Statistica Sinica **34** (2024), 47-66
doi:https://doi.org/10.5705/ss.202021.0235

FAST CONSTRUCTION OF OPTIMAL COMPOSITE LIKELIHOODS

Zhendong Huang* and Davide Ferrari

Free University of Bolzano

Abstract: A composite likelihood is a combination of low-dimensional likelihood objects, and is useful in applications in which the data have a complex structure. The construction of a composite likelihood is crucial, affecting both the computing and the statistical properties of the resulting estimator. Despite this, there is no universal rule for combining low-dimensional likelihood objects that is statistically justified and fast in execution. This study develops a methodology to select and combine the most informative low-dimensional likelihoods from a large set of candidates, while estimating the parameters. The proposed procedure minimizes the distance between composite likelihood and full likelihood scores, subject to a computing cost constraint. The selected composite likelihood is sparse in the sense that it contains relatively few informative sub-likelihoods, and the noisy terms are dropped. The resulting estimator is found to have an asymptotic variance close to that of the minimum-variance estimator constructed using all of the low-dimensional likelihoods.

Key words and phrases: Composite likelihood estimation, composite likelihood selection, O_F-optimality, sparsity-inducing penalty.

1. Introduction

The likelihood function is central to many statistical analyses. However, in some cases, the full likelihood is computationally intractable or difficult to specify. This has motivated the development of composite likelihood methods that avoid intractable full likelihoods by combining a set of low-dimensional likelihood objects. Owing to its flexible framework and computational advantages, composite likelihood inference has become popular in many areas of statistics; see, for example, Varin, Reid and Firth (2011) for an overview and applications.

Let $X \subseteq \mathbb{R}^d$ be a $d \times 1$ vector random variable with density in the family $f(x;\theta)$, where $\theta \in \Theta \subseteq \mathbb{R}^p$ is an unknown parameter. Let θ^* denote the true parameter. Suppose now that the full d-dimensional distribution is difficult to specify or compute, but that it is possible to specify m tractable probability density functions $f_1(s_1;\theta), \ldots, f_m(s_m;\theta)$ for sub-vectors $S_1, \ldots, S_m$ of X, each with dimension smaller than d. For example, S_1 may represent a single element such as X_1, a variable pair such as (X_1, X_2), or a conditional sub-vector such as

*Corresponding author

$X_1|X_2$. The total number of sub-models m may grow quickly with d; for example, taking all variable pairs in X yields $m = d(d-1)/2$. Thus, from one vector X, we form the composite log-likelihood

$$\ell(\theta, w; X) = \sum_{j=1}^{m} w_j \ell_j(\theta; X) = \sum_{j=1}^{m} w_j \log f_j(S_j; \theta),$$

where w is an $m \times 1$ vector of constants determined as the solution to an optimality problem. For n independent and identically distributed (i.i.d.) vectors $X^{(1)}, \ldots, X^{(n)}$, we define

$$\ell(\theta, w; X^{(1)}, \ldots, X^{(n)}) = \sum_{i=1}^{n} \ell(\theta; w, X^{(i)}).$$

The score functions are obtained in the usual way:

$$U(\theta, w; X) = \nabla \ell(\theta, w; X) = \sum_{j=1}^{m} w_j U_j(\theta; X), \quad U_j(\theta; X) = \nabla \ell_j(\theta; X),$$

$$U(\theta, w; X^{(1)}, \ldots, X^{(n)}) = \nabla \ell(\theta, w; X^{(1)}, \ldots, X^{(n)}) = \sum_{j=1}^{m} w_j \sum_{i=1}^{n} U_j(\theta; X^{(i)}),$$

where "∇" denotes the gradient with respect to θ. The maximum composite likelihood estimator $\hat{\theta}(w)$ is defined as the solution to the estimating equation

$$U(\theta, w; X^{(1)}, \ldots, X^{(n)}) = \sum_{j=1}^{m} w_j \sum_{i=1}^{n} U_j(\theta; X^{(i)}) = 0, \tag{1.1}$$

for some appropriate choice of w. In addition to its computational advantages and modeling flexibility, the composite likelihood estimator is popular because it enjoys properties analogous to those of a maximum likelihood (Lindsay (1988); Lindsay, Yi and Sun (2011); Varin, Reid and Firth (2011)). Under typical regularity conditions, the composite likelihood estimator is asymptotically normal with mean θ^* and variance $\{G(\theta^*, w)\}^{-1}$, where

$$G(\theta^*, w) = H(\theta^*, w)\{K(\theta^*, w)\}^{-1} H(\theta^*, w) \tag{1.2}$$

is the so-called Godambe information matrix, and $H(\theta, w) = -E\{\nabla U(\theta, w; X)\}$ and $K(\theta, w) = \mathrm{cov}\{U(\theta, w; X)\}$ are the $p \times p$ sensitivity and variability matrices, respectively. Although the maximum composite likelihood estimator is consistent, in general, $G(\theta^*, w)$ differs from the Fisher information $\mathrm{cov}\{\nabla \log f(X; \theta^*)\}$, with the two coinciding only in special cases where $H(\theta^*, w) = K(\theta^*, w)$.

The choice of w determines both the statistical properties and the computational efficiency of the composite likelihood estimator (Lindsay, Yi and Sun

(2011); Xu and Reid (2011); Huang et al. (2020)). On the one hand, the established theory of unbiased estimating equations suggests finding w that maximizes $tr\{G(\theta^*, w)\}$ (Heyde (2008, Chap. 2)). Although theoretically appealing, these optimal weights depend on an inversion of the score covariance matrix, the estimates of which are often singular. Several selection strategies have been proposed that balance the trade-off between statistical efficiency and computing cost. A common practice is to retain all feasible sub-likelihoods with $w_j = 1$, for all $j \geq 1$. However, this is undesirable from the viewpoints of computational parsimony and statistical efficiency, because the presence of too many correlated scores inflates the variability matrix K (Cox and Reid (2004); Ferrari, Qian and Hunter (2016)). A smaller subset may be selected by setting some w_j equal to zero, but determining a suitable subset remains challenging. Dillon and Lebanon (2010) and Ferrari, Qian and Hunter (2016) develop stochastic approaches in which sub-likelihoods are sampled according to a statistical efficiency criterion. Ad-hoc methods have also been developed; for example, in spatial data analysis, it is often convenient to consider sub-likelihoods corresponding to close-by observations; see Heagerty and Lele (1998), Sang and Genton (2014) and Bevilacqua and Gaetan (2015).

Motivated by this gap in the literature, we develop a methodology for selecting sparse composite likelihoods in large problems by retaining only the most informative scores in the estimating equations (1.1), while dropping the noisy terms. To this end, we propose minimizing the distance between the maximum likelihood score and the composite likelihood score, subject to a computing cost constraint.

The remainder of the paper is organized as follows. In Section 2, we describe the main sub-likelihood selection and combination methodology. In Section 3, we discuss the properties of our method. Theorem 2 shows that the proposed empirical composition rule is asymptotically equivalent to the optimal composition rule based on all available scores, and Theorems 3 and 4 give the consistency and asymptotic normality, respectively, of the resulting parameter estimator. In Section 4, we present examples related to common families of models. In Section 5, we apply our method to real Covid-19 epidemiological data. Finally, Section 6 concludes the paper.

2. Main Methodology

2.1. Penalized score distance minimization

We propose solving equation (1.1) with weights $w = w_\lambda(\theta)$ selected by minimizing the penalized score distance

$$\frac{1}{2} E \left\| U^{ML}(\theta; X) - U(\theta, w; X) \right\|_2^2 + \lambda \sum_{j=1}^{m} |w_j| , \tag{2.1}$$

where $U^{ML}(\theta; x) = \nabla \log f(x; \theta)$ is the maximum likelihood score, $\|\cdot\|_2$ denotes the L_2-norm, and $\lambda \geq 0$ is a regularization parameter. The resulting minimizer, say $w_\lambda(\theta)$, is then used to estimate the parameters by solving the composite likelihood estimating equation (1.1) in θ with $w = w_\lambda(\theta)$.

The vector of coefficients that minimizes (2.1) is allowed to contain positive, negative, or zero values, although negative elements do not cause any concerns in our method. The size of such coefficients is expected to be larger for sub-likelihoods that are strongly correlated with the full likelihood. Thus, a negative coefficient associated with a certain sub-likelihood score does not imply that this sub-likelihood is less informative, but simply means that it has a negative correlation with the maximum likelihood score.

The first term in the objective (2.1) improves the statistical efficiency by finding a composite score close to the maximum likelihood score. Note that minimizing the first term alone (when $\lambda = 0$) corresponds to finding finite-sample optimal O_F-optimal estimating equations. This criterion formalizes the idea of minimizing the variance in estimating equations; see (Heyde (2008, Chap. 1)). Lindsay, Yi and Sun (2011) point out that this type of criterion is suitable in the context of composite likelihood estimation, but that a general computational procedure to minimize such a criterion in large problems is still an open problem.

