EFFICIENCY OF PROFILE LIKELIHOOD IN SEMI-PARAMETRIC MODELS

Yuichi Hirose

School of Mathematics, Statistics and Computer Science,
Victoria University of Wellington, New Zealand

February 12, 2008

Profile likelihood is a popular method of estimation in the presence of a nuisance parameter. It is especially useful for estimation in semi-parametric models, since the method reduces the infinite-dimensional estimation problem to a finite-dimensional one. In this paper we investigate the efficiency of a semi-parametric maximum likelihood estimator based on the profile likelihood. By introducing a new parameterization, we improve on the seminal work of Murphy and van der Vaart (2000) in two ways: we prove the no bias condition in a general semi-parametric model context, and deal with the direct quadratic expansion of the profile likelihood rather than an approximate one. To illustrate the method, an application to two-phase, outcome-dependent sampling design is given.

Key words: Semi-parametric model; Profile likelihood; Two-phase, outcome-dependent sampling; Efficiency; M-estimator; Maximum likelihood estimator; Efficient score; Efficient information bound.

E-mail: Yuichi.Hirose@mcs.vuw.ac.nz
Telephone: 64 - 4 - 4635341 extn 5275
Fax: 64 - 4 - 4635045

1 Introduction

The efficient score function can be defined as the score function for β minus its orthogonal projection onto the nuisance tangent space (cf. Bickel, Klaassen, Ritov and Wellner (1993)). An alternative characterization, due to Newey (1994), is to describe the efficient score in terms of the derivative with respect to β of a "population profile log-likelihood", which is a population version of the ordinary sample profile log-likelihood. Since "profiling out" is a familiar method of dealing with nuisance parameters in likelihood calculations, the Newey characterization is perhaps a more natural method than projection.

Suppose we consider a semiparametric model of the form

$$\mathcal{P} = \{p(x;\beta,\eta) : \ \beta \in \Theta_\beta \subset \mathbb{R}^m, \, \eta \in \Theta_\eta\}$$

where β is the m-dimensional parameter of interest, and η is a nuisance parameter, which may be infinite-dimensional. Let (β_0, η_0) be the true value of (β, η). We assume Θ_β is a compact set containing an open neighborhood of β_0 in $\mathbb{R}^m$, and Θ_η is a convex set containing η_0 in a Banach space $\mathcal{B}$.

We also assume that, for each $\beta \in \Theta_\beta$, the expected log-likelihood $E_{\beta_0,\eta_0} \log p(X;\beta,\eta)$ is uniquely maximized with respect to $\eta \in \Theta_\eta$. For each β, define

$$\hat\eta(\beta) = \text{argmax}_{\eta \in \Theta_\eta} E_{\beta_0,\eta_0} \log p(X;\beta,\eta), \tag{1}$$

then we have $\hat\eta(\beta_0) = \eta_0$ and the derivative

$$\dot\ell^*_\beta(x,\beta_0) = \left.\frac{\partial}{\partial\beta}\right|_{\beta=\beta_0} \log p(x;\beta,\hat\eta(\beta))$$

is the efficient score function (cf. Newey (1994)).

On the other hand, let

$$\hat\eta_n(\beta) = \text{argmax}_{\eta\in\Theta_\eta} \sum_{i=1}^n \log p(X_i;\beta,\eta). \tag{2}$$

The *profile log-likelihood function* for β is the log-likelihood

$$\ell_n(\beta, \hat\eta_n(\beta)) = \sum_{i=1}^n \log p(X_i;\beta,\hat\eta_n(\beta))$$

treated as a function of β only. The solution to the profile likelihood estimating equation $\frac{\partial}{\partial\beta}\ell_n(\hat\beta_n, \hat\eta_n(\hat\beta_n)) = 0$ gives the MLE $\hat\beta_n$.

The purpose of this paper is to investigate the efficiency of the semi-parametric maximum likelihood estimator based on the profile likelihood. The difficulty in the proof of the efficiency of the profile-likelihood estimator is that the corresponding estimating equation cannot be treated using standard M-estimator theory since the estimating functions depend implicitly on the sample size. Murphy and van der Vaart (2000) proved this efficiency by introducing the approximate least favorable submodel to express the upper and lower bounds for the profile log-likelihood. Since these two bounds have the same expression for the asymptotic expansion, so does the one for the profile log-likelihood. This method cleverly avoided the implicit dependence on the sample size n.

The main idea of this paper is the introduction of the new function $\hat{\eta}(\beta, F)$ with an additional parameter F such that the estimating equations based on the profile likelihood and the least favorable submodel can be expressed as

$$\sum_{i=1}^{n} \frac{\partial}{\partial \beta} \log p(X_i; \beta, \hat{\eta}(\beta, F_n)) = 0$$

and

$$\sum_{i=1}^{n} \frac{\partial}{\partial \beta} \log p(X_i; \beta, \hat{\eta}(\beta, F_0)) = 0,$$

respectively, where F_n is the empirical cumulative distribution function (cdf) and F_0 is the cdf for the true distribution. This gives an estimating function which is an explicit function of sample size n, through F. Then, we show that the solutions $\hat{\beta}_n$ to the above estimating equations are asymptotically equivalent. Since the estimator based on the least favorable submodel is efficient, this demonstrates that the estimator based on the profile-likelihood is also efficient. Moreover, the no bias condition, which is one of the assumptions in Murphy and van der Vaart (2000), follows naturally from our approach.

An outline of this paper is as follows. Section 2 presents the main result, in which, we prove the efficiency of the estimator based on the profile likelihood by introducing the empirical cdf as a parameter. In section 3, two-phase, outcome-dependent sampling design is used as an example. This paper is motivated by the question "Is the method of Scott and Wild (1997, 2001) efficient?" and we give the answer to the question in the discussion.

2 Main result

We denote the set of cdf's on the sample space $\mathcal{X}$ by $\mathcal{F}$. The set $\mathcal{F}$ is convex. Let $F_n(x)$ be the empirical cdf and $F_0(x)$ the cdf for the density $p(x; \beta_0, \eta_0)$.

For a map $\hat{\eta} : \Theta_\beta \times \mathcal{F} \to \Theta_\eta$, define a model (called the *induced model*) with density

$$p'(x; \beta, F) = p(x; \beta, \hat{\eta}(\beta, F)), \ \beta \in \Theta_\beta, \ F \in \mathcal{F}.$$

The score function in the induced model is denoted by

$$\phi(x, \beta, F) = \frac{\partial}{\partial \beta} \log p'(x; \beta, F). \tag{3}$$

(Condition $(R1)$ or $(R1)^*$ in SECTION 2.1.2 assume the differentiability of the function $p'(x; \beta, F)$ with respect to β.)

We assume that

(R0) $\hat{\eta}$ satisfies $\hat{\eta}(\beta_0, F_0) = \eta_0$ and the function

$$\dot{\ell}^*_\beta(x, \beta_0) = \phi(x, \beta_0, F_0)$$

is the efficient score function.

We present the main results of this paper.

Theorem 1.*[The main theorem] Suppose sets of assumptions* $\{(R0), (R1), (R2), (R3)\}$ *or* $\{(R0), (R1)^*, (R2), (R3)\}$ *given in* Section 2.1.2*, then, for any random sequence* $\tilde{\beta}_n \xrightarrow{P} \beta_0$ *and the empirical cdf* F_n*, we have*

$$\sqrt{n} E_{\beta_0,\eta_0} \phi(X, \beta_0, F_n) \;=\; o_P(1) \quad (\textit{No bias contion}) \tag{4}$$

and

$$\begin{aligned} \sum_{i=1}^n \log p'(X_i; \tilde{\beta}_n, F_n) \;&=\; \sum_{i=1}^n \log p'(X_i; \beta_0, F_n) + (\tilde{\beta}_n - \beta_0)^T \sum_{i=1}^n \phi(X_i, \beta_0, F_0) \\ &\quad + \frac{1}{2} n (\tilde{\beta}_n - \beta_0)^T I^*_\beta (\tilde{\beta}_n - \beta_0) + o_{P_{\beta_0,\eta_0}} (\sqrt{n} \|\tilde{\beta}_n - \beta_0\| + 1)^2 \end{aligned} \tag{5}$$

where $I^*_\beta = E_{\beta_0,\eta_0}(\dot{\ell}^*_\beta \dot{\ell}^{*T}_\beta)$ *is the efficient information matrix.* □

The proof is given in the next section.

By Corollary 1.1 in Murphy and van der Vaart (2000), we have the following result.

Corollary 1. *A consistent solution* $\hat{\beta}_n$ *to the estimating equation*

$$\sum_{i=1}^n \phi(X_i, \hat{\beta}_n, F_n) = 0 \tag{6}$$

is an asymptotically linear estimator for β_0 *with the efficient influence function*

$$\tilde{\ell}^*_\beta(x, \beta_0) = (I^*_\beta)^{-1} \dot{\ell}^*_\beta(x, \beta_0)$$

so that

$$\sqrt{n}(\hat{\beta}_n - \beta_0) = \frac{1}{\sqrt{n}} \sum_{i=1}^n \tilde{\ell}^*_\beta(X_i, \beta_0) + o_P(1)$$

and

$$\sqrt{n}(\hat{\beta}_n - \beta_0) \xrightarrow{d} N\left(0, (I^*_\beta)^{-1}\right).$$

□

This demonstrates that the profile likelihood MLE $\hat{\beta}_n$ is efficient.

2.1 Assumptions and proof

2.1.1 Path-wise differentiability

A *path* in a convex subset $\mathcal{C}$ of a Banach space $\mathcal{B}$ is a continuously differentiable map $\eta(t) : \Theta_t \to \mathcal{C}$ where Θ_t is a closed interval in $\mathbb{R}$. The derivative of a path $\eta(t)$ is denoted by $\dot{\eta}(t)$. A map

$f(\eta) : \mathcal{C} \to \mathbb{R}^m$ is *path-wise differentiable* with respect to η if there exists a bounded linear operator $d_\eta f(\eta)$, called the *derivative* of $f(\eta)$, such that, for each path $\eta(t)$ and $t \in \Theta_t$,

$$\frac{\partial}{\partial t} f(\eta(t)) = d_\eta f(\eta(t))\dot{\eta}(t).$$

A norm of the derivative $d_\eta f(\eta)$ at η_0 is defined by

$$\|d_\eta f(\eta_0)\| = \sup\left\{ \frac{\|d_\eta f(\eta_0)\dot{\eta}(0)\|}{\|\dot{\eta}(0)\|} : \ \eta(t) \text{ path with } \eta(0) = \eta_0 \text{ and } \dot{\eta}(0) \neq 0 \right\}.$$

A map $f(\eta)$ is *continuously path-wise differentiable* with respect to η if the derivative $d_\eta f(\eta)$ is continuous function of η.

