

Model Selection Based on Maximum Likelihood Estimation: A Jackknife Approach



Hyunsook Lee and G. Jogesh Babu
Department of Statistics
The Pennsylvania State University
{hlee, babu}@stat.psu.edu

Introduction

Maximum likelihood principle plays a role in shaping many well known model selection criteria. The most popular criteria are Akaike information criterion (AIC), Minimum descriptive length (MDL), and Bayesian information criterion (BIC). AIC was introduced to find a model that minimizes the Kullback-Leibler(KL) distance

$$I[g; f_{\hat{\theta}}] := \int g(x) \log g(x) dx - \int g(x) \log f_{\hat{\theta}}(x) dx.$$

For an unknown density g , AIC chooses a model that maximizes the estimate of $E_g[\log f_{\hat{\theta}}(X)]$ from which bias becomes unavoidable. This bias is presented as a penalty term in AIC, in proportion to the number of parameters in a model. MDL was developed to find a model that minimizes code lengths, which is equivalent to maximizing entropy. Maximum entropy is similar to maximizing the log likelihood in AIC; however, MDL measures the complexity of a model, which contributes as a penalty term. On the other hand, BIC was developed from the idea of choosing a model of most probable posterior distribution. MDL and BIC share the same penalty term, $p \log n$. Cross validation is another procedure used in model selection. It was shown to be asymptotically equivalent to AIC by Stone (1977). Several studies were found in bootstrap model selection by estimating the KL distance.

We developed Jackknife information criterion (JIC) and studied its asymptotic properties with minimal assumptions on the likelihood. In contrast to AIC type criteria, the jackknife method reduces the bias substantially. The comparison among information criteria with JIC is presented through simulations.

Preliminaries

Suppose the data points, X_1, X_2, \dots, X_n are from an unknown distribution G and $\mathcal{M} = \{f_{\theta} : \theta \in \Theta^p\}$ is a class of candidate models. The log likelihood $\log f_{\theta}(X_i)$ of X_i is denoted by $l_i(\theta)$, the log likelihood functions of all the observations is $L(\theta) = \sum_{i=1}^n l_i(\theta)$, and the log likelihood function without i^{th} observation is $L_{-i}(\theta) = \sum_{j \neq i} l_j(\theta)$. The followings are assumed;

(J1) There exists a unique parameter θ_o in Θ^p that satisfies $E_g[\nabla_{\theta} l_1(\theta_o)] = 0$.

(J2) Within a neighborhood of θ_o , it is assumed that the first and the second derivatives of the log likelihood, $l_i(\theta)$, with respect to θ exist and they are denoted by $\nabla_{\theta} l_i(\theta)$ and $\nabla_{\theta}^2 l_i(\theta)$, respectively. For any $\theta \in \Theta^p$, those derivatives are bounded, i.e. $|\frac{\partial}{\partial \theta_k} l_i(\theta)| < h(X_i)$ and $|\frac{\partial^2}{\partial \theta_k \partial \theta_l} l_i(\theta)| < h(X_i)$, where $k, l = 1, \dots, p$ for all X_i and $h(X_i)$ is non negative such that $E_g[h(X_i)] < \infty$.

(J3) $E_g[\nabla_{\theta}^2 l_1(\theta_o)] = -\Lambda$, where Λ is a non-singular matrix.

(J4) For $\psi_{\theta}(X_i) = l_i(\theta)$ or $\psi_{\theta}(X_i) = \nabla_{\theta}^T l_i(\theta)$,

$$\nabla_{\theta} E[\psi_{\theta}(X_i)] = E[\nabla_{\theta} \psi_{\theta}(X_i)].$$

(J5) For $\forall \theta \in \Theta$, $E_g[\nabla_{\theta} l_i(\theta) \nabla_{\theta}^T l_i(\theta)]$ and $-E_g[\nabla_{\theta}^2 l_i(\theta)]$ are finite and positive definite $p \times p$ matrices. If $g = f_{\theta_o}$, then $E_{f_{\theta_o}}[\nabla_{\theta} l_i(\theta) \nabla_{\theta}^T l_i(\theta)] = -E_{f_{\theta_o}}[\nabla_{\theta}^2 l_i(\theta)]$.

