2010 IEEE International Conference on Acoustics, Speech, and Signal Processing ICASSP 2010

March 14-19, 2010 • Dallas, Texas, U.S.A.

IEEE Catalog Number: CFP10ICA-CDR
ISBN: 978-1-4244-4296-6 ISSN: 1520-6149

A STUDY OF SEVERAL MODEL SELECTION CRITERIA FOR DETERMINING THE NUMBER OF SIGNALS

*Shikui Tu and Lei Xu**

Department of Computer Science and Engineering,
The Chinese University of Hong Kong, Shatin, N.T., Hong Kong. {sktu,lxu}@cse.cuhk.edu.hk

ABSTRACT

Addressing the problem of detecting the number of source signals as selecting the hidden dimensionality of Factor Analysis (FA) model, we investigate several model selection criteria via a new empirical analyzing tool that examines the joint effect of signal-noise ratio (SNR) and sample size N on the model selection performance. The contours of the model selection accuracies visualize a three-region partition on the space of SNR and N, and a diminishing marginal effect which trades off SNR and N on the performance. Moreover, the newly derived Variational Bayes algorithm and three variants of Bayesian Ying-Yang (BYY) algorithms are more robust against reducing SNR and N, where the BYY with priors' hyperparameters updated is the best in general.

Index Terms— Number of signals, hidden dimensionality, linear model, model selection, criteria

1. INTRODUCTION

It is an essential issue to detect the number of underlying source signals 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., [1]). The observed vector can be modeled as a superposition of a finite number of underlying source signals with an additive noise. The source signals and noise vector sequence are assumed to be two independent ergodic zero-mean Gaussian random processes. Moreover, this issue can also be addressed as a model selection problem of selecting the hidden dimensionality of Factor Analysis (FA) [2] in its special case of Principal Component Analysis (PCA) [3] under the Maximum Likelihood principle.

A classical approach to model selection is the two-stage procedure, i.e., parameter learning is made on a set of candidate models, among which one is selected by a model selection criterion. The existing criteria include Akaike's Information Criterion (AIC)[4], Bozdogan's Consistent Akaike's Information Criterion (CAIC)[5], Hannan-Quinn information criterion (HQC) [6], Schwarz's Bayesian Information Criterion (BIC)[7] (which coincides with Rissanen's Minimum Description Length (MDL)[8]), and recently developed Minka's criterion (MK)[9], Variational Bayes (VB) [10], Bayesian Ying-Yang (BYY) harmony learning criterion[11].

Early from [1], AIC and MDL were introduced to determine the number of signals, and then it was followed by a lot of researches such as [12], with emphasis on the asymptotic properties of the criteria under certain assumptions. Following the above track, this paper aims at a systematic investigation on those criteria via a new empirical analyzing tool that examines the joint effect of the signal-noise ratio (SNR) and sample size N rather than the effect of either SNR or N with the other fixed in previous work, e.g., [13].

The adopted parameterization of FA is different from the common one in e.g., [10]. The two forms are equivalent in Maximum Likelihood learning, but different in model selection as pointed out in [14] under BYY. Actually, the adopted form results in a better model selection ability under BYY and VB with details in another working paper[15]. Here, we implement the EM algorithm [3] for classical AIC etc. For VB, we derive a new VBEM algorithm by imposing appropriate priors on the unknown parameters. In addition to the existing prior-free version (BYYo)[16], BYY is further implemented by not only adopting the same priors (BYYp) as the VBEM but also updating the hyperparameters of the priors (BYYph) under the Hessian based second-order information conservation principle [16].

By varying a wide range of N and SNR in the empirical analysis, we connect the contour of the same model selection accuracy, and the contours actually define a family of model selection performance indifference curves (a term borrowed from economics) for each criterion. Then, we are able to reveal a diminishing marginal effect that the amount of SNR (or N) to trade for a unit of N (or SNR) increases if the performance is kept unchanged, and also able to present a three-region partition on the space of N and SNR, i.e., all methods perform well/bad when SNR and N are too large/small respectively, while within the region with moderate SNR and N, the performances of these methods demonstrate diversity clearly. Moreover, VB and three variants of BYY outperform

* Corresponding author: Lei Xu. Email: lxu@cse.cuhk.edu.hk. 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).

the others in the region of diversity, while BYYph is the best in general.

