Available online at www.sciencedirect.com

SCIENCE DIRECT®

Neural Networks 17 (2004) 1051–1059

Neural Networks

www.elsevier.com/locate/neunet

2004 Special Issue

A comparative investigation on subspace dimension determination

Xuelei Hu*, Lei Xu

Department of Computer Science and Engineering, The Chinese University of Hong Kong, Shatin, N.T., Hong Kong, China

Received 15 December 2003; accepted 20 July 2004

Abstract

It is well-known that constrained Hebbian self-organization on multiple linear neural units leads to the same k-dimensional subspace spanned by the first k principal components. Not only the batch PCA algorithm has been widely applied in various fields since 1930s, but also a variety of adaptive algorithms have been proposed in the past two decades. However, most studies assume a known dimension k or determine it heuristically, though there exist a number of model selection criteria in the literature of statistics. Recently, criteria have also been obtained under the framework of Bayesian Ying–Yang (BYY) harmony learning. This paper further investigates the BYY criteria in comparison with existing typical criteria, including Akaike's information criterion (AIC), the consistent Akaike's information criterion (CAIC), the Bayesian inference criterion (BIC), and the cross-validation (CV) criterion. This comparative study is made via experiments not only on simulated data sets of different sample sizes, noise variances, data space dimensions, and subspace dimensions, but also on two real data sets from air pollution problem and sport track records, respectively. Experiments have shown that BIC outperforms AIC, CAIC, and CV while the BYY criteria are either comparable with or better than BIC. Therefore, BYY harmony learning is a more preferred tool for subspace dimension determination by further considering that the appropriate subspace dimension k can be automatically determined during implementing BYY harmony learning for the principal subspace while the selection of subspace dimension k by BIC, AIC, CAIC, and CV has to be made at the second stage based on a set of candidate subspaces with different dimensions which have to be obtained at the first stage of learning.

Keywords: BYY harmony learning; Subspace dimension; Principal component analysis; Model selection; Data smoothing

1. Introduction

Since 1930s, principal component analysis (PCA) has been used as a popular technique for processing, compressing and visualizing data, with wide applications in psychology, economics, education, management of personnel matters (Diamantaras & Kung, 1996; Jolliffe, 1986). In 1982, Oja (1982) showed that a constrained Hebbian rule based self-organization actually performs PCA adaptively. In 1989, Oja (1989) further showed that constrained Hebbian self-organization on multiple linear neural units leads to the same k-dimensional subspace spanned by the first k principal components. A variety of adaptive algorithms have been proposed to obtain the first k principal components or the subspace spanned by the first k principal components in the past two decades (Baldi & Hornik, 1995; Rubner & Tavan, 1989; Xu, 1994a,b; Xu & Yuille, 1995).

Most studies assume a known dimension k or determine k heuristically (Cattell, 1966; Kaiser, 1970). Actually, the selection of k can be addressed via statistical model selection by noticing that PCA is equivalent to a special case of the conventional factor analysis under the maximum likelihood principle (Anderson & Rubin, 1956) which is recently revisited under the name of the probabilistic PCA model (Roweis, 1998; Tipping & Bishop, 1999). Conventionally, statistical model selection was conducted via a two-phase style implementation that first obtain a set of candidate models under the maximum likelihood (ML) principle and then select the 'optimal' model according to a given model selection criterion. Popular examples of such criteria include the Akaike's (1974, 1987) information

* Corresponding author. Tel: +852 64080568; fax: +852 26035302.
E-mail addresses: xlhu@cse.cuhk.edu.hk (X. Hu), lxu@cse.cuhk.edu.hk (L. Xu).

0893-6080/$ - see front matter © 2004 Elsevier Ltd. All rights reserved.
doi:10.1016/j.neunet.2004.07.005

criterion (AIC), Bozdogan's (1987) consistent Akaike's information criterion (CAIC), Schwarz (1978) Bayesian inference criterion (BIC) which coincides with Rissanen's minimum description length (MDL) criterion (Barron & Rissanen, 1998; Rissanen, 1978), and cross-validation (CV) criterion (Stone, 1974).

The Bayesian Ying–Yang (BYY) learning was proposed as a unified statistical learning framework firstly in 1995 (Xu, 1995) and systematically developed in past years. Providing with a 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 (Xu, 2000, 2001a,b, 2002, 2003a, b). By applying BYY learning to the factor analysis, not only criteria have been obtained for selecting the PCA subspace dimension but also an adaptive algorithm has been developed that performs PCA with an appropriate subspace dimension k automatically determined during this adaptive learning (Xu, 2001b, 2003b). This paper further investigates the BYY criteria in comparison with the criteria of AIC, CAIC, BIC, and CV in the cases of small sample size.

This comparative study is carried out via experiments not only on simulated data sets of different sample sizes, noise variances, data space dimensions and subspace dimensions, etc. but also on two real data sets from air pollution data and sport track records, respectively. Experiments have shown that the performance of BIC is superior to AIC, CAIC, and CV. The BYY criterion that corresponds to BYY harmony empirical learning (BYY-HEC) (Xu, 2000, 2002) and the BYY criterion that corresponds to BYY harmony data smoothing learning (BYY-HDS) criterion (Xu, 2001b, 2002, 2003a) are comparable with that of BIC or even superior to BIC in some cases. From another perspective, selection of subspace dimension k by BIC, AIC, CAIC, and CV has to be made at the second stage on a set of candidate subspaces with different dimensions and these candidate subspaces have to be obtained via learning at the first stage. However, an appropriate subspace dimension k can be automatically determined during implementing BYY harmony learning. Therefore, BYY harmony learning is a more preferred tool for principal subspace dimension determination.

