Ca' Foscari
University
of Venice

**Department
of Economics**

Working Paper

Andrea Pastore and Stefano F.
Tonellato

A merging algorithm for
Gaussian mixture components

ISSN: 1827-3580
No. 04/WP/2013

A merging algorithm for Gaussian mixture components

A. Pastore and S. F. Tonellato

Department of Economics, University Ca' Foscari of Venice

March 8, 2013

Abstract

In finite mixture model clustering, each component of the fitted mixture is usually associated with a cluster. In other words, each component of the mixture is interpreted as the probability distribution of the variables of interest conditionally on the membership to a given cluster. The Gaussian mixture model (GMM) is very popular in this context for its simplicity and flexibility. It may happen, however, that the components of the fitted model are not well separated. In such a circumstance, the number of clusters is often overestimated and a better clustering could be obtained by joining some subsets of the partition based on the fitted GMM. Some methods for the aggregation of mixture components have been recently proposed in the literature. In this work, we propose a hierarchical aggregation algorithm based on a generalisation of the definition of silhouette-width taking into account the Mahalanobis distances induced by the precision matrices of the components of the fitted GMM. The algorithm chooses the number of groups corresponding to the hierarchy level giving rise to the highest average-silhouette-width. Some simulation experiments and real data applications indicate that its performance is at least as good as the one of other existing methods.

Keywords Gaussian mixtures; Model based clustering; Silhouette width; dissimilarity.

JEL Codes C39, C46

Address for correspondence:

Andrea Pastore

Department of Economics
Ca' Foscari University of Venice
Cannaregio 873, Fondamenta S. Giobbe
30121 Venezia - Italy
Phone: (+39) 041 2347419
Fax: (+39) 041 2347444
e-mail: pastore@unive.it

This Working Paper is published under the auspices of the Department of Economics of the Ca' Foscari University of Venice. Opinions expressed herein are those of the authors and not those of the Department. The Working Paper series is designed to divulge preliminary or incomplete work, circulated to favour discussion and comments. Citation of this paper should consider its provisional character.

1 Introduction

Gaussian finite mixture models (GMM) (McLachlan and Peel, 2000), are widely used for clustering purposes. Commonly, maximum likelihood estimates of the parameters are achieved via the Expectation-Maximization (EM) algorithm, whereas the number of components of the mixture is determined by minimising the Bayesian Information Criterion (BIC) (Fraley and Raftery, 2002; Fraley et al., 2012). A quite natural choice is to associate a cluster to each component of the fitted GMM. This methodology, however, often leads to solutions with an unreasonably high number of poorly separated clusters when mixture components overlap. In order to solve this problem, some authors proposed some alternatives to EM-BIC for parameter estimation and for the determination of the number of components. Although such methods might seem suitable for clustering purposes, they lead to parameter estimates that are not based on the maximum likelihood method (see, for instance, Biernacki et al. (2000)).

Baudry et al. (2010) argued that the number of components of the GMM should not necessarily be equal to the number of groups. Assuming that such number is less than or equal to the number of components, a cluster could be associated with a combination of components. Therefore, they propose to retain the mixture model fitted via EM-BIC and to find the number of groups with a hierarchical merging procedure which minimises the entropy of the clustering solution. A similar approach is followed also by Hennig (2010), where some aggregation criteria are proposed, based either on *modality* (i.e. merging components that produce a unimodal distribution) or on *misclassification probabilities* (i.e. merging components in order to minimize the misclassification probabilities).

In this paper we propose an aggregation algorithm for GMM components which explicitly takes into account for within-cluster heterogeneity and between-cluster separation of the involved partitions. The algorithm is based on a suitable extension of the Average Silhouette Width (ASW) measure (Rousseeuw, 1987; Kaufman and Rousseeuw, 1990) to partitions based on a fitted GMM. In section 2 we shall recall the definitions of GMM, silhouette width and average silhouette width, emphasising the difficulties in defining silhouette widths consistently with the structure of GMM's. In section 3 we shall introduce an asymmetric dissimilarity measure which provides the basis for the definition of silhouette widths tailored on a fitted GMM. In section 4 we shall define a silhouette based hierarchical procedure for cluster aggregation and present some applications to simulated and real data. Some concluding

remarks will be given in section 5.

2 Gaussian mixture models and average silhouette width

2.1 Gaussian Mixture Models

Let $U = \{u_1, \dots, u_n\}$ be a set of n objects we want to classify. Assume that on each element of U a p -variate observations x_i is taken, and that x_i is a determination of the random vector X_i . Furthermore, we assume that X_i , $i = 1, \dots, n$, are iid and can be represented as Gaussian mixtures:

$$X_i \sim \sum_{t=1}^g \pi_t N[x|\mu_t, \Sigma_t], \quad (1)$$

where $0 < \pi_t < 1$, $t = 1, \dots, g$, $\sum_{t=1}^g \pi_t = 1$ and $N[x|\mu, \Sigma]$ is the p -variate Gaussian pdf with expected value μ , and covariance matrix Σ . Restrictions on the covariance matrices can be introduced in order to reduce the number of parameters (Banfield and Raftery, 1993).

