REGRESSION AND TIME SERIES MODEL SELECTION USING VARIANTS OF THE SCHWARZ INFORMATION CRITERION

by Andrew A. Neath[†] and Joseph E. Cavanaugh[‡]

[†] Department of Mathematics and Statistics
P.O. Box 1653
Southern Illinois University
Edwardsville, IL 62026

[‡] Department of Statistics
222 Math Sciences Building
University of Missouri
Columbia, MO 65211

Keywords: Bayesian analysis; decision theory; Fisher information; information theory; multiple linear regression; state-space model.

ABSTRACT

The Schwarz (1978) information criterion, SIC, is a widely-used tool in model selection, largely due to its computational simplicity and effective performance in many modeling frameworks. The derivation of SIC (Schwarz, 1978) establishes the criterion as an asymptotic approximation to a transformation of the Bayesian posterior probability of a candidate model. In this paper, we investigate the derivation for the identification of terms which are discarded as being asymptotically negligible, but which may be significant in small to moderate sample-size applications. We suggest several SIC variants based on the inclusion of these terms. The results of a simulation study show that the variants improve upon the performance of SIC in two important areas of application: multiple linear regression and time series analysis.

1. Introduction

One of the most important problems confronting an investigator in statistical modeling is the choice of an appropriate model to characterize the underlying data. This determination can often be facilitated through the use of an information theoretic criterion, which judges the propriety of a fitted model by assessing whether it offers an optimal balance between "goodness of fit" and parsimony.

The first information theoretic criterion to gain wide-spread acceptance as a model selection tool was the Akaike (1973, 1974) information criterion, AIC. Many other criteria have been subsequently introduced, including well-known measures by Schwarz (1978), Rissanen (1978), Akaike (1978), Hannan and Quinn (1979), and Hurvich and Tsai (1989). Although AIC remains arguably the most widely used of the model selection criteria, the Schwarz information criterion, SIC, is a popular competitor. In fact, SIC is often preferred over AIC by practitioners who find appeal in either its Bayesian justification or its tendency to choose more parsimonious models than AIC.

Schwarz (1978) rigorously establishes SIC "for the case of independent, identically distributed observations, and linear models," under the assumption that the likelihood is from the regular exponential family. Haughton (1988) extends the derivation to a context where the likelihood is from the curved exponential family. Cavanaugh, Neath, and Shumway (1995) present a derivation which does not require that the likelihood has any particular form, but only assumes that it satisfies a set of non-restrictive regularity conditions. Additional generalizations of Schwarz's derivation are considered by Stone (1979), Leonard (1982), and Kashyap (1982).

As conveyed in the original derivation and subsequent extensions, the justification of SIC is based on establishing that the criterion serves as an asymptotic approximation to a transformation of the Bayesian posterior probability of a candidate model. Our interest lies in investigating this justification for the possibility of retaining terms in the criterion which are asymptotically negligible. Such terms would be of debatable value in large-sample settings, but may improve the effective-

ness of SIC in small to moderate sample-size applications.

In Section 2, we present a hueristic derivation of SIC and indicate how this derivation suggests certain variants of the criterion. In Section 3, we consider the performance of these variants in two important areas of application: multiple linear regression and time series analysis. Section 4 concludes.

2. Variants of SIC

Let Y_n denote the observed data. Assume that Y_n is to be described using a model M_k selected from a set of candidate models $M_1, M_2, \ldots, M_L$. Assume that each M_k $(1 \leq k \leq L)$ is uniquely parameterized by a vector ${}_k\underline{\theta}$, where ${}_k\underline{\theta}$ is an element of the parameter space $\Theta(k)$.

Let $L({}_k\underline{\theta} \mid Y_n)$ denote the likelihood for Y_n based on M_k. Also, let ${}_k\underline{\hat{\theta}}_n$ denote the maximum likelihood vector obtained by maximizing $L({}_k\underline{\theta} \mid Y_n)$ over $\Theta(k)$. We assume that derivatives of $L({}_k\underline{\theta} \mid Y_n)$ up to order three exist with respect to ${}_k\underline{\theta}$, and are continuous and suitably bounded for all ${}_k\underline{\theta} \in \Theta(k)$.

The motivation behind SIC can be seen through a Bayesian development of the model selection problem. Let $\pi(k)$ $(1 \leq k \leq L)$ denote a discrete prior over the models $M_1, M_2, \ldots, M_L$. Let $g({}_k\underline{\theta} \mid k)$ $(1 \leq k \leq L)$ denote a prior on ${}_k\underline{\theta}$ given the model M_k.

Applying Bayes' Theorem, the joint posterior of M_k and ${}_k\underline{\theta}$ can be written as

$$f((k,\ {}_k\underline{\theta}) \mid Y_n) = \frac{\pi(k)\ g({}_k\underline{\theta} \mid k)\ L({}_k\underline{\theta} \mid Y_n)}{h(Y_n)},$$

where $h(Y_n)$ denotes the marginal distribution of Y_n.