The rest of this paper is organized as follows. In Section 2, we introduce the PCA special case of factor analysis under the maximum likelihood learning. In Section 3, we further introduce not only the criteria BIC, AIC, CAIC, and CV, but also the criteria BYY-HEC and BYY-HDS. Moreover, a technique for effectively implementing data smoothing learning is proposed. Comparative experiments are given in Section 4 and a conclusion is made in Section 5.

2. Factor analysis, PCA and ML learning

Provided that x is a d-dimensional random vector of observable variables, e is a d-dimensional random vector of unobservable noise variables, and y is a k-dimensional random vector of unobservable latent variables. The following is the widely used factor analysis model in the literature of statistics (Anderson & Rubin, 1956; Xu, 1998)

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

where A is a $d\times k$ loading matrix, y is from $G(y|0,I)$,[1] e is also from Gaussian, and y and e are mutually independent. Particularly, when e comes from $G(e|0, \sigma^2 I)$, the maximum likelihood estimation on $p(x)$ leads to A consisting of the first k principal component vectors as columns. This case is equivalent to PCA (Anderson & Rubin, 1956; Xu, 1998). This factor analysis model is also recently re-iterated under the name of PCA probabilistic model (Roweis, 1998; Tipping & Bishop, 1999).

Given k and a set of observations $\{x_t\}_{t=1}^n$, one widely used method for estimating $\theta=\{A, \sigma^2\}$ is maximum likelihood learning. That is

$$\hat{\theta} = \arg\max_{\theta} L(\theta), \tag{2}$$

where $L(\theta)$ is the following log likelihood function (Lawley, 1940; Rubin & Thayer, 1982)

$$L(\theta) = -\frac{n}{2}\ln|AA^{\mathrm{T}} + \sigma^2 I| - \frac{n}{2}\mathrm{tr}(S(AA^{\mathrm{T}} + \sigma^2 I)^{-1}), \tag{3}$$

where

$$S = \frac{1}{n}\sum_{t=1}^{n} x_t x_t^{\mathrm{T}}, \tag{4}$$

is the sample covariance matrix, supposing that the observed variables have been centered at the sample means.

It has been shown that A estimated from Eq. (2) spans the principal subspace of the data (Anderson & Rubin, 1956; Xu, 1998). Let $\lambda_1 > \lambda_2 > \cdots > \lambda_d \geq 0$ be the eigenvalues of S, and let $l_1, l_2, \ldots, l_d$ be the corresponding eigenvectors normalized by $l_j' l_j = 1$. Let $U_k = (l_1, \ldots, l_k)$. The solution of A and σ^2 are given by

$$\hat{\sigma}^2 = \frac{1}{d-k}\sum_{i=k+1}^{d} \lambda_i, \tag{5}$$

$$\hat{A} = U_k \varDelta R, \tag{6}$$

where $\varDelta$ is the $k\times k$ diagonal matrix with the ith diagonal element being $\sqrt{\lambda_i - \hat{\sigma}^2}$ and R is an arbitrary $k\times k$ orthogonal rotation matrix.

In implementation, the above $\hat{\sigma}^2$, $\hat{A}$ can be obtained via making an eigen-decomposition of the sample covariance

[1] $G(x|\mu,\Sigma)$ denotes a multivariate normal (Gaussian) distribution with mean μ and covariance matrix Σ.

matrix S in help of numerical methods such as the QR method. However, the estimated S becomes poor for a small size of samples. Another method is the expectation–maximization (EM) algorithm for PCA (Roweis, 1998; Rubin & Thayer, 1982; Tipping & Bishop, 1999). It is still calculated in a batch way based on all samples. Several adaptive algorithms for PCA were developed in the literature of neural networks. In this paper, we simply use the QR eigen-decomposition method for a large sample size, while for a small sample size of high dimension we use the adaptive algorithm given by Eq. (82) in Section 3.3 of Xu (2002). The details are referred in Section 4.

3. Subspace dimension determination

3.1. Typical model selection criteria

Most studies assume a known dimension k or determine it heuristically (Cattell, 1966; Kaiser, 1970), especially in engineering fields such as signal and image processing, though the task can be performed via several existing statistical model selection criteria. One main reason is that these criteria have to be implemented via a two-stage style that is usually very computationally intensive. First, we need to assume a range of values of k from $k_{\min}$ to $k_{\max}$ which is assumed to contain the optimal k. At each specific k, we estimate the parameters θ under the ML learning principle. Second, we make the following selection

$$\hat{k} = \arg\min_k \{J(\hat{\theta}, k), k = k_{\min}, \ldots, k_{\max}\}, \tag{7}$$

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

Three typical model selection criteria are the Akaike's information criterion (AIC) (Akaike, 1974, 1987), its extension called Bozdogan's consistent Akaike's information criterion (CAIC) (Bozdogan, 1987), Schwarz's Bayesian inference criterion (BIC) (Schwarz, 1978) which coincides with Rissanen's minimum description length (MDL) criterion (Barron & Rissanen, 1998; Rissanen, 1978). These three model selection criteria can be summarized into the following general form (Sclove, 1994)

$$J(\hat{\theta}, k) = -2L(\hat{\theta}) + C(n)D(k), \tag{8}$$

where $L(\hat{\theta})$ is the log likelihood of Eq. (3) based on the ML estimate $\hat{\theta}$ under a given k, and $D(k)$ is the number of independent parameters in that k-components model (Akaike, 1987; Anderson & Rubin, 1956)

$$D(k) = dk + 1 - k(k-1)/2. \tag{9}$$

Moreover, $C(n)$ is a function with respect to the number of observations as follows:

- $C(n)=2$ for Akaike's information criterion (AIC) (Akaike, 1974, 1987),
- $C(n)=\ln(n)+1$ for Bozdogan's consistent Akaike's information criterion (CAIC) (Bozdogan, 1987),
- $C(n)=\ln(n)$ for Schwarz's Bayesian inference criterion (BIC) (Schwarz, 1978).

