Supplementary materials for this article are available online. Please click the JASA link at http://pubs.amstat.org.

Composite Likelihood Bayesian Information Criteria for Model Selection in High-Dimensional Data

Xin GAO and Peter X.-K. SONG

For high-dimensional data sets with complicated dependency structures, the full likelihood approach often leads to intractable computational complexity. This imposes difficulty on model selection, given that most traditionally used information criteria require evaluation of the full likelihood. We propose a composite likelihood version of the Bayes information criterion (BIC) and establish its consistency property for the selection of the true underlying marginal model. Our proposed BIC is shown to be selection-consistent under some mild regularity conditions, where the number of potential model parameters is allowed to increase to infinity at a certain rate of the sample size. Simulation studies demonstrate the empirical performance of this new BIC, especially for the scenario where the number of parameters increases with sample size. Technical proofs of our theoretical results are provided in the online supplemental materials.

KEY WORDS: Consistency; Model selection; Pseudo-likelihood; Variable selection.

1. INTRODUCTION

In the analysis of high-dimensional data with complex dependency structures, exact likelihood inference often leads to computational complexity. A compromise approach is to use simpler pseudolikelihoods, such as the composite likelihood approach (Lindsay 1988; Cox and Reid 2004). A composite likelihood is constructed from low-dimensional likelihood objects defined over small subsets of data. This dimension-reduction methodology on the likelihood function has been successfully applied in many areas, including generalized linear mixed models (Renard, Molenberghs, and Geys 2004), genetics (Fearnhead and Donnelly 2002), spatial statistics (Hjort and Omre 1994; Heagerty and Lele 1998; Varin and Vidoni 2005), and multivariate survival analysis (Parner 2001; Li and Lin 2006). It has been demonstrated to have desirable theoretical properties, including estimation consistency and asymptotic normality, and can be used to establish hypothesis testing procedures in a similar fashion to the classical likelihood ratio test [see the recent review by Varin (2008) and references therein].

There are often many potential candidate models for revealing the data-generating mechanism. Model selection has become a very important issue in statistical modeling. Varin and Vidoni (2005) proposed a composite likelihood information criterion analogous to Akaike's (1973) information criterion (AIC). Their method selects the model with the best prediction power by minimizing a composite Kullback–Leibler (KL) distance for a future experiment. The proposed first-order unbiased selection statistic contains two components, the composite log-likelihood of the data under a candidate model and the penalty related to the effective number of parameters in the model. In particular, when the composite likelihood takes the form of an ordinary likelihood, the penalty term reduces to the exact number of parameters in the model, which coincides with the AIC. Note that the AIC focuses on selecting models with the best prediction power and that it is not a consistent model selection criterion (e.g., Haughton 1988). As a result, the AIC tends to favor overfitting models. In effect, Varin and Vidoni's composite likelihood selection criterion resembles the AIC, in that it penalizes the number of parameters at the rate of $O(1)$. In some applications, building a parsimonious model is critical to proper interpretation of covariate effects. Therefore, although overfitting does not greatly affect prediction power, it is problematic in studies of association.

This article focuses on the development of a Bayes information criterion (BIC) for the composite likelihood methodology. The BIC was first proposed by Schwarz (1978) in the paradigm of maximum likelihood methodology. Subsequently, many authors have extended it to other estimation methods, including Konish, Ando, and Imoto (2004) in the penalized maximum likelihood method (see also Berger, Ghosh, and Mukhopadhyay 2003; Chakrabarti and Ghosh 2006; Jiang 2007). Essentially, the BIC penalizes more heavily on the number of parameters at the rate of $O(\log n)$, and has been shown to be a consistent model selection criterion in many settings, including linear models (Rao and Wu 1989), partially linear models (Wang, Li, and Tsai 2007), change point analysis (Yao 1988; Csörgö and Horváth 1997), and longitudinal data analysis (Wang and Qu 2009). Recently, Chen and Chen (2008) proposed an extended BIC (EBIC) criterion in the setting of linear regression models with high-dimensional covariates, with an extra penalty proposed to penalize the dimension of model space that supposedly increases with increasing sample size. This penalty is essentially to force the selection of sparse models when the number of regression coefficients, P, tends to infinity as the sample size n increases. Such an EBIC has been shown to be a consistent model selection criterion in the case of generalized linear models with large model space (Chen and Chen 2009).

Here we consider a general statistical model for high-dimensional data with complicated correlation structures. One example of high-dimensional data is correlated regression data (e.g., longitudinal or clustered data) with a large number of covariates. When using the composite likelihood method of parameter estimation, it is of interest to investigate whether the BIC is valid for model selection and, if so, how it behaves.

Xin Gao is Associate Professor, Department of Mathematics and Statistics, York University, Toronto, Ontario, Canada M3J 1P3 (E-mail: *xingao@mathstat.yorku.ca*). Peter X.-K. Song is Professor, Department of Biostatistics, University of Michigan, Ann Arbor, MI 48109-2029 (E-mail: *pxsong@umich.edu*).
This research is supported by Natural Science and Engineering Research Council of Canada Grants held by the first author and NSF grant held by the second author. The authors thank the editor, the associate editor and the referees, whose comments have greatly improved this manuscript.

Journal of the American Statistical Association
December 2010, Vol. 105, No. 492, Theory and Methods
DOI: 10.1198/jasa.2010.tm09414

This motivates us to address the following three goals: (1) to define a composite likelihood BIC (CL-BIC) that will be applicable in situations where the number of parameters increases with the sample size; (2) to establish a large-sample property of the model selection consistency for the proposed CL-BIC, which is a key advantage of the BIC or its variants, as shown in the literature; and (3) to compare CL-BIC with Varin and Vidoni's composite likelihood AIC (CL-AIC), to gain insight into the differences in performance between the AIC and BIC in composite likelihood methodology. In the simulation studies presented in Section 4, we include a comparison of the CL-BIC in the full and composite likelihood methods, given the full likelihood method's status as the gold standard in terms of sensitivity and selectivity.

The article is organized as follows. Section 2 presents the BIC in the composite likelihood framework, and Section 3 concentrates on the property of model selection consistency for the proposed CL-BIC. Sections 4 and 5 illustrate the performance of the CL-BIC and report the comparisons with the CL-AIC via simulation studies and real data analysis. Section 6 concludes with some remarks.

2. COMPOSITE LIKELIHOOD BAYESIAN INFORMATION CRITERION

2.1 Composite Likelihood

The CL paradigm (Lindsay 1988) constitutes a rich class of pseudolikelihoods based on marginal likelihood objects. Let $\{f(\mathbf{y};\boldsymbol{\psi}), \boldsymbol{\psi} \in \Psi\}$ be a parametric statistical model, with the parameter space $\Psi \subseteq \mathcal{R}^Q$. Let $\mathbf{Y} = (\mathbf{Y}_1', \ldots, \mathbf{Y}_n')'$ denote the data set, where $\mathbf{Y}_i = (y_{i1}, \ldots, y_{im_i})'$ are the vector of observations sampled independently on unit i, $i = 1, \ldots, n$, from a study population. For convenience, we may regard $\mathbf{Y}$ as vectorized data, in which one observation y_{ij} is indexed by $j = 1, \ldots, m_i$ and $i = 1, \ldots, n$. Because the methodology of composite likelihood lies in the idea of dimension reduction for likelihood function, the parameter $\boldsymbol{\psi}$ would be partitioned as $\boldsymbol{\psi} = (\boldsymbol{\theta}, \boldsymbol{\eta})$, where $\boldsymbol{\theta}$ is the parameter of interest to be estimated and $\boldsymbol{\eta}$ is the nuisance parameter that will not be estimated by the composite likelihood method. Consequently, the model selection in composite likelihood methodology is concerned with parameter $\boldsymbol{\theta}$, and the corresponding parameter space is $\Theta \subseteq \mathcal{R}^P$, with dimension P possibly dependent on the sample size.

To form a composite likelihood, first consider a collection of index subsets $\mathcal{A} = \{A : A \subseteq \Omega\}$, where each element A is a subset of $\Omega = \{(i,j), j = 1, \ldots, m_i, i = 1, \ldots, n\}$. For a given unit i, we similarly denote $\mathcal{A}_i = \{A : A \subseteq \Omega_i\}$ by $\Omega_i = \{(i,j), j = 1, \ldots, m_i\}$. This implies that $\Omega = \bigcup_{i=1}^n \Omega_i$. Clearly, the cardinality of set Ω, $\mathrm{card}(\Omega)$, escalates as the sample size n increases. Let $\mathbf{Y}_A$ denote the subset of the data with respect to set A, namely $\mathbf{Y}_A = \{y_{ij}, (i,j) \in A\}$. According to Lindsay (1988), a composite likelihood function is defined as

$$\mathrm{CL}(\boldsymbol{\theta}; \mathbf{Y}) = \prod_{A \in \mathcal{A}} L_A(\boldsymbol{\theta}; \mathbf{Y})^{w_A} = \prod_{i=1}^{n} \prod_{A \in \mathcal{A}_i} L_A(\boldsymbol{\theta}; \mathbf{Y})^{w_A}, \quad (1)$$

where $L_A(\boldsymbol{\theta}; \mathbf{Y}) = f(\mathbf{Y}_A; \boldsymbol{\theta})$ is the marginal likelihood with respect to composite set A, and $\{w_A\}$ is a set of suitable weights. It is easy to see that a singleton $\mathcal{A}_i = \{\Omega_i\}$ corresponds to the full likelihood, and that $\mathcal{A}_i = \{\{1\}, \ldots, \{m_i\}\}$ gives rise to a composite likelihood of univariate margins. The composite log-likelihood is $\mathrm{cl}(\boldsymbol{\theta}; \mathbf{Y}) = \sum_{i=1}^n \sum_{A \in \mathcal{A}_i} w_A \ell_A(\boldsymbol{\theta}; \mathbf{Y})$, where $\mathrm{cl}(\boldsymbol{\theta}; \mathbf{Y}) = \log \mathrm{CL}(\boldsymbol{\theta}; \mathbf{Y})$ and $\ell_A(\boldsymbol{\theta}; \mathbf{Y}) = \log L_A(\boldsymbol{\theta}; \mathbf{Y})$.

