Bayesian Analysis (2014) **9**, Number 1, pp. 147–170

A Bayesian Nonparametric Approach for Time Series Clustering

Luis E. Nieto-Barajas * and Alberto Contreras-Cristán †

Abstract. In this work we propose a model-based clustering method for time series. The model uses an almost surely discrete Bayesian nonparametric prior to induce clustering of the series. Specifically we propose a general Poisson-Dirichlet process mixture model, which includes the Dirichlet process mixture model as a particular case. The model accounts for typical features present in a time series like trends, seasonal and temporal components. All or only part of these features can be used for clustering according to the user. Posterior inference is obtained via an easy to implement Markov chain Monte Carlo (MCMC) scheme. The best cluster is chosen according to a heterogeneity measure as well as the model selection criterion LPML (logarithm of the pseudo marginal likelihood). We illustrate our approach with a dataset of time series of share prices in the Mexican stock exchange.

Keywords: Bayes nonparametrics, dynamic linear model, model-based clustering, Pitman-Yor process, time series analysis

1 Introduction

Time series analysis usually concentrates on providing flexible models that account for all possible characteristics inherent in a particular dataset. Describing the probabilistic mechanism that generated the data and producing future predictions are the two main objectives (e.g. Chatfield 1989). On the other hand, in this work we aim at producing clusters of time series that present similar behaviours. Clustering time series becomes relevant in several applications. For example, in portfolio theory (Markowitz 1952), the investor wants to diversify the risk by selecting stocks with different regimes; or in co-integration theory (Granger and Newbold 1974), one might be interested in knowing which set of series can present similar behaviour.

Our motivating example is the clustering of business enterprises that are listed in the Mexican stock exchange based on their monthly share prices. The Mexican stock exchange is the second largest stock exchange in Latin America after the Brazilian one. The benchmark stock index, named IPC, is a broad indicator of the stock exchange's overall performance. This index is constructed as a weighted average of shares that are representative of all the shares listed on the exchange from various sectors across the economy. To better determine the representativeness of a share, it is convenient to identify those shares that show a common behaviour and those that present a distinctive behaviour.

*Department of Statistics, ITAM, Mexico lnieto@itam.mx
†Department of Probability and Statistics, IIMAS-UNAM, Mexico alberto@sigma.iimas.unam.mx

 DOI:10.1214/13-BA852

Within the Bayesian approach, the most commonly used model for time series analysis has been the normal dynamic linear model (Harrison and Stevens 1976). Generalizations of this model started fifteen years later and most of them consider a more flexible distribution for the error terms. For example, scale mixture of normals (Carlin et al. 1992) and finite mixture of normals (Carter and Kohn 1994, 1996). These and other proposals are summarized in Chib and Greenberg (1996). More recently, Bayesian nonparametric generalizations have also been considered. For example Caron et al. (2008) proposed Dirichlet process mixture models for the error terms of both the state and space equations. Fox et al. (2011) considered the switching dynamic linear model and placed a Dirichlet process for modeling the switching regimes. Ghosh et al. (2012), on the other hand, generalized the linearity of a dynamic model by assuming nonparametric functions of the coefficients and covariates and in particular they took Gaussian process priors. Although these alternative Bayesian parametric and nonparametric proposals provide an enhanced flexibility for time series modeling, none of them are suitable for producing clusters.

Bayesian methods for classification of data which are ordered in time have been explored by Zhou and Wakefield (2006) who aimed to discover (fission yeast) genes that exhibit similar behaviour. With a time series defined for each gene in the dataset, their hierarchical model assumes a random effects linear model where the random effect is defined by a random walk process to include time dependence. Additionally, Heard et al. (2006) developed a Bayesian hierarchical clustering method which uses Bayesian regression with basis functions to model time dependent data. They study how to group genes that exhibit similar dynamics in anopheline mosquitoes, after their infection with *Salmonella typhi*.

In this work we propose a (hierarchical) linear regression mixed model that accommodates level, trends, seasonal and time dependent components. The temporal effects are modelled with a first order autoregressive process, similar to the evolution equation in the standard dynamic state-space models (Harrison and Stevens 1976). The joint distribution of some coefficients and the random effects of an entire time series are embedded within a hierarchical nonparametric prior. Specifically we use the Poisson-Dirichlet process prior (Pitman and Yor 1997), which is a member of the stick-breaking processes (Ishwaran and James 2001). These processes are almost surely discrete random measures. It is, in fact, this discreteness property of the Poisson-Dirichlet process that will be used to induce the desired clustering of the time series. For the rest of the coefficients, not considered for clustering, we use hierarchical multivariate parametric priors. In summary, our model can be thought of as a multivariate Poisson-Dirichlet mixture model.

The structure of the paper is as follows: In Section 2 we motivate our proposal starting with a dynamic linear model and describe a Bayesian nonparametric mixture framework for clustering. Section 3 deals with the posterior characterization of the model. In Section 4 we propose a model selection criterion for selecting the best clustering structure. In Section 5 we apply our clustering approach to the motivating data of the share prices listed in the Mexican stock exchange. The paper ends with a discussion in Section 6.

Before we proceed we introduce some notation: $\mathrm{N}(\mu, \sigma^2)$ denotes a normal distribution with mean μ and variance σ^2; $\mathrm{N}_n(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ denotes an n-variate multivariate normal distribution with mean vector $\boldsymbol{\mu}$ and variance-covariance matrix $\boldsymbol{\Sigma}$; $\mathrm{Be}(q_0, q_1)$ denotes a beta distribution with mean $q_0/(q_0+q_1)$; $\mathrm{Ga}(q_0, q_1)$ denotes a gamma distribution with mean q_0/q_1; and $\mathrm{IGa}(c_0, c_1)$ denotes an inverse gamma distribution with mean $c_1/(c_0-1)$.

2 The model

2.1 Sampling model

Let $\mathbf{y}_i = \{y_{it} : t = 1, 2, \ldots, T\}$, $i = 1, \ldots, n$, be a set of n time series, each of them observed during T time periods. One of the most powerful Bayesian models for the analysis of time series is the the dynamic linear model (Harrison and Stevens 1976). This model is described in terms of an observation equation and an evolution or system equation as follows:

$$y_{it} = F_{it}\theta_{it} + \epsilon_{it}, \tag{1}$$

$$\theta_{it} = \rho\theta_{i,t-1} + \nu_{it}, \tag{2}$$

together with $\epsilon_{it} \sim \mathrm{N}(0, \sigma^2_{\epsilon i})$ and $\nu_{it} \sim \mathrm{N}(0, \sigma^2_\theta)$ with independence across i and t. The evolution equation (2) describes a dynamic in the coefficients θ_{it} as an autoregressive process of order one (i.e., an AR(1)). For most time series this construction has been proved to be flexible enough (West and Harrison 1999).

Let us concentrate on the evolution equation (2) and drop for the moment the subindex i, i.e., $\theta_t = \rho\theta_{t-1} + \nu_t$. It is well known (e.g. Chatfield 1989, pg. 35) that an AR(1) process is stationary if one allows the time index t to go from $-\infty$ to ∞. If the time index is bounded, as in our case where $t \in \{1, \ldots, T\}$, Ross (2000, pg. 575) suggests changing the variance of the first innovation ν_1 to achieve stationarity. In particular, if we take $\nu_1 \sim \mathrm{N}(0, \sigma^2_\theta)$ and $\nu_t \sim \mathrm{N}(0, \sigma^2_\theta(1-\rho^2))$ for $t > 1$, it is not difficult to prove that by defining $\theta_1 = \nu_1$ and using the re-scaled innovations in (2), we have that $\theta_t \sim \mathrm{N}(0, \sigma^2_\theta)$ marginally and that $\mathrm{Corr}(\theta_t, \theta_s) = \rho^{|t-s|}$. Therefore, we can re-write the evolution equation (2) for a finite time series as $\boldsymbol{\theta}'_i = (\theta_{i1}, \ldots, \theta_{iT}) \sim \mathrm{N}_T(\mathbf{0}, \mathbf{R})$, where the variance-covariance matrix $\mathbf{R} = (R_{jk})$ has typical element $R_{jk} = \sigma^2_\theta \rho^{|j-k|}$. Note that the prime $'$ denotes transpose.

To accommodate level, trends, seasonal and temporal components in the model, we can define an observation equation, as in (1), such that

$$\mathbf{E}(y_{it}) = \mu_i + \boldsymbol{\omega}'_i\mathbf{g}(t) + \boldsymbol{\upsilon}'_i\mathbf{h}(t) + \theta_{it},$$

where μ_i denotes the level of the series, $\boldsymbol{\omega}'_i\mathbf{g}(t)$ denotes a polynomial trend, which for instance, for a quadratic shape is $\omega_{1i}t + \omega_{2i}t^2$. The component $\boldsymbol{\upsilon}'_i\mathbf{h}(t)$ denotes the seasonal component, which can be defined through latent indicators. If, for example, the observation times of the series are months, the j-th monthly effect could be described in terms of a latent $m_j(t) = I(t = j)$, for $j = 1, \ldots, 12$. In this case the seasonal

component would be $\upsilon_{2i}m_2(t) + \cdots + \upsilon_{12,i}m_{12}(t)$, where the first month indicator is not present to avoid singularity problems in the design matrix. Finally, θ_{it} denotes the temporal component and plays the role of a dynamic intercept that accounts for time dependence in the observations.