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 (Stone, 1974). For the ith partition, let D_i be the data subset used for testing and D_{-i} be the remainder of the data used for training, the cross-validated log-likelihood for a k-components model is

$$J(\hat{\theta}, k) = -\frac{1}{m}\sum_{i=1}^{m} L(\hat{\theta}(D_{-i})|D_i), \tag{10}$$

where m is the number of partitions, $\hat{\theta}(D_{-i})$ denotes the ML parameter estimates of k-components model from the ith training subset, and $L(\hat{\theta}(D_{-i})|D_i)$ is the log-likelihood evaluated on the data set D_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 Xu (1995) 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 modeling, with a number of new results. The details are referred to Xu (2000, 2001a, b, 2002, 2003a,b).

Applying BYY harmony learning to the factor analysis at its PCA special case, the following criterion is obtained for selecting the subspace dimension k (Xu, 2002)

$$\text{BYY-HEC}(\hat{\theta}, k) = \frac{d}{2}\ln\hat{\sigma}^2 + \frac{k}{2}(1 + \ln(2\pi)), \tag{11}$$

where $\hat{\theta}$ can be either a ML estimate or obtained from BYY harmony learning by an adaptive algorithm. This criterion is obtained from implementing BYY harmony learning in a situation without considering any regularization, i.e. learning from samples directly or called empirical learning. Shortly, we refer it by BYY harmony empirical learning (BYY-HEC) criterion. In the cases of a small size of sample, BYY harmony learning is implemented with certain regularization measures. One is called data smoothing regularization. The basic idea of this smoothing regularization is to learn a parametric model and a Parzen window nonparametric model with a smoothing parameter h^2 such that parameters in the parametric model and the smoothing parameter h^2 are determined together. Considering this data smoothing regularization, the above criterion is modified as follows

$$\text{BYY-HDS}(\hat{\theta}_h,k) = \frac{d}{2}\ln\hat{\sigma}_h^2 + \frac{k}{2}(1+\ln(2\pi)), \tag{12}$$

where $\hat{\theta}_h = \{\hat{A}_h, \hat{\sigma}_h^2, \hat{h}^2\}$ is obtained from BYY harmony learning with data smoothing regularization. That is

$$\hat{\theta}_h = \arg\max_{\theta_h} H(\theta_h, \hat{k}). \tag{13}$$

It should be noted that h^2 affects the determination of both θ_h and k, and both BYY-HDS $(\hat{\theta}_h,k)$ and $H(\theta_h,\hat{k})$ are obtained from maximizing the same BYY harmony function $H(\theta_h,k)$ (Xu, 2001a, 2002).

On the factor model Eq. (1), it follows from Section 2.2.1 in Xu (2002), especially Eqs. (34), (37) and (38) in Xu (2002), that the specific from of $H(\theta_h,k)$ is given as follows

$$H(\theta_h,k) = -\frac{d}{2}\ln\sigma_h^2 - \frac{1}{2n}\sum_{t=1}^{n}\frac{\|x_t - A_h y_t\|^2}{\sigma_h^2} - \frac{dh^2}{2\sigma_h^2} - \ln\sum_{t=1}^{n}p_h(x_t) - \frac{k}{2}(1+\ln(2\pi)), \tag{14}$$

under the constrain $E(y_t y_t^{\mathrm{T}}) = I$, where

$$y_t = (A_h^{\mathrm{T}}A_h + \sigma_h^2 I)^{-1}A_h^{\mathrm{T}}x_t, \tag{15}$$

and $p_h(x)$ is a Parzen window density

$$p_h(x) = \frac{1}{n}\sum_{t=1}^{n}G(x|x_t, h^2 I). \tag{16}$$

With k fixed, Eq. (13) can be implemented via an adaptive algorithm, e.g. a simplified version of either the one given by Eq. (78) in Xu (2001b) or the one given in Table 1 of Xu (2002). In this paper, since typical model selection criteria are compared based on the ML estimate via the QR eigen-decomposition method that is made in batch. We also implement Eq. (13) in batch. Similar to the procedure given in Table 1 of Xu (2002), we iterate following steps

$$\begin{aligned}
&\text{Yang step:} \quad \text{get } y_t \text{ by Eq. (15)},\\
&\text{Ying step:} \ \text{(a) from} \frac{\partial H(\theta_h,k)}{\partial\sigma_h^2} = 0, \frac{\partial H(\theta_h,k)}{\partial A_h^2} = 0, \text{update}\\
&\qquad \sigma_h^2 = \frac{1}{nd}\sum_{t=1}^{n}\|x_t - A_h y_t\|^2 + h^2,\\
&\qquad A_h = \frac{1}{n}\sum_{t=1}^{n}x_t y_t^{\mathrm{T}},\\
&\quad \text{(b) update } h^2 = \mathrm{e}^{\tilde{h}^2} \text{ by}\\
&\qquad h^{2\text{new}} = \mathrm{e}^{\tilde{h}^{2\text{new}}}, \tilde{h}^{2\text{new}} = \tilde{h}^{2\text{old}} + \eta_0\delta\tilde{h}^2,\\
&\qquad \delta\tilde{h}^2 = \frac{\partial H(\theta_h,k)}{\partial\tilde{h}^2} = \frac{1}{2}\left(d - \frac{dh^2}{\sigma_h^2} - \frac{\gamma(h^2)}{h^2 G(h^2)}\right),\\
&\qquad \gamma(h^2) = \sum_{t=1}^{n}\sum_{r=1}^{n}\exp\left(-\frac{\|x_t - x_r\|^2}{2h^2}\right)\|x_t - x_r\|^2,\\
&\qquad G(h^2) = \sum_{t=1}^{n}\sum_{r=1}^{n}\exp\left(-\frac{\|x_t - x_r\|^2}{2h^2}\right),
\end{aligned} \tag{17}$$

where η_0 is a step length constant. This iterative procedure is guaranteed to converge since it is actually the specific form of the Ying–Ying alternative procedure, see Section 3.5 in Xu (2003b).