Derivative with convex combination: Let η_0 and η be two points in the convex set $\mathcal{C}$. Then the map $t \to \eta_0 + t(\eta - \eta_0)$, $t \in [0,1]$ is a path in $\mathcal{C}$. If a function $f : \mathcal{C} \to \mathbb{R}$ is path-wise differentiable, then, for all $t \in [0,1]$,

$$\frac{\partial}{\partial t} f(\eta_0 + t(\eta - \eta_0)) = d_\eta f(\eta_0 + t(\eta - \eta_0))(\eta - \eta_0).$$

2.1.2 Assumptions

On the set of cdf functions $\mathcal{F}$, we use the sup-norm, i.e., for $F, F_0 \in \mathcal{F}$,

$$\|F - F_0\| = \sup_x |F(x) - F_0(x)|.$$

For $\rho > 0$, let

$$\mathcal{C}_\rho = \{F \in \mathcal{F} : \|F - F_0\| < \rho\}.$$

We assume that:

(R1) The $\sqrt{n}$-consistency of F_n, $\sqrt{n}\|F_n - F_0\| = O_P(1)$, and, for each $(\beta, F) \in \Theta_\beta \times \mathcal{F}$, the log-likelihood function $\log p'(x; \beta, F)$ is twice continuously differentiable with respect to β and continuously path-wise differentiable with respect to F for all x.

(R1)* The empirical process F_n satisfies $n^{1/4}\|F_n - F_0\| = o_P(1)$ and for each $(\beta, F) \in \Theta_\beta \times \mathcal{F}$, the log-likelihood function $\log p'(x; \beta, F)$ is twice continuously differentiable with respect to β and twice continuously path-wise differentiable with respect to F for all x.

(Derivatives are denoted by $\phi(x, \beta, F) = \frac{\partial}{\partial \beta} \log p'(x; \beta, F)$, $\frac{\partial}{\partial \beta}\phi(x, \beta, F)$, $d_F\phi(x, \beta, F)$, and $d_F^2\phi(x, \beta, F)$.)

(R2) The efficient information matrix $I_\beta^* = E_{\beta_0,\eta_0}\dot{\ell}_\beta^*\dot{\ell}_\beta^{*T} = E_{\beta_0,\eta_0}\phi\phi^T(X, \beta_0, F_0)$ is invertible.

(R3) There exist a $\rho > 0$ and a neighborhood Θ_β of β_0 such that the class of functions $\{\phi(x, \beta, F) : \ (\beta, F) \in \Theta_\beta \times \mathcal{C}_\rho\}$ is P_{β_0,η_0}-Donsker with square integrable envelope function, and such that the class of functions $\{\frac{\partial}{\partial \beta}\phi(x, \beta, F) : \ (\beta, F) \in \Theta_\beta \times \mathcal{C}_\rho\}$ is P_{β_0,η_0}-Glivenko-Cantelli with integrable envelope function.

2.1.3 Proof

Suppose $\{(R0), (R1), (R2), (R3)\}$ or $\{(R0), (R1)^*, (R2), (R3)\}$.

First, we prove Equation (4). Since (i) the induced model $p'(x;\beta,F)$ is a probability model, (ii) the range of the score operator $\dot{\ell}_F(X,\beta_0,F_0) = d_F \log p'(x;\beta_0,F_0) = d_F \log p(x;\beta_0,\hat{\eta}(\beta_0,F_0))$ for F is in the nuisance tangent space (the tangent space for η), and (iii) the function $\phi(X,\beta_0,F_0)$ is the efficient score function, we have

$$E_{\beta_0,\eta_0} d_F\phi(X,\beta_0,F_0) = -E_{\beta_0,\eta_0}\phi\dot{\ell}_F(X,\beta_0,F_0) = 0 \ \text{ (the zero operator).} \tag{7}$$

For F_n and F_0 in $\mathcal{F}$, consider a path $F_n^*(t) = F_0 + t(F_n - F_0)$, $t \in [0,1]$. Then $F_n^*(0) = F_0$ and $F_n^*(1) = F_n$. Under assumptions $\sqrt{n}\|F_n - F_0\| = O_P(1)$ (condition (R1)) or $n^{1/4}\|F_n - F_0\| = o_P(1)$ (condition (R1)*), we have that $\sup_{t\in[0,1]}|F_n^*(t) - F_0| = o_P(1)$.

Suppose condition (R1). By the mean value theorem for vector valued function (cf. Hall and Newell (1979)),

$$\begin{aligned}
& \|\sqrt{n}E_{\beta_0,\eta_0}\phi(X,\beta_0,F_n)\| \\
= \ & \|\sqrt{n}E_{\beta_0,\eta_0}\phi(X,\beta_0,F_n^*(1)) - \sqrt{n}E_{\beta_0,\eta_0}\phi(X,\beta_0,F_n^*(0))\| \\
\le \ & \sup_{t\in[0,1]} \|E_{\beta_0,\eta_0} d_F\phi(X,\beta_0,F_n^*(t))\|\sqrt{n}\|F_n - F_0\| \\
= \ & \|E_{\beta_0,\eta_0} d_F\phi(X,\beta_0,F_0) + o_p(1)\|\sqrt{n}\|F_n - F_0\| \ \text{ (since } \sup_{t\in[0,1]}|F_n^*(t) - F_0| = o_P(1)) \\
= \ & o_p(1)\sqrt{n}\|F_n - F_0\| \ \text{ (by Equation (7))} \\
= \ & o_P(1) \ \text{ (since } \sqrt{n}\|F_n - F_0\| = O_P(1)).
\end{aligned}$$

Alternatively, suppose condition (R1)*. We modify the proof of the mean value theorem for vector valued function in Hall and Newell (1979). Let $f_n(t) = \sqrt{n}E_{\beta_0,\eta_0}\phi(X,\beta_0,F_n^*(t))$ and

$$\Phi_n(t) = \frac{\langle f_n(1) - f_n(0), \ f_n(t) - f_n(0)\rangle}{\|f_n(1) - f_n(0)\|}$$

where $\langle u, v\rangle = u^T v$ for $u, v \in \mathbb{R}^m$. Then

$$\begin{aligned}
& \|\sqrt{n}E_{\beta_0,\eta_0}\phi(X,\beta_0,F_n)\| \\
= \ & \|\sqrt{n}E_{\beta_0,\eta_0}\phi(X,\beta_0,F_n^*(1)) - \sqrt{n}E_{\beta_0,\eta_0}\phi(X,\beta_0,F_n^*(0))\| \\
= \ & \Phi_n(1) - \Phi_n(0) \\
= \ & \frac{\partial}{\partial t}\Phi_n(0) + \frac{\partial^2}{\partial t^2}\Phi_n(t_n^*) \ \text{ (for some } t_n^* \in [0,1]\text{, by Taylor's expansion)} \\
= \ & \frac{\langle f_n(1) - f_n(0), \ \frac{\partial}{\partial t}f_n(0) + \frac{\partial^2}{\partial t^2}f_n(t_n^*)\rangle}{\|f_n(1) - f_n(0)\|} \\
\le \ & \sup_{t\in[0,1]}\left\|\sqrt{n}E_{\beta_0,\eta_0}\frac{\partial}{\partial t}\phi(X,\beta_0,F_n^*(0)) + \sqrt{n}E_{\beta_0,\eta_0}\frac{\partial^2}{\partial t^2}\phi(X,\beta_0,F_n^*(t))\right\| \\
& \text{(by the Cauchy-Schwrz inequality)} \\
= \ & \sup_{t\in[0,1]}\|E_{\beta_0,\eta_0}d_F\phi(X,\beta_0,F_0)\sqrt{n}(F_n - F_0) + E_{\beta_0,\eta_0}d_F^2\phi(X,\beta_0,F_n^*(t))\sqrt{n}(F_n - F_0)^2\|
\end{aligned}$$

$$\begin{aligned} & \text{(by the definition of path-wise differentiability)} \\ = & \sup_{t\in[0,1]} \|E_{\beta_0,\eta_0} d_F^2 \phi(X,\beta_0,F_n^*(t))\sqrt{n}(F_n - F_0)^2\| \;\; \text{(by Equation (7))} \\ \leq & \|E_{\beta_0,\eta_0} d_F^2 \phi(X,\beta_0,F_0) + o_p(1)\|\sqrt{n}\|F_n - F_0\|^2 \\ = & o_P(1) \;\; \text{(since } \sqrt{n}\|F_n - F_0\|^2 = o_P(1)). \end{aligned}$$

Thus under assumptions (R1) or (R1)*, we have proved Equation (4).

The rest of the proof is similar to the one for Murphy and van der Vaart (2000).