Jackknife model selection

Definition 1 Let J_n , the jackknife estimator of log likelihood, be

$$J_n = nL(\hat{\theta}) - \sum_{i=1}^n L_{-i}(\hat{\theta}_{-i}).$$

Under (J1)-(J5), $\hat{\theta} \xrightarrow{p} \theta_o$ and $\hat{\theta}_{-i} \xrightarrow{p} \theta_o$ uniformly in i , which leads to the following theorem.

Theorem 1 Let X_1, X_2, \dots, X_n be iid random variables from a density function g of an unknown distribution and $\mathcal{M} = \{f_{\theta} : \theta \in \Theta^p\}$ be a class of candidate models. If the (g, \mathcal{M}) meets (J1)-(J5), then the jackknife estimator of log likelihood, J_n satisfies

$$\begin{aligned} J_n &= nL(\hat{\theta}) - \sum_{i=1}^n L_{-i}(\hat{\theta}_{-i}) \\ &= L(\theta_o) + \frac{1}{n} \sum_{i \neq j} \nabla l_i(\theta_o)^T \Lambda^{-1} \nabla l_j(\theta_o) + \varepsilon_n, \end{aligned} \quad (1)$$

where $E(\varepsilon_n^2) \rightarrow 0$ as $n \rightarrow \infty$. Moreover, $E[\frac{1}{n} \sum_{i \neq j} \nabla l_i(\theta_o)^T \Lambda^{-1} \nabla l_j(\theta_o)] = 0$ so that J_n is an asymptotically unbiased estimator of the log likelihood.

To obtain the stochastic order of J_n , in addition to (J1)-(J5), we assume the followings;

(J6) There exists $\hat{\theta}$ such that $\nabla_{\theta} L(\hat{\theta}) = 0$ and $\hat{\theta}$ tends to θ_o with probability 1.

(J7) There exists $\hat{\theta}_{-i}$ such that $\nabla_{\theta} L_{-i}(\hat{\theta}_{-i}) = 0$ and $\hat{\theta}_{-i}$ tends to θ_o with probability 1 uniformly in i .

Theorem 2 Let X_1, X_2, \dots, X_n be iid random variables from the unknown distribution with density g and \mathcal{M} be a class of candidate models such that $\mathcal{M} = \{f_{\theta} : \theta \in \Theta^p\}$. Under (J1)-(J7), the stochastic orders relating to J_n are:

- (1) $J_n = L(\theta_o) + O(\log \log n)$ a.e.
- (2) $\frac{1}{n} J_n = \int \log f_{\theta}(x) g(x) dx + O(\sqrt{n^{-1} \log \log n})$ a.e.

Popular information criteria involve log likelihood multiplied by -2.

Definition 2 (JIC) Jackknife information criterion is defined as minus twice J_n .

