JOURNAL OF THE AMERICAN STATISTICAL ASSOCIATION
2016, VOL. 111, NO. 516, 1775–1790, Theory and Methods
http://dx.doi.org/10.1080/01621459.2015.1115762

Taylor & Francis
Taylor & Francis Group

Optimal Model Averaging Estimation for Generalized Linear Models and Generalized Linear Mixed-Effects Models

Xinyu Zhang[a], Dalei Yu[b], Guohua Zou[a], and Hua Liang[c]

[a]Academy of Mathematics and Systems Science, Chinese Academy of Sciences, and School of Mathematical Science, Beijing, China; [b]Statistics and Mathematics College, Yunnan University of Finance and Economics, Kunming, China; [c]Department of Statistics, Geoge Washington University, Washington, DC, USA

ABSTRACT
Considering model averaging estimation in generalized linear models, we propose a weight choice criterion based on the Kullback–Leibler (KL) loss with a penalty term. This criterion is different from that for continuous observations in principle, but reduces to the Mallows criterion in the situation. We prove that the corresponding model averaging estimator is asymptotically optimal under certain assumptions. We further extend our concern to the generalized linear mixed-effects model framework and establish associated theory. Numerical experiments illustrate that the proposed method is promising.

ARTICLE HISTORY
Received August 2014
Revised October 2015

KEYWORDS
Asymptotic optimality; Kullback–Leibler loss; Misspecification; Penalized generalized weighted least squares (PGWLS); Prediction accuracy

1. Introduction

Model averaging, unlike most variable selection studies which focus on identifying important predictors, aims to the prediction accuracy given several predictors (Ando and Li 2014). Using variable selection (or model selection), we end up putting all our inferential eggs in one unevenly woven basket (Longford 2005), while model averaging is a smoothed extension of model selection for substantially reducing risk relative to selection (Hansen 2014). Bayesian model averaging (BMA) has long been a popular statistical technique; see Hoeting et al. (1999) for a comprehensive review of this literature. Unlike BMA, where models are usually weighted by their posterior model probabilities, the current article focuses on the method of determining weights from frequentist perspective.

Over the past decade, there has been a substantial amount of interest in development of asymptotically optimal model averaging procedure. Various strategies have been proposed. See, for example, Mallows model averaging (Hansen 2007), optimal mean squared error averaging (Liang et al. 2011), jackknife model averaging (Hansen and Racine 2012), heteroskedasticity-robust C_p (Liu and Okui 2013), and optimal model averaging for linear mixed-effects models (Zhang, Zou, and Liang 2014). These methods are mainly developed for the linear framework. As far as we know, there is little work on developing optimal model averaging methods under generalized linear (mixed-effects) models.

Our attempt in this article is at developing an optimal model averaging method for generalized linear models (GLMs), for which we immediately face two challenges in contrast to the case in the linear setting. First, we need to develop a proper weight choice criterion. In the existing literature on optimal model averaging for the linear setting, the weight choice criterion is generally developed after one obtains an unbiased estimator of the squared prediction risk. This way is hard to manipulate, if not impossible, because of the generality of the GLM link function. Second, we expect to prove that the resultant model averaging estimator is asymptotically optimal, that is, it minimizes predictive error in the large sample case. However, the proof of the asymptotic optimality is much more difficult comparing to the situation of the linear setting, where a commonly used tool is Theorem 2 of Whittle (1960). This theorem cannot be applied anymore because the model averaging estimator is not a linear function of response variables.

To address the first challenge, we will use a plug-in estimator of the Kullback–Leibler (KL) loss plus a penalty term as the weight choice criterion, which is equivalent to penalizing the negative log-likelihood. It is interesting to note that this criterion reduces to the Mallows criterion (Hansen 2007) for the normal distribution situation. For the second one, we alternatively use the theory for consistency of estimators in misspecified models, developed by White (1982) (see Equation (6) in Section 3 for the details) to establish the asymptotic optimality in the view that all candidate models can be misspecified. We assume the number of candidate models to be finite. The asymptotic optimality will be established for the dimension of covariates being either fixed or diverging.

Under the framework of conditional inference (Jiang 1999), we further extend our work to generalized linear mixed-effects models (GLMMs) for analyzing the nonnormal and nonindependent data. The notion of conditional Kullback–Leibler divergence is introduced and the corresponding weight choice criterion is developed. Then, the asymptotic optimality is investigated accordingly.

The remainder of this article is organized as follows. Section 2 introduces the model averaging estimation for GLMs and proposes a weight choice method by considering the KL loss.

CONTACT Hua Liang hliang@gwu.edu Department of Statistics, George Washington University, Washington, DC 20052, and Special Term Professor at School of Statistics, Capital University of Economics and Business, Beijing, China.

Section 3 presents the asymptotic optimality of the proposed method. Section 4 suggests three model screening strategies. Section 5 develops a model averaging method for GLMMs. Sections 6 and 7 present the numerical results in simulation and real data examples. Section 8 concludes the article. Assumptions and technical proofs are contained in the Appendices.

2. Model Averaging Estimation and Weight Choice

Consider the exponential family:

$$f(y_i|\theta_i,\phi)=\exp\left\{\frac{y_i\theta_i-b(\theta_i)}{\phi}+c(y_i,\phi)\right\},\quad i=1,\ldots,n, \tag{1}$$

where θ_i and ϕ are parameters and $b(\cdot)$ and $c(\cdot,\cdot)$ are known functions. The canonical parameter θ_i connects the parameter $\boldsymbol{\beta}$ and the k-dimension covariate vector $\mathbf{X}_i$ in the form $\theta_i=\mathbf{X}_i^{\mathrm{T}}\boldsymbol{\beta}$. We first assume k to be fixed and discuss diverging situation at the end of Section 3. We estimate $\boldsymbol{\beta}$ under different candidate models, that is, with different elements of $\mathbf{X}_i$.

Suppose we have S candidate models and S is finite. Denote the maximum likelihood estimator of $\boldsymbol{\beta}$ under the sth candidate model by $\widehat{\boldsymbol{\beta}}_{(s)}$. Note that some of components of $\widehat{\boldsymbol{\beta}}_{(s)}$ are restricted to be zeros. Let $\theta_{0,i}$ be the true value of θ_i. We do not assume that the true value $\theta_{0,i}$ is indeed a linear combination of $\mathbf{X}_i$. In other words, it is not required that there exist a $\boldsymbol{\beta}$ so that $\theta_{0,i}=\mathbf{X}_i^{\mathrm{T}}\boldsymbol{\beta}$. Thus, each of the candidate models can be misspecified.

Let the weight vector $\mathbf{w}=(w_1,\ldots,w_S)^{\mathrm{T}}$ belong to the set $\mathcal{W}=\{\mathbf{w}\in[0,1]^S:\sum_{s=1}^S w_s=1\}$. Then, the model averaging estimator of $\boldsymbol{\beta}$ can be expressed as

$$\widehat{\boldsymbol{\beta}}(\mathbf{w})=\sum_{s=1}^S w_s\widehat{\boldsymbol{\beta}}_{(s)}.$$

Let $\mathbf{X}=(\mathbf{X}_1,\ldots,\mathbf{X}_n)^{\mathrm{T}}$ which is assumed to be of full column rank, $\mathbf{y}=(y_1,\ldots,y_n)^{\mathrm{T}}$, $\boldsymbol{\theta}=(\theta_1,\ldots,\theta_n)^{\mathrm{T}}$, and $\boldsymbol{\theta}_0$ be the true value of $\boldsymbol{\theta}$. So a model averaging estimator of $\boldsymbol{\theta}_0$ is

$$\boldsymbol{\theta}\left\{\widehat{\boldsymbol{\beta}}(\mathbf{w})\right\}=\left[\theta_1\left\{\widehat{\boldsymbol{\beta}}(\mathbf{w})\right\},\ldots,\theta_n\left\{\widehat{\boldsymbol{\beta}}(\mathbf{w})\right\}\right]^{\mathrm{T}}=\mathbf{X}\widehat{\boldsymbol{\beta}}(\mathbf{w}).$$

2.1. Weight Choice

Our weight choice criterion is motivated by using the Kullback–Leibler (KL) loss and is defined as follows. Let $\boldsymbol{\mu}=E(\mathbf{y})$, $B_0=\sum_{i=1}^n b(\theta_{0i})$, $B\{\widehat{\boldsymbol{\beta}}(\mathbf{w})\}=\sum_{i=1}^n b[\theta_i\{\widehat{\boldsymbol{\beta}}(\mathbf{w})\}]$, and

$$J(\mathbf{w})=\phi^{-1}B\left\{\widehat{\boldsymbol{\beta}}(\mathbf{w})\right\}-\phi^{-1}\boldsymbol{\mu}^{\mathrm{T}}\boldsymbol{\theta}\left\{\widehat{\boldsymbol{\beta}}(\mathbf{w})\right\}.$$

The KL loss of $\boldsymbol{\theta}\{\widehat{\boldsymbol{\beta}}(\mathbf{w})\}$ is

$$\begin{aligned}\mathrm{KL}(\mathbf{w})&=2\sum_{i=1}^n E_{\mathbf{y}^*}\left\{\log\left\{f(\mathbf{y}^*|\boldsymbol{\theta}_0,\phi)\right\}\right.\\&\quad\left.-\log\left(f[\mathbf{y}^*|\boldsymbol{\theta}\left\{\widehat{\boldsymbol{\beta}}(\mathbf{w})\right\},\phi]\right)\right\}\\&=2\phi^{-1}B\left\{\widehat{\boldsymbol{\beta}}(\mathbf{w})\right\}-2\phi^{-1}\boldsymbol{\mu}^{\mathrm{T}}\boldsymbol{\theta}\left\{\widehat{\boldsymbol{\beta}}(\mathbf{w})\right\}\\&\quad-2\phi^{-1}B_0+2\phi^{-1}\boldsymbol{\mu}^{\mathrm{T}}\boldsymbol{\theta}_0\\&=2J(\mathbf{w})-2\phi^{-1}B_0+2\phi^{-1}\boldsymbol{\mu}^{\mathrm{T}}\boldsymbol{\theta}_0,\end{aligned} \tag{2}$$

where $\mathbf{y}^*$ is another realization from $f(\cdot|\boldsymbol{\theta}_0,\phi)$ and independent of $\mathbf{y}$. Assume ϕ is known. Typically, in logistic and Poisson regressions, $\phi=1$.

If $\boldsymbol{\mu}$ were known, we would obtain a weight vector by minimizing $J(\mathbf{w})$ given the relationship between $J(\mathbf{w})$ and $\mathrm{KL}(\mathbf{w})$ in (2). In practice, the minimization of $J(\mathbf{w})$ is infeasible owing to the unknown parameter $\boldsymbol{\mu}$. An intuitive solution is to estimate $J(\mathbf{w})$. That is, we may use $\mathbf{y}$ to estimate $\boldsymbol{\mu}$ directly, that is, we plug $\mathbf{y}$ into $J(\mathbf{w})$. Then, we can obtain weights by minimizing $\phi^{-1}B\{\widehat{\boldsymbol{\beta}}(\mathbf{w})\}-\phi^{-1}\mathbf{y}^{\mathrm{T}}\boldsymbol{\theta}\{\widehat{\boldsymbol{\beta}}(\mathbf{w})\}$. Unfortunately, this intuitive procedure leads to overfitting, that is, weight one is put into the full model which includes all covariates in $\mathbf{X}_i$, because $\phi^{-1}B\{\widehat{\boldsymbol{\beta}}(\mathbf{w})\}-\phi^{-1}\mathbf{y}^{\mathrm{T}}\boldsymbol{\theta}\{\widehat{\boldsymbol{\beta}}(\mathbf{w})\}$ just equals the negative log-likelihood $-\log[\text{likelihood}\{\boldsymbol{\theta}=\mathbf{X}\widehat{\boldsymbol{\beta}}(\mathbf{w})\}]$. To avoid the overfitting, we introduce a penalty term to $2\phi^{-1}B\{\widehat{\boldsymbol{\beta}}(\mathbf{w})\}-2\phi^{-1}\mathbf{y}^{\mathrm{T}}\boldsymbol{\theta}\{\widehat{\boldsymbol{\beta}}(\mathbf{w})\}$; that is, our proposed weight choice criterion is

$$\mathcal{G}(\mathbf{w})=2\phi^{-1}B\{\widehat{\boldsymbol{\beta}}(\mathbf{w})\}-2\phi^{-1}\mathbf{y}^{\mathrm{T}}\boldsymbol{\theta}\{\widehat{\boldsymbol{\beta}}(\mathbf{w})\}+\lambda_n\mathbf{w}^{\mathrm{T}}\mathbf{k}, \tag{3}$$

where $\lambda_n\mathbf{w}^{\mathrm{T}}\mathbf{k}$ is the penalty term, $\mathbf{k}=(k_1,\ldots,k_S)^{\mathrm{T}}$, k_s is the number of columns of $\mathbf{X}$ used in the sth candidate model, and λ_n is a tuning parameter. The resultant weight vector is defined as

$$\widehat{\mathbf{w}}=\mathrm{argmin}_{\mathbf{w}\in\mathcal{W}}\mathcal{G}(\mathbf{w}). \tag{4}$$

The weight choice criterion (3) can also be obtained by penalizing the negative log-likelihood. When $\boldsymbol{\theta}$ is estimated by the model averaging estimator $\mathbf{X}\widehat{\boldsymbol{\beta}}(\mathbf{w})$, the corresponding penalized negative log-likelihood can be written as

$$\begin{aligned}\mathcal{P}(\mathbf{w})&=-2\log[\text{likelihood}\{\boldsymbol{\theta}=\mathbf{X}\widehat{\boldsymbol{\beta}}(\mathbf{w})\}]+\lambda_n\mathbf{w}^{\mathrm{T}}\mathbf{k}\\&=2\phi^{-1}B\{\widehat{\boldsymbol{\beta}}(\mathbf{w})\}-2\phi^{-1}\mathbf{y}^{\mathrm{T}}\boldsymbol{\theta}\{\widehat{\boldsymbol{\beta}}(\mathbf{w})\}+\lambda_n\mathbf{w}^{\mathrm{T}}\mathbf{k}\\&\quad+C(\mathbf{y},\phi),\end{aligned} \tag{5}$$

where $C(\mathbf{y},\phi)=\sum_{i=1}^n c(y_i,\phi)$ is unrelated to $\mathbf{w}$. It is easy to see that $\mathcal{P}(\mathbf{w})$ and $\mathcal{G}(\mathbf{w})$ are equivalent in the sense of choosing weights.

When all elements of $\mathbf{w}$ are 0 or 1, if $\lambda_n=2$, then the criterion $\mathcal{G}(\mathbf{w})$ is the AIC; and if $\lambda_n=\log(n)$, then the criterion $\mathcal{G}(\mathbf{w})$ is the BIC. For the normal distribution, it is easy to check that if $\lambda_n=2$, then the criterion $\mathcal{G}(\mathbf{w})$ is the Mallows criterion of Hansen (2007).

3. Asymptotic Optimality

Let $\boldsymbol{\beta}^*_{(s)}$ be the parameter vector which minimizes the KL divergence between the true model (1) and the sth candidate model. From Theorem 3.2 of White (1982), we know that, under certain regularity conditions,

$$\widehat{\boldsymbol{\beta}}_{(s)}-\boldsymbol{\beta}^*_{(s)}=O_p(n^{-1/2}). \tag{6}$$

The following theorem establishes the asymptotic optimality of the model averaging estimator $\boldsymbol{\theta}\{\widehat{\boldsymbol{\beta}}(\widehat{\mathbf{w}})\}$.

Theorem 1. If Equation (6) and Conditions C.1–C.3 in Appendix A.1 are satisfied, and $n^{-1/2}\lambda_n=O(1)$, then

$$\frac{\mathrm{KL}(\widehat{\mathbf{w}})}{\inf_{\mathbf{w}\in\mathcal{W}}\mathrm{KL}(\mathbf{w})}\to 1 \tag{7}$$

in probability as $n\to\infty$.

Theorem 1 indicates that our model averaging estimator $\boldsymbol{\theta}\{\widehat{\boldsymbol{\beta}}(\widehat{\mathbf{w}})\}$ is optimal in the sense that the KL loss between the model with $\boldsymbol{\theta}_0$ being estimated by $\boldsymbol{\theta}\{\widehat{\boldsymbol{\beta}}(\widehat{\mathbf{w}})\}$ and the true model is asymptotically identical to that by the infeasible best possible model averaging estimator.

Remark 1. Theorem 1 indicates that the proposed estimator is asymptotically optimal when $\lambda_n = \log(n)$, which seemingly contradicts the results obtained for the BIC by Yang (2005). But actually, there is no contradiction, because in Yang (2005), the author demonstrated that any consistent model selection criterion behaves suboptimally (not optimal) for estimating the regression function subject to the true model being included in the candidate set, while the optimality of the current article is established when all candidate models are misspecified and the consistency is not possible in such a situation. By the way, to obtain the asymptotic optimality, we impose Condition C.3 $n\xi_n^{-2} = o(1)$. This condition is slightly stronger than that $\xi_n \to \infty$, a commonly used condition for deriving the asymptotic optimality given $\lambda_n = 2$ (Hansen 2007), but is the same as the third part of Condition (A7) of Zhang, Zou, and Liang (2014) and is also implied by Conditions (7) and (8) of Ando and Li (2014). We further explore this optimality in Setting 1 of Section 6 and confirm this finding numerically.

Remark 2. Note that both $\lambda_n = 2$ and $\log(n)$ satisfy $n^{-1/2}\lambda_n = O(1)$, required in Theorem 1, which means that if one prefers to AIC or BIC, this can be achieved by choosing $\lambda_n = 2$ or $\log(n)$. The simulation results in Section 6 show that the two versions of our model averaging method with $\lambda_n = 2$ and $\lambda_n = \log(n)$ outperform AIC and BIC model selection procedures, and also outperform model averaging procedures by smoothed AIC (SAIC) and smoothed BIC (SBIC) (Buckland, Burnham, and Augustin 1997).

Theorem 2 states that the proposed criterion still has asymptotic optimality for diverging k_s with additional regularity assumptions.

Theorem 2. If Conditions C.2 and C.4–C.6 in Appendix A.1 are satisfied, and $n^{-1/2}\lambda_n = O(1)$, then (7) still holds in probability as $n \to \infty$.

