Ann Inst Stat Math (2011) 63:1247–1275
DOI 10.1007/s10463-010-0280-y

Efficiency of profile likelihood in semi-parametric models

Yuichi Hirose

Received: 26 February 2007 / Revised: 18 November 2009 / Published online: 31 March 2010

Abstract Profile likelihood is a popular method of estimation in the presence of an infinite-dimensional nuisance parameter, as 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 parametrization, we improve on the seminal work of Murphy and van der Vaart (*J Am Stat Assoc*, 95: 449–485, 2000): our improvement establishes the efficiency of the estimator through the direct quadratic expansion of the profile likelihood, which requires fewer assumptions. To illustrate the method an application to two-phase outcome-dependent sampling design is given.

Keywords Semi-parametric model · Profile likelihood · Two-phase outcome-dependent sampling · Efficiency · M-estimator · Maximum likelihood estimator · Efficient score · Efficient information bound

1 Introduction

The method of profile likelihood estimation is a familiar methodology in the presence of a nuisance parameter. It is particularly useful when dealing with applications to semi-parametric models, since the method to calculate the semi-parametric maximum likelihood estimator (MLE) is based on a particular representation of the profile likelihood. For example, Scott and Wild (1997, 2001) use profile likelihood to calculate the semi-parametric maximum likelihood estimator for data from variations of case–control studies.

Y. Hirose (✉)
School of Mathematics, Statistics and Operations Research,
Victoria University of Wellington, Wellington, New Zealand
e-mail: Yuichi.Hirose@msor.vuw.ac.nz

The purpose of this paper is to investigate the efficiency of the 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 sub-model 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.

We take an alternative approach to Murphy and van der Vaart (2000), so that we can treat the expansion of the profile likelihood directly. Suppose we consider a semi-parametric model of the form

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

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 $\mathcal{B}$. The expectation with respect to the density $p(x; \beta, \eta)$ is denoted by $E_{\beta,\eta}$.

We slightly extend the usual definition of profile likelihood as follows: suppose there exists a function $\hat{\eta}(\beta, F)$ of the parameter of interest β and a cdf F such that $\hat{\eta}(\beta_0, F_0) = \eta_0$ and the derivative

$$\left. \frac{\partial}{\partial \beta} \right|_{\beta=\beta_0} \log p(x; \beta, \hat{\eta}(\beta, F_0)) \tag{1}$$

is the efficient score function where F_0 is the cdf for the density $p(x; \beta_0, \eta_0)$. Then for the empirical cdf F_n, the *profile log-likelihood* for β is defined by

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

Under mild regularity conditions with the assumption that

$$\hat{\eta}(\beta) = \mathrm{argmax}_{\eta \in \Theta_\eta} E_{\beta_0,\eta_0} \log p(X; \beta, \eta) \tag{2}$$

exists for all β in some neighborhood of β_0, the usual profile likelihood is a special case of the extended one. To see this, define $\hat{\eta}(\beta, F) = \mathrm{argmax}_{\eta \in \Theta_\eta} \int \log p(x; \beta, \eta) \mathrm{d}F$. Then the relations

$$\frac{1}{n} \sum_{i=1}^{n} \log p(X_i; \beta, \eta) = \int \log p(x; \beta, \eta) \mathrm{d}F_n$$

and

$$E_{\beta_0,\eta_0} \log p(X;\beta,\eta) = \int \log p(x;\beta,\eta)\mathrm{d}F_0$$

imply that $\hat{\eta}_n(\beta) = \hat{\eta}(\beta, F_n)$ and $\hat{\eta}(\beta) = \hat{\eta}(\beta, F_0)$, respectively. In this situation, Eq. 1 is the efficient score function by Newey (1994).

The main purpose of this paper is to introduce the function $\hat{\eta}(\beta, F)$ as having an additional parameter F so that the estimating equation based on the profile likelihood can be expressed as

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

This gives an estimating function which is an explicit function of sample size n, through F_n. Then, we show that the solution $\hat{\beta}_n$ to the estimating equation is efficient. Moreover, the assumption of Conditions (10) and (11) in Murphy and van der Vaart (2000) can be avoided in our approach.

An outline of this paper is as follows: Sect. 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, and in Sect. 3 a two-phase outcome-dependent sampling design is used as an example.

2 Main result

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

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

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

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 Sect. 2.1.2 assumes 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.

The main results of this paper are presented below.

