

A Constrained MLE Formulation for Multinormal Mixture Decomposition

*Una formulazione vincolata per stime di massima verosimiglianza
per modelli miscuglio di distribuzioni normali multivariate*

Filippo Domma

Dipartimento di Economia e Statistica
Università della Calabria
87036 Arcavacata di Rende (CS), Italy
f.domma@unical.it

Salvatore Ingrassia¹

Dipartimento di Economia e Statistica
Università della Calabria
87036 Arcavacata di Rende (CS), Italy
s.ingrassia@unical.it

Riassunto: Il metodo della massima verosimiglianza nel caso di miscugli di distribuzioni normali conduce a problemi di ottimizzazione mal posti. In questo lavoro, generalizzando alcuni risultati di Hathaway (1985), viene proposta una formulazione vincolata della funzione di verosimiglianza per distribuzioni miscuglio di normali multivariate. Viene inoltre proposta una semplice procedura che consente l'implementazione di tale metodo in usuali algoritmi di ottimizzazione. Tale procedura si rivela di particolare interesse nelle applicazioni di tali modelli miscuglio per l'imputazione di dati mancanti.

Keywords: Mixture models, Likelihood, EM algorithm, Missing data.

1. Introduction

Finite mixture distributions play a central role in statistical modeling as they combine much of the flexibility of non parametric models with some analytic properties of parametric models, see e.g. Titterton *et al.* (1985). In this last decade, mixture distributions have found new fields of application in data mining, especially in the area of combining statistical models, see e.g. Jordan and Jacobs (1994), Sharkey (1999).

Let $f(\cdot; \psi)$ be the density function of a normal mixture distribution, where ψ assumes values in some parameter space Ψ and let $\mathcal{L}(\psi)$ be the log-likelihood function corresponding to a sample of size N with law $f(\cdot; \psi)$. The MLE $\hat{\psi}$ is usually computed by means of suitable optimization procedures; in most cases one uses the EM algorithm (see e.g. McLachlan and Krishnan (1997)) which generates a sequence $\{\psi^{(m)}\}_m$ – starting from some initial estimate $\psi^{(0)}$ – such that the corresponding sequence $\{\mathcal{L}(\psi^{(m)})\}_m$ is not decreasing. However, the convergence towards $\hat{\psi}$ is not guaranteed for two main reasons: *i*) the log-likelihood function $\mathcal{L}(\psi)$ may be unbounded and this may cause the failure of the algorithm; *ii*) the log-likelihood function $\mathcal{L}(\psi)$ presents local maxima, so that the final estimate depends on the initial guess $\psi^{(0)}$.

The univariate case has been investigated in a sequence of papers, in particular Hathaway (1985) proved that the likelihood function should be maximised in a suitable subset

¹F. Domma carried out the research within the project *Modelli di Durata: Sviluppi Metodologici ed Applicazioni a Fenomeni Economico-Sociali* (PRIN 2001, resp. F. Domma); S. Ingrassia carried out the research within the project *Metodi Statistici e Reti Neurali per l'Analisi di Dati Complessi* (PRIN 2000, resp. G. Lunetta).

of Ψ . Some ideas will be outlined in Section 2; a larger synthesis is provided in Ingrasias (1992). In this paper we approach the multivariate case. In Section 3 we present a constrained formulation of the maximum likelihood estimation for the decomposition of multivariate normal distributions; in Section 4 we illustrate a simple procedure in order to implement such ideas in practical algorithms. This procedure reveals of some interest in applications of mixture models for the imputation of missing data.

2. The univariate case

Let $f(x; \psi)$ be the density function of a mixture of k univariate normal component:

$$f(x; \psi) = \alpha_1 p(x; \mu_1, \sigma_1^2) + \cdots + \alpha_k p(x; \mu_k, \sigma_k^2) \quad (1)$$

where $p(x; \mu_j, \sigma_j^2)$ is the density function of a normal distribution with parameters μ_j, σ_j^2 ($j = 1, \dots, k$), and $\psi \in \Psi$ with:

$$\Psi = \{(\alpha_j, \mu_j, \sigma_j^2), \alpha_j \geq 0, \alpha_1 + \cdots + \alpha_k = 1, \mu_j \in \mathbb{R}, \sigma_j^2 > 0\} \subset \mathbb{R}^{3k}. \quad (2)$$

Let $\mathcal{L}(\psi)$ be the loglikelihood function of ψ for a given sample $\mathbf{x} = (x_1, \dots, x_N)$ of size N drawn from $f(x; \psi)$. It is well known that $\mathcal{L}(\psi|\mathbf{x})$ may show some singularities, thus the definition of estimate of maximum likelihood as the absolute maximum of $\mathcal{L}(\psi|\mathbf{x})$ lacks numerical sense. Under reasonable hypotheses, there exists only one strongly consistent solution of the likelihood equations and it is a point of local maximum of the likelihood function, see Redner and Walker (1984). In the following as the MLE estimate, we shall mean such a point of Ψ . In this context, Theorem 2.1 in Hathaway (1985) assures that if the sample $\mathbf{x} = (x_1, \dots, x_N)$ drawn with law (1) contains at least $k + 1$ distinct points, then for $c \in (0, 1]$, there exists a point of absolute maximum which is strongly consistent in the subset of Ψ satisfying the constraint $\min_{i \neq j} \left(\frac{\sigma_j^2}{\sigma_i^2} \right) \geq c > 0$.