4. Model Averaging After Model Screening

If the number of candidate models S is large, the computational burden of our model averaging procedure will be heavy. In this case, a model screening step prior to model averaging is desirable. Let $\mathcal{S}^*$ be a subset of $\{1, \ldots, S\}$ and thus $\mathcal{W}^* = \{\mathbf{w} \in [0,1]^S : \sum_{s\in\mathcal{S}^*} w_s = 1 \text{ and } \sum_{s\notin\mathcal{S}^*} w_s = 0\}$ is a subset of $\mathcal{W}$. When model averaging is implemented based on the subset $\mathcal{S}^*$, the resultant weight vector is

$$\widehat{\mathbf{w}}^* = \operatorname{argmin}_{\mathbf{w}\in\mathcal{W}^*} \mathcal{G}(\mathbf{w}). \tag{8}$$

Assumption 1. There exist a nonnegative series of ν_n and a weight series of $\mathbf{w}_n \in \mathcal{W}$ such that $\xi_n^{-1}\nu_n \to 0$, $\inf_{\mathbf{w}\in\mathcal{W}} \mathrm{KL}(\mathbf{w}) = \mathrm{KL}(\mathbf{w}_n) - \nu_n$, and $P(\mathbf{w}_n \in \mathcal{W}^*) \to 1$ as $n \to \infty$, where ξ_n is defined in Appendix A.1.

Theorem 3. If Assumption 0 is satisfied, then under the conditions of Theorem 2,

$$\frac{\mathrm{KL}(\widehat{\mathbf{w}}^*)}{\inf_{\mathbf{w}\in\mathcal{W}} \mathrm{KL}(\mathbf{w})} \to 1 \tag{9}$$

in probability as $n \to \infty$.

Theorem 3 indicates that after adding Assumption 0, our model averaging estimator $\boldsymbol{\theta}\{\widehat{\boldsymbol{\beta}}(\widehat{\mathbf{w}})\}$ is still optimal based on the candidate model set $\mathcal{S}^*$. So we term $\mathcal{S}^*$ as the optimal set. For practical purpose, we suggest three model screening strategies as follows.

STR 1—Threshold model screening (TMS): we use the selected weight vector $\widehat{\mathbf{w}}$ defined in (4) to do model screening. Specially, we remove the models with weights smaller than a threshold constant. This strategy is motivated by the optimal set $\mathcal{S}^*$. Intuitively, Assumption 0 requires that the infeasible optimal weight vector is included by the screened set $\mathcal{W}^*$ with probability approaching to 1. A disadvantage of this screening is still its computational burden because we still need to calculate $\widehat{\mathbf{w}}$.

STR 2—Top m model screening (TmMS): we use a penalized regression with the adaptive LASSO (Zou 2006) to screen out some candidate models. In the implementation of minimizing penalized function, we use the pathwise coordinate descent algorithm (Friedman, Hastie, and Tibshirani 2010) and try m tuning parameters. Different tuning parameters may results in different models, which will be included in our resulting candidate models. This idea is inspired by the work of Yuan and Yang (2005) and Zhang, Lu, and Zou (2013), where they selected m candidate models with smallest information criterion values. If we calculate the information criterion values of all candidate models and the number of candidate models is large, we would still face heavy computational burden.

STR 3—Ordering model screening (OMS): we use the maximum marginal likelihood estimator proposed by Fan and Song (2010) to order the covariates. After ordering, we use k nested models as candidate models, the sth one of which includes the first s covariates. This idea has been used by Claeskens, Croux, and van Kerckhoven (2006) for their model averaging.

Theoretical investigation of the superiority of these three strategies seems very difficult if not impossible. But we will explore their numerical performance in the simulation section.

5. Extension to the Generalized Linear Mixed-Effects Models

In Section 3, we have assumed that all the observations are independent. In practice, responses can be discontinuously distributed (e.g., binary and binomial outcomes and counts) and correlated (Jiang et al. 2001; Molenberghs et al. 2010). With incorporating the random effects into GLM, GLMM is a routine choice for the analysis of nonnormal and correlated data, consists of some of the most useful models in statistics (Lele, Nadeem, and Schmuland 2010), and has been widely used in

various scientific areas. Moreover, even in the independent scenario, the use of random effects allows us to connect a nonexponential family distribution to an exponential family distribution. In Appendix A.5, we will give an example to demonstrate this connection. Given the usefulness of GLMMs aforementioned, it is desirable to consider the model averaging in the context of GLMMs.

Unfortunately, there is no immediate way to extend the procedures developed in Section 2 to the context of GLMMs because we face at least two new challenges. First, in the presence of random effects, the KL loss used in Section 2 is inappropriate since KL loss is based on the marginal likelihood function. So, we need to define a new loss function for developing model averaging method under GLMMs. Second, in $\mathcal{G}(\mathbf{w})$ defined in (3), the penalty term $\lambda_n \mathbf{w}^{\mathrm{T}}\mathbf{k}$ is related to the number of the fixed effects, whereas in GLMMs, we have both fixed and random effects. So we need to use a new penalty term. In this section, we tackle the aforementioned problems within the framework of conditional inference (Jiang 1999). Specifically, we will develop the optimal weight choice schemes of model averaging for GLMM where the estimation/prediction of fixed/random effects are accommodated simultaneously.

Suppose that conditional on the random canonical parameters $\boldsymbol{\theta} = (\theta_{11}, \ldots, \theta_{1n_1}, \ldots, \theta_{m1}, \ldots, \theta_{mn_m})^{\mathrm{T}}$, the observable responses $\{y_{ij}, i = 1, \ldots, m;\ j = 1, \ldots, n_i\}$ have the probability density function

$$f(\mathbf{y} \mid \boldsymbol{\theta}, \phi) = \prod_{i=1}^{m} \prod_{j=1}^{n_i} f(y_{ij} \mid \boldsymbol{\theta}, \phi) = \prod_{i=1}^{m} \prod_{j=1}^{n_i} \exp\left\{\frac{y_{ij}\theta_{ij} - b(\theta_{ij})}{\phi} + c(y_{ij}, \phi)\right\}, \tag{10}$$

where $\mathbf{y} = (y_{11}, y_{12}, \ldots, y_{mn_m})^{\mathrm{T}}$. The correlation structure of $\mathbf{y}$ is featured by the random vector $\boldsymbol{\theta}$ and many typical correlation structures can be written in this form. For example, for data with repeated measurements, y_{ij} represents the jth measurement for the ith subject, then the typical form of θ_{ij} is $\theta_{ij} = \mathbf{x}_{ij}^{\mathrm{T}}\boldsymbol{\beta} + \alpha_i$ where $\boldsymbol{\beta}$ is the population level fixed effects and α_i is the random intercept term.

Given a GLMM, θ_{ij} is modeled by

$$\theta_{ij} = \mathbf{x}_{ij}^{\mathrm{T}}\boldsymbol{\beta} + \mathbf{z}_{ij}^{\mathrm{T}}\boldsymbol{\alpha}_i, \tag{11}$$

where $\mathbf{x}_{ij}$ and $\mathbf{z}_{ij}$ are the covariates vectors, $\boldsymbol{\beta}$ is the fixed-effect vector, and $\boldsymbol{\alpha}_i$, $i = 1, \ldots, m$, are random-effect vectors. Write $\mathbf{X} = (\mathbf{x}_{11}, \ldots, \mathbf{x}_{mn_m})^{\mathrm{T}}$, $\mathbf{Z} = \mathrm{diag}\{(\mathbf{z}_{11}, \ldots, \mathbf{z}_{1n_1})^{\mathrm{T}}, \ldots, (\mathbf{z}_{m1}, \ldots, \mathbf{z}_{mn_m})^{\mathrm{T}}\}$, and $\boldsymbol{\alpha} = (\boldsymbol{\alpha}_1^{\mathrm{T}}, \ldots, \boldsymbol{\alpha}_m^{\mathrm{T}})^{\mathrm{T}}$. We assume $\mathbf{X}$ and $\mathbf{Z}$ have full column ranks. Let $\mathbf{U} = (\mathbf{X}, \mathbf{Z})$ and $\boldsymbol{\gamma} = (\boldsymbol{\beta}^{\mathrm{T}}, \boldsymbol{\alpha}^{\mathrm{T}})^{\mathrm{T}}$ such that

$$\boldsymbol{\theta} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\boldsymbol{\alpha} = \mathbf{U}\boldsymbol{\gamma}.$$

Denote the true value of $\boldsymbol{\theta}$ by $\boldsymbol{\theta}_0 = (\theta_{0,11}, \theta_{0,12}, \ldots, \theta_{0,mn_m})^{\mathrm{T}}$. Note that $\mathbf{U}\boldsymbol{\gamma}$ is not necessarily equal to the true value $\boldsymbol{\theta}_0$, therefore misspecification is allowed. In the rest of the article, we assume that $\mathbf{U}$ is of full column rank.

Remark 3. When $\mathbf{U}$ is not full column ranked, the random canonical parameters vector $\boldsymbol{\theta}$ needs to be reparameterized. Specifically, we reparameterize $\boldsymbol{\theta}$ as $\boldsymbol{\theta} = \mathbf{X}_1\boldsymbol{\beta}_1 + \mathbf{X}_2\boldsymbol{\beta}_2 + \mathbf{Z}\boldsymbol{\alpha}$, where $\mathbf{X}_2$ is one of the largest submatrices of $\mathbf{X}$ such that $(\mathbf{X}_2, \mathbf{Z})$ is of full column rank, and $\mathbf{X}_1$ is the other part of $\mathbf{X}$. Accordingly, there exists a matrix $\mathbf{D}$ such that $\mathbf{X}_1 = (\mathbf{X}_2, \mathbf{Z})\mathbf{D}$. One can reparameterize parameters as $\boldsymbol{\theta} = \mathbf{X}_1\boldsymbol{\beta}_1 + \mathbf{X}_2\boldsymbol{\beta}_2 + \mathbf{Z}\boldsymbol{\alpha} = \mathbf{X}_2\breve{\boldsymbol{\beta}} + \mathbf{Z}\breve{\boldsymbol{\alpha}}$, where $\breve{\boldsymbol{\beta}} = \boldsymbol{\beta}_{11} + \boldsymbol{\beta}_2$, $\breve{\boldsymbol{\alpha}} = \boldsymbol{\beta}_{12} + \boldsymbol{\alpha}$, and $\boldsymbol{\beta}_{11}$ and $\boldsymbol{\beta}_{12}$ are subvectors of $\mathbf{D}\boldsymbol{\beta}_1$ such that $(\boldsymbol{\beta}_{11}^{\mathrm{T}}, \boldsymbol{\beta}_{12}^{\mathrm{T}})^{\mathrm{T}} = \mathbf{D}\boldsymbol{\beta}_1$ and the dimension of $\boldsymbol{\beta}_{12}$ is the same as that of $\boldsymbol{\alpha}$. $\breve{\boldsymbol{\beta}}$ and $\breve{\boldsymbol{\alpha}}$ contain new fixed and random effect vectors, respectively. Under this new parameter setting, the matrix $(\mathbf{X}_2, \mathbf{Z})$ is of full column rank. See Zhao et al. (2006), Tan and David (2013), and Yu, Zhang, and Yau (2014) for details on this reparameterization.

The mixed-effects $\boldsymbol{\gamma}$ can be estimated/predicted simultaneously by maximizing a log penalized likelihood function; see Jiang (1999), Wang, Tsai, and Qu (2012), and Yu, Zhang, and Yau (2014) for details. The resultant estimator is referred to as the penalized generalized weighted least-squares (PGWLS) estimator. The PGWLS estimator is flexible for the mixed-effects model with unknown random effects distributions and it is computationally efficient (Jiang 1999; Wang, Tsai, and Qu 2012).

Assume we have S candidate GLMMs. The sth one uses p_s components of $\mathbf{x}_{ij}$ and d_{si} components of $\mathbf{z}_{ij}$. p_s and d_{si} are assumed to be fixed. Write $\widehat{\boldsymbol{\beta}}_{(s)}$ and $\widehat{\boldsymbol{\alpha}}_{(s)} = (\widehat{\boldsymbol{\alpha}}_{(s),1}^{\mathrm{T}}, \ldots, \widehat{\boldsymbol{\alpha}}_{(s),m}^{\mathrm{T}})^{\mathrm{T}}$ as the estimators of $\boldsymbol{\beta}$ and $\boldsymbol{\alpha}$ under the sth model based on the (restrictive) penalized log-likelihood. Note that some of components of $\widehat{\boldsymbol{\beta}}_{(s)}$ and $\widehat{\boldsymbol{\alpha}}_{(s)}$ are restricted to be zeros. Write $\widehat{\boldsymbol{\gamma}}_{(s)} = (\widehat{\boldsymbol{\beta}}_{(s)}^{\mathrm{T}}, \widehat{\boldsymbol{\alpha}}_{(s)}^{\mathrm{T}})^{\mathrm{T}}$. Then, the corresponding estimator of $\boldsymbol{\theta}_0$ is $\widehat{\boldsymbol{\theta}}_{(s)} = \mathbf{U}\widehat{\boldsymbol{\gamma}}_{(s)}$. Write $\mathbf{U} = (\mathbf{u}_{11}, \mathbf{u}_{12}, \ldots, \mathbf{u}_{mn_m})^{\mathrm{T}}$. The conditional Kullback–Leibler divergence from the true model (10) to the sth candidate model is defined as

$$2\sum_{i=1}^{m}\sum_{j=1}^{n_i}\left(E_{\mathbf{y}|\boldsymbol{\theta}_0}\left\{\log f(y_{ij} \mid \theta_{0,ij}, \phi)\right\} - E_{\mathbf{y}|\boldsymbol{\theta}_0}\left[\log f\left\{y_{ij} \mid \mathbf{u}_{ij}^{\mathrm{T}}\boldsymbol{\gamma}_{(s)}, \phi\right\}\right]\right), \tag{12}$$

where $\boldsymbol{\gamma}_{(s)}$ contains unknown parameters and some elements of it are restricted to be zeros.

The model averaging estimator of $\boldsymbol{\theta}_0$ can be expressed by

$$\widehat{\boldsymbol{\theta}}(\mathbf{w}) = \sum_{s=1}^{S} w_s \widehat{\boldsymbol{\theta}}_{(s)}.$$

Denote $\widehat{\boldsymbol{\theta}}(\mathbf{w}) = (\widehat{\theta}_{\mathbf{w},11}, \widehat{\theta}_{\mathbf{w},12}, \ldots, \widehat{\theta}_{\mathbf{w},mn_m})^{\mathrm{T}}$, $\boldsymbol{\mu} = (\mu_{0,11}, \mu_{0,12}, \ldots, \mu_{0,mn_m})^{\mathrm{T}} = \{E(y_{11} \mid \boldsymbol{\theta}_0), E(y_{12} \mid \boldsymbol{\theta}_0), \ldots, E(y_{mn_m} \mid \boldsymbol{\theta}_0)\}^{\mathrm{T}}$ and $B(\boldsymbol{\theta}) = 2\phi^{-1}\sum_{i=1}^{m}\sum_{j=1}^{n_i} b(\theta_{ij})$. The conditional Kullback–Leibler (cKL) loss using the model averaging estimator $\widehat{\boldsymbol{\theta}}(\mathbf{w})$ is defined as

$$\begin{aligned}
\mathrm{cKL}(\mathbf{w}) &= 2\sum_{i=1}^{m}\sum_{j=1}^{n_i}\left(E_{\mathbf{y}^*|\boldsymbol{\theta}_0}\left\{\log f(y_{ij}^* \mid \theta_{0,ij}, \phi)\right\} - E_{\mathbf{y}^*|\boldsymbol{\theta}_0}\left[\log f\left(y_{ij}^* \mid \widehat{\theta}_{\mathbf{w},ij}, \phi\right)\right]\right) \\
&= 2\phi^{-1}\sum_{i=1}^{m}\sum_{j=1}^{n_i}\mu_{0,ij}\left(\theta_{0,ij} - \widehat{\theta}_{\mathbf{w},ij}\right) + 2\phi^{-1}\sum_{i=1}^{m}\sum_{j=1}^{n_i}\{b(\widehat{\theta}_{\mathbf{w},ij}) - b(\theta_{0,ij})\} \\
&= 2\phi^{-1}\left[\boldsymbol{\mu}^{\mathrm{T}}\boldsymbol{\theta}_0 - B(\boldsymbol{\theta}_0) - \boldsymbol{\mu}^{\mathrm{T}}\widehat{\boldsymbol{\theta}}(\mathbf{w}) + B\left\{\widehat{\boldsymbol{\theta}}(\mathbf{w})\right\}\right],
\end{aligned} \tag{13}$$

where conditional on $\boldsymbol{\theta}_0$, $\mathbf{y}^* = (y^*_{11}, y^*_{12}, \ldots, y^*_{mn_m})^{\mathrm{T}}$ is an independent copy of $\mathbf{y}$.

Similar to the weight choice criterion $\mathcal{G}(\mathbf{w})$ in (3), we propose a criterion

$$\mathcal{G}_c(\mathbf{w}) = 2\phi^{-1}B\{\widehat{\boldsymbol{\theta}}(\mathbf{w})\} - 2\phi^{-1}\mathbf{y}^{\mathrm{T}}\widehat{\boldsymbol{\theta}}(\mathbf{w}) + \mathbf{w}^{\mathrm{T}}\boldsymbol{\Upsilon}\boldsymbol{\varpi}_{m,n}, \quad (14)$$

where $\boldsymbol{\varpi}_{m,n} = (\varpi_0, \ldots, \varpi_m)^{\mathrm{T}}$ is the vector of tuning parameters and

$$\boldsymbol{\Upsilon} = \begin{pmatrix} p_1 & d_{11} & \ldots & d_{1m} \\ \ldots & \ldots & \ldots & \ldots \\ p_S & d_{S1} & \ldots & d_{Sm} \end{pmatrix}.$$

Note that more strictly, $\boldsymbol{\varpi}_{m,n}$ should be replaced by $\boldsymbol{\varpi}_{n_1,\ldots,n_m}$, but for simplicity, we use $\boldsymbol{\varpi}_{m,n}$. Different from $\mathcal{G}(\mathbf{w})$ in (3), the tuning parameters in $\mathcal{G}_c(\mathbf{w})$ can vary for different i, $i = 1, \ldots, m$, and the tuning parameters corresponding to the fixed and random effects can also be different. The obtained weight vector is $\widetilde{\mathbf{w}} = \mathrm{argmin}_{\mathbf{w}\in\mathcal{W}}\mathcal{G}_c(\mathbf{w})$. It is straightforward to show that the criterion $\mathcal{G}_c(\mathbf{w})$ reduces to the criterion $\mathcal{G}(\mathbf{w})$ given by (3).

If we set $\varpi_i = 2$ for $i = 0, \ldots, m$ and the sth element of $\mathbf{w}$ in $\mathcal{G}_c(\mathbf{w})$ as one and others as zeros, then $\mathcal{G}_c(\mathbf{w})$ becomes

$$2\phi^{-1}B\{\widehat{\boldsymbol{\theta}}_{(s)}\} - 2\phi^{-1}\mathbf{y}^{\mathrm{T}}\widehat{\boldsymbol{\theta}}_{(s)} + 2\big(p_s + \sum_{i=1}^{m} d_{si}\big),$$

which is the conditional Akaike information criterion (cAIC) for the sth model. Formal derivations of the cAIC under the current scenario are referred to Yu, Zhang, and Yau (2014).

