GENERALIZING THE DERIVATION OF THE SCHWARZ INFORMATION CRITERION

from *Communications in Statistics – Theory and Methods*, Volume 28, 1999, pages 49–66.

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

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

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

Key Words: Bayes factors; Bayesian analysis; Bayesian information criterion; model selection criterion.

ABSTRACT

The Schwarz information criterion (SIC, BIC, SBC) is one of the most widely known and used tools in statistical model selection. The criterion was derived by Schwarz (1978) to serve as an asymptotic approximation to a transformation of the Bayesian posterior probability of a candidate model. Although the original derivation assumes that the observed data is independent, identically distributed, and arising from a probability distribution in the regular exponential family, SIC has traditionally been used in a much larger scope of model selection problems. To better justify the widespread applicability of SIC, we derive the criterion in a very general framework: one which does not assume any specific form for the likelihood function, but only requires that it satisfies certain non-restrictive regularity conditions.

1. INTRODUCTION

In statistical modeling, an investigator must often choose a suitable model among a collection of viable candidates. Such a determination may be facilitated by the use of a selection criterion, which assigns a score to every fitted model in a candidate class based on some underlying statistical principle. The fitted model which is favored is the one corresponding to the minimum score (or maximum score, depending on how the criterion is defined).

The Schwarz information criterion (SIC, BIC, SBC), introduced by Schwarz (1978) as a competitor to the Akaike (1973, 1974) information criterion (AIC), is one of the most popular and effective of the criteria used for model selection. Schwarz derived SIC to serve as an asymptotic approximation to a transformation of the Bayesian posterior probability of a candidate model. In large-sample settings, the fitted model favored by SIC ideally corresponds to the candidate model which is *a posteriori* most probable; i.e., the model which is rendered most plausible by the data at hand. The computation of SIC is based on the empirical log-likelihood and does not require the specification of priors.

In Bayesian applications, pairwise comparisons between models are often based on Bayes factors. Assuming two candidate models are regarded as equally probable *a priori*, a Bayes factor represents the ratio of the posterior probabilities of the models. The model which is *a posteriori* most probable is determined by whether the Bayes factor is less than or greater than one. In certain settings, model selection based on SIC is roughly equivalent to model selection based on Bayes factors (Kass and Rafferty, 1995; Kass and Wasserman, 1995). Thus, SIC has appeal in many Bayesian modeling problems where priors are hard to set precisely.

Though motivated from a Bayesian perspective, SIC is also used extensively in frequentist applications. Unlike many of its competitors (such as AIC), SIC

has the following consistency property: provided that the family of candidate models under consideration includes the model which generates the data, SIC will asymptotically identify the true model with probability one. In practice, this optimality property is exhibited by the tendency of SIC to select models which are attractively simple.

Schwarz (1978, p. 462) established 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) extended the derivation to the curved exponential family. However, the criterion has long been applied in a more general array of model selection settings. Most notably, it has been successfully used in many time series frameworks, including univariate ARMA modeling, vector AR modeling, and state-space modeling. (See, respectively, Sneek, 1984; Lütkepohl, 1985; Koehler and Murphree, 1988.) Informal generalizations of the criterion, such as those presented by Stone (1979), Kashyap (1982), Leonard (1982), Kass (1983), and Neath and Cavanaugh (1997), suggest that the applicability of SIC extends to a very wide range of modeling settings. However, a rigorous generalization of Schwarz's development seems to be lacking from the literature (cf. Kass and Rafferty, 1995, p. 779).

To better justify the widespread use of SIC, we present a derivation which establishes the validity of the criterion in a very general framework: one which does not assume any specific form for the likelihood function, but only requires that it satisfies certain non-restrictive regularity conditions. Our derivation is presented in the same spirit as the original justification of SIC provided by Schwarz (1978), in that it is based on the same motivation and involves analogous arguments. Thus, our derivation verifies that the broad-based use of SIC is defensible by showing that, under general conditions, the criterion can be developed in the context proposed by Schwarz, rather than by showing that

the criterion can be re-derived from another perspective. (In this regard, Rissanen, 1978, provides an alternate justification of SIC based on the *minimum description length* principle from information theory. Under this principle, the preferred model is the one which permits a reconstruction of the sample utilizing the smallest possible number of bits. Thus, the preferred model corresponds to a codification of the data which is complete yet as concise as possible.)

In the next section, we briefly describe model selection based on SIC. In Section 3, we list the regularity conditions on the likelihood required for the derivation, which follows in Section 4. In Section 5, we discuss a practical modeling framework in which the use of SIC is supported by our derivation, yet not by the original justification provided by Schwarz.

