Iterative Likelihood: A Unified Inference Tool [1]

Haiying Wang[a], Dixin Zhang[b], Hua Liang[c] and David Ruppert[d]
[a]Department of Statistics, University of Connecticut, Storrs, CT 06269-4120, USA
haiying.wang@uconn.edu

[b]Department of Finance, Nanjing University, Nanjing, Jiangsu 210093, China
dixinz01@nju.edu.cn

[c]Department of Statistics, George Washington University, Washington, D.C., USA
hliang@gwu.edu

[d] Department of Statistical Science, Cornell University, Ithaca, New York 14853, USA
dr24@cornell.edu

SUMMARY. We propose a framework for inference based on an "iterative likelihood function", which provides a unified representation for a number of iterative approaches, including the EM algorithm and the generalized estimating equations. The parameters are decoupled to facilitate construction of the inference vehicle, to simplify computation, or to ensure robustness to model misspecification and then recoupled to retain their original interpretations. For simplicity, throughout the paper we will refer to the log-likelihood as the "likelihood". We define the global, local, and stationary estimates of an iterative likelihood and, correspondingly, the global, local, and stationary attraction points of the expected iterative likelihood. Asymptotic properties of the global, local, and stationary estimates are derived under certain assumptions. An iterative likelihood is usually constructed such that the true value of the parameter is a a point of attraction of the expected log-likelihood. Often, one can only verify that the true value of the parameter is a local or stationary attraction, but not a global attraction. We show that when the true value of the parameter is a global attraction, any global estimate is consistent and asymptotically normal; when the true value is a local or stationary attraction, there exists a local or stationary estimate that is consistent and asymptotically normal, with a probability tending to 1. The behavior of the estimates under a misspecified model is also discussed. Our methodology is illustrated with three examples: 1) estimation of the treatment group difference in the level of censored HIV RNA viral load from an AIDS clinical trial; 2) analysis of the relationship between forced expiratory volume and height in girls from a longitudinal pulmonary function study; and 3) investigation of the impact of smoking on lung cancer in the presence of DNA adducts. Two additional examples are in the supplementary materials, GEEs (Generalized Estimating Equations) with missing covariates and an unweighted estimator for big data with subsampling.

KEY WORDS: Asymptotic properties, attraction, censoring, EM algorithm, generalized estimating equations, mean score method.

[1] The authors thank the Editor, an associate editor and two referees for their insightful suggestions and comments that have greatly improved an earlier version of this paper. Liang's research was partially supported by NSF grant DMS-1620898.

1. Introduction

Estimation procedures with iterations are widely used for statistical inference. Examples include the Newton method, the expectation-maximization (EM) algorithm, and the method for alternately solving generalized estimating equations (GEE) (Liang and Zeger, 1986). In most settings, iterations are considered as a numerical technique that circumvents otherwise computationally intractable tasks rather than a statistically meaningful notion. An exception to this may be the EM algorithm, into which extensive research has been conducted for a better understanding of the underlying concepts.

This article proposes a framework for inference based on an “iterative likelihood function”, which provides a unified representation for a number of iterative approaches, including the EM algorithm and the GEE. In fact, any system of estimating functions can be expressed as an iterative likelihood, albeit not uniquely. This approach is closely related to the method of decoupling described by Hand and Crowder (1996). It iteratively alternates between decoupling the parameters and recoupling them. The parameters are decoupled to facilitate construction of the inference vehicle, to simplify computation, or to ensure robustness of estimation to model misspecification and recoupled to retain their original interpretations. For simplicity, we will refer to the log-likelihood as simply the likelihood.

We define the global, local, and stationary estimates from an iterative likelihood and, correspondingly, the global, local, and stationary attraction points of the expectation of the iterative likelihood. An iterative likelihood is usually constructed such that the true value of the parameter is an attraction point of its expectation. Often, it is unrealistic to expect the true value to be a global attraction, but verification for local or stationary attraction is likely. We show that when the true value is a global attraction, any global estimate is consistent and asymptotically normal; when the true value is a local or stationary attraction, there exists a local or stationary estimate that is consistent and asymptotically normal, with a probability tending to 1. Our asymptotic variance reduces to the "sandwich" estimate when the iterative likelihood degenerates to a conventional likelihood-based function and to an estimate of Louis's correction (Louis, 1982), or another simple alternative, in computing the variance for the EM algorithm. The behavior of the estimates under a misspecified model is also discussed. A modified Newton algorithm that is straightforward to implement is proposed to find estimates from an iterative likelihood.

We use three real, but simplified, examples to illustrate the proposed methodology. The first involves estimating the treatment group difference in the level of censored HIV RNA viral load from an AIDS clinical trial, where the EM algorithm for accommodating the censoring is represented in the context of an iterative likelihood. The second is an analysis of the relationship between FEV_1 and height in girls from a longitudinal study of lung function, where an iterative likelihood related to GEE is used. The third is an investigation of the impact of smoking on lung cancer in the presence of DNA adducts, where the mean score method for missing covariates is extended.

In Section 2, we introduce the concept of the iterative likelihood, along with three examples. Section 3 gives the asymptotic properties of an iterative likelihood and discusses the robustness of estimation. In Section 4, we focus on obtaining estimates from an iterative likelihood and the asymptotic variance. Section 5 revisits the three examples with real data followed by extensive simulation studies in Section 6 to exam the performance of the proposed method on the three examples. Some concluding remarks are given in Section 7 and all technical proofs are included in the supplementary materials. The supplementary materials also contain two examples, GEEs (Generalized Estimating Equations) with missing covariates and an unweighted estimator for big data with subsampling.

2. Concept

2.1. Definition of the iterative likelihood and examples

Let Y_i, $i = 1, 2, \ldots, N$, be the response vector from the ith subject, with the distribution of the Ys being partially characterized by parameters $\boldsymbol{\theta}$ in the parameter space Θ. We consider only the case where Θ is a bounded and convex subspace of a K-dimension Euclidean space, noting that the extension beyond Euclidean space is natural, but requires further research. To rule out a parameter or estimate being on the boundary, we assume throughout that the parameter space is open. We use $\boldsymbol{\theta}$ for a general parameter, while θ or θ' indicates a typical parameter value We express any value of $\boldsymbol{\theta}$, θ, and derivatives of scalar functions of θ with respect to θ as $1 \times K$ vectors.

For simplicity, we use the term "likelihood" instead of "log likelihood function", despite the latter being more accurate. Let $l_i(\theta, \theta') = l_i(\theta, \theta'; Y_i)$ be a function of θ, θ', Y_i, and potential covariates, where θ and θ' are two possibly distinct values of $\boldsymbol{\theta}$. We refer to

$$L_N(\theta, \theta') = \sum_{i=1}^{N} l_i(\theta, \theta') = \sum_{i=1}^{N} l_i(\theta, \theta'; Y_i) \tag{1}$$

as an iterative likelihood for $\boldsymbol{\theta}$ if it is used for inference about $\boldsymbol{\theta}$. Unlike a conventional likelihood-based function such as quasi- or pseudo-likelihood, which involves only θ, an iterative likelihood involves both θ and θ'. We use the term "iterative" as θ' represents an intermediate value of θ in the estimation iterations. As detailed later, separation of θ and θ' in notation renders great generality by encompassing a variety of statistical methods with convenience for statistical conceptualization, mathematical representation, and computational programming, while retaining connection to the conventional likelihood-based framework. We consider three examples as follows. Section 5 will revisit them with detailed discussions and analyses of real data.

Example 1 (EM algorithm) Let Z_i and X_i be the complete response and its predictors for the ith subject, respectively, and $f_Z(z|x;\theta)$ be the conditional distribution of Z_i given X_i, characterized by parameters $\boldsymbol{\theta}$. The response Z_i may be subject to missingness. The potentially masked or missing response is denoted by Y_i. The predictors, X_i, are always observed. We may define an iterative likelihood for $\boldsymbol{\theta}$ as

$$L_N(\theta,\theta') = \sum_{i=1}^{N} l_i(\theta,\theta'), \text{ with } l_i(\theta,\theta') = \int \{\log f_Z(z|X_i;\theta)\}\, f_{Z|Y}(z|Y_i,X_i;\theta')dz, \qquad (2)$$

where $f_{Z|Y}(z|y,x;\theta)$ is the conditional distribution of Z_i given Y_i. Here $f_{Z|Y}(z|y,x;\theta)$ is just $f_Z(z|x;\theta)$ if the ith response is totally missing, i.e., Y_i does not contain a measurement.

The EM algorithm (Dempster et al., 1977) for estimating $\boldsymbol{\theta}$ can be represented in the form of $L_N(\theta,\theta')$. Consider only local maximization (maximization over θ with θ' fixed) at the M-step. At each iteration, the E-step substitutes the value of θ from the previous iteration for θ' in $L_N(\theta,\theta')$; the M-step maximizes $L_N(\theta,\theta')$ locally over θ while fixing θ'. At convergence, the final estimate $\widehat{\boldsymbol{\theta}}_N$ will be such that for any value θ in its neighborhood,

$$\begin{aligned} L_N(\widehat{\boldsymbol{\theta}}_N,\widehat{\boldsymbol{\theta}}_N) &= \sum_{i=1}^{N}\int \{\log f_Z(z|X_i;\widehat{\boldsymbol{\theta}}_N)\}\, f_{Z|Y}(z|Y_i,X_i;\widehat{\boldsymbol{\theta}}_N)dz \\ &\geq \sum_{i=1}^{N}\int \{\log f_Z(z|X_i;\theta)\}\, f_{Z|Y}(z|Y_i,X_i;\widehat{\boldsymbol{\theta}}_N)dz = L_N(\theta,\widehat{\boldsymbol{\theta}}_N). \end{aligned}$$

Example 2 (GEE) Let $Y_i = (Y_{i1},\ldots,Y_{iM_i})$ be the responses of the ith subject, where M_i is the number of observations from the ith subject. We model the mean, standard deviation and correlation as,

$$\mathrm{E}Y_{ij} = \mathbf{m}_{ij}(b), \qquad (3a)$$

$$\mathrm{Sd}Y_{ij} = \mathbf{d}_{ij}(b,a), \qquad (3b)$$

$$\text{Corr}(Y_{ij_1}, Y_{ij_2}) = \mathbf{r}_{ij_1j_2}(b, a, c), \tag{3c}$$

respectively, where $\mathbf{m}_{ij}()$, $\mathbf{d}_{ij}()$, and $\mathbf{r}_{ij_1j_2}()$ are known functions characterized by parameter vectors $\mathbf{b}$, $(\mathbf{b}, \mathbf{a})$ and $(\mathbf{b}, \mathbf{a}, \mathbf{c})$. We define an iterative likelihood for $\boldsymbol{\theta} = (\mathbf{b}, \mathbf{a}, \mathbf{c})$ as,

$$L_N(\theta, \theta') \overset{\text{def.}}{=} L_N^{(\text{m})}(b, \theta') + L_N^{(\text{d})}(a, \theta') + L_N^{(\text{r})}(c, \theta'), \tag{4}$$

where $\theta = (b, a, c)$, $\theta' = (b', a', c')$, and

$$L_N^{(\text{m})}(b, \theta') \overset{\text{def.}}{=} -\sum_{i=1}^{N} \sum_{j_1, j_2} \{W_{ij_1j_2}^{(\text{m})}(\theta') \cdot D_{ij_1}^{(\text{m})}(b) \cdot D_{ij_2}^{(\text{m})}(b)\}, \tag{5a}$$

$$L_N^{(\text{d})}(a, \theta') \overset{\text{def.}}{=} -\sum_{i=1}^{N} \sum_{j} [W_{ij}^{(\text{d})}(\theta') \cdot \{D_{ij}^{(\text{d})}(b', a)\}^2], \tag{5b}$$

$$L_N^{(\text{r})}(c, \theta') \overset{\text{def.}}{=} -\sum_{i=1}^{N} \sum_{j_1, j_2} [W_{ij_1j_2}^{(\text{r})}(\theta') \cdot \{D_{ij_1j_2}^{(\text{r})}(b', a', c)\}^2], \tag{5c}$$

with

$$D_{ij}^{(\text{m})}(b) \overset{\text{def.}}{=} Y_{ij} - \mathbf{m}_{ij}(b), \tag{6a}$$

$$D_{ij}^{(\text{d})}(a, b') \overset{\text{def.}}{=} \{Y_{ij} - \mathbf{m}_{ij}(b')\}^2 - \mathbf{d}_{ij}^2(b', a), \tag{6b}$$

$$D_{ij_1j_2}^{(\text{r})}(c, a', b') = \frac{Y_{ij_1} - \mathbf{m}_{ij_1}(b')}{\mathbf{d}_{ij_1}(b', a')} \frac{Y_{ij_2} - \mathbf{m}_{ij_2}(b')}{\mathbf{d}_{ij_2}(b', a')} - \mathbf{r}_{ij_1j_2}(b', a', c), \tag{6c}$$

where the weights $W_{ij_1j_2}^{(\text{m})}(\theta')$, $W_{ij}^{(\text{d})}(\theta')$, and $W_{ij_1j_2}^{(\text{r})}(\theta')$ can take arbitrary nontrivial forms. We choose $W_{ij_1j_2}^{(\text{m})}(\theta')$ to be the (j_1, j_2) element of the inverse of the covariance matrix of $\{D_{i1}^{(\text{m})}(b), \ldots, D_{iM_i}^{(\text{m})}(b)\}$ under a certain working model, and $W_{ij}^{(\text{d})}(\theta')$ and $W_{ij_1j_2}^{(\text{r})}(\theta')$ the inverses of the variances of $D_{ij}^{(\text{d})}(b', a')$ and $D_{ij_1j_2}^{(\text{r})}(b', a', c')$, respectively.

Consider the following algorithm for estimating $\boldsymbol{\theta}$, which is similar to the GEE method (Liang and Zeger, 1986; Zeger and Liang, 1986) when (3b) and (3c) are combined into a covariance model to be estimated empirically. At each step, update the values of $b, a,$ and c to be the maximizers of $L_N^{(\mathrm{m})}(b, \theta')$, $L_N^{(\mathrm{d})}(a, \theta')$, and $L_N^{(\mathrm{r})}(c, \theta')$, respectively, while fixing $\theta' = (b', a', c')$ at its value obtained from the previous iteration. At convergence, the final estimate $\widehat{\boldsymbol{\theta}}_N = (\widehat{\mathbf{b}}_N, \widehat{\mathbf{a}}_N, \widehat{\mathbf{c}}_N)$ will be such that for any $\theta = (b, a, c)$, $L_N(\widehat{\boldsymbol{\theta}}_N,\ \widehat{\boldsymbol{\theta}}_N) = L_N^{(\mathrm{m})}(\widehat{\mathbf{b}}_N,\ (\widehat{\mathbf{b}}_N, \widehat{\mathbf{a}}_N, \widehat{\mathbf{c}}_N)) + L_N^{(\mathrm{d})}(\widehat{\mathbf{a}}_N,\ (\widehat{\mathbf{b}}_N, \widehat{\mathbf{a}}_N, \widehat{\mathbf{c}}_N)) + L_N^{(\mathrm{r})}(\widehat{\mathbf{c}}_N,\ (\widehat{\mathbf{b}}_N, \widehat{\mathbf{a}}_N, \widehat{\mathbf{c}}_N)) \geq L_N^{(\mathrm{m})}(b,\ (\widehat{\mathbf{b}}_N, \widehat{\mathbf{a}}_N, \widehat{\mathbf{c}}_N)) + L_N^{(\mathrm{d})}(a,\ (\widehat{\mathbf{b}}_N, \widehat{\mathbf{a}}_N, \widehat{\mathbf{c}}_N)) + L_N^{(\mathrm{r})}(c,\ (\widehat{\mathbf{b}}_N, \widehat{\mathbf{a}}_N, \widehat{\mathbf{c}}_N)) = L_N(\theta, \widehat{\boldsymbol{\theta}}_N)$.

The iterative likelihood (4) can be naturally extended into nonparametric settings. Consider modeling the mean response using smoothing splines. Let $(X_{ij}^{[1]}, \ldots, X_{ij}^{[K^{(\mathrm{m})}]})$ be the $1 \times K^{(\mathrm{m})}$ row vector of the predictors for the jth response of the ith subject and $\mathbf{b} = (\mathbf{b}_{[1]}(), \ldots, \mathbf{b}_{[K^{(\mathrm{m})}]}())$ be the "parameter", where $\mathbf{b}_{[k]}()$ is an unknown smoothing spline corresponding to $X_{ij}^{[k]}$. For specific splines $b = (b_{[1]}(), \ldots, b_{[K^{(\mathrm{m})}]}())$, one can replace (3a) and (5a) by $\mathrm{E}Y_{ij} = \mathbf{m}_{ij}(b) = \sum_{k=1}^{K^{(\mathrm{m})}} b_{[k]}(X_{ij}^{[k]})$, and $L_N^{(\mathrm{m})}(b, \theta') \overset{\text{def.}}{=} -\sum_{i=1}^{N} \sum_{j_1, j_2} \{W_{ij_1j_2}^{(\mathrm{m})}(\theta') \cdot D_{ij_1}^{(\mathrm{m})}(b) \cdot D_{ij_2}^{(\mathrm{m})}(b)\} - \sum_{i=1}^{N} \sum_{k=1}^{K^{(\mathrm{m})}} \left[s_k \int \left\{ \frac{d^2 b_{[k]}(t)}{dt^2} \right\}^2 dt \right]$, respectively, where s_k is a pre-fixed or unknown nonnegative smoothing parameters reflecting the penalty for non-smoothness of $b_{[k]}(t)$. Even though detailed investigation of its properties requires more research, this extension from a parametric setting to nonparametric smoothing splines is intuitive and straightforward. The derivative of the iterative likelihood defined here can be used to define a system of penalized estimating equations (Fu, 2003). A direct extension of estimating functions into a nonparametric setting would involve higher level mathematics (Hastie and Tibshirani, 1990, p 105-136) even in its formulation.