A Bayesian model selection rule would favor the model M_k which is *a posteriori* most probable. The posterior distribution for M_k is given by

$$P(k \mid Y_n) = h(Y_n)^{-1}\ \pi(k)\ \int L({}_k\underline{\theta} \mid Y_n)\ g({}_k\underline{\theta} \mid k)\ d_k\underline{\theta}.$$

Now consider minimizing $-2 \ln P(k \mid Y_n)$ as opposed to maximizing $P(k \mid Y_n)$. We have

$$-2 \ln P(k \mid Y_n) = 2 \ln \{h(Y_n)\} - 2 \ln \{\pi(k)\} - 2 \ln \left\{ \int L({}_k\underline{\theta} \mid Y_n)\ g({}_k\underline{\theta} \mid k)\ d_k\underline{\theta} \right\}.$$

Since the first of the three terms on the right-hand side of the preceding expression is constant with respect to k, we can discard this term for the purpose of model selection. We then obtain the proportionality

$$\begin{aligned} -2\ln P(k \mid Y_n) &\propto \\ S(k \mid Y_n) &\equiv -2\ln\{\pi(k)\} - 2\ln\left\{\int L({}_k\underline{\theta} \mid Y_n)\; g({}_k\underline{\theta} \mid k)\; d_k\underline{\theta}\right\}. \quad (2.1) \end{aligned}$$

Now consider the integral which appears in (2.1):

$$\int L({}_k\underline{\theta} \mid Y_n)\; g({}_k\underline{\theta} \mid k)\; d_k\underline{\theta}. \quad (2.2)$$

In order to obtain an approximation to this term, we take a second-order Taylor expansion of the log-likelihood about ${}_k\hat{\underline{\theta}}_n$. We have

$$\begin{aligned} \ln L({}_k\underline{\theta} \mid Y_n) &\approx \ln L({}_k\hat{\underline{\theta}}_n \mid Y_n) + ({}_k\underline{\theta} - {}_k\hat{\underline{\theta}}_n)' \; \frac{\partial \ln L({}_k\hat{\underline{\theta}}_n \mid Y_n)}{\partial {}_k\underline{\theta}} \\ &\quad + \frac{1}{2}({}_k\underline{\theta} - {}_k\hat{\underline{\theta}}_n)' \left[\frac{\partial^2 \ln L({}_k\hat{\underline{\theta}}_n \mid Y_n)}{\partial {}_k\underline{\theta} \partial {}_k\underline{\theta}'}\right] ({}_k\underline{\theta} - {}_k\hat{\underline{\theta}}_n) \\ &= \ln L({}_k\hat{\underline{\theta}}_n \mid Y_n) - \frac{n}{2}({}_k\underline{\theta} - {}_k\hat{\underline{\theta}}_n)' \left[I_n({}_k\hat{\underline{\theta}}_n, Y_n)\right] ({}_k\underline{\theta} - {}_k\hat{\underline{\theta}}_n), \end{aligned}$$

where

$$I_n({}_k\hat{\underline{\theta}}_n, Y_n) = -\frac{1}{n}\; \frac{\partial^2 \ln L({}_k\hat{\underline{\theta}}_n \mid Y_n)}{\partial {}_k\underline{\theta} \partial {}_k\underline{\theta}'}$$

is the observed Fisher information matrix. Thus,

$$L({}_k\underline{\theta} \mid Y_n) \approx L({}_k\hat{\underline{\theta}}_n \mid Y_n)\; \exp\left\{-\frac{n}{2}\; ({}_k\underline{\theta} - {}_k\hat{\underline{\theta}}_n)' \left[I_n({}_k\hat{\underline{\theta}}_n, Y_n)\right] ({}_k\underline{\theta} - {}_k\hat{\underline{\theta}}_n)\right\}.$$

We can therefore rewrite (2.2) as follows:

$$\begin{aligned} &\int L({}_k\underline{\theta} \mid Y_n)\; g({}_k\underline{\theta} \mid k)\; d_k\underline{\theta} \approx \\ &L({}_k\hat{\underline{\theta}}_n \mid Y_n) \int \exp\left\{-\frac{n}{2}\; ({}_k\underline{\theta} - {}_k\hat{\underline{\theta}}_n)' \left[I_n({}_k\hat{\underline{\theta}}_n, Y_n)\right] ({}_k\underline{\theta} - {}_k\hat{\underline{\theta}}_n)\right\} g({}_k\underline{\theta} \mid k)\; d_k\underline{\theta}. \quad (2.3) \end{aligned}$$

The preceding Taylor approximation holds when ${}_k\underline{\theta}$ is close to ${}_k\hat{\underline{\theta}}_n$. Thus, the approximation (2.3) should be valid for large n. In this instance, $L({}_k\underline{\theta} \mid Y_n)$ should dominate the prior $g({}_k\underline{\theta} \mid k)$ within a small neighborhood of ${}_k\hat{\underline{\theta}}_n$. Outside of this neighborhood, $L({}_k\underline{\theta} \mid Y_n)$ and the exponential term in (2.3) should be small enough to force the corresponding integrands near zero.

