Selecting Variance Structure in Mixed Effect Models by Information Criteria Based on Monte Carlo Approximations

Wataru Sakamoto

Osaka University, JAPAN

December 16th–19th, 2011
7th IASC-ARS (Joint 2011) in Taipei

Outline

Mixed effect models

Mixed effect model

Measurements = Fixed effect (mean, trend) + Random effect (between-subj.) + Error (within-subj.)

Linear mixed effect (LME) model (Laird and Ware, 1982)

$$y_{ij} = \boldsymbol{x}_{ij}^{\mathrm{T}}\boldsymbol{\beta} + \boldsymbol{z}_{ij}^{\mathrm{T}}\boldsymbol{b}_i + \epsilon_{ij}, \quad i = 1, \ldots, n;\ j = 1, \ldots, m_i$$

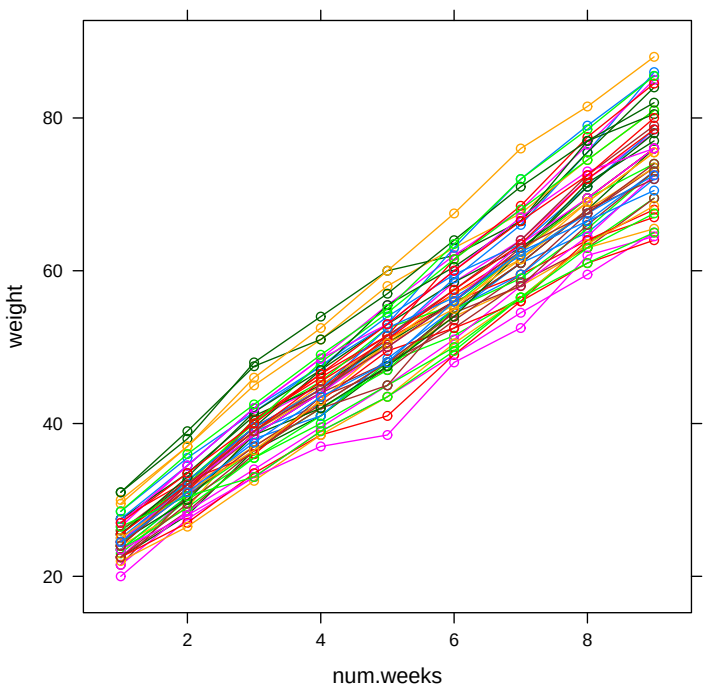
- $\boldsymbol{\beta}$: a parameter vector (itself estimated with ML).
- $\boldsymbol{b}_i$: a random vector (variance estimated with ML or REML).
- ϵ_{ij}: a random variable (variance estimated with ML or REML).

Longitudinal data

An example

Pig weights data (Diggle et al., 2002)

- Available from the R package `SemiPar` (Ruppert et al., 2003).
- Responses: body weights (observed at 1–9 weeks for each of 48 pigs)
- The variability between responses seems to increase as the number of weeks increases.



Mixed effect models for longitudinal data

An example: pig weights data

Three candidates of LME models:

- M_0 (with no random effect)

$$y_{ij} = \beta_0 + \beta_1 t_j + \quad + \epsilon_{ij}$$

- M_1 (with random effect for intercepts)

$$y_{ij} = \beta_0 + \beta_1 t_j + b_{0i} \quad + \epsilon_{ij}$$

- M_2 (with random effect for intercepts and slopes)

$$y_{ij} = \beta_0 + \beta_1 t_j + b_{0i} + b_{1i} t_j + \epsilon_{ij}$$

$i = 1, \ldots, 48$ (pigs); $t_j = j$, $j = 1, \ldots, 9$ (weeks)

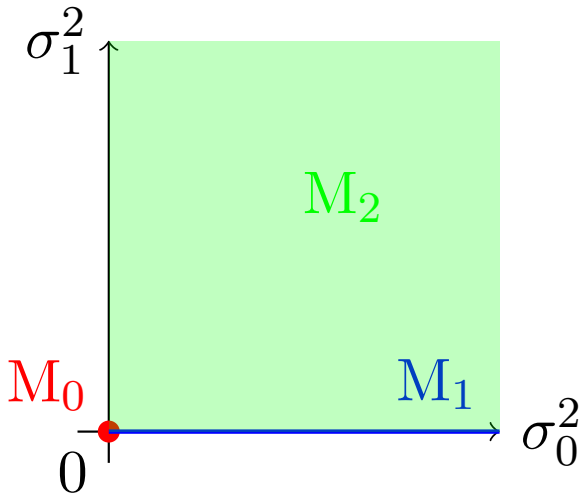
- Assumptions
  - $b_{0i} \sim \mathrm{N}(0, \sigma_0^2)$, $b_{1i} \sim \mathrm{N}(0, \sigma_1^2)$, $\epsilon_{ij} \sim \mathrm{N}(0, \sigma_e^2)$ (IID)

How to select variance structure in LME models

An issue on regularity conditions

The region of variance parameters is constrained. A regularity condition may NOT hold in selecting random effect terms.

- A true value may not exist *inside* the region of parameters.



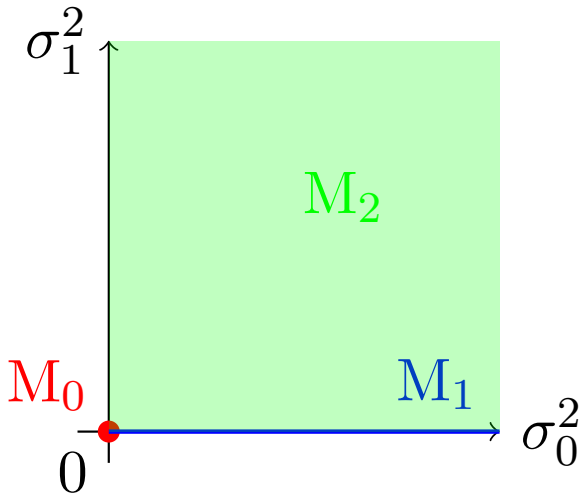
- Asymptotic distribution of the ML estimator
  - Not normal, if the true value is on the edge of the region.
  - Projection of multivariate normal distribution onto the region of parameters (Self and Liang, 1987).

How to select variance structure in LME models

An issue on regularity conditions

The region of variance parameters is constrained. A regularity condition may NOT hold in selecting random effect terms.

- A true value may not exist *inside* the region of parameters.



- Asymptotic distribution of the ML estimator
  - Not normal, if the true value is on the edge of the region.
  - Projection of multivariate normal distribution onto the region of parameters (Self and Liang, 1987).

How to select variance structure in LME models

Likelihood ratio test

Three candidates of LME models ($\mathrm{M}_0 \subset \mathrm{M}_1 \subset \mathrm{M}_2$)

- M_0: $y_{ij} = \beta_0 + \beta_1 t_j + \qquad\qquad + \epsilon_{ij}$
- M_1: $y_{ij} = \beta_0 + \beta_1 t_j + b_{0i} \qquad\quad + \epsilon_{ij}, \quad b_{0j} \sim \mathrm{N}(0, \sigma_0^2)$
- M_2: $y_{ij} = \beta_0 + \beta_1 t_j + b_{0i} + b_{1i} t_j + \epsilon_{ij}, \quad b_{1j} \sim \mathrm{N}(0, \sigma_1^2)$

- Hypotheses (considered on the largest model M_2)
  - H_0: $\sigma_0^2 = 0$ and $\sigma_1^2 = 0$ (with no random effect)
  - H_1: $\sigma_0^2 \neq 0$ and $\sigma_1^2 = 0$ (with random effect for only intercept)
  - H_2: $\sigma_1^2 \neq 0$ (with random effect for slopes)
- Asymptotic distribution of the log-likelihood ratio statistic
  - In balanced design cases: a mixture of chi-square densities (Self and Liang, 1987; Stram and Lee, 1994)
  - Simulation-based method (Crainiceanu and Ruppert, 2004)

How to select variance structure in LME models

Information Criteria

How to select an *optimal* model?