2. MODEL SELECTION BASED ON SIC

Let Y_n denote the observed data. Assume that Y_n is to be described using a model M_k selected from a sequence of candidate models $M_1, \ldots, M_L$, which are not necessarily nested. Assume that each M_k is uniquely parameterized by a vector θ^k, presumed to lie in a parameter space $\Theta(k) \subseteq \Re^k$. Let D_k denote the dimension of M_k; i.e., the number of functionally independent parameters in θ^k which must be estimated.

Let $L(\theta^k \,|\, Y_n)$ represent the likelihood for θ^k based on Y_n. Let $\hat{\theta}_n^k$ denote the estimator of θ^k obtained by maximizing the likelihood $L(\theta^k \,|\, Y_n)$ over $\Theta(k)$. Let $L(\hat{\theta}_n^k \,|\, Y_n)$ denote the corresponding empirical likelihood.

The Schwarz information criterion can be defined as

$$\mathrm{SIC} = -2 \ln L(\hat{\theta}_n^k \,|\, Y_n) + D_k \ln n.$$

In settings where the sample size n is ambiguous, it is often recommended that n be chosen to grow at the same rate as the Hessian of $-\ln L(\theta^k \,|\, Y_n)$ (see, e.g., Kass and Rafferty, 1995, p. 779). (This recommendation is supported by

our forthcoming derivation, where division by n must ensure the convergence of the Hessian to a positive definite matrix.)

In practice, SIC is computed for each of the models $M_1, \ldots, M_L$, and the model corresponding to the minimum value of SIC is selected. In our derivation, we show that SIC provides a large-sample approximation to

$$-2 \ln P(M_k \mid Y_n) - 2 \ln\{h(Y_n)\},$$

where $P(M_k \mid Y_n)$ denotes the posterior probability of the model M_k given the data Y_n, and $h(Y_n)$ denotes the marginal density of Y_n. Since $h(Y_n)$ does not depend on M_k, the model associated with the minimum value of SIC should therefore correspond to the model with the highest posterior probability among $M_1, \ldots, M_L$.

As previously mentioned, the original derivation of SIC by Schwarz (1978) justifies the criterion in a setting where Y_n consists of independent, identically distributed data, and $L(\theta^k \mid Y_n)$ belongs to the regular exponential family. Our generalized derivation extends the justification beyond this context, to applications where $L(\theta^k \mid Y_n)$ need only satisfy a set of non-restrictive regularity conditions. These conditions are listed in the next section.

3. REGULARITY CONDITIONS ON THE LIKELIHOOD

Let

$$V_n(\theta^k) = -\frac{1}{n} \ln L(\theta^k \mid Y_n) \qquad \text{and} \qquad W_n(\theta^k) = E\{V_n(\theta^k)\}.$$

We assume the following regularity conditions on $V_n(\theta^k)$ and $W_n(\theta^k)$.

(1) $V_n(\theta^k)$ has first- and second-order derivatives which are continuous over $\Theta(k)$. Let

$$V_n^{(1)}(\theta^k) = \frac{\partial V_n(\theta^k)}{\partial \theta^k} \qquad \text{and} \qquad V_n^{(2)}(\theta^k) = \frac{\partial^2 V_n(\theta^k)}{\partial \theta^k \, \partial \theta^{k\prime}}.$$

(2) $V_n(\theta^k)$ has a unique global minimum at $\hat{\theta}_n^k$, where $\hat{\theta}_n^k$ is an interior point of $\Theta(k)$.

(3) As $n \to \infty$, $W_n(\theta^k)$ converges to a function $W(\theta^k)$ that has first- and second-order derivatives which are continuous over $\Theta(k)$. The convergence is uniform in θ^k over $\Theta(k)$. Let

$$W^{(2)}(\theta^k) = \frac{\partial^2 W(\theta^k)}{\partial \theta^k \, \partial \theta^{k\,\prime}}.$$

(4) $W(\theta^k)$ has a unique global minimum at θ_*^k, where θ_*^k is an interior point of $\Theta(k)$.

(5) $V_n(\theta^k) \to W(\theta^k)$ almost surely as $n \to \infty$, uniformly in θ^k over $\Theta(k)$.

(6) $V_n^{(2)}(\theta^k) \to W^{(2)}(\theta^k)$ almost surely as $n \to \infty$, uniformly in θ^k over $\Theta(k)$.

(7) $W^{(2)}(\theta^k)$ is positive definite in a neighborhood of θ_*^k. Over this neighborhood, the eigenvalues of $W^{(2)}(\theta^k)$ are bounded and bounded away from zero.

Note that the preceding conditions, in particular (2), (4), and (5), imply that $\hat{\theta}_n^k$ converges almost surely to θ_*^k as $n \to \infty$.

