From Dirichlet Process mixture models to spectral clustering

Stefano Tonellato

Department of Economics

Università Ca' Foscari Venezia

Abstract

This paper proposes a clustering method based on the sequential estimation of the random partition induced by the Dirichlet process. Our approach relies on the Sequential Importance Resampling (SIR) algorithm and on the estimation of the posterior probabilities that each pair of observations are generated by the same mixture component. Such estimates do not require the identification of mixture components, and therefore are not affected by label switching. Then, a similarity matrix can be easily built, allowing for the construction of a weighted undirected graph, where nodes represent individuals and edge weights quantify the similarity between pairs of individuals. The paper shows how, in such a context, spectral clustering techniques can be applied in order to identify homogeneous groups.

Keywords

Dirichlet process priors, sampling importance resampling, weighted graph, laplacian, spectral clustering

JEL Codes

C11, C38

1 Dirichlet process mixtures and clustering

A very important class of models in Bayesian nonparametrics is based on the Dirichlet process and is known as Dirichlet process mixture Antoniak (1974). In this model, the observable random variables, X_i, $i = 1, \ldots, n$, are assumed to be exchangeable and generated by the following hierarchical model:

$$\begin{aligned} X_i|\theta_i &\overset{ind}{\sim} p(\cdot|\theta_i),\ \theta_i \in \Theta \\ \theta_i|G &\overset{iid}{\sim} G \\ G &\sim DP(\alpha, G_0), \end{aligned}$$

where $DP(\alpha, G_0)$ denotes a Dirichlet process (DP) with base measure G_0 and precision parameter $\alpha > 0$. Since the DP generates almost surely discrete random measures on the parameter space Θ, ties among the parameter values have positive probability, leading to a batch of clusters of the parameter vector $\theta = [\theta_1, \ldots, \theta_n]^T$. Exploiting the Pólya urn representation of the DP, the model can be rewritten as

$$\begin{aligned} X_i|s_i, \theta^*_{s_i} &\overset{iid}{\sim} p(\cdot|\theta^*_{s_i}),\ \theta^*_{s_i} \in \Theta && (1) \\ \theta^*_{s_i} &\overset{iid}{\sim} G_0 && (2) \\ p(s_i = j|\mathbf{s}_{<i}) &= \begin{cases} \frac{\alpha}{\alpha+i-1} & j = k \\ \frac{n_j}{\alpha+i-1} & j \in \{k-1\}, \end{cases} && (3) \\ s_i &\perp \theta^*_j \qquad \forall i, j, && (4) \end{aligned}$$

where $\{k\} = \{1, \ldots, k\}$, $\mathbf{s}_{<i} = \{s_j,\ j \in \{i-1\}\}$ (in the rest of the paper, the subscript $< i$ will refer to those quantities that involve all the observations $X_{i'}$ such that $i' < i$), $s_j \in \{k\}$ for $j \in \{k-1\}$, and n_j is the number of θ_i's equal to θ^*_j. In this model representation, the parameter θ can be expressed as $(\mathbf{s}, \theta^*)$, with $\mathbf{s} = \{s_i : s_i \in \{k\},\ i \in \{n\}\}$, $\theta^* = [\theta^*_1, \ldots, \theta^*_k]^T$ with $\theta^*_j \overset{iid}{\sim} G_0$, and $\theta_i = \theta^*_{s_i}$. Consequently, the marginal distribution of X_i is a mixture with k components, where k is an unknown random integer.

