Title	Fast Construction of Optimal Composite Likelihoods
Manuscript ID	SS-2021-0235
URL	http://www.stat.sinica.edu.tw/statistica/
DOI	10.5705/ss.202021.0235
Complete List of Authors	Zhendong Huang and Davide Ferrari
Corresponding Authors	Zhendong Huang
E-mails	huangzhd2014@gmail.com

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 useful in applications where the data have complex structure. Although composite likelihood construction is a crucial aspect influencing both computing and statistical properties of the resulting estimator, currently there does not seem to exist a universal rule to combine low-dimensional likelihood objects that is statistically justified and fast in execution. This paper develops a methodology to select and combine the most informative low-dimensional likelihoods from a large set of candidates while carrying out parameter estimation. The new procedure minimizes the distance between composite likelihood and full likelihood scores subject to a constraint representing the afforded computing cost. The selected composite likelihood is sparse in the sense that it contains a relatively small number of informative sub-likelihoods while the noisy terms are dropped. The resulting estimator is found to have asymptotic variance close to that of the minimum-variance estimator constructed using all 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. There are a number of situations, however, where the full likelihood is computationally intractable or difficult to specify. These challenges have motivated the development of composite likelihood methods, which avoid intractable full likelihoods by combining a set of low-dimensional likelihood objects. Due to its flexible framework and computational advantages, composite likelihood inference has become popular in many areas of statistics; e.g., see Varin et al. (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 it is possible to specify m tractable pdfs $f_1(s_1;\theta), \dots, f_m(s_m;\theta)$ for sub-vectors $S_1, \dots, S_m$ of X, each with dimension smaller than d. For example, S_1 may represent a single element like X_1, a variable pair like (X_1, X_2) or a conditional sub-vector like $X_1|X_2$. 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 a $m \times 1$ vector of constants to be determined as a solution to an optimality problem. For n independent, identically distributed 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:

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

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

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. Besides computational advantages and modeling flexibility, one reason for the popularity of the composite likelihood estimator is that it enjoys properties analogous to maximum likelihood (Lindsay, 1988; Lindsay et al., 2011; Varin et al., 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, $H(\theta, w) = -E\{\nabla U(\theta, w; X)\}$ and $K(\theta, w) = \text{cov}\{U(\theta, w; X)\}$ are the $p \times p$ sensitivity and variability matrices, respectively. Although the maximum composite likelihood estimator is consistent, $G(\theta^*, w)$ is generally different from the Fisher information $\text{cov}\{\nabla \log f(X; \theta^*)\}$, 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 computational efficiency of the composite likelihood estimator (Lindsay et al., 2011; Xu and Reid, 2011; Huang et al., 2020). On one hand, the established theory of unbiased estimating equations would suggest to find w so to maximize $tr\{G(\theta^*, w)\}$ (Heyde, 2008, Chapter 2). Although theoretically appealing, these optimal weights depend on inversion of the score covariance matrix whose estimates are often singular. Different selection strategies to balance the trade-off between statistical efficiency and computing cost have been explored in the literature. A common practice is to retain all feasible sub-likelihoods with $w_j = 1$, for all $j \geq 1$, but this is undesirable from either computational parsimony or statistical efficiency viewpoints, since the presence of too many correlated scores inflates the variability matrix K (Cox and Reid, 2004; Ferrari et al., 2016). A smaller subset may be selected by setting some of the w_js equal to zero, but determining a suitable subset remains challenging. Dillon and Lebanon (2010) and Ferrari et al. (2016) develop stochastic approaches where sub-likelihoods are sampled according to a statistical efficiency criterion. Ad-hoc methods have been developed depending on the type of model under exam; for example, in spatial data analysis it is often convenient to consider sub-likelihoods corresponding to close-by observations; e.g., see Heagerty and Lele (1998); Sang and Genton (2014); Bevilacqua and Gaetan (2015).

Motivated by this gap in the literature, the present paper develops a methodology to select sparse composite likelihoods in large problems by retaining only the most informative scores in the estimating equations (1.1), while dropping the noisy ones. To this end, we propose to minimize the distance between the maximum likelihood score and the composite likelihood score subject to a constraint representing the overall computing cost.

