Second-Order Bias-Corrected AIC for Selecting Structural Equation Models

Kentaro HAYASHI
Department of Psychology, University of Hawaii at Manoa
(E-mail: hayashik@hawaii.edu)

AND

Hirokazu YANAGIHARA
Department of Mathematics, Hiroshima University
(E-mail: yanagi@math.sci.hiroshima-u.ac.jp)

Introduction

We derive a second-order bias correction of Akaike Information Criterion (AIC) in structural equation models (SEM) under the normal assumption when the model is **overspecified**.

Note: "Overspecified" means the candidate models (f's) include the true model (φ).

Contents:

1. **Introduction**

(a) Structural Equation Models (SEM)

(b) AIC and CV (Cross-Validation) Criterion

2. **General Theory**

(a) True and Candidate Models

(b) Likelihood and MLE

(c) Risk, Bias, and Information Criterion

(d) Estimated Bias

3. **Recent and Current Studies**

(a) Notations (Derivatives, Expectation of Moment Matrices, Estimates of Expected Moment Matrices, and Coefficients in Bias-Correction Terms)

(b) Evaluating Bias of Information Criteria

(c) Asymptotic Expansion of Expectation of Estimated “Beta” Term

(d) Bias of AIC (Main Result of Current Study)

(e) Useful Formulas in Obtaining the Coefficient Terms

Structural Equation Models (SEM)

References: Bollen (1989), Bartholomew and Knott (1999), Skrondal and Rabe-Hesketh (2004), Yuan and Bentler (2007)

- SEM is one of the most frequently used multivariate techniques in social sciences.

- SEM aims to express the covariance structure using relatively small number of parameters. Notation: $\boldsymbol{\Sigma}(\boldsymbol{\theta})$

- The single most famous SEM is the confirmatory factor analysis (CFA) model:

  $$\boldsymbol{y} = \boldsymbol{\mu} + \boldsymbol{\Lambda} \boldsymbol{f} + \boldsymbol{\varepsilon},$$

  where

$\boldsymbol{y}$ $(p \times 1)$:	Observed variables,
$\boldsymbol{\mu}$ $(p \times 1)$:	Population means,
$\boldsymbol{\Lambda}$ $(p \times m)$:	Factor loadings (Path coefficients),
$\boldsymbol{f}$ $(m \times 1)$:	Factors,
$\boldsymbol{\varepsilon}$ $(p \times 1)$:	Errors.

  Note: CFA is a linear model but the factors ($\boldsymbol{f}$) are **latent** (NOT observed) variables.

- Typical assumptions: Errors are mutually uncorrelated, and factors and errors are uncorrelated. That is, $\mathrm{Cov}(\boldsymbol{\varepsilon}_i, \boldsymbol{\varepsilon}_j) = 0$ and $\mathrm{Cov}(\boldsymbol{f}_i, \boldsymbol{\varepsilon}_j) = 0$.

- The covariance structure of variables ($\boldsymbol{y}$) is expressed as a function of:

  $\boldsymbol{\Lambda}$ $(p \times m)$: Factor loadings (Path coefficients),

  $\boldsymbol{\Phi}$ $(m \times m)$: Factor correlations, and

  $\boldsymbol{\Psi}$ $(p \times p)$: Error variances.

- That is, the covariance structure under CFA is expressed as:

$$\boldsymbol{\Sigma}(\boldsymbol{\theta}) = \boldsymbol{\Lambda}\boldsymbol{\Phi}\boldsymbol{\Lambda}' + \boldsymbol{\Psi},$$

where $\boldsymbol{\theta} = (\boldsymbol{\lambda}', \boldsymbol{\phi}', \boldsymbol{\psi}')' = (\theta_1, \ldots, \theta_q)'$.

AIC (Akaike Information Criterion)

- When the candidate model is *overspecified*, AIC is the first-order bias-corrected estimator of the risk function based on the expected predictive Kullback-Leibler (KL) discrepancy between the true model and the candidate model. That is,

$$\mathrm{E}[\mathrm{AIC}] = R_{\mathrm{KL}} + O(n^{-1}).$$

- AIC tends to choose the model with many parameters as the best model (when the full model has too many parameters).

