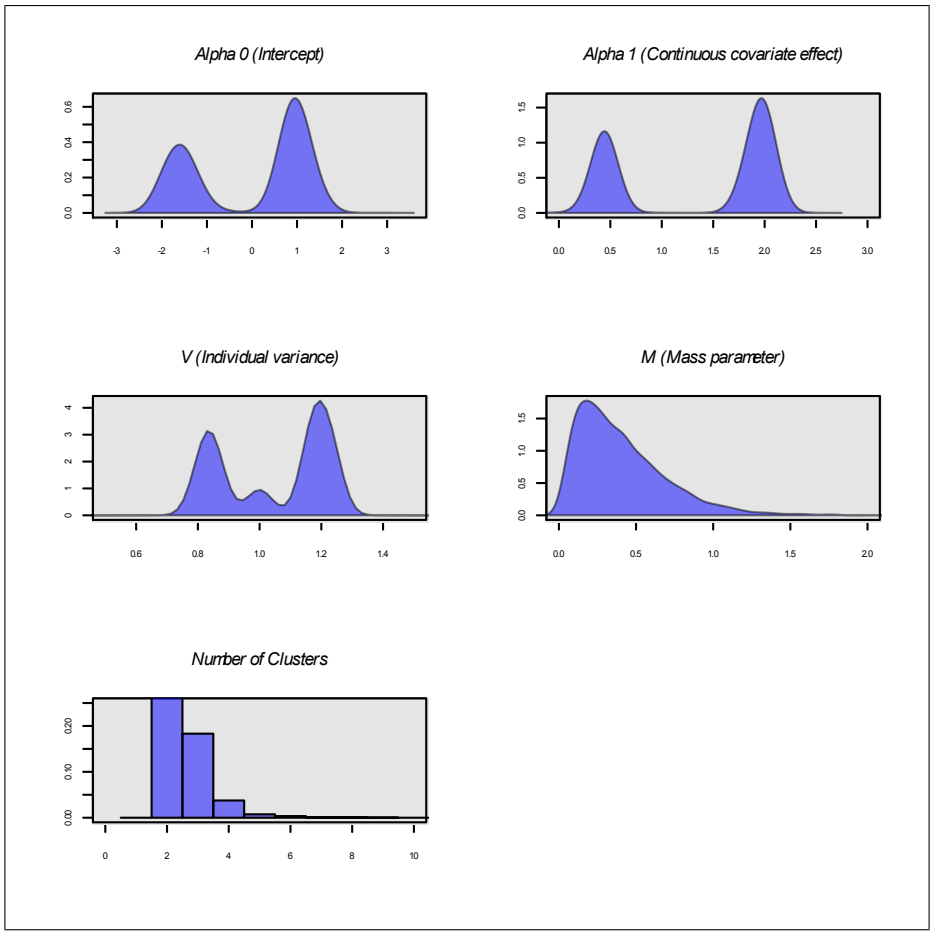

FIGURE 3. Posterior predictive distribution of selected parameters for Model 2.

of similar size along the range of x (see figures 8- 11). It can also be seen that the most autonomous observations are actually the first representatives of their clusters, when the proposed order is considered. Model 2 has other mispecification. Here, the only covariate considered is the continuous one, x. The SM clearly identifies two clusters, which correspond exactly to the simulated groups. Model 3 includes both g and x and constitutes a more reasonable specification, although we know there is an interaction term missing. The SM still shows a partitioning in two groups, although more diffuse, and that is a sign of a possible interaction that has not been taken into account. When the SM is sorted only by x, the SDG shows that, although there is a tendency to form clusters at different values of the

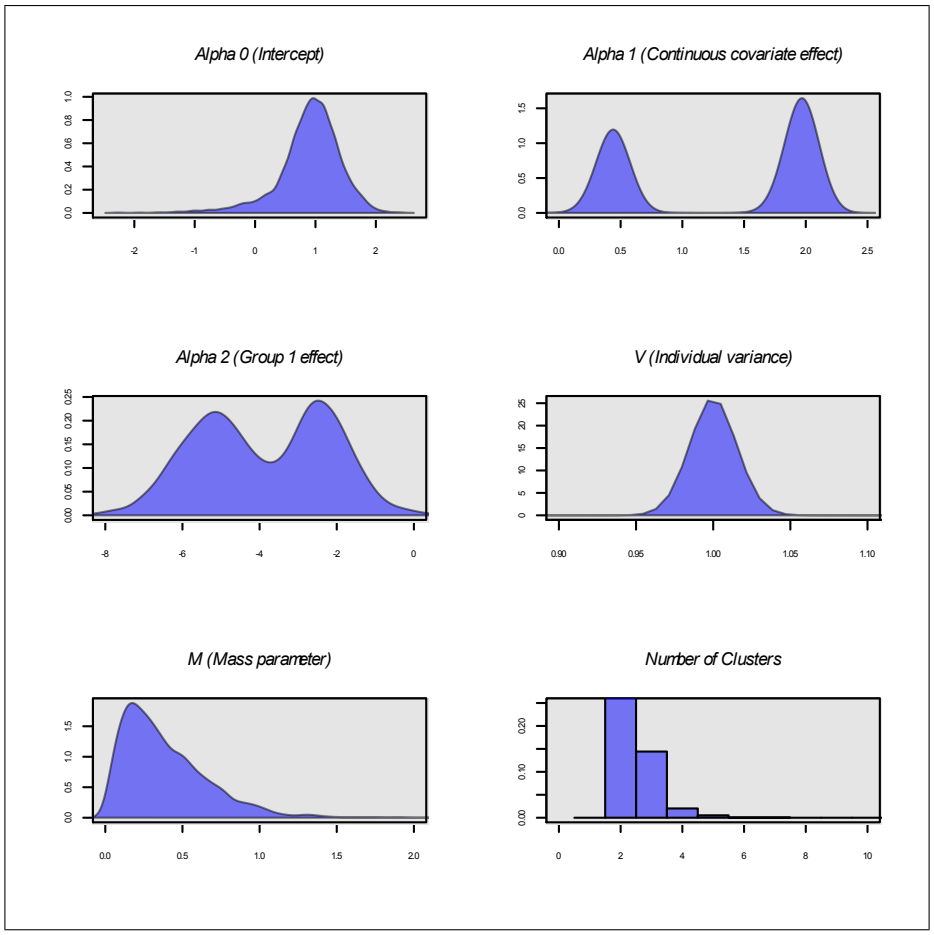

FIGURE 4. Posterior predictive distribution of selected parameters for Model 3.

covariate, it is not fully explained by it. When the SM is additionally sorted by the group effect, the partitions become quite clear. Model 4 adds the mentioned interaction effect. The SM clearly reflects the maximal partition, so, as far as clustering is concerned, the model does not capture any substantial difference in the behavior of the observations based on the statistical model, and one common set of parameters is set for all. There is still a difference between both groups, which consists on the different variances, and it is not captured. This may be explained by the high tolerance to different variances determined by the hyperpriors $\nu_0 = 1$ and $\nu_1 = 1$. In Figure 7, the clustering behavior of the observations in terms of

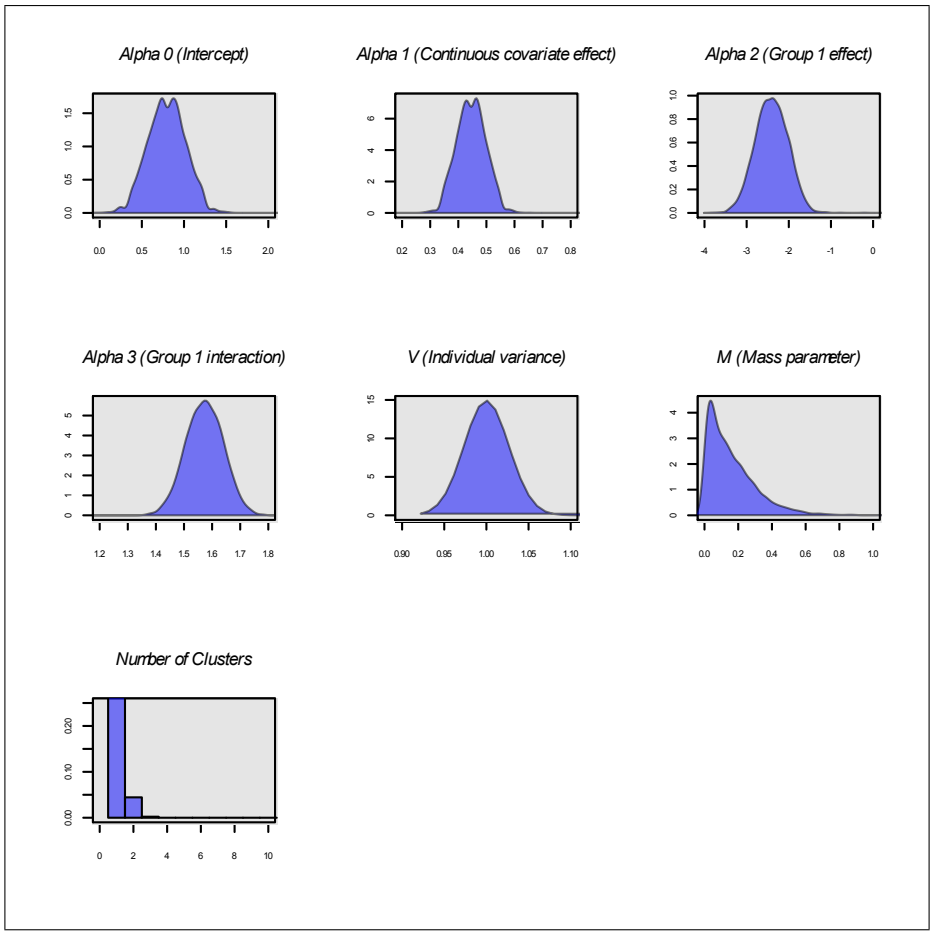

FIGURE 5. Posterior predictive distribution of selected parameters for Model 4.

the covariates can be appreciated clearly for all models. The lines connect observations that belong to the same cluster. The degree of association is related to the intensity of the lines. In fact, pairs of observations that do not belong to the same cluster are joined by transparent lines. The color of the lines is determined by the position of the FCR in the sample, based on sorting by Group and x. Observations are represented by circles, and the radious of the circles is proportional to the autonomy of each observation. FCRs are represented by the biggest circles. In Model 1, the clustering process determines, roughly, three levels for x in Group 0. Group 1 is more homogeneous. In turn, low levels of x determine more variability in the FCRs than high levels. In Model 2, two clusters, represented by their FCRs,

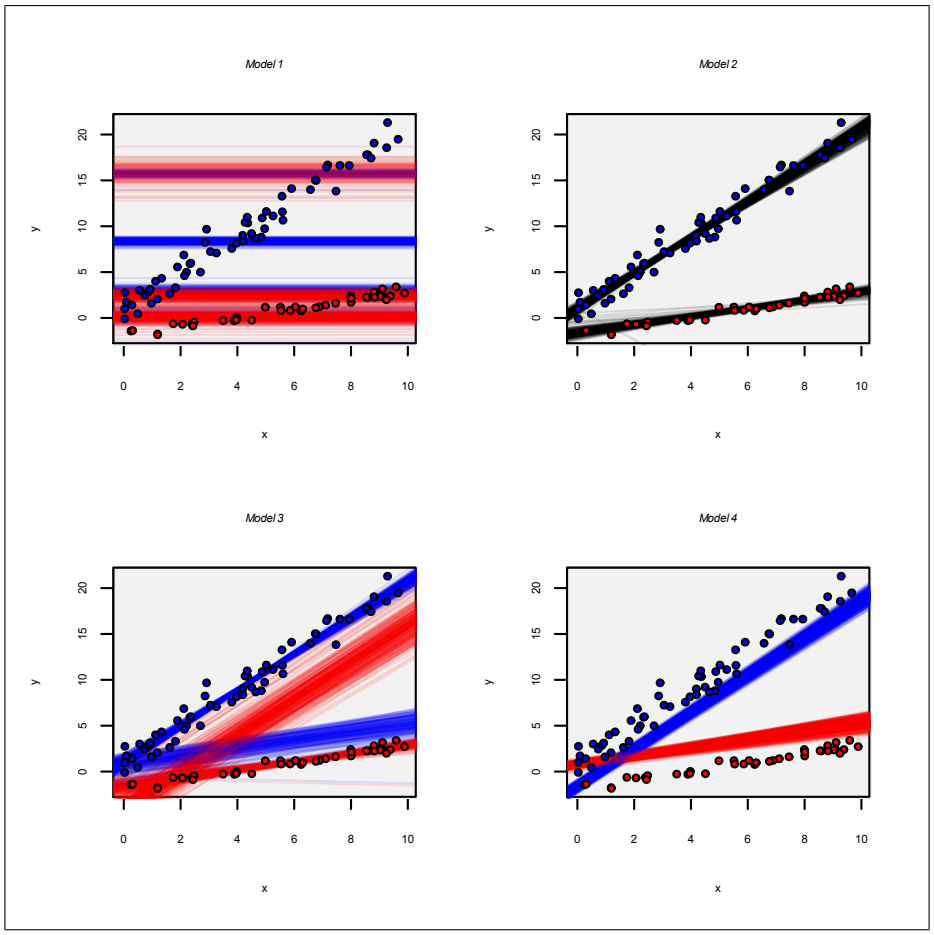

FIGURE 6. Sampled posterior predictive values for simulated data. Group 0 is marked in blue, and Group 1 in red.

can be clearly identified. In Model 3, clusters tend to combine at low levels of x, representing some sort of confussion in the clustering mechanism. Model 4 clearly groups all observations in one cluster, represented by the first oservation.

