Theoretical Analysis and Comparison of Several Criteria on Linear Model Dimension Reduction

Shikui Tu and Lei Xu

Department of Computer Science and Engineering,
The Chinese University of Hong Kong, Hong Kong, P.R. China
{sktu,lxu}@cse.cuhk.edu.hk

Abstract. Detecting the dimension of the latent subspace of a linear model, such as Factor Analysis, is a well-known model selection problem. The common approach is a two-phase implementation with the help of an information criterion. Aiming at a theoretical analysis and comparison of different criteria, we formulate a tool to obtain an order of their approximate underestimation-tendencies, i.e., AIC, BIC/MDL, CAIC, BYY-FA(a), from weak to strong under mild conditions, by studying a key statistic and a crucial but unknown indicator set. We also find that DNLL favors cases with slightly dispersed signal and noise eigenvalues. Simulations agree with the theoretical results, and also indicate the advantage of BYY-FA(b) in the cases of small sample size and large noise.

1 Introduction

Linear model is one of the most common modeling approaches to multivariate data in many scientific fields. Factor Analysis (FA)[1] is a such widely-used linear model that assumes the observations come from a linear mixture of some latent Gaussian factors with additive Gaussian noise. It is usually used for dimension reduction via detecting the hidden structures. Also, as recently revisited in [2], PCA is equivalent to a special case of FA [1] under the Maximum Likelihood (ML) principle. FA is extended to Independent Component Analysis (ICA)[3] by requiring higher order independence, no noise and square mixing matrix.

One of the fundamental tasks in FA modeling is determining the dimension of the latent subspace, i.e., the number of hidden factors. It is a model selection problem in machine learning. Also, it is addressed as the problem of detecting the number of signals through a noisy channel in signal processing [4,5,6,7,8]. One conventional approach is hypothesis tests based on the likelihood ratio statistic [9] and a subjective threshold. Another approach is the two-phase implementation that requires no subjective threshold with the help of an information criterion such as Akaike's Information Criterion (AIC)[10], Bozdogan's Consistent Akaike's Information Criterion (CAIC)[11], Schwarz's Bayesian Information Criterion (BIC)[12] (which coincides with Rissanen's Minimum Description Length (MDL)[13]), and Bayesian Ying-Yang (BYY) harmony learning criterion[14].

Following an early work [4] in signal processing literature, a framework was proposed in [5] for studying criteria such as AIC and MDL, with asymptotic

T. Adali et al. (Eds.): ICA 2009, LNCS 5441, pp. 154–162, 2009.

bounds provided for overestimation and underestimation probabilities, which was further studied in [6,7]. Recently, the behaviors of AIC and MDL in a situation with high dimensional signals but relatively few samples were investigated in [8]. In this track [4,5,6,7,8], FA is considered in its special case of PCA, and the studies are focused on asymptotic properties, such as consistency and asymptotic normality, and the results were shown to be robust for non-Gaussian sources empirically[7]. However, in practical the sample size is finite or even small, and it is intractable to get an exact selection accuracies of different criteria. An easier way is to study their relative selection tendencies for a preliminary comparison.

This paper formulates a tool further developed from[5] for a theoretical comparison of typical criteria in terms of ordered approximate underestimation tendencies. It suffices to study a key statistic and an indicator set which is inherently associated with each criterion and depends on the distribution of samples. The order from weak to strong is shown to be AIC,BIC,CAIC and BYY-FA(a) under mild conditions, while DNLL is found to favor the cases with slightly dispersed signal and noise eigenvalues. Though analytically hard, BYY-FA(b) is shown to be empirically superior for those small-sample-size and large-noise cases.

The rest of the paper is organized as follows. In Section 2, we briefly review FA and several criteria. In Section 3, we formulate a tool for comparisons of different criteria via studying a key statistic and a crucial indicator set, and then conduct simulations in Section 4. The conclusion is made in Section 5.

2 Factor Analysis and Serval Model Selection Criteria

Factor Analysis. Assume $\mathbf{x}$ is an observed n-dimensional random variable, and it is distributed according to the following descriptions:

