Efficient composite likelihood for a scalar parameter of interest

Published version:
DOI: 10.1002/sta4.222

Received: Added at production | Revised: Added at production | Accepted: Added at production

DOI: xxx/xxxx

ORIGINAL ARTICLE

Efficient composite likelihood for a scalar parameter of interest

Luigi Pace*[1] | Alessandra Salvan[2] | Nicola Sartori[2]

[1]Department of Economics and Statistics, University of Udine, Italy

[2]Department of Statistical Sciences, University of Padova, Italy

Correspondence

*Luigi Pace, Department of Economics and Statistics, University of Udine, Via Tomadini 30/A, 33100 Udine, Italy. Email: luigi.pace@uniud.it

Funding Information

This research was supported by the Italian Ministry of Education PRIN 2015 grant 2015EASZFS_003, by the University of Udine grant PRID2017_DIES004 and by the University of Padova grant BIRD185955.

Summary

For inference in complex models, composite likelihood combines genuine likelihoods based on low-dimensional portions of the data, with weights to be chosen. Optimal weights in composite likelihood may be searched following different routes, leading to a solution only in scalar parameter models. Here, after briefly reviewing the main approaches, we show how to obtain first order optimal weights when using composite likelihood for inference on a scalar parameter in the presence of nuisance parameters. These weights depend on the true parameter value and need to be estimated. Under regularity conditions, the resulting likelihood ratio statistic has the standard asymptotic null distribution and improved local power. Simulation results in multivariate normal models show that estimation of optimal weights maintains the standard approximate null distribution and produces a visible gain in power with respect to constant weights.

KEYWORDS:

Asymptotic efficiency, composite likelihood, log likelihood ratio statistic, multivariate normal, nuisance parameter.

1 | INTRODUCTION

Complex models for high-dimensional data often entail that the full likelihood is difficult to specify or computationally intractable. In these cases, inference can be based on possibly dependent genuine log likelihoods from selected low-dimensional portions of the data. Composite log likelihood (Lindsay 1988; Varin et al. 2011) combines these individual log likelihoods through a linear combination with suitable weights.

In the literature, several contributions have considered the choice of weights for a given set of log likelihoods to be combined. Although a popular and straightforward option, equal weights may be inefficient. Unequal weights have been suggested aiming at computational and/or statistical efficiency. Zero-one weights defined through an efficient tapering strategy have been extensively studied for spatial models, see Sang & Genton (2014, Section 3.2) for a review, and also Castruccio et al. (2016).

For more general models, optimal weights have been investigated following two different routes. The first route exploits the theory of optimal combination of estimating equations, as in Lindsay (1988, Section 4), McCullagh & Nelder (1989, Section 9.4.2), Heyde (1997, Chapter 6). Typically, the weights in the optimal estimating equation are a function of the parameter, and therefore the corresponding optimal estimating function is not the score of a composite log likelihood. See Kuk (2007) and Deng et al. (2014) for optimal estimating equations from composite likelihood. See also Li & Sang (2018). A second route is proposed in Fraser & Reid (2018), where optimal weights are obtained for composite likelihood for a scalar parameter, based on a first order asymptotic analysis using score variables. These optimal weights depend on a reference parameter value, which in practice has to be estimated.

In this paper, after a brief review of optimal combination of estimating functions and log likelihoods, we focus on inference on a scalar parameter of interest in the presence of nuisance parameters. We obtain expressions for first order optimal weights when combining individual profile log likelihoods. For inference based on the composite likelihood ratio statistic with estimated optimal weights, we give examples of the optimal combination in multivariate normal models with common marginal mean of interest and various correlation structures depending on nuisance parameters.

Simulation results show that use of estimated optimal weights instead of equal weights maintains the standard approximate null distribution and produces a visible gain in power.

2 | OPTIMAL COMBINATION OF ESTIMATING FUNCTIONS AND LOG LIKELIHOODS

Let y be the data modelled as a realization of the random variable Y with density $p_Y(y;\theta), \theta \in \Theta \subseteq \mathbb{R}^p$. The full log likelihood is $\ell(\theta) = \log p_Y(y;\theta)$, with score vector $u(\theta) = (\partial/\partial\theta)\ell(\theta)$. Let $\ell_j(\theta)$, $j = 1, \ldots, q$, be possibly dependent genuine log likelihoods based on low-dimensional portions of the data. A composite log likelihood (Lindsay 1988; Varin et al. 2011) is

$$\ell_C(\theta) = \sum_{j=1}^{q} w_j \ell_j(\theta)\,, \tag{1}$$

where $w = (w_1, \ldots, w_q)^\top$ is a vector of weights to be chosen. Equal weights, $w_j = 1, j = 1, \ldots, q$, give

$$\ell_I(\theta) = \sum_{j=1}^{q} \ell_j(\theta)\,, \tag{2}$$

the independence combination. This is the optimal combination of the $\ell_j(\theta)$'s when sources are independent.

Weights in (1) are assumed to be non-negative in most of the literature, as in Varin et al. (2011). This restriction guarantees that the Kullback–Leibler information inequality propagates from each individual $\ell_j(\theta)$ to $\ell_C(\theta)$. See Lindsay et al. (2011, Section 4.7) for other reasons in favour of non-negative weights. Use of possibly negative weights is explored in Harden (2013, Section 3.2). For what follows, the assumption $w_j \geq 0$, $j = 1, \ldots, q$, is unnecessarily restrictive.

Let I_d denote the identity matrix of order d. Notation $x = [x_j]$ is used for a vector with entries x_j. Similarly, $A = [a_{jk}]$ is a matrix with entries a_{jk}. For a square matrix $A = [a_{jk}]$, we write $x = \mathrm{diag}(A) = [a_{jj}]$. Conversely, $A = \mathrm{diag}(x)$ is the diagonal matrix having x as its main diagonal.

The score function of $\ell_j(\theta)$, $j = 1, \ldots, q$, is denoted by $u_j(\theta) = (\partial/\partial\theta)\ell_j(\theta) = [u_{jr}(\theta)] = [(\partial/\partial\theta_r)\ell_j(\theta)]$, $r = 1, \ldots, p$. The corresponding Fisher information is $I_{jj}(\theta) = \mathrm{Var}_\theta\{u_j(\theta)\} = E_\theta\{-(\partial/\partial\theta^\top)u_j(\theta)\}$, assumed to be of order $O(n)$. For the last equality, we require that each $\ell_j(\theta)$ satisfies the first and the second Bartlett identities.