The power of the nonparametric part of SSMMs to detect differences on how the observations respond to the model specification depends on several factors. It is advisable that the SSM, which rules the nonparametric part, should be flexible enough to allow detecting differences when there actually are. For that purpose, it is better that the variance specified in the topmost part of the hierarchy should be relatively tight. That allows for flexible cluster location placement, as driven by the data. Otherwise, it is possible to end up with all observations assigned to a

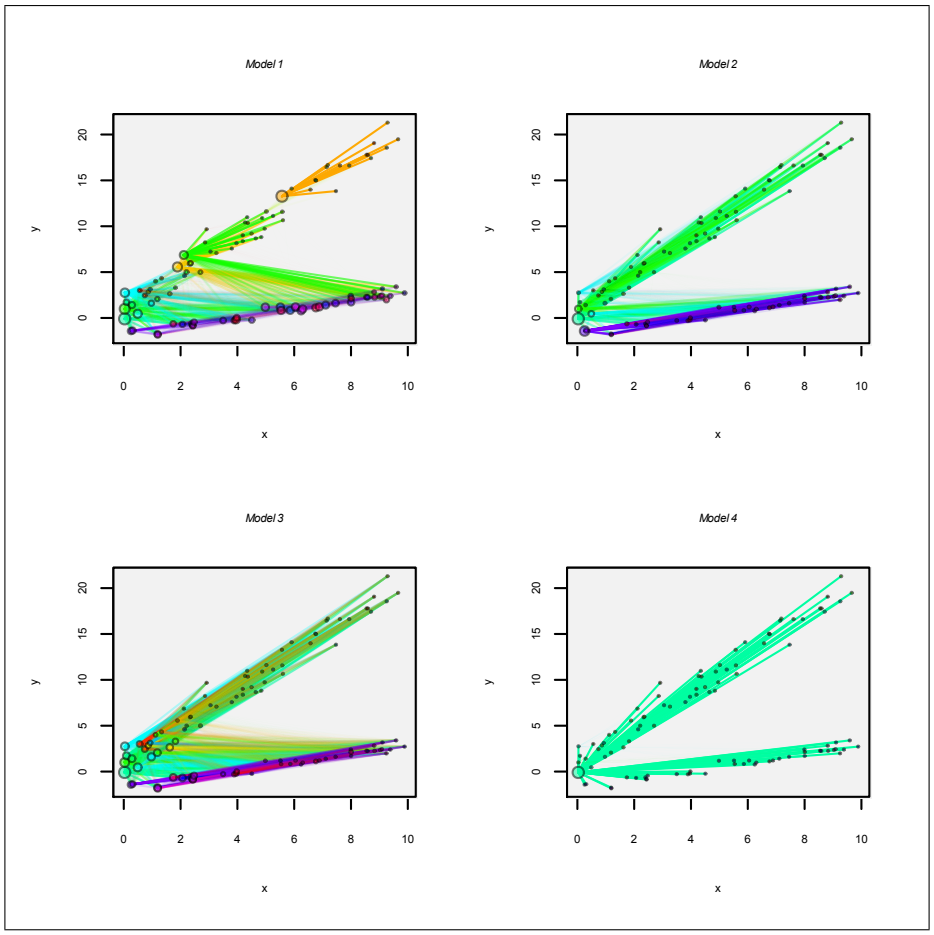

FIGURE 7. Clustering plot based on intrinsic similarity matrix.

single cluster. Other very important aspect is the indirect specification of the prior number of clusters. In the DP, this is adjusted in the mass parameter M, and the mean number of clusters is, a priori, close to $M \log(n)$ (Liu, 1996). As usual in Bayesian statistics, all prior specifications need to be as honest as possible, but it is recommended to give positive prior probability mass to a broad range of number of clusters. For instance, the PY process with $\alpha < 0$ upper bounds the number of clusters with probability one, and this is not recommended in general, unless the researcher is highly convinced on putting all the prior probability on small values for the number of clusters.

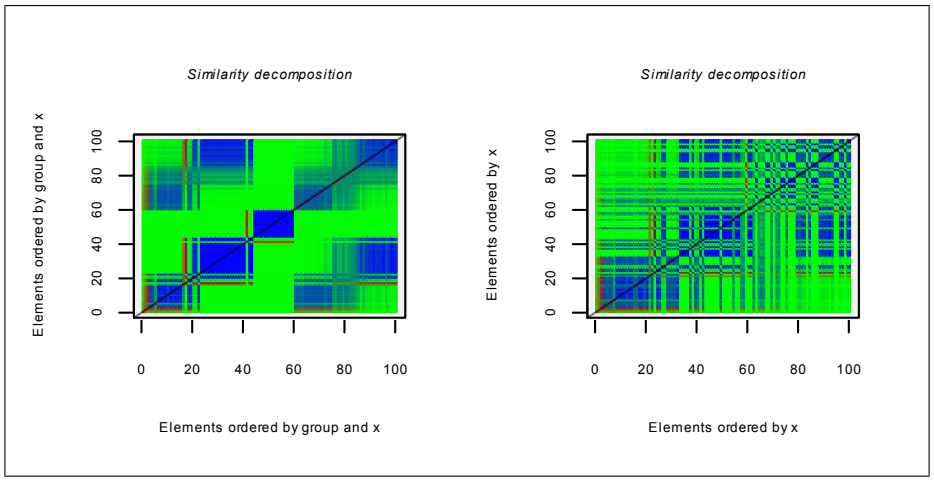

FIGURE 8. Similarity decomposition graph for Model 1.

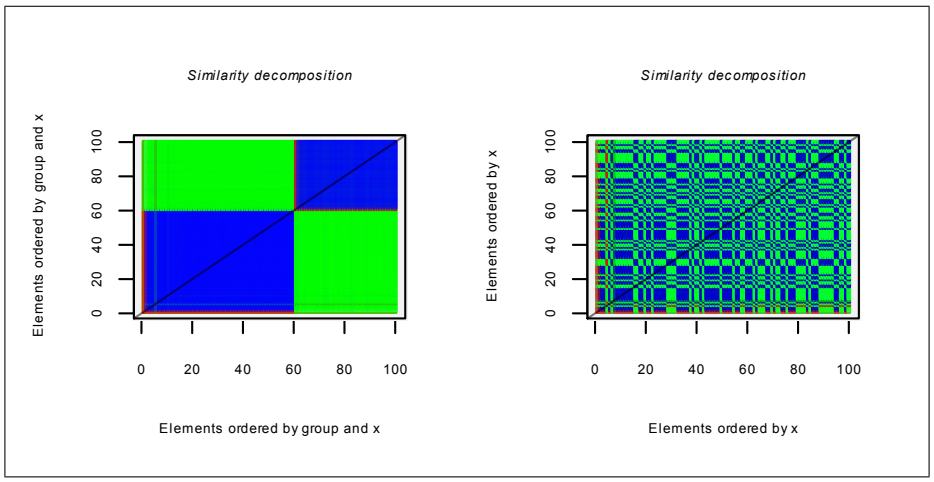

FIGURE 9. Similarity decomposition graph for Model 2.

6. Example: Forbes' data

James Forbes (1857) studied the relationship between atmospheric pressure and the boiling point of water. He was interested in estimating altitude, as he knew that altitude could be determined from atmospheric pressure, with lower pressures corresponding to higher altitudes. He collected data in Scotland and in the Alps, measuring at each of 17 locations pressure in inches of mercury with a barometer and boiling point in degrees Fahrenheit using a thermometer. The data seem to

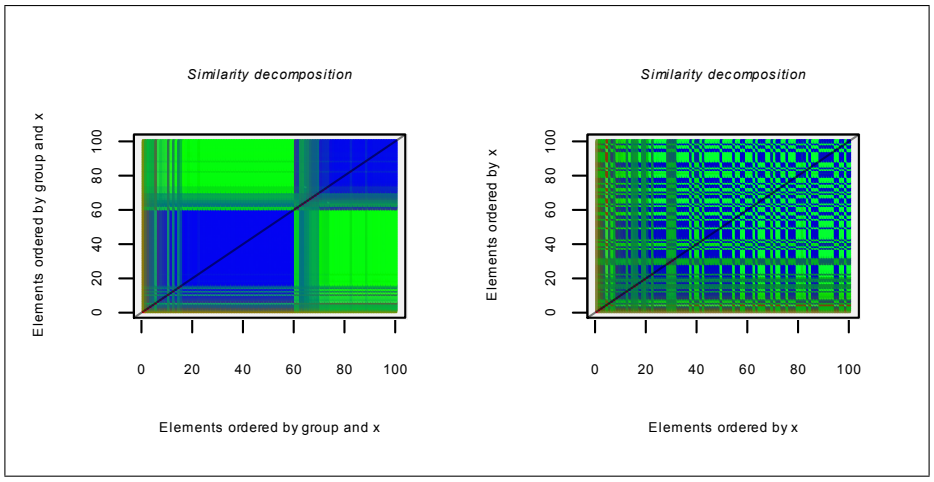

FIGURE 10. Similarity decomposition graph for Model 3.

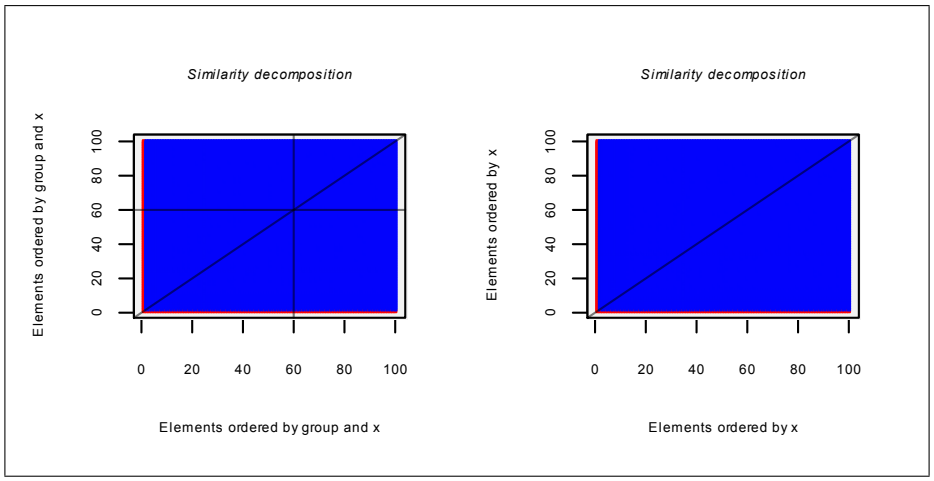

FIGURE 11. Similarity decomposition graph for Model 4.

follow a well defined linear trend (Figure 12). Two observations, however, do not seem to fit in the model like the rest do. Sorting by Temperature, these observations correspond to numbers 12 and 17. These observations present the highest errors, from standard linear regression (Figure 12). Actually, a logarithmic transfomation of Pressure is known to normalize the distribution of the residuals. But these two observations keep certain distance from the rest, and can be considered as *outliers*. The inherent flexibility of our model allows for atypical observations to potentially be allocated to a different cluster, thus avoiding the classical usage of stabilizing

transformations. We are going to fit Model (36) to Forbes' data to see the result of Similarity Analysis applied to this case. Based on standard linear regression, we set $\mu = (-83, 0.5)^T$. The residual variance from linear regression is 0.08. We set $\nu_0 = 100$ and $\nu_1 = 2$ to define a prior expectation of 0.02 in the baseline measure for individual variance parameters V_i. Prior distribution for M is determined by $a = 10$ and $b = 2$, that is, an expectation of 5 in the Gamma prior. The rest of the parameters is specified by $\eta_0 = 4$, $\kappa_0 = 1$, and Σ is set to the Identity matrix.

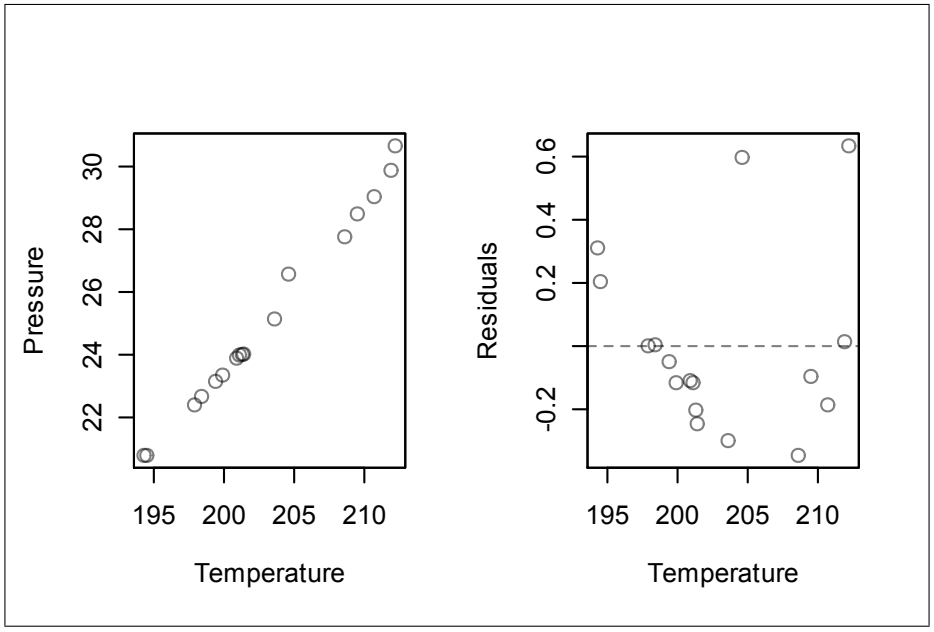

FIGURE 12. Forbes' data and residuals from standard linear regression.

6.1. Results. The posterior distribution for individual parameters can be seen in Figure 15. We can appreciate that observations 1, 12 and 17 show a lower intercept and higher slope, compared with the rest of the observations. Observation 2 also differentiates from the rest, to a lesser extent. Individual variances are set equal, from visual appreciation. The SDG (Figure 13) is consistent with these conclusions. Observations 12 and 17 appear as classified together in one cluster, apart from the rest, although they also appear related with observation 1. Additional information is obtained in relation to observations 1, 2 and 3. Observation 1 looks like an outlier, separate from the second cluster, which in turn is represented

partially by observations 2 and 3. In Figure 15, we can see that the distribution of regression parameters corresponding to observation 2 tend to overlap with the posterior distribution for observations 3 to 11.

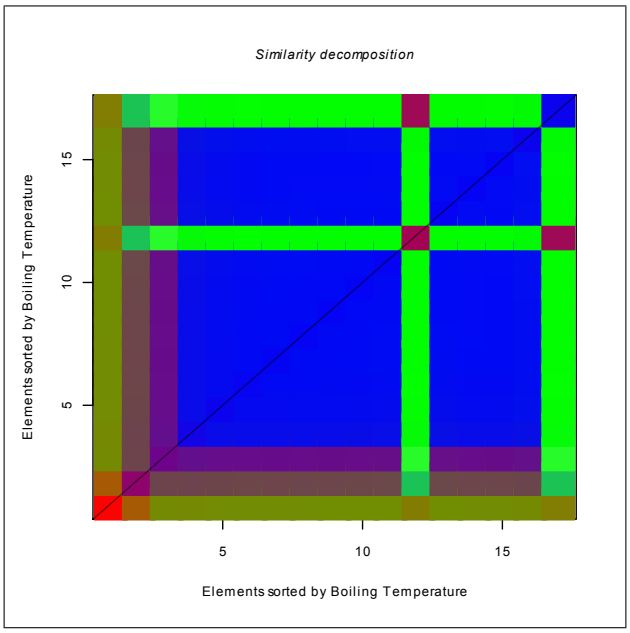

FIGURE 13. Similarity Decomposition Graph for Forbes' data.