Now consider evaluating the right-hand side of (2.3) using the noninformative prior $g({}_k\underline{\theta} \mid k) = 1$. In this case, we have

$$\int \exp\left\{-\frac{n}{2}\ ({}_k\underline{\theta} - {}_k\hat{\underline{\theta}}_n)'\ \left[I_n({}_k\hat{\underline{\theta}}_n, Y_n)\right]\ ({}_k\underline{\theta} - {}_k\hat{\underline{\theta}}_n)\right\}\ d_k\underline{\theta} = (2\pi)^{(\dim({}_k\underline{\theta}\ )/2)}\ |n\ I_n({}_k\hat{\underline{\theta}}_n, Y_n)|^{-1/2}.$$

Substituting this result into (2.3) yields

$$\int L({}_k\underline{\theta} \mid Y_n)\ g({}_k\underline{\theta} \mid k)\ d_k\underline{\theta} \approx L({}_k\hat{\underline{\theta}}_n \mid Y_n)\ (2\pi)^{(\dim({}_k\underline{\theta}\ )/2)}\ |n\ I_n({}_k\hat{\underline{\theta}}_n, Y_n)|^{-1/2}. \quad (2.4)$$

(The preceding can be viewed as a variation on the LaPlace method of approximating the integral (2.2). See Kass and Raftery, 1995. The approximation (2.4) is valid so long as $g({}_k\underline{\theta} \mid k)$ is noninformative or "flat" over the neighborhood of ${}_k\hat{\underline{\theta}}_n$ where $L({}_k\underline{\theta} \mid Y_n)$ is dominant, although the choice of $g({}_k\underline{\theta} \mid k) = 1$ makes our derivation more tractable.)

We can use (2.4) and (2.1) to justify writing

$$S(k \mid Y_n) \approx$$
$$-2 \ln L({}_k\hat{\underline{\theta}}_n \mid Y_n) + \dim({}_k\underline{\theta}\ ) \left\{\ln\left(\frac{n}{2\pi}\right)\right\} + \ln|I_n({}_k\hat{\underline{\theta}}_n, Y_n)| - 2\ln\{\pi(k)\}. \quad (2.5)$$

Ignoring terms in the preceding that are bounded as the sample size grows to infinity, we obtain

$$S(k \mid Y_n) \approx -2 \ln L({}_k\hat{\underline{\theta}}_n \mid Y_n) + (\dim({}_k\underline{\theta}\ ))(\ln n) \equiv \text{SIC}. \quad (2.6)$$

Now suppose we reconsider the step which leads from (2.5) to (2.6). Expression (2.5) involves the terms

$$\ln|I_n({}_k\hat{\underline{\theta}}_n, Y_n)| \quad \text{and} \quad -2\ln\pi(k), \quad (2.7)$$

both of which are discarded for the definition of SIC. The former of these terms depends on the data and the form of the candidate model; the latter depends on the prior over the collection of candidate models. Both terms contain information relevant to assessing the suitability of a fitted candidate model. Moreover, although

both terms are asymptotically negligible, in small to moderate sample-size settings, either may be significant in relation to the two terms which define SIC.

Retaining either or both of the terms in (2.7) leads to variants of SIC which are not only computationally accessible, but are also easily justified through the derivation of the criterion. Define

$$\begin{aligned}
\text{SIC} &= -2 \ln L({}_k\hat{\underline{\theta}}_n \mid Y_n) + (\dim({}_k\underline{\theta}))(\ln n), && (2.8)\\
\text{SIC}_f &= \text{SIC} + \ln |I_n({}_k\hat{\underline{\theta}}_n, Y_n)|, && (2.9)\\
\text{SIC}_p &= \text{SIC} - 2 \ln \pi(k), && (2.10)\\
\text{SIC}_{fp} &= \text{SIC} + \ln |I_n({}_k\hat{\underline{\theta}}_n, Y_n)| - 2 \ln \pi(k). && (2.11)
\end{aligned}$$

Since the Fisher information matrix is important in characterizing the likelihood function for a candidate model, the term $\ln |I_n({}_k\hat{\underline{\theta}}_n, Y_n)|$ provides an intuitively reasonable correction to SIC. The term $-2 \ln \pi(k)$ would also provide a relevant correction in instances where *a priori*, an investigator favors some candidate models over others. (For instance, in modeling a monthly time series, models which contain seasonal components may be favored over ones which do not.) Since a certain degree of prior knowledge is needed to specify the models in the candidate collection, an investigator enters any modeling problem with a preference towards specific types of models. Thus, a meaningful definition of $\pi(k)$ should be possible in many practical applications. Omitting this correction would be equivalent to taking $\pi(k)$ to be a uniform prior over the models $M_1, M_2, \ldots, M_L$.