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

$$\sqrt{n} E_{\beta_0,\eta_0} \phi(X, \beta_0, F_n) = o_P(1) \tag{4}$$

and

$$\begin{aligned} \sum_{i=1}^{n} \log p^*(X_i; \tilde{\beta}_n, F_n) = \sum_{i=1}^{n} \log p^*(X_i; \beta_0, F_n) + (\tilde{\beta}_n - \beta_0)^T \sum_{i=1}^{n} \phi(X_i, \beta_0, F_0) \\ + \frac{1}{2} n (\tilde{\beta}_n - \beta_0)^T I^*_\beta (\tilde{\beta}_n - \beta_0) \\ + o_{P_{\beta_0,\eta_0}}(\sqrt{n}\|\tilde{\beta}_n - \beta_0\| + 1)^2 \end{aligned} \tag{5}$$

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*

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

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 Hadamard differentiability

We say that a map $\psi : B_1 \to B_2$ between two Banach spaces B_1 and B_2 is Hadamard differentiable at x if there is a continuous linear map $d\psi(x) : B_1 \to B_2$ such that

$$\frac{\psi(x + th') - \psi(x)}{t} \to d\psi(x)(h) \quad \text{as } t \to 0 \text{ and } h' \to h.$$

The map $d\psi(x)$ is called derivative of ψ at x, and is continuous in x. [For reference, see Gill (1989) and Shapiro (1990).]

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 Hadamard 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 Hadamard 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 exists 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 Eq. 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

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

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),

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

Alternatively, suppose condition (R1)*. We modify the proof of the mean value theorem for the 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

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

$$\leq \sup_{t\in[0,1]} \left\| \sqrt{n} E_{\beta_0,\eta_0} \frac{\partial}{\partial t} \phi\left(X, \beta_0, F_n^*(0)\right) + \sqrt{n} E_{\beta_0,\eta_0} \frac{\partial^2}{\partial t^2} \phi\left(X, \beta_0, F_n^*(t)\right) \right\|$$

(by the Cauchy–Schwarz inequality)

$$= \sup_{t\in[0,1]} \| E_{\beta_0,\eta_0} d_F \phi(X, \beta_0, F_0) \sqrt{n}(F_n - F_0)$$

$$+ E_{\beta_0,\eta_0} d_F^2 \phi(X, \beta_0, F_n^*(t)) \sqrt{n}(F_n - F_0)^2 \|$$

(by the definition of path-wise differentiability)

$$= \sup_{t\in[0,1]} \| E_{\beta_0,\eta_0} d_F^2 \phi(X, \beta_0, F_n^*(t)) \sqrt{n}(F_n - F_0)^2 \| \quad \text{(by Eq. 7)}$$

$$\leq \left\| E_{\beta_0,\eta_0} d_F^2 \phi(X, \beta_0, F_0) + o_p(1) \right\| \sqrt{n} \, \|F_n - F_0\|^2$$

$$= o_P(1) \quad (\text{since } \sqrt{n}\|F_n - F_0\|^2 = o_P(1)).$$

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

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

and

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

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 $(\beta_n^*, F_n^*) \xrightarrow{P} (\beta_0, F_0)$,

$$\frac{1}{\sqrt{n}} \sum_{i=1}^{n} \left\{ \phi\left(X_i, \beta_n^*, F_n^*\right) - \phi\left(X_i, \beta_0, F_0\right) \right\}$$

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

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

By combining Eqs. 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 $\|\beta_n^*-\beta_0\|\le\|\tilde{\beta}_n-\beta_0\|$,

$$\begin{aligned}
&\sum_{i=1}^{n}\log p^*(X_i;\tilde{\beta}_n,F_n)-\sum_{i=1}^{n}\log p(X_i;\beta_0,F_n)\\
&=\sqrt{n}(\tilde{\beta}_n-\beta_0)^T\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\phi(X_i,\beta_0,F_n)\\
&\quad+\frac{1}{2}n(\tilde{\beta}_n-\beta_0)^T\frac{1}{n}\sum_{i=1}^{n}\frac{\partial}{\partial\beta}\phi\left(X_i,\beta_n^*,F_n\right)(\tilde{\beta}_n-\beta_0)\\
&=\sqrt{n}(\tilde{\beta}_n-\beta_0)^T\left\{\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\phi(X_i,\beta_0,F_0)+o_P(1)\right\}\quad\text{(by Eq. 10)}\\
&\quad+\frac{1}{2}n(\tilde{\beta}_n-\beta_0)^T\left\{-E_{\beta_0,\eta_0}\phi\phi^T(X,\beta_0,F_0)+o_P(1)\right\}(\tilde{\beta}_n-\beta_0)\quad\text{(by Eq. 9)}\\
&=\sqrt{n}(\tilde{\beta}_n-\beta_0)^T\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\phi(X_i,\beta_0,F_0)-\frac{1}{2}n(\tilde{\beta}_n-\beta_0)^TE_{\beta_0,\eta_0}(\phi\phi^T)(\tilde{\beta}_n-\beta_0)\\
&\quad+o_P(\sqrt{n}\|\tilde{\beta}_n-\beta_0)\|+1)^2.
\end{aligned}$$