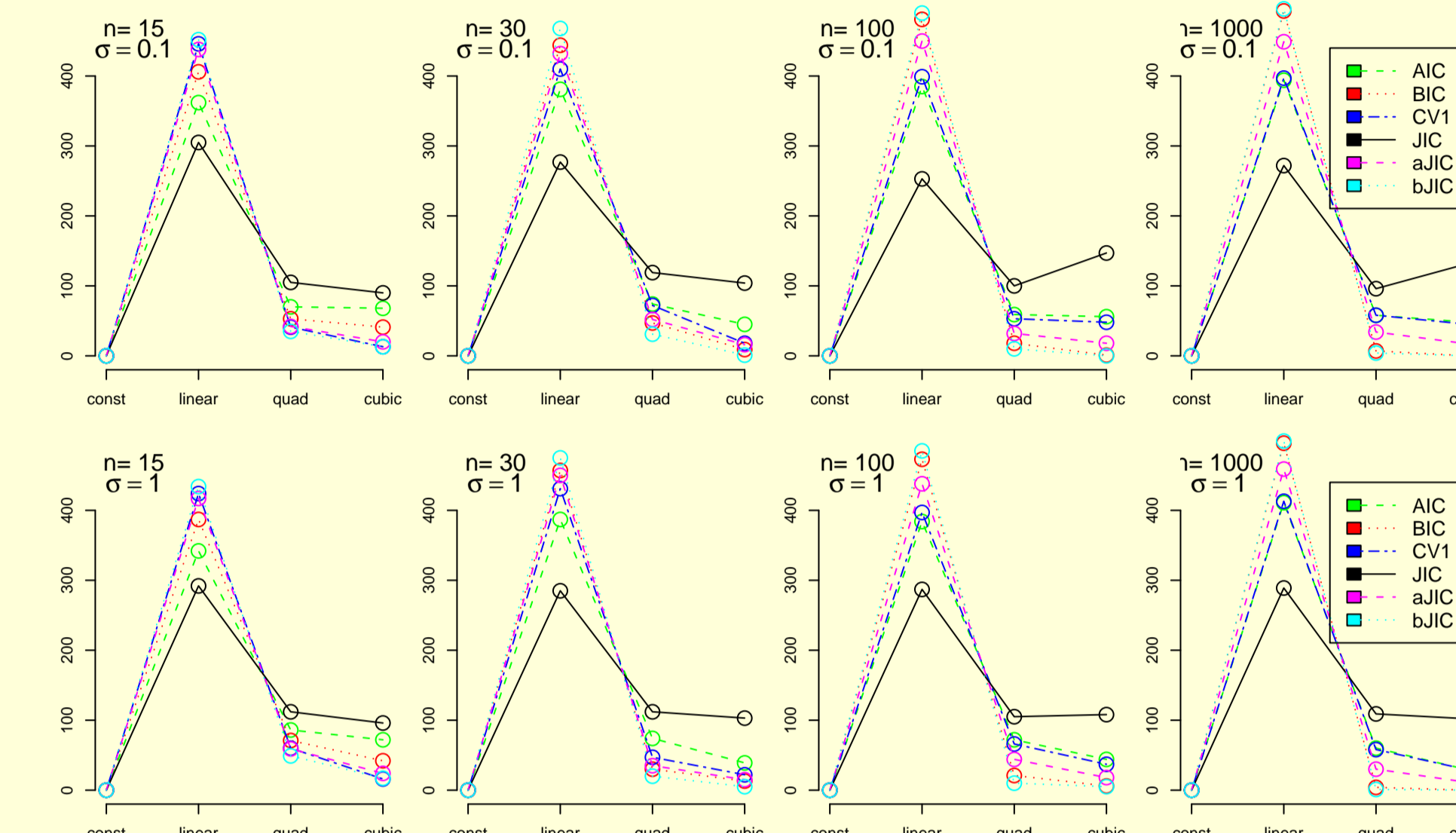
$$\begin{aligned} JIC &= -2J_n \\ &= -2nL(\hat{\theta}) + 2 \sum_{i=1}^n L_{-i}(\hat{\theta}_{-i}). \end{aligned}$$

For nested models, asymptotically $L(\theta_o)$ are identical so that JIC cannot distinguish among models. We suggest two modified JICs related to AIC and BIC.

