

Model comparison with composite likelihood information criteria

CHI TIM NG^{1,2} and HARRY JOE³

¹*Department of Statistics, Seoul National University, Room 430, Building 25, Seoul, South Korea. E-mail: *easterlyng@gmail.com*

²*Department of Statistics, Chonnam National University, Gwangju, 500-757, South Korea*

³*Department of Statistics, University of British Columbia, Room ESB 3138, Earth Sciences Building, Vancouver, Canada. E-mail: Harry.Joe@ubc.ca*

Comparisons are made for the amount of agreement of the composite likelihood information criteria and their full likelihood counterparts when making decisions among the fits of different models, and some properties of penalty term for composite likelihood information criteria are obtained. Asymptotic theory is given for the case when a simpler model is nested within a bigger model, and the bigger model approaches the simpler model under a sequence of local alternatives. Composite likelihood can more or less frequently choose the bigger model, depending on the direction of local alternatives; in the former case, composite likelihood has more “power” to choose the bigger model. The behaviors of the information criteria are illustrated via theory and simulation examples of the Gaussian linear mixed-effects model.

Keywords: Akaike information criterion; Bayesian information criterion; local alternatives; mixed-effects model; model comparison

1. Introduction

Composite likelihood inference based on low-dimensional marginal or conditional distributions is common when the full likelihood is computationally too difficult. It has been increasingly used in recent years for inference with complex models; see Varin [13], Varin *et al.* [14] for reviews.

For model selection with composite likelihood, one might wonder if the use of limited or reduced information leads to different decisions. To understand this, an asymptotic theory based on the theory of a sequence of contiguous local alternatives is developed to compare Akaike information criterion (AIC) and Bayesian information criterion (BIC) in their full likelihood and composite marginal likelihood versions. We show that model selection based on AIC and its composite likelihood counterpart (as proposed in Varin and Vidoni [15]) are sometimes similar (when models under consideration are far apart) and sometimes not similar (when one model is a perturbation of another). The patterns can be explained via local alternatives where the perturbed model is at a distance $n^{-1/2}$ from a “null” or simplified model, with n being the sample size.

We also provide simulation results under models where the maximum likelihood is feasible; one class of such models is the linear mixed-effects models based on the normal distribution. Within different sub-cases of the Gaussian linear mixed-effects models, the simulation results are consistent with the asymptotic theory.

The remainder of the paper is organized as follows. In Section 2, we introduce our notation and state the definitions for the composite marginal likelihood and the information criteria. In

Section 3, asymptotic properties of composite likelihood information criteria are presented. In Section 4, comparisons of decisions between Varin and Vidoni’s composite likelihood information criterion (abbreviated CLAIC as in Varin *et al.* [14]), Gao and Song’s information criterion (abbreviated as CLBIC in Gao and Song [4]), and their full-likelihood counterparts are summarized via simulation studies. Section 5 contains a data example with a mixed-effects model. Section 6 concludes with some discussion and future research. The proofs of the main theorems in Section 3 are given in Appendix A.1.

2. Composite likelihood and information criteria

For the comparison of composite likelihood and full likelihood information criteria, we consider the case of independent multivariate measurements on n subjects, possibly with covariates. Nested statistical models will be considered.

2.1. Model

Let $\mathbf{y}_1, \dots, \mathbf{y}_n$ be the realizations of independent d -dimensional random vectors \mathbf{Y}_i , with respective covariates summarized as matrices $\mathbf{x}_1, \dots, \mathbf{x}_n$. Suppose that the data generating mechanism of \mathbf{Y}_i is governed by the density function $g(\mathbf{y}_i; \mathbf{x}_i)$. Candidate parametric models are $f^{(M)}(\mathbf{y}_i; \mathbf{x}_i, \boldsymbol{\theta}^{(M)})$, for $M = 1, 2, \dots$; M is an index for different models that are considered, and $\boldsymbol{\theta}^{(M)}$ is the parameter vector for model M . Let $p_M = \dim(\boldsymbol{\theta}^{(M)})$ be the dimension of $\boldsymbol{\theta}^{(M)}$ for a generic model M ; the superscript will be omitted unless we are referring to two or more models.

2.2. Composite likelihood

For model M , let $L_{\text{CL}}^{(M)}(\boldsymbol{\theta}^{(M)}) = L_{\text{CL}}^{(M)}(\boldsymbol{\theta}^{(M)}; \mathbf{y}_1, \dots, \mathbf{y}_n; \mathbf{x}_1, \dots, \mathbf{x}_n)$ be a particular composite marginal log-likelihood. We are using the same composite likelihood (same set of marginal density functions) for all competing models. Let $S \subset \{1, 2, \dots, d\}$ be a non-empty subset of indexes. For notation, $f_S^{(M)}$ indicates a marginal density of $f^{(M)}$ with margin S and g_S is the corresponding margin of g . The particular composite likelihood could be based on all bivariate margins, or a subset of bivariate margins, or more generally a set of margins $\{S_1, \dots, S_Q\}$ with corresponding weights w_1, \dots, w_Q . Suppressing the superscript for the model, let

$$L_{\text{CL}}(\boldsymbol{\theta}) = L_{\text{CL}}(\boldsymbol{\theta}; \mathbf{y}_1, \dots, \mathbf{y}_n; \mathbf{x}_1, \dots, \mathbf{x}_n) = \sum_{i=1}^n \ell_{\text{CL}}(\boldsymbol{\theta}; \mathbf{y}_i, \mathbf{x}_i) \tag{2.1}$$

be the log composite likelihood. Here

$$\exp\{\ell_{\text{CL}}(\boldsymbol{\theta}; \mathbf{y}_i, \mathbf{x}_i)\} = \prod_{q=1}^Q f_{S_q}^{w_q}(\mathbf{y}_{i,S_q}; \mathbf{x}_i, \boldsymbol{\theta}), \tag{2.2}$$

S_q is a subset consisting of indexes, and w_q is a positive weight for S_q . For example, if these are the pairs for bivariate composite likelihood, then the cardinality of $\{S_q\}$ is $Q = d(d - 1)/2$. Note that the case of full likelihood is covered with $S_1 = \{1, \dots, d\}$ with the cardinality of $\{S_q\}$ being 1.

2.3. Composite likelihood information criteria

Consider the composite likelihood versions of Akaike information criterion (AIC) and Bayesian information criterion (BIC) described in Varin and Vidoni [15], Gao and Song [4], Varin *et al.* [14]. They are defined as (with superscript for model M omitted):

$$\text{CLAIC} = -2L_{\text{CL}}(\hat{\theta}_{\text{CL}}) + 2\text{tr}\{\mathbf{J}(\hat{\theta}_{\text{CL}})\mathbf{H}^{-1}(\hat{\theta}_{\text{CL}})\} \tag{2.3}$$

and

$$\text{CLBIC} = -2L_{\text{CL}}(\hat{\theta}_{\text{CL}}) + (\log n) \text{tr}\{\mathbf{J}(\hat{\theta}_{\text{CL}})\mathbf{H}^{-1}(\hat{\theta}_{\text{CL}})\}. \tag{2.4}$$

Here, $\hat{\theta}_{\text{CL}} = \hat{\theta}_{n,\text{CL}}$ is the composite likelihood estimator that maximizes (2.1). The matrices $\mathbf{H}(\theta)$ and $\mathbf{J}(\theta)$ are the Hessian matrix and the covariance matrix of the score function, respectively,

$$\mathbf{H}(\theta) = - \lim_{n \rightarrow \infty} n^{-1} \frac{\partial^2 L_{\text{CL}}(\theta; \mathbf{y}_1, \dots, \mathbf{y}_n, \mathbf{x}_1, \dots, \mathbf{x}_n)}{\partial \theta \partial \theta^T}$$

and

$$\mathbf{J}(\theta) = \text{Cov} \left[n^{-1/2} \frac{\partial L_{\text{CL}}(\theta; \mathbf{y}_1, \dots, \mathbf{y}_n, \mathbf{x}_1, \dots, \mathbf{x}_n)}{\partial \theta} \right].$$

When there are several models, the CLAIC (CLBIC) principle selects the model with smallest value of CLAIC (CLBIC). CLAIC has penalty term $2 \text{tr}(\mathbf{JH}^{-1})$ and CLBIC has penalty term $(\log n) \text{tr}(\mathbf{JH}^{-1})$ that depends on the sample size n . With large n , CLBIC might choose smaller models than CLAIC.

3. Main theorems

The main results are presented in this section, with proofs in the [Appendix](#). Consider the nested cases where model 1 is nested within model 2. Proposition 3.1 gives general results of the composite likelihood ratio under nested cases. If the true model is covered by either model 1 or model 2, Theorem 3.1 provides further comparison of the asymptotic properties of CLAIC and CLBIC under a sequence of local alternative hypotheses. Results under model misspecification are summarized in Theorem 3.2.

To describe the theorems, the following notation is used,

- Model 1: $\mathbf{Y}|\mathbf{x} \sim f^{(1)}(\mathbf{y}; \mathbf{x}, \theta), \theta \in \Theta$.
- Model 2: $\mathbf{Y}|\mathbf{x} \sim f^{(2)}(\mathbf{y}; \mathbf{x}, \gamma), \gamma \in \Gamma$.
- True model: $\mathbf{Y}|\mathbf{x} \sim g(\mathbf{y}; \mathbf{x})$.

This notation matches $\theta^{(1)} = \theta$ and $\theta^{(2)} = \gamma$, as used in Section 2, but we are temporarily reducing the number of superscripts. Let θ^* be the parameters for $f^{(1)}(\cdot; \theta)$ such that $f^{(1)}$ is the closest to g in the divergence (see Xu and Reid [17]) based on the composite log-likelihood function $L_{\text{CL}}^{(1)}$. Similarly γ^* is defined. Note that θ^* and γ^* might depend on the composite log-likelihood that is used.

Proposition 3.1 (Asymptotic distribution of the composite likelihood ratio). *Consider the log composite likelihood ratio of two competing models,*

$$\text{LR} = L_{\text{CL}}^{(2)}(\hat{\gamma}) - L_{\text{CL}}^{(1)}(\hat{\theta}). \tag{3.1}$$

Suppose that assumptions A1–A3 (given in Appendix A) hold. If for all (\mathbf{x}, \mathbf{y}) ,

$$f^{(1)}(\mathbf{y}; \mathbf{x}, \theta^*) = f^{(2)}(\mathbf{y}; \mathbf{x}, \gamma^*), \tag{3.2}$$

then the limiting distribution of 2LR has the same law as $\mathbf{Z}^T D \mathbf{Z}$, where \mathbf{Z} is a vector of independent standard normal random variables and D is a diagonal matrix with eigenvalues of the matrix:

$$\mathbf{B} = \begin{pmatrix} -(\mathbf{J}^{(11)})(\mathbf{H}^{(1)})^{-1} & (\mathbf{J}^{(12)})(\mathbf{H}^{(2)})^{-1} \\ -(\mathbf{J}^{(21)})(\mathbf{H}^{(1)})^{-1} & (\mathbf{J}^{(22)})(\mathbf{H}^{(2)})^{-1} \end{pmatrix}. \tag{3.3}$$

Here, $\mathbf{H}^{(1)}$, $\mathbf{J}^{(12)}$, etc., are defined in Appendix A.

In order to understand how different criteria can differ, we do an analysis for a sequence of contiguous alternatives, in which the true model is model 2 and its parameter depends on the sample size n and is closer to the null model as n increases. Such theory helps to explain what happens in finite samples; see Section 4. Suppose that model 2 is $f^{(2)}(\cdot; \theta, \zeta)$ and model 1 (null model) is nested within model 2, that is, $f^{(1)}(\cdot; \theta) = f^{(2)}(\cdot; \theta, \mathbf{0})$. The local alternatives assumption refers to that $g(\cdot) = f^{(2)}(\cdot; \theta_{2n}^*, \zeta_n^*)$ with $\zeta_n^* = a_n \boldsymbol{\epsilon}$ converges to $\zeta^* = \mathbf{0}$ at rate $a_n = n^{-1/2}$ or $a_n = \sqrt{\log n/n}$, and $\theta_{2n}^* \rightarrow \theta^*$. Let θ_{1n}^* be the parameter for $f^{(1)}(\cdot; \theta)$ such that $f^{(1)}$ is closest to g in the divergence (see Xu and Reid [17]) based on the composite log-likelihood function $L_{\text{CL}}^{(1)}$. Assume that θ_{1n}^* and θ_{2n}^* are asymptotically equivalent, that is,

$$\theta_{1n}^* - \theta_{2n}^* \rightarrow \mathbf{0}, \quad n \rightarrow \infty. \tag{3.4}$$

We next state the main theorem for comparing CLAIC, AIC, CLBIC for nested models, when the null model is true, or when the larger model is true under a sequence of local alternatives.

Theorem 3.1. *Consider the model selection problem H_1 : Model 1 $f^{(1)}(\cdot; \theta)$ is the true model versus H_2 : Model 2 $f^{(2)}(\cdot; \theta, \zeta)$ is the true model. Here, γ is p_2 -dimensional and ζ is m -dimensional, where $m = p_2 - p_1$. Let P_1^{AIC} be the probability that AIC selects model 1. Similar notation is used for BIC, CLAIC, and CLBIC.*

- (1) Under H_1 , $P_1^{\text{CLAIC}} \rightarrow C_1 \in (0, 1)$ and $P_1^{\text{CLBIC}} \rightarrow 1$.

