รายงานวิจัยฉบับสมบูรณ์

เรื่อง

The Study of the Probability of Overfitting and the Signal-to-Noise Ratio of the KIC_U Criterion

โดย

ดร.รุจิเรข บุศราวงศ์

ภาควิชาสถิติประยุกต์ คณะวิทยาศาสตร์

สถาบันเทคโนโลยีพระจอมเกล้าเจ้าคุณทหารลาดกระบัง

งานวิจัยนี้ได้รับทุนวิจัยจากรายได้คณะวิทยาศาสตร์

ประจำปีงบประมาณ 2548

ABSTRACT

This research shows the derivation of the probability of overfitting and the signal-to-noise ratio of the $\mathbf{KIC_U}$ criterion. Comparing them with **AIC**, $\mathbf{AIC_C}$, $\mathbf{AIC_U}$, **SBC, KIC** and $\mathbf{KIC_C}$ on AR(1), AR(2), and AR(3) models were examined. The results show that, for small to medium sample sizes, the $\mathbf{KIC_U}$ criterion had the lowest probability of overfitting and the highest signal-to-noise ratio. However, the **SBC** criterion is the best for large sample.

TABLE OF CONTENTS

LIST OF TABLES

CHAPTER 1
INTRODUCTION

1.1 Background

One of the most important problems in statistical modeling is to choose an appropriate model from a class of candidate models in order to characterize the underlying data. A model selection criterion provides a useful tool in this regard by assessing whether the fitted model offers an optimal balance between "goodness-of-fit" and parsimony. In other word, model selection criteria are used to select the best forecasting models.

Boosarawongse and Chongcharoen (2003) proposed the $\mathbf{KIC_C}$ criteria, which is an estimator of a variant of Kullback's symmetric divergence for autoregressive frameworks. Let n be the number of observations, $\hat{\sigma}_p^2$ is the maximum likelihood estimator of error variance of the process, and p is the model order. The $\mathbf{KIC_C}$ criteria is defined as

$$\mathbf{KIC_C}(p) = n\ln\hat{\sigma}_p^2 + n\ln\left(\frac{n}{n-p}\right) + \frac{n[(n+p)(n-p)+(n-p-2)]}{(n-p-2)(n-p)}.$$

The simulation results shown that $\mathbf{KIC_C}$ provides better model order choices than **AIC**(Akaike, 1973), $\mathbf{AIC_C}$(Hurvich and Tsai, 1989), **SBC**(Schwarz, 1978), **KIC** (Cavanaugh, 1999) for small sample. That is, it has the highest frequency of interpreting the correct model order, the lowest probability of overfitting and the highest signal-to-noise-ratio.

From the work of McQuarrie, Shumway and Tsai (1997), they have shown that $n\log\tilde{\sigma}_p^2$, where $\tilde{\sigma}_p^2 = \frac{n\hat{\sigma}_p^2}{n-p-1}$ provides a better estimator than $n\log\hat{\sigma}_p^2$. In

practice, $\tilde{\sigma}_p^2$ can be replaced by $s_p^2 = \frac{(n-p-1)\tilde{\sigma}_p^2}{n-p}$. Then, Boosarawongse (2004) proposed the following criteria

$$\mathbf{KIC_U}(p) = n\ln s_p^2 + n\ln\left(\frac{n}{n-p}\right) + \frac{n[(n+p)(n-p)+(n-p-2)]}{(n-p-2)(n-p)}.$$

The simulation results shown that for small to moderate sample sizes, the $\mathbf{KIC_U}$ criterion had the best performance in identifying the correct model order than the **AIC**, $\mathbf{AIC_C}$, $\mathbf{AIC_U}$, **SBC**, **KIC** and $\mathbf{KIC_C}$ on AR(1), AR(2), AR(3) models.

Encouraged by the preceding findings, this research attempts to find the theoretical property of the $\mathbf{KIC_U}$ criteria for AR model that is the probability of overfitting and the signal-to-noise ratio.

1.2 Objectives of the Study

1. To find the probability of overfitting and the signal-to-noise ratio of the $\mathbf{KIC_U}$ criteria for AR model.
2. To evaluate the selection performance the probability of overfitting and the signal-to-noise ratio of the $\mathbf{KIC_U}$ criteria for AR model with its of other well-known criteria

1.3 Scope of the Study

In this study, the probability of overfitting and the signal-to-noise ratio of the $\mathbf{KIC_U}$ criteria for AR model will be derive and ends with a comparison of the performance of the **AIC**, $\mathbf{AIC_C}$, $\mathbf{AIC_U}$, **SBC**, **KIC**, $\mathbf{KIC_C}$ and $\mathbf{KIC_U}$ on AR(1), AR (2), AR(3) models by comparing the probability of overfitting and the signal-to-noise ratio.

Boosarawongse (2004: 2) proposed the $\mathbf{KIC_C}$ and $\mathbf{KIC_U}$ criteria for AR models respectively.

2.2 Properties of the Criteria

When choosing the best model from the candidate models for each model selection criterion, the best model is assumed to have the lowest selection criterion value. In the following sections, **MSC** is the model selection criterion and **MSC**(p) is the criterion value of a model at order p of an autoregressive model.

2.2.1 Probability of Overfitting

In the case of overfitting, the true model order is k and a candidate model order is p where $p=k+L$ and $L>0$, L being the amount of overfitting. If $\mathbf{MSC}(k+L) < \mathbf{MSC}(k)$, the candidate model with an order $k+L$ is selected instead of the true model of order k. The model with order $k+L$ is said to be *overfitted*. Then the probability of overfitting by L is given by

$$P\{\mathbf{MSC}(k+L) < \mathbf{MSC}(k)\} \quad . \qquad \ldots\ldots (1)$$

The criterion that gives the lower probability of overfitting is the better one.

2.2.2 Signal-to-Noise Ratio

The signal-to-noise ratio is a measurement which is basically a ratio of the expectation to the standard deviation of the difference in criterion values for two models. The ratio tends to assess whether the penalty term is sufficiently strong in relation to the goodness-of-fit term. From the true model order k and a candidate model order p where $p=k+L$ and $L>0$, the true model is considered better than a candidate model if $\mathbf{MSC}(k) < \mathbf{MSC}(k+L)$. McQuarrie and Tsai (1998: 24) defined the signal as $E[\mathbf{MSC}(k+L) - \mathbf{MSC}(k)]$, and the noise as the standard deviation of the difference, $sd[\mathbf{MSC}(k+L) - \mathbf{MSC}(k)]$. Then the signal-to-noise ratio that the true model is selected compared with a candidate model is defined as

$$\frac{E\left[\mathbf{MSC}(k+L)-\mathbf{MSC}(k)\right]}{sd\left[\mathbf{MSC}(k+L)-\mathbf{MSC}(k)\right]}. \qquad \ldots\ldots (2)$$

The criterion that gives a higher signal-to-noise ratio is the better one. Notice that when the amount of overfitting, L, increases, the signal-to-noise ratio will increase to indicate a higher overfit.

