CLADAG 2013

9 th SCIENTIFIC MEETING OF THE CLASSIFICATION AND DATA ANALYSIS GROUP OF THE ITALIAN STATISTICAL SOCIETY

September 18 - 20, 2013
University of Modena and Reggio Emilia
San Geminiano Complex - Modena, Italy

Book of Abstracts

Editors: Tommaso Minerva, Isabella Morlini, Francesco Palumbo
CLEUP
ISBN: 9788867871179

Patronage

Dipartimento di Economia Marco Biagi
Dipartimento di Comunicazione ed Economia
Dipartimento di Scienze Fisiche, Informatiche e Matematiche

Data driven EM constraints for mixtures of factor analyzers

Francesca Greselin and Salvatore Ingrassia

Abstract Mixtures of factor analyzers are becoming more and more popular in the area of model based clustering of high-dimensional data. In this paper we implement a data-driven methodology to maximize the likelihood function in a constrained parameter space, to overcome the well known issue of singularities and to reduce spurious maxima in the EM algorithm. Simulation results and applications to real data show that the problematic convergence of the EM, even more critical when dealing with factor analyzers, can be greatly improved.

Key words: Mixture of Factor Analyzers, Model-Based Clustering, Constrained EM algorithm.

1 Introduction and motivation

Finite mixture distributions, dating back to the seminal work of Newcomb and Pearson, have been receiving a growing interest in statistical modeling all along the last century. Along the lines of Ghahramani and Hilton (1997) we assume that the data have been generated by a linear factor model with latent variables modeled as Gaussian mixtures. Our purpose is to improve the performances of the EM algorithm, giving practical recipes to overcome some of its issues. Following Ingrassia (2004), in this paper we introduce and implement a procedure for the parameters estimation of mixtures of factor analyzers, which maximizes the likelihood function in a constrained parameter space, having no singularities and a reduced number of spurious local maxima. Within the Gaussian Mixture (GM) model-based approach to density estimation and clustering, the density of the d-dimensional random variable $\mathbf{X}$ of interest is modeled as a mixture of a number, say G, of multivariate normal densities in some unknown proportions $\pi_1, \ldots \pi_G$,

Francesca Greselin
Department of Statistics and Quantitative Methods
Milano-Bicocca University
Via Bicocca degli Arcimboldi 8 - 20126 Milano (Italy), e-mail: francesca.greselin@unimib.it

Salvatore Ingrassia
Department of Economics and Business
University of Catania
Corso Italia 55 - Catania (Italy), e-mail: s.ingrassia@unict.it

$$f(\mathbf{x};\theta)=\sum_{g=1}^{G}\pi_g\phi_d(\mathbf{x};\mu_g,\Sigma_g)$$

where $\phi_d(\mathbf{x};\mu,\Sigma)$ denotes the d-variate normal density function with mean μ and covariance matrix Σ. Then, we postulate a finite mixture of linear sub-models $\mathbf{X}_i=\mu_g+\Lambda_g\mathbf{U}_{ig}+\mathbf{e}_{ig}$ with probability π_g $(g=1,\ldots,G)$ for $i=1,\ldots,n$, for the distribution of the full observation vector $\mathbf{X}$, given the (unobservable) latent factors $\mathbf{U}$, where Λ_g is a $d\times q$ matrix of *factor loadings*, the *factors* $\mathbf{U}_{1g},\ldots,\mathbf{U}_{ng}$ are $\mathcal{N}(\mathbf{0},\mathbf{I}_q)$ distributed independently of the *errors* $\mathbf{e}_{ig}$, which are independently $\mathcal{N}(\mathbf{0},\Psi_g)$ distributed, and Ψ_g is a $d\times d$ diagonal matrix $(g=1,\ldots,G)$. We suppose that $q<d$, which means that q latent factors are jointly explaining the d observable features of the statistical units. Under these assumptions, $\Sigma_g=\Lambda_g\Lambda_g'+\Psi_g \quad (g=1,\ldots,G)$. The parameter vector $\theta_{MGFA}(d,q,G)$ now consists of the elements of the component means μ_g, the Λ_g, and the Ψ_g, along with the mixing proportions π_g $(g=1,\ldots,G-1)$.