Information obtained from the SDG can be complemented by looking at the clustering plot in Figure 14. It can be seen there that the general tendency is represented by observations 1 and 2, and observations corresponding to central temperatures are also represented by observation 3. Observations 12 and 17 show a clear tendency to be classified in a cluster represented by observation 12, although the latter can be also represented by observation 1, since they are connected. In fact, observation 1 seems to relate with every observation up to some extent, and it is difficult to decide where to classify it from visual inspection. The diameter of the circles is proportional to the autonomy of each element, and the intensity of the connecting lines is proportional to the intrinsic similarity. Additionally, the graph includes posterior predictive mean response, based on posterior predictive means for the regression parameters, and 95% credibility bands, based on posterior predictive means for the individual variances. Observations 12 and 17 fall out of the bands, confirming their outlier condition. Three observations (10, 11 and 13)

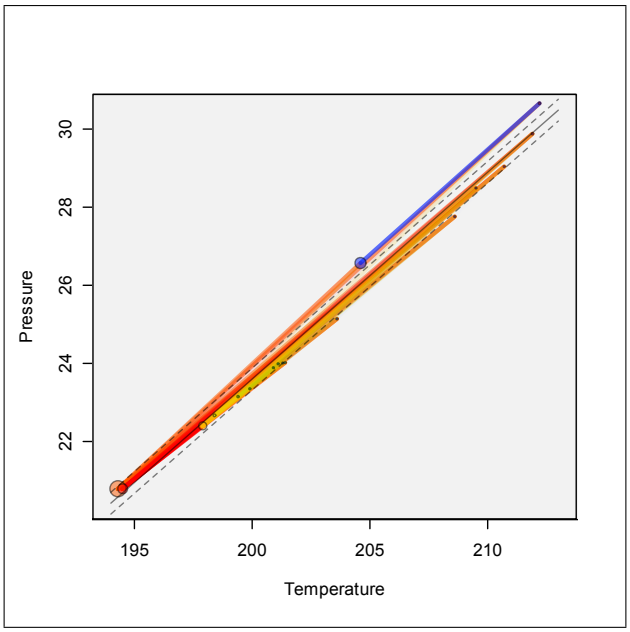

FIGURE 14. Clustering plot for Forbes' data.

fall slightly below the lower confidence band, but the tendency is to classify them in a cluster lead by observation 3, like other observations in the central part of the graph. The sampled partition with minimum expected loss has $EL_{1:1} = 29.05$, and is represented as

$$\pi_1 = (1, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 1, 2, 2, 2, 2, 1),$$

where the positions correspond to observations (in order determined by Temperature) and numbers correspond to cluster labels. Putting observations 10, 11 and 13 in a new cluster results in the following partition:

$$\pi_2 = (1, 2, 2, 2, 2, 2, 2, 2, 2, 3, 3, 1, 3, 2, 2, 2, 1).$$

It has expected loss $EL_{1:1} = 87.91$. To decide about the membership of observation 1, the expected loss for the partition

$$\pi_3 = (1, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 3, 2, 2, 2, 2, 3)$$

is 29.23. So, the optimal decision is to represent the 17 observations by partition π_1. It can be concluded then that the individuals that behave (slightly) different from the rest are 1, 12 and 17.

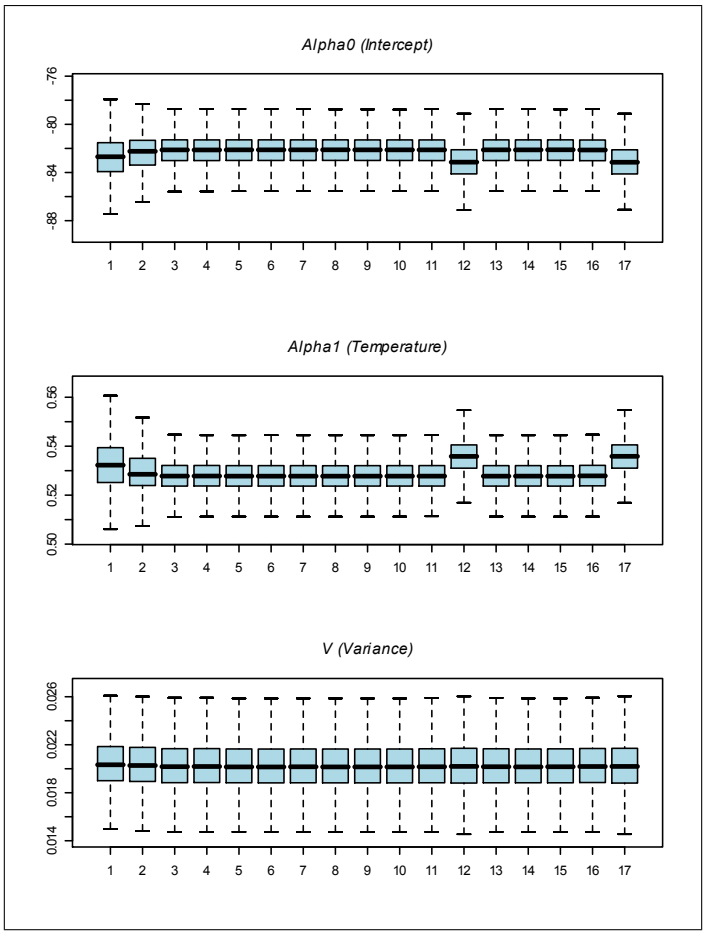

FIGURE 15. Posterior invididual parameters for Forbes' data.

CHAPTER 5

Application: Multivariate Binary Model

The application shown here considers modelling a bivariate binary response, representing the occurrence of an event in two periods, and its relation to covariates measured at the beginning of a study. The response is the occurrence of atrial fibrillation (AF) in 102 patients evaluated at 30 days and 1 year of follow-up. The multivariate binary response is modelled based on latent covariates which are supposed to come from a multivariate Normal distribution, following Albert and Chib (1993). This distribution, in turn, considers individual regression parameters coming from a Dirichlet Process, together with individual variance. Similarity Analysis is based on the same principles as before, and this application gives us the opportunity to observe clustering behavior from a new modelling perspective.

1. Atrial Fibrilation data

We consider data coming from 102 patients that were admitted to the Catholic University Hospital for non-valvular AF between January 2000 and August 2002, prospectively recruited. The data are used here with permission from the researchers (Acevedo et al 2006). Conclusions derived from the application of the statistical models presented here are intended to be subject of further research with participation of physicians and experts in the scientific areas involved.

AF is the most common sustained arrhythmia in clinical practice and it is associated with increased risk of morbidity and mortality. Systemic and/or local inflammation could be involved in the process of thrombogenesis and contribute to the perpetuation of the arrhythmia. There is interest in evaluating wether the presence of inflammation could contribute to predict the cardiac rhythm during the long-term follow-up. The existence of a systemic inflammatory state is characterized by the elevation in C-reactive protein (CRP) plasma levels. Evidence for the presence of inflammation during AF has been suggested by the findings of activation

of the complement system and release of proinflammatory cytokines after cardiac surgery and by the demonstration of inflammation in left atrial biopsies taken during surgery (Bruins et al 1997, Frustaci et al 1997). Chung et al. (2001) and Dernellis et al. (2001) have demonstrated the existence of a systemic inflammatory state, characterized by the elevation in C-reactive protein (CRP) plasma levels, in patients with atrial arrhythmias and non-valvular AF, respectively. Other groups have published about an elevation in other inflammatory markers in the same patients (such as ICAM and IL-6) (Roldan et al 2003). The objective in the present study is to predict the occurrence of AF at 30 days and 1 year follow-up, based on individual covariates measured at baseline. Variables considered and their possible relation with the outcome is herewith explained:

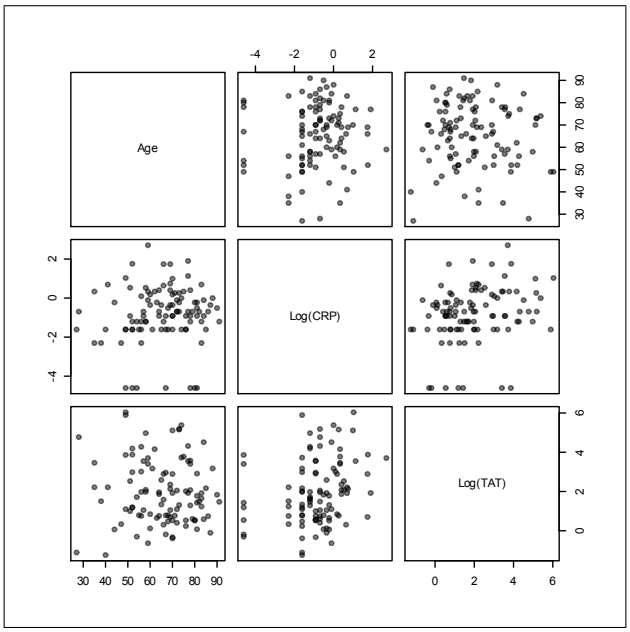

FIGURE 1. Scatterplots of association between covariates.

- Age. It has been suggested that the probability of having recurrence of Atrial fibrillation is incresed in patientes older than 70. In these patients there is a higher prevalence of AF, probably due to the fact that these patients have a greater prevalence of hypertensive cardiomyopathy with diastolic dysfunction and dilated left atrium. In our data, the range of Age is between 27 and 91 years old.

- HBP High blood pressure (HBP) causes dilation of the left atrium, which is associated to systolic dysfunction and AF, since there is an elevated diastolic pressure in the left ventricle which predisposes the left atrium to get enlarged.
- LVD Left ventricular dysfunction (LVD) is a direct cause of AF.
- AA Anticoagulants/Antiarrhythmic (AA) drugs (e.g. aspirin) act as antiinflammatory drugs. In AF, there exists an inflammatory process in the atrium with inflammation and increased hemostatic activation, increasing the possibility to form thrombus inside the atrium. Therefore, it has been seen that when a patient takes aspirin there is less chance for the recurrence of AF.
- Paroxysmal/Persistent state. There is more chance for the recurrence of AF in patients who have chronic AF than for those who had a paroxysmal AF. Persistent and chronic patients have larger left atrium, which predisposes the arrhythmia to stay forever.
- CRP CRP is elevated in patients with chronic and paroxysmal AF and it characterizes the existence of a systemic inflammatory state. CRP determinations were performed with the inmunoturbidimetric method (sensitivity < 0.03 mg/dl). It has been suggested that small increments of CRP, within normal levels, may be useful to detect inflammation.
- TAT Thrombin-antithrombin complex (TAT) plasma levels are associated with the presence of thrombus inside the atrium.

TABLE 1. Basic summary of Atrial Fibrilation data

Variable	Total n=103			Paroxysmal group n=63		Persistent group n=39	
	Mean	St. Dev.	Range	Mean	St. Dev.	Mean	St. Dev.
Age	65.6	14.0	27 - 91	64.8	15.3	66.9	11.6
log(CRP)	-0.81	1.43	(-4.6) - 2.7	-1.01	1.43	-0.49	1.39
log(TAT)	1.98	1.63	(-1.2) - 6.0	1.88	1.61	2.13	1.67
HBP (%)	49.0			50.8		46.2	
LVD (%)	12.7			9.5		17.9	
AA (%)	40.2			41.3		38.5	
30 days FA (%)	40.2			9.5		89.7	
1 year FA (%)	52.0			31.7		84.6	

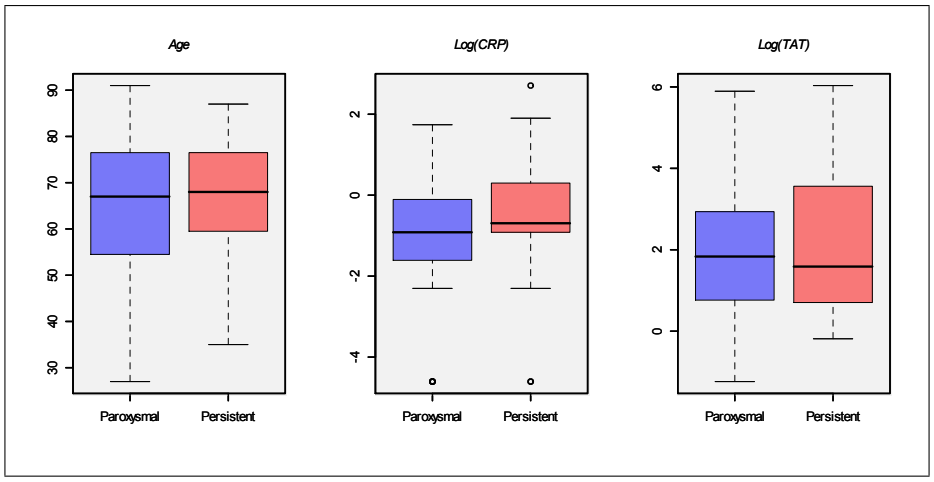

FIGURE 2. Association between Paroxysmal/Permanent groups and covariates.

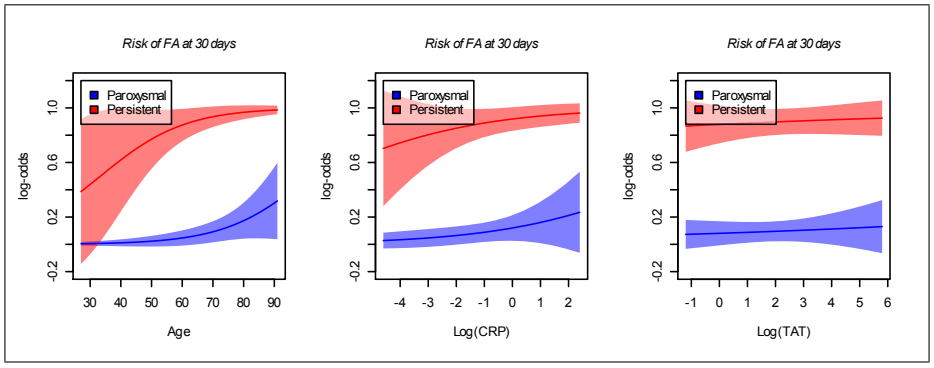

FIGURE 3. Nonparametric estimates of risk at 30 days follow-up.

CRP and TAT levels are entered in logarithmic form in model specifications to standardize its left-skewed distribution. HBP, LVD, AA and Persistent staus are represented with 1 if the risk factor is present, 0 if absent. A general description of the data and the association of the covariates is presented in table 1 and figures 1 and 2. Figures 3 and 4 show nonparametric estimates of risk at 30 days and 1 year follow-up, based on penalized regression splines fit by the *gam* function in *mgcv* package for statistical software R.

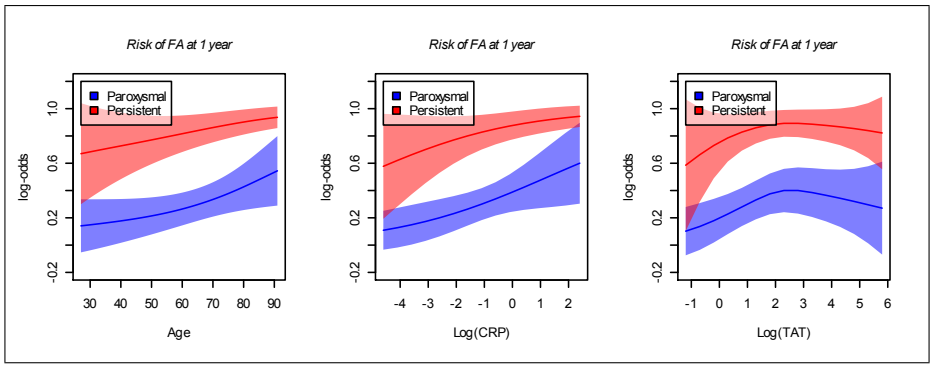

FIGURE 4. Nonparametric estimates of risk at 1 year follow-up.