CHAPTER 3
THE PROBABILITY OF OVERFITTING AND THE SIGNAL-TO-NOISE

This chapter show the derivation of two theoretical properties of the $\mathbf{KIC_U}$ (Boosarawongse, 2000: 3)criterion in autoregressive model. Omitting the constant $n\ln 2\pi$, the $\mathbf{KIC_U}$ criterion for AR model at order p is given by

$$\mathbf{KIC_U}(p) = n\ln\hat{s}_p^2 + n\ln\left(\frac{n}{n-p}\right) + \frac{n[(n+p)(n-p)+(n-p-2)]}{(n-p-2)(n-p)}. \qquad (1)$$

where $\hat{s}_p^2$ is an unbiased estimator of error variance of the process with sample of size n.

3.1 Probability of Overfitting

When choosing the best model from the candidate models for each model selection criterion, the best model is assumed to have the lowest selection criterion value. In the case of overfitting, we assume that the true model order is k and a candidate model order is p where $p=k+L$ and $L>0$, L being the amount of overfitting. If the criterion value of a model of order $k+L$ is less than that of order k, $\mathbf{KIC_U}(k+L) < \mathbf{KIC_U}(k)$, the candidate model with an order $k+L$ is selected instead of the true model of order k. The model with order $k+L$ is said to be *overfitted*. Then the probability of overfitting by L for the $\mathbf{KIC_U}$ criterion is given by

$$P\{\mathbf{KIC_U}(k+L) < \mathbf{KIC_U}(k)\} \ . \qquad \ldots\ldots (2)$$

The criterion that gives the lower probability of overfitting is the better one.

From (1), the $\mathbf{KIC}_{\mathrm{U}}$ criteria for an AR model with orders of k and $k+L$ are

$$\mathbf{KIC}_{\mathrm{U}}(k) = n\ln\hat{s}_k^2 + n\ln\left(\frac{n}{n-k}\right) + \frac{n[(n+k)(n-k)+(n-k-2)]}{(n-k-2)(n-k)} \quad \ldots\ldots (3)$$

and

$$\mathbf{KIC}_{\mathrm{U}}(k+L) = n\ln\hat{s}_{k+L}^2 + n\ln\left(\frac{n}{n-k-L}\right) + \frac{n[(n+k+L)(n-k-L)+(n-k-L-2)]}{(n-k-L-2)(n-k-L)} \quad \ldots\ldots (4)$$

respectively. Substituting in (2), the following is obtained;

$$\begin{aligned}
&P\{\mathbf{KIC}_{\mathrm{U}}(k+L) < \mathbf{KIC}_{\mathrm{U}}(k)\} \\
&= P\left\{ n\ln\hat{s}_{k+L}^2 + n\ln\left(\frac{n}{n-k-L}\right) + \frac{n[(n+k+L)(n-k-L)+(n-k-L-2)]}{(n-k-L-2)(n-k-L)} \right. \\
&\qquad \left. < n\ln\hat{s}_k^2 + n\ln\left(\frac{n}{n-k}\right) + \frac{n[(n+k)(n-k)+(n-k-2)]}{(n-k-2)(n-k)} \right\} \\
&P\{\mathbf{KIC}_{\mathrm{U}}(k+L) < \mathbf{KIC}_{\mathrm{U}}(k)\} \\
&= P\left\{ \ln\hat{s}_{k+L}^2 + \ln\left(\frac{n}{n-k-L}\right) + \frac{[(n+k+L)(n-k-L)+(n-k-L-2)]}{(n-k-L-2)(n-k-L)} \right. \\
&\qquad \left. < \ln\hat{s}_k^2 + \ln\left(\frac{n}{n-k}\right) + \frac{[(n+k)(n-k)+(n-k-2)]}{(n-k-2)(n-k)} \right\}
\end{aligned}$$

Since $SSE_k = (n-k)\hat{s}_k^2$ and therefore $\ln SSE_k = \ln(n-k) + \ln\hat{s}_k^2$, then

$$\begin{aligned}
&P\{\mathbf{KIC}_{\mathrm{U}}(k+L) < \mathbf{KIC}_{\mathrm{U}}(k)\} \\
&= P\left\{ \ln SSE_{k+L} - \ln(n-k-L) - \ln(n-k-L) + \frac{[(n+k+L)(n-k-L)+(n-k-L-2)]}{(n-k-L-2)(n-k-L)} \right. \\
&\qquad \left. < \ln SSE_k - \ln(n-k) - \ln(n-k) + \frac{[(n+k)(n-k)+(n-k-2)]}{(n-k-2)(n-k)} \right\}
\end{aligned}$$

$$= P\left\{\ln\frac{SSE_{k+L}}{SSE_k} - 2\ln(n-k-L) + 2\ln(n-k) + \frac{[(n+k+L)(n-k-L)+(n-k-L-2)]}{(n-k-L-2)(n-k-L)}\right.$$
$$\left. < \frac{[(n+k)(n-k)+(n-k-2)]}{(n-k-2)(n-k)}\right\}$$

$$= P\left\{\ln\frac{SSE_k}{SSE_{k+L}} > 2\ln\frac{(n-k)}{(n-k-L)} + \frac{[(n+k+L)(n-k-L)+(n-k-L-2)]}{(n-k-L-2)(n-k-L)}\right.$$
$$\left. - \frac{[(n+k)(n-k)+(n-k-2)]}{(n-k-2)(n-k)}\right\}. \quad \ldots (5)$$

Using the proof from Appendix A.2,

$$\frac{[(n+k+L)(n-k-L)+(n-k-L-2)]}{(n-k-L-2)(n-k-L)} - \frac{[(n+k)(n-k)+(n-k-2)]}{(n-k-2)(n-k)}$$
$$= \frac{L[2(n-1)(n-k)(n-k-L)+(n-k-2)(n-k-L-2)]}{(n-k-L-2)(n-k-L)(n-k-2)(n-k)}$$

$$\text{let} \quad \mathbf{A} = \frac{L[2(n-1)(n-k)(n-k-L)+(n-k-2)(n-k-L-2)]}{(n-k-L-2)(n-k-L)(n-k-2)(n-k)} \quad \ldots (6)$$

