Journal of Mathematical Modelling and Algorithms (2006) **5:** 447–473
DOI 10.1007/s10852-005-9021-2

A Comparative Investigation on Model Selection in Independent Factor Analysis

YUJIA AN★, XUELEI HU and LEI XU
Department of Computer Science and Engineering, The Chinese University of Hong Kong, Shatin, N.T., Hong Kong. e-mail: {yjan, xlhu, lxu}@cse.cuhk.edu.hk

Abstract. With uncorrelated Gaussian factors extended to mutually independent factors beyond Gaussian, the conventional factor analysis is extended to what is recently called independent factor analysis. Typically, it is called binary factor analysis (BFA) when the factors are binary and called non-Gaussian factor analysis (NFA) when the factors are from real non-Gaussian distributions. A crucial issue in both BFA and NFA is the determination of the number of factors. In the literature of statistics, there are a number of model selection criteria that can be used for this purpose. Also, the Bayesian Ying-Yang (BYY) harmony learning provides a new principle for this purpose. This paper further investigates BYY harmony learning in comparison with existing typical criteria, including Akaik's information criterion (AIC), the consistent Akaike's information criterion (CAIC), the Bayesian inference criterion (BIC), and the cross-validation (CV) criterion on selection of the number of factors. This comparative study is made via experiments on the data sets with different sample sizes, data space dimensions, noise variances, and hidden factors numbers. Experiments have shown that for both BFA and NFA, in most cases BIC outperforms AIC, CAIC, and CV while the BYY criterion is either comparable with or better than BIC. In consideration of the fact that the selection by these criteria has to be implemented at the second stage based on a set of candidate models which have to be obtained at the first stage of parameter learning, while BYY harmony learning can provide not only a new class of criteria implemented in a similar way but also a new family of algorithms that perform parameter learning at the first stage with automated model selection, BYY harmony learning is more preferred since computing costs can be saved significantly.

Mathematical Subject Classifications (2000): 68Q10, 62H25, 68T05, 65C10, 68Q32.

Key words: BYY harmony learning, hidden factor, binary factor analysis, non-Gaussian factor analysis, model selection, automatic model selection.

1. Introduction

Factor analysis (FA) is a well-known multivariate analysis technique in help of the following linear model [3, 35].

$$x = Ay + c + e, \tag{1}$$

★ Corresponding author.

where x is a d-dimensional random vector of observable variables, e is a d-dimensional random vector of unobservable noise variables and is drawn from Gaussian, A is a $d \times k$ loading matrix, c is a d-dimensional mean vector and y is a k-dimensional random vector of unobservable mutually uncorrelated Gaussian factors. y and e are mutually independent. However, FA is not appropriately applicable to the real world data that cannot be described as generated from Gaussian factors. With uncorrelated Gaussian factors extended to mutually independent factors beyond Gaussian, FA is extended to what is recently called independent factor analysis. Typically, it is called binary factor analysis (BFA) when the factors are binary and called non-Gaussian factor analysis (NFA) when the factors are from real non-Gaussian distributions.

In practice, BFA has been widely used in various fields, especially the social science such as the area of political science, educational testing, psychological measurement as well as the tests of disease severity, etc. [26]. Also, it can be used for data reduction [30]. There are many studies made on BFA, under the names of latent trait model (LTA), item response theory (IRT) models as well as latent class model [5, 14]. Another typical example is the multiple cause model that considers the observed samples generated from independent binary hidden factors [13, 22]. One other type of examples includes the auto-association network that is trained by back-propagation via simply copying input as the desired output [8, 9], and the LMSER self-organizing [27].

NFA can avoid the rotation and additive indeterminacies encountered by classical FA. It also relaxes the impractical noise-free assumption for independent component analysis (ICA) [32, 34]. In recent years, studies related to NFA have been carried out. One kind of examples consists of the efforts under the name noisy ICA. One example is given in [15] where a so-called joint maximum likelihood is considered to be maximized. However, a rough approximation is actually used there and also how to specify a scale remains an open problem yet [34]. Some other noisy ICA examples are also referred to [12, 16]. In [19], an approach that exactly implements ML learning for the model Equation (1) was firstly proposed. Similar to [29, 37], they considered modelling each non-Gaussian factor by a Gaussian mixture. In help of a trick that the product of summations is equivalently exchanged into a summation of products, the integral in computing likelihood becomes a summation of a large number of analytically computable integrals on Gaussians, which makes an exact ML learning on Equation (1) implemented by an exact EM algorithm. The same approach has been also published in [4] under the name of independent factor analysis. However, the number of terms of computable integrals on Gaussians grows exponentially with the number of factors and it correspondingly incurs exponentially growing computing costs. In contrast, computing costs of implementing the NFA algorithms in [32, 34] grow only linearly with the number of factors, as demonstrated in [18, 34].

One other crucial issue in implementing BFA and NFA is appropriately determining the number of hidden factors, i.e., the dimension of y in Equation (1), which is a typical model selection problem. In literature of statistical learning, many efforts have been made on model selection via a two-phase style implementation that first conducts parameter learning on a set of candidate models under the maximum likelihood (ML) principle and then select the ‘optimal’ model among the candidates according to a given model selection criterion. Popular examples of such criteria include the Akaike’s information criterion (AIC) [1, 2], Bozdogan’s consistent Akaike’s information criterion (CAIC) [10], Schwarz’s Bayesian inference criterion (BIC) [23] which formally coincides with Rissanen’s minimum description length (MDL) criterion [6, 20], and cross-validation (CV) criterion [25]. Such criteria can also be used to the model selection of BFA and NFA, in help of a two-phase implementation.

The Bayesian Ying-Yang (BYY) harmony learning was proposed as a unified statistical learning framework firstly in 1995 [28] and systematically developed in past years. BYY harmony learning consists of a general BYY system and a fundamental harmony learning principle as a unified guide for developing new regularization techniques, a new class of criteria for model selection, and a new family of algorithms that perform parameter learning with automated model selection [32–34, 36]. By applying BYY learning to BFA and NFA respectively, not only new criteria for model selection in BFA and NFA are obtained, but also adaptive algorithms are developed that perform BFA and NFA with an appropriate number of hidden factors automatically determined during adaptive learning [32]. This paper further investigates the BYY model selection criterion in comparison with the criteria of AIC, CAIC, BIC, and CV for the model selection of BFA and NFA respectively.

This comparative study is carried out via experiments on simulated data sets with different sample sizes, data space dimensions, noise variances, and hidden factors numbers. Experiments on both BFA and NFA have shown that the performance of BIC is superior to AIC, CAIC, and CV in most cases. The BYY criterion is, in most cases, comparable with or even superior to the best among of BIC, AIC, CAIC, and CV. In consideration that the selection by these criteria has to be implemented at the second stage based on a set of candidate models which have to be obtained at the first stage of parameter learning, while BYY harmony learning can also be implemented by algorithms that perform parameter learning with automated model selection with computing costs saved significantly, BYY harmony learning is a more preferred tool for NFA.

The rest of this paper is organized as follows. Section 2 briefly describes the BFA and NFA via the maximum likelihood learning. In Section 3, we further introduce not only the criteria AIC, CAIC, BIC, and CV, but also the BYY criterion for the model selection of BFA and NFA. Comparative experiments are given in Section 4 and a conclusion is made in Section 5.

2. ML Learning for Binary Factor Analysis and Non-Gaussian Factor Analysis

2.1. BINARY FACTOR ANALYSIS AND ML LEARNING

The model Equation (1) is called binary factor analysis (BFA) when y is a binary random vector, which comes from the following multivariate Bernoulli distribution [32, 35]:

$$p(y) = \prod_{j=1}^{k} \left[q_j \delta\left(y^{(j)}\right) + (1 - q_j)\delta\left(1 - y^{(j)}\right) \right], \tag{2}$$

where q_j is the probability that $y^{(j)}$ takes the value 1. In this paper, we consider that e is from a spherical Gaussian distribution for the binary factor analysis model, i.e., $p(x|y)$ has the following form:

$$p(x|\, y) = G\left(x|Ay + c,\, \sigma^2 I\right) \quad (\text{for BFA}) \tag{3}$$

where $G(x|Ay + c, \sigma^2 I)$ denotes a multivariate normal (Gaussian) distribution with mean $Ay + c$ and spherical covariance matrix $\sigma^2 I$, with I being a $d \times d$ identity matrix.

Given k and a set of observations $\{x_t\}_{t=1}^{n}$, the parameters $\theta = \{A, c, \sigma^2\}$ is usually estimated via maximum likelihood learning. That is,

$$\hat{\theta} = \arg\max_{\theta} L(\theta). \tag{4}$$

For BFA, $L(\theta)$ is the following log likelihood function

$$L(\theta) = \sum_{t=1}^{n} \ln\left(p(x_t)\right) = \sum_{t=1}^{n} \ln\left(\sum_{y \in D} p(x_t|y)p(y)\right), \tag{5}$$

where D is the set that contains all possible values of a binary vector y.