The rest of the 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 introduces the two-stage procedure with several model selection criteria whose behaviors are empirically analyzed in section 4 followed by the concluding remarks in section 5.

2. PROBLEM FORMULATION

In signal processing [1], a common model for the received complex-valued signal vector $\mathbf{x}(t)$ from an array of n sensors at time instance t, is $\mathbf{x}(t) = \mathbf{A}\mathbf{s}(t) + \mathbf{e}(t)$, where $\mathbf{A}$ is the steering matrix with full column rank. The m-dimensional source signal vector sequence $\{\mathbf{s}(t)\}$ is assumed to be a stationary and ergodic Gaussian random process with zero mean and positive definite covariance matrix $\mathbf{\Sigma}_s$. The noise sequence $\{\mathbf{e}(t)\}$ is assumed to be a stationary and ergodic Gaussian vector process, independent of the source signals, with zero mean and isotropic covariance matrix $\sigma_e^2\mathbf{I}_n$, where $\mathbf{I}_n$ is the $n \times n$ unit matrix. Determining the number of source signals based on an observed sequence $\{\mathbf{x}(t)\}_{t=1}^N$ is to estimate the rank of $\mathbf{A}\mathbf{\Sigma}_s\mathbf{A}^H$ in $\mathbf{\Sigma}_{x|c} = \mathbf{A}\mathbf{\Sigma}_s\mathbf{A}^H + \sigma_e^2\mathbf{I}_n$ where $\Sigma_{x|c}$ 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 [16, 3] and statistics [2], assumes an observed real-valued n-dimensional random variable $\mathbf{x}$ as follows:

$$\left\{\begin{array}{l} \mathbf{x} = \mathbf{U}\mathbf{y} + \boldsymbol{\mu} + \mathbf{e}, \quad \mathbf{\Theta}_m = \{\mathbf{U}, \boldsymbol{\mu}, \mathbf{\Lambda}, \mathbf{\Sigma}_e\}; \\ q(\mathbf{x}|\mathbf{y}) = G(\mathbf{x}|\mathbf{U}\mathbf{y} + \boldsymbol{\mu}, \mathbf{\Sigma}_e), q(\mathbf{y}) = G(\mathbf{y}|\mathbf{0}, \mathbf{\Lambda}), \\ q(\mathbf{x}|\mathbf{\Theta}_m) = \int p(\mathbf{x}|\mathbf{y})p(\mathbf{y})d\mathbf{y} = G(\mathbf{x}|\boldsymbol{\mu}, \mathbf{\Sigma}_x), \\ \mathbf{\Lambda} = diag[\lambda_1, \ldots, \lambda_m], \mathbf{\Sigma}_x = \mathbf{U}\mathbf{\Lambda}\mathbf{U}^T + \mathbf{\Sigma}_e, \end{array}\right. \quad (1)$$

where $\mathbf{U}^T\mathbf{U} = \mathbf{I}_m$, and $\mathbf{\Sigma}_x$ is the population matrix of the data and $\mathbf{\Sigma}_e = \sigma_e^2\mathbf{I}_n$ which makes FA equivalent to PCA under the maximum likelihood principle [3]. Estimating the hidden dimensionality m is to determine the rank of $\mathbf{U}\mathbf{\Lambda}\mathbf{U}$ in $\mathbf{\Sigma}_x = \mathbf{U}\mathbf{\Lambda}\mathbf{U}^T + \sigma_e^2\mathbf{I}_n$ based on $\{\mathbf{x}_t\}_{t=1}^N$.

The two rank estimation problems are equivalent in a sense that they aim to estimate the (same) rank m in respectively two similar sample covariance equations. Next, we focus on the latter one, which is also widely used as a dimensionality reduction technique for feature extraction.

