

Bayesian principal curve clustering by species-sampling mixture models

Clustering mediante modelli mistura a campionamento di specie di curve principali bayesiane

Raffaele Argiento and Alessandra Guglielmi

Abstract In this work we are interested in clustering data whose support is “curved”. For this purpose, we will follow a Bayesian nonparametric approach by considering a species sampling mixture model. Our first goal is to define a general/flexible class of distributions, such that they can model data from clusters with non standard shape. To this end, we extend the definition of principal curve given in [8] (Tibshirani 1992) into a Bayesian framework. We propose a new hierarchical model, where the data in each cluster are parametrically distributed around the Bayesian principal curve, and the prior cluster assignment is given on the latent variables at the second level of hierarchy according to a species sampling model. As an application we will consider the detection of seismic faults using data coming from Italian earthquake catalogues.

Abstract *In questo lavoro siamo interessati al raggruppamento di dati il cui supporto è “curvo”. Per perseguire questo obiettivo, seguiamo un approccio bayesiano non parametrico utilizzando un modello mistura a campionamento di specie. Il nostro primo obiettivo è quello di definire una classe generale/flessibile di distribuzioni parametriche, in modo che queste possano modellare gruppi con forme non usuali. A tal fine, estendiamo la definizione di curva principale data in [8] (Tibshirani 1992) ad un contesto bayesiano. In conclusione, in questo lavoro proponiamo un nuovo modello gerarchico, nel quale i dati in ciascun gruppo hanno distribuzione parametrica centrata su una curva. L’assegnazione a priori dei dati ai gruppi è invece rappresentata mediante la legge di variabili latenti al secondo livello di gerarchia, le quali sono distribuite secondo un processo a campionamento di specie. Come applicazione consideriamo l’individuazione di faglie sismiche per dati provenienti da un catalogo di terremoti italiano.*

Raffaele Argiento
CNR-IMATI, Via Bassini 15, Milano (Italy) e-mail: raffaele@mi.imati.cnr.it

Alessandra Guglielmi
Politecnico di Milano, Dipartimento di Matematica, Piazza Leonardo da Vinci 32, 20133 Milano (Italy) e-mail: alessandra.guglielmi@polimi.it

Key words: Cluster Analysis, Mixture Models, Principal Curve, Specie Sampling Models

1 Introduction

In a nutshell, model-based cluster analysis means that data are assumed as coming from a several subpopulations, where each subpopulation is modeled separately, so that the overall population is a mixture. In this paper we are going to cluster data whose support is “curved” (in a Euclidean space), in a Bayesian nonparametric model-based framework. To this end we will define a *species sampling* mixture model with kernels belonging to a flexible and large class of densities.

Species sampling mixtures have been introduced in [5] as an extension of the well known Dirichlet process mixture model. They do not require specifying the number of mixture components and the clustering procedure can be viewed as a generalized Chinese restaurant process (see [5]): assume customers arriving sequentially at a Chinese restaurant and randomly assigned to an infinite number of tables which have unlimited seating capacities. When a new customer arrives, she will be seated to a (new or one of the occupied) table according to the current seating arrangement of all previous customers. Under this process, inference on the number of clusters and mixture model parameters estimation are unified and computed by a suitable MCMC algorithm.

Recently (see [1]) we have addressed the problem of clustering curved data, introducing a model called b-DBSCAN. However the prior for the random partition under the b-DBSCAN model, i.e. prior cluster assignment, is based on the geometry of the space of kernel densities rather than a direct random partition prior elicitation. Bayesian statisticians would prefer the latter alternative. With these considerations in mind, as an alternative, a new hierarchical model for clustering is proposed here, where the data in each cluster are parametrically distributed around a curve (*principal curve*), and the prior cluster assignment is given on the latent variables at the second level of hierarchy according to a species sampling model as sketched before. Clustering using principal curves is useful for detecting curvilinear features in spatial point patterns: as an application we will analyze the detection of seismic faults from earthquake catalogs.

2 Species sampling mixture models

A proper *species sampling model* (SSM) (see [6]), on $\Theta \subset \mathbb{R}^d$ is a random probability measure of the form

$$P(\cdot) := \sum_{j=1}^{\infty} w_j \delta_{\tau_j}, \quad (1)$$

where the atoms $\{\tau_j\}$ are chosen iid according to a centering distribution P_0 on Θ , while the distribution of the weights $\{w_j\}$ is characterized by the generalized Chinese process described in the Introduction (see [6] for mathematical details).