Since we are assuming that the observations y_{it} are the result of adding a measurement error ϵ_{it} to a mean level $\mathbf{E}(y_{it})$, our idea is to cluster the (whole) observed time series $\mathbf{y}_i' = (y_{i1}, \ldots, y_{iT})$, $i = 1, \ldots, n$, according to the parameters that determine the mean level (denoised series), that is, $\boldsymbol{\eta}_i = (\mu_i, \boldsymbol{\omega}_i, \boldsymbol{\upsilon}_i, \boldsymbol{\theta}_i)$. However, depending on the data characteristics, not all of the parameters considered in $\boldsymbol{\eta}_i$ will be useful for clustering purposes. For instance two series that share the same trend, seasonalities and temporal components but differ in the level μ_i might be desired to belong to the same cluster. Thus, we will write our general sampling model as

$$\mathbf{y}_i = \mathbf{Z}\boldsymbol{\alpha}_i + \mathbf{X}\boldsymbol{\beta}_i + \boldsymbol{\theta}_i + \boldsymbol{\epsilon}_i, \quad i = 1, 2, \ldots n, \tag{3}$$

where $\mathbf{Z}$ and $\mathbf{X}$ are two design matrices of dimension $T \times p$ and $T \times d$ respectively. The $p \times 1$ dimensional vector $\boldsymbol{\alpha}_i$, the $d \times 1$ dimensional vector $\boldsymbol{\beta}_i$ and the $T \times 1$ dimensional vector $\boldsymbol{\theta}_i$ are parameters of the model such that $\boldsymbol{\eta}_i = (\boldsymbol{\alpha}_i, \boldsymbol{\beta}_i, \boldsymbol{\theta}_i)$, but only $\boldsymbol{\beta}_i$ and $\boldsymbol{\theta}_i$ will be considered for clustering. For example, if the clustering is to be based on everything else rather than the level μ_i then we would take $\boldsymbol{\alpha}_i = \mu_i$ and $\boldsymbol{\beta}_i = (\boldsymbol{\omega}_i, \boldsymbol{\upsilon}_i)$. Finally, $\boldsymbol{\epsilon}_i' = (\epsilon_{i1}, \ldots, \epsilon_{iT}) \sim \mathrm{N}_T(\mathbf{0}, \sigma^2_{\epsilon_i}\mathbf{I})$ is the vector of measurement errors such that $\mathbf{I}$ is the identity matrix of dimension $T \times T$.

2.2 Prior distributions

Let $\boldsymbol{\gamma}_i' = (\boldsymbol{\beta}_i', \boldsymbol{\theta}_i')$ denote the vector of coefficients that will be used for clustering. The idea is to define a joint prior for the whole set $(\boldsymbol{\gamma}_1, \ldots, \boldsymbol{\gamma}_n)$ that allows for ties and at the same time respects the evolution specification (2). In the Bayesian nonparametric setting one of the most widely used priors is the Dirichlet process prior, first introduced by Ferguson (1973). One of the reasons this prior has been so popular is, in fact, its discreteness property that allows for ties in the observations and therefore makes it capable of producing clusters. Moreover, this prior can be centered on any parametric model, univariate or multivariate.

In this paper we propose to use a generalization of the Dirichlet process prior which belongs to the class of stick-breaking priors (Ishwaran and James 2001). In particular we consider the (two parameter) Poisson-Dirichlet (or simply Pitman-Yor) process prior (Pitman and Yor 1997). If a probability measure G has a Poisson-Dirichlet prior with scalar parameters $a \in [0, 1)$, $b > -a$ and mean parameter G_0, which will be denoted $G \sim \mathcal{PD}(a, b, G_0)$, then

$$G(\cdot) = \sum_{k=1}^{\infty} w_k \delta_{\xi_k}(\cdot)$$

is almost surely a discrete random measure with random weights w_k and random locations ξ_k. For this representation $\xi_k \overset{\text{iid}}{\sim} G_0$, $k = 1, 2, \ldots$ and δ_ξ is a point mass at

ξ. The random weights $\{w_k\}$ are defined as $w_1 = v_1$ and $w_k = v_k \prod_{l<k}(1-v_l)$, with $v_k \overset{\text{ind}}{\sim} \text{Be}(1-a, b+ka)$. The specific choice for the distribution of the stick-breaks v_k characterizes the Poisson-Dirichlet process, however the functional parameter G_0 can be specified by the user. The parameter G_0 is known as the centering measure since $\mathbf{E}(G) = G_0$. Two important priors arise as special cases, the Dirichlet process prior when $a = 0$ and the normalized stable process when $b = 0$.

In particular we take

$$\boldsymbol{\gamma}_i | G \overset{\text{iid}}{\sim} G, \quad \text{for} \quad i = 1, \ldots, n \quad \text{with} \quad G \sim \mathcal{PD}(a, b, G_0), \tag{4}$$

and $G_0(\boldsymbol{\gamma}) = G_0(\boldsymbol{\beta}, \boldsymbol{\theta}) = \text{N}_d(\boldsymbol{\beta}|\mathbf{0}, \boldsymbol{\Sigma}_\beta) \times \text{N}_T(\boldsymbol{\theta}|\mathbf{0}, \mathbf{R})$, with $\boldsymbol{\Sigma}_\beta = \text{diag}(\sigma^2_{\beta 1}, \ldots, \sigma^2_{\beta d})$ and $\mathbf{R}$ defined as before. In consequence, this choice of prior implies that the $\boldsymbol{\gamma}_i$'s are exchangeable with marginal distribution $\boldsymbol{\gamma}_i \sim G_0$ for all $i = 1, \ldots, n$. To understand how the ties occur, Pitman (1995) showed that if we integrate out the nonparametric measure G, the joint distribution of the $\boldsymbol{\gamma}_i$'s is characterized by a generalized Polya urn mechanism with conditional distribution that depends on the density g_0 associated to G_0 and given by

$$f(\boldsymbol{\gamma}_i \,|\, \boldsymbol{\gamma}_{-i}) = \frac{b + a m_i}{b + n - 1} g_0(\boldsymbol{\gamma}_i) + \sum_{j=1}^{m_i} \frac{n^*_{j,i} - a}{b + n - 1} \delta_{\boldsymbol{\gamma}^*_{j,i}}(\boldsymbol{\gamma}_i), \tag{5}$$

for $i = 1, \ldots, n$, where $\boldsymbol{\gamma}_i = (\boldsymbol{\gamma}_1, \ldots, \boldsymbol{\gamma}_{i-1}, \boldsymbol{\gamma}_{i+1}, \ldots, \boldsymbol{\gamma}_n)$ denotes the set of all $\boldsymbol{\gamma}_j$'s excluding the i^{th} element, and $(\boldsymbol{\gamma}^*_{1,i}, \ldots, \boldsymbol{\gamma}^*_{m_i,i})$ denote the unique values in $\boldsymbol{\gamma}_{-i}$, each occurring with frequency $n^*_{j,i}$, $j = 1, \ldots, m_i$, which satisfy the condition $n^*_{1,i} + \cdots + n^*_{m_i,i} = n - 1$. Therefore, after integrating the nonparametric measure G, for each pair $\boldsymbol{\gamma}_i = (\boldsymbol{\beta}_i, \boldsymbol{\theta}_i)$, $\boldsymbol{\beta}_i$ and $\boldsymbol{\theta}_i$ are independent with marginal distributions $\text{N}_d(\mathbf{0}, \boldsymbol{\Sigma}_\beta)$ and $\text{N}_T(\mathbf{0}, \mathbf{R})$ respectively, respecting the evolution equation (2), but with dependence across i allowing for ties in the pairs $\boldsymbol{\gamma}_i$. In general, the number of clusters m (unique values in $\boldsymbol{\gamma} = (\boldsymbol{\gamma}_1, \ldots, \boldsymbol{\gamma}_n)$) is determined by the parameters (a, b). Larger values of either a or b, within the valid ranges, produce a larger m (e.g. Navarrete et al. 2008).

Finally, for the parameter vector $\boldsymbol{\alpha}_i$ of the coefficients not considered for clustering, we take a normal prior of the form

$$\boldsymbol{\alpha}_i \overset{\text{iid}}{\sim} \text{N}_p(\mathbf{0}, \boldsymbol{\Sigma}_\alpha), \quad \text{for} \quad i = 1, \ldots, n, \tag{6}$$

with $\boldsymbol{\Sigma}_\alpha = \text{diag}(\sigma^2_{\alpha 1}, \ldots, \sigma^2_{\alpha p})$.

2.3 Hyper-prior distributions

We conclude the specifications of our model by assigning hyper-prior distributions to all hyper-parameters. These are $\sigma^2_{\epsilon_i}$, $i = 1, \ldots n$, $\sigma^2_{\beta_j}$, $j = 1, \ldots d$, $\sigma^2_{\alpha_k}$, $k = 1, \ldots, p$, σ^2_θ, ρ, a and b. For the first three sets of variances, we assign conditionally conjugate priors of the form

$$\sigma^2_{\epsilon_i} \sim \text{IGa}(c^\epsilon_0, c^\epsilon_1), \quad \sigma^2_{\beta_j} \sim \text{IGa}(c^\beta_0, c^\beta_1), \quad \sigma^2_{\alpha_k} \sim \text{IGa}(c^\alpha_0, c^\alpha_1), \tag{7}$$

for $i = 1, \dots n$, $j = 1, \dots d$ and $k = 1, \dots, p$, respectively.

The choice of the prior for (σ_θ^2, ρ) is highly important, since these parameters determine the evolution patterns in the time dependence. For them we propose a joint reference prior derived in Mendoza and Nieto-Barajas (2006), so maximizing the power of the data to determine their best values. This is given by

$$f(\sigma_\theta^2, \rho) \propto (\sigma_\theta^2)^{-1} \frac{\sqrt{1+\rho^2}}{1-\rho^2}, \tag{8}$$

for $\sigma_\theta^2 > 0$ and $\rho \in (-1, 1)$.