We shall denote the estimates of π_t , μ_t and Σ_t , $t = 1, \dots, g$ by $\hat{\pi}_t$, $\hat{\mu}_t$ and $\hat{\Sigma}_t$ respectively. The number of components, g , and the parametrisation of Σ_t , $t = 1, \dots, g$, are determined by minimising the BIC. The estimated posterior probability that the i -th unit belongs to component t will be denoted by $\hat{\tau}_{i,t}$. Furthermore, let $z_i \in \{1, \dots, g\}$ represent the label of the group maximising the i -th unit estimated posterior membership probability distribution, i.e.

$$z_i \in \{1, \dots, g\} : \hat{\tau}_{i,z_i} \geq \hat{\tau}_{i,t}, \quad t \in \{1, \dots, g\} \setminus z_i. \quad (2)$$

Then, the set U will be partitioned in g subsets $C_1, \dots, C_g$, such that $u_i \in C_t \Leftrightarrow z_i = t$. We shall denote such partition as:

$$\mathcal{P}_g = \{C_1, \dots, C_g\}.$$

In such a setting, each component of the mixture is interpreted as the probability distribution characterising a specific cluster in the population. Loosely speaking, each cluster can be identified with its corresponding component in the mixture model.

Such interpretation, however, is sometimes inappropriate. Two typical examples are unimodal mixtures and mixtures with overlapping components. One solution to this problem is given by the aggregation of badly separated clusters in the original classification. This means that the fitted mixture model remains unchanged, but cluster specific probability distributions now can be characterised either by single components or by mixtures of some components of the fitted model. This approach is followed by Baudry et al. (2010) and Hennig (2010). In the following, then, we shall distinguish between *components* and *clusters* (or *groups*).

2.2 Average Silhouette Width

The Average Silhouette Width (ASW, Rousseeuw (1987); Kaufman and Rousseeuw (1990)) is an index which evaluates the quality of a partition, accounting for both heterogeneity within each cluster and separation among clusters. It can be used to validate a clustering solution and to determine the number of groups when hierarchical methods are applied.

Given a set of objects U , ASW depends on the given partition and a distance, or, more generally, on a dissimilarity measure d , $d : U^2 \mapsto \mathbb{R}^+$. Let $D = \{d_{ij}, i, j = 1, \dots, n\}$ be a pairwise dissimilarity matrix between the objects of U and $\{C_1, \dots, C_g\}$ a partition of U . Furthermore, let us denote by n_t the cardinality of C_t and by $z_i \in \{1, \dots, g\}$ the label of the group the i -th unit has been allocated to, i.e. $z_i = t \Leftrightarrow u_i \in C_t$. The silhouette width of u_i is defined as:

$$s_i = \frac{b_i - a_i}{\max\{a_i, b_i\}}. \quad (3)$$

The term a_i in (3) is defined as

$$a_i = \sum_{j: z_j = z_i} \frac{d_{ij}}{n_{z_i} - 1}.$$

It is a measure of the dissimilarity between u_i and C_{z_i} . If the i -th object is a singleton, i.e. if $n_{z_i} = 1$, $a_i = 0$. The term b_i in (3) is defined as

$$b_i = \min_{t \neq z_i} \sum_{j: z_j = t} \frac{d_{ij}}{n_t}.$$

It is a measure of the dissimilarity between u_i and its neighbouring cluster, C_{q_i} , where

$$q_i = \arg \min_{t \in \{1, \dots, g\} \setminus i} \sum_{j: z_j = t} \frac{d_{ij}}{n_t}.$$

The ASW of a given partition of U is simply defined as the average of the individual silhouette widths:

$$ASW = \frac{1}{n} \sum_{i=1}^n s_i.$$

Similarly, cluster specific ASW's can be computed.

The quantity

$$\bar{a}_t = \frac{1}{n_t} \sum_{i: z_i = t} a_i = \frac{1}{n_t(n_t - 1)} \sum_{i \neq j: z_i = z_j = t} d_{ij},$$

known as *mean-of-distances*, can be interpreted as a measure of the heterogeneity within C_t . In the following, we shall use the same term, even when d is a dissimilarity but not a distance. In that case, of course, the mean-of-distances statistic should be read as an average dissimilarity between the objects in the same cluster.