Since the functions $\phi(x,\beta,F)$ and $\frac{\partial}{\partial\beta}\phi(x,\beta,F)$ are continuous at (β_0,F_0), and they are dominated by the square integrable function and the integrable function, respectively, by dominated convergence theorem, for every $(\beta_n^*,F_n^*) \xrightarrow{P} (\beta_0,F_0)$, we have

$$E_{\beta_0,\eta_0}\|\phi(X,\beta_n^*,F_n^*) - \phi(X,\beta_0,F_0)\|^2 \xrightarrow{P} 0.$$

and

$$E_{\beta_0,\eta_0}\|\frac{\partial}{\partial\beta}\phi(X,\beta_n^*,F_n^*) - \frac{\partial}{\partial\beta}\phi(X,\beta_0,F_0)\| \xrightarrow{P} 0.$$

Since $p'(x;\beta,F)$ is a probability model,

$$E_{\beta_0,\eta_0}\frac{\partial}{\partial\beta}\phi(X,\beta_0,F_0) = -E_{\beta_0,\eta_0}\phi\phi^T(X,\beta_0,F_0).$$

Together with condition (R3), this implies that, for every random sequence $(\beta_n^*,F_n^*) \xrightarrow{P} (\beta_0,F_0)$,

$$\frac{1}{\sqrt{n}}\sum_{i=1}^n\{\phi(X_i,\beta_n^*,F_n^*) - \phi(X_i,\beta_0,F_0)\} = \sqrt{n}E_{\beta_0,\eta_0}\{\phi(X,\beta_n^*,F_n^*) - \phi(X,\beta_0,F_0)\} + o_P(1), \tag{8}$$

$$\frac{1}{\sqrt{n}}\sum_{i=1}^n\frac{\partial}{\partial\beta}\phi(X_i,\beta_n^*,F_n^*) \xrightarrow{P} -E_{\beta_0,\eta_0}\phi\phi^T(X,\beta_0,F_0). \tag{9}$$

By combining Equation (4) and (8), we get

$$\frac{1}{\sqrt{n}}\sum_{i=1}^n\phi(X_i,\beta_0,F_n) = \frac{1}{\sqrt{n}}\sum_{i=1}^n\phi(X_i,\beta_0,F_0) + o_P(1). \tag{10}$$

Finally, by Taylor's expansion with respect to β, for some β_n^* with $\|\beta_n^* - \beta_0\| \leq \|\tilde{\beta}_n - \beta_0\|$,

$$\begin{aligned} & \sum_{i=1}^n \log p'(X_i;\tilde{\beta}_n,F_n) - \sum_{i=1}^n \log p(X_i;\beta_0,F_n) \\ = & \sqrt{n}(\tilde{\beta}_n - \beta_0)^T\frac{1}{\sqrt{n}}\sum_{i=1}^n\phi(X_i,\beta_0,F_n) + \frac{1}{2}n(\tilde{\beta}_n - \beta_0)^T\frac{1}{n}\sum_{i=1}^n\frac{\partial}{\partial\beta}\phi(X_i,\beta_n^*,F_n)(\tilde{\beta}_n - \beta_0) \\ = & \sqrt{n}(\tilde{\beta}_n - \beta_0)^T\left\{\frac{1}{\sqrt{n}}\sum_{i=1}^n\phi(X_i,\beta_0,F_0) + o_P(1)\right\} \;\; \text{(by Equation (10))} \\ & +\frac{1}{2}n(\tilde{\beta}_n - \beta_0)^T\left\{-E_{\beta_0,\eta_0}\phi\phi^T(X,\beta_0,F_0) + o_P(1)\right\}(\tilde{\beta}_n - \beta_0) \;\; \text{(by Equation (9))} \\ = & \sqrt{n}(\tilde{\beta}_n - \beta_0)^T\frac{1}{\sqrt{n}}\sum_{i=1}^n\phi(X_i,\beta_0,F_0) - \frac{1}{2}n(\tilde{\beta}_n - \beta_0)^T E_{\beta_0,\eta_0}(\phi\phi^T)(\tilde{\beta}_n - \beta_0) \\ & +o_P(\sqrt{n}\|\tilde{\beta}_n - \beta_0)\| + 1)^2. \end{aligned}$$

This proves Equation (5).

2.2 Useful theorem to identify the efficient score function

To verify Condition ($R0$), the following theorem may be useful. This is a modification of the proof in Breslow, McNeney and Wellner (2000) which originally adapted from Newey (1994).

THEOREM 2. *Suppose $\eta(t)$ is an arbitrary path such that $\eta(0) = \eta_0$ and let $\alpha(t) = \eta(t) - \eta_0$. If*

$$\hat{\eta}(\beta_0, F_0) = \eta_0 \tag{11}$$

and, for each $\beta \in \Theta_\beta$,

$$\left.\frac{\partial}{\partial t}\right|_{t=0} E_{\beta_0,\eta_0}\left[\log p(x; \beta, \hat{\eta}(\beta, F_0) + \alpha(t))\right] = 0, \tag{12}$$

then the function $\phi(x, \beta_0, F_0) = \frac{\partial}{\partial\beta}|_{\beta=\beta_0} \log p(x; \beta, \hat{\eta}(\beta, F_0))$ is the efficient score function.

Proof. Condition (12) implies that

$$\begin{aligned} 0 &= \left.\frac{\partial}{\partial\beta}\right|_{\beta=\beta_0} \left.\frac{\partial}{\partial t}\right|_{t=0} E_{\beta_0,\eta_0}\left[\log p(x; \beta, \hat{\eta}(\beta, F_0) + \alpha(t))\right] \\ &= \left.\frac{\partial}{\partial t}\right|_{t=0} E_{\beta_0,\eta_0}\left[\left.\frac{\partial}{\partial\beta}\right|_{\beta=\beta_0} \log p(x; \beta, \hat{\eta}(\beta, F_0) + \alpha(t))\right]. \end{aligned} \tag{13}$$

By differentiating the identity

$$\int \left(\frac{\partial}{\partial\beta} \log p(x; \beta, \hat{\eta}(\beta, F_0) + \alpha(t))\right) p(x; \beta, \hat{\eta}(\beta, F_0) + \alpha(t))dx = 0$$

with respect to t at $t = 0$ and $\beta = \beta_0$, we get

$$\begin{aligned} 0 &= \left.\frac{\partial}{\partial t}\right|_{t=0,\beta=\beta_0} \int \left(\frac{\partial}{\partial\beta} \log p(x; \beta, \hat{\eta}(\beta, F_0) + \alpha(t))\right) p(x; \beta, \hat{\eta}(\beta, F_0) + \alpha(t))dx \\ &= E_{\beta_0,\eta_0}\left[\phi(x, \beta_0, F_0)\left(\left.\frac{\partial}{\partial t}\right|_{t=0} \log p(x; \beta_0, \eta(t))\right)\right] \quad (\text{ by } (11)) \\ &\quad + \left.\frac{\partial}{\partial t}\right|_{t=0} E_{\beta_0,\eta_0}\left[\left.\frac{\partial}{\partial\beta}\right|_{\beta=\beta_0} \log p(x; \beta, \hat{\eta}(\beta, F_0) + \alpha(t))\right] \\ &= E_{\beta_0,\eta_0}\left[\phi(x, \beta_0, F_0)\left(\left.\frac{\partial}{\partial t}\right|_{t=0} \log p(x; \beta_0, \eta(t))\right)\right] \quad (\text{ by } (13)). \end{aligned} \tag{14}$$

Let $c \in \mathbb{R}^m$ be arbitrary. Then, it follows from Equation (14) that the product $c'\phi(x, \beta_0, F_0)$ is orthogonal to the nuisance tangent space $\dot{\mathcal{P}}_\eta$ which is the closed linear span of score functions of the form $\frac{\partial}{\partial t}|_{t=0} \log p(x; \beta_0, \eta(t))$. Using Condition (11), we have

$$\begin{aligned} \phi(x, \beta_0, F_0) &= \left.\frac{\partial}{\partial\beta}\right|_{\beta=\beta_0} \log p(x; \beta, \eta_0) + \left.\frac{\partial}{\partial\beta}\right|_{\beta=\beta_0} \log p(x; \beta_0, \hat{\eta}(\beta, F_0)) \\ &= \dot{\ell}_\beta(x; \beta_0, \eta_0) - \psi(x; \beta_0, \eta_0), \end{aligned}$$

where $\dot{\ell}_\beta(x;\beta_0,\eta_0) = \frac{\partial}{\partial\beta}|_{\beta=\beta_0} \log p(x;\beta,\eta_0)$ and $\psi(x;\beta_0,\eta_0) = -\frac{\partial}{\partial\beta}|_{\beta=\beta_0} \log p(x;\beta_0,\hat{\eta}(\beta,F_0))$. Finally, $c'\phi(x,\beta_0,F_0) = c'\dot{\ell}_\beta(x;\beta_0,\eta_0) - c'\psi(x;\beta_0,\eta_0)$ is orthogonal to the nuisance tangent space $\dot{\mathcal{P}}_\eta$ and $c'\psi(x;\beta_0,\eta_0) \in \dot{\mathcal{P}}_\eta$ implies that $c'\psi(x;\beta_0,\eta_0)$ is the orthogonal projection of $c'\dot{\ell}_\beta(x;\beta_0,\eta_0)$ onto the nuisance tangent space $\dot{\mathcal{P}}_\eta$. Since $c \in \mathbb{R}^m$ is arbitrary, $\phi(x,\beta_0,F_0)$ is the efficient score function. □

2.3 Comments on the MLE

The MLE is obtained when the function $\hat{\eta}(\beta,F)$ is given by

$$\hat{\eta}(\beta,F) = \mathrm{argmax}_{\eta\in\Theta_\eta} \int \log p(x;\beta,\eta)dF \tag{15}$$

If, for each $\beta \in \Theta_\beta$, the function $\hat{\eta}(\beta,F_0)$ uniquely maximizes the expected log-likelihood

$$E_{\beta_0,\eta_0} \log p(X;\beta,\eta) = \int \log p(x;\beta,\eta)dF_0,$$

then Conditions (11) and (12) in THEOREM 2 hold, provided the function $p(x;\beta,\eta)$ satisfies sufficient differentiability conditions.. It follows, by THEOREM 2, that the efficient score function is given by $\phi(x,\beta_0,F_0) = \frac{\partial}{\partial\beta}|_{\beta=\beta_0} \log p(x;\beta,\hat{\eta}(\beta,F_0))$.

We assumed the parameter η is in a Banach space. Suppose the function $\Phi(\beta,\eta,F) = \int \log p(x;\beta,\eta)dF$ is Fréchet differentiable with respect to η so that the maximizer $\hat{\eta}(\beta,F)$ in Equation (15) is the solution to the operator equation

$$d_\eta\Phi(\beta,\eta,F) = 0$$

where $d_\eta\Phi(\beta,\eta,F)$ is the derivative of $\Phi(\beta,\eta,F)$ with respect to η. Note that $d_\eta\Phi(\beta_0,\eta_0,F_0) = 0$. If $d_\eta\Phi(\beta,\eta,F)$ is Fréchet differentiable with respect to η and the second derivative

$$d^2_\eta\Phi(\beta_0,\eta_0,F_0)$$

is invertible, then by the implicit function theorem (cf. Appendix C), there exist $r > 0$ and $\rho > 0$ such that, for each (β,F) with $\|\beta-\beta_0\| < \rho$ and $\|F-F_0\| \equiv \sup_x |F(x)-F_0(x)| < \rho$, the solution $\hat{\eta}(\beta,F)$ with

$$\|\hat{\eta}(\beta,F)-\eta_0\| < r$$

exists. Moreover, if the operator $d_\eta\Phi(\beta,\eta,F)$ is k-times continuously Fréchet-differentiable with respect to (β,η,F), then the solution $\hat{\eta}(\beta,F)$ is k-times continuously Fréchet differentiable with respect to (β,F). Therefore, according to the implicit function theorem, the function $\Phi(\beta,\eta,F)$ must be at least 2-times continuously Fréchet differentiable with respect to the parameter η to assure the existence of the continuously Fréchet-differentiable function $\hat{\eta}(\beta,F)$. However, we have seen that the path-wise differentiability is sufficient to prove Theorem 1. Since we do not know the conclusions of the implicit function theorem hold when the Fréchet-differentiability is replaced with the path-wise differentiability, there is no guarantee that the maximizer $\hat{\eta}(\beta,F)$ exists and is differentiable under Conditions $(R0)-(R3)$.