The term $\lambda \sum_{j=1}^{m} |w_j|$ is a penalty that discourages overly complex scores. The geometric properties of the L_1-norm penalty ensure that several elements in the solution $w_\lambda(\theta)$ are zero for sufficiently large λ, thus simplifying the resulting estimating equations. This is a key property of the proposed approach, exploited to reduce the computation burden.

The optimal solution $w_\lambda(\theta)$ may be interpreted as one that maximizes statistical accuracy, subject to a given computing cost. Alternatively, $w_\lambda(\theta)$ may be viewed as a composition rule that minimizes the likelihood complexity, subject to some efficiency loss compared with the maximum likelihood. The constant λ balances the trade-off between statistical efficiency and computational cost: $\lambda = 0$ is optimal in terms of asymptotic efficiency, but offers no reduction in the likelihood complexity, whereas high values of $\lambda > 0$ may result in a loss of informative data subsets and their scores.

There are two difficulties related to the direct minimization of (2.1): the presence of the intractable likelihood score function U^{ML}, and the expectation that depends on the unknown parameter θ. Up to an additive term independent of w, the penalized score distance in (2.1) can be expressed as

$$\frac{1}{2} E \left\| U(\theta, w; X) \right\|_2^2 - E \left\{ U^{ML}(\theta; X)^\top U(\theta, w; X) \right\} + \lambda \sum_{j=1}^{m} |w_j| \,. \tag{2.2}$$

Let $M(\theta; X) = \{U_1(\theta; X), \ldots, U_m(\theta; X)\}$ be a $p \times m$ matrix with columns given by the $p \times 1$ score vectors $U_j(\theta; X)$ $(j = 1, \ldots, m)$. Then, the first term of (2.2)

may be expressed as $w^\top J(\theta)w/2$, where $J(\theta)$ is the $m \times m$ score covariance matrix

$$J(\theta) = E\{M(\theta; X)^\top M(\theta; X)\}.$$

Note that at $\theta = \theta^*$, assuming unbiased scores with $E\{U_j(\theta^*; X)\} = 0$ $(j = 1, \ldots, m)$, the second Bartlett equality gives $E\{U^{ML}(\theta^*; X)U_j(\theta^*; X)^\top\} = E\{U_j(\theta^*; X)U_j(\theta^*; X)^\top\} = -E\{\nabla U_j(\theta^*; X)\}$. This implies that the second term in (2.2) is $-w^\top \mathrm{diag}\{J(\theta^*)\}$, where $\mathrm{diag}(A)$ denotes the diagonal vector of the matrix A. Therefore, (2.2) may be approximated by

$$d_\lambda(\theta, w) = \frac{1}{2} w^\top J(\theta) w - w^\top \mathrm{diag}\{J(\theta)\} + \lambda \sum_{j=1}^{m} |w_j| \,. \tag{2.3}$$

For n independent observations $X^{(1)}, \ldots, X^{(n)}$ on X, we obtain the empirical composition rule $\hat{w}_\lambda(\theta)$ by minimizing the empirical criterion

$$\hat{d}_\lambda(\theta, w) = \frac{1}{2} w^\top \hat{J}(\theta) w - w^\top \mathrm{diag}\{\hat{J}(\theta)\} + \lambda \sum_{j=1}^{m} |w_j| \,, \tag{2.4}$$

where $\hat{J}(\theta) = n^{-1} \sum_{i=1}^{n} M(\theta; X^{(i)})^\top M(\theta; X^{(i)})$.

The final composite likelihood estimator is found by replacing $w = \hat{w}_\lambda(\theta)$ in (1.1), and then solving the following estimating equation with respect to θ:

$$U\{\theta, \hat{w}_\lambda(\theta); X^{(1)}, \ldots, X^{(n)}\} = 0. \tag{2.5}$$

Although $\hat{w}_\lambda(\theta)$ is, in general, smooth in a neighborhood of θ^*, it may exibit a number of nondifferentiable points on the parameter space Θ. Therefore, the convergence of standard gradient-based root-finding algorithms, such as the Newton–Raphson algorithm, is not guaranteed.

To address this issue, we propose taking a preliminary root-n consistent estimate $\tilde{\theta}$ and finding the final estimator $\hat{\theta}_\lambda$ by solving the estimating equation

$$U\{\theta, \hat{w}_\lambda(\tilde{\theta}); X^{(1)}, \ldots, X^{(n)}\} = 0, \tag{2.6}$$

where $\hat{w}_\lambda(\tilde{\theta})$ is a quantity that depends fully on the data. A preliminary estimate is often easy to obtain, for example, by solving (1.1) with $w_j = 1$, for all $1 \leq j \leq m$. Alternatively, a computationally cheap root-n consistent estimate may be obtained by setting $w_j = 1$ for j in some random subset $S \subseteq \{1, \ldots, m\}$, and $w_j = 0$ otherwise.

2.2. Computational aspects and selection of λ

The numerical examples presented here are implemented as follows. We first compute the sparse composition rule $\hat{w}_\lambda(\tilde{\theta})$ by minimizing the convex criterion (2.4), with $\theta = \tilde{\theta}$ as a preliminary root-n consistent estimator, using the least-

angle regression algorithm (Efron et al. (2004)). Finally, we obtain the estimator $\hat{\theta}_\lambda$ using a one-step Newton–Raphson update, starting from $\theta = \tilde{\theta}$ and applied to (1.1); see Chapter 5 in Van der Vaart (2000) for an introduction to one-step estimation. As a preliminary estimator $\tilde{\theta}$, one may choose the composite likelihood estimator with the uniform composition rule $w_j = 1$, for all $j \geq 1$. From a theoretical viewpoint, any root-n consistent initial estimator $\tilde{\theta}$ leads to the same asymptotic results for the final estimate. We find that the affect of the initial estimate is negligible in many situations.

Similarly to the least-angle algorithm originally developed by Efron et al. (2004) in the context of sparse linear regression, each step of our implementation includes the score $U_j(\theta; X)$ that has the largest correlation with the residual difference $U_j(\theta; X) - U(\theta, w; X)$, followed by an adjustment step on w. An alternative approach is to solve (2.4) with respect to θ with $w = w_\lambda(\theta)$ using a Newton–Raphson algorithm. Here, $w_\lambda(\theta)$ is updated in each iteration using a coordinate-descent approach, as in Wu and Lange (2008).

The selection of λ is of practical importance because it balances the trade-off between statistical and computational efficiency. In many practical applications, this choice depends on available computing resources and the objective of the analysis. Although we do not seek a universal approach for selection, we consider the following heuristic strategy. Taking a grid Λ of values for λ corresponding to different numbers of selected scores, we consider $\hat{\lambda} = \max\{\lambda \in \Lambda : \phi(\lambda) > \tau\}$, for some user-specified constant $0 < \tau \leq 1$, where $\phi(\lambda) = \mathrm{tr}(\hat{J}_\lambda)/\mathrm{tr}(\hat{J})$. Here, $\hat{J}_\lambda$ denotes the empirical covariance matrix for the selected partial scores evaluated at $\tilde{\theta}$, and $\hat{J}$ is the covariance for all scores. Thus, $\phi(\lambda)$ can be viewed as the approximate proportion of the score variance explained by the selected scores. In practice, one may choose τ to be a sufficiently large value, such as $\tau = 0.75$ or $\tau = 0.90$.

When the covariance for all scores $\hat{J}$ is difficult to obtain because of an excessive computational burden, we propose using an upper bound of $\phi(\lambda)$ instead. Let $\tilde{\lambda} \in \Lambda$ be the next value for $\lambda \in \Lambda$ smaller than $\hat{\lambda}$; that is, we set $\tilde{\lambda} = \hat{\lambda}$ if $\hat{\lambda} = \min(\Lambda)$, and $\tilde{\lambda} = \max\{\lambda \in \Lambda : \lambda < \hat{\lambda}\}$ otherwise. Note that $\phi(\hat{\lambda}) < \mathrm{tr}(\hat{J}_{\hat{\lambda}})/\mathrm{tr}(\hat{J}_{\tilde{\lambda}})$, where the right-hand side represents the relative proportion of the score variance explained by reducing λ from $\hat{\lambda}$ to $\tilde{\lambda}$. Thus, in practice, one may take $\hat{\lambda}$ such that $\mathrm{tr}(\hat{J}_{\hat{\lambda}})/\mathrm{tr}(\hat{J}_{\tilde{\lambda}}) > \delta$, for some relative tolerance level $0 < \delta < 1$.

