The Canadian Journal of Statistics
Vol. 42, No. 4, 2014, Pages 544–562
La revue canadienne de statistique

Reducing the sensitivity to nuisance parameters in pseudo-likelihood functions

Yang NING[1]*, Kung-Yee LIANG[2] and Nancy REID[3]

[1]*Department of Statistics and Actuarial Science, University of Waterloo, Waterloo, Canada*
[2]*Department of Life Sciences and Institute of Genome Sciences, National Yang-Ming University, Taiwan, ROC*
[3]*Department of Statistical Science, University of Toronto, Toronto, Canada*

Key words and phrases: Composite likelihood; higher order inference; invariance; nuisance parameters; pseudo-likelihood.

MSC 2010: Primary 62F03; secondary 62F10

Abstract: In a parametric model, parameters are often partitioned into parameters of interest and nuisance parameters. However, as the data structure becomes more complex, inference based on the full likelihood may be computationally intractable or sensitive to potential model misspecification. Alternative likelihood-based methods proposed in these settings include pseudo-likelihood and composite likelihood. We propose a simple adjustment to these likelihood functions to reduce the impact of nuisance parameters. The advantages of the modification are illustrated through examples and reinforced through simulations. The adjustment is still novel even if attention is restricted to the profile likelihood. *The Canadian Journal of Statistics* 42: 544–562; 2014 © 2014 Statistical Society of Canada

Résumé: Les paramètres d'un modèle sont souvent catégorisés comme nuisibles ou d'intérêt. À mesure que la structure des données devient plus complexe, la vraisemblance peut devenir incalculable ou sensible à des erreurs de spécification. La pseudo-vraisemblance et la vraisemblance composite ont été présentées comme des solutions dans ces situations. Les auteurs proposent un ajustement simple de ces fonctions de vraisemblance afin d'atténuer l'effet des paramètres nuisibles. Les avantages offerts par cette modification sont illustrés par des exemples et appuyés par des simulations. Cet ajustement est inédit même si les auteurs restreignent leur attention aux profils de vraisemblance. *La revue canadienne de statistique* 42: 544–562; 2014 © 2014 Société statistique du Canada

1. INTRODUCTION

Likelihood functions play a key role in statistical inference. However, in many complex models, the likelihood function may be difficult to evaluate, or even to specify. To address these concerns, many alternative likelihood methods, including pseudo-likelihoods (Gong & Samaniego, 1981; Severini, 1998b) and composite likelihoods (Besag, 1974; Lindsay, 1988; Cox & Reid, 2004) have been proposed. One advantage of an alternative likelihood function over an estimating function is that likelihood ratio type statistics can be used to set confidence regions and test statistical hypothesis about parameters (Hanfelt & Liang, 1995).

Suppose that we observe n independent random variables Y_i, $i = 1, \ldots, n$, with density function $f(y; \psi)$, where $\psi = (\psi^1, \ldots, \psi^{p+q}) = (\theta, \phi)$, θ is a p-dimensional parameter of interest and ϕ is a q-dimensional nuisance parameter. The log-likelihood for ψ is given by $L(\psi) = \sum_{i=1}^n \log f(y_i; \psi)$, and $\tilde{L}(\theta) = L(\theta, \tilde{\phi}(\theta))$ is the profile log-likelihood for θ, where y_i is the observed value of Y_i and $\tilde{\phi}(\theta) = \text{argmax}_{\phi} L(\theta, \phi)$. The simplest approach for estimation

* *Author to whom correspondence may be addressed.*
E-mail: y4ning@uwaterloo.ca

and inference about θ is to treat the profile likelihood as an ordinary likelihood. However, this ignores the error associated with estimating the nuisance parameters. Various adjustments to the profile likelihood have been proposed in order to reduce the sensitivity to nuisance parameters; see Barndorff-Nielsen (1983), Barndorff-Nielsen & Cox (1994), Cox & Reid (1987), Fraser (2003), McCullagh & Tibshirani (1990), Pace & Salvan (2006), Severini (1998a) and Stern (1997). As shown by McCullagh & Tibshirani (1990) and Stern (1997), the profile score function for θ is biased with expected value $O(1)$ and a suitable adjustment typically reduces the bias of the associated score function to $O(n^{-1})$. Recently, these types of adjustments have been extended to estimating equations; see Adimari & Ventura (2002), Bellio, Greco, & Ventura (2008), Jorgensen & Knudsen (2004), Rathouz & Liang (1999), Severini (2002) and Wang & Hanfelt (2003).

While adjustments for the profile likelihood are well studied, little has been done with respect to reducing the impact of nuisance parameters in alternative likelihood functions, such as the pseudo-likelihood function and the composite likelihood function. The purpose of the current article is to fill this gap. We propose a simple non-additive adjustment, with the goal of reducing the bias of the score function to $O(n^{-1})$. While the adjustment can be analytically derived in some specific examples, we propose a first order approximation method to calculate the adjustment for general models. When applied to the profile likelihood, the adjustment is still novel, while its asymptotic expansion agrees with the adjustment suggested by Stern (1997) to first order.

Chandler & Bate (2007) and Pace, Salvan, & Sartori (2011) considered adjustments to the composite likelihood. Since the asymptotic null distribution of the composite likelihood ratio test is not chi-squared (Varin, Reid, & Firth, 2011), the aim of their procedures is to calibrate the limiting distribution so that it can be compared to the chi-square distribution. When applied to composite likelihood, the proposed method reduces the impact of nuisance parameters and is thus quite different from adjustments to composite likelihood proposed by Chandler & Bate (2007) and Pace, Salvan, & Sartori (2011). These latter involve completely different procedures.

We introduce the pseudo-likelihood and some examples in Section 2. The adjusted pseudo-likelihood is described in Section 3, and we illustrate the calculation of the adjustment in Section 4. The theoretical properties of the adjusted pseudo-likelihood are established in Section 5. In Section 6, we revisit the examples and conduct simulation studies to evaluate the finite sample performance of the adjustment. Our method is applied to adjust the profile composite likelihood and profile likelihood in Sections 7 and 8, respectively. Section 9 contains a general discussion.

2. BACKGROUND ON PSEUDO-LIKELIHOOD

The profile likelihood is obtained by maximizing out the nuisance parameter ϕ for fixed θ. As an alternative, Gong & Samaniego (1981) proposed a pseudo-likelihood approach, in which ϕ is estimated by a single convenient estimator, say $\tilde{\phi}$. For instance, $\tilde{\phi}$ can be obtained by solving an unbiased estimating equation. The log pseudo-likelihood $L(\theta, \tilde{\phi})$, can be used for estimation and inference about θ. Under some regularity conditions, Gong & Samaniego (1981) showed that the maximizer of $L(\theta, \tilde{\phi})$ remains consistent and asymptotically normal, provided $\tilde{\phi}$ is consistent and asymptotically normal. The limiting distribution of the associated log-likelihood ratio test is derived in Liang & Self (1996); see also Chen & Liang (2010), which considers testing boundary parameters.

Example 1: Variance component model

Consider the model

$$Y = X\beta + \epsilon, \epsilon \sim N\{0, \Sigma(\theta)\},$$

where Y is a $n \times 1$ vector of response variables, X is a $n \times q$ design matrix, $\Sigma(\theta)$ is a $n \times n$ matrix indexed by an unknown parameter of interest θ with dimension p, and β is the nuisance parameter. In pedigree analysis, it has been a standard procedure to estimate β by the conventional least square estimator $\tilde{\beta} = (X^{\mathrm{T}}X)^{-1}X^{\mathrm{T}}Y$, and conduct inference about θ based on $L(\theta, \tilde{\beta})$, where $L(\theta, \beta)$ is the log likelihood based on data $Y = y$. This pseudo-likelihood approach is widely used in the field of genetic epidemiology; see Abney, McPeek, & Ober (2000), Khoury, Beaty, & Cohen (1993) and Santamaria et al. (2007).

Example 2: A general setting for pseudo-likelihood

Chen & Liang (2010) introduced a general setting, in which inference based on the pseudo-likelihood is a natural choice. Suppose that the log-likelihood can be decomposed into two parts $L(\theta, \phi) = L_1(\theta, \phi) + L_2(\phi)$, where both $L_1(\theta, \phi)$ and $L_2(\phi)$ are log-likelihood functions for some observed random variables. We could estimate ϕ by $\tilde{\phi} = \text{argmax}_\phi L_2(\phi)$ and then plug it in $L_1(\theta, \phi)$ to construct the log pseudo-likelihood $L_1(\theta, \tilde{\phi})$. An example in the field of statistical genetics is studied in Section 6 (Liang, Rathouz, & Beaty, 1996).

3. ADJUSTED PSEUDO-LIKELIHOOD

Let $\tilde{L}(\theta) = L(\theta, \tilde{\phi})$ denote the log pseudo-likelihood for θ. Assume that indices $a, b, c, \ldots$ range over $1, \ldots, p$, indices $e, f, g, e', f', g' \ldots$ range over $p+1, \ldots, p+q$ and indices $r, s, t \ldots$ range over $1, \ldots, p+q$. Derivatives are indicated by subscripts; for instance, $L_r(\psi) = \partial L(\psi)/\partial\psi^r$ and $L_{rs}(\psi) = \partial^2 L(\psi)/\partial\psi^r\partial\psi^s$. Let the true value of the parameter be $\psi_0 = (\theta_0, \phi_0)$. The expectation evaluated at the true model is denoted by $E_0(\cdot) = E(\cdot; \psi_0)$. Denote $\lambda_{rs} = E_0\{L_{rs}(\psi_0)\}$, $\lambda_{rst} = E_0\{L_{rst}(\psi_0)\}$ and $\lambda_{r,s} = E_0\{L_r(\psi_0)L_s(\psi_0)\}$. Note that λ_{rs}, λ_{rst} and $\lambda_{r,s}$, etc. are of order $O(n)$ under standard regularity conditions (Severini, 2000). Let λ^{ef} be the $q \times q$ matrix inverse of λ_{ef}. Einstein summation is adopted, for instance, $x_a y^a$ denotes $\sum x_a y^a$, where the summation is over all possible values of a.

