Statistica Sinica **21** (2011), 71-105

# ISSUES AND STRATEGIES IN THE SELECTION OF COMPOSITE LIKELIHOODS

Bruce G. Lindsay[1], Grace Y. Yi[2] and Jianping Sun[1]

[1]*Penn State University and* [2]*University of Waterloo*

*Abstract:* The composite likelihood method has been proposed and systematically discussed by Besag (1974), Lindsay (1988), and Cox and Reid (2004). This method has received increasing interest in both theoretical and applied aspects. Compared to the traditional likelihood method, the composite likelihood method may be less statistically efficient, but it can be designed so as to be significantly faster to compute and it can be more robust to model misspecification. Although there are a number of ways to formulate a composite likelihood to balance the trade-off between the efficiency and computational price, there does not seem to exist a universal rule for constructing a combination of composite likelihoods that is both computationally convenient and statistically appealing. In this article we present some thoughts on the composite likelihood, drawing on basic knowledge about likelihood and estimating functions. A new efficiency result based on the Hoeffding decomposition of $U$-statistics is given. A recommendation is given to consider the construction of surrogate density functions as a way to better bridge the gap between likelihood methods and composite likelihood methods.

*Key words and phrases:* Estimating functions, Fisher consistency, Hoeffding scores, inference functions, information-unbiased, likelihood functions, linear combinations.

## 1. Introduction

The likelihood function has become a centerpiece of statistical inference since it was turned into a powerful tool by Fisher (1922). Modern high-dimensional data, such as spatial data, image data, complex structured longitudinal data, and long-sequence genetic data, have generated significant new challenges to the use of likelihood-based methods. There are two facets to the challenge; one involves model-building, and the need to build reasonable models; the second involves computing, and the need to produce answers in reasonable time. These challenges have helped to generate considerable interest in alternative estimation methods that are not based on full likelihood specification, such as quasi-likelihood (Wedderburn (1974); McCullagh (1983)), and in estimating functions more generally (e.g., Godambe (1960); Durbin (1960)). In this paper, we consider a methodology based on partial specification of the full likelihood that is called *composite likelihood*.

Let $\mathbf{Y} = (Y_1, \ldots, Y_d)^{\mathrm{T}}$ be a $d$-dimensional random vector with the probability model $f(\mathbf{y}; \theta)$, where $\theta$ is a parameter taking values in a parameter space $\Theta$ that is a subset of a Euclidean space of dimension $p$. To narrow the scope of this paper, we assume that there is a significant computational challenge in evaluating $f(\mathbf{y}; \theta)$, or the corresponding likelihood $L(\theta)$, and this challenge increases dramatically as the data dimension $d$ increases. Thus, while it might be possible to compute a lower dimensional marginal distribution, such as that of $(Y_i, Y_j)$, it might be impossible to compute $f(\mathbf{y}; \theta)$.

We use the symbol $N_{ops}(D)$ to indicate the number of computer operations, in order of magnitude, needed to calculate a marginal distribution for $D$ variables a single time. For example, if $f(\mathbf{y}; \theta)$ is the density of the multivariate normal distribution $N(\mu, \Sigma)$, the calculation of the inverse of $\Sigma$ for any subset of $D$ variables makes $N_{ops}(D) = O(D^3)$ unless $\Sigma$ is of special form. In more severe cases, $N_{ops}(D)$ grows exponentially in $D$. This occurs, for instance, when calculation of the distribution of the marginal subset requires integration over a set of $D$ unobserved random variables. We assume that calculating the densities is so expensive that one can only consider their calculation for smaller values of $D$. There are obviously other computational costs associated with our problem, but we assume that the one time calculation of the density for a marginal density grows so quickly in dimension $D$ that computation for the full dimension $d$ is not possible.

One solution to this computational problem is to consider estimation methods based on objective functions called *composite likelihoods*, which simply means a product of sub-likelihoods,

$$C(\theta) = \prod_{k=1}^{N_{cl}} L(\theta; S_k), \tag{1.1}$$

where $N_{cl}$ is the number of factors in $C(\theta)$ and where each $L(\theta; S_k)$ is a user-selected sub-likelihood generated from $f(\mathbf{y}; \theta)$ by considering a particular conditional or marginal set of variables $S_k$. That is, $S_k$ could be a sub-vector of the data, such as $(y_1, y_2)$, or it could be a "conditioned" pair of vectors, such as $(y_1, y_2)|y_1$. Ideally, $C(\theta)$ would be constructed so that the parameter $\theta$ is identified.

Composite likelihood methods date back to the pseudo-likelihood of Besag (1974) and the partial likelihood of Cox (1975). Lindsay (1988) coined the term composite likelihood for "product of likelihood" $C(\theta)$ constructions; he reviewed the small available literature to that date. In the 1990's, the pairwise likelihood (Hjort and Omre (1994); Heagerty and Lele (1998); Nott and Rydén (1999)) rose to prominence in the world of spatial statistics. The use of the composite

likelihood, especially the pairwise likelihood, has received increasing attention in recent years due to the simplicity in defining the objective function and computational advantages when dealing with data with complex structure (e.g., Kuk and Nott (2000); Zhao and Joe (2005); Fieuws and Verbeke (2006); Dillon and Lebanon (2009)). A recent review can be found in Varin (2008) and Varin, Reid, and Firth (2011). There are also many uses in the complex likelihoods of genomic data (e.g., Devlin, Risch, and Roeder (1996); Fearnhead and Donnelly (2002); Engler et al. (2006)). A review on this topic is given in Larribe and Fearnhead (2011).

There are two particularly attractive features about the composite likelihood. For the first, the standard Kullback-Leibler inequality applies to each sub-likelihood, so we have:

$$E_{\theta_0}\Big[\log\Big\{\frac{C(\theta_0)}{C(\theta)}\Big\}\Big] = E_{\theta_0}\left[\sum_{k=1}^{N_{cl}} \log\Big\{\frac{L(\theta_0; S_k)}{L(\theta; S_k)}\Big\}\right] \geq 0,$$

where $\theta_0$ denotes the true value of $\theta$. This implies that maximization based on $\log C(\theta)$ gives a Fisher consistent estimation method for $\theta$. (With added regularity conditions, Fisher consistency often implies consistency in probability.) This reliability is a key factor in making composite likelihood a reasonable choice in a complex situation. For the second attractive feature, $E\{\nabla \log L(\theta; S_k)\} = 0$ under standard regularity conditions, so that $c(\theta) = \nabla \log C(\theta)$ is a mean zero estimating equation for $\theta$. Here $\nabla$ is the gradient operator, that is, the operation of differentiation with respect to the vector $\theta$. Since these two properties match those of the full likelihood, they might be called *first order* likelihood properties.

Thus the validity of using a chosen composite likelihood to perform inference about $\theta$ can be justified either from the standpoint of unbiased estimating functions or the Kullback-Leibler criterion. For details, see Lindsay (1988), Cox and Reid (2004) and Varin (2008). These properties are quite important, as they give us a guarantee that the $\theta$ value that we are estimating with the composite likelihood $C(\theta)$ is exactly the same as the one in the complete model.

Unfortunately, the second order properties of likelihood are not possessed by composite likelihood except under special circumstances. The key second order property of a likelihood is the *second Bartlett identity*,

$$\mathcal{I}(\theta) = E\{-\nabla^2 \log L(\theta)\} = E[\{\nabla \log L(\theta)\}\{\nabla \log L(\theta)\}^{\mathrm{T}}],$$

where $\nabla^2$ denotes the Hessian, or the operation of twice differentiation with respect to $\theta$. This property holds for all the sub-likelihoods in the composite likelihood, so that there is a meaningful *Fisher information matrix* for each term in the composite likelihood,

$$\mathcal{I}_{S_k}(\theta) = E\{-\nabla^2 \log L(\theta; S_k)\} = E[\{\nabla \log L(\theta; S_k)\}\{\nabla \log L(\theta; S_k)\}^{\mathrm{T}}].$$

However, this identity does not hold in general for the log composite likelihood, $\log C(\theta)$. As a result, many of the nice features of likelihood inference do not hold for composite likelihood inference. Most importantly for this paper, one cannot count on asymptotic efficiency for the maximum composite likelihood estimators.

If we let $S_k$ be a single variable $y_i$, then we obtain the marginal likelihood for $Y_i$ with $D = 1$, $L(\theta; y_i)$, which we call a *one-wise* likelihood. Setting $S_k = (y_i, y_j)$, yields the marginal likelihood for a pair of observations, $L(\theta; y_i, y_j)$, called a *pairwise* likelihood. If we let $S_k = y_i|y_j$, we obtain $L(\theta; y_i|y_j)$, which we will call a *pairwise conditional likelihood*. The last two likelihoods have data dimension $D = 2$. Since the cost of marginal density calculation is largely determined by $D$, we let $D_{cl}$ be the largest data dimension among the $S_k$ included in $C(\theta)$ in (1.1). Our assumption about computational expense implies that we should first consider using small values of $D_{cl}$.

If we fix an upper bound, say $D^*$, on $D_{cl}$, then we have partially controlled for our computational expense. However, it is clear that one can also control the computational expense by limiting $N_{cl}$, the number of sub-likelihoods that one uses.

There are many possible composite likelihoods for any particular likelihood problem. This paper is concerned with developing strategies that a researcher might use for the design of a composite likelihood, where by composite likelihood design we mean the selection of the sub-likelihoods in $C(\theta)$, including the selection of the number $N_{cl}$, the dimension $D_{cl}$, and the particular $S_k$ (i.e., marginal or conditional densities). We will do so by considering a variety of statistical and computational aspects, as well as overall ease of use and interpretation.

## 2. Basic Issues in Design

The very rich environment of composite likelihood makes it very difficult to make sweeping statements about how they should be constructed. For example, Cox devised the partial likelihood as a justification for an estimation method that eliminated the unknown baseline hazard function from the estimation problem (Cox (1975)). On the other hand, Besag's pseudo-likelihood was an equally clever method for creating straightforward estimation in a model with simple low order conditional densities but an impossibly expensive full likelihood. In keeping with Varin (2008) and Varin, Reid, and Firth (2011), we focus here on issues generated when the lower order marginals are significantly easier to compute than the full density or any of its higher order conditionals.

Given a composite likelihood design with $N_{cl}$ sub-likelihoods each of dimension $D_{cl}$, the overall number of operations needed is of order

$$N_{cl} \times N_{ops}(D_{cl}).$$

Thus if one were to fix computational costs by holding the number of operations fixed, one would face a clear design trade-off. Increasing $D_{cl}$ clearly forces a decrease in $N_{cl}$, and the change needed can be substantial if, for instance, $N_{ops}(D_{cl})$ grows exponentially in $D_{cl}$.

There are, of course, other factors that play into the overall computational challenge that is faced. One is the dimension of the parameter space, $p$. The expense for this depends on the efficiency of optimization algorithms. Further, inference about the parameter $\theta$ could be based on a single realization $\mathbf{y}$, or on the information in a sample $\mathbf{y}_1, \ldots, \mathbf{y}_n$ that is independently drawn from $f(\mathbf{y}; \theta)$. Although growth in the replication variable $n$ increases computational effort, it does so linearly in $n$. For this reason we do not consider varying $n$ explicitly here.

If we construct $C(\theta)$ using all ordered pairs, the *all-pairwise composite likelihood* is

$$C_{apw}(\theta) = \prod_{i<j} L(\theta; y_i, y_j),$$

i.e., the value of $N_{cl}$ is $\binom{d}{2}$ and so grows quadratically in $d$. If the computational effort to produce the full likelihood has calculations that are $O(d^3)$, as in the multivariate normal, then all-pairwise likelihood would have computational advantages for large $d$, while an all-three-wise calculation, with $N_{cl} = \binom{d}{3} = O(d^3)$ might not. The important point here is that increasing $D_{cl}$ without controlling the number of terms $N_{cl}$ can lead to an explosive growth in computational effort.

When we fix computational expense, and turn to other statistical factors, we find the situation is still very complex. Although it would be nice to focus on statistical efficiency, there are also more basic aspects to consider. For example, is the entire parameter identifiable from the composite likelihood? Ensuring that the parameters of interest are identifiable is an important prerequisite for selection of a method. As in Cox's partial likelihood, however, one might be happy with eliminating nuisance parameters. Also, efficiency is a pointwise concept, and we would generally prefer a composite likelihood that had *reliable* efficiency across the parameter space to one that was highly efficient at some parameter values but disastrous at others.

