Available online at www.sciencedirect.com
ScienceDirect
Computational Statistics & Data Analysis 51 (2007) 2573–2585
COMPUTATIONAL STATISTICS & DATA ANALYSIS
www.elsevier.com/locate/csda

A mixture of mixture models for a classification problem: The unity measure error

Marco Di Zio[a,*], Ugo Guarnera[a], Roberto Rocci[b]

[a] *Istituto Nazionale di Statistica, via Cesare Balbo 16, 00184 Roma, Italy*
[b] *Università di Tor Vergata, via Columbia 2, 00133 Roma, Italy*

Received 19 May 2005; received in revised form 5 January 2006; accepted 5 January 2006
Available online 20 January 2006

Abstract

A mixture of Gaussian mixture models is proposed to deal with the identification of survey respondents providing values in a wrong unity measure. The "two-level" mixture model allows effective classification in a non-normal setting. The natural constraints of the problem make the model identifiable. The effectiveness of the proposal is shown by simulation studies and an application to the *1997 Italian Labour Cost Survey*.

Keywords: Editing; Systematic error; Mixture models; EM algorithm

1. Introduction

The assumption of data from a Gaussian distribution is sometimes too restrictive. Different techniques can be used to relax this hypothesis. Among them, the use of finite mixture models is interesting. Marron and Wand (1992) show that a finite Gaussian mixture can reasonably approximate a wide class of probability distributions. As far as the classification problems are concerned, it can be useful to consider groups as components of a finite mixture and, within each group in turn, to model data as a Gaussian mixture, thus obtaining a mixture of mixture models. The idea of a two-level mixture model is appealing because it allows to model non-Gaussian distribution within groups, i.e., the different components of the finite mixture (first level). The drawback of this approach is that identifiability problems may arise. Willse and Boik (1999) discuss this issue underlining that, under the imposition of appropriate constraints, the mixture of mixtures model is identifiable. Hastie and Tibshirani (1996) use a two-level mixture model for discriminant analysis. However, their approach has not any identifiability problem because membership function is known. An important application field where constraints may be imposed, making the mixture of mixture models identifiable, is that illustrated in Di Zio et al. (2005). The problem is the localization of unity measure error in data, i.e., the identification of survey respondents that provide values in a wrong unity measure. It is typical of the Official Statistics production, and it is generally treated in the data editing phase. This phase consists of localizing non-sampling errors in data (*editing*) and treating them, often substituting each value classified as erroneous with a more plausible one (*imputation*). Data editing is important both in terms of data quality and survey cost. Thus, techniques introduced to clean data are essentially

* Corresponding author. Tel.: +39 06 4673 2871; fax: +39 06 4673 2955.
E-mail address: dizio@istat.it (M. Di Zio).

0167-9473/$ - see front matter © 2006 Elsevier B.V. All rights reserved.
doi:10.1016/j.csda.2006.01.001

required to balance the trade off between these aspects (Granquist, 1995). Recently, with the advances in computers capabilities, the automatic editing approach, based on the Fellegi-Holt paradigm (Fellegi and Holt, 1976; De Waal, 2003), has increased its popularity. While this approach is appropriate for dealing with random errors, it requires data free of systematic errors. A particular systematic error, that frequently appears in surveys collecting numerical data, is the unity measure error causing the "true" value to be multiplied by a constant factor (e.g., 100 or 1000). This is due to the misunderstanding, by some respondents, of the unity measure, e.g., a respondent is request to report the amount of money in thousands but he expresses it (erroneously) in millions. This error highly affects both data accuracy (bias) and editing and imputation costs. In fact, all the automatic data editing process cannot be performed satisfactorily if this error is not removed preliminarily. In the National Statistical Institutes, this error is generally treated through ad hoc poorly automated procedures, using mainly graphical analysis and ratio edits, i.e., bounds on ratios between pairs of variables. The limit is both in terms of quality and cost. Quality is limited by the fact that the traditional approaches may take into account no more than a pairwise relationship between variables. Costs are essentially influenced by the fact that for each survey a new ad hoc procedure must be set up. To overcome these limits, Di Zio et al. (2005) defined the unity measure error as a clustering problem, and proposed an approach based on mixture modelling. They assume that, in the log scale, error-free data follow a Gaussian distribution. Although they show that this approach performs quite satisfactorily also in the presence of non-Gaussian data, its behaviour is expected to be poorer when data are far from the Gaussianity and clusters highly overlap each other. In this paper, the method proposed in Di Zio et al. (2005) is generalized by modelling the distribution of error-free data through a mixture of Gaussians. The resulting model becomes a mixture of Gaussian mixtures, hereafter mixture of mixtures unity measure error (UME).

The plan of the paper is the following. In Section 2, the model is formalized. In Section 3, the identifiability issue is discussed. In Section 4, an EM algorithm is introduced to compute the maximum likelihood estimates of model parameters. Effectiveness of the proposal is shown in Section 5 through simulation experiments, and in Section 6 by an application to a subset of real data from the *1997 Italian Labour Cost Survey* (LCS). Some final conclusions are drawn in Section 7.

2. The model

According to the approach of Di Zio et al. (2005), error-free data are considered as independent realizations of a random J-vector $\tilde{\mathbf{X}} = \left(\tilde{X}_1, \ldots, \tilde{X}_J\right)'$ with probability density function (p.d.f.) $\tilde{g}_0(\mathbf{x})$. The unity measure error acts on each variable $\tilde{X}_j$ $(j = 1, \ldots, J)$ through the transformation $\tilde{X}_j \to \tilde{c}\,\tilde{X}_j$ where $\tilde{c}$ is a constant factor (the generalization to the case of different constant factors for different variables is straightforward). In presence of non-negative-valued variables (a typical situation in economic surveys), it may be useful to work in the logarithmic scale. Actually, if we let $\mathbf{X} = \log\left(\tilde{\mathbf{X}}\right)$, and denote by $g_0(\mathbf{x})$ the p.d.f. associated with $\mathbf{X}$, the unity measure error can be represented through the transformation $X_j \to X_j + c$ where $c = \log(\tilde{c})$. For each subset of indices $l = \{j_1, \ldots, j_k\} \subseteq \{1, \ldots, J\}$, the observations affected by a unity measure error in the variables $\mathbf{X}_l = \left(X_{j_1}, \ldots, X_{j_k}\right)'$ define a cluster l similar in shape to the cluster of error-free units, but with a different location. More precisely, the units of cluster l can be thought of as generated by the density $g_l(\mathbf{x}) = g_0(\mathbf{x} - \mathbf{c}_l)$ where $\mathbf{c}_l$ is a vector whose components c_{lj} $(j = 1, \ldots, J)$ are equal to c if $j \in l$, and zero otherwise. In this framework, data can be modelled through the mixture density

$$f(\mathbf{x}) = \sum_{l=1}^{L} p_l g_l(\mathbf{x}), \tag{1}$$

where L is the number of the distinct error patterns. For each observation $\mathbf{x}_i$ $(i = 1, \ldots, n)$, the probability of belonging to a particular cluster can be computed by

$$\Pr(l|\mathbf{x}_i) = \frac{p_l g_l(\mathbf{x}_i)}{\sum_{l=1}^{L} p_l g_l(\mathbf{x}_i)}. \tag{2}$$

Thus, if the probability $\Pr(l|\mathbf{x}_i)$ can be estimated for all the groups l, the observation $\mathbf{x}_i$ can be assigned to the cluster (i.e., error pattern) with the highest estimated probability.