3. The multivariate case

Let $f(\mathbf{x}; \psi)$ be the density of a mixture of k multinormal distribution:

$$f(\mathbf{x}; \psi) = \alpha_1 p(\mathbf{x}; \boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1) + \cdots + \alpha_k p(\mathbf{x}; \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) \quad (3)$$

where the α_j are the mixing weights and $p(\mathbf{x}; \boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j)$ is the density function of the multivariate normal distribution with mean vector $\boldsymbol{\mu}_j$ and covariance matrix $\boldsymbol{\Sigma}_j$ which is required to be positive definite (denoted by $\boldsymbol{\Sigma}_j > 0$). Finally we set $\psi = \{(\alpha_j, \boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j), j = 1, \dots, k\} \in \Psi$, where Ψ is the parameter space given by:

$$\Psi = \{(\alpha_j, \boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j), \alpha_j \geq 0, \alpha_1 + \cdots + \alpha_k = 1, \boldsymbol{\mu}_j \in \mathbb{R}^p, \boldsymbol{\Sigma}_j > 0\} \quad (4)$$

so that $\Psi \subset \mathbb{R}^{k(1+p+(p^2+p)/2)}$. It is well known that the matrix $\boldsymbol{\Sigma}_j$ is positive definite if and only if all its eigenvalues are strictly positive. Thus, if we denote by $\lambda_i(\boldsymbol{\Sigma}_j)$ the i -th eigenvalue of the j -th covariance matrix $\boldsymbol{\Sigma}_j$, the parameter space (4) can be written as:

$$\Psi = \{(\alpha_j, \boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j^2), \alpha_j \geq 0, \alpha_1 + \cdots + \alpha_k = 1, \boldsymbol{\mu}_j \in \mathbb{R}^p, \lambda_i(\boldsymbol{\Sigma}_j) > 0\} \quad (5)$$

Hathaway (1985) states that for a mixture of k p -variate normals, constraining all eigenvalues of $\Sigma_i \Sigma_j^{-1}$ ($1 \leq i < j \leq p$) to be greater than or equal to some minimum positive value (satisfied by the true parameter) leads to a constrained (global) maximum-likelihood formulation as the assumptions of Kiefer and Wolfowitz (1956) are satisfied. In order to make this result applicable in practical algorithms this constraint is here satisfied by imposing bounds on the eigenvalues of each covariance matrix Σ_j ($j = 1, \dots, k$). Indeed we prove the following result.

Proposition 1 Let $\mathbf{A}, \mathbf{B}$ two $p \times p$ symmetric and positive definite marices. Then we have:

$$\lambda_{\max}(\mathbf{A}\mathbf{B}^{-1}) \leq \frac{\lambda_{\max}(\mathbf{A})}{\lambda_{\min}(\mathbf{B})} \quad (6)$$

$$\lambda_{\min}(\mathbf{A}\mathbf{B}^{-1}) \geq \frac{\lambda_{\min}(\mathbf{A})}{\lambda_{\max}(\mathbf{B})} \quad (7)$$

where $\lambda_{\min}(\cdot)$ and $\lambda_{\max}(\cdot)$ are respectively the smallest and the largest eigenvalue of the matrix $(\cdot)$.

In particular the eigenvalues of Σ_j give the variances along the principal axes; thus imposing bounds on the eigenvalues of Σ_j amounts to imposing bounds on these variances. Let a, b two strictly positive constants and assume that the covariance matrices Σ_j satisfy the constraints:

$$a \leq \lambda_i(\Sigma_j) \leq b \quad i = 1, \dots, p, j = 1, \dots, k. \quad (8)$$

Then we consider the following constrained parameter space $\Psi_{a,b}$ of Ψ :

$$\Psi_{a,b} = \{(\alpha_j, \boldsymbol{\mu}_j, \Sigma_j), \alpha_j \geq 0, \alpha_1 + \dots + \alpha_k = 1, \boldsymbol{\mu}_j \in \mathbb{R}^p, a \leq \lambda_i(\Sigma_j) \leq b\}.$$

4. A practical recipe

In order to implement the constraints (8), we can apply some simple results of the matrix theory stating that, if $\mathbf{A}$ and $\mathbf{B}$ are two $p \times p$ symmetric matrices, then we have:

$$\lambda_{\min}(\mathbf{A} + \mathbf{B}) \geq \lambda_{\min}(\mathbf{A}) + \lambda_{\min}(\mathbf{B}) \quad (9)$$

$$\lambda_{\max}(\mathbf{A} + \mathbf{B}) \leq \lambda_{\max}(\mathbf{A}) + \lambda_{\max}(\mathbf{B}) \quad (10)$$

see e.g. Horn and Johnson (1985).

