Contents lists available at ScienceDirect

Neural Networks

journal homepage: www.elsevier.com/locate/neunet

A note on variational Bayesian factor analysis

Jian-hua Zhao [a,b,*], Philip L.H. Yu [a]

[a] *Department of Statistics and Actuarial Science, The University of Hong Kong, Hong Kong*

[b] *School of Mathematics and Statistics, Yunnan University, Kunming 650091, China*

ARTICLE INFO

Article history:
Received 31 March 2008
Received in revised form
17 November 2008
Accepted 18 November 2008

Keywords:
Factor analysis
VB
BIC

ABSTRACT

Existing works on variational bayesian (VB) treatment for factor analysis (FA) model such as [Ghahramani, Z., & Beal, M. (2000). Variational inference for Bayesian mixture of factor analysers. In *Advances in neural information proceeding systems*. Cambridge, MA: MIT Press; Nielsen, F. B. (2004). Variational approach to factor analysis and related models. *Master's thesis*, The Institute of Informatics and Mathematical Modelling, Technical University of Denmark.] are found theoretically and empirically to suffer two problems: ① penalize the model more heavily than BIC and ② perform unsatisfactorily in low noise cases as redundant factors can not be effectively suppressed. A novel VB treatment is proposed in this paper to resolve the two problems and a simulation study is conducted to testify its improved performance over existing treatments.

1. Introduction

Factor analysis (FA) is a powerful multivariate analysis technique that uncovers the latent common characteristics (or factors) among a set of variables and has been widely used for dimension reduction, feature extraction, time series prediction and so forth.

There has been a long-standing interest in Bayesian treatment for FA since it can avoid the over-fitting problem of maximum likelihood (ML) method and lead to determine the number of factors automatically. However, conventional Baysian treatment usually depends on sampling-based techniques, which is computationally demanding and is often limited to small-scale problems. In recent years, Baysian approaches have achieved increasing applications, due to advances of computationally more efficient approximate inference methods such as *Variational Bayesian* (VB) that can tackle not only small-scale problems but also large-scale ones.

VB has been applied to FA model and mixtures of FAs or related models in Beal (2003) and Ghahramani and Beal (2000) and further investigated in Nielsen (2004). Comparing with Bayesian information criterion (BIC), however, Nielsen (2004, p. 58) observes that VB for FA (henceforth denoted as VBFA1) tends to obtain a under-fitting result; (Beal, 2003, p. 142–p. 143) also observes that a mixture version of VBFA1 tends to penalize model complexity too heavily, i.e., choose a model with relatively fewer components. Since it is well known that BIC is simply a crude approximation to Baysian evidence, it seems reasonable to see that VB should perform comparable to or better than BIC. These two somewhat surprising findings motivate us to examine whether there are problems in the formulation of VBFA1.

In this paper, we find two problems associated with VBFA1: ① the large sample limit of VBFA1 does not correspond to BIC and has a heavier penalty than BIC. This means that VBFA1 contradicts with the general theoretical result on latent variable models developed in Attias (1999) that BIC emerges as a limiting case of VB; ② the performance of VBFA1 in a low noise case is not very satisfactory since redundant factors can not be effectively suppressed.

To resolve the above two problems, we propose a new VB treatment in this paper. The remainder of the paper is organized as follows: Section 2 gives a review of FA model and some results recently obtained in Zhao, Yu, and Jiang (2008), which motivates our solution to the second problem of VBFA1; Section 3 details the two problems in VBFA1 and Section 4 proposes our new VB treatment. Section 5 conducts a simulation study to assess its performance; Section 6 closes the paper with a conclusion.

2. FA model

Suppose that a d-dimensional data vector $\mathbf{x}_i$ in an i.i.d sample $\mathbf{X}_N = \{\mathbf{x}_i\}_{i=1}^N$ follows a q-factor model:

$$\begin{cases} \mathbf{x}_i = \mathbf{A}\mathbf{y}_i + \boldsymbol{\mu} + \boldsymbol{\epsilon}_i, \\ \mathbf{y}_i \sim \mathcal{N}(\mathbf{0}, \mathbf{I}), \quad \boldsymbol{\epsilon}_i \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Psi}), \end{cases} \tag{1}$$

where $\boldsymbol{\mu}$ is a d-dimensional mean vector, $\mathbf{A} = (\mathbf{a}_1, \ldots, \mathbf{a}_q)$ is a $d \times q$ factor loadings matrix, $\mathbf{y}_i$ is a q-dimensional latent vector,

[*] Corresponding address: Department of Statistics and Actuarial Science, The University of Hong Kong, Meng Wah Complex, Room 518, 5/F, Hong Kong. Tel.: +852 28597942; fax: +852 28589041.

E-mail addresses: jhzhao.ynu@gmail.com (J.-h. Zhao), plhyu@hku.hk (P.L.H. Yu).

0893-6080/$ – see front matter © 2008 Elsevier Ltd. All rights reserved.
doi:10.1016/j.neunet.2008.11.002

representing those factors common to all components of $\mathbf{x}_i$, and $\mathbf{\Psi} = \text{diag}\{\psi_1, \psi_2, \ldots, \psi_d\}$ is a positive diagonal matrix. We use $\mathbf{I}$ to denote an unit matrix whose dimension should be apparent from the context.

Under the model (1), $\mathbf{x}_i \sim \mathcal{N}(\boldsymbol{\mu}, \mathbf{\Psi} + \mathbf{A}\mathbf{A}')$. Let

$$\bar{\mathbf{x}} = \frac{1}{N}\sum_{i=1}^{N}\mathbf{x}_i, \qquad \mathbf{S} = \frac{1}{N}\sum_{i=1}^{N}(\mathbf{x}_i - \bar{\mathbf{x}})(\mathbf{x}_i - \bar{\mathbf{x}})'$$

be the sample mean vector and sample covariance matrix of $\mathbf{x}$. As the global maximum likelihood estimator (MLE) of $\boldsymbol{\mu}$ is trivially the sample mean $\bar{\mathbf{x}}$, the MLE of $\mathbf{A}$ and $\mathbf{\Psi}$ can be obtained by maximizing

$$\mathcal{L}(\boldsymbol{\theta}) = -\frac{N}{2}\left\{\ln|\mathbf{\Psi} + \mathbf{A}\mathbf{A}'| + \text{tr}\left([\mathbf{\Psi} + \mathbf{A}\mathbf{A}']^{-1}\mathbf{S}\right)\right\} \tag{2}$$

via e.g. the popular EM algorithm (Dempster, Laird, & Rubin, 1977). Despite the advantage of simplicity and stability, convergence of EM is only linear and can be painfully slow. Recently, Zhao et al. (2008) propose a conditional maximization (CM) algorithm, which shares the same advantage of EM but possesses quadratic convergence. The CM-step 1 in this algorithm is to maximize (2) over $\mathbf{A}$ given $\mathbf{\Psi}$. In addition to achieve faster convergence, this step provides more insight into the property of MLE of A. We write the theoretical result in this step as Theorem 1 below.

Theorem 1. *Let $\mathbf{S}_N$ denote the normalized sample covariance matrix: $\mathbf{S}_N = \mathbf{\Psi}^{-1/2}\mathbf{S}\mathbf{\Psi}^{-1/2}$ and $(\lambda_k, \mathbf{u}_k)$, $k = 1, \ldots, d$ be its eigenvalue-eigenvector pairs so that $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_d$. Given that $\mathbf{\Psi}$ is known, the MLE of $\mathbf{A}$ is obtained by*

$$\mathbf{A} = \mathbf{\Psi}^{1/2}\mathbf{U}_{q'}\left(\mathbf{\Lambda}_{q'} - \mathbf{I}\right)^{1/2}\mathbf{V}, \tag{3}$$

where q' is the unique integer satisfying $\lambda_{q'} > 1 \geq \lambda_{q'+1}$, $\mathbf{U}_{q'} = (\mathbf{u}_1, \mathbf{u}_2, \ldots, \mathbf{u}_{q'})$ and $\mathbf{\Lambda}_{q'} = \text{diag}(\lambda_1, \lambda_2, \ldots, \lambda_{q'})$. The matrix $\mathbf{V}$ can be arbitrarily chosen except for the requirement $\mathbf{V}\mathbf{V}' = \mathbf{I}$.

If $\mathbf{V}$ is set to be the first q' rows of a $q \times q$ identity matrix $\mathbf{I}$, from (3) we have

$$\mathbf{A}'\mathbf{\Psi}^{-1}\mathbf{A} = \mathbf{\Lambda}_{q'} - \mathbf{I}, \quad \text{or equivalently, } \mathbf{a}_k'\mathbf{\Psi}^{-1}\mathbf{a}_k = \lambda_k - 1. \tag{4}$$

Substituting (3) into (2), we obtain

$$-\frac{2}{N}\cdot\mathcal{L} = \sum_{k=1}^{q'}(\ln\lambda_k - \lambda_k + 1) + \ln|\mathbf{\Psi}| + \sum_{j=1}^{d}\lambda_j. \tag{5}$$

Given $\mathbf{\Psi}$ in (5), $\ln|\mathbf{\Psi}| + \sum_{j=1}^{d}\lambda_j$ is obviously a constant. Since the function $f(\lambda) = \ln\lambda - \lambda + 1$ is negative and strictly decreasing in the interval $(1, \infty)$, the closer λ_k (among $\lambda_1, \ldots, \lambda_{q'}$) in (5) is to 1, the less the factor k contributes to $\mathcal{L}$. Therefore, Theorem 1 in fact suggests using $\mathbf{a}_k'\mathbf{\Psi}^{-1}\mathbf{a}_k$ to measure the significance of factor k from the maximum likelihood perspective. Note that up to a constant, (5) equals to equation (18) in Jöreskog (1967, p. 448).

3. VBFA1

In this section, we briefly review the existing VB treatments for FA model (Ghahramani & Beal, 2000; Nielsen, 2004), for convenience, denoted as VBFA1 henceforth.