In the case of finite mixtures with k components, with k fixed and known, under a frequentist perspective it would be quite straightforward to cluster the data by maximising the probability of the allocation of each datum to one of the k components, conditionally on the observed sample (McLachlan and Peel, 2000). Under a Bayesian perspective, the same results can be achieved, provided that either some identifiability constraints on the parameters are introduced, or a suitable risk function is minimised (Stephens, 2000). Unfortunately, under the assumptions we made, such computations are not feasible even numerically, due to the well known label switching problem (Frühwirth-Schnatter, 2006) that persists when the number of mixture components is not known, nor finite, as in the case of Dirichlet process mixtures. Nevertheless, equations (1)–(4) are very helpful in estimating posterior pairwise similarities and building hierarchical clustering algorithms as in Medvedovic and Sivaganesan (2002) and Medvedovic and Guo (2004). In section 2, a sequential estimation algorithm analogous to the one in Maceachern et al. (1999) is developed. In section 3, individuals are represented as nodes of a weighted undirected graph. Nodes can then be classified throug a spectral clustering technique as in von Luxburg (2007). The approach proposed in sections 2 and 3 has a double benefit. On one hand, the sequential estimation algorithm guarantees a fast estimation of pairwise similarities. On the other hand, the construction of the random walk on the graph mentioned above, allows us to choose the optimal partition by a minimum description length algorithm, so avoiding the subjective choice of a cut of the dendrogram usually associated to hierarchical clustering algorithms. Furthermore, as a byproduct, the entropy of any partition of the data can be computed and it is closely linked to the fitted model. This allows for a model based comparison of any pair of partitions.

2 Sampling importance resampling

Under the assumptions we introduced above, following the arguments of Maceachern et al. (1999), we can write the conditional posterior distribution of s_i given $x_1, \ldots, x_i$, as

$$p(s_i = j | \mathbf{s}_{<i}, \theta^*, \mathbf{x}_{<i}^{(j)}, x_i) = \begin{cases} \frac{n_j}{\alpha+i-1} p(x_i | \theta_j^*, \mathbf{s}_{<i}, \mathbf{x}_{<i}^{(j)}) & j \in \{k\} \\ \frac{\alpha}{\alpha+i-1} p(x_i | \theta_{k+1}^*) & j = k+1, \end{cases}$$

where $\mathbf{x}_{<i}^{(j)} = \{x_{i'} : i' < i, s_{i'} = j\}$, $j = 1, \ldots, k$, and $\mathbf{x}_{<i}^{(k+1)} = \emptyset$, since $\forall i' < i$, $s_{i'} \in \{k\}$.

We can marginalise the conditional posterior of s_i with respect to θ^*, obtaining

$$p(s_i = j | \mathbf{s}_{<i}, \mathbf{x}_{<i}^{(j)}, x_i) = \begin{cases} \frac{n_j}{\alpha+i-1} p(x_i | s_i = j, \mathbf{s}_{<i}, \mathbf{x}_{<i}^{(j)}) & j \in \{k\} \\ \frac{\alpha}{\alpha+i-1} p(x_i | s_i = k+1, \mathbf{s}_{<i}, \mathbf{x}_{<i}) & j = k+1, \end{cases}$$

where

$$p(x_i | s_i = j, \mathbf{s}_{<i}, \mathbf{x}_{<i}) = \int_\Theta p(x_i | \theta, s_i = j, \mathbf{s}_{<i}, \mathbf{x}_{<i}^{(j)}) p(\theta | s_i = j, \mathbf{s}_{<i}, \mathbf{x}_{<i}^{(j)}) d\theta \tag{5}$$

and

$$p(x_i | s_i = k+1, \mathbf{s}_{<i}, \mathbf{x}_{<i}) = \int_\Theta p(x_i | \theta) dG_0(\theta). \tag{6}$$

Notice that when G_0 is a conjugate prior for (1), the computation of (5) and (6) is often straightforward.

The following importance sampler has been introduced in Maceachern et al. (1999).

SIR algorithm. For $i = 1, \ldots, n$, repeat steps (A) and (B)

(A) Compute

$$g(x_i | \mathbf{s}_{<i}, \mathbf{x}_{<i}) \propto \sum_{j=1}^{k+1} \frac{n_j}{\alpha + i - 1} p(x_i | s_i = j, \mathbf{s}_{<i}, \mathbf{x}_{<i}^{(j)}),$$

with $n_{k+1} = \alpha$.

