Investigations on Non-Gaussian Factor Analysis

Zhi-Yong Liu, *Student Member, IEEE*, Kai-Chun Chiu, *Student Member, IEEE*, and Lei Xu, *Fellow, IEEE*

***Abstract*—This letter further explores the Bayesian Ying-Yang learning based non-Gaussian factor analysis (NFA) via investigating its key yet analytically intractable factor estimating step. Among the three suggested numerical approaches we empirically show that the so-called *iterative fixed posteriori approximation* approach is the most optimal, as well as theoretically prove that the *iterative fixed posteriori approximation* is another type of EM-algorithm, with the proof of its convergence also shown.**

***Index Terms*—BYY harmony learning, EM-algorithm, factor analysis, independent component analysis, non-Gaussian factor analysis.**

I. Introduction

THE recently proposed non-Gaussian factor analysis (NFA) [1]–[3] generalizes the well-known factor analysis (FA)

$$x(t) = Ay(t) + e(t), \quad x \in \mathbb{R}^n, \quad y \in \mathbb{R}^m \tag{1}$$

by assuming that each factor y_j follows non-Gaussian distribution. In effect, it not only avoids the rotation and additive indeterminacies encountered by classical FA [2], but also relaxes the impractical noise-free assumption for independent component analysis (ICA) [2]. Other efforts in literature on the noisy ICA model can be referred to, for instance, [4], [5], and a recent systematic review on ICA and its extensions with noise and temporal dependence structure is referred to [3].

Provided that the noise e is independent of y as assumed by FA and NFA, x generated by (1) can then be modeled by

$$p(x) = \int p_e(x - Ay)p(y)dy. \tag{2}$$

Since conventional FA assumes each y_j as Gaussian, the model can be analytically estimated by the well-known expectation–maximization (EM) algorithm [6]. In contrast, the EM algorithm cannot be applied in a similar way for the NFA model to estimate each non-Gaussian factor y_j as the integral in (2) is analytically intractable [7]. To remove the integral, one has to resort to either numerical integration or Monte-Carlo stochastic integration. Yet, such techniques are very computationally intensive [7].

Manuscript received March 26, 2003; revised November 4, 2003. This work was supported by the Research Grant Council of the Hong Kong SAR under Project CUHK 4336/02E. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Richard C. Kavanagh.

The authors are with the Department of Computer Science and Engineering, The Chinese University of Hong Kong, Shatin, N.T., Hong Kong, China (e-mail: zyliu@cse.cuhk.edu.hk).

Digital Object Identifier 10.1109/LSP.2004.828928

Alternatively, each non-Gaussian factor y_j may be modeled by a Gaussian mixture [7]–[9]

$$p(y \mid \zeta) = \prod_j p_j(y_j), \quad p_j(y_j) = \sum_r \alpha_{j,r} G\left(y_j \mid m_{j,r}, \sigma_{j,r}^2\right) \tag{3}$$

where $G(y_j \mid m_{j,r}, \sigma_{j,r}^2)$ denotes a Gaussian probability density function with mean $m_{j,r}$ and variance $\sigma_{j,r}^2$. As discussed by [7], via introducing a missing data to indicate the factor generated by the respective Gaussian component and transforming $p(y)$ in (3) to a mixture of Gaussian products, the EM algorithm can then be performed analytically. Such a technique was also discussed in [9] for estimating the independent factor analysis (IFA) model. Nonetheless, such approach suffers from the curse of dimensionality. That is, the computational complexity grows exponentially with the number of factors [9].

From the perspective of BYY harmony learning [2], [3], by nature the analytically intractable integral can be avoided by finding $\hat{y}(t)$ that maximizes the *a posteriori* for each $x(t)$ as

$$\hat{y}(t) = \arg\max_y \ln[p(x(t) \mid y(t))p(y(t))] \tag{4}$$

where $p(y(t))$ is given by (3) and $p(x(t) \mid y(t)) = G(x(t) \mid Ay(t), \Sigma_e)$ with Σ_e denoting the covariance matrix of e. Since numerical approaches should be adopted to solve the MAP factor estimate problem in (4), this letter compares three suggested numerical approaches in [2] for it. Also, we theoretically prove that the *iterative fixed posteriori approximation* approach is another type of EM-algorithm, based on which we also prove its convergence.