3. Properties

3.1. Conditions for uniqueness

This section gives an explicit expression for the minimizer of the penalized score distance criterion, and provides sufficient conditions for its uniqueness. The main requirement for uniqueness is that each partial score cannot be fully

determined by a linear combination of other scores. Specifically, we require the following condition:

Condition 1. Define $U_j = U_j(\theta, X)$. For any $\lambda > 0$ and $\theta \in \Theta$, the random vectors $(U_1^\top, U_1^\top U_1 \pm \lambda), \ldots, (U_m^\top, U_m^\top U_m \pm \lambda)$ are linearly independent.

Note that the condition is satisfied automatically, unless some partial score is perfectly correlated with the others, which is rarely the case in practice.

For a vector $a \in \mathbb{R}^m$, we use $a_{\mathcal{E}}$ to denote the sub-vector corresponding to index $\mathcal{E} \subseteq \{1, \ldots, m\}$, and $A_{\mathcal{E}}$ denotes the sub-matrix of the squared matrix A formed by taking the rows and columns corresponding to $\mathcal{E}$. We use $\mathrm{sign}(w)$ to denote the vector sign function, with the jth element taking the value -1, 0, or 1 if $w_j < 0$, $w_j = 0$, or $w_j > 0$, respectively. Let $\eta = 0$ if $\hat{J}(\theta)$ is positive definite, or else $\eta = \max_{x \in \mathbb{R}^q} [\mathrm{diag}\{\hat{J}(\theta)\}^\top V(\theta) x / \|V(\theta)x\|_1]$, where q denotes the number of zero eigenvalues of $\hat{J}(\theta)$, and $V(\theta)$ is an $m \times q$ matrix that collects the eigenvectors corresponding to zero eigenvalues.

Theorem 1. *Under Condition* 1*, for any $\theta \in \Theta$ and $\lambda > \eta$, the minimizer of the penalized distance $\hat{d}_\lambda(\theta, w)$ defined in* (2.4) *is unique with probability one, and is given by*

$$\hat{w}_{\lambda,\hat{\mathcal{E}}}(\theta) = \left\{ \hat{J}_{\hat{\mathcal{E}}}(\theta) \right\}^{-1} \left[diag\left\{ \hat{J}_{\hat{\mathcal{E}}}(\theta) \right\} - \lambda \; sign\{\hat{w}_{\lambda,\hat{\mathcal{E}}}(\theta)\} \right], \quad \hat{w}_{\lambda,\setminus\hat{\mathcal{E}}}(\theta) = 0,$$

where $\hat{\mathcal{E}} \subseteq \{1, \ldots, m\}$ is the index set defined as

$$\hat{\mathcal{E}} = \left\{ j : \left| n^{-1} \sum_{i=1}^{n} U_j(\theta; X^{(i)})^\top U_j(\theta; X^{(i)}) \right. \right. \\ \left. \left. - n^{-1} \sum_{i=1}^{n} U_j(\theta; X^{(i)})^\top U(\theta, \hat{w}_\lambda(\theta); X^{(i)}) \right| \geq \lambda \right\}, \tag{3.1}$$

and $\setminus\hat{\mathcal{E}}$ denotes the complement index set $\{1, \ldots, m\} \setminus \hat{\mathcal{E}}$. Moreover, $\hat{w}_{\lambda,\hat{\mathcal{E}}}(\theta)$ contains at most $np \wedge m$ nonzero elements.

When $\lambda = 0$, we have $\hat{\mathcal{E}} = \{1, \ldots, m\}$, meaning that the corresponding composition rule $\hat{w}_{\lambda,\hat{\mathcal{E}}}(\theta)$ does not contain any zero elements. In this case, we require the empirical score covariance matrix $\hat{J}(\theta)$ to be nonsingular, which is certainly violated when $np < m$. Even for the case $np > m$, $\hat{J}(\theta)$ may be singular because of the presence of largely correlated partial scores. On the other hand, setting $\lambda > \eta$ always gives a nonsingular score covariance matrix and guarantees the existence of $\hat{w}_{\lambda,\hat{\mathcal{E}}}(\theta)$. For sufficiently large λ, a relatively small subset of scores is selected. The formula in (3.1) suggests that the jth score is selected when it contributes enough information to the overall composite likelihood. In particular, the jth score is included if the estimated absolute difference between its Fisher

information $E\{U_j(\theta;X)^\top U_j(\theta;X)\}$ and the covariance with the overall composite likelihood score $E\{U_j(\theta;X)^\top U(\theta,w;X)\}$ is greater than λ.

3.2. Asymptotic optimality of the empirical composition rule

In this section, we investigate the asymptotic behavior of the sparse composition rule $\hat{w}_\lambda(\theta)$ as the sample size n diverges. The main result is the convergence of $\hat{w}_\lambda(\theta)$ to the ideal composition rule $w_\lambda(\theta)$, which is defined as the minimizer of the population criterion $d_\lambda(\theta,w)$ specified in (2.3). Letting $\lambda \to 0$ as n increases implies that the sparse rule $\hat{w}_\lambda(\theta)$ is asymptotically equivalent to the optimal rule $w_0(\theta)$ in terms of the variance of the estimator, with the latter, however, involving all m scores. To show this, we introduce an additional technical requirement on the covariance between the sub-likelihood scores.

Condition 2. For all $j,k \geq 1$, $sup_{\theta\in\Theta}|\hat{J}(\theta)_{jk} - J(\theta)_{jk}| \to 0$ in probability as $n \to \infty$, where $\hat{J}(\theta)_{jk}$ and $J(\theta)_{jk}$ are the $\{j,k\}$th elements of $\hat{J}(\theta)$ and $J(\theta)$, respectively. Moreover, each element of $J(\theta)$ is continuous and bounded, and the smallest eigenvalue of $J(\theta)$ is uniformly bounded away from zero for all $\theta \in \Theta$.

Theorem 2. *Under Conditions* 1 *and* 2*, for any* $\lambda > 0$ *and* $\theta \in \Theta$*, we have* $\sup_{\theta\in\Theta} \|\hat{w}_\lambda(\theta) - w_\lambda(\theta)\|_1 \to 0$ *in probability, as* $n \to \infty$*.*

Because the preliminary estimate $\tilde{\theta}$ is consistent, the continuity of $w_\lambda(\theta)$ (shown in Lemma A.2 in the Appendix) implies immediately that $\hat{w}_\lambda(\tilde{\theta})$ converges to $w_\lambda(\theta^*)$; that is, the empirical composition rule converges to the ideal composition rule evaluated at the true parameter. Theorem 2 implies that the proposed sparse composite likelihood score is a suitable approximation for the optimal score involving m sub-likelihoods. Specifically, under regularity conditions, we have

$$\sup_{\theta\in\Theta} \left\| \frac{1}{n}\sum_{i=1}^{n} U\{\theta, \hat{w}_\lambda(\tilde{\theta}); X^{(i)}\} - \frac{1}{n}\sum_{i=1}^{n} U\{\theta, w_0(\theta^*); X^{(i)}\} \right\|_1 \to 0 \tag{3.2}$$

in probability as $n \to \infty$ and $\lambda \to 0$. Note that it is often difficult to compute the optimal composition rule $w_0(\theta^*)$ and the related Godambe information matrix $G\{\theta, w_0(\theta^*)\}$, owing to the need to invert the $m \times m$ score covariance matrix with entries $E\{U_j(\theta;X)^\top U_k(\theta;X)\}$, for $1 \leq j,k \leq m$. On the other hand, the sparse composition rule $\hat{w}_\lambda(\tilde{\theta})$ and the implied Godambe information have the advantage of being computationally tractable, because they involve only a fraction of the scores.

3.3. Limit behavior of the estimator $\hat{\theta}_\lambda$ and the standard errors

The final estimator $\hat{\theta}_\lambda$ is an M-estimator that solves estimating equations of the form

$$\frac{1}{n}\sum_{i=1}^{n} U\{\theta, \hat{w}_\lambda(\tilde{\theta}); X^{(i)}\} = 0. \tag{3.3}$$

Because the vector $\hat{w}_\lambda(\tilde{\theta})$ converges to $w_\lambda^* = w_\lambda(\theta^*)$, by Theorem 2 and Lemma A.2, the limit of (3.3) may be written as

$$E\{U(\theta, w_\lambda^*; X)\} = 0. \tag{3.4}$$

To show that $\hat{\theta}_\lambda$ is consistent for the solution of (3.4), we assume additional regularity conditions to ensure a unique root of the ideal composite likelihood score and stochastic equicontinuity on each of the sub-likelihood scores.

