Outlier Detection via Parsimonious Mixtures of Contaminated Gaussian Distributions

Antonio Punzo * Paul D. McNicholas †

Abstract

For multivariate continuous data, the contaminated Gaussian distribution — having two parameters indicating the proportion of outliers and the degree of contamination — represents a convenient and natural way to model and detect outliers. In this paper, we introduce a mixture model whereby each mixture component is itself a contaminated Gaussian distribution. To introduce parsimony, a family of fourteen mixtures of contaminated Gaussian distributions is developed by applying constraints to eigen-decomposed component covariance matrices. This approach is, amongst other things, an effective alternative to trimmed clustering. Although these models could be used for model-based clustering, classification, and discriminant analysis, we focus on the more general model-based classification framework. An expectation-conditional maximization algorithm is used to find maximum likelihood estimates of the parameters and thereby give classifications for the observations. A simulation study is performed to evaluate the behaviour of the Bayesian information criterion and the integrated completed likelihood in model selection. This novel family of models is applied to artificial and real data in order to illustrate some of its advantages. Amongst them, and in contrast to the trimmed clustering approach, we have: 1) each observation has a posterior probability of belonging to a particular group and, inside each group, of being an outlier or not, 2) the models do not require pre-specification of quantities such as the proportion of observations to trim, 3) the approach can be easily used in high dimensions, and 4) model-based classification is permitted in addition to clustering.

Key words: Mixture models, Model-based classification, EM algorithm, Contaminated Gaussian distribution, Outlier detection, Robust estimates, Trimmed clustering.

*Department of Economics and Business, University of Catania, Corso Italia 55, 95129 Catania, Italy. Tel.: +39-095-7537640, e.mail: antonio.punzo@unict.it

†Department of Mathematics and Statistics, University of Guelph, Guelph, Ontario, Canada, N1G 2W1. Tel.: +1-519-8244120, ext. 53136, e.mail: paul.mcnicholas@uoguelph.ca

1 Introduction

Finite mixtures of distributions are commonly employed in statistical modelling with two different purposes (Titterington *et al.*, 1985, pp. 2–3). In *indirect applications*, they are used as semiparametric competitors of nonparametric density estimation techniques (see Titterington *et al.*, 1985, pp. 28–29, McLachlan and Peel, 2000, p. 8, and Escobar and West, 1995). On the other hand, in *direct applications*, finite mixture models are considered as a powerful device for clustering, classification, and discriminant analysis by assuming that one or more mixture components represent a group (or class or cluster) within the original data (see McLachlan and Basford, 1988 and Fraley and Raftery, 1998).

For continuous multivariate random variables, attention is commonly focused on mixtures of Gaussian distributions because of their computational and theoretical convenience. Unfortunately, real data are often "contaminated" by outliers that affect the estimation of the component means and covariance matrices (see, e.g., Bock, 2002). Thus, the detection of these outliers, and the development of robust methods of parameter estimation insensitive to the presence of outliers, are important practical problems. Following Gallegos and Ritter (2009), the mixture modelling literature on this topic can be summarized as follows (for the alternative trimmed clustering approach see, e.g., García-Escudero *et al.*, 2008, 2010).

1. Campbell (1984), McLachlan and Basford (1988, Section 2.8), and De Veaux and Krieger (1990) use M-estimates of the means and covariance matrices of the Gaussian components of the mixture model.

2. McLachlan and Peel (1998) and Peel and McLachlan (2000) introduce mixtures of t-distributions (see also Greselin and Ingrassia, 2010 and Andrews and McNicholas, 2011).

3. Fraley and Raftery (2002) add, to the mixture of Gaussian distributions, a uniform component on the convex hull of the data in order to accommodate outliers.

The performance of these aforementioned methods is analyzed in Hennig (2004).

4. Browne *et al.* (2012) introduce a mixture model whereby each mixture component is itself a mixture of a Gaussian and a uniform distribution.

However, these mixture-based approaches have some drawbacks. In direct applications, the first two methods do not allow for the direct detection of outliers. The approaches considering the uniform distribution, if used for discriminant analysis, cannot recognize a new noisy observation (an observation that has not been used to fit the model) if it lies outside the support defined by the fitted uniform distribution(s); this is paradoxical because the new observation should be the strongest available outlier in the philosophy of the corresponding model. Finally, in indirect applications, mixtures having one or

more uniform distributions do not provide an overall smooth density, which is a fundamental requirement in the nonparametric paradigm (Silverman, 1981).

To overcome these problems, a mixture of contaminated Gaussian distributions is proposed in Section 2.1. A contaminated Gaussian distribution is a two-component Gaussian mixture in which one of the components, with a large prior probability, represents the "good" observations, and the other, with a small prior probability, the same mean, and an inflated covariance matrix, represents the "bad" observations (Aitkin and Wilson, 1980). It represents a common and simple theoretical model for the occurrence of outliers. Furthermore, parsimonious variants of the proposed model are introduced in the fashion of Celeux and Govaert (1995) by imposing constraints on eigen-decomposed component covariance matrices (Section 2.2). The most general model-based classification framework is considered (Section 2.3) and an expectation-conditional maximization (ECM) algorithm for parameter estimation is outlined (Section 3). Further computational/operational aspects are detailed in Section 4. In Section 5, the Bayesian information criterion (BIC; Schwarz, 1978) and the integrated completed likelihood (ICL; Biernacki *et al.*, 2000) are compared for model selection for our novel family of mixtures of contaminated Gaussian distributions. Applications on artificial and real data are presented in Section 6 and a comparison with trimmed clustering, as implemented in the `tclust` package (Fritz *et al.*, 2012) of R (R Core Team, 2013), is discussed in Section 7.

2 Methodology

2.1 The general model

The distribution of a p-variate random vector $\boldsymbol{X}$, according to a parametric finite mixture model with k components, can be written as

$$p\left(\boldsymbol{x};\boldsymbol{\Psi}\right)=\sum_{j=1}^{k}\pi_j f\left(\boldsymbol{x};\boldsymbol{\vartheta}_j\right), \tag{1}$$

where π_j is the weight (mixing proportion) of the jth component, with $\pi_j>0$ and $\sum_{j=1}^{k}\pi_j=1$, $f\left(\boldsymbol{x};\boldsymbol{\vartheta}_j\right)$ is the parametric (with respect to $\boldsymbol{\vartheta}_j$) distribution associated with the jth component, and $\boldsymbol{\Psi}=\{\boldsymbol{\pi},\boldsymbol{\vartheta}\}$, with $\boldsymbol{\pi}=\{\pi_j\}_{j=1}^{k}$ and $\boldsymbol{\vartheta}=\{\boldsymbol{\vartheta}_j\}_{j=1}^{k}$, contains all of the parameters of the mixture. As usual, model (1) implicitly assumes that the component distributions should all belong to the same parametric family.

In this paper, as component density in (1), we adopt the *contaminated Gaussian distribution*

$$f\left(\boldsymbol{x};\boldsymbol{\vartheta}_j\right)=\alpha_j\phi\left(\boldsymbol{x};\boldsymbol{\mu}_j,\boldsymbol{\Sigma}_j\right)+\left(1-\alpha_j\right)\phi\left(\boldsymbol{x};\boldsymbol{\mu}_j,\eta_j\boldsymbol{\Sigma}_j\right),$$

where $\alpha_j \in [0,1]$, $\eta_j > 0$, $\boldsymbol{\vartheta}_j = \left\{\alpha_j, \boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j, \eta_j\right\}$, and

$$\phi\left(\boldsymbol{x}; \boldsymbol{\mu}, \boldsymbol{\Sigma}\right) = \left(2\pi\right)^{-\frac{p}{2}} \left|\boldsymbol{\Sigma}\right|^{-\frac{1}{2}} \exp\left\{-\frac{1}{2}\delta\left(\boldsymbol{x}, \boldsymbol{\mu}; \boldsymbol{\Sigma}\right)\right\} \tag{2}$$

is the distribution of a p-variate Gaussian random vector with mean $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$. In (2),

$$\delta\left(\boldsymbol{x}, \boldsymbol{\mu}; \boldsymbol{\Sigma}\right) = \left(\boldsymbol{x} - \boldsymbol{\mu}\right)' \boldsymbol{\Sigma}^{-1} \left(\boldsymbol{x} - \boldsymbol{\mu}\right)$$

denotes the Mahalanobis distance between $\boldsymbol{x}$ and $\boldsymbol{\mu}$ with covariance matrix $\boldsymbol{\Sigma}$. The result is the *mixture of contaminated Gaussian distributions*, given by

$$p\left(\boldsymbol{x}; \boldsymbol{\Psi}\right) = \sum_{j=1}^{k} \pi_j \left[\alpha_j \phi\left(\boldsymbol{x}; \boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j\right) + \left(1 - \alpha_j\right) \phi\left(\boldsymbol{x}; \boldsymbol{\mu}_j, \eta_j \boldsymbol{\Sigma}_j\right)\right], \tag{3}$$

where $\boldsymbol{\Psi} = \{\boldsymbol{\pi}, \boldsymbol{\alpha}, \boldsymbol{\vartheta}\}$, with $\boldsymbol{\alpha} = \{\alpha_j\}_{j=1}^{k}$. Previous work on mixtures of Gaussian mixtures can be found, for example, in Orbanz and Buhmann (2005) and Di Zio *et al.* (2007).