2.4 Comments on Murphy and van der Vaart (2000)

Murphy and van der Vaart (2000) proved the efficiency of the profile likelihood by introducing an approximately least favorable submodel which we describe below: They assume that, for each $(\beta', \eta') \in \Theta_\beta \times \Theta_\eta$, there is a function

$$\beta \to \hat{\eta}(\beta, \beta', \eta') \tag{16}$$

such that

(P1) the function $\ell(x, \beta, \beta', \eta') = \log p(x; \beta, \hat{\eta}(\beta, \beta', \eta'))$ is twice continuously differentiable with respect to β and continuous with respect to (β', η') (Denote $\dot{\ell}^*_\beta(x, \beta, \beta', \eta') = \frac{\partial}{\partial \beta}\ell(x, \beta, \beta', \eta')$ and $\ddot{\ell}^*_\beta(x, \beta, \beta', \eta') = \frac{\partial^2}{\partial \beta^2}\ell(x, \beta, \beta', \eta')$);

(P2) $\hat{\eta}(\beta, \beta, \eta) = \eta$ for all $(\beta, \eta) \in \Theta_\beta \times \Theta_\eta$.

(P3) The efficient score function is given by $\dot{\ell}^*_\beta(x, \beta_0, \beta_0, \eta_0)$.

Let $\hat{\eta}_n(\beta)$ be the function defined by Equation (2). They also assume that, for any $\beta^*_n \xrightarrow{P} \beta_0$:

(P4) $\hat{\eta}_n(\beta^*_n) \xrightarrow{P} \eta_0$;

(P5) $E_{\beta_0, \eta_0}\dot{\ell}^*_\beta(x, \beta_0, \beta^*_n, \hat{\eta}_n(\beta^*_n)) = o_P(\|\beta^*_n - \beta_0\| + 1/\sqrt{n})$ (the no-bias condition);

(P6) the functions $\dot{\ell}^*_\beta(x, \beta, \beta', \eta')$ and $\ddot{\ell}^*_\beta(x, \beta, \beta', \eta')$ are continuous with respect to (β, β', η') at $(\beta_0, \beta_0, \eta_0)$;

(P7) the class of functions $\{\dot{\ell}^*_\beta(x, \beta, \beta', \eta') : (\beta, \beta', \eta') \in \Theta_\beta \times \Theta_\beta \times \Theta_\eta\}$ is Donsker with square integrable envelope function;

(P8) the class of functions $\{\ddot{\ell}^*_\beta(x, \beta, \beta', \eta') : (\beta, \beta', \eta') \in \Theta_\beta \times \Theta_\beta \times \Theta_\eta\}$ is Glivenko-Cantelli with integrable envelope function.

Under conditions (P1)–(P8), Murphy and van der Vaart (2000) show that the asymptotic expansion of the profile log-likelihood, for any $\beta^*_n \xrightarrow{P} \beta_0$, is

$$\begin{aligned}
\sum_{i=1}^n \log p(X_i; \beta^*_n, \hat{\eta}_n(\beta^*_n)) &= \sum_{i=1}^n \log p(X_i; \beta_0, \hat{\eta}_n(\beta_0)) + (\beta^*_n - \beta_0)^T \sum_{i=1}^n \dot{\ell}^*_\beta(X_i, \beta_0, \beta_0, \eta_0) \\
&+ \frac{n}{2}(\beta^*_n - \beta_0)^T \left[E_{\beta_0, \eta_0}(\dot{\ell}^*_\beta \dot{\ell}^{*T}_\beta)\right](\beta^*_n - \beta_0) \\
&+ o_P(\sqrt{n}\|\beta^*_n - \beta_0\| + 1)^2.
\end{aligned}$$

This equation leads to the asymptotic linearity of the estimator $\hat{\beta}_n$ with the efficient influence function.

In their proof, the function $\hat{\eta}(\beta, \beta', \eta')$ is used to create the upper and lower bounds for the expansion of the profile likelihood which is a function of $\hat{\eta}_n(\beta)$. They conclude that since the two bounds converge to the same expression, the expansion of the profile likelihood must converge to the same limit. Thus, they do not have to treat the function $\hat{\eta}_n(\beta)$, directly.

3 Example: Two-phase outcome-dependent sampling

In this section, we demonstrate that the estimator constructed by the method of Scott and Wild (1997, 2001) is efficient. We apply the method to the two-phase outcome-dependent sampling design of Lawless, Kalbfleish and Wild (1999). Breslow, McNeney and Wellner (2003) used the approach in Murphy and van der Vaart (2000) to demonstrate the efficiency of the estimator based on the profile likelihood in a variation of this example. On the contrary, we apply THEOREM 1 to show the same result.

Remark 3.1: In THEOREM 1, we proved the no bias condition (P5) in general context. Therefore, when we apply THEOREM 1, verification of (P5) is no longer needed. See Breslow, McNeney and Wellner (2003) for the verification of (P5) in the case of two-phase outcome-dependent sampling.

Two-phase outcome-dependent sampling: We assume that the underlying data generating process on the sample space $\mathcal{Y} \times \mathcal{X}$ is a model

$$\mathcal{Q} = \{p(y,x;\theta) = f(y|x;\theta)g(x) : \ \theta \in \Theta, \ g \in \mathcal{G}\}.$$

Here $f(y|x;\theta)$ is a conditional density of Y given X which depends on a finite dimensional parameter θ, $g(x)$ is an unspecified density of X which is an infinite-dimensional nuisance parameter. We assume the set Θ is a compact set containing a neighborhood of the true value θ_0 and $\mathcal{G}$ is the set of all densities of x. The variable Y may be discrete or continuous variable.

For a partition of the sample space $\mathcal{Y} \times \mathcal{X} = \cup_{s=1}^{S} \mathcal{S}_s$, let

$$Q_s(\theta, g) = \mathbb{P}\{(Y,X) \in \mathcal{S}_s\} = \int f(y|x;\theta)\, g(x)\, 1_{(y,x)\in\mathcal{S}_s}\, dy\, dx$$

and

$$Q_{s|X}(x;\theta) = \mathbb{P}\{(Y,x) \in \mathcal{S}_s|x\} = \int f(y|x;\theta) 1_{(y,x)\in\mathcal{S}_s} dy.$$

In each stratum $s = 1, \ldots, s$, let m_s be the number of fully observed units, $n_s - m_s$ be the number of subjects whose only information retained is the identity of the stratum. Lawless, Kalbfleish and Wild (1999) discussed variations of the two-phase, outcome-dependent sampling design (the variable probability sampling (VPS1,VPS2), the basic stratified sampling (BSS)). For all sampling schemes (VPS1,VPS2, and BSS), the resulting likelihoods are equal to

$$L(\theta, g) = \prod_{s=1}^{S} \left\{ \prod_{i=1}^{m_s} f(y_{si}|x_{si};\theta) g(x_{si}) \right\} Q_s(\theta, g)^{n_s - m_s} = \prod_{s=1}^{S} \left\{ \prod_{i=1}^{m_s} \frac{f(y_{si}|x_{si};\theta) g(x_{si})}{Q_s(\theta, g)} \right\} Q_s(\theta, g)^{n_s}. \tag{17}$$

The likelihood motivates us to interpret the observed data as an i.i.d. sample from the mixture of models

$$\mathcal{P}_s = \left\{ p_s(y,x;\theta,g) = \frac{f(y|x;\theta)g(x)1_{(y,x)\in\mathcal{S}_s}}{Q_s(\theta,g)} : \ \theta \in \Theta, \ g \in \mathcal{G} \right\}, \quad s = 1, \ldots, S,$$

and

$$\mathcal{P}_{S+1} = \{p_{S+1}(j;\theta,g) = Q_j(\theta,g) : \ \theta \in \Theta, \, g \in \mathcal{G}\},$$

where $j \in \{1,\ldots,S\}$ indicates the stratum. We denote the corresponding mixture probability density function as

$$p(s,z;\theta,g) = 1_{s\in\{1,\ldots,S\}} w_s p_s(y,x;\theta,g) + 1_{s=S+1} w_{S+1} p_{S+1}(j;\theta,g)$$

where $(s,z) = (s, 1_{s\in\{1,\ldots,S\}}(y,x) + 1_{s=S+1}j)$ and $w_s > 0$, $s = 1,\ldots,S,S+1$, with $\sum_{s=1}^{S+1} w_s = 1$.

Let F_{s0} and F_{sn} be the cdf for the true distribution and the empirical cdf in the model $\mathcal{P}_s$, respectively, $s = 1,\ldots,S+1$. Then the cdf for the true distribution and the empirical cdf in the mixture model are, respectively,

$$F_0(s,z) = w_s F_{s0}(z)$$

and

$$F_n(s,z) = 1_{s\in\{1,\ldots,S\}} \frac{m_s}{n_T} F_{sn} + 1_{s=S+1} \frac{n}{n_T} F_{(S+1)n}$$

where $n = \sum_{s=1}^{S} n_s$ and $n_T = n + \sum_{s=1}^{S} m_s$. We assume that $(\frac{m_1}{n_T},\ldots,\frac{m_S}{n_T},\frac{n}{n_T}) \to (w_1,\ldots,w_S,w_{S+1})$, and the $\sqrt{n}$- or $n^{1/4}$-consistency of the empirical cdf, i.e.,

$$\sqrt{n}\|F_n(s,z) - F_0(s,z)\| = O_P(1),$$

or

$$n^{1/4}\|F_n(s,z) - F_0(s,z)\| = o_P(1),$$

where $\|F_n(s,z) - F_0(s,z)\| = \sup_{s,z} |F_n(s,z) - F_0(s,z)|$.

