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 $\varepsilon > 0$, $\lim_{N\to\infty} \Pr(\sup_{\alpha\in\Gamma} ||B_N(\alpha)|| \leq \varepsilon) = 1$.

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

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

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,

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

For any $\varepsilon > 0$, we have

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

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 $\lim_{N\to\infty}\Pr(||\widehat{\boldsymbol{\theta}}_N - \theta_{0N}|| \leq \varepsilon) = 1$, 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 $\Omega_\varepsilon = \{ \sup_{||e|| \leq \delta(\varepsilon)} ||G_N(\theta_{0N} + e) - G_{0N}(\theta_{0N} + e)|| \leq \varepsilon_0 \delta(\varepsilon)/2\} \cap \{ \sup_{||e|| \leq \delta(\varepsilon)} ||H_N(\theta_{0N} + e) - H_{0N}(\theta_{0N} + e)|| \leq \varepsilon_0/4\}$. 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

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

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

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

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

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

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

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,

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

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

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

$$L_N^{(\text{m})}(b, \theta') \overset{\text{def.}}{=} -\sum_{i=1}^{N} \sum_{j_1, j_2} \{W_{ij_1j_2}^{(\text{m})}(\theta') \cdot D_{ij_1}^{(\text{m})}(b) \cdot D_{ij_2}^{(\text{m}\ )}(b)\} \frac{\delta_{ij_1}}{\pi(Y_{ij_1}, \zeta')} \frac{\delta_{ij_2}}{\pi(Y_{ij_2}, \zeta')}, \tag{s.3a}$$

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

$$L_N^{(\text{r})}(c, \theta') \overset{\text{def.}}{=} -\sum_{i=1}^{N} \sum_{j_1, j_2} [W_{ij_1j_2}^{(\text{r})}(\theta') \cdot \{D_{ij_1j_2}^{(\text{r})}(b', a', c)\}^2] \frac{\delta_{ij_1}}{\pi(Y_{ij_1}, \zeta')} \frac{\delta_{ij_2}}{\pi(Y_{ij_2}, \zeta')}, \tag{s.3c}$$

$$L_N^{(\pi)}(\zeta, \theta') \overset{\text{def.}}{=} -\sum_{i=1}^{N} \sum_{j_1, j_2} \{W_{ij_1j_2}^{(\pi)}(\theta') \cdot D_{ij_1}^{(\pi)}(\zeta) \cdot D_{ij_2}^{(\pi)}(\zeta)\}, \tag{s.3d}$$

with

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

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

$$D_{ij_1j_2}^{(\text{r})}(c, a', b') = \frac{Y_{ij_1} - \mathbf{m}_{ij_1}(X_{ij}; b')}{\mathbf{d}_{ij_1}(bi', a')} \frac{Y_{ij_2} - \mathbf{m}_{ij_2}(X_{ij}; b')}{\mathbf{d}_{ij_2}(X_{ij}; b', a')} - \mathbf{r}_{ij_1j_2}(b', a', c) \tag{s.4c}$$

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

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.