# A "Density-Based" Algorithm for Cluster Analysis Using Species Sampling Gaussian Mixture Models

Raffaele ARGIENTO, Andrea CREMASCHI, and Alessandra GUGLIELMI

## 1. INTRODUCTION

In this article we introduce a Bayesian nonparametric model for cluster analysis. Typically, clustering means to partition a set of $n$ objects (i.e., data) into $k$ groups, even if the common features of the objects in each group are unknown or unobservable (i.e., latent). In general, data do not belong to a unique correct clustering, but, depending on the application, we would like to estimate a "true" one.

There are plenty of cluster analysis algorithms or models that, in the last decades, have been proposed. Here, we find it useful to distinguish between model-based and heuristic clustering techniques. The former class refers to those methods that require a statistical model to describe the problem, that is, mixture modeling; see, for instance, McLachlan and Peel (2000). The latter class includes those algorithms defined from a given starting partition, and carried on following some heuristic scheme. Very popular examples are

Raffaele Argiento is Researcher, CNR-IMATI, Milano 20133, Italy (E-mail: *raffaele@mi.imati.cnr.it*). Andrea Cremaschi is Ph.D. Student, School of Mathematics, Statistics and Actuarial Science, University of Kent, Canterbury, UK (E-mail: *ac429@kent.ac.uk*). Alessandra Guglielmi is Associate Professor, Dipartimento di Matematica, Politecnico di Milano, Milano 20133, Italy (E-mail: *alessandra.guglielmi@polimi.it*).

the hierarchical clustering (Johnson 1967), and $k$-means (MacQueen 1967). While these methods have been widely used in practice, they may suffer from serious limitations. For example, a distance between the objects must always be available, but in general it depends on problem features and data characterization. Moreover, for some of these methods, the number of clusters must be fixed in advance.

Here we propose a Bayesian nonparametric model, that combines two ingredients: species sampling mixture models of Gaussian distributions (in a nutshell, mixtures of infinite Gaussian densities with random weights), and a heuristic clustering procedure, called DBSCAN. The DBSCAN algorithm (Ester, Kriegel, and Xu 1996) is a density-based clustering technique, where the word *density* refers to the spatial disposition of data points, that are *dense* when forming a group: two data points are in the same cluster if their distance is smaller than the threshold. As far as the species sampling mixture model is concerned, it is well known that this model is convenient to assign a prior directly on the partition of the data, representing the natural parameter in the cluster analysis context. Moreover, the number of clusters is not fixed a priori, but it is estimated as a feature of the partition of the observations.

To summarize, our model is based on the slackness of the natural clustering rule of species sampling mixture models of parametric densities, when we mean that two observations $X_i$ and $X_j$ are in the same cluster if, and only if, the latent parameters $\theta_i$ and $\theta_j$ are equal. We say instead that two observations share the same cluster if the distance between the densities corresponding to their latent parameters is smaller than a threshold $\epsilon$. We complete the definition to provide an equivalence relation among data labels. The resulting new random partition parameter $\rho_\epsilon$ is coarser than the original $\rho$ induced by the species sampling mixture model, that is, the cardinality of $\rho_\epsilon$ is smaller than that of $\rho$. Of course, this procedure depends on the value of the threshold $\epsilon$ and the distance between densities. As far as the latter choice is concerned, we consider the symmetrized Kullback–Leibler I-divergence and the $L^2$ distance, since they are easy to interpret, and have a closed analytical form under Gaussian kernels. On the other hand, the elicitation of a value for $\epsilon$ has a key role in our model, since this threshold greatly affects the cluster estimate. Here we suggest to fix a grid of reasonable values for $\epsilon$, and choose the value maximizing the posterior expectation of a function of the random partition. In the applications, we consider some posterior predictive distribution summary statistics, as well as more standard tools like the silhouette coefficient and the adjusted Rand index. The Bayesian cluster estimate is given in terms of this new random partition, and results from the minimization of the posterior expectation of a loss function, as usually done in the literature (see, e.g., Lau and Green 2007).

We discuss implementation and applications of the model to a simulated bivariate dataset from a mixture of two densities with a curved cluster, and to a dataset consisting of gene expression profiles measured at nine different times, known in literature as yeast cell cycle data. We have paid special attention to comparing our model with "competitors" in the literature. In particular, for the simulated bivariate example, we have compared our estimates not only to those from heuristic DBSCAN, $k$-means and hierarchical algorithms, but also with estimation provided by the nested Dirichlet process mixture model (Rodriguez, Dunson, and Gelfand 2008), particularly helpful in clustering in nested settings. In both applications, thanks to this empirical thorough comparison, we have concluded that cluster

estimates from our model turn out to be more effective. Our estimates fit data particularly well when they come from heavy tailed or curved clusters.

The rest of this article is organized as follows. Section 2 describes the underlying species sampling mixture models. In Section 3 we describe the model under the new parameterization in details, discussing some of its main features. Section 4 illustrates the choice of the threshold parameter $\epsilon$. In Section 5 the simulated bivariate "curved" dataset is analyzed, with a comparison analysis with competing methods, while Section 6 addresses the yeast cell cycle data in Cho et al. (1998). We conclude with a discussion in Section 7.

## 2. THE MODEL

We set up a Bayesian model in which the partition of data is a random variable, distributed according to some prior distribution. If $(X_1, \ldots, X_n)$ represents the data, its conditional distribution is

$$(X_1, \ldots, X_n)|C_1, \ldots, C_k, \phi_1, \ldots, \phi_k \sim \prod_{j=1}^{k} \left\{ \prod_{i \in C_j} f(x_i; \phi_j) \right\}, \tag{1}$$

where $\boldsymbol{\rho} := \{C_1, \ldots, C_k\}$ is a partition of the data label set $\{1, \ldots, n\}$ and $\{f(\cdot; \phi), \phi \in \Theta\}$ is a family of densities on $\mathbb{R}^p$. We assume that the family of densities is identifiable, that is, $P_{\phi_1} = P_{\phi_2}$ implies $\phi_1 = \phi_2$, where $P_\phi$ is the probability measure corresponding to the density $f(\cdot; \phi)$. Observe that here $k$ is the number of clusters in the partition $\boldsymbol{\rho}$. From (1), it is clear that, conditionally on $\boldsymbol{\rho}$, data are independent between different clusters and are independent and identically distributed (iid) within each cluster. To complete the Bayesian model we need to assign a prior for $(\boldsymbol{\rho}, \boldsymbol{\phi})$. We assume that

$$\pi(\boldsymbol{\rho}) = \mathbb{P}(\boldsymbol{\rho} = \{C_1, \ldots, C_k\}) = p(\#C_1, \ldots, \#C_k), \tag{2}$$

where $p(\cdot)$ is an *infinite* exchangeable partition probability function (eppf). Moreover, conditionally on $\boldsymbol{\rho}$, we assume that the parameters in $\boldsymbol{\phi} := (\phi_1, \ldots, \phi_k)$ in (1) are iid from some fixed distribution $P_0$ on $\Theta \subset \mathbb{R}^s$. By Pitman (1996), for each distribution $P_0$ and eppf $p(\cdot)$, there exists a unique species sampling prior $\Pi(\cdot; p, P_0)$ on the space of all probabilities on $\Theta$, such that model (1) under the specified prior is equivalent to

$$\begin{aligned} &X_i|\theta_i \overset{\text{ind}}{\sim} f(\cdot; \theta_i) \quad i = 1, \ldots, n \\ &\theta_i|P \overset{\text{iid}}{\sim} P \quad i = 1, \ldots, n \qquad P \sim \Pi(\cdot; p, P_0), \end{aligned} \tag{3}$$

where $P_0$ represents the expectation of $P$. In this model every $X_i$ has density $f(\cdot, \theta_i)$, which is univocally determined by the value of $\theta_i$. In this case, we say that $\theta_i$ is the latent variable corresponding to $X_i$ in the mixture model (3).