To treat FA model in a Bayesian way, VBFA1 introduces a prior distribution $p(\boldsymbol{\theta})$ over model parameter $\boldsymbol{\theta} = (\boldsymbol{\mu}, \mathbf{A}, \boldsymbol{\varphi})$, where $\boldsymbol{\varphi} = \mathbf{\Psi}^{-1}$, and make prediction by marginalizing over $\boldsymbol{\theta}$ with respect to the posterior distribution $p(\boldsymbol{\theta}|\mathbf{X})$. In order to determine q automatically during Bayesian learning process, VBFA1 utilizes the idea of automatic relevance determination (ARD) by further introducing a *hierarchical* prior $p(\mathbf{A}|\boldsymbol{\omega})$, where $\boldsymbol{\omega} = (\omega_1, \ldots, \omega_q)$, over factor loading matrix $\mathbf{A}$:

$$p(\mathbf{A}|\boldsymbol{\omega}) = \prod_{k=1}^{q}\left(\frac{\omega_k}{2\pi}\right)^{d/2}\exp\left(-\frac{\omega_k}{2}\mathbf{a}_k'\mathbf{a}_k\right). \tag{6}$$

If the posterior distribution over hyperparameter ω_k concentrates on a large value, the corresponding column $\mathbf{a}_k$ tends to be close to zero. Thus unnecessary factors will be effectively switched off and the number of factors q can then be determined. Such resulting fully Bayesian treatment is however intractable and hence VBFA1 resorts to use VB to make approximate inference.

VB aims to optimize a lower bound $\mathcal{F}$ of the model evidence $\ln p(\mathbf{X})$, which is obtained by Jensen's inequality:

$$\begin{aligned}\ln p(\mathbf{X}) &= \ln\int p(\mathbf{X}, \mathbf{Y}, \boldsymbol{\theta})\mathrm{d}\mathbf{Y}\mathrm{d}\boldsymbol{\theta} \\ &\geq \int q(\mathbf{Y}, \boldsymbol{\theta})\ln\frac{p(\mathbf{X}, \mathbf{Y}, \boldsymbol{\theta})}{q(\mathbf{Y}, \boldsymbol{\theta})}\mathrm{d}\mathbf{Y}\mathrm{d}\boldsymbol{\theta} = \mathcal{F}(q).\end{aligned} \tag{7}$$

Here $\boldsymbol{\theta} = \{\theta_i\}$ denotes the set of parameters and hyperparameters. Eq. (7) can also be expressed in terms of Kullback–Leibler (KL) divergence.

$$\begin{aligned}\text{KL}(q \parallel p) &= \ln p(\mathbf{X}) - \mathcal{F}(q) \\ &= -\int q(\mathbf{Y}, \boldsymbol{\theta})\ln\frac{p(\mathbf{Y}, \boldsymbol{\theta}|\mathbf{X})}{q(\mathbf{Y}, \boldsymbol{\theta})}\mathrm{d}\mathbf{Y}\mathrm{d}\boldsymbol{\theta}.\end{aligned} \tag{8}$$

It can be seen from (8) that maximizing $\mathcal{F}$ is equivalent to minimizing KL divergence between q and true posterior $p(\mathbf{Y}, \boldsymbol{\theta}|\mathbf{X})$. Since true posterior p is usually intractable, VB utilizes a tractable q to approximate p. This is achieved by assuming that q factorizes over latent variable $\mathbf{Y}$ and component variables θ_i of $\boldsymbol{\theta}$

$$q(\mathbf{Y}, \boldsymbol{\theta}) = q(\mathbf{Y})\prod_i q(\theta_i). \tag{9}$$

According to Bishop (2006), the use of factorization in (9) is purely for achieving tractability and less factorization could achieve a tighter lower bound $\mathcal{F}$ in (7). Substituting (9) into (7) and maximizing $\mathcal{F}$ over $q(\mathbf{Y})$ and $q(\theta_i)$ leads to the following VBEM updating steps:

- VBE-step:

$$q(\mathbf{Y}) = \prod_n q(\mathbf{y}_n) \propto \prod_n \exp\langle\ln p(\mathbf{x}_n, \mathbf{y}_n, \boldsymbol{\theta})\rangle_{q(\boldsymbol{\theta})}. \tag{10}$$

- VBM-step:

$$q(\theta_i) \propto \exp\langle\ln p(\mathbf{X}, \mathbf{Y}, \boldsymbol{\theta})\rangle_{q(\mathbf{Y})\prod_{j\neq i} q(\theta_j)} \quad \forall i. \tag{11}$$

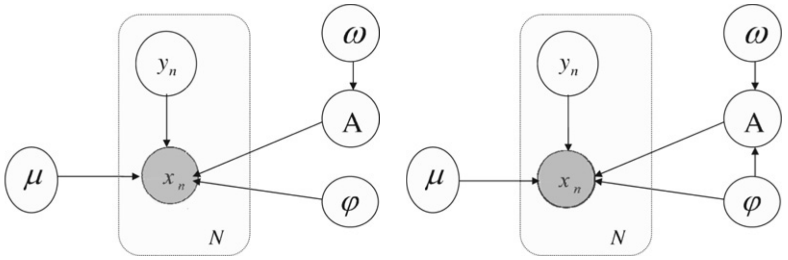
In (10), the factorization of $q(\mathbf{Y})$ into $q(\mathbf{y}_n)$ is because the data $\mathbf{y}_n, i = 1, \ldots, N$ are i.i.d.. $\langle\cdot\rangle_{q(\cdot)}$ denotes an expectation taken with respect to the distribution $q(\cdot)$. For notation convenience, the distribution $q(\cdot)$ will be dropped in what follows if it is apparent from the context.

The probabilistic graphical model of VBFA1 is shown in Fig. 1(a). For tractability, VBFA1 utilizes conjugate priors for the other nodes:

$$p(\boldsymbol{\mu}|\mathbf{0}, \beta^{-1}\mathbf{I}) = \mathcal{N}(\boldsymbol{\mu}|\mathbf{0}, \beta^{-1}\mathbf{I}), \tag{12}$$

$$p(\boldsymbol{\omega}|\mathbf{a}^{\omega}, \mathbf{b}^{\omega}) = \prod_{k=1}^{q}\Gamma(\omega_k|a^{\omega}, b^{\omega}), \tag{13}$$

$$p(\boldsymbol{\varphi}|\mathbf{a}^{\varphi}, \mathbf{b}^{\varphi}) = \prod_{j=1}^{d}\Gamma(\varphi_j|a^{\varphi}, b^{\varphi}), \tag{14}$$



(a) VBFA1.

(b) VBFA2.

Fig. 1. Probabilistic graphical models for VBFA1 and VBFA2.

where $\Gamma(\cdot)$ stands for a Gamma distribution. For simplicity, VBFA1 assumes that $q(\boldsymbol{\theta})$ is factorized into

$$q(\boldsymbol{\theta}) = q(\boldsymbol{\mu})q(\mathbf{A})q(\boldsymbol{\varphi})q(\boldsymbol{\omega}). \tag{15}$$

Due to the choice of conjugate priors and factorization assumption (15), it is easy to derive the VBE-step and VBM-steps in VBFA1. To save space, the expressions for $q(\boldsymbol{\varphi})$, $q(\boldsymbol{\mu})$ and $q(\mathbf{Y})$ are omitted here. $q(\mathbf{A})$ and $q(\boldsymbol{\omega})$ are given below for later use.

$$q(\mathbf{A}) = \prod_j \mathcal{N}(\tilde{\mathbf{a}}_j|\mathbf{m}_{\mathbf{a}}^{(j)}, \boldsymbol{\Sigma}_{\mathbf{a}}^{(j)}), \tag{16}$$

$$q(\boldsymbol{\omega}) = \prod_k \Gamma(\omega_k|\tilde{a}_k^{\omega}, \tilde{b}_k^{\omega}), \tag{17}$$

where $\tilde{\mathbf{a}}_k$ is a column vector corresponding to k-th row of $\mathbf{A}$, and

$$\boldsymbol{\Sigma}_{\mathbf{a}}^{(j)} = \langle\varphi_j\rangle^{-1}\boldsymbol{\Delta}_1^{-1}, \qquad \mathbf{m}_{\mathbf{a}}^{(j)} = \boldsymbol{\Delta}_1^{-1}\left(\sum_i \langle\mathbf{y}_i\rangle(\mathbf{x}_{ij} - \langle\mu_j\rangle)\right), \tag{18}$$

$$\tilde{a}_k^{\omega} = a_k^{\omega} + \frac{d}{2}, \qquad \tilde{b}_k^{\omega} = b_k^{\omega} + \frac{1}{2}\langle\mathbf{a}_k'\mathbf{a}_k\rangle, \tag{19}$$

where $\boldsymbol{\Delta}_1 = \langle\varphi_j\rangle^{-1}\mathrm{diag}\langle\boldsymbol{\omega}\rangle + \sum_i\langle\mathbf{y}_i\mathbf{y}_i'\rangle$ ($\mathrm{diag}\langle\boldsymbol{\omega}\rangle = \mathrm{diag}(\langle\omega_1\rangle, \ldots, \langle\omega_k\rangle)$) and

$$\langle\omega_k\rangle = \tilde{a}_k^{\omega}/\tilde{b}_k^{\omega}. \tag{20}$$

Note that compared with VBFA1, Ghahramani and Beal (2000) does not require factorization assumption $q(\mathbf{A}, \boldsymbol{\mu}) = q(\mathbf{A})q(\boldsymbol{\mu})$ and can obtain a tighter lower bound $\mathcal{F}$. Thus VBFA1 described in this section is in fact a slightly simplified version of Ghahramani and Beal (2000). Nevertheless, both treatments are based on the same prior (6) for factor loading $\mathbf{A}$ that has $d \times q$ free parameters. Below we show that such an assumption has two problems.

3.1. Problem 1 of VBFA1

An investigation of (18) and (19) is helpful for understanding the iterative dynamics of ARD mechanism. A large $\langle\omega_k\rangle$ tends to make element k of $\mathbf{m}_{\mathbf{a}}^{(j)}$, $j = 1, \ldots, d$, i.e., column k of $\langle\mathbf{A}\rangle$ (or $\langle\mathbf{a}_k\rangle$) and thus $\langle\mathbf{a}_k'\mathbf{a}_k\rangle$ in (19) approach zero, which in turn makes $\langle\omega_k\rangle$ in (20) larger. Unfortunately, we also observe that VBFA1 has problem 1: a small $\langle\varphi_j\rangle^{-1}$ or noise ψ_j tends to prevent $\mathbf{m}_{\mathbf{a}}^{(j)}$ (or row j of $\langle\mathbf{A}\rangle$) and thus all $\langle\mathbf{a}_k'\mathbf{a}_k\rangle$'s from zero, which in turn restrains all $\langle\omega_k\rangle$'s in (19) from being desiredly large. In other words, low noises have a negative effect on ARD mechanism.