Let $N = \sum_{i=1}^{m} n_i$. If we set $\varpi_0 = \log(N)$, $\varpi_i = \log(n_i)$ for $i = 1, \ldots, m$, and the sth element of $\mathbf{w}$ in $\mathcal{G}_c(\mathbf{w})$ as one and others as zeros, then $\mathcal{G}_c(\mathbf{w})$ reduces to the conditional Bayesian information criterion (cBIC) for the sth model as follows:

$$2\phi^{-1}B\{\widehat{\boldsymbol{\theta}}_{(s)}\} - 2\phi^{-1}\mathbf{y}^{\mathrm{T}}\widehat{\boldsymbol{\theta}}_{(s)} + \log(N)p_s + \sum_{i=1}^{m}\log(n_i)d_{si}.$$

The derivation of cBIC is provided in Appendix A.7.

Let $n_* = \min_{i=1,\ldots,m} n_i$ and $\|\mathbf{v}\|_\infty$ be the supremum norm of a general vector $\mathbf{v}$. We have the following asymptotic optimality for $\widehat{\boldsymbol{\theta}}(\widetilde{\mathbf{w}})$.

Theorem 4. If Conditions C.7–C.11 in Appendix A.1 are satisfied and $n_*^{-1/2}\|\boldsymbol{\varpi}_{m,n}\|_\infty = O(1)$, then

$$\frac{\mathrm{cKL}(\widetilde{\mathbf{w}})}{\inf_{\mathbf{w}\in\mathcal{W}}\mathrm{cKL}(\mathbf{w})} \to 1 \quad (15)$$

in probability, as $\boldsymbol{n_*} \to \infty$.

Theorem 4 indicates that with appropriate tuning parameters, the model averaging estimator $\widehat{\boldsymbol{\theta}}(\widetilde{\mathbf{w}})$ is optimal in the sense that the conditional KL loss using $\widehat{\boldsymbol{\theta}}(\widetilde{\mathbf{w}})$ is asymptotically identical to that by the infeasible best possible model averaging estimator.

Remark 4. On the choice of the tuning parameters vector $\boldsymbol{\varpi}_{m,n}$, we can take $\boldsymbol{\varpi}_{m,n} = 2 \times \mathbf{1}_{(m+1)\times 1}$, because in this case, we can connect the criterion $\mathcal{G}_c(\mathbf{w})$ to cAIC. Moreover, we can also take $\boldsymbol{\varpi}_{m,n} = \{\log(N), \log(n_1), \ldots, \log(n_m)\}^{\mathrm{T}}$, which is the vector of logarithms of sizes for fixed effects and random effects, respectively. As such, we can then link the criterion $\mathcal{G}_c(\mathbf{w})$ with cBIC. The simulation results in Section 6 show that these two versions of averaging methods outperform model selection based on cAIC and cBIC and model averaging based on smoothed cAIC (ScAIC) and smoothed cBIC (ScBIC), respectively.

6. Simulation Study

In this section, we conduct simulation experiments to demonstrate the finite sample performance of the model averaging methods. Because they achieve the asymptotic optimality, we name them OPT. For GLMs, we use two versions of OPT: OPT1 with $\lambda_n = 2$ and OPT2 with $\lambda_n = \log(n)$. For GLMMs, we also try two versions: OPT1 with $\boldsymbol{\varpi}_{m,n} = 2 \times \mathbf{1}_{(m+1)\times 1}$ and OPT2 with $\boldsymbol{\varpi}_{m,n} = \{\log(N), \log(n_1), \ldots, \log(n_m)\}^{\mathrm{T}}$.

6.1. Alternative Methods

For GLMs, we compare the OPT methods with several popular model selection methods: AIC and BIC, and model averaging methods using scores of information criteria: SAIC and SBIC (Buckland, Burnham, and Augustin 1997). These methods are also advocated by Hjort and Claeskens (2003), Claeskens et al. (2006), and Zhang and Liang (2011). The SAIC method assigns weights

$$w_{\mathrm{AIC},s} = \exp(-\mathrm{AIC}_s/2)/\sum_{s=1}^{S}\exp(-\mathrm{AIC}_s/2)$$

to model s and the SBIC allocates weights similarly. For GLMMs, we compare OPT methods (see Formula (14) for the weight choice criterion) with model selection methods by cAIC, cBIC and their smoothed versions: smoothed cAIC (ScAIC) and smoothed cBIC (ScBIC) whose weights are set by the same way as the SAIC and SBIC, respectively.

6.2. Simulation Designs and Results

We consider three simulation settings with the first two for the GLMs, and the third for the GLMMs. In the first setting, all candidate models are misspecified. In the second setting, the number of candidate models is large, the true model can be one of the candidate models, and we apply the model screening methods proposed in Section 4 to this setting. In Settings 1 and 2, we set sample size $n \in \{100, 200\}$, and use KL loss (divided by n) defined in (2) for assessment. In Setting 3, we set $n_i = n$ and $(m, n) \in \{(8, 16), (8, 20), (10, 28), (10, 32)\}$, and use the conditional KL loss (divided by mn) defined in (13) for assessment. For each setting, we generate 500 simulated data.

Setting 1. We generate y_i from Binomial$(1, p_{0i})$ with

$$p_{0i} = \exp\left(\mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta}\right)\left\{1 + \exp\left(\mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta}\right)\right\},$$

where $\boldsymbol{\beta} = (1, 0.2, -1.2, -0.5, 0.7)^{\mathrm{T}}$ and the components of $\mathbf{x}_i$ follow normal distributions with mean zeros, variance ones and the correlation between different components of $\mathbf{x}_i$ being 0.75. To mimic the situation that all candidate models are misspecified, we pretend the last covariate missed in all candidate models. As a result, we have $S = 2^4 - 1$ logistic regression candidate models. This design is also applied to Poisson(p_{0i}) with $p_{0i} = \exp(\mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta})$ to generate responses.

Table 1. Simulation results for Setting 1 ($\times 10^{-1}$).

Regression	n		OPT1	AIC	SAIC	OPT2	BIC	SBIC
Logistic	100	Median	0.647	1.204	0.991	0.675	1.389	1.351
		Mean	0.732	1.116	0.989	0.782	1.416	1.352
		SD	0.305	0.454	0.378	0.333	0.220	0.260
	200	Median	0.493	0.535	0.555	0.522	1.323	1.258
		Mean	0.525	0.670	0.636	0.566	1.237	1.161
		SD	0.135	0.318	0.232	0.165	0.312	0.323
Poisson	100	Median	2.598	3.085	2.973	2.757	3.349	3.340
		Mean	2.686	3.151	3.009	2.826	3.489	3.456
		SD	0.641	0.719	0.686	0.681	0.943	0.882
	200	Median	2.419	2.678	2.590	2.491	3.198	3.158
		Mean	2.466	2.728	2.660	2.558	3.277	3.243
		SD	0.457	0.577	0.521	0.477	0.589	0.581

Simulation results are shown in Table 1. It is seen that both OPT1 and OPT2 always yield smaller mean and median values than their competitors AIC/SAIC and BIC/SBIC. This pattern is also almost true regarding standard deviation (SD) values except one case of the logistic regression with $n = 100$, in which OPT2 yields larger SD than BIC and SBIC.

Next, following a referee's suggestion, we use a nonexponential family distribution to generate data. Specifically, we generate y_i by a negative binomial distribution

$$\Pr(y_i = k) = \binom{k + r_i - 1}{k} p^k (1-p)^{r_i}, \text{ for } k = 0, 1, 2, \ldots, \tag{16}$$

where $p = 0.5$ is the success probability in each experiment, $r_i = \lceil 2\exp(\mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta})\rceil$ is the number of failures until the experiment is stopped, $\lceil a \rceil$ returns the smallest integer larger than a, and $\boldsymbol{\beta} = (1, 0.2, -1.2, -0.5, 0)^{\mathrm{T}}$. We take Poisson regression Poisson$\{\exp(\mathbf{x}_i{}^{\mathrm{T}}\boldsymbol{\beta})\}$ to fit the data. We still omit the last covariate and consider $S = 2^4 - 1$ candidate models. Note that the last element of $\boldsymbol{\beta}$ is zero, so the largest candidate model does not omit covariate, but it is still misspecified, because the response follows the negative binomial distribution instead of the Poisson distribution. The KL loss defined in (2) is based on exponential family, so instead, we use $n^{-1}\sum_{i=1}^{n}\{\widehat{E}(y_i|\mathbf{x}_i) - E(y_i|\mathbf{x}_i)\}^2$ to evaluate methods, where $\widehat{E}(y_i|\mathbf{x}_i)$ is obtained by model selection or averaging methods. Simulation results are shown in Table 2. It is seen that in all situations OPT1 always performs better than AIC and SAIC, and OPT2 always performs better than BIC and SBIC.

Table 2. Simulation results for Setting 1: Nonexponential distribution.

n		OPT1	AIC	SAIC	OPT2	BIC	SBIC
100	Median	7.794	8.115	8.075	7.745	9.319	9.015
	Mean	8.955	9.475	9.411	8.844	11.085	10.759
	SD	4.869	5.248	5.235	4.650	6.492	6.265
200	Median	7.993	8.211	8.102	7.845	8.548	8.521
	Mean	8.732	8.923	8.903	8.616	9.281	9.184
	SD	3.845	4.001	3.974	3.736	4.181	4.111

We calculate the means of $\mathrm{KL}(\widehat{\mathbf{w}})/\inf_{\mathbf{w}\in\mathcal{W}}\mathrm{KL}(\mathbf{w})$ with $\lambda_n = 2$ or $\log(n)$ based on the 500 replications. The mean curves, shown in Figure 1, decrease and approach to 1 with the increase of sample size n. The feature confirms asymptotic optimality numerically as we claimed in Remark 1.

Setting 2. This setup is adopted from Zou (2006). Specially,

$$\Pr(y_i = 1) = \exp(\mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta})/\{1 + \exp(\mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta})\},$$

where the components of $\mathbf{x}_i$ follow normal distributions with mean zeros, variance ones and the correlation between different components of $\mathbf{x}_i$ being 0.75, and $\boldsymbol{\beta} = (3, 0, 0, 1.5, 0, 0, 7, 0, 0)^{\mathrm{T}}$ or $\boldsymbol{\beta} = (3, 0.1, -0.2, 1.5, 0.05, 0.08, 7, 0.1, 0.3)^{\mathrm{T}}$, which represents sparse and nonsparse cases, respectively. Since there are $2^9 - 1 = 511$ submodels, we implement three screening strategies, TMS, TmMS, and OMS, proposed in Section 4 to screen models prior to averaging. In TMS, we use the weighs from OPT2 to screen models and use 0.01 as the threshold constant. In TmMS, we set $m = 100$, which is the default number in the pathwise coordinate descent method of Friedman, Hastie, and Tibshirani (2010).

Simulation results for sparse and nonsparse cases are shown in Tables 3 and 4, respectively. It is seen from Table 3 that OPT1 always performs better than AIC and SAIC, and OPT2 performs better than OPT1 in most of cases. When using TmMS or OMS, the BIC and SBIC sometimes yield smaller medians than OPT2; see, for example, the cases with $n = 100$. But, depending on mean or SD, OPT2 performs much better than BIC and SBIC,

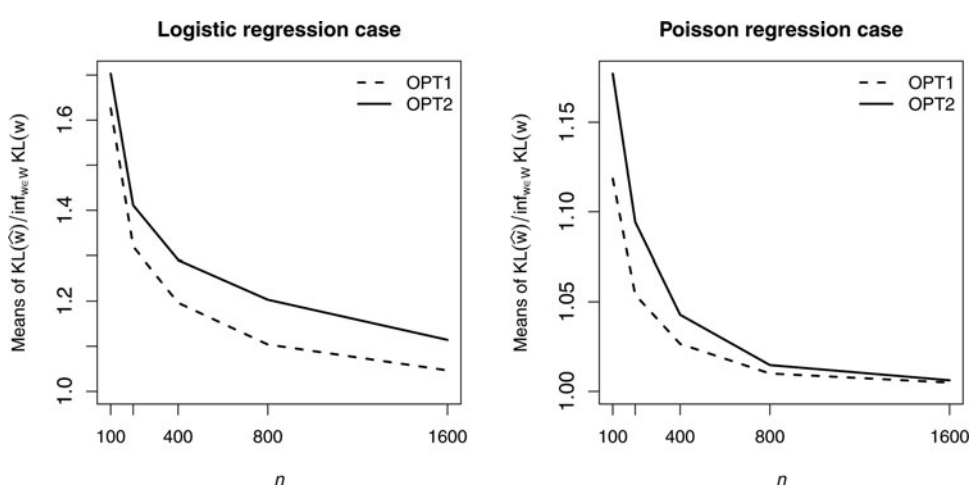

Figure 1. Assessing the asymptotic optimality of OPT1 ($\lambda_n = 2$) and OPT2 ($\lambda_n = \log(n)$) using simulation Setting 1.

Table 3. Simulation results for Setting 2: Sparse model ($\times 10^{-1}$).

Model Screening	n		OPT1	AIC	SAIC	OPT2	BIC	SBIC
TMS	100	Median	0.776	1.430	1.321	0.645	1.200	0.989
		Mean	3.126	16.839	17.138	1.986	15.070	15.031
		SD	4.321	24.804	25.407	2.578	24.307	24.317
	200	Median	0.293	0.427	0.384	0.265	0.349	0.315
		Mean	0.419	0.816	0.788	0.314	0.667	0.636
		SD	0.566	2.030	2.073	0.277	1.976	1.989
TmMS	100	Median	0.886	1.192	1.033	0.723	0.631	0.609
		Mean	4.096	14.743	13.899	2.569	6.539	6.334
		SD	6.171	24.521	23.055	3.603	17.780	14.476
	200	Median	0.309	0.403	0.343	0.271	0.320	0.267
		Mean	0.497	0.699	0.664	0.377	0.438	0.403
		SD	0.777	1.606	1.519	0.398	0.527	0.549
OMS	100	Median	0.877	0.833	0.883	0.753	0.629	0.622
		Mean	3.655	14.636	12.811	2.256	3.395	3.451
		SD	5.227	24.121	21.154	2.971	10.092	8.898
	200	Median	0.290	0.354	0.313	0.292	0.337	0.295
		Mean	0.487	0.706	0.651	0.383	0.413	0.379
		SD	0.886	1.886	1.823	0.377	0.346	0.327

Table 4. Simulation results for Setting 2: Nonsparse model ($\times 10^{-1}$).

Model Screening	n		OPT1	AIC	SAIC	OPT2	BIC	SBIC
TMS	100	Median	0.780	1.560	1.376	0.685	1.312	1.116
		Mean	3.214	17.509	17.609	2.075	16.227	15.980
		SD	4.190	25.310	25.253	2.530	24.955	24.457
	200	Median	0.289	0.445	0.413	0.268	0.394	0.358
		Mean	0.439	0.922	0.893	0.328	0.770	0.728
		SD	0.646	2.784	2.830	0.307	2.733	2.742
TmMS	100	Median	0.906	1.204	1.147	0.779	0.709	0.679
		Mean	4.137	15.037	14.431	2.675	6.676	6.418
		SD	5.680	24.119	22.669	3.469	16.822	13.331
	200	Median	0.326	0.415	0.362	0.305	0.351	0.294
		Mean	0.533	0.877	0.727	0.409	0.471	0.418
		SD	0.978	2.803	1.904	0.446	0.560	0.452
OMS	100	Median	0.872	1.094	0.983	0.856	0.742	0.668
		Mean	3.848	15.124	13.235	2.413	4.103	4.044
		SD	5.285	24.214	20.755	3.038	11.201	9.868
	200	Median	0.301	0.390	0.340	0.331	0.383	0.336
		Mean	0.510	0.756	0.666	0.418	0.479	0.433
		SD	0.903	2.426	1.705	0.413	0.411	0.387

which means OPT2 is more stable. Different from TmMS and OMS, when using TMS, OPT2 always yields smaller median than the BIC and SBIC. The TMS performs better than the TmMS and OMS, but, as discussed in Section 4, the TMS has serious computational burden when the number of candidate models is very large. In most cases, the OMS outperforms the TmMS, so in the following simulation studies, we only use the OMS to screen models.

For nonsparse situation shown in Table 4, the ordinal rankings of the model selection and averaging methods, generally follow the same pattern as that of the sparse situation shown in Table 3.

Next, we consider a situation that the number of covariates is diverging with sample size. Specifically, we set $\boldsymbol{\beta} = (3, 0, 0, 1.5, 0, 0, 7, 0, 0, 0.2, \ldots, 0.2)^{\mathrm{T}}_{[5.8n^{1/3}]\times 1}$. Other settings are the same as those in the above paragraph. Simulation result is shown in Table 5. It is seen that OPT2, BIC and SBIC perform much better than OPT1, AIC and SAIC, respectively. Again, we find that the BIC and SBIC yield smaller medians than OPT2, but, much larger mean and SD.

Finally, following a referee's suggestion, we explore our model averaging methods with other values of λ_n differing from 2 and $\log(n)$. Specifically, we vary λ_n in a set $\{0.5, 1, \ldots, 6\}$ and set $\boldsymbol{\beta} = (3, 0, 0, 1.5, 0, 0, 7, 0, 0)^{\mathrm{T}}$. Figure 2 shows the relationship between the median of KL loss and λ_n. For $n = 100$ and $n = 200$, the medians of KL loss are minimized at $\lambda_n = 4$ and $\lambda_n = 3$, respectively. When λ_n increases from 0.5 to 2, the medians decrease rapidly, but when λ_n increases from 4 to 6, the medians increase slightly.

Table 5. Simulation results for Setting 2: High-dimensional model ($\times 10^{-1}$).