Let $\widetilde{\mathrm{cl}}_r^{(1)}(\boldsymbol{\theta})$ denote $\partial\, \mathrm{cl}(\boldsymbol{\theta}; \mathbf{Y})/\partial\theta_{[r]}$, with $\theta_{[r]}$ corresponding to the rth element in $\boldsymbol{\theta}$. Let the composite score vector $\mathbf{U}(\boldsymbol{\theta}; \mathbf{Y}) = \widetilde{\mathbf{cl}}^{(1)}(\boldsymbol{\theta})$ correspond to the vector of first derivatives. The maximum CLE is given by

$$\hat{\boldsymbol{\theta}}^c = \arg\max_{\boldsymbol{\theta} \in \Theta} \mathrm{cl}(\boldsymbol{\theta}; \mathbf{Y}).$$

Because each term in (1) is a likelihood object, the resulting composite score equation $\mathbf{U}(\boldsymbol{\theta}; \mathbf{Y}) = \mathbf{0}$ is unbiased under the assumption that these likelihood objects are valid marginal densities of the underlying joint parametric model $f(\mathbf{y}; \boldsymbol{\psi})$. As usual, the composite likelihood estimate is obtained as a solution to this composite score equation. From the classical theory of estimating functions (e.g., Song 2007, chapter 3), the associated CLE is consistent and asymptotically normally distributed under some mild regularity conditions.

2.2 Bayes Information Criterion

Denote the true full parameter by $\boldsymbol{\psi}_T = (\boldsymbol{\theta}_T, \boldsymbol{\eta}_T) \in \mathrm{int}(\Psi)$ and the true marginal parameter by $\boldsymbol{\theta}_T \in \mathrm{int}(\Theta)$. Consequently, the true full model is $f(\mathbf{y}; \boldsymbol{\psi}_T)$ and the true marginal model constitutes a set of true composite marginal densities $\{f(\mathbf{y}_A; \boldsymbol{\theta}_T), A \in \mathcal{A}\}$.

To derive BIC in the composite likelihood framework, we need some additional notations. Let $P = \dim(\Theta)$, and let s be a subset of $\{1, \ldots, P\}$. Denote by $\boldsymbol{\theta}_s$ the parameter $\boldsymbol{\theta}$ with those elements outside s being pre-specified as 0 or some known values. Because set s and a candidate marginal submodel $\{f(\mathbf{y}_A; \boldsymbol{\theta}_s), A \in \mathcal{A}\}$ correspond to each other uniquely, this submodel is simply denoted by s for convenience. Consequently, set $T \subseteq \{1, \ldots, P\}$ denotes the true marginal model.

Let d_s be the number of parameters under a marginal submodel s. Let $\mathcal{S}$ denote the model space of all possible submodels being considered. Associated with each submodel s, let $p(s)$ be the prior probability of the occurrence of the submodel defined on space $\mathcal{S}$.

In the conventional setting where the number of parameters P is fixed (or not dependent on the sample size n), it is commonly assumed that each submodel s has an equal probability of being selected, namely a uniform prior over the model space, $p(s) = 1/\mathrm{card}(\mathcal{S})$, where $\mathrm{card}(\mathcal{S})$ is the cardinality of $\mathcal{S}$. Under the full likelihood framework, assuming equal priors for different submodels, Schwarz (1978) proposed the BIC criterion to select the best model among all the candidate models. The first term in BIC is minus twice the log-likelihood evaluated at the maximum likelihood estimate and the second term is $\log(n)$ times the number the parameters in the model.

A much more challenging task of model selection in high-dimensional data analysis is that P is not fixed but increases as the sample size rises. Suppose that $P = O(n^\kappa)$, with $\kappa > 0$. In this case, the equal probability prior will actually favor models with more parameters; see for example Chen and Chen (2008). In many practical studies, important attributes are typically only a handful, in spite of a large P. This naturally necessitates the imposition of lower preferences on models with a large number of parameters; in other words, an additional penalty is required

in BIC to ensure an increasing chance of selecting models with sparsity. This can be done by assigning priors through a stratified sampling scheme proposed by Chen and Chen (2008). To proceed, first partition the model space into submodel spaces $\mathcal{S} = \bigcup_{k=1}^{P} \mathcal{S}_k$, where each $\mathcal{S}_k$ contains models with k parameters. For example, $\mathcal{S}_1$ is a collection of all the models containing one parameter. Let $\tau(\mathcal{S}_k) = \text{card}(\mathcal{S}_k)$ be the size of $\mathcal{S}_k$. Obviously, $\tau(\mathcal{S}_1) = P$. Within a given subspace $\mathcal{S}_k$, an equal probability prior is imposed as $p(s|\mathcal{S}_k) = 1/\tau(\mathcal{S}_k)$, $s \in \mathcal{S}_k$. Moreover, specifying prior probabilities for these subspaces proportional to their sizes, say $p(\mathcal{S}_k) \propto \{\tau(\mathcal{S}_k)\}^{\xi}$ for some $\xi \le 1$, we obtain that the prior probability of a submodel s being selected via this stratified sampling procedure is proportional to $\tau(\mathcal{S}_k)^{-\gamma}$, with $\gamma = 1 - \xi > 0$. Using such prior probabilities, Chen and Chen (2008) have proposed an extended BIC criterion which has an extra penalty term $2\gamma \log \tau(\mathcal{S}_k)$ for $s \in \mathcal{S}_k$ on the model space complexity.

When the full likelihood is numerically prohibitive to compute, we aim to develop an analogue of extended BIC criterion based on the composite likelihood. It is natural to generalize the approach of BIC and select the model with the highest composite posterior probability. In this spirit, for any model s, we define the composite posterior probability as follows:

$$P_c(s|\mathbf{Y}) = \frac{p(s) \int \text{CL}(\mathbf{Y}|\boldsymbol{\theta}_s)\pi_s(\boldsymbol{\theta}_s)\, d\boldsymbol{\theta}_s}{\sum_{s\in\mathcal{S}} p(s) \int \text{CL}(\mathbf{Y}|\boldsymbol{\theta}_s)\pi_s(\boldsymbol{\theta}_s)\, d\boldsymbol{\theta}_s},$$

with $\pi_s(\boldsymbol{\theta}_s)$ denoting the prior density of $\boldsymbol{\theta}_s$. It is assumed that $\log \pi_s(\boldsymbol{\theta}_s) = O(1)$, and $\pi_s(\boldsymbol{\theta}_s)$ is sufficiently flat in the neighborhood of $\hat{\boldsymbol{\theta}}_s^c$. Using the Laplace approximation (Tierney and Kadane 1986; Tierney, Kass, and Kadane 1989), and ignoring $O_p(1)$ terms, we have the resulting criterion simplified as:

$$-2\log \text{CL}(\hat{\boldsymbol{\theta}}_s^c; \mathbf{Y}) + d_s \log(n) + 2\gamma \log\{\tau(\mathcal{S}_{d_s})\}. \quad (2)$$

Extra consideration is needed regarding the measure of model complexity in the context of composite likelihood methodology. The asymptotic distribution of composite log-likelihood ratio statistic is a weighted sum of χ_1^2 distribution, with the total weights equal to the effective degrees of freedom $d_s^* = \text{trace}(\mathbf{H}_s^{-1}\mathbf{V}_s)$, where

$$\begin{aligned} \mathbf{H}_s &= \text{E}_{\boldsymbol{\psi}_{T,0}}\{-\widetilde{\mathbf{cl}}^{(2)}(\boldsymbol{\theta}_s)\} \qquad \text{and} \\ \mathbf{V}_s &= \text{var}_{\boldsymbol{\psi}_{T,0}}\{\widetilde{\mathbf{cl}}^{(1)}(\boldsymbol{\theta}_s)\}. \end{aligned} \quad (3)$$

Here $\widetilde{\mathbf{cl}}^{(2)}(\boldsymbol{\theta})$ denotes the matrix of second order derivatives $\widetilde{\text{cl}}_{rt}^{(2)}(\boldsymbol{\theta}) = \partial^2 \text{cl}(\boldsymbol{\theta}; \mathbf{Y})/\partial\theta_{[r]}\,\partial\theta_{[t]}$. The expected value and the variance of the composite log-likelihood ratio statistic are d_s^* and $2d_s^*$, respectively. The d_s^* has been accepted as a measure of model complexity in composite likelihood setting (Varin and Vidoni 2005). Such modification is necessary to ensure the selection consistency as shown in a subsequent section. Thus we propose to replace d_s in (2) by d_s^*. The proposed CL-BIC for model selection is given as follows:

$$\text{CL-BIC}(s) = -2\log \text{CL}(\hat{\boldsymbol{\theta}}_s^c; \mathbf{Y}) + d_s^* \log(n) + 2\gamma \log\{\tau(\mathcal{S}_{d_s^*})\}, \quad (4)$$

with the cardinality term $\tau(\mathcal{S}_{d_s^*}) = P^{d_s^*}$. In (4), the first term is minus twice of the composite log-likelihood that reflects the goodness-of-fit for a given model s, the second term is the penalty for the model complexity; and the third term is the penalty on the model space complexity that enforces sparsity on any model selected. The coefficient γ tunes the degree of preference for large sized models. The larger the γ, the more favorable a sparse model becomes.