- To maximize the expected log-likelihood $\mu(\boldsymbol{\theta}) = \mathrm{E}_{\boldsymbol{y}}\{\log(\boldsymbol{y}|\boldsymbol{\theta})\}$ (that is, to minimize the Kullback–Leibler divergence)
- The maximum log-likelihood $l(\hat{\boldsymbol{\theta}}|\boldsymbol{y}_{\mathrm{obs}}) = \log f(\boldsymbol{y}_{\mathrm{obs}}|\hat{\boldsymbol{\theta}})$ is biased as an estimate of the expected log-likelihood.

$$\mathrm{Bias}\{l(\hat{\boldsymbol{\theta}})\} = \mathrm{E}_{\boldsymbol{y}_{\mathrm{obs}}}\{l(\hat{\boldsymbol{\theta}}|\boldsymbol{y}_{\mathrm{obs}}) - \mu(\hat{\boldsymbol{\theta}})\}$$

Information criteria

$$-2\,l(\hat{\theta}) + 2\,\widehat{\mathrm{Bias}}\{l(\hat{\theta})\}$$

- Consider a penalty term to adjust the bias in estimating the expected log-likelihood by the maximum log-likelihood.

How to select variance structure in LME models

Information Criteria

How to select an *optimal* model?

- To maximize the expected log-likelihood $\mu(\boldsymbol{\theta}) = \mathrm{E}_{\boldsymbol{y}}\{\log(\boldsymbol{y}|\boldsymbol{\theta})\}$ (that is, to minimize the Kullback–Leibler divergence)
- The maximum log-likelihood $l(\hat{\boldsymbol{\theta}}|\boldsymbol{y}_{\mathrm{obs}}) = \log f(\boldsymbol{y}_{\mathrm{obs}}|\hat{\boldsymbol{\theta}})$ is biased as an estimate of the expected log-likelihood.

$$\mathrm{Bias}\{l(\hat{\boldsymbol{\theta}})\} = \mathrm{E}_{\boldsymbol{y}_{\mathrm{obs}}}\{l(\hat{\boldsymbol{\theta}}|\boldsymbol{y}_{\mathrm{obs}}) - \mu(\hat{\boldsymbol{\theta}})\}$$

Information criteria

$$-2\,l(\hat{\boldsymbol{\theta}}) + 2\,\widehat{\mathrm{Bias}}\{l(\hat{\boldsymbol{\theta}})\}$$

- Consider a penalty term to adjust the bias in estimating the expected log-likelihood by the maximum log-likelihood.

How to select variance structure in LME models

Information Criteria

- AIC (Akaike information criteria) (Akaike, 1973)

$$\text{AIC} = -2\, l(\hat{\boldsymbol{\theta}}|\boldsymbol{y}_{\text{obs}}) + 2\text{dim}(\boldsymbol{\theta})$$

  - In choosing random effect terms, AIC does not give an asymptotically unbiased estimate of the expected log-likelihood (Greven and Kneib, 2009).

- CAIC (Conditional AIC) (Vaida and Blanchard, 2005)

$$\text{CAIC} = -2\, \log(\boldsymbol{y}_{\text{obs}}|\hat{\boldsymbol{\theta}}, \hat{\boldsymbol{b}}) + 2\Phi_0(\boldsymbol{y}_{\text{obs}})$$

  - Gives an unbiased estimate of the conditional expected log-likelihood if variance parameter is known.
  - Tends to choose a larger model if unknown variance parameters are estimated (Greven and Kneib, 2009).
  - Modified versions (Liang *et al.*, 2008; Greven and Kneib, 2009): are very complicated in computation.

How to select variance structure in LME models

Information Criteria

- AIC (Akaike information criteria) (Akaike, 1973)

$$\text{AIC} = -2\, l(\hat{\boldsymbol{\theta}}|\boldsymbol{y}_{\text{obs}}) + 2\text{dim}(\boldsymbol{\theta})$$

  - In choosing random effect terms, AIC does not give an asymptotically unbiased estimate of the expected log-likelihood (Greven and Kneib, 2009).

- CAIC (Conditional AIC) (Vaida and Blanchard, 2005)

$$\text{CAIC} = -2\, \log(\boldsymbol{y}_{\text{obs}}|\hat{\boldsymbol{\theta}}, \hat{\boldsymbol{b}}) + 2\Phi_0(\boldsymbol{y}_{\text{obs}})$$

  - Gives an unbiased estimate of the conditional expected log-likelihood if variance parameter is known.
  - Tends to choose a larger model if unknown variance parameters are estimated (Greven and Kneib, 2009).
  - Modified versions (Liang *et al.*, 2008; Greven and Kneib, 2009): are very complicated in computation.

How to select variance structure in LME models

Information Criteria

- BIC (Bayes information criteria) (Schwartz, 1978)

$$\text{BIC} = -2\, l(\hat{\boldsymbol{\theta}}|\boldsymbol{y}_{\text{obs}}) + \dim(\boldsymbol{\theta}) \cdot \log N$$

  - Has consistency in selecting the true model (if it exists).
  - Tends to choose a smaller model wrongly (more critical!).
- GIC (Generalized information criteria) (Pu and Niu, 2006)
  - The bias correction term is chosen in such a way that conditions to keep consistency in selecting the true model are satisfied.
- Bootstrap IC (Shang and Cavanaugh, 2008)
  - Would take too much time in computation.

Our proposal

A simpler information criterion that gives a consistent estimate of the expected log-likelihood even under parameter constraints.

How to select variance structure in LME models

Information Criteria

- BIC (Bayes information criteria) (Schwartz, 1978)

$$\text{BIC} = -2\, l(\hat{\boldsymbol{\theta}}|\boldsymbol{y}_{\text{obs}}) + \dim(\boldsymbol{\theta}) \cdot \log N$$

  - Has consistency in selecting the true model (if it exists).
  - Tends to choose a smaller model wrongly (more critical!).
- GIC (Generalized information criteria) (Pu and Niu, 2006)
  - The bias correction term is chosen in such a way that conditions to keep consistency in selecting the true model are satisfied.
- Bootstrap IC (Shang and Cavanaugh, 2008)
  - Would take too much time in computation.

Our proposal

A simpler information criterion that gives a consistent estimate of the expected log-likelihood even under parameter constraints.

How to select variance structure in LME models

Information Criteria

- BIC (Bayes information criteria) (Schwartz, 1978)

$$\text{BIC} = -2\, l(\hat{\boldsymbol{\theta}}|\boldsymbol{y}_{\text{obs}}) + \dim(\boldsymbol{\theta}) \cdot \log N$$

  - Has consistency in selecting the true model (if it exists).
  - Tends to choose a smaller model wrongly (more critical!).
- GIC (Generalized information criteria) (Pu and Niu, 2006)
  - The bias correction term is chosen in such a way that conditions to keep consistency in selecting the true model are satisfied.
- Bootstrap IC (Shang and Cavanaugh, 2008)
  - Would take too much time in computation.

Our proposal

A simpler information criterion that gives a consistent estimate of the expected log-likelihood even under parameter constraints.

How to select variance structure in LME models

Information Criteria

- BIC (Bayes information criteria) (Schwartz, 1978)

$$\text{BIC} = -2\, l(\hat{\boldsymbol{\theta}}|\boldsymbol{y}_{\text{obs}}) + \dim(\boldsymbol{\theta}) \cdot \log N$$

  - Has consistency in selecting the true model (if it exists).
  - Tends to choose a smaller model wrongly (more critical!).
- GIC (Generalized information criteria) (Pu and Niu, 2006)
  - The bias correction term is chosen in such a way that conditions to keep consistency in selecting the true model are satisfied.
- Bootstrap IC (Shang and Cavanaugh, 2008)
  - Would take too much time in computation.

Our proposal

A simpler information criterion that gives a consistent estimate of the expected log-likelihood even under parameter constraints.

Bias evaluation

Notations and Assumptions

Consider a balanced design

- Observation: $\boldsymbol{y}_i = (y_{i1}, \ldots, y_{im})^{\mathrm{T}}$, $i = 1, \ldots, m$
- True distribution: $\boldsymbol{y}_1, \ldots, \boldsymbol{y}_n \sim g(\boldsymbol{y})$ (IID)
- Fitted distribution: $\boldsymbol{y}_1, \ldots, \boldsymbol{y}_n \sim f(\boldsymbol{y}|\boldsymbol{\theta})$ (IID)
- Region of parameters: $\boldsymbol{\theta} \in \boldsymbol{\Theta} (\subset \mathbf{R}^d)$
  - $\boldsymbol{\Theta}$ is not necessarily open.

Assumptions on $f(\boldsymbol{y}|\boldsymbol{\theta})$ (regularity conditions)

- Identifiable: $f(\boldsymbol{y}|\boldsymbol{\theta}_1) \not\equiv f(\boldsymbol{y}|\boldsymbol{\theta}_2)$ if $\boldsymbol{\theta}_1 \neq \boldsymbol{\theta}_2$
- The support $\{\boldsymbol{y} : f(\boldsymbol{y}|\boldsymbol{\theta}) > 0\}$ does not depend on $\boldsymbol{\theta}$.
- Three-times differentiable w.r.t. $\boldsymbol{\theta}$.
- Expectation and differentiation w.r.t. $\boldsymbol{\theta}$ are interchangeable.

