ISSN 1673-3460

Volume 6 · Number 2 · June 2011

Available online
www.springerlink.com

Frontiers of Electrical and Electronic Engineering in China

高等教育出版社
HIGHER EDUCATION PRESS

Springer

Front. Electr. Electron. Eng. China 2011, 6(2): 245–255
DOI 10.1007/s11460-011-0146-y

RESEARCH ARTICLE

Shikui TU, Lei XU

An investigation of several typical model selection criteria for detecting the number of signals

Abstract Based on the problem of detecting the number of signals, this paper provides a systematic empirical investigation on model selection performances of several classical criteria and recently developed methods (including Akaike's information criterion (AIC), Schwarz's Bayesian information criterion, Bozdogan's consistent AIC, Hannan-Quinn information criterion, Minka's (MK) principal component analysis (PCA) criterion, Kritchman & Nadler's hypothesis tests (KN), Perry & Wolfe's minimax rank estimation thresholding algorithm (MM), and Bayesian Ying-Yang (BYY) harmony learning), by varying signal-to-noise ratio (SNR) and training sample size N. A family of model selection indifference curves is defined by the contour lines of model selection accuracies, such that we can examine the joint effect of N and SNR rather than merely the effect of either of SNR and N with the other fixed as usually done in the literature. The indifference curves visually reveal that all methods demonstrate relative advantages obviously within a region of moderate N and SNR. Moreover, the importance of studying this region is also confirmed by an alternative reference criterion by maximizing the testing likelihood. It has been shown via extensive simulations that AIC and BYY harmony learning, as well as MK, KN, and MM, are relatively more robust than the others against decreasing N and SNR, and BYY is superior for a small sample size.

Keywords number of signals, array processing, factor analysis, principal component analysis (PCA), model selection criteria

1 Introduction

Detecting the number of source signals is an essential issue in many signal processing problems such as sensor array processing, the poles retrieval of a system response, the direction of arrival estimation by a smart antenna system, retrieving the overlapping echoes from radar backscatter, and so on (see e.g., Refs. [1,2]). The observed vector can be modeled as a superposition of a finite number of underlying Gaussian source signals with an additive Gaussian noise [2]. The signal-number determination is also addressed as a model selection problem [1,3] in machine learning and statistics, i.e., selecting the latent dimensionality of a factor analysis (FA) [4] model. Revisited in Ref. [5], FA implements principal component analysis (PCA) as a special case under the maximum likelihood (ML) principle.

To tackle this model selection problem, a traditional approach is a two-stage implementation, i.e., parameter learning is repeated on a set of candidate latent dimensionalities among which one is selected by a model selection criterion. Existing classical criteria, such as Akaike's information criterion (AIC) [6] etc., trade off between the likelihood-based goodness of fit and model complexity, subject to noise and uncertainty in a finite number of observations. Recently, on one hand, Minka's criterion (MK) [7] is a further developed Bayesian model selection method for PCA, while Bayesian Ying-Yang (BYY) harmony learning [8,9] is another statistical learning framework for model selection. On the other hand, based on the recent results of sample covariance asymptotics stemming from random matrix theory [10–12], new algorithms were proposed for the related rank estimation problem, including Kritchman & Nadler's hypothesis test (KN) [13] and Perry & Wolfe's minimax rank estimation thresholding algorithm (MM) [14].

It is important to examine the relative strengths and weaknesses of these model selection methods. One way [3,15,16] is to empirically examine their model selection performances by varying signal structure, training sample size N, etc. The other way is to formally analyze their statistical properties such as consistency. Initialized from Ref. [1] and followed by Refs. [17–22], AIC

Received February 28, 2011; accepted March 24, 2011

Shikui TU, Lei XU (✉)
Department of Computer Science and Engineering, The Chinese University of Hong Kong, Hong Kong, China
E-mail: lxu@cse.cuhk.edu.hk

and minimum description length (MDL) [23] were introduced to determine the number of signals with efforts on approximating the underestimation (or overestimation) probability and asymptotical consistency under an infinite N. Moreover, recent progress in random matrix theory demonstrates the existence of a phase transition threshold (for the eigenvalues of the covariance), below which the effective number of signals is reduced [10–12,22] under the limit $N, n \to \infty$ with $n/N \to c$, where n is the dimensionality of observations, and $c > 0$ is a constant. In this context, MK and KN were respectively analyzed in Refs. [13,24], while MM was shown to admit asymptotic minimax optimality [14].

Following the track on the problem of detecting the underlying signal number, this paper aims at a systematic empirical comparison on a wide scope of their performances of the above methods in model selection and testing likelihood. We examine the *joint* effect of training sample size N and signal-to-noise ratio (SNR, defined as the ratio of the smallest signal eigenvalue to the noise eigenvalue of the population covariance matrix), rather than the single effect of either of SNR and N with the other fixed, which is not adequate for a systematic comparison.

In experiments, model selection accuracies are collected under extensive configurations of N and SNR. The contours of equal accuracies are connected to obtain a family of model selection indifference curves (a term borrowed from economics). With the help of indifference curves, we are able to visually reveal a diminishing marginal effect that the amount of SNR (or N) to trade for a unit of N (or SNR) grows if the model selection accuracy is kept at the same level, and also able to identify a three-region partition in the configuration space. All methods perform well within the range that N and SNR are large, but unavoidably suffer from underestimation due to a reduction of the effective number of signals [10,12,22] within the range that N and SNR are too small. Interestingly, the relative strengths and weaknesses of these methods are obviously demonstrated within a region with moderate N and SNR. This region deserves more investigations.

Moreover, a further verification is made with help of a reference criterion that selects a model with the maximum testing likelihood (TLL) as an approximation of generalization risk. We have observed not only the reduction of the number of effective signals when N and SNR are too small, but also the importance of this region with the following features: 1) AIC and BYY, as well as MK, KN, and MM, are more robust against decreasing N and SNR; 2) BYY is superior in small-sample-size area for model selection and with smaller generalization error than TLL.

The rest of this paper is organized as follows. Section 2 formulates the problem of determining the number of signals as estimating the hidden dimensionality of FA. Section 3 reviews the model selection methods to be investigated. Section 4 is devoted to a systematic empirical analysis. Section 5 concludes this paper.

2 Problem description

In several important problems, such as sensor array processing in signal processing [1,2], a common model for the signal vector $\boldsymbol{x}(t)$ from an array of n sensors at time instance t is $\boldsymbol{x}(t) = \boldsymbol{A}_\phi \boldsymbol{s}(t) + \boldsymbol{e}(t)$, where $\boldsymbol{A}_\phi$ is often referred to as the steering matrix with full column rank. The m-dimensional source signal vector sequence $\{\boldsymbol{s}(t)\}$ is assumed to be a stationary and ergodic Gaussian random process with zero mean and positive definite covariance matrix $\boldsymbol{\Sigma}_s$. The additive noise vector sequence $\{\boldsymbol{e}(t)\}$ is assumed to be a stationary and ergodic Gaussian vector process, independent of the source signals, with zero mean and covariance matrix $\boldsymbol{\Sigma}_e = \sigma_e^2 \boldsymbol{I}_n$, where σ_e^2 is an unknown scalar and $\boldsymbol{I}_n$ is the $n \times n$ identity matrix. The signals and unknown parameters are complex-valued. Associated with this model, a key problem which has received much attention in signal processing literature (e.g., Refs. [1,17,18,20,21]) is to determine the number of source signals based on an observed sequence $\boldsymbol{x}(t), t = 1, 2, \ldots, N$, or the rank of $\boldsymbol{A}_\phi \boldsymbol{\Sigma}_s \boldsymbol{A}_\phi^{\mathrm{H}}$ in the following equation:

$$\boldsymbol{\Sigma}_x = \boldsymbol{A}_\phi \boldsymbol{\Sigma}_s \boldsymbol{A}_\phi^{\mathrm{H}} + \sigma_e^2 \boldsymbol{I}_n, \tag{1}$$

where $\boldsymbol{\Sigma}_x$ is the population covariance matrix of the received data, and the superscript "H" means the complex conjugate transpose.

