Pattern Recognition Letters 33 (2012) 1117–1126

Contents lists available at SciVerse ScienceDirect

Pattern Recognition Letters

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

A theoretical investigation of several model selection criteria for dimensionality reduction

Shikui Tu, Lei Xu *

Department of Computer Science and Engineering, The Chinese University of Hong Kong, Hong Kong, PR China

ARTICLE INFO

Article history:
Received 11 March 2011
Available online 1 February 2012
Communicated by G. Moser

Keywords:
Factor analysis
PCA
Dimensionality reduction
Model selection criteria

ABSTRACT

Based on the problem of determining the hidden dimensionality (or the number of latent factors) of Factor Analysis (FA) model, this paper provides a theoretic comparison on several classical model selection criteria, including Akaike's Information Criterion (AIC), Bozdogan's Consistent Akaike's Information Criterion (CAIC), Hannan–Quinn information criterion (HQC), Schwarz's Bayesian Information Criterion (BIC). We focus on building up a partial order of the relative underestimation tendency. The order is shown to be AIC, HQC, BIC, and CAIC, indicating the underestimation probabilities from small to large. This order indicates an order of model selection performances to great extent, because underestimations usually take the major proportion of wrong selections when the sample size and the population signal-to-noise ratio (SNR, defined as the ratio of the smallest variance of the hidden dimensions to the variance of noise) decrease. Synthetic experiments by varying the values of the SNR and the training sample size N verify the theoretical results.

1. Introduction

Dimensionality reduction plays an important role in feature selection and extraction for pattern recognition problems (Tubbs et al., 1982; Wang and Paliwal, 2003). Especially for those high-dimensional pattern classification tasks, dimensionality reduction based methods can well improve the classification accuracy (Kim, 2011), or improve classifiers' computational efficiency (Villegas and Paredes, 2011). Factor Analysis (FA) (Anderson and Rubin, 1956) is a widely-used linear technique of dimensionality reduction (Jolliffe, 2002), by modeling the observed multidimensional variable as a low dimensional Gaussian latent variable (or factor) through a linear transform by a factor loading matrix, plus a Gaussian noise vector. Revisited in (Tipping and Bishop, 1999), the maximum likelihood solution of FA with an isotropic noise covariance matrix extracts principal components of the observed data as given by Principal Component Analysis (PCA) (Jolliffe, 2002). In this paper, the number of latent factors and the number of principal components are all referred to as the hidden dimensionality, whose value is usually unknown but important in many areas and applications. This paper focuses on this model selection problem of determining the hidden dimensionality.

To tackle this model selection problem, a traditional approach is a two-stage implementation, in which parameter learning is repeated on a set of candidate model scales among which one is selected by a model selection criterion. Existing criteria include Akaike's Information Criterion (AIC) (Akaike, 1974), Bozdogan's Consistent Akaike's Information Criterion (CAIC) (Bozdogan, 1987), Hannan–Quinn information criterion (HQC) (Hannan and Quinn, 1979), Schwarz's Bayesian Information Criterion (BIC) (Schwarz, 1978) (which coincides with Rissanen's Minimum Description Length (MDL) Rissanen, 1978). These classical criteria attempt to select a model with small generalization error, by trading off between the likelihood-based goodness of fit and model complexity subject to noise and uncertainty in a finite number of observations.

It is important to examine the relative strength and weakness of various model selection criteria. One way (e.g., in Hu and Xu (2004), Chen et al. (2008), and Tu and Xu (2011)) is to empirically examine their model selection performances by varying the structure of the population eigenvalues, training sample size N, etc. The other way is to formally analyze the probability of accurate estimation. Initialized from Wax and Kailath (1985) and followed by e.g., (Zhang et al., 1989; Xu and Kaveh, 1995; Liavas and Regalia, 2001; Fishler and Poor, 2005; Nadakuditi and Edelman, 2008), AIC and MDL were introduced to determine the number of source signals (or the latent factors) with efforts on approximating the underestimation (or overestimation) probability and asymptotical consistency under an infinite N. BIC/MDL was found to be consistent while AIC tended to overestimation as $N \to +\infty$. Moreover, in Nadler (2010), a more accurate expression for the performance of MDL was derived and the overestimation probability of AIC was analyzed from a random matrix theory viewpoint (Johnstone,

* Corresponding author.
E-mail addresses: sktu@cse.cuhk.edu.hk (S. Tu), lxu@cse.cuhk.edu.hk (L. Xu).

0167-8655/$ - see front matter © 2012 Elsevier B.V. All rights reserved.
doi:10.1016/j.patrec.2012.01.010

2001), in the joint limit N, $n \to \infty$ with $n/N \to c$, where n is the dimensionality of observations, and $c > 0$ is a constant.

Following the above track, this paper aims at a comparative investigation on model selection behavior of AIC, HQC, BIC/MDL and CAIC. We focus on building up a partial order of the relative underestimation tendency of these criteria, i.e., AIC, HQC, BIC/MDL and CAIC from weak to strong. The paper (Tu and Xu, 2009) only preliminarily reported the theoretical results without details due to the space limit. In this paper, we further provide a systematic support with more details and insights for the main claims in (Tu and Xu, 2009). Moreover, we also conduct an extensive experimental justification, which coincides with this partial order.

The rest of this paper is organized as follows. Section 2 formulates the problem of determining the hidden dimensionality of a probabilistic model FA, and also introduces a two-stage implementation with classical model selection criteria. Section 3 analyzes these criteria in terms of the relative underestimation tendency. Experimental results in Section 4 verify the relative order. Finally, concluding remarks are provided in Section 5. Proofs are left in Appendix A.

2. Problem formulation and classical model selection criteria

Factor Analysis (FA) (Anderson and Rubin, 1956; Tipping and Bishop, 1999) assumes an n-dimensional observation $\mathbf{x}$ to be distributed as follows:

$$\begin{cases} \mathbf{x} = \mathbf{A}\mathbf{y} + \boldsymbol{\mu} + \mathbf{e}, \quad \boldsymbol{\Theta}_m = \{\mathbf{A}, \boldsymbol{\mu}, \boldsymbol{\Sigma}_y, \boldsymbol{\Sigma}_e\}; \\ p(\mathbf{x}|\mathbf{y}) = G(\mathbf{x}|\mathbf{A}\mathbf{y} + \boldsymbol{\mu}, \boldsymbol{\Sigma}_e), \quad p(\mathbf{y}) = G(\mathbf{y}|\mathbf{0}, \boldsymbol{\Sigma}_y), \\ p(\mathbf{x}|\boldsymbol{\Theta}_m) = \int p(\mathbf{x}|\mathbf{y})p(\mathbf{y})d\mathbf{y} = G(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma}_x), \\ \boldsymbol{\Sigma}_y = \mathbf{I}_m (m \times m \text{ identity matrix}), \quad \boldsymbol{\Sigma}_x = \mathbf{A}\boldsymbol{\Sigma}_y\mathbf{A}^T + \boldsymbol{\Sigma}_e, \end{cases} \tag{1}$$

where $\mathbf{y}$ is an $m \times 1$ hidden factor vector, $\mathbf{A}$ is an $n \times m$ factor loading matrix with full column rank, the noise covariance matrix $\boldsymbol{\Sigma}_e$ is diagonal, and $G(\bullet|\boldsymbol{\mu}, \boldsymbol{\Sigma})$ denotes a Gaussian distribution with the mean vector $\boldsymbol{\mu}$ and the covariance matrix $\boldsymbol{\Sigma}$. We set FA in its special case by $\boldsymbol{\mu} = \mathbf{0}$, $\boldsymbol{\Sigma}_e = \sigma_e^2\mathbf{I}_n$, which is equivalent to PCA (Anderson and Rubin, 1956; Tipping and Bishop, 1999) under the maximum likelihood principle. Then, the population covariance matrix of the observations is

$$\boldsymbol{\Sigma}_x = \mathbf{A}\boldsymbol{\Sigma}_y\mathbf{A}^T + \sigma_e^2\mathbf{I}_n. \tag{2}$$

The problem of determining the hidden dimensionality of $\mathbf{y}$ is to estimate the rank of $\mathbf{A}\boldsymbol{\Sigma}_y\mathbf{A}^T$ based on an i.i.d sample set $\mathscr{X}_N = \{\mathbf{x}_t\}_{t=1}^N$.