3.2. Problem 2 of VBFA1

In this section we analyze the large sample limit of VBFA1. $\mathcal{F}$ in (7) can be rewritten as a sum of two terms

$$\mathcal{F} = \underbrace{\left\langle \ln\frac{p(\mathbf{X}, \mathbf{Y}|\boldsymbol{\theta})}{q(\mathbf{Y})}\right\rangle_{q(\mathbf{Y})q(\boldsymbol{\theta})}}_{\mathcal{F}_{\mathcal{D}}} - \underbrace{\mathrm{KL}[q(\boldsymbol{\theta}) \parallel p(\boldsymbol{\theta})]}_{\mathcal{F}_{P}}. \tag{21}$$

According to the theoretical results on general latent variable models (including FA model) (Attias, 1999), we have that as $N \to \infty$, ① the first term $\mathcal{F}_{\mathcal{D}} \to \ln p(\mathbf{X}|\hat{\boldsymbol{\theta}})$, where $\hat{\boldsymbol{\theta}}$ is the ML estimator, and ② the second term $\mathcal{F}_P \to C_1(q)\ln N + O(1)$ ($C_1(q) = d(q+2)/2$ is the number of free parameters of $\boldsymbol{\theta}$ in VBFA1). Combining ① with ②, we have the limit of VBFA1

$$\mathcal{F}_1 \longrightarrow \ln p(\mathbf{X}|\hat{\boldsymbol{\theta}}) - C_1(q)\ln N + O(1), \quad N \to \infty. \tag{22}$$

In fact, (22) can also be obtained by explicitly calculating the large sample limit of the right-hand side of (21). To save space, we omit the detail in this paper. Recall that the Bayesian information criterion (BIC) (Schwarz, 1978) for FA model is in the form

$$\mathrm{BIC} = \ln p(\mathbf{X}|\hat{\boldsymbol{\theta}}) - C_2(q)\ln N, \tag{23}$$

where $C_2(q) = (d(q+2) - q(q+1)/2)/2$. The number of free parameters drops by $q(q+1)/2$ as the factor loading $\mathbf{A}$ can only be determined up to a rotation. Note that VBFA1 fails to consider this fact. It is well known that BIC is consistent in selecting the correct model (Shao, 1997). If the candidate models contains the true one, the probability of choosing the true model by BIC approaches one as sample size $N \to \infty$. Comparing (22) with (23), we find that VBFA1 has problem 2: $\mathcal{F}_1$ penalizes model complexity more heavily than BIC.

4. VBFA2

In this section, we propose a new VB treatment to resolve the two problems suffered by VBFA1. Note that the prior (6) in VBFA1 uses $\mathbf{a}_k'\mathbf{a}_k$ to measure factor k. A disadvantage of using this prior is that it does not consider the influence from parameter $\boldsymbol{\varphi}$ and may result in problem 1 mentioned in Section 3. Instead of using $\mathbf{a}_k'\mathbf{a}_k$, Theorem 1 suggests using $\mathbf{a}_k'\boldsymbol{\varphi}\mathbf{a}_k$ to measure the significance of factor k, which naturally involves the impact from $\boldsymbol{\varphi}$. This motivates us to consider the following form of prior for $\mathbf{A}$:

$$p(\mathbf{A}|\boldsymbol{\varphi}, \boldsymbol{\omega}) = \prod_{k=1}^{q}\left(\frac{\omega_k}{2\pi}\right)^{d/2}|\boldsymbol{\varphi}|^{1/2}\exp\left(-\frac{\omega_k}{2}\mathbf{a}_k'\boldsymbol{\varphi}\mathbf{a}_k\right). \tag{24}$$

Notice that prior (24) with $\boldsymbol{\omega}$ fixed has been used in Bayesian FA proposed by Akaike (1987), but he did not consider ARD. Actually, (24) with ARD has also been used by Oba, Sato, and Ishii (2003) in the context of probabilistic principal component analysis. However, they did not study the advantages using (24) over (6) theoretically or empirically. To our knowledge, this remains unknown and is one of our focuses in this paper.

The problem 2 of VBFA1 stated in Section 3.2, i.e. the rotation problem, is well known in ML FA and commonly solved by imposing a computationally convenient constraint, e.g. in the CM algorithm of Zhao et al. (2008), $\mathbf{A}$ has the decomposition in the form $\mathbf{A} = \mathbf{BV}$ (see (3)), in which the orthogonal matrix $\mathbf{V}$ is set to be $\mathbf{I}$. Since such a constraint is intractable for Bayesian methods, we follow Fokoué and Titterington (2003) to preassign values to

$q(q+1)/2$ arguments of $\mathbf{A}$ so as to reduce the number of free parameters directly. We use the following lower triangular matrix for $\mathbf{A}$:

$$\mathbf{A} = \begin{pmatrix} a_{11} & 0 & 0 & \cdots & 0 & 0 \\ a_{21} & a_{22} & 0 & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ a_{q-1,1} & a_{q-1,2} & a_{q-1,3} & \cdots & a_{q-1,q-1} & 0 \\ a_{q,1} & a_{q,2} & a_{q,3} & \cdots & a_{q,q-1} & a_{q,q} \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ a_{d,1} & a_{d,2} & a_{d,3} & \cdots & a_{d,q-1} & a_{d,q} \end{pmatrix}. \tag{25}$$

The justification of using constraint (25) is detailed in Appendix B. A discussion concerning the rotation problem with VBFA1-ARD and VBFA2-ARD can be found in Appendix C.

Using (25), the prior (24) becomes

$$p(\mathbf{A}|\boldsymbol{\varphi},\boldsymbol{\omega}) = \prod_{k=1}^{q} \left(\frac{\omega_k}{2\pi}\right)^{(d-k+1)/2} \prod_{j=k}^{d} \varphi_j^{1/2} \exp\left(-\frac{\omega_k}{2}\bar{\mathbf{a}}_k'\boldsymbol{\varphi}_k\bar{\mathbf{a}}_k\right), \tag{26}$$

Here $\boldsymbol{\varphi}_k = \mathrm{diag}(\varphi_k,\ldots,\varphi_d)$ and $\bar{\mathbf{a}}_k = (a_{kk},\ldots,a_{dk})'$, $k = 1,\ldots,q$.

Given prior (26), it is tractable to consider less factorization (27) rather than (15) with the advantage of achieving a tighter lower bound $\mathcal{F}$:

$$q(\boldsymbol{\theta}) = q(\boldsymbol{\mu})q(\mathbf{A}|\boldsymbol{\varphi})q(\boldsymbol{\varphi})q(\boldsymbol{\omega}). \tag{27}$$

The probabilistic graphical model of VBFA1 is shown in Fig. 1(b) and the resulting VB treatment from (26), (27) and (12)–(14) is called VBFA2. The VBEM updating steps in VBFA2 are as follows:

- VBE-step:

$$q(\mathbf{Y}) = \prod_i \mathcal{N}\left(\mathbf{y}_i|\mathbf{m}_{\mathbf{y}}^{(i)}, \boldsymbol{\Sigma}_{\mathbf{y}}\right). \tag{28}$$

- VBM-step 1:

$$q(\boldsymbol{\mu}) = \mathcal{N}\left(\boldsymbol{\mu}|\mathbf{m}_{\mu}, \boldsymbol{\Sigma}_{\mu}\right). \tag{29}$$

- VBM-step 2: $q(\mathbf{A}|\boldsymbol{\varphi})q(\boldsymbol{\varphi})$ can be further factorized into $\prod_{j=1}^{d} q(\tilde{\mathbf{a}}_j|\varphi_j)q(\varphi_j)$ and $q(\tilde{\mathbf{a}}_j|\varphi_j)q(\varphi_j)$ turns out to be a Gaussian–Gamma distribution, so that

$$q(\mathbf{A}|\boldsymbol{\varphi})q(\boldsymbol{\varphi}) = \prod_{j=1}^{d} \mathcal{N}(\bar{\tilde{\mathbf{a}}}_j|\mathbf{m}_{\mathbf{a}}^{(j)}, \boldsymbol{\Sigma}_{\mathbf{a}}^{(j)})\Gamma(\varphi_j|\tilde{a}_j^{\varphi}, \tilde{b}_j^{\varphi}), \tag{30}$$

  where $\bar{\tilde{\mathbf{a}}}_j$ consists of the first $j*$ elements of $\tilde{\mathbf{a}}_j$, with $j* = \min(j,q)$.