Finally, for the Poisson-Dirichlet process parameters (a, b), we consider a joint prior taking ideas from Jara et al. (2010). Since $a \in [0, 1)$ marginally, we take a mixture prior for a with a continuous distribution on $(0, 1)$ and a point mass at zero of the form

$$f(a) = \pi I_{\{0\}}(a) + (1-\pi)\text{Be}(a|q_0^a, q_1^a). \tag{9}$$

Conditionally on a, we incorporate the constraint $b > -a$ by taking a shifted gamma, i.e.,

$$f(b|a) = \text{Ga}(b + a|q_0^b, q_1^b). \tag{10}$$

3 Posterior characterization

If we let $\boldsymbol{\alpha}' = (\boldsymbol{\alpha}_1', \dots, \boldsymbol{\alpha}_n')$, $\boldsymbol{\gamma}' = (\boldsymbol{\gamma}_1', \dots, \boldsymbol{\gamma}_n')$ and $\boldsymbol{\sigma}_\epsilon' = (\sigma_{\epsilon_1}^2, \dots, \sigma_{\epsilon_n}^2)$ then the likelihood function is given by

$$f(\mathbf{y} \,|\, \boldsymbol{\alpha}, \boldsymbol{\gamma}, \boldsymbol{\sigma}_\epsilon) = \prod_{i=1}^{n} \text{N}_T(\mathbf{y}_i \,|\, \mathbf{Z}\boldsymbol{\alpha}_i + \mathbf{X}\boldsymbol{\beta}_i + \boldsymbol{\theta}_i,\, \sigma_{\epsilon_i}^2 \mathbf{I}). \tag{11}$$

The posterior distribution for $\boldsymbol{\alpha}_i$ can be obtained analytically by marginalizing the likelihood with respect to the marginal prior distribution of $\boldsymbol{\gamma}_i$. After some algebra,

(i) the marginal posterior distribution of $\boldsymbol{\alpha}_i$ becomes

$$f(\boldsymbol{\alpha}_i \,|\, \mathbf{y}, \sigma_{\epsilon_i}^2, \boldsymbol{\Sigma}_\alpha, \boldsymbol{\Sigma}_\beta) = \text{N}_p(\boldsymbol{\alpha}_i \,|\, \boldsymbol{\mu}_\alpha, \mathbf{V}_\alpha),$$

for $i = 1, \dots, n$, where $\boldsymbol{\mu}_\alpha = \mathbf{V}_\alpha \mathbf{Z}'\mathbf{W}_i^{-1}\mathbf{y}_i$ and $\mathbf{V}_\alpha = (\mathbf{Z}'\mathbf{W}_i^{-1}\mathbf{Z} + \boldsymbol{\Sigma}_\alpha^{-1})^{-1}$ with matrices

$$\mathbf{Q}_i = \sigma_{\epsilon_i}^2 \mathbf{I} + \mathbf{R}, \tag{12}$$

$$\mathbf{V}_{\beta_i} = (\mathbf{X}'\mathbf{Q}_i^{-1}\mathbf{X} + \boldsymbol{\Sigma}_\beta^{-1})^{-1}, \tag{13}$$

$$\mathbf{W}_i = (\mathbf{Q}_i^{-1} + \mathbf{Q}_i^{-1}\mathbf{X}\mathbf{V}_{\beta_i}\mathbf{X}'\mathbf{Q}_i^{-1})^{-1}, \tag{14}$$

of dimensions $T \times T$, $d \times d$ and $T \times T$, respectively.

Posterior behaviour of the rest of the parameters will be characterized by their full conditional distributions. We now concentrate on $\boldsymbol{\gamma}$. Recall that $\boldsymbol{\gamma}_{-i}$ denotes the set of all $\boldsymbol{\gamma}_j$'s excluding the i^{th} element, and $\boldsymbol{\gamma}^*_{j,i}$'s denote the unique values in $\boldsymbol{\gamma}_{-i}$, each occurring with frequency $n^*_{j,i}$, $j = 1, \ldots, m_i$. We use the generalized Polya urn representation of the prior (5), once the nonparametric part G has been integrated out, and rely on usual posterior computations (e.g. Escobar and West 1998; Ishwaran and James 2001). Then,

(ii) the posterior distribution for $\boldsymbol{\gamma}$, is again characterized by a generalized Polya urn, which gives the full conditional distribution for $\boldsymbol{\gamma}'_i = (\boldsymbol{\beta}'_i, \boldsymbol{\theta}'_i)$ as

$$f(\boldsymbol{\gamma}_i \,|\, \mathbf{y}, \boldsymbol{\gamma}_{-i}, \boldsymbol{\sigma}_\epsilon, \boldsymbol{\Sigma}_\beta, \mathbf{R}) = q_0\, g_0(\boldsymbol{\gamma}_i \,|\, \mathbf{y}_i, \sigma^2_{\epsilon_i}, \boldsymbol{\Sigma}_\beta, \mathbf{R}) + \sum_{j=1}^{m_i} q_j \delta_{\boldsymbol{\gamma}^*_{j,i}}(\boldsymbol{\gamma}_i),$$

for $i = 1, \ldots, n$, where

$$g_0(\boldsymbol{\gamma}_i \,|\, \mathbf{y}_i, \sigma^2_{\epsilon_i}, \boldsymbol{\Sigma}_\beta, \mathbf{R}) = \mathrm{N}_T(\boldsymbol{\theta}_i \,|\, \boldsymbol{\mu}_{\theta_i}, \mathbf{S}_{\theta_i}) \times \mathrm{N}_d(\boldsymbol{\beta}_i \,|\, \boldsymbol{\mu}_{\beta_i}, \mathbf{V}_{\beta_i}),$$

with variance-covariance matrices $\mathbf{S}_{\theta_i} = ((\sigma^2_{\epsilon_i}\mathbf{I})^{-1} + \mathbf{R}^{-1})^{-1}$ and $\mathbf{V}_{\beta_i}$ given in (13), and vectors $\boldsymbol{\mu}_{\theta_i} = \mathbf{S}_{\theta_i}(\sigma^2_{\epsilon_i}\mathbf{I})^{-1}(\mathbf{y}_i - \mathbf{Z}\boldsymbol{\alpha}_i - \mathbf{X}\boldsymbol{\beta}_i)$ and $\boldsymbol{\mu}_{\beta_i} = \mathbf{V}_{\beta_i}\mathbf{X}'\mathbf{Q}_i^{-1}(\mathbf{y}_i - \mathbf{Z}\boldsymbol{\alpha}_i)$, with $\mathbf{Q}_i$ given in (12). The weights q_0 and q_j are computed by setting $D_0 = (b + am_i)\mathrm{N}(\mathbf{y}_i \,|\, \mathbf{Z}\boldsymbol{\alpha}_i, \mathbf{W}_i)$, with $\mathbf{W}_i$ given in (14) and $D_j = (n^*_{j,i} - a)\mathrm{N}(\mathbf{y}_i \,|\, \mathbf{Z}\boldsymbol{\alpha}_i + \mathbf{X}\boldsymbol{\beta}^*_{j,i} + \boldsymbol{\theta}^*_{j,i}, \sigma^2_{\epsilon_i}\mathbf{I})$, so that

$$q_0 = \frac{D_0}{D_0 + \sum_{j=1}^{m_i} D_j}, \quad q_j = \frac{D_j}{D_0 + \sum_{j=1}^{m_i} D_j}, \quad j = 1, \ldots, m_i.$$

The conditional posterior distribution for the variances $\sigma^2_{\epsilon_i}$, $i = 1, \ldots n$, $\sigma^2_{\beta_j}$, $j = 1, \ldots d$ and $\sigma^2_{\alpha_k}$, $k = 1, \ldots, p$ given the data and the rest of the parameters are all conditionally conjugate.

(iii) The conditional posterior distribution for $\sigma^2_{\epsilon_i}$ has the form

$$f(\sigma^2_{\epsilon_i} \,|\, \mathbf{y}, \text{rest}) = \mathrm{IGa}\left(\sigma^2_{\epsilon_i} \,\middle|\, c^\epsilon_0 + \frac{T}{2}\,,\; c^\epsilon_1 + \frac{1}{2}\mathbf{M}'_i\mathbf{M}_i\right),$$

where $\mathbf{M}_i = \mathbf{y}_i - \mathbf{Z}\boldsymbol{\alpha}_i - \mathbf{X}\boldsymbol{\beta}_i - \boldsymbol{\theta}_i$, for $i = 1, \ldots, n$.

(iv) The conditional posterior distribution for $\sigma^2_{\alpha_j}$ has the form

$$f(\sigma^2_{\alpha_j} \,|\, \mathbf{y}, \text{rest}) = \mathrm{IGa}\left(\sigma^2_{\alpha_j} \,\middle|\, c^\alpha_0 + \frac{n}{2},\;\; c^\alpha_1 + \frac{1}{2}\sum_{i=1}^{n} \alpha^2_{ij}\right),$$

for $j = 1, 2, \ldots, p$.

For $m \leq n$ we denote by $(\boldsymbol{\gamma}^*_1, \ldots, \boldsymbol{\gamma}^*_m)$ the set of unique values in $\boldsymbol{\gamma} = (\boldsymbol{\gamma}_1, \ldots, \boldsymbol{\gamma}_n)$, accordingly $\boldsymbol{\gamma}^*_j = (\boldsymbol{\beta}^*_j, \boldsymbol{\theta}^*_j)$, $j = 1, \ldots, m$, so that

(v) The conditional posterior distribution for $\sigma^2_{\beta_k}$ has the form

$$f(\sigma^2_{\beta_k} \,|\, \mathbf{y}, \text{rest}) = \text{IGa}\left(\sigma^2_{\beta_k} \,\middle|\, c_0^{\beta} + \frac{m}{2}\,,\, c_1^{\beta} + \frac{1}{2}\sum_{j=1}^{m}(\beta^*_{jk})^2\right),$$

for $k = 1, 2, \ldots, d$. For this conditional posterior distribution, $\beta^*_{j,k}$ is the k-th component in $\boldsymbol{\beta}^*_j$, $k = 1, 2, \ldots, d$.