(B) Generate s_i from the multinomial distribution with

$$p(s_i = j|\mathbf{s}_{<i}, \mathbf{x}_{<i}^{(j)}, x_i) \propto \frac{n_j}{\alpha + i - 1} p(x_i|s_i = j, \mathbf{s}_{<i}, \mathbf{x}_{<i}^{(j)}).$$

Taking R independent replicas of this algorithm we obtain $s_i^{(r)}$, $i = 1, \ldots, n$, $r = 1, \ldots, R$, and $\theta_j^* \sim p(\theta|\mathbf{x}^{(j)})$, with $\mathbf{x}^{(j)} = \{x_i : i \in \{n\}, s_i = j\}$, and compute the importance weights

$$w_r \propto \prod_{i=1}^{n} g(x_i|\mathbf{s}_{<i}, \mathbf{x}_{<i})$$

such that $\sum_{r=1}^{R} w_r = 1$. Should the variance of the importance weights be too small, the efficiency of the sampler could be improved by resampling as follows (Cappé et al., 2005):

1. compute $N_{\text{eff}} = (\sum_{r=1}^{R} w_r^2)^{(-1)}$;
2. if $N_{\text{eff}<\frac{R}{2}}$, draw R particles from the current particle set with probabilities equal to their weights, replace the old particle with the new ones and assign them constant weights $w_r = \frac{1}{R}$.

3 Pairwise similarities and spectral clustering

3.1 Pairwise similarities

Intuitively, we can state that two individuals, i and j, are similar if x_i and x_j are generated by the same mixture component, i.e. if $s_i = s_j$. Label switching prevents us from identifying mixture components, but not from assessing similarities among individuals. In fact, the algorithm introduced in the previous section may help us in estimating pairwise similarities between individuals. The posterior probability that x_i and x_j are generated by the same component, i.e. the posterior probability of the event $\{s_i = s_j\}$, can be estimated as

$$\hat{p}_{ij} = \sum_{r=1}^{R} w_r I\left(s_i^{(r)}, s_j^{(r)}\right),$$

where $I(x, y) = 1$ if $x = y$ and $I(x, y) = 0$ otherwise. We can then define a similarity matrix S with ij-th element $s_{ij} = \hat{p}_{ij}$.

3.2 Graph representation

The matrix S can be used to build the weighted undirected graph $G = (V, E)$, where each node in the set V represents an individual in the sample, i.e. $V = \{n\}$, and the set E contains all the edges in G. Furthermore, the weight of the generic edge (i, j) is given by $w_{ij} = s_{ij}$ if $i \neq j$, and $w_{ij} = 0$ otherwise. We want to find a partition of the graph such that the edges between different groups have low weight and edges within clusters have high weight.

Let us define the degree of node i as $d_i = \sum_{j=1}^{n} w_{ij}$, $i = 1, \dots, n$ and the degree matrix as $D = \mathrm{diag}(d_1, \dots, d_n)$. Furtehrmore, let $A \subseteq V$ and $\bar{A} = V \setminus A$. Then A is identified by the membership vector

$$\mathbb{1}_A = (f_1, \dots, f_n)' \in \mathbb{R}^n : f_i = 1 \Leftrightarrow i \in A,\ f_i = 0 \Leftrightarrow\ i \in \bar{A}.$$

The subset A is connected if any pair of nodes in A can be joined by a path containing only nodes in A; A is a connected component if it is connected and there are no connections between A and $\bar{A}$. The family of subsets $\{A_1, \dots, A_k\}$ form a partition of V if $\cup_{i=1}^{n} A_i = V$ and $A_i \cap A_j = \emptyset$.

The unnormalised laplacian is defined as $L = D - S$ and it has the following properties:

1. For any $f \in \mathbb{R}^n$
$$f'Lf = \frac{1}{n} \sum_{i,j=1}^{n} s_{ij}(f_i - f_j)^2$$

2. L is symmetric and positive semi-definite

3. The smallest eigenvalue of L is 0 and the corresponding eigenvector is $\mathbb{1}$

4. L has non-negative real valued eigenvalues $0 = \lambda_1 \leq \lambda_2 \leq \dots \leq \lambda_n$

The multiplicity k of the eigenvalue 0 of L equals the number of connected components $A_1, \dots, A_k$ in G. The eigenspace of the eigenvalue 0 is spanned by the indicator vectors $\mathbb{1}_{A_1}, \dots, \mathbb{1}_{A_k}$. This result is important, since, as we shall see later, the values taken by the eigenvalues of a suitably normalised laplacian allow us to set the number of components in the optimal clustering.