The learning task consists of estimating the parameters $\boldsymbol{\Theta}_m$ and selecting the hidden dimensionality m, traditionally tackled by the following two-stage procedure:

- **Stage I:** Compute $\widehat{\boldsymbol{\Theta}}_m = \widehat{\boldsymbol{\Theta}}(\mathscr{X}_N, m)$ for each candidate $m \in \mathscr{M}$ with a given candidate set $\mathscr{M}$. Normally, $\widehat{\boldsymbol{\Theta}}_m$ is given by the Maximum Likelihood (ML) estimator $\widehat{\boldsymbol{\Theta}}_m^{ML} = \arg\max_{\Theta_m} \ln p(\mathscr{X}_N|\boldsymbol{\Theta}_m) = \arg\min_{\Theta_m} \mathscr{E}_L(\mathscr{X}_N|\boldsymbol{\Theta}_m)$, where $\mathscr{E}_L(\mathscr{X}_N|\boldsymbol{\Theta}_m) = -\frac{2}{N} \ln p(\mathscr{X}_N|\boldsymbol{\Theta}_m)$ is denoted as **NLL** (negative log-likelihood).
- **Stage II:** Estimate $\hat{m} = \arg\min_m \mathscr{E}_{Cri}$, where $\mathscr{E}_{Cri}$ is a model selection criterion (Cri), e.g.,

$$\mathscr{E}_{Cri}(\mathscr{X}_N, \widehat{\boldsymbol{\Theta}}_m) = \mathscr{E}_L(\mathscr{X}_N, \widehat{\boldsymbol{\Theta}}_m) + \frac{\rho_N d_m}{N}, \tag{3}$$

$$d_m = nm + 1 - \frac{m(m-1)}{2}, \text{ number of free parameters of FA,}$$

$$\rho_N = \begin{cases} 2; & \text{for } \mathbf{AIC}(\text{Akaike, 1974}) \\ \ln N; & \text{for } \mathbf{BIC}/\mathbf{MDL}(\text{Schwarz, 1978; Rissanen, 1978}) \\ \ln N + 1; & \text{for } \mathbf{CAIC}(\text{Bozdogan, 1987}) \\ 2\ln(\ln N); & \text{for } \mathbf{HQC}(\text{Hannan and Quinn, 1979}) \end{cases}, \tag{4}$$

The criterion by Eq. (3) aims to trade off between the log likelihood and the model/sample complexity. An appropriate balance allows to detect the true hidden dimensionality, if any, subject to noise and small sample size.

3. A theoretic underestimation partial order

3.1. Events of estimating the hidden dimensionality

Denote the events of *underestimation* (U), *overestimation* (O) and *successful-selection* (S) in order as:

$$A_1: \quad \hat{m} < m^*; \qquad B_1: \quad \hat{m} > m^*; \qquad C_1: \quad \hat{m} = m^* \tag{5}$$

where $\hat{m} = \hat{m}(\mathscr{X}_N)$ is an estimator of the underlying true hidden dimensionality m^* under a model selection criterion, i.e.,

$$\hat{m}(\mathscr{X}_N) = \arg\min_m \mathscr{E}_{cri}(\mathscr{X}_N, m). \tag{6}$$

For the model selection behavior of a criterion, it suffices to compute all the probabilities $P(A_1)$, $P(B_1)$ or $P(C_1)$, which are difficult for a finite N. One way is to estimate them empirically by the rates of three categories U/S/O (i.e., underestimation, successful-selection and overestimation, respectively); another way is to estimate them under certain assumptions and approximations.

Since A_1 and B_1 are disjoint, they contribute the event of wrong selections as follows:

$$P(\overline{C}_1) = P(\overline{\overline{A}_1 \cap \overline{B}_1}) = P(A_1 \cup B_1) = P(A_1) + P(B_1). \tag{7}$$

which motivates us to study $P(C_1)$ via analyzing $P(A_1)$ and $P(B_1)$ respectively. Since computing $P(A_1)$ or $P(B_1)$ is still hard, we turn to an investigation on the relative underestimation or overestimation tendency which is also an insightful investigation.

3.2. The structural property of the criterion function

Assuming a zero mean, the sample covariance matrix is a sufficient statistic for the hidden dimensionality estimation problem. Many criteria are based on the eigenvalues of the sample covariance matrix. The difficulty is to distinguish between the small yet significant sample eigenvalues due to weak hidden factors, and large yet insignificant sample eigenvalues due to noise.

As in Eq. (6), $\hat{m}$ is determined by a discrete optimization, in which a continuous optimization for parameter learning is nested as in Fig. 1(a). To locate the minima, it is reasonable to study the backward difference function, i.e.,

$$\nabla_m \mathscr{E}_{Cri} = \mathscr{E}_{Cri}(\mathscr{X}_N, m) - \mathscr{E}_{Cri}(\mathscr{X}_N, m-1), \tag{8}$$

whose sign actually determines the local preference over two consecutive models $\{m-1, m\}$ as in Fig. 1(b)(d). The difference function Eq. (8) can be formulated as a function of a statistic $\bar{\gamma}(\mathscr{X}_N)$ which is a subset or a function of the sample eigenvalues. The statistic $\bar{\gamma}(\mathscr{X}_N)$ is a preprocessed input to each criterion. For those criteria in Eq. (3), $\bar{\gamma}(\mathscr{X}_N)$ can be the ratio of a sample eigenvalue to the mean of the rest smaller sample eigenvalues. Through studying the sign change of the difference function with respect to this key statistic, we are able to provide a partial order of the relative underestimation tendency (U-tendency) of Eq. (3) in Fig. 2.

We use Eq. (1) in FA's special case of PCA by letting $\boldsymbol{\mu} = \mathbf{0}$, $\boldsymbol{\Sigma}_e = \sigma_e^2\mathbf{I}_n$. Let $s_1 \geqslant \ldots \geqslant s_n$ be the sample eigenvalues, then the Maximum Likelihood (ML) parameter estimate $\widehat{\boldsymbol{\Theta}}_m^{ML}$ of FA by Eq. (1) is given to be (Anderson and Rubin, 1956; Wax and Kailath, 1985; Bishop, 1999):

$$\begin{cases} \widehat{\mathbf{A}}_{n\times m}^{ML} = \mathbf{U}_{n\times m}(\mathbf{D}_m - \hat{\sigma}_e^2)^{\frac{1}{2}}\mathbf{R}^T, \\ \hat{\sigma}_e^{2,ML} = \frac{1}{n-m}\sum_{i=m+1}^{n} s_i, \end{cases} \tag{9}$$

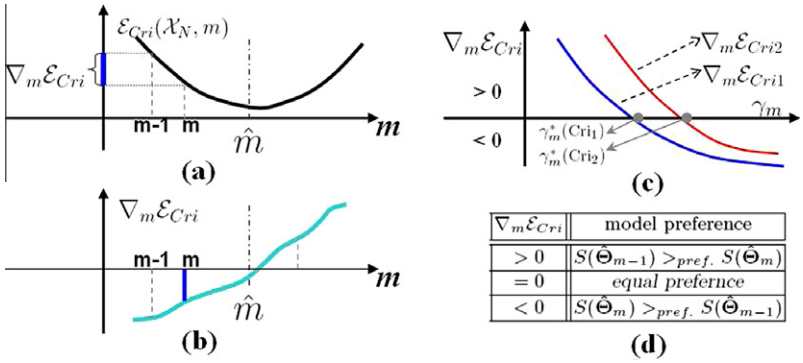


Fig. 1. For a given $\mathcal{X}_N$, graphs of $\mathcal{E}_{Cri}$ and $\nabla_m \mathcal{E}_{Cri}$ w.r.t. m are sketched in (**a**) and (**b**), while for two criteria, Cri_1 and Cri_2, the graphs of $\nabla_m \mathcal{E}_{Cri_1}, \nabla_m \mathcal{E}_{Cri_2}$ w.r.t. γ_m given m are sketched in (**c**), as well as its corresponding local preference defined in (**d**), where $S(\boldsymbol{\Theta}_m, m)$ represents a family of statistical models $p(\mathbf{x}|\boldsymbol{\Theta}_m)$ for FA. We call $\gamma_m^*(Cri_1)$ or $\gamma_m^*(Cri_2)$ as **indicator**.