Note that the issue of asymptotic efficiency calls for an analysis of the *composite score*

$$c(\theta) = \nabla \log C(\theta)$$

as an estimating function, so much of this paper concerns composite likelihoods as generators of score functions. Since all such scores fit within the class of linear combinations of sub-likelihood scores, we will consider them in that framework.

In the next three sections we examine the statistical efficiency of the composite score function $c(\theta)$ within the class of unbiased estimating functions for

$\theta$. We consider an optimal weighting strategy in Section 3, and in Section 4 we consider the problem of finding the optimal estimating function within the class of additive estimating equations. The optimal methods are generally not acceptable because they add significant computational costs, but they do suggest new strategies for construction of composite likelihoods and estimating functions that are reliable in efficiency. These issues are examined more carefully in several examples in Section 5.

## 3. Composite Scores as Estimating Functions

We now turn to theory of unbiased estimating functions in order to better understand the efficiency of composite likelihoods. We appeal to the powerful optimality theory that can be used in this setting.

To emphasize its special role in this theory, we use $u_{mle}$ as notation for $\nabla \log f(\mathbf{y};\theta)$. A $p \times 1$ unbiased estimating function $g(\mathbf{y};\theta)$ is one that satisfies $E(g) = 0$. Differentiation under the integral gives use the important relationship $E(u_{mle}g^{\mathrm{T}}) = E(-\nabla g)$. The *Godambe information* in $g$ is defined to be

$$J(\theta; g) = E(-\nabla g)\{\mathrm{var}(g)\}^{-1}E\{-(\nabla g)^{\mathrm{T}}\}.$$

The inverse of $J(\theta; g)$ is the nominal asymptotic covariance matrix for the parameter estimator obtained from a sample $\mathbf{y}_1, \ldots, \mathbf{y}_n$ using $g$. The following concept is important when we consider the reliability of an estimating function. Given any unbiased $g$, consider minimizing over $p \times p$ matrices $R$ in the least squares criterion $E\{(u_{mle} - Rg)(u_{mle} - Rg)^{\mathrm{T}}\}$. The solution is $R_{min} = E(u_{mle}g^{\mathrm{T}})\{\mathrm{var}(g)\}^{-1}$, because $E\{(u_{mle} - R_{min}g)g^{\mathrm{T}}\} = 0$. If we let $g^* = R_{min}g$, then we have

$$E(u_{mle}g^{*\mathrm{T}}) = E(-\nabla g^*) = E(g^*g^{*\mathrm{T}}).$$

We say an estimating function $g^*$ is *information-unbiased* if it satisfies these equalities. We have just shown that an arbitrary estimating function $g$ can be converted to $g^*$ that is information-unbiased by suitable matrix multiplication. Such a multiplication does not change the point estimator. In this form we have the information identity $J(\theta; g) = J(\theta; g^*) = E(-\nabla g^*)$. Every sub-likelihood score is automatically information-unbiased by the second Bartlett identity.

We use the symbol $f$ to represent all marginal and conditional densities derived from $f(\mathbf{y};\theta)$, using the variables with subscripts to indicate which density we mean. So $f(y_i|y_j;\theta)$ will be the density for $Y_i$ given $Y_j = y_j$, for example. With $S_k$ standing for a choice of conditioning or marginal model, we use the notation $u(\theta; S_k)$ for the score $\nabla \log L(\theta; S_k)$.

Given a particular choice of sub-likelihoods, $N_{cl}$ in number, the corresponding composite score is

$$\sum_{k=1}^{N_{cl}} u(\theta; S_k).$$

The following result gives a basic calculation tool that we use repeatedly in our analysis.

**Proposition 1.** *If $S_1$ is a marginal or conditional event and $S_2$ is a marginal event such that all variables appearing in $S_1$ also appear in $S_2$, then*

$$E\{u(\theta; S_1)u^{\mathrm{T}}(\theta; S_2)\} = E\{u(\theta; S_1)u^{\mathrm{T}}(\theta; S_1)\} = J(\theta; u(\theta; S_1)).$$

*In particular, by setting $S_2$ to be the whole data set, the result holds for $u(\theta; S_2) = u_{mle}$.*

We can create a class of estimating functions from the composite scores by considering all their linear combinations. We note that this class is larger than the class of weighted composite likelihood scores, but smaller than the class of all unbiased estimating functions. Let $u(\theta)$ be the $pN_{cl} \times 1$ vector formed by stacking the vectors $u(\theta, S_1), \ldots, u(\theta, S_{N_{cl}})$. Let $W(\theta)$ be an arbitrary $p \times pN_{cl}$ matrix, not depending on data but possibly depending on parameter $\theta$. Then $W(\theta)u(\theta)$ is a $p \times 1$ dimensional set of estimating functions for $\theta$. The composite likelihood estimator corresponds to using $N_{cl}$ side-by-side $p \times p$ identity matrices for $W(\theta)$; call this $W_{cl}(\theta)$.

The estimating function formulation now provides the tools necessary to do an information analysis for the point estimators $\widehat{\theta}_W$ that solve

$$W(\theta)u(\theta) = 0.$$

In this case the Godambe information has the familiar sandwich form

$$J_W(\theta) = W(\theta)G(\theta)[W(\theta)V(\theta)\{W(\theta)\}^{\mathrm{T}}]^{-1}\{W(\theta)G(\theta)\}^{\mathrm{T}}.$$

Here $G(\theta)$ is the $pN_{cl} \times p$ matrix formed by stacking up the information matrices $\mathcal{I}_{S_k}(\theta)$ of the subscores $u(\theta; S_k)$ over $k$ and $V$ is the $pN_{cl} \times pN_{cl}$ covariance matrix of vector $u(\theta)$. There are $p \times p$ blocks along the diagonal of $V(\theta)$ that equal $\mathcal{I}_{S_k}(\theta)$. Thus if the scores $u(\theta; S_1), \ldots, u(\theta; S_{N_{cl}})$ are uncorrelated, the Godambe information of the composite likelihood, using $W_{cl}(\theta)$, is $\mathcal{I}_{S_1}(\theta) + \cdots + \mathcal{I}_{S_{N_{cl}}}(\theta)$. For example, Cox's partial likelihood (Cox (1975)), by its construction, has this ideal property. It is the correlation between the scores that creates a departure from this ideal state.

Unfortunately, there are considerable computational challenges in carrying out the information calculation. Assuming we have a known $W(\theta)$, such as might be derived from a composite likelihood choice such as the all-pairwise likelihood $C_{apw}(\theta)$, one key numerical task is calculating the entries of $V(\theta)$. If we consider the set of all pairwise scores, for example, the calculations require the calculation of many terms like $E(u_{ij}u_{kl})$ for all pairs of pairs. This is a calculation involving

the joint distribution of $(Y_i, Y_j, Y_k, Y_l)$, having data dimension four that we have assumed is expensive. In addition, there are now potentially $\binom{pN_{cl}}{2}$ terms to calculate, where $N_{cl}$ is itself $\binom{d}{2}$.

In this formulation we can also construct the optimal weights, in the sense of information, to be

$$W_{opt}(\theta) = G^{\mathrm{T}}(\theta)V^{-1}(\theta),$$

giving the optimally weighted composite score as

$$w_{opt}(\theta) = W_{opt}(\theta)u(\theta)$$

which has Godambe information

$$J_{opt}(\theta) = G^{\mathrm{T}}(\theta)V^{-1}(\theta)G(\theta).$$

Finding the optimal weights $W_{opt}(\theta)$ can be an enormous numerical challenge as we must invert the $pN_{cl} \times pN_{cl}$ dimensional matrix $V(\theta)$, a calculation that is $O\{(pN_{cl})^3\}$.

We conclude that the theory of optimally weighted estimating equations has limited usefulness for the efficiency problem we address. If we choose to use weighted composite scores, then the optimal solution is very expensive to compute, and the complex structure of the information calculation does not immediately point to simple ways to choose a composite likelihood whose score would have good efficiency properties. However, we should note that one can employ a mixed strategy that increases efficiency with weightings while holding down computational cost. See Kuk (2007) for a particularly clever design, where optimally weighted estimating functions were used for parameters of interest, and flat weights/composite likelihood were used for nuisance parameters.

Returning to composite likelihood, we note that even if one could compute the optimal weights $W_{opt}$, it is very unlikely that there exists a composite likelihood for which $W_{opt}u$ is the score function. Moreover, if we generalize the composite likelihood to a weighted composite likelihood of the form

$$C_w(\theta) = \prod_{k=1}^{N_{cl}} \{L(\theta; S_k)\}^{w_k},$$

then the weights $w_k$ must be positive and constant in $\theta$ if we are to retain the Fisher consistency property. Given a choice between a weighted score approach and a composite likelihood approach, the score approach offers the possibility of increased efficiency. However, if this is impossible to obtain for computational reasons, it would seem that there are advantages to sticking to a composite likelihood. Historically, building inference methods based on an objective function,

such as the least squares criterion or the likelihood function, has been viewed as giving a clear conceptual framework for inference. It can lead to reliable algorithms for optimization that provide resolutions for irregularity problems, such as the multiple solution problem that can arise in solving equations (e.g., Heyde and Morton (1998)), or can give solutions on a boundary when the equations have no interior point solutions (e.g., Self and Liang (1987)). In addition, using a likelihood-like function opens up the possibility of applying Bayesian methodology in the composite framework (e.g., Heagerty and Lele (1998); Christensen and Waagepetersen (2002)).

## 4. Optimal Additive Estimating Functions

Sometimes greater insights to a problem can arise by building a more general framework. In this section this occurs when we generalize the results of Section 3 by changing the class of estimating functions under consideration from linear combinations of sub-likelihood scores to additive estimating functions.

Let $\mathcal{G}_i$ be the class of all $p \times 1$ unbiased estimating functions of the form $g_i(y_i;\theta)$ that have bounded covariance, and $\mathcal{G}_{ij}$ be the class of all $p \times 1$ unbiased estimating functions $g_{ij}(y_i, y_j;\theta)$ that have bounded covariance. We consider all estimating functions in the linear classes of estimating functions

$$\mathcal{L}_1 = \{c + \sum_i g_i(y_i;\theta) : c \text{ is a constant vector, and } g_i \in \mathcal{G}_i\},$$

$$\mathcal{L}_2 = \{c + \sum_i g_i(y_i;\theta) + \sum_{i<j} g_{ij}(y_i, y_j;\theta) : c \text{ is a constant vector}, g_i \in \mathcal{G}_i, \text{ and } g_{ij} \in \mathcal{G}_{ij}\}.$$

We call these classes the *additive functions* of orders 1 and 2, respectively. Here the functions $g$ are $p \times 1$ vector-valued functions. It is easily seen that every linear combination of one-wise, pairwise, and pairwise conditional scores lies in $\mathcal{L}_2$. It is also clear how to generalize $\mathcal{L}_2$ to $\mathcal{L}_k$ $(3 \le k \le d)$, the order $k$ additive functions.

We consider the problem of deriving the optimal estimating function in $\mathcal{L}_k$, using the criterion of Godambe. Once again, the optimal function $A_{k,opt}$ is given by the projection of the full score $u_{mle}$ onto $\mathcal{L}_k$ $(k = 1, 2)$:

$$A_{k,opt} = \arg\min_{A \in \mathcal{L}_k} E\{(u_{mle} - A)(u_{mle} - A)^{\mathrm{T}}\}.$$

Here the minimization takes place in the Loewner ordering; a global minimizer exists because $\mathcal{L}_k$ is a closed linear space. We can identify the solution by the fact that the residuals $u_{mle} - A_{k,opt}$ must be orthogonal to the basis functions $g_i$ and $g_{ij}$.

We note that the minimization needed to find $A_{k,opt}$ is pointwise, in that it is done separately for each parameter value $\theta$. We start the discussion by identifying certain parameter values where the minimization can be carried out explicitly.

**Definition 1.** An "Independence parameter value" $\theta_{ind}$ is any value of $\theta$ in the parameter space $\Theta$ such that when $\theta = \theta_{ind}$ all the variable components $Y_i$ are independent.

### 4.1. Examples of independence parameter values

We clarify what is meant by Definition 1 through examples, and then proceed to giving $A_{k,opt}(\theta_{ind})$, the optimal function at $\theta_{ind}$.