2.2 Parsimonious variants of the general model

Because there are $p\left(p+1\right)/2$ free parameters for each $\boldsymbol{\Sigma}_j$, it is usually necessary to introduce parsimony into the general model (3) for real applications. To this end, and following Celeux and Govaert (1995), we consider the eigen decomposition

$$\boldsymbol{\Sigma}_j = \lambda_j \boldsymbol{\Gamma}_j \boldsymbol{\Delta}_j \boldsymbol{\Gamma}_j', \qquad j = 1, \ldots, k, \tag{4}$$

where $\lambda_j = \left|\boldsymbol{\Sigma}_j\right|^{1/p}$, $\boldsymbol{\Delta}_j$ is the scaled ($\left|\boldsymbol{\Delta}_j\right| = 1$) diagonal matrix of the eigenvalues of $\boldsymbol{\Sigma}_j$ sorted in decreasing order, and $\boldsymbol{\Gamma}_j$ is a $p \times p$ orthogonal matrix whose columns are the normalized eigenvectors of $\boldsymbol{\Sigma}_j$, ordered according to their eigenvalues. Each component in the right side of (4) also has a different geometric interpretation: λ_j determines the volume of the cluster, $\boldsymbol{\Delta}_j$ its shape, and $\boldsymbol{\Gamma}_j$ its orientation.

The constraints we impose on the three components of (4) generate the family of fourteen parsimonious mixtures of contaminated Gaussian distributions (PMCGD) models summarized in Table 1. The same constraints are also applied by Celeux and Govaert (1995) to the classical mixtures of Gaussian distributions; the result is the well-known family of Gaussian parsimonious clustering (GPC) models. As shown in the first column of Table 1, the family of fourteen models can be further split into three subfamilies: *spherical*, *diagonal*, and *general*.

Table 1: Nomenclature, covariance structure, type of ML solution in the first CM-step of the ECM algorithm (CF=closed form and IP=iterative procedure), and number of free covariance parameters for each member of the PMCGD family.

Family	Model	Volume	Shape	Orientation	$\boldsymbol{\Sigma}_j$	ML	Free covariance parameters
Spherical	EII	Equal	Spherical	-	$\lambda \boldsymbol{I}$	CF	1
	VII	Variable	Spherical	-	$\lambda_j \boldsymbol{I}$	CF	k
Diagonal	EEI	Equal	Equal	Axis-Aligned	$\lambda \boldsymbol{\Delta}$	CF	p
	VEI	Variable	Equal	Axis-Aligned	$\lambda_j \boldsymbol{\Delta}$	IP	$k+p-1$
	EVI	Equal	Variable	Axis-Aligned	$\lambda \boldsymbol{\Delta}_j$	CF	$1+k\,(p-1)$
	VVI	Variable	Variable	Axis-Aligned	$\lambda_j \boldsymbol{\Delta}_j$	CF	kp
General	EEE	Equal	Equal	Equal	$\lambda \boldsymbol{\Delta\Gamma\Delta}'$	CF	$p\,(p+1)\,/2$
	VEE	Variable	Equal	Equal	$\lambda_j \boldsymbol{\Delta\Gamma\Delta}'$	IP	$k+p-1+p\,(p-1)\,/2$
	EVE	Equal	Variable	Equal	$\lambda \boldsymbol{\Delta}_j \boldsymbol{\Gamma} \boldsymbol{\Delta}_j'$	IP	$1+k\,(p-1)+p\,(p-1)\,/2$
	EEV	Equal	Equal	Variable	$\lambda \boldsymbol{\Delta\Gamma}_j \boldsymbol{\Delta}'$	CF	$p+kp\,(p-1)\,/2$
	VVE	Variable	Variable	Equal	$\lambda_j \boldsymbol{\Delta}_j \boldsymbol{\Gamma} \boldsymbol{\Delta}_j'$	IP	$kp+p\,(p-1)\,/2$
	VEV	Variable	Equal	Variable	$\lambda_j \boldsymbol{\Delta\Gamma}_j \boldsymbol{\Delta}'$	IP	$k+p-1+kp\,(p-1)\,/2$
	EVV	Equal	Variable	Variable	$\lambda \boldsymbol{\Delta}_j \boldsymbol{\Gamma}_j \boldsymbol{\Delta}_j'$	CF	$1+k\,(p-1)+kp\,(p-1)\,/2$
	VVV	Variable	Variable	Variable	$\lambda_j \boldsymbol{\Delta}_j \boldsymbol{\Gamma}_j \boldsymbol{\Delta}_j'$	CF	$kp\,(p+1)\,/2$

2.3 Modelling framework: model-based classification

Model-based classification is receiving renewed attention (see, e.g., Dean *et al.*, 2006, McNicholas, 2010, Andrews *et al.*, 2011, Browne and McNicholas, 2012a, and Subedi *et al.*, 2013). However, despite being the most general framework within which to present and analyze direct applications of mixture models, it remains the "poor cousin" of model-based clustering within the literature.

Consider n observations $\{\boldsymbol{x}_i\}_{i=1}^n$ of which, without loss of generality, are ordered so that the first m are known to belong to one of k groups; these are the so-called labeled observations. Let $\boldsymbol{z}_i$ be the k-dimensional component-label vector in which the jth element $z_{ij} = 1$ if $\boldsymbol{x}_i$ belongs to component j and $z_{ij} = 0$ otherwise, $j = 1, \ldots, k$. If the ith observation is labeled, denote with $\widetilde{\boldsymbol{z}}_i = (\widetilde{z}_{i1}, \ldots, \widetilde{z}_{ik})'$ its component-membership indicator.

In model-based classification, we use all n observations to estimate the parameters of the mixture; the fitted model is so adopted to classify each of the $n-m$ unlabelled observations through the corresponding maximum *a posteriori* (MAP) probability. We obtain the model-based clustering scenario as a special case when $m = 0$. In discriminant analysis, we use only the m labeled observations to estimate the parameters of the mixture; the fitted model is then adopted to classify, using the MAP criterion, each of the $n - m$ unlabelled observations.

3 Maximum likelihood estimation

To fit the PMCGD models, we adopt the ECM algorithm of Meng and Rubin (1993), a variant of the classical expectation-maximization (EM) algorithm (Dempster *et al.*, 1977) that is a natural approach

for maximum likelihood (ML) estimation when data are missing. In our case, there are two sources of missing data: one arises from the fact that we do not know some of the group memberships, that is, we do not know $\boldsymbol{z}_i$ for $i = m+1, \ldots, n$; the other arises from the fact that we do not know whether an observation in group j is "good" or "bad". To denote this second source of missing data, we introduce v_{ij} so that $v_{ij} = 1$ if observation i in group j is "good", and $v_{ij} = 0$ if observation i in group j is "bad". Thus, with $\mathcal{S}_l = \{\boldsymbol{x}_i, \widetilde{\boldsymbol{z}}_i, \boldsymbol{v}_i\}_{i=1}^{m}$ and $\mathcal{S}_u = \{\boldsymbol{x}_i, \boldsymbol{z}_i, \boldsymbol{v}_i\}_{i=m+1}^{n}$ we denote the labeled complete-data and the unlabeled complete-data, respectively. The complete-data are so indicated with $\mathcal{S} = \{\mathcal{S}_l, \mathcal{S}_u\}$. Accordingly, the complete-data log-likelihood, $l_c(\boldsymbol{\Psi}|\mathcal{S})$, can be decomposed as

$$l_c(\boldsymbol{\Psi}|\mathcal{S}) = l_{1c}(\boldsymbol{\pi}|\mathcal{S}) + l_{2c}(\boldsymbol{\alpha}|\mathcal{S}) + l_{3c}(\boldsymbol{\vartheta}|\mathcal{S}), \tag{5}$$

where

$$l_{1c}(\boldsymbol{\pi}|\mathcal{S}) = l_{1cl}(\boldsymbol{\pi}|\mathcal{S}_l) + l_{1cl}(\boldsymbol{\pi}|\mathcal{S}_u), \tag{6}$$

$$l_{2c}(\boldsymbol{\alpha}|\mathcal{S}) = l_{2cl}(\boldsymbol{\alpha}|\mathcal{S}_l) + l_{2cu}(\boldsymbol{\alpha}|\mathcal{S}_u), \tag{7}$$

and

$$l_{3c}(\boldsymbol{\vartheta}|\mathcal{S}) = l_{3cl}(\boldsymbol{\vartheta}|\mathcal{S}_l) + l_{3cu}(\boldsymbol{\vartheta}|\mathcal{S}_u). \tag{8}$$