$$\begin{array}{l} \mathbf{x} = \mathbf{A}\mathbf{y} + \boldsymbol{\mu} + \mathbf{e},\ p(\mathbf{x}|\mathbf{y}) = G(\mathbf{x}|\mathbf{A}\mathbf{y} + \boldsymbol{\mu}, \boldsymbol{\Sigma}_e), p(\mathbf{y}) = G(\mathbf{y}|\mathbf{0}, \boldsymbol{\Sigma}_y), \\ \begin{cases} \boldsymbol{\Theta}_m = \{\mathbf{A}, \boldsymbol{\Sigma}_e\} \text{ if } \boldsymbol{\Sigma}_y = \mathbf{I}_m \text{ (the } m \times m \text{ identity matrix)}, & \textit{for } \mathbf{FA(a)}; \\ \boldsymbol{\Theta}_m = \{\mathbf{A}, \boldsymbol{\Lambda}_m, \boldsymbol{\Sigma}_e\} \text{ if } \boldsymbol{\Sigma}_y = \boldsymbol{\Lambda}_m \text{ (diagonal) and } \mathbf{A}^T\mathbf{A} = \mathbf{I}_m, & \textit{for } \mathbf{FA(b)}; \end{cases} \\ p(\mathbf{x}) = \int p(\mathbf{x}|\mathbf{y})p(\mathbf{y})d\mathbf{y} = G(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma}_x),\ \boldsymbol{\Sigma}_x = \mathbf{A}\mathbf{A}^T + \boldsymbol{\Sigma}_e \end{array} \quad (1)$$

where $\mathbf{y}$ is an $m \times 1$ hidden factor, $\boldsymbol{\Theta}_m$ is the unknown parameter set including an $n \times m$ factor loading matrix $\mathbf{A}$ and a diagonal noise covariance matrix $\boldsymbol{\Sigma}_e$, and $G(\bullet|\boldsymbol{\mu}, \boldsymbol{\Sigma})$ denotes a Gaussian distribution with the mean vector $\boldsymbol{\mu}$ and the covariance matrix $\boldsymbol{\Sigma}$. The two formulations, i.e., FA(a) and FA(b), are equivalent under the Maximum Likelihood principle for parameter learning, but they are different under the BYY harmony learning [14] for selecting m which will be introduced in Section 3.3&4.1. In the sequel, we assume $\boldsymbol{\mu} = \mathbf{0}, \boldsymbol{\Sigma}_e = \sigma_e^2\mathbf{I}_n$.

Several Criteria and Two-phase Implementation. The task of FA modeling consists of parameter learning and selecting m, based on a sample set $\mathcal{X}_N = \{\mathbf{x}_t\}_{t=1}^N$, and it is tackled by the following two-phase implementation:

- **Phase I:** Compute $\hat{\boldsymbol{\Theta}}_m = \hat{\boldsymbol{\Theta}}(\mathcal{X}_N, m)$ for each $m \in [m_{low}, m_{up}]$ with m_{low} and m_{up} given. Normally, $\hat{\boldsymbol{\Theta}}_m$ is the Maximum Likelihood (ML) estimator $\hat{\boldsymbol{\Theta}}_m^{ML} = \arg\max_{\Theta_m} \ln p(\mathcal{X}_N|\boldsymbol{\Theta}_m) = \arg\min_{\Theta_m} \mathcal{E}_L(\mathcal{X}_N|\boldsymbol{\Theta}_m)$, where $\mathcal{E}_L(\mathcal{X}_N|\boldsymbol{\Theta}_m) = -\frac{2}{N} \ln p(\mathcal{X}_N|\boldsymbol{\Theta}_m)$ is denoted as **NLL**(negative log-likelihood).

– **Phase II:** Estimate $\hat{m} = \arg\min_m \mathcal{E}_{Cri}(\mathcal{X}_N, \hat{\boldsymbol{\Theta}}_m)$, where $\mathcal{E}_{Cri}$ is formulated according to a criterion (Cri), e.g.,

$$\mathcal{E}_{Cri}(\mathcal{X}_N, \hat{\boldsymbol{\Theta}}_m) = \mathcal{E}_L(\mathcal{X}_N, \hat{\boldsymbol{\Theta}}_m) + \rho_{cri} d_m,$$
$$d_m = nm+1, \quad \rho_{cri} = \begin{cases} \rho_L = 0; & \text{for } \mathbf{NLL} \\ \rho_{AIC} = \frac{2}{N}; & \text{for } \mathbf{AIC} \\ \rho_{BIC} = \frac{\ln N}{N}; & \text{for } \mathbf{BIC} \\ \rho_{CIC} = \frac{\ln N+1}{N}; & \text{for } \mathbf{CAIC} \end{cases} \tag{2}$$

3 Theoretical Analysis and Comparisons

3.1 A Tool for Comparisons

Based on Sec.2&[5], this subsection further formulate a tool aiming at analysis and comparisons of different criteria for FA modeling, and provides a summarized guidance for the detailed analysis in the subsequent subsections.

