Department of Social Systems and Management

Discussion Paper Series

No.1303

Variable Selection for Bayesian Linear Regression Model in a Finite Sample Size

by

Satoshi KABE and Yuichiro KANAZAWA

February 2013

UNIVERSITY OF TSUKUBA
Tsukuba, Ibaraki 305-8573
JAPAN

Variable Selection for Bayesian Linear Regression Model in a Finite Sample Size

Satoshi KABE * Yuichiro KANAZAWA †

Abstract

In Bayesian data analysis, a deviance information criterion (DIC) proposed by Spiegelhalter et al. (2002) is widely used for the model selection, since this criterion is relatively easy to calculate and applicable to a wide range of statistical models. Spiegelhalter et al. (2002) gave an asymptotic justification of DIC in the case where the number of observations grows with respect to the number of parameters. In small-sample cases, however, the estimated asymptotic bias of DIC might underestimate the true bias (Burnham, 2002). In this paper, we propose a finite-sample bias corrected information criterion (IC_{BL}) for the Bayesian linear regression models with conjugate priors, as AIC_C proposed by Sugiura (1978) in frequentist framework. We examine the performance of the proposed information criterion relative to the DIC for small-sample cases by simulation, and found that our proposed information criterion outperforms DIC.

*Doctoral Student, Graduate School of Systems and Information, University of Tsukuba. E-mail: k0420214@sk.tsukuba.ac.jp.

†Professor of Statistics, Department of Social Systems and Management, University of Tsukuba. Email: kanazawa@sk.tsukuba.ac.jp.

1 Introduction

The data analysis often involves a comparison of several candidate models. Because true model is seldom known a priori, there is a need for a simple, effective, and objective methods for the selection of the best approximating model. Akaike (1973) proposed an information criterion later known to be Akaike's information criterion (AIC) to be an extension of likelihood theory. AIC is based on the concept of minimizing the Kullback-Leibler Information, a measure of discrepancy between the true density (or model) and approximating model. The discrepancy between two models or probability densities is expressed by the expected log-likelihood with respect to the true density. AIC is designed to be an approximately unbiased estimator of the expected log-likelihood under the assumption that "true model" exists and it is one of the candidate models being considered (see Appendix D for detail).

Hurvich and Tsai (1989) showed that AIC can be dramatically biased when the sample size is small. The finite bias-correction version of AIC (AIC_C) for linear regression model with normally distributed error was proposed by Sugiura (1978). This criterion adjusts the AIC to be an exact unbiased estimator of the expected log-likelihood. AIC_C is extended to autoregressive (AR) model and autoregressive moving average (ARMA) model (Hurvich and Tsai, 1989) and multivariate regression model (Bedrick and Tsai, 1994).

The Bayesian information criterion (BIC) was proposed by Schwarz (1978). BIC selects the best approximating model with the highest posterior probability from the candidate models given the *i.i.d.* observed data $\boldsymbol{x} \equiv (x_1, x_2, \ldots, x_N)$. This criterion is derived from the marginal likelihood $p(\boldsymbol{x}|M)$, where M stands for model, under the assumption that each competing parametric model has the same prior probability $p(M)$ and prior distribution on

the parameters $p(\boldsymbol{\theta}|M)$ is very vague. Because of this "ignorance" prior, many statisticians consider BIC as a variation of AIC, and not a 'fully' Bayesian model selection criterion that allows the incorporation of prior information. In BIC, the marginal likelihood are approximated by the Laplace's method which requires a large number of observations. Unlike AIC, however derivation of BIC does not require an assumption that the true model is in one of the candidate models.

In the 'fully' Bayesian data analysis, marginal likelihood is often used to evaluate the goodness of fit of approximating models. To evaluate an evidence in favor of one model against another, Bayes factor is widely used according to Kass and Raftery (1995) and they proposed a method of assessing the strength of evidence extending the method proposed by Jeffreys (1961).

The Bayes factor is simply expressed by the ratio of marginal likelihoods with the same prior probability on each model. The large-sample distribution theory for Bayes factor is not yet available unlike the standard likelihood ratio test in the non-Bayesian approach.

For computing the marginal likelihoods in the Bayes factor, one needs to integrate over parameters and in general, this integration is difficult when the number of parameters is large. Thus, under the finite sample size, approximation methods for marginal likelihood using the posterior distributions are proposed by several researchers (e.g., Newton and Raftery, 1994; Gelfand and Dey, 1994).

To resolve a problem of comparing complex hierarchical Bayesian models in which the number of parameters can be open to interpretation, Spiegelhalter et al. (2002) suggested effective number of parameters p_D for the Bayesian models as the difference between the posterior mean of the deviance and the deviance at the posterior mean of the parameters, and they used it as a

Bayesian measure of complexity in a model. As the number of data is sufficiently large, a deviance information criterion (DIC) is given by adding p_D to the posterior mean of the deviance.

In this paper, we propose a finite-sample bias corrected information criterion (IC_{BL}) for the Bayesian linear regression models with conjugate priors, because Spiegelhalter et al. (2002) gave an asymptotic justification of DIC in the case where the number of observations is large with respect to the number of parameters. We examine the finite-sample bias correction when the posterior mean of observed log-likelihood is used as an estimator of posterior mean of expected log-likelihood such as Sugiura (1978) has done to AIC in frequentist case. We evaluate the results of simulation studies based on the proposed information criterion relative to the DIC, and as an empirical example, we estimate the cost functions for the U.S. electric power industry and select the best approximating model via our proposed information criterion in the set of the candidate models.

The rest of this paper is organized as follows: Next section introduces the bias correction for the posterior mean of observed log-likelihood and propose our information criterion for the variable selection in the Bayesian linear regression model. Section 3 shows the results of simulation studies to show the validity of our proposed information criterion when the sample size is small. Section 4 presents the empirical example of the U.S. electric power industry, and the last section contains the conclusions of our study.

2 Variable Selection for Bayesian Linear Regression Model in a Finite Sample Size

Let us denote unknown true density as $f_{\mathbf{Y}}(\cdot)$ and approximating model as $g(\cdot|\boldsymbol{\theta})$ with parameter vector $\boldsymbol{\theta}$. Then Kullback-Leibler Information between $f_{\mathbf{Y}}$ and g can be expressed as follows

$$I(f_{\mathbf{Y}}, g(\cdot|\boldsymbol{\theta})) = \int f_{\mathbf{Y}}(\boldsymbol{z}) \log \left\{ \frac{f_{\mathbf{Y}}(\boldsymbol{z})}{g(\boldsymbol{z}|\boldsymbol{\theta})} \right\} d\boldsymbol{z} \tag{2.1}$$

and (2.1) can be rewritten as

$$I(f_{\mathbf{Y}}, g(\cdot|\boldsymbol{\theta})) = \mathbf{E}_z \left[\log\{f_{\mathbf{Y}}(\boldsymbol{z})\}\right] - \mathbf{E}_z \left[\log\{g(\boldsymbol{z}|\boldsymbol{\theta})\}\right]. \tag{2.2}$$

Even though the true density $f_{\mathbf{Y}}(\cdot)$ is unknown, the first term on the right-hand side of Kullback-Leibler Information in (2.2) can be regarded as a constant since the variable $\boldsymbol{z}$ is integrated out.

In Bayesian perspective, parameters follow the posterior distributions estimated by observed data $\boldsymbol{y}$. Hence we consider the posterior mean of (2.2):

$$\mathbf{E}_{\theta|y} \left[I(f_{\mathbf{Y}}, g(\cdot|\boldsymbol{\theta}))\right] = \mathbf{E}_z \left[\log\{f_{\mathbf{Y}}(\boldsymbol{z})\}\right] - \mathbf{E}_{\theta|y} \left[\mathbf{E}_z \left[\log\{g(\boldsymbol{z}|\boldsymbol{\theta})\}\right]\right] \tag{2.3}$$

and as in Spiegelhalter et al. (2002) and Ando (2007), our proposed information criterion is constructed based on the posterior mean of expected log-likelihood.

As in Bayesian linear regression model in (A.1), we use $\boldsymbol{y}$ as observed data of sample size N obtained from the unknown true density $f_{\mathbf{Y}}(\boldsymbol{y})$ to estimate the posterior distributions of parameters $\boldsymbol{\beta}$ and σ^{-2}, while we also use $\boldsymbol{z}$ as replicate data of sample size N generated from the unknown true density $f_{\mathbf{Y}}(\boldsymbol{z})$ to evaluate the goodness of fit of approximating model $g(\boldsymbol{z}|\boldsymbol{X}, \boldsymbol{\beta}, \sigma^{-2})$. Then we select the best approximating model with maximizing the posterior

mean of expected log-likelihood T as in the second term on the right-hand side of (2.3)

$$T \equiv \mathbf{E}_{\beta,\sigma^{-2}|y,X}\left[\mathbf{E}_z\left[\log\left\{g(\boldsymbol{z}|\boldsymbol{X},\boldsymbol{\beta},\sigma^{-2})\right\}\right]\right] \tag{2.4}$$

where we assume that expectation with respect to the joint posterior distribution $\mathbf{E}_{\beta,\sigma^{-2}|y,X}[\,\cdot\,]$ can be calculated by $\mathbf{E}_{\beta,\sigma^{-2}|y,X}[\,\cdot\,] \equiv \mathbf{E}_{\sigma^{-2}|y,X}[\ \mathbf{E}_{\beta|\sigma^{-2}y,X}[\,\cdot\,]\ ]$ from (A.15) and (A.16).

To estimate the posterior mean of expected log-likelihood T in (2.4), we use the posterior mean of observed log-likelihood $\widehat{T}_N$:

$$\widehat{T}_N \equiv \mathbf{E}_{\beta,\sigma^{-2}|y,X}\left[\log\left\{g(\boldsymbol{y}|\boldsymbol{X},\boldsymbol{\beta},\sigma^{-2})\right\}\right] \tag{2.5}$$

and its bias $b_\Theta \equiv \mathbf{E}_y[\widehat{T}_N - T] \neq 0$ to obtain the bias-corrected estimator $\widehat{T}_N - \widehat{b}_N$, where $\widehat{b}_N$ is the estimate of b_Θ. Then we propose information criterion (IC) of the form

$$\mathrm{IC} \equiv -2\widehat{T}_N + 2\widehat{b}_N \tag{2.6}$$

as in (D.15) and (D.16), so that we can choose the best approximating model that minimizes IC in (2.6).

Ignoring the constant term, we can express the log-likelihood function for the replicate data $\boldsymbol{z}$ such as

$$\log\left\{g(\boldsymbol{z}|\boldsymbol{X},\boldsymbol{\beta},\sigma^{-2})\right\} = \frac{N}{2}\log\sigma^{-2} - \frac{\sigma^{-2}}{2}(\boldsymbol{z}-\boldsymbol{X}\boldsymbol{\beta})'(\boldsymbol{z}-\boldsymbol{X}\boldsymbol{\beta}) \tag{2.7}$$

where parameters $\boldsymbol{\beta}$ and σ^{-2} follow the posterior distributions estimated by observed data $\boldsymbol{y}$ and $\boldsymbol{X}$. Then the posterior mean of expected log-likelihood T in (2.4) is expressed as

$$\begin{aligned} T &= \mathbf{E}_{\beta,\sigma^{-2}|y,X}\left[\mathbf{E}_z\left[\log\left\{g(\boldsymbol{z}|\boldsymbol{X},\boldsymbol{\beta},\sigma^{-2})\right\}\right]\right] \\ &= \mathbf{E}_{\beta,\sigma^{-2}|y,X}\left[\mathbf{E}_z\left[\frac{N}{2}\log\sigma^{-2} - \frac{\sigma^{-2}}{2}(\boldsymbol{z}-\boldsymbol{X}\boldsymbol{\beta})'(\boldsymbol{z}-\boldsymbol{X}\boldsymbol{\beta})\right]\right] \end{aligned}$$

$$
\begin{aligned}
&= \mathbf{E}_{\beta,\sigma^{-2}|y,X}\left[\frac{N}{2}\log\sigma^{-2} - \frac{\sigma^{-2}}{2}(\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta})'(\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta})\right] \\
&\quad - \mathbf{E}_{\beta,\sigma^{-2}|y,X}\left[\mathbf{E}_z\left[\frac{\sigma^{-2}}{2}(\boldsymbol{z}-\boldsymbol{X}\boldsymbol{\beta})'(\boldsymbol{z}-\boldsymbol{X}\boldsymbol{\beta})\right]\right] \\
&\quad + \mathbf{E}_{\beta,\sigma^{-2}|y,X}\left[\frac{\sigma^{-2}}{2}(\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta})'(\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta})\right] \\
&= \widehat{T}_N - C_1 + C_2. \qquad (2.8)
\end{aligned}
$$

When the posterior mean of observed log-likelihood $\widehat{T}_N$ is used as an estimator of posterior mean of expected log-likelihood T, the bias b_Θ with respect to the true density $f_{\mathbf{Y}}(\boldsymbol{y})$ is obtained by $b_\Theta \equiv \mathbf{E}_y[\widehat{T}_N - T] = \mathbf{E}_y[C_1 - C_2]$.

First we evaluate C_1 in (2.8). However, true density $f_{\mathbf{Y}}(\boldsymbol{z})$ is seldom known in practice, so that expectation with respect to the true density is not analytically obtained. In the previous studies, Kitagawa (1997) replaced the unknown true density by the prior predictive density to construct the predictive information criterion (PIC) for the Bayesian linear Gaussian model, while Laud and Ibrahim (1995), Gelfand and Ghosh (1998), and Ibrahim et al. (2001) considered using the posterior predictive density to generate the replicate data $\boldsymbol{z}$ for model assessment. In this paper, we use the posterior predictive density to evaluate the expectation with respect to the true density $f_{\mathbf{Y}}(\boldsymbol{z})$ in C_1 because the prior predictive density is far more sensitive to the selection of prior distribution.

To evaluate C_1 in (2.8), we assume that true density $f_{\mathbf{Y}}(\cdot)$ is a N-dimensional multivariate normal distribution with unknown true parameters and replace the true density with a (conditional) posterior predictive density

$$
\boldsymbol{z}|\sigma^{-2}, \boldsymbol{y}, \boldsymbol{X} \sim \mathcal{N}\left(\boldsymbol{X}\boldsymbol{b}_1, \sigma^2\boldsymbol{\Sigma}_0\right), \qquad (2.9)
$$

where σ^{-2} follows the posterior distribution in (A.16) and $\boldsymbol{\Sigma}_0 = \boldsymbol{I}_N + \boldsymbol{X}\boldsymbol{B}_1\boldsymbol{X}'$ (see Appendix A).

From (2.9), we estimate C_1 in (2.8) as

$$\begin{aligned}
\widehat{C}_1 &= \mathbf{E}_{\beta,\sigma^{-2}|y,X}\left[\mathbf{E}_{z|\sigma^{-2},y,X}\left[\frac{\sigma^{-2}}{2}(\boldsymbol{z}-\boldsymbol{X}\boldsymbol{\beta})'(\boldsymbol{z}-\boldsymbol{X}\boldsymbol{\beta})\right]\right] \\
&= \mathbf{E}_{\beta,\sigma^{-2}|y,X}\left[\mathbf{E}_{z|\sigma^{-2},y,X}\left[\frac{\sigma^{-2}}{2}(\boldsymbol{z}-\boldsymbol{X}\boldsymbol{b}_1+\boldsymbol{X}\boldsymbol{b}_1-\boldsymbol{X}\boldsymbol{\beta})'(\boldsymbol{z}-\boldsymbol{X}\boldsymbol{b}_1+\boldsymbol{X}\boldsymbol{b}_1-\boldsymbol{X}\boldsymbol{\beta})\right]\right] \\
&= \mathbf{E}_{\beta,\sigma^{-2}|y,X}\left[\mathbf{E}_{z|\sigma^{-2},y,X}\left[\frac{\sigma^{-2}}{2}(\boldsymbol{z}-\boldsymbol{X}\boldsymbol{b}_1)'(\boldsymbol{z}-\boldsymbol{X}\boldsymbol{b}_1)\right]\right] \\
&\quad + \mathbf{E}_{\beta,\sigma^{-2}|y,X}\left[\frac{\sigma^{-2}}{2}(\boldsymbol{\beta}-\boldsymbol{b}_1)'(\boldsymbol{X}'\boldsymbol{X})(\boldsymbol{\beta}-\boldsymbol{b}_1)\right] \\
&= \mathbf{E}_{\beta,\sigma^{-2}|y,X}\left[\mathbf{E}_{z|\sigma^{-2},y,X}\left[\frac{\sigma^{-2}}{2}(\boldsymbol{z}-\boldsymbol{X}\boldsymbol{b}_1)'(\boldsymbol{z}-\boldsymbol{X}\boldsymbol{b}_1)\right]\right] \\
&\quad + \mathbf{E}_{\sigma^{-2}|y,X}\left[\mathrm{tr}\left\{\frac{\sigma^{-2}}{2}(\boldsymbol{X}'\boldsymbol{X})\mathbf{E}_{\beta|\sigma^{-2},y,X}\left[(\boldsymbol{\beta}-\boldsymbol{b}_1)(\boldsymbol{\beta}-\boldsymbol{b}_1)'\right]\right\}\right].
\end{aligned} \tag{2.10}$$

From (2.9), the first term on the right-hand side of (2.10) can be rewritten as follows

$$\begin{aligned}
&\mathbf{E}_{\beta,\sigma^{-2}|y,X}\left[\mathbf{E}_{z|\sigma^{-2},y,X}\left[\frac{\sigma^{-2}}{2}(\boldsymbol{z}-\boldsymbol{X}\boldsymbol{b}_1)'(\boldsymbol{z}-\boldsymbol{X}\boldsymbol{b}_1)\right]\right] \\
&= \mathbf{E}_{\sigma^{-2}|y,X}\left[\frac{\sigma^{-2}}{2}\mathrm{tr}\Big\{\mathbf{E}_{z|\sigma^{-2},y,X}\left[(\boldsymbol{z}-\boldsymbol{X}\boldsymbol{b}_1)(\boldsymbol{z}-\boldsymbol{X}\boldsymbol{b}_1)'\right]\Big\}\right] \\
&= \mathbf{E}_{\sigma^{-2}|y,X}\left[\frac{\sigma^{-2}}{2}\mathrm{tr}\Big\{\sigma^2\boldsymbol{\Sigma}_0\Big\}\right] \\
&= \frac{1}{2}\mathrm{tr}\left\{\boldsymbol{I}_N+\boldsymbol{X}\boldsymbol{B}_1\boldsymbol{X}'\right\} \\
&= \frac{N}{2}+\frac{1}{2}\mathrm{tr}\left\{(\boldsymbol{X}'\boldsymbol{X})\boldsymbol{B}_1\right\}.
\end{aligned} \tag{2.11}$$

The second term on the right-hand side of (2.10) is obtained similarly from (A.11) by

$$\begin{aligned}
&\mathbf{E}_{\sigma^{-2}|y,X}\left[\mathrm{tr}\left\{\frac{\sigma^{-2}}{2}(\boldsymbol{X}'\boldsymbol{X})\mathbf{E}_{\beta|\sigma^{-2},y,X}\left[(\boldsymbol{\beta}-\boldsymbol{b}_1)(\boldsymbol{\beta}-\boldsymbol{b}_1)'\right]\right\}\right] \\
&= \mathbf{E}_{\sigma^{-2}|y,X}\left[\mathrm{tr}\left\{\frac{\sigma^{-2}}{2}(\boldsymbol{X}'\boldsymbol{X})\sigma^2\boldsymbol{B}_1\right\}\right] \\
&= \frac{1}{2}\mathrm{tr}\left\{(\boldsymbol{X}'\boldsymbol{X})\boldsymbol{B}_1\right\}.
\end{aligned} \tag{2.12}$$

From (2.11) and (2.12), $\widehat{C}_1$ in (2.10) is evaluated as follows

$$\begin{aligned}
\widehat{C}_1 &= \mathbf{E}_{\beta,\sigma^{-2}|y,X}\left[\mathbf{E}_{z|\sigma^{-2},y,X}\left[\frac{\sigma^{-2}}{2}(\boldsymbol{z}-\boldsymbol{X}\boldsymbol{b}_1)'(\boldsymbol{z}-\boldsymbol{X}\boldsymbol{b}_1)\right]\right] \\
&\quad + \mathbf{E}_{\sigma^{-2}|y,X}\left[\mathrm{tr}\left\{\frac{\sigma^{-2}}{2}(\boldsymbol{X}'\boldsymbol{X})\mathbf{E}_{\beta|\sigma^{-2},y,X}\left[(\boldsymbol{\beta}-\boldsymbol{b}_1)(\boldsymbol{\beta}-\boldsymbol{b}_1)'\right]\right\}\right] \\
&= \frac{N}{2} + \frac{1}{2}\mathrm{tr}\left\{(\boldsymbol{X}'\boldsymbol{X})\boldsymbol{B}_1\right\} + \frac{1}{2}\mathrm{tr}\left\{(\boldsymbol{X}'\boldsymbol{X})\boldsymbol{B}_1\right\} \\
&= \frac{N}{2} + \mathrm{tr}\left\{(\boldsymbol{X}'\boldsymbol{X})\boldsymbol{B}_1\right\}. \qquad (2.13)
\end{aligned}$$

Since $\widehat{C}_1$ does not depend on any data $\boldsymbol{y}$, we have $\mathbf{E}_y(\widehat{C}_1) = \widehat{C}_1$.

Suppose that interchange of order of integrations is valid, we can rewrite $\mathbf{E}_y(C_2)$ such as

$$\begin{aligned}
\mathbf{E}_y\left(C_2\right) &= \mathbf{E}_y\left[\mathbf{E}_{\beta,\sigma^{-2}|y,X}\left[\frac{\sigma^{-2}}{2}(\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta})'(\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta})\right]\right] \\
&= \mathbf{E}_{\beta,\sigma^{-2}}\left[\mathbf{E}_{y|X,\beta,\sigma^{-2}}\left[\frac{\sigma^{-2}}{2}(\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta})'(\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta})\right]\right] \qquad (2.14)
\end{aligned}$$

where $\mathbf{E}_{\beta,\sigma^{-2}}[\,\cdot\,]$ is an expectation with respect to joint prior distribution and $\mathbf{E}_{y|X,\beta,\sigma^{-2}}[\,\cdot\,]$ is an expectation with respect to N-dimensional multivariate normal distribution with mean vector $\boldsymbol{X}\boldsymbol{\beta}$ and variance covariance matrix $\sigma^2\boldsymbol{I}_N$. Since $\mathbf{E}_{y|X,\beta,\sigma^{-2}}[(\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta})(\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta})'] = \sigma^2\boldsymbol{I}_N$, $\mathbf{E}_y(C_2)$ can be evaluated as

$$\begin{aligned}
\mathbf{E}_y\left(C_2\right) &= \mathbf{E}_{\beta,\sigma^{-2}}\left[\mathbf{E}_{y|X,\beta,\sigma^{-2}}\left[\mathrm{tr}\left\{\frac{\sigma^{-2}}{2}(\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta})(\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta})'\right\}\right]\right] \\
&= \frac{1}{2}\mathrm{tr}\left\{\boldsymbol{I}_N\right\} \\
&= \frac{N}{2}. \qquad (2.15)
\end{aligned}$$

Therefore bias $\widehat{b}_N$ is obtained by

$$\begin{aligned}
\widehat{b}_N &= \mathbf{E}_y\left[\widehat{C}_1 - C_2\right] \\
&= \widehat{C}_1 - \mathbf{E}_y\left(C_2\right)
\end{aligned}$$

$$= \frac{N}{2} + \mathrm{tr}\left\{(\boldsymbol{X}'\boldsymbol{X})\boldsymbol{B}_1\right\} - \frac{N}{2}$$
$$= \mathrm{tr}\left\{(\boldsymbol{X}'\boldsymbol{X})\boldsymbol{B}_1\right\}. \tag{2.16}$$

Then the bias $\widehat{b}_N$ in (2.16) can be regarded as a ratio of variance covariance matrices $\sigma^2(\boldsymbol{X}'\boldsymbol{X})^{-1}$ and $\sigma^2\boldsymbol{B}_1$.

Multiplying -2 to the bias-corrected estimator $\widehat{T}_N - \widehat{b}_N$, our proposed information criterion for variable selection in the Bayesian linear regression model (IC_{BL}) is obtained by

$$\mathrm{IC}_{BL} = -2\widehat{T}_N + 2\mathrm{tr}\left\{(\boldsymbol{X}'\boldsymbol{X})\boldsymbol{B}_1\right\} \tag{2.17}$$

where $\boldsymbol{B}_1 = \left(\boldsymbol{X}'\boldsymbol{X} + \boldsymbol{B}_0^{-1}\right)^{-1}$.

For simplicity, let us denote the parameter $\boldsymbol{B}_0$ as $\boldsymbol{B}_0 = \kappa_0\boldsymbol{I}_K$, $(\kappa_0 > 0)$ and the bias term $\widehat{b}_N$ can be rewritten as

$$\widehat{b}_N = K - \mathrm{tr}\{(\boldsymbol{X}'\boldsymbol{X}\boldsymbol{B}_0 + \boldsymbol{I}_K)^{-1}\} \tag{2.18}$$

from the matrix inversion lemma [1]. Then if the sample size $N \to \infty$, the last term in (2.18) is expected to be zero because $\mathrm{tr}\{(\kappa_0\boldsymbol{X}'\boldsymbol{X}/N + \boldsymbol{I}_N/N)^{-1}/N\} \to 0$ (i.e., $\widehat{b}_N \to K$) when each element of $\boldsymbol{X}'\boldsymbol{X}/N$ does not diverge under the standard set of assumptions. Furthermore, if κ_0 is sufficiently large (i.e., non-informative prior), we also have $\mathrm{tr}\{(\kappa_0\boldsymbol{X}'\boldsymbol{X} + \boldsymbol{I}_K)^{-1}\} \to 0$ in (2.18).

[1] For any matrices $\boldsymbol{A}$ $(m \times m)$, $\boldsymbol{B}$ $(m \times n)$, $\boldsymbol{C}$ $(n \times m)$, and $\boldsymbol{D}$ $(n \times n)$, we have

$$\left(\boldsymbol{A} - \boldsymbol{B}\boldsymbol{D}^{-1}\boldsymbol{C}\right)^{-1} = \boldsymbol{A}^{-1} + \boldsymbol{A}^{-1}\boldsymbol{B}\left(\boldsymbol{D} - \boldsymbol{C}\boldsymbol{A}^{-1}\boldsymbol{B}\right)^{-1}\boldsymbol{C}\boldsymbol{A}^{-1}$$

where $\boldsymbol{A}$ and $\boldsymbol{D}$ are nonsingular matrices.

3 Simulation study

We conduct a simulation study to compare the small-sample performance of our proposed information criterion (IC_{BL}) in (2.17) and deviance information criterion (DIC) which is computed as

$$\mathrm{DIC} = -2\widehat{T}_N + p_D, \tag{3.1}$$

where Spiegelhalter et al. (2002) termed p_D the effective number of parameters defined as $p_D \equiv 2\log\{g(\boldsymbol{y}|\boldsymbol{X}, \bar{\boldsymbol{\beta}}, \bar{\sigma}^{-2})\} - 2\widehat{T}_N$ evaluated at the posterior means of parameters $\bar{\boldsymbol{\beta}}$ $(= \boldsymbol{b}_1)$ and $\bar{\sigma}^{-2}$ $(= \nu_1/\lambda_1)$ in (A.15) and (A.16).