This proves Eq. 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 et al. (2000a) which was 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 $\phi(x, \beta_0, F_0) = \frac{\partial}{\partial\beta}\big|_{\beta=\beta_0} \log p(x; \beta, \hat{\eta}(\beta, F_0))$ *is the efficient score function.*

Proof Condition (12) implies that

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

By differentiating the identity

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

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

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

Let $c \in \mathbb{R}^m$ be arbitrary. Then, it follows from Eq. 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 $\frac{\partial}{\partial t}\big|_{t=0} \log p(x; \beta_0, \eta(t))$. Using Condition (11), we have

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

where $\dot{\ell}_\beta(x; \beta_0, \eta_0) = \frac{\partial}{\partial\beta}\big|_{\beta=\beta_0} \log p(x; \beta, \eta_0)$ and $\psi(x; \beta_0, \eta_0) = -\frac{\partial}{\partial\beta}\big|_{\beta=\beta_0} \log p(x; \beta_0, \hat{\eta}(\beta, F_0))$. 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 Murphy and van der Vaart (2000)

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

$$\beta \to \hat{\eta}(\beta, \beta', \eta') \tag{15}$$

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 Eq. (2). They also assume that, for any $\beta^*_n \xrightarrow{P} \beta_0$:

(P4) $\hat{\eta}_n(\beta^*_n) \xrightarrow{P} \eta_0$ [Condition (10) in Murphy and van der Vaart (2000)];
(P5) $E_{\beta_0,\eta_0}\dot{\ell}^*_\beta(x, \beta_0, \beta^*_n, \hat{\eta}_n(\beta^*_n)) = o_P(\|\beta^*_n - \beta_0\| + 1/\sqrt{n})$ [Condition (11) in Murphy and van der Vaart (2000)];
(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

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

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.

2.4 Comments on NPMLE

Kiefer and Wolfowitz (1956) defined a non-parametric maximum likelihood estimator (NPMLE) $\hat{\theta}$ based on data $X_1, \ldots, X_n$ from the model $\{P_\theta : \theta \in \Theta\}$, where Θ may be infinite-dimensional and P_θ is the distribution of $X_1, \ldots, X_n$ under θ, if

$$\frac{dP_{\hat{\theta}}}{d(P_{\hat{\theta}} + P_\theta)}(X_1, \ldots, X_n) \geq \frac{dP_\theta}{d(P_{\hat{\theta}} + P_\theta)}(X_1, \ldots, X_n)$$

for all $\theta \in \Theta$.

A NPMLE defined above often does not exist in the original parameter space so the model needs to be extended to include the estimator. For example, the Nelson-Aalen estimator $\hat{\Lambda}(t) = \int_0^t \frac{N(ds)}{Y(s)}$ is a well known NPMLE of the true cumulative hazard function $\Lambda(t) = \int_0^t h(s)ds$. Since the cumulative hazard function $\Lambda(t)$ is absolutely continuous, whereas the estimator $\hat{\Lambda}(t)$ is discrete, we need to extend the model to allow discrete cumulative hazard functions. However, several different discrete extensions of the model can be constructed, each leading to a different NPMLE. In this example it can be shown that, at an underlying continuous point in the model, the estimators are asymptotically equivalent (see Andersen et al. (1993), pp. 221–229).

Motivated by this example, Gill (1989) addressed that searching for "the correct discrete extension" of a given continuous model is a wrong approach. Instead, one should extend score functions from continuous to discrete points in the parameter space in as analytically smooth a way as possible.

The approach in this paper follows the line of Gill's idea with the efficient score function in a semi-parametric model. We assume existence of a function $\hat{\eta}(\beta, F)$ such that the efficient score function in the original model is given by Eq. (1) and its discrete extension is given by

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

where the empirical cdf F_n is in the set $\mathcal{F}$ of cdf functions on the sample space $\mathcal{X}$. The result of the paper implies that, under the conditions (R0–R3) given previously, estimators $\hat{\beta}_n$ based on discrete extensions of this form are asymptotically linear estimators with the efficient influence function, therefore, they are asymptotically equivalent and efficient.

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 et al. (1999). Breslow et al. (2003) used the approach of Murphy and van der Vaart (2000) to demonstrate the efficiency of the estimator based on the profile likelihood in a variation of this example. In contrast, we apply Theorem 1 to show the same result.

Remark 3.1 Theorem 1 does not require verification of (P4) and (P5) [Conditions (10) and (11) in Murphy and van der Vaart (2000)]. See Breslow et al. (2003) and van der Vaart and Wellner (2001) for the verification of (P4) and (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 $\Theta \subset \mathbb{R}^k$ 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 a discrete or continuous variable.