– **select m via discrete optimization.** Consider $S(\boldsymbol{\Theta}_m, m)$ to be a family of statistical models $p(\mathbf{x}|\boldsymbol{\Theta}_m)$ for FA(a) given in eq.(1) with $\boldsymbol{\Theta}_m = \{A_{n\times m}, \sigma_e^2\}$. Given a criterion(Cri) with $\mathcal{E}_{Cri} = \mathcal{E}_{Cri}(\mathcal{X}_N, \hat{\boldsymbol{\Theta}}(\mathcal{X}_N, m)) = \mathcal{E}_{Cri}(\mathcal{X}_N, m)$, an estimator of m^*(the underlying true dimension) is given by $\hat{m}(\mathcal{X}_N) = \arg\min_m \mathcal{E}_{Cri}$. To locate the minima w.r.t. discrete m, no derivative can be used. However, it is reasonable to study instead the backward difference function $\nabla_m \mathcal{E}_{Cri} = \mathcal{E}_{Cri}(\mathcal{X}_N, m) - \mathcal{E}_{Cri}(\mathcal{X}_N, m-1)$, as shown in Fig.1(a)(b).
– **from $\nabla_m \mathcal{E}_{Cri}$ to local preference.** It is intractable to study $\nabla_m \mathcal{E}_{Cri}$ as a function of $\mathcal{X}_N, m$. Fortunately, $\nabla_m \mathcal{E}_{Cri}$ from several criteria for FA can be formulated as $\nabla_m \mathcal{E}_{Cri}(\gamma_m, m)$(Fig.1(b)(c)), a function of m and a statistic γ_m given in eq.(8), which will be shown in Sec.3.2&3.3. The medium γ_m extracts and transmits sufficient information from samples to selecting m, and also determines the *local preference* over each $\{m-1, m\}$ as in Fig.1(d). Γ_m^* and its element γ_m^* are separately termed **indicator set** and **indicator** at m. Note that γ_m is closely related to the signal-to-noise ratio.
– **approximate underestimation tendency.** *Underestimation* refers to an event "$\hat{m} < m^*$". Considering the *Local preference* defined in Fig.1(d) over $\{m^*-1, m^*\}$, if $\gamma_{m^*} \in \Gamma_{m^*}^+$, then m^*-1 is preferred to m^*, which indicates that "$\hat{m} < m^*$" is *likely* to happen (though not guaranteed). Therefore, it is reasonable to approximate the underestimation tendency by the probability $Pr\{\gamma_{m^*} \in \Gamma_{m^*}^+\}$. Its exact evaluation is intractable for a finite or small N, but the relative tendencies of different criteria can be determined as follows.
– **A TOOL for comparisons.** Fixing $m = m^*$, assume $\nabla_m \mathcal{E}_{Cri_1}(\gamma_m)$ and $\nabla_m \mathcal{E}_{Cri_2}(\gamma_m)$, sketched in Fig.1(c), are strictly monotone decreasing in domain Γ_D with their indicators satisfying $\gamma_m^*(\text{Cri}_1) < \gamma_m^*(\text{Cri}_2)$. Actually, these assumptions hold for several criteria as in Sec.3.2. Then, $\Gamma_m^+(\text{Cri}_i) = (-\infty, \gamma_m^*(\text{Cri}_i)) \bigcap \Gamma_D$, $i=1,2$, and $Pr\{\gamma_m \in \Gamma_m^+(\text{Cri}_2)\} - Pr\{\gamma_m \in \Gamma_m^+(\text{Cri}_1)\} = Pr\{\gamma_m^*(\text{Cri}_1) < \gamma_m < \gamma_m^*(\text{Cri}_2)\} \geq 0$. So, "*approximately* the underestimation tendency of Cri_2 is stronger than that of Cri_1" or $Cri_1 \prec_u Cri_2$. Similar analysis on overestimation can be performed at $m = m^* + 1$.

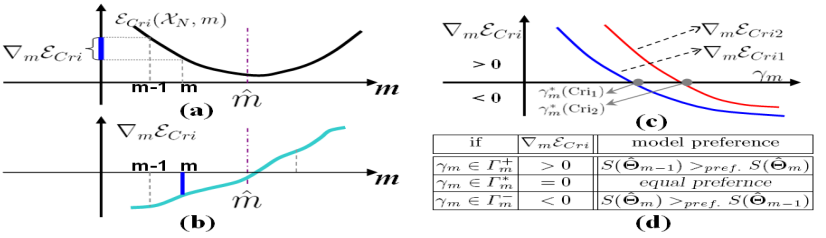

Fig. 1. For a given $\mathcal{X}_N$, graphs of $\mathcal{E}_{Cri}$ and $\nabla_m\mathcal{E}_{Cri}$ w.r.t. m are sketched in (a)&(b), while for two criteria, Cri_1 and Cri_2, the graphs of $\nabla_m\mathcal{E}_{Cri_1}, \nabla_m\mathcal{E}_{Cri_2}$ w.r.t. γ_m given m are sketched in (c), as well as its corresponding local preference defined in (d)

