Bayesian Analysis (2011) **6**, Number 4, pp. 573–602

Model based clustering for three-way data structures

Cinzia Viroli*

Abstract. The technological progress of the last decades has made a huge amount of information available, often expressed in unconventional formats. Among these, three-way data occur in different application domains from the simultaneous observation of various attributes on a set of units in different situations or locations. These include data coming from longitudinal studies of multiple responses, spatio-temporal data or data collecting multivariate repeated measures. In this work we propose model based clustering for the wide class of continuous three-way data by a general mixture model which can be adapted to the different kinds of three-way data. In so doing we also provide a tool for simultaneously performing model estimation and model selection. The effectiveness of the proposed method is illustrated on a simulation study and on real examples.

Keywords: Mixture models, Birth and death process, Matrix-variate normal distribution, Three-way data.

1 Introduction

We are living in an era characterized by an exponentially increasing availability of information. This explosion of data, which can have complex structure, leads to an increasing demand for appropriate new strategies of statistical analysis. Although in principle complex data structures can take a myriad of formats, they can be often arranged in a three-way data structure. A three-way data set is characterized by three modes, namely rows, columns and layers. Depending on the entity indexed in each of the three modes, different data examples may be considered. Thus, for instance, longitudinal data on multiple response variables can be arranged in a three-way data set where n observed units are represented in rows, a set of p variables are indexed in columns and a set of r times are the layers. Some of the most common three-way data structures are illustrated in Table 1.

We focus on the problem of clustering the n observed objects represented in rows

*Department of Statistics, University of Bologna, Italy, cinzia.viroli@unibo.it

DOI:10.1214/11-BA622

Mode	*Rows*	*Columns*	*Layers*
Three-way structure			
Multivariate longitudinal data	units	variables	times
Multivariate repeated measures	units	variables	situations
Multivariate spatial data	units	variables	locations
Multivariate time-series	units/locations	times	variables
Multivariate spatio-temporal data	locations	variables	times
...	...	...	...

Table 1: Some common three-way data structures.

that can be a set of independent units or a set of spatially correlated locations. In a geometrical perspective, a unit of a three-way data set can be viewed as a point in the Euclidean space $\mathcal{R}^{(r \times p)}$ (in contrast to the conventional two-way data, where each unit is a point in the p-dimensional space $\mathcal{R}^p$). By denoting with j the generic observation (where $j = 1, \dots, n$), we have an $r \times p$ observed matrix, Y_j, for each statistical unit. Thus, the challenge of the cluster analysis is to suitably classify realizations coming from random matrices (instead of the conventional random univariate or p-variate variables) in some k unknown groups, with $k < n$. Considering the peculiarity of the data, a clustering strategy should address the following objectives:

i) modeling the possible (spatial) correlation between the observations (when units are not *i.i.d.*);

ii) defining two different covariance matrices for describing the variable correlations separately from the temporal (or spatial) correlations;

iii) modeling possible temporal (or spatial) covariance/correlation structures;

iv) estimating simultaneously the unknown number of groups k.

In this work we introduce, develop and explore a model based clustering approach, which simultaneously addresses all the above objectives. In so doing, the proposed approach represents a unified and general strategy for classifying the different types of three-way data described in Table 1.

2 Review of existing methods

Different solutions for clustering three-way data have been proposed in the statistical literature. The simplest scheme consists in applying some dimension reduction techniques, such as principal component analysis, to one of the modes, so as to convert the three-way data set to a two-way data set, and thereby to apply conventional clustering techniques. However, it can be shown that the leading components do not necessarily preserve the clustering structure of the data (Chang 1983). Gordon and Vichi (1998) and Vichi (1999) have developed a strategy based on a least-square approach, which has been recently extended in order to combine clustering and data reduction (Vichi et al. 2007). These methodologies do not require an explicit distributional assumption on the clusters and therefore do not allow one to explicitly model the correlation structures along the two modes of interest. In a model-based perspective, Basford and McLachlan (1985) adapted the Gaussian mixture likelihood approach to three-way data. In this approach they assumed that the component mean vectors might vary between groups and one of the two modes (for instance between the variables). On the contrary, the within component covariance matrices are not taken to depend on the modes. This implies that, coming back to the multivariate longitudinal data example, the correlations between and within variables and times are not explicitly modeled and this represents the main drawback of the method. In a different perspective, the Dirichlet process mixture models (Reich and Bondell 2010; Gelfand et al. 2005) provide an interesting approach for cluster analysis of multivariate spatial data, although they have not been specifically developed for three-way data.

More recently, Mixtures of Matrix Normal distributions (MMN) have been proposed and investigated (Viroli 2011) with the aim of taking into account the full information on the two modes, separately but simultaneously. This purpose is achieved by modeling the distribution of the observed matrices according to a matrix-variate normal distribution (Nel 1977; Dutilleul 1999). This approach represents a very general framework that includes, as special cases, both the conventional mixtures of multivariate normals and the variant proposed by Basford and McLachlan (1985) for the analysis of three-way data. The MMN model can be quite easily estimated by an EM algorithm (Dempster et al. 1977) under the hypothesis that the number of mixture components is fixed or known in advance and the observations are *i.i.d.* Thus, with reference to the four objectives described above, MMN addresses only objective *ii*.

In this work we propose a generalized MMN model (GMMN). With respect to MMN, the proposed GMMN model is developed in a Bayesian framework. This has the great advantage of extending MMN in order to address the remaining objectives described above. More specifically, we will show that GMMN provides a unified tool for model estimation and inference on the number of mixture components simultaneously (objective *iv*), with possible correlated observations (objective *i*) and temporal structured covariance matrices (objective *iii*). Model inference is solved via the Gibbs sampler (Geman and Geman 1984). For the general case of an unknown number of components, an ergodic Markov chain with a stationary distribution given by the posterior distribution of the enlarged parameter space to include k is constructed. This means that the dimension of the parameters is allowed to vary throughout the MCMC iterative procedure.

A very popular and conceptually elegant algorithm for inference of spaces with varying dimension is the reversible jump MCMC algorithm (Green 1995), which has been successfully applied to univariate Gaussian mixtures (Richardson and Green 1997). Its extension to mixtures of multivariate Gaussians is not straightforward because of the mathematical complexity of the split and combine moves and of the Jacobian. The problem has been solved via marginalized likelihood (Tadesse et al. 2005) or, alternatively, by posing specific constraints on the eigenvalue decomposition of the component covariance matrices (see Zhang et al. (2004), and Dellaportas and Papageorgiou (2006)).

As an alternative to the reversible jump MCMC algorithm, Stephens (2000a) developed a birth and death MCMC algorithm (BDMCMC). The key idea is to construct a Markov birth-death process in which the number of components can vary by allowing new components to be born and existing components to die in continuous time. BDMCMC can be thought of as a continuous-time version of the reversible jump MCMC (Cappé et al. 2002) in which both good and bad births may occur but they can survive for a different time interval according to their likelihood function. Stephens (2000a) has successfully applied the algorithm to mixtures of univariate and bivariate normals and t distributions. In this paper we adapt the BDMCMC scheme to the GMMN.

The paper is organized as follows. In the next Section, we introduce the proposed model. In Section 4 we show, from a theoretical point of view, the application to multivariate repeated measures, longitudinal data on multiple response variables and multivariate spatio-temporal data. Section 5 covers the difficult issue of model selection by the construction of a BDMCMC chain. In Section 6 we present some numerical experiments with simulated and real data. Discussion and concluding remarks are presented in Section 7.