Model Screening	n		OPT1	AIC	SAIC	OPT2	BIC	SBIC
OMS	100	Median	5.041	35.614	33.185	3.570	1.594	2.337
		Mean	5.376	36.293	34.141	3.832	12.135	11.265
		S.D.	3.142	24.021	21.103	2.324	20.252	16.397
	200	Median	4.570	26.114	24.139	2.801	1.102	1.085
		Mean	4.734	25.846	24.564	2.926	4.135	3.798
		S.D.	1.417	11.742	10.099	0.893	8.715	7.135

Setting 3. Conditional on the random effects $\alpha_{0,i} \sim \mathrm{Normal}(0, 0.5^2)$, we generate the response by $y_{ij} \mid \alpha_{0,i} \sim \mathrm{Binomial}(1, p_{0,ij})$, where $i = 1 \ldots, m$, $j = 1, \ldots, n$, $p_{0,ij} = \exp(\theta_{0,ij})/\{1 + \exp(\theta_{0,ij})\}$, $\theta_{0,ij} = \beta_1 + \beta_2 x_{ij,1} + \beta_3 x_{ij,2}^2 + \beta_4 \exp(x_{ij,3}/3) + \beta_5 \sin(x_{ij,4}) + \varrho_i \alpha_{0,i}$, $(\beta_1, \ldots, \beta_5) = (0.5, -0.2, -0.1, 0.1, -0.2)$, $x_{ij,1}, \ldots, x_{ij,4}$ follow normal distributions with mean zeros, variance ones and the correlation between different components of $\mathbf{x}_{ij}$ being 0.75, and $\varrho_i = 1$ for $i \le m/2$ and 0 otherwise. We combine $2^4 - 1 = 15$ mixed-effects logistic regression candidate models, in which the sth

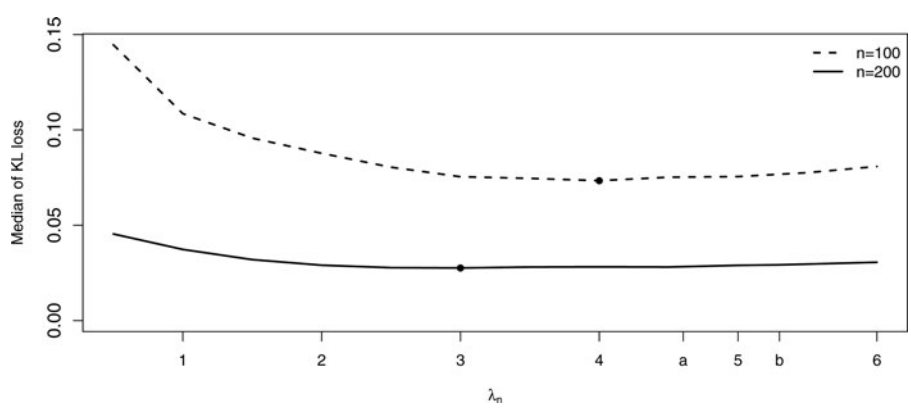

Figure 2. The relationship between the median of KL loss and λ_n in simulation Setting 2. $a = \log(100)$ and $b = \log(200)$. The points with the smallest losses are indicated by the filled circle •.

Table 6. Simulation results for Setting 3 ($\times 10^{-1}$).

Regression	m,n		OPT1	cAIC	ScAIC	OPT2	cBIC	ScBIC
Logistic	8,16	Median	0.419	0.540	0.423	0.388	0.449	0.388
		Mean	0.480	0.616	0.496	0.441	0.535	0.443
		S.D.	0.250	0.355	0.284	0.226	0.308	0.244
	8,20	Median	0.379	0.490	0.401	0.370	0.435	0.371
		Mean	0.424	0.553	0.453	0.402	0.484	0.411
		S.D.	0.200	0.281	0.230	0.177	0.252	0.193
	10,16	Median	0.291	0.397	0.320	0.302	0.372	0.329
		Mean	0.324	0.432	0.363	0.333	0.407	0.361
		S.D.	0.136	0.196	0.165	0.138	0.188	0.157
	10,20	Median	0.283	0.387	0.311	0.299	0.368	0.332
		Mean	0.300	0.407	0.338	0.316	0.392	0.349
		S.D.	0.114	0.158	0.139	0.124	0.161	0.142
Poisson	8,16	Median	1.005	1.214	1.108	1.003	1.232	1.102
		Mean	1.057	1.272	1.156	1.053	1.298	1.162
		S.D.	0.363	0.418	0.398	0.370	0.467	0.406
	8,20	Median	0.925	1.088	0.983	0.929	1.111	1.014
		Mean	0.958	1.116	1.021	0.977	1.156	1.052
		S.D.	0.306	0.362	0.317	0.321	0.390	0.338
	10,16	Median	0.739	0.813	0.770	0.754	0.849	0.799
		Mean	0.746	0.833	0.794	0.770	0.891	0.829
		S.D.	0.182	0.209	0.196	0.193	0.258	0.215
	10,20	Median	0.700	0.759	0.729	0.724	0.815	0.763
		Mean	0.725	0.786	0.755	0.755	0.857	0.794
		S.D.	0.184	0.205	0.189	0.198	0.242	0.215

model:

$$\Pr(y_{ij} = 1 \mid \alpha_i) = \exp\left(\beta_1 + \sum_{\tau \in T_s} \beta_{\tau+1} x_{ij,\tau} + \alpha_i\right) \Bigg/ \left\{1 + \exp\left(\beta_1 + \sum_{\tau \in T_s} \beta_{\tau+1} x_{ij,\tau} + \alpha_i\right)\right\},$$

with T_s being a subset of $\{1, \ldots, 4\}$. Because we are uncertain whether the random effect should be included, we also combine 15 fixed effects logistic regression candidate models with $\alpha_i \equiv 0$. All candidate models are misspecified. Similar to the situation in Setting 1, we also apply this set-up to Poisson$(p_{0,ij})$ with $p_{0,ij} = \exp(\theta_{0,ij})$ to generate responses.

The simulation results under different combinations of (m, n) are shown in Table 6. In all the cases, both OPT1 and OPT2 yield smaller mean, median, and SD values than their competitors: cAIC/ScAIC and cBIC/ScBIC.

In Setting 1, we once explored asymptotic optimality for GLMs numerically. We now investigate this performance for GLMMs as well. We calculate the means of $\mathrm{cKL}(\widetilde{\mathbf{w}})/\inf_{\mathbf{w}\in\mathcal{W}} \mathrm{cKL}(\mathbf{w})$ based on the 500 replications, where $\widetilde{\mathbf{w}}$ is the weight vector determined by OPT1 or OPT2, and sample size $(m, n) \in \{(6, 32), (10, 64), (14, 96), (18, 128), (22, 160)\}$. The simulation results are provided in Figure 3. It is observed that when (m, n) increases, the means decrease and are closer to 1. This finding further echoes the theoretical results in Theorem 4.

7. Real-World Data Examples

7.1. Prediction on Standard and Poor's Credit Ratings Data

The dataset, taken from Compustat, and analyzed by Ashbaugh-Skaife, Collins, and LaFond (2006) and Verbeek (2007), contains observations of Standard and Poor's credit ratings of 921 U.S. firms in 2005, in which there are 435 firms with investment-grade rating and 486 firms with speculative-grade rating. The following independent variables are available: working capital of the firm, which proxies the firm's short-term liquidity; retained earnings and earnings before interest and taxes, which proxy historical profitability and current profitability respectively; book leverage, the ratio of the firm's debt to assets; and log sales volume which proxies the firm's size. We scale the first three of these variables by total assets in our analysis. The interaction terms combined by any two of the above variables are also used as independent variables.

Logistic regression models and TmMS and OMS model screening strategies are used in the prediction. We randomly select 200 observations as training data and leave out the remaining observations as test data. This process is repeated 500 times. To measure the prediction accuracy, we use the following KL-type loss function

$$L_{\mathrm{KL}} = -2n_0^{-1} \sum_{i=1}^{n_0} \log f\left\{y_{\mathrm{test},i} \mid \widehat{\theta}_{\mathrm{test},i}(\widehat{\mathbf{w}}), \phi\right\}, \tag{17}$$

where $\{y_{\mathrm{test},1}, \ldots, y_{\mathrm{test},n_0}\}$ are testing observations, $\widehat{\theta}_{\mathrm{test},i}(\widehat{\mathbf{w}}) = \mathbf{x}_{\mathrm{test},i}^{\mathrm{T}} \widehat{\boldsymbol{\beta}}(\widehat{\mathbf{w}})$, $\mathbf{x}_{\mathrm{test},i}$ is the vector of covariates of the ith testing observation, $\widehat{\boldsymbol{\beta}}(\widehat{\mathbf{w}})$ is obtained based on the training data, and the function f is defined in (1). This loss function finds its root

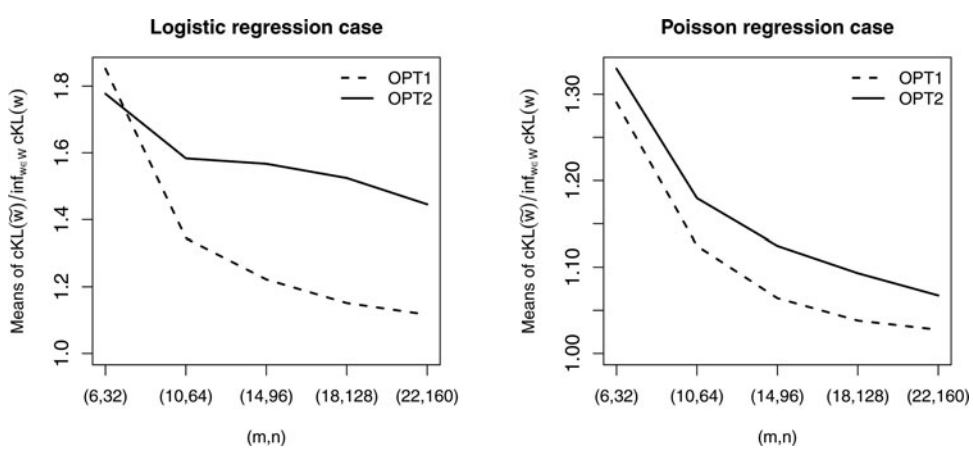

Figure 3. Assessing the asymptotic optimality of OPT1 and OPT2 using simulation Setting 3.

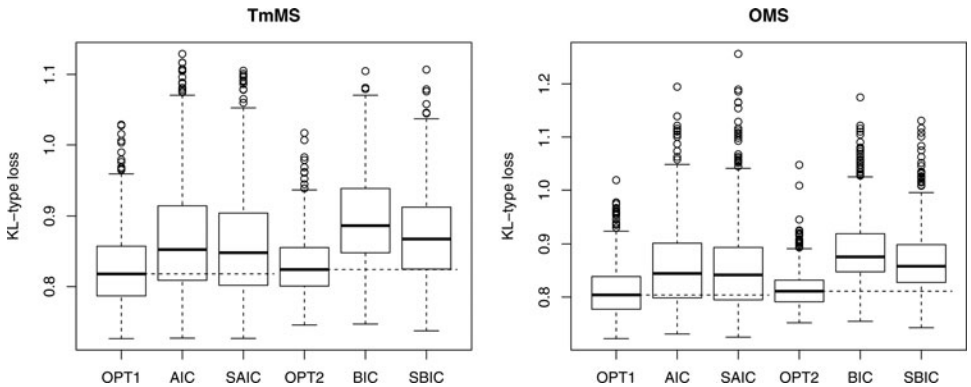

Figure 4. Boxplot of KL-type prediction loss in the Standard and Poor's credit ratings data example. The dashed lines indicate the medians of KL-type prediction loss by the OPT method.

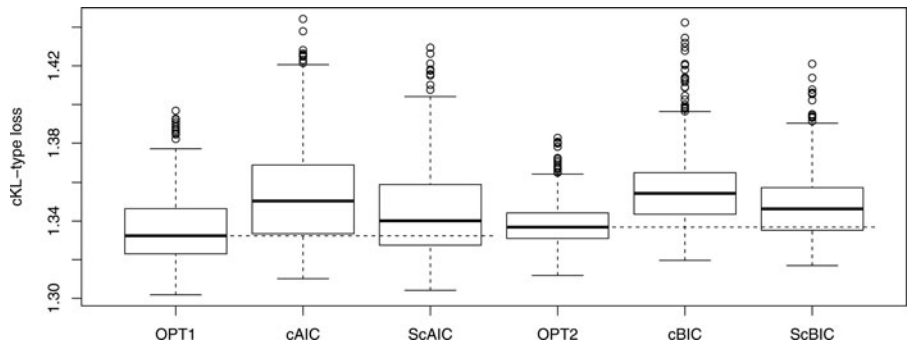

Figure 5. Boxplots of cKL-type prediction loss in European Currency Opinion Survey data example. The dashed lines indicate the medians of cKL-type prediction loss by the OPT methods.

in Akaike information and can be viewed as a measure of predicting error of $\widehat{\boldsymbol{\beta}}(\widehat{\mathbf{w}})$ in the prediction of $\{y_{\text{test},1}, \ldots, y_{\text{test},n_0}\}$.

Figure 4 shows the box plots of all KL-type prediction losses by model selection and averaging methods. No matter using TmMS or OMS, OPT1 and OPT2 always perform better than AIC, SAIC, BIC, and SBIC methods.

7.2. Prediction on European Currency Opinion Survey Data

A subset of data was extracted from the European Currency Opinion Survey conducted by GESIS Leibniz Institute for the Social Sciences. The opinion survey was run on a yearly basis from 2003 to 2007 for 12 European nations with 12078 respondents in total. The purpose of the survey is to study people's perception of European Currency (EUR) (Conflitti 2011). There are four possible choices for response variable: "disadvantageous," "don't know/no answer," "neither one neither the other, no change," and "advantageous." For illustration purpose, we combine the response variable as "1=advantageous" or "0=otherwise." Also, we take a subset of data from six countries (Spain, Belgium, France, Ireland, Germany, and Finland) in 2006 (with 1294 observations in total, and 716 of them has answered "advantageous"). There are five groups of variables: "gender" denoted by $\mathbf{x}_{ij,1}$, "age" (consists of 3 dummy variables denoted by $\mathbf{x}_{ij,2}$), "education" (consists of 3 dummy variables denoted by $\mathbf{x}_{ij,3}$), "occupation" (consists of 3 dummy variables denoted by $\mathbf{x}_{ij,4}$) and "locality" (consists of 2 dummy variables denoted by $\mathbf{x}_{ij,5}$). Depending on these five groups of variables, we consider $2^5 - 1 = 31$ candidate mixed-effects logistic models, the sth of which is defined by

$$\theta_{s,ij} = \beta_0 + \sum_{\tau \in T_s} \mathbf{x}_{ij,\tau}^{\mathrm{T}} \boldsymbol{\beta}_\tau + \alpha_i, \quad i = 1, \ldots, 6, \quad j = 1, \ldots, n_i$$

where $\{n_i, i = 1, \ldots, 6\} = \{215, 213, 228, 226, 170, 242\}$, α_i represents random effect of the ith country, the set T_s contains indexes of variables used in this candidate model, and β_0 and $\boldsymbol{\beta}_\tau$ are fixed effects. We also combine 31 fixed effects logistic regression candidate models because of the uncertainty in determining whether the random effect should be included. In each country, we randomly take 150 observations to compose the test data and the remaining as the training data. This process is repeated 500 times. We use cKL-type loss function similar to (17) to measure the prediction accuracy.

Figure 5 shows the box plots of all cKL-type prediction losses by model selection and averaging methods. It is observed that both OPT1 and OPT2 perform better than their competitors cAIC/ScAIC and cBIC/ScBIC.

8. Concluding Remarks

In this article, we have proposed model averaging methods for GLMs and GLMMs. The resultant methods were shown to be asymptotically optimal in the sense of achieving the lowest KL loss or conditional KL loss. Also, three model screening strategies prior to model averaging are suggested and investigated. Numerical analysis in comparison with existing methods strongly favors our model averaging methods.

In the current article, we have assumed the number of candidate models S to be fixed. When this number increases with n,

stronger conditions may need for establishing asymptotic optimality and the procedure needs further investigation. In addition, we did not study the choice of the tuning parameters in this article. For example in GLMs, although the two versions of our method with $\lambda_n = 2$ and $\lambda_n = \log(n)$ outperform the AIC and BIC, respectively, how to choose between 2 and $\log(n)$, or choose an optimal λ_n is still very challenging and warrants a future investigation.

Appendices A:

A1. Assumptions

Assumption A. We list the regularity conditions required for Theorem 1, where all the limiting properties here and throughout the text hold under $n \to \infty$. Let $\boldsymbol{\epsilon} = (\epsilon_1, \ldots, \epsilon_n)^{\mathrm{T}} = \mathbf{y} - \boldsymbol{\mu}$, $\bar{\sigma}^2 = \max_{i\in\{1,\ldots,n\}} \mathrm{var}(\epsilon_i)$, $\boldsymbol{\beta}^*(\mathbf{w}) = \sum_{s=1}^{S} w_s \boldsymbol{\beta}^*_{(s)}$,

$$\mathrm{KL}^*(\mathbf{w}) = 2\phi^{-1}B\{\boldsymbol{\beta}^*(\mathbf{w})\} - 2\phi^{-1}B_0 - 2\phi^{-1}\boldsymbol{\mu}^{\mathrm{T}}[\boldsymbol{\theta}\{\boldsymbol{\beta}^*(\mathbf{w})\} - \boldsymbol{\theta}_0],$$

and $\xi_n = \inf_{\mathbf{w}\in\mathcal{W}} \mathrm{KL}^*(\mathbf{w})$.

Condition C.1. $\|\mathbf{X}^{\mathrm{T}}\boldsymbol{\mu}\| = O(n)$, $\|\mathbf{X}^{\mathrm{T}}\boldsymbol{\epsilon}\| = O_p(n^{1/2})$, and uniformly for $\mathbf{w} \in \mathcal{W}$,

$$\|\partial B(\boldsymbol{\beta})/\partial\boldsymbol{\beta}^{\mathrm{T}}|_{\boldsymbol{\beta}=\widetilde{\boldsymbol{\beta}}(\mathbf{w})}\| = O_p(n)$$

for every $\widetilde{\boldsymbol{\beta}}(\mathbf{w})$ between $\widehat{\boldsymbol{\beta}}(\mathbf{w})$ and $\boldsymbol{\beta}^*(\mathbf{w})$.

Condition C.2. Uniformly for $s \in \{1, \ldots, S\}$, $n^{-1}\bar{\sigma}^2\|\boldsymbol{\theta}(\boldsymbol{\beta}^*_{(s)})\|^2 = O(1)$.

Condition C.3. $n\xi_n^{-2} = o(1)$.

The formulas in Conditions C.1 and C.2 can be verified from original conditions in the existing literature such as White (1982). This verification is simple, but tedious and out of the focus of the current article. So here we use Conditions C.1 and C.2 directly instead of some more original conditions. Condition C.3 requires that ξ_n grows at a rate no slower than $n^{1/2}$, is the same as the third part of Condition (A7) of Zhang, Zou, and Liang (2014) and is also implied by Conditions (7) and (8) of Ando and Li (2014).

Assumption B. Let $\widehat{\boldsymbol{\beta}}_s \in R^{k_s\times 1}$ be a vector containing estimators in $\widehat{\boldsymbol{\beta}}_{(s)}$ without restriction, $\boldsymbol{\beta}^*_s \in R^{k_s\times 1}$ be the corresponding subvector of $\boldsymbol{\beta}^*_{(s)}$ and $\mathbf{X}_{(s)}$ be the corresponding conformal covariates matrix of size $n \times k_s$ whose rows are $\mathbf{X}^{\mathrm{T}}_{(s),1}, \ldots, \mathbf{X}^{\mathrm{T}}_{(s),n}$. Define

$$\mathcal{B}_n\left(\boldsymbol{\beta}^*_s \mid \delta\right) = \left\{\boldsymbol{\beta}_s \in R^{k_s} : \left\|\frac{n^{1/2}}{k^{1/2}}\left(\boldsymbol{\beta}_s - \boldsymbol{\beta}^*_s\right)\right\| \le \delta\right\}.$$

Let $b^{(t)}(x) = d^t b(x)/dx^t$ and $\mathbf{D}_s = \mathrm{diag}\{b^{(2)}(\mathbf{X}^{\mathrm{T}}_{(s),i}\boldsymbol{\beta}_s)\}_{i=1,\ldots,n}$.