- (2) Under H_1 , $P_1^{\text{CLAIC}} < P_1^{\text{AIC}}$.
- (3) Under H_2 with $\zeta = \zeta_n^* = \epsilon n^{-1/2}$ and $\epsilon = O(1)$, and assuming (3.4), $P_1^{\text{CLAIC}} \rightarrow C_2 \in (0, 1)$ and $P_1^{\text{CLBIC}} \rightarrow 1$.
- (4) Under H_2 with $\zeta = \zeta_n^* = \epsilon \sqrt{\log n/n}$ and $\epsilon = O(1)$, and assuming (3.4), $P_1^{\text{CLAIC}} \rightarrow 0$ and $P_1^{\text{CLBIC}} \rightarrow C_3 \in (0, 1)$.

To be more specific, we have $C_1 = P(\lambda_1 U_1 + \dots + \lambda_m U_m < 2(\lambda_1 + \dots + \lambda_m))$, where U_1, U_2, \dots, U_m are independent χ_1^2 random variables and $\lambda_1, \lambda_2, \dots, \lambda_m$ are the non-zero eigenvalues of \mathbf{B} defined in (3.3). If the full-likelihood is used, $\lambda_1 = \dots = \lambda_m = 1$.

In Theorem 3.1, (1) is a special case of (3) with $\epsilon = \mathbf{0}$. The asymptotic results (1) and (3) are natural. Intuitively speaking, if less parameters than the true model are selected, the composite likelihood decreases by a positive quantity of $O(n)$. Such a decrease dominates the CLAIC (CLBIC) penalty term so the penalty term is ignorable. This guarantees that the true model is better than the smaller models in terms of CLAIC (CLBIC). On the other hand, if more parameters are involved than necessary, the increase in composite likelihood is just $O(1)$. For CLAIC, the change in penalty term is also $O(1)$, so the model is correctly selected only with some positive probability. For CLBIC, provided that the penalty term is monotonic (see Lemma A.2), it is guaranteed that the change in penalty term is positive and is $O(\log n)$, dominating the increase in composite likelihood. Then, the true model is better than any other bigger model.

If model 2 is the true model and the two models are sufficiently far apart from each other, that is, $\zeta = O(1) \neq \mathbf{0}$, then all the criteria asymptotically choose the correct model. On the contrary, if the two models differ by only a small perturbation, for example, $\zeta = O(1/\sqrt{n})$ or $\zeta = O(\sqrt{\log n/n})$, it can be seen from results (3) and (4) that the behavior of CLAIC and CLBIC differ. CLBIC is less likely select the correct model than CLAIC.

Comparing CLAIC and its full-likelihood counterpart, CLAIC has greater probability of selecting the larger model. The difference in such probabilities depends on the eigenvalues $\lambda_1, \dots, \lambda_m$. Roughly speaking, if $(\lambda_1, \dots, \lambda_m)$, after standardization is closer to $(1, \dots, 1)$, the “loss of information” due to the use of composite likelihood is less significant. It is natural to consider C_1 in Theorem 3.1 as a measurement of closeness of the composite likelihood to the full-likelihood. It is interesting to note that C_1 does not depend on the parameters for full-likelihood. For composite likelihood, it is possible that C_1 depends on the parameters through $\lambda_1, \dots, \lambda_m$. The dependence of C_1 on the parameters will be illustrated via simulation examples in Section 4.

Part of the results in Theorem 3.1 can be generalized to the situation of model misspecification.

Theorem 3.2 (The same notation as in Proposition 3.1 and Theorem 3.1 is used). *Suppose that model 1 is nested within model 2 but neither model 1 nor model 2 is the true model. Let $(\theta_{2n}^*, \zeta_n^*)$ be the parameter under model 2 that is the closest to the true model in the divergence (see Xu and Reid [17]) based on the composite likelihood.*

- If equation (3.2) holds, (1) $P_1^{\text{CLAIC}} \rightarrow C_1 \in (0, 1)$ and $P_1^{\text{CLBIC}} \rightarrow 1$.*
- If equation (3.2) does not hold,*
- (2) *If $\zeta_n^* = O(1)$, then $P_1^{\text{CLAIC}} \rightarrow 0$ and $P_1^{\text{CLBIC}} \rightarrow 0$.*
- (3) *If $\zeta_n^* = \epsilon n^{-1/2}$ and $\epsilon = O(1)$, and assuming (3.4), $P_1^{\text{CLAIC}} \rightarrow C_2 \in (0, 1)$ and $P_1^{\text{CLBIC}} \rightarrow 1$.*

(4) If $\zeta_n^* = \varepsilon \sqrt{\log n/n}$ and $\varepsilon = O(1)$, and assuming (3.4), $P_1^{\text{CLAIC}} \rightarrow 0$ and $P_1^{\text{CLBIC}} \rightarrow C_3 \in (0, 1)$.

In the model misspecification cases, it is more difficult to compare analytically the probabilities of selecting model 1 for AIC and CLAIC. To compare AIC and CLAIC, simulation examples are provided in Section 4.

4. Simulation studies

In this section, we show simulation results of the following comparisons in their decisions among competing models,

1. CLAIC versus CLBIC,
2. CLAIC versus AIC,
3. CLBIC versus BIC.

To do this, we choose models where the maximum likelihood estimators are also computationally feasible. The analysis is different from that in Gao and Song [4] in that our concern is not in whether the correct model is asymptotically chosen with probability 1. If models being compared are close to each other, then any of the models could be chosen with positive probability, and we are interested in where CLAIC and AIC might differ.

One general model that allows a variety of univariate and dependence parameters is the mixed-effects model (see Laird and Ware [7]); it is defined via:

$$\begin{aligned} \mathbf{Y}_i &= \mathbf{x}_i \boldsymbol{\beta} + \mathbf{z}_i \mathbf{b}_i + \boldsymbol{\varepsilon}_i, & i = 1, 2, \dots, n, \\ \mathbf{b}_i &\sim N(\mathbf{0}, \Psi), & \boldsymbol{\varepsilon}_i \sim N(\mathbf{0}, \phi \mathbf{I}_d), \end{aligned}$$

where $\boldsymbol{\beta}$ is $(s + 1)$ -dimensional vector of fixed effects, \mathbf{b}_i is r -dimensional vector of random effects. \mathbf{x}_i and \mathbf{z}_i are $d \times (s + 1)$ and $d \times r$ observable matrices, \mathbf{x}_i has a first column of 1s, ϕ is a variance parameter, Ψ is a $r \times r$ covariance matrix. Both full likelihood and composite likelihood of the mixed-effects model can be expressed explicitly with the matrix algebra notation (see, e.g., Fackler [3], Magnus and Neudecker [8]). This model leads to closed form expressions where \mathbf{H} and \mathbf{J} can be computed (see Appendix B).

A special case is the clustered data model with exchangeable dependence structure. It is defined by setting $\mathbf{z}_i = (1, 1, \dots, 1)^T$, $\Psi = \sigma^2 \rho$, and $\phi = \sigma^2(1 - \rho)$, and closed forms for \mathbf{H} and \mathbf{J} can be found in Joe and Lee [5].

The three examples given below are representative cases to show patterns in the decisions from various criteria and in the penalty term $\text{tr}(\mathbf{J}\mathbf{H}^{-1})$; the patterns were seen over different parameter settings and dimension d . In the following examples, the composite likelihood corresponding to the pairwise likelihood or bivariate composite likelihood (BCL) is specified via

$$S_q = \{(i, j) \text{ for all } i < j\}.$$

In Example 2, trivariate composite likelihood (TCL) is also used. The sets S_q for defining TCL are

$$S_q = \{(i, j, k) \text{ for all } i < j < k\}.$$

Table 1. Comparison of decisions for AIC versus CLAIC for different β vectors, and distribution of $\text{tr}(\mathbf{JH}^{-1})$. Cluster size $d = 4$; $\beta_0 = (0.3, 1.3, 0.00, 0.00)$, $\beta_1 = (0.3, 1.3, 0.05, 0.02)$, $\beta_2 = (0.3, 1.3, 0.15, 0.05)$, $\beta_3 = (0.3, 1.3, 0.15, 0.10)$. Covariates are drawn from $N(0, \mathbf{I})$

CLAIC\AIC	β_0			β_1			β_2			β_3		
	1	2	3	1	2	3	1	2	3	1	2	3
$n_c = 1$	472	39	11	439	37	19	5	4	1	1	0	1
$n_c = 2$	23	335	11	24	309	16	1	500	50	1	112	25
$n_c = 3$	16	5	88	6	13	137	0	33	406	0	13	847

#covariates	Lower quartile Q1 to upper quartile Q3 of $\text{tr}(\mathbf{JH}^{-1})$							
	β_0		β_1		β_2		β_3	
	Q1	Q3	Q1	Q3	Q1	Q3	Q1	Q3
1	13.7	14.1	13.7	14.1	13.6	14.0	13.6	14.0
2	16.4	16.7	16.4	16.7	16.4	16.7	16.3	16.7
3	19.1	19.3	19.1	19.3	19.1	19.3	19.1	19.3

In order that decisions based on AIC and CLAIC are not always for one model, parameters are chosen appropriately so that the simpler model has some chance to be chosen. In Example 1, we consider smaller beta versus larger beta values.

Example 1 (Cluster model with exchangeable covariance matrix, regression vector β at varying distance from 0). The true number of covariates is 3. Let $\beta_0 = (0.3, 1.3, 0.00, 0.00)$, $\beta_1 = (0.3, 1.3, 0.05, 0.02)$, $\beta_2 = (0.3, 1.3, 0.15, 0.05)$, and $\beta_3 = (0.3, 1.3, 0.15, 0.10)$, with first element of the β vectors being the intercept. Because the last two parameters (regression coefficients for second and third covariates are smaller), for model selection, simpler models without the additional covariates might be chosen for any information criteria. The parameters $\sigma^2 = 1$ and $\rho = 0.5$ are fixed.

For each of the four β vectors, 1000 replicates with sample size $n = 100$ and cluster size $d = 4$ are generated. Three different settings are used to simulate the covariates and the random effects. In settings (i) and (ii), the covariates $\mathbf{x}_i = (\mathbf{x}_{i1}, \mathbf{x}_{i2}, \mathbf{x}_{i3})^T$ are independent random vectors from $N(0, \Sigma_X)$ with $\Sigma_X = \mathbf{I}$, the identity matrix and $\Sigma_X = 0.2\mathbf{I} + 0.8\mathbf{1}\mathbf{1}^T$, respectively. The random effect \mathbf{b}_i is obtained from normal distribution. In setting (iii), t -distribution with degree of freedom 3 is used for \mathbf{b}_i instead so that the robustness of the information criteria under model misspecification can therefore be investigated. That is, $\mathbf{b}_i = \rho^{1/2}t_i$, where t_i are independent t -distributed random variables. We then compare the decisions of AIC and CLAIC for regression models with the first, the first two or all three covariates ($n_c = 1, 2$ or 3). For setting (i), summaries in Table 1 show patterns in the decisions and in the amount of variation in the CLAIC penalty term $\text{tr}(\mathbf{JH}^{-1})$. As an example, for β_1 , there were 137 cases where both AIC and CLAIC chose the 3-covariate model. Table 1 shows that the decisions for CLAIC are the same as with AIC in a high proportion of cases; both tend to choose a regression model with more covariates

Table 2. Comparison of decisions for AIC versus CLAIC for different β vectors, and distribution of $\text{tr}(\mathbf{JH}^{-1})$. Cluster size $d = 4$; $\beta_0 = (0.3, 1.3, 0.00, 0.00)$, $\beta_1 = (0.3, 1.3, 0.05, 0.02)$, $\beta_2 = (0.3, 1.3, 0.15, 0.05)$, $\beta_3 = (0.3, 1.3, 0.15, 0.10)$. Covariates are drawn from $N(0, 0.2\mathbf{I} + 0.8\mathbf{11}^T)$

CLAIC\AIC	β_0			β_1			β_2			β_3		
	1	2	3	1	2	3	1	2	3	1	2	3
$n_c = 1$	608	43	16	567	41	18	77	24	5	33	9	5
$n_c = 2$	28	205	7	18	234	8	2	617	37	3	441	39
$n_c = 3$	15	2	76	11	3	100	3	28	207	1	29	440

#covariates	Lower quartile Q1 to upper quartile Q3 of $\text{tr}(\mathbf{JH}^{-1})$							
	β_0		β_1		β_2		β_3	
	Q1	Q3	Q1	Q3	Q1	Q3	Q1	Q3
1	13.7	14.1	13.7	14.1	13.6	14.0	13.6	14.0
2	16.4	16.7	16.4	16.7	16.4	16.7	16.4	16.7
3	19.1	19.3	19.1	19.3	19.1	19.3	19.1	19.3

if the true β vector has more coefficients farther from 0. The results of BIC and CLBIC are similar, and are not shown. The variation in $\text{tr}(\mathbf{JH}^{-1})$ is not too much when the sample size is large enough. As implied by Lemma A.2, $\text{tr}(\mathbf{JH}^{-1})$ tends to increase for models with additional parameters. Similar results of settings (ii) and (iii) are given in Tables 2 and 3, respectively. In this example, it can be seen that all information criteria have a higher chance to select the smaller model in the presence of strong correlations (say, 0.8) in the covariates. In the case where the distribution of ε_i is misspecified, the decisions from all information criteria are very similar to the counterpart without misspecification.

