Statistica Sinica Preprint No: SS-2022-0167	
Title	A Note on Information Bias and Efficiency of Composite Likelihood
Manuscript ID	SS-2022-0167
URL	http://www.stat.sinica.edu.tw/statistica/
DOI	10.5705/ss.202022.0167
Complete List of Authors	Libai Xu, Nancy Reid and Ximing Xu
Corresponding Authors	Ximing Xu
E-mails	ximing@nankai.edu.cn
Notice: Accepted version subject to English editing.	

Statistica Sinica

A note on information bias and efficiency of composite likelihood

Libai Xu, Nancy Reid and Ximing Xu

University of Toronto and Chongqing Medical University

Abstract: The properties of inference based on composite likelihood (CL) are well-established, but can be surprising, and intuition based on likelihood inference can be misleading. In this note, we show by example that the variance of a maximum composite likelihood estimator (MCLE) can increase when nuisance parameters are known, rather than estimated; that estimators based on more independent component likelihoods can be less efficient than those based on fewer; and that incorporating higher-dimensional marginal densities can also lead to less efficient inference. The role of information bias is highlighted to understand the occurrence of these paradoxical phenomena.

Key words and phrases: Bartlett's second identity, Estimating function, Godambe information matrix, Nuisance parameter, Pairwise likelihood.

1. Introduction

Suppose $\mathbf{y} = (y_1, ..., y_p)^{\mathrm{T}}$ is a p-dimensional random vector with probability density $f(\mathbf{y}; \theta)$, where θ is in a q-dimensional parameter space Θ. The CL

(Lindsay, 1988) is defined as $CL(\theta;\mathbf{y}) = \prod_{k=1}^{K} L_k(\theta;\mathbf{y})^{w_k}$, where the sub-likelihoods $L_k(\theta;\mathbf{y})$'s are usually the joint or conditional densities of some sub-vectors of $\mathbf{y}$; the weights w_k's could be positive or negative (Yi, 2014). Given n random samples, $\mathbf{y}^{(i)}, i = 1, ..., n$, the composite log-likelihood is $c\ell(\theta;\mathbf{y}) = \sum_{i=1}^{n} \log CL(\theta, \mathbf{y}^{(i)})$, and the MCLE is $\hat{\theta}_{CL} = \arg\max_\theta c\ell(\theta;\mathbf{y})$.

CLs lead to inference that is similar to that based on genuine likelihoods. Under some regularity conditions, $\hat{\theta}_{CL}$ is consistent, and asymptotically normally distributed with variance equal to the Godambe information matrix, $G(\theta) = H(\theta)J^{-1}(\theta)H(\theta)$ (Varin et al., 2011), where $H(\theta) = E\{-\nabla_\theta u_c(\theta;\mathbf{y})\}, J(\theta) = \text{var}\{u_c(\theta;\mathbf{y})\}$, with the composite score function $u_c(\theta;\mathbf{y}) = \nabla_\theta c\ell(\theta;\mathbf{y})$. However, there are aspects of inference based on CLs that are qualitatively different from inference based on the full likelihood. In this note we describe three such unexpected properties by examples that allow us to calculate the Godambe information or asymptotic variances analytically, and show how information bias plays a key role. Note that a CL is *information-unbiased* if $H(\theta) = J(\theta)$, and *information-biased* otherwise (Lindsay, 1982).

2. Three noticeable properties of CL with illustrative examples

2.1 An information-biased CL may lead to less efficient estimators of the parameters of interest when the nuisance parameters are known than when they are unknown and estimated. Suppose $\mathbf{y}^{(1)}, \ldots, \mathbf{y}^{(n)}$ are n independent observations from $N(0, \Sigma)$, where $\Sigma = \sigma^2\{(1-\rho)I_p + \rho J_p\}$, I_p is the $p \times p$ identity matrix and J_p is a $p \times p$ matrix with all entries equal to 1, with the parameter of interest $\rho \in [1/(1-p), 1]$ and a nuisance parameter $\sigma^2 > 0$.

