BAYESIAN YING-YANG LEARNING ON ORTHOGONAL BINARY FACTOR ANALYSIS

*Ke Sun, Lei Xu**

Abstract: Binary Factor Analysis (BFA) aims to discover latent binary structures in high dimensional data. Parameter learning in BFA faces an exponential computational complexity and a large number of local optima. The model selection to determine the latent binary dimension is therefore difficult. Traditionally, it is implemented in two separate stages with two different objectives. First, parameter learning is performed for each candidate model scale to maximise the likelihood; then the optimal scale is selected to minimise a model selection criterion. Such a two-phase implementation suffers from huge computational cost and deteriorated learning performance on large scale structures. In contrast, the Bayesian Ying-Yang (BYY) harmony learning starts from a high dimensional model and automatically deducts the dimension during learning. This paper investigates model selection on a subclass of BFA called Orthogonal Binary Factor Analysis (OBFA). The Bayesian inference of the latent binary code is analytically solved, based on which a BYY machine is constructed. The harmony measure that serves as the objective function in BYY learning is more accurately estimated by recovering a regularisation term. Experimental comparison with the two-phase implementations shows superior performance of the proposed approach.

Key words: *Binary factor analysis, model selection, Bayesian Ying-Yang learning*

Received: April 4, 2009
Revised and accepted: July 27, 2009

1. Introduction

Latent structures often exist in high dimensional observations. Depending on the characteristics of data, such a structure can be a low dimensional manifold or just a discrete set. Discovering the underlying structure with an appropriate scale is a central problem in statistical learning. In classical Factor Analysis (FA) [1], the underlying structure is assumed to be a low dimensional Gaussian distribution. A variant model called Binary Factor Analysis (BFA) adopts a vector of independent

*Ke Sun, Lei Xu
Department of Computer Science and Engineering, the Chinese University of Hong Kong, Hong Kong, P. R. China, E-mail: lxu@cse.cuhk.edu.hk

Bernoulli distributions as the latent model. As in FA, BFA also faces the difficulty of model selection, that is, to determine the number of binary factors so that the resulting model represents the regularity well but does not overfit the training data. The problem is even more difficult in BFA because of the combinatorial complexity and a large number of local optima. Research on BFA under the names of latent trait model, item response theory, latent class model or multiple cause model [1, 2, 3] is widely used in data reduction, psychological measurement, political science, etc.

This paper focuses on Orthogonal Binary Factor Analysis (OBFA) which restricts the loading matrix in BFA to be orthogonal. OBFA constructs an orthogonal co-ordinate system in a m-dimensional linear subspace of the observation space $\Re^n$ $(m < n)$, so that the representative clusters lie around $\{-1, 1\}^m$ and therefore are separated by the co-ordinate planes. For example, in a 2-level Orthogonal Experiment Design (OED) [4], OBFA can be applied to extract 2^m representative experimental inputs based on a training set. In information theory, a parallel binary channel with additive Gaussian noise and rotational transformation at the output end results in a OBFA model. In a single hidden layer free-forward stochastic neural network with sigmoid activation function, the hidden layer encodes space regions according to linear separations and thus can be initialised by OBFA. In psychological measurement, the questionnaires can be analysed with OBFA if the latent trait is binary, such as sex or a threshold based property. Moreover, analysis on OBFA is useful to construct parametric structures in BFA, which does not have analytical Bayesian inverse in general.

Given finite observations $\mathcal{X}_N = \{\boldsymbol{x}_t\}_{t=1}^N$, learning is traditionally implemented in two separate phases. The first phase estimates for each model scale m within a candidate set $\mathcal{K}$ the unknown parameters $\boldsymbol{\Theta}_m$ through maximising the log-likelihood $\log q(\mathcal{X}_N \mid \boldsymbol{\Theta}_m)$; the second phase selects the optimal scale $m^\star$ with

$$m^\star = \arg\min_{m \in \mathcal{K}} J(\hat{\boldsymbol{\Theta}}_m, m), \tag{1}$$

where $\hat{\boldsymbol{\Theta}}_m$ is the maximum likelihood solution, $J(\hat{\boldsymbol{\Theta}}_m, m)$ is a model selection criterion, such as Akaike's Information Criterion (AIC) [5], Schwarz's Bayesian Information Criterion (BIC) [6], Hannan-Quinn Information Criterion (HQC) [7], and Bozdogan's Consistent Akaike's Information Criterion (CAIC) [8]. For OBFA, these criteria are listed as follows.

$$\begin{aligned} J(\hat{\boldsymbol{\Theta}}_m, m) &= -2\log q(\mathcal{X}_N \mid \hat{\boldsymbol{\Theta}}_m) + \rho_N d_m, \\ d_m &= mn - m(m-1)/2 + m + n + 1, \\ \rho_N &= \begin{cases} 2 & \text{for AIC;} \\ \log(N) & \text{for BIC;} \\ \log\log(N) & \text{for HQC;} \\ \log(N) + 1 & \text{for CAIC,} \end{cases} \end{aligned} \tag{2}$$

where d_m is the number of free parameters in the OBFA model, $n = \dim(\boldsymbol{x})$ is the data dimension, and $m = \dim(\boldsymbol{y})$ is the latent binary dimension. For all criteria listed above, the term $-2\log q(\mathcal{X}_N \mid \hat{\boldsymbol{\Theta}}_m)$ represents the description error of the candidate model, while the term $\rho_N d_m$ represents the penalty of model scale. The model selection in Eq. (2) tries to balance the descriptive capacity and model complexity. This two phase approach suffers from huge computational cost, especially for intractable problems such as BFA. Moreover, the parameter learning performance deteriorates rapidly as m increases, which will make $J(\hat{\Theta}_m, m)$ estimated unreliably.

