AN ITERATIVE PROJECTIONS ALGORITHM FOR ML FACTOR ANALYSIS

Abd-Krim Seghouane

National ICT Australia,
Canberra Research Laboratory
And
The Australian National University,
Research School of Information Sciences and Engineering
E-mail: Abd-krim.seghouane@nicta.com.au

ABSTRACT

Alternating minimization of the information divergence is used to derive an effective algorithm for maximum likelihood (ML) factor analysis. The proposed algorithm is derived as an iterative alternating projections procedure on a model family of probability distributions defined on the factor analysis model and a desired family of probability distributions constrained to be concentrated on the observed data. The algorithm presents the advantage of being simple to implement and stable to converge. A simulation example that illustrates the effectiveness of the proposed algorithm for ML factor analysis is presented.

1. INTRODUCTION

The most common and simple example of a latent variable model is that of factor analysis [1] which has been widely used in many disciplines such as biology, social sciences, economics and engineering. In this model, $f(.,W)$ is a linear function of $\mathbf{x}$

$$\mathbf{y} = W\mathbf{x} + \mu + \varepsilon. \quad (1)$$

Conventionally, the latent variables also known as the factors are defined to be independent and Gaussian with unit variance, so $\mathbf{x} \sim N(0, I_q)$. The noise is also Gaussian such that $\varepsilon \sim N(0, \Psi)$, with Ψ diagonal and the $p \times q$ parameter matrix W contains the factor loadings. Generally, $q < p$ such that the latent variables or factors identifies the common characteristics among the observed data. The parameter μ permits the observations to have nonzero mean. With this formulation, the observation vectors are also normal $N(\mu, WW^\top + \Psi)$. Note that given the factors $\mathbf{x}$, the observation variables are independent. This assumption of conditional independence is the key one in the factor analysis model. In the terminology of factor analysis, this model is called exploratory factor analysis model.

Building a factor analysis model for the observation vectors requires the estimation of W, Ψ and μ for which no closed form analytic solution exists. Direct ML estimation has been widely used for fitting factor analysis models. A variety of iterative algorithms to perform ML estimation have been proposed in the literature [2], however, they present several practical problems [3]. ML estimation of factor analysis can be conceptualized as ML estimation in a multivariate normal model with missing data. In this case, the easiest algorithm to implement ML estimation and the most stable in the sense of monotonically increasing the likelihood, is the EM algorithm [4]. EM for ML factor analysis was described in [5]. Despite its reliable monotone convergence, the convergence rate of EM can be impractically slow in factor analysis models [6]. To obtain ML estimates more efficiently in factor analysis models the ECME algorithm [7] was proposed in [8][9].

The EM algorithm for ML estimation is one of the most widely used parameter estimation procedures from incomplete data. In [10], the EM algorithm was described in geometric terms. The framework used in [10] for studying the EM algorithm is referred to as being information geometric because it is based on a geometric property of the information divergence, which is treated as a distance measure between probability distributions. Under this information geometric framework, the EM algorithm can be viewed as an alternating minimization procedure of the Kullback-Leibler (KL) [11] between a parameter family and a desired family of probability distributions [12][13]..

Based on this similarity between information geometric alternating projections and the EM algorithm, ML estimation for fitting factor analysis models is approached as a double minimization of the KL divergence between two probability distributions in this paper. Using the Gaussian assump-

National ICT Australia is funded by the Australian Department of Communications, Information Technology and the Arts and the Australian Research Council through Backing Australia's Ability and the ICT Center of Excellence Program.

1-4244-2376-7/08/$20.00 ©2008 IEEE

tion on the factors closed form solutions for these alternative projections are developed. It is not attempted in this paper to provide a complete comparison with the array of competing algorithms for ML factor analysis. The emphasis is to give a new insight from an information geometric point of view to the ML factor analysis problem and its relationship to the ML estimation using EM algorithm. Moreover, the derived closed form solutions for the double minimization are very simple to implement and stable to converge.