Now we briefly discuss an efficient estimation of d_s^*. Let a consistent estimator of d_s^* be denoted as $\hat{d}_s^* = \text{trace}(\hat{\mathbf{H}}_s^{-1}\hat{\mathbf{V}}_s)$. For matrix $\mathbf{H}_s$, under standard regularity conditions, a consistent estimator is the negative Hessian matrix evaluated at the maximum composite likelihood estimator:

$$\hat{\mathbf{H}}_s = -\widetilde{\mathbf{cl}}^{(2)}(\boldsymbol{\theta}_s)\big|_{\hat{\boldsymbol{\theta}}_s^c}.$$

If the Hessian is difficult to compute, an alternative estimator is

$$\hat{\mathbf{H}}_s = -\sum_{i=1}^{n}\sum_{A\in\mathcal{A}_i} w_A \left(\frac{\partial \ell_A(\boldsymbol{\theta}_s, \mathbf{Y})}{\partial \boldsymbol{\theta}_s}\bigg|_{\hat{\boldsymbol{\theta}}_s^c}\right)\left(\frac{\partial \ell_A(\boldsymbol{\theta}_s, \mathbf{Y})}{\partial \boldsymbol{\theta}_s}\bigg|_{\hat{\boldsymbol{\theta}}_s^c}\right)',$$

as the second Bartlett identity remains true for each subset.

The estimation of $\mathbf{V}_s$ poses more difficulties, since the corresponding naive estimator

$$\hat{\mathbf{V}}_s = \left(\sum_{i=1}^{n}\sum_{A\in\mathcal{A}_i} w_A \frac{\partial \ell_A(\boldsymbol{\theta}_s, \mathbf{Y})}{\partial \boldsymbol{\theta}_s}\bigg|_{\hat{\boldsymbol{\theta}}_s^c}\right) \times \left(\sum_{i=1}^{n}\sum_{A\in\mathcal{A}_i} w_A \frac{\partial \ell_A(\boldsymbol{\theta}_s, \mathbf{Y})}{\partial \boldsymbol{\theta}_s}\bigg|_{\hat{\boldsymbol{\theta}}_s^c}\right)'$$

vanishes when evaluated at the maximum composite likelihood estimator. If all the $\mathbf{Y}_1, \ldots, \mathbf{Y}_n$ are independent, $\mathbf{V}_s$ can be estimated by the sample variances of the individual contributions to the composite score function. An interesting alternative is to perform jackknife (Zhao and Joe 2005) for the evaluation of the variance matrix. For non-independent samples, one might partition the sample $\mathbf{Y}$ so that the corresponding contributions to the composite score function are approximately uncorrelated. Then the empirical and jackknife estimation can be derived based on these contributions. A more detailed discussion on the estimation of $\mathbf{V}_s$, especially for time series and spatial data, may be found in Varin (2008).

3. MODEL SELECTION CONSISTENCY

Given all the competing models in model space, the notion of consistent model selection is about identifying the smallest correct model with probability tending to one as the sample size increases. We further assume that under the data generating mechanism, the number of parameters in the true model is bounded by a constant K. In what follows, the model space $\mathcal{S}$ is restricted to $\{s : d_s \le K\}$. Given an arbitrary model s, it may be one of the following three scenarios: (i) the true marginal model T, with the parameter vector $\boldsymbol{\theta}_T$ containing d_T components; (ii) an under-fitting marginal model $s-$ with $\boldsymbol{\theta}_{s-} \not\supseteq \boldsymbol{\theta}_T$; (iii) an over-fitting marginal model $s+$ with $\boldsymbol{\theta}_T \subset \boldsymbol{\theta}_{s+}$ and $\boldsymbol{\theta}_{s+} \neq \boldsymbol{\theta}_T$. The corresponding sets of under-fitting models and over-fitting models are denoted as S_- and S_+, respectively. Let CL-BIC(s), $s = T, s-, s+$, denote the composite likelihood BIC criteria obtained under the true (T), under-fitting ($s-$), and over-fitting marginal models ($s+$). It is also worthy to emphasize that throughout this article, we focus on composite likelihood, thus we are only dealing with marginal models that are parameterized in composite likelihood formulation.

In this paper, we assume the conventional regularity conditions required for consistency and asymptotic normality of the maximum likelihood estimator (Cox and Hinkley 1974). Furthermore, we assume several additional regularity conditions needed by composite likelihood estimation in connection to model misspecification (White 1982; Varin and Vidoni 2005), detailed as follows.

Assumption 1 (A1). For each submodel s, the parameter space Θ_s is a compact subset of $\mathcal{R}^{d_s}$, and for fixed $\mathbf{Y}$, $\mathrm{cl}(\boldsymbol{\theta}_s; \mathbf{Y})$ is twice continuously differentiable with respect to $\boldsymbol{\theta}_s$.

Assumption 2 (A2). (a) For each submodel s, $|\mathrm{cl}(\boldsymbol{\theta}_s; \mathbf{Y})|$, $|\widetilde{\mathrm{cl}}_i^{(1)}(\boldsymbol{\theta}_s; \mathbf{Y}) \cdot \widetilde{\mathrm{cl}}_j^{(1)}(\boldsymbol{\theta}_s; \mathbf{Y})|$, $|\widetilde{\mathrm{cl}}_{ij}^{(2)}(\boldsymbol{\theta}_s; \mathbf{Y})|, i, j = 1, \ldots, d_s$, are dominated by functions integrable with respect to the probability measure of the true full model for all $\boldsymbol{\theta}_s \in \Theta_s$. (b) Denote the composite log-likelihood ratio (CLR) between two marginal submodels s and s' by

$$\lambda_{s'|s}(\mathbf{Y}; \boldsymbol{\theta}_{s'}, \boldsymbol{\theta}_s) = \log\left\{\frac{\mathrm{CL}(\boldsymbol{\theta}_{s'}; \mathbf{Y})}{\mathrm{CL}(\boldsymbol{\theta}_s; \mathbf{Y})}\right\} = \mathrm{cl}(\boldsymbol{\theta}_{s'}; \mathbf{Y}) - \mathrm{cl}(\boldsymbol{\theta}_s; \mathbf{Y}). \tag{5}$$

Assume $\mathrm{E}_{\boldsymbol{\psi}_{T,0}}\{\lambda_{T|s}(\mathbf{Y}; \boldsymbol{\theta}_{T,0}, \boldsymbol{\theta}_s)\}$ exists for all $\boldsymbol{\theta}_s$, and has a unique minimum at $\boldsymbol{\theta}_{s,0} \in \mathrm{int}(\Theta_s)$. Here $\boldsymbol{\psi}_{T,0}$ is the true value of the parameter $\boldsymbol{\psi}_T$ under the true full model $f(\mathbf{y}; \boldsymbol{\psi})$.

It is easy to see that this $\boldsymbol{\theta}_{s,0}$ effectively defines the pseudo true value of parameter $\boldsymbol{\theta}_s$ in Θ_s under a misspecified model s, which minimizes the expected composite KL distance (Varin and Vidoni 2005) between the true marginal model and a marginal submodel s. That is, $\boldsymbol{\theta}_{s,0} = \arg\min_{\boldsymbol{\theta}_s \in \Theta_s} \mathrm{E}_{\boldsymbol{\psi}_{T,0}}\{\lambda_{T|s}(\mathbf{Y}; \boldsymbol{\theta}_{T,0}, \boldsymbol{\theta}_s)\}$.

