

Statistica Sinica Preprint No: SS-2026-0168	
Title	Generalized Aic for Componentwise Model Selection
Manuscript ID	SS-2026-0168
URL	http://www.stat.sinica.edu.tw/statistica/
DOI	10.5705/ss.202026.0168
Complete List of Authors	Youngjo Lee and Kwok Pui Choi
Corresponding Authors	Kwok Pui Choi
E-mails	stackp@nus.edu.sg
Notice: Accepted author version.	

Generalized AIC for componentwise model selection

Youngjo Lee and Kwok Pui Choi

Seoul National University and National University of Singapore

Abstract: A statistical model can be decomposed into a number of interconnected component models. Model selection can be carried out componentwise to meet each component's inferential objectives. We propose a generalized Akaike information criterion that provides a unified framework for optimal parameter estimation for components of interest while ensuring consistent selection of the remaining components. This leads to an extension of the definition of degrees of freedom to random parameters. We illustrate model selection using various information criteria and elucidate correspondences between Bayesian and likelihood approaches.

Key words and phrases: AIC; BIC; CAIC; DIC; Lasso; Degrees of freedom

1. Introduction

The Akaike information criterion (AIC; Akaike, 1973) seeks to minimize the Kullback–Leibler divergence between the true model and a candidate model, thereby achieving minimax-rate optimality for parameter estimation. In contrast, the Bayesian information criterion (BIC; Schwarz, 1978) is consistent for selecting the true model. Yang (2005) showed that any consistent model selection procedure cannot simultaneously be minimax-rate optimal for parameter estimation. He noted that “traditionally, the goal of statistical inference is often not considered until after the ‘optimal’ model is found,” and further recommended against identifying the true model when the primary objective is to estimate the parameters of interest. Motivated by these results, we develop a class of likelihood-based information

criteria designed to meet these seemingly conflicting inferential objectives.

Lee and Nelder (2001) demonstrated that a statistical model can be decomposed into interconnected component models. Importantly, inferential objectives may differ across components: some are of primary interest for estimation or prediction, while others are subject to consistent model selection. Motivated by this distinction, we combine the strengths of AIC for optimal estimation of components of interest with those of BIC for consistent selection of the remaining components. As will be shown, this can be achieved by applying a generalized AIC (GAIC) to an appropriate likelihood. The entire model can then be chosen by selecting appropriate component models.

In this paper, we propose a likelihood-based GAIC that applies either AIC or BIC to each component according to its inferential objective, providing both frequentist and Bayesian interpretations. The proposed criterion unifies information criteria such as AIC and BIC and offers a coherent interpretation of information criteria derived from marginal (Akaike, 1973), restricted (Patterson and Thompson, 1971), conditional (Vaida and Blanchard, 2005), and other likelihoods including the integrated and model likelihoods considered here. We investigate the properties of these likelihood-based criteria for componentwise model selection. For illustration, we consider generalized linear models (GLMs; McCullagh and Nelder, 1989), generalized linear mixed models (GLMMs; Breslow and Clayton, 1993), hierarchical generalized linear models (HGLMs; Lee and Nelder, 1996), and robust HGLMs (Lee and Nelder, 2006a; Noh and Lee, 2007). GLMMs are a special case of HGLMs with normal random effects, while robust HGLMs allow heavy-tailed random effects. GLMs consists of only fixed parameters, whereas the others include random parameters. This investigation also motivates a new information criterion for subject-specific prediction and clarifies how likelihood

bridges Bayesian and frequentist approaches to model selection. In Section 5, we show the correspondence between deviance information criterion (DIC) of Spiegelhalter et al. (2002) and GAIC. GAICs lead naturally to a definition of the degrees of freedom associated with random components. We also discuss lasso-type estimation in GAICs.

2. Random effects models and their associated likelihoods

Statistical models may involve both fixed and random parameters. Classical frequentist likelihood models typically include only fixed parameters, whereas Bayesian models treat all unknown parameters as random. As a result, model selection procedures under the two paradigms are fundamentally different. To develop a unified framework for model selection, we consider random-effects models that contain both fixed and random parameters. For ease of exposition, we focus on the linear mixed model (LMM), whose likelihoods admit explicit forms. Nevertheless, the results of this paper extend naturally to a wider class of models, including log-normal, Poisson, and gamma random-effects models, as discussed in Section 5. Consider the LMM

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{v} + \mathbf{e},$$

where $\mathbf{X}$ and $\mathbf{Z}$ are $n \times p$ and $n \times q$ design matrices, respectively, $\mathbf{e} \sim N(0, \phi\mathbf{I}_n)$ and $\mathbf{v} \sim N(0, \boldsymbol{\Lambda})$. Let $\boldsymbol{\lambda}$ be a t -vector of parameters in $\boldsymbol{\Lambda}$. The linear predictor, $\boldsymbol{\eta}$, of LMM is

$$\boldsymbol{\eta} = \boldsymbol{\mu} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{v}. \quad (2.1)$$

The first two conditional moments of LMM are denoted by

$$\boldsymbol{\mu} = E(\mathbf{y}|\mathbf{v}) = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{v} \quad \text{and} \quad \text{var}(\mathbf{y}|\mathbf{v}) = \phi\mathbf{I}_n,$$

2.1 Preliminaries and notation

which allow subject-specific prediction of $\boldsymbol{\mu}$. This leads to the multivariate model with the first two marginal moments denoted by

$$\boldsymbol{\mu}_m = E(\mathbf{y}) = \mathbf{X}\boldsymbol{\beta} \quad \text{and} \quad \text{var}_m(\mathbf{y}) = \boldsymbol{\Sigma} = \mathbf{Z}\boldsymbol{\Lambda}\mathbf{Z}^\top + \phi\mathbf{I}_n,$$

which allow estimation of $\boldsymbol{\mu}_m$ for population. Random effects models have four components $\boldsymbol{\beta}, \mathbf{v}, \phi, \boldsymbol{\Lambda}$, and each component can be modeled as a GLM (Lee and Nelder, 2001). In this paper, we study how to select component GLMs to meet inferential objectives.

Model (2.1) is sufficiently general to accommodate a general multi-component random effects model: Suppose that the distribution of $\mathbf{y}|\mathbf{v}$ follows a GLM with linear predictor: $\eta = m(\boldsymbol{\mu}) = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}_1\mathbf{v}_1 + \cdots + \mathbf{Z}_M\mathbf{v}_M$, where $m(\cdot)$ is the link function, $\mathbf{X}$, and $\mathbf{Z}_i, 1 \leq i \leq M$ are $n \times p$, and $n \times q_i (1 \leq i \leq M)$ design matrices, and $\mathbf{v}_i \sim N(0, \boldsymbol{\Lambda}_i)$ for $i = 1, \dots, M$. By taking $\mathbf{Z} = [\mathbf{Z}_1 \cdots \mathbf{Z}_M]$, $\mathbf{v} = (\mathbf{v}_1^\top, \dots, \mathbf{v}_M^\top)^\top$ and $\boldsymbol{\Lambda} = \text{diag}(\boldsymbol{\Lambda}_1, \dots, \boldsymbol{\Lambda}_M)$, this multi-component model can be reduced to Model (2.1) with $q = q_1 + \cdots + q_M$. Thus, we use Model (2.1) and assume that $\mathbf{X}$ has full column rank.

2.1 Preliminaries and notation

Let $\boldsymbol{\theta} = (\phi, \boldsymbol{\Lambda}, \boldsymbol{\beta})$ be the fixed parameters, $\mathbf{v}$ the random parameters, and $\boldsymbol{\omega} = (\boldsymbol{\theta}, \mathbf{v})$ the entire parameters. We denote the model, marginal, restricted, conditional, integrated, extended, and h-likelihoods by $L_0, L_m, L_R, L_c, L_i, L_e$, and L_n , respectively; whose definitions are given upon first use. Some of these likelihoods can be viewed as adjusted profile likelihoods or penalized likelihoods. The corresponding log-likelihoods are denoted by $\ell_0, \ell_m, \ell_R, \ell_c, \ell_i, \ell_e$, and ℓ_n . Let $\hat{\boldsymbol{\theta}}_{\text{MLE}}$ denote the maximum likelihood estimator (MLE) maximizing ℓ_m ; $(\hat{\boldsymbol{\theta}}_{\text{MHLE}}, \hat{\mathbf{v}}_{\text{MHLE}})$ the maximum h-likelihood (MHLE) maximizing ℓ_n ; and $\hat{\boldsymbol{\psi}}_{\text{REMLE}}$ the restricted MLE (REMHLE) maximizing ℓ_R where $\boldsymbol{\psi} = (\phi, \boldsymbol{\Lambda})$. Let $\hat{\boldsymbol{\varsigma}}(\boldsymbol{\psi})$ denote an esti-

2.2 Integrated, extended, h-, and adjusted profile likelihoods

mator of ς for a given ψ . For example, $\widehat{\varsigma}_{\text{MLE}}(\psi)$ is the MLE of ς for a given ψ .

All vectors are column vectors, denoted by boldface lower case letters, while matrices are denoted by boldface upper case letters. The determinant and trace of a square matrix, $\mathbf{A}$, are denoted by $|\mathbf{A}|$ and $\text{tr}(\mathbf{A})$, respectively. The norm of a vector $\mathbf{v}$ with respect to an invertible matrix $\mathbf{A}$ is defined as $\|\mathbf{v}\|_{\mathbf{A}} = (\mathbf{v}^\top \mathbf{A}^{-1} \mathbf{v})^{1/2}$. If $\mathbf{A}$ is an identity matrix, the norm is simply denoted as $\|\mathbf{v}\|$. All proofs are provided in the Supplementary Material.

2.2 Integrated, extended, h-, and adjusted profile likelihoods

We introduce the likelihood framework for estimation and model selection. Throughout this paper, ψ denotes the parameters of interest for estimation or prediction, ς the remaining parameters for consistent model selection. We define ς as null when the primary objective is to estimate all the parameters. Similarly, ψ is said to be null when the inferential aim focuses on consistent model selection. To focus on estimating ψ , we consider the integrated likelihood, $L_i(\psi; \mathbf{y})$, obtained by integrating out ς ,

$$L_i(\psi; \mathbf{y}) = \int L_e(\psi, \varsigma; \mathbf{y}) d\varsigma.$$

Here, $L_e(\psi, \varsigma; \mathbf{y}) = L_e(\beta, \phi, \lambda, \mathbf{v}; \mathbf{y}) = f(\mathbf{y}, \mathbf{v} | \theta)$ denotes the extended likelihood as defined in Bjørnstad (1996). In this paper, we use of integrated likelihood for consistent model selection of components ς and estimation of components ψ .

According to the likelihood principle (Birnbbaum, 1962), the marginal likelihood, which is also an integrated likelihood,

$$L_m(\theta; \mathbf{y}) = f(\mathbf{y} | \theta) = \int f(\mathbf{y}, \mathbf{v} | \theta) d\mathbf{v},$$

captures all the information in the data regarding the fixed parameters θ . Bjørnstad (1996) generalized this likelihood principle to the extended likelihood, asserting that it captures

2.2 Integrated, extended, h-, and adjusted profile likelihoods

all the information in the data regarding the entire set of fixed and random parameters $\omega = (\boldsymbol{\theta}, \mathbf{v})$. In random effects models, the integrated likelihood is often intractable. Thus, in formulating the information criteria we use the adjusted profile extended likelihood L_{ea} , which is a Laplace approximation of the integrated likelihood

$$L_{\text{ea}}(\boldsymbol{\psi}; \mathbf{y}) = L_{\text{e}}(\hat{\boldsymbol{\varsigma}}_{\text{MHLE}}(\boldsymbol{\psi}), \boldsymbol{\psi}; \mathbf{y}) m(\boldsymbol{\psi}, \mathbf{y}) = L_{\text{I}}(\boldsymbol{\psi}; \mathbf{y}) \{1 + O_p(n^{-1/2})\}, \quad (2.2)$$

where $m(\boldsymbol{\psi}, \mathbf{y}) = \left| \{-\partial^2 \ell_{\text{e}}(\boldsymbol{\varsigma}, \boldsymbol{\psi}; \mathbf{y}) / \partial \boldsymbol{\varsigma} \partial \boldsymbol{\varsigma}^\top\} / (2\pi) \right|_{\boldsymbol{\varsigma}=\hat{\boldsymbol{\varsigma}}_{\text{MHLE}}(\boldsymbol{\psi})}^{-1/2}$. This gives asymptotically equivalent information criteria based on integrated likelihoods. Lee and Nelder (2001) proposed using the adjusted profile extended likelihood L_{ea} to compute the marginal likelihood and restricted likelihood.