For obtaining the conditional posterior distribution of the hyper parameters σ^2_θ and ρ, we note that their likelihood is given by the joint prior distribution of the $\boldsymbol{\gamma}_i$'s, which is given by $\text{lik}(\sigma^2_\theta, \rho \,|\, \boldsymbol{\gamma}) \propto \prod_{j=1}^{m} \text{N}_d(\boldsymbol{\beta}^*_j \,|\, \mathbf{0}, \boldsymbol{\Sigma}_\beta)\text{N}_T(\boldsymbol{\theta}^*_j \,|\, \mathbf{0}, \mathbf{R})$, where $\mathbf{R} = \sigma^2_\theta \mathbf{P}$ and $\mathbf{P}$ is the $T \times T$ dimensional matrix with elements $P_{ij} = \rho^{|i-j|}$, for $i, j = 1, \ldots, T$. This likelihood depends only on the distinct pairs $\boldsymbol{\gamma}^*_j = (\boldsymbol{\beta}^*_j, \boldsymbol{\theta}^*_j)$, $j = 1, \ldots, m$ in $\boldsymbol{\gamma} = (\boldsymbol{\gamma}_1, \ldots, \boldsymbol{\gamma}_n)$, with $m \leq n$. Therefore, combining this with the reference prior (8),

(vi) the conditional posterior distribution for σ^2_θ is proper as long as $m \geq 1$, which is true if $n \geq 1$, and is given by

$$f(\sigma^2_\theta \,|\, \mathbf{y}, \text{rest}) = \text{IGa}\left(\sigma^2_\theta \,\middle|\, \frac{mT}{2}\,,\, \frac{1}{2}\sum_{j=1}^{m}(\boldsymbol{\theta}^*_j)'\mathbf{P}^{-1}\boldsymbol{\theta}^*_j\right), \quad \text{and}$$

(vii) the conditional posterior distribution of ρ becomes

$$f(\rho \,|\, \mathbf{y}, \text{rest}) \propto |\mathbf{P}|^{-m/2}\exp\left\{-\frac{1}{2\sigma^2_\theta}\sum_{j=1}^{m}(\boldsymbol{\theta}^*_j)'\mathbf{P}^{-1}\boldsymbol{\theta}^*_j\right\}\frac{\sqrt{1+\rho^2}}{1-\rho^2},$$

for $\rho \in (-1, 1)$.

For the parameters (a, b), their prior distribution is updated with the EPPF (exchangeable partition probability function), induced by the Poisson-Dirichlet process (Pitman 1995), which acts as a likelihood and is given by

$$f(n^*_1, \ldots, n^*_m | a, b) = \frac{\Gamma(b+1)}{\Gamma(b+n)}\left\{\prod_{j=1}^{m-1}(b+ja)\right\}\left\{\prod_{j=1}^{m}\frac{\Gamma(n^*_j - a)}{\Gamma(1-a)}\right\}.$$

(viii) The conditional posterior distribution for a is

$$f(a|b, \text{rest}) = \left\{\prod_{j=1}^{m-1}(b+ja)\right\}\left\{\prod_{j=1}^{m}\frac{\Gamma(n^*_j - a)}{\Gamma(1-a)}\right\}f(a),$$

for $a \in [\max\{-b, 0\}, 1)$, and $f(a)$ given in (9).

(ix) The conditional posterior distribution for b becomes

$$f(b|a,\text{rest}) = \frac{\Gamma(b+1)}{\Gamma(b+n)}\left\{\prod_{j=1}^{m-1}(b+ja)\right\} f(b|a),$$

for $b > -a$, and $f(b|a)$ given in (10).

Posterior inference can be done by obtaining posterior draws from the marginal posterior distribution for $\boldsymbol{\alpha}_i$, as in (i), together with a Gibbs sampler (Smith and Roberts 1993) with the full conditional distributions (ii)–(ix). With the exception of (vii), (viii) and (ix) which require Metropolis within Gibbs steps (Tierney 1994), the rest of the conditional distributions are of standard form and so can be sampled directly. As noted by Jara et al. (2010), sampling from (viii) requires special attention. Since the prior for a, as in (9), is a mixture of a point mass and a continuous distribution on $(0,1)$, the Metropolis-Hastings proposal must define an irreducible chain. For that we suggest taking proposal draws independently from a mixture distribution of the form $f(a) = 0.5I_{\{0\}}(a) + 0.5\text{Be}(a|1,1)$.

When dealing with Dirichlet process mixture models, which are particular cases of our Poisson-Dirichlet process model, MacEachern (1994) noticed that a "sticky clusters" effect appears when sampling from the nonparametric components, which in our case are the $\boldsymbol{\gamma}$ parameters. To overcome this problem it was suggested to introduce an acceleration step to improve the chain mixing. This step consists of resampling the unique $\boldsymbol{\gamma}_i$'s values $\boldsymbol{\gamma}_j^*$, $j = 1,\ldots,m$. The corresponding conditional posterior distribution, conditional on the cluster configuration (c.c.) $I_j = \{i : \boldsymbol{\gamma}_i = \boldsymbol{\gamma}_j^*\}$ is given by

$$f(\boldsymbol{\gamma}_j^* \,|\, \mathbf{y}, \text{c.c., rest}) \propto \left\{\prod_{i\in I_j} \text{N}_T(\mathbf{y}_i \,|\, \mathbf{Z}\boldsymbol{\alpha}_i + \mathbf{X}\boldsymbol{\beta}_i + \boldsymbol{\theta}_i,\, \sigma_{\epsilon_i}^2\mathbf{I})\right\} \times g_0(\boldsymbol{\gamma}_j^*).$$

Again, sampling from this distribution is easier by sampling from $\boldsymbol{\beta}_j^*$ and $\boldsymbol{\theta}_j^*$ separately.

(x) The corresponding full conditional for $\boldsymbol{\theta}_j^*$ is given by

$$f(\boldsymbol{\theta}_j^* \,|\, \mathbf{y}, \boldsymbol{\beta}_j^*, \text{rest}) = \text{N}_T(\boldsymbol{\theta}_j^* \,|\, \boldsymbol{\mu}_\theta^*,\, \mathbf{S}_\theta^*),$$

where $\mathbf{S}_\theta^* = (\sum_{i\in I_j}(\sigma_{\epsilon_i}^2\mathbf{I})^{-1} + \mathbf{R}^{-1})^{-1}$ and $\boldsymbol{\mu}_\theta^* = \mathbf{S}_\theta^* \sum_{i\in I_j}(\sigma_{\epsilon_i}^2\mathbf{I})^{-1}(\mathbf{y}_i - \mathbf{Z}\boldsymbol{\alpha}_i - \mathbf{X}\boldsymbol{\beta}_j^*)$.

(xi) The corresponding full conditional for $\boldsymbol{\beta}_j^*$ is

$$f(\boldsymbol{\beta}_j^* \,|\, \mathbf{y}, \boldsymbol{\theta}_j^*, \text{rest}) = \text{N}_d(\boldsymbol{\beta}_j^* \,|\, \boldsymbol{\mu}_\beta^*, \mathbf{S}_\beta^*),$$

where $\mathbf{S}_\beta^* = (\mathbf{X}'\sum_{i\in I_j}(\sigma_{\epsilon_i}^2\mathbf{I})^{-1}\mathbf{X} + \boldsymbol{\Sigma}_\beta^{-1})^{-1}$ and $\boldsymbol{\mu}_\beta^* = \mathbf{S}_\beta^*\mathbf{X}'\sum_{i\in I_j}(\sigma_{\epsilon_i}^2\mathbf{I})^{-1}(\mathbf{y}_i - \mathbf{Z}\boldsymbol{\alpha}_i - \boldsymbol{\theta}_j^*)$.

Including this acceleration step to sample from (x) and (xi) is straightforward since both distributions are standard multivariate normals. Neal (2000) discusses different algorithms for sampling from Dirichlet process mixture models which are also applicable to more general processes. Ours would correspond to Neal's algorithm 2. This computational algorithm was implemented in Fortran and is available upon request.

4 Clustering selection and fitting measures

As mentioned before, posterior inference for our model is obtained by implementing a Gibbs sampler. When convergence is attained, the posterior samples of the parameters can be used to determine a clustering structure for the data set $(\mathbf{y}_1, \ldots, \mathbf{y}_n)$. At each iteration, the Gibbs sampler produces an implicit clustering of the parameters $\boldsymbol{\gamma} = (\boldsymbol{\gamma}_1, \boldsymbol{\gamma}_2, \ldots \boldsymbol{\gamma}_n)$, for which each $\boldsymbol{\gamma}_i$ (partially) characterizes the time series $\mathbf{y}_i$, thus inducing a clustering of the time series $\mathbf{y}_i$'s. To avoid the label switching problem, that naturally arises when using mixture models (Stephens 2000), we summarize the clustering information not by registering the cluster membership, but by counting the number of times (iterations) that two parameters, say $\boldsymbol{\gamma}_i$ and $\boldsymbol{\gamma}_j$, belong to the same cluster.

With this information we build a similarity matrix containing the relative frequencies (relative to the number of posterior samples) of pairwise clustering corresponding to the event that $\mathbf{y}_i$ and $\mathbf{y}_j$ share the same $\boldsymbol{\gamma}$ parameter values, that is $\boldsymbol{\gamma}_i = \boldsymbol{\gamma}_j$. Each cell, (i, j), of this pairwise clustering matrix can be interpreted as the probability of two time series $\mathbf{y}_i$ and $\mathbf{y}_j$ belonging to the same cluster. The problem now is how to determine a single clustering structure based on this similarity matrix. Medvedovic and Sivaganesan (2002), for example, use the pairwise clustering matrix as an input of a (classical) hierarchical clustering procedure, and with a selection of an appropriate link function, this produces a dendrogram from which a single clustering can be chosen. Alternatively, Argiento et al. (2013) define the similarity matrix based on distances between the corresponding parameter densities, thereby inducing a coarser clustering structure.