Bias evaluation

Notations and Assumptions

Consider a balanced design

- Observation: $\boldsymbol{y}_i = (y_{i1}, \ldots, y_{im})^{\mathrm{T}}$, $i = 1, \ldots, m$
- True distribution: $\boldsymbol{y}_1, \ldots, \boldsymbol{y}_n \sim g(\boldsymbol{y})$ (IID)
- Fitted distribution: $\boldsymbol{y}_1, \ldots, \boldsymbol{y}_n \sim f(\boldsymbol{y}|\boldsymbol{\theta})$ (IID)
- Region of parameters: $\boldsymbol{\theta} \in \boldsymbol{\Theta} (\subset \mathbf{R}^d)$
  - $\boldsymbol{\Theta}$ is not necessarily open.

Assumptions on $f(\boldsymbol{y}|\boldsymbol{\theta})$ (regularity conditions)

- Identifiable: $f(\boldsymbol{y}|\boldsymbol{\theta}_1) \not\equiv f(\boldsymbol{y}|\boldsymbol{\theta}_2)$ if $\boldsymbol{\theta}_1 \neq \boldsymbol{\theta}_2$
- The support $\{\boldsymbol{y} : f(\boldsymbol{y}|\boldsymbol{\theta}) > 0\}$ does not depend on $\boldsymbol{\theta}$.
- Three-times differentiable w.r.t. $\boldsymbol{\theta}$.
- Expectation and differentiation w.r.t. $\boldsymbol{\theta}$ are interchangeable.

Bias evaluation

Log-likelihood function: quadratic approximation

Log-likelihood function:

$$l_n(\boldsymbol{\theta}) = \sum_{i=1}^{n} \log f(\boldsymbol{y}_i|\boldsymbol{\theta})$$
$$\approx l_n(\boldsymbol{\theta}_0) + S_n(\boldsymbol{\theta}_0)^{\mathrm{T}}(\boldsymbol{\theta} - \boldsymbol{\theta}_0) - \tfrac{1}{2}(\boldsymbol{\theta} - \boldsymbol{\theta}_0)^{\mathrm{T}} H_n(\boldsymbol{\theta}_0)(\boldsymbol{\theta} - \boldsymbol{\theta}_0)$$

- $\boldsymbol{\theta}_0$: A maximizer of the expected log-likelihood (next slide)
- $\hat{\boldsymbol{\theta}}_n = \arg\max_{\boldsymbol{\theta} \in \boldsymbol{\Theta}} l_n(\boldsymbol{\theta})$: Maximum likelihood estimator
  - Not necessarily a solution of the likelihood equation $S_n(\boldsymbol{\theta}) = 0$.
- $S_n(\boldsymbol{\theta}) = \sum \frac{\partial}{\partial \boldsymbol{\theta}} \log f(\boldsymbol{y}_i|\boldsymbol{\theta})$: Score function (vector)
- $H_n(\boldsymbol{\theta}) = -\sum \frac{\partial^2}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^{\mathrm{T}}} \log f(\boldsymbol{y}_i|\boldsymbol{\theta})$
  : Observed information (matrix)

Bias evaluation

Expected log-likelihood function: quadratic approximation

Expected log-likelihood function:

$$\begin{aligned}\mu(\boldsymbol{\theta}) &= \mathrm{E}_g\{\log f(\boldsymbol{y}|\boldsymbol{\theta})\} \\ &\approx \mu(\boldsymbol{\theta}_0) + \delta(\boldsymbol{\theta}_0)^{\mathrm{T}}(\boldsymbol{\theta}-\boldsymbol{\theta}_0) - \tfrac{1}{2}(\boldsymbol{\theta}-\boldsymbol{\theta}_0)^{\mathrm{T}}\mathcal{J}(\boldsymbol{\theta}_0)(\boldsymbol{\theta}-\boldsymbol{\theta}_0)\end{aligned}$$

- $\boldsymbol{\theta}_0 = \arg\max_{\boldsymbol{\theta}\in\Theta} \mu(\boldsymbol{\theta})$, which gives minimum KL divergence.
- $\delta(\boldsymbol{\theta}) = \mathrm{E}_g\left\{\frac{\partial}{\partial\boldsymbol{\theta}}\log f(\boldsymbol{y}|\boldsymbol{\theta})\right\}$
- $\mathcal{J}(\boldsymbol{\theta}) = \mathrm{E}_g\left\{-\frac{\partial^2}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^{\mathrm{T}}}\log f(\boldsymbol{y}|\boldsymbol{\theta})\right\}$

If the true model is included in the class of fitted models ($g(\boldsymbol{y}) \equiv f(\boldsymbol{y}|\boldsymbol{\theta}_0)$),

- $\delta(\boldsymbol{\theta}_0) = 0$
- $\mathcal{J}(\boldsymbol{\theta}_0) = \mathcal{I}(\boldsymbol{\theta}_0) \equiv \mathrm{E}_{\boldsymbol{\theta}_0}\left\{\frac{\partial}{\partial\boldsymbol{\theta}}\log f(\boldsymbol{y}|\boldsymbol{\theta}_0)\cdot\frac{\partial}{\partial\boldsymbol{\theta}^{\mathrm{T}}}\log f(\boldsymbol{y}|\boldsymbol{\theta}_0)\right\}$
  : Fisher information matrix (assumed to be pos. def.)

Bias evaluation

Expected log-likelihood function: quadratic approximation

Expected log-likelihood function:

$$\begin{aligned}\mu(\boldsymbol{\theta}) &= \mathrm{E}_g\{\log f(\boldsymbol{y}|\boldsymbol{\theta})\} \\ &\approx \mu(\boldsymbol{\theta}_0) + \delta(\boldsymbol{\theta}_0)^{\mathrm{T}}(\boldsymbol{\theta}-\boldsymbol{\theta}_0) - \tfrac{1}{2}(\boldsymbol{\theta}-\boldsymbol{\theta}_0)^{\mathrm{T}}\mathcal{J}(\boldsymbol{\theta}_0)(\boldsymbol{\theta}-\boldsymbol{\theta}_0)\end{aligned}$$

- $\boldsymbol{\theta}_0 = \arg\max_{\boldsymbol{\theta}\in\Theta} \mu(\boldsymbol{\theta})$, which gives minimum KL divergence.
- $\delta(\boldsymbol{\theta}) = \mathrm{E}_g\left\{\frac{\partial}{\partial\boldsymbol{\theta}}\log f(\boldsymbol{y}|\boldsymbol{\theta})\right\}$
- $\mathcal{J}(\boldsymbol{\theta}) = \mathrm{E}_g\left\{-\frac{\partial^2}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^{\mathrm{T}}}\log f(\boldsymbol{y}|\boldsymbol{\theta})\right\}$

If the true model is included in the class of fitted models ($g(\boldsymbol{y}) \equiv f(\boldsymbol{y}|\boldsymbol{\theta}_0)$),

- $\delta(\boldsymbol{\theta}_0) = 0$
- $\mathcal{J}(\boldsymbol{\theta}_0) = \mathcal{I}(\boldsymbol{\theta}_0) \equiv \mathrm{E}_{\boldsymbol{\theta}_0}\left\{\frac{\partial}{\partial\boldsymbol{\theta}}\log f(\boldsymbol{y}|\boldsymbol{\theta}_0)\cdot\frac{\partial}{\partial\boldsymbol{\theta}^{\mathrm{T}}}\log f(\boldsymbol{y}|\boldsymbol{\theta}_0)\right\}$
  : Fisher information matrix (assumed to be pos. def.)