In LMM, Henderson et al. (1959) proposed using joint density $f(\mathbf{y}, \mathbf{v} | \boldsymbol{\theta})$ to obtain the best linear unbiased predictor of $\mathbf{v}$. To extend the ML procedures to random effects models, Lee and Nelder (1996) introduced h-likelihood, which is an extended likelihood. However, unlike MLEs, maximum extended likelihood estimators are generally not asymptotically optimal; see the discussion in Meng (2009). To address this, Lee and Lee (2025) defined a new h-likelihood $L_{\text{n}}(\omega; \mathbf{y})$ as an adjusted extended likelihood,

$$L_{\text{n}}(\omega; \mathbf{y}) = L_{\text{e}}(\omega; \mathbf{y}) a(\boldsymbol{\theta}, \mathbf{y}), \quad (2.3)$$

where $a(\boldsymbol{\theta}, \mathbf{y})$ is an adjustment term. In random effects models, they showed that the MHLE $(\hat{\boldsymbol{\theta}}_{\text{MHLE}}, \hat{\mathbf{v}}_{\text{MHLE}})$ of the new h-likelihood (2.3) satisfies $\hat{\boldsymbol{\theta}}_{\text{MHLE}} = \hat{\boldsymbol{\theta}}_{\text{MLE}}$, and that $\hat{\mathbf{v}}_{\text{MHLE}}$ is asymptotically the best unbiased predictor of $\mathbf{v}$. They also showed how to obtain MHLE and their variance estimators in general models. Specifically, in LMM, $a(\boldsymbol{\theta}, \mathbf{y}) = |\boldsymbol{\Lambda}^{-1} + \mathbf{Z}^\top \mathbf{Z} / \phi|^{-1/2}$, and the random effects estimator $\hat{\mathbf{v}}_{\text{MHLE}}(\boldsymbol{\theta}) = \text{E}(\mathbf{v} | \mathbf{y})|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}_{\text{MLE}}}$ is the best linear unbiased predictor of Henderson et al. (1959). In this paper, we adopt the MHLE

2.2 Integrated, extended, h-, and adjusted profile likelihoods

procedure for parameter estimation and variance estimation. For random effects model, Lee and Lee (2025) showed that

$$L_m(\boldsymbol{\theta}; \mathbf{y}) = \int L_e(\boldsymbol{\omega}; \mathbf{y}) d\mathbf{v} = \int L_e(\boldsymbol{\theta}, \mathbf{v}; \mathbf{y}) d\mathbf{v} = L_n(\boldsymbol{\theta}, \hat{\mathbf{v}}_{\text{MHLE}}(\boldsymbol{\theta}); \mathbf{y}). \quad (2.4)$$

This shows that the marginal likelihood can be viewed either as an integrated likelihood obtained by integrating out the random effects $\mathbf{v}$ from the extended likelihood (2.2) or as a profile h-likelihood obtained by profiling out the random effects $\mathbf{v}$ from the h-likelihood (2.3). Thus, $\hat{\boldsymbol{\theta}}_{\text{MHLE}} = \hat{\boldsymbol{\theta}}_{\text{MLE}}$ and $\ell_n(\hat{\boldsymbol{\theta}}_{\text{MHLE}}, \hat{\mathbf{v}}_{\text{MHLE}}; \mathbf{y}) = \ell_m(\hat{\boldsymbol{\theta}}_{\text{MLE}}; \mathbf{y})$.

Adjusted profile likelihoods have been used in higher-order asymptotic theories. Specifically, in marginal models where $\boldsymbol{\psi}$ and $\boldsymbol{\varsigma}$ are parameter orthogonal, Cox and Reid (1987) applied the adjusted profile marginal likelihood (2.5) below to reduce the bias of the estimator $\hat{\boldsymbol{\psi}}_{\text{MLE}}$ in the presence of the remaining parameters $\boldsymbol{\varsigma}$,

$$L_{\text{ma}}(\boldsymbol{\psi}; \mathbf{y}) = L_m(\hat{\boldsymbol{\varsigma}}_{\text{MLE}}(\boldsymbol{\psi}), \boldsymbol{\psi}; \mathbf{y})m(\boldsymbol{\psi}, \mathbf{y}), \quad (2.5)$$

where $m(\boldsymbol{\psi}, \mathbf{y}) = |\{-\partial^2 \ell_m(\boldsymbol{\varsigma}, \boldsymbol{\psi}; \mathbf{y}) / \partial \boldsymbol{\varsigma} \partial \boldsymbol{\varsigma}^\top\} / (2\pi)|_{\boldsymbol{\varsigma}=\hat{\boldsymbol{\varsigma}}_{\text{MLE}}(\boldsymbol{\psi})}^{-1/2}$. In fact, $L_{\text{ma}}(\boldsymbol{\psi}; \mathbf{y})$ is also the Laplace approximation to the integrated likelihood $\int L_m(\boldsymbol{\psi}, \boldsymbol{\varsigma}; \mathbf{y}) d\boldsymbol{\varsigma}$. They showed that $\hat{\boldsymbol{\varsigma}}_{\text{MLE}}(\hat{\boldsymbol{\psi}}_{\text{MLE}}) - \hat{\boldsymbol{\varsigma}}_{\text{MLE}}(\boldsymbol{\psi}) = O_p(1/n)$ under parameter orthogonality. Care must be exercised when applying equations (2.2) and (2.5), as partitioning of parameters depends on the full model under consideration. Suppose that the parameters of interest $\boldsymbol{\psi} = (\boldsymbol{\phi}, \boldsymbol{\lambda})$ in both marginal model and random effects model as the full model, the definition of the remaining parameters, $\boldsymbol{\varsigma}$, differs between the two: $\boldsymbol{\varsigma} = \boldsymbol{\beta}$ in marginal models, whereas $\boldsymbol{\varsigma} = (\boldsymbol{\beta}, \mathbf{v})$ in random effects models.

2.3 Likelihoods for linear mixed models

In LMM, the extended likelihood is

$$L_e(\boldsymbol{\omega}; \mathbf{y}) = (2\pi\phi)^{-n/2} |2\pi\boldsymbol{\Lambda}|^{-1/2} \exp \left\{ -\|\mathbf{y} - \mathbf{X}\boldsymbol{\beta} - \mathbf{Z}\mathbf{v}\|^2 / (2\phi) - \|\mathbf{v}\|_{\boldsymbol{\Lambda}}^2 / 2 \right\},$$

and the marginal likelihood is

$$L_m(\boldsymbol{\theta}; \mathbf{y}) = |2\pi\boldsymbol{\Sigma}|^{-1/2} \exp \left\{ -\|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|_{\boldsymbol{\Sigma}}^2 / 2 \right\}. \quad (2.6)$$

With $e_i = y_i - \mu_i$,

$$\begin{aligned} L_c(\boldsymbol{\beta}, \phi, \mathbf{v}; \mathbf{y}) &= f(\mathbf{y}|\boldsymbol{\omega}) = (2\pi\phi)^{-n/2} \exp \left\{ -\|\mathbf{y} - \mathbf{X}\boldsymbol{\beta} - \mathbf{Z}\mathbf{v}\|^2 / (2\phi) \right\} \\ &= (2\pi\phi)^{-n/2} \exp \left\{ -\sum e_i^2 / (2\phi) \right\} \end{aligned}$$

is called the conditional likelihood of the linear model with fixed effects $\mathbf{v}$, equivalently written as $L_c(\boldsymbol{\mu}, \phi; \mathbf{y})$ since $\boldsymbol{\mu} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{v}$ with fixed $(\boldsymbol{\beta}, \mathbf{v})$.

Inconsistency of MLEs when the number of the remaining parameters increases with the sample size is well known as the Neyman-Scott problem (Neyman and Scott, 1948). In LMM, Patterson and Thompson (1971) introduced the REMLE for estimating the variance components $\boldsymbol{\psi} = (\phi, \boldsymbol{\Lambda})$ when the dimension of $\boldsymbol{\beta}$ increases with the sample size. The restricted likelihood is the adjusted profile marginal likelihood

$$\ell_R(\boldsymbol{\psi}; \mathbf{y}) = \ell_{ma}(\boldsymbol{\psi}; \mathbf{y}) = -\left\{ \log |2\pi\boldsymbol{\Sigma}| + \log |\mathbf{X}^\top \boldsymbol{\Sigma}^{-1} \mathbf{X} / (2\pi)| + \mathbf{y}^\top \mathbf{P} \mathbf{y} \right\} / 2,$$

where $m(\boldsymbol{\psi}, \mathbf{y}) = |\mathbf{X}^\top \boldsymbol{\Sigma}^{-1} \mathbf{X} / (2\pi)|^{-1/2}$ in (2.5), and

$\mathbf{P} = \boldsymbol{\Sigma}^{-1} - \boldsymbol{\Sigma}^{-1} \mathbf{X} (\mathbf{X}^\top \boldsymbol{\Sigma}^{-1} \mathbf{X})^{-1} \mathbf{X}^\top \boldsymbol{\Sigma}^{-1}$. Note that $\mathbf{y}^\top \mathbf{P} \mathbf{y} = \|\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}_{MLE}(\boldsymbol{\psi})\|_{\boldsymbol{\Sigma}}^2$, and $\hat{\boldsymbol{\beta}}_{MLE}(\boldsymbol{\psi}) = (\mathbf{X}^\top \boldsymbol{\Sigma}^{-1} \mathbf{X})^{-1} \mathbf{X}^\top \boldsymbol{\Sigma}^{-1} \mathbf{y}$ is the MLE of $\boldsymbol{\beta}$ for a given $\boldsymbol{\psi}$.

3. GAIC for component model selection

3.1 GAIC for marginal model selection

Lee and Nelder (2001) showed that fitting a random effects model is reduced to fitting a four-component GLM: one for each of the $\beta, \mathbf{v}, \phi$, and λ components. This procedure can be further extended to fit a robust random effects model by allowing random effects for ϕ and λ (Lee et al., 2017). Thus, one can view component models as lego blocks for building and fitting a complex model. Suppose that ψ is the component of interest for estimation, while ς denotes the remaining components for consistent model selection. Such inferential aims should guide the overall model selection process through componentwise selection. We shall show that if the aim is to estimate all the fixed parameters $\psi = \theta$, then AIC is used; if the aim is consistent model selection of all the fixed parameters $\varsigma = \theta$, then BIC is used. In this paper, we propose synthesizing a final model that satisfies these inferential aims by applying either AIC or BIC suitably to each component model.

Let $\mathbf{y} = (y_1, \dots, y_n)^\top$ denote the observed data, and $\mathbf{y}^*$ denote future data generated by the true but unknown density $g(\mathbf{y}^*)$. We assume $\mathbf{y}$ and $\mathbf{y}^*$ are independent. Suppose that the inferential objective of model selection is to estimate only the fixed parameters in a class of marginal models, $M_k: L_*(\psi_k; \mathbf{y}) = f(\mathbf{y}|\psi_k)$, $k = 1, \dots, K$, where ψ_k is a subset of θ_k . Here, L_* denotes a generic likelihood introduced in Section 2.1. For example, $L_* = L_m$ for AIC. Let $\hat{\psi}_{\text{MLE},k} = \hat{\psi}_{\text{MLE},k}(\mathbf{y})$ be the MLE under the model M_k . The Kullback–Leibler information is

$$\begin{aligned} I(g(\cdot), f(\cdot|\hat{\psi}_{\text{MLE},k})) &= E_{\mathbf{y}^*}[\log\{g(\mathbf{y}^*)/f(\mathbf{y}^*|\hat{\psi}_{\text{MLE},k})\}] \\ &= E_{\mathbf{y}^*}\{\log g(\mathbf{y}^*)\} - E_{\mathbf{y}^*}\{\log f(\mathbf{y}^*|\hat{\psi}_{\text{MLE},k})\}, \end{aligned}$$

3.1 GAIC for marginal model selection

where $\int g(\mathbf{y}^*) d\mathbf{y}^* = 1$ and $E_{\mathbf{y}^*} \{\log f(\mathbf{y}^* | \hat{\boldsymbol{\psi}}_{\text{MLE},k})\} = \int \log f(\mathbf{y}^* | \hat{\boldsymbol{\psi}}_{\text{MLE},k}) g(\mathbf{y}^*) d\mathbf{y}^*$. Thus, minimizing the Kullback–Leibler information is equivalent to maximizing the expected log-likelihood, $E_{\mathbf{y}^*} \{\log f(\mathbf{y}^* | \hat{\boldsymbol{\psi}}_{\text{MLE},k})\}$:

$$E_{\mathbf{y}^*} \{\ell_*(\hat{\boldsymbol{\psi}}_{\text{MLE},k}; \mathbf{y}^*)\} = \int \ell_*(\hat{\boldsymbol{\psi}}_{\text{MLE},k}; \mathbf{y}^*) g(\mathbf{y}^*) d\mathbf{y}^*,$$

where $L_*(\boldsymbol{\psi}_k; \mathbf{y}^*) = c f(\mathbf{y}^* | \boldsymbol{\psi}_k)$ with $c = \int L_*(\boldsymbol{\psi}_k; \mathbf{y}^*) d\mathbf{y}^*$. This extends the original definition of the Akaike information to the generalized Akaike information for marginal model when $\boldsymbol{\psi}$ is a subset of $\boldsymbol{\theta}$:

$$\text{GAI} = -2E_{\mathbf{y}} E_{\mathbf{y}^*} \ell_*(\hat{\boldsymbol{\psi}}_{\text{MLE}}; \mathbf{y}^*) = -2 \iint \ell_*(\hat{\boldsymbol{\psi}}_{\text{MLE}}; \mathbf{y}^*) g(\mathbf{y}^*) d\mathbf{y}^* g(\mathbf{y}) d\mathbf{y}. \quad (3.7)$$

We shall show that AIC and BIC can be extended to random effects model by replacing $\hat{\boldsymbol{\psi}}_{\text{MLE}}$ by $\hat{\boldsymbol{\psi}}_{\text{MHLE}}$.