2. Statistical model

Suppose we have a random sample $y_1, \ldots, y_n$ of m-dimensional binary responses, which in turn may be explained by a set of $q-1$ real-valued baseline covariates $x_1, \ldots, x_{q-1}$. We propose the model

$$\begin{aligned} y_i &= (I(z_{i1} \geq 0), I(z_{i2} \geq 0), \cdots, I(z_{im} \geq 0))^T \;\; i = 1, \ldots, n \\ z_i|(\alpha_i, \Lambda_i), d_i &\sim \mathrm{N}_m\left((\alpha_{i1}^T d_i \cdots \alpha_{im}^T d_i)^T, \Lambda_i\right) \;\; i = 1, \ldots, n \\ (\alpha_i, \Lambda_i)|G &\sim G. \end{aligned} \tag{41}$$

with

$$\begin{aligned} \alpha_i &= (\alpha_{i1}^T \; \alpha_{i2}^T \; \cdots \; \alpha_{im}^T)^T \; i = 1, \ldots, n \\ \alpha_{ij} &= (\alpha_{ij0} \; \alpha_{ij1} \; \cdots \; \alpha_{ij(q-1)})^T \; j = 1, \ldots, m \\ d_i &= (1 \; x_{1i} \; x_{2i} \; \cdots \; x_{(q-1)i})^T \end{aligned} \tag{42}$$

In (41), $y_i, i = 1, \ldots, n$ are the observed multivariate binary responses, and they are defined on the basis of latent vectors z_i, whose distribution is supposed to be multivariate Normal, following Albert and Chib (1993). The prior mean of z_i is defined as a linear combination of individual covariates grouped in the vector d_i, associated with an unknown vector of regression coefficients α_i and covariance

matrix Λ_i. In turn, G comes from a RPM defined as

$$\begin{aligned} G &\sim \mathrm{DP}(M, G_0) \\ G_0(\alpha_i) &\equiv \mathrm{N}_{mq}(\mu, \Sigma) \\ G_0(\Lambda_i) &\equiv \mathrm{IW}_{\eta}(\Lambda_0) \\ (\mu, \Sigma) &\sim \mathrm{N}_{mq}\mathrm{IW}_{\eta_0}(\mu_1, \Sigma/\kappa_1; \eta_0, \Sigma_0) \\ M &\sim \Gamma(a, b) \end{aligned} \tag{43}$$

Note that the covariates in d_i are the same for every component from 1 to m in the multivariate response vector, and we are considering q coefficients for each individual component of the response, including the intercept.

Our interest focuses on the estimation of individual relative risks (RR) of presenting an episode of AF at 30 days or 1 year, based on individual covariates. We define a set of reference values for the covariates x_0. These values can be, for instance, theoretical indicators of low risk, or mean values for a certain group of reference individuals. The RR for a patient i at stage j ($j = 1$ for 30 days, $j = 2$ for 1 year) compared with the reference values is calculated as

$$RR_{ij} = \frac{p(z_{ij} \geq 0 | x_i, \theta_i)}{p(z_{ij} \geq 0 | x_0, \theta_i)}. \tag{44}$$

The posterior predictive relative risk for an hypothetical patient with covariate values x_h can also be calculated, based on the posterior predictive distribution $p(z_{n+1}|y_1, \ldots, y_n)$. It is estimated by

$$RR^*_j(x_h) = \frac{r_j(x_h)}{r_j(x_0)}. \tag{45}$$

To calculate the predictive risk $r_j(x_h)$, we sample posterior predictive values (θ^*, Λ^*) for the regression coefficients and individual variance from (29). This is then evaluated in (41) with the covariate value of interest x_h.

The model presented here follows, in principle, an approach similar to the mixture models presented in the works of Kottas, Müller and Quintana (2005) for multivariate ordinal data, and Jara, García-Zattera and Lesaffre (2007) for multivariate binary outcomes. The difference is made in the extreme flexibility in the specification of the variances. The covariance matrix for the latent vectors z_i is not constrained at all and, moreover, the base measure specified for the Dirichlet

process that defines the random distribution for the multivariate regression parameters considers, additionally, a prior Normal - Inverse Wishart distribution for even more flexibility. It is important to note that the variance specified in (41) leads to parameter estimations that are not likelihood identified (Chib and Greenberg, 1998). This identification problem affects the estimations in terms of scale, and it is solved specifying the covariance in terms of correlations, or specifying constraints as discussed in Jara et al (2007). We do not consider such restrictions here, and the possible implications of this problem need to be further investigated like, for instance, how this affects the MCMC sampling. RRs defined in (44) and (45), on the other hand, do not depend directly on the non-identifiable scale of the parameters, but the specific influence of non-identifiability on these estimations needs to be investigated, too. Other important consequence of the chosen parameterization is the dimension of the parameter space. Sampling from all parameters involved is a computationally intensive task. In the context of this application, the dimension of the multivariate binary response vector is $m = 2$. With 102 observations and considering 3 covariates plus intercept, the dimension of the regression parameters for each level of the multivariate response is $q = 4$. The values to estimate in each individual covariance matrix is 3, plus 2 values for the latent z. That gives a total of 13 parameters per individual. In addition, the posterior distribution of the baseline parameters μ and Σ is also sampled. In our case, μ is an 8-dimensional vector, making necessary 36 parameters to estimate in Σ. Posterior predictive values for α and Λ are also sampled. In summary, including also the posterior sampling of the mass parameter M, 1384 parameters are obtained at each iteration. In return of this cost, we obtain a model capable to adapt itself to a broad range of possible relations between response and covariates, plus the additional consideration to partition structures in the data. Information about the data that is possible to obtain include detection of outliers, adaptation of the model to subgroups of data, modelling the relation between response and covariates specifically for each dimension in the response, and more.

2.1. Posterior computations. Every component of the binary observations y_i constrains, *a posteriori*, the corresponding component of the latent vectors z_i to be positive or negative with probability one. Conditional on the data and the

rest of the parameters, the posterior distribution of $z_1, \ldots, z_n$ is then truncated multivariate normal of the form

$$\begin{aligned}
(z_i|y_i, \alpha_i, D_i, \Lambda_i, B_i) &\sim TN_m(D_i\alpha_i, \Lambda_i, B_i) \\
D_i &= \begin{pmatrix} d_i^T & 0 & \cdots & \cdots & 0 \\ 0 & d_i^T & 0 & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & \cdots & \cdots & \cdots & d_i^T \end{pmatrix} \\
B_i &= B_{i1} \times B_{i2} \times \cdots \times B_{im} \\
B_{ij} &= \begin{cases} \mathbb{R}^+ & \text{if } y_{ij} = 1 \\ \mathbb{R}^- & \text{if } y_{ij} = 0 \end{cases}
\end{aligned} \tag{46}$$

The posterior distribution of z_i is truncated to the m-dimensional product of positive or negative real intervals, depending on the observed value y_i. Simulations for the truncated multivariate normal distribution in (46) were done following Geweke (1991).

The partition for the parameters coming from the DP is updated as follows. Posterior predictive rule results in putting observation i in cluster j with probability proportional to

$$n_j(2\pi)^{-mn_j/2}|\Lambda_j^*|^{-n_j/2} \exp\left[-\frac{1}{2}\sum_{c_i=j}(z_i - D_i\alpha_j^*)^T\Lambda_j^{*-1}(z_i - D_i\alpha_j^*)\right] \tag{47}$$

where α_j^* and Λ_j^* are the vector of parameters and covariance matrix associated with cluster j, that is, each cluster location is $(mq + mq(mq+1)/2)$-dimensional. The probability of generating a new cluster for observation i is proportional to

$$M\int (2\pi)^{-m/2}|\Lambda|^{-1/2} \exp\left[-\frac{1}{2}(z_i - D_i\alpha)^T\Lambda^{-1}(z_i - D_i\alpha)\right] dG_0(\alpha, \Lambda) \tag{48}$$

This result requires using algorithms designed for non-conjugate specifications. See Appendix A for details. For cluster locations, we have

$$\begin{aligned} \alpha_j^*|\Lambda_j^* &\sim \mathrm{N}_{mq}(M^*, V^*) \\ \text{with } V^* &= \left(\sum_{s_i=j} D_i^T \Lambda_j^{*-1} D_i + \Sigma^{-1}\right)^{-1} \\ M^* &= V^* \left(\sum_{s_i=j} z_i^T \Lambda_j^* D_i + \mu\Sigma^{-1}\right) \end{aligned} \tag{49}$$

$$\Lambda_j^*|\alpha_j^* \sim \mathrm{IW}_{n_j+\eta_0}\left(\left(\sum_{s_i=j}(z_i - D_i\alpha_j^*)(z_i - D_i\alpha_j^*)^T + \Sigma\right)^{-1}\right) \tag{50}$$

Parameters (μ, Σ), which determine the baseline distribution in the DP for the regression parameters, are in turn updated using (40), but replacing q by mq, κ_0 by κ_1, μ_0 by μ_1 and Λ_0 by Σ.

2.2. Model specification. Of principal interest is the relationship between CRP and the response. Our basic model (Model 1) considers Age, Log(CRP), and Persistent state as baseline covariates. After knowing the conclusions from this model, we extend the specification to include the effects of HBP, LVD, AA and Log(TAT) in Model 2. Prior specification for the baseline measure G_0 was defined by a standard multivariate Normal distribution for α, and Λ_0 was defined as the Identity matrix. This prior specifications were intended to represent conservative initial beliefs on the effect of the covariates, but *informative* enough to have parameter estimations in a reasonable scale. In order to obtain posterior predictions for the parameters, the number of clusters does not have to be too high, because this would give too much importance to the baseline measure in the sampling. In initial test runs, the proposed specification gave reasonable results concerning this point, too. Hyperparameters η and η_0 were set to the dimension of the corresponding vectors, plus 2, to ensure the finiteness of its density and expectation, and κ was set to 1.

3. Results

Table 2 summarizes the posterior predictive distribution for parameters coming from the DP for Model 1. In order to obtain conclusions from these distributions,

all the available information must be considered. These results constitute an initial reference and do not show the different interactions involved. At first glance, we can see that the effect of age appears to be positive, although the negative mean at 30 days indicates fluctuations that need further exploration. At both 30 days and 1 year of follow-up, Persistent patients have clearly more risk of AF. It can also be seen that the risk of AF increases with log(CRP), specially at 1 year, although the negative mean at the latter time indicates variations as in the case of Age. The correlation between both periods is small, as it is seen in the value of λ_{12}. Detailed information about the implications of this model in terms of predictive relative risks for different combinations of covariate values can be obtained from Figure 6. From (41), the probability of having an AF episode is obtained from the quadrants of the bivariate Normal distribution defined by the individual parameters coming from the DP. The first quadrant defines the probability of having AF at both 30 days and 1 year of follow-up. The second quadrant represents having an episode at 1 year but not at 30 days. The third one, the probability of not having an episode at any time. The fourth quadrant is related with the probability of having AF at 30 days but not at 1 year. The two-dimensional relative risks in Figure 6 are based on the quadrants obtained for every sample of the parameters, and the 95% credibility ellipsoids are based on a bivariate normal approximation for the mean relative risks obtained for each MCMC sample. All relative risks are based on taking as reference patients with Age=30, CRP=0.01, TAT=0.3, Paroxysmal, not hypertensive, with no left atrial dysfunction and no aspirin. Several conclusions can be obtained from Figure 6. The effect of the Persistent status is more notorious at 1 year. At 30 days, it is slightly noticeable at high levels of CRP. The effect of CRP is clear for the risk at 1 year, but not at 30 days. This information is complemented in Figure 9, where marginal 30 days and 1 year risks are presented for different levels of CRP and mean levels of the remaining covariates. The effect of Age changes depending on CRP levels. At low levels of CRP, increments in Age result in increased risk at both 30 days and 1 year follow-ups, and the increase is higher for older patients. At intermediate and high levels of CRP, the effect of Age inverts. This is consistent with the fact that older patients in the sample present, in average, lower levels of CRP (Figure 1). 50% of patients older that 60 years old use aspirin, compared with

TABLE 2. Summary statistics for posterior predictive distribution of parameters in Model 1. Columns indicate mean, standard deviation, percentiles and probability of being positive.

	Mean	S.Dev.	P05	P25	P50	P75	P95	P(>0)
30 days								
Intercept	-0.046	0.216	-0.33	-0.109	-0.059	-0.007	0.327	0.224
Age	-0.004	0.092	-0.019	0.000	0.000	0.001	0.008	0.647
Persistent	0.058	0.198	-0.249	0.043	0.065	0.089	0.303	0.907
Log(CRP)	0.007	0.18	-0.236	-0.005	0.002	0.01	0.28	0.559
1 year								
Intercept	-0.028	0.221	-0.278	-0.085	-0.036	0.015	0.3	0.314
Age	0.002	0.082	-0.019	0.000	0.000	0.001	0.009	0.649
Persistent	0.035	0.195	-0.263	0.017	0.038	0.061	0.277	0.847
Log(CRP)	-0.001	0.198	-0.272	-0.003	0.005	0.012	0.165	0.664
Covariance								
Λ_{11}	0.21	0.922	0.005	0.006	0.007	0.008	0.846	1.000
Λ_{12}	-0.017	0.382	-0.09	-0.001	0.000	0.000	0.069	0.330
Λ_{22}	0.191	0.847	0.005	0.006	0.007	0.008	0.637	1.000

20% in the rest. This reduces inflammation and hence CRP values. Their risk may be explained by other factors, for instance, the prevalence of HBP is 59.7% in these patients, compared with 28.5% in the younger ones.

The predictive results from Model 2 are summarized in Table 3. Essentially, the conclusions for the Persistent group and the effect of CRP persist. The use of Anticoagulants (AA) represent a slight decrease in the risk at 30 days and 1 year. For the rest of the parameters (HBP, LVD and Log(TAT)), the results in Table 3 seem inconclusive at 30 days. The same can be said for Age. At 1 year, Age and LVD cleary express as risk factors, but the effect of HBP and Log(TAT) remain unclear. Again, the covariance parameter Λ_{12} indicates low association between 30 days and 1 year responses. Figure 8 details predictive relative risk at 30 days and 1 year for different levels of CRP, Age and Paroxysmal/Persistent groups. Compared with Model 1, the adjustment for the additional effects sharpens the effect of CRP. It can be seen that the risk gets stable for high levels of CRP (see also Figure 10), and the main increase of risk manifests in relation with small increases at low levels of CRP. In this segment, the effect of Age dilutes in the presence of the additional factors. At high levels of CRP, the inverse effect of Age seen in Model 1 becomes more clear.

TABLE 3. Summary statistics for posterior predictive distribution of parameters in Model 2. Columns indicate mean, standard deviation, percentiles and probability of being positive.