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

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

and

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

In each stratum $s = 1, \ldots, s$, let m_s be the number of fully observed units, and $n_s - m_s$ be the number of subjects whose only information retained is the identity of the stratum. Lawless et al. (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

$$\begin{aligned} L(\theta, g) &= \prod_{s=1}^{S} \left\{ \prod_{i=1}^{m_s} f(y_{si}|x_{si}; \theta) g(x_{si}) \right\} Q_s(\theta, g)^{n_s - m_s} \\ &= \prod_{s=1}^{S} \left\{ \prod_{i=1}^{m_s} \frac{f(y_{si}|x_{si}; \theta) g(x_{si})}{Q_s(\theta, g)} \right\} Q_s(\theta, g)^{n_s}. \end{aligned} \tag{16}$$

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, Eq. 16, as the one for an i.i.d. sample from the density

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

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

$$\int \log p(s,z;\theta,g)\mathrm{d}F(s,z)$$
$$= \sum_{s=1}^{S}\left\{a_s\int(\log f(y|x;\theta)+\log g(x))\,\mathrm{d}F_s + (a_{S+1}\mathrm{d}F_{S+1}(s)-a_s)\log Q_s(\theta,g)\right\}. \tag{17}$$

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

$$\int \log p(s,z;\theta,g)\mathrm{d}F_0(s,z)$$
$$= \sum_{s=1}^{S}\{w_{S+1}Q_s(\theta_0,g_0)\log Q_s(\theta,g) + w_s E_{s,\theta_0,g_0}[\log f(Y|X;\theta)+\log g(X)-\log Q_s(\theta,g)]\}$$

and

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

Theorem A [The efficient score function] *Let*

$$\hat{g}(x,\theta,F) = \frac{f^*(x,F)}{\sum_{s=1}^{S} a_s^*(\theta,F)\frac{Q_{s|X}(x;\theta)}{\hat{Q}_s(\theta,F)}}, \tag{18}$$

where

$$f^*(x,F) = \sum_{s=1}^{S} a_s \int \frac{\mathrm{d}F_s}{\mathrm{d}(y,x)}\mathrm{d}y,$$
$$a_s^*(\theta,F) = a_{S+1}[\hat{Q}_s(\theta,F) - \mathrm{d}F_{S+1}(s)] + a_s,$$

and

$$\hat{Q}_s(\theta,F) = \int Q_{s|X}(x;\theta)\hat{g}(x,\theta,F)\mathrm{d}x. \tag{19}$$

Then the efficient score function is given by

$$\phi(s, z, \theta_0, F_0) = \frac{\partial}{\partial\theta}\bigg|_{\theta=\theta_0} \log p(s, z; \theta, \hat{g}(x, \theta, F_0)). \tag{20}$$

The proof of THEOREM A is given in Appendix A.

Remark 3.3 Recall that, for each $s = 1, \ldots, S+1$, F_{s0} 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

$$\begin{aligned} f^*(x, F_0) &= \sum_{s=1}^{S} w_s \int \frac{\mathrm{d}F_{s0}}{\mathrm{d}(y, x)}\mathrm{d}y \\ &= \sum_{s=1}^{S} w_s \int \frac{f(y|x;\theta)1_{(y,x)\in\mathcal{S}_s} g_0(x)}{Q_s(\theta_0, g_0)}\mathrm{d}y \\ &= \sum_{s=1}^{S} w_s \frac{Q_{s|X}(x;\theta_0)}{Q_s(\theta_0, g_0)} g_0(x). \end{aligned} \tag{21}$$

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

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

In addition to the relationships given above, we assume that

(T0) For each $s = 1, \ldots, S$,

$$\frac{Q_s(\theta_0, g_0)}{1 + Q_s(\theta_0, g_0)} < \frac{w_s}{w_{S+1}} < \frac{Q_s(\theta_0, g_0)}{1 - Q_s(\theta_0, g_0)}$$

This condition is used to verify (R1) in Theorem 1 (differentiability of the score function in the induced model). This means, since $\frac{w_s}{w_{S+1}} \approx \frac{m_s}{n}$, if sample size proportions $\frac{m_s}{n}$ differ greatly from the probability $Q_s(\theta_0, g_0)$, then there is no guarantee that the model satisfies Condition (R1).

Remark 3.4 Let $\mathcal{F}$ be the set of cdf's on the sample space and, for $\rho > 0$, let

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

By Eqs. (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 $s = 1, \ldots, S$ 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 3.5 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 verification 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.