This problem Equation (4) can be implemented by the EM algorithm that iterates the following steps [7, 30]:

Step 1: calculate $p(y|x_t)$ by

$$p(y|x_t) = \frac{p(x_t|y)p(y)}{\sum_{y \in D} p(x_t|y)p(y)}. \tag{6}$$

Step 2: update A, c and σ^2 by

$$A = \left(\sum_{t=1}^{n}\sum_{y \in D} p(y|x_t)(x_t - c)y^T\right)\left(\sum_{t=1}^{n}\sum_{y \in D} p(y|x_t)yy^T\right)^{-1}, \tag{7}$$

$$c = \frac{1}{n}\sum_{t=1}^{n}\sum_{y\in D} p(y|x_t)(x_t - Ay), \tag{8}$$

and

$$\sigma^2 = \frac{1}{dn}\sum_{t=1}^{n}\sum_{y\in D} p(y|x_t)\|e_t\|^2 \tag{9}$$

where $e_t = x_t - A_y - c$.

2.2. NON-GAUSSIAN FACTOR ANALYSIS AND ML LEARNING

The model Equation (1) is called non-Gaussian factor analysis (NFA) when y is a non-Gaussian random real vector. In this paper we consider that each non-Gaussian factor $y^{(j)}$ is described by a Gaussian mixture distribution [19, 32, 34]:

$$p(y) = \prod_{j=1}^{k}\left[p_j\left(y^{(j)}\right)\right], \; p_j\left(y^{(j)}\right) = \sum_{q_j=1}^{k_j}\beta_{j,q_j} G\left(y^{(j)}|\mu_{j,q_j}, \sigma^2_{j,q_j}\right). \tag{10}$$

Also, we consider $p(x|y)$ in the following form:

$$p(x|y) = G(x|Ay, \Sigma), \tag{11}$$

where Σ is the covariance matrix of error e. For simplification, we preprocess the observable data to be zero mean such that the unknown parameter c in Equation (1) can be ignored.

In implementation of NFA, the unknown parameters θ consists of the mixing matrix A, the covariance matrix Σ, and the parameters $\theta_j = \left\{\beta_{j,q_j}, \mu_{j,q_j}, \sigma^2_{j,q_j}\right\}$ for each factor $y^{(j)}$. Given a number k of factors and the number k_j for each factor $y^{(j)}$ (denoted by $K = \{k, \{k_j\}\}$) as well as a set of observations $\{x_t\}_{t=1}^{n}$, maximum likelihood learning by Equation (4) is still used for estimating θ, with the following log likelihood function:

$$L(\theta) = \sum_{t=1}^{n}\ln\left(p(x_t)\right) = \sum_{t=1}^{n}\ln\left(\int p(x_t|y)p(y)dy\right), \tag{12}$$

which cannot be implemented by the EM algorithm as the integral in this function is analytically intractable. In [19], a missing data $q = [q_1, q_2, \ldots, q_k]$ is used to indicate which factor is generated by the corresponding Gaussian component, and then $p(y)$ in Equation (10) is expressed as a mixture of Gaussian products. As a result, the integral becomes a summation of a large number of analytically computable integrals on Gaussians, which makes an exact ML learning on Equation (1) implemented by an exact EM algorithm. The same approach has been also published in [4] under the name of independent factor analysis.

Specifically, each state q_j in q indicates the factor $y^{(j)}$ generated by the q_jth Gaussian component and each state q corresponds to a k-dimensional Gaussian density with the mixing proportion β_q, mean μ_q, and diagonal covariance matrix V_q as follows:

$$\begin{aligned} \beta_q &= \prod_{j=1}^{k} \beta_{j,q_j} = \beta_{1,q_1} \cdot \ldots \cdot \beta_{k,q_k}, \\ \mu_q &= \left[\mu_{1,q_1}, \ldots, \mu_{k,q_k}\right]^T, \\ V_q &= \operatorname{diag}\left(\sigma^2_{1,q_1}, \ldots, \sigma^2_{k,q_k}\right), \end{aligned} \tag{13}$$

where the notation diag$(d_1, \ldots, d_k)$ denotes a diagonal matrix with the diagonal elements being $d_1, \ldots, d_k$.

Thus, the form of $p(y)$ in Equation (10) can be rewritten as

$$p(y) = \sum_q \beta_q G\left(y\middle|\mu_q, V_q\right), \tag{14}$$

where the summation $\sum_q = \sum_{q_1}, \ldots, \sum_{q_k}$. Also, the form of $p(q)$ and $p(x|q)$ can be written as

$$p(q) = \beta_q, \quad p(x|q) = G\left(x|A_{\mu_q}, AV_qA^T + \Sigma\right). \tag{15}$$

The EM algorithm for solving Equation (12) is given in the following steps [4, 19]:

Step E: calculate $p(q|x_t)$ by

$$p(q|x_t) = \frac{p(q)p(x_t|q)}{\sum_q p(q)p(x_t|q)}, \tag{16}$$

$$p\left(q_j|x_t\right) = \sum_{\{q_i\}_{i \neq j}} p(q|x_t), \tag{17}$$

where $\sum_{\{q_i\}_{i \neq j}}$ denotes summation over $\{q_{i \neq j}\}$, holding q_j fixed.

Step M: update A, Σ and β_{j,q_j}, μ_{jq_j}, σ^2_{j,q_j} by

$$A = \sum_{t=1}^{n} x_t \langle y_t^T \rangle \left(\sum_{t=1}^{n} \langle y_t y_t^T \rangle \right)^{-1}, \tag{18}$$

$$\Sigma = \frac{1}{n} \sum_{t=1}^{n} x_t x_t^T - \frac{1}{n} \sum_{t=1}^{n} x_t \langle y_t^T \rangle A^T, \tag{19}$$

$$\mu_{j,q_j} = \frac{\sum_{t=1}^{n} p(q_j|x_t)\left\langle y_t^{(j)}\right\rangle}{\sum_{t=1}^{n} p(q_j|x_t)}, \tag{20}$$

$$\sigma_{j,q_j}^2 = \frac{\sum_{t=1}^{n} p(q_j|x_t)\left\langle y_t^{(j)2}\right\rangle}{dn} - \mu_{j,q_j}^2, \tag{21}$$

$$\beta_{j,q_j} = \tfrac{1}{n}\sum_{t=1}^{n} p(q_j|x_t), \tag{22}$$

respectively, where

$$\begin{aligned}
\langle y_t^T \rangle &= \sum_q p(q|x_t)h_q, \\
\langle y_t y_t^T \rangle &= \sum_q p(q|x_t)\left(h_q h_q^T + M_q\right), \\
\left\langle y_t^{(j)} \right\rangle &= \sum_{\{q_i\}_{i\neq j}} p(q|x_t)\left(h_q\right)_j, \\
\left\langle y_t^{(j)2} \right\rangle &= \sum_{\{q_i\}_{i\neq j}} p(q|x_t)\left(h_q h_q^T + M_q\right)_{jj}, \\
M_q &= \left(A^T \sum\nolimits^{-1} A + V_q^{-1}\right)^{-1}, \\
h_q &= M_q\left(A^T \sum\nolimits^{-1} x_t + V_q^{-1}\mu_q\right),
\end{aligned} \tag{23}$$

where $(h_q)_j$ means the jth element of vector h_q.

Obviously, the number of terms in the summation $\sum_q = \sum_{q_1}, \ldots, \sum_{q_k}$ grows exponentially with k, and correspondingly incurs that computing costs exponentially grows with k. This is a serious disadvantage of this approach.

3. Factor Number Determination

3.1. TYPICAL MODEL SELECTION CRITERIA

The other crucial issue in implementing BFA and NFA is appropriately determining k for BFA and $K = \{k, \{k_j\}\}$ for NFA, which is a typical model selection problem. In literature of statistical learning, many efforts have been made on model selection via a two-phase style implementation that first conducts maximum likelihood learning on a set of candidate models and then selects the 'optimal' model among the candidates according to a given model selection criterion. Here, we only describe the detail for NFA as an example. BFA is similar except that only k need to be determined.

Given a range of values of k from k_{min} to k_{max} and a range of values of each k_j from $k_{j_{\min}}$ to $k_{j_{\max}}$, which forms a domain D for $K = \{k, k_j\}$. First, we estimate the parameters θ under the ML learning principle at each specific k and $K \in D$. Second, we make the following selection

$$\hat{K} = \arg \min_K \left\{ J\left(\hat{\theta}, K\right), K \in D \right\}, \tag{24}$$

where $J(\hat{\theta}, K)$ is a given model selection criterion.