Assumption 3 (A3). The random vectors $\mathbf{Y}_1, \ldots, \mathbf{Y}_n$ are all independently and identically distributed. The composite likelihood estimator $\hat{\boldsymbol{\theta}}_s^c$ is consistent, $\hat{\boldsymbol{\theta}}_s^c \xrightarrow{p} \boldsymbol{\theta}_{s,0}$, and asymptotically normally distributed, $\sqrt{n}(\hat{\boldsymbol{\theta}}_s^c - \boldsymbol{\theta}_{s,0}) \xrightarrow{d} \mathrm{N}_{d_s}(0, \mathbf{G}_s^{-1})$, with $\mathbf{G}_s = \mathbf{H}_s^{-1}\mathbf{V}_s\mathbf{H}_s^{-1}$.

Next we provide the assumption to ensure model identifiability. Between the true model and a competing model s, we examine the standardized expected composite KL distance:

$$\mathrm{E}_{\boldsymbol{\psi}_{T,0}}\{\lambda_{T|s}(\mathbf{Y}; \boldsymbol{\theta}_{T,0}, \boldsymbol{\theta}_{s,0})\} / \left[\mathrm{var}_{\boldsymbol{\psi}_{T,0}}\left\{\lambda_{T|s}(\mathbf{Y}; \boldsymbol{\theta}_{T,0}, \boldsymbol{\theta}_{s,0})\right\}\right]^{1/2}.$$

In the traditional setting of P being fixed, the space of all possible candidate models is also fixed. The model at which the minimum of such distance is attained remains the same as sample size increases. Therefore the minimum of such distance between the true model and other competing models is of order $O(\sqrt{n})$. When P increases with the sample size n, the true model is still fixed but the potential candidate model space increases with the sample size. This imposes the problem that the model at which the minimum distance is attained will change as the model space changes. The minimum distance between the true model and any other competing model may be growing at a rate slower than $\sqrt{n}$. In order to ensure the identifiability of the true model, a necessary lower bound is needed for such minimum distance.

Assumption 4 (A4). Both of the following conditions are assumed to hold:

$$\lim_{n\to\infty} \min_{s\in S_-} \left\{ (\log n)^{-1/2} \frac{\mathrm{E}_{\boldsymbol{\psi}_{T,0}}\{\lambda_{T|s}(\mathbf{Y}; \boldsymbol{\theta}_{T,0}, \boldsymbol{\theta}_{s,0})\}}{[\mathrm{var}_{\boldsymbol{\psi}_{T,0}}\{\lambda_{T|s}(\mathbf{Y}; \boldsymbol{\theta}_{T,0}, \boldsymbol{\theta}_{s,0})\}]^{1/2}} \right\} = \infty, \tag{6}$$

$$\liminf_{n\to\infty} \min_{s\in S_-} \left\{ (\log n)^{-1/2} \left[\mathrm{var}_{\boldsymbol{\psi}_{T,0}}\left\{\lambda_{T|s}(\mathbf{Y}; \boldsymbol{\theta}_{T,0}, \boldsymbol{\theta}_{s,0})\right\}\right]^{1/2} \right\} \geq C_1 \tag{7}$$

for a positive constant C_1.

In effect, Equation (6) in Assumption 4 implies that as n increases, the minimum standardized expected KL distance should increase at a rate greater than $\sqrt{\log n}$. Equation (7) in Assumption 4 requires that the minimum variance of KL distance should increase at least at a rate of $\log n$. It is a generalization of the asymptotical identifiability condition given by Chen and Chen (2008) in the linear model setting. Consider a model $\mathbf{Y} = \mathbf{X}\boldsymbol{\theta} + \boldsymbol{\epsilon}$, where $\boldsymbol{\epsilon} \sim \mathrm{N}_n(0, \sigma^2\mathbf{I})$. Let $\mathbf{X}_T$ and $\mathbf{X}_s$ denote the design matrices of the true model and a candidate model with respective vectors of the regression coefficients $\boldsymbol{\theta}_T$ and $\boldsymbol{\theta}_s$. Denote the true null value as $\boldsymbol{\theta}_{T,0}$ under the true model and the pseudo null value as $\boldsymbol{\theta}_{s,0}$ under the candidate model. Then Assumption 4 given in (6) and (7) reduces to the condition in Chen and Chen (2008):

$$\lim_{n\to\infty} \min_{s\in S_-} \{(\log n)^{-1} \Delta_n(s)\} = \infty, \tag{8}$$

with $\Delta_n(s) = \|\mathbf{X}_T\boldsymbol{\theta}_T - \mathbf{X}_s(\mathbf{X}_s'\mathbf{X}_s)^{-1}\mathbf{X}_s'\mathbf{X}_T\boldsymbol{\theta}_{T,0}\|_2^2$.

To establish the consistency result, we need a set of additional regularity assumptions regarding the uniform boundedness of the moments of the derivatives of the composite log-likelihood. This set of Assumptions A.1–A.2 are listed in the Appendix. Next we state the main results.

Theorem 1. Under the regularity conditions (Assumptions 1–4, A.1, A.2),

$$P_{\psi_{T,0}}\left\{\min_{s\in S_-} \text{CL-BIC}(s) > \text{CL-BIC}(T)\right\} \to 1,$$

as $n \to \infty$.

Next we consider the over-fitting scenario. Define the model space $S_+(m) \subset S_+$, with $S_+(m) = \{s : s \in S_+, d_s - d_T = m\}$, $m = 1, \ldots, K - d_T$. For any over-fitting model s, define a matrix $\mathbf{D}_s = (\mathbf{I}_{d_T}, \mathbf{0}_{d_T, d_s - d_T})$, with $\mathbf{I}_{d_T}$ being an identity matrix of dimension $d_T \times d_T$, and $\mathbf{0}_{d_T, d_s - d_T}$ denoting a matrix of zeros with dimension $d_T \times (d_s - d_T)$. Let $\mathbf{M}_{s/T}$ denote the difference matrix $(\mathbf{H}_s(\boldsymbol{\theta}_{s,0})^{-1} - \mathbf{D}_s'\mathbf{H}_T^{-1}(\boldsymbol{\theta}_{T,0})\mathbf{D}_s)$. Let $\lambda_{s[1]}, \ldots, \lambda_{s[m]}$ denote the nonzero eigenvalues of $\mathbf{M}_{s/T}^{1/2}\mathbf{V}_s(\boldsymbol{\theta}_{s,0})\mathbf{M}_{s/T}^{1/2}$ in ascending order and $\overline{\lambda}_s = \sum_{j=1}^m \lambda_{s[j]}/m$. Define

$$\varpi = \limsup_{n\to\infty} \max_{s\in S_+} \left(\lambda_{s[m]}/\overline{\lambda}_s\right).$$

When all the eigenvalues are equal, the ratio of the maximum eigenvalue over the mean eigenvalue, $\lambda_{s[m]}/\overline{\lambda}_s$, is one. On the other hand, $\lambda_{s[m]}/\overline{\lambda}_s < m$. Thus ϖ resides in interval $[1, K - d_T)$.

Theorem 2. Under the regularity conditions (Assumptions 1–4, A.1, A.2), when $\gamma > \varpi - 1/(2\kappa)$,

$$P_{\psi_{T,0}}\left\{\min_{s\in S_+} \mathrm{CL}-\mathrm{BIC}(s) > \mathrm{CL}-\mathrm{BIC}(T)\right\} \to 1,$$

as $n \to \infty$.

4. SIMULATION

To examine the performance of the proposed CL-BIC, we conduct three Monte Carlo simulation experiments to select significant parameters in the setting of high-dimensional data.

4.1 Multivariate Normal Model

We consider the multivariate familial data analysis discussed in Zhao and Joe (2005). The sample is drawn from families with inter-correlations among individuals in a family. Denote the numbers of families and members in each family by n and m. The response vector of measurements for the ith family is denoted by $\mathbf{Y}_i = (y_{i1}, \ldots, y_{im})'$. Associated is a set of covariates at the individual level, $\mathbf{X}_i = (\mathbf{x}_{i1}, \ldots, \mathbf{x}_{im})'$, with $\mathbf{x}_{ik} = (\mathbf{x}_{ik1}, \ldots, \mathbf{x}_{ikP})'$, representing the P covariates observed for the kth individual in the ith family. The first simulation is focused on a multivariate normal model, in which $\mathbf{Y}_i$ follows a multivariate normal distribution, $\mathrm{N}_m(\boldsymbol{\mu}_i, \boldsymbol{\Sigma})$, where the mean vector is governed by a linear model, $\boldsymbol{\mu}_i = \mathbf{X}_i\boldsymbol{\beta}$, with $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_P)'$. The covariance matrix $\boldsymbol{\Sigma}$ is specified according to an exchangeable dependence structure, $\sigma_{k,k'} = \rho$. In addition, among all the covariate coefficients, most of them are zero while a small subset are nonzero. We wish to select the significant covariates among the P candidates.