The reminder 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. Particularly, while Theorem 2 shows that the proposed empirical composition rule is asymptotically equivalent to the optimal composition rule that uses all the available scores, Theorems 3 and 4 give consistency and asymptotic normality of the resulting parameter estimator. In Section 4, we discuss some illustrative examples related to common families of models. In Section 5, we apply our method to real Covid-19 epidemiological data. Finally, in Section 6 we conclude and provide final remarks.

2. Main methodology

2.1 Penalized score distance minimization

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

$$\frac{1}{2} E \left\| U^{ML}(\theta; X) - U(\theta, w; X) \right\|_2^2 + \lambda \sum_{j=1}^{m} |w_j| \,, \tag{2.3}$$

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

The vector of coefficients minimizing (2.3) is allowed to contain positive, negative or zero values, although negative elements do not cause any specific concerns in our method. The size of such coefficients is expected to be larger for those sub-likelihoods that are strongly correlated with the full likelihood. Thus, a negative coefficient associated with certain sub-likelihood score does not imply that such a sub-likelihood is less informative, but simply means that it has negative correlation with the maximum likelihood score.

The first term in the objective (2.3) aims at improving 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 minimization of the variance in estimating equations; see (Heyde, 2008, Ch.1). Lindsay et al. (2011) point out that this type of criterion is suitable in the context of composite likelihood estimation, although finding a general computational procedure to minimize such a criterion in large probelms is still an open problem.

The term $\lambda \sum_{j=1}^{m} |w_j|$ is a penalty discouraging 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 which is 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 level of afforded computing. Alternatively, $w_\lambda(\theta)$ may be viewed as a composition rule that minimizes the likelihood complexity subject to some afforded efficiency loss compared to 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 likelihood complexity, while for increasing $\lambda > 0$ informative data subsets and their scores might be tossed away.

Difficulties related to the direct minimization of (2.3) are the presence of the intractable likelihood score function U^{ML} and the expectation depending on the unknown parameter θ. Up to an additive term independent of w, the penalized score distance in (2.3) can be expressed as

$$\frac{1}{2} E \left\| U(\theta, w; X) \right\|_2^2 - E\left\{ U^{ML}(\theta; X)^\top U(\theta, w; X) \right\} + \lambda \sum_{j=1}^{m} |w_j| \,. \tag{2.4}$$

Let $M(\theta; X) = (U_1(\theta; X), \ldots, U_m(\theta; X))$ be the $p \times m$ matrix with columns given by $p \times 1$ score vectors $U_j(\theta; X)$ $(j = 1, \ldots, m)$. Then, the first term of (2.4) 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 (2.4) is $-w^\top \mathrm{diag}\{J(\theta^*)\}$, where $\mathrm{diag}(A)$ denotes the diagonal vector of the matrix A. Therefore, (2.4) may be approximated by

$$d_\lambda(\theta, w) = \frac{1}{2} w^\top J(\theta) w - w^\top \mathrm{diag}\{J(\theta)\} + \lambda \sum_{j=1}^{m} |w_j| \,. \tag{2.5}$$

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

$$\hat{d}_\lambda(\theta, w) = \frac{1}{2} w^\top \hat{J}(\theta) w - w^\top \mathrm{diag}\{\hat{J}(\theta)\} + \lambda \sum_{j=1}^{m} |w_j| \,, \tag{2.6}$$

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

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

$$U(\theta, \hat{w}_\lambda(\theta); X^{(1)}, \ldots, X^{(n)}) = 0. \tag{2.7}$$

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

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

$$U(\theta, \hat{w}_\lambda(\tilde{\theta}); X^{(1)}, \ldots, X^{(n)}) = 0. \tag{2.8}$$

where $\hat{w}_\lambda(\tilde{\theta})$ is a quantity fully dependent 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

For the numerical examples in this paper, the following implementation is considered. We first compute the sparse composition rule $\hat{w}_\lambda(\tilde{\theta})$, by minimizing the convex criterion (2.6) with $\theta = \tilde{\theta}$ being a preliminary root-n consistent estimator. Minimization of (2.6) is implemented through the least-angle regression algorithm (Efron et al., 2004). Finally, the estimator $\hat{\theta}_\lambda$ is obtained by a one-step Newton-Raphson update starting from $\theta = \tilde{\theta}$ applied to (1.1). See Chapter 5 in Van der Vaart (2000) for an introduction to one-step estimation. As preliminary estimator $\tilde{\theta}$, one may choose the composite likelihood estimator with uniform composition rule $w_j = 1$, for all $j \geq 1$. From theoretical view point, any root-n consistent initial estimator $\tilde{\theta}$ leads to the same asymptotic results of the final estimate. We find that the impact of the initial estimates is negligible in many situations.