When $\boldsymbol{\psi}$ contains random parameters $\mathbf{v}$ with $\ell_* = \ell_c$, Vaida and Blanchard (2005) defined the conditional Akaike information

$$\begin{aligned} \text{CAI} &= -2E_{(\mathbf{y}, \mathbf{u})} E_{\mathbf{y}^* | \mathbf{u}} \ell_c(\hat{\boldsymbol{\psi}}_{\text{MHLE}}; \mathbf{y}^*) \\ &= -2 \iint \left\{ \int \ell_c(\hat{\boldsymbol{\psi}}_{\text{MHLE}}; \mathbf{y}^*) g(\mathbf{y}^* | \mathbf{u}) d\mathbf{y}^* \right\} g(\mathbf{y}, \mathbf{u}) d\mathbf{y} d\mathbf{u}, \end{aligned}$$

where $\mathbf{v}$ and $\mathbf{u}$ denote possibly different random effects and $g(\mathbf{y}, \mathbf{u}) = g(\mathbf{y} | \mathbf{u}) g(\mathbf{u})$ is the joint density of $(\mathbf{y}, \mathbf{u})$. In their derivation, the true density of $\mathbf{u}$, $g(\mathbf{u})$, may be different from the density of random effects $f(\mathbf{u} | \boldsymbol{\lambda})$ in the extended likelihood of the candidate model. Therefore, an advantage of their approach is that model selection is independent of the true density of $\mathbf{u}$. In this paper, we define the generalized Akaike information when $\boldsymbol{\psi}$ contains random parameters:

$$\text{GAI} = -2E_{\mathbf{y} | \mathbf{u}} E_{\mathbf{y}^* | \mathbf{u}} \ell_*(\hat{\boldsymbol{\psi}}_{\text{MHLE}}; \mathbf{y}^*) = -2 \iint \ell_*(\hat{\boldsymbol{\psi}}_{\text{MHLE}}; \mathbf{y}^*) g(\mathbf{y}^* | \mathbf{u}) g(\mathbf{y} | \mathbf{u}) d\mathbf{y}^* d\mathbf{y}.$$

3.1 GAIC for marginal model selection

This definition is more aligned with the Akaike information (3.7). Furthermore, the same advantage remains because $g(\mathbf{u})$ can differ from $f(\mathbf{u}|\boldsymbol{\lambda})$ in the candidate model.

When the true model is contained in the class of candidate models, let $\boldsymbol{\psi}_0$ be the true values of the fixed parameters and the realized values of the random parameters. Otherwise, $\boldsymbol{\psi}_0$ denotes the pseudo-true values and the pseudo-realized values. We propose the GAIC based on various likelihoods

$$\text{GAIC} = -2\ell_*(\hat{\boldsymbol{\psi}}_{\text{MHLE}}; \mathbf{y}) + 2 \text{ df}, \quad (3.8)$$

where, throughout the paper, the degrees of freedom for components $\boldsymbol{\psi}$ are

$$\text{df} = \dim(\boldsymbol{\psi}) = \text{tr}\{E[n^{-1}\{-\partial^2\ell_*(\boldsymbol{\psi}; \mathbf{y})/\partial\boldsymbol{\psi}\partial\boldsymbol{\psi}^\top\}|_{\boldsymbol{\psi}=\boldsymbol{\psi}_0}] \text{var}(\sqrt{n}(\hat{\boldsymbol{\psi}}_{\text{MHLE}} - \boldsymbol{\psi}_0))\}. \quad (3.9)$$

This leads to the definition of the degrees of freedom associated with the random effects. In this paper, we estimate df by $\text{tr}\{E[\mathbf{B}_1\mathbf{B}_2]\}$ where

$$\mathbf{B}_1 = -\frac{1}{n} \frac{\partial^2\ell_*(\boldsymbol{\psi}; \mathbf{y})}{\partial\boldsymbol{\psi}\partial\boldsymbol{\psi}^\top} \Big|_{\boldsymbol{\psi}=\hat{\boldsymbol{\psi}}_{\text{MHLE}}} \quad \text{and} \quad \mathbf{B}_2 = n \text{var}(\hat{\boldsymbol{\psi}}_{\text{MHLE}} - \boldsymbol{\psi}_0). \quad (3.10)$$

It is known that regularity conditions (R1)-(R6) below; and identifiability condition (i.e., $\boldsymbol{\theta}_1 \neq \boldsymbol{\theta}_2$ implies the distributions of $\mathbf{y}$ under $\boldsymbol{\theta}_1$ and $\boldsymbol{\theta}_2$ are different) are sufficient for the asymptotic optimality of MLEs to hold. Lee and Nelder (1996) proposed to impose constraints on random parameters for identifiability condition. With this identifiability condition, and the same regularity conditions (R1)-(R6), Lee and Lee (2025) shows that asymptotic optimality of MHLEs for fixed and for random parameters will hold.

(R1) The true parameter is an interior point of the parameter space.

(R2) The support $\{\mathbf{y} : f(\mathbf{y}|\boldsymbol{\theta}) > 0\}$ is independent of $\boldsymbol{\theta}$.

(R3) The log-likelihood function is continuously differentiable up to the third order with respect to $\boldsymbol{\theta}$ in an open neighborhood of $\boldsymbol{\theta}$.

3.1 GAIC for marginal model selection

(R4) Integration and differentiation can be interchanged up to second order.

(R5) In a neighborhood of the true parameter, the third derivatives are bounded by an integrable function almost surely with respect to $\mathbf{y}$.

(R6) The Fisher information matrix for an observation is finite and positive definite.

Theorem 1. *Under regularity conditions (R1)–(R6) for the h -likelihood, if the true model is contained in the class of candidate models, we have*

$$E[\text{GAIC}] = \text{GAI} + o(1). \quad (3.11)$$

When $\boldsymbol{\psi}$ is a subset of $\boldsymbol{\theta}$, $E = E_{\mathbf{y}}$ and $\text{var} = \text{var}_{\mathbf{y}}$ in (3.9); and when $\boldsymbol{\psi}$ contains $\mathbf{v}$, $E = E_{\mathbf{y}|\mathbf{u}}$ and $\text{var} = \text{var}_{\mathbf{y}|\mathbf{u}}$ in (3.9).

Theorem 1 shows that likelihood-based model selection GAIC satisfies the frequentist property based on future observations $\mathbf{y}^*$.

Donohue et al. (2011) introduced the conditional AIC (CAIC), which is a GAIC with $\boldsymbol{\psi} = (\boldsymbol{\beta}, \mathbf{v})$. In the random-effects models, their estimator is asymptotically equivalent to the MHLE in Lee and Lee (2025). Their regularity conditions are for consistency, asymptotic normality of MHLE, and for Bartlett identities of their variance estimation to hold. It is well known that these three properties hold for MLE under regularity conditions for marginal likelihood $L_{\mathbf{m}}(\boldsymbol{\theta}; \mathbf{y})$. Imposing the same set of conditions, (R1)–(R6), on the h -likelihood $L_{\mathbf{n}}(\boldsymbol{\omega}; \mathbf{y})$, Lee and Lee (2025) showed that the MHLE continued to satisfy these three properties, and consistent variance estimator of MHLE could be obtained from the Hessian matrix of the h -likelihood beyond the class of random effects models considered in this paper. No extra conditions are necessary for Bartlett identities (Meng, 2009) of models with continuous random effects. In this paper, let the degrees of freedom for the component $\boldsymbol{\psi} = (\boldsymbol{\beta}, \mathbf{v})$ be

3.1 GAIC for marginal model selection

$\rho = \text{df}$, which satisfies

$$\text{rank}(\mathbf{X}) \leq \rho = \dim(\boldsymbol{\psi}) \leq \text{rank}(\mathbf{X}, \mathbf{Z}), \quad (3.12)$$

where $\rho = \text{rank}(\mathbf{X})$ if $\mathbf{v} = 0$; and $\rho = \text{rank}(\mathbf{X}, \mathbf{Z})$ if $\mathbf{v}$ are fixed parameters. In one-way random-effects models, $\text{rank}(\mathbf{X}) + \text{rank}(\mathbf{Z}) - \text{rank}(\mathbf{X}, \mathbf{Z}) = 1$, showing that upper bound in (3.12) is tighter than that given in Donohue et al. (2011).

For the rest of Section 3, we consider optimal estimation of fixed parameters $\boldsymbol{\psi}_k$. By Equation (2.4), $\hat{\boldsymbol{\psi}}_{\text{MLE},k} = \hat{\boldsymbol{\psi}}_{\text{MHLE},k}$ and obviously $\ell_*(\hat{\boldsymbol{\psi}}_{\text{MLE},k}; \mathbf{y}) = \ell_*(\hat{\boldsymbol{\psi}}_{\text{MHLE},k}; \mathbf{y})$, which implies the following proposition.

Proposition 1. *If $\boldsymbol{\psi}_k$ is a subset of $\boldsymbol{\theta}_k$, then the model selected from the class of the subject-specific random-effects models $\ell_n(\boldsymbol{\omega}_k; \mathbf{y})$ coincides with that selected from the class of the marginal models $\ell_m(\boldsymbol{\theta}_k; \mathbf{y})$.*

Suppose there are two different LMMs with the same conditional mean, $\boldsymbol{\mu}(\mathbf{v}) = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{v}$ and $\boldsymbol{\mu}(\mathbf{u}) = \mathbf{X}\boldsymbol{\beta} + \mathbf{P}\mathbf{u}$, where $\dim(\mathbf{v})$ and $\dim(\mathbf{u})$ need not be equal, but having the same marginal model with $\boldsymbol{\mu}_m = \mathbf{X}\boldsymbol{\beta}$ and $\text{var}_m(\mathbf{y}) = \phi\mathbf{I}_n + \mathbf{Z}\text{var}(\mathbf{v})\mathbf{Z}^\top = \phi\mathbf{I}_n + \mathbf{P}\text{var}(\mathbf{u})\mathbf{P}^\top$. Lee and Nelder (2006b) showed that the two LMMs lead to the same conditional prediction of $\boldsymbol{\mu} = \mathbf{E}(\mathbf{y}|\mathbf{v}) = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{v}$ satisfying $\mathbf{X}\hat{\boldsymbol{\beta}}_{\text{MHLE}} + \mathbf{Z}\hat{\mathbf{v}}_{\text{MHLE}} = \mathbf{X}\hat{\boldsymbol{\beta}}_{\text{MLE}} + \mathbf{P}\hat{\mathbf{u}}_{\text{MHLE}}$, and the same marginal prediction of mean $\boldsymbol{\mu}_m = \mathbf{E}(\mathbf{y}) = \mathbf{X}\boldsymbol{\beta}$ because $\hat{\boldsymbol{\beta}}_{\text{MLE}} = \hat{\boldsymbol{\beta}}_{\text{MHLE}}$ and $\mathbf{Z}\hat{\mathbf{v}}_{\text{MHLE}} = \mathbf{P}\hat{\mathbf{u}}_{\text{MHLE}}$. Thus, random effects model allows both conditional and marginal inferences (Lee and Nelder, 2004). Proposition 1 further generalizes this result to model selection for random-effects models: both the marginal and the extended likelihoods give the same information criterion for fixed parameter $\boldsymbol{\psi}$, provided that any extended likelihood leads to the same marginal likelihood. This equivalence is a consequence of using the integrated likelihood from (2.4).

3.2 AIC

Let $\ell_* = \ell_m$, integrating out $\mathbf{v}$. Let $\boldsymbol{\psi} = \boldsymbol{\theta}$ and $\boldsymbol{\varsigma}$ be null. This yields the adjusted profile marginal likelihood, $L_{\text{ma}}(\boldsymbol{\psi}; \mathbf{y}) = L_m(\boldsymbol{\theta}; \mathbf{y})$ with the degrees of freedom

$$\text{df} = \text{tr}\{\mathbb{E}[n^{-1}\{-\partial^2 \ell_m(\boldsymbol{\theta}; \mathbf{y})/\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^\top\} | \boldsymbol{\theta} = \boldsymbol{\theta}_0]\} \text{var}(\sqrt{n}(\hat{\boldsymbol{\theta}}_{\text{MLE}} - \boldsymbol{\theta}_0)) = \dim(\boldsymbol{\theta}).$$

So, AIC is a GAIC. Akaike introduced his information,

$$\text{MAI} = -2\mathbb{E}_{\mathbf{y}}\mathbb{E}_{\mathbf{y}^*}\{\ell_m(\hat{\boldsymbol{\theta}}_{\text{MLE},k}; \mathbf{y}^*)\},$$

and his information criterion, AIC

$$\text{MAIC} = -2\ell_m(\hat{\boldsymbol{\theta}}_{\text{MLE},k}; \mathbf{y}) + 2\text{df}_k,$$

where $\text{df}_k = p_k + t_k + 1$, $p_k = \dim(\boldsymbol{\beta}_k)$ and $t_k = \dim(\boldsymbol{\lambda}_k)$. He showed Theorem 1, $\mathbb{E}_{\mathbf{y}}(\text{MAIC}) = \text{MAI} + o(1)$ without requiring the true model is to be in the class of candidate models in some cases. We prefix AI and AIC with M to emphasize the whole marginal parameter estimation of $\boldsymbol{\psi}_k = \boldsymbol{\theta}_k$.