Let $\Sigma_j^{(m)}$ be the j -th covariance matrix evaluated at the m -th iteration of the algorithm and denote respectively by $\lambda_{\min}(\Sigma_j^{(m)})$ and $\lambda_{\max}(\Sigma_j^{(m)})$ the smallest and the largest eigenvalue of $\Sigma_j^{(m)}$ ($j = 1, \dots, k$). Afterwards let us consider:

$$\begin{aligned} \lambda_*^{(m)} &= \min_{j=1, \dots, k} \left\{ \lambda_{\min}(\Sigma_1^{(m)}), \dots, \lambda_{\min}(\Sigma_k^{(m)}) \right\} \\ \lambda^*{}^{(m)} &= \max_{j=1, \dots, k} \left\{ \lambda_{\max}(\Sigma_1^{(m)}), \dots, \lambda_{\max}(\Sigma_k^{(m)}) \right\}. \end{aligned}$$

Let us set $\epsilon_*^{(m)} = a - \lambda_*^{(m)}$. The matrix $\epsilon_*^{(m)} \mathbf{I}_p$, where $\mathbf{I}_p$ is the p -dimensional identity matrix, has the eigenvalue $\epsilon_*^{(m)}$ with multiplicity p . Thus if $\lambda_{\min}(\Sigma_j^{(m)}) < a$ for some j ,

the matrix $\Sigma_j^{(m)} + \epsilon_*^{(m)} \mathbf{I}_p$ satisfies the lower bound in (8). As matter of fact, according to the relation (9), we have:

$$\lambda_{\min}(\Sigma_j^{(m)} + \epsilon_*^{(m)} \mathbf{I}_p) \geq \lambda_{\min}(\Sigma_j^{(m)}) + \epsilon_*^{(m)} \geq a .$$

Analogous arguments apply in order to impose the upper bound in (8) according to the relation (10). If $\lambda_{\max}(\Sigma_j^{(m)}) > b$ for some j , then let us set $\epsilon^{*(m)} = \lambda^{*(m)} - b$; thus the matrix $\Sigma_j^{(m)} - \epsilon^{*(m)} \mathbf{I}_p$ satisfies the constraint:

$$\lambda_{\max}(\Sigma_j^{(m)} - \epsilon^{*(m)} \mathbf{I}_p) \leq \lambda_{\max}(\Sigma_j^{(m)}) - \epsilon^{*(m)} \leq b .$$

The two constraints described above can be implemented quite easily in optimization algorithms, like the EM algorithm, performing some steps at the m -th iteration, after computing the eigenvalues of $\Sigma_j^{(m)}$.

An important field of application concerns the missing data imputation by means of a mixture of multivariate normal distributions, see Little and Schluchter (1985) especially in neural network training from incomplete datasets, see e.g. Domma and Ingrassia (2001). Indeed in such cases, due to the imputation of the missing data on the basis of the mixture models, the variation is underestimated. In many applications using the EM algorithm, we observed that this procedure converged to some spurious local maxima of the likelihood function characterized by a (practically) singular covariance matrix. The lower bound in (8) is intended to prevent just this phenomenon. Some computer simulations showed that the method presented in this paper has been applied successfully.

References

- Domma F. and Ingrassia S. (2001) Mixture models for maximum likelihood estimation from incomplete values, in: *Advances in Classification and Data Analysis*, Borra S., Rocci M., Vichi M. and Schader M., (Eds.), Springer-Verlag, Berlin, 201–208.
- Hathaway R. (1985) A constrained formulation of maximum-likelihood estimation for normal mixture distributions, *The Annals of Statistics*, 13, 795–800.
- Horn R. and Johnson C. (1985) *Matrix Analysis*, Cambridge University Press, Cambridge.
- Ingrassia S. (1992) A comparison between the simulated annealing and the EM algorithms in normal mixture decompositions, *Statistics and Computing*, 2, 203–211.
- Jordan M. and Jacobs R. (1994) Hierarchical mixtures of experts and the EM algorithm, *Neural Computation*, 3, 79–97.
- Kiefer J. and Wolfowitz J. (1956) Consistency of the maximum likelihood estimator in the presence of infinitely many incidental parameters, *Annals of Mathematical Statistics*, 27, 888–906.
- Little R. and Schluchter M. (1985) Maximum likelihood estimation for mixed continuous and categorical data with missing values, *Biometrika*, 72, 497–512.
- McLachlan G. and Krishnan T. (1997) *The EM Algorithm and Extensions*, John Wiley & Sons, New York.
- Redner R. and Walker H. (1984) Mixture densities, maximum likelihood and the EM algorithm, *SIAM Reviews*, 46, 195–239.
- Sharkey A. (1999) *Combining Artificial Neural Nets*, Springer-Verlag, Berlin.
- Titterton D., Smith A. and Makov U. (1985) *Statistical Analysis of Finite Mixture Distributions*, John Wiley & Sons, Chichester.