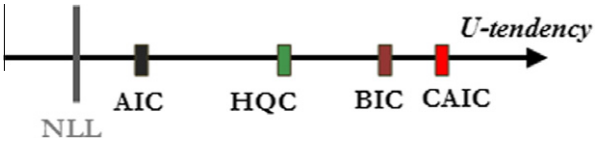


Fig. 2. The theoretical relative underestimation tendency (U-tendency) of Eq. (3) from weak to strong.

where the i-th column of $\mathbf{U}_{n\times m}$ is the eigenvector of the sample covariance matrix corresponding to s_i, the elements of $\mathbf{D}_m$ are all zero except diagonals to be $s_1,\ldots,s_m$ or shortly $\mathbf{D}_m = diag[s_1,\ldots,s_m]$, and $\mathbf{R}$ is an arbitrary rotation matrix. It follows from Eq. (9) that

$$\mathcal{E}_L\left(\mathcal{X}_N, \hat{\boldsymbol{\Theta}}_m^{ML}\right) = (n-m)\ln\sum_{i=m+1}^{n} s_i - (n-m)\ln(n-m) - \sum_{i=m+1}^{n}\ln s_i. \tag{10}$$

As a result, the difference function of Eq. (3) is

$$\nabla_m \mathcal{E}_{cri} = \nabla_m \mathcal{E}_L(\mathcal{X}_N, m) + \alpha(m, N), \tag{11}$$

where

$$\nabla_m \mathcal{E}_L(\mathcal{X}_N, m) = -(n-m+1)\ln\left(1 + \frac{\gamma_m - 1}{n-m+1}\right) + \ln\gamma_m, \tag{12}$$

$$\alpha(m,N) = \rho_N \frac{n-m+1}{N},\quad \mathcal{A}_m^n = \sum_{i=m}^{n}\frac{s_i}{n-m+1},\quad \gamma_m = \frac{s_m}{\mathcal{A}_{m+1}^n} \geqslant 1, \tag{13}$$

Fixing m at the underlying true hidden dimensionality m^*, then s_{m^*} and $\mathcal{A}_{m^*}^n$ are ML estimators of m^*-th eigenvalue of the population covariance matrix $\mathbf{A}\boldsymbol{\Sigma}_y\mathbf{A}^T + \sigma_e^2\mathbf{I}_n$ and noise variance σ_e^2 respectively, and thus γ_{m^*} is the ML estimator of the underlying ratio $\gamma_o = \frac{\lambda_{m^*}}{\sigma_e^2} + 1$ which we call signal–noise ratio (SNR)[1] in this paper, where λ_{m^*} is the m^*-th largest eigenvalue of $\mathbf{A}\boldsymbol{\Sigma}_y\mathbf{A}^T$. Thus, we call γ_{m^*} as *empirical* signal–noise ratio (eSNR). In this context, the $\bar{\gamma}(\mathcal{X}_N)$ becomes a scalar γ_m for Eq. (3). Generally, $\bar{\gamma}(\mathcal{X}_N)$ can be a multivariate vector, e.g., a two-variable vector for the criterion DNLL (difference of NLL), which is left in Appendix A.

The difference function Eq. (11) is contributed by two terms on the right hand side. When $m \leqslant m^*$, the dominant term is the former one which is the fitting error caused by the insufficient scale of structure. When $m > m^*$, the latter term plays a main role, for the cost of increasing model structure. The model selection performance of a criterion depends on how a criterion balances the two terms around the critical point m^*. Concerning about the mathematical properties of $\nabla_m \mathcal{E}_{Cri}$, the following lemmas holds (with proof in Appendix A):

Table 1
For each of the three scenarios in the experiments, there are $|V_N| \times |V_{\gamma_o}|$ configurations, where $V_N = \{17,25,50,75,100,200,400,800\}$, $V_{\gamma_o} = \{1.2, 1.5, 2, 2.5, 3, 3.5, 4, 8, 16\}$, and $\boldsymbol{U}$ is randomly generated and normalized to be $\boldsymbol{U}^T\boldsymbol{U} = \boldsymbol{I}_m$, and $\lambda_i \sim [1,10]$ means λ_i is uniformly drawn from the interval [1,10] so that the signal eigenvalues can vary.

scenario (τ)	*settings of each scenario* $\forall N \in V_N,\ \forall\gamma_o \in V_{\gamma_o}$
I: $n = 15$, $m^* = 5$, $\lambda_i = 1$, $\forall i$	$\{(I, N, \gamma_o)\}$
II: $n = 30$, $m^* = 10$, $\lambda_i = 1$, $\forall i$	$\{(II, N, \gamma_o)\}$
III: $n = 15$, $m^* = 5$, $\lambda_i \sim [1,10]$, $\forall i$	$\{(III, N, \gamma_o)\}$
IV: $n = 15$, $m^* = 5$, $\lambda_i = 1$, $\forall i$, Uniform noise	$\{(IV, N, \gamma_o)\}$

Lemma 1. *Considering γ_m as an independent variable γ for $\nabla_m \mathcal{E}_{cri}$ when m fixed, $\nabla_m \mathcal{E}_{cri}(\gamma)$ is a monotone decreasing function of $\gamma \in [1, +\infty)$ upper bounded by $\alpha(m,N)$.*

Lemma 2. *Let γ^* be the root of $\nabla_m \mathcal{E}_{cri}(\gamma) = 0$ or equivalently $\nabla_m \mathcal{E}_L(\gamma) = -\alpha$, then*

1. *Given $\alpha > 0$, γ^* is unique for $\gamma > 1$ and bounded in $(\gamma_{low}, \gamma_{up})$, where $\gamma_{low} = (k+1)C_0 - k$, and $\gamma_{up} = \gamma_{low} + \sqrt{2(k+1)C_0(C_0-1)}$, and $C_0 = \exp\{\frac{\alpha}{k}\} = \exp\left\{\frac{(k+1)\rho}{kN}\right\}$, $k = n - m$.*
2. *If $\rho(Cri_1) > \rho(Cri_2) > 0$ for criteria Cri_1 and Cri_2, then we have $\gamma^*(Cri_1) > \gamma^*(Cri_2) > 1$.*

Remarks 1. γ is an independent variable corresponding to γ_m for a fixed m, and ρ or $\rho(Cri)$ corresponds to ρ_N in Eq. (3) for any criterion (Cri) with the subscript N omitted.

We transform the problem of ordering the relative U-tendency to sorting the indicators, which are the critical points (roots) of the sign change of the difference functions, under the following assumptions:

- $P(A_1) \approx P(A_2)$, where A_2 denotes the event of $\{\mathcal{E}(\mathcal{X}_N, m^* - 1) < \mathcal{E}(\mathcal{X}_N, m^*)\}$. It actually means *underestimation* is approximated by *locally preferring $m^* - 1$ to m^**. This assumption has been used in e.g., (Zhang et al., 1989; Fishler et al., 2002), and it will be verified via experiments later.

[1] SNR is a term borrowed from signal processing literature, because its definition coincides with those SNR given in signal processing literature. There may be many other definitions for SNR, e.g., $\gamma_o = \lambda_{m^*}/\sigma_e^2$, $\gamma_o = 10\log[\text{power of one signal}/\sigma_e^2]dB$. Here, the definition is the ratio of the smallest factor eigenvalue of the population covariance matrix to the noise eigenvalue, which is equivalent to λ_{m^*}/σ_e^2 up to a constant.

- Assume the probability $P\{\gamma_m \in \Gamma\} > 0$, for any non degenerate interval Γ in the support of γ_m. Otherwise, $P\{\gamma_m \in \Gamma\} = 0$.

At $m = m^*$, suppose Cri_1 and Cri_2 comes from Eq. (3) and satisfy the above assumptions with their difference functions sketched in Fig. 1(c) as well as their indicators (i.e., roots of $\nabla_m \mathscr{E}_{cri}(\gamma_m) = 0$) satisfying $\gamma^*_m(\mathrm{Cri}_1) < \gamma^*_m(\mathrm{Cri}_2)$. Then, Cri_2 has a bigger chance to get a smaller dimensionality, because

$$P\{\gamma_m \in \Gamma^+_m(\mathrm{Cri}_2)\} - P\{\gamma_m \in \Gamma^+_m(\mathrm{Cri}_1)\} = P\{\gamma^*_m(\mathrm{Cri}_1) \leqslant \gamma_m < \gamma^*_m(\mathrm{Cri}_2)\} > 0, \quad (14)$$

where $\Gamma^+_m(\mathrm{Cri}_i) = [1, \gamma^*_m(\mathrm{Cri}_i))$, $i = 1,2$. That is, "the underestimation tendency of Cri_2 is stronger than that of Cri_1" or $Cri_1 \prec$ textsubscript$uCri_2$. Similar analysis on overestimation can be performed at $m = m^* + 1$. Therefore, for AIC, BIC, HQC, CAIC given by Eq. (3), we obtain the following partial order of underestimation tendency from weak to strong:

Theorem 1. *If $N > 16$, then we have: "(NLL$\prec_u$) AIC $\prec_u$ HQC $\prec_u$ BIC $\prec_u$ CAIC"* (Fig. 2).

Remarks 2. The indicator $\gamma^*_m(\mathrm{Cri})$ actually characterizes a lower bound of eSNR γ_m for the criterion (Cri) to avoid the risk of underestimation. In other words, a criterion with a large lower bound requires a higher SNR γ_o, namely stronger factors, because the eSNR γ_m is close to the SNR γ_o with high probability for a large sample size. Actually, for a factor to be identified, the population SNR γ_o must be larger than a critical value according to a phase transition phenomenon in eigenvalues (see e.g., Baik and Silverstein, 2006; Johnstone, 2006; Paul, 2007), when the sample size is relatively small.

