Journal of Statistical Planning and Inference 141 (2011) 3047–3054

Contents lists available at ScienceDirect

Journal of Statistical Planning and Inference

journal homepage: www.elsevier.com/locate/jspi

On the robustness of maximum composite likelihood estimate

Ximing Xu *, N. Reid

Department of Statistics, University of Toronto, 100 St. George St. Toronto, Ontario, Canada M5S 3G3

ARTICLE INFO

Article history:
Received 9 October 2010
Received in revised form
23 March 2011
Accepted 29 March 2011
Available online 2 April 2011

Keywords:
Pseudo-likelihood
Consistency
Godambe information
Model misspecification

ABSTRACT

Composite likelihood methods have been receiving growing interest in a number of different application areas, where the likelihood function is too cumbersome to be evaluated. In the present paper, some theoretical properties of the maximum composite likelihood estimate (MCLE) are investigated in more detail. Robustness of consistency of the MCLE is studied in a general setting, and clarified and illustrated through some simple examples. We also carry out a simulation study of the performance of the MCLE in a constructed model suggested by Arnold (2010) that is not multivariate normal, but has multivariate normal marginal distributions.

1. Introduction

The likelihood function plays a critical role in statistical inference in both frequentist and Bayesian frameworks. However, with the current explosion in the size of data sets and the increase in complexity of the dependencies among variables in many realistic models, it is often impractical or cumbersome to construct the full likelihood. In these situations, composite likelihoods, which are usually constructed by compounding some lower dimensional likelihoods, can be considered as a convenient surrogate. Suppose Y is a p-dimensional random vector with probability density function $f(y;\theta)$ for some q-dimensional parameter vector $\theta \in \Theta$, and suppose $\{A_1,\ldots,A_K\}$ is a set of events with associated likelihood functions $L_k(\theta;y) \propto f(y \in A_k;\theta)$ $(k=1,2,\ldots,K)$. Following Lindsay (1988), the composite likelihood function is defined as

$$CL(\theta;y) = \prod_{k=1}^{K} L_k(\theta;y)^{w_k}, \tag{1}$$

where $\{w_k\}$ is a set of non-negative weights. Note that $L_k(\theta;y)$ might depend only on a sub-vector of θ. The choice of the component likelihoods $L_k(\theta;y)$ and the weights $\{w_k\}$ may be critical to improve the accuracy and efficiency of the resulting statistical inference (Lindsay, 1988; Joe and Lee, 2009; Varin et al., 2011). From the above definition it is easy to see that the full likelihood is a special case of composite likelihood; however, composite likelihood will not usually be a genuine likelihood function, that is, it may not be proportional to the density function of any random vector.

The most commonly used versions of composite likelihood are composite marginal likelihood and composite conditional likelihood. Two examples of composite conditional likelihood functions are the pairwise composite conditional likelihood function,

$$\mathcal{L}_{PC}(\theta;y) = \prod_{r=1}^{p} \prod_{s \neq r} f(y_r|y_s;\theta)^{w_{rs}}, \tag{2}$$

* Corresponding author.
E-mail address: ximing@utstat.utoronto.ca (X. Xu).

0378-3758/$ - see front matter © 2011 Elsevier B.V. All rights reserved.
doi:10.1016/j.jspi.2011.03.026

and the full conditional likelihood composite likelihood function,

$$\mathcal{L}_{FC}(\theta;y)=\prod_{r=1}^{p} f(y_r|y_{(-r)};\theta)^{w_r}, \tag{3}$$

where $y_{(-r)}$ denotes the random vector with y_r deleted. Two particularly useful composite marginal likelihood functions are the independence marginal likelihood function,

$$\mathcal{L}_{ind}(\theta;y)=\prod_{r=1}^{p} f(y_r;\theta)^{w_r}, \tag{4}$$

and the pairwise likelihood function

$$\mathcal{L}_{pair}(\theta;y)=\prod_{r=1}^{p-1}\prod_{s=r+1}^{p} f(y_r,y_s;\theta)^{w_{rs}}. \tag{5}$$

With a sample of independent observations $\mathbf{y}=(y^{(1)},\ldots,y^{(n)})$, the overall composite log-likelihood function is

$$c\ell(\theta;\mathbf{y})=\sum_{i=1}^{n} c\ell(\theta;y^{(i)})=\sum_{i=1}^{n}\log CL(\theta;y^{(i)}), \tag{6}$$

and the maximum composite likelihood estimator (MCLE) is defined by

$$\hat{\theta}_{CL}=\underset{\theta}{\operatorname{argmax}}\, c\ell(\theta;\mathbf{y}). \tag{7}$$

Composite likelihood methods have proved useful in a range of complex applications, including models for spatial processes, models for statistical genetics and models for clustered data; several of these are surveyed in Varin et al. (2011). In addition to computational convenience, inference based on the composite likelihood may have good properties. For example, because each of the components of the composite likelihood is based on a density, the estimating equation obtained from the derivative of the composite log-likelihood function is unbiased. In modelling only lower dimensional marginal or conditional densities, composite conditional or marginal likelihood inference is widely viewed as robust, in the sense that the inference is valid for a range of statistical models consistent with the component densities. In the following sections we will study the consistency and robustness of the maximum composite likelihood estimator in more detail.