The preceding conditions are characteristic of those which arise in establishing the asymptotic normality of the maximum likelihood estimator when the underlying data is not necessarily independent and the model being fit to the data is possibly misspecified. See, for example, Ljung and Caines (1979), Stoffer and Wall (1991), Cavanaugh and Shumway (1997). Demonstrating the asymptotic normality of $\sqrt{n}\,(\hat{\theta}_n^k - \theta_*^k)$ often proceeds by taking a first-order expansion of $0 = V_n^{(1)}(\hat{\theta}_n^k)$ about θ_*^k, using the convergence of $V_n^{(2)}(\theta_*^k)$ to $W^{(2)}(\theta_*^k)$, and establishing the asymptotic normality of $\sqrt{n}\,V_n^{(1)}(\theta_*^k)$. Our justification of SIC requires a set of conditions which will allow us to take a

second-order expansion of $V_n(\theta^k)$ about $\hat{\theta}_n^k$, establish asymptotic bounds for the second-order term over a neighborhood of θ_*^k, and characterize the convergence of the expansion. Our development does not require assumptions regarding the model structure or the data dependency, generally utilized in demonstrations of asymptotic normality pertaining to the behavior of $\sqrt{n}\, V_n^{(1)}(\theta_*^k)$. Thus, the conditions we require are less technical than those required to demonstrate asymptotic normality.

4. GENERALIZED DERIVATION OF SIC

The motivation behind SIC can be outlined as follows.

Suppose that the conditional density of the data Y_n given both M_k (the k^{th} candidate model) and θ^k (the parameter vector for M_k) is denoted by $f(Y_n \,|\, (M_k,\, \theta^k))$. Let $\pi(M_k)$ denote a discrete prior over the set of the candidate models which assigns a prior probability to each of the models $M_1, \ldots, M_L$. Let $g(\theta^k \,|\, M_k)$ denote a prior on the parameter vector θ^k given the model M_k.

We assume that the prior $\pi(M_k)$ assigns a positive probability to each model M_k, $1 \leq k \leq L$. Further, we assume that the prior $g(\theta^k \,|\, M_k)$ is bounded over $\Theta(k)$, and is bounded away from zero over a neighborhood of θ_*^k.

Applying Bayes' theorem, the joint posterior of $(M_k,\, \theta^k)$ can be written as

$$f((M_k,\, \theta^k) \,|\, Y_n) = \frac{\pi(M_k)\, g(\theta^k \,|\, M_k)\, f(Y_n \,|\, (M_k,\, \theta^k))}{h(Y_n)}. \tag{4.1}$$

A Bayesian model selection procedure could then be based on choosing the model M_k which is *a posteriori* most probable; i.e., choosing the M_k which maximizes $P(M_k \,|\, Y_n)$, where

$$P(M_k \,|\, Y_n) = \int f((M_k,\, \theta^k) \,|\, Y_n)\, d\theta^k. \tag{4.2}$$

Since $f(Y_n \,|\, (M_k,\, \theta^k)) = L(\theta^k \,|\, Y_n)$, we can use (4.1) to write (4.2) as

$$P(M_k \,|\, Y_n) = h(Y_n)^{-1}\, \pi(M_k) \int L(\theta^k \,|\, Y_n)\, g(\theta^k \,|\, M_k)\, d\theta^k. \tag{4.3}$$

Now consider minimizing $-2 \ln P(M_k \,|\, Y_n)$ as an alternative to maximizing $P(M_k \,|\, Y_n)$. From (4.3), we have

$$\begin{aligned} -2 \ln P(M_k \,|\, Y_n) &= 2 \ln \{h(Y_n)\} - 2 \ln \{\pi(M_k)\} \\ &\quad -2 \ln \left\{ \int L(\theta^k \,|\, Y_n)\, g(\theta^k \,|\, M_k)\, d\theta^k \right\}. \end{aligned} \tag{4.4}$$

The first of the three terms on the right-hand side of (4.4) does not depend on the model M_k, and is therefore irrelevant for the purpose of model selection. Consider the third term on the right-hand side of (4.4). Through the application of two lemmas, we will demonstrate that this term is asymptotically bounded between

$$-2 \ln L(\hat{\theta}_n^k \,|\, Y_n) + D_k \ln n + R_2(D_k) \tag{4.5}$$

and

$$-2 \ln L(\hat{\theta}_n^k \,|\, Y_n) + D_k \ln n + R_1(D_k), \tag{4.6}$$

where $R_2(D_k) < R_1(D_k)$, and $R_1(D_k), R_2(D_k)$ do not depend on n. (Recall that D_k denotes the dimension of M_k.) If we then ignore the terms in (4.5), (4.6), and (4.4) which do not grow in magnitude as $n \to \infty$, we obtain from these expressions the following approximate large-sample relation:

$$-2 \ln P(M_k \,|\, Y_n) - 2 \ln \{h(Y_n)\} \approx -2 \ln L(\hat{\theta}_n^k \,|\, Y_n) + D_k \ln n.$$

This motivates the use of SIC for model selection, where we choose the candidate model M_k by finding the fitted model which minimizes

$$\text{SIC} = -2 \ln L(\hat{\theta}_n^k \,|\, Y_n) + D_k \ln n.$$

In large-sample applications, this fitted model should correspond to the candidate model which maximizes $P(M_k \,|\, Y_n)$.

