Revista de Análisis Económico, Vol. 28, Nº 1, pp. 47-64 (Abril 2013)

DOES THE BIC ESTIMATE AND FORECAST BETTER THAN THE AIC?*

¿ESTIMA Y PREDICE EL BIC MEJOR QUE EL AIC?

CARLOS A. MEDEL
Central Bank of Chile

SERGIO C. SALGADO
University of Minnesota

Abstract

We test two questions: (i) Is the Bayesian Information Criterion (BIC) more parsimonious than Akaike Information Criterion (AIC)? and (ii) Is BIC better than AIC for forecasting purposes? By using simulated data, we provide statistical inference of both hypotheses individually and then jointly with a multiple hypotheses testing procedure to control better for type-I error. Both testing procedures deliver the same result: The BIC shows an in- and out-of-sample superiority over AIC only in a long-sample context.

Keywords: *AIC, BIC, information criteria, time-series models, overfitting, forecast comparison, joint hypothesis testing.*

JEL Classification: *C22, C51, C52, C53.*

Resumen

Contestamos dos preguntas: (i) ¿Es el criterio de información bayesiano (BIC) más parsimonioso que el criterio de información de Akaike (AIC)? y (ii) ¿Es el BIC mejor que el AIC para fines predictivos? Mediante el uso de

* We thank Yan Carrière-Swallow, Carlos García (Editor), Mario Giarda, Michael Pedersen, Pablo Pincheira, Felipe Saffie, and an anonymous referee for their kind help and comments. We also thank the comments of seminar participants at Central Bank of Chile. Any errors or omissions are responsibility of the authors. The views and ideas expressed in this paper do not necessarily represent those of the Central Bank of Chile or its authorities. E-mails: cmedel@bcentral.cl (corresponding author); salga101@um.edu.

datos simulados, proporcionamos inferencia estadística con respecto a ambas hipótesis de manera individual y luego conjunta con un procedimiento de pruebas de hipótesis múltiples para controlar mejor el error tipo I. Ambas pruebas entregan el mismo resultado: el BIC muestra una superioridad dentro y fuera de la muestra sobre el AIC sólo en un contexto de muestra larga.

Palabras clave: *AIC, BIC, criterios de información, modelos de series de tiempo, sobreajuste, evaluación de proyecciones, prueba de hipótesis conjunta.*

Clasificación JEL: *C22, C51, C52, C53.*

1. INTRODUCTION

The success of many economic decisions relies on the forecast accuracy of certain key variables. Often, economic theory is not clear about the relationship between two or more variables, and a data snooping analysis is performed prior to modeling. A useful model-building procedure in circumstances with lower levels of knowledge about the fundamental variables behind the dynamics of the true data generating process is the use of the so-called information criteria –measures of goodness of fit based on the log likelihood function (ℓ), the number of regressors (p), and the sample size (T). However, is not clear when –especially sample size, given the different asymptotic behavior– their model-based forecast may dominate.

The aim of this paper is to test two questions: (i) Is the Bayesian Information Criterion (BIC) more parsimonious than the Akaike Information Criterion (AIC)? and (ii) Is BIC better than AIC for forecasting purposes[1]? We provide statistical inference on both hypotheses individually with a significance test –based on Diebold and Mariano (1995), and West (1996)– and jointly with a multiple hypotheses test following White (2000) approach with some considerations of Hansen's (2005) superior predictive ability test[2]. The exercise consists in the simulation of a large stationary dataset, containing 1,000 series generated by an autoregressive process (AR) of order $p = 6$. We then compute and compare the order determined by each criteria, which often differs from the true order. Then, for each series, we generate 1-step ahead forecasts and evaluate their accuracy based on the root of the squared forecast error (RSFE). We perform this exercise several times, each one considering a different sample size of the same 1,000 series, to basically account for the different asymptotic behavior of each information criteria.

---

[1] More details on derivation and comparison between both criterion can be found in Akaike (1974), Shibata (1976), Rissasen (1978), Schwarz (1978), Stone (1979), Lütkepohl (1985), Koehler and Murphree (1988), Zucchini (2000), Kuha (2004), and Weakliem (2004).

[2] These procedures are related to those used in Wolak (1987, 1989), and Sullivan, Timmermann, and White (1999). We use a version closer to that used in Pincheira (2011a, 2011b, 2012). A recent survey can be found in Corradi and Distaso (2011).

The AIC is defined as $T \cdot \log \ell + 2p^{AIC}$, while the BIC as $T \cdot \log \ell + p^{BIC} \cdot \log T$. A lower score reflects a better fit. The difference in the chosen lag length comes exclusively from the penalty term imposed on the number of regressors of the fitted model. As is shown in Granger and Jeon (2004), it is expected for a sample size $T \geq 8$ and a given value of ℓ that $p^{BIC} \leq p^{AIC}$. The results reveal the existence of (in-sample) overfitting by AIC compared with BIC across different estimation sample sizes. From a predictive point of view, BIC beats AIC yielding a smaller RSFE on average, only in a long-sample context. When we test both hypotheses together controlling better for type-I error, our results supports this long-sample BIC superiority.

The remaining work proceed as follows. In Section 2, we describe our dataset, and discuss some asymptotic properties of information criteria. In Section 3, we report univariate in- and out-of-sample test results. In Section 4, we describe and analyze the results of joint test. Also, we provide some intuition about the different type-I error control used by our testing approaches. Finally, Section 5 concludes.

2. ESTIMATION SETUP

2.1. Data

The simulated stationary data is generated as realizations of the AR(6) process:

$$y_t = 0.09y_{t-1} + 0.08y_{t-2} + 0.07y_{t-3} + 0.06y_{t-4} + 0.05y_{t-5} + 0.04y_{t-6} + \varepsilon_t,$$

where $\varepsilon_t \sim iidN(0, 2\%)$, using a random numbers generator. Note that the ratio of persistency to variance (0.39/0.02) reaches 20 times, a value similarly achieved with the maximum level of persistency allowed and a variance of 5% (0.99/0.05). Thus, describing a vast majority of economic time-series. The order $p = 6$ does not depends on itself. It is chosen in accordance to the relative slack between the maximum order of autoregression with which the search of the best model is made and the true order –in this case, as $p^{\max} = 24$, the gap is four times the order of the true model. The number of replications is $I = 1{,}000$, and the complete sample size is $T = 5{,}000$, adding one observation for forecasting evaluation. We perform the same exercise four times, each one with a different sample size varying according to $\tau = \{50; 100; 1{,}000; 5{,}000\}$. By doing this, we analyze the behavior of each $\{y_t^{i \in I}\}_{t=1}^{t=\tau+1}$ process four times, carrying out an empirical insight about asymptotic behavior of both information criteria. As $I = 1{,}000$ may represent a number of replications which may not describe population parameters, we carry out a backup simulation with $I' = 10{,}000$ for the more sensitive case ($\tau = 50$). This, to have a measure of how far we are from a case more closely to population parameters. As

the results are both numerical and qualitative maintained, we keep $I = 1{,}000$ for the sake of computational efficiency[3].