As in Hurvich and Tsai (1989), we consider the nested candidate models by using seven explanatory variables $\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_7$. In this paper, $\boldsymbol{x}_1$ is a $N \times 1$ vector whose elements are ones (i.e., intercept term) and the other $N \times 1$ vectors $\boldsymbol{x}_i$ $(2 \leq i \leq 7)$ are generated from the uniform distribution $\mathcal{U}(-2, 2)$. These variables $\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_7$ are included into the candidate models in a sequentially nested fashion. The candidate models are linear regression models given by $\boldsymbol{y} = \beta_1\boldsymbol{x}_1 + \beta_2\boldsymbol{x}_2 + \cdots + \beta_K\boldsymbol{x}_K + \boldsymbol{\varepsilon}$, where $\boldsymbol{\varepsilon} \sim \mathcal{N}(\boldsymbol{0}, \sigma^2\boldsymbol{I}_N)$. The candidate model with $K = 1$ has only intercept term and with $K = 7$ is the full model. In this simulation study, we determine the number of variables K by using our proposed information criterion (IC_{BL}) in (2.17) and DIC in (3.1) in small-sample cases $N = 25, 50, 100$ with informative $(\kappa_0 = 0.1)$ and non-informative $(\kappa_0 = 100)$ priors. To examine the small-sample performance in the Bayesian linear regression case, we generate a sample of $\boldsymbol{y}$ from the true model $(K = 3)$,

$$\boldsymbol{y} = 1.0\boldsymbol{x}_1 + 2.0\boldsymbol{x}_2 + 3.0\boldsymbol{x}_3 + \boldsymbol{\varepsilon}, \quad \boldsymbol{\varepsilon} \sim \mathcal{N}\left(\boldsymbol{0}, 1.0\boldsymbol{I}_N\right). \tag{3.2}$$

In Table 1, we examine the performance of IC_{BL} and DIC for the small-sample cases $(N = 25, 50, 100)$. The parameters of prior distributions in

(A.2) and (A.3) are set to be $\boldsymbol{b}_0 = \mathbf{0}$, $\boldsymbol{B}_0 = \kappa_0 \boldsymbol{I}_K$ ($\kappa_0 = 0.1$ or 100), and $\nu_0 = \lambda_0 = 0.1$. The simulation considered each combination of $N = 25, 50, 100$ and $\kappa_0 = 0.1, 100$. 50,000 MCMC draws are generated from the posterior distributions in (A.15) and (A.16) to compute the posterior mean of observed log-likelihood $\widehat{T}_N$ in (2.5). For each combination of (N, κ_0), we generate 100 observations of IC_{BL} and DIC, and record the number of selected models (i.e., the candidate model with minimum value for the two criteria).

Table 1 shows that our proposed information criterion (IC_{BL}) identifies the true model ($K = 3$) for the small-sample cases ($N = 25, 50, 100$) with informative prior ($\kappa_0 = 0.1$) far better than DIC because DIC tends to overfit the model for these small-sample cases. On the other hand, for the non-informative prior ($\kappa_0 = 100$), both criteria tend to overfit the model for the sample size $N = 25$ and 50, but nevertheless our proposed information criterion (IC_{BL}) far outperformes DIC at the sample size $N = 100$.

In Tables 2 and 3, we show the results of average criteria in 100 observations for the small-sample cases ($N = 25, 50, 100$) with informative ($\kappa_0 = 0.1$) and non-informative ($\kappa_0 = 100$) priors. Both criteria selected the true model ($K = 3$) in all cases, but the difference between $2\widehat{b}_N$ and p_D becomes more apparent along with an increase in the number of explanatory variables. Hence the effective number of parameters p_D in DIC tends to underestimate the complexity of candidate model as compared with the bias term $2\widehat{b}_N$ in IC_{BL}.

Table 1: The number of selected models by IC_{BL} and DIC for small-sample cases $N = 25, 50, 100$ with informative ($\kappa_0 = 0.1$) and non-informative ($\kappa_0 = 100$) priors.

	Informative prior ($\kappa_0 = 0.1$)						Non-informative prior ($\kappa_0 = 100$)					
	$N = 25$		$N = 50$		$N = 100$		$N = 25$		$N = 50$		$N = 100$	
Model (K)	IC_{BL}	DIC	IC_{BL}	DIC	IC_{BL}	DIC	IC_{BL}	DIC	IC_{BL}	DIC	IC_{BL}	DIC
1	0	0	0	0	0	0	0	0	0	0	0	0
2	0	0	0	0	0	0	0	0	0	0	0	0
3	97	79	95	74	95	76	81	65	89	75	90	75
4	2	9	5	15	5	14	8	10	5	13	7	12
5	1	6	0	7	0	3	7	9	4	8	3	3
6	0	5	0	2	0	5	0	3	0	1	0	6
7	0	1	0	2	0	2	4	13	2	3	0	4

Table 2: Average values of IC_{BL} and DIC in 100 observations for small-sample cases $N = 25, 50, 100$ with informative prior ($\kappa_0 = 0.1$).

	Informative prior ($\kappa_0 = 0.1$)														
	$N = 25$					$N = 50$					$N = 100$				
Model (K)	IC_{BL}	DIC	$\widehat{T}_N$	$2\widehat{b}_N$	p_D	IC_{BL}	DIC	$\widehat{T}_N$	$2\widehat{b}_N$	p_D	IC_{BL}	DIC	$\widehat{T}_N$	$2\widehat{b}_N$	p_D
1	99.304	99.558	-48.938	1.429	1.683	196.367	196.519	-97.350	1.667	1.819	394.064	394.147	-196.123	1.818	1.902
	(6.596)	(6.595)	(3.298)	(0.000)	(0.008)	(9.404)	(9.404)	(4.702)	(0.000)	(0.009)	(12.972)	(12.972)	(6.486)	(0.000)	(0.009)
2	93.196	92.664	-45.134	2.928	2.396	181.946	181.215	-89.275	3.396	2.665	361.471	360.616	-178.897	3.677	2.822
	(5.943)	(5.945)	(2.974)	(0.077)	(0.041)	(7.360)	(7.362)	(3.682)	(0.030)	(0.021)	(9.739)	(9.737)	(4.868)	(0.012)	(0.013)
3	55.425	54.115	-25.502	4.420	3.111	86.241	84.633	-40.565	5.111	3.503	136.230	134.440	-65.349	5.531	3.742
	(2.984)	(2.984)	(1.486)	(0.112)	(0.065)	(5.324)	(5.326)	(2.663)	(0.043)	(0.046)	(10.446)	(10.448)	(5.223)	(0.019)	(0.038)
4	56.745	54.676	-25.436	5.874	3.805	87.682	85.198	-40.430	6.823	4.338	137.879	135.149	-65.248	7.382	4.653
	(3.030)	(3.029)	(1.516)	(0.146)	(0.075)	(5.385)	(5.383)	(2.693)	(0.058)	(0.050)	(10.578)	(10.589)	(5.289)	(0.027)	(0.042)
5	58.004	55.200	-25.357	7.290	4.486	89.188	85.832	-40.333	8.521	5.165	139.577	135.916	-65.171	9.234	5.573
	(3.088)	(3.075)	(1.536)	(0.208)	(0.105)	(5.474)	(5.477)	(2.743)	(0.063)	(0.051)	(10.477)	(10.480)	(5.239)	(0.031)	(0.046)
6	59.228	55.703	-25.272	8.684	5.159	90.776	86.568	-40.283	10.211	6.002	141.136	136.540	-65.026	11.083	6.487
	(3.141)	(3.131)	(1.560)	(0.234)	(0.119)	(5.456)	(5.465)	(2.735)	(0.084)	(0.052)	(10.495)	(10.492)	(5.248)	(0.034)	(0.041)
7	60.470	56.221	-25.201	10.068	5.820	92.206	87.137	-40.160	11.886	6.817	142.716	137.185	-64.895	12.925	7.395
	(3.094)	(3.069)	(1.530)	(0.269)	(0.139)	(5.446)	(5.454)	(2.729)	(0.095)	(0.057)	(10.512)	(10.519)	(5.256)	(0.038)	(0.046)

Standard deviations in parentheses

Table 3: Average values of IC_{BL} and DIC in 100 observations for small-sample cases $N = 25, 50, 100$ with non-informative prior ($\kappa_0 = 100$).

	Non-informative prior ($\kappa_0 = 100$)														
	$N = 25$					$N = 50$					$N = 100$				
Model (K)	IC_{BL}	DIC	$\widehat{T}_N$	$2\widehat{b}_N$	p_D	IC_{BL}	DIC	$\widehat{T}_N$	$2\widehat{b}_N$	p_D	IC_{BL}	DIC	$\widehat{T}_N$	$2\widehat{b}_N$	p_D
1	98.416	98.386	-48.208	1.999	1.970	197.876	197.861	-97.938	2.000	1.984	393.852	393.846	-195.926	2.000	1.993
	(6.588)	(6.589)	(3.294)	(0.000)	(0.010)	(8.148)	(8.147)	(4.074)	(0.000)	(0.010)	(12.035)	(12.034)	(6.017)	(0.000)	(0.010)
2	93.019	91.952	-44.510	3.999	2.932	182.333	181.298	-89.167	3.999	2.965	360.553	359.537	-178.277	4.000	2.984
	(5.732)	(5.732)	(2.866)	(1.65E-04)	(0.011)	(7.243)	(7.243)	(3.622)	(4.67E-05)	(0.011)	(10.999)	(10.998)	(5.500)	(1.49E-05)	(0.011)
3	31.506	29.409	-12.754	5.998	3.901	56.363	54.310	-25.182	5.999	3.946	108.119	106.092	-51.060	5.999	3.973
	(7.097)	(7.097)	(3.549)	(2.90E-04)	(0.012)	(12.093)	(12.090)	(6.047)	(7.37E-05)	(0.013)	(13.496)	(13.498)	(6.748)	(2.42E-05)	(0.012)
4	33.519	30.392	-12.761	7.997	4.869	58.419	55.352	-25.210	7.999	4.932	110.077	107.040	-51.039	7.999	4.962
	(7.281)	(7.282)	(3.641)	(3.70E-04)	(0.014)	(12.074)	(12.074)	(6.037)	(1.05E-04)	(0.016)	(13.381)	(13.381)	(6.691)	(2.78E-05)	(0.014)
5	35.083	30.928	-12.544	9.996	5.841	60.275	56.190	-25.138	9.998	5.914	111.959	107.913	-50.980	9.999	5.952
	(7.372)	(7.373)	(3.686)	(4.98E-04)	(0.015)	(12.271)	(12.270)	(6.136)	(1.35E-04)	(0.015)	(13.369)	(13.370)	(6.685)	(3.29E-05)	(0.016)
6	37.234	32.048	-12.620	11.995	6.809	62.496	57.395	-25.249	11.998	6.897	113.978	108.924	-50.990	11.999	6.945
	(7.633)	(7.634)	(3.817)	(6.20E-04)	(0.016)	(12.272)	(12.273)	(6.136)	(1.66E-04)	(0.020)	(13.766)	(13.766)	(6.883)	(3.61E-05)	(0.016)
7	38.753	32.543	-12.379	13.994	7.784	64.278	58.159	-25.140	13.997	7.878	115.613	109.551	-50.807	13.999	7.937
	(8.174)	(8.177)	(4.087)	(7.31E-04)	(0.018)	(12.289)	(12.289)	(6.144)	(1.85E-04)	(0.018)	(14.019)	(14.020)	(7.010)	(3.84E-05)	(0.019)

Standard deviations in parentheses

4 Empirical example

As an empirical example, we estimate the cost function in the U.S. electric power industry. The data set in Nerlove (1961) includes total cost "TC" (million \$), output "Q" (billion kwh), wage rate "PL" (\$/hr), fuel price "PF" (¢/million Btu), and capital price "PC" (index) for 145 firms in 44 states in the year 1955. He divided the 145 firms into five groups of 29 firms, ordered by output. The total costs of the first 29 firms with lower outputs are widely scattered, hence we removed these firms from the data to avoid a failure of the homoskedasticity assumption that the error variance does not depend on the explanatory variables.

Nerlove (1961) fitted a log-linear cost function (i.e., Cobb-Douglas form). The Cobb-Douglas form is a very convenient parameterization to represent the cost-minimization problem, and coefficients in the log-linear form are elasticities. However as discussed in Hayashi (2000) and Greene (2000), there exists a nonlinear relationship between $\log(TC)$ and $\log(Q)$. Hence we include the polynomial regression term into the log-linear cost function as

$$\log(TC_i) = \alpha_0+\alpha_1 \log(PL_i)+\alpha_2 \log(PF_i)+\alpha_3 \log(PC_i)+\sum_{m=1}^{M} \beta_m \left[\log(Q_i)\right]^m+\varepsilon_i \tag{4.1}$$

where $\varepsilon_i \sim \mathcal{N}(0, \sigma^2)$, $i = 1, 2, \ldots, 116$, and we determine order M in (4.1) via information criterion.

Table 4 shows the results of model selection for the cost function in (4.1). The parameters of priors are set to be $\boldsymbol{b}_0 = \mathbf{0}$, $\kappa_0 = 10^4$, $\nu_0 = \lambda_0 = 0.1$ and we draw $50,000$ samples of parameters to estimate the posterior mean of observed log-likelihood $\widehat{T}_N$. In Table 4, our proposed information criterion (IC_{BL}) selects model 2 ($M = 2$) as the best approximating model, while DIC

selects model 4 ($M = 4$). Since the number of sample is only 116, is relatively small, and our simulation study shows that IC_{BL} performs better than DIC for small sizes, we are inclined to believe that DIC tries to overfit the data relative to IC_{BL}.

Table 4: Model selection for the cost function in the U.S. electric power industry.

Model (M)	IC_{BL}	DIC	$\widehat{T}_N$	$2\widehat{b}_N$	p_D
1	-40.175	-44.214	25.085	9.994	5.955
2	-62.785	-67.822	37.388	11.991	6.954
3	-60.085	-66.099	36.974	13.862	7.848
4	-61.856	-68.134	38.155	14.453	8.175

5 Conclusion

In Bayesian data analysis, DIC (Spiegelhalter et al., 2002) is widely used for the model selection, since this criterion is relatively easy to calculate and applicable to a wide range of statistical models. Spiegelhalter et al. (2002) gave an asymptotic justification of DIC in the case where the number of observations grows with respect to the number of parameters. In the small-sample cases, however, the estimated asymptotic bias of DIC might underestimate the true bias (Burnham, 2002). In this paper, we have focused on the variable selection criterion for the Bayesian linear regression models in a finite sample case, as AIC_C proposed by Sugiura (1978) in frequentist framework,

and examined the performance of our proposed information criterion (IC_{BL}) relative to the DIC for small-sample cases.

In our simulation study, DIC often shows a tendency to overfit the model (see Table 1). On the other hand, our proposed information criterion (IC_{BL}) performs well for small-sample cases ($N = 25, 50, 100$). We also find that the measure of model complexity $2\widehat{b}_N$ is mostly larger than effective number of parameters p_D (see Tables 2 and 3). Hence, the bias correction of DIC is likely to underestimate the model complexity in small-sample cases.

To show the applicability of our proposed information criterion (IC_{BL}) to the empirical study when the sample size is small, we estimate the cost function on the U.S. electric power industry. We find that selected model by DIC (i.e., model 4) has too many parameters relative to the model selected by IC_{BL} (i.e., model 2). Therefore this result shows that DIC tends to overfit the model in small-sample case.

In this paper, we successfully showed that our proposed information criterion (IC_{BL}) outperforms DIC in small-sample cases. Interesting directions for the further research would be to extend our information criterion to the several types of Bayesian linear regression models (e.g., the hierarchical Bayesian linear regression model, Bayesian linear regression model with serially correlated error, and Bayesian linear regression model with structural change).

A Bayesian Linear Regression Model

We consider the linear regression model as follows

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}, \quad \boldsymbol{\varepsilon} \sim \mathcal{N}\left(\boldsymbol{0}, \sigma^2 \boldsymbol{I}_N\right) \tag{A.1}$$

where $\boldsymbol{y}$ is a $N \times 1$ vector and $\boldsymbol{X}$ is a $N \times K$ non-stochastic matrix. The parameter vector $\boldsymbol{\beta}$ is a $K \times 1$ vector and error term $\boldsymbol{\varepsilon}$ follows a N-dimensional multivariate normal distribution $\mathcal{N}\left(\boldsymbol{0}, \sigma^2 \boldsymbol{I}_N\right)$. We assume that prior distribution of $\boldsymbol{\beta}$ is a K-dimensional multivariate normal distribution and that of σ^{-2} is a gamma distribution:

$$\boldsymbol{\beta}|\sigma^{-2} \sim \mathcal{N}\left(\boldsymbol{b}_0, \sigma^2 \boldsymbol{B}_0\right) \tag{A.2}$$

$$\sigma^{-2} \sim \mathcal{G}\left(\frac{\nu_0}{2}, \frac{\lambda_0}{2}\right) \tag{A.3}$$

and joint prior distribution $p(\boldsymbol{\beta}, \sigma^{-2})$ is expressed as

$$\begin{aligned} p(\boldsymbol{\beta}, \sigma^{-2}) &= p(\boldsymbol{\beta}|\sigma^{-2}) p(\sigma^{-2}) \\ &\propto (\sigma^{-2})^{K/2} \exp\left[-\frac{\sigma^{-2}}{2} (\boldsymbol{\beta} - \boldsymbol{b}_0)' \boldsymbol{B}_0^{-1} (\boldsymbol{\beta} - \boldsymbol{b}_0)\right] \\ &\quad \times (\sigma^{-2})^{\frac{\nu_0}{2} - 1} \exp\left[-\frac{\lambda_0 \sigma^{-2}}{2}\right] \end{aligned} \tag{A.4}$$

where $\boldsymbol{b}_0$, $\boldsymbol{B}_0$, $\nu_0/2$, and $\lambda_0/2$ are assumed to be known.

Let us denote $\hat{\boldsymbol{\beta}}_N = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{y}$, then we have

$$\begin{aligned} &(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})'(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}) \\ &= (\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}_N + \boldsymbol{X}\hat{\boldsymbol{\beta}}_N - \boldsymbol{X}\boldsymbol{\beta})'(\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}_N + \boldsymbol{X}\hat{\boldsymbol{\beta}}_N - \boldsymbol{X}\boldsymbol{\beta}) \\ &= (\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}_N)'(\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}_N) + (\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}_N)'\boldsymbol{X}'\boldsymbol{X}(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}_N) \end{aligned} \tag{A.5}$$

where $\boldsymbol{X}'(\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}) = \boldsymbol{0}$. Hence the likelihood function is written as

$$\mathcal{L}\left(\boldsymbol{\beta}, \sigma^{-2}|\boldsymbol{y}, \boldsymbol{X}\right)$$

$$
\begin{aligned}
&\propto (\sigma^{-2})^{N/2} \exp\left[-\frac{\sigma^{-2}}{2}(\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta})'(\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta})\right] \\
&= (\sigma^{-2})^{N/2} \exp\left[-\frac{\sigma^{-2}}{2}\left\{(\boldsymbol{y}-\boldsymbol{X}\hat{\boldsymbol{\beta}}_N)'(\boldsymbol{y}-\boldsymbol{X}\hat{\boldsymbol{\beta}}_N)+(\boldsymbol{\beta}-\hat{\boldsymbol{\beta}}_N)'\boldsymbol{X}'\boldsymbol{X}(\boldsymbol{\beta}-\hat{\boldsymbol{\beta}}_N)\right\}\right].
\end{aligned} \tag{A.6}
$$

From (A.4) and (A.6), posterior distribution is calculated as

$$
p(\boldsymbol{\beta},\sigma^{-2}|\boldsymbol{y},\boldsymbol{X}) \propto \mathcal{L}\left(\boldsymbol{\beta},\sigma^{-2}|\boldsymbol{y},\boldsymbol{X}\right) p(\boldsymbol{\beta},\sigma^{-2}). \tag{A.7}
$$

Therefore we have

$$
\begin{aligned}
&p(\boldsymbol{\beta},\sigma^{-2}|\boldsymbol{y},\boldsymbol{X}) \\
&\propto (\sigma^{-2})^{N/2} \exp\left[-\frac{\sigma^{-2}}{2}\left\{(\boldsymbol{y}-\boldsymbol{X}\hat{\boldsymbol{\beta}}_N)'(\boldsymbol{y}-\boldsymbol{X}\hat{\boldsymbol{\beta}}_N)+(\boldsymbol{\beta}-\hat{\boldsymbol{\beta}}_N)'\boldsymbol{X}'\boldsymbol{X}(\boldsymbol{\beta}-\hat{\boldsymbol{\beta}}_N)\right\}\right] \\
&\quad \times (\sigma^{-2})^{K/2} \exp\left[-\frac{\sigma^{-2}}{2}(\boldsymbol{\beta}-\boldsymbol{b}_0)'\boldsymbol{B}_0^{-1}(\boldsymbol{\beta}-\boldsymbol{b}_0)\right] \\
&\quad \times (\sigma^{-2})^{\frac{\nu_0}{2}-1} \exp\left[-\frac{\lambda_0\sigma^{-2}}{2}\right].
\end{aligned} \tag{A.8}
$$

Lemma A.1. [2]

$$
\begin{aligned}
&(\boldsymbol{\beta}-\hat{\boldsymbol{\beta}}_N)'\boldsymbol{X}'\boldsymbol{X}(\boldsymbol{\beta}-\hat{\boldsymbol{\beta}}_N)+(\boldsymbol{\beta}-\boldsymbol{b}_0)'\boldsymbol{B}_0^{-1}(\boldsymbol{\beta}-\boldsymbol{b}_0) \\
&= (\boldsymbol{\beta}-\boldsymbol{b}_1)'\boldsymbol{B}_1^{-1}(\boldsymbol{\beta}-\boldsymbol{b}_1)+\left(\boldsymbol{b}_0-\hat{\boldsymbol{\beta}}_N\right)'\left((\boldsymbol{X}'\boldsymbol{X})^{-1}+\boldsymbol{B}_0\right)^{-1}\left(\boldsymbol{b}_0-\hat{\boldsymbol{\beta}}_N\right)
\end{aligned} \tag{A.9}
$$

where

$$
\boldsymbol{b}_1 = \boldsymbol{B}_1\left(\boldsymbol{X}'\boldsymbol{y}+\boldsymbol{B}_0^{-1}\boldsymbol{b}_0\right) \tag{A.10}
$$

$$
\boldsymbol{B}_1 = \left(\boldsymbol{X}'\boldsymbol{X}+\boldsymbol{B}_0^{-1}\right)^{-1}. \tag{A.11}
$$

[2]The proof is in Appendix B.

From Lemma A.1, the joint posterior distribution $p(\boldsymbol{\beta}, \sigma^{-2}|\boldsymbol{y}, \boldsymbol{X})$ is obtained by

$$p(\boldsymbol{\beta}, \sigma^{-2}|\boldsymbol{y}, \boldsymbol{X}) \propto \exp\left[-\frac{\sigma^{-2}}{2}\left(\boldsymbol{\beta}-\boldsymbol{b}_1\right)'\boldsymbol{B}_1^{-1}\left(\boldsymbol{\beta}-\boldsymbol{b}_1\right)\right] \times \left(\sigma^{-2}\right)^{\frac{\nu_1}{2}-1}\exp\left[-\frac{\lambda_1\sigma^{-2}}{2}\right] \tag{A.12}$$

where ν_1 and λ_1 are defined as follows:

$$\nu_1 = \nu_0 + N + K \tag{A.13}$$

$$\lambda_1 = \lambda_0 + (\boldsymbol{y}-\boldsymbol{X}\hat{\boldsymbol{\beta}}_N)'(\boldsymbol{y}-\boldsymbol{X}\hat{\boldsymbol{\beta}}_N) + \left(\boldsymbol{b}_0-\hat{\boldsymbol{\beta}}_N\right)'\left(\left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}+\boldsymbol{B}_0\right)^{-1}\left(\boldsymbol{b}_0-\hat{\boldsymbol{\beta}}_N\right). \tag{A.14}$$

From (A.12), posterior distributions [3] of parameters $\boldsymbol{\beta}$ and σ^{-2} are expressed as

$$\boldsymbol{\beta}|\sigma^{-2}, \boldsymbol{y}, \boldsymbol{X} \sim \mathcal{N}\left(\boldsymbol{b}_1, \sigma^2\boldsymbol{B}_1\right) \tag{A.15}$$

$$\sigma^{-2}|\boldsymbol{y}, \boldsymbol{X} \sim \mathcal{G}\left(\frac{\nu_1}{2}, \frac{\lambda_1}{2}\right). \tag{A.16}$$

The marginal posterior distribution $p(\boldsymbol{\beta}|\boldsymbol{y}, \boldsymbol{X})$ is derived as

$$\begin{aligned} p(\boldsymbol{\beta}|\boldsymbol{y}, \boldsymbol{X}) &= \int_0^\infty p(\boldsymbol{\beta}, \sigma^{-2}|\boldsymbol{y}, \boldsymbol{X})d\sigma^2 \\ &\propto \int_0^\infty \exp\left[-\frac{\left(\boldsymbol{\beta}-\boldsymbol{b}_1\right)'\boldsymbol{B}_1^{-1}\left(\boldsymbol{\beta}-\boldsymbol{b}_1\right)}{2\sigma^2}\right](\sigma^2)^{-(\frac{\nu_1}{2}-1)}\exp\left[-\frac{\lambda_1}{2\sigma^2}\right]d\sigma^2 \\ &= \int_0^\infty (\sigma^2)^{-(\frac{\nu_1}{2}-1)}\exp\left[-\frac{\lambda_1+\left(\boldsymbol{\beta}-\boldsymbol{b}_1\right)'\boldsymbol{B}_1^{-1}\left(\boldsymbol{\beta}-\boldsymbol{b}_1\right)}{2\sigma^2}\right]d\sigma^2. \end{aligned} \tag{A.17}$$

[3] We notice that posterior mean $\boldsymbol{b}_1$ is rewritten as $\boldsymbol{b}_1 = \boldsymbol{B}_1(\boldsymbol{X}'\boldsymbol{X}\hat{\boldsymbol{\beta}}_N + \boldsymbol{B}_0^{-1}\boldsymbol{b}_0)$ by using MLE $\hat{\boldsymbol{\beta}}_N = \left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\boldsymbol{X}'\boldsymbol{y}$. Then posterior mean $\boldsymbol{b}_1$ is a weighted average of $\hat{\boldsymbol{\beta}}_N$ and prior mean $\boldsymbol{b}_0$ with weights inversely proportional to the variance covariance matrices, $\sigma^{-2}(\boldsymbol{X}'\boldsymbol{X})$ and $\sigma^{-2}\boldsymbol{B}_0^{-1}$, of $\hat{\boldsymbol{\beta}}_N$ and $\boldsymbol{b}_0$.