Proposed firstly in 1995 [9] and systematically developed in the past decade [10, 11, 12], the BYY harmony learning provides a general statistical learning framework, under which model selection is performed automatically during parameter learning. To characterise the parametric structure of a learning system, both the external observation $\boldsymbol{X}$ and its internal presentation $\boldsymbol{Y}$ are considered within two complementary Bayesian decompositions

$$q(\boldsymbol{X},\boldsymbol{R}) = q(\boldsymbol{X}\,|\,\boldsymbol{R})q(\boldsymbol{R}) \quad \text{and} \quad p(\boldsymbol{R},\boldsymbol{X}) = p(\boldsymbol{R}\,|\,\boldsymbol{X})p(\boldsymbol{X}), \tag{3}$$

where $\boldsymbol{R} = \{\boldsymbol{Y},\boldsymbol{\Theta}\}$ consists of both the inner state $\boldsymbol{Y}$ and all the unknown parameters $\boldsymbol{\Theta}$. In the BYY system, a *Ying Machine* models the internal representations via a distribution $q(\boldsymbol{Y})$ and describes the generation process with a conditional distribution $q(\boldsymbol{X}\,|\,\boldsymbol{Y})$; the complementary part, the *Yang Machine*, characterises the observations $\{\boldsymbol{x}_t\}_{t=1}^N$ as a distribution $p(\boldsymbol{X})$ and describes the inference process with a conditional distribution $p(\boldsymbol{Y}\,|\,\boldsymbol{X})$. Once these parametric substructures are fixed, the BYY harmony learning is implemented by maximising the harmony measure

$$H(p\,||\,q) = \int p(\mathbf{R}\,|\,\mathbf{X})p(\mathbf{X}) \log\left[q(\mathbf{X}\,|\,\mathbf{R})q(\mathbf{R})\right] d\mathbf{X}d\mathbf{R}, \tag{4}$$

with a large initial model scale, which is usually represented by the dimension of $\boldsymbol{Y}$ and will be deducted during learning [10].

Much work has been dedicated to FA, BFA and other related models using the BYY harmony learning [10, 11, 13]. When the other structures are fixed, several typical choices of $p(\boldsymbol{Y}\,|\,\boldsymbol{X})$, an important component in the joint distribution $p(\boldsymbol{X},\boldsymbol{R})$, result in several implementation scenarios with different model selection performances [11, 12]. The paper [13] investigates BFA with $p(\boldsymbol{Y}\,|\,\boldsymbol{X})$ chosen free of structure. A systematical comparison with several typical model selection criteria shows the BYY approach has superior performance [13]. As a follow-up work of [13], the present paper studies BFA by considering $p(\boldsymbol{Y}\,|\,\boldsymbol{X})$ in a Bayesian structure, restricted to OBFA so that analysis and computation are feasible. The harmony measure in Eq. (4) is more accurately estimated by recovering a regularisation term that is missing in [13]. According to experimental comparison with the two phase implementations, the BYY approach shows the best performance in model selection.

The rest of this paper is organised as follows. Section 2 introduces the OBFA model. Section 3 constructs a BYY machine for OBFA and presents an OBFA learning algorithm. Section 4 evaluates the proposed algorithm through experimental comparison with traditional two phase implementations for model selection. Section 5 concludes.

2. Orthogonal Binary Factor Analysis

Factor Analysis in the literature of statistics [1] is formulated as

$$x = Ay + c + e, \tag{5}$$

where $x \in \Re^n$ is the observations, $y \in \Re^m (m < n)$ represents the latent factors, $A_{n\times m}$ is a loading matrix, $c_{n\times 1}$ is a mean offset and $e_{n\times 1}$ is a zero-mean Gaussian noise that is independent of y. Therefore

$$q(x \,|\, y) = G(x \,|\, Ay + c, \Sigma), \tag{6}$$

where Σ is the covariance matrix of e, $G(\cdot \,|\, \mu, \Sigma)$ denotes a Gaussian distribution with mean μ and covariance Σ. In the classical FA, y is assumed to be Gaussian distributed with indepent dimensions. As a variation, BFA assumes $y \in \{-1, 1\}^m$ is distributed according to

$$q(y) = \prod_{i=1}^{m} \theta_i^{(1+y_i)/2}(1-\theta_i)^{(1-y_i)/2} \quad (0 < \theta_i < 1,\ i = 1, \dots, m). \tag{7}$$

Intuitively, FA models the observations with one single independent Gaussian distribution for each of a set of directions, while BFA adopts a Bernoulli distribution on two possible positions along each direction. It tries to adapt the model to real world data with complementary characteristics, such as sex, extreme joint angle, etc.

The combinatorial complexity induced by such a variation results in huge computational cost. The Maximum Likelihood (ML) learning through the Estimation Maximization (EM) in Algorithm 1 has to estimate the 2^m-point posterior distributions $q(y \,|\, x_t)$ for each training sample $x_t \in \mathcal{X}_N$ and thus requires exponential space and time. Moreover, it can only train a model given a prefixed model scale m. For model selection, EM has to be repeated several times on a candidate set of different scales, which leads to much more computation. A BYY-AUTO algorithm for BFA have avoided this ML training enumeration [13] by deducting the model scale during training. However, there is still a combinatorial problem to compute

$$y_t = \mathop{\arg\max}_{y \in \{-1,1\}^m} \log q(y \,|\, x_t) \tag{8}$$

in the BYY-AUTO learning algorithm. As $\log q(y \,|\, x)$ is quadratic in y, Eq. (8) is a binary quadratic programming (BQP) problem that falls in the category of NP-hard.

Orthogonality is a natural condition in FA where an orthonormal basis always exists within a linear subspace of $\Re^n$. In BFA the loading matrix A is not necessarily orthogonal because the set $\{-1, 1\}^m \in \Re^m$ is not closed under an orthogonal transformation. Nevertheless, orthogonal BFA still makes sense in wide applications where the linear transformation of the binary factors mainly consists of rotation, as discussed in Section 1. In this case, $A_{n\times m} = Q_{n\times m}\Lambda_{m\times m}$, where the columns of Q are orthonormal and $\Lambda = diag(\lambda_1, \lambda_2, \dots, \lambda_m), \lambda_i > 0, i = 1, \dots, m$.