The maps $x \to \frac{1}{x}$,

$$g \to f(y|x;\theta) g(x) 1_{(y,x) \in \mathcal{S}_s}$$

and

$$g \to \int Q_{s|X}(x;\theta)g(x)\mathrm{d}x$$

are Hadamard differentiable (cf. Gill 1989). By the chain rule and product rule of Hadamard differentiable maps, the densities

$$p_s(y,x;\theta,g) = \frac{f(y|x;\theta)g(x)1_{(y,x)\in\mathcal{S}_s}}{\int Q_{s|X}(x;\theta)g(x)\mathrm{d}x}$$

and

$$p_{S+1}(i;\theta,g) = Q_i(\theta,g) = \int Q_{i|X}(x;\theta)g(x)\mathrm{d}x$$

are Hadamard differentiable with respect to $g = g(x)$. By continuing a similar argument, we can show that the derivatives of these densities are Hadamard differentiable with respect to $g = g(x)$. By Condition (T2), these densities are twice continuously differentiable with respect to θ. Therefore, to verify condition (R1) all we need is the differentiability of the function $\hat{g}(x,\theta,F)$.

We show the Hadamard differentiability of $\hat{g}(x,\theta,F)$ with respect to F. For $\theta \in \mathbb{R}^k$, $F = \sum_{s=1}^{S+1} a_s F_s$ and function $g(x)$, define

$$\Psi_{\theta,F}(g) = \frac{f^*(x,F)}{\sum_{s=1}^{S} a_s^{\#}(\theta,F,g)\frac{Q_{s|X}(x;\theta)}{Q_s(\theta,g)}},$$

where $a_s^{\#}(\theta,F,g) = a_{S+1}[Q_s(\theta,g) - dF_{S+1}(s)] + a_s$. Then, by Eqs. (18) and (19), the function $\hat{g}(x,\theta,F)$ is the solution to the equation

$$g(x) = \Psi_{\theta,F}(g)(x). \tag{24}$$

Suppose a map $t \to F_t$ is such that $F_t \to F$ and $t^{-1}(F_t - F) \to h$ as $t \downarrow 0$. Since $f^*(x,F) = \sum_{s=1}^{S} a_s \int \frac{\mathrm{d}F_s}{\mathrm{d}(y,x)}\mathrm{d}y = \int \frac{\mathrm{d}F}{\mathrm{d}(y,x)}\mathrm{d}y$ is linear with respect to F,

$$t^{-1}\{f^*(x,F_t) - f^*(x,F)\} \to \int \frac{\mathrm{d}h}{\mathrm{d}(y,x)}\mathrm{d}y \quad \text{as} \quad t \downarrow 0.$$

Let $\pi_s : F = \sum_{s'=1}^{S+1} a_{s'}F_{s'} \to a_s F_s$ be a projection, then

$$\begin{aligned} t^{-1}\left\{a_s^{\#}(\theta,F_t,g) - a_s^{\#}(\theta,F,g)\right\} &= -a_{S+1}d\pi_{S+1}t^{-1}\{F_t - F\}(s) \\ &\to -a_{S+1}d\pi_{S+1}h(s) \quad \text{as} \quad t \downarrow 0. \end{aligned}$$

It follows that

$$
\begin{aligned}
&t^{-1}\left\{f^*(x,F_t)\sum_{s=1}^{S}a_s^{\#}(\theta,F,g)\frac{Q_{s|X}(x;\theta)}{Q_s(\theta,g)}-f^*(x,F)\sum_{s=1}^{S}a_s^{\#}(\theta,F_t,g)\frac{Q_{s|X}(x;\theta)}{Q_s(\theta,g)}\right\}\\
&\quad=t^{-1}\left\{f^*(x,F_t)-f^*(x,F)\right\}\sum_{s=1}^{S}a_s^{\#}(\theta,F,g)\frac{Q_{s|X}(x;\theta)}{Q_s(\theta,g)}\\
&\qquad-f^*(x,F)\sum_{s=1}^{S}t^{-1}\left\{a_s^{\#}(\theta,F_t,g)-a_s^{\#}(\theta,F,g)\right\}\frac{Q_{s|X}(x;\theta)}{Q_s(\theta,g)}\\
&\quad\to\int\frac{\mathrm{d}h}{\mathrm{d}(y,x)}\mathrm{d}y\sum_{s=1}^{S}a_s^{\#}(\theta,F,g)\frac{Q_{s|X}(x;\theta)}{Q_s(\theta,g)}\\
&\qquad+f^*(x,F)\sum_{s=1}^{S}a_{S+1}d\pi_{S+1}h(s)\frac{Q_{s|X}(x;\theta)}{Q_s(\theta,g)},
\end{aligned}
$$