The quantities in (6), (7), and (8) are defined as

$$l_{1cl}(\boldsymbol{\pi}|\mathcal{S}_l) = \sum_{i=1}^{m}\sum_{j=1}^{k} \widetilde{z}_{ij} \ln \pi_j,$$

$$l_{1cu}(\boldsymbol{\pi}|\mathcal{S}_u) = \sum_{i=m+1}^{n}\sum_{j=1}^{k} z_{ij} \ln \pi_j, \tag{9}$$

$$l_{2cl}(\boldsymbol{\alpha}|\mathcal{S}_l) = \sum_{i=1}^{m}\sum_{j=1}^{k} \widetilde{z}_{ij} \left[v_{ij} \ln \alpha_j + (1 - v_{ij}) \ln (1 - \alpha_j)\right], \tag{10}$$

$$l_{2cu}(\boldsymbol{\alpha}|\mathcal{S}_u) = \sum_{i=m+1}^{n}\sum_{j=1}^{k} z_{ij} \left[v_{ij} \ln \alpha_j + (1 - v_{ij}) \ln (1 - \alpha_j)\right], \tag{11}$$

$$l_{3cl}(\boldsymbol{\vartheta}|\mathcal{S}_l) = -\frac{1}{2}\sum_{i=1}^{m}\sum_{j=1}^{k}\left\{\widetilde{z}_{ij} \ln |\boldsymbol{\Sigma}_j| + p\widetilde{z}_{ij}(1 - v_{ij}) \ln \eta_j + \widetilde{z}_{ij}\left(v_{ij} + \frac{1 - v_{ij}}{\eta_j}\right)\delta\left(\boldsymbol{x}_i, \boldsymbol{\mu}_j; \boldsymbol{\Sigma}_j\right)\right\}, \tag{12}$$

and

$$l_{3cu}(\boldsymbol{\vartheta}|\mathcal{S}_u) = -\frac{1}{2}\sum_{i=m+1}^{n}\sum_{j=1}^{k}\left\{z_{ij} \ln |\boldsymbol{\Sigma}_j| + pz_{ij}(1 - v_{ij}) \ln \eta_j + z_{ij}\left(v_{ij} + \frac{1 - v_{ij}}{\eta_j}\right)\delta\left(\boldsymbol{x}_i, \boldsymbol{\mu}_j; \boldsymbol{\Sigma}_j\right)\right\}. \tag{13}$$

The ECM algorithm iterates between three steps — an E-step and two CM-steps — until convergence. The only difference from the EM algorithm is that each M-step is replaced by two simpler CM-steps. They arise from the partition $\boldsymbol{\Psi} = \{\boldsymbol{\Psi}_1, \boldsymbol{\Psi}_2\}$, where $\boldsymbol{\Psi}_1 = \left\{\pi_j, \alpha_j, \boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j\right\}_{j=1}^{k}$ and $\boldsymbol{\Psi}_2 = \{\eta_j\}_{j=1}^{k}$.

3.1 Model VVV

Here, we detail the ECM algorithm for the most general PMCGD model, i.e., the VVV model (3).

3.1.1 E-step

The E-step, on the $(r+1)$th iteration of the ECM algorithm requires the calculation of $Q\left(\boldsymbol{\Psi}|\boldsymbol{\Psi}^{(r)}\right)$, the current conditional expectation of $l_c\left(\boldsymbol{\Psi}|\mathcal{S}\right)$. In order to do this, we need to calculate $E_{\boldsymbol{\Psi}^{(r)}}\left(Z_{ij}|\boldsymbol{x}_i\right)$, for $i = m+1, \ldots, n$ and $j = 1, \ldots, k$, and $E_{\boldsymbol{\Psi}^{(r)}}\left(V_{ij}|\boldsymbol{x}_i, \boldsymbol{z}_i\right)$, for $i = 1, \ldots, n$ and $j = 1, \ldots, k$. They are given by

$$E_{\boldsymbol{\Psi}^{(r)}}\left(Z_{ij}|\boldsymbol{x}_i\right) = z_{ij}^{(r)} = \frac{\pi_j^{(r)} f\left(\boldsymbol{x}_i; \boldsymbol{\vartheta}_j^{(r)}\right)}{p\left(\boldsymbol{x}_i; \boldsymbol{\Psi}^{(r)}\right)}$$

and

$$E_{\boldsymbol{\Psi}^{(r)}}\left(V_{ij}|\boldsymbol{x}_i, \boldsymbol{z}_i\right) = v_{ij}^{(r)} = \frac{\alpha_j^{(r)} \phi\left(\boldsymbol{x}_i; \boldsymbol{\mu}_j^{(r)}, \boldsymbol{\Sigma}_j^{(r)}\right)}{f\left(\boldsymbol{x}_i; \boldsymbol{\vartheta}_j^{(r)}\right)},$$

respectively. Then, by substituting z_{ij} with $z_{ij}^{(r)}$ in (9), (11), and (13), v_{ij} with $v_{ij}^{(r)}$ in (10), (11), (12), and (13), and putting them together in (5), we obtain $Q\left(\boldsymbol{\Psi}|\boldsymbol{\Psi}^{(r)}\right)$.

3.1.2 CM-step 1

The first CM-step on the $(r+1)$th iteration of the ECM algorithm requires the calculation of $\boldsymbol{\Psi}_1^{(r+1)}$ as the value of $\boldsymbol{\Psi}_1$ that maximizes $Q\left(\boldsymbol{\Psi}|\boldsymbol{\Psi}^{(r)}\right)$ with $\boldsymbol{\Psi}_2$ fixed at $\boldsymbol{\Psi}_1^{(r)}$. In particular, we obtain

$$\pi_j^{(r+1)} = \frac{n_j^{(r)}}{n},$$

$$\alpha_j^{(r+1)} = \frac{1}{n_j^{(r)}} \left(\sum_{i=1}^{m} \widetilde{z}_{ij} v_{ij}^{(r)} + \sum_{i=m+1}^{n} z_{ij}^{(r)} v_{ij}^{(r)} \right), \tag{14}$$

$$\boldsymbol{\mu}_j^{(r+1)} = \frac{1}{s_j^{(r)}} \left[\sum_{i=1}^{m} \widetilde{z}_{ij} \left(v_{ij}^{(r)} + \frac{1 - v_{ij}^{(r)}}{\eta_j^{(r)}} \right) \boldsymbol{x}_i + \sum_{i=m+1}^{n} z_{ij}^{(r)} \left(v_{ij}^{(r)} + \frac{1 - v_{ij}^{(r)}}{\eta_j^{(r)}} \right) \boldsymbol{x}_i \right],$$

and

$$\boldsymbol{\Sigma}_j^{(r+1)} = \frac{1}{n_j^{(r)}} \boldsymbol{W}_j^{(r)},$$

where

$$n_j^{(r)} = \sum_{i=1}^{m} \widetilde{z}_{ij} + \sum_{i=m+1}^{n} z_{ij}^{(r)},$$

$$s_j^{(r)} = \sum_{i=1}^{m} \widetilde{z}_{ij} \left(v_{ij}^{(r)} + \frac{1 - v_{ij}^{(r)}}{\eta_j^{(r)}} \right) + \sum_{i=m+1}^{n} z_{ij}^{(r)} \left(v_{ij}^{(r)} + \frac{1 - v_{ij}^{(r)}}{\eta_j^{(r)}} \right),$$

and

$$\boldsymbol{W}_j^{(r+1)} = \boldsymbol{W}_{j,l}^{(r+1)} + \boldsymbol{W}_{j,u}^{(r+1)}$$

with

$$\boldsymbol{W}_{j,l}^{(r+1)} = \sum_{i=1}^{m} \widetilde{z}_{ij} \left(v_{ij}^{(r)} + \frac{1 - v_{ij}^{(r)}}{\eta_j^{(r)}} \right) \left(\boldsymbol{x}_i - \boldsymbol{\mu}_j^{(r+1)} \right) \left(\boldsymbol{x}_i - \boldsymbol{\mu}_j^{(r+1)} \right)'$$

and

$$\boldsymbol{W}_{j,u}^{(r+1)} = \sum_{i=m+1}^{n} z_{ij}^{(r)} \left(v_{ij}^{(r)} + \frac{1 - v_{ij}^{(r)}}{\eta_j^{(r)}} \right) \left(\boldsymbol{x}_i - \boldsymbol{\mu}_j^{(r+1)} \right) \left(\boldsymbol{x}_i - \boldsymbol{\mu}_j^{(r+1)} \right)'.$$

Given $\eta_j^{(r)} > 1$, in updating $\boldsymbol{\mu}_j$ and $\boldsymbol{\Sigma}_j$ the observations are downweighted in line with $(1 - v_{ij}^{(r)})$; see Little (1988) for a discussion on downweighting for the contaminated Gaussian distribution. This is an important aspect for robust estimation of $\boldsymbol{\mu}_j$ and $\boldsymbol{\Sigma}_j$.