Algorithm 1: EM algorithm for BFA.

```
Input  : $\mathcal{X}_N = \{\boldsymbol{x}_t\}_{t=1}^N \subset \Re^n$; a fixed binary dimension $m = \dim(\boldsymbol{y})$
Output: $\boldsymbol{\Theta} = \{\boldsymbol{\theta}, \boldsymbol{A}, \boldsymbol{c}, \boldsymbol{\Sigma}\}$
Initialise $\boldsymbol{\Theta}_0 = \{\boldsymbol{\theta}_0, \boldsymbol{A}_0, \boldsymbol{c}_0, \boldsymbol{\Sigma}_0\}$;
repeat
    (E-step)  for $\boldsymbol{x}_t \in \mathcal{X}_N, \boldsymbol{y} \in \{-1,1\}^m$ do
        $q(\boldsymbol{y} \,|\, \boldsymbol{x}_t) = q(\boldsymbol{y})q(\boldsymbol{x}_t \,|\, \boldsymbol{y}) / \sum_{\boldsymbol{y}\in\{-1,1\}^m} \left[q(\boldsymbol{y})q(\boldsymbol{x}_t \,|\, \boldsymbol{y})\right]$;
    (M-step) $\boldsymbol{\theta} = \frac{1}{N}\sum_{t=1}^{N}\sum_{\boldsymbol{y}\in\{-1,1\}^m} q(\boldsymbol{y} \,|\, \boldsymbol{x}_t)(\boldsymbol{y}+1)/2$;
        $\boldsymbol{c} = \frac{1}{N}\sum_{t=1}^{N}\sum_{\boldsymbol{y}\in\{-1,1\}^m} q(\boldsymbol{y} \,|\, \boldsymbol{x}_t)(\boldsymbol{x}_t - \boldsymbol{A}\boldsymbol{y})$;
        $\boldsymbol{A} = \left(\sum_{t=1}^{N}\sum_{\boldsymbol{y}\in\{-1,1\}^m} q(\boldsymbol{y} \,|\, \boldsymbol{x}_t)(\boldsymbol{x}_t - \boldsymbol{c})\boldsymbol{y}^T\right) \times \left(\sum_{t=1}^{N}\sum_{\boldsymbol{y}\in\{-1,1\}^m} q(\boldsymbol{y} \,|\, \boldsymbol{x}_t)\boldsymbol{y}\boldsymbol{y}^T\right)^{-1}$;
    $\boldsymbol{\Sigma} = \frac{1}{N}\sum_{t=1}^{N}\sum_{\boldsymbol{y}\in\{-1,1\}^m} q(\boldsymbol{y} \,|\, \boldsymbol{x}_t) diag\left((\boldsymbol{x}_t - \boldsymbol{A}\boldsymbol{y} - \boldsymbol{c})(\boldsymbol{x}_t - \boldsymbol{A}\boldsymbol{y} - \boldsymbol{c})^T\right)$;
until $\log q(\mathcal{X}_N \,|\, \boldsymbol{\Theta})$ converges ;
```

From Bayes's rule,

$$\log q(\boldsymbol{y} \,|\, \boldsymbol{x}) = \log q(\boldsymbol{y}) + \log q(\boldsymbol{x} \,|\, \boldsymbol{y}) - \log \left[\sum_{\boldsymbol{y}\in\{-1,1\}^m} q(\boldsymbol{y})q(\boldsymbol{x} \,|\, \boldsymbol{y})\right]. \tag{9}$$

Further from Eq. (6) and Eq. (7), $\log q(\boldsymbol{y} \,|\, \boldsymbol{x})$ is a quadratic function of $\boldsymbol{y}$. If the observation error $\boldsymbol{e}$ is spherical such that $\boldsymbol{\Sigma} = \sigma^2 \boldsymbol{I}$, where $\boldsymbol{I}$ denotes the identity matrix, the quadratic term of $\boldsymbol{y}$ in $\log q(\boldsymbol{y} \,|\, \boldsymbol{x})$ becomes

$$-\frac{1}{2}\boldsymbol{y}^T\boldsymbol{A}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{A}\boldsymbol{y} = -\frac{1}{2\sigma^2}\boldsymbol{y}^T\boldsymbol{\Lambda}^T\boldsymbol{Q}^T\boldsymbol{Q}\boldsymbol{\Lambda}\boldsymbol{y} = -\frac{1}{2}\sum_{i=1}^{m}\frac{\lambda_i^2}{\sigma^2} \tag{10}$$

and thus vanishes. Hence $\log q(\boldsymbol{y} \,|\, \boldsymbol{x})$ is linear in $\boldsymbol{y}$ and the posterior distribution $q(\boldsymbol{y} \,|\, \boldsymbol{x})$ has independent dimensions. We assume

$$q(\mathbf{y} \,|\, \mathbf{x}) = \prod_{i=1}^{m} \hat{\theta}_i(\mathbf{x})^{(1+y_i)/2}(1-\hat{\theta}_i(\mathbf{x}))^{(1-y_i)/2}. \tag{11}$$

Extracting the linear coefficients of $\boldsymbol{y}$ on both sides of Eq. (9) yields

$$\frac{1}{2}\log\frac{\hat{\theta}_i(\boldsymbol{x})}{1-\hat{\theta}_i(\boldsymbol{x})} = \frac{1}{2}\log\frac{\theta_i}{1-\theta_i} + \boldsymbol{A}_i^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{c}), \quad i = 1, 2, \ldots, m, \tag{12}$$

where $\boldsymbol{A}_i$ is the ith column of $\boldsymbol{A}$. Hence

$$\hat{\theta}_i(\mathrm{x}) = \frac{1}{1+e^{-\xi_i(\boldsymbol{x})}},\ \xi_i(\boldsymbol{x}) = \log\frac{\theta_i}{1-\theta_i} + 2\boldsymbol{A}_i^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{c}),\ i=1,2,\ldots,m. \quad (13)$$

This analysis provides evidence for using sigmoid neurons to simulate non-linearity in Independent Component Analysis (ICA) [14], also for using the BI-structure BFA learning [13] with $p(\boldsymbol{y}\,|\,\boldsymbol{x})$ constructed in a similar form of Eq. (11) and Eq. (13).

Taking benefit of the posterior independence of $\boldsymbol{y}$ in $\log q(\boldsymbol{y}\,|\,\boldsymbol{x})$, parameter learning in OBFA can be performed efficiently. In the EM algorithm, $q(\boldsymbol{y}\,|\,\boldsymbol{x})$ can be computed according to Eq. (11) and Eq. (13). Both $E_{q(\boldsymbol{y}\,|\,\boldsymbol{x})}(\boldsymbol{y})$ and $E_{q(\boldsymbol{y}\,|\,\boldsymbol{x})}(\boldsymbol{y}\boldsymbol{y}^T)$, which have to be solved by exhaustive enumeration in Algorithm 1, can be directly computed from $\hat{\boldsymbol{\theta}}(\boldsymbol{x})$. The exponential space-time complexity in EM is saved. Also, the BYY-AUTO algorithm used in [13] becomes computationally feasible because $\log q(\boldsymbol{y}\,|\,\boldsymbol{x}_t)$ in Eq. (8) deteriorates to be linear in $\boldsymbol{y}$ and the BQP can be solved via a greedy algorithm which determines one bit at one time.