	Mean	S.Dev.	P05	P25	P50	P75	P95	P(>0)
30 days								
Intercept	-0.038	0.289	-0.530	-0.145	-0.057	0.050	0.490	0.336
Age	-0.010	0.132	-0.073	-0.002	0.001	0.002	0.048	0.586
HBP	-0.013	0.280	-0.521	-0.065	-0.012	0.036	0.485	0.417
LVD	-0.016	0.290	-0.564	-0.076	0.009	0.062	0.469	0.548
AA	0.000	0.295	-0.517	-0.055	-0.001	0.049	0.514	0.494
Persistent	0.040	0.269	-0.477	-0.010	0.059	0.101	0.522	0.741
Log(CRP)	0.002	0.278	-0.490	-0.021	0.003	0.028	0.469	0.546
Log(TAT)	-0.006	0.283	-0.500	-0.027	0.000	0.017	0.460	0.509
1 year								
Intercept	-0.021	0.307	-0.513	-0.122	-0.043	0.063	0.517	0.367
Age	0.000	0.111	-0.060	-0.002	0.001	0.002	0.054	0.603
HBP	0.006	0.284	-0.440	-0.053	-0.005	0.039	0.496	0.462
LVD	0.017	0.277	-0.488	-0.045	0.020	0.080	0.487	0.608
AA	0.000	0.288	-0.471	-0.057	-0.009	0.050	0.512	0.435
Persistent	0.025	0.279	-0.448	-0.019	0.035	0.081	0.508	0.698
Log(CRP)	0.003	0.281	-0.498	-0.018	0.004	0.025	0.509	0.586
Log(TAT)	-0.002	0.274	-0.469	-0.016	0.001	0.021	0.490	0.519
Covariance								
Λ_{11}	0.470	1.366	0.005	0.007	0.010	0.257	3.045	1.000
Λ_{12}	0.010	0.620	-0.365	-0.002	0.000	0.001	0.396	0.413
Λ_{22}	0.456	1.280	0.005	0.006	0.009	0.261	2.845	1.000

Similarity Analysis shows a general tendency of the clustering process to group observations in one cluster for Model 1 (Figure 11), with some outliers, which correspond to the red points in the individual posterior relative risk pairs of Figure 5. A more sensitive description of the clustering process is revealed in Figure 14. It can be seen there that the two outliers in the rigth part of the graph have some relation with the principal cluster. Two points in the bottom, which appear as outliers in Figure 5, are revealed as FCRs of clusters of little representation. These observations, together with the third lowest point in the graph, and the main FCR represented with the big circle in Figure 14, have 4 of the 7 lowest CRP values in the sample (0.01), a value that is relatively far from the next sorted value of 0.1. In model 2 (Figures 12 and 14), approximately three low representation clusters are identified, together with the same outliers as before. Figure 13 shows autonomy plots for Models 1 and 2. It can be seen that the identification of outliers, and highly auonomous individuals in general, is consistent in both models. Because of

its specification, which considers more covariates related to the response, Model 2 captures more individual differences thatn Model 1. The mean number of clusters obtained from Model 1 is 2.2 (1.5 s.d.), and 5.2 (2.9 s.d.) for Model 2. The mean for mass parameter M was 0.38 (0.39 s.d.) for Model 1, and 1.04 (0.80 s.d.) for Model 2.

Comparison between Models 1 and 2 was done based on conditional predictive ordinates (CPO) (Gelfand, Dey, Chang, 1992). Denoting as y_{-i} the set of observations without considering observation i, the CPO is defined as the conditional $p(y_i|y_{-i})$. The CPO for each observation i can be approximated by Monte Carlo integration as

$$C\hat{P}O_i = \left(\frac{1}{N} \sum_{k=1}^{N} \frac{1}{p(y_i|\theta_i^{(k)}, x_i)} \right)^{-1}$$

where N is the size of the MCMC sample. See Chen, Shao and Ibrahim (2000) for further discussion on this topic. For the application presented here, let $q_{i,11} = p(z_{i1} \geq 0, z_{i2} \geq 0)$, $q_{i,01} = p(z_{i1} < 0, z_{i2} \geq 0)$, $q_{i,00} = p(z_{i1} < 0, z_{i2} < 0)$ and $q_{i,10} = p(z_{i1} \geq 0, z_{i2} < 0)$. These numbers are obtained from the posterior bivariate Normal probability for z_i based on the sampled value of the vector θ_i. Then we have

$$p(y_i|\theta_i, x_i) = y_{i1}y_{i2}q_{i,11} + (1-y_{i1})y_{i2}q_{i,01} + (1-y_{i1})(1-y_{i2})q_{i,00} + y_{i1}(1-y_{i2})q_{i,10}.$$

We compare models based on the quotient

$$\frac{\prod_i p(y_i|y_{-i})_{\text{Model 1}}}{\prod_i p(y_i|y_{-i})_{\text{Model 2}}}.$$

In our case, we obtained -1483 for the numerator in logarithmic scale, and -7656 for the denominator, so in terms of cross-validation model assessment, we choose Model 1.

4. Discussion

The main interest in this application was to study the influence of CRP in the risk of presenting AF. We have obtained evidence that, from all possible risk factors considered, inflammation, measured by CRP, and the condition of presenting persistent AF, represent an increase in the risk, although moderate. This can be concluded even after considering additional factors that may relate to AF, namely

TABLE 4. Predictive power of proposed models. Values in %.

	Se	Sp	PPV	NPV
30 days				
Model 1	49.9	52.1	41.3	60.9
Model 2	49.3	52.1	41.4	60.9
1 year				
Model 1	50.4	53.3	52.6	51.3
Model 2	49.5	53.7	52.8	51.2

LVD, HBP, AA and TAT levels. The formulation based on Species Sampling mixture allows the expression of variations in the observed relationship between response and covariates for each individual. Each subject can express the relation in its own way, or tend to behave like another observation in the sample, in terms of the covariates. As a result, we conclude that each subject may have an idiosyncratic relation, or tend to follow the model in the same way, except for some few observations, all from the Paroxysmal group. Individuals that behave different from the majority differentiate between them, too. This result is consistent on both proposed models, suggesting the need to investigate what makes these individuals behave differently. In order to assess the overall predictive power of the models, for each stage j, we can calculate Sensitivity (Se) from $p(z_j >= 0|y_j = 1)$ and Specificity (Sp) from $p(z_j < 0|y_j = 0)$. Positive predictive value (PPV) and Negative predictive value (NPV) can be calculated by means of Bayes theorem. We have $PPV = \frac{Se}{p(z_j \geq 0)} p(y_j = 1)$, and, similarly, $NPV = \frac{Sp}{p(z_j < 0)} p(y_j = 0)$. Collapsing over all individuals, and approximating $p(y_j = 1)$ by the empirical estimation $\sum_i y_{ij}/n$, we obtain the results shown in Table 4. Both models show a similar predictive power, indicating that the additional covariates included in Model 2 do not improve much the association between predicted and observed responses. It would be interesting to see if a different parameterization of the variance that takes into account identifiability restrictions can improve these results. It can be concluded from our results that CRP and Persistent state are positively related with the risk of presenting AF at 30 days and 1 year, although further research has to be done to identify more individual characteristics that could predict this outcome.

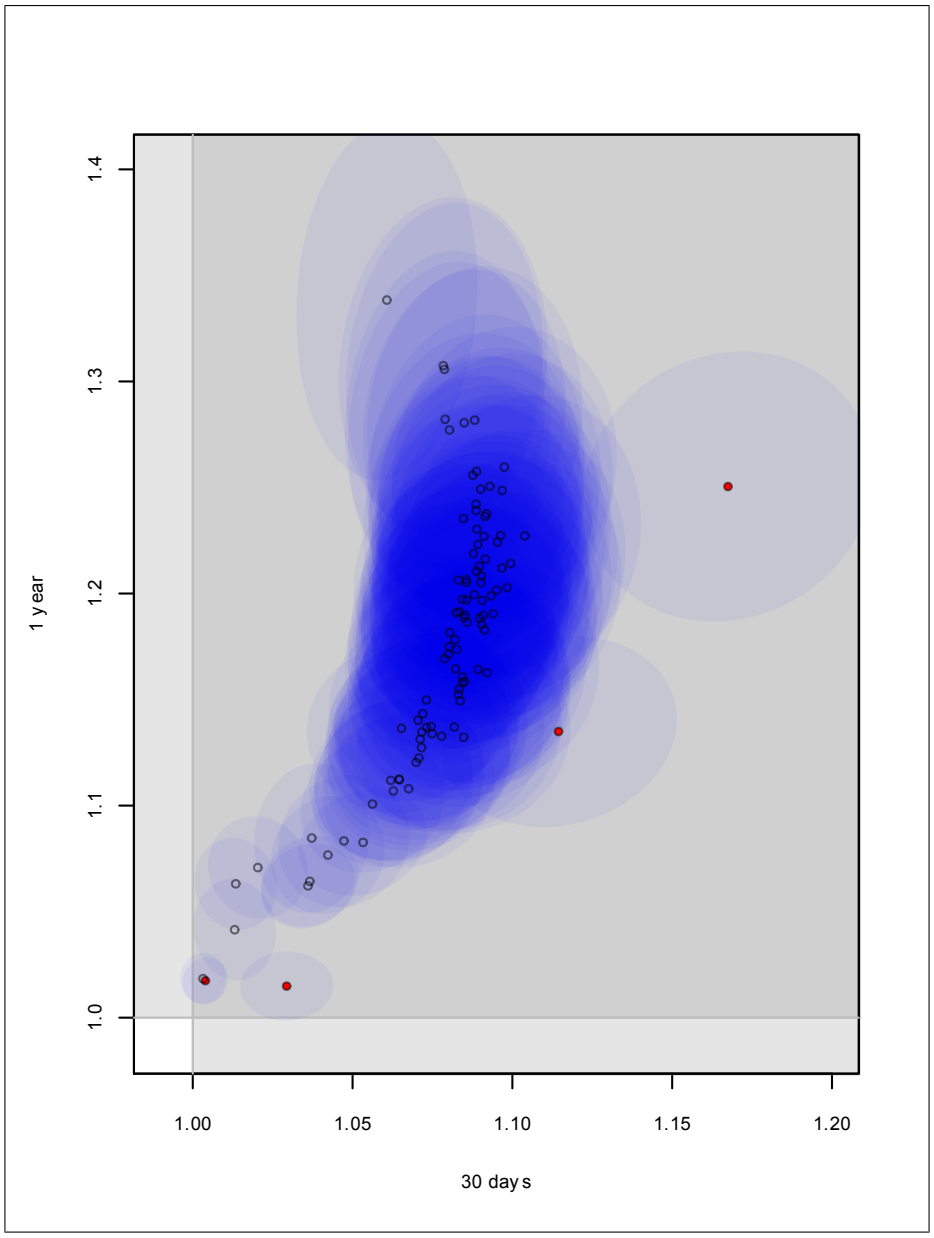

FIGURE 5. Posterior 30 days and 1 year estimated relative risk of FA according to Model 1. Ellipses represent 95% credibility region.

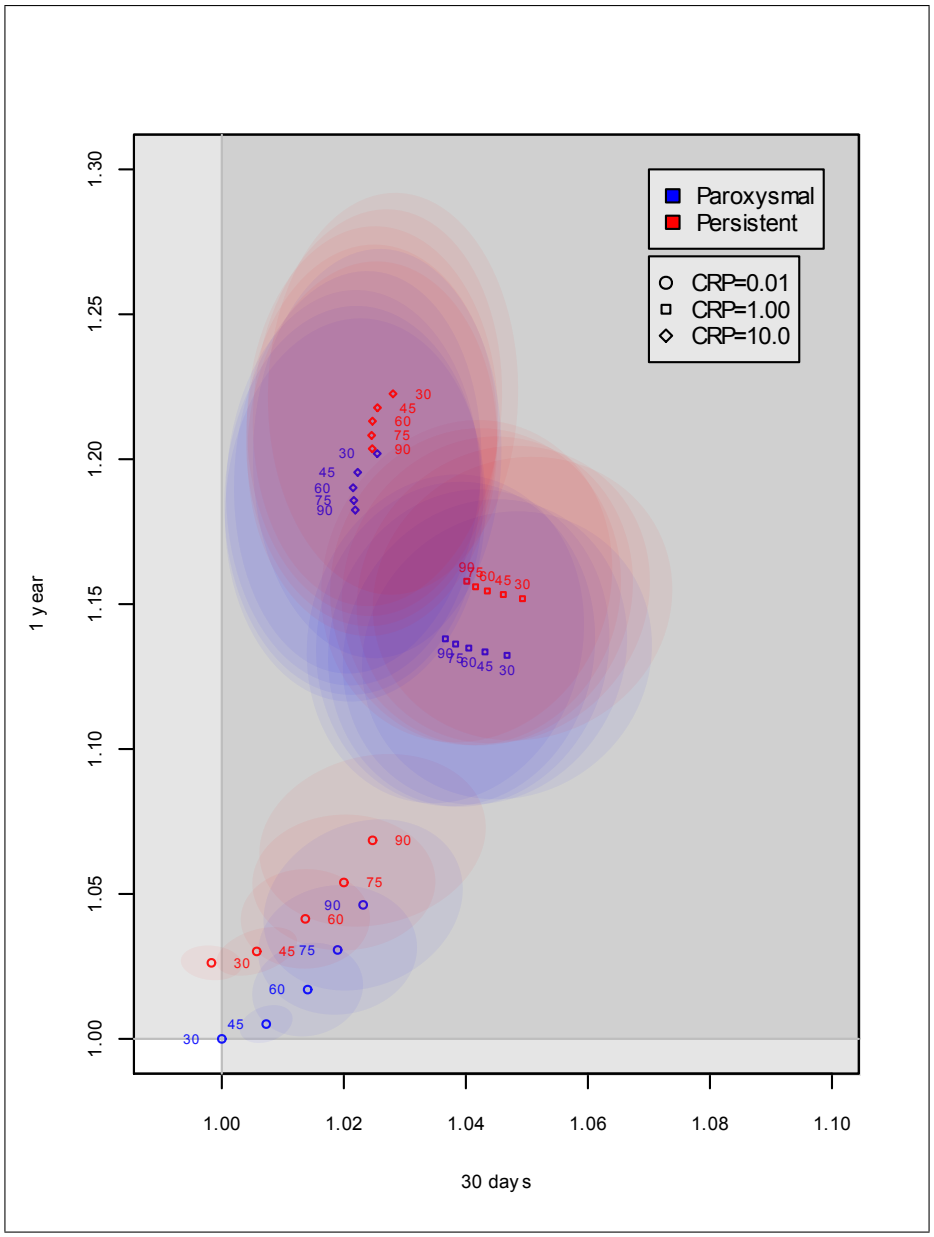

FIGURE 6. Posterior 30 days and 1 year predictive relative risk of FA according to Model 1, by Age, Paroxysmal/Persistent status and Log(CRP). Numbers denote Age. Ellipses represent 95% credibility region.