Integration in (A.17) is obtained by using the gamma function [4]:

$$\begin{aligned}
&\left[\frac{\lambda_1+(\boldsymbol{\beta}-\boldsymbol{b}_1)'\,\boldsymbol{B}_1^{-1}\,(\boldsymbol{\beta}-\boldsymbol{b}_1)}{2}\right]^{-(\frac{\nu_1}{2}-1)}\\
&\times\int_0^\infty\left[\frac{\lambda_1+(\boldsymbol{\beta}-\boldsymbol{b}_1)'\,\boldsymbol{B}_1^{-1}\,(\boldsymbol{\beta}-\boldsymbol{b}_1)}{2\sigma^2}\right]^{\frac{\nu_1}{2}-1}\exp\left[-\frac{\lambda_1+(\boldsymbol{\beta}-\boldsymbol{b}_1)'\,\boldsymbol{B}_1^{-1}\,(\boldsymbol{\beta}-\boldsymbol{b}_1)}{2\sigma^2}\right]d\sigma^2\\
&=\Gamma\left(\frac{\nu_1}{2}\right)\left[\frac{\lambda_1+(\boldsymbol{\beta}-\boldsymbol{b}_1)'\,\boldsymbol{B}_1^{-1}\,(\boldsymbol{\beta}-\boldsymbol{b}_1)}{2}\right]^{-\left(\frac{\nu_1}{2}-1\right)}\\
&\propto\Gamma\left(\frac{\nu_0+N+K}{2}\right)\left[1+\frac{1}{\nu_0+N}\,(\boldsymbol{\beta}-\boldsymbol{b}_1)'\left(\frac{\lambda_1}{\nu_0+N}\boldsymbol{B}_1\right)^{-1}(\boldsymbol{\beta}-\boldsymbol{b}_1)\right]^{-\frac{\nu_0+N+K}{2}}
\end{aligned}\tag{A.18}$$

then we notice that (A.18) is proportional to the K-dimensional multivariate t-distribution [5]. Hence we have

$$\boldsymbol{\beta}|\boldsymbol{y},\boldsymbol{X}\sim\mathcal{T}_K\left(\boldsymbol{b}_1,\ \frac{\lambda_1}{\nu_0+N}\boldsymbol{B}_1,\ \nu_0+N\right).\tag{A.19}$$

Let us denote the predictive value of $\boldsymbol{y}$ as $\boldsymbol{y}_0$. Then predictive density $p(\boldsymbol{y}_0|\boldsymbol{y},\boldsymbol{X})$ is derived as

$$\begin{aligned}
p(\boldsymbol{y}_0|\boldsymbol{y},\boldsymbol{X})&=\int_0^\infty\int_{\mathbb{R}^K}p(\boldsymbol{y}_0|\boldsymbol{X},\boldsymbol{\beta},\sigma^{-2})p(\boldsymbol{\beta},\sigma^{-2}|\boldsymbol{y},\boldsymbol{X})d\boldsymbol{\beta}d\sigma^2\\
&=\int_0^\infty\left[\int_{\mathbb{R}^K}p(\boldsymbol{y}_0|\boldsymbol{X},\boldsymbol{\beta},\sigma^{-2})p(\boldsymbol{\beta}|\sigma^{-2},\boldsymbol{y},\boldsymbol{X})d\boldsymbol{\beta}\right]\\
&\qquad\times p(\sigma^{-2}|\boldsymbol{y},\boldsymbol{X})d\sigma^2.
\end{aligned}\tag{A.20}$$

[4]The gamma function is denoted by

$$\Gamma(m)=\int_0^\infty x^{m-1}e^{-x}dm$$

where $m>0$ and $\Gamma(m+1)=m\Gamma(m)$.

[5]The K-dimensional multivariate t-distribution $\mathcal{T}_K(\boldsymbol{\mu},\boldsymbol{\Sigma},\nu)$ is denoted by

$$p(\boldsymbol{x}|\boldsymbol{\mu},\boldsymbol{\Sigma},\nu)=\frac{\Gamma\left(\frac{\nu+K}{2}\right)}{\Gamma\left(\frac{\nu}{2}\right)\nu^{\frac{K}{2}}\pi^{\frac{K}{2}}}\boldsymbol{\Sigma}^{-\frac{1}{2}}\left[1+\frac{1}{\nu}(\boldsymbol{x}-\boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})\right]^{-\frac{\nu+K}{2}}$$

where $\boldsymbol{x}$ is a K-dimensional random variable, and $\mathbf{E}(\boldsymbol{x})=\boldsymbol{\mu}$ and $\mathbf{Var}(\boldsymbol{x})=\frac{\nu}{\nu-2}\boldsymbol{\Sigma}$.

First, we calculate the integration with respect to $\boldsymbol{\beta}$:

$$\int_{\mathbb{R}^K} p(\boldsymbol{y}_0|\boldsymbol{X},\boldsymbol{\beta},\sigma^{-2})p(\boldsymbol{\beta}|\sigma^{-2},\boldsymbol{y},\boldsymbol{X})d\boldsymbol{\beta}. \tag{A.21}$$

where

$$\begin{aligned}
&p(\boldsymbol{y}_0|\boldsymbol{X},\boldsymbol{\beta},\sigma^{-2})p(\boldsymbol{\beta}|\sigma^{-2},\boldsymbol{y},\boldsymbol{X}) \\
&\propto (\sigma^{-2})^{N/2}\exp\left[-\frac{\sigma^{-2}}{2}(\boldsymbol{y}_0-\boldsymbol{X}\boldsymbol{\beta})'(\boldsymbol{y}_0-\boldsymbol{X}\boldsymbol{\beta})\right] \\
&\qquad\times(\sigma^{-2})^{K/2}\exp\left[-\frac{\sigma^{-2}}{2}(\boldsymbol{\beta}-\boldsymbol{b}_1)'\boldsymbol{B}_1^{-1}(\boldsymbol{\beta}-\boldsymbol{b}_1)\right] \\
&=(\sigma^{-2})^{(N+K)/2}\exp\left[-\frac{\sigma^{-2}}{2}(\boldsymbol{y}_0-\boldsymbol{X}\boldsymbol{\beta})'(\boldsymbol{y}_0-\boldsymbol{X}\boldsymbol{\beta})-\frac{\sigma^{-2}}{2}(\boldsymbol{\beta}-\boldsymbol{b}_1)'\boldsymbol{B}_1^{-1}(\boldsymbol{\beta}-\boldsymbol{b}_1)\right].
\end{aligned} \tag{A.22}$$

Lemma A.2.

$$\begin{aligned}
&(\boldsymbol{y}_0-\boldsymbol{X}\boldsymbol{\beta})'(\boldsymbol{y}_0-\boldsymbol{X}\boldsymbol{\beta})+(\boldsymbol{\beta}-\boldsymbol{b}_1)'\boldsymbol{B}_1^{-1}(\boldsymbol{\beta}-\boldsymbol{b}_1) \\
&=(\boldsymbol{y}_0-\boldsymbol{X}\boldsymbol{b}_1)'\boldsymbol{\Sigma}_0^{-1}(\boldsymbol{y}_0-\boldsymbol{X}\boldsymbol{b}_1)+(\boldsymbol{\beta}-\boldsymbol{b}_2)'\boldsymbol{B}_2^{-1}(\boldsymbol{\beta}-\boldsymbol{b}_2)
\end{aligned} \tag{A.23}$$

where

$$\boldsymbol{b}_2=\boldsymbol{B}_2\left(\boldsymbol{X}'\boldsymbol{y}_0+\boldsymbol{B}_1^{-1}\boldsymbol{b}_1\right) \tag{A.24}$$

$$\boldsymbol{B}_2=\left(\boldsymbol{X}'\boldsymbol{X}+\boldsymbol{B}_1^{-1}\right)^{-1} \tag{A.25}$$

$$\boldsymbol{\Sigma}_0=\boldsymbol{I}_N+\boldsymbol{X}\boldsymbol{B}_1\boldsymbol{X}'. \tag{A.26}$$

From Lemma A.2, we have the integration with respect to $\boldsymbol{\beta}$ in (A.21) as

$$\begin{aligned}
&p(\boldsymbol{y}_0|\sigma^{-2},\boldsymbol{y},\boldsymbol{X}) \\
&=\int_{\mathbb{R}^K} p(\boldsymbol{y}_0|\boldsymbol{X},\boldsymbol{\beta},\sigma^{-2})p(\boldsymbol{\beta}|\sigma^{-2},\boldsymbol{y},\boldsymbol{X})d\boldsymbol{\beta} \\
&\propto\int_{\mathbb{R}^K}(\sigma^{-2})^{(N+K)/2}\exp\left[-\frac{\sigma^{-2}}{2}(\boldsymbol{y}_0-\boldsymbol{X}\boldsymbol{b}_1)'\boldsymbol{\Sigma}_0^{-1}(\boldsymbol{y}_0-\boldsymbol{X}\boldsymbol{b}_1)\right. \\
&\qquad\left.-\frac{\sigma^{-2}}{2}(\boldsymbol{\beta}-\boldsymbol{b}_2)'\boldsymbol{B}_2^{-1}(\boldsymbol{\beta}-\boldsymbol{b}_2)\right]d\boldsymbol{\beta}
\end{aligned}$$

$$\propto (\sigma^{-2})^{N/2} \exp\left[-\frac{\sigma^{-2}}{2}(\boldsymbol{y}_0 - \boldsymbol{X}\boldsymbol{b}_1)'\boldsymbol{\Sigma}_0^{-1}(\boldsymbol{y}_0 - \boldsymbol{X}\boldsymbol{b}_1)\right]. \tag{A.27}$$

Hence the predictive density conditional on σ^{-2} is N-dimensional multivariate normal distribution:

$$\boldsymbol{y}_0|\sigma^{-2}, \boldsymbol{y}, \boldsymbol{X} \sim \mathcal{N}\left(\boldsymbol{X}\boldsymbol{b}_1, \sigma^2\boldsymbol{\Sigma}_0\right). \tag{A.28}$$

From (A.16) and (A.27), integration with respect to the posterior distribution of σ^2 for $p(\boldsymbol{y}_0|\sigma^{-2}, \boldsymbol{y}, \boldsymbol{X})$ is

$$\begin{aligned}
&\int_0^\infty p(\boldsymbol{y}_0|\sigma^{-2}, \boldsymbol{y}, \boldsymbol{X}) p(\sigma^{-2}|\boldsymbol{y}, \boldsymbol{X}) d\sigma^2 \\
&\propto \int_0^\infty (\sigma^{-2})^{N/2} \exp\left[-\frac{\sigma^{-2}}{2}(\boldsymbol{y}_0 - \boldsymbol{X}\boldsymbol{b}_1)'\boldsymbol{\Sigma}_0^{-1}(\boldsymbol{y}_0 - \boldsymbol{X}\boldsymbol{b}_1)\right] \\
&\qquad \times (\sigma^{-2})^{\frac{\nu_1}{2}-1} \exp\left[-\frac{\lambda_1\sigma^{-2}}{2}\right] d\sigma^2 \\
&= \int_0^\infty (\sigma^{-2})^{\frac{\nu_1+N}{2}-1} \exp\left[-\frac{\lambda_1 + (\boldsymbol{y}_0 - \boldsymbol{X}\boldsymbol{b}_1)'\boldsymbol{\Sigma}_0^{-1}(\boldsymbol{y}_0 - \boldsymbol{X}\boldsymbol{b}_1)}{2\sigma^2}\right] d\sigma^2 \\
&= \left[\frac{\lambda_1 + (\boldsymbol{y}_0 - \boldsymbol{X}\boldsymbol{b}_1)'\boldsymbol{\Sigma}_0^{-1}(\boldsymbol{y}_0 - \boldsymbol{X}\boldsymbol{b}_1)}{2}\right]^{-\left(\frac{\nu_1+N}{2}-1\right)} \\
&\qquad \times \int_0^\infty \left[\frac{\lambda_1 + (\boldsymbol{y}_0 - \boldsymbol{X}\boldsymbol{b}_1)'\boldsymbol{\Sigma}_0^{-1}(\boldsymbol{y}_0 - \boldsymbol{X}\boldsymbol{b}_1)}{2\sigma^2}\right]^{\frac{\nu_1+N}{2}-1} \\
&\qquad\qquad \times \exp\left[-\frac{\lambda_1 + (\boldsymbol{y}_0 - \boldsymbol{X}\boldsymbol{b}_1)'\boldsymbol{\Sigma}_0^{-1}(\boldsymbol{y}_0 - \boldsymbol{X}\boldsymbol{b}_1)}{2\sigma^2}\right] d\sigma^2 \\
&= \Gamma\left(\frac{\nu_1+N}{2}\right)\left[\frac{\lambda_1 + (\boldsymbol{y}_0 - \boldsymbol{X}\boldsymbol{b}_1)'\boldsymbol{\Sigma}_0^{-1}(\boldsymbol{y}_0 - \boldsymbol{X}\boldsymbol{b}_1)}{2}\right]^{-\left(\frac{\nu_1+N}{2}-1\right)} \\
&\propto \Gamma\left(\frac{\nu_1+N}{2}\right)\left[1 + \left(\frac{1}{\nu_1}\right)(\boldsymbol{y}_0 - \boldsymbol{X}\boldsymbol{b}_1)'\left(\frac{\lambda_1}{\nu_1}\boldsymbol{\Sigma}_0\right)^{-1}(\boldsymbol{y}_0 - \boldsymbol{X}\boldsymbol{b}_1)\right]^{-\frac{\nu_1+N}{2}}.
\end{aligned} \tag{A.29}$$

Hence we notice that (A.29) is proportional to N-dimensional multivariate t-distribution:

$$\boldsymbol{y}_0|\boldsymbol{y}, \boldsymbol{X} \sim \mathcal{T}_N\left(\boldsymbol{X}\boldsymbol{b}_1, \boldsymbol{\Sigma}_0, \nu_1\right). \tag{A.30}$$

B Proof of Lemma A.1

Recall (A.9):

$$
\begin{aligned}
&(\boldsymbol{\beta}-\hat{\boldsymbol{\beta}}_N)'\boldsymbol{X}'\boldsymbol{X}(\boldsymbol{\beta}-\hat{\boldsymbol{\beta}}_N)+(\boldsymbol{\beta}-\boldsymbol{b}_0)'\boldsymbol{B}_0^{-1}(\boldsymbol{\beta}-\boldsymbol{b}_0)\\
&=(\boldsymbol{\beta}-\boldsymbol{b}_1)'\boldsymbol{B}_1^{-1}(\boldsymbol{\beta}-\boldsymbol{b}_1)+\left(\boldsymbol{b}_0-\hat{\boldsymbol{\beta}}_N\right)'\left((\boldsymbol{X}'\boldsymbol{X})^{-1}+\boldsymbol{B}_0\right)^{-1}\left(\boldsymbol{b}_0-\hat{\boldsymbol{\beta}}_N\right).
\end{aligned}
$$

where $\boldsymbol{B}_1=\left(\boldsymbol{X}'\boldsymbol{X}+\boldsymbol{B}_0^{-1}\right)^{-1}$ and $\boldsymbol{b}_1=\boldsymbol{B}_1\left(\boldsymbol{X}'\boldsymbol{y}+\boldsymbol{B}_0^{-1}\boldsymbol{b}_0\right)$.

Proof. To prove (A.9), we show the following equation:

$$
\begin{aligned}
&(\boldsymbol{\theta}-\boldsymbol{c})'\boldsymbol{A}(\boldsymbol{\theta}-\boldsymbol{c})+(\boldsymbol{d}-\boldsymbol{\theta})'\boldsymbol{B}(\boldsymbol{d}-\boldsymbol{\theta})\\
&=\left[\boldsymbol{\theta}-(\boldsymbol{A}+\boldsymbol{B})^{-1}(\boldsymbol{A}\boldsymbol{c}+\boldsymbol{B}\boldsymbol{d})\right]'(\boldsymbol{A}+\boldsymbol{B})\left[\boldsymbol{\theta}-(\boldsymbol{A}+\boldsymbol{B})^{-1}(\boldsymbol{A}\boldsymbol{c}+\boldsymbol{B}\boldsymbol{d})\right]\\
&\quad+(\boldsymbol{c}-\boldsymbol{d})'\left(\boldsymbol{B}^{-1}+\boldsymbol{A}^{-1}\right)(\boldsymbol{c}-\boldsymbol{d})
\end{aligned}
\tag{B.1}
$$

where $\boldsymbol{A}$ and $\boldsymbol{B}$ are $K\times K$ symmetric and invertible matrices, and $\boldsymbol{\theta}$, $\boldsymbol{c}$ and $\boldsymbol{d}$ are $K\times 1$ vectors.

Taking a trace of matrices for both sides, we have

$$
\begin{aligned}
&(\boldsymbol{\theta}-\boldsymbol{c})'\boldsymbol{A}(\boldsymbol{\theta}-\boldsymbol{c})+(\boldsymbol{d}-\boldsymbol{\theta})'\boldsymbol{B}(\boldsymbol{d}-\boldsymbol{\theta})\\
&=\operatorname{tr}\left\{\boldsymbol{A}(\boldsymbol{\theta}-\boldsymbol{c})(\boldsymbol{\theta}-\boldsymbol{c})'+\boldsymbol{B}(\boldsymbol{d}-\boldsymbol{\theta})(\boldsymbol{d}-\boldsymbol{\theta})'\right\}\\
&=\operatorname{tr}\left\{(\boldsymbol{A}+\boldsymbol{B})\boldsymbol{\theta}\boldsymbol{\theta}'-2(\boldsymbol{A}\boldsymbol{c}+\boldsymbol{B}\boldsymbol{d})\boldsymbol{\theta}'+\boldsymbol{A}\boldsymbol{c}\boldsymbol{c}'+\boldsymbol{B}\boldsymbol{d}\boldsymbol{d}'\right\}\\
&=\operatorname{tr}\left\{(\boldsymbol{A}+\boldsymbol{B})\left[\boldsymbol{\theta}\boldsymbol{\theta}'-2(\boldsymbol{A}+\boldsymbol{B})^{-1}(\boldsymbol{A}\boldsymbol{c}+\boldsymbol{B}\boldsymbol{d})\boldsymbol{\theta}'\right]+\boldsymbol{A}\boldsymbol{c}\boldsymbol{c}'+\boldsymbol{B}\boldsymbol{d}\boldsymbol{d}'\right\}\\
&=\operatorname{tr}\Big\{(\boldsymbol{A}+\boldsymbol{B})\left[\boldsymbol{\theta}-(\boldsymbol{A}+\boldsymbol{B})^{-1}(\boldsymbol{A}\boldsymbol{c}+\boldsymbol{B}\boldsymbol{d})\right]\left[\boldsymbol{\theta}-(\boldsymbol{A}+\boldsymbol{B})^{-1}(\boldsymbol{A}\boldsymbol{c}+\boldsymbol{B}\boldsymbol{d})\right]'\\
&\qquad-(\boldsymbol{A}+\boldsymbol{B})\left[(\boldsymbol{A}+\boldsymbol{B})^{-1}(\boldsymbol{A}\boldsymbol{c}+\boldsymbol{B}\boldsymbol{d})\right]\left[(\boldsymbol{A}+\boldsymbol{B})^{-1}(\boldsymbol{A}\boldsymbol{c}+\boldsymbol{B}\boldsymbol{d})\right]'\\
&\qquad+\boldsymbol{A}\boldsymbol{c}\boldsymbol{c}'+\boldsymbol{B}\boldsymbol{d}\boldsymbol{d}'\Big\}\\
&=\left[\boldsymbol{\theta}-(\boldsymbol{A}+\boldsymbol{B})^{-1}(\boldsymbol{A}\boldsymbol{c}+\boldsymbol{B}\boldsymbol{d})\right]'(\boldsymbol{A}+\boldsymbol{B})\left[\boldsymbol{\theta}-(\boldsymbol{A}+\boldsymbol{B})^{-1}(\boldsymbol{A}\boldsymbol{c}+\boldsymbol{B}\boldsymbol{d})\right]\\
&\qquad-\left[(\boldsymbol{A}+\boldsymbol{B})^{-1}(\boldsymbol{A}\boldsymbol{c}+\boldsymbol{B}\boldsymbol{d})\right]'(\boldsymbol{A}+\boldsymbol{B})\left[(\boldsymbol{A}+\boldsymbol{B})^{-1}(\boldsymbol{A}\boldsymbol{c}+\boldsymbol{B}\boldsymbol{d})\right]\\
&\qquad+\boldsymbol{c}'\boldsymbol{A}\boldsymbol{c}+\boldsymbol{d}'\boldsymbol{B}\boldsymbol{d}.
\end{aligned}
\tag{B.2}
$$

Next last three terms of the above expression in (B.2) can be rewritten as

$$
\begin{aligned}
&- \left[(\boldsymbol{A}+\boldsymbol{B})^{-1}(\boldsymbol{A}\boldsymbol{c}+\boldsymbol{B}\boldsymbol{d})\right]'(\boldsymbol{A}+\boldsymbol{B})\left[(\boldsymbol{A}+\boldsymbol{B})^{-1}(\boldsymbol{A}\boldsymbol{c}+\boldsymbol{B}\boldsymbol{d})\right] \\
&\quad + \boldsymbol{c}'\boldsymbol{A}\boldsymbol{c}+\boldsymbol{d}'\boldsymbol{B}\boldsymbol{d} \\
&= -(\boldsymbol{A}\boldsymbol{c}+\boldsymbol{B}\boldsymbol{d})'(\boldsymbol{A}+\boldsymbol{B})^{-1}(\boldsymbol{A}\boldsymbol{c}+\boldsymbol{B}\boldsymbol{d}) \\
&\quad + \boldsymbol{c}'\boldsymbol{A}\boldsymbol{c}+\boldsymbol{d}'\boldsymbol{B}\boldsymbol{d} \\
&= -\left(\boldsymbol{c}+\boldsymbol{A}^{-1}\boldsymbol{B}\boldsymbol{d}\right)'\boldsymbol{A}(\boldsymbol{A}+\boldsymbol{B})^{-1}\boldsymbol{B}\left(\boldsymbol{B}^{-1}\boldsymbol{A}\boldsymbol{c}+\boldsymbol{d}\right) \\
&\quad + \boldsymbol{c}'\boldsymbol{A}\boldsymbol{c}+\boldsymbol{d}'\boldsymbol{B}\boldsymbol{d} \\
&= -\left(\boldsymbol{c}+\boldsymbol{A}^{-1}\boldsymbol{B}\boldsymbol{d}\right)'\left(\boldsymbol{B}^{-1}+\boldsymbol{A}^{-1}\right)^{-1}\left(\boldsymbol{B}^{-1}\boldsymbol{A}\boldsymbol{c}+\boldsymbol{d}\right) \\
&\quad + \boldsymbol{c}'\boldsymbol{A}\boldsymbol{c}+\boldsymbol{d}'\boldsymbol{B}\boldsymbol{d} \\
&= \left(\boldsymbol{c}-\boldsymbol{d}+\boldsymbol{d}+\boldsymbol{A}^{-1}\boldsymbol{B}\boldsymbol{d}\right)'\left(\boldsymbol{B}^{-1}+\boldsymbol{A}^{-1}\right)^{-1}\left(\boldsymbol{c}-\boldsymbol{d}-\boldsymbol{B}^{-1}\boldsymbol{A}\boldsymbol{c}-\boldsymbol{c}\right) \\
&\quad + \boldsymbol{c}'\boldsymbol{A}\boldsymbol{c}+\boldsymbol{d}'\boldsymbol{B}\boldsymbol{d} \\
&= (\boldsymbol{c}-\boldsymbol{d})'\left(\boldsymbol{B}^{-1}+\boldsymbol{A}^{-1}\right)^{-1}(\boldsymbol{c}-\boldsymbol{d}) \\
&\quad + (\boldsymbol{c}-\boldsymbol{d})'\left(\boldsymbol{B}^{-1}+\boldsymbol{A}^{-1}\right)^{-1}\left(-\boldsymbol{B}^{-1}\boldsymbol{A}\boldsymbol{c}-\boldsymbol{c}\right) \\
&\quad + \left(\boldsymbol{d}+\boldsymbol{A}^{-1}\boldsymbol{B}\boldsymbol{d}\right)'\left(\boldsymbol{B}^{-1}+\boldsymbol{A}^{-1}\right)^{-1}(\boldsymbol{c}-\boldsymbol{d}) \\
&\quad + \left(\boldsymbol{d}+\boldsymbol{A}^{-1}\boldsymbol{B}\boldsymbol{d}\right)'\left(\boldsymbol{B}^{-1}+\boldsymbol{A}^{-1}\right)^{-1}\left(-\boldsymbol{B}^{-1}\boldsymbol{A}\boldsymbol{c}-\boldsymbol{c}\right) \\
&\quad + \boldsymbol{c}'\boldsymbol{A}\boldsymbol{c}+\boldsymbol{d}'\boldsymbol{B}\boldsymbol{d} \\
&= (\boldsymbol{c}-\boldsymbol{d})'\left(\boldsymbol{B}^{-1}+\boldsymbol{A}^{-1}\right)^{-1}(\boldsymbol{c}-\boldsymbol{d})+\boldsymbol{R}_0. \qquad \text{(B.3)}
\end{aligned}
$$

where

$$
\begin{aligned}
\boldsymbol{R}_0 = &(\boldsymbol{c}-\boldsymbol{d})'\left(\boldsymbol{B}^{-1}+\boldsymbol{A}^{-1}\right)^{-1}\left(-\boldsymbol{B}^{-1}\boldsymbol{A}\boldsymbol{c}-\boldsymbol{c}\right) \\
&+ \left(\boldsymbol{d}+\boldsymbol{A}^{-1}\boldsymbol{B}\boldsymbol{d}\right)'\left(\boldsymbol{B}^{-1}+\boldsymbol{A}^{-1}\right)^{-1}(\boldsymbol{c}-\boldsymbol{d}) \\
&+ \left(\boldsymbol{d}+\boldsymbol{A}^{-1}\boldsymbol{B}\boldsymbol{d}\right)'\left(\boldsymbol{B}^{-1}+\boldsymbol{A}^{-1}\right)^{-1}\left(-\boldsymbol{B}^{-1}\boldsymbol{A}\boldsymbol{c}-\boldsymbol{c}\right) \\
&+ \boldsymbol{c}'\boldsymbol{A}\boldsymbol{c}+\boldsymbol{d}'\boldsymbol{B}\boldsymbol{d}. \qquad \text{(B.4)}
\end{aligned}
$$

Finally we show that the remainder term in (B.3) become zero (i.e., $\boldsymbol{R}_0 = 0$).