Usually, for clustering purposes, the following normalised graph Laplacians are considered:

$$
\begin{aligned}
L_{Sym} &= D^{-1/2} L D^{-1/2} = I - D^{-1/2} S D^{-1/2} \\
L_{RW} &= D^{-1} L = I - D^{-1} S.
\end{aligned}
$$

Here, we shall focus our attention on L_{RW}, the random walk normalised laplacian. and consider the clustering algorithm introduced in Shi and Malik (2000). Let $S \in \mathbb{R}^{n \times n}$, $k =$ number of clusters

1. Build a similarity graph with adjacency matrix S
2. Compute the unnormalised Laplacian, $L = D - S$
3. Compute the first k eigenvectors of $L_{RW} = I - D^{-1}S$, $u_1, \dots, u_k$
4. Let $U = [u_1, \dots, u_k] \in \mathbb{R}^{n \times k}$
5. For $i = 1, \dots, n$ let y_i denote the i-th row of U
6. Cluster the y_i's in $\mathbb{R}^k$ with the k-means algorithm into k clusters, $C_1, \dots, C_k$

Output: clusters $A_1, \dots, A_k$ with $A_j = i : y_i \in C_j$

How many clusters? (von Luxburg, Statistics and Computing 2007)

Choose k such that $\lambda_1, \dots, \lambda_k$ are very small, but λ_{k+1} is relatively large

The eigengap euristic

- Compute $\Delta_i = \lambda_i - \lambda_{i+1}$, $i = 1, n - 1$
- Set $k : \Delta_k = \max_i \Delta_i$.

4 Examples

In this section we apply the spectral clustering methods based on the Dirichlet process prior to some datasets for which the true clustering is known. We compare the the clusterings provided by our method with the ones produced by standard spectral clustering techniques, where pairwise similarities are defined as the inverse Euclidean distances between observations, and with the MAP classifications produced by finite Gaussian mixtures estimated via maximum likelihood (Fraley and Raftery, 2000). Comparisons are made by computing the Rand and the adjusted Rand indeces between each partition and the known true clustering.

4.1 Example 1

Figure 1 shows a simulated data set composed by observations generated by a bivariate Gaussian distribution (red), a banana shaped cluster (black), and a uniform noise (green). The data have been standardised and the following model has been fitted:

$$\begin{aligned} X_i|\mu_i, \mathbf{\Psi}_i &\overset{ind}{\sim} N(\mu_i, \mathbf{\Psi}_i) \\ (\mu_i, \mathbf{\Psi}_i)|G &\overset{iid}{\sim} G \\ G &\sim DP(\alpha, NW(\mu_0, \kappa_0, \nu_0, \mathbf{S}_0)) \end{aligned}$$

with $\alpha = 0.5, \boldsymbol{\mu}_0 = \mathbf{0}, \kappa_0 = 0.05, \nu_0 = 4, \mathbf{S}_0 = \mathbf{I}_2$; $R = 5000$, $n = 410$.

The eigenvalues of L_{RW} suggest a classification in four clusters for the spectral clustering based on the DP prior, and in two clusters for the standard spectral clustering technique (Figure 2). The finite mixture model identifies three clusters, as shown in Figure 3. Table 1 shows that the DP based spectral clustering outperforms the two competitors in terms of both Rand and Adjusted Rand index.