3.1.3 CM-step 2

The second CM-step on the $(r+1)$th iteration of the ECM algorithm requires the calculation of $\boldsymbol{\Psi}_2^{(r+1)}$ as the value of $\boldsymbol{\Psi}_2$ that maximizes $Q\left(\boldsymbol{\Psi}|\boldsymbol{\Psi}^{(r)}\right)$ with $\boldsymbol{\Psi}_1$ fixed at $\boldsymbol{\Psi}_1^{(r+1)}$. In particular, we have to maximize

$$\begin{aligned} &-\frac{p}{2} \sum_{i=1}^{m} \widetilde{z}_{ij} \left(1 - v_{ij}^{(r)}\right) \ln \eta_j - \frac{p}{2} \sum_{i=m+1}^{n} z_{ij}^{(r)} \left(1 - v_{ij}^{(r)}\right) \ln \eta_j \\ &-\frac{1}{2} \sum_{i=1}^{m} \widetilde{z}_{ij} \frac{1 - v_{ij}^{(r)}}{\eta_j} \delta\left(\boldsymbol{x}_i, \boldsymbol{\mu}_j^{(r+1)}; \boldsymbol{\Sigma}_j^{(r+1)}\right) \\ &-\frac{1}{2} \sum_{i=m+1}^{n} \widetilde{z}_{ij}^{(r)} \frac{1 - v_{ij}^{(r)}}{\eta_j} \delta\left(\boldsymbol{x}_i, \boldsymbol{\mu}_j^{(r+1)}; \boldsymbol{\Sigma}_j^{(r+1)}\right) \end{aligned} \tag{15}$$

with respect to η_j, $j = 1, \ldots, k$. The updated estimate of η_j, which exists in closed form, is given by

$$\eta_j^{(r+1)} = \frac{\sum_{i=1}^{m} \widetilde{z}_{ij} \left(1 - v_{ij}^{(r)}\right) \delta\left(\boldsymbol{x}_i, \boldsymbol{\mu}_j^{(r+1)}; \boldsymbol{\Sigma}_j^{(r+1)}\right)}{p \sum_{i=1}^{m} \widetilde{z}_{ij} \left(1 - v_{ij}^{(r)}\right)} + \frac{\sum_{i=m+1}^{n} z_{ij}^{(r)} \left(1 - v_{ij}^{(r)}\right) \delta\left(\boldsymbol{x}_i, \boldsymbol{\mu}_j^{(r+1)}; \boldsymbol{\Sigma}_j^{(r+1)}\right)}{p \sum_{i=m+1}^{n} z_{ij}^{(r)} \left(1 - v_{ij}^{(r)}\right)},$$

which, by definition, is a positive quantity.

3.2 Parsimonious models

The ECM algorithm for the other PMCGD models changes only with respect to the way the terms of the decomposition of $\boldsymbol{\Sigma}_j$ are obtained in the first CM-step. In particular, these updates are analogous to those given by Celeux and Govaert (1995) for the GPC models; the only difference is that, on the $(r+1)$th iteration of the algorithm, $\boldsymbol{W}_j^{(r+1)}$ is used instead of the classical scattering matrix

$$\sum_{i=1}^{m} \widetilde{z}_{ij} \left(\boldsymbol{x}_i - \boldsymbol{\mu}_j^{(r+1)}\right) \left(\boldsymbol{x}_i - \boldsymbol{\mu}_j^{(r+1)}\right)' + \sum_{i=m+1}^{n} z_{ij}^{(r)} \left(\boldsymbol{x}_i - \boldsymbol{\mu}_j^{(r+1)}\right) \left(\boldsymbol{x}_i - \boldsymbol{\mu}_j^{(r+1)}\right)'.$$

4 Further aspects

Code for the ECM algorithm was written in the R computing environment. We used the recent `mixture` package (Browne and McNicholas, 2013b), which gives a flexible implementation of the EM algorithm for the GPC models, as a basis to implement this code. The `mixture` differs from the `Rmixmod` package (Biernacki *et al.*, 2008 and Lebret *et al.*, 2012) with respect to the algorithm used in the M-step to estimate parameters for the EVE and VVE models. In particular, the `Rmixmod` package adopts the classical FG-algorithm of Flury and Gautschi (1986) while the `mixture` package uses the majorization-minimization (MM) algorithm of Browne and McNicholas (2013a), which is preferable especially in high dimensions. For the alternative use of accelerated line search (ALS) algorithms to solve the same problem, see Browne and McNicholas (2012b).

4.1 Initialization

The choice of the starting values for the ECM algorithm constitutes an important issue. The standard initialization consists of selecting a value for $\boldsymbol{\Psi}^{(0)}$. In particular, a random initialization is usually repeated t times, from different random positions, and the solution maximizing the observed-data log-likelihood among these t runs is selected (see Biernacki *et al.* 2003, Karlis and Xekalaki 2003, and Bagnato and Punzo, 2012 for other more complicated strategies).

Instead of selecting $\boldsymbol{\Psi}^{(0)}$ randomly, we suggest the following technique. Each GPC model can be seen as nested in the corresponding PMCGD model. In particular, each GPC model can be obtained from the corresponding PMCGD model by fixing $\alpha_j = 1$ and/or $\eta_j = 1$, $j = 1, \ldots, k$; thus, with these constraints, the VEV-PMCGD and the VEV-GPC model are equivalent. Then, the EM estimates of the parameters for each GPC model — obtained with the `gpcm()` function of `mixture` package — along with the constraint $\alpha_j = \eta_j = 1$, $j = 1, \ldots, k$, can be used to initialize the corresponding PMCGD model. From an operational point of view, thanks to the monotonicity property of the ECM algorithm (see,

e.g., McLachlan and Krishnan, 2007, p. 33) this also guarantees that the observed-data log-likelihood of a PMCGD model will be always greater than or equal to the observed-data log-likelihood of the corresponding GPC model. This is a fundamental consideration for the use of likelihood-based model selection criteria for choosing between a PMCGD model and the corresponding GPC model.

4.2 Convergence criterion

The Aitken acceleration (Aitken, 1926) is used to estimate the asymptotic maximum of the log-likelihood at each iteration of the ECM algorithm. Based on this estimate, we can decide whether or not the algorithm has reached convergence; i.e., whether or not the log-likelihood is sufficiently close to its estimated asymptotic value. The Aitken acceleration at iteration $r+1$ is given by

$$a^{(r+1)} = \frac{l^{(r+2)} - l^{(r+1)}}{l^{(r+1)} - l^{(r)}},$$

where $l^{(r+2)}$, $l^{(r+1)}$, and $l^{(r)}$ are the observed-data log-likelihood values from iterations $r+2$, $r+1$, and r, respectively. Then, the asymptotic estimate of the log-likelihood at iteration $r+2$ is given by

$$l_{\infty}^{(r+2)} = l^{(r+1)} + \frac{1}{1 - a^{(r+1)}} \left(l^{(r+2)} - l^{(r+1)} \right),$$

cf. Böhning *et al.* (1994). The ECM algorithm can be considered to have converged when $l_{\infty}^{(r+2)} - l^{(r+1)} < \epsilon$ (see Lindsay, 1995 and McNicholas *et al.*, 2010).

4.3 Automatic detection of outliers

For a PMCGD model, the classification of an observation $\boldsymbol{x}_i$ means:

step 1. determine its group of membership;

step 2. establish if it is either a "good" or a "bad" observation in that group.

Let $\widehat{\boldsymbol{z}}_i$ and $\widehat{\boldsymbol{v}}_i$ denote, respectively, the expected values of $\boldsymbol{z}_i$ and $\boldsymbol{v}_i$ arising from the ECM algorithm. Then, to evaluate the group of membership of $\boldsymbol{x}_i$, $i = m+1, \ldots, n$, we can use the MAP classification induced by

$$\text{MAP}\left(\widehat{z}_{ij}\right) = \begin{cases} 1 & \text{if } \max_h \left\{\widehat{z}_{ih}\right\} \text{ occurs at component } j \\ 0 & \text{otherwise.} \end{cases}$$

Analogously, to detect if $\boldsymbol{x}_i$, $i = 1, \ldots, n$, is "good" or "bad", we can compute $\text{MAP}\left(\widehat{v}_{ij}\right)$, with j such that $\text{MAP}\left(\widehat{z}_{ij}\right) = 1$. The resulting information can be used to eliminate the outliers, if such an outcome

is desired (Berkane and Bentler, 1988). The remaining data may then be treated as effectively being distributed according to a mixture of Gaussian distributions.