**Example 1.** If $\mathbf{Y}$ has a multivariate normal distribution $N(\mu, \Sigma)$ with $\Sigma = [\sigma_{ij}]$, where $\sigma_{ii} = \sigma^2$ and $\sigma_{ij} = \rho\sigma^2$ $(i \neq j)$, then $\theta = (\mu, \rho, \sigma^2)^{\mathrm{T}}$ is the parameter, and $\theta_{ind} = (\mu, 0, \sigma^2)^{\mathrm{T}}$.

**Example 2.** Consider the Bahadur representation (Bahadur (1961)) for a $d$-dimensional binary response variable $\mathbf{Y} = (Y_1, \ldots, Y_d)^{\mathrm{T}}$. Let $\mu_j = E(Y_j) = P(Y_j = 1)$ be the mean for the $j$th component, and $Z_j = (Y_j - \mu_j)/\sqrt{\mu_j(1-\mu_j)}$ be the standardized variable, $j = 1, \ldots, d$. Let $\rho_{d_1 \cdots d_k} = E(Z_{d_1} \cdots Z_{d_k})$ be the $k$th order correlation among $Z_{d_1}, \ldots$, and $Z_{d_k}$, for any subset $(d_1, \ldots, d_k)$ of the index set $(1, \ldots, d)$. The Bahadur representation (Bahadur (1961); Cox (1972)) of the joint distribution of $(Y_1, \ldots, Y_d)$ is then given by

$$f(y_1, \ldots, y_d) = \prod_{k=1}^{d} \mu_k^{y_k}(1-\mu_k)^{1-y_k} \cdot (1 + \sum_{d_1 < d_2} \rho_{d_1 d_2} \cdot z_{d_1} z_{d_2} + \sum_{d_1 < d_2 < d_3} \rho_{d_1 d_2 d_3} \cdot z_{d_1} z_{d_2} z_{d_3} + \cdots + \rho_{1 \cdots d} \cdot z_1 \cdots z_d). \quad (4.1)$$

A nice property of this representation is that the structure is retained for any subsets of $Y_1, \ldots, Y_d$. That is, for a non-empty subset $D$ of $(1, \ldots, d)$, the induced marginal density for the subvector $(y_j, j \in D)$ is

$$f(y_j, j \in D) = \prod_{j \in D} \mu_j^{y_j}(1-\mu_j)^{1-y_j} \cdot (1 + \sum_{Q \subset D, |Q| \geq 2} \rho_Q \prod_{k \in Q} z_k), \quad (4.2)$$

where $\rho_Q$ represents $\rho_{j_1 \cdots j_k}$ if $Q = \{j_1, \ldots, j_k\}$, and $|Q|$ denotes the number of elements in $Q$. That is, the marginal sub-likelihood for the variables in $D$ is given by (4.2).

If considering the case with an exchangeable correlation structure and a common marginal mean $\mu$, i.e., $\rho_{d_1 \cdots d_k} = \rho$ for any subset $(d_1, \ldots, d_k)$, and $\mu_j = \mu$ for $j = 1, \ldots, d$, then the parameter vector indexing the distribution is $\theta = (\rho, \mu)^{\mathrm{T}}$. It is easily seen that $\theta_{ind} = (0, \mu)^{\mathrm{T}}$.

**Example 3.** Suppose $\mathbf{Y} = (Y_1, \ldots, Y_d)^{\mathrm{T}}$ is a vector of multivariate survival times with marginal survivor function $S_j(y_j)$, $j = 1, \ldots, d$. Let $S(y_1, \ldots, y_d)$ be the joint survivor function for $\mathbf{Y}$. In survival analysis, copula models are commonly used to link the joint and marginal survivor functions. That is,

$$S(y_1, \ldots, y_d) = C\{S_1(y_1), \ldots, S_d(y_d); \phi\},$$

where $C(u_1, \ldots, u_d; \phi)$ is a copula function indexed by parameter $\phi$ (e.g., Joe (1997)). This structure has the appeal that the marginal and association parameters in the joint survivor function can be separately expressed by distinct parameters. If $\beta$ denotes the parameters in the marginal survivor functions $S_j(y_j), j = 1, \ldots, d$, then the parameter vector associated with the joint model is $\theta = (\beta^{\mathrm{T}}, \phi^{\mathrm{T}})^{\mathrm{T}}$. For example, taking the copula function to be

$$C(u_1, \ldots, u_d; \phi) = \psi^{-1}\{\sum_{j=1}^{d} \psi(u_j; \phi); \phi\}, \ 0 \le u_j \le 1, j = 1, \ldots, d$$

leads to the so-called Archimedean family, where the range of $\psi(u; \phi)$ is between 0 and 1, and $\psi(u; \phi)$ is completely monotonic, i.e., $\psi(u; \phi)$ is differentiable with $(-1)^r \psi^{(r)}(u; \phi) > 0$, $r = 1, 2, \ldots$. In particular, setting $\psi(u; \phi) = u^{-1/\phi} - 1$ with $\phi > 0$ leads to the Clayton models with

$$S(y_1, \ldots, y_d; \theta) = [\{S_1(y_1; \beta)\}^{-\phi} + \cdots + \{S_d(y_d; \beta)\}^{-\phi}]^{-1/\phi}.$$

As $\phi \to 0$, $Y_1, \ldots, Y_d$ become independent. So we can write $\theta_{ind} = (\beta^{\mathrm{T}}, 0)^{\mathrm{T}}$. Similarly, setting $\psi(u; \phi) = (-\log u)^{1/\phi}$ with $0 < \phi < 1$, or $\psi(u; \phi) = -\log\{(\phi^u - 1)/(\phi - 1)\}$ with $1 \le \phi$ yields the positive stable frailty model, or Frank model, respectively. In both cases, $Y_1, \ldots, Y_d$ become independent as $\phi \to 1$. Thus, $\theta_{ind}$ can be taken as $(\beta^{\mathrm{T}}, 1)^{\mathrm{T}}$.

### 4.2. Hoeffding scores

We return to the problem of finding $A_{opt,k}$, the optimal additive estimating function. We first do so at values of $\theta_{ind}$. We start by defining the Hoeffding one-wise, pairwise, and three-wise scores by

$$\begin{aligned} u_i^{hfd} &= u_i, \\ u_{ij}^{hfd} &= u_{ij} - u_i - u_j, \\ u_{ijk}^{hfd} &= u_{ijk} - u_{ij} - u_{ik} - u_{jk} + u_i + u_j + u_k, \end{aligned}$$

where $i, j$, and $k$ are distinct integers between 1 and $d$. We can similarly define the Hoeffding $k$-wise score $u^{hfd}_{i_1\cdots i_k}$. We call these the Hoeffding scores because they are derived from the Hoeffding decomposition of a $U$-statistic (Hoeffding (1948); Lee (1990, p.26)). We can then define the *Hoeffding additive scores* of various orders by

$$h_1(\theta) = \sum_i u^{hfd}_i,$$
$$h_2(\theta) = h_1(\theta) + \sum_{i<j} u^{hfd}_{ij},$$
$$h_3(\theta) = h_2(\theta) + \sum_{i<j<k} u^{hfd}_{ijk}.$$

The $k$th order Hoeffding additive scores $h_k(\theta)$ $(1 \le k \le d)$ can be defined recursively.

**Theorem 1.** *For $1 \le k \le d$, $h_k(\theta_{ind}) = A_{k,opt}(\theta_{ind})$. That is, at $\theta_{ind}$, the Hoeffding additive scores $h_k(\theta)$ are the information optimal estimating functions in $\mathcal{L}_k$.*

The proof involves showing the residuals are orthogonal to the basis; see Appendix A. The following theorem tells us a bit more about this solution.

**Theorem 2.** *Under $\theta_{ind}$, the Hoeffding $k$-wise scores are mutually orthogonal, over all orders, so that, for example, $E\{u^{hfd}_{ij}(u^{hfd}_{ijk})^{\mathrm{T}}\} = 0$. Moreover, they are all information-unbiased, so that*

$$E(-\nabla u^{hfd}_{ij}) = E\{u^{hfd}_{ij}(u^{hfd}_{ij})^{\mathrm{T}}\} = J(\theta_{ind}; u^{hfd}_{ij}).$$

*As a result, there is a decomposition of the information in $h_k(\theta_{ind})$ of the form*

$$J\{\theta_{ind}; h_k(\theta_{ind})\} = \sum_i J(\theta_{ind}; u^{hfd}_i) + \sum_{i<j} J(\theta_{ind}; u^{hfd}_{ij}) + \cdots + \sum_{i_1<\cdots<i_k} J(\theta_{ind}; u^{hfd}_{i_1\cdots i_k}).$$

The proof of this result is straightforward. Notice that the calculation of the individual information terms is also likely to be less expensive due to the independence that occurs at $\theta_{ind}$.

To describe these results in a more suggestive way, we might say the optimal order-2 score arises by first constructing the optimal one-wise score, and then adding the Hoeffding pairwise scores, which contain only the *new* information $J(\theta; u^{hfd}_{ij})$ that is found in the pairwise density but not the one-wise densities. That is, $u^{hfd}_{ij}$ is orthogonal to all one-wise scores $u'_i$s and so contains only the non-redundant information found in the pair $(y_i, y_j)$. We think it is important

to keep in mind this basic paradigm when constructing either scores or composite likelihood: each additive term should, if possible, provide new, orthogonal, information only.

It is clear that this result must have implications for the structure of the optimal weighting matrix for one-wise and pairwise scores at $\theta_{ind}$ as well. If we construct the optimal weighting matrix $W_{opt}(\theta)$ for the Hoeffding one-wise and pairwise scores $u_i^{hfd}$ and $u_{ij}^{hfd}$, then $W_{opt}(\theta)$ is a stack of $p \times p$ identity matrices, which shows that the Hoeffding additive score $h_2(\theta)$ is the optimal weighting solution (this is a weaker result than being optimal additive). Assuming $W_{opt}(\theta)$ is continuous in $\theta$, this means $W_{opt}(\theta)$ should be near the Hoeffding weights when $\theta$ is near $\theta_{ind}$.

As we will see later, the Hoeffding optimality results can have very limited value except for parameter values that are near the independence parameters. On the other hand the theorem holds for absolutely any model, and so represents a breakthrough of sorts in a problem that has so far seen little in the way of general theory.

### 4.3. Parameter classification

The Hoeffding optimality results enable us to create a classification of the parameters based on their Hoeffding scores. We say that an element of the parameter vector has *no Hoeffding one-wise information* if the corresponding diagonal element of $J(\theta_{ind}, u_i)$ is zero for all $i$. We say that the parameter element has *no Hoeffding pairwise information* if the corresponding diagonal element of $J(\theta, u_{ij}^{hfd})$ is zero for all $i$ and $j$.

These definitions are important to the theory as follows. If there is no one-wise information about a parameter, then each Hoeffding pairwise score $u_{ij} - u_i - u_j$ is the corresponding pairwise score $u_{ij}$. It follows that the second order Hoeffding additive score $h_2(\theta)$ is just the all-pairwise score $c_{apw}(\theta)$ for that parameter. On the other hand, if there is no pairwise information in any pair, then $h_2(\theta)$ must equal $c_{aow}(\theta)$, the all-one-wise score. However the corresponding equality $u_{ij} = u_i + u_j$ now implies that $h_2(\theta)$ itself is a multiple of $c_{apw}(\theta)$. Thus in either case, $h_2(\theta)$ is equivalent to the all-pairwise score.

We give an example of each of these two parameter types in our examples. It follows from the above remarks that at $\theta_{ind}$, the Hoeffding scores are no more efficient than pairwise for these parameter types. However, if we consider a neighborhood of $\theta_{ind}$, we know that the optimal weighting matrix should stay close to the weights given by Hoeffding scores, and so Hoeffding scores should be superior in efficiency to pairwise scores in some neighborhood of $\theta_{ind}$. Our examples bear this out.

### 4.4. Results about general $\theta$

The rather surprising nature of the optimal additive scores at $\theta_{ind}$ leads one to consider the consequences for non-independence values of $\theta$. For example, are the optimal additive scores always linear combinations of marginal scores? The answer to this question is no. In fact the best additive scores appear to be generally more difficult to compute than the optimal weighted combination of scores. To illustrate this point, consider the multivariate normal model, $N\{\mu(\theta), \Sigma(\theta)\}$. The log likelihood is quadratic in the data $\mathbf{y}$, and hence the score $u_{mle}(\theta)$ is as well. This means that $u_{mle}(\theta)$ lies in $\mathcal{L}_2$. It follows that the second order optimal additive estimating function is $u_{mle}(\theta)$ itself, as nothing else could be more efficient. That is, we have gained no computational advantage in using the additive estimating function approach - unless perhaps the result inspires a new algorithm for finding the maximum likelihood estimator (MLE).