Condition 3. For all constants $c > 0$, $\inf_{\{\theta:\|\theta-\theta^*\|_1 \geq c\}} \|E\{U(\theta, w_\lambda^*; X)\}\|_1 > 0 = \|E\{U(\theta^*, w_\lambda^*; X)\}\|_1$. Moreover, assume $\sup_{\theta\in\Theta}\|\sum_{i=1}^n U_j(\theta; X^{(i)})/n - E\{U_j(\theta; X)\}\|_1 \to 0$ in probability as $n \to \infty$, for all $1 \leq j \leq m$.

Theorem 3. *Under Conditions* 1–3, $\hat{\theta}_\lambda$ *converges in probability to* θ^*.

To obtain the asymptotic normality of the final estimator $\hat{\theta}_\lambda$, we assume the following condition for the sub-likelihood scores.

Condition 4. Assume for all $j \geq 1$, $\mathrm{var}\{U_j(\theta^*; X)\} < \infty$. In a neighborhood of θ^*, each $U_j(\theta; x)$ is twice continuously differentiable in θ, and the partial derivatives are dominated by some fixed integrable functions depending only on x. Moreover, assume $H(\theta^*, w_\lambda^*)$ defined in (1.2) is nonsingular.

Theorem 4. *Under Conditions* 1–4, *we have*

$$n^{1/2} G_\lambda(\theta^*)^{1/2}\left(\hat{\theta}_\lambda - \theta^*\right) \to N_p(0, I_p) \tag{3.5}$$

as $n \to \infty$, *where* $G_\lambda(\theta) = G(\theta, w_\lambda^*)$ *is the* $p \times p$ *Godambe information matrix defined in* (1.2).

We estimate the $p \times p$ Godambe information matrix $G_\lambda(\theta)$ in (3.5) using the sandwich estimator $\hat{G}_\lambda = \hat{H}_\lambda \hat{K}_\lambda^{-1} \hat{H}_\lambda$, and obtain the $p \times p$ matrices $\hat{H}_\lambda$ and $\hat{K}_\lambda$ by replacing $\theta = \hat{\theta}_\lambda$ in

$$\hat{H}_\lambda(\theta) = -\frac{1}{n}\sum_{i=1}^{n} \nabla U\{\theta, \hat{w}_\lambda(\tilde{\theta}); X^{(i)}\},$$
$$\hat{K}_\lambda(\theta) = \frac{1}{n}\sum_{i=1}^{n} U\{\theta, \hat{w}_\lambda(\tilde{\theta}); X^{(i)}\} U\{\theta, \hat{w}_\lambda(\theta); X^{(i)}\}^\top.$$

Practical advantages of using the sparse composition rule $\hat{w}_\lambda(\theta)$ are the reduction of the computational cost and the increased stability of the standard error calculations. Although the score variance matrix $\hat{K}_\lambda(\theta)$ may be difficult to obtain when $\lambda = 0$, owing to potentially $O(m^2)$ covariance terms, choosing a sufficiently

large value for $\lambda > 0$ avoids this situation by reducing the number of terms in the composite likelihood score.

4. Examples

4.1. Common location for heterogeneous variates

Let $X \sim N_m(\theta 1_m, \Sigma)$, where the $m \times m$ covariance matrix Σ has off-diagonal elements σ_{jk} $(j \neq k)$ and diagonal elements σ_k^2 $(j = k)$. Computing the maximum likelihood estimator of θ requires Σ^{-1}, and Σ is usually estimated by the sample covariance $\hat{\Sigma}$. However, when $n < m$, the maximum likelihood estimator is not available, because the sample covariance $\hat{\Sigma}$ is singular; on the other hand, the composite likelihood estimator is still feasible. The jth marginal score is $U_j(\theta; x) = (x_j - \theta)/\sigma_j^2$, and the composite likelihood estimating equation based on the sample $X^{(1)}, \ldots, X^{(n)}$ is

$$0 = \sum_{j=1}^{m} \frac{w_j}{\sigma_j^2} \sum_{i=1}^{n} (X_j^{(i)} - \theta). \tag{4.1}$$

Then, the population and empirical score covariances are $m \times m$ matrices with jkth entries

$$J(\theta)_{jk} = E\left\{ \frac{(X_j - \theta)(X_k - \theta)}{\sigma_j^2 \sigma_k^2} \right\} \text{ and } \hat{J}(\theta)_{jk} = \frac{1}{n} \sum_{i=1}^{n} \left\{ \frac{(X_j^{(i)} - \theta)(X_k^{(i)} - \theta)}{\sigma_j^2 \sigma_k^2} \right\},$$

respectively.

It is useful to inspect the special case in which X has independent components ($\sigma_{jk} = 0$, for all $j \neq k$). This setting corresponds to the fixed-effect meta-analysis model that combines estimators from m independent studies to improve accuracy. From Theorem 1, the optimal composition rule is

$$\hat{w}_{\lambda,j}(\theta) = \left\{ 1 - \frac{\lambda n \sigma_j^4}{\sum_{i=1}^{n} (X_j - \theta)^2} \right\} I\left\{ \frac{\sum_{i=1}^{n} (X_j^{(i)} - \theta)^2}{n\sigma_j^4} > \lambda \right\}.$$

The optimal population composition rule $w_{\lambda,j}(\theta)$ is the same as the above expression, with the sample averages replaced by expectations. The composition rule $w_{\lambda,j}(\theta)$ evaluated at the true parameter is $w_{\lambda,j}^* = (1 - \lambda\sigma_j^2) I(\sigma_j^{-2} > \lambda)$ $(j = 1, \ldots, m)$. This highlights that overly noisy data subsets with variance $\sigma_j^2 > \lambda^{-1}$ are dropped, and thus do not influence the final estimator for θ. In particular, the number of nonzero elements in w_λ^* is $\sum_{j=1}^{m} I(\sigma_j^2 < \lambda^{-1})$. Finally, substituting $w_j = \hat{w}_{\lambda,j}(\theta)$ in (4.1) gives the following fixed-point equation:

$$\theta = \left\{ \sum_{j=1}^{m} \frac{\hat{w}_{\lambda,j}(\theta)}{\sigma_j^2} \bar{X}_j \right\} \Big/ \left\{ \sum_{k=1}^{m} \frac{\hat{w}_{\lambda,k}(\theta)}{\sigma_k^2} \right\}, \tag{4.2}$$

which is a weighted average of the marginal sample means $\bar{X}_j = n^{-1}\sum_{i=1}^{n} X_j^{(i)}$ $(j = 1, \ldots, m)$. We obtain the final composite likelihood estimator $\hat{\theta}_\lambda$ by solving equation (4.2).

When $\lambda = 0$, we have uniform weights $w_0^* = (1, \ldots, 1)^\top$ and the corresponding estimator $\hat{\theta}_0$ is the usual optimal meta-analysis solution. Although the implied estimator $\hat{\theta}_0$ has a minimum variance, it offers no control over the overall computational cost, because all m sub-scores are selected. On the other hand, choosing $\lambda > 0$ judiciously may lead to a low computational burden with negligible efficiency loss for the resulting estimator. For instance, assuming $\sigma_j^2 = j^2$, we have

$$\frac{1}{2}E\left\{U(\theta, w_\lambda^*; X) - U(\theta, w_0^*; X)\right\}^2 \leq \lambda^2 \sum_{j\in\mathcal{E}} j^2 + \sum_{j\notin\mathcal{E}} j^{-2}, \tag{4.3}$$

where $\mathcal{E} = \{j : j^2 < \lambda^{-1}\}$ and θ is the true parameter. Because the number of selected scores is $\sum_{j=1}^{m} I\left(j^2 < \lambda^{-1}\right) = \lfloor \lambda^{-1/2} \rfloor$, we can write $\lambda^2 \sum_{j\in\mathcal{E}} j^2 \leq \lambda^2\lambda^{-1}\lambda^{-1/2} = \lambda^{1/2}$, which converges to zero as $\lambda \to 0$; additional calculations show that $\sum_{j\notin\mathcal{E}} j^{-2} \to 0$ as $\lambda \to 0$. Hence, the left-hand side of (4.3) goes to zero as long as $\lambda \to 0$. This suggests that the truncated composite likelihood score suitably approximates the optimal score, while containing relatively few terms. If the elements of X are correlated with $\sigma_{jk} \neq 0$, for $j \neq k$, the partial scores contain overlapping information on θ. In this case, discarding highly correlated partial scores would improve the computational cost, while maintaining satisfactory statistical efficiency for the final estimator.

Figure 1 shows the solution path of w_λ^*. That is, it shows the trajectory of the elements of the optimal composition rule w_λ^* for different values of λ, and the asymptotic relative efficiency of the corresponding composite likelihood estimator $\hat{\theta}_\lambda$ compared with the maximum likelihood estimator for different values of λ. When m is large ($m = 1{,}000$), the asymptotic relative efficiency drops gradually until only a few scores remain. This example shows that we can achieve relatively high efficiency by using the selected composite likelihood equations, when a few partial scores contain the majority of information about θ. In such cases, the final estimator $\hat{\theta}_\lambda$ with a sparse composition rule is expected to achieve a good trade-off between computational cost and statistical efficiency.

