Prior Free Bayesian Estimation through the AIC

Junnan He, Werner Ploberger *

Department of Economics

Washington University in St. Louis

Campus Box 1208, St. Louis MO 63130-4899

October 8, 2018

Abstract

In this paper we show that, in linear models with an increasing number of parameters, the estimator resulting from the maximization of Akaike's Information criterion is asymptotically equivalent to some Bayesian estimators. The family of prior distributions which generates our estimators are normal distributions, defined on the space of all sequence, and is characterized by an exponential decay of the variance for the higher order components of the parameter.

1 Introduction

In the case of finite-dimensional parameters, the theory of optimal estimation is already well established and the theory is well presented in textbooks like van der Vaardt (2000) and Strasser (1996). However, there is no comprehensive theory of optimal estimators in the case of infinite number of estimators. One of the classical optimality results is that of Shibata (1973), which shows under some mild regularity conditions, the AIC criterion (Akaike, 1973) chooses the best estimator among sieve estimators.

In this paper, we show the equivalence between a class of Bayesian estimators and the AIC selection of sieve estimators. We make very strong assumptions on the class of prior distributions. Suppose the parameters are drawn from

*We would like to thank Whitney Newey, Ivana Komunjer, Yuping Chen, Li Zhang for helpful discussions. An earlier version of the abstract was published in the Proceedings of the Spring 2016 Info-metrics Conference

such a prior distribution, the best estimator is then simply the estimator that minimizes posterior risks given the data. Therefore when such priors are appropriate, through the equivalence results, the AIC estimator is asymptotically optimal among all other estimators. Many papers in the literature have justified certain information criterion based on a Bayesian rationale. Such as the Bayesian Information Criterion (Schwarz, 1978) and the Posterior information criterion (Phillips and Ploberger, 1994). Our paper differs from these literature because our prior distribution is not dismissible asymptotically.

Consider the linear model of the following form

$$y_{(t)} = x_{(t)}\prime\beta + u_{(t)},^{1}$$

where the data is generated in the following way. There is a constant $\lambda \in (0,1)$ such that $\beta_{(i)} \sim \mathcal{N}(0, \lambda^i)$ independent over i, and $\beta_{(i)}$ does not vary with t. The random error term $u_{(t)}$ is iid $\mathcal{N}(0,1)$. We assume that there are a number of regressors $x_{(t)} = \{x_{(t,n)}\}_{n=1,\ldots}$ and they are taken as given. Let the X be an $T \times n$ matrix with orthogonal columns, and each column has Euclidean norm $\sqrt{T}$. The number of regressors n is diverging in T that $n = O(\sqrt{T})$. We call the dth model the one that include all the first d regressors X_d. Our main result shows that asymptotically, the AIC optimal model choose the $-\ln T/\ln\lambda$-th model. If n grows slower than $-\ln T/\ln\lambda$, our result implies the largest model is the best model.

The assumptions on the design matrix can be naturally satisfied in several applications. For example, in estimating non-linear function using higher order polynomials, one can apply the Gram-Schmidt process to X and obtain a set of orthogonal regressors of polynomials in increasing degree. As an other example, in factor models, usually the factors are estimated as principal components, which are orthogonal regressors. The principal components are also ordered naturally according to their variances.

The family of prior distributions is the following data generating process.

[1] We use subscript parathesis $X_{(t)}$ to indicate the tth item in the vector. If we use a subscript X_t without parenthesis, it means the sub-vector consisting of the 1st to tth item.

2 AIC and the OLS regression problem

We use the simple OLS estimator, when we select the first d regressors, the estimator is

$$(X_d\prime X_d)^{-1}X_d\prime Y.$$

and the AIC for d regressors is defined to be $\frac{T}{2}\ln(\hat{\sigma}^2(d)) + d$ where $\hat{\sigma}^2(d)$ is simply the MLE estimator of the error when d regressors are included into the regression. Hence

$$\begin{aligned}\hat{\sigma}^2(d) =& \frac{1}{T}(Y - X_d(X_d\prime X_d)^{-1}X_d\prime Y)\prime(Y - X_d(X_d\prime X_d)^{-1}X_d\prime Y)\\ =& \frac{1}{T}\left((I - Proj_d)(X_n\beta_n + u)\right)\prime\left((I - Proj_d)(X_n\beta_n + u)\right)\\ =& \frac{1}{T}(\beta_n\prime X_n\prime Proj_d^{\perp}X_n\beta_n + u\prime Proj_d^{\perp}u - 2u\prime Proj_d^{\perp}X_n\beta_n)\end{aligned}$$

Where $Proj_d := X_d(X_d\prime X_d)^{-1}X_d$ and $Proj_d^{\perp} := I - Proj_d$. It follows from the orthogonality of X that

$$X_n\prime Proj_d^{\perp}X_n = T\begin{bmatrix}\mathbf{0}_d & * \\ * & \mathbf{1}_{n-d}\end{bmatrix}$$

where the $\mathbf{0}_d$ represents a $d \times d$ zero matrix and $\mathbf{1}_{n-d}$ is an $(n-d)\times(n-d)$ identity matrix. Hence

$$\hat{\sigma}^2(d) = \sum_{i=d+1}^{n}\beta_{(i)}^2 + \frac{1}{T}\sum_{i=d+1}^{T}u_{(i)}^2 - \frac{2}{T}u\prime Proj_d^{\perp}X_n\beta_n.^2$$

We observe that

- since $\beta_{(i)}$ is normal of variance λ^i, $\beta_{(i)}^2$ is a χ_1^2 distribution scaled by a factor λ^i, therefore it has mean λ^i and variance $2\lambda^{2i}$.
- Moreover, each $u_{(i)}^2$ is a χ_1^2 random variable hand has expectation 1 variance 2.

[2] Since u is a standard normal vector, we can specify the coordinate in whichever way we want, hence the subscript (i) picks the dimensions that is exactly in the basis of $Proj_d^{\perp}$ and should not be confused with the subscript of the β. Hence for example, $au\prime Proj_d^{\perp}u - u\prime Proj_{d+C}^{\perp}u$ is the sum of $n-d$ independent $\chi^2(1)$ random variables scaled by $(a-1)$ plus another C independent $\chi^2(1)$ random variables.