The additive decomposition of information in Theorem 2 is limited to independence, but we can create an extension that sheds some light on the construction of the optimal scores more widely. Suppose we have created an estimating function $g(\mathbf{y};\theta)$. We consider how we might improve its information by adding a subset of the variables in $\mathbf{Y}$, say $S$. This problem is made somewhat easier by generalizing it, and finding the matrix $R$ and function $a$ that maximize $J(\theta; Rg + a)$ over all $p \times p$ matrices $R$ and all functions $a(S;\theta)$. The solution is as follows.

**Proposition 2.** *The matrix $R_{opt}$ is the matrix that makes $R_{opt} \cdot [g(\mathbf{y};\theta) - E\{g(\mathbf{Y};\theta)|S_k\}]$ information-unbiased. The function $a_{opt}(S_k;\theta)$ is then given by*

$$a_{opt}(S_k;\theta) = u(\theta; S_k) - E\{R_{opt} g(\mathbf{Y};\theta)|S_k\}.$$

*In this case*

$$J(\theta; R_{opt} g + a_{opt}) = J(\theta; [g(\mathbf{y};\theta) - E\{g(\mathbf{Y};\theta)|S_k\}]) + J\{\theta; u(\theta; S_k)\}.$$

The proof is outlined in Appendix B. There are several important conceptual lessons in this result. First, this clearly shows that the marginal score $u(\theta; S_k)$ is always a fundamental building block for adding the information in a set of variables $S_k$. But secondly, if we wish to add a new score $a(S_k;\theta)$ to an existing estimating equation, it is best to do so by first constructing $g^*$ so as to remove the information in $g(\mathbf{y};\theta)$ that can be attributed to $S_k$, then adding it back in through $u(\theta; S_k)$. This result also gives the following corollary that provides a non-trivial characterization of the optimal estimating functions in any additive class.

**Corollary 1.** *The estimating function $g$ is optimal additive estimating function in the additive class of estimating functions $\mathcal{G}^* = \{\Sigma_k g_k(S_k;\theta)\}$ if and only if*

$$E(R_{opt}g|S_k) = u(\theta;S_k)$$

*with probability one for every $k$.*

We note that one can in theory create an algorithm similar to the backfitting algorithm of generalized additive models (Hastie and Tibshirani (1990)) that would iteratively determine the optimal additive function. One would cycle through the subsets of interest $S_k$, adding adjustments one at a time and improving information at each step. However, this presents numerous new computational challenges.

### 4.5. Hoeffding scores and likelihood

We next examine whether the Hoeffding additive scores $h_k(\theta)$ are themselves composite likelihood scores for some true or weighted composite likelihood. If true, then these composite likelihoods have an attractive local optimality property near $\theta_{ind}$ relative to other composite likelihoods. The answer to this question is mostly negative, which is somewhat disappointing.

To be more precise about this, we first construct the likelihood-based objective function whose differentiation yields the Hoeffding scores. Recursively, define the *Hoeffding likelihoods* by

$$\begin{aligned}
L_{H_1}(\theta) &= \prod_i f(y_i;\theta),\\
L_{H_2}(\theta) &= L_{H_1}(\theta)\cdot\prod_{i<j}\frac{f(y_i,y_j;\theta)}{f(y_i;\theta)f(y_j;\theta)},\\
L_{H_3}(\theta) &= L_{H_2}(\theta)\cdot\prod_{i<j<k}\frac{f(y_i,y_j,y_k;\theta)f(y_i;\theta)f(y_j;\theta)f(y_k;\theta)}{f(y_i,y_j;\theta)f(y_i,y_k;\theta)f(y_j,y_k;\theta)}.
\end{aligned}$$

The higher order Hoeffding likelihoods $L_{H_k}$ are determined in a similar fashion from the higher order Hoeffding scores. Note that $h_k(\theta) = \nabla L_{H_k}(\theta)$ is the $k$th order Hoeffding additive score.

It is easily seen that the Hoeffding likelihood $L_{H_1}(\theta)$ based on the one-wise marginal likelihoods is a true composite likelihood. On the other hand, the second-order Hoeffding additive score is quite surprising:

$$\begin{aligned}
h_2(\theta) &= \sum_{i<j} u_{ij} - (d-2)\sum_i u_i\\
&= c_{apw}(\theta) - (d-2)c_{aow}(\theta). \qquad (4.3)
\end{aligned}$$

The striking feature here is the large negative weight on the one-wise scores. Because a composite likelihood score is a sum of marginal and conditional scores, the true composite likelihood that comes closest to having such a relatively large negative weight on the one-wise scores is the all-pairwise-conditionals composite likelihood, $C_{apc}(\theta) = \prod_{i \neq j} L(\theta; y_i|y_j)$, whose score is

$$c_{apc}(\theta) = 2\left\{ \sum_{i<j} u_{ij} - \left(\frac{d-1}{2}\right) \sum_i u_i \right\}.$$

The details are given in Appendix C. We note that when $d = 3$, $c_{apc}$ and $h_2$ are proportional, and the likelihoods are as well. However, as $d$ increases to larger values, there is a significant difference in weighting of the one-wise scores. When we go beyond pairwise to three-wise or more, we find that there is no overlap between the Hoeffding additive scores and the scores for any composite likelihood except for the trivial case when $k = d$, i.e., when the $k$-wise score is the full likelihood score.

### 4.6. Modified Hoeffding scores

We now know that the Hoeffding likelihoods $L_{H_k}(\theta)$ are not composite likelihoods for $k > 1$, but also that they are superior in information to any composite likelihood at $\theta_{ind}$. How well would they work if used for arbitrary $\theta$? The answer is "very poorly". Our examples in Section 5 show that, as the density $f$ deviates from independence, the Hoeffding likelihoods can behave catastrophically worse than the composite likelihoods we consider. Before proceeding to these results, we consider what might go wrong with the Hoeffding additive scores, and in the process we show how one can improve them while retaining local optimality at $\theta_{ind}$.

Recall that the Hoeffding pairwise scores $u_{ij}^{hfd} = u_{ij} - u_i - u_j$ have the property that they are orthogonal at $\theta_{ind}$ to all the other scores. Thus by their construction, they represent only the new information that was not in those scores. Their weakness is that this orthogonality does not hold away from independence.

In the spirit of Proposition 2, we propose to replace $u_{ij}^{hfd}$ with a pairwise term we call the *centered pairwise score*,

$$u_{ij}^{cen} = u_{ij} - B_i u_i - B_j u_j.$$

Here the $p \times p$ weight matrices $B_i(\theta)$ and $B_j(\theta)$ are used to remove the marginal effects of $u_i$ and $u_j$ at each $\theta$ value, making the centered pairwise score orthogonal to those marginal scores at that $\theta$.

The idea is to force $u_{ij}^{cen}$ to represent the new information in the pair that was not already present in the one-wise scores $u_i$ and $u_j$. If $u_{ij}^{cen} = 0$ for all $i$ and

$j$ and for some parameter element $\theta_1$, we can then say that there is no *additional information (beyond the two one-wise scores) in the pairwise density* about $\theta_1$. With this definition, we have the intuitively pleasing result that when we have a parameter $\theta_1$ in the $N\{\mu(\theta), \Sigma(\theta)\}$ model that only affects the mean parameter $\mu(\theta)$, then there is no additional pairwise information about $\theta_1$. A further remark on this is placed at the end of the section.

We can solve for $B_i$ and $B_j$ by letting them be the $\beta_i$ and $\beta_j$ that minimize the matrix least squares criterion

$$E\{(u_{ij} - \beta_i u_i - \beta_j u_j)(u_{ij} - \beta_i u_i - \beta_j u_j)^{\mathrm{T}}\}.$$

The minimizers in the Loewner ordering satisfy the matrix equalities:

$$E\{(u_{ij} - B_i u_i - B_j u_j)u_i^{\mathrm{T}}\} = 0,$$

$$E\{(u_{ij} - B_i u_i - B_j u_j)u_j^{\mathrm{T}}\} = 0.$$

Assuming the regularity conditions in Section 1 and noting that $E(u_{ij}u_r^{\mathrm{T}}) = E(u_r u_r^{\mathrm{T}}), r = i, j$, with a little algebraic manipulation one obtains the formulas

$$B_i = (C_{ij}\mathcal{I}_j^{-1} - \mathcal{I}_i C_{ji}^{-1})^{-1}(I_p - \mathcal{I}_i C_{ji}^{-1}),$$

$$B_j = (C_{ji}\mathcal{I}_i^{-1} - \mathcal{I}_j C_{ij}^{-1})^{-1}(I_p - \mathcal{I}_j C_{ij}^{-1}),$$

where $I_p$ is the $p \times p$ identity matrix, and

$$\begin{aligned} C_{ij} &= E(u_i u_j^{\mathrm{T}}), \\ C_{ji} &= C_{ij}^{\mathrm{T}}, \\ \mathcal{I}_i &= E(u_i u_i^{\mathrm{T}}), \\ \mathcal{I}_j &= E(u_j u_j^{\mathrm{T}}). \end{aligned}$$

With these choices of $B_i$ and $B_j$, the centered pairwise score $u_{ij}^{cen}$ satisfies the second Bartlett identity, and is orthogonal to any linear combination of $u_i$ and $u_j$, the two scores with which $u_{ij}$ is intrinsically dependent. (This is proved later in this section).

On the computational side, note that the weight matrices $B_i$ and $B_j$ are constructed from the pairwise density, nothing higher order, so the problem still has $D_{cl} = 2$. Of course, if the parameter dimension $p$ is large, there could be significant computational burden in finding the needed inverses.

In order to assess the usefulness of the centered pairwise scores, we propose to examine the *modified Hoeffding score* $h_2(\theta)$,

$$h_2^{mod}(\theta) = \sum u_i + \sum u_{ij}^{cen}. \tag{4.4}$$

This proposal is again based on the assumption that we must avoid any weight matrix calculations involving the full set of scores. When $h_2^{mod}(\theta)$ is constructed in this way, it matches the Hoeffding additive score $h_2(\theta)$ at $\theta_{ind}$, and so is locally optimal.

We conclude this section by showing that $u_{ij}^{cen}$ mimics the score for a sub-likelihood in a manner that $u_{ij}^{hfd}$ does not, namely information unbiasedness. In the next section we give an argument for why this should make it a more reliable method.

We next verify that $u_{ij}^{cen}$ has the information-unbiased property.

**Proposition 3.**

$$E(-\nabla u_{ij}^{cen}) = E(u_{ij}^{cen} u_{mle}^{\mathrm{T}}) = E\{u_{ij}^{cen}(u_{ij}^{cen})^{\mathrm{T}}\}.$$

**Proof.** The first equality is a standard interchange of derivative and integral. If we calculate the middle expression by first conditioning on $(Y_i, Y_j)$, we find that it is equal to $E(u_{ij}^{cen} u_{ij}^{\mathrm{T}})$. This is then equal to $E\{u_{ij}^{cen}(u_{ij}^{cen})^{\mathrm{T}}\}$ by the orthogonality of $u_{ij}^{cen}$ to $u_i$ and $u_j$.

**Remark 1.** At $\theta_{ind}$, the Hoeffding score decomposition gives an ideal orthogonal decomposition of scores so that information is additive over pieces. In such a case there is no ambiguity when we say that when $u_{ij}^{hfd} = 0$, there is no additional information about a parameter in the pairwise score after accounting for one-wise information; $u_{ij}^{hfd}$ is not just orthogonal to $u_i$ and $u_j$, it is orthogonal to any $u_k$. At other values of $\theta$, we do not have $u_{ij}^{cen}$ orthogonal to $u_k$ for other $k$ than $i$ and $j$. Our definition above treats the pair $(Y_i, Y_j)$ as an island even though the additional information in the pair over the one-wise could, for example, be defined through Proposition 2 as $J[\theta; c_{aow} - E\{c_{aow}|(Y_i, Y_j)\}] + J(\theta; u_{ij}) - J(\theta; c_{aow})$. Our choice is a computationally practical simplification of this that has some intuitive appeal.