as $t\downarrow 0$. Altogether, as $t\downarrow 0$,

$$
\begin{aligned}
&t^{-1}\left\{\Psi_{\theta,F_t}(g)-\Psi_{\theta,F}(g)\right\}\\
&\quad=t^{-1}\left\{\frac{f^*(x,F_t)}{\sum_{s=1}^{S}a_s^{\#}(\theta,F_t,g)\frac{Q_{s|X}(x;\theta)}{Q_s(\theta,g)}}-\frac{f^*(x,F)}{\sum_{s=1}^{S}a_s^{\#}(\theta,F,g)\frac{Q_{s|X}(x;\theta)}{Q_s(\theta,g)}}\right\}\\
&\quad=t^{-1}\frac{f^*(x,F_t)\sum_{s=1}^{S}a_s^{\#}(\theta,F,g)\frac{Q_{s|X}(x;\theta)}{Q_s(\theta,g)}-f^*(x,F)\sum_{s=1}^{S}a_s^{\#}(\theta,F_t,g)\frac{Q_{s|X}(x;\theta)}{Q_s(\theta,g)}}{\sum_{s=1}^{S}a_s^{\#}(\theta,F_t,g)\frac{Q_{s|X}(x;\theta)}{Q_s(\theta,g)}\sum_{s=1}^{S}a_s^{\#}(\theta,F,g)\frac{Q_{s|X}(x;\theta)}{Q_s(\theta,g)}}\\
&\quad\to d_F\Psi_{\theta,F}(g)h
\end{aligned}
\tag{25}
$$

where the map $d_F\Psi_{\theta,F}(g)$ is given by

$$
\begin{aligned}
&d_F\Psi_{\theta,F}(g)h\\
&\quad=\frac{\int\frac{\mathrm{d}h}{\mathrm{d}(y,x)}\mathrm{d}y\sum_{s=1}^{S}a_s^{\#}(\theta,F,g)\frac{Q_{s|X}(x;\theta)}{Q_s(\theta,g)}+f^*(x,F)\sum_{s=1}^{S}a_{S+1}d\pi_{S+1}h(s)\frac{Q_{s|X}(x;\theta)}{Q_s(\theta,g)}}{\left\{\sum_{s=1}^{S}a_s^{\#}(\theta,F,g)\frac{Q_{s|X}(x;\theta)}{Q_s(\theta,g)}\right\}^2}.
\end{aligned}
$$

Hence, the map $F\to\Psi_{\theta,F}(g)$ is Hadamard differentiable at (θ,g,F) with derivative $d_F\Psi_{\theta,F}(g)$ (clearly, the derivative is linear in h, we omit the proof of boundedness of $d_F\Psi_{\theta,F}(g)$).

Now, suppose a map $t \to g_t$ is such that $g_t \to g$ and $t^{-1}(g_t - g) \to h'$ as $t \downarrow 0$. As $t \downarrow 0$,

$$t^{-1}\{Q_s(\theta, g_t) - Q_s(\theta, g)\} = \int Q_{s|X}(x;\theta)t^{-1}\{g_t(x) - g(x)\}\mathrm{d}x \to \int Q_{s|X}(x;\theta)h'(x)\mathrm{d}x$$

and

$$t^{-1}\left\{a_s^{\#}(\theta, F, g_t) - a_s^{\#}(\theta, F, g)\right\} = a_{S+1}t^{-1}\{Q_s(\theta, g_t) - Q_s(\theta, g)\} \to a_{S+1}\int Q_{s|X}(x;\theta)h'(x)\mathrm{d}x.$$

Hence,

$$\begin{aligned}
&t^{-1}\left\{\sum_{s=1}^{S} a_s^{\#}(\theta, F, g)\frac{Q_{s|X}(x;\theta)}{Q_s(\theta, g)} - \sum_{s=1}^{S} a_s^{\#}(\theta, F, g_t)\frac{Q_{s|X}(x;\theta)}{Q_s(\theta, g_t)}\right\}\\
&= -\sum_{s=1}^{S} Q_{s|X}(x;\theta)\\
&\quad\times\frac{t^{-1}\left\{a_s^{\#}(\theta, F, g_t) - a_s^{\#}(\theta, F, g)\right\}Q_s(\theta, g) - a_s^{\#}(\theta, F, g)t^{-1}\{Q_s(\theta, g_t) - Q_s(\theta, g)\}}{Q_s(\theta, g_t)Q_s(\theta, g)}\\
&\to \sum_{s=1}^{S}\frac{\left\{a_s^{\#}(\theta, F, g) - a_{S+1}Q_s(\theta, g)\right\}Q_{s|X}(x;\theta)\int Q_{s|X}(x;\theta)h'(x)\mathrm{d}x}{(Q_s(\theta, g))^2}.
\end{aligned}$$