On the other hand, a model called Factor Analysis (FA) in machine learning [25] and statistics [4], assumes an n-dimensional real-valued observation $\boldsymbol{x}$ to be distributed as follows:

$$\begin{cases} \boldsymbol{x} = \boldsymbol{A}\boldsymbol{y} + \boldsymbol{\mu} + \boldsymbol{e}, \quad \boldsymbol{\Theta}_m = \{\boldsymbol{A}, \boldsymbol{\mu}, \boldsymbol{\Sigma}_y, \boldsymbol{\Sigma}_e\}, \\ p(\boldsymbol{x}|\boldsymbol{y}) = G(\boldsymbol{x}|\boldsymbol{A}\boldsymbol{y} + \boldsymbol{\mu}, \boldsymbol{\Sigma}_e), \\ p(\boldsymbol{y}) = G(\boldsymbol{y}|\boldsymbol{0}, \boldsymbol{\Sigma}_y), \\ p(\boldsymbol{x}|\boldsymbol{\Theta}_m) = \int p(\boldsymbol{x}|\boldsymbol{y}) p(\boldsymbol{y}) \mathrm{d}\boldsymbol{y} = G(\boldsymbol{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma}_x), \\ \boldsymbol{\Sigma}_y = \boldsymbol{I}_m \ (m \times m \text{ identity matrix}) \ , \\ \boldsymbol{\Sigma}_x = \boldsymbol{A}\boldsymbol{\Sigma}_y \boldsymbol{A}^{\mathrm{T}} + \boldsymbol{\Sigma}_e, \end{cases} \tag{2}$$

where $\boldsymbol{y}$ is an $m \times 1$ hidden factor vector, $\boldsymbol{\Theta}_m$ denotes the set of the model parameters of FA with the hidden dimensionality being m, $\boldsymbol{A}$ 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 a mean vector $\boldsymbol{\mu}$ and a covariance matrix $\boldsymbol{\Sigma}$. Following the track in signal processing, we set FA in its special case by $\boldsymbol{\mu} = \boldsymbol{0}, \boldsymbol{\Sigma}_e = \sigma_e^2 \boldsymbol{I}_n$, which is equivalent to PCA [4,5] under the maximum likelihood principle. Then, the population covariance matrix of the observations is

$$\boldsymbol{\Sigma}_x = \boldsymbol{A}\boldsymbol{\Sigma}_y\boldsymbol{A}^{\mathrm{T}} + \sigma_e^2\boldsymbol{I}_n. \tag{3}$$

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

The two rank estimation problems by Eqs. (1) and (3) are equivalent, because they both seek a similar decomposition of a sample covariance matrix with a matrix rank of $\boldsymbol{A}_\phi\boldsymbol{\Sigma}_s\boldsymbol{A}_\phi^{\mathrm{H}}$ (or $\boldsymbol{A}\boldsymbol{\Sigma}_y\boldsymbol{A}^{\mathrm{T}}$) being smaller than n, i.e., $m < n$, though the estimated matrix $\boldsymbol{A}_\phi$ in Eq. (1) and $\boldsymbol{A}$ in Eq. (3) may not be equivalent due to certain indeterminacies in the decompositions of Eqs. (1) and (3).

3 Methods

Given a sample set $\mathcal{X}_N = \{\boldsymbol{x}_t\}_{t=1}^N$, where the mean is assumed to be zero, the task of FA modeling consists of estimating the parameters $\boldsymbol{\Theta}_m$ and selecting the number of factors m, traditionally tackled by a two-stage procedure:

1) Compute $\hat{\boldsymbol{\Theta}}_m = \hat{\boldsymbol{\Theta}}(\mathcal{X}_N, m)$ for each candidate m. Normally, $\hat{\boldsymbol{\Theta}}_m$ is an ML estimator $\hat{\boldsymbol{\Theta}}_m^{\mathrm{ML}} = \arg\min_{\boldsymbol{\Theta}_m} \mathcal{J}_{\mathrm{nll}}(\mathcal{X}_N|\boldsymbol{\Theta}_m)$, where $\mathcal{J}_{\mathrm{nll}} = -\frac{2}{N}\ln p(\mathcal{X}_N|\boldsymbol{\Theta}_m)$ is denoted as negative log-likelihood (NLL).

2) Estimate $\hat{m} = \arg\min_m \mathcal{J}_{\mathrm{Cri}}$, where $\hat{m}$ gives an estimate of the true hidden dimensionality m^*, and $\mathcal{J}_{\mathrm{Cri}}$ is a model selection criterion (Cri), e.g.,

$$\mathcal{J}_{\mathrm{Cri}}(\mathcal{X}_N, \hat{\boldsymbol{\Theta}}_m) = \mathcal{J}_{\mathrm{nll}}(\mathcal{X}_N, \hat{\boldsymbol{\Theta}}_m) + (\rho_N d_m)/N, \tag{4}$$

$$d_m = nm + 1 - [m(m-1)]/2, \quad n = \dim(\boldsymbol{x}), \tag{5}$$

$$\rho_N = \begin{cases} 2, & \text{for AIC [6]}, \\ \ln N, & \text{for BIC/MDL [23,26]}, \\ \ln N + 1, & \text{for CAIC [27]}, \\ 2\ln(\ln N), & \text{for HQC [28]}, \end{cases} \tag{6}$$

where d_m in Eq. (5) is the number of free parameters of the real-valued FA model, and $d_m = m(2n-m)+1$ for the complex-valued FA [1]. In the sequel, we focus on the real-valued case.

The criterion by Eq. (4) aims to trade off between the log likelihood and the model/sample complexity. Besides, the difference of the NLL of two FA models, denoted as DNLL, allows model selection by calculating the log ratio of the likelihood values. The DNLL estimate $\hat{m} = \arg\min_m \mathcal{J}_{\mathrm{DNLL}}(\mathcal{X}_N, \hat{\boldsymbol{\Theta}}_m^{\mathrm{ML}})$, where $\mathcal{J}_{\mathrm{DNLL}} = \mathcal{J}_{\mathrm{nll}}(\mathcal{X}_N, \hat{\boldsymbol{\Theta}}_m^{\mathrm{ML}}) - \mathcal{J}_{\mathrm{nll}}(\mathcal{X}_N, \hat{\boldsymbol{\Theta}}_{m-1}^{\mathrm{ML}})$.

Model selection aims to avoid overfitting by minimizing the generalization error $\mathrm{KL}(p_o\|p(\hat{\boldsymbol{\Theta}}_m^{\mathrm{ML}})) = \int p_o(\boldsymbol{x})\ln[p_o(\boldsymbol{x})/p(\boldsymbol{x}|\hat{\boldsymbol{\Theta}}_m^{\mathrm{ML}})]\,\mathrm{d}\boldsymbol{x}$, where $p_o(\boldsymbol{x})$ denotes the true probability density function. Suppose $X'_{N'}$ is a test set of a large size N' from p_o, then $\max_m \ln p(X'_{N'}|\hat{\boldsymbol{\Theta}}_m^{\mathrm{ML}})$ and $\min_m \mathrm{KL}(p_o\|p(\boldsymbol{x}|\hat{\boldsymbol{\Theta}}_m^{\mathrm{ML}}))$ are approximately equivalent. We use TLL to denote the following criterion:

$$\hat{m}(\mathrm{TLL}) = \arg\max_m \ln p(X'_{N'}|\hat{\boldsymbol{\Theta}}_m^{\mathrm{ML}}). \tag{7}$$

TLL is regarded as a reference criterion which selects the optimal ML estimate by maximizing testing log-likelihood as an approximation of generalization risk.

For a wide scope of model selection methods, we conduct the performance evaluation on not only several classical criteria such as AIC [6], Bayesian information criterion (BIC) [26], but also three recent approaches as follows:

1) Kritchman & Nadler's estimator [13] and Perry & Wolfe's minimax rank estimator [14], two examples of development based on the sample covariance asymptotics stemming from random matrix theory [10–12];

2) Minka's criterion [7], a further approximation for Bayesian model selection;

3) Bayesian Ying-Yang harmony learning, developed from another direction for model selection.

3.1 Kritchman & Nadler's hypothesis test (KN)

Kritchman & Nadler's algorithm [13] performs a sequence of hypothesis tests on the hidden dimensionality, with the help of a matrix perturbation approach for the interactions between noise and signal eigenvalues. The KN test algorithm is based on a result that the distribution of the largest eigenvalue s_1 of a d-dimensional sample covariance matrix, computed from pure Gaussian noise observations with the zero mean and the population covariance being $\sigma^2\boldsymbol{I}_d$, converges to a Tracy-Widom distribution in the joint limit $d, N \to \infty$ with $d/N = c$ fixed [29], i.e.,

$$\Pr\{s_1 < \sigma^2(\mu_{N,d} + s\sigma_{N,d})\} \to F_\beta(s), \tag{8}$$

where F_β denotes the Tracy-Widom distribution[1] of order β, and $\beta = 1, 2$ respectively corresponds to real or complex-valued observations. As described in Ref. [11], for real-valued observations the following equations:

$$\begin{cases} \mu_{N,d} = \frac{1}{2n}(\sqrt{N-1} + \sqrt{d-1})^2, \\ \sigma_{N,d} = \sqrt{\mu_{N,d/N}}\left(\sqrt{2/(N-1)} + \sqrt{2/(d-1)}\right)^{1/3}, \end{cases} \tag{9}$$

give an $O_P(d^{-2/3})$ rate of convergence in Eq. (8). Moreover, for the case of n-dimensional population covariance with m signal eigenvalues, Kritchman and Nadler proposed in Ref. [13] a consistent estimator σ^2_{KN} (in the joint limit $n, N \to \infty$ with $n/N = c$) for the unknown

[1] The Tracy-Widom distribution F_β can be explicitly computed from the solution of a second order Painlevé ordinary differential equation [29]. A Matlab code for computing F_β is given at http://momardieng.com/mathematics.