2.2. Asymptotic properties

Both criteria have different asymptotics properties: AIC is not consistent while BIC it is, and when $k > 1$ it will choose the correct model almost sure (becoming strongly consistent)[4]. As is pointed out by Canova (2007), intuitively AIC is not consistent because the penalty function used does not simultaneously goes to infinity as $T \to \infty$, and to zero when scaled by T. This led us to the use of different values of τ, and stands for our conclusion with univariate tests[5]. Note that consistency is not a must for forecast accuracy; the true model may underperform out-of-sample against a nested benchmark. Hansen (2009) finds that it is expected that a model with an autoregressive order smaller than true may beat out-of-sample, as a consequence of underfitting.

The asymptotic properties of AIC and BIC are derived in Shibata (1976, 1980, 1981), Bhansali and Downham (1977), Sawa (1978), Stone (1979), Geweke and Meese (1981), Pötscher and Srinivasan (1991), Markon and Krueger (2004), and Karagrigoriou, Mattheou, and Vonta (2011). Recently, Xu and McLeod (2012) derive the asymptotics properties of the Generalized Information Criteria (GIC) which nests the criterion considered in this paper. In Appendix A we show the asymptotic properties of AIC and BIC based on Nishii (1984)[6].

3. UNIVARIATE RESULTS

3.1. In-sample results

As pointed out by Lütkepohl (1985), Nickelsburg (1985), Yi and Judge (1988), Clark (2004), Granger and Jeon (2004), Raffalovich *et al*. (2008), and Shittu and

---

[3] We perform our simulations using an *ad hoc* Matlab code for $I = 1{,}000$. We then perform our backup simulation using the more specific commands provided in Econometrics Toolbox 2.1. The latter estimates takes a prohibitive debugging time with $I' = 10{,}000$ and four values of τ. Another tool used was Eviews 7.2, but its pseudo-random numbers generator was not so powerful as the generated by Matlab. We provide statistical inference of each comparison to check the robustness of our results.

[4] See more details on Bozdogan (1987), Bickel and Zhang (1992), and Wasserman (2000). Some authors has proposed several modifications to AIC to improve its long-sample behavior, as Hurvich and Tsai (1993), and Burnham and Anderson (1998).

[5] There is no specific definition for short-sample. Thus, we find that, for example, are used as 45 observations in Sargent and Sims (1977), 14 in Miller, Supel, and Turner (1980), 15 in Nickelsburg (1985), 23 in Sims (1980), 68 in Fischer (1981), 56 in Gordon and King (1982), and many other candidates.

[6] Along this paper we keep fixed the variance of the data generating process. Other cases of asymptotic properties, besides when $T \to \infty$, are derived for instance in Stone (1979) and Shibata (1981). Empirically, Yang (2003) and Chen, Giannakouros, and Yang (2007) analyze some cases where the variance becoming larger.

Asemota (2009), AIC is prone to selecting more dynamic models than is the BIC –a fact that is supported theoretically. In Figure 1, we report the relative frequency of the number of regressors chosen by each criterion with different sample size, showing the common finding. These lag length orders are chosen by computing the lowest score achieved by each criterion fitting the AR(6) process choosing $p \in \mathbb{N}\,[1,24]$. The results of Figure 1 are summarized in Table 1, which reflects a consistent overfitting of AIC and the alignment of BIC through the true order as sample size increases.

TABLE 1

STATISTICS OF THE NUMBER OF REGRESSORS CHOSEN BY EACH CRITERION

	$\tau = 50$		$\tau = 100$		$\tau = 1{,}000$		$\tau = 5{,}000$	
	AIC	BIC	AIC	BIC	AIC	BIC	AIC	BIC
Median	19	17	10	1	12	4	12	6
Maximum	24	24	24	10	24	9	24	11
Minimum	1	1	1	1	2	1	5	4
Standard deviation	6.36	9.67	7.80	1.31	6.88	1.35	6.67	0.59
Skewness	–1.49	–0.04	0.29	1.92	0.22	0.08	0.26	0.84
Kurtosis	4.19	1.11	1.58	7.12	1.55	3.17	1.52	13.21

Source: Authors' computations.

For inference purposes, we define the variable $\Delta N_{il\tau}$ for the i^{th} replication as the difference between the number of regressors chosen by AIC and by BIC given a sample size τ: $\Delta N_{il\tau} = NReg_{il\tau}^{AIC} - NReg_{il\tau}^{BIC}$. Naturally, the variable $\Delta N_{il\tau}$ has a fixed sample size of 1,000 observations (the number of replications). We estimate the regression $\Delta N_{il\tau} = c_\tau + \upsilon_{il\tau}$, where $\upsilon_{il\tau} \sim iidN(0,\sigma_\upsilon^2)$ and test the one-sided null hypothesis (NH) that $NH_\tau^{In-Sample} : E[c_\tau] \leq 0$, following the Diebold and Mariano (1995) and West (1996) approach. Rejecting the NH will confirm the statistical significance of AIC's overfitting compared with BIC[7]. The estimates by ordinary least squares (OLS) are presented in Table 2.

[7] This finding is not necessarily bad for the AIC. There an extensive empirical literature that finds that AIC outperforms BIC in many contexts. Moreover, Kilian (2001) finds that it is a better criterion for identifying the true impulse response function.