In Di Zio et al. (2005) the density g_0 of the error-free data is taken as J-variate normal with parameters $(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, so that each component g_l is obtained from g_0 simply by substituting $\boldsymbol{\mu}$ with $\boldsymbol{\mu} + \mathbf{c}_l$. In the proposed approach, the normality assumption is no longer required. This is accomplished by assuming that each density g_l is, in turn, expressed as a mixture of M Gaussians

$$g_l(\mathbf{x}) = \sum_{m=1}^{M} q_m h\left(\mathbf{x} - \mathbf{c}_l; \boldsymbol{\mu}_m, \boldsymbol{\Sigma}_m\right), \tag{3}$$

where $h\left(\cdot; \boldsymbol{\mu}_m, \boldsymbol{\Sigma}_m\right)$ is a J-variate normal density with mean vector $\boldsymbol{\mu}_m$ and covariance matrix $\boldsymbol{\Sigma}_m$. From Eqs. (1) and (3), it follows that the density of $\mathbf{X}$ can be written as

$$f(\mathbf{x}) = \sum_{l=1}^{L}\sum_{m=1}^{M} p_l q_m h\left(\mathbf{x} - \mathbf{c}_l; \boldsymbol{\mu}_m, \boldsymbol{\Sigma}_m\right) = \sum_{l=1}^{L}\sum_{m=1}^{M} p_l q_m h\left(\mathbf{x}; \boldsymbol{\mu}_m + \mathbf{c}_l, \boldsymbol{\Sigma}_m\right), \tag{4}$$

where

$$h\left(\mathbf{x}; \boldsymbol{\mu}_m + \mathbf{c}_l, \boldsymbol{\Sigma}_m\right) = (2\pi)^{-J/2}|\boldsymbol{\Sigma}_m|^{-1/2} \times \exp\left[-\tfrac{1}{2}\left(\mathbf{x} - \boldsymbol{\mu}_m - \mathbf{c}_l\right)'\boldsymbol{\Sigma}_m^{-1}\left(\mathbf{x} - \boldsymbol{\mu}_m - \mathbf{c}_l\right)\right].$$

Formula (4) is the density of a mixture of $L \times M$ Gaussian distributions with mean vectors suitably constrained, and global mixing proportions given by the products of the first- and second-level mixing proportions. The multiplicative structure of the mixing proportions is a consequence of the nature of the UME problem. In fact, apart the location parameters, the second-level mixture is the same for all the first-level groups. Note that further constraints can be introduced by assuming that all the M components have the same covariance matrix, $\boldsymbol{\Sigma}_m = \boldsymbol{\Sigma}$ for each $m = 1, \ldots, M$ (homoscedastic model). Constraints have an impact also on the number of parameters to be estimated. Actually, only one second-level mixture has to be estimated, (i.e., M mixing proportions q_m, M mean vectors $\boldsymbol{\mu}_m$, and M covariance matrices $\boldsymbol{\Sigma}_m$). Nevertheless, the number of error patterns and hence the number of first-level mixing proportions p_l increases exponentially as the number p of variables affected by UME increases. This aspect limits the applicability of the model when the number of variables is high with respect to the number of observations. This forces the researcher to work on subsets of variables separately rather than on all the variables simultaneously. The double level of mixtures reflects the two goals of this method: the first represents the classification, the second the density estimation. This approach has the advantage to be mathematically tractable and, at the same time, quite flexible to deal with data far from normality. A discussion on mixture of mixtures modelling is in McLachlan and Peel (2000).

3. Identifiability

In this section, the identifiability of the model (4) is investigated. We remind that, if Ψ is the parameter space and $\Im = \{f(\mathbf{x}; \psi), \psi \in \Psi\}$ a parametric family of probability distributions over Ψ, then $\Im$ is said to be identifiable if the mapping $\psi \longrightarrow f(\cdot; \psi)$ is a one to one map of Ψ onto $\Im$. In the context of finite mixture models, uniqueness of representation is required only up to relabelling of group indices. Thus, if we let $\boldsymbol{\pi} = (\pi_1, \ldots, \pi_K)'$, and $\boldsymbol{\Theta} = \{\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_K\}$, identifiability of the mixture model $f(\mathbf{x}; \boldsymbol{\pi}, \boldsymbol{\Theta}) = \sum_{k=1}^{K} \pi_k g(\mathbf{x}; \boldsymbol{\theta}_k)$ means that for any p.d.f. $\tilde{f}\left(\mathbf{x}; \tilde{\boldsymbol{\pi}}, \tilde{\boldsymbol{\theta}}\right) = \sum_{k=1}^{\tilde{K}} \tilde{\pi}_k g\left(\mathbf{x}; \tilde{\boldsymbol{\theta}}_k\right)$, the equality $\tilde{f} \equiv f$ implies $\tilde{K} = K$ and the existence of a component relabelling such that $\tilde{\pi}_k = \pi_k$ and $\tilde{\boldsymbol{\theta}}_k = \boldsymbol{\theta}_k$, $k = 1, \ldots, K$. Identifiability has been proved for some important class of distribution families such as gamma or multivariate Gaussian (see Teicher, 1963; Yakowitz and Spragins, 1968). In absence of any constraint, it is easy to show that in general a mixture of mixtures is not identifiable. This is essentially due to the possibility of interchanging component labels between the two levels of the model. In our case, the particular structure of the UME model does not allow such interchange. In particular it will be shown that the model in Eq. (4) is identifiable provided that all of the Gaussian densities $h\left(\mathbf{x}; \boldsymbol{\mu}_m + \mathbf{c}_l, \boldsymbol{\Sigma}_m\right)$ are distinct. Henceforth, $\mathbf{a} \succeq \mathbf{b}$ [$\mathbf{a} \succ \mathbf{b}$] denotes that the components of the vector

$\mathbf{a} - \mathbf{b}$ are all non-negative (but at least one positive), and $\|\cdot\|_1$ denotes the ℓ_1 norm, i.e., $\|\mathbf{a}\|_1 = \sum_{j=1}^{J} |a_j|$ where $\mathbf{a} = (a_1, a_2, \ldots, a_J)'$.