In this work we consider only proper species sampling models, that is, $P(\cdot) = \sum_{i=1}^{\infty} \xi_i \delta_{\boldsymbol{\tau}_i}(\cdot)$, where $(\xi_i)$ and $(\boldsymbol{\tau}_i)$ are independent, $(\boldsymbol{\tau}_i) \overset{\text{iid}}{\sim} P_0(\cdot)$ and the law of $(\xi_i)$ is induced from $\Pi(\cdot; p, P_0)$ (see Pitman 1996, for its explicit description); here $\delta_\tau(\cdot)$ is the degenerate probability measure on $\tau$. An interesting example is the *normalized generalized gamma* (NGG) process prior, introduced by Regazzini, Lijoi, and Prünster (2003), encompassing the Dirichlet process. See Lijoi, Mena, and Prünster (2007) and Argiento, Guglielmi, and

Pievatolo (2010) for more details, and, in particular, for the analytic expression of the eppf under NGG mixture models. On the other hand, when the NGG process prior reduces to the Dirichlet measure, the eppf turns out to be a variant of Ewens sampling formula $p(n_1, \ldots, n_k) = (\Gamma(\alpha + 1)/\Gamma(\alpha + n))\alpha^{k-1} \prod_{j=1}^{k} (n_j - 1)!$, with $n = n_1 + \cdots + n_k$.

Representation (3) is useful to compute posterior inference. On the other hand, formulation (1)–(2) is the most expressive here, since the random parameter contains $\boldsymbol{\rho}$, which is the object of the statistical analysis.

Finally, observe that equivalence between models (1)–(2) and (3) holds thanks to the natural clustering rule and identifiability of $\{f(\cdot, \theta), \theta \in \Theta\}$. By *natural clustering rule* we mean the following: given $\theta_1, \ldots, \theta_n$, $X_i$ and $X_j$ belong to the same cluster if, and only if, $\theta_i = \theta_j$. In this case we write $X_i \leftrightarrow X_j$. The partition $\boldsymbol{\rho} = \{C_1, \ldots, C_k\}$ represents the quotient set of the data label set $\{1, \ldots, n\}$ by the equivalence relation $\leftrightarrow$, and $\boldsymbol{\phi} = (\phi_1, \ldots, \phi_k)$ are the unique values among the $\theta_i$'s.

## 3. RELAXING THE EQUALITY CONSTRAINT IN THE CLUSTERING RULE

The sensitivity of cluster estimates to hyperparameters in species sampling mixture models is a well-known issue. First of all, when the tails of the "true" distribution are heavy, to fit the data, the Bayesian estimate tipically adopts many kernels to reconstruct the "true" density shape. In this case, a deeper analysis on the prior elicitation could be accomplished. See for instance, Argiento, Guglielmi, and Soriano (2013) and Griffin (2010). Second, if the "true" distribution has nonconvex contour lines, as in Section 5 here, the hierarchical mixture model generally yields cluster estimates where the cluster components do not represent real data clusters, unless the kernel density has a proper nonconvex shape.

To overcome these problems, here we propose a new rule to assign observations to clusters, relaxing the equality constraint imposed by the natural clustering rule under the species sampling mixture model (3). If $d(\cdot, \cdot)$ is any distance between densities, the natural clustering rule can be restated as

$$X_i \leftrightarrow X_j \quad \Leftrightarrow \quad d(f(\cdot, \theta_i), f(\cdot, \theta_j)) = 0$$

when the family $\{f(\cdot; \theta), \theta \in \Theta\}$ is identifiable. To relax the rule, instead of grouping elements whose kernel densities are equal, we assign those data whose densities are "close" to the same cluster.

*Definition 1.* Given a configuration $(\theta_1, \ldots, \theta_n)$, a threshold $\epsilon > 0$, and a distance between densities $d(\cdot, \cdot)$, two observations $X_i$ and $X_j$ are *directly reachable* if

$$d(f(\cdot, \theta_i), f(\cdot, \theta_j)) < \epsilon.$$

We write $X_i \overset{\epsilon}{\rightsquigarrow} X_j$; since transitivity does not hold in this case, $\overset{\epsilon}{\rightsquigarrow}$ is not an equivalence relation.

*Definition 2.* Given a configuration $(\theta_1, \ldots, \theta_n)$, a threshold $\epsilon > 0$, and a distance between densities $d(\cdot, \cdot)$, two observations are *reachable* if there exists a finite sequence

$X_{j_1}, \ldots, X_{j_m}$ such that

$$X_i \overset{\epsilon}{\rightsquigarrow} X_{j_1} \overset{\epsilon}{\rightsquigarrow} X_{j_2} \overset{\epsilon}{\rightsquigarrow} \cdots \overset{\epsilon}{\rightsquigarrow} X_{j_m} \overset{\epsilon}{\rightsquigarrow} X_j.$$

We write $X_i \overset{\epsilon}{\leftrightarrow} X_j$. It is straightforward to prove that $\overset{\epsilon}{\leftrightarrow}$ is an equivalence relation among the data, so that we define $\boldsymbol{\rho}_\epsilon = \{C_1^{(\epsilon)}, \ldots, C_m^{(\epsilon)}\}$ as the quotient set of $\{1, \ldots, n\}$ by $\overset{\epsilon}{\leftrightarrow}$. Here $m := m(\epsilon) \leq k$ denotes the new number of clusters.

These definitions are based on the DBSCAN algorithm (Ester, Kriegel, and Xu 1996). DBSCAN (*density-based spatial clustering of applications with noise*) is a well-known algorithm in the data mining community; in brief, it clusters data at hand through a notion of distance between items and two parameters, an integer $N$ and a positive real $\epsilon$. In this work we consider only the case $N = 1$, since if $N > 1$, the relation $\overset{\epsilon}{\leftrightarrow}$ induced by DBSCAN among the data labels is not an equivalence.

By DBSCAN($\{f(\cdot;\theta_1), \ldots, f(\cdot;\theta_n)\}, d, \epsilon$) we mean a deterministic function where the input values are: (i) $(\theta_1, \ldots, \theta_n)$, the latent variables in model (3) corresponding to the data; (ii) a distance $d$ between densities; and (iii) a threshold $\epsilon > 0$. The input values $(\theta_1, \ldots, \theta_n)$ can be equivalently described as $(\{C_1, \ldots, C_k\}, (\phi_1, \ldots, \phi_k))$, while (ii) can be substituted by a matrix of the distances between $f(\cdot;\phi_i)$ and $f(\cdot;\phi_j)$, $i, j = 1, \ldots, k$. The output values are: a partition $(C_1^{(\epsilon)}, \ldots, C_m^{(\epsilon)})$ of the index set $\{1, \ldots, n\}$, obtained grouping the subsets $\{C_1, \ldots, C_k\}$ according to the equivalence relation $\overset{\epsilon}{\leftrightarrow}$ given by Definitions 1 and 2, and the vectors $(\boldsymbol{\phi}_1^{(\epsilon)}, \ldots, \boldsymbol{\phi}_m^{(\epsilon)})$ (of the latent variables associated to each $C_j^{(\epsilon)}$, $j = 1, \ldots, m$) and $(\boldsymbol{n}_1^{(\epsilon)}, \ldots, \boldsymbol{n}_m^{(\epsilon)})$ (size vectors of the sets among $\{C_1, \ldots, C_k\}$ composing $C_1^{(\epsilon)}, \ldots, C_m^{(\epsilon)}$). Specifically, we are applying the deterministic DBSCAN procedure to the species sampling mixture model, obtaining a new model, called $b$-DBSCAN.