It follows that, as $t \downarrow 0$,

$$\begin{aligned}
&t^{-1}\left\{\Psi_{\theta,F}(g_t) - \Psi_{\theta,F}(g)\right\}\\
&= t^{-1}\left\{\frac{f^*(x, F)}{\sum_{s=1}^{S} a_s^{\#}(\theta, F, g_t)\frac{Q_{s|X}(x;\theta)}{Q_s(\theta,g_t)}} - \frac{f^*(x, F)}{\sum_{s=1}^{S} a_s^{\#}(\theta, F, g)\frac{Q_{s|X}(x;\theta)}{Q_s(\theta,g)}}\right\}\\
&= \frac{f^*(x, F)t^{-1}\left\{\sum_{s=1}^{S} a_s^{\#}(\theta, F, g)\frac{Q_{s|X}(x;\theta)}{Q_s(\theta,g)} - \sum_{s=1}^{S} a_s^{\#}(\theta, F, g_t)\frac{Q_{s|X}(x;\theta)}{Q_s(\theta,g_t)}\right\}}{\sum_{s=1}^{S} a_s^{\#}(\theta, F, g_t)\frac{Q_{s|X}(x;\theta)}{Q_s(\theta,g_t)}\sum_{s=1}^{S} a_s^{\#}(\theta, F, g)\frac{Q_{s|X}(x;\theta)}{Q_s(\theta,g)}}\\
&\to d_g\Psi_{\theta,F}(g)h' \qquad (26)
\end{aligned}$$

where

$$d_g\Psi_{\theta,F}(g)h' = \frac{f^*(x, F)\sum_{s=1}^{S}\frac{\left\{a_s^{\#}(\theta,F,g) - a_{S+1}Q_s(\theta,g)\right\}}{Q_s(\theta,g)}\frac{Q_{s|X}(x;\theta)}{Q_s(\theta,g)}\int Q_{s|X}(x;\theta)h'(x)\mathrm{d}x}{\left\{\sum_{s=1}^{S} a_s^{\#}(\theta, F, g)\frac{Q_{s|X}(x;\theta)}{Q_s(\theta,g)}\right\}^2}.$$

Since the limit is linear in h', the map $g \to \Psi_{\theta,F}(g)$ is Hadamard differentiable provided the map $d_g\Psi_{\theta,F}(g)$ is bounded. Next, we show the boundedness of the derivative $d_g\Psi_{\theta,F}(g)$. Let L_1 be the space of all real valued measurable functions $h(x)$ with $\|h\|_1 = \int |h(x)|\mathrm{d}x < \infty$. Then L_1 is a Banach space with norm $\|\cdot\|_1$.

By Eq. (21)

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

It follows that, at $(\theta, g, F) = (\theta_0, g_0, F_0)$, we have

$$d_g\Psi_{\theta_0,F_0}(g_0)h' = g_0(x)\frac{\sum_{s=1}^S \frac{\{w_s - w_{S+1}Q_s(\theta_0,g_0)\}}{Q_s(\theta_0,g_0)}\frac{Q_{s|X}(x;\theta_0)}{Q_s(\theta_0,g_0)}\int Q_{s|X}(x;\theta_0)h'(x)\mathrm{d}x}{\sum_{s=1}^S w_s \frac{Q_{s|X}(x;\theta_0)}{Q_s(\theta_0,g_0)}}$$

and

$$\|d_g\Psi_{\theta_0,F_0}(g_0)h'\|_1 < \|h'\|_1,$$

where we used $\|g_0\|_1 = 1$,

$$\int Q_{s|X}(x;\theta_0)\frac{|h'(x)|}{\|h'\|_1}\mathrm{d}x \le 1,$$

and by Condition (T0) in Remark 3.3,

$$\frac{Q_s(\theta_0,g_0)}{1+Q_s(\theta_0,g_0)} < \frac{w_s}{w_{S+1}} < \frac{Q_s(\theta_0,g_0)}{1-Q_s(\theta_0,g_0)} \Rightarrow \left|\frac{w_s - w_{S+1}Q_s(\theta_0,g_0)}{Q_s(\theta_0,g_0)}\right| < w_s.$$

Thus the linear map $d_g\Psi_{\theta_0,F_0}(g_0) : L_1 \to L_1$ has the operator norm $\|d_g\Psi_{\theta_0,F_0}(g_0)\| < 1$. We assume that

(T3) There is a neighborhood $\Theta \times \mathcal{G} \times \mathcal{C}_\rho$ of (θ_0, g_0, F_0) such that $\|d_g\Psi_{\theta,F}(g)\| < 1$ for all $(\theta, g, F) \in \Theta \times \mathcal{G} \times \mathcal{C}_\rho$.

Let $I : L_1 \to L_1$ be the identity operator. In the neighborhood $\Theta \times \mathcal{G} \times \mathcal{C}_\rho$ the map $(I - d_g\Psi_{\theta,F}(g)) : L_1 \to L_1$ has the inverse $(I - d_g\Psi_{\theta,F}(g))^{-1}$, which is also a bounded linear map (cf. Kolmogorov and Fomin 1975, Theorem 4 on p. 231). In the following, we assume the inverse $(I - d_g\Psi_{\theta,F}(g))^{-1}$ exists.