It should be noted that the appearance of the observed Fisher information $I_n({}_k\hat{\underline{\theta}}_n, Y_n)$ in the approximation of (2.2) leading to SIC has been observed in various discussions and developments of SIC: e.g., Leonard, 1982; Haughton, 1988; Kass and Raftery, 1995. In particular, Kashyap (1982) recommends a criterion similar to SIC_f for order selection in autoregressive moving-average modeling. Yet the inclusion of the term $\ln |I_n({}_k\hat{\underline{\theta}}_n, Y_n)|$ in the evaluation of SIC is not practiced, despite the fact that it could be easily computed in many applications. This is most likely due to the fact that the term is of a constant order, whereas the two terms comprising SIC

are of orders n and $\ln n$.

We should also note that both SIC and the proposed variants involve the approximation of the integral (2.2) for the purpose of approximating (2.1). Certainly, one may choose to directly evaluate (2.2) using a computationally intensive technique such as Gibbs sampling, importance sampling, Gaussian quadrature, etc. (See Kass and Raftery, 1995.) Such an approach would be a very important component of a formal Bayesian analysis, since it would lead to an exact determination of (2.1). We argue, however, that part of the popularity of SIC stems from its computational simplicity and its wide-spread applicability. The proposed SIC variants share these same two advantages, both of which would be lost if the direct evaluation of (2.2) was attempted.

3. Simulation Study

We examine the performance of the SIC variants (2.9), (2.10), (2.11) against traditional SIC (2.8) in a simulation study which focuses on two important modeling frameworks: multiple linear regression and time series analysis. We consider small to moderate sample-size settings.

3.1 Regression

Consider the ordinary linear regression model

$$\underline{y} = \mathbf{X}\underline{\beta} + \underline{\epsilon}, \qquad \underline{\epsilon} \sim N_n(\underline{0}, \sigma^2\ \mathbf{I}), \tag{3.1}$$

where $\underline{y}$ is an $n \times 1$ observation vector, $\underline{\epsilon}$ is an $n \times 1$ error vector, $\underline{\beta}$ is a $(r+1) \times 1$ parameter vector, and $\mathbf{X}$ is an $n \times (r+1)$ design matrix of full-column rank.

The goal is to determine which potential independent variables should be included in $\mathbf{X}$ in order to adequately describe the response variable y. For ease of exposition, we will assume our candidate models are nested. This corresponds to a practical setting where the predictor variables can be listed in some order of importance.

The design matrices for the models under consideration will have the following layouts:

$$\mathbf{X}_0 = \begin{bmatrix} 1 \\ \vdots \\ 1 \end{bmatrix}, \quad \mathbf{X}_1 = \begin{bmatrix} 1 & x_{11} \\ \vdots & \vdots \\ 1 & x_{n1} \end{bmatrix}, \quad \ldots, \quad \mathbf{X}_R = \begin{bmatrix} 1 & x_{11} & \ldots & x_{1R} \\ \vdots & \vdots & & \vdots \\ 1 & x_{n1} & \ldots & x_{nR} \end{bmatrix}.$$

We will denote the corresponding candidate models (3.1) by $M_0, M_1, \ldots, M_R$, respectively.

The quantities necessary to calculate SIC and its variants are easily obtained. Since the likelihood is given by

$$L(\underline{\beta}, \sigma \mid \underline{y}) = \left(2\pi\sigma^2\right)^{-\frac{n}{2}} \exp\left\{-\frac{1}{2\sigma^2} \, ||\underline{y} - \mathbf{X}\underline{\beta}||^2\right\},$$

we have

$$-2 \ln L(\hat{\underline{\beta}}, \hat{\sigma} \mid \underline{y}) = n \ln(2\pi \mathrm{e}) + n \ln(\hat{\sigma}^2)$$

where

$$\hat{\sigma}^2 = \frac{1}{n} \, ||\underline{y} - \mathbf{X}\hat{\underline{\beta}}||^2.$$

For the observed Fisher information matrix for the parameter vector $\underline{\theta} = (\underline{\beta}^{'} \sigma)^{'}$, one can easily verify

$$-\frac{1}{n} \frac{\partial^2 \ln L(\underline{\beta}, \sigma \mid \underline{y})}{\partial \underline{\theta} \partial \underline{\theta}^{'}} \Big|_{\underline{\theta} = (\hat{\underline{\beta}}^{'} \hat{\sigma})^{'}} = \frac{1}{\hat{\sigma}^2} \begin{bmatrix} \frac{1}{n}\mathbf{X}^{'}\mathbf{X} & 0 \\ 0 & 2 \end{bmatrix} \equiv I_n(\hat{\sigma}).$$

Since $I_n(\hat{\sigma})$ is $(r+2) \times (r+2)$ and $(\mathbf{X}^{'}\mathbf{X})$ is $(r+1) \times (r+1)$, we have

$$|I_n(\hat{\sigma})| = \left(\frac{1}{\hat{\sigma}^2}\right)^{r+2} \left\{2 \left(\frac{1}{n}\right)^{r+1} \left|\mathbf{X}^{'}\mathbf{X}\right|\right\}$$

and

$$\ln |I_n(\hat{\sigma})| = \ln 2 + \ln |\mathbf{X}^{'}\mathbf{X}| - (r+1) \ln n - (r+2) \ln \hat{\sigma}^2.$$

As for choosing a prior over the class of candidate models, one possibility is to use the Poisson distribution with the mean set equal to the prior estimate of the number of predictor variables. As a right-skewed distribution with a mode at this prior estimate, the Poisson may well serve as a suitable quantifier of subjective

information. Since we are placing probabilities on a finite set of models, a truncated Poisson would be used.