### 4.7. Reliable versus optimal

We have argued that the Fisher consistency property of the true composite likelihood is a desirable property: it arises because we add together terms, each of which satisfies the Kullback-Leibler inequality. However, if we were to include sub-likelihoods with negative weights, the guarantee of Fisher consistency would be lost. That is, adding together the log sub-likelihoods is a reliable mechanism, even though it is not necessarily most *efficient* - we know this because the Hoeffding result implies that one needs to use negative weights for optimal efficiency near $\theta_{ind}$. Given the high computational cost of optimality, we propose

to direct our attention towards lower cost methods that are inherently reliable in consistency and efficiency.

If log sub-likelihoods can be reliably combined by summing with positive weights, how can we reliably combine estimating functions $g_1, \ldots, g_k$? Even at the most primitive level, one should know which ones have negative signs and which ones positive when forming $g_1 \pm \cdots \pm g_k$. Our proposed rule for reliability is that *one should add information-unbiased scores (or at least use positive weights).* We recall that any estimating function $g_i$ that is not information-unbiased can be put into this form by matrix multiplication, so our proposed rule provides a general way to combine estimating functions.

We can claim no optimality theorem for this rule, but we can make the following observations about reliability. First, if the scores are uncorrelated and information-unbiased, then adding them together with equal weight is the most efficient of all weighting schemes. Secondly, if this rule is applied to sub-likelihood scores, it exactly replicates the class of composite likelihood scores, as the sub-likelihood scores are automatically information-unbiased.

There is a third, and deeper, reason that this rule is reasonable. An unbiased estimating function $g$ that is *information-unbiased* satisfies

$$E(-\nabla g) = E(g u_{mle}^{\mathrm{T}}) = E(g g^{\mathrm{T}}) \geq 0,$$

where the inequality is in a sense of the Loewner ordering. If we examine the last display, we note that the second expression $E(g u_{mle}^{\mathrm{T}})$ gives the covariances of the $g$ score with $u_{mle}$, while the third expression $E(g g^{\mathrm{T}})$ is a non-negative definite matrix. That is, information-unbiasedness implies that the estimating function $g$ has a positive association with the full likelihood score. We call this *positive likelihood association.* If $g$ were the gradient of an objective function, such as $u(\theta; S_k)$, positive likelihood association has a physical interpretation: the steepest ascent direction on $\log L(\theta; S_k)$, namely $u(\theta; S_k)$, has a positive association with the steepest ascent direction on $\log L(\theta)$, namely $u_{mle}$.

Now if we add together estimating functions all with positive association with $u_{mle}$, the sum has positive association. However, if we were to add some with negative weights, this guarantee would be lost. So our proposed rule for summing estimating functions guarantees positive likelihood association.

**Remark 2.** One can also view our rule as "semi-optimizing" the Godambe information formula $J(\theta; g) = E(u_{mle} g^{\mathrm{T}})\{\mathrm{var}(g)\}^{-1} E(g u_{mle}^{\mathrm{T}})$. By adding together terms with positive association, we are making sure that the expressions $E(-\nabla g) = E(g u_{mle}^{\mathrm{T}})$ increase as we add terms. Using a negative sign would make these "numerator" terms decrease, so might be a bad idea. Our naivety, from the efficiency point of view, comes from ignoring the costs that are associated with $\{\mathrm{var}(g)\}^{-1}$ when we use this rule.

We now can introduce a corollary to Proposition 3.

**Corollary 2.** *The modified second order Hoeffding score $h_2^{mod}(\theta)$ is a sum of terms with positive likelihood association.*

In contrast, the Hoeffding pairwise score $u_{ij}^{hfd}$ is not guaranteed to have positive likelihood association, as we show in the following example. Therefore, $h_2(\theta)$ itself has no guarantee of having positive association. For this reason we might find its reliability suspect.

**Example 4.** Suppose

$$\mathbf{Y} = (Y_1, \ldots, Y_d)^{\mathrm{T}} \sim N(\theta 1_d, \Sigma),$$

where $1_d$ is the $d \times 1$ unit vector, and $\Sigma$ is a $d \times d$ matrix with diagonal elements 1 and off-diagonal elements $\rho$. Here $\theta$ is an unknown parameter, but $\rho$ is assumed known.

For $i \neq j$, we have the pairwise and one-wise scores

$$u_{ij} = \frac{1}{1+\rho}(y_i + y_j - 2\theta), \quad u_i = y_i - \theta, \quad u_j = y_j - \theta,$$

which yield

$$E(\nabla u_{ij}) = -\frac{2}{1+\rho}, \quad E(\nabla u_i) = E(\nabla u_j) = -1.$$

Therefore,

$$\begin{aligned} E(u_{ij}^{hfd} u_{mle}) &= E(u_{ij}u_{mle}) - E(u_i u_{mle}) - E(u_j u_{mle}) \\ &= E(-\nabla u_{ij}) - E(-\nabla u_i) - E(-\nabla u_j) \\ &= \frac{-2\rho}{1+\rho}, \end{aligned}$$

so the Hoeffding pairwise score has negative likelihood association when the correlation coefficient $\rho$ is positive.

We can construct the centered pairwise scores as follows. Since $E(u_i^2) = E(u_j^2) = 1$, and $E(u_i u_j) = \rho$, we have $B_i(\theta) = B_j(\theta) = 1/(1+\rho)$, leading to the centered pairwise score

$$u_{ij}^{cen} = u_{ij} - \left(\frac{1}{1+\rho}\right) u_i - \left(\frac{1}{1+\rho}\right) u_j.$$

That is, how much one-wise scores should be removed from the pairwise score depends on how strongly the one-wise scores are related.

In fact, in this example $u_{ij}^{cen} = 0$ at $\rho = 0$, showing that all the pairwise information about the mean parameter $\theta$ is in the one-wise scores at this value.

## 5. Numerical Illustrations

In this section we provide a numerical study to compare the performance of different ways to combine composite scores. We consider a multivariate normal distribution $N(0, \Sigma)$ with the covariance matrix $\Sigma = [\sigma_{ij}]$. There is a single scalar parameter $\theta$, so that $p = 1$. The examples are chosen so that we can do relatively simple exact efficiency calculations, not for their intrinsic interest. We restrict attention here to $D_{cl} = 2$, and so will focus on various combinations of one-wise and pairwise scores.

We compare the following five methods.

1. All-pairwise likelihood, with $c_{apw}(\theta) = \sum_{i<j} u_{ij}$.
2. All-pairwise conditionals, with $c_{apc}(\theta) = 2[c_{apw}(\theta) - \{(d-1)/2\} \cdot c_{aow}(\theta)]$, where $c_{aow}(\theta)$ is the all-one-wise score $\sum_i u_i$.
3. The second order Hoeffding additive scores $h_2(\theta)$ defined in (4.3).
4. Modified second order Hoeffding scores $h_2^{mod}(\theta)$ defined in (4.4).
5. Maximum likelihood, with score $u_{mle}$.

In each of our examples there is an exchangeable distribution for the vector $\mathbf{Y}$. Under such circumstances, it is an easy argument to show that the optimal weights for each $u_i$ are the same, and the weights for each $u_{ij}$ are the same. That is to say, the linear optimal estimating function $w_{opt}(\theta)$ with $D_{cl} = 2$ is of the form

$$\alpha_1 \sum_i u_i + \alpha_2 \sum_{i<j} u_{ij} \tag{5.1}$$

for constants $\alpha_1$ and $\alpha_2$. Moreover, the score $u_{mle}$ in the normal likelihood lies in the linear class of estimating functions with $D_{cl} = 2$ (as it depends on the data linearly through $y_i$ and $y_i y_j$ products only), so in fact $u_{mle}$ is $w_{opt}(\theta)$. Since all the methods we consider have the form (5.1), they are competitors to be equivalent to MLE, provided they correspond to good choices of $\alpha_1$ and $\alpha_2$.

If the parameter $\theta$ has dimension one, then the efficiency of the various estimators is determined by their relative weight

$$RW = -\frac{\alpha_1}{\alpha_2}.$$

For example, for all-pairwise $c_{apw}(\theta)$, this ratio is always zero.

Table 1. Relative weights comparison for constant correlation model.

| Score | $RW$ | Same as MLE |
|---|---|---|
| True | $\frac{d-2}{1+\rho}$ | all $\rho$ |
| Pairwise | 0 | never |
| All cond'ls | $\frac{d-1}{2}$ | $\rho = 1 - \frac{2}{d-1}$ |
| Hoeffding | $d-2$ | $\rho = 0$ |
| Mod. Hoeffding | $-\frac{\rho^2}{1+\rho^2} + \frac{d-2}{1+\rho^2}$ | $\rho = 0, \rho \approx 1 - \frac{1}{d-1}$ |

In such a simple setting one might expect all methods to do well. However, we will see that ramping up the data dimension $d$ severely erodes efficiency because, in fact, none of the methods comes close to having the same relative weights as the MLE.

### 5.1. Exchangeable normal: mean and correlation known

Suppose $\mathbf{Y} = (Y_1, \ldots, Y_d)^{\mathrm{T}} \sim N(0, \Sigma)$, where $\Sigma = \sigma^2 \cdot \{(1-\rho)I_d + \rho 1_d 1_d^{\mathrm{T}}\}$. Here, $\sigma^2$ is the unknown parameter and $\rho$ is assumed to be a known constant. We start by displaying in Table 1 the relative weights $RW$ described above. True to our local optimality theory, the relative weights for MLE and Hoeffding are identically $d-2$ at the independence model $\rho = 0$, as is the modified Hoeffding relative weight. That is, Hoeffding and modified Hoeffding are both locally equivalent to the MLE in this case, while all-pairwise and all-pairwise conditionals are not. However, we are interested in behaviour away from $\rho = 0$, so we also check to see if there are any other values of $\rho$ for which a method might be locally equivalent to the MLE, and so be fully efficient. Here we see that both all-pairwise conditionals and modified Hoeffding have this property at exactly one value of $\rho$. These values are near 1 if $d$ is large. Thus we might expect the latter two methods to be reliably efficient at larger values of $\rho$.

To carry out the efficiency calculations, we note that under the normal distribution, all above five types of scores can also be written as $Y^{\mathrm{T}}AY - E(Y^{\mathrm{T}}AY)$ with suitable matrices $A$. In addition, the matrix $A$ can be expressed as a linear combination of $A_{aow}$ for one-wise marginal score and $A_{apw}$ for pairwise score according to the same weights $\alpha_1$ and $\alpha_2$ mentioned above. Therefore, we can compare the relative efficiency, $r$, between composite score estimator, $\widehat{\theta}_{cs}$, and MLE, $\widehat{\theta}_{mle}$, as

$$\begin{aligned} r &= \frac{\mathrm{avar}(\widehat{\theta}_{mle})}{\mathrm{avar}(\widehat{\theta}_{cs})} \\ &= \frac{\{\mathrm{tr}(A_{mle}\Sigma A_{cs}\Sigma)\}^2}{\mathrm{tr}(A_{mle}\Sigma A_{mle}\Sigma) \cdot \mathrm{tr}(A_{cs}\Sigma A_{cs}\Sigma)}. \end{aligned}$$

We compare the relative efficiencies of all methods when $d = 3, 20, 50$ for values of $\rho$ in its range $(-1/(d-1), 1)$. Note that when $d = 3$, all-pairwise conditionals and Hoeffding are identical. Note also that this is a model in which the Hoeffding pairwise scores $u_{ij} - u_i - u_j$ are all zero at $\rho = 0$. That is, there is no additional information about the mean in the pairwise densities at this value. So at this $\rho$, all the methods we are considering are equivalent to the all-one-wise score, and so all have relative efficiency of 1 at $\rho = 0$.

At $d = 3$, it is clear that all-pairwise is inferior to the other methods, especially so for large $\rho$. The local optimality of the other methods near $\rho = 0$ is also apparent.

At $d = 20$, there seems to be less benefit to having locally optimal weights in $h_2(\theta)$ and $h_2^{mod}(\theta)$, as their regions of high efficiency are considerably smaller. It also becomes clear that we cannot consider $h_2(\theta)$ an acceptable method for estimation, as it has disastrous efficiency beyond $\rho = 0.2$. The comparison between all-pairwise and modified Hoeffding is something of a tie, with the latter better for large $\rho$. And all-pairwise-conditionals is a clear overall winner. The plot for $d = 50$ provides nearly identical information, but now the worst case efficiency for all-pairwise conditionals has degraded to about 0.50.