Condition C.4. There exists a constant $C_0 > 0$ such that for any fixed $\delta > 0$, any $\boldsymbol{\beta}_s \in \mathcal{B}_n(\boldsymbol{\beta}^*_s \mid \delta)$ and every $s = 1, \ldots, S$, the minimum eigenvalue of $\mathbf{X}^{\mathrm{T}}_{(s)}\mathbf{D}_s\mathbf{X}_{(s)}/n$ is bounded below by C_0 for all sufficiently large n, and there exists a constant C_1 such that

$$\frac{\sum_{i=1}^{n}\bar{\sigma}^2\,\|\mathbf{X}_i\|^2}{kn} \le C_1 < \infty.$$

Condition C.5. $k^{-1/2}\|\mathbf{X}^{\mathrm{T}}\boldsymbol{\mu}\| = O(n)$ and uniformly for $\mathbf{w} \in \mathcal{W}$,

$$k^{-1/2}\|\partial B(\boldsymbol{\beta})/\partial\boldsymbol{\beta}^{\mathrm{T}}|_{\boldsymbol{\beta}=\widetilde{\boldsymbol{\beta}}(\mathbf{w})}\| = O_p(n)$$

for every $\widetilde{\boldsymbol{\beta}}(\mathbf{w})$ between $\widehat{\boldsymbol{\beta}}(\mathbf{w})$ and $\boldsymbol{\beta}^*(\mathbf{w})$.

Condition C.6. $k^2 n\xi_n^{-2} = o(1)$.

Condition C.4 is used to prove that $\|\widehat{\boldsymbol{\beta}}_s - \boldsymbol{\beta}^*_s\|$ converges to zero at certain rate and is similar to Conditions (A.1′) and (A.2′) of Flynn, Hurvich, and Simonoff (2013) and Conditions 1 and 2 of Lv and Liu (2014). Conditions C.5 and C.6 are extensions of C.1 and C.3 under the diverging k situation. Note that we allow k to increase with n but cannot arbitrarily. For instance, if ξ_n has order n, then Condition C.6 implies $k = o(n^{1/2})$.

Assumption C. From Yu, Zhang, and Yau (2014), we know that under certain regularity conditions there exists a unique $\boldsymbol{\gamma}^*_{(s)} = ({\boldsymbol{\beta}^*_{(s)}}^{\mathrm{T}}, {\boldsymbol{\alpha}^*_{(s)}}^{\mathrm{T}})^{\mathrm{T}} = ({\boldsymbol{\beta}^*_{(s)}}^{\mathrm{T}}, {\boldsymbol{\alpha}^*_{(s),1}}^{\mathrm{T}}, \ldots, {\boldsymbol{\alpha}^*_{(s),m}}^{\mathrm{T}})^{\mathrm{T}}$ which minimizes the conditional Kullback–Leibler divergence defined in (12). Let $\boldsymbol{\theta}^*(\mathbf{w}) = \sum_{s=1}^{S} w_s\mathbf{U}\boldsymbol{\gamma}^*_{(s)}$, $\mathrm{cKL}^*(\mathbf{w}) = 2\phi^{-1}[\boldsymbol{\mu}^{\mathrm{T}}\boldsymbol{\theta}_0 - B(\boldsymbol{\theta}_0) - \boldsymbol{\mu}^{\mathrm{T}}\boldsymbol{\theta}^*(\mathbf{w}) + B\{\boldsymbol{\theta}^*(\mathbf{w})\}]$, $\xi_N = \inf_{\mathbf{w}\in\mathcal{W}} \mathrm{cKL}^*(\mathbf{w})$, and $p(\boldsymbol{\theta}_0)$ be the probability density function of $\boldsymbol{\theta}_0$. We list the regularity conditions required for Theorem 4.

Condition C.7. All the elements in the design matrices $\mathbf{X}$ and $\mathbf{Z}$ are dominated by a positive constant uniformly.

Condition C.8. For every $\widetilde{\boldsymbol{\theta}}(\mathbf{w}) = (\widetilde{\theta}_{\mathbf{w},11}, \ldots, \widetilde{\theta}_{\mathbf{w},mn_m})^{\mathrm{T}}$ lying between $\widehat{\boldsymbol{\theta}}(\mathbf{w})$ and $\boldsymbol{\theta}^*(\mathbf{w})$,

$$\sup_{\mathbf{w}\in\mathcal{W}} \frac{1}{N}\sum_{i=1}^{m}\sum_{j=1}^{n_i}\left|b^{(1)}\left(\widetilde{\theta}_{\mathbf{w},ij}\right)\right|^2 = O_p(1).$$

Condition C.9. There exists a random variable $\bar{\theta}_4$ such that for every $1 \le s \le S$

$$\frac{\sum_{i=1}^{m}\sum_{j=1}^{n_i}\left|\theta^*_{s,ij}\right|^4}{N} \le \bar{\theta}_4 \tag{A.1}$$

holds $p(\boldsymbol{\theta}_0)$-a.s. and $E_{\boldsymbol{\theta}_0}|\bar{\theta}_4| < \infty$. There exists a function $\bar{\mu}(\boldsymbol{\theta}_0)$ such that

$$\frac{1}{N}\sum_{i=1}^{m}\sum_{j=1}^{n_i}\left|\mu_{0,ij}\right|^2 \le \bar{\mu}(\boldsymbol{\theta}_0), \tag{A.2}$$

holds $p(\boldsymbol{\theta}_0)$-a.s. and $E_{\boldsymbol{\theta}_0}|\bar{\mu}(\boldsymbol{\theta}_0)| < \infty$. Let $\sigma^2_{ij}(\theta_{0,ij}) = E_{y_{ij}|\boldsymbol{\theta}_0}(y_{ij} - \mu_{0,ij})^2$. There exists a function $\bar{\sigma}^2(\boldsymbol{\theta}_0)$ such that

$$\frac{\sum_{i=1}^{m}\sum_{j=1}^{n_i}\left|\sigma^2_{ij}(\theta_{0,ij})\right|^2}{N} \le \bar{\sigma}^4(\boldsymbol{\theta}_0), \tag{A.3}$$

holds $p(\boldsymbol{\theta}_0)$-a.s. and $E_{\boldsymbol{\theta}_0}\bar{\sigma}^4(\boldsymbol{\theta}_0) < \infty$.

Condition C.10. For every $s \in \{1, \ldots, S\}$,

$$E\left\|N^{1/2}\left(\widehat{\boldsymbol{\beta}}_{(s)} - \boldsymbol{\beta}^*_{(s)}\right)\right\|^2 = O(1)$$

and

$$\max_{1\le i\le m} E\left\| n_i^{1/2}\left(\widehat{\boldsymbol{\alpha}}_{(s),i}-\boldsymbol{\alpha}^*_{(s),i}\right)\right\|^2 = O(1).$$

Condition C.11. $N^2 n_*^{-1}\xi_N^{-2} = o_p(1)$.

Conditions C.7–C.10 play a similar role to Conditions C.1 and C.2 in Theorem 1. Condition C.11 can be viewed as an analogous version of C.3 in the context of GLMMs. To be more special, Conditions C.7 and C.8 play similar role as $\|\partial B(\boldsymbol{\beta})/\partial\boldsymbol{\beta}^{\mathrm{T}}|_{\boldsymbol{\beta}=\widetilde{\boldsymbol{\beta}}(\mathbf{w})}\| = O_p(n)$ in Condition C.1. Condition C.7 and (A.2) in Condition C.9 are analogous to $\|\mathbf{X}^{\mathrm{T}}\boldsymbol{\mu}\| = O(n)$. Condition C.7 and (A.3) in Condition C.9 can be viewed as a variant of $\|\boldsymbol{\epsilon}^{\mathrm{T}}\boldsymbol{\mu}\| = O_p(n)$. In addition, (A.1) and (A.3) can be viewed as the extended version of Condition C.2 in the presence of random effects. Condition C.10 guarantees that $N^{1/2}(\widehat{\boldsymbol{\beta}}_{(s)}-\boldsymbol{\beta}^*_{(s)})$ and $n_i^{1/2}(\widehat{\boldsymbol{\alpha}}_{(s),i}-\boldsymbol{\alpha}^*_{(s),i})$ have finite second-order moment. Similar condition can also be found in the model selection literatures (Lv and Liu 2014; Yu, Zhang, and Yau 2014; Donohue et al. 2011).

A2. Proof of Theorem 1

Let $\widetilde{\mathcal{G}}(\mathbf{w}) = \mathcal{G}(\mathbf{w}) - 2\phi^{-1}B_0 + 2\phi^{-1}\boldsymbol{\mu}^{\mathrm{T}}\boldsymbol{\theta}_0$. It is obvious that $\widehat{\mathbf{w}} = \operatorname{argmin}_{\mathbf{w}\in\mathcal{W}}\widetilde{\mathcal{G}}(\mathbf{w})$. From the proof of Theorem 1' in Wan, Zhang, and Zou (2010), Theorem 1 is valid if the following holds:

$$\sup_{\mathbf{w}\in\mathcal{W}}\frac{|\mathrm{KL}(\mathbf{w})-\mathrm{KL}^*(\mathbf{w})|}{\mathrm{KL}^*(\mathbf{w})} = o_p(1) \tag{A.4}$$

and

$$\sup_{\mathbf{w}\in\mathcal{W}}\frac{|\widetilde{\mathcal{G}}(\mathbf{w})-\mathrm{KL}^*(\mathbf{w})|}{\mathrm{KL}^*(\mathbf{w})} = o_p(1). \tag{A.5}$$

By (6), we know that uniformly for $\mathbf{w}\in\mathcal{W}$,

$$\widehat{\boldsymbol{\beta}}(\mathbf{w})-\boldsymbol{\beta}^*(\mathbf{w}) = \sum_{s=1}^{S} w_s(\widehat{\boldsymbol{\beta}}_{(s)}-\boldsymbol{\beta}^*_{(s)}) = O_p(n^{-1/2}). \tag{A.6}$$

It follows from (A.6), Condition C.1, and Taylor expansion that uniformly for $\mathbf{w}\in\mathcal{W}$,

$$\begin{aligned}\left|B\left\{\widehat{\boldsymbol{\beta}}(\mathbf{w})\right\}-B\left\{\boldsymbol{\beta}^*(\mathbf{w})\right\}\right| &\le \left\|\frac{\partial B(\boldsymbol{\beta})}{\partial\boldsymbol{\beta}^{\mathrm{T}}}|_{\boldsymbol{\beta}=\widetilde{\boldsymbol{\beta}}(\mathbf{w})}\right\|\left\|\widehat{\boldsymbol{\beta}}(\mathbf{w})-\boldsymbol{\beta}^*(\mathbf{w})\right\| \\ &= O_p(n^{1/2}),\\ \boldsymbol{\mu}^{\mathrm{T}}\left[\boldsymbol{\theta}\left\{\widehat{\boldsymbol{\beta}}(\mathbf{w})\right\}-\boldsymbol{\theta}\left\{\boldsymbol{\beta}^*(\mathbf{w})\right\}\right] &\le \|\boldsymbol{\mu}^{\mathrm{T}}\mathbf{X}\|\left\|\widehat{\boldsymbol{\beta}}(\mathbf{w})-\boldsymbol{\beta}^*(\mathbf{w})\right\| \\ &= O_p(n^{1/2})\end{aligned}$$

and

$$\boldsymbol{\epsilon}^{\mathrm{T}}[\boldsymbol{\theta}\{\widehat{\boldsymbol{\beta}}(\mathbf{w})\}-\boldsymbol{\theta}\{\boldsymbol{\beta}^*(\mathbf{w})\}] \le \|\boldsymbol{\epsilon}^{\mathrm{T}}\mathbf{X}\|\|\widehat{\boldsymbol{\beta}}(\mathbf{w})-\boldsymbol{\beta}^*(\mathbf{w})\| = O_p(1),$$

where $\widetilde{\boldsymbol{\beta}}(\mathbf{w})$ is a vector between $\widehat{\boldsymbol{\beta}}(\mathbf{w})$ and $\boldsymbol{\beta}^*(\mathbf{w})$. In addition, using the central limit theorem and Condition C.2, we know that uniformly for $\mathbf{w}\in\mathcal{W}$,

$$\boldsymbol{\epsilon}^{\mathrm{T}}\boldsymbol{\theta}\{\boldsymbol{\beta}^*(\mathbf{w})\} = \sum_{s=1}^{S} w_s\boldsymbol{\epsilon}^{\mathrm{T}}\boldsymbol{\theta}(\boldsymbol{\beta}^*_{(s)}) = O_p(n^{1/2}). \tag{A.7}$$

These arguments indicate that

$$\begin{aligned}&\sup_{\mathbf{w}\in\mathcal{W}}|\mathrm{KL}(\mathbf{w})-\mathrm{KL}^*(\mathbf{w})| \\ &\quad\le 2\phi^{-1}\sup_{\mathbf{w}\in\mathcal{W}}|B\{\widehat{\boldsymbol{\beta}}(\mathbf{w})\}-B\{\boldsymbol{\beta}^*(\mathbf{w})\}| \\ &\qquad +2\phi^{-1}\sup_{\mathbf{w}\in\mathcal{W}}|\boldsymbol{\mu}^{\mathrm{T}}[\boldsymbol{\theta}\{\widehat{\boldsymbol{\beta}}(\mathbf{w})\}-\boldsymbol{\theta}\{\boldsymbol{\beta}^*(\mathbf{w})\}]| \\ &\quad= O_p(n^{1/2})\end{aligned} \tag{A.8}$$

and

$$\begin{aligned}&\sup_{\mathbf{w}\in\mathcal{W}}\left|\widetilde{\mathcal{G}}(\mathbf{w})-\mathrm{KL}^*(\mathbf{w})\right| \le 2\phi^{-1}\sup_{\mathbf{w}\in\mathcal{W}}\left|B\left\{\widehat{\boldsymbol{\beta}}(\mathbf{w})\right\}-B\left\{\boldsymbol{\beta}^*(\mathbf{w})\right\}\right| \\ &\quad+2\phi^{-1}\sup_{\mathbf{w}\in\mathcal{W}}\left|\mathbf{y}^{\mathrm{T}}\boldsymbol{\theta}\left\{\widehat{\boldsymbol{\beta}}(\mathbf{w})\right\}-\boldsymbol{\mu}^{\mathrm{T}}\boldsymbol{\theta}\left\{\boldsymbol{\beta}^*(\mathbf{w})\right\}\right|+\lambda_n\mathbf{w}^{\mathrm{T}}\mathbf{k} \\ &\le 2\phi^{-1}\sup_{\mathbf{w}\in\mathcal{W}}\left|B\left\{\widehat{\boldsymbol{\beta}}(\mathbf{w})\right\}-B\{\boldsymbol{\beta}^*(\mathbf{w})\}\right| \\ &\quad+2\phi^{-1}\sup_{\mathbf{w}\in\mathcal{W}}\left|\boldsymbol{\mu}^{\mathrm{T}}\left[\boldsymbol{\theta}\left\{\widehat{\boldsymbol{\beta}}(\mathbf{w})\right\}-\boldsymbol{\theta}\left\{\boldsymbol{\beta}^*(\mathbf{w})\right\}\right]\right| \\ &\quad+2\phi^{-1}\sup_{\mathbf{w}\in\mathcal{W}}\left|\boldsymbol{\epsilon}^{\mathrm{T}}\boldsymbol{\theta}\left\{\boldsymbol{\beta}^*(\mathbf{w})\right\}\right| \\ &\quad+2\phi^{-1}\sup_{\mathbf{w}\in\mathcal{W}}\left|\boldsymbol{\epsilon}^{\mathrm{T}}\left[\boldsymbol{\theta}\left\{\widehat{\boldsymbol{\beta}}(\mathbf{w})\right\}-\boldsymbol{\theta}\left\{\boldsymbol{\beta}^*(\mathbf{w})\right\}\right]\right|+\lambda_n\mathbf{w}^{\mathrm{T}}\mathbf{k} \\ &= O_p(n^{1/2})+\lambda_n\mathbf{w}^{\mathrm{T}}\mathbf{k}.\end{aligned} \tag{A.9}$$

Now, from (A.8) and (A.9), $n\xi_n^{-2} = o(1)$, and $n^{-1/2}\lambda_n = O(1)$, we can obtain (A.4) and (A.5). This completes the proof.

A3. Proof of Theorem 2

We first show that for any fixed $\varepsilon > 0$, there exists a $\delta_\varepsilon > 0$ such that for all sufficiently large n

$$P\left(\left\|\frac{n^{1/2}}{k^{1/2}}(\widehat{\boldsymbol{\beta}}_s-\boldsymbol{\beta}^*_s)\right\|\le\delta_\varepsilon\right)\ge 1-\varepsilon. \tag{A.10}$$

Write $\boldsymbol{\mu}^*_s = \{b^{(1)}(\mathbf{X}^{\mathrm{T}}_{(s),1}\boldsymbol{\beta}^*_s),\ldots,b^{(1)}(\mathbf{X}^{\mathrm{T}}_{(s),n}\boldsymbol{\beta}^*_s)\}^{\mathrm{T}}$. The quasi-true value $\boldsymbol{\beta}^*_s$ minimizes the KL divergence so that $\partial\{B(\boldsymbol{\beta}_s)-\boldsymbol{\mu}^{\mathrm{T}}\boldsymbol{\theta}(\boldsymbol{\beta}_s)\}/\partial\boldsymbol{\beta}_s\,|_{\boldsymbol{\beta}_s=\boldsymbol{\beta}^*_s} = 0_{k_s\times 1}$, which implies that $\mathbf{X}^{\mathrm{T}}_{(s)}\boldsymbol{\mu} = \mathbf{X}^{\mathrm{T}}_{(s)}\boldsymbol{\mu}^*_s$. Therefore, by using the first-order Taylor expansion of $\partial\log f(\mathbf{y}\mid\mathbf{X}_{(s)}\boldsymbol{\beta}_s,\phi)/\partial\boldsymbol{\beta}_s = \mathbf{0}_{k_s\times 1}$ at $\boldsymbol{\beta}^*_s$, we obtain that

$$\frac{n^{1/2}}{k^{1/2}}\left(\widehat{\boldsymbol{\beta}}_s-\boldsymbol{\beta}^*_s\right) = \left\{\frac{1}{n}\mathbf{X}^{\mathrm{T}}_{(s)}\mathbf{D}_s\mathbf{X}_{(s)}\,|_{\boldsymbol{\beta}_s=\widetilde{\boldsymbol{\beta}}_s}\right\}^{-1}\frac{\mathbf{X}^{\mathrm{T}}_{(s)}(\mathbf{y}-\boldsymbol{\mu})}{k^{1/2}n^{1/2}},$$

where $\widetilde{\boldsymbol{\beta}}_s$ lies between $\widehat{\boldsymbol{\beta}}_s$ and $\boldsymbol{\beta}^*_s$. Then with Condition C.4, it follows that for sufficiently large n

$$\begin{aligned}P\left(\left\|\frac{n^{1/2}}{k^{1/2}}\left(\widehat{\boldsymbol{\beta}}_s-\boldsymbol{\beta}^*_s\right)\right\|\le\delta\right) &\ge P\left(C_0^{-1}\left\|\frac{\mathbf{X}^{\mathrm{T}}_{(s)}(\mathbf{y}-\boldsymbol{\mu})}{k^{1/2}n^{1/2}}\right\|\le\delta\right) \\ &\ge 1-\frac{\sum_{i=1}^n\left\|\mathbf{X}_{(s),i}\right\|^2\operatorname{var}(\epsilon_i)}{C_0^2\delta^2 nk} \\ &\ge 1-\frac{\sum_{i=1}^n\|\mathbf{X}_i\|^2\operatorname{var}(\epsilon_i)}{C_0^2\delta^2 nk} \\ &\ge 1-\frac{C_1}{C_0^2\delta^2},\end{aligned}$$

and thus (A.10) is proved by taking $\delta = \delta_\varepsilon = C_1^{1/2}/(\varepsilon^{1/2}C_0)$.