For $r = 0, 1, \ldots, R$, the criteria (2.8) through (2.11) can be expressed as

$$\begin{aligned} \text{SIC} &= n \ln \hat{\sigma}^2 + (r+1) \ln n, \\ \text{SIC}_f &= (n-r-2) \ln \hat{\sigma}^2 + \ln |\mathbf{X}'\mathbf{X}|, \\ \text{SIC}_p &= n \ln \hat{\sigma}^2 + (r+1) \ln n - 2 \ln \pi(r), \\ \text{SIC}_{fp} &= (n-r-2) \ln \hat{\sigma}^2 + \ln |\mathbf{X}'\mathbf{X}| - 2 \ln \pi(r). \end{aligned}$$

(Here, we have ignored constants which would not vary over different fitted models within the candidate class.)

We compare the behavior of the four proposed criteria by simulating a setting where one must decide among seven candidate models $M_0, M_1, \ldots, M_6$. (Thus, $R = 6$.) One thousand sets of data are generated from a true model in the candidate class. For every data set, the seven models in the candidate class are fit to the data; SIC, SIC_f, SIC_p, and SIC_{fp} are evaluated; and the favored model for each criterion is recorded. Over the one thousand data sets, the selections are tabulated, summarized, and reported.

For the prior $\pi(r)$, a Poisson distribution is used where the mean is set equal to the number of predictor variables in the true model. The distribution is truncated at 6.

Three simulation sets are run using various true models and sample sizes. The results are reported in Tables I to III.

The first set (Table I) features a true model with three predictor variables, where each variable has the same influence on the response. The sample size is 20. SIC_f greatly outperforms SIC: SIC_f correctly chooses the true model 98.1% of the time, compared to a 78.8% correct selection rate for SIC. Note that SIC_f does not exhibit the same tendency as SIC to choose models which have too many predictor variables.

The second set (Table II) features the same true model and sample size as in the first set, except that the candidate predictor variables are correlated. This set is

included to check whether the presence of multicollinearity affects the performance of the criteria. The correct selection rates decline slightly from the first set, yet the rate for SIC_f remains quite high at 94.8%.

The third set (Table III) features a true model where the second predictor variable has less influence on the response than the first, and the third has less influence than the second. The sample size is 15. The performance of SIC_f is not diminished by either the form of the true model or the small sample size: SIC_f obtains a 97.5% correct selection rate. On the other hand, SIC chooses the correct model only 65.1% of the time, and exhibits a strong tendency towards choosing models with too many predictor variables.

Although SIC and SIC_f are asymptotically equivalent, the simulation results indicate that their selection properties are different for small to moderate sample sizes. Models chosen by SIC often tend to be larger than necessary. In many applications, the additional term in SIC_f corrects this propensity towards overfitting without overcompensating and leading the criterion towards choosing undersized models. Since underfitting is often regarded as a more serious error than overfitting, the fact that SIC_f often guards against the latter without leaning towards the former makes the variant an attractive alternative to SIC.

The selection criteria depending on the Poisson prior (SIC_p and SIC_{fp}) outperform their counterparts that implicitly assume a uniform prior (SIC and SIC_f, respectively). In each case, our prior is chosen so that the prior estimate of the true model dimension is correct; thus, it is not surprising that the inclusion of the prior improves the selection performance of the criteria.

Naturally, whether the choice of the prior improves or degrades the performance of a Bayesian procedure depends upon both the validity and the form of the prior. Thus, the selection of $\pi(r)$ is an important practical issue, yet one which is outside the scope of the present paper.

TABLE I: Summary of Simulation Set 1.

$$\text{True Model:} \qquad y_i = 1 + x_{i1} + x_{i2} + x_{i3} + \epsilon_i,$$
$$\epsilon_i \sim \ iid\ N(0, 0.25).$$

- Predictors $x_{i1}, x_{i2}, \ldots, x_{i6}$: Generated from a uniform distribution on the interval (0,6); i.e., $x_{ij} \sim \ iid\ U(0, 6)$.
- Sample Size: $n = 20$.
- Prior $\pi(r)$: Poisson distribution with a mean of 3 truncated for $r > 6$.

Criterion Selections

r	SIC	SIC_f	SIC_p	SIC_{fp}
0	0	0	0	0
1	0	0	0	0
2	0	0	0	0
3	788	981	845	991
4	108	18	95	8
5	61	1	36	1
6	43	0	24	0

TABLE II: Summary of Simulation Set 2.

$$\text{True Model:} \qquad y_i = 1 + x_{i1} + x_{i2} + x_{i3} + \epsilon_i,$$
$$\epsilon_i \sim \ iid\ N(0, 0.25).$$

- Predictors $x_{i1}, x_{i2}, \ldots, x_{i6}$: Each 6×1 vector $\underline{\mathbf{x}}_i = (x_{i1}, x_{i2}, \ldots, x_{i6})'$ is generated from a multivariate normal distribution. For this distribution, the mean vector is $(3, 3, \ldots, 3)$, and the variance/covariance matrix is such that the diagonal elements are 1.50 and the off-diagonal elements are 0.75.
- Sample Size: $n = 20$.
- Prior $\pi(r)$: Poisson distribution with a mean of 3 truncated for $r > 6$.