### 5.2. Exchangeable normal: mean and variance known

We consider a second simple example. Suppose $\mathbf{Y} = (Y_1, \ldots, Y_d)^{\mathrm{T}} \sim N(0, \Sigma)$, where $\Sigma^{-1} = \sigma^{-2}\{(1-\beta)I_d + \beta 1_d 1_d^{\mathrm{T}}\}$. Here, $\sigma^2$ is treated as a known parameter and $\beta$ is assumed to be unknown, $\beta = 0$ corresponding to $\theta_{ind}$. In this case, the weights for the MLE and for the modified Hoeffding are more complex, and so we do not repeat Table 1. However, we should note that at $\beta = 0$, the one-wise scores are zero, so in contrast with the first example, there is no one-wise information at $\beta = 0$. However, this again implies that all the competing methods are equivalent to pairwise, and hence fully efficient, at $\beta = 0$.

Again, we compare the relative efficiencies for our five methods when $d = 3, 20, 50$, under different values for $\beta \in (-1/(d-1), 1)$. Starting with $d = 3$, we see that the locally optimal methods are both superior to pairwise in a neighborhood of $\beta = 0$. Other than this, pairwise is a little better than Hoeffding/all-pairwise conditionals at the two extremes of $\beta'$s range, but much worse for $\beta \approx 0.3$. Modified Hoeffding is an overall winner. All methods become fully efficient as $\beta$ goes to 1.

For $d = 20$, the performance of Hoeffding $h_2$ is so bad that we left it off the plot. Examining the other three, we can see that pairwise is largely inferior in efficiency to the other two. Both all-pairwise conditionals and pairwise deteriorate badly near $\beta = 0.1$. We do not have a simple explanation for this weak performance. On the other hand, modified Hoeffding performs the best, with clear

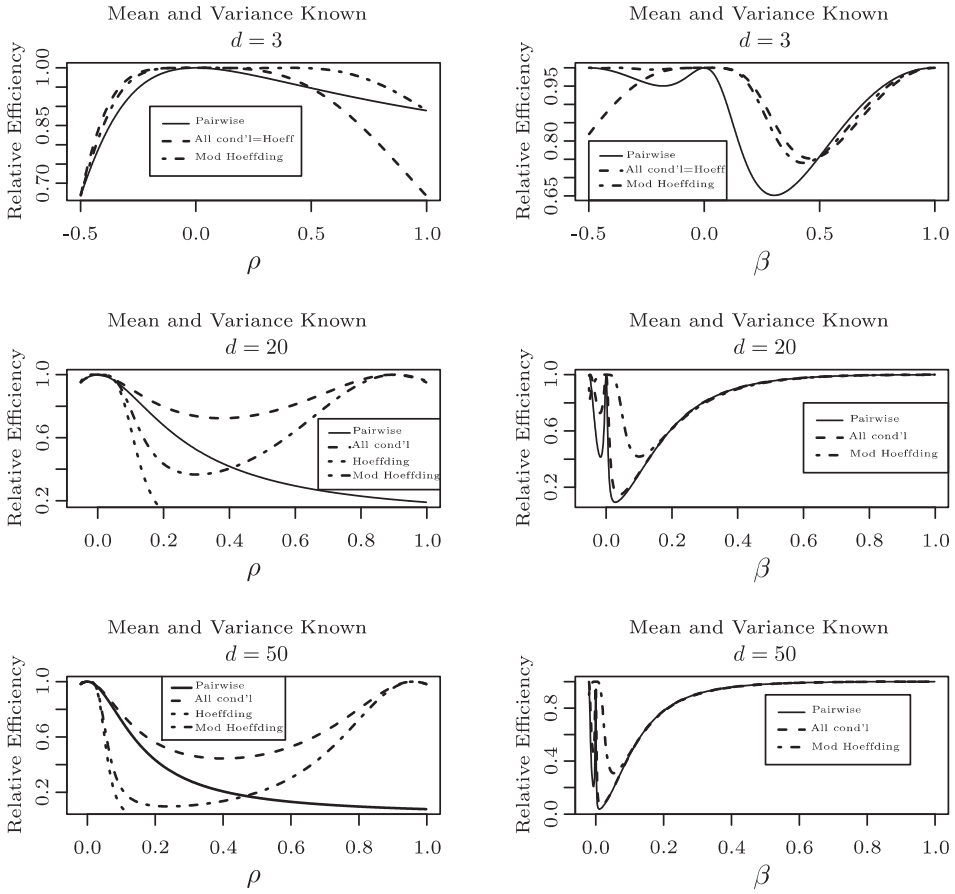


Figure 1. Comparisons of the relative efficiency for different composite likelihoods under various settings.

local optimality. Among the true composite likelihoods, all-pairwise conditionals was best. The same conclusions hold for $d = 50$, except that the worst case efficiencies decay further, with modified Hoeffding going down to about 30%.

## 6. Strategies

The Hoeffding optimality result and the seeming success of the modified Hoeffding scores offer some strategic insights. One clear insight is that the pairwise likelihood is generally not "conditional" enough. That is, it contains too many copies of the one-wise information, in fact $d - 2$ times too much near $\theta_{ind}$. Another lesson is that we can correct for this feature somewhat at other values of $\theta$ than $\theta_{ind}$ by using $u_{ij}^{cen}$, and obtain methods that were clearly superior to pairwise in our examples. It is also clear that all-pairwise conditionals perform

better than pairwise near the independence model.

However, these results do not directly address a question that we view as becoming ever more important in composite likelihood: when and how should one move to a higher order strategy (with $D_{cl}$ larger) while keeping $N_{cl}$ small. Our information decomposition in Theorem 2 does provide some insights. The additive decomposition of information at $\theta_{ind}$ means that we can precisely determine the contribution to the information of higher order marginals relative to lower order ones. Thus, for example, one could take the full Hoeffding decomposition, list the information content of each score for a parameter of interest, and then select the ones that give the greatest return in information per dollar cost in computing. Of course, such an analysis would be naive given the very local nature of Hoeffding optimality.

However, there are examples that show that this thinking can apply more widely. Suppose the model $f(\mathbf{y};\theta)$ is a first order Markov chain for the sequence $Y_1,\ldots,Y_d$. In such a case, it is clear that the true score $u_{mle}$ has $D_{cl} = 2$, regardless of the value of $\theta$, and that it only uses adjacent pairs $(Y_i, Y_{i+1})$. It also follows that the Hoeffding additive score $h_2(\theta)$ must be $u_{mle}$ at $\theta_{ind}$. This can be checked by direct calculation: first show that at $\theta_{ind}$, $u_{ij}^{hfd} = 0$ for any $(i,j)$ pair that are not adjacent, then plug these scores into $h_2(\theta)$ to give the result. That is, the number of scores we need to consider is only of order $N_{cl} = d$. More generally, the optimal additive score at any $\theta$, with $D_{cl} = 2$, must also be $u_{mle}$. That is, many of the terms in the best additive decomposition are identically zero; namely all functions of $(Y_i, Y_j)$ where $i$ and $j$ are not neighbors. Similar remarks apply to higher order Markov chain models, with $N_{cl}$ staying at $d$, but $D_{cl}$ increasing with the order of the Markov chain. That is, assumptions of Markovian structure can greatly reduce computational burden by keeping $N_{cl}$ of order $d$.

Markov chains are thus examples where many of the additive score terms are exactly zero. (See Hjort and Varin (2008), for a more detailed comparison of the efficiency of various composite likelihood choices in Markov chain models.) There are many other models of interest where one might not have exact zeroes (except possibly for certain parameter values), but one would still expect some ordering in the information content of the various scores due to some spatial or temporal structure, and therefore expect there to be significant value to using higher order marginals with fewer composite likelihood factors $N_{cl}$. In the following, we describe some strategies for doing this. These strategies are being implemented and evaluated by the authors in a separate ongoing project.

### 6.1. Surrogate densities and likelihoods

Motivated by our Markov chain example, we now consider a special class of composite likelihoods $C(\theta)$ that have the property that they are the actual

likelihoods for true density functions. We say that the density function $s(y;\theta)$ is a surrogate density for $f(y;\theta)$ if $s(y;\theta) = C(\theta)$ for some composite likelihood $C$. We then call $C(\theta)$ a *surrogate composite likelihood*. In this section we show how one can construct a sequence of surrogates $s_m(\mathbf{y};\theta)$ such that $D_{cl}$ increases in $m$, but $N_{cl}$ is fixed at $d$, while statistical efficiency increases as well. Thus one could select the parameter $m$ to optimize efficiency for any given set of computational limitations.

Surrogate composite likelihoods have some conceptual advantages. Most commonly used constructions of composite likelihoods, such as pairwise, are not true likelihoods for any density, and so the parallels to likelihood theory are limited to first order properties. When calculations are done treating the surrogate model as true instead of the nominal model, surrogate likelihoods obey all the standard properties of likelihood, such as full efficiency for parameter estimation, estimation of information from the information matrix, sensible Bayesian analysis, and so forth. Thus to the extent that the surrogate density $s(\mathbf{y};\theta)$ is a reasonable approximation to the full model $f(\mathbf{y};\theta)$, these features should be approximately true for the surrogate composite likelihood.

### 6.2. Surrogate Markov random field

The universal representation of a density $s$ in the form

$$s(\mathbf{y}) = s(y_1)s(y_2|y_1)s(y_3|y_1,y_2) \times \cdots \times s(y_d|y_1,\ldots,y_{d-1})$$

shows that one can define a density function $s$ that will be a surrogate for $f(\mathbf{y};\theta)$ by defining each of the conditional densities $s(y_k|y_{<k>})$ for $k = 2,\ldots,d$, where $y_{<k>}$ stands for the set $\{y_1, y_2, \ldots, y_{k-1}\}$. We here define these conditional densities using the true model $f$ by setting

$$s(y_k|y_{<k>}) = f\{y_k|y^*(k)\},$$

where $y^*(k)$ is a chosen subset of $y_{<k>}$. It follows from the product representation of the density that the surrogate composite likelihood is

$$C_s(\theta) = \prod s(y_k|y_{<k>};\theta) = \prod f\{y_k|y^*(k);\theta\}.$$

If the size of the subset $y^*(k)$ is fixed to be $m$ or less, then the calculations for this surrogate density/likelihood are from marginals of order no greater than $D_{cl} = m+1$. Ideally, the density $f\{y_k|y^*(k)\}$ equals $f(y_k|y_{<k>})$, as then we have reproduced $f$ exactly. Thus the closeness of the surrogate $s$ to $f$ depends on how strongly the conditional density of $Y_k$ given $y_{<k>}$ depends on the "neighbors" of $y_k$ that are specified by $y^*(k)$. The worst case situation for this type of composite

likelihood occurs if the data is exchangeable in the model, then all points are equal neighbors of each other, and so one is unlikely to do a good job of approximating $f(y_k|y_{<k>})$ with fewer conditioning variables.

If we were to choose $y^*(k)$ to be $y_{k-1}, \ldots, y_{k-m}$, then the surrogate model would be an $m$th order Markov chain model for the $\mathbf{Y}$ sequence. In this setting one can show that each increment of $m$ leads to a surrogate density that is closer, in Kullback-Leibler discrepancy, to the model $f$ (A similar remark holds for any strategy that increments $y^*(k)$ in a nested fashion). In a surrogate Markov chain, it can also be shown that the surrogate marginal densities $s(y_k, \ldots, y_{k+m})$ for $m+1$ adjacent variables will also equal $f(y_k, \ldots, y_{k+m})$, giving another piece of evidence about how $s$ becomes an increasingly accurate approximation for $f$ as $m$ increases. The details are presented in Appendix D.

More generally, and more appropriately for data that has spatial structure, one can create a *surrogate Markov random field* using lower order marginal densities. For each data point $y_k$, we first specify a set of *neighbors* of $y_k$ , a subset of $\{y_1, \ldots, y_d\} \backslash \{y_k\}$ that we denote by $N(y_k)$. We assume that if $y_i$ is a neighbor for $y_j$, then $y_j$ is a neighbor of $y_i$. For each $k$, let the conditioning event $y^*(k)$ be the intersection of $\{y_l, \ldots, y_{k-1}\}$ and $N(y_k)$. Further, let the induced neighbors of $y_k$ be all those $y_j$ that appear with $y_k$ in any of the terms $f\{y_m|y^*(m)\}, m = 1, \ldots, d$. This set includes all neighbors of $y_k$, plus possibly some neighbors of $y_k'$s neighbors. When this is done, the surrogate density $s$ has the following property: the conditional density of $Y_k$ given the remaining data is the same as the conditional density of $Y_k$ given its induced neighbors. This is called the "local Markov property" and a density with this characteristic is called a *Markov random field* (Rue and Held (2005)).