4.2. Covariance estimation

Suppose X follows a multivariate normal distribution with zero mean vector and covariance $\Sigma(\theta)$, with elements $\Sigma(\theta)_{jk} = \exp(-\theta\delta_{jk})$ $(j \neq k)$ and $\Sigma(\theta)_{jk} = 1$ $(j = k)$. The quantity δ_{jk} may be regarded as the distance between the spatial locations j and k. Here, the case of equally distant points corresponds to the covariance estimation for exchangeable variables described in Cox and Reid

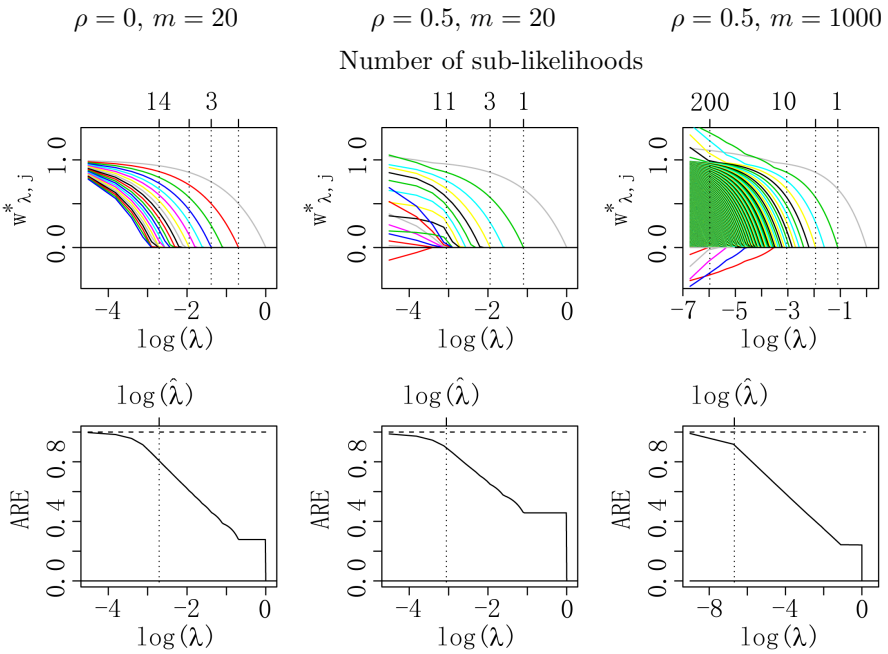

Figure 1. Top Row: Solution paths for the minimizer w^*_λ of $d_\lambda(\theta, w)$ defined in (2.3) at the true parameter for different values of λ, with the corresponding numbers of sub-likelihoods shown on the top axis. Bottom Row: Asymptotic relative efficiency (ARE) of the estimator $\hat{\theta}_\lambda$ compared with that of the maximum likelihood estimator. The vertical dashed lines represent $\hat{\lambda}$ selected as described in Section 2.2, using $\tau = 0.9$. The results correspond to the location model $X \sim N_m(\theta 1_m, \Sigma)$, with $\Sigma_{jk} = j$ $(j = k)$ and $\Sigma_{jk} = \rho(jk)^{1/2}$ $(j \neq k)$.

(2004). The maximum composite likelihood estimator solves

$$
\begin{aligned}
0 &= \sum_{j<k} w_{jk} \sum_{i=1}^{n} U_{jk}(\theta; X_j^{(i)}, X_k^{(i)}) \\
&= \sum_{j<k} w_{jk} \sum_{i=1}^{n} \left[\frac{\Sigma(\theta)_{jk}\{{X_j^{(i)}}^2 + {X_k^{(i)}}^2 - 2X_j^{(i)}X_k^{(i)}\Sigma(\theta)_{jk}\}}{\{1 - \Sigma(\theta)_{jk}^2\}^2} \right] \Sigma(\theta)_{jk}\delta_{jk} \\
&\quad - \sum_{j<k} w_{jk} \sum_{i=1}^{n} \left\{ \frac{\Sigma(\theta)_{jk} + X_j^{(i)}X_k^{(i)}}{1 - \Sigma(\theta)_{jk}^2} \right\} \Sigma(\theta)_{jk}\delta_{jk},
\end{aligned}
$$

where $U_{jk}(\theta; x_j, x_k)$ is the score of a bivariate normal distribution for the pair (X_j, X_k). Figure 2 shows the solution path of the optimal composition rule w^*_λ for different values of λ, and the asymptotic relative efficiency of the estimator $\hat{\theta}_\lambda$ compared with the maximum likelihood estimator. Here, we consider variable pairs ranging from $m = 45$ to $m = 1{,}225$. When $\lambda = 0$, the proposed estimator has relatively high asymptotic efficiency. Interestingly, the efficiency stays at

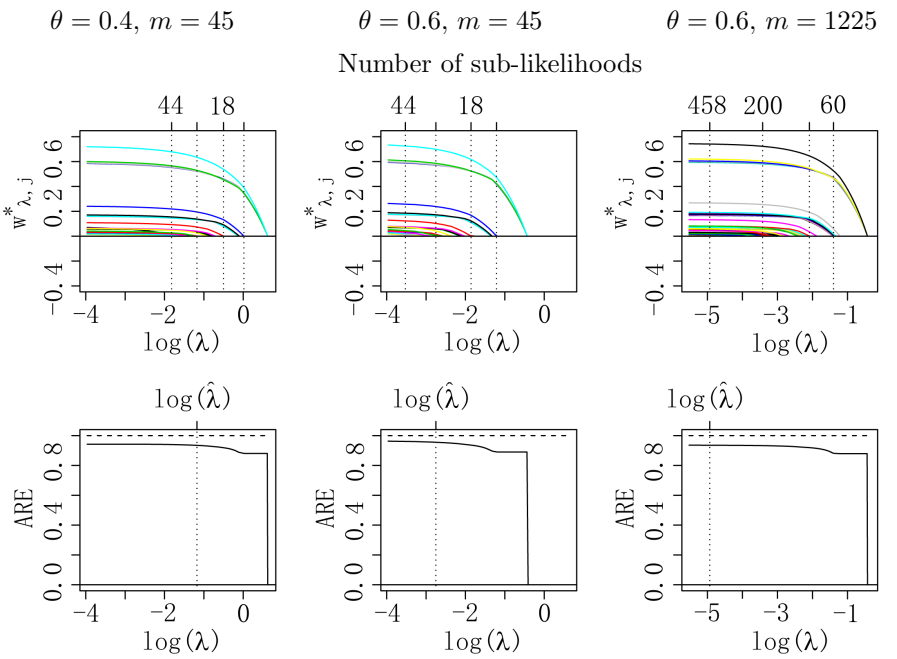

Figure 2. Top Row: Solution paths for the minimizer w^*_λ of $d_\lambda(\theta, w)$ defined in (2.3) at the true parameter for different values of λ, with the corresponding numbers of sub-likelihoods shown on the top axis. Bottom Row: Asymptotic relative efficiency (ARE) of the estimator $\hat{\theta}_\lambda$ compared with that of the maximum likelihood estimator. The vertical dashed lines correspond to $\hat{\lambda}$ selected as described in Section 2.2, with $\tau = 0.9$. The results correspond to the model $X \sim N_d(0, \Sigma(\theta))$, with $\Sigma(\theta)_{jk} = \exp\{-\theta(2|j-k|)^{1/2}\}$.

around 90% until only a few sub-likelihoods remain, suggesting that a very small proportion of partial-likelihood components contain most of the information about θ. In such cases, the proposed selection procedure is useful, because it reduces the computational burden, while retaining satisfactory efficiency for the final estimator.

4.3. Location estimation for exchangeable variates

For $X \sim N_m(\theta 1_m, \Sigma)$, with $\Sigma = (1-\rho)I_m + \rho 1_m 1_m^\top$, for $0 < \rho < 1$, the marginal scores $U_j(\theta; X) = X_j - \theta$ $(j = 1, \ldots, m)$ are identically distributed with equal correlation. As $n \to \infty$, the optimal composition rule converges to

$$w^*_{\lambda,j} = \frac{1-\lambda}{\rho(m-1)+1} I(\lambda < 1) \quad (j = 1, \ldots, m),$$

so the corresponding composite likelihood estimator is $\hat{\theta}_\lambda = \sum_{j=1}^m \bar{X}_j/m$ and is independent of λ. This suggests that the partial scores are selected randomly in the empirical composition rule $\hat{w}_\lambda(\theta)$. However, we can take a sufficiently large

value for λ, such that the sparse solution containing only a few zero elements still ensures relatively high statistical efficiency for the corresponding estimator $\hat{\theta}_\lambda$. To see this, first note that the eigenvalue of the score covariance $J(\theta)$ is $\rho(m-1)+1$, while the remaining $m-1$ eigenvalues are all equal to $1-\rho$, suggesting that the first score contains a relatively large amount of information on θ compared with that of the other scores. Furthermore, because $\mathrm{var}(\hat{\theta}_\lambda) = \{\rho(m-1)+1\}/(mn)$, the asymptotic relative efficiency of the composite likelihood with $m < \infty$ compared with that with $m \to \infty$ is $\rho m/\{\rho(m-1)+1\}$; this is 0.83, 0.90, and 0.98 for $m = 5$, 9, and 50, respectively, when $\rho = 0.75$.