As we have already pointed out, the silhouette width depends on both the dissimilarity matrix D and the partition. An essential coherence condition, is that the dissimilarity measure used, implicitly or explicitly, to partition U , be the same one leading to the matrix D used in the computation of silhouette widths. In the next subsection we shall pursue this objective. Clearly, if $\mathcal{P}_g$ represents the clustering solution based on a GMM, this condition is not met. In the next section, we shall pursue the objective of defining a dissimilarity measure, as well as its corresponding silhouette widths, well suited to a fitted GMM.

3 Silhouette widths for Gaussian mixtures

3.1 A dissimilarity measure coherent with the fitted GMM

Suppose that a partition of the set U based on a GMM has been obtained. Does there exist a dissimilarity measure d coherent with the fitted GMM? In order to answer this question, we first focus on pairwise dissimilarities. We assume, for the

moment, that distinct clusters are associated to distinct mixture components, and consider two generic individuals, u_i and u_j , in U . If such units have been assigned to the same group, say C_t , then a coherent dissimilarity measure is the squared Mahalanobis distance (SDM) defined by the t -th component in the mixture model:

$$d_{i,j,t} = (x_i - x_j)' \Sigma_t^{-1} (x_i - x_j), \quad (4)$$

where Σ_t is defined as in (1). If, instead, $z_i \neq z_j = t$ the SMD would make sense only if $\Sigma_{z_i} = \Sigma_{z_j}$. When such condition does not hold, we define an asymmetric pairwise dissimilarity measure:

$$\tilde{d}_{i,j} = d_{i,j,z_i}. \quad (5)$$

The asymmetry is due to the fact that $\tilde{d}_{j,i} = d_{j,i,z_j}$ generally differs from $\tilde{d}_{i,j}$.

3.2 Silhouette widths for Gaussian mixtures

Let us define the average dissimilarity between u_i and C_t as:

$$\bar{d}_{i,t} = \begin{cases} 0 & \text{if } t = z_i \text{ and } n_t = 1 \\ \frac{1}{n_t - 1} \sum_{j: z_j = t} \tilde{d}_{i,j} & \text{if } t = z_i \text{ and } n_t > 1 \\ \frac{1}{n_t} \sum_{j: z_j = t} \tilde{d}_{i,j} & \text{if } t \neq z_i \end{cases}$$

Clearly, the average dissimilarity $\bar{d}_{i,t}$ is determined by the SMD induced by Σ_{z_i} , with Σ_{z_i} and z_i as defined in (1) and (2). Therefore, if $t = z_i$, then $\bar{d}_{i,z_i}$ measures how heterogeneous u_i is with respect to its cluster, C_{z_i} . When, instead, $z_i \neq t$, $\bar{d}_{i,t}$ measures the separation of u_i from C_t , using the square of the metric induced by the variance matrix of the z_i -th component. We can then define:

(a) a measure of the dissimilarity between u_i and C_{z_i} ,

$$a'_i = \begin{cases} 0 & \text{if } n_{z_i} = 1 \\ \bar{d}_{i,z_i} & \text{otherwise;} \end{cases}$$

(b) a measure of the dissimilarity between u_i and its neighbouring cluster C_{q_i} ,

$$b'_i = \min_{t \in \{1, \dots, g\} \setminus i} \bar{d}_{i,t}$$

- (c) a Gaussian mixture model silhouette-width (GMM-SW) combining a'_i and b'_i as in (3), i.e.

$$s'_i = \frac{b'_i - a'_i}{\max(a'_i, b'_i)}. \quad (6)$$

Remark 1. Notice that in this case the mean of distances is

$$\bar{a}'_t = \frac{1}{n_t} \sum_{i: z_i=t} a'_i = \frac{1}{n_t(n_t - 1)} \sum_{i \neq j: z_i=z_j=t} d_{i,j,t}.$$

Remark 2. The silhouette width defined above depends on the dissimilarity measure $\tilde{d}_{i,j}$ introduced in (5), which itself depends on unknown model parameters. In the following applications, they will be computed by plugging parameter estimates in (4).

Example 1. We present two simulated datasets from bivariate 4-component GMM's with different variance matrices. Figures 1 and 2 show the scatterplots of these data and the charts of both the GMM-SW and the silhouette width based on the Euclidean distance (SW) for such datasets. In both simulations GMM-SW take sensibly higher values than the SW's. This feature is particularly apparent in the second simulation. In fact, in the black group in Figure 2 SW's take even negative values. This is due mainly to the fact that the SMD induced by the variance matrix of that component is sensibly different from the Euclidean distance. In such a circumstance, then, being GMM-SW's more consistent with the fitted model, they provide a better evaluation a model based classification.

3.3 Silhouette widths for aggregate components

When some Gaussian mixture components are not well separated, the method illustrated in subsection 2.1 might produce quite unsatisfactory partitions. In order to improve the classification in such circumstances, we are aiming at a generalization of the GMM-SW allowing us to produce a better partition by aggregating some clusters. In the following, then, we will distinguish between mixture components and groups, the latter coinciding with clusters in $\mathcal{P}_g$ or with the aggregation of some of them.