Let $u_V(\theta)$ be the qp-dimensional vector obtained by concatenating the score vectors $u_j(\theta)$. The covariance matrix of $u_V(\theta)$ is denoted by $\Sigma_V(\theta)$. It is a square matrix of order pq with diagonal blocks $I_{jj}(\theta)$, and off-diagonal blocks $I_{jk}(\theta) = \mathrm{Cov}_\theta(u_j(\theta), u_k(\theta))$, $j, k = 1, \ldots, q$, accounting for correlation between sources. We assume that $\Sigma_V(\theta)$ has full rank. Moreover, let $i_V(\theta)$ be the (qp, p) matrix obtained by stacking the information matrices $I_{jj}(\theta)$. Under regularity conditions, $i_V(\theta) = E_\theta\{u_V(\theta)u(\theta)^\top\}$ (see e.g. Jørgensen & Knudsen 2004, formula (1)).

For an unbiased estimating function $q(\theta) = q(\theta; y)$, i.e. a function such that $E_\theta\{q(\theta; Y)\} = 0$, the sensitivity and variability matrices are $H(\theta) = E_\theta\{-(\partial/\partial\theta^\top)q(\theta; Y)\}$ and $J(\theta) = \mathrm{Var}_\theta\{q(\theta; Y)\}$, respectively. Godambe information is $G(\theta) = H(\theta)^\top J(\theta)^{-1} H(\theta)$. Under regularity conditions, $G(\theta)^{-1}$ is the asymptotic covariance matrix of the estimator defined as a consistent root of $q(\theta) = 0$. If $H(\theta) = J(\theta)$, $G(\theta)^{-1} = H(\theta)^{-1}$ and $q(\theta)$ is called information unbiased (Jørgensen & Knudsen 2004; Lindsay 1982).

The estimating function from (1), $u_C(\theta) = \sum_{j=1}^{q} w_j u_j(\theta)$, is unbiased. Writing

$$u_C(\theta) = (w_1 I_p, \ldots, w_q I_p)\, u_V(\theta)\,, \tag{3}$$

sensitivity and variability of $u_C(\theta)$ are

$$J_C(\theta) = (w_1 I_p, \ldots, w_q I_p)\, \Sigma_V(\theta)(w_1 I_p, \ldots, w_q I_p)^\top\,, \qquad H_C(\theta) = (w_1 I_p, \ldots, w_q I_p)\, i_V(\theta)\,.$$

If a vector of weights is such that $H_C(\theta)$ is positive definite, then $E_{\theta_0}(\ell_C(\theta))$ has a local maximum at $\theta = \theta_0$. This condition is crucial for consistency of the maximum composite likelihood estimator and can hold even if $w_j < 0$ for some j (see Example 1).

When $p = 1$, $J_C(\theta) = w^\top \Sigma_V(\theta) w$ and $H_C(\theta) = w^\top i_V(\theta)$, so that Godambe information is

$$G_C(\theta) = \frac{\{w^\top i_V(\theta)\}^2}{w^\top \Sigma_V(\theta) w} = \frac{\mathrm{Cov}^2_\theta\{u_C(\theta), u(\theta)\}}{\mathrm{Var}_\theta\{u_C(\theta)\}}\,. \tag{4}$$

The theory of optimal combination of estimating functions (Lindsay, 1988, Section 4; McCullagh & Nelder, 1989, Section 9.4.2; Heyde, 1997, Chapter 6), applied to $u_j(\theta)$, $j = 1, \ldots, q$, leads to

$$q^*(\theta) = i_V(\theta)^\top \Sigma_V(\theta)^{-1} u_V(\theta)\,. \tag{5}$$

The combination $q^*(\theta)$ is the best linear predictor of $u(\theta)$ based on $u_V(\theta)$. Indeed, $i_V(\theta)^\top\, \Sigma_V(\theta)^{-1}$ is the matrix of the regression coefficients of the full likelihood score, $u(\theta)$, on $u_V(\theta)$. The estimating function $q^*(\theta)$ is both unbiased and information unbiased.

From (3), we see that $u_C(\theta)$ is of the form (5) only if

$$(w_1 I_p, \dots, w_q I_p) = i_V(\theta)^\top \Sigma_V(\theta)^{-1} \tag{6}$$

for some $w_1, \dots, w_q$ not depending on θ. When $p > 1$, condition (6) is severely restrictive, and, even when $p = 1$, it requires that the vector of optimal weights does not depend on θ.

For a scalar θ, Fraser & Reid (2018) obtain optimal weights for $\ell_C(\theta)$ through a first order asymptotic analysis of the $u_j(\theta)$'s at a reference parameter value θ_0. The optimal weights depend on θ_0. They are

$$w^*(\theta_0) = [w_j^*(\theta_0)] = \Sigma_V(\theta_0)^{-1} i_V(\theta_0) \tag{7}$$

and the optimally weighted composite log likelihood is

$$\ell_C^*(\theta) = \sum_{j=1}^{q} w_j^*(\theta_0)\ell_j(\theta)\,. \tag{8}$$

In practice, θ_0 has to be estimated, e.g. from $\ell_I(\theta)$.

The score $u_C^*(\theta) = (\partial/\partial\theta)\ell_C^*(\theta)$ is an unbiased estimating function. At $\theta = \theta_0$ it is information unbiased and optimal, being $u_C^*(\theta_0) = q^*(\theta_0)$. Writing the information unbiasedness condition $J_C(\theta_0) = H_C(\theta_0)$ as

$$w^\top \{\Sigma_V(\theta_0) w - i_V(\theta_0)\} = 0\,, \tag{9}$$

we see that weights $w^*(\theta_0)$ make the vector in brackets equal to the zero vector. Supplementing illustrations in Fraser & Reid (2018, Section 3), we give below three examples of optimal weights for composite likelihood for a scalar θ in some special settings.