A species sampling mixture model, can be defined in a hierarchical way as follow:

$$X_i | \theta_i \stackrel{\text{ind.}}{\sim} f(\cdot; \theta_i) \quad i = 1, \dots, n \quad (2)$$

$$\theta_i | P \stackrel{\text{i.i.d.}}{\sim} P \quad i = 1, \dots, n, \quad P \sim \Pi(\cdot; \cdot, P_0) \quad (3)$$

where $\Pi(\cdot; \cdot, P_0)$ is the law of of a species sampling model as in (1). Here $\mathbb{E}(P(A)) = P_0(A)$ for each measurable $A \subset \Theta$. As a particular case, notation $P \sim DP(\alpha, P_0)$ means that P is given a Dirichlet process prior with parameters $\alpha > 0$ and P_0 . Species sampling mixture models are particularly suitable for cluster analysis. Indeed, under (2)-(3) the natural clustering rule is: X_i and X_j share the same cluster if and only if $\theta_i = \theta_j$. We denote with $(\phi_1, \dots, \phi_k)$ the unique values among the θ_i 's, and with $\rho = \{C_1, \dots, C_k\}$ the partition induced by the rule $i \in C_j$ iff $\theta_i = \psi_j$, $i = 1, \dots, n$ and $j = 1, \dots, k$. The parameter ρ (random partition) is the main object of our analysis.

3 Bayesian principal curve mixtures

The aim of this section is to define a general and flexible class of kernels densities, such that they can model data on clusters with non standard shape. The main idea is to represent data in the same *curved cluster* using one single distribution centered around a curve $\mathbf{m}(\cdot) := (m_1(t), m_2(t))$, $t \in (0, 1)$. To be more formal, suppose that $X \in \mathbb{R}^2$, and consider the model $X | \mathbf{m}(\cdot), \Sigma \sim f_{\mathbf{m}(\cdot), \Sigma}$ with

$$f_{\mathbf{m}(\cdot), \Sigma}(x) = \int_0^1 \mathcal{N}_2(x; \mathbf{m}(t), \Sigma) dt. \quad (4)$$

The triplet

$$\{\mathcal{U}(0, 1), \mathcal{N}_2(\mathbf{m}(\cdot), \Sigma), \mathbf{m}(\cdot)\} \quad (5)$$

is a *principal curve* according to [8]. We can write our model as

$$X = \mathbf{m}(T) + \varepsilon;$$

where $T \sim \mathcal{U}(0, 1)$, $\varepsilon \sim \mathcal{N}_2(0, \Sigma)$ and T and ε are independent. Then, we can describe our model as a nonparametric regression, where the mean of each data is a random point on a curve, and the error is bivariate Gaussian with covariance matrix Σ . More in details we represented the components of the curve $\mathbf{m}(\cdot)$ through two independent *low-rank thin-plate* splines [2], that is:

$$\mathbf{m}(t) = \begin{pmatrix} m_1(t) \\ m_2(t) \end{pmatrix} = \begin{pmatrix} \beta_0 + \beta_1 t + \sum_{k=1}^K u_k |t - \kappa_k|^3 \\ \gamma_0 + \gamma_1 t + \sum_{k=1}^K v_k |t - \kappa_k|^3 \end{pmatrix} \quad \Sigma = \begin{pmatrix} \sigma_1^2 & 0 \\ 0 & \sigma_2^2 \end{pmatrix},$$

where $\{\kappa_k, k = 1, \dots, K\}$ are fixed knots. Finally, we observe as the parameters of the kernel $f_{\mathbf{m}(\cdot), \Sigma}$ can be summarized as the real valued vector $\theta = (\boldsymbol{\beta}, \mathbf{u}, \boldsymbol{\gamma}, \mathbf{v}, \sigma_1^2, \sigma_2^2)$. Since our aim is to perform cluster analysis for data “belonging” to different curved clusters, we add a level of hierarchy to $f_{\mathbf{m}(\cdot), \Sigma}$ according to a Bayesian nonparametric approach. By Bayesian principal curve mixture model we mean the following:

$$\begin{aligned} X_i | \theta_i &\stackrel{\text{i.i.d.}}{\sim} f(x_i | \theta_i) \\ \theta_i &= (\mathbf{m}_i(\cdot), \Sigma_i) | P \stackrel{\text{i.i.d.}}{\sim} P, \quad i = 1, \dots, n, \\ P &\sim \Pi(\cdot; \cdot, P_0) \end{aligned}$$