Example 2 (Multivariate normal regression model, different covariance structures). This example shows local alternatives or perturbations of different types, either in univariate or in dependence parameters. We compare exchangeable (exch) versus unstructured (unstr) dependence when true covariance matrix has different deviations from exchangeable. The choices of the true covariance matrices are:

$$\Sigma_1 = \begin{pmatrix} 1 & 0.5 & 0.5 & 0.5 \\ 0.5 & 1 & 0.5 & 0.5 \\ 0.5 & 0.5 & 1 & 0.5 \\ 0.5 & 0.5 & 0.5 & 1 \end{pmatrix},$$

$$\Sigma_2 = \begin{pmatrix} 1 & 0.5 + \varepsilon_1/\sqrt{n} & 0.5 & 0.5 \\ 0.5 + \varepsilon_1/\sqrt{n} & 1 & 0.5 & 0.5 \\ 0.5 & 0.5 & 1 & 0.5 + \varepsilon_1/\sqrt{n} \\ 0.5 & 0.5 & 0.5 + \varepsilon_1/\sqrt{n} & 1 \end{pmatrix},$$

Table 3. Comparison of decisions for AIC versus CLAIC for different β vectors, and distribution of $\text{tr}(\mathbf{JH}^{-1})$. Cluster size $d = 4$; $\beta_0 = (0.3, 1.3, 0.00, 0.00)$, $\beta_1 = (0.3, 1.3, 0.05, 0.02)$, $\beta_2 = (0.3, 1.3, 0.15, 0.05)$, $\beta_3 = (0.3, 1.3, 0.15, 0.10)$. Covariates are drawn from multivariate t distribution with $\Sigma_X = \mathbf{I}$

CLAIC\AIC	β_0			β_1			β_2			β_3		
	1	2	3	1	2	3	1	2	3	1	2	3
$n_c = 1$	484	43	8	449	42	8	3	5	1	1	0	1
$n_c = 2$	28	314	13	21	311	17	1	539	43	1	129	33
$n_c = 3$	12	16	82	18	9	125	0	32	376	0	17	818

#covariates	Lower quartile Q1 to upper quartile Q3 of $\text{tr}(\mathbf{JH}^{-1})$							
	β_0		β_1		β_2		β_3	
	Q1	Q3	Q1	Q3	Q1	Q3	Q1	Q3
1	14.8	15.2	14.8	15.2	14.7	15.1	14.7	15.1
2	17.2	17.5	17.2	17.5	17.2	17.5	17.2	17.5
3	19.7	19.8	19.7	19.8	19.7	19.8	19.7	19.8

$$\Sigma_{ka} = \text{diag}(1, 1, 1 + \varepsilon_2/\sqrt{n}, 1 + \varepsilon_2/\sqrt{n})\Sigma_k \text{diag}(1, 1, 1 + \varepsilon_2/\sqrt{n}, 1 + \varepsilon_2/\sqrt{n})$$

for $k = 1, 2$, where $\varepsilon_1 = 0.07\sqrt{200}$ and $\varepsilon_2 = 0.05\sqrt{200}$. Σ_2 changes some correlation parameters, Σ_{1a} changes some variance (univariate) parameters, and Σ_{2a} changes both correlation and variance parameters. The regression vector $\beta = (0.3, 1.3)$ is fixed and the covariates x_i are independent standard normal random variables. Summaries in Table 4 are from 1000 replicates with different sample sizes n and cluster size $d = 4$.

The patterns are similar to above for larger cluster size $d = 5, 6, 7$ and perturbations of a different exchangeable correlation matrix. That is, CLAIC tends to more often than AIC choose the unstructured dependence when the perturbation is only in the variances (i.e., Σ_{1a}), and AIC tends to more often than CLAIC choose the unstructured dependence when the perturbation is only in the correlations (i.e., Σ_2). For perturbations in the correlations, going to trivariate composite likelihood makes CLAIC closer to AIC in the decision between the two models.

For Σ_1 , CLAIC selects bigger model more often than AIC in all three settings (see Table 4). However, the probabilities P_1^{CLAIC} and P_1^{AIC} are very close to each other. In this example, CLAIC and AIC give very similar decisions. The outcome is consistent with Theorem 3.1(2). Under H_1 , AIC selects model 1 with probability approximately $\Pr(Z_1^2 + \dots + Z_m^2 < 2m)$. For the TCL with $S_q = \{(i, j, k)\}$, $n = 500$, $\Sigma = \Sigma_1$, CLAIC selects model 1 with probability approximately $\Pr(\lambda_1 Z_1^2 + \dots + \lambda_m Z_m^2 < 2(\lambda_1 + \dots + \lambda_m))$. Here, $\lambda_1, \dots, \lambda_m$ are

3.34, 2.87, 2.73, 2.52, 2.07, 2.03, 1.61, 1.50.

Table 4. Comparison of decisions for AIC versus CLAIC under different perturbations of the exchangeable dependence model

CLAIC\AIC	Σ_1		Σ_{1a}		Σ_2		Σ_{2a}	
	exch.	unstr.	exch.	unstr.	exch.	unstr.	exch.	unstr.
<i>n</i> = 200, <i>d</i> = 4, BCL								
exch.	919	16	813	15	668	211	574	162
unstr.	40	25	95	77	21	100	45	219
<i>n</i> = 500, <i>d</i> = 4, BCL								
exch.	911	12	825	10	699	175	593	168
unstr.	50	27	108	57	18	108	42	197
<i>n</i> = 500, <i>d</i> = 4, TCL								
exch.	944	6	890	6	710	94	617	84
unstr.	17	33	43	61	7	189	18	281

Since these λ values differ from each other, Lemma A.1(2) guarantees that

$$\Pr(\lambda_1 Z_1^2 + \dots + \lambda_m Z_m^2 < 2(\lambda_1 + \dots + \lambda_m)) < \Pr(Z_1^2 + \dots + Z_m^2 < 2m).$$

Indeed, for the eigenvalues λ in the example, we have

$$\Pr(\lambda_1 Z_1^2 + \dots + \lambda_8 Z_8^2 > 2(\lambda_1 + \dots + \lambda_8)) = 0.0468 \quad \text{and} \quad \Pr(Z_1^2 + \dots + Z_8^2 > 16) = 0.0424.$$

Here, the numerical method proposed in Rice [11] is used to obtain the first probability. The first probability is slightly greater than the second probability.

Example 3 (Multivariate normal regression model, different covariance structures). This example shows the exchangeable (exch) dependence model and its local alternatives with perturbations of different sizes in dependence parameters. Information criteria AIC, BIC, CLAIC, and CLBIC are compared. The choices of the true covariance matrices are:

$$\Sigma(\delta) = \begin{pmatrix} 1 & 0.5 + \delta & 0.5 & 0.5 \\ 0.5 + \delta & 1 & 0.5 & 0.5 \\ 0.5 & 0.5 & 1 & 0.5 + \delta \\ 0.5 & 0.5 & 0.5 + \delta & 1 \end{pmatrix}.$$

Define $\Sigma_1 = \Sigma(0)$, $\Sigma_2 = \Sigma(n^{-1/2})$, $\Sigma_3 = \Sigma(n^{-1/2} \log n)$, and $\Sigma_4 = \Sigma(0.2)$. The regression vector $\beta = (0.3, 1.3)$ is fixed and the covariates \mathbf{x}_i are independent standard normal random variables.

Summaries in Table 5 are from 1000 replicates sample size $n = 500$ and cluster size $d = 4$. The frequencies of selecting the exchangeable dependence model are reported. We see that BIC/CLBIC tends to select the exchangeable dependence model more often than AIC/CLAIC. Under the assumption of exchangeable dependence model, BIC/CLBIC have greater chance of selecting the correct model. However, BIC/CLBIC are less sensitive to small perturbations than AIC/CLAIC. The results are consistent with Theorem 3.1.

Table 5. Comparison of decisions for AIC, BIC, CLAIC, and CLBIC under different perturbations of the exchangeable dependence model. Sample size $n = 500$

Info. crit.	Frequency of selecting exchangeable			
	Σ_1	Σ_2	Σ_3	Σ_4
AIC	961	712	5	0
CLAIC	950	803	28	0
BIC	1000	1000	705	0
CLBIC	1000	1000	927	0

Example 4 (Comparison of information criteria under model misspecification). To see the effect of model misspecification, we repeat Example 3 with the following changes: (i) $\mathbf{b}_i = \mathbf{C}\mathbf{u}_i$, where $\Psi = \mathbf{C}\mathbf{C}^T$ is the Cholesky decomposition and \mathbf{u}_i are vectors of independent Laplace random variables with mean zero and variance one. (ii) \mathbf{b}_i is generated from normal distribution but

$$\mathbf{Y}_i = (0.3, 0.6, 0.9, 1.2)^T + \mathbf{x}_i\boldsymbol{\beta} + \mathbf{z}_i\mathbf{b}_i + \boldsymbol{\varepsilon}_i, \quad i = 1, 2, \dots, n.$$

The results under (i) and (ii) are summarized in Tables 6 and 7, respectively. The decisions under (i) are comparable to (1), (3), (4) in Theorem 3.2. The decisions under (ii) are similar to that described in (2) in Theorem 3.2. Comparing with Example 3, under both (i) and (ii), the alternative model is more likely to be selected.

5. Spruce tree growth data

In this section, we study the spruce tree growth data in Example 1.3 in Diggle *et al.* [2]. The decisions from AIC (BIC) and their composite likelihood counterparts are compared.

The dataset consists of the data from $n = 79$ trees and is available in the R package MEMSS (Pinheiro and Bates [10]). For each tree, the logarithm of the volume of the tree trunk was estimated and recorded in $d = 13$ chosen days t_1, t_2, \dots, t_{13} from the beginning of the experiment

Table 6. Comparison of decisions for AIC, BIC, CLAIC, and CLBIC under perturbation in the distribution law. Sample size $n = 500$

Info. crit.	Frequency of selecting exchangeable			
	Σ_1	Σ_2	Σ_3	Σ_4
AIC	768	479	3	0
CLAIC	712	537	12	0
BIC	1000	998	575	0
CLBIC	1000	1000	795	0

Table 7. Comparison of decisions for AIC, BIC, CLAIC, and CLBIC under perturbation in the mean. Sample size $n = 500$

Info. crit.	Frequency of selecting exchangeable			
	Σ_1	Σ_2	Σ_3	Σ_4
AIC	0	0	0	0
CLAIC	0	0	0	0
BIC	24	0	0	0
CLBIC	68	3	0	0

over a period of 674 days. The trees were grown in four different plots, labeled 1, 2, 3, 4, respectively. The days are 152, 174, 201, 227, 258, 469, 496, 528, 556, 579, 613, 639, 674 days since 1988-01-01, corresponding to roughly beginning of June to mid-August in 1988 and mid-April to the end of October in 1989. The first two plots represent an ozone-controlled atmosphere and the last two plots represent a normal atmosphere. From the plots in Diggle *et al.* [2], the growth rates in the two time periods are different.

A linear mixed-effects model accounts for different growth rates in the two periods is the following. For a given tree, with $y = \log$ size has growth and $t = \text{day since 1988-01-01}$,

$$y_{ij} = a_0 + a_1(t_j - 152)/100 + \varepsilon_i, \quad 152 \leq t_j \leq 258,$$

$$y_{ij} = [a_0 + a_1(258 - 152)/100] + a_2(t_j - 445)/100 + \varepsilon_i, \quad 469 \leq t_j \leq 674.$$

To introduce fixed and random effects, $a_0 = \beta_0 + \beta_3 I(\text{ozone}) + b_0$, where b_0 is random with normal distribution; in addition, $a_1 = \beta_1 + \beta_4 I(\text{ozone}) + b_1$, $a_2 = \beta_2 + \beta_5 I(\text{ozone}) + b_2$, where b_1, b_2 are also random and normally distributed. There was little growth in between the two periods so the use of $445 = 469 - 24$ treats the days 258 and 469 as one measurement unit apart.

Estimates of regression coefficients for the fixed effects and SDs of the random effects are shown in Table 8; the standard errors of these parameter estimates are obtained with the delete-one jackknife as mentioned in Varin *et al.* [14] for composite likelihood methods. Based on the estimates in this table, for submodels we consider setting $\beta_5, \beta_3, \beta_4$ in turn to zero for the effects of ozone in the second period, initial point, and first period. Hence, we have submodels with 5, 4 and 3 regression parameters. In Table 9, the decisions of the difference full likelihood and composite likelihood information criteria are shown.

For these four models, all of the information criteria chose the same best model with a significant β_4 , the effect of ozone for the growth rate in the first period. Based on these criteria and standard errors, the effect β_5 of the ozone for the growth in the second period is much more negligible, and the effect β_3 of ozone for the period before day 152 is also non-significant. Note that the model with $\beta_5 = 0$ and five non-zero β 's, the AIC/BIC values are relatively closer to those for the best model than the corresponding CLAIC/CLBIC values; this is also seen in the corresponding z -statistics: for β_3 , the ratio of estimate and SE is $-0.118/0.162 = -0.73$ for full likelihood, $-0.097/0.171 = -0.57$ for TCL, and $-0.094/0.175 = -0.54$ for BCL.