Substituting in (5), the following is obtained;

$$P\{\mathbf{KIC_U}(k+L) < \mathbf{KIC_U}(k)\} = P\left\{\ln\frac{SSE_k}{SSE_{k+L}} > 2\ln\frac{(n-k)}{(n-k-L)} + \mathbf{A}\right\}$$
$$= P\left\{\frac{SSE_k}{SSE_{k+L}} > \left\{\frac{(n-k)}{(n-k-L)}\right\}^2 \exp(\mathbf{A})\right\}$$
$$= P\left\{\frac{SSE_k}{SSE_{k+L}} - 1 > \left\{\frac{(n-k)}{(n-k-L)}\right\}^2 \exp(\mathbf{A}) - 1\right\}$$
$$= P\left\{\frac{SSE_k - SSE_{k+L}}{SSE_{k+L}} > \left\{\frac{(n-k)}{(n-k-L)}\right\}^2 \exp(\mathbf{A}) - 1\right\}.$$

From McQuarrie and Tsai (1998: 66), $SSE_k - SSE_{k+L} \sim \sigma_k^2\chi_L^2$, $SSE_{k+L} \sim \sigma_k^2\chi_{n-k-L}^2$ and $SSE_k - SSE_{k+L}$ are independent of SSE_{k+L}.

$$P\{\mathbf{KIC}_{\mathrm{U}}(k+L) < \mathbf{KIC}_{\mathrm{U}}(k)\}$$

$$= P\left\{\frac{\chi_L^2}{\chi_{n-k-L}^2} > \left\{\frac{(n-k)}{(n-k-L)}\right\}^2 \exp(\mathbf{A}) - 1\right\}$$

$$= P\left\{F_{L,n-k-L} > \left(\frac{n-k-L}{L}\right) \times \left(\left\{\frac{(n-k)}{(n-k-L)}\right\}^2 \exp(\mathbf{A}) - 1\right)\right\} \quad \ldots (7)$$

3.2 The asymptotic probability of overfitting

From McQuarrie and Tsai (1998: 41), for a fixed k and L, and where n approaches infinity, $\frac{\chi_{n-k-L}^2}{(n-k-L)} \to 1$, $F_{L,n-k-L} \to \frac{\chi_L^2}{L}$, and if $\lim_{n\to\infty} f_n \to f$ then $\lim_{n\to\infty} P\{F_{L,n-k-L} > f_n\} \to P\{\chi_L^2 > fL\}$, and $\exp(x) = 1 + x + \sum_{i=2}^{\infty} \frac{x^i}{i!}$. is replaced by $\exp(x) = 1 + x + o(x^2)$ for $0 \le x \le 1$. Then from (7), if

$$f_n = \left(\frac{n-k-L}{L}\right) \times \left(\left\{\frac{(n-k)}{(n-k-L)}\right\}^2 \exp(\mathbf{A}) - 1\right)$$

where $\mathbf{A} = \frac{L[2(n-1)(n-k)(n-k-L) + (n-k-2)(n-k-L-2)]}{(n-k-L-2)(n-k-L)(n-k-2)(n-k)}$.

then

$$\lim_{n\to\infty} f_n \simeq \lim_{n\to\infty}\left[\left(\frac{n-k-L}{L}\right) \times \left(\left\{\frac{(n-k)}{(n-k-L)}\right\}^2 \exp(\mathbf{A}) - 1\right)\right]$$

$$= \lim_{n\to\infty}\left[\left(\frac{n-k-L}{L}\right) \times \left(\left\{\frac{(n-k)}{(n-k-L)}\right\}^2 \left\{1 + \mathbf{A} + o\left(\mathbf{A}^2\right)\right\} - 1\right)\right]$$

$$= \lim_{n\to\infty}\left[\left(\frac{n-k-L}{L}\right) \times \left\{\left\{\frac{(n-k)}{(n-k-L)}\right\}^2 + \mathbf{A}\left\{\frac{(n-k)}{(n-k-L)}\right\}^2 - 1\right\}\right]$$

$$= \lim_{n\to\infty}\left[\left(\frac{n-k-L}{L}\right)\times\left\{\frac{(n-k)^2-(n-k-L)^2}{(n-k-L)^2}+\mathbf{A}\left\{\frac{(n-k)}{(n-k-L)}\right\}^2\right\}\right]$$

$$= \lim_{n\to\infty}\left[\left(\frac{n-k-L}{L}\right)\times\left(\frac{2(n-k)L-L^2}{(n-k-L)^2}+\mathbf{A}\left\{\frac{(n-k)}{(n-k-L)}\right\}^2\right)\right]$$

$$= \lim_{n\to\infty}\left\{\left[\frac{n-k-L}{L}\times\frac{2(n-k)L-L^2}{(n-k-L)^2}\right]+\mathbf{A}\left(\frac{n-k-L}{L}\right)\times\left\{\frac{(n-k)}{(n-k-L)}\right\}^2\right\}$$

$$= \lim_{n\to\infty}\left\{\left[\frac{2(n-k)-L}{(n-k-L)}\right]+\mathbf{A}\left[\frac{(n-k)^2}{L(n-k-L)}\right]\right\}.$$

By substituting **A**, the following is obtained;

$$\lim_{n\to\infty} f_n = \lim_{n\to\infty}\left[\frac{2(n-k)-L}{(n-k-L)}\right]+\lim_{n\to\infty}\left\{\mathbf{A}\left[\frac{(n-k)^2}{L(n-k-L)}\right]\right\}$$

$$\simeq 2+\lim_{n\to\infty}\left\{\frac{L\left[2(n-1)(n-k)(n-k-L)+(n-k-2)(n-k-L-2)\right]}{(n-k-L-2)(n-k-L)(n-k-2)(n-k)}\left[\frac{(n-k)^2}{L(n-k-L)}\right]\right\}$$

$$= 2+\lim_{n\to\infty}\left\{\frac{\left[2(n-1)(n-k)^2(n-k-L)+(n-k-2)(n-k-L-2)(n-k)\right]}{(n-k-L-2)(n-k-L)^2(n-k-2)}\right\}$$

$$= 2+2$$

$$= 4\,.$$

Now, $\lim_{n\to\infty} f_n \to 4$ is obtained giving rise to

$$\lim_{n\to\infty} P\left\{F_{L,n-k-L} > f_n\right\} \to P\left\{\chi_L^2 > 4L\right\}. \qquad \ldots\ldots (8)$$

Therefore, the asymptotic probability of overfitting of the $\mathbf{KIC_U}$ criterion by L is $P\left\{\chi_L^2 > 4L\right\}$. Notice that $P\left\{\chi_L^2 > 4L\right\}$ decreases as the amount of overfitting L increases.