Remark 3.2: It is possible to interpret the likelihood (Equation 17) as the one for an i.i.d. sample from the density

$$p(s,z;\theta,g) = \{w_1 f(y|x;\theta)g(x)\}^{1_{s=1}} \{w_2 Q_j(\theta,g)\}^{1_{s=2}}.$$

where $(s,z) = (s, 1_{s=1}(y,x) + 1_{s=2}j)$, $w_1, w_2 > 0$ and $w_1 + w_2 = 1$.

3.1 The efficient score function

Let F_s $(s = 1,\ldots,S,S+1)$ be a cdf function in the model $\mathcal{P}_s$ and $a_s > 0$ be such that $\sum_{s=1}^{S+1} a_s = 1$. For a mixture cdf $F(s,z) = a_s F_s(z)$, the integral of the log-likelihood is

$$\begin{aligned}&\int \log p(s,z;\theta,g)dF(s,z)\\ =& \sum_{s=1}^{S} \left\{ a_s \int (\log f(y|x;\theta) + \log g(x))\, dF_s + (a_{S+1}dF_{S+1}(s) - a_s) \log Q_s(\theta,g) \right\}. \quad (18)\end{aligned}$$

Then, the expected log-likelihood and the averaged log-likelihood are

$$\begin{aligned}&\int \log p(s,z;\theta,g)dF_0(s,z)\\ =& \sum_{s=1}^{S} \{w_{S+1}Q_s(\theta_0,g_0)\log Q_s(\theta,g) + w_s E_{s,\theta_0,g_0}[\log f(Y|X;\theta) + \log g(X) - \log Q_s(\theta,g)]\}\end{aligned}$$

and

$$\begin{aligned}\frac{1}{n_T}\ell_n(\theta, g) &= \int \log p(s,z;\theta,g)dF_n(s,z) \\ &= \sum_{s=1}^{S}\left\{\frac{n}{n_T}\frac{n_s}{n}\log Q_s(\theta,g) + \frac{1}{n_T}\sum_{i=1}^{m_s}[\log f(Y_{si}|X_{si};\theta) + \log g(X_{si}) - \log Q_s(\theta,g)]\right\}.\end{aligned}$$

Theorem A.*[The efficient score function] Let*

$$\hat{g}(x,\theta,F) = \frac{f^*(x,F)}{\sum_{s=1}^{S} a_s^*(\theta,F)\frac{Q_{s|X}(x;\theta)}{\hat{Q}_s(\theta,F)}}, \tag{19}$$

where

$$f^*(x,F) = \sum_{s=1}^{S} a_s \int \frac{dF_s}{d(y,x)}dy,$$

$$a_s^*(\theta,F) = a_{S+1}[\hat{Q}_s(\theta,F) - dF_{S+1}(s)] + a_s,$$

and

$$\hat{Q}_s(\theta,F) = \int Q_{s|X}(x;\theta)\hat{g}(x,\theta,F)dx. \tag{20}$$

Then the efficient score function is given by

$$\phi(s,z,\theta_0,F_0) = \frac{\partial}{\partial\theta}\bigg|_{\theta=\theta_0} \log p(s,z;\theta,\hat{g}(x,\theta,F_0)). \tag{21}$$

The proof of Theorem A is given in Appendix A.

Remark 3.3: Recall that, for each $s = 1,\ldots,S+1$, $F_{s,0}$ is the cdf for the true distribution in the model $\mathcal{P}_s$. In particular, when $s = S+1$, the function $F_{(S+1)0}$ is the cdf for the true distribution on the sample space $\{1,\ldots,S\}$, i.e., $dF_{(S+1)0}(i) = Q_i(\theta_0, g_0)$, $i = 1,\ldots,S$. Note that

$$\begin{aligned} f^*(x,F_0) &= \sum_{s=1}^{S} w_s \int \frac{dF_{s,0}}{d(y,x)}dy \\ &= \sum_{s=1}^{S} w_s \int \frac{f(y|x;\theta)1_{(y,x)\in\mathcal{S}_s} g_0(x)}{Q_s(\theta_0,g_0)}dy = \sum_{s=1}^{S} w_s \frac{Q_{s|X}(x;\theta_0)}{Q_s(\theta_0,g_0)}g_0(x).\end{aligned}$$

Therefore, if $\hat{Q}_s(\theta_0,F_0) = Q_s(\theta_0,g_0)$ then

$$1 > a_s^*(\theta_0,F_0) = w_{S+1}[\hat{Q}_s(\theta_0,F_0) - Q_s(\theta_0,g_0)] + w_s = w_s > 0 \tag{22}$$

and

$$\hat{g}(x,\theta_0,F_0) = \frac{f^*(x,F_0)}{\sum_{s=1}^{S} w_s \frac{Q_{s|X}(x;\theta_0)}{\hat{Q}_s(\theta_0,F_0)}} = g_0(x).$$

On the other hand, if $\hat{g}(x,\theta_0,F_0) = g_0(x)$ then

$$1 > \hat{Q}_s(\theta_0,F_0) = \int Q_{s|X}(x;\theta_0)g_0(x)dx = Q_s(\theta_0,g_0) > 0. \tag{23}$$

Remark 3.4: The function $\hat{g}(x,\theta,F)$ given by Equation (19) induces a version of the approximate least favorable submodel in Murphy and van der Vaart (2000). For $s = 1,\dots,S+1$, let $F_s(z;\theta',g')$ be the cdf for the density $p_s(z;\theta',g')$ in the model $\mathcal{P}_s$. Let $F(\theta',g') = w_s F_s(z;\theta',g')$. Then a version of the approximate least favorable submodel is given by

$$\hat{g}(x,\theta,F(\theta',g')) = \frac{f^*(x,F(\theta',g'))}{\sum_{s=1}^{S} a_s^*(\theta,F(\theta',g'))\frac{Q_{s|X}(x;\theta)}{\hat{Q}_s(\theta,F(\theta',g'))}}$$

where

$$\begin{aligned} f^*(x,F(\theta',g')) &= \sum_{s=1}^{S} w_s \int \frac{f(y|x;\theta')1_{(y,x)\in\mathcal{S}_s} g'(x)}{Q_s(\theta',g')}dy = \sum_{s=1}^{S} w_s \frac{Q_{s|X}(x;\theta')}{Q_s(\theta',g')} g'(x), \\ a_s^*(\theta,F(\theta',g')) &= w_{S+1}[\hat{Q}_s(\theta,F(\theta',g')) - Q_s(\theta',g')] + w_s \end{aligned}$$

and

$$\hat{Q}_s(\theta,F(\theta',g')) = \int Q_{s|X}(x;\theta)\hat{g}(x,\theta,F(\theta',g'))dx.$$

Remark 3.5: Let $\mathcal{F}$ be the set of cdf's on the sample space and, for $\rho > 0$, let

$$\mathcal{C}_\rho = \{F \in \mathcal{F} : \sup_{s,z} |F(s,z) - F(s,z)| < \rho\}.$$

By Equations (22) and (23) and continuity of the functions with respect to (θ, F) (continuity will be verified in the next section), the following assumption should hold:

(T1) there are $\rho > 0$ and compact set Θ containing a neighborhood of θ_0 such that, for $s = 1,\dots,S$ and for all $(\theta,F) \in \Theta \times \mathcal{C}_\rho$,

$$1 > a_s^*(\theta,F) \geq \delta > 0$$

and

$$1 > \hat{Q}_s(\theta,F) \geq \delta > 0.$$

This condition will be used to verify Condition (R3) in Theorem 1.

3.2 Asymptotic normality

We assume the $\sqrt{n}$-consistency of the empirical cdf

$$\sqrt{n}\|F_n(s,z) - F_0(s,z)\| = O_P(1),$$

and we verify conditions $(R0)$, $(R1)$, $(R2)$, and $(R3)$ so that we can apply Theorem 1 to show the efficiency of the MLE based on the profile likelihood in this example.

Remark 3.6: In this example, we could assume the $n^{1/4}$-consistency of the empirical cdf,

$$n^{1/4}\|F_n(s,z) - F_0(s,z)\| = o_P(1),$$

and verify conditions $(R0)$, $(R1)^*$, $(R2)$, and $(R3)$ to apply Theorem 1. Since the vitrification of both cases are similar, we present only one of them.

Condition (R0): This condition is verified by Theorem A.

Condition (R1): We assume that

(T2) for all $\theta \in \Theta$, the function $f(y|x;\theta)$ is twice continuously differentiable with respect to θ.

For any path $g(t) = g(x,t)$, the densities

$$p_s(y,x;\theta,g(t)) = \frac{f(y|x;\theta)g(x,t)1_{(y,x)\in\mathcal{S}_s}}{\int f(y|x;\theta)g(x,t)1_{(y,x)\in\mathcal{S}_s}dydx}$$

and

$$p_{S+1}(i;\theta,g(t)) = Q_i(\theta,g) = \int f(y|x;\theta)g(x,t)1_{(y,x)\in\mathcal{S}_i}dydx$$

are twice continuously differentiable with respect to θ and continuously differentiable with respect to t. Therefore to verify condition (R1), all we need is the differentiability of the function $\hat{g}(x,\theta,F)$.

Because Equation (19) and Equation (20) form a system of equations, the differentiability of these equations depends on the differentiability of the other. The function $f^*(x,F) = \sum_{s=1}^S a_s \int \frac{dF_s}{d(y,x)}dy$ is linear with respect to F. This implies that it is continuously path-wise differentiable with respect to F. The function $a_s^*(\theta,F) = a_{S+1}[\hat{Q}_s(\theta,F) - dF_{S+1}(s)] + a_s$ is twice continuously differentiable with respect to θ and continuously path-wise differentiable with respect to F if $\hat{Q}_s(\theta,F)$ is. By assumption (T2), $Q_{s|X}(x;\theta) = \int f(y|x;\theta)1_{(y,x)\in\mathcal{S}_s}dy$ is twice continuously differentiable with respect to θ. By Equation (19) and Equation (20), these differentiabilities imply that the maximizer $\hat{g}(x,\theta,F)$ is twice continuously differentiable with respect to θ and continuously path-wise differentiable with respect to F if $\hat{Q}_s(\theta,F)$ is. Conversely, if the function $\hat{g}(x,\theta,F)$ is twice continuously differentiable with respect to θ and continuously path-wise differentiable with respect to F, then so is the function $\hat{Q}_s(\theta,F)$.