3. A BYY Machine for Orthogonal BFA

Under the BYY framework, a previous work [13] on BFA assumes the inference process $p(\boldsymbol{Y}\,|\,\boldsymbol{X})$ to be free of structure. This paper assumes $p(\boldsymbol{Y}\,|\,\boldsymbol{X}) = q(\boldsymbol{Y}\,|\,\boldsymbol{X})$ is the Bayesian inverse of the generation process $\boldsymbol{Y} \to \boldsymbol{X}$, as in Eq. (11). Let

$$p(\boldsymbol{X}) = \frac{1}{N}\sum_{t=1}^{N}\delta(\boldsymbol{X}-\boldsymbol{x}_t) \quad (14)$$

be the empirical distribution, where $\delta(.)$ is the Dirac delta. It further follows from Eq. (4), Eq. (6), Eq. (7) and Eq. (11) that

$$H(p\,||\,q) = \frac{1}{N}\sum_{t=1}^{N}\tilde{H}(\boldsymbol{x}_t,\,\hat{\boldsymbol{y}}(\boldsymbol{x}_t),\,\boldsymbol{\Theta}) \quad (15)$$

and

$$\begin{aligned}\tilde{H}(\boldsymbol{x},\,\boldsymbol{y},\,\boldsymbol{\Theta}) = &\underbrace{-\frac{n}{2}\log 2\pi - \frac{1}{2}\log|\boldsymbol{\Sigma}| - \frac{1}{2}\left[\boldsymbol{A}\boldsymbol{y}+\boldsymbol{c}-\boldsymbol{x}\right]^T\boldsymbol{\Sigma}^{-1}\left[\boldsymbol{A}\boldsymbol{y}+\boldsymbol{c}-\boldsymbol{x}\right]}_{\text{❶}} \\ &\underbrace{+\sum_{i=1}^{m}\left[\frac{1+y_i}{2}\log\theta_i + \frac{1-y_i}{2}\log(1-\theta_i)\right]}_{\text{❷}}\underbrace{-\frac{1}{2}Tr\left[\boldsymbol{A}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{A}\cdot\left(\boldsymbol{I}-diag(\boldsymbol{y}\boldsymbol{y}^T)\right)\right]}_{\text{❸}},\end{aligned} \quad (16)$$

where

$$\hat{\boldsymbol{y}}(\boldsymbol{x}) = E_{q(\boldsymbol{y}\,|\,\boldsymbol{x})}(\boldsymbol{y}) = 2\hat{\boldsymbol{\theta}}(\boldsymbol{x}) - 1 \quad (17)$$

and $\boldsymbol{\Theta} = \{\boldsymbol{\theta}, \boldsymbol{A}, \boldsymbol{c}, \boldsymbol{\Sigma}\}$. Term ❶
$\log q(\boldsymbol{X}\,|\,\boldsymbol{Y})$ from the inference results $\boldsymbol{Y}$ to the observations $\boldsymbol{X}$; term ❷

the negative entropy, or simplicity, of the internal structure; term ❸ sation term resulting from the Bayesian structure of $p(\boldsymbol{Y}\,|\,\boldsymbol{X})$. With a large initial $\dim(\boldsymbol{y})$, both parameter learning and model selection are implemented by an optimisation process to maximise $H(p\,||\,q)$ in Eq. (15), during which a redundant dimension y_i may be identified when $\theta_i \to \{0,1\}$ or $||\boldsymbol{A}_i||_2 \to 0$ and discarded at a dimension deduction threshold, while parameter learning proceeds at the next lower dimensional model. Eventually, the learning algorithm will determine all the unknown parameters as well as an appropriate model scale [10, 11, 12].

Compared with the objective function used in [13], Eq. (15) is a more accurate expression of the harmony measure due to the term ❸ regularisation strength to enlarge $\boldsymbol{\Sigma}$ and reduce $||\boldsymbol{A}_i||_2$. As a result, dimension deduction in BFA learning is in the following two scenarios:

1. The internal structure $q(\boldsymbol{Y})$ in Eq. (7) deteriorates to a $(m-1)$ dimensional independent Bernoulli distribution when a bit y_i tends to be deterministic, that is, $\theta_i \to 0$ or 1;

2. The information carried by the ith bit y_i vanishes during the generation process $q(\boldsymbol{X}\,|\,\boldsymbol{Y})$ in Eq. (6), that is, $||\boldsymbol{A}_i|| \to \boldsymbol{0}$ when the random variable y_i is multiplied by zero and thus become deterministic.

Alternatively, the task is also equivalent to a neural network learnings as shown in Fig. 1. The BYY learning adjusts the weights in a three-layer feed-forward neural network to maximise $H(p\,||\,q)$ in Eq. (15), with saturated neurons discarded during learning. To compute $\partial H/\partial\zeta$, $\forall\zeta \in \{\boldsymbol{\theta}, \boldsymbol{A}, \boldsymbol{c}, \boldsymbol{\Sigma}\}$ and maximise $H(p\,||\,q)$ through a gradient based optimisation, we consider the total differential form

$$dH = \frac{1}{N}\sum_{t=1}^{N}\left\{\left[\frac{\partial\tilde{H}}{\partial\boldsymbol{y}}(\boldsymbol{x}_t,\,\hat{\boldsymbol{y}}(\boldsymbol{x}_t),\,\boldsymbol{\Theta})\right]^T d\hat{\boldsymbol{y}}(\boldsymbol{x}_t) + \left[\frac{\partial\tilde{H}}{\partial\boldsymbol{\Theta}}(\boldsymbol{x}_t,\,\hat{\boldsymbol{y}}(\boldsymbol{x}_t),\,\boldsymbol{\Theta})\right]^T d\boldsymbol{\Theta}\right\} \tag{18}$$

obtained from Eq. (15). From Eq. (13) and Eq. (17),

$$d\hat{\boldsymbol{y}}(\boldsymbol{x}_t) = \frac{1}{2}\left(\boldsymbol{\iota} - \hat{\boldsymbol{y}}(\boldsymbol{x}_t)\right) \circ \left(\boldsymbol{\iota} + \hat{\boldsymbol{y}}(\boldsymbol{x}_t)\right) \circ d\boldsymbol{\xi}(\boldsymbol{x}_t), \tag{19}$$

where $\boldsymbol{\iota}_{m\times 1} = (1, 1, \ldots, 1)^T$, "$\circ$" is the element-by-element multiplication. From Eq. (16),

$$\frac{\partial\tilde{H}}{\partial\boldsymbol{y}}(\boldsymbol{x}_t, \hat{\boldsymbol{y}}(\boldsymbol{x}_t), \boldsymbol{\Theta}) = \frac{1}{2}\log(\boldsymbol{\theta}) - \frac{1}{2}\log(\boldsymbol{\iota} - \boldsymbol{\theta}) + \boldsymbol{A}^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}_t - \boldsymbol{c}) = \frac{1}{2}\boldsymbol{\xi}(\boldsymbol{x}_t). \tag{20}$$

Plugging Eq. (20) and Eq. (19) into Eq. (18) yields

$$\frac{\partial H}{\partial\zeta} = \frac{1}{N}\sum_{t=1}^{N}\left\{\frac{1}{4}\boldsymbol{\xi}(\boldsymbol{x}_t) \circ (\boldsymbol{\iota} - \hat{\boldsymbol{y}}(\boldsymbol{x}_t)) \circ (\boldsymbol{\iota} + \hat{\boldsymbol{y}}(\boldsymbol{x}_t)) \circ \frac{\partial\boldsymbol{\xi}}{\partial\zeta}(\boldsymbol{x}_t) + \frac{\partial\tilde{H}}{\partial\zeta}(\boldsymbol{x}_t, \hat{\boldsymbol{y}}(\boldsymbol{x}_t), \boldsymbol{\Theta})\right\}, \forall\zeta \in \boldsymbol{\Theta}. \tag{21}$$

A gradient ascend learning based on Eq. (21) is given in Algorithm 2.