On the other hand, Dahl (2006) criticizes the previous selection criteria by arguing that when following a model based approach, as in our case, the model itself produces a series of clusters (one at each Gibbs sampler iteration), so why not select one iteration as a representative clustering structure. He therefore suggests choosing the cluster (iteration) that minimizes the square deviations with respect to the pairwise clustering matrix. Here we follow Dahl's approach.

To compare among the clusters obtained by different prior specifications, we summarize the heterogeneity of a clustering by considering the heterogeneity measure (HM). If $G_1, \ldots, G_m$ denote the sets of indices for a clustering of m clusters with sizes $n_1, \ldots, n_m$ then

$$\mathrm{HM}(G_1, \ldots, G_m) = \sum_{k=1}^{m} \frac{2}{n_k - 1} \sum_{i<j \in G_k} \sum_{t=1}^{T} (y_{it} - y_{jt})^2.$$

The larger the value of HM the more heterogeneous a clustering is. These values should be compared with care across different m's since in the extreme case that each series forms its own cluster then HM takes the value of zero. So it is preferably a clustering with small HM and small m.

Additionally we assess model fit by computing the logarithm of the pseudo marginal likelihood (LPML), which is a predictive measure for model performance. This measure is based on the conditional predictive ordinate (CPO) statistics (Geisser and Eddy

1979). Given posterior samples $\boldsymbol{\alpha}_i^{(l)}, \boldsymbol{\beta}_i^{(l)}, \boldsymbol{\theta}_i^{(l)}, \sigma_{\epsilon_i}^{(l)\,2}$, $l = 1, \ldots, L$, from $\boldsymbol{\alpha}_i, \boldsymbol{\gamma}_i = (\boldsymbol{\beta}_i, \boldsymbol{\theta}_i)$ and $\sigma_{\epsilon_i}^2$, a Monte Carlo estimate $\widehat{\text{CPO}}_i$ for CPO_i, $i = 1, \ldots, n$, is obtained as

$$\widehat{\text{CPO}}_i = \left(\frac{1}{L} \sum_{l=1}^{L} \frac{1}{f(\mathbf{y}_i | \boldsymbol{\alpha}^{(l)}, \boldsymbol{\gamma}^{(l)}, \boldsymbol{\sigma}_\epsilon^{(l)})} \right)^{-1}.$$

As suggested by Mukhopadhyay and Gelfand (1997), the CPO is computationally more stable if we evaluate the conditional density in terms of the whole mixture as

$$f(\mathbf{y}_0 | \boldsymbol{\alpha}^{(l)}, \boldsymbol{\gamma}^{(l)}, \boldsymbol{\sigma}_\epsilon^{(l)}) = \sum_{j=1}^{m} \frac{n_j^* - a}{b+n} \text{N}(\mathbf{y}_0 | \mathbf{Z}\boldsymbol{\alpha}_0^{(l)} + \mathbf{X}\boldsymbol{\beta}_j^{*(l)} + \boldsymbol{\theta}_j^{*(l)}, \sigma_{\epsilon_0}^{2(l)} \mathbf{I})$$

$$+ \frac{b + am}{b+n} \text{N}(\mathbf{y}_0 | \mathbf{Z}\boldsymbol{\alpha}_0^{(l)}, \mathbf{W}_0).$$

Alternatively, this conditional density can be computed by evaluating in the corresponding mixture component. Although this latter is computationally simpler, for the Mexican stock exchange data to be analysed in Section 5, longer chains are required to obtain the same values as those obtained with Mukhopadhyay and Gelfand (1997)'s approach.

Finally, these values are summarized to define

$$\widehat{\text{LPML}} = \sum_{i=1}^{n} \log(\widehat{\text{CPO}}_i).$$

Larger values of LPML indicate better fit.

5 Application: Mexican stock exchange data

In this section we apply the proposed method to our motivating data. The objective is to produce clusters of companies listed in the Mexican stock exchange. We note that one company can have more than one type of share, but for the purpose of this analysis those will be considered as different companies. The information consists of monthly adjusted closing share prices of $n = 58$ companies, available from September 2006 to August 2011. That is, the length of the time series is $T = 60$ months. This information was obtained from the Factiva database which is part of the Dow Jones News Corporation (http://www.dowjones.com/factiva/).

The observed share prices take values in different scales. Producing a cluster using the original values would result in clustering only those series with similar observed scales, leaving apart some series with similar patterns but different scales. Aiming to produce a more objective clustering, we work with the same scale by linearly transforming the data so that each series takes values in the interval $(0, 1)$. The 58 scaled series are presented in Figure 1. From this figure we observe that the scaled series do show different patterns (tendencies and periodicities) and therefore clustering them is a challenge. It

is worth mentioning that in the financial literature (e.g. Campbell et al. 1997) a usual transformation of the data leading to a scale-free representation is given by computing the returns or log-returns. However, this operation would eliminate structures in the data, like trends, and then we would not be able to use such structures for clustering, as our proposal suggests.

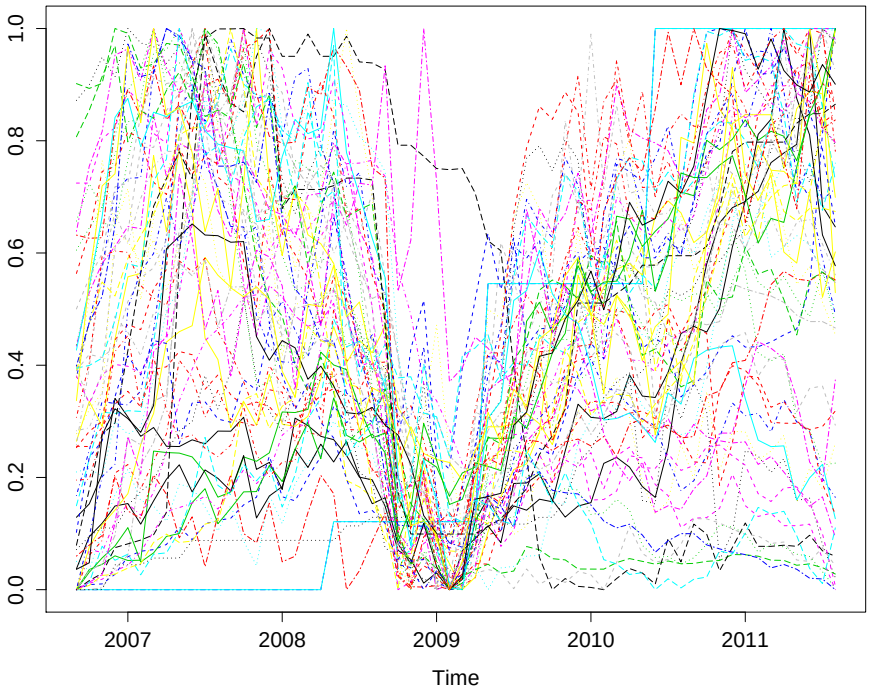

Figure 1: Time series plot of the (scaled) share prices of 58 companies listed in the Mexican stock exchange.

We first attempted to produce a naive clustering by computing the Pearson's correlation matrix of the series and using it as a similarity matrix in a hierarchical clustering. The complete linkage agglomerative clustering procedure produces the dendrogram shown in Figure 2. Additionally, Figure 2 includes a heatmap of the correlation matrix. Apart from two somehow homogeneous clusters (darker well formed squares in the center of the heatmap) of 18 and 11 companies respectively, the rest of the clusters seem to be quite heterogeneous. Another characteristic that can be derived from this figure is that the dendrogram suggests up to 6 "clear" clusters in the data.

We now implemented our clustering proposal described in Section 2. As is well known in model based clustering with Bayesian nonparametric mixtures, (e.g. Barrios et al. 2013),

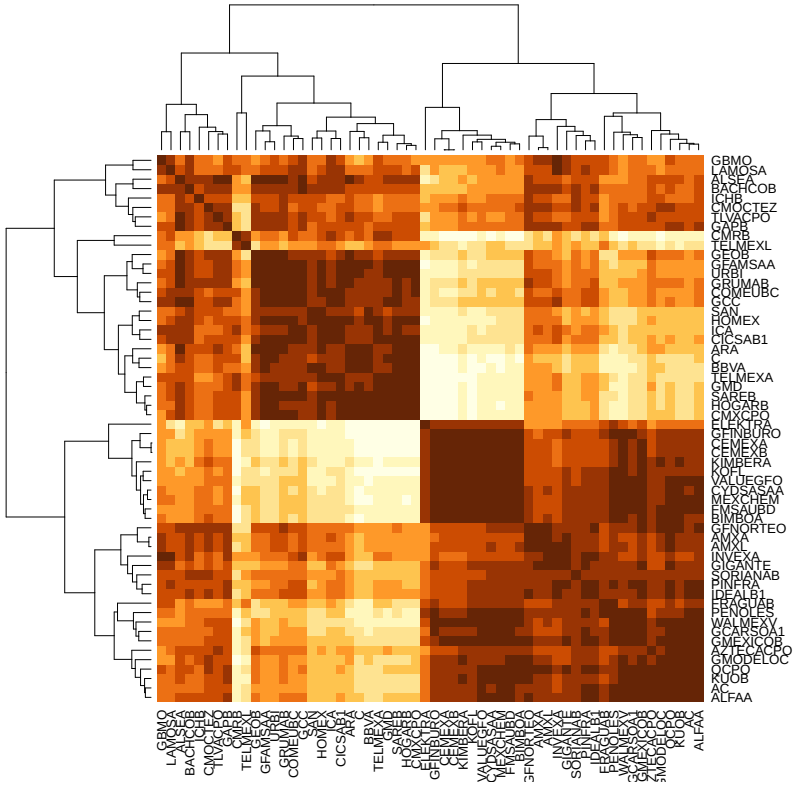

Figure 2: Mexican stock exchange companies. Heatmap and dendrogram for complete linkage hierarchical clustering. Correlation matrix was used as similarity matrix.