With k enumerated as in Eq. (7) and its corresponding parameters obtained by the above Eq. (17), we can select a best value of k by BYY-HDS$(\hat{\theta}_h,k)$ in Eq. (12) that is simplified from $-H(\theta_h,k)$ by Eq. (14) by noticing $\hat{\sigma}_h^2 = \frac{1}{nd} \times \sum_{t=1}^{n}\|x_t - A_h y_t\|^2 + h^2$ and ignoring the term $\ln\sum_{t=1}^{n}p_h(x_t)$. Shortly, we refer it by BYY harmony data smoothing learning (BYY-HDS) criterion.

Besides the above criterion based selection, adaptive algorithms have also been developed from BYY harmony learning such that an appropriate subspace dimension k is automatically determined during adaptively learning (Xu, 2001a, 2003b). The subspace dimensions obtained via either this automatic determination or the above criterion have no difference. The difference is that the automatic determination saves significantly computational costs of implementing the conventional two stage style of statistical model selection. Thus, if the performances by the criteria from BYY harmony learning are comparable or even superior to typical statistical model selection criteria of AIC, CAIC, BIC, and CV in the cases of a small sample size, we certainly prefer to use BYY harmony learning as a tool for determining principal subspace.

4. Empirical comparative studies

We investigate the experimental performances of the model selection criteria AIC, CAIC, BIC, 10-fold CV, BYY-HEC, and BYY-HDS, on both synthetic data sets and two real world data sets. In implementation, estimating A and σ^2 is made either via the QR eigen-decomposition method by the Matlab function eig (Anderson, 1999) for implementing the ML learning or via Eq. (17) for implementing BYY harmony learning, with simply $h^2=0$ for BYY-HEC. In addition, to clearly illustrate the curve of each criterion in one figure we normalize the values of each curve to zero mean and unit variance.

4.1. Experiments on simulated data sets

We design four groups simulation experiments to illustrate the performance of each criterion on the data sets with different sample sizes, noise variances, data dimensions, and numbers of components. The observations x_t, $t=1,\ldots,n$ are generated from $x_t=Ay_t+e_t$ with y_t randomly generated from $G(y_t|0,I)$ and e_t randomly generated from $G(e_t|0,\sigma^2 I)$. Each simulation is repeated 100 times. Experiments are repeated over 100 times to facilitate our observing on statistical behaviors. Each elements of A is generated from $G(a_{ij}|0,1)$. Usually we set $k_{\min}=1$ and $k_{\max}=2k-1$ where k is the true number of components.

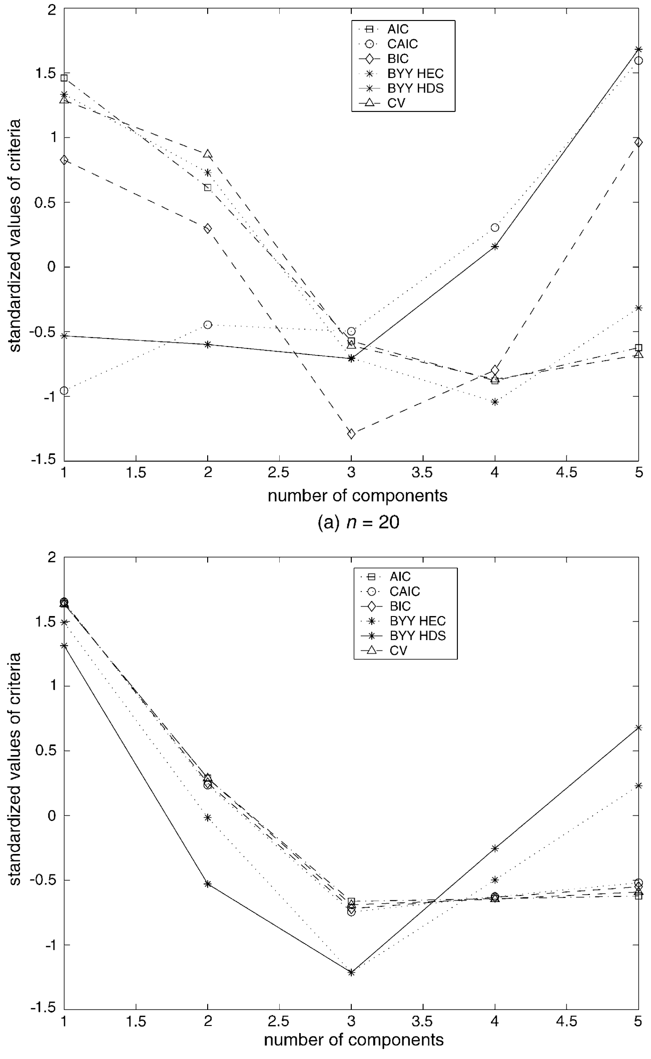
4.1.1. Effects of sample size on model selection criteria

We investigate the performances of every criterion on the data sets with different sample sizes $n=20$, 40, and 100. In this example, the dimension of x is $d=10$ and the dimension of y is $k=3$. The noise variance σ^2 is equal to $0.2\psi_k$ where ψ_k denotes the smallest positive eigenvalue of $A^{\mathrm{T}}A$. The results are shown in Fig. 1. Table 1 illustrates the rates of underestimating, success, and overestimating of each methods in 100 experiments.