where P is a species sampling model as in (1). Under this model, conditioning to the random partition ρ (induced by the SSM), data points in each cluster are centered around a curve, and are (conditionally) Gaussian distributed; their means are bound to lie on a curve $\mathbf{m}_i(t)$, $t \in (0, 1)$. Observe that we will be able to provide not only cluster estimates, but also principal curve estimates (see Section bho). As far as the choice of P_0 is concerned, we assume that the component of $\theta = (\boldsymbol{\beta}, \mathbf{u}, \boldsymbol{\gamma}, \mathbf{v}, \sigma_1^2, \sigma_2^2)$ are independent and

$$\begin{aligned} \beta_0, \beta_1, \gamma_0, \gamma_1 &\stackrel{\text{i.i.d.}}{\sim} \mathcal{N}(0, 1000) & \sigma_1^2, \sigma_2^2 &\stackrel{\text{i.i.d.}}{\sim} \text{inv-gamma}(a, b) \\ u_i | \sigma_u &\stackrel{\text{i.i.d.}}{\sim} \mathcal{N}(0, \sigma_u^2), & \sigma_u &\sim \text{inv-gamma}(a_u, b_u) \\ v_i | \sigma_v &\stackrel{\text{i.i.d.}}{\sim} \mathcal{N}(0, \sigma_v^2), & \sigma_v &\sim \text{inv-gamma}(a_v, b_v). \end{aligned}$$

This prior choice for the parameters $\mathbf{u}$ and $\mathbf{v}$ is motivated by the equivalence between penalized splines and mixed models in the regression context, as shown for instance in [2]. As a final remark, we observe that, any smoother depends heavily on the choice of the smoothing parameter: in our model the smoothing parameters are $\lambda_u = \sigma_u^2 / \sigma_1^2$ and $\lambda_v = \sigma_v^2 / \sigma_2^2$: small (large) λ 's corresponds to oversmoothing (undersmoothing). As we will show in the next section, the hyperparameters of the inverse-gamma distributions should be accurately fixed heavy oversmoothing or undersmoothing.

4 Application

We have implemented our model for the simple case of a Dirichlet process mixture model (DPM); posterior estimates were obtained through a modification of the Blocked Gibbs sampler based on a finite dimensional approximation of the Dirichlet process proposed in [4]. Moreover, all Bayesian cluster estimates here are based on the posterior distribution $\pi(\rho | \text{data})$ of the random partition ρ , as the partition minimizing the posterior expectation of Binder's loss function, i.e. as the loss function assigning cost b when two elements are wrongly clustered together and cost a when two elements are erroneously assigned to different clusters. In particular we have assumed $a = b$.

Simulated data

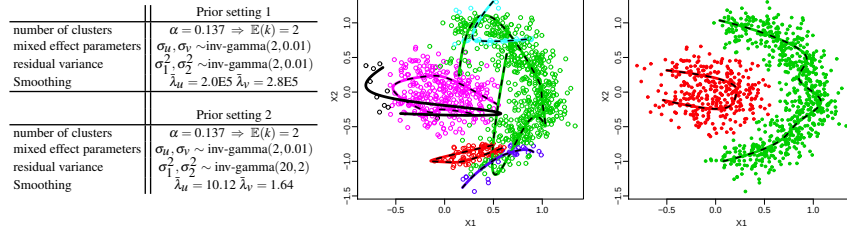


Fig. 1 Cluster and principal curve estimates when the smoothing parameters are large (figure on the left) or small (figure on the right)

In this section we illustrate our model with application to a simulated dataset of size $n = 1000$. Data are shown in Figure 1; there are two main groups of observations: the first one has a sharp round shape and it is located around the point $(0,0)$, while the second group lays on a semicircular region on the right of the first group.

We report here the result of two significative experiments with hyperparameters reported in the table on the left panel of Figure 1. From the figure we see that, if an “informative” (Prior setting 1) distribution for the covariance matrix Σ_i in each cluster is elicited, the cluster and curve estimates in the right panel are better than those in the left panel, corresponding to a “non-informative” (Prior setting 2) prior for the covariance matrix. This can even be understood in terms of the two smoothing parameters. If we consider the posterior means of the smoothing parameters for each observation $(\hat{\lambda}_{ui}, \hat{\lambda}_{vi}) := \mathbb{E}((\lambda_{ui}, \lambda_{vi})|data)$, with $i = 1, \dots, n$, and consider $\bar{\lambda}_u = \frac{1}{n} \sum_{i=1}^n \hat{\lambda}_{ui}$, $\bar{\lambda}_v = \frac{1}{n} \sum_{i=1}^n \hat{\lambda}_{vi}$ as two overall poster smoothing indexes, we see that the first setting of the prior leads to an oversmoothing estimation of the principal curves within the clusters (middle panel of the Figure 1).