Bias evaluation

Quadratic approximation of MLE under parameter constraints

Expected log-likelihood function:

$$\mu(\boldsymbol{\theta}) \approx \mu(\boldsymbol{\theta}_0) - \tfrac{1}{2}(\boldsymbol{\theta} - \check{\boldsymbol{\theta}}_0)^{\mathrm{T}} \mathcal{J}(\boldsymbol{\theta}_0)(\boldsymbol{\theta} - \check{\boldsymbol{\theta}}_0) + \tfrac{1}{2}\delta(\boldsymbol{\theta}_0)^{\mathrm{T}} \mathcal{J}(\boldsymbol{\theta}_0)^{-1}\delta(\boldsymbol{\theta}_0)$$

- $\check{\boldsymbol{\theta}}_0 = \boldsymbol{\theta}_0 + \mathcal{J}(\boldsymbol{\theta}_0)^{-1}\delta(\boldsymbol{\theta}_0)$
- An approximation of $\boldsymbol{\theta}_0 = \arg\max_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} \mu(\boldsymbol{\theta})$:

$$\tilde{\boldsymbol{\theta}}_0 = \arg\min_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} (\boldsymbol{\theta} - \check{\boldsymbol{\theta}}_0)^{\mathrm{T}} \mathcal{J}(\boldsymbol{\theta}_0)(\boldsymbol{\theta} - \check{\boldsymbol{\theta}}_0)$$

Log-likelihood function:

$$l_n(\boldsymbol{\theta}) \approx l_n(\boldsymbol{\theta}_0) - \tfrac{1}{2}(\boldsymbol{\theta} - \check{\boldsymbol{\theta}}_n)^{\mathrm{T}} H_n(\boldsymbol{\theta}_0)(\boldsymbol{\theta} - \check{\boldsymbol{\theta}}_n) + \tfrac{1}{2} S_n(\boldsymbol{\theta}_0)^{\mathrm{T}} H_n(\boldsymbol{\theta}_0)^{-1} S_n(\boldsymbol{\theta}_0)$$

- $\check{\boldsymbol{\theta}}_n = \boldsymbol{\theta}_0 + H_n(\boldsymbol{\theta}_0)^{-1} S_n(\boldsymbol{\theta}_0)$
- An approximation of MLE $\hat{\boldsymbol{\theta}}_n = \arg\max_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} l_n(\boldsymbol{\theta})$:

$$\tilde{\boldsymbol{\theta}}_n = \arg\min_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} (\boldsymbol{\theta} - \check{\boldsymbol{\theta}}_n)^{\mathrm{T}} H_n(\boldsymbol{\theta}_0)(\boldsymbol{\theta} - \check{\boldsymbol{\theta}}_n)$$

Bias evaluation

Quadratic approximation of MLE under parameter constraints

Expected log-likelihood function:

$$\mu(\boldsymbol{\theta}) \approx \mu(\boldsymbol{\theta}_0) - \tfrac{1}{2}(\boldsymbol{\theta} - \check{\boldsymbol{\theta}}_0)^{\mathrm{T}} \mathcal{J}(\boldsymbol{\theta}_0)(\boldsymbol{\theta} - \check{\boldsymbol{\theta}}_0) + \tfrac{1}{2}\delta(\boldsymbol{\theta}_0)^{\mathrm{T}} \mathcal{J}(\boldsymbol{\theta}_0)^{-1}\delta(\boldsymbol{\theta}_0)$$

- $\check{\boldsymbol{\theta}}_0 = \boldsymbol{\theta}_0 + \mathcal{J}(\boldsymbol{\theta}_0)^{-1}\delta(\boldsymbol{\theta}_0)$
- An approximation of $\boldsymbol{\theta}_0 = \arg\max_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} \mu(\boldsymbol{\theta})$:

$$\tilde{\boldsymbol{\theta}}_0 = \arg\min_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} (\boldsymbol{\theta} - \check{\boldsymbol{\theta}}_0)^{\mathrm{T}} \mathcal{J}(\boldsymbol{\theta}_0)(\boldsymbol{\theta} - \check{\boldsymbol{\theta}}_0)$$

Log-likelihood function:

$$l_n(\boldsymbol{\theta}) \approx l_n(\boldsymbol{\theta}_0) - \tfrac{1}{2}(\boldsymbol{\theta} - \check{\boldsymbol{\theta}}_n)^{\mathrm{T}} H_n(\boldsymbol{\theta}_0)(\boldsymbol{\theta} - \check{\boldsymbol{\theta}}_n) + \tfrac{1}{2} S_n(\boldsymbol{\theta}_0)^{\mathrm{T}} H_n(\boldsymbol{\theta}_0)^{-1} S_n(\boldsymbol{\theta}_0)$$

- $\check{\boldsymbol{\theta}}_n = \boldsymbol{\theta}_0 + H_n(\boldsymbol{\theta}_0)^{-1} S_n(\boldsymbol{\theta}_0)$
- An approximation of MLE $\hat{\boldsymbol{\theta}}_n = \arg\max_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} l_n(\boldsymbol{\theta})$:

$$\tilde{\boldsymbol{\theta}}_n = \arg\min_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} (\boldsymbol{\theta} - \check{\boldsymbol{\theta}}_n)^{\mathrm{T}} H_n(\boldsymbol{\theta}_0)(\boldsymbol{\theta} - \check{\boldsymbol{\theta}}_n)$$

Bias evaluation

Asymptotic distribution of MLE under parameter constraints

For simplicity, consider the case $g(\boldsymbol{y}) \equiv f(\boldsymbol{y}|\boldsymbol{\theta}_0)$, so $\delta(\boldsymbol{\theta}_0) = 0$.

Theorem (Self and Liang, 1987)

Under some convexity conditions on $\boldsymbol{\Theta}$ in the neighborhood of $\boldsymbol{\theta}_0$,

- $\sqrt{n}(\hat{\boldsymbol{\theta}}_n - \tilde{\boldsymbol{\theta}}_n) = o_p(1)$
- The asymptotic distribution of $\hat{\boldsymbol{\theta}}_n$ is the same as that of

$$\arg\min_{\boldsymbol{\theta} \in \boldsymbol{\Theta}} (\boldsymbol{X} - \boldsymbol{\theta})^{\mathrm{T}} \mathcal{J}(\boldsymbol{\theta}_0)(\boldsymbol{X} - \boldsymbol{\theta}),$$

  where $\boldsymbol{X} \sim \mathrm{N}_d(\boldsymbol{\theta}_0, n^{-1}\mathcal{J}(\boldsymbol{\theta}_0)^{-1})$.
  - The MLE of $\boldsymbol{\theta} \in \boldsymbol{\Theta}$ for a sample from $\mathrm{N}_d(\boldsymbol{\theta}, n^{-1}\mathcal{J}(\boldsymbol{\theta}_0)^{-1})$.

- $n^{-1}S_n(\boldsymbol{\theta}_0) \xrightarrow{L} \mathrm{N}_d(\mathbf{0}, n^{-1}\mathcal{I}(\boldsymbol{\theta}_0))$ (by CLT)
- $n^{-1}H_n(\boldsymbol{\theta}_0) \xrightarrow{P} \mathcal{J}(\boldsymbol{\theta}_0) = \mathcal{I}(\boldsymbol{\theta}_0)$ (by LLN)

Bias evaluation

Asymptotic bias under parameter constraints

The bias of maximum log-likelihood as an estimate of the expected log-likelihood:

$$
\begin{aligned}
b_n(\hat{\boldsymbol{\theta}}_n) &= \mathrm{E}_g\{l_n(\hat{\boldsymbol{\theta}}_n) - n\mu(\hat{\boldsymbol{\theta}}_n)\} \\
&= \mathrm{E}_g\{l_n(\hat{\boldsymbol{\theta}}_n) - l_n(\boldsymbol{\theta}_0)\} + \mathrm{E}_g\{l_n(\boldsymbol{\theta}_0) - n\mu(\boldsymbol{\theta}_0)\} + n\mathrm{E}_g\{\mu(\boldsymbol{\theta}_0) - \mu(\hat{\boldsymbol{\theta}}_n)\} \\
&= \mathrm{E}_g[\{S_n(\boldsymbol{\theta}_0) - n\delta(\boldsymbol{\theta}_0)\}^{\mathrm{T}}(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0)] \\
&\quad - \tfrac{1}{2}\mathrm{E}_g[(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0)^{\mathrm{T}}\{H_n(\boldsymbol{\theta}_0) - n\mathcal{J}(\boldsymbol{\theta}_0)\}(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0)] + o_p(n^{-1/2}) \\
&= \mathrm{E}_g[\{S_n(\boldsymbol{\theta}_0) - n\delta(\boldsymbol{\theta}_0)\}^{\mathrm{T}}(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0)] + o_p(1)
\end{aligned}
$$

The quadratic form term is $o_p(1)$ because

- $\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0 = O_p(n^{-1/2})$, and
- $n^{-1}H_n(\boldsymbol{\theta}_0) = \mathcal{J}(\boldsymbol{\theta}_0) + o_p(1)$.