Example 3 (Missing covariates problem) Let Y_i be the response of the ith subject and X_i and Z_i be its predictors. While Y_i and X_i are observed for all subjects, Z_i is missing for some subjects with a probability depending on Y_i and X_i. Let R_i be 1 if Z_i is observed and 0 otherwise,

$f_{Y|Z}(y|z, x; b)$ be the conditional distribution of Y_i given Z_i and X_i, characterized by parameters $\mathbf{b}$ and $f_{Z|Y}(z|y, x; a)$ be the conditional distribution of Z_i given Y_i and X_i, characterized by parameters $\mathbf{a}$. Note that $f_{Y|Z}$ and $f_{Z|Y}$ do not constitute a likelihood decomposition and thus $\mathbf{b}$ and $\mathbf{a}$ are not functionally independent. Even so, we treat them as separate parameters and define an iterative likelihood for $\boldsymbol{\theta} = (\mathbf{b}, \mathbf{a})$ as

$$L_N(\theta, \theta') = \sum_{i=1}^{N} l_i(\theta, \theta'), \text{ with } l_i(\theta, \theta') = l_i((b, a), (b', a')) \overset{\text{def.}}{=} l_i^{(0)}(b, a') + l_i^{(1)}(a), \tag{7}$$

where $\theta = (b, a)$, $\theta' = (b', a')$ and $l_i^{(0)}(b, a') \overset{\text{def.}}{=} R_i{\cdot}\log f_{Y|Z}(Y_i|Z_i, X_i; b) + (1-R_i){\cdot}\int\{\log f_{Y|Z}(Y_i|z, X_i; b)\} f_{Z|Y}(z|Y_i, X_i; a')dz$, $l_i^{(1)}(a) \overset{\text{def.}}{=} R_i{\cdot}\log f_{Z|Y}(Z_i|Y_i, X_i; a)$. We denote $L_N^{(0)}(b, a') \overset{\text{def.}}{=} \sum_{i=1}^{N} l_i^{(0)}(b, a')$ and $L_N^{(1)}(a) \overset{\text{def.}}{=} \sum_{i=1}^{N} l_i^{(1)}(a)$.

When Y_i is conditionally independent of X_i given Z_i, and $(Y_i,\ X_i)$ is discrete with a small number of possible realizations and $\mathbf{a}$ is nonparametric, this reduces to the mean score method proposed by Reilly and Pepe (1995). In general, the estimate $\widehat{\boldsymbol{\theta}}_N = (\widehat{\mathbf{b}}_N, \widehat{\mathbf{a}}_N)$ satisfies

$$\frac{\partial L_N(\theta,\ \theta')}{\partial \theta}\Big|_{\theta=\theta'=\widehat{\theta}_N} = \left\{ \frac{\partial L_N^{(0)}(b,\ a')}{\partial b}\Big|_{b=\widehat{\mathbf{b}}_N, a'=\widehat{\mathbf{a}}_N}, \ \frac{\partial L_N^{(1)}(a)}{\partial a}\Big|_{a=\widehat{\mathbf{a}}_N} \right\} = 0.$$

The above three examples illustrate the generality of the iterative likelihood in representing a number of seemingly unrelated statistical approaches.

2.2. Estimates from an iterative likelihood

A simple algorithm for estimating $\boldsymbol{\theta}$ from an iterative likelihood $L_N(\theta; \theta')$ is as follows.

Algorithm 1. Step 1. Set $l = 0$ and $\theta^{(0)}$ to be an initial value of θ;

Step 2. Increment l to $l+1$ and set $\theta^{(l+1)}$ to be (2.a) a value of θ that maximizes $L_N(\theta, \theta^{(l)})$, (2.b) a value of θ that locally maximizes $L_N(\theta, \theta^{(l)})$, or (2.c) a root of the equations $\partial L_N(\theta, \theta^{(l)})/\partial\theta = 0$;

Step 3. Repeat Step 2.a, 2.b, or 2.c until convergence and set $\widehat{\boldsymbol{\theta}}_N$ to be $\theta^{(l+1)}$.

Depending on whether Step 2a, 2b, or 2c is used in the iterations, we can give a hierarchy of definitions for the estimate $\widehat{\boldsymbol{\theta}}_N$ from the iterative likelihood $L_N(\theta, \theta')$.

Definition 1. An estimate of $\boldsymbol{\theta}$ is referred to as:

(a) a global estimate from $L_N(\theta, \theta')$ if

$$L_N(\widehat{\boldsymbol{\theta}}_N, \widehat{\boldsymbol{\theta}}_N) - L_N(\theta, \widehat{\boldsymbol{\theta}}_N) \geq 0 \text{ for any } \theta \in \Theta; \tag{8a}$$

(b) a local estimate from $L_N(\theta, \theta')$ if an open neighborhood of $\widehat{\boldsymbol{\theta}}_N$ exists such that

$$L_N(\widehat{\boldsymbol{\theta}}_N, \widehat{\boldsymbol{\theta}}_N) - L_N(\theta, \widehat{\boldsymbol{\theta}}_N) \geq 0 \text{ for any } \theta \text{ in the neighborhood;} \tag{8b}$$

(c) a stationary estimate from $L_N(\theta, \theta')$ if

$$\frac{\partial L_N(\theta, \theta')}{\partial \theta}\bigg|_{\theta=\theta'=\widehat{\theta}_N} = 0. \tag{8c}$$

By definition, a global estimate must be a local estimate, which must be a stationary estimate. In Example 1, 2, or 3, Algorithm 1 with Step 2.b, 2.a, or 2.c is used, yielding a local, global, or stationary estimate, respectively. Further, we define a neighborhood of the estimates in each of the three types.

Definition 2. For an iterative likelihood $L_N(\theta, \theta')$ and a given $\varepsilon > 0$, we refer to the subsets of the parameter space Θ,

$$\widehat{\Theta}_{\varepsilon N}^{(\text{global})} \overset{\text{def.}}{=} \{\theta \in \Theta : \text{ there exists a global estimate } \widehat{\boldsymbol{\theta}}_N \text{ such that } ||\theta - \widehat{\boldsymbol{\theta}}_N|| \leq \varepsilon\},$$

$$\widehat{\Theta}_{\varepsilon N}^{(\text{local})} \overset{\text{def.}}{=} \{\theta \in \Theta : \text{ there exists a local estimate } \widehat{\boldsymbol{\theta}}_N \text{ such that } ||\theta - \widehat{\boldsymbol{\theta}}_N|| \leq \varepsilon\},$$

$$\widehat{\Theta}_{\varepsilon N}^{(\text{stationary})} \overset{\text{def.}}{=} \{\theta \in \Theta : \text{ there exists a stationary estimate } \widehat{\boldsymbol{\theta}}_N \text{ such that } ||\theta - \widehat{\boldsymbol{\theta}}_N|| \leq \varepsilon\},$$

as $\varepsilon-$neighborhoods of the global, local, and stationary estimates, respectively, where $|| \cdot ||$ is the Euclidean norm for a vector or square matrix.

By definition, $\widehat{\Theta}_{0N}^{(\text{global})}$, $\widehat{\Theta}_{0N}^{(\text{local})}$, and $\widehat{\Theta}_{0N}^{(\text{stationary})}$ are the neighborhoods of global, local, and stationary estimates, respectively. As detailed later, one has to search in one of these three sets for a consistent and asymptotically normal estimate, depending on the condition for the true value of $\boldsymbol{\theta}$ to satisfy in the expectation of $L_N(\theta, \theta')$. Note that (8c) is a system of estimating equations, which is implied by (8b), which is implied by (8a). Thus, there always exists a system of estimating functions corresponding to any iterative likelihood. Likewise, any system of estimating equations can be expressed in the form of an iterative likelihood. To see this, let $\boldsymbol{\theta} = (\boldsymbol{\theta}_{[1]}, \ldots, \boldsymbol{\theta}_{[K]})$ be the parameters and $G_N(\theta) = (G_N^{[1]}(\theta), \ldots, G_N^{[K]}(\theta))$ be a system of estimating functions. Correspondingly, we can define

$$L_N(\theta, \theta') = \sum_{k=1}^{K} \left\{ w_k \int_{-\infty}^{\theta_{[k]}} G_N^{[k]}(\theta^*)|_{\theta^*_{[k]}=\alpha,\, \theta^*_{[-k]}=\theta'_{[-k]}} \, d\alpha \right\},$$

where w_k is a non-zero weight and the subscript $[-k]$ indexes the elements in a vector except the k-th one. Note $\partial L_N(\theta, \theta')/\partial\theta|_{\theta=\theta'=\widehat{\theta}_N} = 0$. However, the expression of an iterative likelihood corresponding to a system of estimating equations may not be unique because one may obtain different iterative likelihoods by using different weights or permuting the order of elements in $\boldsymbol{\theta}$. Of note, unless the derivative matrix of estimating equations with respect to θ is symmetric, a system of estimating equations cannot be expressed as a likelihood-based function such as quasi-likelihoods (McCullagh and Nelder, 1989, p. 333-334), which does not involve θ'. Thus, introducing θ' is critical for ensuring the existence of an iterative likelihood corresponding to every system of estimating equations.

Despite the connection, an iterative likelihood may be preferred to estimating equations in representing an inference vehicle. 1) An iterative likelihood is more directly related to the underlying statistical model and can possibly be given meaningful interpretations while estimating equations are often treated as a computational tool. For example, in a specific case of the weighted least squares estimation, an iterative likelihood corresponds to a weighted sum of squared losses and it can be interpreted as a measure of the goodness-of-fit of the underlying statistical model, while a system of estimating equations does not have this direct interpretation. 2) Extending an iterative likelihood to a more general parameter space, for example, discrete or functional, is natural and straightforward while defining estimating equations beyond Euclidean space may be challenging. 3) The representation of iterative likelihood is richer than that of estimating equations because multiple iterative likelihoods may correspond to one system of estimating equations. This is helpful to characterize the inference vehicle from more vantage points. Reducing the representation from an iterative likelihood to estimating equations may result in loss of some useful characteristics of the inference vehicle. For example, the representation of iterative likelihood allows for a distinction among global, local, and stationary estimates, which are consistent and asymptotically normal in different senses under different conditions, as detailed in Section 3.2.

3. THEORY

This section provides a general theory on iterative likelihood. We start with some definitions in Section 3.1. Section 3.2 focuses on the asymptotic properties for estimates from an iterative likelihood. Section 3.3 examines estimates under a misspecified model.

3.1. Attraction in the expectation of iterative likelihoods

Consider a sequence of iterative likelihoods $\{L_N(\theta, \theta')\}$ as defined in (1). Throughout, we assume the following regularity conditions:

1) all Ys are independent of each other;

2) the function $l_i(\theta, \theta')$ is twice differentiable with respect to θ and θ' in the interior of Θ, the

expectations of its first two derivatives exist, and the order of the expectation and differentiation operators can be exchanged;

3) both $\partial l_i(\theta,\theta')/\partial\theta$ and $\partial^2 l_i(\theta,\theta')/\partial\theta^2$ are bounded and continuous functions of θ and θ' uniformly for i, θ, θ' and the data.

We denote

$$G_N(\theta) \overset{\text{def.}}{=} \frac{1}{N}\sum_{i=1}^{N} g_i(\theta), \text{ where } g_i(\theta) \overset{\text{def.}}{=} \frac{\partial l_i(\theta,\theta')}{\partial\theta}\Big|_{\theta'=\theta}, \tag{9a}$$

$$H_N^{(0)}(\theta) \overset{\text{def.}}{=} \frac{1}{N}\sum_{i=1}^{N} h_i^{(0)}(\theta), \text{ where } h_i^{(0)}(\theta) \overset{\text{def.}}{=} -\frac{\partial^2 l_i(\theta,\theta')}{\partial\theta^2}\Big|_{\theta'=\theta}, \tag{9b}$$

$$H_N^{(1)}(\theta) \overset{\text{def.}}{=} \frac{1}{N}\sum_{i=1}^{N} h_i^{(1)}(\theta), \text{ where } h_i^{(1)}(\theta) \overset{\text{def.}}{=} \frac{\partial^2 l_i(\theta,\theta')}{\partial\theta\partial\theta'}\Big|_{\theta'=\theta}, \tag{9c}$$

$$H_N(\theta) \overset{\text{def.}}{=} H_N^{(0)}(\theta) - H_N^{(1)}(\theta) = \frac{1}{N}\sum_{i=1}^{N} h_i(\theta), \text{ where } h_i(\theta) \overset{\text{def.}}{=} h_i^{(0)}(\theta) - h_i^{(1)}(\theta), \tag{9d}$$

$$U_N(\theta) \overset{\text{def.}}{=} \frac{1}{N}\sum_{i=1}^{N} u_i(\theta), \text{ where } u_i(\theta) \overset{\text{def.}}{=} \left[\left\{\frac{\partial l_i(\theta,\theta')}{\partial\theta}\right\}^t \frac{\partial l_i(\theta,\theta')}{\partial\theta}\right]\Big|_{\theta'=\theta}. \tag{9e}$$

Here, $G_N(\theta)$ is the system of estimating functions corresponding to $L_N(\theta,\theta')$, with $H_N(\theta)$ being the negative Hessian matrix, possibly asymmetric, and $U_N(\theta)$ the outer product of the elementwise estimating functions. In $H_N(\theta)$, $H_N^{(1)}(\theta)$ corresponds to the variability due to θ' being unknown. By the notation, $\widehat{\boldsymbol{\theta}}_N$ is a stationary estimate from $L_N(\theta,\theta')$ if $G_N(\widehat{\boldsymbol{\theta}}_N) = 0$ and it is a local estimate if, in addition, $\rho(H_N^{(0)}(\widehat{\boldsymbol{\theta}}_N)) \geq 0$, where $\rho(A)$ is the smallest eigenvalue of $(A + A^t)/2$ for a square matrix A.

To understand the behavior of an estimate from an iterative likelihood, we examine the "attraction point", which represents an "estimate in expectation", with a hierarchy of definitions given below, corresponding to the hierarchy of definitions for estimates in Definition 1. When the sample size is large, an estimate is expected to be close to an attraction. In practice, one constructs an iterative likelihood such that the true value of the parameter is an attraction in the expectation of the iterative likelihood under the model assumptions, thereby ensuring that an estimate is close to the true value in large samples when the model is correctly specified. However, since the behavior of estimates under a misspecified model is also our interest, we define a sequence of attractions instead of a fixed attraction, noting that when the model is misspecified, the "estimate in expectation" may depend on N. Consider an example of modeling the mean of Y_i, m_i as $\boldsymbol{\theta}$ and use $\widehat{\boldsymbol{\theta}}_N = (Y_1 + \cdots + Y_N)/N$ to estimate $\boldsymbol{\theta}$. Under the model, all m_is reduce to a common value, which is an attraction point. However, when the model is misspecified, the expectation of $\widehat{\boldsymbol{\theta}}_N$, $\theta_{0N} = (m_1 + \cdots + m_N)/N$, may depend on N. In fact, $\{\theta_{0N}\}$ may not even converge.

Definition 3. Let $\{L_N(\theta, \theta')\}$ be a sequence of iterative likelihoods for $\boldsymbol{\theta}$ as defined in (1), with $G_N(\theta)$ and $H_N(\theta)$ defined in (9a) and (9d), respectively. We call a sequence of values of $\boldsymbol{\theta}$, $\{\theta_{0N}\}$,

(a) a sequence of global attraction points in $\{\mathrm{E}L_N(\theta, \theta')\}$ if there exists a constant $C > 0$ such that

$$||\theta - \theta_{0N}|| \le \{\mathrm{E}L_N(\theta_{0N}, \theta) - \mathrm{E}L_N(\theta, \theta)\} \cdot C/N \quad \text{for all } \theta \in \Theta \text{ and } N; \tag{10a}$$

(b) a sequence of local attraction points in $\{\mathrm{E}L_N(\theta, \theta')\}$ if there exist constants $C > 0$ and $\delta_0 > 0$ such that

$$||\theta - \theta_{0N}|| \le \{\mathrm{E}L_N(\theta_{0N}, \theta) - \mathrm{E}L_N(\theta, \theta)\} \cdot C/N \quad \text{for all } \theta \in \Theta,\ ||\theta - \theta_{0N}|| < \delta_0 \text{ and } N; \tag{10b}$$

(c) a sequence of stationary attraction points in $\{\mathrm{E}L_N(\theta, \theta')\}$ if there exists a positive number $\varepsilon_0 > 0$ such that for all N,

$$\mathrm{E}G_N(\theta_{0N}) = 0, \tag{10c}$$

$$\sigma(\mathrm{E}H_N(\theta_{0N})) > \varepsilon_0, \tag{10d}$$

where $\sigma(A)$ denotes the smallest absolute eigenvalue of $(A + A^t)/2$ for a square matrix A.

When all θ_{0N}s reduce to a common value θ_0, we say θ_0 is a global, local, or stationary attraction when (a), (b), or (c) holds, respectively. By definition, a sequence of global attractions must be a sequence of local attractions, which must be a sequence of stationary attractions. Note that $\{\theta_{0N}\}$ is a sequence of local attractions if and only if 1) (10c) holds and 2) there exists $\varepsilon_0 > 0$ such that for all N,

$$\rho(\mathrm{E}H_N(\theta_{0N})) > \varepsilon_0. \tag{11}$$

The latter is immediate from the Taylor expansion under the regularity condition: $\frac{1}{N}\{\mathrm{E}L_N(\theta_{0N}, \theta) - \mathrm{E}L_N(\theta, \theta)\} = -(\theta - \theta_{0N})\, \mathrm{E}G^t_N(\theta_{0N}) + \frac{1}{2}(\theta - \theta_{0N})\, \frac{\mathrm{E}H_N(\theta_{0N}) + \mathrm{E}H^t_N(\theta_{0N})}{2}\, (\theta - \theta_{0N})^t + r_N(\theta_{0N})$, where $r_N(\theta_{0N})/(||\theta - \theta_{0N}||^2)$ converge to 0 uniformly for N as $||\theta - \theta_{0N}||$ tends to 0. We refer to (10d) and (11) as uniform nonsingularity and uniform positive definiteness, respectively.

When $\{\theta_{0N}\}$ is a sequence of global attractions, 1) $\mathrm{E}L_N(\theta_{0N}, \theta)) \geq \mathrm{E}L_N(\theta, \theta)$ holds, ensuring that the corresponding estimating equations are unbiased, and 2) θ is close to θ_{0N} if $\mathrm{E}L_N(\theta, \theta)$ nears $\mathrm{E}L_N(\theta_{0N}, \theta)$, amounting to convexity and then identifiability of θ at θ_{0N} in $\mathrm{E}L_N(\theta, \theta')/N$ in some global sense. When $\{\theta_{0N}\}$ is a sequence of local attractions, 1) the estimating equations are unbiased, and 2) local convexity of $\mathrm{E}L_N(\theta, \theta')/N$ at $\theta = \theta_{0N}$, and then local identifiability holds. When $\{\theta_{0N}\}$ is a sequence of stationary attractions, 1) the estimating equations are unbiased, and 2) local identifiability holds.

3.2. Asymptotic properties

The asymptotic properties of estimates from a sequence of iterative likelihoods are as follows.