Expanding the remainder term (B.4), we have

$$\begin{aligned}
\boldsymbol{R}_0 = & -\boldsymbol{c}'\left(\boldsymbol{I}_K + \boldsymbol{B}\boldsymbol{A}^{-1}\right)^{-1}\boldsymbol{A}\boldsymbol{c} - \boldsymbol{c}'\left(\boldsymbol{B}^{-1} + \boldsymbol{A}^{-1}\right)^{-1}\boldsymbol{c} \\
& + \boldsymbol{d}'\left(\boldsymbol{I}_K + \boldsymbol{B}\boldsymbol{A}^{-1}\right)^{-1}\boldsymbol{A}\boldsymbol{c} + \boldsymbol{d}'\left(\boldsymbol{B}^{-1} + \boldsymbol{A}^{-1}\right)^{-1}\boldsymbol{c} \\
& + \boldsymbol{d}'\left(\boldsymbol{B}^{-1} + \boldsymbol{A}^{-1}\right)^{-1}\boldsymbol{c} - \boldsymbol{d}'\left(\boldsymbol{B}^{-1} + \boldsymbol{A}^{-1}\right)^{-1}\boldsymbol{d} \\
& + \boldsymbol{d}'\boldsymbol{B}\left(\boldsymbol{I}_K + \boldsymbol{B}^{-1}\boldsymbol{A}\right)^{-1}\boldsymbol{c} - \boldsymbol{d}'\boldsymbol{B}\left(\boldsymbol{I}_K + \boldsymbol{B}^{-1}\boldsymbol{A}\right)^{-1}\boldsymbol{d} \\
& - \boldsymbol{d}'\left(\boldsymbol{I}_K + \boldsymbol{B}\boldsymbol{A}^{-1}\right)^{-1}\boldsymbol{A}\boldsymbol{c} - \boldsymbol{d}'\left(\boldsymbol{B}^{-1} + \boldsymbol{A}^{-1}\right)^{-1}\boldsymbol{c} \\
& - \boldsymbol{d}'\left(\boldsymbol{B}^{-1} + \boldsymbol{A}^{-1}\right)^{-1}\boldsymbol{c} - \boldsymbol{d}'\boldsymbol{B}\left(\boldsymbol{I}_K + \boldsymbol{B}^{-1}\boldsymbol{A}\right)^{-1}\boldsymbol{c} \\
& + \boldsymbol{c}'\boldsymbol{A}\boldsymbol{c} + \boldsymbol{d}'\boldsymbol{B}\boldsymbol{d} \\
= & \boldsymbol{c}'\boldsymbol{A}\boldsymbol{c} - \boldsymbol{c}'\left(\boldsymbol{I}_K + \boldsymbol{B}\boldsymbol{A}^{-1}\right)^{-1}\boldsymbol{A}\boldsymbol{c} - \boldsymbol{c}'\left(\boldsymbol{B}^{-1} + \boldsymbol{A}^{-1}\right)^{-1}\boldsymbol{c} \\
& + \boldsymbol{d}'\boldsymbol{B}\boldsymbol{d} - \boldsymbol{d}'\boldsymbol{B}\left(\boldsymbol{I}_K + \boldsymbol{B}^{-1}\boldsymbol{A}\right)^{-1}\boldsymbol{d} - \boldsymbol{d}'\left(\boldsymbol{B}^{-1} + \boldsymbol{A}^{-1}\right)^{-1}\boldsymbol{d} \\
= & \boldsymbol{c}'\left[\boldsymbol{A} - \left(\boldsymbol{I}_K + \boldsymbol{B}\boldsymbol{A}^{-1}\right)^{-1}\boldsymbol{A}\right]\boldsymbol{c} - \boldsymbol{c}'\left(\boldsymbol{B}^{-1} + \boldsymbol{A}^{-1}\right)^{-1}\boldsymbol{c} \\
& + \boldsymbol{d}'\left[\boldsymbol{B} - \boldsymbol{B}\left(\boldsymbol{I}_K + \boldsymbol{B}^{-1}\boldsymbol{A}\right)^{-1}\right]\boldsymbol{d} - \boldsymbol{d}'\left(\boldsymbol{B}^{-1} + \boldsymbol{A}^{-1}\right)^{-1}\boldsymbol{d}. \qquad \text{(B.5)}
\end{aligned}$$

To show that remainder term $\boldsymbol{R}_0 = 0$, we need to prove that

$$\left(\boldsymbol{B}^{-1} + \boldsymbol{A}^{-1}\right)^{-1} - \left[\boldsymbol{A} - \left(\boldsymbol{I}_K + \boldsymbol{B}\boldsymbol{A}^{-1}\right)^{-1}\boldsymbol{A}\right] = \boldsymbol{0}_{K\times K}, \qquad \text{(B.6)}$$

$$\left(\boldsymbol{B}^{-1} + \boldsymbol{A}^{-1}\right)^{-1} - \left[\boldsymbol{B} - \boldsymbol{B}\left(\boldsymbol{I}_K + \boldsymbol{B}^{-1}\boldsymbol{A}\right)^{-1}\right] = \boldsymbol{0}_{K\times K}. \qquad \text{(B.7)}$$

Multiplying (B.6) on the left by $\left(\boldsymbol{B}^{-1} + \boldsymbol{A}^{-1}\right)$, we have

$$\begin{aligned}
& \boldsymbol{I}_K - \left(\boldsymbol{B}^{-1} + \boldsymbol{A}^{-1}\right)\left[\boldsymbol{A} - \left(\boldsymbol{I}_K + \boldsymbol{B}\boldsymbol{A}^{-1}\right)^{-1}\boldsymbol{A}\right] \\
& = \boldsymbol{I}_K - \left(\boldsymbol{B}^{-1} + \boldsymbol{A}^{-1}\right)\left[\boldsymbol{A} - \left(\boldsymbol{B}^{-1} + \boldsymbol{A}^{-1}\right)^{-1}\boldsymbol{B}^{-1}\boldsymbol{A}\right] \\
& = \boldsymbol{I}_K - \boldsymbol{B}^{-1}\boldsymbol{A} - \boldsymbol{I}_K + \boldsymbol{B}^{-1}\boldsymbol{A} \\
& = \boldsymbol{0}_{K\times K}, \qquad \text{(B.8)}
\end{aligned}$$

and multiplying (B.7) on the right by $\left(\boldsymbol{B}^{-1} + \boldsymbol{A}^{-1}\right)$, we have

$$\boldsymbol{I}_K - \left[\boldsymbol{B} - \boldsymbol{B}\left(\boldsymbol{I}_K + \boldsymbol{B}^{-1}\boldsymbol{A}\right)^{-1}\right]\left(\boldsymbol{B}^{-1} + \boldsymbol{A}^{-1}\right)$$

$$
\begin{aligned}
&= \boldsymbol{I}_K - \left[\boldsymbol{B} - \boldsymbol{B}\boldsymbol{A}^{-1}\left(\boldsymbol{B}^{-1} + \boldsymbol{A}^{-1}\right)^{-1}\right]\left(\boldsymbol{B}^{-1} + \boldsymbol{A}^{-1}\right) \\
&= \boldsymbol{I}_K - \boldsymbol{I}_K - \boldsymbol{B}\boldsymbol{A}^{-1} + \boldsymbol{B}\boldsymbol{A}^{-1} \\
&= \boldsymbol{0}_{K\times K}. \qquad \text{(B.9)}
\end{aligned}
$$

From (B.8) and (B.9), (B.6) and (B.7) hold. Hence the remainder term $\boldsymbol{R}_0$ is zero and we can show that (B.1) is correct. Substituting $\boldsymbol{A} = \boldsymbol{X}'\boldsymbol{X}$, $\boldsymbol{B} = \boldsymbol{B}_0^{-1}$, $\boldsymbol{\theta} = \boldsymbol{\beta}$, $\boldsymbol{c} = \hat{\boldsymbol{\beta}}_N$ and $\boldsymbol{d} = \boldsymbol{b}_0$ into (B.1), then we have (A.9). □

C Proof of Lemma A.2

Recall (A.23):

$$
\begin{aligned}
&(\boldsymbol{y}_0 - \boldsymbol{X}\boldsymbol{\beta})'(\boldsymbol{y}_0 - \boldsymbol{X}\boldsymbol{\beta}) + (\boldsymbol{\beta} - \boldsymbol{b}_1)'\boldsymbol{B}_1^{-1}(\boldsymbol{\beta} - \boldsymbol{b}_1) \\
&= (\boldsymbol{y}_0 - \boldsymbol{X}\boldsymbol{b}_1)'\boldsymbol{\Sigma}_0^{-1}(\boldsymbol{y}_0 - \boldsymbol{X}\boldsymbol{b}_1) + (\boldsymbol{\beta} - \boldsymbol{b}_2)'\boldsymbol{B}_2^{-1}(\boldsymbol{\beta} - \boldsymbol{b}_2)
\end{aligned}
$$

where

$$
\begin{aligned}
\boldsymbol{b}_2 &= \boldsymbol{B}_2\left(\boldsymbol{X}'\boldsymbol{y}_0 + \boldsymbol{B}_1^{-1}\boldsymbol{b}_1\right) \\
\boldsymbol{B}_2 &= \left(\boldsymbol{X}'\boldsymbol{X} + \boldsymbol{B}_1^{-1}\right)^{-1} \\
\boldsymbol{\Sigma}_0 &= \boldsymbol{I}_N + \boldsymbol{X}\boldsymbol{B}_1\boldsymbol{X}'.
\end{aligned}
$$

Proof. Expanding the left-hand side of (A.23), we have

$$
\begin{aligned}
&(\boldsymbol{y}_0 - \boldsymbol{X}\boldsymbol{\beta})'(\boldsymbol{y}_0 - \boldsymbol{X}\boldsymbol{\beta}) + (\boldsymbol{\beta} - \boldsymbol{b}_1)'\boldsymbol{B}_1^{-1}(\boldsymbol{\beta} - \boldsymbol{b}_1) \\
&= \boldsymbol{y}_0'\boldsymbol{y}_0 - \boldsymbol{y}_0'\boldsymbol{X}\boldsymbol{\beta} - \boldsymbol{\beta}'\boldsymbol{X}'\boldsymbol{y}_0 + \boldsymbol{\beta}'\boldsymbol{X}'\boldsymbol{X}\boldsymbol{\beta} \\
&\quad + \boldsymbol{\beta}'\boldsymbol{B}_1^{-1}\boldsymbol{\beta} - \boldsymbol{\beta}'\boldsymbol{B}_1^{-1}\boldsymbol{b}_1 - \boldsymbol{b}_1'\boldsymbol{B}_1^{-1}\boldsymbol{\beta} + \boldsymbol{b}_1'\boldsymbol{B}_1^{-1}\boldsymbol{b}_1 \\
&= \operatorname{tr}\left\{\boldsymbol{y}_0\boldsymbol{y}_0' - 2\boldsymbol{X}'\boldsymbol{y}_0\boldsymbol{\beta}' + (\boldsymbol{X}'\boldsymbol{X})\boldsymbol{\beta}\boldsymbol{\beta}'\right. \\
&\qquad \left. + \boldsymbol{B}_1^{-1}\boldsymbol{\beta}\boldsymbol{\beta}' - 2\boldsymbol{B}_1^{-1}\boldsymbol{b}_1\boldsymbol{\beta}' + \boldsymbol{B}_1^{-1}\boldsymbol{b}_1\boldsymbol{b}_1'\right\} \\
&= \operatorname{tr}\left\{\boldsymbol{y}_0\boldsymbol{y}_0' + \boldsymbol{B}_1^{-1}\boldsymbol{b}_1\boldsymbol{b}_1'\right.
\end{aligned}
$$

$$
\begin{aligned}
&\quad + \left(\boldsymbol{X}'\boldsymbol{X} + \boldsymbol{B}_1^{-1}\right)\boldsymbol{\beta}\boldsymbol{\beta}' - 2\left(\boldsymbol{X}'\boldsymbol{y}_0 + \boldsymbol{B}_1^{-1}\boldsymbol{b}_1\right)\boldsymbol{\beta}'\big\} \\
&= \operatorname{tr}\big\{\boldsymbol{y}_0\boldsymbol{y}_0' + \boldsymbol{B}_1^{-1}\boldsymbol{b}_1\boldsymbol{b}_1' \\
&\quad + \left(\boldsymbol{X}'\boldsymbol{X} + \boldsymbol{B}_1^{-1}\right)(\boldsymbol{\beta} - \boldsymbol{b}_2)(\boldsymbol{\beta} - \boldsymbol{b}_2)' \\
&\quad - \left(\boldsymbol{X}'\boldsymbol{X} + \boldsymbol{B}_1^{-1}\right)\boldsymbol{b}_2\boldsymbol{b}_2'\big\} \\
&= (\boldsymbol{\beta} - \boldsymbol{b}_2)'\,\boldsymbol{B}_2^{-1}\,(\boldsymbol{\beta} - \boldsymbol{b}_2) + \boldsymbol{R}_1 \qquad \text{(C.1)}
\end{aligned}
$$

where

$$
\begin{aligned}
\boldsymbol{R}_1 &= \boldsymbol{y}_0'\boldsymbol{y}_0 + \boldsymbol{b}_1'\boldsymbol{B}_1^{-1}\boldsymbol{b}_1 - \boldsymbol{b}_2'\left(\boldsymbol{X}'\boldsymbol{X} + \boldsymbol{B}_1^{-1}\right)\boldsymbol{b}_2 \\
\boldsymbol{b}_2 &= \boldsymbol{B}_2\left(\boldsymbol{X}'\boldsymbol{y}_0 + \boldsymbol{B}_1^{-1}\boldsymbol{b}_1\right) \\
\boldsymbol{B}_2 &= \left(\boldsymbol{X}'\boldsymbol{X} + \boldsymbol{B}_1^{-1}\right)^{-1}.
\end{aligned}
$$

Next the remainder term $\boldsymbol{R}_1$ in (C.1) can be rewritten as

$$
\begin{aligned}
\boldsymbol{R}_1 &= \boldsymbol{y}_0'\boldsymbol{y}_0 + \boldsymbol{b}_1'\boldsymbol{B}_1^{-1}\boldsymbol{b}_1 - \boldsymbol{b}_2'\left(\boldsymbol{X}'\boldsymbol{X} + \boldsymbol{B}_1^{-1}\right)\boldsymbol{b}_2 \\
&= \boldsymbol{y}_0'\boldsymbol{y}_0 + \boldsymbol{b}_1'\boldsymbol{B}_1^{-1}\boldsymbol{b}_1 \\
&\quad - \left[\left(\boldsymbol{X}'\boldsymbol{X} + \boldsymbol{B}_1^{-1}\right)^{-1}\left(\boldsymbol{X}'\boldsymbol{y}_0 + \boldsymbol{B}_1^{-1}\boldsymbol{b}_1\right)\right]'\left(\boldsymbol{X}'\boldsymbol{X} + \boldsymbol{B}_1^{-1}\right) \\
&\quad \times \left[\left(\boldsymbol{X}'\boldsymbol{X} + \boldsymbol{B}_1^{-1}\right)^{-1}\left(\boldsymbol{X}'\boldsymbol{y}_0 + \boldsymbol{B}_1^{-1}\boldsymbol{b}_1\right)\right] \\
&= \boldsymbol{y}_0'\boldsymbol{y}_0 + \boldsymbol{b}_1'\boldsymbol{B}_1^{-1}\boldsymbol{b}_1 \\
&\quad - \left(\boldsymbol{X}'\boldsymbol{y}_0 + \boldsymbol{B}_1^{-1}\boldsymbol{b}_1\right)'\left(\boldsymbol{X}'\boldsymbol{X} + \boldsymbol{B}_1^{-1}\right)^{-1}\left(\boldsymbol{X}'\boldsymbol{y}_0 + \boldsymbol{B}_1^{-1}\boldsymbol{b}_1\right) \\
&= \boldsymbol{y}_0'\boldsymbol{y}_0 + \boldsymbol{b}_1'\boldsymbol{B}_1^{-1}\boldsymbol{b}_1 \\
&\quad - \left(\boldsymbol{X}'\boldsymbol{y}_0 - \boldsymbol{X}'\boldsymbol{X}\boldsymbol{b}_1 + \boldsymbol{X}'\boldsymbol{X}\boldsymbol{b}_1 + \boldsymbol{B}_1^{-1}\boldsymbol{b}_1\right)' \\
&\quad \times \left(\boldsymbol{X}'\boldsymbol{X} + \boldsymbol{B}_1^{-1}\right)^{-1} \\
&\quad \times \left(\boldsymbol{X}'\boldsymbol{y}_0 - \boldsymbol{X}'\boldsymbol{X}\boldsymbol{b}_1 + \boldsymbol{X}'\boldsymbol{X}\boldsymbol{b}_1 + \boldsymbol{B}_1^{-1}\boldsymbol{b}_1\right) \\
&= \boldsymbol{y}_0'\boldsymbol{y}_0 + \boldsymbol{b}_1'\boldsymbol{B}_1^{-1}\boldsymbol{b}_1 \\
&\quad - \left(\boldsymbol{X}'\boldsymbol{y}_0 - \boldsymbol{X}'\boldsymbol{X}\boldsymbol{b}_1 + \boldsymbol{B}_2^{-1}\boldsymbol{b}_1\right)'
\end{aligned}
$$

$$
\begin{aligned}
&\times \boldsymbol{B}_2 \\
&\times \left(\boldsymbol{X}'\boldsymbol{y}_0 - \boldsymbol{X}'\boldsymbol{X}\boldsymbol{b}_1 + \boldsymbol{B}_2^{-1}\boldsymbol{b}_1\right) \\
=\ & \boldsymbol{y}_0'\boldsymbol{y}_0 + \boldsymbol{b}_1'\boldsymbol{B}_1^{-1}\boldsymbol{b}_1 \\
&- \left(\boldsymbol{X}'\boldsymbol{y}_0 - \boldsymbol{X}'\boldsymbol{X}\boldsymbol{b}_1\right)' \boldsymbol{B}_2 \left(\boldsymbol{X}'\boldsymbol{y}_0 - \boldsymbol{X}'\boldsymbol{X}\boldsymbol{b}_1\right) \\
&- \left(\boldsymbol{X}'\boldsymbol{y}_0 - \boldsymbol{X}'\boldsymbol{X}\boldsymbol{b}_1\right)' \boldsymbol{B}_2 \left(\boldsymbol{B}_2^{-1}\boldsymbol{b}_1\right) \\
&- \left(\boldsymbol{B}_2^{-1}\boldsymbol{b}_1\right)' \boldsymbol{B}_2 \left(\boldsymbol{X}'\boldsymbol{y}_0 - \boldsymbol{X}'\boldsymbol{X}\boldsymbol{b}_1\right) \\
&- \left(\boldsymbol{B}_2^{-1}\boldsymbol{b}_1\right)' \boldsymbol{B}_2 \left(\boldsymbol{B}_2^{-1}\boldsymbol{b}_1\right) \\
=\ & \boldsymbol{y}_0'\boldsymbol{y}_0 + \boldsymbol{b}_1'\boldsymbol{B}_1^{-1}\boldsymbol{b}_1 \\
&- \left(\boldsymbol{y}_0 - \boldsymbol{X}\boldsymbol{b}_1\right)' \boldsymbol{X}\boldsymbol{B}_2\boldsymbol{X}' \left(\boldsymbol{y}_0 - \boldsymbol{X}\boldsymbol{b}_1\right) \\
&- \left(\boldsymbol{y}_0 - \boldsymbol{X}\boldsymbol{b}_1\right)' \boldsymbol{X}\boldsymbol{b}_1 \\
&- \boldsymbol{b}_1'\boldsymbol{X}' \left(\boldsymbol{y}_0 - \boldsymbol{X}\boldsymbol{b}_1\right) \\
&- \boldsymbol{b}_1'\boldsymbol{B}_2^{-1}\boldsymbol{b}_1 \\
=\ & \boldsymbol{y}_0'\boldsymbol{y}_0 + \boldsymbol{b}_1'\boldsymbol{B}_1^{-1}\boldsymbol{b}_1 \\
&+ \left(\boldsymbol{y}_0 - \boldsymbol{X}\boldsymbol{b}_1\right)' \left(\boldsymbol{I}_N - \boldsymbol{X}\boldsymbol{B}_2\boldsymbol{X}'\right) \left(\boldsymbol{y}_0 - \boldsymbol{X}\boldsymbol{b}_1\right) \\
&- \left(\boldsymbol{y}_0 - \boldsymbol{X}\boldsymbol{b}_1\right)' \left(\boldsymbol{y}_0 - \boldsymbol{X}\boldsymbol{b}_1\right) \\
&- 2\left(\boldsymbol{y}_0 - \boldsymbol{X}\boldsymbol{b}_1\right)' \boldsymbol{X}\boldsymbol{b}_1 \\
&- \boldsymbol{b}_1'\boldsymbol{B}_2^{-1}\boldsymbol{b}_1. \qquad \text{(C.2)}
\end{aligned}
$$

Substituting

$$
\begin{aligned}
\boldsymbol{I}_N - \boldsymbol{X}\boldsymbol{B}_2\boldsymbol{X}' &= \boldsymbol{I}_N - \boldsymbol{X}\left(\boldsymbol{X}'\boldsymbol{X} + \boldsymbol{B}_1^{-1}\right)^{-1}\boldsymbol{X}' \\
&= \left(\boldsymbol{I}_N + \boldsymbol{X}\boldsymbol{B}_1\boldsymbol{X}'\right)^{-1} \qquad \text{(C.3)}
\end{aligned}
$$

into (C.2), we have

$$
\begin{aligned}
\boldsymbol{R}_1 = \ & \boldsymbol{y}_0'\boldsymbol{y}_0 + \boldsymbol{b}_1'\boldsymbol{B}_1^{-1}\boldsymbol{b}_1 \\
&+ \left(\boldsymbol{y}_0 - \boldsymbol{X}\boldsymbol{b}_1\right)' \left(\boldsymbol{I}_N + \boldsymbol{X}\boldsymbol{B}_1\boldsymbol{X}'\right)^{-1} \left(\boldsymbol{y}_0 - \boldsymbol{X}\boldsymbol{b}_1\right)
\end{aligned}
$$

$$
\begin{aligned}
&- (\boldsymbol{y}_0 - \boldsymbol{X}\boldsymbol{b}_1)' (\boldsymbol{y}_0 - \boldsymbol{X}\boldsymbol{b}_1) \\
&- 2 (\boldsymbol{y}_0 - \boldsymbol{X}\boldsymbol{b}_1)' \boldsymbol{X}\boldsymbol{b}_1 \\
&- \boldsymbol{b}_1' \left(\boldsymbol{X}'\boldsymbol{X} + \boldsymbol{B}_1^{-1}\right) \boldsymbol{b}_1 \\
&= (\boldsymbol{y}_0 - \boldsymbol{X}\boldsymbol{b}_1)' (\boldsymbol{I}_N + \boldsymbol{X}\boldsymbol{B}_1\boldsymbol{X}')^{-1} (\boldsymbol{y}_0 - \boldsymbol{X}\boldsymbol{b}_1) \\
&= (\boldsymbol{y}_0 - \boldsymbol{X}\boldsymbol{b}_1)' \boldsymbol{\Sigma}_0^{-1} (\boldsymbol{y}_0 - \boldsymbol{X}\boldsymbol{b}_1) . \qquad \text{(C.4)}
\end{aligned}
$$

Hence, from (C.1) and (C.4), we have (A.23). □

D Classical Measure of Model Selection

D.1 Akaike's Information Criterion (AIC)

We denote a true density (or model) $f_X(x)$ and its approximating model $g(x|\boldsymbol{\theta})$ with K-dimensional parameter vector $\boldsymbol{\theta}$. Given the *i.i.d.* observed data $\boldsymbol{y} \equiv (y_1, y_2, \ldots, y_N)$, MLE of parameter $\boldsymbol{\theta}$ is defined as $\hat{\boldsymbol{\theta}}_N \equiv \hat{\boldsymbol{\theta}}_N(\boldsymbol{y})$ under the regularity condition and the true parameter $\boldsymbol{\theta}_0$ satisfies $f_X(x) \equiv g(x|\boldsymbol{\theta}_0)$. Then Kullback-Leibler Information is calculated as

$$
\begin{aligned}
I(f, g(\cdot|\hat{\boldsymbol{\theta}}_N)) &= \int f_X(x) \log \left\{ \frac{f_X(x)}{g(x|\hat{\boldsymbol{\theta}}_N)} \right\} dx. \\
&= constant - \mathbf{E}_x \left[\log\{g(x|\hat{\boldsymbol{\theta}}_N)\} \right] \qquad \text{(D.1)}
\end{aligned}
$$

where the expected log-likelihood $\mathbf{E}_x[\log\{g(x|\hat{\boldsymbol{\theta}}_N)\}]$ in (D.1) is a measure of goodness of fit of approximating model $g(\cdot|\hat{\boldsymbol{\theta}}_N)$ relative to the true density $f_X(x)$.

Now we consider a problem of finding the model that maximizes the estimate of the expected log-likelihood $\mathbf{E}_x[\log\{g(x|\hat{\boldsymbol{\theta}}_N)\}]$ in (D.1) among several competing models. The most natural estimator of the expected log-likelihood is its sample counterpart based on $\boldsymbol{y} \equiv (y_1, y_2, \ldots, y_N)$, that is,

$\frac{1}{N}\sum_{i=1}^{N}\log\{g(y_i|\hat{\boldsymbol{\theta}}_N)\}$. Since we do not know if this is an unbiased estimator of the expected log-likelihood, we examine the asymptotic bias b_Θ when the observed log-likelihood is used as an estimator of expected log-likelihood, which is decomposed in three terms

$$\begin{aligned}
b_\Theta &\equiv \mathbf{E}_y\left[\frac{1}{N}\sum_{i=1}^{N}\log\{g(y_i|\hat{\boldsymbol{\theta}}_N)\} - \mathbf{E}_x\left[\log\{g(x|\hat{\boldsymbol{\theta}}_N)\}\right]\right] \\
&= \mathbf{E}_y\left[\frac{1}{N}\sum_{i=1}^{N}\log\{g(y_i|\hat{\boldsymbol{\theta}}_N)\} - \frac{1}{N}\sum_{i=1}^{N}\log\{g(y_i|\boldsymbol{\theta}_0)\}\right] \\
&\quad + \mathbf{E}_y\left[\frac{1}{N}\sum_{i=1}^{N}\log\{g(y_i|\boldsymbol{\theta}_0)\} - \mathbf{E}_x\left[\log\{g(x|\boldsymbol{\theta}_0)\}\right]\right] \\
&\quad + \mathbf{E}_y\left[\mathbf{E}_x\left[\log\{g(x|\boldsymbol{\theta}_0)\}\right] - \mathbf{E}_x\left[\log\{g(x|\hat{\boldsymbol{\theta}}_N)\}\right]\right] \\
&= \mathbf{E}_y(D_1) + \mathbf{E}_y(D_2) + \mathbf{E}_y(D_3). \qquad \text{(D.2)}
\end{aligned}$$

The three components, D_1, D_2, and D_3 are respectively: The discrepancy between the average observed log-likelihoods of the approximating model $g(\cdot|\boldsymbol{\theta})$ under the MLE $\hat{\boldsymbol{\theta}}_N$ and the true parameter $\boldsymbol{\theta}_0$; The sampling bias of the log-likelihoods of the approximating model $g(\cdot|\boldsymbol{\theta})$ under the true parameter $\boldsymbol{\theta}_0$; The discrepancy between the expected log-likelihoods of the approximating model $g(\cdot|\boldsymbol{\theta})$ under the MLE $\hat{\boldsymbol{\theta}}_N$ and the true parameter $\boldsymbol{\theta}_0$.