- *Reason*: AIC tends to underestimate the bias when the candidate model has many parameters (because the bias term of AIC is derived based on the asymptotic theory of $\hat{\boldsymbol{\theta}}$).

- The (negative) property of AIC that the candidate model having too many parameters is chosen as the best model tends to appear in the *overspecified* models.

Cross-Validation (CV)

- Even when the candidate model is misspecified, by correcting the bias of the cross validation (CV) criterion (Stone, 1974), the *second-order bias-corrected* estimators of the risk function have been proposed under the general condition (e.g., Yanagihara, Tonda and Matsumoto, 2006). That is,

$$\mathrm{E}[\mathrm{CCV}] = R_{\mathrm{KL}} + O(n^{-2}).$$

- However, the many computational tasks are need for obtaining bias-corrected criteria and these criteria have <u>large variance</u>.

True and Candidate Models (General)

Let $\boldsymbol{y}_1,\ldots,\boldsymbol{y}_n$ be p-dimensional random observation vectors, where n is the sample size.

- The true model: $M_\varphi : \boldsymbol{y}_1,\ldots,\boldsymbol{y}_n \sim i.i.d.\, \varphi(\boldsymbol{y})$,

where $\varphi(\boldsymbol{y})$ is an *unknown* probability density function.

- The candidate model: $M_f : \boldsymbol{y}_1,\ldots,\boldsymbol{y}_n \sim i.i.d.\, f(\boldsymbol{y} \mid \boldsymbol{\theta})$,

where $\mathcal{F} = \{f(\boldsymbol{y} \mid \boldsymbol{\theta}); \boldsymbol{\theta} \subseteq \Theta\}$, $\boldsymbol{\theta} = (\theta_1,\ldots,\theta_q)'$.

Candidate Model in SEM

If the candidate model is specified, then

$$N\boldsymbol{S} \sim W_p(N, \boldsymbol{\Sigma}(\boldsymbol{\theta}_0))$$

$$\Rightarrow N\boldsymbol{S} = \boldsymbol{W}_1 + \cdots + \boldsymbol{W}_N,\ \boldsymbol{W}_1, \ldots, \boldsymbol{W}_N \sim i.i.d. W_p(1, \boldsymbol{\Sigma}(\boldsymbol{\theta}_0)),$$

where $N = n - 1$.

$\boldsymbol{W}_1, \ldots, \boldsymbol{W}_N$ can be regarded as independent observations. Therefore, the candidate model is:

$$M_f : \boldsymbol{W}_1, \ldots, \boldsymbol{W}_N \sim i.i.d. W_p(1, \boldsymbol{\Sigma}(\boldsymbol{\theta})).$$

Likelihood and MLE (General)

- **Log-likelihood function**: $L(\boldsymbol{\theta} \mid \boldsymbol{Y}) = \sum_{i=1}^{n} \log f(\boldsymbol{y}_i \mid \boldsymbol{\theta})$,

  where $\boldsymbol{Y} = (\boldsymbol{y}_1, \ldots, \boldsymbol{y}_n)'$.

- **Maximum likelihood estimator** (MLE) of $\boldsymbol{\theta}$: $\hat{\boldsymbol{\theta}} = \arg\max_{\boldsymbol{\theta}} L(\boldsymbol{\theta} \mid \boldsymbol{Y})$

- Convergence of MLE in the misspecified model (White, 1982) :

$$\lim_{n \to \infty} \hat{\boldsymbol{\theta}} = \boldsymbol{\theta}_0, \quad E_{\boldsymbol{y}}\left[\frac{\partial}{\partial \boldsymbol{\theta}} \log f(\boldsymbol{y} \mid \boldsymbol{\theta}) \Big|_{\boldsymbol{\theta} = \boldsymbol{\theta}_0} \right] = \mathbf{0}$$

  where $E_{\boldsymbol{y}}$ denotes an expectation with respect to $\boldsymbol{y}$ under the true model $\varphi(\boldsymbol{y})$.