Three typical model selection criteria are the Akaike's information criterion (AIC) [1, 2], its extension called Bozdogan's consistent Akaike's information criterion (CAIC) [10], and Schwarz's Bayesian inference criterion (BIC) [23] which coincides with Rissanen's minimum description length (MDL) criterion [6, 20]. These three criteria can be summarized into the following general form [24]

$$J\left(\hat{\theta}, K\right) = -2L\left(\hat{\theta}\right) + W(n)U(K) \tag{25}$$

where $L(\hat{\theta})$ is the log likelihood Equation (4) based on the ML estimation $\widehat{\theta}$ under a given K, and $U(K) = (d - 2)k + 0.5d(d + 1) + \sum_{j}^{k} = {}_1(3k_j - 1) + 1$ is the number of free parameters in the K-model for NFA. Moreover, $W(n)$ is a function with respect to the number of observations as follows:

- $W(n) = 2$ for Akaike's information criterion (AIC) [1, 2],
- $W(n) = \ln(n) + 1$ for Bozdogan's consistent Akaike's information criterion (CAIC) [10],
- $W(n) = \ln(n)$ for Schwarz's Bayesian inference criterion (BIC) [23].

For BFA, the number of free parameters in a k-factor model is $U(k) = d(k + 1)+1$.

Another well-known model selection technique is cross-validation (CV), by which data are repeatedly partitioned into two sets, one is used to build the model and the other is used to evaluate the statistic of interest [25]. For the ith partition, let E_i be the data subset used for testing and E_{-i} be the remainder of the data used for training, the cross-validated log likelihood for a K-factor model is

$$J\left(\hat{\theta}, K\right) = -\frac{1}{m}\sum_{i=1}^{m} L\left(\hat{\theta}(E_{-i})|E_i\right) \tag{26}$$

where m is the number of partitions, $\hat{\theta}(E_{-i})$ denotes the ML parameter estimates from the ith training subset, and $L(\hat{\theta}(E_{-i})|E_i)$ is the log likelihood evaluated on the data set E_i. Featured by m, it is usually referred as making a m-fold cross-validation or shortly m-fold CV.

3.2. BYY HARMONY LEARNING

Bayesian Ying-Yang (BYY) learning was proposed as a unified statistical learning framework firstly in [28] and systematically developed in past years.

From the perspective of general learning framework, the BYY harmony learning consists of a general BYY system and a fundamental harmony learning principle as a unified guide for developing new regularization techniques, a new class of criteria for model selection, and a new family of algorithms that perform parameter learning with automated model selection. From the perspective of specific learning paradigms, the BYY learning with specific structures applies to unsupervised learning, supervised learning, and state space approach for temporal modelling, with a number of new results. The details are referred to [31–36].

3.2.1. *Determining Hidden Factors by BYY Criterion for BFA*

Applying BYY harmony learning to a BFA model, the following criterion is obtained for selecting the factor number k [30, 35]

$$J\left(\hat{\theta}, K\right) = k \ln 2 + 0.5d \ln \hat{\sigma}^2 \;(\text{for BFA}). \tag{27}$$

We refer it shortly by BYY criterion, where $\widehat{\sigma}^2$ can be obtained via BYY harmony learning, i.e.,

$$\hat{\theta} = \arg\max_{\theta} H\left(\theta, \hat{k}\right). \tag{28}$$

According to Section 4.1, especially Equations (21), (30)–(32), in [35], the specific form of $H(\theta, k)$ is given as follows

$$H(\theta, k) = \frac{d}{2} \ln \sigma^2 + \frac{1}{2n} \sum_{t=1}^{n} \frac{\|x_t - Ay_t - c\|^2}{\sigma^2} - \frac{1}{n} \sum_{t=1}^{n} \sum_{j=1}^{k} \left[y_t^{(j)} lnq_j + \left(1 - y_t^{(j)}\right) ln\left(1 - q_j\right)\right], \tag{29}$$

where $y_t^{(j)}$ is the j-th element in vector y_t, and

$$y_t = \arg\max_{y} H(\theta, k). \tag{30}$$

In a two-stage implementation, q_j is simply preset as 0.5.

The above Equation (28) can be implemented by either a batch or an adaptive algorithm. Specifically, with k fixed, BFA can be implemented via the adaptive algorithm given by Table I in [35]. However, typical model selection criteria are evaluated in this paper basing on the ML estimation via the EM algorithm made in batch, we write the procedure given in Table I of [35] into its corresponding batch version that iterates the following steps:

Step 1: get y_t by Equation (30),

Table I. Numbers of underestimating (U), success (S), and overestimating (O) by each criterion on the data sets with different sample sizes for BFA in 100 experiments

Criteria	n = 20			n = 40			n = 100		
	U	S	O	U	S	O	U	S	O
AIC	5	53	42	1	78	21	0	89	11
BIC	15	81	4	3	97	0	0	100	0
CAIC	32	67	1	10	90	0	1	99	0
10-fold CV	3	62	35	1	79	20	0	90	10
BYY	4	77	19	1	93	6	0	100	0

Step 2: from $\frac{\partial H(\theta,k)}{\partial\theta} = 0$, ($\theta$ include A, c, σ^2), update

$$\begin{aligned} e_t &= x_t - Ay_t - c, \\ \sigma^2 &= \frac{\sum_{t=1}^{n} \|e_t\|^2}{dn}, \\ A &= \left(\sum_{t=1}^{n} (x_t - c)y_t^T\right)\left(\sum_{t=1}^{n} y_t y_t^T\right)^{-1}, \\ c &= \frac{1}{n}\sum_{t=1}^{n} (x_t - Ay_t). \end{aligned} \tag{31}$$

This iterative procedure is guaranteed to converge since it is actually the specific form of the Ying-Yang alternative procedure, see Section 3.2 in [34].

With k enumerated as in Equation (24) and its corresponding parameters obtained by the above Equation (31), we can select a best value of k for BFA by the BYY criterion in Equation (27).

3.2.2. *Determining Hidden Factors by BYY Criterion for NFA*

Applying the BYY harmony learning to a NFA model, the following criterion for NFA is obtained for selecting $K = \{k, k_j\}$ [34, 36]

$$\begin{aligned} J\left(\hat{\theta}, K\right) &= \frac{1}{2}\ln\left|\hat{\Sigma}\right| + \frac{k}{2}(1 + \ln(2\pi)) \\ &\quad + \sum_{j=1}^{k}\sum_{q_j=1}^{k_j} \hat{\beta}_{j,q_j}\left(\frac{1}{2}\ln\hat{\sigma}^2{}_{j,q_j} - \ln\hat{\beta}_{j,q_j}\right). \end{aligned} \tag{32}$$

We refer it shortly by BYY criterion.

According to Section IVA and especially Equation (52) in [36], the specific form of $H(\theta, K)$ for NFA is given as follows

$$H(\theta,\, k) = -0.5\ln|\Sigma| - \frac{1}{n}\sum_{t=1}^{n}\sum_{j=1}^{k}\ln p_j\left(y_t^{(j)}\right), \tag{33}$$

$$\Sigma = \tfrac{1}{n}\sum_{t=1}^{n} e_t e_t^T,\ e_t = x_t - Ay_t,$$

where $p_j(\, y_t^{(j)})$ is same as given in Equation (10) and $y_t^{(j)}$ is the j-th element in vector y_t, and we have

$$y_t = y(x_t) = \arg\max_{y} H(\theta,\, K). \tag{34}$$

which is specifically implemented via a nonlinear optimization algorithm. In this paper, we adopt a so-called fixed posterior approximation technique that iterates the following two steps [18, 32, 36]:

$$\begin{aligned}
\text{Step(a): } p_{j,q_j} &= \frac{\beta_{j,q_j} G\left(y_t^{(j)}|\mu_{j,q_j}, \sigma^2_{j,q_j}\right)}{\sum_{q_j=1}^{k_j}\beta_{j,q_j} G\left(y_t^{(j)}|\mu_{j,q_j}, \sigma^2_{j,q_j}\right)}, \\
b_j &= \sum_{q_j=1}^{k_j}\frac{p_{j,q_j}}{\sigma^2_{j,q_j}}, d_j = \sum_{q_j=1}^{k_j}\frac{p_{j,q_j}\mu_{j,q_j}}{\sigma^2_{j,q_j}}, \\
\text{Step(b): } y_t^{\text{new}} &= \left(A^T\textstyle\sum^{-1}A + \text{diag}(b_1, \ldots, b_k)\right)^{-1} \\
&\quad \times\left(A^T\textstyle\sum^{-1}x_t + [d_1, \ldots, d_k]^T\right).
\end{aligned} \tag{35}$$

Moreover, with $K = \{k, k_j\}$ fixed, $\max_\theta H\left(\theta,\, \hat{k}\right)$ can be implemented via the algorithm given by Equation (57) in [36]. Here, we rewrite the algorithm that iterate the following steps:

Yang step: get y_t by Equation (35),
Ying step: (a) updating parameters in $p(x\,|\,y)$

$$\begin{aligned}
&e_t = x_t - A_{y_t}, \\
&\Sigma^{\text{new}} = (1-\eta)\Sigma^{\text{old}} + \eta e_t e_t^T, \\
&A^{\text{new}} = A^{\text{old}} + \eta e_t y_t^T.
\end{aligned}$$