Derivatives of log-likelihood: The log-likelihood function for one observation is

$$\begin{aligned}\log p(s,z;\theta,\hat{g}(x,\theta,F)) &= \{1_{s=S+1}1_{i\in\{1,\ldots,S\}} - 1_{s\in\{1,\ldots,S\}}1_{i=s}\}\log\hat{Q}_i(\theta,F)\\ &\quad +1_{s\in\{1,\ldots,S\}}\{\log f(y|x;\theta) + \log\hat{g}(x,\theta,F)\}. \qquad (24)\end{aligned}$$

The induced score function is

$$\begin{aligned}\phi(s,z,\theta,F) &= \frac{\partial}{\partial\theta}\log p(s,z;\theta,\hat{g}(x,\theta,F))\\ &= \{1_{s=S+1}1_{i\in\{1,\ldots,S\}} - 1_{s\in\{1,\ldots,S\}}1_{i=s}\}\frac{\dot{\hat{Q}}_{i,\theta}}{\hat{Q}_i}(\theta,F)\\ &\quad +1_{s\in\{1,\ldots,S\}}\left\{\frac{\dot{f}}{f}(y|x;\theta) + \frac{\dot{\hat{g}}_\theta}{\hat{g}}(x,\theta,F)\right\} \qquad (25)\end{aligned}$$

where $\dot{f} = \frac{\partial}{\partial\theta}f$, $\dot{\hat{Q}}_{i,\theta} = \frac{\partial}{\partial\theta}\hat{Q}_i$ and $\dot{\hat{g}}_\theta = \frac{\partial}{\partial\theta}\hat{g}$. The derivatives of the induced score function with respect to θ is

$$\begin{aligned}\frac{\partial}{\partial\theta}\phi(s,z,\theta,F) &= \{1_{s=S+1}1_{i\in\{1,\ldots,S\}} - 1_{s\in\{1,\ldots,S\}}1_{i=s}\}\left\{\frac{\ddot{\hat{Q}}_{i,\theta}}{\hat{Q}_i} - \left(\frac{\dot{\hat{Q}}_{i,\theta}}{\hat{Q}_i}\right)^2\right\}\\ &\quad +1_{s\in\{1,\ldots,S\}}\left\{\frac{\ddot{f}}{f} - \left(\frac{\dot{f}}{f}\right)^2 + \frac{\ddot{\hat{g}}_\theta}{\hat{g}} - \left(\frac{\dot{\hat{g}}_\theta}{\hat{g}}\right)^2\right\}. \qquad (26)\end{aligned}$$

where $\ddot{f} = \frac{\partial^2}{\partial\theta^2} f$, $\ddot{\hat{Q}}_{i,\theta} = \frac{\partial^2}{\partial\theta^2}\hat{Q}_i$, and $\ddot{\hat{g}}_\theta = \frac{\partial^2}{\partial\theta^2}\hat{g}$.

Condition (R2): We assume that

(T3) There is no $a \in \mathbb{R}^m$ such that $a^T \frac{\dot{f}}{f}(y|x;\theta)$ is constant in y for almost all x.

The term $\frac{\dot{Q}_{i,\theta}}{Q_i}(\theta_0, F_0)$ is a nonrandom vector and $\frac{\dot{\hat{g}}_\theta}{\hat{g}}(x, \theta_0, F_0)$ is a function of x. Therefore, by Equation (25) and assumption (T3), there is no $a \in \mathbb{R}^m$ such that $a^T\phi(s, z, \theta_0, F_0)$ is constant in y for almost all x. By Theorem 1.4 in Seber and Lee (2003), $\sum_{s=1}^S w_s E_{s,\beta_0,F_0}(\phi\phi^T)$ is nonsingular with the bounded inverse.

Conditions (R3): We assume that

(T4) envelope functions

$$\sup_{\theta\in\Theta}\left\|\dot{f}(y|x;\theta)\right\|,\ \sup_{\theta\in\Theta}\left\|\ddot{f}(y|x;\theta)\right\|,\ \sup_{\theta\in\Theta}\left\|\frac{\dot{f}}{f}(y|x;\theta)\right\|,$$
$$\sup_{\theta\in\Theta}\left(\int\left\|\dot{f}(y|x;\theta)\right\|dy\right)^2,\ \sup_{\theta\in\Theta}\left(\int\left\|\frac{\dot{f}}{f}(y|x;\theta)\right\|dy\right)\left(\int\left\|\dot{f}(y|x;\theta)\right\|dy\right)$$

are integrable;

(T5) non-random functions $\|\dot{\hat{Q}}_{i,\theta}\|$ and $\|\ddot{\hat{Q}}_{i,\theta}\|$ are bounded by some positive constant L on the set $\Theta \times \mathcal{C}_\rho$ which we defined in (T1);

(T6) the classes

$$\left\{\frac{\dot{f}}{f}(y|x;\theta):\ \theta\in\Theta\right\},\ \left\{Q_{s|X}(x;\theta):\ \theta\in\Theta\right\},\ \left\{\dot{Q}_{s|X}(x;\theta) = \frac{\partial}{\partial\theta}Q_{s|X}(x;\theta):\ \theta\in\Theta\right\}$$

are P_{θ_0,g_0}-Donsker classes of functions.

Function $\frac{\partial}{\partial\theta}\phi(s, z, \theta, F)$ is continuous in the parameters (θ, F), the set Θ is compact, and the set $\mathcal{C}_\rho$ in condition (T1) is a P_{θ_0,g_0}-Donsker class (cf. van der Vaart (1998), page 273). By Theorem 3 in van der Vaart and Wellner (2000), the class

$$\left\{\frac{\partial}{\partial\theta}\phi(s, z, \theta, F):\ (\theta, F)\in\Theta\times\mathcal{C}_\rho\right\}$$

is P_{θ_0,g_0}-Glivenko-Cantelli if it has an integrable envelope function. In Appendix B, we show that the class has integrable envelope function.

Also, in Appendix B, we show that the class of function

$$\{\phi(s, z, \theta, F):\ (\theta, F)\in\Theta\times\mathcal{C}_\rho\}$$

is P_{θ_0,g_0}-Donsker with square integrable envelope function.

Remark 3.7: Conditions (T2), (T3), (T4) and (T6) are satisfied by the logistic regression model

$$f(y|x;\theta) = \frac{e^{y(\theta^T x)}}{1 + e^{\theta^T x}}$$

where $y \in \{0, 1\}$, $x \in \mathbb{R}^m$, $\theta \in \mathbb{R}^m$.

4 Discussion

We have shown the efficiency of the estimator based on the profile likelihood in general semiparametric model. But this does not answer the question that "Is the estimator based on the profile likelihood by the method of Lagrange multipliers in Scott and Wild (1997, 2001) efficient?"

In the example of the two-phase, outcome-dependent sampling design, the method in Scott and Wild (1997, 2001) gives us a candidate function $\hat{g}(x, \theta, F)$. In Theorem A, we showed that the corresponding induced score function (Equation (21)) gives the efficient score function in the example. Then Theorem 1 and its Corollary can be applied to showed that the estimator is an asymptotically linear estimator with the efficient influence function. Thus the estimator is efficient in this example. However, for general case, Theorem 1 can not be applied, since we do not know that the candidate function $\hat{g}(x, \theta, F)$ given by the method gives the efficient score function.

Future work remains to prove or disprove the efficiency of the estimator based on the profile likelihood by the method of Lagrange multipliers.

Appendix A: Proof of Theorem A

In Step 1, we find a function $\hat{g}(\theta, F) = \hat{g}(x, \theta, F)$ by using the method of Scott and Wild (1997, 2001). In Step 2, we show that $\int \log p(s, z; \theta, \hat{g}(\theta, F_0)) dF_0(s, z)$ satisfies Conditions (11) and (12) in Theorem 2 so that the claim follows form this theorem.

Step 1: First, we find a function $\hat{g}(x, \theta, F)$ under the assumption that the support of the distribution of X is finite: i.e. $\mathrm{supp}(X) = \{v_1, \ldots, v_K\}$. Let $(g_1, \ldots, g_K) = (g(v_1), \ldots, g(v_K))$, then $\log g(x)$ and $Q_s(\theta, g)$ can be expressed as $\log g(x) = \sum_{k=1}^K 1_{x=v_k} \log g_k$ and $Q_s(\theta, g) = \int Q_{s|X}(x; \theta) g(x) dx = \sum_{k=1}^K Q_{s|X}(v_k; \theta) g_k$.

To find the maximizer $(g_1, \ldots, g_K)$ of

$$\int \log p(\theta, g) dF \;=\; \sum_{s=1}^S \left\{ a_s \int \left(\log f(y|x; \theta) + \log g(X) \right) dF_s + (a_{S+1} dF_{S+1}(s) - a_s) \log Q_s(\theta, g) \right\}$$

at θ, differentiate $\int \log p(\theta, g) dF$ with respect to g_k,

$$\frac{\partial}{\partial g_k} \int \log p(\theta, g) dF = \sum_{s=1}^S \left\{ a_s \frac{\int 1_{X=v_k} dF_s}{g_k} + (a_{S+1} dF_{S+1}(s) - a_s) \frac{Q_{s|X}(v_k; \theta)}{Q_s(\theta, g)} \right\}.$$

Let η be a Lagrange multiplier to account for $\sum_k g_k = 1$. Set $\frac{\partial}{\partial g_k} \int \log p(\theta, g) dF + \eta = 0$. Multiply by g_k and sum over $k = 1, \ldots, K$. Then $\sum_{s=1}^S \{a_s + (a_{S+1} dF_{S+1}(s) - a_s)\} + \eta = 0$ or $\eta = -a_{S+1} \sum_{s=1}^S dF_{S+1}(s) = -a_{S+1}$. Therefore $\frac{\partial}{\partial g_k} \int \log p(\theta, g) dF - a_{S+1} = 0$ or

$$\hat{g}(v_k, \theta, F) = g_k \;=\; \frac{\sum_{s=1}^S a_s \int 1_{X=v_k} dF_s}{a_{S+1} - \sum_{s=1}^S (a_{S+1} dF_{S+1}(s) - a_s) \frac{Q_{s|X}(v_k; \theta)}{Q_s(\theta, g)}}.$$

This function is of the form in Equation (19).