2. PROBLEM FORMULATION

Let Y be the $n \times p$ observed data matrix corresponding to n p-dimensional i.i.d observed data vectors. Constructing a factor analysis model for the observation consists in approximating $\mathbf{y}_i$, $i = 1, ..., n$ using the model

$$\mathbf{y}_i = W\mathbf{x}_i + \mu + \varepsilon_i, \quad i = 1, ..., n \tag{2}$$

where μ is a p-dimensional mean vector, W is the $p \times q$ factor loading matrix, $\mathbf{x}_i$ is the unobservable variable vector consisting of $q < p$ factors that follows a $N(0, I_q)$ and the noise vectors ε_i, $i = 1, ..., n$ are i.i.d $N(0, \Psi)$ where $\Psi = \text{diag}(\psi_1, ..., \psi_p)$ is a diagonal matrix. In the terminology of factor analysis, the vector $(\psi_1, ..., \psi_p)$ is called vector of uniquenesses.

ML estimation has been popular for fitting factor analysis models, it is obtained by minimizing the negative log-likelihood

$$\begin{aligned} l(W, \Psi, \mu|Y) &= \frac{n}{2}\left\{\log|\Sigma| + \text{tr}(\Sigma^{-1}S)\right. \\ &+ \left.(\bar{\mathbf{y}} - \mu)^\top \Sigma^{-1}(\bar{\mathbf{y}} - \mu)\right\} \end{aligned} \tag{3}$$

where $\Sigma = \Psi + WW^\top$,

$$\bar{\mathbf{y}} = \frac{1}{n}\sum_{i-1}^{n}\mathbf{y}_i \quad \text{and} \quad S = \frac{1}{n}\sum_{i=1}^{n}(\mathbf{y} - \bar{\mathbf{y}})(\mathbf{y} - \bar{\mathbf{y}})^\top.$$

The ML estimator of μ is the sample mean $\mathbf{y}$, and hence $\theta = (W, \Psi)$ can be estimated by minimizing

$$l(\theta|Y) = \frac{n}{2}\left\{\log|\Sigma| + \text{tr}(\Sigma^{-1}S)\right\}. \tag{4}$$

As described above in the Introduction, different iterative algorithms have been introduced to generate estimators of W and Ψ. However, these algorithms present the inconvenient of having a convergence none guaranteed. Indeed, these iterative algorithms do not define an EM algorithm. Then, they do not necessarily enjoy the general convergence properties of EM algorithms [14] which are more adapted to this kind of estimation problem as it is a missing data problem since the observations alone are incomplete for estimating θ.

Minimizing $l(\theta|Y)$ with respect W and Ψ is equivalent to minimizing

$$\begin{aligned} l(\theta|Y) &= \frac{1}{2}\left\{\log|\Sigma| + \text{tr}(\Sigma^{-1}S) - \log|S| - n\right\} \\ &= \text{KL}(N(0, S) \parallel N(0, \Sigma)), \end{aligned} \tag{5}$$

which represents the KL divergence between two Gaussian distributions with zero means. Therefore, fitting factor analysis models based on ML estimation is equivalent to searching for a best approximation model according to the criterion which consists in minimizing the KL divergence between the observations generating distribution and the parametric approximation distribution. This KL divergence minimization takes into account only distributions characterizing the observed incomplete data. To minimize the KL divergence between probability distributions that describe the complete data, the view of the EM algorithm as an alternating minimization procedure using information geometric framework can be adopted.

In this paper, the ML factor analysis problem is approached using an information geometric framework [15]. More specifically, ML estimation is posed as a double projection onto two sets of probability distributions or as a double minimization of the KL divergence between two probability distributions.

3. INFORMATION GEOMETRIC APPROACH TO ML FACTOR ANALYSIS

The information geometric principle [10] is applied to ML factor analysis to derive an efficient iterative algorithm for which convergence properties can be derived [15].