(b) updating parameters in $p(y)$

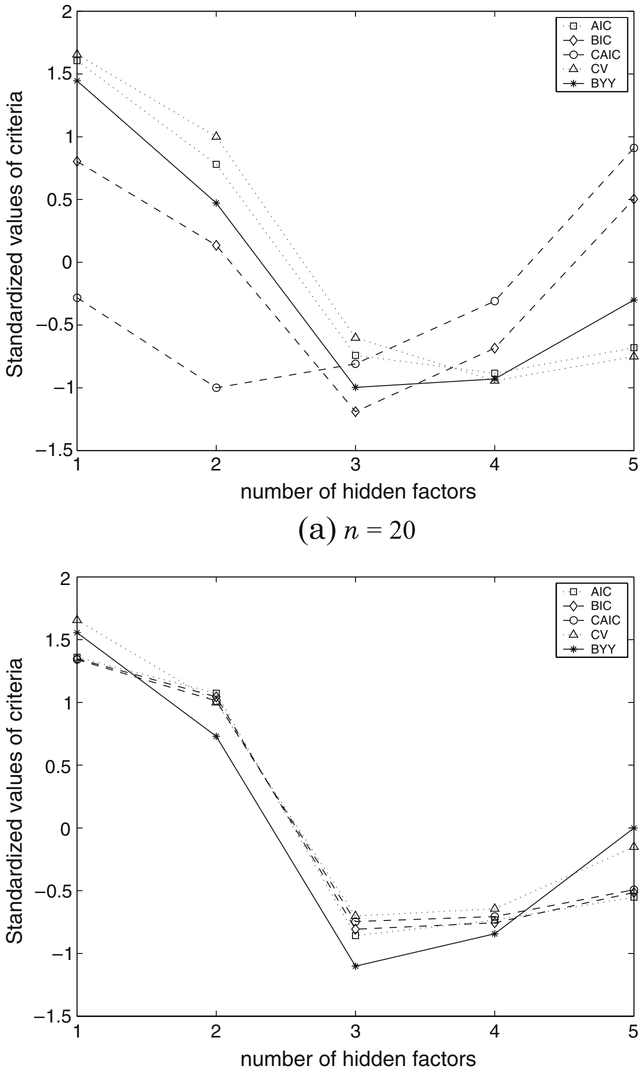
$$
\begin{aligned}
p_{j,q_j} &= \frac{\beta_{j,q_j} G\left(y_t^{(j)} | \mu_{j,q_j}, \sigma_{j,q_j}^2\right)}{\sum_{q_j=1}^{k_j} \beta_{j,q_j} G\left(y_t^{(j)} | \mu_{j,q_j}, \sigma_{j,q_j}^2\right)}, \\
\beta_{j,q_j}^{\text{new}} &= (1-\eta_0)\beta_{j,q_j}^{\text{old}} + \eta_0 p_{j,q_j}, \\
\mu_{j,q_j}^{\text{new}} &= \mu_{j,q_j}^{\text{old}} + \eta_0 p_{j,q_j}\left(y_t^{(j)} - \mu_{j,q_j}^{\text{old}}\right), \\
\sigma_{j,q_j}^{2\ \text{new}} &= \left(1-\eta_0 p_{j,q_j}\right)\sigma_{j,q_j}^{2\ \text{old}} + \eta_0 p_{j,q_j}\left(y_t^{(j)} - \mu_{j,q_j}^{\text{old}}\right)^2, \\
\mu_j &= \sum_{q_j=1}^{k_j} \beta_{j,q_j}^{\text{new}} \mu_{j,q_j}^{\text{new}}, \\
\sigma_j^2 &= \sum_{q_j=1}^{k_j} \beta_{j,q_j}^{\text{new}} \sigma_{j,q_j}^{2\ \text{new}}, \\
\mu_{j,q_j}^{\text{new}} &= \frac{\mu_{j,q_j}^{\text{new}} - \mu_j}{\sigma_j}, \sigma_{j,q_j}^{2\ \text{new}} = \frac{\sigma_{j,q_j}^{\text{new}}}{\sigma_j^2},
\end{aligned}
\tag{36}
$$

where η, η_0 are step length constants. In this paper, for simplification, we set all k_js to be a same integer. This iterative procedure is guaranteed to converge since it is actually the specific form of the Ying-Yang alternative procedure, see Section III in [32]. With K enumerated as in Equation (24) and its corresponding parameters obtained as the above, we can select a best K by BYY criterion in Equation (32).

Besides the above criteria based selection, adaptive algorithms have also been developed from BYY harmony learning to implement BFA and NFA such that an appropriate k or K can be automatically determined during adaptive learning [32, 34, 36]. The hidden factors obtained via either the automatic determination or the above criterion have no difference in performances. The difference lays in that the automatic determination can save significantly computational costs because avoiding the conventional two stage implementation. Therefore, as long as the performances by the criterion from BYY harmony learning are comparable to typical criteria AIC, CAIC, BIC, and CV, we prefer to use BYY harmony learning as a tool for determining k for BFA and K for NFA.

4. Comparative Empirical Studies

We investigate the experimental performances of the model selection criteria AIC, BIC, CAIC, 10-fold CV, BYY criterion for BFA and NFA on four types of data sets with different settings. In implementation of BFA, A, c and σ^2 are estimated via Equation (31) for BYY criterion, and via the EM algorithm in Equations (7)–(9) for AIC, BIC, CAIC and 10-fold CV. In implementation of



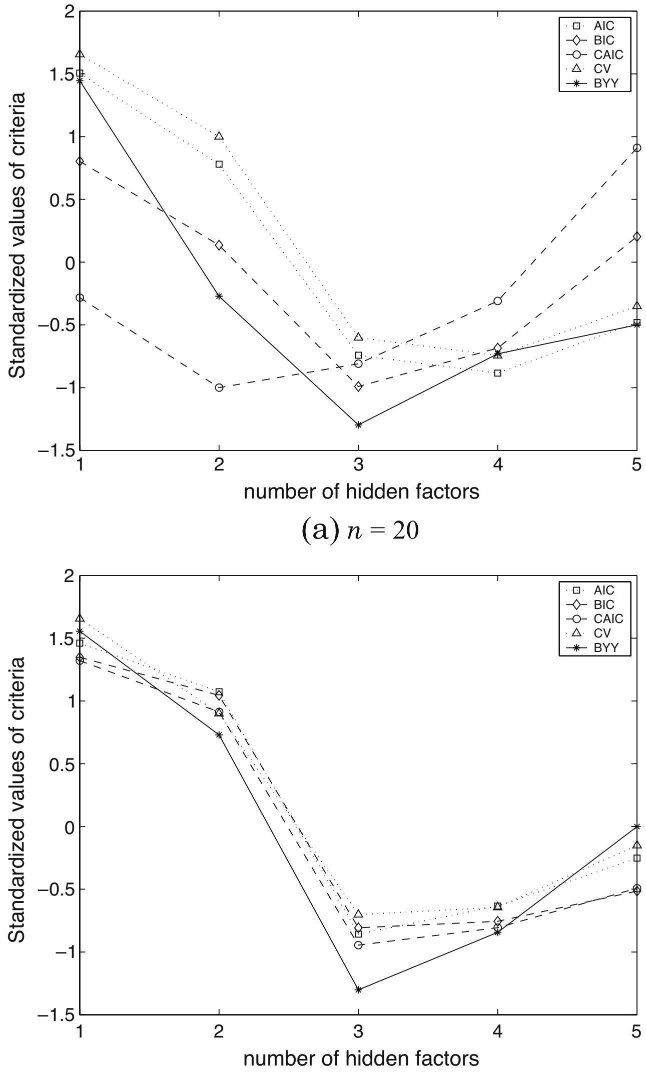
(a) $n = 20$

(b) $n = 100$

Figure 1. The curves obtained by the criteria AIC, BIC, CAIC, 10-fold CV and BYY on the data sets of a nine-dimensional x ($d = 9$) generated from a three-dimensional y ($k = 3$) with different sample sizes for BFA.

NFA, we implement parameters learning to obtain all the parameters θ via the algorithm Equation (36) for BYY criterion, and via the EM algorithm Equations (16)–(23) for AIC, BIC, CAIC and 10-fold CV. In addition, to clearly illustrate the curve of each criterion within one same figure we normalize the values of each curve to zero mean and unit variance.

