Universal inference with composite likelihoods

Hien D Nguyen*[1] Jessica Bagnall-Guerreiro[1] Andrew T Jones[1]

March 22, 2021

Abstract

Maximum composite likelihood estimation is a useful alternative to maximum likelihood estimation when data arise from data generating processes (DGPs) that do not admit tractable joint specification. We demonstrate that generic composite likelihoods consisting of marginal and conditional specifications permit the simple construction of composite likelihood ratio-like statistics from which finite-sample valid confidence sets and hypothesis tests can be constructed. These statistics are universal in the sense that they can be constructed from any estimator for the parameter of the underlying DGP. We demonstrate our methodology via a simulation study using a pair of conditionally specified bivariate models.

Key words: Composite likelihoods; Pseudolikelihoods, Confidence sets; Hypothesis tests; Conditional models

1 Introduction

Likelihood-based methods are among the most important tools for conducting statistical inference. However, data generating processes (DGPs) of complex models often do not admit tractable likelihood functions. In such cases, a potential remedy is to specify the model based on more amenable marginal and conditional probability density/mass functions (PDFs/PMFs) of the DGP, instead. This joint specification is often referred to as the composite likelihood (CL) or pseudolikelihood.

The literature regarding CL-based inference has its roots in the works of Besag [1975] and Lindsay [1988]. Further developments regarding the theory and application of CL methods can be found in Arnold and Strauss [1991], Molenberghs and Verbeke [2005], Varin et al. [2011], Yi [2014], and Nguyen [2018], among other works.

We build upon the recent work of Wasserman et al. [2020] who demonstrated the construction of sample splitting and sample swapping likelihood ratio statistics that yield finite-sample valid confidence sets and hypothesis tests, and are universal in the sense that they are agnostic to parameter estimators. The inferential constructions are similar to the recently popularized e-values of Vovk and Wang [2021], as well as the s-values of Grunwald et al. [2020] and the betting scores of Shafer [2021]. We demonstrate how our CL-based methods can be used via applications to constructing confidence sets and tests for a pair of conditionally specified bivariate models. Here, we consider simulations study regarding the exponential conditional model of Arnold et al. [1999] and the log-normal conditional model of Sarabia et al. [2007].

The paper proceeds as follows. In Section 2, we present the CL framework and the universal confidence set and hypothesis test constructions. A simulation study of our methodology is presented in Section 3.

2 Universal inference via composite likelihoods

Let $\boldsymbol{X} \in \mathbb{X} \subseteq \mathbb{R}^d$ be a random variable arising from a parametric distribution characterized by the PDF/PMF (generically, PDF) $p(\boldsymbol{x}; \boldsymbol{\theta})$, where $\boldsymbol{\theta} \in \Theta \subseteq \mathbb{R}^q$ is a parameter vector ($d, q \in \mathbb{N}$). We shall write $\boldsymbol{X}^\top = (X_1, \dots, X_d)$ to indicate a random variable and $\boldsymbol{x}^\top = (x_1, \dots, x_d)$ to indicate its realization.

Let $2^{[d]}$ be the power set of $[d] = \{1, \dots, d\}$, and let $\mathbb{S}_d = 2^{[d]} \setminus \{\varnothing\}$. For each $\mathcal{S} \in \mathbb{S}_d$, let $\mathcal{S} = \left\{s_1, \dots, s_{|\mathcal{S}|}\right\} \subseteq [d]$, where $|\mathcal{S}|$ is the cardinality of $\mathcal{S}$. Further, let $\mathbb{T}_d$ be the set of all divisions of $[d]$ into two nonempty subsets. We represent each element of $\mathbb{T}_d$ as a pair $\mathcal{T} = \left(\overleftarrow{\mathcal{T}}, \overrightarrow{\mathcal{T}}\right)$, where $\overleftarrow{\mathcal{T}} = \left\{\overleftarrow{t}_1, \dots, \overleftarrow{t}_{|\overleftarrow{\mathcal{T}}|}\right\} \subset [d]$ and $\overrightarrow{\mathcal{T}} = \left\{\overrightarrow{t}_1, \dots, \overrightarrow{t}_{|\overrightarrow{\mathcal{T}}|}\right\} \subset [d] \setminus \overleftarrow{\mathcal{T}}$ are the 'left-hand' and 'right-hand' subsets of the division $\mathcal{T}$, respectively. We note that $|\mathbb{S}_d| = 2^d - 1$ and $|\mathbb{T}_d| = 3^d - 2^{d+1} + 1$.