σ^2, which amounts to solving the following $m+1$ equations involving the unknowns $\hat{\rho}_1, \hat{\rho}_2, \ldots, \hat{\rho}_m$ and σ^2_{KN}:

$$\begin{cases} \sigma^2_{\mathrm{KN}} - \frac{1}{n-m}\left[\sum_{j=m+1}^{n} s_j + \sum_{j=1}^{m}(s_j - \hat{\rho}_j)\right] = 0, \\ \hat{\rho}_j^2 - \hat{\rho}_j\left[s_j + \sigma^2_{\mathrm{KN}}(1 - \frac{n-m}{N})\right] + s_j\sigma^2_{\mathrm{KN}} = 0. \end{cases} \qquad (10)$$

Reference [13] suggested to solve Eq. (10) iteratively starting from an initial guess $\left(\sum_{i=m+1}^{n} s_i\right)/[(n-m)(1-m/N)]$ for σ^2_{KN}, where $s_1 \geqslant s_2 \geqslant \cdots \geqslant s_n$ are the eigenvalues of the sample covariance matrix $\boldsymbol{S}_N = \frac{1}{N}\sum_{t=1}^{N} \boldsymbol{x}_t\boldsymbol{x}_t^{\mathrm{T}}$.

Based on Eq. (8) and the estimate σ^2_{KN} obtained from Eq. (10), the KN hypothesis test is: "$\mathcal{H}_0$: *at least* m *hidden dimensions* ***vs*** $\mathcal{H}_1$: *at most* $m-1$ *hidden dimensions*". If for a chosen confidence level α,

$$s_m > \sigma^2_{\mathrm{KN}}(m)\left[\mu_{N,n-m} + s(\alpha)\sigma_{N,n-m}\right] \qquad (11)$$

is satisfied, then $\mathcal{H}_0$ is accepted and m is increased by one; otherwise, the output is $\hat{m} = m-1$, where $s(\alpha)$ is the corresponding value computed by inversion of the Tracy-Widom distribution. Reference [13] showed that the misidentification probability of the KN estimator converges to the significant level α in the joint limit $n, N \to \infty$, $n/N \to c \geqslant 0$. This paper uses the code from Nadler's web site[2)] to implement the KN algorithm with $\alpha = 0.1\%$.

3.2 Minimax rank estimation (MM)

Based on the existence of a phase transition threshold below which the sample eigenvalues are irrelevant to the population eigenvalues [10–12], a decision-theoretic method for rank estimation was proposed in Ref. [14]. Considering first the problem of differentiating between observing no signal at all ($m=0$ with the population covariance being $\sigma^2\boldsymbol{I}_n$) and observing a single signal ($m=1$ with the largest population eigenvalue being $\lambda_1+\sigma^2$, $\lambda_1>0$). When $m=0$, the asymptotic null distribution of the largest sample eigenvalue s_1 is the Tracy-Widom distribution as given in Eq. (8). When $m=1$, the asymptotic alternate distribution[3)] is

$$\Pr\{s_1 \leqslant \mu_{N,n}(\lambda_1) + x\sigma_{N,n}(\lambda_1)\} \to \Phi(x), \qquad (12)$$

where $\lambda_1 > \sigma^2\sqrt{n/N}$ holds, $\Phi(\cdot)$ denotes the standard normal distribution function, and

$$\begin{aligned} \mu_{N,n}(\lambda) &= (\lambda+\sigma^2)(1+\tfrac{n\sigma^2}{N\lambda}), \\ \sigma_{N,n}(\lambda) &= (\lambda+\sigma^2)\sqrt{\tfrac{2}{\beta N}(1-n\sigma^4/(N\lambda^2))}. \end{aligned}$$

A thresholding decision rule is defined as $\delta_T(s_1)=1$ if $s_1>T$; $\delta_T(s_1)=0$ otherwise. Based on the asymptotic null and alternate distributions, the risk of $\delta_T(s_1)$ is asymptotically as

$$R(\lambda_1, \delta_T) \to \begin{cases} c_{\mathrm{I}}\left(1 - F_\beta\left(\frac{T(j)-\mu_{N,n}}{\sigma_{N,n}}\right)\right), & \text{when } \lambda_1 = 0, \\ c_{\mathrm{E}}\Phi\left(\frac{T(j)-\mu_{N,n}(\lambda_1)}{\sigma_{N,n}(\lambda_1)}\right), & \text{otherwise}, \end{cases}$$

where $c_{\mathrm{I}} > 0$ is an "inclusion" penalty for the false-positive decision, and $c_{\mathrm{E}} > 0$ is an "exclusion" penalty for the false-negative decision, and

$$\begin{aligned} \mu_{N,n} &= \tfrac{\sigma^2}{N}(\sqrt{n}+\sqrt{N})^2, \\ \sigma_{N,n} &= \tfrac{\sigma^2}{N}(\sqrt{n}+\sqrt{N})\left(\tfrac{1}{\sqrt{n}}+\tfrac{1}{\sqrt{N}}\right)^{1/3}. \end{aligned}$$

Supposing $\lambda_1 \geqslant \lambda_0$ is a priori known if $m=1$, the minimax rank estimate algorithm chooses a threshold T to minimize the maximum risk $R(0,\delta_T) \vee R(\lambda_0,\delta_T)$, which implies $R(0,\delta_T) = R(\lambda_0,\delta_T)$, or

$$c_{\mathrm{I}}\left(1 - F_\beta\left(\frac{T-\mu_{N,n}}{\sigma_{N,n}}\right)\right) = c_{\mathrm{E}}\Phi\left(\frac{T-\mu_{N,n}(\lambda_0)}{\sigma_{N,n}(\lambda_0)}\right). \qquad (13)$$

The above thresholding approach is extended for an arbitrary $m>0$ with a sequence $\{c_{\mathrm{E}}(i)\}_{i=1}^{n}$ of exclusion cost, and then the MM thresholding procedure is

$$\hat{m} = \arg\max_{1\leqslant i\leqslant n} i, \ \text{ s.t. } s_j > T(j), \ \forall 1 \leqslant j \leqslant i, \qquad (14)$$

where with the corresponding exclusion cost as $\sum_{\ell=j}^{n} c_{\mathrm{E}}(\ell)$, $T(j)$ is determined by Eq. (13) which implies the MM thresholding algorithm is minimax-optimal [14]. In the implementation, we compute T in Eq. (13) numerically using bisection, set $c_{\mathrm{I}} = c_{\mathrm{E}}(1) = \cdots = c_{\mathrm{E}}(n)$ with $\lambda_0 = \sqrt{n/N}\sigma^2 + N^{-1/3}$ as in Ref. [14], and estimate σ^2 by the maximum likelihood estimator.

3.3 Minka's criterion (MK) for PCA

Bayesian model selection requires computing the marginal likelihood, which involves a difficult integral over a high dimensional parameter space. One way is to approximate the marginal likelihood by Laplace's method, a simplification of which is the BIC approximation [26]. Recently, Minka [7] proposed a new Laplace approximation to the marginal likelihood $p(X_N|m) = \int p(X_N|\boldsymbol{\Theta})p(\boldsymbol{\Theta})\mathrm{d}\boldsymbol{\Theta}$ for FA in Eq. (2), where $\boldsymbol{\Sigma}_y = \boldsymbol{I}_m$ and $\boldsymbol{\Sigma}_e = \sigma_e^2\boldsymbol{I}_n$, and the prior $p(\boldsymbol{\Theta})$ for the parameter $\boldsymbol{\Theta} = \{\boldsymbol{A}, \sigma_e^2\}$ is adopted as

$$p(\boldsymbol{A}, \sigma_e^2) \propto |\boldsymbol{A}\boldsymbol{A}^{\mathrm{T}} + \sigma_e^2\boldsymbol{I}|^{-\frac{\alpha+2}{2}}\mathrm{e}^{-\frac{\alpha}{2}\mathrm{Tr}[(\boldsymbol{A}\boldsymbol{A}^{\mathrm{T}}+\sigma_e^2\boldsymbol{I})^{-1}]},$$

where α is a hyperparameter, and for a noninformative prior, α should be small.

[2)] The KN estimate package is downloaded from http://www.wisdom.weizmann.ac.il/∼nadler/, where the inversion of the Tracy-Widom distribution is approximated by a simple formula (refer to the package for details).

[3)] Please refer to Theorem 2 of Ref. [14] for details.