- VBM-step 3:

$$q(\boldsymbol{\omega}) = \prod_k \Gamma\left(\omega_k|\tilde{a}_k^{\omega}, \tilde{b}_k^{\omega}\right), \tag{31}$$

The corresponding moments in (28)–(31) are given by

$$\boldsymbol{\Sigma}_{\mathbf{y}} = \left(\mathbf{I} + \langle\mathbf{A}'\boldsymbol{\varphi}\mathbf{A}\rangle\right)^{-1}, \qquad \mathbf{m}_{\mathbf{y}}^{(i)} = \boldsymbol{\Sigma}_{\mathbf{y}}\langle\mathbf{A}\rangle'\langle\boldsymbol{\varphi}\rangle(\mathbf{x}_i - \langle\boldsymbol{\mu}\rangle), \tag{32}$$

$$\boldsymbol{\Sigma}_{\mu} = (N\langle\boldsymbol{\varphi}\rangle + \beta)^{-1}\mathbf{I}, \qquad \mathbf{m}_{\mu} = \boldsymbol{\Sigma}_{\mu}\langle\boldsymbol{\varphi}\rangle\sum_i(\mathbf{x}_i - \langle\mathbf{A}\rangle\langle\mathbf{y}_i\rangle), \tag{33}$$

$$\boldsymbol{\Sigma}_{\mathbf{a}}^{(j)} = \varphi_j^{-1}\boldsymbol{\Delta}_{j*}^{-1}, \qquad \mathbf{m}_{\mathbf{a}}^{(j)} = \boldsymbol{\Delta}_{j*}^{-1}\left(\sum_i\langle\mathbf{y}_i\rangle_{j*}(\mathbf{x}_{ij} - \langle\mu_j\rangle)\right), \tag{34}$$

$$\begin{aligned} \tilde{a}_j^{\varphi} &= a_j^{\varphi} + \frac{n}{2}, \\ \tilde{b}_j^{\varphi} &= b_j^{\varphi} + \frac{1}{2}\left(\sum_i(x_{ij} - \langle\mu_j\rangle)^2 - [\mathbf{m}_a^{(j)}]'\boldsymbol{\Delta}_{j*}\mathbf{m}_a^{(j)}\right). \end{aligned} \tag{35}$$

$$\tilde{a}_k^{\omega} = a_k^{\omega} + \frac{d-k+1}{2}, \qquad \tilde{b}_k^{\omega} = b_k^{\omega} + \frac{1}{2}\langle\bar{\mathbf{a}}_k'\boldsymbol{\varphi}_k\bar{\mathbf{a}}_k\rangle, \tag{36}$$

where $\boldsymbol{\Delta} = \mathrm{diag}\langle\boldsymbol{\omega}\rangle + \sum_i\langle\mathbf{y}_i\mathbf{y}_i'\rangle$ and $\boldsymbol{\Delta}_{j*}$ consists of the first $j*$ rows and $j*$ columns of $\boldsymbol{\Delta}$. $\langle\mathbf{y}_i\rangle_{j*}$ and $\mathbf{m}_{\mathbf{a}}^{(j)}$ stand for the first $j*$ elements of $\langle\mathbf{y}_i\rangle$ and $\langle\tilde{\mathbf{a}}_j\rangle$. Comparing (18) and (34), we see that a major difference between VBFA2 and VBFA1 is that element k of $\mathbf{m}_{\mathbf{a}}^{(j)}$, $j = 1,\ldots,d$, i.e., column k of $\langle\mathbf{A}\rangle$ (or $\langle\bar{\mathbf{a}}_k\rangle$) in VBFA2 only depends on $\langle\omega_k\rangle$, rather than $\langle\omega_k\rangle$ and $\langle\varphi_j\rangle$ in VBFA1. A large $\langle\omega_k\rangle$ tends to make column k of $\langle\mathbf{A}\rangle$ (or $\langle\bar{\mathbf{a}}_k\rangle$) and thus $\langle\bar{\mathbf{a}}_k'\boldsymbol{\varphi}_k\bar{\mathbf{a}}_k\rangle$ go to zero, which in turn makes $\langle\omega_k\rangle$ in (20) larger. In short, the ARD mechanism in VBFA2 is much more insensitive to low noises. In Appendix A, we give the detailed expression of the lower bound $\mathcal{F}$ in VBFA2 (denoted by $\mathcal{F}_2$ hereafter) as the bound is typically useful for monitoring convergence of VB algorithm and model comparison.

4.1. The large sample limit of VBFA2

As the number of free parameters is reduced by $q(q+1)/2$, the second term in (21) $\mathcal{F}_P$ tends to $C_2(q)\ln N$. Therefore, the lower bound of VBFA2

$$\mathcal{F}_2 \longrightarrow \ln p(\mathbf{X}|\hat{\boldsymbol{\theta}}) - C_2(q)\ln N + O(1), \tag{37}$$

which is approximately equal to BIC.

4.2. Backward learning of VBFA2

An advantage of VB is that the lower bound $\mathcal{F}$ can be used as model selection criterion since it approximates the model evidence $\ln p(\mathbf{X})$. Given a range for the number of factors q in FA, e.g., $[q_{\min}, q_{\max}]$, we can choose the minimizer of $\mathcal{F}(q)$ as the best solution for q. However, it would be time consuming to perform such crude search if the range is not small. To alleviate this problem, we propose a backward learning algorithm for VBFA2. The algorithm consists of two stages.

Stage1. Set $q = q_{\max}$. Iterate (32)–(36) until a convergence criterion is met. During iterations, any column k of $\mathbf{A}$ with $\langle\omega_k\rangle$ going to infinity (over some threshold, in practice) is removed (as long as $\mathcal{F}_2$ increases). Set $q_{\text{best}} = q_{\text{rem}}$, and $\mathcal{F}_{\text{best}} = \mathcal{F}_2$, where q_{rem} is the number of factors finally remained.

Stage2. Step 1. Set $q = q_{\text{rem}} - 1$. Remove the column of $\mathbf{A}$ with the largest $\langle\omega_k\rangle$ and iterate (32)–(36) until a convergence criterion is met. If $\mathcal{F}_2 > \mathcal{F}_{\text{best}}$, set $\mathcal{F}_{\text{best}} = \mathcal{F}_2$ and $q_{\text{best}} = q_{\text{rem}}$.
Step 2. If $q = q_{\min}$, then quit; else go to step 1.

Due to the lower triangular structure of $\mathbf{A}$, the free parameters in $\mathbf{A}$ should be redefined carefully when some column of $\mathbf{A}$ is removed. For example, if we remove column 2 of $\mathbf{A}$, then the zeros marked in red from column 3 to d in (25) will be reactivated as free parameters.

In Stage 1, it seems better to perform removal only when the condition that $\mathcal{F}_2$ increases is satisfied. However, from our experience, such a condition does not lead to improved performance but more iterations. Therefore, we does not perform the condition in our simulation study.

Note also q_{best} obtained in Stage 1 purely depends on ARD mechanism and any inactive factor has been removed. Intuitively, the retained factors should be more or less useful. Therefore, it would be interesting to investigate the performance of ARD (i.e., only using the output of Stage 1) on model selection and generalized performance, which is detailed in Section 5.

5. Simulations

We examine the performance of VBFA1, VBFA2, and MLFA for low ($d = 10$) and high ($d = 100$) dimensional data with Normal or Low noise. Hence there are four data settings in all:

- *data* 1: $d = 10$, Normal noise
  The data is generated from 5-factor FA model with
  $$\boldsymbol{\mu}_1 = (0\,0 \cdots 0)', \qquad \boldsymbol{\Psi}_1 = \mathrm{diag}\{0.1(1\,2 \cdots 10)\},$$
  $$\mathbf{A}_1' = \begin{pmatrix} 2 & -6 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 3 & 4 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & -2 & -1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & -5 & -2 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & -0.5 & -3 & 4 \end{pmatrix}$$
- *data* 2: $d = 10$, Low noise
  $\boldsymbol{\mu}_2 = \boldsymbol{\mu}_1$, $\boldsymbol{\Psi}_2 = \boldsymbol{\Psi}_1$ and $\mathbf{A}_2 = \mathbf{A}_1$ except that $\psi_{2ii} = 0.01\psi_{1ii}$, $i = 1, \ldots, 5$.
- *data* 3: $d = 100$, Normal noise
  The FA model has 5 factors (or latent dimension) with parameters: $\boldsymbol{\mu}_3 = \mathbf{0}_{100\times 1}$, $\boldsymbol{\Psi}_3 = \mathrm{diag}\{0.05(1, 2, \ldots, 100)\}$, $\mathbf{A}_3 = \mathbf{0}_{100\times 5}$, except that some elements in each column (detailedly, $a_{(j-1)*20+1\,j}, \ldots, a_{\min((j-1)*20+21,100)\,j}$ in column j, $j = 1, \ldots, 5$) are generated from normal distribution with mean 0 and variance 1.
- *data* 4: $d = 100$, Low noise
  $\boldsymbol{\mu}_4 = \boldsymbol{\mu}_3$, $\boldsymbol{\Psi}_4 = \boldsymbol{\Psi}_3$ and $\mathbf{A}_4 = \mathbf{A}_3$ except that $\psi_{4ii} = 0.001\psi_{3ii}$, $i = 91, \ldots, 100$.

It seems that we simply consider the data generated from FA model (1) with $\mathbf{A}$ satisfying constraint (25) (denoted as $\mathbf{A}_c$). In fact, we have taken into account the data from general $\mathbf{A}$ no matter satisfying constraint (25) or not, because the rotation invariance of FA model leads that $\mathbf{x}_i$ generated using general $\mathbf{A}$ can always be viewed as generated using $\mathbf{A}_c$ as detailed in Appendix B.

In our simulation study, we use the following settings:

- *Prior settings:* $\beta = a^{\varphi} = b^{\varphi} = 10^{-3}$; $a^{\omega} = b^{\omega} = 10^{-3}/N$, where N is the sample size of the training data.
- *Convergence criterion:* Stop algorithms if $|1 - \mathcal{F}^{(t)}/\mathcal{F}^{(t+1)}| < tol$ or $t > K_{\max}$ with tol $= 10^{-9}$ and the maximal number of iterations: $K_{\max} = 1000$. To avoid poor local maxima, the best solution from r runs of VBFA1 or VBFA2 is chosen. For *data* 1 and 2, $r = 10$ and for *data* 3 and 4, $r = 5$.

5.1. VBFA2 vs. VBFA1

In this subsection, we compare VBFA1 and VBFA2 on generalization performance and ARD performance, i.e., whether unnecessary factors can be suppressed.

5.1.1. ARD performance

VBFA1 and VBFA2 with $q = 9$ are fitted to *data* 2. When the *convergence criterion* is satisfied, the typical ARD output from VBFA1 and VBFA2 are shown in Fig. 2. Obviously, for this data, ARD works well only in VBFA2 and ARD in VBFA1 cannot effectively suppress the unnecessary factors. Consequently, VBFA1 may improperly suggest the latent dimension in the data and mislead the subsequent model explanation, which is important for factor analysis. Similar observations can be obtained for *data* 4 when $N \geq 300$. For *data* 1 and 3, ARD in VBFA1 and VBFA2 perform more or less similarly.