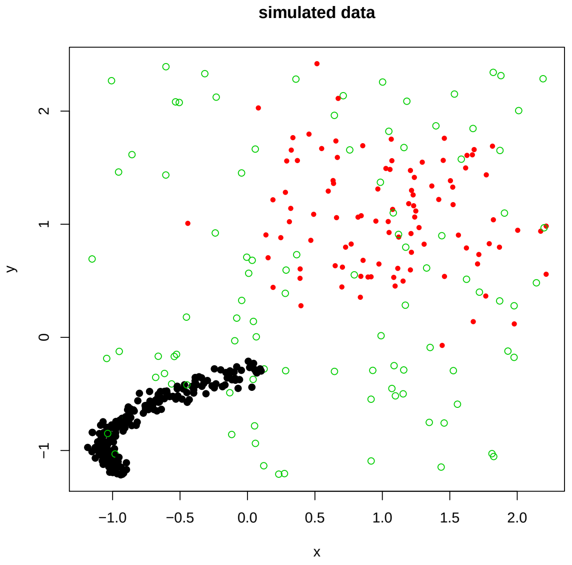

Figure 1: Example 1. A simulated data set composed by observations generated by a bivariate Gaussian distribution (red), a banana shaped cluster (black), and a uniform noise (green).

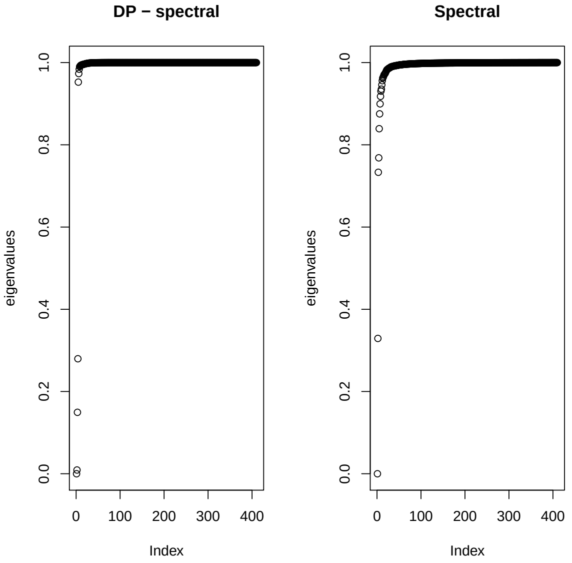

Figure 2: Example 1. Eigenvalues of L_{RW} in the banana shaped cluster application.

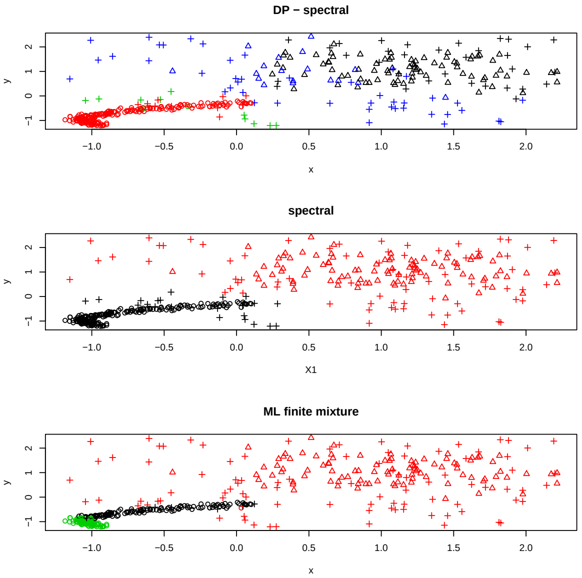

Figure 3: Example 1. Clusterings produced by the alternative methods

	ARI	RI
DP - spectral	0.71	0.86
Spectral	0.65	0.82
ML mixture	0.45	0.74

Table 1: Example 1. Comparison of the alternative classifications with the true clustering

4.2 Example 2.

In this example we consider a dataset presented in Jain and Law (2005). The data are shown in Figure 4. After standardisation, the following model has been fitted:

$$\begin{aligned} X_i|\mu_i, \mathbf{\Psi}_i &\overset{ind}{\sim} N(\mu_i, \mathbf{\Psi}_i) \\ (\mu_i, \mathbf{\Psi}_i)|G &\overset{iid}{\sim} G \\ G &\sim DP(\alpha, NW(\mu_0, \kappa_0, \nu_0, \mathbf{S}_0)) \end{aligned}$$

with $\alpha = 0.3, \boldsymbol{\mu}_0 = \mathbf{0}, \kappa_0 = 0.1, \nu_0 = 4, \mathbf{S}_0 = 5\mathbf{I}_2$; $R = 5000$, $n = 373$.