For a very small α and a reasonably large sample size N, the approximation formula is simplified as

$$\begin{aligned} p(\mathcal{X}_N|m) \approx 2^{-m}\prod\nolimits_{i=1}^{m}\left[\Gamma\left(\frac{n-i+1}{2}\right)\pi^{-\frac{n-i+1}{2}}\right] \\ \cdot\left(\prod\nolimits_{i=1}^{m} s_i\right)^{-\frac{N}{2}}(\hat{\sigma}_e^2)^{-\frac{N(n-m)}{2}}(2\pi)^{\frac{d_m-1}{2}}N^{-\frac{m}{2}} \\ \cdot\left[\prod\nolimits_{i=1}^{m}\prod\nolimits_{j=i+1}^{n}(\hat{s}_j^{-1}-\hat{s}_i^{-1})(s_i-s_j)N\right]^{-\frac{1}{2}}, \end{aligned} \tag{15}$$

where $\hat{s}_i = s_i$ if $i \leqslant m$; otherwise $\hat{s}_i = \hat{\sigma}_e^2$ with $\hat{\sigma}_e^2 = \frac{1}{n-m}\sum_{j=m+1}^{n} s_j$. The hidden dimensionality is estimated as $\hat{m} = \arg\min_m\{-\ln p(\mathcal{X}_N|m)\}$.

3.4 Bayesian Ying-Yang (BYY) harmony learning

First proposed in Ref. [8] and systematically developed over a decade, BYY harmony learning theory is a general statistical learning framework that provides not only a set of new model selection criteria but also a class of automatic model selection algorithms, under a best harmony principle, which is to maximize the following harmony functional:

$$\begin{aligned} H(p\|q) &= \int p(R|X)p(X)\ln[q(X|R)q(R)]\,\mathrm{d}X\mathrm{d}R \\ &= \int p(\boldsymbol{\Theta}|X)H(p\|q,\boldsymbol{\Theta})\,\mathrm{d}\boldsymbol{\Theta}, \qquad (16) \\ H(p\|q,\boldsymbol{\Theta}) &= \int p(Y|X,\boldsymbol{\Theta})p(X)\ln[q(X|Y,\boldsymbol{\Theta}) \\ &\quad \cdot q(Y|\boldsymbol{\Theta})]\,\mathrm{d}Y\mathrm{d}X + \ln q(\boldsymbol{\Theta}), \end{aligned}$$

where the data X is regarded as generated from its inner representation $R = \{Y, \boldsymbol{\Theta}\}$, with Y and $\boldsymbol{\Theta}$ being latent variables and parameters respectively. The two types of Bayesian decompositions, i.e., $p(R|X)p(X)$ and $q(X|R)q(R)$, are called Yang machine and Ying machine respectively. As interpreted in Ref. [30], maximizing $H(p\|q)$ leads to not only a best Ying-Yang matching (may not really reach) $q(X|R)q(R) = p(R|X)p(X)$ which turns $H(p\|q)$ into a negative entropy, but also a least model complexity by further maximizing this negative entropy. The embedded Ying-Yang matching is in a sense of best harmony not maximum likelihood.

We adopt a two-stage iterative procedure given in Ref. [9] to implement BYY harmony learning on FA, as in the third row of Table 1. It needs to be noted that the BYY harmony learning is featured by not only its model selection criterion for FA but also its learning with automatic model selection, that is, determining m automatically during estimating $\boldsymbol{A}$, $\boldsymbol{\Sigma}_y$ and σ_e^2. Usually, further improvements will be achieved if we implement the BYY harmony learning with automatic model selection. Readers are referred to Sect. 3.2 in Ref. [30]. Here we only consider a two-stage implementation for comparing with typical model selection criteria that are all made in a two-stage implementation.

Table 1 BYY harmony learning on FA

	criterion
Ying	$q(\boldsymbol{\Theta}_m) = 1$, $q(X|Y,\boldsymbol{\Theta}_m) = \prod_t G(\boldsymbol{x}_t|\boldsymbol{A}\boldsymbol{y}_t, \sigma_e^2\boldsymbol{I}_n)$, $q(Y|\boldsymbol{\Theta}_m) = \prod_t G(\boldsymbol{y}_t|\boldsymbol{0}, \boldsymbol{\Sigma}_y)$
Yang	$p(Y|X,\boldsymbol{\Theta}_m) = \prod_t G(\boldsymbol{y}_t|\widetilde{\boldsymbol{W}}\boldsymbol{x}_t, \boldsymbol{\Sigma}_{y|x})$, $p(X) = \delta(X - \mathcal{X}_N)$
H	$\hat{m} = \arg\max_m\{H(p\|q, \hat{\boldsymbol{\Theta}}_m^{\mathrm{H}}) - \frac{1}{2}d_m\}$, $\hat{\boldsymbol{\Theta}}_m^{\mathrm{H}} = \arg\max_{\boldsymbol{\Theta}_m} H(p\|q, \boldsymbol{\Theta}_m)$, $H(p\|q,\boldsymbol{\Theta}_m) = \sum_t \ln G(\boldsymbol{x}_t|\boldsymbol{0}, \boldsymbol{\Sigma}_x) - N\ln\sqrt{(2\pi\mathrm{e})^m|\boldsymbol{\Sigma}_{y|x}|}$ $-\frac{1}{2}\mathrm{Tr}[\boldsymbol{\Delta}^{\mathrm{T}}\boldsymbol{\Sigma}_{y|x}^{-1}\boldsymbol{\Delta}\boldsymbol{S}_N]$
where	$\mathcal{X}_N = \{\boldsymbol{x}_t\}_{t=1}^{N}$, $\boldsymbol{\Delta} = \widetilde{\boldsymbol{W}} - \boldsymbol{W}$, $\boldsymbol{W} = \boldsymbol{\Sigma}_y\boldsymbol{A}\boldsymbol{\Sigma}_x^{-1}$, $\boldsymbol{\Sigma}_x = \boldsymbol{A}\boldsymbol{\Sigma}_y\boldsymbol{A}^{\mathrm{T}} + \sigma_e^2\boldsymbol{I}_n$, $\boldsymbol{\Sigma}_{y|x} = \boldsymbol{\Sigma}_y^{-1} + \mathbf{A}^{\mathrm{T}}(\sigma_e^{-2}\boldsymbol{I}_n)\boldsymbol{A}$, $\boldsymbol{A}^{\mathrm{T}}\boldsymbol{A} = \boldsymbol{I}_m$, $\boldsymbol{\Sigma}_y$ is relaxed to be diagonal, $\widetilde{\boldsymbol{W}}$ is free

It should be noted that the criterion in Table 1 is actually equivalent to Eq. (18) in Ref. [30] but written in an alternative expression that shares the same format of Eq. (4). Putting the equations of $\boldsymbol{\Sigma}_x$ and $\boldsymbol{\Sigma}_{y|x}$ in the last row of Table 1 into the third row, it follows that

$$\begin{aligned} \mathcal{J}_{\mathrm{BYY}} &= -H(p\|q,\boldsymbol{\Theta}_m) + \tfrac{1}{2}d_m \\ &= \tfrac{1}{2}\left\{\ln|\sigma_e^2\boldsymbol{I}_n| + \ln|\boldsymbol{\Sigma}_y| + m + d_m + m\ln(2\pi\mathrm{e})\right\}, \end{aligned}$$

by noticing that the third trace-term in the third row of Table 1 eventually tends to zero as

$$p(Y|X,\boldsymbol{\Theta}) \rightarrow \frac{q(X|Y,\boldsymbol{\Theta})q(Y|\boldsymbol{\Theta})}{\int q(X|Y,\boldsymbol{\Theta})q(Y|\boldsymbol{\Theta})\mathrm{d}Y}$$

during learning. This $\mathcal{J}_{\mathrm{BYY}}$ is identical to Eq. (18) in Ref. [30] by letting $h = 0$ without considering data smoothing.

Additionally, it should be also noted that in Table 1 $\boldsymbol{A}$ and $\boldsymbol{\Sigma}_y$ are reparameterized differently from that in Eq. (2). Though the two parameterizations have no difference on the likelihood function, but it results in a better model selection ability than the one by Eq. (2) under BYY. Details are referred to Ref. [31].

Moreover, readers are referred to Sect. 2.2 in Ref. [32] for further improvements via a co-dimensional matrix pair nature for both an improved model selection criterion, e.g., Eq. (29) in Ref. [32], and automatic model selection.

4 Empirical analysis

4.1 A method for empirical study

Most existing experimental investigations merely study how the model selection performance varies as either of SNR and sample size N changes. We propose a new method for empirical analysis in order to systematically compare the model selection performances of all above methods and to demonstrate how SNR (defined as the ratio of the smallest signal eigenvalue of the population covariance matrix to the noise eigenvalue) and N *jointly* affect the performance.