Remarks 3. Based on Eq. (7), when underestimation plays a key role in wrong selections, e.g., at the cases of small sample size and weak structure, this partial order of relative U-tendency largely implies an order of accurate selection performance. AIC pays a high risk of overestimation for its robustness against underestimation, while CAIC greatly avoid overestimation through a large penalty but at the expense of being liable to underestimation. Besides, NLL tends to select large m in probability one unless $\gamma_m = 1$, $\forall m > m^*$ or $s_i = \sigma^2$ ($\forall i \geqslant m^*$) which requires $N \to +\infty$.

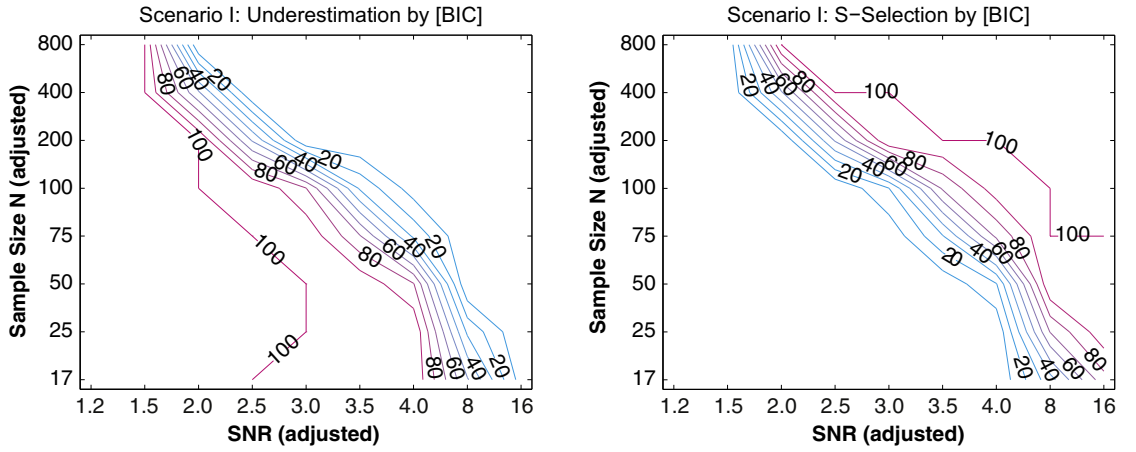
4. Empirical analysis

With the help of a wide scope of controlled experiments on synthetic data, we are able to verify the estimated hidden dimensionality with a known true one, so as to examine the relative strengths and weaknesses of various model selection criteria. Synthetic data are generated according to the FA model with the population covariance matrix $\mathbf{\Sigma}_x = \boldsymbol{U}\boldsymbol{\Lambda}\boldsymbol{U}^T + \sigma_e^2\boldsymbol{I}_n$, where $\boldsymbol{U}^T\boldsymbol{U} = \boldsymbol{I}_{m^*}$, m^* denotes the true underlying hidden dimensionality, $\boldsymbol{\Lambda}$ is a diagonal matrix with $\lambda_1 \geqslant \ldots \geqslant \lambda_{m^*} > 0$ as its diagonal elements, and σ_e^2 is the noise variance as in Eq. (2).

Similar to Tu and Xu (2011), experiments are conducted by varying the values of the sample size N, and the signal–noise-ratio (SNR) $\gamma_o = \lambda_{m^*}/\sigma_e^2 + 1$, under two scenarios of dimensionalities given in Table 1. For each configuration (τ, N, γ_o) in Table 1, 10^3 independent trials are implemented. For every trial, a synthetic data set $\mathscr{X}_N$ is randomly generated. The two-stage procedure is implemented on $\mathscr{X}_N$ for every candidate integer m in $[1, 2m^* - 1]$, among which one is selected by a criterion. The points in the configuration space (N, γ_o) having the same model selection results, including Successful-selection (S) $\hat{m} = m^*$ and Underestimation (U), are connected, and thus we obtain performance contour maps, as given in Fig. 3. Fig. 3 shows that as the sample N and/or γ_o become small (i.e., moving towards the left-bottom of the contour map), the model selection accuracies decrease to zero, while the underestimation rates grow to 100%, which implies that underestimation takes the major proportion of wrong selections, because the effective number of latent factors may be reduced due to a phase transition threshold according to the random matrix theory (Baik and Silverstein, 2006; Johnstone, 2006).

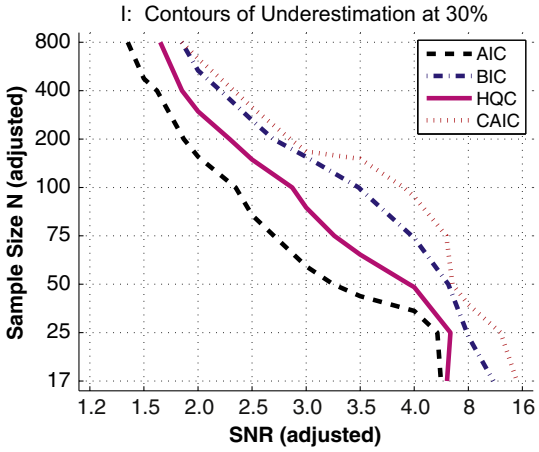
We verify Theorem 1 through experiments on the four scenarios in Table 1. Figs. 4 and 5 present several underestimation contour (U-contour) curves of the same levels in the contour maps of the four criteria in Eq. (3). The result coincides with Theorem 1. Moreover, examples of the successful-selection contours from the four scenarios are reported in Fig. 6, respectively, which indicates a reverse performance order in contrast to the underestimation order in Theorem 1, except when SNR is large but N is very small.

In addition, we verify the effectiveness of an approximation $P(A_1) \approx P(A_2)$ used in the theoretical analysis in Section 3. This approximation has also been adopted as an assumption without comprehensive verifications in (Zhang et al., 1989; Xu and Kaveh, 1995 and Fishler and Poor, 2005). Based on the above experiments,

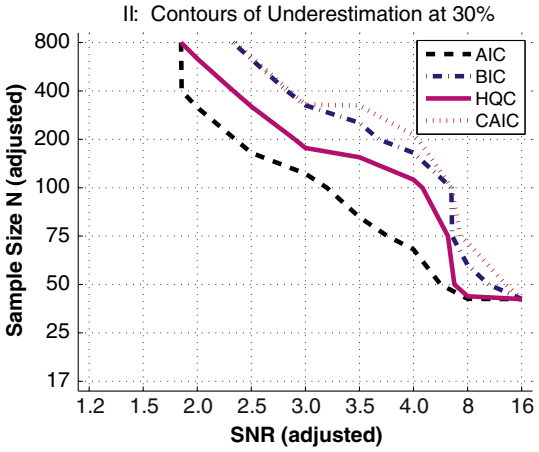


(a) Underestimation contour map of BIC (b) Success-selection contour map of BIC

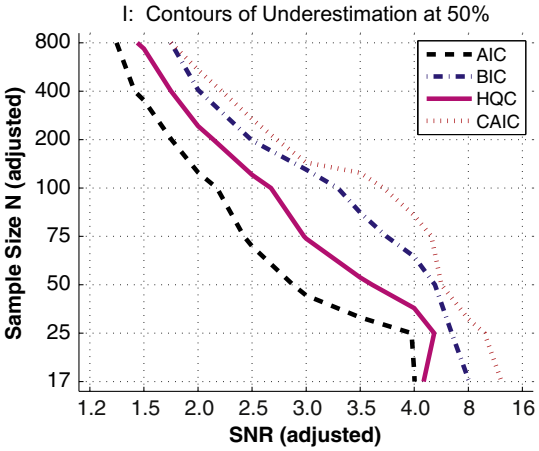
Fig. 3. Examples of contour maps with adjusted axes, i.e., $1.2, \ldots, 16 \in V_{\gamma_o}$ being equally spaced in horizontal-axis, and $25, \ldots, 800 \in V_N$ being equally spaced in vertical-axis (If moving from the top-right to the bottom-left of the contour map, the rates of underestimation increase, whereas the rates of success-selection (S-selection) decrease.).