To derive an estimator of θ, the appropriate sets of probability distributions P and Q have to be introduced first. From the factor analysis model (1), each member $q(\mathbf{y}, \mathbf{x}; \theta)$ of Q is a Gaussian distribution $N(\alpha, \Delta)$, where

$$\alpha = \begin{bmatrix} \mu \\ 0 \end{bmatrix} \in R^{(p+q)\times 1} \quad \text{and} \tag{6}$$

$$\Delta = \begin{bmatrix} WW^\top + \Psi & W \\ W^\top & I \end{bmatrix} \in R^{(p+q)\times(p+q)}. \tag{7}$$

The members $p(\mathbf{y}, \mathbf{x})$ of the set of generating distributions P are also Gaussian with marginal distribution $p(\mathbf{y}) = N(\mu, \Sigma)$ for which $\bar{\mathbf{y}}$ and S are consistent estimators of the mean and the covariance matrix respectively.

Having $q(\mathbf{y}, \mathbf{x}; \theta^{(k)})$ obtained from the previous iteration, the first step of the algorithm consists in constructing the approximation of the complete data distribution

$$\begin{aligned} p^{(k+1)}(\mathbf{y}, \mathbf{x}) &= \arg\min_{p\in P} KL(p(\mathbf{y}, \mathbf{x}) \parallel q(\mathbf{y}, \mathbf{x}; \theta^{(k)})) \\ &= q(\mathbf{x}|\mathbf{y}; \theta^{(k)})p(\mathbf{y}). \end{aligned}$$

However, since only an estimate $\hat{p}(\mathbf{y}) = N(\bar{\mathbf{y}}, S)$ of $p(\mathbf{y})$ is available the approximation of the complete data distribution is given by

$$p^{(k+1)}(\mathbf{y}, \mathbf{x}) = q(\mathbf{x}|\mathbf{y}; \theta^{(k)})\hat{p}(\mathbf{y}). \tag{8}$$

Since P is a set of Gaussian distributions, this step is resumed by finding the mean and covariance of $p^{(k+1)}(\mathbf{y}, \mathbf{x}) = N(\lambda, \Omega)$. From (6) and (7), we have

$$q(\mathbf{x}|\mathbf{y}, \theta^{(k)}) = N(\mu^k_{\mathbf{x}|\mathbf{y}}, \Sigma^k_{\mathbf{x}|\mathbf{y}})$$

where

$$\begin{aligned} \mu^k_{\mathbf{x}|\mathbf{y}} &= (I + W_k^\top \Psi_k^{-1} W_k)^{-1} W_k^\top \Psi_k^{-1}(\mathbf{y} - \mu) \\ \Sigma^k_{\mathbf{x}|\mathbf{y}} &= (I + W_k^\top \Psi_k^{-1} W_k)^{-1} \end{aligned}$$

and with simple mathematical manipulations detailed in the Appendix, we obtain

$$\lambda = \begin{bmatrix} \bar{\mathbf{y}} \\ \Sigma^k_{\mathbf{x}|\mathbf{y}} W_k^\top \Psi_k^{-1}(\bar{\mathbf{y}} - \mu) \end{bmatrix} \quad \text{and} \tag{9}$$

$$\Omega = \begin{bmatrix} \omega_{11} & \omega_{12} \\ \omega_{21} & \omega_{22} \end{bmatrix}, \tag{10}$$

where

$$\begin{aligned} \omega_{11} &= S \\ \omega_{12} &= S(W_k W_k^\top + \Psi_k)^{-1} W_k \\ \omega_{21} &= W_k^\top (W_k W_k^\top + \Psi_k)^{-1} S \\ \omega_{22} &= I - W_k^\top (W_k W_k^\top + \Psi_k)^{-1} W_k \\ &+ W_k^\top (W_k W_k^\top + \Psi_k)^{-1} S (W_k W_k^\top + \Psi_k)^{-1} W_k \end{aligned}$$