From (A.10), we have

$$\left\|\widehat{\boldsymbol{\beta}}_s - \boldsymbol{\beta}_s^*\right\| = \left\|\widehat{\boldsymbol{\beta}}_{(s)} - \boldsymbol{\beta}_{(s)}^*\right\| = O_p(k^{1/2}n^{-1/2}),$$

and thus

$$\|\widehat{\boldsymbol{\beta}}(\mathbf{w}) - \boldsymbol{\beta}^*(\mathbf{w})\| \le \sum_{s=1}^{S} w_s \|\widehat{\boldsymbol{\beta}}_{(s)} - \boldsymbol{\beta}_{(s)}^*\| = O_p(k^{1/2}n^{-1/2}). \tag{A.11}$$

By the proof of Theorem 1, (A.11), and Condition C.5, we have that uniformly for $\mathbf{w} \in \mathcal{W}$,

$$\|B\{\widehat{\boldsymbol{\beta}}(\mathbf{w})\} - B\{\boldsymbol{\beta}^*(\mathbf{w})\}\| = O_p(kn^{1/2}) \tag{A.12}$$

and

$$\boldsymbol{\mu}^{\mathrm{T}}[\boldsymbol{\theta}\{\widehat{\boldsymbol{\beta}}(\mathbf{w})\} - \boldsymbol{\theta}\{\boldsymbol{\beta}^*(\mathbf{w})\}] = O_p(kn^{1/2}). \tag{A.13}$$

In addition, the second part of Condition C.4 also implies that $\|\mathbf{X}^{\mathrm{T}}\boldsymbol{\epsilon}\| = O_p(k^{1/2}n^{1/2})$, which along with (A.11) implies

$$\boldsymbol{\epsilon}^{\mathrm{T}}[\boldsymbol{\theta}\{\widehat{\boldsymbol{\beta}}(\mathbf{w})\} - \boldsymbol{\theta}\{\boldsymbol{\beta}^*(\mathbf{w})\}] = O_p(k). \tag{A.14}$$

Then, using (A.7), (A.12), (A.13), (A.14) and the steps in (A.8) and (A.9), we have $\sup_{\mathbf{w}\in\mathcal{W}} |\mathrm{KL}(\mathbf{w}) - \mathrm{KL}^*(\mathbf{w})| = O_p(kn^{1/2})$ and $\sup_{\mathbf{w}\in\mathcal{W}} |\widetilde{\mathcal{G}}(\mathbf{w}) - \mathrm{KL}^*(\mathbf{w})| = O_p(kn^{1/2}) + \lambda_n \mathbf{w}^{\mathrm{T}}\mathbf{k}$, which, together with Condition C.6 and $\lambda_n n^{-1/2} = O(1)$, imply (A.4) and (A.5), and thus we can obtain (7).

A4. Proof of Theorem 3

Let $a(\mathbf{w}) = \widetilde{\mathcal{G}}(\mathbf{w}) - \mathrm{KL}(\mathbf{w})$. Following (A.4) and (A.5), Assumption 0, and Condition C.6, we have

$$\sup_{\mathbf{w}\in\mathcal{W}} \left|\frac{a(\mathbf{w})}{\mathrm{KL}^*(\mathbf{w})}\right| = o_p(1), \tag{A.15}$$

$$\sup_{\mathbf{w}\in\mathcal{W}} \left|\frac{\nu_n}{\mathrm{KL}^*(\mathbf{w})}\right| = o_p(1), \tag{A.16}$$

$$\begin{aligned}\sup_{\mathbf{w}\in\mathcal{W}} \left|\frac{\mathrm{KL}^*(\mathbf{w})}{\mathrm{KL}(\mathbf{w})}\right| &= \left\{\inf_{\mathbf{w}\in\mathcal{W}} \left|\frac{\mathrm{KL}(\mathbf{w})}{\mathrm{KL}^*(\mathbf{w})}\right|\right\}^{-1} \\ &\le \left\{1 - \sup_{\mathbf{w}\in\mathcal{W}} \left|\frac{\mathrm{KL}(\mathbf{w}) - \mathrm{KL}^*(\mathbf{w})}{\mathrm{KL}^*(\mathbf{w})}\right|\right\}^{-1} \to 1\end{aligned} \tag{A.17}$$

in probability, and

$$\begin{aligned}\sup_{\mathbf{w}\in\mathcal{W}} \left|\frac{\mathrm{KL}^*(\mathbf{w})}{\mathrm{KL}(\mathbf{w}) - \nu_n}\right| &= \left\{\inf_{\mathbf{w}\in\mathcal{W}} \left|\frac{\mathrm{KL}(\mathbf{w}) - \nu_n}{\mathrm{KL}^*(\mathbf{w})}\right|\right\}^{-1} \\ &\le \left\{1 - \sup_{\mathbf{w}\in\mathcal{W}} \left|\frac{\mathrm{KL}(\mathbf{w}) - \mathrm{KL}^*(\mathbf{w})}{\mathrm{KL}^*(\mathbf{w})}\right| \right. \\ &\quad \left. - \sup_{\mathbf{w}\in\mathcal{W}} \left|\frac{\nu_n}{\mathrm{KL}^*(\mathbf{w})}\right|\right\}^{-1} \\ &\to 1\end{aligned} \tag{A.18}$$

in probability, as $n \to \infty$. It follows from Assumption 0 and (A.15)–(A.18) that for any $\delta > 0$,

$$\begin{aligned}
&P\left\{\left|\frac{\inf_{\mathbf{w}\in\mathcal{W}} \mathrm{KL}(\mathbf{w})}{\mathrm{KL}(\widehat{\mathbf{w}}^*)} - 1\right| > \delta\right\} \\
&= P\left\{\left|\frac{\mathrm{KL}(\widehat{\mathbf{w}}^*) - \inf_{\mathbf{w}\in\mathcal{W}} \mathrm{KL}(\mathbf{w})}{\mathrm{KL}(\widehat{\mathbf{w}}^*)}\right| > \delta\right\} \\
&= P\left[\left|\frac{\inf_{\mathbf{w}\in\mathcal{W}^*}\{\mathrm{KL}(\mathbf{w}) + a(\mathbf{w})\} - a(\widehat{\mathbf{w}}^*) - \inf_{\mathbf{w}\in\mathcal{W}} \mathrm{KL}(\mathbf{w})}{\mathrm{KL}(\widehat{\mathbf{w}}^*)}\right| > \delta\right] \\
&= P\left[\left|\frac{\inf_{\mathbf{w}\in\mathcal{W}^*}\{\mathrm{KL}(\mathbf{w}) + a(\mathbf{w})\} - a(\widehat{\mathbf{w}}^*) - \inf_{\mathbf{w}\in\mathcal{W}} \mathrm{KL}(\mathbf{w})}{\mathrm{KL}(\widehat{\mathbf{w}}^*)}\right| \right. \\
&\quad \left. > \delta, \mathbf{w}_n \in \mathcal{W}^*\right] \\
&\quad + P\left[\left|\frac{\inf_{\mathbf{w}\in\mathcal{W}^*}\{\mathrm{KL}(\mathbf{w}) + a(\mathbf{w})\} - a(\widehat{\mathbf{w}}^*) - \inf_{\mathbf{w}\in\mathcal{W}} \mathrm{KL}(\mathbf{w})}{\mathrm{KL}(\widehat{\mathbf{w}}^*)}\right| \right. \\
&\quad \left. > \delta, \mathbf{w}_n \notin \mathcal{W}^*\right] \\
&\le P\left[\left|\frac{\inf_{\mathbf{w}\in\mathcal{W}^*}\{\mathrm{KL}(\mathbf{w}) + a(\mathbf{w})\} - a(\widehat{\mathbf{w}}^*) - \inf_{\mathbf{w}\in\mathcal{W}} \mathrm{KL}(\mathbf{w})}{\mathrm{KL}(\widehat{\mathbf{w}}^*)}\right| \right. \\
&\quad \left. > \delta \mid \mathbf{w}_n \in \mathcal{W}^*\right] P(\mathbf{w}_n \in \mathcal{W}^*) + P(\mathbf{w}_n \notin \mathcal{W}^*) \\
&\le P\left\{\left|\frac{\mathrm{KL}(\mathbf{w}_n) + a(\mathbf{w}_n) - a(\widehat{\mathbf{w}}^*) - \inf_{\mathbf{w}\in\mathcal{W}} \mathrm{KL}(\mathbf{w})}{\mathrm{KL}(\widehat{\mathbf{w}}^*)}\right| \right. \\
&\quad \left. > \delta \mid \mathbf{w}_n \in \mathcal{W}^*\right\} P(\mathbf{w}_n \in \mathcal{W}^*) + P(\mathbf{w}_n \notin \mathcal{W}^*) \\
&\le P\left\{\left|\frac{\mathrm{KL}(\mathbf{w}_n) + a(\mathbf{w}_n) - a(\widehat{\mathbf{w}}^*) - \inf_{\mathbf{w}\in\mathcal{W}} \mathrm{KL}(\mathbf{w})}{\mathrm{KL}(\widehat{\mathbf{w}}^*)}\right| > \delta\right\} \\
&\quad + P(\mathbf{w}_n \notin \mathcal{W}^*) \\
&= P\left\{\left|\frac{\mathrm{KL}(\mathbf{w}_n) + a(\mathbf{w}_n) - a(\widehat{\mathbf{w}}^*) - \mathrm{KL}(\mathbf{w}_n) + \nu_n}{\mathrm{KL}(\widehat{\mathbf{w}}^*)}\right| > \delta\right\} \\
&\quad + P(\mathbf{w}_n \notin \mathcal{W}^*) \\
&\le P\left\{\left|\frac{a(\mathbf{w}_n)}{\mathrm{KL}(\widehat{\mathbf{w}}^*)}\right| + \left|\frac{a(\widehat{\mathbf{w}}^*)}{\mathrm{KL}(\widehat{\mathbf{w}}^*)}\right| + \left|\frac{\nu_n}{\mathrm{KL}(\widehat{\mathbf{w}}^*)}\right| > \delta\right\} + P(\mathbf{w}_n \notin \mathcal{W}^*) \\
&\le P\left\{\left|\frac{a(\mathbf{w}_n)}{\inf_{\mathbf{w}\in\mathcal{W}} \mathrm{KL}(\mathbf{w})}\right| + \sup_{\mathbf{w}\in\mathcal{W}}\left|\frac{a(\mathbf{w})}{\mathrm{KL}(\mathbf{w})}\right| + \sup_{\mathbf{w}\in\mathcal{W}}\left|\frac{\nu_n}{\mathrm{KL}(\mathbf{w})}\right| > \delta\right\} \\
&\quad + P(\mathbf{w}_n \notin \mathcal{W}^*) \\
&\le P\left\{\left|\frac{a(\mathbf{w}_n)}{\mathrm{KL}(\mathbf{w}_n) - \nu_n}\right| + \sup_{\mathbf{w}\in\mathcal{W}}\left|\frac{a(\mathbf{w})}{\mathrm{KL}^*(\mathbf{w})}\right| \sup_{\mathbf{w}\in\mathcal{W}}\left|\frac{\mathrm{KL}^*(\mathbf{w})}{\mathrm{KL}(\mathbf{w})}\right| \right. \\
&\quad \left. + \sup_{\mathbf{w}\in\mathcal{W}}\left|\frac{\nu_n}{\mathrm{KL}^*(\mathbf{w})}\right| \sup_{\mathbf{w}\in\mathcal{W}}\left|\frac{\mathrm{KL}^*(\mathbf{w})}{\mathrm{KL}(\mathbf{w})}\right| > \delta\right\} + P(\mathbf{w}_n \notin \mathcal{W}^*) \\
&\le P\left\{\sup_{\mathbf{w}\in\mathcal{W}}\left|\frac{a(\mathbf{w})}{\mathrm{KL}^*(\mathbf{w})}\right| \sup_{\mathbf{w}\in\mathcal{W}}\left|\frac{\mathrm{KL}^*(\mathbf{w})}{\mathrm{KL}(\mathbf{w}) - \nu_n}\right| \right. \\
&\quad + \sup_{\mathbf{w}\in\mathcal{W}}\left|\frac{a(\mathbf{w})}{\mathrm{KL}^*(\mathbf{w})}\right| \sup_{\mathbf{w}\in\mathcal{W}}\left|\frac{\mathrm{KL}^*(\mathbf{w})}{\mathrm{KL}(\mathbf{w})}\right| \\
&\quad \left. + \sup_{\mathbf{w}\in\mathcal{W}}\left|\frac{\nu_n}{\mathrm{KL}^*(\mathbf{w})}\right| \sup_{\mathbf{w}\in\mathcal{W}}\left|\frac{\mathrm{KL}^*(\mathbf{w})}{\mathrm{KL}(\mathbf{w})}\right| > \delta\right\} + P(\mathbf{w}_n \notin \mathcal{W}^*) \\
&\to 0,
\end{aligned} \tag{A.19}$$

as $n \to \infty$, which implies (9).

A5. An Example of Connecting a Nonexponential Family Distribution with an Exponential Family Distribution via a Hierarchical Framework

Suppose y_i comes from a negative binomial distribution with probability density function $f(y_i) = p_i^{y_i}(1-p_i)^r\Gamma(r+y_i)/\{y_i!\Gamma(1/2)\}$, where $\Gamma(\cdot)$ is the Gamma function, p_i is the success probability in each experiment, and r is the number of failures until the experiment is stopped. It is well known that negative binomial distribution does not belong to exponential family. But if we introduce the random canonical parameters vector, then we can build a connection between the negative binomial distribution and a Poisson distribution via a hierarchical framework. Specifically, let $\theta_i = \log(\lambda_i)$, $\lambda_i \sim \text{Gamma}\{r, p_i/(1-p_i)\}$ and conditionally on θ_i, $y_i \mid \theta_i \sim \text{Poisson}(\lambda_i)$. Then, one can verify that the marginal distribution of y_i follows the negative binomial distribution function.

In addition, by the connection discussed above, when the negative binomial response comes from the hierarchical framework and we use Poisson$\{\exp(\theta_i)\}$ to fit the data, following (13), we can define a loss function as

$$2\left(\sum_{i=1}^{n}\lambda_i\left\{\log(\lambda_i)-\widehat{\theta}_i\right\}+\sum_{i=1}^{n}\left[\exp(\widehat{\theta}_i)-\lambda_i\right]\right),$$

where $\widehat{\theta}_i$ is the estimator of θ_i in fitting the data by Poisson$\{\exp(\theta_i)\}$.

A6. Proof of Theorem 4

In Sections A.6 and A.7, all the limiting processes hold with respect to $n_* \to \infty$ and that m remains bounded or grows as a function of n_*.

