Introducing spatio-temporal dependence in clustering: from a parametric to a nonparametric approach.

Introduzione di dipendenza spazio-temporale nel problema di clustering: da un approccio parametrico a un approccio non-parametrico

Clara Grazian, Gianluca Mastrantonio and Enrico Bibbona

Abstract A huge literature about clustering spatial time data exists. The problem has been studied both in a parametric and in a nonparametric setting. There are several problems in defining a proper clustering procedure, depending on the type of relationship between the clusters. From a parametric point of view, a classic approach is to introduce mixture models and studying the posterior distribution of the mixture weights. We propose a mixture model where the mixing probabilities are time specific and are assumed to follow a Logistic-Normal distribution. We introduce dependence between the vectors of mixing probabilities by means of a Gaussian processes representation. We briefly propose a nonparametric extension of this approach.

Abstract *Esiste un'estesa letteratura riguardante problemi di clustering temporale e spaziale. Il problema è stato studiato sia in ambito parametrico che nonparametrico. I problemi principali delle procedure di clustering spazio-temporali dipendono dal tipo di dipendenza tra i cluster stessi. Da un punto di vista parametrico, l'approccio classico è quello di assumere un modello mistura e studiare la distribuzione a posteriori dei pesi della mistura. In questo articolo proponiamo ancora un modello mistura dove i pesi dipendono dal tempo e dove si assume che seguano un modello logistico-normale. La dipendenza temporale tra i cluster è introdotta attraverso una rappresentazione in base a processi gaussiani. Nell'articolo proponiamo brevemente anche un'estensione nonparametrica dell'approccio.*

Key words: Gaussian processes, logitN, temporal clustering

Clara Grazian
Clara Grazian, University of Oxford, e-mail: `clara.grazian@ndm.ox.ac.uk`

Gianluca Mastrantonio
Politecnico di Torino e-mail: `gianluca.mastrantonio@polito.it`

Enrico Bibbona
Politecnico di Torino e-mail: `enrico.bibbona@polito.it`

1 Introduction

There is more and more interest in spatial data, i.e. data where the response variable is measured at spatial locations or data where the response variable is defined as a set of spatial coordinates. This is due to the increased ability to store and collect this type of data.

A central problem in the analysis of spatial data is spatial clustering, which groups similar spatial objects into classes. Standard applications are the identification of land areas for usage purposes in agricultural sciences or weather patterns in environmental sciences. The goal of spatial clustering may be multiple, focusing on the study of the characteristics of each cluster and on a better understanding and description of the data and, ultimately, influencing policies in public health and environment. Due to its huge usefulness in applied sciences, spatial clustering has been a very active subject, with many contributions from different fields.

In this work, we will focus on the problem of modelling spatial coordinates (or transformation of them) and, in particular, clustering them through time. Direct applications of this may be seen in the modelling of wind directions [16], ocean streams [8], identification of three-dimensional protein structures [12] and animal movements [3].

The problem of clustering in this settings relates to the description of structural changes in the time series. For instance, it is generally assumed that the animal behaviour changes according to a natural cycle in the observational period, for example the resting/feeding cycle, or to out-of-ordinary situations, such as the assault of a predator or changement in human activities.

The joint distribution of the coordinates, or of an appropriate transformation of them, can be seen as a mixture process where the mixture components are the different behaviours or regimes. It is generally assumed that the switching between regimes is temporally structured [17], and sometime also spatially, as in [2], often ruled by a non-observed Markov process leading to the class of hidden Markov models (HMMs) [25].

In this paper, we propose a mixture-type model, as the hidden Markov model, but with a higher level of flexibility given by the assumption that the vector of probabilities is marginally distributed as a Logistic-Normal model (*LogitN*) [1] and the structured temporal dependence is induced via a coregionalization over a multivariate Gaussian process [10]. This is not the first proposal where a structured dependence is introduced over vectors *LogitN* distributed, however, as we will show, our proposal is more general since it focuses on the dependence structure of the probabilities vectors rather than that of the Gaussian process, as in [22] and [4].

The rest of the paper is organized as follows: Section 2 provides the notation and gives some preliminaries, Section 3 describes the proposed method; Section 4 focuses on a real application on data of animal movement. Section 5 concludes the paper.

2 Notation and preliminaries

Suppose the data are indexed by temporal indices $(t_1,\dots,t_T)' \equiv \mathscr{T}$, assuming that t_1 is the starting observational time and t_T the ending observational time, by allowing that $t_i - t_{i-1} = c_i$ which may vary depending on i.

