The joint role of trimming and constraints in robust estimation for mixtures of Gaussian factor analyzers

CrossMark

Luis Angel García-Escudero [a], Alfonso Gordaliza [a], Francesca Greselin [b,*],
Salvatore Ingrassia [c], Agustín Mayo-Iscar [a]

[a] *Department of Statistics and Operational Research and IMUVA, University of Valladolid, Spain*
[b] *Department of Statistics and Quantitative Methods, Milano-Bicocca University, Italy*
[c] *Department of Economics and Business, University of Catania, Italy*

ARTICLE INFO

ABSTRACT

Mixtures of Gaussian factors are powerful tools for modeling an unobserved heterogeneous population, offering – at the same time – dimension reduction and model-based clustering. The high prevalence of spurious solutions and the disturbing effects of outlying observations in maximum likelihood estimation may cause biased or misleading inferences. Restrictions for the component covariances are considered in order to avoid spurious solutions, and trimming is also adopted, to provide robustness against violations of normality assumptions of the underlying latent factors. A detailed AECM algorithm for this new approach is presented. Simulation results and an application to the AIS dataset show the aim and effectiveness of the proposed methodology.

© 2016 Elsevier B.V. All rights reserved.

1. Introduction and motivation

Factor analysis is an effective method of summarizing the variability between a number of correlated features, through a much smaller number of unobservable, hence named *latent*, factors. It originated from the consideration that, in many phenomena, several observed variables could be explained by a few unobserved ones. Under this approach, each single variable (among the p observed ones) is assumed to be a linear combination of d underlying common factors with an accompanying error term to account for that part of the variability which is unique to it (not in common with other variables). Ideally, d should be substantially smaller than p, to achieve parsimony.

Clearly, the effectiveness of this method is limited by its global linearity, as happens for principal components analysis. Hence, Ghahramani and Hilton (1997), Tipping and Bishop (1999) and McLachlan and Peel (2000a) solidly widened the applicability of these approaches by combining local models of Gaussian factors in the form of finite mixtures. The idea is to employ latent variables to perform dimensional reduction in each component, thus providing a statistical method which concurrently performs clustering and, within each cluster, local dimensionality reduction.

In the literature, error and factors are routinely assumed to have a Gaussian distribution because of their mathematical and computational tractability: however, statistical methods which ignore departure from normality may cause biased or misleading inference. Moreover, it is well known that maximum likelihood estimation for mixtures often leads to

---

* Corresponding author.

E-mail addresses: lagarcia@eio.uva.es (L.A. García-Escudero), alfonsog@eio.uva.es (A. Gordaliza), francesca.greselin@unimib.it (F. Greselin), s.ingrassia@unict.it (S. Ingrassia), agustin@med.uva.es (A. Mayo-Iscar).

ill-posed problems because of the unboundedness of the objective function to be maximized, which favors the appearance of non-interesting local maximizers and degenerate or *spurious* solutions.

The lack of robustness in mixture fitting arises whenever the sample contains a certain proportion of data that does not follow the underlying population model. Spurious solutions can even appear when ML estimation is applied to artificial data drawn from a given finite mixture model, i.e. without adding any kind of contamination. Hence, more robust estimation is needed. Many contributions in this sense can be found in the literature: from the Mclust model with a noise component in Fraley and Raftery (1998), mixtures of *t*-distributions in McLachlan and Peel (2000c), the trimmed likelihood mixture fitting method in Neykov et al. (2007), the trimmed ML estimation of contaminated mixtures in Gallegos and Ritter (2009), and the robust improper ML estimator introduced in Coretto and Hennig (2011), among many others. Some important applications in computer vision, pattern recognition, analysis of microarray gene expression data, or tomography (see, for example, Stewart (1999), Campbell et al. (1997), Bickel (2003) and Maitra (2001), respectively) suggest that more attention should be paid to robustness, because noise in the datasets may be frequent in all these fields.

Different types of constraints have been traditionally applied in Gaussian mixtures of factor analyzers, for instance, some authors propose taking a common (diagonal) error matrix (as for the Mixtures of Common Factor Analyzers, denoted by MCFA, in Baek et al., 2010) or imposing an isotropic error matrix (Bishop and Tipping, 1998). This strategy has proven to be effective in many cases, at the expenses of stronger distributional restrictions on the data. To avoid singularities and spurious solutions, under milder conditions, Greselin and Ingrassia (2015) recently proposed maximizing the likelihood by constraining the eigenvalues of the covariance matrices, following previous work of Ingrassia (2004) and going back to Hathaway (1985). Furthermore, mixtures of *t*-analyzers have been considered (see McLachlan and Bean, 2005; Baek and McLachlan, 2011; Steane et al., 2012; Lin et al., 2014b, and references therein) in an attempt to make the model less sensitive to outliers, but they, too, are not robust against very extreme outliers (Hennig, 2004).