The idea of the adjustment is to consider a transformation of θ, such that after adjustment the pseudo-score function for θ has mean 0. As shown in Appendix, $E_0\{\tilde{L}_a(\theta_0)\} = O(1)$, implying that the true parameter θ_0 is not the solution of $E_0\{\tilde{L}_a(\theta)\} = 0$. We define the target parameter $\theta_n^*(\psi_0)$ as the solution to

$$E_0\{\tilde{L}_a(\theta)|_{\theta=\theta_n^*(\psi_0)}\} = 0. \tag{1}$$

The pseudo-score function becomes exactly unbiased if we evaluate the parameter at $\theta_n^*(\psi_0)$ instead of θ_0. Let $\tilde{\theta}$ be the solution of $\partial\tilde{L}(\theta)/\partial\theta = 0$. The target parameter $\theta_n^*(\psi_0)$ can be interpreted as the target of the estimator $\tilde{\theta}$, in the sense that $\theta_n^*(\psi_0)$ satisfies (1) and $\tilde{\theta}$ satisfies the sample version of (1), where the expectation is replaced with the sample average. Note that $\theta_n^*(\psi_0)$ may depend on the nuisance parameter ϕ_0, which can be estimated by $\tilde{\phi}$. The explicit form of $\theta_n^*(\psi_0)$ is often unknown, but it can be estimated using the first order approximation method described in Section 4. Hereafter, for notational simplicity, we drop the dependence of $\theta_n^*(\psi_0)$ on n, and write $\theta^*(\psi_0)$ or θ^*. The transformation θ^* is also known as the bridge function in the model misspecification literature introduced by Jiang & Turnbull (2004) and Yi & Reid (2010) in models without nuisance parameters. A reviewer has asked if θ^* is always well-defined, and in general the answer to this question seems difficult. We think it should be studied case by case; for instance, in Example 1, we have $\theta^* = \theta\text{tr}(Q\Sigma^{-1}Q\Sigma)/n$. Hence, θ^* exists and is unique.

To better understand the impact of nuisance parameters on estimation, we consider the following decomposition of $\tilde{\theta} - \theta_0$,

$$\tilde{\theta} - \theta_0 = (\tilde{\theta} - \theta^*) + (\theta^* - \theta_0).$$

The first term $\tilde{\theta} - \theta^*$ represents the random fluctuation of $\tilde{\theta}$ around the target parameter θ^* due to the random noise from the data. The second term $\theta^* - \theta_0$ characterizes the bias of estimation in finite samples. The presence of $\theta^* - \theta_0$ is due to the estimation of nuisance parameters. To see this, when the model contains no nuisance parameters, the pseudo-score functions $\tilde{L}_a(\theta)$ in (1) are replaced by the score functions $L_a(\theta)$, and therefore, by definition, we obtain $\theta^* = \theta_0$. While we will show that $\theta^* - \theta_0 = o(n^{-1/2})$ in Section 4, it can be large for small n.

To reduce the impact of nuisance parameters, we would like θ^* to be as close as possible to θ_0, after adjusting the location of the likelihood surface. More precisely, our adjustment is given by replacing θ in $\tilde{L}(\theta)$ with θ^*, and we define the adjusted log pseudo-likelihood for θ as

$$\tilde{L}^*(\theta) = \tilde{L}\left[\theta^*\{\tilde{\psi}(\theta)\}\right],$$

where $\tilde{\psi}(\theta) = (\theta, \tilde{\phi})$. The adjusted pseudo-likelihood is constructed by embedding the bridge function into the log pseudo-likelihood. This differs from the additive form of other adjustments to the profile log-likelihood (McCullagh & Tibshirani, 1990; Chandler & Bate, 2007).

4. CALCULATION OF THE ADJUSTED PSEUDO-LIKELIHOOD

In this section, we derive a simple formula for the first order approximation to θ^*. Starting from Equation (1), a Taylor expansion about θ_0 gives

$$0 = E_0\{\tilde{L}_a(\theta^*)\} = E_0\{\tilde{L}_a(\theta_0)\} + (\theta^{*b} - \theta_0^b)E_0\{\tilde{L}_{ab}(\theta_0)\} + O(n||\theta^* - \theta_0||^2), \tag{2}$$

where $||\cdot||$ is the Euclidean norm. Under mild regularity conditions (Severini, 2000), $E_0\{\tilde{L}_{ab}(\theta_0)\} = O(n)$, and by Appendix, $E_0\{\tilde{L}_a(\theta_0)\} = O(1)$. Therefore, from Equation (2), we have $\theta^* - \theta_0 = O(n^{-1})$. Combining this with the expansion in Appendix, we obtain

$$\theta^{*a} - \theta_0^a = \rho^a(\psi_0) + O(n^{-2}) = \alpha_b\beta^{ab} + O(n^{-2}). \tag{3}$$

Here, $\rho^a(\psi_0) = \alpha_b\beta^{ab} = O(n^{-1})$, where α_b is an approximation to $E_0\{\tilde{L}_b(\theta_0)\}$, and β^{ab} is a $p \times p$ matrix inverse of β_{ab}, where β_{ab} is an approximation to the pseudo-information matrix $E_0\{-\tilde{L}_{ab}(\theta_0)\}$. In Appendix, we derive explicit expressions for α_b and β_{ab} under two scenarios. The first scenario is that the explicit form of the estimator $\tilde{\phi}$ is given directly, and the second scenario is that $\tilde{\phi}$ is defined implicitly as the solution of a set of unbiased estimation functions. These two situations can cover a broad range of estimation methods for nuisance parameters used in pseudo-likelihood inference.

From Equation (3), ρ characterizes the leading term in the amount of shift from the true parameter to the new target parameter. Therefore, a first order approximation to the adjusted log pseudo-likelihood is given by

$$\tilde{L}^{**}(\theta) = \tilde{L}\{\theta + \tilde{\rho}(\theta)\},$$

where $\tilde{\rho}(\theta) = \rho(\theta, \tilde{\phi})$. Note that $\tilde{\rho}(\theta)$ is obtained by plugging $\tilde{\phi}$ into $\rho(\theta, \phi)$. To make our notation system more coherent, we use $\tilde{K}(\theta)$ to denote the quantity $K(\theta, \phi)$ if ϕ is estimated by $\tilde{\phi}$. Under the assumption that $\partial\rho(\psi_0)/\partial\psi = O(n^{-1})$, the resulting estimation error of replacing ϕ_0 in $\rho(\psi_0) = \rho(\theta_0, \phi_0)$ with $\tilde{\phi}$ is $o(n^{-1})$, that is, $\tilde{\rho}(\theta_0) - \rho(\psi_0) = o(n^{-1})$.

5. PROPERTIES OF THE ADJUSTED PSEUDO-LIKELIHOOD

In this section, we present the asymptotic properties associated with the first order approximation to the adjusted log pseudo-likelihood $\tilde{L}^{**}(\theta)$, and indicate that the same results hold for the exact version $\tilde{L}^*(\theta)$ as well.

5.1. First Order Properties of Adjusted Pseudo-Likelihood

Let $\tilde{\theta}^{**}$ denote the root of $\tilde{L}_a^{**}(\theta) = 0$, for $a = 1, \ldots, p$. The Taylor expansion of $\tilde{L}_a^{**}(\tilde{\theta}^{**}) = 0$ about θ_0 yields

$$\tilde{\theta}^{**a} - \theta_0^a = -\tilde{L}^{**ab}(\theta_0)\tilde{L}_b^{**}(\theta_0) + o_p(n^{-1/2}),$$

where $\tilde{L}^{**ab}(\theta_0)$ is the $p \times p$ matrix inverse of $\tilde{L}_{ab}^{**}(\theta_0)$. Since by assumption $\partial\rho(\psi_0)/\partial\psi = O(n^{-1})$ and by definition $\tilde{L}^{**}(\theta) = \tilde{L}\{\theta + \tilde{\rho}(\theta)\}$, we have $\tilde{L}^{**ab}(\theta_0) - \tilde{L}^{ab}(\theta_0) = o_p(n^{-1})$, where $\tilde{L}^{ab}(\theta_0)$ is the matrix inverse of $\tilde{L}_{ab}(\theta_0)$. In addition,

$$\tilde{L}_b^{**}(\theta_0) = \tilde{L}_b(\theta_0) + \tilde{\rho}^c(\theta_0)\tilde{L}_{bc}(\theta_0) + o_p(1),$$

where $\tilde{\rho}^c(\theta_0)\tilde{L}_{bc}(\theta_0) = O_p(1)$. These together imply that

$$\tilde{\theta}^{**a} - \theta_0^a = -\tilde{L}^{ab}(\theta_0)\tilde{L}_b(\theta_0) + o_p(n^{-1/2}). \tag{4}$$

Recall that $\tilde{\theta}$ is the solution of $\partial\tilde{L}(\theta)/\partial\theta = 0$. By the asymptotic expansion (4), we find that $\tilde{\theta}^{**}$ is asymptotically equivalent to $\tilde{\theta}$, that is, $W_e^{**}(\theta_0) - W_e(\theta_0) = o_p(1)$, where $W_e^{**}(\theta_0) = n^{1/2}(\tilde{\theta}^{**} - \theta_0)$ and $W_e(\theta_0) = n^{1/2}(\tilde{\theta} - \theta_0)$. Define the pseudo-likelihood ratio test (PLRT) as $W(\theta_0) = 2\{\tilde{L}(\tilde{\theta}) - \tilde{L}(\theta_0)\}$ and the adjusted pseudo-likelihood ratio test (APLRT) as $W^{**}(\theta_0) = 2\{\tilde{L}^{**}(\tilde{\theta}^{**}) - \tilde{L}^{**}(\theta_0)\}$. The following theorem establishes the asymptotic distribution of APLRT.