Although the four models in Table 9 are ranked the same on all information criteria, this is not the case when we also consider other models with additional binary variables to handle four plots

Table 8. Spruce data: Comparison of parameter estimates from maximizing full likelihood, TCL, BCL; the correlation of the random effects are small and not included. Standard errors (SEs) are obtained via the delete-one jackknife

	Full (SE)	TCL (SE)	BCL (SE)
β_0	4.272 (0.154)	4.310 (0.152)	4.311 (0.152)
β_1	1.415 (0.064)	1.371 (0.062)	1.373 (0.062)
β_2	0.371 (0.021)	0.383 (0.021)	0.382 (0.021)
β_3	-0.101 (0.173)	-0.097 (0.171)	-0.097 (0.171)
β_4	-0.223 (0.076)	-0.228 (0.074)	-0.227 (0.075)
β_5	-0.012 (0.027)	-0.012 (0.027)	-0.012 (0.027)
residSD	0.138 (0.005)	0.126(0.005)	0.118 (0.006)
SD(b_0)	0.616 (0.051)	0.625 (0.050)	0.630 (0.050)
SD(b_1)	0.270 (0.031)	0.323 (0.030)	0.353 (0.034)
SD(b_2)	0.098 (0.018)	0.110 (0.017)	0.118 (0.017)

(two plots for each of ozone and control). That is, to relate to what we found in the simulation examples in Section 4, if we consider many models and some of them are quite close in fit because of some regression coefficients being near zero, then the rankings can be different for full and composite likelihood information criteria.

6. Discussion

In this paper, we have results that show how decisions from CLAIC compare with those from AIC for nested models. This was mostly based on the theory of local alternatives applied to composite likelihood; this is the theory that is most relevant to understand how model selection performs for models that are not far apart.

Table 9. Spruce data: Comparison of decisions for AIC, BIC, CLAIC, and CLBIC. The decision is the number of β 's in the model with smallest information criterion value. The values of CLAIC and CLBIC have been divided by $\binom{13}{3} = 286$ for TCL and by $\binom{13}{2} = 78$ for BCL in order that they are smaller

# β 's	Full likelihood		TCL		BCL	
	AIC	BIC	CLAIC	CLBIC	CLAIC	CLBIC
6	-2319.5	-2288.7	-291.9	-276.7	-124.0	-112.1
5 ($\beta_5 = 0$)	-2321.3	-2292.9	-292.6	-278.3	-124.4	-113.1
4 ($\beta_5 = \beta_3 = 0$)	-2322.7	-2296.6	-293.8	-281.4	-125.4	-115.6
3 ($\beta_5 = \beta_3 = \beta_4 = 0$)	-2314.7	-2291.0	-288.4	-277.5	-120.9	-112.7
Decision	4	4	4	4	4	4

The theory of this paper can be applied to other models to understand better how CLAIC compares with AIC for different types of perturbations that may involve univariate or dependence parameters. This can be done if the \mathbf{J} and \mathbf{H} can be computed, possibly based on simulation methods. Further analysis will help in the understanding of conditions for which CLAIC has more “power” to detect a more complex model. The results have some analogies with those in Joe and Maydeu-Olivares [6], where it is shown that there are directions of local alternatives for which goodness-of-fit statistics based on low-dimensional margins can have more power.

Although analysis in this paper is with composite marginal likelihood, we expect many of the results apply to composite conditional likelihood.

Another topic of research is further study of the extension of the procedure of Vuong [16] for composite likelihood to understand its potential usefulness for comparing prediction similarity for non-nested models.

Appendix A: Proofs

A.1. Assumptions

The following assumptions are used, similar to Vuong [16].

A1: Θ, Γ are compact subsets of a Euclidean space.

A2: Let $\boldsymbol{\vartheta} = \boldsymbol{\theta}$ for model 1 and $\boldsymbol{\vartheta} = \boldsymbol{\gamma}$ for model 2. For $M = 1, 2$, under the true model, we have almost surely for all (\mathbf{x}, \mathbf{y}) , $\log f_{S_q}^{(M)}(\mathbf{y}_{S_q}; \mathbf{x}, \boldsymbol{\theta})$ is twice continuously differentiable over the parameter space. In addition, there exist integrable (under the true model) functions $K_q^{(M)}(\mathbf{x}, \mathbf{y})$, $K_{qj}^{(M)}(\mathbf{x}, \mathbf{y})$, $K_{qjk}^{(M)}(\mathbf{x}, \mathbf{y})$, where ϑ_j, ϑ_k are components in the parameter $\boldsymbol{\vartheta}$, such that

$$\begin{aligned} \sup \left| \log f_{S_q}^{(M)}(\mathbf{y}_{S_q}; \mathbf{x}_i, \boldsymbol{\vartheta}) \right|^2 &< K_q^{(M)}(\mathbf{x}, \mathbf{y}), \\ \sup \left| \frac{\partial}{\partial \vartheta_j} \log f_{S_q}^{(M)}(\mathbf{y}_{S_q}; \mathbf{x}_i, \boldsymbol{\vartheta}) \right|^2 &< K_{qj}^{(M)}(\mathbf{x}, \mathbf{y}), \\ \sup \left| \frac{\partial^2}{\partial \vartheta_j \partial \vartheta_k} \log f_{S_q}^{(M)}(\mathbf{y}_{S_q}; \mathbf{x}_i, \boldsymbol{\vartheta}) \right| &< K_{qjk}^{(M)}(\mathbf{x}, \mathbf{y}), \end{aligned}$$

where the suprema are over the parameter space Θ or Γ .

A3: Under the true model, for $f^{(1)}$, the local maximum point

$$\boldsymbol{\theta}^* = \arg \max_{\Theta} \lim_{n \rightarrow \infty} n^{-1} \sum_{i=1}^n \mathbb{E} \left\{ \sum_{q=1}^Q \log f_{S_q}^{(1)}(\mathbf{Y}_{i,S_q}; \mathbf{x}_i, \boldsymbol{\theta}) \right\}$$

is unique and $\boldsymbol{\theta}^*$ is an interior point of Θ . Similarly $\boldsymbol{\gamma}^*$ is defined for $f^{(2)}$ and is an interior point of Γ .

Assumption A2 guarantees the existence of positive definite matrices $\mathbf{H}^{(1)}$, $\mathbf{H}^{(2)}$, \mathbf{J} given below. For the matrices defined below, all expectations below are taken under the true model.

$$\mathbf{H}^{(1)}(\boldsymbol{\theta}) = - \lim_{n \rightarrow \infty} n^{-1} \sum_{i=1}^n \text{E}_g \left\{ \sum_{q=1}^Q \frac{\partial^2}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^T} \log f_{S_q}^{(1)}(\mathbf{Y}_{i,S_q}; \mathbf{x}_i, \boldsymbol{\theta}) \right\},$$

$$\mathbf{J}^{(11)}(\boldsymbol{\theta}) = \lim_{n \rightarrow \infty} n^{-1} \sum_{i=1}^n \text{E}_g \left\{ \frac{\partial}{\partial \boldsymbol{\theta}} \sum_{q=1}^Q \log f_{S_q}^{(1)}(\mathbf{Y}_{i,S_q}; \mathbf{x}_i, \boldsymbol{\theta}) \cdot \frac{\partial}{\partial \boldsymbol{\theta}^T} \sum_{q=1}^Q \log f_{S_q}^{(1)}(\mathbf{Y}_{i,S_q}; \mathbf{x}_i, \boldsymbol{\theta}) \right\},$$

$$\mathbf{J}^{(12)}(\boldsymbol{\theta}, \boldsymbol{\gamma}) = \lim_{n \rightarrow \infty} n^{-1} \sum_{i=1}^n \text{E}_g \left\{ \frac{\partial}{\partial \boldsymbol{\theta}} \sum_{q=1}^Q \log f_{S_q}^{(1)}(\mathbf{Y}_{i,S_q}; \mathbf{x}_i, \boldsymbol{\theta}) \cdot \frac{\partial}{\partial \boldsymbol{\gamma}^T} \sum_{q=1}^Q \log f_{S_q}^{(2)}(\mathbf{Y}_{i,S_q}; \mathbf{x}_i, \boldsymbol{\gamma}) \right\}.$$

Similarly $\mathbf{H}^{(2)}(\boldsymbol{\gamma})$, $\mathbf{J}^{(22)}(\boldsymbol{\gamma})$, $\mathbf{J}^{(21)}(\boldsymbol{\gamma}, \boldsymbol{\theta})$ can be defined. Let

$$\mathbf{J} = \begin{pmatrix} \mathbf{J}^{(11)}(\boldsymbol{\theta}^*) & \mathbf{J}^{(12)}(\boldsymbol{\theta}^*, \boldsymbol{\gamma}^*) \\ \mathbf{J}^{(21)}(\boldsymbol{\gamma}^*, \boldsymbol{\theta}^*) & \mathbf{J}^{(22)}(\boldsymbol{\gamma}^*) \end{pmatrix}.$$

Applying the law of large numbers and the Central Limit theorem, we have as $n \rightarrow \infty$,

$$-n^{-1} \begin{bmatrix} \frac{\partial^2}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^T} L_{\text{CL}}^{(1)}(\boldsymbol{\theta}^*) & \frac{\partial^2}{\partial \boldsymbol{\gamma} \partial \boldsymbol{\gamma}^T} L_{\text{CL}}^{(2)}(\boldsymbol{\gamma}^*) \end{bmatrix} \rightarrow^{a.s.} [\mathbf{H}^{(1)} \quad \mathbf{H}^{(2)}], \quad (\text{A.1})$$

$$\begin{bmatrix} \frac{1}{\sqrt{n}} \frac{\partial}{\partial \boldsymbol{\theta}^T} L_{\text{CL}}^{(1)}(\boldsymbol{\theta}^*) & \frac{1}{\sqrt{n}} \frac{\partial}{\partial \boldsymbol{\gamma}^T} L_{\text{CL}}^{(2)}(\boldsymbol{\gamma}^*) \end{bmatrix} \rightarrow^d N(\mathbf{0}, \mathbf{J}). \quad (\text{A.2})$$

A.2. Proof of Proposition 3.1

The proof can be established following the same arguments as in Vuong [16], so that most details are omitted. Below, the asymptotic covariance matrix is obtained in a heuristic way.

Based on (A.1) and (A.2), and the assumptions A1–A3 (see Appendix A), Taylor expansions to second order are valid and lead to:

$$2 \text{LR} = n(\hat{\boldsymbol{\gamma}} - \boldsymbol{\gamma}^*)^T \mathbf{H}^{(2)}(\hat{\boldsymbol{\gamma}} - \boldsymbol{\gamma}^*) - n(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*)^T \mathbf{H}^{(1)}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*) + o_p(1),$$

and the matrix of the (asymptotic) quadratic form in independent standard normal random variables is $\mathbf{V}^{1/2} \text{diag}(-\mathbf{H}^{(1)}, \mathbf{H}^{(2)}) \mathbf{V}^{1/2}$, where

$$\mathbf{V} = \text{diag}((\mathbf{H}^{(1)})^{-1}, (\mathbf{H}^{(2)})^{-1}) \mathbf{J} \text{diag}((\mathbf{H}^{(1)})^{-1}, (\mathbf{H}^{(2)})^{-1}) = \begin{pmatrix} \mathbf{V}^{(11)} & \mathbf{V}^{(12)} \\ \mathbf{V}^{(21)} & \mathbf{V}^{(22)} \end{pmatrix}$$

is the asymptotic covariance matrix of $n^{1/2}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*, \hat{\boldsymbol{\gamma}} - \boldsymbol{\gamma}^*)$. The eigenvalues of this matrix are the same as those of

$$\mathbf{A} = \begin{pmatrix} -\mathbf{H}^{(1)} \mathbf{V}^{(11)} & -\mathbf{H}^{(1)} \mathbf{V}^{(12)} \\ \mathbf{H}^{(2)} \mathbf{V}^{(21)} & \mathbf{H}^{(2)} \mathbf{V}^{(22)} \end{pmatrix}.$$

Let

$$\mathbf{K} = \begin{pmatrix} -\mathbf{I}_{p_1} & \mathbf{0} \\ \mathbf{0} & \mathbf{I}_{p_2} \end{pmatrix}.$$

Then

$$\mathbf{KAK} = \begin{pmatrix} -(\mathbf{J}^{(11)})(\mathbf{H}^{(1)})^{-1} & (\mathbf{J}^{(12)})(\mathbf{H}^{(2)})^{-1} \\ -(\mathbf{J}^{(21)})(\mathbf{H}^{(1)})^{-1} & (\mathbf{J}^{(22)})(\mathbf{H}^{(2)})^{-1} \end{pmatrix},$$

and the eigenvalues of this matrix and \mathbf{A} are the same.