We consider two different scenarios. In the first scenario, we set $P = 30$, $n = 200$ and $m = 4$. The covariates are generated from a multivariate normal with the standard normal $\mathrm{N}(0, 1)$ marginals and inter-correlation $\mathrm{Cov}(x_{ikp}, x_{ikp'}) = 0.2$. The within-family correlation ρ is set to either 0.3 or 0.6. The regression coefficients of the true marginal model are set to two cases $\boldsymbol{\beta}_1(T) = (0.1, 0.2, 0.4, 0.1, 0.4, 0.2, 0.3, 0.4, 0.5, 0.3)'$, or $\boldsymbol{\beta}_2(T) = (0.5, 0.1, 0.4, 0.3, 0.5, 0.1, 0.004, 0.04, 0.03, 0.003)'$, with the other 20 coefficients set to zero. The second case contains 10 nonzero coefficients, while four of them are too small and considered to be not useful and not be used to compute the positive selection rates. This setup can help us evaluate the performance of the model selection criteria when the covariates have different levels of effect. In the second scenario, we set $P = 1000$, $n = 200$, and $m = 4$. The covariates are partitioned into 20 blocks of 50 each. Within each block, the covariates are generated from a multivariate normal with univariate standard normal marginals and equal inter-covariate correlation 0.2, and covariates from different blocks have zero correlations. Similarly, the within-family correlation ρ is set to either 0.3 or 0.6. The regression coefficients of the true marginal model are set to the same values as those in scenario I with the other 990 coefficients set to zero.

We impose penalization on the composite likelihood with L_1 penalty. We gradually increase the tuning parameter in the penalty term and obtain a sequence of nested models. Under scenario II with $P \gg n$, we randomly partition the 1000 covariates into 8 disjoint subsets of 125 covariates each and apply the penalized composite likelihood on each subset. We then pool the reduced subsets of covariates together and perform the same procedure to obtain the sequence of nested models, at which the CL-BIC is computed to determine the optimal tuning.

For each candidate model, the CL-BIC is evaluated under either the univariate composite likelihood $\sum_{i=1}^{n}\sum_{k=1}^{m} \mathrm{cl}(y_{ik}; \beta)$, or the pairwise composite log-likelihood $\sum_{i=1}^{n}\sum_{k<k'} \mathrm{cl}(y_{ik}, y_{ik'}; \beta)$. The resulting two CL-BIC criteria are denoted as CL_U-BIC or CL_B-BIC, respectively. For the purpose of comparison, we also include Varin and Vidoni's CL_U-AIC based on the univariate composite likelihood, and Chen and Chen's EBIC based on the full likelihood that serves as the gold standard. The two versions of CL-BIC and the EBIC are calculated with $\gamma = 0$, and 0.5 for scenario I and with $\gamma = 0$, 0.5 1.0 for scenario II.

For each setting, 100 simulated data sets are generated. Tables 1 and 2 summarize the performance of the different information criteria. The positive selection rate (PSR) is defined as the ratio of identified significant predictors among all the significant predictors. The false discovery rate (FDR) is defined as the ratio of falsely identified predictors among all the identified predictors. In a multiple testing framework, the positive selection rates reflect the power or sensitivity of the test, and the false discovery rate reflects the error rate or selectivity of the test.

Table 1 provides the performance of different methods when $n = 200$, and $P = 30$. we observe that the strength of correlation does mildly affect the performance of different methods. But the relative comparison among different methods exhibits the same pattern at different correlation levels. As CL_U-AIC has less penalty on the model complexity, it always achieves higher PSR than CL_U-BIC and CL_B-BIC. Under such a modest sample size and small P setting, all the information criteria have satisfactory FDR control. With regard to the size of γ for

Table 1. Positive selection rates (PSR) and false discovery rates (FDR) on multivariate normal model with $P = 30$ and $N = 200$

β	ρ_y	Rate	CL-AIC	CL_U-BIC_0	CL_U-$\mathrm{BIC}_{0.5}$	CL_B-BIC_0	CL_B-$\mathrm{BIC}_{0.5}$	EBIC_0	$\mathrm{EBIC}_{0.5}$
β_1	0.3	PSR	0.878	0.739	0.655	0.911	0.875	0.914	0.877
		FDR	0.035	0.003	0.002	0.037	0.011	0.034	0.008
	0.6	PSR	0.873	0.727	0.668	0.946	0.903	0.949	0.913
		FDR	0.026	0.002	0.002	0.053	0.014	0.055	0.017
β_2	0.3	PSR	0.852	0.697	0.667	0.892	0.818	0.892	0.825
		FDR	0.108	0.005	0.004	0.116	0.045	0.128	0.041
	0.6	PSR	0.845	0.695	0.663	0.938	0.890	0.940	0.888
		FDR	0.095	0.014	0.006	0.142	0.065	0.135	0.060

Table 2. Positive selection rates (PSR) and false discovery rates (FDR) on multivariate normal model with $P = 1000$ and $n = 200$

β	ρ_y	Rate	CL-AIC	CL_U-BIC$_0$	CL_U-BIC$_{0.5}$	CL_U-BIC$_{1.0}$	CL_B-BIC$_0$	CL_B-BIC$_{0.5}$	CL_B-BIC$_{1.0}$	EBIC$_0$	EBIC$_{0.5}$	EBIC$_{1.0}$
β_1	0.3	PSR	0.896	0.758	0.611	0.505	0.893	0.819	0.789	0.889	0.818	0.789
		FDR	0.472	0.035	0.002	0.000	0.439	0.039	0.012	0.378	0.037	0.011
	0.6	PSR	0.894	0.766	0.601	0.508	0.894	0.837	0.814	0.881	0.838	0.809
		FDR	0.456	0.046	0.004	0.000	0.346	0.052	0.026	0.211	0.052	0.023
β_2	0.3	PSR	0.868	0.687	0.622	0.582	0.850	0.717	0.693	0.842	0.710	0.692
		FDR	0.780	0.025	0.009	0.002	0.641	0.044	0.014	0.545	0.032	0.014
	0.6	PSR	0.873	0.688	0.612	0.575	0.847	0.728	0.703	0.815	0.722	0.702
		FDR	0.783	0.033	0.009	0.006	0.535	0.064	0.020	0.316	0.053	0.020

the CL-BIC criteria, $\gamma = 0.5$ or higher seems unnecessary, and it attenuates the power. Therefore, using $\gamma = 0$ is recommended here by both CL-BIC and EBIC. The CL_B-BIC always achieve higher PSR than CL_U-BIC, demonstrating the power gain by using the pairwise likelihoods rather than the univariate likelihoods. Compared to the full likelihood based EBIC, CL_B-BIC has shown PSR and FDR very close to those of EBIC. This demonstrates that under the exchangeable correlation structure, the discrepancy between the pairwise likelihood and the full likelihood is very little.

Table 2 provides the performance of different methods when $n = 200$, and $P = 1000$. With such a large number of covariates, the CL_U-AIC does not adequately control the FDR rate. It seems that CL_B-BIC$_{0.5}$ has a satisfactory performance and controls the FDR rate very well. The penalty with $\gamma = 1$ seems too harsh, and it attenuates the power. Therefore, when $P = 1000$ and $n = 200$, CL_B-BIC$_{0.5}$ is recommended. The CL_B-BIC always achieves higher PSR than does CL_U-BIC, suggesting the importance of incorporating correlation in the composite likelihood. The performance of CL_B-BIC is very close to that of EBIC.

4.2 Multivariate Probit Model

The second simulation study is based on a multivariate probit model, in which the binary response vector arises from a dichotomization of an underlying multivariate normally distributed random vector. Under the same setup in Section 4.1, binary correlated responses are obtained by dichotomizing the continuous multivariate normal measurements. Also, the two scenarios of $P < n$ and $P \gg n$ are considered. For a multivariate probit model with many covariates, the full likelihood involves high dimensional integration and is computationally prohibitive. We thus compare the performance of the different information criteria under only composite likelihood methodology, including CL_U-AIC, CL_U-BIC, and CL_B-BIC. For each setting, 100 simulated data sets are generated. Results are summarized in Tables 3 and 4. It is noted that even with $P = 30$, and $n = 100$, the over-fitting effect of CL_U-AIC is exhibited. When $P = 1000$, the FDR of CL_U-AIC is about 50 to 70 percent, indicating an inadequate control of the error rate. The CL_B-BIC always has higher PSR than does CL_U-BIC because of the advantage of using pairwise likelihoods over univariate likelihoods. When $P = 30$, the penalty term with $\gamma = 0$ is sufficient to maintain a good FDR for CL_B-BIC. When $P = 1000$, the penalty term with $\gamma = 0.5$ is needed to control the error rate. Thus for the multivariate probit model, the CL_B-BIC is recommended for its computational feasibility and simplicity compared to the full likelihood approach, and also it clearly provides a satisfactory performance in terms of sensitivity and selectivity.

4.3 Quadratic Exponential Model

In the above two simulations, we examine the performance of CL-BIC when univariate and pairwise likelihoods are involved in the composite likelihood formulation. In the third simulation, we will consider a less simple formulation involving conditional likelihoods which has been discussed by Geys, Molenberghs, and Ryan (1997, 1999) and Hanfelt (2004), among others.