- Denote $\frac{2}{T}u\prime Proj_d^{\perp}X_n\beta_n$ by $\epsilon(d)$. It has expectation 0, whether one takes β_n fixed or not. Hence its variance is simply

$$\frac{4}{T^2}\boldsymbol{E}\left[\beta_n\prime X_n\prime Proj_d^{\perp}uu\prime Proj_d^{\perp}X_n\beta_n\right].$$

When we take β_n as given and take expectation over u, we get the conditional variance given β_n. When we take expectation over β and u we get the unconditional variance. They are respectively

$$\frac{4}{T}\sum_{i=d+1}^{n}\beta_{(i)}^2 \text{ and } \frac{4}{T}\sum_{i=d+1}^{n}\lambda^i = \frac{4}{T}\frac{1-\lambda^{n-d}}{1-\lambda}\lambda^{d+1}.$$

2.1 Comparing asymptotic $AIC(d)$ and $AIC(d+C)$ when $d := -\ln T/\ln\lambda$

We want to give a bound on the probability that for a fixed large C, the probability that AIC is minized at $d+C$ as $T\to\infty$. Such probability is bounded above by $\lim_{T\to\infty}\Pr(AIC(d)\geq AIC(d+C))$. It is easy to see that $AIC(d)\geq AIC(d+C)$ for some constant C if and only if

$$\hat{\sigma}^2(d)\geq e^{2C/T}\hat{\sigma}^2(d+C),$$

In other words

$$\sum_{i=d+1}^{n}\beta_{(i)}^2+\frac{1}{T}\sum_{i=d+1}^{T}u_{(i)}^2-\epsilon(d)\geq e^{2C/T}\left(\sum_{i=d+C+1}^{n}\beta_{(i)}^2+\frac{1}{T}\sum_{i=d+C+1}^{T}u_{(i)}^2-\epsilon(d+C)\right).$$

We rewrite it with a normalization T/C on both hand sides and get

$$\begin{aligned}&\frac{T}{C}\sum_{i=d+1}^{d+C}\beta_{(i)}^2\\ \geq&\frac{T}{C}(e^{2C/T}-1)\left(\sum_{i=d+C+1}^{n}\beta_{(i)}^2+\frac{1}{T}\sum_{i=d+C+1}^{T}u_{(i)}^2\right)-\frac{1}{C}\sum_{i=d+1}^{d+C}u_{(i)}^2\\ &-\frac{T}{C}e^{2C/T}\epsilon(d+C)+\frac{T}{C}\epsilon(d)\end{aligned}$$

The expectation of LHS is

$$\frac{T}{C}\sum_{i=d+1}^{d+C}\lambda^i = \frac{T}{C}\lambda^{d+1}\frac{1-\lambda^C}{1-\lambda} = \frac{\lambda}{C}\frac{1-\lambda^C}{1-\lambda}$$

when we plug in $d := -\ln T/\ln\lambda$. The variance of LHS is

$$\frac{T^2}{C^2}\sum_{i=d+1}^{d+C} 2\lambda^{2i} = 2\frac{T^2}{C^2}\lambda^{2d+2}\frac{1-\lambda^{2C}}{1-\lambda^2} = 2\left(\frac{\lambda}{C}\right)^2\frac{1-\lambda^{2C}}{1-\lambda^2}$$

when we plug in $d := -\ln T/\ln\lambda$. Hence the LHS is of order $O(\frac{1}{C})$.

On the other hand, the RHS has expectation

$$\begin{aligned}
&\frac{e^{2C/T}-1}{C}\left(T\sum_{i=d+C+1}^{n}\lambda^i + (T-d-C)\right) - 1\\
=&\frac{e^{2C/T}-1}{C}\left(T\lambda^{d+C+1}\frac{1-\lambda^{n-d-C}}{1-\lambda} + T - d - C\right) - 1\\
\to&\frac{2}{T}\left(\lambda^{C+1}\frac{1-\lambda^{n-d-C}}{1-\lambda} + T - d - C\right) - 1 \to 1
\end{aligned}$$

by first plug in $d := -\ln T/\ln\lambda$ and take $T\to\infty$.

The variance of RHS is bounded by the following term multiplied by 2 (to take care of the covariances)

$$\begin{aligned}
&\left(\frac{e^{2C/T}-1}{C}\right)^2\left(T^2 2\sum_{i=d+C+1}^{n}\lambda^{2i} + \sum_{i=d+C+1}^{T} 2\right) + \sum_{i=d+1}^{d+C}\frac{2}{C}\\
&+\left(\frac{T}{C}\right)^2 e^{4C/T}\boldsymbol{V}ar[\epsilon(d+C)] + \left(\frac{T}{C}\right)^2\boldsymbol{V}ar[\epsilon(d)]\\
\to&\frac{4}{T^2}\left(2\lambda^{2C+2}\frac{1-\lambda^{2(n-d-C)}}{1-\lambda} + 2(T-d-C)\right) + \frac{2}{C}\\
&+\left(\frac{T}{C}\right)^2 e^{4C/T}\frac{4}{T}\frac{1-\lambda^{n-d-C}}{1-\lambda}\lambda^{d+C+1} + \left(\frac{T}{C}\right)^2\frac{4}{T}\frac{1-\lambda^{n-d}}{1-\lambda}\lambda^{d+1}\\
\to&\frac{2}{C} + \frac{4}{C^2}\frac{1-\lambda^{n-d-C}}{1-\lambda}\lambda^{C+1} + \frac{4}{C^2}\frac{1-\lambda^{n-d}}{1-\lambda}\lambda
\end{aligned}$$

by first plug in $d := -\ln T/\ln\lambda$ and take $T\to\infty$.

It can be seen that LHS is of order $O(\frac{1}{C})$. The first two terms in the RHS equals $2 - \frac{\chi^2(C)}{C}$ which is a Chi-square C degree of freedom variable multiplied by a factor of $-1/C$ and then translated two units to the right. The last two

terms is of order $O(\frac{1}{C})$. Moreover, it is clear that the above limits are uniform for all $C \in [0, n]$ as long as $n/T \to 0$.

Therefore, for any large enough $C \leq n$, the probability that LHS $\geq$ RHS is approximately

$$\begin{aligned}
\Pr(0 \geq 2 - 1/C\chi^2(C)) =& \Pr(\chi^2(C) > 2C) \\
=& \int_{2C}^{\infty} \frac{1}{2^{C/2}\Gamma(C/2)} x^{C/2-1} e^{-x/2} \boldsymbol{d}x \\
=& \frac{\Gamma(C/2, C)}{\Gamma(C/2)} \\
\leq& \frac{\lceil C/2 - 1 \rceil !}{\lfloor C/2 - 1 \rfloor !} e^{-C} \sum_{k=0}^{\lceil C/2-1 \rceil} \frac{C^k}{k!} \\
\leq& 2e^{-C} \sum_{k=0}^{\lceil C/2-1 \rceil} \frac{C^k}{k!} \\
\leq& 2e^{-C} \frac{C/2}{\sqrt{\pi C}} \left(\frac{e}{2}\right)^{C/2} \\
=& \sqrt{\frac{C}{\pi}} (2e)^{-C/2},
\end{aligned}$$

where we used properties of the incomplete Gamma function[3] and Stirling's approximation. Hence we conclude that as $T \to \infty$ the probability that $d + C$ minimizes AIC is bounded by $\sqrt{\frac{C}{\pi}}(2e)^{-C/2}$ for all large C.