2. Aspects of robustness for the MCLE

This section and the next is a complement to the discussions on the robustness of composite likelihood inference in Varin (2008) and Varin et al. (2011). To formulate ideas about robustness we distinguish between the true data-generating model, and the model used for inference, following Kent (1982). We suppose the random vector Y has distribution function $G(y)$; the marginal distribution function for a sub-vector $Y_k \subset Y$ is $G_k(y_k)$ and the corresponding density function is $g_k(y_k), k=1,\ldots,K$, with respect to some dominating measure μ. Now consider the family of modelled distributions for Y_k, with common support and family of density functions $\{f_k(y_k;\theta);\theta\in\Omega\}$ with respect to the same dominating measure μ. We restrict attention to the unweighted composite marginal likelihood:

$$CL(\theta;y)=\prod_{k=1}^{K} f_k(y_k;\theta). \tag{8}$$

The family of densities is correctly specified if there exists $\theta_0\in\Omega$ such that $f(y;\theta_0)=g(y)$; if no such θ_0 exists, the model is misspecified. The composite marginal likelihood (8) is correctly specified if all component families $\{f_k(y_k;\theta);\theta\in\Omega\}$ are correctly specified.

If the full model is misspecified, then as in Kent (1982) and White (1982), we define θ^*_{ML} as the parameter which minimizes the Kullback–Leibler divergence between the specified full model and the true model $g(\cdot)$. Similarly, for misspecified composite likelihood, θ^* is a parameter point which minimizes the composite Kullback–Leibler divergence (Varin and Vidoni, 2005):

$$\theta^*=\underset{\theta}{\operatorname{argmin}} E_g\left\{\log\frac{\prod_{k=1}^{K} g_k(Y_k)}{CL(\theta;Y)}\right\}=\underset{\theta}{\operatorname{argmin}}\sum_{k=1}^{K} E_g\left\{\log\frac{g_k(Y_k)}{f_k(Y_k;\theta)}\right\}. \tag{9}$$

Consistency of the maximum composite likelihood estimator is claimed in several papers, although without detailed proof; see for example Lindsay (1988), Molenberghs and Verbeke (2005) and Jin (2009). Asymptotic results on misspecified full likelihood functions, as in White (1982), cannot be applied to the case of composite likelihood directly, since the composite likelihood function will not usually be a genuine likelihood function, as mentioned in Section 1. In the Appendix we adapt Wald's classical approach (Wald, 1949) to establish the result that the MCLE converges almost surely to θ^* defined in (9), taking model misspecification into account. The regularity conditions are analogous to those given in Wald's proof, but applied to the component likelihoods without explicit assumptions on the full likelihood.

Table 1
Model specification.

Model	Full likelihood	Composite likelihood
Correctly specified	$f(y;\theta_0)=g(y)$	$f_k(y;\theta_0)=g_k(y)$ for all k
	$\hat{\theta}_{ML}\to\theta_0$	$\hat{\theta}_{CL}\to\theta_0$
Misspecified	$f(y;\theta)\neq g(y)$,	$f_k(y;\theta)\neq g_k(y)$ for some k
	$\hat{\theta}_{ML}\to\theta^*_{ML}$	$\hat{\theta}_{CL}\to\theta^*$

Given consistency, the usual results on estimating equations, and some further regularity conditions, imply that the MLE and MCLE are asymptotically normally distributed as the sample size $n\to\infty$. The MLE has asymptotic variance determined by the expected Fisher information, and the asymptotic variance of the MCLE is calculated as the inverse of the Godambe information matrix $G(\theta)=H(\theta)J(\theta)^{-1}H(\theta)$ (Lindsay, 1988; Varin, 2008) where $H(\theta)=E\{-\nabla_\theta u(\theta;\mathbf{Y})\}$, $J(\theta)=\mathrm{var}_\theta\{u(\theta;\mathbf{Y})\}$ and $u(\theta;\mathbf{Y})=\nabla_\theta c\ell(\theta;\mathbf{Y})$ with $c\ell(\theta)$ defined as (6). Here ∇_θ is the operation of differentiation with respect to the parameter θ.

Model specifications under different mechanisms and their impact on the convergence of the resulting maximum likelihood estimators are illustrated schematically in Table 1. The first row illustrates the result that has been most studied: when the model and sub-models are correctly specified, the resulting MCLE and MLE are both consistent for the true parameter value, under some regularity conditions, and the MCLE will be less efficient than the MLE, although a number of examples indicate that the loss of efficiency can be quite small.