3. MODEL SELECTION CRITERIA

The two-stage procedure performs parameter learning over a set of candidate models among which one is selected by a model selection criterion. Typical examples include the classical AIC [4], BIC/MDL [7, 8], CAIC [5], HQC[6], and recent Minka's criterion (MK) [9] and VB [10], and BYY [16, 17], as well as the difference of negative log-likelihood (DNLL). They are briefly summarized in Tab.1.

criteria	Stage-I	Stage-II: $\hat{m} = \arg\min_m J(m)$
DNLL	EM alg.	$J(m) = -\mathcal{L}(\hat{\mathbf{\Theta}}_m^{ML}) + \mathcal{L}(\hat{\mathbf{\Theta}}_{m-1}^{ML})$
AIC		$J(m) = -\mathcal{L}(\hat{\mathbf{\Theta}}_m^{ML}) + d_m$
BIC		$J(m) = -\mathcal{L}(\hat{\mathbf{\Theta}}_m^{ML}) + \frac{d_m}{2}\ln N$
CAIC		$J(m) = -\mathcal{L}(\hat{\mathbf{\Theta}}_m^{ML}) + \frac{d_m}{2}(\ln N + 1)$
HQC		$J(m) = -\mathcal{L}(\hat{\mathbf{\Theta}}_m^{ML}) + d_m \ln(\ln N)$
MK	eig	$J(m)$ by equation (30) in [9].
VB	VBEM	$J(m) = -\mathcal{F}(\hat{p}_{\mathbf{U}}, \hat{p}_\nu, \hat{p}_\phi, \hat{p}_Y, m)$
BYY	eq.(2)	$J(m) = -H(p\|q, \mathbf{\Theta}_m^*, \Xi^*) + \frac{1}{2}d_m$

Table 1. The two-stage procedures for several criteria are given, where $\mathcal{L}(\hat{\mathbf{\Theta}}_m^{ML}) = \max_{\mathbf{\Theta}_m} \ln q(X_N|\mathbf{\Theta}_m)$ is the maximized log-likelihood of data set X_N, and $d_m = nm + 1 - \frac{m(m-1)}{2}$ is the number of free parameters in FA. In Stage-I, the "EM alg." denotes the Expectation Maximization (EM) algorithm for FA; the "eig" means estimating the sample eigenvalues for MK; the $\mathcal{F}(\hat{p}_{\mathbf{U}}, \hat{p}_\nu, \hat{p}_\phi, \hat{p}_Y, m)$ is the resulted variational lower bound by VBEM; the $H(p\|q, \mathbf{\Theta}_m^*, \Xi^*)$ is the resulted harmony functional by implementing eq.(2).

One difficulty in Bayesian model selection is to compute the marginal likelihood which incorporates priors on the parameters and involves a high dimensional integration. To approximate the marginal likelihood, Minka [9] proposed a criterion (MK) via Laplace approximation. Variational Bayes (VB) [10] is another way to approximate the (log) marginal likelihood with a lower bound by means of the variational methods. Since no VB algorithm exists for FA by eq.(1), we derive one in this paper by adopting a uniform prior over the Stiefel manifold used in [9] for $\mathbf{U}$, i.e., $q(\mathbf{U}) = 2^{-m}\prod_i \Gamma(\frac{n-i+1}{2})\pi^{-\frac{n-i+1}{2}}$, the commonly used Gamma density as priors for the precision parameters, i.e., $q(\boldsymbol{\nu}|a^\nu, b^\nu) = \prod_{i=1}^m \Gamma(\nu_i|a_i^\nu, b_i^\nu)$, $q(\varphi|a^\varphi, b^\varphi) = \Gamma(\varphi|a^\varphi, b^\varphi)$, with $\boldsymbol{\nu} = \mathbf{\Lambda}^{-1}$ and $\varphi = (\sigma_e^2)^{-1}$. Then, the VBEM algorithm implements $\max_{\{p_{\mathbf{U}}, p_\nu, p_\varphi, p_Y\}} \mathcal{F}$ for each candidate scale m, and $\mathcal{F}$ is the variational lower bound:

$$\mathcal{F} = \int p_{\mathbf{U}}p_{\boldsymbol{\nu}}p_\varphi p_Y \ln\left[\frac{q(X_N, Y|\mathbf{\Theta})q(\mathbf{\Theta}|\Xi)}{p_{\mathbf{U}}p_{\boldsymbol{\nu}}p_\varphi p_Y}\right] dY\, d\mathbf{U}\, d\boldsymbol{\nu}\, d\varphi,$$

where the posterior is constrained to be in a factorized form of $p_{\mathbf{U}}p_{\boldsymbol{\nu}}p_\varphi p_Y$, and $q(X_N, Y|\mathbf{\Theta}) = \prod_t q(\mathbf{x}_t|\mathbf{y}_t)q(\mathbf{y}_t)$ is given by eq.(1), and $q(\mathbf{\Theta}|\Xi) = q(\mathbf{U})\, q(\boldsymbol{\nu}|a^\nu, b^\nu)q(\varphi|a^\varphi, b^\varphi)$.

Firstly proposed in [11] and systematically developed over a decade [16, 17], 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. Given in [16, 17], the general two-stage procedure of BYY harmony learning is summarized in Tab.1, and the Stage-I is to implement

$$\begin{array}{l} \text{I(a): } \mathbf{\Theta}_m^{(\tau)} = \arg\max / \text{incr}_{\mathbf{\Theta}_m} H(p\|q, \mathbf{\Theta}_m, \Xi^{(\tau-1)}), \\ \text{I(b): } \Xi^{(\tau)} = \arg\max / \text{incr}_{\Xi}\{H(p\|q, \mathbf{\Theta}_m^{(\tau)}, \Xi) + \frac{1}{2}d(\Xi)\}, \\ d(\Xi) = -d_m + (\mathbf{\Theta}_m^{(\tau-1)} - \mathbf{\Theta}_m^{(\tau)})\Omega(\mathbf{\Theta}_m^{(\tau-1)} - \mathbf{\Theta}_m^{(\tau)}), \end{array} \quad (2)$$

where $\Omega = \nabla^2_{\mathbf{\Theta}\mathbf{\Theta}^T} H(p\|q, \mathbf{\Theta}_m^{(\tau)}, \Xi)$. Specifically the har-

band (%)	very good	not good
$80 \sim 100$	most criteria	DNLL, AIC
$40 \sim 80$	BYY,VB	BIC,CAIC,DNLL
$0 \sim 40$	BYY,VB	the rest criteria
low SNR $\{1.5, 2.0\}$	VB $\{7$ red $*\}$ BYY $\{8$ red $*$ = 1(BYYo)+ 3(BYYp)+4(BYYph)$\}$	the rest criteria get fewer than 3 red $*$
small N $<= 75$	BYYph gets most red $*$	the rest criteria get few red $*$

Table 2. The comparisons are based on (1) the band area between the specified contour lines $(\%)$ (the bigger and closer to left corner, the better), (2) the number of red asterisk $(*)$ (the more, the better).

mony functional for FA by eq.(1) is

$$H(p\|q,\mathbf{\Theta}_m,\Xi) = \prod_t \ln G(\mathbf{x}_t|\mathbf{0},\Sigma_x) + N\ln\sqrt{(2\pi e)^n|\Sigma_{y|x}|} + d_r(\widetilde{W}) + \ln q(\mathbf{\Theta}|\Xi), \quad (3)$$

where $d_r(\widetilde{W}) = -Tr[\Delta_W^T(\Sigma_{y|x})^{-1}\Delta_W S_N]$ will vanish as the difference $\Delta_W = \widetilde{W} - W$ converges to zero, i.e. the free parameter $\widetilde{W}$ converges to $W = \mathbf{\Lambda U}\mathbf{\Sigma}_x^{-1}$, with $\Sigma_{y|x} = \mathbf{\Lambda}^{-1} + \mathbf{U}^T\mathbf{\Sigma}_e^{-1}\mathbf{U}$ and $\mathbf{\Sigma}_x$ given in eq.(1).