When the sample size is only 20, we see that BYY-HDS and BIC select the right number 3. CAIC selects the number 1. AIC, 10-fold CV and BYY-HEC select 4. When the sample size is 100, all the criteria lead to the right number. We also observe that the value of BYY-HDS is similar with the value of BYY-HEC when the sample size is large. Similar observations can be observed in Table 1. For a small size, CAIC tends to underestimate the number while AIC, 10-fold CV, and BYY-HEC tend to overestimate the number. Again, BYY-HDS illustrates the highest successful rate.

4.1.2. Effects of noise variance on model selection criteria

We further investigate the performance of each criterion on the data sets with different scale of noise added. In this example, the dimension of x is $d=10$, the dimension of y is $k=3$, and the sample size is $n=50$. The noise variance σ^2 is equal to $0.5\psi_k\approx 1.64$, $0.25\psi_k\approx 0.82$, and $0.125\psi_k\approx 0.41$. The results are shown in Fig. 2. Table 2 illustrates the rates of underestimating, success, and overestimating of each methods in 100 experiments.



(a) $n=20$

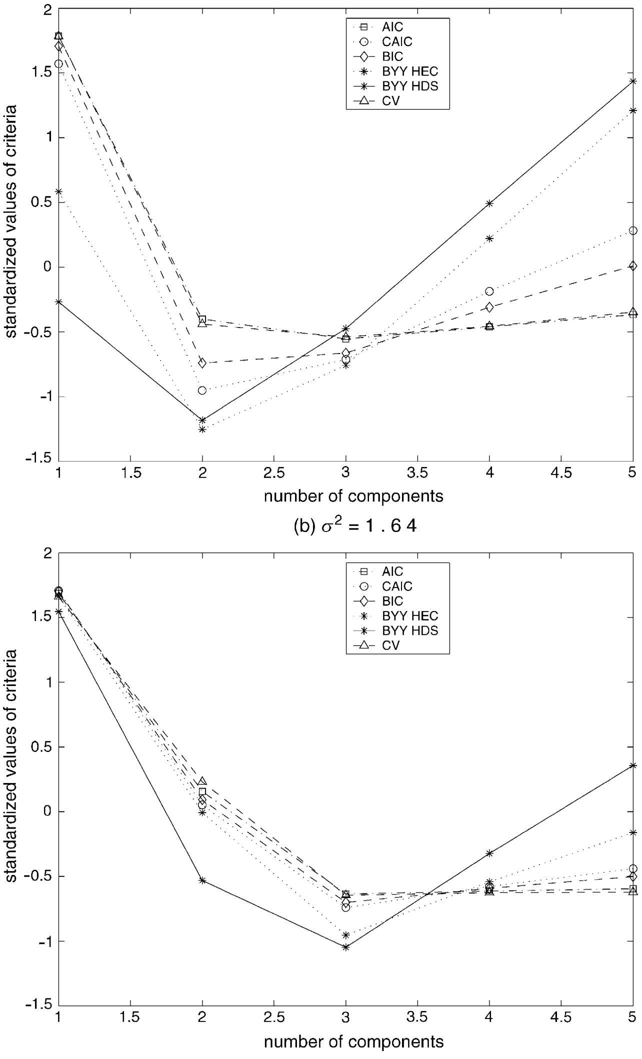
(b) $n=100$

Fig. 1. The curves obtained by the criteria AIC, CAIC, BIC, 10-fold CV, BYY-HEC and BYY-HDS on the data sets of a 10-dimensional x ($d=10$) generated from a 3-dimensional y ($k=3$) with different sample sizes. (a) $n=20$ and (b) $n=100$.

When the noise variance is 1.64, we see that only AIC and 10-fold CV select the right number 3, CAIC, BIC, BYY-HEC and BYY-HDS select two components. When the noise variance is 0.41, all the criteria lead to the right number. Similar observations can be observed in Table 2. Thus, for a large noise variance, CAIC, BIC, BYY-HEC and

Table 1
Rates of underestimating (U), success (S), and overestimating (O) by each criteria on simulation data sets with different sample sizes in 100 experiments

Criteria	$n=20$			$n=40$			$n=100$		
	U	S	O	U	S	O	U	S	O
AIC	2	68	30	0	81	19	0	85	15
CAIC	26	73	1	2	98	0	0	100	0
BIC	10	84	6	1	99	0	0	100	0
BYY-HEC	6	74	20	0	98	2	0	100	0
BYY-HDS	11	86	3	1	99	0	0	100	0
10-Fold CV	3	71	26	0	87	13	0	92	8



(b) σ^2 = 1 . 6 4

(b) σ^2 = 0 .41

Fig. 2. The curves obtained by the criteria AIC, CAIC, BIC, 10-fold CV, BYY-HEC and BYY-HDS on the data sets of a 10-dimensional x ($d=10$) generated from a 3-dimensional y ($k=3$) with different noise variances. (a) $\sigma^2=1.64$ and (b) $\sigma^2=0.41$.

BYY-HDS are high likely to underestimate the number while AIC and 10-fold CV have a slight risk of overestimating the number. For a small noise variance, CAIC, BIC, BYY-HEC and BYY-HDS have high successful rates while AIC and 10-fold CV still have a slight risk of overestimating the number.