2.2 Comparing asymptotic $AIC(d)$ and $AIC(d - C)$ when $d := -\ln T / \ln \lambda$

On the other hand, we want to give a bound on the probability that for a fixed large C, the probability that AIC is minized at $d - C$ as $T \to \infty$. Such probability is bounded above by $\lim_{T\to\infty} \Pr(AIC(d) \geq AIC(d - C))$.

$AIC(d) \geq AIC(d - C)$ if and only if $e^{2C/T}\hat{\sigma}^2(d) \geq \hat{\sigma}^2(d - C)$. Apply the scaling $\lambda^{C-1}T$ to both hand sides, we rewrite the inequality in the following

[3] Weisstein, Eric W., "Incomplete Gamma Function", *MathWorld*. (equation 2)

way:

$$\begin{aligned}
&\lambda^{C-1}\left[T(e^{2C/T}-1)\left(\sum_{i=d+1}^{n}\beta_{(i)}^2+\frac{1}{T}\sum_{i=d+1}^{T}u_{(i)}^2\right)-\sum_{i=d-C+1}^{d}u_{(i)}^2\right]\\
&-\lambda^{C-1}Te^{2C/T}\epsilon(d)+\lambda^{C-1}T\epsilon(d-C)\\
\geq&\lambda^{C-1}T\sum_{i=d-C+1}^{d}\beta_{(i)}^2
\end{aligned}$$

Expectation of LHS is

$$\begin{aligned}
&\lambda^{C-1}\left[T(e^{2C/T}-1)\left(\sum_{i=d+1}^{n}\lambda^i+\frac{1}{T}(T-d)\right)-C\right]\\
=&\lambda^{C-1}\left[2C\left(\lambda^{d+1}\frac{1-\lambda^{n-d}}{1-\lambda}+\frac{T-d}{T}\right)-C\right]\\
\to&\lambda^{C-1}C
\end{aligned}$$

when we take $d=-\ln T/\ln\lambda$ and then take $T\to\infty$. The variance of LHS is bounded by two times the following:

$$\begin{aligned}
&\lambda^{2C-2}\left[(T(e^{2C/T}-1))^2\left(\sum_{i=d+1}^{n}2\lambda^{2i}+\frac{1}{T^2}2(T-d)\right)+2C\right]\\
&+\lambda^{2C-2}\left[(T(e^{2C/T}))^2\frac{4}{T}\lambda^{d+1}\frac{1-\lambda^{n-d}}{1-\lambda}+T^2\frac{4}{T}\lambda^{d-C+1}\frac{1-\lambda^{n-d+C}}{1-\lambda}\right]\\
\to&\lambda^{2C-2}4C^2\left(2\frac{\lambda^2}{T^2}\frac{1-\lambda^{2n-2d}}{1-\lambda}+\frac{2(T-d)}{T^2}\right)+4\lambda^{2C-1}\frac{1-\lambda^{2n-2d}}{1-\lambda}+4\lambda^{C-1}\frac{1-\lambda^{n-d+C}}{1-\lambda}\\
\to&4\lambda^{2C-1}\frac{1}{1-\lambda}+4\lambda^{C-1}\frac{1}{1-\lambda}
\end{aligned}$$

when we take $d=-\ln T/\ln\lambda$ and then take $T\to\infty$. Hence the LHS is of order $O(\lambda^{C/2})$

Now consider RHS, it can be seen that

$$RHS=\lambda^{C-1}T\sum_{i=d-C+1}^{d}\beta_{(i)}^2>\lambda^{C-1}T\beta_{(d-C+1)}^2\sim\chi^2(1)$$

after taking $d=-\ln T/\ln\lambda$.

Hence it can be seen that for any fixed large C, the probability that LHS≥RHS

is bounded above by the probability that LHS$\geq \chi^2(1)$. This is approximately

$$\Pr(\lambda^{C/2} > \chi^2(1)) = \int_0^{\lambda^{C/2}} \frac{1}{\sqrt{2}\Gamma(1/2)} x^{-1/2} e^{-x/2} \boldsymbol{d}x \leq \int_0^{\lambda^{C/2}} x^{-1/2} \boldsymbol{d}x = \lambda^{C/4}$$

for all large C. Hence we conclude that as $T \to \infty$ the probability that $d - C$ minimizes AIC is bounded by $\lambda^{C/4}$ for all large C.

3 The Bayesian problem

Our Bayesian problem can be formulated in a slightly more general case of the infinite dimensional space. However since there is no density function available in infinite dimensional situations, finding the posterior measure in infinite dimensional space requires Radon-Nikodym derivative. A general treatment of the subject can be found in Stuart (2010).

The standard result gives that the posterior mean vector of β is

$$\hat{\beta} := \Sigma^{1/2} \left(\Sigma^{1/2} X\prime X \Sigma^{1/2} + I \right)^{-1} \Sigma^{1/2} X\prime (X\beta + u)$$

Let $Q := \Sigma^{1/2} \left(\Sigma^{1/2} X\prime X \Sigma^{1/2} + I \right)^{-1} \Sigma^{1/2}$, then $\hat{\beta} = QX\prime(X\beta + u)$.
and the posterior variance covariance

$$\boldsymbol{E}_{posterior}[(\beta - \hat{\beta})\prime(\beta - \hat{\beta})] = \Sigma^{1/2}(\Sigma^{1/2} X\prime X \Sigma^{1/2} + I)^{-1} \Sigma^{1/2}.$$

Since $X\prime X = TI$, $\hat{\beta}_{(i)} = \frac{T\lambda^i}{T\lambda^i+1}\beta_{(i)} + \frac{\lambda^i}{T\lambda^i+1} X_{(i)}\prime u$. For any given i, the second term goes to 0 as $T \to \infty$ since $X_{(i)}$ and u are not dependent. On the other hand $\frac{T\lambda^i}{T\lambda^i+1}$ is approximately 1 for large T and small i, and approximately 0 for large i. It can be easily check that the first term is approximately $\beta_{(i)}$ for the first $-\ln T/\ln\lambda - C$ coordinates, and they are approximately 0 for $i \geq -\ln T/\ln\lambda + C$ for some C depends only on λ. This shows that the AIC from the previous section would select approximately the same number of regressors asymptotically.