3 Model specification

3.1 Generalized Mixture of Matrix Normals

Let k be a set of sub-populations (or groups) of unknown proportions from which the observed units are supposed to come. In a general perspective we consider the observed sample of matrices $Y_1, Y_2, \dots, Y_n$, as a set of conditionally independent and not identically distributed observations coming from the mixture model

$$f(Y_j|k, \boldsymbol{\pi}, \boldsymbol{\Theta}_1, \dots, \boldsymbol{\Theta}_k) = \sum_{i=1}^{k} \pi_{ij} \mathcal{M}_{(r\times p)}(Y_j; \boldsymbol{\Theta}_i), \tag{1}$$

where $j = 1, \dots, n$ and $\boldsymbol{\Theta}_i$ denotes the set of parameters of each component distribution. The weights $\boldsymbol{\pi} = [\pi_{ij}]_{i=1,\dots,k;j=1,\dots,n}$ satisfy $\pi_{ij} > 0$ with $\sum_{i=1}^{k} \pi_{ij} = 1$ for all j. They vary with j because the observations are not necessarily taken independently. Obviously, in the case of an independent sample of observations, $\pi_{ij} = \pi_i$ for all j. We focus on continuous observed matrices. The distribution of the generic i-th component should allow for a separate treatment of the variability of the second and third mode, in order to model possible auto-correlated temporal or spatial covariance structures when necessary. To this purpose the i-th density is assumed to be a matrix-variate normal distribution. More specifically, the density of the $r \times p$ matrix of observations, Y_j, is the matrix normal distribution of parameters $\boldsymbol{\Theta}_i = \{M_i, \Phi_i, \Omega_i\}$:

$$\begin{aligned}\mathcal{M}_{(r\times p)}(Y_j; M_i, \Phi_i, \Omega_i) \quad = \quad & (2\pi)^{-\frac{rp}{2}} |\Phi_i|^{-\frac{p}{2}} |\Omega_i|^{-\frac{r}{2}} \\ & \exp\left\{-\frac{1}{2}\mathrm{tr}\left(\Phi_i^{-1}(Y_j - M_i)\Omega_i^{-1}(Y_j - M_i)^\top\right)\right\}\end{aligned} \tag{2}$$

where M_i is an $r \times p$ matrix of means; Φ_i an $r \times r$ covariance matrix containing the variances and covariances between the r entities within the third mode; and Ω_i is a $p \times p$ covariance matrix containing the variance and covariances of the p variables (or times) indexed by the second mode. The Kronecker product of the two covariance matrices $\Sigma_i = \Phi_i \otimes \Omega_i$ contains the $pr \times pr$ covariances between the entities of the two modes.

3.2 Hierarchical formulation of GMMN

Being a typical incomplete-data problem, the GMMN model can be rephrased according to a hierarchical formulation.

We introduce n independent latent variables, $\{z_1, \dots, z_n\}$ called *allocation* variables, that identify the sub-population (or group) from which each observed matrix comes.

More precisely, z_j (with $j = 1, \dots, n$) is a vector of dimension k which assumes value equal to 1 if the observation belongs to one of the k sub-populations and 0 elsewhere. Therefore z_j follows a multinomial distribution

$$f(z_j|\boldsymbol{\pi}, k) = \prod_{i=1}^{k} \pi_{ij}^{z_{ij}}, \tag{3}$$

from which $f(z_{ij} = 1|\boldsymbol{\pi}, k) = \pi_{ij}$.

For correlated observations we assume that $Y_1, \dots, Y_n$ are independent given the set of latent variables $\mathbf{z} = \{z_1, \dots, z_n\}$. For unconditionally independent observed matrices the n components of $\mathbf{z}$ are all equal to each other; that is to say, $\mathbf{z}$ reduces to a single latent vector, z, of length k.

The conditional density of the random matrix, Y_j, given the allocation variable, z_j, is the matrix-variate normal distribution in the form:

$$f(Y_j|z_j, \boldsymbol{\Theta}, k) = \prod_{i=1}^{k} \left[\mathcal{M}_{(r\times p)}(Y_j; M_i, \Phi_i, \Omega_i)\right]^{z_{ij}}. \tag{4}$$

Given k and the set of parameters $\boldsymbol{\pi}$ and $\boldsymbol{\Theta}$, the complete joint distribution of Y and $\mathbf{z}$ can be decomposed into the product of two conditional densities

$$f(Y, \mathbf{z}|\boldsymbol{\pi}, \boldsymbol{\Theta}, k) = f(Y|\mathbf{z}, \boldsymbol{\Theta}, k) f(\mathbf{z}|\boldsymbol{\pi}, k). \tag{5}$$

3.3 Posterior density

We allow additional layers to the hierarchy by adding a set of hyperparameters $\boldsymbol{\omega}$ for $\boldsymbol{\Theta}$ and $\boldsymbol{\pi}$. In the GMMN model the distribution of interest is the posterior distribution of the allocation variables, of the parameters and hyperparameters (including k) given the observed data Y. By using formulation (5) it can be expressed as

$$f(\mathbf{z}, \boldsymbol{\pi}, \boldsymbol{\Theta}, \boldsymbol{\omega}, k|Y) \propto f(Y|\mathbf{z}, \boldsymbol{\Theta}, k) f(\mathbf{z}|\boldsymbol{\pi}, k) f(\boldsymbol{\pi}|\boldsymbol{\omega}, k) f(\boldsymbol{\Theta}|\boldsymbol{\omega}, k) f(\boldsymbol{\omega}|k) f(k), \tag{6}$$

where $f(\boldsymbol{\pi}|\boldsymbol{\omega}, k)$, $f(\boldsymbol{\Theta}|\boldsymbol{\omega}, k)$, $f(\boldsymbol{\omega}|k)$ and $f(k)$ are the prior distributions of parameters and hyperparameters. Considering a sample of n observations, the posterior distribution can be expressed as follows:

$$\begin{aligned} f(\mathbf{z},\boldsymbol{\pi},\boldsymbol{\Theta},\boldsymbol{\omega},k|Y) \quad &\propto \quad \left\{\prod_{j=1}^{n}\prod_{i=1}^{k}|\Phi_i|^{-\frac{p}{2}z_{ij}}\right\}\left\{\prod_{j=1}^{n}\prod_{i=1}^{k}|\Omega_i|^{-\frac{r}{2}z_{ij}}\right\} \\ &\times \quad \left\{\exp\left[-\frac{1}{2}\sum_{j=1}^{n}\sum_{i=1}^{k}\left(\mathrm{tr}\left(\Phi_i^{-1}(Y_j-M_i)\Omega_i^{-1}(Y_j-M_i)^\top\right)\right)z_{ij}\right]\right\} \\ &\times \quad \left\{\prod_{j=1}^{n}\prod_{i=1}^{k}\pi_{ij}^{z_{ij}}\right\} f(\boldsymbol{\pi}|\boldsymbol{\omega},k)f(\boldsymbol{\Theta}|\boldsymbol{\omega},k)f(\boldsymbol{\omega}|k)f(k). \end{aligned} \tag{7}$$

3.4 Class prediction

In a model based perspective the problem is to allocate the random sample of n observed matrices $Y_1,\ldots,Y_n$ to the component from which they are assumed to come, under the assumption that each component corresponds to each sub-population or group. Classification of units may be performed by computing the conditional probabilities $f(z_{ij}=1|Y,\boldsymbol{\pi},\boldsymbol{\Theta},k)=\tau_{ij}(\boldsymbol{\pi},\boldsymbol{\Theta})$ which can be derived by the Bayes' rule:

$$\tau_{ij}(\boldsymbol{\pi},\boldsymbol{\Theta}) \quad = \quad \frac{\pi_{ij}\mathcal{M}_{(r\times p)}(Y_j;M_i,\Phi_i,\Omega_i)}{\sum_{h=1}^{k}\pi_{hj}\mathcal{M}_{(r\times p)}(Y_j;M_h,\Phi_h,\Omega_h)}. \tag{8}$$

3.5 Identifiability

The model identifiability is crucial to obtain unique and consistent parameter estimates. The model is not identified when different parameter vectors parameterize the same distribution and therefore they are equivalent. There are two different identifiability aspects to be considered in the proposed GMMN. The first one arises from the lack of uniqueness of the Kronecker product of the two covariance matrices, which is $\Sigma_i=\Phi_i\otimes\Omega_i$, for all $i=1,\ldots,k$. Given a multiplicative constant a, different from zero, $\Sigma_i=(a\Phi_i)\otimes(\frac{1}{a}\Omega_i)$. This ambiguity can be avoided by assuming that the trace of each Φ_i, for all i, is equal to r, or alternatively, by imposing the trace of the matrices Ω_i to be equal to p.

The second aspect is that, being a mixture model, the proposed GMMN could be affected by the so-called label-switching problem discussed by Stephens (2000b). The label-switching problem arises since the likelihood is invariant under relabeling of the mixture components and so under any possible permutation. Various solutions to the label-switching problem have been proposed, including the k-means clustering algorithm and a transportation algorithm for optimization (see, among the others, Celeux (1998) and Stephens (2000b)). In this work we adopt the solution to postprocess the MCMC output by relabeling the sampled parameters according to the ranking of $\mathrm{tr}\left(\Phi_i^{-1} M_i \Omega_i^{-1} M_i^\top\right)$ for all i.

4 Specific Applications

4.1 Multivariate repeated measures

In multivariate repeated measures or longitudinal data on multiple responses, we observe a set of independent but not identically distributed matrices, $Y_1, \ldots, Y_n$ containing the multiple responses in the different time points or situations. Within the i-th subpopulation Ω_i is the covariance matrix of the variables and Φ_i the covariance matrix between the repeated situations. We assume they are unstructured matrices. The matrix $\Sigma_i = \Phi_i \otimes \Omega_i$ of dimension $rp \times rp$ contains the covariances between the p variables in the r repeated situations.