Criterion Selections

r	SIC	SIC_f	SIC_p	SIC_{fp}
0	0	0	0	0
1	0	0	0	0
2	0	0	0	0
3	750	948	819	964
4	133	46	116	34
5	63	5	34	1
6	54	1	31	1

TABLE III: Summary of Simulation Set 3.

$$\text{True Model:} \qquad y_i = 1 + 6x_{i1} + 3x_{i2} + 0.5x_{i3} + \epsilon_i,$$
$$\epsilon_i \sim \ iid\ N(0, 0.25).$$

- Predictors $x_{i1}, x_{i2}, \ldots, x_{i6}$: Generated from a uniform distribution on the interval (0,6); i.e., $x_{ij} \sim \ iid\ U(0, 6)$.
- Sample Size: $n = 15$.
- Prior $\pi(r)$: Poisson distribution with a mean of 3 truncated for $r > 6$.

Criterion Selections

r	SIC	SIC_f	SIC_p	SIC_{fp}
0	0	0	0	0
1	0	0	0	0
2	1	5	1	5
3	651	975	747	980
4	122	19	111	15
5	108	1	72	0
6	118	0	69	0

3.2 Time Series

The state-space model is becoming an increasingly popular tool in time series analysis due to its versatility and generality. Shumway (1988, page 173) points out that "[the model] seems to subsume a whole class of special cases of interest in much the same way that linear regression does."

Our second collection of simulations focuses on two important models which are part of the state-space family: the univariate autoregressive model, and the univariate autoregressive model with observation noise. The univariate autoregressive model of order p can be written as

$$y_t = \phi_1 y_{t-1} + \phi_2 y_{t-2} + \ldots + \phi_p y_{t-p} + \epsilon_t, \quad \epsilon_t \sim iid\ N(0, \sigma_o^2).$$

We denote this model as AR(p). The univariate autoregressive model of order p with observation noise can be written as

$$y_t = z_t + v_t, \quad v_t \sim iid\ N(0, \sigma_o^2),$$

$$z_t = \phi_1 z_{t-1} + \phi_2 z_{t-2} + \ldots + \phi_p z_{t-p} + \epsilon_t, \quad \epsilon_t \sim iid\ N(0, \sigma_s^2).$$

We denote this model as ARN(p). (We note that this model is equivalent to a univariate autoregressive moving-average model of orders p and p with parameters that satisfy certain constraints.)

The parameter vectors for the AR(p) and ARN(p) models are, respectively, the $(p+1) \times 1$ and $(p+2) \times 1$ vectors

$$\underline{\theta} = (\phi_1, \phi_2, \ldots, \phi_p, \sigma_s^2)' \text{ and } \underline{\theta} = (\phi_1, \phi_2, \ldots, \phi_p, \sigma_o^2, \sigma_s^2)'.$$

Our goal is to use a collection of observations $y_1, y_2, \ldots, y_n$ to determine an appropriate order p for the autoregression. We will assume the choices for p are 1 through P, corresponding to the candidate models $M_1, M_2, \ldots, M_P$.

To fit the candidate models, the models are expressed in state-space form. The EM algorithm (Shumway and Stoffer, 1982) is used to find the parameter estimates.

The empirical log-likelihood is evaluated using the innovations form of the likelihood. (See Shumway, 1988, page 178.) The observed Fisher information matrix is computed using the algorithm described in Harvey (1989, pages 140 to 142).

We compare the behavior of the four proposed criteria by simulating a setting where one must choose from among the nested candidate models $M_1, M_2, \ldots, M_P$. One thousand sets of data are generated from a true model in the candidate class. For every data set, the P models in the candidate class are fit to the data; SIC, SIC$_f$, SIC$_p$, and SIC$_{fp}$ are evaluated; and the favored model for each criterion is recorded. Over the one thousand data sets, the selections are tabulated, summarized, and reported.

For the prior $\pi(p)$, a Poisson distribution is used where the mean is set equal to the order of the true model. The maximum order P is determined by finding the point at which the distribution is zero out to three significant digits. The distribution is then truncated so that mass is only assigned to the points $p = 1, 2, \ldots, P$.

Four simulation sets are run using various true models and sample sizes. The results are reported in Tables IV to VII.