Consider an experiment involving n clusters, the ith of which contains n_i binary measurements. Suppose $y_{ij} = 1$ when the outcome is success and $y_{ij} = -1$ when the outcome is failure. Let $\mathbf{Y}_i$ represent the vector of outcomes for the ith cluster. Geys,

Table 3. Positive selection rates (PSR) and false discovery rates (FDR) on multivariate probit model with $P = 30$ and $n = 100$

β	ρ_y	Rate	CL-AIC	CL_U-BIC$_0$	CL_U-BIC$_{0.5}$	CL_B-BIC$_0$	CL_B-BIC$_{0.5}$
β_1	0.3	PSR	0.846	0.710	0.670	0.768	0.682
		FDR	0.248	0.068	0.060	0.111	0.063
	0.6	PSR	0.850	0.713	0.675	0.769	0.707
		FDR	0.233	0.067	0.052	0.104	0.063
β_2	0.3	PSR	0.812	0.693	0.687	0.707	0.692
		FDR	0.394	0.079	0.071	0.111	0.078
	0.6	PSR	0.813	0.703	0.695	0.735	0.693
		FDR	0.363	0.089	0.065	0.130	0.069

Table 4. Positive selection rates (PSR) and false discovery rates (FDR) on multivariate probit model with $P = 1000$ and $n = 100$

β	ρ_y	Rate	CL-AIC	CL_U-BIC_0	CL_U-$BIC_{0.5}$	CL_U-$BIC_{1.0}$	CL_B-BIC_0	CL_B-$BIC_{0.5}$	CL_B-$BIC_{1.0}$
β_1	0.3	PSR	0.790	0.766	0.593	0.393	0.782	0.647	0.475
		FDR	0.522	0.431	0.118	0.029	0.494	0.169	0.055
	0.6	PSR	0.778	0.756	0.571	0.398	0.775	0.640	0.500
		FDR	0.540	0.448	0.095	0.024	0.516	0.181	0.058
β_2	0.3	PSR	0.865	0.840	0.635	0.540	0.863	0.692	0.588
		FDR	0.703	0.587	0.092	0.012	0.696	0.163	0.031
	0.3	PSR	0.868	0.828	0.637	0.518	0.858	0.718	0.592
		FDR	0.711	0.590	0.087	0.012	0.678	0.167	0.036

Molenberghs, and Ryan (1997) used the following model for the joint distribution of clustered binary data:

$$f_{Y_i}(y_i) \propto \exp\left\{\sum_{j=1}^{n_i} \mu_{ij} y_{ij} + \sum_{j \le j'} w_{ijj'} y_{ij} y_{ij'}\right\}, \tag{9}$$

which belongs to the quadratic exponential family discussed in Zhao and Prentice (1990).

It is more convenient to express the joint distribution in terms of z_i, the number of successes for the ith cluster. Assuming $\mu_{ij} \equiv \mu_i$ and $w_i \equiv w$, and through reparametrization $\mu_i^* = 2\mu_i$, and $w^* = 2w$, model (9) is transformed into:

$$f_{\mathbf{Y}_i}(\mathbf{y}_i) = \exp\{\mu_i^* z_i + w^*(-z_i(n_i - z_i)) - C(\mu_i^*, w^*)\},$$

with $C(\mu_i^*, w^*)$ being the normalizing constant. A positive interaction effect w^* corresponds to classical clustering or over-dispersion, while a negative value corresponds to under-dispersion.

Using traditional likelihood approach to analyze such data will inevitably involve highly intensive calculation of the normalizing constant $C(\mu_i^*, w^*)$, which varies across clusters of different sizes. As an appealing alternative method, we formulate the composite likelihood in the form of conditional likelihoods.

$$\mathrm{cl} = \sum_{i=1}^{n} \sum_{j=1}^{n_i} \log f(y_{ij} | \{y_{ij'}\}, j' \neq j).$$

To agree with the general form in Equation (1), this formulation can be viewed as having weight n_i for the index subset $(1, 2, \ldots, n_i)$, and having weight -1 for all the index subsets containing $n_i - 1$ distinct indices, and having weight 0 for any other subsets. Within each cluster, there are n_i conditional probabilities of observing the outcome for the jth measurement, given the outcome for the other $n_i - 1$ measurements. Under the assumption of the exchangeable nature of the measurement, there are two types of contributions: (i) the conditional probability of an additional success result, given there are $z_i - 1$ successes and $n_i - z_i$ failures:

$$p_{is} = \frac{\exp\{\mu_i^* - w^*(n_i - z_i + 1)\}}{1 + \exp\{\mu_i^* - w^*(n_i - z_i + 1)\}},$$

(ii) the conditional probability of an additional failure, given there are z_i successes and $n_i - z_i - 1$ failures:

$$p_{if} = \frac{\exp\{-\mu_i^* + w^*(n_i - z_i - 1)\}}{1 + \exp\{-\mu_i^* + w^*(n_i - z_i - 1)\}}.$$

Thus, the composite likelihood can be expressed as $\mathrm{cl} = \sum_{i=1}^{n} \{z_i \log p_{is} + (n - z_i) \log p_{if}\}$.

Modelling in terms of covariate effect can be achieved using the linear model $\mu_i^* = \mathbf{X}_i \boldsymbol{\beta}$, where $\mathbf{X}_i$ is a $1 \times P$ vector containing the covariate values and $\boldsymbol{\beta}$ is a $P \times 1$ vector of regression coefficients. Under this conditional likelihood framework, we conduct the following simulation to select the true covariates that affect the parameters μ_i^*. Within each simulated data set, n clusters are generated. Within each cluster, n_i binary measurements are simulated, where n_i ranges from 4 to 8. The binary measurements are simulated according to the quadratic exponential model (9), with $\mu_i^* = X_i \beta$. The design matrix entry x_{ij}s are randomly drawn from $\mathrm{N}(0, 1)$. The vector β contains 1000 regression coefficients, in which only ten coefficients are nonzero and equal to $0.2, 0.2, 0.4, 0.3, 0.4, 0.2, 0.3, 0.4, 0.5, 0.3$; the remaining 990 covariates effects are set to zero. We impose penalization on the composite likelihood with L_1 penalty. We gradually increase the tuning parameter in the penalty term and obtain a sequence of models. For each candidate model s, we compute the ratio of the maximum eigenvalue over the mean eigenvalue of the matrix $\hat{\mathbf{H}}_s^{-1/2} \hat{\mathbf{V}}_s \hat{\mathbf{H}}_s^{-1/2}$. The maximum ratio over all the models being examined offers an ad-hoc estimator $\hat{\varpi}$ of the quantity ϖ. The CL-BIC is evaluated with $\gamma = \hat{\varpi} - 0.5$. The CL-AIC is evaluated for comparison.

Table 5 summarizes the performance of the two different information criteria based on 100 simulated data sets under each

Table 5. Positive selection rates (PSR) and false discovery rates (FDR) on quadratic exponential model with $P = 1000$

n	w	Rate	CL-AIC	CL-BIC
500	0.2	PSR	0.937	0.933
		FDR	0.777	0.218
	0.3	PSR	0.921	0.915
		FDR	0.742	0.250
	0.4	PSR	0.923	0.914
		FDR	0.728	0.394
1000	0.2	PSR	0.936	0.936
		FDR	0.809	0.016
	0.3	PSR	0.923	0.923
		FDR	0.783	0.032
	0.4	PSR	0.914	0.914
		FDR	0.757	0.139

setting. The number of clusters are set to be 500 or 1000, and the interaction effects are set to be 0.2, 0.3, and 0.4. Across different settings, both methods can achieve PSR about 90 percent. However, the over-fitting effect of CL-AIC is exhibited. The FDR of CL-AIC is about 70 to 80 percent, indicating an inadequate control of the error rate. In contrast, when $n = 1000$, the FDR of CL-BIC is well under 5 percent with $w = 0.2$ or 0.3, and about 13.9 percent with $w = 0.4$, demonstrating a good control of error rate. It is also shown that the performance of CL-BIC greatly improves when the cluster size n increases from 500 to 1000. When the interaction effect increases from 0.2 to 0.4, the CL-BIC will have lower PSR and higher FDR, showing the influence of the interaction effect on the performance of the proposed method. Overall, for quadratic exponential model, the CL-BIC criterion has been shown to perform very well with the formulation of conditional likelihoods.

4.4 Some Practical Guidelines

Here we provide a remark regarding the selection of γ in the penalty term. We implemented two approaches in three simulations above. One varies the magnitude of γ and selects the optimum value that offers the best balance between sensitivity and selectivity as shown in Sections 4.1 and 4.2. The other approach uses $\gamma = \hat{\varpi} - 1/(2\kappa)$, as shown in Section 4.3 under the circumstances that the sample size is fairly sufficient to obtain a good estimate of ϖ.