3.3 Signal-to-Noise Ratio

The signal-to-noise ratio is a measurement which is basically a ratio of the expectation to the standard deviation of the difference in criterion values for two models. The ratio tends to assess whether the penalty term is sufficiently strong in relation to the goodness-of-fit term. From the true model order k and a candidate model order p where $p=k+L$ and $L>0$, the true model is considered better than a candidate model if the criterion value of a model of order k is less than that of order $k+L$, $\mathbf{KIC}_{\mathrm{U}}(k) < \mathbf{KIC}_{\mathrm{U}}(k+L)$. For the $\mathbf{KIC}_{\mathrm{U}}$ criterion, McQuarrie and Tsai (1998: 24) defined the signal as $E\left[\mathbf{KIC}_{\mathrm{U}}(k+L)-\mathbf{KIC}_{\mathrm{U}}(k)\right]$, and the noise as the standard deviation of the difference, $sd\left[\mathbf{KIC}_{\mathrm{U}}(k+L)-\mathbf{KIC}_{\mathrm{U}}(k)\right]$. Then the signal-to-noise ratio that the true model is selected compared with a candidate model is defined as

$$\frac{E\left[\mathbf{KIC}_{\mathrm{U}}(k+L)-\mathbf{KIC}_{\mathrm{U}}(k)\right]}{sd\left[\mathbf{KIC}_{\mathrm{U}}(k+L)-\mathbf{KIC}_{\mathrm{U}}(k)\right]}. \qquad \ldots\ldots (9)$$

The criterion that gives a higher signal-to-noise ratio is the better one. Notice that when the amount of overfitting, L, increases, the signal-to-noise ratio will increase to indicate a higher overfit.

Using (3), (4) and $E(\ln \hat{s}_k^2) \simeq \ln \sigma_k^2 - \frac{1}{n-k}$ from the proof in Appendix A.3, the following is obtained;

$$\begin{aligned} E(\mathbf{KIC}_{\mathrm{U}}(k)) &\simeq n\ln\sigma_k^2 - \frac{n}{n-k} + n\ln\left(\frac{n}{n-k}\right) \\ &\quad + \frac{n[(n+k)(n-k)+(n-k-2)]}{(n-k-2)(n-k)} \\ &= n\ln\sigma_k^2 - \frac{n}{n-k}n\ln\left(\frac{n}{n-k}\right) + \frac{n[(n+k)(n-k)+(n-k-2)]}{(n-k-2)(n-k)} \end{aligned} \qquad \ldots\ldots (10)$$

and

$$E(\mathbf{KIC}_{\mathrm{U}}(k+L)) \simeq n\ln\sigma_k^2 - \frac{n}{n-k-L} + n\ln\left(\frac{n}{n-k-L}\right) + \frac{n[(n+k+L)(n-k-L)+(n-k-L-2)]}{(n-k-L-2)(n-k-L)}$$

$$E(\mathbf{KIC}_{\mathrm{U}}(k+L)) \simeq n\ln\sigma_k^2 - \frac{n}{n-k-L} + n\ln\left(\frac{n}{n-k-L}\right) + \frac{n[(n+k+L)(n-k-L)+(n-k-L-2)]}{(n-k-L-2)(n-k-L)}. \quad \ldots\ldots (11)$$

Then,

$$\begin{aligned}
&E(\mathbf{KIC}_{\mathrm{U}}(k+L)) - E(\mathbf{KIC}_{\mathrm{U}}(k)) \\
&\simeq \left(n\ln\sigma_k^2 - \frac{n}{n-k-L} + n\ln\left(\frac{n}{n-k-L}\right) + \frac{n[(n+k+L)(n-k-L)+(n-k-L-2)]}{(n-k-L-2)(n-k-L)}\right) \\
&\quad - \left(n\ln\sigma_k^2 - \frac{n}{n-k} + n\ln\left(\frac{n}{n-k}\right) + \frac{n[(n+k)(n-k)+(n-k-2)]}{(n-k-2)(n-k)}\right) \\
&= -\frac{n}{n-k-L} + \frac{n}{n-k} + n\ln\left(\frac{n-k}{n-k-L}\right) \\
&\quad + \frac{nL[2(n-1)(n-k)(n-k-L)+(n-k-2)(n-k-L-2)]}{(n-k-L-2)(n-k-L)(n-k-2)(n-k)} \\
&= \frac{-nL}{(n-k-L)(n-k)} + n\ln\left(\frac{n-k}{n-k-L}\right) \\
&\quad + \frac{nL[2(n-1)(n-k)(n-k-L)+(n-k-2)(n-k-L-2)]}{(n-k-L-2)(n-k-L)(n-k-2)(n-k)}.
\end{aligned}$$

Therefore, the signal is evaluated as

$$\text{signal} \simeq \frac{-nL}{(n-k-L)(n-k)} + n\ln\left(\frac{n-k}{n-k-L}\right) + \frac{nL[2(n-1)(n-k)(n-k-L)+(n-k-2)(n-k-L-2)]}{(n-k-L-2)(n-k-L)(n-k-2)(n-k)}. \quad \ldots\ldots (12)$$

To measure the noise;

$$
\begin{aligned}
&sd\left[\mathbf{KIC}_{\mathrm{U}}(k+L)-\mathbf{KIC}_{\mathrm{U}}(k)\right] \\
&= sd\left[\left(n\ln\hat{s}^2_{k+L} + n\ln\left(\frac{n}{n-k-L}\right) + \frac{n[(n+k+L)(n-k-L)+(n-k-L-2)]}{(n-k-L-2)(n-k-L)}\right)\right. \\
&\qquad \left. -\left(n\ln\hat{s}^2_{k} + n\ln\left(\frac{n}{n-k}\right) + \frac{n[(n+k)(n-k)+(n-k-2)]}{(n-k-2)(n-k)}\right)\right] \\
&= sd\left[n\ln\hat{s}^2_{k+L} - n\ln\hat{s}^2_{k} + \text{constant}\right]
\end{aligned}
$$

Again, since $SSE_k = (n-k)\hat{s}^2_k$ then $\ln SSE_k = \ln(n-k) + \ln\hat{s}^2_k$.

$$
\begin{aligned}
&= sd\left[n\ln SSE_{k+L} - n\ln SSE_k \right] \\
&= sd\left[n\ln\frac{SSE_{k+L}}{SSE_k}\right].
\end{aligned}
$$