The interesting case for studying robustness is when the components of composite likelihood, such as lower dimensional marginal densities, are correctly specified, but the full likelihood is misspecified; we call this robustness of consistency. In this case the MLE will not usually be consistent for the true parameter value. On the other hand the MCLE, which is calculated from the composite likelihood making use of the correctly specified lower dimensional margins only, still converges to the true parameter value without depending on the joint model. However, the asymptotic variance of the MCLE may vary dramatically according to different true joint models.

Finally, if both the composite and the full likelihood are not correctly specified, the MCLE or MLE will converge not to the true parameter, but to θ^* or to θ^*_{ML}.

Jin (2009, Ch. 5) considered robustness of efficiency, in a particular construction for multivariate binary data, through simulations comparing the efficiency of the MCLE to that of the MLE.

3. Some examples

We illustrate some of the points above with some simple examples constructed to highlight aspects of robustness.

Example 1 (*Estimation of association parameters*). This example is due to Andrei and Kendziorski (2009). Suppose $Y_1\sim N(\mu_1,\sigma_1^2)$, $Y_2\sim N(\mu_2,\sigma_2^2)$ and $\varepsilon\sim N(0,1)$ are independent random variables. Let $Y_3=Y_1+Y_2+bY_1Y_2+\varepsilon$, $b\neq 0$. We can show that all full conditional distributions i.e. $f(Y_1|Y_2,Y_3)$, $f(Y_2|Y_1,Y_3)$ and $f(Y_3|Y_1,Y_2)$ are normal, but the joint distribution is not multivariate normal due to the non-zero interaction term bY_1Y_2. If we misspecify the joint model as multivariate normal, b will be estimated as 0 directly. If we use the full conditional distribution $f(Y_3|Y_1,Y_2)$, the MCLE of b is $\hat{b}_{CL}=\sum_{i=1}^n y_{1i}y_{2i}(y_{3i}-y_{1i}-y_{2i})/\sum_{i=1}^n(y_{1i}y_{2i})^2$, which is consistent for b. We can also use $f(Y_1|Y_2,Y_3)$ or $f(Y_2|Y_1,Y_3)$, but the resulting MCLE cannot be expressed in a closed form and some numerical methods are needed.

Example 2 (*Estimation of the correlation*). The random vector $(Y_1,Y_2,Y_3,Y_4)'$ follows a multivariate normal distribution with mean vector $(0,0,0,0)'$ and covariance matrix

$$\Sigma=\begin{pmatrix}1 & \rho_0 & 2\rho_0 & 2\rho_0\\ \rho_0 & 1 & 2\rho_0 & 2\rho_0\\ 2\rho_0 & 2\rho_0 & 1 & \rho_0\\ 2\rho_0 & 2\rho_0 & \rho_0 & 1\end{pmatrix}.$$

Suppose we know the correlation between Y_1 and Y_2 is the same as the correlation between Y_3 and Y_4. If we model the joint distribution of $(Y_1,Y_2,Y_3,Y_4)'$ as multivariate normal with zero mean vector and all correlations equal, the covariance matrix is then misspecified and the resulting MLE will not be consistent for ρ_0. On the other hand, if we only use the correct information about the pairs (Y_1,Y_2) and (Y_3,Y_4) and construct the composite likelihood

$$CL(\rho;y_1,y_2,y_3,y_4)=f_{12}(y_1,y_2;\rho)f_{34}(y_3,y_4;\rho) \tag{10}$$

where both f_{12} and f_{34} are the density functions for a bivariate normal with mean vector $(0,0)'$ and covariance matrix

$$\Sigma=\begin{pmatrix}1 & \rho\\ \rho & 1\end{pmatrix},$$

then by Corollary 1 in the Appendix, the resulting MCLE is consistent for ρ_0.

It is of interest to note that the parameter constraint needed to ensure that the covariance matrix is non-negative definite in the correct full likelihood is $-1/5 \le \rho \le 1/3$, whereas in the composite likelihood (10) the parameter constraint is $-1 \le \rho \le 1$. The composite likelihood (10) can also be thought as the full likelihood for a multivariate normal distribution with a block diagonal covariance matrix, which is obviously different from the true full model.

From this example we can see that even if different parameter constraints are imposed or the composite likelihood is compatible with different full models, the MCLE will be consistent as long as all of the component likelihoods are correctly specified.

Example 3 (*No compatible joint density exists*). Suppose the true model for the random vector (Y_1,Y_2,Y_3) is multivariate normal with mean vector $(\mu_0,\mu_0,\mu_0)'$ $(\mu_0 > 0)$, and covariance matrix equal to the identity matrix. Now consider the following pairwise likelihood

$$CL(\mu; y_1,y_2,y_3) = f_{12}(y_1,y_2;\mu) f_{13}(y_1,y_3;\mu) f_{23}(y_2,y_3;\mu) \tag{11}$$

where both f_{12} and f_{23} are the density functions for a bivariate normal density with unknown mean vector $(\mu,\mu)'$ and covariance matrix equal to the 2×2 identity matrix. However, $f_{13}(y_1,y_3;\mu)$ is misspecified as

$$f_{13}(y_1,y_3;\mu) = \frac{1}{\mu}\exp\left(-\frac{y_1}{\mu}\right)\frac{1}{\sqrt{2\pi}}\exp\left(-\frac{(y_3-\mu)^2}{2}\right)$$

It is easy to see the no compatible joint density exists for the composite likelihood (11) since from f_{12} and f_{13} we get different marginal densities for Y_1.