5. Real-Data Example: Spatial Covariance Estimation for Covid-19 Data

In this section, we apply the proposed methodology to public health data on the Covid-19 pandemic supplied by the Italian Civil Protection Department. The data consist of $n = 60$ observations of daily new Covid-19 cases for the period February 24 to April 23, 2020, across $d = 90$ Italian provinces corresponding to capital cities in regions or autonomous territories. Let $X_j^{(i)} = \mu_j^{(i)} + \varepsilon_j^{(i)}$ be the number of new Covid-19 cases observed on day i in province j, where $\mu_j^{(i)}$ is a location-specific deterministic trend and $\sigma_j^2 = \mathrm{var}(\varepsilon_j^{(i)})$. Assume normally distributed errors $\varepsilon_1^{(i)}, \ldots, \varepsilon_m^{(i)}$, with covariance specified by the gravity model

$$\mathrm{cov}(\varepsilon_j^{(i)}, \varepsilon_k^{(i)}) = \sigma_j \sigma_k \exp\left\{-\frac{\theta t_{jk}}{(m_j m_k)}\right\} \quad (j,k = 1, \ldots, d), \tag{5.1}$$

where m_j denotes the population size (in millions) of the jth site, and t_{jk} is the distance between sites, computed as $t_{jk} = \{(\mathrm{lat}_j - \mathrm{lat}_k)^2 + (\mathrm{lon}_j - \mathrm{lon}_k)^2\}^{1/2}$, where lat_j and lon_j represent the latitude and longitude, respectively, of the jth site. Then, each $X^{(i)} = (X_1^{(1)}, \ldots, X_d^{(i)})$ may be regarded as a set of d observations on a Gaussian random field, with the exponential spatial covariance function given in (5.1); see Bevilacqua and Gaetan (2015) for more information on pairwise likelihood estimations for Gaussian random fields. Clearly, correlation depends on population density and potential flows of people across pairs of provinces. To control for these aspects, we include both population size and distance between provinces in the gravity model specified in (5.1). This means that θ should be interpreted as a covariance parameter for new cases between provinces, conditional on population size and distance between provinces. Population size and distance are the two main factors often used when formulating distance decay laws, such as gravity models, to represent spatial interactions related to disease diffusion processes.

The main interest is in estimating the covariance parameter θ in order to monitor contagion across provinces. To this end, the data are first de-trended and normalized. The local trends $\hat{\mu}_j$ are estimated using a Nadaraya–Watson

Table 1. Estimates for the spatial covariance parameter for the Covid-19 data with corresponding standard errors and number of selected sub-likelihoods (# sub-likelihoods).

λ (10^5)	8.64	8.57	8.29	8.12	7.82	7.69	7.59	7.54	7.44	7.22	7.13
$\hat{\theta}_\lambda$ (10^{-2})	2.58	3.53	3.58	3.90	4.03	4.12	4.23	4.25	4.25	4.25	4.25
SE (10^{-2})	4.50	2.34	2.17	1.26	0.84	0.69	0.47	0.46	0.11	0.10	0.08
# sub-likelihoods	40	42	44	46	48	50	52	54	56	58	60

kernel smoother, implemented by the R function `ksmooth`. The error variance σ_j^2 is estimated by $\hat{\sigma}_j^2 = \sum_{i=1}^n (X_j^{(i)} - \hat{\mu}_j^{(i)})^2/n$. The covariance parameter θ is subsequently estimated using the pair-wise likelihood approach in Section 4.2, with $\delta_{jk} = t_{jk}/(m_j m_k)$. Table 1 shows estimates corresponding to a sequence of decreasing values for λ. As the number of pair-wise likelihood terms increases, the estimator $\hat{\theta}_\lambda$ tends to approach some stable value and its standard error decreases. For instance, taking $\tau = 0.75$, as defined in Section 2.2, the final estimate is $\hat{\theta}_\lambda = 4.25 \times 10^{-2}$, with standard error 1.01×10^{-3}, corresponding to 58 pairs of cities selected out of $\binom{90}{2} = 4{,}005$ pairs. Furthermore, both $\hat{\theta}_\lambda$ and the standard error converge when only about 58 sub-likelihoods have been selected. In comparison, the uniform composite likelihood estimator using all pairs of sites with equal weights is $\hat{\theta}_{\text{unif}} = 6.15 \times 10^{-2}$, with standard error 2.42×10^{-3}.

6. Conclusion

Composite likelihood inference plays an important role as a remedy to the drawbacks of traditional likelihood approaches, with advantages in terms of computing and modeling flexibility. Nevertheless, There does not appear to be a universal procedure for constructing composite likelihoods that is statistically justified and fast to execute (Lindsay, Yi and Sun (2011)). Motivated by this gap in the literature, we propose a selection methodology that results in composite likelihood estimating equations with good statistical properties. The selected equations are sparse for sufficiently large λ, meaning that they contain only the most informative sub-likelihood score terms. This sparsity-promoting mechanism is useful when the sub-likelihood scores are heterogeneous in terms of their information, or when the ideal O_F-optimal score is difficult to obtain. Remarkably, the sparse score is shown to approximate the O_F-optimal score in large samples under reasonable conditions; see Theorem 2 and Equation (3.2).

For implementation, we have proposed a selection criterion for choosing λ, which performed well in the examples, rather than providing a universal approach. In practice, it could be feasible to use any alternative criteria to choose λ, depending on the problem. For example, when the full score covariance is not available owing to its computational burden, one may use the upper bound provided in Section 2.2, or choose λ up to some given level of information gain. The latter is defined as the ratio of the smallest eigenvalue to the trace of

the current selected score covariance, which is decreasing with λ, by the min-max theorem in linear algebra. As another idea, by the Karush–Kuhn–Tucker condition of quadratic optimization, λ represents the norm of the estimated covariance between the current selected sub-score $U_j(\theta; X)$ and the residual $\{U(\theta, w; X) - U_j(\theta; X)\}$. One can choose λ such that the covariance is smaller than some predetermined value.

Building on the recent success of shrinkage methods for the full likelihood, many works have proposed using sparsity-inducing penalties in the composite likelihood framework; for example, see Bradic, Fan and Wang (2011), Xue, Zou and Ca (2012), and Gao and Carroll (2017). However, the spirit of our approach differs from these methods, because our penalty focuses on entire sub-likelihood functions, rather than on elements of θ. In contrast to the aforementioned approaches, our penalization strategy has the advantage of retaining asymptotically unbiased estimating equations, thus leading to desirable asymptotic properties of the related parameter estimator.

Several theoretical and applied avenues are open to future research. First, the current study focuses on the case where p is finite. However, penalties able to deal with situations where both m and p are allowed to grow with n may be useful for analyses of high-dimensional data. Implementations of the convex efficiency criterion (2.1) beyond the current i.i.d. setting offer another useful future research direction. For example, this would be valuable when analyzing spatial or spatio-temporal data, where often the overwhelming number of sub-likelihoods poses a challenge to traditional composite likelihood methods.

Acknowledgements

The authors acknowledge the financial support from the Italian Ministry MIUR under the PRIN project Hi-Di NET-Econometric Analysis of High Dimensional Models with Network Structures in Macroeconomics and Finance (grant 2017TA7TYC).

Appendix: Proofs

For convenience, we use $U_j(\theta; X^{(i)})$ for $U_j^{(i)}$ and let $M^{(i)}$ be the $p \times m$ matrix collecting $U_j^{(i)}$ for $j = 1, \ldots, m$. Let $M_{\hat{\varepsilon}}^{(i)}$ be the sub-matrix of $M^{(i)}$ with columns indexed by the set $\hat{\varepsilon}$ and denote $U^{(i)} = U\{\theta; \hat{w}_\lambda(\theta), X^{(i)}\} = M^{(i)}\hat{w}_\lambda(\theta)$.