Let $\widehat{\boldsymbol{\beta}}_{\mathbf{w}} = \sum_{s=1}^{S} w_s\widehat{\boldsymbol{\beta}}_{(s)}$, $\boldsymbol{\beta}^*_{\mathbf{w}} = \sum_{s=1}^{S} w_s\boldsymbol{\beta}^*_{(s)}$, $\widehat{\boldsymbol{\alpha}}_{\mathbf{w},i} = \sum_{s=1}^{S} w_s\widehat{\boldsymbol{\alpha}}_{(s),i}$ and $\boldsymbol{\alpha}^*_{\mathbf{w},i} = \sum_{s=1}^{S} w_s\boldsymbol{\alpha}^*_{(s),i}$ such that

$$\widehat{\boldsymbol{\beta}}_{\mathbf{w}} - \boldsymbol{\beta}^*_{\mathbf{w}} = \sum_{s=1}^{S} w_s(\widehat{\boldsymbol{\beta}}_{(s)} - \boldsymbol{\beta}^*_{(s)})$$

and

$$\widehat{\boldsymbol{\alpha}}_{\mathbf{w},i} - \boldsymbol{\alpha}^*_{\mathbf{w},i} = \sum_{s=1}^{S} w_s(\widehat{\boldsymbol{\alpha}}_{(s),i} - \boldsymbol{\alpha}^*_{(s),i}).$$

Under Condition C.10, by repeatedly using the Minkowski's inequality, we have

$$E^{1/2}\left\{\sup_{\mathbf{w}\in\mathcal{W}}\left\|N^{1/2}\left(\widehat{\boldsymbol{\beta}}_{\mathbf{w}} - \boldsymbol{\beta}^*_{\mathbf{w}}\right)\right\|\right\}^2 \le \sum_{s=1}^{S} E^{1/2}\left\|N^{1/2}\left(\widehat{\boldsymbol{\beta}}_{(s)} - \boldsymbol{\beta}^*_{(s)}\right)\right\|^2 = O(1),$$

therefore,

$$\sup_{\mathbf{w}\in\mathcal{W}}\left\|\widehat{\boldsymbol{\beta}}_{\mathbf{w}} - \boldsymbol{\beta}^*_{\mathbf{w}}\right\| = O_p(N^{-1/2}). \tag{A.20}$$

Similarly, with Condition C.10,

$$\max_{1\le i\le m} E^{1/2}\left\{\sup_{\mathbf{w}\in\mathcal{W}}\left\|n_i^{1/2}\left(\widehat{\boldsymbol{\alpha}}_{\mathbf{w},i} - \boldsymbol{\alpha}^*_{\mathbf{w},i}\right)\right\|\right\}^2 \le \max_{1\le i\le m}\sum_{s=1}^{S} E^{1/2}\left\|n_i^{1/2}\left(\widehat{\boldsymbol{\alpha}}_{(s),i} - \boldsymbol{\alpha}^*_{(s),i}\right)\right\|^2 = O(1).$$

Thus

$$\frac{E\left(\sup_{\mathbf{w}\in\mathcal{W}}\sum_{i=1}^{m}\left\|n_i^{1/2}\left(\widehat{\boldsymbol{\alpha}}_{\mathbf{w},i} - \boldsymbol{\alpha}^*_{\mathbf{w},i}\right)\right\|^2\right)}{N} \le \frac{\sum_{i=1}^{m} E\left\{\sup_{\mathbf{w}\in\mathcal{W}}\left\|n_i^{1/2}\left(\widehat{\boldsymbol{\alpha}}_{\mathbf{w},i} - \boldsymbol{\alpha}^*_{\mathbf{w},i}\right)\right\|\right\}^2}{mn_*} = O(n_*^{-1}),$$

which implies that

$$\sup_{\mathbf{w}\in\mathcal{W}}\left\{\frac{\sum_{i=1}^{m}\left\|n_i^{1/2}\left(\widehat{\boldsymbol{\alpha}}_{\mathbf{w},i} - \boldsymbol{\alpha}^*_{\mathbf{w},i}\right)\right\|^2}{N}\right\}^{1/2} = O_p\left(n_*^{-1/2}\right). \tag{A.21}$$

Recall that

$$\begin{aligned}
&\sup_{\mathbf{w}\in\mathcal{W}}\left|\text{cKL}(\mathbf{w}) - \text{cKL}^*(\mathbf{w})\right| \\
&= 2\phi^{-1}\sup_{\mathbf{w}\in\mathcal{W}}\left|B\{\widehat{\boldsymbol{\theta}}(\mathbf{w})\} - B\{\boldsymbol{\theta}^*(\mathbf{w})\} - \boldsymbol{\mu}^{\mathrm{T}}\left\{\widehat{\boldsymbol{\theta}}(\mathbf{w}) - \boldsymbol{\theta}^*(\mathbf{w})\right\}\right| \\
&\le 2\phi^{-1}\sup_{\mathbf{w}\in\mathcal{W}}\sum_{i=1}^{m}\sum_{j=1}^{n_i}\left|b^{(1)}\left(\widetilde{\theta}_{\mathbf{w},ij}\right)\right|\left|\widehat{\theta}_{\mathbf{w},ij} - \theta^*_{\mathbf{w},ij}\right| \\
&\quad + 2\phi^{-1}\sup_{\mathbf{w}\in\mathcal{W}}\sum_{i=1}^{m}\sum_{j=1}^{n_i}\left|\mu_{0,ij}\right|\left|\widehat{\theta}_{\mathbf{w},ij} - \theta^*_{\mathbf{w},ij}\right|.
\end{aligned} \tag{A.22}$$

It follows from (A.20), (A.21), and Conditions C.7 and C.8 that

$$\begin{aligned}
&\sup_{\mathbf{w}\in\mathcal{W}}\sum_{i=1}^{m}\sum_{j=1}^{n_i}\left|b^{(1)}\left(\widetilde{\theta}_{\mathbf{w},ij}\right)\right|\left|\widehat{\theta}_{\mathbf{w},ij} - \theta^*_{\mathbf{w},ij}\right| \\
&\le \sup_{\mathbf{w}\in\mathcal{W}} N\left[\left\{\frac{\sum_{i=1}^{m}\sum_{j=1}^{n_i}\left|b^{(1)}\left(\widetilde{\theta}_{\mathbf{w},ij}\right)\right|^2\left\|\mathbf{x}_{ij}\right\|^2}{N}\right\}^{1/2}\left\|\widehat{\boldsymbol{\beta}}_{\mathbf{w}} - \boldsymbol{\beta}^*_{\mathbf{w}}\right\|\right. \\
&\quad + \left\{\frac{\sum_{i=1}^{m}\sum_{j=1}^{n_i}\left|b^{(1)}\left(\widetilde{\theta}_{\mathbf{w},ij}\right)\right|^2\left\|\mathbf{z}_{ij}\right\|^2}{N}\right\}^{1/2} \\
&\quad \left. \times\left\{\frac{\sum_{i=1}^{m}\left\|n_i^{1/2}\left(\widehat{\boldsymbol{\alpha}}_{\mathbf{w},i} - \boldsymbol{\alpha}^*_{\mathbf{w},i}\right)\right\|^2}{N}\right\}^{1/2}\right] \\
&= O_p(Nn_*^{-1/2}).
\end{aligned} \tag{A.23}$$

In a similar way, from (A.20), (A.21), and Conditions C.7, C.9 and C.10, we have

$$\sup_{\mathbf{w}\in\mathcal{W}}\sum_{i=1}^{m}\sum_{j=1}^{n_i}\left|\mu_{0,ij}\right|\left|\widehat{\theta}_{\mathbf{w},ij} - \theta^*_{\mathbf{w},ij}\right| = O_p\left(Nn_*^{-1/2}\right). \tag{A.24}$$

Combining (A.22)–(A.24), we have

$$\sup_{\mathbf{w}\in\mathcal{W}}\left|\text{cKL}(\mathbf{w}) - \text{cKL}^*(\mathbf{w})\right| = O_p(Nn_*^{-1/2}). \tag{A.25}$$

Let $\widetilde{\mathcal{G}}_c(\mathbf{w}) = \mathcal{G}_c(\mathbf{w}) + 2\phi^{-1}\{\boldsymbol{\mu}^{\mathrm{T}}\boldsymbol{\theta}_0 - B(\boldsymbol{\theta}_0)\}$. It is straightforward to show that

$$\begin{aligned}
&\sup_{\mathbf{w}\in\mathcal{W}} \left|\widetilde{\mathcal{G}}_c(\mathbf{w}) - \mathrm{cKL}^*(\mathbf{w})\right| \\
&\quad\le 2\phi^{-1} \sup_{\mathbf{w}\in\mathcal{W}} \sum_{i=1}^{m}\sum_{j=1}^{n_i} \left|b^{(1)}\left(\widetilde{\theta}_{\mathbf{w},ij}\right)\right| \left|\widehat{\theta}_{\mathbf{w},ij} - \theta^*_{\mathbf{w},ij}\right| \\
&\qquad +2\phi^{-1} \sup_{\mathbf{w}\in\mathcal{W}} \sum_{i=1}^{m}\sum_{j=1}^{n_i} \left|\mu_{0,ij}\right| \left|\widehat{\theta}_{\mathbf{w},ij} - \theta^*_{\mathbf{w},ij}\right| \\
&\qquad +2\phi^{-1} \sup_{\mathbf{w}\in\mathcal{W}} \left|\left(\mathbf{y}-\boldsymbol{\mu}\right)^{\mathrm{T}} \left\{\widehat{\boldsymbol{\theta}}(\mathbf{w}) - \boldsymbol{\theta}^*(\mathbf{w})\right\}\right| \\
&\qquad +2\phi^{-1} \sup_{\mathbf{w}\in\mathcal{W}} \left|\left(\mathbf{y}-\boldsymbol{\mu}\right)^{\mathrm{T}} \boldsymbol{\theta}^*(\mathbf{w})\right| + \sup_{\mathbf{w}\in\mathcal{W}} \mathbf{w}^{\mathrm{T}}\boldsymbol{\Upsilon}\boldsymbol{\varpi}_{m,n}.
\end{aligned} \tag{A.26}$$

It follows from (A.20), (A.21), and Conditions C.7 and C.9 that

$$\begin{aligned}
&\sup_{\mathbf{w}\in\mathcal{W}} \left|\left(\mathbf{y}-\boldsymbol{\mu}\right)^{\mathrm{T}} \left\{\widehat{\boldsymbol{\theta}}(\mathbf{w}) - \boldsymbol{\theta}^*(\mathbf{w})\right\}\right| \\
&\quad\le \sum_{s=1}^{S} N^{1/2} \left\|\frac{\mathbf{y}-\boldsymbol{\mu}}{N^{1/2}}\right\| \left(\sum_{i=1}^{m}\sum_{j=1}^{n_i} \left|\widehat{\theta}_{s,ij} - \theta^*_{s,ij}\right|^2\right)^{1/2} \\
&\quad= O_p\left(Nn_*^{-1/2}\right).
\end{aligned} \tag{A.27}$$

Since y_{ij} for $i = 1, \ldots, m$ and $j = 1, \ldots, n_i$ are conditionally independent given $\boldsymbol{\theta}_0$, from Condition C.9, we obtain that

$$\begin{aligned}
&\left[E_{\mathbf{y}|\boldsymbol{\theta}_0}\left\{\sum_{s=1}^{S}\left|\frac{\sum_{i=1}^{m}\sum_{j=1}^{n_i}(y_{ij}-\mu_{0,ij})\theta^*_{s,ij}}{N^{1/2}}\right|\right\}^2\right]^{1/2} \\
&\quad\le \sum_{s=1}^{S}\left\{\frac{\sum_{i=1}^{m}\sum_{j=1}^{n_i} E_{\mathbf{y}|\boldsymbol{\theta}_0}(y_{ij}-\mu_{0,ij})^2\left|\theta^*_{s,ij}\right|^2}{N}\right\}^{1/2} \\
&\quad\le \sum_{s=1}^{S}\left\{\frac{\sum_{i=1}^{m}\sum_{j=1}^{n_i}\left|\sigma^2_{ij}\left(\theta_{0,ij}\right)\right|^2}{N}\right\}^{1/4}\left\{\frac{\sum_{i=1}^{m}\sum_{j=1}^{n_i}\left|\theta^*_{s,ij}\right|^4}{N}\right\}^{1/4} \\
&\quad\le S\left\{\bar{\sigma}^4(\boldsymbol{\theta}_0)\bar{\theta}_4\right\}^{1/4}
\end{aligned}$$

holds $p(\boldsymbol{\theta}_0)$-a.s., which yields that

$$\sup_{\mathbf{w}\in\mathcal{W}} \left|\left(\mathbf{y}-\boldsymbol{\mu}\right)^{\mathrm{T}} \boldsymbol{\theta}^*(\mathbf{w})\right| = O_p(N^{1/2}). \tag{A.28}$$

It follows from (A.23) and (A.24) and (A.26)–(A.28) that

$$\begin{aligned}
\sup_{\mathbf{w}\in\mathcal{W}} \left|\widetilde{\mathcal{G}}_c(\mathbf{w}) - \mathrm{cKL}^*(\mathbf{w})\right| &= O_p(Nn_*^{-1/2}) + O_p(N^{1/2}) \\
&\quad + \sup_{\mathbf{w}\in\mathcal{W}} \mathbf{w}^{\mathrm{T}}\boldsymbol{\Upsilon}\boldsymbol{\varpi}_{m,n} \\
&= O_p(Nn_*^{-1/2}) + \sup_{\mathbf{w}\in\mathcal{W}} \mathbf{w}^{\mathrm{T}}\boldsymbol{\Upsilon}\boldsymbol{\varpi}_{m,n}.
\end{aligned} \tag{A.29}$$

Denote $r_n(\mathbf{w}) = \mathrm{cKL}^*(\mathbf{w}) - \mathrm{cKL}(\mathbf{w}) + \widetilde{\mathcal{G}}_c(\mathbf{w}) - \mathrm{cKL}^*(\mathbf{w})$. Then

$$\widetilde{\mathcal{G}}_c(\mathbf{w}) = \mathrm{cKL}(\mathbf{w}) + r_n(\mathbf{w}). \tag{A.30}$$

With (A.25) and (A.29), it is readily seen that

$$\begin{aligned}
&\sup_{\mathbf{w}\in\mathcal{W}} \left|\frac{r_n(\mathbf{w})}{\mathrm{cKL}(\mathbf{w})}\right| \\
&\quad\le \frac{Nn_*^{-1/2}}{\xi_N} \sup_{\mathbf{w}\in\mathcal{W}} \left|\frac{\mathrm{cKL}^*(\mathbf{w}) - \mathrm{cKL}(\mathbf{w}) + \widetilde{\mathcal{G}}_c(\mathbf{w}) - \mathrm{cKL}^*(\mathbf{w})}{Nn_*^{-1/2}}\right| \\
&\qquad \times \left\{\sup_{\mathbf{w}\in\mathcal{W}} \left|\frac{\mathrm{cKL}^*(\mathbf{w})}{\mathrm{cKL}(\mathbf{w})}\right|\right\} \\
&\quad\le \frac{Nn_*^{-1/2}}{\xi_N}\left\{O_p(1) + \sup_{\mathbf{w}\in\mathcal{W}}\left|\frac{\mathbf{w}^{\mathrm{T}}\boldsymbol{\Upsilon}\boldsymbol{\varpi}_{m,n}}{Nn_*^{-1/2}}\right|\right\} \\
&\qquad \times \left\{-\frac{Nn_*^{-1/2}}{\xi_N}\sup_{\mathbf{w}\in\mathcal{W}}\left|\frac{\mathrm{cKL}(\mathbf{w}) - \mathrm{cKL}^*(\mathbf{w})}{Nn_*^{-1/2}}\right| + 1\right\}^{-1} \\
&\quad= o_p(1),
\end{aligned} \tag{A.31}$$

where we have used the condition $n_*^{-1/2}\|\boldsymbol{\varpi}_{m,n}\|_\infty = O(1)$ and the fact that

$$\begin{aligned}
\sup_{\mathbf{w}\in\mathcal{W}} \left|\frac{\mathbf{w}^{\mathrm{T}}\boldsymbol{\Upsilon}\boldsymbol{\varpi}_{m,n}}{Nn_*^{-1/2}}\right| &= \sup_{\mathbf{w}\in\mathcal{W}} \frac{n_*^{1/2}\sum_{s=1}^{S} w_s\left\{\varpi_0 p_s + \sum_{i=1}^{m}\varpi_i d_{si}\right\}}{\sum_{i=1}^{m} n_i} \\
&= O\left(\frac{\|\boldsymbol{\varpi}_{m,n}\|_\infty}{n_*^{1/2}}\right).
\end{aligned}$$

From (A.31) we can get (15).

A7. Derivation of Conditional BIC

In this subsection, we give the details on the derivation the conditional BIC (cBIC) for GLMM; that is,

$$c\mathrm{BIC} = -2l_c(\mathbf{y} \mid \widehat{\boldsymbol{\gamma}}) + \log(N)p + \sum_{i=1}^{m}\log(n_i)d_i, \tag{A.32}$$

where $l_c(\mathbf{y} \mid \boldsymbol{\gamma}) = \sum_{i=1}^{m}\sum_{j=1}^{n_i}[\phi^{-1}\{y_{ij}\mathbf{u}_{ij}^{\mathrm{T}}\boldsymbol{\gamma} - b(\mathbf{u}_{ij}^{\mathrm{T}}\boldsymbol{\gamma})\} + c(y_{ij}, \phi)]$, and p and d_i are the dimensions of fixed effects vector $\boldsymbol{\beta}$ and random effects vector $\boldsymbol{\alpha}_i$, respectively. The design matrix $(\mathbf{X}, \mathbf{Z})$ is full column ranked (if it is not, one needs to adopt the reparameterization commented in Remark 3) and in this case $p = \mathrm{rank}(\mathbf{X}_2)$ and $d_i = \mathrm{rank}(\mathbf{Z}_i)$. In GLMM, the penalized log-likelihood function corresponding to the PGWLS method is

$$l_p(\mathbf{y} \mid \boldsymbol{\gamma}) = l_c(\mathbf{y} \mid \boldsymbol{\gamma}) + p_R(\boldsymbol{\alpha} \mid \nu), \tag{A.33}$$

where $p_R(\boldsymbol{\alpha} \mid \nu)$ is the penalty on the random effects, and ν is the prefixed Lagrange multiplier (Jiang 1999; Wang, Tsai, and Qu 2012). In the derivation of cBIC, for simplicity, we assume that the fixed effects have a noninformative prior and the marginal probability density function of $\mathbf{y}$ is

$$f(\mathbf{y}) = \int \exp\left\{l_c\left(\mathbf{y} \mid \boldsymbol{\gamma}\right)\right\}\pi\left(\boldsymbol{\gamma}\right)d\boldsymbol{\gamma},$$

where $\pi(\boldsymbol{\gamma}) = C_q\exp\{p_R(\boldsymbol{\alpha} \mid \nu)\}$ and C_q is a term related to q. Under this setup, the cBIC is essentially the first-order Laplace approximation of $-2\log\{f(\mathbf{y})\}$ (Konishi and Kitagawa 2007, Chapter 9). We introduce the following notations: $\widehat{\boldsymbol{\gamma}}$ is the solution of

$\partial l_p(\boldsymbol{\gamma})/\partial\boldsymbol{\gamma} = \mathbf{0}_{(p+q)\times 1}$, $q = \sum_{i=1}^m d_i$,

$$\begin{aligned}
\mathbf{Q}_n &= \text{blockdiag}(n_1\mathbf{I}_{d_1}, \ldots, n_m\mathbf{I}_{d_m}),\\
\mathbf{Q}_{mn} &= \text{blockdiag}\left(N\mathbf{I}_p, \mathbf{Q}_n\right),\\
\mathbf{H}_{00}(\boldsymbol{\gamma}) &= -\frac{\partial^2 l_c(\mathbf{y}\mid\boldsymbol{\gamma})}{\partial\boldsymbol{\beta}\partial\boldsymbol{\beta}^{\mathrm{T}}},\quad \mathbf{H}_{0\boldsymbol{\alpha}}(\boldsymbol{\gamma}) = -\frac{\partial^2 l_c(\mathbf{y}\mid\boldsymbol{\gamma})}{\partial\boldsymbol{\beta}\partial\boldsymbol{\alpha}^{\mathrm{T}}},\\
\mathbf{H}_{ii}(\boldsymbol{\gamma}) &= -\frac{\partial^2 l_c(\mathbf{y}\mid\boldsymbol{\gamma})}{\partial\boldsymbol{\alpha}_i\partial\boldsymbol{\alpha}_i^{\mathrm{T}}},\\
\mathbf{H}_{\boldsymbol{\alpha}\boldsymbol{\alpha}}(\boldsymbol{\gamma}) &= \text{blockdiag}\left\{-\frac{\partial^2 l_c(\mathbf{y}\mid\boldsymbol{\gamma})}{\partial\boldsymbol{\alpha}_i\partial\boldsymbol{\alpha}_i^{\mathrm{T}}}\right\}_{i=1,\ldots,m},\\
\mathbf{J}_{00}(\boldsymbol{\gamma}) &= -\frac{1}{N}\frac{\partial^2 l_p(\mathbf{y}\mid\boldsymbol{\gamma})}{\partial\boldsymbol{\beta}\partial\boldsymbol{\beta}^{\mathrm{T}}} = \frac{1}{N}\mathbf{H}_{00}(\boldsymbol{\gamma}),\\
\mathbf{J}_{\boldsymbol{\alpha}\boldsymbol{\alpha}}(\boldsymbol{\gamma}) &= -\mathbf{Q}_n^{-1/2}\frac{\partial^2 l_p\left(\mathbf{y}\mid\boldsymbol{\gamma}\right)}{\partial\boldsymbol{\alpha}\partial\boldsymbol{\alpha}^{\mathrm{T}}}\mathbf{Q}_n^{-1/2}\\
&= \mathbf{Q}_n^{-1/2}\left\{\mathbf{H}_{\boldsymbol{\alpha}\boldsymbol{\alpha}}(\boldsymbol{\gamma}) - \frac{\partial^2 p_R(\boldsymbol{\alpha}\mid\nu)}{\partial\boldsymbol{\alpha}\partial\boldsymbol{\alpha}^{\mathrm{T}}}\right\}\mathbf{Q}_n^{-1/2},\\
\mathbf{J}_{0\boldsymbol{\alpha}}(\boldsymbol{\gamma}) &= -\frac{1}{N^{1/2}}\frac{\partial^2 l_p\left(\mathbf{y}\mid\boldsymbol{\gamma}\right)}{\partial\boldsymbol{\beta}\partial\boldsymbol{\alpha}^{\mathrm{T}}}\mathbf{Q}_n^{-1/2} = \frac{1}{N^{1/2}}\mathbf{H}_{0\boldsymbol{\alpha}}(\boldsymbol{\gamma})\mathbf{Q}_n^{-1/2},
\end{aligned}$$

and

$$\mathbf{V}_{00}(\boldsymbol{\gamma}) = \mathbf{J}_{00}(\boldsymbol{\gamma}) - \mathbf{J}_{0\boldsymbol{\alpha}}(\boldsymbol{\gamma})\mathbf{J}_{\boldsymbol{\alpha}\boldsymbol{\alpha}}^{-1}(\boldsymbol{\gamma})\mathbf{J}_{\boldsymbol{\alpha}0}(\boldsymbol{\gamma}).$$

To obtain the desired result, we need the additional regularity conditions. Let $\lambda_{\min}(\mathbf{C})$ and $\lambda_{\max}(\mathbf{C})$ denote the smallest and largest eigenvalues of the matrix $\mathbf{C}$, respectively.