In order to understand more how ARD works in VBFA1 and VBFA2, Fig. 3 plots the training curves of ω_k, $k = 1, \ldots, 9$ using 5000 iterations. Looking at Fig. 3 (a) and (c) ($N = 70$), VBFA2 requires only less than 1000 iterations to identify the correct dimension q, much faster than VBFA1 (>5000 iterations). As N becomes larger, VBFA1 becomes even worse, see Fig. 3 (b) and (d) ($N = 600$), while VBFA2 performs consistently well. Fig. 3 (e) shows a realization of the lower bound $\mathcal{F}_1$ for VBFA1 and $\mathcal{F}_2$ for VBFA2.

5.1.2. Generalization performance

Following Nielsen (2004), the predictive performance of VBFA1 and VBFA2 is investigated via *learning curves*. A *learning curve* plots the negative log-likelihood evaluated on an independent test data set (test error) using the fitted model based on training data sets with varying sample size N. For each data setting, we generate an independent test data set with size 10^6. For each N under consideration, we generate 50 training data sets. For comparison, MLFA is also included. Fig. 4(a)–(d) show the obtained average learning curves by VBFA1, VBFA2 and ML methods for *data* 1–4, respectively. Fig. 4 (a) and (b) show the results of FA models with $q = \{5, 9\}$. Fig. 4 (c) and (d) show the results of FA with $q = 5$ for VBFA1 and VBFA2, and the results of FAs with $q = \{3, 5, 10\}$ for ML method. From Fig. 4, we conclude:

1. VBFA2 is in general better than VBFA1, especially, for small N in *data* 1 and *data* 2. The superiority in *data* 2 is more obvious than that in *data* 1. VBFA2 performs consistently better than ML method but VBFA1 is not the case, e.g., for $N = 40$ in *data* 1 and 2.
2. In ML method, the choice of q is crucial for the model prediction. This can be observed more clearly from Fig. 4 (c). When $N < 100$, FA with $q = 3$ best fits the data and FAs with $q = 5$ and $q = 9$ overfit, while when $100 \leq N \leq 600$, FA with $q = 5$ best fits the data, FA with $q = 3$ underfits and that with $q = 9$ overfits. In contrast, VB method is insensitive to the choice of q as long as q is large enough, e.g., FAs with $q = 5$ and $q = 9$ give similar performance. In fact we also plot the curves of VB methods for FA with $q = 9$ in Fig. 4 (c) and (d) as those in Fig. 4 (a) and (b), but FA with $q = 9$ superposes upon FA with $q = 5$. This could be ascribed to the ARD mechanism in VB methods since it automatically penalizes the unneeded latent dimensions.

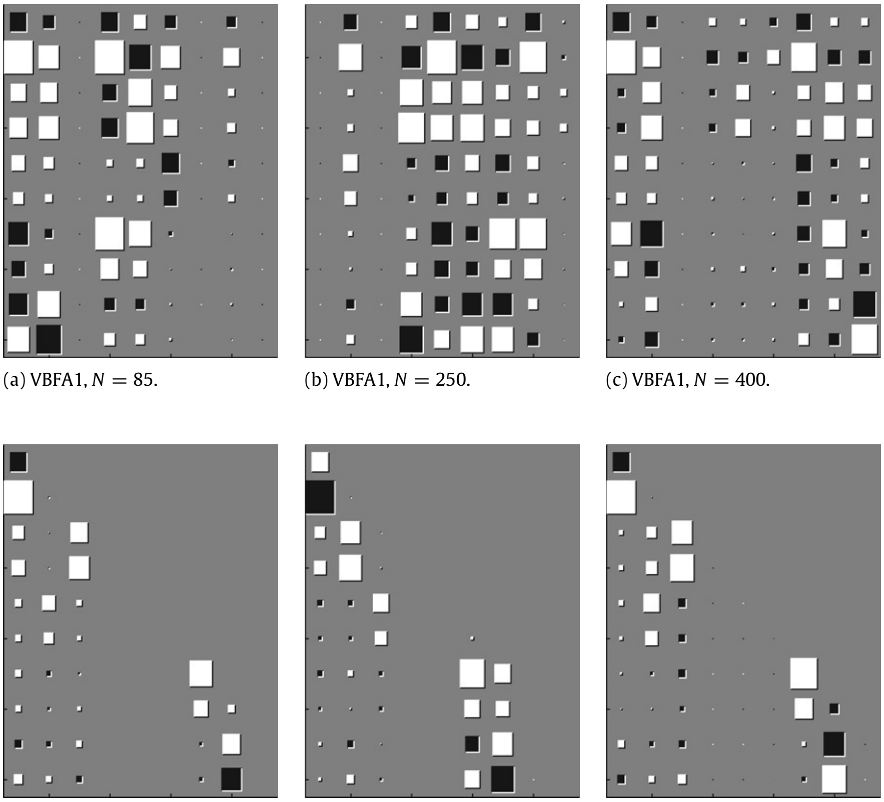
5.2. Model selection: VB vs. BIC

We compare the model selection ability using VB lower bound ($\mathcal{F}_1$ and $\mathcal{F}_2$) and BIC. In addition, for VBFA2, we also examine the model selection performance purely based on ARD ($\mathcal{F}_2$-ARD, for short). In order to mitigate the computation burden, the backward learning algorithm for VBFA2 presented in Section 4.2 is used. In Stage 1, we remove column k of $\mathbf{A}$ if its corresponding $\langle\omega_k\rangle > N$, where N is the sample size of the training data. Fig. 3(f) shows a realization of the algorithm for VBFA2. The detailed model selection results for *data* 1 and 3 are summarized in Table 1 (To save space, the results for *data* 2 and 4 are omitted since results are similar). The predictive performance of chosen models for all four data settings is visualized in Fig. 5. Some conclusions are drawn as follows.

(1) $\mathcal{F}_2$ vs. $\mathcal{F}_1$: From Table 1, compared with $\mathcal{F}_2$ for $40 \leq N \leq 125$ in *data* 1 and $60 \leq N \leq 200$ in *data* 3, it is clear that $\mathcal{F}_1$ tends to underfit the data as it often chooses a simpler model that has smaller q and larger test error.

(2) $\mathcal{F}_2$ vs. BIC: BIC in *data* 1 gives satisfactory dimension estimation and test error for $25 \leq N \leq 70$ in data 1. Nevertheless, its performance for *data* 3 is rather unsatisfactory compared with $\mathcal{F}_2$. In fact, when $N \geq 100$ for *data* 3, from Fig. 4 (c), FA with $q = 5$ performs the best but BIC can not choose this best dimension and underesimates the data dimension until $N \geq 250$. In contrast, the performance of $\mathcal{F}_2$ seems more stable than BIC for different data settings.

(3) $\mathcal{F}_2$-ARD vs. $\mathcal{F}_2$ and BIC: From Fig. 5 and Table 1: ① In large N case (e.g. $N = 600$) all three criteria perform very similarly. ② In small N case ($N \leq 125$ in *data* 1 and $N \leq 250$ in *data* 3), $\mathcal{F}_2$-ARD performs the best in prediction and dimension estimation, though BIC for $N = 25$ and $N = 40$ in *data* 1 chooses a more plausible dimension (closer to true one) but



(a) VBFA1, $N = 85$.

(b) VBFA1, $N = 250$.

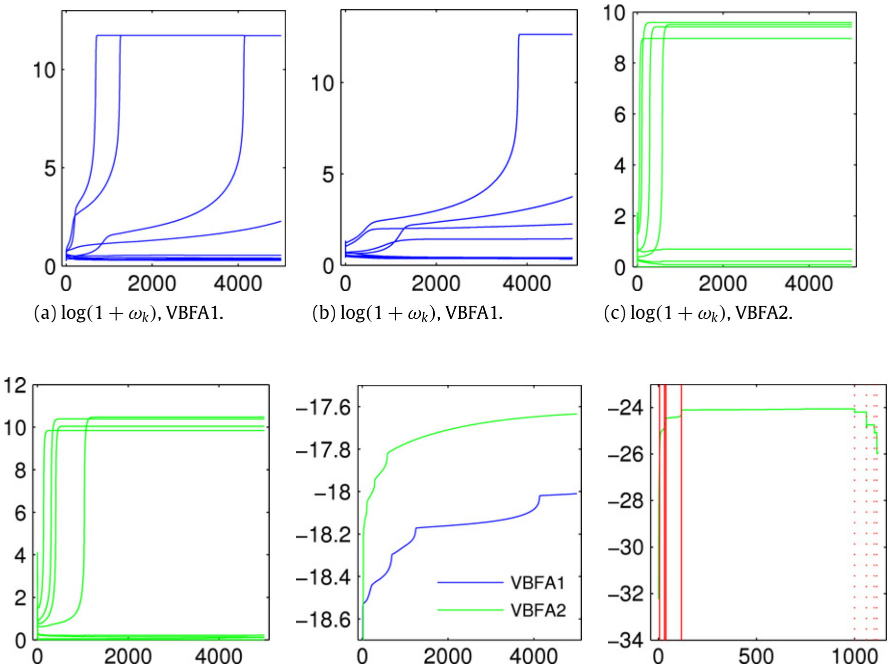
(c) VBFA1, $N = 400$.

(d) VBFA2, $N = 85$.

(e) VBFA2, $N = 250$.

(f) VBFA2, $N = 400$.

Fig. 2. Typical Hinton diagrams of factor loadings by fitting VBFA1 and VBFA2 to *data* 2 with different N.



(a) $\log(1 + \omega_k)$, VBFA1.

(b) $\log(1 + \omega_k)$, VBFA1.