Likelihood and MLE in SEM

In SEM, the discrepancy function is:

$$F_{\mathrm{KL}}(\boldsymbol{\theta} \mid \boldsymbol{S}) = -\log \mid \boldsymbol{S}\Sigma^{-1} \mid + tr(\boldsymbol{S}\Sigma^{-1}) - p$$

$$\Rightarrow \frac{1}{N} \sum_{i=1}^{N} F_{\mathrm{KL}}(\boldsymbol{\theta} \mid \boldsymbol{W}_i).$$

Therefore, the log-likelihood is: $-2L(\boldsymbol{\theta} \mid \boldsymbol{S}) = (N)F_{\mathrm{KL}}(\boldsymbol{\theta} \mid \boldsymbol{S})$,
and the MLE is:

$$\hat{\boldsymbol{\theta}} = \arg\max_{\boldsymbol{\theta}} L(\boldsymbol{\theta} \mid \boldsymbol{S}).$$

Risk Function, Bias, and Information Criterion

- **Risk function based on the expected predictive KL discrepancy**:

$$R_{\mathrm{KL}} = E_{\boldsymbol{y}} E_{\boldsymbol{u}} \left[-2L(\hat{\boldsymbol{\theta}} \mid \boldsymbol{U}) \right],$$

where $\boldsymbol{U} = (\boldsymbol{u}_1, \ldots, \boldsymbol{u}_n)'$ is an $n \times p$ future observation matrix (independent of $\boldsymbol{Y}$), and $\boldsymbol{u}_i$ is distributed according to the same distribution of $\boldsymbol{y}_i$ $(i = 1, \ldots, n)$.

- Bias: $B = R_{\mathrm{KL}} - E_{\boldsymbol{y}} \left[-2L(\hat{\boldsymbol{\theta}} \mid \boldsymbol{Y}) \right]$.

- Information criterion (IC):

$$\mathrm{IC} = -2L(\hat{\boldsymbol{\theta}} \mid \boldsymbol{Y}) + \hat{B},$$

where $\hat{B}$ is a consistent estimator of B.

Note: The ICs are specified by different terms of $\hat{B}$.

Estimated Bias in Information Criteria

- AIC: $\hat{B} = 2q$.
- TIC (Takeuchi information criterion; Takeuchi, 1976):

$$\hat{B} = 2\mathrm{tr}\{\hat{\boldsymbol{I}}(\hat{\boldsymbol{\theta}})\hat{\boldsymbol{J}}(\hat{\boldsymbol{\theta}})^{-1}\}.$$

- CV: $\hat{B} = -2\sum_{i=1}^{n} \log f(\boldsymbol{y}_i \mid \hat{\boldsymbol{\theta}}_{[-i]}) + 2L(\hat{\boldsymbol{\theta}} \mid \boldsymbol{Y})$,

  where $\boldsymbol{\theta}_{[-i]}$ is Jackknife estimator of $\boldsymbol{\theta}$ defined by

$$\hat{\boldsymbol{\theta}}_{[-i]} = \arg\max_{\boldsymbol{\theta}} \left\{ \sum_{j \neq i}^{n} \log f(\boldsymbol{y}_j \mid \boldsymbol{\theta}) \right\}.$$

Notations

A. Derivatives:

1. *First*-order (**Gradient**): $\boldsymbol{g}(\boldsymbol{y} \mid \boldsymbol{\theta}) = -\frac{\partial}{\partial \boldsymbol{\theta}} \log f(\boldsymbol{y} \mid \boldsymbol{\theta}),$

2. *Second*-order (**Hessian**): $\boldsymbol{H}(\boldsymbol{y} \mid \boldsymbol{\theta}) = -\frac{\partial^2}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}'} \log f(\boldsymbol{y} \mid \boldsymbol{\theta}),$

3. *Third*-order: $\boldsymbol{C}(\boldsymbol{y} \mid \boldsymbol{\theta}) = -\left(\frac{\partial}{\partial \boldsymbol{\theta}'} \otimes \frac{\partial^2}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}'}\right) \log f(\boldsymbol{y} \mid \boldsymbol{\theta}),$

4. *Fourth*-order: $\boldsymbol{Q}(\boldsymbol{y} \mid \boldsymbol{\theta}) = -\left(\frac{\partial^2}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}'} \otimes \frac{\partial^2}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}'}\right) \log f(\boldsymbol{y} \mid \boldsymbol{\theta}),$

where $\otimes$ is the Kronecker product.