Example 1 *Optimal weights need not all be positive.*
Consider a $\Sigma_V(\theta_0)$ having Toeplitz, i.e. diagonal-constant, structure. Let $q = 10$ and first row entries of $\Sigma_V(\theta_0)$ be $I_{1k} = \{(10 - k + 1)/10\}^{1/2}$, $k = 1, \dots, 10$. Then, using (7), $w^*(\theta_0) = (1.942, -0.490, -0.258, -0.183, -0.156, -0.156, -0.183, -0.258, -0.490, 1.942)^\top$. The symmetry relation $w^*(\theta_0)_j = w^*(\theta_0)_{q-j+1}$ holds whenever $\Sigma_V(\theta_0)$ has Toeplitz structure. From (4), the asymptotic relative efficiency of the maximizer of (2) with respect to the maximizer of (8) is 0.729.

Example 2 *Two sources.*
When $q = 2$, $u_V(\theta)^\top = (u_1(\theta), u_2(\theta))$, $i_V(\theta)^\top = (I_{11}(\theta), I_{22}(\theta))$ and

$$\Sigma_V(\theta) = \begin{pmatrix} I_{11}(\theta) & I_{12}(\theta) \\ I_{12}(\theta) & I_{22}(\theta) \end{pmatrix}\,.$$

Supposing, without loss of generality, $I_{11}(\theta) \geq I_{22}(\theta)$, and $-1 < \rho(\theta) < 1$, where $\rho(\theta) = I_{12}(\theta)/\sqrt{I_{11}(\theta) I_{22}(\theta)}$, optimal weights are given by

$$w^*(\theta_0) = \frac{1}{1 - \rho^2(\theta_0)} \begin{pmatrix} 1 - \rho(\theta_0)\sqrt{I_{22}(\theta_0)/I_{11}(\theta_0)} \\ 1 - \rho(\theta_0)\sqrt{I_{11}(\theta_0)/I_{22}(\theta_0)} \end{pmatrix}\,.$$

From (4), the asymptotic relative efficiency of the maximizer of (2) with respect to the maximizer of (8) is equal to $(1 - \rho^2)/(1 - 4\rho^2 Q/(1 + Q)^2)$, where $\rho = \rho(\theta_0)$, $Q = I_{11}(\theta_0)/I_{22}(\theta_0) \geq 1$. It approaches zero as ρ^2 approaches 1, the faster the larger $Q - 1$ is.

Example 3 *Correlation exchangeable sources.*
Suppose that the q sources are correlation exchangeable, i.e. $I_{jj}(\theta) = \sigma^2(\theta)$ and $I_{jk}(\theta) = \rho(\theta)\sigma^2(\theta)$ for every $j, k \in \{1, \dots, q\}$ with $j \neq k$. Then $\Sigma_V(\theta) = \sigma^2(\theta) R_V(\theta)$, where

$$R_V(\theta) = (1 - \rho(\theta)) I_q + \rho(\theta) 1_q 1_q^\top\,,$$

where 1_q is a q-dimensional vector of ones. The matrix $R_V(\theta)$ is positive definite provided that $\rho(\theta) > -1/(q - 1)$. Equation (9) is satisfied if $R_V(\theta_0) w = 1_q$. The sum of the entries of each row of $R_V(\theta_0)$ is constant and equal to $1 + \rho(\theta_0)(q - 1)$. Therefore, the optimal weights are

$$w^*(\theta_0) = \frac{1}{1 + \rho(\theta_0)(q - 1)} 1_q\,,$$

so that $\ell_C^*(\theta) = \ell_I(\theta)/\{1 + \rho(\theta_0)(q - 1)\}$ and the maximizer of (8) coincides with the maximizer of (2).

The composite log likelihood ratio statistic from $\ell_C^*(\theta)$,

$$W_C^*(\theta_0) = 2\left\{\ell_C^*(\hat\theta_C^*) - \ell_C^*(\theta_0)\right\}\,, \tag{10}$$

with $\hat{\theta}^*_C$ the maximizer of $\ell^*_C(\theta)$, has a χ^2_1 asymptotic null distribution under regularity conditions. Optimality of $u^*_C(\theta_0)$ translates into maximization of the local power of (10).

In contrast, when $J_C(\theta_0) \neq H_C(\theta_0)$, the log likelihood ratio statistic $W_C(\theta_0)$ from $\ell_C(\theta_0)$ has a rescaled χ^2_1 asymptotic null distribution, νZ^2, where the $Z \sim N(0,1)$ and $\nu = J_C(\theta_0)H_C(\theta_0)^{-1}$ (see e.g. Molenberghs & Verbeke 2005, Section 9.3.3). Therefore,

$$W_A(\theta_0) = \frac{H_C(\theta_0)}{J_C(\theta_0)} W_C(\theta_0) \tag{11}$$

is asymptotically χ^2_1 under θ_0. The local power of $W_A(\theta_0)$ is usually inferior to that of $W^*_C(\theta_0)$. Indeed, standard expansions show that the local non-null asymptotic distribution of $W_A(\theta_0)$, when the true parameter is $\theta = \theta_0 + \delta/\sqrt{n}$, is non-central chi-squared on one degree of freedom and non-centrality parameter $\delta^2 G_C(\theta_0)/n$, which is maximum when $\ell_C(\theta)$ is $\ell^*_C(\theta)$. In the special case when the q sources are correlation exchangeable, as in Example 3, $W^*_C(\theta_0)$ coincides with the multiplicative adjustment (11) of the log likelihood ratio from $\ell_I(\theta)$. Multiplicative adjustments of $W_C(\theta_0)$ when $p > 1$ are proposed in Chandler & Bate (2007) and Pace et al. (2011).

3 | INFERENCE ON A SCALAR PARAMETER OF INTEREST

Suppose that the log likelihoods to be combined are $\ell_j(\psi, \lambda)$, $j = 1, \ldots, q$, where ψ is a scalar parameter of interest and λ is a nuisance parameter. Let $u_{j\psi}(\psi,\lambda) = (\partial/\partial\psi)\ell_j(\psi,\lambda)$ and $u_{j\lambda}(\psi,\lambda) = (\partial/\partial\lambda)\ell_j(\psi,\lambda)$ be the components of the score function of $\ell_j(\psi,\lambda)$.