In summary,

- (i) the MAIC provides ML estimation of $\boldsymbol{\psi}_k = \boldsymbol{\theta}_k$, achieving minimax-rate optimally in fixed parameter estimation (Yang, 2005). We shall use $\ell_{\text{ea}}(\boldsymbol{\theta}_k; \mathbf{y})$ when $\ell_m(\boldsymbol{\theta}_k; \mathbf{y})$ is difficult to obtain.
- (ii) From (2.4), the MAIC based on the h-likelihood

$$\text{MAIC} = -2\ell_n(\boldsymbol{\theta}_k, \mathbf{v}_k; \mathbf{y}) + 2(p_k + t_k + 1), \quad (3.13)$$

provides optimal MLE of $\boldsymbol{\theta}_k$ under the random effects model (Lee and Lee, 2025).

Thus, AIC can be extended to random-effects models with $\boldsymbol{\varsigma}_k = \mathbf{v}_k$ and $\boldsymbol{\psi}_k = \boldsymbol{\theta}_k$, which is a consequence of Proposition 1.

3.3 BIC and likelihood model information criterion

Setting $\boldsymbol{\psi}$ null and $\boldsymbol{\varsigma} = \boldsymbol{\theta}$ yields the adjusted profile marginal likelihood (2.5) $L_{\text{ma}}(\boldsymbol{\psi}; \mathbf{y}) = L_{\text{m}}(\hat{\boldsymbol{\theta}}_{\text{MLE}}; \mathbf{y})m(\mathbf{y})$ with $m(\mathbf{y}) = |\{-\partial^2 \ell_{\text{m}}(\boldsymbol{\theta}; \mathbf{y})/\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^\top\}/(2\pi)|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}_{\text{MLE}}}^{-1/2}$, which is the Laplace approximation to the integrated likelihood $\int L_{\text{m}}(\boldsymbol{\theta})d\boldsymbol{\theta}$. The degrees of freedom here is $\text{df} = \dim(\boldsymbol{\psi}) = 0$ as there is no parameter to estimate. Theorem 2 (ii) below shows that the resulting information criterion will be asymptotically equivalent to BIC. Thus, BIC is a GAIC.

By selecting a Bayesian candidate model with the largest marginal posterior for $\boldsymbol{\theta}$, Schwarz (1978) showed that BIC consistently selects the true Bayesian model. We show that BIC consistently selects the true likelihood model in this paper. Without assuming prior, we only need to consider the class of marginal candidate models $M_k : L_{\text{m}}(\boldsymbol{\theta}_k; \mathbf{y})$, $k = 1, \dots, K$. Suppose the true model is among the candidate models, $M_1, \dots, M_K$. There exists a parametrization of $\boldsymbol{\theta}_k$ such that $\boldsymbol{\delta}_k = \delta(\boldsymbol{\theta}_k)$, which may depend on the model M_k and $f(\boldsymbol{\theta}_k) = |\partial(\delta_1, \dots, \delta_{p_k})/\partial(\theta_1, \dots, \theta_{p_k})| = |\partial \boldsymbol{\delta}_k/\partial \boldsymbol{\theta}_k|$ denotes the Jacobian of the parametrization, $\boldsymbol{\delta}_k$. We define the integrated likelihood for model selection

$$L_0(M_k; \mathbf{y}, \delta) = \int L_{\text{m}}(\boldsymbol{\theta}_k; \mathbf{y}) d\boldsymbol{\delta}_k = \int L_{\text{m}}(\boldsymbol{\theta}_k; \mathbf{y}) |\partial \boldsymbol{\delta}_k/\partial \boldsymbol{\theta}_k| d\boldsymbol{\theta}_k \quad (3.14)$$

to satisfy $\int L_0(M_k; \mathbf{y}, \delta) d\mathbf{y} = 1$ subject to the constraint, $\int |\partial \boldsymbol{\delta}_k/\partial \boldsymbol{\theta}_k| d\boldsymbol{\theta}_k = 1$. Thus, $L_0(M_k; \mathbf{y}, \delta)$ is a density function of $\mathbf{y}$. The integrated likelihood corresponds to the marginal posterior, $P(M_k|\mathbf{y})$, by taking the Jacobian term as the prior $f(\boldsymbol{\theta}_k) = |\partial \boldsymbol{\delta}_k/\partial \boldsymbol{\theta}_k|$. This gives the density function $g(\mathbf{y}^*)$ of generating $\mathbf{y}^*$. It is well known that MLE is invariant to the parametrization. However, due to the Jacobian term, parametrization now plays an important role in model selection based on integrated likelihood $L_0(M_k; \mathbf{y}, \delta)$.

3.3 BIC and likelihood model information criterion

Let MI be the likelihood model information

$$\text{MI}(\delta) = -2\text{E}_{\mathbf{y}}\text{E}_{\mathbf{y}^*}\{\ell_0(\widehat{M}; \mathbf{y}^*, \delta)\}.$$

For likelihood model selection, we propose the model information criterion (MIC):

$$\text{MIC}(\delta) = -2\ell_0(\widehat{M}(\delta); \mathbf{y}, \boldsymbol{\delta}) + 2\text{df},$$

where $\text{df} = \dim(\boldsymbol{\psi}) = 0$ and the MLE is

$$\widehat{M}(\delta) = \arg \max\{\ell_0(M_k; \mathbf{y}, \boldsymbol{\delta}_k) : 1 \leq k \leq K\},$$

which maximizes $\ell_0(M_k; \mathbf{y}, \boldsymbol{\delta}_k) = \sum_{i=1}^n \ell_0(M_k; y_i, \boldsymbol{\delta}_k)$. The proof of Theorem 1 requires the asymptotic normality of MLE, which cannot be carried over to MLE, $\widehat{M}$, for discrete class of candidate models $\{M_1, \dots, M_K\}$.

Theorem 2. (i) *Suppose the true model is contained in the class of candidate models. If n is large enough, we have*

$$\text{E}_{\mathbf{y}}[\text{MIC}(\delta)] = \text{MI}(\delta).$$

(ii) *Asymptotically, $\text{MIC}(\delta)$ selects the same model as BIC does, which is independent of the unknown parametrization δ , and*

$$\text{BIC}(k) = -2\ell_m(\widehat{\boldsymbol{\theta}}_{\text{MLE},k}; \mathbf{y}) + (p_k + t_k + 1) \log n = \text{MIC}(\delta_k) + O_p(1),$$

where $\text{BIC} = \arg \min\{\text{BIC}(k) : 1 \leq k \leq K\}$.

Theorem 2 (i) shows that MIC is a GAIC with $\ell_* = \ell_0$. Thus, consistent likelihood model selection can be achieved by selecting the largest model likelihood just as consistent Bayesian model selection is achieved by selecting the largest marginal posterior. However, MIC depends on an unknown parametrization $\boldsymbol{\delta}$, which limits its use as a model selection

3.3 BIC and likelihood model information criterion

criterion. Theorem 2 (ii) implies that the BIC provides consistent model selection independent of the parametrization. Consequently, BIC selects the most parsimonious candidate model that contains the true data-generating model (Yang, 2005). We therefore adopt BIC for consistent likelihood model selection that is independent of the parametrization, which also ensures consistent Bayesian model selection independent of the choice of priors.

In summary,

- (i) letting $\boldsymbol{\psi}_k$ be null and $\boldsymbol{\varsigma}_k = \boldsymbol{\theta}_k$ gives simultaneous ML estimation and consistent marginal model selection;
- (ii) letting $\boldsymbol{\psi}_k$ be null and $\boldsymbol{\varsigma}_k = \boldsymbol{\omega}_k$ gives simultaneous MHL estimation and consistent random effects model selection.

This shows that BIC can be extended to random-effects models, which is again a consequence of Proposition 1 with $\boldsymbol{\psi}$ null.

For variable selection and estimation of $\boldsymbol{\theta}_k$, the penalized log-marginal-likelihood

$$\ell_{\text{mp}}(\boldsymbol{\theta}_k, \mathbf{y}) = \ell_{\text{m}}(\boldsymbol{\theta}_k; \mathbf{y}) - \kappa \text{df}_k,$$

has been considered. In penalized likelihood approach with $\text{df}_k = \dim(\boldsymbol{\theta}_k)$, $\kappa = 2$ and $\kappa = \log n$ in MAIC and BIC, respectively. Choosing the penalty factor κ appropriately, a lasso-type estimation (Tibshirani, 1996) can be obtained through a penalized likelihood approach for the component $\boldsymbol{\beta}_k$. The choice of a data-dependent κ is therefore crucial in penalized likelihood methods. In contrast, $\kappa = 2$ in all the GAICs for componentwise model selection and estimation while the degrees of freedom, $\text{df}_k = \dim(\boldsymbol{\psi}_k)$, varies according to the component of interest $\boldsymbol{\psi}_k$. In MAIC, $\text{df}_k = \dim(\boldsymbol{\psi}_k) = p_k + t_k + 1$. whereas in BIC, $\text{df}_k = \dim(\boldsymbol{\psi}_k) = 0$. In BIC, as a GAIC with $\boldsymbol{\varsigma}_k = \boldsymbol{\theta}_k$ and null $\boldsymbol{\psi}_k$, the term, $(p_k + t_k + 1) \log n$,

3.4 GAIC based on restricted likelihood

is due to the asymptotic order of integrating $\boldsymbol{\theta}_k$ out. We shall elaborate how to accommodate the lasso-type estimation in GAICs in Section 6.3.

3.4 GAIC based on restricted likelihood

Setting $\boldsymbol{\psi} = (\phi, \boldsymbol{\lambda})$, $\boldsymbol{\varsigma} = \boldsymbol{\beta}$ and $\ell_* = \ell_R$ shows that restricted likelihood is an integrated likelihood, $L_R(\boldsymbol{\psi}; \mathbf{y}) = \int L_m(\boldsymbol{\theta}; \mathbf{y}) d\boldsymbol{\varsigma} = L_i(\boldsymbol{\psi}; \mathbf{y})$. For LMM, the REML estimation was developed by Patterson and Thompson (1971) to reduce bias in estimating $\boldsymbol{\psi}$. For random-effects models, via simulation studies, Han and Lee (2024) demonstrated that integrated likelihood, $L_i(\boldsymbol{\psi}; \mathbf{y})$, significantly reduces the bias of estimating $\boldsymbol{\psi}$ in extreme binary data with temporal-spatially correlated random effects. Lee et al. (2017) showed that in LMM

$$L_R(\boldsymbol{\psi}; \mathbf{y}) = L_i(\boldsymbol{\psi}; \mathbf{y}) = L_{ma}(\boldsymbol{\psi}; \mathbf{y}) = L_{ea}(\boldsymbol{\psi}; \mathbf{y}),$$

which is also the same as the conditional density $f(\mathbf{y}|\mathbf{t}, \boldsymbol{\psi})$, conditioning on the sufficient statistic $\mathbf{t}$ for $\boldsymbol{\beta}$ (Smyth and Verbyla, 1996; Lee et al., 2017). From (3.9), $df = t + 1 = \dim(\boldsymbol{\psi})$. Thus, a GAIC can be constructed based on the restricted likelihood

$$\text{RAIC} = -2\ell_R(\hat{\boldsymbol{\psi}}_{\text{REMLE}}; \mathbf{y}) + 2(t + 1). \quad (3.15)$$

Asymptotically, using the restricted likelihood is equivalent to applying BIC for model selection of $\boldsymbol{\beta}$ for given $\boldsymbol{\psi}$:

$$-2\ell_R(\boldsymbol{\psi}; \mathbf{y}) = \{-2\ell_m(\hat{\boldsymbol{\beta}}_{\text{MLE}}(\boldsymbol{\psi}), \boldsymbol{\psi}; \mathbf{y}) + p \log n\}(1 + o(1)). \quad (3.16)$$

In random-effects models, the marginal likelihood is often difficult to compute. Lee and Nelder (2001) proposed to compute the restricted likelihood L_R by $L_{ea}(\boldsymbol{\psi}; \mathbf{y})$ as defined in (2.2).