*Corresponding author—Email: h.nguyen5@latrobe.edu.au. [1]Department of Mathematics and Statistics, La Trobe University, Melbourne, Australia.

For each $\mathcal{S}$ and $\mathcal{T}$, we assign a non-negative coefficient $\sigma_{\mathcal{S}}$ and $\tau_{\mathcal{T}}$, respectively. We call these coefficients the weights, and we put these weights into the vectors $\boldsymbol{\sigma} = (\sigma_{\mathcal{S}})_{\mathcal{S}\in\mathbb{S}_d}$ and $\boldsymbol{\tau} = (\tau_{\mathcal{T}})_{\mathcal{T}\in\mathbb{T}_d}$, respectively. Assume that $\upsilon = \sum_{\mathcal{S}\in\mathbb{S}_d}\sigma_{\mathcal{S}} + \sum_{\mathcal{T}\in\mathbb{T}_d}\tau_{\mathcal{T}} > 0$.

Given weights $\boldsymbol{\sigma}$ and $\boldsymbol{\tau}$, we define the individual CL (ICL) function for $\boldsymbol{X}$ as

$$p_{\boldsymbol{\sigma},\boldsymbol{\tau}}(\boldsymbol{x};\boldsymbol{\theta}) = \prod_{\mathcal{S}\in\mathbb{S}_d}\left[p(\boldsymbol{x}_{\mathcal{S}};\boldsymbol{\theta})\right]^{\sigma_{\mathcal{S}}/\upsilon}\prod_{\mathcal{T}\in\mathbb{T}_d}\left[p\left(\boldsymbol{x}_{\overleftarrow{\mathcal{T}}}|\boldsymbol{x}_{\overrightarrow{\mathcal{T}}};\boldsymbol{\theta}\right)\right]^{\tau_{\mathcal{T}}/\upsilon},$$

where $\boldsymbol{x}_{\mathcal{S}}^\top = \left(x_{s_1},\dots,x_{s_{|\mathcal{S}|}}\right)$, $\boldsymbol{x}_{\overleftarrow{\mathcal{T}}} = \left(x_{\overleftarrow{t}_1},\dots,x_{\overleftarrow{t}_{|\overleftarrow{\mathcal{T}}|}}\right)$, and $\boldsymbol{x}_{\overrightarrow{\mathcal{T}}} = \left(x_{\overrightarrow{t}_1},\dots,x_{\overrightarrow{t}_{|\overrightarrow{\mathcal{T}}|}}\right)$. Here, $p(\boldsymbol{x}_{\mathcal{S}};\boldsymbol{\theta})$ is the marginal PDF of $\boldsymbol{X}_{\mathcal{S}}$, and $p\left(\boldsymbol{x}_{\overleftarrow{\mathcal{T}}}|\boldsymbol{x}_{\overrightarrow{\mathcal{T}}};\boldsymbol{\theta}\right)$ is the conditional PDF of $\boldsymbol{X}_{\overleftarrow{\mathcal{T}}}$ conditioned on $\boldsymbol{X}_{\overrightarrow{\mathcal{T}}} = \boldsymbol{x}_{\overrightarrow{\mathcal{T}}}$.

2.1 Sample splitting and sample swapping

Let $\mathbf{X}_n = (\boldsymbol{X}_i)_{i=1}^n$ be a sequence of n IID random variables with the same DGP as $\boldsymbol{X}$, and split $\mathbf{X}_n$ into two subsamples $\mathbf{X}_n^1 = \left(\boldsymbol{X}_i^1\right)_{i=1}^{n_1}$ and $\mathbf{X}_n^2 = \left(\boldsymbol{X}_i^2\right)_{i=1}^{n_2}$ of sizes n_1 and n_2, respectively, where $n = n_1 + n_2$. We assume that $\boldsymbol{X}$ has a DGP that is characterized by the PDF $p(\boldsymbol{x};\boldsymbol{\theta}_0)$, for some $\boldsymbol{\theta}_0\in\Theta$, and we let $\Pr_{\boldsymbol{\theta}_0}$ be its corresponding probability measure. Let $\hat{\boldsymbol{\theta}}_n^1$ and $\hat{\boldsymbol{\theta}}_n^2$ be a pair of generic estimators of $\boldsymbol{\theta}_0$, computed using only $\mathbf{X}_n^1$ or $\mathbf{X}_n^2$, respectively.