The new partition $\boldsymbol{\rho}_\epsilon = \{C_1^{(\epsilon)}, \ldots, C_m^{(\epsilon)}\}$ is such that, for each $h = 1, \ldots, m$:

$$C_h^{(\epsilon)} = C_{l_1(h)} \cup \cdots \cup C_{l_{k_h^\epsilon}(h)}, \qquad \{l_1(h), \ldots, l_{k_h^\epsilon}(h)\} \subseteq \{1, \ldots, k\}. \tag{4}$$

In brief, (4) states that every element $C_h^{(\epsilon)}$ of the partition $\boldsymbol{\rho}_\epsilon$ is a finite union of some elements of the partition $\boldsymbol{\rho}$, depending on $\epsilon$ and the index $h$. Let us consider now, for each $h = 1, \ldots, m$, the vector $\boldsymbol{\phi}_h^{(\epsilon)} := (\phi_{l_1(h)}, \ldots, \phi_{l_{k_h^\epsilon}(h)})$ and the vector $\boldsymbol{n}_h^{(\epsilon)} := (\#C_{l_1(h)}, \ldots, \#C_{l_{k_h^\epsilon}(h)}) = (n_{l_1(h)}, \ldots, n_{l_{k_h^\epsilon}(h)})$, and define $\boldsymbol{\phi}^{(\epsilon)} := (\boldsymbol{\phi}_1^{(\epsilon)}, \ldots, \boldsymbol{\phi}_m^{(\epsilon)})$, $\boldsymbol{n}^{(\epsilon)} := (\boldsymbol{n}_1^{(\epsilon)}, \ldots, \boldsymbol{n}_m^{(\epsilon)})$. Some calculations (see the Appendix in the supplementary materials) show that the $b$-DBSCAN model can be expressed as

$$\begin{aligned}
&X_1, \ldots, X_n | C_1^{(\epsilon)}, \ldots, C_m^{(\epsilon)}, \boldsymbol{\phi}^{(\epsilon)}, \boldsymbol{n}^{(\epsilon)} \sim \prod_{h=1}^{m} \tilde{f}\left(\boldsymbol{X}_{C_h^{(\epsilon)}}; \boldsymbol{\phi}_h^{(\epsilon)}, \boldsymbol{n}_h^{(\epsilon)}\right) \\
&\left(C_1^{(\epsilon)}, \ldots, C_m^{(\epsilon)}, \boldsymbol{\phi}^{(\epsilon)}, \boldsymbol{n}^{(\epsilon)}\right) = \text{DBSCAN}(\{C_1, \ldots, C_k\}, (\phi_1, \ldots, \phi_k), d, \epsilon) \\
&\phi_1, \ldots, \phi_k | k \overset{\text{iid}}{\sim} P_0 \qquad \boldsymbol{\rho} \sim \pi(\boldsymbol{\rho}) = p(\#C_1, \ldots, \#C_k),
\end{aligned} \tag{5}$$

where the density $\tilde{f}(\cdot; \boldsymbol{\phi}_h^{(\epsilon)}, \boldsymbol{n}_h^{(\epsilon)})$ is a finite mixture of densities in $\{f(\cdot, \theta), \theta \in \Theta\}$ and its expression has been given in the online Appendix (see (14)). Observe that we elicit the prior on the parameter of interest $(\boldsymbol{\rho}_\epsilon, \boldsymbol{\phi}^{(\epsilon)}, \boldsymbol{n}^{(\epsilon)}) := \text{DBSCAN}(\{C_1, \ldots, C_k\}, (\phi_1, \ldots, \phi_k), d, \epsilon)$ as

the prior induced by a deterministic transformation of $(\boldsymbol{\rho}, \boldsymbol{\phi})$. We refer to the three equations (5) as $b$-DBSCAN model in the rest of the article.

To make inference, we need to sample from the posterior distribution $\mathcal{L}(\boldsymbol{\rho}_\epsilon, \boldsymbol{\phi}^{(\epsilon)}, \boldsymbol{n}^{(\epsilon)}|\,\text{data})$ which, augmenting the state space, can be expressed as the marginal distribution of

$$\mathcal{L}\Big(\boldsymbol{\rho}_\epsilon, \boldsymbol{\phi}^{(\epsilon)}, \boldsymbol{n}^{(\epsilon)}, \boldsymbol{\rho}, \boldsymbol{\phi}|\text{data}\Big) = \mathcal{L}\Big(\boldsymbol{\rho}_\epsilon, \boldsymbol{\phi}^{(\epsilon)}, \boldsymbol{n}^{(\epsilon)}|\boldsymbol{\rho}, \boldsymbol{\phi}, \text{data}\Big)\, \mathcal{L}(\boldsymbol{\rho}, \boldsymbol{\phi}|\text{data}). \tag{6}$$

Since $(\boldsymbol{\rho}_\epsilon, \boldsymbol{\phi}^{(\epsilon)}, \boldsymbol{n}^{(\epsilon)})$ is a deterministic function of $(\boldsymbol{\rho}, \boldsymbol{\phi})$, the first factor on the right-hand side of (6) is degenerate on DBSCAN$(\{C_1, \ldots, C_k\}, (\phi_1, \ldots, \phi_k), d, \epsilon)$, while the second is the posterior distribution of the parameter $(\boldsymbol{\rho}, \boldsymbol{\phi})$ in models (1)–(2).

In the rest of the article we fix $\epsilon$, without assuming it random. In fact, from the proof of (5) (see (10) in the online Appendix), it is clear that, conditionally to $(\boldsymbol{\rho}, \boldsymbol{\phi})$, the distribution of data does not depend on $\epsilon$. This also implies that, as far as density estimation is concerned, models (3) and (5) are equivalent. As a consequence, posterior computation (for fixed $\epsilon$) is standard here by means of the augmentation step (6). However, to be "transparent," the MCMC scheme and the Bayesian estimate of the random partition $\boldsymbol{\rho}_\epsilon$ are briefly described in the Appendix (see the online Appendix B, Computational Details).

## 4. CLUSTERING VALIDATION TOOLS

It is clear that one of the main issues in our approach is the choice of $\epsilon$. As the application sections will show, this hyperparameter strongly affects the posterior cluster estimate. Here we propose to fix $\epsilon$ to optimize some suitable posterior functionals. Our approach is the following: (a) fix a grid of values $\epsilon_1, \ldots, \epsilon_r$; (b) evaluate the posterior expectation $\mathbb{E}(\mathcal{H}(\boldsymbol{\rho}_{\epsilon_j})|\text{data})$ for a suitable function $\mathcal{H}$ for $j = 1, \ldots, r$; (c) choose the optimal $\epsilon_j$ among $\epsilon_1, \ldots, \epsilon_r$. As $\mathcal{H}$, we use either standard tools as the silhouette coefficient (Rousseeuw 1987) and the adjusted Rand index (Hubert and Arabie 1985), or new indexes built from the posterior predictive distribution under our model.

As far as the latter indexes are concerned, let $X_{\text{new}}$ be a new observation from (1)–(2), and, for $i = 1, \ldots, n$, let $Y^\epsilon_{\text{new}}(X_i) = 1$ if $X_{\text{new}}$ is in the same cluster as $X_i$, and 0 otherwise. Therefore, for each $i$, we consider $\mathcal{L}(X_{\text{new}}|Y^\epsilon_{\text{new}}(X_i) = 1, \text{data})$, that is the posterior predictive law of a new observation conditionally to the event that this observation share the same cluster with $X_i$. For a fixed $\epsilon$, we compute conditional posterior predictive residuals defined as

$$r_i^{(\epsilon)} := r_i = \frac{X_i - \mathbb{E}\big(X_{\text{new}}|Y^\epsilon_{\text{new}}(X_i) = 1, \text{data}\big)}{\big(\text{var}\big(X_{\text{new}}|Y^\epsilon_{\text{new}}(X_i) = 1, \text{data}\big)\big)^{1/2}}. \qquad i = 1, \ldots, n. \tag{7}$$

When data are multivariate, the previous square root of the matrix stands for its Cholesky decomposition matrix. For each data component $j = 1, \ldots, p$, we compute

$$\text{Ind}_j^{(\epsilon)} := \frac{1}{n}\sum_{i=1}^{n} r_{i,j}^2, \tag{8}$$

which represents a predictive goodness-of-fit index of the $b$-DBSCAN-mixture model on the $j$th data component as a function of $\epsilon$.

Moreover, we consider the following posterior predictive probabilities, for any fixed $\epsilon > 0$:

$$\mathbb{P}\left(Y^{\epsilon}_{\text{new}}(X_i) = 1 | X_{\text{new}} = X_i, \text{data}\right), \quad i = 1, \ldots, n. \tag{9}$$

In words, for each $i$, (9) is the probability that a new observation is assigned to the same cluster as $X_i$, conditionally to the event that $X_{\text{new}}$ and $X_i$ assume exactly the same value. However, for a fixed $i$, the value assumed by such an index cannot be interpreted "per se," but it must be compared to all the other values ($j \neq i$). High values denote that $X_i$ is "nested" in its cluster, while small values suggest that $X_i$ is a "frontier" point in the cluster it has been assigned to. Hence, those probabilities have an interpretation as misclassification indexes.

The online Appendix shows how to compute (7) and (9) by means of a posterior sample of $(\boldsymbol{\rho}, \boldsymbol{\phi})$.