Prior formulation and hyperparameters

In this setting $\pi_{ij} = \pi_i$ for all j and distribution (3) simplifies to

$$f(z|\boldsymbol{\pi}, k) = \prod_{i=1}^{k} \pi_i^{z_i}, \tag{9}$$

because of the independence between the observations. The set of hyper-parameters is $\boldsymbol{\omega} = (\beta, \rho)^\top$, and we further assume $f(\boldsymbol{\pi}|\boldsymbol{\omega}, k) = f(\boldsymbol{\pi}|k)$. The role of the two hyperparameters β and ρ is to parameterize the prior distributions of Φ_i and Ω_i respectively, for all i. We can choose non-informative prior distributions for the model parameters. More precisely:

$$\begin{aligned}
M_i &\sim \mathcal{M}_{(r\times p)}(M_0, \Phi_0, \Omega_0) && (10)\\
\Phi_i^{-1}|\beta &\sim \mathcal{W}_r\left(2\alpha, (2\beta)^{-1}\right) && (11)\\
\beta &\sim \mathcal{W}_r\left(2g, (2h)^{-1}\right) && (12)\\
\Omega_i^{-1}|\rho &\sim \mathcal{W}_p\left(2\zeta, (2\rho)^{-1}\right) && (13)\\
\rho &\sim \mathcal{W}_p\left(2l, (2m)^{-1}\right) && (14)\\
\boldsymbol{\pi} &\sim \mathcal{D}(\varrho, \ldots, \varrho) && (15)
\end{aligned}$$

for $i = 1, \dots, k$. In the previous expressions $\mathcal{M}_{(r\times p)}$ denotes the matrix-variate normal distribution of order $r \times p$, $\mathcal{W}$ denotes the multivariate Wishart distribution and $\mathcal{D}$ the symmetric Dirichlet. Moreover, β, Φ_0 and h are $r \times r$ matrices, ρ, Ω_0 and m are $p \times p$ matrices, M_0 is an $r \times p$ matrix and α, ζ, l and g are scalars. The prior distribution for k is assumed to be a truncated Poisson:

$$f(k) \propto \frac{\lambda^k}{k!}, \qquad k \in \{1, 2, \dots, k_{\max}\}$$

where the constants λ and $k_{\max}$ are data-driven choices.

Full conditionals

The full conditional distributions are proportional to known distributions. By using $|\dots$ to denote conditioning on all other variables, they are:

$$f(z_{ij} = 1|\dots) \propto \pi_i \mathcal{M}_{(r\times p)}(Y_j; M_i, \Phi_i, \Omega_i) \tag{16}$$

$$\mathrm{vec}(M_i)|\dots \sim \mathcal{N}_{rp}\left(\Upsilon^{-1}\xi, \Upsilon^{-1}\right) \tag{17}$$

$$\Phi_i^{-1}|\dots \sim \mathcal{W}_r\left(2\alpha + pn_i, \left[2\beta + \sum_{j:z_j=i}(Y_j - M_i)\Omega_i^{-1}(Y_j - M_i)^\top\right]^{-1}\right) \tag{18}$$

$$\Omega_i^{-1}|\dots \sim \mathcal{W}_p\left(2\rho + rn_i, \left[2\zeta + \sum_{j:z_j=i}(Y_j - M_i)^\top\Phi_i^{-1}(Y_j - M_i)\right]^{-1}\right) \tag{19}$$

$$\beta|\dots \sim \mathcal{W}_r\left(2g + 2k\alpha, \left[2h + 2\sum_{i=1}^{k}\Psi_i^{-1}\right]^{-1}\right) \tag{20}$$

$$\rho|\dots \sim \mathcal{W}_p\left(2l + 2k\zeta, \left[2m + 2\sum_{i=1}^{k}\Omega_i^{-1}\right]^{-1}\right) \tag{21}$$

$$\boldsymbol{\pi}|\dots \sim \mathcal{D}(\varrho + n_1, \dots, \varrho + n_k) \tag{22}$$

where $n_i = \sum_{j=1}^n z_{ij}$, $\xi = (\Phi_i \otimes \Omega_i)^{-1}\sum_{j=1}^n \mathrm{vec}(Y_j)z_{ij} + (\Phi_0 \otimes \Omega_0)^{-1}\mathrm{vec}(M_0)$ and $\Upsilon = n_i(\Phi_i \otimes \Omega_i)^{-1} + (\Phi_0 \otimes \Omega_0)$. The analytical derivation of full conditional (17) is demonstrated in the Appendix. The other previous expressions can be easily obtained by combining equation (7) with the priors previously described.

4.2 Multivariate longitudinal data

In this second framework, the sample $Y_1, \dots, Y_n$ represents a set of independent but not identically distributed observed matrices at different r times; therefore, again, we have $\pi_{ij} = \pi_i$ for all j. Within each sub-population i, Φ_i is assumed to be a temporal structured covariance matrix. There are several popular correlation structures, including the compound symmetry structure, the first-order autoregressive AR(1) structure or the Toeplitz structure (see, for more examples, Jennrich and Schluchter (1986)). In this paper we confine attention to the AR(1) structure for all the Φ_i covariance matrices. The other common types of temporal structures could be considered and adapted to the proposed setting with little mathematical treatment. In our setting, within each component i, the covariance matrix Φ_i can be decomposed as

$$\Phi_i(\beta_i) = (\sigma_i \mathbf{I}_r) R_i(\beta_i)(\sigma_i \mathbf{I}_r), \tag{23}$$

where $R_i(\beta_i)$ is a correlation matrix having the AR(1) structure:

$$R_i(\beta_i) = [\beta_i]^{|u-v|} \qquad \text{with } u, v = 1, \dots, r.$$

Identifiability constraint

It is worth noting that in this framework the estimation problem is simplified under the identifiability constraint that the trace of each Φ_i is equal to r. In fact, expression (23) becomes $\Phi_i(\beta_i) = R_i(\beta_i)$ because the identifiability condition is equivalent to imposing $\sigma_i = 1$, for all $i = 1, \dots, k$.

Prior formulation and hyperparameters

In this situation, the set of hyperparameters is $\boldsymbol{\omega} = (\beta_1, \dots, \beta_k, \sigma_1, \dots, \sigma_k, \rho)^\top$. We consider non-informative prior distributions for each β_i given by uniform distributions in [-1,1]. The prior distribution for σ_i^{-1} is a Gamma distribution with parameters a and b, for all i, with $i = 1, \dots, k$. The prior distribution for ρ is given in (14). Being a deterministic function of β_i and σ_i, there is no prior distribution for Φ_i. It is worth noting that, by fixing $\beta_i = \beta$ and $\sigma_i = \sigma$, a GMMN model with homoscedastic temporal components could be estimated.

Full conditionals

The full conditionals for β_i and σ_i are derived in the Appendix. The other posterior distributions are not changed with respect to the setting of the previous situation and are given in expressions (17), (19), (21) and (22).

4.3 Multivariate spatio-temporal data

In this setting, we consider p variables observed at r times for n different locations. The set of observed matrices, $Y_1, \dots, Y_n$, are not independent but they can be spatially correlated. Spatial relationships are taken into account by allowing the weights of the mixture in (1) to vary from one location to another. This solution is inspired by the spatial mixture formulation for Poisson distributed two-way data proposed by Fernández and Green (2002). It consists of introducing k independent additional latent variables to capture spatial correlation. The weights are a function of these latent variables via the logistic transform so as to incorporate the spatial dependence in the mixture model. The temporal correlation between the observed matrices is modelled by estimating AR(1) structured covariance matrices, as presented in the previous section.

Prior formulation and hyperparameters

In this formulation, the set of hyperparameters $\boldsymbol{\omega}$ includes also the spatial latent variables, denoted by $x_1, \dots, x_k$, and an additional non-negative parameter ζ. Each x_i (with $i = 1, \dots, k$) is a Markov random field with density function:

$$f(x_i|\zeta) = (2\pi)^{-n/2} \prod_{j=1}^{n} (1 + \zeta \upsilon_j)^{1/2} \exp\left[-\frac{1}{2} \left(\zeta \sum_{j \sim j'} (x_{ij} - x_{ij'})^2 + \sum_{j=1}^{n} x_j^2 \right) \right] \tag{24}$$

where $\upsilon_1, \dots, \upsilon_n$ denote the eigenvalues of a spatial matrix which contains the number of neighbours of each location in the diagonal, the value -1 if two locations are neighbours and zero otherwise. $\sum_{j \sim j'}$ denotes the sum over all pairs of neighbours with each pair counted only once. Finally ζ is a hyperparameter with uniform prior distribution between 0 and $\zeta_{\max}$. When $\zeta = 0$ there is independence between locations, as ζ increases neighbouring locations have ever more similar values of the spatial latent variable x. Given $x_1, \dots, x_k$ and ζ the weights for location j take the form

$$\pi_{ij} = \frac{e^{x_{ij}}}{\sum_{h=1}^{k} e^{x_{hj}}}.$$

Full conditionals

The full conditional for $\mathbf{z}$ is now

$$f(z_{ij} = 1|\ldots) \propto \pi_{ij}\mathcal{M}_{(r\times p)}(Y_j; M_i, \Phi_i, \Omega_i) \tag{25}$$

instead of (16). The posterior distributions of $x_1, \ldots, x_k$ and ζ are derived in the Appendix.