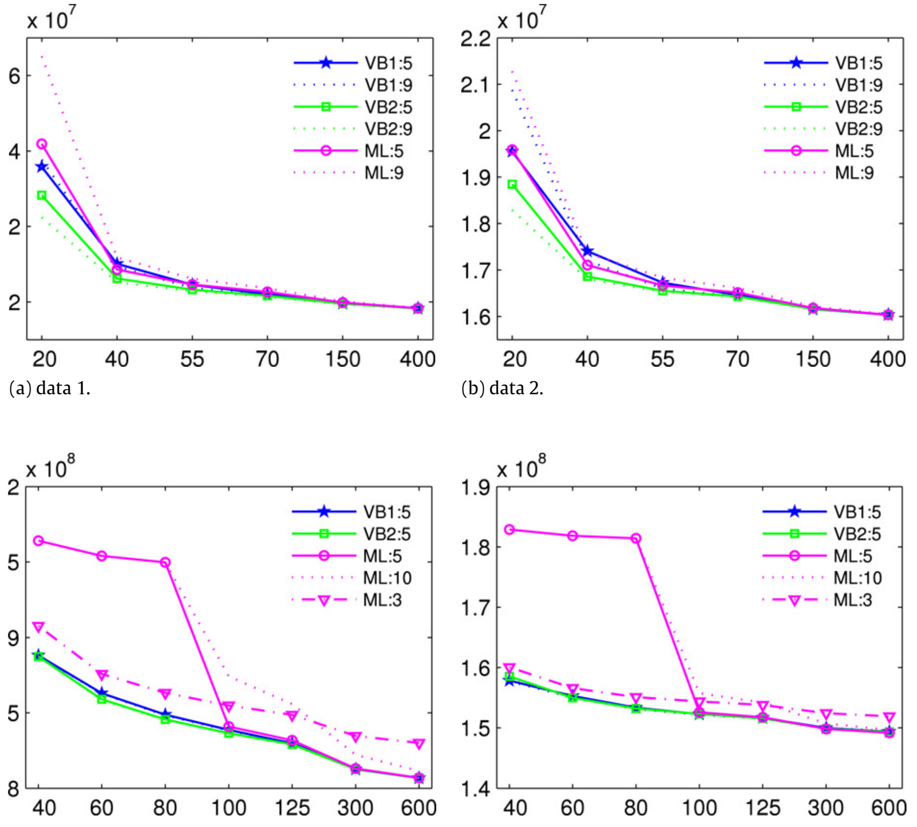
(c) $\log(1 + \omega_k)$, VBFA2.

(d) $\log(1 + \omega_k)$, VBFA2.

(e) evolvement of F_1 and F_2.

(f) evolvement of F_2.

Fig. 3. (a)–(d), the fitting of VBFA1 and VBFA2 with $q = 9$ to *data* 2. (a) and (c) ($N = 70$), (b) and (d) ($N = 600$): evolvements of $\ln(1 + \omega_k)$, $k = 1, \ldots, 9$, for VBFA1 and VBFA2. (e) Evolvement of $\mathcal{F}$ for VBFA1 and VBFA2; (f) evolvement of $\mathcal{F}_2$ by the backward learning algorithm. Solid and dotted lines signal the removals of inactive factors via threshold of $\langle\omega_k\rangle$ and the enforced removals of active factors, starting from the least active one (the largest $\langle\omega_k\rangle$).



(a) data 1.

(b) data 2.

(c) data 3.

(d) data 4.

Fig. 4. Learning curves of VBFA1, VBFA2 and MLFA with various q for four different data settings. VB1:k, VB2:k and ML:k denote learning FA model with $q = k$ by VBFA1, VBFA2 and ML respectively.

such choice does not lead better prediction. $\mathcal{F}_2$-ARD instead woks well by automatically penalizing the factors which can not receive enough support from the small size data and thus obtains better prediction. Note also that $\mathcal{F}_2$-ARD in Table 1 has one overestimation for $N = 600$ in *data* 1 and *data* 3. This is because they reached the maximal number of iterations $K_{\max} = 1000$. Actually, setting $K_{\max} = 2000$ can make $\mathcal{F}_2$-ARD output the correct dimension for the overestimated data.

6. Concluding remarks and future work

We have developed a novel VB treatment for FA model. Theoretically, we show that the resulting model selection criterion from our method is approximately equivalent to BIC while the existing methods in the literature are not the case. Empirically, we focus on the data generated from 'sparse' factor loadings matrices, namely there are many zeros on each column like $\mathbf{A}_1 - \mathbf{A}_4$ in Section 5, because such data bears deep origin in history of FA model in which each factor is expected to capture the information of several variables. Under the 'sparse' case, we find that ① our method performs better than the existing methods on both model selection and prediction, particularly for the data in which some data dimensions are contaminated by low noises; ② our method performs more stable than BIC in terms of different data settings. Moreover, our empirical results reveal that in determining the number of factors, using ARD only performs better than using $\mathcal{F}_1$ or $\mathcal{F}_2$ and is more reliable particularly for the data sets with small sample size.

As suggested by one reviewer, we also investigated the performance of VBFA1 and VBFA2 using the data from a 'non-sparse' factor loadings matrix, e.g. a dense matrix generated from Gaussian distribution, and found that VBFA2 in this case performs slightly better than or nearly same to VBFA1 (results not shown in this paper). Under the 'non-sparse' case, both VBFA1 and VBFA2 choose a similar latent dimension that leads to similar generalized performance. The better performance of VBFA2 under the 'sparse' case might be ascribed to the fact that $\mathcal{F}_2$ approaches to BIC while $\mathcal{F}_1$ has a heavier penalty than BIC because compared with VBFA1, VBFA2 in this case determines a latent dimension that is closer to true one and can lead better generalized performance as discussed in the first paragraph of Section 6. Therefore it can be concluded that VBFA2 is a safer tool than VBFA1 for applications because the performance of VBFA2 is either better than VBFA1 or comparable with VBFA1.

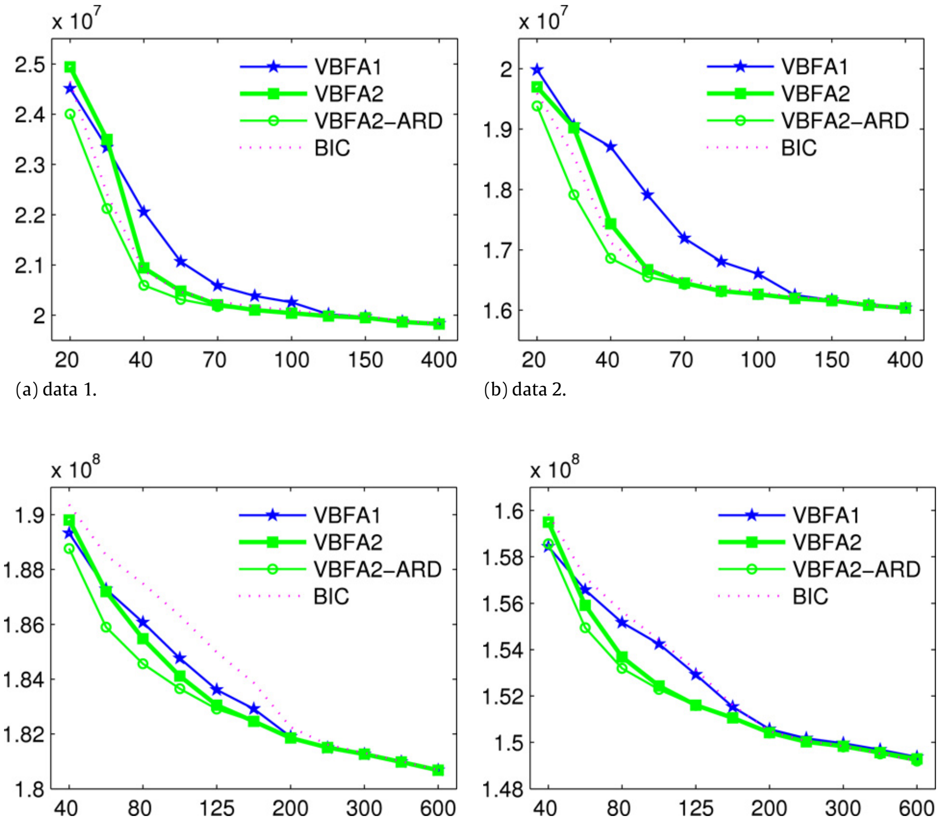
Although we simply focus on FA model in this paper, the idea could be extended to many other latent variable models for which existing VB formulations seemingly suffer similar problems, e.g., mixtures of FAs (Ghahramani & Beal, 2000), probabilistic canonical correlation analysis (Wang, 2007), extended or parallel factor analysis (Nielsen, 2004), rectified factor analysis (Harva & Kabán, 2007), etc. Of course, its efficiency in these cases needs to be further investigated.

Acknowledgements

The authors would like to thank the reviewers for their insightful comments that led to an improved article with a more general and in-depth simulation study. In addition, Appendix C was added under the suggestions of the reviewers. The authors are supported by a small project funding from The University of Hong Kong.

Table 1
Comparisons on latent dimension estimation among $\mathcal{F}_1$, $\mathcal{F}_2$, $\mathcal{F}_2$-ARD and BIC. U, C, O: Total number of underestimation, correct estimation, overestimation of dimensions among 50 simulations; A: The average estimated latent dimension from 50 simulations. Test: the average test error per data point from 50 simulations. Bold face signals that a criterion starts to obtain a satisfactory dimension estimation (≥40).

N	Method	*data* 1					*data* 3					
		U	C	O	A	Test	*N*	U	C	O	A	Test
25	BIC	6	**43**	1	4.9	22.41	60	50	0	0	1.1	188.52
	$\mathcal{F}_1$	50	0	0	1.7	23.34		50	0	0	1.9	187.28
	$\mathcal{F}_2$	50	0	0	1.4	23.93		50	0	0	2.6	187.22
	$\mathcal{F}_2$-ARD	46	4	0	3.7	22.12		9	**41**	0	4.8	185.90
40	BIC	1	48	1	5.0	20.88	100	49	1	0	2.1	186.32
	$\mathcal{F}_1$	50	0	0	2.5	22.05		50	0	0	3.2	184.76
	$\mathcal{F}_2$	46	4	0	3.9	21.02		31	19	0	4.3	184.17
	$\mathcal{F}_2$-ARD	4	**46**	0	4.9	20.60		0	50	0	5.0	183.65
70	BIC	0	49	1	5.0	20.27	125	49	1	0	2.9	184.98
	$\mathcal{F}_1$	49	1	0	4.0	20.59		45	5	0	3.9	183.61
	$\mathcal{F}_2$	9	**41**	0	4.8	20.23		10	**40**	0	4.8	183.05
	$\mathcal{F}_2$-ARD	1	49	0	5.0	20.17		0	50	0	5.0	182.91
100	BIC	0	49	1	5.0	20.10	200	17	33	0	4.6	182.25
	$\mathcal{F}_1$	27	23	0	4.5	20.26		4	**46**	0	4.9	181.90
	$\mathcal{F}_2$	0	50	0	5.0	20.04		0	50	0	5.0	181.85
	$\mathcal{F}_2$-ARD	0	50	0	5.0	20.04		0	50	0	5.0	181.85
125	BIC	0	50	0	5.0	20.02	250	3	**47**	0	4.9	181.62
	$\mathcal{F}_1$	4	**46**	0	4.9	20.02		0	50	0	5.0	181.49
	$\mathcal{F}_2$	0	50	0	5.0	19.98		0	50	0	5.0	181.50
	$\mathcal{F}_2$-ARD	0	50	0	5.0	19.98		0	50	0	5.0	181.50
600	BIC	0	50	0	5.0	19.81	600	0	50	0	5.0	180.69
	$\mathcal{F}_1$	0	50	0	5.0	19.81		0	50	0	5.0	180.68
	$\mathcal{F}_2$	0	50	0	5.0	19.80		0	50	0	5.0	180.68
	$\mathcal{F}_2$-ARD	0	49	1	5.0	19.80		0	49	1	5.0	180.68



(a) data 1.