Condition C.12. For sufficiently large n_*

$$0 < c_0 \le \lambda_{\min}\{\mathbf{V}_{00}(\widehat{\boldsymbol{\gamma}})\} \le \lambda_{\max}\{\mathbf{V}_{00}(\widehat{\boldsymbol{\gamma}})\} \le c_0^* < \infty,$$

and for every $i = 1, \ldots, m$

$$0 < c_1 \le \lambda_{\min}\{\mathbf{H}_{ii}(\widehat{\boldsymbol{\gamma}})/n_i\} \le \lambda_{\max}\{\mathbf{H}_{ii}(\widehat{\boldsymbol{\gamma}})/n_i\} \le c_1^* < \infty,$$

where c_i and c_i^* $(i = 0, 1)$ are positive constants.

Condition C.13.

$$m/n_* = o(1), \tag{A.34}$$

$$\begin{aligned}
0 < c_2 &\le \lambda_{\min}\left\{-\frac{\partial^2 p_R(\widehat{\boldsymbol{\alpha}}\mid\nu)}{\partial\boldsymbol{\alpha}\partial\boldsymbol{\alpha}^{\mathrm{T}}}\right\}\\
&\le \lambda_{\max}\left\{-\frac{\partial^2 p_R(\widehat{\boldsymbol{\alpha}}\mid\nu)}{\partial\boldsymbol{\alpha}\partial\boldsymbol{\alpha}^{\mathrm{T}}}\right\} \le c_2^* < \infty,
\end{aligned} \tag{A.35}$$

uniformly for every $\nu > 0$, where c_2 and c_2^* are positive constants.

Condition C.12 can be verified for some usually used GLMMs and similar conditions have been discussed in Nie (2007) and Yu, Zhang, and Yau (2014). In Condition C.13, (A.34) guarantees that we have sufficient information to predict random effects. (A.35) is a very basic condition and can be verified for certain GLMMs. For example, in the random intercept model: $\theta_{ij} = \beta_0 + \beta_1 x_{ij} + b_i$, where $x_{ij} \sim N(0, 1)$, as stated above in Remark 3, we need to consider reparameterize $\alpha_i = \beta_0 + b_i$ and $\theta_{ij} = \beta_1 x_{ij} + \alpha_i$. In PGWLS method, according to Jiang (1999) and Wang, Tsai, and Qu (2012),

$$p_R(\boldsymbol{\alpha}\mid\nu) = -\frac{\nu}{2}\left\{\|\mathbf{P_A}(\boldsymbol{\alpha} - \beta_0\mathbf{1}_{m\times 1})\|^2 + \|\boldsymbol{\alpha} - \beta_0\mathbf{1}_{m\times 1}\|^2\right\}$$

with $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_m)^{\mathrm{T}}$ and $\mathbf{P_A} = \mathbf{1}_{m\times 1}\mathbf{1}_{m\times 1}^{\mathrm{T}}/m$. It is clear that

$$\lambda_{\min}\left\{-\frac{\partial^2 p_R(\boldsymbol{\alpha}\mid\nu)}{\partial\boldsymbol{\alpha}\partial\boldsymbol{\alpha}^{\mathrm{T}}}\right\} = \lambda_{\min}\{\nu(\mathbf{P_A} + \mathbf{I}_m)\} = \nu > 0$$

and

$$\lambda_{\max}\left\{-\frac{\partial^2 p_R(\boldsymbol{\alpha}\mid\nu)}{\partial\boldsymbol{\alpha}\partial\boldsymbol{\alpha}^{\mathrm{T}}}\right\} \le 2\nu < \infty,$$

for each $\boldsymbol{\alpha}$ and $\nu > 0$.

We now derive cBIC for GLMMs. By the definition of $\widehat{\boldsymbol{\gamma}}$, it is clear that the first-order Laplace approximation of $f(\mathbf{y})$ is

$$\begin{aligned}
&C_q\int\exp\left\{l_p\left(\mathbf{y}\mid\widehat{\boldsymbol{\gamma}}\right) - \frac{(\boldsymbol{\gamma}-\widehat{\boldsymbol{\gamma}})^{\mathrm{T}}\mathbf{Q}_{mn}^{1/2}\mathbf{J}(\widehat{\boldsymbol{\gamma}})\mathbf{Q}_{mn}^{1/2}(\boldsymbol{\gamma}-\widehat{\boldsymbol{\gamma}})}{2}\right\}d\boldsymbol{\gamma}\\
&= \frac{C_q(2\pi)^{\frac{p+q}{2}}\exp\left\{l_p\left(\mathbf{y}\mid\widehat{\boldsymbol{\gamma}}\right)\right\}}{|\mathbf{Q}_{mn}|^{1/2}|\mathbf{J}(\widehat{\boldsymbol{\gamma}})|^{1/2}},
\end{aligned}$$

where $\mathbf{J}(\boldsymbol{\gamma}) = -\mathbf{Q}_{mn}^{-1/2}\partial^2 l_p(\boldsymbol{\gamma})/(\partial\boldsymbol{\gamma}\partial\boldsymbol{\gamma}^{\mathrm{T}})\mathbf{Q}_{mn}^{-1/2}$. Therefore,

$$\begin{aligned}
&-2\log\left[C_q\int\exp\left\{l_p\left(\mathbf{y}\mid\widehat{\boldsymbol{\gamma}}\right)\right.\right.\\
&\quad\left.\left.-\frac{(\boldsymbol{\gamma}-\widehat{\boldsymbol{\gamma}})^{\mathrm{T}}\mathbf{Q}_{mn}^{1/2}\mathbf{J}(\widehat{\boldsymbol{\gamma}})\mathbf{Q}_{mn}^{1/2}(\boldsymbol{\gamma}-\widehat{\boldsymbol{\gamma}})}{2}\right\}d\boldsymbol{\gamma}\right]\\
&= -2l_p\left(\mathbf{y}\mid\widehat{\boldsymbol{\gamma}}\right) + \log|\mathbf{Q}_{mn}| + \log|\mathbf{J}(\widehat{\boldsymbol{\gamma}})|\\
&\quad -(p+q)\log(2\pi) - 2\log(C_q)\\
&= -2l_p\left(\mathbf{y}\mid\widehat{\boldsymbol{\gamma}}\right) + \log(N)p + \sum_{i=1}^m\log(n_i)d_i\\
&\quad +\log|\mathbf{V}_{00}(\widehat{\boldsymbol{\gamma}})| + \log|\mathbf{J}_{\boldsymbol{\alpha}\boldsymbol{\alpha}}(\widehat{\boldsymbol{\gamma}})| - (p+q)\log(2\pi) - 2\log(C_q).
\end{aligned} \tag{A.36}$$

Under Condition C.13, it can be verified that

$$\begin{aligned}
\log|\mathbf{J}_{\boldsymbol{\alpha}\boldsymbol{\alpha}}(\widehat{\boldsymbol{\gamma}})| &= \sum_{i=1}^m\log|\mathbf{H}_{ii}(\widehat{\boldsymbol{\gamma}})/n_i| + O_p(m/n_*)\\
&= \sum_{i=1}^m\log|\mathbf{H}_{ii}(\widehat{\boldsymbol{\gamma}})/n_i| + o_p(1).
\end{aligned} \tag{A.37}$$

Then by (A.36) and (A.37) we have

$$\begin{aligned}
&-2\log\left[C_q\int\exp\left\{l_p\left(\mathbf{y}\mid\widehat{\boldsymbol{\gamma}}\right)\right.\right.\\
&\quad\left.\left.-\frac{(\boldsymbol{\gamma}-\widehat{\boldsymbol{\gamma}})^{\mathrm{T}}\mathbf{Q}_{mn}^{1/2}\mathbf{J}(\widehat{\boldsymbol{\gamma}})\mathbf{Q}_{mn}^{1/2}(\boldsymbol{\gamma}-\widehat{\boldsymbol{\gamma}})}{2}\right\}d\boldsymbol{\gamma}\right]\\
&= -2l_p\left(\mathbf{y}\mid\widehat{\boldsymbol{\gamma}}\right) + \log(N)p + \sum_{i=1}^m\log(n_i)d_i
\end{aligned}$$

$$+\log|\mathbf{V}_{00}(\widehat{\boldsymbol{\gamma}})| + \sum_{i=1}^{m}\log|\mathbf{H}_{ii}(\widehat{\boldsymbol{\gamma}})/n_i| - (p+q)\log(2\pi) - 2\log(C_q) + o_p(1). \quad (A.38)$$

From Conditions C.12 and C.13, we have $\log|\mathbf{V}_{00}(\widehat{\boldsymbol{\gamma}})| + \sum_{i=1}^{m}\log|\mathbf{H}_{ii}(\widehat{\boldsymbol{\gamma}})/n_i| = O(p+q)$ and $C_q = O_p(q)$, which along with $\log(N)p + \sum_{i=1}^{m}\log(n_i)d_i \geq \log(n_*)(p+q)$, leads to (A.32) after ignoring the terms in (A.38) which are $o_p\{\log(n_*)(p+q)\}$.

Acknowledgments

The work of Zhang was performed during his visit to Prof. Raymond J. Carroll at Texas A&M University, whose support is gratefully appreciated. The authors thank the co-editors, an associate editor and three referees for their constructive suggestions and comments that have substantially improved an earlier version of this article.

Funding

Zhang's work was partially supported by National Natural Science Foundation of China (NNSFC, Grant nos. 71522004, 11471324 and 11271355). Yu's work was partially supported by NNSFC (Grant no. 11301463). Zou's work was partially supported by NNSFC (Grant no. 11331011) and the Beijing High-level Talents Program. Liang's work was partially supported by NSF grants DMS-1440121 and DMS-1418042, and Award Number 11529101, made by National Natural Science Foundation of China.

References

Ando, T., and Li, K.-C. (2014), "A Model-Averaging Approach for High-dimensional Regression," *Journal of the American Statistical Association*, 109, 254–265. [1775,1777,1784]

Ashbaugh-Skaife, H., Collins, D. W., and LaFond, R. (2006), "The Effects of Corporate Governance on Firms' Credit Ratings," *Journal of Accounting and Economics*, 42, 203–243. [1782]

Buckland, S. T., Burnham, K. P., and Augustin, N. H. (1997), "Model Selection: An Integral Part of Inference," *Biometrics*, 53, 603–618. [1777,1779]

Claeskens, G., Croux, C., and van Kerckhoven, J. (2006), "Variable Selection for Logistic Regression Using a Prediction-focused Information Criterion," *Biometrics*, 62, 972–979. [1779]

Conflitti, C. (2011), "Opinion Surveys on the Euro: A Multilevel Multinomial Logistic Analysis," *SSRN*. [1783]

Donohue, M., Overholser, R., Xu, R., and Vaida, F. (2011), "Conditional Akaike Information Under Generalized Linear and Proportional Hazards Mixed Models," *Biometrika*, 98, 685–700. [1785]

Fan, J., and Song, R. (2010), "Sure Independence Screening in Generalized Linear Models With NP-dimensionality," *The Annals of Statistics*, 38, 3567–3604. [1777]

Flynn, C., Hurvich, C., and Simonoff, J. (2013), "Efficiency for Regularization Parameter Selection in Penalized Likelihood Estimation of Misspecified Models," *Journal of the American Statistical Association*, 108, 1031–1043. [1784]

Friedman, J., Hastie, T., and Tibshirani, R. (2010), "Regularization Paths for Generalized Linear Models via Coordinate Descent," *Journal of Statistical Software*, 33, 1–22. [1780]

Hansen, B. E. (2007), "Least Squares Model Averaging," *Econometrica*, 75, 1175–1189. [1775,1776,1777]

——— (2014), "Model Averaging, Asymptotic Risk, and Regressor Groups," *Quantitative Economics*, 5, 495–530. [1775]

Hansen, B. E., and Racine, J. (2012), "Jackknife Model Averaging," *Journal of Econometrics*, 167, 38–46. [1775]

Hjort, N. L., and Claeskens, G. (2003), "Frequentist Model Average Estimators," *Journal of the American Statistical Association*, 98, 879–899. [1779]

Hoeting, J. A., Madigan, D., Raftery, A. E., and Volinsky, C. T. (1999), "Bayesian Model Averaging: A Tutorial," *Statistical Science*, 14, 382–417. [1775]

Jiang, J. (1999), "Conditional Inference About Generalized Linear Mixed Models," *The Annals of Statistics*, 27, 1974–2007. [1775,1778,1788,1789]

Jiang, J., Jia, H., and Chen, H. (2001), "Maximum Posterior Estimation of Random Effects in Generalized Linear Mixed Models," *Statistica Sinica*, 11, 97–120. [1777]

Konishi, S., and Kitagawa, G. (2007), *Information Criteria and Statistical Modeling*, New York: Springer. [1788]

Lele, S. R., Nadeem, K., and Schmuland, B. (2010), "Estimability and Likelihood Inference for Generalized Linear Mixed Models Using Data Cloning," *Journal of the American Statistical Association*, 105, 1617–1625. [1777]

Liang, H., Zou, G., Wan, A. T. K., and Zhang, X. (2011), "Optimal Weight Choice for Frequentist Model Average Estimators," *Journal of the American Statistical Association*, 106, 1053–1066. [1775]

Liu, Q., and Okui, R. (2013), "Heteroskedasticity-robust Cp Model Averaging," *The Econometrics Journal*, 16, 463–472. [1775]

Longford, N. T. (2005), "Editorial: Model Selection and Efficiency—is 'Which Model?' the Right Question?" *Journal of the Royal Statistical Society*, Series A, 168, 469–472. [1775]

Lv, J., and Liu, J. (2014), "Model Selection Principles in Misspecified Models," *Journal of the Royal Statistical Society*, Series B, 76, 141–167. [1784,1785]

Molenberghs, G., Verbeke, G., Demétrio, C. G. B., and Vieira, A. (2010), "A Family of Generalized Linear Models for Repeated Measures With Normal and Conjugate Random Effects," *Statistical Science*, 25, 325–347. [1777]

Nie, L. (2007), "Convergence Rate of MLE in Generalized Linear and Nonlinear Mixed-effects Models: Theory and Applications," *Journal of Statistical Planning and Inference*, 137, 1787–1804. [1789]

Tan, L. S. L., and David, J. (2013), "Variational Inference for Generalized Linear Mixed Models Using Partially Noncentered Parametrizations," *Statistical Science*, 28, 168–188. [1778]

Verbeek, M. (2007), *A Guide to Modern Econometrics* (3rd ed.), New York: Wiley. [1782]

Wan, A. T. K., Zhang, X., and Zou, G. (2010), "Least Squares Model Averaging by Mallows Criterion," *Journal of Econometrics*, 156, 277–283. [1785]

Wang, P., Tsai, G.-F., and Qu, A. (2012), "Conditional Inference Functions for Mixed-effects Models With Unspecified Random-effects Distribution," *Journal of the American Statistical Association*, 107, 725–736. [1778,1788,1789]

White, H. (1982), "Maximum Likelihood Estimation of Misspecified Models," *Econometrica*, 50, 1–25. [1775,1776,1784]

Whittle, P. (1960), "Bounds for the Moments of Linear and Quadratic Forms in Independent Variables," *Theory of Probability & Its Applications*, 5, 302–305. [1775]

Yang, Y. (2005), "Can the Strengths of AIC and BIC be Shared? A Conflict Between Model Indentification and Regression Estimation," *Biometrika*, 92, 937–950. [1777]

Yu, D., Zhang, X., and Yau, K. (2014), "Information Criterion for Misspecificed Random Effects Models," under review. [1778,1779,1784,1785,1789]

Yuan, Z., and Yang, Y. (2005), "Combining Linear Regression Models: When and How?" *Journal of the American Statistical Association*, 100, 1202–1214. [1777]

Zhang, X., and Liang, H. (2011), "Focused Information Criterion and Model Averaging for Generalized Additive Partial Linear Models," *The Annals of Statistics*, 39, 174–200. [1779]

Zhang, X., Lu, Z., and Zou, G. (2013), "Adaptively Combined Forecasting for Discrete Response Time Series," *Journal of Econometrics*, 176, 80–91. [1777]

Zhang, X., Zou, G., and Liang, H. (2014), "Model Averaging and Weight Choice in Linear Mixed-effects Models," *Biometrika*, 101, 205–218. [1775,1777,1784]

Zhao, Y., Staudenmayer, J., Coull, B. A., and Wand, M. P. (2006), "General Design Bayesian Generalized Linear Mixed Models," *Statistical Science*, 21, 35–51. [1778]

Zou, H. (2006), "The Adaptive Lasso and its Oracle Properties," *Journal of the American Statistical Association*, 101, 1418–1429. [1780]