Ignoring priors by letting $q(\mathbf{\Theta}|\Xi) = 1$, BYY (denoted as "BYYo") still possesses a good model selection ability[16]. By eq.(2), we further implement BYY (denoted as "BYYp") by adopting the same priors as used in VB, so that $\ln q(\mathbf{\Theta}|\Xi)$ in eq.(3) plays a role of regularization. By I(b) in eq.(2), the hyperparameters are updated in BYY (named "BYYph") to further increase the harmony functional. All BYY algorithms are implemented by the gradient method.

4. EMPIRICAL ANALYSIS

Empirical analysis is based on a series of controlled experiments by varying a wide range of the sample size $N \in \{25, 50, 75, 100, 200, 400, 800\}$, SNR $\gamma_o = \frac{\lambda_{m^*}}{\sigma_e^2} + 1 \in \{1.2, 1.5, 2, 2.5, 3, 3.5, 4, 8, 16\}$, where $\lambda_i = \ldots = \lambda_{m^*} = 1$, and n, m^* (i.e., $\dim(\mathbf{x}), \dim(\mathbf{y})$) are respectively fixed at $15, 5$ due to the space limit. For each of 10^2 independent runs, the two-stage procedure for every criterion is made on the set of candidate models $\mathcal{M} = \{1, \ldots, 9\}$ based on a data set X_N randomly generated according to a chosen setting for N, γ_o. We report the percentages of the successful selections, i.e., $\hat{m} = m^*$, in the form of contour maps in Fig.1.

The contour maps define a family of model selection *indifference curves* that visualize the performance over the space of SNR and N. The performances decrease as N and SNR reduce. Also, it can be observed from Fig.1 that (1) a *three-region partition*, i.e., all criteria perform well/bad when SNR and N are large/too small, while the region with moderate SNR and N differentiates those criteria well; (2) a *diminishing marginal effect*, i.e., the amount of SNR (or N) to trade for a unit loss of N (or SNR) increases as moving down an indifference curve.

Detailed observations from Fig.1,2 are listed in Tab.3. VB and three variants of BYY are relatively more robust than the other criteria against reducing the sample size and SNR, where BYYph performs the best in general.

All methods are also evaluated on a real world dataset Pendigits[1] (16 attributes, 10 classes, 10992 instances). Similarly, we vary the training sample size N. The classification results basically coincide with the model selection performance on synthetic data.

%	$N = 16$	$N = 30$	$N = 100$
AIC	56.46 ± 6.20	88.91 ± 1.72	96.61 ± 0.39
BIC	56.46 ± 6.19	87.72 ± 1.92	$\mathbf{96.64 \pm 0.40}$
HQC	56.46 ± 6.20	88.51 ± 1.59	96.62 ± 0.31
CAIC	48.34 ± 12.7	87.63 ± 1.62	96.63 ± 0.33
DNLL	79.18 ± 9.01	86.48 ± 1.71	91.26 ± 0.58
VB	87.02 ± 2.45	93.86 ± 1.31	96.19 ± 0.31
BYYph	$\mathbf{88.57 \pm 1.04}$	$\mathbf{94.15 \pm 0.33}$	96.29 ± 0.16

Table 3. Classification accuracies *mean*$\pm$*stdev* of 10^2 runs.

5. CONCLUSION

Based on the problem of determining the number of underlying source signals, we have investigated the relative strengths and weaknesses of not only the classical AIC, BIC/MDL, CAIC, HQC, but also recently developed Minka's criterion, VB and BYY. We derive a new VB algorithm for FA by imposing appropriate priors, which are also adopted in BYY for further implementations. The investigation is made via a new empirical analyzing tool featured by model selection indifference curves which reveal a three-region partition and a diminishing marginal effect. Moreover, the BYY with the priors' hyperparaemters updated is the best in general.