Theorem 1. Let $\{L_N(\theta, \theta')\}$ be a sequence of iterative likelihoods for $\boldsymbol{\theta}$ as defined in (1), with $G_N(\theta)$, $H_N(\theta)$, and $U_N(\theta)$ denoted in (9a), (9d), and (9e), respectively. Let $\{\theta_{0N}\}$ be a sequence of values of $\boldsymbol{\theta}$. Under the regularity conditions 1)-3), we have:

(a) If $\{\theta_{0N}\}$ is a sequence of global attractions in $\{\mathrm{E}L_N(\theta, \theta')\}$, then for any sequence of global estimates $\{\widehat{\boldsymbol{\theta}}_N\}$ from $\{L_N(\theta, \theta')\}$, $\{\widehat{\boldsymbol{\theta}}_N - \theta_{0N}\}$ converges in probability to 0;

(b) If $\{\theta_{0N}\}$ is a sequence of local attractions in $\{\mathrm{E}L_N(\theta, \theta')\}$, then there exists a sequence of estimates $\{\widehat{\boldsymbol{\theta}}_N\}$ from $\{L_N(\theta, \theta')\}$ such that for any $\varepsilon > 0$,

$$\lim_{N\to\infty} \Pr(||\widehat{\boldsymbol{\theta}}_N - \theta_{0N}|| \leq \varepsilon \text{ and } \widehat{\boldsymbol{\theta}}_N \text{ is a local estimate}) = 1; \tag{12a}$$

(c) If $\{\theta_{0N}\}$ is a sequence of stationary attractions in $\{\mathrm{E}L_N(\theta, \theta')\}$, then there exists a sequence of estimates $\{\widehat{\boldsymbol{\theta}}_N\}$ from $\{L_N(\theta, \theta')\}$ such that for any $\varepsilon > 0$,

$$\lim_{N\to\infty} \Pr(||\widehat{\boldsymbol{\theta}}_N - \theta_{0N}|| \leq \varepsilon \text{ and } \widehat{\boldsymbol{\theta}}_N \text{ is a stationary estimate}) = 1; \tag{12b}$$

(d) If $\{\theta_{0N}\}$ is a sequence of global, local, or stationary attractions in $\{\mathrm{E}L_N(\theta, \theta')\}$ and $\{\widehat{\boldsymbol{\theta}}_N\}$ is a sequence of estimates from $\{L_N(\theta, \theta')\}$ such that $\{\widehat{\boldsymbol{\theta}}_N - \theta_{0N}\}$ converges in probability to 0, then $\{\sqrt{N}(\widehat{\boldsymbol{\theta}}_N - \theta_{0N})\}$ is asymptotically normally distributed with mean 0 and variance matrix,

$$\{\mathrm{E}H_N(\theta_{0N})\}^{-1} \cdot \mathrm{E}U_N(\theta_{0N}) \cdot \left[\{\mathrm{E}H_N(\theta_{0N})\}^t\right]^{-1}. \tag{12c}$$

The proof of the theorem is provided in the supplementary materials.

Results (a) and (d) imply that if the true value of $\boldsymbol{\theta}$ is a global attraction, any sequence of global estimates will be consistent and asymptotically normal, but the theorem does not ensure the existence of a global estimate. When $L_N(\theta, \theta')$ is a true likelihood, which does not involve θ', the true value is generally a global attraction, even though exceptions do exist (e.g. Ferguson, 1982). Often, an iterative likelihood is constructed such that the true value can be verified as a local or stationary attraction, but not a global attraction. When the true value is a local attraction, Results (b) and (d) imply that there exists a consistent and asymptotically normal estimate that is a local estimate, with a probability tending to 1. Similarly, when the true value is a stationary attraction, Results (c) and (d) imply that there exists a consistent and asymptotically normal estimate that is a stationary estimate, with a probability tending to 1. In either case, the global estimate is not necessarily consistent; but the set of local or stationary estimates will include a consistent and asymptotically normal estimate, with a probability tending to 1. When the set includes only one estimate, one can be quite certain that that estimate is consistent and asymptotically normal if the sample size is large. When the set includes multiple estimates, however, the iterative likelihood may not lend itself useful for identifying a consistent and asymptotically normal estimate from the set and one may rely on the substantive knowledge, the statistical model, or other criteria (e.g., Heyde and Morton, 1998) to choose an estimate. This might be considered the price for using an iterative likelihood as opposed to a true likelihood, although such problems sometimes occur even when a true likelihood is used. In conclusion, when the true value of the parameter is a global attraction, any global estimate is consistent and asymptotically normal; when the true value can only be verified as a local or stationary attraction, one has to search in the set of all possible local or stationary estimates to find a consistent and asymptotically normal estimate.

To better understand the practical implication of the true value of the parameter being a local or stationary attraction, we may rewrite (12a) and (12b) as

$$\lim_{N\to\infty} \Pr\{\theta_{0N} \in \widehat{\Theta}_{\varepsilon N}^{(\text{local})}\} = 1, \text{ and } \lim_{N\to\infty} \Pr\{\theta_{0N} \in \widehat{\Theta}_{\varepsilon N}^{(\text{stationary})}\} = 1,$$

respectively, saying the probability that any neighborhood of local or stationary estimates covers the true value is close to 1 when the sample size is large enough.

Result (d) says that for a sequence of estimates $\{\widehat{\boldsymbol{\theta}}_N\}$ with $\{\widehat{\boldsymbol{\theta}}_N - \theta_{0N}\}$ converging in probability to 0, $\{\sqrt{N}(\widehat{\boldsymbol{\theta}}_N - \theta_{0N})\}$ is asymptotically normally distributed with mean 0 and variance matrix (12c). When $(\widehat{\boldsymbol{\theta}}_N - \theta_{0N})$ converges in probability to 0, we may use

$$H_N(\widehat{\boldsymbol{\theta}}_N) = H_N^{(0)}(\widehat{\boldsymbol{\theta}}_N) - H_N^{(1)}(\widehat{\boldsymbol{\theta}}_N) = -\sum_{i=1}^{N} \left\{ \frac{\partial^2 l_i(\theta, \theta')}{\partial \theta^2} + \frac{\partial^2 l_i(\theta, \theta')}{\partial \theta \partial \theta'} \right\} \Big|_{\theta=\theta'=\widehat{\theta}_N}, \text{ and}$$

$$U_N(\widehat{\boldsymbol{\theta}}_N) = \sum_{i=1}^{N} \left[\left\{ \frac{\partial l_i(\theta, \theta')}{\partial \theta} \right\}^t \frac{\partial l_i(\theta, \theta')}{\partial \theta} \right] \Big|_{\theta=\theta'=\widehat{\theta}_N},$$

to approximate $H_N(\theta_{0N})$ and $U_N(\theta_{0N})$, respectively. It follows from the law of large numbers and Result (a) of Lemma 1 in the Appendix that both $H_N(\theta_{0N}) - \mathrm{E}H_N(\theta_{0N})$ and $U_N(\theta_{0N}) - \mathrm{E}U_N(\theta_{0N})$ converge in probability to 0. Thus, the asymptotic variance (12c) can be approximated by

$$\{H_N^{(0)}(\widehat{\boldsymbol{\theta}}_N) - H_N^{(1)}(\widehat{\boldsymbol{\theta}}_N)\}^{-1} \cdot U_N(\widehat{\boldsymbol{\theta}}_N) \cdot [\{(H_N^{(0)}(\widehat{\boldsymbol{\theta}}_N)\}^t - \{H_N^{(1)}(\widehat{\boldsymbol{\theta}}_N)\}^t]^{-1}, \tag{13}$$

where $H_N^{(1)}$ reflects the additional variability in the estimation of $\boldsymbol{\theta}$ due to θ' being estimated. When θ' is absent from $L_N(\theta, \theta')$, where $H_N^{(1)}(\widehat{\boldsymbol{\theta}}_N) = 0$, (13) reduces to the "sandwich" estimate (e.g. Huber, 1967). We term (13) robust variance.

If all θ_{0N}s are equal to θ_0 and for all i,

$$\mathrm{E}\{u_i(\theta_0) - h_i(\theta_0)\} = 0, \tag{14}$$

then $\mathrm{E}U_N(\theta_0) = \mathrm{E}H_N(\theta_0)$, and (12c) will be equal to $\{EU_N(\theta_0)\}^{-1}$ or $\{H_N(\theta_0)\}^{-1}$, implying that the asymptotic variance of $\sqrt{N}(\widehat{\boldsymbol{\theta}}_N - \theta_{0N})$ can be approximated by either

$$2\{H_N(\widehat{\boldsymbol{\theta}}_N) + H_N^t(\widehat{\boldsymbol{\theta}}_N)\}^{-1}, \text{ or} \tag{15}$$

$$\{U_N(\widehat{\boldsymbol{\theta}}_N)\}^{-1}. \tag{16}$$

We refer to (15) and (16) as H- and U- variances, respectively. The example given in Section 5.1 satisfies condition (14) and thus the asymptotic variance can also be approximated by the H- or U- variance. Lindsay (1982) termed (14) as information unbiasedness, which allows an estimating function to be treated as a score function up to the second order. Gan and Jiang (1999) used (14) to test consistency and asymptotically efficiency of an estimate from a true likelihood.

3.3. Estimates under a misspecified model

Usually, an iterative likelihood is constructed such that the true value of $\boldsymbol{\theta}$ is an attraction under a working model. Because Theorem 1 ensures that an estimate $\widehat{\boldsymbol{\theta}}_N$ is close to an attraction θ_{0N} without alluding to correctness of a model, we can discuss the behavior of $\widehat{\boldsymbol{\theta}}_N$ under a misspecified model by examining θ_{0N} under a model relaxed from the working one. We give a general treatment here with a specific example discussed in Section 5.2.

Let φ_0 denote the true distribution of the data. Let Φ^* denote a working statistical model, which is a set of possible distributions that may or may not include φ_0. The model Φ^* is said to be misspecified if it does not include φ_0. Let Φ^{**} be a model relaxed from Φ^*, that is, a superset of Φ^* assumed to include φ_0. When θ_{0N} is either a global, local, or stationary attraction in $\mathrm{E}L_N(\theta, \theta')$ under Φ^{**}, θ_{0N} satisfies

$$\frac{\partial \mathrm{E}L_N(\theta, \theta')}{\partial \theta}\Big|_{\theta=\theta'=\theta_{0N}} = \mathrm{E}G_N(\theta_{0N}) = 0, \tag{17}$$

where the expectation is taken with respect to the unknown true distribution φ_0 in Φ^{**}. If θ_{0N} is uniquely identified by (17) under Φ^{**} along with the respective condition for global, local, or stationary attraction, one can express θ_{0N} as a function of φ_0 under Φ^{**}, $\theta_{0N} = \boldsymbol{\theta}_N[\varphi_0; \Phi^{**}]$. For the working model Φ^*, let $\boldsymbol{\theta}_N[\varphi_0; \Phi^*]$ be the expression of $\boldsymbol{\theta}_N[\varphi_0; \Phi^{**}]$ under the constraints given by Φ^*. Depending on whether $\boldsymbol{\theta}_N[\varphi_0; \Phi^*]$ is a global, local, or stationary attraction under Φ^{**}, the respective asymptotic property in Theorem 1 follows. Note that $\boldsymbol{\theta}_N[\varphi_0; \Phi^*]$ may not always be global, local, or stationary attraction under Φ^{**}. For example in GEE, if one is only interested in the mean parameters and they are correctly specified, then $\boldsymbol{\theta}_N[\varphi_0; \Phi^*]$ can be global, local, or stationary attraction even when the variance and data distribution are misspecified. However, if variance parameters are also of interest, then $\boldsymbol{\theta}_N[\varphi_0; \Phi^*]$ may not be a global, local, or stationary attraction. In this case, θ_{0N} in Theorem 1 need to be replaced by $\boldsymbol{\theta}_N[\varphi_0; \Phi^*]$. A specific example along with more detailed discussions on this will be presented in Section 5.2.

Let $\{\widehat{\boldsymbol{\theta}}_N \}$ be a sequence of estimates such that $\{\widehat{\boldsymbol{\theta}}_N - \theta_{0N}\} = \{\widehat{\boldsymbol{\theta}}_N - \boldsymbol{\theta}_N[\varphi_0; \Phi^{**}]\}$ converges in probability to 0 under Φ^*. Consider a subvector of $\boldsymbol{\theta}$, $\boldsymbol{\beta}$, with corresponding components β_{0N}, $\widehat{\boldsymbol{\beta}}_N$, $\boldsymbol{\beta}_N[\varphi_0; \Phi^{**}]$ and $\boldsymbol{\beta}_N[\varphi_0; \Phi^*]$ in θ_{0N}, $\widehat{\boldsymbol{\theta}}_N$, $\boldsymbol{\theta}_N[\varphi_0; \Phi^{**}]$ and $\boldsymbol{\theta}_N[\varphi_0; \Phi^*]$, respectively. Because $\boldsymbol{\beta}_N[\varphi_0; \Phi^*]$ can be treated as the true value of $\boldsymbol{\beta}$ when the working model is correctly specified, it would be independent of N. When $\boldsymbol{\beta}_N[\varphi_0; \Phi^{**}]$ is also independent of N and can be attached with the same interpretation as $\boldsymbol{\beta}_N[\varphi_0; \Phi^*]$, we can say that $\widehat{\boldsymbol{\beta}}_N$, which is obtained from the working model Φ^*, is consistent and asymptotically normal under a relaxed model Φ^{**}.

Often, one can achieve this robustness with an iterative likelihood in the form $L_N(\theta, \theta') = L_N^{(0)}(\beta, \theta') + L_N^{(1)}(\lambda, \theta')$, where $\boldsymbol{\theta} = (\boldsymbol{\beta}, \boldsymbol{\lambda})$ and $\partial \mathrm{E} L_N^{(0)}(\beta, \theta')/\partial \beta = 0$ holds at $\beta = \beta_0$ with λ being any value under Φ^{**}. This is facilitated by the separation of θ and θ', which allows the construction of estimating functions for β independently of λ while fixing $\theta' = (\beta', \lambda')$, thereby ensuring the robustness of estimating $\boldsymbol{\beta}$ under a misspecified model Φ^*.

4. Estimation

This section focuses on finding estimates from an iterative likelihood and their asymptotic variance. Even though Algorithm 1 can be used to find estimates, it is inadequate: 1) It requires the availability of a procedure for performing maximization or solving equations; 2) It may be inefficient in that at each iteration θ may be updated to a value further away from θ' while θ' and θ are eventually required to be the same; 3) It does not ensure that all global, local, or stationary estimates can be found.

We present an algorithm for finding estimates from $L_N(\theta, \theta')$ in (1), with $G_N(\theta)$, $H_N^{(0)}(\theta)$, $H_N^{(1)}(\theta)$, and $H_N(\theta\ )$ defined in (9a), (9b), (9c) and (9d), respectively. Section 4.1 describes a modified Newton method, which is easy to program and possibly more efficient than Algorithm 1. In Section 4.2, we discuss evaluations of derivatives, which are required in implementing the algorithms and computing the asymptotic variance.

4.1. A modified Newton algorithm

A modified Newton method for finding a stationary estimate from $L_N(\theta, \theta')$ is as follows.

Algorithm 2. Step 1. Set $l = 0$ and $\theta^{(0)}$ to be an initial value of θ;

Step 2. Increment l to $l + 1$ and set $\theta^{(l+1)}$ to be $K(\theta^{(l)})$, where

$$K(\theta) = \theta + G_N(\theta)\ \{H_N^{(0)}(\theta)\}^{-1}, \tag{18}$$

Step 3. Repeat Step 2 until convergence and set $\widehat{\boldsymbol{\theta}}_N$ to be $\theta^{(l+1)}$.

While Step 2 of Algorithm 1 involves a multiple-step search for a maximizer or a stationary value of $L(\theta, \theta^{(l)})$ with respect to θ, Step 2 of Algorithm 2 uses a single-step search. Algorithm 2 may be more efficient in that it saves computation for updating θ to a value further away from $\theta' = \theta^{(l)}$. We may impose some constraints on (18) such as $L_N(\theta^{(l+1)}, \theta^{(l)}) > L_N(\theta^{(l)}, \theta^{(l)})$ for finding estimates of a specific type. We may also multiply by a constant, say $\zeta \in (0, 1]$, to the last term of (18) to make the algorithm more stable. In Example 1 of the EM algorithm, the constraint $L_N(\theta^{(l+1)}, \theta^{(l)}) > L_N(\theta^{(l)}, \theta^{(l)})$ would ensure each iteration increases the observed data likelihood. If $H_N^{(0)}(\theta)$ were replaced by $H_N(\theta)$ in (18), Algorithm 2 would reduce to the regular

Newton procedure. The following theorem gives sufficient conditions for an iteration sequence from Algorithm 2 to converge.

Theorem 2. Let $\widehat{\boldsymbol{\theta}}_N$ be a stationary estimate from $L_N(\theta, \theta')$. Then, there exists an open neighborhood of $\widehat{\boldsymbol{\theta}}_N$ such that for any initial value $\theta^{(0)}$ in that neighborhood, the iteration sequence $\theta^{(0)}, \theta^{(1)}, \theta^{(2)}, \ldots$, from Algorithm 2, where $\theta^{(l+1)} = K(\theta^{(l)})$, $l = 1, 2, \ldots$, and $K()$ is defined in (18), converges to $\widehat{\boldsymbol{\theta}}_N$ provided that (i) $H_N^{(0)}(\theta)$ and $G_N(\theta)$ are differentiable with respect to θ at $\theta = \widehat{\boldsymbol{\theta}}_N$; (ii) The largest absolute eigenvalue of the matrix $H_N^{(1)}(\widehat{\boldsymbol{\theta}}_N) \cdot \{H_N^{(0)}(\widehat{\boldsymbol{\theta}}_N)\}^{-1}$ is smaller than 1.

The proof of the theorem is provided in the supplementary materials. Condition (i) holds when the iterative likelihood is smooth enough. Heuristically, Condition (ii) means that $L_N(\theta, \theta')$ is more sensitive to θ than to θ'. As shown in Section 5 for our examples, this often holds in applications. The importance of this theorem is more theoretical than practical. It ensures that, given the conditions, any iteration sequence from Algorithm 2 with an initial value close enough to a stationary estimate will converge to that estimate. In practice, one may generate a sequence of initial values. Under Conditions (i) and (ii), if Algorithm 2 is repeated over this sequence of initial values long enough, the probability that all the estimates can be found is close to 1. If no sequence converges, the regular Newton method may be used.