We note that GAIC based on restricted likelihood uses BIC (3.16) for β and uses AIC (3.15) for ψ . This leads to consistent selection of β and optimal parameter estimation of ψ by iterating the following two steps:

Step (1): For a given ψ , apply the restricted likelihood (3.16) for consistent model selection of β .

Step (2): For a given β , apply RAIC (3.15) for optimal parameter estimation of ψ .

Throughout the paper, $\hat{\beta}_{\text{MLE}} = \hat{\beta}_{\text{MLE}}(\hat{\psi}_{\text{REMLE}})$ maximizes ℓ_{m} whereas $\hat{\psi}_{\text{REMLE}} = \hat{\psi}_{\text{REMLE}}(\hat{\beta}_{\text{MLE}})$ maximizes ℓ_{R} . RAIC selects a model to estimate ψ optimally and β consistently. GAIC for a component model adopts a stepwise procedure for model selection and parameter estimation. Restricted likelihood has been developed originally as an adjusted profile marginal likelihood to reduce bias. When the marginal likelihood is intractable to obtain, RAIC can be defined as a GAIC for random-effects models with $\psi = (\phi, \lambda)$, $\varsigma = (\beta, \mathbf{v})$ and $\ell_* = \ell_{\text{ea}}$.

4. GAICs for subject-specific model selection

4.1 Degrees of freedom for components

Given ϕ , Lee and Nelder (1996) defined the scaled deviance for a random effects model as

$$D(\hat{\mu}_{\text{MHLE}}, \phi; \mathbf{y}) = -2\ell_{\text{c}}(\hat{\mu}_{\text{MHLE}}, \phi; \mathbf{y}) + 2\ell_{\text{c}}(\mathbf{y}, \phi; \mathbf{y}) = \sum_{i=1}^n \hat{e}_i^2 / \phi, \quad (4.17)$$

where $\hat{e}_i = y_i - \hat{y}_i$, is the i -th residual in LMM, and the i -th deviance residual in random-effects models. They showed that

$$\text{E}[D(\hat{\mu}_{\text{MHLE}}, \phi; \mathbf{y})] = n - \rho + o(1),$$

where $\rho = \text{df}$, the degrees of freedom (3.9) for the component $\psi = (\beta, \mathbf{v})$, can be computed as in (4.18) below. For the goodness-of-fit tests for binomial and Poisson random effect models

4.1 Degrees of freedom for components

where ϕ is 1, they proposed the use of scaled deviance test. In this paper, we extend this result for $\boldsymbol{\psi}$ to any subset of $\boldsymbol{\omega}$. Spiegelhalter et al. (2002) introduced the Bayesian version of the scaled deviance for BIC.

The use of random-effects models is often criticised because the resulting estimates can be sensitive to the assumed distribution of the random effects, which is difficult to check in practice (Glidden and Liang, 2002). Noh et al. (2005) showed that such sensitivity can be mitigated by adopting heavy-tailed random effects. This can be done by adding another random effects $\mathbf{u}$ in the dispersion parameters as illustrated in the examples in Section 5. This leads to a t -like heavy-tailed distribution with the conditional moments $E(\mathbf{v}|\mathbf{u}) = \mathbf{0}$ and $\text{var}(\mathbf{v}|\mathbf{u}) = \boldsymbol{\Lambda}(\mathbf{u})$.

Let $\boldsymbol{\psi} = (\boldsymbol{\beta}, \mathbf{v})$ for prediction of the conditional mean $\boldsymbol{\mu}$, and $\mathbf{W} = \{\text{var}(\mathbf{y}|\mathbf{v}, \mathbf{u})\{d\boldsymbol{\eta}/d\boldsymbol{\mu}\}^2\}^{-1}$. Then, $\mathbf{B}_1$ and $\mathbf{B}_2$ in (3.10) can be written as:

$$\begin{aligned} n\mathbf{B}_1 &= \begin{bmatrix} \mathbf{X}^\top \mathbf{W} \mathbf{X} & \mathbf{X}^\top \mathbf{W} \mathbf{Z} \\ \mathbf{X}^\top \mathbf{W} \mathbf{Z} & \mathbf{Z}^\top \mathbf{W} \mathbf{Z} \end{bmatrix}, \\ \frac{1}{n}\mathbf{B}_2 &= \text{var}(\hat{\boldsymbol{\psi}}_{\text{MHLE}} - \boldsymbol{\psi}_0) = \begin{bmatrix} \mathbf{X}^\top \mathbf{W} \mathbf{X} & \mathbf{X}^\top \mathbf{W} \mathbf{Z} \\ \mathbf{X}^\top \mathbf{W} \mathbf{Z} & \mathbf{Z}^\top \mathbf{W} \mathbf{Z} + \boldsymbol{\Lambda}(\mathbf{u})^{-1} \end{bmatrix}^{-1}, \end{aligned} \quad (4.18)$$

where $\boldsymbol{\psi} = (\boldsymbol{\beta}^\top, \mathbf{v}^\top)^\top$ and $\boldsymbol{\psi}_0 = (\boldsymbol{\beta}_0^\top, \mathbf{v}_0^\top)^\top$. In LMM, $\mathbf{W} = \mathbf{I}_n/\phi$ and $\boldsymbol{\Lambda}(\mathbf{u}) = \boldsymbol{\Lambda}$. The degrees of freedom, ρ , can be found in Lee and Nelder (1996). They relaxed the normality assumption on the random effects by considering independent non-normal random effects for HGLMs. Hodges and Sargent (2001) investigated various degrees of freedom used in smoothing literature as in Hastie and Tibshirani (1990). Spiegelhalter et al. (2002) derived the degrees of freedom, ρ , as a measure of Bayesian model complexity. For LMM, Vaida and Blanchard (2005) defined CAIC for GLMMs, which is a GAIC with $\ell_* = \ell_c$. For

4.2 Conditional information criterion

correlated normal random effects with some nonzero off-diagonal elements in $\mathbf{\Lambda}$, they found that $\rho = \text{df}$ satisfying (3.11). Ha et al. (2007) applied CAIC to frailty models, which can be fitted by Poisson HGLMs (Ha and Lee, 2003). Donohue et al. (2011) and Yu and Yau (2012) justified the use of CAIC for GLMMs with $\phi = 1$. It can be seen now that the degrees of freedom, $\rho = \text{df}$, can be extended to robust HGLMs. In robust HGLMs, $(\mathbf{v}, \mathbf{u})$ are random effects. For subject-specific prediction of $\boldsymbol{\psi} = (\boldsymbol{\beta}, \mathbf{v})$, the random effects $\mathbf{v}$, not the random effects $\mathbf{u}$, are of interest to predict in this paper. In (4.18), $\text{var}(\hat{\boldsymbol{\psi}}_{\text{MHLE}} - \boldsymbol{\psi}_0)$, based on $-\partial^2 \ell_n(\boldsymbol{\omega}; \mathbf{y}) / \partial \boldsymbol{\psi} \partial \boldsymbol{\psi}^\top$, is the variance estimator when $\boldsymbol{\varsigma}$ is known (Lee and Nelder, 1996). Lee and Lee (2025) showed that the full Hessian matrix provides consistent variance estimator of $\hat{\boldsymbol{\psi}}_{\text{MHLE}}$. For a similar adjustment in LMM, see Liang et al. (2008).

4.2 Conditional information criterion

We have studied information criteria for marginal models with fixed parameters only. The conditional likelihood $L_c(\boldsymbol{\psi}, \phi; \mathbf{y}) = L_c(\boldsymbol{\mu}, \phi; \mathbf{y}) = f(\mathbf{y}|\boldsymbol{\omega}) = f(\mathbf{y}|\boldsymbol{\beta}, \mathbf{v}, \phi)$ is the GLM likelihood with $(\boldsymbol{\beta}, \mathbf{v}, \phi)$ as fixed parameters, leading to the GLM MLE. To improve estimation of treatment effects $\boldsymbol{\beta}$ in linear models, Yates (1940) showed that fixed block effects $\mathbf{v}$ should be treated as random to recover inter-block information. Similarly, in binary matched pairs, the bias of $\hat{\boldsymbol{\beta}}_{\text{MLE}}$, can be reduced by treating pair effects as random (Lee, 2001). In this section, we employ information criteria to evaluate if including random effects improves subject-specific predictions

Suppose ϕ is known and subject-specific prediction for $\boldsymbol{\mu}$ is of interest with $\boldsymbol{\psi} = (\boldsymbol{\beta}, \mathbf{v})$. If we apply GAIC to models with fixed $\mathbf{v}$, CAIC is equivalent to MAIC with $\text{df} = \dim(\boldsymbol{\psi}) = \dim(\boldsymbol{\beta}) + \dim(\mathbf{v}) = \text{rank}(\mathbf{X}, \mathbf{Z})$. If we apply GAIC to random $\mathbf{v}$ for subject-specific prediction, parameter and variance estimators are not derived from the GLM by taking $\ell_* = \ell_c$ but from

4.2 Conditional information criterion

the random effects model to meet the inferential aim. Even if the true model is a GLM with fixed $\mathbf{v}$, the GAIC could prefer to treat $\mathbf{v}$ random for better subject-specific prediction with $\text{df} = \rho$. Following Vaida and Blanchard (2005) we require cluster focus asymptotic, where $\rho \leq \text{rank}(\mathbf{X}, \mathbf{Z})$ remains fixed as $n \rightarrow \infty$, which is key for consistency and valid variance estimation of MHLE under the GLM score equation from the conditional likelihood $L_* = L_c$.

Proposition 1 is not applicable because $\boldsymbol{\psi}$ contains $\mathbf{v}$. The CAIC is the GAIC of the form (3.8) with $\ell_*(\boldsymbol{\psi}, \phi; \mathbf{y}) = \ell_c(\boldsymbol{\psi}, \phi; \mathbf{y})$. By using conditional likelihood, parameter $\boldsymbol{\lambda}$ is eliminated. For LMM, Vaida and Blanchard (2005) proposed the use of CAIC for selection of subject-specific component $\boldsymbol{\psi} = (\boldsymbol{\beta}, \mathbf{v})$. For random-effects models with $\phi = 1$ and $\boldsymbol{\Lambda}(\mathbf{u}) = \boldsymbol{\Lambda}$, Donohue et al. (2011) and Yu and Yau (2012) defined the CAIC as

$$\text{CAIC} = -2\ell_c(\hat{\boldsymbol{\psi}}_{\text{MHLE}}; \mathbf{y}) + 2\rho,$$

and showed Theorem 1 for CAIC, $E_{\mathbf{y}|\mathbf{u}}[\text{CAIC}] = \text{CAI} + o(1)$. Thus, CAIC is a GAIC with $\ell_* = \ell_c$ and $\boldsymbol{\psi} = (\boldsymbol{\beta}, \mathbf{v})$. In LMM, the MHLE can be consistent even if $\mathbf{u}$ is non-normal as Vaida and Blanchard (2005) noted. However, in random-effects models, if the distribution of random effects is misspecified, there is no guarantee that the MHLE can be consistent (Glidden and Liang, 2002). For random-effects models, we require that true model is contained in a candidate model even if $E_{\mathbf{y}|\mathbf{u}}$ is independent of the distributional assumption on $\mathbf{u}$.

The degrees of freedom for a fixed parameter is 1, whereas it is between 0 and 1 for a random effect (Lee et al., 2017). By (3.12), $\boldsymbol{\psi} = (\boldsymbol{\beta}, \mathbf{v})$ implies that $0 \leq \dim(\mathbf{v}) \leq \text{rank}(\mathbf{X}, \mathbf{Z}) - \text{rank}(\mathbf{X})$. Thus, the degrees of freedom for random effects, $\dim(\mathbf{v})$, can be estimated by $\hat{\rho}_{\text{MHLE}} - \text{rank}(\mathbf{X})$. If $\hat{\rho}_{\text{MHLE}}$ is a consistent estimator of ρ , $\rho - \text{rank}(\mathbf{X})$ is the degrees of freedom for a random component $\mathbf{v}$, which depends on the model assumption and its parameters. Note that $\rho = \dim(\boldsymbol{\psi})$ is the degrees of freedom for the component $\boldsymbol{\psi}$, but

4.3 Information criteria for subject-specific model selection

not for the entire component ω .

When ϕ is unknown, difficulty arises. With an additional assumption that the ratio Λ/ϕ is known, Vaida and Blanchard (2005) derived ρ as given in (3.9) for LMM. When ϕ is unknown, as in Vaida and Blanchard (2005), let $\text{df} = \hat{\rho}_{\text{MHLE}} + 1$ to have

$$\text{CAIC} = -2\ell_{\text{c}}(\hat{\psi}_{\text{MHLE}}; \mathbf{y}) + 2(\hat{\rho}_{\text{MHLE}} + 1).$$

We view their CAIC as a GAIC with $\psi = (\beta, \mathbf{v}, \phi)$. Here, $\text{df} = \dim(\psi) = \dim(\beta) + \dim(\phi) + \dim(\mathbf{v})$. From Theorem 1, the CAIC can be extended to log-normal and gamma random effects with unknown ϕ .