First, we evaluate $\mathbf{E}_y(D_1)$ in (D.2). Second-order Taylor series expansion around MLE $\hat{\boldsymbol{\theta}}_N$ gives

$$\begin{aligned}
\frac{1}{N}\sum_{i=1}^{N}\log\{g(y_i|\boldsymbol{\theta})\} \approx \frac{1}{N}\sum_{i=1}^{N}\log\{g(y_i|\hat{\boldsymbol{\theta}}_N)\} \\
+ (\boldsymbol{\theta}-\hat{\boldsymbol{\theta}}_N)'\left[\frac{1}{N}\sum_{i=1}^{N}\frac{\partial\log\{g(y_i|\hat{\boldsymbol{\theta}}_N)\}}{\partial\boldsymbol{\theta}}\right] \\
- \frac{1}{2}(\boldsymbol{\theta}-\hat{\boldsymbol{\theta}}_N)'\left[\frac{1}{N}\sum_{i=1}^{N}-\frac{\partial^2\log\{g(y_i|\hat{\boldsymbol{\theta}}_N)\}}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}'}\right](\boldsymbol{\theta}-\hat{\boldsymbol{\theta}}_N). \qquad \text{(D.3)}
\end{aligned}$$

Assuming that interchange of integral and derivative is valid under the regularity conditions, we have

$$\begin{aligned}\frac{\partial}{\partial \boldsymbol{\theta}} \int g(x|\boldsymbol{\theta}) \frac{\partial \log g(x|\boldsymbol{\theta})}{\partial \boldsymbol{\theta}'} dx &= \int \frac{\partial}{\partial \boldsymbol{\theta}} \left\{ g(x|\boldsymbol{\theta}) \frac{\partial \log g(x|\boldsymbol{\theta})}{\partial \boldsymbol{\theta}'} \right\} dx \\ &= \int g(x|\boldsymbol{\theta}) \left\{ \frac{\partial \log g(x|\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} \frac{\partial \log g(x|\boldsymbol{\theta})}{\partial \boldsymbol{\theta}'} \right\} dx \\ &\quad + \int g(x|\boldsymbol{\theta}) \left\{ \frac{\partial^2 \log g(x|\boldsymbol{\theta})}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}'} \right\} dx. \end{aligned} \tag{D.4}$$

Since the left-hand side of (D.4) is zero because

$$\begin{aligned}\frac{\partial}{\partial \boldsymbol{\theta}} \int g(x|\boldsymbol{\theta}) \frac{\partial \log g(x|\boldsymbol{\theta})}{\partial \boldsymbol{\theta}'} dx &= \frac{\partial}{\partial \boldsymbol{\theta}} \int g(x|\boldsymbol{\theta}) \frac{1}{g(x|\boldsymbol{\theta})} \frac{\partial g(x|\boldsymbol{\theta})}{\partial \boldsymbol{\theta}'} dx \\ &= \frac{\partial}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}'} \int g(x|\boldsymbol{\theta}) dx \\ &= 0, \end{aligned}$$

when evaluated at $\boldsymbol{\theta} = \boldsymbol{\theta}_0$ in (D.4), we have

$$\boldsymbol{J}(\boldsymbol{\theta}_0) = \boldsymbol{I}(\boldsymbol{\theta}_0) \tag{D.5}$$

where due to the fact that $g(x|\boldsymbol{\theta}_0) = f_X(x)$,

$$\boldsymbol{J}(\boldsymbol{\theta}_0) \equiv \int f_X(x) \left\{ -\frac{\partial^2 \log g(x|\boldsymbol{\theta}_0)}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}'} \right\} dx \tag{D.6}$$

$$\boldsymbol{I}(\boldsymbol{\theta}_0) \equiv \int f_X(x) \left\{ \frac{\partial \log g(x|\boldsymbol{\theta}_0)}{\partial \boldsymbol{\theta}} \frac{\partial \log g(x|\boldsymbol{\theta}_0)}{\partial \boldsymbol{\theta}'} \right\} dx. \tag{D.7}$$

Since $\hat{\boldsymbol{\theta}}_N \xrightarrow{p} \boldsymbol{\theta}_0$ as $N \to \infty$, we have

$$\frac{1}{N} \sum_{i=1}^{N} \left[-\frac{\partial^2 \log\{g(y_i|\hat{\boldsymbol{\theta}}_N)\}}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}'} \right] \xrightarrow{p} \boldsymbol{J}(\boldsymbol{\theta}_0) = \boldsymbol{I}(\boldsymbol{\theta}_0). \tag{D.8}$$

Moreover, asymptotic normality of MLE shows that

$$\sqrt{N}(\hat{\boldsymbol{\theta}}_N - \boldsymbol{\theta}_0) \overset{w}{\rightsquigarrow} \mathcal{N}\left(\mathbf{0}, \boldsymbol{I}^{-1}(\boldsymbol{\theta}_0)\right). \tag{D.9}$$

Since the second term on the right-hand side of (D.3) is zero because $\hat{\boldsymbol{\theta}}_N$ is MLE, $\mathbf{E}_y(D_1)$ can be asymptotically evaluated in view of Slutsky's theorem by substituting $\boldsymbol{\theta} = \boldsymbol{\theta}_0$ into (D.3) as follows

$$\begin{aligned}\mathbf{E}_y(D_1) &\xrightarrow{p} \frac{1}{2}\mathrm{tr}\left\{\boldsymbol{I}(\boldsymbol{\theta}_0)\mathbf{E}_y\left[(\hat{\boldsymbol{\theta}}_N - \boldsymbol{\theta}_0)(\hat{\boldsymbol{\theta}}_N - \boldsymbol{\theta}_0)'\right]\right\} \\ &= \frac{1}{2N}\mathrm{tr}\left\{\boldsymbol{I}(\boldsymbol{\theta}_0)\boldsymbol{I}^{-1}(\boldsymbol{\theta}_0)\right\} \\ &= \frac{K}{2N}. \end{aligned} \tag{D.10}$$

Next we can evaluate $\mathbf{E}_y(D_2)$ in (D.2) as

$$\begin{aligned}\mathbf{E}_y(D_2) &= \mathbf{E}_y\left[\frac{1}{N}\sum_{i=1}^{N}\log\{g(y_i|\boldsymbol{\theta}_0)\} - \mathbf{E}_x\left[\log\{g(x|\boldsymbol{\theta}_0)\}\right]\right] \\ &= \mathbf{E}_y\left[\log\{g(y|\boldsymbol{\theta}_0)\}\right] - \mathbf{E}_x\left[\log\{g(x|\boldsymbol{\theta}_0)\}\right] \\ &= \mathbf{E}_y\left[\log\{f_X(y)\}\right] - \mathbf{E}_x\left[\log\{f_X(x)\}\right] \\ &= 0. \end{aligned} \tag{D.11}$$

Finally, we evaluate $\mathbf{E}_y(D_3)$ in (D.2) based on the second-order Taylor series expansion around true parameter $\boldsymbol{\theta}_0$:

$$\begin{aligned}\mathbf{E}_x\left[\log\{g(x|\hat{\boldsymbol{\theta}}_N)\}\right] \approx \mathbf{E}_x\left[\log\{g(x|\boldsymbol{\theta}_0)\}\right] &+ (\hat{\boldsymbol{\theta}}_N - \boldsymbol{\theta}_0)'\mathbf{E}_x\left[\frac{\partial\log\{g(x|\boldsymbol{\theta}_0)\}}{\partial\boldsymbol{\theta}}\right] \\ &- \frac{1}{2}(\hat{\boldsymbol{\theta}}_N - \boldsymbol{\theta}_0)'\mathbf{E}_x\left[-\frac{\partial^2\log\{g(x|\boldsymbol{\theta}_0)\}}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}'}\right](\hat{\boldsymbol{\theta}}_N - \boldsymbol{\theta}_0). \end{aligned} \tag{D.12}$$

The second term on the right-hand side of (D.12) is zero because the true parameter $\boldsymbol{\theta}_0$ is a point solution of $\mathbf{E}_x\left[\partial\log\{g(x|\boldsymbol{\theta})\}/\partial\boldsymbol{\theta}\right] = \mathbf{0}$. Since $\boldsymbol{J}(\boldsymbol{\theta}_0) = \boldsymbol{I}(\boldsymbol{\theta}_0)$, we can evaluate $\mathbf{E}_y(D_3)$ as follows

$$\begin{aligned}\mathbf{E}_y(D_3) &\xrightarrow{p} \frac{1}{2}\mathrm{tr}\left\{\boldsymbol{I}(\boldsymbol{\theta}_0)\mathbf{E}_y\left[(\hat{\boldsymbol{\theta}}_N - \boldsymbol{\theta}_0)(\hat{\boldsymbol{\theta}}_N - \boldsymbol{\theta}_0)'\right]\right\} \\ &= \frac{1}{2N}\mathrm{tr}\left\{\boldsymbol{I}(\boldsymbol{\theta}_0)\boldsymbol{I}^{-1}(\boldsymbol{\theta}_0)\right\}\end{aligned}$$

$$= \frac{K}{2N}. \tag{D.13}$$

Therefore the asymptotic bias is

$$\begin{aligned}\mathbf{E}_y(D_1) + \mathbf{E}_y(D_2) + \mathbf{E}_y(D_3) &\xrightarrow{p} \frac{K}{2N} + 0 + \frac{K}{2N} \\ &= \frac{K}{N} \\ &= b_\Theta. \end{aligned} \tag{D.14}$$

Then we have an asymptotically unbiased estimator of expected log-likelihood as

$$\frac{1}{N}\sum_{i=1}^{N} \log\{g(y_i|\hat{\boldsymbol{\theta}}_N)\} - \frac{K}{N} \tag{D.15}$$

As a matter of convention the criterion is often stated as that of minimizing

$$-2\log\{g(\boldsymbol{y}|\hat{\boldsymbol{\theta}}_N)\} + 2K.$$

Therefore AIC can be computed as

$$\mathrm{AIC} = -2\log\{\mathcal{L}_g(\hat{\boldsymbol{\theta}}_N|\boldsymbol{y})\} + 2K \tag{D.16}$$

where $\mathcal{L}_g(\hat{\boldsymbol{\theta}}_N|\boldsymbol{y}) \equiv g(\boldsymbol{y}|\hat{\boldsymbol{\theta}}_N)$.

D.2 Corrected AIC in a Finite Sample Size (AIC_C)

Sugiura (1978) derived the corrected AIC when the number of data N is small for the linear regression model with normally distributed error:

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}, \quad \boldsymbol{\varepsilon} \sim \mathcal{N}\left(\mathbf{0}, \sigma^2 \boldsymbol{I}_N\right) \tag{D.17}$$

where $\boldsymbol{y}$ is a $N \times 1$ vector, $\boldsymbol{X}$ is a $N \times K$ matrix, and $\boldsymbol{\beta}$ is a $K \times 1$ parameter vector. The maximized log-likelihood of candidate linear regression model $g(\cdot|\boldsymbol{X}, \hat{\boldsymbol{\beta}}_N, \hat{\sigma}_N^{-2})$ can be taken as

$$\log\left\{g(\boldsymbol{y}|\boldsymbol{X}, \hat{\boldsymbol{\beta}}_N, \hat{\sigma}_N^{-2})\right\} = -\frac{N}{2}\log\hat{\sigma}_N^2 - \frac{1}{2}\frac{(\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}_N)'(\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}_N)}{\hat{\sigma}_N^2} \tag{D.18}$$

where MLEs in (D.18) are well-known:

$$\hat{\boldsymbol{\beta}}_N = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{y} \tag{D.19}$$

$$\hat{\sigma}^2_N = (\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}_N)'(\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}_N)/N. \tag{D.20}$$

Let us denote the sample $\boldsymbol{z} \equiv (z_1, z_2, \ldots, z_N)$ from the true density $f_{\mathbf{Y}}(\boldsymbol{z})$ to evaluate expected log-likelihood and the true density is assumed to be a N-dimensional multivariate normal distribution $\boldsymbol{z} \sim \mathcal{N}(\boldsymbol{\mu}, \sigma^2\boldsymbol{I}_N)$, where we assume that $K \times 1$ true parameter $\boldsymbol{\beta}$ satisfies $\boldsymbol{\mu} = \boldsymbol{X}\boldsymbol{\beta}$. We define the expected log-likelihood $\mathcal{T}$ as

$$\begin{aligned}\mathcal{T} &\equiv \int \log\left\{g(\boldsymbol{z}|\boldsymbol{X}, \hat{\boldsymbol{\beta}}_N, \hat{\sigma}_N^{-2})\right\} f_{\mathbf{Y}}(\boldsymbol{z})d\boldsymbol{z} \\ &= \mathbf{E}_z\left[-\frac{N}{2}\log\hat{\sigma}^2_N - \frac{1}{2}\frac{(\boldsymbol{z} - \boldsymbol{X}\hat{\boldsymbol{\beta}}_N)'(\boldsymbol{z} - \boldsymbol{X}\hat{\boldsymbol{\beta}}_N)}{\hat{\sigma}^2_N}\right]\end{aligned} \tag{D.21}$$

and define the observed log-likelihood $\mathcal{T}_N$ in (D.18) as

$$\begin{aligned}\mathcal{T}_N &\equiv \log\left\{g(\boldsymbol{y}|\boldsymbol{X}, \hat{\boldsymbol{\beta}}_N, \hat{\sigma}_N^{-2})\right\} \\ &= -\frac{N}{2}\log\hat{\sigma}^2_N - \frac{1}{2}\frac{(\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}_N)'(\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}_N)}{\hat{\sigma}^2_N} \\ &= -\frac{N}{2}\log\hat{\sigma}^2_N - \frac{N}{2}.\end{aligned} \tag{D.22}$$

From (D.21) and (D.22), bias b_Θ when the observed log-likelihood $\mathcal{T}_N$ is used as an estimator of expected log-likelihood $\mathcal{T}$ is defined as

$$\begin{aligned}b_\Theta &\equiv \mathbf{E}_y\left[\mathcal{T}_N - \mathcal{T}\right] \\ &= -\frac{N}{2} + \frac{1}{2}\mathbf{E}_y\mathbf{E}_z\left[\frac{(\boldsymbol{z} - \boldsymbol{X}\hat{\boldsymbol{\beta}}_N)'(\boldsymbol{z} - \boldsymbol{X}\hat{\boldsymbol{\beta}}_N)}{\hat{\sigma}^2_N}\right].\end{aligned} \tag{D.23}$$

The second term on the right-hand of (D.23) is evaluated as

$$\frac{1}{2}\mathbf{E}_y\mathbf{E}_z\left[\frac{(\boldsymbol{z} - \boldsymbol{X}\hat{\boldsymbol{\beta}}_N)'(\boldsymbol{z} - \boldsymbol{X}\hat{\boldsymbol{\beta}}_N)}{\hat{\sigma}^2_N}\right]$$

$$
\begin{aligned}
&= \frac{1}{2}\mathbf{E}_y\mathbf{E}_z\left[\frac{(\boldsymbol{z}-\boldsymbol{X}\boldsymbol{\beta}+\boldsymbol{X}\boldsymbol{\beta}-\boldsymbol{X}\hat{\boldsymbol{\beta}}_N)'(\boldsymbol{z}-\boldsymbol{X}\boldsymbol{\beta}+\boldsymbol{X}\boldsymbol{\beta}-\boldsymbol{X}\hat{\boldsymbol{\beta}}_N)}{\hat{\sigma}_N^2}\right] \\
&= \frac{1}{2}\mathbf{E}_y\mathbf{E}_z\left[\frac{(\boldsymbol{z}-\boldsymbol{X}\boldsymbol{\beta})'(\boldsymbol{z}-\boldsymbol{X}\boldsymbol{\beta})}{\hat{\sigma}_N^2}\right] + \frac{1}{2}\mathbf{E}_y\mathbf{E}_z\left[\frac{(\boldsymbol{X}\boldsymbol{\beta}-\boldsymbol{X}\hat{\boldsymbol{\beta}}_N)'(\boldsymbol{X}\boldsymbol{\beta}-\boldsymbol{X}\hat{\boldsymbol{\beta}}_N)}{\hat{\sigma}_N^2}\right] \\
&= \frac{1}{2}\mathbf{E}_y\left[\frac{N\sigma^2}{\hat{\sigma}_N^2}\right] + \frac{1}{2}\mathbf{E}_y\left[\frac{(\boldsymbol{\beta}-\hat{\boldsymbol{\beta}}_N)'(\boldsymbol{X}'\boldsymbol{X})(\boldsymbol{\beta}-\hat{\boldsymbol{\beta}}_N)}{\hat{\sigma}_N^2}\right] \\
&= \frac{N^2}{2}\mathbf{E}_y\left[\frac{\sigma^2}{N\hat{\sigma}_N^2}\right] + \frac{N}{2}\mathbf{E}_y\left[\frac{(\boldsymbol{\beta}-\hat{\boldsymbol{\beta}}_N)\sigma^{-2}(\boldsymbol{X}'\boldsymbol{X})(\boldsymbol{\beta}-\hat{\boldsymbol{\beta}}_N)}{N\hat{\sigma}_N^2/\sigma^2}\right]. \qquad \text{(D.24)}
\end{aligned}
$$

Using the fact that $N\hat{\sigma}_N^2/\sigma^2 \sim \chi^2_{N-K}$, we have

$$
\mathbf{E}_y\left[\frac{N\sigma^2}{\hat{\sigma}_N^2}\right] = \mathbf{E}_y\left[\frac{1}{\chi^2_{N-K}}\right] = \frac{1}{N-K-2}. \qquad \text{(D.25)}
$$

Moreover, since asymptotic normality $\hat{\boldsymbol{\beta}}_N \sim \mathcal{N}\left(\boldsymbol{\beta}, \sigma^2(\boldsymbol{X}'\boldsymbol{X})^{-1}\right)$ holds with respect to sample $\boldsymbol{y} \equiv (y_1, y_2, \ldots, y_N)$ generated from the true density $f_{\mathbf{Y}}(\boldsymbol{y})$, we have

$$
\left(\frac{N-K}{K}\right)\frac{(\boldsymbol{\beta}-\hat{\boldsymbol{\beta}}_N)\sigma^{-2}(\boldsymbol{X}'\boldsymbol{X})(\boldsymbol{\beta}-\hat{\boldsymbol{\beta}}_N)'}{N\hat{\sigma}_N^2/\sigma^2} \sim F(K, N-K) \qquad \text{(D.26)}
$$

and expectation of (D.26) with respect to $f_{\mathbf{Y}}(\boldsymbol{y})$ is derived by

$$
\mathbf{E}_y\left[\frac{(\boldsymbol{\beta}-\hat{\boldsymbol{\beta}}_N)\sigma^{-2}(\boldsymbol{X}'\boldsymbol{X})(\boldsymbol{\beta}-\hat{\boldsymbol{\beta}}_N)'}{N\hat{\sigma}_N^2/\sigma^2}\right] = \left(\frac{K}{N-K}\right)\frac{N-K}{N-K-2}. \qquad \text{(D.27)}
$$

Therefore

$$
\begin{aligned}
\mathcal{T}_N - b_\Theta &= \mathcal{T}_N + \frac{N}{2} - \frac{N^2}{2(N-K-2)} - \frac{NK}{2(N-K-2)} \\
&= -\frac{N}{2}\log\hat{\sigma}_N^2 - \frac{N(N+K)}{2(N-K-2)}. \qquad \text{(D.28)}
\end{aligned}
$$

Consequently, multiplying -2 to (D.28) we have

$$
\mathrm{AIC}_C = N\left\{\log\hat{\sigma}_N^2 + 1\right\} + 2\frac{N(K+1)}{N-K-2}. \qquad \text{(D.29)}
$$

D.2.1 Proof of Eq.(D.25)

We show that

$$\frac{N\hat{\sigma}_N^2}{\sigma^2} \sim \chi^2_{N-K}$$

where $\hat{\sigma}_N^2 = (\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}_N)'(\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}_N)/N$.

Proof. We consider the QR decomposition of the $N \times K$ matrix $\boldsymbol{X}$ such as

$$\boldsymbol{X} = \boldsymbol{Q}\boldsymbol{R} \tag{D.30}$$

where $\boldsymbol{Q}$ is an orthogonal $N \times N$ matrix and $\boldsymbol{R}$ is a $N \times K$ matrix defined as

$$\boldsymbol{R} = \begin{bmatrix} \boldsymbol{R}_K \\ \boldsymbol{0}_{(N-K)\times K} \end{bmatrix} \tag{D.31}$$

by using a $K \times K$ invertible upper triangular matrix $\boldsymbol{R}_K$.

We can rewrite the linear regression model with normally distributed error $\boldsymbol{\varepsilon}$ as follows

$$\boldsymbol{\varepsilon} = \left(\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}_N\right) + \boldsymbol{X}\left(\hat{\boldsymbol{\beta}}_N - \boldsymbol{\beta}\right), \quad \boldsymbol{\varepsilon} \sim \mathcal{N}\left(\boldsymbol{0}, \sigma^2\boldsymbol{I}_N\right). \tag{D.32}$$

where $\hat{\boldsymbol{\beta}}_N = \left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\boldsymbol{X}'\boldsymbol{y}$.

Multiplying $\boldsymbol{Q}'$ and substituting $\boldsymbol{X} = \boldsymbol{Q}\boldsymbol{R}$ into (D.32), we have

$$\begin{aligned} \boldsymbol{Q}'\boldsymbol{\varepsilon} &= \boldsymbol{Q}'\boldsymbol{y} - \boldsymbol{Q}'\left(\boldsymbol{Q}\boldsymbol{R}\right)\hat{\boldsymbol{\beta}}_N + \boldsymbol{Q}'\left(\boldsymbol{Q}\boldsymbol{R}\right)\left(\hat{\boldsymbol{\beta}}_N - \boldsymbol{\beta}\right) \\ &= \boldsymbol{Q}'\boldsymbol{y} - \boldsymbol{R}\hat{\boldsymbol{\beta}}_N + \boldsymbol{R}\left(\hat{\boldsymbol{\beta}}_N - \boldsymbol{\beta}\right). \end{aligned} \tag{D.33}$$

Let us denote $\boldsymbol{q}^{\varepsilon}_{N\times 1} = \boldsymbol{Q}'\boldsymbol{\varepsilon}$ and $\boldsymbol{q}^{y}_{N\times 1} = \boldsymbol{Q}'\boldsymbol{y}$. (D.33) can be written as

$$\boldsymbol{q}^{\varepsilon}_{N\times 1} = \boldsymbol{q}^{y}_{N\times 1} - \boldsymbol{R}\hat{\boldsymbol{\beta}}_N + \boldsymbol{R}\left(\hat{\boldsymbol{\beta}}_N - \boldsymbol{\beta}\right) \tag{D.34}$$

where

$$\boldsymbol{q}^{y}_{N\times 1} - \boldsymbol{R}\hat{\boldsymbol{\beta}}_N = \begin{bmatrix} \boldsymbol{q}^{y}_{K\times 1} - \boldsymbol{R}_K\hat{\boldsymbol{\beta}}_N \\ \boldsymbol{q}^{y}_{(N-K)\times 1} \end{bmatrix} \tag{D.35}$$

$$\boldsymbol{R}\left(\hat{\boldsymbol{\beta}}_N - \boldsymbol{\beta}\right) = \begin{bmatrix} \boldsymbol{R}_K\left(\hat{\boldsymbol{\beta}}_N - \boldsymbol{\beta}\right) \\ \mathbf{0}_{(N-K)\times 1} \end{bmatrix}. \tag{D.36}$$

Then $\boldsymbol{q}^y_{K\times 1} - \boldsymbol{R}_K\hat{\boldsymbol{\beta}}_N$ in (D.35) and $\boldsymbol{R}_K(\hat{\boldsymbol{\beta}}_N - \boldsymbol{\beta})$ in (D.36) can be obtained as

$$\begin{aligned} \boldsymbol{q}^y_{K\times 1} - \boldsymbol{R}_K\hat{\boldsymbol{\beta}}_N &= \boldsymbol{q}^y_{K\times 1} - \boldsymbol{R}_K\left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\boldsymbol{X}'\boldsymbol{y} \\ &= \boldsymbol{q}^y_{K\times 1} - \boldsymbol{R}_K\left(\boldsymbol{R}'\boldsymbol{Q}'\boldsymbol{Q}\boldsymbol{R}\right)^{-1}\boldsymbol{R}'\boldsymbol{Q}'\boldsymbol{y} \\ &= \boldsymbol{q}^y_{K\times 1} - \boldsymbol{q}^y_{K\times 1} \\ &= \mathbf{0}_{K\times 1} \end{aligned} \tag{D.37}$$

$$\begin{aligned} \boldsymbol{R}_K\left(\hat{\boldsymbol{\beta}}_N - \boldsymbol{\beta}\right) &= \boldsymbol{R}_K\left(\boldsymbol{\beta} + \left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\boldsymbol{X}'\boldsymbol{\varepsilon} - \boldsymbol{\beta}\right) \\ &= \boldsymbol{R}_K\left(\boldsymbol{R}'\boldsymbol{Q}'\boldsymbol{Q}\boldsymbol{R}\right)^{-1}\boldsymbol{R}'\boldsymbol{Q}'\boldsymbol{\varepsilon} \\ &= \boldsymbol{q}^{\varepsilon}_{K\times 1}. \end{aligned} \tag{D.38}$$

Hence we have

$$\boldsymbol{q}^{\varepsilon}_{N\times 1} = \begin{bmatrix} \boldsymbol{q}^{\varepsilon}_{K\times 1} \\ \boldsymbol{q}^{\varepsilon}_{(N-K)\times 1} \end{bmatrix} \tag{D.39}$$

$$\begin{aligned} &= \boldsymbol{Q}'\boldsymbol{y} - \boldsymbol{R}\hat{\boldsymbol{\beta}}_N + \boldsymbol{R}(\hat{\boldsymbol{\beta}}_N - \boldsymbol{\beta}) \\ &= \begin{bmatrix} \mathbf{0}_{K\times 1} \\ \boldsymbol{q}^y_{(N-K)\times 1} \end{bmatrix} + \begin{bmatrix} \boldsymbol{q}^{\varepsilon}_{K\times 1} \\ \mathbf{0}_{(N-K)\times 1} \end{bmatrix}. \end{aligned} \tag{D.40}$$

Since we notice that

$$\boldsymbol{Q}'\boldsymbol{y} - \boldsymbol{R}\hat{\boldsymbol{\beta}}_N = \boldsymbol{Q}'\left(\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}_N\right) \tag{D.41}$$

$$\boldsymbol{q}^y_{(N-K)\times 1} = \boldsymbol{q}^{\varepsilon}_{(N-K)\times 1} \sim \mathcal{N}\left(\mathbf{0}, \sigma^2\boldsymbol{I}_{N-K}\right) \tag{D.42}$$

then

$$\left\{\boldsymbol{Q}'\boldsymbol{y} - \boldsymbol{R}\hat{\boldsymbol{\beta}}_N\right\}'\left\{\boldsymbol{Q}'\boldsymbol{y} - \boldsymbol{R}\hat{\boldsymbol{\beta}}_N\right\} = \left(\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}_N\right)'\left(\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}_N\right)$$

$$= (q^{\varepsilon}_{K+1})^2 + (q^{\varepsilon}_{K+2})^2 + \cdots + (q^{\varepsilon}_{N})^2. \quad \text{(D.43)}$$

Therefore

$$\sigma^{-2}\left(\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}_N\right)'\left(\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}_N\right) = \sigma^{-2}\left[(q^{\varepsilon}_{K+1})^2 + (q^{\varepsilon}_{K+2})^2 + \cdots + (q^{\varepsilon}_{N})^2\right]$$
$$\sim \chi^2_{N-K}. \quad \text{(D.44)}$$

□

D.3 Bayesian Information Criterion (BIC)

The posterior probability of selecting a candidate model M is obtained as

$$\Pr(M|\boldsymbol{x}) = \frac{p(\boldsymbol{x}|M)p(M)}{p(\boldsymbol{x})} \quad \text{(D.45)}$$

where $\boldsymbol{x} \equiv (x_1, \ldots, x_N)$ is the *i.i.d.* observed data.

Assuming that candidate models have same prior probabilities $p(M)$, the critical quantity to be approximated is the marginal likelihood of the candidate model M:

$$p(\boldsymbol{x}|M) = \int \left[\prod_{i=1}^{N} g(x_i|\boldsymbol{\theta}, M)\right] p(\boldsymbol{\theta}|M)d\boldsymbol{\theta} \quad \text{(D.46)}$$

where parameter $\boldsymbol{\theta}$ has dimension K.