## 5. SIMULATED DATA

In this section we illustrate the proposed model with application to a simulated dataset of size $n = 1000$. In particular, we simulated iid observations from a mixture of bivariate densities (see Figure 1); there are two main groups of observations, from the two components of the mixture: one has a sharp round shape and it is located around the point (0, 0) (cluster $A$), while the other lays on a semicircular region on the right of the previous group (cluster $B$). The Bayesian cluster estimates here (and in the next section) are defined as the partition minimizing posterior Binder's loss function with equal misclassification costs; see the online Appendix B in the supplementary materials for more details.

### 5.1 $b$-DBSCAN Clustering

As far as the $b$-DBSCAN mixture model (5) is concerned, we assume Gaussian kernels for $\{f(\cdot;\theta)\}$, while the mixing measure is the Dirichlet process. In particular, we complete the prior specification by assuming a prior $\alpha \sim \text{gamma}(\gamma_1, \gamma_2)$, while $P_0(d\theta) = N(d\mu|m_0, \Sigma/\kappa_0) \times \text{Inv-Wishart}(d\Sigma|\nu_1, \Psi_1)$, where $\theta = (\mu, \Sigma)$. Here, $\mathbb{E}(\alpha) = \gamma_1/\gamma_2$ and $\mathbb{E}(\Sigma) = \Psi_1/(\nu_1 - p - 1)$. At first, we fixed $\epsilon = 0$, that is, when model (5) reduces to a Dirichlet process mixture (DPM) model. As far as the hyperparameters are concerned, we have performed a robustness analysis, choosing different values for $\gamma_1$, $\gamma_2$, $m_0$, $\kappa_0$, $\nu_1$, and $\Psi_1$. The analyses, not reported here, show that, regardless of the values of the hyperparameters, the estimated number of clusters is larger than the true one (i.e., 2). This is an expected result, since many Gaussian densities are needed to fit cluster $B$ (see Figure 1(a)).

On the other hand, when $\epsilon$ is larger than 0, to make the $b$-DBSCAN model more flexible, we recommend to fix hyperparameters so that the conditional variance of $f(\cdot;\theta)$ is small, and the prior expected number of mixture components is large. In particular, we fixed $\gamma_1$ and $\gamma_2$ such that $E(\alpha) = 11$, $\text{var}(\alpha) = 4$, $m_0 = (0, 0)^T$, $\kappa_0 = 0.001$, $\nu_1 = 10$ and $\Psi_1 = \text{diag}(0.1)$. Of course, we also need to choose a distance $d(\cdot, \cdot)$ between the distributions. Here we report estimates when $d$ is the $L^2$ distance and the Kullback–Leibler I-divergence (it can be symmetrized to become a pseudodistance). Figures 1 and 2 show the estimates for different values of $\epsilon$ under these distances. As we expected, the estimated number of

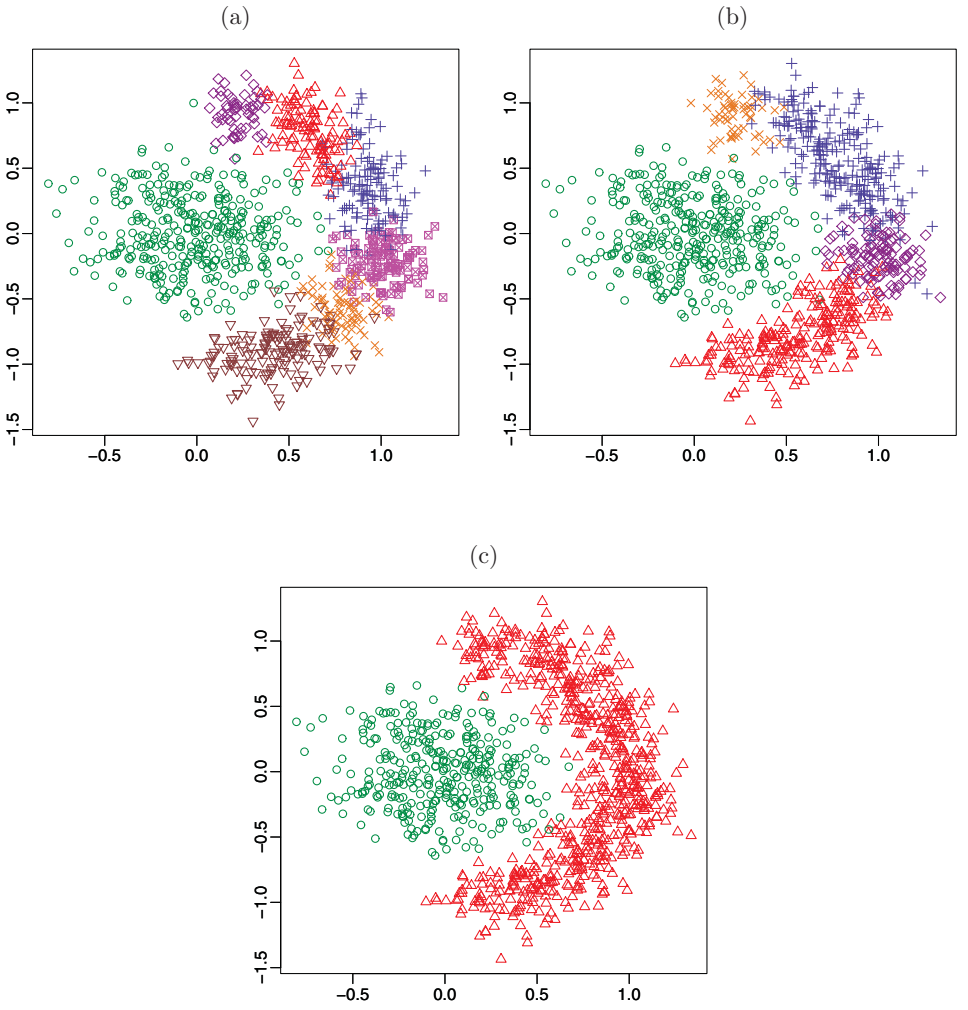


Figure 1. Simulated dataset: posterior estimates of the random partition when $d$ is the symmetrized Kullback–Leibler I-divergence; (a): $\epsilon = 0$, $\hat{m} = 7$, (b): $\log(1+\epsilon) = 2.5$, $\hat{m} = 5$, (c): $\log(1+\epsilon) = 3$, $\hat{m} = 2$. Each symbol in the graph represents a cluster.

clusters $\hat{m}$ reduces as $\epsilon$ increases: in fact, the model groups the DPM clusters into new bigger clusters.

The choice of the distance can greatly affect the posterior cluster estimate: for the Kullback–Leibler I-divergence, as well as for the Hellinger distance case (not reported here), as $\epsilon$ increases, groups with similar mean parameters are merged, and we find good posterior estimates (Figure 1(c)). In contrast, under the $L^2$ distance, as $\epsilon$ increases, groups with similar covariance matrices are merged, and in this case the clusters follow a different grouping path, leading to unsatisfactory estimated partitions (Figure 2(c)). This is clear since, for fixed variances, the $L^2$ distance is sensitive to differences between the means, while the Kullback–Leibler I-divergence, for fixed means, is sensitive to differences between the variance matrices. Mathematical details are given in the online Appendix D.