We now state and prove the two lemmas which will be used to establish the asymptotic bounds (4.5) and (4.6).

Lemma 1

For some positive constants λ_1 and λ_2, the following will hold for θ^k in a neighborhood of θ_*^k provided that n is sufficiently large:

$$V_n(\hat{\theta}_n^k) + \frac{\lambda_2}{2}(\theta^k - \hat{\theta}_n^k)'(\theta^k - \hat{\theta}_n^k)$$
$$\leq V_n(\theta^k) \leq$$
$$V_n(\hat{\theta}_n^k) + \frac{\lambda_1}{2}(\theta^k - \hat{\theta}_n^k)'(\theta^k - \hat{\theta}_n^k).$$

Proof:

Expanding $V_n(\theta^k)$ about $\hat{\theta}_n^k$, the point at which $V_n(\theta^k)$ attains its global minimum, yields

$$\begin{aligned} V_n(\theta^k) &= V_n(\hat{\theta}_n^k) + (\theta^k - \hat{\theta}_n^k)' V_n^{(1)}(\hat{\theta}_n^k) \\ &\quad + \frac{1}{2}(\theta^k - \hat{\theta}_n^k)' V_n^{(2)}(\gamma_n^k)(\theta^k - \hat{\theta}_n^k) \\ &= V_n(\hat{\theta}_n^k) + \frac{1}{2}(\theta^k - \hat{\theta}_n^k)' V_n^{(2)}(\gamma_n^k)(\theta^k - \hat{\theta}_n^k), \end{aligned} \tag{4.7}$$

where γ_n^k is between θ^k and $\hat{\theta}_n^k$.

Let $\lambda_n^{min}(\theta^k)$ and $\lambda_n^{max}(\theta^k)$ represent, respectively, the smallest and largest eigenvalues of the matrix $V_n^{(2)}(\theta^k)$. For any θ^k in $\Theta(k)$, we have that

$$\begin{aligned} (\theta^k - \hat{\theta}_n^k)'(\theta^k - \hat{\theta}_n^k)\lambda_n^{min}(\gamma_n^k) &\leq (\theta^k - \hat{\theta}_n^k)' V_n^{(2)}(\gamma_n^k)(\theta^k - \hat{\theta}_n^k) \\ &\leq (\theta^k - \hat{\theta}_n^k)'(\theta^k - \hat{\theta}_n^k)\lambda_n^{max}(\gamma_n^k). \end{aligned} \tag{4.8}$$

Now the regularity conditions imply that for every θ^k in $\Theta(k)$, $\lambda_n^{min}(\theta^k)$ and $\lambda_n^{max}(\theta^k)$ respectively converge, almost surely, to the smallest and largest eigenvalues of $W^{(2)}(\theta^k)$. Within a neighborhood of θ_*^k, $W^{(2)}(\theta^k)$ is positive definite and has its eigenvalues bounded between two positive constants. It is therefore possible to choose an n_1 and a neighborhood $N(\theta_*^k)$ of θ_*^k such that for some constants λ_1, λ_2 satisfying $0 < \lambda_2 < \lambda_1 < \infty$, we have

$$\lambda_2 \leq \inf_{n>n_1}\left\{\inf_{\theta^k \in N(\theta_*^k)} \lambda_n^{min}(\theta^k)\right\} \quad \text{and} \quad \lambda_1 \geq \sup_{n>n_1}\left\{\sup_{\theta^k \in N(\theta_*^k)} \lambda_n^{max}(\theta^k)\right\}. \tag{4.9}$$

Thus, by (4.8) and (4.9), for all $n > n_1$ and for $\gamma_n^k \in N(\theta_*^k)$, it follows that

$$\begin{aligned} 0 \leq (\theta^k - \hat{\theta}_n^k)' (\theta^k - \hat{\theta}_n^k) \lambda_2 &\leq (\theta^k - \hat{\theta}_n^k)' V_n^{(2)}(\gamma_n^k) (\theta^k - \hat{\theta}_n^k) \\ &\leq (\theta^k - \hat{\theta}_n^k)' (\theta^k - \hat{\theta}_n^k) \lambda_1. \end{aligned} \tag{4.10}$$

Now since $\hat{\theta}_n^k$ converges to θ_*^k almost surely, $\hat{\theta}_n^k \in N(\theta_*^k)$ for all n exceeding some n_2. Moreover, since γ_n^k is between $\hat{\theta}_n^k$ and θ^k, whenever $\hat{\theta}_n^k \in N(\theta_*^k)$ and $\theta^k \in N(\theta_*^k)$, it must also hold that $\gamma_n^k \in N(\theta_*^k)$. Thus by (4.7) and (4.10), the bounds stated in the lemma will hold for all $\theta^k \in N(\theta_*^k)$, provided that $n > n_0 = \max\{n_1, n_2\}$. □