The purpose of the present work is to introduce an estimating procedure for the mixture of Gaussian factors analyzers that can resist the effect of outliers and avoid spurious local maximizers. The proposed constraints can also be used to take into account prior information about the scatter parameters.

Trimming has been shown to be a simple, powerful, flexible and computationally feasible way to provide robustness in many different statistical frameworks. The basic idea behind trimming here is the removal of a small proportion α of observations whose values would be the most unlikely to occur if the fitted model were true. In this way, trimming avoids a small fraction of outlying observations exerting a harmful effect on the estimation. Incorporating constraints into the mixture estimation provides a well-posed setting for the mathematical problem and reduces the risk of incurring spurious solutions. Moreover, the formulation of the problem according to Section 3.1 allows the estimators to obtain the desired statistical properties, such as existence and consistency, as in García-Escudero et al. (2008).

The rest of the paper has been organized as follows. In Section 2, the notation is introduced and the main ideas about Gaussian Mixtures of Factor Analyzers (hereafter denoted by MFA) are summarized. Then, in Section 3 the trimmed likelihood for MFA is presented, and fairly extensive notes are provided concerning the EM algorithm, with incorporated trimming and constrained estimation. In Section 4, the performance of the new procedure is discussed, on the grounds of some numerical results obtained from simulated and real data. In particular, the bias and MSE of robustly estimated model parameters for different cases of data contamination, are compared using Monte Carlo experiments. The application to the Australian Institute of Sports dataset shows how classification and factor analysis can be developed using the new model. Section 5 contains concluding notes and provides ideas for further research.

2. Gaussian mixtures of factor analyzers

The density of the p-dimensional random variable $\mathbf{X}$ of interest is modeled as a mixture of G multivariate normal densities in some unknown proportions $\pi_1, \ldots, \pi_G$, whenever each data point is taken to be a realization of the following density function,

$$f(\mathbf{x}; \boldsymbol{\theta}) = \sum_{g=1}^{G} \pi_g \phi_p(\mathbf{x}; \boldsymbol{\mu}_g, \boldsymbol{\Sigma}_g) \tag{1}$$

where $\phi_p(\mathbf{x}; \boldsymbol{\mu}, \boldsymbol{\Sigma})$ denotes the p-variate normal density function with mean vector $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$. Here the vector $\boldsymbol{\theta} = \boldsymbol{\theta}_{GM}(p, G)$ of unknown parameters consists of the $(G-1)$ mixing proportions π_g, the Gp elements of the component means $\boldsymbol{\mu}_g$, and the $\frac{1}{2}Gp(p+1)$ distinct elements of the component-covariance matrices $\boldsymbol{\Sigma}_g$. MFA postulates a finite mixture of linear sub-models for the distribution of the full observation vector $\mathbf{X}$, given the (unobservable) factors $\mathbf{U}$. That is, MFA provides local dimensionality reduction by assuming that the distribution of the observation $\mathbf{X}_i$ can be given as

$$\mathbf{X}_i = \boldsymbol{\mu}_g + \boldsymbol{\Lambda}_g \mathbf{U}_{ig} + \mathbf{e}_{ig} \quad \text{with probability } \pi_g (g = 1, \ldots, G) \quad \text{for } i = 1, \ldots, n, \tag{2}$$

where $\boldsymbol{\Lambda}_g$ is a $p \times d$ matrix of *factor loadings*, the *factors* $\mathbf{U}_{1g}, \ldots, \mathbf{U}_{ng}$ are $\mathcal{N}(\mathbf{0}, \mathbf{I}_d)$ distributed independently of the *errors* $\mathbf{e}_{ig}$. The latter are independently $\mathcal{N}(\mathbf{0}, \boldsymbol{\Psi}_g)$ distributed, and $\boldsymbol{\Psi}_g$ is a $p \times p$ diagonal matrix $(g = 1, \ldots, G)$. The diagonality of $\boldsymbol{\Psi}_g$ is one of the key assumptions of factor analysis: the observed variables are independent given the factors. Note that the factor variables $\mathbf{U}_{ig}$ model correlations between the elements of $\mathbf{X}_i$, while the errors $\mathbf{e}_{ig}$ account for independent