Model based clustering of multivariate spatio-temporal data: a matrix-variate approach

Cinzia Viroli

Abstract Multivariate spatio-temporal data arise from the observation of a set of measurements in different times on a sample of spatially correlated locations. They can be arranged in a three-way data structure characterized by rows, columns and layers. In this perspective each observed statistical unit is a matrix of observations instead of the conventional p-dimensional vector. In this work we propose model based clustering for this wide class of continuous three-way data by a general mixture model with components modelled by matrix-variate Gaussian distributions. The effectiveness of the proposed method is illustrated on multivariate crime data collected on the Italian provinces in the years 2005-2009.

Key words: Gibbs sampler, random matrix, three-way data

1 Introduction

Multivariate spatio-temporal data occur in different scientific fields from the observation of multiple measurements in repeated situations (or times) on spatially correlated locations. Due to their complexity they cannot be arranged in a conventional matrix of observations but they may be represented in a three-way data structure, where the, say n, locations are represented in rows, a set of p variables are indexed in columns and the different, say r, times are the layers.

We focus on the problem of clustering the n spatially correlated locations. By denoting with j the generic observation (where $j = 1,\ldots,n$), 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

Dipartimento di Scienze Statistiche, Università di Bologna, e-mail: cinzia.viroli@unibo.it

the joint objectives of modeling the spatial correlation between the observations (since units are not *i.i.d.*), defining two different covariance matrices for describing the variable correlations separately from the temporal (or spatial) correlations and modeling possible temporal correlation structures.

Different solutions for clustering three-way data have been proposed in the statistical literature. 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 and Rocci, 2007). These methodologies do not require an explicit distributional assumption on the clusters and therefore they 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 would imply that 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 (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 (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.

In this work we propose a generalized MMN model (GMMN) for properly modelling multivariate spatio-temporal data in a Bayesian framework. This has the advantage of extending MMN in order to model spatially correlated observations and temporal structured covariance matrices. Model inference is solved via the Gibbs sampler (Geman and Geman, 1984).

2 Generalized mixture of matrix-normals

Let $Y_1, \ldots, Y_n$ be the dependent location matrices of dimension $r \times p$. They are assumed to belong to a set k of sub-populations or groups of unknown proportions. In a general perspective, we consider the observed sample of n matrices as a set of conditionally independent and not identically distributed observations coming from the mixture model

$$f(Y_j|k,\pi,\Theta_1,\dots,\Theta_k)=\sum_{i=1}^{k}\pi_{ij}\mathscr{M}_{(r\times p)}(Y_j;\Theta_i), \tag{1}$$

where $j=1,\dots,n$ and Θ_i denotes the set of parameters of each component distribution. The weights $\pi=[\pi_{ij}]_{i=1,\dots,k;j=1,\dots,n}$ satisfy $\pi_{ij}>0$ with $\sum_{i=1}^{k}\pi_{ij}=1$ for all j. They vary with j in order to take into account the spatial correlation. 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 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 covariance structures. 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 $\Theta_i=\{M_i,\Phi_i,\Omega_i\}$:

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

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.

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. In this work 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

$$\Phi_i(\beta_i)=(\sigma_i\mathbf{I}_r)R_i(\beta_i)(\sigma_i\mathbf{I}_r), \tag{3}$$

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

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

2.1 Hierarchical formulation of GMMN

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 from which $f(z_{ij}=1|\pi,k)=\pi_{ij}$. In order to deal with correlated observations, we assume that $Y_1,\dots,Y_n$ are independent given the set of latent variables $\mathbf{z}=\{z_1,\dots,z_n\}$.

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

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

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

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

We allow additional layers to the hierarchy by adding a set of hyperparameters ω for Θ and π. In so doing we formulate the estimation problem in a Bayesian framework. In the GMMN model the distribution of interest is the posterior distribution of the allocation variables, of the parameters and hyperparameters (for fixed k) given the observed data Y. By using formulation (5) it can be expressed as

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

where $f(\pi|\omega,k)$, $f(\Theta|\omega,k)$ and $f(\omega|k)$ are the prior distributions of parameters and hyperparameters.

2.2 Prior formulation and hyperparameters

In this setting the set of hyper-parameters is $\omega=(\beta_1,\dots,\beta_k,\sigma_1,\dots,\sigma_k,\rho,x_1,\dots,x_k,\zeta)^\top$, where x_i (with $i=1,\dots,k$) are spatial latent variables. Each x_i is a Markov random field with density function:

$$f(x_i|\zeta)=(2\pi)^{-n/2}\prod_{j=1}^{n}(1+\zeta\upsilon_j)^{1/2}\exp\left[-\frac{1}{2}\left(\zeta\sum_{j\sim j'}(x_{ij}-x_{ij'})^2+\sum_{j=1}^{n}x_j^2\right)\right] \tag{7}$$