B. Expectation of Moment Matrices

1. **Information**: $\boldsymbol{I}(\boldsymbol{\theta}) = E_{\boldsymbol{y}}[\boldsymbol{g}(\boldsymbol{y} \mid \boldsymbol{\theta})\boldsymbol{g}(\boldsymbol{y} \mid \boldsymbol{\theta})']$,

2. **Jacobian**: $\boldsymbol{J}(\boldsymbol{\theta}) = E_{\boldsymbol{y}}[\boldsymbol{H}(\boldsymbol{y} \mid \boldsymbol{\theta})]$,

3. Expected *third*-order moment matrix: $\boldsymbol{K}(\boldsymbol{\theta}) = E_{\boldsymbol{y}}[\boldsymbol{C}(\boldsymbol{y} \mid \boldsymbol{\theta})]$,

4. Expected *fourth*-order moment matrix: $\boldsymbol{L}(\boldsymbol{\theta}) = E_{\boldsymbol{y}}[\boldsymbol{Q}(\boldsymbol{y} \mid \boldsymbol{\theta})]$.

C. Estimates of Expected Moment Matrices:

1. Estimated **Information:** $\hat{\boldsymbol{I}}(\boldsymbol{\theta}) = \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{g}(\boldsymbol{y}_i \mid \boldsymbol{\theta})\boldsymbol{g}(\boldsymbol{y}_i \mid \boldsymbol{\theta})'$,

2. Estimated **Jacobian:** $\hat{\boldsymbol{J}}(\boldsymbol{\theta}) = \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{H}(\boldsymbol{y}_i \mid \boldsymbol{\theta})$,

3. Estimated 3rd-order moment: $\hat{\boldsymbol{K}}(\boldsymbol{\theta}) = \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{C}(\boldsymbol{y}_i \mid \boldsymbol{\theta})$,

4. Estimated 4th-order moment: $\hat{\boldsymbol{L}}(\boldsymbol{\theta}) = \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{Q}(\boldsymbol{y}_i \mid \boldsymbol{\theta})$.

D(a). Coefficients in Bias-Correction Terms (General Case)

$$\alpha_1 = \mathrm{tr}\{\boldsymbol{I}(\boldsymbol{\theta}_0)\boldsymbol{J}(\boldsymbol{\theta}_0)^{-1}\},$$

$$\alpha_2 = E_{\boldsymbol{y}}[\boldsymbol{g}(\boldsymbol{y} \mid \boldsymbol{\theta}_0)'\boldsymbol{J}(\boldsymbol{\theta}_0)^{-1}\boldsymbol{H}(\boldsymbol{y} \mid \boldsymbol{\theta}_0)\boldsymbol{J}(\boldsymbol{\theta}_0)^{-1}\boldsymbol{g}(\boldsymbol{y} \mid \boldsymbol{\theta}_0)],$$

$$\alpha_3 = E_{\boldsymbol{y}}[\boldsymbol{g}(\boldsymbol{y} \mid \boldsymbol{\theta}_0)'\boldsymbol{J}(\boldsymbol{\theta}_0)^{-1}\boldsymbol{K}(\boldsymbol{\theta}_0)\{\boldsymbol{J}(\boldsymbol{\theta}_0)^{-1}\boldsymbol{g}(\boldsymbol{y} \mid \boldsymbol{\theta}_0) \otimes \boldsymbol{J}(\boldsymbol{\theta}_0)^{-1}\boldsymbol{g}(\boldsymbol{y} \mid \boldsymbol{\theta}_0)\}],$$