Analogous 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 $U_j(\theta; X)$ having the largest correlation with the residual difference $U_j(\theta; X) - U(\theta, w; X)$, followed by an adjustment step on w. An alternative computing approach is to solve (2.6) with respect to θ with $w = w_\lambda(\theta)$ using a Newton-Raphson algorithm with $w_\lambda(\theta)$ updated through a coordinate-descent approach as in Wu et al. (2008) in each iteration.

Selection of λ is an aspect of practical importance since it balances the trade-off between statistical and computational efficiency. In many practical applications this choice ultimately depends on the available computing resources and the objective of one's analysis. Although a universal approach for selection is not sought in this paper, the following heuristic strategy may be considered. 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 $\tilde{\theta}$, whilst $\hat{J}$ is the covariance for all scores. Thus $\phi(\lambda)$ can be viewed as the approximate proportion of 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 due to excessive computational burden, we propose to use an upper bound of $\phi(\lambda)$ instead. Let $\tilde{\lambda} \in \Lambda$ be the next value for $\lambda \in \Lambda$ smaller than $\hat{\lambda}$, i.e. 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 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 $(U_1^\top, U_1^\top U_1 \pm \lambda), \dots, (U_m^\top, U_m^\top U_m \pm \lambda)$ are linearly independent.*

We note that the condition is automatically satisfied, unless some partial score is perfectly correlated to the others, which is rarely the case in real applications.

For a vector $a \in \mathbb{R}^m$, we use $a_{\mathcal{E}}$ to denote the sub-vector corresponding to index $\mathcal{E} \subseteq \{1, \ldots, m\}$, while $A_{\mathcal{E}}$ denotes the sub-matrix of the squared matrix A formed by taking rows and columns corresponding to $\mathcal{E}$. The notation $\mathrm{sign}(w)$ is used for the vector sign function with jth element taking values -1, 0 and 1 if $w_j < 0$, $w_j = 0$ and $w_j > 0$. 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 a $m \times q$ matrix collecting 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 (2.6) is unique with probability one and given by*

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

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

$$\hat{\mathcal{E}} = \left\{ j : \left| n^{-1} \sum_{i=1}^n U_j(\theta; X^{(i)})^\top U_j(\theta; X^{(i)}) \right.\right.$$
$$\left.\left. - n^{-1} \sum_{i=1}^n U_j(\theta; X^{(i)})^\top U(\theta, \hat{w}_\lambda(\theta); X^{(i)}) \right| \geq \lambda \right\}, \qquad (3.9)$$

and $\backslash\hat{\mathcal{E}}$ *denotes the complement index set* $\{1,\ldots,m\}\backslash\hat{\mathcal{E}}$. *Moreover,* $\hat{w}_{\lambda,\hat{\mathcal{E}}}(\theta)$ *contains at most* $np \wedge m$ *non-zero 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 zero elements. In this case, the empirical score covariance matrix $\hat{J}(\theta)$ is required to be non-singular which is certainly violated when $np < m$. Even for the case $np > m$, $\hat{J}(\theta)$ may be singular due to the presence of largely correlated partial scores. On the other hand, setting $\lambda > \eta$ always gives a non-singular score covariance matrix and guarantees existence of $\hat{w}_{\lambda,\hat{\mathcal{E}}}(\theta)$. For sufficiently large λ, a relatively small subset of scores is selected. The formula in (3.9) suggests that the jth score is selected when it contributes enough information in the overall composite likelihood. Particularly, 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

The asymptotic behavior of the sparse composition rule $\hat{w}_\lambda(\theta)$ is investigated in this section 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 (2.5). 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 estimator's variance, with the latter, however, involving all m scores. To show this, an additional technical requirement on the covariance between sub-likelihood scores is introduced.