TABLE 2

ESTIMATES OF DIFFERENCES IN NUMBER OF REGRESSORS

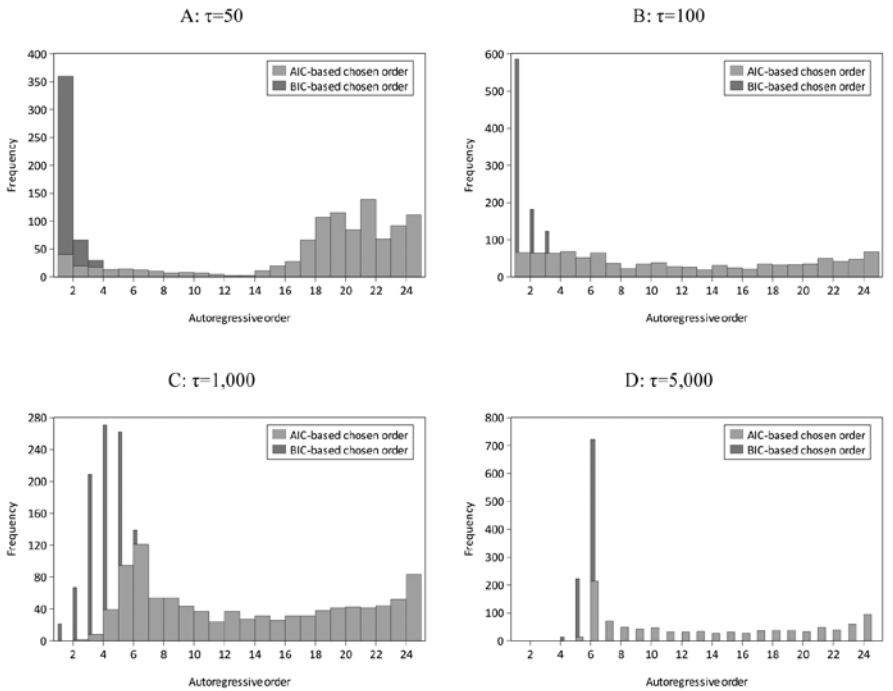
	$\tau = 50$	$\tau = 100$	$\tau = 1{,}000$	$\tau = 5{,}000$
c_τ	6.30	9.75	8.94	7.81
Standard deviation	0.28	0.25	0.22	0.20
One-sided p-value	0.00	0.00	0.00	0.00

Source: Authors' computations.

The statistic $t_{\Delta N} = \Delta\overline{N} / [\sigma_{\Delta N} / \sqrt{Obs.}]$ is statistically significant at traditional levels of significance. This implies that the AIC chooses consistently more dynamic models than those chosen by BIC.

FIGURE 1

HISTOGRAMS OF IN-SAMPLE AUTOREGRESSIVE ORDER ESTIMATES



Source: Authors' computations.

3.2. Out-of-sample results

Lütkepohl (1985) shows that BIC outperforms AIC among other criteria in a 1-step ahead out-of-sample simulation exercise with vector autoregressions. Other authors, such as Koehler and Murphree (1988), and Granger and Jeon (2004), also find BIC to be superior to AIC when using macroeconomic data, and at multiple horizons. We replicate this finding in our setup by performing 1-step ahead forecasts for each $\{y_t^{i\in I}\}_{t=1}^{t=\tau+1}$ replication. The results for each criterion are depicted in Table 3, where BIC-based forecasts show a better fit with $\tau = 50$ and along with less volatile errors only with $\tau = 5{,}000$. The columns of Table 3 corresponds to descriptive statistics of root squared forecast error (RSFE) measure, defined as:

$$RSFE = \left[(y_{t|t}^{i} - \hat{y}_{t|t-1}^{i|\tau,criterion})^2 \right]^{\frac{1}{2}},$$

where $\hat{y}_{t|t-1}^{i|\tau,criterion}$ is the 1-step ahead forecast of $y_{t|t}^{i}$ based on a model estimated with a sample size τ and the criterion AIC or BIC.

We then evaluate the accuracy by computing the statistical significance of the difference between the squared forecast error (SFE) achieved by both criteria, using the series,

$$\Delta SFE_{i|\tau} = SFE_{i|\tau}^{AIC} - SFE_{i|\tau}^{BIC} = (y_{t|t}^{i} - \hat{y}_{t|t-1}^{i|\tau,AIC})^2 - (y_{t|t}^{i} - \hat{y}_{t|t-1}^{i|\tau,BIC})^2 .$$

We test the one-sided null hypothesis that $NH_{\tau}^{Out-of-Sample} : E[d_{\tau}] \leq 0$ over the regression $\Delta SFE_{i|\tau} = d_{\tau} + \xi_{i|\tau}$, with $\xi_{i|\tau} \sim iidN(0, \sigma_{\xi}^2)$. Estimates by OLS are presented in Table 4. There is evidence of predictive BIC-superiority only with long-sample estimates. For short-sample we can not determine about predictive fit between both information criteria; even more, with $\tau = 100$ the statistic d_{τ} is negative but not significant.

TABLE 3

STATISTICS OF THE FORECASTING EVALUATION SERIES

	$\tau = 50$		$\tau = 100$		$\tau = 1{,}000$		$\tau = 5{,}000$	
	AIC	BIC	AIC	BIC	AIC	BIC	AIC	BIC
Mean	0.65	0.64	0.65	0.66	0.68	0.66	0.99	0.91
Median	0.56	0.53	0.56	0.57	0.57	0.58	0.45	0.42
Maximum	9.00	9.24	10.50	8.52	8.61	7.94	10.31	9.69
Minimum	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Standard deviation	0.76	0.77	0.76	0.75	0.74	0.76	1.33	1.21
Skewness	5.48	5.21	6.04	5.14	4.32	5.07	2.40	2.31
Kurtosis	44.68	40.12	57.07	40.12	29.94	38.31	10.75	10.21

Source: Authors' computations.

TABLE 4

ESTIMATES OF DIFFERENCES IN SFE

	$\tau = 50$	$\tau = 100$	$\tau = 1{,}000$	$\tau = 5{,}000$
d_τ	0.01	–0.01	0.02	0.08
Standard deviation	0.01	0.02	0.02	0.02
One–sided *p*-value	0.16	0.27	0.12	0.00

Source: Authors' computations.