2 The likelihood function and the EM algorithm for MGFA

In this section we summarize the main steps of the EM algorithm for mixtures of Factor analyzers, see e.g. McLachlan *et al*. (2003) for details. Let $\mathbf{x}_i$ $(i=1,\ldots,n)$ denotes the realization of $\mathbf{X}_i$. Then, the complete-data likelihood function for a sample $\underset{\sim}{\mathbf{X}}$ of size n can be written as

$$L_c(\theta;\underset{\sim}{\mathbf{X}})=\prod_{i=1}^{n}\prod_{g=1}^{G}\left[\phi_d\left(\mathbf{x}_i|\mathbf{u}_i;\mu_g,\Lambda_g,\Psi_g\right)\phi_q(\mathbf{u}_{ig})\pi_g\right]^{z_{ig}}. \tag{1}$$

Due to the factor structure of the model, we consider the alternating expectation-conditional maximization (AECM) algorithm, consisting of the iteration of two conditional maximizations, until convergence. There is one E-step and one CM-step, alternatively i) considering $\theta_1=\{\pi_g,\mu_g,g=1,\ldots,G\}$ where the missing data are the unobserved group labels $\underset{\sim}{\mathbf{Z}}=(\mathbf{z}_1',\ldots,\mathbf{z}_n')$ and ii) considering $\theta_2=\{(\Lambda_g,\Psi_g),g=1,\ldots,G\}$ where the missing data are the group labels $\mathbf{Z}$ and the unobserved latent factors $\mathbf{U}=(\mathbf{U}_{11},\ldots,\mathbf{U}_{nG})$. In the First Cycle, after updating the $z_{ig}^{(k+1)}$ in the E-step, the M-step provides new values for $\pi_g^{(k+1)},\mu_g^{(k+1)},n_g^{(k+1)}$. In the Second Cycle, after writing the complete data log-likelihood, some algebras lead to the following estimate of $\{(\Lambda_g,\Psi_g),g=1,\ldots,G\}$

$$\hat{\Lambda}_g=\mathbf{S}_g^{(k+1)}\gamma_g^{(k)\prime}[\Theta_g^{(k)}]^{-1}\qquad \hat{\Psi}_g=\operatorname{diag}\left\{\mathbf{S}_g^{(k+1)}-\hat{\Lambda}_g\gamma_g^{(k)}\mathbf{S}_g^{(k+1)}\right\},$$

where

$$\mathbf{S}_g^{(k+1)}=(1/n_g^{(k+1)})\sum_{i=1}^{n}z_{ig}^{(k+1)}(\mathbf{x}_i-\mu_g^{(k+1)})(\mathbf{x}_i-\mu_g^{(k+1)})'$$

$$\gamma_g^{(k)}=\Lambda_g^{(k)\prime}(\Lambda_g^{(k)}\Lambda_g^{(k)\prime}+\Psi_g^{(k)})^{-1}\quad\text{and}\quad\Theta_g^{(k)}=\mathbf{I}_q-\gamma_g^{(k)}\Lambda_g^{(k)}+\gamma_g^{(k)}S_g^{(k+1)}\gamma_g^{(k)\prime}.$$

3 Likelihood maximization in constrained parametric spaces

Along the lines of Ingrassia (2004) let us consider the constrained parameter space

$$\Theta_c = \{(\pi_1,\ldots,\pi_G,\mu_1,\ldots,\mu_G,\Sigma_1,\ldots,\Sigma_G) \in \mathbb{R}^{k[1+d+(d^2+d)/2]} : \\ \pi_g \geq 0, \pi_1+\cdots+\pi_G = 1, a \leq \lambda_{ig} \leq b, \quad g = 1,\ldots,G \;\; i = 1,\ldots,d\}. \quad (2)$$

Applying the eigenvalue decomposition to the square $d \times d$ matrix $\Lambda_g\Lambda_g'$ we can find Γ_g and Δ_g such that $\Lambda_g\Lambda_g' = \Gamma_g\Delta_g\Gamma_g'$ where Γ_g is the orthonormal matrix whose rows are the eigenvectors of $\Lambda_g\Lambda_g'$ and $\Delta_g = \text{diag}(\delta_{1g},\ldots,\delta_{dg})$ is the sorted diagonal matrix of the eigenvalues of $\Lambda_g\Lambda_g'$, i.e. $\delta_{1g} \geq \ldots \geq \delta_{qg} \geq 0$, and $\delta_{(q+1)g} = \cdots = \delta_{dg} = 0$. Further, applying now the singular value decomposition to Λ_g, we get $\Lambda_g = \mathbf{U}_g\mathbf{D}_g\mathbf{V}_g'$. This yields $\Lambda_g\Lambda_g' = (\mathbf{U}_g\mathbf{D}_g\mathbf{V}_g')(\mathbf{V}_g\mathbf{D}_g'\mathbf{U}_g') = \mathbf{U}_g\mathbf{D}_g\mathbf{D}_g'\mathbf{U}_g'$ hence $\text{diag}(\delta_{1g},\ldots,\delta_{qg}) = \text{diag}(d_{1g}^2,\ldots,d_{qg}^2)$. We can now modify the EM algorithm in such a way that the eigenvalues of the covariances Σ_g (for $g = 1,\ldots,G$) are confined into suitable ranges. To this aim we exploit the following inequalities

