Bayesian nonparametric covariate driven clustering

Un modello bayesiano nonparametrico per clustering in presenza di covariate

Raffaele Argiento, Ilaria Bianchini, Alessandra Guglielmi and Ettore Lanzarone

Abstract In this paper we introduce a Bayesian model for clustering individuals with covariates. This model combines the joint distribution of data in the sample, given the parameter and covariates, with a prior for this parameter. Here, the partition of the sample subjects is the parameter, and the prior we assume encourages two subjects to co-cluster when they have similar covariates. Cluster estimates are based on the posterior distribution of the random partition, given data. As an application, we fit our model to a dataset on gap times between recurrent blood donations from AVIS (Italian Volunteer Blood-donors Association), the largest provider of blood donations in Italy.

Abstract *Introduciamo un modello per il clustering di individui in presenza di covariate. Il modello combina la distribuzione congiunta dei dati, condizionatamente al parametero e alle covariate, con una prior per il parametro stesso, secondo l'approccio bayesiano. Qui il parametro è la partizione dei soggetti nel campione. La prior che introduciamo incoraggia due soggetti a stare nello stesso gruppo se hanno covariate simili. Applicheremo il nostro modello ad un dataset che riguarda le donazioni di sangue ripetute nel tempo.*

Key words: Bayesian nonparametrics, clustering, normalized completely random measures, regression models

Raffaele Argiento
Dipartimento ESOMAS, Università di Torino, e-mail: raffaele.argiento@unito.it

Ilaria Bianchini, Alessandra Guglielmi
Dipartimento di Matematica, Politecnico di Milano, e-mail: alessandra.guglielmi@polimi.it

Ettore Lanzarone
CNR-IMATI, Milano, e-mail: ettore.lanzarone@cnr.it

1 Introduction

In this paper, we introduce a Bayesian model for clustering individuals with covariates. Typically, in the Bayesian framework, there are two approaches for clustering. The first one assumes data to be independently distributed according to a mixture of parametric densities (possibly depending on covariates), with k components; k may be finite, with a prior on k, or infinite ($k = +\infty$), corresponding to a Bayesian nonparametric mixture. The second approach assumes a prior directly on the partition of sample subjects into clusters. Dirichlet process mixtures (DPMs), popularized by [4], are an example of Bayesian nonparametric models, where the weights of the infinite mixture are constructed using the stick-breaking representation (see [14]). Advantages of DPMs over finite mixtures with a prior on the number of components are the existence of generic Markov chain Monte Carlo (MCMC) algorithms for posterior inference to adapt to various applications and elegant mathematical properties of the nonparametric model; see [11] for a discussion on both models. Any DPM induces a random partition of the subject labels $\{1,2,\ldots,n\}$ through the values of the parameters $(\theta_1,\ldots,\theta_n)$ identifying the mixture component the observations are sampled from; in fact, since the mixing measure is almost surely discrete, there are ties in $(\theta_1,\ldots,\theta_n)$ with positive probability. Two subjects i and l share the same cluster if and only if $\theta_i = \theta_l$. In general, this relationship holds for any Bayesian nonparametric mixture model where the mixing measure is an almost surely discrete random probability measure.

In this paper, we follow the second Bayesian approach to clustering, which is more direct, since the random parameter of the model is the subject of the inference itself, i.e. the partition of the sample subjects. To define the model, we assign the joint conditional distribution of data in the sample, given the random partition, and the prior for this parameter. Corresponding cluster estimates are summaries of the posterior distribution of the random partition, given data. In particular, our prior depends on covariates, and we encourage, a priori, two subjects to co-cluster when they have similar covariates.