For $k\in\{1,2\}$, we let

$$L_{\boldsymbol{\sigma},\boldsymbol{\tau}}\left(\boldsymbol{\theta};\mathbf{X}_n^k\right) = \prod_{i=1}^{n_k} p_{\boldsymbol{\sigma},\boldsymbol{\tau}}\left(\boldsymbol{X}_i^k\right)$$

be the CL function of $\mathbf{X}_n^k$, as a function of $\boldsymbol{\theta}$. We write the split sample CL ratio statistics (spCLRSs) and the swapped sample CL ratio statistic (swCLRS) as

$$U_{\boldsymbol{\sigma},\boldsymbol{\tau}}^k(\boldsymbol{\theta};\mathbf{X}_n) = L_{\boldsymbol{\sigma},\boldsymbol{\tau}}\left(\hat{\boldsymbol{\theta}}^{3-k};\mathbf{X}_n^k\right)/L_{\boldsymbol{\sigma},\boldsymbol{\tau}}\left(\boldsymbol{\theta};\mathbf{X}_n^k\right),$$

for each $k\in\{1,2\}$, and

$$\bar{U}_{\boldsymbol{\sigma},\boldsymbol{\tau}}(\boldsymbol{\theta};\mathbf{X}_n) = \left\{U_{\boldsymbol{\sigma},\boldsymbol{\tau}}^1(\boldsymbol{\theta};\mathbf{X}_n) + U_{\boldsymbol{\sigma},\boldsymbol{\tau}}^2(\boldsymbol{\theta};\mathbf{X}_n)\right\}/2,$$

respectively.

For $\alpha\in(0,1)$, let

$$\mathcal{C}^\alpha(\mathbf{X}_n) = \left\{\boldsymbol{\theta}\in\Theta : U_{\boldsymbol{\sigma},\boldsymbol{\tau}}^1(\boldsymbol{\theta};\mathbf{X}_n)\le 1/\alpha\right\} \text{ and } \bar{\mathcal{C}}^\alpha(\mathbf{X}_n) = \left\{\boldsymbol{\theta}\in\Theta : \bar{U}_{\boldsymbol{\sigma},\boldsymbol{\tau}}(\boldsymbol{\theta};\mathbf{X}_n)\le 1/\alpha\right\}$$

be confidence sets based on the spCLRS and the swCLRS, respectively. We have the following result regarding the validity of $\mathcal{C}^\alpha(\mathbf{X}_n)$ and $\bar{\mathcal{C}}^\alpha(\mathbf{X}_n)$ (all theoretical results in this work are proved in Nguyen, 2020).

Proposition 1. *The set estimators $\mathcal{C}^\alpha(\mathbf{X}_n)$ and $\bar{\mathcal{C}}^\alpha(\mathbf{X}_n)$ are finite sample-valid $100(1-\alpha)\%$ confidence sets for $\boldsymbol{\theta}_0$ in the sense that*

$$\Pr_{\boldsymbol{\theta}_0}(\boldsymbol{\theta}_0\in\mathcal{C}^\alpha(\mathbf{X}_n))\ge 1-\alpha, \text{ and } \Pr_{\boldsymbol{\theta}_0}\left(\boldsymbol{\theta}_0\in\bar{\mathcal{C}}^\alpha(\mathbf{X}_n)\right)\ge 1-\alpha,$$

for any $n\in\mathbb{N}$.