Figure 4 shows also the clusterings provided by the alternative methods. Notice that the eigenvalues of L_{RW} suggest a unique cluster, as shown in Figure 5.

Table 2 shows that DP based spectral clustering outperforms the classification pro-

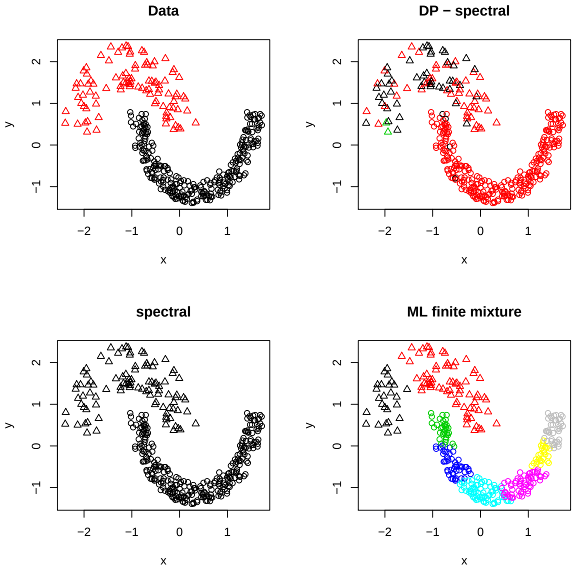

Figure 4: Example 2. The data and the clusterings produced by the alternative methods.

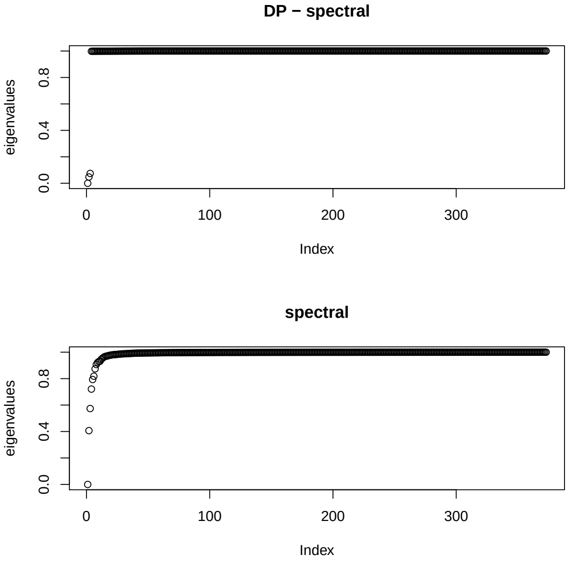

Figure 5: Example 2. Eignevalues of L_{RW} for DP based and standard spectral clustering.

	ARI	RI
DP	0.53	0.77
ML mixture	0.06	0.46

Table 2: Comparison with the classification produced by the finite mixture model.

duced by the finite Gaussian mixture model.

4.3 Example 3.

The dataset we consider in this example consists 178 measurements of 13 variables (Alcohol, Malic acid, Ash, Alcalinity, Magnesium, Phenols, Flavanoids, Nonflavanoids, Proanthocyanins, Color intensity, Hue, OD280.OD315Dilution, Proline) on three types of wine(Barolo, Grignolino and Barbera) (Forina et al., 2008). The data are shown in Figure 6. Five clustering variables have been selected by applying the method suggested in Raftery and Dean (2006). After standardisation, the following model has been fitted:

$$\begin{aligned} X_i|\mu_i, \mathbf{\Psi}_i &\overset{ind}{\sim} N(\mu_i, \mathbf{\Psi}_i) \\ (\mu_i, \mathbf{\Psi}_i)|G &\overset{iid}{\sim} G \\ G &\sim DP(\alpha, NW(\mu_0, \kappa_0, \nu_0, \mathbf{S}_0)) \end{aligned}$$

with $\alpha = 0.1, \boldsymbol{\mu}_0 = \mathbf{0}, \kappa_0 = 0.01, \nu_0 = 100, \mathbf{S}_0 = 50\mathbf{I}_5$; $R = 5000$, $n = 178$.