Bias evaluation

Asymptotic bias under parameter constraints

In the case $g(\boldsymbol{y}) \equiv f(\boldsymbol{y}|\boldsymbol{\theta}_0)$, so that $\delta(\boldsymbol{\theta}_0) = 0$, then

Asymptotic bias under parameter constraints

$$b_n(\hat{\boldsymbol{\theta}}_n) = n\mathrm{E}_{\boldsymbol{\theta}_0}[(\check{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0)^{\mathrm{T}} \mathcal{J}(\boldsymbol{\theta}_0)(\tilde{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0)] + o_p(1)$$

- $\check{\boldsymbol{\theta}}_n = \boldsymbol{\theta}_0 + H_n(\boldsymbol{\theta}_0)^{-1} S_n(\boldsymbol{\theta}_0)$ (not necessarily in $\boldsymbol{\Theta}$)
- $\tilde{\boldsymbol{\theta}}_n = \arg\min_{\boldsymbol{\theta} \in \boldsymbol{\Theta}} (\boldsymbol{\theta} - \check{\boldsymbol{\theta}}_n)^{\mathrm{T}} H_n(\boldsymbol{\theta}_0)(\boldsymbol{\theta} - \check{\boldsymbol{\theta}}_n)$

If $\Pr(\check{\boldsymbol{\theta}}_n = \tilde{\boldsymbol{\theta}}_n) \to 1$ ($\tilde{\boldsymbol{\theta}}_n$ is almost surely in the interior of $\boldsymbol{\Theta}$),

$$\begin{aligned} b_n(\hat{\boldsymbol{\theta}}_n) &= n\mathrm{E}_{\boldsymbol{\theta}_0}\{(\tilde{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0)^{\mathrm{T}} \mathcal{J}_n(\boldsymbol{\theta}_0)(\tilde{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0)\} + o_p(1) \\ &= \mathrm{tr}\{\mathcal{J}_n(\boldsymbol{\theta}_0)\mathcal{I}_n(\boldsymbol{\theta}_0)^{-1}\} + o_p(1) = p + o_p(1), \end{aligned}$$

which coincides with the bias correction term in the AIC without parameter constraints.

Bias evaluation

Asymptotic bias under parameter constraints

In the case $g(\boldsymbol{y}) \equiv f(\boldsymbol{y}|\boldsymbol{\theta}_0)$, so that $\delta(\boldsymbol{\theta}_0) = 0$, then

Asymptotic bias under parameter constraints

$$b_n(\hat{\boldsymbol{\theta}}_n) = n\mathrm{E}_{\boldsymbol{\theta}_0}[(\check{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0)^{\mathrm{T}} \mathcal{J}(\boldsymbol{\theta}_0)(\tilde{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0)] + o_p(1)$$

- $\check{\boldsymbol{\theta}}_n = \boldsymbol{\theta}_0 + H_n(\boldsymbol{\theta}_0)^{-1} S_n(\boldsymbol{\theta}_0)$ (not necessarily in $\boldsymbol{\Theta}$)
- $\tilde{\boldsymbol{\theta}}_n = \arg\min_{\boldsymbol{\theta} \in \boldsymbol{\Theta}} (\boldsymbol{\theta} - \check{\boldsymbol{\theta}}_n)^{\mathrm{T}} H_n(\boldsymbol{\theta}_0)(\boldsymbol{\theta} - \check{\boldsymbol{\theta}}_n)$

If $\Pr(\check{\boldsymbol{\theta}}_n = \tilde{\boldsymbol{\theta}}_n) \to 1$ ($\tilde{\boldsymbol{\theta}}_n$ is almost surely in the interior of $\boldsymbol{\Theta}$),

$$\begin{aligned} b_n(\hat{\boldsymbol{\theta}}_n) &= n\mathrm{E}_{\boldsymbol{\theta}_0}\{(\tilde{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0)^{\mathrm{T}} \mathcal{J}_n(\boldsymbol{\theta}_0)(\tilde{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0)\} + o_p(1) \\ &= \mathrm{tr}\{\mathcal{J}_n(\boldsymbol{\theta}_0)\mathcal{I}_n(\boldsymbol{\theta}_0)^{-1}\} + o_p(1) = p + o_p(1), \end{aligned}$$

which coincides with the bias correction term in the AIC without parameter constraints.

Bootstrap estimate of the bias

Bootstrap estimate of the expected log-likelihood function (quadratic approximation):

$$\begin{aligned}\mu_*(\boldsymbol{\theta}) &= \mathrm{E}_*\{\log f(\boldsymbol{y}^*|\boldsymbol{\theta})\} \\ &\approx \mu_*(\hat{\boldsymbol{\theta}}) + \delta_*(\hat{\boldsymbol{\theta}})^{\mathrm{T}}(\boldsymbol{\theta}-\hat{\boldsymbol{\theta}}) - \tfrac{1}{2}(\boldsymbol{\theta}-\hat{\boldsymbol{\theta}})^{\mathrm{T}}\mathcal{J}_*(\hat{\boldsymbol{\theta}})(\boldsymbol{\theta}-\hat{\boldsymbol{\theta}})\end{aligned}$$

where

- $\boldsymbol{y}^*$: An observation from the bootstrap distribution
- $\hat{\boldsymbol{\theta}}$: MLE from the original sample
- $\delta_*(\boldsymbol{\theta}) = \mathrm{E}_*\left\{\frac{\partial}{\partial\boldsymbol{\theta}}\log f(\boldsymbol{y}^*|\boldsymbol{\theta})\right\}$
- $\mathcal{J}_*(\boldsymbol{\theta}) = \mathrm{E}_*\left\{-\frac{\partial^2}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^{\mathrm{T}}}\log f(\boldsymbol{y}^*|\boldsymbol{\theta})\right\}$
  - In case of nonparametric bootstrap,

$$\delta_*(\boldsymbol{\theta}) = n^{-1}S_n(\boldsymbol{\theta}), \quad \text{and} \quad \mathcal{J}_*(\boldsymbol{\theta}) = n^{-1}H_n(\boldsymbol{\theta}).$$

Bootstrap estimate of the bias

Log-likelihood function for a bootstrap sample (quadratic approximation):

$$
\begin{aligned}
l_*(\boldsymbol{\theta}) &= \sum_{i=1}^{n} \log f(\boldsymbol{y}_i^* | \boldsymbol{\theta}) \\
&\approx l_*(\hat{\boldsymbol{\theta}}) + S_*(\hat{\boldsymbol{\theta}})^{\mathrm{T}}(\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}) - \tfrac{1}{2}(\boldsymbol{\theta} - \hat{\boldsymbol{\theta}})^{\mathrm{T}} H_*(\hat{\boldsymbol{\theta}})(\boldsymbol{\theta} - \hat{\boldsymbol{\theta}})
\end{aligned}
$$

- $S_*(\boldsymbol{\theta}) = \frac{\partial}{\partial \boldsymbol{\theta}} l_*(\boldsymbol{\theta})$: Score function (vector)
- $H_*(\boldsymbol{\theta}) = \frac{\partial^2}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^{\mathrm{T}}} l_*(\boldsymbol{\theta})$: Observed information (matrix)
- $\hat{\boldsymbol{\theta}}_* = \arg\max_{\boldsymbol{\theta} \in \boldsymbol{\Theta}} l_*(\boldsymbol{\theta})$: MLE for the bootstrap sample

Bootstrap estimate of the bias

Bootstrap estimate of the bias of maximum log-likelihood (MLL) as an estimate of the expected log-likelihood (ELL)
(Shang and Cavanaugh, 2008)

$$b_*(\hat{\boldsymbol{\theta}}_*) = \mathrm{E}_*\{l_*(\hat{\boldsymbol{\theta}}_*) - n\mu_*(\hat{\boldsymbol{\theta}}_*)\}$$

Quadratic approximation around the original MLE $\hat{\boldsymbol{\theta}}$:

$$\begin{aligned} b_*(\hat{\boldsymbol{\theta}}_*) &= \mathrm{E}_*\{l_*(\hat{\boldsymbol{\theta}}_*) - l_*(\hat{\boldsymbol{\theta}})\} + \mathrm{E}_*\{l_*(\hat{\boldsymbol{\theta}}) - n\mu_*(\hat{\boldsymbol{\theta}})\} + n\mathrm{E}_*\{\mu_*(\hat{\boldsymbol{\theta}}) - \mu_*(\hat{\boldsymbol{\theta}}_*)\} \\ &\approx \mathrm{E}_*[\{S_*(\hat{\boldsymbol{\theta}}) - n\delta_*(\hat{\boldsymbol{\theta}})\}^{\mathrm{T}}(\hat{\boldsymbol{\theta}}_* - \hat{\boldsymbol{\theta}})] \end{aligned}$$