Our model is a generalization of the PPMx model proposed in [12], a product partition model with covariates information, extending the product partition model by [7]. This latter assumes a prior probability mass function for the random partition $\rho_n = \{A_1,\ldots,A_{k_n}\}$ proportional to the product of functions defined over the clusters, which are called cohesion. In [12], as well as in its generalizations, the cohesion function is restricted to be the one induced by the Dirichlet process, namely $c(S_j) = \kappa(n_j-1)!$, where κ is a positive constant and n_j is the size of cluster A_j. However, this cohesion inherits “the rich-gets-richer” property of the Dirichlet process, i.e. sampled posterior partitions consist of a small number of large clusters, where new observations are more likely to join already-large clusters with probability proportional to the cardinality of the cluster.

To overcome this limitation, we introduce a more general class of PPMx models, with cohesion induced by normalized completely random measures (NormCRMs); see [13]. We perturb the product partition expression of the prior of the random partition via a similarity function g which depends on all the covariates associated to subjects in each cluster. Such g can be any non-negative function of some similarity measure guaranteeing that the prior probability that two items belong to the same cluster increases if their similarity increases. Note that we call our model "nonparametric", even though the random partition parameter has finite dimension; however its dimension is huge and increases with sample size. We are also able to build a general MCMC sampler to perform posterior analysis that does not depend on the specific choice of similarity. We test our model on a simulated dataset, and on a dataset on gap times between recurrent blood donations from AVIS (Italian Volunteer Blood-donors Association), the largest provider of blood donations in Italy. For the latter application, the problem is to find suitable methods to cluster recurrent event data, and predict a new recurrent event, using covariates describing personal characteristics of the sample individuals. In this paper, we model the sequence of gap times between recurrent events (blood donations) since donors are expected to donate blood not before a fixed amount of time imposed by the law. According to AVIS standard practice, the gap time between donations is the quantity that can be influenced for a better planning of the overall daily blood supply.

2 Bayesian covariate driven clustering

In a regression context, let $\boldsymbol{y}_i, \boldsymbol{x}_i$, $i = 1, \ldots, n$ be the vector of responses and covariates for subject i, with $dim(\boldsymbol{y}_i) = n_i$; we assume $n_i = 1$ for all i in this section for greater clarity. We denote by $\boldsymbol{y}_j^*$ (and $\boldsymbol{x}_j^*$) the set of all responses y_i (and covariate vectors $\boldsymbol{x}_i$) in cluster A_j; the notation will be used later in the paper. We start from a family of regression models $f(\cdot; \boldsymbol{x}, \boldsymbol{\theta})$, $\boldsymbol{\theta} \in \Theta \subset \mathbb{R}^l$, and specify a hierarchical model that encourages subjects with similar covariates to be in the same cluster, using a data dependent prior for the random partition of data. We assume that data are independent across groups, conditionally on covariates and the cluster specific parameters; these are i.i.d from a base distribution P_0. Covariates enter both in the likelihood and the prior in our model. Concretely, we assume:

$$Y_1, \ldots Y_n | \mathbf{x}_1, \ldots, \mathbf{x}_n, \boldsymbol{\theta}_1^*, \ldots, \boldsymbol{\theta}_{k_n}^*, \rho_n \sim \prod_{j=1}^{k_n} f(\mathbf{y}_j^* | \mathbf{x}_j^*, \boldsymbol{\theta}_j^*) \tag{1}$$

$$\boldsymbol{\theta}_1^*, \ldots, \boldsymbol{\theta}_{k_n}^* | \rho_n \overset{\text{iid}}{\sim} P_0 \tag{2}$$

$$p(\rho_n = \{A_1, \ldots, A_{k_n}\} | \mathbf{x}_1, \ldots, \mathbf{x}_n) \propto \int_0^{+\infty} D(u,n) \prod_{j=1}^{k_n} c(u, n_j) g(\mathbf{x}_j^*) du \tag{3}$$

where n_j denotes the size of cluster A_j, $g(\mathbf{x}_j^*)$ is the similarity function on cluster A_j such that $g(\emptyset) = 1$, and P_0 is a diffuse probability on the parameter space. Here $D(u,n)$ and $c(u,n_j)$ are defined as:

$$D(u,n) = \frac{u^{n-1}}{\Gamma(n)} \exp\{-\kappa \int_0^{+\infty} \left(1 - \mathrm{e}^{-us}\right) \rho(s) ds\}$$

where $\rho(ds) = \dfrac{1}{\Gamma(1-\sigma)} s^{-1-\sigma} \mathrm{e}^{-s} \mathbb{1}_{(0,+\infty)}(s)ds$ and

$$c(u,n_j) = \int_0^{+\infty} \kappa s^{n_j} \mathrm{e}^{-us} \rho(s) ds = \frac{\kappa\,\Gamma(n_j-\sigma)}{\Gamma(1-\sigma)} \frac{1}{(1+u)^{n_j-\sigma}}. \tag{4}$$

The intensity $\rho(ds)$, the positive parameter κ and the probability P_0 define a specific class of normalized completely random measures, called normalized generalized gamma process (NGG). Parameter σ has a deep influence on the clustering behavior. In particular, the discount parameter σ affects the variance: the larger it is, the more disperse is the distribution on the number of clusters. This feature mitigates "the rich-gets-richer" effect, typical of the Dirichlet process, leading to more homogeneous clusters. For more details on the behavior of σ in NGG's, see for instance [3], [10] and [2].

The likelihood specification in (1) may be any model, from simple regression models to the more complex models for gap times of recurrent events as in the AVIS application. The prior (3) is a perturbation of a prior for ρ_n, called product partition model (PPM) and introduced in [7]. When $g \equiv 1$, i.e. there are no extra information from covariates, the prior mass of each cluster depends only on its size through $c(u,n_j)$; when g is a proper function, the higher is the value of $g(\mathbf{x}_j^*)$, i.e. the more similar are covariates in cluster j, the higher is the prior probability mass of that cluster. This interpretation is justified since the prior $p(\rho_n|\mathbf{x}_1,\ldots,\mathbf{x}_n)$ in (3) can be equivalently written as

$$p(\rho_n|\mathbf{x}_1,\ldots,\mathbf{x}_n,u) \propto M(u) \prod_{j=1}^{k_n} c(u,n_j) g(\mathbf{x}_j^*) \tag{5}$$

for some prior density on the auxiliary variable $u > 0$. In other words, our model is an extension on the PPMx model, namely, it is a mixture of PPMx models (5).

It is quite natural to let the similarity to be a non-increasing function of the distance among covariates in the cluster, namely

$$\mathscr{D}_{A_j} = \sum_{i \in A_j} d(\mathbf{x}_i, \mathbf{c}_{A_j}) \tag{6}$$

where $\mathbf{c}_{A_j}$ is the centroid of the set of covariates in cluster j and d is a suitable distance function that we discuss later. Moreover, we assume $g(\mathscr{D}_{A_j}) := 1$ if the size of the set A_j is 1, i.e. $|A_j| = 1$.

The choice of the similarity is crucial, since this function controls how covariates affect the clustering. For this reason, we propose a list of similarity functions that proved to work reasonably well in practice; among those, here we list:

$g_A(\mathbf{x}_j^*;\lambda) = \mathrm{e}^{-t^{\alpha}}$, for $\alpha > 0$ ($\alpha = 0.5, 1, 2$), with $t = \lambda \mathscr{D}_{A_j}$;

$g_C(\mathbf{x}_j^*;\lambda)$ equals to $\mathrm{e}^{-t\log t}$ if $t \geq \frac{1}{e}$, or to $\frac{\mathrm{e}^{1/e-1}}{t}$ if $t < \frac{1}{e}$, where $t = \lambda \mathscr{D}_{A_j}$.