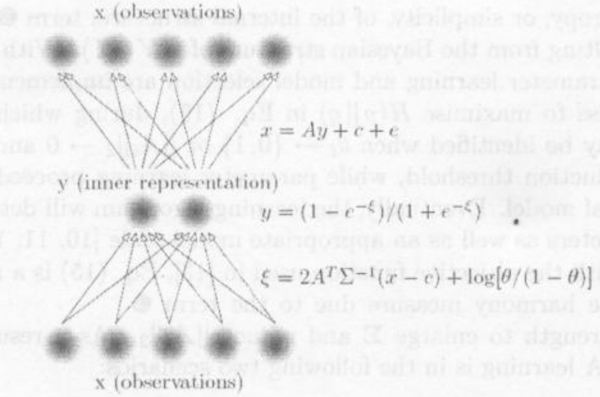

x (observations)
$x = Ay + c + e$
y (inner representation)
$y = (1 - e^{-\xi})/(1 + e^{-\xi})$
$\xi = 2A^T\Sigma^{-1}(x - c) + \log[\theta/(1-\theta)]$
x (observations)

Fig. 1 *A neural work structure for OBFA Learning.*

4. Simulation

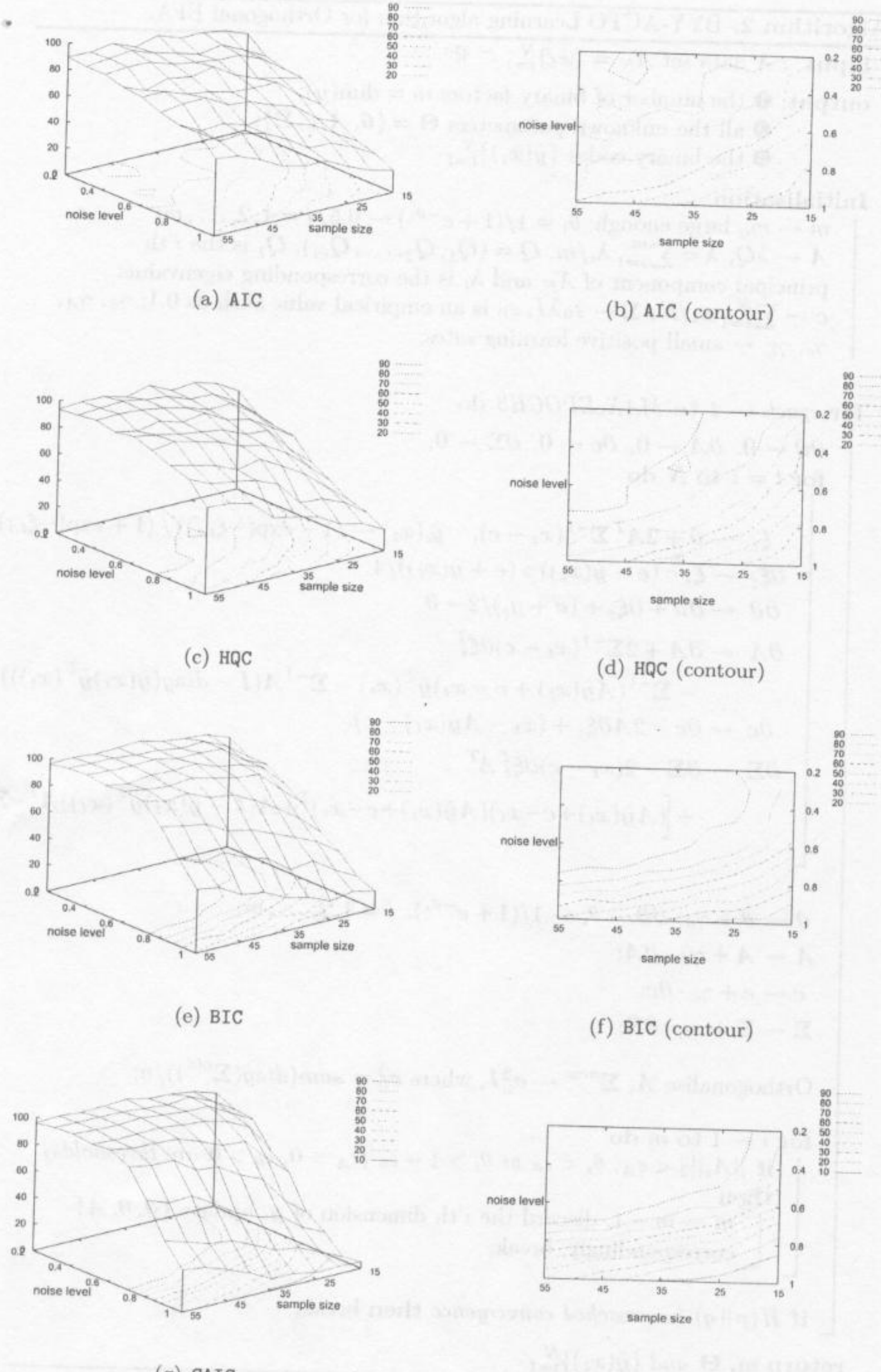
This section evaluates the proposed BFA algorithm, denoted as BYY-Bayes, through experimental comparison with the two phase model selection implementations listed in Section 1 and the free structure BYY-AUTO algorithm in pp. 151 [13] named BYY-free. To investigate the performance of these algorithms under different configurations, we fix $\dim(\boldsymbol{x}) = 8$, $\dim(\boldsymbol{y}^\star) = 3$ and vary the sample size N and noise level σ over a 9×9 grid $\{15, 20, 25, \dots, 55\} \times \{0.2, 0.3, 0.4, \dots, 1.0\}$. For each configuration $(\dim(\boldsymbol{x}), \dim(\boldsymbol{y}^\star), N, \sigma)$ in such a grid, synthetic data is generated according to Eq. (6) and Eq. (7) with $\boldsymbol{\theta} \sim Beta(5,5)^1$, $\boldsymbol{A} = \boldsymbol{Q}_{n\times m}\boldsymbol{\Lambda}_{m\times m}$, $\boldsymbol{Q}$ randomly generated with orthonormal columns, $\boldsymbol{\Lambda} = diag(\lambda_1, \lambda_2, \dots, \lambda_m)$, $\lambda_i \sim Uni(1,2)^2$, $i = 1, 2, \dots, m$ and $\boldsymbol{e} \sim G(\boldsymbol{e}\,|\,0, \sigma^2\boldsymbol{I})$. In total 200 data sets are independently generated for each grid point, and then passed to all six algorithms investigated for model selection. For the two phase approaches, the candidate set of binary dimensions is fixed as $\{1, 2, 3, 4, 5\}$. For BYY-free and BYY-Bayes, the initial binary dimension is chosen as 5.