Theorem 1. *Assume that $\tilde{\phi}$ has the following asymptotic distribution $n^{1/2}(\tilde{\phi} - \phi_0) \rightarrow N(0, \Sigma_{22})$. Then under the same regularity conditions (A1)–(A6) in Gong & Samaniego (1981),*

$$n^{1/2}(\tilde{\theta}^{**} - \theta_0) \rightarrow_d N\{0, I_{11}^{-1}(I_{11} + I_{12}\Sigma_{22}I_{12}^{\mathrm{T}})I_{11}^{-1}\},$$

where

$$I_{11} = \lim_{n\rightarrow\infty} E_0\left\{-\frac{1}{n}\frac{\partial^2 L(\theta_0, \phi_0)}{\partial\theta\partial\theta^{\mathrm{T}}}\right\}, \quad I_{12} = \lim_{n\rightarrow\infty} E_0\left\{-\frac{1}{n}\frac{\partial^2 L(\theta_0, \phi_0)}{\partial\theta\partial\phi^{\mathrm{T}}}\right\},$$

and $W^{**}(\theta_0) \rightarrow_d \sum_{j=1}^p \delta_j U_j$, where U_j's are independent χ_1^2 variables and δ_j's are the eigenvalues of $I_{11}^{-1}(I_{11} + I_{12}\Sigma_{22}I_{12}^{\mathrm{T}})$.

The proof of Theorem 1 follows by Theorem 2.2 in Gong & Samaniego (1981) and our Equation (4) and hence is omitted. When the score function for θ is orthogonal to the score function for ϕ, that is, $I_{12} = 0$, the asymptotic distribution of $W^{**}(\theta_0)$ reduces to χ_p^2. While $W^{**}(\theta_0)$ and $W(\theta_0)$ have the same limiting distribution with the same eigenvalue, $W^{**}(\theta_0)$ has better finite sample performance, because our adjustment alleviates the impact of nuisance parameters.

5.2. Higher Order Properties of the Adjusted Pseudo-Likelihood

To demonstrate the theoretical advantage of our adjustment, we consider higher order properties of the pseudo-score function. To show the following two theorems, we require that $\partial\rho(\theta_0, \phi_0)/\partial\phi$ is $O(n^{-1})$, which is true if $\partial\lambda_{rs}/\partial\psi$ and $\partial\lambda_{r,s}/\partial\psi$, etc. are $O(n)$. The proofs of the theorems are shown in the Supplementary Materials. As shown in Appendix, without any adjustment, we have

$E_0\{\partial\tilde{L}(\theta_0)/\partial\theta\} = O(1)$. The following theorem shows that the bias of the adjusted pseudo-score function is reduced by an order of magnitude.

Theorem 2. *The bias of the adjusted pseudo-score function is $O(n^{-1})$, that is,*

$$E_0\{\partial\tilde{L}^{**}(\theta_0)/\partial\theta\} = O(n^{-1}).$$

The pseudo-score function falls into the catalogue of estimating functions. In the estimating function framework, a well-known criterion for measuring the amount of information loss due to the unknown parameters is the mean square error criterion (MSE) (Liang, 1987; Wang & Hanfelt, 2003). The MSE of the adjusted pseudo-score function in our context can be defined as

$$\mathrm{MSE}(\partial\tilde{L}^{**}(\theta)/\partial\theta) = E_0||\partial\tilde{L}^{**}(\theta)/\partial\theta - \partial L(\theta,\phi)/\partial\theta||^2, \tag{5}$$

where $||\cdot||$ is the Euclidean norm. Note that the score function $\partial L(\theta,\phi_0)/\partial\theta$ with $\phi = \phi_0$ is the optimal estimating equation for θ, in the sense that it maximizes the Godambe information (Godambe, 1960). In general, the MSE is a measure of the closeness of the estimating equation to its counterpart assuming nuisance parameters are known. Given the MSE criterion, the following theorem establishes the superiority of the adjusted pseudo-score function under the parameter orthogonality condition, that is, $E_0\{\partial^2 L(\psi_0)/\partial\theta\partial\phi\} = 0$.

Theorem 3. *Ignoring the terms of order $o(1)$, we have*

$$\mathrm{MSE}\{\partial\tilde{L}^{**}(\theta_0)/\partial\theta\} \le \mathrm{MSE}\{\partial\tilde{L}(\theta_0)/\partial\theta\},$$

if θ is orthogonal to ϕ.

In the proof of Theorem 3, we have shown that $\mathrm{MSE}\{\partial\tilde{L}^{**}(\theta_0)/\partial\theta\}$ and $\mathrm{MSE}\{\partial\tilde{L}(\theta_0)/\partial\theta\}$ are $O(1)$. Thus, Theorem 3 demonstrates that the adjusted pseudo-score function is asymptotically closer to the oracle score function $\partial L(\theta,\phi_0)/\partial\theta$ under the MSE criterion. When θ is a scalar, the orthogonality of θ and ϕ can be always achieved, after a reparametrization of ϕ (Cox & Reid, 1987). Similar properties are established by Liang (1987) and by Wang & Hanfelt (2003) under orthogonality conditions, for estimating equations derived from adjusted profile log-likelihoods.

6. EXAMPLES

6.1. Example 1: Variance Component Models

The log pseudo-likelihood is given by

$$\tilde{L}(\theta) = L(\theta,\tilde{\beta}) = -\frac{1}{2}\log\left[\det\{\Sigma(\theta)\}\right] - \frac{1}{2}Y^{\mathrm{T}}Q^{\mathrm{T}}\Sigma^{-1}(\theta)QY, \tag{6}$$

where det represents the determinant of a matrix and $Q = I - X(X^{\mathrm{T}}X)^{-1}X^{\mathrm{T}}$. The derivative of the log pseudo-likelihood with respect to θ is

$$\frac{\partial L(\theta,\tilde{\beta})}{\partial\theta^a} = -\frac{1}{2}\mathrm{tr}\left\{\Sigma^{-1}(\theta)\frac{\mathrm{d}\Sigma(\theta)}{\mathrm{d}\theta^a}\right\} + \frac{1}{2}Y^{\mathrm{T}}Q^{\mathrm{T}}\Sigma^{-1}(\theta)\frac{\mathrm{d}\Sigma(\theta)}{\mathrm{d}\theta^a}\Sigma^{-1}(\theta)QY,$$

where tr is the trace of a matrix. By Theorem 1, the limiting distribution of the corresponding adjusted pseudo-likelihood ratio test is χ^2_p, due to the orthogonality of β and θ. By Theorem 3, under the orthogonality condition, the adjusted pseudo-score function has smaller mean square

TABLE 1: Empirical rejection rate of Wald, score, pseudo-likelihood ratio test (PLRT), adjusted pseudo-likelihood ratio test (APLRT) and adjusted score test (AS) for $H_0 : \theta = 1$ in Example 1 (nominal type I error 0.05), where $\Sigma(\theta) = \theta\Sigma$, Σ is an exchangeable matrix with correlation parameter ρ.

	$\rho = 0.3$					$\rho = 0.7$				
n	Wald test	Score test	PLRT	APLRT	AS	Wald test	Score test	PLRT	APLRT	AS
5	0.471	0.011	0.221	0.140	0.081	0.462	0.012	0.220	0.153	0.077
7	0.345	0.013	0.146	0.092	0.068	0.332	0.016	0.142	0.090	0.067
10	0.275	0.017	0.110	0.084	0.063	0.281	0.018	0.116	0.080	0.061
20	0.172	0.034	0.073	0.060	0.052	0.176	0.032	0.076	0.057	0.051
30	0.127	0.041	0.071	0.059	0.053	0.129	0.039	0.070	0.059	0.052
40	0.112	0.047	0.069	0.060	0.056	0.113	0.044	0.066	0.058	0.054
50	0.098	0.044	0.063	0.055	0.055	0.096	0.047	0.060	0.053	0.054

error than the pseudo-score function. Thus, similar to the standard score test, we construct an adjusted score test (AS) based on the adjusted pseudo-score function.

In our first simulation scenario, we assume $X_{1i} = 1$, $X_{2i} \sim N(0, 1)$, $\beta = (1, 1)$, $\Sigma(\theta) = \theta\Sigma$, where Σ is an exchangeable matrix with known correlation ρ, and $\theta = 1$. The parameter of interest is θ. The number of simulations is 10,000.