Here $\lambda > 0$ is a tuning parameter. The similarity function g_A is intuitive, i.e. its behaviour for $t \to +\infty$ is exponential. As far as the expression of g_C is concerned, we have proposed the expression $\mathrm{e}^{-t\log t}$ in such a way that, for large t, we contrast the asymptotic behavior of the cohesion function (4) induced by the NGG process. In fact, our model works well if the prior is not completely driven by covariates, because otherwise we could lose all the advantages of a Bayesian model-based clustering approach (e.g., uncertainty quantification, prediction, sharing information across clusters).

When the similarity is g_A, if we choose a very small λ, we concentrate the values of $\lambda\mathscr{D}_A$ around the origin, and hence we obtain similar values for $g_A(\cdot)$: in this case, the effect of covariate information on the prior of ρ_n will be very mild, since the range of values that the similarity can assume is very limited. A similar argument is valid for large values of λ. In conclusion, we calibrate λ such that g_A is evaluated in the range, say, $(0,3)$, for this particular choice of similarity.

In the applications we consider later, covariates will always be continuous or binary; categorical or ordinal covariates are translated into dummies. Hence, if $\mathbf{x}_1$ and $\mathbf{x}_2$ are vectors of covariates, $\mathbf{x}_j=(\mathbf{x}_j^c,\mathbf{x}_j^b)$, where $\mathbf{x}_j^c$ is the sub-vector of all the continuous covariates and $\mathbf{x}_j^b$ is the sub-vector of all binary covariates, we define the function d in (6) as

$$d(\mathbf{x}_1,\mathbf{x}_2)=d^c(\mathbf{x}_1^c,\mathbf{x}_2^c)+d^b(\mathbf{x}_1^b,\mathbf{x}_2^b), \tag{7}$$

where d^c is the Malahanobis distance between vectors, i.e. the Euclidean distance between *standardized* vectors of covariates, and d^b is the Hamming distance between vectors of binary covariates. The choice of the distance in (7) is not unique, but alternatives are among the subject of current research.

The way we define g does not increase the complexity of the algorithm for posterior inference. Indeed, we are able to devise a general MCMC sampler to perform posterior analysis that does not depend on the specific choice of similarity. The full-conditionals of the Gibbs sampler are relatively easy to implement in this case, since our algorithm generalizes the augmented marginal algorithm for mixture models in [9] and [5].

3 Simulated data

We apply model (1)-(3) in the regression context. Here $f(\mathbf{y}_j^*|\mathbf{x}_j^*,\theta_j^*)$ is the Gaussian regression model. We simulated a dataset of points $(y_i,x_{i1},\ldots,x_{ip})$ for $i=1,\ldots,n$, with $n=200$ and $p=4$. The last 2 covariates are binary and were generated from the Bernoulli distribution, while the first 2 were generated from Gaussian densities. The responses y_i's were generated from a linear regression model with linear predictor $\boldsymbol{x}_i^T\boldsymbol{\beta}$, where $\boldsymbol{\beta}^0:=(\beta_0^0,\beta_1^0,\beta_2^0,\beta_3^0,\beta_4^0)$ and variance $\sigma_e^2=0.5$. We have generated 3 different groups by generating both covariates and responses from distributions with different parameters.

We run the Gibbs sampler algorithm to obtain 5,000 final iterations from the full posterior distribution, with initial burn-in of 2,000 and thinning of 10 iterations.

A-posteriori we classified all datapoints according to the optimal partition, under the different similarity functions; results are summarized in Table 1. By optimal partition we mean the realization, in the MCMC chain, of the random partition ρ_n which minimizes posterior expected value of the Binder's loss function with equal missclassification weights [8]. Observe that there are no missclassified data using similarity g_A, while 4% of data are missclassified using g_C, while the missclassification error increases to 16% if we do not assume covariate information ($g \equiv 1$).