Let, then, $\mathbf{s}_{t_i} = (s_{t_i,1}, s_{t_i,2})' \in \mathbb{R}^2$ be spatial location at time t_i with $\mathbf{s} = (\mathbf{s}_{t_1},\dots,\mathbf{s}_{t_T})'$.

A standard transformation considered when analysing spatial coordinates is looking at the projections of $\mathbf{s}_{t_i}$ on the x- and y- axes of the cartesian coordinates system centred on $\mathbf{s}_{t_{i-1}}$, say $\mathbf{r}_{t_i} = (r_{t_i,1}, r_{t_i,2}) \in \mathbb{R}^2$ and then defining

$$\mathbf{y}_{t_i} = \frac{\mathbf{r}_{t_i}}{d(t_{i+1},t_i)}.$$

The variable $\mathbf{y}_{t_i}$ contains all the sufficient information to recover the trajectory without loosing any significant property. In particular, the sign of $y_{t_i,1}$ provides information about the fact that the movement is on the same direction of the previous one, while the sign of $y_{t_i,2}$ indicates if it turns to the right or to the left. Moreover, $||\mathbf{y}_{t_i}||$ represents the step-length and $\theta_{t_i} = \text{atan2}(y_{t_i,2}, y_{t_i,1})$ the turning angle of the movement, respectively.

A standard clustering methodology is to introduce information indicating the cluster membership as a latent variable $\mathbf{z} = \{z_t\}_{t\in\mathscr{T}}$, with $z_t \in \{1,2,\dots,K\} \equiv \mathscr{K}$, and K the total number of clusters.

The data are assumed to come from a mixture-type model based on bivariate Gaussian densities:

$$f(\mathbf{y}|\mathbf{z}\{\xi_k,\Omega_k\}) = \prod_{t\in\mathscr{T}} f(\mathbf{y}_t|\xi_{z_t},\Omega_{z_t})$$

where

$$\mathbf{y}_t|\xi_{z_t},\Omega_{z_t} \sim N_2(\xi_{z_t},\Omega_{z_t})$$

i.e. given the latent variables $\mathbf{z}_t$, the observations $\mathbf{y}_t$ are independent. The hidden Markov models (HMMs) [5] assume the latent variables follow the Markov rule

$$\Pr(z_{t_i} \mid z_{t_1},\cdots,z_{t_{i-1}}) = \Pr(z_{t_i} \mid z_{t_{i-1}})$$

and, in particular,

$$z_{t_i}|z_{t_{i-1}},\{\pi_{k,k'}\}_{k,k'\in\mathscr{K}} \sim \sum_{k\in\mathscr{K}} \pi_{z_{t_{i-1}},k}\delta_k.$$

Although HMMs are widely used ans easy to implement, the Markov structure may be too restrictive. For instance, the assumption that the probabilities of switching cluster is fixed and independent from time is difficult to accept in all the contexts. In some works, see for example [19], [17], [13] and [11], problems like these are tackled using covariates that model probabilities, but not always these are available and may not be enough for describing the complexity of reality. [18] proposes a mixture model where the latent spatial process is allowed to evolve dynamically over time.

We proposed to consider a more complex model where the mixing probabilities follow a *LogitN* model; this has been proposed by [1] as a distribution for independent compositional data, i.e. vectors of positive proportions with a constant sum. Beyond the obvious fact that it can model data which are positive and less than a constant, the constant-sum constraint induces some more insidious characteristics which will be described in the next Section.

3 The model

We propose to introduce a dependence on time of the vector of probabilities, instead of using a HMM

$$z_t|\pi_t \sim \sum_{k=1}^{K} \pi_{t,k}\delta_k,$$

where $0 \leq \pi_{t,k} \leq 1$, $\sum_{k=1}^{K} \pi_{t,k} = 1$. The probability vector π_t is then defined with the logistic transformation of real valued variables

$$\pi_{t,k} = \frac{e^{\omega_{t,k}}}{1+\sum_{j=1}^{K} e^{\omega_{t,j}}}, k = 1,\dots,K \tag{1}$$

with the last element defined as

$$\pi_{t,K} = \frac{1}{1+\sum_{j=1}^{K-1} e^{\omega_{t,j}}}.$$

It has to be noticed that adding a constant c to each $\omega_{t,k}$ produces the same vector of probabilities, and then an identifiability constraint is needed; without loss of generality, we set to zero the K^{th}.