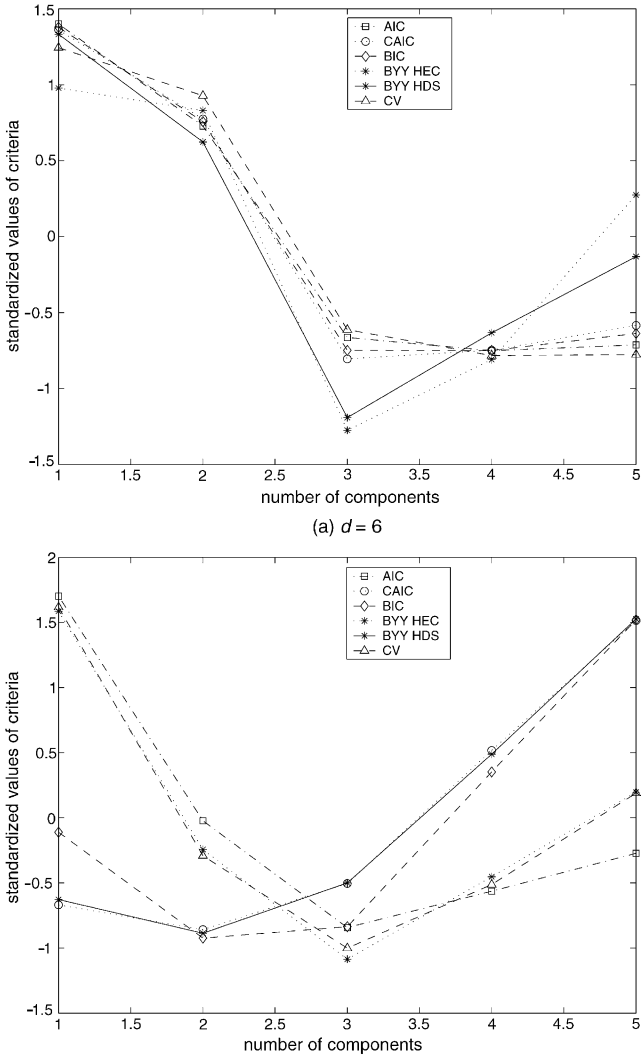
Table 2
Rates of underestimating (U), success (S), and overestimating (O) by each criteria on simulation data sets with different noise variances in 100 experiments

Criteria	$\sigma^2=1.64$			$\sigma^2=0.82$			$\sigma^2=0.41$		
	U	S	O	U	S	O	U	S	O
AIC	3	77	20	0	82	18	0	84	16
CAIC	54	46	0	1	99	0	0	100	0
BIC	49	51	0	1	98	1	0	99	1
BYY-HEC	66	34	0	7	93	0	0	100	0
BYY-HDS	79	21	0	11	89	0	0	100	0
10-Fold CV	3	78	19	0	87	13	0	88	12

4.1.3. *Effects of data dimension on model selection criteria*

Next, we investigate the effect of data dimension on each criteria. The dimension of y is $k=3$, the noise variance σ^2 is equal to $0.2\psi_k$, and the sample size in $n=50$. The dimension of x is $d=6$, 12, and 30. The results are shown in Fig. 3. Table 3 illustrates the rates of underestimating, success, and overestimating of each methods in 100 experiments.

When the dimension of x is 6, we observe that CAIC, BIC, BYY-HEC and BYY-HDS select the right number 3, while AIC and 10-fold CV select the number 4. When the dimension of x is 30, BYY-HEC, 10-fold CV and AIC get the right number 3, but CAIC, BIC and BYY-HDS choose the number 2. Similar observations can be obtained in Table 3. For a low dimensional x, CAIC, BIC, and



(a) $d=6$

(b) $d=30$

Fig. 3. The curves obtained by the criteria AIC, CAIC, BIC, 10-fold CV, BYY-HEC and BYY-HDS on the data sets of a x with different dimensions generated from a 3-dimensional y ($k=3$). (a) $d=6$ and (b) $d=30$.

Table 3
Rates of underestimating (U), success (S), and overestimating (O) by each criteria on simulation data sets with different data dimensions in 100 experiments

Criteria	$d=6$			$d=12$			$d=30$		
	U	S	O	U	S	O	U	S	O
AIC	0	87	13	0	86	14	1	89	10
CAIC	0	100	0	2	98	0	65	35	0
BIC	0	99	1	1	99	0	30	70	0
BYY-HEC	0	100	0	0	100	0	2	96	2
BYY-HDS	2	98	0	1	99	0	28	72	0
10-Fold CV	0	80	20	0	85	15	0	93	7

Table 4
Rates of underestimating (U), success (S), and overestimating (O) by each criteria on simulation data sets with different subspace dimensions in 100 experiments

Criteria	$k=2$			$k=5$			$k=10$		
	U	S	O	U	S	O	U	S	O
AIC	0	90	10	0	86	14	0	62	38
CAIC	10	90	0	15	85	0	4	96	0
BIC	2	98	0	4	96	0	1	99	0
BYY-HEC	1	99	0	1	99	0	0	83	17
BYY-HDS	3	97	0	13	87	0	31	69	0
10-Fold CV	0	92	8	0	96	4	0	95	5

BYY-HEC get high successful rates. For a high dimensional x BYY-HEC still has high successful rates, but CAIC, BIC and BYY-HDS tend to underestimating the dimension of subspace, while AIC and 10-fold CV always get a slight risk of overestimating.

4.1.4. *Effect of subspace dimension on model selection criteria*

Finally, we consider the effect of subspace dimension, that is, the dimension of y on each criterion. In this example, we set $n=50$, $d=20$, and $\sigma^2=0.2\psi_k$. The dimension of y is $k=2$, 5, and 10. Table 4 illustrates the rates of underestimating, success, and overestimating of each methods in 100 experiments.