We now consider the testing of null and alternative hypotheses

$$\mathrm{H}_0 : \boldsymbol{\theta}\in\Theta_0 \text{ and } \mathrm{H}_1 : \boldsymbol{\theta}\in\Theta_1,$$

where $\Theta_0,\Theta_1\subseteq\Theta$. Let

$$\mathbb{M}\left(\mathbf{X}_n^k\right) = \left\{\boldsymbol{\theta}\in\Theta_0 : L_{\boldsymbol{\sigma},\boldsymbol{\tau}}\left(\boldsymbol{\theta};\mathbf{X}_n^k\right) = \max_{\boldsymbol{\vartheta}\in\Theta_0} L_{\boldsymbol{\sigma},\boldsymbol{\tau}}\left(\boldsymbol{\vartheta};\mathbf{X}_n^k\right)\right\}$$

be the set of maximizers of the CL function $L_{\boldsymbol{\sigma},\boldsymbol{\tau}}\left(\boldsymbol{\theta};\mathbf{X}_n^k\right)$, for each $k\in\{1,2\}$, and write $\tilde{\boldsymbol{\theta}}_n^k\in\mathbb{M}\left(\mathbf{X}_n^k\right)$. We then write the sample splitting and sample swapping test statistics as

$$V_{\boldsymbol{\sigma},\boldsymbol{\tau}}^k(\mathbf{X}_n) = U_{\boldsymbol{\sigma},\boldsymbol{\tau}}^k\left(\tilde{\boldsymbol{\theta}}_n^k\right), \text{ and } \bar{V}_{\boldsymbol{\sigma},\boldsymbol{\tau}}(\mathbf{X}_n) = \left\{U_{\boldsymbol{\sigma},\boldsymbol{\tau}}^1\left(\tilde{\boldsymbol{\theta}}_n^1\right) + U_{\boldsymbol{\sigma},\boldsymbol{\tau}}^2\left(\tilde{\boldsymbol{\theta}}_n^2\right)\right\}/2,$$

respectively. Further, define the split sample CL ratio test (spCLRT) and the swapped sample CL ratio test (swCLRT) by the rejection rules: reject H_0 if $V_{\boldsymbol{\sigma},\boldsymbol{\tau}}^1(\mathbf{X}_n)\ge 1/\alpha$ or if $\bar{V}_{\boldsymbol{\sigma},\boldsymbol{\tau}}(\mathbf{X}_n)\ge 1/\alpha$, respectively. We have the following result regarding the finite sample-validity of the tests.

Proposition 2. *The spCLRT and swCLRT control the Type I error for all $\alpha\in(0,1)$ and $n\in\mathbb{N}$ in the sense that*

$$\sup_{\boldsymbol{\theta}_0\in\Theta_0}\Pr_{\boldsymbol{\theta}_0}\left(V_{\boldsymbol{\sigma},\boldsymbol{\tau}}^1(\mathbf{X}_n) > 1/\alpha\right)\le\alpha, \text{ and } \sup_{\boldsymbol{\theta}_0\in\Theta_0}\Pr_{\boldsymbol{\theta}_0}\left(\bar{V}_{\boldsymbol{\sigma},\boldsymbol{\tau}}(\mathbf{X}_n) > 1/\alpha\right)\le\alpha.$$

Table 1: Simulation results.

(a) CP and AS results for the spCLRS and swCLRS 95% confidence sets.

		CP		n_1	AS		n_1
	θ_0	100	1000	10000	100	1000	10000
spCLRS	1	1	1	1	1.43	0.46	0.14
	5	1	1	1	4.60	1.49	0.47
	10	1	1	1	8.32	2.57	0.82
swCLRS	1	1	1	1	1.28	0.40	0.12
	5	1	0.99	1	4.13	1.29	0.40
	10	1	1	1	7.40	2.31	0.73

(b) Proportion of rejections by the spCLRT and swCLRT.

		Rej.		n_1
	c_0	100	1000	10000
spCLRT	0	0	0	0
	1	0.26	1	1
	5	0.98	1	1
swCLRT	0	0	0	0
	1	0.32	1	1
	5	1	1	1

3 Simulation study

All numerical computation were conducted in the R programming environment [R Core Team, 2020]. The code for the analyses are made available at `hiendn/CompositeLikelihoodISI`.