Lemma 2

Consider two sequences of positive random variables $\{T_n\}$ *and* $\{U_n\}$ *and a convergent positive sequence* $\{\rho_n\}$ *defined such that* $[T_n \geq U_n]$ *whenever* $[U_n > \rho_n]$. *Suppose there exists two positive constants* γ *and* ε *such that*

$$P[(T_n - \rho_n) \geq \gamma] \geq \varepsilon \quad \textit{for all} \quad n.$$

For any positive constant δ, *the following holds for sufficiently large* n:

$$\{\ln E(T_n^n) - \ln E(U_n^n)\} > -\delta.$$

Proof:

Let

$$R_n = \begin{cases} U_n & \text{for } U_n \leq \rho_n \\ T_n & \text{for } U_n > \rho_n. \end{cases}$$

Consider the difference $(R_n^n - T_n^n)$. If $R_n = T_n$, then this difference is zero. If $R_n = U_n$ and $U_n \neq T_n$, then we must have $R_n = U_n \leq \rho_n$, and this difference cannot exceed ρ_n^n. Thus, it follows that $R_n^n - T_n^n \leq \rho_n^n$. This relation implies

$$E(R_n^n) \leq E(T_n^n) + \rho_n^n,$$

which leads to

$$\frac{E(R_n^n)}{E(T_n^n)} \leq 1 + \frac{\rho_n^n}{E(T_n^n)}. \tag{4.11}$$

Now using the Markov inequality (Billingsley, p. 74, 1986), one can show that for any positive sequence $\{c_n\}$,

$$\{E(T_n^n)\}^{1/n} \geq c_n \ \{P\left[T_n \geq c_n\right]\}^{1/n},$$

meaning

$$\frac{\{E(T_n^n)\}^{1/n}}{\rho_n} \geq \frac{c_n}{\rho_n} \ \{P\left[T_n \geq c_n\right]\}^{1/n}. \tag{4.12}$$

Choose the sequence $\{c_n\}$ by setting $c_n = \rho_n + \gamma$. For this c_n, we have that

$$\lim_{n\to\infty} \frac{c_n}{\rho_n} > 1, \tag{4.13}$$

and since

$$P\left[T_n \geq c_n\right] = P\left[(T_n - \rho_n) \geq \gamma\right] \geq \varepsilon > 0,$$

we also have that

$$\lim_{n\to\infty} \{P\left[T_n \geq c_n\right]\}^{1/n} = 1. \tag{4.14}$$

Thus, by (4.12), (4.13), and (4.14), there exists an $\alpha > 0$ and an n_0 such that whenever $n > n_0$,

$$\frac{\{E(T_n^n)\}^{1/n}}{\rho_n} > 1 + \alpha.$$

This implies

$$\lim_{n\to\infty} \left\{1 + \frac{\rho_n^n}{E(T_n^n)}\right\} = 1. \tag{4.15}$$

Now from (4.11) and (4.15), for any $\delta > 0$, we can find an n_1 such that for all $n > n_1$,

$$\frac{E(R_n^n)}{E(T_n^n)} < \exp(\delta).$$

Yet since $R_n \geq U_n$, the preceding implies that whenever $n > n_1$,

$$\frac{E(T_n^n)}{E(U_n^n)} > \exp(-\delta),$$

or equivalently,

$$\{\ln E(T_n^n) - \ln E(U_n^n)\} > -\delta. \qquad \square$$

Note the result of Lemma 2 implies $\lim_{n\to\infty}\{\ln E(T_n^n) - \ln E(U_n^n)\} \geq 0$, provided that the limit exists.

We now use our lemmas to establish the asymptotic bounds (4.5) and (4.6). Let

$$\begin{aligned} Z_n(\theta^k) &= \exp\{-V_n(\theta^k)\}, \\ X_{1,n}(\theta^k) &= \exp\left\{-V_n(\hat{\theta}_n^k) - \frac{\lambda_1}{2}(\theta^k - \hat{\theta}_n^k)'(\theta^k - \hat{\theta}_n^k)\right\}, \\ X_{2,n}(\theta^k) &= \exp\left\{-V_n(\hat{\theta}_n^k) - \frac{\lambda_2}{2}(\theta^k - \hat{\theta}_n^k)'(\theta^k - \hat{\theta}_n^k)\right\}. \end{aligned}$$