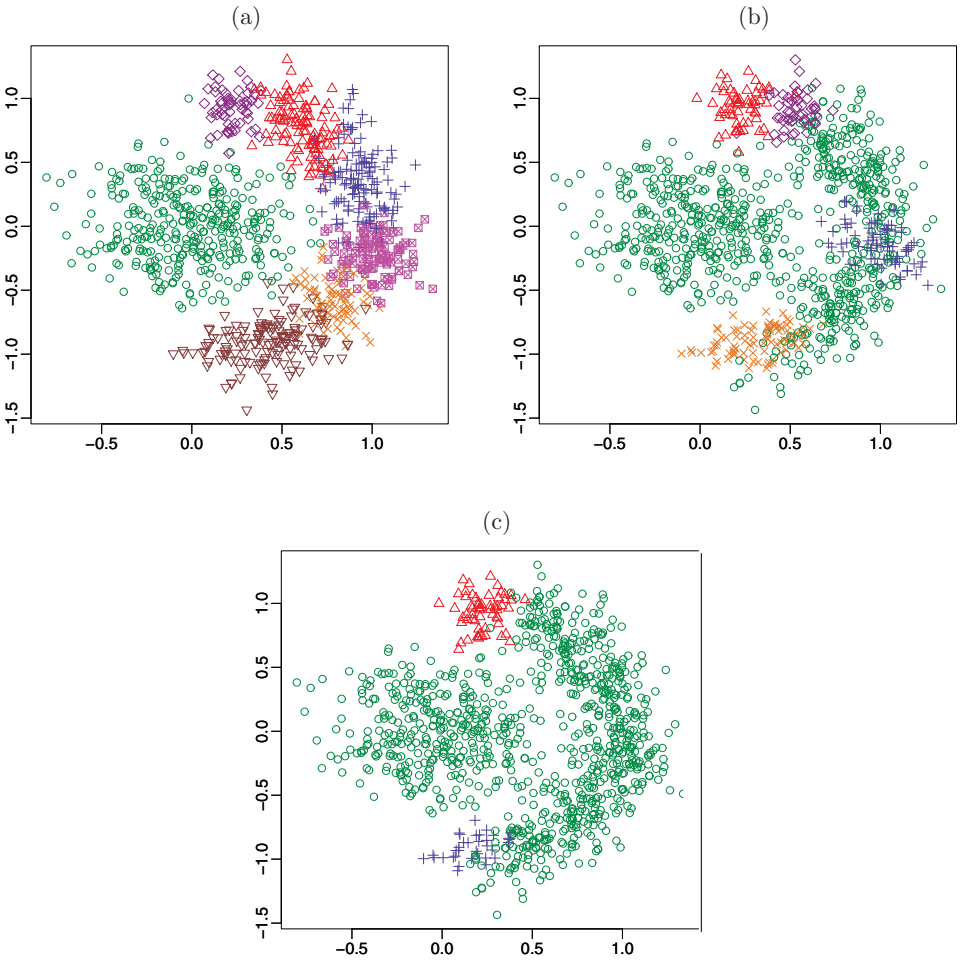


Figure 2. Simulated dataset: posterior estimates of the random partition when $d$ is the $L^2$ distance; (a): $\epsilon = 0$, $\hat{m} = 7$, (b): $\epsilon = 1.925$, $\hat{m} = 5$, (c): $\epsilon = 2.2$, $\hat{m} = 3$. Each symbol in the graph represents a cluster.

To choose a value for $\epsilon$, for instance when $d$ is the symmetrized Kullback–Leibler I-divergence, we fixed a grid of values of $\epsilon$, that is, $\log(1+\epsilon) \in \{0.5, 1.5, 2, 2.5, 2.75, 3, 3.5, 4\}$. For each $j = 1, \ldots, 8$ we computed $\mathbb{E}(\mathcal{H}(\boldsymbol{\rho}_{\epsilon_j})|\text{data})$ through the MCMC method, where $\mathcal{H}$ is the silhouette coefficient or the adjusted Rand index. Figure 3(a) shows the two posterior functionals, as $\epsilon$ varies. Both lines lead to the same conclusion: $\log(1+\epsilon) = 3$ is the optimal choice. Figure 1(c) shows that, under this choice, our estimate is very close to the true partition.

We computed the misclassification error under the optimal estimated partition. We found that 337 points in cluster $A$, and 644 in cluster $B$, were correctly classified; the overall misclassification rate is 1.9%. Moreover we computed the misclassification probability index (9) for each data points. To compare all these individual values, we first computed $q_{0.25}$ and $q_{0.75}$, the first and the third sample quartile of the values of the index. We classified as boundary points all data such that the probability index was smaller than

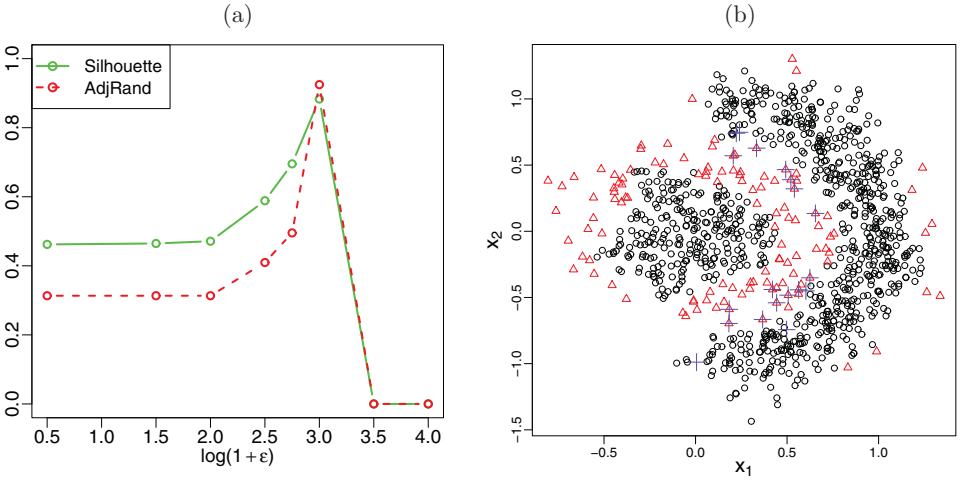


Figure 3. Simulated dataset: posterior expectation of the adjusted Rand and silhouette indexes as a function of $\epsilon$ (panel (a)) and misclassification graph under the optimal estimated partition (panel (b)).

$q_{0.25} - 1.5(q_{0.75} - q_{0.25})$. A summary of the results is depicted in Figure 3(b), where the boundary points are represented by (red) triangles, while misclassified data are represented by (blue) crosses. Observe as misclassified data lie in the middle of the two main groups, where there is uncertainty between the two clusters membership.

### 5.2 Comparison With Other Methods

As we mentioned in the Introduction, there are many model-based and heuristic clustering techniques that can be compared with the Bayesian cluster estimate under the $b$-DBSCAN mixture model. Among the latter, our main "competitors" are the heuristic DBSCAN algorithm, since ours is its Bayesian nonparametric mixture counterpart, and the hierarchical and $k$-means algorithms (widely known even outside the statistical community). On the other hand, we showed in (5) that our model can be interpreted as a mixture of mixtures; hence, as a (Bayesian) model-based alternative, we have considered the nested Dirichlet process (NDP) mixture model (Rodriguez, Dunson, and Gelfand 2008), that has a similar mixture interpretation.

To apply the heuristic DBSCAN procedure to the simulated data, we considered the Euclidean distance among points in $\mathbb{R}^2$ to build clusters and fixed the minimum number $N$ of points to define a group to be a cluster equal to 1 and 6. In Figure 4(a) and (b) we report clustering results for two different values of $(N, \epsilon)$, choosing these two pairs among those better reflecting the true partition. When $N = 1$, noise elements are not allowed, and the algorithm identifies many "singleton" clusters (panel (a)), in addition to two main ones. In contrast, when $N = 6$, less clusters are found, but many points are classified as "noise" by the algorithm; see the diamonds in Figure 4(b). In the same plot, the three full squares seem to form a cluster on their own, despite that $N = 6$ is the minimum number of points to define a cluster. The inconsistency is a consequence of the nonuniqueness of the DBSCAN-partition when $N > 1$.