Our goal is to identify the sets of global, local, and stationary estimates, namely, $\widehat{\Theta}_{0N}^{(\text{global})}$, $\widehat{\Theta}_{0N}^{(\text{local})}$, and $\widehat{\Theta}_{0N}^{(\text{stationary})}$. After each application of Algorithm 2 that results in a converging sequence, we include the estimate $\widehat{\boldsymbol{\theta}}_N$ in $\widehat{\Theta}_{0N}^{(\text{stationary})}$; if further $\rho(H_N^{(0)}(\widehat{\boldsymbol{\theta}}_N)) > 0$, we include $\widehat{\boldsymbol{\theta}}_N$ in $\widehat{\Theta}_{0N}^{(\text{local})}$. Finally, we include $\widehat{\boldsymbol{\theta}}_N$ in $\widehat{\Theta}_{0N}^{(\text{global})}$ if it is confirmed to be a global estimate. This may be achieved probabilistically by generating a random but ergodic sequence of values for θ and comparing $L_N(\theta, \widehat{\boldsymbol{\theta}}_N)$ with $L_N(\widehat{\boldsymbol{\theta}}_N, \widehat{\boldsymbol{\theta}}_N)$. However, this confirmation is unnecessary when the true value is not verified as a global attraction or $\widehat{\Theta}_{0N}^{(\text{stationary})}$ or $\widehat{\Theta}_{0N}^{(\text{local})}$ can be otherwise narrowed down to a reasonable estimate.

4.2. Evaluations of derivatives and asymptotic variance

In Algorithm 2, $G_N(\theta)$ and $H_N^{(0)}(\theta)$ are evaluated while in a regular Newton method, $G_N(\theta)$ and $H_N(\theta)$ are evaluated. Algorithm 2 may be preferred in that $H_N^{(0)}(\theta)$ is easier to evaluate than $H_N(\theta)$. Specifically, Algorithm 2 only requires evaluations of the first two derivatives of $L_N(\theta, \theta')$ with respect to θ while fixing θ'. In all the three examples given, these evaluations are straightforward. Even so, we indicate that this advantage may be less important as tools such as automatic differentiation (Griewank and Corliss, 1991) for evaluating derivatives become readily available. Automatic differentiation allows one to find accurate derivatives of any programmable function without recoding the function. To obtain the asymptotic variance (13), one needs to compute $H_N^{(0)}(\widehat{\boldsymbol{\theta}}_N)$, $H_N^{(1)}(\widehat{\boldsymbol{\theta}}_N)$, and $U_N(\widehat{\boldsymbol{\theta}}_N)$. One obtains $H_N^{(0)}(\widehat{\boldsymbol{\theta}}_N)$ and $U_N(\widehat{\boldsymbol{\theta}}_N)$ in evaluating $\partial l_i(\theta, \theta')/\partial\theta$ and $\partial^2 l_i(\theta, \theta')/\partial^2\theta$ at $\theta = \theta' = \widehat{\boldsymbol{\theta}}_N$, which are routinely provided in our estimation algorithms; one may obtain $H_N^{(1)}(\widehat{\boldsymbol{\theta}}_N)$ in evaluating $\partial^2 l_i(\theta, \theta')/\partial\theta\partial\theta'$ at $\theta = \theta' = \widehat{\boldsymbol{\theta}}_N$ via automatic differentiation or the finite difference method.

Of note, separation of θ and θ' in defining an iterative likelihood is helpful for computational programming. The key is to program a function for the iterative likelihood, which takes two arguments θ and θ' and returns individual iterative likelihoods. Evaluating this function and its derivatives provides all necessary information for implementing the estimation algorithm and for computing the asymptotic variance.

5. APPLICATION

5.1. EM algorithm: censored HIV RNA in an AIDS clinical trial

Consider the iterative likelihood (2) for the EM algorithm formulated in Example 1. Note

$$\begin{aligned} g_i(\theta) = \frac{\partial l_i(\theta, \theta')}{\partial \theta}\Big|_{\theta'=\theta} &= \left[\frac{\partial}{\partial \theta}\int \{\log f_Z(z|X_i;\theta)\}\cdot f_{Z|Y}(z|Y_i, X_i;\theta')dz\right]\Big|_{\theta'=\theta} \\ &= \frac{1}{f_Y(Y_i|X_i;\theta)}\left\{\frac{\partial}{\partial \theta}\int f_Z(Y_i, z|X_i;\theta)dz\right\} = \frac{\partial}{\partial \theta}\{\log f_Y(Y_i|X_i;\theta)\} \end{aligned}$$

is actually the score function for the observed data. Note $h_i(\theta) = -\partial g_i(\theta)/\partial \theta$ and $u_i(\theta) = g_i^t(\theta)$ $g_i(\theta)$. Thus, for the true value of $\boldsymbol{\theta}$, θ_0, $\mathrm{E}g_i(\theta_0) = 0$ and $\mathrm{E}h_i(\theta_0) = \mathrm{E}u_i(\theta_0)$, which ensures that, under mild conditions, θ_0 is a local attraction and that (14) holds. However, θ_0 may not be a global attraction. This is expected because Result (a) of Theorem 1 would have otherwise implied that any estimate obtained via the EM algorithm with global maximization at the M-step is consistent, contradicting to the observation (Boyles, 1983; Wu, 1983) that an estimate obtained via the EM algorithm is not necessarily even a maximum likelihood estimate nor consistent.

It follows from Results (b) and (d) of Theorem 1 that one needs to search in the set of all local estimates for a consistent and asymptotically normal estimate. In this particular application to the EM algorithm, such an estimate can be identified by comparing the likelihood

$$\sum_{i=1}^{N}\log f_Y(Y_i|X_i;\theta) = \sum_{i=1}^{N}\log\left[\int\left\{f_Z(z|X_i;\theta)\cdot f_{Y|Z}(Y_i|X_i, z;\theta\right\}dz\right]$$

for the values of θ in the finite set of local estimates. The iterative likelihood (2) itself, like the EM algorithm, however, does not provide a direct way to identify such an estimate.

Because (14) holds, the asymptotic variance can be approximated by either the robust variance (13), the H-variance (15) or the U-variance (16). Note $H_N(\theta_0) = H_N^{(0)}(\theta_0) - H_N^{(1)}(\theta_0)$, where $H_N^{(0)}(\theta_0)$ is the observed complete data information and $H_N^{(1)}(\theta_0)$ is the observed missing data information (Louis, 1982). Thus, the H-variance is an estimate for the Louis' corrected variance for the EM algorithm. The U-variance is easier to compute because $U_N(\theta_0) = \sum\limits_{i=1}^{N}\{g_i^t(\theta_0)g_i(\theta_0)\}$ is a direct output from every iteration. This simple formula for computing the variance was noted by Meilijson (1989). Finally, when information unbiasedness (14) is questionable, one can always use the robust variance (13).

Note that both (2) and $\sum\limits_{i=1}^{N} \log f_Y(Y_i|X_i;\theta)$ are iterative likelihoods corresponding to the estimating equations, $\sum_{i=1}^{N} \frac{\partial\{\log f_Y(Y_i|X_i;\theta)\}}{\partial\theta} = 0$, where the latter is a degenerate iterative likelihood that does not involve θ'. The advantages for using the former over the latter in computation comes at the expense of the need to search the set of local estimates for a consistent and asymptotically normal estimate.

Consider a specific example from the AIDS Clinical Trial Group Study 398 (Hammer et al., 2002), a randomized trial to compare antiviral strategies in treating HIV-infected subjects with HIV RNA viral load greater than 1,000 copies/mL at randomization (week 0). Our interest is to compare the HIV RNA level at week 24 between the two treatment groups: those who were randomized to take double protease inhibitor (PI) and those to take single PI. For simplicity, we only use the 231 subjects who had not taken non-nucleoside reverse transcriptase inhibitor (NNRTI) for more than 7 days by week 0 and who had HIV RNA measurements at weeks 0 and 24, with 152 and 79 subjects in the double PI and single PI groups, respectively. Due to technical limitations to the assay, some HIV RNA values are not directly observed, but are known to be below or above a certain detection limit. In our data, all HIV RNA values at week 0 are actually observed, but 26.4% of the measurements at week 24 were censored below detection limits, with the detection limits ranging from 7 to 45 copies/mL.

For subject i, let Z_i denote the complete data, the actual level of HIV RNA in $\log_{10}$ at week

24 and C_i be the censoring limit. We may reasonably assume that Z_i and C_i are independent of each other and thus effectively treat C_i as fixed. The observed response Y_i can be written as $\{\max(Z_i, C_i), \mathrm{I}(Z_i \leq C_i)\}$. Let $X_i = (X_i^{[1]}, X_i^{[2]})$ be the covariates, where $X_i^{[1]}$ is the treatment indicator (1 = double PI and 0 = single PI), and $X_i^{[2]}$ is the HIV RNA value in $\log_{10}$ at week 0. We assume that Z_i is normally distributed with mean $b_1 + b_2X_i^{[1]} + b_3X_i^{[2]}$ and variance $\exp(2a_1)$. We define an iterative likelihood for $\boldsymbol{\theta} = (\mathbf{b}_0, \mathbf{b}_1, \mathbf{b}_2, \mathbf{a}_0)$, $L_N(\theta, \theta')$ as in (2), with

$$
\begin{aligned}
f_Z(z|x;\theta) &= \frac{1}{\sqrt{2\pi}\exp(a_1)} \exp\left\{-\frac{(z - b_1 - b_2x^{[1]} - b_3x^{[2]})^2}{2\exp(2a_1)}\right\}, \\
f_{Z|Y}(z|y, x;\theta) &= \mathrm{I}(z > C) + \mathrm{I}(z \leq C) \cdot \exp\left\{-\frac{(z - b_1 - b_2x^{[1]} - b_3x^{[2]})^2}{2\exp(2a_1)}\right\} \cdot \\
&\qquad \left[\int_{-\infty}^{C} \exp\left\{-\frac{(s - b_1 - b_2x^{[1]} - b_3x^{[2]})^2}{2\exp(2a_1)}\right\} ds\right]^{-1}.
\end{aligned}
$$

Starting with a reasonable initial value, we use Algorithm 2 to find a stationary estimate from $L_N(\theta, \theta')$. We then define an initial box as this estimate plus and minus 20 times the robust standard deviation. We repeat Algorithm 2 for 100 initial values randomly generated from this initial box. It turns out all the sequences that meet the convergence criteria converge to the same stationary estimate and we conclude that the stationary estimate is unique. By Theorem 1, it is consistent and asymptotically normal, with a probability tending to 1. Because $H_N^{(0)}(\widehat{\boldsymbol{\theta}}_N)$ is the observed complete data information and $H_N^{(1)}(\widehat{\boldsymbol{\theta}}_N)$ is the observed missing data information, Condition (ii) of Theorem 2 would hold in expectation. For the AIDS data, the matrix $H_N^{(1)}(\widehat{\boldsymbol{\theta}}_N) \cdot \{H_N^{(0)}(\widehat{\boldsymbol{\theta}}_N)\}^{-1}$ is positive definite and its largest eigenvalue is 0.4, and Condition (ii) holds. The numerical results are presented in Table 1 with the standard deviations obtained via the robust, H- and U- variances given in (13), (15) and (16), respectively. It turns out that patients taking double protease inhibitor did have lower HIV RNA levels at week 24 at the significance level of 0.05, and the initial HIV RNA value at week 0 has a positive effect on the HIV RNA level at week 24. We also apply the original EM algorithm to analyze this data set and the results are identical to those obtained from the iterative likelihood using Algorithm 2.

5.2. GEE: FEV$_1$ in a pulmonary function study

Consider the iterative likelihood $L_N(\theta, \theta')$ in (4) for the model formulated in Example 2. We are interested in the case when the model is misspecified. Consider a hierarchy of four statistical models: $\Phi_0 = \{\mathrm{E}Y_{ij} = m_{ij}, \mathrm{Sd}Y_{ij} = d_{ij}, \mathrm{Corr}(Y_{ij_1}, Y_{ij_2}) = r_{ij_1j_2}\}$, $\Phi_1 = \{\mathrm{E}Y_{ij} = \mathbf{m}_{ij}(b), \mathrm{Sd}Y_{ij} = d_{ij},\ \mathrm{Corr}(Y_{ij_1}, Y_{ij_2}) = r_{ij_1j_2}\}$, $\Phi_2 = \{\mathrm{E}Y_{ij} = \mathbf{m}_{ij}(b), \mathrm{Sd}Y_{ij} = \mathbf{d}_{ij}(b,a), \mathrm{Corr}(Y_{ij_1}, Y_{ij_2}) = r_{ij_1j_2}\}$, $\Phi_3 = \{\mathrm{E}Y_{ij} = \mathbf{m}_{ij}(b), \mathrm{Sd}Y_{ij} = \mathbf{d}_{ij}(b,a), \mathrm{Corr}(Y_{ij_1}, Y_{ij_2}) = \mathbf{r}_{ij_1j_2}(b,a,c)\}$, where m_{ij}, d_{ij}, and $r_{ij_1j_2}$ denote the unconstrained individual values for mean, standard deviation, and correlation, respectively, while $\mathbf{m}_{ij}()$, $\mathbf{d}_{ij}()$, and $\mathbf{r}_{ij}()$ are the mean, standard deviation, and correlation models, respectively, which denote the known functions mapping the parameter values b, (b,a), and (b,a,c) to the mean, standard deviation, and correlation, respectively. Models Φ_1, Φ_2, and Φ_3 correspond to the following three assumptions. (i) The mean model is correctly specified; (ii) The mean and standard deviation models are correctly specified; and (iii) The mean, standard deviation, and correlation models are correctly specified.

Note that Φ_3 is a subset of Φ_2, which is a subset of Φ_1, which is a subset of Φ_0. Even though the construction of the iterative likelihood (4) for $\boldsymbol{\theta} = (\mathbf{b}, \mathbf{a}, \mathbf{c})$ is based on the working statistical model Φ_3, it follows from (5a-5c) and (6a-6c) that under Φ_0,

$$\begin{aligned}
&\mathrm{E}L_N(\theta,\ \theta') = \mathrm{E}L_N((b,a,c),\ (b',a',c')) \\
&= C(\theta') - \sum_{i=1}^{N}\sum_{j_1,j_2}\Big(W^{(\mathrm{m})}_{ij_1j_2}(\theta')[\{m_{ij_1} - \mathbf{m}_{ij_1}(b)\}\{m_{ij_2} - \mathbf{m}_{ij_2}(b)\}]\Big) \\
&-\sum_{i=1}^{N}\sum_{j}\Big(W^{(\mathrm{d})}_{ij}(\theta')\ [d^2_{ij} + \{m_{ij} - \mathbf{m}_{ij}(b')\}^2 - \mathbf{d}^2_{ij}(b',a)]^2\Big)
\end{aligned}$$

$$-\sum_{i=1}^{N}\sum_{j_1j_2}\left(W^{(\mathrm{r})}_{ij_1j_2}(\theta')\left[\frac{r_{ij_1j_2}\cdot d_{ij_1}\ d_{ij_2}+\{m_{ij_1}-\mathbf{m}_{\ ij_1}(b')\}\ \{m_{ij_2}-\mathbf{m}_{ij_2}(b')\}}{\mathbf{d}_{ij_1}(b',a')\ \mathbf{d}_{ij_1}(b',a')}-\mathbf{r}_{ij_1j_2}(b',a',c)\right]^2\right),$$

where $\theta=(b,a,c)$, $\theta'=(b',a',c')$, and $C(\theta')$ depends on θ' but not on θ. Assume that $\mathbf{m}_{ij}$, $\mathbf{d}_{ij}$, and $\mathbf{r}_{ij}$ are chosen so that there exists a sequence of stationary attraction points in $\{\mathrm{E}L_N(\theta,\theta')\}$. Let $\theta_{0N}=(b_{0N},a_{0N},c_{0N})$ denote the value of $\boldsymbol{\theta}=(\mathbf{b},\mathbf{a},\mathbf{c})$ such that

$$\frac{\partial\mathrm{E}L_N(\theta,\theta')}{\partial b}\Big|_{\theta'=\theta=\theta_{0N}}=0,\quad \frac{\partial\mathrm{E}L_N(\theta,\theta')}{\partial a}\Big|_{\theta'=\theta=\theta_{0N}}=0,\quad \frac{\partial\mathrm{E}L_N(\theta,\theta')}{\partial c}\Big|_{\theta'=\theta=\theta_{0N}}=0. \tag{19}$$

Under model Φ_3, θ_{0N} would degenerate to the value $\theta_0=(b_0,a_0,c_0)$ such that

$$m_{ij}-\mathbf{m}_{ij}(b_0)=0, \tag{20a}$$

$$d_{ij}-\mathbf{d}_{ij}(b_0,a_0)=0, \tag{20b}$$

$$r_{ij_1j_2}-\mathbf{r}_{ij_1j_2}(b_0,a_0,c_0)=0. \tag{20c}$$

In that case, the condition for $\theta_{0N}=\theta_0$ to be a global attraction can be easily characterized. For example, the following conditions are sufficient: 1) For any $\varepsilon>0$, there exists $\delta(\varepsilon)>0$ such that $||(b-b_0,a-a_0,c-c_0)||>\varepsilon$ would imply that $\mathbf{m}_{ij}(b),\mathbf{d}_{ij}(b,a)$, and $\mathbf{r}_{ij_1j_2}(b,a,c)$ differ from $\mathbf{m}_{ij}(b_0),\mathbf{d}_{ij}(b_0,a_0)$, and $\mathbf{r}_{ij_1j_2}(b_0,a_0,c_0)$, respectively, by at least $\delta(\varepsilon)$ in Euclidean metric; and 2) There exist ξ,η, $0<\xi<\eta$, such that as a $M_i\times M_i$ matrix, $\{W^{(\mathrm{m})}_{ij_1j_2}(\theta')-\xi{\cdot}\mathrm{I}(j_1=j_2)\}$ is positive definite, $\{W^{(\mathrm{m})}_{ij_1j_2}(\theta')-\eta{\cdot}\mathrm{I}(j_1=j_2)\}$ is negative definite, and $W^{(\mathrm{d})}_{ij}(\theta')$ and $W^{(\mathrm{r})}_{ij_1j_2}(\theta')$ are within $[\xi,\eta]$. Under model Φ_2 or Φ_1, θ_{0N} would degenerate to $\theta_{0N}=(b_0,a_0,c_{0N})$ or $\theta_{0N}=(b_0,a_{0N},c_{0N})$, respectively, where b_0 and a_0 satisfy (20a) and (20b). In such a case, the condition for (b_0,a_0) or b_0 to be a global attraction when c_{0N} or (a_{0N},c_{0N}) is fixed at any unknown values can also be similarly characterized, but formalizing these conditions requires extended definitions of attraction from the entire parameter vector into a subvector, for which Theorem 1 can be improved. To limit the length of the article, we assume here directly that θ_{0N} is a global attraction under Φ_1,

Φ_2, or Φ_3. Thus, by Theorem 1, for any sequence of global estimates $\{\widehat{\boldsymbol{\theta}}_N\}$, $\{\widehat{\boldsymbol{\theta}}_N - \theta_{0N}\} = \{(\widehat{\mathbf{b}}_N - b_{0N}, \widehat{\mathbf{a}}_N - a_{0N}, \widehat{\mathbf{c}}_N - c_{0N})\}$ converges in probability to 0, with asymptotic normality.