If we are creating a surrogate Markov random field, then we need to consider which data point is to be labelled $y_1$, which $y_2$, and so forth. Clearly our sequential factorization depends on this choice. Given an initial labelling of the data, $(y_1, \ldots, y_d)$, let $(y_1', \ldots, y_d')$ be a reordered version. A useful selection of this ordering can reasonably be based on computational considerations, as we desire to keep the data dimension $D_{cl}$ small. To achieve this, the conditioning sets, $y^*(k) = y_{<k>}$ intersected with $N(y_k)$, should be kept as small as possible.

We propose a sequential selection, starting with the last variable $y_d$. The choice for the last variable is obvious because its conditioning set does not depend on ordering. Minimizing the size of $y^*(d) = N(y_d)$ means that one should choose the variable $y_d'$ from among those $y_k$ with the smallest number of neighbors. Next, note that once $y_d'$ is selected, one can proceed to choosing $y_{d-1}'$ in the same fashion, but now the set of conditioning neighbors must be in the set $\{y_1, \ldots, y_d\} \backslash \{y_d'\}$. That is, this new optimization problem is the same as the first, only $y_d'$ has been eliminated from the data set. Continuing in this manner gives a recursive

algorithm that starts with selecting the last variable $y'_d$ and, at each following step, we choose $y'_{d-k}$, for $k = 1, \ldots, d-1$, based on it having the fewest neighbors among the remaining variables.

As an example, suppose the data are realizations of a spatial process where the observations are taken at points in the rectangular lattice $\{(i, j) : i = 1, \ldots, I,$ and $j = 1, \ldots, J\}$. Suppose that the neighbors of $y_k$ are defined to be those $y$-values whose lattice locations $a', b'$ are within a fixed Euclidean distance of $(a_k, b_k)$, say $\leq \sqrt{2}$. Then in the rectangular lattice, observations at the four corner points each have 3 neighbors, the edges of the rectangle each have 5 neighbors, and points in the interior of the lattice each have 8 neighbors. If we follow the above algorithm, then the first four points to be selected, namely $y'_d, \ldots, y'_{d-3}$, are the four corners, each having three neighbors. After they are removed, the edge locations next to the corners have one less neighbor, and so they would be selected next because they have four remaining neighbors. The algorithm would continue in this way, "peeling off" the corners and edges, while never taking an interior point. As a result, the largest conditioning set has size 4, which can be shown to be the smallest maximal size that could be obtained by any ordering of the factorization. In this example, the number of points in an induced neighborhood is never more than one element larger than the original neighborhoods.

**Remark 3.** One could also build a sequence of nested surrogate models, and use likelihood ratio testing to compare their quality of fit.

### 6.3. Hidden surrogate densities

Many applications of composite likelihood are in settings where the complexity in computation arises from integration or summation over many hidden variables. (e.g., Breslow and Clayton (1993); Diggle, Tawn, and Moyeed (1998); Fieuws, Verbeke, and Molenberghs (2007)). Suppose we are doing inference for a model that has the structure

$$f(\mathbf{y}; \theta) = \int f(\mathbf{y}, \mathbf{h}; \theta) d\mathbf{h},$$

where the hidden random variables $\mathbf{H}$ are high-dimensional, and the densities $f(\mathbf{y}, \mathbf{h}; \theta)$ are each individually simple to compute. We here have represented the hidden variables as continuous, but we mean to allow discrete variables $\mathbf{H}$ as well. We assume it is the high-dimensional integration or summation that creates the computational problem. Rather than creating a surrogate density for $f(\mathbf{y}; \theta)$

directly, we could instead create surrogate densities for $f(\mathbf{y}, \mathbf{h}; \theta)$, say $g(\mathbf{y}, \mathbf{h}; \theta)$, and then for an overall surrogate use

$$g(\mathbf{y}; \theta) = \int g(\mathbf{y}, \mathbf{h}; \theta) d\mathbf{h},$$

where the $g'$s are densities that not only easier to compute, but also chosen so that the "complete data" likelihood $g(\mathbf{y}, \mathbf{h}; \theta)$ is also equal to a complete data composite likelihood for the density $f(\mathbf{y}, \mathbf{h}; \theta)$. We call $g(\mathbf{y}; \theta)$ a *hidden surrogate* for $f(\mathbf{y}; \theta)$. It is now critical, if this construction is to be computationally efficient, that the integration over $\mathbf{h}$ to create $g(\mathbf{y}; \theta)$ becomes more efficient to implement than the integration for $f(\mathbf{y}; \theta)$.

One possible way to do this is to let $g(\mathbf{y}, \mathbf{h}; \theta)$ be an order-$m$ Markov chain surrogate for $f(\mathbf{y}, \mathbf{h}; \theta)$. As before, this guarantees a growing Kullback-Leibler similarity between $g$ and $f$. The resulting observed data surrogate $g(\mathbf{y}; \theta)$ is then a hidden Markov chain, and benefits from the computational simplifications for its calculation that are found in the forward and backward algorithms. This can reduce an integration whose operations grow exponentially in $d$ to calculations that are quadratic in $d$ (Ewens and Grant (2001)). More generally, if we use a surrogate Markov random field for $g(\mathbf{y}, \mathbf{h}; \theta)$, we obtain a similar savings in computation.

We foresee considerable advantages to the hidden surrogate approach due to one's ability to use the surrogate densities for inference about the hidden variables. It also provides a natural method to create proposal distributions for a Monte Carlo likelihood analysis (Geyer and Thompson (1992); Gilks, Richardson, and Spoegelhalter (1996); Robert and Casella (1997)). However, there is one note of caution: even though the surrogate exactly matches the chosen complete data marginal densities, it does not imply that the score functions from $g(\mathbf{y}; \theta)$ satisfy $E_\theta\{\nabla \log g(\mathbf{Y}; \theta)\} = 0$. That is, if one uses a hidden surrogate for a composite likelihood, one needs to be cautious about the bias that is introduced into the parameter estimates.

**Remark.** It has long been commonplace in statistical practice to use a statistical model that has simple computations and analysis even though it is clearly false. Prominent examples include the wide variety of models based on linearity assumptions and on normality assumptions. These methods are usually justified on the basis that the analysis is likely to be approximately valid if the model assumptions are approximately true. Another way to put this is that we can think of the normal/linear model as a useful surrogate for a model that would more accurately capture our state of knowledge. One of the virtues of composite likelihood methods is that they have that same simplifying nature. Indeed, we have shown above that there are useful ways to create computationally easier surrogate models within the composite likelihood framework.

## 7. Discussion

In the first part of this paper, we restricted our attention to the determining the optimal estimating functions based on using only marginal densities with data dimension less than some fixed value. In the process, $N_{cl}$ was left unconstrained. We did this in two ways, by examining the optimal weights for the composite likelihood scores and by considering the optimal additive estimating functions. In the process, our optimality result showed that the pairwise scores could suffer in efficiency loss from an implicit overuse of one-wise information.

However, simply using the composite likelihood generated by the Hoeffding scores turned out to be too "local" to the independence models. In this paper, we provided motivations for the use of "all pairwise conditionals" and modified Hoeffding scores. Modified Hoeffding scores could serve as a start toward a method with easy computation as well as balanced optimality and positive likelihood association. The pairwise conditional likelihood seems to be competitive, has the advantage of being a true composite likelihood, and serves as a reasonable compromise between efficiency and cheapness. Our numerical assessments here were quite limited. It would be interesting to compare the performance of various composite likelihoods on more complex models.

In Section 6, we developed several new composite likelihood constructions based on using Markov random field surrogate densities. These models are a natural way to create density approximations using an increasing hierarchy of conditional independence assumptions. They are fully efficient when the true model has the needed conditional independence relationships, and otherwise provides a computationally feasible way to construct a composite likelihood for models where the data has some natural neighborhood structure to its dependencies.

We consider these results a bare beginning. Much more research is warranted to investigate the use of the composite likelihood regarding these questions.

## Acknowledgements

Lindsay's and Sun's research was supported by the National Science Foundation through NSF-DMS 0714839. Yi's research was supported by the Natural Sciences and Engineering Research Council of Canada. The authors thank Nancy Reid for helpful comments on the first version.

## Appendix A: Proof of Theorem 1

We need to show that at $\theta = \theta_{ind}$, $u_{mle} - h_k(\theta)$ is orthogonal to all the basis elements of $\mathcal{L}_k$. First consider the case where $k = 2$ and we wish to show

orthogonality to any function of $g(y_1, y_2; \theta) \in \mathcal{G}_{12}$. We have

$$\begin{aligned} &E[\{u_{mle} - h_2(\theta)\}g^{\mathrm{T}}(Y_1, Y_2; \theta)|(Y_1, Y_2)] \\ &= E(E[\{u_{mle} - h_2(\theta)\}g^{\mathrm{T}}(Y_1, Y_2; \theta)|(Y_1, Y_2)]) \\ &= E[u_{12} - E\{h_2(\theta)|(Y_1, Y_2)\}g^{\mathrm{T}}(Y_1, Y_2; \theta)]. \end{aligned}$$

Thus we are done if we show $E\{h_2(\theta)|(Y_1, Y_2)\} = u_{12}$. We consider the conditional expectation of all the terms in $h_2(\theta)$. First, $E\{u_1|(Y_1, Y_2)\} = u_1$ and $E\{u_2|(Y_1, Y_2)\} = u_2$, but for every other index $j$ we have $E\{u_j|(Y_1, Y_2)\} = 0$ due to the independence of $Y_j$ and $(Y_1, Y_2)$ at $\theta_{ind}$. Next, clearly $E\{u_{12}^{hfd}|(Y_1, Y_2)\} = u_{12}^{hfd}$. For $j > 2$, we have

$$E\{u_{1j}^{hfd}|(Y_1, Y_2)\} = E\{u_{1j} - u_1 - u_j|(Y_1, Y_2)\} = E\{u_{j|1} - u_j|(Y_1, Y_2)\} = 0.$$

We have here used the fact that $E(u_{j|1}|Y_1)$ is necessarily zero by the mean zero property of a score, but the independence of the variables means that the conditional distributions of one subset given any other subset are just the marginals.

More generally, we show that at $\theta = \theta_{ind}$, for $3 \leq k \leq d$, $u_{mle} - h_k(\theta)$ is orthogonal to any function of $r$-wise subset $S_r = (y_{i1}, \ldots, y_{ir})$, $g(S_r; \theta) \in \mathcal{G}_{S_r}$, $r = 1, \ldots, k$. Indeed, as

$$\begin{aligned} E[\{u_{mle} - h_k(\theta)\}g^{\mathrm{T}}(S_r; \theta)] &= E(E[\{u_{mle} - h_k(\theta)\}g^{\mathrm{T}}(S_r; \theta)|S_r]) \\ &= E([u(\theta; S_r) - E\{h_k(\theta)|S_r\}]g^{\mathrm{T}}(S_r; \theta)), \end{aligned}$$

we are done if we show that $E\{h_k(\theta_{ind})|S_r\} = u(\theta_{ind}; S_r)$. This can be proved by applying the following identities to $h_k(\theta)$ iteratively, following the same arguments as above: for any subset $S \subset \{y_1, \ldots, y_d\}$,

$$E\{u(\theta; S)|S_r\} = \begin{cases} u(\theta; S), & \text{if } S \subset S_r \\ u(\theta; S \cap S_r), & \text{if } S \cap S_r \neq \oslash,\ \theta = \theta_{ind} \\ 0, & \text{if } S \cap S_r = \oslash,\ \theta = \theta_{ind}. \end{cases}$$