Having generated $p^{(k+1)}(\mathbf{y}, \mathbf{x})$ from the first I-projection, the second partial minimization consists in generating the ML estimation of the parameters $\theta^{(k+1)}$ using the updated complete data distribution. This is equivalent to finding the I-projection of $p^{(k+1)}(\mathbf{y}, \mathbf{x})$ onto Q

$$q(\mathbf{y}, \mathbf{x}, \theta^{(k+1)}) = \arg\min_{\theta \in \Theta} KL(p^{(k+1)}(\mathbf{y}, \mathbf{x}) \parallel q(\mathbf{y}, \mathbf{x}; \theta)).$$

As described by (6) and (7) the members of Q are Gaussian distributions $N(\alpha, \Delta)$ parameterized by θ and μ, if, for simplicity, we denote by

$$\lambda = \begin{bmatrix} \lambda_1 \\ \lambda_2 \end{bmatrix}, \tag{11}$$

the mean and of $p^{(k+1)}(\mathbf{y}, \mathbf{x})$ which is defined by (9), then $KL(p^{(k+1)}(\mathbf{y}, \mathbf{x}) \parallel q(\mathbf{y}, \mathbf{x}; \theta^{(k+1)}))$

$$\begin{aligned} &= KL(N(\lambda, \Omega), N(\alpha, \Delta_{k+1})) \\ &= \mathrm{tr}(\Delta_{k+1}^{-1}\Omega) + \ln \mid \Delta_{k+1} \mid - \ln \mid \Omega \mid \\ &+ (\alpha - \lambda)^\top \Delta_{k+1}^{-1}(\alpha - \lambda) - (p + q) \end{aligned} \tag{12}$$

with

$$\begin{aligned} \mathrm{tr}(\Delta_{k+1}^{-1}\Omega) &= \mathrm{tr}(\Psi_{k+1}^{-1}\omega_{11}) - 2\mathrm{tr}(\Psi_{k+1}^{-1} W_{k+1}\omega_{12}^\top) \\ &+ \mathrm{tr}((I + W_{k+1}^\top \Psi_{k+1}^{-1} W_{k+1})\omega_{22}), \end{aligned}$$

$$(\alpha - \lambda)^\top \Delta_{k+1}^{-1}(\alpha - \lambda) = (\bar{\mathbf{y}} - \mu)^\top (\Psi_{k+1} + W_{k+1} W_{k+1}^\top)^{-1}(\bar{\mathbf{y}} - \mu)$$

and

$$\begin{aligned} \ln \mid \Delta_{k+1} \mid &= -\ln \mid \Delta_{k+1}^{-1} \mid \\ &= -\ln \mid \Psi_{k+1}^{-1} \mid \\ &\cdot \mid I + W_{k+1}^\top \Psi_{k+1}^{-1} W_{k+1} \\ &- W_{k+1}^\top \Psi_{k+1}^{-1} \Psi_{k+1} \Psi_{k+1}^{-1} W_{k+1} \mid \\ &= -\ln \mid \Psi_{k+1}^{-1} \mid . \end{aligned}$$

Finding the ML estimate of θ_{k+1} and μ is equivalent to the minimization of (12) with respect to μ, W_{k+1} and Ψ_{k+1} which easily gives

$$\mu = \bar{\mathbf{y}} \tag{13}$$

$$W_{k+1} = S\Phi_k^{-1} W_k \Gamma_k^{-1} \tag{14}$$

$$\Psi_{k+1} = \mathrm{diag}(S - W_{k+1} W_k^\top \Phi_k^{-1} S). \tag{15}$$

where $\Phi_k = \Psi_k + W_k W_k^\top$ and $\Gamma_k = I - W_k^\top \Phi_k^{-1} W_k + W_k^\top \Phi_k^{-1} S \Phi_k^{-1} W_k$. Therefore, the iterative application of (14) and (15) generates in the limit the ML estimates for the factor analysis model parameters W and Ψ. From (15) and the fact that $\Psi_{k+1} > 0$, the matrix $W_{k+1} W_k^\top \Phi_k^{-1} S$ is strictly dominated by S in the sense of positive matrices.