By application of Lemma 1, we can find an n_0 and a neighborhood $N(\theta_*^k)$ of θ_*^k such that for all $n > n_0$ and for all $\theta^k \in N(\theta_*^k)$, we have

$$X_{1,n}(\theta^k) \leq Z_n(\theta^k) \leq X_{2,n}(\theta^k). \tag{4.16}$$

Now, consider applying Lemma 2 to the positive random variables $U_n = X_{1,n}(\theta^k)$ and $T_n = Z_n(\theta^k)$, where the relevant probability measure is $g(\theta^k \mid M_k)$. Define

$$\begin{aligned} X_1(\theta^k) &= \exp\left\{-W(\theta_*^k) - \frac{\lambda_1}{2}(\theta^k - \theta_*^k)'(\theta^k - \theta_*^k)\right\}, \\ Z(\theta^k) &= \exp\{-W(\theta^k)\}. \end{aligned}$$

The regularity conditions indicate that both $X_1(\theta^k)$ and $Z(\theta^k)$ attain a common global maximum at θ_*^k. The conditions also ensure

$$\begin{aligned} &X_{1,n}(\theta^k) \to X_1(\theta^k) \text{ almost surely, uniformly in } \theta^k, \text{ and} \\ &Z_n(\theta^k) \to Z(\theta^k) \text{ almost surely, uniformly in } \theta^k. \end{aligned}$$

Thus, there exists an n_1 and a convergent positive sequence $\{\rho_n\}$ such that

$$\left\{\theta^k \mid X_{1,n}(\theta^k) > \rho_n \text{ for } n > n_1\right\} \subseteq N(\theta_*^k). \tag{4.17}$$

Moreover, n_1 and $\{\rho_n\}$ can be chosen so that for some positive constants γ and ε, we have

$$P\left[\left(Z_n(\theta^k) - \rho_n\right) \geq \gamma\right] \geq \varepsilon \quad \text{for all} \quad n > n_1. \tag{4.18}$$

Now take $n_2 = \max\{n_0, n_1\}$. Note that whenever both $X_{1,n}(\theta^k) > \rho_n$ and $n > n_2$, (4.17) will apply, and the ordering (4.16) will hold: i.e., we will have $X_{1,n}(\theta^k) \leq Z_n(\theta^k)$. By virtue of this fact and (4.18), application of Lemma 2 (with $U_n = X_{1,n}(\theta^k)$, $T_n = Z_n(\theta^k)$, $n > n_2$) guarantees that for any $\delta_* > 0$,

$$\left[\ln E\{Z_n^n(\theta^k)\} - \ln E\{X_{1,n}^n(\theta^k)\}\right] > -\frac{\delta_*}{2} \tag{4.19}$$

provided that n is sufficiently large. One can appeal to Lemma 2 in a similar manner to establish that

$$\left[\ln E\{X_{2,n}^n(\theta^k)\} - \ln E\{Z_n^n(\theta^k)\}\right] > -\frac{\delta_*}{2} \tag{4.20}$$

when n is sufficiently large.

Thus, by (4.19) and (4.20), there exists an n_* such that for all $n > n_*$,

$$\ln E\{X_{1,n}^n(\theta^k)\} - \frac{\delta_*}{2} < \ln E\{Z_n^n(\theta^k)\} < \ln E\{X_{2,n}^n(\theta^k)\} + \frac{\delta_*}{2},$$

or equivalently,

$$-2\ln E\{X_{2,n}^n(\theta^k)\} - \delta_* < -2\ln E\{Z_n^n(\theta^k)\} < -2\ln E\{X_{1,n}^n(\theta^k)\} + \delta_*. \tag{4.21}$$

The relation (4.21) will lead to the justification of the asymptotic bounds (4.5) and (4.6). Note that the middle term in (4.21) can be written as

$$\begin{aligned} -2\ln E\{Z_n^n(\theta^k)\} &= -2\ln \int \left[\exp\{-V_n(\theta^k)\}\right]^n g(\theta^k \mid M_k)\, d\theta^k \\ &= -2\ln \int L(\theta^k \mid Y_n)\, g(\theta^k \mid M_k)\, d\theta^k. \end{aligned} \tag{4.22}$$

We will reduce and bound the left-hand and right-hand terms in (4.21) by utilizing the previously mentioned boundedness requirements on the prior $g(\theta^k \mid M_k)$. Specifically, we assume that for some constants $0 < b \leq B < \infty$,

$$0 \leq g(\theta^k \mid M_k) \leq B \text{ for all } \theta^k \in \Theta(k), \tag{4.23}$$

$$b \leq g(\theta^k \mid M_k) \text{ for all } \theta^k \text{ within a neighborhood } N_*(\theta_*^k) \text{ of } \theta_*^k. \tag{4.24}$$

Further, we assume that the n_* which ensures (4.21) holds whenever $n > n_*$ is large enough to also ensure $\hat{\theta}_n^k \in N_*(\theta_*^k)$ whenever $n > n_*$.