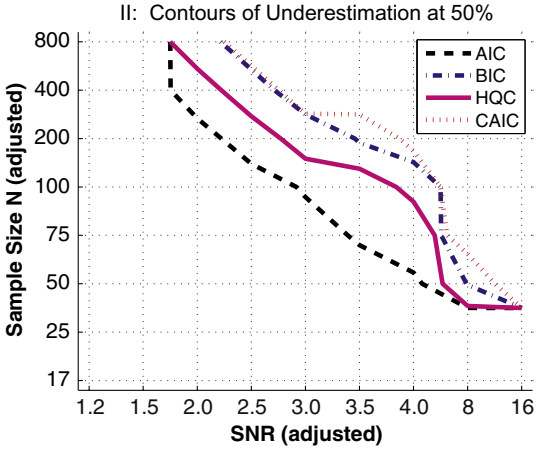
(a) Scenario I: U-Contour 30%



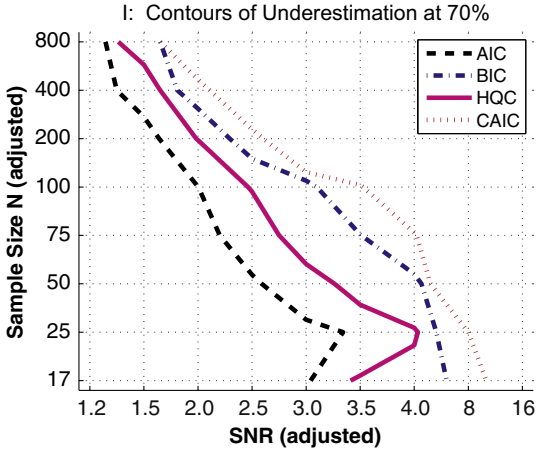
(b) Scenario II: U-Contour 30%



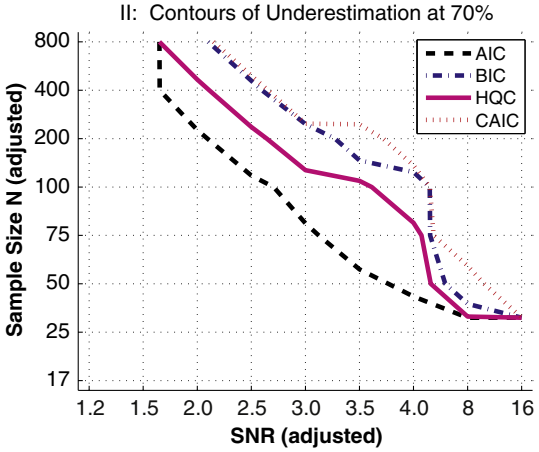
(c) Scenario I: U-Contour 50%



(d) Scenario II: U-Contour 50%



(e) Scenario I: U-Contour 70%.



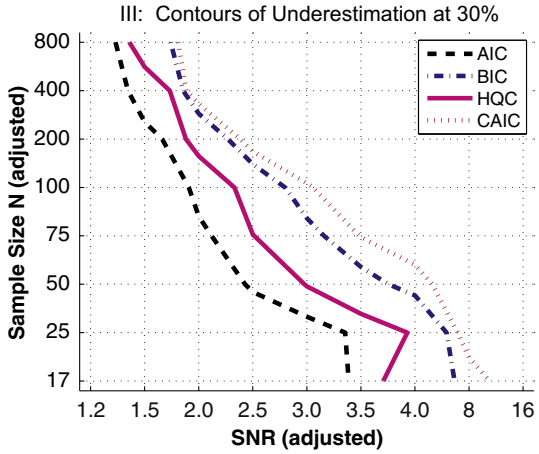
(f) Scenario II: U-Contour 70%.

Fig. 4. The underestimation contours (U-contours) of AIC, BIC, HQC and CAIC at the same level (30%, 50% or 70%) for scenario I on the left and scenario II on the right (In Fig. 4(e), the contour curves of AIC and HQC turn left sharply when moving from $N = 25$ to $N = 17$ where SNR is large, which indicates a rapid reduction on underestimation rates, due to a quick rise of overestimation of AIC and HQC. The sharp turns can also be observed in Fig. 5.)

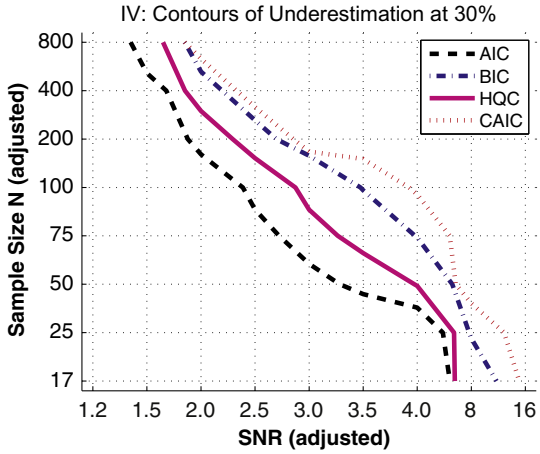
the effectiveness is verified by empirically estimating the difference

$$d(N,\gamma_o) = P(A_1) - P(A_2) \approx [n(A_1) - n(A_2)]/10^3, \tag{15}$$

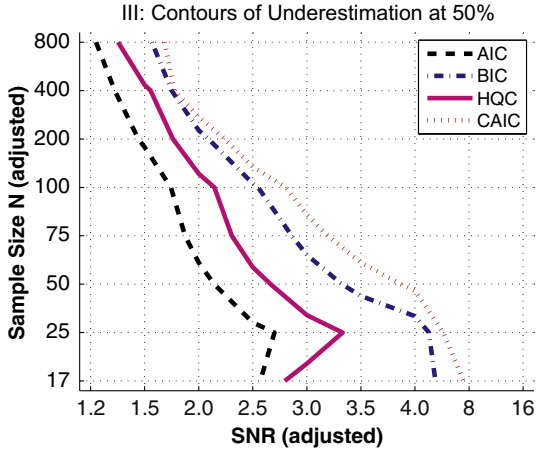
and the relative difference $d(N,\gamma_o)/P(A_1)$, where $n(A_1)$ and $n(A_2)$ are respectively the number of occurrences of event A_1 and A_2 in 10^3 independent trials, for each $(I,N,\gamma_o) \in \{(I,N,\gamma_o)|\forall N \in V_N, \gamma_o \in V_{\gamma_o}\}$. Taking AIC for example, the values of $d(N,\gamma_o)$ and $d(N,\gamma_o)/P(A_1)$ are given in Table 2. The values show that the approximation $P(A_1) \approx P(A_2)$ is effective for most cases, though it is not very good when N is very small and SNR is relatively large. Instead of listing all empirically computed values of $d(N,\gamma_o)$ for each criterion and



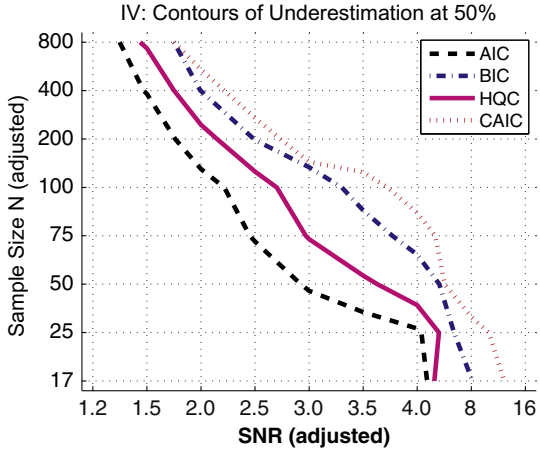
(a) Scenario III: U-Contour 30%



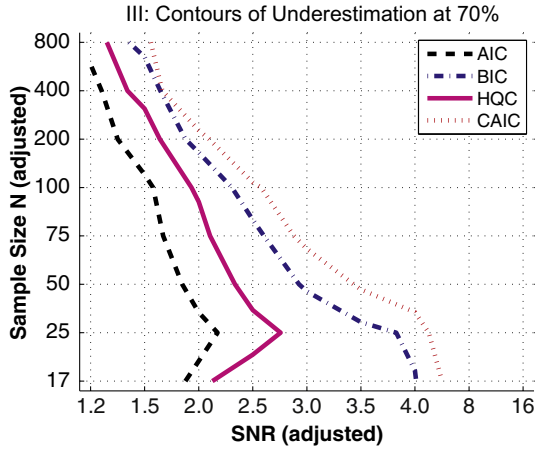
(b) Scenario IV: U-Contour 30%



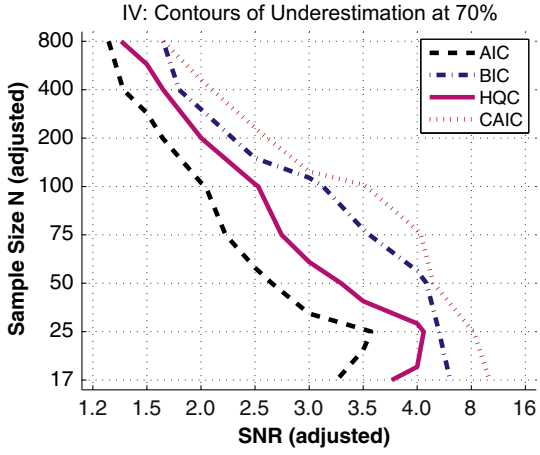
(c) Scenario III: U-Contour 50%



(d) Scenario IV: U-Contour 50%



(e) Scenario III: U-Contour 70%.