McQuarrie and Tsai (1998: 70) suggested an approximation of $Var\left(n\ln\frac{SSE_{k+L}}{SSE_k}\right)$ by $\frac{2n^2L}{(n-k-L)(n-k+2)}$. Then the noise becomes

$$
\text{noise} \simeq \frac{n\sqrt{2L}}{\sqrt{(n-k-L)(n-k+2)}}. \qquad \ldots\ldots (13)
$$

Therefore, the signal-to-noise ratio of the $\mathbf{KIC}_{\mathrm{U}}$ criterion for the AR(k) model is given by

$$
\begin{aligned}
&\frac{E(\mathbf{KIC}_{\mathrm{U}}(k+L)) - E(\mathbf{KIC}_{\mathrm{U}}(k))}{sd\left[\mathbf{KIC}_{\mathrm{U}}(k+L)-\mathbf{KIC}_{\mathrm{U}}(k)\right]} \\
&\qquad \simeq \frac{\sqrt{(n-k-L)(n-k+2)}}{n\sqrt{2L}} \times \left[\frac{-nL}{(n-k-L)(n-k)} + n\ln\left(\frac{n-k}{n-k-L}\right)\right. \\
&\qquad \left. + \frac{nL\left[2(n-1)(n-k)(n-k-L)+(n-k-2)(n-k-L-2)\right]}{(n-k-L-2)(n-k-L)(n-k-2)(n-k)}\right].
\end{aligned} \qquad (14)
$$

3.4 The asymptotic signal-to-noise ratio

For a fixed k and L, and where n approaches infinity,

$$\lim_{n\to\infty}\left\{\frac{\sqrt{(n-k-L)(n-k+2)}}{n\sqrt{2L}}\times\left[\frac{-nL}{(n-k-L)(n-k)}+n\ln\left(\frac{n-k}{n-k-L}\right)\right.\right.$$
$$\left.\left.+\frac{nL\left[2(n-1)(n-k)(n-k-L)+(n-k-2)(n-k-L-2)\right]}{(n-k-L-2)(n-k-L)(n-k-2)(n-k)}\right]\right\}$$

$$=\lim_{n\to\infty}\left\{\frac{\sqrt{(n-k-L)(n-k+2)}}{n\sqrt{2L}}\times\left\{\frac{-nL}{(n-k-L)(n-k)}+n\ln\left(\frac{n-k}{n-k-L}\right)\right\}\right.$$
$$+\left[\frac{\sqrt{(n-k-L)(n-k+2)}}{n\sqrt{2L}}\right]\times\left[\frac{2nL(n-1)(n-k)(n-k-L)}{(n-k-L-2)(n-k-L)(n-k-2)(n-k)}\right]$$
$$\left.+\left[\frac{\sqrt{(n-k-L)(n-k+2)}}{n\sqrt{2L}}\right]\frac{nL(n-k-2)(n-k-L-2)}{(n-k-L-2)(n-k-L)(n-k-2)(n-k)}\right]\right\}$$

For a fixed k and L, and where n approaches infinity, McQuarrie and Tsai (1998: 70) suggested an approximation of $\ln\left(1-\frac{L}{n}\right)\simeq\frac{-L}{n}$ when $L<< n$. Then

$$=\lim_{n\to\infty}\frac{\sqrt{(n-k-L)(n-k+2)}}{n\sqrt{2L}}\times\left\{n\ln\left(\frac{n-k}{n-k-L}\right)\right\}$$
$$+\lim_{n\to\infty}\left\{\left[\frac{\sqrt{(n-k-L)(n-k+2)}}{n\sqrt{2L}}\right]\times\left[\frac{2nL(n-1)(n-k)(n-k-L)}{(n-k-L-2)(n-k-L)(n-k-2)(n-k)}\right]\right\}$$
$$\simeq\lim_{n\to\infty}\frac{\sqrt{(n-k-L)(n-k+2)}}{\sqrt{2L}}\times\frac{L}{n}$$
$$+\lim_{n\to\infty}\left\{\frac{2nL\sqrt{(n-k-L)(n-k+2)}(n-1)(n-k)(n-k-L)}{n\sqrt{2L}(n-k-L-2)(n-k-L)(n-k-2)(n-k)}\right\}$$
$$=\frac{L}{\sqrt{2L}}+\frac{2L}{\sqrt{2L}}$$
$$=\frac{3L}{\sqrt{2L}}. \qquad \ldots\ldots (15)$$

Therefore, the asymptotic signal-to-noise ratio of the $\mathbf{KIC_U}$ criterion for the AR model is $\frac{3L}{\sqrt{2L}}$. Notice that $\frac{3L}{\sqrt{2L}}$ increases as the amount of overfitting L increases.

CHAPTER 4
SIMULATION STUDY AND CONCLUSIONS

4.1 Simulation Study

In this section, the probability of overfitting and the signal-to-noise ratio of the $\mathbf{KIC_U}$ criterion were tested by comparing them with a number of considered criteria using simulation. The considered criteria are **AIC** (Akaike, 1973: 273), $\mathbf{AIC_C}$ (Hurvich and Tsai, 1989: 301), $\mathbf{AIC_U}$ (Hurvich, Shumway and Tsai, 1990: 711), **SBC** (Schwarz, 1978: 462), **KIC** (Cavanaugh, 1999: 340) and $\mathbf{KIC_C}$ (Boosarawongse, 2004: 2) which their probability of overfitting and signal-to-noise ratio are given in McQuarrie and Tsai(1998: 25-40).

AR(1), AR(2) and AR(3) models with sample sizes 25, 40, 60 and 100, ranging from small to large, were used. All results are shown in tables 1-6.

The data shown in table 1 are the probability of overfitting for the AR(1) model by each criteria for different amounts of overfitting L and different sample sizes, e.g. for $\mathbf{KIC_U}$ where $n = 25$ and $L = 1$ the probability of overfitting was 0.0382. This means that this criterion would select the model whose order is higher by one order than true model, AR(1), with a probability of 0.0382. It can be seen that, where $n = 25$, $\mathbf{KIC_U}$ is the best among the selection criteria as it had the lowest probability of overfitting; 0.0382 and 0.0128 for $L = 1$ and 2 respectively. Where $n = 40$, $\mathbf{KIC_U}$ is the best with the lowest probability of overfitting; 0.0411, 0.0150 and 0.0053 for $L = 1$, 2 and 3 respectively. Where $n = 60$, $\mathbf{KIC_U}$ performs the best with the lowest probability of overfitting; 0.0427, 0.0161 and 0.0060 for $L = 1$, 2 and 3 respectively. Where $n = 100$, the best criterion is **SBC** with the lowest probability of overfitting; 0.0341, 0.0115 and 0.0040 for $L = 1$, 2 and 3 respectively.