4. CONVERGENCE ANALYSIS

It is straightforward to establish that

$$\begin{aligned} &KL(p^{(k+1)}(\mathbf{y}, \mathbf{x}) \parallel q(\mathbf{y}, \mathbf{x}; \theta^{(k+1)})) \\ &\quad \leq KL(p^{(k+1)}(\mathbf{y}, \mathbf{x}) \parallel q(\mathbf{y}, \mathbf{x}; \theta^{(k)})) \\ &\quad \leq KL(p^{(k)}(\mathbf{y}, \mathbf{x}) \parallel q(\mathbf{y}, \mathbf{x}; \theta^{(k)})). \end{aligned}$$

Therefore, the sequence of parameters generated by the algorithm decreases the KL divergence between $p(\mathbf{y}, \mathbf{x})$ and $q(\mathbf{y}, \mathbf{x}; \theta)$ and then increases the likelihood. Since $KL(p(\mathbf{y}, \mathbf{x}) \parallel q(\mathbf{y}, \mathbf{x}; \theta))$ is bounded below by zero, it converges to a local minimum when $k \to \infty$.

Analysis of convergence of the proposed algorithm to a global minimum is similar to that of the EM algorithm and is beyond the scope of this paper [14]. However, it can easily be demonstrated that the convergence points for $p(\mathbf{y}, \mathbf{x})$ and $q(\mathbf{y}, \mathbf{x}; \theta))$ are similar. Indeed, S and $\bar{\mathbf{y}}$ are efficient estimators of μ and Σ, then,

$$S \to \Sigma \quad \text{and} \quad \bar{\mathbf{y}} \to \mu \quad \text{when} \quad n \to \infty,$$

and for $\Sigma = WW^\top + \Psi$ the algorithm described by (14) and (15) stops since $\Omega = \Delta$ which corresponds to the intersection of the two families of probability distributions P and Q since $\lambda = \alpha$. Therefore,

$$KL(p^{(k)}(\mathbf{y},\mathbf{x}) \parallel q(\mathbf{y},\mathbf{x};\theta^{(k)})) \longrightarrow 0.$$
$$n,k \longrightarrow \infty$$

5. SIMULATION RESULTS

To illustrate the performance of the proposed algorithm in fitting factor analysis models to observed data, a set of artificial data were generated from model (1) with

$$\Psi = \text{diag}\left[\begin{array}{cccccccccc} 1 & 2 & 3 & 4 & 5 & 6 & 7 & 8 & 9 & 10 \end{array}\right],$$

$$\mu = \left[\begin{array}{cccccccccc} 3 & 3 & 3 & 3 & 7 & 7 & 7 & 7 & 7 & 7 \end{array}\right].$$

and W'=

$$\left[\begin{array}{cccccccccc} 1.3 & 1 & 1.5 & 2.3 & 1.8 & 1.2 & 1.5 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1.8 & 2.2 & 1 & 1.8 & 1.2 & 1.5 \\ 3.5 & 2 & 2.5 & 1.5 & 2 & 3 & 2.5 & 1.8 & 1.4 & 1.3 \\ 4 & 2.2 & 1.3 & 2.4 & 0 & 0 & 0 & 2 & 3.1 & 2.7 \end{array}\right].$$

To examine the performance of this proposed algorithm, factor analysis models with different number of factors q were fitted to the same data sets which consists of $n = 500$ and $n = 1000$. To show how the algorithm performs in these different settings, the Frobenius norm of the error in estimating Ψ

$$\text{Err} = \|\hat{\Psi}_k - \Psi\|_F^2$$

was used as a performance measure.

Figure 1(a) and (c) show the evolvement of the error Err as a function of the number of iterations k, in estimating Ψ for factor analysis models with $q = 1$, $q = 2$ and $q = 3$ respectively and $n = 500$. Figure 2(a) and (c) show the same error for $n = 1000$.

From these figures we observe that convergence of the algorithm occurs after 20-100 iterations. Therefore, the proposed algorithm appears to be an attractive algorithm for ML factor analysis. The convergence results of figures 2 are better than the ones of figures 1, this is due to the larger number of data used in the experiment used to generate figures 2. This result confirms the convergence results established in the previous section. The relative performance of the proposed algorithm to other existing algorithms was not tackled here since it was not our objective to develop a competing algorithm.