The MCLE of μ from the composite likelihood function (11) can be obtained by solving the score equation

$$5n\mu^3 - S_n\mu^2 + n\mu - S_{1n} = 0 \tag{12}$$

where $S_n = \sum_{i=1}^n (Y_{1i}+2Y_{2i}+2Y_{3i})$ and $S_{1n} = \sum_{i=1}^n Y_{1i}$. As $n\to\infty$, a direct argument using the consistency of sample means for the population mean shows that the unique real root of (12) converges to μ_0. The asymptotic variance of $\hat{\mu}_{CL}$ can be calculated using the Godambe information function $G(\theta)$, and the ratio of the asymptotic variance of $\hat{\mu}_{ML}$ to that of $\hat{\mu}_{CL}$ is $r = \{5+(1/\mu^2)\}^2/3[8+\{1+(1/\mu^2)\}^2]$. It is easy to check $r \le 1$ and equality holds only for $\mu = 1$.

From this artificial example, we can see that although no compatible joint density exists, the limit of the MCLE may still be meaningful, even consistent for the true value of parameter. In general the MCLE converges to θ^* which minimizes the composite Kullback–Leibler divergence whether the specified sub-models are compatible or not. If the specified sub-models are very close to the corresponding true sub-models, we can imagine that θ^* should be a good estimate of the true parameter value even if those specified sub-models are incompatible.

Example 4 (*A class of distributions with normal margins, Arnold, 2010*). Suppose the random vector $Y = (Y_1,Y_2,\ldots,Y_p)$ has the following density function:

$$f(Y) = \phi_p(Y;\mu,\Sigma) + g(\mu,\Sigma)\left(\prod_{i=1}^p Y_i\right) I_A(Y), \tag{13}$$

where $\phi^{(p)}(Y;\mu,\Sigma)$ is the density function of p-dimensional multivariate normal with mean vector μ and covariance matrix Σ, $g(\cdot)$ is a function of parameters chosen to guarantee that $f(Y)\ge 0$, $A = \{Y : -t \le Y_i \le t, i = 1,2,\ldots,p\}$, t is a threshold parameter, and $I_A(Y) = 1$ if $Y \in A$ and 0 otherwise. All $k < p$ dimensional sub-vectors of Y follow k-dimensional multivariate normal distributions with corresponding mean vectors and covariance matrices. When $t = 0$, $f(Y)$ becomes $\phi^{(p)}(Y;\mu,\Sigma)$. This example also provides a general approach to construct a density with the same margins as a pre-specified density. In model (13), depending on the complexity of the function $g(\cdot)$, the calculation of the MLE may be very difficult. In the simulation study, we let $t=1$, $\mu = \mathbf{0}$ and $\Sigma = (1-\rho)I_p + \rho J_p$, where I_p is identity matrix, J_p is a $p\times p$ matrix with all entries equal to 1, and ρ is the common correlation coefficient for $p \ge 3$. Since $A \subseteq \{Y : Y'Y \le p\}$, we can choose the function g as

$$g(\mu,\Sigma) = \inf_{Y'Y\le p} \phi^{(p)}(Y;\mu,\Sigma) \le \inf_{Y\in A} \phi^{(p)}(Y;\mu,\Sigma)$$

To calculate $g(\mu,\Sigma)$, we use the fact that

$$\sup_{Y'Y\le p} Y'\Sigma^{-1}Y = p\lambda_p,$$

where λ_p is the largest eigenvalue of Σ^{-1}, and is $1/(1-\rho)$ if $0 \le \rho < 1$, and $1/\{1+(p-1)\rho\}$ if $1/(1-p) < \rho \le 0$.

We begin with $p=3$ and consider three different estimators of ρ: the MLE $\hat{\rho}$; the MCLE, $\hat{\rho}_{CL}$ obtained by maximizing the pairwise likelihood (5) with equal weights, and the simple unbiased estimator based on the method of moments,

$$\tilde{\rho} = \frac{2S_2}{np(p-1)}, \quad \text{where } S_2 = \sum_{i=1}^n \sum_{s>r}^p Y_r^{(i)} Y_s^{(i)}.$$

The last two estimators are free of the function $g(\cdot)$ and are more computationally convenient than the MLE.

Table 2
Performances of $\hat{\rho}$, $\hat{\rho}_{CL}$ and $\tilde{\rho}$ when $n=100$, $M=10\,000$, $p=3$ and $t=1$.