A.3. Proof of Theorem 3.1

Consider the nested case where $f^{(1)}(\cdot; \mathbf{x}, \boldsymbol{\theta}) = f^{(2)}(\cdot; \mathbf{x}, \boldsymbol{\theta}, \mathbf{0})$. Suppose that $\boldsymbol{\gamma} = (\boldsymbol{\theta}, \boldsymbol{\zeta})$ is p_2 -dimensional and $\boldsymbol{\zeta}$ is m -dimensional, where $m = p_2 - p_1$. (Note: the maximum composite likelihood estimator for model 1 is $\hat{\boldsymbol{\theta}}$, and it is not the sub-vector of $\hat{\boldsymbol{\gamma}}$, the maximum composite likelihood estimator for model 2.) For convenience, the following notation is used throughout the proof,

$$\mathbf{H}^{(1)} = \mathbf{H}_{\theta\theta} \quad \text{and} \quad \mathbf{H}^{(2)} = \begin{pmatrix} \mathbf{H}_{\theta\theta} & \mathbf{H}_{\theta\zeta} \\ \mathbf{H}_{\zeta\theta} & \mathbf{H}_{\zeta\zeta} \end{pmatrix},$$

$$\mathbf{J}^{(11)} = \mathbf{J}_{\theta\theta}, \quad \mathbf{J}^{(22)} = \begin{pmatrix} \mathbf{J}_{\theta\theta} & \mathbf{J}_{\theta\zeta} \\ \mathbf{J}_{\zeta\theta} & \mathbf{J}_{\zeta\zeta} \end{pmatrix}, \quad \mathbf{J}^{(21)} = \begin{pmatrix} \mathbf{J}_{\theta\theta} \\ \mathbf{J}_{\zeta\theta} \end{pmatrix}, \quad \mathbf{J}^{(12)} = (\mathbf{J}_{\theta\theta} \quad \mathbf{J}_{\theta\zeta}).$$

Proof of (1). For CLBIC, it is a special case of Theorem 1 and 2 in Gao and Song [4]. A detailed treatment on the order consistency can be found in Gao and Song [4]. Below, we complete the proof by showing that P_1^{CLAIC} , the probability that CLAIC selects model 1 under H_1 has the form $P(\lambda_1 U_1 + \dots + \lambda_m U_m < 2(\lambda_1 + \dots + \lambda_m))$.

Let $\boldsymbol{\gamma}^* = (\boldsymbol{\theta}^*, \boldsymbol{\zeta}^*)$ be the true value. Under the null hypothesis, $\boldsymbol{\zeta}^* = \mathbf{0}$. From Taylor expansions of $L_{\text{CL}}^{(2)}(\boldsymbol{\gamma}^*)$ and $L_{\text{CL}}^{(1)}(\boldsymbol{\theta}^*)$ around $\hat{\boldsymbol{\gamma}}$ and $\hat{\boldsymbol{\theta}}$, we have the composite log-likelihood ratio:

$$0 \leq \text{LR} = \frac{1}{2}n(\hat{\boldsymbol{\gamma}} - \boldsymbol{\gamma}^*)^T \mathbf{H}^{(2)}(\hat{\boldsymbol{\gamma}} - \boldsymbol{\gamma}^*) - \frac{1}{2}n(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*)^T \mathbf{H}^{(1)}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*) + o_p(1).$$

From Proposition 3.1, it has asymptotically the same distribution as $\mathbf{Z}^T D \mathbf{Z}$ where \mathbf{Z} is a $(p_1 + p_2)$ -vector of independent $N(0, 1)$ random variables and D is a diagonal matrix with diagonal elements equal to the eigenvalues of \mathbf{B} (defined in (3.3)). In addition, the penalty terms $-(\mathbf{J}^{(11)})(\mathbf{H}^{(1)})^{-1}$ and $(\mathbf{J}^{(22)})(\mathbf{H}^{(2)})^{-1}$ are the two main diagonal blocks in the partitioned matrix \mathbf{B} , respectively. Therefore,

$$\text{tr}[(\mathbf{J}^{(22)})(\mathbf{H}^{(2)})^{-1}] - \text{tr}[(\mathbf{J}^{(11)})(\mathbf{H}^{(1)})^{-1}] = \text{tr} \mathbf{B}.$$

We claim that the number of non-zero eigenvalues λ_i of \mathbf{B} is m . To verify this, the characteristic equation $|\mathbf{B} - \lambda \mathbf{I}_{p_1+p_2}| = 0$ can be written as

$$\begin{vmatrix} \mathbf{J}_{\theta\theta} + \lambda \mathbf{H}_{\theta\theta} & \mathbf{J}_{\theta\theta} & \mathbf{J}_{\theta\zeta} \\ \mathbf{J}_{\theta\theta} & \mathbf{J}_{\theta\theta} - \lambda \mathbf{H}_{\theta\theta} & \mathbf{J}_{\theta\zeta} - \lambda \mathbf{H}_{\theta\zeta} \\ \mathbf{J}_{\zeta\theta} & \mathbf{J}_{\zeta\theta} - \lambda \mathbf{H}_{\theta\zeta} & \mathbf{J}_{\zeta\zeta} - \lambda \mathbf{H}_{\zeta\zeta} \end{vmatrix} = 0.$$

Subtract the second column from the first column, and then subtract the first row from the second row,

$$0 = \begin{vmatrix} \lambda \mathbf{H}_{\theta\theta} & \mathbf{J}_{\theta\theta} & \mathbf{J}_{\theta\zeta} \\ \mathbf{0} & -\lambda \mathbf{H}_{\theta\theta} & -\lambda \mathbf{H}_{\theta\zeta} \\ \lambda \mathbf{H}_{\zeta\theta} & \mathbf{J}_{\zeta\theta} - \lambda \mathbf{H}_{\theta\zeta} & \mathbf{J}_{\zeta\zeta} - \lambda \mathbf{H}_{\zeta\zeta} \end{vmatrix} = (-1)^{p_1} \lambda^{2p_1} \begin{vmatrix} \mathbf{H}_{\theta\theta} & \mathbf{J}_{\theta\theta} & \mathbf{J}_{\theta\zeta} \\ \mathbf{0} & \mathbf{H}_{\theta\theta} & \mathbf{H}_{\theta\zeta} \\ \mathbf{H}_{\zeta\theta} & \mathbf{J}_{\zeta\theta} - \lambda \mathbf{H}_{\theta\zeta} & \mathbf{J}_{\zeta\zeta} - \lambda \mathbf{H}_{\zeta\zeta} \end{vmatrix}.$$

If AIC is considered, the \mathbf{J} matrices are the same as the \mathbf{H} matrices. Subtract the second row from the first, and then subtract the second column multiplied by $\mathbf{H}_{\theta\theta}^{-1} \mathbf{H}_{\theta\zeta}$ from the third to get:

$$0 = \lambda^{2p_1} \begin{vmatrix} \mathbf{H}_{\theta\theta} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{H}_{\theta\theta} & \mathbf{0} \\ \mathbf{H}_{\zeta\theta} & (1 - \lambda) \mathbf{H}_{\theta\zeta} & (1 - \lambda) (\mathbf{H}_{\zeta\zeta} - \mathbf{H}_{\zeta\theta} \mathbf{H}_{\theta\theta}^{-1} \mathbf{H}_{\theta\zeta}) \end{vmatrix}.$$

The eigenvalues are $\lambda = 0$ (multiplicity = $2p_1$) and $\lambda = 1$ (multiplicity = m).

Proof of (2). The required result is a direct consequence of (1) and Lemma A.1.

Proof of (3). For CLAIC, we show that P_1^{CLAIC} is asymptotically equivalent to a non-central chi-square probability. Note that CLAIC selects model 1 if the CLAIC comparison is:

$$\Pr\{2[L_{\text{CL}}^{(2)}(\hat{\boldsymbol{y}}_n) - L_{\text{CL}}^{(1)}(\hat{\boldsymbol{\theta}}_{1n})] < 2\{\text{tr}(\mathbf{J}^{(2)}[\mathbf{H}^{(2)}]^{-1}) - \text{tr}(\mathbf{J}^{(1)}[\mathbf{H}^{(1)}]^{-1})\}\}. \tag{A.3}$$

Here $2[L_{\text{CL}}^{(2)}(\hat{\boldsymbol{y}}_n) - L_{\text{CL}}^{(1)}(\hat{\boldsymbol{\theta}}_{1n})]$ is a non-negative quadratic form, and a representation for it is obtained below.

Write $L_{\text{CL}}(\hat{\boldsymbol{\theta}}_{2n}, \hat{\boldsymbol{\zeta}}_n) = L_{\text{CL}}^{(2)}(\hat{\boldsymbol{y}}_n)$ and $L_{\text{CL}}(\tilde{\boldsymbol{\theta}}_n(\boldsymbol{\zeta}^*), \boldsymbol{\zeta}^*) = L_{\text{CL}}^{(1)}(\hat{\boldsymbol{\theta}}_{1n})$, where $\boldsymbol{\zeta}^* = \mathbf{0}$. Let $\tilde{\boldsymbol{\theta}}_n(\boldsymbol{\zeta})$ be the maximum composite likelihood estimate when $\boldsymbol{\zeta}$ is fixed, so that $L_{\text{CL}}(\tilde{\boldsymbol{\theta}}_n(\boldsymbol{\zeta}), \boldsymbol{\zeta})$ is the profile composite log-likelihood.

Assume that all of the regularity conditions for maximum likelihood apply to all of the marginal densities in the composite likelihood. The derivation below is similar to a result in Cox and Hinkley ([1], Section 9.3) for the full log-likelihood. For the difference of composite log-likelihoods in (A.3), we take an expansion to second order:

$$\begin{aligned} & 2[L_{\text{CL}}(\hat{\boldsymbol{\theta}}_{2n}, \hat{\boldsymbol{\zeta}}_n) - L_{\text{CL}}(\tilde{\boldsymbol{\theta}}_n(\boldsymbol{\zeta}^*), \boldsymbol{\zeta}^*)] \\ &= n(\hat{\boldsymbol{\theta}}_{2n} - \boldsymbol{\theta}^*)^T \mathbf{H}_{\theta\theta}(\hat{\boldsymbol{\theta}}_{2n} - \boldsymbol{\theta}^*) \\ & \quad + 2n(\hat{\boldsymbol{\theta}}_{2n} - \boldsymbol{\theta}^*)^T \mathbf{H}_{\theta\zeta}(\hat{\boldsymbol{\zeta}}_n - \boldsymbol{\zeta}^*) + n(\hat{\boldsymbol{\zeta}}_n - \boldsymbol{\zeta}^*)^T \mathbf{H}_{\zeta\zeta}(\hat{\boldsymbol{\zeta}}_n - \boldsymbol{\zeta}^*) \\ & \quad - n(\tilde{\boldsymbol{\theta}}_n(\boldsymbol{\zeta}^*) - \boldsymbol{\theta}^*)^T \mathbf{H}_{\theta\theta}(\tilde{\boldsymbol{\theta}}_n(\boldsymbol{\zeta}^*) - \boldsymbol{\theta}^*) + o_p(1). \end{aligned} \tag{A.4}$$

For the profile likelihood, by differentiating $\partial L_{\text{CL}}(\tilde{\theta}_n(\zeta), \zeta)/\partial \theta = \mathbf{0}$, one gets:

$$\frac{\partial^2 L_{\text{CL}}}{\partial \theta \partial \theta^T}(\tilde{\theta}_n(\zeta), \zeta) \frac{\partial \tilde{\theta}_n}{\partial \zeta^T} + \frac{\partial^2 L_{\text{CL}}}{\partial \theta \partial \zeta^T}(\tilde{\theta}_n(\zeta), \zeta) = \mathbf{0},$$

so that as $n \rightarrow \infty$,

$$\left. \frac{\partial \tilde{\theta}_n(\zeta)}{\partial \zeta^T} \right|_{\zeta^*} = -\mathbf{H}_{\theta\theta}^{-1} \mathbf{H}_{\theta\zeta} + o_p(1).$$

Expand $\tilde{\theta}_n(\zeta)$ around $\zeta = \zeta^*$ at $\zeta = \hat{\zeta}_n$ to get

$$\tilde{\theta}_n(\zeta^*) = \tilde{\theta}_n(\hat{\zeta}_n) + \left. \frac{\partial \tilde{\theta}_n(\zeta)}{\partial \zeta^T} \right|_{\zeta^*} (\zeta^* - \hat{\zeta}_n) + o_p(1) = \hat{\theta}_{2n} + \mathbf{H}_{\theta\theta}^{-1} \mathbf{H}_{\theta\zeta} (\hat{\zeta}_n - \zeta^*) + o_p(1).$$

Hence,

$$\tilde{\theta}_n(\zeta^*) - \theta^* = \tilde{\theta}_n(\zeta^*) - \hat{\theta}_{2n} + \hat{\theta}_{2n} - \theta^* = \mathbf{H}_{\theta\theta}^{-1} \mathbf{H}_{\theta\zeta} (\hat{\zeta}_n - \zeta^*) + (\hat{\theta}_{2n} - \theta^*) + o_p(1).$$

Substitute into (A.4) to get

$$\begin{aligned} & 2[L_{\text{CL}}(\hat{\theta}_{2n}, \hat{\zeta}_n) - L_{\text{CL}}(\tilde{\theta}_n(\zeta^*), \zeta^*)] \\ & = n(\hat{\zeta}_n - \zeta^*)^T [\mathbf{H}_{\zeta\zeta} - \mathbf{H}_{\zeta\theta} \mathbf{H}_{\theta\theta}^{-1} \mathbf{H}_{\theta\zeta}] (\hat{\zeta}_n - \zeta^*) + o_p(1). \end{aligned} \tag{A.5}$$

Under a sequence of contiguous alternatives, $n^{1/2}(\mathbb{E}[\hat{\zeta}_n] - \zeta_n) \rightarrow \mathbf{0}$ and $n^{1/2}(\zeta_n - \zeta^*) \rightarrow \boldsymbol{\varepsilon}$ as $n \rightarrow \infty$. So marginally $n^{1/2}(\hat{\zeta}_n - \zeta^*)$ is asymptotically $N(\boldsymbol{\delta}_\zeta, \mathbf{V}_\zeta)$, where $\boldsymbol{\delta}_\zeta = \boldsymbol{\varepsilon}$ and \mathbf{V}_ζ is the (2, 2) block of the partitioned covariance matrix,

$$\begin{pmatrix} \mathbf{H}^{\theta\theta} & \mathbf{H}^{\theta\zeta} \\ \mathbf{H}^{\zeta\theta} & \mathbf{H}^{\zeta\zeta} \end{pmatrix}^{-1} \begin{pmatrix} \mathbf{J}_{\theta\theta} & \mathbf{J}_{\theta\zeta} \\ \mathbf{J}_{\zeta\theta} & \mathbf{J}_{\zeta\zeta} \end{pmatrix} \begin{pmatrix} \mathbf{H}^{\theta\theta} & \mathbf{H}^{\theta\zeta} \\ \mathbf{H}^{\zeta\theta} & \mathbf{H}^{\zeta\zeta} \end{pmatrix}^{-1}.$$

Then, (A.5) is asymptotically a quadratic form based on a random vector with $N(\boldsymbol{\delta}_\zeta, \mathbf{V}_\zeta)$ distribution.