The estimation of d_s^* is needed in the implementation of the methods, which has been outlined in Section 2. To demonstrate the relationship between d_s and d_s^*, we follow the setup in Section 4.2 and consider a sequence of models under the setting of multivariate probit model with $P = 1000$, and the inter-family correlation equal to 0.3. Similarly, we use the setup in Section 4.3 and examine a sequence of models under the quadratic exponential model with $P = 1000$, and $w = 0.2$. The degrees of freedom d_s of the models being examined range from 1 to 80. The estimated $\hat{d}_s^*$ is obtained from $n = 1000$ data points to ensure reliability of estimation. It is observed from the simulations that $\hat{d}_s^*$ increases in an approximately linear pattern with d_s. For instance, in the setting of multivariate probit model using composite univariate likelihood, when d_s takes the values of $5, 6, 9, 11, 79$, the corresponding $\hat{d}_s^*$ takes the values of $3.94, 4.85, 7.92, 10.02, 77.87$, respectively. In the setting of quadratic exponential model using composite conditional likelihood, when d_s takes the values of $5, 8, 10, 14, 78$, the corresponding $\hat{d}_s^*$ takes the values $3.44, 6.19, 8.08, 10.69, 60.85$. The difference $\hat{d}_s^* - d_s$ is plotted versus d_s in Figure 1. When d_s is below 20, the difference is below 2 for composite univariate likelihood and below 5 for composite conditional likelihood. Thus, as an empirical guideline, if the sample size is not large enough to give a reliable estimate of $\hat{d}_s^*$, d_s may be used as a convenient replacement of d_s^*, when d_s is not too large.

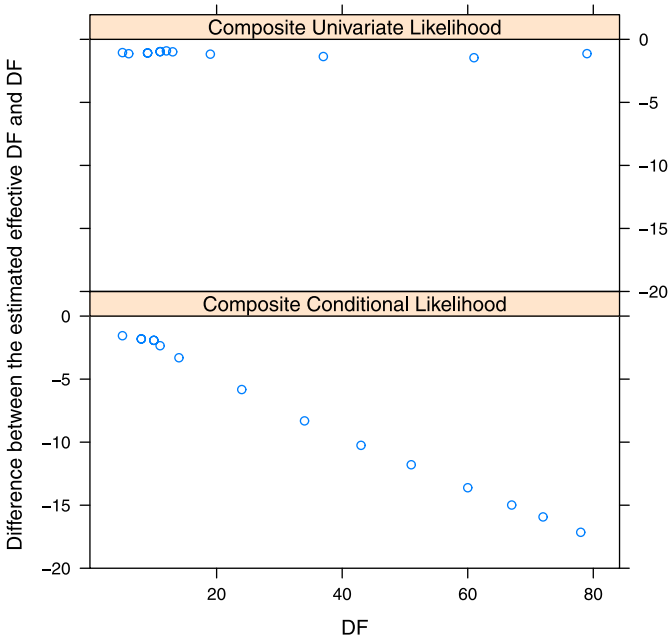

Figure 1. Comparison of estimated effective degrees of freedom d_s^* versus the degrees of freedom d_s. The online version of this figure is in color.

5. REAL DATA ANALYSIS

To examine the empirical performance of the proposed method, we analyze a data from a diabetic nephropathy (DN) study at University of Michigan. In this data set, 35 DN abnormal patients were followed for a period of time ranging from 6.91 to 10.89 years. During the period, their renal functions were measured at multiple time points and the treatment results were classified into binary outcomes as either successful or failure. Each of the patients' renal tissue had undergone a microarray analysis to obtain gene expression data. The purpose of the study is to determine if there are any biomarkers among the 500 candidate genes that have important influence on the risk of exacerbation through a certain therapeutic program. The challenge imposed by this data analysis is the presence of correlation among repeated measurements within each patient. Furthermore, the number of repeated measurements varies across the 35 patients. The total number of measurements is 402, while the total number of candidate covariates is 500. This data set is a practical example containing a large number of covariates and strong dependency among the clustered binary outcomes. We analyze the data using the composite conditional likelihood (Geys, Molenberghs, and Ryan 1997, 1999; Hanfelt 2004).

In order to perform the model selection, we impose penalization on the composite likelihood with L_1 penalty. We gradually increase the tuning parameter in the penalty term and obtain a sequence of models, at which both CL-AIC and CL-BIC are computed. We choose $\gamma = 1 - 1/(2\kappa) = 0.75$, setting $\kappa = 2$, and $\hat{\varpi} = 1$, the lower bound of ϖ. When the tuning parameter increases from 0.04 to 0.26, the number of parameters in the sequence of selected models are $3, 4, 5, 8, 9, 11$, respectively. Under the assumption that only a handful of gene covariates really influence the renal function, we do not further increase the bound to obtain more complicated models. For the sequence of selected models, the CL-BIC takes the values of $557.85, 551.21, 559.21, 578.96, 585.45, 567.32$, whereas the CL-AIC takes the value of $536.09, 518.5732, 515.7058, 502.8211, 498.4278, 458.5472$. As shown, CL-BIC is minimized at an intermediate model with 4 parameters including intercept, interaction and 2 gene covariates. Among all the models being examined, CL-AIC is minimized at the most complicated

model with 11 parameters. The CL-BIC shows its advantage of balancing the model fitting and model complexity when dealing with a large model space.

6. CONCLUDING REMARKS

Model selection is difficult when the number of parameters in the model increases with the sample size. Recently, EBIC (Chen and Chen 2008) has been advocated to address the difficulty through adding an extra penalization term on the dimensionality of the model space. The selection consistency of the EBIC has been established in the linear regression and generalized linear model settings. The proposed CL-BIC may be regarded as an extension of EBIC, but it is applicable to a much broader range of likelihood or quasi-likelihood methods. The model selection consistency of CL-BIC remains true under mild regularity conditions. This is illustrated numerically via three important statistical models. Obviously, a key advantage of the CL-BIC is that it makes the variable selection possible even if the full likelihood is not feasible to compute.

APPENDIX

Additional Regularity Assumptions

Throughout the rest of the appendices, let $\|\cdot\|$ denote the supremum norm. Let $\widetilde{\mathrm{cl}}^{(3)}_{rtu}(\boldsymbol{\theta})$ denote the mixed partial derivative $\partial^3 \mathrm{cl}(\boldsymbol{\theta};\mathbf{Y})/\partial\theta_{[r]}\,\partial\theta_{[t]}\,\partial\theta_{[u]}$, with the subscripts $[r]$, $[t]$, and $[u]$ denoting the elements in the parameter vector $\boldsymbol{\theta}$. Let $R = r_1,\ldots,r_m$ denote a set of coordinate indices with $m \le 3$. Then $\widetilde{\mathrm{cl}}^{(m)}_R(\boldsymbol{\theta})$ denotes $\partial^m \mathrm{cl}(\boldsymbol{\theta};\mathbf{Y})/\partial\theta_{[r_1]}\cdots\partial\theta_{[r_m]}$. For instance, when $R = r_1, r_2$, $\widetilde{\mathrm{cl}}^{(2)}_R(\boldsymbol{\theta})$ denotes $\partial^2 \mathrm{cl}(\boldsymbol{\theta};\mathbf{Y})/\partial\theta_{[r_1]}\theta_{[r2]}$. The following assumptions mainly require the moments (up to the third moments) of the derivatives (up to the third order) of the composite likelihood functions are uniformly bounded in the model space.

Assumption A.1 (A5). (a) Let $R = r_1,\ldots,r_m$ denote a set of coordinate indices with $m \le 3$. It is assumed that for all $s \in \mathcal{S}$ with $d_s \le K$,

$$n^{-1}E_{\boldsymbol{\psi}_{T,0}}\big[\big\{\widetilde{\mathrm{cl}}^{(m)}_R(\boldsymbol{\theta}_{s,0})\big\}^2\big] \le C_2.$$

(b) It is assumed that for all $s \in \mathcal{S}$ with $d_s \le K$,

$$0 < C_3 \le \lambda_{\min}(n^{-1}\mathbf{H}_s(\boldsymbol{\theta}_{s,0})); \qquad 0 < C_4 \le \lambda_{\min}(n^{-1}\mathbf{V}_s(\boldsymbol{\theta}_{s,0})),$$

with $\lambda_{\min}$ denoting the smallest eigenvalue. Furthermore, for all $s \in S_+$, $0 < C_5 \le \lambda_{s[1]}$, where $\lambda_{s[1]}$ denotes the smallest nonzero eigenvalues of the matrix $\mathbf{M}_{s/T}^{1/2}\mathbf{V}_s\mathbf{M}_{s/T}^{1/2}$.

(c) Let $R = r_1,\ldots,r_m$ denote a set of coordinate indices with $m = 2$, or 3. It is assumed that, for any given $\epsilon > 0$, there exists a constant $\eta > 0$, such that

$$\begin{aligned}(1-\epsilon)\big|E_{\boldsymbol{\psi}_{T,0}}\big[n^{-1}\widetilde{\mathrm{cl}}^{(m)}_R(\boldsymbol{\theta}_{s,0})\big]\big| &\le \big|E_{\boldsymbol{\psi}_{T,0}}\big[n^{-1}\widetilde{\mathrm{cl}}^{(m)}_R(\boldsymbol{\theta}^*)\big]\big| \\ &\le (1+\epsilon)\big|E_{\boldsymbol{\psi}_{T,0}}\big[n^{-1}\widetilde{\mathrm{cl}}^{(m)}_R(\boldsymbol{\theta}_{s,0})\big]\big|\end{aligned}$$

for all $s \in \mathcal{S}$ with $d_s \le K$, and $\|\boldsymbol{\theta}^* - \boldsymbol{\theta}_{s,0}\| \le \eta$.