Suppose that a partition, $\mathcal{P}_k$, in k groups has been obtained by joining some clusters in $\mathcal{P}_g$. Let us denote with $A_{h,k}$ the h -th cluster in $\mathcal{P}_k$, $h = 1, \dots, k$. Define

$$J_{h,k,t} = \begin{cases} 1 & \text{if } C_t \subseteq A_{h,k}, \\ 0 & \text{otherwise,} \end{cases}$$

with $h = 1, \dots, k$ and $t = 1, \dots, g$. Each cluster $A_{h,k}$ can then be defined as

$$A_{h,k} = \bigcup_{t: J_{h,k,t}=1} C_t,$$

where $\mathcal{P}_g = \{C_1, \dots, C_g\}$ is the partition obtained from the fitted model by applying the rule defined in (2).

Remark 3. In order to be coherent with the notation we have just introduced, we can write $\mathcal{P}_g = \{A_{1,g}, \dots, A_{g,g}\}$, with $A_{t,g} = C_t$.

Let

$$Z_i \in \{1, \dots, k\} : A_{Z_i,k} \supseteq C_{z_i},$$

i.e. Z_i is the label of the cluster (in $\mathcal{P}_k$) which includes C_{z_i} , $i = 1, \dots, n$ (notice that $A_{Z_i,k}$ might coincide with C_{z_i} for some i). Furthermore, let $N_{i,k}$ be the cardinality of $A_{Z_i,k}$, that is:

$$N_{i,k} = \sum_{t: C_t \subseteq A_{Z_i,k}} n_t$$

We can then measure the dissimilarity between u_i and $A_{Z_i,k}$ with

$$a_i^* = \sum_{j: C_{z_j} \subseteq A_{Z_i,k}} \frac{\tilde{d}_{i,j}}{N_{i,k} - 1} = \frac{1}{N_{i,k} - 1} \left[(n_{z_i} - 1) \bar{d}_{i,z_i} + \sum_{\substack{t: C_t \subseteq A_{Z_i,k} \\ t \neq z_i}} n_t \bar{d}_{i,t} \right],$$

and its separation from the clusters $A_{h,k}$, $h \neq Z_i$, with

$$b_i^* = \min_{t: C_t \not\subseteq A_{Z_i,k}} \bar{d}_{i,t} \quad (7)$$

We can then generalise the definition of GMM-SW to this context by defining

$$s_i^* = \frac{b_i^* - a_i^*}{\max(a_i^*, b_i^*)}. \quad (8)$$

Notice that, when we consider the partition $\mathcal{P}_g$, equations (8) and (3) coincide.

Remark 4. In this case, the mean-of-distances statistic for the cluster $A_{Z_i,k}$ is given by

$$\begin{aligned}
\bar{a}_{Z_i}^* &= \frac{1}{\sum_{t: C_t \subseteq A_{Z_i,k}} n_t} \sum_{i: C_{z_i} \subseteq A_{Z_i,k}} a_i \\
&= \frac{1}{N_{i,k}} \sum_{i: C_{z_i} \subseteq A_{Z_i,k}} a_i \\
&= \frac{1}{N_{i,k}} \sum_{i: C_{z_i} \subseteq A_{Z_i,k}} \sum_{j: C_{z_j} \subseteq A_{Z_i,k}} \frac{d_{i,j,z_i}}{N_{i,k} - 1} \\
&= \frac{1}{N_{i,k}} \sum_{i: C_{z_i} \subseteq A_{Z_i,k}} \sum_{j: C_{z_j} \subseteq A_{Z_i,k}} \left(\frac{d_{i,j,z_i} + d_{j,i,z_j}}{2} \right) \\
&= \frac{1}{N_{i,k}} \sum_{i: C_{z_i} \subseteq A_{Z_i,k}} \sum_{j: C_{z_j} \subseteq A_{Z_i,k}} \left(\frac{\tilde{d}_{i,j} + \tilde{d}_{j,i}}{2} \right)
\end{aligned}$$

It is worth to note that $\bar{a}_{Z_i}^*$ is the average dissimilarity within the Z_i -th cluster in $\mathcal{P}_k$ in terms of the symmetric dissimilarity measure $d_{i,j}^* = \frac{\tilde{d}_{i,j} + \tilde{d}_{j,i}}{2}$.

The average silhouette width of $\mathcal{P}_k$ is defined as

$$\text{GMM-ASW}(\mathcal{P}_k) = \frac{\sum_{i=1}^n s_i^*}{n}. \quad (9)$$

Obviously, we can compute, in the same way, cluster specific average silhouette widths.