4 Asymptotic equivalence

4.1 l^2 equivalence

Let $\tilde{\beta}$ be the AIC estimate and $\hat{\beta}$ be the bayesian estimate. Then we have the following two theorems.

Theorem 1 *Under our assumptions, we have*

$$\boldsymbol{E}_u||\tilde{\beta}-\hat{\beta}||_2^2 = o\left(\boldsymbol{E}_u||\beta-\hat{\beta}||_2^2\right)$$

e.g. the difference between the estimators is a magnitude smaller than the estimation error.

For some d^* is optimally chosen by AIC between $1, 2, \ldots, n$, it is readily seen that

$$\tilde{\beta} = (X_{d^*}\prime X_{d^*})^{-1}X_{d^*}\prime(X\beta+u) \text{ and } \hat{\beta} = QX\prime(X\beta+u)$$

Notice that the two estimates is of different dimensions, $\tilde{\beta}$ has d^* non-trivial dimensions and we would fill the remaining dimensions with 0. Notice that by definition, Q is a diagonal matrix. Hence we can write Q_{d^*} be the top left $d^* \times d^*$ square submatrix and Q_{d^*+} be the $(n-d^*)\times(n-d^*)$ be the submatrix at the bottom right cornor. Hence

$$\hat{\beta}_{d^*} = Q_{d^*}X\prime_{d^*}(X\beta+u) \text{ and } \hat{\beta}_{d^*+} = Q_{d^*+}X\prime_{d^*+}(X\beta+u).$$

And therefore, $||\tilde{\beta}-\hat{\beta}||_2^2 = ||\tilde{\beta}-\hat{\beta}_{d^*}||_2^2 + ||\hat{\beta}_{d^*+}||_2^2$.

Proof.

We can expand the expression and get

$$\begin{aligned}
||\tilde{\beta}-\hat{\beta}||_2^2 =& ||\tilde{\beta}-\hat{\beta}_{d^*}||_2^2 + ||\hat{\beta}_{d^*+}||_2^2 \\
=& ||(X_{d^*}\prime X_{d^*})^{-1}X_{d^*}\prime(X\beta+u) - Q_{d^*}X\prime_{d^*}(X\beta+u)||_2^2 + ||Q_{d^*+}X\prime_{d^*+}(X\beta+u)||_2^2 \\
\leq& ||((X_{d^*}\prime X_{d^*})^{-1}-Q_{d^*})X_{d^*}\prime X\beta||_2^2 + ||Q_{d^*+}X\prime_{d^*+}X\beta||_2^2 \\
&+ ||((X_{d^*}\prime X_{d^*})^{-1}-Q_{d^*})X_{d^*}\prime u||_2^2 + ||Q_{d^*+}X\prime_{d^*+}u||_2^2 \\
=& \sum_{i=1}^{d^*}\left(\frac{\beta_{(i)}}{1+T\lambda^i}\right)^2 + \sum_{i=d^*+1}^{n}\left(\frac{T\lambda^i\beta_{(i)}}{1+T\lambda^i}\right)^2 \\
&+ ||((X_{d^*}\prime X_{d^*})^{-1}-Q_{d^*})X_{d^*}\prime u||_2^2 + ||Q_{d^*+}X\prime_{d^*+}u||_2^2
\end{aligned}$$

The third and the fourth term can be separated into the norm contributed from the first d^* terms in u and the remaining terms. i.e.

$$\begin{aligned}
&||((X_{d^*}\prime X_{d^*})^{-1}-Q_{d^*})X_{d^*}\prime u||_2^2 + ||Q_{d^*+}X\prime_{d^*+}u||_2^2 \\
=& u\prime X_{d^*}\left((X_{d^*}\prime X_{d^*})^{-1}-Q_{d^*}\right)^2 X_{d^*}\prime u + u\prime X_{d^*+}Q_{d^*+}^2 X_{d^*+}\prime u
\end{aligned}$$

Take any $C = \ln\ln d$ for $d := -\ln T/\ln\lambda$, by our previous analysis, we have that $d - C < d^* < d + C$ as $T \to \infty$. Hence for T large enough, we can bound the above expression by

$$
\begin{aligned}
&||((X_{d^*}\prime X_{d^*})^{-1} - Q_{d^*})X_{d^*}\prime u||_2^2 + ||Q_{d^*+}X\prime_{d^*+}u||_2^2 \\
<&u\prime X_{d+C}\left((X_{d+C}\prime X_{d+C})^{-1} - Q_{d+C}\right)^2 X_{d+C}\prime u + u\prime X_{(d-C)+}Q_{(d-C)+}^2 X_{(d-C)+}\prime u
\end{aligned}
$$

Taking expectation over u the above expression can be expressed in terms of trace, i.e. from $\mathbb{E}[uu\prime] = I$ we have

$$
\begin{aligned}
=&tr\left(\left((X_{d+C}\prime X_{d+C})^{-1} - Q_{d+C}\right)^2 X_{d+C}\prime\mathbb{E}[uu\prime]X_{d+C}\right) + tr\left(Q_{(d-C)+}^2 X_{(d-C)+}\prime\mathbb{E}[uu\prime]X_{(d-C)+}\right) \\
=&\sum_{i=1}^{d+C}\left(\frac{1}{T} - \frac{\lambda^i}{1+\lambda^i T}\right)^2 T + \sum_{i=d-C+1}^{T}\frac{T\lambda^{2i}}{(1+\lambda^i T)^2} \\
\le&\sum_{i=1}^{d+C}\frac{1}{T(T\lambda^i+1)^2} + \sum_{i=d-C+1}^{T}\lambda^{2i}T \\
\le&\sum_{i=1}^{d+C}\frac{\lambda^{-i}}{T^2} + T\lambda^{2d-2C+2}\frac{1-\lambda^{T-d+C}}{1-\lambda} \\
=&\lambda^{-1}\frac{1}{T^2}\frac{\lambda^{-d}\lambda^{-C}-1}{\lambda^{-1}-1} + T\lambda^{2d}\lambda^{-2C}\lambda^2\frac{1-\lambda^{T-d+C}}{1-\lambda}
\end{aligned}
$$