6. REFERENCES

[1] M. Wax and T. Kailath, "Detection of signals by information theoretic criteria," *IEEE Trans. on Acoustics, Speech and Signal Processing*, vol. ASSP-33, no. 2, pp. 387, 1985.

[2] T.W. Anderson and H. Rubin, "Statistical inference in factor analysis," in *Proc. of third Berkeley symposium on mathematical statistics and probability*, 1956, vol. 5, pp. 111–150.

[3] Michael E. Tipping and Christopher M. Bishop, "Mixtures of probabilistic principal component analyzers," *Neural Computation*, vol. 11, no. 2, pp. 443–482, 1999.

[4] H. Akaike, "A new look at the statistical model identification," *IEEE Transactions on Automatic Control*, vol. 19, no. 6, pp. 716–723, Dec 1974.

[5] Hamparsum Bozdogan, "Model selection and Akaike's Information Criterion (AIC): The general theory and its analytical extensions," *Psychometrika*, vol. 52, no. 3, pp. 345–370, 1987.

[6] E. J. Hannan, A. J. McDougall, and D. S. Poskitt, "The determination of the order of an autoregression," *Journal of the Royal Statistical Society. Series B (Methodological)*, vol. 51, pp. 217–233, 1989.

[7] Gideon Schwarz, "Estimating the Dimension of a Model," *The Annual of Statistics*, vol. 6, no. 2, pp. 461–464, 1978.