5 Stochastic model selection for GMMN

In many real applications the number of mixture components k is unknown and this model uncertainty has to be addressed. Conventionally in model based clustering each mixture component is interpreted as a cluster even if non-homogeneous clusters could be themselves described by a mixture of two or more component distributions. This latter case can be addressed by merging mixture components according to some criterion (Hennig 2010; Baudry et al. 2008). Making inference on k is of interest, regardless the role of each component in a clustering perspective.

In the framework of GMMN, the BDMCMC scheme essentially consists of randomly jumping between mixture models with a different number of components. Consider the set of varying-dimension parameters in compact form denoted by $\boldsymbol{\xi}^{(k)} = \{k, \boldsymbol{\pi}^{(k)}, \boldsymbol{\Theta}^{(k)}\}$ in order to make explicit that the dimension of the parameters changes with k. The aim is to combine the previously described posterior distributions in order to construct an irreducible Markov chain with stationary distribution $f(z, \boldsymbol{\pi}^{(k)}, \boldsymbol{\Theta}^{(k)}, \boldsymbol{\omega}^{(k)}, k|Y) = f(z, \boldsymbol{\xi}^{(k)}, \boldsymbol{\omega}^{(k)}|Y)$. The unknown quantities of interest $\boldsymbol{\xi}^{(k)}$ and $\boldsymbol{\omega}^{(k)}$ may be estimated by the sample path averages of the chain.

5.1 A birth-death process for GMMN

Following the scheme proposed by Stephens (2000a), we first construct a continuous time Markov birth-death process with stationary distribution $f(\boldsymbol{\xi}^{(k)}|Y, \boldsymbol{\omega}^{(k)})$ (see Algorithm 1). The continuous birth-death process is run for a virtual time of length t_0, in which many birth and death events may occur. Births and deaths are assumed to be independent Poisson processes. As a consequence, the time passing between each birth/death event is exponentially distributed with parameter depending on the birth/death rates. In the following we present the birth-death algorithm. By combining it with the full conditionals presented in Section 4, a Markov chain with stationary distribution $f(z, \boldsymbol{\xi}^{(k)}, \boldsymbol{\omega}^{(k)}|Y)$ is then implemented (see Algorithm 2).

Algorithm 1 In order to simulate a process with stationary distribution $f(\boldsymbol{\xi}^{(k)}|Y, \boldsymbol{\omega}^{(k)})$, a set of initial values for GMMN parameters in $\boldsymbol{\xi}^{(k)}$ must be chosen. Then iterate the following steps for an interval of time t_0:

1. Let the birth rate be equal to λ;

2. Calculate the death rate for each component:

$$\delta_i = \lambda \frac{f(Y|\boldsymbol{\xi}_{-i}^{(k-1)}, \boldsymbol{\omega}^{(k-1)})}{f(Y|\boldsymbol{\xi}^{(k)}, \boldsymbol{\omega}^{(k)})} \frac{p(k-1)}{kp(k)} \qquad (i = 1, \ldots, k).$$

3. Calculate the total death rate $\delta = \sum_{i=1}^{k} \delta_i$.

4. Simulate the time to the next jump from an exponential distribution with mean $1/(\lambda + \delta)$.

5. Simulate the type of jump: birth or death with respective probabilities $\frac{\lambda}{\lambda+\delta}$ and $\frac{\delta}{\lambda+\delta}$

6. Adjust the parameters to reflect the birth or death:
   Birth: Simulate each π_j from a Beta distribution with parameters $(1, k)$ and $\boldsymbol{\Theta}$ from the prior densities described in Section 4.
   Death: Select a component to die with probability $\frac{\delta_i}{\delta}$, for $i = 1, \ldots, k$.

In the algorithm $f(Y|\boldsymbol{\xi}^{(k)}, \boldsymbol{\omega}^{(k)})$ is the likelihood of the GMMN model:

$$f(Y|\boldsymbol{\xi}^{(k)}, \boldsymbol{\omega}^{(k)}) = \prod_{j=1}^{n} \left(\pi_{1j} \mathcal{M}_{(r\times p)}(Y_j; M_1, \Phi_1, \Omega_1) +, \ldots, +\pi_{kj} \mathcal{M}_{(r\times p)}(Y_j; M_k, \Phi_k, \Omega_k)\right)$$

and $f(Y|\boldsymbol{\xi}_{-i}^{(k-1)}, \boldsymbol{\omega}^{(k-1)})$ is the likelihood of the GMMN model without the i-th component. Therefore the chain may lead to many births of components which do not contribute to describe the data, but such components will have a higher death rate. The convergence of this process to the stationary distribution $f(\boldsymbol{\xi}^{(k)}|Y, \boldsymbol{\omega}^{(k)})$ follows from Stephens (2000a) (Theorem 3.1).

5.2 The BDMCMC algorithm

By augmenting the data Y by the allocation variable z and combining Algorithm 1 with the previously described full conditionals an algorithm with stationary distribution $f(z, \boldsymbol{\xi}^{(k)}, \boldsymbol{\omega}^{(k)}|Y)$ can be implemented.

Algorithm 2 Assuming a current set $(\boldsymbol{\xi}_t^{(k)}, z_t, \boldsymbol{\omega}_t^{(k)})$ of unknown quantities at iteration t, simulate a new set for iteration $t+1$ as follows:

1. Simulate $\boldsymbol{\xi}_{t'}^{(k)}$ by running Algorithm 1 for a virtual fixed time t_0 starting from $\boldsymbol{\xi}_t^{(k)}$ and fixing $\boldsymbol{\omega}^{(k)}$ to be $\boldsymbol{\omega}_t^{(k)}$. Set $\boldsymbol{\xi}_{t+1}^{(k)} = \boldsymbol{\xi}_{t'}^{(k)}$.
2. Simulate z_{t+1} from its full conditional given the current set of parameters.
3. Simulate $\boldsymbol{\omega}_{t+1}^{(k)}$ from the full conditionals given the current set of parameters.
4. Simulate $\boldsymbol{\pi}_{t+1}^{(k)}$ and $\boldsymbol{\Theta}_{t+1}^{(k)}$ from the full conditionals given the current set of parameters.

Note that step 4 is not strictly necessary, since $\boldsymbol{\pi}^{(k)}$ and $\boldsymbol{\Theta}^{(k)}$ are also updated at step 6 of Algorithm 1. This step has been included to improve mixing as suggested by Stephens (2000a). The algorithm requires the specification of the time t_0 for which the birth and death process is run. This choice is related to the choice of the value for the birth rate λ since the time between each birth/death event is distributed according to an exponential distribution with mean $1/(\lambda + \delta)$. In our applications, we have fixed $t_0 = 1$ and set different values for λ.

6 Experimental results

In the following, some numerical experiments with simulated and real data are presented. The utilized code has been implemented in `R` (R Development Core Team 2008) and is available at the author homepage.

6.1 Simulated data

The proposed GMMN is here evaluated on a simulation study. The aim of the simulated example is twofold. First we implement a Gibbs sampling MCMC algorithm in simulated data with supposed known k. Then the BDMCMC algorithm is applied in order to make inference on k. We also compare the obtained results with those obtained by applying the MMN (Viroli 2011) on the same data.