In the first scenario, the bridge can be shown to be $\theta^*(\theta) = \theta \text{tr}(Q\Sigma^{-1}Q\Sigma)/n$, which is independent of nuisance parameters. The adjusted pseudo-likelihood is $\tilde{L}^*(\theta) = \tilde{L}(\theta^*(\theta), \tilde{\beta})$ and the adjusted pseudo-score function is $\partial\tilde{L}^*(\theta)/\partial\theta$. Table 1 summarizes the empirical rejection rate of various tests. It has been well documented in the literature that the Wald-type inference can be ill behaved, and this concern is alleviated by using the likelihood ratio type statistic (Hauck, Walter, & Donner, 1977). The results in Table 1 confirm this conclusion. We find that the Wald test tends to underestimate the nominal level, even if the sample size is as large as 50. On the other hand, the rejection rate for the score test is much smaller than the nominal level for small sample size. It is seen that our proposed APLRT performs consistently better than PLRT. Moreover, the adjusted score test (AS) performs best among these five procedures under all scenarios considered, and it produces satisfactory results even when $n = 5$.

In this example, it is instructive to compare the confidence intervals. The interval $[\theta_1, \theta_2]$ is a 95% confidence interval based on the PLRT, if any $\theta \in [\theta_1, \theta_2]$ satisfies $2\{\tilde{L}(\tilde{\theta}) - \tilde{L}(\theta)\} \leq q_{0.95}$, where $\tilde{\theta}$ is the maximizer of $\tilde{L}(\theta)$ and $q_{0.95}$ is the 95% quantile of the chi-square distribution with 1 degree of freedom. Interestingly, we observe that if $[\theta_1, \theta_2]$ denotes the 95% confidence interval based on the PLRT, then $[\theta_1\delta, \theta_2\delta]$ is the 95% confidence interval based on the APLRT, where $\delta = n/\text{tr}(Q\Sigma^{-1}Q\Sigma)$. Thus, calculating the 95% confidence interval based on the APLRT involves no additional steps, and δ can be interpreted as an inflation factor which represents the ratio of the lengths of the two confidence intervals. As shown in Figure 1, the confidence interval based on the APLRT is wider than that based on the PLRT, especially when n is small. From a hypothesis testing perspective, the PLRT produces a confidence interval that is too narrow, and tends to reject the null hypothesis more frequently than the nominal type I error. On the other hand, the APLRT has slightly larger confidence intervals and more accurate empirical rejection rate than PLRT; see Table 1. As n increases, the ratio of the lengths of two confidence intervals approaches 1.

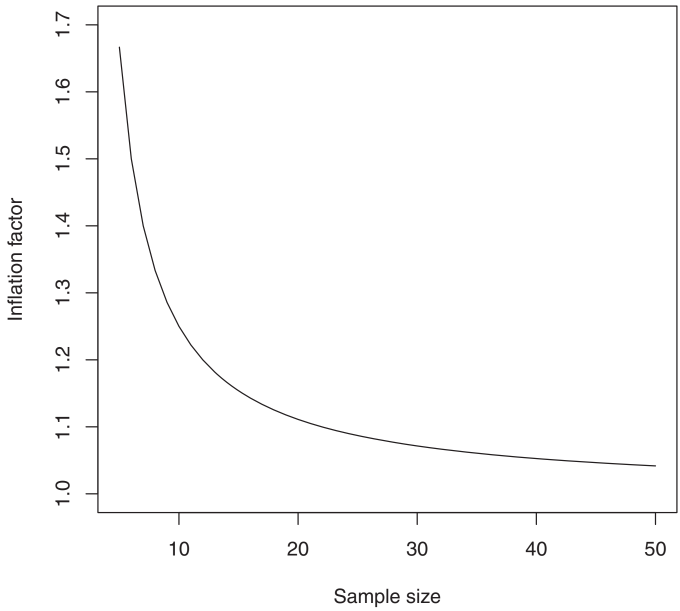

FIGURE 1: Plot of the inflation factor δ (ratio of the lengths of confidence interval based on APLRT to that based on PLRT).

To illustrate the usage of the proposed method for multidimensional parameter of interest, we consider a second simulation scenario. As above we suppose that $\Sigma(\theta, \rho) = \theta\Sigma(\rho)$, where $\Sigma(\rho)$ is an exchangeable matrix with unknown correlation parameter ρ, but we assume that a two dimensional parameter (θ, ρ) is the parameter of interest. The same data generating process as in the first scenario is adopted. After some algebra, we can show that the target parameter (θ^*, ρ^*) is the solution of the following equations

$$\theta^* = \theta \text{tr}\{Q\Sigma^{-1}(\rho)Q\Sigma(\rho^*)\} \text{ and } \text{tr}\{\Sigma^{-1}(\rho)A\} = \text{tr}\{Q\Sigma^{-1}(\rho)A\Sigma^{-1}(\rho)Q\Sigma(\rho^*)\}\theta^*/\theta, \quad (7)$$

where A is a $n \times n$ matrix with diagonal elements 0 and off-diagonal elements 1. Given values of (θ, ρ), (7) can be solved numerically. The adjusted pseudo-likelihood function is $\tilde{L}(\theta, \rho) = L(\theta^*(\theta, \rho), \rho^*(\theta, \rho), \tilde{\beta})$, where $L(\theta, \rho, \tilde{\beta})$ is given by (6), and $\Sigma(\theta)$ is replaced with $\Sigma(\theta, \rho)$. The comparison of the empirical rejection rate based on various tests is shown in Table 2. We find that the APLRT and the AS have similar performance and both outperform the Wald, score and pseudo-likelihood ratio tests.

6.2. Example 2: A General Setting for Pseudo-Likelihood with Applications to Genetic Linkage Analysis

To illustrate the pseudo-likelihood method in the general setting, we consider the following example from the linkage analysis, where the data consist of genetic marker data M and trait data T. The likelihood function is proportional to $f(M|T; \theta, \phi)f(T; \phi)$, where ϕ represents an unknown penetrance parameter for the trait locus and θ represents the recombination fraction between a trait locus and a marker. Liang, Rathouz, & Beaty (1996) demonstrated that the pseudo-likelihood estimator is more efficient than the conditional likelihood estimator based on the model $f(M|T; \theta, \phi)$ and almost as efficient as the maximum likelihood estimator by theoretical analysis and simulations. More importantly, the pseudo-likelihood approach is more

TABLE 2: Empirical rejection rate of Wald, score, pseudo-likelihood ratio test (PLRT), adjusted pseudo-likelihood ratio test (APLRT) and adjusted score test (AS) for $H_0 : \theta = 1, \rho = 0.3$ and $H_0 : \theta = 1, \rho = 0.7$ in Example 1 (nominal type I error 0.05), where $\Sigma(\theta) = \theta\Sigma$, Σ is an exchangeable matrix with correlation parameter ρ.

	$\rho = 0.3$					$\rho = 0.7$				
n	Wald test	Score test	PLRT	APLRT	AS	Wald test	Score test	PLRT	APLRT	AS
10	0.314	0.006	0.145	0.084	0.083	0.325	0.010	0.147	0.087	0.079
20	0.232	0.028	0.083	0.069	0.065	0.220	0.030	0.078	0.064	0.066
30	0.146	0.040	0.075	0.061	0.063	0.138	0.034	0.076	0.063	0.059
40	0.103	0.045	0.064	0.053	0.056	0.110	0.042	0.070	0.055	0.058
50	0.083	0.042	0.060	0.054	0.056	0.087	0.045	0.065	0.054	0.054

flexible in the sense that one is allowed to estimate ϕ externally through segregation analysis. Thus, the pseudo-likelihood approach is often preferred in linkage analysis.

For illustration, we consider a simple situation in which the trait is governed by a single autosomal dominant locus with two alleles, say D and d, and the markers have two codominant alleles, say C and B. The joint probabilities and observed frequencies of (T, M) are given in Table 3. It is easily seen that the full log-likelihood $L(\theta, \phi)$ and the marginal log-likelihood $L_2(\phi) = \log f(T; \phi)$ are given by

$$L(\theta, \phi) = n_{11} \log\{(1-\theta)\phi/2\} + n_{12} \log(\theta\phi/2) + n_{21} \log\{(1-\phi+\theta\phi)/2\} + n_{22} \log\{(1-\theta\phi)/2\},$$

and

$$L_2(\phi) = n_{1+} \log(\phi/2) + n_{2+} \log(1-\phi/2),$$

respectively. The conditional log-likelihood of the marker data M given the trait data T is $L_1(\theta, \phi) = L(\theta, \phi) - L_2(\phi)$. Maximizing the marginal log-likelihood $L_2(\phi)$ yields $\tilde{\phi} = 2n_{1+}/n$. After some algebra, the first order approximation in Equation (3) is given by $\rho = A/\lambda$, where

$$A = \frac{\phi(1-2\theta)}{2(1-(1-\theta)\phi)^2} + \frac{\phi(1-2\theta)}{2(1-\theta\phi)^2},$$
$$\lambda = -\frac{n\phi}{2(1-\theta)\theta} - \frac{n\phi^2(2-\phi)}{2(1-\theta\phi)(1-(1-\theta)\phi)}.$$

The adjusted log pseudo-likelihood is therefore given by $L_1(\theta + \tilde{\rho}(\theta), \tilde{\phi})$. The range of θ is from 0 to 0.5 and higher values of θ imply weaker linkage. In this example, θ and ϕ are not orthogonal. The limiting distribution of PLRT and APLRT is given by Theorem 5.1, in which the elements of information matrix are

$$I_{11} = \frac{\phi}{2(1-\theta)} + \frac{\phi}{2\theta} + \frac{\phi^2}{2(1-\phi+\theta\phi)} + \frac{\phi^2}{2(1-\theta\phi)}$$

TABLE 3: Observed frequencies and probabilities of combinations of T and M.