The most straightforward route to obtain first order optimal weights for inference about ψ is to define a composite log likelihood for ψ by combining profile log likelihoods from each source. This gives

$$\ell_C(\psi) = \sum_{j=1}^{q} w_j \ell_j(\psi, \hat{\lambda}_{j\psi}), \tag{12}$$

where $\hat{\lambda}_{j\psi}$ is the maximizer of $\ell_j(\psi,\lambda)$ with respect to λ with ψ fixed.

First order optimal weights for (12) may be derived as a generalization of (7) because the log likelihoods $\ell_j(\psi, \hat{\lambda}_{j\psi})$ satisfy to $O(1)$ the first two Bartlett identities. In particular, they are obtained by replacing $\Sigma_V(\theta_0)$ and $i_V(\theta_0)$ in (7) by a first order approximation, Σ_{VP}, of the covariance matrix of $[u_{j\psi}(\psi, \hat{\lambda}_{j\psi})]$ and of its main diagonal, respectively. Off-diagonal entries of Σ_{VP} involve correlation between sources that may depend on an additional parameter ν. The overall parameter is then $\theta = (\psi, \lambda, \nu)$, so that generally $\Sigma_{VP} = \Sigma_{VP}(\theta)$. We suppose that $\Theta = \Psi \times \Lambda \times N$, where $\psi \in \Psi, \lambda \in \Lambda, \nu \in N$.

The first order approximation Σ_{VP} is obtained using the standard expansion

$$u_{j\psi}(\psi, \hat{\lambda}_{j\psi}) = u_{j\psi}(\psi,\lambda) - I_{j\psi,j\lambda}(\psi,\lambda) I_{j\lambda,j\lambda}(\psi,\lambda)^{-1} u_{j\lambda}(\psi,\lambda) + O_p(1),$$

where $I_{j\psi,j\lambda}(\psi,\lambda) = E_{\psi,\lambda}\left\{u_{j\psi}(\psi,\lambda)u_{j\lambda}(\psi,\lambda)^\top\right\}$ and $I_{j\lambda,j\lambda}(\psi,\lambda) = E_{\psi,\lambda}\left\{u_{j\lambda}(\psi,\lambda)u_{j\lambda}(\psi,\lambda)^\top\right\}$ are blocks of the Fisher information from $\ell_j(\psi,\lambda)$, as well as $I_{j\psi,j\psi}(\psi,\lambda) = \mathrm{Var}_{\psi,\lambda}\left\{u_{j\psi}(\psi,\lambda)\right\}$, $j = 1, \ldots, q$. The diagonal entries of Σ_{VP} are then

$$(\Sigma_{VP})_{jj} = I_{j\psi,j\psi}(\psi,\lambda) - I_{j\psi,j\lambda}(\psi,\lambda)\, I_{j\lambda,j\lambda}(\psi,\lambda)^{-1}\, I_{j\lambda,j\psi}(\psi,\lambda).$$

The off-diagonal entries of Σ_{VP} are expressed using, for $j, k = 1, \ldots, q$, $j \neq k$, the quantities

$$\begin{aligned} I_{j\psi,k\psi}(\theta) &= E_\theta\left\{u_{j\psi}(\psi,\lambda)u_{k\psi}(\psi,\lambda)\right\}, \quad I_{j\lambda,k\psi}(\theta) = E_\theta\left\{u_{j\lambda}(\psi,\lambda)u_{k\psi}(\psi,\lambda)\right\}, \\ I_{j\psi,k\lambda}(\theta) &= E_\theta\left\{u_{j\psi}(\psi,\lambda)u_{k\lambda}(\psi,\lambda)^\top\right\}, \quad I_{j\lambda,k\lambda}(\theta) = E_\theta\left\{u_{j\lambda}(\psi,\lambda)u_{k\lambda}(\psi,\lambda)^\top\right\}. \end{aligned}$$

We obtain

$$\begin{aligned} (\Sigma_{VP})_{jk} \;=\; & I_{j\psi,k\psi}(\theta) - I_{k\psi,k\lambda}(\psi,\lambda) I_{k\lambda,k\lambda}(\psi,\lambda)^{-1} I_{k\lambda,j\psi}(\theta) - I_{j\psi,j\lambda}(\psi,\lambda) I_{j\lambda,j\lambda}(\psi,\lambda)^{-1} I_{j\lambda,k\psi}(\theta) \\ & + I_{j\psi,j\lambda}(\psi,\lambda) I_{j\lambda,j\lambda}(\psi,\lambda)^{-1} I_{j\lambda,k\lambda}(\theta) I_{k\lambda,k\lambda}(\psi,\lambda)^{-1} I_{k\lambda,k\psi}(\theta). \end{aligned}$$

Hence, denoting by θ_0 the true value of θ, first order optimal weights in (12) are

$$w^*_P(\theta_0) = \Sigma_{VP}(\theta_0)^{-1} \mathrm{diag}(\Sigma_{VP}(\theta_0)).$$

When ψ and λ are orthogonal for all sources, i.e. when, for $j = 1, \ldots, q$, $E_{\psi,\lambda}\left\{u_{j\psi}(\psi,\lambda)\, u_{j\lambda}(\psi,\lambda)\right\} = I_{j\psi,j\lambda}(\psi,\lambda) = 0$, we have $\Sigma_{VP}(\theta_0) = [I_{j\psi,k\psi}(\theta_0)]$, where $I_{j\psi,j\psi}(\theta_0) = I_{j\psi,j\psi}(\psi_0, \lambda_0)$. Therefore, first order optimal weights become

$$w^*_P(\theta_0) = [I_{j\psi,k\psi}(\theta_0)]^{-1} [I_{j\psi,j\psi}(\theta_0)]. \tag{13}$$

Moreover, under orthogonality, first order inference based on $\ell_j(\psi, \hat{\lambda}_{j\psi})$ is equivalent to first order inference based on $\ell_j(\psi, \hat{\lambda}_j)$ or even on $\ell_j(\psi, \tilde{\lambda})$, where $\hat{\lambda}_j$ is the unconstrained maximum likelihood estimate of λ from $\ell_j(\psi,\lambda)$ and $\tilde{\lambda}$ is a $\sqrt{n}$-consistent estimate of λ. A first order optimal combination is, therefore,

$$\ell_C^*(\psi) = w_P^*(\theta_0)^\top \ell_V(\psi, \tilde{\lambda}) \,, \tag{14}$$

with $\ell_V(\psi, \lambda) = [\ell_j(\psi, \lambda)]$.