Our interest is the behavior of $\widehat{\boldsymbol{\theta}}_N$ when the working model Φ_3 misspecifies the distribution of the data. We use the notation defined in Section 3.3. First, consider a relaxed model Φ_1 and $\boldsymbol{\beta} = \mathbf{b}$. It follows from (19) that $\boldsymbol{\beta}_N[\varphi_0; \Phi_1]$ is the value of $\mathbf{b}$, b_0 such that (20a) holds for all i and j and so is $\boldsymbol{\beta}_N[\varphi_0; \Phi_3]$. Hence, both $\boldsymbol{\beta}_N[\varphi_0; \Phi_1]$ and $\boldsymbol{\beta}_N[\varphi_0; \Phi_3]$ are independent of N and can be attached with the same interpretation as parameters characterizing means m_{ij}s. Second, consider a relaxed model Φ_2 and $\boldsymbol{\beta} = (\mathbf{b}, \mathbf{a})$. It follows from (19) that $\boldsymbol{\beta}_N[\varphi_0; \Phi_1]$ is the value of $(\mathbf{b}, \mathbf{a})$, (b_0, a_0) such that (20a) and (20b) hold for all i and j and so is $\boldsymbol{\beta}_N[\varphi_0; \Phi_2]$. Hence, both $\boldsymbol{\beta}_N[\varphi_0; \Phi_2]$ and $\boldsymbol{\beta}_N[\varphi_0; \Phi_3]$ are independent of N and can be attached with the same interpretation as parameters characterizing means m_{ij}s and standard deviations d_{ij}s. Under either the relaxed model Φ_1 or Φ_2, the global estimate $\widehat{\mathbf{b}}_N$ or $(\widehat{\mathbf{b}}_N, \widehat{\mathbf{a}}_N)$, obtained from the working model Φ_3, will still be consistent and asymptotically normal.

In conclusion, under some mild conditions, estimates $\widehat{\mathbf{b}}_N$, $(\widehat{\mathbf{b}}_N, \widehat{\mathbf{a}}_N)$, and $(\widehat{\mathbf{b}}_N, \widehat{\mathbf{a}}_N, \widehat{\mathbf{c}}_N)$ converge in probability to b_0, (b_0, a_0), and (b_0, a_0, c_0), as they are originally defined, with asymptotic normality, when assumptions (i), (ii), and (iii) hold, respectively. Therefore, our estimation procedure based on the iterative likelihood yields a consistent and asymptotically normal estimate for the mean parameters even when the standard deviation and correlation models are misspecified, and for the standard deviation parameters even when the correlation model is misspecified.

Consider a specific example from an observational study (Laird et al., 1992; Wang et al., 1993) designed to characterize the pulmonary function growth in children aged from 6 to 18 years old. In this study, 300 girls were enrolled in grade 1 or 2 and seen annually until high school graduation or loss to follow up. At each visit, spirometry of the pulmonary function was conducted by asking the participants to exhale air with maximal force and velocity into a closed chamber. We focus on the forced expiratory volume (FEV_1), the volume of air (in liter) exhaled in the first second of the manoeuvre, a measure widely used as an indicator of respiratory health. Our objective is to characterize the dependence of FEV_1 on height. Let Y_{ij} be the FEV_1 in log for the ith subject at the jth visit and X_{ij} be the corresponding height in meters. We use the iterative likelihood $L_N(\theta, \theta')$ defined in (4) with the following model specification:

$$\mathrm{E}Y_{ij} = \mathbf{m}_{ij}(b) = b_1 + b_2 X_{ij}, \tag{21a}$$

$$\mathrm{Sd}Y_{ij} = \mathbf{d}_{ij}(a,b) = \exp(a_1 + a_2 \mathbf{m}_{ij}(b)) = \exp(a_1 + a_2(b_1 + b_2 X_{ij})), \tag{21b}$$

$$\mathrm{Corr}(Y_{ij_1}, Y_{ij_2}) = \mathbf{r}_{ij_1 j_2}(c,a,b) = \frac{\mathrm{I}(j_1 = j_2)}{1+\exp(2c_1)} + \frac{\exp(2c_1)\cdot \exp\{-\exp(c_2)|X_{ij_1} - X_{ij_2}|\}}{1+\exp(2c_1)}, \tag{21c}$$

with $\boldsymbol{\theta} = (\mathbf{b}, \mathbf{a}, \mathbf{c})$, where $\mathbf{b} = (\mathbf{b}_1, \mathbf{b}_2)$, $\mathbf{a} = (\mathbf{a}_1, \mathbf{a}_2)$, and $\mathbf{c} = (\mathbf{c}_1, \mathbf{c}_2)$.

We first find an initial box in a similar manner as in Section 5.1. We repeat Algorithm 2 for 100 initial values randomly generated from this initial box. It turns out all the sequences that meet the convergency criteria converge to the same stationary estimate. So the obtained stationary estimate is a global estimate. By Theorem 1, this estimate is consistent and asymptotically normal. Because the expectation of $H_N^{(1)}(\theta)$ at the true value of θ is 0 and so will $H_N^{(1)}(\widehat{\boldsymbol{\theta}}_N)$ be close to 0, Condition (ii) of Theorem 2 would hold in expectation. For this FEV_1 dataset, the largest absolute eigenvalue of $H_N^{(1)}(\widehat{\boldsymbol{\theta}}_N) \cdot \{H_N^{(0)}(\widehat{\boldsymbol{\theta}}_N)\}^{-1}$ is 0.25, and Condition (ii) holds. However, the information unbiasedness (14) does not hold because $g_i^t(\theta)$ is not a score function. We even do not fully specify the distribution of the data. So the H- and U- variances are not applicable here. The numerical results are presented in Table 2 with the standard deviations obtained via the robust variance. Height has a significantly positive effect on the FEV_1. We also use the original GEE approach with various correlation structures to analyze the data. The results for parameter b are all similar. The GEE approach does not have an established procedure for estimating the variances of estimators for parameters in the correlation structures.

5.3. Mean score method: smoking induced lung cancer

Consider the iterative likelihood $L_N(\theta, \theta')$ defined in (7) for the model formulated in Example 3. We assume $f_{Y|Z}(y|z, x; b)$ and $f_{Z|Y}(z|y, x; b)$ are correctly specified with $\theta_0 = (b_0, a_0)$ being the true value of $\boldsymbol{\theta} = (\mathbf{b}, \mathbf{a})$. Noting that R_i and Z_i are conditionally independent of each other given X_i and Y_i, we have

$$\begin{aligned}
\mathrm{E} g_i(\theta_0) &= \left\{ \mathrm{E} \frac{\partial l_i^{(0)}(b, a_0)}{\partial b}\Big|_{b=b_0}, \ \mathrm{E} \frac{\partial l_i^{(1)}(a)}{\partial a}\Big|_{a=a_0} \right\} \\
&= \Bigg(\mathrm{E}_Y \left[\mathrm{E}_{RZ|Y} \left\{ R_i \frac{\partial \log f_{Y|Z}(Y_i|Z_i, X_i; b)}{\partial b} \Big| Y_i \right\} + \right. \\
&\qquad\qquad \left. \mathrm{E}_{R|Y}(1 - R_i|Y_i) \cdot \mathrm{E}_{Z|Y} \left\{ \frac{\partial \log f_{Y|Z}(Y_i|Z_i, X_i; b)}{\partial b} \Big| Y_i \right\} \right] \Big|_{b=b_0}, \\
&\qquad \mathrm{E}_Y \left[\mathrm{E}_{R|Y}(R_i|Y_i) \cdot \mathrm{E}_{Z|Y} \left\{ \frac{\partial \log f_{Z|Y}(Z_i|Y_i, X_i; a)}{\partial a} \Big| Y_i \right\} \right] \Big|_{a=a_0} \Bigg) \\
&= 0.
\end{aligned}$$

Therefore, (10c) holds. Further, we have

$$\mathrm{E} h_i(\theta_0) = -\mathrm{E} \left\{ \frac{\partial^2 l_i(\theta, \theta')}{\partial \theta^2} + \frac{\partial^2 l_i(\theta, \theta')}{\partial \theta \partial \theta'} \right\} \Big|_{\theta'=\theta=\theta_0} = \mathrm{E}_Y \begin{vmatrix} h_{i10} & 0 \\ h_{i20} & h_{i30} \end{vmatrix}, \qquad (22)$$

where $h_{i10} = \mathrm{E}_{Z|Y}\{G_0^t(Y_i|Z_i, X_i) G_0(Y_i|Z_i, X_i)\}$, $h_{i20} = \{r(Y_i, X_i) - 1\} \, \mathrm{E}_{Z|Y}\{G_1^t(Y_i|Z_i, X_i) G_0(Z_i|Y_i, X_i)\}$, $h_{i30} = r(Y_i, X_i) \, \mathrm{E}_{Z|Y}\{G_1^t(Z_i|Y_i, X_i) G_1(Z_i|Y_i, X_i)\}$, and

$$\begin{aligned}
G_0(y|z, x) &= \frac{\partial \log f_{Y|Z}(y|z, x; b)}{\partial b}\Big|_{b=b_0}, \ G_1(z|y, x) = \frac{\partial \log f_{Z|Y}(z|y, x; a)}{\partial a}\Big|_{a=a_0}, \\
r(y, x) &= \mathrm{E}_{R|Y}(R|Y = y, X = x).
\end{aligned}$$

When a is null, that is, the conditional distribution of Z_i given Y_i and X_i is known, $\mathrm{E}[h_i(\theta_0) + h_i^t(\theta_0)]/2$ will reduce to the upper left expression in the matrix in (22), which will be nonnegative definite. When a is unknown, $\mathrm{E}[h_i(\theta_0) + h_i^t(\theta_0)]/2$ will be nonnegative definite if $4\, r(Y_i, X_i) \geq (r(Y_i, X_i) - 1)^2$, namely, $r(Y_i, X_i) > 3 - \sqrt{8} \approx 17.2\%$. Under some conditions ensuring uniformity with respect to N, we may say that when the conditional distribution of Z given Y and X is known or the probability of observing Z is greater than 17.2%, the true value (b_0, a_0) is a local attraction; otherwise, (b_0, a_0) may only be verified as a stationary attraction. In the latter case, one may not exclude a stationary estimate as a consistent and asymptotically estimate even if it is not a local estimate. That is, a consistent and asymptotically normal estimate could be $\widehat{\boldsymbol{\theta}}_N = (\widehat{\mathbf{b}}_N, \widehat{\mathbf{a}}_N)$, at which the negative Hessian matrix, $\partial^2 L_N(\theta, \theta')/\partial\theta^2|_{\theta=\theta'=\widehat{\theta}_N}$ is not positive definite. However, when a_0 is held at any fixed, unknown, value, we can verify that b_0 is a local attraction and this refined condition may be used to narrow down the set of candidate estimates. That again requires an extension of definitions for attractions and a subsequent extension of Theorem 1 from the entire parameter vector to a subvector.

For a specific example, consider a case-control study of smoking-induced lung cancer (García-Closas et al., 1997). In this study, cases were patients diagnosed with primary lung cancer, and admitted for thoracic surgery; controls were friends or spouses of lung cancer, cardiac, or thoracic surgery patients. Information on smoking history was available on 846 cases and 938 controls. The study assessed a variety of biomarkers believed to be associated with the pathway from smoking exposure to lung cancer. We focus on DNA adducts (number per 10^8 nucleotides) in blood cells, measured on a sample of consecutively available cases and on controls matched to cases based on smoking status (never, former and current) and other covariates. In total, 80 subjects have measurements on DNA adducts. The availability of marker data depends not only on exposure and other covariates, but also on outcome. For example, selection of controls for measurements on DNA adducts by matching to cases on smoking status is a process in which the decision depends on exposure (smoking status) and outcome (case-control status). Our objective is to evaluate the usefulness of DNA adducts in characterizing the pathway from smoking exposure to lung cancer, while appropriately accounting for the dependence of the missingness of marker data on the exposure and outcome. Clearly, a complete case analysis would be biased because the missing data are not missing completely at random (Little and Rubin, 2002). Such an analysis would also be highly inefficient because it would use only 80 subjects with DNA adducts out of a total of 1784.

For the ith subject, let Y_i be the indicator of lung cancer (1 = case, 0 = control), X_i be the total packyears, a measure of smoking exposure, and Z_i be the level of DNA adducts. For simplicity, we ignore other covariates. Our interest is to estimate how Y_i depends on X_i and Z_i. We use the iterative likelihood (7) with the following model specification, noting that this is a case-control study and Z_i is nonnegative, but can be zero,

$$\begin{aligned} f_{Y|Z}(y|z,x;b) &= \frac{\mathrm{I}(y=0)+\mathrm{I}(y=1)\cdot\exp(b_1+b_2x+b_3z+b_4xz)}{1+\exp(b_1+b_2x+b_3z+b_4xz)}, \\ f_{Z|Y}(z|y,x;a) &= \mathrm{I}(z>0)\cdot\frac{1}{\sqrt{2\pi}\exp(a_5)}\exp\left\{-\frac{(z-a_1-a_2y-a_3x-a_4yx)^2}{2\exp(2a_5)}\right\} \\ &\quad +\mathrm{I}(z=0)\cdot\int_{-\infty}^{0}\frac{1}{\sqrt{2\pi}\exp(a_5)}\exp\left\{-\frac{(s-a_1-a_2y-a_3x-a_4yx)^2}{2\exp(2a_5)}\right\}ds. \end{aligned}$$

Because the percentage of subjects with measurements on Z is 4.5% (80/1784), which is smaller than 17.2%, the true value of the parameter is not necessarily a local attraction. Thus, we need to search in the set of all stationary estimates for a consistent and asymptotically normal estimate.

To implement the method, numerical approximation of the integral in (7) is necessary as there is no closed-form expression. We adopt the Gauss-Legendre quadrature with 100 points. The results in this example and the simulation studies in Section 6 indicate that the Gauss-Legendre quadrature is efficient and accurate for the problem. If (7) involves multidimensional integration, the numerical integration on sparse grids method proposed by Heiss and Winschel (2008) can be used which is an extension of the Gaussian quadrature to multiple dimensions.

We find an initial box in a similar manner as in Section 5.1. We repeat Algorithm 2 for 100 initial values randomly generated from this initial box. It turns out all the iterations that meet the convergency criteria converge to the same stationary estimate, so the stationary estimate is unique. By Theorem 1, this estimate is consistent and asymptotically normal. Note that $H_N^{(0)}(\widehat{\boldsymbol{\theta}}_N)$ is a block diagonal matrix and $H_N^{(1)}(\widehat{\boldsymbol{\theta}}_N)$ has non-zero elements only in the lower left corner. The eigenvalues of $H_N^{(1)}(\widehat{\boldsymbol{\theta}}_N) \cdot \{H_N^{(0)}(\widehat{\boldsymbol{\theta}}_N)\}^{-1}$ are all 0, so Condition (ii) of Theorem 2 holds. However, the information unbiasedness (14) does not hold because $\mathrm{E}\{h_i(\theta_0)\}$ is not symmetric, and thus the H- and U- variances are not applicable. The numerical results are presented in Table 3 with the standard deviations obtained via the robust variance. We find smoking exposure and the level of DNA adducts have an interaction effect. For this data set, since the variable X_i is continuous, the original mean score method is not applicable.

6. SIMULATION STUDIES

In this section, we present the results from extensive simulation studies to access the finite sample performance of the proposed inferential method and numerical algorithm. The first study mimics the AIDS Clinical Trial Group Study 398.

Example 4 We generate the complete date Z_i from a normal distribution with mean $b_1 + b_2 X_i^{[1]} + b_3 X_i^{[2]}$ and variance $\exp(2a_1)$, where $X_i^{[1]}$ is a Bernoulli random variable with probability of success being 0.5, $X_i^{[2]}$ is a normal random variable $N(0,2)$, $b_1 = -1.5$, $b_2 = -0.5$, $b_3 = 1$ and $a_1 = 0.5$. $X_i^{[1]}$ and $X_i^{[2]}$ are independent. The censoring limit C_i is a uniform random variable on $[-3, -1]$

so that the censor rate is about 46%.

We first ignore the censoring and used least squares method to find initial values. Then we implement Algorithm 2 to solve the iterative likelihood defined in (2). Table 4 gives the simulation results. The biases and the standard errors of the estimators decrease as the sample size increases. The three standard error estimators for parameter b yield reasonable estimates and the confidence intervals have coverage probabilities close to the nominal level. The coverage probability of the confidence interval of a is slightly lower than the nominal level when $n = 100$, but the situation is ameliorated rapidly as the sample size increases. We have considered various choices of parameter values, sample sizes and censor rates. The results are similar to those presented in Table 4. We also implement the original EM algorithm for all the simulation settings and obtained results identical to those from our method. For the model considered in this example, there are closed-form expressions in both the E and M steps, so it takes the EM algorithm a slightly shorter time to finish the simulation than Algorithm 2.

Example 5 The second simulation setup is designed to mimic the study of pulmonary function growth in children. For subject i, the number of observations, M_i, is sampled from integers between 1 and 12 with equal probabilities. Then Y_{ij}s, $j = 1, .., M_i$, are generated from a multivariate normal distribution with mean, variance and correlation structure specified in (21a)- (21c), where X_{ij}s are independent normal random variables with mean 0 and variance 0.5, $b_1 = -2$, $b_2 = 2$, $a_1 = -2$, $a_2 = 0.1$, $c_1 = -0.5$ and $c_2 = 1$.

We use unweighted estimating equations to obtain initial values and then implement Algorithm 2 to solve the iterative likelihood in (4). Table 5 summarizes the simulation results. The estimators of b and a are nearly unbiased and the bias of the estimator of c drops rapidly as n increases. There are some deviations between the coverage probabilities and the nominal level for parameters a and c, but the performance of the confidence intervals improves with the sample size. For the same reason as described in Section 5.2, the H- and U- variances are not applicable.