Result 1. Let $\mathbf{N} = \{\boldsymbol{\eta}_{lm}; l = 1, \ldots, L, m = 1, \ldots, M\}$ be a finite set of distinct vectors such that $\boldsymbol{\eta}_{lm} = \mathbf{a}_l + \mathbf{b}_m$, where $\mathbf{A} = \{\mathbf{a}_l; l = 1, \ldots, L\}$ and $\mathbf{B} = \{\mathbf{b}_m; m = 1, \ldots, M\}$ are sets of vectors such that $\mathbf{a}_l - \mathbf{a}_1 \succeq \mathbf{0}$, $(l = 1, \ldots, L)$. If there exists a relabelling of the elements of $\mathbf{N} = \{\tilde{\boldsymbol{\eta}}_{st}; s = 1, \ldots, L, t = 1, \ldots, M\}$ and a set $\tilde{\mathbf{B}} = \left\{\tilde{\mathbf{b}}_t; t = 1, \ldots, M\right\}$, such that it is possible to write $\tilde{\boldsymbol{\eta}}_{st} = \mathbf{a}_s + \tilde{\mathbf{b}}_t$, then $\tilde{\mathbf{B}} = \mathbf{B}$ and the equality $\boldsymbol{\eta}_{lm} = \tilde{\boldsymbol{\eta}}_{st}$, i.e.,

$$\mathbf{a}_l + \mathbf{b}_m = \mathbf{a}_s + \tilde{\mathbf{b}}_t, \tag{5}$$

implies $l = s$ and $\mathbf{b}_m = \tilde{\mathbf{b}}_t$.

Proof. For the proof of the statement it is sufficient to show that $\mathbf{B} = \tilde{\mathbf{B}}$. In this case we can find an invertible function $r(\cdot)$ such that $\tilde{\mathbf{b}}_t = \mathbf{b}_{r(t)}$. In fact, the elements of $\mathbf{B}$ and $\tilde{\mathbf{B}}$ are distinct since the elements of $\mathbf{N}$ are. The equality $\boldsymbol{\eta}_{lm} = \tilde{\boldsymbol{\eta}}_{st}$ implies

$$\boldsymbol{\eta}_{lm} = \tilde{\boldsymbol{\eta}}_{st} = \mathbf{a}_s + \tilde{\mathbf{b}}_t = \mathbf{a}_s + \mathbf{b}_{r(t)} = \boldsymbol{\eta}_{sr(t)}. \tag{6}$$

By the distinctness of the elements of $\mathbf{N}$, it follows that $(l, m) = (s, r(t))$, i.e., $l = s$ and $\mathbf{b}_m = \tilde{\mathbf{b}}_t$.

In order to prove that $\mathbf{B} = \tilde{\mathbf{B}}$, we consider a vector $\boldsymbol{\eta}_{\min}$ such that $\boldsymbol{\eta} - \boldsymbol{\eta}_{\min} \succeq \mathbf{0}$ for each $\boldsymbol{\eta} \in \mathbf{N}$. For instance, by using the following notation for the vectors $\boldsymbol{\eta}_{\min} = \left(\eta_{1;\min}, \ldots, \eta_{J;\min}\right)$ and $\boldsymbol{\eta}_{lm} = \left(\eta_{1;lm}, \ldots, \eta_{J;lm}\right)$, $\boldsymbol{\eta}_{\min}$ can be chosen as the vector whose components are defined by

$$\eta_{j;\min} = \min_{lm} \eta_{j;lm} \; j = 1, \ldots, J.$$

Let $\boldsymbol{\eta}_{l'm'}$ be the nearest vector to $\boldsymbol{\eta}_{\min}$ in $\mathbf{N}$, that is,

$$\boldsymbol{\eta}_{l'm'} = \operatorname*{argmin}_{\boldsymbol{\eta} \in \mathbf{N}} \|\boldsymbol{\eta} - \boldsymbol{\eta}_{\min}\|_1. \tag{7}$$

If there are two or more vectors fulfilling Eq. (7), we choose the one with the lowest second index. Noting that for any two vectors $\mathbf{a}, \mathbf{b} \succeq \mathbf{0}$ we have $\|\mathbf{a} + \mathbf{b}\|_1 = \|\mathbf{a}\|_1 + \|\mathbf{b}\|_1$, and also that $\boldsymbol{\eta}_{1m'} - \boldsymbol{\eta}_{\min} \succeq \mathbf{0}$, it follows

$$\begin{aligned} \|\boldsymbol{\eta}_{l'm'} - \boldsymbol{\eta}_{\min}\|_1 &= \|\mathbf{a}_{l'} - \mathbf{a}_1 + \mathbf{a}_1 + \mathbf{b}_{m'} - \boldsymbol{\eta}_{\min}\|_1 \\ &= \|\mathbf{a}_{l'} - \mathbf{a}_1\|_1 + \|\boldsymbol{\eta}_{1m'} - \boldsymbol{\eta}_{\min}\|_1 \\ &\geqslant \|\boldsymbol{\eta}_{1m'} - \boldsymbol{\eta}_{\min}\|_1, \end{aligned} \tag{8}$$

where the equality holds if and only if $\|\mathbf{a}_{l'} - \mathbf{a}_1\|_1 = 0$. By construction of $\boldsymbol{\eta}_{l'm'}$, we conclude that $l' = 1$.

Let $\tilde{\boldsymbol{\eta}}_{s't'}$ be $\boldsymbol{\eta}_{1m'}$ relabelled. With the same arguments, it is possible to show that $s' = 1$ and

$$\boldsymbol{\eta}_{1m'} = \tilde{\boldsymbol{\eta}}_{1t'} \Leftrightarrow \mathbf{b}_{m'} = \tilde{\mathbf{b}}_{t'}.$$

In this way, it is proved that $\mathbf{B}$ and $\tilde{\mathbf{B}}$ must have at least one element in common. The procedure can be repeated by excluding from $\mathbf{N}$ the elements $\boldsymbol{\eta}_{lm'}$ for $l = 1, 2, \ldots, L$, thus showing that $\mathbf{B}$ and $\tilde{\mathbf{B}}$ must have another element in common, say $\mathbf{b}_{m''} = \tilde{\mathbf{b}}_{t''}$. Of course, $\mathbf{b}_{m''} \neq \mathbf{b}_{m'}$, otherwise $\boldsymbol{\eta}_{1m''} = \boldsymbol{\eta}_{1m'}$ and this is not possible because the elements of $\mathbf{N}$ are all distinct. It follows that $\mathbf{B}$ and $\tilde{\mathbf{B}}$ must have at least two elements in common. By iterating the procedure $M - 2$ times again, it is proved that $\mathbf{B} = \tilde{\mathbf{B}}$. □

Result 1 allows to give sufficient conditions for the identification of model in Formula (4).