To use $\ell_C^*(\psi)$ in practice, weights $w_P^*(\theta_0)$ are to be estimated by a $\sqrt{n}$-consistent estimate $\tilde{w}_P^*$, giving $\ell_{EC}^*(\psi) = (\tilde{w}_P^*)^\top \ell_V(\psi, \tilde{\lambda})$. The composite log likelihood $\ell_{EC}^*(\psi)$ satisfies to $O(1)$ the first two Bartlett identities. As a consequence, log likelihood ratio statistic

$$W_{EC}^*(\psi_0) = 2\left\{\sup_{\psi\in\Psi} \ell_{EC}^*(\psi) - \ell_{EC}^*(\psi_0)\right\} \tag{15}$$

has an asymptotic χ^2_1 null distribution.

4 | ILLUSTRATIONS: INFERENCE ON A COMMON MEAN IN MULTIVARIATE NORMAL MODELS

Let $Y_1, \ldots, Y_n$ be independent random variables with $Y_i = (Y_{i1}, \ldots, Y_{iq})^\top \sim N_q(\mu 1_q, \sigma^2 R)$, where $\mu \in \mathbb{R}$ is the parameter of interest, $\sigma^2 > 0$ is a nuisance parameter, and the correlation matrix R is provisionally supposed to be known. Univariate margins provide the log likelihoods

$$\ell_j(\mu, \sigma^2) = -\frac{n}{2}\log\sigma^2 - \frac{1}{2\sigma^2}\sum_{i=1}^n (y_{ij} - \mu)^2 \,, \qquad j = 1, \ldots, q \,.$$

The parameter of interest μ is orthogonal to covariance parameters. To compute first order optimal weights (13), only covariances of scores $u_{j\mu}(\mu, \sigma^2) = \sum_{i=1}^n (y_{ij} - \mu)/\sigma^2, j = 1, \ldots, q$, are needed. We have $I_{j\mu,k\mu}(\mu, \sigma^2) = n\mathrm{Cov}(Y_{1j}, Y_{1k})/\sigma^4, j, k = 1, \ldots, q$, so that $\Sigma_{VP}(\mu_0, \sigma_0^2) = nR/\sigma_0^2$ and

$$w_P^*(\theta_0) = w_P^* = R^{-1}\mathrm{diag}(R) = R^{-1}1_q \,.$$

Log likelihood (14) for μ is then

$$\begin{aligned} \ell_C^*(\mu) &= -\frac{n}{2}(1_q^\top w_P^*)\log\tilde{\sigma}^2 - \frac{1}{2\tilde{\sigma}^2}\sum_{j=1}^q (w_P^*)_j \sum_{i=1}^n (y_{ij} - \mu)^2 \\ &= -\frac{n}{2}(1_q^\top R^{-1} 1_q)\log\tilde{\sigma}^2 - \frac{1}{2\tilde{\sigma}^2}\sum_{i=1}^n 1_q^\top R^{-1}\mathrm{diag}\left\{(y_i - 1_q\mu)(y_i - 1_q\mu)^\top\right\} \,, \end{aligned}$$

where $\tilde{\sigma}^2$ is a moment estimate σ^2, and is maximized by

$$\hat{\mu}_C = \frac{(w_P^*)^\top \bar{y}_V}{(w_P^*)^\top 1_q} \,,$$

where $\bar{y}_V^\top = (\bar{y}_1, \ldots, \bar{y}_q)$, with $\bar{y}_j = \sum_{i=1}^n y_{ij}/n$. A direct check shows that $\hat{\mu}_C$ is also the maximum of the full log likelihood

$$\ell(\mu) = -\frac{nq}{2}\log\sigma^2 - \frac{n}{2}\log|R| - \frac{1}{2\sigma^2}\sum_{i=1}^n (y_i - 1_q\mu)^\top R^{-1}(y_i - 1_q\mu) \,.$$

The likelihood ratio statistic from $\ell_C^*(\mu)$ is

$$W_C^*(\mu_0) = \frac{n}{\tilde{\sigma}^2}(1_q^\top w_P^*)(\hat{\mu}_C - \mu_0)^2 \,. \tag{16}$$

We will use $W_{EC}^*(\mu_0)$ to denote the likelihood ratio statistic (16) computed with estimated weights $\tilde{w}_P^*$, that is with estimated R.

Because of orthogonality, σ^2 in the independence log likelihood

$$\ell_I(\mu, \sigma^2) = -\frac{nq}{2}\log\sigma^2 - \frac{1}{2\sigma^2}\sum_{j=1}^q\sum_{i=1}^n (y_{ij} - \mu)^2$$

may be replaced with a consistent estimate $\tilde{\sigma}^2$. Adjustment of $\ell_I(\mu, \tilde{\sigma}^2)$ requires

$$H_I = H_I(\mu_0, \sigma_0^2) = 1_q^\top i_V(\mu_0, \sigma_0^2) = \frac{nq}{\sigma_0^2}$$

and

$$J_I = J_I(\mu_0, \sigma_0^2) = 1_q^\top \Sigma_V(\mu_0, \sigma_0^2) 1_q = \frac{n}{\sigma_0^2} 1_q^\top R 1_q \,.$$

The adjustment factor H_I/J_I is $q/(1_q^\top R 1_q)$ and the corresponding adjusted likelihood ratio statistic is

$$W_A(\mu_0) = \frac{nq^2}{\tilde{\sigma}^2\, 1_q^\top R 1_q}(\hat{\mu}_I - \mu_0)^2 \tag{17}$$

TABLE 1 Equicorrelated multivariate normal model with common marginal means. Empirical rejection probabilities (%) for $W^*_{EC}(\mu)$ for testing $\mu = 0$ at nominal levels 10%, 5%, 1%, 10^4 replications, $n = 5, 10, 20$, $q = 5$, true values $\sigma^2 = 1$ and $\rho = 0.1, 0.5, 0.9$, (a) $\mu = 0$, (b) $\mu = 0.5$.

		n = 5			n = 10			n = 20		
	ρ	10	5	1	10	5	1	10	5	1
	0.1	13.10	7.75	2.82	11.72	6.82	2.09	11.34	6.36	1.75
(a)	0.5	17.61	11.95	5.87	13.55	8.18	2.93	11.43	6.75	2.15
	0.9	17.43	12.33	6.37	13.85	8.20	2.88	11.22	6.33	2.07
	0.1	66.65	55.42	36.38	89.69	83.22	65.26	99.32	98.41	93.87
(b)	0.5	46.78	37.56	23.61	65.94	55.03	35.23	88.86	82.21	63.35
	0.9	38.14	29.34	17.44	51.81	40.81	23.24	75.64	65.49	42.86

with $\hat{\mu}_I = \bar{y} = (nq)^{-1}\sum_{i=1}^n\sum_{j=1}^q y_{ij}$. We will denote by $W_{EA}(\mu_0)$ the likelihood ratio statistic (17) computed with estimated R.