3.2 AIC, BIC, CAIC

Assume the eigenvalues of the sample covariance matrix, i.e., $S_N = \frac{1}{N}\sum_{t=1}^{N}\mathbf{x}_t\mathbf{x}_t^T$, are $\{s_i : 1 \le i \le n\}$ with $s_1 \ge \ldots \ge s_n$. The Maximum Likelihood (ML) estimate $\hat{\mathbf{\Theta}}_m^{ML}$ for FA(a) in eq.(1) is given to be ([1,4,2]):

$$\begin{cases} \hat{\mathbf{A}}_{n\times m}^{ML} = \mathbf{U}_{n\times m}(\mathbf{D}_m - \hat{\sigma}_e^2)^{\frac{1}{2}}\mathbf{R}^T, \ \ \mathbf{D}_m = diag[s_1, \ldots, s_m], \\ \hat{\sigma}_e^{2,ML} = \frac{1}{n-m}\sum_{i=m+1}^{n} s_i, \end{cases} \tag{3}$$

where the i-th column of $\mathbf{U}_{n\times m}$ is the eigenvector of S_N corresponding to s_i, and $\mathbf{R}$ is an arbitrary rotation matrix. According to eq.(2) and eq.(3), the NLL and the difference functions of some criteria, are further formulated as:

$$\mathcal{E}_L(\mathcal{X}_N, \hat{\mathbf{\Theta}}_m^{ML}) = k\ln\sum_{i=m+1}^{n} s_i - k\ln k - \sum_{i=m+1}^{n}\ln s_i, \ \ k = n-m, \tag{4}$$

$$\nabla_m\mathcal{E}_L(\gamma_m, m) \doteq \nabla_m\mathcal{E}_L(\mathcal{X}_N, \hat{\mathbf{\Theta}}_m^{ML}) = -(k+1)\ln\left(1 + \frac{\gamma_m - 1}{k+1}\right) + \ln\gamma_m \tag{5}$$

$$\frac{\partial\nabla_m\mathcal{E}_L(\gamma_m, m)}{\partial\gamma_m} = -\frac{k(\gamma_m - 1)}{(k+\gamma_m)\gamma_m} \le 0, \ \forall\gamma_m \in [1, +\infty). \tag{6}$$

$$\nabla_m\mathcal{E}_{Cri}(\gamma_m, m) \doteq \nabla_m\mathcal{E}_{Cri}(\mathcal{X}_N, \hat{\mathbf{\Theta}}_m^{ML}) = \nabla_m\mathcal{E}_L(\mathcal{X}_N, \hat{\mathbf{\Theta}}_m^{ML}) + n\rho_{cri} \tag{7}$$

where ρ_{cri} is given in eq.(2), and γ_m is explicitly formulated by

$$\gamma_m = \gamma_{m,m}, \ \ \gamma_{i,m} = s_i/\mathcal{A}_{m+1}^n \ge 1, i = 1, \ldots, m; \quad \mathcal{A}_m^n = \frac{1}{n-m+1}\sum_{i=m}^{n} s_i, \tag{8}$$

Due to the space limit, all theoretical results are given without proofs.

Lemma 1. *(1). Given $\rho > 0$, the root γ^* of $\nabla_m\mathcal{E}_L(\gamma) = -n\rho$ 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 $C_0 = \exp\{\frac{n\rho}{k}\}$, $k = n-m$. (2). For $\rho_1 > \rho_2 > 0$, we have $\gamma^*(\rho_1) > \gamma^*(\rho_2) > 1$.*

Remarks: Similar bounds were provided in[5,6] by two kinds of Taylor approximations w.r.t. two formulated variables separately, while Lemma 1(1) was derived by a second-order Taylor approximation (as in[6]) w.r.t to γ (as in[5]).

Theorem 1. *Since the indicator $\gamma_m^*(Cri)$ is a root of $\nabla_m \mathcal{E}_{Cri}(\gamma_m) = 0$, then*

1. $1 = \gamma_m^*(\text{NLL}) < \gamma_m^*(\text{AIC}) < \gamma_m^*(\text{BIC}) < \gamma_m^*(\text{CAIC})$, *if* $N \geq 8 > e^2$.
2. $\Gamma_m^+ = [1, \gamma_m^*(Cri))$, $\Gamma_m^- = (\gamma_m^*(Cri), +\infty)$, *and indicator set* $\Gamma_m^* = \{\gamma_m^*(Cri)\}$.
3. *Applying* $C_0 \approx 1 + \frac{n\rho}{k}$ *to* γ_{up} *in Lemma 1, we get a further approximation:*

$$\gamma_m^*(Cri) \approx 1 + (n + n/k) \cdot \frac{c}{N} + \frac{n}{k}\sqrt{2(k+1)\left(\frac{k}{n} + \frac{c}{N}\right)\frac{c}{N}} + O(\frac{c}{N}), \quad (9)$$

where $c = 2, \ln N, \ln N + 1$ *for* ***AIC***, ***BIC***, ***CAIC*** *separately, and* $k = n - m$.