Seismic sources data

In the construction of seismic risk maps is important to identify faults segments inside the seismogenic sources. From the literature (see [7]) is well known that in Italy there are inherent difficulties to this achievement. As an application we deal with this problem by fitting our model to data coming from the CPTI04 Italian earthquake catalogue (CPTI04 2004), from which we have considered all the events classified by geologists as belonging to the three composite sources CSS-25, CSS-27, CSS-37 of the Database of Italian Seismogenic Sources [3]. Data are arranged on a long stretch of the Apennines mountains ranging from Modena district (norther Italy) up to Frosinone district (middle Italy). We have run the analysis under various hyperparameter settings, here we report one of the most significant (see the table in Figure 2). Here the mass parameter of the Dirichlet process α has an interpretation as smoothing parameter controlling the length of the principal curve in each cluster. Large values of α mean more groups and shorter principal curves and vice versa.

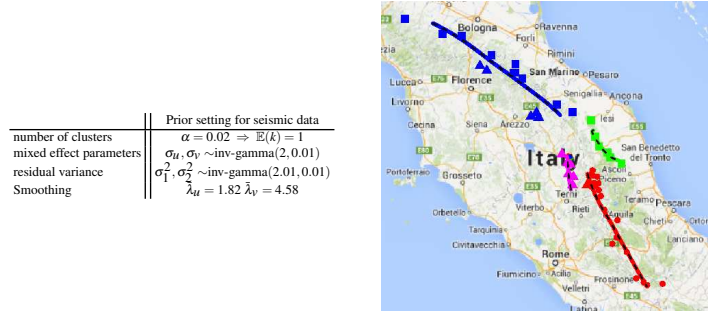


Fig. 2 Earthquakes from CPTI04 catalog classified by source 27 (squares), source 37 (triangles) and source 25 (circles). Cluster and principal curve estimates are represented by colors

The posterior estimates are reported in Figure 2, and can be summarized as follows: the model is able to recover source CSS-25; source CSS-27 is split into two parts (north and south); the southern part of source CSS-37 is grouped, while the northern events are wrongly classified as belonging to source CSS-27. Indeed, since the sample size is very low northern events of source CSS-37 are far from the southern points on the same source but geographically close to source CSS-27. In addition, when geologists associate earthquakes with seismogenic sources, they take into account many elements, as the focal mechanism, and not only the geographical location of the epicentre as done by our model. On the other hand, the association of the events related to the source CSS-27 with two fault segments may be supported by geological considerations as reported in [7].

Acknowledgments: we thank Renata Rotondi for the seismic dataset and very interesting discussions.

References

1. Argiento, R., Cremaschi, A., Guglielmi, A.: A “Density-Based” Algorithm for Cluster Analysis Using Species Sampling Gaussian Mixture Models. *J. Comp. Graph. Stat.* (2013) doi:10.1080/10618600.2013.856796
2. Crainiceanu, C.M., Ruppert, D., Wand, M.P.: Bayesian Analysis for Penalized Spline Regression Using WinBUGS. *J. Stat. Softw.*, **14**, (2005)
3. DISS Working Group: Database of Individual Seismogenic Sources INGV, Italy (2007)
4. Ishwaran, H. and James, L. F.: Gibbs sampling methods for stick-breaking priors. *J.A.S.A.* **96**, 161–173 (2001).
5. Ishwaran, H. and James, L. F.: Generalized weighted Chinese restaurant processes for species sampling mixture models. *Stat. Sinica* **13**, 1211–1236 (2003)
6. Pitman, J.: Some developments of the Blackwell-MacQueen urn scheme. Volume 30 of *IMS Lecture Notes-Monograph Series*, pp. 245-267. Hayward (1996)
7. Rotondi, R.: Retrospective validation of renewal-based, medium-term earthquake forecasts. *Geophys. J. Int.* **195**, 381-407 (2013)
8. Tibshirani, R.: Principal Curves Revisited. *Stat. Comput.*, **2**, 183–190 (1992)