The asymptotic efficiency of $\hat{\mu}_I$ relative to $\hat{\mu}_C$ depends on R only and is

$$\mathrm{ARE} = \frac{q^2}{(1_q^\top R 1_q)\,(1_q^\top R^{-1} 1_q)}\,. \tag{18}$$

In the following, we consider three correlation structures with parameters to be estimated. Simulation results are given to compare the null distribution and power of $W^*_{EC}(\mu)$ and $W_{EA}(\mu)$.

Example 4 *Equicorrelated multivariate normal model.*

Let $R = (1-\rho)I_q + \rho 1_q 1_q^\top$, with $\rho \geq 0$. The process generating data y_{ij}, $i = 1, \dots, n$, $j = 1, \dots, q$, can be seen as $Y_{ij} = \mu + \xi_i + \epsilon_{ij}$, where ξ_i and ϵ_{ij} are independent random variables having marginal distributions $\xi_i \sim N(0, \sigma^2_\xi)$ and $\epsilon_{ij} \sim N(0, \sigma^2_\epsilon)$. Then $\sigma^2 = \sigma^2_\xi + \sigma^2_\epsilon$, $\rho = \sigma^2_\xi/(\sigma^2_\xi + \sigma^2_\epsilon)$.

Sources $\ell_j(\mu, \sigma^2)$ are correlation exchangeable, being $u_{j\mu}(\mu, \sigma^2) = \sum_{i=1}^n (y_{ij} - \mu)/\sigma^2$, and $I_{j\mu,j\mu}(\mu, \sigma^2) = n\sigma^2$, $I_{j\mu,k\mu}(\mu, \sigma^2, \rho) = n\sigma^2\rho$. Results of Example 3 apply, so that $\ell^*_C(\mu) = 1/(1 + \rho(q-1))\ell_I(\mu, \tilde{\sigma}^2)$ and

$$W^*_{EC}(\mu_0) = \frac{nq}{\tilde{\sigma}^2(1 + \tilde{\rho}(q-1))}(\bar{y} - \mu_0)^2\,,$$

that coincides with $W_{EA}(\mu_0)$ computed using the same estimates $\tilde{\sigma}^2$ and $\tilde{\rho}$. Moment estimates $\tilde{\sigma}^2$ and $\tilde{\rho}$ of σ^2 and ρ may be used (see e.g. Searle et al. 1992, Section 3.5) given by

$$\tilde{\sigma}^2 = \frac{SS_B}{(n-1)q} + \frac{SS_E}{nq}\,, \qquad \tilde{\rho} = \max\left(\frac{SS_B/(n-1) - SS_E/(n(q-1))}{SS_E/n + SS_B/(n-1)}, 0\right)\,,$$

where

$$SS_E = \sum_{i=1}^n\sum_{j=1}^q (y_{ij} - \bar{y}_i)^2\,, \qquad SS_B = q\sum_{i=1}^n (\bar{y}_i - \bar{y})^2\,,$$

with $\bar{y}_i = \sum_{j=1}^q y_{ij}/q$.

The empirical distribution of $W^*_{EC}(\mu)$ for testing $\mu = 0$ has been evaluated through a simulation study. The results with 10^4 replications, $n = 5, 10, 20$, $q = 5$, true values $\sigma^2 = 1$, $\rho = 0.1, 0.5, 0.9$, and (a) $\mu = 0$, (b) $\mu = 0.5$, are displayed in Table 1. The null distribution of $W^*_{EC}(\mu)$ approaches the nominal χ^2_1 as n increases or ρ gets closer to zero. From $n = 20$, empirical power is reasonably close to values, not shown in the table, obtained with known σ^2 and ρ using the statistic $W^*_C(\mu_0) = nq(\bar{y} - \mu_0)^2/\{\sigma_0^2(1 + \rho_0(q-1))\}$ whose null distribution is exactly χ^2_1.

Example 5 *AR(1) model.*

Let us consider an autoregressive specification of order 1, i.e. $R = [\rho^{|j-k|}]$, where $|\rho| < 1$. Let ϵ_{ij} be independent random variables with marginal distribution $\epsilon_{ij} \sim N(0, \sigma^2_\epsilon)$, $\sigma^2_\epsilon > 0$. The data generating process for y_{ij}, $i = 1, \dots, n$, $j = 1, \dots, q$, can be seen as $Y_{i1} \sim N(\mu, \sigma^2_\epsilon/(1-\rho^2))$ and $Y_{ij} - \mu = \rho(Y_{ij-1} - \mu) + \epsilon_{ij}$, $j = 2, \dots, q$. Here, $\sigma^2 = \sigma^2_\epsilon/(1-\rho^2)$. From

$$R^{-1} = \frac{1}{1-\rho^2}\begin{pmatrix} 1 & -\rho & 0 & \cdots & 0 & 0 \\ -\rho & 1+\rho^2 & -\rho & 0 & \cdots & 0 \\ 0 & -\rho & 1+\rho^2 & -\rho & \cdots & 0 \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ 0 & 0 & \cdots & 0 & -\rho & 1 \end{pmatrix}$$

TABLE 2 AR(1) model. Empirical rejection probabilities (%) for $W_{EA}(\mu)$ and $W^*_{EC}(\mu)$ for testing $\mu = 0$ at nominal levels 10%, 5%, 1%, 10^4 replications, $q = 10$, true values $\sigma^2 = 1$, $\rho = -0.8$, $\mu = 0$ and $\mu = 0.5$ with $n = 1$, $\mu = 0$ and $\mu = 0.25$ with $n = 5$, $\mu = 0$ and $\mu = 0.20$ with $n = 10$.