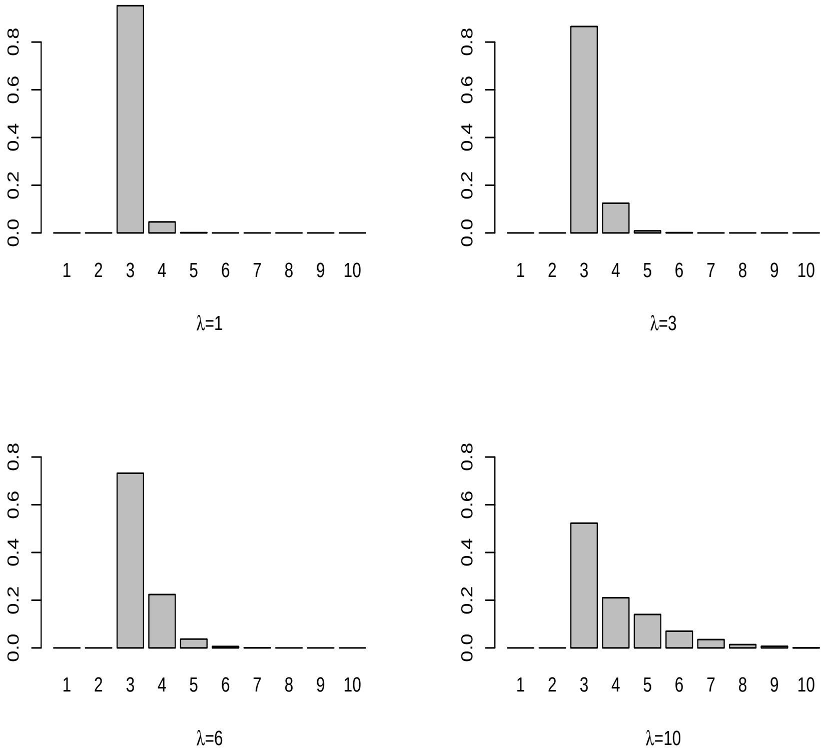

Figure 1: *Simulated data*. The figure shows the barplots of the posterior distribution of k (ranging from 1 to 10) according to different priors ($\lambda = 1, 3, 6, 10$).

A sample of 500 matrix observations with $r = 3$ and $p = 5$ has been generated from a mixture of three matrix normals with mean matrices reported in Table 2 and weights given by $\pi = \{0.3, 0.4, 0.3\}$. The six (3×2) covariance matrices have been randomly generated through the methodology proposed by Joe (2006). The simulated data can be cast in the scheme of multivariate repeated measures. On this data we have implemented a Gibbs sampling using the full conditionals derived in Section 4 for this kind of three-way data. The values of the fixed quantities at the lowest level of the hierarchical model have been chosen in order to guarantee relatively flat priors. M_0

has been set equal to the mid-point of the range of the observations. The values of the other constants are:

$$\Phi_0 = \operatorname{diag}\left(V_1^2, \ldots, V_r^2\right), \qquad \Omega_0 = \operatorname{diag}\left(S_1^2, \ldots, S_p^2\right)$$
$$\alpha = 1 + (r \cdot k), \quad \rho = 1 + (p \cdot k) \qquad g = \frac{r}{2}, \qquad l = \frac{p}{2}$$
$$h = \operatorname{diag}\left(\frac{100g}{\alpha R_1^2}, \ldots, \frac{100g}{\alpha R_r^2}\right), \qquad m = \operatorname{diag}\left(\frac{100l}{\rho S_1^2}, \ldots, \frac{100l}{\rho S_p^2}\right) \qquad \varrho = \frac{(k-1)\cdot n}{p \cdot r},$$

where $V_1, \ldots, V_r$ and $S_1, \ldots, S_p$ are the ranges along the two modes. In order to monitor convergence to stationarity, graphical diagnoses and a multiple sequence diagnostic analysis have been explored. We have run $m = 10$ parallel Gibbs sampling chains, each of length $n = 10,000$, with randomly selected starting points. The parallel sequences have been compared using the analysis of variance approach by Gelman and Rubin (1992). The potential scale reduction factor, $PSRF$, for each value of the component mean matrices, is reported in Table 2. Since these values are very close to 1, no particular convergence problem has been observed. The good convergence behaviour was quite evident by looking at some graphical diagnoses as well, like the traceplots and the density plots of the sampled parameters, the running mean plots and the autocorrelation plots (not reported here for space reasons). On the basis of these diagnostic criteria, we decided to discard the first 2,000 values as burn-in.

The Gibbs sampler estimates for the component mean matrices have been obtained by computing the posterior mean across the $m = 10$ parallel outputs. These values (denoted by $\hat{M}_{\mathrm{GS}}$) are reported in Table 2. As shown in the table, estimates are quite accurate.

The BDMCMC algorithm has been applied with random starting points for the model parameters and with values $\lambda = 1, 3, 6, 10$. We have run 10,000 iterations of the algorithm for each of the four settings. The run took about an hour on `R 2.8.1` under Microsoft Windows XP professional (CPU X86 Family 15, Model 4, ∼2666 Mhz). In order to have an idea of the mixing of the chain, we have computed the percentages of iterations which changed k, which in this case were 9.3% for $\lambda = 1$, 23.9% for $\lambda = 3$, 40.9% for $\lambda = 6$ and 45.4% for $\lambda = 10$.

Inference on k may be based on the estimates of the marginal posterior distribution. Figure 1 shows the barplots of the posterior distribution of k (ranging from 1 to 10) according to the different priors ($\lambda = 1, 3, 6, 10$). These results indicate that inference for k is quite sensitive to the choice of the priors. As λ increases the mixing behavior increases as well and the algorithm explores more states of the space. However, the

	$i=1$			$i=2$			$i=3$		
	M	$\hat{M}_{\text{GS}}$	$PSRF$	M	$\hat{M}_{\text{GS}}$	$PSRF$	M	$\hat{M}_{\text{GS}}$	$PSRF$
M_{11}	-0.18	-0.18	1.00	-0.44	-0.41	1.00	-0.19	-0.20	1.00
M_{21}	-0.88	-0.86	1.00	0.28	0.21	1.00	0.99	0.96	0.99
M_{31}	-0.03	-0.10	1.00	-0.63	-0.61	1.00	-0.25	-0.30	1.00
M_{12}	-0.25	-0.20	0.99	0.34	0.33	0.99	-0.28	-0.27	1.00
M_{22}	0.08	0.01	1.00	-0.80	-0.71	1.01	-0.13	-0.25	1.00
M_{32}	0.26	0.25	1.00	-0.15	-0.10	1.00	0.43	0.43	1.00
M_{13}	-0.42	-0.32	1.00	-0.46	-0.52	1.01	0.91	0.77	1.00
M_{23}	-0.47	-0.48	1.03	-0.85	-0.66	1.00	0.67	0.09	1.00
M_{33}	-0.69	-0.72	1.00	-0.60	-0.61	1.00	-0.08	-0.27	1.00
M_{14}	-0.78	-0.58	1.00	0.72	0.68	1.00	-0.15	-0.02	1.00
M_{24}	-0.45	-0.43	1.00	0.78	0.74	1.00	0.19	0.15	1.03
M_{34}	0.09	0.04	1.02	0.02	0.05	1.00	0.85	0.84	1.00
M_{15}	-0.36	-0.33	1.00	-0.42	-0.38	1.02	0.50	0.30	1.00
M_{25}	-1.00	-0.92	1.00	0.11	0.13	1.00	0.76	0.61	1.00
M_{35}	0.37	0.31	1.01	0.02	-0.01	1.02	-0.19	-0.05	1.00

Table 2: *Simulated data*. The table shows the true values of the mean matrices, M, of the three components ($i = 1, 2, 3$) and the corresponding Gibbs sampling estimates denoted by $\hat{M}_{\text{GS}}$. $PSRF$ is the potential scale reduction factor (Gelman and Rubin 1992).

true value of components, $k = 3$, is suggested in all the four settings. For comparative purposes, on the same data we have applied the MMN (Viroli 2011) with different values of k ranging from 1 to 4 and with different starting points, randomly generated, in the EM-algorithm. It is well known that the EM-algorithm is quite sensitive to the choice of its initial values and this aspect was evident also in our analysis. Table 3 shows the best values of the Bayesian Information Criterion (BIC), the Akaike Information Criterion (AIC) and the Integrated Classification Likelihood Criterion (ICL-BIC) obtained in a sequence of 100 multistart estimation procedures.

Table 3: Frequencies with which each model is selected according to the information criteria BIC, AIC and ICL-BIC.

	$k = 1$	$k = 2$	$k = 3$	$k = 4$
BIC	0	20	41	39
AIC	0	17	43	40
ICL-BIC	0	20	41	39

Results show that in this simulation study all the criteria suggest the correct number of components, $k = 3$, most of the time, but $k = 4$ seems to be a plausible choice, as well. From a computational point of view, the computational time required for this extensive exploratory analysis was about three hours and hence significantly higher than those required for running the BDMCMC algorithm.

6.2 Real example 1: Headache data

This data set derives from a study of C. Philips and M. Jahanshahi of the London University Institute of Psychiatry (Hand and Taylor 1987). The aim of the study was to investigate the reactions to noise of $n = 75$ headache sufferers, distinguished by two types of headache: migraine or tension. Each subject was exposed to a sequence of operations; first an initial measurement of sensitivity scores followed by a relaxation pause and treatment, then again the measurement of sensitivity scores. The sensitivity scores were obtained by listening to a tone which was gradually increasing in volume. The levels at which the noise became uncomfortable (first variable of the analysis) and definitely unpleasant (second variable of the analysis) have been recorded. Information collected in this study represents an example of multivariate repeated measures where $p = 2$ variables are observed in $r = 2$ time points. We want to investigate if the reaction to the noise listening is different for the two kinds of headache sufferers or in other terms if the observed subjects cluster into $k = 2$ groups. To answer this question, we have applied the proposed GMMN model with unstructured covariance matrices. Since here k is known, we have directly implemented a Gibbs sampling using the full conditionals derived in Section 4. Results are based on runs of length 10,000 with the first 5,000 iterations being discarded as burn-in. This burn-in period is larger than necessary, since from all graphical tools for monitoring convergence it was evident that stationarity is quickly achieved.