Remarks: This theorem indicates: (1). NLL tends to select large m in probability one unless $\gamma_m = 1, \forall m > m^*$ or $s_i = \sigma^2$ $(\forall i \geq m^*)$ which requires $N \to +\infty$; (2). Fixing $m = m^*$, "AIC$\prec_u$BIC$\prec_u$CAIC" holds according to Sec.3.1.

3.3 DNLL and BYY-FA(a)

The likelihood-ratio test is a conventional approach to model selection in statistics [9]. The logarithm of the likelihood-ratio or the difference of the Negative-Log-likelihood (NLL) is denoted as **DNLL**, and the corresponding objective function is $\mathcal{E}_{\text{DNLL}}(\mathcal{X}_N, \hat{\Theta}_m^{ML}) = \nabla_m \mathcal{E}_L$, where $\nabla_m \mathcal{E}_L$ is given in eq.(5). Then,

$$\begin{aligned}\nabla_m(\mathcal{E}_{\text{DNLL}}(\mathcal{X}_N, \hat{\Theta}_m^{ML})) = \nabla_m^2 \mathcal{E}_L = -2(k+1)\ln\left(1 + \tfrac{\gamma_{m,m}-1}{k+1}\right) \\ +(k+2)\ln\left(1 + \tfrac{\gamma_{m-1,m}+\gamma_{m,m}-2}{k+2}\right) - \ln\tfrac{\gamma_{m-1,m}}{\gamma_{m,m}}\end{aligned} \quad (10)$$

where $\gamma_{m-1,m}, \gamma_{m,m}$ are formulated in eq.(8), and $\gamma_{m-1,m} \geq \gamma_{m,m} \geq 1$. According to the Formulation 1, $\gamma(\mathcal{X}_N, m)$ is generalized to a two-variable vector $(\gamma_{m-1,m}, \gamma_{m,m})$, and the indicator set Γ_m^* becomes a 2-dimensional boundary.

Theorem 2. *Define $s_p, \ldots, s_q$ to be "slightly dispersed", if $|s_i - \mathcal{A}_p^q| < \delta$ holds for any $i \in [p, q]$ and a very small $\delta(> 0)$. The criterion* DNLL *captures the variations of* NLL*, and especially at the unknown true dimension m^* we have*

1. When $m = m^*$: *If $s_{m-1} \approx s_m \gg \mathcal{A}_{m+1}^n$, then $\gamma_{m-1,m} \approx \gamma_{m,m} \gg 1$, which implies $\nabla_m \mathcal{E}_{DNLL} < 0$, i.e., m^* is preferred to $m^* - 1$.*
2. When $m - 1 = m^*$: *If $s_{m-1} \gg s_m \approx \mathcal{A}_{m+1}^n$, then $\gamma_{m-1,m} \gg \gamma_{m,m} \approx 1$, which implies $\nabla_m \mathcal{E}_{DNLL} > 0$, i.e., m^* is preferred to $m^* + 1$.*
3. *If $s_1, \ldots, s_{m^*}$ are slightly dispersed, $s_{m^*+1}, \ldots, s_n$ are also slightly dispersed, and $s_{m^*} \gg s_{m^*+1}$, then m^* is the global minimum of $\mathcal{E}_{DNLL}$.*

Remarks: Instead of strict mathematical formulations, the conditions in Theorem 2 are stated in an intuitive way. It implies DNLL favors slightly dispersed signal and noise eigenvalues, as well as a large signal-to-noise ratio (SNR). However, the conditions will be probably violated when N and SNR is small.

Another approach to tackling model selection problems is the Bayesian Ying-Yang (BYY) harmony learning theory[14]. We defer its detailed introduction to

the next section. With γ_m again formulated in eq.(8), a BYY criterion (denoted as **BYY-FA(a)**) for FA(a) in eq.(1) as well as its difference function is

$$\begin{aligned} &\mathcal{E}_H^a(\mathcal{X}_N, \hat{\mathbf{\Theta}}_m^{ML}) = m\ln(2\pi e) + n\ln\left(\tfrac{1}{n-m}\textstyle\sum_{i=m+1}^{n} s_i\right), \\ &\nabla_m \mathcal{E}_H^a(\gamma_m, m) = \ln(2\pi e) - n\left\{\ln\left(1+\tfrac{\gamma_m}{n-m}\right) + \ln\left(1+\tfrac{1}{n-m}\right)\right\}, \end{aligned} \tag{11}$$