4.4 Constraints for detection of outliers

When the PMCGD models are used for detection of outliers in each group, $(1-\alpha_j)$ should represent the percentage of outliers and η_j should denote the degree of contamination. Then, for the latter parameter, we could prefer the assumption $\eta_j > 1$ so that η_j can be meant as the increase in variability due to the "bad" observations (i.e., an inflation parameter). Operationally, on the $(r+1)$th iteration of the first CM-step, the `optim()` function, of the `stats` package of `R`, is used for a numerical search of the maximum $\eta_j^{(r+1)}$ of (15) over the interval $(1, \eta^*)$, with $\eta^* > 1$. In the analyses of Section 6, we fix $\eta^* = 1000$ in order to facilitate faster convergence. Furthermore, one could require that in the jth group, $j = 1, \ldots, k$, the proportion of "good" data is at least equal to a pre-determined value α_j^* (e.g., we might set $\alpha_j^* = 0.5$). In this case, the `optim()` function is also used for a numerical search of the maximum $\alpha_j^{(r+1)}$, over the interval $(\alpha_j^*, 1)$, of the function

$$\sum_{i=1}^{m} \widetilde{z}_{ij} \left[v_{ij}^{(r)} \ln \alpha_j + \left(1 - v_{ij}^{(r)}\right) \ln\left(1-\alpha_j\right) \right] + \sum_{i=m+1}^{n} z_{ij}^{(r)} \left[v_{ij}^{(r)} \ln \alpha_j + \left(1 - v_{ij}^{(r)}\right) \ln\left(1-\alpha_j\right) \right]. \qquad (16)$$

The function in (16) has already been used to obtain the updates of α_j in (14). Finally, to simplify parameter estimation, α_j and/or η_j may be fixed *a priori* by the user. With reference to α_j, this does not represent a strong restriction if one thinks that the trimmed clustering approach, for the detection of outliers, typically requires one to specify the proportion of outliers (the so-called trimming proportion) in advance (see, e.g., Fritz *et al.*, 2012).

5 Model selection

5.1 Model selection criteria

The PMCGD models, in addition to $\boldsymbol{\Psi}$, are also characterized by the particular covariance structure and by the number of components k. So far, these quantities have been treated as *a priori* fixed. Nevertheless, for practical purposes, model selection is usually required. One way to perform model selection is via computation of a convenient (likelihood-based) model selection criterion across all fourteen models and over a reasonable range of values for k, and then choosing the model associated with the best value of the adopted criterion.

The BIC is by far the most popular such criterion within the literature to date. We have

$$\text{BIC} = 2l\left(\widehat{\boldsymbol{\Psi}}\right) - \rho \ln n, \tag{17}$$

where ρ is the overall number of free parameters in the model. One alternative that has been used by several authors of late (e.g., McNicholas and Subedi, 2012) is the ICL, given by

$$\text{ICL} \approx \text{BIC} + \sum_{i=m+1}^{n} \sum_{j=1}^{k} \text{MAP}\left(\widehat{z}_{ij}\right) \ln \widehat{z}_{ij}, \tag{18}$$

where $\sum_{i=m+1}^{n} \sum_{j=1}^{k} \text{MAP}\left(\widehat{z}_{ij}\right) \ln \widehat{z}_{ij}$ is the estimated mean entropy, which reflects the uncertainty in the classification of observation i into component j.

5.2 Comparing the BIC and the ICL

We present here the results of a simulation study implemented in R with the aim of comparing the performance of the BIC and the ICL for PMCGD model selection. Because many factors come into play (e.g., the number of components k, the dimension p of the observed variables, the overall sample size n, the number m of labeled observations), some of them are necessarily considered fixed for our purposes.

5.2.1 Design

One-hundred data sets are generated from each model in our family. We fix: $p = 2$, $k = 2$, $n = 300$, $\pi_1 = \pi_2 = 0.5$, $\alpha_1 = \alpha_2 = 0.9$, $\eta_1 = \eta_2 = 10$, and $\boldsymbol{\mu}_1 = \mathbf{0}$. With regard to the remaining parameters of the PMCGD models, in the bivariate case we have

$$\boldsymbol{\Sigma}_j = \lambda_j \boldsymbol{\Gamma}_j \boldsymbol{\Delta}_j \boldsymbol{\Gamma}_j' = \lambda_j \boldsymbol{R}\left(\gamma_j\right) \begin{pmatrix} 1/\delta_j & 0 \\ 0 & \delta_j \end{pmatrix} \boldsymbol{R}\left(\gamma_j\right)', \tag{19}$$

where

$$\boldsymbol{R}\left(\gamma_j\right) = \begin{pmatrix} \cos\gamma_j & -\sin\gamma_j \\ \sin\gamma_j & \cos\gamma_j \end{pmatrix}$$

is the rotation matrix of angle γ_j, and $\delta_j \in (0, 1]$. Note that the elements in the shape matrix arise from the constraint $|\boldsymbol{\Delta}_j| = 1$. Hence, we have a single parameter for each component of the eigen-decomposition: λ_j is the volume parameter, δ_j is the shape parameter, and γ_j is the orientation parameter (see Greselin and Punzo, 2013, for further details). To generate data from each model, we preliminarily set $\boldsymbol{\Sigma}_1$ according to the subfamilies in Table 1. In particular we consider: $\lambda_1 = 1$, $\delta_1 = 1$, and $\gamma_1 = 0$ (0°

in degrees) for the spherical family, $\lambda_1 = 1$, $\delta_1 = 0.7$, and $\gamma_1 = 0$ for the diagonal family, and $\lambda_1 = 1$, $\delta_1 = 0.7$, and $\gamma_1 = \pi/6$ (30° in degrees) for the general family. With regard to $\boldsymbol{\Sigma}_2$ we choose: $\lambda_2 = 3$ for models with variable volume, $\delta_2 = 0.3$ for models with variable size, and $\gamma_2 = \pi/6 + \pi/4$ (30° + 45° in degrees) for models with variable orientation. In order to make a fair comparison in clustering/classification terms among these different configurations, the second variate μ_{22} of $\boldsymbol{\mu}_2 = (0, \mu_{22})'$ is computed, via a numerical procedure, to guarantee a fixed overlap between the good observations of the two groups. In line with Bagnato *et al.* (2013), we adopt the well-known measure of overlap of Bhattacharyya (1943),

$$B = \frac{1}{8}\delta\left(\boldsymbol{\mu}_1, \boldsymbol{\mu}_2; \boldsymbol{\Sigma}\right) + \frac{1}{2}\ln\left(\frac{|\boldsymbol{\Sigma}|}{\sqrt{\boldsymbol{\Sigma}_1 + \boldsymbol{\Sigma}_2}}\right),$$

with $\boldsymbol{\Sigma} = \left(\boldsymbol{\Sigma}_1 + \boldsymbol{\Sigma}_2\right)/2$, which takes values between 0 (absence of overlap) and 1 (maximum overlap). In particular, we consider four scenarios: $B = 0.05$, $B = 0.15$, $B = 0.25$, and $B = 0.35$. Finally, two values for the number m of unlabeled observations are also used: $m = 0$ and $m = 50$. In the latter case, fifty numbers are randomly generated, without replacement, from the set $\{1, \ldots, 300\}$ and then used as indicators of the observations to consider as labeled (using the true labels for them). All eight combinations of the factors B and m are taken into account in the simulations.

5.2.2 Results

Table 2 shows the percentage of times the BIC and the ICL discover the true generating model in the corresponding column. Note that, in each replication, the fourteen models of the PMCGD family are

Table 2: Percentage of times the BIC and the ICL select the true PMCGD model in the corresponding column.

			EII	VII	EEI	VEI	EVI	VVI	EEE	VEE	EVE	EEV	VVE	VEV	EVV	VVV	mean
$m = 0$	$B = 0.05$	BIC	96	91	96	93	96	93	90	81	95	63	94	75	89	87	88.50
		ICL	94	91	95	94	97	93	89	84	94	66	95	76	90	86	88.86
	$B = 0.15$	BIC	96	87	95	81	95	93	84	75	95	56	94	56	87	80	83.86
		ICL	92	88	93	79	93	93	86	87	92	54	96	51	88	83	83.93
	$B = 0.25$	BIC	97	87	83	81	93	89	67	68	97	44	90	49	84	71	78.57
		ICL	86	83	79	82	93	88	70	72	90	42	91	32	85	72	76.07
	$B = 0.35$	BIC	95	82	55	72	90	88	51	59	91	38	81	27	77	68	69.57
		ICL	59	73	56	65	72	74	56	58	72	25	76	12	73	51	58.71
$m = 50$	$B = 0.05$	BIC	87	95	97	86	97	91	97	86	96	56	97	79	86	82	88.00
		ICL	88	95	95	85	97	91	97	89	96	58	97	82	87	84	88.64
	$B = 0.15$	BIC	94	95	97	85	93	93	90	82	93	48	97	60	82	80	84.93
		ICL	94	96	94	89	93	93	93	90	94	55	97	59	83	85	86.79
	$B = 0.25$	BIC	96	94	89	94	97	96	79	79	90	56	95	60	84	80	84.93
		ICL	95	92	91	89	95	92	88	81	94	56	90	53	88	84	84.86
	$B = 0.35$	BIC	93	90	89	83	93	94	71	77	95	52	94	58	82	83	82.43
		ICL	87	85	88	79	89	94	81	71	94	48	91	45	87	83	80.14

directly fitted with $k = 2$. The last column of Table 2 helps us to view the overall results by showing

the means of the percentages computed by row. We can see that, apart from the pair $(m = 0, B = 0.35)$, where the BIC outperforms the ICL, the performance of the two model selection criteria is similar. This similarity is also corroborated by the overall mean of 82.60% for the BIC and 81.00% for the ICL. Furthermore, we can see that the behavior of the BIC and the ICL deteriorates with the increase in B when $m = 0$, while it remains more stable with the increase in B when $m = 50$; this means that, as expected, knowledge of the labels for some of the observations helps the BIC and the ICL in discovering the true model in situations of increasing overlap between groups. It is also interesting to note that, for some models such as the EEV and the VEV, the considerable overlap between groups makes assessment of the true model challenging. In summary, it is difficult to establish the best model selection criteria among those considered; thus, we will use both in the data analysis presented in Section 6.