In this way, it is possible to model the latent variable ω_t. In particular, we propose to define it as

$$\omega_t = (\mathbf{I}_{K-1} \otimes \mathbf{X}_t)\beta + \mathbf{A}\eta_t \tag{2}$$

where $\mathbf{I}_d$ is the identity matrix of dimension $d \times d$, $\mathbf{X}_t$ is a set of p covariates changing over time and β the corresponding coefficients such that each $\omega_{t,k}$ shares the same set of covariates but different coefficient β. $\mathbf{A}$ is the term introducing dependence among the vectors $\omega_{\cdot,k} = \{\omega_{t,k}\}_{t\in\mathscr{T}}$, while η_t is the error term such that it is defined as a Gaussian process, i.e. $\eta_{\cdot k} \sim GP(\mathbf{0}, C_k(\cdot;\psi_k))$ where C_k models the structured dependence. Consequently, the covariance between ω_t and $\omega_{t'}$ is given by $\Sigma_{t,t'} = \mathbf{A}\mathbf{C}_{\eta,|t-t'|}\mathbf{A}'$.

We say that marginally the vector of probabilities π_t follows a logit-normal model [1], i.e. $\pi_t \sim LogitN(\mu_t, \Sigma)$, where $\mu_t = (\mathbf{I}_{K-1} \otimes \mathbf{X}_t)\beta$ and $\Sigma = \mathbf{A}\mathbf{A}'$.

The nature of a compositional vector, in particular its property to sum to a constant term, implies a lack of interpretability since the correlations are not completely free to vary in $(-1,1)$. We propose a different parametrisation to deal with this problem.

Instead of modelling ω_t, we propose to introduce another variable

$$\gamma_t = (\mathbf{I}_K \otimes \mathbf{X}_t)\,\beta + \mathbf{A}^* \eta_t$$

in similar way than ω_t, with the difference that all the K variables are defined, i.e. $\mathbf{I}_K$ is the identity matrix of dimension $K \times K$, $\mathbf{A}$ is a $K \times K$ matrix introducing dependence among the $\gamma_{.,k} = \{\gamma_{t,k}\}_{t \in \mathscr{T}}$ while η_t is the error term such that $\eta_{.k} \sim GP(\mathbf{0}, C_k^*(\cdot;\psi_k^*))$. Then, γ_t has covariance matrix given by

$$\Sigma^* = \mathbf{A}^*(\mathbf{A}^*)'.$$

Therefore, it is possible to define a different vector of probabilities with elements

$$\pi^*_{t,k} = \frac{\exp(\gamma_{t,k})}{\sum_{j=1}^K \exp(\gamma_{t,j})}.$$

It is evident, however, that the model is not identifiable. Notice that π^*_t follows a marginal distribution different from π_t, with a different temporal dependence.

However, if we let

$$\omega_{t,k} \coloneqq \gamma_{t,k} - \gamma_{t,K}$$

we create a link between the two parametrizations that induces

$$\mathbf{A} = [\mathbf{A}^*]_{1:(K-1),1:K} - [\mathbf{A}^*]_{K,1:K},$$

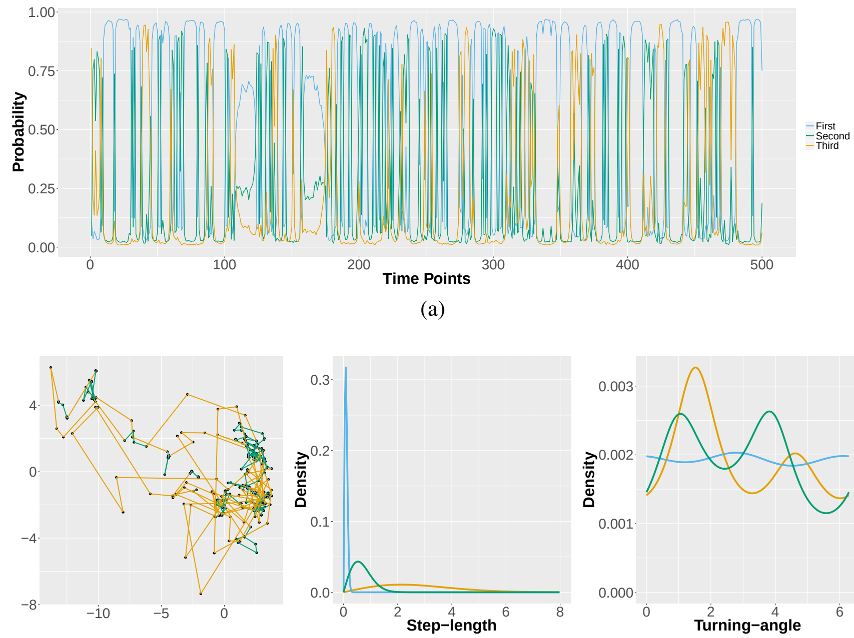
and, consequently,

$$[\Sigma^*]_{1:(K-1),1:(K-1)} + \mathbf{1}_{K-1}[\Sigma^*]_{K,K}\mathbf{1}'_{K-1} - [\Sigma^*]_{1:(K-1),K}\mathbf{1}'_{K-1} - \mathbf{1}_{K-1}[\Sigma^*]'_{1:(K-1),K}.$$