True value of ρ	−0.49	−0.25	0	0.25	0.5	0.75	0.99
Sim. mean of $\hat{\rho}_{CL}$	−0.4924	−0.2512	−0.0012	0.2515	0.4986	0.7487	0.9899
Sim. mean of $\hat{\rho}$	−0.4900	−0.2481	0.0019	0.2489	0.4983	0.7502	0.9900
Sim. mean of $\tilde{\rho}$	−0.4908	−0.2479	−0.0015	0.2521	0.4998	0.7511	0.9874
Sim. variance of $\hat{\rho}_{CL}$	0.0008	0.0013	0.0036	0.0037	0.0024	0.0006	10^{-6}
Sim. variance of $\hat{\rho}$	10^{-6}	0.0012	0.0036	0.0036	0.0023	0.0059	10^{-6}
Sim. variance of $\tilde{\rho}$	0.0025	0.0023	0.0035	0.0057	0.0092	0.0155	0.0215
$S_{\hat{\rho}}/S_{\hat{\rho}_{CL}}$	0.0025	0.9231	1.0000	0.9730	0.9583	0.9833	1.0000
$S_{\hat{\rho}_{CL}}/S_{\tilde{\rho}}$	0.3334	0.5614	1.0252	0.6521	0.2599	0.0402	10^{-5}

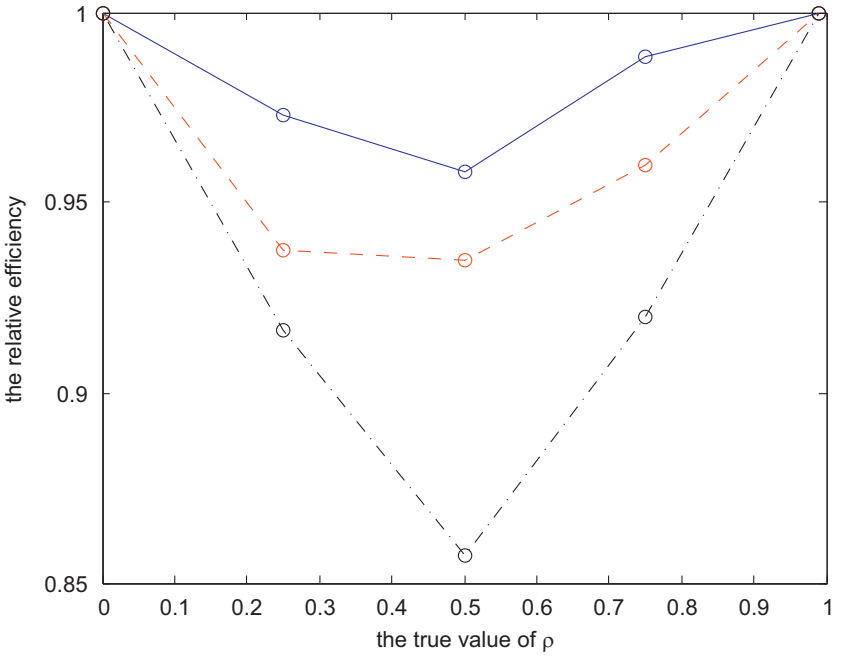

Fig. 1. The ratio of the simulated variances, $S_{\hat{\rho}}/S_{\hat{\rho}_{CL}}$, as a function of ρ. The lines shown are for $p=3,6,8$ (descending).

We used rejection sampling to generate n sample points from the joint distribution (13), using the fact that

$$f(Y) \le \phi^{(p)}(Y;\mu,\Sigma)\left\{1+I_A(Y)\prod_{i=1}^{p} Y_i\right\} \le 2\phi^{(p)}(Y;\mu,\Sigma).$$

We used numerical methods to calculate $\hat{\rho}$ and $\hat{\rho}_{CL}$, solving the relevant score equations, and calculated simulation means and variances of $\hat{\rho}$, $\hat{\rho}_{CL}$ and $\tilde{\rho}$. In Table 2 the notations $S_{\hat{\rho}}$, $S_{\hat{\rho}_{CL}}$ and $S_{\tilde{\rho}}$ are used for the simulation variances. The ratios $S_{\hat{\rho}_{CL}}/S_{\tilde{\rho}}$ and $S_{\hat{\rho}}/S_{\hat{\rho}_{CL}}$ are used to compare the efficiencies of the three estimators.

The results for sample size $n=100$, simulation size $M=10\,000$, threshold $t=1$ and dimension $p=3$ are presented in Table 2. All three methods produce accurate point estimates. With the exception of $\rho=-0.49$, $\mathrm{var}(\hat{\rho}_{CL})$ is very close to $\mathrm{var}(\hat{\rho})$, and $\mathrm{var}(\hat{\rho}_{CL})$ seems smaller than $\mathrm{var}(\tilde{\rho})$ for any value of ρ except $\rho=0$. We also performed the simulation for values of $t=2,4,8$ and observed the same phenomenon. Fig. 1 illustrates the efficiency of $\hat{\rho}_{CL}$ with increasing p. For $p=6$ and 8, $S_{\hat{\rho}_{CL}}/S_{\tilde{\rho}}$ exhibits the same pattern as that at $p=3$; see Fig. 1.