6 Data analysis

In this section, we will evaluate the performance of the PMCGD models on artificial and real data sets. Particular attention will be devoted to the problem of detecting outliers.

6.1 Artificial data: overall uniform noise

In this first analysis, a sample of $n = 180$ simulated bivariate points is generated from an EVE-GPC model with $k = 2$ components of equal size ($n_1 = n_2 = 90$). The parameters of the mixture components are given in Table 3. Twenty noise points are also added from a uniform distribution over the range -10 to 10 on each variate. The true grouping is shown in Figure 1.

Table 3: Means and components of the eigen decompostion in (4) for the EVE-GPC model of Section 6.1.

Group 1	Group 2
$\boldsymbol{\mu}_1 = \begin{pmatrix} -2 \\ -2 \end{pmatrix}$	$\boldsymbol{\mu}_1 = \begin{pmatrix} 2 \\ 2 \end{pmatrix}$
$\lambda_1 = 1$	$\lambda_2 = 1$
$\boldsymbol{\Delta}_1 = \begin{pmatrix} 1/0.7 & 0 \\ 0 & 0.7 \end{pmatrix}$	$\boldsymbol{\Delta}_2 = \begin{pmatrix} 1/0.3 & 0 \\ 0 & 0.3 \end{pmatrix}$
$\boldsymbol{\Gamma}_1 = \begin{pmatrix} \sqrt{3}/2 & 1/2 \\ -1/2 & \sqrt{3}/2 \end{pmatrix}$	$\boldsymbol{\Gamma}_2 = \begin{pmatrix} \sqrt{3}/2 & 1/2 \\ -1/2 & \sqrt{3}/2 \end{pmatrix}$

We now consider the results obtained by fitting the family of GPC models. With this aim, the `gpcm()` function, of the `mixture` package for R, is used. When $k = 2$, the best GPC model according to the BIC

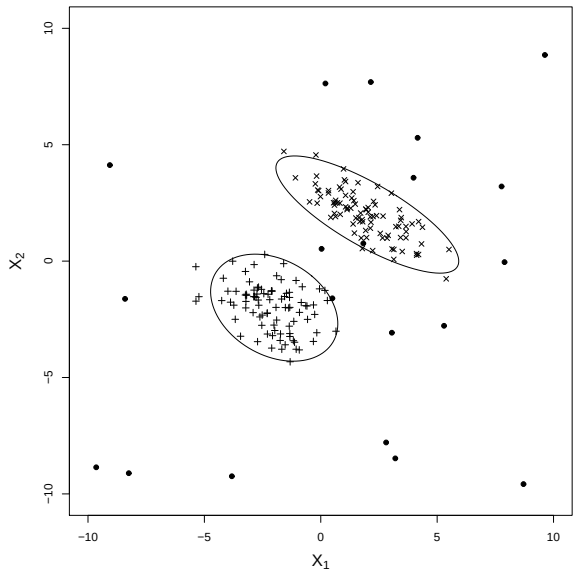
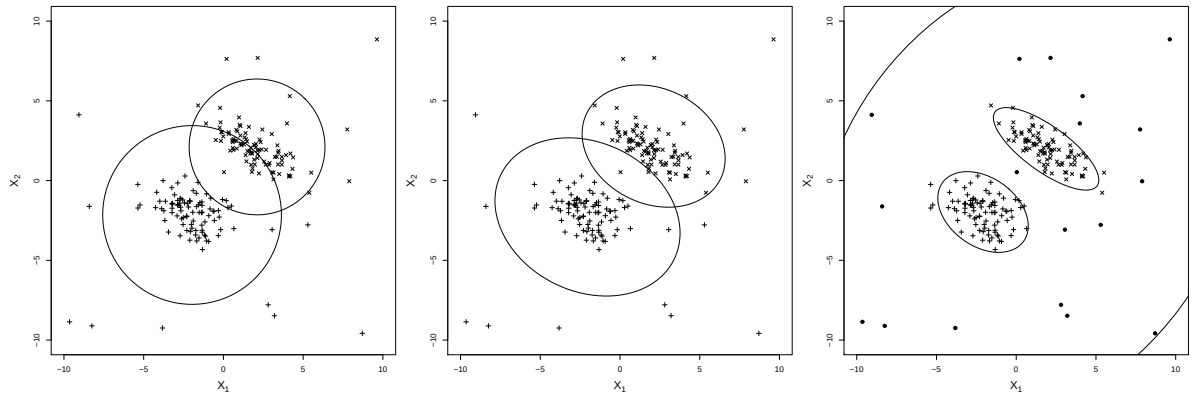


Figure 1: Scatter plot and true ellipses of equal (95%) concentration of the "uniform-contaminated" EVE-GPC model of Section 6.1. Uniform noise points are denoted by •.

(−1957.293) is VII while according to the ICL (−1960.651) is VEE. These two models are displayed in Figure 2(a) and Figure 2(b), respectively. As expected, these models are affected by noise. When $k = 3$,



(a) VII - best BIC when $k = 2$ (b) VEE - best ICL when $k = 2$ (c) VVE - best BIC and ICL when $k = 3$

Figure 2: Scatter plots and ellipses of equal (95%) concentration of some GPC models fitted on the simulated data of Section 6.1.

the best GPC model according to both the BIC (−1729.945) and the ICL (−1736.400) is the VVE in Figure 2(c). Obviously, the additional third component is attempting to model the background noise. However, this attempt affects the detection of the underlying EVE model.

On the contrary, by fitting our family of models with $k = 2$, the best model according to both the BIC (−1729.442) and the ICL (−1735.658) is the true one, with corresponding clustering represented in Figure 3. It compares very well with the true grouping (Figure 1). In particular, from the clustering results of Table 4, we can see that the model recognizes 18 out of 20 noise observations. Finally, note

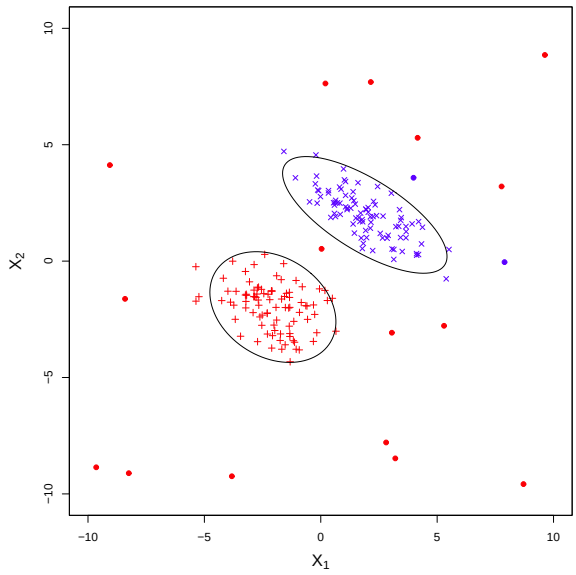


Figure 3: Scatter plot and ellipses of equal (95%) concentration of the EVE-PMCGD model fitted on the simulated data of Section 6.1. Detected outliers are denoted with bullets.

that among the BIC and ICL values seen so far, the highest ones refer to the EVE-PMCGD model, which represents the best model from this point of view as well.

Table 4: Clustering from the EVE-PMCGD model for the artificial data of Section 6.1.

True \ Fitted	Group 1	Group 2	Noise
Group 1	90	–	–
Group 2	–	90	–
Noise	1	1	18

6.2 Real data: `geyser2` data set

The second analysis considers the Old Faithful Geyser data set, which contains 272 observation of eruption length (see, e.g., Azzalini and Bowman, 1990). In line with García-Escudero *et al.* (2003) and Fritz *et al.* (2012), a bivariate data set can be constructed considering the eruption lengths and the corresponding previous eruption lengths. This data set, named `geyser2`, accompanies the `tclust` package for R.