Fig. 2(a, c, e, g, i, k) shows the model selection accuracy over the 9×9 configuration grid in the percentage $\kappa(N, \sigma)$ that $\dim(\boldsymbol{y})$ is correctly estimated. The corresponding contour graphs are drawn on the right hand side for a more accurate comparison of different methods. For example, in Fig. 2(f), the inner contour line circles a region where BIC gives a model selection accuracy greater than 90%. For all approaches investigated, model selection becomes inaccurate when N goes small or σ goes large. By comparing the area of the region $\{(N, \sigma) : \kappa(N, \sigma) > 90\%\}$, we clearly see BIC, CAIC, BYY-free and BYY-Bayes provide a better estimation than AIC and HQC on large N and small σ, especially AIC. Even when N is large, AIC does not converge to the true model scale. There is a slight performance drop of BYY-free and BYY-Bayes when σ is very small, because a proper noise level can form a natural regularisation to aid automatic dimension reduction. As N decreases and σ increases, the performance of all the two phase implementations deteriorates

[1] $Beta(\alpha, \beta)$ denotes a beta distribution with parameters (α, β)

[2] $Uni(a, b)$ denotes a continuous uniform distribution over the interval (a, b)

```
Algorithm 2: BYY-AUTO Learning algorithm for Orthogonal BFA.
input  : A data set $\mathcal{X}_N = \{x_t\}_{t=1}^N \subset \Re^n$
output: ❶ the number of binary factors $m \Rightarrow \dim(y)$;
        ❷ all the unknown parameters $\Theta = \{\theta, A, c, \Sigma\}$;
        ❸ the binary codes $\{\hat{y}(x_t)\}_{t=1}^N$

Initialisation
    $m \leftarrow m_0$ large enough: $\theta_i = 1/(1+e^{-\vartheta_i}) \leftarrow 0.5$, $i = 1, 2, \ldots, m$;
    $A \leftarrow \bar{\lambda} Q$, $\bar{\lambda} = \sum_{i=1}^m \lambda_i / m$, $Q = (Q_1, Q_2, \ldots, Q_m)$, $Q_i$ is the $i$'th
    principal component of $\mathcal{X}_N$ and $\lambda_i$ is the corresponding eigenvalue;
    $c \leftarrow \sum_{t=1}^N x_t / N$; $\Sigma \leftarrow \varepsilon_0 \bar{\lambda} I$, $\varepsilon_0$ is an empirical value such as 0.1; $\gamma_\vartheta$, $\gamma_A$,
    $\gamma_c$, $\gamma_\Sigma \leftarrow$ small positive learning rates;

for $epoch \leftarrow 1$ to $MAX\_EPOCHS$ do
    $\partial\vartheta \leftarrow 0$, $\partial A \leftarrow 0$, $\partial c \leftarrow 0$, $\partial\Sigma \leftarrow 0$;
    for $t = 1$ to $N$ do
        $\xi_t \leftarrow \vartheta + 2A^T \Sigma^{-1}(x_t - c)$,   $\hat{y}_i(x_t) \leftarrow (1 - \exp(-\xi_{t,i})) / (1 + \exp(-\xi_{t,i}))$
        $\partial\xi_t \leftarrow \xi_t \circ (e - \hat{y}(x_t)) \circ (e + \hat{y}(x_t))/4$
        $\partial\vartheta \leftarrow \partial\vartheta + \partial\xi_t + (e + y_t)/2 - \theta$
        $\partial A \leftarrow \partial A + 2\Sigma^{-1}(x_t - c)\partial\xi_t^T$
                $- \Sigma^{-1}(A\hat{y}(x_t) + c - x_t)\hat{y}^T(x_t) - \Sigma^{-1} A(I - diag(\hat{y}(x_t)\hat{y}^T(x_t)))$
        $\partial c \leftarrow \partial c - 2A\partial\xi_t + (x_t - A\hat{y}(x_t) - c)$
        $\partial\Sigma \leftarrow \partial\Sigma - 2(x_t - c)\partial\xi_t^T A^T$
                $+ \left[(A\hat{y}(x_t) + c - x_t)(A\hat{y}(x_t) + c - x_t)^T + A(I - \hat{y}(x_t)\hat{y}^T(x_t))A^T - \Sigma\right]/2$

    $\vartheta \leftarrow \vartheta + \gamma_\vartheta \cdot \partial\vartheta$,   $\theta_i \leftarrow 1/(1+e^{-\vartheta_i})$, $i = 1, 2, \ldots, m$;
    $A \leftarrow A + \gamma_A \cdot \partial A$;
    $c \leftarrow c + \gamma_c \cdot \partial c$;
    $\Sigma \leftarrow \Sigma + \gamma_\Sigma \cdot \partial\Sigma$;

    Orthogonalise $A$, $\Sigma^{new} \leftarrow \sigma_0^2 I$, where $\sigma_0^2 = sum(diag(\Sigma^{old}))/n$;

    for $i \leftarrow 1$ to $m$ do
        if $||A_i||_2 < \epsilon_A$, $\theta_i < \epsilon_\theta$ or $\theta_i > 1 - \epsilon_\theta$ ($\epsilon_A > 0$, $\epsilon_\theta > 0$ are thresholds)
        then
            $m \leftarrow m - 1$, discard the $i$'th dimension of $y$, update $\{\vartheta, \theta, A\}$
            correspondingly, break;

    if $H(p||q)$ has reached convergence then break;

return $m$, $\Theta$ and $\{\hat{y}(x_t)\}_{t=1}^N$
```



(a) AIC

(b) AIC (contour)