Step 2: Condition (11) is verified in Remark 3.3. Now, we verify Condition (12). Let $g(x,t)$ be a path in the space of density functions with $g(x,0) = g_0(x)$. Define $\alpha(t) = \alpha(x,t) = g(x,t) - g_0(t)$ and write $\alpha'(x,0) = \frac{\partial}{\partial t}|_{t=0}\,\alpha(x,t)$. Then

$$\begin{aligned}
& \frac{\partial}{\partial t}\bigg|_{t=0} \int \log p(s,z;\theta,\hat{g}(\theta,F_0)+\alpha(t))dF_0(s,z) \\
= & \frac{\partial}{\partial t}\bigg|_{t=0} \sum_{s=1}^{S} \left\{ w_s \int \log(\hat{g}(x,\theta,F_0)+\alpha(t))dF_{s,0} + (w_{S+1}Q_s(\theta_0,g_0)-w_s)\log Q_s(\theta,\hat{g}(\theta,F_0)+\alpha(t)) \right\} \\
= & \frac{\partial}{\partial t}\bigg|_{t=0} \int \log(\hat{g}(x,\theta,F_0)+\alpha(t))f^*(x,F_0)dx \\
& + \frac{\partial}{\partial t}\bigg|_{t=0} \sum_{s=1}^{S}(w_{S+1}Q_s(\theta_0,g_0)-w_s)\log Q_s(\theta,\hat{g}(\theta,F_0)+\alpha(t)) \\
= & \int \frac{\alpha'(x,0)}{\hat{g}(x,\theta,F_0)} f^*(x,F_0)dx + \sum_{s=1}^{S}(w_{S+1}Q_s(\theta_0,g_0)-w_s)\frac{\int Q_{s|X}(x;\theta)\alpha'(x,0)dx}{\hat{Q}_s(\theta,F_0)} \\
= & \sum_{s=1}^{S} a_s^*(\theta,F_0)\frac{\int Q_{s|X}(x;\theta)\alpha'(x,0)dx}{\hat{Q}_s(\theta,F_0)} + \sum_{s=1}^{S}(w_{S+1}Q_s(\theta_0,g_0)-w_s)\frac{\int Q_{s|X}(x;\theta)\alpha'(x,0)dx}{\hat{Q}_s(\theta,F_0)} \\
= & \sum_{s=1}^{S}\{a_s^*(\theta,F_0)+(w_{S+1}Q_s(\theta_0,g_0)-w_s)\}\frac{\int Q_{s|X}(x;\theta)\alpha'(x,0)dx}{\hat{Q}_s(\theta,F_0)} \\
= & w_{S+1}\sum_{s=1}^{S}\hat{Q}_s(\theta,F_0)\frac{\int Q_{s|X}(x;\theta)\alpha'(x,0)dx}{\hat{Q}_s(\theta,F_0)} = w_{S+1}\sum_{s=1}^{S}\int Q_{s|X}(x;\theta)\alpha'(x,0)dx \\
= & w_{S+1}\int \alpha'(x,0)dx = w_{S+1}\frac{\partial}{\partial t}\bigg|_{t=0}\int g(x,t)dx = 0.
\end{aligned}$$

where we used

$$\begin{aligned}
a_s^*(\theta,F_0)+(w_{S+1}Q_s(\theta_0,g_0)-w_s) &= w_{S+1}[\hat{Q}_s(\theta,F_0)-Q_s(\theta_0,g_0)]+w_s+(w_{S+1}Q_s(\theta_0,g_0)-w_s) \\
&= w_{S+1}\hat{Q}_s(\theta,F_0).
\end{aligned}$$

Appendix B: Verification of conditions (R3) continued

The function $B(x,\theta,F)$ and its derivatives: Let

$$B(x,\theta,F) = \sum_{s=1}^{S} a_s^*(\theta,F)\frac{Q_{s|X}(x;\theta)}{\hat{Q}_s(\theta,F)}.$$

Then the maximizer (Equation (19)) is $\hat{g}(x,\theta,F) = \frac{f^*(x,F)}{B(x,\theta,F)}$.

Note that, since $1 > a_s^*(\theta,F) \geq \delta > 0$ and $1 > \hat{Q}_s(\theta,F) \geq \delta > 0$ (assumption (T1)), for all $(\theta,F) \in \Theta \times \mathcal{C}_\rho$,

$$\delta = \delta\sum_{s=1}^{S} Q_{s|X}(x;\theta) \leq B(x,\theta,F) \leq \frac{1}{\delta}\sum_{s=1}^{S} Q_{s|X}(x;\theta) = \frac{1}{\delta}. \tag{27}$$

The first and second derivatives of $B(x,\theta,F)$ with respect to θ are

$$\dot{B}_\theta(x,\theta,F) = \frac{\partial}{\partial\theta}B(x,\theta,F) = \sum_{s=1}^{S}\left\{\dot{a}^*_{s,\theta}\frac{Q_{s|X}}{\hat{Q}_s} + a^*_s\frac{\dot{Q}_{s|X}\hat{Q}_s - Q_{s|X}\dot{\hat{Q}}_{s,\theta}}{\hat{Q}_s^2}\right\} \tag{28}$$

and

$$\begin{aligned}\ddot{B}_\theta(x,\theta,F) &= \frac{\partial^2}{\partial\theta^2}B(x,\theta,F)\\ &= \sum_{s=1}^{S}\left\{\ddot{a}^*_{s,\theta}\frac{Q_{s|X}}{\hat{Q}_s} + 2\dot{a}^*_{s,\theta}\frac{\dot{Q}_{s|X}\hat{Q}_s - Q_{s|X}\dot{\hat{Q}}_{s,\theta}}{\hat{Q}_s^2}\right.\\ &\qquad \left. + a^*_s\frac{\ddot{Q}_{s|X}\hat{Q}_s^2 - 2\dot{Q}_{s|X}\dot{\hat{Q}}_{s,\theta}\hat{Q}_s - Q_{s|X}\ddot{\hat{Q}}_{s,\theta}\hat{Q}_s + 2Q_{s|X}\dot{\hat{Q}}_{s,\theta}^{\,2}}{\hat{Q}_s^3}\right\}.\end{aligned}$$

Verifying the class $\{\phi(s,z,\theta,F):\ (\theta,F)\in\Theta\times\mathcal{C}_\rho\}$ is Donsker:

By assumptions (T1) and (T6), the classes

$$\{B(x,\theta,F):\ (\theta,F)\in\Theta\times\mathcal{C}_\rho\} \text{ and } \{\dot{B}_\theta(x,\theta,F):\ (\theta,F)\in\Theta\times\mathcal{C}_\rho\}$$

are uniformly bounded P_{θ_0,g_0}-Donsker classes. By Equation (27) and Example 2.10.9, page 192, van der Vaart and Wellner (1996), the class

$$\left\{\frac{1}{B(x,\theta,F)}:\ (\theta,F)\in\Theta\times\mathcal{C}_\rho\right\}$$

is a uniformly bounded P_{θ_0,g_0}-Donsker class. By Example 2.10.8, page 192, van der Vaart and Wellner (1996), it follows that the class

$$\left\{\frac{\dot{\hat{g}}_\theta}{\hat{g}}(x,\theta,F) = -\frac{\dot{B}_\theta}{B}(x,\theta,F):\ (\theta,F)\in\Theta\times\mathcal{C}_\rho\right\}$$

is a uniformly bounded P_{θ_0,g_0}-Donsker class.

Finally, since the class $\left\{\frac{\dot{f}}{f}(y|x;\theta):\ \theta\in\Theta\right\}$ is P_{θ_0,g_0}-Donsker (assumption (T6)) and bounded in $L_1(P_{\theta_0,g_0})$ (assumption (T4)), and the functions $\frac{\dot{\hat{Q}}_{i,\theta}}{\hat{Q}_i}(\theta,F)$, $i=1,\ldots,S$, are bounded nonrandom continuous functions on the set $\theta\times\mathcal{C}_\rho$ (see (a) below), by Equation (25) and Example 2.10.7, page 192, van der Vaart and Wellner (1996), we have the class $\{\phi(s,z,\theta,F):\ (\theta,F)\in\Theta\times\mathcal{C}_\rho\}$ is P_{θ_0,g_0}-Donsker.

Verifying the classes have integrable and square integrable envelope functions:

We show that the class $\{\phi(s,z,\theta,F):\ (\theta,F)\in\Theta\times\mathcal{C}_\rho\}$ has square integrable function and the class$\{\frac{\partial}{\partial\theta}\phi(s,z,\theta,F):\ (\theta,F)\in\Theta\times\mathcal{C}_\rho\}$ has integrable function.

By Equation (25) and Equation (26), it is enough to show that

(a) $\left\|\frac{\dot{Q}_{i,\theta}}{Q_i}\right\|$, $\left\|\frac{\ddot{Q}_{i,\theta}}{Q_i}\right\|$, $\left\|\frac{\dot{Q}_{i,\theta}}{Q_i}\right\|\times\left\|\frac{\dot{Q}_{j,\theta}}{Q_j}\right\|$, $i,j=1,\ldots,S$, are bounded by some constant;

(b) the classes $\left\{\frac{\dot{\hat{g}}_\theta}{\hat{g}}(x,\theta,F):\ (\theta,F)\in\Theta\times\mathcal{C}_\rho\right\}$, $\left\{\frac{\ddot{\hat{g}}_\theta}{\hat{g}}(x,\theta,F):\ (\theta,F)\in\Theta\times\mathcal{C}_\rho\right\}$, $\left\{\left(\frac{\dot{\hat{g}}_\theta}{\hat{g}}(x,\theta,F)\right)^2:\ (\theta,F)\in\Theta\times\mathcal{C}_\rho\right\}$, $\left\{\left(\frac{\dot{f}}{f}(y|x;\theta)\right)^T\frac{\dot{\hat{g}}_\theta}{\hat{g}}(x,\theta,F):\ (\theta,F)\in\Theta\times\mathcal{C}_\rho\right\}$ have an integrable envelope function.

Derivatives of $Q_s(\theta,F)$: Derivatives $\dot{\hat{Q}}_{i,\theta}=\frac{\partial}{\partial\theta}\hat{Q}_i$ and $\ddot{\hat{Q}}_{i,\theta}=\frac{\partial^2}{\partial\theta^2}\hat{Q}_i$ are non-random functions of (θ,F) on a compact set $\Theta\times\mathcal{C}_\rho$. By assumption (T1), $\hat{Q}_s(\theta,F)\geq\delta>0$ for all $(\theta,F)\in\Theta\times\mathcal{C}_\rho$. This and (T5) imply $\left\|\frac{\dot{Q}_{i,\theta}}{Q_i}\right\|\leq\frac{L}{\delta}$, $\left\|\frac{\ddot{Q}_{i,\theta}}{Q_i}\right\|\leq\frac{L}{\delta}$, $\left\|\frac{\dot{Q}_{i,\theta}}{Q_i}\right\|\times\left\|\frac{\dot{Q}_{j,\theta}}{Q_j}\right\|\leq\frac{L^2}{\delta^2}$. Therefore we have (a).