	Marker		
Trait	CB	BB	Marginal
Affected	n_{11}	n_{12}	n_{1+}
	$(1-\theta)\phi/2$	$\theta\phi/2$	$\phi/2$
Unaffected	n_{21}	n_{22}	n_{2+}
	$(1-(1-\theta)\phi)/2$	$(1-\theta\phi)/2$	$1-\phi/2$
Marginal	n_{+1}	n_{+2}	n
	1/2	1/2	1

and

$$I_{12} = -\frac{1}{2(1-\phi+\theta\phi)} + \frac{1}{2(1-\theta\phi)},$$

respectively.

In the simulation study, we set $\theta = 0.2$ and $\theta = 0.3$. Note that when θ is close to the boundary, the behaviour of PLRT is not regular (Chen & Liang, 2010). In Theorem 1, I_{11} and I_{12} can be computed algebraically with parameters replaced by their estimates. Then the corresponding quantiles of the asymptotic distributions in Theorem 1 are used as the critical values to calculate the empirical rejection rate. The results from 10,000 simulations are shown in Table 4. The following facts are observed. First, the Wald test is inaccurate especially when n is small. Second, the score test has an accurate rejection rate and often outperforms the PLRT. Third, the APLRT outperforms the PLRT and the AS outperforms the score test. This suggests that our adjustment by reducing the impact of nuisance parameters works well in finite samples.

TABLE 4: Empirical rejection rate of Wald, score, pseudo-likelihood ratio test (PLRT), adjusted pseudo-likelihood ratio test (APLRT) and adjusted pseudo-score test (AS) in Example 2 (nominal type I error 0.05).

n	Wald test	Score test	PLRT	APLRT	AS
		$\theta = 0.2, \phi = 0.4$			
20	0.017	0.038	0.029	0.038	0.042
40	0.025	0.039	0.044	0.050	0.046
60	0.037	0.042	0.064	0.055	0.046
		$\theta = 0.3, \phi = 0.6$			
20	0.022	0.044	0.028	0.045	0.047
40	0.028	0.047	0.042	0.050	0.048
60	0.032	0.048	0.062	0.051	0.049

7. ADJUSTED PROFILE COMPOSITE LIKELIHOOD

The composite likelihood, defined as the product of likelihoods for a set of marginal or conditional events, has been successfully applied in many areas of statistics. Well-known examples include independence likelihoods and pairwise likelihoods. Let $L^c(\psi) = \sum_{i=1}^n \sum_{k=1}^K \log f(A_k(y_i), \psi)$ be a composite log-likelihood for $\psi = (\theta, \phi)$, where $f(A_k(y_i), \psi)$ is the probability density of the kth event about random variable Y_i, that is, $A_k(Y_i)$. In many applications, the sample size n is large and the number of events K is small or moderate. Hence, the asymptotic regime considered for composite likelihood is $n \to \infty$ and K is fixed. For inference on θ, we can construct the profile composite log-likelihood $\widetilde{L^c}(\theta) = L^c(\theta, \tilde{\phi}(\theta))$, where $\tilde{\phi}(\theta) = \text{argmax}_\phi L^c(\theta, \phi)$. A review of composite likelihood methods is given in Varin, Reid, & Firth (2011).

In contrast to the pseudo-likelihood, which is based on a genuine likelihood function, and a consistent estimator of nuisance parameters, the composite likelihood is not a genuine likelihood function. Similar to the pseudo-score function, the profile composite score function $\partial\widetilde{L^c}(\theta)/\partial\theta$ has mean $O(1)$. Hence, we can use the same method to adjust the profile composite log-likelihood $\widetilde{L^c}(\theta)$. Specifically, the adjusted profile composite log-likelihood is given by $\widetilde{L^c}^*(\theta) = \widetilde{L^c}[\theta^*\{\tilde{\psi}(\theta)\}]$, where $\tilde{\psi}(\theta) = \{\theta, \tilde{\phi}(\theta)\}$, and the target parameter $\theta^*(\psi_0)$ satisfies $E_0\{\widetilde{L^c}_a(\theta)|_{\theta=\theta^*(\psi_0)}\} = 0$.

The adjusted profile composite log-likelihood $\widetilde{L^c}^*(\theta)$ can be calculated using the same first order approximation method in Section 4. Similar to (3), we have

$$\theta^{*a} - \theta_0^a = \rho^a(\theta_0) + O(n^{-2}) = \alpha_b\beta^{ab} + O(n^{-2}), \tag{8}$$

where $\rho^a(\theta_0) = \alpha_b\beta^{ab}$. For composite likelihoods, α_b is an approximation to $E_0\{\widetilde{L^c}_b(\theta_0)\}$, and β_{ab} is an approximation to $E_0\{-\widetilde{L^c}_{ab}(\theta_0)\}$. Under the asymptotic framework that the sample size n goes to infinity and the number of events K is fixed, the cumulants such as $E_0\{L^c_{ab}(\theta_0)\}$ and $E_0\{L^c_{abe}(\theta_0)\}$ are of order n. The derivation of α_b and β_{ab} is sketched in the Supplementary Materials.

In contrast to the pseudo-likelihood method, with composite likelihood, the corresponding full likelihood may involve additional nuisance parameters besides ϕ. For instance, association parameters are typically not present in the independence likelihood. In this case, θ^* as well as the adjusted profile composite log-likelihood may depend on some unidentified nuisance parameters. This identifiability issue can be addressed by applying the first order approximation method in Section 4, as that approximation only relies on the lower order cumulants of composite likelihood functions, which could be estimated empirically, even if θ^* is not identifiable. More detailed discussion on the identifiability issue is provided in the Supplementary Materials.

Let $\tilde{\theta}$ be the maximizer of $\widetilde{L^c}(\theta)$, $\tilde{\theta}^{**}$ be the maximizer of $\widetilde{L^c}^{**}(\theta)$, $W(\theta_0) = 2\{\widetilde{L^c}(\tilde{\theta}) - \widetilde{L^c}(\theta_0)\}$ denote the composite likelihood ratio test (CLRT) and $W^{**}(\theta_0) = 2\{\widetilde{L^c}^{**}(\tilde{\theta}^{**}) - \widetilde{L}^{**}(\theta_0)\}$ denote the adjusted composite likelihood ratio test (ACLRT). Let

$$I = \lim_{n\to\infty} E_0\left\{-\frac{1}{n}\frac{\partial^2 L^c(\psi_0)}{\partial\psi\partial\psi^{\mathrm{T}}}\right\}, \quad J = \lim_{n\to\infty} E_0\left\{\frac{1}{n}\frac{\partial L^c(\psi_0)}{\partial\psi}\left(\frac{\partial L^c(\psi_0)}{\partial\psi}\right)^{\mathrm{T}}\right\},$$

denote the sensitivity matrix and variability matrix (Varin, Reid, & Firth, 2011). The Godambe information matrix is given by $G = IJ^{-1}I$. With the partition $\psi = (\theta, \phi)$, we partition I and I^{-1} as

$$I = \begin{pmatrix} I_{11} & I_{12} \\ I_{21} & I_{22} \end{pmatrix}, \qquad I^{-1} = \begin{pmatrix} I^{11} & I^{12} \\ I^{21} & I^{22} \end{pmatrix},$$

and similarly for G and G^{-1}. Similar to Section 5.1, it is easily seen that $\tilde{\theta}$ and $\tilde{\theta}^{**}$ are asymptotically first order equivalent. The following theorem analogous to Theorem 1 establishes the asymptotic distribution of ACLRT.

Theorem 4. *Under the same regularity conditions in Molenberghs & Verbeke (2005),*

$$n^{1/2}(\tilde{\theta}^{**} - \theta_0) \to_d N(0, G^{11}),$$

and $W^{**}(\theta_0) \to_d \sum_{j=1}^{p} \delta_j U_j$, *where* U_j'*s are independent* χ_1^2 *variables and* δ_j'*s are the eigenvalues of* $(I^{11})^{-1}G^{11}$.

This theorem is useful for determining the critical values of the adjusted composite likelihood ratio test. Note that, due to the information bias from the composite likelihood, we typically find $I \neq J$, which implies $G^{11} \neq I^{11}$. Thus, unlike the (adjusted) pseudo-likelihood ratio test, the (adjusted) composite likelihood ratio test may not have the limiting χ^2 distribution, even under the parameter orthogonality condition, defined by $I_{12} = 0$. For the higher order properties described in Section 5.2, it is easily seen that Theorems 2 and 3 remain valid for the adjusted profile composite likelihood.