Result 2. Let

$$f(\mathbf{x}) = \sum_{l=1}^{L} \sum_{m=1}^{M} p_l q_m h\left(\mathbf{x}; \mathbf{c}_l + \boldsymbol{\mu}_m, \boldsymbol{\Sigma}_m\right),$$

and

$$f(\mathbf{x}) = \sum_{s=1}^{L} \sum_{t=1}^{\tilde{M}} \tilde{p}_s \tilde{q}_t h\left(\mathbf{x}; \mathbf{c}_s + \tilde{\boldsymbol{\mu}}_t, \tilde{\boldsymbol{\Sigma}}_t\right)$$

be two different parametrizations of the same mixture of mixtures UME model. Furthermore, let us suppose that $\mathbf{c}_l - \mathbf{c}_1 \geq \mathbf{0}$, $l = 1, \ldots, L$.

If $(l, m) \neq (l_1, m_1)$ implies

$$\left\|\left(\mathbf{c}_l + \boldsymbol{\mu}_m\right) - \left(\mathbf{c}_{l_1} + \boldsymbol{\mu}_{m_1}\right)\right\|_2^2 + \left\|\boldsymbol{\Sigma}_m - \boldsymbol{\Sigma}_{m_1}\right\|_2^2 \neq 0, \tag{9}$$

where $\|\cdot\|_2$ is the Froebenius norm, then there exists a relabelling such that

$$p_l = \tilde{p}_l, \quad q_m = \tilde{q}_m, \quad \boldsymbol{\mu}_m = \tilde{\boldsymbol{\mu}}_m, \quad \boldsymbol{\Sigma}_m = \tilde{\boldsymbol{\Sigma}}_m.$$

Proof. Let us denote

$$\mathbf{a}_l = \left(\mathbf{c}_l', \mathbf{0}'\right)', \quad \mathbf{b}_m = \left(\boldsymbol{\mu}_m', vec(\boldsymbol{\Sigma}_m)'\right)'$$

and

$$\tilde{\mathbf{b}}_t = \left(\tilde{\boldsymbol{\mu}}_t', vec\left(\tilde{\boldsymbol{\Sigma}}_t\right)'\right)',$$

where the null vector $\mathbf{0}$ has the same number of components as $vec\,(\boldsymbol{\Sigma}_m)$. Let us consider the $L \times M$ vectors

$$\boldsymbol{\eta}_{lm} = \mathbf{a}_l + \mathbf{b}_m$$

which are distinct because of condition (9), and the $L \times \tilde{M}$ vectors

$$\tilde{\boldsymbol{\eta}}_{st} = \mathbf{a}_s + \tilde{\mathbf{b}}_t.$$

The identifiability of Gaussian mixture guarantees that $L \times M = L \times \tilde{M}$, i.e., $M = \tilde{M}$, and also that for each pair (l, m) there exists a pair (s, t) such that

$$p_l q_m = \tilde{p}_s \tilde{q}_t, \quad \boldsymbol{\eta}_{lm} = \tilde{\boldsymbol{\eta}}_{st}.$$

By Result 1, the second of the previous equalities implies $l = s$ and $\mathbf{b}_m = \tilde{\mathbf{b}}_t$. Hence, by a suitable relabelling,

$$p_l q_m = \tilde{p}_l \tilde{q}_m, \quad \boldsymbol{\mu}_m = \tilde{\boldsymbol{\mu}}_m, \quad \boldsymbol{\Sigma}_m = \tilde{\boldsymbol{\Sigma}}_m$$

which implies

$$p_l = \tilde{p}_l, \quad q_m = \tilde{q}_m$$

by the uniqueness of the independence model. □

A sufficient condition for (9) to be fulfilled is that all of the covariance matrices $\boldsymbol{\Sigma}_m$ $(m = 1, \ldots, M)$ are distinct. It is worth mentioning this case because, in situations where the distribution of the error-free data is symmetric but not normal, it can be managed by using a mixture of Gaussians having the same location but different covariance matrices. The identifiability of mixture of mixtures UME model has been proved under the assumption that the known vectors $\mathbf{c}_1, \mathbf{c}_2, \ldots, \mathbf{c}_L$ are such that $\mathbf{c}_l - \mathbf{c}_1$ has non-negative components for each l. In the present context, where the vectors $\mathbf{c}_l$ are associated with unity measure error patterns, this assumption seems quite reasonable. For instance, for $K = 2$, the error patterns corresponding to the vectors $\mathbf{c}_1 = (-\log(1000), -\log(1000))'$, $\mathbf{c}_2 = (\log(1000), 0)'$, $\mathbf{c}_3 = (0, \log(1000))'$, $\mathbf{c}_4 = (\log(1000), \log(1000))'$, $\mathbf{c}_5 = (-\log(1000), 0)'$, $\mathbf{c}_6 = (0, -\log(1000))'$, $\mathbf{c}_7 = (0, 0)'$, satisfy the sufficient conditions for the identifiability. Nevertheless, if we exclude the error pattern given by the vector $\mathbf{c}_1$, then the sufficient conditions are no longer fulfilled. However, the model is still identified since the sufficient conditions can be recovered by multiplying the variables by -1. The critical situation is when neither $\mathbf{c}_1$ nor $\mathbf{c}_4$ are present.

This situation seems to be quite unrealistic in that it would correspond to the possibility a specific unity measure error occurs on each of the two variables but not on both simultaneously. An important remark is that Results 1 and 2 are based on the distinctness of all the Gaussian densities appearing in the model definition. In the following example, the role played by the distinctness of the mixture densities in ensuring identifiability is shown. Let us consider a homoscedastic mixture of mixtures model with density

$$f(\mathbf{x}) = p\left[qh\left(\mathbf{x};\boldsymbol{\mu}_1,\boldsymbol{\Sigma}\right)+(1-q)h\left(\mathbf{x};\boldsymbol{\mu}_2,\boldsymbol{\Sigma}\right)\right] + (1-p)\left[qh\left(\mathbf{x};\boldsymbol{\mu}_1+\mathbf{c},\boldsymbol{\Sigma}\right)+(1-q)h\left(\mathbf{x};\boldsymbol{\mu}_2+\mathbf{c},\boldsymbol{\Sigma}\right)\right],$$

with $0\leqslant p\leqslant 1$, $0\leqslant q\leqslant 1$ and $\mathbf{c}$ a known vector, and let us suppose that $\boldsymbol{\mu}_2=\boldsymbol{\mu}_1+\mathbf{c}$. The second density of the first group, and the first density of the second are the same. Thus, the hypothesis of distinctness is not fulfilled. In fact, this model can be viewed as a homoscedastic Gaussian mixture of three components with mixing proportions $\pi_1=pq$, $\pi_2=p(1-q)+(1-p)q$, $\pi_3=(1-p)(1-q)$, and mean vectors $\boldsymbol{\mu}_1$, $\boldsymbol{\mu}_1+\mathbf{c}$, $\boldsymbol{\mu}_1+2\mathbf{c}$. Since the weights π_1, π_2, π_3 are symmetric functions of p and q, the same model can be obtained by interchanging p and q, thus the model is not identifiable.