Making experiments on data sets generated by varying SNR and N simultaneously, empirical model selection accuracies measured by each criterion (cri) are visualized as a function of both SNR and N, namely,

$$f_{\text{cri}}(\text{SNR}, N) \in [0, 1], \tag{17}$$

which increases as SNR or N grows. For an effective visualization, we plot contour lines of such functions in 2D figures as shown in Figs. 1(a) and 1(b). Actually, the contour lines define a family of model selection indifference curves, which is a term borrowed from economics. Similarly as in economics, the indifference curve implies a diminishing marginal effect. When we move down (up) an indifference curve, the amount of SNR (or N) is needed to compensate for a unit loss of N (or SNR). According to the values of the contours, the configuration space of SNR and N can be divided into three regions as sketched in Fig. 1(c), namely, a very good performance region for all methods, a very bad performance region for all methods, and a diversity region where different methods

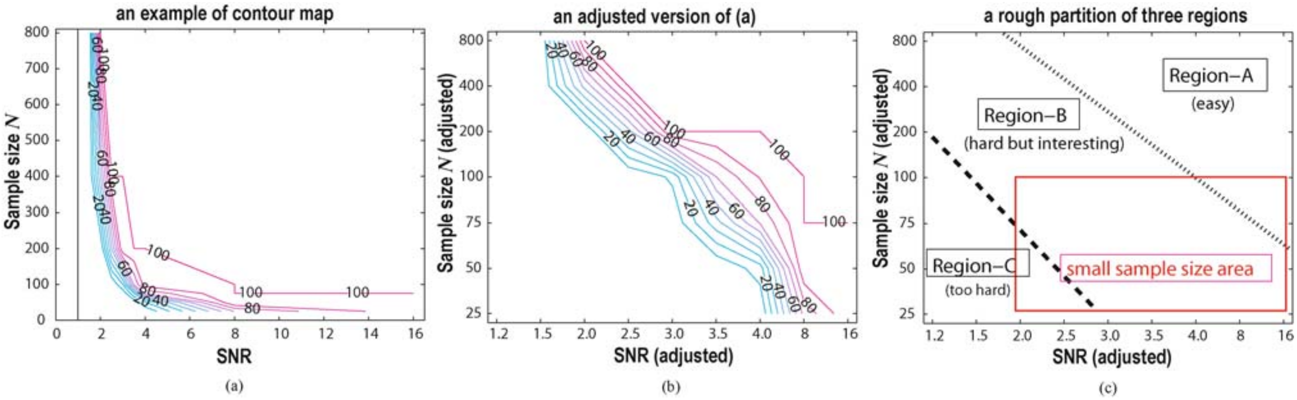


Fig. 1 (a) An example of contour map; (b) an adjusted version of (a), where $1.2, \ldots, 16 \in V_{\gamma_o}$ are equally spaced in horizontal-axis, and $25, \ldots, 800 \in V_N$ are equally spaced in vertical-axis; (c) a rough three-region partition (for Scenario I)

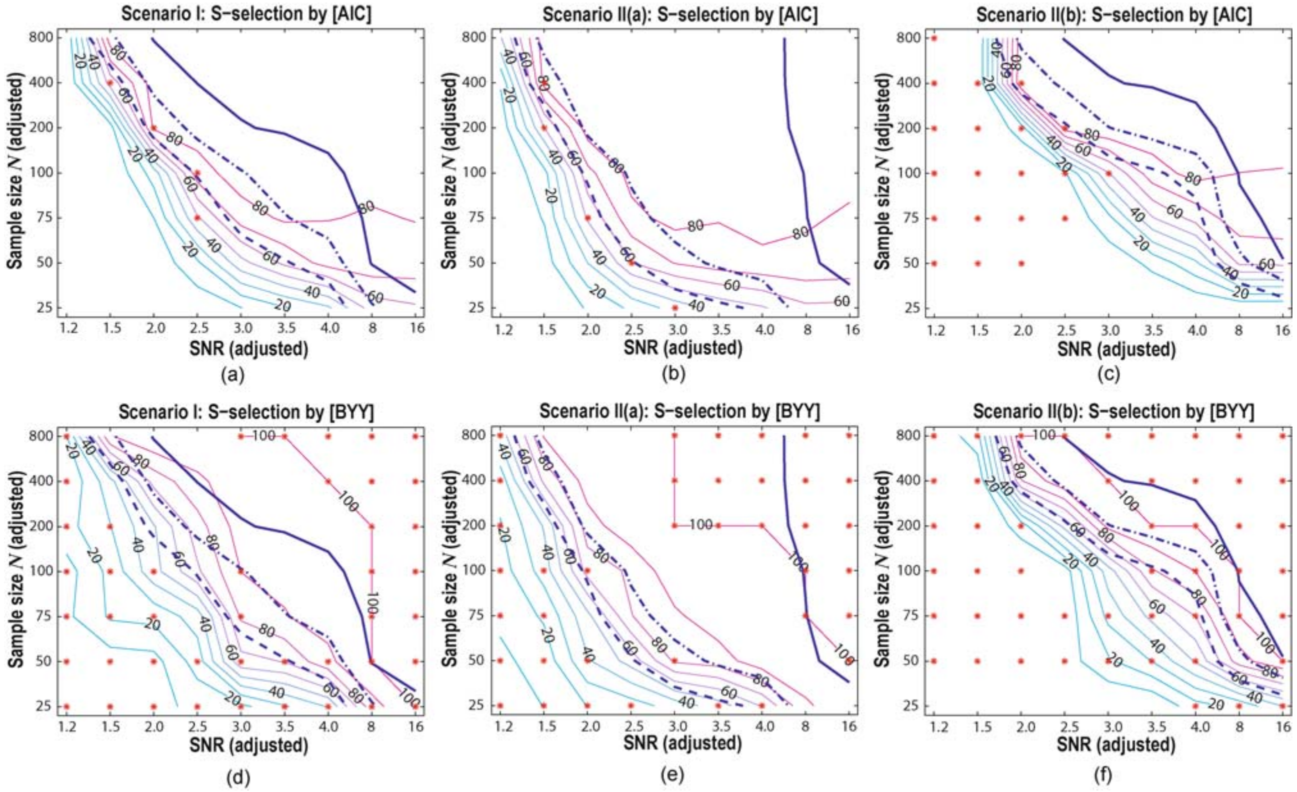


Fig. 2 Adjusted contour maps of successful-selection (S-selection) rates of AIC and BYY over the three scenarios. (We also define an "average" criterion (AvgC) that averages the successful-selection rates among all the model selection methods. Three thick blue contour lines of AvgC are added, i.e., dashed line (30%), dashed-dotted line (60%) and solid line (90%). Moreover, we put a red asterisk (*) in (τ, N, γ_o) to indicate that the corresponding criterion or method gets the highest rate. Among all three scenarios, the 30% and 60% contour lines of AIC and BYY are far closer to the bottom-left than those of AvgC, which means that BYY and AIC are very robust as the experimental conditions deteriorate.)

demonstrate the relative strengths and weaknesses obviously. This three-region partition indicates that it suffices to make a comparison within the diversity region whose importance is further verified by the reference criterion TLL in Eq. (7).

4.2 Investigation on model selection performance

Experiments are conducted on synthetic data which enable us to verify the estimated hidden dimensionality with a known true one, and to study a parameter estimate in comparison with a true one by the testing likelihood. Synthetic data are generated according to the FA model with the population covariance matrix $\boldsymbol{\Sigma}_x = \boldsymbol{U}\boldsymbol{\Lambda}\boldsymbol{U}^{\mathrm{T}} + \sigma_e^2\boldsymbol{I}_n$, where $\boldsymbol{U}^{\mathrm{T}}\boldsymbol{U} = \boldsymbol{I}_{m^*}$, m^* denotes the true underlying hidden dimensionality, $\boldsymbol{\Lambda}$ is a diagonal matrix with $\lambda_1 \geqslant \lambda_2 \geqslant \cdots \geqslant \lambda_{m^*} > 0$ as its diagonal elements, and σ_e^2 is the noise variance as in Eq. (3).

A configuration is featured by a triple (τ, N, γ_o), where τ is the scenario number defined in Table 2, N is the training sample size, and γ_o is the SNR mathematically given by $\gamma_o = \lambda_{m^*}/\sigma_e^2 + 1$. For each configuration in Table 2, 10^3 independent trials are implemented. For every trial, a synthetic data set $\mathcal{X}_N$ is randomly generated according to (τ, N, γ_o). The two-stage procedure is implemented on $\mathcal{X}_N$ for every candidate integer m in $[1, 2m^* - 1]$. The contour maps of the successful model selection rates by Eq. (17) are shown in Figs. 2, 3, and 4 with respect to adjusted axes for a more detailed view.