7.1. Example 3: Pairwise Likelihood for the Probit Model

Let $Y_i = (Y_{i1}, \ldots, Y_{im})$ denote a m-dimensional vector of binary measurements on subject i, $i = 1, \ldots, n$. The probability of a positive response Y_{ij}, for $j = 1, \ldots, m$, conditionally on a random effect a_i, is specified as

$$\Phi^{-1}\{P(Y_{ij} = 1|a_i)\} = X_{ij}^{\mathrm{T}}\theta + a_i,$$

where $\Phi(x)$ denotes the standard normal distribution function, X_{ij} is a set of covariates associated with the fixed effect θ and $a_i \sim N(0, \sigma^2)$ is the random effect. Inference about θ and σ^2 can be based on the pairwise log-likelihood (Renard, Molenberghs, & Geys, 2004) given by

$$\begin{aligned} L^c(\theta, \sigma^2) = \sum_{i=1}^{n} \sum_{1 \le k < k' \le m} \Big[& I(Y_{ik'} = 1, Y_{ik} = 1) \log \Phi_2(\eta_{ik'}, \eta_{ik}; \rho) \\ & + I(Y_{ik'} = 1, Y_{ik} = 0) \log\{\Phi(\eta_{ik'}) - \Phi_2(\eta_{ik'}, \eta_{ik}; \rho)\} \\ & + I(Y_{ik'} = 0, Y_{ik} = 1) \log\{\Phi(\eta_{ik}) - \Phi_2(\eta_{ik'}, \eta_{ik}; \rho)\} \\ & + I(Y_{ik'} = 0, Y_{ik} = 0) \log\{1 - \Phi(\eta_{ik}) - \Phi(\eta_{ik'}) + \Phi_2(\eta_{ik'}, \eta_{ik}; \rho)\} \Big], \end{aligned}$$

where $\Phi_2(x, y; \rho)$ denotes the standardized bivariate normal distribution function with correlation $\rho = \sigma^2/(1 + \sigma^2)$, and $\eta_{ik} = X_{ik}^{\mathrm{T}}\theta/(1 + \sigma^2)^{1/2}$.

Assume that we are interested in the possibly multidimensional parameter θ and σ^2 is treated as a nuisance parameter. Let $\Phi'(z)$ denote the derivative of $\Phi(z)$ with respect to z, and $\Phi_2^{(j)}(z_1, z_2; z_3)$ denote the derivative of $\Phi_2(z_1, z_2; z_3)$ with respect to z_j for $j = 1, 2, 3$. It is seen that the composite score function for θ is given by

$$\begin{aligned} \frac{\partial L^c(\theta, \sigma^2)}{\partial \theta} = \sum_{i=1}^{n} \sum_{1 \le k < k' \le m} \Big[& I(Y_{ik'} = 1, Y_{ik} = 1) \frac{\Phi_{2\theta}(\eta_{ik'}, \eta_{ik}; \rho)}{\Phi_2(\eta_{ik'}, \eta_{ik}; \rho)} \\ & + I(Y_{ik'} = 1, Y_{ik} = 0) \frac{\Phi_\theta(\eta_{ik'}) - \Phi_{2\theta}(\eta_{ik'}, \eta_{ik}; \rho)}{\Phi(\eta_{ik'}) - \Phi_2(\eta_{ik'}, \eta_{ik}; \rho)} \end{aligned}$$

$$+ I(Y_{ik'} = 0, Y_{ik} = 1)\frac{\Phi_\theta(\eta_{ik}) - \Phi_{2\theta}(\eta_{ik'}, \eta_{ik}; \rho)}{\Phi(\eta_{ik}) - \Phi_2(\eta_{ik'}, \eta_{ik}; \rho)}$$
$$\left. + I(Y_{ik'} = 0, Y_{ik} = 0)\frac{-\Phi_\theta(\eta_{ik}) - \Phi_\theta(\eta_{ik'}) + \Phi_{2\theta}(\eta_{ik'}, \eta_{ik}; \rho)}{1 - \Phi(\eta_{ik}) - \Phi(\eta_{ik'}) + \Phi_2(\eta_{ik'}, \eta_{ik}; \rho)}\right],$$

where

$$\Phi_\theta(\eta_{ik}) = \frac{X_{ik}^{\mathrm{T}}\Phi'(\eta_{ik})}{(1+\sigma^2)^{1/2}}, \qquad \Phi_{2\theta}(\eta_{ik'}, \eta_{ik}; \rho) = \frac{X_{ik'}^{\mathrm{T}}\Phi_2^{(1)}(\eta_{ik'}, \eta_{ik}; \rho) + X_{ik}^{\mathrm{T}}\Phi_2^{(2)}(\eta_{ik'}, \eta_{ik}; \rho)}{(1+\sigma^2)^{1/2}}.$$

The composite score function for σ^2 has the same form as that for θ, in which $\Phi_\theta(\eta_{ik})$ and $\Phi_{2\theta}(\eta_{ik'}, \eta_{ik}; \rho)$ are replaced with

$$\Phi_{\sigma^2}(\eta_{ik}) = -\frac{X_{ik}^{\mathrm{T}}\Phi'(\eta_{ik})}{2(1+\sigma^2)^{3/2}},$$

and

$$\Phi_{2\sigma^2}(\eta_{ik'}, \eta_{ik}; \rho) = -\frac{X_{ik'}^{\mathrm{T}}\Phi_2^{(1)}(\eta_{ik'}, \eta_{ik}; \rho)\theta + X_{ik}^{\mathrm{T}}\Phi_2^{(2)}(\eta_{ik'}, \eta_{ik}; \rho)\theta}{2(1+\sigma^2)^{3/2}} + \frac{\Phi_2^{(3)}(\eta_{ik'}, \eta_{ik}; \rho)}{(1+\sigma^2)^2}.$$

Calculating the explicit forms for the second and third order derivatives of $L^c(\theta, \sigma^2)$ is possible but tedious. Instead, we applied numerical differentiation to compute the empirical estimates of the Godambe information matrix G in Theorem 4 and the first order approximation ρ in the Supplementary Materials.

In the simulation study, we assume that there is only one covariate $X_{ij} \sim N(0, 1)$ and $(\theta_0, \theta_1) = (0, 1)$. The number of simulations is 5000. We compare the performance of the Wald test, composite score test, composite likelihood ratio test (CLRT), two types of adjusted composite likelihood by Chandler & Bate (2007) and Pace, Salvan, & Sartori (2011) (CLRT-C and CLRT-P) and our proposed adjusted composite likelihood ratio test (ACLRT). Note that the critical values for CLRT and ACLRT are obtained from the corresponding quantiles of the linear combination of χ_1^2 in Theorem 4, and the critical values for CLRT-C and CLRT-P are obtained from the quantiles of χ_2^2. In Table 5, we find that the empirical rejection rate from the Wald and score tests is considerably higher than the nominal level. The adjusted composite likelihood by Chandler & Bate (2007) and Pace, Salvan, & Sartori (2011) (CLRT-C and CLRT-P) seems to slightly improve the composite likelihood ratio test (CLRT). However, both CLRT-C and CLRT-P are not accurate enough for small sample sizes. As expected, ACLRT yields the most accurate estimated type I error among the six testing procedures.

8. APPLICATIONS TO THE PROFILE LIKELIHOOD

The method described in Section 3 can be applied to the profile likelihood as well. Let $M(\theta) = L\{\theta, \hat{\phi}(\theta)\}$ be the profile log-likelihood for θ, where $\hat{\phi}(\theta)$ is the maximizer of $L(\theta, \phi)$ for a fixed θ. Our adjusted profile log-likelihood is $M^*(\theta) = M[\theta^*\{\hat{\psi}(\theta)\}]$, where $\hat{\psi}(\theta) = \{\theta, \hat{\phi}(\theta)\}$ and θ^* is defined by equation (1) with $\tilde{L}(\theta)$ replaced by $M(\theta)$. In contrast to the adjustment proposed in McCullagh & Tibshirani (1990), $M^*(\theta)$ is well defined, even if θ is multidimensional. By a Taylor series expansion, we have

$$M^*(\theta) = M(\theta) + \rho^a M_a + O_p(n^{-1}),$$

TABLE 5: Empirical rejection rate of Wald, score, CLRT, CLRT-C, CLRT-P and ACLRT in Example 3 (nominal type I error 0.05), where $m = 3$, $\theta = (0, 1)$.

σ^2	n	Wald test	Score test	CLRT	CLRT-C	CLRT-P	ACLRT
1	10	0.164	0.132	0.108	0.121	0.098	0.078
	20	0.147	0.142	0.101	0.115	0.098	0.082
	30	0.119	0.115	0.088	0.085	0.082	0.063
	40	0.088	0.079	0.070	0.069	0.062	0.054
2	10	0.173	0.147	0.095	0.123	0.092	0.077
	20	0.148	0.133	0.088	0.093	0.085	0.070
	30	0.140	0.139	0.089	0.093	0.085	0.068
	40	0.095	0.059	0.052	0.060	0.060	0.055

and this can be shown to be equivalent to Stern's adjustment (Stern, 1997) ignoring the error of order $O_p(n^{-1})$. From this perspective, our derivation explains how Stern's adjustment is constructed and it also provides a well interpreted target likelihood $M^*(\theta)$ to which Stern's adjustment approximates.

8.1. Profile Likelihood Inference in Example 1

To illustrate the usage of the proposed adjusted profile log-likelihood $M^*(\theta)$, we revisit the variance component model. Assume that $X_i = 1$ and $\Sigma(\theta) = \theta I_n$, where I_n is the $n \times n$ identity matrix. As shown by McCullagh & Tibshirani (1990), the modified profile log-likelihood is given by $L_{mp}(\theta) = -(n-1)\log\theta/2 - s^2/(2\theta)$, with the maximizer $\tilde{\theta} = s^2/(n-1)$, where $s^2 = \sum_{i=1}^n (Y_i - \bar{Y})^2/n$. By the definition of θ^*, it is easily shown that $\theta^*(\theta) = \theta(n-1)/n$. Then, the adjusted profile log-likelihood is given by $M^*(\theta) = L_{mp}(\theta)n/(n-1)$, with the same maximizer $\tilde{\theta}$. By the definition of ρ, the first order approximation to θ^* is given by $\theta(n-3)/(n-2)$, which differs from the exact result by $O(n^{-2})$.