Proof of Theorem 1. We first note that for any $\theta \in \Theta$, $\hat{d}_\lambda(\theta, w)$ is lower bounded, thus the minimizer exists. This is implied by taking the eigen decomposition of the real Hermitian matrix $\hat{J}(\theta)$ and then re-organize $\hat{d}_\lambda(\theta, w)$ as a summation of perfect square terms corresponding to nonzero eigenvalues, a non-negative first order term corresponding to zero eigenvalues and a constant. We also note that $U^{(i)}$ is unique due to the strict convexity of the first term of

$\hat{d}_\lambda(\theta, w)$ with respect to $U^{(i)}$, the convexity of the rest of the terms with respect to w, and the linearity of $U^{(i)}$ with respect to w. By the Karush-Kuhn-Tucker conditions for quadratic optimization, the solution must satisfy

$$\frac{1}{n}\sum_{i=1}^{n} {U_j^{(i)}}^\top U^{(i)} - \frac{1}{n}\sum_{i=1}^{n} {U_j^{(i)}}^\top U_j^{(i)} + \lambda\gamma_j = 0, \qquad \text{for } j = 1, \ldots, m, \tag{A.1}$$

where $\gamma_j = \text{sign}(w_j)$ if $w_j \neq 0$ and $\gamma_j \in [-1, 1]$ if $w_j = 0$. This implies that $\hat{\varepsilon}$ defined in (3.1) is unique. Note that the rank of $\hat{J}_{\hat{\varepsilon}} \equiv n^{-1}\sum_{i=1}^{n} {M_{\hat{\varepsilon}}^{(i)}}^\top M_{\hat{\varepsilon}}^{(i)}$ is at most $\min(m, np)$. Next we show that $\hat{J}_{\hat{\varepsilon}}$ has full rank. Otherwise, by the rank equality of the Gram matrix, there exists a subset $\tilde{\varepsilon} \subseteq \hat{\varepsilon}$, $|\tilde{\varepsilon}| \leq \min(m, np+1)$ and some $k \in \tilde{\varepsilon}$, such that $({U_k^{(1)}}^\top, \ldots, {U_k^{(n)}}^\top)$ can be written as a linear combination of $({U_j^{(1)}}^\top, \ldots, {U_j^{(n)}}^\top)$, for $j \in \tilde{\varepsilon}$ and $j \neq k$. Together with the Karush-Kuhn-Tucker condition, there exist constants a_j, $j \in \tilde{\varepsilon}$ and $j \neq k$ (with $a_j \neq 0$ for some j) such that $U_k^{(i)} = \sum_{j\in\tilde{\varepsilon}, j\neq k} a_j U_j^{(i)}$, for all $i = 1, \ldots, n$, and

$$\frac{1}{n}\sum_{i=1}^{n} {U_k^{(i)}}^\top U_k^{(i)} + \lambda\gamma_k = \sum_{j\in\tilde{\varepsilon}, j\neq k} a_j \frac{1}{n}\sum_{i=1}^{n} {U_j^{(i)}}^\top U_j^{(i)} + \lambda\gamma_j.$$

This represents a linear system with $(np + 1)$ equations but only $|\tilde{\varepsilon}| - 1$ degrees of freedom, meaning that the rank of the $(np + 1) \times |\tilde{\varepsilon}|$ matrix generated by columns $({U_j^{(1)}}^\top, \ldots, {U_j^{(n)}}^\top, (1/n)\sum_{i=1}^{n} {U_j^{(i)}}^\top U_j^{(i)} + \lambda\gamma_j)$, $j \in \tilde{\varepsilon}$ is smaller or equal to $|\tilde{\varepsilon}| - 1$. Since $|\tilde{\varepsilon}| \leq np + 1$, we have that the $|\tilde{\varepsilon}|$ columns are linearly dependent. Under Condition 1, this event has zero probability, which is a contradiction. The statement in the theorem then follows by solving the Karush-Kuhn-Tucker equations in (A.1).

Lemma A.1. *Under Conditions* 1 *and* 2, *for any* $\lambda > 0$, $\sup_{\theta\in\Theta} \|\hat{d}_\lambda\{\theta, \hat{w}_\lambda(\theta)\} - d_\lambda\{\theta, \hat{w}_\lambda(\theta)\}\|_1 \to 0$ *in probability, as* $n \to \infty$.

Proof of Lemma A.1. By definition, it suffices to show that for all $j, k \geq 1$, $sup_{\theta\in\Theta}|\hat{J}(\theta)_{jk} - J(\theta)_{jk}| \to 0$ in probability as $n \to \infty$, and that $\|\hat{w}_\lambda(\theta)\|_1$ is uniformly bouned with probability tending to one. The first part is ensured by Condition 2. For the second part, by Theorem 1, it suffices to show that $\hat{J}_{\hat{\epsilon}}(\theta)$ and $\hat{J}_{\hat{\epsilon}}(\theta)^{-1}$ are uniformly bounded entry-wise with probability tending to 1, which is guarenteed by the uniform convergence of $\hat{J}(\theta)$ in probability, the boundedness of each element of $J(\theta)$ and the invertibility of $\hat{J}_{\hat{\epsilon}}(\theta)$ accoridng to the min-max theorem in linear algebra.

Proof of Theorem 2. Recall that $\hat{w}_\lambda(\theta) = \text{argmin}_w \hat{d}_\lambda(\theta, w)$ and $w_\lambda(\theta) = \text{argmin}_w d_\lambda(\theta, w)$. Let ξ be the smallest eigenvalue of $J(\theta)$. By definition, we have

$$\begin{aligned}
&\sup_{\theta}\left[\frac{1}{2}\xi\|\hat{w}_\lambda(\theta)-w_\lambda(\theta)\|_2^2\right]\\
&\le \sup_{\theta}\left[\frac{1}{2}\{\hat{w}_\lambda(\theta)-w_\lambda(\theta)\}^T J(\theta)\{\hat{w}_\lambda(\theta)-w_\lambda(\theta)\}\right]\\
&\le \sup_{\theta}\left[|d\{\theta,\hat{w}_\lambda(\theta)\}-d\{\theta,w_\lambda(\theta)\}|\right]\\
&\le \sup_{\theta}\left[|d\{\theta,\hat{w}_\lambda(\theta)\}-\hat{d}\{\theta,\hat{w}_\lambda(\theta)\}|\right]+\sup_{\theta}\left[|\hat{d}\{\theta,w_\lambda(\theta)\}-d\{\theta,w_\lambda(\theta)\}|\right]
\end{aligned}$$

where the second inequality is due to the Karush-Kuhn-Tucker conditions of quadratic optimization, and the second last inequality is due to that $\hat{w}_\lambda(\theta)$ and $w_\lambda(\theta)$ are the corresponding minimizers. By Lemma A.1, the first term of the last inequality converges to zero in probability, and the same holds for the second term. Under Condition 2, $\xi > 0$. Since m is fixed, it concludes the proof.

Lemma A.2. *Under Conditions* 1 *and* 2, $w_\lambda(\theta)$ *is continuous with respect to both* λ *and* θ *on* $\lambda \ge 0$ *and* $\theta \in \Theta$.

Proof of Lemma A.2. For simplicity, here we show the continuity of $w_\lambda(\theta)$ with respect to θ. The proof for continuity with respect to λ is the same and thus omitted. For any $c > 0$ and $\theta_1 \in \Theta$, it suffices to show that there exist some $\delta > 0$, such that $\|\theta-\theta_1\|_1 < \delta$ implies $\|w_\lambda(\theta)-w_\lambda(\theta_1)\| < c$. To find δ, recall that $w_\lambda(\theta)$ is the minimizer of $d_\lambda(\theta, w)$ defined in (2.3). Under Condition 2, $d_\lambda(\theta, w)$ is strictly convex with respect to w. Thus, there exists $c_1 = \inf_{\{w:\|w-w_\lambda(\theta_1)\|_1=c\}} d_\lambda(\theta_1, w) > d_\lambda\{\theta_1, w_\lambda(\theta_1)\}$. Moreover, $d_\lambda(\theta, w)$ is uniformly continuous on the closed domain $\{w \in \mathbb{R}^m : \|w - w_\lambda(\theta_1)\|_1 \le c\} \times \{\theta \in \mathbb{R}^p : \|\theta - \theta_1\|_1 \le \delta_1\}$ for some $\delta_1 > 0$. Thus we can find $\delta \in (0, \delta_1)$ such that for any $\{\theta : \|\theta - \theta_1\|_1 < \delta\}$ and $\{w : \|w - w_\lambda(\theta_1)\|_1 \le c\}$, $\|d_\lambda(\theta, w) - d_\lambda(\theta_1, w)\|_1 < \{c_1 - d_\lambda(\theta_1, w)\}/2$. This implies that when $\|\theta-\theta_1\|_1 < \delta$, $d_\lambda(\theta, w) > d_\lambda\{\theta, w_\lambda(\theta_1)\}$ for all $\{w : \|w-w_\lambda(\theta_1)\|_1 = c\}$. Since $d_\lambda(\theta, w)$ is strictly convex, we have $\|w_\lambda(\theta) - w_\lambda(\theta_1)\| < c$.