For the left-hand term in (4.21), utilizing (4.23), we have

$$\begin{aligned}
& -2 \ln E\{X_{2,n}^n(\theta^k)\} \\
= \; & -2 \ln \int_{\Theta(k)} \left[\exp\left\{ -V_n(\hat{\theta}_n^k) - \frac{\lambda_2}{2} (\theta^k - \hat{\theta}_n^k)' (\theta^k - \hat{\theta}_n^k) \right\} \right]^n g(\theta^k \mid M_k)\, d\theta^k \\
\geq \; & -2 \ln \int_{\Re^k} B \exp\left\{ -n V_n(\hat{\theta}_n^k) - \frac{n\lambda_2}{2} (\theta^k - \hat{\theta}_n^k)' (\theta^k - \hat{\theta}_n^k) \right\} d\theta^k \\
= \; & 2n\, V_n(\hat{\theta}_n^k) - 2 \ln B \\
& \quad -2 \ln \left[\left(\frac{2\pi}{n\lambda_2} \right)^{D_k/2} \int_{\Re^k} \left(\frac{n\lambda_2}{2\pi} \right)^{D_k/2} \exp\left\{ -\frac{1}{2} \frac{(\theta^k - \hat{\theta}_n^k)' (\theta^k - \hat{\theta}_n^k)}{(1/n\lambda_2)} \right\} d\theta^k \right] \\
= \; & -2 \ln L(\hat{\theta}_n^k \mid Y_n) - 2 \ln B - 2 \ln \left(\frac{2\pi}{n\lambda_2} \right)^{D_k/2} \\
= \; & -2 \ln L(\hat{\theta}_n^k \mid Y_n) + D_k \ln n + R_2(D_k), \qquad (4.25)
\end{aligned}$$

where $R_2(D_k) = D_k \ln \lambda_2 - D_k \ln 2\pi - 2 \ln B$.

For the right-hand term in (4.21), we must argue the existence of a positive lower bound for the integral

$$\int_{N_*} \left(\frac{n\lambda_1}{2\pi} \right)^{D_k/2} \exp\left\{ -\frac{1}{2} \frac{(\theta^k - \hat{\theta}_n^k)' (\theta^k - \hat{\theta}_n^k)}{(1/n\lambda_1)} \right\} d\theta^k, \qquad (4.26)$$

where N_* denotes the neighborhood $N_*(\theta_*^k)$ referenced in (4.24). Note that the integrand in (4.26) is a D_k-dimensional Gaussian density with mean $\hat{\theta}_n^k$ and variance/covariance matrix $(1/n\lambda_1)$I. As $n \to \infty$, $\hat{\theta}_n^k$ converges almost surely to θ_*^k and $(1/n\lambda_1)$ converges to 0; as a result, the density becomes increasingly concentrated about θ_*^k and the integral (4.26) converges almost surely to 1. Moreover, since $\hat{\theta}_n^k \in N_*(\theta_*^k)$ whenever $n > n_*$, there exists an $\upsilon > 0$ (depending on $N_*(\theta_*^k)$ and λ_1) such that (4.26) is no less than υ for any $n > n_*$. Utilizing (4.24), we therefore have for all $n > n_*$

$$\begin{aligned}
& -2 \ln E\{X_{1,n}^n(\theta^k)\} \\
= \; & -2 \ln \int_{\Theta(k)} \left[\exp\left\{ -V_n(\hat{\theta}_n^k) - \frac{\lambda_1}{2} (\theta^k - \hat{\theta}_n^k)' (\theta^k - \hat{\theta}_n^k) \right\} \right]^n g(\theta^k \mid M_k)\, d\theta^k
\end{aligned}$$

$$
\begin{aligned}
&\leq -2\ln \int_{N_*} b\ \exp\left\{-nV_n(\hat{\theta}_n^k) - \frac{n\lambda_1}{2}(\theta^k - \hat{\theta}_n^k)'(\theta^k - \hat{\theta}_n^k)\right\} d\theta^k \\
&= 2n\,V_n(\hat{\theta}_n^k) - 2\ln b \\
&\quad -2\ln\left[\left(\frac{2\pi}{n\lambda_1}\right)^{D_k/2} \int_{N_*} \left(\frac{n\lambda_1}{2\pi}\right)^{D_k/2} \exp\left\{-\frac{1}{2}\frac{(\theta^k - \hat{\theta}_n^k)'(\theta^k - \hat{\theta}_n^k)}{(1/n\lambda_1)}\right\} d\theta^k\right] \\
&\leq -2\ln L(\hat{\theta}_n^k \,|\, Y_n) - 2\ln b - 2\ln\left(\frac{2\pi}{n\lambda_1}\right)^{D_k/2} - 2\ln \upsilon \\
&= -2\ln L(\hat{\theta}_n^k \,|\, Y_n) + D_k \ln n + R_1(D_k), \qquad (4.27)
\end{aligned}
$$

where $R_1(D_k) = D_k \ln \lambda_1 - D_k \ln 2\pi - 2\ln b - 2\ln \upsilon$.