4. Maximum likelihood estimates

By assuming the independence of observations, we can write the log-likelihood of the whole sample as

$$\ell(\boldsymbol{\vartheta}) = \sum_{i=1}^{n}\log f\left(\mathbf{x}_i\right) = \sum_{i=1}^{n}\log\left\{\sum_{l=1}^{L}\sum_{m=1}^{M}p_l q_m h_{ilm}\right\}, \tag{10}$$

where $\boldsymbol{\vartheta}$ is the whole set of parameters to be estimated and $h_{ilm}=h\left(\mathbf{x}_i;\boldsymbol{\mu}_m+\mathbf{c}_l,\boldsymbol{\Sigma}_m\right)$. To compute the maximum likelihood estimates of model parameters, we note that (see Hathaway, 1986) the maximization of (10) is equivalent to the maximization of the "fuzzy" function

$$\ell_f(\boldsymbol{\vartheta}) = \sum_{i=1}^{n}\sum_{l=1}^{L}\sum_{m=1}^{M}u_{ilm}\log\left(p_l q_m h_{ilm}\right) - \sum_{i=1}^{n}\sum_{l=1}^{L}\sum_{m=1}^{M}u_{ilm}\log\left(u_{ilm}\right), \tag{11}$$

where the u_{ilm}'s are non-negative and such that $\sum_{lm}u_{ilm}=1$ for $i=1,2,\ldots,n$. To maximize ℓ_f we adopt a coordinate ascent method, where in each step the objective function is maximized with respect to a subset of parameters given the current values of the others. In this way each parameter, or subset of parameters, is in turn updated and the algorithm increases the value of the objective function at each step. The algorithm stops when the function increment in a particular step is lower than a given threshold. The fundamental steps of our algorithm are the following:

(a) *Update of* u_{ilm}: It can be easily shown that (11) has a maximum with respect to the u's when

$$u_{ilm} = \frac{p_l q_m h_{ilm}}{\sum_{lm}p_l q_m h_{ilm}}. \tag{12}$$

(b) *Update of* p_l: By rewriting (11) as

$$\ell_f(\boldsymbol{\vartheta}) = \sum_{ilm}u_{ilm}\log\left(p_l\right)+const., \tag{13}$$

where *const.* indicates a term that does not depend on the p's, we deduce that (11) is maximized with respect to the p's when

$$p_l = \frac{1}{n}\sum_{im}u_{ilm}. \tag{14}$$

(c) *Update of* q_m: As in the previous step, it can be shown that (11) obtains a maximum with respect to the q's when

$$q_m = \frac{1}{n}\sum_{il}u_{ilm}. \tag{15}$$

(d) *Update of* $\boldsymbol{\mu}_m$: First, we rewrite (11) as

$$\ell_f(\boldsymbol{\vartheta}) = \sum_{ilm} u_{ilm} \left[-\frac{1}{2} \left(\mathbf{x}_i - \boldsymbol{\mu}_m - \mathbf{c}_l\right)' \boldsymbol{\Sigma}_m^{-1} \left(\mathbf{x}_i - \boldsymbol{\mu}_m - \mathbf{c}_l\right) \right] + const., \tag{16}$$

where *const*. indicates a term independent of the $\boldsymbol{\mu}$'s. Then it simply follows that

$$\boldsymbol{\mu}_m = \frac{1}{\sum_{il} u_{ilm}} \sum_{il} u_{ilm} \left(\mathbf{x}_i - \mathbf{c}_l\right). \tag{17}$$

(e) *Update of* $\boldsymbol{\Sigma}_m$: By rewriting (11) as

$$\ell_f(\boldsymbol{\vartheta}) = -\frac{1}{2} \sum_{ilm} u_{ilm} \left[\log \left(|\boldsymbol{\Sigma}_m| + \mathbf{d}_{ilm}' \boldsymbol{\Sigma}_m^{-1} \mathbf{d}_{ilm} \right) \right] + const., \tag{18}$$

where *const*. indicates a term independent of the $\boldsymbol{\Sigma}_m$'s and $\mathbf{d}_{ilm} = \mathbf{x}_i - \boldsymbol{\mu}_m - \mathbf{c}_l$, we deduce that the update of $\boldsymbol{\Sigma}_m$ is

$$\boldsymbol{\Sigma}_m = \frac{1}{\sum_{il} u_{ilm}} \sum_{il} u_{ilm} \mathbf{d}_{ilm} \mathbf{d}_{ilm}', \tag{19}$$

while in the homoscedastic case, i.e., $\boldsymbol{\Sigma} = \boldsymbol{\Sigma}_m$, we have

$$\boldsymbol{\Sigma} = \frac{1}{n} \sum_{ilm} u_{ilm} \mathbf{d}_{ilm} \mathbf{d}_{ilm}'. \tag{20}$$

By iterating the above described steps we obtain a monotone algorithm which can be easily shown to be of ECM type (Meng and Rubin, 1993).

In practical applications, it turns out that a crucial role is played by the choice of the starting points, as usual in the EM algorithms (see Biernacki et al., 2003). We developed an initialization strategy based on a constrained version of the *K*-means clustering technique that should approximate the model, i.e., the $L \times M$ centroids are constrained to be of the form $\boldsymbol{\mu}_m + \mathbf{c}_l$. More in detail, the algorithm consists of two distinct phases corresponding to the two levels of the mixture model. In the first phase, the first-level groups (i.e., those corresponding to error patterns) are determined subject to the constraint that their centroids are obtained from the centroid of error-free data by translation for the appropriate vector $\mathbf{c}_l$ $l = 1, \ldots, L$. This is accomplished through the following simple iterative scheme:

(1) use the overall sample-median vector as initial guess of the error-free data centroid $\mathbf{p}_0^0$;
(2) compute the centroid of group l by adding to $\mathbf{p}_0^0$ the corresponding translation vector: $\mathbf{p}_l^0 = \mathbf{p}_0^0 + \mathbf{c}_l$;
(3) assign each unit to the group with the nearest centroid;
(4) translate each unit in cluster l by the corresponding vector $-\mathbf{c}_l$;
(5) determine the new centroid $\mathbf{p}_0^1$ of the error-free units as the overall sample mean vector computed on the translated data.