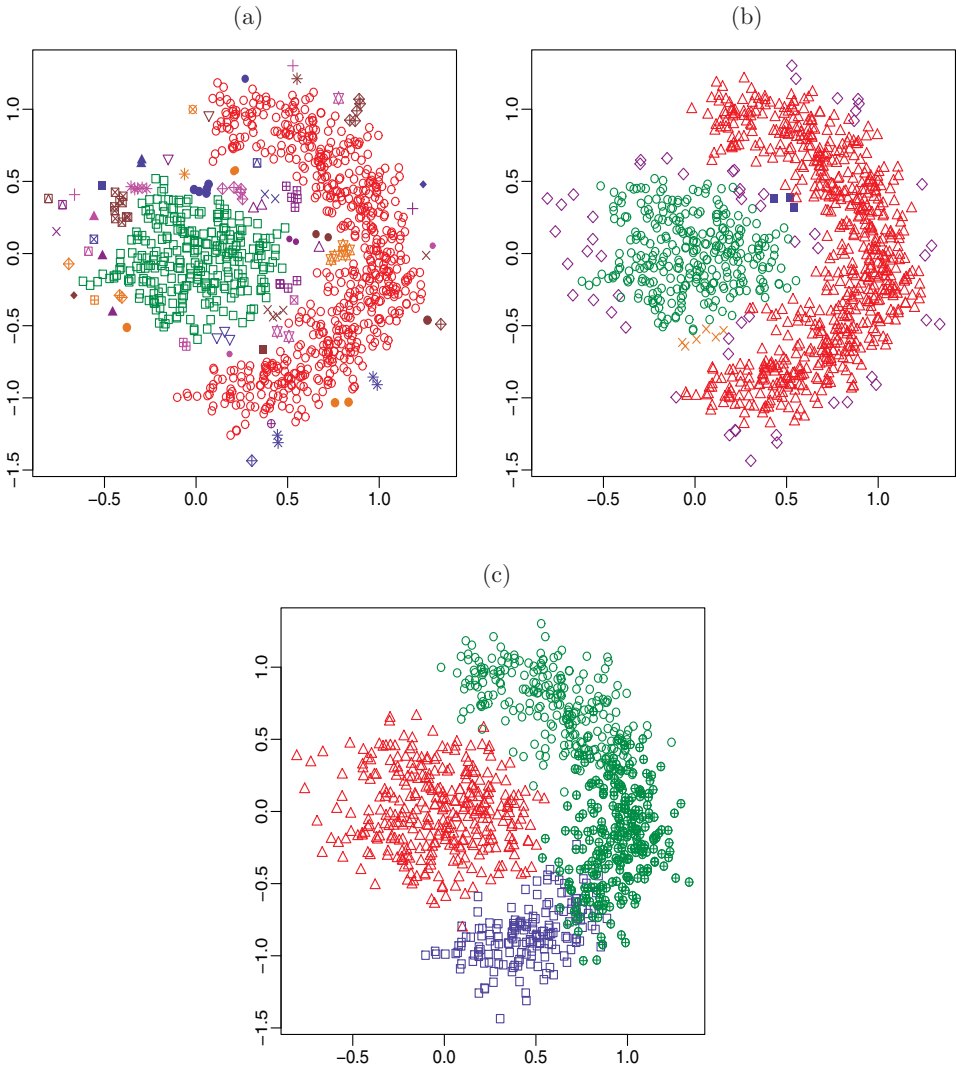


Figure 4. Simulated dataset: heuristic DBSCAN clustering results of the simulated dataset when $N = 1$, $\epsilon = 0.075$ (a) and $N = 6$, $\epsilon = 0.1$ (b), NDP mixture model estimate (c). The empty and crossed circles in (c) belong to the same cluster, but to different subclusters of the NDP mixture model.

On the other hand, we considered the following NDP mixture model:

$$X_i|G_i \overset{\text{ind}}{\sim} \int f(\cdot;\theta)G_i(d\theta) \qquad G_i|Q \overset{\text{iid}}{\sim} Q \quad i = 1,\ldots,n,$$

where $Q \overset{\text{a.s.}}{=} \sum_1^\infty \omega_l \delta_{\tau_l}$, $\{\omega_l\}$ are Dirichlet process stick breaking weights with parameter $\alpha > 0$ and the location points $\tau_l$'s are iid from a Dirichlet process with parameters $\beta > 0$ and probability measure $P_0$. Here $f(\cdot;\theta)$ is the Gaussian kernel and the mean mixing distribution $P_0$ is assigned as under the $b$-DBSCAN model. Under this model, the prior cluster

Table 1. Adjusted Rand index for the cluster estimates under different techniques

| | Fixed $m$ | | | | |
|---|---|---|---|---|---|
| | 2 | 3 | 4 | | |
| $k$-means | 0.005 | 0.474 | 0.364 | DBSCAN ($N = 1, \epsilon = 0.075$) | 0.789 ($\hat{m} = 2$) |
| Hierarchical - Single | −0.001 | 0.001 | 0.006 | DBSCAN ($N = 6, \epsilon = 0.1$) | 0.826 ($\hat{m} = 3$) |
| Hierarchical - Complete | 0.278 | 0.136 | 0.329 | NDP mixture | 0.625 ($\hat{m} = 3$) |
| Hierarchical - Average | 0.057 | 0.549 | 0.545 | $b$-DBSCAN | 0.925 ($\hat{m} = 2$) |

assignment consists of two levels of hierarchy: one assigning data to random mixing distributions (and forming subclusters), and the other grouping the mixing distributions themselves, building clusters. However, even if this model yields clustering on both observations and distributions, here we are interested only in grouping the distributions $\{G_1, \ldots, G_n\}$, which should represent *clusters of subclusters*. We completed the prior specification fixing $\alpha$ so that the prior number of clusters of distributions is 2 and the parameter $\beta$ is a priori gamma-distributed; we did some robustness analysis on hyperparameters, and we report the best cluster estimate in Figure 4(c). It is clear that the model is not able to allocate data points which belong to cluster B in one single cluster.

To finish up and summarize the comparison, Table 1 displays the adjusted Rand index between the true clustering and the estimates under different clustering techniques or models. For $k$-means and hierarchical algorithms, where the number $m$ of clusters must be fixed in advance, we choose $m = 2, 3, 4$, while for the other methods the table displays (in parenthesis) the estimated number of clusters. Table 1 points out that $k$-means and hierarchical algorithms are not able to recover cluster B; this conclusion is also supported by the plots of cluster estimates, not included here. The heuristic DBSCAN performs pretty well, however finding many singleton or noise clusters (see Figure 4(a) and (b)). On the other hand, the $b$-DBSCAN performs outstandingly well for this dataset.

In conclusion, we recommend Bayesian nonparametric clustering models, first because they seem to be more effective (as shown by Table 1), and second because inference is "richer," that is, not limited to point estimation. Moreover, not only we are able to assign a data point to a cluster, but we also quantify the uncertainty among cluster assignments through posterior predictive probabilities (9).

## 6. YEAST CELL CYCLE DATA

We fitted our model to a dataset, very popular in the literature on clustering of gene expression profiles, usually called yeast cell cycle data (see Cho et al. 1998). A gene expression dataset from a microarray experiment can be represented by a real-valued matrix $[X_{ij},\ 1 \le i \le n, 1 \le j \le p]$, where the rows $(X_1, \ldots, X_n)$ contain the expression patterns of genes and are our data points. Each cell $X_{ij}$ is the measured expression level of gene $i$ in sample (or at time) $j$. The yeast cell cycle data contain $n = 389$ gene expression profiles, observed at 17 different time values, one every 10 min from time

zero. We chose only some components of the whole expressions, representing the second cell cycle ($j = 9, \ldots, 17$). The final dataset ($n = 389$, $p = 9$) has been obtained by standardizing each row of the gene expression matrix to have zero mean and unit variance. By visual inspection, Cho et al. (1998) grouped the data according the peak times of expression levels; see Figure 5. They detected five peaking points, corresponding to five phases of the cell cycle. As in the previous example, we assume the Gaussian kernel as $f(\cdot;\theta)$ and the Dirichlet process as mixing measure. The latent variable here is $\theta = (\mu, \sigma^2)$, where $\mu$ is the mean and $\sigma^2\mathbb{I}_p$ is the covariance matrix of the Gaussian distribution. Moreover, conditionally on the total mass parameter $\alpha$, $P \sim \text{Dirichlet}(\alpha, P_0)$, with $\alpha \sim \text{gamma}(\gamma_1, \gamma_2)$, and $P_0(d\mu, d\sigma^2) = N(d\mu|m_0, \sigma^2/\kappa_0\mathbb{I}_p) \times \text{inv-gamma}(d\sigma|a, b)$. Observe that the Gaussian kernel densities have diagonal covariance matrices, which greatly simplifies computation, since only diagonal matrices must be inverted in the MCMC algorithm. On the other hand, this implies that data are modeled from a mixture of Gaussian kernels with spherical contour lines, a very strong assumption when $\epsilon = 0$, but not in case $\epsilon > 0$, when the clusters are modeled as finite unions of round shaped groups.