Lemma 2. *According to eq.(11) and the tool defined in Sec.3.1, we have*

1. *Since the indicator $\gamma_m^*(H_a)$ is the root of $\nabla_m \mathcal{E}_H^a(\gamma_m) = 0$, then $\gamma_m^*(H_a) = (n-m+1)\left[(2\pi e)^{\frac{1}{n}} - 1\right] + 1 > 1$, e.g., $\gamma_m^*(H_a) \approx 3.595$ when $n = 9, m = 3$.*
2. *$\Gamma_m^+ = [1, \gamma_m^*(H_a))$, $\Gamma_m^- = (\gamma_m^*(H_a), +\infty)$, the indicator set $\Gamma_m^* = \{\gamma_m^*(H_a)\}$.*

Theorem 3. *There exists an equivalent ρ_{H_a} for* BYY-FA(a)*, and then we indirectly compare the indicator $\gamma_m^*(H_a)$ of* BYY-FA(a) *with $\gamma_m^*(Cri)$ of another criterion* (Cri) *by approximately comparing ρ_{H_a} with ρ_{cri} as follows:*

1. *ρ_{H_a} is bounded in $\left(\rho_{H_a}^{(low)}, \rho_{H_a}^{(up)}\right)$, where $\rho_{H_2}^{(up)} = \frac{n-m}{n}\ln c_n$ and $\rho_{H_a}^{(low)} = c_n + \frac{2c_n-1}{k-1} - \frac{\sqrt{2(k+1)c_n(c_n-1)+1}}{k-1}$, $c_n = \sqrt[n]{(2\pi e)}$, $k = n-m$.*
2. *Given n, m, $\exists N_{cri} > 1$ such that $\rho_{H_a} < \rho_{cri}$ iff $1 < N < N_{cri}$. Also, N_{cri} is lower bounded by N_{up}, which is the largest N that satisfies $\rho_{H_a}^{(up)} < \rho_{cri}$. E.g., $N_{up} = 14, 23, 31$ for* AIC*,* BIC*, and* CAIC *respectively, when $n = 9, m = 5$.*

Remarks: Consider $m = m^*$. (1). Since $\gamma_{m^*}^*(H_a)$ is irrelevant to N and γ_{m^*} is the ML estimator for the true unknown SNR $\gamma_o = \lambda_{m^*}/\sigma^2$, then BYY-FA(a) tends to underestimate m regardless of N as long as $\gamma_{m^*}^*(H_a) > \gamma_o$. (2). We compare BYY-FA(a) with other criteria (Cri), such as AIC, BIC and CAIC, directly by calculating each indicator $\gamma_{m^*}^*(\text{Cri})$ as in Lemma 1&2 or indirectly in form of ρ_{cri} as in Theorem 3. (3). There exists a small N_{cri}, such that if $N < N_{cri}$ then BYY-FA(a)$\prec_u$Cri, otherwise Cri$\prec_u$BYY-FA(a), according to Sec.3.1.

4 Empirical Study and BYY-FA(b)

4.1 BYY-FA(b)

The criteria analyzed above are relatively easy for a theoretical analysis, while BYY Harmony Learning Theory on another formulation of FA, i.e., FA(b) in eq.(1), is difficult. However, via an empirical comparison, we still provide insights of its model selection performances.

Firstly proposed in 1995 and systematically developed in the past decade, Bayesian Ying-Yang (BYY) harmony learning theory is a general statistical learning framework that can handle both parameter learning and model selection under a best harmony principle. The BYY harmony learning leads us not only a set of new model selection criteria for typical structures, but also a class of automatic model selection algorithms. For more details, please refer to a recent systematic review[14].

FA(a) and FA(b) in eq.(1) are equivalent under ML principle but different under the BYY harmony learning theory [14]. The former leads to BYY-FA(a) in Sec.3.3, while the latter leads to BYY-FA(b) as follows, with a similar two-phase procedure (see eq.(7) in[14]) implemented,

$$\mathcal{E}_H^b = m\ln(2\pi e) + \ln|\mathbf{\Lambda}| + n\ln\sigma_e^2 + h^2 Tr\left[(\mathbf{A}\mathbf{\Lambda}\mathbf{A}^T + \sigma_e^2\mathbf{I}_n)^{-1}\right]. \quad (12)$$