For BFA, we design four groups of experiments to illustrate the performance of each criterion on data sets with different sample sizes, data dimensions, noise

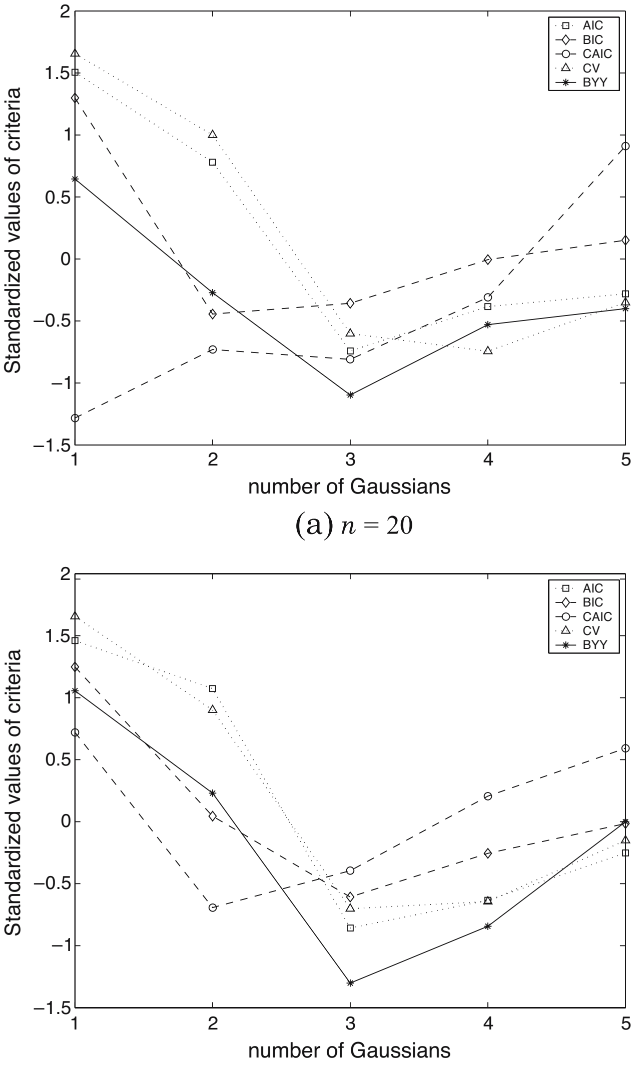


(a) $n = 20$

(b) $n = 100$

Figure 2. The curves obtained by the criteria AIC, BIC, CAIC, 10-fold CV and BYY for selecting the factor number k with $k_j = 3$ fixed on the data sets of a seven-dimensional x ($d = 7$) generated from a three-dimensional y ($k = 3$) with different sample sizes for NFA.

variances, and numbers of hidden factors. For NFA, since the computational complexity grows exponentially with the number of factors, we only make comparisons on three groups of data with different sample sizes, data dimensions, and noise variances. The observations x_t, $t = 1, \ldots, n$ for BFA are generated from $x_t = Ay_t + c + e_t$ with y_t randomly generated from a Bernoulli distribution with $q_j = 0.5$ and e_t randomly generated from $\mathcal{N}(0, \sigma^2 I)$. For NFA, the observations are generated from $x_t = Ay_t + e_t$ with each $y_t^{(j)}$ randomly



(a) $n = 20$

(b) $n = 100$

Figure 3. The curves obtained by the criteria AIC, BIC, CAIC, 10-fold CV and BYY for selecting the Gaussian number k_j with $k = 3$ fixed on the data sets of a seven-dimensional x ($d = 7$) generated from a three-dimensional y ($k = 3$) with different sample sizes for NFA.

generated from a Gaussian mixture with three Gaussians and e_t randomly generated from $\mathcal{N}(0, \Sigma)$.

Experiments are repeated over 100 times to facilitate our observing on statistical behaviors. Each element of A is generated from $\mathcal{N}(0, 1)$.

For BFA, we set $k_{\min} = 1$ and $k_{\max} = 2k - 1$ where k is the true number of hidden factors. For NFA, we simplify the task by setting the number k_j for each factor $y^{(j)}$ to a same number such that what to be determined are only two

Table II. Numbers of underestimating (U), success (S), and overestimating (O) by each criterion on the data sets with different sample sizes for selecting hidden factor number k with $k_j = 3$ fixed for NFA in 50 experiments

Criteria	$n = 20$			$n = 40$			$n = 100$		
	U	S	O	U	S	O	U	S	O
AIC	4	25	21	2	32	16	1	40	9
BIC	10	38	2	8	42	0	1	49	0
CAIC	19	31	0	14	36	0	4	46	0
10-fold CV	1	27	22	1	34	15	0	42	8
BYY	3	34	13	1	41	8	0	48	2

numbers k and k_j. Also, k and k_j are determined separately, i.e., holding $k_j = 3$ fixed when determining k and holding k fixed when determining the number k_j. Usually, we set $k_{\min} = 1$ and $k_{\max} = 2k - 1$ where k is the true number of hidden factors and $k_{j\min} = 1$ and $k_{j\max} = 5$ since the true number of k_j is 3.

4.1. EFFECTS OF SAMPLE SIZES

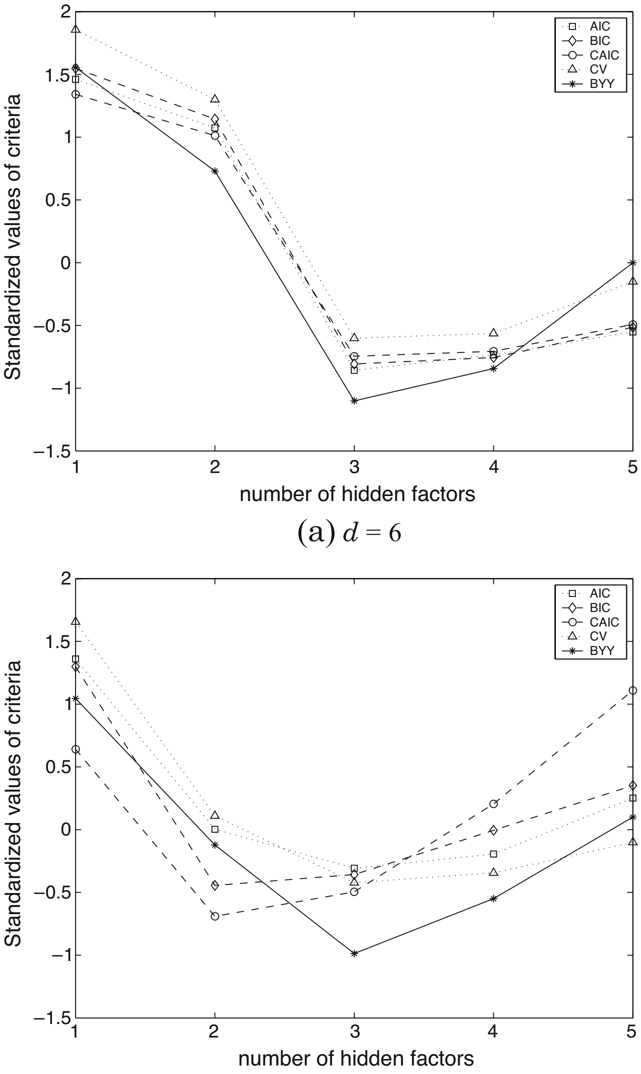
4.1.1. *For Binary Factor Analysis*

We investigate the performances of every criterion on the data sets with different sample sizes $n = 20$, $n = 40$, and $n = 100$ for implementing BFA. In this experiment, the dimension of x is $d = 9$ and the dimension of y is $k = 3$. The noise variance σ^2 is equal to 0.1. The results are shown in Figure 1. Table I illustrates the numbers of underestimating, success, and overestimating of each criterion in 100 experiments.

When the sample size is only 20, BYY and BIC select the correct number 3. CAIC selects the number 2. AIC, 10-fold CV select 4. When the sample size is 100, all the criteria lead to the correct number. Similar observations can be observed in Table I. For a small sample size, CAIC tends to underestimate the

Table III. Numbers of underestimating (U), success (S), and overestimating (O) by each criterion on the data sets with different sample sizes for selecting Gaussian number k_j with $k = 3$ fixed for NFA in 50 experiments

Criteria	$n = 20$			$n = 40$			$n = 100$		
	U	S	O	U	S	O	U	S	O
AIC	9	31	10	4	34	12	1	41	8
BIC	22	26	2	17	32	1	5	43	2
CAIC	28	22	0	21	29	0	15	35	0
10-fold CV	6	29	15	3	32	15	0	39	11
BYY	7	31	12	4	37	9	1	44	5



(a) $d = 6$

(b) $d = 25$

Figure 4. The curves obtained by the criteria AIC, BIC, CAIC, 10-fold CV and BYY on the data sets of a x with different dimensions generated from a three-dimensional y ($k = 3$) for BFA.

number while AIC, 10-fold CV tend to overestimate the number, while BYY criterion has a risk of overestimation.

4.1.2. *For Non-Gaussian Factor Analysis*

We investigate the performances of every criterion on the data sets with different sample sizes $n = 20$, $n = 40$, and $n = 100$ for implementing NFA. In this experiment, the dimension of x is $d = 7$ and the dimension of y is $k = 3$. The noise

Table IV. Numbers of underestimating (U), success (S), and overestimating (O) by each criterion on the data sets with different data dimensions for BFA in 100 experiments

Criteria	$d = 6$			$d = 15$			$d = 25$		
	U	S	O	U	S	O	U	S	O
AIC	0	89	11	0	86	14	2	83	16
BIC	0	98	2	3	96	1	30	69	1
CAIC	0	100	0	7	93	0	48	52	0
10-fold CV	0	90	10	0	86	14	0	89	11
BYY	0	99	1	1	95	4	10	89	1

covariance matrix Σ is equal to $0.1I$ (I is a 7×7 identity matrix). The results with different factor number k and fixing k_j as 3 are shown in Figure 2. The results with different k_j and fixing k as 3 are shown in Figure 3. Tables II and III illustrate the numbers of underestimating, success, and overestimating of each method for selecting k and k_j respectively in 50 experiments.