$$\begin{aligned} aJIC &= JIC + 2p \\ bJIC &= JIC + p \log n \end{aligned}$$

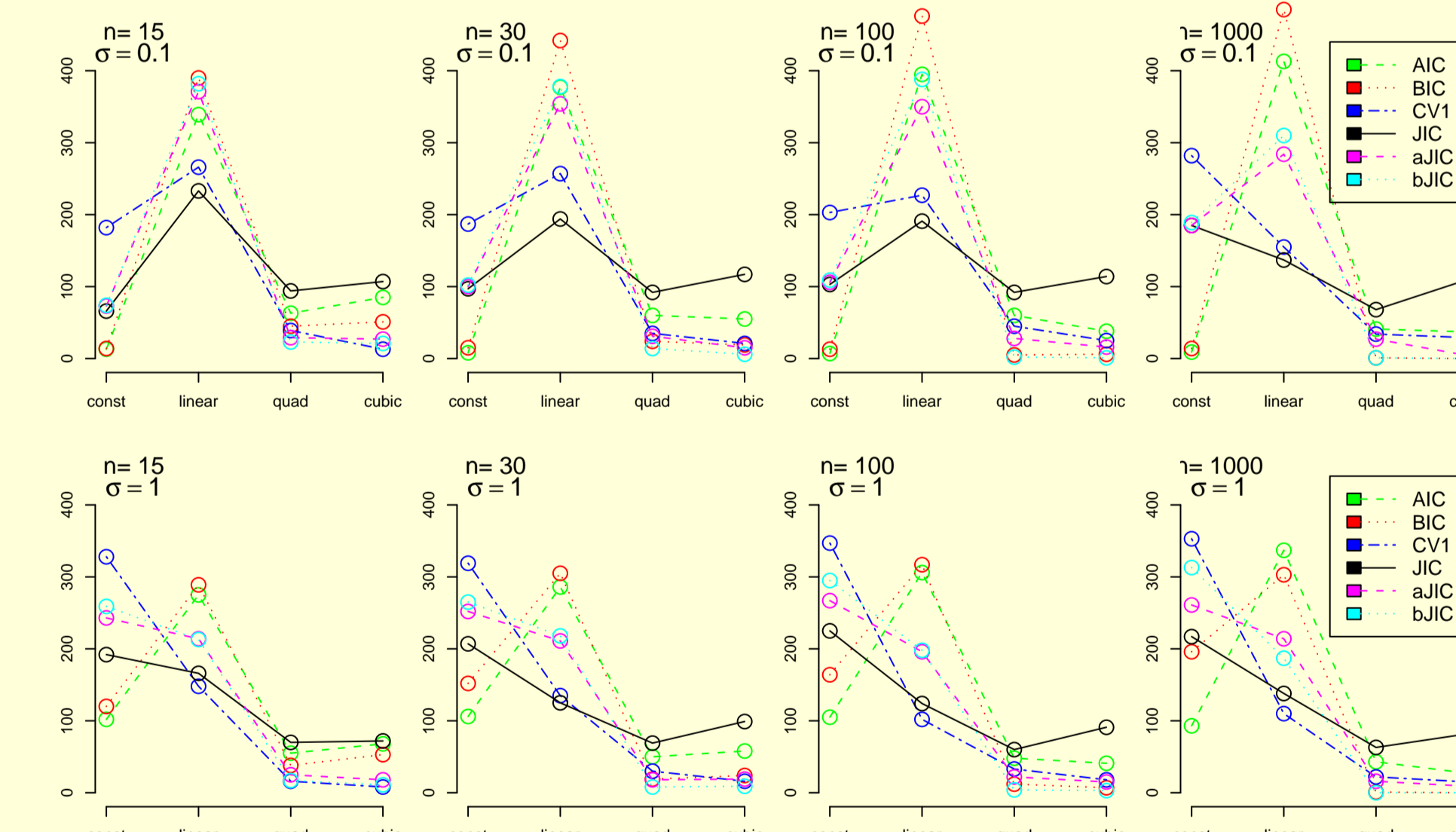
Simulations

In this section, the performance of JIC is compared with other information criteria. First, we considered a very simple linear model. Suppose the true model is $Y = 2X + 5 + e$, where $e \sim N(0, \sigma^2)$ and candidate models are polynomials of different orders. We investigated the frequencies of model selection via several information criteria while X and e were randomly generated. There were 500 trials.



From the above figure, as expected, JIC was least effective in this nested case. However, aJIC and bJIC slightly outperformed AIC and BIC, respectively.

Now, we considered a case when the true model does not exist in the set of nested candidate models. Consider the true model is $Y = 2X + 5 + e$, where $e \sim C(0, \sigma^2)$, from which C stands for Cauchy distribution, and candidate models are polynomials of different orders with normal error distribution.



From the above figure, AIC and BIC outperformed the information criteria of resampling techniques.