4.2 Simulations

We design $3 \times 3 = 9$ cases of experimental environments by considering three levels of sample size N and noise σ_e^2 respectively, with $n = 9$ and $m^* = 3$ fixed. Three levels are $100, 50, 25$ for N or $0.1\lambda_{m^*}, 0.3\lambda_{m^*}, 0.5\lambda_{m^*}$ for σ_e^2 (equivalently $\gamma_o = \lambda_{m^*}/\sigma_e^2 = 10, 3.33, 2$), where λ_{m^*} is the m^*-th largest Gaussian signal's variance. We randomly generate samples according to each setting of FA in eq.(1) for each of 100 independent repeated runs, in which two-phase procedure is implemented by setting $[m_{low}, m_{up}] = [1, 6]$ and randomly initializing $\mathbf{\Theta}_m$. The selection percentage rates are reported in Table 1. The indicators $\gamma_{m^*}^*(Cri)$ are approximately calculated by eq.(9), and $\gamma_{m^*}^*(H_a)$ by Lemma 2.

The simulations suggest the following observations. (1). The performances of all criteria are comparable when N,γ_o are large, but they decline at different speeds as N, γ_o reduce. (2). For a large $N(= 100)$, BIC and CAIC is consistent but AIC risks an overestimation. Let Cri be AIC, BIC or CAIC, and then $\gamma_{m^*}^*(Cri)$ grows as N reduces. When $\gamma_{m^*}^*(Cri)$ exceeds γ_o, Cri tends

Table 1. We report the percentage rates of model selection of 9 combinations in three categories, i.e., *underestimation*(**U**),*successful selection*(**S**) and *overestimation*(**O**). The indicator $\gamma_m^*(Cri)$ is calculated at $m = m^* = 3$. Note that $\gamma_m^*(Cri)$ by eq.(9) approximates γ_m^{num} well, where γ_m^{num} is the numerical solution of $\nabla_m \mathcal{E}_{Cri}(\gamma) = 0$.

(a). Sample size $N = \mathbf{100}$, $\boldsymbol{\gamma_o = \lambda_{m^*}/\sigma_e^2}$ (3 levels)

noise level:	$\gamma_o = \mathbf{10}$			$\gamma_o = \mathbf{3.33}$			$\gamma_o = \mathbf{2}$			γ_m^*(Cri) approximated	γ_m^{num} is the
Cri \ *rates*	U	S	O	U	S	O	U	S	O	by eq.(9).	numerical sol.
AIC	0	99	1	0	96	4	0	**97**	3	$\gamma_3^*(\text{AIC}) \approx 1.87$	$\gamma_3^{num}(\text{AIC}) = 1.83$
BIC	0	100	0	1	99	4	9	91	0	$\gamma_3^*(\text{BIC}) \approx 2.50$	$\gamma_3^{num}(\text{BIC}) = 2.43$
CAIC	0	100	0	1	99	4	22	78	0	$\gamma_3^*(\text{CAIC}) \approx 2.72$	$\gamma_3^{num}(\text{CAIC}) = 2.65$
DNLL	2	98	0	39	61	0	63	27	0	*not available*	*not available*
BYY-FA(a)	0	100	0	30	70	0	98	2	0	$\gamma_3^*(H_a) = 3.59$	$\gamma_3^*(H_a) = 3.59$
BYY-FA(b)	0	100	0	1	99	0	1	95	4	*not available*	*not available*

(b). Sample size $N = \mathbf{50}$, $\boldsymbol{\gamma_o = \lambda_{m^*}/\sigma_e^2}$ (3 levels)

(same as (a))	U	S	O	U	S	O	U	S	O	γ_m^*(Cri) by eq.(9)	γ_3^{num}(Cri)
AIC	0	98	2	0	**99**	1	18	79	3	$\gamma_3^*(\text{AIC}) \approx 2.36$	$\gamma_3^{num}(\text{AIC}) = 2.30$
BIC	0	100	0	6	94	2	76	24	0	$\gamma_3^*(\text{BIC}) \approx 3.18$	$\gamma_3^{num}(\text{BIC}) = 3.10$
CAIC	0	100	0	20	80	2	91	9	0	$\gamma_3^*(\text{CAIC}) \approx 3.57$	$\gamma_3^{num}(\text{CAIC}) = 3.50$
DNLL	5	95	0	49	51	0	86	14	0	*not available*	*not available*
BYY-FA(a)	0	100	0	35	65	2	96	4	0	$\gamma_3^*(H_a) = 3.59$	$\gamma_3^*(H_a) = 3.59$
BYY-FA(b)	0	100	0	3	97	0	5	**84**	11	*not available*	*not available*

(b). Sample size $N = \mathbf{25}$, $\boldsymbol{\gamma_o = \lambda_{m^*}/\sigma_e^2}$ (3 levels)