Since $\lambda^d = 1/T$ and $\lambda^C = (\ln d)^{\ln\lambda}$, the above expression becomes

$$
\frac{\lambda^{-1}}{\lambda^{-1}-1}\frac{(\ln d)^{-\ln\lambda}}{T} - \frac{\lambda^{-1}}{\lambda^{-1}-1}\frac{1}{T^2} + \frac{(\ln d)^{-2\ln\lambda}}{T}\frac{1-\lambda^{T-d+C}}{1-\lambda} = O\left(\frac{(\ln d)^{-2\ln\lambda}}{T}\right)
$$

Therefore,

$$
\begin{aligned}
\mathbb{E}_u||\tilde{\beta} - \hat{\beta}||_2^2 =& \sum_{i=1}^{d^*}\left(\frac{\beta_{(i)}}{1+T\lambda^i}\right)^2 + \sum_{i=d^*+1}^{n}\left(\frac{T\lambda^i\beta_{(i)}}{1+T\lambda^i}\right)^2 \\
&+ \mathbb{E}_u||((X_{d^*}\prime X_{d^*})^{-1} - Q_{d^*})X_{d^*}\prime u||_2^2 + \mathbb{E}_u||Q_{d^*+}X\prime_{d^*+}u||_2^2 \\
\le& \sum_{i=1}^{d+C}\left(\frac{\beta_{(i)}}{1+T\lambda^i}\right)^2 + \sum_{i=d-C}^{n}\left(\frac{T\lambda^i\beta_{(i)}}{1+T\lambda^i}\right)^2 + O\left(\frac{(\ln d)^{-2\ln\lambda}}{T}\right) \\
\le& \sum_{i=1}^{d+C}\left(\frac{\beta_{(i)}}{T\lambda^i}\right)^2 + \sum_{i=d-C+1}^{n}\left(T\lambda^i\beta_{(i)}\right)^2 + O\left(\frac{(\ln d)^{-2\ln\lambda}}{T}\right)
\end{aligned}
$$

for large enough T. The first term has mean $\frac{1}{T^2}\lambda^{-1}\frac{\lambda^{-d-C+1}-1}{\lambda^{-1}-1} = O\big(\frac{(\ln d)^{-\ln\lambda}}{T}\big)$

and variance $\frac{2}{T^4}\lambda^{-2}\frac{\lambda^{-2d-2C}-1}{\lambda^{-2}-1} = O\big(\frac{(\ln d)^{-2\ln\lambda}}{T^2}\big)$, hence the first term is of order $O\big(\frac{(\ln d)^{-\ln\lambda}}{T}\big)$. The second term has mean $T^2\lambda^{3d}\lambda^{-3C}\lambda^3\frac{1-\lambda^{3(n-d+C)}}{1-\lambda^3} = O\big(\frac{(\ln d)^{-3\ln\lambda}}{T}\big)$ and variance $2T^4\lambda^{6d}\lambda^{-6C}\lambda^6\frac{1-\lambda^{6(n-d+C)}}{1-\lambda^6} = O\big(\frac{(\ln d)^{-6\ln\lambda}}{T^2}\big)$, hence the second term is of order $O\big(\frac{(\ln d)^{-3\ln\lambda}}{T}\big)$. Therefore we conclude that

$$\mathbb{E}_u||\tilde{\beta}-\hat{\beta}||_2^2 \leq O\bigg(\frac{(\ln d)^{-\ln\lambda}}{T}\bigg)+O\bigg(\frac{(\ln d)^{-3\ln\lambda}}{T}\bigg)+O\bigg(\frac{(\ln d)^{-2\ln\lambda}}{T}\bigg) = O\bigg(\frac{(\ln d)^{-3\ln\lambda}}{T}\bigg).$$

On the other hand, we can get a Chi-square lower bound by comparing the first d^* terms in the true parameter β and AIC estimate $\tilde{\beta}$.

$$\begin{aligned}
||\tilde{\beta}-\beta||_2^2 \geq& ||(X_{d^*}\prime X_{d^*})^{-1}X_{d^*}\prime(X\beta+u)-\beta_{d^*}||_2^2\\
=& u\prime X_{d^*}(X_{d^*}\prime X_{d^*})^{-1}(X_{d^*}\prime X_{d^*})^{-1}X_{d^*}\prime u\\
\geq& u\prime X_{d-C}(X_{d-C}\prime X_{d-C})^{-1}(X_{d-C}\prime X_{d-C})^{-1}X_{d-C}\prime u,
\end{aligned}$$

hence the lower bound follows some scaled Chi-square distribution of $d-C$ degree of freedom. Taking expectation over u we have

$$\begin{aligned}
\mathbb{E}_u||\tilde{\beta}-\beta||_2^2 \geq& \mathbb{E}_u[u\prime X_{d-C}(X_{d-C}\prime X_{d-C})^{-1}(X_{d-C}\prime X_{d-C})^{-1}X_{d-C}\prime u]\\
=& tr((X_{d-C}\prime X_{d-C})^{-1}X_{d-C}\prime\mathbb{E}_u[uu\prime]X_{d-C}(X_{d-C}\prime X_{d-C})^{-1})\\
=& \frac{d-C}{T}
\end{aligned}$$

Therefore

$$\mathbb{E}_u||\tilde{\beta}-\hat{\beta}||_2^2 = o(\mathbb{E}_u||\tilde{\beta}-\beta||_2^2)$$

as $T\to\infty$. We have therefore shown that $\hat{\beta}$ and $\tilde{\beta}$ are asymptotically equivalent under l^2 norm. ■

4.2 Equivalence under linear projections

Not only the global distance between the two estimators is smaller than the estimation error, this is also true for many of the linear projections of the estimator. For all vectors B, $B\prime(\beta-\hat{\beta}))$ is normal. We show that for all vectors B satisfying some restrictions $B\prime(\tilde{\beta}-\hat{\beta})$ is of smaller order than the standard deviation of $B\prime(\beta-\hat{\beta}))$. For this to hold, we need to require that the components of $B=(b_{(i)})_{i=1}^{\infty}$ are all of the same order of magnitude.

Definition 2 *We say the partial sum of a sequence* $S_n := \sum_{i=1}^n b_i^2$ *is of* slow growth *if for any constant* C

$$\lim_{n\to\infty} \frac{S_{n+C}}{S_n} = 1.^4$$

Theorem 3 *If* $B = (b_{(i)})_{i=1}^{\infty}$ *whose squared partial sum is of slow growth, then* $B\prime(\tilde{\beta} - \hat{\beta})$ *is of smaller order than the standard deviation of* $B\prime(\beta - \hat{\beta})$.

