Statistica Sinica **34** (2024), 523-526
doi:https://doi.org/10.5705/ss.202022.0167

A NOTE ON INFORMATION BIAS AND EFFICIENCY OF COMPOSITE LIKELIHOOD

Libai Xu, Nancy Reid and Ximing Xu*

*Soochow University, University of Toronto
and Chongqing Medical University*

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

Key words and phrases: Bartlett's second identity, estimating function, godambe information matrix, nuisance parameter, pairwise likelihood.

1. Introduction

Suppose $\mathbf{y} = (y_1, \ldots, y_p)^{\mathrm{T}}$ is a p-dimensional random vector with probability density $f(\mathbf{y};\theta)$, where θ is in a q-dimensional parameter space Θ. The composite likelihood (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 $L_k(\theta;\mathbf{y})$ are usually the joint or conditional densities of some sub-vectors of $\mathbf{y}$, and the weights w_k can be positive or negative (Yi (2017)). Given n random samples, $\mathbf{y}^{(i)}$, for $i = 1, \ldots, n$, the composite log-likelihood is $c\ell(\theta;\mathbf{y}) = \sum_{i=1}^{n} \log CL(\theta, \mathbf{y}^{(i)})$, and the maximum composite likelihood estimator (MCLE) is $\hat{\theta}_{CL} = \mathrm{argmax}_{\theta} c\ell(\theta;\mathbf{y})$.

CLs lead to inferences similar to those 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, Reid and Firth (2011)), where $H(\theta) = E\{-\nabla_{\theta} u_c(\theta;\mathbf{y})\}$, $J(\theta) = \mathrm{var}\{u_c(\theta;\mathbf{y})\}$, with the composite score function $u_c(\theta;\mathbf{y}) = \nabla_{\theta} c\ell(\theta;\mathbf{y})$. However, several aspects of inferences based on CLs, are qualitatively different from those based on a full likelihood. In this note, we describe three such properties using examples that allow us to calculate the Godambe information or asymptotic variances analytically, and show how information bias plays a key role.

*Corresponding author.

Note that a CL is *information-unbiased* if $H(\theta) = J(\theta)$, and is *information-biased* otherwise (Lindsay (1982)).

2. Three properties of CLs

Property 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 $N(0, \Sigma)$. Here $\Sigma = \sigma^2\{(1-\rho)I_p + \rho J_p\}$, where I_p is the $p \times p$ identity matrix, J_p is a $p \times p$ matrix with all entries equal to one, the parameter of interest $\rho \in [1/(1-p), 1]$, and $\sigma^2 > 0$ is a nuisance parameter.

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, it is fully efficient, with 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 $\text{avar}(\hat{\rho})$, we have

$$r(\rho) = \frac{\text{avar}(\tilde{\rho})}{\text{avar}(\hat{\rho})} = \frac{c(p, \rho)}{(1+\rho^2)^2\{1+(p-1)\rho\}^2}, \tag{2.1}$$

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. Comparisons for different p yield the same results. 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 zero as $\rho \to 1/(1-p)$ or one, and 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 higher-dimensional component likelihoods.

Property 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 $\Sigma = \text{diag}(\Sigma_1, \sigma^2)$ 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 information from the independent variable Y_3 and is used 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 $\hat{\mu}_{123} = \{\sigma^2(\bar{y}_1 + \bar{y}_2) + \bar{y}_3\}/(1+2\sigma^2)$, respectively, 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, because (Y_1, Y_2) determines μ exactly, with $\mu \equiv (Y_1 + Y_2)/2$. However, 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 $[-1, -1/2)$.

Property 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 the covariance structures, and we can change the value of k to adjust the strength of the dependence. Here, Y_4 is determined completely by $1 - \sum_{i=1}^{3} Y_i$. We estimate θ based on the independent triplets $(y_1^{(i)}, y_2^{(i)}, y_3^{(i)})^{\mathrm{T}}$, for $i = 1, \ldots, n$. Comparing the independence likelihood and the pairwise likelihood of all independent triplets, we obtain the following ratio of Godambe information:

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

Detailed calculations of H_{ind} and J_{ind} and of H_{pair} and J_{pair} are presented in the Supplementary Material, Section S2. In particular, for $k = 5$, the ratio as a function of θ is plotted in Figure 2 in Section S1 of the Supplementary Material, and $r(\theta) = 1$ has a solution $\theta = 1/3$. When $\theta < 1/3$, then $r(\theta) < 1$, and when $\theta > 1/3$, then $r(\theta) > 1$. Specifically, both the independence likelihood and the pairwise likelihood are fully efficient when $k = 1$; 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 serves as a reminder that inferences based on a CL require care, beyond adjusting the variance of the MCLEs or the limiting distribution of the CL ratio test. Furthermore, we presented an example in which a CL based on the marginal density of the components, such as the independence and pairwise CLs, may not be consistent with a unique multivariate distribution Yi (2017). In contrast, for a CL constructed from a conditional distribution, the Hammersley–Clifford theorem ensures there is a unique joint distribution compatible with these conditional components (Besag (1975)).

Supplementary Material

The online Supplementary Material includes two figures for Examples 1 and 3, and detailed calculations for Example 3.

Acknowledgments

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**, 179–195.

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

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

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

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

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

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

Yi, G. Y. (2017). Composite likelihood/pseudolikelihood. In *Wiley StatsRef: Statistics Reference Online* (Edited by N. Balakrishnan, T. Colton, B. Everitt, W. Piegorsch, F. Ruggeri and J.L. Teugels). Web: `https://doi.org/10.1002/9781118445112.stat07855`.

Libai Xu

School of Mathematical Sciences, Soochow University, Suzhou, 215006, China.

E-mail: libai.xu@utoronto.ca

Nancy Reid

Department of Statistical Sciences, University of Toronto, Toronto, ON M5S 1A1, Canada.

E-mail: nancym.reid@utoronto.ca

Ximing Xu

Big Data Center for Children's Medical Care, Children's Hospital of Chongqing Medical University, Chongqing 400016, China.

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

(Received May 2022; accepted June 2022)