4. A JOINT TEST

4.1. A reality check

We now test the two null hypotheses together in a standardized version for each sample size τ:

$$\begin{bmatrix} NH_\tau^{In-sample} \\ NH_\tau^{Out-of-sample} \end{bmatrix} = E\begin{bmatrix} NReg_{\tau,Standardized}^{AIC} - NReg_{\tau,Standardized}^{BIC} \\ SFE_{\tau,Standardized}^{AIC} - SFE_{\tau,Standardized}^{BIC} \end{bmatrix} = E\left[\mathbf{Z}_\tau\right] \leq 0 .$$

It is expected that a vector $\mathbf{x}$ that contains all the NHs has nonpositive values, implying that BIC is the best in estimation and forecasting. When the number of replications (I) goes to infinity, we have $\sqrt{I}\left(\overline{\mathbf{Z}} - E[\mathbf{Z}]\right) \xrightarrow{A} N(\mathbf{0}, \mathbf{\Omega})$ where $\mathbf{Z}$ is a standardized vector $\mathbf{x}$ ($\mathbf{Z} = [\mathbf{x} - \overline{\mathbf{x}}]' \Sigma_{\mathbf{x}}^{-1}$, with $\mathbf{\Sigma}$ the covariance matrix of $\mathbf{x}$), and $\mathbf{\Omega}$ is the long-run covariance matrix. While I goes to infinity, we are able to build the following statistic,

$$\max_{m \in \{1,\ldots,H\}} \left[\sqrt{I} \frac{1}{I} \sum_{i=1}^{I} (\overline{\mathbf{Z}}_{mi} - E\left[\mathbf{Z}_{mi}\right]) \right]_{H \times 1}$$

where m is the m^{th} row of a vector $\mathbf{Z}$ that contains all the hypotheses to be tested. Nevertheless, as the maximum of a Gaussian process is not Gaussian, we have to use any methodology able to deliver asymptotically valid *p*-values for the least favorable configuration (LFC). As White (2000) pointed out, there two ways in which we can compute the *p*-values for LFC: (i) a simulation-based approach, and (ii) a bootstrap-based approach. We use the former, but in a less conservative manner as in Hansen (2005)[8].

[8] A brief review about divergences of both methods are discussed in Corradi and Distaso (2011).

Consider the diagonal matrix $\mathbf{D}$, defined as $\mathbf{D}_{mm} = \sigma_m^{-1}$; $m = 1, \ldots, H$, in which $\sigma_m^2 = \Omega_{mm}$. Then, it must be fulfilled that $\sqrt{I}\mathbf{D}\left(\overline{\mathbf{Z}} - E[\mathbf{Z}]\right) \xrightarrow{A} N(\mathbf{0}, \mathbf{D\Omega D})$, with the advantage that now $[\mathbf{D\Omega D}]_{mm} = 1$; $\forall m = 1,...,H$. However, the terms $E[\mathbf{Z}]$, $\mathbf{D}$, and $\mathbf{\Omega}$ are unknown. Regarding the first unknown term, note that the NH can be written as $NH : E[\mathbf{Z}] \leq \mathbf{0}$, and, as the number of vectors that are coherent with this NH goes to infinity, we can pick the LFC, $E[\mathbf{Z}] = \mathbf{0}$, and work in a bounded test that allows for the identification of unknown terms. For the remaining two, we can use the Newey and West (1987) method to obtain a positive definite consistent estimator of $\mathbf{\Omega}$, generating an estimation of $\mathbf{D}$ using $\mathbf{D}_{mm} = \mathbf{\Omega}_{mm}^{-0.5}$ [9].

Embedding all the identified terms, under the NH we have $\sqrt{I}\hat{\mathbf{D}}\overline{\mathbf{Z}} \xrightarrow{A} N(\mathbf{0}, \hat{\mathbf{\Omega}})$ where $\hat{\mathbf{\Omega}} \equiv \hat{\mathbf{D}}\hat{\mathbf{\Omega}}\hat{\mathbf{D}}$. Then, the statistic can be written as,

$$\max_{m \in \{1,\ldots,H\}} \sqrt{I}\hat{\mathbf{D}}\overline{\mathbf{Z}},$$

where m-elements represent the components of the vector $\hat{\mathbf{D}}\overline{\mathbf{Z}}$.

The critical values of the statistic are derived from Monte Carlo simulations according to White's (2000) procedure, following these steps: (i) calculate the Cholesky decomposition of $\hat{\mathbf{D}}\hat{\mathbf{\Omega}}\hat{\mathbf{D}}=\mathbf{G'G}$, with $\mathbf{G}$ being a superior triangular matrix, (ii) define a number of replications, representing the number of realizations of the experiment, in this case, 1,000,000, (iii) for each replication, calculate an independent realization v of a multivariate normal distribution $N(\mathbf{0}, \mathbf{I}_{H\times H})$, (iv) define ω as $\omega = \mathbf{G}'v$, such that ω is an independent realization of $N(\mathbf{0}, \hat{\mathbf{D}}\hat{\mathbf{\Omega}}\hat{\mathbf{D}})$, (v) define s as:

$$s = \max_{m \in \{1,\ldots,H\}} \{\omega_m\}$$

and finally, (vi) sort the m terms and define the critical values according to the corresponding quantiles.

4.2. Estimates results

The estimates of $\mathbf{Z}$ and $\hat{\mathbf{\Omega}}$ with the Newey-West estimator gives the next pairwise results,

---

[9] As $\mathbf{\Omega}$ is a positive semidefinite matrix, at least one hypothesis has to be nonnested. There is no available test for multiple nested hypotheses with $m > 2$ at the time. However, the test proposed in Clark and McCracken (2001) can be used for pairwise comparisons ($m = 2$).

$$\mathbf{Z}_{\tau=50}=\begin{bmatrix} -8.03\times10^{-17} \\ -1.44\times10^{-16} \end{bmatrix}, \Omega_{\tau=50}=\begin{bmatrix} 1.00 & 0.06 \\ 0.06 & 1.00 \end{bmatrix},$$

$$\mathbf{Z}_{\tau=100}=\begin{bmatrix} -1.84\times10^{-17} \\ -1.90\times10^{-16} \end{bmatrix}, \Omega_{\tau=100}=\begin{bmatrix} 1.00 & 0.07 \\ 0.07 & 1.00 \end{bmatrix},$$

$$\mathbf{Z}_{\tau=1,000}=\begin{bmatrix} -1.31\times10^{-16} \\ -2.15\times10^{-16} \end{bmatrix}, \Omega_{\tau=1,000}=\begin{bmatrix} 1.00 & 0.10 \\ 0.10 & 1.00 \end{bmatrix},$$

$$\mathbf{Z}_{\tau=5,000}=\begin{bmatrix} 1.65\times10^{-16} \\ 2.74\times10^{-16} \end{bmatrix}, \Omega_{\tau=5,000}=\begin{bmatrix} 1.00 & 0.08 \\ 0.08 & 1.00 \end{bmatrix},$$