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 element 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 bounded away from zero for all $\theta \in \Theta$ uniformly.*

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 $n \to \infty$.*

Since the preliminary estimate $\tilde{\theta}$ is consistent, continuity of $w_\lambda(\theta)$ (shown in Lemma 2 in the Appendix) implies immediately that $\hat{w}_\lambda(\tilde{\theta})$ converges to $w_\lambda(\theta^*)$, i.e. 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

$$\sup_{\theta\in\Theta}\left\|\frac{1}{n}\sum_{i=1}^{n}U(\theta,\hat{w}_\lambda(\tilde{\theta});X^{(i)})-\frac{1}{n}\sum_{i=1}^{n}U(\theta,w_0(\theta^*);X^{(i)})\right\|_1\to 0, \tag{3.10}$$

in probability as $n\to\infty$ and $\lambda\to 0$. Note the optimal composition rule $w_0(\theta^*)$ and the related Godambe information matrix $G\{\theta,w_0(\theta^*)\}$ are typically hard to compute due to inverting the $m\times m$ score covariance matrix with entries $E[U_j(\theta;X)^\top U_k(\theta;X)]$, $1\le j,k\le m$. On the other hand, the sparse composition rule $\hat{w}_\lambda(\tilde{\theta})$ and the implied Godambe information have the advantage to be computationally tractable since they only involve a fraction of scores.

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

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

$$\frac{1}{n}\sum_{i=1}^{n}U\{\theta,\hat{w}_\lambda(\tilde{\theta});X^{(i)}\}=0. \tag{3.11}$$

Since the vector $\hat{w}_\lambda(\tilde{\theta})$ converges to $w^*_\lambda=w_\lambda(\theta^*)$ by Theorem 2 and Lemma 2, the limit of (3.11) may be written as

$$E\left\{U(\theta,w^*_\lambda;X)\right\}=0. \tag{3.12}$$

To show that $\hat{\theta}_\lambda$ is consistent for the solution of (3.12), we assume additional regularity conditions, particularly, to ensure a unique root of the ideal composite likelihood score and stochastic equicontinuity on each 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* $\sup_{\theta\in\Theta} \| \sum_{i=1}^n U_j(\theta; X^{(i)})/n - EU_j(\theta; X)\|_1 \to 0$ *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 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 only depending on* x. *Moreover, assume* $H(\theta^*, w_\lambda^*)$ *defined in (1.2) is nonsingular.*

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

$$n^{1/2} G_\lambda(\theta^*)^{1/2} \left(\hat{\theta}_\lambda - \theta^*\right) \to N_p(0, I_p), \tag{3.13}$$

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

The $p\times p$ Godambe information matrix $G_\lambda(\theta)$ in (3.13) can be estimated by the sandwich estimator $\hat{G}_\lambda = \hat{H}_\lambda \hat{K}_\lambda^{-1} \hat{H}_\lambda$, where $p \times p$ matrices $\hat{H}_\lambda$ and

$\hat{K}_\lambda$ are obtained, respectively, 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 computational cost and increased stability of the standard error calculations. Although the score variance matrix $\hat{K}_\lambda(\theta)$ may be difficult to obtain when $\lambda = 0$ due 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 usually Σ is estimated by the sample covariance $\hat{\Sigma}$. When $n < m$, however, the maximum likelihood estimator is not available since 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 $X^{(1)}, \dots, X^{(n)}$ is

$$0 = \sum_{j=1}^{m} \frac{w_j}{\sigma_j^2} \sum_{i=1}^{n} (X_j^{(i)} - \theta). \tag{4.14}$$

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

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

respectively.

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

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

whilst the optimal population composition rule $w_{\lambda,j}(\theta)$ is the same as the above expression with sample averages replaced by expectations. The composition rule $w_{\lambda,j}(\theta)$ evaluated at the true parameter is $w^*_{\lambda,j} = (1 - \lambda \sigma_j^2) I(\sigma_j^{-2} > \lambda)$ $(j = 1, \dots, m)$. 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. Particularly, the number of non-zero 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 (4.14) gives the following fixed-point equation

$$\theta = \left\{ \sum_{j=1}^{m} \frac{\hat{w}_{\lambda,j}(\theta)}{\sigma_j^2} \bar{X}_j \right\} / \left\{ \sum_{k=1}^{m} \frac{\hat{w}_{\lambda,k}(\theta)}{\sigma_k^2} \right\}, \tag{4.15}$$

which is a weighted average of marginal sample means $\bar{X}_j = n^{-1} \sum_{i=1}^{n} X_j^{(i)}$ $(j = 1, \ldots, m)$. The final composite likelihood estimator $\hat{\theta}_\lambda$ may be obtained by solving equation (4.15).