Envelope functions for derivatives:

Since $0<a_s<1$ $(s=1,\ldots,S,S+1)$, with assumption $(T5)$, we have

$$\|\dot{a}^*_{s,\theta}(\theta,F)\|\leq\|\dot{\hat{Q}}_{s,\theta}(\theta,F)\|\leq L. \tag{29}$$

Combine assumption (T1), (T5), Equation (28), (29) and $\sum_{s=1}^S Q_{s|X}(x;\theta)=1$ to get

$$\begin{aligned}
\|\dot{B}_\theta(x,\theta,F)\| &\leq \sum_{s=1}^S\left\{\|\dot{a}^*_{s,\theta}\|\frac{Q_{s|X}}{\hat{Q}_s}+a^*_s\frac{\|\dot{Q}_{s|X}\|\hat{Q}_s+Q_{s|X}\|\dot{\hat{Q}}_{s,\theta}\|}{\hat{Q}_s^2}\right\}\\
&\leq \sum_{s=1}^S\left\{L\frac{Q_{s|X}}{\delta}+1\cdot\frac{\|\dot{Q}_{s|X}\|\cdot 1+Q_{s|X}L}{\delta^2}\right\}\\
&\leq \frac{L}{\delta}+\frac{\int\|\dot{f}(y|x;\theta)\|dy+L}{\delta^2}\\
&= c_1\int\|\dot{f}(y|x;\theta)\|dy+c_2
\end{aligned} \tag{30}$$

where $c_1=\frac{1}{\delta^2}>0$ and $c_2=\frac{L}{\delta}+\frac{L}{\delta^2}>0$.

Similarly, for some positive constants c_1,c_2,c_3,

$$\|\ddot{B}_\theta(x,\theta,F)\|\leq c_1\int\|\ddot{f}(y|x;\theta)\|dy+c_2\int\|\dot{f}(y|x;\theta)\|dy+c_3. \tag{31}$$

Since $B(x,\theta,F)\geq\delta>0$, Equations (30) and (31) imply that, for some positive constants c_1,c_2,c_3,c_4,

$$\left\|\frac{\dot{\hat{g}}_\theta}{\hat{g}}(x,\theta,F)\right\|=\left\|-\frac{\dot{B}_\theta}{B}(x,\theta,F)\right\|\leq c_1\int\|\dot{f}(y|x;\theta)\|dy+c_2,$$

$$\left\|\frac{\dot{\hat{g}}_\theta}{\hat{g}}(x,\theta,F)\right\|^2\leq c_1\left(\int\|\dot{f}(y|x;\theta)\|dy\right)^2+c_2\int\|\dot{f}(y|x;\theta)\|dy+c_3,$$

$$\begin{aligned}
\left\|\frac{\ddot{\hat{g}}_\theta}{\hat{g}}(x,\theta,F)\right\| &= \left\|-\frac{\ddot{B}_\theta}{B}+2\left(\frac{\dot{B}_\theta}{B}\right)^2\right\|\\
&\leq c_1\int\|\ddot{f}(y|x;\theta)\|dy+c_2\left(\int\|\dot{f}(y|x;\theta)\|dy\right)^2+c_3\int\|\dot{f}(y|x;\theta)\|dy+c_4,
\end{aligned}$$

and

$$\left\|\left(\frac{\dot{f}}{f}(y|x;\theta)\right)^T\frac{\dot{\hat{g}}_\theta}{\hat{g}}(x,\theta,F)\right\|=\left\|\left(\frac{\dot{f}}{f}(y|x;\theta)\right)^T\frac{\dot{B}_\theta}{B}(x,\theta,F)\right\|\leq\left\|\frac{\dot{f}}{f}(y|x;\theta)\right\|\left(c_1\int\|\dot{f}(y|x;\theta)\|dy+c_2\right)$$

(by the Cauchy-Schwarz inequality). By assumption (T4), we have condition (b).

Appendix C: Implicit function theorem

This version of the implicit function theorem is taken from Theorem 4E, Zielder, 1995, page 250. Let X, Y, and Z be Banach spaces, and let $F(u,v)$ is an n-times continuously Fréchet differentiable map from an open neighborhood $U(u_0,v_0) \subset X \times Y$ of (u_0,v_0) to Z such that

$$F(u_0,v_0) = 0$$

and

$$F_v(u_0,v_0) : Y \to Z \text{ is bijective.}$$

Then there exist $r >$ and $\rho > 0$ such that, for each given $u \in X$ with $\|u - u_0\| < \rho$, the equation

$$F(u,v) = 0$$

has a solution v, denoted by $v(u)$, such that

$$\|v - v_0\| < r.$$

Moreover, the function $u \to v(u)$ is n-times continuously Fréchet differentiable.

References

Begun, J. M., Hall, W. J., Huang, W. M. and Wellner, J. A. (1983). Information and asymptotic efficiency in parametric–nonparametric models. *Ann. Statist.* **11** 432–452.

Bickel, P.J., Klaassen, C.A.J., Ritov, Y. and Wellner, J.A. (1993). *Efficient and Adaptive Estimation for Semiparametric Models.* Johns Hopkins Univ. Press, Baltimore.

Breslow, N.E. and Chatterjee, N. (1999). Design and analysis of two-phase studies with binary outcome applied to Wilms tumor prognosis. *Appl. Statist.* **48** 457–468.

Breslow, N.E. and Holubkov, R. (1997). Maximum likelihood estimation of logistic regression parameters under two-phase, outcome-dependent sampling. *J. Roy. Statist. Soc. ser. B* **59** 447–461.

Breslow, N.E., McNeney, B. and Wellner, J.A. (2003). Large sample theory for semiparametric regression models with two-phase, outcome dependent sampling. *Ann. Statist.* **31** 1110–1139.

Breslow, N.E., McNeney, B. and Wellner, J.A. (2000). Large sample theory for semiparametric regression models with two-phase, outcome dependent sampling. Technical Report 381, Dept. Statistics, Univ. Washington.

Breslow, N.E., Robins, J.M. and Wellner, J.A. (2000). On the semi-parametric efficiency of logistic regression under case-control sampling. *Bernoulli* **6** 447–455.

Dudley, R.M. (1989). *Real analysis and probability.* Pacific Grove, California.

Godambe, V.P. (1991). Orthogonality of estimating functions and nuisance parameters. *Biometrika* **78** 143–151.

Hall, W.S. and Newell, M.L. (1979). The mean value theorem for vector valued functions: a simple proof. *Mathematics Magazine* **52** 157–158.

Hirose, Y. (2005). Efficiency of the semi-parametric maximum likelihood estimator in generalized case-control studies. Ph.D. dissertation, Univ. Auckland.

Lawless, J.L., Kalbfleish, J.D. and Wild, C.J. (1999). Semiparametric methods for response-selective and missing data problems in regression. *J. Roy. Statist. Soc. Ser. B* **61** 413–438.

Lee, A.J. (2004). Semi-parametric efficiency bounds for regression models under choice-based sampling. Unpublished manuscript, Univ. Auckland.

Lee, A.J. and Hirose, Y. (2005). Semi-parametric efficiency bounds for regression models under case-control sampling: the profile likelihood approach. Unpublished manuscript, Univ. Auckland.

McLeish, D. L. and Small, C. G. (1992). A projected likelihood function for semiparametric models. *Biometrika* **79** 93–102.

McNeney, W.B. (1998). Asymptotic efficiency in semiparametric models with non-i.i.d. data. Ph.D. dissertation, Univ. Washington.

Murphy, S.A. and van der Vaart, A.W. (1999). Observed information in semi-parametric models. *Bernoulli* **5** 381–412.

Murphy, S.A. and van der Vaart, A.W. (2000). On profile likelihood (with discussion). *J. Amer. Statist. Assoc.* **95** 449–485.

Nan, B., Emond, M. and Wellner, J.A. (2004). Information bounds for cox regression models with missing data. *Ann. Statist.* **32** 723–753.

Newey, W.K. (1990). Semi-parametric efficiency bounds. *J. Appl. Econ* **5** 99–135.

Newey, W.K. (1994). The asymptotic variance of semi-parametric estimators. *Econometrica* **62** 1349–1382.

Pollard, D. (2002). *A user's guide to measure theoretic probability.* Cambridge Univ. Press, Cambridge.

Prentice, R.L. and Pyke, R. (1979). Logistic disease incidence models and case-control studies. *Biometrika* **66** 403–411.

Robins, J.M., Hsieh, F. and Newey, W.K. (1995). Semiparametric efficient estimation of a conditional density with missing or mismeasured covariates. *J. Roy. Statist. Soc. Ser. B* **57** 409–424.

Robins, J.M., Rotnitzky, A. and Zhao, L.P. (1994). Estimation of regression coefficients when some regressors are not always observed. *J. Amer. Statist. Assoc.* **89** 846–866.

Seber, G.A.F. and Lee, A.J. (2003). *Linear Regression Analysis, Second Edition.* Wiley, New York.

Scott, A.J. and Wild, C.J. (1997). Fitting regression models to case-control data by maximum likelihood. *Biometrika* **84** 57–71.

Scott, A.J. and Wild, C.J. (2001). Maximum likelihood for generalised case-control studies. *J. Stat. Plann. Inference* **96** 3–27.

Tsiatis, A.B. (2006). *Semiparametric Theory and Missing Data.* Springer, New York.

van de Geer, S.A. (2000). *Empirical Processes in M-Estimation.* Cambridge Univ. Press, Cambridge.

van der Vaart, A.W. and Wellner, J.A. (1996). *Weak Convergence and Empirical Processes.* Springer, New York.

van der Vaart, A.W. (1998). *Asymptotic Statistics.* Cambridge Univ. Press, Cambridge.

van der Vaart, A.W. and Wellner, J.A. (2000). Preservation theorems for Glivenko-Cantelli and uniform Glivenko-Cantelli class, *High Dimensional Probability, vol. II*, 115-134, (Eds: E. Gine, D.M. Mason, J.A. Wellner). Birkhuser, Boston.