Table 1 Missclassification rates for the simulated dataset

missclassif rate	g_A	g_C	$g \equiv 1$
	0%	4%	16%

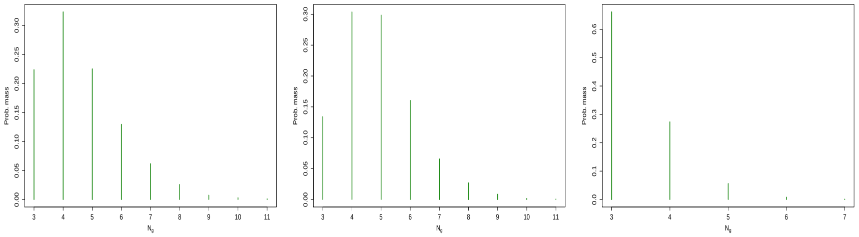

Fig. 1 Posterior distribution of K_n under g_A (left), g_C (center) and $g \equiv 1$ (right).

We computed the posterior distribution of K_n, the number of clusters, in the three cases; see Figure 1. Figure 2 displays the predictive distribution corresponding to covariates $\boldsymbol{x}_1$ of the first subject. The green vertical line corresponds to the actual observation y_1. It is clear that in the last case, i.e. when we do not include covariate information in the prior for the random partition, the predictive law is not able to distinguish to which of the three groups the subject belongs (thus, we have three peaks in the law). In cases A and C the predictive law exhibits only one main peak: the covariate information helps, in this case, in selecting the right group for the observation. This is also proved by the missclassification table above.

We underline that our prior encourages subjects with the same covariates to be in the same cluster, so that the posterior will generally allocate these subjects in the same cluster as well. On the other hand, if two subjects have very different covariates, our prior would classify them to different clusters, even if their responses are similar. However, the likelihood, i.e. the conditional distribution of data given the parameter, could correct the prior probability, if this is the case, and could allocate two subjects with different covariates to the same cluster.

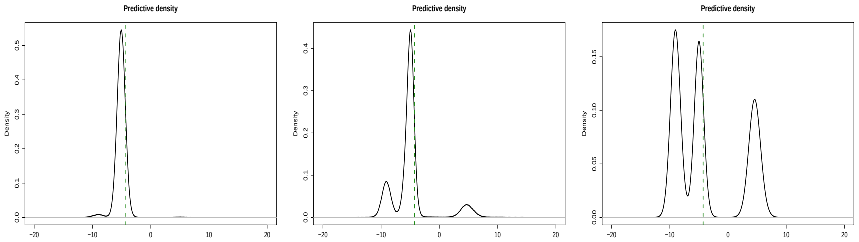

Fig. 2 Predictive distribution of Y_1 under g_A (left), g_C (center) and $g \equiv 1$ (right); vertical lines denote the true value

4 Blood donation data

Our data concern new donors of whole blood donating in a fixed time window in the main building of AVIS in Milano. Data are recurrent donation times, with extra information summarized in a set of covariates, collected by AVIS physicians. The last gap times are administratively censored for almost all the donors, except those having their last donation exactly on that date. The dataset contains 17198 donations, made by 3333 donors.

The statistical focus here is the clustering of donors according to the trajectories of gap times. Figure 3 reports the histogram of this variable (in the log-scale) for men and women. The skewness of these histograms can be explained since, according to the Italian law, the maximum number of whole blood donations is 4 per year for men and 2 for women, with a minimum of 90 days between a donation and the next one. Note that the minimum for men is around 4.5 ($e^{4.5} \simeq 90$ days). For women,

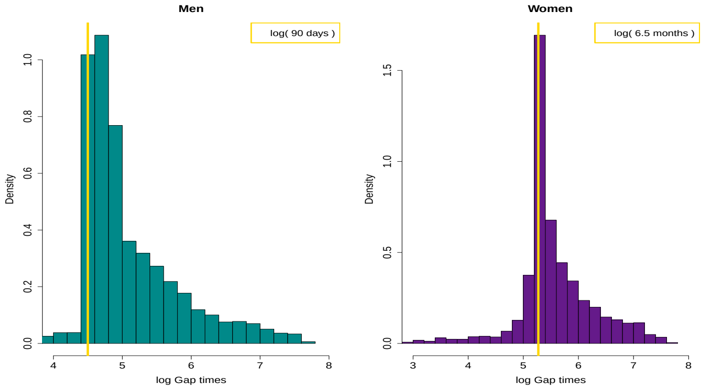

Fig. 3 Histogram of the logarithm of the observed gap-times divided according to gender, male (left) and women (right).

the distribution has a mode approximately in 5.3 in the log scale: this means 200 days, that corresponds to about 6 month and a half. Observe that donors may donate before the minimum imposed by law, under good donor's health conditions and the physician's consent.