As shown in Table 4, when subspace dimension is small all criteria have good performance. When subspace dimension is large BYY-HDS gets a risk of underestimating, while AIC gets a risk of overestimating.

4.2. *Experiments on real world data*

We further apply these model selection criteria on two real world data sets. For real data sets the true subspace dimension is unknown, and thus we just show some differences between these model selection criteria.

4.2.1. *Air pollution data*

The air pollution date from Table 1.3 in Johnson and Wichern (1998) consists of 42 measurements ($n=42$) on seven air-pollution variables ($d=7$) such as wind, solar radiation and so on recorded at 12:00 noon in the Los Angeles area on different days. We want to see how many invisible causes affect the air quality. We implement each criterion on the original data set. The results are shown in Fig. 4. The subspace dimension determined by BYY-HDS is three, by CAIC is five and by other criteria is six. According to some research on air pollution problem, it seems that five or less components maybe more reasonable.

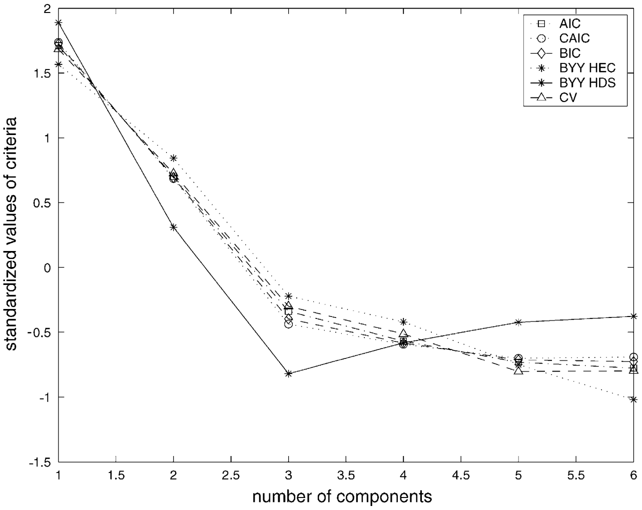


Fig. 4. The curves obtained by the criteria AIC, CAIC, BIC, 10-fold CV, BYY-HEC and BYY-HDS on the air pollution data set.

4.2.2. *Track records data*

This data consists of the national track records for man in eight items ($d=8$) from 100 m to marathon of 55 countries ($n=55$). It is come from Table 8.6 in Johnson and Wichern (1998). The task is to find how many principal components should be used to present the data. The variables are given in different measures and the scales of each viable are quite different, e.g. the time for 100 m is quiet different with the time of Marathon. So we do the experiments after

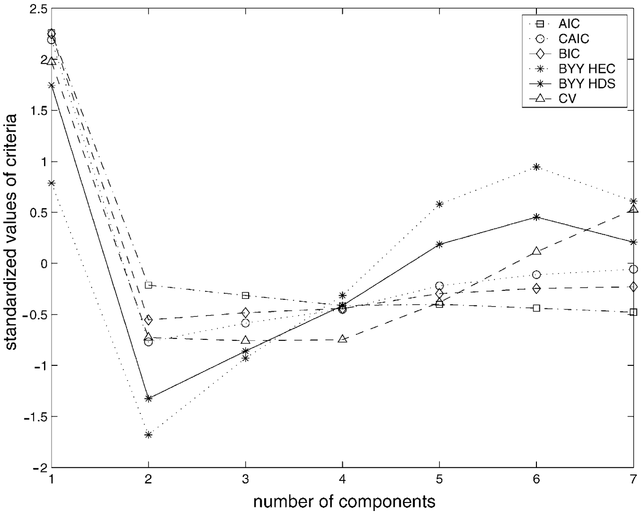


Fig. 5. The curves obtained by the criteria AIC, CAIC, BIC, 10-fold CV, BYY-HEC and BYY-HDS on the data set of track records for man.

Table 5
CPU time results on the simulation data sets with $n=20$, $d=10$, and $k=3$ by using the algorithm Eq. (17) in 100 experiments

Method	CPU time (in seconds)
The algorithm with $h^2=0$	0.01
The algorithm with $h^2>0$	0.07
AIC, CAIC, BIC, and BYY-HEC	0.27
BYY-HDS	0.82
10-Fold CV	3.12

normalizing the data. The results are shown in Fig. 5. All the methods select two components except for AIC which selects seven components and 10-fold CV which selects three components.

4.3. Discussion on computational cost

All the experiments were carried out using MATLAB R12.1 v.6.1 on a p4 1.4 GHz 512 KB RAM PC. We illustrate the computational results in Table 5 for the first example described in Section 4.1.1 with sample size $n=20$ by using the batch algorithm Eq. (17). It should be noted that all values given in Table 5 are the average of 100 experiments.

Subspace dimension determination by AIC, CAIC, BIC, BYY-HEC, and BYY-HDS takes similar CPU time. BYY-HDS has a little more computational cost for learning the smoothing parameter h^2. The m-fold cross-validation method requires the highest computational cost because for each candidate model the parameters have to be estimated m times. The computational costs of all these criteria are much more than that of using the algorithm for a single model because all of them require to obtain a whole set of candidate subspaces. Since experiments have shown that the two BYY criteria are superior or comparable to typical statistical model selection criteria, corresponding BYY harmony learning based techniques are considered more favorable because they have automatic model selection ability such that they are less computationally intensive.