3.4 Examples

Example 1. We consider the dataset in sec. 4.1 of Baudry et al. (2010). Such dataset consists of simulated data from a six-component bivariate gaussian mixture and it is included in the library **Mclust**, version 4.0 (Fraley et al., 2012), of GNU-R (R Development Core Team, 2011).

The **mclust** function estimates a six-component model with unconstrained variance-covariance matrices. The scatterplot in Figure 3 shows the simulated data, jointly with the GMM-SW and the SW. Clearly, in this case some of the components of the mixture cannot be interpreted as a cluster, since they are poorly separated.

Figure 4 shows the same data and a new partition obtained by merging C_3 and C_4 in a single cluster. Notice that the new cluster (the green one) has a good average GMM-SW (about 0.84), although lower than the average silhouette of the red (0.95) and the cyan (0.89) single-component clusters. Figures 5 and 6 are similar to Figure 4, with the difference that in the former C_1 and C_6 have been merged, whereas in the latter C_3 , C_4 and C_5 have been joined. In Figure 7, a single cluster is defined by the aggregation of C_2 and C_5 . This cluster is completely unsatisfactory, and this fact is correctly represented by the GMM-ASW (-0.74). Finally, Figure 8 shows the partition where two clusters are defined by the aggregation of C_1 and C_6 , and C_3 and C_4 respectively. The result is a four-cluster solution obtained from a fitted six-component mixture model. Notice that this is the solution (within those proposed in these figures) with the best GMM-ASW (0.87).

4 A hierarchical merging procedure based on the GMM-ASW

Starting from a fitted g component Gaussian mixture model, we wish to obtain a partition consisting of a number of groups lower than g , by merging some of the clusters in $\mathcal{P}_g$. We then need to decide how many clusters, as well as which ones, need to be merged. A common approach is to define a hierarchical procedure such that, for $k = 1, \dots, g-2, g-1$, a $g-k$ cluster solution, $\mathcal{P}_{g-k}$ is obtained by merging two clusters in $\mathcal{P}_{g-k+1}$. At each step of the hierarchy, we have to merge *single-component* clusters, i.e. those clusters modelled by a specific component in the fitted Gaussian mixture, or *sub-mixture* clusters, i.e. the aggregate clusters, modelled by a mixture of components of the fitted model. Clearly, the candidate groups for aggregation are the ones corresponding to poorly separated components. Such poorly separated clusters can be identified by exploiting the information provided by the GMM-SW and the GMM-ASW. In fact, objects with low silhouette necessarily determine a poor separation between clusters. At the end of the procedure, a partition $\mathcal{P}_{k^*}$, with $1 \leq k^* \leq g$ should be chosen accordingly with some optimality criteria.

Let us assume we joined some clusters in $\mathcal{P}_g$, obtaining the partition $\mathcal{P}_k$, recall the separation measure introduced in (7) and define the following family of clusters $\mathcal{P}_g$:

$$\mathcal{F}_{i,k} = \{C_t \in \mathcal{P}_g : C_t \cap A_{Z_i,k} = \emptyset, t = 1, \dots, g\}$$

We define $q_{i,k}$ as the label of the mixture component such that $b_i^* = \bar{d}_{i,q_{i,k}}$. In

other words, $C_{q_{i,k}}$ in $\mathcal{F}_{i,k}$ is the closest subset to the i -th object. Furthermore, let $A_{Q_{i,k},k}$ be the aggregate cluster which includes C_{q_i} . We point out that, whereas z_i is uniquely determined by the fitted model, $Z_{i,k}$, $q_{i,k}$ and $Q_{i,k}$ depend on the aggregation level, i.e. they vary depend on $\mathcal{P}_k$. In the following, we will show that these quantities play a key role in determining the optimal partition, $\mathcal{P}_{k^*}$.

4.1 Badly allocated objects and poorly separated clusters

Assume we defined a threshold $\delta > 0$ such that a generic object $u_i \in U$ with GMM-SW value $s_i^* \leq \delta$ can be considered badly allocated.

For each pair of subsets $A_{j,k}$ and $A_{s,k}$ in $\mathcal{P}_k$, we define a function that simply counts the number of badly allocated objects in the set

$$G_{j,s,k} = \left\{ u_i \in U : A_{Z_{i,k},k} = A_{j,k} \text{ and } A_{Q_{i,k},k} = A_{s,k}, \text{ or } \right. \\ \left. A_{Z_{i,k},k} = A_{s,k} \text{ and } A_{Q_{i,k},k} = A_{j,k}, \ i = 1, \dots, n \right\}.$$

Such function can be defined as:

$$B_\delta(A_{j,k}, A_{s,k}) = \sum_{i=1}^n \mathbb{1}_{[-1,\delta]}(s_i) \cdot \left(\mathbb{1}_{A_{j,k} \cup A_{s,k}}(C_{z_i}) \cdot \mathbb{1}_{A_{j,k} \cup A_{s,k}}(C_{q_{i,k}}) \right).$$

If $B_\delta(A_{j,k}, A_{s,k}) = 0$, the subsets $A_{j,k}$ and $A_{s,k}$ are well separated. Otherwise, a number $B_\delta(A_{j,k}, A_{s,k})$ of badly allocated objects are *between* $A_{j,k}$ and $A_{s,k}$. Therefore, a high value of $B_\delta(A_{j,k}, A_{s,k})$ indicates that $A_{j,k}$ and $A_{s,k}$ are poorly separated.