the prior choice of the variance $\sigma^2_{\epsilon_i}$ is crucial for determining an appropriate clustering. Slightly informative priors sometimes produce better clusterings. On the other hand, Gelman (2006) points out the importance of the prior distribution on the variance parameters of a hierarchical model, which in our case are $\sigma^2_{\beta_j}$ and $\sigma^2_{\alpha_k}$. We therefore consider two sets of values for the hyper parameters of these priors, say, $(c_0^k, c_1^k) \in \{(0.01, 0.01), (2, 1)\}$, for $k = \epsilon, \beta, \alpha$. This implies that the inverse gamma priors have infinite mean and variance in the first choice, and mean 1 with infinite variance in the second choice. For the specification of the Poisson-Dirichlet process parameters a and b we considered three options: $a = 0$ and $(q_0^b, q_1^b) = (1, 1)$ to define a Dirichlet process; $(q_0^a, q_1^a, \pi) = (1, 1, 0.5)$ and $b = 0$ to define a normalized stable process; and $(q_0^a, q_1^a, \pi) = (1, 1, 0.5)$ together with $(q_0^b, q_1^b) = (1, 1)$ to define a Poisson-Dirichlet (non Dirichlet, nor normalized stable) process.

Recall that the parameters of the model $\boldsymbol{\eta}_i$ for individual i, are divided into three blocks, $\boldsymbol{\alpha}_i$ of dimension p, $\boldsymbol{\beta}_i$ of dimension d, and $\boldsymbol{\theta}_i$ of dimension T, and only the last two blocks, $\boldsymbol{\gamma}_i = (\boldsymbol{\beta}_i, \boldsymbol{\theta}_i)$, are used for clustering purposes. We consider two different model

specifications and thus different sets of explanatory variables. The first set contains level, linear trend, monthly seasonal components, and temporal components, that is, $\mathbf{E}(y_{it}) = \mu_i + \omega_{1i}t + \sum_{j=2}^{12} \upsilon_{j,i} m_j(t) + \theta_{it}$, with $m_j(t)$ the month indicator as described in Section 1. In this scenario we have a total of $p + d = 13$ parameters plus the T temporal components for each $\mathbf{y}_i$. We consider two cases for clustering, everything but the level ($p = 1$), and everything but the level and linear trend ($p = 2$). The second scenario adds to the first scenario a quadratic trend term, $\omega_{2i}t^2$, thus having a total of $p + d = 14$ parameters. We also try with different possibilities varying $p \in \{1, 2, 3\}$.

With the previous model specifications we carried out posterior inference by implementing a Gibbs sampler with 10000 iterations, 1000 as burn-in period, and keep one of every 5th iteration to reduce the autocorrelation of the chain. Convergence of the chain was assessed informally by looking at ergodic mean plots of the baseline parameters. Running time for the 10000 iterations varies from 20 to 40 minutes according to the prior specifications. In particular, the choice $(c_0^k, c_1^k) = (0.01, 0.01)$ makes the algorithm run slower. Table 1 summarizes the goodness of fit statistic, LPML, as well as the heterogeneity measure, HM, of the optimal clustering, $\hat{m}$, obtained with Dahl (2006)'s procedure.

Table 1: Mexican stock exchange dataset. Logarithm of the pseudo marginal likelihood (LPML) statistic, clustering heterogeneity measure (HM), and optimal number of clusters ($\hat{m}$), for different prior selections of (c_0^k, c_1^k) for $k = \epsilon, \beta, \alpha$, (p, d) and $\mathcal{PD}(a, b)$ prior processes.

		$(c_0^k, c_1^k) = (0.01, 0.01)$			$(c_0^k, c_1^k) = (2, 1)$		
(p, d)	Model	LPML	HM	$\hat{m}$	LPML	HM	$\hat{m}$
$(1, 12)$	Dir	1623.18	63.90	24	674.41	183.37	4
$(1, 12)$	Nstable	1584.99	74.17	22	632.23	183.48	4
$(1, 12)$	Po-Dir	1638.01	83.59	21	671.85	183.48	4
$(2, 11)$	Dir	2143.34	233.08	14	801.95	400.72	4
$(2, 11)$	Nstable	2110.66	220.35	15	847.56	392.40	5
$(2, 11)$	Po-Dir	2066.61	205.56	16	823.84	392.40	5
$(1, 13)$	Dir	1819.27	109.56	15	702.04	179.86	4
$(1, 13)$	Nstable	1705.70	113.04	16	663.02	168.88	5
$(1, 13)$	Po-Dir	1741.20	110.56	15	708.63	179.86	4
$(2, 12)$	Dir	2231.01	233.08	14	926.40	400.72	4
$(2, 12)$	Nstable	2205.38	244.37	12	930.57	400.72	4
$(2, 12)$	Po-Dir	2255.87	233.08	14	912.10	400.72	4
$(3, 11)$	Dir	2476.69	247.47	13	917.74	372.66	6
$(3, 11)$	Nstable	2478.74	217.97	15	938.59	350.16	8
$(3, 11)$	Po-Dir	2408.26	247.47	13	927.81	350.16	8

Several conclusions can be derived from Table 1. A better fit is achieved when the number of parameters p in $\boldsymbol{\alpha}$, not used for clustering, is larger ($p \geq 2$). This makes sense since the parameters $\boldsymbol{\alpha}_i$ are not bound to be tight among individuals, allowing them to take the best possible value for each individual i producing a better fit. In most

of the cases, the fitting is slightly better (larger LPML) for the Dirichlet case compared with the other two cases. However, looking at the heterogeneity measure HM, those cases that produce a better fit ($p > 1$) also produce the more heterogeneous clusters. In fact, when examining the clustering structure induced by these heterogeneous cases, they form one big cluster with most of the time series and many singleton clusters. That is, once the level, the linear ($p = 2$) (and quadratic, $p = 3$) trend are removed, most of the series follow the same seasonal and temporal effects.

On the other hand, the most homogeneous clusters but with bad fit are produced when the clustering is produced with everything else but the level of the series ($p = 1$). Comparing the scenarios with ($d = 13$) and without ($d = 12$) quadratic trend in the clustering part, the fitting is slightly better when quadratic trend is considered in the model.

Now, comparing the prior variances selection (c_0^k, c_1^k), across the columns of Table 1, we notice a huge difference both in the fitting and heterogeneity measures. The fitting is dramatically better when choosing $(0.01, 0.01)$. Additionally, the heterogeneity of the clustering structure is a lot smaller with this same choice. However, the reduction in heterogeneity is due to an increment in the number of groups ($\hat{m}$). The smallest HM measure achieved with $(0.01, 0.01)$ produces 24 clusters; in contrast, the smallest HM measure with $(2, 1)$ produces 5 clusters. To determine the best clustering we need to find a balance between the heterogeneity and the number of groups.

Studying with more detail the different clustering structures produced, we notice that the clustering with HM = 63.90 and 24 groups only has 9 groups (less than 40% of the groups) with more than one company, that is, 15 groups are singletons. We show these 9 groups in Figure 3. The groups are well formed, and apparently different from each other. An intermediate clustering, with a smaller number of total groups, is that with an HM = 109.56 and 15 groups. Here, 7 groups (less than 50% of the groups) have more than one company and the remaining 8 groups are singletons. These 7 groups are shown in Figure 4. Visually, these 7 significant groups look somehow homogeneous. Since we now have 8 singletons, at least 7 of the 15 singletons of the previous clustering must have been assigned to another group.

Most of the clusterings obtained with $(c_0^k, c_1^k) = (2, 1)$ have 4 or 5 groups. However, they have different values of HM. Note that all those clusterings with the same HM correspond to the same clustering structure. To understand the different clusterings obtained, we concentrate on the clusterings with the four lowest HM values (last column in Table 1). Three of them have 4 clusters and one has 5. The clustering sizes are C1={22,20,15,1}, C2={21,20,16,1}, C3={21,19,16,2} and C4={21,19,16,1,1} with HM 183.48, 183.37, 179.86 and 168.88, respectively.

Clusterings C1 and C2 differ by one allocation, series S7 = AZTECACPO, which is allocated in group 1 of C1 and in group 3 of C2. To better appreciate the two different allocations, Figure 5 graphically represents series S7 in these two groups. From the graph it is perfectly understandable why the two model specifications have problems allocating this series. We prefer S7 to be allocated in group 3 of C2 since C2 achieves a slightly smaller HM. Now clusterings C2 and C3 differ by allocating series S17 =

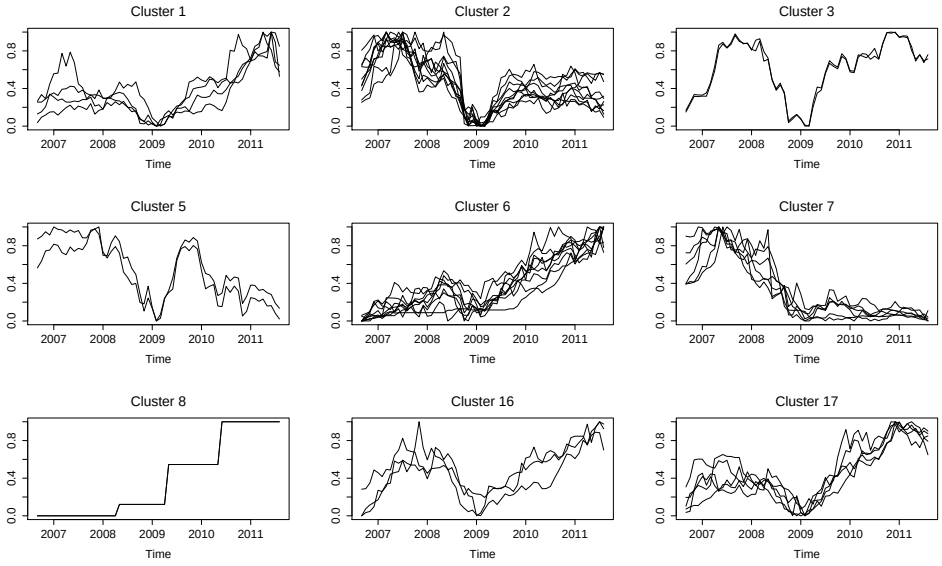

Figure 3: Mexican stock exchange companies. Clustering with 24 groups. Shown are 9 groups with more than one company.