Iterate steps 2–5 until no change in units assignments is made in two successive iterations. Once the data are assigned to first-level groups, each unit in cluster l is translated as in step (4) by $-\mathbf{c}_l$ and the second stage of the initialization procedure consists of an ordinary *K*-means algorithm corresponding to the specified number of second-level clusters. The sample mean vectors and variance–covariance matrices of the second-level clusters are then taken as initial values of the model parameters.

5. Simulation experiments

In order to compare the simple Gaussian mixture model and the mixture of homoscedastic and heteroscedastic Gaussian mixture models, three different groups of experiments have been carried out. The elements varying in the experiments are the generating probability distribution, the sample size, and the translation vectors. In the first group

Table 1
Results for the experiments where the data distribution is (1) bivariate normal distribution ($S0$), (2) bivariate T distribution with 1 d.f. ($T0$), and (3) bivariate skew-T distribution ($T1$), the translation constant is based on log(1000) and sample size $n = 1000$

Data distribution		$S0$	$T0$	$T1$
benchmark		985.18	975.65	972.57
mixt	Correct Class	985.08	973.20	954.37
	Freq BIC	100	0	0
homo2	Correct Class	985.14	971.35	952.32
	Freq BIC	0	0	0
homo3	Correct Class	984.92	970.00	957.98
	Freq BIC	0	2	0
homo4	Correct Class	984.76	968.65	958.23
	Freq BIC	0	6	3
homo5	Correct Class	983.90	967.20	958.214
	Freq BIC	0	20	32
homo6	Correct Class	983.16	964.72	956.76
	Freq BIC	0	72	65
hete2	Correct Class	985.02	975.21	968.22
	Freq BIC	100	0	0
hete3	Correct Class	985.02	975.09	969.43
	Freq BIC	0	63	0
hete4	Correct Class	984.92	974.53	969.39
	Freq BIC	0	25	57
hete5	Correct Class	984.18	973.93	969.76
	Freq BIC	0	10	27
hete6	Correct Class	983.12	973.72	969.72
	Freq BIC	0	2	16

of experiments, a sample of 1000 observations is drawn from a four component mixtures, respectively of: (1) bivariate normal distributions ($S0$), (2) bivariate T distributions with 1 d.f. ($T0$), and (3) bivariate skew-T distributions ($T1$) (see for details Azzalini and Capitanio, 2003). The location parameter is $(-2.5, -2.6)$ for all the distributions, the covariance matrix Σ^0 has components $\sigma^0_{11} = 3.040$, $\sigma^0_{12} = 2.698$ and $\sigma^0_{22} = 2.880$ for $S0$ and $T0$, while $T1$ is characterized by the dispersion matrix Ω with components $\omega_{11} = 7.600$, $\omega_{12} = 6.745$, $\omega_{22} = 7.200$ and shape parameter vector $\alpha = (20, 0)$. For the meaning of dispersion matrix and shape parameter see Azzalini and Capitanio (2003). The mixture components correspond to the four different translation vectors $(0, 0)'$, $(0, \log(1000))'$, $(\log(1000), 0)'$, $(\log(1000), \log(1000))'$, with mixing proportions $\pi_1 = 0.5$, $\pi_2 = 0.1$, $\pi_3 = 0.1$, $\pi_4 = 0.3$. For each sample, the simple Gaussian mixture, the homoscedastic and the heteroscedastic mixture of mixture models are estimated. The corresponding classification is computed also varying the number of the second-level mixture components. In order to choose among the models in the same family, the corresponding BIC is computed (see Keribin, 2000). As benchmark, the optimal classification, i.e., the classification obtained by using the true generating mixture distribution, is also computed. This process is iterated 100 times. To evaluate the performance of the different models, the following quantities are computed (Table 1): (a) the average number of correct classifications ('Correct Class'), and (b) the frequency that a certain model is chosen according to BIC ('Freq BIC'). The homoscedastic mixture of two-component mixture model is indicated with 'homo2', 'hete2' is the heteroscedastic mixture of two-component mixture model, and analogously the others. Note that the frequencies sum to 100 within the two groups of models, i.e., simple mixture plus homoscedastic, and heteroscedastic. The second group of experiments is the same as the first, apart that the sample size is 500. The results are shown in Table 2.

In the third group of experiments, samples of size 1000 are drawn from a bivariate skew-T distribution ($T2$) with the location parameter set to $(-2.5, -2.6)$, the dispersion matrix Ω has components $\omega_{11} = 30.40$, $\omega_{12} = 26.98$, $\omega_{22} = 28.80$, the shape parameter $\alpha = (10, 50)$, and the non-zero components of the translation vectors are log(100) instead of log(1000). This last experiment is performed to see the behaviour of such a modelling in a very unfavourable situation. A sample drawn from this distribution is depicted in Fig. 1. Different symbols (circles, triangles, crosses and squares) are used to indicate the four $T2$ generating distributions corresponding to the different error patterns.

Table 2
Results for the experiments where the data distribution is (1) bivariate normal distribution ($S0$), (2) bivariate T distribution with 1 d.f. ($T0$), and (3) bivariate skew-T distribution ($T1$), the translation constant is based on log(1000) and sample size $n = 500$

Data distribution		$S0$	$T0$	$T1$
benchmark		492.450	487.610	439.520
mixt	Correct Class	492.570	487.370	340.290
	Freq BIC	100	0	0
homo2	Correct Class	492.480	485.940	413.060
	Freq BIC	0	0	0
homo3	Correct Class	492.440	485.120	418.930
	Freq BIC	0	7	7
homo4	Correct Class	492.080	484.620	419.130
	Freq BIC	0	13	11
homo5	Correct Class	491.580	483.350	419.990
	Freq BIC	0	36	44
homo6	Correct Class	490.520	481.880	421.420
	Freq BIC	0	44	38
hete2	Correct Class	492.360	487.500	408.110
	Freq BIC	100	3	0
hete3	Correct Class	492.030	487.040	427.510
	Freq BIC	0	82	25
hete4	Correct Class	492.100	486.700	431.260
	Freq BIC	0	13	52
hete5	Correct Class	491.350	485.950	430.920
	Freq BIC	0	2	21
hete6	Correct Class	490.530	484.870	430.160
	Freq BIC	0	0	2