As far as the choice of the hyperparameters is concerned, to make the model more flexible, we fixed them so that the prior number of mixture components is large. In particular, we fixed $(m_0, \kappa_0, a, b)$ so that the prior variance for $\mu$ is large ($10\mathbb{I}_p$), but the prior mean and variance of $\sigma^2$ are small (both equal to 0.1). Furthermore, we set $(\gamma_1, \gamma_2) = (2, 0.01)$, to obtain a vague prior for the total mass parameter $\alpha$.

In our experiments we considered $d$ as the Kullback–Leibler I-divergence (the Hellinger distance provided very similar results). We applied the $\epsilon$-strategy described in Section 4: we fixed a grid of values of $\epsilon$ and computed $\text{Ind}_j^{(\epsilon)}$ in (8) for each $j \in \{10, 11, 12, 14, 16\}$, that is, when data have peaks according to Cho et al. (1998). We found that, except for the late G1 phase (Figure 5(b), $j = 11$), all the index trajectories have a minimum around $\epsilon = 2.8$; furthermore, the sum of the indexes has a minimum exactly at $\epsilon = 2.8$. Figure 6 shows our cluster estimate for such a value of $\epsilon$. Observe that we found eight clusters, instead of five detected by the reference partition in Figure 5. Our clusterization takes into account not only the different peaks, but also the entire trajectories of the gene expressions. As a final remark, note that a positive feature of our procedure is the ability to classify nonstandard data: see the "outlier" trajectories grouped into clusters 7 and 8 (Figure 6, panels (g) and (h)). We have also checked robustness of these results to different choices of hyperparameters; for brevity, this analysis is not reported here.

Finally, we made a comparison with some heuristic methods ($k$-means and hierarchical). As criterion that applies under both heuristic and model-based techniques, we evaluated the silhouette coefficient of the estimated clustering. We found that, for the same number $m$ of clusters (estimated or fixed), when $m = 2, \ldots, 11$, the silhouette coefficient prefers our model. In particular, the silhouette coefficient of our "best" cluster estimate ($\epsilon = 2.8$, $\hat{m} = 8$) is 0.282, while it is at most 0.187 for the heuristic models with $m = 8$. On the other hand, if $m = 5$ the best silhouette coefficient of the heuristic algorithms is 0.259, while its value for the $b$-DBSCAN estimates when $\epsilon = 3.15$ ($\hat{m} = 5$) is 0.372. The silhouette coefficient of the reference partition is 0.129.

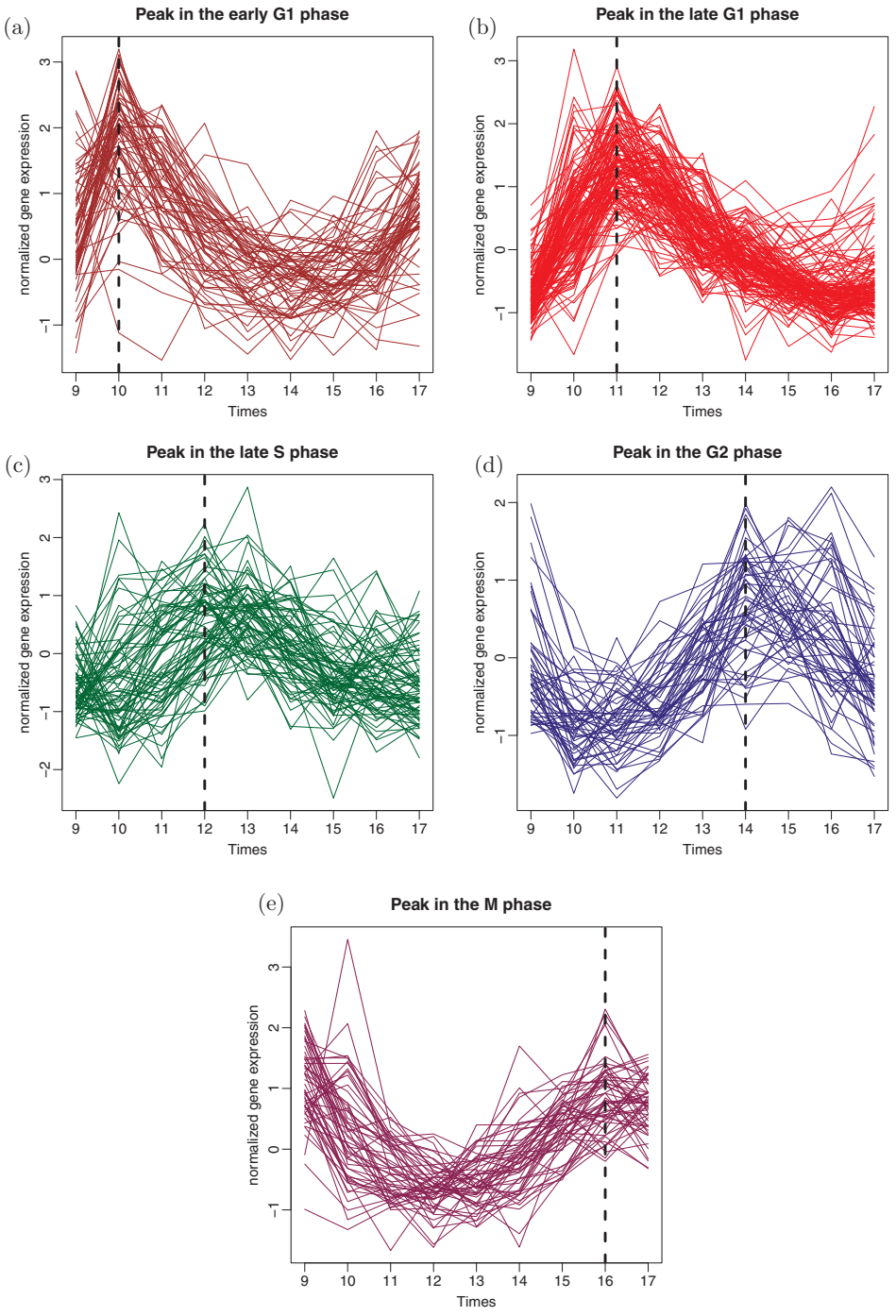


Figure 5. Yeast cell cycle dataset: reference partition by Cho et al. (1998).