4.2 A hierarchical merging procedure

The k -th step of the procedure starts from a $(g - k + 1)$ -group partition, $\mathcal{P}_{g-k+1}$, with $k \leq g - 2$ (where g represents, as usual, the number of components of the fitted model) and provides a $(g - k)$ -group partition, $\mathcal{P}_{g-k}$.

The k -th step of the hierarchical merging algorithm can be described as follows:

- k.i. Starting from $\mathcal{P}_{g-k+1}$, set $\delta = \delta_0$, where δ_0 is a threshold fixed by the researcher (in the following examples, we put $\delta_0 = 0.25$, assuming that an average silhouette width lower than this value indicates no structure, as suggested by Kaufman and Rousseeuw (1990), p. 88, table 4);

- k.ii. for each pair $A_{j,g-k+1}, A_{s,g-k+1}$ in $\mathcal{P}_{g-k+1}$, with $j, s = 1, \dots, g - k + 1$, and $j < s$:
- compute $B_\delta(A_{j,g-k+1}, A_{s,g-k+1})$;
 - if $B_\delta(A_{j,g-k+1}, A_{s,g-k+1}) > 0$, join $A_{j,g-k+1}$ and $A_{s,g-k+1}$ and compute the GMM-ASW, as defined in (9), on the resulting $(g - k)$ -group partition;
 - if $B_\delta(A_{j,g-k+1}, A_{s,g-k+1}) = 0$ for all pairs $A_{j,g-k+1}$ and $A_{s,g-k+1}$, set $\delta = \delta + \epsilon$ (we found empirically that with $\epsilon = 0.05$ the algorithm performs satisfactorily) and restart 1.ii;
- k.iii. join the two clusters which maximise GMM-ASW among the $(g - k)$ -group partitions you have previously obtained and define the resulting partition $\mathcal{P}_{g-k}$;
- k.iv.
- if $k < g - 2$ go to $k + 1$
 - else, compute the GMM-ASW on $\mathcal{P}_2$: if it is greater than δ_0 stop; else set $\mathcal{P}_1 = U$ (which is equivalent to state that there is no structure in the data).

At the end of this procedure we need to compute

$$k^* = \arg \max_{2 \leq k \leq g} \text{GMM-ASW}(\mathcal{P}_k).$$

If $\text{GMM-ASW}(\mathcal{P}_{k^*}) > \delta_0$, the optimal partition is $\mathcal{P}_{k^*}$; otherwise we conclude that there is no clear structure in the data and identify a unique cluster, which coincides with U .

4.3 Examples

Monte Carlo experiment We set up a Monte Carlo experiment simulating datasets which mimic the data of the example in sec. 4.1 of Baudry et al. (2010). Each dataset has size $n = 600$ and observations have been simulated according to a 6-component bivariate GMM. The model has been parametrised in a way such that four cluster could be defined: two of them can be identified with two distinct components whereas each of the remaining ones can be identified by merging two overlapping components. More specifically the parametrisation is:

- (component corresponding to the first cluster)

$$\pi_1 = 0.2, \mu_1 = \begin{pmatrix} 0 \\ 0 \end{pmatrix}, \Sigma_1 = \begin{pmatrix} 0.25 & 0.3 \\ 0.3 & 1 \end{pmatrix}$$

- (components corresponding to the second cluster)

$$\pi_2 = 0.2, \mu_2 = \begin{pmatrix} 1 \\ 5 \end{pmatrix}, \Sigma_2 = \begin{pmatrix} 0.4 & 0 \\ 0 & 1 \end{pmatrix}$$

$$\pi_3 = 0.2, \mu_3 = \begin{pmatrix} 1 \\ 5 \end{pmatrix}, \Sigma_3 = \begin{pmatrix} 1 & 0 \\ 0 & 0.4 \end{pmatrix}$$

- (component corresponding to the third cluster)

$$\pi_4 = 0.2, \mu_4 = \begin{pmatrix} 8 \\ 5 \end{pmatrix}, \Sigma_4 = \begin{pmatrix} 0.25 & -0.3 \\ -0.3 & 1 \end{pmatrix}$$