(c) HQC

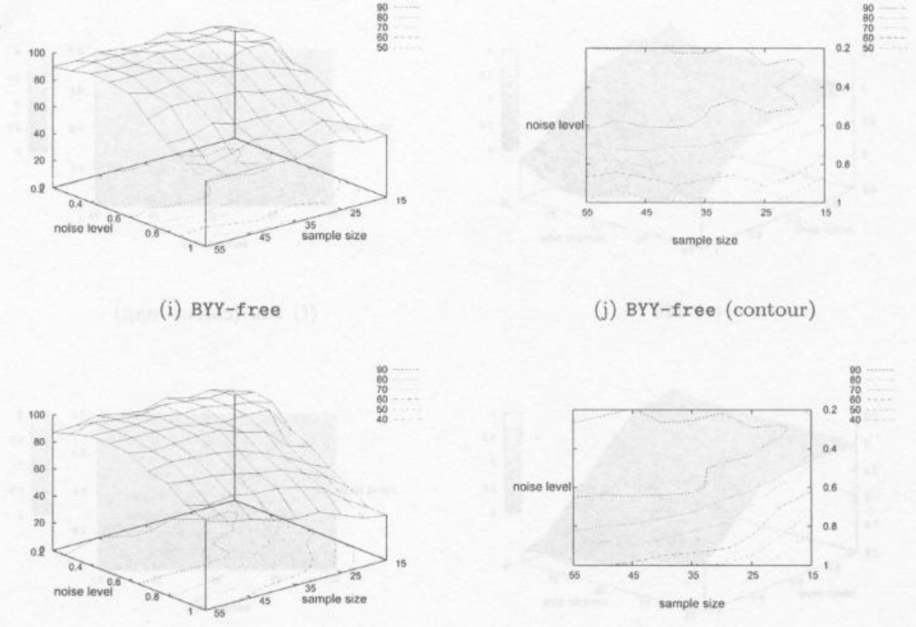
(d) HQC (contour)

(e) BIC

(f) BIC (contour)

(g) CAIC

(h) CAIC (contour)



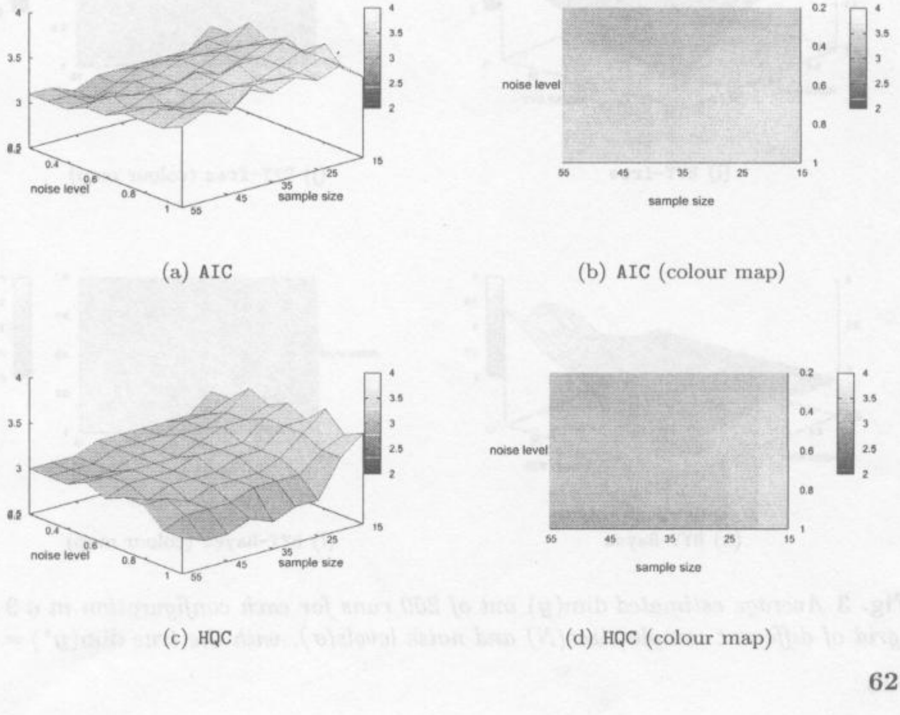
(i) BYY-free

(j) BYY-free (contour)

(k) BYY-Bayes

(l) BYY-Bayes (contour)

Fig. 2 *Percentage of correctly estimated* $\dim(\boldsymbol{y})$ *out of 200 runs for each configuration in a* 9×9 *grid of different sample sizes(N) and noise levels(*σ*).*