We model gap times of successive donations as a regression model for recurrent gap times with two linear predictor terms, involving fixed-time and time varying covariates. The distribution of each gap time, in the log scale, is assumed to be skew-normal (see Figure 3); using parameterization in [6], we model the logarithm of the t-th gap time of donor i as Gaussian distributed. Cluster specific parameters are the intercept, the skewness parameter, and the variance of the response. We assume the prior for the random partition as in (5). Among donor's covariates, we include gender, blood type and RH factor, age, body mass index (BMI) and other information.

Preliminary analysis shows that, a posteriori, age and BMI (time-varying) have an effect on the gap time, as well as gender and RH factor. Details on the cluster estimate will be given during the talk.

References

1. Argiento, R., Guglielmi, A., Hsiao, C. K., Ruggeri, F., Wang, C.: Modeling the association between clusters of SNPs and disease responses. In: Müller, Mitra (eds.) Nonparametric Bayesian Inference in Biostatistics, pp. 115-134. Springer, New York (2015)
2. Argiento, R., Guglielmi, A., Pievatolo, A.; Bayesian density estimation and model selection using nonparametric hierarchical mixtures. Computational Statistics & Data Analysis, **54**, 816-832 (2010)
3. Argiento, R., Bianchini, I., Guglielmi, A.: Posterior sampling from ε-approximation of normalized completely random measure mixtures. Electronic Journal of Statistics, **10**, 3516-3547 (2016)
4. Escobar, M. D., West, M.: Bayesian density estimation and inference using mixtures. Journal of the American statistical association, **90**, 577–588 (1995)
5. Favaro, S., Teh, Y. W.: MCMC for normalized random measure mixture models. Statistical Science, **28**, 335–359 (2013)
6. Frühwirth-Schnatter, S., Pyne, S. (2010). Bayesian inference for finite mixtures of univariate and multivariate skew-normal and skew-t distributions. Biostatistics, 11(2), 317-336.
7. Hartigan, J. A.: Partition models. Communications in statistics-Theory and methods, **19**, 2745–2756 (1990)
8. Lau, J. W., Green, P. J.: Bayesian model-based clustering procedures. Journal of Computational and Graphical Statistics, **16**, 526–558 (2007)
9. Lijoi, A., Prünster, I.: Models beyond the Dirichlet process. In: Hiort, Holmes, Müller, Walker (eds.). Bayesian nonparametrics, pp. 80–136. Cambridge University Press, Cambridge (2010)
10. Lijoi, A., Mena, R. H., Prünster, I.: Controlling the reinforcement in Bayesian nonparametric mixture models. Journal of the Royal Statistical Society: Series B (Statistical Methodology), **69**, 715-740 (2007)
11. Miller, J. W., Harrison, M. T.: Mixture models with a prior on the number of components. Journal of the American Statistical Association, 1-17, 10.1080/01621459.2016.1255636 (2017)
12. Müller, P., Quintana, F., Rosner, G. L.: A product partition model with regression on covariates. Journal of Computational and Graphical Statistics, **20**, 260–278 (2011)

13. Regazzini, E., Lijoi, A., Prünster, I.: Distributional results for means of normalized random measures with independent increments. Annals of Statistics, **31**, 560–585 (2003)
14. Sethuraman, J.: A constructive definition of Dirichlet priors. Statistica sinica, **4**, 639–650 (1994).