Section II briefly describes the NFA algorithm and three MAP factor estimating approaches. Section III proves that the *iterative fixed posteriori approximation* approach is another type of EM-algorithm, with the proof of its convergence also shown. Section IV empirically analyzes the three MAP factor estimate approaches and shows that the *iterative fixed posteriori approximation* is the optimal approach. Section V concludes the letter.

II. NFA Algorithm: Estimating Map Factors and Three Numerical Approaches

The NFA algorithm proposed in [2] is

Step 1) Fix $\{\alpha_{j,r}, m_{j,r}, \sigma_{j,r}^2\}$ and $\{A, \Sigma_e\}$, find the factor $\hat{y}(t)$ according to (4).

Step 2) Fix $\{A, \Sigma_e\}$ and $y(t)$, update $\alpha_{j,r} = (e^{c_{j,r}})/(\sum_l e^{c_{j,l}})$, $c_{j,r}^{\text{new}} = c_{j,r}^{\text{old}} + \eta \sum_l h_{j,l}[I_{l,r} - \alpha_{j,r}]$, $m_{j,r}^{\text{new}} = m_{j,r}^{\text{old}} + \eta h_{j,r} e_{j,r}$, $\sigma_{j,r}^{2,\text{new}} = \sigma_{j,r}^{2,\text{old}} + \eta(e_{j,r}^2 - \sigma_{j,r}^2)$, $e_{j,r} = \hat{y}_j - m_{j,r}$, $h_{j,r} = (\alpha_{j,r} G(\hat{y}_j \mid m_{j,r}, \sigma_{j,r}^2))/(\sum_l \alpha_{j,l} G(\hat{y}_j \mid m_{j,l}, \sigma_{j,l}^2))$, $I_{i,r} = \begin{cases} 1 & i = r \\ 0 & \text{otherwise} \end{cases}$.

Step 3) Fix $\{\alpha_{j,r}, m_{j,r}, \sigma_{j,r}^2\}$ and $y(t)$, update $A^{\text{new}} = A^{\text{old}} + \eta e(t)\hat{y}(t)^T$, and $\Sigma_e^{\text{new}} = (1-\eta)\Sigma_e^{\text{old}} + \eta e(t)e(t)^T$, where $e(t) = x(t) - A\hat{y}(t)$.

For the BYY harmony learning based NFA algorithm, the Yang machine by nature requires finding the y that maximizes the posterior, as in step 1. Then, based on the sample $x(t)$ and the estimated $\hat{y}(t)$ in step 1, steps 2 and 3 update the remaining parameters via the typical least mean square (LMS) criterion. Moreover, as discussed in [2], due to the least complexity property of the BYY harmony learning, the NFA algorithm is capable of selecting the number of factors. Furthermore, the problem of local optimization could be alleviated by the two newly introduced regularization techniques—data smoothing and normalization learning. Readers interested are referred to [2] and [10] for further details.

Although the MAP task in step 1 plays a key role for the whole learning process, it is analytically intractable. Below we review the three numerical approaches adopted for estimating the MAP factor scores.

A. Iterative Fixed Posteriori Approximation Approach

The so-called *fixed posteriori approximation* proposed by [2] is

$$\hat{y} = (A^T {\Sigma_e}^{-1} A + \text{diag}[b_1, \ldots, b_k])^{-1}[A^T {\Sigma_e}^{-1} x + d] \quad (5)$$

where $b_j = \sum_r (h_{j,r})/(\sigma_{j,r}^2), d_j = \sum_r (h_{j,r} m_{j,r})/(\sigma_{j,r}^2)$, and the posteriori

$$h_{j,r} = \frac{\alpha_{j,r} G\left(y_j \,|\, m_{j,r}, \sigma_{j,r}^2\right)}{\sum_r \alpha_{j,r} G\left(y_j \,|\, m_{j,r}, \sigma_{j,r}^2\right)} \quad (6)$$

is approximately regarded as being irrelevant to y.

Based on it, the following iterative procedure can be used to find a solution of (4).