6. CONCLUSION

Under the information geometric framework as developed in [10], an algorithm for ML factor analysis which is easy to implement and stable to converge is proposed. Unlike the EM approach to ML factor analysis, this framework allows an intuitive understanding of the algorithm. The proposed algorithm is developed by using a relaxation procedure, lifting the original ML estimation problem from the minimization in one probability density set to the minimization in two probability density sets. In these two probability sets, the ML factor analysis problem is formulated as an alternating minimization procedure of the KL divergence between a model family and a desired family of probability densities leading in a natural way to an iterative alternating projections algorithm. Alternating minimization of the information divergence is a powerful computational procedure for ML estimation. The detailed presentation of its application to ML factor analysis can be viewed as a simple illustration of its use when searching for structure in multivariate data with latent variables treated as missing data. The performance of the proposed algorithm in a simulation example of ML factor analysis was investigated. Simulation results illustrating the convergence behavior were given.

Appendix

As in [16](page 91),the mean λ and the covariance matrix Ω of $p^{(k+1)}(\mathbf{y},\mathbf{x})$ can be obtained by deriving the first and second order terms of

$$\log p^{(k+1)}(\mathbf{y},\mathbf{x}) = \log q(\mathbf{x}|\mathbf{y};\theta^{(k)}) + \log \hat{p}(\mathbf{y})$$

which are obtained from the expression

$$\begin{aligned} & - \frac{1}{2}(\mathbf{y}-\bar{\mathbf{y}})^\top S^{-1}(\mathbf{y}-\bar{\mathbf{y}}) \\ & - \frac{1}{2}\left(\mathbf{x} - \Sigma^k_{\mathbf{x}|\mathbf{y}} W_k^\top \Psi_k^{-1}(\mathbf{y}-\mu)\right)^\top (\Sigma^k_{\mathbf{x}|\mathbf{y}})^{-1} \\ & \cdot \left(\mathbf{x} - \Sigma^k_{\mathbf{x}|\mathbf{y}} W_k^\top \Psi_k^{-1}(\mathbf{y}-\mu)\right). \end{aligned}$$

After some manipulations, the second order can be written as

$$\left[\begin{array}{c} \mathbf{y} \\ \mathbf{x} \end{array}\right]^\top \cdot \left[\begin{array}{cc} S^{-1} + \Psi_k^{-1} W_k \Sigma^k_{\mathbf{x}|\mathbf{y}} W_k^\top \Psi^{-1} & -\Psi_k^{-1} W_k \\ -W_k^\top \Psi_k^{-1} & (\Sigma^k_{\mathbf{x}|\mathbf{y}})^{-1} \end{array}\right] \cdot \left[\begin{array}{c} \mathbf{y} \\ \mathbf{x} \end{array}\right]$$

and the covariance matrix Ω given in (10) is obtained by inverting the above precision matrix with the following inversion result

$$\left[\begin{array}{cc} A & B \\ C & D \end{array}\right]^{-1} = \left[\begin{array}{cc} E & F \\ G & H \end{array}\right]$$

where

$$\begin{aligned} E &= (A - BD^{-1}C)^{-1} \\ F &= -(A - BD^{-1}C)^{-1}BD^{-1} \\ G &= -D^{-1}C(A - BD^{-1}C)^{-1} \\ H &= D^{-1} + D^{-1}C(A - BD^{-1}C)^{-1}BD^{-1} \end{aligned}$$

and this identity involving matrix inverses

$$(P^{-1} + B^\top R^{-1} B) B^\top R^{-1} = P B^\top (B P B^\top + R)^{-1}$$

From [16](page 92), we have

$$E \begin{bmatrix} \mathbf{y} \\ \mathbf{x} \end{bmatrix} =$$

$$\begin{bmatrix} S & S\Psi_k^{-1} W_k \Sigma_{\mathbf{x}|\mathbf{y}}^k \\ \Sigma_{\mathbf{x}|\mathbf{y}}^k W_k^\top \Psi_k^{-1} S & \Sigma_{\mathbf{x}|\mathbf{y}}^k + \Sigma_{\mathbf{x}|\mathbf{y}}^k W_k^\top \Psi_k^{-1} S \Psi_k^{-1} W_k \Sigma_{\mathbf{x}|\mathbf{y}}^k \end{bmatrix}$$