4. Discussion

This paper sets out some issues in the study of robustness of composite likelihood inference; specifically emphasizing robustness of consistency. Robustness in inference usually means obtaining the same inferential result under a range of models. In point estimation the range of models is often considered to be small-probability perturbations of the assumed model, to reflect the sampling notion of occasional outliers.

In composite likelihood, the range of models is, loosely speaking, all models consistent with the specified set of sub-models $f_k(y \in A_k;\theta)$. For example if pairwise likelihood is used, the range of models is those consistent with the assumed bivariate distributions. In many, or even most, applications of composite likelihood, it is not immediately clear what that range of models looks like, and indeed whether there is even a single model compatible with the assumed sub-models.

The Wald assumptions set out in the Appendix are sufficient to ensure consistency of the MCLE, although they may be stronger than necessary. The most restrictive of these assumptions is (A7): that there exists a unique point $\theta^* \in \Omega$ that minimizes the Kullback–Leibler divergence (9). For each component likelihood the assumption that there is a unique $\theta_k^* \in \Omega_k$ would be more closely analogous to the usual Wald assumption for the MLE.

However, even in cases where both the MLE and the MCLE are not consistent, the MCLE might still be more reliable than the MLE, since mis-specifying a high dimensional complex joint density may be much more likely than mis-specifying some simpler lower dimensional densities.

The MCLE also has a type of robustness of efficiency. In computing the asymptotic variance, the composite likelihood is always treated as a "misspecified" model even if all component likelihoods are correctly specified. On the other hand, the inverse of the Fisher information matrix $I(\theta)=E\{-\ell''(\theta)\}$, which is used as the asymptotic variance of the MLE, is sensitive to model misspecification.

Composite likelihood also has a type of computational robustness, discussed in Varin et al. (2011); there is some evidence from applied work that the composite likelihood surface is smoother, and hence easier to maximize, than the likelihood surface.

There is also some evidence that composite likelihood inference is robust to missing data, although there is still much work to be done in this area. Recent papers discussing this include Yi et al. (2011), Molenberghs et al. (2011) and He and Yi (2011).

Acknowledgments

We are grateful to Professor Barry Arnold for suggesting a version of Example 4, and to Grace Yi, Keith Knight and Muni Srivastava for helpful suggestions. This research was partially supported by the Natural Sciences and Engineering Research Council of Canada.

Appendix A. Consistency of the MCLE

A.1. Introduction and assumptions

For analytical simplicity we only treat the composite marginal likelihoods with equal weights; however, the results obtained here should be easily generalized to more general situations.

Following Wald (1949) we introduce some notation for the needed assumptions. For any θ and for $\rho,r>0$ let $f(y;\theta,\rho)=\sup\{f(y;\theta'):\|\theta'-\theta\|\le\rho\}$, where $\|\cdot\|$ means Euclidean norm; $\varphi(y,r)=\sup\{f(y;\theta):\|\theta\|>r\}$; $f^*(y;\theta,\rho)=\max\{f(y;\theta,\rho),1\}$; $\varphi^*(y,r)=\max\{\varphi(y,r),1\}$.

For each $k\in(1,2,\ldots,K)$, we make the following assumptions, analogous to Assumptions 1–8, in Wald (1949):

(A0): The parameter space Ω is a closed subset of q-dimensional Cartesian space.
(A1): $f_k(Y_k;\theta,\rho)$ is a measurable function of Y_k for any θ and ρ.
(A2): The density function $f_k(Y_k;\theta)$ is distinct for different values of θ, i.e. if $\theta_1\neq\theta_2$ then $\mu[\{Y_k: f_k(Y_k;\theta_1)\neq f_k(Y_k;\theta_2)\}]>0$
(A3): For sufficiently small ρ and sufficiently large r, the expected values $\int\log f_k^*(Y_k;\theta,\rho)g_k(Y_k)d\mu(Y_k)$ and $\int\log\varphi_k^*(Y_k,r)g_k(Y_k)d\mu(Y_k)$ are finite.
(A4): For any $\theta\in\Omega$, there exist a set B_θ^k, such that $\int_{B_\theta^k}g_k(Y_k)d\mu(Y_k)=0$ and $f_k(Y_k;\theta')\to f_k(Y_k;\theta)$ as $\theta'\to\theta$ for $Y_k\in\overline{B_\theta^k}$ (the complement set of B_θ^k).
(A5): The expectation of $\log g_k(Y_k)$ exists.
(A6): There exists a set A_k, such that $\int_{A_k}g_k(Y_k)d\mu(Y_k)=0$ and $\lim_{\|\theta\|\to\infty}f_k(Y_k;\theta)=0$ for $Y_k\in\overline{A_k}$.
(A7): There exists a unique point $\theta^*\in\Omega$ which minimizes the composite Kullback–Leibler divergence defined in (9).

A.2. The main theorem

Theorem 1. *Assume that $Y^{(1)},\ldots,Y^{(n)}$ are independently and identically distributed with distribution function $G(Y)$. Under the regularity conditions* (A0)–(A7), *the maximum composite likelihood estimator $\hat\theta_{CL}$ converges almost surely to θ^* defined in* (9).