In Fritz *et al.* (2012), the number of groups is fixed to $k = 3$ and a pre-assigned proportion 0.03 of the data is trimmed. Following this road, we fix $k = 3$ and $\alpha_j = 0.95$, $j = 1, 2, 3$. The best PMCGD model, according to both the BIC (-1113.730) and the ICL (-1121.339), is VII, with corresponding clustering displayed in Figure 4. Among the detected outliers, we can see six anomalous "short followed by short" eruption lengths. Notice that two other observations are detected as noise, one of which is situated

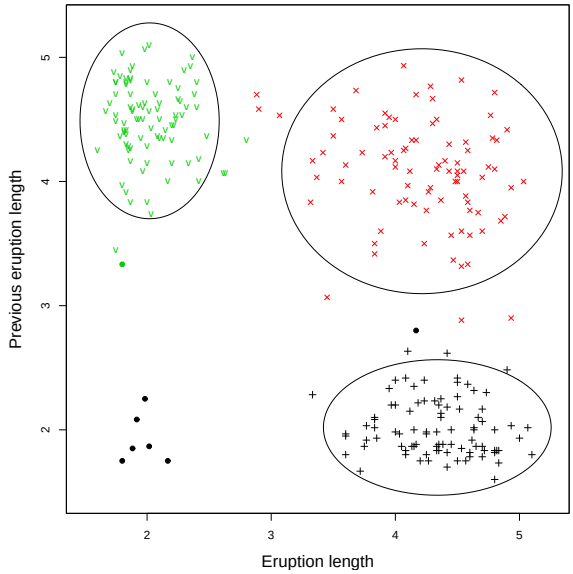


Figure 4: Scatter plot and ellipses of equal (95%) concentration of the VVI-PMCGD model fitted to the `geyser2` data set. Detected outliers are denoted with bullets.

between the two clusters on the right. Contrary to what one might expect, this analysis underlines that the PMCGD models, in addition to being able to find surrounding noisy data as in Section 6.1, can also detect outliers when they are located in a "separated" part of the space.

6.3 Real data: blue crab data set

The third analysis is based on the very popular crab data set of Campbell and Mahon (1974). Attention is focused on the sample of $n = 100$ blue crabs of the genus *Leptograpsus*, of which there are $n_1 = 50$ males (group 1) and $n_2 = 50$ females (group 2). For each specimen, we consider $p = 2$ measurements (in millimeters), namely the rear width (RW) and the length along the midline of the carapace (CL). The scatter plot of these data is shown in Figure 5.

Following the scheme of Peel and McLachlan (2000), eight "perturbed" data sets are generated by substituting the original value of CL for the 25th point (highlighted by a yellow bullet in Figure 5) with eight atypical values. The aim of Peel and McLachlan (2000) was to show that the t mixture-based clustering is robust to these perturbations unlike Gaussian mixture-based clustering. Here, we will show that our approach to clustering systematically outperforms the t mixture-based approach.

Ceteris paribus with the two approaches considered in Peel and McLachlan (2000), we will directly fit the VVV-PMCGD model with $k = 2$. For each of the three competing techniques, Table 5 reports the number of misallocated observations for each perturbed data set. Results for the Gaussian mixture and the t mixture are taken from Peel and McLachlan (2000, Table 1). It can be seen that the VVV-PMCGD model is systematically the most robust to these perturbations, with the number of misallocated

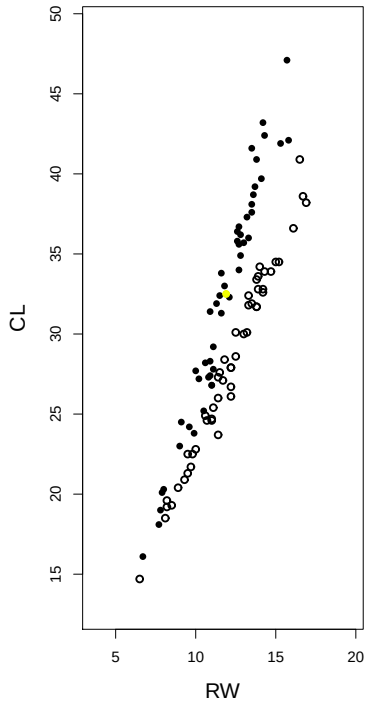


Figure 5: Scatter plot of the blue crab data (○ denotes male and ● female; ● denotes the observation perturbed for the analysis of Section 6.3).

Table 5: Number of misallocated blue crabs ($n = 100$), with respect to gender, for three model-based approaches to clustering. The last column reports the estimated value of the inflation parameter η in the group containing the outlier.

Value	GPC (VVV)	t mixture	PMCGD (VVV)	$\widehat{\eta}$
-15	49	19	13	480.062
-10	49	19	13	386.877
-5	21	20	13	299.083
-0	19	18	13	222.148
5	21	20	13	156.084
10	50	20	13	100.949
15	47	20	13	56.680
20	49	20	13	22.793

observations remaining fixed at 13 regardless of the particular value perturbed. Furthermore, our model always detects only one outlier, the true one, and it always belongs to the true group of female blue crabs (denoted with ● in Figure 5). Finally, by recalling that the original value of CL for the 25th point was 32.5, it is also interesting to note how the estimated values of η (in the group containing the outlier) move in line with the departure of the perturbed point from the group of membership.

7 Comparison with trimmed clustering

Fritz *et al.* (2012) recently introduced the `tclust` package for R as a non-heirarchical and model-based trimming approach to clustering. In this approach, a proportion p of the most outlying observations is trimmed by assuming Gaussian groups, as we did here. In particular, the `tclust()` function implements different algorithms aimed at maximizing the likelihood of the so-called *spurious outliers model* (see Gallegos, 2002 and Gallegos and Ritter, 2005) under different types of constraints (specified by the argument `restr`) and different possibilities for their strength (as controlled by the argument `restr.fact`). Although the approach is flexible in terms of possibilities given to the user, it has the following drawbacks when compared to our approach:

1. `tclust` adopts a hard clustering approach, meaning that each observation is either trimmed or fully assigned to a cluster. In comparison, our approach is "double-soft", meaning that each observation has a posterior probability of cluster membership and, in each cluster, a posterior probability to be either "good" or "bad". While our soft classification can always be converted into a hard partition (see Section 4.3), the reverse is not true.

2. In `tclust`, we have to pre-specify four quantities: k, p, `restr`, and `restr.fact`. Although some general guidelines are given in Fritz *et al.* (2012), no automatic criterion is provided to select them. This is a significant problem because we cannot expect the user to know these quantities in advance, especially for high-dimensional spaces; on the contrary, the user should ask the statistical method to discover the best specification for these quantities. Our approach is based on k, α_j, η_j, and the type of constraint for $\boldsymbol{\Sigma}_j$. Differently from `tclust`, ML can be used to estimate α_j and η_j (see Section 3), and automatic criteria, such as the BIC and the ICL, can be adopted to select k and the type of constraint for $\boldsymbol{\Sigma}_j$ (see Section 5). However, as mentioned in Section 4.4, we also allow the user the possibility of specifying these quantities if preliminary information about them is available.

3. While `tclust` only allows for clustering, the PMCGD models can also be used for model-based classification and discriminant analysis.

8 Conclusions

A family of fourteen parsimonious mixtures of contaminated Gaussian distributions (PMCGD) models has been introduced within the general model-based classification framework. These models generalize the well-known family of fourteen Gaussian parsimonious clustering (GPC) models of Celeux and Govaert

(1995). This generalization provides a more robust approach to the fitting of GPC models, as observations that are atypical of a component are given reduced weight in the calculation of its mean vector and covariance matrix. Moreover, natural detection of outliers is facilitated without the use of exogenous trimming methods. Application of these models to simulated and real data demonstrates their superior performance compared to the well-established GPC models as well as to other models within the mixture modelling framework. Future work will focus on contamination of non-elliptical component densities to allow an even more flexible modelling paradigm.

Acknowledgements

This work was carried out while the first author was visiting the University of Guelph. This visit was funded by the University Research Chair in Computational Statistics and by a Discovery Grant from the Natural Sciences and Engineering Research Council of Canada.

References

Aitken, A. (1926). On Bernoulli's numerical solution of algebraic equations. In *Proceedings of the Royal Society of Edinburgh*, volume 46, pages 289–305.

Aitkin, M. and Wilson, G. T. (1980). Mixture models, outliers, and the EM algorithm. *Technometrics*, **22**(3), 325–331.

Andrews, J. L. and McNicholas, P. D. (2011). Extending mixtures of multivariate t-factor analyzers. *Statistics and Computing*, **21**(3), 361–373.

Andrews, J. L., McNicholas, P. D., and Subedi, S. (2011). Model-based classification via mixtures of multivariate t-distributions. *Computational Statistics and Data Analysis*, **55**(1), 520–529.

Azzalini, A. and Bowman, A. W. (1990). A look at some data on the Old Faithful geyser. *Applied Statistics*, **39**(3), 357–365.

Bagnato, L. and Punzo, A. (2012). Finite mixtures of unimodal beta and gamma densities and the k-bumps algorithm. *Computational Statistics*.

Bagnato, L., Greselin, F., and Punzo, A. (2013). On the spectral decomposition in normal discriminant analysis. *Communications in Statistics - Simulation and Computation*.

Berkane, M. and Bentler, P. M. (1988). Estimation of contamination parameters and identification of outliers in multivariate data. *Sociological Methods & Research*, **17**(1), 55–64.

Bhattacharyya, A. (1943). On a measure of divergence between two statistical populations defined by their probability distributions. *Bulletin of the Calcutta Mathematical Society*, **35**(4), 99–109.