5. Conclusion

We have made an experimental comparison on several typical model selection criteria by using them to determine the dimension of principal subspace. The considered criteria include four typical model selection criteria AIC, CAIC, BIC, and 10-fold CV, and two model selection criteria obtained from BYY harmony learning, namely BYY-HEC and BYY-HDS. We observe that BYY-HDS demonstrated the best performance when sample size is small. BYY-HEC is superior to other methods when data dimension is large. Both BYY-HDS and BYY-HEC have high successful rate in most cases. BIC also 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. The cross-validation method requires a highest computing cost. In addition, we provide a batch algorithm to learn A_h, σ_k^2 and the smoothing parameter h^2 based on BYY harmony learning with data smoothing regularization on PCA.

Acknowledgements

The work described in this paper was fully supported by a grant from the Research Grant Council of the Hong Kong SAR (project no: CUHK 4184/03E). The first author would like to express thanks to Dr Zhiyong Liu and Dr Kai-Chun Chiu for their useful helps.

References

Akaike, H. (1974). A new look at statistical model identification. *IEEE Transactions on Automatic Control*, *19*, 716–723.

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

Anderson, E. Z. B. (1999). *LAPACK user's guide* (3rd ed.). Philadelphia, PA: SIAM.

Anderson, T. W., & Rubin, H. (1956). Statistical inference in factor analysis. In: *Proceedings of the third Berkeley symposium on mathematical statistics and probability* (Vol. 5) (pp. 111–150). Berkeley.

Baldi, P., & Hornik, K. (1995). Learning in linear neural networks: a survey. *IEEE Transactions on Neural Networks*, *6*, 837–858.

Barron, A., & Rissanen, J. (1998). The minimum description length principle in coding and modeling. *IEEE Transactions on Information Theory*, *44*, 2743–2760.

Bozdogan, H. (1987). Model selection and Akaike's information criterion (AIC): The general theory and its analytical extensions. *Psychometrika*, *52*(3), 345–370.

Cattell, R. (1966). The screen test for the number of factors. *Multivariate Behavioral Research*, *1*, 245–276.

Diamantaras, D., & Kung, S. (1996). *Principal component neural networks: Theory and applications*. New York: Wiley.

Johnson, R., & Wichern, D. (1998). *Applied multivariate statistical analysis* (4th ed.). New York: Prentice Hall.

Jolliffe, I. (1986). *Principal component analysis*. New York: Springer.

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

Lawley, D.N. (1940). The estimation of factor loadings by the method of maximum likelihood. In: *Proceedings of the Royal Society of Edinburgh* (Vol. 60) (pp. 64–82).

Oja, E. (1982). A simplified neuron model as a principal component analyzer. *Journal of Mathematical Biology*, *15*, 267–273.

Oja, E. (1989). Neural networks, principal components, and subspaces. *International Journal of Neural Systems*, *1*, 61–68.

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

Roweis, S. (1998). EM algorithms for PCA and SPCA. In M. I. Jordan, M. J. Kearns, & S. A. Solla, *Advances in Neural information processing systems* (Vol. 10) (pp. 626–632). Cambridge, MA: The MIT Press.

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

Rubner, J., & Tavan, P. (1989). A self-organizing network for principal-component analysis. *Europhysics Letters*, *10*, 693–698.

Schwarz, G. (1978). Estimating the dimension of a model. *The Annals of Statistics*, *6*(2), 461–464.

Sclove, S. L. (1994). Some aspects of model-selection criteria. In H. Bozdogan, *Proceedings of the first US/Japan conference on the frontiers of statistical modeling: An informational approach* (Vol. 2) (pp. 37–67). Dordrecht, The Netherlands: Kluwer Academic Publishers.

Stone, M. (1974). Use of cross-validation for the choice and assessment of a prediction function. *Journal of the Royal Statistical Society B*, *36*, 111–147.

Tipping, M. E., & Bishop, C. M. (1999). Mixtures of probabilistic principal component analysers. *Neural Computation*, *11*(2), 443–482.

Xu, L. (1994a). Beyond PCA learning: From linear to nonlinear and from global representation to local representation. In: *Proceedings of the international conference on neural information processing (ICONIP94)* (Vol. 2) (pp. 943–949). Seoul, Korea.

Xu, L. (1994b). Theories for unsupervised learning: PCA and its nonlinear extensions. In: *Proceedings of IEEE ICNN94* (Vol. II) (pp. 1252–1257) Orlando, Florida.

Xu, L. (1995). Bayesian–Kullback coupled Ying–Yang machines: Unified learings and new results on vector quantization. In: *Proceedings of the international conference on neural information processing (ICONIP95)* (pp. 977–988). Beijing, China.

Xu, L. (1998). Bayesian Ying–Yang learning theory for data dimension reduction and determination. *Journal of Cumputational Intelligence in Finance*, *6*(5), 6–18.

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

Xu, L. (2001a). Best harmony, unified RPCL and automated model selection for unsupervised and supervised on gaussian mixtures, three-layer nets and me-rbf-svm models. *International Journal of Neural Systems*, *11*(1), 43–69.

Xu, L. (2001b). BYY harmony learning, independent state space, and generalized APT financial analyses. *IEEE Transactions on Neural Netwroks*, *12*(4), 822–849.

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

Xu, L. (2003a). Data smoothing regularization, multi-sets-learning, and problem solving strategies. *Neural Networks*, *16*(5/6), 817–825.

Xu, L. (2003b). Independent component analysis and extensions with noise and time: A Bayesian Ying–Yang learning perspective. *Neural Information Processing Letters and Reviews*, *1*(1), 1–52.

Xu, L., & Yuille, A. (1995). Robust principal component analysis by self-organizing rules based on statistical physics approach. *IEEE Transactions on Neural Networks*, *6*(1), 131–143.