Table 4 shows the misclassification rates of the fitted model. For comparative purposes, we have fitted the mixture model proposed by Basford and McLachlan (1985) and some conventional clustering methods on the data collapsed by taking the mean of the variables between the two time points. In the table we have reported the misclassification rates obtained with the Basford and McLachlan model (BMclust), the Gaussian mixtures by the `mclust` package of `R` (Fraley and Raftery 2002, 2006) and the classical hierarchical clustering (HC) according to different methods (complete linkage, single linkage and Ward method).

GMMN	BMclust	Complete HC	Single HC	Ward HC	Mclust
0.067	0.213	0.173	0.293	0.200	0.133

Table 4: Misclassification rates for headache data according to different clustering methods: BMclust refers to the Basford and McLachlan model, Complete HC, Single HC and Ward HC denote the hierarchical clustering according to the complete linkage, the single linkage and the Ward method, respectively; Mclust refers to the conventional Gaussian mixtures.

Results indicate that collapsing the full information contained in the two modes leads to less flexible methods, while GMMN leads to the lowest error rate; only 5 out of 75 subjects are misclassified.

6.3 Real example 2: Crime in the 103 Italian provinces

Every year, an Italian financial newspaper, *Il Sole 24 Ore*, analyzes the quality of life in the 103 provinces of Italy through several indicators collected in different thematic areas (`www.ilsole24ore.com`). This data set consists of $p = 4$ measurements on crime in the Italian provinces collected and published in $r = 5$ years, from 2005 to 2009. The $p = 4$ indicators are: home-invasion robberies (per 100,000 residents), teenage crime rate (per 1,000 residents), the number of reported robberies (per 100,000 residents) and rate of muggings and pickpockets (per 100,000 residents). These are not violent crime measurements but they could still offer a useful indication on the safety level in the different geographical areas. Since Italy is a complex and heterogeneous country characterized by a deep income inequality between the dynamic, industrialized North and the less developed, agricultural-based Centre-South, we expect a deep territorial heterogeneity in terms of safety and quality of life.

The aim of this study is to cluster the Italian provinces on the basis of the four crime indicators taking into account the entire period of the five years 2005-2009.

In order to measure the strength of the spatial dependence among the provinces we have computed the Moran's I autocorrelation coefficient (Gittleman and Kot 1990) separately for year and variable (see Table 5). The null hypothesis of no correlation is tested assuming normality of the Moran's I statistic under this null hypothesis. In brackets the obtained p-values are reported. From the results it is clear that the first and second variables are always positively spatially correlated, while the other two variables are spatially correlated only in certain years. However, overall, there is a significant correlation between the provinces and the observations cannot be considered independent. Therefore this data set represents an example of multivariate spatio-temporal data with dependent observations.

	2005	2006	2007	2008	2009
home robberies	0.54 (0.000)	0.56 (0.000)	0.57 (0.000)	0.56 (0.000)	0.50 (0.000)
teenage crime	0.56 (0.000)	0.57 (0.000)	0.46 (0.000)	0.49 (0.000)	0.46 (0.000)
robberies	0.08 (0.051)	0.10 (0.033)	0.08 (0.063)	0.08 (0.062)	0.06 (0.137)
muggings	0.22 (0.000)	0.17 (0.002)	0.08 (0.093)	0.12 (0.029)	0.10 (0.059)

Table 5: *Crime in the Italian provinces.* Moran's I autocorrelation coefficient. In brackets the p-values are reported.

We have modelled the territorial dependence through priors on the mixture weights and the temporal correlations among the five years with an AR(1) structure. What differentiates our cluster analysis from a classification on a single year only is the fact that we model simultaneously the correlations of variables within and between the different years. In fact it could easily happen that clustering of provinces observed in 2005 could be quite different from that obtained in 2009, since in the considered years the political action to reduce these criminal activities could have achieved different results across the provinces.

	home robberies	teenage crime	robberies	muggings
		$i = 1$		
2005	239.02	16.84	33.24	200.70
2006	263.27	19.37	34.34	196.97
2007	283.84	17.03	35.11	215.85
2008	327.22	17.45	37.91	248.82
2009	282.23	18.22	34.26	202.91
		$i = 2$		
2005	161.00	9.36	30.85	61.94
2006	166.29	10.00	29.22	74.03
2007	197.34	9.77	29.69	86.77
2008	226.33	9.91	33.01	92.10
2009	209.35	10.76	29.61	77.18
		$i = 3$		
2005	216.66	16.48	106.34	481.76
2006	242.83	21.35	104.06	546.84
2007	294.87	20.57	119.53	689.74
2008	322.60	20.16	129.16	676.33
2009	274.54	19.34	112.57	494.50
		$i = 4$		
2005	138.48	4.36	351.25	220.38
2006	147.51	6.25	341.74	243.79
2007	162.82	6.51	354.39	240.09
2008	182.91	7.88	315.71	236.51
2009	166.66	11.84	273.59	220.48

Table 6: *Crime in the Italian provinces.* Values of the four ($i = 1, 2, 3, 4$) component mean matrices.

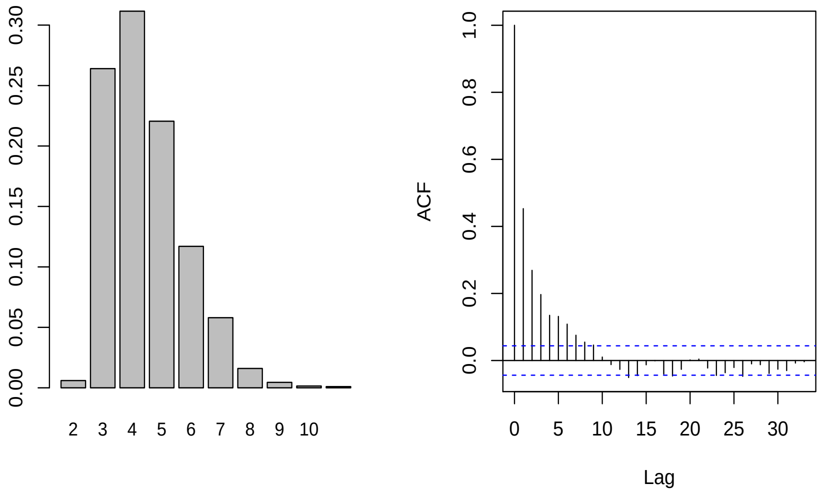

Figure 2: *Crime in the Italian provinces.* Results from BDMCMC algorithm to fit the GMMN model with $\lambda = 2$. The first plot shows the posterior probabilities of different values of k. The second graph depicts the autocorrelation plot for sampled values of k.

The analysis has been performed by running a BDMCMC chain with $\lambda = 2$, representing the prior belief that provinces could be clustered into safe areas and dangerous ones. We applied the BDMCMC algorithm to obtain a sample of size 20,000 from random starting points, and discarded the first 10,000 observations as burn-in. The mixing behavior of the chains over k was quite good since the percentage of sample points for which k changed was 67.23%. The posterior probabilities and the autocorrelation for sampled values of k are shown in Figure 2. As shown from the first graph, the mode is $k = 4$. Therefore a GMMN model with four components has been fitted to this data by running 20,000 iterations of the Gibbs sampler algorithm (with a burn in of 10,000 iterations). The estimated value of ζ is 0.12, thus denoting that a certain proportion of spatial dependence has affected the probabilities of group membership.

In order to interpret the estimated four groups of provinces, we can consider the component mean matrices of the GMMN classification, reported in Table 6, for the four groups ($i = 1, 2, 3, 4$).