Example 6 In this study we evaluate the performance of the extended mean score method. To be able to know the true parameters, data are generated from the following model setup.

$$f_{Y|Z}(y|z,x;b) = \frac{\mathrm{I}(y=0) + \mathrm{I}(y=1)\cdot \exp(b_1 + b_2 x + b_3 z + b_4 xz)}{1+\exp(b_1 + b_2 x + b_3 z + b_4 xz)}, \tag{23a}$$

$$f_{Z|Y}(z|y,x;a) = \frac{f_{Y|Z}(y|z,x;b_0) f_Z(z|x;a)}{\int f_{Y|Z}(y|z,x;b_0) f_Z(z|x;a)dz}, \tag{23b}$$

where

$$\begin{aligned} f_Z(z|x;a) = {} & \mathrm{I}(z>0)\cdot \frac{1}{\sqrt{2\pi}\exp(a_3)} \exp\left\{ -\frac{(z-a_1-a_2x)^2}{2\exp(2a_3)} \right\} \\ & + \mathrm{I}(z=0)\cdot \int_{-\infty}^{0} \frac{1}{\sqrt{2\pi}\exp(a_3)} \exp\left\{ -\frac{(s-a_1-a_2x)^2}{2\exp(2a_3)} \right\} ds \end{aligned} \tag{23c}$$

In equation (23b), we assume that a is the only unknown parameter in the conditional density of Z. The true value of b, b_0, is used to specify a conditional density of Z so that data can be generated from the true model. We first generate Z_i from (23c) with $a_1 = 0.3$, $a_2 = 0.2$ and $a_3 = 0.2$, and then generate Y_i from (23b) with $b_1 = -1$, $b_2 = 1$, $b_3 = 1$, $b_4 = -1$. For missing probabilities, we use $P(R_i = 0|Y_i = 0) = 0.7$ and $P(R_i = 0|Y_i = 1) = 0.6$ which yield a missing rate close to 66%. Integrals in (7) and (23b) are approximated numerically by Gauss-Legendre quadrature with 100 points.

In each repetition, the observed data are used to obtain an initial value of b by fitting a logistic regression and an initial value of a by fitting a liner regression. Algorithm 2 is then used to solve the iterative likelihood defined in (7). The simulation results are summarized in Table 6. The parameter estimators have small biases and standard errors, the variance estimators are fairly accurate, and the confidence intervals have reasonable coverage probabilities. The performance of each estimator improves as sample size increases. To exam the performance of the proposed method at a situation similar to the smoking-induced lung cancer study, we also conducted the simulation experiment with a large sample size ($N = 1000$) and a high missing rate around 90%. The method performs well at this scenario. For the same reason as described in Section 5.3, the H- and U- variances are not applicable here.

7. Conclusion

The framework of iterative likelihood provides a unified representation for a number of statistical methods. This representation renders great convenience in constructing inference vehicle, simplifying computation, or maintaining robustness of estimation to model misspecification. As for asymptotic theory, we present a hierarchy of realistically verifiable conditions, with corresponding strategies for finding a consistent and asymptotically normal estimate, for which the algorithm described is useful. Aside from illustrating our methodology, the three examples given are also interesting in their own right. In the first example, all stationary estimates for the EM algorithm are sought; different asymptotic variances are presented. The second example involves a general model for unbalanced data, where robustness of the estimates are discussed. In the third example, the mean score method is extended, where a parametric model is used to impute the missing covariates. For the data analysis, positive measure with zero values are handled.

Many issues remained to be addressed. 1) Multiple expressions in terms of iterative likelihood often exist for a specific system of estimating equations, but how to choose a particular expression over another is unclear. Presumably, one should take into account both computational convenience in finding estimates and theoretical consideration in verifying the conditions for asymptotic properties. 2) Our asymptotic theory and estimation algorithm are limited to the case where the parameter space is Euclidean and data on different subjects are independent. Extension beyond Euclidean space is natural, but requires higher level mathematics. For example, the stationarity of the estimates may be expressed as self-consistency for the iteration operators in a functional space. Further, the independence among subjects may be relaxed by replacing $1/N$ by a more general scaling sequence that stabilizes the second order moments of the iterative likelihood. 3) As seen in Sections 5.2 and 5.3, one may often verify a higher level of attraction condition for a specific component of the parameter vector while holding the rest of the parameters fixed at an unknown value. It is hopeful to use this refined characterization of an iterative likelihood to improve our asymptotic theory. 4) Even though the repeated use of Algorithm 2 as described above can theoretically yield all estimates and confirm the global estimates, with a probability close to 1, this approach is computationally expensive and even insurmountable. Further, the assurance is probabilistic and it is hard to gauge whether the search is thorough enough to be terminated. Further research is needed on optimization algorithms that find all roots of equations and the global maximizers of a scaler function efficiently. A possible direction is the application of interval arithmetic.

REFERENCES

Boyles, R. A. (1983), "On the convergence of the EM algorithm," *Journal of the Royal Statistical Society. Series B (Methodological)*, 47–50.

Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977), "Maximum likelihood from incomplete data via the EM algorithm," *Journal of the Royal Statistical Society. Series B (Methodological)*, 1–38.

Ferguson, T. S. (1982), "An inconsistent maximum likelihood estimate," *Journal of the American Statistical Association*, 77, 831–834.

Fu, W. J. (2003), "Penalized estimating equations," *Biometrics*, 59, 126–132.

Gan, L. and Jiang, J. (1999), "A test for global maximum," *Journal of the American Statistical Association*, 94, 847–854.

García-Closas, M., Kelsey, K. T., Wiencke, J. K., Xu, X., Wain, J. C., and Christiani, D. C. (1997), "A case-control study of cytochrome P450 1A1, glutathione S-transferase M1, cigarette smoking and lung cancer susceptibility (Massachusetts, United States)," *Cancer Causes & Control*, 8, 544–553.

Griewank, A. and Corliss, G. F. (1991), *Automatic Differentiation of Algorithm: Theorem, Implementation, and Application*, Philadelphia: SIAM.

Hammer, S. M., Vaida, F., Bennett, K. K., Holohan, M. K., Sheiner, L., Eron, J. J., Wheat, L. J., Mitsuyasu, R. T., Gulick, R. M., Valentine, F. T., et al. (2002), "Dual vs single protease inhibitor therapy following antiretroviral treatment failure," *Journal of the American Medical Association*, 288, 169–180.

Hand, D. J. and Crowder, M. J. (1996), *Practical Longitudinal Data Analysis*, vol. 34, Chapman & Hall/CRC.

Hastie, T. and Tibshirani, R. (1990), *Generalized Additive Models*, Monographs on statistics and applied probability, London ; New York: Chapman and Hall, 1st ed.

Heiss, F. and Winschel, V. (2008), "Likelihood approximation by numerical integration on sparse grids," *Journal of Econometrics*, 144, 62–80.

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

Huber, P. J. (1967), "The behavior of maximum likelihood estimates under nonstandard conditions," in *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability*, vol. 1, pp. 221–33.

Laird, N. M., Donnelly, C., and Ware, J. H. (1992), "Longitudinal Studies with Continuous Responses," *Statistical Methods in Medical Research*, 1, 225–247.

Liang, K. Y. and Zeger, S. L. (1986), "Longitudinal Data Analysis Using Generalized Linear Models," *Biometrika*, 73, 13–22.

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

Little, R. J. A. and Rubin, D. B. (2002), *Statistical analysis with missing data*, Wiley Series in Probability and Statistics, Hoboken, NJ: Wiley-Interscience, 2nd ed.

Louis, T. A. (1982), "Finding the observed information matrix when using the EM algorithm," *Journal of the Royal Statistical Society. Series B (Methodological)*, 226–233.

McCullagh, P. and Nelder, J. A. (1989), *Generalized linear models*, London New York: Chapman and Hall, 2nd ed.

Meilijson, I. (1989), "A fast improvement to the EM algorithm on its own terms," *Journal of the Royal Statistical Society. Series B (Methodological)*, 127–138.

Reilly, M. and Pepe, M. S. (1995), "A mean score method for missing and auxiliary covariate data in regression models," *Biometrika*, 82, 299–314.

Wang, X., Dockery, D. W., Wypij, D., Fay, M. E., and Ferris, B. G. (1993), "Pulmonary function between 6 and 18 years of age," *Pediatric pulmonology*, 15, 75–88.

Wu, C. (1983), "On the convergence properties of the EM algorithm," *The Annals of Statistics*, 11, 95–103.

Zeger, S. L. and Liang, K.-Y. (1986), "Longitudinal data analysis for discrete and continuous outcomes," *Biometrics*, 121–130.

Table 1: Example 1: Estimates of the parameters for censored HIV RNA data

Parameter	Interpretation	Estimate	SE (Robust)	SE (H)	SE (U)
b_1	intercept in HIV RNA*	-2.046	0.913	0.91	0.939
b_2	treatment in HIV RNA*	-0.715	0.287	0.288	0.295
b_3	initial HIV RNA in HIV RNA*	1.097	0.187	0.186	0.197
a_1	log SD of HIV RNA*	0.698	0.043	0.057	0.087

$*$:at week 24

Table 2: Example 2: Estimates of the parameters for FEV$_1$ data

Parameter	Interpretation	Estimate	SE (Robust)
b_1	intercept in mean FEV$_1$	-2.259	0.031
b_2	height in mean FEV$_1$	2.053	0.021
a_1	intercept in log SD of FEV$_1$	-2.135	0.114
a_2	height in log SD of FEV$_1$	0.027	0.117
c_1	relative magnitude of serial correlation	0.726	0.494
c_2	depenence of serial correlation on lag	0.459	0.277

Table 3: Example 3: Estimates of the parameters for smoking-induced lung cancer data with DNA adducts

Parameter	Interpretation	Estimate	SE (Robust)
b_1	intercept in lung cancer	-2.219	0.783
b_2	packyears in lung cancer	0.473	0.557
b_3	DNA adducts in lung cancer	0.464	0.616
b_4	(packyears)*(DNA adducts) in lun cancer	0.945	0.476
a_1	intercept in DNA adducts	1.166	0.422
a_2	packyears in DNA adducts	-0.410	0.309
a_3	lung cancer in DNA adducts	0.520	0.574
a_4	(packyears)*(lung cancer) in DNA adducts	0.477	0.429
a_5	log SD of DNA adducts	-0.095	0.110

Table 4: Simulation results for Example 4.

	$N = 100$				$N = 200$			
Parameter	b_1	b_2	b_3	a_1	b_1	b_2	b_3	a_1
Estimate	-1.515	-0.474	1.002	0.478	-1.504	-0.507	1.001	0.489
Bias	-0.015	0.026	0.002	-0.022	-0.004	-0.007	0.001	-0.011
StD	0.275	0.391	0.152	0.103	0.192	0.274	0.108	0.073
SE (Robust)	0.272	0.374	0.148	0.098	0.192	0.265	0.105	0.071
SE (H)	0.272	0.375	0.151	0.102	0.192	0.265	0.107	0.072
SE (U)	0.279	0.388	0.162	0.111	0.195	0.270	0.111	0.075
CP (Robust)	0.959	0.930	0.938	0.916	0.950	0.943	0.946	0.937
CP (H)	0.958	0.932	0.942	0.937	0.948	0.943	0.948	0.945
CP (U)	0.966	0.944	0.952	0.952	0.948	0.948	0.951	0.950

Note: The estimated values, associated bias, standard error (StD), mean of the standard error estimator (SE) and the coverage probability (CP) of the 95% confidence interval. H or U indicates that the the standard error estimator is based on the H- or U- variance estimator.

Table 5: Simulation results for Example 5.

	b_1	b_2	a_1	a_2	c_1	c_2
	$N = 100$					
Estimate	-2.000	2.000	-2.003	0.100	-0.487	1.014
Bias	0.000	0.000	-0.003	0.000	0.013	0.014
StD	0.005	0.009	0.063	0.029	0.242	0.641
SE (Robust)	0.005	0.009	0.067	0.030	0.278	0.676
CP (Robust)	0.944	0.955	0.931	0.920	0.986	0.908
	$N = 200$					
Estimate	-2.000	2.000	-2.002	0.099	-0.500	0.988
Bias	0.000	0.000	-0.002	-0.001	0.000	-0.012
StD	0.004	0.007	0.045	0.020	0.162	0.419
SE (Robust)	0.004	0.007	0.045	0.020	0.186	0.425
CP (Robust)	0.945	0.955	0.932	0.931	0.980	0.945

Note: The legend is the same as in Table 4.

Table 6: Simulation results for Example 6

	b_1	b_2	b_3	b_4	a_1	a_2	a_3
				$N = 200$			
Estimate	-1.021	1.016	1.018	-1.017	0.298	0.189	0.181
Bias	-0.021	0.016	0.018	-0.017	-0.002	-0.011	-0.019
StD	0.208	0.213	0.142	0.155	0.179	0.191	0.123
SE (Robust)	0.191	0.207	0.139	0.153	0.181	0.179	0.115
CP (Robust)	0.932	0.945	0.932	0.940	0.955	0.924	0.923
				$N = 400$			
Estimate	-1.001	1.001	1.000	-1.001	0.300	0.201	0.194
Bias	-0.001	0.001	0.000	-0.001	0.000	0.001	-0.006
StD	0.089	0.097	0.061	0.070	0.078	0.082	0.054
SE (Robust)	0.085	0.091	0.062	0.067	0.081	0.081	0.053
CP (Robust)	0.937	0.948	0.954	0.933	0.964	0.945	0.938

Note: The legend is the same as in Table 4.

Supplemental Materials for "Iterative Likelihood: A Unified Inference Tool"

Haiying Wang[a], Dixin Zhang[b], Hua Liang[c] and David Ruppert[d]

[a]Department of Statistics, University of Connecticut, Storrs, CT 06269-4120, USA
haiying.wang@uconn.edu

[b]Department of Finance, Nanjing University, Nanjing, Jiangsu 210093, China
dixinz01@nju.edu.cn

[c]Department of Statistics, George Washington University, Washington, D.C., USA
hliang@gwu.edu

[d] Department of Statistical Science, Cornell University, Ithaca, New York 14853, USA
dr24@cornell.edu

SUMMARY. In this document, we present the detailed proofs for the main results and additional examples.

S.1. A PRELIMINARY LEMMA

Lemma 1. Let $\{B_N(\alpha)\}$ be a sequence of random numbers (scalers, vectors, or matrices) indexed by α, $\alpha \in \Gamma$, where Γ is a bounded subspace of an Euclidean space. For notation, we treat $B_N(\alpha)$ as an $1 \times M$ vector, where M is the number of elements in $B_N(\alpha)$.

(a) If $B_N(\alpha)$ is a continuous function of α uniformly for α and the data and for any fixed α, $\{B_N(\alpha)\}$ converges in probability to 0, then, for any $\varepsilon > 0$, $\lim\limits_{N\to\infty} \Pr(\sup\limits_{\alpha\in\Gamma} ||B_N(\alpha)|| \le \varepsilon) = 1$.

(b) Suppose that for any $\varepsilon > 0$, there exists $\delta(\varepsilon) > 0$ such that $||\alpha_2 - \alpha_1|| \le \delta(\varepsilon)$, α_1, $\alpha_2 \in \Gamma$ implies $||\mathrm{E}[(B_N\{\alpha_2) - B_N(\alpha_1)\}^t \cdot \{B_N(\alpha_2) - B_N(\alpha_1)\}]|| \le \varepsilon$ for all N. If for any

α, $\{B_N(\alpha)\}$ converges in distribution to $F(b;\alpha)$, which is continuous with respect to b, uniformly for α, then for any b, $b \in (-\infty,\infty)^M$, $\lim_{N\to\infty}\sup_{\alpha\in\Gamma}|\Pr(B_N(\alpha) < b) - F(b;\alpha)| = 1$.

Proof: (a) Because $B_N(\alpha)$ is a continuous function of α uniformly for α and the data, for any $\varepsilon > 0$, there exists $\delta(\varepsilon) > 0$ such that for any α_1 and α_2 satisfying $||\alpha_2 - \alpha_1|| \le \delta(\varepsilon)$,

$$||B_N(\alpha_2) - B_N(\alpha_1)|| \le \varepsilon.$$

Because Γ is a bounded Euclidean space, there exists a finite number $J(\varepsilon)$ such that we can divide Γ into subsets $\Gamma_1, \ldots, \Gamma_{J(\varepsilon)}$ with fixed points $\alpha_1 \in \Gamma_1, \ldots, \alpha_{J(\varepsilon)} \in \Gamma_{J(\varepsilon)}$ such that $\sup_{\alpha\in\Gamma_j}||\alpha - \alpha_j|| \le \delta(\varepsilon)$ holds for all $j = 1, \ldots, J(\varepsilon)$. Thus,

$$\begin{aligned}\sup_{\alpha\in\Gamma}||B_N(\alpha)|| &= \sup_{\alpha\in\Gamma}\left\{\sum_{j=1}^{J(\varepsilon)}[\mathrm{I}(\alpha\in\Gamma_j)\cdot||B_N(\alpha)||]\right\} \\ &\le \sup_{\alpha\in\Gamma}\left(\sum_{j=1}^{J(\varepsilon)}\left[\mathrm{I}(\alpha\in\Gamma_j)\cdot\{||B_N(\alpha_j)|| + ||B_N(\alpha) - B_N(\alpha_j)||\}\right]\right) \\ &\le \max_{j=1,\ldots,J(\varepsilon)}||B_N(\alpha_j)|| + \varepsilon.\end{aligned}$$

Letting N tend to ∞ and then ε tend to 0, we have the result.

