

Composite likelihood: Issues in Efficiency

Bruce G. Lindsay, Jianping Sun, Department of Statistics, Penn State University

Motivation

The motivation is to estimate the unknown ancestor DNA sequences from the observed descendent DNA sequences with the consideration of some biological realistic, such as *mutation* and *recombination*, so that to construct an evolution tree.

Model

To estimate K different unknown ancestors, μ_1, \dots, μ_K , and their corresponding unknown weights π_1, \dots, π_K , let $A_i = l$ denote the potential ancestor chosen at site i is μ_l , with prob. $P(A_i = l) = \pi_l$, where $i = 1, \dots, L$, L is the sequence length. Define the prob. of recombination between site j and site $j + 1$ is $P(R_j = 1) = q$, for $j = 1, \dots, L - 1$. Then, we have

$$P(Y = (y_1, \dots, y_L)) = \sum \sum [P(Y = (y_1, \dots, y_L) | A_1 = i_1, \dots, R_1 = r_1, \dots) \times P(A_1 = i_1, \dots, A_L = i_L) P(R_1 = r_1, \dots, R_{L-1} = r_{L-1})]$$

where $P(R_1 = r_1, \dots, R_{L-1} = r_{L-1})$ is the prob. of independent Bernoulli trials with recombination rate q , and $P(A_1 = i_1, \dots, A_L = i_L)$ is the prob. of independent drawing corresponding ancestor sequence. In addition, we have

$$P(Y = (y_1, \dots, y_L) | A_1 = i_1, \dots, R_1 = r_1, \dots) = \kappa_{\{y_1, \dots, y_L\} | \{\mu_{B_1(1)}, \dots, \mu_{B_L(L)}\}}$$

where $\kappa_{\{y|\mu\}}$ is the *mutation kernel* defined in Chen and Lindsay (2006), denoting the prob. of observing descendent y when ancestor is μ ; the value of (B_1, \dots, B_L) depends on the recombination indicator R and the potential ancestor at each site.

- $B_1 = i_1$;
- $B_2 = i_1$ if $R_1 = 0$; $B_2 = i_2$ if $R_1 = 1$;
- $B_3 = B_2$ if $R_2 = 0$; $B_3 = i_3$ if $R_2 = 1$; ...

Therefore, the log-likelihood function is

$$l = \sum n_{y_1, \dots, y_L} \log P(Y = (y_1, \dots, y_L))$$

Computation Complexity

Huge computation challenges exist for the above likelihood, because of enormous recombination possibilities when L is large. The EM algorithm need thousands of steps to converge even for simple case $L = 3$. Hence, we need some other methods to reduce the complexity of computation.

Composite Likelihood (Lindsay 1988) is one possible solution.

Composite Likelihood Method

Let $\mathbf{Y} = (Y_1, Y_2, \dots, Y_d)^T \sim f(\mathbf{y}; \theta)$, $\theta \in \mathbb{R}^p$. Note that Y_1, Y_2, \dots, Y_d may not be independent. Let $L(\theta; \mathbf{y}) = f(\mathbf{y}; \theta)$, and $U_{mle}(\theta; \mathbf{y}) = \nabla \log f(\mathbf{y}; \theta)$ denote the corresponding likelihood and score functions respectively.

Now define the following subset likelihood and scores:

- Onewise marginal: L_i and U_i for all i
- Pairwise marginal: L_{ij} and U_{ij} for $i < j$
- All-pairwise-conditional: $L_{i|j}$ and $U_{i|j}$, $i \neq j$
- variables indexed by s : L_s and U_s

The *Composite Likelihood* (CL) and *Composite Score* (CS) are defined as

$$CL = \prod_s L_s \text{ and } CS = \sum_s U_s$$

More general, the *Weighted Composite Likelihood* can be defined as $\prod_s L_s^{w_s}$, where $w_s \geq 0$ is called *weight*.

Construction of Composite Scores

Typically, there is *information loss* for the CL when compared with the true likelihood (TL). The problem of information loss could also cause *efficiency loss* for CL method. Hence, one interesting problem is how to construct a proper CL so that it maintains efficiency as much as possible and keeps the computations economical meanwhile.