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. The hyperparameter 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 $x_1,\ldots,x_k$ and ζ the weights for location j take the form

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

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 $i = 1,\ldots,k$. 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. Finally, the role of ρ is to parameterize the prior distributions of Ω_i for all i. We can choose non-informative prior distributions for ρ and the model parameters. More precisely:

$$M_i \sim \mathscr{M}_{(r\times p)}(M_0, \Phi_0, \Omega_0) \tag{8}$$

$$\Omega_i^{-1}|\rho \sim \mathscr{W}_p\left(2\zeta, (2\rho)^{-1}\right) \tag{9}$$

$$\rho \sim \mathscr{W}_p\left(2l, (2m)^{-1}\right) \tag{10}$$

$$\tag{11}$$

for $i = 1,\ldots,k$. In the previous expressions $\mathscr{M}_{(r\times p)}$ denotes the matrix-variate normal distribution of order $r\times p$ and $\mathscr{W}$ denotes the multivariate Wishart distribution. Moreover, Φ_0 is a $r\times r$ matrix, ρ, Ω_0 and m are $p\times p$ matrices, M_0 is an $r\times p$ matrix and ζ and l are scalars.

With this setting the full conditionals are proportional to known distribution and a Gibbs sampler algorithm can be applied. The full conditional for $\mathbf{z}$ is

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

The posterior distributions of $x_1,\ldots,x_k$ and ζ are

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

and

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

The analytical derivation of full conditional of β_i is

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

where $P_i = \sum_{j:z_j=i}(Y_j - M_i)\Omega_i^{-1}(Y_j - M_i)^\top$, C_1 is a tridiagonal matrix with 0 on the diagonal and 1 on the lower and upper diagonals and $C_2 = \mathrm{diag}(0,1,\ldots,1,0)$. 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:

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

from which it follows

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

where G represents the Gamma distribution. Finally, the other posterior distributions can be analytically derived by combining equation (6) with the priors previously described:

$$\begin{aligned} &\mathrm{vec}(M_i)|\ldots \sim \mathcal{N}_{rp}\left(\Upsilon^{-1}\xi, \Upsilon^{-1}\right) \\ &\Omega_i^{-1}|\ldots \sim \mathcal{W}_p\left(2\rho + rn_i, \left[2\zeta + \sum_{j:z_j=i}(Y_j - M_i)^\top \Phi_i^{-1}(Y_j - M_i)\right]^{-1}\right) \\ &\rho|\ldots \sim \mathcal{W}_p\left(2l + 2k\zeta, \left[2m + 2\sum_{i=1}^{k}\Omega_i^{-1}\right]^{-1}\right) \\ &\pi|\ldots \sim \mathcal{D}(\rho + n_1, \ldots, \rho + n_k) \end{aligned}$$

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 $\Upsilon = n_i(\Phi_i \otimes \Omega_i)^{-1} + (\Phi_0 \otimes \Omega_0)$.

2.3 Example: 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.

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

A GMMN model with k components ranging from 2 to 5, has been fitted to this data by running 20,000 iterations of the Gibbs sampler algorithm (with a burn in of 10,000 iterations). For space reasons, we describe here the $k = 4$ solution. 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 1, for the four groups ($i = 1,2,3,4$).

As shown from the table, the first cluster is characterized by high values for home 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. The third cluster includes the $n_3 = 9$ biggest and most touristic provinces, like Rome, Turin, Florence and Milan. 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 $\beta_4 = 0.55$.

References

1. Basford, K.E. & Mclachlan, G.J., 1985, The Mixture Method of Clustering applied to three-way data. *Journal of Classification*, 2, 109-125.
2. Dutilleul, P., 1999, The MLE algorithm for the matrix normal distribution. *Journal of Statistical Computation and Simulation*, 64, 105-123.
3. Fernández C. & Green, P.J., 2002, Modelling spatially correlated data via mixtures: a Bayesian approach. *Journal of the Royal Statistical Society - Series B*, 64, 805-826.
4. Gelfand, A.E. & Kottas, A. & MacEachern, S. N., 2005, Bayesian Nonparametric Spatial Modeling With Dirichlet Process Mixing. *Journal of the American Statistical Association*, 100, 1021-1035.
5. Geman, S, & Geman, D., 1984, Stochastic relaxation, Gibbs distributions and the Bayesian restoration of images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 6, 721-741.
6. Gordon, A.D. & Vichi, M., 1998, Partitions of Partitions. *Journal of Classification*, 15, 265-285.
7. Vichi, M., 1999, One mode classification of a three-way data set. *Journal of Classification*, 16, 27-44.
8. Vichi, M. & Rocci, R. & Kiers, A.L., 2007, Simultaneous Component and Clustering Models for three-way data: within and between approaches. *Journal of Classification*, 24, 71-98.
9. Viroli, C., 2011, Finite mixtures of matrix normal distributions for classifying three-way data. *Statistics and Computing*, 21, 511-522.