3.1 Bivariate distribution with exponential conditional distributions

We first consider a simulation study regarding data generated from the bivariate exponential distribution of Arnold et al. [1999, Sec. 4.4]. Here the random variable $\boldsymbol{X}^\top = (X_1, X_2)$ has joint PDF

$$p(\boldsymbol{x};\theta) = \kappa(\theta)\exp\{-x_1 - x_2 - \theta x_1 x_2\},$$

where $\theta \geq 0$ is the parameter of interest, and $\kappa(\theta) = \theta\exp\{-1/\theta\} / \int_{1/\theta}^{\infty} w^{-1}\exp(-w)\,\mathrm{d}w$ is an intractable normalization constant. However, the conditional PDFs of $X_k | X_{3-k} = x_{3-k}$, for $k \in \{1, 2\}$, can be specified by

$$p(x_k | x_{3-k};\theta) = f_{\mathrm{Exp}}(x_k; 1 + \theta x_{3-k}),$$

where $f_{\mathrm{Exp}}(x;\lambda) = \lambda\exp(-\lambda x)$ is the PDF of the exponential distribution with rate $\lambda > 0$. Thus, we can conduct inference regarding this DGP by considering ICLs of the form

$$p_{\boldsymbol{\sigma},\boldsymbol{\tau}}(\boldsymbol{x};\theta) = [p(x_1|x_2;\theta)]^{1/2}[p(x_2|x_1;\theta)]^{1/2},$$

where $\boldsymbol{\sigma} = \mathbf{0}$ and $\boldsymbol{\tau} = (1/2)\,\mathbf{1}$.

For data $\mathbf{X}_n$ with identical DGP to $\boldsymbol{X}$, characterized by $\theta_0 \in \{1, 5, 10\}$, where $n_1 = n_2 \in \{100, 1000, 10000\}$, we consider the use of the spCLRS and swCLRS confidence sets at the $\alpha = 0.05$ level. Here, each confidence set is constructed using the maximum composite likelihood estimator (MCLE).

For each pair (n_1, θ), we replicate the simulation $r = 100$ times and compute the coverage proportion (CP) and average size (AS) of the confidence intervals for the two set constructions. Here, CP and AS are computed as $r^{-1}\sum_{j=1}^{r} [\![\theta_0 \in \mathcal{C}_j]\!]$ and $r^{-1}\sum_{j=1}^{r} \mathrm{diam}(\mathcal{C}_j)$, where $\mathcal{C}_j$ is a stand-in for a confidence set constructed from the rth replicate, $[\![\cdot]\!]$ are Iverson brackets, and $\mathrm{diam}(\cdot)$ is the metric set diameter operator.

The results are presented in Table 1(a). We observe that CP was near perfect, with only one scenario yielding a confidence set that did not contain θ_0. This supports Proposition 1, although it indicates that the confidence sets are fairly conservative. We observe that AS is decreasing in n_1, as expected, and increasing in θ_0. We also find that the swCLRS sets are smaller than the spCLRS sets, which suggests a more efficient use of the data.

3.2 Bivariate distribution with log-normal conditional distributions

We now consider the bivariate distribution of Sarabia et al. [2007], which is specified by the PDF

$$p(\boldsymbol{x};\boldsymbol{\theta}) = \frac{\kappa(c)}{2\pi\sigma_1\sigma_2 x_1 x_2}\exp\left\{-\frac{1}{2}\left[\left(\frac{\log x_1 - \mu_1}{\sigma_1}\right)^2 + \left(\frac{\log x_2 - \mu_2}{\sigma_2}\right)^2 + c\left(\frac{\log x_1 - \mu_1}{\sigma_1}\right)^2\left(\frac{\log x_2 - \mu_2}{\sigma_2}\right)^2\right]\right\}, \quad (1)$$

where $\boldsymbol{\theta}^\top = (\mu_1, \sigma_1^2, \mu_2, \sigma_2^2, c)$, with $\mu_1, \mu_2 \in \mathbb{R}$, $\sigma_1^2, \sigma_2^2 > 0$, and $c \geq 0$. Here, $\kappa(c) = \sqrt{2c}/U\left(1/2, 1, (2c)^{-1}\right)$, where $U(a, b, z)$ is the confluence hypergeometric function, defined as per Abramowitz and Stegun [1972, Eqn. 13.2.5].