$$\alpha_4 = E_{\boldsymbol{y}}\left[\mathrm{tr}\left\{\boldsymbol{J}(\boldsymbol{\theta}_0)^{-1}\boldsymbol{I}(\boldsymbol{\theta}_0)\boldsymbol{J}(\boldsymbol{\theta}_0)^{-1}\boldsymbol{H}(\boldsymbol{y} \mid \boldsymbol{\theta}_0)\boldsymbol{J}(\boldsymbol{\theta}_0)^{-1}\boldsymbol{H}(\boldsymbol{y} \mid \boldsymbol{\theta}_0)\right\}\right],$$

$$\alpha_5 = E_{\boldsymbol{y}}\left[\mathrm{tr}\{\boldsymbol{C}(\boldsymbol{y} \mid \boldsymbol{\theta}_0)(\boldsymbol{J}(\boldsymbol{\theta}_0)^{-1}\boldsymbol{g}(\boldsymbol{y} \mid \boldsymbol{\theta}_0) \otimes \boldsymbol{J}(\boldsymbol{\theta}_0)^{-1}\boldsymbol{I}(\boldsymbol{\theta}_0)\boldsymbol{J}(\boldsymbol{\theta}_0)^{-1})\}\right],$$

$$\alpha_6 = E_{\boldsymbol{y}}\left[\boldsymbol{g}(\boldsymbol{y} \mid \boldsymbol{\theta}_0)'\boldsymbol{J}(\boldsymbol{\theta}_0)^{-1}\boldsymbol{H}(\boldsymbol{y} \mid \boldsymbol{\theta}_0)\right]\boldsymbol{J}(\boldsymbol{\theta}_0)^{-1}E_{\boldsymbol{y}}\left[\boldsymbol{H}(\boldsymbol{y} \mid \boldsymbol{\theta}_0)\boldsymbol{J}(\boldsymbol{\theta}_0)^{-1}\boldsymbol{g}(\boldsymbol{y} \mid \boldsymbol{\theta}_0)\right],$$

$$\alpha_7 = E_{\boldsymbol{y}}\left[\mathrm{tr}\left\{\boldsymbol{K}(\boldsymbol{\theta}_0)(\boldsymbol{J}(\boldsymbol{\theta}_0)^{-1}\boldsymbol{H}(\boldsymbol{y} \mid \boldsymbol{\theta}_0)\boldsymbol{J}(\boldsymbol{\theta}_0)^{-1}\boldsymbol{g}(\boldsymbol{y} \mid \boldsymbol{\theta}_0) \otimes \boldsymbol{J}(\boldsymbol{\theta}_0)^{-1}\boldsymbol{I}(\boldsymbol{\theta}_0)\boldsymbol{J}(\boldsymbol{\theta}_0)^{-1})\right\}\right],$$

$$\alpha_8 = \mathrm{tr}\left\{\boldsymbol{K}(\boldsymbol{\theta}_0)(\boldsymbol{J}(\boldsymbol{\theta}_0)^{-1}\boldsymbol{K}(\boldsymbol{\theta}_0)\mathrm{vec}(\boldsymbol{J}(\boldsymbol{\theta}_0)^{-1}\boldsymbol{I}(\boldsymbol{\theta}_0)\boldsymbol{J}(\boldsymbol{\theta}_0)^{-1}) \otimes \boldsymbol{J}(\boldsymbol{\theta}_0)^{-1}\boldsymbol{I}(\boldsymbol{\theta}_0)\boldsymbol{J}(\boldsymbol{\theta}_0)^{-1})\right\},$$

$$\alpha_9 = E_{\varphi}\left[\boldsymbol{g}(\boldsymbol{y}_1 \mid \boldsymbol{\theta}_0)'\boldsymbol{J}(\boldsymbol{\theta}_0)^{-1}\boldsymbol{H}(\boldsymbol{y}_2 \mid \boldsymbol{\theta}_0)\boldsymbol{J}(\boldsymbol{\theta}_0)^{-1}\boldsymbol{H}(\boldsymbol{y}_1 \mid \boldsymbol{\theta}_0)\boldsymbol{J}(\boldsymbol{\theta}_0)^{-1}\boldsymbol{g}(\boldsymbol{y}_2 \mid \boldsymbol{\theta}_0)\right],$$