Step 1) Fix h, update y according to (5).

Step 2) Fix y, update h according to (6).

B. Gradient Descent Approach

The derivative of $f(y) = \ln[p(x \,|\, y)p(y)]$ with respect to y is

$$\nabla_y(f) = A^T {\Sigma_e}^{-1}(x - Ay) + g \quad (7)$$

where $g = [g_1, \ldots, g_m]^T$ with $g_j = \sum_r h_{j,r}(m_{j,r} - y_j)/(\sigma_{j,r}^2)$, and $h_{j,r}$ is the same as in Section II-A.

C. Conjugate Gradient Approach

The *conjugate gradient approach* is considered superior to the quasi-Newton method in the sense that it avoids the difficulty of having to compute the Hessian matrix, but still possesses the super-linear rate of convergence. The algorithm is:

Initialize y, set $s = -g = -\nabla_y(f)$ by (7).

Step 1) $y^{\text{new}} = y^{\text{old}} + \alpha s^{\text{old}}, g^{\text{new}} = \nabla_{y^{\text{new}}}(f)$.

Step 2) $s^{\text{new}} = -g^{\text{new}} + \beta s^{\text{old}}, \beta = ((g^{\text{new}})^T g^{\text{new}})/((g^{\text{old}})^T g^{\text{old}})$.

In this letter, we choose the learning rate α based on [11] which satisfies the well-known Wolfe condition.

III. Proof of the Iterative Fixed Posteriori Approximation as a Type of EM-Algorithm

The EM algorithm [12] is commonly adopted for tackling the incomplete-data problem. In short, the E-step concerns finding the expectation of the complete-data based cost function with respective to the "missing" data, which is typically the log-likelihood function while the M-step concerns updating the parameters via maximization of the expectation.

Here, the EM algorithm is adopted for solving the MAP factor estimate problem

$$\hat{y} = \arg\max_y \ln p(x \,|\, y)p(y) = \arg\max_y \ln p(x, y) \quad (8)$$

where $p(y) = \prod_j \sum_r \alpha_{j,r} G(y_j \,|\, m_{j,r}, \sigma_{j,r}^2), p(x|y) = G(x \,|\, Ay, \Sigma_e)$. It should be noted that unlike the conventional EM algorithm which is applied to maximize the likelihood function with respect to the parameters, here the EM algorithm is adopted to maximize the function with respect to the factor.

To show how the EM-algorithm can be derived, we regard the data $\{x, y\}$ incomplete and introduce a "missing" indicator $z = \{z_j\}_{j=1}^m$, defined as $z_j = r$ iff y_j was generated by component r, such that $P(z_j = r) = \alpha_{j,r}$.

Moreover, since each z_j is related only to the mutually independent $y_j, \{z_j\}_{j=1}^m$ are also mutually independent, which makes $P(z) = \prod_{j=1}^m P(z_j)$. Consequently, the joint distribution of the complete data set $\omega = \{x, y, z\}$ can be obtained as [12] $p(\omega) = p(x \,|\, y, z)p(y \,|\, z)P(z) = p(x \,|\, y)p(y \,|\, z)P(z)$, since given y, according to the definition of z, x is independent of z. Meantime, because $y = \{y_j\}_{j=1}^m$ and $z = \{z_j\}_{j=1}^m$ both are mutually independent, we have $p(y \,|\, z) = \prod_{j=1}^m p(y_j \,|\, z_j)$. Because z_j indicates y_j is generated by the respective component, thus, given z_j the conditional density of y_j is

$$p(y_j \,|\, z_j = r) = G\left(y_j \,|\, m_{j,r}, \sigma_{j,r}^2\right). \quad (9)$$

Consequently, the logarithm of the joint distribution $p(x, y, z)$ can be obtained as

$$\begin{aligned} \ln p(\omega) &= \ln p(x \,|\, y) + \ln p(y \,|\, z) + \ln P(z) \\ &= \ln G(x \,|\, Ay, \Sigma_e) + \sum_{j=1}^m (\ln p(y_j \,|\, z_j) + \ln P(z_j)). \end{aligned}$$