When the sample size is only 20, we can find the similar results with that of BFA. We see that BYY and BIC select the correct hidden factors number 3, CAIC selects the number 2, AIC and 10-fold CV select 4. When the sample size is 100, all the criteria lead to the correct number. Similar observations can be observed in Table II. For a small sample size, CAIC tends to underestimate the number while AIC, 10- fold CV tend to overestimate the number. BYY criterion has a little risk of overestimation while BIC has a little risk of underestimation.

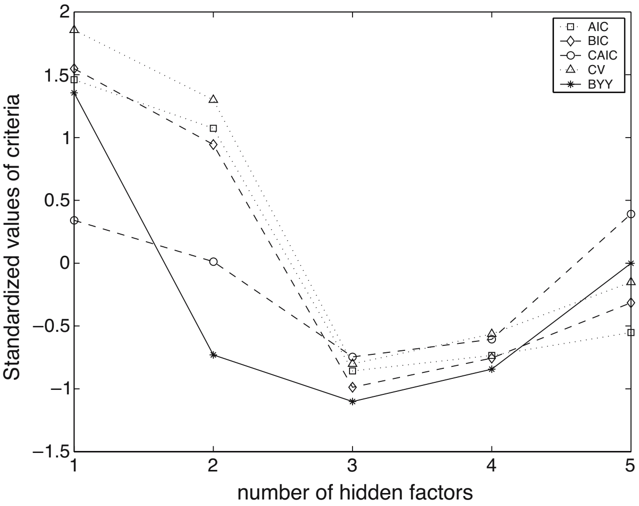
When the sample size is only 20, AIC and BYY select the correct Gaussian number 3 for each Gaussian mixture of y_j. BIC selects the number 2 and CAIC select the number 1, and 10-fold CV select the number 4. When the sample size is 100, only CAIC leads to the number 2, all the other criteria select the correct number 3. Similar observations can be observed in Table III. CAIC tends to underestimate even the sample size is large enough.

4.2. EFFECTS OF DATA DIMENSIONS

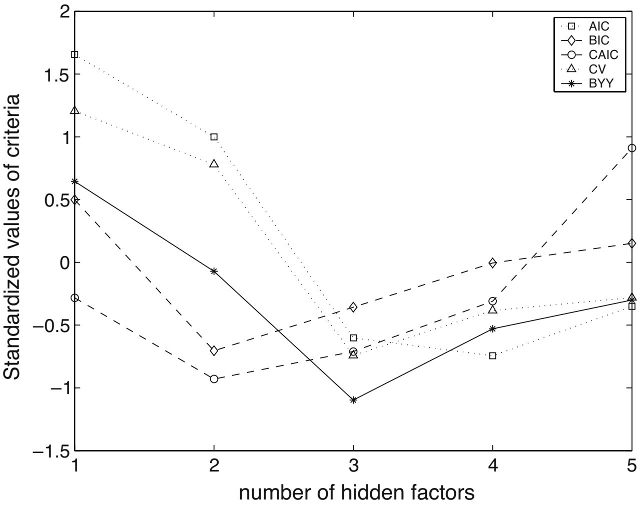
4.2.1. *For Binary Factor Analysis*

Next we investigate the effect of data dimension on each method for implementing BFA. The dimension of y is $k = 3$, the noise variance σ^2 is equal to 0.1, and the sample size is $n = 50$. The dimension of x is $d = 6$, $d = 15$, and $d = 25$. The results are shown in Figure 4. Table IV illustrates the numbers of underestimating, success, and overestimating of each method in 100 experiments.

When the dimension of x is 6, we observe that all these criteria tend to select the right number 3. However, when the dimension of x is increased to 25, BYY, 10-fold CV and AIC get the right number 3, but CAIC and BIC choose the number 2. Similar observations can be obtained in Table IV. For a high dimensional x, BYY, and 10- fold CV still have high successful rates but CAIC



(a) $d = 5$



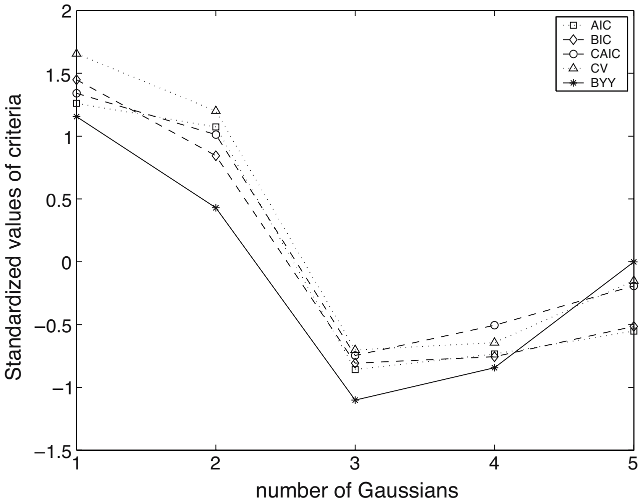
(b) $d = 20$

Figure 5. The curves obtained by the criteria AIC, BIC, CAIC, 10-fold CV and BYY for selecting the factor number k with fixing $k_j = 3$ on the data sets of a x with different dimensions generated from a three-dimensional y ($k = 3$) for NFA.

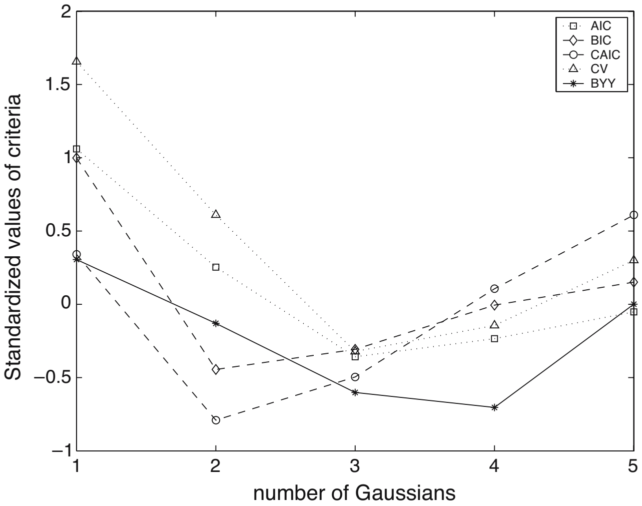
and BIC tend to underestimating the hidden factors number k. AIC has a slight risk to overestimate the hidden factors number.

4.2.2. *For Non-Gaussian Factor Analysis*

Now we investigate the effect of data dimension on each criterion for NFA. The dimension of y is $k = 3$, the noise covariance matrix Σ is equal to $0.1I$, and



(a) $d = 5$



(a) $d = 20$

Figure 6. The curves obtained by the criteria AIC, BIC, CAIC, 10-fold CV and BYY for selecting the Gaussian number k_j with fixing $k = 3$ on the data sets of a x with different dimensions generated from a three-dimensional y ($k = 3$) for NFA.

the sample size is $n = 80$. The dimension of x is $d = 5$, $d = 10$, and $d = 20$. The results with different factor number k while fixing $k_j = 3$ are shown in Figure 5. The results with different k_j while fixing $k = 3$ are shown in Figure 6. Tables V and VI illustrate the numbers of underestimating, success, and overestimating of each criterion for selecting k and k_j respectively in 50 experiments.

Table V. NFAs of underestimating (U), success (S), and overestimating (O) by each criterion on the data sets with different data dimensions for selecting hidden factor number k with $k_j = 3$ fixed for NFA in 50 experiments

Criteria	$d = 5$			$d = 10$			$d = 20$		
	U	S	O	U	S	O	U	S	O
AIC	2	42	8	0	39	11	0	36	14
BIC	3	47	0	10	40	0	13	34	3
CAIC	4	46	0	13	37	0	20	29	1
10-fold CV	0	40	10	0	40	10	0	39	11
BYY	0	47	3	1	44	5	2	40	8

When the dimension of x is 5, we observe that all these criteria tend to select the correct factor number 3. However, when the dimension of x is increased to 20, the result is similar with that of BFA except AIC. That is, BYY and 10-fold CV get the correct number 3, but CAIC and BIC tend to underestimate the factor number and AIC tend to overestimate the factor number. Similar observations can be obtained in Table V.

When the dimension of x is 20, AIC and 10-fold CV get the correct Gaussian number 3 of k_j, BIC and CAIC tend to underestimation while BYY criterion has a risk of overestimation. Similar observations can be obtained in Table VI.