$$\alpha_{10} = E_{\varphi}\left[\mathrm{tr}\{\boldsymbol{K}(\boldsymbol{\theta}_0)(\boldsymbol{J}(\boldsymbol{\theta}_0)^{-1}\boldsymbol{g}(\boldsymbol{y} \mid \boldsymbol{\theta}_0) \otimes \boldsymbol{J}(\boldsymbol{\theta}_0)^{-1}\boldsymbol{H}(\boldsymbol{y} \mid \boldsymbol{\theta}_0)\boldsymbol{J}(\boldsymbol{\theta}_0)^{-1}\boldsymbol{I}(\boldsymbol{\theta}_0)\boldsymbol{J}(\boldsymbol{\theta}_0)^{-1}\}\right],$$

$$\alpha_{11} = \mathrm{tr}\{\boldsymbol{K}(\boldsymbol{\theta}_0)'\boldsymbol{J}(\boldsymbol{\theta}_0)^{-1}\boldsymbol{K}(\boldsymbol{\theta}_0)(\boldsymbol{J}(\boldsymbol{\theta}_0)^{-1}\boldsymbol{I}(\boldsymbol{\theta}_0)\boldsymbol{J}(\boldsymbol{\theta}_0)^{-1} \otimes \boldsymbol{J}(\boldsymbol{\theta}_0)^{-1}\boldsymbol{I}(\boldsymbol{\theta}_0)\boldsymbol{J}(\boldsymbol{\theta}_0)^{-1})\},$$

$$\alpha_{12} = \mathrm{tr}\{\boldsymbol{L}(\boldsymbol{\theta}_0)(\boldsymbol{J}(\boldsymbol{\theta}_0)^{-1}\boldsymbol{I}(\boldsymbol{\theta}_0)\boldsymbol{J}(\boldsymbol{\theta}_0)^{-1} \otimes \boldsymbol{J}(\boldsymbol{\theta}_0)^{-1}\boldsymbol{I}(\boldsymbol{\theta}_0)\boldsymbol{J}(\boldsymbol{\theta}_0)^{-1})\}.$$

D(b). Case with *Overspecified* Models $\boldsymbol{I}(\boldsymbol{\theta}_0) = \boldsymbol{J}(\boldsymbol{\theta}_0)$

$$\alpha_1 = q,$$

$$\alpha_2 = E_{\boldsymbol{y}}[\boldsymbol{g}(\boldsymbol{y} \mid \boldsymbol{\theta}_0)'\boldsymbol{I}(\boldsymbol{\theta}_0)^{-1}\boldsymbol{H}(\boldsymbol{y} \mid \boldsymbol{\theta}_0)\boldsymbol{I}(\boldsymbol{\theta}_0)^{-1}\boldsymbol{g}(\boldsymbol{y} \mid \boldsymbol{\theta}_0)],$$

$$\alpha_3 = E_{\boldsymbol{y}}[\boldsymbol{g}(\boldsymbol{y} \mid \boldsymbol{\theta}_0)'\boldsymbol{I}(\boldsymbol{\theta}_0)^{-1}\boldsymbol{K}(\boldsymbol{\theta}_0)\{\boldsymbol{I}(\boldsymbol{\theta}_0)^{-1}\boldsymbol{g}(\boldsymbol{y} \mid \boldsymbol{\theta}_0) \otimes \boldsymbol{I}(\boldsymbol{\theta}_0)^{-1}\boldsymbol{g}(\boldsymbol{y} \mid \boldsymbol{\theta}_0)\}],$$

$$\alpha_4 = E_{\boldsymbol{y}}\left[\mathrm{tr}\left\{\boldsymbol{I}(\boldsymbol{\theta}_0)^{-1}\boldsymbol{H}(\boldsymbol{y} \mid \boldsymbol{\theta}_0)\boldsymbol{I}(\boldsymbol{\theta}_0)^{-1}\boldsymbol{H}(\boldsymbol{y} \mid \boldsymbol{\theta}_0)\right\}\right],$$