First we rewrite (D.46) as

$$\int g(\boldsymbol{x}|\boldsymbol{\theta}, M)p(\boldsymbol{\theta}|M)d\boldsymbol{\theta} = \int \exp\left[Q(\boldsymbol{\theta})\right] d\boldsymbol{\theta} \quad \text{(D.47)}$$

where

$$Q(\boldsymbol{\theta}) = \log\{g(\boldsymbol{x}|\boldsymbol{\theta}, M)p(\boldsymbol{\theta}|M)\}. \quad \text{(D.48)}$$

From the second-order Taylor series expansion of $Q(\boldsymbol{\theta})$, we can approximate it around the MLE $\hat{\boldsymbol{\theta}}_N \equiv \hat{\boldsymbol{\theta}}(\boldsymbol{x})$ as follows:

$$Q(\boldsymbol{\theta}) \approx Q(\hat{\boldsymbol{\theta}}_N) + (\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}_N)' \left.\frac{\partial Q(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}\right|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}_N} + \frac{1}{2}(\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}_N)' \left.\frac{\partial^2 Q(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}\partial \boldsymbol{\theta}'}\right|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}_N} (\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}_N). \quad \text{(D.49)}$$

Assuming that prior distribution $p(\boldsymbol{\theta}|M)$ is a non-informative flat prior and sample size N is large, we can treat $p(\boldsymbol{\theta}|M)$ as a constant. Then we have

$$\left.\frac{\partial Q(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}\right|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}_N} = \left.\frac{\partial \log\{g(\boldsymbol{x}|\boldsymbol{\theta}, M)\}}{\partial \boldsymbol{\theta}}\right|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}_N} + \left.\frac{\partial \log\{p(\boldsymbol{\theta}|M)\}}{\partial \boldsymbol{\theta}}\right|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}_N} = \mathbf{0} \tag{D.50}$$

and

$$\begin{aligned}\left.\frac{\partial^2 Q(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}\partial \boldsymbol{\theta}'}\right|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}_N} &= \left.\frac{\partial^2 \log\{g(\boldsymbol{x}|\boldsymbol{\theta}, M)\}}{\partial \boldsymbol{\theta}\partial \boldsymbol{\theta}'}\right|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}_N} + \left.\frac{\partial \log\{p(\boldsymbol{\theta}|M)\}}{\partial \boldsymbol{\theta}\partial \boldsymbol{\theta}'}\right|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}_N} \\ &= \left.\frac{\partial^2 \log\{g(\boldsymbol{x}|\boldsymbol{\theta}, M)\}}{\partial \boldsymbol{\theta}\partial \boldsymbol{\theta}'}\right|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}_N}. \end{aligned} \tag{D.51}$$

Hence, $\exp[Q(\boldsymbol{\theta})]$ in (D.47) can be approximated as

$$\begin{aligned}&g(\boldsymbol{x}|\boldsymbol{\theta}, M)p(\boldsymbol{\theta}|M) \\ &\approx g(\boldsymbol{x}|\hat{\boldsymbol{\theta}}_N, M)p(\hat{\boldsymbol{\theta}}_N|M)\exp\left[-\frac{1}{2}(\boldsymbol{\theta}-\hat{\boldsymbol{\theta}}_N)'(N\widehat{\boldsymbol{J}}(\hat{\boldsymbol{\theta}}_N))(\boldsymbol{\theta}-\hat{\boldsymbol{\theta}}_N)\right]\end{aligned} \tag{D.52}$$

where

$$\widehat{\boldsymbol{J}}(\hat{\boldsymbol{\theta}}_N) = -\frac{1}{N}\frac{\partial^2 \log\{g(\boldsymbol{x}|\hat{\boldsymbol{\theta}}_N, M)\}}{\partial \boldsymbol{\theta}\partial \boldsymbol{\theta}'} = \frac{1}{N}\sum_{i=1}^{N}\left[-\frac{\partial^2 \log\{g(x_i|\hat{\boldsymbol{\theta}}_N, M)\}}{\partial \boldsymbol{\theta}\partial \boldsymbol{\theta}'}\right].$$

From (D.52), the marginal likelihood in (D.46) is approximately as follows:

$$g(\boldsymbol{x}|\hat{\boldsymbol{\theta}}_N, M)p(\hat{\boldsymbol{\theta}}_N|M)\int \exp\left[-\frac{1}{2}(\boldsymbol{\theta}-\hat{\boldsymbol{\theta}}_N)'(N\widehat{\boldsymbol{J}}(\hat{\boldsymbol{\theta}}_N))(\boldsymbol{\theta}-\hat{\boldsymbol{\theta}}_N)\right]d\boldsymbol{\theta}.$$

The needed integral is directly related to the underlying K-dimensional multivariate normal distribution and can be evaluated because we know the needed normalizing constant:

$$\int (2\pi)^{-K/2}|N\widehat{\boldsymbol{J}}(\hat{\boldsymbol{\theta}}_N)|^{1/2}\exp\left[-\frac{1}{2}(\boldsymbol{\theta}-\hat{\boldsymbol{\theta}}_N)'(N\widehat{\boldsymbol{J}}(\hat{\boldsymbol{\theta}}_N))(\boldsymbol{\theta}-\hat{\boldsymbol{\theta}}_N)\right]d\boldsymbol{\theta} = 1,$$

where $|\cdot|$ denotes the determinant of matrix. Therefore, we have

$$p(\boldsymbol{x}|M) \approx g(\boldsymbol{x}|\hat{\boldsymbol{\theta}}_N, M)p(\hat{\boldsymbol{\theta}}_N|M)\left[(2\pi)^{K/2}|N\widehat{\boldsymbol{J}}(\hat{\boldsymbol{\theta}}_N)|^{-1/2}\right]$$

$$= g(\boldsymbol{x}|\hat{\boldsymbol{\theta}}_N, M)p(\hat{\boldsymbol{\theta}}_N|M)\left[(2\pi)^{K/2}N^{-K/2}|\widehat{\boldsymbol{J}}(\hat{\boldsymbol{\theta}}_N)|^{-1/2}\right].$$

Taking -2 times the log of the right-hand side above, we have essentially the BIC:

$$-2\log\{g(\boldsymbol{x}|\hat{\boldsymbol{\theta}}_N, M)\}+K\log(N)-2\log\{p(\hat{\boldsymbol{\theta}}_N|M)\}-K\log(2\pi)+\log\{|\widehat{\boldsymbol{J}}(\hat{\boldsymbol{\theta}}_N)|\}.$$

The last three terms of above expression are dropped because $-2\log\{p(\hat{\boldsymbol{\theta}}_N|M)\}$, $K\log(2\pi)$, and $\widehat{\boldsymbol{J}}(\hat{\boldsymbol{\theta}}_N) \xrightarrow{p} \boldsymbol{J}(\boldsymbol{\theta}_0)$ are constants. Therefore BIC can be computed as

$$\text{BIC} = -2\log\{\mathcal{L}_g(\hat{\boldsymbol{\theta}}_N|\boldsymbol{x})\} + K\log(N) \tag{D.53}$$

where $\mathcal{L}_g(\hat{\boldsymbol{\theta}}_N|\boldsymbol{x}) \equiv g(\boldsymbol{x}|\hat{\boldsymbol{\theta}}_N, M)$.

D.4 Likelihood Ratio Test and Asymptotic property of Likelihood Ratio Statistic

Given *i.i.d.* data $\boldsymbol{x} \equiv (x_1, x_2, \ldots, x_N)$, the likelihood function with respect to model $g(\boldsymbol{x}|\boldsymbol{\theta})$ is defined as

$$\mathcal{L}_g(\boldsymbol{\theta}|\boldsymbol{x}) \equiv \prod_{i=1}^{N} g(x_i|\boldsymbol{\theta})$$

where $\boldsymbol{\theta}$ is a K-dimensional parameter.

Considering the null hypothesis for true parameter $\boldsymbol{\theta}_0$:

$$H_0 : \theta_{0,j} = 0 \quad \text{for } 1 \le j \le R, \tag{D.54}$$

we define the $K \times 1$ parameter vector $\boldsymbol{\theta}_0$ and restricted MLE $\hat{\boldsymbol{\theta}}_N^{res}$:

$$\boldsymbol{\theta}_0 = \begin{bmatrix} \mathbf{0}_{R\times 1} \\ \boldsymbol{\theta}_0^* \end{bmatrix}, \quad \hat{\boldsymbol{\theta}}_N^{res} = \begin{bmatrix} \mathbf{0}_{R\times 1} \\ \hat{\boldsymbol{\theta}}_N^* \end{bmatrix} \tag{D.55}$$

where $\mathbf{0}_{R\times 1}$ is a $R \times 1$ zero vector, and $\boldsymbol{\theta}_0^*$, $\hat{\boldsymbol{\theta}}_N^*$ are $(K-R) \times 1$ non-zero vectors.

Given K-dimensional unrestricted MLE $\hat{\boldsymbol{\theta}}_N$, the second-order Taylor series expansion of log-likelihood function $\log\{\mathcal{L}_g(\boldsymbol{\theta}_0|\boldsymbol{x})\} \equiv \sum_{i=1}^N \log\{g(x_i|\boldsymbol{\theta}_0)\}$ around the MLE $\hat{\boldsymbol{\theta}}_N$ shows that

$$\sum_{i=1}^N \log\{g(x_i|\boldsymbol{\theta}_0)\} \approx \sum_{i=1}^N \log\{g(x_i|\hat{\boldsymbol{\theta}}_N)\} + (\boldsymbol{\theta}_0 - \hat{\boldsymbol{\theta}}_N)' \sum_{i=1}^N \left.\frac{\partial \log\{g(x_i|\boldsymbol{\theta})\}}{\partial \boldsymbol{\theta}}\right|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}_N} - \frac{1}{2}\sqrt{N}(\boldsymbol{\theta}_0 - \hat{\boldsymbol{\theta}}_N)' \frac{1}{N}\sum_{i=1}^N \left[-\left.\frac{\partial^2 \log\{g(x_i|\boldsymbol{\theta})\}}{\partial \boldsymbol{\theta}\partial \boldsymbol{\theta}'}\right|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}_N} \right] \sqrt{N}(\boldsymbol{\theta}_0 - \hat{\boldsymbol{\theta}}_N). \tag{D.56}$$

Notice that

$$\sum_{i=1}^N \left.\frac{\partial \log\{g(x_i|\boldsymbol{\theta})\}}{\partial \boldsymbol{\theta}}\right|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}_N} = \mathbf{0}, \tag{D.57}$$

$$\frac{1}{N}\sum_{i=1}^N \left[-\frac{\partial^2 \log\{g(x_i|\hat{\boldsymbol{\theta}}_N)\}}{\partial \boldsymbol{\theta}\partial \boldsymbol{\theta}'} \right] \xrightarrow{p} \mathbf{E}_x \left[-\frac{\partial^2 \log\{g(x|\boldsymbol{\theta}_0)\}}{\partial \boldsymbol{\theta}\partial \boldsymbol{\theta}'} \right] \equiv \boldsymbol{I}(\boldsymbol{\theta}_0), \tag{D.58}$$

then (D.56) can be rewritten as

$$\log\{\mathcal{L}_g(\boldsymbol{\theta}_0|\boldsymbol{x})\} \approx \log\{\mathcal{L}_g(\hat{\boldsymbol{\theta}}_N|\boldsymbol{x})\} - \frac{1}{2}\sqrt{N}(\hat{\boldsymbol{\theta}}_N - \boldsymbol{\theta}_0)'\boldsymbol{I}(\boldsymbol{\theta}_0)\sqrt{N}(\hat{\boldsymbol{\theta}}_N - \boldsymbol{\theta}_0) \tag{D.59}$$

where $\boldsymbol{I}(\boldsymbol{\theta}_0)$ is the Fisher information matrix.

Given the K-dimensional restricted MLE $\hat{\boldsymbol{\theta}}_N^{res}$, the second-order Taylor series expansion of log-likelihood function $\log\{\mathcal{L}_g(\boldsymbol{\theta}_0|\boldsymbol{x})\} \equiv \sum_{i=1}^N \log\{g(x_i|\boldsymbol{\theta}_0)\}$ around the restricted MLE $\hat{\boldsymbol{\theta}}_N^{res}$ shows that

$$\sum_{i=1}^N \log\{g(x_i|\boldsymbol{\theta}_0)\} \approx \sum_{i=1}^N \log\{g(x_i|\hat{\boldsymbol{\theta}}_N^{res})\} + (\boldsymbol{\theta}_0 - \hat{\boldsymbol{\theta}}_N^{res})' \sum_{i=1}^N \left.\frac{\partial \log\{g(x_i|\boldsymbol{\theta})\}}{\partial \boldsymbol{\theta}}\right|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}_N^{res}} - \frac{1}{2}\sqrt{N}(\boldsymbol{\theta}_0 - \hat{\boldsymbol{\theta}}_N^{res})' \frac{1}{N}\sum_{i=1}^N \left[-\left.\frac{\partial^2 \log\{g(x_i|\boldsymbol{\theta})\}}{\partial \boldsymbol{\theta}\partial \boldsymbol{\theta}'}\right|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}_N^{res}} \right] \sqrt{N}(\boldsymbol{\theta}_0 - \hat{\boldsymbol{\theta}}_N^{res}). \tag{D.60}$$

Recall (D.55):

$$\boldsymbol{\theta}_0 = \begin{bmatrix} \mathbf{0}_{R\times 1} \\ \boldsymbol{\theta}_0^* \end{bmatrix} \quad \text{and} \quad \hat{\boldsymbol{\theta}}_N^{res} = \begin{bmatrix} \mathbf{0}_{R\times 1} \\ \hat{\boldsymbol{\theta}}_N^* \end{bmatrix}$$

and let us denote

$$\boldsymbol{\nabla}_R \equiv \begin{bmatrix} \frac{\partial}{\partial\theta_1} \\ \frac{\partial}{\partial\theta_2} \\ \vdots \\ \frac{\partial}{\partial\theta_R} \end{bmatrix} \quad \text{and} \quad \boldsymbol{\nabla}_{K-R} \equiv \begin{bmatrix} \frac{\partial}{\partial\theta_{R+1}} \\ \frac{\partial}{\partial\theta_{R+2}} \\ \vdots \\ \frac{\partial}{\partial\theta_K} \end{bmatrix}.$$

Since the restricted MLE $\hat{\boldsymbol{\theta}}_N^{res}$ does not include parameters $\theta_1, \theta_2, \ldots, \theta_R$ in the model $g(\cdot|\hat{\boldsymbol{\theta}}_N^{res})$, $\boldsymbol{\nabla}_R \log\{g(x_i|\hat{\boldsymbol{\theta}}_N^{res})\}$ is a $R \times 1$ zero vector. Moreover, $(K-R)$-dimensional parameter $\hat{\boldsymbol{\theta}}_N^*$ can be regarded as a solution of vector equation $\sum_{i=1}^N \boldsymbol{\nabla}_{K-R} \log\{g(x_i|\boldsymbol{\theta})\} = \mathbf{0}_{(K-R)\times 1}$. Therefore the first derivative of log-likelihood in (D.60) is zero:

$$\begin{aligned} \sum_{i=1}^N \frac{\partial \log\{g(x_i|\boldsymbol{\theta})\}}{\partial\boldsymbol{\theta}}\bigg|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}_N^{res}} &= \sum_{i=1}^N \begin{bmatrix} \boldsymbol{\nabla}_R \\ \boldsymbol{\nabla}_{K-R} \end{bmatrix} \log\{g(x_i|\boldsymbol{\theta})\}\bigg|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}_N^{res}} \\ &= \begin{bmatrix} \sum_{i=1}^N \boldsymbol{\nabla}_R \log\{g(x_i|\hat{\boldsymbol{\theta}}_N^{res})\} \\ \sum_{i=1}^N \boldsymbol{\nabla}_{K-R} \log\{g(x_i|\hat{\boldsymbol{\theta}}_N^{res})\} \end{bmatrix} \\ &= \begin{bmatrix} \mathbf{0}_{R\times 1} \\ \sum_{i=1}^N \boldsymbol{\nabla}_{K-R} \log\{g(x_i|\hat{\boldsymbol{\theta}}_N^{*})\} \end{bmatrix} \\ &= \mathbf{0}_{K\times 1}. \end{aligned} \tag{D.61}$$

The second derivative of log-likelihood in (D.60) is denoted as

$$\begin{aligned} &\frac{1}{N}\sum_{i=1}^N \left[-\frac{\partial^2 \log\{g(x_i|\boldsymbol{\theta})\}}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}'}\bigg|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}_N^{res}} \right] \\ &= \frac{1}{N}\sum_{i=1}^N \begin{bmatrix} -\boldsymbol{\nabla}_R\boldsymbol{\nabla}_R' \log\{g(x_i|\hat{\boldsymbol{\theta}}_N^{res})\} & -\boldsymbol{\nabla}_R\boldsymbol{\nabla}_{K-R}' \log\{g(x_i|\hat{\boldsymbol{\theta}}_N^{res})\} \\ -\boldsymbol{\nabla}_{K-R}\boldsymbol{\nabla}_R' \log\{g(x_i|\hat{\boldsymbol{\theta}}_N^{res})\} & -\boldsymbol{\nabla}_{K-R}\boldsymbol{\nabla}_{K-R}' \log\{g(x_i|\hat{\boldsymbol{\theta}}_N^{res})\} \end{bmatrix} \\ &\xrightarrow{p} \begin{bmatrix} \mathbf{0}_{R\times R} & \mathbf{0}_{R\times(K-R)} \\ \mathbf{0}_{(K-R)\times R} & \boldsymbol{I}(\boldsymbol{\theta}_0^*) \end{bmatrix} \end{aligned} \tag{D.62}$$

where $\boldsymbol{I}(\boldsymbol{\theta}_0^*)$ is a $(K-R)\times(K-R)$ Fisher information matrix evaluated at parameter $\boldsymbol{\theta}_0^*$.

From (D.61) and (D.62), the second-order Taylor series expansion of log-likelihood in (D.60) can be rewritten as

$$
\begin{aligned}
&\sum_{i=1}^{N}\log\{g(x_i|\boldsymbol{\theta}_0)\}\\
&\approx \sum_{i=1}^{N}\log\{g(x_i|\hat{\boldsymbol{\theta}}_N^{res})\}\\
&\quad -\frac{1}{2}\sqrt{N}(\hat{\boldsymbol{\theta}}_N^{res}-\boldsymbol{\theta}_0)'\begin{bmatrix}\mathbf{0}_{R\times R} & \mathbf{0}_{R\times(K-R)}\\ \mathbf{0}_{(K-R)\times R} & \boldsymbol{I}(\boldsymbol{\theta}_0^*)\end{bmatrix}\sqrt{N}(\hat{\boldsymbol{\theta}}_N^{res}-\boldsymbol{\theta}_0)\\
&= \sum_{i=1}^{N}\log\{g(x_i|\hat{\boldsymbol{\theta}}_N^{res})\}\\
&\quad -\frac{1}{2}\begin{bmatrix}\mathbf{0}'_{R\times 1} & \sqrt{N}(\hat{\boldsymbol{\theta}}_N^{*}-\boldsymbol{\theta}_0^*)'\end{bmatrix}\begin{bmatrix}\mathbf{0}_{R\times R} & \mathbf{0}_{R\times(K-R)}\\ \mathbf{0}_{(K-R)\times R} & \boldsymbol{I}(\boldsymbol{\theta}_0^*)\end{bmatrix}\begin{bmatrix}\mathbf{0}_{R\times 1}\\ \sqrt{N}(\hat{\boldsymbol{\theta}}_N^{*}-\boldsymbol{\theta}_0^*)\end{bmatrix}\\
&= \sum_{i=1}^{N}\log\{g(x_i|\hat{\boldsymbol{\theta}}_N^{res})\}\\
&\quad -\frac{1}{2}\sqrt{N}(\hat{\boldsymbol{\theta}}_N^{*}-\boldsymbol{\theta}_0^*)'\boldsymbol{I}(\boldsymbol{\theta}_0^*)\sqrt{N}(\hat{\boldsymbol{\theta}}_N^{*}-\boldsymbol{\theta}_0^*) \qquad \text{(D.63)}
\end{aligned}
$$

then (D.63) can be rewritten as

$$\log\{\mathcal{L}_g(\boldsymbol{\theta}_0|x)\}\approx\log\{\mathcal{L}_g(\hat{\boldsymbol{\theta}}_N^{res}|x)\}-\frac{1}{2}\sqrt{N}(\hat{\boldsymbol{\theta}}_N^{*}-\boldsymbol{\theta}_0^*)'\boldsymbol{I}(\boldsymbol{\theta}_0^*)\sqrt{N}(\hat{\boldsymbol{\theta}}_N^{*}-\boldsymbol{\theta}_0^*). \qquad \text{(D.64)}$$

Recall (D.59) and (D.64):

$$
\begin{aligned}
\log\{\mathcal{L}_g(\boldsymbol{\theta}_0|x)\}&\approx\log\{\mathcal{L}_g(\hat{\boldsymbol{\theta}}_N|x)\}-\frac{1}{2}\sqrt{N}(\hat{\boldsymbol{\theta}}_N-\boldsymbol{\theta}_0)'\boldsymbol{I}(\boldsymbol{\theta}_0)\sqrt{N}(\hat{\boldsymbol{\theta}}_N-\boldsymbol{\theta}_0),\\
\log\{\mathcal{L}_g(\boldsymbol{\theta}_0|x)\}&\approx\log\{\mathcal{L}_g(\hat{\boldsymbol{\theta}}_N^{res}|x)\}-\frac{1}{2}\sqrt{N}(\hat{\boldsymbol{\theta}}_N^{*}-\boldsymbol{\theta}_0^*)'\boldsymbol{I}(\boldsymbol{\theta}_0^*)\sqrt{N}(\hat{\boldsymbol{\theta}}_N^{*}-\boldsymbol{\theta}_0^*),
\end{aligned}
$$

then subtracting (D.64) from (D.59), we have

$$\log\{\mathcal{L}_g(\hat{\boldsymbol{\theta}}_N^{res}|x)\}-\log\{\mathcal{L}_g(\hat{\boldsymbol{\theta}}_N|x)\}\approx-\frac{1}{2}\sqrt{N}(\hat{\boldsymbol{\theta}}_N-\boldsymbol{\theta}_0)'\boldsymbol{I}(\boldsymbol{\theta}_0)\sqrt{N}(\hat{\boldsymbol{\theta}}_N-\boldsymbol{\theta}_0)$$

$$+\frac{1}{2}\sqrt{N}(\hat{\boldsymbol{\theta}}_N^* - \boldsymbol{\theta}_0^*)'\boldsymbol{I}(\boldsymbol{\theta}_0^*)\sqrt{N}(\hat{\boldsymbol{\theta}}_N^* - \boldsymbol{\theta}_0^*). \tag{D.65}$$

Therefore the left-hand side in (D.65) can be rewritten as the likelihood ratio statistic by multiplying -2 to the both sides in (D.65), and then we need to evaluate the asymptotic distribution of the right-hand side.

$$\begin{aligned}-2\log\left[\frac{\mathcal{L}_g(\hat{\boldsymbol{\theta}}_N^{res}|x)}{\mathcal{L}_g(\hat{\boldsymbol{\theta}}_N|x)}\right] &\approx \sqrt{N}(\hat{\boldsymbol{\theta}}_N - \boldsymbol{\theta}_0)'\boldsymbol{I}(\boldsymbol{\theta}_0)\sqrt{N}(\hat{\boldsymbol{\theta}}_N - \boldsymbol{\theta}_0) \\ &\quad - \sqrt{N}(\hat{\boldsymbol{\theta}}_N^* - \boldsymbol{\theta}_0^*)'\boldsymbol{I}(\boldsymbol{\theta}_0^*)\sqrt{N}(\hat{\boldsymbol{\theta}}_N^* - \boldsymbol{\theta}_0^*). \end{aligned} \tag{D.66}$$

Redefine the Fisher information matrix $\boldsymbol{I}(\boldsymbol{\theta}_0)$ as

$$\begin{aligned}&\frac{1}{N}\sum_{i=1}^{N}\begin{bmatrix} -\boldsymbol{\nabla}_R\boldsymbol{\nabla}_R'\log\{g(x_i|\hat{\boldsymbol{\theta}}_N)\} & -\boldsymbol{\nabla}_R\boldsymbol{\nabla}_{K-R}'\log\{g(x_i|\hat{\boldsymbol{\theta}}_N)\} \\ -\boldsymbol{\nabla}_{K-R}\boldsymbol{\nabla}_R'\log\{g(x_i|\hat{\boldsymbol{\theta}}_N)\} & -\boldsymbol{\nabla}_{K-R}\boldsymbol{\nabla}_{K-R}'\log\{g(x_i|\hat{\boldsymbol{\theta}}_N)\}\end{bmatrix} \\ &\xrightarrow{p}\begin{bmatrix} \mathbf{E}_x\left[-\boldsymbol{\nabla}_R\boldsymbol{\nabla}_R'\log\{g(x|\boldsymbol{\theta}_0)\}\right] & \mathbf{E}_x\left[-\boldsymbol{\nabla}_R\boldsymbol{\nabla}_{K-R}'\log\{g(x|\boldsymbol{\theta}_0)\}\right] \\ \mathbf{E}_x\left[-\boldsymbol{\nabla}_{K-R}\boldsymbol{\nabla}_R'\log\{g(x|\boldsymbol{\theta}_0)\}\right] & \mathbf{E}_x\left[-\boldsymbol{\nabla}_{K-R}\boldsymbol{\nabla}_{K-R}'\log\{g(x|\boldsymbol{\theta}_0)\}\right]\end{bmatrix} \\ &=\begin{bmatrix}\boldsymbol{A}(\boldsymbol{\theta}_0^*) & \boldsymbol{B}(\boldsymbol{\theta}_0^*) \\ \boldsymbol{B}'(\boldsymbol{\theta}_0^*) & \boldsymbol{I}(\boldsymbol{\theta}_0^*)\end{bmatrix} \\ &\equiv \boldsymbol{I}(\boldsymbol{\theta}_0)\end{aligned} \tag{D.67}$$

where $\boldsymbol{A}(\boldsymbol{\theta}_0^*)$ is a $R \times R$ matrix and $\boldsymbol{B}(\boldsymbol{\theta}_0^*)$ is a $R \times (K-R)$ matrix.