Recall that we have

$$\tilde{\beta}-\hat{\beta} = \begin{bmatrix} \cdots \\ \frac{1}{1+\lambda^i T}\beta_{(i)} + \frac{1}{1+\lambda^i T}\frac{1}{T}(X\prime u)_{(i)} \\ \cdots \\ -\frac{T\lambda^j}{1+\lambda^j T}\beta_{(j)} - \frac{\lambda^j}{1+\lambda^j T}(X\prime u)_{(j)} \\ \cdots \end{bmatrix} \text{ and } \beta-\hat{\beta} = \begin{bmatrix} \cdots \\ \frac{1}{1+\lambda^k T}\beta_{(k)} - \frac{\lambda^k}{1+\lambda^k T}(X\prime u)_{(i)} \\ \cdots \end{bmatrix}$$

where $1 \le i \le d^* < j \le n$ and $1 \le k \le n$.

When $B\prime = (b_1, b_2, \ldots)$ is just a row vector, consider $B\prime(\beta - \hat{\beta})$, it follows a mean zero normal distribution with variance

$$\begin{aligned} \sum_{i=1}^{n}\left(\left(\frac{b_i}{1+\lambda^i T}\right)^2 \lambda^i + \left(\frac{\lambda^i b_i}{1+\lambda^i T}\right)^2 T\right) \ge & \sum_{i=1}^{d}\left(\frac{\lambda^i b_i}{1+\lambda^i T}\right)^2 T \\ \ge & O\left(\frac{1}{T}\right)\sum_{i=1}^{d} b_i^2 \end{aligned}$$

hence $B\prime(\beta - \hat{\beta})$ is of order greater or equal to $O(\frac{1}{\sqrt{T}})\sqrt{\sum_{i=1}^d b_i^2}$ for $d := -\ln T/\ln\lambda$.

To prove the theorem, we will show that $B\prime(\tilde{\beta} - \hat{\beta}) = o\left(\sqrt{\frac{\sum_{i=1}^d b_i^2}{T}}\right)$ under the assumption that $\sum_{i=1}^n b_i^2$ is of slow growth. Before we present the proof, we first prepare the following lemma.

Lemma 4 *Suppose* $S_d := \sum_i^d b_i^2$ *is of slow growth, then for any* $\lambda \in (0,1)$*, and any constant* C*, and* n *such that* $\lim_{d\to\infty} d/n \to 0$*, the following limits hold as* $d \to \infty$*:*

$$\frac{\sum_{i=1}^{d+C} b_i^2 \lambda^{d+C-i}}{\sum_{i=1}^d b_i^2} \to 0; \text{ and } \frac{\sum_{j=1}^{n-d+C} b_{j+d-C}^2 \lambda^j}{\sum_{i=1}^d b_i^2} \to 0.$$

[4] Kapoor and Nautiyal (1981) studied classes of functions of various speeds of growth, our definition of slow growth here satisfies the more general hypothesis (H, ii) in their paper, but not necessarily the more restrictive ones $(H, iii) - (H, v)$.

Proof. To establish the first limit, observe that for any $\lambda \in (0,1)$

$$\lambda \times \lambda^{k-i} = \lambda^{k+1} + (1-\lambda) \sum_{j=k-i+1}^{k} \lambda^j.$$

Hence

$$\begin{aligned}\frac{\sum_{i=1}^{d+C} b_i^2 \lambda^{d+C-i}}{\sum_{i=1}^{d} b_i^2} =& \frac{\sum_{i=1}^{d+C} b_i^2 \lambda^{d+C+1}}{\lambda S_d} + \frac{(1-\lambda)\sum_{i=1}^{d+C} b_i^2 \sum_{j=d+C-i+1}^{d+C} \lambda^j}{\lambda S_d} \\ =& \frac{S_{d+C}\lambda^{d+C+1}}{\lambda S_d} + \frac{(1-\lambda)\sum_{j=1}^{d+C}\sum_{i=d+C-j+1}^{d+C} b_i^2 \lambda^j}{\lambda S_d}\end{aligned}$$

The first term goes to 0 as $d \to \infty$. The second term can be decomposed into two parts

$$\begin{aligned}&\frac{\sum_{j=1}^{d+C}\sum_{i=d+C-j+1}^{d+C} b_i^2 \lambda^j}{S_d} \\ =& \frac{\sum_{j=1}^{K}\sum_{i=d+C-j+1}^{d+C} b_i^2 \lambda^j}{S_d} + \frac{\sum_{j=K+1}^{d+C}\sum_{i=d+C-j+1}^{d+C} b_i^2 \lambda^j}{S_d} \\ \leq& \frac{\sum_{j=1}^{K} \lambda^j (S_{d+C} - S_{d+C-K})}{S_d} + \frac{\sum_{j=K+1}^{d+C} \lambda^j \sum_{i=d+C-j+1}^{d+C} b_i^2}{S_{d+C}} \frac{S_{d+C}}{S_d} \\ \leq& \frac{\sum_{j=1}^{K} \lambda^j (S_{d+C} - S_{d+C-K})}{S_d} + \left(\sum_{j=K+1}^{d+C} \lambda^j\right) \frac{S_{d+C}}{S_d}.\end{aligned}$$

For any fixed K, the first term goes to 0, the second term can be arbitrarily small by choosing K large enough and that $\frac{S_{d+C}}{S_d} \to 1$. This establishes the first limit.

To obtain the second limit, observe the following identity

$$\lambda^j = \lambda^{k+1} + (1-\lambda)\sum_{i=j}^{k} \lambda^i.$$

Therefore

$$
\begin{aligned}
&\frac{\sum_{j=1}^{n-d+C} b_{j+d-C}^2 \lambda^j}{\sum_{i=1}^{d} b_i^2} \\
=&\frac{\sum_{j=1}^{n-d+C} b_{j+d-C}^2 \lambda^{n-d+C+1}}{S_d} + \frac{(1-\lambda)\sum_{j=1}^{n-d+C} b_{j+d-C}^2 \sum_{i=j}^{n-d+C} \lambda^i}{S_d} \\
\leq&\lambda^{n-d+C+1}\frac{S_n}{S_d} + \frac{(1-\lambda)\sum_{i=1}^{n-d+C} \sum_{j=1}^{i} b_{j+d-C}^2 \lambda^i}{S_d} \\
\leq&\lambda^{n-d+C+1}\frac{S_n}{S_d} + \frac{(1-\lambda)\sum_{i=1}^{K} \sum_{j=1}^{i} b_{j+d-C}^2 \lambda^i}{S_d} + \frac{(1-\lambda)\sum_{i=K+1}^{n-d+C} \sum_{j=1}^{i} b_{j+d-C}^2 \lambda^i}{S_d} \\
\leq&\lambda^{n-d+C+1}\frac{S_n}{S_d} + \frac{(1-\lambda)\sum_{i=1}^{K} \lambda^i (S_{K+d-C} - S_{d-C})}{S_d} + \frac{(1-\lambda)\sum_{i=K+1}^{n-d+C} \lambda^i S_{i+d-C}}{S_d}
\end{aligned}
$$