Figure 7 shows that DP based spectral clustering identifies three groups, whereas standard spectral clustering does not capture any cluster in the dataset. Figures 8 and 9 show the classifications produced by the DB based clustering and the clustering based on the finite mixture model. Table 3 shows that these two partitions are equivalent in terms of both Rand and Adjusted Rand index.

	ARI	RI
DP - spectral	0.78	0.90
Spectral	0.00	0.34
ML mixture	0.78	0.90

Table 3: Example 3. Comparison of the alternative classifications with the true clustering

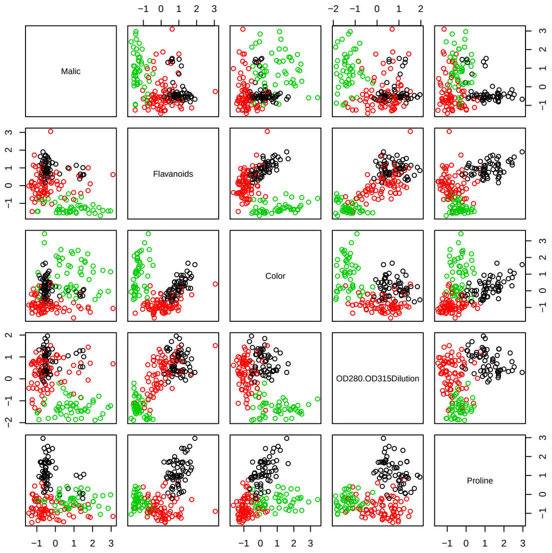

Figure 6: Example 3. The data: colours identify the three different wine types.

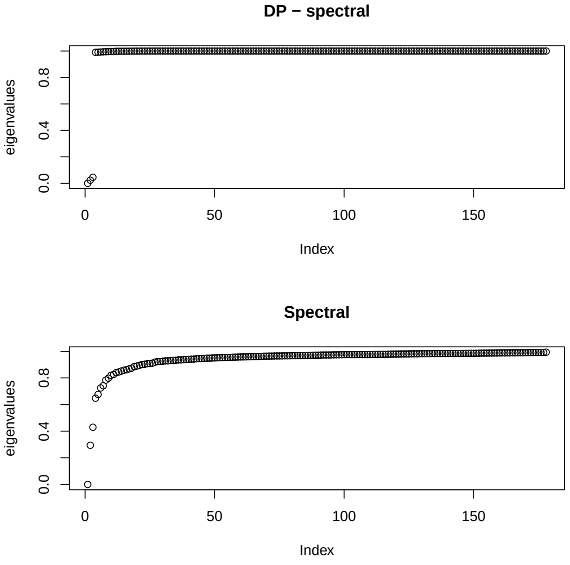

Figure 7: Example 3. Eigenvalues of L_{rw} for DP based and standard spectral clustering

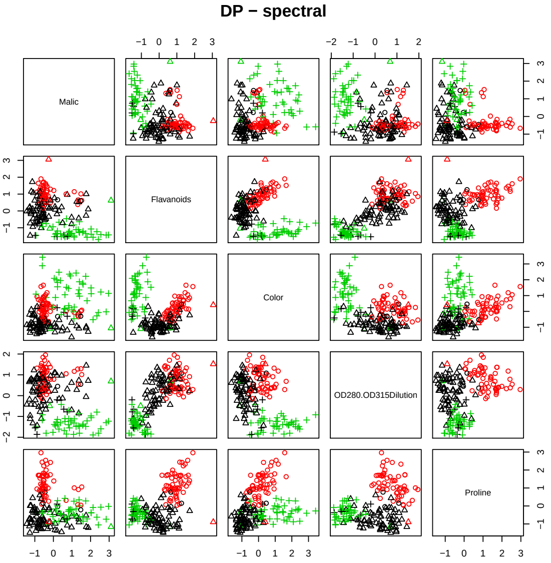

Figure 8: Example 3. DP based spectral clustering.