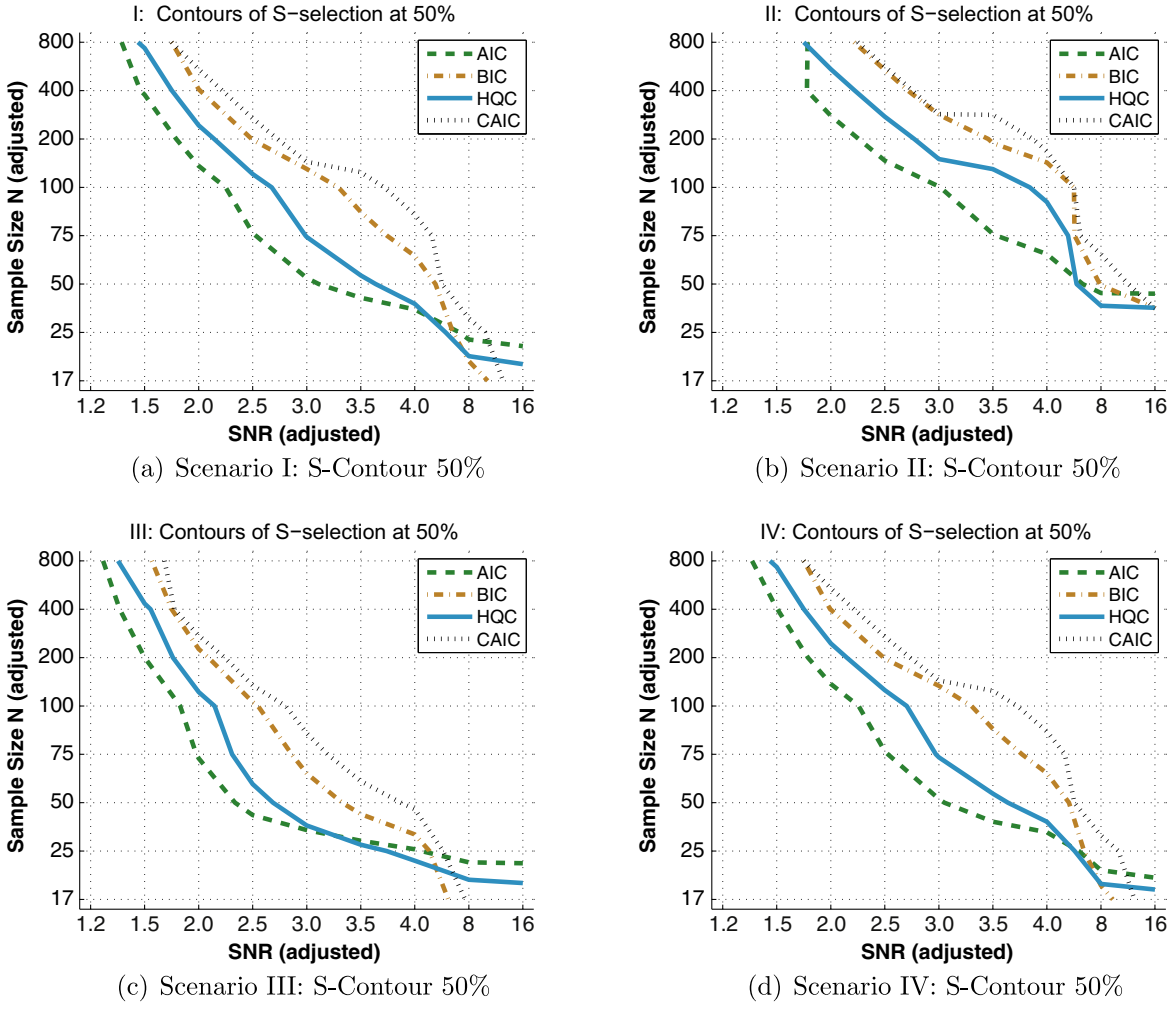


(f) Scenario IV: U-Contour 70%.

Fig. 5. The underestimation contours (U-contours) of AIC, BIC, HQC and CAIC at the same level (30%, 50% or 70%) for scenario III on the left and scenario IV on the right.

for each scenario, we calculate the means and standard deviations of $d(N,\gamma_o)$ on Scenario I for each of the four criteria. Table 3 shows that the approximation is also effective for BIC, HQC, and CAIC.

All criteria are also evaluated on two UCI (Asuncion and Newman, 2007) real world data sets, i.e., Pen-Based Recognition of Handwritten Digits Data Set (denoted as PEN, 16 attributes, 10 classes, 10992 instances), and Waveform Database Generator (Version 1) Data Set (denoted as WAVE, 21 attributes, 3 classes, 5000 instances). Since we do not know the true hidden dimensionality (if any) of this data set, the four criteria are evaluated by their classification performances instead. In practice, FA or PCA is usually used to extract features for tasks like classification, clustering, and regression, so that the computation may be more efficient with the dimensionality and noise reduced.



(a) Scenario I: S-Contour 50% (b) Scenario II: S-Contour 50%

(c) Scenario III: S-Contour 50% (d) Scenario IV: S-Contour 50%

Fig. 6. Example successful-selection contour (S-contour) curves of AIC, BIC, HQC and CAIC at the same level (50%) for the four scenarios. (In contrast to the underestimation contours, the levels of S-contours decrease from the upper-right to the bottom-left of the figure. Therefore, the closer of the same-level S-contours are to the bottom-left, the better that criterion performs.).

Table 2
The empirical values ($\times 10^{-2}$) of the approximation difference $d(N, \gamma_o)$ computed by Eq. (15) and the relative difference $d(N, \gamma_o)/P(A_1)$ for AIC on Scenario I are listed below.

N\SNR	1.2	1.5	2.0	2.5	3.0	3.5	4.0	7	16
Values of difference $d(N, \gamma_o)$									
17	2.4	4.0	4.5	0.8	−2.0	−9.6	−7.6	−7.9	−1.0
25	2.5	3.5	5.0	5.4	5.4	1.7	−1.2	−0.8	0
50	0.4	0.9	3.8	5.2	1.8	−0.1	−0.5	0	0
75	0.1	0.9	4.5	2.0	−0.3	−0.1	0	0	0
100	0.2	1.0	3.7	0.1	−0.1	0	0	0	0
200	0.1	2.5	−0.2	0	0	0	0	0	0
400	0.2	1.0	0	0	0	0	0	0	0
800	1.4	0	0	0	0	0	0	0	0
Values of relative difference $d(N, \gamma_o)/P(A_1)$									
17	2.5	4.3	5.1	1.0	−2.8	−16.5	−15.1	−86.8	−200.0
25	2.5	3.5	5.1	5.7	6.7	2.6	−2.4	−19.5	0
50	0.4	0.9	3.9	6.8	4.3	−0.5	−7.8	0	0
75	0.1	0.9	5.1	4.6	−2.9	−4.5	0	0	0
100	0.2	1.0	5.2	0.8	−7.1	0	0	0	0
200	0.1	2.6	−2.9	0	0	0	0	0	0
400	0.2	2.5	0	0	0	0	0	0	0
800	1.4	0	0	0	0	0	0	0	0

Similarly, we vary the training sample size $N \in \{17, 20, 30, 40, 50, 60, 70, 80, 90, 100, 200\}$. For each of 10^2 runs, we randomly select N instances from every class for training an FA model, based on which a Bayesian classifier is built to classify the remaining instances. The two-stage procedure is implemented on a candidate set $\mathscr{M} = \{1, 2, \ldots, 15\}$. Although the true hidden dimensionality

Table 3
We calculate e.g., the means and standard deviations, of $d(N, \gamma_o)$ on Scenario I to justify the effectiveness of the approximation, where "mean ($|d|$)" is calculated by $\frac{1}{|V_N|\cdot|V_{\gamma_o}|}\sum_{N\in V_N}\sum_{\gamma_o\in V_{\gamma_o}}|d(N,\gamma_o)|$, and so on and so forth.