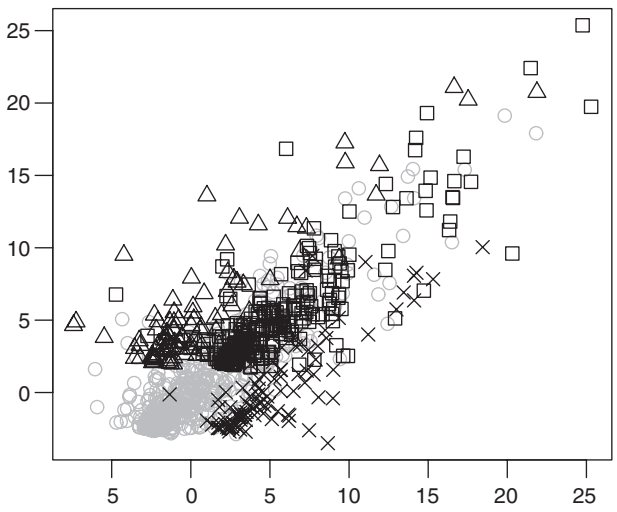

Fig. 1. A sample drawn from a four component mixture of $T2$ translated by log(100).

The situation is critical, as it is almost impossible to distinguish the four mixture components. The results of the simulation are reported in Table 3.

The experiments provide evidence to draw some conclusions. A simple normal mixture works satisfactorily in situation not so far from the normality assumption, like $S0$ and $T0$ (Tables 1 and 2). In these cases, the use of a more complex model does not cause any improvement. Moreover, heteroscedastic models seem to perform slightly better than homoscedastic ones.

Table 3
Results for the experiments where the data distribution is bivariate skew-T distribution ($T2$), the translation constant is based on log(100) and sample size $n = 1000$

Data distribution		$T2$
benchmark		779.660
mixt	Correct Class	596.110
	Freq BIC	0
homo2	Correct Class	629.920
	Freq BIC	0
homo3	Correct Class	701.240
	Freq BIC	1
homo4	Correct Class	709.770
	Freq BIC	14
homo5	Correct Class	720.660
	Freq BIC	19
homo6	Correct Class	726.840
	Freq BIC	66
hete2	Correct Class	718.480
	Freq BIC	0
hete3	Correct Class	757.200
	Freq BIC	0
hete4	Correct Class	765.600
	Freq BIC	31
hete5	Correct Class	768.540
	Freq BIC	45
hete6	Correct Class	769.230
	Freq BIC	24

When the generating probability distribution function is ($T1$) (skew and heavy tails), the gain obtained by the heteroscedastic model with respect to the simple normal mixture is sensible, and it approaches the optimal classification. Also the homoscedastic modelling behaves better than the simple normal mixture (a low gain in the case of 1000 observations experiment), but the gain is always less than that obtained by the heteroscedastic one.

In the last experiment, the most critical situation, there is a sensible gain by using the homoscedastic model with respect to the simple normal mixture, but the behaviour of the heteroscedastic mixture is much better, approaching the benchmark given by the optimal classification. A final consideration regards the choice of the number of components. For the heteroscedastic case, the results show that the BIC does not always address the best model. However, even when the mode of the frequency of choice of a certain model does not correspond to the best average classification, the absolute difference between the average number of correctly classified units is very low. The same can be stated for the homoscedastic model, apart the pathological behaviour in the case $T0$. Actually, the BIC suggests the six-component model, while the simple mixture has the best classification behaviour. Furthermore, this simple model is never chosen in the 100 experiments.

A further experiment is devoted to test the starting point method chosen for the experiments so far described. The EM algorithm is initialized by using a constrained version of K-means clustering algorithm, as described in Section 4. In order to evaluate this initialization method, two samples of 1000 observations from $T1$ and $T2$ are drawn. For each sample, the mixture of mixture models with four heteroscedastic components is estimated 100 times with the constrained K-means initialization. For all the estimates, the value reached by the likelihood is the same.

6. Application to real data

To illustrate the effectiveness and test the performance of our proposal, we carry out an experiment on a subset of the *1997 Italian* LCS. The LCS is a periodic sample survey that collects information on employment, hours worked, wages, salaries and labour cost on about 12,000 firms with more than 10 employees. The survey is subjected to a specific European Regulation requiring all the European Community Member States to collect every four years

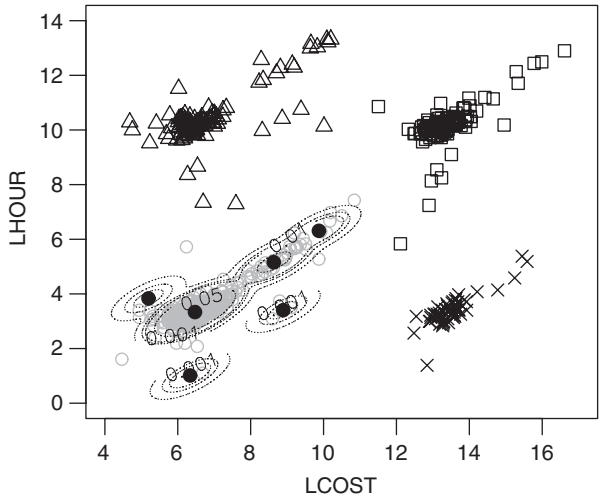

Fig. 2. Classification through homoscedastic mixture with $M = 6$ within cluster components.

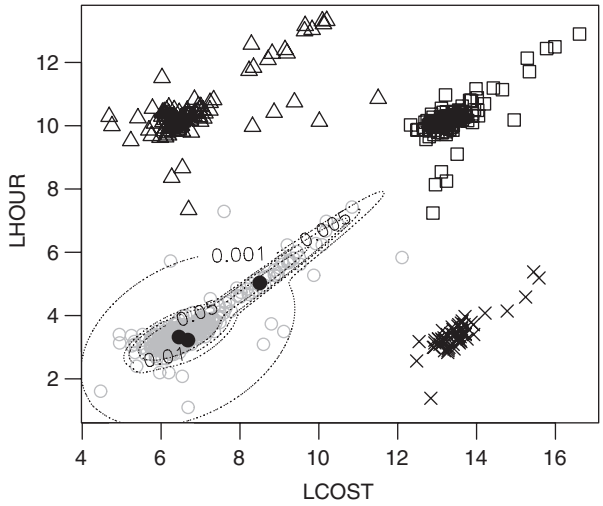

Fig. 3. Classification through a heteroscedastic mixture with $M = 3$ within cluster components.

detailed information about the labour cost and employment structure in some specific industries. Our data-set consists of 744 units that belong to the metallurgic economic activity sector. In particular, we analyse two main variables measuring the *Total Labour Cost* (LCOST) and *Total Hours Worked* (LHOUR). These variables are affected by the *1000-factor error*, since some respondents have expressed the LCOST in thousand of Italian Lira instead of millions, and similarly the hours have not been reported in thousand as requested. Details on the error profile and the impact of systematic error on data accuracy can be found in Cirianni et al. (2000).