(b) For any two $1 \times K$ random vectors, Z and Z' with respective distributions $P(b)$ and $P'(b)$ and $\epsilon > 0$, we have $P(b-\epsilon) - \Pr(Z' - Z \ge \epsilon) \le P'(b) \le P(b+\epsilon) + \Pr(Z' - Z \ge \epsilon)$. Hence,

$$P(b-\epsilon) - \frac{||\mathrm{E}\{(Z'-Z)^t(Z'-Z)\}||}{||\epsilon||^2} \le P'(b) \le P(b+\epsilon) + \frac{||\mathrm{E}\{(Z'-Z)^t(Z'-Z)\}||}{||\epsilon||^2}.$$

Hence,

$$|P'(b) - P(b)| \le \frac{||\mathrm{E}\{(Z'-Z)^t(Z'-Z)\}||}{||\epsilon||^2} + \max\{P(b+\epsilon) - P(b), P(b) - P(b-\epsilon)\}.$$

For any $\varepsilon > 0$, let $\delta^*(\varepsilon) > 0$ be such that for any α_1 and α_2 in Γ satisfying $||\alpha_2 - \alpha_1|| \le \delta^*(\varepsilon)$,

$$||E[\{B_N(\alpha_1) - B_N(\alpha_2)\}^t\{B_N(\alpha_1) - B_N(\alpha_2)\}]|| \le \varepsilon \text{ and } |F(b;\alpha_1) - F(b;\alpha_2)| \le \varepsilon.$$

Let $F_N(b;\alpha)$ denote the distribution of $B_N(\alpha)$. Because Γ is a bounded Euclidean space, there exists a finite number $J(\varepsilon)$ such that we can divide Γ into subsets $\Gamma_1, \ldots, \Gamma_{J(\varepsilon)}$ with fixed points $\alpha_1 \in \Gamma_1, \ldots, \alpha_{J(\varepsilon)} \in \Gamma_{J(\varepsilon)}$ such that $\sup\limits_{\alpha \in \Gamma_j} ||\alpha - \alpha_j|| \leq \delta^*(\varepsilon)$ holds for all $j = 1, \ldots, J(\varepsilon)$. Hence, for any $\epsilon > 0$, $\epsilon \in (-\infty, \infty)^M$, we have

$$\begin{aligned}
&\sup_{\alpha \in \Gamma} |(B_N(\alpha) < b) - F(b;\alpha)| = \sup_{\alpha \in \Gamma} \left| \sum_{j=1}^{J(\varepsilon)} \mathrm{I}(\alpha \in \Gamma_j) \cdot \{F_N(b;\alpha) - F(b;\alpha)\} \right| \\
&\quad \leq \sup_{\alpha \in \Gamma} \sum_{j=1}^{J(\varepsilon)} \Big[\mathrm{I}(\alpha \in \Gamma_j) \cdot \{|F_N(b, \alpha_j) - F(b;\alpha_j)| + |F(b;\alpha_j) - F(b;\alpha)| \\
&\qquad\quad + |F_N(b;\alpha) - F_N(b;\alpha_j)|\} \Big] \\
&\quad \leq \max_{j=1,\ldots,J(\varepsilon)} \{|F_N(b, \alpha_j) - F(b;\alpha_j)|\} + \varepsilon \\
&\qquad\quad + \sup_{\alpha \in \Gamma} \sum_{j=1}^{J(\varepsilon)} \mathrm{I}(\alpha \in \Gamma_j) \cdot \Bigg(\frac{||\mathrm{E}[\{(B_N(\alpha) - B_N(\alpha_j)\}^t \{B_N(\alpha) - B_N(\alpha_j)\}]||}{||\epsilon||^2} \\
&\qquad\quad + \max(F_N\{b + \epsilon;\alpha_j) - F_N(b;\alpha_j),\ F_N(b;\alpha_j) - F_N(b - \epsilon;\alpha_j)\} \Bigg) \\
&\quad \leq \max_{j=1,\ldots,J(\varepsilon)} |F_N(b, \alpha_j) - F(b;\alpha_j)| + \varepsilon + \frac{\varepsilon}{||\epsilon||^2} \\
&\qquad\quad + \max_{j=1,\ldots,J(\varepsilon)} [\max\{F(b + \epsilon;\alpha_j) - F(b;\alpha_j),\ F(b;\alpha_j) - F(b - \epsilon;\alpha_j)\} \\
&\qquad\quad + |F_N(b + \epsilon;\alpha_j) - F(b + \epsilon;\alpha_j)| + |F_N(b;\alpha_j) - F(b;\alpha_j)| \\
&\qquad\quad + |F_N(b - \epsilon;\alpha_j) - F(b - \epsilon;\alpha_j)| \Big].
\end{aligned}$$

This term can be further bounded by $\max\limits_{j=1,\ldots,J(\varepsilon)} |F_N(b, \alpha_j) - F(b;\alpha_j)| + \varepsilon + \varepsilon/||\epsilon||^2 + \sup_{\alpha \in \Gamma} [\max\{F(b + \epsilon;\alpha) - F(b;\alpha), F(b;\alpha) - F(b - \epsilon;\alpha)\}] + \max\limits_{j=1,\ldots,J(\varepsilon)} \{|F_N(b + \epsilon;\alpha_j) - F(b + \epsilon;\alpha_j)| + |F_N(b;\alpha_j) - F(b;\alpha_j)| + |F_N(b - \epsilon;\alpha_j) - F(b - \epsilon;\alpha_j)|\}$. Letting N tend to ∞, ε to 0, and then ϵ to 0, we have the result.

S.2. Proof of Theorem 1

We define for any θ and θ' in Θ, $Q_N(\theta,\theta') \overset{\text{def.}}{=} \partial L_N(\theta,\theta')/\partial\theta$, $Q_{0N}(\theta,\theta') \overset{\text{def.}}{=} \mathrm{E}Q_N(\theta,\theta')$, $\quad L_{0N}(\theta,\theta') \overset{\text{def.}}{=} \mathrm{E}L_N(\theta,\theta')$, $G_{0N}(\theta) \overset{\text{def.}}{=} \mathrm{E}G_N(\theta)$, $H_{0N}(\theta) \overset{\text{def.}}{=} \mathrm{E}H_N(\theta)$, $U_{0N}(\theta) \overset{\text{def.}}{=} \mathrm{E}U_N(\theta)$, $T_N(\theta,\theta') \overset{\text{def.}}{=} L_N(\theta,\theta') - \mathrm{E}L_N(\theta,\theta')$, and $A(\delta) \overset{\text{def.}}{=} T_N(\theta+\delta\ (\theta'-\theta),\ \theta')-T_N(\theta,\theta')$. It follows from the regularity condition that, for any θ and θ' in Θ, $dA(\delta)/d\delta = N(\theta'-\theta)\cdot[Q_N(\theta+\delta(\theta'-\theta),\theta') - Q_{0N}(\theta+\delta(\theta'-\theta),\theta')]^t$ is a continuous function of δ. Thus, there exists $\delta^* \in (0,1)$ such that $T_N(\theta',\theta') - T_N(\theta,\theta') = A(1) - A(0) = \frac{dA(\delta)}{d\delta}|_{\delta=\delta^*} = N(\theta-\theta')\cdot\{Q_N(\theta'',\theta') - Q_{0N}(\theta'',\theta')\}^t$, where $\theta'' = \theta + \delta^*(\theta'-\theta)$ is also within Θ because Θ is convex. Hence,

$$\frac{|T_N(\theta',\ \theta') - T_N(\theta,\ \theta')|}{N||\theta'-\theta||} \le ||Q_N(\theta'',\theta') - Q_{0N}(\theta'',\theta')|| \le \sup_{\theta^*,\theta^{**}\in\Theta} ||Q_N(\theta^*,\theta^{**}) - Q_{0N}(\theta^*,\theta^{**})||.$$

For any $\varepsilon > 0$, we have

$$\begin{aligned}
\Pr(||\widehat{\boldsymbol{\theta}}_N - \theta_{0N}|| \le \varepsilon) &\ge \Pr\left(\frac{C\{L_{0N}(\theta_{0N},\widehat{\boldsymbol{\theta}}_N) - L_{0N}(\widehat{\boldsymbol{\theta}}_N,\widehat{\boldsymbol{\theta}}_N)\}}{N||\widehat{\boldsymbol{\theta}}_N - \theta_{0N}||} \le \varepsilon\right)\\
&\ge \Pr\left(\frac{C\{L_N(\theta_{0N},\widehat{\boldsymbol{\theta}}_N) - L_N(\widehat{\boldsymbol{\theta}}_N,\ \widehat{\boldsymbol{\theta}}_N)\}}{\sqrt{N}||\widehat{\boldsymbol{\theta}}_N - \theta_{0N}||} \le 0\right)\\
&\quad - \Pr\left(\frac{C|(L_N\{\widehat{\boldsymbol{\theta}}_N,\widehat{\boldsymbol{\theta}}_N) - L_{0N}(\widehat{\boldsymbol{\theta}}_N,\widehat{\boldsymbol{\theta}}_N)\} - \{L_N(\theta_{0N},\widehat{\boldsymbol{\theta}}_N) - L_{0N}(\theta_{0N},\widehat{\boldsymbol{\theta}}_N)\}|}{N||\widehat{\boldsymbol{\theta}}_N - \theta_{0N}||} > \varepsilon\right)\\
&= 1 - \Pr\left(\frac{|T_N(\widehat{\boldsymbol{\theta}}_N;\widehat{\boldsymbol{\theta}}_N) - T_N(\theta_{0N},\widehat{\boldsymbol{\theta}}_N)|}{N||\widehat{\boldsymbol{\beta}}_N - \theta_{0N}||} > \frac{\varepsilon}{C}\right)\\
&\ge 1 - \Pr\left(\sup_{\theta^*,\theta^{**}\in\Theta}|Q_N(\theta^*,\ \theta^{**}) - Q_{0N}(\theta^*,\ \theta^{**})| > \frac{\varepsilon}{C}\right).
\end{aligned}$$

The regularity condition ensures that $Q_N(\theta^*,\ \theta^{**}) - Q_{0N}(\theta^*,\ \theta^{**})$ is a continuous function of (θ^*,θ^{**}) uniformly for $(\theta^*,\theta^{**}) \in \Theta\times\Theta$ and the law of large numbers ensures that $Q_N(\theta^*,\ \theta^{**}) - Q_{0N}(\theta^*,\ \theta^{**})$ converges in probability to 0 as N tends to ∞. Letting N tend to ∞ and then ε to 0, and applying Lemma 1 (a), we have $\lim_{N\to\infty}\Pr(||\widehat{\boldsymbol{\theta}}_N - \theta_{0N}|| \le \varepsilon) = 1$, which yields (a).

We now prove (b). For an $1 \times K$ vector e, we denote $r_N(e) \overset{\text{def.}}{=} G_N(\theta_{0N} + e) - G_{0N}(\theta_{0N}) + e\, H_{0N}(\theta_{0N})$. Let $\varepsilon_0 > 0$ be the lower bound of $\rho(\mathrm{E}H_N(\theta_{0N})) = \rho(H_{0N}(\theta_{0N}))$. It follows from the regularity condition and the Taylor expansion of $G_{0N}(\theta_{0N} + e)$ around $e = 0$ that for any ε, $0 < \varepsilon \leq \varepsilon_0$, we can find $\delta(\varepsilon)$, $0 < \delta(\varepsilon) \leq \varepsilon$ such that for all N, θ, and e, $||e|| < \delta(\varepsilon) \leq \varepsilon$, $||G_{0N}(\theta_{0N} + e) - G_{0N}(\theta_{0N}) + e \cdot H_{0N}(\theta_{0N})|| \leq \varepsilon ||e||/2 \leq \varepsilon_0 ||e||/2$, and $||H_{0N}(\theta_{0N} + e) - H_{0N}(\theta_{0N})|| \leq \varepsilon/4 \leq \varepsilon_0/4$.

We now consider a probability subspace $\Omega_\varepsilon = \{\sup\limits_{||e|| \leq \delta(\varepsilon)} ||G_N(\theta_{0N} + e) - G_{0N}(\theta_{0N} + e)|| \leq \varepsilon_0 \delta(\varepsilon)/2\} \cap \{\sup\limits_{||e|| \leq \delta(\varepsilon)} ||H_N(\theta_{0N} + e) - H_{0N}(\theta_{0N} + e)|| \leq \varepsilon_0/4\}$. For any data point in Ω_ε, $1 \times K$ vectors e and η, $||e|| \leq \delta(\varepsilon) \leq \varepsilon$, $||\eta|| = 1$, we have $||r_N(e)|| = ||G_{0N}(\theta_{0N} + e) - G_{0N}(\theta_{0N}) + e\, H_{0N}(\theta_{0N}) + G_N(\theta_{0N} + e) - G_{0N}(\theta_{0N} + e)|| \leq \varepsilon_0 ||e||/2 + \varepsilon_0 \delta(\varepsilon)/2 = \varepsilon_0 \delta(\varepsilon)$, and

$$\begin{aligned} \eta\, H_N(\theta_{0N} + e)\, \eta^t &= \eta\, H_{0N}(\theta_{0N})\, \eta^t + \eta\, [H_N(\theta_{0N} + e) - H_{0N}(\theta_{0N} + e)]\, \eta^t \\ &\qquad + \eta \{H_{0N}(\theta_{0N} + e) - H_{0N}(\theta_{0N})\}\, \eta^t \\ &\geq \rho(H_{0N}(\theta_{0N} + e)) - \varepsilon_0/4 - \varepsilon_0/4 \geq \varepsilon_0 - \varepsilon_0/4 - \varepsilon_0/4 = \varepsilon_0/2. \end{aligned}$$

Thus, $\rho(H_N(\theta_{0N} + e)) \geq \varepsilon_0/2$. For the operator S that maps e to $r_N(e)\{H_{0N}(\theta_{0N})\}^{-1}$,

$$\begin{aligned} ||S(e)|| &= ||r_N(e) \cdot \{H_{0N}(\theta_{0N})\}^{-1}|| \leq ||r_N(e)|| \cdot ||\{H_{0N}(\theta_{0N})\}^{-1}|| \\ &\leq \varepsilon_0 \delta(\varepsilon) \cdot \rho(H_{0N}(\theta_{0N}))^{-1} \leq \delta(\varepsilon). \end{aligned}$$

Thus, S is a continuous function from $\{e : ||e|| \leq \delta(\varepsilon)\}$ to $\{e : ||e|| \leq \delta(\varepsilon)\}$. The Brouwer fixed point theorem ensures that there exists $\widehat{e}_N$, $||\widehat{e}_N|| \leq \delta(\varepsilon) \leq \varepsilon$, such that $\widehat{e}_N = S(\widehat{e}_N) = r_N(\widehat{e}_N) \cdot [H_{0N}(\theta_{0N})]^{-1}$. We define $\widehat{\boldsymbol{\theta}}_N$ as a statistic such that for any data point in Ω_ε, $\widehat{\boldsymbol{\theta}}_N = \theta_{0N} + \widehat{e}_N$. Hence, $G_N(\widehat{\boldsymbol{\theta}}_N) = -\widehat{e}_N \cdot H_{0N}(\theta_{0N}) + r_N(\widehat{e}_N) + G_{0N}(\theta_{0N}) = -r_N(\widehat{e}_N) \cdot \{H_{0N}(\theta_{0N})\}^{-1} \cdot H_{0N}(\theta_{0N}) + r_N(\widehat{e}_N) = 0$.

Because $\rho(H_{0N}(\widehat{\boldsymbol{\theta}}_N)) = \rho(H_{0N}(\theta_{0N} + \widehat{e}_N)) > \varepsilon_0/2 > 0$ and

$$L_N(\theta,\, \widehat{\boldsymbol{\theta}}_N) = L_N(\widehat{\boldsymbol{\theta}}_N,\, \widehat{\boldsymbol{\theta}}_N) + (\theta - \widehat{\boldsymbol{\theta}}_N) \cdot G_N^t(\widehat{\boldsymbol{\theta}}_N) \\ - \frac{1}{2}(\theta - \widehat{\boldsymbol{\theta}}_N) \cdot \frac{H_N(\widehat{\boldsymbol{\theta}}_N) + \{H_N(\widehat{\boldsymbol{\theta}}_N)\}^{-1}}{2} \cdot (\theta - \widehat{\boldsymbol{\theta}}_N)^t + o(||\theta - \widehat{\boldsymbol{\theta}}_N||),$$

there exists a neighborhood of $\widehat{\boldsymbol{\theta}}_N$ such that for any θ in the neighborhood, $L_N(\widehat{\boldsymbol{\theta}}_N, \widehat{\boldsymbol{\theta}}_N) \geq L_N(\theta, \widehat{\boldsymbol{\theta}}_N)$, indicating that for any fixed data point in Ω_ε, $\widehat{\boldsymbol{\theta}}_N$ is a local estimate from $L_N(\theta, \theta')$. Hence,

$$\begin{aligned} &\Pr(||\widehat{\boldsymbol{\theta}}_N - \theta_{0N}|| \leq \varepsilon \text{ and } \widehat{\boldsymbol{\theta}}_N \text{ is a local estimate}) \\ &\quad = \Pr(||\widehat{e}_N|| \leq \varepsilon \text{ and } \widehat{\boldsymbol{\theta}}_N \text{ is a local estimate}) \\ &\quad \geq \Pr(\Omega_\varepsilon) \\ &\quad = \Pr\Big(\{\sup_{||e|| \leq \delta(\varepsilon)} ||G_N(\theta_{0N} + e) - G_{0N}(\theta_{0N} + e)|| \leq \varepsilon_0 \delta(\varepsilon)/2\} \\ &\qquad\quad \cap \{\sup_{||e|| \leq \delta(\varepsilon)} ||H_N(\theta_{0N} + e) - H_{0N}(\theta_{0N})|| \leq \varepsilon_0/4\}\Big). \end{aligned}$$

The regularity condition ensures that both $G_N(\theta_{0N} + e) - G_{0N}(\theta_{0N} + e)$ and $H_N(\theta_{0N} + e) - H_{0N}(\theta_{0N})$ are continuous functions of e uniformly for e, $||e|| \leq \delta(\varepsilon)$, and the law of large numbers ensures that for any fixed e both converge in probability to 0 as N tends to ∞. Letting N tend to ∞ and applying Lemma 1(a), we have $\lim_{N\to\infty} \Pr(||\widehat{\boldsymbol{\theta}}_N - \theta_{0N}|| < \varepsilon$ and $\widehat{\boldsymbol{\theta}}_N$ is a local estimate$) = 1$.