4.3 Information criteria for subject-specific model selection

For estimating $\psi = (\beta, \mathbf{v})$, we may use either $L_{\text{i}}(\psi; \mathbf{y}) = \int L_{\text{e}}(\omega; \mathbf{y}) d\zeta$ with $\zeta = (\phi, \lambda)$ or $L_{\text{i}}(\psi; \mathbf{y}) = \int L_{\text{c}}(\psi, \zeta; \mathbf{y}) d\zeta$ with $\zeta = \phi$. Using the latter, we have the following proposition.

Proposition 2. In LMM, for $a = \Gamma((n-2)/2)/(2\pi^{n/2})$,

$$L_{\text{i}}(\mu; \mathbf{y}) = \int L_{\text{c}}(\mu, \phi; \mathbf{y}) d\phi = a\{\|\mathbf{y} - \mu\|^2\}^{-(n-2)/2} = a\left\{\sum e_i^2\right\}^{-(n-2)/2}. \quad (4.19)$$

Suppose that we want to construct a GAIC with $\ell_* = \ell_{\text{i}}$ and $\psi = (\beta, \mathbf{v})$ for subject-specific prediction. We have the generalized Akaike information

$$\text{SAI} = -2\text{E}_{\mathbf{y}|\mathbf{u}}\text{E}_{\mathbf{y}^*|\mathbf{u}}\{\ell_{\text{i}}(\hat{\mu}_{\text{MHLE}}; \mathbf{y}^*)\} = (n-2)\text{E}_{\mathbf{y}|\mathbf{u}}\text{E}_{\mathbf{y}^*|\mathbf{u}}\log\{\|\mathbf{y}^* - \hat{\mu}_{\text{MHLE}}\|^2\} - 2\log a,$$

and the corresponding GAIC,

$$\text{SAIC} = -2\ell_{\text{i}}(\hat{\psi}_{\text{MHLE}}; \mathbf{y}) + 2\rho = (n-2)\log\{\|\mathbf{y} - \hat{\mu}_{\text{MHLE}}\|^2\} - 2\log a + 2\rho, \quad (4.20)$$

where

$$-2\ell_{\text{i}}(\psi; \mathbf{y}) = \{-2\ell_{\text{c}}(\psi, \hat{\phi}_{\text{MHLE}}(\psi); \mathbf{y}) + \log n\}(1 + o(1)) \quad (4.21)$$

4.3 Information criteria for subject-specific model selection

by applying Laplace approximation to $\ell_i(\boldsymbol{\psi}; \mathbf{y})$. Since a only depends on n , $\log a$ will be ignored for the rest of this paper as it has no effect on the model selection. For a given $\boldsymbol{\psi} = (\boldsymbol{\beta}, \mathbf{v})$, the integrated likelihood (4.21) provides consistent model selection by using BIC for the component ϕ . Then, SAIC (4.20) provides optimal estimation by using AIC for the component $\boldsymbol{\psi}$.

Theorem 3. *In LMM, under appropriate conditions,*

$$E[\text{SAIC}] = \text{SAI} + o(1).$$

In LMM, $\mathbf{v}$ may not be necessarily normal, i.e., true model may not be a LMM. From (4.20),

$$\text{SAIC} = (n - 2) \log \left\{ \sum_{i=1}^n \hat{e}_i^2 \right\} + 2\hat{\rho}_{\text{MHLE}}, \quad (4.22)$$

where $\hat{e}_i = y_i - \hat{\mu}_{\text{MHLE},i}$ is the i -th residual in the LMM. This shows that the degrees of freedom for random effects $\mathbf{v}$ is again estimated by $\hat{\rho}_{\text{MHLE}} - \text{rank}(\mathbf{X})$ because $\dim(\boldsymbol{\psi}) = \dim(\boldsymbol{\beta}) + \dim(\mathbf{v})$.

Suppose that $\mathbf{v}$ can be modeled as either random or fixed parameters. The residual sum of squares, $\sum \hat{e}_i^2$, is minimized when $(\hat{\boldsymbol{\beta}}, \hat{\mathbf{v}})$ are the MLEs for linear model with fixed $(\boldsymbol{\beta}, \mathbf{v})$. Let $\text{SAIC}(F)$ and $\text{SAIC}(R)$ be the SAIC under fixed and random $\mathbf{v}$, respectively. When $\mathbf{v}$ is fixed, we denote the right-hand-side of (4.22) by $\text{SAIC}(F)$. Similarly, define $\text{SAIC}(R)$ when $\mathbf{v}$ is random. Modeling $\mathbf{v}$ as random is preferred to fixed iff $\text{SAIC}(F) \geq \text{SAIC}(R)$ iff $2\{\text{rank}(\mathbf{X}, \mathbf{Z}) - \rho\}/(n - 2) \geq \log \left\{ \sum \hat{e}_{ri}^2 / \sum \hat{e}_{fi}^2 \right\}$, where $\hat{e}_{fi}$ and $\hat{e}_{ri}$ denote the i -th residual when $\mathbf{v}$ is modeled as fixed and random, respectively. Thus, treating $\mathbf{v}$ as random is preferred to treating $\mathbf{v}$ as fixed if an increase in the residual sum of squares is offset by a decrease in the degrees of freedom for $\mathbf{v}$, $\dim(\mathbf{v}) = \rho - \text{rank}(\mathbf{X})$, by assuming $\mathbf{v}$ random. Degree of freedom for a fixed parameter is well known to be one, whereas that for a random parameter

is fractional, between 0 and 1 (Lee et al., 2017). In this paper, the degrees of freedom of random parameters can be defined explicitly so that conventional model selection can be extended to models with random parameters. For fixed $\mathbf{v}$, $\dim(\mathbf{v}) = \text{rank}(\mathbf{X}, \mathbf{Z}) - \text{rank}(\mathbf{X})$. Even if a GLM is selected by other information criteria, SAIC and CAIC could favor treating $\mathbf{v}$ as random for a better subject-specific prediction.

By conditioning on $\mathbf{v}$, CAIC retains the effects $\mathbf{v}$ and allows us to determine whether to treat $\mathbf{v}$ as random or fixed for estimating $(\boldsymbol{\beta}, \phi, \mathbf{v})$. Theorem 3 shows that model selection for subject-specific prediction $(\boldsymbol{\beta}, \mathbf{v})$ can be achieved by conditioning $\mathbf{v}$ followed by integrating out ϕ .

5. Likelihood model selection as a Bayesian model selection

Lee (2026) clarified the relationship among likelihood, Bayesian posterior, and frequentist confidence procedure. Under suitable parametrization, a normalized likelihood corresponds to a Bayesian posterior, with the associated Jacobian term acting as a prior. Consequently, the credibility level of a Bayesian credible interval corresponds to the frequentist coverage probability of a confidence interval based on the normalized likelihood, and the two intervals are asymptotically equal. This paper extends this asymptotic property to model selection, allowing likelihood-based model selection to be interpreted as Bayesian model selection for given data $\mathbf{y}$. This observation and frequentist Theorems 1–3 imply that likelihood bridges the Bayesian and the frequentist approaches for model selection.

In the Bayesian approach with random parameters only, the extended likelihood is $L_e(\boldsymbol{\omega}; \mathbf{y}) = f(\mathbf{y}|\boldsymbol{\omega})f(\boldsymbol{\omega})$, where the conditional likelihood $L_c(\boldsymbol{\omega}; \mathbf{y}) = f(\mathbf{y}|\boldsymbol{\omega})$ is the Bayesian likelihood and $f(\boldsymbol{\omega})$ is the prior. It corresponds to assuming a joint prior $f(\boldsymbol{\omega}) = f(\mathbf{v}|\boldsymbol{\theta})f(\boldsymbol{\theta})$

5.1 DIC versus Epistemic CAIC

with $f(\mathbf{v}|\boldsymbol{\theta}) = f(\mathbf{v}|\boldsymbol{\lambda})$ and a uniform prior for fixed parameters, $f(\boldsymbol{\theta}) = 1$. Thus, the integrated likelihood for $\boldsymbol{\psi}$ corresponds to the marginal posterior $f(\boldsymbol{\psi}|\mathbf{y})$. In the likelihood-based approach, the assumption on the random effects distribution is checkable (Lee et al., 2017), and so $f(\mathbf{v}|\boldsymbol{\lambda})$ is treated as part of the model. The MLE and MHLE correspond to the posterior modes under appropriate priors. Analogous to the marginal posterior of a model M_k , $P(M_k|\mathbf{y})$, we refer to $L_0(M_k; \mathbf{y}, \boldsymbol{\delta}_k)$ in this paper as the likelihood of the model M_k , which is obtained by integrating out $\boldsymbol{\theta}_k$ in (3.14). The frequentist CAIC corresponds to the Bayesian DIC based on the posterior $f(\boldsymbol{\omega}|\mathbf{y})$ (Spiegelhalter et al., 2002), while the SAIC corresponds to model selection based on the marginal posterior $f(\boldsymbol{\psi}|\mathbf{y})$, potentially leading to componentwise Bayesian model selection. This implies that, from a Bayesian perspective, CAIC allows a choice between the two priors: a uniform prior for $\mathbf{v}$, corresponding to fixed $\mathbf{v}$, and $f(\mathbf{v}|\boldsymbol{\lambda})$ as a prior for $\mathbf{v}$, corresponding to random $\mathbf{v}$. Thus in the next section, we shall that GAICs support not only frequentist model selection based on future data but also Bayesian model selection based on the observed data.

5.1 DIC versus Epistemic CAIC

We have shown that likelihood facilitates both frequentist and Bayesian model selections. Lee and Nelder (1996) derived ρ in the context of a frequentist chi-square deviance test for goodness-of-fit; here, it serves as a likelihood-based measure of degrees of freedom, $\dim(\boldsymbol{\beta}, \mathbf{v})$, for the subject-specific component $\boldsymbol{\psi} = (\boldsymbol{\beta}, \mathbf{v})$. Spiegelhalter et al. (2002) interpreted ρ as a Bayesian measure of model complexity in DIC. A key implication of asymptotic invariance—both to parametrization in the likelihood framework and to prior choice in the Bayesian framework—is that the two approaches yield the same measure of model complexity for random-effects models, an observation that appears to have been largely overlooked.

5.1 DIC versus Epistemic CAIC

While integration is commonly used in Bayesian inference to eliminate nuisance parameters, we show that integrated likelihoods enable simultaneous componentwise estimation of parameters of interest and consistent selection of remaining parameters. Parametrization in the likelihood framework corresponds to prior specification in the Bayesian framework, suggesting a useful parallel between the two approaches. In Section 3.3, we further show that the model likelihood $L_0(M_k; \mathbf{y})$ can be interpreted as an analogue of the marginal posterior of model M_k . This perspective also suggests a natural extension of DIC using marginal posteriors for componentwise model selection.

In this section, we study epistemic information criteria for given data. Let $\hat{\boldsymbol{\psi}}$ be the solution of a likelihood estimating equation $\ell'_*(\hat{\boldsymbol{\psi}}) = 0$, which leads to

$$\ell_*(\boldsymbol{\psi}; y) = \ell_*(\hat{\boldsymbol{\psi}}; y) - \frac{1}{2}(\hat{\boldsymbol{\psi}} - \boldsymbol{\psi})^\top \{-\ell''_*(\hat{\boldsymbol{\psi}}; y)\}(\hat{\boldsymbol{\psi}} - \boldsymbol{\psi}) + O_p(n^{-1/2}).$$

We may consider an epistemic information criterion

$$\text{GAI}_{\mathbf{y}} = -2\text{E}_{\mathbf{y}}\{\ell_*(\boldsymbol{\psi}; \mathbf{y})\} \quad \text{and} \quad \text{GAIC}_{\mathbf{y}} = -2\ell_*(\hat{\boldsymbol{\psi}}; \mathbf{y}) + \rho,$$

where $\rho = \text{tr}\{(-\ell''_*(\hat{\boldsymbol{\psi}}; \mathbf{y}))\text{E}_{\mathbf{y}}(\hat{\boldsymbol{\psi}} - \boldsymbol{\psi})(\hat{\boldsymbol{\psi}} - \boldsymbol{\psi})^\top\}$. Then, we have

$$\text{E}_{\mathbf{y}}(\text{GAIC}_{\mathbf{y}}) = \text{GAI}_{\mathbf{y}} + o(1). \quad (5.23)$$

By maximizing likelihood measure for adequacy of fit, $\text{E}_{\mathbf{y}}\{\ell_*(\boldsymbol{\psi}; \mathbf{y})\}$, we can simultaneously have a MLE $\hat{\boldsymbol{\psi}}$ maximizing $\ell_*(\boldsymbol{\psi}; \mathbf{y})$ and an epistemic measure for adequacy of fit $\text{GAIC}_{\mathbf{y}}$ to select a model.