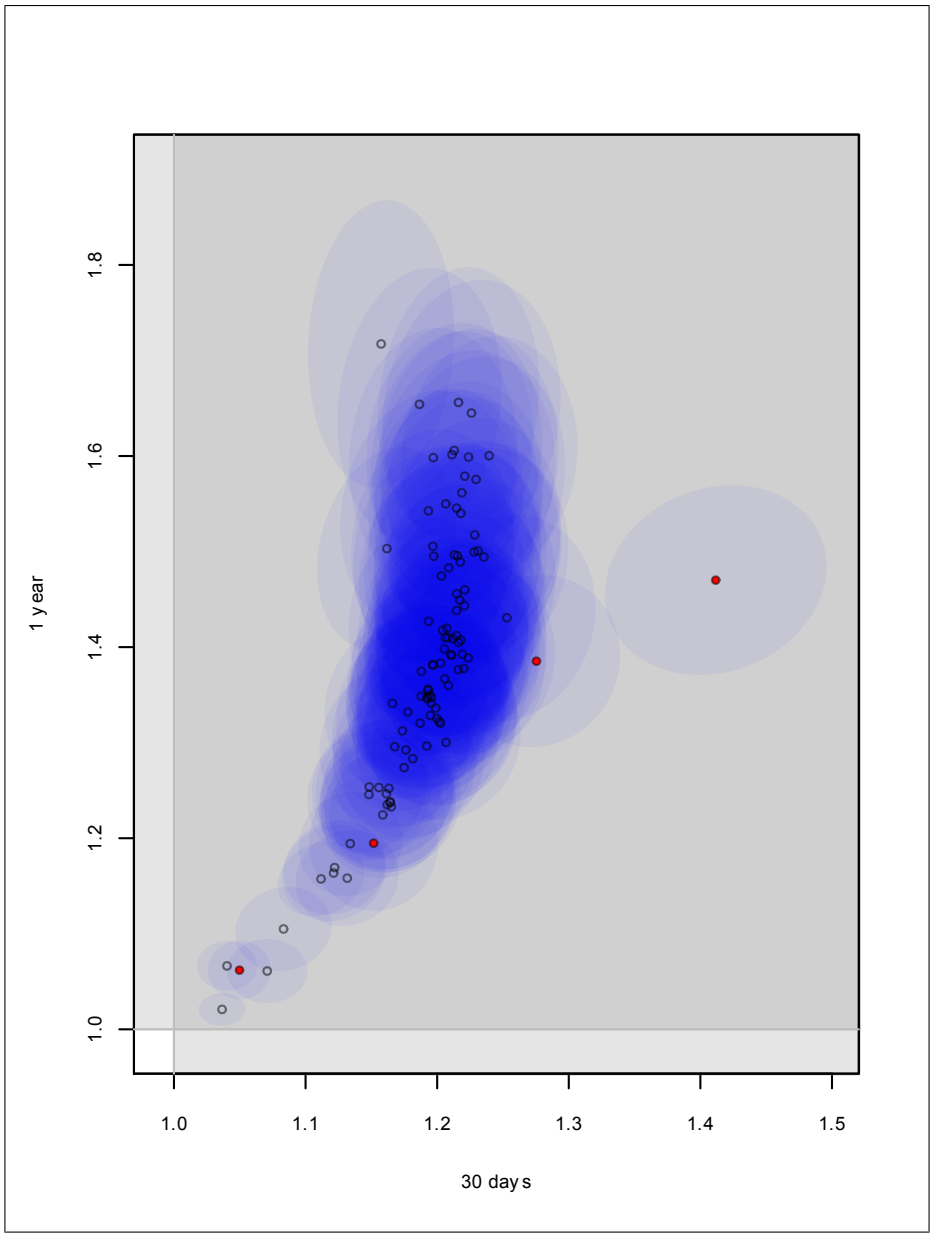

FIGURE 7. Posterior 30 days and 1 year estimated relative risk of FA according to Model 2. Ellipses represent 95% credibility region.

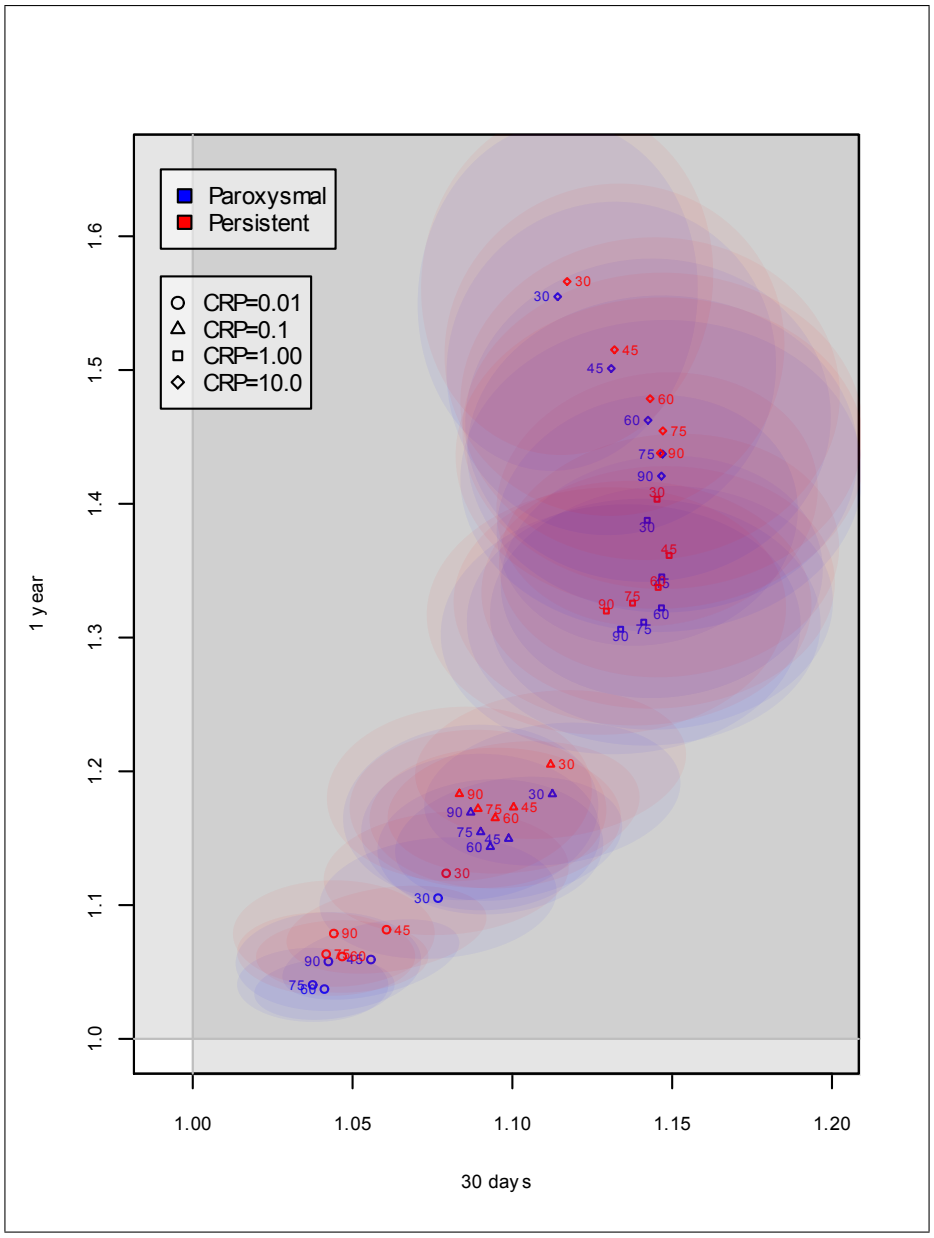

FIGURE 8. Posterior 30 days and 1 year predictive relative risk of FA according to Model 2, by Age, Paroxysmal/Persistent status and Log(CRP). Numbers denote Age. Ellipses represent 95% credibility region.

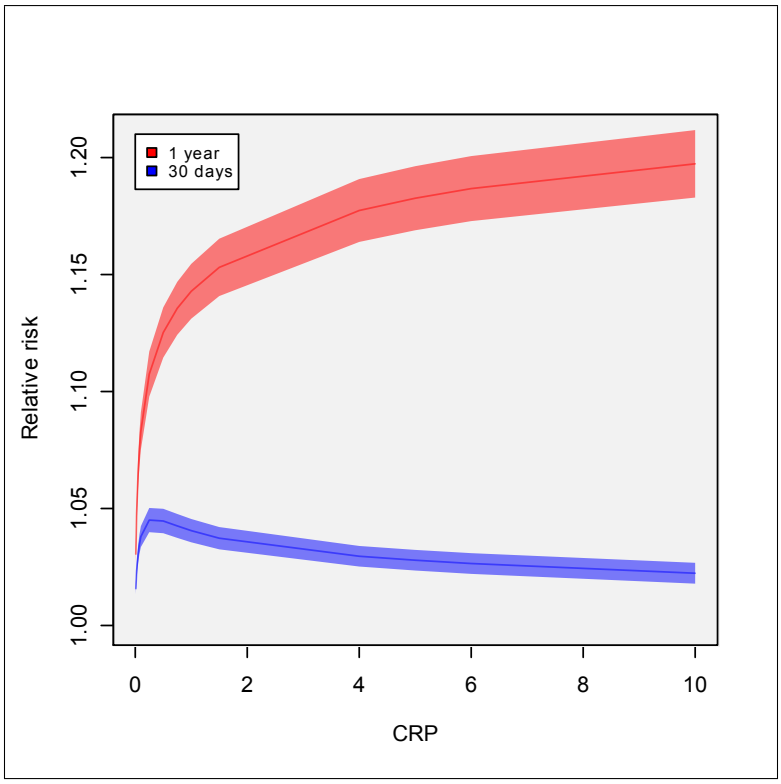

FIGURE 9. Posterior predictive 30 days and 1 year marginal relative risk of FA according to Model 1, associated with different values for Log(CRP). Bands represent 95% credibility.

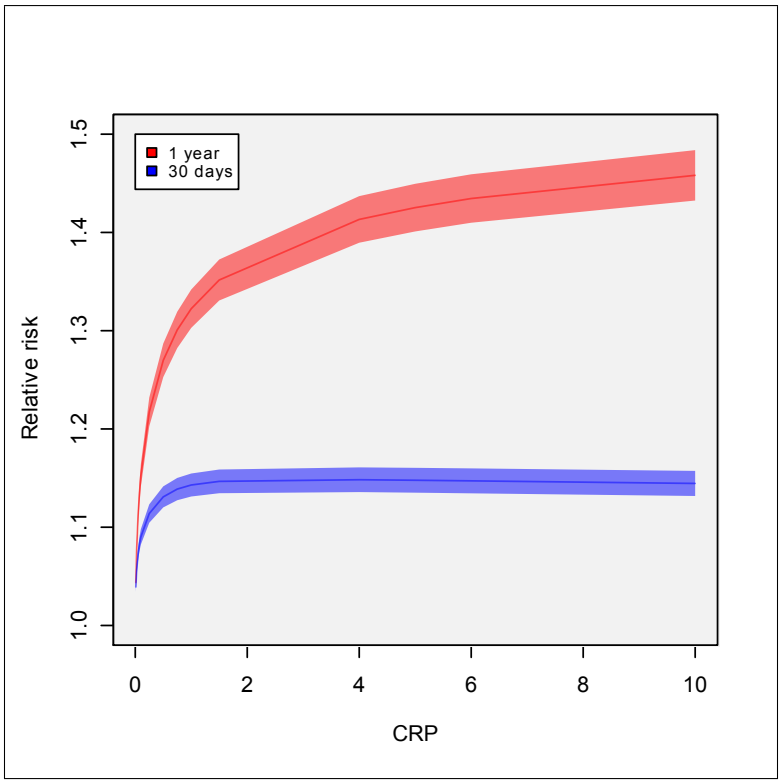

FIGURE 10. Posterior predictive 30 days and 1 year marginal relative risk of FA according to Model 2, associated with different values for Log(CRP). Bands represent 95% credibility.

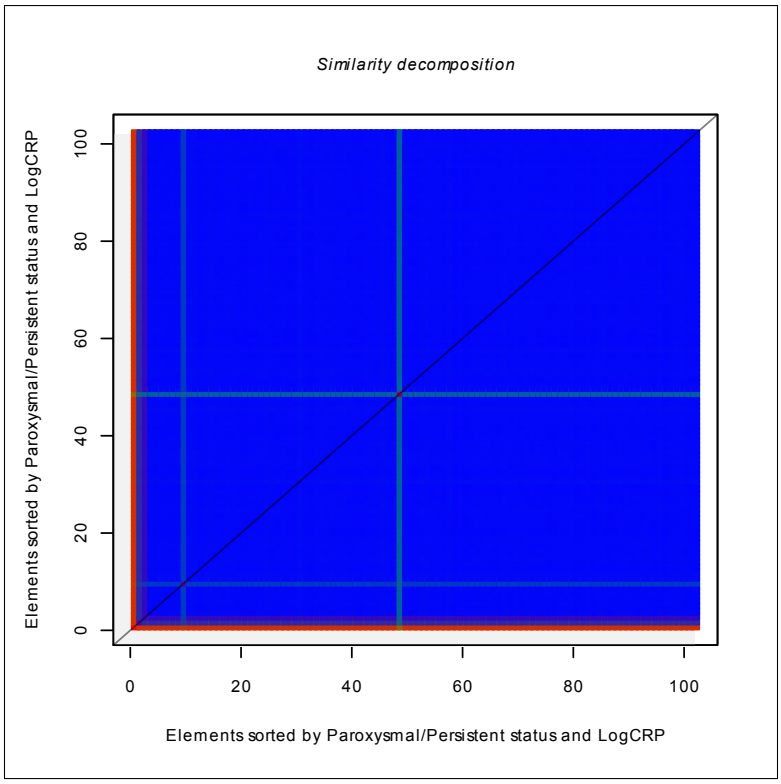

FIGURE 11. SDG for Model 1.

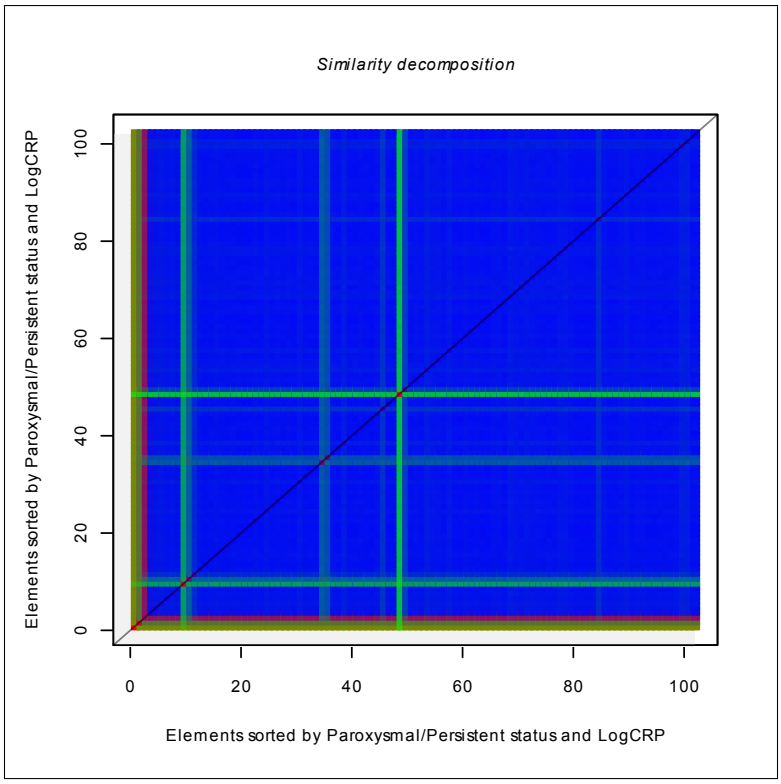

FIGURE 12. SDG for Model 2.

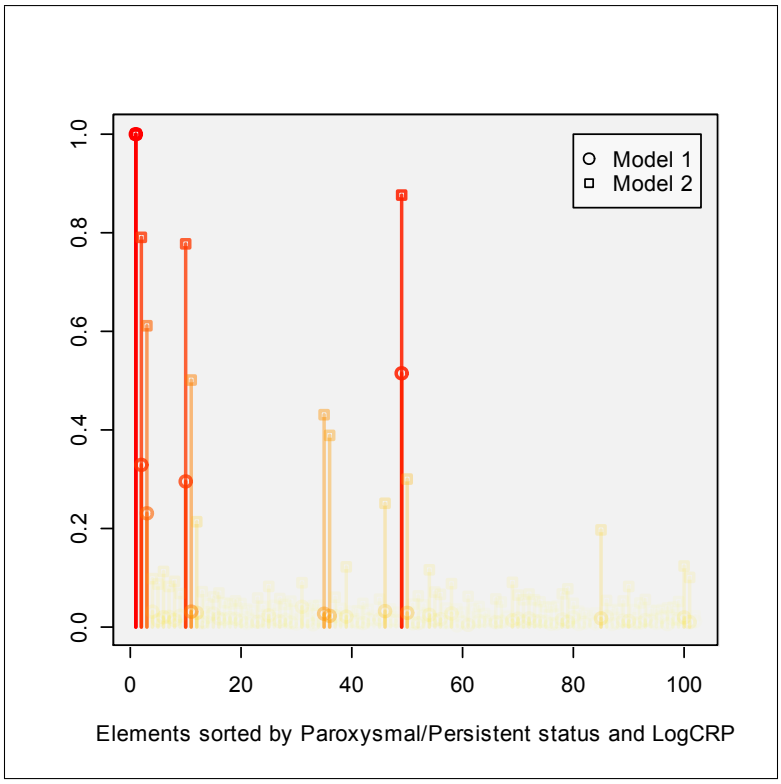

FIGURE 13. Autonomy plot for Models 1 and 2.