As shown from the table, the first cluster is characterized by high values for home

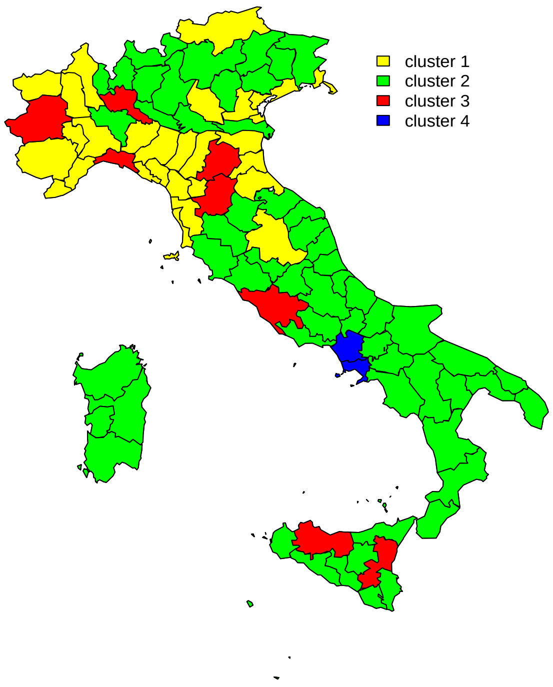

Figure 3: *Crime in the Italian provinces.* Province classification into $k = 4$ groups.

robberies and teenage crime and relatively low values for the other two measurements. This group consists of $n_1 = 31$ provinces. The estimated temporal correlation is $\beta_1 = 0.71$. On the contrary, the second cluster consists of $n_2 = 61$ relatively safe cities (all the crime measurements are lower than those of the other groups), with higher correlations between the years ($\beta_2 = 0.83$). In line with the economic and territorial differences mentioned above, the first cluster of provinces corresponds to some of the most industrialized and rich provinces of the North and Center of Italy, while provinces of the second cluster are mainly located in the Center and South of Italy. This clustering is shown in Figure 3, which represents the map of Italy with the 103 provinces. The third cluster includes the $n_3 = 9$ biggest and most touristic provinces, like Rome, Turin, Florence and Milan (see Figure 3). These are the provinces with the highest values of home robberies, teenage crimes and reported muggings, and therefore the most dangerous ones in terms of the crime indicators considered in this analysis. The temporal correlation is $\beta_3 = 0.66$. Cluster 4 consists of only two provinces (Naples and Caserta) of the South of Italy, which are notoriously and particularly unsafe in terms of robberies and muggings. The estimated temporal correlation for this last cluster is $\beta_4 = 0.55$.

7 Conclusions

We believe that the mixture model that has been introduced and discussed in this paper provides an interesting new tool for clustering the most common classes of three-way data which include, as examples, multivariate repeated measures, spatio-temporal data or longitudinal studies of multiple responses. The proposed GMMN has several advantages compared to the conventional statistical approaches for clustering three-way data. First, by defining two different covariance matrices for the two modes, it can describe the variable correlations separately from the temporal (or spatial) correlations. This also allows one to model structured temporal or spatial correlation. As a consequence, the GMMN model seems to be more flexible and it can outperform other clustering methods in terms of classification performance, as shown in the first real example. Second, the model can be easily adapted to deal with non-independent observations, like in the context of spatio-temporal data where the statistical units are represented by spatially correlated locations. Finally, and more importantly, the GMMN model has been developed in a Bayesian framework thus providing a tool for model estimation of several classes of three-way data and inference on the number of mixture components. Model inference can be obtained by means of a conventional MCMC approach based on the Gibbs sampling, while the issue of model selection has been addressed by constructing a Markov chain through the simulation of a continuous-time stochastic birth-and-death point process. From the computational point of view, the algorithm could be expensive and the computational burden increases as the dimensionality or the sample size increase. To give an example, running 10,000 iterations of the BDMCMC algorithm with $n = 500$, $r = 3$ and $p = 5$ in the simulation study took about an hour on R 2.8.1 under Microsoft Windows XP professional (CPU X86 Family 15, Model 4, ∼2666 Mhz). Of course the computational time could be considerably reduced under different platforms. However, in some circumstances it could be useful to apply some dimensionality reduction techniques to reduce the model complexity when r or p increases, without the need of collapsing the three-way structure into a two-way data set. Finally, it would be interesting to explore wider classes of covariance structures for modeling the spatial or temporal correlations in order to broaden the application fields of the proposed method.

Appendix: Posterior distributions

Full conditional for M

The full conditional (17) of the mean component matrix M_i is obtained by combining (10) and (7):

$$\begin{aligned}
f(M_i|\dots) \quad &\propto \quad \exp\left\{-\frac{1}{2}\sum_{j=1}^{n}\mathrm{tr}\left(\Psi_i^{-1}(Y_j-M_i)\Psi_i^{-1}(Y_j-M_i)^\top\right)z_{ij}\right\}\times\\
&\qquad \exp\left\{-\frac{1}{2}\mathrm{tr}\left(\Psi_0^{-1}(M_i-M_0)\Psi_0^{-1}(M_i-M_0)^\top\right)\right\}\\
&\propto \quad \exp\left\{-\frac{1}{2}\sum_{j=1}^{n}\mathrm{tr}\left(-2\Psi_i^{-1}Y_j\Omega_i^{-1}M_i^\top+\Psi_i^{-1}M_i\Omega_i^{-1}M_i^\top\right)z_{ij}\right\}\times\\
&\qquad \exp\left\{-\frac{1}{2}\mathrm{tr}\left(-2\Psi_0^{-1}M_0\Omega_0^{-1}M_i^\top+\Psi_0^{-1}M_i\Omega_0^{-1}M_i^\top\right)\right\}\\
&= \quad \exp\left\{-\frac{1}{2}\left[-2\mathrm{tr}\left(\Psi_i^{-1}\left(\sum_{j=1}^{n}Y_jz_{ij}\right)\Omega_i^{-1}+\Psi_0^{-1}M_0\Omega_0^{-1}\right)M_i^\top\right]\right\}\times\\
&\qquad \exp\left\{-\frac{1}{2}\mathrm{tr}\left(n_i\Psi_i^{-1}M_i\Omega_i^{-1}M_i^\top+\Psi_0^{-1}M_i\Omega_0^{-1}M_i^\top\right)\right\}.
\end{aligned}$$

Now by using the properties of trace we obtain

$$\begin{aligned}
&\exp\left\{-\frac{1}{2}\left[-2\mathrm{vec}(M_i)^\top\mathrm{vec}\left(\Psi_i^{-1}\left(\sum_{j=1}^{n}Y_jz_{ij}\right)\Omega_i^{-1}+\Psi_0^{-1}M_0\Omega_0^{-1}\right)\right]\right\}\times\\
&\exp\left\{-\frac{1}{2}\mathrm{vec}(M_i)^\top\left[n_i\Psi_i^{-1}\otimes\Omega_i^{-1}+\Psi_0^{-1}\otimes\Omega_0^{-1}\right]\mathrm{vec}(M_i)\right\},
\end{aligned}$$

which is proportional to the multivariate Gaussian distribution in (17).

Full conditional for β_i and σ_i

By putting (23) into (7) it is possible to derive the posteriors of the two parameters of interest β_i and σ_i given the other parameters and variables:

$$f(\beta_i, \sigma_i| \ldots) \propto (\sigma_i^2)^{-\frac{rp}{2}n_i} |R_i(\beta_i)|^{-\frac{p}{2}n_i} \exp\left[-\frac{1}{2}\sigma_i^{-2}\mathrm{tr}\left(R_i(\beta_i)^{-1}P_i\right)\right] f(\beta_i)f(\sigma_i)$$

with $P_i = \sum_{j:z_j=i}(Y_j - M_i)\Omega_i^{-1}(Y_j - M_i)^\top$. In order to obtain the full conditional for β_i, we consider $|R_i(\beta_i)| = (1-\beta_i^2)^{r-1}$ and

$$R_i(\beta_i)^{-1} = \frac{1}{(1-\beta_i^2)}\left(\mathbf{I}_r - \beta_i C_1 + \beta_i^2 C_2\right),$$

where C_1 is a tridiagonal matrix with 0 on the diagonal and 1 on the lower and upper diagonals and $C_2 = \mathrm{diag}(0, 1, \ldots, 1, 0)$. Therefore,

$$\mathrm{tr}\left(R_i(\beta_i)^{-1}P_i\right) = \frac{1}{(1-\beta_i^2)}\left(\mathrm{tr}(P_i) - \beta_i \mathrm{tr}(C_1P_i) + \beta_i^2\mathrm{tr}(C_2P_i)\right).$$

The analytical derivation of full conditional of β_i is

$$\begin{aligned} f(\beta_i|\ldots) \quad \propto \quad & \frac{1}{2}(1-\beta_i^2)^{\frac{1}{2}(r-1)pn_i} \\ & \exp\left[-\frac{1}{2}\frac{\sigma_i^{-1}}{1-\beta_i^2}\left(\mathrm{tr}(P_i) - \beta_i\mathrm{tr}(C_1P_i) + \beta_i^2\mathrm{tr}(C_2P_i)\right)\right]. \end{aligned}$$

This expression is not a known distribution but realizations from it can be generated according to a self-normalized importance sampling scheme.