For any fixed K the second term goes to 0 as $d \to \infty$ due to the slow growth assumption. Now observe that

$$S_k = S_d \prod_{i=1}^{k-d} \frac{S_{d+i}}{S_{d+i-1}},$$

by the slow growth assumption, there exists $\underline{d}$ such that if $d > \underline{d}$, $\frac{S_d}{S_{d-1}} \leq \lambda^{-1/2}$. Let $k > d > \underline{d}$, then

$$S_k \leq S_d \lambda^{-(k-d)/2}.$$

Hence by choosing any $K > C$, the first and the third term is bounded by

$$
\begin{aligned}
&\lambda^{n-d+C+1}\frac{S_n}{S_d} + \frac{\sum_{i=K+1}^{n-d+C} \lambda^i S_{i+d-C}}{S_d} \\
\leq&\lambda^{n-d+C+1}\lambda^{-(n-d)/2} + \sum_{i=K+1}^{n-d+C} \lambda^i \lambda^{-(i-C)/2}
\end{aligned}
$$

where the first term goes to 0 as $n, d \to \infty$ since $d/n \to 0$. The second term can be arbitrarily small by choosing K large enough. This completes the proof. ■

Now we proceed to the proof of Theorem 3.

Proof. Observe that $B\prime(\tilde{\beta} - \hat{\beta})$ can be separated into four terms. We will show

that each of the four terms is of order $o\left(\sqrt{\frac{S_d}{T}}\right)$.

$$|B\prime(\tilde{\beta}-\hat{\beta})| \leq \sum_{i=1}^{d^*} |\frac{b_i}{1+\lambda^i T}\beta_{(i)}| + \sum_{i=1}^{d^*} |\frac{b_i}{1+\lambda^i T}\frac{1}{T}(X\prime u)_{(i)}| + \sum_{j=d^*+1}^{n} |\frac{b_j T\lambda^j}{1+\lambda^j T}\beta_{(j)}| + \sum_{j=d^*+1}^{n} |\frac{b_j \lambda^j}{1+\lambda^j T}(X\prime u)_{(j)}|$$

where the d^* is determined optimally by AIC.

The First Term

Consider the first term, to show that it is of order $o\left(\sqrt{\frac{S_d}{T}}\right)$, we need to show that for any $M > 0$ and for any $\epsilon > 0$ arbitrarily small, there exists $\underline{T}$ such that if $T > \underline{T}$,

$$\Pr\left(\sqrt{\frac{T}{S_d}}\sum_{i=1}^{d^*} |\frac{b_i}{1+\lambda^i T}\beta_{(i)}| > M\right) < \epsilon.$$

From Section 2.1, we know that for any given large C,

$$\Pr\left(\sum_{i=1}^{d^*} |\frac{b_i}{1+\lambda^i T}\beta_{(i)}| > \sum_{i=1}^{d+C} |\frac{b_i}{1+\lambda^i T}\beta_{(i)}|\right) \leq \delta^C$$

for some fixed $\delta \in (0,1)$, $d := -\ln T \ln\lambda$ as $T \to \infty$. For convenience, denote $\sum_{i=1}^{d^*} |\frac{b_i}{1+\lambda^i T}\beta_{(i)}|$ by Σ^* and $\sum_{i=1}^{d+C} |\frac{b_i}{1+\lambda^i T}\beta_{(i)}|$ by Σ^C.

$$\begin{aligned}
&\Pr\left(\sqrt{\frac{T}{S_d}}\Sigma^* > M\right) \\
=&\Pr\left(\{\sqrt{\frac{T}{S_d}}\Sigma^* > M\}\bigcap\{\Sigma^* > \Sigma^C\}\right) + \Pr\left(\{\sqrt{\frac{T}{S_d}}\Sigma^* > M\}\bigcap\{\Sigma^* \leq \Sigma^C\}\right) \\
\leq&\Pr\left(\Sigma^* > \Sigma^C\right) + \Pr\left(\sqrt{\frac{T}{S_d}}\Sigma^C > M\right) \leq \delta^C + \Pr\left(\sqrt{\frac{T}{S_d}}\Sigma^C > M\right)
\end{aligned}$$

Hence it is sufficient to show that $\Pr\left(\sqrt{\frac{T}{S_d}}\Sigma^C > M\right)$ goes to 0 for any C, M.

Since Σ^C is a positive random variable, by markov inequality,

$$\Pr\left(\sqrt{\frac{T}{S_d}}\Sigma^C > M\right) \leq \frac{\sqrt{\frac{T}{S_d}}\mathbb{E}[\Sigma^C]}{M}$$

Since Σ^C is a sum of half-normal random variable. We can calculate their expectations.

$$\begin{aligned}
\mathbb{E}[\Sigma^C] \leq & \sqrt{\frac{2}{\pi}} \sum_{i=1}^{d+C} \frac{|b_i|\lambda^{i/2}}{\lambda^i T} \\
= & \sqrt{\frac{2}{\pi}} \frac{\lambda^{-(d+C)/2}}{T} \sum_{i=1}^{d+C} |b_i|\lambda^{(d+C-i)/2} \\
\leq & \sqrt{\frac{2}{\pi}} \frac{\lambda^{-C/2}}{\sqrt{T}} \sqrt{\frac{\sum_{i=1}^{d+C} b_i^2 \lambda^{(d+C-i)/2}}{\sum_{i=j}^{d+C} \lambda^{(d+C-j)/2}}} \sum_{j=1}^{d+C} \lambda^{(d+C-j)/2} \\
= & \sqrt{\frac{2}{\pi}} \frac{\lambda^{-C/2}}{\sqrt{T}} \sqrt{\sum_{i=1}^{d+C} b_i^2 \lambda^{(d+C-i)/2}} \sqrt{\frac{1}{1-\lambda^{1/2}}}
\end{aligned}$$

where we applied quadratic mean inequality to get the second inequality. Therefore

$$\begin{aligned}
\Pr\left(\sqrt{\frac{T}{S_d}}\Sigma^C > M\right) \leq & \frac{\sqrt{\frac{T}{S_d}}\mathbb{E}[\Sigma^C]}{M} \\
\leq & \sqrt{\frac{2}{\pi}} \sqrt{\frac{1}{1-\lambda^{1/2}}} \frac{\lambda^{-C/2}}{M} \sqrt{\frac{\sum_{i=1}^{d+C} b_i^2 \lambda^{(d+C-i)/2}}{S_d}}
\end{aligned}$$

which goes to 0 by Lemma 1. Therefore, $\Pr\left(\sqrt{\frac{T}{S_d}}\Sigma^* > M\right)$ is arbitrarily small for any M as $T \to \infty$. This shows the first term is of order $o\left(\sqrt{\frac{S_d}{T}}\right)$.