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 minimum variance, it offers no control for the overall computational cost since all m sub-scores are selected. On the other hand, choosing judiciously $\lambda > 0$ may lead to low computational burden with negligible efficiency loss for the resulting estimator. For instance, assuming $\sigma_j^2 = j^2$, a calculation shows

$$\frac{1}{2} E \left\{ U(\theta, w_\lambda^*; X) - U(\theta, w_0^*; X) \right\}^2 \leq \lambda^2 \sum_{j \in \mathcal{E}} j^2 + \sum_{j \notin \mathcal{E}} j^{-2}, \tag{4.16}$$

where $\mathcal{E} = \{j : j^2 < \lambda^{-1}\}$ and θ here is the true parameter. Since the number of selected scores is $\sum_{j=1}^{m} I\left(j^2 < \lambda^{-1}\right) = \lfloor \lambda^{-\frac{1}{2}} \rfloor$, we can write $\lambda^2 \sum_{j \in \mathcal{E}} j^2 \leq \lambda^2 \lambda^{-1} \lambda^{-\frac{1}{2}} = \lambda^{\frac{1}{2}}$, which converges to zero as $\lambda \to 0$; additional calculations also show $\sum_{j \notin \mathcal{E}} j^{-2} \to 0$ as $\lambda \to 0$. Hence, the left hand side of (4.16) goes to zero as long as $\lambda \to 0$. This suggests that the truncated composite likelihood score approximates suitably the optimal score, while containing a relatively small number of 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, tossing away highly correlated partial scores would improve computing while maintaining satisfactory statistical efficiency for the final estimator.

Figure 1 shows the solution path of w^*_λ; that is, 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 to the maximum likelihood estimator for different values of λ. When m is large ($m = 1000$), the asymptotic relative efficiency drops gradually until only a few scores are left. This example illustrates that relatively high efficiency can be reached by the selected composite likelihood equations, when a few partial scores already 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

$\rho = 0$, $m = 20$ $\qquad$ $\rho = 0.5$, $m = 20$ $\qquad$ $\rho = 0.5$,

$m = 1000$

Number of sub-likelihoods

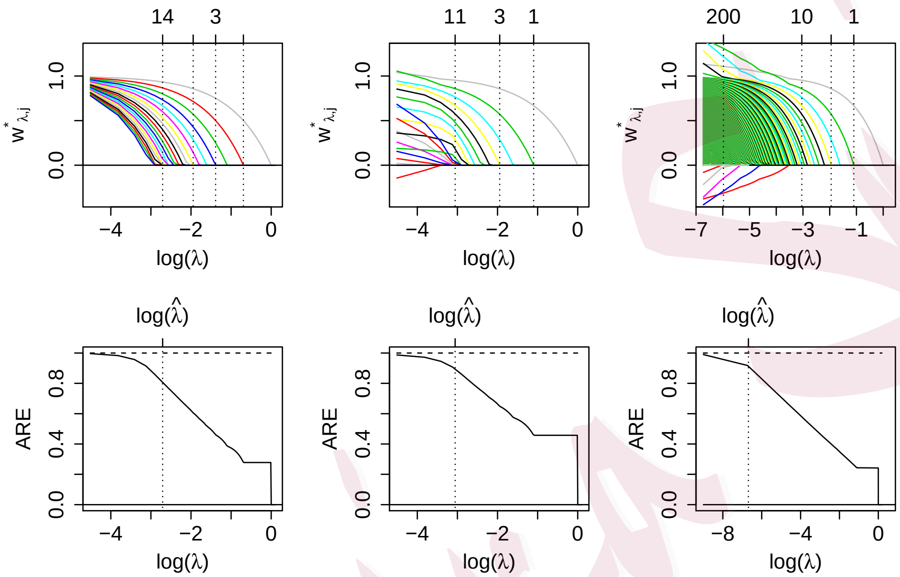

Figure 1: Top Row: Solution paths for the minimizer w^*_λ of $d_\lambda(\theta, w)$ defined in (2.5) at the true parameter for different values of λ with corresponding number of sub-likelihoods shown in the top axis. Bottom Row: Asymptotic relative efficiency (ARE) of $\hat{\theta}_\lambda$ compared to the maximum likelihood estimator. The vertical dashed lines represent $\hat{\lambda}$ selected as described in §2.2 using $\tau = 0.9$. Results correspond to the location model $X \sim N_m(\theta 1_m, \Sigma)$ with $\Sigma_{jk} = j$ $(j = k)$ and $\Sigma_{jk} = \rho(jk)^{1/2}$ $(j \neq k)$.