The logarithmic transformation of the LCOST and LHOUR is taken, and the clusters associated with the four different error patterns are defined as follows: cluster1 = no errors, cluster2 = only LHOUR in error, cluster3 = only LCOST in error, cluster4 = both variables in error.

In order to classify firms according to their unity measure error pattern, we follow the approach described in the previous sections, modelling data through a homoscedastic and heteroscedastic Gaussian mixture of mixtures. The starting points for the EM algorithm are determined through the constrained K-means algorithm described in Section 4. Different experiments are performed by varying the number M of the components of the within-cluster mixture.

The optimal number of components is chosen according to the BIC criterion. It results that the appropriate choice for the homoscedastic model is $M = 6$, while for the heteroscedastic model is $M = 3$. The resulting classifications are reported in Figs. 2, and 3, respectively. The four estimated clusters are represented with different symbols (circles, triangles, crosses and squares) as in Fig. 1, and the contour plot (at levels 0.05, 0.01, 0.005, 0.001) of the estimated mixture of mixtures, together with their mean vectors (black dots) are reported. Note that, for the sake of simplicity, only the means of the cluster1 are reported, while the others can be easily obtained by the appropriate translation. The classification is almost the same with homoscedastic and heteroscedastic models. This is mainly due to the fact that the clusters are well separated and identifiable. It is worthwhile noting that, in the heteroscedastic model (Fig. 3), there are two mixture components that are located almost at the same point. The two components differ mainly for the estimated covariance matrix. As noted in Section 3, this kind of mixture is useful to approximate symmetric distribution with heavy tails. A stronger sensitivity to the model, in terms of classification performance, is expected in presence of clusters more overlapping each other, as noted in Section 5. The estimation of the mixture model parameters and the clustering has been carried out through a code developed in the R environment (R Development Core Team, 2004) available upon request.

7. Discussion

In this paper, a method to identify observations affected by unity measure errors is proposed. The problem is reinterpreted in a probabilistic clustering framework. The p.d.f. of the observations is modelled as a finite mixture where each component corresponds to a particular error pattern. The density of each component is, in turn, estimated by using a finite mixture of Gaussians in order to allow a more general setting. The resulting model, a 'mixture of mixtures', is proved to be identifiable under suitable conditions. A similar approach has been also used by Hastie and Tibshirani (1996) in the discriminant analysis context. The maximum likelihood estimates of model parameters are computed using an EM algorithm. An initialization strategy based on a constrained version of *K*-means algorithm is proposed. This technique seems to be appropriate in providing starting points for the EM. Indeed, simulation experiments (not reported here) show that classification of error patterns based only on *K*-means algorithm performs quite satisfactorily even though the number of units correctly classified is lower than that obtained using mixture model. These results can be explained noting that classification based on *K*-means algorithm is equivalent to classification based on a Gaussian mixture whose components have spherical covariance matrices (Gordon, 1999). Nevertheless, whenever within-group independence cannot be assumed, models corresponding to more complex association structures are expected to perform better. Two different settings, with equal covariance matrices at the second level (homoscedastic) and possibly different covariance matrices (heteroscedastic) are tested and compared through simulation studies. The approach is shown to be useful to deal with data distributed far from the Gaussianity. Results suggest a strategy based on the use of a constrained *K*-means algorithm for the inizialization step of the EM estimation phase, the use of heteroscedastic models, and the BIC as penalizing function for choosing the number of components of the mixture.

References

Azzalini, A., Capitanio, A., 2003. Distributions generated by perturbation of symmetry with emphasis on a multivariate skew-*t* distribution. J. Roy. Statist. Soc. Ser. B 65, 367–389.

Biernacki, C., Celeux, G., Govaert, G., 2003. Choosing starting values for the EM algorithm for getting the highest likelihood in multivariate Gaussian mixture models. Comput. Statist. Data Anal. 41, 561–575.

Cirianni, A., Di Zio, M., Luzi, O., Seeber, A.C., 2000. The new integrated data procedure for the Italian Labour Cost survey: measuring the effects on data of combined techniques. Proceedings of the International Conference on Establishment Surveys II (ICES), June 17–21, 2000, Buffalo, USA.

De Waal, T., 2003. Solving the error localization problem by means of vertex generation. Surv. Methodol. 1, 71–79.

Di Zio, M., Guarnera, U., Luzi, O., 2005. Editing systematic unity measure errors through mixture modelling. Surv. Methodol. 31, 53–63.

Fellegi, I.P., Holt, D., 1976. A systematic approach to edit and imputation. J. Amer. Statist. Assoc. 71, 17–35.

Gordon, A.D., 1999. Classification. second ed.. Chapman & Hall, London.

Granquist, L., 1995. Improving the traditional editing process. In: Cox, B.G., Binder, D.A., Chinappa, B.N., Christianson, A., Colledge, M.J., Kott, P.S. (Eds.), Business Survey Methods. Wiley, New York.

Hastie, T., Tibshirani, R., 1996. Discriminant analysis by Gaussian mixtures. J. Roy. Statist. Soc. Ser. B 58, 155–176.
Hathaway, R.J., 1986. Another interpretation of the EM algorithm for mixture distributions. Statist. Probab. Lett. 4, 53–56.
Keribin, C., 2000. Consistent estimation of the order of mixture models. Sankya The Indian J. Statist. 62, 49–66.
Marron, J.S., Wand, M.P., 1992. Exact mean integrated squared error. Ann. Statist. 20, 712–736.
McLachlan, G.J., Peel, D., 2000. Finite Mixture Models. Wiley, New York.
Meng, X.L., Rubin, D.B., 1993. Maximum likelihood estimation via the ECM algorithm: a general framework. Biometrika 80, 267–278.
R: A language and environment for statistical computing. R Foundation for Statistical Computing. Vienna, Austria, 2004.
Teicher, H., 1963. Identifiability of finite mixtures. Ann. Math. Statist. 34, 1265–1269.
Willse, A., Boik, R.J., 1999. Identifiable finite mixtures of location models for clustering mixed-mode data. Statist. Comput. 9, 111–121.
Yakowitz, S.J., Spragins, J.D., 1968. On the identifiability of finite mixtures. Ann. Math. Statist. 39, 209–214.