The model selection indifference curves, given by the contour lines of successful-selection rates, visualize the

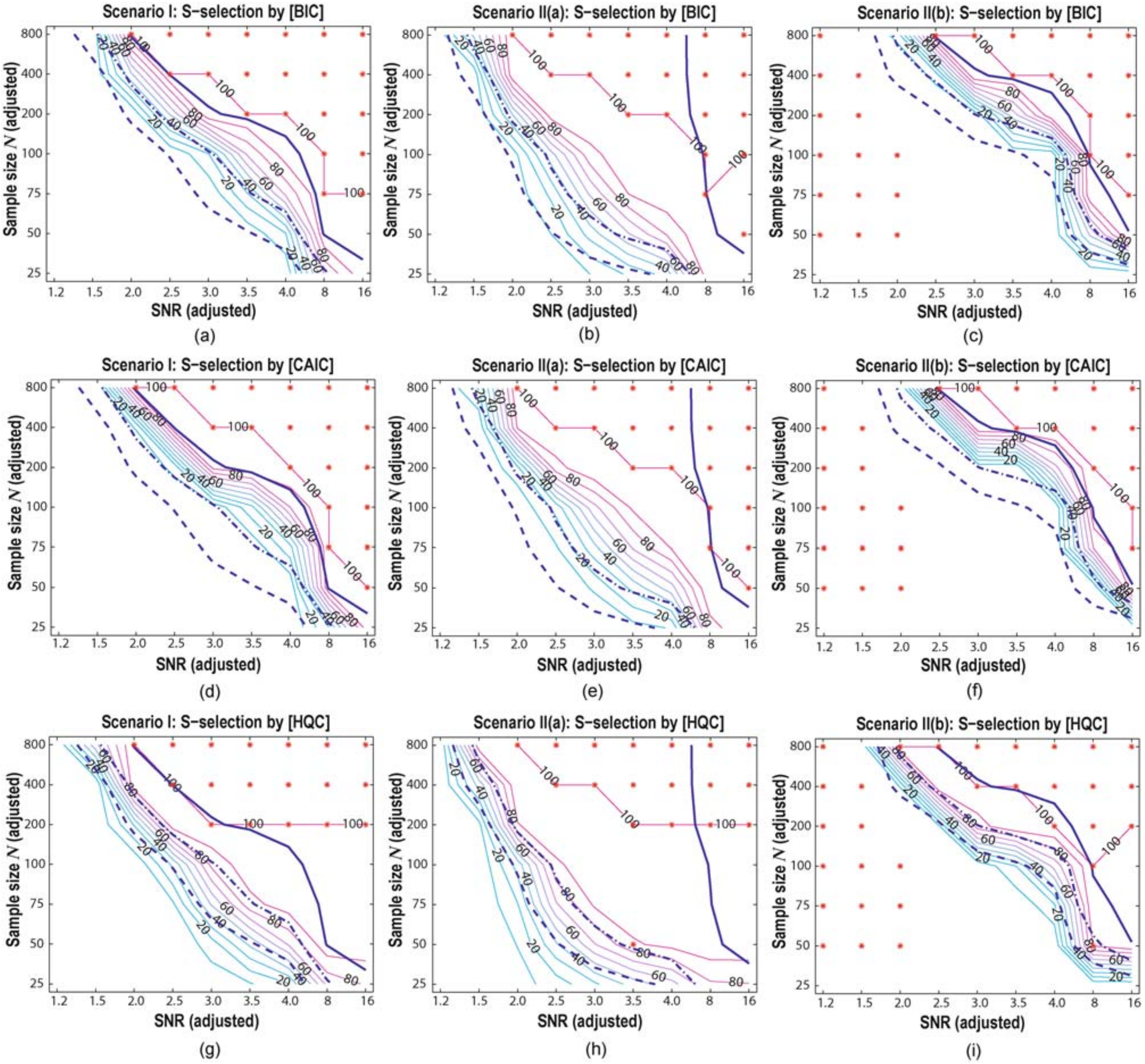


Fig. 3 Adjusted contour maps of successful-selection rates of BIC, CAIC, and HQC across the three scenarios. (Refer to the caption of Fig. 2 for notation details. According to the positions of 30% and 60% contour lines, BIC and CAIC are inferior to AvgC, while HQC slightly outperforms AvgC.)

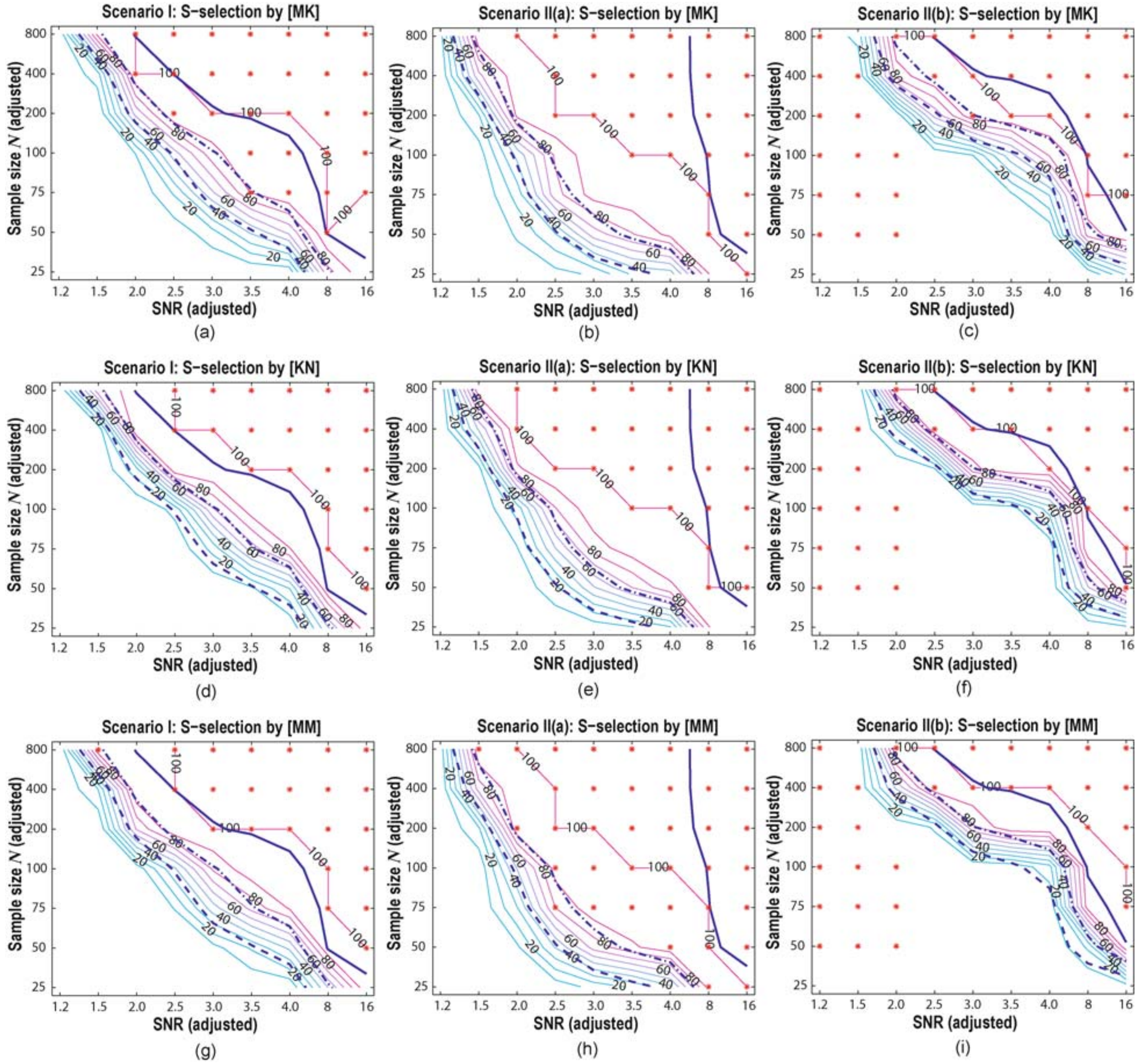
joint effect of SNR and N on the performance. The performance degrades along the bulge toward the lower left of the map with small SNR and N. With the help of the indifference curves, we observe:

Table 2 Configurations of different scenarios (For each scenario, there are $|V_N| \times |V_{\gamma_o}|$ configurations, where $V_N = \{25, 50, 75, 100, 200, 400, 800\}$, $V_{\gamma_o} = \{1.2, 1.5, 2, 2.5, 3, 3.5, 4, 8, 16\}$, and $\boldsymbol{U}$ is randomly generated and normalized to be $\boldsymbol{U}^{\mathrm{T}}\boldsymbol{U} = \boldsymbol{I}_m$, and $\lambda_i \sim [1, 10]$ means λ_i is uniformly drawn from the interval $[1, 10]$ so that the signal eigenvalues can vary a little but not much.)

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

1) The negative slope of the indifference curves as in Fig. 1(a) implies that as N (or γ_o) becomes larger, the model selection accuracy increases at a decreasing rate, i.e., the additions to the successful-selections are successively smaller. Thus, this confirms the *diminishing marginal effect* introduced in Sect. 4.1. It is a joint effect of SNR and training sample size on model selection performance.