The first set (Table IV) is based on a setting considered by Hurvich and Tsai (1989) in a simulation to investigate the performance of the "corrected" Akaike information criterion, AIC$_c$. The true model is an AR(2) and the sample size is 23. In one hundred replications with a maximum order of $P = 10$, Hurvich and Tsai report correct selection rates for SIC, AIC$_c$, and AIC of 78%, 80%, and 52% (respectively). In our set, one thousand replications are considered with a maximum order of $P = 7$. Our correct selection rate for SIC is 87.8%, about 10% higher than that reported by Hurvich and Tsai. (This may not only be due to the different maximum orders considered, but also due to the different fitting procedures used and the different definitions of SIC employed. Hurvich and Tsai use $n \ln \hat{\sigma}_s^2$ as a "goodness of fit" term as opposed to $-2 \ln L(\hat{\underline{\theta}} \mid Y_n)$.) The correct selection rate for SIC$_f$ is somewhat higher than that for SIC: 92.8%.

The second set (Table V) is also based on a setting considered by Hurvich and

Tsai (1989). Here, the true model is a nonstationary AR(3). In one hundred replications with a sample size of 15 and a maximum order of P=6, Hurvich and Tsai report disappointing correct selection rates for each of the criteria considered: based on one hundred replications, the rates for SIC, AIC_c, and AIC are listed as 19%, 45%, and 10% (respectively). In our set, we consider a larger sample size of 25, and a higher maximum order of P=10. Based on one thousand replications, SIC obtains a correct selection rate of 78.1%. The rate for SIC_f is again higher than that for SIC at 83.7%.

The third set (Table VI) features an ARN(1) model with a sample size of 15. The maximum order is $P = 6$. Here, SIC obtains a 68.5% correct selection rate, and exhibits a propensity towards overfitting, choosing a model of order three or higher 19.1% of the time. SIC_f obtains an impressive 91.0% correct selection rate, and chooses a model of order three or higher in only one instance out of one thousand.

The preceding simulation sets may create the impression that the inclusion of the Fisher information term in SIC provides additional protection against overfitting at the cost of marginally increasing the likelihood of underfitting. Although the term often behaves in this manner, it would be incorrect to characterize the correction as merely an additional penalty term. To illustrate this point, our last simulation set (Table VII) features a setting where the correction provides protection against both overfitting *and* underfitting.

Here, the true model is an ARN(2). The sample size is 25 and the maximum order is P=7. Without the Fisher information term, SIC obtains only a 46.4% correct selection rate, choosing a model of order one 42.2% of the time and a model of order three or higher 11.4% of the time. On the other hand, SIC_f obtains a 67.4% correct selection rate, choosing a model of order one 26.3% of the time and a model of order three or higher 6.3% of the time. Clearly, the poor performance of SIC in this simulation set is mainly due to its propensity to underfit; a tendency which is markedly reduced by the inclusion of the Fisher information correction.

Note that in all four sets, the use of the Poisson prior results in a noteworthy

improvement in selection performance.

The time series simulations reinforce the same conclusions as the regression simulations. SIC_f consistently outperforms SIC in terms of correct order selections, often by a substantial degree. Also, the use of the correctly specified Poisson prior improves the selection performance of both SIC and SIC_f.

4. Conclusion

The Schwarz information criterion is derived as an asymptotic approximation to a transformation of the Bayesian posterior probability of a candidate model. Through the investigation of the derivation, we have proposed corrected variants of SIC which seemingly improve upon the small to moderate sample-size performance of the criterion. These variants are based on the inclusion of two asymptotically negligible terms: one which involves the observed Fisher information matrix for the model parameters, and the other which depends on a prior over the collection of candidate models.

The inclusion of the first term often improves the performance of SIC by decreasing the likelihood of overfitting without unduly increasing the likelihood of underfitting. In some instances, the term may even decrease the likelihood of underfitting.

The inclusion of the second term has the potential to improve the performance of SIC, depending on the quality and the form of the prior. At the very least, the option of incorporating the second term may be attractive to those who object to the fact that traditional SIC discards all prior information on the grounds that it is asymptotically negligible.

The appeal of the proposed SIC variants lie in their computational simplicity and wide-spread applicability. A thorough Bayesian analysis might opt for an exact evaluation of the posterior probabilities of the various candidate models. We do not view the suggested variants as being competitors to this approach. Rather, we see them as incorporating reasonable corrections which have the potential to improve the effectiveness of traditional SIC in small to moderate sample settings.

TABLE IV: Summary of Simulation Set 4.

True Model: $y_t = 0.99y_{t-1} - 0.80y_{t-2} + \epsilon_t,$

$\epsilon_t \sim \ iid\ N(0, 1.00).$

- Sample Size: $n = 23$.
- Maximum Order P: 7.
- Prior $\pi(p)$: Poisson distribution with a mean of 2 truncated so that its mass is distributed only among the points $p = 1, 2, \ldots, 7$. (The Poisson(2) distribution is zero out to three significant digits when evaluated for arguments exceeding 7.)

Criterion Selections

p	SIC	SIC_f	SIC_p	SIC_{fp}
1	11	18	12	19
2	878	928	933	953
3	59	38	40	23
4	23	13	13	5
5	16	3	2	0
6	9	0	0	0
7	4	0	0	0

TABLE V: Summary of Simulation Set 5.