Thus, by (4.21), (4.25), and (4.27), expression (4.22) is bounded between (4.5) and (4.6) whenever $n > n_*$. This completes the proof.

5. CONCLUSION

In addition to extending Schwarz's derivation to a large collection of likelihoods, our derivation features other important generalizations of the original development. Unlike Schwarz's justification, ours does not assume the underlying data is independent and identically distributed. Also, the asymptotic arguments in Schwarz's derivation assume that the data (exhibited in the form of a sufficient statistic) is *fixed* while the sample size goes to infinity (see Schwarz, 1978, Proposition, p. 462). This simplifies the derivation, since the fixed data translates to a fixed set of parameter estimates for the fitted model. In our justification, such an assumption is not employed.

As mentioned in the introduction, SIC has long been successfully used as a selection criterion in time series applications. To illustrate a setting which is within the scope of our derivation yet beyond the scope of the original justification, consider the state-space framework. The state-space model is becoming increasingly popular in time series analysis due to its versatility and generality. Shumway (1988, p. 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." The successful application of SIC in the state-space framework is illustrated in Koehler and Murphree (1988) and Neath and Cavanaugh (1997), among others.

Estimation in the state-space setting is routinely accomplished by maximizing the Gaussian log-likelihood in its innovation form (cf. Shumway, 1988, p. 178). Ljung and Caines (1979) present an asymptotic theory which can be used to justify the strong consistency and asymptotic normality of the Gaussian maximum likelihood estimator, even in the absence of normally distributed errors or a correctly specified model. (See Caines, 1988, p. 499; Harvey, 1989, pp. 128-130.) Our regularity conditions in Section 3 are implied by the assumptions under which this theory holds. This can be easily verified, since our notation is quite similar to that used by Ljung and Caines (1979). Note that our regularity condition (1) is implied by the initial requirement in their last subsection of Section 2 (see also their definition (2.3)); our condition (2) is assumed in the statement of their Theorem 1; and our conditions (3), (4), and (7) follow from the assumptions in the statement of their Corollary to Theorem 1. Our conditions (5) and (6) are established utilizing their assumptions (2.4) through (2.10) along with (3.11); see their results (3.2) and (A.11).

The asymptotic theory of Ljung and Caines (1979) therefore encompasses the regularity conditions used in our derivation: if the requirements of the theory are accepted to justify maximum likelihood estimation in the state-space setting, those same requirements will justify the application of SIC. Our derivation should similarly support the use of SIC in other frameworks where our regularity conditions are enveloped by a theoretical structure conducive to maximum likelihood estimation.

ACKNOWLEDGEMENTS

The authors wish to thank the Associate Editor for carefully reading the original manuscript, and for preparing a helpful and constructive review which served to greatly improve the exposition and content. The authors also wish to thank Professor Robert H. Shumway for his insights and contributions. The work of the first author was supported by a grant from the National Science Foundation (DMS-9704436).

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.

Billingsley, P. (1986). *Probability and Measure (Second Edition)*, New York: John Wiley.

Caines, P. E. (1988). *Linear Stochastic Systems*, New York: John Wiley.

Cavanaugh, J. E. and Shumway, R. H. (1997). "A bootstrap variant of AIC for state-space model selection," *Statistica Sinica*, 7, 473–496.

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

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.

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. (1993). "Bayes factors in practice," *The Statistician*, 42, 551–560.

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

Kass, R. E. and Wasserman, L. (1995). "A reference Bayesian test for nested hypotheses and its relationship to the Schwarz criterion," *Journal of the American Statistical Association*, 90, 928–934.

Koehler, A. B. and Murphree, E. S. (1988). "A comparison of the Akaike and Schwarz criteria for selecting model order," *Applied Statistics*, 37, 187–195.

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.

Ljung, L. and Caines, P. E. (1979). "Asymptotic normality of prediction error estimators for approximate system models," *Stochastics*, 3, 29–46.

Lütkepohl, H. (1985). "Comparison of criteria for estimating the order of a vector autoregressive process," *Journal of Time Series Analysis*, 6, 35–52.

Neath, A. A. and Cavanaugh, J. E. (1997). "Regression and time series model selection using variants of the Schwarz information criterion," *Communications in Statistics – Theory and Methods*, 26, 559–580.

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*, New Jersey: Prentice-Hall.

Sneek, J. M. (1984). *Modelling Procedures for Univariate Economic Time Series*, Amsterdam: Free University Press.

Stoffer, D. S. and Wall, K. D. (1991). "Bootstrapping state-space models: Gaussian maximum likelihood estimation and the Kalman filter," *Journal of the American Statistical Association*, 86, 1024–1033.

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