$\Sigma(\theta)_{jk} = 1$ $(j = k)$. The quantity δ_{jk} may be regarded as the distance between spatial locations j and k, and the case of equally distant points corresponds to covariance estimation for exchangeable variables described in detail by Cox and Reid (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 to the maximum likelihood estimator. A number of variable pairs ranging from $m = 45$ to $m = 1225$ is considered. When $\lambda = 0$, the proposed estimator has relatively high asymptotic efficiency. Interestingly, efficiency stays at around 90% until only a few sub-likelihoods are left, suggesting that a very small proportion of partial-likelihood components already contains the majority of the information about θ. In such cases, the proposed selection procedure is useful by reducing the computing 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 $\Sigma = (1-\rho)I_m + \rho 1_m 1_m^\top$, $0 < \rho < 1$, 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, taking a sufficiently large value for λ, so 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 eigenvalue of the score covariance $J(\theta)$ is $\rho(m-1)+1$, whilst the remaining $m-1$ eigenvalues are all equal to $1-\rho$, suggesting that the first score contains relatively large information on θ compared to the other scores. Furthermore, since $\text{var}(\hat{\theta}_\lambda) = \{\rho(m-1)+1\}/(mn)$ the asymptotic relative efficiency of the composite likelihood with $m < \infty$ compared to 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 $\rho = 0.75$.

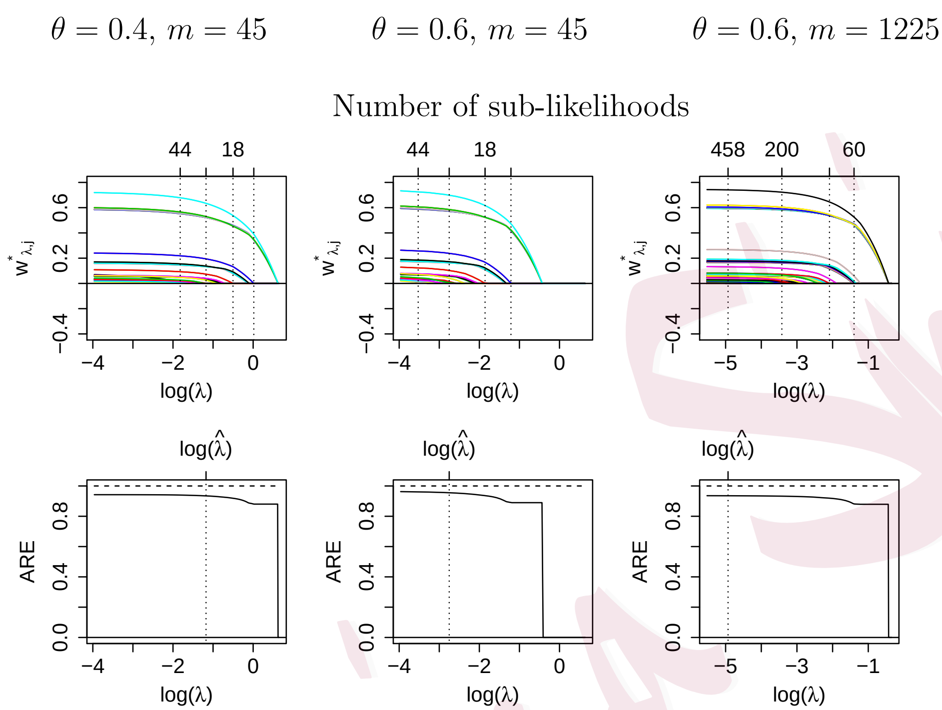

Figure 2: Top Row: Solution paths for the minimizer w^*_λ of $d_\lambda(\theta, w)$ defined in (2.5) at the true parameter for different values of λ with corresponding number of sub-likelihoods shown in the top axis. Bottom Row: Asymptotic relative efficiency (ARE) of the estimator $\hat{\theta}_\lambda$ compared to the maximum likelihood estimator. The vertical dashed lines correspond to $\hat{\lambda}$ selected as described in §2.2 with $\tau = 0.9$. Results correspond to the model $X \sim N_d(0, \Sigma(\theta))$ with $\Sigma(\theta)_{jk} = \exp\{-\theta(2|j-k|)^{1/2}\}$.