## Appendix B: Proof of Proposition 2

The class of functions defined by $Rg + a$, for arbitrary $p \times p$ matrix $R$, is linear, so the optimal estimating function in this class comes from minimizing the least squares criterion $E\{(u_{mle} - Rg - a)(u_{mle} - Rg - a)^{\mathrm{T}}\}$. We start by replacing $g$ with $g^* = g(\mathbf{y}; \theta) - E\{g(\mathbf{Y}; \theta)|S_k\}$. We can now minimize the equivalent criterion

$$E([U - Rg^* - RE\{g(\mathbf{Y}; \theta)|S_k\} - a(S_k; \theta)] \cdot [u_{mle} - Rg^* - RE\{g(Y; \theta)|S_k\} - a(S_k; \theta)]^{\mathrm{T}})$$

over $R$ and $a$. If we define the new function $a^*(S_k;\theta) = a(S_k;\theta) - RE\{g(Y;\theta)|S_k\}$, then the objective is to minimize

$$E\{(U - Rg^* - a^*)(U - Rg^* - a^*)^{\mathrm{T}}\}$$

over matrices $R$ and functions $a^*$ of $S_k$. We can do this by showing that for $R_{opt}$ and $a_{opt}$ as given in Proposition 2, and $a^*_{opt} = a_{opt} + R_{opt}E\{g(y;\theta)|S_k\} = u(\theta;S_k)$, we have the residuals $u_{mle} - R_{opt}g - a_{opt}$, or $u_{mle} - R_{opt}g^* - a^*_{opt}$ orthogonal to all linear functions of $g^*$ and all functions of $S_k$. Now as a function of $S_k$, $a^*_{opt}$ is orthogonal to $g^*$ (which is conditionally mean zero given $S_k$) and so any linear functions of it. Since $R_{opt}$ makes $R_{opt}g^*$ information-unbiased, we have

$$E\{(u_{mle} - R_{opt}g^* - a^*_{opt})g^{*\mathrm{T}}\} = 0.$$

Next, we consider arbitrary basis functions $h(S_k)$. Note that $u_{mle} - a^*_{opt} = u_{mle} - u(\theta;S_k)$ is conditionally mean zero given $S_k$, as is $R_{opt}g^*$. Hence we have

$$E\{(u_{mle} - R_{opt}g^* - a^*_{opt})h^{\mathrm{T}}(S_k)\} = 0,$$

as required. The orthogonality properties of $g^*$ and $a^*_{opt}$ give the information decomposition.

## Appendix C: Expression of All-Pairwise Conditionals

The product of *all* pairwise conditionals $\prod_{i\neq j} L_{i|j}$ yields the sum of pairwise conditional scores

$$\nabla \log(\prod_{i\neq j} L_{i|j}) = \sum_{i\neq j} u_{i|j} = \sum_{i\neq j} u_{ij} - \sum_{i\neq j} u_j.$$

Noting that

$$\begin{aligned}\sum_{i\neq j} u_j &= \sum_{i=1}^{d} \sum_{1\leq j\leq d, j\neq i} u_j \\ &= (u_2 + u_3 + u_4 + \cdots + u_d) + (u_1 + u_3 + u_4 + \cdots + u_d) \\ &\quad + (u_1 + u_2 + u_4 + \cdots + u_d) + \cdots + (u_1 + u_2 + u_3 + \cdots + u_{d-1}) \\ &= (d-1)\sum_i u_i,\end{aligned}$$

we obtain that

$$\sum_{i\neq j} u_{i|j} = \sum_{i\neq j} u_{ij} - (d-1)\sum_i u_i = 2\left\{\sum_{i<j} u_{ij} - \left(\frac{d-1}{2}\right)\sum_i u_i\right\}.$$

## Appendix D: On Surrogate Markov Chain Models

**Proposition 4.** *The Markov chain surrogate density of order $m$ satisfies*

$$s(y_k, \ldots, y_{k+m}) = f(y_k, \ldots, y_{k+m}).$$

**Proof.** Consider a simple example, where $m = 1$ and $d = 4$. Then we can write

$$\begin{aligned} s(y_1, y_2, y_3, y_4) &= f(y_4|y_3)f(y_3|y_2)f(y_2|y_1)f(y_1) \\ &= \frac{f(y_1, y_2)f(y_2, y_3)f(y_3, y_4)}{f(y_2)f(y_3)}. \end{aligned}$$

To find the marginal density for two variables, we integrate out the other two. When written appropriately, these integrations are clear. For example, if we wish to integrate out $y_1$ and $y_4$ to get the marginal for $Y_2$ and $Y_3$, we rewrite $s$ in the form

$$s(y_1, y_2, y_3, y_4) = f(y_1|y_2)f(y_2, y_3)f(y_4|y_3),$$

where the integrals over $y_1$ and $y_4$ in the first and last terms are, for fixed $y_2$ and $y_3$, exactly one. This argument clearly extends to larger $m$ and $d$.

**Proposition 5.** *The Markov chain surrogate densities $s_m$ satisfy*

$$\int f(\mathbf{y}) \cdot \log\left\{\frac{s_{m+1}(\mathbf{y})}{s_m(\mathbf{y})}\right\} d\mathbf{y} \geq 0$$

*so that the Kullback-Leibler divergence $\int f(\mathbf{y}) \log\{f(\mathbf{y})/s_m(\mathbf{y})\} d\mathbf{y}$ is monotonically decreasing in $m$.*

**Proof.** We consider a simple example first. Suppose $d = 4$ and we compare $m = 0$ and $m = 1$. Then we wish to find

$$\begin{aligned} &\int f(\mathbf{y}) \cdot \log\left\{\frac{s_{m+1}(\mathbf{y})}{s_m(\mathbf{y})}\right\} d\mathbf{y} \\ &= E\left[\log\left\{\frac{s_{m+1}(\mathbf{Y})}{s_m(\mathbf{Y})}\right\}\right] \\ &= E\left[\log\left\{\frac{f(Y_1, Y_2)f(Y_2, Y_3)f(Y_3, Y_4)}{f(Y_1)f^2(Y_2)f^2(Y_3)f(Y_4)}\right\}\right] \\ &= E\left[\log\left\{\frac{f(Y_1, Y_2)}{f(Y_1)f(Y_2)}\right\}\right] + E\left[\log\left\{\frac{f(Y_2, Y_3)}{f(Y_2)f(Y_3)}\right\}\right] + E\left[\log\left\{\frac{f(Y_3, Y_4)}{f(Y_3)f(Y_4)}\right\}\right]. \end{aligned}$$

Now each summand is nonnegative by the information inequality, as applied to each two dimensional marginal density. This proof generalizes to arbitrary $m$ and $d$.

## References

Bahadur, R. R. (1961). A representation of the joint distribution of response to $n$ dichotomous items. In *Studies in Item Analysis and Prediction*, (Edited by H. Solomon), 158-168. Stanford Mathematical Studies in the Social Sciences VI. Stanford University Press.

Besag, J. (1974). Spatial interaction and the statistical analysis or lattice systems (with discussion). *J. Roy. Statist. Soc. Ser. B* **36**, 192-236.

Breslow, N. E. and Clayton, D. G. (1993). Approximate inference in generalized linear mixed models. *J. Amer. Statist. Assoc.* **88**, 9-24.

Christensen, O. F. and Waagepetersen, R. (2002). Bayesian prediction of spatial count data using generalized linear mixed models. *Biometrics* **58**, 280-286.

Cox, D. R. (1972). The analysis of multivariate binary data. *Appl. Statist.* **21**, 113-120.

Cox, D. R. (1975). Partial likelihood. *Biometrika* **62**, 269-276.

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

Devlin, B., Risch, N. and Roeder, K. (1996). Disequilibrium mapping: composite likelihood for pairwise disequilibrium. *Genomics* **36**, 1-16.

Diggle, P. J., Tawn, J. A. and Moyeed, R. A. (1998). Model-based geostatistics. *J. Roy. Statist. Soc. Ser. C* **47**, 299-350.

Dillon, J. V. and Lebanon, G. (2009). Statistical and computational tradeoffs in stochastic composite likelihood. *Proceedings of the* 12*th International Conference on Artificial Intelligence and Statistics* (*AISTATS*) **5**, 129-136.

Durbin, J. (1960). Estimation of parameters in time-series regression models. *Biometrika* **47**, 139-153.

Engler, D. A., Mohapatra, M., Louis, D. N. and Betensky, R. A. (2006). A pseudolikelihood approach for simultaneous analysis of array comparative genomic hybridizations. *Biostatistics* **7**, 399-321.

Ewens, W. J. and Grant, G. R. (2001). *Statistical Methods in Bioinformatics*. Springer, New York.

Fearnhead, P. and Donnelly, P. (2002). Approximate likelihood methods for estimating local recombination rates (with discussion). *J. Roy. Statist. Soc. Ser. B* **64**, 657-680.

Fieuws, S. and Verbeke, G. (2006). Pairwise fitting of mixed models for the joint modeling of multivariate longitudinal profiles. *Biometrics* **62**, 424-431.

Fieuws, S., Verbeke, G. and Molenberghs, G. (2007). Random-effects models for multivariate repeated measures. *Statist. Meth. Medical Res.* **16**, 387-397.

Fisher, R. A. (1922). On the mathematical foundations of theoretical statistics. *Philos. Trans. R. Soc. Ser. A* **222**, 309-368.

Geyer, C. J. and Thompson, E. A. (1992). Constrained Monte Carlo maximum likelihood for dependent data (with discussion). *J. Roy. Statist. Soc. Ser. B* **54**, 657-699.

Gilks, W. R., Richardson, D. and Spoegelhalter, D. J. (1996). *Markov Chain Monte Carlo in Practice*. Chapman and Hall/CRC, New York.

Godambe, V. P. (1960). An optimum property of regular maximum likelihood estimation. *Ann. Statist.* **31**, 1208-1212.

Hastie, T. J. and Tibshirani, R. J. (1990). *Generalized Additive Models*. Chapman and Hall/CRC.

Heagerty, P. J. and Lele, S. R. (1998). A composite likelihood approach to binary spatial data. *J. Amer. Statist. Assoc.* **93**, 1099-1111.

Heyde, C. C. and Morton, R. (1998). Multiple roots in general estimating equations. *Biometrika* **85**, 954-959.

Hjort, N. and Omre, H. (1994). Topics in spatial statistics (with discussion, comments and rejoinder). *Scand. J. Statist.* **21**, 289-357.

Hjort, N. L. and and Varin, C. (2008). ML, PL, QL in Markov chain models. *Scand. J. Statist.* **35**, 64-82.

Hoeffding, W. (1948). A class of statistics with asymptotically normal distribution. *Ann. Math. Statist.* **19**, 293-325.

Joe, H. (1997). *Multivariate Models and Dependence Concepts*. Chapman and Hall/CRC.

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

Kuk, A. Y. C. and Nott, D. J. (2000). A pairwise likelihood approach to analyzing correlated binary data. *Statist. Probab. Lett.* **47**, 329-335.

Larribe, F. and Fearnhead, P. (2011). On composite likelihoods in statistical genetics. *Statist. Sinica* **21**, 43-69.

Lee, A. J. (1990). *U-Statistics: Theory and Practice*. CRC Press, Taylor & Francis Group, LLC.

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

McCullagh, P. (1983). Quasi-likelihood functions. *Ann. Statist.* **11**, 59-67.

Nott, D. and Rydén, T. (1999). Pairwise likelihood methods for inference in image models. *Biometrika* **86**, 661-676.

Robert, C. P. and Casella, G. (1997). *Monte Carlo Statistical Methods*. Wiley, New York.

Rue, H. and Held, L. (2005). *Gaussian Markov Random Fields: Theory and Applications*. Chapman & Hall/CRC Press.

Self, S. G. and Liang, K. Y. (1987). Asymptotic properties of maximum likelihood estimators and likelihood ratio tests under nonstand conditions. *J. Amer. Statist. Assoc.* **82**, 605-610.

Varin, C. (2008). On composite marginal likelihoods. *Adv. Statist. Anal.* **92**, 1-28.

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

Wedderburn, R. W. M. (1974). Quasi-likelihood, generalized linear models, and the Gauss-Newton method. *J. Roy. Statist. Soc. Ser. B* **61**, 439-447.

Zhao, Y. and Joe, H. (2005). Composite likelihood estimation in multivariate data analysis. *Canad. J. Statist.* **33**, 335-356.

Department of Statistics, Penn State University, University Park, PA 16802, USA.

E-mail: bgl@psu.edu

Department of Statistics and Actuarial Science, University of Waterloo, Waterloo, Ontario, Canada N2L 3G1.

E-mail: yyi@uwaterloo.ca

Department of Statistics, Penn State University, University Park, PA 16802, USA.

E-mail: jxs1021@psu.edu

(Received October 2009; accepted September 2010)