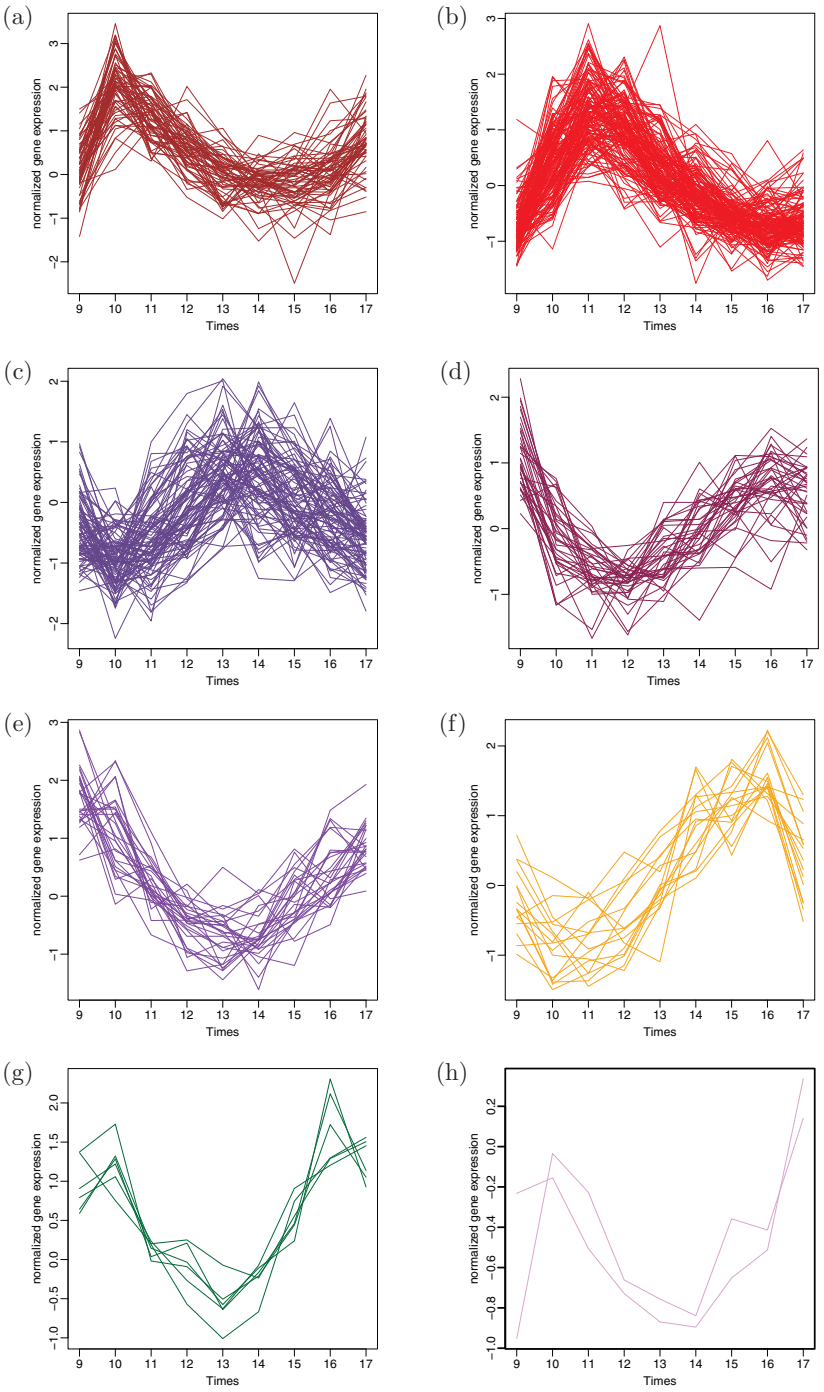


Figure 6. Yeast cell cycle dataset: optimal posterior cluster estimate.

## 7. DISCUSSION

We have presented a Bayesian nonparametric framework for model-based clustering. Data have been initially modeled through a species sampling mixture model. The core of our work lies in defining the data partition parameter in a new way: two observations are in the same cluster if the distance between densities corresponding to their latent parameters is smaller than a threshold $\epsilon$. This definition is made mathematically coherent introducing the *reachability* property in Definitions 1 and 2. We call the proposed model *b*-DBSCAN mixture. This model can be interpreted as a mixture whose components are themselves mixtures of parametric densities (for instance, Gaussian kernels). Crucial ingredients are the (pseudo-)distance $d$ between densities and the hyperparameter $\epsilon$.

We have discussed implementation and applications of the *b*-DBSCAN mixture model to two datasets. Our model works very well; however, it must be mentioned that, as with all Bayesian nonparametric models, it has a high computational cost, compared to heuristic clustering techniques. As far as the elicitation of $\epsilon$ is concerned, we have suggested a strategy to fix it, as the optimal value of the posterior expectation of a function of the random partition.

An interesting property of our model is that it is not directly affected by "curse of dimensionality." Even when data are high dimensional, as long as the underlying species sampling mixture model is able to detect the subclusters which are candidate to be grouped, our model works well. The computational burden due to high dimensionality of the probability distributions remains unchanged from the underlying species sampling mixture models. In fact, under the *b*-DBSCAN model, reclustering of data is achieved through distances among densities $\{f(\cdot, \theta)\}$, which are scalar objects, whatever is the dimension of data ($p$) or of the parameter $\theta$ ($s$).

In the two applications here, we have always assumed a DPM model. Interested readers should refer to Cremaschi (2012) for an application with a proper NGG-mixture model. Besides, in this work, we have focused on Gaussian kernels, but of course other parametric families could be fixed as well. The choice of the Gaussian distribution is essentially due to nice theoretical properties (mixtures of Gaussians are dense in the space of densities on an Euclidean space), low computational effort (conjugacy), and closed form of some distances ($L^2$, Kullback–Leibler and Hellinger).

Finally, extensions to the current approach include further work on the elicitation of $\epsilon$ and categorical formulations of this clustering model. These and other topics are the subject of current research.

## SUPPLEMENTARY MATERIALS

**Data and codes:** Data, R, and C codes to perform posterior inference under *b*-DBSCAN model for the two applications in Sections 5 and 6. Please read file README contained in the tar.gz file for more details. (DataCodes-bDBSCAN.tar.gz)

**Appendix:** The supplementary files include the Appendix, which gives the derivation of (5) (Appendix A), computational details on the MCMC algorithm to compute posterior inference under the *b*-DBSCAN model (Appendix B), computation of (7) and (9)

(Appendix C), mathematical details about derivation of the $KL$, and $L^2$ distances between Gaussian densities (Appendix D). (FourAppendices.pdf)

## ACKNOWLEDGMENTS

The authors thank the two reviewers and the Editor for their valuable help in improving this article.

*[Received December 2012. Revised July 2013]*

## REFERENCES

Argiento, R., Guglielmi, A., and Pievatolo, A. (2010), "Bayesian Density Estimation and Model Selection Using Nonparametric Hierarchical Mixtures," *Computational Statistics and Data Analysis*, 54, 816–832. [1129]

Argiento, R., Guglielmi, A., and Soriano, J. (2013), "A Semiparametric Bayesian Generalized Linear Mixed Model for the Reliability of Kevlar Fibres," *Applied Stochastic Models in Business and Industry*, 29, 410–423. [1129]

Cho, R. J., Campbell, M. J., Winzeler, E. A., Steinmetz, L., Conway, A., Wodicka, L., Wolfsberg, T. G., Gabrielian, A. E., Landsman, D., Lockhart, D. J., and Davis, R. W. (1998), "A Genome-Wide Transcriptional Analysis of the Mitotic Cell Cycle," *Molecular Cell*, 2, 65–73. [1128,1137]

Cremaschi, A. (2012), "Model-Based Clustering via Bayesian Nonparametric Mixture Models," Tesi di laurea magistrale, Ingegneria Matematica, Politecnico di Milano. [1141]

Ester, M., Kriegel, H. P., and Xu, X. (1996), "Knowledge Discovery in Large Spatial Databases: Focusing Techniques for Efficient Class Identification," in *Proceedings of the 4th International Symposium on Large Spatial Databases, Portland, ME, 1995, Lecture Notes in Computer Science*, volume 951, New York: Springer, pp. 67–82. [1127,1130]

Griffin, J. (2010), "Default Priors for Density Estimation With Mixture Models," *Bayesian Analysis*, 5, 45–64. [1129]

Hubert, L. J., and Arabie, P. (1985), "Comparing Partitions," *Journal of Classification*, 2, 193–218. [1131]

Johnson, S. (1967), "Hierarchical Clustering Schemes," *Psychometrika*, 32, 241–254. [1127]

Lau, J. W., and Green, P. J. (2007), "Bayesian Model Based Clustering Procedures," *Journal of Computational and Graphical Statistics*, 16, 526–558. [1127]

Lijoi, A., Mena, R. H., and Prünster, I. (2007), "Controlling the Reinforcement in Bayesian Nonparametric Mixture Models," *Journal of the Royal Statistical Society*, Series B, 69, 715–740. [1128]

MacQueen, J. (1967), "Some Methods for Classification and Analysis of Multivariate Observations," in *Proceedings of 5th Berkeley Symposium on Mathematical Statistics and Probability*, volume 1, University of California Press, pp. 281–297. [1127]

McLachlan, G., and Peel, D. (2000), *Finite Mixture Models*, Hoboken, NJ: Wiley. [1126]

Pitman, J. (1996), "Some Developments of the Blackwell-Macqueen Urn Scheme," in *Statistics, Probability and Game Theory: Papers in Honor of David Blackwell,* volume 30 of *IMS Lecture Notes-Monograph Series*, eds. T. S., Ferguson, L. S. Shapley, and M. J. B., Hayward, CA: Institute of Mathematical Statistics, pp. 245–267. [1128]

Regazzini, E., Lijoi, A., and Prünster, I. (2003), "Distributional Results for Means of Random Measures With Independent Increments," *The Annals of Statistics*, 31, 560–585. [1128]

Rodriguez, A., Dunson, D. B., and Gelfand, A. E. (2008), "The Nested Dirichlet Process," *Journal of the American Statistical Association*, 103, 1131–1154. [1127,1135]

Rousseeuw, P. A. (1987), "Silhouettes: A Graphical Aid to the Interpretation and Validation of Cluster Analysis," *Journal of Computational and Applied Mathematics*, 20, 53–65. [1131]