5. Real data example: Spatial covariance estimation for Covid-19 data

The methodology is applied to public health data on the Covid-19 epidemic supplied by the Italian Civil Protection Department. The data considered here consists of $n = 60$ observations on daily new Covid-19 cases from 24 February to 23 April 2020 observed across the $d = 90$ Italian provinces corresponding to capital cities in regions or autonomous territories. Data is available as supplementary material to this paper. 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

$$\mathrm{cov}(\varepsilon_j^{(i)}, \varepsilon_k^{(i)}) = \sigma_j \sigma_k \exp\{-\theta t_{jk}/(m_j m_k)\} \quad (j, k = 1, \ldots, d), \tag{5.17}$$

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 latitude and longitude for the jth site, respectively. 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 exponential spatial covariance function given in (5.17); see Bevilacqua and Gaetan (2015) for more details and comparisons on pairwise likelihood estimation 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 included both population size and distance between provinces are included in the gravity model specified in (5.17). 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 in 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 to monitor contagion across provinces. To this end, the data are first de-trended and normalized; estimation of the local trends $\hat{\mu}_j$ are obtained by a Nadaraya-Watson kernel smoother, implemented in the R function `ksmooth`. The error variance σ_j^2 are 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 §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

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
# sub0likelihoods	40	42	44	46	48	50	52	54	56	58	60

§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 $\binom{90}{2} = 4005$ pairs. We also see from the table that both $\hat{\theta}_\lambda$ and the standard error is converging with only about 58 sub-likelihoods 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 and final remarks

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, a universal procedure to construct composite likelihoods that is statistically justified and fast to execute does not seem to exist (Lindsay et al., 2011). Motivated by this gap in the literature, this paper introduces a selection methodology resulting 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 found to be useful in common situations where 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 O_F-optimal score in large samples under reasonable conditions; see Theorem 2 and Equation 3.10.

For implementation, we proposed a selection criterion to choose λ which perform well in the examples, instead of providing a universal approach. In practice, it could be feasible to use any alternative criterions to choose λ, according to the realization of the problems. For example, when the full score covariance is not available due to computational burden, one may consider to use the upper bound provided in Section 2.2, or to choose λ up to some given level of information gain, defined by the ratio of the smallest eigenvalue to the trace of the current selected score covariance, which is decreasing with λ by min-max theorem in linear algebra. As another idea, we note that 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 pre-fixed value.

Building on the recent success of shrinkage methods for the full likelihood, many works have proposed the use of sparsity-inducing penalties in the composite likelihood framework; e.g., see Bradic et al. (2011); Xue et al. (2012); Gao and Carroll (2017). However, the spirit of our approach is entirely different from these methods, since our penalty focuses on entire sub-likelihood functions rather than on elements of θ. In contrast to the above approaches, our penalization strategy has the advantage of retaining asymptotically unbiased estimating equations, thus leading to desirable asymptotic properties of the related parameter estimator.

A number of developments of the present study may be pursued in the future from either theoretical or applied viewpoints. Although the current paper focuses on the case where p is finite, penalties able to deal with situations where both m and p are allowed to grow with n may be useful for the analysis of high-dimensional data. Implementations of the convex efficiency criterion (2.3) beyond the current i.i.d. setting would be another useful future research direction. For example, this would be valuable for the analysis of spatial or spatio-temporal data, where often the overwhelming number of sub-likelihoods poses a challenge to traditional composite likelihood methods.

Acknowledgement

The authors acknowledge the financial support from 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

$$\frac{1}{n}\sum_{i=1}^{n} U_j^{(i)\top} U^{(i)} - \frac{1}{n}\sum_{i=1}^{n} U_j^{(i)\top} U_j^{(i)} + \lambda\gamma_j = 0, \qquad \text{for } j = 1, \ldots, m, \quad (6.18)$$

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 3.9 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 (6.18). □

Lemma 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 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

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

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 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. $\square$

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

Proof of Lemma 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 2.5. 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 $\{w \in \mathbb{R}^m : \|w - w_\lambda(\theta_1)\|_1 \leq c\} \times \{\theta \in \mathbb{R}^p : \|\theta - \theta_1\|_1 \leq \delta_1\}$ 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 $\{w : \|w - w_\lambda(\theta_1)\|_1 \leq c\}$, $\|d_\lambda(\theta, w) - d_\lambda(\theta_1, w)\|_1 < \{c_1 - d_\lambda(\theta_1, w)\}/2$. 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 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

$$\sup_{\theta\in\Theta} \left\| \frac{1}{n} U\{\theta, \hat{w}_\lambda(\tilde{\theta}); X^{(1)}, \ldots, X^{(n)}\} - E\{U(\theta, w^*_\lambda; X)\} \right\|_1 \to 0, \qquad (6.19)$$

in probability as $n \to \infty$.