Then, the EM algorithm needs to find the expectation of the complete data joint $\log p(\omega)$ function shown above with respect to the the "missing" data z given y^* as

$$\begin{aligned} Q(y, y^*) &= E_{z \,|\, y^*}[\ln p(\omega) \,|\, y^*] \\ &= \ln G(x \,|\, Ay, \Sigma_e) \\ &\quad + \sum_{j=1}^m E_{z_j y_j^*}[\ln p(y_j \,|\, z_j) + \ln P(z_j)] \\ &= \ln G(x \,|\, Ay, \Sigma_e) \\ &\quad + \sum_{j=1}^m \sum_{i=1}^k \ln\left[G\left(y_j \,|\, m_{j,i}, \sigma_{j,i}^2\right) \alpha_{j,i}\right] P(z_j = i \,|\, y_j^*). \end{aligned}$$

Where the second equality is also due to the mutual independence for $y = \{y_j\}_{j=1}^m$ and $z = \{z_j\}_{j=1}^m$, for which the detail is

$$\begin{aligned} &E_{z\,|\,y^*}[\ln p(y_j\,|\,z_j) + \ln P(z_j)] \\ &\quad = \sum_{z_1=1}^{k} \cdots \sum_{z_j=1}^{k} \cdots \sum_{z_m=1}^{k} [\ln p(y_j\,|\,z_j) + \ln P(z_j)] \\ &\quad\quad \times \prod_{l=1}^{m} P(z_l\,|\,y_l^*) \\ &\quad = \sum_{z_j=1}^{k} [\ln p(y_j\,|\,z_j) + \ln P(z_j)] P(z_j\,|\,y_j^*) \\ &\quad = E_{z_j\,|\,y_j^*}[\ln p(y_j z_j) + \ln P(z_j)]. \end{aligned}$$

Given the current y_j^*, the probability $P(z_j = i\,|\,y_j^*)$, which we denote by $\gamma_{j,i}$ is

$$\gamma_{j,i} \triangleq P(z_j = i\,|\,y_j^*) = \frac{\alpha_{j,i} G\left(y_j^*\,|\,m_{j,i}, \sigma_{j,i}^2\right)}{\sum_r \alpha_{j,r} G\left(y_j^*\,|\,m_{j,r}, \sigma_{j,r}^2\right)}. \quad (11)$$

Thus, the function $Q(y, y^*)$ can be finally obtained as

$$\begin{aligned} Q(y, y^*) = \ln G(x\,|\,Ay, \Sigma_e) + \sum_{j=1}^{m}\sum_{r=1}^{k} \gamma_{j,r} \ln \alpha_{j,r} \\ \times G\left(y_j\,|\,m_{j,r}, \sigma_{j,r}^2\right). \quad (12) \end{aligned}$$

Since the M-step requires finding y^{new} to maximize the above expectation $Q(y, y^*)$ as $y^{\text{new}} = \arg\max_y Q(y, y^*)$, we differentiate $Q(y, y^*)$ with respect to y to get $(\partial Q(y, y^*))/(\partial y) = (\sum_{r=1}^k (m_{1,r})/(\sigma_{1,r}^2)\gamma_{1,r}, \ldots, \sum_{r=1}^k (m_{m,r})/(\sigma_{m,r}^2)\gamma_{m,r})^T + y \cdot \text{diag}[b_1, \ldots, b_k] + A^T\Sigma_e^{-1}(x - Ay)$, where $b_j = \sum_{r=1}^k (\gamma_{j,r})/(\sigma_{j,r}^2) = \sum_{r=1}^k (h_{j,r})/(\sigma_{j,r}^2)$. Thus, $(\partial Q(y, y^*))/(\partial y) = 0$ makes

$$y^{\text{new}} = \left(A^T\Sigma_e^{-1}A + \text{diag}[b_1, \ldots, b_k]\right)^{-1}\left[A^T\Sigma_e^{-1}x + d\right] \quad (13)$$

where $d \in \mathbb{R}^m$ is with element $d_j = \sum_{r=1}^k (\gamma_{j,r} m_{j,r})/(\sigma_{j,r}^2) = \sum_{r=1}^k (h_{j,r} m_{j,r})/(\sigma_{j,r}^2)$ and $h_{j,r}$ is defined by (6). So (13) is exactly (5). Previously we assumed the number of components of mixture models all equals k. However, (13) (including b and d) can be extended to the case with different component numbers.