$$\alpha_5 = E_{\boldsymbol{y}}\left[\mathrm{tr}\{\boldsymbol{C}(\boldsymbol{y} \mid \boldsymbol{\theta}_0)(\boldsymbol{I}(\boldsymbol{\theta}_0)^{-1}\boldsymbol{g}(\boldsymbol{y} \mid \boldsymbol{\theta}_0) \otimes \boldsymbol{I}(\boldsymbol{\theta}_0)^{-1})\}\right],$$

$$\alpha_6 = E_{\boldsymbol{y}}\left[\boldsymbol{g}(\boldsymbol{y} \mid \boldsymbol{\theta}_0)'\boldsymbol{I}(\boldsymbol{\theta}_0)^{-1}\boldsymbol{H}(\boldsymbol{y} \mid \boldsymbol{\theta}_0)\right]\boldsymbol{I}(\boldsymbol{\theta}_0)^{-1}E_{\boldsymbol{y}}\left[\boldsymbol{H}(\boldsymbol{y} \mid \boldsymbol{\theta}_0)\boldsymbol{J}(\boldsymbol{\theta}_0)^{-1}\boldsymbol{g}(\boldsymbol{y} \mid \boldsymbol{\theta}_0)\right],$$

$$\alpha_7 = E_{\boldsymbol{y}}\left[\mathrm{tr}\left\{\boldsymbol{K}(\boldsymbol{\theta}_0)(\boldsymbol{I}(\boldsymbol{\theta}_0)^{-1}\boldsymbol{H}(\boldsymbol{y} \mid \boldsymbol{\theta}_0)\boldsymbol{I}(\boldsymbol{\theta}_0)^{-1}\boldsymbol{g}(\boldsymbol{y} \mid \boldsymbol{\theta}_0) \otimes \boldsymbol{I}(\boldsymbol{\theta}_0)^{-1})\right\}\right],$$

$$\alpha_8 = \mathrm{tr}\left\{\boldsymbol{K}(\boldsymbol{\theta}_0)(\boldsymbol{J}(\boldsymbol{\theta}_0)^{-1}\boldsymbol{K}(\boldsymbol{\theta}_0)\mathrm{vec}(\boldsymbol{J}(\boldsymbol{\theta}_0)^{-1}\boldsymbol{I}(\boldsymbol{\theta}_0)\boldsymbol{J}(\boldsymbol{\theta}_0)^{-1}) \otimes \boldsymbol{I}(\boldsymbol{\theta}_0)^{-1})\right\},$$

$$\alpha_9 = E_{\boldsymbol{y}}\left[\boldsymbol{g}(\boldsymbol{y}_1 \mid \boldsymbol{\theta}_0)'\boldsymbol{I}(\boldsymbol{\theta}_0)^{-1}\boldsymbol{H}(\boldsymbol{y}_2 \mid \boldsymbol{\theta}_0)\boldsymbol{I}(\boldsymbol{\theta}_0)^{-1}\boldsymbol{H}(\boldsymbol{y}_1 \mid \boldsymbol{\theta}_0)\boldsymbol{I}(\boldsymbol{\theta}_0)^{-1}\boldsymbol{g}(\boldsymbol{y}_2 \mid \boldsymbol{\theta}_0)\right],$$

$$\alpha_{10} = E_{\boldsymbol{y}}\left[\mathrm{tr}\{\boldsymbol{K}(\boldsymbol{\theta}_0)(\boldsymbol{I}(\boldsymbol{\theta}_0)^{-1}\boldsymbol{g}(\boldsymbol{y} \mid \boldsymbol{\theta}_0) \otimes \boldsymbol{I}(\boldsymbol{\theta}_0)^{-1}\boldsymbol{H}(\boldsymbol{y} \mid \boldsymbol{\theta}_0)\boldsymbol{I}(\boldsymbol{\theta}_0)^{-1}\}\right],$$