Before we prove Theorem 1, we state the following lemmas. By the expected value $E_g(\cdot)$, we shall mean the expected value determined under the true distribution $G(Y)$.

Lemma 1. *For any $\theta \neq \theta^*$, we have*

$$E_g\left\{\sum_{k=1}^{K} \log f_k(Y_k;\theta)\right\} < E_g\left\{\sum_{k=1}^{K} \log f_k(Y_k;\theta^*)\right\} \leq E_g\left\{\sum_{k=1}^{K} \log g_k(Y_k)\right\} \tag{14}$$

Lemma 2.

$$\lim_{\rho\to 0} E_g\left\{\sum_{k=1}^{K} \log f_k(Y_k;\theta,\rho)\right\} = E_g\left\{\sum_{k=1}^{K} \log f_k(Y_k;\theta)\right\} \tag{15}$$

Lemma 3.

$$\lim_{r\to\infty} E_g\left\{\sum_{k=1}^{K} \log \varphi_k(Y_k,r)\right\} = -\infty \tag{16}$$

The three Lemmas follow immediately from Assumption (A7) and Lemmas 1–3 in Wald (1949).

Proof of Theorem 1. First we shall prove that

$$\Pr\left\{\lim_{n\to\infty} \frac{\sup_{\theta\in\omega} \prod_{i=1}^{n}\prod_{k=1}^{K} f_k(Y_k^{(i)};\theta)}{\prod_{i=1}^{n}\prod_{k=1}^{K} f_k(Y_k^{(i)};\theta^*)} = 0\right\} = 1 \tag{17}$$

for any closed subset ω which belongs to Ω and does not contain θ^* defined in (A7).

From Lemma 3, for each i, we can choose $r_0 > 0$ such that

$$E_g\left\{\sum_{k=1}^{K} \log \varphi_k(Y_k^{(i)},r_0)\right\} < E_g\left\{\sum_{k=1}^{K} \log f_k(Y_k^{(i)};\theta^*)\right\} \tag{18}$$

Let $\omega_0 = \{\theta : \theta \in \omega \text{ and } \|\theta\| \leq r_0\} \subseteq \omega$. From Lemma 1 and 2, for each $\theta \in \omega_0$, we can find a ρ_θ such that

$$E_g\left\{\sum_{k=1}^{K} \log f_k(Y_k^{(i)};\theta,\rho_\theta)\right\} < E_g\left\{\sum_{k=1}^{K} \log f_k(Y_k^{(i)};\theta^*)\right\} \tag{19}$$

Since ω_0 is compact, by the finite-covering theorem there exists a finite number of points $\{\theta_1,\ldots,\theta_h\}$ in ω_0 such that $S(\theta_1,\rho_{\theta_1}) \cup \ldots \cup S(\theta_h,\rho_{\theta_h}) \supseteq \omega_0$, where $S(\theta,\rho)$ denotes the sphere with center θ and radius ρ. Clearly, we have

$$\sup_{\theta\in\omega} \prod_{i=1}^{n}\prod_{k=1}^{K} f_k(Y_k^{(i)};\theta) \leq \sum_{l=1}^{h}\left\{\prod_{i=1}^{n}\prod_{k=1}^{K} f_k(Y_k^{(i)};\theta_l,\rho_{\theta_l})\right\} + \prod_{i=1}^{n}\prod_{k=1}^{K} \varphi_k(Y_k^{(i)},r_0) \tag{20}$$

Hence (17) is proved if we can show that

$$\Pr\left\{\lim_{n\to\infty} \frac{\prod_{i=1}^{n}\prod_{k=1}^{K} f_k(Y_k^{(i)};\theta_l,\rho_{\theta_l})}{\prod_{i=1}^{n}\prod_{k=1}^{K} f_k(Y_k^{(j)};\theta^*)} = 0\right\} = 1, \quad (l=1,\ldots,h) \tag{21}$$

and

$$\Pr\left\{\lim_{n\to\infty} \frac{\prod_{i=1}^{n}\prod_{k=1}^{K} \varphi_k(Y_k^{(i)},r_0)}{\prod_{i=1}^{n}\prod_{k=1}^{K} f_k(Y_k^{(i)};\theta^*)} = 0\right\} = 1. \tag{22}$$

Proving the above two equations is equivalent to showing that for $l=1,\ldots,h$

$$\Pr\left\{\lim_{n\to\infty} \sum_{i=1}^{n}\left[\log\prod_{k=1}^{K} f_k(Y_k^{(i)};\theta_l,\rho_{\theta_l}) - \log\prod_{k=1}^{K} f_k(Y_k^{(i)};\theta^*)\right] = -\infty\right\} = 1 \tag{23}$$

and

$$\Pr\left\{\lim_{n\to\infty} \sum_{i=1}^{n}\left[\log\prod_{k=1}^{K} \varphi_k(Y_k^{(i)},r_0) - \log\prod_{k=1}^{K} f_k(Y_k^{(i)};\theta^*)\right] = -\infty\right\} = 1 \tag{24}$$

These equations follow immediately from (18) and (19) and the strong law of large numbers.