$$\lambda_{\min}(\Lambda_g\Lambda_g' + \Psi_g) \geq \lambda_{\min}(\Lambda_g\Lambda_g') + \lambda_{\min}(\Psi_g) \geq a$$
$$\lambda_{\max}(\Lambda_g\Lambda_g' + \Psi_g) \leq \lambda_{\max}(\Lambda_g\Lambda_g') + \lambda_{\max}(\Psi_g) \leq b$$

which enforce (2) when imposing the following constraints

$$d_{ig}^2 + \psi_{ig} \geq a \qquad i = 1,\ldots,d \quad (3)$$
$$d_{ig} \leq \sqrt{b - \psi_{ig}} \qquad i = 1,\ldots,q \quad (4)$$
$$\psi_{ig} \leq b \qquad i = q+1,\ldots,d \quad (5)$$

for $g = 1,\ldots,G$, where ψ_{ig} denotes the i-th diagonal entry of Ψ_g. In particular, we note that condition (3) reduces to $\psi_{ig} \geq a$ for $i = (q+1),\ldots,d$.

It is important to remark that the resulting EM algorithm is monotone, once the initial guess, say Σ_g^0, satisfies the constraints. Further, as shown in the case of gaussian mixtures in Ingrassia and Rocci (2007), the maximization of the complete log-likelihood is guaranteed. On another note, a data driven method to gauge the bounds a and b is needed.

4 Numerical studies

A brief numerical study is presented here, to compare the performance of the constrained vs unconstrained EM algorithm. More simulations have been performed, also with real datasets are available in (see Greselin and Ingrassia, 2013). A sample of $N = 150$ data has been generated with weights $\alpha = (0.3, 0.4, 0.3)'$ with parameters such that $\max_{i,g}\lambda_i(\Sigma_g) = 4.18$. We run 100 times both the unconstrained and the constrained AECM algorithms (for different values of the constraints a, b) using a common random initial clusterings.The unconstrained algorithm attains the right solution in 24% of cases; summary statistics about the misclassification error, over

the 100 runs, are reported in Table 1. To compare how a and b influences the performance of the constrained EM, different pairs of values has been considered, and Table 2 shows the more interesting cases. Further results are reported in Figure 1, where the boxplots of the distribution of the misclassification errors show the poor performance of the unconstrained algorithm compared to its constrained version. For all values of the upper bound b, the third quartile of the misclassification error is steadily equal to 0. Indeed, for $b = 6, 10$ and 15 we had no misclassification error, while we observed very low and rare misclassification errors only for $b = 20$ and $b = 25$ (respectively 3 and 11 not null values, over 100 runs). Moreover, the robustness of the results with respect to the choice of the upper constraint is apparent. A data driven method to select the bounds can be derived from the observed results, by running the constrained EM for increasing values of the upper bound, till a decrease in the final likelihood. The value of b before the decrease, observed over a series of run, will be chosen as upper value for the constrained estimation.

Table 1 Summary statistics for the Misclassif Error over 100 runs of the unconstrained EM alg

min	Q_1	Q_2	Q_3	max
0%	17%	36%	45.3%	60%

Table 2 Percentage of convergence to the right maximum of the constrained EM algorithm for $a = 0.01$ and some values of the upper constraint b

$b: +\infty$	6	10	15	20	25
24%	100%	100%	100%	97%	89%

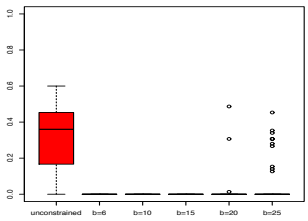

Fig. 1 Boxplots of the misclassification error. From left to right, boxplots refer to the unconstrained algorithm, then to the constrained algorithm, for $a = 0.01$ and $b = 6, 10, 15, 20, 25$.

References

Ghahramani, Z. and Hilton, G. (1997). The EM algorithm for mixture of factor analyzers. *Tech. Rep. CRG-TR-96-1*.

Greselin, F. and Ingrassia, S. (2013) Maximum likelihood estimation in constrained parameter spaces for mixtures of factor analyzers, *http://arxiv.org/abs/1301.1505*.

Ingrassia, S. (2004). A likelihood-based constrained algorithm for multivariate normal mixture models. *Stat. Meth. & Appl.*, **13**, 151–166.

Ingrassia, S. and Rocci, R. (2007). Constrained monotone em algorithms for finite mixture of multivariate gaussians. *Comp. Stat. & Data Anal.*, **51**, 5339–5351.

McLachlan, G., Peel, D., and Bean, R. (2003). Modelling high-dimensional data by mixtures of factor analyzers. *Comp. Stat. and Data Anal.*, **41**, 379–388.