2) From Fig. 1(b), the configuration space (SNR × N) can be partitioned into *three regions* as sketched in Fig. 1(c), i.e., *region-A* for the cases of relatively large SNR and N, *region-B* for the moderate cases, and *region-C* for the cases of very small SNR and N. The successful-selection rates of most criteria are comparably high in *region-A*, and degenerate to a very low or even zero in *region-C* due to severe underestimation which may attribute to the reduction of effective number of underlying



(a) (b) (c)

(d) (e) (f)

(g) (h) (i)

Fig. 4 Adjusted contour maps of successful-selection rates of MK, KN, and MM across the three scenarios. (Refer to the caption of Fig. 2 for notations details. According to the positions of 30% and 60% contour lines, MK and MM are slightly better than AvgC, while KN is comparable to AvgC.)

signals [10–12]. This reduction is also demonstrated in Figs. 5(a)–5(d) from the perspective of the testing log-likelihood[4]. $\hat{m}$(TLL) by Eq. (7) to select the optimal ML estimate is significantly smaller than the true one in *region-C*. According to Figs. 2 to 5, we observe that in *region-B*, BYY and AIC are more robust against decreasing N and SNR, while in *region-A* most methods are comparably good except AIC. As a whole, BYY is superior in general; DNLL is almost always the worst.

3) For each scenario at *region-B,C*, AIC, HQC, KN, MK, MM, and BYY are relatively more robust than the others. BYY is generally superior in the small-sample-size area as sketched in Fig. 1(c), while KN, MK, and MM are robust for a small SNR with a sample size not too small. Across scenarios, the performances of all methods (except DNLL) improves as we move from Scenario I to II(a) with stronger signals, and decrease from Scenario I to II(b) with higher dimensionalities.

In addition, the contour maps are averaged along one axis and then projected along the other, helping to explore the marginalized effect of either of SNR and N. The results in Fig. 6 show that BYY, MK, MM, and KN are better than HQC, BIC, and CAIC. AIC is robust when N and SNR are small but suffers from about 20% wrong selection as N and SNR become large. Moreover, the projected N-axis demonstrates the diversity of the performances more obviously than the projected SNR-axis does.

5 Concluding remarks

The relative strengths and weaknesses of several model selection methods have been investigated systematically

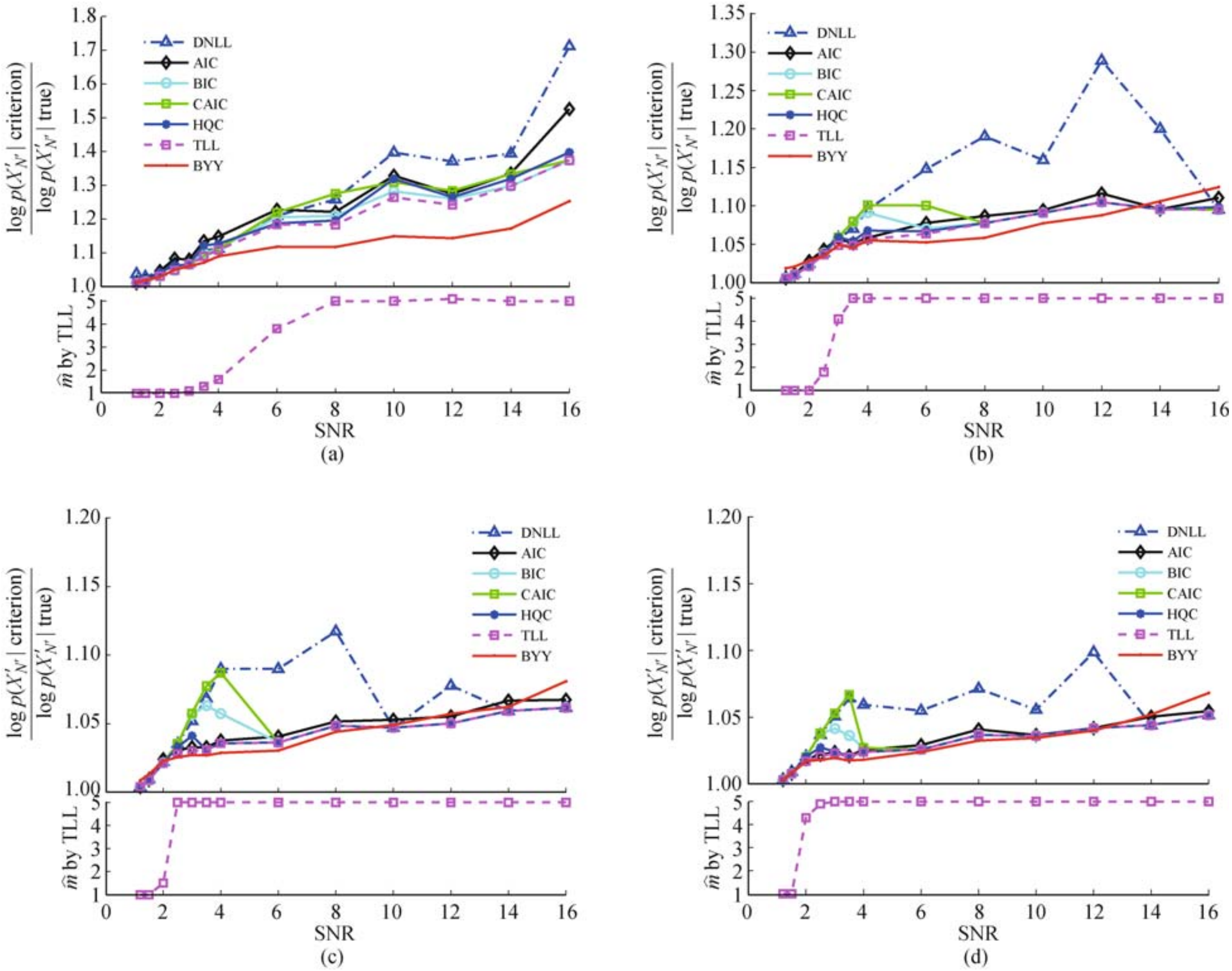


Fig. 5 (a) Ratio and $\hat{m}$(TLL), when $N = 25$; (b) ratio and $\hat{m}$(TLL), $N = 50$; (c) ratio and $\hat{m}$(TLL), $N = 75$; (d) ratio and $\hat{m}$(TLL), $N = 100$ (The vertical axis in the upper part presents the ratio $\ln p(X'_{N'}|\hat{\boldsymbol{\Theta}})/\ln p(X'_{N'}|\boldsymbol{\Theta}_o)$ of the estimated parameter $\hat{\boldsymbol{\Theta}}$ to that of the true parameter $\boldsymbol{\Theta}_o$ of Scenario I obtained from a synthetic testing set $X'_{N'}$ (with $N' = 10^4$), where $X'_{N'}$ is generated from the same FA that generates the training set. The $\hat{\boldsymbol{\Theta}}$ is repeatedly estimated on 100 training sets from the same FA with the true parameter $\boldsymbol{\Theta}_o$, and the mean of the resulting ratios is reported. In experiments, the candidate scales are in $\{1, 2, \ldots, 9\}$, and true scale $m^* = 5$, and $\ln p(X'_{N'}|\hat{\boldsymbol{\Theta}}) \leqslant \ln p(X'_{N'}|\boldsymbol{\Theta}_o) < 0$ which implies a ratio closer to 1.0 indicates a smaller generalization error.)

[4] The MK, MM, and KN are omitted because they do not estimate the parameters directly, while the ML estimate $\hat{\boldsymbol{\Theta}}^{\mathrm{ML}}_{\hat{m}}$ can be used to get the testing log-likelihood instead with $\hat{m}$ estimated by MK, MM, or KN.