In the case $g(\boldsymbol{y}) \equiv f(\boldsymbol{y}|\theta_0)$,

$$b_n(\hat{\boldsymbol{\theta}}_*) \approx \mathrm{E}_*\{(\check{\boldsymbol{\theta}}_* - \hat{\boldsymbol{\theta}})^{\mathrm{T}} \mathcal{J}_*(\hat{\boldsymbol{\theta}})(\tilde{\boldsymbol{\theta}}_* - \hat{\boldsymbol{\theta}})\}$$

where

- $\check{\boldsymbol{\theta}}_* = \hat{\boldsymbol{\theta}} + H_*(\hat{\boldsymbol{\theta}})^{-1} S_*(\hat{\boldsymbol{\theta}})$ (not necessarily in Θ)
- $\tilde{\boldsymbol{\theta}}_* = \arg\min_{\boldsymbol{\theta}\in\Theta}(\boldsymbol{\theta} - \check{\boldsymbol{\theta}}_*)^{\mathrm{T}} H_*(\hat{\boldsymbol{\theta}})(\boldsymbol{\theta} - \check{\boldsymbol{\theta}}_*)$

Bootstrap estimate of the bias

Bootstrap estimate of the bias of maximum log-likelihood (MLL) as an estimate of the expected log-likelihood (ELL)
(Shang and Cavanaugh, 2008)

$$b_*(\hat{\boldsymbol{\theta}}_*) = \mathrm{E}_*\{l_*(\hat{\boldsymbol{\theta}}_*) - n\mu_*(\hat{\boldsymbol{\theta}}_*)\}$$

Quadratic approximation around the original MLE $\hat{\boldsymbol{\theta}}$:

$$\begin{aligned} b_*(\hat{\boldsymbol{\theta}}_*) &= \mathrm{E}_*\{l_*(\hat{\boldsymbol{\theta}}_*) - l_*(\hat{\boldsymbol{\theta}})\} + \mathrm{E}_*\{l_*(\hat{\boldsymbol{\theta}}) - n\mu_*(\hat{\boldsymbol{\theta}})\} + n\mathrm{E}_*\{\mu_*(\hat{\boldsymbol{\theta}}) - \mu_*(\hat{\boldsymbol{\theta}}_*)\} \\ &\approx \mathrm{E}_*[\{S_*(\hat{\boldsymbol{\theta}}) - n\delta_*(\hat{\boldsymbol{\theta}})\}^{\mathrm{T}}(\hat{\boldsymbol{\theta}}_* - \hat{\boldsymbol{\theta}})] \end{aligned}$$

In the case $g(\boldsymbol{y}) \equiv f(\boldsymbol{y}|\boldsymbol{\theta}_0)$,

$$b_n(\hat{\boldsymbol{\theta}}_*) \approx \mathrm{E}_*\{(\check{\boldsymbol{\theta}}_* - \hat{\boldsymbol{\theta}})^{\mathrm{T}}\mathcal{J}_*(\hat{\boldsymbol{\theta}})(\tilde{\boldsymbol{\theta}}_* - \hat{\boldsymbol{\theta}})\}$$

where

- $\check{\boldsymbol{\theta}}_* = \hat{\boldsymbol{\theta}} + H_*(\hat{\boldsymbol{\theta}})^{-1}S_*(\hat{\boldsymbol{\theta}})$ (not necessarily in $\boldsymbol{\Theta}$)
- $\tilde{\boldsymbol{\theta}}_* = \arg\min_{\boldsymbol{\theta}\in\boldsymbol{\Theta}}(\boldsymbol{\theta} - \check{\boldsymbol{\theta}}_*)^{\mathrm{T}}H_*(\hat{\boldsymbol{\theta}})(\boldsymbol{\theta} - \check{\boldsymbol{\theta}}_*)$

Monte Carlo approximation of the bias correction term

We wouldn't need to conduct bootstrapping actually.

According to the asymptotic distribution of $\check{\boldsymbol{\theta}}_n$ and $\tilde{\boldsymbol{\theta}}_n$, generate the following random samples for $b = 1, \ldots, B$:

- $\check{\boldsymbol{\theta}}_{(b)} = \hat{\boldsymbol{\theta}} + n^{-1/2} \mathcal{J}(\hat{\boldsymbol{\theta}})^{-1/2} \boldsymbol{Z}_{(b)}$, where $\boldsymbol{Z}_{(b)} \sim \mathrm{N}_d(\mathbf{0}, \mathbf{I}_d)$, and
- $\tilde{\boldsymbol{\theta}}_{(b)} = \arg\min_{\boldsymbol{\theta} \in \boldsymbol{\Theta}} (\boldsymbol{\theta} - \check{\boldsymbol{\theta}}_{(b)})^{\mathrm{T}} \mathcal{J}(\hat{\boldsymbol{\theta}}) (\boldsymbol{\theta} - \check{\boldsymbol{\theta}}_{(b)})$

Then, estimate the asymptotic bias as follows:

$$\hat{b}(\hat{\boldsymbol{\theta}}) = B^{-1} \sum_{b=1}^{B} \{ (\check{\boldsymbol{\theta}}_{(b)} - \hat{\boldsymbol{\theta}})^{\mathrm{T}} \mathcal{J}(\hat{\boldsymbol{\theta}}) (\tilde{\boldsymbol{\theta}}_{(b)} - \hat{\boldsymbol{\theta}}) \}$$

An information criterion under parameter constraints

$$\mathrm{IC}_{\mathrm{PC}} = -2\, l_n(\hat{\boldsymbol{\theta}}) + 2\hat{b}(\hat{\boldsymbol{\theta}})$$

Computation of the bias correction term for LME models

Maximum likelihood (ML) estimation

LME model to fit: $\boldsymbol{y}_i = \boldsymbol{X}_i\boldsymbol{\beta} + \boldsymbol{Z}_i\boldsymbol{b}_i + \boldsymbol{e}_i,\ \ i = 1, \ldots, n$

- $\boldsymbol{b}_i \sim \mathrm{N}_r(\boldsymbol{0}, \sigma^2\boldsymbol{G})$, $\boldsymbol{G} = \boldsymbol{G}(\boldsymbol{\psi})$, $\boldsymbol{\psi} = (\psi_1, \ldots, \psi_q)^{\mathrm{T}}$
- $\boldsymbol{e}_i \sim \mathrm{N}_m(\boldsymbol{0}, \sigma^2\mathbf{I}_m)$

Fisher information matrix ($\boldsymbol{\theta} = (\boldsymbol{\beta}, \boldsymbol{\psi}, \sigma^2)^{\mathrm{T}}$):

$$\mathcal{J}(\boldsymbol{\theta}) = \begin{pmatrix} \sum_i \boldsymbol{X}_i^{\mathrm{T}}\boldsymbol{V}_i^{-1}\boldsymbol{X}_i & O & O \\ O & \mathcal{J}_{\psi\psi} & \mathcal{J}_{\psi\sigma^2} \\ O & \mathcal{J}_{\sigma^2\psi} & \mathcal{J}_{\sigma^2\sigma^2} \end{pmatrix}$$

- $\boldsymbol{V}_i = \mathrm{Var}(\boldsymbol{y}_i) = \sigma^2(\boldsymbol{Z}_i\boldsymbol{G}\boldsymbol{Z}_i^{\mathrm{T}} + \mathbf{I}_m)$
- $(\mathcal{J}_{\psi\psi})_{jk} = \frac{1}{2}\sum_i \mathrm{tr}\left(\boldsymbol{V}_i^{-1}\frac{\partial \boldsymbol{V}_i}{\partial \psi_j}\boldsymbol{V}_i^{-1}\frac{\partial \boldsymbol{V}_i}{\partial \psi_k}\right)$ etc.

Estimate of the asymptotic bias:

if $\boldsymbol{\beta}_0 \in \mathbf{R}^p$ and $\sigma_0^2 (> 0)$ are supposed to be at an interior point,

$$\hat{b}(\hat{\boldsymbol{\theta}}) = (p+1) + B^{-1}\sum_{b=1}^{B}[(\check{\boldsymbol{\psi}}_{(b)} - \hat{\boldsymbol{\psi}})^{\mathrm{T}}\{\mathcal{J}^{\psi\psi}(\hat{\boldsymbol{\theta}})\}^{-1}(\tilde{\boldsymbol{\psi}}_{(b)} - \hat{\boldsymbol{\psi}})],$$

where $\mathcal{J}^{\psi\psi}(\hat{\boldsymbol{\theta}})$ is the elements of $\mathcal{J}(\hat{\boldsymbol{\theta}})^{-1}$ corresponding to $\boldsymbol{\psi}$.