Consider an epistemic Bayesian measure for adequacy of fit in Spiegelhalter et al. (2002)

$$\text{DI}_{\mathbf{y}} = -2\text{E}_{\boldsymbol{\psi}|\mathbf{y}}\{\ell_*(\boldsymbol{\psi}; \mathbf{y})\},$$

5.1 DIC versus Epistemic CAIC

where $E_{\psi|\mathbf{y}}(\cdot)$ is the expectation under the marginal posterior of ψ . Then, the corresponding DIC is

$$\text{DIC} = -2\ell_*(\hat{\psi}; \mathbf{y}) + \rho = \text{GAIC}_{\mathbf{y}} + o_p(1),$$

provided $\hat{\psi}$ is a consistent estimator, for example, the marginal posterior mean $\hat{\psi} = E(\psi|\mathbf{y})$.

Since

$$\text{GAI}_{\mathbf{y}} = E_{\mathbf{y}}\text{DI}_{\mathbf{y}} + o(1)$$

if $\ell_* = \ell_{\text{c}}$ and $\hat{\psi} = \hat{\psi}_{\text{MHLE}}$, DIC corresponds to $\text{GAIC} = \text{GAIC}_{\mathbf{y}} \propto D(\hat{\mu}_{\text{MHLE}}, \phi; \mathbf{y}) + \rho$ where $D(\hat{\mu}_{\text{MHLE}}, \phi; \mathbf{y})$ is the scaled deviance (4.17) for goodness-of-fit. Thus, DIC is analogous to the epistemic likelihood-based measure of adequacy of fit $\text{GAIC}_{\mathbf{y}}$. Recently, the epistemic interpretation of confidence for the observed interval based on given data is reconciled with aleatory frequentist coverage probability interpretation for interval-generating procedure based on future data (Pawitan et al., 2023; Lee, 2025, 2026); epistemic confidence is analogous to epistemic Bayesian credibility. Thus, DIC could lead to an epistemic interpretation of the frequentists' GAIC, which provides frequentist interpretation of DIC: see the analogy between marginal posterior and confidence likelihood (Lee and Lee, 2025). Under regularity conditions (R1)-(R6), $\ell'_c(\hat{\psi}_{\text{MHLE}}) = 0$ and the frequentist GAI, based on future data, requires asymptotic normality, whereas epistemic $\text{GAI}_{\mathbf{y}}$ for given data does not require it. Frequentists' GAI in Section 3.1 could be independent of the true distribution of the random effects, whereas epistemic $\text{GAI}_{\mathbf{y}}$ may not. Thus, the epistemic information criterion $\text{GAI}_{\mathbf{y}}$ satisfies (5.23) under weaker conditions than those of the frequentists' GAI. And so it could remain effective in high-dimensional settings where degrees of freedom increase with sample size: see analogous discussion in Lee and Lee (2025). Further studies are of interest.

6. Real data examples

Various information criteria have been proposed for model selection. However, difficulties arise when different criteria select different models. Criteria should therefore align with specific inferential objectives. We illustrate this using four models derived from (2.1). M1, GLM with $\mathbf{v} = \mathbf{0}$; M2, GLM with fixed $\mathbf{v}$; M3, GLMM with $\mathbf{v} \sim N(0, \mathbf{\Lambda})$, where $\mathbf{\Lambda} = \lambda_0 \mathbf{G}$; and M4, robust HGLM with $\mathbf{v}|\mathbf{u} \sim N(0, \mathbf{\Lambda}(\mathbf{u}))$, inverse-gamma($\alpha + 1, \alpha$), u_i , $E(u_i) = 1$ and $\text{var}(u_i) = \delta$, where u_i is the i -th entry of $\mathbf{u}$ in the examples. This paper restricts u_i s to the inverse gamma distribution for robust HGLMs. We apply these GAICs for component model selection in the four real datasets.

GLMs with independent errors, M1 and M2 correspond to the random effects models with $\lambda_1 = 0$ and $\lambda_1 = \infty$, respectively (Lee and Nelder, 1996). Thus, random effects model is a compromise between two GLMs, M1 and M2. The robust HGLM, M4, being shrunk more than M3, leans more towards M1 than ordinary random effects model, M3. Models M1, M4, M3 and M2 increase in order of $\dim(\mathbf{v})$; whereas M1, M4, M3 and M2 decrease in the (deviance) residual sum of squares when effects $\mathbf{v}$ are present.

6.1 Model selection using GAIC

The first three real data examples—Cadralazine, Crack Growth, and Exam—are drawn from the LMM framework and illustrate the impact of model selection across increasing sample sizes ($n = 60$, 241, and 4059, respectively). The fourth example, the Epilepsy dataset ($n = 236$), involves count data. Additionally, the second Crack Growth dataset is used to demonstrate log-normal and gamma random-effects models. [Detailed descriptions of all examples and sources of datasets are provided in §S5 in the supplementary materials.](#)

6.1 Model selection using GAIC

The components ψ and ς corresponding to each GAIC—BIC, MAIC, RAIC, CAIC (DIC) and SAIC—are provided in Supplementary Table A. Except for the newly proposed SAIC, these criteria have been widely used, albeit typically in isolation. Here, we investigate model selection based on these GAICs jointly. Table 6.1 summarizes the results for selecting models that estimate ψ while consistently selecting the remaining component ς .

In the Cadralazine data, BIC and MAIC select the GLM M2; RAIC and CAIC select the robust HGLM M4; and SAIC selects the GLMM M3. This lack of agreement indicates no clear consensus among the criteria, possibly due to the relatively small sample size ($n = 60$), which may be insufficient for asymptotic model selection properties to hold. However, in the Exam data, which has a large sample size, all information criteria select the robust HGLM M4.

In the Crack Growth data, under the LMM framework, all the information criteria except SAIC select the robust LMM M4 whereas SAIC selects the linear model M2. CAIC targets estimation of β , $\mathbf{v}$, and ϕ , whereas SAIC focuses on β and $\mathbf{v}$. This suggests that model selection is sensitive to the estimation of the single parameter ϕ , which is undesirable. Consequently, the LMM may not be suitable for analyzing the Crack Growth data. This motivates consideration of alternative models such as log-normal and gamma random-effects models. All the GAICs favor the log-normal random-effects model over the LMM, with all criteria selecting the robust log-normal HGLM M4. For the gamma random-effects model, the conditional likelihood involves a complex function of ϕ . Since Theorem 3 has yet to be established for this model, SAIC values are not reported. All the GAICs prefer the gamma random-effects model over the log-normal model, and all criteria select the robust HGLM M4. Thus, the robust gamma HGLM M4 is selected to satisfy the inferential aims of all

GAICs.

In the Epilepsy data, a Poisson model is applied. Since $\phi = 1$, SAIC is not applicable. CAIC is therefore used to estimate the subject-specific parameters $\boldsymbol{\psi} = (\boldsymbol{\beta}, \mathbf{v})$. BIC selects the GLMM M3, corresponding to a marginal model for correlated errors with fixed parameters $\boldsymbol{\beta}, \boldsymbol{\lambda}$ and $\phi = 1$. If the primary objective is to estimate fixed parameters $\boldsymbol{\beta}, \boldsymbol{\lambda}$, and $\phi = 1$ under a GLM with independent errors and fixed $\mathbf{v}$, the parsimonious GLM with $\lambda_1 = \infty$ is selected based on MAIC and RAIC. If the objective is subject-specific prediction, CAIC selects the robust HGLM M4. These results highlight that the inferential objective should guide the choice of information criterion.

6.2 Simulation study

Several GAICs such as AIC, BIC, and RIC have been extensively studied both theoretically and through simulation studies, recent work on CAIC has focused on LMMs (Vaida and Blanchard, 2005), frailty models (Ha et al., 2007), and Poisson GLMMs (Donohue et al., 2011). However, no such simulation studies currently exist for gamma HGLMs. To address this gap, we present a targeted simulation study for this setting. We fitted the crack growth data using the gamma GLM (M2). We generated fixed unknown constant vectors $\mathbf{c}$ from $\text{Uniform}(-1, 1)$. Then, by taking $\beta_0 = \hat{\beta}_{\text{MLE}}$, $\phi_0 = \hat{\phi}_{\text{MLE}}$ and $\mathbf{v}_0 = \alpha \mathbf{c}$ as true values, where $\alpha = 0.01, 0.05, 0.1, 0.2, 0.3, 0.5, 1, 2$ and 5 , 100 simulated data were generated from M2 and compute the predicted squared deviation (PSD) $\sum (\mu_{0ij} - \hat{\mu}_{ij})^2 / \mu_{0ij}^2$. In the literature, the term ‘oracle’ is used in various contexts. Here (i) Oracle is an estimation method as if the true model M2 is known in advance. In practice, the true model is unknown but it is presented for baseline comparison. (ii) BIC and CAIC select model using BIC and CAIC, respectively. (iii) Two-stage is to first select a model using BIC, then $\hat{\mu}_{ij}$ is computed from

Table 1: Likelihood based model selection for components in the four datasets

IC	M1	M2	M3	M4	M1	M2	M3	M4
Cadralazine data, $n = 60$, LMM				Crack growth data, $n = 241$, log-normal				
BIC	128.2	<u>1.2</u>	23.7	7.1	57.1	-0.05	-32.6	<u>-211.3</u>
MAIC	121.9	<u>-42.8</u>	11.1	-7.6	46.6	-80.2	-46.5	<u>-228.7</u>
RAIC	128.1	34.1	16.8	<u>9.3</u>	51.7	-39.6	-42.2	<u>-224.0</u>
CAIC	121.9	-42.8	-44.4	<u>-51.8</u>	46.6	-80.2	-81.7	<u>-82.1</u>
SAIC	188.9	26.7	<u>25.8</u>	26.1	676.8	550.4	549.0	<u>546.6</u>
Crack growth data, $n = 241$, LMM				Crack growth data, $n = 241$, gamma model				
BIC	150.0	149.8	106.1	<u>-4.5</u>	47.7	-10.0	-39.5	<u>-221.6</u>
MAIC	139.6	69.7	92.2	<u>-21.9</u>	37.3	-90.1	-53.4	<u>-239.0</u>
RAIC	156.0	229.3	107.4	<u>-5.8</u>	42.7	-45.0	-49.3	<u>-234.1</u>
CAIC	139.6	69.7	67.0	<u>55.4</u>	37.3	-90.1	-91.6	<u>-99.1</u>
SAIC	781.1	<u>711.1</u>	712.6	714.3	--	--	--	--
Exam data, $n = 4059$, LMM				Epilepsy data, $n = 236$, Poisson model				
BIC	9785.4	9697.1	9390.9	<u>9084.5</u>	1668.7	1484.4	<u>1369.7</u>	1371.2
MAIC	9766.5	9274.4	9365.7	<u>9053.0</u>	1647.9	<u>1276.6</u>	1345.5	1343.5
RAIC	9776.3	9326.6	9373.1	<u>9224.9</u>	1664.7	<u>1217.6</u>	1350.4	1349.1
CAIC	9776.5	9274.4	9268.3	<u>9109.8</u>	1647.9	1276.6	1272.4	<u>1271.2</u>
SAIC	31954.8	31462.4	31456.4	<u>31454.3</u>	--	--	--	--

6.3 Lasso-type versus random-effects estimation

the M1 if M1 ($\mathbf{v} = \mathbf{0}$) is selected; otherwise, it is computed from the model selected by CAIC. Consequently, if M1 is not selected by BIC, the two-stage and CAIC are the same. BIC could select random-effects model. BIC tends to select the true M2 as α increases. When $\alpha \geq 0.2$, M1 is not selected by BIC, and so the two-stage and the CAIC give the same PSD. When $\alpha \leq 0.05$, M1 with $\mathbf{v} = \mathbf{0}$ has the smallest PSD, and therefore BIC tends to give a smaller PSD than that of CAIC by selecting M1. When $1 > \alpha \geq 0.1$, CAIC tends to give smallest PSD. When $0.05 \leq \alpha \leq 0.1$, the two-stage differs from both BIC and CAIC. When α is large ($\alpha \geq 1$), the oracle method (M2) gives the smallest PSD. When $\alpha = 5$, BIC selects the true model M2 with the highest probability whereas CAIC gives the smallest PSD. Even if the true model is a fixed-effects model M2, we prefer to use CAIC if the aim is a subject-specific prediction of μ_{0ij} . Simulation results show that CAIC is generally better for subject-specific prediction whereas BIC is better for consistent model selection when α is large. Thus, judicious choice of GAIC allows model selection to meet model selection objectives.

6.3 Lasso-type versus random-effects estimation

Table 6.2 shows that the BIC could select a random-effects model with a smaller PSD if the true unknown $\mathbf{v}_0$ were regarded as a random sample. We used inverse gamma distribution for the u_i s, leading to a heavy-tailed t -type density $g(\psi_i)$ for the random component $\psi_i = v_i$. Exponential u_i leads to lasso estimation in both Bayesian (Park and Casella, 2008) and frequentist approaches (Tibshirani, 1996). Thus, in componentwise model selection, lasso-type estimators can replace $\hat{\psi}_{\text{MHLE}}$ in GAIC. The use of exponentially distributed u_i gives lasso estimation for $\psi_i = \beta_i$, leading to a heavy-tailed distribution $g(\psi_i)$ with an undefined $g'(0)$. This allows the thresholding in estimation, $\hat{\psi}_i = \hat{\beta}_i = 0$. The estimation procedure is

6.3 Lasso-type versus random-effects estimation

Table 2: Simulated results: Entries under M1 to M4 columns are in percentages.