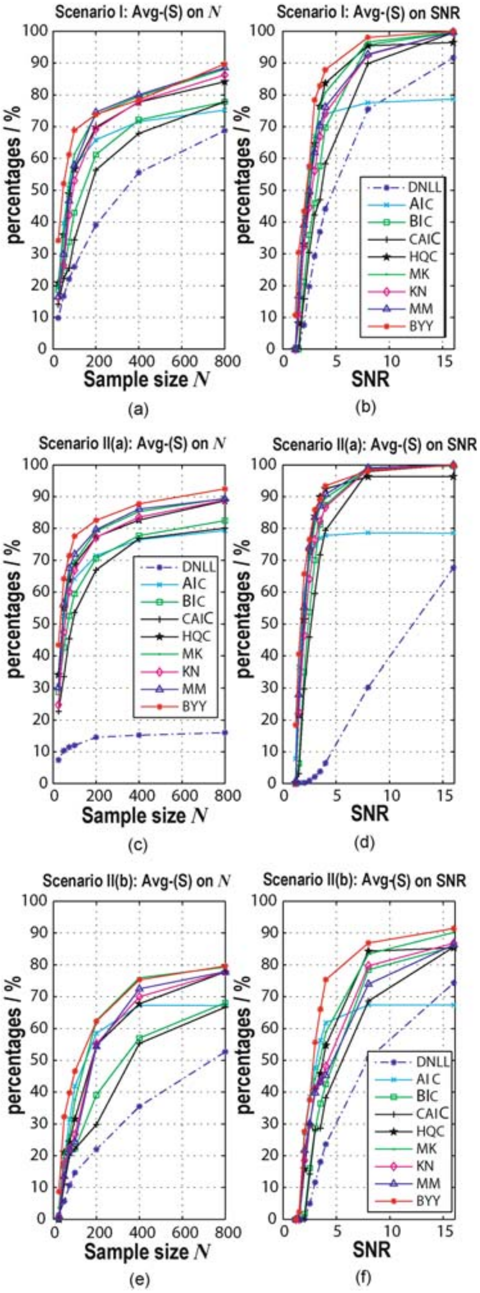


Fig. 6 Curves of average successful-selection (Avg-(S)) rates by averaging the contour maps along one axis and then projecting along the other

in terms of determining the hidden dimensionality of FA. Different from the existing empirical analysis, we have studied the joint effect of training sample size N and SNR on model selection performance. By connecting the contour of equal model selection accuracies, the obtained model selection indifference curves visually reveal the following:

1) A diminishing marginal effect that the amount of SNR (or N) is needed to compensate for the unit loss of N (or SNR) if the same model selection accuracy is maintained;

2) A three-region partition in the configuration space, i.e., all methods perform well within the range of large N and SNR, but unavoidably suffer from underestimation within the range of too small N and SNR due to a significant reduction of effective number of signals, whereas the performances of all methods demonstrate obviously different within the region with moderate N and SNR.

In addition, the comparison has been made in terms of testing likelihoods with an alternative reference criterion TLL (testing log-likelihood) which selects the optimal ML estimate. TLL further confirms the importance of studying the region of diversity. In this region, AIC and BYY are more robust against decreasing N and SNR, while BYY's estimate is recommended according to its robust model selection performance.

Acknowledgements The work described in this paper was fully supported by a grant from the Research Grant Council of the Hong Kong SAR (No. CUHK4177/07E).

References

1. Wax M, Kailath T. Detection of signals by information theoretic criteria. IEEE Transactions on Acoustics, Speech and Signal Processing, 1985: 33(2): 387–392
2. Schmidt R. Multiple emitter location and signal parameter estimation. IEEE Transactions on Antennas and Propagation, 1986, 34(3): 276–280
3. Tu S, Xu L. Theoretical analysis and comparison of several criteria on linear model dimension reduction. In: Proceedings of the 8th International Conference on Independent Component Analysis and Signal Separation. 2009, 154–162
4. Anderson T, Rubin H. Statistical inference in factor analysis. In: Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability. 1956, 5: 111–150
5. Tipping M E, Bishop C M. Mixtures of probabilistic principal component analyzers. Neural Computation, 1999, 11(2): 443–482
6. Akaike H. A new look at the statistical model identification. IEEE Transactions on Automatic Control, 1974, 19(6): 716–723
7. Minka T P. Automatic choice of dimensionality for PCA. Advances in Neural Information Processing Systems, 2001, 13: 598–604
8. Xu L. Bayesian-Kullback coupled Ying-Yang machines: unified learnings and new results on vector quantization. In: Proceedings of International Conference on Neural Information Processing. 1995, 977–988
9. Xu L. Bayesian Ying Yang learning. Scholarpedia, 2007, 2(3): 1809
10. Baik J, Silverstein J W. Eigenvalues of large sample covariance matrices of spiked population models. Journal of Multivariate Analysis, 2006, 97(6): 1382–1408
11. Johnstone I M. High dimensional statistical inference and random matrices. In: Proceedings of International Congress of Mathematicians. 2006, 1–28
12. Paul D. Asymptotics of sample eigenstruture for a large dimensional spiked covariance model. Statistica Sinica, 2007, 17(4): 1617–1642

13. Kritchman S, Nadler B. Determining the number of components in a factor model from limited noisy data. Chemometrics & Intelligent Laboratory Systems, 2008, 94(1): 19–32

14. Perry P O, Wolfe P J. Minimax rank estimation for subspace tracking. Selected Topics in Signal Proceesing, 2010, 4(3): 504–513

15. Hu X, Xu L. A comparative investigation on subspace dimension determination. Neural Networks, 2004, 17(8–9): 1051–1059

16. Chen P, Wu T J, Yang J. A comparative study of model selection criteria for the number of signals. IET Radar, Sonar and Navigation, 2008, 2(3): 180–188

17. Zhang Q T, Wong K, Yip P, Reilly J. Statistical analysis of the performance of information theoretic criteria in the detection of the number of signals in array processing. IEEE Transactions on Acoustics, Speech and Signal Processing, 1989, 37(10): 1557–1567

18. Xu W, Kaveh M. Analysis of the performance and sensitivity of eigendecomposition-based detectors. IEEE Transactions on Signal Processing, 1995, 43(6): 1413–1426

19. Liavas A, Regalia P. On the behavior of information theoretic criteria for model order selection. IEEE Transactions on Signal Processing, 2001, 49(8): 1689–1695

20. Fishler E, Grosmann M, Messer H. Detection of signals by information theoretic criteria: general asymptotic performance analysis. IEEE Transactions on Signal Processing, 2002, 50(5): 1027–1036

21. Fishler E, Poor H. Estimation of the number of sources in unbalanced arrays via information theoretic criteria. IEEE Transactions on Signal Processing, 2005, 53(9): 3543–3553

22. Nadakuditi R, Edelman A. Sample eigenvalue based detection of high-dimensional signals in white noise using relatively few samples. IEEE Transactions on Signal Processing, 2008, 56(7): 2625–2638

23. Rissanen J. Modelling by the shortest data description. Automatica, 1978, 14(5): 465–471

24. Hoyle D C. Automatic PCA dimension selection for high dimensional data and small sample sizes. Journal of Machine Learning Research, 2008, 9(12): 2733–2759

25. Bishop C M. Variational principal components. In: Proceedings of the Ninth International Conference on Artificial Neural Networks. 1999, 1: 509–514

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

27. Bozdogan H. Model selection and Akaike's Information Criterion (AIC): the general theory and its analytical extensions. Psychometrika, 1987, 52(3): 345–370

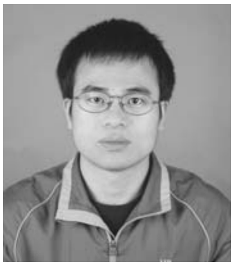
28. Hannan E, Quinn B. The determination of the order of an autoregression. Journal of the Royal Statistical Society. Series B, 1979, 41(2): 190–195

29. Johnstone I M. On the distribution of the largest eigenvalue in principal component anslysis. Annals of Statistics, 2001, 29(2): 295–327

30. Xu L. Bayesian Ying-Yang system, best harmony learning, and five action circling. Frontiers of Electrical and Electronic Engineering in China, 2010, 5(3): 281–328

31. Tu S, Xu L. Parameterizations make different model selections: empirical findings from factor analysis. Frontiers of Electrical and Electronic Engineering in China (in Press)

32. Xu L. Codimensional matrix pairing perspective of BYY harmony learning: hierarchy of bilinear systems, joint decomposition of data-covariance, and applications of network biology. Frontiers of Electrical and Electronic Engineering in China, 2011, 6(1): 86–119



Shikui TU is a Ph.D candidate of the Department of Computer Science and Engineering, The Chinese University of Hong Kong. He obtained his Bachelor degree from School of Mathematical Science, Peking University, in 2006. His research interests include statistical learning, pattern recognition, and bioinformatics.

Lei XU, chair professor of The Chinese University of Hong Kong (CUHK), Fellow of IEEE (2001–), Fellow of International Association for Pattern Recognition (2002–), and Academician of European Academy of Sciences (2002–). He completed his Ph.D thesis at Tsinghua University by the end of 1986, became postdoc at Peking University in 1987, then promoted to associate professor in 1988 and a professor in 1992. During 1989–1993 he was research associate and postdoc in Finland, Canada and USA, including Harvard and MIT. He joined CUHK as senior lecturer in 1993, professor in 1996, and chair professor in 2002. He published several well-cited papers on neural networks, statistical learning, and pattern recognition, e.g., his papers got over 3400 citations (SCI) and over 6300 citations by Google Scholar (GS), with the top-10 papers scored over 2100 (SCI) and 4100 (GS). One paper scored 790 (SCI) and 1351 (GS). He served as a past governor of International Neural Network Society (INNS), a past president of APNNA, and a member of Fellow Committee of IEEE CI Society. He received several national and international academic awards (e.g., 1993 National Nature Science Award, 1995 INNS Leadership Award and 2006 APNNA Outstanding Achievement Award).