		$\mu = 0$			$\mu = 0.5$		
		10	5	1	10	5	1
$n = 1$	$W_{EA}(\mu)$	13.75	8.29	3.12	77.85	68.59	48.77
	$W^*_{EC}(\mu)$	14.50	9.14	3.72	84.31	76.49	57.67
		$\mu = 0$			$\mu = 0.25$		
		10	5	1	10	5	1
$n = 5$	$W_{EA}(\mu)$	10.82	5.97	1.25	84.67	76.36	54.46
	$W^*_{EC}(\mu)$	10.78	5.76	1.42	91.02	84.33	66.38
		$\mu = 0$			$\mu = 0.20$		
		10	5	1	10	5	1
$n = 10$	$W_{EA}(\mu)$	10.40	5.24	0.97	91.48	85.79	67.46
	$W^*_{EC}(\mu)$	10.26	4.97	0.97	96.02	92.35	79.40

we get

$$(w^*_P)^\top = \frac{1}{1-\rho^2}\left(1-\rho, (1-\rho)^2, \ldots, (1-\rho)^2, 1-\rho\right) .$$

We see that, as q diverges, $\ell^*_C(\mu) = \{(1-\rho)/(1+\rho)\}\ell_I(\mu) + O(1)$. Moment estimates of ρ and σ^2_ϵ are

$$\tilde{\rho} = \frac{q\sum_{i=1}^n \sum_{j=1}^{q-1}(y_{ij} - \bar{y})(y_{ij+1} - \bar{y})}{(q-1)\sum_{i=1}^n \sum_{j=1}^{q}(y_{ij} - \bar{y})^2} ,$$

$$\tilde{\sigma}^2_\epsilon = \frac{1}{n(q-1)}\sum_{i=1}^{n}\sum_{j=1}^{q-1}\left\{(y_{ij+1} - \bar{y}) - \tilde{\rho}(y_{ij} - \bar{y})\right\}^2 .$$

Simulation results based on 10^4 replications for testing $\mu = 0$ with $n = 1, 5, 10$ and $q = 10$ using $W^*_{EC}(\mu)$ and $W_{EA}(\mu)$ are summarized in Table 2. Covariance parameters have been set to $\sigma^2 = 1$ and $\rho = -0.8$. These values give ARE in (18) equal to 0.786, so that a tangible gain in power is expected when using $W^*_{EC}(\mu)$ in place of $W_{EA}(\mu)$. Samples were generated with μ values $\mu = 0$ and $\mu = 0.5$ for $n = 1$, $\mu = 0$ and $\mu = 0.25$ for $n = 5$, $\mu = 0$ and $\mu = 0.20$ for $n = 10$. Empirical rejection probabilities displayed in Table 2 show a satisfactory agreement with nominal levels when $n \geq 5$ both for $W^*_{EC}(\mu)$ and for $W_{EA}(\mu)$, while the power of the former is visibly larger.

Example 6 *General correlation model.*
With R unstructured and $n > q$, first order optimal weights can be estimated from the sample correlation matrix $\tilde{R}$ of data y_{ij}. Reasonable accuracy requires n to be much larger than q.

For $W^*_{EC}(\mu)$ and $W^*_{EA}(\mu)$ with $\mu = 0$, simulation results based on 10^4 replications are summarized in Table 3. Data with $n = 50, 100, 200$, $q = 5$, were generated from an AR(1) model having $\sigma^2 = 1$, $\rho = -0.8$, and μ values $\mu = 0$ and $\mu = 0.1$ for $n = 50$, $\mu = 0$ and $\mu = 0.075$ for $n = 100$, $\mu = 0$ and $\mu = 0.05$ for $n = 200$. Although empirical rejection probabilities displayed in Table 3 show a faster convergence to nominal levels for $W_{EA}(\mu)$, the power gain using $W^*_{EC}(\mu)$ is clearly seen.

TABLE 3 General correlation model. Empirical rejection probabilities (%) for $W_{EA}(\mu)$ and $W^*_{EC}(\mu)$ for testing $\mu = 0$ at nominal levels 10%, 5%, 1%, 10^4 replications, $q = 5$, data generated from AR(1) with true values $\sigma^2 = 1$, $\rho = -0.8$, $\mu = 0$ and $\mu = 0.1$ for $n = 50$, $\mu = 0$ and $\mu = 0.075$ for $n = 100$, $\mu = 0$ and $\mu = 0.05$ for $n = 200$, .

		$\mu = 0$			$\mu = 0.1$		
		10	5	1	10	5	1
$n = 50$	$W_{EA}(\mu)$	9.47	4.68	0.97	61.81	49.07	25.98
	$W^*_{EC}(\mu)$	12.81	7.14	1.95	83.43	74.99	54.81
		$\mu = 0$			$\mu = 0.075$		
		10	5	1	10	5	1
$n = 100$	$W_{EA}(\mu)$	9.94	5.11	1.09	65.25	53.42	29.48
	$W^*_{EC}(\mu)$	11.25	6.00	1.40	86.07	78.54	57.63
		$\mu = 0$			$\mu = 0.05$		
		10	5	1	10	5	1
$n = 200$	$W_{EA}(\mu)$	9.66	5.02	0.89	60.33	47.82	25.21
	$W^*_{EC}(\mu)$	10.28	5.41	1.16	83.40	73.62	50.43

5 | DISCUSSION

In this paper, we obtain first order optimal weights when combining individual profile log likelihoods for inference on a scalar parameter in the presence of nuisance parameters. The examples in Section 4 indicate that replacing nuisance parameters with moment estimates preserves gains due to optimal weights. These results are in line with findings in Fraser & Reid (2018, in particular, Example 7).

Practical implementation of (15), especially outside orthogonality, requires computation of $\Sigma_{VP}(\theta_0)$. Apart from special cases as those considered in Section 4 for illustration, exact calculations are generally unfeasible. One possibility is to evaluate $\Sigma_{VP}(\theta_0)$ via simulation from the full model with estimated parameters, as done in Cattelan & Sartori (2016) for multiplicative adjustments of composite likelihood ratio. Parameter estimates can be obtained from an independence combination of marginal likelihoods from low-dimensional portions of the data allowing estimation of θ_0.