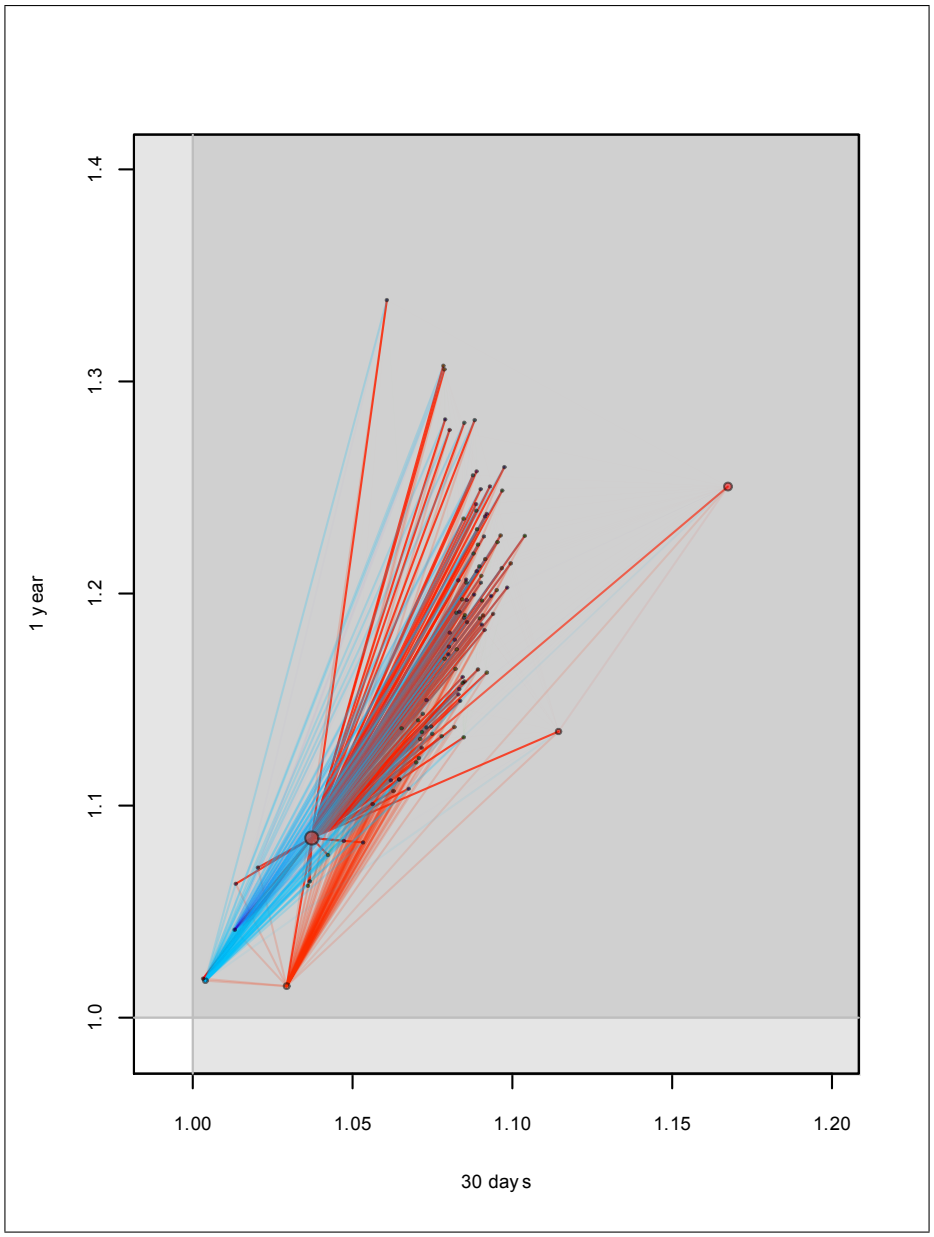

FIGURE 14. Posterior partition structure for Model 1.

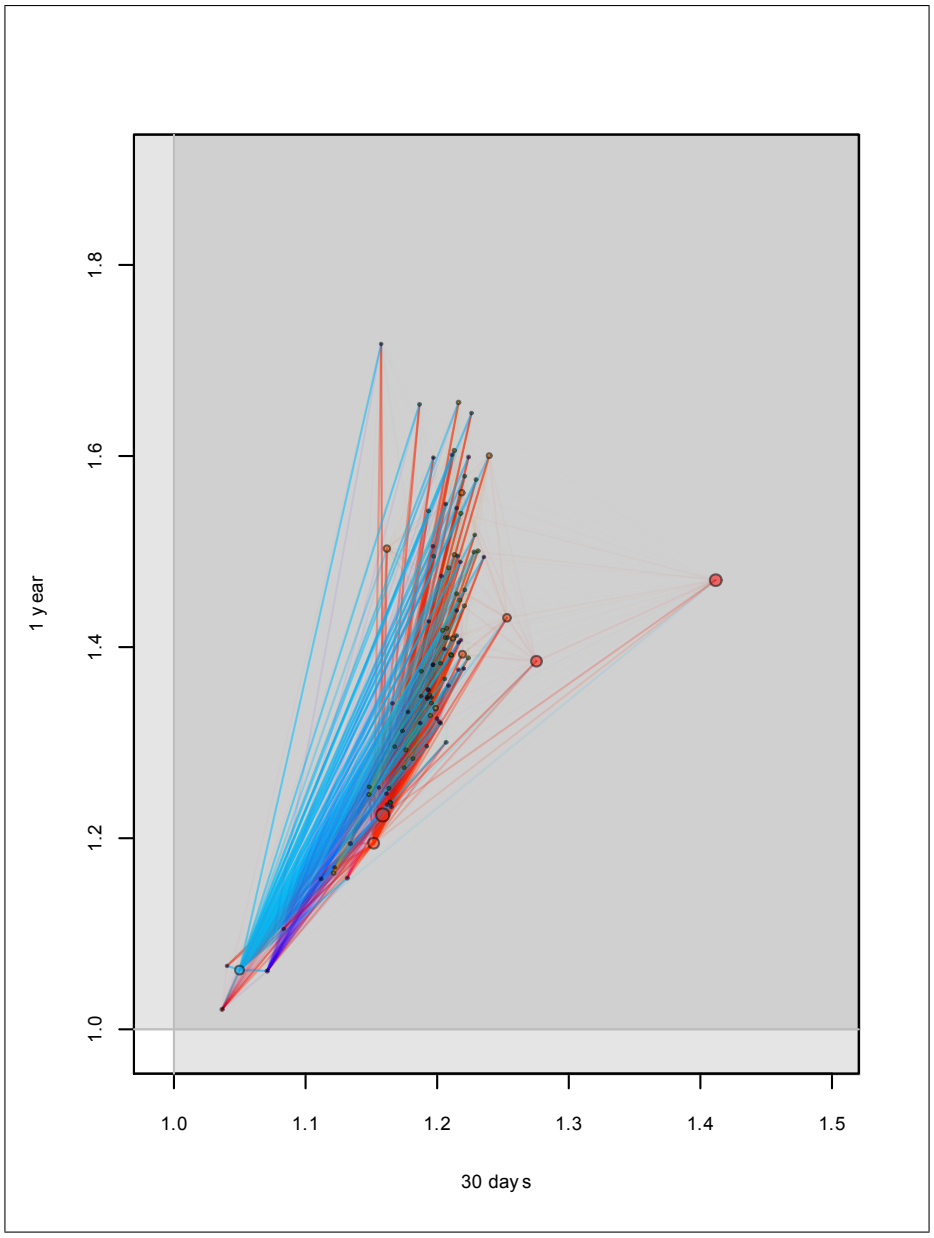

FIGURE 15. Posterior partition structure for Model 2.

CHAPTER 6

Conclusions

SSMMs constitute a type of statistical models capable of adapting to individual and/or grupal differences in the form that the covariates relate with the response. This ability is based on an inherent clustering mechanism that considers partitions for the individual parameters, and tends to cluster together individuals based on the likelihood. The dynamic nature of the process comes from taking into account the latent partition of the data, no matter what it is, and its variability. Understanding partitions and the role of individual characteristics in their formation is crucial to make good use of the power and flexibility of these models. These models make it possible for any difference in how individuals follow the statistical model to manifest. That is, they *adjust* the model specification to take into account characteristics that may not have been considered explicitly in the formulation of the likelihood. So if we are able to understand and interpret this adjustment, we will learn much more about the data.

The representation of partitions by PMs and the explicit identification of FCRs in PMs by means of the decomposition of PMs in intrinsic and extrinsic parts provide a link between partitions and individual characteristics (covariates). This link is closely related to the nature of the probability model assumed for the partitions, based on predictive rules, and the likelihood of the data given the partition and the locations of the clusters that form it.

In order to obtain the same kind of information that we are able to extract from one single partition from the whole partition process, we proposed a method, called Similarity Analysis. This method extends the decomposition of PMs to their expectation, the SM. Based on the Choleski decomposition of the SM, similarity decomposition proved to be an informative and sensitive tool to identify individual influence in the clustering process. We applied the proposed technique to the

Galaxy data, dealing specifically with clustering, and to models considering covariates, based on simulated and real data. Similarity Analysis applied to Galaxy data allowed us to show this analysis in practice, and gave plaussible answers to the clustering problem itself, that is, we were able to make inference about the partition of the data. With the application to simulated data, we were able to clearly identify the influence of the covariates in a controlled situation. This experience was used to improve the quality of information obtained after specifying a SSMM to a real data set. This application was very useful to see the power and flexibility of SSMMs in action, apart from the interpretation provided by the clustering, from which we were able to identify individuals that showed a different behavior in relation to the specified model, consistently across two different specifications.

This work intends to contribute one step in the research of a vast and fascinating field. Further steps that can be seen from current perspective are:

- Application of SSMM priors and Similarity Analysis to model specifications based on more flexible distributions than the Normal case considered here, for instance, Skew-Normal models, considering parameters related with skewness in the nonparametric part of the hierarchy.
- Explore measures of sensitivity of the clustering process to individual characteristics.
- Application of Similarity Analysis to models based on mixtures of other SSMs, and incorporate learning about the parameter α of the PY process.
- Extend the application of Similarity Analysis to other prior specifications for partitions, apart from SSMs.

APPENDIX A

Computational Issues

To obtain samples from the posterior distribution of the parameters in the context of SSM mixtures, MCMC based on Gibbs sampling, sometimes with Metropolis-Hastings steps within, are considered here. The process can be divided in two complementary sampling tasks, for the *parametric* and the *non-parametric* parts of the models. For the parametric part, all available methods should work here. The main discussion is related on how to sample parameters coming from nonparametric distributions. The latter problem consists, in practice, in sampling a partition, which is ruled by a posterior discrete distribution on a finite but huge space. The main problem of the early sampling schemes was the difficult to cover the posterior distribution of cluster locations, due to "sticky" model parameters. The initial schemes proposed moving elements to the existing clusters one at a time (individual reallocation). Some newer aproaches propose merge/split techniques like Dahl's SAMS algorithm (Dahl 2005).

1. Gibbs sampling by individual allocation

One choice considered here to update parameters whose distribution come from the nonparametric part of the models is MacEachern & Müller (1998, 2000) algorithm. Distributions coming from a Dirichlet process are discrete with probability 1 (Ferguson, 1973, Blackwell & MacQueen, 1973). The same result extends to SSM (Pitman 1996). Let's suppose we have a sample $(\alpha_1, \ldots, \alpha_n)$ from a $DP(M, G_0)$ or, in general, a SSM. From (5) and the exchangeability of θ_i, $i = 1, 2, \ldots, n$, we have that the distribution of one θ_i conditional on the rest $\theta_{-i} = (\theta_1, \ldots, \theta_{i-1}, \theta_{i+1}, \ldots, \theta_n)$ is

$$p(\theta_i|\theta_{-i}) = \sum_{j=1}^{k^-} p_j(N_n^-)\delta_{\theta_j^*}(\theta_i) + p_{k^-+1}(N_n^-)G_0(\theta_i) \tag{51}$$

where k^- and N_n^- are the number of clusters and the vector of cluster sizes without considering element i. With a Gibbs sampling scheme in mind, we repeat the j-th sampled value for θ with probability proportional to $p_j(N_n)$, $j = 1, \ldots, k^-$, or we sample from the base distribution G_0 with probability proportional to $p_{k+1}(N_n)$. MacEachern & Müller algorithm is based on updating the vector of cluster memberships $(c_1, \ldots, c_n)$ and then the cluster locations $(\theta_1^*, \ldots, \theta_k^*)$ conditional on the sampled partition configuration. If we combine Bayes theorem with (51), we have

$$p(c_i = j|\theta^*, y, \nu) \propto \begin{cases} p_j(N_n^-)p(y_i|\theta_j^*, \nu) & j = 1, \ldots, k^- \\ \\ p_{k^-+1}(N_n^-) \int p(y_i|\theta, \nu) dG_0(\theta) & j = k^- + 1 \end{cases} \tag{52}$$

Note that $k^- = k$ if the current value of n_{c_i} is greater than one, and $k-1$ otherwise. Conditional on cluster configuration, cluster locations can be sampled from

$$p(\theta_j^*|s, y, \nu) \propto \prod_{s_i=j} p(y_i|\theta_j^*, \nu) G_0(\theta_j^*) \tag{53}$$

1.1. Conjugate models. When G_0 is conjugate with $p(y|\theta, \nu)$, Gibbs sampling goes as follows. Sample a starting configuration from prior distribution (51). Then repeat the following steps:

(1) Sample $(\theta_1^*, \ldots, \theta_k^*)$ from (53)
(2) For $i = 1, \ldots, n$, sample c_i from (52). Note that, at each step, clusters can be created or deleted. When a new cluster is created, increment k and sample θ_k^* from (53). To program the algorithm in a computer, *deleting* a cluster j is equivalent to leaving it empty, that is, $n_j = 0$. Non-empty clusters must be kept together, due to the restrictions for partition representation, so deleting a cluster is equivalent to exchanging it with the last one. Then it must be assured that cluster labels are sorted by their first element, so exchanging cluster labels may be necessary.
(3) Continue the Gibbs sampling scheme for the rest of the parameters in the usual way.

1.2. Non-conjugate models. In non-conjugate models, there is no explicit form for the integral in (52). In that situation, MacEachern & Müller (1998) propose the *No Gaps* algorithm. It is based on having latent $(\theta_{k+1}, \ldots, \theta_n)$, which are in

fact sampled only when needed, from the base distribution G_0. Replace the second step of the previous algorithm with the following one:

(2) If $n_{s_i} > 1$, set $k^- = k$ and sample c_i from

$$p(c_i = j|c_{-i}, \theta) \propto \begin{cases} n_j^- p(y_i|\theta_j^*, \nu) & j = 1, \ldots, k^- \\ \\ \frac{M}{k^-+1} p(y_i|\theta_{k^-}^*, \nu) & j = k^- + 1 \end{cases} \tag{54}$$

If $n_{c_i} = 1$ then with probability $(k-1)/k$ leave s_i unchanged. Otherwise, set $k^- = k - 1$ and resample s_i from (54). If the new s_i equals $k^- + 1 = k$, then θ_k^* remains unchanged. Otherwise, keep θ_k^* for a future new cluster, and decrement k.