The data shown in table 4 are the signal-to-noise ratios for the AR(1) model for each criteria using different amounts of overfitting L for different sample sizes, e.g. for $\mathbf{KIC_U}$, where $n = 25$ and $L = 1$, the signal-to-noise ratio was 2.5325. This

means that this criterion will select the model whose order is higher by one order than the true model, AR(1), with a signal-to-noise ratio of 2.5325. It was found that, where $n = 25$, $\mathbf{KIC_U}$ has the highest signal-to-noise ratio; 2.5325 and 3.6494 for $L = 1$ and 2 respectively. Where $n = 40$, the highest signal-to-noise ratio was generated by $\mathbf{KIC_U}$; 2.3594, 3.3711 and 4.1729 for $L = 1$, 2 and 3 respectively. Where $n = 60$, $\mathbf{KIC_U}$ generated the highest signal-to-noise ratio; 2.2738, 3.2362 and 3.9894 for $L = 1$, 2 and 3 respectively. Where $n = 100$, the highest signal-to-noise ratio was generated by **SBC**; 2.5182, 3.5378 and 4.3041 for $L = 1$, 2 and 3 respectively.

4.2 Conclusions

For sample small to medium sizes, the $\mathbf{KIC_U}$ criterion had the lowest probability of overfitting and the highest signal-to-noise-ratio. However, as the sample size increases, **SBC** had the lowest probability of overfitting and the highest signal-to-noise-ratio.

Therefore, for small to medium sample sizes, the $\mathbf{KIC_U}$ performed the best, and the sample size is large **SBC** is the best.

Table 1 the probability of overfitting for AR(1) model

n	L	probability of overfitting						
		AIC	**AIC_C**	**AIC_U**	**SBC**	**KIC**	**KIC_C**	**KIC_U**
25	1	0.1796	0.1262	0.0787	0.0887	0.1003	0.0677	0.0382
	2	0.1720	0.0907	0.0509	0.0589	0.0714	0.0334	0.0128
	3	0.1608	0.0602	0.0363	0.0403	0.0519	0.0155	0.0041
	4	0.1524	0.0380	0.0312	0.0291	0.0396	0.0069	0.0013
40	1	0.1709	0.1383	0.0511	0.0629	0.0935	0.0738	0.0411
	2	0.1572	0.1077	0.0230	0.0330	0.0623	0.0396	0.0150
	3	0.1402	0.0790	0.0108	0.0178	0.0419	0.0205	0.0053
	4	0.1260	0.0565	0.0091	0.0101	0.0290	0.0104	0.0019
60	1	0.1662	0.1448	0.0489	0.0476	0.0900	0.0771	0.0427
	2	0.1496	0.1170	0.0301	0.0205	0.0578	0.0431	0.0161
	3	0.1300	0.0898	0.0198	0.0090	0.0372	0.0234	0.0060
	4	0.1133	0.0678	0.0074	0.0042	0.0245	0.0126	0.0022
100	1	0.1626	0.1499	0.0431	0.0341	0.0872	0.0796	0.0438
	2	0.1437	0.1244	0.0128	0.0115	0.0545	0.0458	0.0170
	3	0.1223	0.0985	0.0091	0.0040	0.0338	0.0257	0.0066
	4	0.1041	0.0771	0.0053	0.0014	0.0213	0.0144	0.0025

Table 2 the probability of overfitting for AR(2) model

n	L	probability of overfitting						
		AIC	**AIC_C**	**AIC_U**	**SBC**	**KIC**	**KIC_C**	**KIC_U**
25	1	0.1896	0.1169	0.0873	0.0961	0.1081	0.0630	0.0356
	2	0.1864	0.0800	0.0567	0.0669	0.0805	0.0295	0.0113
	3	0.1781	0.0503	0.0421	0.0479	0.0610	0.0129	0.0034
	4	0.1723	0.0299	0.0268	0.0362	0.0484	0.0054	0.0010
40	1	0.1767	0.1328	0.0529	0.0665	0.0980	0.0710	0.0396
	2	0.1653	0.1009	0.0230	0.0362	0.0672	0.0371	0.0140
	3	0.1496	0.0722	0.0178	0.0202	0.0463	0.0187	0.0048
	4	0.1362	0.0504	0.0081	0.0118	0.0330	0.0092	0.0017
60	1	0.1699	0.1412	0.0495	0.0496	0.0928	0.0753	0.0417
	2	0.1546	0.1125	0.0245	0.0219	0.0608	0.0414	0.0155
	3	0.1357	0.0851	0.0109	0.0099	0.0398	0.0221	0.0057
	4	0.1194	0.0634	0.0063	0.0047	0.0266	0.0117	0.0021
100	1	0.1648	0.1478	0.0442	0.0350	0.0889	0.0785	0.0433
	2	0.1466	0.1217	0.0215	0.0120	0.0561	0.0448	0.0166
	3	0.1255	0.0956	0.0180	0.0042	0.0352	0.0249	0.0064
	4	0.1074	0.0743	0.0054	0.0015	0.0224	0.0139	0.0024

Table 5 signal to noise ratio for AR(2) model

n	L	signal to noise ratio						
		AIC	**AIC_C**	**AIC_U**	**SBC**	**KIC**	**KIC_C**	**KIC_U**
25	1	0.5567	1.1253	1.8624	1.3652	1.2201	1.8569	2.5792
	2	0.7434	1.6668	2.7090	1.8605	1.6599	2.6439	3.6435
	3	0.8555	2.1422	3.4181	2.1907	1.9510	3.2623	4.4591
	4	0.9231	2.6014	4.0736	2.4258	2.1559	3.7982	5.1474
40	1	0.6153	0.9392	1.6646	1.7639	1.2954	1.6628	2.3791
	2	0.8438	1.3646	2.3905	2.4460	1.7925	2.3570	3.3569
	3	1.0006	1.7181	2.9743	2.9355	2.1463	2.8937	4.1019
	4	1.1171	2.0407	3.4909	3.3191	2.4210	3.3500	4.7258
60	1	0.6467	0.8526	1.5718	2.0901	1.3359	1.5710	2.2842
	2	0.8973	1.2270	2.2440	2.9206	1.8634	2.2239	3.2238
	3	1.0776	1.5296	2.7751	3.5335	2.2502	2.7264	3.9404
	4	1.2195	1.7982	3.2363	4.0294	2.5611	3.1513	4.5407
100	1	0.6712	0.7904	1.5046	2.4855	1.3676	1.5044	2.2151
	2	0.9390	1.1293	2.1394	3.4916	1.9188	2.1282	3.1282
	3	1.1375	1.3974	2.6346	4.2474	2.3312	2.6074	3.8258
	4	1.2989	1.6305	3.0590	4.8709	2.6700	3.0118	4.4114