After 1,000,000 of replications of each $\mathbf{G}'\nu$ matrix, we have the following estimations of $\hat{\mathbf{D}}\hat{\Omega}\hat{\mathbf{D}}$,

$$\hat{\mathbf{D}}\hat{\Omega}\hat{\mathbf{D}}_{\tau=50}=\begin{bmatrix} 23.78 & 2.79 \\ 0.00 & 24.19 \end{bmatrix}, \hat{\mathbf{D}}\hat{\Omega}\hat{\mathbf{D}}_{\tau=100}=\begin{bmatrix} 24.48 & 3.34 \\ 0.00 & 24.73 \end{bmatrix},$$

$$\hat{\mathbf{D}}\hat{\Omega}\hat{\mathbf{D}}_{\tau=1,000}=\begin{bmatrix} 23.25 & 4.47 \\ 0.00 & 21.56 \end{bmatrix}, \hat{\mathbf{D}}\hat{\Omega}\hat{\mathbf{D}}_{\tau=5,000}=\begin{bmatrix} 22.25 & 3.52 \\ 0.00 & 24.56 \end{bmatrix}.$$

Given that the results of tabulated $\left(m_{\tau=\tau_0}^{90\%}, \tau_0 \in \tau\right)$ and calculated critical value of the maximum element of $\mathbf{Z}_\tau\left(t_{\mathbf{Z}_m}^{\tau=\tau_0}=\max_{m=1,\ldots,H}\sqrt{I}\mathbf{Z}_{m|\tau}\right)$ are:

τ	$m_{\tau=\tau_0}^{90\%}$	$t_{\mathbf{Z}_m}^{\tau=\tau_0}$
50	–1.13 x 10^{-16}	–2.93 x 10^{-17}
100	–1.49 x 10^{-16}	–3.82 x 10^{-17}
1,000	–1.69 x 10^{-16}	–4.63 x 10^{-17}
5,000	2.14 x 10^{-16}	5.52 x 10^{-17}

the $NH: E\left[\mathbf{Z}_\tau\right] \leq \mathbf{0}$ is not rejected at typical significance levels for $\tau = \{50;100;1,000\}$. But, when $\tau = 5,000$ the results leads us to state that BIC is a dominant criteria for modeling stationary autoregressive processes for forecasting purposes.

4.3. Type-I error control analysis

According to White (2000), Hansen (2005), Corradi and Distaso (2011), and Pincheira (2011a, 2012), when interest is centered on testing more than one univariate hypothesis jointly, there are generally two strategies for statistical inference. On one hand, we may determine the superiority in- and out-of-sample of BIC over AIC by stating that, given the results of both individual tests, we may reject or not both NH[10]. On the other hand, we can perform a joint test that controls better for the type-I error (this is, reject a true null hypothesis), as is summarized in the derivation of asymptotic valid *p*-values for LFC statistic. Obviously, both strategies will have the same outcome when the hypotheses are fully independent.

The first strategy –in this case, that based on the separate regressions– may present shortcomings handling type-I error, that is, rejection of a true NH. To figure this out, we will follow closely the next example proposed in Pincheira (2011a, 2012).

Assume that $NH: E\left[\mathbf{Y}\right] = \mathbf{0}_{L\times L}$, $\mathrm{L} \in \mathbb{N}$, and the alternative hypothesis (AH) states that at least one component of $\mathbf{Y}$ is positive, $AH: \exists l \in \{1,...,L\} \mid E\left[\mathbf{Y}_l\right] > 0$. Let's suppose now that we have a collection of tests T_l that depends on sample size (Ψ), and is assigned to test $NH^{(l)}: E\left[\mathbf{Y}_l\right] = 0$, with one-sided $AH^{(l)}: E\left[\mathbf{Y}_l\right] > 0$, implying that *any* T_l will reject the $NH^{(l)}$ at a determined confidence level $0 \leq \alpha \leq 1$ when $T_l(\Psi) > \delta$. In this case, δ represents a tabulated value coming from the distribution function to which contrast the NH. If the elements of $\vec{T} = (T_1,...,T_L)'$ are orthogonal, we have that,

$$\Pr\left(\exists l \in \{1,...,L\} \ni T_l(\Psi) > \delta \mid NH\right) = \Pr(\sum_{l=1}^{L} \Upsilon_l > 0 \mid NH),$$

in which $\Upsilon_l = 1$ if $T_l(\Psi) > \delta$, or 0 otherwise. Then, Υ_l is a random variable that follows a Bernoulli distribution function of parameter $p \geq \alpha$, $0 \leq p \leq 1$. Under the NH, $\sum_{l=1}^{L} \Upsilon_l$ follows a binomial distribution with parameters *L* and *p*. By using this terms, we have that,

---

[10] In this class of tests we found approaches like Bonferroni bounds and the proposed by Holm (1979).

$$\Pr\left(\sum_{l=1}^{L}\Upsilon_l > 0 \mid NH\right) = 1 - \Pr\left(\sum_{l=1}^{L}\Upsilon_l = 0 \mid NH\right),$$
$$= 1 - \Pr\left(T_l(\Psi) \leq \delta \quad \forall l \in \{1,...,L\} \mid NH\right),$$
$$= 1 - (1-p)^L \to 1 \quad when \quad L \to \infty.$$

In other words, the strategy that tests the NH under the assumption of orthogonality between the elements of $\vec{T}$, looses the control of type-I error as the number of hypotheses to be tested goes to infinity[11]. Instead, this will not happen with a joint test that takes into account the interactions between the elements of $\vec{T}$.

5. CONCLUDING REMARKS

This document addresses the overfitted in-sample estimation of the AIC relative to BIC, and forecast accuracy using autoregressive models based on both information criteria. We formally test two null hypotheses: (i) Is the BIC more parsimonious than the AIC? and (ii) Is BIC better than AIC for forecasting purposes? The exercise consists of a simulation of a stationary dataset of 1,000 series generated by an AR(6) process, and then computing and comparing the order determined by each criterion chosen from a maximum order of 24 lags. Then, for each model, we generate 1-step ahead forecasts and evaluate their accuracy. We perform this exercise four times, each one with a different estimation sample size varying according to 50, 100, 1,000, and 5,000 observations.

We test both null hypotheses individually with standard significance tests, and jointly with a multiple hypotheses test. The results show that the AIC chooses more dynamic models than those chosen with the BIC, and that BIC-based models have better out-of-sample performance than those based on AIC only with long-sample estimates. Furthermore, it is also shown that when the type-I error is controlled with a multiple hypotheses testing procedure, such as that the one developed in White (2000) and Hansen (2005), the results are robust. This leads us to conclude that BIC is a dominant criteria for modeling stationary autoregressive processes and for forecasting purposes exclusively in a long-sample context.