When σ^2 is unknown, the maximum pairwise likelihood estimate (MPLE) $\hat{\rho}$ is identical to the MLE of ρ (Mardia et al., 2009), hence, fully efficient, with the asymptotic variance $\text{avar}(\hat{\rho}) = 2(1-\rho)^2\{1 + (p-1)\rho\}^2/\{np(p-1)\}$; when σ^2 is known, the MPLE $\tilde{\rho}$ is less efficient than the MLE of ρ (Cox and Reid, 2004). Comparing $\text{avar}(\tilde{\rho})$ and $\text{avar}(\hat{\rho})$,

$$r(\rho) = \text{avar}(\tilde{\rho})/\text{avar}(\hat{\rho}) = c(p, \rho)/[(1+\rho^2)^2\{1 + (p-1)\rho\}^2], \qquad (2.1)$$

where $c(p, \rho) = (1-\rho)^2(3\rho^2 + p^2\rho^2 + 1) - p\rho(3\rho^3 - 8\rho^2 + 3\rho - 2)$. The ratio $r(\rho)$, as a function of ρ, is plotted in S1 Figure 1 for $p = 3$. When ρ is positive, $\tilde{\rho}$ is more efficient than $\hat{\rho}$; when $\rho < 0$, the opposite direction is observed. We performed the comparisons for different p and observed the same phenomenon. It can be shown that the asymptotic covariance

between $\hat{\rho}$ and the MPLE $\hat{\sigma}^2$ is $2\rho(1-\rho)\{1+(p-1)\rho\}\sigma^2/(np)$ which goes to 0 as $\rho \to 1/(1-p)$ or 1, while the asymptotic covariance between $\tilde{\rho}$ and $\hat{\sigma}^2$ is not equal to zero at $\rho = 1/(1-p)$. This may explain why the paradox occurs when $\rho \to 1/(1-p)$ by Theorem 1 of Henmi and Eguchi (2004).

An information-biased CL may also lead to less efficient estimators by incorporating more independent CLs or by using component likelihoods with higher dimension, as seen in the following two subsections.

2.2 Information additivity may not hold for the product of independent information-biased CLs. Suppose the random vector $(Y_1, Y_2, Y_3)^{\mathrm{T}}$ follows a normal distribution $N(\mu, \Sigma)$, where $\Sigma = \mathrm{diag}(\Sigma_1, \sigma^2)$ and $\Sigma_1 = (1-\rho)I_2 + \rho J_2$. Assume that σ^2 is known, μ and ρ are unknown, and μ is the only parameter of interest. Consider the independence likelihood $CL_{12}(\mu) = f(y_1;\mu)f(y_2;\mu)$, which is free of the nuisance parameter ρ, and the CL, $CL_{123}(\mu) = CL_{12}(\mu)f(y_3;\mu)$, which incorporates the information from the independent variable Y_3, to estimate μ. Given a random sample of size n, the MCLEs from CL_{12} and CL_{123} are $\hat{\mu}_{12} = (\bar{y}_1 + \bar{y}_2)/2$ and $\hat{\mu}_{123} = \{\sigma^2(\bar{y}_1 + \bar{y}_2) + \bar{y}_3\}/(1+2\sigma^2)$, with variances $(1+\rho)/(2n)$ and $\{2(1+\rho)\sigma^4 + \sigma^2\}/\{n(1+2\sigma^2)^2\}$, respectively, where $\bar{y}_j = \sum_{i=1}^n y_j^{(i)}/n$ for $j = 1, 2, 3$.

We can compare the variances of the two MCLEs directly. For example,

when $\sigma^2 = 2$, the variance of $\hat{\mu}_{123}$ is $(10 + 8\rho)/(25n)$ which is smaller than $(1 + \rho)/(2n)$ if and only if $\rho > -5/9$. Note that if $\rho = -1$ this result is expected as (Y_1, Y_2) determines μ exactly with $\mu \equiv (Y_1 + Y_2)/2$; but the dependence on σ^2 of the range of ρ over which Y_3 degrades the inference is surprising; as σ^2 increases this range approaches $[-1, -1/2)$.