Let us denote the score function as

$$\frac{1}{\sqrt{N}}\sum_{i=1}^{N}\frac{\partial\log\{g(x_i|\boldsymbol{\theta})\}}{\partial\boldsymbol{\theta}}\bigg|_{\boldsymbol{\theta}=\boldsymbol{\theta}_0} \equiv \begin{bmatrix}\boldsymbol{s}_r(\boldsymbol{\theta}_0) \\ \boldsymbol{s}_u(\boldsymbol{\theta}_0)\end{bmatrix} \tag{D.68}$$

where $\boldsymbol{s}_r(\boldsymbol{\theta}_0)$ is a $R \times 1$ vector and $\boldsymbol{s}_u(\boldsymbol{\theta}_0)$ is a $(K-R) \times 1$ vector. Under the null hypothesis $H_0 : \theta_{0,j}$ for $1 \le j \le R$, the Taylor series expansion of the first derivative of log-likelihood divided by $\sqrt{N}$ shows that

$$\frac{1}{\sqrt{N}}\sum_{i=1}^{N}\frac{\partial\log\{g(x_i|\hat{\boldsymbol{\theta}}_N)\}}{\partial\boldsymbol{\theta}} \approx \frac{1}{\sqrt{N}}\sum_{i=1}^{N}\frac{\partial\log\{g(x_i|\boldsymbol{\theta}_0)\}}{\partial\boldsymbol{\theta}}$$

$$- \left[-\frac{1}{N} \sum_{i=1}^{N} \frac{\partial^2 \log\{g(x_i|\boldsymbol{\theta}_0)\}}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}'} \right] \sqrt{N}(\hat{\boldsymbol{\theta}}_N - \boldsymbol{\theta}_0). \tag{D.69}$$

As the number of samples N is large, (D.69) can be rewritten as

$$\mathbf{0}_{K\times 1} \approx \begin{bmatrix} \boldsymbol{s}_r(\boldsymbol{\theta}_0) \\ \boldsymbol{s}_u(\boldsymbol{\theta}_0) \end{bmatrix} - \boldsymbol{I}(\boldsymbol{\theta}_0)\sqrt{N}(\hat{\boldsymbol{\theta}}_N - \boldsymbol{\theta}_0) \tag{D.70}$$

then we have

$$\begin{bmatrix} \boldsymbol{s}_r(\boldsymbol{\theta}_0) \\ \boldsymbol{s}_u(\boldsymbol{\theta}_0) \end{bmatrix} \approx \boldsymbol{I}(\boldsymbol{\theta}_0)\sqrt{N}(\hat{\boldsymbol{\theta}}_N - \boldsymbol{\theta}_0). \tag{D.71}$$

As for restricted MLE $\hat{\boldsymbol{\theta}}_N^{res}$, under the null hypothesis $H_0 : \theta_{0,j}$ for $1 \leq j \leq R$, the Taylor series expansion of the first derivative of log-likelihood divided by $\sqrt{N}$ shows that

$$\begin{aligned} \frac{1}{\sqrt{N}} \sum_{i=1}^{N} \frac{\partial \log\{g(x_i|\hat{\boldsymbol{\theta}}_N^{res})\}}{\partial \boldsymbol{\theta}} &\approx \frac{1}{\sqrt{N}} \sum_{i=1}^{N} \frac{\partial \log\{g(x_i|\boldsymbol{\theta}_0)\}}{\partial \boldsymbol{\theta}} \\ &\quad - \left[-\frac{1}{N} \sum_{i=1}^{N} \frac{\partial^2 \log\{g(x_i|\boldsymbol{\theta}_0)\}}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}'} \right] \sqrt{N}(\hat{\boldsymbol{\theta}}_N^{res} - \boldsymbol{\theta}_0) \\ &\approx \frac{1}{\sqrt{N}} \sum_{i=1}^{N} \frac{\partial \log\{g(x_i|\boldsymbol{\theta}_0)\}}{\partial \boldsymbol{\theta}} \\ &\quad - \sqrt{N} \begin{bmatrix} \boldsymbol{A}(\boldsymbol{\theta}_0^*) & \boldsymbol{B}(\boldsymbol{\theta}_0^*) \\ \boldsymbol{B}'(\boldsymbol{\theta}_0^*) & \boldsymbol{I}(\boldsymbol{\theta}_0^*) \end{bmatrix} \begin{bmatrix} \mathbf{0}_{R\times 1} \\ \hat{\boldsymbol{\theta}}_N^* - \boldsymbol{\theta}_0^* \end{bmatrix}. \end{aligned} \tag{D.72}$$

Since the left-hand side of (D.72) is zero, (D.72) can be rewritten as

$$\begin{bmatrix} \boldsymbol{s}_r(\boldsymbol{\theta}_0) \\ \boldsymbol{s}_u(\boldsymbol{\theta}_0) \end{bmatrix} \approx \begin{bmatrix} \boldsymbol{B}(\boldsymbol{\theta}_0^*)\sqrt{N}(\hat{\boldsymbol{\theta}}_N^* - \boldsymbol{\theta}_0^*) \\ \boldsymbol{I}(\boldsymbol{\theta}_0^*)\sqrt{N}(\hat{\boldsymbol{\theta}}_N^* - \boldsymbol{\theta}_0^*) \end{bmatrix}, \tag{D.73}$$

hence we have

$$\boldsymbol{s}_r(\boldsymbol{\theta}_0) \approx \boldsymbol{B}(\boldsymbol{\theta}_0^*)\sqrt{N}(\hat{\boldsymbol{\theta}}_N^* - \boldsymbol{\theta}_0^*)$$

$$\approx \boldsymbol{B}(\boldsymbol{\theta}_0^*)\boldsymbol{I}^{-1}(\boldsymbol{\theta}_0^*)\boldsymbol{s}_u(\boldsymbol{\theta}_0). \tag{D.74}$$

From (D.71) and (D.73), we can denote the first two terms on the right-hand side of (D.66) as follows

$$\begin{aligned}\sqrt{N}(\hat{\boldsymbol{\theta}}_N - \boldsymbol{\theta}_0)'\boldsymbol{I}(\boldsymbol{\theta}_0)\sqrt{N}(\hat{\boldsymbol{\theta}}_N - \boldsymbol{\theta}_0) &\approx \left\{ \boldsymbol{I}^{-1}(\boldsymbol{\theta}_0) \begin{bmatrix} \boldsymbol{s}_r(\boldsymbol{\theta}_0) \\ \boldsymbol{s}_u(\boldsymbol{\theta}_0) \end{bmatrix} \right\}' \begin{bmatrix} \boldsymbol{s}_r(\boldsymbol{\theta}_0) \\ \boldsymbol{s}_u(\boldsymbol{\theta}_0) \end{bmatrix} \\ &= \begin{bmatrix} \boldsymbol{s}_r(\boldsymbol{\theta}_0) \\ \boldsymbol{s}_u(\boldsymbol{\theta}_0) \end{bmatrix}' \boldsymbol{I}^{-1}(\boldsymbol{\theta}_0) \begin{bmatrix} \boldsymbol{s}_r(\boldsymbol{\theta}_0) \\ \boldsymbol{s}_u(\boldsymbol{\theta}_0) \end{bmatrix},\end{aligned} \tag{D.75}$$

$$\begin{aligned}\sqrt{N}(\hat{\boldsymbol{\theta}}_N^* - \boldsymbol{\theta}_0^*)'\boldsymbol{I}(\boldsymbol{\theta}_0^*)\sqrt{N}(\hat{\boldsymbol{\theta}}_N^* - \boldsymbol{\theta}_0^*) &\approx \left[\boldsymbol{I}^{-1}(\boldsymbol{\theta}_0^*)\boldsymbol{s}_u(\boldsymbol{\theta}_0)\right]' \boldsymbol{I}(\boldsymbol{\theta}_0^*) \left[\boldsymbol{I}^{-1}(\boldsymbol{\theta}_0^*)\boldsymbol{s}_u(\boldsymbol{\theta}_0)\right] \\ &= \boldsymbol{s}_u'(\boldsymbol{\theta}_0)\boldsymbol{I}^{-1}(\boldsymbol{\theta}_0^*)\boldsymbol{s}_u(\boldsymbol{\theta}_0)\end{aligned} \tag{D.76}$$

where Fisher information matrix $\boldsymbol{I}(\boldsymbol{\theta}_0)$ is symmetric and invertible. Using (D.74), vector of score function in (D.75) is rewritten as

$$\begin{aligned}\begin{bmatrix} \boldsymbol{s}_r(\boldsymbol{\theta}_0) \\ \boldsymbol{s}_u(\boldsymbol{\theta}_0) \end{bmatrix} &\approx \left[\begin{pmatrix} \boldsymbol{s}_r(\boldsymbol{\theta}_0) - \boldsymbol{B}(\boldsymbol{\theta}_0^*)\boldsymbol{I}^{-1}(\boldsymbol{\theta}_0^*)\boldsymbol{s}_u(\boldsymbol{\theta}_0) \\ \boldsymbol{0}_{(K-R)\times 1} \end{pmatrix} + \begin{pmatrix} \boldsymbol{B}(\boldsymbol{\theta}_0^*)\boldsymbol{I}^{-1}(\boldsymbol{\theta}_0^*)\boldsymbol{s}_u(\boldsymbol{\theta}_0) \\ \boldsymbol{s}_u(\boldsymbol{\theta}_0) \end{pmatrix} \right] \\ &= [\boldsymbol{s}_a(\boldsymbol{\theta}_0) + \boldsymbol{s}_b(\boldsymbol{\theta}_0)]\end{aligned} \tag{D.77}$$

and inverse matrix of $\boldsymbol{I}(\boldsymbol{\theta}_0)$ is expressed as

$$\begin{aligned}&\boldsymbol{I}^{-1}(\boldsymbol{\theta}_0) \\ &= \begin{bmatrix} \boldsymbol{D}(\boldsymbol{\theta}_0^*) & -\boldsymbol{D}(\boldsymbol{\theta}_0^*)\boldsymbol{B}(\boldsymbol{\theta}_0^*)\boldsymbol{I}^{-1}(\boldsymbol{\theta}_0^*) \\ -\boldsymbol{I}^{-1}(\boldsymbol{\theta}_0^*)\boldsymbol{B}'(\boldsymbol{\theta}_0^*)\boldsymbol{D}(\boldsymbol{\theta}_0^*) & \boldsymbol{I}^{-1}(\boldsymbol{\theta}_0^*) + \boldsymbol{I}^{-1}(\boldsymbol{\theta}_0^*)\boldsymbol{B}'(\boldsymbol{\theta}_0^*)\boldsymbol{D}(\boldsymbol{\theta}_0^*)\boldsymbol{B}(\boldsymbol{\theta}_0^*)\boldsymbol{I}^{-1}(\boldsymbol{\theta}_0^*) \end{bmatrix}\end{aligned} \tag{D.78}$$

where $\boldsymbol{D}(\boldsymbol{\theta}_0^*) \equiv \left(\boldsymbol{A}(\boldsymbol{\theta}_0^*) - \boldsymbol{B}(\boldsymbol{\theta}_0^*)\boldsymbol{I}^{-1}(\boldsymbol{\theta}_0^*)\boldsymbol{B}'(\boldsymbol{\theta}_0^*)\right)^{-1}$.

From (D.75) and (D.77), we have

$$\begin{bmatrix} \boldsymbol{s}_r(\boldsymbol{\theta}_0) \\ \boldsymbol{s}_u(\boldsymbol{\theta}_0) \end{bmatrix}' \boldsymbol{I}^{-1}(\boldsymbol{\theta}_0) \begin{bmatrix} \boldsymbol{s}_r(\boldsymbol{\theta}_0) \\ \boldsymbol{s}_u(\boldsymbol{\theta}_0) \end{bmatrix} = [\boldsymbol{s}_a'(\boldsymbol{\theta}_0) + \boldsymbol{s}_b'(\boldsymbol{\theta}_0)]\, \boldsymbol{I}^{-1}(\boldsymbol{\theta}_0)\, [\boldsymbol{s}_a(\boldsymbol{\theta}_0) + \boldsymbol{s}_b(\boldsymbol{\theta}_0)]$$

$$
\begin{aligned}
&= \boldsymbol{s}_a'(\boldsymbol{\theta}_0)\boldsymbol{I}^{-1}(\boldsymbol{\theta}_0)\boldsymbol{s}_a(\boldsymbol{\theta}_0) + \boldsymbol{s}_a'(\boldsymbol{\theta}_0)\boldsymbol{I}^{-1}(\boldsymbol{\theta}_0)\boldsymbol{s}_b(\boldsymbol{\theta}_0) \\
&\quad + \boldsymbol{s}_b'(\boldsymbol{\theta}_0)\boldsymbol{I}^{-1}(\boldsymbol{\theta}_0)\boldsymbol{s}_a(\boldsymbol{\theta}_0) + \boldsymbol{s}_b'(\boldsymbol{\theta}_0)\boldsymbol{I}^{-1}(\boldsymbol{\theta}_0)\boldsymbol{s}_b(\boldsymbol{\theta}_0).
\end{aligned} \tag{D.79}
$$

Using the fact that

$$
\begin{aligned}
&\left[\begin{array}{cc} \boldsymbol{B}'(\boldsymbol{\theta}_0^*) & \boldsymbol{I}(\boldsymbol{\theta}_0^*) \end{array}\right] \boldsymbol{I}^{-1}(\boldsymbol{\theta}_0) \\
&= \left[\begin{array}{cc} \boldsymbol{B}'(\boldsymbol{\theta}_0^*) & \boldsymbol{I}(\boldsymbol{\theta}_0^*) \end{array}\right] \\
&\qquad \times \left[\begin{array}{cc} \boldsymbol{D}(\boldsymbol{\theta}_0^*) & -\boldsymbol{D}(\boldsymbol{\theta}_0^*)\boldsymbol{B}(\boldsymbol{\theta}_0^*)\boldsymbol{I}^{-1}(\boldsymbol{\theta}_0^*) \\ -\boldsymbol{I}^{-1}(\boldsymbol{\theta}_0^*)\boldsymbol{B}'(\boldsymbol{\theta}_0^*)\boldsymbol{D}(\boldsymbol{\theta}_0^*) & \boldsymbol{I}^{-1}(\boldsymbol{\theta}_0^*) + \boldsymbol{I}^{-1}(\boldsymbol{\theta}_0^*)\boldsymbol{B}'(\boldsymbol{\theta}_0^*)\boldsymbol{D}(\boldsymbol{\theta}_0^*)\boldsymbol{B}(\boldsymbol{\theta}_0^*)\boldsymbol{I}^{-1}(\boldsymbol{\theta}_0^*) \end{array}\right] \\
&= \left[\begin{array}{cc} \boldsymbol{0}_{(K-R)\times R} & \boldsymbol{I}_{K-R} \end{array}\right],
\end{aligned} \tag{D.80}
$$

we have

$$
\begin{aligned}
\boldsymbol{s}_b'(\boldsymbol{\theta}_0)\boldsymbol{I}^{-1}(\boldsymbol{\theta}_0) &= \left[\begin{array}{cc} \left\{\boldsymbol{B}(\boldsymbol{\theta}_0^*)\boldsymbol{I}^{-1}(\boldsymbol{\theta}_0^*)\boldsymbol{s}_u(\boldsymbol{\theta}_0)\right\}' & \boldsymbol{s}_u'(\boldsymbol{\theta}_0) \end{array}\right] \boldsymbol{I}^{-1}(\boldsymbol{\theta}_0) \\
&= \left[\begin{array}{cc} \left\{\boldsymbol{I}^{-1}(\boldsymbol{\theta}_0^*)\boldsymbol{s}_u(\boldsymbol{\theta}_0)\right\}'\boldsymbol{B}'(\boldsymbol{\theta}_0^*) & \boldsymbol{s}_u'(\boldsymbol{\theta}_0) \end{array}\right] \boldsymbol{I}^{-1}(\boldsymbol{\theta}_0) \\
&= \left[\boldsymbol{I}^{-1}(\boldsymbol{\theta}_0^*)\boldsymbol{s}_u(\boldsymbol{\theta}_0)\right]' \left[\begin{array}{cc} \boldsymbol{B}'(\boldsymbol{\theta}_0^*) & \boldsymbol{I}(\boldsymbol{\theta}_0^*) \end{array}\right] \boldsymbol{I}^{-1}(\boldsymbol{\theta}_0) \\
&= \left[\boldsymbol{I}^{-1}(\boldsymbol{\theta}_0^*)\boldsymbol{s}_u(\boldsymbol{\theta}_0)\right]' \left[\begin{array}{cc} \boldsymbol{0}_{(K-R)\times R} & \boldsymbol{I}_{K-R} \end{array}\right] \\
&= \left[\begin{array}{cc} \boldsymbol{0}_{1\times R} & \boldsymbol{s}_u'(\boldsymbol{\theta}_0)\boldsymbol{I}^{-1}(\boldsymbol{\theta}_0^*) \end{array}\right].
\end{aligned} \tag{D.81}
$$

Hence we notice that

$$
\begin{aligned}
\boldsymbol{I}^{-1}(\boldsymbol{\theta}_0)\boldsymbol{s}_b(\boldsymbol{\theta}_0) &= \left[\boldsymbol{s}_b'(\boldsymbol{\theta}_0)\boldsymbol{I}^{-1}(\boldsymbol{\theta}_0)\right]' \\
&= \left[\begin{array}{c} \boldsymbol{0}_{R\times 1} \\ \boldsymbol{I}^{-1}(\boldsymbol{\theta}_0^*)\boldsymbol{s}_u(\boldsymbol{\theta}_0) \end{array}\right].
\end{aligned} \tag{D.82}
$$

From (D.81) and (D.82), the cross product terms on the right-hand side of (D.79) are calculated as follows:

$$
\boldsymbol{s}_a'(\boldsymbol{\theta}_0)\boldsymbol{I}^{-1}(\boldsymbol{\theta}_0)\boldsymbol{s}_b(\boldsymbol{\theta}_0)
$$

$$= \begin{bmatrix} \boldsymbol{s}_r(\boldsymbol{\theta}_0) - \boldsymbol{B}(\boldsymbol{\theta}_0^*)\boldsymbol{I}^{-1}(\boldsymbol{\theta}_0^*)\boldsymbol{s}_u(\boldsymbol{\theta}_0) \\ \mathbf{0}_{(K-R)\times 1} \end{bmatrix}' \begin{bmatrix} \mathbf{0}_{R\times 1} \\ \boldsymbol{I}^{-1}(\boldsymbol{\theta}_0^*)\boldsymbol{s}_u(\boldsymbol{\theta}_0) \end{bmatrix}$$

$$= 0, \tag{D.83}$$

$$\boldsymbol{s}_b'(\boldsymbol{\theta}_0)\boldsymbol{I}^{-1}(\boldsymbol{\theta}_0)\boldsymbol{s}_a(\boldsymbol{\theta}_0)$$

$$= \begin{bmatrix} \mathbf{0}_{1\times R} & \boldsymbol{s}_u'(\boldsymbol{\theta}_0)\boldsymbol{I}^{-1}(\boldsymbol{\theta}_0^*) \end{bmatrix} \begin{bmatrix} \boldsymbol{s}_r(\boldsymbol{\theta}_0) - \boldsymbol{B}(\boldsymbol{\theta}_0^*)\boldsymbol{I}^{-1}(\boldsymbol{\theta}_0^*)\boldsymbol{s}_u(\boldsymbol{\theta}_0) \\ \mathbf{0}_{(K-R)\times 1} \end{bmatrix}$$

$$= 0. \tag{D.84}$$

The other terms on the right-hand side of (D.79) are computed as follows:

$$\begin{aligned} &\boldsymbol{s}_a'(\boldsymbol{\theta}_0)\boldsymbol{I}^{-1}(\boldsymbol{\theta}_0)\boldsymbol{s}_a(\boldsymbol{\theta}_0) \\ &= \begin{bmatrix} \boldsymbol{s}_r(\boldsymbol{\theta}_0) - \boldsymbol{B}(\boldsymbol{\theta}_0^*)\boldsymbol{I}^{-1}(\boldsymbol{\theta}_0^*)\boldsymbol{s}_u(\boldsymbol{\theta}_0) \\ \mathbf{0}_{(K-R)\times 1} \end{bmatrix}' \\ &\quad \times \begin{bmatrix} \boldsymbol{D}(\boldsymbol{\theta}_0^*) & -\boldsymbol{D}(\boldsymbol{\theta}_0^*)\boldsymbol{B}(\boldsymbol{\theta}_0^*)\boldsymbol{I}^{-1}(\boldsymbol{\theta}_0^*) \\ -\boldsymbol{I}^{-1}(\boldsymbol{\theta}_0^*)\boldsymbol{B}'(\boldsymbol{\theta}_0^*)\boldsymbol{D}(\boldsymbol{\theta}_0^*) & \boldsymbol{I}^{-1}(\boldsymbol{\theta}_0^*) + \boldsymbol{I}^{-1}(\boldsymbol{\theta}_0^*)\boldsymbol{B}'(\boldsymbol{\theta}_0^*)\boldsymbol{D}(\boldsymbol{\theta}_0^*)\boldsymbol{B}(\boldsymbol{\theta}_0^*)\boldsymbol{I}^{-1}(\boldsymbol{\theta}_0^*) \end{bmatrix} \\ &\quad \times \begin{bmatrix} \boldsymbol{s}_r(\boldsymbol{\theta}_0) - \boldsymbol{B}(\boldsymbol{\theta}_0^*)\boldsymbol{I}^{-1}(\boldsymbol{\theta}_0^*)\boldsymbol{s}_u(\boldsymbol{\theta}_0) \\ \mathbf{0}_{(K-R)\times 1} \end{bmatrix} \\ &= \left[\boldsymbol{s}_r(\boldsymbol{\theta}_0) - \boldsymbol{B}(\boldsymbol{\theta}_0^*)\boldsymbol{I}^{-1}(\boldsymbol{\theta}_0^*)\boldsymbol{s}_u(\boldsymbol{\theta}_0)\right]' \boldsymbol{D}(\boldsymbol{\theta}_0^*) \left[\boldsymbol{s}_r(\boldsymbol{\theta}_0) - \boldsymbol{B}(\boldsymbol{\theta}_0^*)\boldsymbol{I}^{-1}(\boldsymbol{\theta}_0^*)\boldsymbol{s}_u(\boldsymbol{\theta}_0)\right], \end{aligned} \tag{D.85}$$

$$\boldsymbol{s}_b'(\boldsymbol{\theta}_0)\boldsymbol{I}^{-1}(\boldsymbol{\theta}_0)\boldsymbol{s}_b(\boldsymbol{\theta}_0)$$

$$= \begin{bmatrix} \mathbf{0}_{1\times R} & \boldsymbol{s}_u'(\boldsymbol{\theta}_0)\boldsymbol{I}^{-1}(\boldsymbol{\theta}_0^*) \end{bmatrix} \begin{bmatrix} \boldsymbol{B}(\boldsymbol{\theta}_0^*)\boldsymbol{I}^{-1}(\boldsymbol{\theta}_0^*)\boldsymbol{s}_u(\boldsymbol{\theta}_0) \\ \boldsymbol{s}_u(\boldsymbol{\theta}_0) \end{bmatrix}$$

$$= \boldsymbol{s}_u'(\boldsymbol{\theta}_0)\boldsymbol{I}^{-1}(\boldsymbol{\theta}_0^*)\boldsymbol{s}_u(\boldsymbol{\theta}_0). \tag{D.86}$$

From (D.83), (D.84), (D.85) and (D.86), (D.75) can be approximated to

$$\sqrt{N}(\hat{\boldsymbol{\theta}}_N - \boldsymbol{\theta}_0)'\boldsymbol{I}(\boldsymbol{\theta}_0)\sqrt{N}(\hat{\boldsymbol{\theta}}_N - \boldsymbol{\theta}_0)$$

$$\approx \left[\boldsymbol{s}_r(\boldsymbol{\theta}_0) - \boldsymbol{B}(\boldsymbol{\theta}_0^*)\boldsymbol{I}^{-1}(\boldsymbol{\theta}_0^*)\boldsymbol{s}_u(\boldsymbol{\theta}_0)\right]' \boldsymbol{D}(\boldsymbol{\theta}_0^*) \left[\boldsymbol{s}_r(\boldsymbol{\theta}_0) - \boldsymbol{B}(\boldsymbol{\theta}_0^*)\boldsymbol{I}^{-1}(\boldsymbol{\theta}_0^*)\boldsymbol{s}_u(\boldsymbol{\theta}_0)\right]$$
$$+ \boldsymbol{s}_u'(\boldsymbol{\theta}_0)\boldsymbol{I}^{-1}(\boldsymbol{\theta}_0^*)\boldsymbol{s}_u(\boldsymbol{\theta}_0). \tag{D.87}$$

Recall (D.76):

$$\sqrt{N}(\hat{\boldsymbol{\theta}}_N^* - \boldsymbol{\theta}_0^*)'\boldsymbol{I}(\boldsymbol{\theta}_0^*)\sqrt{N}(\hat{\boldsymbol{\theta}}_N^* - \boldsymbol{\theta}_0^*) \approx \left[\boldsymbol{I}^{-1}(\boldsymbol{\theta}_0^*)\boldsymbol{s}_u(\boldsymbol{\theta}_0)\right]' \boldsymbol{I}(\boldsymbol{\theta}_0^*) \left[\boldsymbol{I}^{-1}(\boldsymbol{\theta}_0^*)\boldsymbol{s}_u(\boldsymbol{\theta}_0)\right]$$
$$= \boldsymbol{s}_u'(\boldsymbol{\theta}_0)\boldsymbol{I}^{-1}(\boldsymbol{\theta}_0^*)\boldsymbol{s}_u(\boldsymbol{\theta}_0).$$

From (D.76) and (D.87), (D.66) is expressed as

$$-2\log\left[\frac{\mathcal{L}_g(\hat{\boldsymbol{\theta}}_N^{res}|x)}{\mathcal{L}_g(\hat{\boldsymbol{\theta}}_N|x)}\right]$$
$$\approx \left[\boldsymbol{s}_r(\boldsymbol{\theta}_0) - \boldsymbol{B}(\boldsymbol{\theta}_0^*)\boldsymbol{I}^{-1}(\boldsymbol{\theta}_0^*)\boldsymbol{s}_u(\boldsymbol{\theta}_0)\right]' \boldsymbol{D}(\boldsymbol{\theta}_0^*) \left[\boldsymbol{s}_r(\boldsymbol{\theta}_0) - \boldsymbol{B}(\boldsymbol{\theta}_0^*)\boldsymbol{I}^{-1}(\boldsymbol{\theta}_0^*)\boldsymbol{s}_u(\boldsymbol{\theta}_0)\right] \tag{D.88}$$

where

$$\boldsymbol{D}^{-1}(\boldsymbol{\theta}_0^*) \equiv \boldsymbol{A}(\boldsymbol{\theta}_0^*) - \boldsymbol{B}(\boldsymbol{\theta}_0^*)\boldsymbol{I}^{-1}(\boldsymbol{\theta}_0^*)\boldsymbol{B}'(\boldsymbol{\theta}_0^*). \tag{D.89}$$

From the central limit theorem for score function in (D.68), we have

$$\frac{1}{\sqrt{N}}\sum_{i=1}^{N} \frac{\partial \log\{g(x_i|\boldsymbol{\theta})\}}{\partial \boldsymbol{\theta}}\Big|_{\boldsymbol{\theta}=\boldsymbol{\theta}_0} \equiv \begin{bmatrix} \boldsymbol{s}_r(\boldsymbol{\theta}_0) \\ \boldsymbol{s}_u(\boldsymbol{\theta}_0) \end{bmatrix} \sim \mathcal{N}\left(\mathbf{0}_{K\times 1}, \boldsymbol{I}(\boldsymbol{\theta}_0)\right). \tag{D.90}$$