Computation of the bias correction term for LME models

Restricted maximum likelihood (REML) estimation

LME model to fit: $\boldsymbol{y}_i = \boldsymbol{X}_i\boldsymbol{\beta} + \boldsymbol{Z}_i\boldsymbol{b}_i + \boldsymbol{e}_i,\ \ i = 1, \ldots, n$

- $\boldsymbol{b}_i \sim \mathrm{N}_r(\boldsymbol{0}, \sigma^2\boldsymbol{G})$, $\boldsymbol{G} = \boldsymbol{G}(\boldsymbol{\psi})$, $\boldsymbol{\psi} = (\psi_1, \ldots, \psi_q)^{\mathrm{T}}$
- $\boldsymbol{e}_i \sim \mathrm{N}_m(\boldsymbol{0}, \sigma^2\mathbf{I}_m)$

Fisher information matrix ($\boldsymbol{\theta} = (\boldsymbol{\psi}, \sigma^2)^{\mathrm{T}}$):

$$\mathcal{J}_R(\boldsymbol{\theta}) = \begin{pmatrix} \mathcal{J}_{R\psi\psi} & \mathcal{J}_{R\psi\sigma^2} \\ \mathcal{J}_{R\sigma^2\psi} & \mathcal{J}_{R\sigma^2\sigma^2} \end{pmatrix}$$

- $\boldsymbol{V}_i = \mathrm{Var}(\boldsymbol{y}_i) = \sigma^2(\boldsymbol{Z}_i\boldsymbol{G}\boldsymbol{Z}_i^{\mathrm{T}} + \mathbf{I}_m)$
- $\boldsymbol{P}_V = \boldsymbol{V}^{-1} - \boldsymbol{V}^{-1}\boldsymbol{X}(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{V}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{V}^{-1}$
- $(\mathcal{J}_{R\psi\psi})_{jk} = \frac{1}{2}\mathrm{tr}\left(\boldsymbol{P}_V^{-1}\frac{\partial \boldsymbol{P}_V}{\partial \psi_j}\boldsymbol{P}_V^{-1}\frac{\partial \boldsymbol{P}_V}{\partial \psi_k}\right)$ etc.

Estimate of the asymptotic bias:

if $\sigma_0^2(>0)$ is supposed to be at an interior point,

$$\hat{b}_R(\hat{\boldsymbol{\theta}}_R) = 1 + B^{-1}\sum_{b=1}^{B}[(\check{\boldsymbol{\psi}}_R^{(b)} - \hat{\boldsymbol{\psi}}_R)^{\mathrm{T}}\{\mathcal{J}_R^{\psi\psi}(\hat{\boldsymbol{\theta}}_R)\}^{-1}(\tilde{\boldsymbol{\psi}}_R^{(b)} - \hat{\boldsymbol{\psi}}_R)],$$

where $\mathcal{J}_R^{\psi\psi}(\hat{\boldsymbol{\theta}}_R)$ is the elements of $\mathcal{J}_R(\hat{\boldsymbol{\theta}}_R)^{-1}$ corresponding to $\boldsymbol{\psi}$.

Example
Pig weights data

Values of estimated bias and informatin criteria in fitting LME models to pig weights data ($B = 10000$):

Estimation	ML			REML		
Fitted model	M_0	M_1	M_2	M_0	M_1	M_2
Num. pars	3	4	5	1	2	3
$\hat{b}(\hat{\boldsymbol{\theta}})$	(3)	3.980	5.036	(1)	2.003	2.992
$\mathrm{IC}_{\mathrm{PC}}$	2508.50	2037.81	1748.15	2494.91	2037.80	1747.01
AIC	2508.50	2037.85	1748.08	2494.91	2037.80	1747.03
CAIC	—	1914.91	1518.97	—	1915.03	1518.95
BIC	2520.71	2054.13	1768.42	2498.98	2045.93	1759.24

- The estimated bias $\hat{b}(\hat{\boldsymbol{\theta}})$ was close to the number of parameters.
- All of the IC selected M_2 (with random intercepts and slopes).

Simulation experiment
Objective and design

Objectives

- To evaluate bias correction term numerically
- To compare the performance of information criteria on selecting a true model

Design

- Models (true/fitted)

  $y_{ij} = \beta_0 + \beta_1 t_{ij} + b_{0j} + b_{1j} t_{ij} + e_{ij},\ i = 1, \ldots, n;\ j = 1, \ldots, m$

  - $t_{ij} = j\ (j = 1, \ldots, m)$
  - $b_{0j} \sim \mathrm{N}(0, \sigma^2 \phi_0)$, $b_{1j} \sim \mathrm{N}(0, \sigma^2 \phi_1)$ (independent)
  - $e_{ij} \sim \mathrm{N}(0, \sigma^2)$
- True values (to appear in this talk)
  - Common values: $(\beta_0, \beta_1) = (0, 1)$, $\sigma^2 = 1$
  - $\mathrm{M_{T0}}$: $(\psi_0, \psi_1) = (0, 0)$ (without random effect)
  - $\mathrm{M_{T1}}$: $(\psi_0, \psi_1) = (1/4, 0)$ (with random intercept)
  - $\mathrm{M_{T2}}$: $(\psi_0, \psi_1) = (1/4, 1/8)$ (with random slope and intercept)

Simulation experiment
Objective and design

Objectives

- To evaluate bias correction term numerically
- To compare the performance of information criteria on selecting a true model

Design

- Models (true/fitted)

  $y_{ij} = \beta_0 + \beta_1 t_{ij} + b_{0j} + b_{1j} t_{ij} + e_{ij},\ i = 1, \ldots, n;\ j = 1, \ldots, m$

  - $t_{ij} = j\ (j = 1, \ldots, m)$
  - $b_{0j} \sim \mathrm{N}(0, \sigma^2 \phi_0),\ b_{1j} \sim \mathrm{N}(0, \sigma^2 \phi_1)$ (independent)
  - $e_{ij} \sim \mathrm{N}(0, \sigma^2)$
- True values (to appear in this talk)
  - Common values: $(\beta_0, \beta_1) = (0, 1),\ \sigma^2 = 1$
  - $\mathrm{M_{T0}}$: $(\psi_0, \psi_1) = (0, 0)$ (without random effect)
  - $\mathrm{M_{T1}}$: $(\psi_0, \psi_1) = (1/4, 0)$ (with random intercept)
  - $\mathrm{M_{T2}}$: $(\psi_0, \psi_1) = (1/4, 1/8)$ (with random slope and intercept)

Simulation experiment
Methods

Methods (replicate 1000 times)

- Model to be fitted (by ML and REML estimation)

$$b_{0j} \sim \mathrm{N}(0, \sigma^2\phi_0),\ b_{1j} \sim \mathrm{N}(0, \sigma^2\phi_1) \text{ (independent)}$$
$$(j = 1, \ldots, m)$$

  - M_0: $(\psi_0, \psi_1) = (0, 0)$ (without random effect)
  - M_1: $\psi_0 = 0,\ \psi_1 \geq 0$ (with random intercept)
  - M_2: $\psi_0 \geq 0,\ \psi_1 \geq 0$ (with random slope and intercept)
- Information criteria to select a model
  - $\mathrm{IC}_{\mathrm{PC}}$ (Proposal)
  - AIC (Akaike, 1973)
  - CAIC (Vaida and Blanchard, 2005)
  - BIC (Schwartz, 1978)
- How to compare
  - The bias of IC as an estimate of $-2\times$ELL (except CAIC)
  - The proportion of samples for which each model was chosen

Simulation experiment
Results

The bias of IC as an estimate of $-2\times$ELL
True model: $\mathrm{M_{T0}}$ (without random effect), $n = 20$, $m = 10$

Estimation	ML			REML		
Fitted	$\mathrm{M_0}$	$\mathrm{M_1}$	$\mathrm{M_2}$	$\mathrm{M_0}$	$\mathrm{M_1}$	$\mathrm{M_2}$
$\mathrm{IC_{PC}}$	0.34	0.93	0.95	-0.25	0.19	0.14
AIC	0.34	1.76	3.51	-0.25	1.01	2.68
BIC	10.23	14.95	20.00	3.04	7.61	12.57

- $\mathrm{IC_{PC}}$ gave smaller bias than AIC and BIC.