	M1	M2	M3	M4	PSD	M1	M2	M3	M4	PSD	M1	M2	M3	M4	PSD
	$\alpha = 0.01$					$\alpha = 0.05$					$\alpha = 0.1$				
Oracle	0	100	0	0	0.842	0	100	0	0	0.841	0	100	0	0	0.798
BIC	48	0	33	19	0.326	46	0	34	20	0.431	38	0	36	24	0.823
Two-stage	48	0	47	5	0.246	46	0	50	4	0.357	38	0	55	7	0.766
CAIC	22	0	61	17	0.332	18	0	65	17	0.435	14	0	73	23	0.603
	$\alpha = 0.2$					$\alpha = 0.3$					$\alpha = 0.5$				
Oracle	0	100	0	0	0.842	0	100	0	0	0.857	0	100	0	0	0.842
BIC	0	0	82	18	0.775	0	11	71	18	0.824	0	23	65	12	0.853
Two-stage	0	0	77	23	0.743	0	5	74	21	0.822	0	8	71	21	0.841
CAIC	0	0	77	23	0.743	0	5	74	21	0.822	0	8	71	21	0.841
	$\alpha = 1$					$\alpha = 2$					$\alpha = 5$				
Oracle	0	100	0	0	0.889	0	100	0	0	0.798	0	100	0	0	0.889
BIC	0	50	32	17	0.946	0	55	28	17	0.836	0	63	20	17	0.908
Two-stage	0	21	58	21	0.927	0	37	42	21	0.812	0	41	30	29	0.893
CAIC	0	21	58	21	0.927	0	37	42	21	0.812	0	41	30	29	0.893

called oracle when $\hat{\psi}_i = 0$ as if it were known in advance that $\psi_i = 0$. Lee and Oh (2014) generalized the exponential distribution in lasso to the gamma distribution for u_i , leading to $g(0) = \infty$, which enhances the oracle estimation for high-dimensional sparse fixed parameters ψ . Compared with contemporary lasso-type estimating procedures, this gives the best oracle estimation in high-dimensional sparse cases for principal components (Lee et al., 2010), for $\psi = \beta$ (Lee et al. (2011); Lee and Oh (2014); Ng, Lee and Lee (2018); Kaymaz et al. (2026)), and for $\psi = \phi$ (Zeng et al., 2025). Thus, lasso type estimation is effective for estimation of sparse high-dimensional fixed parameters, whereas random-effects estimation is effective

when effects are regarded as a random sample: see Lee et al. (2017) for further discussion.

7. Concluding Remarks

GAIC provides a unifying framework for existing information criteria in model selection and offers a principled way to define the degrees of freedom for random parameters. Each GAIC selects a model that achieves optimal estimation of the parameters of interest ψ while consistently selecting the remaining component ς . The availability of a class of GAICs is therefore advantageous for addressing diverse inferential objectives. GAICs encompass established criteria such as AIC, BIC, CAIC, DIC, and RAIC, while also lead to new criteria SAIC and MIC. Even if the objective is to select a model from a class of marginal candidate models, CAIC and SAIC could recommend random-effects models to improve subject-specific prediction. As the sample size increases, all GAICs tend to converge to a common selected model.

Correlation among covariates in $\mathbf{X}$ could cause a potentially wrong model selection. However, correlation among components leads to a totally different marginal model for a model selection. For example, consider the one-way random-effects model with a heavy-tailed error

$$y_i = \mu + v_i + e_i,$$

where $v_i \sim N(0, \lambda)$, $e_i|u_i \sim N(0, \phi_i)$ and $\phi_i = u_i\phi$ with $E(u_i) = 1$ and $\text{var}(u_i) = \delta$; and v_i, e_i are independent given u_i . This leads to $E(y_i) = \mu$, $\text{var}(y_i) = \lambda + \phi$, $E[(y_i - \mu)^3] = E[(v_i + e_i)^3] = E[v_i^3 + 3v_i^2e_i + 3v_ie_i^2 + e_i^3] = 3E(v_ie_i^2) = 3E[E(v_i|u_i)E(e_i^2|u_i)] = 3\phi E(v_iu_i) = 3\phi\rho$ with $\rho = \text{cov}(v_i, u_i)$. Thus, correlation between the components v_i and ϕ_i results in different skewness in the marginal model through the additional parameter ρ . Investigating the effect

of this correlation on model selection is interesting and will be left for future research.

Conditioning has been used in likelihood approach to eliminate parameters not of interest. Here, conditional likelihood allows us determine whether $\mathbf{v}$ should be treated as random or fixed for improved subject-specific prediction. SAIC combines integration and conditioning within likelihood-based model selection. We have studied consistent model selection of $\boldsymbol{\varsigma}$ and estimation of $\boldsymbol{\psi}$. Lasso-type estimators could replace $\hat{\boldsymbol{\psi}}_{\text{MHLE}}$ within GAICs to enable oracle selection and estimation $\boldsymbol{\psi}$. Future work may explore combining integration, conditioning, and the choice of $\boldsymbol{\psi}$ and its estimation procedures for more general multi-component model selection.

Supplementary Materials

This online supplementary material consists of the following: Proofs of Theorems 1–3, Proposition 2, detailed description of the four real data examples and Table A for the components $\boldsymbol{\psi}$ and $\boldsymbol{\varsigma}$.

Acknowledgements

We thank the Associate Editor and the two anonymous reviewers for their careful reading of the manuscript and for their constructive and helpful comments, which have greatly improved the paper.

References

- Akaike, H. (1973). Information theory as an extension of the maximum likelihood principle. In Second International Symposium on Information Theory, pp. 276–281. (B.N. Petrov and F. Csaki, ed.) Akademiai Kiado, Budapest, 267–281.

REFERENCES

- Birnbaum, A. (1962). On the foundations of statistical inference. Journal of the American Statistical Association 57(298), 269–306.
- Bjørnstad, J. F. (1996). On the generalization of the likelihood function and the likelihood principle. Journal of the American Statistical Association 91(434), 791–806.
- Breslow, N. and D. Clayton (1993). Generalized linear mixed models. Journal of the American Statistical Association 88, 9–25.
- Cox, D. R. and N. Reid (1987). Parameter orthogonality and approximate conditional inference. Journal of the Royal Statistical Society Series B: Statistical Methodology 49(1), 1–39.
- Donohue, M., R. Overholser, R. Xu, and F. Vaida (2011). Conditional akaike information under generalized linear and proportional hazards mixed models. Biometrika 98(3), 685–700.
- Glidden, D. V. and K. Y. Liang (2002). Ascertainment adjustment in complex diseases. Genetic Epidemiology: The Official Publication of the International Genetic Epidemiology Society 23(3), 201–208.
- Ha, I. D. and Y. Lee (2003). Estimating frailty models via poisson hierarchical generalized linear models. Journal of Computational and Graphical Statistics 12(3), 663–681.
- Ha, I. D., Y. Lee, and G. MacKenzie (2007). Model selection for multi-component frailty models. Statistics in Medicine 26(26), 4790–4807.
- Han, J. and Y. Lee (2024). Enhanced Laplace approximation. Journal of Multivariate Analysis 202, 105321.
- Hastie, T. J. and R. J. Tibshirani (1990). Generalized Additive Models, Volume 43 of Monographs on Statistics and Applied Probability. Chapman & Hall.
- Henderson, C. R., O. Kempthorne, S. R. Searle, and C. Von Krosigk (1959). The estimation of environmental and genetic trends from records subject to culling. Biometrics 15(2), 192–218.
- Kaymaz, O., D. Vionanda, F. Dogru, Y. Lee, H. Wood and A. Gusnanto (2026). Sparse logistic regression on genomic data for prediction of tumour pathological subtype. To appear in IEEE/ACM Transactions on Computational

REFERENCES

Biology and Bioinformatics.

Lee, D., W. Lee, Y. Lee and Y. Pawitan (2010). Super sparse principal analysis for high-throughput genomic data.

BMC Bioinformatics, **11**(1), 296.

Lee, D., W. Lee, Y. Lee and Y. Pawitan (2011). Sparse partial least-squares regression and its applications to high-throughput data analysis. Chemometrics and Intelligent Laboratory Systems, **109**(1), 1–8.

Lee, Y. (2001). Can we recover information from concordant pairs in binary matched pairs? Journal of Applied Statistics **28**(2), 239–246.

Lee, Y. (2026). Resolving the induction problem: Can we state with complete confidence that the sun rises forever? To appear in International Statistical Review.

Lee, H. and Y. Lee (2025). Statistical inference for random unknowns via modifications of extended likelihood. *arXiv:2310.09955v3*.

Lee, Y. and J. A. Nelder (1996). Hierarchical generalized linear models. Journal of the Royal Statistical Society Series B: Statistical Methodology **58**(4), 619–656.

Lee, Y. and J. A. Nelder (2001). Hierarchical generalised linear models: a synthesis of generalised linear models, random-effect models and structured dispersions. Biometrika **88**(4), 987–1006.

Lee, Y. and J. A. Nelder (2004). Conditional and marginal models: another view. Statistical Science **19**(2), 219–238.

Lee, Y. and J. A. Nelder (2006a). Double hierarchical generalized linear models (with discussion). Journal of the Royal Statistical Society Series C: Applied Statistics **55**(2), 139–185.

Lee, Y. and J. A. Nelder (2006b). Fitting via alternative random-effect models. Statistics and Computing **16**, 69–75.

Lee, Y., J. A. Nelder, and Y. Pawitan (2017). Generalized linear models with random effects: unified analysis via H-likelihood. Chapman and Hall/CRC.

Lee, Y. and M. Noh (2026). On the use of Gaussian process and deep neural network for semi-parametric Poisson inference. A manuscript submitted for publication.

REFERENCES

- Lee, Y. and H.-S. Oh (2014). A new sparse variable selection via random-effect model. Journal of Multivariate Analysis 125, 89–99.
- Liang, H., H. Wu, and G. Zou (2008). A note on conditional AIC for linear mixed-effects models. Biometrika 95(3), 773–778.
- Meng, X.-L. (2009). Decoding the h-likelihood. Statistical Science 24(3), 280–293.
- Neyman, J. and E. L. Scott (1948). Consistent estimates based on partially consistent observations. Econometrica: Journal of the Econometric Society, 1–32.
- Ng, C. T., W. Lee and Y. Lee (2018). Change-point estimators with true identification property. Bernoulli, **24**(1), 616–660.
- Noh, M. and Y. Lee (2007). Robust modeling for inference from generalized linear model classes. Journal of the American Statistical Association 102(479), 1059–1072.
- Noh, M., Y. Lee, and Y. Pawitan (2005). Robust ascertainment-adjusted parameter estimation. Genetic Epidemiology: The Official Publication of the International Genetic Epidemiology Society 29(1), 68–75.
- Park, T. and G. Casella (2008). The Bayesian Lasso. Journal of the American Statistical Association 103(482), 681–686.
- Patterson, H. D. and R. Thompson (1971). Recovery of inter-block information when block sizes are unequal. Biometrika 58(3), 545–554.
- Schwarz, G. (1978). Estimating the dimension of a model. The Annals of Statistics 6(2), 461–464.
- Smyth, G. K. and A. P. Verbyla (1996). A conditional likelihood approach to residual maximum likelihood estimation in generalized linear models. Journal of the Royal Statistical Society Series B: Statistical Methodology 58(3), 565–572.
- Spiegelhalter, D. J., N. G. Best, B. P. Carlin, and A. Van Der Linde (2002). Bayesian measures of model complexity and fit. Journal of the Royal Statistical Society Series B: Statistical Methodology 64(4), 583–639.

REFERENCES

- Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. Journal of the Royal Statistical Society Series B: Statistical Methodology 58(1), 267–288.
- Vaida, F. and S. Blanchard (2005). Conditional Akaike information for mixed effects models. Biometrika 92, 351–370.
- Yang, Y. (2005). Can the strengths of AIC and BIC be shared? A conflict between model identification and regression estimation. Biometrika 92(4), 937–950.
- Yates, F. (1940). The recovery of inter-block information in balanced incomplete block designs. Annals of Eugenics 10(1), 317–325.
- Yu, D. and K. K. Yau (2012). Conditional akaike information criterion for generalized linear mixed models. Computational Statistics & Data Analysis 56(3), 629–644.
- Zeng, Y., Z. Wang, Y. Lee, and N. Tang (2025). Simultaneous outlier detection and prediction for kriging with true identification. Journal of Computational and Graphical Statistics, 1–12.

Department of Statistics, Seoul National University

E-mail: (youngjo@snu.ac.kr)

Department of Statistics and Data Science, National University of Singapore

E-mail: (choikp@nus.edu.sg)