$$\alpha_{11} = \mathrm{tr}\{\boldsymbol{K}(\boldsymbol{\theta}_0)'\boldsymbol{I}(\boldsymbol{\theta}_0)^{-1}\boldsymbol{K}(\boldsymbol{\theta}_0)(\boldsymbol{I}(\boldsymbol{\theta}_0)^{-1} \otimes \boldsymbol{I}(\boldsymbol{\theta}_0)^{-1})\},$$

$$\alpha_{12} = \mathrm{tr}\{\boldsymbol{L}(\boldsymbol{\theta}_0)(\boldsymbol{I}(\boldsymbol{\theta}_0)^{-1} \otimes \boldsymbol{I}(\boldsymbol{\theta}_0)^{-1})\}.$$

Bias of AIC (Main Result of Current Study)

$$B_{\text{AIC}} = 2(\alpha_1 - q) + \frac{1}{n}\Big(-\alpha_1 - 3\alpha_2 + \alpha_3 + 4\alpha_4 + 4\alpha_5 + 4\alpha_6 - 4\alpha_7 + \alpha_8 + 4\alpha_9 - 8\alpha_{10} + 2\alpha_{11} - \alpha_{12}\Big) + O(n^{-2}).$$

Useful Formulas in Obtaining the Coefficient Terms

If $\boldsymbol{W} \sim W_p(1, \boldsymbol{\Sigma}(\boldsymbol{\theta}_0))$, for any symmetric matrices $\boldsymbol{A}$, $\boldsymbol{B}$, and $\boldsymbol{C}$,

1. $E[\mathrm{tr}(\boldsymbol{WA})] = \mathrm{tr}(\boldsymbol{A})$,

2. $E[\mathrm{tr}(\boldsymbol{WA})\mathrm{tr}(\boldsymbol{WA})] = \mathrm{tr}(\boldsymbol{A})\mathrm{tr}(\boldsymbol{B}) + 2\mathrm{tr}(\boldsymbol{AB})$,

3. $$\begin{aligned} &E[\mathrm{tr}(\boldsymbol{WA})\mathrm{tr}(\boldsymbol{WB})\mathrm{tr}(\boldsymbol{WC})] = \mathrm{tr}(\boldsymbol{A})\mathrm{tr}(\boldsymbol{B})\mathrm{tr}(\boldsymbol{C}) \\ &+2\left\{\mathrm{tr}(\boldsymbol{A})\mathrm{tr}(\boldsymbol{BC}) + \mathrm{tr}(\boldsymbol{B})\mathrm{tr}(\boldsymbol{AC}) + \mathrm{tr}(\boldsymbol{C})\mathrm{tr}(\boldsymbol{AB})\right\} + 8\mathrm{tr}(\boldsymbol{ABC}). \end{aligned}$$

These expectations imply:

4. $E[\{\mathrm{tr}(\boldsymbol{A}-\boldsymbol{WA})\}\{\mathrm{tr}(\boldsymbol{B}-\boldsymbol{WB})\}] = 2\mathrm{tr}(\boldsymbol{AB})$,

5. $E[\{\mathrm{tr}(\boldsymbol{A}-\boldsymbol{WA})\}\{\mathrm{tr}(\boldsymbol{B}-\boldsymbol{WB})\}\{\mathrm{tr}(\boldsymbol{C}-\boldsymbol{WC})\}] = -8\mathrm{tr}(\boldsymbol{ABC})$.

Note: If $\boldsymbol{B}$ is not symmetric, then:

5(b).
$$E[\{\mathrm{tr}(\boldsymbol{A}-\boldsymbol{WA})\}\{\mathrm{tr}(\boldsymbol{B}-\boldsymbol{WB})\}\{\mathrm{tr}(\boldsymbol{C}-\boldsymbol{WC})\}]$$
$$= -4\{\mathrm{tr}(\boldsymbol{ABC}) + \mathrm{tr}(\boldsymbol{AB'C})\}.$$

Thank you!