Biernacki, C., Celeux, G., and Govaert, G. (2000). Assessing a mixture model for clustering with the integrated completed likelihood. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, **22**(7), 719–725.

Biernacki, C., Celeux, G., and Govaert, G. (2003). Choosing starting values for the EM algorithm for getting the highest likelihood in multivariate Gaussian mixture models. *Computational Statistics & Data Analysis*, **41**(3-4), 561–575.

Biernacki, C., Celeux, G., Govaert, G., Langrognet, F., Noulin, G., and Vernaz, Y. (2008). MIXMOD - *Statistical Documentation*. downloadable from `http://www.mixmod.org/IMG/pdf/statdoc_2_1_1.pdf`.

Bock, H. (2002). Clustering methods: From classical models to new approaches. *Statistics in Transition*, **5**(5), 725–758.

Böhning, D., Dietz, E., Schaub, R., Schlattmann, P., and Lindsay, B. (1994). The distribution of the likelihood ratio for mixtures of densities from the one-parameter exponential family. *Annals of the Institute of Statistical Mathematics*, **46**(2), 373–388.

Browne, R. P. and McNicholas, P. D. (2012a). Model-based clustering, classification, and discriminant analysis of data with mixed type. *Journal of Statistical Planning and Inference*, **142**(11), 2976–2984.

Browne, R. P. and McNicholas, P. D. (2012b). Orthogonal Stiefel manifold optimization for eigen-decomposed covariance parameter estimation in mixture models. *Statistics and Computing*.

Browne, R. P. and McNicholas, P. D. (2013a). Estimating common principal components in high dimensions. Arxiv preprint arXiv:1302.2102.

Browne, R. P. and McNicholas, P. D. (2013b). ***mixture**: Mixture Models for Clustering and Classification*.

Browne, R. P., McNicholas, P. D., and Sparling, M. D. (2012). Model-based learning using a mixture of mixtures of Gaussian and uniform distributions. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **34**(4), 814–817.

Campbell, N. (1984). Mixture models and atypical values. *Mathematical Geology*, **16**(5), 465–477.

Campbell, N. A. and Mahon, R. J. (1974). A multivariate study of variation in two species of rock crab of genus Leptograpsus. *Australian Journal of Zoology*, **22**(3), 417–425.

Celeux, G. and Govaert, G. (1995). Gaussian parsimonious clustering models. *Pattern Recognition*, **28**(5), 781–793.

De Veaux, R. D. and Krieger, A. M. (1990). Robust estimation of a normal mixture. *Statistics & Probability Letters*, **10**(1), 1–7.

Dean, N., Murphy, T. B., and Downey, G. (2006). Using unlabelled data to update classification rules with applications in food authenticity studies. *Journal of the Royal Statistical Society: Series C*, **55**(1), 1–14.

Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society. Series B (Methodological)*, **39**(1), 1–38.

Di Zio, M., Guarnera, U., and Rocci, R. (2007). A mixture of mixture models for a classification problem: The unity measure error. *Computational Statistics & Data analysis*, **51**(5), 2573–2585.

Escobar, M. and West, M. (1995). Bayesian density estimation and inference using mixtures. *Journal of the American Statistical Association*, **90**(430), 577–588.

Flury, B. N. and Gautschi, W. (1986). An algorithm for simultaneous orthogonal transformation of several positive definite matrices to nearly diagonal form. *SIAM Journal on Scientific and Statistical Computing*, **7**(1), 169–184.

Fraley, C. and Raftery, A. E. (1998). How many clusters? Which clustering method? Answers via model-based cluster analysis. *Computer Journal*, **41**(8), 578–588.

Fraley, C. and Raftery, A. E. (2002). Model-based clustering, discriminant analysis, and density estimation. *Journal of the American Statistical Association*, **97**(458), 611–631.

Fritz, H., García-Escudero, L. A., and Mayo-Iscar, A. (2012). **tclust**: An `r` package for a trimming approach to cluster analysis. *Journal of Statistical Software*, **47**(12), 1–26.

Gallegos, M. T. (2002). Maximum likelihood clustering with outliers. In K. Jajuga, A. Sokolowski, and H.-H. Bock, editors, *Classification, Clustering, and Data Analysis: Recent Advances and Applications*, Studies in Classification, Data Analysis, and Knowledge Organization, pages 247–255. Springer, Berlin - Heidelberg.

Gallegos, M. T. and Ritter, G. (2005). A robust method for cluster analysis. *The Annals of Statistics*, **33**(1), 347–380.

Gallegos, M. T. and Ritter, G. (2009). Trimmed ML estimation of contaminated mixtures. *Sankhyā: The Indian Journal of Statistics, Series A*, **71**(2), 164–220.

García-Escudero, L. A., Gordaliza, A., and Matrán, C. (2003). Trimming tools in exploratory data analysis. *Journal of Computational and Graphical Statistics*, **12**(2).

García-Escudero, L. A., Gordaliza, A., Matrán, C., and Mayo-Iscar, A. (2008). A general trimming approach to robust cluster analysis. *The Annals of Statistics*, **36**(3), 1324–1345.

García-Escudero, L. A., Gordaliza, A., Matrán, C., and Mayo-Iscar, A. (2010). A review of robust clustering methods. *Advances in Data Analysis and Classification*, **4**(2), 89–109.

Greselin, F. and Ingrassia, S. (2010). Constrained monotone EM algorithms for mixtures of multivariate t distributions. *Statistics and Computing*, **20**(1), 9–22.

Greselin, F. and Punzo, A. (2013). Closed likelihood ratio testing procedures to assess similarity of covariance matrices. *The American Statistician (in press)*.

Hennig, C. (2004). Breakdown points for maximum likelihood estimators of location-scale mixtures. *The Annals of Statistics*, **32**(4), 1313–1340.

Karlis, D. and Xekalaki, E. (2003). Choosing initial values for the EM algorithm for finite mixtures. *Computational Statistics & Data Analysis*, **41**(3–4), 577–590.

Lebret, R., Iovleff, S., Langrognet, F., Biernacki, C., Celeux, G., and Govaert, G. (2012). ***Rmixmod**: The* R *Package of the Model-Based Unsupervised, Supervised and Semi-Supervised Classification Mixmod Library*.

Lindsay, B. (1995). *Mixture Models: Theory, Geometry and Applications*, volume 5. NSF-CBMS Regional Conference Series in Probability and Statistics, Institute of Mathematical Statistics, Hayward, California.

Little, R. J. A. (1988). Robust estimation of the mean and covariance matrix from data with missing values. *Applied Statistics*, **37**(1), 23–38.

McLachlan, G. J. and Basford, K. E. (1988). *Mixture Models: Inference and Applications to Clustering*. Marcel Dekker, New York.

McLachlan, G. J. and Krishnan, T. (2007). *The EM algorithm and extensions*. John Wiley & Sons, New York.

McLachlan, G. J. and Peel, D. (1998). Robust cluster analysis via mixtures of multivariate t-distributions. In A. Amin, D. Dori, P. Pudil, and H. Freeman, editors, *Advances in Pattern Recognition*, volume 1451 of *Lecture Notes in Computer Science*, pages 658–666. Springer, Berlin - Heidelberg.

McLachlan, G. J. and Peel, D. (2000). *Finite Mixture Models*. John Wiley & Sons, New York.

McNicholas, P. D. (2010). Model-based classification using latent Gaussian mixture models. *Journal of Statistical Planning and Inference*, **140**(5), 1175–1181.

McNicholas, P. D. and Subedi, S. (2012). Clustering gene expression time course data using mixtures of multivariate t-distributions. *Journal of Statistical Planning and Inference*, **142**(5), 1114–1127.

McNicholas, P. D., Murphy, T. B., McDaid, A. F., and Frost, D. (2010). Serial and parallel implementations of model-based clustering via parsimonious Gaussian mixture models. *Computational Statistics & Data Analysis*, **54**(3), 711–723.

Meng, X.-L. and Rubin, D. B. (1993). Maximum likelihood estimation via the ECM algorithm: A general framework. *Biometrika*, **80**(2), 267–278.

Orbanz, P. and Buhmann, J. M. (2005). SAR images as mixtures of Gaussian mixtures. *IEEE International Conference on Image Processing*, **2**, 209–212.

Peel, D. and McLachlan, G. J. (2000). Robust mixture modelling using the t distribution. *Statistics and Computing*, **10**(4), 339–348.

Schwarz, G. (1978). Estimating the dimension of a model. *The Annals of Statistics*, **6**(2), 461–464.

R Core Team (2013). R*: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Silverman, B. (1981). Using kernel density estimates to investigate multimodality. *Journal of the Royal Statistical Society, Series B: Methodological*, **43**, 97–99.

Subedi, S., Punzo, A., Ingrassia, S., and McNicholas, P. D. (2013). Clustering and classification via cluster-weighted factor analyzers. *Advances in Data Analysis and Classification*, **7**(1), 5–40.

Titterington, D. M., Smith, A. F. M., and Makov, U. E. (1985). *Statistical Analysis of Finite Mixture Distributions*. John Wiley & Sons, New York.