Considering a $R \times K$ nonrandom matrix $\boldsymbol{C}$:

$$\boldsymbol{C}(\boldsymbol{\theta}_0^*) \equiv \left[\begin{array}{cc} \boldsymbol{I}_R & -\boldsymbol{B}(\boldsymbol{\theta}_0^*)\boldsymbol{I}^{-1}(\boldsymbol{\theta}_0^*) \end{array}\right], \tag{D.91}$$

then we know that

$$\boldsymbol{C}(\boldsymbol{\theta}_0^*) \begin{bmatrix} \boldsymbol{s}_r(\boldsymbol{\theta}_0) \\ \boldsymbol{s}_u(\boldsymbol{\theta}_0) \end{bmatrix} = \left[\boldsymbol{s}_r(\boldsymbol{\theta}_0) - \boldsymbol{B}(\boldsymbol{\theta}_0^*)\boldsymbol{I}^{-1}(\boldsymbol{\theta}_0^*)\boldsymbol{s}_u(\boldsymbol{\theta}_0)\right]$$
$$\sim \mathcal{N}\left(\mathbf{0}_{R\times 1}, \boldsymbol{C}(\boldsymbol{\theta}_0^*)\boldsymbol{I}(\boldsymbol{\theta}_0)\boldsymbol{C}'(\boldsymbol{\theta}_0)\right) \tag{D.92}$$

where the Fisher information matrix $\boldsymbol{I}(\boldsymbol{\theta}_0)$:

$$\boldsymbol{I}(\boldsymbol{\theta}_0) \equiv \begin{bmatrix} \boldsymbol{A}(\boldsymbol{\theta}_0^*) & \boldsymbol{B}(\boldsymbol{\theta}_0^*) \\ \boldsymbol{B}'(\boldsymbol{\theta}_0^*) & \boldsymbol{I}(\boldsymbol{\theta}_0^*) \end{bmatrix}$$

and

$$
\begin{aligned}
&\boldsymbol{C}(\boldsymbol{\theta}_0^*)\boldsymbol{I}(\boldsymbol{\theta}_0)\boldsymbol{C}'(\boldsymbol{\theta}_0^*) \\
&= \left[\begin{array}{cc} \boldsymbol{I}_R & -\boldsymbol{B}(\boldsymbol{\theta}_0^*)\boldsymbol{I}^{-1}(\boldsymbol{\theta}_0^*) \end{array}\right] \left[\begin{array}{cc} \boldsymbol{A}(\boldsymbol{\theta}_0^*) & \boldsymbol{B}(\boldsymbol{\theta}_0^*) \\ \boldsymbol{B}'(\boldsymbol{\theta}_0^*) & \boldsymbol{I}(\boldsymbol{\theta}_0^*) \end{array}\right] \left[\begin{array}{c} \boldsymbol{I}_R \\ \left\{-\boldsymbol{B}(\boldsymbol{\theta}_0^*)\boldsymbol{I}^{-1}(\boldsymbol{\theta}_0^*)\right\}' \end{array}\right] \\
&= \boldsymbol{A}(\boldsymbol{\theta}_0^*) - \boldsymbol{B}(\boldsymbol{\theta}_0^*)\boldsymbol{I}^{-1}(\boldsymbol{\theta}_0^*)\boldsymbol{B}'(\boldsymbol{\theta}_0^*) \\
&= \boldsymbol{D}^{-1}(\boldsymbol{\theta}_0^*). \qquad \text{(D.93)}
\end{aligned}
$$

Therefore asymptotic distribution of likelihood ratio statistic in (D.88) is chi-squared distribution with degrees of freedom R:

$$-2\log\left[\frac{\mathcal{L}(\hat{\boldsymbol{\theta}}_N^{res}|x)}{\mathcal{L}(\hat{\boldsymbol{\theta}}_N|x)}\right] \sim \chi_R^2. \qquad \text{(D.94)}$$

E Bayesian Measure of Model Selection

E.1 Marginal Likelihood and Bayes Factor

Given the parameter $\boldsymbol{\theta} \in \boldsymbol{\Theta}$, marginal likelihood $p(\boldsymbol{x}|M)$ for *i.i.d.* data $\boldsymbol{x} \equiv (x_1, x_2, \ldots, x_N)$ conditional on model M is defined as

$$p(\boldsymbol{x}|M) = \int_{\boldsymbol{\Theta}} \mathcal{L}_g(\boldsymbol{\theta}|\boldsymbol{x}, M)p(\boldsymbol{\theta}|M)d\boldsymbol{\theta} \qquad \text{(E.1)}$$

where

$$\mathcal{L}_g(\boldsymbol{\theta}|\boldsymbol{x}, M) = \prod_{i=1}^{N} g(x_i|\boldsymbol{\theta}, M)$$

and $p(\boldsymbol{\theta}|M)$ is a prior distribution given model M. To evaluate the marginal likelihood $p(\boldsymbol{x}|M)$ in (E.1), we suppose that $p(\boldsymbol{\theta}|M)$ is a proper density. Then we notice that

$$1 = \int_{\boldsymbol{\Theta}} p(\boldsymbol{\theta}|M)d\boldsymbol{\theta}$$

$$= \int_{\boldsymbol{\Theta}} \left[\frac{p(\boldsymbol{x}|M)p(\boldsymbol{\theta}|\boldsymbol{x}, M)}{\mathcal{L}_g(\boldsymbol{\theta}|\boldsymbol{x}, M)p(\boldsymbol{\theta}|M)} \right] p(\boldsymbol{\theta}|M)d\boldsymbol{\theta}$$
$$= p(\boldsymbol{x}|M) \int_{\boldsymbol{\Theta}} \frac{1}{\mathcal{L}_g(\boldsymbol{\theta}|\boldsymbol{x}, M)} p(\boldsymbol{\theta}|\boldsymbol{x}, M)d\boldsymbol{\theta} \tag{E.2}$$

where

$$\frac{p(\boldsymbol{x}|M)p(\boldsymbol{\theta}|\boldsymbol{x}, M)}{\mathcal{L}_g(\boldsymbol{\theta}|\boldsymbol{x}, M)p(\boldsymbol{\theta}|M)} = \frac{p(\boldsymbol{\theta}|\boldsymbol{x}, M)}{\frac{\mathcal{L}_g(\boldsymbol{\theta}|\boldsymbol{x}, M)p(\boldsymbol{\theta}|M)}{p(\boldsymbol{x}|M)}} = \frac{p(\boldsymbol{\theta}|\boldsymbol{x}, M)}{p(\boldsymbol{\theta}|\boldsymbol{x}, M)} = 1.$$

Given the MCMC draws after burn-in period $\{\boldsymbol{\theta}^{(j)}\}_{j=1}^{n}$ from the posterior distribution $p(\boldsymbol{\theta}|\boldsymbol{x}, M)$, Newton and Raftery (1994) estimates the marginal likelihood $p(x|M)$ in (E.2) as

$$\hat{p}(\boldsymbol{x}|M) = \left[\frac{1}{n} \sum_{j=1}^{n} \frac{1}{\mathcal{L}_g(\boldsymbol{\theta}^{(j)}|\boldsymbol{x}, M)} \right]^{-1} \tag{E.3}$$

and the Bayes factor for model M_i against model M_j is obtained as

$$\mathrm{BF}_{ij} = \frac{\hat{p}(\boldsymbol{x}|M_i)}{\hat{p}(\boldsymbol{x}|M_j)}. \tag{E.4}$$

The BIC in (D.53) gives a rough approximation to the logarithm of the Bayes factor (Kass and Raftery, 1995) as follows:

$$\log \mathrm{BF}_{ij} = \log\{p(\boldsymbol{x}|M_i)\} - \log\{p(\boldsymbol{x}|M_j)\}$$
$$\approx \frac{1}{2} \left[\mathrm{BIC}(M_i) - \mathrm{BIC}(M_j) \right]. \tag{E.5}$$

E.2 Deviance Information Criterion (DIC)

Given observed data $\boldsymbol{x} \equiv (x_1, x_2, \ldots, x_N)$ and K-dimensional parameter vector $\boldsymbol{\theta}$, Spiegelhalter et al. (2002) defined the reduction in uncertainty due to the estimation of parameter $\boldsymbol{\theta}$ as

$$d_{\Theta}\{\boldsymbol{x}, \boldsymbol{\theta}_0, \tilde{\boldsymbol{\theta}}(\boldsymbol{x})\} = -2\log\{p(\boldsymbol{x}|\boldsymbol{\theta}_0)\} + 2\log\{p(\boldsymbol{x}|\tilde{\boldsymbol{\theta}}(\boldsymbol{x}))\} \tag{E.6}$$

where $\boldsymbol{\theta}_0$ is a true parameter and $p(\boldsymbol{x}|\tilde{\boldsymbol{\theta}}(\boldsymbol{x}))$ is an approximating model. In the classical measure of model selection based on the MLE $\hat{\boldsymbol{\theta}}_N$, expectation of $d_\Theta\{\boldsymbol{x}, \boldsymbol{\theta}_0, \tilde{\boldsymbol{\theta}}(\boldsymbol{x})\}$ with respect to the unknown true model is used to express the complexity of non-Bayesian model.

In Bayesian perspective, true parameter $\boldsymbol{\theta}_0$ can be replaced by a random quantity $\boldsymbol{\theta}$. Then Spiegelhalter et al. (2002) defined a posterior mean of $d_\Theta\{\boldsymbol{x}, \boldsymbol{\theta}, \tilde{\boldsymbol{\theta}}(\boldsymbol{x})\}$ as the effective number of parameters $p_D\{\boldsymbol{x}, \boldsymbol{\Theta}, \tilde{\boldsymbol{\theta}}(\boldsymbol{x})\}$:

$$\begin{aligned} p_D\{\boldsymbol{x}, \boldsymbol{\Theta}, \tilde{\boldsymbol{\theta}}(\boldsymbol{x})\} &= \mathbf{E}_{p(\theta|x)}\left[d_\Theta\{\boldsymbol{x}, \boldsymbol{\theta}, \tilde{\boldsymbol{\theta}}(\boldsymbol{x})\}\right] \\ &= \mathbf{E}_{p(\theta|x)}\left[-2\log\{p(\boldsymbol{x}|\boldsymbol{\theta})\}\right] + 2\log\{p(\boldsymbol{x}|\tilde{\boldsymbol{\theta}}(\boldsymbol{x}))\}. \end{aligned} \tag{E.7}$$

Taking $\tilde{\boldsymbol{\theta}}(\boldsymbol{x}) = \mathbf{E}(\boldsymbol{\theta}|\boldsymbol{x}) = \bar{\boldsymbol{\theta}}$, effective number of parameters $p_D\{\boldsymbol{x}, \boldsymbol{\Theta}, \tilde{\boldsymbol{\theta}}(\boldsymbol{x})\}$ in (E.7) can be rewritten as

$$p_D = \overline{D(\boldsymbol{\theta})} - D(\bar{\boldsymbol{\theta}}) \tag{E.8}$$

where Spiegelhalter et al. (2002) termed $D(\boldsymbol{\theta})$ the 'Bayesian deviance'.

Suppose that we wish to make predictions on a replicate data set X_{rep} which has an identical design to the observed data $\boldsymbol{x} \equiv (x_1, x_2, \ldots, x_N)$, we set the true model $p(X_{rep}|\boldsymbol{\theta})$. Then deviance information criterion (DIC) selects a model for which $\mathbf{E}_{p(\theta|x)}\mathbf{E}_{p(X_{rep}|\theta)}\left[-2\log\{p(X_{rep}|\bar{\boldsymbol{\theta}})\}\right]$ is expected to be small. To derive the DIC, we define c_Θ such as

$$\begin{aligned} c_\Theta &= \mathbf{E}_{p(X_{rep}|\theta)}\left[-2\log\{p(X_{rep}|\bar{\boldsymbol{\theta}})\}\right] - \left[-2\log\{p(x|\bar{\boldsymbol{\theta}})\}\right] \\ &= \mathbf{E}_{p(X_{rep}|\theta)}\left[D_{rep}(\bar{\boldsymbol{\theta}})\right] - D(\bar{\boldsymbol{\theta}}). \end{aligned} \tag{E.9}$$

To evaluate (E.9), it is convenient to expand c_Θ into three terms:

$$\begin{aligned} c_\Theta = \mathbf{E}_{p(X_{rep}|\theta)}\left[D_{rep}(\bar{\boldsymbol{\theta}}) - D_{rep}(\boldsymbol{\theta})\right] + \mathbf{E}_{p(X_{rep}|\theta)}\Big[D_{rep}(\boldsymbol{\theta}) - D(\boldsymbol{\theta})\Big] \\ + \left[D(\boldsymbol{\theta}) - D(\bar{\boldsymbol{\theta}})\right] \end{aligned}$$

$$= L_1(\bar{\boldsymbol{\theta}}, \boldsymbol{\theta}) + L_2(\boldsymbol{\theta}, \boldsymbol{\theta}) + \left[D(\boldsymbol{\theta}) - D(\bar{\boldsymbol{\theta}})\right], \tag{E.10}$$

where we denote the first two terms by $L_1(\bar{\boldsymbol{\theta}}, \boldsymbol{\theta})$ and $L_2(\boldsymbol{\theta}, \boldsymbol{\theta})$ respectively.

In practice, Spiegelhalter et al. (2002) approximated the true Bayes estimator by the posterior mean $\bar{\boldsymbol{\theta}}$. Expanding the Bayesian deviance $D_{rep}(\bar{\boldsymbol{\theta}})$ to the second order shows that

$$D_{rep}(\bar{\boldsymbol{\theta}}) \approx D_{rep}(\boldsymbol{\theta}) + (\bar{\boldsymbol{\theta}} - \boldsymbol{\theta})' \frac{\partial D_{rep}(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} + \frac{1}{2}(\bar{\boldsymbol{\theta}} - \boldsymbol{\theta})' \frac{\partial^2 D_{rep}(\boldsymbol{\theta})}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}'} (\bar{\boldsymbol{\theta}} - \boldsymbol{\theta}) \tag{E.11}$$

and $L_1(\bar{\boldsymbol{\theta}}, \boldsymbol{\theta})$ in (E.10) can be approximated to

$$\begin{aligned} L_1(\bar{\boldsymbol{\theta}}, \boldsymbol{\theta}) &\approx \mathbf{E}_{p(X_{rep}|\theta)} \left[(\bar{\boldsymbol{\theta}} - \boldsymbol{\theta})' \frac{\partial D_{rep}(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} + \frac{1}{2}(\bar{\boldsymbol{\theta}} - \boldsymbol{\theta})' \frac{\partial^2 D_{rep}(\boldsymbol{\theta})}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}'} (\bar{\boldsymbol{\theta}} - \boldsymbol{\theta}) \right] \\ &= \mathbf{E}_{p(X_{rep}|\theta)} \left[-2(\bar{\boldsymbol{\theta}} - \boldsymbol{\theta})' \frac{\partial \log\{p(X_{rep}|\boldsymbol{\theta})\}}{\partial \boldsymbol{\theta}} \right. \\ &\qquad \left. -(\bar{\boldsymbol{\theta}} - \boldsymbol{\theta})' \frac{\partial^2 \log\{p(X_{rep}|\boldsymbol{\theta})\}}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}'} (\bar{\boldsymbol{\theta}} - \boldsymbol{\theta}) \right] \\ &= \mathrm{tr}\left\{ \boldsymbol{J}_{rep}(\boldsymbol{\theta}) (\boldsymbol{\theta} - \bar{\boldsymbol{\theta}})(\boldsymbol{\theta} - \bar{\boldsymbol{\theta}})' \right\} \end{aligned} \tag{E.12}$$

where

$$\mathbf{E}_{p(X_{rep}|\theta)} \left[\frac{\partial \log\{p(X_{rep}|\boldsymbol{\theta})\}}{\partial \boldsymbol{\theta}} \right] = \mathbf{0}_{K \times 1}, \tag{E.13}$$

$$\mathbf{E}_{p(X_{rep}|\theta)} \left[-\frac{\partial^2 \log\{p(X_{rep}|\boldsymbol{\theta})\}}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}'} \right] = \boldsymbol{J}_{rep}(\boldsymbol{\theta}). \tag{E.14}$$

Since $\boldsymbol{J}_{rep}(\boldsymbol{\theta})$ is assumed to be the Fisher information matrix $\boldsymbol{I}_{rep}(\boldsymbol{\theta})$, then (E.12) can be rewritten as

$$\mathrm{tr}\left\{ \boldsymbol{J}_{rep}(\boldsymbol{\theta}) (\boldsymbol{\theta} - \bar{\boldsymbol{\theta}})(\boldsymbol{\theta} - \bar{\boldsymbol{\theta}})' \right\} = \mathrm{tr}\left\{ \boldsymbol{I}_{rep}(\boldsymbol{\theta}) (\boldsymbol{\theta} - \bar{\boldsymbol{\theta}})(\boldsymbol{\theta} - \bar{\boldsymbol{\theta}})' \right\}. \tag{E.15}$$

Using (E.10) and (E.15), we have a posterior mean of c_Θ in (E.10):

$$\begin{aligned} \mathbf{E}_{p(\theta|x)}\left(c_\Theta\right) \approx{}& \mathbf{E}_{p(\theta|x)} \left[\mathrm{tr}\left\{ \boldsymbol{I}_{rep}(\boldsymbol{\theta}) (\boldsymbol{\theta} - \bar{\boldsymbol{\theta}})(\boldsymbol{\theta} - \bar{\boldsymbol{\theta}})' \right\} \right] \\ &+ \mathbf{E}_{p(\theta|x)} \left[L_2(\boldsymbol{\theta}, \boldsymbol{\theta}) \right] + \left[\overline{D(\boldsymbol{\theta})} - D(\bar{\boldsymbol{\theta}}) \right], \end{aligned}$$

$$= \mathrm{tr}\left\{\boldsymbol{I}_{rep}(\boldsymbol{\theta})\boldsymbol{\Sigma}_\theta\right\} + \mathbf{E}_{p(\theta|x)}\left[L_2(\boldsymbol{\theta},\boldsymbol{\theta})\right] + p_D \tag{E.16}$$

where

$$\boldsymbol{\Sigma}_\theta = \mathbf{E}_{p(\theta|x)}\left[(\boldsymbol{\theta}-\bar{\boldsymbol{\theta}})(\boldsymbol{\theta}-\bar{\boldsymbol{\theta}})'\right] \tag{E.17}$$

and Spiegelhalter et al. (1998, 2002) suggested that posterior mean of $L_2(\boldsymbol{\theta},\boldsymbol{\theta})$ on the right-hand side of (E.16) is zero:

$$\begin{aligned}\mathbf{E}_{p(\theta|x)}\left[L_2(\boldsymbol{\theta},\boldsymbol{\theta})\right] &= \mathbf{E}_{p(\theta|x)}\mathbf{E}_{p(X_{rep}|\theta)}\left[-2\log\{p(X_{rep}|\boldsymbol{\theta})\}\right] + 2\mathbf{E}_{p(\theta|x)}\left[\log\{p(\boldsymbol{x}|\boldsymbol{\theta})\}\right] \\ &= 0. \end{aligned} \tag{E.18}$$

Next we expand $D(\boldsymbol{\theta}) = -2\log\{p(\boldsymbol{x}|\boldsymbol{\theta})\}$ around $\bar{\boldsymbol{\theta}}$ to the second order:

$$\begin{aligned} D(\boldsymbol{\theta}) &\approx D(\bar{\boldsymbol{\theta}}) + (\boldsymbol{\theta}-\bar{\boldsymbol{\theta}})'\left.\frac{\partial D(\boldsymbol{\theta})}{\partial\boldsymbol{\theta}}\right|_{\boldsymbol{\theta}=\bar{\boldsymbol{\theta}}} + \frac{1}{2}(\boldsymbol{\theta}-\bar{\boldsymbol{\theta}})'\left.\frac{\partial^2 D(\boldsymbol{\theta})}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}'}\right|_{\boldsymbol{\theta}=\bar{\boldsymbol{\theta}}}(\boldsymbol{\theta}-\bar{\boldsymbol{\theta}}) \\ &= D(\bar{\boldsymbol{\theta}}) - 2(\boldsymbol{\theta}-\bar{\boldsymbol{\theta}})'\frac{\partial\log\{p(\boldsymbol{x}|\bar{\boldsymbol{\theta}})\}}{\partial\boldsymbol{\theta}} - (\boldsymbol{\theta}-\bar{\boldsymbol{\theta}})'\frac{\partial^2\log\{p(\boldsymbol{x}|\bar{\boldsymbol{\theta}})\}}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}'}(\boldsymbol{\theta}-\bar{\boldsymbol{\theta}}). \end{aligned} \tag{E.19}$$

The posterior mean of (E.19) gives

$$\begin{aligned}\mathbf{E}_{p(\theta|x)}\left[D(\boldsymbol{\theta})\right] &\approx D(\bar{\boldsymbol{\theta}}) + \mathrm{tr}\left\{-\boldsymbol{H}(\bar{\boldsymbol{\theta}})\mathbf{E}_{p(\theta|x)}\left[(\boldsymbol{\theta}-\bar{\boldsymbol{\theta}})(\boldsymbol{\theta}-\bar{\boldsymbol{\theta}})'\right]\right\} \\ &= D(\bar{\boldsymbol{\theta}}) + \mathrm{tr}\left\{-\boldsymbol{H}(\bar{\boldsymbol{\theta}})\boldsymbol{\Sigma}_\theta\right\} \end{aligned} \tag{E.20}$$

where

$$\begin{aligned}&\mathbf{E}_{p(\theta|x)}\left[\boldsymbol{\theta}-\bar{\boldsymbol{\theta}}\right] = \mathbf{E}_{p(\theta|x)}\left(\boldsymbol{\theta}\right) - \bar{\boldsymbol{\theta}} = \mathbf{0}, \\ &\boldsymbol{H}(\bar{\boldsymbol{\theta}}) = \frac{\partial^2\log\{p(\boldsymbol{x}|\bar{\boldsymbol{\theta}})\}}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}'}. \end{aligned}$$

Since $\mathbf{E}_{p(\theta|x)}\left[D(\boldsymbol{\theta})\right] = \overline{D(\boldsymbol{\theta})}$, (E.20) can be rewritten as

$$\overline{D(\boldsymbol{\theta})} - D(\bar{\boldsymbol{\theta}}) \approx \mathrm{tr}\left\{-\boldsymbol{H}(\bar{\boldsymbol{\theta}})\boldsymbol{\Sigma}_\theta\right\}. \tag{E.21}$$

Recall (E.16):

$$\mathbf{E}_{p(\theta|x)}\left(c_\Theta\right) \approx \mathbf{E}_{p(\theta|x)}\left[\mathrm{tr}\left\{\boldsymbol{I}_{rep}(\boldsymbol{\theta})(\boldsymbol{\theta}-\bar{\boldsymbol{\theta}})(\boldsymbol{\theta}-\bar{\boldsymbol{\theta}})'\right\}\right]$$

$$
\begin{aligned}
&+ \mathbf{E}_{p(\theta|x)}\left[L_2(\boldsymbol{\theta},\boldsymbol{\theta})\right] + \left[\overline{D(\boldsymbol{\theta})} - D(\bar{\boldsymbol{\theta}})\right] \\
&= \mathrm{tr}\left\{\boldsymbol{I}_{rep}(\boldsymbol{\theta})\boldsymbol{\Sigma}_\theta\right\} + p_D.
\end{aligned}
$$

Spiegelhalter et al. (1998, 2002) approximated the first term on the right-hand side of (E.16) as

$$
\begin{aligned}
\mathrm{tr}\left\{\boldsymbol{I}_{rep}(\boldsymbol{\theta})\boldsymbol{\Sigma}_\theta\right\} &\approx \mathrm{tr}\left\{-\boldsymbol{H}(\bar{\boldsymbol{\theta}})\boldsymbol{\Sigma}_\theta\right\} \\
&\approx \overline{D(\boldsymbol{\theta})} - D(\bar{\boldsymbol{\theta}}) \\
&= p_D. \qquad \text{(E.22)}
\end{aligned}
$$

Suppose that (E.16) and (E.22) hold, posterior mean of c_Θ in (E.16) is approximated to

$$
\mathbf{E}_{p(\theta|x)}\left(c_\Theta\right) \approx 2p_D. \qquad \text{(E.23)}
$$

Using (E.9) and (E.23), we have

$$
\begin{aligned}
\mathbf{E}_{p(\theta|x)}\mathbf{E}_{p(X_{rep}|\theta)}\left[-2\log\{p(X_{rep}|\bar{\boldsymbol{\theta}})\}\right] &= -2\log\{p(\boldsymbol{x}|\bar{\boldsymbol{\theta}})\} + \mathbf{E}_{p(\theta|x)}\left(c_\Theta\right) \\
&\approx D(\bar{\boldsymbol{\theta}}) + 2p_D. \qquad \text{(E.24)}
\end{aligned}
$$

Therefore DIC is given by

$$
\begin{aligned}
\mathrm{DIC} &= D(\bar{\boldsymbol{\theta}}) + 2p_D \\
&= \overline{D(\boldsymbol{\theta})} + p_D \qquad \text{(E.25)}
\end{aligned}
$$

where $p_D = \overline{D(\boldsymbol{\theta})} - D(\bar{\boldsymbol{\theta}})$ and $\overline{D(\boldsymbol{\theta})} = \mathbf{E}_{p(\theta|x)}\left[-2\log p(\boldsymbol{x}|\boldsymbol{\theta})\right]$.

References

Akaike, H. (1973) 'Information theory as an extension of the maximum likelihood principle.' Second International Symposium on Information Theory Akademiai Kiado, Budapest pp. 267–281

Ando, T. (2007) 'Bayesian predictive information criterion for the evaluation of hierarchical bayesian and empirical bayes models.' *Biometrika* 94(2), 443–458

Bedrick, E. J., and C. L. Tsai (1994) 'Model selection for multivariate regression in small samples.' *Biometrics* pp. 226–231

Burnham, K. P. (2002) 'Discussion of a paper by D.J. Spiegelhalter et al.' *Journal of the Royal Statistical Society. Series B (Statistical Methodology)* 64(4), 629

Gelfand, A. E., and D. K. Dey (1994) 'Bayesian model choice: asymptotics and exact calculations.' *Journal of the Royal Statistical Society. Series B (Methodological)* pp. 501–514

Gelfand, A. E., and S. K. Ghosh (1998) 'Model choice: A minimum posterior predictive loss approach.' *Biometrika* 85(1), 1–11

Greene, W. H. (2000) *Econometric Analysis 4th Edition* (New York, PrenticeHall)

Hayashi, F. (2000) *Econometrics* (Princeton University Press. Section)

Hurvich, C.M., and C.L. Tsai (1989) 'Regression and time series model selection in small samples.' *Biometrika* 76(2), 297–307

Ibrahim, J. G., M. H. Chen, and D. Sinha (2001) 'Criterion-based methods for bayesian model assessment.' *Statistica Sinica* 11(2), 419–444

Jeffreys, H. (1961) *Theory of probability* (Oxford University Press, USA)

Kass, R. E., and A. E. Raftery (1995) 'Bayes factors.' *The Journal of the American Statistical Association* 90(430), 773–795

Kitagawa, G. (1997) 'Information criteria for the predictive evaluation of bayesian models.' *Communications in statistics-theory and methods* 26(9), 2223–2246

Laud, P. W., and J. G. Ibrahim (1995) 'Predictive model selection.' *Journal of the Royal Statistical Society. Series B (Methodological)* pp. 247–262

Nerlove, M. (1961) *RETURNS TO SCALE IN ELECTRICITY SUPPLY.* (STANFORD UNIV CALIF APPLIED MATHEMATICS AND STATISTICS LABS. Defense Technical Information Center)

Newton, M. A., and A. E. Raftery (1994) 'Approximate bayesian inference with the weighted likelihood bootstrap.' *Journal of the Royal Statistical Society, Series B* 56(1), 3–48

Schwarz, G. (1978) 'Estimating the dimension of a model.' *Annals of Mathematical Statistics* 42(3), 1003–1009

Spiegelhalter, D. J., N. G. Best, B. P. Carlin, and A. van der Linde (1998) 'Bayesian deviance, the effective number of parameters, and the comparison of arbitrarily complex models.' *Research report*

_ (2002) 'Bayesian measures of model complexity and fit.' *Journal of the Royal Statistical Society. Series B (Statistical Methodology)* 64(4), 583–639

Sugiura, N. (1978) 'Further analysts of the data by akaike's information criterion and the finite corrections.' *Communications in Statistics-Theory and Methods* 7(1), 13–26