$$\times \begin{bmatrix} S^{-1}\bar{\mathbf{y}} + \Psi_k^{-1} W_k \Sigma_{\mathbf{x}|\mathbf{y}}^k W_k^\top \Psi_k^{-1} \mu \\ -W_k^\top \Psi_k^{-1} \mu \end{bmatrix}$$

$$= \begin{bmatrix} \bar{\mathbf{y}} \\ \Sigma_{\mathbf{x}|\mathbf{y}}^k W_k^\top \Psi_k^{-1} (\bar{\mathbf{y}} - \mu) \end{bmatrix}$$

where corresponds to the mean λ given in (9).

7. REFERENCES

[1] B. S. Everitt, *An Introduction to Latent Variable Models*, Chapman and Hall, London, 1984.

[2] K. G. Jörekog, "Some contribution to maximum likelihood factor analysis," *Psychometrika*, vol. 32, pp. 433–484, 1967.

[3] R. I. Jennirch and S. M. Bobinson, "A newton-raphson algorithm for maximum likelihood factor analysis," *Psychometrika*, vol. 34, pp. 111–123, 1969.

[4] A. P. Dempster, N. M. Laird, and D. B. Rubin, "Maximum likelihood from incomplete data using the em algorithm," *Journal of the Royal Statistical Society, Serie B*, vol. 39, pp. 1–38, 1977.

[5] D. B. Rubin and D. T. Thayer, "[em," .

[6] P. M. Bentler and J. S. Tanaka, "Problems with EM algorithms for ML factor analysis," *Psychometrika*, vol. 48, pp. 247–251, 1983.

[7] C. Liu, "The ECME algorithm: a simple extention of [em," .

[8] C. Liu and D. B. Rubin, "Maximum likelihood estimation of factor analysis using the ECME algorithm with complete and incomplete data," *Statistica Sinica*, vol. 8, pp. 729–747, 1998.

[9] L. Finesso and P. Spreij, "Factor analysis and alternating minimization," *Alessandro Chiuso, Stefano Pinzoni and Augusto Ferrante Eds, Springer Lecture Notes in Control and Information Sciences*, vol. 364, pp. 85–96, 2007.

[10] I. Csiszár and G. Tusnády, "Information geometry and alternating minimization procedures," *Statistics and Decisions*, vol. 1, pp. 205–237, 1984.

[11] S. Kullback and R. A. Leibler, "On information and sufficiency," *The Annals of Mathematics Statistics*, vol. 22, pp. 76–86, 1951.

[12] S. I. Amari, "Information geometry of the EM and em algorithms for neural networks," *Neural Networks*, vol. 8, pp. 1379–1408, 1995.

[13] A. Zia, J. P. R. Reilly, J. Manton, and S. Shirani, "An information geometry approach to ML estimation with incomplete data: application to semiblind mimo channel identification," *IEEE Transactions on Signal Processing*, vol. 55, pp. 3975–3985, 2007.

[14] C. F. J. Wu, "On the convergence properties of the EM algorithm," *Annals of Statistics*, vol. 11, pp. 95–103, 1983.

[15] I. Csiszár, "I-divergence geometry of probability distributions and minimization problems," *Annals of Probability*, vol. 3, pp. 146–158, 1975.

[16] C. M. Bishop, *Pattern recognition and machine learning*, Springer, New York, 2006.