- (components corresponding to the fourth cluster)

$$\pi_5 = 0.1, \mu_5 = \begin{pmatrix} 8 \\ 0 \end{pmatrix}, \Sigma_5 = \begin{pmatrix} 1 & 0 \\ 0 & 0.4 \end{pmatrix}$$

$$\pi_6 = 0.1, \mu_6 = \begin{pmatrix} 8 \\ 0 \end{pmatrix}, \Sigma_6 = \begin{pmatrix} 0.4 & 0 \\ 0 & 1 \end{pmatrix}$$

For each dataset, parameters have been estimated using the EM algorithm, the number of components has been determined via BIC and then the hierarchical merging procedure based on the GMM-ASW has been applied. The results of the Monte Carlo experiment are shown in tables 1 and 2. Table 1 gives, for each simulated data set, the values of the number of components in the fitted GMM and the value of k^* obtained with the hierarchical merging procedure. Table 2 gives the average value of the (non-adjusted) Rand index between the classification obtained with the hierarchical procedure and the *true* one described above. It is worth to note that such average is never lower than 0.842 (and very close to 1 for $k^* = 4$ and $k^* = 5$). Furthermore, the hierarchical procedure never produces a partition in 6 clusters, whereas the EM-BIC method leads to classifications with at least 6 clusters in 69,6% of the simulated datasets.

k	value of g provided by EM-BIC							tot
	4	5	6	7	8	9	10	
3	3	10	0	22	11	7	0	53
4	23	116	172	75	36	18	3	443
5	0	0	0	0	4	0	0	4
tot	26	126	172	97	51	25	3	500

Table 1: Results of a Monte Carlo experiment

	k		
	3	4	5
average Rand Index	0.842	0.994	0.993

Table 2: Results of a Monte Carlo experiment

Olive oil dataset. We consider the dataset presented by Forina et al. (1983), containing the measurements of $p = 8$ chemical characteristics of $n = 572$ specimens of olive oil produced in various areas of Italy. Specimens originate from nine different areas that can be grouped in three *regions*: Centre-North, South and Sardinia. Many authors analysed this dataset to illustrate clustering procedure. We refer, in particular, to the analysis by Azzalini and Torelli (2007), who consider the first five principal components after an additive log-ratio transformation of the original data. For the same data, we present the results of four classifications:

1. the classification obtained with a GMM without constraints, which gives a 5-group partition;
2. the optimal classification obtained with the hierarchical merging algorithm, which consists of two clusters;
3. the classification in 3 groups obtained with the hierarchical merging procedure, starting from the solution obtained in 1.;
4. the classification obtained with a GMM, with a constraint on the number of components, fixing $g = 3$.

The constraints in the number of groups in 3. and 4. above, are motivated by the fact that we take the classification in three regions as the true one. The clustering solutions are described in Table 3. In Table 4 we report, for each classification,

groups	South	Sardinia	Centre-North
GMM (5 groups)			
1	125	0	0
2	198	0	0
3	0	97	0
4	0	1	80
5	0	0	71
GMM + merge (2 groups)			
1	323	97	0
2	0	1	151
GMM + merge (3 groups)			
1	323	0	0
2	0	97	0
3	0	1	151
Constained GMM (3 groups)			
1	323	0	0
2	0	97	0
3	0	1	151

Table 3: Classifications obtained for the olive-oil dataset.

the values of: GMM-ASW, ASW, Adjusted Rand Index (ARI) with respect to the classification in three regions. The values of ARI are comparable with those reported in Table 3 by Azzalini and Torelli (2007), and obtained with k -means, Hierarchical classification with complete linkage and with Ward method, and with the density-based method proposed in that paper. The optimal classification produced by our hierarchical algorithm consists of two clusters: the first one essentially coincides with the union of *South* and *Sardinia*, whereas the second overlaps with *Centre-North*, with the exception of one individual from *Sardinia*. Notice that, if we constrain the number of groups to be equal to 3, the classification based on the GMM coincides with the one produced by our algorithm (of course, GMM-ASW differ on the two partitions because of differences in parameter estimates) and the ARI computed with respect to the true classification is very close to 1.

Classification	groups	GMM-ASW	ASW	ARI
GMM	5	0.6258	0.4152	0.5882
GMM + merge	2	0.7181	0.3478	0.6268
GMM + merge	3	0.6995	0.3479	0.9969
GMM (constr.)	3	0.8398	0.3488	0.9969

Table 4: Comparison of the clustering results for the olive-oil dataset.

groups	GMM-ASW
9	0.45569
8	0.50073
7	0.51268
6	0.52102
5	0.56051
4	0.58281
3	0.62450
2	0.60279

Table 5: Crabs dataset: values of GMM-ASW obtained with the hierarchical aggregation procedure