For CLBIC, the arguments are similar to that of CLAIC. Here, we highlight the differences between CLBIC and CLAIC. The result is established based on the following comparison

$$2\{L_{\text{CL}}^{(2)}(\hat{\boldsymbol{\gamma}}_n) - L_{\text{CL}}^{(1)}(\hat{\boldsymbol{\theta}}_{1n})\} < \begin{cases} \log n \{ \text{tr}(\mathbf{J}^{(2)}[\mathbf{H}^{(2)}]^{-1}) - \text{tr}(\mathbf{J}^{(1)}[\mathbf{H}^{(1)}]^{-1}) \}, & \text{CLBIC,} \\ 2\{ \text{tr}(\mathbf{J}^{(2)}[\mathbf{H}^{(2)}]^{-1}) - \text{tr}(\mathbf{J}^{(1)}[\mathbf{H}^{(1)}]^{-1}) \}, & \text{CLAIC.} \end{cases} \tag{A.6}$$

The left-hand side has order $O_p(1)$. For CLAIC, the right-hand side is just $O_p(1)$, so there is positive probability that CLAIC selects model 2. On the contrary, for CLBIC, the right-hand side is $O_p(\log n)$. Together with the asymptotic positiveness of the penalty term difference (see Lemma A.2), the increase in the likelihood is offset by the increase in the penalty. Therefore, asymptotically CLBIC cannot select model 2.

Proof of (4). It is similar to the proof of (3) and is omitted here.

A.4. Proof of Theorem 3.2

This is similar to the proof of Theorem 3.1.

A.5. Technical lemmas

Lemma A.1. (1) Let $Z_1^2, Z_2^2, \dots, Z_m^2$ be independent χ_1^2 random variables. Suppose that $m' < m$. Then,

$$P(Z_1^2 + \dots + Z_{m'}^2 < 2m') < P(Z_1^2 + \dots + Z_m^2 < 2m).$$

(2) Further let $\lambda_1, \lambda_2, \dots, \lambda_m$ be non-negative constants. Then,

$$P(\lambda_1 Z_1^2 + \dots + \lambda_m Z_m^2 < 2(\lambda_1 + \dots + \lambda_m)) \leq P(Z_1^2 + \dots + Z_m^2 < 2m).$$

The equality sign holds if and only if $\lambda_1 = \lambda_2 = \dots = \lambda_m$.

Proof. (1) Let \bar{U}_m be the sample average of Z_1^2, \dots, Z_m^2 . Below, we compare the probabilities $P(\bar{U}_m > 2)$ and $P(\bar{U}_{m+1} > 2)$. It can be checked that

$$P(\bar{U}_m > 2) = \int_2^\infty g_m(t) dt = \frac{m}{2^{m/2}\Gamma(m/2)} \int_2^\infty (mt)^{m/2-1} e^{-mt/2} dt.$$

Consider the ratio between the integrands $g_m(t)$ and $g_{m+1}(t)$,

$$R(t) = \frac{g_{m+1}(t)}{g_m(t)} = \frac{(m+1)^{(m+1)/2}\Gamma(m/2)}{\sqrt{2}m^{m/2}\Gamma((m+1)/2)} \sqrt{t}e^{-t/2}.$$

Note that $\sqrt{t}e^{-t/2}$ is monotonic decreasing for $t > 2$, it suffices to show that $R(2) < 1$. To achieve that, the Binet's formula (see Sasvári [12]) can be employed,

$$\frac{\Gamma(m/2)}{\Gamma((m+1)/2)} = \sqrt{2}e^{1/2} \frac{(m-2)^{(m-1)/2}}{(m-1)^{m/2}} \exp[\theta((m-2)/2) - \theta((m-1)/2)],$$

where

$$\theta(x) = \int_0^\infty \left(\frac{1}{e^t - 1} - \frac{1}{t} + \frac{1}{2} \right) e^{-xt} \frac{1}{t} dt.$$

The following bound is also used (see Lemma 2 of Sasvári [12]); for $x > 0$,

$$\theta(x) - \theta(x + 1/2) < \theta(x) - \theta(x + 1) = \left(x + \frac{1}{2} \right) \log \left(1 + \frac{1}{x} \right) - 1.$$

Then

$$\begin{aligned}
 R(2) &\leq e^{-3/2} \sqrt{2} \frac{m}{m-1} \left(1 + \frac{1}{m}\right)^{m/2} \left(1 + \frac{1}{m-2}\right)^{-(m-2)/2} \left(1 + \frac{2}{m-2}\right)^{(m-2)/2} \\
 &\leq \sqrt{2} \frac{m^{1/2}(m+1)^{1/2}}{m-1} \left(1 + \frac{1}{m-2}\right)^{-(m-2)/2}.
 \end{aligned}$$

The right-hand side is monotonic decreasing series of m converging to $\sqrt{2}e^{-1/2} \approx 0.8578 < 1$. It is smaller than 1 when $m \geq 12$. We complete the proof by reporting the numerical values of $P(\bar{U}_m > 2) = P(Z_1^2 + \dots + Z_m^2 > 2m)$ for $m = 1, \dots, 12$. One can see that the monotonic decreasing pattern also holds for $m \leq 12$.

m	1	2	3	4	5	6
$P(\bar{U}_m > 2)$	0.157	0.135	0.112	0.092	0.075	0.062
m	7	8	9	10	11	12
$P(\bar{U}_m > 2)$	0.051	0.042	0.035	0.029	0.024	0.020

(2) Let Ω be the event $\{\lambda_1 Z_1^2 + \dots + \lambda_m Z_m^2 < c(\lambda_1 + \dots + \lambda_m)\}$ and

$$G(\boldsymbol{\lambda}) = P(\lambda_1 Z_1^2 + \dots + \lambda_m Z_m^2 < c(\lambda_1 + \dots + \lambda_m)) = P(\Omega),$$

where the constant c is 2. Without loss of generality, fix the value of $\lambda_1 + \dots + \lambda_m = m$, and let $G^*(\lambda_2, \dots, \lambda_m) = G(m - \lambda_2 - \dots - \lambda_m, \lambda_2, \dots, \lambda_m)$, which we abbreviate below as $G^*(\boldsymbol{\lambda})$. We will consider (i) the stationary points of G^* and (ii) boundary points of G^* .

First, we give the first-order conditions for the stationary points. Rewrite

$$G^*(\boldsymbol{\lambda}) = K \int_0^{cm} \int_0^{cm-v_m} \dots \int_0^{cm-v_m-\dots-v_2} \left\{ \prod_{k=1}^m \lambda_k^{-1/2} v_k^{-1/2} e^{-v_k/2\lambda_k} \right\} dv_1 \dots dv_m.$$

Here, K is a proportionality constant. Let

$$E_i(\boldsymbol{\lambda}) = K \int_0^{cm} \int_0^{cm-v_m} \dots \int_0^{cm-v_m-\dots-v_2} v_i \left\{ \prod_{k=1}^m \lambda_k^{-1/2} v_k^{-1/2} e^{-v_k/2\lambda_k} \right\} dv_1 \dots dv_m,$$

$$E_{ij}(\boldsymbol{\lambda}) = K \int_0^{cm} \int_0^{cm-v_m} \dots \int_0^{cm-v_m-\dots-v_2} v_i v_j \left\{ \prod_{k=1}^m \lambda_k^{-1/2} v_k^{-1/2} e^{-v_k/2\lambda_k} \right\} dv_1 \dots dv_m.$$

Differentiating $G^*(\boldsymbol{\lambda})$ with respect to λ_i for $i \neq 1$, we have

$$-\frac{1}{2\lambda_i} G^*(\boldsymbol{\lambda}) + \frac{1}{2\lambda_i^2} E_i(\boldsymbol{\lambda}) = -\frac{1}{2\lambda_1} G^*(\boldsymbol{\lambda}) + \frac{1}{2\lambda_1^2} E_1(\boldsymbol{\lambda}) = v, \quad i = 2, \dots, m,$$

where v is the Lagrange multiplier. To simplify the first-order conditions, it is convenient to introduce the following notation. Define

$$\begin{aligned}
 h_1(\boldsymbol{\lambda}) &= K \int_0^{cm} \int_0^{cm-v_m} \cdots \int_0^{cm-v_m-\cdots-v_3} \lambda_1^{-3/2} (cm - v_m - \cdots - v_2)^{1/2} \\
 &\quad \times e^{-(cm-v_m-\cdots-v_2)/2\lambda_1} \\
 &\quad \times \left\{ \prod_{k=2}^m \lambda_k^{-1/2} v_k^{-1/2} e^{-v_k/2\lambda_k} \right\} dv_2 \cdots dv_m, \\
 h_{11}(\boldsymbol{\lambda}) &= K \int_0^{cm} \int_0^{cm-v_m} \cdots \int_0^{cm-v_m-\cdots-v_3} \lambda_1^{-5/2} (cm - v_m - \cdots - v_2)^{3/2} \\
 &\quad \times e^{-(cm-v_m-\cdots-v_2)/2\lambda_1} \\
 &\quad \times \left\{ \prod_{k=2}^m \lambda_k^{-1/2} v_k^{-1/2} e^{-v_k/2\lambda_k} \right\} dv_2 \cdots dv_m, \\
 h_{12}(\boldsymbol{\lambda}) &= K \int_0^{cm} \int_0^{cm-v_m} \cdots \int_0^{cm-v_m-\cdots-v_3} \lambda_1^{-3/2} \lambda_2^{-3/2} (cm - v_m - \cdots - v_2)^{1/2} v_2^{1/2} \\
 &\quad \times e^{-(cm-v_m-\cdots-v_2)/2\lambda_1} e^{-v_2/2\lambda_2} \\
 &\quad \times \left\{ \prod_{k=3}^m \lambda_k^{-1/2} v_k^{-1/2} e^{-v_k/2\lambda_k} \right\} dv_2 \cdots dv_m.
 \end{aligned}$$

Similarly, define h_i , h_{ii} , and h_{ij} for other i, j . Below are some useful results obtained from integration by parts over variable v_1 ,

$$\begin{aligned}
 E_1(\boldsymbol{\lambda}) &= \lambda_1 G^*(\boldsymbol{\lambda}) - 2\lambda_1^2 h_1(\boldsymbol{\lambda}), \\
 E_2(\boldsymbol{\lambda}) &= \lambda_2 G^*(\boldsymbol{\lambda}) - 2\lambda_2^2 h_2(\boldsymbol{\lambda}), \\
 E_{11}(\boldsymbol{\lambda}) &= 3\lambda_1 E_1(\boldsymbol{\lambda}) - 2\lambda_1^3 h_{11}(\boldsymbol{\lambda}), \\
 E_{12}(\boldsymbol{\lambda}) &= \lambda_1 E_2(\boldsymbol{\lambda}) - 2\lambda_1^2 \lambda_2 h_{12}(\boldsymbol{\lambda}).
 \end{aligned}$$

Then, the first order conditions becomes $h_1 = h_2 = \cdots = h_m = -v$.