Moreover, by Condition 3, $\inf_{\{\theta:\|\theta-\theta^*\|_1\geq c\}} \|E\{U(\theta, w^*_\lambda; X)\}\|_1 > 0$ 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

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

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 (6.19). This concludes the proof. $\square$

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 1). Taking the entry-wise product of the empirical composition rule and the sub-liekilhood scores implies

$$\sqrt{n}\frac{1}{n}U\{\theta^*, \hat{w}_\lambda(\tilde{\theta}); X^{(1)}, \dots, X^{(n)}\}$$
$$= -\sum_{j=1}^{m} \hat{w}_\lambda(\tilde{\theta})_j \frac{1}{n}\sum_{i=1}^{n} \nabla U_j(\theta^*; X^{(i)}) \left\{\sqrt{n}(\hat{\theta}_\lambda - \theta^*)\right\} + o_p(\sqrt{n}\|\hat{\theta}_\lambda - \theta^*\|_1 1_p), \tag{6.20}$$

where $\hat{w}_\lambda(\tilde{\theta})_j$ is the jth element of $\hat{w}_\lambda(\tilde{\theta})$. Note that by Theorem 2 and Lemma 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 (6.20) 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 (6.20) 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 (6.20) implies the desired result. $\square$

References

Bevilacqua, M. and C. Gaetan (2015). Comparing composite likelihood methods based on pairs for spatial gaussian random fields. *Statistics and Computing 25*(5), 877–892.

Bradic, J., J. Fan, and W. Wang (2011). Penalized composite quasi-likelihood for ultrahigh dimensional variable selection. *Journal of the Royal Statistical Society: Series B (Statistical Methodology) 73*(3), 325–349.

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

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

Efron, B., T. Hastie, I. Johnstone, R. Tibshirani, et al. (2004). Least angle regression. *The Annals of statistics 32*(2), 407–499.

Ferrari, D., G. Qian, and T. Hunter (2016). Parsimonious and efficient likelihood composition by gibbs sampling. *Journal of Computational and Graphical Statistics 25*(3), 935–953.

Gao, X. and R. J. Carroll (2017). Data integration with high dimensionality. *Biometrika 104*(2), 251–272.

Heagerty, P. J. and S. R. Lele (1998). A composite likelihood approach to binary spatial data. *Journal of the American Statistical Association 93*(443), 1099–1111.

Heyde, C. C. (2008). *Quasi-likelihood and its application: a general approach to optimal parameter estimation.* Springer Science & Business Media.

Huang, J., Y. Ning, N. Reid, and Y. Chen (2020). On specification tests for composite likelihood inference. *Biometrika 107*(4), 907–917.

Lindsay, B. G. (1988). Contemporary mathematics volume 80, 1988. Volume 80, pp. 221–239.

Lindsay, B. G., G. Y. Yi, and J. Sun (2011). Issues and strategies in the selection of composite likelihoods. *Statistica Sinica 21*(1), 71–105.

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

Van der Vaart, A. W. (2000). *Asymptotic statistics*, Volume 3. Cambridge university press.

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

Wu, T. T., K. Lange, et al. (2008). Coordinate descent algorithms for lasso penalized regression. *Annals of Applied Statistics 2*(1), 224–244.

Xu, X. and N. Reid (2011). On the robustness of maximum composite likelihood estimate. *Journal of Statistical Planning and Inference 141*(9), 3047–3054.

Xue, L., H. Zou, T. Cai, et al. (2012). Nonconcave penalized composite conditional likelihood estimation of sparse Ising models. *The Annals of Statistics 40*(3), 1403–1429.

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

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

Davide Ferrari, Faculty of Economics and Management, University of Bolzano, Piazza Università 1, Bolzano, 39100, Italy

E-mail: (davferrari@unibz.it)