[1] from UCI repository: http://archive.ics.uci.edu/ml/datasets.html

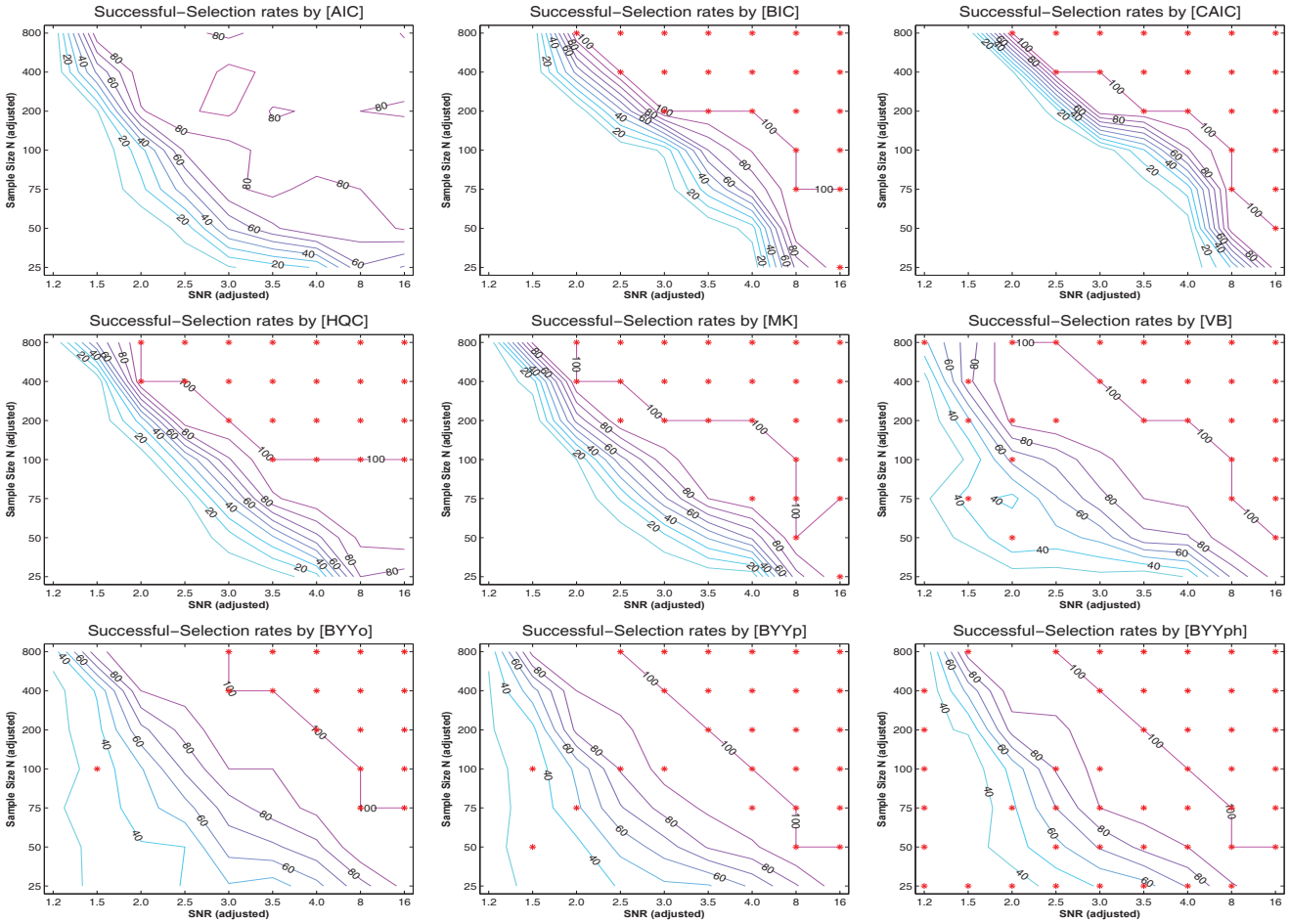

Fig. 1. Contour maps of successful selection rates of all criteria are drawn against to adjusted axes (i.e., equal space among setting values). A red asterisk (∗) indicates the corresponding criterion gets the highest successful selection rate at that setting.

[8] J Rissanen, "Modelling by the shortest data description," *Automatica*, vol. 14, pp. 465–471, 1978.

[9] Thomas P. Minka, "Automatic choice of dimensionality for PCA," in *Advances in Neural Information Processing Systems 13*, 2001, pp. 598–604.

[10] Matthew J. Beal, *Variational Algorithms for Approximate Bayesian Inference*, Ph.D. thesis, Gatsby Computational Neuroscience Unit, University College London, 2003.

[11] Lei Xu, "Bayesian-Kullback coupled Ying-Yang machines: Unified learnings and new results on vector quantization," in *International Conference on Neural Information Processing (ICONIP)*, 1995, pp. 977–988.

[12] E. Fishler, M. Grosmann, and H. Messer, "Detection of signals by information theoretic criteria: general asymptotic performance analysis," *IEEE Transactions on Signal Processing*, vol. 50, no. 5, pp. 1027–1036, May 2002.

[13] Shikui Tu and Lei Xu, "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*, Berlin, Heidelberg, 2009, pp. 154–162, Springer-Verlag.

[14] Lei Xu, "Bayesian Ying-Yang learning theory for data dimension reduction and determination," *Journal of Computational Intelligence in Finance*, vol. 6, no. 5, pp. 6–18, 1998.

[15] Shikui Tu and Lei Xu, "On the two parameterizations of factor analysis: which one is better?," *(In preparation)*, 2009.

[16] Lei Xu, "Bayesian Ying Yang Learning," in *Scholarpedia 2(3):1809, http://scholarpedia.org/article/Bayesian_Ying_Yang_Learning*, 2007.

[17] Lei Xu, "Bayesian Ying-Yang System, Best Harmony Learning, and Five Action Circling," to appear in an invited special issue on *Emerging Themes on Information Theory and Bayesian Approach*, Frontiers of Electrical and Electronic Engineering in China, a journal jointly published by Higher Education Press of China and Springer, 2010.

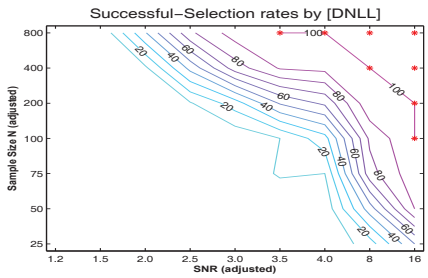

Fig. 2. Continue Fig.1 for DNLL.