We consider the following simulation scenario. Let $\beta = 1$, $\theta = 1$, and the number of simulations be 10,000. We compare the performance of the profile likelihood ratio test (PLRT), modified profile likelihood ratio test (MPLRT), exact adjusted profile likelihood ratio test (eAPLRT) and adjusted profile likelihood ratio test via first order approximations (apAPLRT) in Figure 2. As expected, apAPLRT is very close to eAPLRT as long as $n > 5$. Moreover, eAPLRT and apAPLRT perform better than PLRT, while MPLRT is the best in terms of the empirical rejection rate. This is because $L_{mp}(\theta)$ corresponds to the restricted likelihood which is a valid marginal likelihood function for θ alone McCullagh & Tibshirani (1990). From this perspective, the impact of nuisance parameters is entirely eliminated.

Pace & Salvan (2006) proposed a novel least favourable target log-likelihood $L(\theta, \phi_\theta)$, where ϕ_θ is defined as the maximizer of $E\{L(\theta, \phi)\}$ with respect to ϕ for fixed θ. As discussed in Section 3, our method also provides an interpretable target log-likelihood $M\{\theta^*(\psi)\}$. Both likelihoods can be called as target likelihoods, and they are not available in practice, because ϕ_θ in $L(\theta, \phi_\theta)$ and $\theta^*(\psi)$ in $M\{\theta^*(\psi)\}$ are unknown. However, there exist key differences between these two likelihoods. Note that, a natural estimator of ϕ_θ is $\hat{\phi}(\theta)$, that is the maximizer of $L(\theta, \phi)$ for a fixed θ. By replacing ϕ_θ with $\hat{\phi}(\theta)$ in $L(\theta, \phi_\theta)$, we obtain the profile log-likelihood $L\{\theta, \hat{\phi}(\theta)\}$. As shown by McCullagh & Tibshirani (1990), the profile score function has the bias of order $O(1)$, and therefore, a further bias correction procedure, such as that in McCullagh & Tibshirani (1990), is required in order to reduce the bias. On the other hand, the adjusted profile log-likelihood $M^*(\theta)$

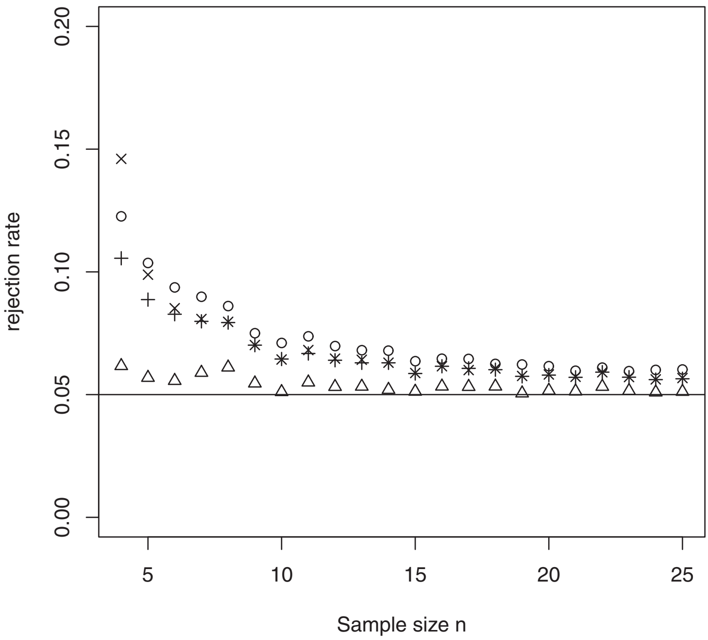

FIGURE 2: Empirical rejection rate of PLRT, MPLRT, eAPLRT and apAPLRT as a function of sample size n. The horizontal line represents the nominal type I error 0.05. The $\circ$, $\triangle$, $+$ and $\times$ symbols correspond to the PLRT, MPLRT, eAPLRT and apAPLRT, respectively. The $+$ and $\times$ symbols may overlap.

as a plug-in counterpart of $M\{\theta^*(\psi)\}$, only has a bias of order $O(n^{-1})$. From this perspective, the nuisance parameter effect has been reduced by the proposed target likelihood $M\{\theta^*(\psi)\}$ but not by the least favourable target likelihood $L(\theta, \phi_\theta)$.

9. DISCUSSION

In this article, we have described a simple non-additive adjustment to reduce the impact of nuisance parameters. A first order approximation method is proposed to calculate the adjustment for general models. The adjustment reduces the bias as well as the MSE of the score type functions. Numerical studies demonstrate satisfactory performance of the method under various models.

Consider the orthogonal parameter preserving transformations, that is, $(\theta, \phi) \mapsto (w(\theta), k(\phi))$, where $w(\theta)$ and $k(\phi)$ are one to one and smooth. The pseudo-likelihood under the new parametrization $(w(\theta), k(\phi))$ can be constructed by estimating $k(\phi)$ by $k(\tilde{\phi})$, since $k(\phi)$ does not involve θ. Hence, the pseudo-likelihood under the new parametrization is well defined. We can show from the definition that the adjusted pseudo-likelihood is invariant under such reparameterization. Similarly, we can show that the adjusted profile composite log-likelihood $\widetilde{L^c}^*(\theta)$ is invariant under more general transformations that preserve the parameter of interest, that is, $(\theta, \phi) \mapsto (w(\theta), k(\theta, \phi))$, where $w(\theta)$ is one to one and smooth, and $k(\theta, \phi)$ is smooth. As discussed in the end of Section 5, when θ is a scalar, there exist parameter of interest preserving transformations such that $w(\theta)$ and $k(\theta, \phi)$ are orthogonal.

Our proposed procedure has several extensions. First, the method can be applied to adjust the generalized profile likelihood (Severini, 1998b). The generalized profile log-likelihood is given by $L\{\theta, \tilde{\phi}(\theta)\}$, where $L(\theta, \phi)$ is the log-likelihood and $\tilde{\phi}(\theta)$ is an estimate of ϕ which may not be identical to the constrained maximum likelihood estimator $\hat{\phi}(\theta)$ defined in Section 8. Second, it is of interest to apply the proposed method to the composite likelihood in which the nuisance parameters are estimated by an estimator other than the constrained maximum composite likelihood estimator. Third, our procedure should be applicable to other types of likelihood functions, including partial likelihood and empirical likelihood. What is needed is a score function that is not unbiased, possibly due to the estimation of nuisance parameters. If the bridge function θ^* can be calculated analytically, then correction is straightforward. In the more general case when it cannot, we need to rely on the cumulants of the alternative likelihood function having the same order in n as assumed in Theorem 2.

In the current article, we focus on correcting the bias of the score type functions. However, the adjusted profile log-likelihood $M^*(\theta)$ does not reduce the information bias to second order (Diciccio et al., 1996). Due to the information bias of the composite likelihood, our proposed adjustment does not simplify the asymptotic distribution of the ACLRT, which is a possible limitation of the method. Recently, Pace, Salvan, & Sartori (2011) proposed a new type of adjustment with the purpose of correcting the information bias. Based on our adjusted profile composite log-likelihood, a further adjustment similar in spirit to Pace, Salvan, & Sartori (2011) can be used to calibrate the ACLRT. The performance of this procedure is worth further investigation.

APPENDIX

In this section, we will illustrate how to derive the leading term in (3), that is α_a and β_{ab}. For simplicity, we omit the parameter ψ_0 in some likelihood quantities, when they are evaluated at ψ_0, such as $L_a = L_a(\psi_0)$. A Taylor expansion of $\tilde{L}_a(\theta_0) = L_a(\theta_0, \tilde{\phi})$ about ϕ_0 yields,

$$\begin{aligned}\tilde{L}_a(\theta_0) &= L_a + L_{ae}(\tilde{\phi}^e - \phi_0^e) + \frac{1}{2}L_{aef}(\tilde{\phi}^e - \phi_0^e)(\tilde{\phi}^f - \phi_0^f) + O_p(n^{-1/2}) \\ &= L_a + \lambda_{ae}(\tilde{\phi}^e - \phi_0^e) + (L_{ae} - \lambda_{ae})(\tilde{\phi}^e - \phi_0^e) + \frac{1}{2}\lambda_{aef}(\tilde{\phi}^e - \phi_0^e)(\tilde{\phi}^f - \phi_0^f) + O_p(n^{-1/2}).\end{aligned}$$

In the first scenario, where the explicit form of $\tilde{\phi}$ is given, we obtain

$$E_0\{\tilde{L}_a(\theta_0)\} = \alpha_a + o(1),$$

where

$$\alpha_a = \lambda_{ae}E_0(\tilde{\phi}^e - \phi_0^e) + E_0\{(L_{ae} - \lambda_{ae})(\tilde{\phi}^e - \phi_0^e)\} + \frac{1}{2}\lambda_{aef}E_0\{(\tilde{\phi}^e - \phi_0^e)(\tilde{\phi}^f - \phi_0^f)\}.$$

In the second scenario, if $\tilde{\phi}$ is the solution of a q-dimensional unbiased estimating equation $H(y; \phi) = 0$ with $E_0\{H(y; \phi)\} = 0$, the result can be simplified. First, we need to introduce more notation. Let $H(y; \phi) = \{H_e(y; \phi), H_f(y; \phi), \ldots\}$. Note that the subscript here merely indicates the component of H, and we denote $H_{ef}(\phi) = \partial H_e(y; \phi)/\partial\phi^f$, and $\kappa_{ef} = E_0(H_{ef}(\phi_0))$. Since the estimating equation $H(y; \phi)$ may not be the derivative of an objective function, the first index is not exchangeable with the latter ones in $H_{ef}(\phi)$ and κ_{ef}. Let $\kappa_{e,f} = E_0(H_e H_f)$, $\kappa_{eg,f} = E_0(H_{eg}H_f)$, $\gamma_{ae,g} = E_0(L_{ae}H_g)$ and κ^{ef} be the $q \times q$ matrix inverse of κ_{ef}. Note that

$$\gamma_{a,g} = E_0(L_a H_g) = \int \frac{\partial f(y; \theta_0, \phi_0)}{\partial\theta^a} H_g(y; \phi_0)\, \mathrm{d}y = \frac{\partial}{\partial\theta^a} E_0\{H_g(y; \phi_0)\} = 0.$$