So in the E-step, we just need to calculate $\gamma_{j,r} = h_{j,r}$ according to (11) or (6) while in the M-step, update y according to (13). It is exactly the *iterative fixed posteriori approximation* discussed previously. Based on the proof in [12], we proceed to show that for each iteration of the EM algorithm, $\ln p(x, y^{\text{new}}) \geq \ln p(x, y^*)$ can be guaranteed for $\ln p(x, y)$. This shows the convergence of this type of EM algorithm. First we introduce the conditional density of ω given x, y as

$$k(\omega\,|\,x, y) = \frac{p(\omega)}{p(x, y)} \Rightarrow \ln p(x, y) = \ln p(\omega) - \ln k(\omega\,|\,x, y)$$

and then define

$$\begin{aligned} H(y, y^*) &= E_{z\,|\,y^*}[\ln k(\omega\,|\,x, y)\,|\,y^*] \\ &\Rightarrow \ln p(x, y) = Q(y, y^*) - H(y, y^*). \quad (14) \end{aligned}$$

TABLE I
COMPARATIVE RESULTS ON 3 MAP APPROACHES

	Mean Square Error		Noise covariance	Time-cost	Time-cost
	signal 1	signal 2	$\hat{\Sigma}_e$	of NFA	of a MAP
#1	0.0687	0.0086	$\begin{pmatrix} 0.017 & -0.002 \\ -0.002 & 0.020 \end{pmatrix}$	17.36s	3ms
#2	0.1397	0.1667	$\begin{pmatrix} 0.021 & 0.001 \\ 0.001 & 0.018 \end{pmatrix}$	232.49s	61ms
#3	0.1004	0.0865	$\begin{pmatrix} 0.012 & -0.003 \\ -0.003 & 0.026 \end{pmatrix}$	62.43s	13ms
#4	0.4782	0.3347	$\begin{pmatrix} 0.021 & 0.015 \\ 0.015 & 0.016 \end{pmatrix}$	5.21s	1ms

#1,#2, #3, #4 denote *iterative fi xed posteriori approximation, gradient descent, conjugate gradient, gaussian approximation* respectively.

Where the expectation of $\ln p(x, y)$ with respect to z is equal to itself because it doesn't involve z. After one iteration by the EM algorithm, the change of $\ln p(x, y)$ is: $\ln p(x, y^{\text{new}}) - \ln p(x, y^*) = Q(y^{\text{new}}, y^*) - Q(y^*, y^*) + H(y^*, y^*) - H(y^{\text{new}}, y^*)$. According to the definition of M-step, we have $Q(y^{\text{new}}, y^*) \geq Q(y^*, y^*)$. Based on the well-known Jensen's inequality and the concave property of the logarithm function, we have for any given pair $(y, y^*), H(y, y^*) \leq H(y^*, y^*)$ and the equality holds iff $k(\omega\,|\,x, y) = k(\omega\,|\,x, y^*)$ almost everywhere (the Lemma 1 given in [12]). Thus, it can be guaranteed that

$$\ln p(x, y^{\text{new}}) - \ln p(x, y^*) \geq 0 \quad (15)$$

for EM iteration. Thus, such an EM-algorithm can ensure the algorithm converge to, at lease, a local optimal value.

IV. EMPIRICAL COMPARISONS ON THREE MAP FACTOR ESTIMATE APPROACHES

Comparisons in both dimensions of effectiveness and efficiency are based on the synthetic two-factor model, with a typical factor estimating process discussed in detail.