Hoeffding Score

Suppose TL is difficulty to specify, but specifications of L_i , L_{ij} , and $L_{i|j}$ are possible. Define *independent parameter value*, θ_{ind} , to be any value of θ for which all Y_i 's are independent. Define two classes of *additive estimating functions*:

- $\mathcal{H}_1 = \{\mu + \sum_{i=1}^n g_i(Y_i)\}$,
- $\mathcal{H}_2 = \{\mu + \sum_{i=1}^n g_i(Y_i) + \sum_{i < j} g_{ij}(Y_i, Y_j)\}$,

where $\mu \in \mathbb{R}$, g_i and g_{ij} , are arbitrary finite variance functions with mean zero under θ . Under θ_{ind} , the projections of U_{mle} on \mathcal{H}_k , $k = 1, 2$, are

- $h_1(\theta, Y) = \sum_{i=1}^n U_i$
- $h_2(\theta, Y) = h_1 + \sum_{i < j} (U_{ij} - U_i - U_j)$

We call $h_1(\theta, Y)$ and $h_2(\theta, Y)$ as the first and second order *Hoeffding Scores* (HS). They are *optimal* in $\mathcal{H}_1, \mathcal{H}_2$ when θ_{ind} holds.

Hoeffding Score (Continue)

In addition, we call $U_{ij}^* = U_{ij} - U_i - U_j$ as *Corrected Pairwise Score*, which remove overused marginal information from the pairwise score. Hence, *Hoeffding likelihood* (HL) generated from Hoeffding scores is defined as

- $L_{h_1} = \prod_i f_i(Y_i)$
- $L_{h_2} = L_{h_1} \prod_{i < j} \frac{f_{ij}(Y_i, Y_j)}{f_i(Y_i)f_j(Y_j)} = \frac{\prod_{i < j} f_{i,j}(Y_i, Y_j)}{[\prod_i f_i(Y_i)]^{d-2}}$

Note there is negative weight in L_{h_2} , so it's not a true CL. Then one big question is: *Can we use $\log L_{h_k}$ as a sensible inference function?* The answer is **NOT necessarily**, because

- HL does **NOT satisfy local max-consistency**. Inference function G satisfies *local max-consistency* if $-\nabla^2 E_{\theta_\tau} [G(\theta_\tau; \mathbf{Y})] \geq 0$. Local max-consistency is a **key part of a consistency proof**.
- HS does **NOT satisfy Information identity**. Estimating function g satisfies *Information identity* if $E\{-\nabla g\} = E\{gg^T\} \geq 0$. Information identity is a **key tool for local max-consistency**.

Even when we try some weaken concept of information identity, say *Positive Likelihood Association*, we still have some problems.

- Positive Likelihood Association: An estimating function g will be said to have *positive likelihood association* if $E(gU_{mle}^T) = E[-\nabla g] \geq 0$.
- **Problem**: U_{ij}^* does **not necessarily** have positive likelihood association.

Modified Hoeffding Score

We want to create a computationally simple modification of Hoeffding score that has positive likelihood association.