$\times 10^{-2}$	AIC	BIC	HQC	CAIC
mean($\|d\|$)	1.339	1.942	1.532	2.372
mean(d)	0.467	1.931	1.082	2.372
std(d)	2.507	3.896	2.557	6.016
min(d)	−9.60	−0.30	−7.10	0.0
max(d)	5.40	14.90	7.80	23.30

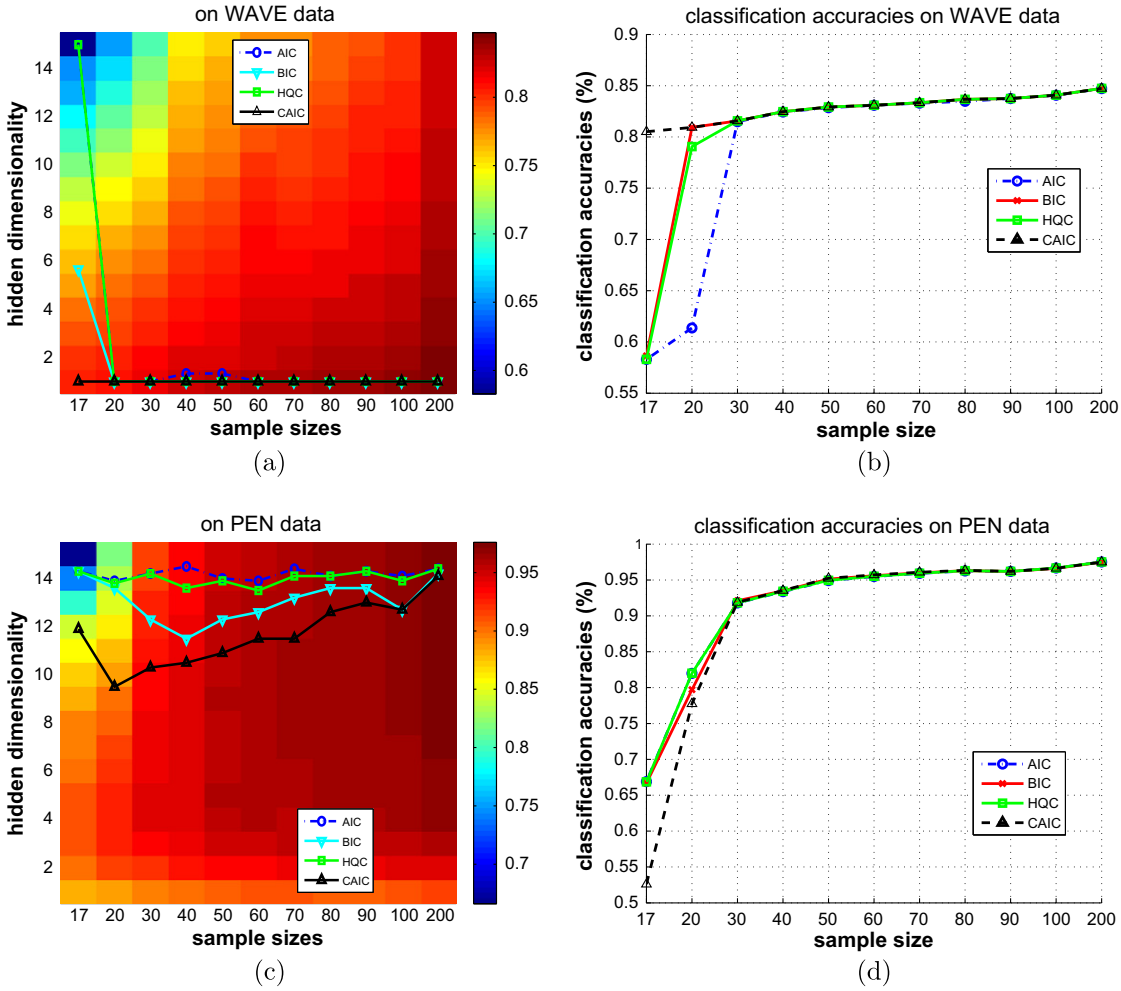


Fig. 7. Results on real world data sets: (a) classification accuracy heat map of the cases that constrain the same hidden dimensionality for each class on WAVE data, as well as the curves of averaged estimated hidden dimensionality by AIC, BIC, HQC, and CAIC; (b) classification accuracies of the two-stage procedures by AIC, BIC, HQC, and CAIC on WAVE data; (c) heat map on PEN data; (The low performance for high-dimensional small-sample-size case may be due to the reduction in effective dimensionality Baik and Silverstein, 2006; Johnstone, 2006.) (d) accuracies on PEN data.

is unknown, the optimal one m_c^* for classification can be estimated according to the classification performances of all FA models trained by enumerating the candidates in $\mathscr{M}$.

The results are reported in Fig. 7. To estimate m_c^*, since it is too expensive to train $|\mathscr{M}|^C$ possible models when there are $|\mathscr{M}|$ candidate dimensions and C classes, we only compute $|\mathscr{M}|$ cases by constraining the hidden dimensionalities to be the same for each class. The classification accuracies of the $|\mathscr{M}|$ cases (vertical axis) are shown by the heat maps in Fig. 7(a) and (c) against a varying sample size (horizontal axis), which approximately implies $m_c^* = 1$ for WAVE and $m_c^* = 14$ for PEN. For a comparison, the estimated hidden dimensionality for each class is averaged and plotted in the heat maps as four curves (corresponding to AIC, BIC, HQC, and CAIC), which show that the four criteria tend to correctly selecting m_c^* as sample size grows, while for a small sample size they choose different dimensionalities but in the same order as Fig. 2. Moreover, the classification accuracies of the four criteria are reported in Fig. 7(b) and (d), which coincide in order with the four curves in the Fig. 7(a) and (c), respectively. It should be noted that Fig. 7(b) shows a reversed performance order from Fig. 7(d), because AIC, BIC, and HQC tends to overestimation for WAVE.

5. Concluding remarks

Based on the problem of determining the hidden dimensionality of FA, the relative strength and weakness of several model selection methods has been investigated theoretically. We concentrate on building up a partial order of relative underestimation tendency

of several criteria. We found that the order is AIC, HQC, BIC/MDL and CAIC from weak to strong. This order is an intrinsic relation of criterion functions to the latent factors' strength. Experiments have been conducted to verify this order and the effectiveness of a common assumption used in this paper and literature.

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: CUHK418012E). The first author would like to thank Prof. Daming Shi, and Mr. Ke Sun, Mr. Lei Shi for their helpful comments and suggestions.

Appendix A. Proofs

Proof of Lemma 1. Replacing γ_m with an independent variable γ, we regard $\nabla_m \mathscr{E}_L$ by Eq. (12) as a function of γ, and then $\forall \gamma \in [1,+\infty)$, we have $\frac{\partial \nabla_m \mathscr{E}_L(\gamma)}{\partial \gamma} = -\frac{k(\gamma-1)}{(k+\gamma)\gamma} \leqslant 0$, $k = n-m$, which implies that $\nabla_m \mathscr{E}_{cri}(\gamma)$ is monotone decreasing, because $\alpha(m,N)$ is irrelevant to γ. Then, $\nabla_m \mathscr{E}_{cri}(\gamma) \leqslant \nabla_m \mathscr{E}_{cri}(1) = \alpha$. □

Proof of Lemma 2. After some straightforward manipulations, $\nabla_m \mathscr{E}_L(\gamma) = -\alpha$ is equivalent to $f(z) = z^{k+1} - (k+1)C_1 z + kC_1 = 0$ with $z = k+\gamma \geqslant k+1$, and $C_1 = (k+1)^k e^{\alpha}$.

Solving $f'(z) = (k+1)(z^k - C_1) = 0$ we get $z_{min} = \sqrt[k]{C_1} > k+1$, and then $f(z)$ has a local minimum at z_{min} because $f''(z) = k(k+1)z^{k-1} > 0$. Since $f(k+1) = (k+1)^{k+1}(1-e^{\alpha}) < 0$, there is no root of $f(z)$ in its decreasing interval $[k+1, z_{min}]$. As $f(z)$ is monotonic increasing in $[z_{min}, +\infty)$, $f(z)$ will have a unique root in $[z_{min}, +\infty)$, i.e., $z^* > z_{min}$.

Approximating $f(z)$ by Taylor-expansion at z_{min} to the second-order, we have $f(z) \approx \hat{f}(z) = f(z_{min}) + \frac{1}{2}f''(z_{min})(z - z_{min})^2$, and get $z_{up} = \sqrt[k]{C_1} + \sqrt{2(k+1)C_0(C_0-1)}$ by solving $\hat{f}(z) = 0$, where $C_0 = \exp\{\alpha/k\} = \exp\left\{\frac{(k+1)\rho}{kN}\right\}$. As $f^{(j)}(z) = (k+1)!z^{t-j+1}/(t-j+1)! > 0$, $f(z)$ grows faster than $\hat{f}(z)$ at $[z_{min}, +\infty)$, which implies that $z^* < z_{up}$. Notice that $\gamma = z - k$, then the Lemma 2(1) holds.