A good explanation to this algorithm and some extensions is given in Neal (1998).

2. SAMS algorithm

Other algorithms consider Metropolis-Hastings steps to update the partitions. One of such algorithms is Dahl's SAMS algorithm. This algorithm was used in the application of section 5, providing faster convergence and a noticeable lower computational burden in the partition update process. As in the preceeding algorithm, there is one version for conjugate and other for non-conjugate SSM specifications.

2.1. Conjugate version.

(1) Let π be the current partition of $[n]$. Form a new partition π^* by means of uniformly selecting a pair of distinct indices i and j and:

(2) If i and j belong to the same cluster C, split the cluster, forming two new singleton clusters with i and j, namely C_i and C_j. Then, for each k in a random permutation of the remaining indices in C, add k to C_i with probability

$$p(k \in C_i|C_i, C_j, y) = \frac{N_{C_i} \int p(y_k|\theta)p(\theta|y_{C_i})dG_0(\theta)}{N_{C_i} \int p(y_k|\theta)p(\theta|y_{C_i})dG_0(\theta) + N_{C_j} \int p(y_k|\theta)p(\theta|y_{C_j})dG_0(\theta)} \tag{55}$$

where $p(\theta|y_C)$ is the posterior distribution of a cluster location based on the base measure $G_0(\theta)$ and the observations corresponding to the indices in C. N_C is the size of cluster C. With probability complementary to

(55), add k to C_j. Note that at each step, either C_i or C_j gains an index, and $p(\theta|y_{C_i})$ and $p(\theta|y_{C_i})$ change accordingly. At each step, relabeling the clusters may be necessary to comply the partition representation rules.

(3) If i and j belong to different clusters C_i and C_j, merge those clusters, and update the location for the joined cluster from (53).

(4) In either case, accept the new partition formed with probability

$$a(\pi^*|\pi) = \min\left\{1, \frac{p(\pi^*|y)q(\pi|\pi^*)}{p(\pi|y)q(\pi^*|\pi)}\right\} \tag{56}$$

where $p(\pi^*|y)$ and $p(\pi|y)$ are the posterior probabilities for the respective partitions, proportional to $p(y|\pi^*)p(\pi^*)$ and $p(y|\pi)p(\pi)$, and $p(pi^*)$ and $p(pi)$ are the prior probabilities for the partitions, defined from (6). $q(\pi^*|\pi)$ and $q(\pi|\pi^*)$ are the probabilities of proposing π or π^* from the respective state, and come from multiplying (55) or its compliment according to the partitions.

2.2. Nonconjugate version. When the model is nonconjugate, as was the case in section 5, replace (55) with

$$p(k \in C_i|C_i, C_j, y) = \frac{N_{C_i}p(y_k|\theta_{C_i})}{N_{C_i}p(y_k|\theta_{C_i}) + N_{C_j}p(y_k|\theta_{C_j})} \tag{57}$$

where θ_{C_i} and θ_{C_j} are new locations for their respective clusters, that can be sampled from (53). The Metropolis-Hastings ratio is similar to 56, but considering the newly sampled locations.

3. Sampling M

Following Escobar and West (1995), we consider $M \sim \Gamma(a_0, b_0)$. Having current values for M and k (the number of clusters), we sample a latent parameter η from

$$\eta|M, k \sim \mathrm{B}(M+1, n), \tag{58}$$

calculate π_η from

$$\frac{\pi_\eta}{1-\pi_\eta} = \frac{a+k-1}{n(b-\log(\eta))} \tag{59}$$

and then sample M from

$$p(M|\eta, k) \sim \pi_\eta \mathrm{G}(a+k, b-\log(\eta)) + (1-\pi_\eta)\mathrm{G}(a+k-1, b-\log(\eta)) \tag{60}$$

4. Software

Algorithms used in this work were programmed in c# and will be available for public use in the near future. The rest of the statistical analysis was done using R2.6.0, Copyright 2007 The R Foundation for Statistical Computing. Concerning this software, additional to the *base* package, the following extensions were used. For penalized regression splines, we used *mgcv-package* (Wood, S.N., 2006). For plots, the *Cairo* package (Urbanek, S., Horner, J., 2007). For analysis of the MCMC chains, the *BOA* package (Smith, B. 2004). For confidence ellipsoids, the *ellipse* package (Murdoch, D., 2006). For multivariate Normal probability calculations, *mvtnorm* package (Hothor, T., Bretz, F., Genz, A., 2006). Some comparisons for Bayesian density estimation based on DP mixture models were done using the *DP* package by Alejandro Jara (2007).

Bibliography

[1] Acevedo, M., Pereira, J., Corbalán, R., Braun, S.,Navarrete, C., Gonzalez, I (2006). C-reactive protein and atrial fibrillation: Evidence for the presence of inflammation in the perpetuation of the arrhythmia, *International Journal of Cardiology* 108, 326331.

[2] Albert, James H., Chib, Siddhartha (1993). Bayesian Analysis of Binary and Polychotomous Response Data, *Journal of the American Statistical Association* , Vol. 88, No. 422, pp. 669-679.

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

[4] Bernardo, Jose M., Smith, Adrian F. M. (1994). Bayesian Theory. (c) 1994 by John Wiley & Sons, Ltd.

[5] Binder, D. A. (1978). Bayesian Cluster Analysis, *Biometrika*, 65, 31-38.

[6] Blackwell, D., MacQueen, J. B. (1973). Ferguson distributions via Plya urn schemes, *The Annals of Statistics* **1**:353-355.

[7] Bruins, Te Velthuis, Yazdanbakhsh, Jansen, van Hardevelt, De Beaumont, et al. Activation of the complement system during and after cardiopulmonary bypass surgery: postsurgery activation involves C-reactive protein and is associated with postoperative arrhythmia. *Circulation* 1997 (Nov);96:35428.

[8] Bush, Cristopher A., MacEachern, Steven. A Semiparametric Bayesian Model for Randomised Block Designs, *Biometrika*, Vol. 83, No. 2 (Jun. 1996), 275-285.

[9] Chen, M., Shao, Q. and Ibrahim, J. Monte Carlo Methods in Bayesian Computation. Springer Series in Statistics, 2000.

[10] Chung, Martin, Sprecher, Wazni, Kanderian, Carnes, et al. C-reactive protein elevation in patients with atrial arrhythmias: inflammatory mechanisms and persistence of atrial fibrillation. *Circulation* 2001 (Dec); 104:288691.

[11] Dahl, David (2005). Sequentially-Allocated Merge-Split Sampler for Conjugate and Nonconjugate Dirichlet ProcessMixture Models. Unpublished work.

[12] De Iorio, Maria, Müller, Peter, Rosner, Gary L., MacEachern, Steven N. An ANOVA model for Dependent Random Measures, *Journal of the American Statistical Association* March 2004, Vol. 99, No. 465.

[13] De La Cruz, R., Quintana, F.A., Müller, P. (2007). Semiparametric Bayesian Classification with Longitudinal Markers. *Applied Statistics, Journal of the Royal Statitical Society, Series C*, 56 (2), 119-137.

[14] Dernellis J, Panaretou M. C-reactive protein and paroxysmal atrial fibrillation: evidence of the implication of an inflammatory process in paroxysmal atrial fibrillation. *Acta Cardiol* 2001 (Dec);56(6):37580.

[15] Dunson, David B., Pillai, Natesh. Bayesian Density Regression. *Journal of the Royal Statistics Society B* (2007) 69, Part 2, pp. 163-183.

[16] Escobar, Michael D. (1994). Estimating normal means with a Dirichlet process prior, *Journal of the American Statistical Association* , Vol. 89, No. 425, pp. 268-277.

[17] Escobar, Michael D., West, Mike (1995). Bayesian density estimation and inference using mixtures, *Journal of the American Statistical Association*, Vol. 90, N0. 430, pp. 577-588.

[18] Ferguson, T. S. (1973). A Bayesian analysis of some nonparametric problems, *The Annals of Statistics* **1**:209-230.

[19] Forbes, J.D. (1857). Further experiments and remarks on the measurement of heights by the boiling of water. *Transactions of the Royal Society* 21, 135-143.

[20] Frustaci, Chimente, Bellocci, Morgante, Russo, Maseri. Histological substrate of atrial biopsies in patients with lone atrial fibrillation. *Circulation* 1997 (August);96(4):11804.

[21] Gelfand, A.E., Dey, D.K., Chang, H. (1992). Model determinating using predictive distributions with implementation via sampling-based methods (with Discussion). In Bayesian Statistics 4 (Eds. J.M. Bernado,J.O. Berger, A.P. Dawid, and A.F.M. Smith). Oxford: Oxford University Press, pp. 147 167.

[22] Geweke, John (1991). Efficient Simulation from the Multivariate Normal and Student-t Distributions Subject to Linear Constraints and the Evaluation of Constraint Probabilities. *Computing Science and Statistics: the Twenty-Third Symposium on the Interface, Seattle, April 22-24, 1991.*

[23] Griffin, J. E., Steel, M. F. J. (2006). Order-based dependent Dirichlet processes, *Journal of the American Statistical Association*, 101, 179194.

[24] Hansen, Ben, Pitman, Jim. Prediction Rules for Exchangeable Sequences Related to Species Sampling, *Statistics & Probability Letters* 46 (2000) 251-256.

[25] Heller, Katherine A., Ghahramani, Zoubin (2005). Byesian Hierarchical Clustering, *Proceedings of the 22nd. International Conference on Machine Learning, Bonn, Germany, 2005.*

[26] Ishwaran, H., James, L. F. (2001). Gibbs Sampling Methods for Stick-Breaking Priors, *Journal of the American Statistical Association* 96: 161173.

[27] Ishwaran, H., James, L. F. (2003). Generalized weighted Chinese restaurant processes for species sampling mixture models, *Statistica Sinica* 13: 12111235.

[28] Ishwaran, H. Zarepour, M. (2000). Markov chain Monte Carlo in approximate Dirichlet and beta two-parameter process hierarchical models, *Biometrika* 87: 371390.

[29] Jara, A., García-Zattera, M. J., Lesaffre, E. (2007). A Dirichlet process mixture model for the analysis of correlated binary responses, *Computational Statistics & Data Analysis* 51 (2007) 54025415.

[30] Klaska, J. Transitivity and partial order, *Math. Bohemica* 122 (1997) 7582.

[31] Kottas, A., Müller, P., Quintana, F. (2005). Nonparametric Bayesian modeling for multivariate ordinal data, *Journal of Computational and Graphical Statistics* 14(3): 610625.

[32] Lau, John W., Green, Peter J. Bayesian Model-Based Clustering Procedures, *Journal of Computational and Graphical Statistics*, Volume 16, Number 3, p.p. 526-558.

[33] Li, Baibing (2006). A new approach to cluster analysis: the clustering-function-based-method, *Journal of the Royal Statistical Society B (2006)* 68, Part 3, pp. 457-476.

[34] Liu, J. S. (1996). Nonparametric hierarchical Bayes via sequential imputations, *The Annals of Statistics* 24: 911930.

[35] Lijoi, Antonio, Mena, Ramsés H., Prünster, Igor. Controlling the Reinforcement in bayesian Nonparametric Mixture Models, *Journal of the Royal Statistical Society B (2007)* 69, Part 4, pp. 715-740.

[36] Lo, Albert Y. (1984). On a class of Bayesian nonparametric estimates: I. Density estimates, *The Annals of Statistics* , Vol. 12, No. 1, pp. 351-357.

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

[38] MacEachern, S., Müller, P. (1998). Estimating mixture of Dirichlet process models, *Journal of computational and graphical statistics*, 7, 223-239.

[39] MacEachern, S., Müller, P. (2000). Efficient MCMC schemes for robust model extensions using encompassing Dirichlet process mixture models, in *Robust Bayesian analysis*, eds. F. Ruggeri and D. R. Insua, New York: Springer-Verlag.

[40] Müller, P., Quintana, F.A. (2004). Nonparametric Bayesian Data Analysis. *Statistical Science*, 19(1), 95-110.

[41] Muliere, P., Secchi, P. (1995). A note on a proper Bayesian Bootstrap, Technical Report 18, Universita degli Sudi di Pavia, Dipartamento di Economia Politica e Metodi Quantitativi.

[42] Navarrete, C., Quintana, F. A., Müller, P. (2008). Some Issues on Nonparametric Bayesian Modelling Using Species Sampling Models. To appear in Statistical Modelling International Journal.

[43] Neal, Radford M. Markov Chain Sampling Methods for Dirichlet Process Mixture Models, Technical Report No. 9815, Department of Statistics, University of Toronto.

[44] Neal, Radford M., Jain, Sonia (2000). A Split-Merge Markov Chain Monte Carlo Procedure for the Dirichlet Process Mixture Model. Technical Report No. 2003, Department of Statistics, University of Toronto.

[45] Nuutila, E. Efficient Transitive Closure Computation in Large Digraphs. *Acta Polytechnica Scandinavica, Mathematics and Computing in Engineering Series No 74, Helsinki 1995.*

Published by the Finnish Academy of Technology ISBN 951-666-451-2 ISSN 1237-2404. UDC 681.3

[46] Pitman, J. (1996). Some developments of the Blackwell-Macqueen urn scheme, *Statistics, Probability and Game Theory* IMS Lecture Notes - Monograph Series (1996) Volume 30, pp. 245-267.

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

[48] Quintana, F. A. (2006). A predictive view of Bayesian clustering, *Journal of Statistical Planning and Inference* 136: 24072429.

[49] Richardson, S., Green, P. J. (1997). On Bayesian analysis of mixtures with an unknown number of components (with discussion), *Journal of the Royal Statistical Society, Series B* 59: 731792.

[50] Roeder, K.(1990). Density estimation with confidence sets exemplified by superclusters and voids in the galaxies, *Journal of the American Statistical Association* 85: 617624.

[51] Roldan, Marin, Blann, Garcia, Marco, Sogorb, et al. Interleukin-6, endothelial activation and thrombogenesis in chronic atrial fibrillation. Eur Heart J 2003 (Jul);24(14):137380.

[52] Rota, Gian-Carlo (1964). The Number of Partitions of a Set, *American Mathematical Monthly* 71(5): 498504.

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

[54] Stephens, M. (2000). Bayesian analysis of mixture models with an unknown number of components - alternative to reversible jump methods, *Annals of Statistics* 28: 4074.

[55] Warshall, Stephen (1962). A Theorem on Boolean Matrices, *Journal of the ACM*, Volume 9 , Issue 1 pp. 11 - 12

[56] Walker, Stephen G. (2007). Sampling the Dirichlet Mixture Model with Slices, *Communications in Statistics - Simulation and Computation* 36, 45-54.

[57] Yoeli, Michael (1961). A Note on a Generalization of Boolean Matrix Theory, *The American Mathematical Monthly*, Vol. 68, No. 6. (Jun. -Jul., 1961), pp. 552-557.