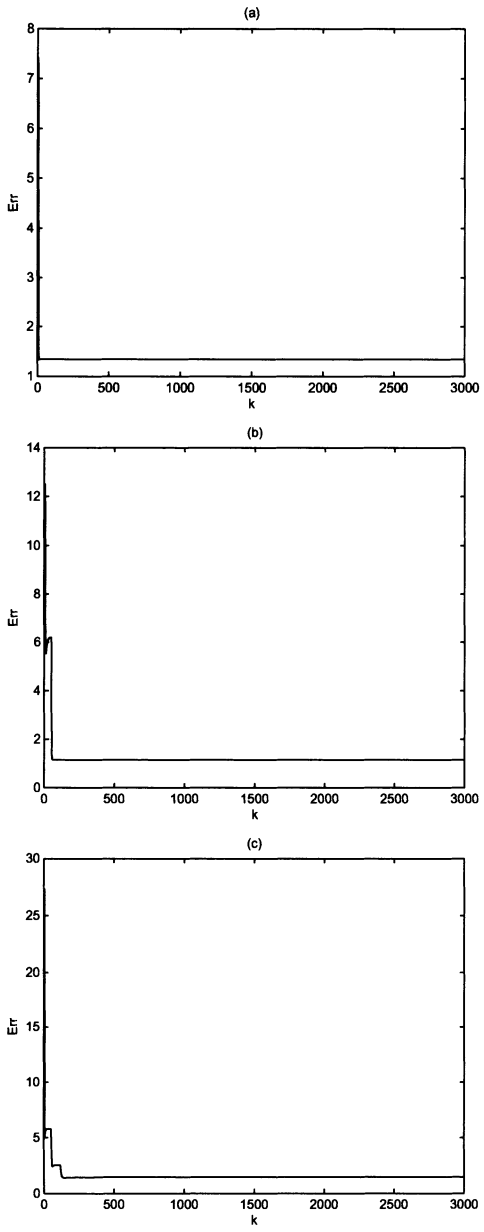

Fig. 1. Evolvement of Err with the number of iterations k for $n = 500$ and (a) $q = 1$, (b) $q = 2$ and (c) $q = 3$.

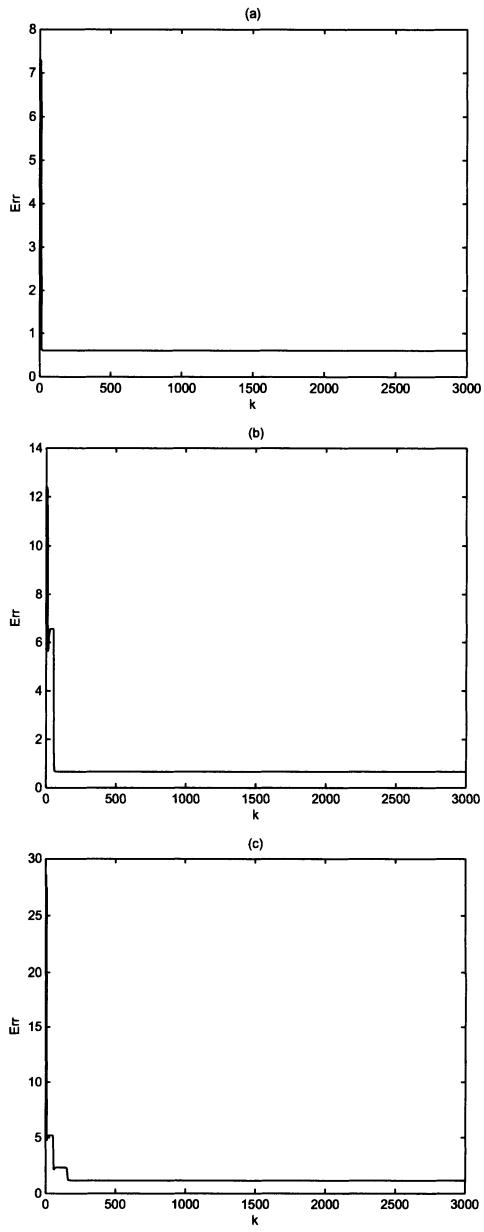

Fig. 2. Evolvement of Err with the number of iterations k for $n = 1000$ and (a) $q = 1$, (b) $q = 2$ and (c) $q = 3$.