[11] Notice that even with $L = 2$ the test size could be distorted.

REFERENCES

AKAIKE, H. (1974). "A New Look at the Statistical Model Identification", *IEEE Transactions on Automatic Control* 19 (6), pp. 716-723.

BHANSALI, R.J. and D.Y. DOWNHAM (1977). "Some Properties of the Order of an Autoregressive Model Selected by a Generalization of Akaike's EPF Criterion", *Biometrika* 64 (3), pp. 547-551.

BICKEL, P. and P. ZHANG (1992). "Variable Selection in Nonparametric Regression with Categorical Covariates", *Journal of the American Statistical Association* 87, pp. 90-97.

BOZDOGAN, H. (1987). "Model Selection and Akaike's Information Criterion (AIC): The General Theory and its Analytical Extensions", *Psychometrika* 52 (3), pp. 345-370.

BURHNAM, K.P. and D.R. ANDERSON (1998). *Model Selection and Inference: A Practical Information Theoretic Approach*, Springer, New York.

CANOVA, F. (2007). *Methods for Applied Macroeconomic Research*, Princeton University Press, USA.

CHEN, L., P. GIANNAKOUROS and Y. YANG (2007). "Model Combining in Factorial Data Analysis", *Journal of Statistical Planning and Inference* 137 (9), pp. 2920-2934.

CLARK, T.E. and M. McCRACKEN (2001). "Tests of Equal Forecast Accuracy and Encompassing for Nested Models", *Journal of Econometrics* 105 (1), pp. 85-110.

CLARK, T.E. (2004). "Can Out-of-Sample Forecast Comparisons Help to Prevent Overfitting?", *Journal of Forecasting* 23 (2), pp. 115-139.

CORRADI, V. and W. DISTASO (2011). "Multiple Forecast Model Evaluation", in M.P. Clements and D.F. Hendry (eds.), *The Oxford Handbook of Economic Forecasting*, Oxford University Press, USA, pp. 391-414.

DIEBOLD, F.X. and R. MARIANO (1995). "Comparing Predictive Accuracy", *Journal of Business and Economic Statistics* 13 (3), pp. 253-263.

FISCHER, S. (1981). "Relative Shocks, Relative Price Volatility, and Inflation", *Brooking Papers on Economic Activity* 2, pp. 381-431.

GEWEKE, J. and R. MEESE (1981). "Estimating Regression Models of Finite but Unknown Order", *International Economic Review* 22 (1), pp. 55-70.

GORDON, R.J. and S.R. KING (1982). "The Output Cost of Desinflation in Traditional and Vector Autoregressive Models", *Brooking Papers on Economic Activity* 13 (1), pp. 205-244.

GRANGER, C.W.J. and Y. JEON (2004). "Forecasting Performance of Information Criteria with Many Macro Series", *Journal of Applied Statistics* 31 (10), pp. 1227-1240.

HANSEN, P.R. (2005). "A Test of Superior Predictive Ability", *Journal of Business and Economic Statistics* 23, pp. 365-380.

HANSEN, P.R. (2009). "*In-Sample Fit and Out-of-Sample Fit: Their Joint Distribution and its Implications for Model Selection*", preliminary version April 23, 2009, Department of Economics, Stanford University, USA.

HOLM, S. (1979). "A Simple Sequentially Rejective Multiple Test Procedure", *Scandinavian Journal of Statistics* 6, pp. 65-70.

HURVICH, C.M. and C.-L. TSAI (1993). "A Corrected Akaike Information Criterion for Vector Autoregressive Model Selection", *Journal of Time Series Analysis* 14, pp. 271-279.

KARAGRIGORIOU, A., K. MATTHEOU and I. VONTA (2011). "On Asymptotic Properties of AIC Variants with Applications", *American Open Journal of Statistics* 1, pp. 105-109.

KILIAN, L. (2001). "Impulse Response Analysis in Vector Autoregressions with Unknown Lag Order", *Journal of Forecasting* 20 (3), pp. 161-179.

KOEHLER, A.B. and E.S. MURPHREE (1988). "A Comparison of the Akaike and Schwarz Criteria for Selecting Model Order", *Journal of the Royal Statistical Society, Series C (Applied Statistics)* 37 (2), pp. 187-195.

KUHA, J. (2004). "AIC and BIC: Comparison of Assumptions and Performance", *Sociological Methods and Research* 33 (2), pp. 188-229.

LÜTKEPOHL, H. (1985). "Comparison of Criteria for Estimating the Order of a Vector Autoregressive Process", *Journal of Time Series Analysis* 6 (1), pp. 35-52.

MARKON, K.E. and R.F. KRUEGER (2004). "An Empirical Comparison of Information – Theoretic Selection Criteria for Multivariate Behavior Genetic Models", *Behavior Genetics* 34 (6), pp. 593-609.