Last, we considered the case when candidate models are not nested. The true model is $C(0, 1)$ and candidates are $C(\cdot, 1)$, $N(\cdot, 1)$, and $N(\cdot, \cdot)$, where \cdot denotes an estimated parameter. Scale was adjusted to match two distributions. From the table, JIC performed similarly to AIC and CV1 performed well in small sample. Although the cauchy distribution and normal distribution are symmetric, all information criteria pick the correct distribution within relatively small sample size.

	n = 15			n = 30			n = 50				
	$C_{(\cdot, 1)}$	$N_{(\cdot, 1)}$	$N_{(\cdot, \cdot)}$	$C_{(\cdot, 1)}$	$N_{(\cdot, 1)}$	$N_{(\cdot, \cdot)}$	$C_{(\cdot, 1)}$	$N_{(\cdot, 1)}$	$N_{(\cdot, \cdot)}$		
AIC	382	47	71	AIC	471	13	16	AIC	496	3	1
BIC	385	55	60	BIC	472	19	9	BIC	496	4	0
CV1	395	67	38	CV1	475	19	6	CV1	496	4	0
JIC	381	44	75	JIC	470	12	18	JIC	496	3	1
aJIC	394	66	40	aJIC	472	19	9	aJIC	496	4	0
bJIC	396	71	33	bJIC	475	19	6	bJIC	496	4	0

Discussion

The strength of JIC lies in its applicability to various families of densities unlike popular model selection criteria, which are applicable only in a restricted distribution family. The simulation study showed that JIC performed similarly among popular selection criteria when candidates models are not nested.

In general, consistent model selection criteria were highly valued in choosing a correct model. The strong consistency comes from a penalty term, when nested models are competing. Some attempts were made to improve AIC and BIC by careful examination of estimating bias. In this context, resampling techniques that reduce bias does not seem to be successful.

A couple of studies presented that the bias of bootstrap model selection is proportional to the number of parameters in a model (Chung *et.al.*, 1997; Ishiguro *et.al.*, 1997; and Shibata, 1997). Chung *et.al.*(1997) added that bootstrap after bootstrap was asymptotically unbiased. Compared to bootstrap after bootstrap, JIC is quite simpler computation-wise and it is asymptotically unbiased.

From eq.(??), JIC could be improved by (a) finding an estimator of $\frac{1}{n} \sum_{i \neq j} \nabla l_i(\theta_o)^T \Lambda^{-1} \nabla l_j(\theta_o)$ or (b) considering 2^{nd} order bias. Resampling techniques allow to be applied to more general sets of candidate models which are not necessarily nested.

A smaller quantity of information criterion does not pertain a correctly specified model; nonetheless, a correct model attains the smallest quantity of IC (Sin and White, 1996). Model selection methods become successful after better understanding of data sets and restricting possible candidates (Burnham and Anderson, 2002).

References

1. Burnham, K.P. and Anderson, D.R. (2002) *Model selection and Inference, A practical Information-Theoretic Approach*, 2nd Ed., Springer-Verlag, New York.
2. Chung, H., Lee, K., and Koo, J. (1996) A note on bootstrap model selection criterion *Stat. & Prob. Letters*, 26, 35-41
3. Ishiguro, M., Sakamoto, Y., and Kitagawa, G. (1997). Bootstrapping Log Likelihood and EIC, an Extension of AIC *Annals of the Institute of Statistical Mathematics*, 49(3), 411-434
4. Shibata, R. (1997). Bootstrap estimate of Kullback-Leibler information for model selection *Statistica Sinica*, 7, 375-394
5. Sin, C. and White, H. (1996) Information criteria for selecting possibly misspecified parametric models *J. Econometrics*, 71, 207-225
6. Stone, M. (1977). An asymptotic equivalence of choice of model by cross-validation and Akaike's criterion *J. Roy. Stat. Soc. B*, 39(1), 44-47

Acknowledgment

This work is supported in part by the NSF Focused Research Group grant DMS-0101360 (P.I.: G. J. Babu).