Next, we show that stationary points of $G^*(\boldsymbol{\lambda})$ without satisfying $\lambda_1 = \cdots = \lambda_m$ do not have semi-negative definite Hessian matrix. Differentiating $G^*(\boldsymbol{\lambda})$ with respect to λ_i twice,

$$\begin{aligned}
 \frac{\partial^2 G^*(\boldsymbol{\lambda})}{\partial \lambda_i^2} &= \frac{1}{4\lambda_i^2} \left(3G - \frac{6E_i}{\lambda_i} + \frac{E_{ii}}{\lambda_i^2} \right) + \frac{1}{4\lambda_1^2} \left(3G - \frac{6E_1}{\lambda_1} + \frac{E_{11}}{\lambda_1^2} \right) \\
 &\quad + \frac{2}{4\lambda_i \lambda_1} \left(G - \frac{E_i}{\lambda_i} - \frac{E_1}{\lambda_1} + \frac{E_{1i}}{\lambda_i \lambda_1} \right) \\
 &= \frac{h_i}{\lambda_i} + \frac{h_1}{\lambda_1} + \frac{1}{2\lambda_i} (h_i - h_{ii}) + \frac{1}{2\lambda_1} (h_1 - h_{11}) - \frac{1}{\lambda_i} (h_1 - h_{1i}).
 \end{aligned}$$

Below, we see that the right-hand side must be positive if $\lambda_1 \neq \lambda_i$ and therefore cannot be a local maximum. By definitions, the first two terms are positive. For the third term, consider the quantities defined below,

$$R_{ii} = G^*(\boldsymbol{\lambda}) - \frac{2E_i(\boldsymbol{\lambda})}{\lambda_i} + \frac{E_{ii}(\boldsymbol{\lambda})}{\lambda_i^2}.$$

It can be rewritten as the integration of the product of $(1 - v_i/\lambda_i)^2$ and some positive terms. Therefore, R_{ii} must be positive. In addition, $R_{ii} = 2\lambda_i(h_i - h_{ii})$. Then, we show that $h_i - h_{ii} > 0$. The fourth term can be handled in the same manner. For the last term, the symmetry $E_{li} = E_{il}$ implies $\lambda_i h_i + \lambda_l h_{li} = \lambda_l h_l + \lambda_i h_{il}$; then using the first order condition for a stationary point and the symmetry of h_{ij} , $(\lambda_l - \lambda_i)(h_l - h_{li}) = 0$. If $\lambda_l \neq \lambda_i$, then $h_l - h_{li} = 0$. The stationary point must not be a local maximum.

Now, we have shown that $\boldsymbol{\lambda} = \mathbf{1}_m$ is the only stationary point of $G^*(\boldsymbol{\lambda})$ that could be a local maximum. It should be noted that such stationary point is not necessarily a local maximum. To avoid the difficulties in checking the negative-definiteness of the Hessian matrix, an indirect approach is adopted. Here, we compare the unique stationary point with the boundary points. The boundary is defined by $\{\lambda_i = 0 \text{ for some } i = 1, 2, \dots, m\}$. Result (2) on the boundary points can be established by applying result (1) and result (2) for stationary points inductively. (Note: for any c , $\boldsymbol{\lambda} = \mathbf{1}_m$ is always a stationary point. However, result (1) is not necessarily valid for all c , so, the local maximality does not always hold for any c .) \square

Lemma A.2 (Monotonicity of the penalty term $\text{tr}(\mathbf{H}^{-1}\mathbf{J})$). *If model 1 is nested within model 2, $\text{tr}(\mathbf{H}^{(1)})^{-1}\mathbf{J}^{(1)} < \text{tr}(\mathbf{H}^{(2)})^{-1}\mathbf{J}^{(2)}$, if $\mathbf{H}^{(1)}, \mathbf{J}^{(1)}$ are evaluated at $\boldsymbol{\theta}^*$ and $\mathbf{H}^{(2)}, \mathbf{J}^{(2)}$ are evaluated at $(\boldsymbol{\zeta}^*, \boldsymbol{\theta}^*)$.*

Proof. Suppose that the parameters are $(\boldsymbol{\zeta}^*, \boldsymbol{\theta}^*)$ for model 2 and $\boldsymbol{\theta}^*$ for model 1. Below, if not specified, the arguments of the \mathbf{H}, \mathbf{J} matrices are $(\boldsymbol{\zeta}^*, \boldsymbol{\theta}^*)$. For model 1, the penalty term is $\text{tr}(\mathbf{H}_{\boldsymbol{\theta}\boldsymbol{\theta}}^{-1}\mathbf{J}_{\boldsymbol{\theta}\boldsymbol{\theta}})$.

Next, we consider the partitioning:

$$\begin{pmatrix} \mathbf{H}_{\zeta\zeta} & \mathbf{H}_{\zeta\theta} \\ \mathbf{H}_{\theta\zeta} & \mathbf{H}_{\theta\theta} \end{pmatrix}^{-1} \begin{pmatrix} \mathbf{J}_{\zeta\zeta} & \mathbf{J}_{\zeta\theta} \\ \mathbf{J}_{\theta\zeta} & \mathbf{J}_{\theta\theta} \end{pmatrix}.$$

We have (see Morrison [9], Section 2.11)

$$\begin{pmatrix} \mathbf{H}_{\zeta\zeta} & \mathbf{H}_{\zeta\theta} \\ \mathbf{H}_{\theta\zeta} & \mathbf{H}_{\theta\theta} \end{pmatrix}^{-1} = \begin{pmatrix} (\mathbf{H}_{\zeta\zeta} - \mathbf{H}_{\zeta\theta}\mathbf{H}_{\theta\theta}^{-1}\mathbf{H}_{\theta\zeta})^{-1} & \\ -\mathbf{H}_{\theta\theta}^{-1}\mathbf{H}_{\theta\zeta}(\mathbf{H}_{\zeta\zeta} - \mathbf{H}_{\zeta\theta}\mathbf{H}_{\theta\theta}^{-1}\mathbf{H}_{\theta\zeta})^{-1} & \\ & -(\mathbf{H}_{\zeta\zeta} - \mathbf{H}_{\zeta\theta}\mathbf{H}_{\theta\theta}^{-1}\mathbf{H}_{\theta\zeta})^{-1}\mathbf{H}_{\zeta\theta}\mathbf{H}_{\theta\theta}^{-1} \\ & & \mathbf{H}_{\theta\theta}^{-1} + \mathbf{H}_{\theta\theta}^{-1}\mathbf{H}_{\theta\zeta}(\mathbf{H}_{\zeta\zeta} - \mathbf{H}_{\zeta\theta}\mathbf{H}_{\theta\theta}^{-1}\mathbf{H}_{\theta\zeta})^{-1}\mathbf{H}_{\zeta\theta}\mathbf{H}_{\theta\theta}^{-1} \end{pmatrix}.$$

The change in the penalty term is therefore

$$\begin{aligned} & \text{tr}(\mathbf{H}_{\zeta\zeta} - \mathbf{H}_{\zeta\theta}\mathbf{H}_{\theta\theta}^{-1}\mathbf{H}_{\theta\zeta})^{-1}(\mathbf{J}_{\zeta\zeta} - \mathbf{H}_{\zeta\theta}\mathbf{H}_{\theta\theta}^{-1}\mathbf{J}_{\theta\zeta}) \\ & + \text{tr}\mathbf{H}_{\theta\theta}^{-1}\mathbf{H}_{\theta\zeta}(\mathbf{H}_{\zeta\zeta} - \mathbf{H}_{\zeta\theta}\mathbf{H}_{\theta\theta}^{-1}\mathbf{H}_{\theta\zeta})^{-1}(\mathbf{H}_{\zeta\theta}\mathbf{H}_{\theta\theta}^{-1}\mathbf{J}_{\theta\theta} - \mathbf{J}_{\zeta\theta}) \\ & = \text{tr}(\mathbf{H}_{\zeta\zeta} - \mathbf{H}_{\zeta\theta}\mathbf{H}_{\theta\theta}^{-1}\mathbf{H}_{\theta\zeta})^{-1}(\mathbf{J}_{\zeta\zeta} - \mathbf{H}_{\zeta\theta}\mathbf{H}_{\theta\theta}^{-1}\mathbf{J}_{\theta\zeta} - \mathbf{J}_{\zeta\theta}\mathbf{H}_{\theta\theta}^{-1}\mathbf{H}_{\theta\zeta} + \mathbf{H}_{\zeta\theta}\mathbf{H}_{\theta\theta}^{-1}\mathbf{J}_{\theta\theta}\mathbf{H}_{\theta\theta}^{-1}\mathbf{H}_{\theta\zeta}). \end{aligned}$$

Note that the term

$$\mathbf{J}_{\zeta\zeta} - \mathbf{H}_{\zeta\theta}\mathbf{H}_{\theta\theta}^{-1}\mathbf{J}_{\theta\zeta} - \mathbf{J}_{\zeta\theta}\mathbf{H}_{\theta\theta}^{-1}\mathbf{H}_{\theta\zeta} + \mathbf{H}_{\zeta\theta}\mathbf{H}_{\theta\theta}^{-1}\mathbf{J}_{\theta\theta}\mathbf{H}_{\theta\theta}^{-1}\mathbf{H}_{\theta\zeta}$$

must be positive definite because \mathbf{J} has the form $\text{E}[\nabla\nabla^T]$. It is the same as

$$\text{E}[\nabla_{\zeta} - \mathbf{H}_{\zeta\theta}\mathbf{H}_{\theta\theta}^{-1}\nabla_{\theta}][\nabla_{\zeta} - \mathbf{H}_{\zeta\theta}\mathbf{H}_{\theta\theta}^{-1}\nabla_{\theta}]^T,$$

where ∇_{θ} and ∇_{ζ} are the gradients of the composite log-likelihood with respect to θ and ζ , respectively. The term $(\mathbf{H}_{\zeta\zeta} - \mathbf{H}_{\zeta\theta}\mathbf{H}_{\theta\theta}^{-1}\mathbf{H}_{\theta\zeta})^{-1}$ is also positive definite because it is a principal block of the matrix

$$\begin{pmatrix} \mathbf{H}_{\zeta\zeta} & \mathbf{H}_{\zeta\theta} \\ \mathbf{H}_{\theta\zeta} & \mathbf{H}_{\theta\theta} \end{pmatrix}^{-1}.$$

For any two positive definite matrices A and B , the trace $\text{tr}(AB)$ must be positive. To see this, consider eigenvalue decomposition $A = P\Lambda P^T$. The trace $\text{tr}(AB) = \text{tr}(\Lambda P^T B P)$ is the dot product of the diagonals of Λ and $P^T B P$. Since $P^T B P$ is positive definite, all diagonal elements must be positive. We have the desired results that the penalty term is monotonic increasing. \square

Appendix B: Full and composition likelihoods of the linear mixed-effects model

For the multivariate normal mixed-effects model Laird and Ware [7], both the full likelihood and composite likelihood can be computed readily, after making use of results on vec and vech operations (see Fackler [3], Magnus and Neudecker [8]).

Model:

$$\begin{aligned} \mathbf{Y}_i &= \mathbf{x}_i\boldsymbol{\beta} + \mathbf{z}_i\mathbf{b}_i + \boldsymbol{\varepsilon}_i, & i = 1, 2, \dots, n, \\ \mathbf{b}_i &\sim N(\mathbf{0}, \Psi), & \boldsymbol{\varepsilon}_i \sim N(\mathbf{0}, \phi\mathbf{I}_d), \end{aligned}$$

where $\boldsymbol{\beta}$ is $(s + 1)$ -dimensional vector of fixed effects, \mathbf{b}_i is r -dimensional vector of random effects. \mathbf{x}_i and \mathbf{z}_i are $d \times s$ and $d \times r$ observable matrices, \mathbf{x}_i has a first column of 1s, ϕ is a variance parameter, Ψ is a $r \times r$ covariance matrix.

Conventions: Define the duplication matrix \mathbf{D}_r such that for any $r \times r$ symmetric matrix \mathbf{A} , we have $\text{vec } \mathbf{A} = \mathbf{D}_r \text{vech } \mathbf{A}$. Define permutation matrices \mathbf{T}_{rr} such that for any $r \times r$ matrix \mathbf{A} , we have $\mathbf{T}_{rr} \text{vec } \mathbf{A} = \text{vec } \mathbf{A}^T$. Define the duplication matrix \mathbf{D}_r and elimination matrix \mathbf{E}_r such

that for any $r \times r$ symmetric matrix \mathbf{A} , we have $\mathbf{E}_r \text{vec } \mathbf{A} = \text{vech } \mathbf{A}$ and $\text{vec } \mathbf{A} = \mathbf{D}_r \text{vech } \mathbf{A}$. The duplication matrix is unique but not the elimination matrix; for the latter, it is convenient to operate on the lower triangle. Let \mathbf{I}_r be the $r \times r$ identity matrix. Some properties of the above-mentioned matrices are as follows. (1) $(\mathbf{I}_r + \mathbf{T}_{rr})\mathbf{D}_r = 2\mathbf{D}_r$, $\mathbf{D}_r\mathbf{E}_r(\mathbf{I}_r + \mathbf{T}_{rr}) = \mathbf{I}_r + \mathbf{T}_{rr}$. (2) If \mathbf{C} is lower-triangular, we have $\text{vec } \mathbf{C} = \mathbf{E}_r^T \text{vech } \mathbf{C}$.

Details for the full likelihood and the pairwise composite likelihood are given in two subsections below. The ideas are similar for other composite likelihoods.

B.1. Full likelihood

Define

$$\Omega_i = \mathbf{z}_i \Psi \mathbf{z}_i^T + \phi \mathbf{I} \quad \text{and} \quad \mathbf{S}_i = (\mathbf{y}_i - \mathbf{x}_i \boldsymbol{\beta})(\mathbf{y}_i - \mathbf{x}_i \boldsymbol{\beta})^T.$$

The likelihood function is

$$L(\boldsymbol{\beta}, \Psi, \phi) = \sum_{i=1}^n \ell_i(\boldsymbol{\beta}, \Psi, \phi; \mathbf{y}_i, \mathbf{x}_i),$$

where

$$\ell_i(\boldsymbol{\beta}, \Psi, \phi) = \ell_i(\boldsymbol{\beta}, \Psi, \phi; \mathbf{y}_i, \mathbf{x}_i) = -\frac{1}{2} \{ \text{tr}(\Omega_i^{-1} \mathbf{S}_i) + \log |\Omega_i| \} - \frac{1}{2} \log(2\pi).$$

The following alternative parameterization is beneficial to numerical computation. Consider $\Psi = \mathbf{C}\mathbf{C}^T$ and $\phi = \kappa^2$, where \mathbf{C} is lower triangular matrix. We have

$$\frac{d \text{vech } \Psi}{d \text{vech } \mathbf{C}} = \mathbf{E}_r(\mathbf{I}_r + \mathbf{T}_{rr})(\mathbf{C} \otimes \mathbf{I}_r)\mathbf{E}_r^T.$$

Under the (\mathbf{C}, κ) parametrization, the score function and Fisher's information matrix are given as follows.