Other terms

For other terms, we can use similar arguments as above, by observing all the following probabilities are exponentially small in C. I.e. by Section 2, for all T

large enough, there exists a fixed $\delta \in (0,1)$ such that

$$\Pr\left(\sum_{i=1}^{d^*} |\frac{b_i}{1+\lambda^i T}\frac{1}{T}(X\prime u)_{(i)}| > \sum_{i=1}^{d+C} |\frac{b_i}{1+\lambda^i T}\frac{1}{T}(X\prime u)_{(i)}|\right) \leq \delta^C;$$

$$\Pr\left(\sum_{j=d^*+1}^{n} |\frac{b_j T\lambda^j}{1+\lambda^j T}\beta_{(j)}| > \sum_{j=d-C+1}^{n} |\frac{b_j T\lambda^j}{1+\lambda^j T}\beta_{(j)}|\right) \leq \delta^C;$$

$$\Pr\left(\sum_{j=d^*+1}^{n} |\frac{b_j \lambda^j}{1+\lambda^j T}(X\prime u)_{(j)}| > \sum_{j=d-C+1}^{n} |\frac{b_j \lambda^j}{1+\lambda^j T}(X\prime u)_{(j)}|\right) \leq \delta^C.$$

In addition, observe that

$$\mathbb{E}\left[\sum_{i=1}^{d+C} |\frac{b_i}{1+\lambda^i T}\frac{1}{T}(X\prime u)_{(i)}|\right] \leq \sqrt{\frac{2}{\pi}}\frac{\lambda^{-C}}{\sqrt{(1-\lambda)T}}\sqrt{\sum_{i=1}^{d+C} b_i^2 \lambda^{d+C-i}};$$

$$\mathbb{E}\left[\sum_{j=d-C+1}^{n} |\frac{b_j T\lambda^j}{1+\lambda^j T}\beta_{(j)}|\right] \leq \sqrt{\frac{2}{\pi}}\frac{\lambda^{-C/2}}{\sqrt{(1-\lambda^{1/2})T}}\sqrt{\sum_{j=1}^{n-d+C} b_{j+d-C}^2 \lambda^{j/2}};$$

$$\begin{aligned}\mathbb{E}\left[\sum_{j=d-C+1}^{n} |\frac{b_j \lambda^j}{1+\lambda^j T}(X\prime u)_{(j)}|\right] &\leq \mathbb{E}\left[\sum_{j=d-C+1}^{n} |b_j \lambda^j (X\prime u)_{(j)}|\right] \\ &\leq \sqrt{\frac{2}{\pi}}\frac{\lambda^{-C}}{\sqrt{(1-\lambda)T}}\sqrt{\sum_{j=1}^{n-d+C} b_{d-C+j}^2 \lambda^j}.\end{aligned}$$

All these expectations go to 0 after multiplied with $\sqrt{\frac{S_d}{T}}$, hence by similar arguments for the first term, all four terms are of order $o\left(\sqrt{\frac{S_d}{T}}\right)$. ■

We therefore conclude that

$$(\tilde{\beta}-\hat{\beta})\prime BB\prime(\tilde{\beta}-\hat{\beta}) = o\left((\beta-\hat{\beta})\prime BB\prime(\beta-\hat{\beta})\right)$$

for all B of finite number of columns and each column $B_{(i)}$ satisfies the slow growth condition, i.e. $\sum_{j=1}^{n} B_{(ij)}^2$ is of slow growth in n.

5 Conclusion

We have shown that the AIC model selection would select approximately the same number of parameters as the Bayesian method. The interpretation is that

suppose the information we have about the data generating is as described in the introduction, then given our knowledge about the decreasing nature of the $\beta(i)$'s, our best estimator would be the bayesian estimator $\hat{\beta}$. However usually we cannot know the exact rate of decrease in the $\beta(i)$'s, and hence there is usually no way of constructing such bayesian estimator in practice. The above analysis shows that we do not need such a bayesian estimator because applying the AIC to sieve estimators results in a good approximation to the Bayesian estimator. Therefore it is optimal compared to every other estimator.

Moreover, although we have analyzed when $\sigma^2(\beta_{(i)}) = \lambda^i$ for $\lambda \in (0,1)$, it can be seen that the above argument carries through as long as $\sigma^2(\beta_{(i)})$ is decreasing even faster than exponentially in i. Hence AIC is approximately the best estimator as long as the prior knowledge for $\beta(i)$ indicates an at least exponentially decreasing variances in i.

References

[1] Akaike, H. (1973) Information theory and an extension of the maximum likelihood principle. In *Proc. 2nd Int. Symp. Info. Theory*, 267-281, eds. B.N. Petrov and F. Csaki, Akademia Kiado, Budapest.

[2] Phillips, P. C. B. and Ploberger, W. (1994) Posterior Odds Testing for a Unit Root with Data-Based Model Selection. *Econometric Theory*, Vol. 10, 774-808.

[3] Kapoor, G.P. and Nautiyal, A. (1981) Polynomial Approximation of an Entire Function of Slow Growth. *Journal of Approximation Theory*, Vol. 32, 64-75.

[4] Kim, J. Y. (1998) Large Sample Properties of Posterior Densities, Bayesian Information Criterion and the Likelihood Principle in Nonstationary Time Series Models. *Econometrica*, 66, 359-380.

[5] Schwarz, Gideon E. (1978), "Estimating the dimension of a model", *Annals of Statistics*, 6 (2): 461?464

[6] Shibata, R. (1983) Asymptotic mean efficiency of a selection of regression variables. *Ann. Inst. Statisti. Math.* 35, 415-423.

[7] Strasser, H. (1985) *Mathematical theory of statistics: statistical experiments and asymptotic decision theory*, Walter de Gruyter.

[8] Stuart, A. M. (2010) Inverse problems: A Bayesian perspective. *Acta Numerica*, 19, 451-559.

[9] van der Vaart, A. W. (2000) *Asymptotic Statistics*, Cambridge University Press.