2.3 Pairwise likelihood may be less efficient than independence likelihood. Suppose $(Y_1, Y_2, Y_3, Y_4)^{\mathrm{T}}$ follows a Multinomial$(1; \theta, \theta, \theta/k, 1 - 2\theta - \theta/k)$, where $k > 0$ and $0 \leq \theta \leq k/(2k + 1)$. The parameter θ controls both the mean and covariance structures, and we can change the value of k to adjust the strength of dependence. Y_4 is completely determined by $1 - \sum_{i=1}^{3} Y_i$. We estimate θ based on the independent triplets $(y_1^{(i)}, y_2^{(i)}, y_3^{(i)})^{\mathrm{T}}$, $i = 1, \ldots, n$. Comparing the independence likelihood and the pairwise likelihood of all independent triplets, we can get the ratio of Godambe information

$$r(\theta) = G(\theta_{ind})/G(\theta_{pair}) = H_{ind}^2(\theta) J_{pair}(\theta) / \{H_{pair}^2(\theta) J_{ind}(\theta)\}. \qquad (2.2)$$

Detailed calculations of H_{ind}, J_{ind} and H_{pair}, J_{pair} are presented in the supplementary material S2. Particularly, for $k = 5$, the ratio as a function of θ is plotted in S1 Figure 2, and $\mathrm{r}(\theta) = 1$ has a solution $\theta = 1/3$. When $\theta < 1/3$, $r(\theta) < 1$ and when $\theta > 1/3$, $r(\theta) > 1$. Specifically, both the independence likelihood and the pairwise likelihood are fully efficient with

$k = 1$; when $k \to 0$, the pairwise likelihood is more efficient than the independence likelihood and $r(\theta) \to 1$; when $k \to \infty$, the independence likelihood is more efficient than the pairwise likelihood and $r(\theta) \to 1$.

3. Discussion

This note is meant to serve as a reminder that inference based on CL does need some care, beyond adjusting the variance of MCLEs or the limiting distribution of the CL ratio test. Another point worth remembering, although not emphasized here, is that CL based on the marginal density of components, such as the independence and pairwise CLs, may not be consistent with a unique multivariate distribution. An example of this was presented Yi (2014). In contrast, CL constructed from conditional distributions can rely on the Hammersley-Clifford theorem to ensure there is a unique joint distribution compatible with these conditional components (Besag, 1975).

4. Supplementary Materials

The supplementary material contains two Sections: S1 includes two figures for Examples 1 and 3; S2 presents detailed calculations of Example 3.

Acknowledgements

We are grateful to Professor Grace Yi for helpful comments on an earlier draft. This work was supported in part by the Natural Sciences and Engineering Research Council of Canada, and Chongqing Innovation Program for Returned Overseas Chinese Scholars (ex2021112).

References

Besag, J. (1975). Statistical analysis of non-lattice data. *Journal of the Royal Statistical Society: Series D (The Statistician) 24*(3), 179–195.

Cox, D. R. and N. Reid (2004). A note on pseudolikelihood constructed from marginal densities. *Biometrika 91*(3), 729–737.

Henmi, M. and S. Eguchi (2004). A paradox concerning nuisance parameters and projected estimating functions. *Biometrika 91*(4), 929–941.

Lindsay, B. (1982). Conditional score functions: some optimality results. *Biometrika 69*(3), 503–512.

Lindsay, B. G. (1988). Composite likelihood methods. *Contemporary mathematics 80*(1), 221–239.

Mardia, K. V., J. T. Kent, G. Hughes, and C. C. Taylor (2009). Maximum likelihood estimation using composite likelihoods for closed exponential families. *Biometrika 96*(4), 975–982.

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

Yi, G. Y. (2014). Composite likelihood/pseudolikelihood. *Wiley StatsRef: Statistics Reference Online*, 1–14.

Department of Statistical Sciences, University of Toronto, Toronto, Canada,
School of Mathematics and Statistics, Jiangsu Normal University, XuZhou, China
E-mail: libai.xu@utoronto.ca

REFERENCES

Department of Statistical Sciences, University of Toronto, Toronto, Canada,

E-mail: nancym.reid@utoronto.ca

Big Data Center for Childrens Medical Care, Childrens Hospital of Chongqing Medical University, Chongqing, China

National Clinical Research Center for Child Health and Disorders, Ministry of Education Key Laboratory of Child Development and Disorders, Chongqing, China

E-mail: ximing@hospital.cqmu.edu.cn