MILLER, P., T.M. SUPEL and T.H. TURNER (1980). "Estimating the Effects of the Oil-Price Shock", *Quarterly Review*, Federal Reserve Bank of Minneapolis.
NEWEY, W. and K. WEST (1987). "A Simple, Positive Semi-definite, Heteroskedasticity and Autocorrelation Consistent Covariance Matrix", *Econometrica* 55 (3), pp. 703-708.
NICKELSBURG, G. (1985). "Small-Sample Properties of Dimensionality Statistics for Fitting VAR Models to Aggregate Economic Data –A Monte Carlo Study", *Journal of Econometrics* 28 (2), pp. 183-192.
NISHII, R. (1984). "Asymptotic Properties of Criteria for Selection of Variables in Multiple Regression", *Annals of Statistics* 12 (2), pp. 758-765.
PINCHEIRA, P. (2011a). *A Joint Test of Superior Predictive Ability for Chilean Inflation*, (in Spanish) Working Paper 620, Central Bank of Chile.
PINCHEIRA, P. (2011b). *A Bunch of Models, a Bunch of Nulls and Inference About Predictive Ability*, Working Paper 607, Central Bank of Chile.
PINCHEIRA, P. (2012). "Un Test Conjunto de Superioridad Predictiva para los Pronósticos de Inflación Chilena", *Journal Economía Chilena (The Chilean Economy)* 15 (3), pp. 4-39.
PÖTSCHER, B.M. and S. SRINIVASAN (1991). "A Comparison of Order Estimation Procedures for ARMA Models", *Statistica Sinica* 4, pp. 29-50.
RAFFALOVICH, L.E., G.D. DEANE, D. ARMSTRONG and H.-S. TSAO (2008). "Model Selection Procedures in Social Research: Monte-Carlo Simulation Results", *Journal of Applied Statistics* 35 (10), pp. 1094-1114.
RISSASEN, J. (1978). "Modeling by Shortest Data Description", *Automatica* 14 (5), pp. 465-471.
SAWA, T. (1978). "Information Criteria for Discriminating Among Alternative Regression Models", *Econometrica* 46 (6), pp. 1273-1282.
SARGENT, T. and C. SIMS (1977). "*Business Cycle Modeling Without Pretending to Have too Much a priori Economic Theory*", Working Paper 55, Federal Reserve Bank of Minneapolis, USA.
SCHWARZ, G.E. (1978). "Estimating the Dimension of a Model", *Annals of Statistics* 6 (2), pp. 461-464.
SHIBATA, R. (1976). "Selection of the Order of an Autoregressive Model by Akaike Information Criterion", *Biometrika* 63 (1), pp. 117-126.
SHIBATA, R. (1980). "Asymptotically Efficient Selection of the Order of the Model for Estimating Parameters of a Linear Process", *Annals of Statistics* 8 (1), pp. 147-164.
SHIBATA, R. (1981). "An Optimal Selection of Regression Variables", *Biometrika* 68, pp. 45-54.
SHITTU, O.I. and M.J. ASEMOTA (2009). "Comparison of Criteria for Estimating the Order of Autoregressive Process: A Monte Carlo Approach", *European Journal of Scientific Research* 30 (3), pp. 409-416.
SIMS, C. (1980). "Macroeconomics and Reality", *Econometrica* 48 (1), pp. 1-48.
SULLIVAN, R., A. TIMMERMANN and H. WHITE (1999). "Data-Snooping, Technical Trading Rule Performance, and the Bootstrap", *Journal of Finance* 54 (5), pp. 1647-1691.
STONE, M. (1979). "Comments on Model Selection Criteria of Akaike and Schwarz", *Journal of the Royal Statistical Society, Series B (Methodological)* 41 (2), pp. 276-278.
WASSERMAN, L. (2000). "Bayesian Model Selection and Model Averaging", *Journal of Mathematical Psychology* 44, pp. 92-107.
WEAKLIEM, L.D. (2004). "Introduction to the Special Issue on Model Selection", *Sociological Methods and Research* 33, pp. 167-186.
WEST, K. (1996). "Asymptotic Inference about Predictive Ability", *Econometrica* 64 (5), pp. 1067-1084.
WHITE, H. (2000). "A Reality Check for Data Snooping", *Econometrica* 68, pp. 1097-1126.
WOLAK, F.A. (1987). "An Exact Test for Multiple Inequality and Equality Constraints in the Linear Regression Model", *Journal of the American Statistical Association* 82, pp. 782-793.
WOLAK, F.A. (1989). "Testing Inequality Constraints in Linear Econometric Models", *Journal of Econometrics* 31, pp. 205-235.
XU, C.J. and I. McLEOD (2012). "Further Asymptotic Properties of the Generalized Information Criteria", *Electronic Journal of Statistics* 6, pp. 656-663.
YANG, Y. (2003). "Regression with Multiple Candidate Models: Selecting or Mixing?", *Statistica Sinica* 13, pp. 783-809.
YI, G. and G. JUDGE (1988). "Statistical Model Selection Criteria", *Economic Letters* 28 (1), pp. 47-51.
ZUCCHINI, W. (2000). "An Introduction to Model Selection", *Journal of Mathematical Psychology* 44, pp. 41-46.

APPENDIX A: ASYMPTOTIC PROPERTIES OF AIC AND BIC

This appendix constitutes a reduced version of Nishii (1984). No more elements than those derived on Nishii's paper have been added.

A.1. Preliminaries

Consider the stationary regression model $\mathbf{y} = \Phi_P \mathbf{y}_{t-P} + \boldsymbol{\varepsilon}$, where $\mathbf{y}$ is an $T \times 1$ vector of observations, Φ_P is a coefficient matrix, $\Phi_P = (\phi_1, ..., \phi_P)'$, and $\boldsymbol{\varepsilon}$ is assumed to be independently normally distributed, $\boldsymbol{\varepsilon} \sim iidN(\mathbf{0}, \sigma^2 \mathbf{I}_T)$. We index a set of models with $j = \{j_1, ..., j_p\}$, sorted according to $1 \le j_1 \le ... \le j_p \le P$, if and only if $\Phi_i \neq 0$, for all $i = j$. The number of unknowns parameters achieves $p_j = p + 1$, because σ^2 is unknown.

Define $\mathbf{D}_j$ the matrix of order $P \times p$, of zeros and ones, that depicts the model *j*. Thus, the model *j*, $\mathbf{y} = \Phi_j \mathbf{y}_{t-j} + \boldsymbol{\varepsilon}$, has an estimated vector parameter $\Phi_j = \mathbf{D}_j \mathbf{D}_j' \Phi_P$. Consider a family of nested models, *J*, thus, we state the following assumption:

ASSUMPTION *If J contains the true model,* $j_0 = \{1, ..., p_0\}$, *the matrix* $\mathbf{y'y}$ *is positive definite, and* $M = \lim_{T \to \infty} T^{-1}(\mathbf{y'y})$ *exists and is positive definite.*

This assumption implies that $rank(\mathbf{y}\mathbf{D}_j) = p$, in other words, that $\mathbf{D}'_j \mathbf{y}'_{t-p} \mathbf{y}_{t-p} \mathbf{D}_j$ is positive definite. For the model $j \in J$ we define the following quantities:

$$\hat{\Phi}_j = \mathbf{D}_j (\mathbf{D}'_j \mathbf{y}'_{t-P} \mathbf{y}_{t-P} \mathbf{D}_j)^{-1} \mathbf{D}'_j \mathbf{y}'_{t-P} \mathbf{y},$$

$$\mathbf{Q}_j = \mathbf{y}_{t-P} \mathbf{D}_j (\mathbf{D}_j' \mathbf{y}'_{t-P} \mathbf{y}_{t-P} \mathbf{D}_j)^{-1} \mathbf{D}' \mathbf{y}_{t-P},$$

$$\hat{\sigma}_j^2 = T^{-1} \cdot \mathbf{y}' \left[\mathbf{I}_T - \mathbf{Q}_j \right] \mathbf{y},$$

where $\hat{\Phi}_j$ is the maximum likelihood estimator of Φ_j, $\mathbf{Q}_j$ is the projection operator with respect to column space of $\mathbf{y}_{t-P}\mathbf{D}_j$, and $\hat{\sigma}_j^2$ is the maximum likelihood estimator of σ_j^2. We discuss the asymptotic properties of the Generalized Information Criteria (GIC) defined as $GIC_j = T \cdot \log \hat{\sigma}_j^2 + g(T) \cdot p_j$, that nests both AIC and BIC. Thus,

$$GIC = \begin{cases} AIC & if\ g(T) = 2, \\ BIC & if\ g(T) = \log(T). \end{cases}$$

Along this work we consider only the case where P and Φ are kept fixed as $T \to \infty$. Some alternative cases are presented in Stone (1979) and Shibata (1981).