Table 6 signal to noise ratio for AR(3) model

n	L	signal to noise ratio						
		AIC	**AIC_C**	**AIC_U**	**SBC**	**KIC**	**KIC_C**	**KIC_U**
25	1	0.4971	1.2324	1.9708	1.2711	1.1321	1.9630	2.6859
	2	0.6588	1.8273	2.8714	1.7270	1.5352	2.7968	3.7964
	3	0.7516	2.3516	3.6297	2.0268	1.7978	3.4538	4.6493
	4	0.8027	2.8602	4.3349	2.2358	1.9784	4.0249	5.3711
40	1	0.5789	0.9908	1.7168	1.6976	1.2413	1.7143	2.4308
	2	0.7922	1.4400	2.4665	2.3522	1.7159	2.4302	3.4301
	3	0.9374	1.8135	3.0705	2.8204	2.0524	2.9841	4.1918
	4	1.0440	2.1546	3.6057	3.1862	2.3124	3.4552	4.8299
60	1	0.6227	0.8827	1.6021	2.0414	1.3001	1.6011	2.3144
	2	0.8633	1.2704	2.2878	2.8517	1.8127	2.2665	3.2665
	3	1.0360	1.5839	2.8298	3.4490	2.1882	2.7788	3.9926
	4	1.1714	1.8621	3.3006	3.9318	2.4894	3.2121	4.6009
100	1	0.6569	0.8067	1.5210	2.4528	1.3463	1.5207	2.2315
	2	0.9188	1.1526	2.1629	3.4453	1.8886	2.1513	3.1513
	3	1.1128	1.4264	2.6636	4.1907	2.2942	2.6357	3.8541
	4	1.2703	1.6643	3.0929	4.8055	2.6273	3.0445	4.4440

BIBLIOGRAPHY

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

_____. 1978. Time Series analysis and control through parametic models, in : D.F. Findley, ed., **Applied Time Series Analysis.** New York, Academic Press, Inc. : 1-23.

Bhansali, R.J. and Downham, D.Y. 1977. Some properties of the order of an autoregressive model selected by a generalization of Akaike's FPE criterian. **Biometrika.** 64: 547-551.

Bhansali, R.J. 1986. Asymptotically efficient selection of the order by the criterian autoregressive transfer function. **Annals of Statistics.** 14.: 315-325.

_____. 1993. Order selection for linear time series models: A review, in: T.S. Rao, ed. **Developments in Time Series Analysis.** (Chapman and Hall, London), 50-66.

Boosarawongse R. and Chongcharoen S. 2003. Model Selection Criteria for Autoregressive Models. **The Proceeding of the 8th APDSI,** 4th– 8th July, China.

Boosarawongse R., 2004. The $\mathbf{KIC_U}$ Model Selection Criteria for Autoregressive Models. **The Proceeding of the 9th APDSI,** 1st– 4th July, Korea.

Box, G.E.P., Jenkins, G.M. and Reinsel, G.C. 1994. **Time Series Analysis, Forcasting and control.** 3rd ed. New Jersey: Prentice-Hall, Inc.

Brokwell, P.J. and Davis, R.A. 1991. **Time Series : Theory and Methods.** 2nd ed. New York: Springer-Verlag, Inc.

Cavanaugh, J.E. 1997. Unifying the derivations of the Akaike and Corrected Akaike information criteria. **Statistics & Probability Letters.** 33: 201-208.

Cavanaugh, J.E. and Shumway, R.H. 1997. A bootstrap varriant of AIC for state-space model selection. **Statistica Sinica.** 7: 473-796.

Cavanaugh, J.E. 1999. A large-sample model selection criterion based on Kullback's symmetric divergence. **Statistics & Probability letter.** 42: 333-343.

_____. 2000. Criterion for linear model selection based on Kullback's symmetric divergence. Technical Report, Department of Statistics. University of Missouri- Columbia.

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

Tsay, S.R. 1984. Order selection in nonstationary autoregressive models. **Annals of Statistics.** 12: 1425-1433.

Wei, W.S. 1990. **Time Series Analysis**. New York, Addison-Wesley, Inc.

Whittle, P . 1953. The analysis of multiple stationary time series. **Journal of the Royal Statistical Society, Series B.** 15: 125-139.

APPENDICES

A.1 $\left[(n+k+L)(n-k-2)-(n+k)(n-k-L-2)\right]=2L(n-1)$

Proof.

$$\begin{aligned}&\left[(n+k+L)(n-k-2)-(n+k)(n-k-L-2)\right]\\&\qquad=n^2-nk-2n+nk-k^2-2k+nL-kL-2L\\&\qquad\quad-(n^2-nk-nL-2n+nk-k^2-kL-2k)\\&\qquad=2L(n-1).\end{aligned}$$

A.2

$$\frac{[(n+k+L)(n-k-L)+(n-k-L-2)]}{(n-k-L-2)(n-k-L)} - \frac{[(n+k)(n-k)+(n-k-2)]}{(n-k-2)(n-k)}$$
$$= \frac{L\left[2(n-1)(n-k)(n-k-L)+(n-k-2)(n-k-L-2)\right]}{(n-k-L-2)(n-k-L)(n-k-2)(n-k)}$$

Proof.

$$\frac{[(n+k+L)(n-k-L)+(n-k-L-2)]}{(n-k-L-2)(n-k-L)} - \frac{[(n+k)(n-k)+(n-k-2)]}{(n-k-2)(n-k)}$$
$$= \frac{\left[(n+k+L)(n-k-L)+(n-k-L-2)\right]\left[(n-k-2)(n-k)\right]}{(n-k-L-2)(n-k-L)(n-k-2)(n-k)}$$
$$- \frac{[(n+k)(n-k)+(n-k-2)]\left[(n-k-L-2)(n-k-L)\right]}{(n-k-L-2)(n-k-L)(n-k-2)(n-k)}$$
$$= \frac{\left[(n+k+L)(n-k-L)(n-k-2)(n-k)\right]+\left[(n-k-L-2)(n-k-2)(n-k)\right]}{(n-k-L-2)(n-k-L)(n-k-2)(n-k)}$$
$$- \frac{\left[(n+k)(n-k)(n-k-L-2)(n-k-L)\right]+\left[(n-k-2)(n-k-L-2)(n-k-L)\right]}{(n-k-L-2)(n-k-L)(n-k-2)(n-k)}$$
$$= \frac{(n-k)(n-k-L)\left[(n+k+L)(n-k-2)-(n+k)(n-k-L-2)\right]}{(n-k-L-2)(n-k-L)(n-k-2)(n-k)}$$
$$+ \frac{(n-k-2)(n-k-L-2)\left[(n-k)-(n-k-L)\right]}{(n-k-L-2)(n-k-L)(n-k-2)(n-k)}.$$

$$= \frac{(n-k)(n-k-L)\left[(n+k+L)(n-k-2)-(n+k)(n-k-L-2)\right]}{(n-k-L-2)(n-k-L)(n-k-2)(n-k)}$$
$$+ \frac{(n-k-2)(n-k-L-2)\left[(n-k)-(n-k-L)\right]}{(n-k-L-2)(n-k-L)(n-k-2)(n-k)}$$