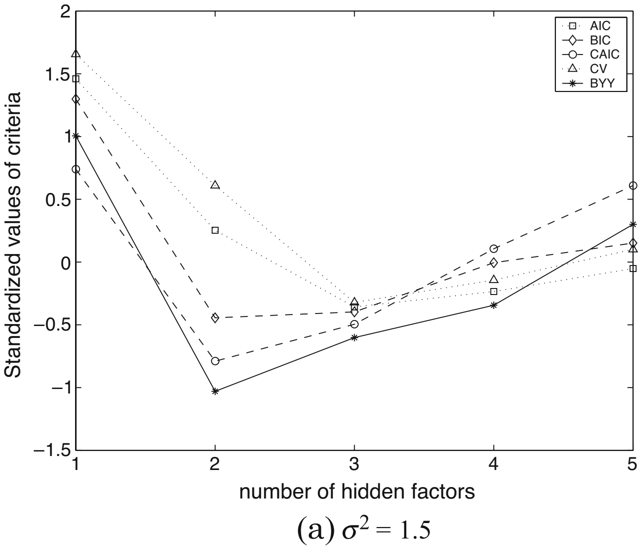
4.3. EFFECTS OF NOISE VARIANCES

4.3.1. *For Binary Factor Analysis*

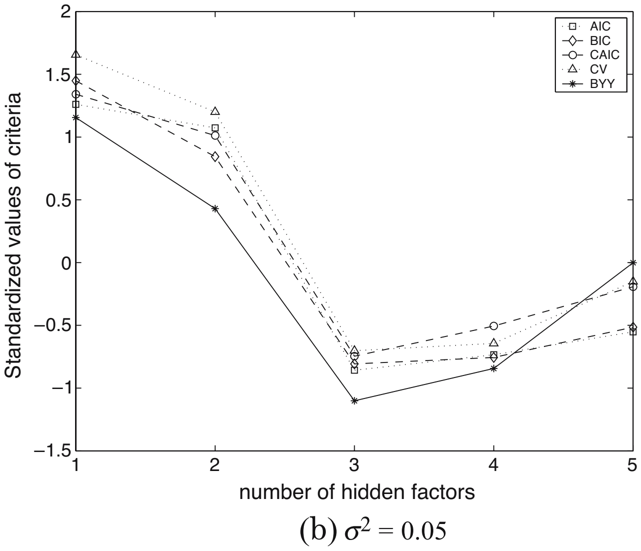
We further investigate the performance of each criterion on the data sets with different scales of noise added for implementing BFA. In this example, the dimension of x is $d = 9$, the dimension of y is $k = 3$, and the sample size is $n = 50$.

Table VI. Numbers of underestimating (U), success (S), and overestimating (O) by each criterion on the data sets with different data dimensions for selecting Gaussian number k_j with $k = 3$ fixed for NFA in 50 experiments

Criteria	$d = 5$			$d = 10$			$d = 20$		
	U	S	O	U	S	O	U	S	O
AIC	0	40	10	2	38	10	1	37	12
BIC	5	43	2	11	38	1	13	33	4
CAIC	10	40	0	14	35	1	22	24	4
10-fold CV	1	39	10	0	39	11	0	40	10
BYY	3	43	4	4	37	9	5	32	13



(a) $\sigma^2 = 1.5$



(b) $\sigma^2 = 0.05$

Figure 7. The curves obtained by the criteria AIC, BIC, CAIC, 10-fold CV and BYY on the data sets of a nine-dimensional x ($d = 9$) generated from a three-dimensional y ($k = 3$) with different noise variances for BFA.

The noise variance σ^2 is equal to 0.05, 0.5, and 1.5. The results are shown in Figure 7. Table VII illustrates the rates of underestimating, success, and overestimating of each method in 100 experiments.

When the noise variance is 1.5, only AIC and 10-fold CV select the correct number 3, BIC, CAIC and BYY select 2 factors. When the noise variance is 0.05 or 0.5, all the criteria lead to the correct number. Similar observations can be observed in Table VII. From this table we can find, for a large noise variance,

Table VII. Numbers of underestimating (U), success (S), and overestimating (O) by each criterion on the data sets with different noise variances for BFA in 100 experiments

Criteria	$\sigma^2 = 0.05$			$\sigma^2 = 0.5$			$\sigma^2 = 1.5$		
	U	S	O	U	S	O	U	S	O
AIC	0	89	11	0	82	18	6	78	16
BIC	0	100	0	2	97	1	40	58	2
CAIC	1	99	0	6	94	0	57	43	0
10-fold CV	0	86	14	0	86	14	1	81	18
BYY	0	100	0	6	94	0	39	55	6

CAIC is high likely to underestimate the number, AIC and 10-fold CV have a risk of overestimate the number.

4.3.2. *For Non-Gaussian Factor Analysis*

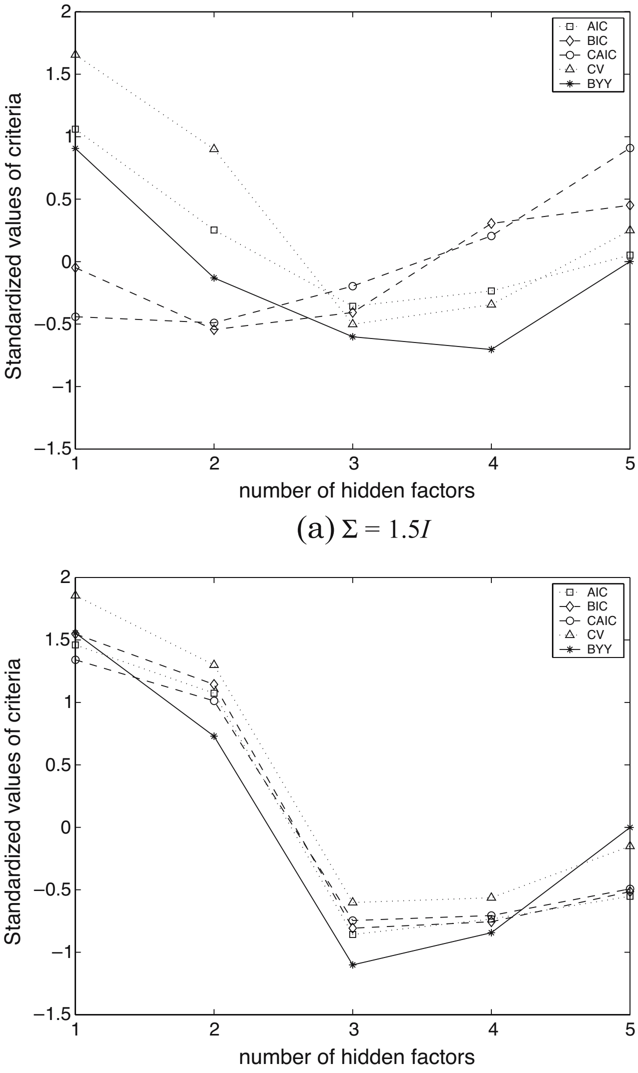
Now we investigate the performance of each criterion on the data sets with different scales of noise added for NFA. In this example, the dimension of x is $d = 7$, the dimension of y is $k = 3$, and the sample size is $n = 80$. The noise covariance matrix Σ is equal to $0.05I$, $0.5I$, and $1.5I$. Because the results with different k_j are similar to the results with number k, they are shown in the same Figure 8. Tables VIII and IX illustrate the numbers of underestimating, success, and overestimating of each method for selecting k and k_j respectively in 50 experiments.

When the noise covariance matrix is $1.5I$, AIC, BIC, CAIC, and 10-fold CV have the similar results with that of BFA. That is, AIC and 10-fold CV select the correct factor number 3, BIC, CAIC select 2 factors while BYY criterion select 4 factors. Similar observations can be observed in Table VIII. From this table we can find, for a large noise variance, CAIC is high likely to underestimate the number, AIC and 10-fold CV have a slight risk of overestimate the hidden factor number.

For the Gaussians number k_j, Table IX shows that the results are similar to the results of determining the hidden factors number k.

4.4. EFFECTS OF HIDDEN FACTOR NUMBERS

Finally, we consider the effect of hidden factor number, that is, the dimension of y on each criterion. Since the computational complexity grows exponentially with the number of factors for NFA, k cannot be larger than 5 for the practical implementation. Therefore, we only consider this comparison on BFA. In this example we set $n = 50$, $d = 15$, and $\sigma^2 = 0.1$. The dimension of y is $k = 3$, $k = 6$,



(a) Σ = 1.5*I*

(b) Σ = 0.05*I*

Figure 8. The curves obtained by the criteria AIC, BIC, CAIC, 10-fold CV and BYY for selecting the factor number k and the Gaussian number k_j on the data sets of a seven-dimensional x ($d = 7$) generated from a three-dimensional y ($k = 3$) with different noise variances for NFA.

and $k = 10$. Table X illustrates the numbers of underestimating, success, and overestimating of each method in 100 experiments.