Score function:

$$\begin{aligned} \frac{d\ell_i}{d\boldsymbol{\beta}} &= (\mathbf{y}_i - \mathbf{x}_i \boldsymbol{\beta})^T \Omega_i^{-1} \mathbf{x}_i, \\ \frac{d\ell_i}{d \text{vech } \mathbf{C}} &= \text{vec}^T [\mathbf{z}_i^T \Omega_i^{-1} (\mathbf{S}_i - \Omega_i) \Omega_i^{-1} \mathbf{z}_i \mathbf{C}] \mathbf{E}_r^T, \\ \frac{d\ell_i}{d\kappa} &= \kappa \text{tr} \{ \Omega_i^{-2} (\mathbf{S}_i - \Omega_i) \}. \end{aligned}$$

Fisher information matrix:

$$\begin{aligned} \mathbf{E} \frac{d}{d\boldsymbol{\beta}} \left(\frac{d\ell_i}{d\boldsymbol{\beta}} \right) &= -\mathbf{x}_i^T \Omega_i^{-1} \mathbf{x}_i, \\ \mathbf{E} \frac{d}{d \text{vech } \mathbf{C}} \left(\frac{d\ell_i}{d\boldsymbol{\beta}} \right) &= \mathbf{0}, \end{aligned}$$

$$\begin{aligned}
 E \frac{d}{d\boldsymbol{\beta}} \left(\frac{d\ell_i}{d\boldsymbol{\beta}} \right) &= \mathbf{0}, \\
 E \frac{d}{d \text{vech } \mathbf{C}} \left(\frac{d\ell_i}{d \text{vech } \mathbf{C}} \right) &= -\mathbf{E}_r(\mathbf{C}^T \otimes \mathbf{I}_r) \{ [\mathbf{z}_i^T \Omega_i^{-1} \mathbf{z}_i] \otimes [\mathbf{z}_i^T \Omega_i^{-1} \mathbf{z}_i] \} (\mathbf{I}_r + \mathbf{T}_{rr})(\mathbf{C} \otimes \mathbf{I}_r) \mathbf{E}_r^T, \\
 E \frac{d}{d\kappa} \left(\frac{d\ell_i}{d \text{vech } \Psi} \right) &= -2\kappa \mathbf{E}_r(\mathbf{C}^T \otimes \mathbf{I}_r) \text{vec} \{ [\mathbf{z}_i^T \Omega_i^{-2} \mathbf{z}_i] \}, \\
 E \frac{d}{d\kappa} \left(\frac{d\ell_i}{d\kappa} \right) &= -2\kappa^2 \text{tr}(\Omega_i^{-2}).
 \end{aligned}$$

B.2. Composite likelihood

We show details of the pairwise composite log-likelihood for the multivariate Gaussian linear mixed-effects model. Define the composite likelihood as

$$L_{\text{CL}}(\boldsymbol{\beta}, \Psi, \phi) = \sum_{i=1}^n \sum_{1 \leq j < k \leq d} \log f_{jk}(y_{ij}, y_{ik}; \boldsymbol{\beta}, \Psi, \phi, \mathbf{x}_i),$$

where

$$\begin{aligned}
 \log f_{jk}(y_{ij}, y_{ik}; \boldsymbol{\beta}, \Psi, \phi, \mathbf{x}_i) &= \ell_{i,jk}(\boldsymbol{\beta}, \Psi, \phi) \\
 &= -\frac{1}{2} \{ \text{tr}[(e_{jk}^T \Omega_i e_{jk})^{-1} (e_{jk}^T \mathbf{S}_i e_{jk})] + \log |e_{jk}^T \Omega_i e_{jk}| \} - \frac{1}{4} d(d-1) \log(2\pi).
 \end{aligned}$$

Let $\ell_{i\text{CL}} = \sum_{1 \leq j < k \leq d} \ell_{i,jk}(\boldsymbol{\beta}, \Psi, \phi)$. For convenience, for $a = 1, 2, 3$, define

$$\begin{aligned}
 \mathbf{A}_{ai} &= \sum_{jk} e_{jk} (e_{jk}^T \Omega_i e_{jk})^{-a} e_{jk}^T, \\
 \mathbf{B}_i &= \sum_{jk} \{ [e_{jk} (e_{jk}^T \Omega_i e_{jk})^{-1} e_{jk}^T] \otimes [e_{jk} (e_{jk}^T \Omega_i e_{jk})^{-1} e_{jk}^T] \},
 \end{aligned}$$

where e_{jk} is the $d \times 2$ matrix that has 1 in the $(j, 1)$ and $(k, 2)$ positions and 0 elsewhere (premultiplying by e_{jk}^T and postmultiplying by e_{jk} extracts the appropriate 2×2 subcovariance matrix).

Score function: With the above alternative parameterization of \mathbf{C} and κ , we have

$$\begin{aligned}
 \frac{d\ell_{i\text{CL}}}{d\boldsymbol{\beta}} &= (\mathbf{y}_i - \mathbf{x}_i \boldsymbol{\beta})^T \mathbf{A}_{1i} \mathbf{x}_i, \\
 \frac{d\ell_{i\text{CL}}}{d \text{vech } \mathbf{C}} &= \text{vec}^T(\mathbf{S}_i - \Omega_i) \mathbf{B}_i (\mathbf{z}_i \otimes \mathbf{z}_i) (\mathbf{C} \otimes \mathbf{I}_r) \mathbf{E}_r^T, \\
 \frac{d\ell_{i\text{CL}}}{d\kappa} &= \kappa \text{tr} \{ \mathbf{A}_{2i} (\mathbf{S}_i - \Omega_i) \}.
 \end{aligned}$$

Second moment \mathbf{J} of score function:

$$\begin{aligned} E\left(\frac{d\ell_{i\text{CL}}}{d\boldsymbol{\beta}}\right)^T \left(\frac{d\ell_{i\text{CL}}}{d\boldsymbol{\beta}}\right) &= \mathbf{x}_i^T \mathbf{A}_{1i} \Omega_i \mathbf{A}_{1i} \mathbf{x}_i, \\ E\left(\frac{d\ell_{i\text{CL}}}{d\boldsymbol{\beta}}\right)^T \left(\frac{d\ell_{i\text{CL}}}{d \text{vech } \mathbf{C}}\right) &= \mathbf{0}, \\ E\left(\frac{d\ell_{i\text{CL}}}{d\boldsymbol{\beta}}\right)^T \left(\frac{d\ell_{i\text{CL}}}{d\kappa}\right) &= \mathbf{0}, \\ E\left(\frac{d\ell_{i\text{CL}}}{d \text{vech } \mathbf{C}}\right)^T \left(\frac{d\ell_{i\text{CL}}}{d \text{vech } \mathbf{C}}\right) &= \mathbf{E}_r(\mathbf{C}^T \otimes \mathbf{I}_r)(\mathbf{z}_i^T \otimes \mathbf{z}_i^T) \mathbf{B}_i(\Omega_i \otimes \Omega_i) \mathbf{B}_i(\mathbf{z}_i \otimes \mathbf{z}_i)(\mathbf{I}_r + T_{rr})(\mathbf{C} \otimes \mathbf{I}_r) \mathbf{E}_r^T, \\ E\left(\frac{d\ell_{i\text{CL}}}{d \text{vech } \mathbf{C}}\right)^T \left(\frac{d\ell_{i\text{CL}}}{d\kappa}\right) &= 2\kappa \mathbf{E}_r(\mathbf{C}^T \otimes \mathbf{I}_r)(\mathbf{z}_i^T \otimes \mathbf{z}_i^T) \mathbf{B}_i(\Omega_i \otimes \Omega_i) \text{vec}(\mathbf{A}_{2i}), \\ E\left(\frac{d\ell_{i\text{CL}}}{d\kappa}\right) \left(\frac{d\ell_{i\text{CL}}}{d\kappa}\right) &= 2\kappa^2 \text{tr}(\mathbf{A}_{2i} \Omega_i \mathbf{A}_{2i} \Omega_i). \end{aligned}$$

Expectation of Hessian matrix \mathbf{H} :

$$\begin{aligned} E \frac{d}{d\boldsymbol{\beta}} \left(\frac{d\ell_{i\text{CL}}}{d\boldsymbol{\beta}}\right) &= -\mathbf{x}_i^T \mathbf{A}_{1i} \mathbf{x}_i, \\ E \frac{d}{d \text{vech } \mathbf{C}} \left(\frac{d\ell_{i\text{CL}}}{d\boldsymbol{\beta}}\right) &= \mathbf{0}, \\ E \frac{d}{d\kappa} \left(\frac{d\ell_{i\text{CL}}}{d\boldsymbol{\beta}}\right) &= \mathbf{0}, \\ E \frac{d}{d \text{vech } \mathbf{C}} \left(\frac{d\ell_{i\text{CL}}}{d \text{vech } \mathbf{C}}\right) &= -\mathbf{E}_r(\mathbf{C}^T \otimes \mathbf{I}_r)(\mathbf{z}_i^T \otimes \mathbf{z}_i) \mathbf{B}_i(\mathbf{z}_i \otimes \mathbf{z}_i)(\mathbf{I}_r + \mathbf{T}_{rr})(\mathbf{C} \otimes \mathbf{I}_r) \mathbf{E}_r^T, \\ E \frac{d}{d\kappa} \left(\frac{d\ell_{i\text{CL}}}{d \text{vech } \mathbf{C}}\right) &= -2\kappa \mathbf{E}_r(\mathbf{C}^T \otimes \mathbf{I}_r) \text{vec}(\mathbf{z}_i^T \mathbf{A}_{2i} \mathbf{z}_i), \\ E \frac{d}{d\kappa} \left(\frac{d\ell_{i\text{CL}}}{d\kappa}\right) &= -2\kappa^2 \text{tr}(\mathbf{A}_{2i}). \end{aligned}$$

Acknowledgements

This research has been supported by an NSERC Discovery grant. We are grateful to the referees for comments that have led to an improved presentation.

References

- [1] Cox, D.R. and Hinkley, D.V. (1974). *Theoretical Statistics*. London: Chapman & Hall. [MR0370837](#)
- [2] Diggle, P.J., Liang, K.Y. and Zeger, S.L. (1994). *Analysis of Longitudinal Data*. Oxford: Oxford Univ. Press.
- [3] Fackler, P.L. (2005). Notes on matrix calculus. Available at <http://www.stat.duke.edu/~zo2/shared/resources/matrixc1.pdf>.
- [4] Gao, X. and Song, P.X.K. (2010). Composite likelihood Bayesian information criteria for model selection in high-dimensional data. *J. Amer. Statist. Assoc.* **105** 1531–1540. [MR2796569](#)
- [5] Joe, H. and Lee, Y. (2009). On weighting of bivariate margins in pairwise likelihood. *J. Multivariate Anal.* **100** 670–685. [MR2478190](#)
- [6] Joe, H. and Maydeu-Olivares, A. (2010). A general family of limited information goodness-of-fit statistics for multinomial data. *Psychometrika* **75** 393–419. [MR2719935](#)
- [7] Laird, N. and Ware, H.H. (1982). Random-effect models for longitudinal data. *Biometrics* **38** 963–974.
- [8] Magnus, J.R. and Neudecker, H. (1999). *Matrix Differential Calculus with Applications in Statistics and Econometrics*. Wiley Series in Probability and Statistics. Chichester: Wiley. [MR1698873](#)
- [9] Morrison, D.F. (2005). *Multivariate Statistical Methods*. Belmont, CA: Thomson/Brooks/Cole.
- [10] Pinheiro, J.C. and Bates, D.M. (2000). *Mixed-Effects Models in S and S-PLUS*. New York: Springer.
- [11] Rice, S.O. (1980). Distribution of quadratic forms in normal random variables—evaluation by numerical integration. *SIAM J. Sci. Statist. Comput.* **1** 438–448. [MR0610756](#)
- [12] Sasvári, Z. (1999). An elementary proof of Binet’s formula for the gamma function. *Amer. Math. Monthly* **106** 156–158. [MR1671869](#)
- [13] Varin, C. (2008). On composite marginal likelihoods. *Adv. Stat. Anal.* **92** 1–28. [MR2414624](#)
- [14] Varin, C., Reid, N. and Firth, D. (2011). An overview of composite likelihood methods. *Statist. Sinica* **21** 5–42. [MR2796852](#)
- [15] Varin, C. and Vidoni, P. (2005). A note on composite likelihood inference and model selection. *Biometrika* **92** 519–528. [MR2202643](#)
- [16] Vuong, Q.H. (1989). Likelihood ratio tests for model selection and nonnested hypotheses. *Econometrica* **57** 307–333. [MR0996939](#)
- [17] Xu, X. and Reid, N. (2011). On the robustness of maximum composite likelihood estimate. *J. Statist. Plann. Inference* **141** 3047–3054. [MR2796010](#)

Received November 2012 and revised June 2013