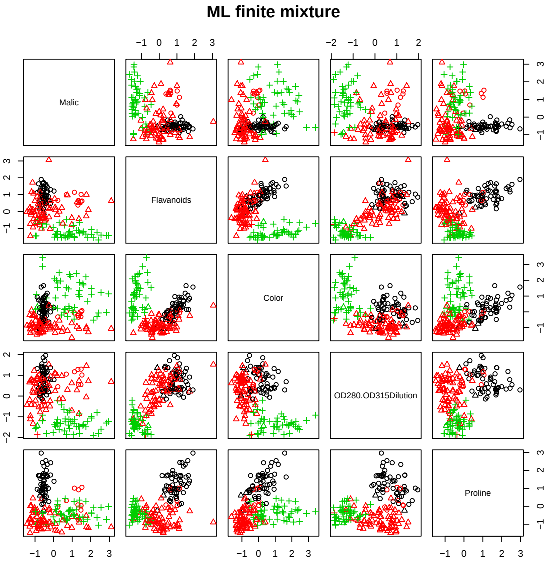

Figure 9: Example 3. Clustering produced by the finite mixture model

5 Discussion

The flexibility of Bayesian nonparametric models improves robustness of classification with respect to finite mixture models. Sampling importance resampling algorithms allow for efficient computations, particularly when the base measure is conjugate to model likelihood. The DP based spectral clustering does not require any restrictions on the parameters or post processing of the posterior simulations. Furthermore, in the examples we have considered, it always outperform the performance of standard spectral clustering. It has also shown to be competitive with the mixture model based classification method.

One limitation of the DP based spectral clustering is the selection of the clustering variables when a high number of attribute measurements is collected. Research on this topic is under way.

References

Antoniak, C. E. (1974, 11). Mixtures of dirichlet processes with applications to bayesian nonparametric problems. *Ann. Statist. 2*(6), 1152–1174.

Cappé, O., E. Moulines, and T. Ryden (2005). *Inference in Hidden Markov Models (Springer Series in Statistics).* Berlin, Heidelberg: Springer-Verlag.

Forina, M., S. Lanteri, C. Armanino, C. Casolino, M. Casale, and P. Oliveri (2008). An extendible pachage of programs for esplorative data analysis, classification and regression analysis. Technical report, Dip. Chimica e Tecnologie Farmaceutiche ed Alimentari, Università di Genova.

Fraley, C. and A. E. Raftery (2000). Model-based clustering, discriminant analysis, and density estimation. *JOURNAL OF THE AMERICAN STATISTICAL ASSOCIATION 97*, 611–631.

Frühwirth-Schnatter, S. (2006). *Finite mixture and Markov switching models.* Springer.

Jain, A. K. and M. H. C. Law (2005). Data clustering: A user's dilemma. In *PReMI*, Volume 3776 of *Lecture Notes in Computer Science*, pp. 1–10. Springer.

Maceachern, S. N., M. Clyde, and J. S. Liu (1999). Sequential importance sampling for nonparametric Bayes models: The next generation. *Can J Statistics 27*(2), 251–267.

McLachlan, G. J. and D. Peel (2000). *Finite mixture models.* New York: Wiley Series in Probability and Statistics.

Medvedovic, M. and J. Guo (2004). Bayesian model-averaging in unsupervised learning from microarray data. In *BIOKDD*, pp. 40–47.

Medvedovic, M. and S. Sivaganesan (2002). Bayesian infinite mixture model based clustering of gene expression profiles. *Bioinformatics 18*(9), 1194–1206.

Raftery, A. E. and N. Dean (2006). Variable selection for model-based clustering. *Journal of the American Statistical Association 101*(473), 168–178.

Shi, J. and J. Malik (2000). Normalized cuts and image segmentation. *IEEE Trans. Pattern Anal. Mach. Intell. 22*(8), 888–905.

Stephens, M. (2000). Dealing with label switching in mixture models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology) 62*(4), 795–809.

von Luxburg, U. (2007). A tutorial on spectral clustering. *Statistics and Computing 17*(4), 395–416.