(b) data 2.

(c) data 3.

(d) data 4.

Fig. 5. Learning curves based on $\mathcal{F}_1$, $\mathcal{F}_2$, $\mathcal{F}_2$-ARD and BIC for different data settings.

Appendix A. The lower bound of VBFA2

Substituting (26), (27) and (12)–(14) into (21), we obtain

$$\begin{aligned}\mathcal{F}_2 = {} & \langle \ln p(\mathbf{X}|\mathbf{Y}, \mathbf{A}, \boldsymbol{\mu}, \boldsymbol{\varphi})\rangle_{q(\mathbf{Y})q(\mathbf{A},\boldsymbol{\varphi})q(\boldsymbol{\mu})} - \mathrm{KL}(q(\mathbf{Y}) \parallel p(\mathbf{Y})) \\ & - \mathrm{KL}(q(\boldsymbol{\mu}) \parallel p(\boldsymbol{\mu})) - \langle \mathrm{KL}(q(\mathbf{A}|\boldsymbol{\varphi}) \parallel p(\mathbf{A}|\boldsymbol{\varphi}, \boldsymbol{\omega}))\rangle_{q(\boldsymbol{\varphi})q(\boldsymbol{\omega})} \\ & - \mathrm{KL}(q(\boldsymbol{\omega}) \parallel p(\boldsymbol{\omega}|\mathbf{a}^{\omega}, \mathbf{b}^{\omega})) - \mathrm{KL}(q(\boldsymbol{\varphi}) \parallel p(\boldsymbol{\varphi}|\mathbf{a}^{\varphi}, \mathbf{b}^{\varphi})), \quad \text{(A.1)}\end{aligned}$$

where

$$\begin{aligned}\langle \ln p(\mathbf{X}|\mathbf{Y}, \mathbf{A}, \boldsymbol{\mu}, \boldsymbol{\varphi})\rangle = {} & -\frac{nd}{2}\ln 2\pi - \frac{1}{2}\sum_{j=1}^{d} j^* + \frac{n}{2}\sum_{j=1}^{d}\langle \ln \varphi_j\rangle \\ & + \sum_{j=1}^{d}(\langle\varphi_j\rangle b_j^{\varphi} - \tilde{a}_j^{\varphi}) + \frac{1}{2}\mathrm{tr}(\langle \mathbf{A}'\boldsymbol{\varphi}\mathbf{A}\rangle \mathrm{diag}\langle\boldsymbol{\omega}\rangle), \qquad \text{(A.2)}\end{aligned}$$

$$\mathrm{KL}(q(\mathbf{Y}) \parallel p(\mathbf{Y})) = -\frac{n}{2}\ln|\boldsymbol{\Sigma}_{\mathbf{y}}| - \frac{n}{2}q + \frac{1}{2}\mathrm{tr}\left(\sum_i \langle \mathbf{y}_i\mathbf{y}_i'\rangle\right), \qquad \text{(A.3)}$$

$$\mathrm{KL}(q(\boldsymbol{\mu}) \parallel p(\boldsymbol{\mu})) = -\frac{1}{2}\ln|\boldsymbol{\Sigma}_{\boldsymbol{\mu}}| - \frac{d}{2} - \frac{d}{2}\ln\beta + \frac{\beta}{2}\langle\boldsymbol{\mu}'\boldsymbol{\mu}\rangle, \qquad \text{(A.4)}$$

$$\begin{aligned}\langle \mathrm{KL}(q(\mathbf{A}|\boldsymbol{\varphi}) \parallel p(\mathbf{A}|\boldsymbol{\varphi}, \boldsymbol{\omega}))\rangle = {} & -\sum_{k=1}^{q}\frac{d-k+1}{2}\langle \ln\tilde{\omega}_k\rangle \\ & + \frac{1}{2}\sum_{j=1}^{d}\left(j^*\langle\ln\varphi_j\rangle + \ln|\boldsymbol{\Delta}_{j^*}| - j^*\right) \\ & - \frac{1}{2}\sum_{j=1}^{d} j^*\langle \ln\varphi_j\rangle + \frac{1}{2}\mathrm{tr}(\mathrm{diag}\langle\mathbf{A}'\boldsymbol{\varphi}\mathbf{A}\rangle\mathrm{diag}\langle\widetilde{\boldsymbol{\omega}}\rangle), \qquad \text{(A.5)}\end{aligned}$$

$$\begin{aligned}& \mathrm{KL}(q(\boldsymbol{\varphi}) \parallel p(\boldsymbol{\varphi}|\mathbf{a}^{\varphi}, \mathbf{b}^{\varphi})) \\ & = \sum_{j=1}^{d}\Big(\tilde{a}_j^{\varphi}\ln\tilde{b}_j^{\varphi} - a_j^{\varphi}\ln b_j^{\varphi} + \ln\Gamma(a_j^{\varphi}) - \ln\Gamma(\tilde{a}_j^{\varphi}) \\ & \quad + b_j^{\varphi}\langle\varphi_j\rangle - \tilde{a}_j^{\varphi} + (\tilde{a}_j^{\varphi} - a_j^{\varphi})\langle\ln\varphi_j\rangle\Big), \qquad \text{(A.6)}\end{aligned}$$

$$\begin{aligned}& \mathrm{KL}(q(\boldsymbol{\omega}) \parallel p(\boldsymbol{\omega}|\mathbf{a}^{\omega}, \mathbf{b}^{\omega})) \\ & = \sum_{k=1}^{q}\Big(\tilde{a}_k^{\omega}\ln\tilde{b}_k^{\omega} - a_k^{\omega}\ln b_k^{\omega} + \ln\Gamma(a_k^{\omega}) - \ln\Gamma(\tilde{a}_k^{\omega}) \\ & \quad + b_k^{\omega}\langle\tilde{\omega}_k\rangle - \tilde{a}_k^{\omega} + (\tilde{a}_k^{\omega} - a_k^{\omega})\langle\ln\tilde{\omega}_k\rangle\Big). \qquad \text{(A.7)}\end{aligned}$$

Note $\mathrm{diag}\langle\boldsymbol{\omega}\rangle$ in (A.2) is computed by the old distribution $q(\boldsymbol{\omega})$. (A.7) and $\mathrm{diag}\langle\widetilde{\boldsymbol{\omega}}\rangle$ in (A.5) are computed by the new (or updated) distribution $q(\boldsymbol{\omega})$.

Appendix B. Justification of using constraint (25) on A.

Let $\mathbf{T}$ be an orthogonal rotation matrix. From FA model definition (1), it is easy to see that FA model is invariant if we replace $\mathbf{A}$ by $\mathbf{AT}$ and $\mathbf{y}_i$ by $\mathbf{T}'\mathbf{y}_i$. This enables us to reparametrize $\mathbf{A}$ via rotation. Let $\mathbf{x}_i$ is generated from FA model in which $\mathbf{A}$ is in the general form no matter satisfying constraint (25) or not, i.e. $\mathbf{x}_i = \mathbf{A}\mathbf{y}_i + \boldsymbol{\epsilon}_i$. By QR decomposition, we have $\mathbf{A} = \mathbf{A}_c\mathbf{R}$, where $\mathbf{A}_c$ is a lower triangular matrix and $\mathbf{R}$ is an orthogonal matrix. Let $\tilde{\mathbf{y}}_i = \mathbf{R}\mathbf{y}_i$. We have $\mathbf{x}_i = \mathbf{A}_c\tilde{\mathbf{y}}_i + \boldsymbol{\epsilon}_i$ and $\tilde{\mathbf{y}}_i \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$. Therefore, $\mathbf{x}_i$ generated using general $\mathbf{A}$ can always be regarded as generated using $\mathbf{A}_c$. Due to this reason, we can use constraint (25) to estimate $\mathbf{A}_c$.

Appendix C. On the rotation problem of VBFA-ARD.

Let's start with ML estimation method that is well known for rotation invariance, which could enlighten us on how to judge whether VBFA-ARD bears rotation invariance or not. Let $\hat{\mathbf{A}}$ be a MLE of $\mathbf{A}$ and $\mathbf{R}$ be a rotation matrix. If we replace $\hat{\mathbf{A}}$ by $\hat{\mathbf{A}}\mathbf{R}$, then the log-likelihood $\mathcal{L}$ in (2) is invariant, i.e., $\mathcal{L}(\hat{\mathbf{A}}) = \mathcal{L}(\hat{\mathbf{A}}\mathbf{R})$. Notice that the rotation of $\mathbf{A}$ corresponds to a rotation of factors from $\mathbf{y}_i$ to $\mathbf{R}'\mathbf{y}_i$.