Since $\alpha = \frac{k+1}{N}\rho$, then $\alpha(Cri_1) > \alpha(Cri_2) > 0$ if $\rho(Cri_1) > \rho(Cri_2) > 0$. Then, we obtain Lemma 2(2), i.e., $\gamma^*(Cri_1) > \gamma^*(Cri_2) > 1$, because of the monotone decreasing property of $\nabla_m \mathscr{E}_{cri}(\gamma)$. Moreover, an approximation formula is obtained by substituting $C_0 \approx 1 + \frac{\alpha}{k}$ into γ_{up}, i.e., $\gamma_m^* \approx 1 + \alpha + \frac{\alpha}{k} + \sqrt{2(k+1)(1+\frac{\alpha}{k})\frac{\alpha}{k}} + O(\frac{\alpha}{k})$. □

Proof of Theorem 1. By Eq. (4), $0 = \rho(\text{NLL}) < \rho(\text{AIC}) < \rho(\text{HQC}) < \rho(\text{BIC}) < \rho(\text{CAIC})$, if the sample size $N > 16$. Then, it follows from Lemma 2(2) that the indicators satisfy $1 = \gamma^*_{m^*}(\text{NLL}) < \gamma^*_{m^*}(\text{AIC}) < \gamma^*_{m^*}(\text{HQC}) < \gamma^*_{m^*}(\text{BIC}) < \gamma^*_{m^*}(\text{CAIC})$, which implies the relative U-tendency order according to Eq. (14). □

An Example that $\bar{\gamma}(\mathscr{X}_N)$ is a two-variable vector:

We define corresponding objective of the difference of negative log-likelihood (DNLL) is $\mathscr{E}_{\text{DNLL}}(\mathscr{X}_N, \widehat{\boldsymbol{\Theta}}_m^{ML}) = \nabla_m \mathscr{E}_L$, where $\nabla_m \mathscr{E}_L$ is given in Eq. (12). Then, its difference function is

$$\nabla_m(\mathscr{E}_{\text{DNLL}}) = \nabla_m^2 \mathscr{E}_L = -2(k+1)\ln\left(1 + \frac{\gamma_m - 1}{k+1}\right) + (k+2)\ln\left(1 + \frac{\beta_m + \gamma_m - 2}{k+2}\right) - \ln\frac{\beta_m}{\gamma_m}, \tag{A.1}$$

where $\beta_m = s_{m-1}/\mathscr{A}^n_{m+1}$, γ_m and $\mathscr{A}^n_{m+1}$ are given in Eq. (13), and $\beta_m \geqslant \gamma_m \geqslant 1$.

Then, the statistic $\bar{\gamma}(\mathscr{X}_N)$ is generalized to be a two variable vector (β_m, γ_m) which characterizes a 2-dimensional boundary of the sign change of the difference function. According to the property of Eq. (A.1), we have the following result: *(1)* When $m = m^*$: If $s_{m-1} \approx s_m \gg \mathscr{A}^n_{m+1}$, then $\gamma_{m-1,m} \approx \gamma_{m,\ m} \gg 1$, which implies $\nabla_m \mathscr{E}_{\text{DNLL}} < 0$, i.e., m^* is preferred to $m^* - 1$.*(2)* When $m - 1 = m^*$: If $s_{m-1} \gg s_m \approx \mathscr{A}^n_{m+1}$, then $\gamma_{m-1,m} \gg \gamma_{m,\ m} \approx 1$, which implies $\nabla_m \mathscr{E}_{\text{DNLL}} > 0$, i.e., m^* is preferred to $m^* + 1$. The above results implies that DNLL favors the case with slightly dispersed latent factors and noise eigenvalues (i.e., prefers scenario I to II (a)) with large SNR. This agrees with the experimental results for DNLL (not shown here). Note that $\nabla_m \mathscr{E}_{\text{DNLL}} = 0$ if $\gamma_{m-1,m} = \gamma_{m,\ m} = 1$ or $s_i = const.$, $\forall i$. Let $k = n - m$, $\delta_\gamma(m) = \gamma_{m-1,m} - \gamma_{m,\ m} > 0$. Then the above results hold because:

$$\frac{\partial \nabla_m \mathscr{E}_{\text{DNLL}}}{\partial \gamma_{m,m}} = \frac{\left\{\begin{matrix} -k(k + \gamma_{m-1,m} + \gamma_{m,m})(\gamma_{m,m} - 1) \\ -\gamma_{m,m}[k(\gamma_{m-1,m} - 1) + \delta_\gamma] \end{matrix}\right\}}{(k + \gamma_{m-1,m} + \gamma_{m,m})(k + \gamma_{m,m})\gamma_{m,m}} \leqslant 0,$$

$$\frac{\partial \nabla_m \mathscr{E}_{\text{DNLL}}}{\partial \gamma_{m-1,m}} = \frac{k(\gamma_{m-1,m} - 1) + \delta_\gamma}{(k + \gamma_{m-1,m} + \gamma_{m,m})\gamma_{m-1,m}} \geqslant 0,$$

$$\frac{\partial \nabla_m \mathscr{E}_{\text{DNLL}}|_{\delta_\gamma = 0}}{\partial \gamma_{m,m}} = \frac{-2k(\gamma_{m,m} - 1)}{(l + 2\gamma_{m,m})(k + \gamma_{m,m})} \leqslant 0.$$

References

Akaike, H., 1974. A new look at the statistical model identification. IEEE Trans. Autom. Control 19, 716–723.

Anderson, T., Rubin, H., 1956. Statistical inference in factor analysis. In: Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability, pp. 111–150.

Asuncion, A., Newman, D., 2007. UCI Machine Learning Repository.

Baik, J., Silverstein, J.W., 2006. Eigenvalues of large sample covariance matrices of spiked population models. J. Multivar. Anal. 97, 1382–1408.

Bishop, C.M., 1999. Variational principal components. In: IEE Conference Publication on Artificial Neural Networks ICANN99, pp. 509–514.

Bozdogan, H., 1987. Model selection and Akaike's Information Criterion (AIC): the general theory and its analytical extensions. Psychometrika 52, 345–370.

Chen, P., Wu, T.J., Yang, J., 2008. A comparative study of model selection criteria for the number of signals. IET Radar Son. Nav. 2, 180–188.

Fishler, E., Grosmann, M., Messer, H., 2002. Detection of signals by information theoretic criteria: general asymptotic performance analysis. IEEE Trans. Signal Process. 50, 1027–1036.

Fishler, E., Poor, H., 2005. Estimation of the number of sources in unbalanced arrays via information theoretic criteria. IEEE Trans. Signal Process. 53, 3543–3553.

Hannan, E., Quinn, B., 1979. The determination of the order of an autoregression. J. Roy. Statist. Soc. Ser. B 41, 190–195.

Hu, X., Xu, L., 2004. A comparative investigation on subspace dimension determination. Neural Netw. 17, 1051–1059.

Johnstone, I.M., 2001. On the distribution of the largest eigenvalue in principal component analysis. Ann. Statist. 29, 295–327.

Johnstone, I.M., 2006. High dimensional statistical inference and random matrices. In: Proceedings International Congress of Mathematicians.

Jolliffe, I.T., 2002. Principal Component Analysis, second ed. Springer.

Kim, S.W., 2011. An empirical evaluation on dimensionality reduction schemes for dissimilarity-based classifications. Pattern Recognition Lett. 32, 816–823.

Liavas, A., Regalia, P., 2001. On the behavior of information theoretic criteria for model order selection. IEEE Trans. Signal Process. 49, 1689–1695.

Nadakuditi, R., Edelman, A., 2008. Sample eigenvalue based detection of high-dimensional signals in white noise using relatively few samples. IEEE Trans. Signal Process. 56, 2625–2638.

Nadler, B., 2010. Nonparametric detection of signals by information theoretic criteria: performance analysis and an improved estimator. IEEE Trans. Signal Process. 58, 2746–2756.

Paul, D., 2007. Asymptotics of sample eigenstructure for a large dimensional spiked covariance model. Statistica Sinica 17, 1617–1642.

Rissanen, J., 1978. Modelling by the shortest data description. Automatica 14, 465–471.

Schwarz, G., 1978. Estimating the dimension of a model. Ann. Statist. 6, 461–464.

Tipping, M.E., Bishop, C.M., 1999. Mixtures of probabilistic principal component analyzers. Neural Comput. 11, 443–482.

Tu, S., Xu, L., 2009. Theoretical analysis and comparison of several criteria on linear model dimension reduction. In: ICA '09: Proceedings of the 8th International Conference on Independent Component Analysis and Signal Separation, Springer-Verlag, Berlin, Heidelberg. pp. 154–162.

Tu, S., Xu, L., 2011. An investigation of several typical model selection criteria for detecting the number of signals. Frontiers of Electrical and Electronic Engineering in China.

Tubbs, J., Coberly, W., Young, D., 1982. Linear dimension reduction and bayes classification with unknown population parameters. Pattern Recognit. 15, 167–172.

Villegas, M., Paredes, R., 2011. Dimensionality reduction by minimizing nearest-neighbor classification error. Pattern Recognition Lett. 32, 633–639.

Wang, X., Paliwal, K.K., 2003. Feature extraction and dimensionality reduction algorithms and their applications in vowel recognition. Pattern Recognit. 36, 2429–2439.

Wax, M., Kailath, T., 1985. Detection of signals by information theoretic criteria. IEEE Trans. Acoust. Speech Signal Process. ASSP-33, 387.

Xu, W., Kaveh, M., 1995. Analysis of the performance and sensitivity of eigendecomposition-based detectors. IEEE Trans. Signal Process. 43, 1413–1426.

Zhang, Q.T., Wong, K., Yip, P., Reilly, J., 1989. Statistical analysis of the performance of information theoretic criteria in the detection of the number of signals in array processing. IEEE Trans. Acoust. Speech Signal Process. 37, 1557–1567.