Assumption A.2 (A6). (a) Let $R = r_1,\ldots,r_m$ denote a set of coordinate indices with $m \le 3$. Let $J_s^{(R)}(\mathbf{Y}_i;\boldsymbol{\theta}^*) = (\frac{\partial^m \mathrm{cl}(\boldsymbol{\theta}_s,\mathbf{Y}_i)}{\partial\theta_{[R]}}|_{\boldsymbol{\theta}_s=\boldsymbol{\theta}^*} - E[\frac{\partial^m \mathrm{cl}(\boldsymbol{\theta}_s,\mathbf{Y}_i)}{\partial\theta_{[R]}}|_{\boldsymbol{\theta}_s=\boldsymbol{\theta}^*}])/\operatorname{var}(\frac{\partial^m \mathrm{cl}(\boldsymbol{\theta}_s,\mathbf{Y}_i)}{\partial\theta_{[R]}}|_{\boldsymbol{\theta}_s=\boldsymbol{\theta}^*})^{1/2}$. There exist constants ζ and δ, such that the absolute value of the third derivative of the cumulant generating function $|g^{(3)}(t)|$ of $J_s^{(R)}(\mathbf{Y}_i;\boldsymbol{\theta}^*)$ is bounded by constant C_6, for all the $\|\boldsymbol{\theta}^* - \boldsymbol{\theta}_{s,0}\| \le \zeta$, $0 \le |t| \le \delta$, and all the $s \in \mathcal{S}, d_s \le K$.

(b) Let

$$W_s(\mathbf{Y}_i) = \frac{\lambda_{T|s}(\mathbf{Y}_i;\boldsymbol{\theta}_{T,0},\boldsymbol{\theta}_{s,0}) - \mathrm{E}_{\boldsymbol{\psi}_{T,0}}\{\lambda_{T|s}(\mathbf{Y}_i;\boldsymbol{\theta}_{T,0},\boldsymbol{\theta}_{s,0})\}}{[\operatorname{var}_{\boldsymbol{\psi}_{T,0}}\{\lambda_{T|s-}(\mathbf{Y}_i;\boldsymbol{\theta}_{T,0},\boldsymbol{\theta}_{s-,0})\}]^{1/2}}.$$

There exists constant δ, such that the absolute value of the third derivative of the cumulant generating function $|g^{(3)}(t)|$ of $W_s(\mathbf{Y}_i)$ is bounded by constant C_7, for all $0 \le |t| \le \delta$, and all the $s \in \mathcal{S}, d_s \le K$.

SUPPLEMENTAL MATERIALS

Technical details: The web appendix provides technical proofs for all of our theoretical results. (CLBICJuly2010supp.pdf)

[Received July 2009. Revised July 2010.]

REFERENCES

Akaike, H. (1973), "Information Theory and an Extension of the Maximum Likelihood Principle," in *Proceedins of the Second International Symposium on Information Theory*, eds. B. N. Petrov and F. Caski, Budapest: Akademiai Kiado, pp. 267–281. [1531]

Berger, J. O., Ghosh, J. K., and Mukhopadhyay, N. (2003), "Approximations and Consistency of Bayes Factors as Model Dimension Grows," *Journal of Statistical Planning and Inference*, 112, 241–258. [1531]

Chakrabarti, A., and Ghosh, J. K. (2006), "A Generalization of BIC for the General Exponential Family," *Journal of Statistical Planning and Inference*, 136, 2847–2872. [1531]

Chen, J. H., and Chen, Z. H. (2008), "Extended Bayesian Information Criteria for Model Selection With Large Model Spaces," *Biometrika*, 95, 759–771. [1531-1534,1539]

——— (2009), "Extended BIC for Small-*n*-Large-*P* Sparse GLM," manuscript, available at *http://www.stat.nus.edu.sg/~stachenz/ChenChen.pdf*. [1531]

Cox, D. R., and Hinkley, D. V. (1974), *Theoretical Statistics*, London: Chapman & Hall. [1534]

Cox, D. R., and Reid, N. (2004), "A Note on Pseudolikelihood Constructed From Marginal Densities," *Biometrika*, 91, 729–737. [1531]

Csörgö, M., and Horváth, L. (1997), *Limit Theorems in Change-Point Analysis*, New York: Wiley. [1531]

Fearnhead, P., and Donnelly, P. (2002), "Approximate Likelihood Methods for Estimating Local Recombination Rates," *Journal of the Royal Statistical Society, Ser. B*, 64, 657–680. [1531]

Geys, H., Molenberghs, G., and Ryan, L. M. (1997), "Pseudo-Likelihood Inference for Clustered Binary Data," *Communications in Statistics: Theory and Methods*, 26, 2743–2767. [1536-1538]

——— (1999), "Pseudolikelihood Modeling of Multivariate Outcomes in Developmental Toxicology," *Journal of the American Statistical Association*, 94, 734–745. [1536,1538]

Hanfelt, J. J. (2004), "Composite Conditional Likelihood for Sparse Clustered Data," *Journal of the Royal Statistical Society, Ser. B*, 66, 259–273. [1536, 1538]

Haughton, D. M. A. (1988), "On the Choice of a Model to Fit Data From an Exponential Family," *The Annals of Statistics*, 16, 342–355. [1531]

Heagerty, P. J., and Lele, S. R. (1998), "A Composite Likelihood Approach to Binary Spatial Data," *Journal of the American Statistical Association*, 93, 1099–1111. [1531]

Hjort, N. L., and Omre, H. (1994), "Topics in Spatial Statistics," *Scandinavian Journal of Statistics*, 21, 289–357. [1531]

Jiang, W. (2007), "Bayesian Variable Selection for High Dimensional Generalized Linear Models: Convergence Rates of the Fitted Densities," *The Annals of Statistics*, 35, 1487–1511. [1531]

Konishi, S., Ando, T., and Imoto, S. (2004), "Bayesian Information Criteria and Smoothing Parameter Selection in Radial Basis Function Networks," *Biometrika*, 91, 27–43. [1531]

Li, Y., and Lin, X. (2006), "Semiparametric Normal Transformation Models for Spatially Correlated Survival Data," *Journal of the American Statistical Association*, 101, 591–603. [1531]

Lindsay, B. (1988), "Composite Likelihood Methods," in *Statistical Inference From Stochastic Processes*, ed. N. U. Prabhu, Providence, RI: American Mathematical Society, pp. 221–239. [1531,1532]

Parner, E. T. (2001), "A Composite Likelihood Approach to Multivariate Survival Data," *Scandinavian Journal of Statistics*, 28, 295–302. [1531]

Rao, C. R., and Wu, Y. H. (1989), "A Strongly Consistent Procedure for Model Selection in a Regression Problem," *Biometrika*, 76, 369–374. [1531]

Renard, D., Molenberghs, G., and Geys, H. (2004), "A Pairwise Likelihood Approach to Estimation in Multilevel Probit Models," *Computational Statistics & Data Analysis*, 44, 649–667. [1531]

Schwarz, G. (1978), "Estimating the Dimension of a Model," *The Annals of Statistics*, 6, 461–464. [1531,1532]

Song, P. X.-K. (2007), *Correlated Data Analysis: Modeling, Analytics and Applications*, New York: Springer. [1532]

Tierney, L., and Kadane, J. B. (1986), "Accurate Approximations for Posterior Moments and Marginal Densities," *Journal of the American Statistical Association*, 81, 82–86. [1533]

Tierney, L., Kass, R. E., and Kadane, J. B. (1989), "Fully Exponential Laplace Approximation to Expectations and Variances of Nonpositive Functions," *Journal of the American Statistical Association*, 84, 710–716. [1533]

Varin, C. (2008). "On Composite Marginal Likelihoods," *Advances in Statistical Analysis*, 92, 1–28. [1531,1533]

Varin, C., and Vidoni, P. (2005), "A Note on Composite Likelihood Inference and Model Selection," *Biometrika*, 92, 519–528. [1531,1533,1534]

Wang, H., Li, R., and Tsai, C.-L. (2007), "Tuning Parameter Selector for the Smoothly Clipped Absolute Deviation Method," *Biometrika*, 94, 553–568. [1531]

Wang, L., and Qu, A. (2009), "Consistent Model Selection and Data-Driven Smooth Tests for Longitudinal Data in the Estimating Equations Approach," *Journal of the Royal Statistical Society, Ser. B*, 71, 177–190. [1531]

White, H. (1982), "Maximum Likelihood Estimation of Misspecified Models," *Econometrica*, 50, 1–25. [1534]

Yao, Y. C. (1988), "Estimating the Number of Change-Points via Schwarz Criterion," *Statistics & Probability Letters*, 6, 181–189. [1531]

Zhao, L. P., and Prentice, R. (1990), "Correlated Binary Regression Using a Quadratic Exponential Mode," *Biometrika*, 77, 642–648. [1537]

Zhao, Y., and Joe, H. (2005), "Composite Likelihood Estimation in Multivariate Data Analysis," *Canadian Journal of Statistics*, 33, 335–356. [1533,1535]