Like in the previous example, the normalizing constant of the joint PDF makes it intractable. However, we may again specify the conditional PDFs of $X_k|X_{3-k} = x_{3-k}$, for $k \in \{1, 2\}$, by

$$p(x_k|x_{3-k};\boldsymbol{\theta}) = f_{\mathrm{LN}}\left(x_k;\mu_k,\sigma_k^2/\left\{1+c\left(\frac{\log x_{3-k}-\mu_{3-k}}{\sigma_{3-k}}\right)^2\right\}\right),$$

where

$$f_{\mathrm{LN}}\left(x;\mu,\sigma^2\right) = \frac{1}{x\sqrt{2\pi\sigma^2}}\exp\left\{-\frac{1}{2}\left(\frac{\log x-\mu}{\sigma}\right)^2\right\}$$

is the PDF of a log-normal distribution with location and scale parameters $\mu \in \mathbb{R}$ and $\sigma^2 > 0$, respectively. We can use the conditional PDFs to conduct CL inference via the ICLs of the form

$$p_{\boldsymbol{\sigma},\boldsymbol{\tau}}(\boldsymbol{x};\boldsymbol{\theta}) = \left[p(x_1|x_2;\boldsymbol{\theta})\right]^{1/2}\left[p(x_2|x_1;\boldsymbol{\theta})\right]^{1/2},$$

where $\boldsymbol{\sigma} = \mathbf{0}$ and $\boldsymbol{\tau} = (1/2)\,\mathbf{1}$.

We simulate data $\mathbf{X}_n$, $n_1 = n_2 \in \{100, 1000, 10000\}$ from DGPs that are characterized by the PDF (1), with parameter vector $\boldsymbol{\theta}_0 = (2, 1, 2, 1, c_0)$, where $c_0 \in \{0, 1, 5\}$. For each pair (n_1, c_0), we use the spCLRT and swCLRT to test the hypotheses $\mathrm{H}_0 : c_0 = 0$ versus $\mathrm{H}_1 : c_0 > 0$, at the $\alpha = 0.05$ level. We repeat each simulation pair $r = 100$ times and compute the proportion of times the null hypothesis was rejected. Here, we again make use of the MCLE.

The results are reported in Table 1(b). Notice that no false rejections were made when $c_0 = 0$, thus the size of the test is conservatively controlled, as predicted by Proposition 2. We also see that the tests become increasingly powerful as c_0 increases and as n_1 increases, as would be expected. There is some evidence that the swCLRT is more powerful than the spCLRT, conforming to observations from the previous study.

References

M Abramowitz and I A Stegun, editors. *Handbook of Mathematical Functions with Formulas, Graphs, and Mathematical Tables.* National Bureau of Standards, Washington, 1972.

B C Arnold and D Strauss. Pseudolikelihood estimation: some examples. *Sankhya B*, 53:233–243, 1991.

B C Arnold, E Castillo, and J M Sarabia. *Conditional Specification of Statistical Models.* Springer, New York, 1999.

J Besag. Statistical analysis of non-lattice data. *Journal of the Royal Statistical Society D*, 24:179–195, 1975.

P Grunwald, R de Heide, and W M Koolen. Safe testing. In *IEEE Information Theory and Applications Workshop (ITA)*, 2020.

B Lindsay. Composite likelihood methods. *Contemporary Mathematics*, 8:221–239, 1988.

G Molenberghs and G Verbeke. *Models For Discrete Longitudinal Data.* Springer, New York, 2005.

H D Nguyen. Nearly universal consistency of maximum likelihood in discrete models. *Journal of the Korean Statistical Society*, 47:90–98, 2018.

H D Nguyen. Universal inference with composite likelihoods. *ArXiv:2009.00848*, 2020.

R Core Team. *R: a language and environment for statistical computing.* R Foundation for Statistical Computing, 2020.

J M Sarabia, E Castillo, M Pascual, and M Sarabia. Bivariate income distributions with lognormal conditionals. *Journal of Economic Inequality*, 5:371–383, 2007.

G Shafer. Testing by betting: a strategy for statistical and scientific communication. *Journal of the Royal Statistical Society B*, To appear, 2021.

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

V Vovk and R Wang. E-values: calibration, combination, and application. *Annals of Statistics*, To appear, 2021.

L Wasserman, A Ramdas, and S Balakrishnan. Universal inference. *Proceedings of the National Academy of Sciences*, 117:16880–16890, 2020.

G Yi. Composite likelihood/pseudolikelihood. In *Wiley StatsRef: Statistics Reference Online*, pages 1–14. Wiley, 2014.