A Taylor expansion of $H(x, \tilde{\phi}) = 0$ around ϕ gives

$$\tilde{\phi}^e - \phi_0^e = -\kappa^{ge}H_g + \kappa^{ge}\kappa^{hf}(H_{gf} - \kappa_{gf})H_h - \frac{1}{2}\kappa^{ge}\kappa^{g'e'}\kappa^{hf}\kappa_{ge'f}H_{g'}H_h + o_p(n^{-1}).$$

Following the arguments in the Appendix of Severini (2002), we obtain

$$E_0(\tilde{\phi}^e - \phi_0^e) = \kappa^{ge}\kappa^{hf}\kappa_{gf,h} - \frac{1}{2}\kappa^{ge}\kappa^{g'e'}\kappa^{hf}\kappa_{ge'f}\kappa_{g',h} + O(n^{-2}), \tag{9}$$

$$E_0(\tilde{\phi}^e - \phi_0^e)(L_{ae} - \lambda_{ae}) = -\kappa^{ge}\gamma_{ae,g} + O(n^{-1}), \tag{10}$$

$$E_0(\tilde{\phi}^e - \phi_0^e)(\tilde{\phi}^f - \phi_0^f) = \kappa^{ge}\kappa^{hf}\kappa_{g,h} + O(n^{-2}). \tag{11}$$

From Equations (9)–(11), we obtain,

$$\alpha_a = \lambda_{ae}\kappa^{ge}\kappa^{hf}(\kappa_{gf,h} - \frac{1}{2}\kappa^{g'e'}\kappa_{ge'f}\kappa_{g',h}) - \kappa^{ge}\gamma_{ae,g} + \frac{1}{2}\lambda_{aef}\kappa^{ge}\kappa^{hf}\kappa_{g,h}. \tag{12}$$

Next, we will derive β_{ab}. By Taylor expansion, we have

$$\tilde{L}_{ab}(\theta_0) = L_{ab}(\psi_0) + L_{abe}(\psi_0)(\tilde{\phi}^e - \phi_0^e) + o_p(n^{1/2}).$$

Since $L_{abe}(\psi_0)(\tilde{\phi}^e - \phi_0^e) = O_p(n^{1/2})$, we obtain $E_0\{\tilde{L}_{ab}(\theta_0)\} = \lambda_{ab} + O(n^{1/2})$. This result holds under both scenarios. Thus, as an approximation to $E_0\{-\tilde{L}_{ab}(\theta_0)\}$, we have $\beta_{ab} = -\lambda_{ab}$.

ACKNOWLEDGEMENTS

This article is based on the first author's PhD thesis research. Financial support from the Department of Biostatistics at Johns Hopkins University and the Natural Sciences and Engineering Council of Canada is acknowledged. We are grateful to the Associate Editor and referees for their detailed and constructive comments on an earlier version.

BIBLIOGRAPHY

Abney, M., McPeek, M. S., & Ober, C. (2000). Estimation of variance components of quantitative traits in inbred populations. *The American Journal of Human Genetics*, 66, 629–650.

Adimari, G. & Ventura, L. (2002). Quasi-profile loglikelihood for unbiased estimating functions. *Annals of the Institute of Statistical Mathematics*, 54, 235–244.

Barndorff-Nielsen, O. E. (1983). On a formula for the distribution of the maximum likelihood estimator. *Biometrika*, 70, 343–365.

Barndorff-Nielsen, O. E. & Cox, D. R. (1994). *Inference and Asymptotics*, Chapman & Hall, London.

Bellio, R., Greco, L., & Ventura, L. (2008). Modified quasi-profile likelihoods from estimating functions. *Journal of Statistical Planning and Inference*, 138, 3059–3068.

Besag, J. (1974). Spatial interaction and the statistical analysis of lattice systems. *Journal of the Royal Statistical Society, Series B*, 36, 192–236.

Chandler, R. E. & Bate, S. (2007). Inference for clustered data using the independence loglikelihood. *Biometrika*, 94, 167–183.

Chen, Y. & Liang, K. Y. (2010). On the asymptotic behaviour of the pseudolikelihood ratio test statistic with boundary problems. *Biometrika*, 97, 603–620.

Cox, D. R. & Reid, N. (1987). Parameter orthogonality and approximate conditional inference. *Journal of the Royal Statistical Society, Series B*, 49, 1–39.

Cox, D. R. & Reid, N. (2004). A note on pseudolikelihood constructed from marginal densities. *Biometrika*, 91, 729–737.

Diciccio, T., Martin, M. A., Stern, S., & Young, G. A. (1996). Information bias and adjusted profile likelihoods. *Journal of the Royal Statistical Society, Series B*, 58, 189–203.

Fraser, D. (2003). Likelihood for component parameters. *Biometrika*, 90, 327–339.

Godambe, V. P. (1960). An optimum property of regular maximum likelihood estimation. *The Annals of Mathematical Statistics*, 31, 1208–1211.

Gong, G. & Samaniego, F. J. (1981). Pseudo maximum likelihood estimation: Theory and applications. *The Annals of Statistics*, 9, 861–869.

Hanfelt, J. J. & Liang, K.-Y. (1995). Approximate likelihood ratios for general estimating functions. *Biometrika*, 82, 461–477.

Hauck, J., Walter, W., & Donner, A. (1977). Wald's test as applied to hypotheses in logit analysis. *Journal of the American Statistical Association*, 72, 851–853.

Jiang, W. & Turnbull, B. (2004). The indirect method: Inference based on intermediate statistics: A synthesis and examples. *Statistical Science*, 19, 239–263.

Jorgensen, B. & Knudsen, S. J. (2004). Parameter orthogonality and bias adjustment for estimating functions. *Scandinavian Journal of Statistics*, 31, 93–114.

Khoury, M. J., Beaty, T. H., & Cohen, B. H. (1993). *Fundamentals of Genetic Epidemiology*, Oxford University Press, New York.

Liang, K. Y. (1987). Estimating functions and approximate conditional likelihood. *Biometrika*, 74, 695–702.

Liang, K. Y., Rathouz, P. J., & Beaty, T. H. (1996). Determining linkage and mode of inheritance: Mod scores and other methods. *Genetic Epidemiology*, 13, 575–593.

Liang, K. Y. & Self, S. (1996). On the asymptotic behaviour of the pseudolikelihood ratio test statistic. *Journal of the Royal Statistical Society, Series B*, 58, 785–796.

Lindsay, B. G. (1988). Composite likelihood methods. *Contemporary Mathematics*, 80, 221–239.

McCullagh, P. & Tibshirani, R. (1990). A simple method for the adjustment of profile likelihoods. *Journal of the Royal Statistical Society, Series B*, 52, 325–344.

Molenberghs, G. & Verbeke, G. (2005). *Models for Discrete Longitudinal Data*, Springer, New York.

Pace, L. & Salvan, A. (2006). Adjustments of the profile likelihood from a new perspective. *Journal of Statistical Planning and Inference*, 10, 3554–3564.

Pace, L., Salvan, A., & Sartori, N. (2011). Adjusting composite likelihood ratio statistics. *Statistica Sinica*, 21, 129–148.

Rathouz, P. & Liang, K. Y. (1999). Reducing sensitivity to nuisance parameters in semiparametric models: A quasiscore method. *Biometrika*, 86, 857–869.

Renard, D., Molenberghs, G., & Geys, H. (2004). A pairwise likelihood approach to estimation in multilevel probit models. *Computational Statistics and Data Analysis*, 44, 649–667.

Santamaria, A., Diego, V., Almasy, L., Rainwater, D. L., Mahaney, M. C., Comuzzie, A. G., Cole, S., Dyer, T. D., Tracy, R., & Stern, M. (2007). Quantitative trait locus on chromosome 12q14.1 influences variation in plasma plasminogen levels in the San Antonio family heart study. *Human Biology*, 79, 515–523.

Severini, T. A. (1998a). An approximation to the modified profile likelihood function. *Biometrika*, 85, 403–411.

Severini, T. A. (1998b). Likelihood functions for inference in the presence of a nuisance parameter. *Biometrika*, 85, 507–522.

Severini, T. A. (2000). *Likelihood Methods in Statistics*, Oxford University Press, New York.

Severini, T. A. (2002). Modified estimating functions. *Biometrika*, 89, 333–344.

Stern, S. E. (1997). A second-order adjustment to the profile likelihood in the case of a multidimensional parameter of interest. *Journal of the Royal Statistical Society, Series B*, 59, 653–665.

Varin, C., Reid, N., & Firth, D. (2011). An overview of composite likelihood methods. *Statistica Sinica*, 21, 5–42.

Wang, M. & Hanfelt, J. J. (2003). Adjusted profile estimating function. *Biometrika*, 90, 845–858.

Yi, G. Y. & Reid, N. (2010). A note on misspecified estimating functions. *Statistica Sinica*, 20, 1749–1769.

Received 6 June 2013
Accepted 30 August 2014