Let $\overline{\theta}_n(Y^{(1)},\ldots,Y^{(n)})$ be any function of the observations $Y^{(1)},\ldots,Y^{(n)}$ such that

$$\frac{\prod_{i=1}^{n}\prod_{k=1}^{K} f_k(Y_k^{(i)};\overline{\theta}_n)}{\prod_{i=1}^{n}\prod_{k=1}^{K} f_k(Y_k^{(i)};\theta^*)} \geq c > 0 \text{ for all } n \text{ and all } Y^{(1)},\ldots,Y^{(n)} \tag{25}$$

If we can show that

$$\Pr\Big\{\lim_{n\to\infty}\overline{\theta}_n=\theta^*\Big\}=1 \tag{26}$$

the proof of Theorem 1 is completed since the maximum composite estimator $\hat{\theta}_{CL}$ satisfies (25). To prove (26) it is sufficient to show that for any $\varepsilon>0$ the probability is one that all limit points $\overline{\theta}$ of the sequence $\{\overline{\theta}_n\}$ satisfy that $\|\overline{\theta}-\theta^*\|\le\varepsilon$. If there exists a limit point $\overline{\theta}_0$ such that $\|\overline{\theta}_0-\theta^*\|>\varepsilon$, we have

$$\sup_{\|\theta-\theta^*\|\ge\varepsilon}\prod_{i=1}^{n}\prod_{k=1}^{K}f_k(Y_k^{(i)};\theta)\ge\prod_{i=1}^{n}\prod_{k=1}^{K}f_k(Y_k^{(i)};\overline{\theta}_n)\ \text{for infinitely many } n \tag{27}$$

Then

$$\frac{\sup_{\|\theta-\theta^*\|\ge\varepsilon}\prod_{i=1}^{n}\prod_{k=1}^{K}f_k(Y_k^{(i)};\theta)}{\prod_{i=1}^{n}\prod_{k=1}^{K}f_k(Y_k^{(i)};\theta^*)}\ge c>0\ \text{for infinitely many } n \tag{28}$$

According our previous result (17) this is an event with probability zero. We have shown that the probability is one that all limit points $\overline{\theta}$ of the sequence $\{\overline{\theta}_n\}$ satisfy that $\|\overline{\theta}-\theta^*\|\le\varepsilon$. Thus Eq. (26) is obtained. □

Since the ordinary likelihood function is a special case of composite likelihood, the consistency of maximum likelihood estimator under a misspecified model (Theorem 2.2 in White, 1982) follows immediately from Theorem 1.

Corollary 1. *If the composite likelihood* (8) *is correctly specified, under the assumptions* (A0)–(A6), *the maximum composite likelihood estimator* $\hat{\theta}_{CL}$ *converges to the true parameter point* θ_0 *almost surely*.

References

Andrei, A., Kendziorski, C., 2009. An efficient method for identifying statistical interactors in gene association networks. Biostatistics 10, 706–718.

Arnold, B., 2010. Example of a non-normal distribution with normal marginals. Unpublished, personal communication.

He, W., Yi, G.Y., 2011. A pairwise likelihood method for correlated binary data with/without missing observations under generalized partially linear single-index models. Statist. Sinica 21, 207–229.

Jin, Z., 2009. Aspects of composite likelihood inference. Ph.D. Thesis, University of Toronto.

Joe, H., Lee, Y., 2009. On weighting of bivariate margins in pairwise likelihood. J. Multivariate Anal. 100, 670–685.

Kent, J.T., 1982. Robust properties of likelihood ratio tests. Biometrika 69, 19–27.

Lindsay, B.G., 1988. Composite likelihood methods. Contemp. Math. 80, 221–239.

Molenberghs, G., Verbeke, G., 2005. Models for Discrete Longitudinal Data. Springer, New York.

Molenberghs, G., Kenward, M., Verbeke, G., Berhanu, T., 2011. Pseudo-likelihood estimation for incomplete data. Statist. Sinica 21, 187–206.

Varin, C., Vidoni, P., 2005. A note on composite likelihood inference and model selection. Biometrika 92, 519–528.

Varin, C., 2008. On composite marginal likelihoods. Adv. Statist. Anal. 92, 1–28.

Varin, C., Reid, N., Firth, D., 2011. An overview of composite likelihood methods. Statist. Sinica 21, 5–42.

Wald, A., 1949. Note on the consistency of the maximum likelihood estimate. Ann. Math. Statist. 20, 595–601.

White, H., 1982. Maximum likelihood estimation of misspecified models. Econometrica 50, 1–25.

Yi, G.Y., Zeng, L., Cook, R.J., 2011. A robust pairwise likelihood method for incomplete longitudinal binary data arising in clusters. Canad. J. Statist. 39, 34–51.