As shown in Table X, when hidden factor number is small all criteria have good performance. When hidden factor number is large AIC, 10-fold CV, BYY criterion gets a risk of overestimating.

Table VIII. Numbers of underestimating (U), success (S), and overestimating (O) by each criterion on the data sets with different noise variances for selecting hidden factor number k with fixing $k_j = 3$ for NFA in 50 experiments

Criteria	$\Sigma = 0.05I$			$\Sigma = 0.5I$			$\Sigma = 1.5I$		
	U	S	O	U	S	O	U	S	O
AIC	0	41	9	1	41	8	1	37	12
BIC	3	47	0	4	45	1	15	28	5
CAIC	5	45	0	6	44	0	25	21	4
10-fold CV	0	40	10	0	41	9	1	39	10
BYY	0	47	3	3	43	4	10	26	14

Table IX. Numbers of underestimating (U), success (S), and overestimating (O) by each criterion on the data sets with different noise variances for selecting Gaussian number k_j with fixing $k = 3$ for NFA in 50 experiments

Criteria	$\Sigma = 0.05I$			$\Sigma = 0.5I$			$\Sigma = 1.5I$		
	U	S	O	U	S	O	U	S	O
AIC	0	40	10	1	39	10	2	38	10
BIC	6	43	1	7	43	0	20	25	5
CAIC	10	40	0	9	41	0	29	21	0
10-fold CV	2	38	10	1	39	10	0	40	10
BYY	0	45	5	2	43	5	8	26	16

Table X. Numbers of underestimating (U), success (S), and overestimating (O) by each criterion on simulation data sets with different hidden factor numbers in 100 experiments

Criteria	$k = 3$			$k = 6$			$k = 10$		
	U	S	O	U	S	O	U	S	O
AIC	0	86	14	0	85	15	0	72	28
BIC	2	98	0	4	96	0	7	93	0
CAIC	6	94	0	9	91	0	11	89	0
10-fold CV	0	85	15	0	85	15	0	81	19
BYY	2	96	4	1	93	6	0	80	20

5. Conclusion

We have made an experimental comparison on several typical model selection criteria by using them to determine the model scales for BFA and NFA. The considered criteria include four typical criteria AIC, BIC, CAIC, 10-fold CV and the criteria obtained from BYY harmony learning. From the comparison results for both BFA and NFA, we observe that BIC got a high successful rate when the data dimension is not too high, CAIC has an underestimation tendency while AIC and 10-fold CV have an overestimation tendency. In most cases, BYY criterion are superior or comparable to other methods. In consideration that the model selection by these criteria have to be made via a two stage implementation with expensive costs, while BYY harmony learning can be implemented with automated model selection during parameter learning, BYY harmony learning is more preferred since computing costs can be saved significantly.

Acknowledgements

The work was fully supported by a grant from the Research Grant Council of the Hong Kong SAR (project No: CUHK4225/04E).

References

1. Akaike, H.: A new look at statistical model identification, *IEEE Trans. Automat. Contr.* **19** (1974), 716–723.
2. Akaike, H.: Factor analysis and AIC, *Psychometrika* **52**(3) (1987), 317–332.
3. Anderson, T. W. and Rubin, H.: Statistical inference in factor analysis, *Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability*, Vol. 5, Berkeley, 1956, pp. 111–150.
4. Attias, H.: Independent factor analysis, *Neur. Comput.* **11** (1999), 803–851.
5. Bartholomew, D. J. and Knott, M.: Latent variable models and factor analysis, *Kendall's Library of Satistics*, Vol. 7, Oxford University Press, New York, 1999.
6. Barron, A. and Rissanen, J.: The minimum description length principle in coding and modeling, *IEEE Trans. Inf. Theory* **44** (1998), 2743–2760.
7. Belouchrani, A. and Cardoso, J.: Maximum likelihood source separation by the expectation-maximization technique: deterministic and stochastic implementation, *Proc. NOLTA95* (1995), 49–53.
8. Bertin, E. and Arnouts, S.: SExtractor: Software for source extraction, *Astron. Astrophys., Suppl. Ser.* **117** (1996).
9. Bourlard, H. and Kamp, Y.: Auto-association by multilayer perceptrons and sigular value decomposition, *Biol. Cybern.* **59** (1988), 291–294.
10. Bozdogan, H.: Model selection and Akaike's information criterion (AIC): the general theory and its analytical extensions, *Psychometrika* **52**(3) (1987), 345–370.
11. Cattell, R.: The scree test for the number of factors, *Multivariate Behav. Res.* **1** (1966), 245–276.
12. Cichocki, A. and Amari, S. I.: *Adaptive Blind Signal and Image Processing*, Wiley, New York, 2002.
13. Dayan, P. and Zemel, R. S.: Competition and multiple cause models, *Neural. Comput.* **7** (1995), 565–579.
14. Heinen, T.: *Latent Class and Discrete Latent Trait Models: Similarities and Differences*, Sage, Thousand Oaks, CA, 1996.
15. Hyvarinen, A.: Independent component analysis in the presence of Gaussian noise by maximizing joint likelihood, *Neurocomputing* **22** (1998), 49–67.
16. Hyvarinen, A., Karhunen, J. and Oja, E.: *Independent Component Analysis*, Wiley, New York, 2001.

17. Kaiser, H.: A second generation little jiffy, *Psychometrika* **35** (1970), 401–415.

18. Liu, Z. Y., Chiu, K. C. and Xu, L.: Investigations on non-Gaussian factor analysis, *IEEE Signal Process. Lett.* **11**(7) (2004), 597–600.

19. Moulines, E., Cardoso, J. and Gassiat, E.: Maximum likelihood for blind separation and deconvolution of noisy signals using mixture models, *Proc. ICASSP97* (1997), 3617–3620.

20. Rissanen, J.: Modeling by shortest data description, *Automatica* **14** (1978), 465–471.

21. Rubin, D. and Thayer, D.: EM algorithms for ML factor analysis, *Psychometrika* **47**(1) (1982), 69–76.

22. Saund, E.: A multiple cause mixture model for unsupervised learning, *Neural Comput.* **7** (1995), 51–71.

23. Schwarz, G.: Estimating the dimension of a model, *Ann. Stat.* **6**(2) (1978), 461–464.

24. Sclove, S. L.: Some aspects of model-selection criteria, *Proceedings of the First US/Japan Conference on the Frontiers of Statistical Modeling: An Informational Approach*, Vol. 2 Kluwer, Dordrecht, The Netherlands, 1994, pp. 37–67.

25. Stone, M.: Use of cross-validation for the choice and assessment of a prediction function, *Journal R. Stat. Soc., B* **36** (1974), 111–147.

26. Treier, S. and Jackman, S.: Beyond factor analysis: modern tools for social measurement, *Presented at the 2002 Annual Meetings of the Western Political Science Association and the Midwest Political Science Association*, 2002.

27. Xu, L.: Least mean square error reconstruction for self-organizing neural-nets, *Neural Netw.* **6** (1993), 627–648. Its early version on *Proc. IJCNN91'Singapore* (1991), 2363–2373.

28. Xu, L.: Bayesian-Kullback coupled Ying-Yang machines: Unified learnings and new results on vector quantization, *Proc. Intl. Conf. on Neural Information Processing (ICONIP95)*, Beijing, China, 1995, pp. 977–988.

29. Xu, L.: 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 K. M. Wong, et al. (eds): *Theoretical Aspects of Neural Computation: A Multidisciplinary Perspective*, Springer, 1997, pp. 43–60.

30. Xu, L.: Bayesian Kullback Ying-Yang dependence reduction theory, *Neurocomputing*, **22** (1–3) (1998), 81–112.

31. Xu, L.: Temporal BYY learning for state space approach, hidden Markov model and blind source separation, *IEEE Trans Signal Process.* **48** (2000), 2132–2144.

32. Xu, L.: BYY harmony learning, independent state space, and generalized APT financial analyses, *IEEE Trans. Neural Netw.* **12**(4) (2001), 822–849.

33. Xu, L.: BYY harmony learning, structural RPCL, and topological self-organizing on mixture models, *Neural Netw.* **15** (2002), 1125–1151.

34. Xu, L.: Independent component analysis and extensions with noise and time: A Bayesian Ying-Yang learning perspective, *Neural Inf. Process. Lett. Rev.* **1**(1) (2003), 1–52.

35. Xu, L.: BYY learning, regularized implementation, and model selection on modular networks with one hidden layer of binary units, *Neurocomputing* **51** (2003), 277–301.

36. Xu, L.: Advances on BYY harmony learning: Information theoretic perspective, generalized projection geometry, and independent factor autodetermination, *IEEE Trans. Neural Netw.* **15**(4) (2004), 885–902.

37. Xu, L., Yang, H. H. and Amari, S. I.: Signal source separation by mixtures: accumulative distribution functions or mixture of bell-shape density distribution functions. *Presentation at FRONTIER FORUM. Japan: Institute of Physical and Chemical Research*, April, 1996.