From Appendix A.1, $\left[(n+k+L)(n-k-2)-(n+k)(n-k-L-2)\right]=2L(n-1)$

$$= \frac{(n-k)(n-k-L)\left[2L(n-1)\right]}{(n-k-L-2)(n-k-L)(n-k-2)(n-k)}$$
$$+ \frac{(n-k-2)(n-k-L-2)\left[L\right]}{(n-k-L-2)(n-k-L)(n-k-2)(n-k)}$$
$$= \frac{L\left[2(n-1)(n-k)(n-k-L)+(n-k-2)(n-k-L-2)\right]}{(n-k-L-2)(n-k-L)(n-k-2)(n-k)}.$$

A.3 $E(\ln \hat{s}_k^2) \simeq \ln \sigma_k^2 - \frac{1}{n-k}$ and $E_{\theta_k}\left[\ln\ s_p^2\right] \simeq \ln \sigma_k^2 - \frac{1}{n-p}$.

Proof.

From Brockwell and Davis (1991: 302), $\frac{n\hat{\sigma}_p^2}{\sigma_k^2}$ is approximately distributed as $\chi^2_{(n-p)}$. Let $n\hat{\sigma}_p^2 = SSE_p$, then SSE_p is approximately distributed as $\sigma_k^2 \chi^2_{(n-p)}$, and

$$\ln SSE_p = \ln \sigma_k^2 + \ln x \qquad \text{where } x \sim \chi^2_{n-p}.$$

Let $z \sim \chi_n^2$

$$E(\ln z) = \int_0^\infty \ln z \frac{1}{2^{\frac{n}{2}}\Gamma(\frac{n}{2})} z^{\frac{n}{2}-1} e^{\frac{-z}{2}} dz$$

Applying the solution from Gradshteyn (1965: 576);

$$\int_0^\infty \ln z \frac{1}{2^{\frac{n}{2}}\Gamma(\frac{n}{2})} z^{\frac{n}{2}-1} e^{\frac{-z}{2}} dz = \psi(\frac{n}{2}) - \ln(\frac{1}{2})$$

$$= \psi(\frac{n}{2}) + \ln 2$$

where $\psi(\frac{n}{2}) = -c - \sum_{j=0}^{\infty}\left(\frac{1}{j+\frac{n}{2}} - \frac{1}{j+1}\right)$, $c = 0.577215664901$ is Euler's constant

and ψ is Euler's psi function. Then there is no closed form solution of E (ln z).

McQuarrie and Tsai (1998: 69) suggested using Taylor expansions for ln z at $E(z)$, for $z \sim \chi_n^2$, E (z) = n, which gives

$$\ln z \simeq \ln(E(z)) + \ln'(E(z))(z - E(z)) - \frac{1}{2}(\ln' E(z))^2 (z - E(z))^2$$

$$= \ln n + \frac{1}{n}(z-n) - \frac{1}{2n^2}(z-n)^2$$

$$E(\ln z) = \ln n + \frac{1}{n}E(z-n) - \frac{1}{2n^2}E(z-n)^2$$
$$= \ln n - \frac{1}{2n^2}E(z-n)^2$$
$$= \ln n - \frac{1}{2n^2}(E(z^2) - 2nE(z) + n^2)$$
$$= \ln n - \frac{1}{2n^2}(E(z^2) - n^2)$$

To find $E(z^2)$;

$$E(z^2) = \int_0^\infty z^2 \frac{1}{2^{\frac{n}{2}}\Gamma(\frac{n}{2})} z^{\frac{n}{2}-1} e^{\frac{-z}{2}} dz$$
$$= \frac{2^{(\frac{n}{2}+2)}\Gamma(\frac{n}{2}+2)}{2^{\frac{n}{2}}\Gamma(\frac{n}{2})} \int_0^\infty \frac{z^{(\frac{n}{2}+2)-1}}{2^{(\frac{n}{2}+2)}\Gamma(\frac{n}{2}+2)} e^{\frac{-z}{2}} dz$$
$$= 2^2(\frac{n}{2}+1)(\frac{n}{2})$$
$$= 4(\frac{n^2}{4} + \frac{n}{2})$$
$$= n^2 + 2n.$$

Then
$$E(\ln z) \simeq \ln n - \frac{1}{2n^2}(n^2 + 2n - n^2)$$
$$= \ln n - \frac{1}{n}.$$

By applying the solution where $x \sim \chi^2_{n-p}$, then $E(\ln x) \simeq \ln(n-p) - \frac{1}{(n-p)}$.

Therefore,

$$E_{\theta_k}(\ln SSE_p) \simeq \ln \sigma_k^2 + \ln(n-p) - \frac{1}{n-p}$$

or

$$E_{\theta_k}(\ln \hat{\sigma}_p^2) \simeq \ln \sigma_k^2 + \ln \frac{(n-p)}{n} - \frac{1}{n-p}.$$

From $s_p^2 = \frac{(n-p-1)\tilde{\sigma}_p^2}{n-p}$ where $\tilde{\sigma}_p^2 = \frac{n\hat{\sigma}_p^2}{n-p-1}$, then $\frac{n-p}{n}s_p^2 = \hat{\sigma}_p^2$ and

$$E_{\theta_k}\left[\ln\left(\frac{n-p}{n}s_p^2\right)\right] \simeq \ln\sigma_k^2 + \ln\frac{(n-p)}{n} - \frac{1}{n-p}.$$

$$E_{\theta_k}\left[\ln s_p^2 + \ln\frac{(n-p)}{n}\right] \simeq \ln\sigma_k^2 + \ln\frac{(n-p)}{n} - \frac{1}{n-p}.$$

$$E_{\theta_k}\left[\ln s_p^2\right] \simeq \ln\sigma_k^2 - \frac{1}{n-p}.$$

and

$$E(\ln\hat{s}_k^2) \simeq \ln\sigma_k^2 - \frac{1}{n-k}$$