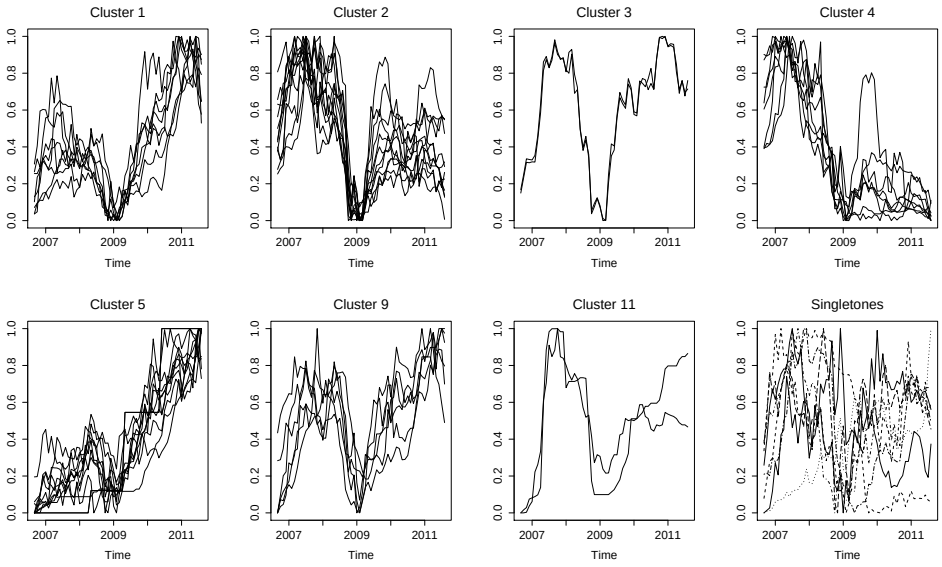

Figure 4: Mexican stock exchange companies. Clustering with 15 groups.

CMRB in groups 2 and 4 respectively. Figure 6 graphically shows series S17 in these two groups. It is not clear from the graph that S17 should belong to any of the two groups. Moreover, clustering C4 differs from clusterings C2 and C3 in allocating series S17 to its own group. In fact C4 leaves the two series shown in the right panel of Figure 6, S17 and S54 = TELMEXL, allocated into two separate groups. We suggest taking C4 as the final clustering. The 58 series divided into the final 5 groups of C4 are presented in Figure 7. It is remarkable how homogeneous these final groups look.

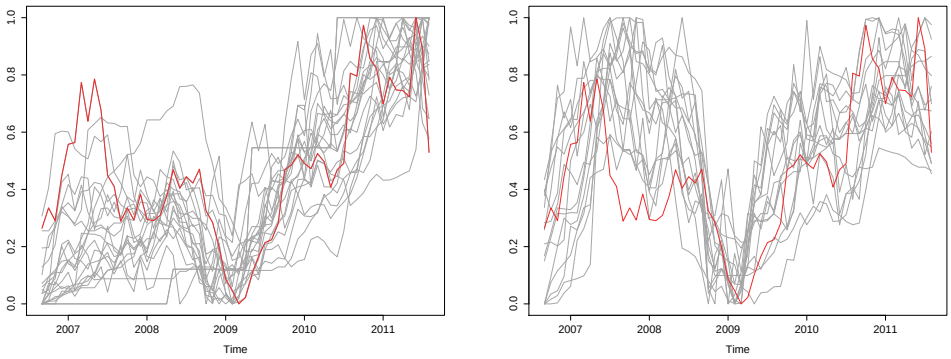

Figure 5: Series S7=AZTECACPO in groups 1 of C1 (left) and group 3 of C2 (right).

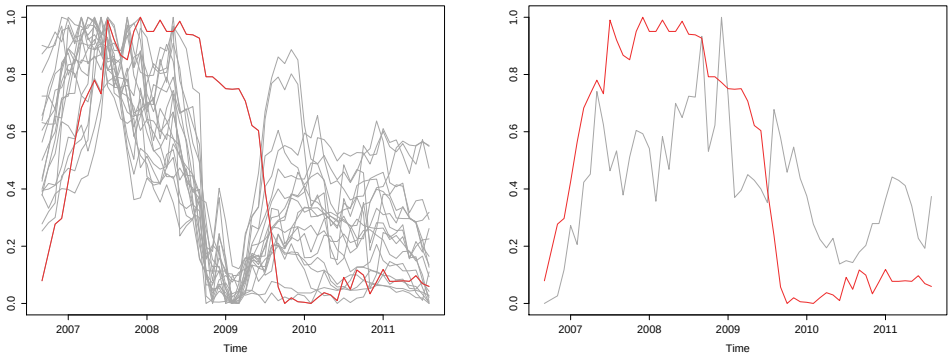

Figure 6: Series S17=CMRB in groups 2 of C2 (left) and group 4 of C3 (right).

Clustering C4 was produced by only one model specification. A normalized stable process ($b = 0$) with $(q_0^a, q_1^a, \pi) = (1, 1, 0.5)$, $(c_0^k, c_1^k) = (2, 1)$ and $(p, d) = (1, 13)$. In order to assess the clarity of the final clustering selected, we present a heat map of the relative frequencies matrix of pairwise clustering in Figure 8. The three big squares correspond to the large groups with 21, 16 and 19 series (companies), from bottom right to top left, respectively. The two separate dark dots in the upper left corner correspond

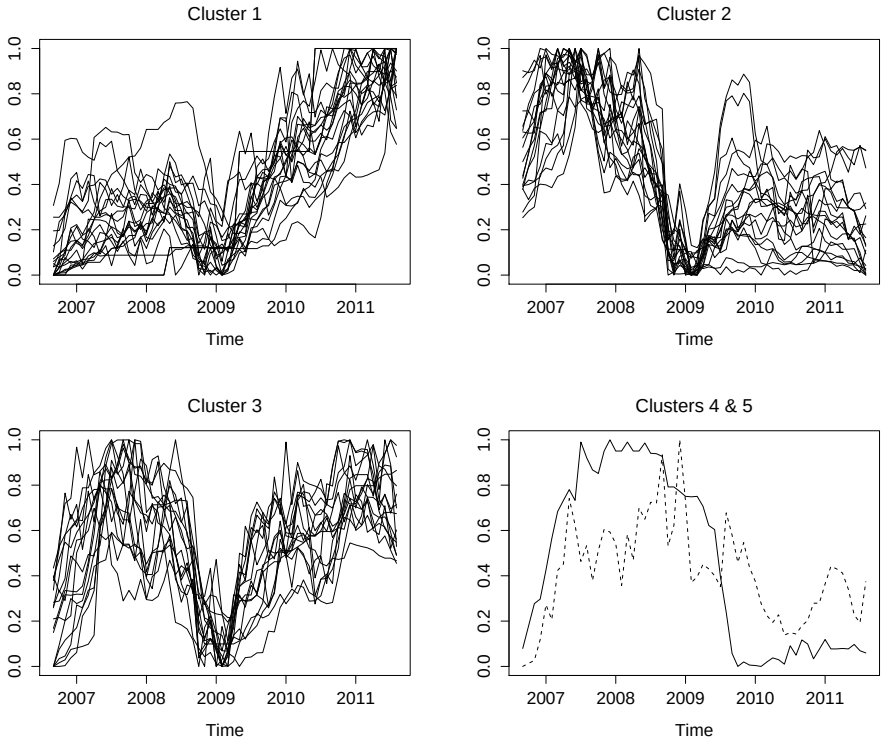

Figure 7: Mexican stock exchange companies. Final clustering with 5 groups.

to those two series that form singleton groups. Comparing this heat map with that produced by the Pearson correlation matrix (Figure 2) makes clear the advantage of using our model based clustering procedure proposed in this paper.

Finally, we pursue an interpretation of the final clusters formed, in the context of the Mexican economy. It is worth noticing that the behaviour of all 58 series is marked by the 2008 world crisis. This crisis started in the USA in September and spread out to the rest of the world afterwards. This effect can be appreciated in Figure 1 where the majority of the series drop down at the end of the year 2008.

The companies in clusters 1,2, and 3 (Figure 7) drop down close to the end of 2008 and beginning of 2009, whereas for clusters 4 and 5 a decay period starts at the end of 2009. We can see that the way in which our clustering methodology forms the groups agrees with the way in which the series behave before and after the crisis period. Cluster 1 is mainly formed by companies which are engaged in the production and marketing of fast moving consumer goods (groceries, baked goods, sodas) as well as some companies devoted to the exploration and exploitation of mineral and metal fields. The upper left panel in Figure 7 shows that companies in cluster 1 have a constant rate

Figure 8: Mexican stock exchange companies. Heatmap and dendrogram for complete linkage hierarchical clustering. Matrix of relative frequencies of pairwise clustering was used as similarity matrix. Model specifications are: $(p, d) = (1, 13)$, $b = 0$, $(q_0^a, q_1^a, \pi) = (1, 1, 0.5)$ and $(c_0^k, c_0^k) = (2, 1)$.

of recovery. This is not the case for the other clusters. The upper right panel of the same Figure shows cluster 2 series. This cluster is formed by a number of companies engaged in construction, i.e., residential housing industries, as well as some banking institutions. These series do not show as clear a recovery as those series in cluster 1. We believe, however, that the companies in this cluster 2 are a more important reflection of the Mexican economy, since economy growth is mostly linked to the development of infrastructure and construction as well as banking.