The proof of (c) is similar. We now prove (d). By the definition of global, local, or stationary attractions, there exists $\varepsilon_0 > 0$ such that (10c) and (10d) hold. For $\{\widehat{\boldsymbol{\theta}}_N\}$ with $\{\widehat{\boldsymbol{\theta}}_N - \theta_{0N}\}$ converging in probability to θ_1, using Taylor expansion along with the regularity condition, we have

$$\begin{aligned} 0 &= G_N(\widehat{\boldsymbol{\theta}}_N) = G_N(\theta_{0N}) + (\widehat{\boldsymbol{\theta}}_N - \theta_{0N}) \cdot H_N(\theta_{0N}) + o_p\left(\left\|\widehat{\boldsymbol{\theta}}_N - \theta_{0N}\right\|\right) \\ &= G_N(\theta_{0N}) + (\widehat{\boldsymbol{\theta}}_N - \theta_{0N}) \cdot \{H_{0N}(\theta_{0N}) + o_p(1)\}. \end{aligned}$$

It follows from the regularity condition and (10d) that

$$\begin{aligned}\sqrt{N}(\widehat{\boldsymbol{\theta}}_N - \theta_{0N}) &= -\sqrt{N} \cdot G_N(\theta_{0N}) \cdot \{H_{0N}(\theta_{0N}) + o_p(1)\}^{-1} \\ &= -\sqrt{N} \cdot G_N(\theta_{0N}) \cdot \{H_{0N}(\theta_{0N})\}^{-1} \cdot \{1 + o_p(1)\} \qquad \text{(s.1)}\end{aligned}$$

We denote $B_N(\alpha) = -\sqrt{N} \cdot \{G_N(\alpha) - G_{0N}(\alpha)\}$, $\alpha \in \Theta$. For a fixed $\alpha \in \Theta$, we have

$$\mathrm{E}\{B_N(\alpha)\} = \mathrm{E}[-\sqrt{N} \cdot \{G_N(\alpha) - G_{0N}(\alpha)\}] = 0$$

$$\begin{aligned}\mathrm{Var}\{B_N(\alpha)\} &= \mathrm{Var}[-\sqrt{N} \cdot \{G_N(\alpha) - G_{0N}(\alpha)\}] \\ &= \frac{1}{N}\sum_{i=1}^{N} \mathrm{Var}\{g_i(\theta)\} = \mathrm{E}U_N(\alpha) = U_{0N}(\alpha).\end{aligned}$$

It follows from the central limit theorem that for any fixed α, $B_N(\alpha)$ converges in distribution to $F(b;\alpha)$, the multivariate normal distribution with mean $\mathrm{E}B_N(\alpha) = 0$ and variance matrix $\mathrm{Var}\{B_N(\alpha)\} = U_{0N}(\alpha)$. Further, for any α_1, α_2 in Θ, we have

$$\begin{aligned}&||\mathrm{E}\{(B_N(\alpha_1) - B_N(\alpha_2)\}^t\{B_N(\alpha_1) - B_N(\alpha_2))\}|| \\ &= ||\frac{1}{N}\sum_{i=1}^{N} \mathrm{Var}\{g_i(\alpha_2) - g_i(\alpha_1)\}|| \\ &\leq \frac{1}{N}\sum_{i=1}^{N} ||\mathrm{Var}\{g_i(\alpha_2) - g_i(\alpha_1)\}||.\end{aligned}$$

We denote $h_i^*(\alpha) \overset{\text{def.}}{=} h_i(\alpha) - \mathrm{E}h_i(\alpha) = -[\partial g_i(\alpha)/\partial\alpha - \mathrm{E}\partial g_i(\alpha)/\partial\alpha]$. It follows from the regularity condition and the Taylor expansion that

$$\{g_i(\alpha_2) - g_i(\alpha_1)\} - \mathrm{E}\{g_i(\alpha_2) - g_i(\alpha_1)\} = h_i^*(\alpha_1)\,(\alpha_1 - \alpha_2) + e_i(\alpha_1, \alpha_2),$$

where $||e_i(\alpha_1, \alpha_2)||/||\alpha_2 - \alpha_1||$ converges to 0 uniformly for i and α_1 as $||\alpha_2 - \alpha_1||$ tends to 0. Thus, we have from the above equation,

$$\mathrm{Var}\{g_i(\alpha_2) - g_i(\alpha_1)\} = \mathrm{E}[\{h_i^*(\alpha_1)\,(\alpha_1 - \alpha_2) + e_i(\alpha_1, \alpha_2)\}^t\,\{h_i^*(\alpha_1)\,(\alpha_1 - \alpha_2) + e_i(\alpha_1, \alpha_2)\}]$$

$$
\begin{aligned}
&= \mathrm{E}[(\alpha_1-\alpha_2)^t\ \{h_i^*(\alpha_1)\}^t\ h_i^*(\alpha_1)\ (\alpha_1-\alpha_2)] \\
&\qquad +\mathrm{E}[(\alpha_1-\alpha_2)^t\ \{h_i^*(\alpha_1)\}^t\ e_i(\alpha_1,\alpha_2)] \\
&\qquad +\mathrm{E}\{e_i^t(\alpha_1,\alpha_2)\ h_i^*(\alpha_1)\ (\alpha_1-\alpha_2)\}+\mathrm{E}\{e_i^t(\alpha_1,\alpha_2)\ e_i(\alpha_1,\alpha_2)\}.
\end{aligned}
$$

It follows from the regularity condition that $||\mathrm{Var}[g_i(\alpha_2)-g_i(\alpha_1)]||$ converges to 0 uniformly for i and α_1 as $||\alpha_2-\alpha_1||$ tends to 0. Therefore, for any $\varepsilon>0$, there exists $\delta(\varepsilon)>0$ such that $||\alpha_2-\alpha_1||\le\delta(\varepsilon)$ implies $||E[\{B_N(\alpha_2)-B_N(\alpha_1)\}^t\{B_N(\alpha_2)-B_N(\alpha_1)\}]||\le\varepsilon$. Applying Lemma 1(b) and the regularity condition, we have that $\lim_{N\to\infty}\sup_{\alpha\in\Theta}|\Pr(B_N(\alpha)<b)-F(b;\alpha)|=1$. Thus, $\lim_{N\to\infty}|\Pr(B_N\{\theta_{0N})<b\}-F(b;\theta_{0N})|=1$. Finally, it follows from (s.1) and (10c) that $\sqrt{N}(\widehat{\boldsymbol{\theta}}_N-\theta_{0N})=B_N(\theta_{0N})\cdot[\mathrm{E}H_N(\theta_{0N})]^{-1}\cdot[1+o_p(1)]$ is asymptotically normally distributed with mean 0 and variance matrix (12c).

S.3. Proof of Theorem 2

Note from (8c) and (9a) that $\widehat{\boldsymbol{\theta}}_N$ satisfies $K(\widehat{\boldsymbol{\theta}}_N)=\widehat{\boldsymbol{\theta}}_N$. It follows from Condition (i) that $K(\theta)$ is differentiable at $\widehat{\boldsymbol{\theta}}_N$. Further,

$$
\begin{aligned}
\frac{\partial}{\partial\theta}K(\theta)|_{\theta=\widehat{\theta}_N} &= I+\left[\frac{\partial G_N(\theta)}{\partial\theta}\cdot\{H_N^{(0)}(\theta)\}^{-1}\right]\Big|_{\theta=\widehat{\theta}_N}+\left[G_N(\theta)\cdot\frac{\partial}{\partial\theta}\{H_N^{(0)}(\theta)\}^{-1}\right]\Big|_{\theta=\widehat{\theta}_N} \\
&= I-H_N(\widehat{\boldsymbol{\theta}}_N)\cdot\{H_N^{(0)}(\widehat{\boldsymbol{\theta}}_N)\}^{-1}=H_N^{(1)}(\widehat{\boldsymbol{\theta}}_N)\cdot\{H_N^{(0)}(\widehat{\boldsymbol{\theta}}_N)\}^{-1}.
\end{aligned}
$$

It follows Condition (ii) that the largest absolute eigenvalue of $\partial K(\theta)/\partial\theta|_{\theta=\widehat{\theta}_N}$ is smaller than 1. Applying Ostrowski theorem (Ortega, 1987, ,p. 145) to $K()$ yields the result.

S.4. Additional Examples

Example 7 (GEE with missing covariates) Further consider the GEE with missing covariates. Along the notation in Example 2 of the manuscript; i.e., let $Y_i=(Y_{i1},\ldots,Y_{iM_i})$ be the responses,

$X_i = (X_{i1}, \ldots, X_{iM_i})$ be the covariates (just one-dimensional for notational simplicity) of the ith subject, where M_i is the number of observations from the ith subject. Let $\delta_{ij} = 1$ if X_{ij} is observed and $\delta_{ij} = 0$ otherwise. Assume that the X's are missing at random (MAR) in the sense that

$$\pi(Y_{ij}, \zeta) = P(\delta_{ij} = 1 | X_{ij}, Y_{ij}) = P(\delta_{ij} = 1 | Y_{ij}),$$

which is parameterized by ζ. We model the mean, standard deviation and correlation as in Example 2.

We define an iterative likelihood for $\boldsymbol{\theta} = (\mathbf{b}, \mathbf{a}, \mathbf{c}, \zeta)$ as,

$$L_N(\theta, \theta') \overset{\text{def.}}{=} L_N^{(\text{m})}(b, \theta') + L_N^{(\text{d})}(a, \theta') + L_N^{(\text{r})}(c, \theta'), \tag{s.2}$$

where $\theta = (b, a, c, \zeta)$, $\theta' = (b', a', c', \zeta')$, and

$$L_N^{(\text{m})}(b, \theta') \overset{\text{def.}}{=} -\sum_{i=1}^{N} \sum_{j_1, j_2} \{W_{ij_1 j_2}^{(\text{m})}(\theta') \cdot D_{ij_1}^{(\text{m})}(b) \cdot D_{ij_2}^{(\text{m }) }(b)\} \frac{\delta_{ij_1}}{\pi(Y_{ij_1}, \zeta')} \frac{\delta_{ij_2}}{\pi(Y_{ij_2}, \zeta')}, \tag{s.3a}$$

$$L_N^{(\text{d})}(a, \theta') \overset{\text{def.}}{=} -\sum_{i=1}^{N} \sum_{j} [W_{ij}^{(\text{d})}(\theta') \cdot \{D_{ij}^{(\text{d})}(b', a)\}^2] \frac{\delta_{ij_1}}{\pi(Y_{ij_1}, \zeta')} \frac{\delta_{ij_2}}{\pi(Y_{ij_2}, \zeta')}, \tag{s.3b}$$

$$L_N^{(\text{r})}(c, \theta') \overset{\text{def.}}{=} -\sum_{i=1}^{N} \sum_{j_1, j_2} [W_{ij_1 j_2}^{(\text{r})}(\theta') \cdot \{D_{ij_1 j_2}^{(\text{r})}(b', a', c)\}^2] \frac{\delta_{ij_1}}{\pi(Y_{ij_1}, \zeta')} \frac{\delta_{ij_2}}{\pi(Y_{ij_2}, \zeta')}, \tag{s.3c}$$

$$L_N^{(\pi)}(\zeta, \theta') \overset{\text{def.}}{=} -\sum_{i=1}^{N} \sum_{j_1, j_2} \{W_{ij_1 j_2}^{(\pi)}(\theta') \cdot D_{ij_1}^{(\pi)}(\zeta) \cdot D_{ij_2}^{(\pi)}(\zeta)\}, \tag{s.3d}$$

with

$$D_{ij}^{(\text{m})}(b) \overset{\text{def.}}{=} Y_{ij} - \mathbf{m}_{ij}(X_{ij}; b), \tag{s.4a}$$

$$D_{ij}^{(\text{d})}(a, b') \overset{\text{def.}}{=} \{Y_{ij} - \mathbf{m}_{ij}(X_{ij}; b')\}^2 - \mathbf{d}_{ij}^2(b', a), \tag{s.4b}$$

$$D_{ij_1 j_2}^{(\text{r})}(c, a', b') = \frac{Y_{ij_1} - \mathbf{m}_{ij_1}(X_{ij}; b')}{\mathbf{d}_{ij_1}(bi', a')} \frac{Y_{ij_2} - \mathbf{m}_{ij_2}(X_{ij}; b')}{\mathbf{d}_{ij_2}(X_{ij}; b', a')} - \mathbf{r}_{ij_1 j_2}(b', a', c) \tag{s.4c}$$

$$D_{ij}^{\pi}(\zeta) = \delta_{ij} - \pi(Y_{ij}, \zeta). \tag{s.4d}$$

Example 8 (Unweighted estimator for big data subsampling) Let $\{(X_i, Y_i)\}_{i=1}^N$ be the independent full data of size N from the joint distribution of (X, Y), where Y is the response variable and X is the covariate variable. Let the joint density of (X, Y) be $f_{XY}(x, y; \theta) = f_{Y|X}(y|x; \beta) f_X(x; \alpha)$, where $\theta = (\beta, \alpha)$, $f_{Y|X}(y|x; \beta)$ is the conditional density of Y given X, and $f_X(x; \alpha)$ is the density of X. With big data where N is super large, using the full data to estimate θ is computationally expensive, so a popular practical solution is to select a smaller subsample to perform calculation (e.g. Avron et al., 2010; Ma et al., 2015; Mahoney, 2011; Meng et al., 2014; Zhang et al., 2020). For estimation efficiency, nonuniform sampling probabilities are recommended where the sampling probabilities depend on the data. For example, optimal subsampling assigns larger probabilities to more informative data points (Wang et al., 2018). Let $\pi(X_i, Y_i)$ be the sampling probability such that $\pi_n(X_i, Y_i) = \Pr(\delta_i = 1|X_i, Y_i)$, $i = 1, ..., N$, where n is the expected subsample size so that $E\{\pi_n(X_i, Y_i)\} = n$ and δ_i is the indicator variable signifying if (X_i, Y_i) is included in the subsample ($\delta_i = 1$ if the i-th data point is selected in the subsample and $\delta_i = 0$ otherwise). Although uniform sampling is often used, there is increasing interest in optimal subsampling where a more inforrmative data point is given a larger value of $\pi(X_i, Y_i)$ (Mahoney, 2011; Zhang et al., 2020). For a selected subsample, a commonly used estimator is the inverse probability weighted estimator, the maximizer of

$$\sum_{i=1}^{N} \frac{\delta_i \log f_{XY}(X_i, Y_i; \theta)}{\pi(X_i, Y_i)}. \tag{s.5}$$

However, the estimator $\hat{\theta}_W$ gives smaller weights to more informative data points, so it is not efficient. To solve this issue, methods have been proposed to correct the bias in the naive unweighted estimator (Fithian and Hastie, 2014; Scott and Wild, 1986; Wang, 2019), and Wang (2019) has proved that the unweighted estimator with bias correction has a higher estimation efficiency. However, the aforementioned investigations exclusively focused on the logistic regression because the bias correction terms depends on the special structure of the logistic regression. A general approach to avoid the inefficient inverse probability weighting is not available for optimal subsampling.

The proposed iterative likelihood framework gives general solutions beyond logistic regression for subsampled data. From Bayes' theorem, the density of (X, Y) for the sampled observation with $\delta = 1$ is

$$f_{XY}(x, y|\delta = 1; \theta) = \frac{f_{Y|X}(y|x; \beta) f_X(x; \alpha) \pi_n(x, y)}{\int \bar{\pi}(x; \beta) f_X(x; \alpha) dx} \tag{s.6}$$

where

$$\bar{\pi}_n(x; \beta) = \int f_{Y|X}(y|x; \theta) \pi_n(x, y) dy \tag{s.7}$$

often have closed form expression in optimal subsampling.

Letting $\theta = (\beta, \alpha)$ and $\theta' = (\beta', \alpha')$, we define an iterative likelihood as

$$L_N(\theta, \theta') = \sum_{i=1}^{N} \delta_i l_i(\theta, \theta'), \tag{s.8}$$

where

$$l_i(\theta, \theta') = \log f_{Y|X}(y|x; \beta) + \log f_X(x; \alpha') - \log \int \bar{\pi}_n(x; \beta) f_X(x; \alpha') dx. \tag{s.9}$$

The above iterative likelihood procedure is innovative in multiple aspects. 1) It gives a general solution to avoid the inverse probability weighting. In addition, our theoretical results in the paper apply, assuming n and N goes to infinity. Note that in subsampling for a given expected subsample size n, the density of (X, Y) given $\delta = 1$ is a sequence that changes with n and N, so the standard i.i.d. argumentation for MLE does not directly applies. 2) Our theoretical results are unconditional, and it is about the true parameter. This is different from existing results for optimal subsampling estimators where the distributional results are often conditional on the observed data, and the theoretical properties are about approximating the full data estimator instead of estimating the true parameter, e.g., Ai et al. (2020); Keret and Gorfine (2020); Wang et al. (2018); Yao and Wang (2019); Yu et al. (2020); Zhang and Wang (2021); Zuo et al. (2021), among others. 3) We believe the resulting estimator has the highest estimation efficiency among regular asymptotically unbiased estimators. However, a rigorous proof needs further investigations.

Of course for the additional estimation efficiency, the iterative likelihood has to pay a price in regression problems. If β is the only parameter of interest, then the weighted estimator can be obtain thorough maximizing

$$\sum_{i=1}^{N} \frac{\delta_i \log f_{Y|X}(y|x;\beta)}{\pi(X_i, Y_i)}, \tag{s.10}$$

without estimating α. We point out this because we do not want oversell iterative likelihood. Every method has its advantages and disadvantages. From the above example, we see that the iterative likelihood provides a general solution to an important problem by looking at it from a broader view. This is definitely one of the advantages of iterative likelihood. Our paper is the first paper about iterative likelihood and we do not expect it to solve all the problems. We hope the paper can be a start in this direction.

References

Ai, M., Yu, J., Zhang, H., and Wang, H. (2020), "Optimal Subsampling Algorithms for Big Data Generalized Linear Models," *Statistica Sinica*, DOI:10.5705/ss.202018.0439.

Avron, H., Maymounkov, P., and Toledo, S. (2010), "Blendenpik: Supercharging LAPACK's least-squares solver," *SIAM Journal on Scientific Computing*, 32, 1217–1236.

Fithian, W. and Hastie, T. (2014), "Local case-control sampling: Efficient subsampling in imbalanced data sets," *Annals of statistics*, 42, 1693.

Keret, N. and Gorfine, M. (2020), "Optimal Cox Regression Subsampling Procedure with Rare Events," *arXiv preprint arXiv:2012.02122*.

Ma, P., Mahoney, M., and Yu, B. (2015), "A Statistical Perspective on Algorithmic Leveraging," *Journal of Machine Learning Research*, 16, 861–911.

Mahoney, M. W. (2011), “Randomized algorithms for matrices and data,” *Foundations and Trends® in Machine Learning*, 3, 123–224.

Meng, X., Saunders, M., and Mahoney, M. (2014), “LSRN: A parallel iterative solver for strongly over- or under- determined systems,” *SIAM Journal on Scientific Computing*, 36, C95–C118.

Ortega, J. M. (1987), *Numerical Analysis: A Second Course*, vol. 3, Society for Industrial and Applied Mathematics.

Scott, A. J. and Wild, C. J. (1986), “Fitting Logistic Models Under Case-Control or Choice Based Sampling,” *Journal of the Royal Statistical Society. Series B*, 48, 170–182.

Wang, H. (2019), “More Efficient Estimation for Logistic Regression with Optimal Subsamples,” *Journal of Machine Learning Research*, 20, 1–59.

Wang, H., Zhu, R., and Ma, P. (2018), “Optimal subsampling for large sample logistic regression,” *Journal of the American Statistical Association*, 113, 829–844.

Yao, Y. and Wang, H. (2019), “Optimal subsampling for softmax regression,” *Statistical Papers*, 60, 235–249.

Yu, J., Wang, H., Ai, M., and Zhang, H. (2020), “Optimal Distributed Subsampling for Maximum Quasi-Likelihood Estimators with Massive Data,” *Journal of the American Statistical Association*, https://doi.org/10.1080/01621459.2020.1773832.

Zhang, H. and Wang, H. (2021), “Distributed subdata selection for big data via sampling-based approach,” *Computational Statistics & Data Analysis*, `https://doi.org/10.1016/j.csda.2020.107072`.

Zhang, T., Ning, Y., and Ruppert, D. (2020), “Optimal Sampling for Generalized Linear Models under Measurement Constraints,” *Journal of Computational and Graphical Statistics*, to appear, now published online.

Zuo, L., Zhang, H., Wang, H., and Liu, L. (2021), “Sampling-Based Estimation for Massive Survival Data with Additive Hazards Model,” *Statistics in Medicine*, 40, DOI:10.1002/sim.8783.