A. Initialization for the Numerical Approaches

All of the three approaches above require a proper initialization. We choose the following *Gaussian approximation* [2]:

$$y = A_y^{-1}(x_A + \Lambda^{-1}d) \quad (16)$$

where $x_A = A^T\Sigma_e^{-1}x, A_y = A^T\Sigma_e^{-1}A + \Lambda^{-1}$, and $d = [d_1, \ldots, d_k]^T, \Lambda = \text{diag}[\lambda_1, \ldots, \lambda_k]$, with $d_j = \Sigma_r \alpha_{j,r} m_{j,r}$ and $\lambda_j = \Sigma_r \alpha_{j,r}[\sigma_{j,r}^2 + (m_{j,r} - d_j)^2]$. The key idea here is that a Gaussian density is used to approximate the Gaussian mixture in (3) such that an analytic solution can be obtained by (16).

B. Simulation Via Synthetic Two-Factor Model

We consider 50 observations generated according to (1) with the parameters as follows, $A = \begin{pmatrix} 1.2 & -1.0 \\ 0.6 & 1.4 \end{pmatrix}, y = [y_1 y_2]^T$, with y_1 being generated from a uniform distribution in the interval $[-0.5 0.5]$, and y_2 from a bimodal symmetric $\beta(2, 2)$ distribution with mean removed, e is randomly generated with pdf $G(e\,|\,0, 0.01\text{I}_2)$, where and hereafter I_n denotes the n-dimensional identity matrix.

The experiment is repeated 20 times with random initializations. All signals are normalized with zero mean and unit variance. The mean square error (MSE) (average over 20 runs) between the normalized estimated state $\hat{y}_j$ and true state y_j are listed in Column 2 of Table I. Three typical estimated noise covariance matrices are listed in Column 3. The cost-time (average over 20 runs) of the three MAP approaches are listed in Column

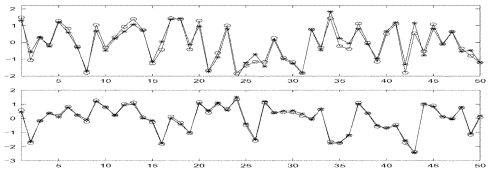

Fig. 1. Sources recovered by *iterative fixed posteriori approximation*, where the source signals are denoted by "o" and recovered signals by "$*$".

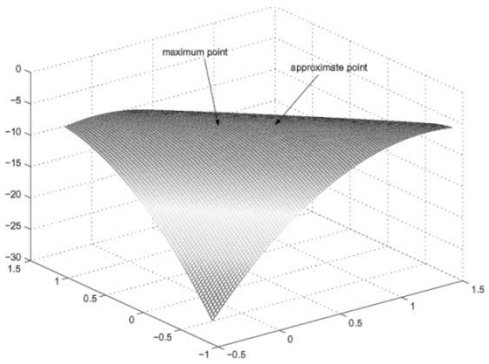

Fig. 2. Typical MAP problem: coordinates making up the horizontal plane are y_1 and y_2, and the erected one is $\ln[p(x \mid y)p(y)]$.

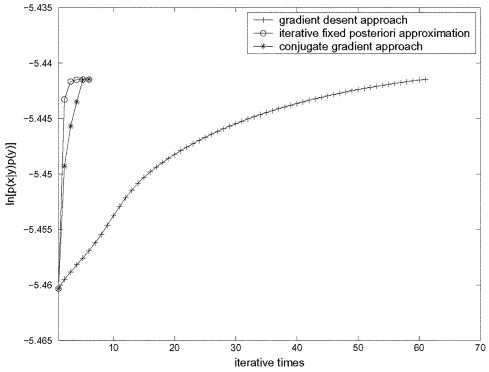

Fig. 3. Convergence process of the three MAP approaches.

4 of Table I. For brevity we only illustrate the results obtained by the *iterative fixed posteriori approximation* (Fig. 1).

In addition, we also directly use the *Gaussian approximation* as the MAP factor estimate. The highest MSE (as shown in Table I) resulted implies the failure of the direct use of the *Gaussian approximation* as the factor estimating approach.

C. Analysis of the MAP Factor Estimating Process

Consider the typical synthetic two-factor model. The optimal MAP factor score as denoted by the maximum point and the value initialized by *Gaussian approximation* are shown by two arrows in Fig. 2. The three MAP estimating processes are shown in Fig. 3, and the corresponding time-cost are respectively 3, 61, and 13 ms (Column 5 of Table I).