Given this link, it follows that $\pi_{t,k} = \pi^*_{t,k}$ for any $k = 1,\cdots,K$ and $t = 1,\cdots,T$, with the advantage to be able to define a covariance structure on the variables γ_t which induces the desired covariance structure on the corresponding π_t.

There are several interesting issues to highlight at this point. First, our proposed model given in (2) generalizes most of the models available in the literature where transformations as in (1) are used to define probabilities vectors. For instance, the proposal of [15] is obtained by assuming $\eta_t \equiv \mathbf{0}_{K-1}$. The model of [18] is obtained letting η_t be a spatio-temporal process with autoregressive temporal increments and a diagonal $\mathbf{A}$. We can also reduce to the proposals of [24], [14] and [21]. On the other hand, models using the cokriging, such as the ones of [20], can not be expressed with our formulation. However, one may notice that the complexity of our approach is reduced with respect to the cokriging.

Secondly, computational issues often arise for models based on Gaussian processes. We make use of the approach proposed by [7], i.e. a scalable nearest-



(a)

(b) (c) (d)

Fig. 1: Real data: (a) probabilities time series; (b) observed trajectory; (c) posterior estimates of step-length; (d) posterior estimates of turning-angle

neighbour Gaussian process (NNGP) which may be seen as a hierarchical sparse prior and allows for efficient MCMC algorithms to be performed without storing or decomposing large covariance matrices. While [7] empirically shows that 25 neighbours are needed to obtain an approximation close to the complete model, the temporal nature of the data analysed in this paper allow us to use just one neighbour, with a even highest level of saving of computational time.

4 Real Data

Data on 6 free-ranging Maremma sheepdogs positions are recorded by tracking collars every 30 minutes. The behaviour of the dogs is unknown because there is minimal supervision by their owners and the animals are allowed to range freely.

We select a time series of 500 points of one dog and characterize the hidden behaviours using the model proposed above; in Figure 1 (b) we show the observed trajectories, i.e. the observed coordinates.

We estimate model with values of K in $[2,3,\ldots,10]$ and, using DIC_5 [6], we select the best model as the one with three components. The three behaviours can be easily characterized looking at the distribution of step-length and turning-angle, Figure 1 (b) and (c). In the first behaviour, the dog has low speed and the distribution of the angle is almost circular uniform, i.e. it moves randomly. In the second one, the speed increases and the direction has two modes with same height, one at $\pi/2$ and one at $3\pi/2$, meaning that it changes direction clockwise and anticlockwise with the same probability. In the last one, the speed is high and it changes direction mainly anticlockwise. From Figure 1 (a) we can see the temporal series of the probabilities and we appreciate that behaviour one is the one that has more temporal stabilities, i.e. it has high values for a long time-period. In Figure 1 (b) we see where the behaviours are spatially localized. The third one occurs in all spatial domain, while the other two behaviours are on the top left (where the house of the owner is localized) and the right part of the map (where there is the livestock).

5 Conclusion

In this work, we propose a model to perform clustering for spatial data, based on a new parametrization of the logit-normal process, with parameters allowing an easier interpretability.

However, the possibility to use this model derives from the knowledge of the exact number of clusters. The nonparametric extension of a mixture model is the Dirichlet process (DP) [9], a stochastic process defined over a measurable space whose random paths are probability measures with probability one. The hierarchical Dirichlet process (HDP), proposed by [23], is an extension of the DP that makes possible to have processes sharing the same set of atoms. In the original construction of the HDP the random vectors $\pi_{j.}$ and $\pi_{k.}$ for any $j \neq k$ are independent. The nonparametric extension of the work proposed in this paper, which will be the subject of further research, focuses on how to introduce dependence in hierarchical Dirichlet processes.