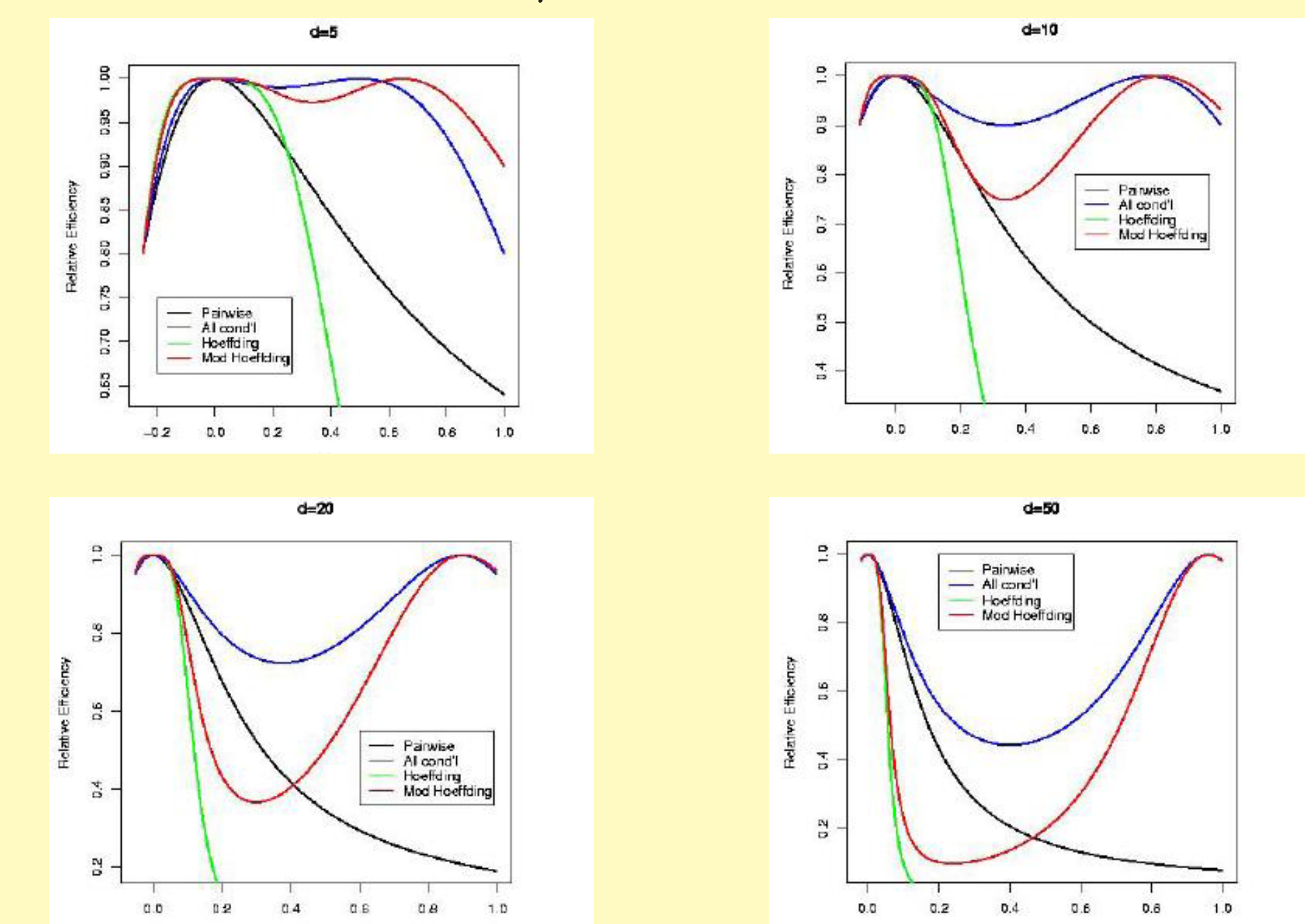
Let $U_{ij}^{**} = U_{ij} - \beta_i(\theta)U_i - \beta_j(\theta)U_j$, where β 's are $p \times p$ weight matrices which remove the marginal U_i and U_j effect. Hence, U_{ij}^{**} is *orthogonal* to U_i and U_j under *general θ* . Moreover, one fact is that U_{ij}^{**} *satisfies* information identity, hence has positive likelihood association. Thus, we define *Modified Hoeffding scores* (MHS) as

$$\sum_i U_i + \sum_{i < j} U_{ij}^{**}$$

Multivariate Normal Example

Suppose $\mathbf{Y} = (Y_1, \dots, Y_d)^T \sim N_d(0, \Sigma)$, where $\Sigma = \sigma^2[(1 - \rho)I_{d \times d} + \rho \mathbf{1}\mathbf{1}^T]$. Here, σ^2 is an unknown parameter and ρ is assumed to be a known constant.

We compare the relative efficiencies for four types of scores: Pairwise, All-Conditional-Pairwise, Hoeffding, and Modified Hoeffding, with respect to true score. The numerical results when $d = 5, 10, 20, 50$ are showing in the following plots. The X-axis is the value of ρ .



Future Work

- Applying CL to the genetic model above. Select candidate composite likelihoods/scores; Find algorithm. (eg. EM and Newton-Raphson); Design a comparison experiment to compare CL with its possible competitors.
- Comparing CL method with other computation method, for example, EM algorithm, Simplex gradient method, and Other MC likelihood method.

Reference

1. Chen, S. and Lindsay, B. G. (2006). *Building mixture trees from binary sequence data*, Biometrika, 93, 4. pp. 843-860.
2. Lindsay, B. G. (1988). *Composite likelihood method*. Contemporary Mathematics, 80, pp. 221-39.