As shown in Fig. 3, all three approaches can approximately arrive at the optimal point starting with the same initialization. This explains why the estimating accuracy of the three approaches does not differ greatly from each other. However, the process of the *gradient descent approach* is too long to get to the optimal solution due to its linear convergence speed. For instance, there is still a small error even after 62 iterations in Fig. 3. The convergence rate of *iterative fixed posteriori approximation*, a type of EM-algorithm, is generally linear [13] while the *conjugate gradient approach* has at least a linear rate. In Fig. 3, however, the *iterative fixed posteriori approximation* converged as quickly as the *conjugate gradient approach*. This may be because the *Gaussian approximation* usually makes the initialization close to the true result, and thus echoes the conclusions in [13] that state "for Gaussian mixtures locally around the true solution...and when the overlap in the mixture is small...the convergence rate for the EM-algorithm tends to be asymptotically superlinear." That is, the *iterative fixed posteriori approximation* here shares this nature. Also, it should be mentioned that the process of finding a proper learning rate in *conjugate gradient approach* is time-consuming although the convergence rates of the two approaches are close. As expected, the *conjugate gradient approach* is worse than the *iterative fixed posteriori approximation* on computing efficiency.

We can conclude that the *iterative fixed posteriori approximation* approach is the most optimal considering both estimating accuracy and computing efficiency.

V. Conclusion

We comparatively study the three suggested MAP factor estimates approaches for NFA and empirically find that the so-called *iterative fixed posteriori approximation* approach is the most optimal when both estimating accuracy and computing efficiency are taken into account. Specifically, the *iterative fixed posteriori approximation* approach is proved to be another type of EM-algorithm, with the proof of convergence shown.

References

[1] L. Xu, "Bayesian Kullback Ying-Yang dependence reduction theory," *Neurocomputing*, vol. 22, pp. 81–111, 1998.

[2] ——, "Byy harmony learning, independent state space and generalized APT financial analysis," *IEEE Trans. Neural Networks*, vol. 12, no. 4, pp. 822–849, 2001.

[3] ——, "Mining dependence structures (ii): An independence analysis perspective," in *Proc. IEEE ICDM'02 Workshop on The Foundation of Data Mining and Discovery*, 2002.

[4] A. Hyvarinen, J. Karhunen, and E. Oja, *Independent Component Analysis*. New York: Wiley, 2001.

[5] A. Cichocki and S. I. Amari, *Adaptive Blind Signal and Image Processing*. New York: Wiley, 2002.

[6] D. Rubi and D. Thayer, "EM algorithm for ML factor analysis," *Psychometrika*, vol. 57, pp. 69–76, 1976.

[7] E. Moulines, J. Cardoso, and E. Gassiat, "Maximum likelihood for blind separation and deconvolution of noisy signals using mixture models," in *Proc. ICASSP'97*, 1997, pp. 3617–3620.

[8] L. Xu, "Bayesian Ying-Yang system and theory as a unified statistical learning approach (iii): Models and algorithms for dependence reduction, data dimension reduction, ICA and supervised learning," in *Theoretical Apsects of Neural Computation: A Multidisciplinary Perspective*, K. M. Wong *et al.*, Ed. New York: Springer-Verlag, 1997, vol. 43–60.

[9] H. Attias, "Independent factor analysis," *Neural Computation*, vol. 11, pp. 803–851, 1999.

[10] L. Xu, "Best harmony, unified RPCL and automated model selection for unsupervised and supervised learning on Gaussian mixtures, three-layer nets and ME-RBF-SVM models," *Int. J. Neural Syst.*, vol. 11, pp. 43–70, 2001.

[11] R. Fletcher, *Practical Methods of Optimization*, 2nd ed. New York: Wiley, 1987.

[12] A. P. Dempster, N. M. Laird, and D. B. Rubin, "Maximum-likelihood from incomplete data via the EM algorithm," *J. R. Statist. Soc. B*, vol. 39, pp. 1–38, 1977.

[13] J. Ma, L. Xu, and M. I. Jordan, "Asymptotic convergence rate of the EM algorithm for Gaussian mixtures," *Neural Comput.*, vol. 12, no. 12, pp. 2881–2907, 2000.