References

1. Aitchison, J. (1986). *The Statistical Analysis of Compositional Data*. London, UK, UK: Chapman & Hall, Ltd.
2. Blackwell, P. G. (2003). Bayesian inference for Markov processes with diffusion and discrete components. *Biometrika*, 90(3): 613–627.
3. Brownlee, J. (1912). The mathematical theory of random migration and epidemic distribution. *Proceedings of the Royal Society of Edinburgh*, 31: 262–289.
4. Brunsdon, T. M. and Smith, T. (1998). The time series analysis of compositional data. *Journal of Official Statistics*, 14(3): 237.
5. Cappé, O., Moulines, E., and Rydén, T. (2009). Inference in Hidden Markov Models. *Proceedings of EUSFLAT Conference.*
6. Celeux, G., Forbes, F., Robert, C. P., Titterington, D. M., Futurs, I., and Rhône-alpes, I. (2006). Deviance information criteria for missing data models. *Bayesian Analysis*, 4: 651–674.

7. Datta, A., Banerjee, S., Finley, A. O., and Gelfand, A. E. (2016). Hierarchical Nearest-Neighbor Gaussian Process Models for Large Geostatistical Datasets. *Journal of the American Statistical Association*, 111(514): 800–812.
8. Ewing, J. A. (1990). Wind, wave and current data for the design of ships and offshore structures. *Marine Structures*, 3(6): 421–459.
9. Ferguson, T. S. (1973). A Bayesian Analysis of Some Nonparametric Problems. *The Annals of Statistics*, 1(2): 209–230.
10. Gelfand, A., Diggle, P., Fuentes, M., and Guttorp, P. (2010). *Handbook of Spatial Statistics*. Chapman and Hall.
11. Gupta, M., Qu, P., and Ibrahim, J. G. (2007). A temporal hidden Markov regression model for the analysis of gene regulatory networks. *Biostatistics*, 8(4): 805–820.
12. Hughes, G. (2007). Multivariate and time series models for circular data with applications to protein conformational angles. *Doctoral dissertation, University of Leeds*.
13. Lagona, F., Maruotti, A., and Picone, M. (2011). A non-homogeneous hidden Markov model for the analysis of multi-pollutant exceedances data. In *Hidden Markov Models, Theory and Applications*. InTech.
14. Martins, A. B. T., Bonat, W. H., and Ribeiro Jr, P. J. (2016). Likelihood analysis for a class of spatial geostatistical compositional models. *Spatial Statistics*, 17: 121 – 130.
15. Maruotti, A., Bulla, J., Lagona, F., Picone, M., and Martella, F. (to appear). Dynamic mixture of factor analyzers to characterize multivariate air pollutant exposures. *Annals of Applied Statistics*, 11(3), 1617–1648.
16. Masseran, N., Razali, A. M., Ibrahim, K., and Latif, M. T. (2013). Fitting a mixture of von Mises distributions in order to model data on wind direction in Peninsular Malaysia. *Energy Conversion and Management*, 72: 94–102.
17. Morales, J. M., Haydon, D. T., Frair, J., Holsinger, K. E., and Fryxell, J. M. (2004). Extracting more out of relocation data: building movement models as mixtures of random walks. *Ecology*, 85(9): 2436–2445.
18. Paci, L. and Finazzi, F. (2017). Dynamic model-based clustering for spatio-temporal data. *Statistics and Computing*, 1–16.
19. Patterson, T. A., Basson, M., Bravington, M. V., and Gunn, J. S. (2009). Classifying movement behaviour in relation to environmental conditions using hidden Markov models. *Journal of Animal Ecology*, 78(6): 1113–1123.
20. Pawlowsky, V. and Burger, H. (1992). Spatial structure analysis of regionalized compositions. *Mathematical Geology*, 24(6): 675–691.
21. Pirzamanbein, B., Lindström, J., Poska, A., and Gaillard, M.-J. (2018). Modelling Spatial Compositional Data: Reconstructions of past land cover and uncertainties. *Spatial Statistics*.
22. Quintana, J. M. and West, M. (1988). Time series analysis of compositional data. *Bayesian Statistics*, 3: 747–756.
23. Teh, Y. W., Jordan, M. I., Beal, M. J., and Blei, D. M. (2006). Hierarchical Dirichlet Processes. *Journal of the American Statistical Association*, 101(476): 1566–1581.
24. Tjelmeland, H. and Lund, K. V. (2003). Bayesian modelling of spatial compositional data. *Journal of Applied Statistics*, 30(1): 87–100.
25. Zucchini, W. and MacDonald, I. (2009). *Hidden Markov Models for Time Series: An Introduction Using R*. Chapman & Hall/CRC Monographs on Statistics & Applied Probability. Taylor & Francis.