Proof of Theorem 3. Note that $\hat{\theta}_\lambda$ and θ^* are the solutions of the estimating equations $U(\theta, w; X^{(1)}, \ldots, X^{(n)})/n = 0$ and $E\{U(\theta, w; X)\} = 0$, with w replaced by $\hat{w}_\lambda(\tilde{\theta})$ and w^*_λ, respectively. By Theorem 2 and Lemma A.2, we have $\|\hat{w}_\lambda(\tilde{\theta}) - w^*_\lambda\|_1 \to 0$ in probability, as $n \to \infty$. Under Condition 2, $E\{U_j(\theta, X)\}$ is bounded. Moreover, under Condition 3, each sub-likelihood score $\sum_{i=1}^n U_j(\theta; X^{(i)})/n \to E\{U_j(\theta; X)\}$ uniformly with probability tending to one. Since $U\{\theta, \hat{w}_\lambda(\tilde{\theta}); X^{(1)}, \ldots, X^{(n)}\}/n$ and $E\{U(\theta, w^*_\lambda; X)\}$ are the product of $\hat{w}_\lambda(\tilde{\theta})$, w^*_λ and the sub-likelihood scores, we have

$$\sup_{\theta\in\Theta}\left\|\frac{1}{n}U\{\theta, \hat{w}_\lambda(\tilde{\theta}); X^{(1)}, \ldots, X^{(n)}\} - E\{U(\theta, w^*_\lambda; X)\}\right\|_1 \to 0, \tag{A.2}$$

in probability as $n \to \infty$.

Moreover, by Condition 3, $\inf_{\{\theta:\|\theta-\theta^*\|_1\ge c\}} \|E\{U(\theta, w^*_\lambda; X)\}\|_1 > 0$ for any constant $c > 0$. Thus, for any $c > 0$, there exists a $\delta > 0$ such that the event

$\|\hat{\theta}_\lambda - \theta^*\|\|_1 \geq c$ implies the event $\|E\{U(\hat{\theta}_\lambda, w^*_\lambda; X)\}\|_1 > \delta$. We have

$$\begin{aligned}
&\Pr\{\|\hat{\theta}_\lambda - \theta^*\|\|_1 \geq c\} \\
&\leq \Pr\{\|E\{U(\hat{\theta}_\lambda, w^*_\lambda; X)\}\|_1 > \delta\} \\
&= \Pr\left\{\|E\{U(\hat{\theta}_\lambda, w^*_\lambda; X)\} - \frac{1}{n}U\{\hat{\theta}_\lambda, \hat{w}_\lambda(\tilde{\theta}); X^{(1)}, \ldots, X^{(n)}\}\|_1 > \delta\right\} \\
&\to 0,
\end{aligned}$$

as $n \to \infty$, where the equality is due to that $U\{\hat{\theta}_\lambda, \hat{w}_\lambda(\tilde{\theta}); X^{(1)}, \ldots, X^{(n)}\} = 0$ and the last line is implied by (A.2). This concludes the proof.

Proof of Theorem 4. Note that $U\{\hat{\theta}_\lambda, \hat{w}_\lambda(\tilde{\theta}); X^{(1)}, \ldots, X^{(n)}\}/n = 0$, and that $\|\hat{\theta}_\lambda - \theta^*\|_1 \to 0$ in probability as $n \to \infty$. By Condition 4 and applying the law of large number to the remainder, we obtain the following expansion of jth sub-likelihood score at θ^*,

$$\begin{aligned}
&\frac{1}{n}\sum_{i=1}^{n} U_j(\hat{\theta}_\lambda; X^{(i)}) - \frac{1}{n}\sum_{i=1}^{n} U_j(\theta^*; X^{(i)}) \\
&= \frac{1}{n}\sum_{i=1}^{n} \nabla U_j(\theta^*; X^{(i)})(\hat{\theta}_\lambda - \theta^*) + o_p(\|\hat{\theta}_\lambda - \theta^*\|_1 1_p),
\end{aligned}$$

where 1_p is the p-dimensional vector with elements equal to one. Note that under Condition 2, $\|\hat{w}_\lambda(\theta)\|_1$ is uniformly bounded (also see the proof of Lemma A.1). Taking the entry-wise product of the empirical composition rule and the sub-liekilhood scores implies

$$\begin{aligned}
&\sqrt{n}\frac{1}{n}U\{\theta^*, \hat{w}_\lambda(\tilde{\theta}); X^{(1)}, \ldots, X^{(n)}\} \\
&= -\sum_{j=1}^{m} \hat{w}_\lambda(\tilde{\theta})_j \frac{1}{n}\sum_{i=1}^{n} \nabla U_j(\theta^*; X^{(i)}) \left\{\sqrt{n}(\hat{\theta}_\lambda - \theta^*)\right\} + o_p(\sqrt{n}\|\hat{\theta}_\lambda - \theta^*\|_1 1_p),
\end{aligned} \tag{A.3}$$

where $\hat{w}_\lambda(\tilde{\theta})_j$ is the jth element of $\hat{w}_\lambda(\tilde{\theta})$. Note that by Theorem 2 and Lemma A.2, $\|\hat{w}_\lambda(\tilde{\theta}) - w^*_\lambda\|_1 \to 0$ in probability. Under Condition 4, by the Central Limit Theorem and Slustky's Theorem, the left-hand side of (A.3) converges in distribution to a multivariate normal random vector with mean zero and covariance $K(\theta^*, w^*_\lambda) = \mathrm{cov}\{U(\theta^*, w^*_\lambda; X)\}$ defined in (1.2). The $p \times p$ matrix $-\sum_{j=1}^{m} \hat{w}_\lambda(\tilde{\theta})_j \sum_{i=1}^{n} \nabla U_j(\theta^*; X^{(i)})/n$ in right hand side of (A.3) converges in probability to $H(\theta^*, w^*_\lambda)$ defined in (1.2) by the Law of Large Numbers and Slustky's Theorem. The invertibility of $H(\theta^*, w^*_\lambda)$ implies that $\hat{\theta}_\lambda$ is root-n consistent. Re-organizing (A.3) implies the desired result.

References

Bevilacqua, M. and Gaetan, C. (2015). Comparing composite likelihood methods based on pairs for spatial Gaussian random fields. *Statistics and Computing* **25**, 877–892.

Bradic, J., Fan, J. and Wang, W. (2011). Penalized composite quasi-likelihood for ultrahigh dimensional variable selection. *Journal of the Royal Statistical Society. Series B (Statistical Methodology)* **73**, 325–349.

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

Dillon, J. V. and Lebanon, G. (2010). Stochastic composite likelihood. *Journal of Machine Learning Research* **11**, 2597–2633.

Efron, B., Hastie, T., Johnstone, I. and Tibshirani, R. (2004). Least angle regression. *The Annals of Statistics* **32**, 407–499.

Ferrari, D., Qian, G. and Hunter, T. (2016). Parsimonious and efficient likelihood composition by Gibbs sampling. *Journal of Computational and Graphical Statistics* **25**, 935–953.

Gao, X. and Carroll, R. J. (2017). Data integration with high dimensionality. *Biometrika* **104**, 251–272.

Heagerty, P. J. and Lele, S. R. (1998). A composite likelihood approach to binary spatial data. *Journal of the American Statistical Association* **93**, 1099–1111.

Heyde, C. C. (2008). *Quasi-Likelihood and its Application: A General Approach to Optimal Parameter Estimation*. Springer Science & Business Media.

Huang, J., Ning, Y., Reid, N. and Chen, Y. (2020). On specification tests for composite likelihood inference. *Biometrika* **107**, 907–917.

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

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

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

Van der Vaart, A. W. (2000). *Asymptotic Statistics*. Cambridge University Press.

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

Wu, T. T. and Lange, K. (2008). Coordinate descent algorithms for Lasso penalized regression. *Annals of Applied Statistics* **2**, 224–244.

Xu, X. and Reid, N. (2011). On the robustness of maximum composite likelihood estimate. *Journal of Statistical Planning and Inference* **141**, 3047–3054.

Xue, L., Zou, H. and Cai, T. (2012). Nonconcave penalized composite conditional likelihood estimation of sparse Ising models. *The Annals of Statistics* **40**, 1403–1429.

Zhendong Huang

School of Mathematics and Statistics, University of Melbourne, Peter Hall Building, Parkville 3010, Australia.

E-mail: huang.z@unimelb.edu.au

Davide Ferrari

Faculty of Economics and Management, University of Bolzano, Bolzano, Trentino-Alto Adige 39100, Italy.

E-mail: davferrari@unibz.it

(Received June 2021; accepted April 2022)