Simulation experiment
Results

The bias of IC as an estimate of $-2\times$ELL
True model: $\mathrm{M_{T1}}$ (with random intercept), $n = 20$, $m = 10$

Estimation	ML			REML		
Fitted	$\mathrm{M_0}$	$\mathrm{M_1}$	$\mathrm{M_2}$	$\mathrm{M_0}$	$\mathrm{M_1}$	$\mathrm{M_2}$
$\mathrm{IC_{PC}}$	-3.76	-0.35	-0.35	-1.28	-1.16	-1.28
AIC	-3.76	-0.28	0.87	-1.28	-1.09	-0.07
BIC	6.14	12.91	17.36	2.02	5.51	9.82

- In ML, $\mathrm{IC_{PC}}$ gave smaller bias than AIC and BIC in fitting $\mathrm{M_1}$ and $\mathrm{M_2}$.
- The bias in fitting $\mathrm{M_0}$ is larger, but this would give little influence on model selection with IC because of poor fitting.

Simulation experiment

Results

The bias of IC as an estimate of $-2\times$ELL
True model: M_{T2} (with random slope), $n = 20$, $m = 10$

Estimation	ML			REML		
Fitted	M_0	M_1	M_2	M_0	M_1	M_2
$\mathrm{IC}_{\mathrm{PC}}$	-27.69	-13.09	-0.14	-11.19	-2.61	0.31
AIC	-27.69	-13.09	0.51	-11.19	-2.61	0.89
BIC	-17.80	0.11	17.01	-7.90	3.99	10.79

- $\mathrm{IC}_{\mathrm{PC}}$ gave smaller bias than AIC and BIC in fitting M_2.
- The bias in fitting M_0 and M_1 is larger, but this would give little influence on model selection with IC because of poor fitting.

Simulation experiment
Results

The proportion of samples for which each model was chosen
True model: $\mathrm{M_{T0}}$ (without random effect), $n = 20$, $m = 10$

Estimation	ML			REML		
Fitted model	$\mathrm{M_0}$	$\mathrm{M_1}$	$\mathrm{M_2}$	$\mathrm{M_0}$	$\mathrm{M_1}$	$\mathrm{M_2}$
Max. ELL	.464	.262	.274	.705	.135	.160
Min. $\mathrm{IC_{PC}}$	.905	.060	.035	.901	.059	.040
Min. AIC	.940	.052	.008	.933	.057	.010
Min. CAIC	.421	.169	.410	.000	.388	.612
Min. BIC	.995	.004	.001	.993	.007	.000

- $\mathrm{IC_{PC}}$ tended to choose slightly larger models than AIC and BIC (not so critical).
- CAIC tended to choose large models more remarkably.

Simulation experiment
Results

The proportion of samples for which each model was chosen
True model: $\mathrm{M_{T1}}$ (with random intercept), $n = 20$, $m = 10$

Estimation	ML			REML		
Fitted model	$\mathrm{M_0}$	$\mathrm{M_1}$	$\mathrm{M_2}$	$\mathrm{M_0}$	$\mathrm{M_1}$	$\mathrm{M_2}$
Max. ELL	.000	.653	.347	.000	.691	.309
Min. $\mathrm{IC_{PC}}$	.013	.871	.116	.017	.844	.139
Min. AIC	.022	.904	.074	.023	.899	.078
Min. CAIC	.000	.358	.642	.000	.340	.660
Min. BIC	.097	.898	.005	.096	.893	.011

- $\mathrm{IC_{PC}}$ gave smaller risk of choosing smaller models wrongly, although it tended to choose slightly larger models than AIC and BIC.
- CAIC tended to choose large models more remarkably.

Simulation experiment
Results

The proportion of samples for which each model was chosen
True model: M_{T2} (with random slope), $n = 10$, $m = 10$

Estimation	ML			REML		
Fitted model	M_0	M_1	M_2	M_0	M_1	M_2
Max. ELL	.000	.018	.982	.000	.013	.987
Min. $\mathrm{IC}_{\mathrm{PC}}$	.008	.111	.881	.005	.121	.874
Min. AIC	.012	.151	.837	.012	.166	.822
Min. CAIC	.001	.019	.980	.000	.018	.992
Min. BIC	.047	.274	.679	.043	.266	.691

- $\mathrm{IC}_{\mathrm{PC}}$ gave smaller risk of choosing smaller models wrongly, although it tended to choose slightly larger models than AIC and BIC.
- CAIC chose M_2 for almost all samples.

Concluding Remarks

Summary

Proposal

- We evaluated an asymptotic bias of MLL as an estimate of ELL under parameter constraints, which would be applicable to any parameter constrained models.
- We proposed a new information criterion (IC) based on Monte Carlo approximation (bootstrap evaluation) of the bias.
  - We use multivariate normal random samples to obtain the IC.
  - Simpler computation than some recently proposed IC.
  - Faster than ordinary bootstrapping method.

Result of the simulation experiment (only random effect selection)

- The proposed IC gives a consistent estimate of $-2\times$ ELL, and numerically less bias than the ordinary AIC and BIC.
- It tends to select slightly larger model than other AIC and BIC. However, it gives smaller risk of choosing smaller models wrongly than AIC and BIC.

Concluding Remarks

Summary

Proposal

- We evaluated an asymptotic bias of MLL as an estimate of ELL under parameter constraints, which would be applicable to any parameter constrained models.
- We proposed a new information criterion (IC) based on Monte Carlo approximation (bootstrap evaluation) of the bias.
  - We use multivariate normal random samples to obtain the IC.
  - Simpler computation than some recently proposed IC.
  - Faster than ordinary bootstrapping method.

Result of the simulation experiment (only random effect selection)

- The proposed IC gives a consistent estimate of $-2\times$ ELL, and numerically less bias than the ordinary AIC and BIC.
- It tends to select slightly larger model than other AIC and BIC. However, it gives smaller risk of choosing smaller models wrongly than AIC and BIC.

Concluding Remarks

Future works

- More comprehensive simulation
  - To select both fixed and random effect terms
- Comparison between ML and REML

Thank you for your attention!

sakamoto@sigmath.es.osaka-u.ac.jp

References

- Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. In: *The 2nd Internaltional Symposium on Information Theory*, B. N. Petrov and F. Csaki (eds.), pp. 267–281. Budapest: Akademiai Kiado.
- Crainiceanu, C. M. and Ruppert, D. (2004). Likelihood ratio tests in linear mixed models with one variance component. *J. Roy. Statist. Soc. Ser. B*, **66**, 165–185.
- Diggle, P.J., Heagerty, P., Liang, K.-Y. and Zeger, S.L. (2002). *Analysis of Longitudinal Data, Second Edition*. Oxford University Press.
- Dimova, R. B., Markatou, M. and Talal, A. H. (2011). Information methods for model selection in linear mixed effect models with application to HCV data. *Comput. Statist. Data Anal.*, **55**, 2677–2697.
- Greven, S. and Kneib, T. (2009). On the behaviour of marginal and conditional Akaike information criteria in linear mixed models. Johns Hopkins University, Dept. of Biostatistics Working Papers, No. 202.
- Laird, N. M. and Ware, J. H. (1982). Random-effects models for longitudinal data. *Biometrics*, **38**, 963–974.

References

- Liang, H., Wu, H. and Zou, G. (2008). A note on conditional AIC for linear mixed-effects models. *Biometrika*, **95**, 773–778.
- Pu, W. and Niu, X.-F. (2006). Selecting mixed-effects models based on a generalized information criterion. *J. Mult. Anal.*, **97**, 733–758.
- Ruppert, D., Wand, M. P. and Carroll, R. J. (2003). *Semiparametric Regression*. Cambridge University Press.
- Schwarz, G. E. (1978). Estimating the dimension of a model. *Annals of Statistics*, **6**, 461–464.
- Self, S. G. and Liang, K. Y. (1987). Asymptotic properties of maximum likelihood estimators and likelihood ratio tests under nonstandard conditions. *J. Amer. Statist. Assoc.*, **82**, 605–610.
- Shang, J. and Cavanaugh, J. E. (2008). Bootstrap variants of the Akaike information criterion for mixed model selection. *Comput. Statist. Data Anal.*, **52**, 2004–2021.
- Stram, D. O. and Lee, J. W. (1994). Variance components testing in the longitudinal mixed effects model. *Biometrics*, **50**, 1171–1177.
- Vaida, F. and Blanchard, S. (2005). Conditional Akaike information for mixed-effects models. *Biometrika*, **92**, 351–370.