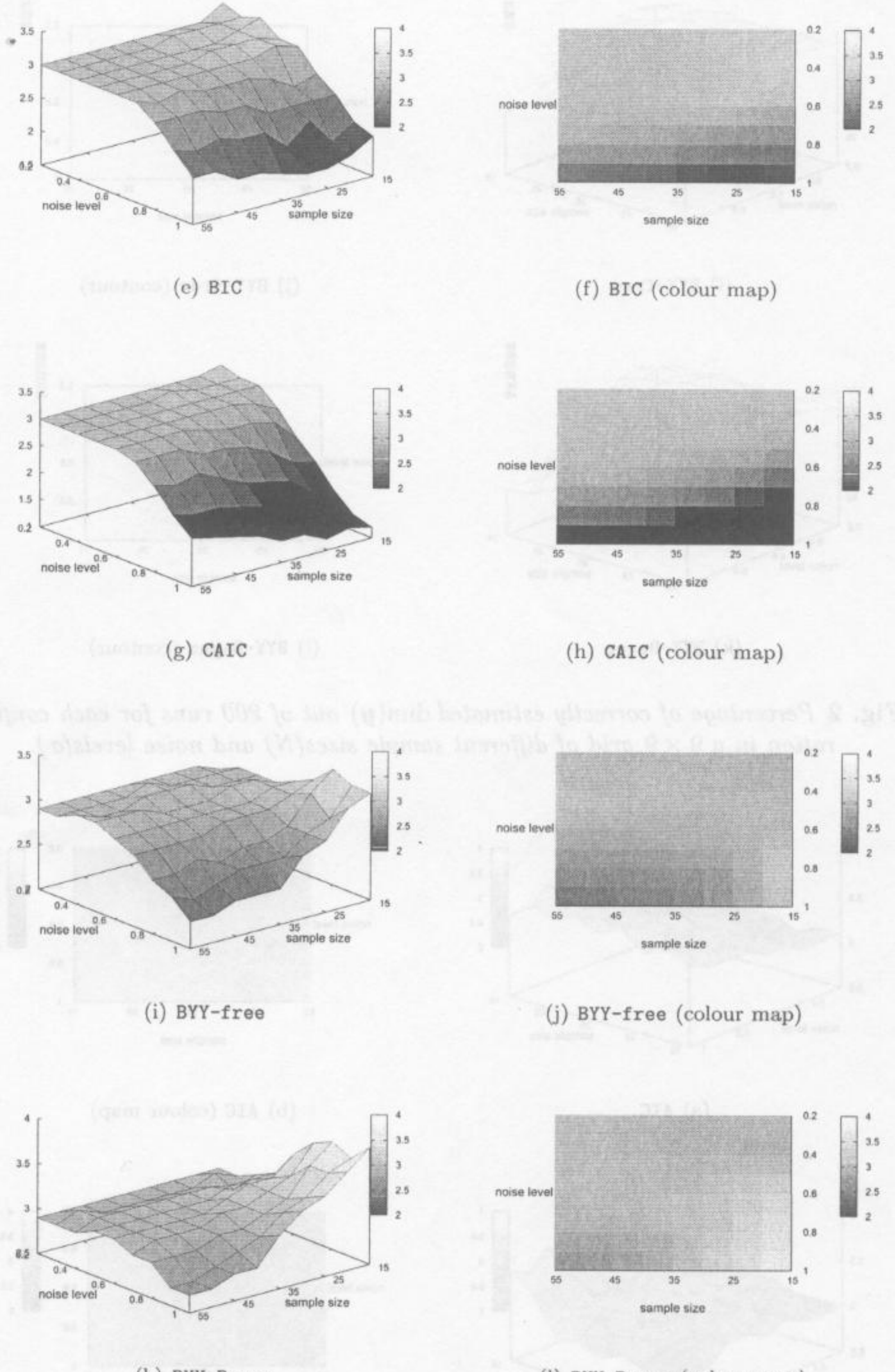


(a) AIC

(b) AIC (colour map)

(c) HQC

(d) HQC (colour map)



(e) BIC

(f) BIC (colour map)

(g) CAIC

(h) CAIC (colour map)

(i) BYY-free

(j) BYY-free (colour map)

(k) BYY-Bayes

(l) BYY-Bayes (colour map)

Fig. 3 *Average estimated* $\dim(\boldsymbol{y})$ *out of 200 runs for each configuration in a* 9×9 *grid of different sample sizes(N) and noise levels(σ), with the true* $\dim(\boldsymbol{y}^{\star}) = 3$.

to be less than 20%. `CAIC` even falls down 10% in the worst case. `BYY-free` and `BYY-Bayes` show superior performance in this area. The configuration grid is almost covered by $\{(N, \sigma) : \kappa_{BYY}(N, \sigma) > 50\%\}$, as shown in Fig. 2(j) and Fig. 2(l). Overall speaking, `BYY-free` and `BYY-Bayes` are more accurate than the two phase approaches. Between the two, `BYY-free` performs better on small sample size; `BYY-Bayes` further improves `BYY-free` on large sample size and provides a better performance in general.

Fig. 3 presents the averge estimated binary dimension $\overline{\dim(\boldsymbol{y})}$ over the same configuration grid. On the right hand side of each graph is a corresponding colour map to show the over/under-estimation tendency. When σ is small, all methods investigated give an accurate estimation in average around $\overline{\dim(\boldsymbol{y})} = 3$. As σ increases, `AIC` tends to an over-estimation; `BIC` and `CAIC` tend to an under-estimation; `HQC`, `BYY-free` and `BYY-Bayes` are relatively accurate in $\overline{\dim(\boldsymbol{y})}$. This intuitively shows `BIC` induces more penalty in the model complexity than `AIC` and `HQC`, while `CAIC` induces even more such penalty. `BYY-free` studied in [13] tends to underestimate the model scale on large N and large σ, while `BYY-Bayes` has avoided this underestimation and thus is more accurate as shown in Fig. 2.

5. Concluding Remarks

The presented work investigates model selection on orthogonal BFA. Taking advantage of the orthogonality, both the generation process from latent binary structures to external observations and its Bayesian inverse have analytical solutions, from which a BYY machine is constructed. Unlike traditional learning that separates model selection from parameter learning, the BYY learning performs model selection and parameter learning simultaneously by maximising a harmony measure, which is more accurately estimated in the Bayesian structure by recovering a regularisation term that is ommited in a previous study [13]. Two different dimension deduction thresholds in BFA are discussed. In the experiments, the proposed approach shows superior performance in model selection, especially on small sample size and large noise. Moreover, learning efficiency is improved due to the automatic dimension deduction.

References

[1] Fruchter B.: Introduction to Factor Analysis, New York: Van Nostrand, 1954.

[2] Heinen T.: Latent Class and Discrete Latent Trait Models: Similarities and Differences (Advanced Quantitative Techniques in the Social Sciences), Thousand Oaks, California: Sage, 1996.

[3] Bartholomew D. J., Knott M.: Latent Variable Models and Factor Analysis, Kendalls Library of Statistics **7**, Oxford University Press, New York, 1999.

[4] Taguchi G.: The System of Experimental Design: Engineering Methods to Optimize Quality and Minimize Costs, UNIPUB/Kraus International Publications, 1987.

[5] Akaike H.: A New Look at the Statistical Model Identification, IEEE Trans. Automatic Control, **19**, 6, 1974, pp. 716–723.

[6] Schwarz G.: Estimating the Dimension of a Model, Annals of Statistics, **6**, 2, 1978, pp. 461–464.

[7] Hannan E. J., Quinn B. G.: The determination of the order of an autoregression, Journal of the Royal Statistical Society B, **41**, 2, 1979, pp. 190-195.

[8] Bozdogan H.: Model Selection and Akaikes Information Criterion (AIC): The General Theory and its Analytical Extensions, Psychometrika, **52**, 3, 1987, pp. 345–370.

[9] Xu L.: A Unified Learning Scheme: Bayesian-Kullback Ying-Yang Machine, Advances in NIPS **8**, eds., D. S. Touretzky et al., MIT Press, 1996, pp. 444–450. A preliminary version in Proc. ICONIP 95, Beijing, 1995, pp. 977–988.

[10] Xu L.: A Trend on Regularization and Model Selection in Statistical Learning: A Bayesian Ying Yang Learning Perspective, Studies in Computational Intelligence, **63**, 2007, pp. 365–406.

[11] Xu L.: Bayesian Ying Yang learning, Scholarpedia, **2**, 3, 1809, http://www.scholarpedia.org/article/Bayesian_Ying_Yang_learning, 2007.

[12] Xu L.: Bayesian Ying Yang System, Best Harmony Learning, and Gaussian Manifold Based Family, eds., J. M. Zurada et al., Intelligence: Research Frontiers, WCCI 2008 Plenary/Invited Lectures, LNCS, **5050**, 2008, pp. 48–78.

[13] An Y. et al.: A Comparative Investigation on Model Selection in Binary Factor Analysis, Studies in Computational Intelligence, **6**, 2005, pp. 145–160.

[14] Xu L.: Independent Component Analysis and Extensions with Noise and Time: A Bayesian Ying-Yang Learning Perspective, Neural Information Processing Letters and Reviews, **1**, 1, 2003, pp. 1–52.