We continue with cluster 3, shown at the lower left panel in Figure 7. This is formed by companies engaged in telecommunications and broadcasting. After the beginning of 2009, the rate of recovery of these companies is somewhere in between the rate of recovery for the two previous clusters. The fact that cluster 1 features a higher growing rate than clusters 2 and 3 is not surprising since most of the Mexican population consumes fast moving consumer goods on a regular basis, and also mining is still a

profitable activity in Mexico. On the other hand it is known that, since the financial crisis started, the development of infrastructure and construction is not growing well in Mexico, a message that tells us that Mexican economy is indeed affected.

Lastly, the companies in cluster 4, operation of restaurants (CMRB), and cluster 5, telecommunications (TELMEX), started dropping down a bit later in time than companies from the other clusters. These last two series correspond to what used to be strong companies in Mexico, at least before the crisis, which explains a delay in their decay after the crisis.

6 Concluding remarks

Clustering time series has several practical uses and is not a simple task. In this article we address the problem by proposing a model based clustering procedure that relies on a Bayesian semiparametric mixture model centred in a state-space model. The model allows for selecting different features of the series for clustering purposes.

We assign to the coefficients of a linear predictor and to a dynamic component a Poisson-Dirichlet process prior. The advantage of using an almost surely discrete nonparametric prior, as the Poisson-Dirichlet process, is the fact that the coefficients naturally cluster into groups of the same value. This, in turn, is used to cluster the observed time series.

For the particular application studied in this article, the Dirichlet process choice ($a = 0$) mostly achieved a better fit to the data. On the other hand, the normalized stable process specification ($b = 0$) produced the final clustering that we chose. Other studies with species sampling models (Lijoi et al. 2007), where the Dirichlet and normalized stable processes are particular cases, suggest that the normalized stable specification, $b = 0$ with a close to 1 in the Poisson-Dirichlet process, produces a clustering structure with a larger number of groups, compared with that of a Dirichlet process, whose size tends to be small. We therefore advise considering several prior specifications, as the ones considered here, in order to find the best clustering structure.

The main objective of this article was to produce a clustering of time series in terms of a selection of simple features such as as trends, seasonality and temporal components. Our method disregarded the explanatory power of the observations. Alternative models can be proposed to achieve a dual objective: a good clustering and good explanatory power. For this purpose, some of the generalizations of the linear dynamic model, discussed in the introduction, could be used. We anticipate that a complicated compromise needs to be tackled. Having a complicated model with good fitting properties, able to explain all the different characteristics in a set of time series, might have the problem of being so good that the clustering induced would be formed by all singletons. Anyway it might be worth trying.

Acknowledgments

The first author acknowledges support to grant I130991-F from the National Council for Science and Technology of Mexico (CONACYT).

References

Argiento, R., Cremaschi, A. and Guglielmi, A. (2013). A Bayesian nonparametric mixture model for cluster analysis. Technical report *Quaderno Imati CNR, 2012 3-MI*, Milano. ISSN 1722-8964. 156

Barrios, E., Lijoi, A., Nieto-Barajas, L.E. and Prünster, I. (2013). Modeling with normalized random measure mixture models. *Statistical Science*. To appear. 158

Campbell, J.Y., Lo A.W. and MacKinlay, A.C. (1997). *The econometrics of financial markets*. Princeton University Press, Princeton, New Jersey. 158

Carlin, B.P., Polson N.G. and Stoffer, D.S. (1992). A Monte Carlo Approach to Nonnormal and Nonlinear State-Space Modeling. *Journal of the American Statistical Association* **87**, 493-500. 148

Caron, F., Davy, M., Doucet, A., Duflos, E. and Vanheeghe, P. (2008). Bayesian Inference for Linear Dynamic Models with Dirichlet Process Mixtures. *IEEE Transactions on Signal Processing* **56**, 71-84. 148

Carter, C.K. and Kohn, R. (1994). On Gibbs Sampling for State-Space Models. *Biometrika* **81**, 541-553. 148

Carter, C.K. and Kohn, R. (1996). Markov Chain Monte Carlo in Conditionally Gaussian State-Space Models. *Biometrika* **83**, 589-601. 148

Chatfield, C. (1989). *The analysis of time series: an introduction*. Chapman and Hall, London. 147, 149

Chib, S. and Greenberg, E. (1996). Markov Chain Monte Carlo Simulation Methods in Econometrics. *Econometric Theory* **12**, 409-431. 148

Dahl, D.B. (2006). Model based clustering for expression data via a Dirichlet process mixture model. In *Bayesian Inference for Gene Expression and Proteomics*, Eds. M. Vanucci, K.-A. Do and P. Müller. Cambridge University Press, Cambridge. 156, 160

Escobar, M.D. and West, M. (1998). Computing nonparametric hierarchical models. In *Practical Nonparametric and Semiparametric Bayesian Statistics*, Eds. D. Dey, P. Müller and Sinha, D. Springer, New-York. 153

Ferguson, T.S. (1973). A Bayesian analysis of some nonparametric problems. *Annals of Statistics* **1**, 209–230. 150

Fox, E., Sudderth, E.B., Jordan, M.I. and Willsky, A.S. (2011). Bayesian Nonparametric Inference of Switching Dynamic Linear Models. *IEEE Transactions on Signal Processing* **59**, 1569-1585. 148

Geisser, S. and Eddy, W.F. (1979). A predictive approach to model selection. *Journal of the American Statistical Association* **74**, 153–160. 156

Gelman, A. (2006). Prior distributions for variance parameters in hierarchical models. *Bayesian Analysis* **1**, 515-533. 159

Ghosh, A., Mukhopadhyay., S., Roy, S. and Bhattacharya, S. (2012). Bayesian Inference in Nonparametric Dynamic State-Space Models. arXiv:1108.3262[stat.ME]. 148

Granger, C.W.J. and Newbold, P. (1974). Spurious regressions in econometrics. *Journal of Econometrics* **2**, 111–120. 147

Harrison, P.J. and Stevens, P.F. (1976). Bayesian forecasting. *Journal of the Royal Statistical Society, Series B* **38**, 205–247. 148, 149

Heard, N.A., Holmes, C.C. and Stephens, D.A. (2006). A quantitative study of gene regulation involved in the immune response of anopheline mosquitoes: An application of Bayesian hierarchical clustering of curves. *Journal of the American Statistical Association* **101**, 18–29. 148

Ishwaran, H. and James, L.F. (2001). Gibbs sampling methods for stick-breaking priors. *Journal of the American Statistical Association* **96**, 161–173. 148, 150, 153

Jara, A., Lesaffre, E., De Iorio, M. and Quintana, F. (2010). Bayesian semiparametric inference for multivariate doubly-interval-censored data. *The Annals of Applied Statistics* **4**, 2126–2149. 152, 155

Lijoi, A., Mena, R.H. and Prünster, I. (2007). Controlling the reinforcement in Bayesian nonparametric mixture models. *Journal of the Royal Statistical Society, Series B* **69**, 715-740. 166

MacEachern, S.N. (1994). Estimating normal means with a conjugate style Dirichlet process prior. *Communications in Statistics B* **23**, 727–741. 155

Markowitz, H.M. (1952). Portfolio selection. *The Journal of Finance* **7**, 77-91. 147

Medvedovic, M. and Sivaganesan, S. (2002). Bayesian infinite mixture model based clustering of gene expression profiles. *Bioinformatics* **18**, 1194–1206. 156

Mendoza, M. and Nieto-Barajas, L.E. (2006). Bayesian solvency analysis with autocorrelated observations. *Applied Stochastic Models in Business and Industry* **22**, 169–180. 152

Mukhopadhyay, S. and Gelfand, A.E. (1997). Dirichlet process mixed generalized linear models. *Journal of the American Statistical Association* **92**, 633–639. 157

Navarrete, C., Quintana, F.A. and Müller, P. (2008). Some issues in nonparametric Bayesian modeling using species sampling models. *Statistical Modelling* **8**, 3–21. 151

Neal, R.M. (2000). Markov chain sampling methods for Dirichlet process mixture models. *Journal of Computational and Graphical Statistics* **9**, 249–265. 155

Pitman, J. (1995). Exchangeable and partially exchangeable random partitions. *Probability Theory and Related Fields* **102**, 145–158. 151, 154

Pitman, J. and Yor, M. (1997). The two-parameter Poisson-Dirichlet distribution derived from a stable subordinator. *The Annals of Probability* **25**, 855–900. 148, 150

Ross, S.M. (2000). *Introduction to probability models*. 7^{th} edition. Harcourt Academic Press, San Diego. 149

Smith, A. and Roberts, G. (1993). Bayesian computations via the Gibbs sampler and related Markov chain Monte Carlo methods. *Journal of the Royal Statistical Society, Series B* **55**, 3-23. 155

Stephens, M. (2000). Dealing with label switching in mixture models. *Journal of the Royal Statistical Society, Series B* **62**, 795-809. 156

Tierney, L. (1994). Markov chains for exploring posterior distributions. *Annals of Statistics* **22**, 1701-1722. 155

West, M. and Harrison, J. (1999). *Bayesian forecasting and dynamic models*. 2^{nd} edition. Springer, New York. 149

Zhou, C. and Wakefield, J. (2006). A Bayesian mixture model for partitioning gene expression data. *Biometrics* **62**, 515–525. 148