Crabs dataset. The crabs dataset (Campbell and Mahon, 1974) consists of five morphological measurements on $n = 200$ crabs, 50 each of two color forms (Blue and Orange) and both sexes. We can identify the four groups with A (Male, Orange), B (Male, Blue), C (Female, Orange) and D (Female, Blue). The joint application of EM algorithm for parameter estimation and BIC for model selection, leads to a 9-component mixture. However, as discussed in Hennig (2010), graphical diagnostic tools suggest a partition in less than nine clusters. Our hierarchical aggregation procedure leads to an optimal partition in three groups. Table 5 reports the value of GMM-ASW at each level of the hierarchy. Table 6 compares the optimal classification with the four-group partition in terms of sex and colour. It is worth to notice that we can discriminate the sex gender within the Orange species, whereas we can identify a unique cluster for the Blue one.

groups	A	B	C	D
1	0	50	0	50
2	0	0	45	0
3	50	0	5	0

Table 6: Crabs dataset: results of the 3-group classification obtained with the hierarchical aggregation procedure

5 Discussion

In this paper we propose a hierarchical procedure aimed at reducing the number of groups when a GMM based classification, identifying distinct clusters with distinct mixture components, overestimates the number of clusters. We introduced a generalisation of the silhouette width, the GMM-SW, which allows us to quantify the *within group* homogeneity and the *between group* separation. It is based on an asymmetric dissimilarity measure dependent on the squared Mahalanobis distances induced by the precision matrices of the components of the fitted GMM. Using GMM-ASW, we can evaluate the quality of both the most common GMM based classification, where a one-one correspondence between clusters and mixture components is assumed and the number of components is estimated by minimising the BIC, and classifications based on the aggregation of such clusters. The hierarchical aggregation procedure we propose optimises the GMM-ASW retaining the originally fitted GMM. For this reason, we can see it as a compromise between model based classification, which is mostly based on the likelihood function for both parameter estimation and model selection (although the penalty in the BIC is likelihood independent), and the silhouette based classification which looks for groups of similar objects, where similarity depends on a given dissimilarity measure which is usually defined independently of any statistical model.

References

- Azzalini, A. and N. Torelli (2007). Clustering via nonparametric density estimation. *Statistics and Computing* 17, 71–80. 10.1007/s11222-006-9010-y.
- Banfield, J. D. and A. E. Raftery (1993). Model-based gaussian and non-gaussian clustering. *Biometrics* 49(3), pp. 803–821.

- Baudry, J.-P., A. E. Raftery, G. Celeux, K. Lo, and R. Gottardo (2010). Combining mixture components for clustering. *Journal of Computational and Graphical Statistics* 19(2), 332–353.
- Biernacki, C., G. Celeux, and G. Govaert (2000, July). Assessing a mixture model for clustering with the integrated completed likelihood. *IEEE Trans. Pattern Anal. Mach. Intell.* 22(7), 719–725.
- Campbell, N. and R. Mahon (1974). A multivariate study of variation in two species of rock crab of the genus *leptograpsus*. *Australian Journal of Zoology* 22(3), 417–425.
- Forina, M., C. Armanino, S. Lanteri, and E. Tiscornia (1983). Classification of olive oils from their fatty acid composition. *Food Research and Data Analysis*, 189–214.
- Fraley, C., A. Raftery, T. Murphy, and L. Scrucca (2012). mclust Version 4 for R: Normal Mixture Modeling for Model-Based Clustering, Classification, and Density Estimation. Technical report, Technical Report.
- Fraley, C. and A. E. Raftery (2002). Model-based clustering, discriminant analysis, and density estimation. *Journal of the American Statistical Association* 97(458), pp. 611–631.
- Hennig, C. (2010). Methods for merging gaussian mixture components. *Advances in Data Analysis and Classification* 4, 3–34. 10.1007/s11634-010-0058-3.
- Kaufman, L. and P. J. Rousseeuw (1990). *Finding groups in data: an introduction to cluster analysis*. New York: John Wiley and Sons.
- McLachlan, G. and D. Peel (2000). *Finite Mixture Models*. Wiley series in probability and statistics: Applied probability and statistics. Wiley.
- R Development Core Team (2011). *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. ISBN 3-900051-07-0.
- Rousseeuw, P. J. (1987). Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics* 20(0), 53 – 65.

Figure 1: Example 1: first simulated data-set, GMM-SW and SW

Figure 2: Example1: second simulated data-set, GMM-SW and SW

Figure 3: Example 1, simulated data and single-component clusters

Figure 4: Example 1, a cluster is defined by the aggregation of the components 3 and 4

Figure 5: Example 1, a cluster is defined by the aggregation of the components 1 and 6

Figure 6: Example 1, a cluster is defined by the aggregation of the components 3, 4, and 5

Figure 7: Example 1, a cluster is defined by the aggregation of the components 2 and 5

Figure 8: Example 1, two clusters are defined by the aggregation of components 1 and 6, and 3 and 4 respectively