As an alternative to combining profile log likelihoods from each source, a starting point for finding first order optimal weights for inference about ψ is the overall profile from $\sum_{j=1}^q w_j \ell_j(\psi, \lambda)$, that is

$$\ell_{OC}(\psi) = \sum_{j=1}^q w_j \ell_j(\psi, \hat\lambda_{\psi C}),$$

where $\hat\lambda_{\psi C}$ is the constrained maximizer of $\sum_{j=1}^q w_j \ell_j(\psi, \lambda)$ with respect to λ for a given ψ. The corresponding score $u_{OC}(\psi) = (\partial/\partial\psi)\ell_{OC}(\psi)$ is

$$u_{OC}(\psi) = \sum_{j=1}^q w_j u_{j\psi}(\psi, \hat\lambda_{\psi C}).$$

Standard asymptotic calculations show that, to first order, sensitivity and variability of $u_{OC}(\psi)$ are

$$H_{OC} = E_\psi\left\{-(\partial/\partial\psi)u_{OC}(\psi)\right\} = \sum_{j=1}^q w_j I_{j\psi,j\psi} - \left(\sum_{j=1}^q w_j I_{j\psi,j\lambda}\right)\left(\sum_{j=1}^q w_j I_{j\lambda,j\lambda}\right)^{-1}\left(\sum_{j=1}^q w_j I_{j\lambda,j\psi}\right)$$

and

$$
\begin{aligned}
J_{OC} = \mathrm{Var}_\theta\{u_{OC}(\psi)\} = {} & w^\top \mathrm{Var}_\theta([u_{j\psi}])\, w \\
& + \left(\sum_{j=1}^q w_j I_{j\psi,\, j\lambda}\right)\left(\sum_{j=1}^q w_j I_{j\lambda, j\lambda}\right)^{-1} \mathrm{Var}_\theta\left(\sum_{j=1}^q w_j u_{j\lambda}\right)\left(\sum_{j=1}^q w_j I_{j\lambda,\, j\lambda}\right)^{-1}\left(\sum_{j=1}^q w_j I_{j\lambda,\, j\psi}\right) \\
& -2\left(\sum_{j=1}^q w_j I_{j\psi, j\lambda}\right)\left(\sum_{j=1}^q w_j I_{j\lambda,\, j\lambda}\right)^{-1}\left(\sum_{j=1}^q\sum_{k=1}^q w_j w_k I_{j\lambda,\, k\psi}\right) .
\end{aligned}
$$

Above, for ease of notation, arguments of u and I quantities are suppressed. Moreover, $\mathrm{Var}_\theta([u_{j\psi}])$ denotes the covariance matrix of the vector with entries $u_{j\psi}$ and $\mathrm{Var}_\theta(\sum_{j=1}^q w_j u_{j\lambda})$ is the covariance matrix of the vector $\sum_{j=1}^q w_j u_{j\lambda}$. Under orthogonality of ψ and λ for all sources, the condition $H_{OC} = J_{OC}$ may be written in the form (9) and is satisfied by the weights $w_P^*(\theta_0)$ in (13). Outside orthogonality, search for optimal weights would require numerical maximization of the Godambe information from H_{OC} and J_{OC} at θ_0.

Examples in Section 4 are limited to one-wise log likelihoods, so that q is not large. Pairwise log likelihoods from d-dimensional observations Y_i would involve $q = d(d-1)/2$ dependent sources, and q rapidly becomes large, making optimal combination unfeasible. A reduction may arise when a structured covariance matrix induces equality constraints on the component of w_P^*, as happens when $\Sigma_{VP}(\theta_0)$ has a Toeplitz structure, like in Examples 4 and 5. The number of sources could thus be reduced by first adding the log likelihoods having equal weights, and then optimally combining the resulting sources. Preliminary computation of $\Sigma_{VP}(\theta)$ at some trial parameter values might be useful to detect sources having the same weight in the combination, at least approximately.

ACKNOWLEDGEMENTS

We express our thanks to the referees for valuable suggestions that helped us improve the presentation. This research was partially supported by the Italian Ministry of Education PRIN 2015 grant 2015EASZFS_003, by the University of Udine grant PRID2017_DIES004 and by the University of Padova grant BIRD185955.

DATA AVAILABILITY STATEMENT

The R code that produced the simulation results in this paper is available on request from the authors.

References

Castruccio, S., Huser, R. & Genton, M. G. (2016). High-order composite likelihood inference for max-stable distributions and processes. *Journal of Computational and Graphical Statistics*, 25, 1212–1229.

Cattelan, M. & Sartori, N. (2016). Empirical and simulated adjustments of composite likelihood ratio statistics. *Journal of Statistical Computation and Simulation*, 86, 1056–1067.

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

Deng, C., Waagepetersen, R. P. & Guan, Y. (2014). A combined estimating function approach for fitting stationary point process models. *Biometrika*, 101, 393–408.

Fraser, D. A. S. & Reid, N. (2018). Combining likelihood and significance functions. *Statistica Sinica*, to appear, doi:10.5705/ss.202016.0508.

Harden, S. (2013). *Weighted Composite Likelihoods*. Ph.D. thesis, University College London.

Heyde, C. C. (1997). *Quasi-Likelihood and its Applications*. Springer, New York.

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

Kuk, A. Y. C. (2007). A hybrid pairwise likelihood method. *Biometrika*, 94, 939–952.

Li, F. & Sang, H. (2018). On approximating optimal weighted composite likelihood method for spatial models. *Stat*, 7:e194.

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

Lindsay, B. G. (1988). Composite likelihood methods. *Contemporary Mathematics*, 80, 221–240.

Lindsay, B. G. Yi, G. & Sun, J. (2011). Issues and strategies in the selection of composite likelihoods. *Statistica Sinica*, 21, 71–105.

McCullagh, P. & Nelder, J. (1989). *Generalized Linear Models*, 2nd edn. Chapman and Hall, London.

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

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

Sang, H. & Genton, M. G. (2014). Tapered composite likelihood for spatial max-stable models. *Spatial Statistics*, 8, 86–103.

Searle, S., Casella, G. & McCulloch, C. (1992). *Variance Components*. Wiley, New York.

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

□