(same as (a))	U	S	O	U	S	O	U	S	O	γ_m^*(Cri) by eq.(9)	γ_3^{num}(Cri)
AIC	1	92	7	13	85	2	58	40	2	$\gamma_3^*(\text{AIC}) \approx 3.21$	$\gamma_3^{num}(\text{AIC}) = 3.13$
BIC	1	99	0	49	51	0	94	6	0	$\gamma_3^*(\text{BIC}) \approx 4.15$	$\gamma_3^{num}(\text{BIC}) = 4.10$
CAIC	1	99	0	84	16	0	100	0	0	$\gamma_3^*(\text{CAIC}) \approx 4.88$	$\gamma_3^{num}(\text{CAIC}) = 4.91$
DNLL	11	89	0	62	38	0	89	11	0	*not available*	*not available*
BYY-FA(a)	1	92	7	35	63	2	85	12	3	$\gamma_3^*(H_a) = 3.59$	$\gamma_3^*(H_a) = 3.59$
BYY-FA(b)	0	99	1	6	**89**	5	21	**66**	13	*not available*	*not available*

to underestimates m, where AIC remains more robust. These agree with Theorem 1. (3). DNLL fails as N, γ_o reduce, which agrees with Theorem 2. (4). BYY-FA(a) tends to underestimate m when $\gamma_o < \gamma^*_{m^*}(H_b)$, which is worse than BIC and CAIC for $N = 100$ but better for $N = 25$. This coincides with Theorem 3. (5). BYY-FA(b) becomes evidently superior when $N \leq 50$ and $\gamma_o \leq 3.33$. For example, it improves by $4.7\%, 6.3\%, 65\%$ relative to AIC when $(N, \gamma_o) = (25, 3.33), (50, 2), (25, 2)$ respectively.

5 Conclusion

We have provided a preliminary theoretical comparison of several criteria based on the problem of selecting the hidden dimension of FA in its special case of PCA. It suffices to study a statistic and a crucial but unknown indicator set for each criterion. Due to the difficulty in exact evaluation of selection accuracy for a finite or small sample size N, the model selection behavior is preliminarily characterized by an order of the approximate underestimation tendencies, i.e., AIC$\prec_u$BIC$\prec_u$CAIC$\prec_u$BYY-FA(a) . DNLL requires a proper dispersion of signal and noise eigenvalues. The simulations agree with the theoretical results and also indicates that BYY-FA(b) becomes superior as N reduces and noise increases.

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

References

1. Anderson, T., Rubin, H.: Statistical inference in factor analysis. In: Proceedings of the third Berkeley symposium on mathematical statistics and probability, vol. 5, pp. 111–150 (1956)
2. Tipping, M.E., Bishop, C.M.: Mixtures of probabilistic principal component analyzers. Neural Computation 11(2), 443–482 (1999)
3. Hyvärinen, A.: Survey on Independent Component Analysis. Neural Computing Surveys 2, 94–128 (1999)
4. Wax, M., Kailath, T.: Detection of signals by information theoretic criteria. IEEE Trans. Acoustics, Speech and Signal Processing ASSP-33(2), 387 (1985)
5. Xu, W., Kaveh, M.: Analysis of the performance and sensitivity of eigendecomposition-based detectors. IEEE Transactions on Signal Processing 43(6), 1413–1426 (1995)
6. Liavas, A., Regalia, P.: On the behavior of information theoretic criteria for model order selection. IEEE Transactions on Signal Processing 49(8), 1689–1695 (2001)
7. Fishler, E., Poor, H.: Estimation of the number of sources in unbalanced arrays via information theoretic criteria. IEEE Transactions on Signal Processing 53(9), 3543–3553 (2005)
8. Nadakuditi, R., Edelman, A.: Sample eigenvalue based detection of high-dimensional signals in white noise using relatively few samples. IEEE Transactions on Signal Processing 56(7), 2625–2638 (2008)
9. Cox, D., Hinkley, D.: Theoretical Statistics. Chapman and Hall, London (1974)

10. Akaike, H.: A new look at the statistical model identification. IEEE Transactions on Automatic Control 19(6), 716–723 (1974)
11. Bozdogan, H.: Model selection and Akaike's Information Criterion (AIC): The general theory and its analytical extensions. Psychometrika 52(3), 345–370 (1987)
12. Schwarz, G.: Estimating the Dimension of a Model. The Annual of Statistics 6(2), 461–464 (1978)
13. Rissanen, J.: Modelling by the shortest data description. Automatica 14, 465–471 (1978)
14. Xu, L.: Bayesian ying yang system, best harmony learning, and gaussian manifold based family. In: Zurada, J.M., Yen, G.G., Wang, J. (eds.) Computational Intelligence: Research Frontiers. LNCS, vol. 5050, pp. 48–78. Springer, Heidelberg (2008)