The full conditional for σ_i can be obtained as follows:

$$\begin{aligned} f(\sigma_i|\ldots) \quad & \propto \quad (\sigma_i^{-2})^{\frac{rp}{2}n_i} \exp\left[-\frac{1}{2}\sigma_i^{-2}\mathrm{tr}\left(R_i(\beta_i)^{-1}P_i\right)\right](\sigma_i^{-2})^{a-1}\exp[-b\sigma_i^{-2}] \\ & = \quad (\sigma_i^{-2})^{\frac{rp}{2}n_i+a-1}\exp\left[-\frac{1}{2}\mathrm{tr}\left(R_i(\beta_i)^{-1}P_i + b\right)\sigma_i^{-2}\right], \end{aligned}$$

from which it follows

$$f(\sigma_i^{-1}|\ldots) \quad \sim \quad G\left(a + \frac{rp}{2}n_i, b + \frac{1}{2}\mathrm{tr}(R_i(\beta_i)^{-1}P_i)\right),$$

where G represents the Gamma distribution.

Full conditionals for $x_1, \ldots, x_k$ and ζ

In order to derive the posterior distribution for the spatial latent variables, they are regarded as n vectors of length k and updated sequentially. By considering prior (24) into (7) we get

$$\begin{aligned} f(x_j|\ldots) &\propto \prod_{i=1}^{k}\left\{\left[\frac{e^{x_{ij}}}{\sum_{h=1}^{k} e^{x_{hj}}}\right]^{z_{ij}} \exp\left[-\frac{1}{2}\left(\zeta\sum_{j\sim j'}(x_{ij}-x_{ij'})^2+\sum_{j=1}^{n}x_j^2\right)\right]\right\} \\ &\propto \prod_{i=1}^{k}\left\{\left[\frac{e^{x_{ij}}}{\sum_{h=1}^{k} e^{x_{hj}}}\right]^{z_{ij}} \exp\left[-\frac{1}{2}\zeta\sum_{j\sim j'}(x_{ij}-x_{ij'})^2\right]\prod_{j=1}^{n}\exp\left[-\frac{1}{2}x_j^2\right]\right\} \\ &\propto \prod_{i=1}^{k}\left\{\left[\frac{e^{x_{ij}}}{\sum_{h=1}^{k} e^{x_{hj}}}\right]^{z_{ij}} \exp\left[-\frac{1}{2}\zeta\sum_{j\sim j'}(x_{ij}-x_{ij'})^2\right]\exp\left[-\frac{1}{2}x_j^2\right]\right\}. \end{aligned}$$

The product between the last two terms within the bracket is the product between two normals which is still normally distributed. Therefore we obtain

$$f(x_j|\ldots) \propto \prod_{i=1}^{k}\left\{\left[\frac{e^{x_{ij}}}{\sum_{h=1}^{k} e^{x_{hj}}}\right]^{z_{ij}} \mathcal{N}\left(\frac{\zeta\sum_{j\sim j'}x_{ij'}}{1+\zeta\upsilon_j},\frac{1}{1+\zeta\upsilon_j}\right)\right\}.$$

The full conditional for ζ simply derives by evaluating (7) and takes the form

$$f(\zeta|\ldots) \quad\propto\quad \left(\prod_{j=1}^{n}(1+\zeta\upsilon_j)^{k/2}\right)\exp\left[-\frac{\zeta}{2}\sum_{i=1}^{k}\sum_{j\sim j'}(x_{ij}-x_{ij'})^2\right].$$

References

Basford, K. E. and McLachlan, G. J. (1985). "The Mixture Method of Clustering applied to three-way data." *Journal of Classification*, 2: 109–125.

Baudry, J. P., Raftery, A. E., Celeux, G., Lo, K., and Gottardo, R. (2008). "Combining mixture components for clustering." *University of Washington, Seattle*, (Technical report 540).

Cappé, O., Robert, C., and Rydén, T. (2002). "Reversible jump MCMC converging to birth-and-death MCMC and more general continuous time samplers." *Journal of the Royal Statistical Society - Series B*, 65: 679–700.

Celeux, G. (1998). "Bayesian Inference for Mixtures: The Label-Switching Problem." In Payne, R. and Green, P. (eds.), *COMPSTAT 1998*, 227–232. Heidelberg and Vienna: Physica-Verlag.

Chang, W. C. (1983). "On using principal components before separating a mixture of two multivariate normal distributions." *Applied Statistics*, 32: 267–275.

Dellaportas, P. and Papageorgiou, I. (2006). "Multivariate mixtures of normals with unknown number of components." *Statistics and Computing*, 16: 1573–1375.

Dempster, N. M., Laird, A. P., and Rubin, D. B. (1977). "Maximum likelihood from incomplete data via the EM algorithm (with discussion)." *Journal of the Royal Statistical Society - Series B*, 39: 1–38.

Dutilleul, P. (1999). "The MLE algorithm for the matrix normal distribution." *Journal of Statistical Computation and Simulation*, 64: 105–123.

Fernández, C. and Green, P. J. (2002). "Modelling spatially correlated data via mixtures: a Bayesian approach." *Journal of the Royal Statistical Society - Series B*, 64: 805–826.

Fraley, C. and Raftery, A. E. (2002). "Model-based clustering, discriminant analysis, and density estimation." *Journal of the American Statistical Association*, 97: 611–631.

— (2006). "MCLUST Version 3 for R: Normal Mixture Modeling and Model-based Clustering." *Department of Statistics, University of Washington*, (Technical report 504).

Gelfand, A. E., Kottas, A., and MacEachern, S. N. (2005). "Bayesian Nonparametric Spatial Modeling With Dirichlet Process Mixing." *Journal of the American Statistical Association*, 100: 1021–1035.

Gelman, A. and Rubin, D. B. (1992). "Inference from iterative simulation using multiple sequences (with discussion)." *Statistical Science*, 7: 457–511.

Geman, S. and Geman, D. (1984). "Stochastic relaxation, Gibbs distributions and the Bayesian restoration of images." *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 6: 721–741.

Gittleman, J. L. and Kot, M. (1990). "Adaptation: statistics and a null model for estimating phylogenetic effects." *Systematic Zoology*, 39: 227241.

Gordon, A. D. and Vichi, M. (1998). "Partitions of Partitions." *Journal of Classification*, 15: 265–285.

Green, P. J. (1995). "Reversible jump Markov chain, Monte Carlo computation and Bayesian model determination." *Biometrika*, 82: 711–732.

Hand, D. J. and Taylor, C. C. (1987). *Multivariate analysis of variance and repeated measures*. London: Chapman & Hall.

Hennig, C. (2010). "Methods for merging Gaussian mixture components." *Advances in Data Analysis and Classification*, 4: 3–34.

Jennrich, R. I. and Schluchter, M. D. (1986). "Unbalanced repeated-measures models with structured covariance matrices." *Biometrics*, 42: 305–320.

Joe, H. (2006). "Generating Random Correlation Matrices Based on Partial Correlations." *Journal of Multivariate Analysis*, 97: 2177–2189.

Nel, H. M. (1977). "On distributions and moments associated with matrix normal distributions." *Mathematical Statistics Department, University of the Orange Free State, Bloemfontein, South Africa*, (Technical report 24).

R Development Core Team (2008). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0.
URL http://www.R-project.org

Reich, B. J. and Bondell, H. D. (2010). "A Spatial Dirichlet Process Mixture Model for Clustering Population Genetics Data." *Biometrics*, DOI: 10.1111/j.1541–0420.2010.01484.x.

Richardson, S. and 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: 731–792.

Stephens, M. (2000a). "Bayesian analysis of mixtures models with an unknown number of components - an alternative to reversible jump methods." *Annals of Statistics*, 28: 40–74.

— (2000b). "Dealing with label-switching in mixture models." *Journal of the Royal Statistical Society - Series B*, 62: 795–809.

Tadesse, M. G., Sha, N., and Vannucci, M. (2005). "Bayesian variable selection in clustering high-dimensional data." *Journal of the American Statistical Association*, 100: 602–617.

Vichi, M. (1999). "One mode classification of a three-way data set." *Journal of Classification*, 16: 27–44.

Vichi, M., Rocci, R., and Kiers, A. L. (2007). "Simultaneous Component and Clustering Models for three-way data: within and between approaches." *Journal of Classification*, 24: 71–98.

Viroli, C. (2011). "Finite mixtures of matrix normal distributions for classifying three-way data." *Statistics and Computing*, 21: 511–522.

Zhang, Z., Chan, K. L., Wu, Y., and Chen, C. (2004). "Learning a multivariate Gaussian mixture model with the reversible jump MCMC algorithm." *Statistics and Computing*, 14: 343–355.