Consider VBFA1-ARD first. To do this, we simply need to replace, in the expression of lower bound $\mathcal{F}_1$, $\mathbf{A}$ and $\mathbf{y}_i$ by $\mathbf{AR}$ and $\mathbf{R}'\mathbf{y}_i$, respectively, (correspondingly, the posterior distributions $q(\mathbf{A})$ and $q(\mathbf{y}_i)$ by $q(\mathbf{AR})$ and $q(\mathbf{R}'\mathbf{y}_i)$) to see whether $\mathcal{F}_1$ will be changed or not. Note that $q(\mathbf{AR})$ and $q(\mathbf{R}'\mathbf{y}_i)$ here are obtained by transformations of $q(\mathbf{A})$ and $q(\mathbf{y}_i)$ with the rotation $\mathbf{R}$.

The terms related with $\mathbf{A}$ and $\mathbf{y}_i$ in the expression of $\mathcal{F}_1$ are given by

$$\begin{aligned}& \langle \ln p(\mathbf{X}|\mathbf{Y}, \mathbf{A}, \boldsymbol{\mu}, \boldsymbol{\varphi})\rangle_{q(\mathbf{Y})q(\mathbf{A})q(\boldsymbol{\varphi})q(\boldsymbol{\mu})} - \mathrm{KL}(q(\mathbf{Y}) \parallel p(\mathbf{Y})) \\ & \quad - \langle \mathrm{KL}(q(\mathbf{A}) \parallel p(\mathbf{A}|\boldsymbol{\omega}))\rangle_{q(\boldsymbol{\omega})}. \qquad \text{(C.1)}\end{aligned}$$

The first two terms in (C.1), discarding the part irrespective to $\mathbf{A}$ and $\mathbf{y}_i$, is

$$\begin{aligned}& \frac{1}{2}\sum_{i=1}^{N}\langle(\mathbf{x}_i - \mathbf{A}\mathbf{y}_i - \boldsymbol{\mu})'\boldsymbol{\varphi}(\mathbf{x}_i - \mathbf{A}\mathbf{y}_i - \boldsymbol{\mu})\rangle_{q(\mathbf{Y})q(\mathbf{A})q(\boldsymbol{\varphi})q(\boldsymbol{\mu})} \\ & \quad - \frac{1}{2}\sum_{i=1}^{N}\langle\mathbf{y}_i'\mathbf{y}_i\rangle - \langle\ln q(\mathbf{Y})\rangle. \qquad \text{(C.2)}\end{aligned}$$

Since $\langle \ln q(\mathbf{Y})\rangle = \frac{1}{2}\ln|\mathbf{I} + \langle\mathbf{A}'\boldsymbol{\varphi}\mathbf{A}\rangle|+$ const, replacing $\mathbf{A}$ and $\mathbf{y}_i$ in (C.2) by $\mathbf{AR}$ and $\mathbf{R}'\mathbf{y}_i$, the value of (C.2) will not change.

Next we analyze $\langle\mathrm{KL}(q(\mathbf{A}) \parallel p(\mathbf{A}|\boldsymbol{\omega}))\rangle_{q(\boldsymbol{\omega})}$. The part related with $\mathbf{A}$ and $\mathbf{y}_i$ is given by

$$\frac{1}{2}\left(\mathrm{tr}(\mathrm{diag}\langle\mathbf{A}'\mathbf{A}\rangle\mathrm{diag}\langle\boldsymbol{\omega}\rangle) - \sum_{j=1}^{d}\ln|\boldsymbol{\Sigma}_{\mathbf{a}}^{(j)}|\right), \qquad \text{(C.3)}$$

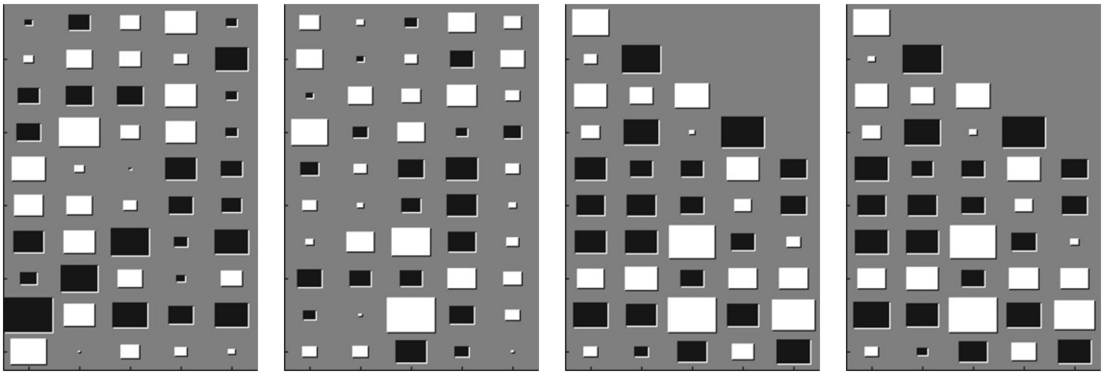
where $\boldsymbol{\Sigma}_{\mathbf{a}}^{(j)} = \left(\mathrm{diag}\langle\boldsymbol{\omega}\rangle + \langle\varphi_j\rangle\sum_i\langle\mathbf{y}_i\mathbf{y}_i'\rangle\right)^{-1}$. Since $\mathrm{diag}\langle\boldsymbol{\omega}\rangle$ is a diagonal matrix, replacing $\mathbf{A}$ and $\mathbf{y}_i$ in (C.3) by $\mathbf{AR}$ and $\mathbf{R}'\mathbf{y}_i$, the value of (C.3) would generally change except that $\mathbf{R}$ is a diagonal matrix with diagonal entries ± 1. This could be easier to observe when the prior $p(\boldsymbol{\omega})$ is so weak that $\langle\omega_k\rangle \approx \frac{d}{\langle\mathbf{a}_k'\mathbf{a}_k\rangle}$ (see (19) and (20)), and (C.3) becomes

$$-\frac{1}{2}\left(\sum_{j=1}^{d}\ln|\boldsymbol{\Sigma}_{\mathbf{a}}^{(j)}|\right) + \mathrm{const}, \qquad \text{(C.4)}$$

which is clearly not invariant against general rotation. A similar conclusion can be drawn for VBFA2-ARD, e.g. with respect to a rotation from lower triangular form (25) to upper triangular one.

Given an initialization of $q(\cdot)'s$, the above analysis implies that VBFA-ARD will generally lead to an unique $q(\mathbf{A})$ in the sense that a rotation of $q(\mathbf{A})$ would generally decrease the value of the lower bound $\mathcal{F}$. It is natural to ask whether all different initializations of $q(\cdot)'s$ will result in the same $q(\mathbf{A})$. In other words, whether the true $\mathbf{A}$ could be fully recovered or not. To answer this, we conduct a small simulation study. The estimated factor loadings matrix $\langle\mathbf{A}\rangle$ by VBFA1 shown in Fig. C.1(b) is not similar to the true factor loadings matrix drawn from Gaussian distribution shown in Fig. C.1(a). Nevertheless, results of rotations from $\mathbf{A}$ to $\mathbf{A}_c$ and $\langle\mathbf{A}\rangle$ to $\langle\mathbf{A}_c\rangle$ by QR decomposition shown in Fig. C.1(c) and (d) are in fact very similar to each other. This indicates that the factor loadings matrix using VBFA1-ARD can only be determined up to a rotation.

In summary, ① our simulation study reveals that the factor loadings matrix in VBFA-ARD, like ML method, can be determined up to a rotation. VBFA2-ARD estimates the ones satisfying constraint (25) while VBFA1-ARD estimates the ones that might depend on initialization (similar to EM in ML). ② unlike ML, VBFA-ARD is variant with rotation in general.



(a) A.

(b) $\langle A \rangle$ by VBFA1.

(c) A_c.

(d) $\langle A_c \rangle$ by VBFA1.

Fig. C.1. Hinton diagrams of factor loadings matrices. (a) True $\mathbf{A}$, (b) $\langle \mathbf{A} \rangle$ by VBFA1, (c) $\mathbf{A}_c$ by QR decomposition of $\mathbf{A}$ and (d) $\langle \mathbf{A}_c \rangle$ by QR decomposition of $\langle \mathbf{A} \rangle$ from VBFA1. The sample size used is 1000.

References

Akaike, H. (1987). Factor analysis and AIC. *Psychometrika*, *52*(3), 317–332.

Attias, H. (1999). Inferring parameters and structure of latent variable models by Variational Bayes. In *Proceedings of the fifteenth conference on uncertainty in artificial intelligence*.

Beal, M. J. (2003). Variational Algorithms for approximation Bayesian inference. *Ph.D. thesis*. The University of London.

Bishop, C. M. (2006). *Pattern recognition and machine learning*. Springer.

Dempster, A. P., Laird, N. M., & Rubin, D. B. (1977). Maximum likelihood from incomplete data using the EM algorithm. *Journal of the Royal Statistical Society Series B*, *39*(1), 1–38. (with discussion).

Fokoué, E., & Titterington, D. M. (2003). Mixtures of factor analysers: Bayesian estimation and inference by stochastic simulation. *Machine Learning*, *50*, 73–94.

Ghahramani, Z., & Beal, M. (2000). Variational inference for Bayesian mixture of factor analysers. In *Advances in neural information proceeding systems*. Cambridge, MA: MIT Press.

Harva, M., & Kabán, A. (2007). Variational learning for rectified factor analysis. *Signal Processing*, *87*(3), 509–527.

Jöreskog, K. G. (1967). Some contributions to maximum likelihood factor analysis (32) (4) 433–482.

Nielsen, F. B. (2004). Variational approach to factor analysis and related models. *Master's thesis*. The Institute of Informatics and Mathematical Modelling, Technical University of Denmark.

Oba, S., Sato, M., & Ishii, S. (2003). Prior hyperparameters in Bayesian PCA. In *Lecture notes in computer science*: *Vol. 2714* (pp. 271–279).

Schwarz, G. (1978). Estimating the dimension of a model. *The Annals of Statistics*, *6*, 461–464.

Shao, J. (1997). An asymptotic theory for linear model selection. *Statistica Sinica*, *7*, 221–242. with discussion.

Wang, C. (2007). Variational Bayesian approach to canonical correlation analysis. *IEEE Transctions on Neural Networks*, *18*(3), 905–910.

Zhao, J., Yu, P. L. H., & Jiang, Q. (2008). ML estimation for factor analysis: EM or non-EM? *Statistics and Computing*, *18*(2), 109–123.