A.2. Goodness of fit measures

Consider j a model selected of an information criterion of all J possible specifications. We define the following two measures of goodnes of fit to whom derive its asymptotic properties:

$$\textit{(i):} \quad \Pr_{j|T} = \Pr\left[\hat{j} = j\right],$$

$$\textit{(ii):} \quad R_T = E_{\mathrm{y}}\left[\left|\Phi_P \mathbf{y}_{t-P} - \Phi_j \mathbf{y}_{t-j}\right|^2\right],$$

We can redefine the second term by expressing R_T as a sum of $R_{j|T}$ across j, $R_T = \sum_{j\in J} R_{j|T} = \sum_{j\in J} E_{\mathrm{y}}\left[\left|\Phi_P \mathbf{y}_{t-P} - \Phi_j y_{t-j}\right|^2 \Gamma_{\hat{j}=j}\right]$, where Γ_{ϖ} act as indicator function of ϖ[12]. Now, let's define two groups of models, $J_1 = \{j \in J \mid j \not\supset j_0\}$, and $J_2 = \{j \in J \mid j \supseteq j_0\}$. Then, for any criterion, the next conditions must be fulfilled:

CONDITION 1: $\lim_{T\to\infty} T \cdot \Pr_{j|T} = 0$ *for* $j \in J_1$.

CONDITION 2: $\lim_{T\to\infty} \Pr_{j|T} = 0$ *for* $j \in J_2 - \{j_0\}$.

These conditions implies for $R_{j|T}$ the following:

THEOREM 1 (Nishii, 1984, p. 760):

- *If a criterion satisfies Condition 1, then* $\lim_{T\to\infty} R_{j|T} = 0$ *for* $j \in J_1$.
- *If a criterion satisfies Condition 2, then* $\lim_{T\to\infty} R_{j|T} = 0$ *for* $j \in J_2 - \{j_0\}$.

PROOF: See Nishii (1984), p. 760.

REMARK: *For a criterion that jointly satisfies Condition 1 and 2, we have* $\lim_{T\to\infty} R_T = \lim_{T\to\infty} R_{j_0|T} = p_0\sigma^2$.

[12] These same measures are used in Shibata (1976) for AIC case.

A.3. Asymptotic properties

We now show the asymptotic distribution of the model $\hat{j}$ and the limit of R_T for both criteria. Let $\mathbf{M}^{0.5}$ be a squared matrix of order P such that $\mathbf{M}^{0.5\prime}\mathbf{M}^{0.5} = \mathbf{M}$, and, for a $\tilde{j} \in J_2$, let $\mathbf{L}_{\tilde{j}}$ be a $(P - p_0) \times p^*_{\tilde{j}}$ matrix defined as (row and column orders depicted around matrix):

$$\mathbf{M}^{0.5}\mathbf{D}_j = \begin{array}{c} \\ p_0 \\ \vdots \\ P - p_0 \end{array} \begin{array}{c} \begin{array}{ccc} p_0 & \cdots & p^*_{\tilde{j}} \end{array} \\ \begin{bmatrix} 0 & \dots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \dots & \mathbf{L}_{\tilde{j}} \end{bmatrix} \end{array},$$

where $p^*_{\tilde{j}} = p_{\tilde{j}} - p_{j_0}$. For $\tilde{j} \in J_2$, we define the following squared matrix of order $(P - p_0)$, $\xi_{\tilde{j}} = \mathbf{z}'\mathbf{L}_{\tilde{j}}(\mathbf{L}'_{\tilde{j}}\mathbf{L}_{\tilde{j}})^{-1}\mathbf{L}'_{\tilde{j}}\mathbf{z}$, where $\mathbf{z} \sim N(\mathbf{0}, \mathbf{I}_{P-p_0})$, and $\xi^i_{\tilde{j}} = \xi_{\tilde{j}} - i \cdot p^*_{\hat{j}}$. When $p_0 = P$, the matrices $\mathbf{L}_{\tilde{j}}$ and $\mathbf{z}$ are set to zeros.

LEMMA: *For a model* $\tilde{j} \in J_2$, $AIC_{j_0} - AIC_{\tilde{j}}$ *converges in law to the random variable* $\xi^i_{\tilde{j}}$ *as* $T \to \infty$.

THEOREM 2 (Asymptotic properties of $\Pr_{j|T}$ and R_T for AIC, Nishii, 1984, p. 761):

- *For a model* $j \in J_1$, *and any positive constant* λ, $\lim_{T\to\infty} T^{-\lambda} \Pr_{j|T} = 0$.
- *For a model* $j \in J_2$, $\Pr_{j|T}$ *converges to* $\overline{\Pr}_j = \Pr\left[\xi^i_j \geq \xi^i_{\tilde{j}}\right]$, *for* $\tilde{j} \in J_2$.
- *The function* R_T *converges to,* $\overline{R} = \sigma^2\left(p_0 + \sum_{j\in J} E\left[\xi_j \Gamma_{(\xi^i_j \geq \xi^i_{\tilde{j}})}\right]\right)$ *for* $\tilde{j} \in J_2$.

PROOF: See Nishii (1984), pp. 761-762.

Asymptotically, AIC has a positive probability of selecting models that properly include the true model. However, BIC has slightly different asymptotic properties; is a consistent estimator of the true model as follows:

THEOREM 3 (Asymptotic properties of $\Pr_{j|T}$ and R_T for BIC, Nishii, 1984, p. 764):

- *For a model* $j \in J_1$, $\Pr_{j|T} = o(T^{-\lambda})$, *for any positive constant* λ.
- *For a model* $j \in J_2 - \{j_0\}$, $\Pr_{j|T} = o(1)$.
- *The function* R_T *converges to* $p_0\sigma^2$ *as* $T \to \infty$.

PROOF: See Nishii (1984), pp. 764-765.