True Model: $y_t = -0.95y_{t-1} + y_{t-2} + 0.95y_{t-3} + \epsilon_t,$

$\epsilon_t \sim \ iid\ N(0, 1.00).$

- Sample Size: $n = 25$.
- Maximum Order P: 10.
- Prior $\pi(p)$: Poisson distribution with a mean of 3 truncated so that its mass is distributed only among the points $p = 1, 2, \ldots, 10$. (The Poisson(3) distribution is zero out to three significant digits when evaluated for arguments exceeding 10.)

Criterion Selections

p	SIC	SIC_f	SIC_p	SIC_{fp}
1	33	52	27	39
2	36	41	38	44
3	781	837	850	874
4	70	55	59	41
5	29	10	15	2
6	18	4	6	0
7	14	1	3	0
8	7	0	1	0
9	3	0	0	0
10	9	0	1	0

TABLE VI: Summary of Simulation Set 6.

True Model:

$$y_t = z_t + v_t,$$
$$v_t \sim iid\ N(0, 0.20),$$
$$z_t = 0.60 z_{t-1} + \epsilon_t,$$
$$\epsilon_t \sim iid\ N(0, 1.00).$$

- Sample Size: $n = 15$.
- Maximum Order P: 6.
- Prior $\pi(p)$: Poisson distribution with a mean of 1 truncated so that its mass is distributed only among the points $p = 1, 2, \ldots, 6$. (The Poisson(1) distribution is zero out to three significant digits when evaluated for arguments exceeding 6.)

Criterion Selections

p	SIC	SIC_f	SIC_p	SIC_{fp}
1	685	910	895	980
2	124	89	83	20
3	32	1	6	0
4	45	0	12	0
5	40	0	1	0
6	74	0	3	0

TABLE VII: Summary of Simulation Set 7.

True Model:
$$y_t = z_t + v_t,$$
$$v_t \sim iid\ N(0, 0.05),$$
$$z_t = -0.40 z_{t-1} + 0.55 z_{t-2} + \epsilon_t,$$
$$\epsilon_t \sim iid\ N(0, 1.00).$$

- Sample Size: $n = 25$.
- Maximum Order P: 7.
- Prior $\pi(p)$: Poisson distribution with a mean of 2 truncated so that its mass is distributed only among the points $p = 1, 2, \ldots, 7$. (The Poisson(2) distribution is zero out to three significant digits when evaluated for arguments exceeding 7.)

Criterion Selections

p	SIC	SIC_f	SIC_p	SIC_{fp}
1	422	263	435	276
2	464	674	505	702
3	54	51	40	22
4	23	11	11	0
5	18	0	4	0
6	9	0	4	0
7	10	1	1	0

BIBLIOGRAPHY

Akaike, H. (1973). "Information theory and an extension of the maximum likelihood principle," *Second International Symposium on Information Theory* (B. N. Petrov and F. Csaki, Eds.). Budapest, Hungary: Akademia Kiado, 267–281.

Akaike, H. (1974). "A new look at the statistical model identification," *IEEE Transactions on Automatic Control*, AC-19, 716–723.

Akaike, H. (1978). "Time series analysis and control through parametric methods," *Applied Time Series Analysis* (D. Findley, Ed.). New York: Academic Press, 1–24.

Cavanaugh, J. E., Neath, A. A. and Shumway, R. H. (1995). A generalized derivation of the Schwarz information criterion. Technical Report. University of Missouri, Columbia, Department of Statistics. Submitted for publication.

Hannan, E. J. and Quinn, B. G. (1979). "The determination of the order of an autoregression," *Journal of the Royal Statistical Society, B*, 41, 190–195.

Harvey, A. C. (1989). *Forecasting, Structural Time Series Models and the Kalman Filter*, Cambridge University Press, New York.

Haughton, D. M. A. (1988). "On the choice of a model to fit data from an exponential family," *The Annals of Statistics*, 6, 342 – 355.

Hurvich, C. M. and Tsai, C. L. (1989). "Regression and time series model selection in small samples," *Biometrika*, 76, 297–307.

Kashyap, R. L. (1982). "Optimal choice of AR and MA parts in autoregressive moving-average models," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 4, 99–104.

Kass, R. E. and Raftery, A. E. (1995). "Bayes factors," *Journal of the American Statistical Association*, 90, 773–795.

Leonard, T. (1982). "Comments on 'A simple predictive density function,"' by M. LeJeune and G. D. Faulkenberry, *Journal of the American Statistical Association*, 77, 657–658.

Rissanen, J. (1978). "Modeling by shortest data description," *Automatica*, 14, 465–471.

Schwarz, G. (1978). "Estimating the dimension of a model," *The Annals of Statistics*, 6, 461–464.

Shumway, R. H. (1988). *Applied Statistical Time Series Analysis*, Prentice-Hall, New Jersey.

Shumway, R. H. and Stoffer, D. S. (1982). "An approach to time series smoothing and forecasting using the EM algorithm," *Journal of Time Series Analysis*, 3, 253–264.

Stone, M. (1979). "Comments on model selection criteria of Akaike and Schwarz," *Journal of the Royal Statistical Society, B*, 41, 276–278.