By Eqs. 24, 25 and 26, as $t \downarrow 0$,

$$\begin{aligned}
t^{-1}\{\hat{g}(x,\theta,F_t) - \hat{g}(x,\theta,F)\} &= t^{-1}\{\Psi_{\theta,F_t}(\hat{g}(x,\theta,F_t)) - \Psi_{\theta,F}(\hat{g}(x,\theta,F))\} \\
&= t^{-1}\{\Psi_{\theta,F_t}(\hat{g}(x,\theta,F_t)) - \Psi_{\theta,F}(\hat{g}(x,\theta,F_t))\} \\
&\quad + t^{-1}\{\Psi_{\theta,F}(\hat{g}(x,\theta,F_t)) - \Psi_{\theta,F}(\hat{g}(x,\theta,F))\} \\
&= d_F\Psi_{\theta,F}(\hat{g}(x,\theta,F))h + d_g\Psi_{\theta,F}(\hat{g}(x,\theta,F)) \\
&\quad \times t^{-1}\{\hat{g}(x,\theta,F_t)) - \hat{g}(x,\theta,F)\} + o(1).
\end{aligned}$$

It follows that

$$[I - d_g \Psi_{\theta,F}(\hat{g}(x,\theta,F))]t^{-1}\{\hat{g}(x,\theta,F_t)) - \hat{g}(x,\theta,F)\} = d_F \Psi_{\theta,F}(\hat{g}(x,\theta,F))h + o(1)$$

and

$$t^{-1}\{\hat{g}(x,\theta,F_t)) - \hat{g}(x,\theta,F)\} \to [I - d_g \Psi_{\theta,F}(\hat{g}(x,\theta,F))]^{-1} d_F \Psi_{\theta,F}(\hat{g}(x,\theta,F))h$$

as $t \downarrow 0$. Since the map $[I - d_g \Psi_{\theta,F}(\hat{g}(x,\theta,F))]^{-1} d_F \Psi_{\theta,F}(\hat{g}(x,\theta,F))$ is bounded and linear, the function $\hat{g}(x,\theta,F)$ is Hadamard differentiable with respect to F. Similarly, we can show that differentiability of the function $\hat{g}(x,\theta,F)$ with respect to θ.

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

$$\log p(s,z;\theta,\hat{g}(x,\theta,F)) = \{1_{s=S+1}1_{i\in\{1,\ldots,S\}} - 1_{s\in\{1,\ldots,S\}}1_{i=s}\}\log \hat{Q}_i(\theta,F) + 1_{s\in\{1,\ldots,S\}}\left\{\log f(y|x;\theta) + \log \hat{g}(x,\theta,F)\right\}. \quad (27)$$

The induced score function is

$$\begin{aligned}\phi(s,z,\theta,F) &= \frac{\partial}{\partial\theta}\log p(s,z;\theta,\hat{g}(x,\theta,F)) \\ &= \{1_{s=S+1}1_{i\in\{1,\ldots,S\}} - 1_{s\in\{1,\ldots,S\}}1_{i=s}\}\frac{\dot{\hat{Q}}_{i,\theta}}{\hat{Q}_i}(\theta,F) \\ &\quad + 1_{s\in\{1,\ldots,S\}}\left\{\frac{\dot{f}}{f}(y|x;\theta) + \frac{\dot{\hat{g}}_\theta}{\hat{g}}(x,\theta,F)\right\} \end{aligned} \quad (28)$$

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 derivative of the induced score function with respect to θ is

$$\begin{aligned}\frac{\partial}{\partial\theta}\phi(s,z,\theta,F) &= \{1_{s=S+1}1_{i\in\{1,\ldots,S\}} - 1_{s\in\{1,\ldots,S\}}1_{i=s}\}\left\{\frac{\ddot{\hat{Q}}_{i,\theta}}{\hat{Q}_i} - \left(\frac{\dot{\hat{Q}}_{i,\theta}}{\hat{Q}_i}\right)^2\right\} \\ &\quad + 1_{s\in\{1,\ldots,S\}}\left\{\frac{\ddot{f}}{f} - \left(\frac{\dot{f}}{f}\right)^2 + \frac{\ddot{\hat{g}}_\theta}{\hat{g}} - \left(\frac{\dot{\hat{g}}_\theta}{\hat{g}}\right)^2\right\}. \end{aligned} \quad (29)$$

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

(T4) 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 non-random vector and $\frac{\dot{\hat{g}}_\theta}{\hat{g}}(x, \theta_0, F_0)$ is a function of x. Therefore, by Eq. 28 and assumption (T4), 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 non-singular with the bounded inverse.

Conditions (R3): We assume that

(T5) Envelope functions

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

are integrable;

(T6) 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);

(T7) 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 with respect to 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, p. 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

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

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

Remark 3.6 Conditions (T2), (T3), (T5) and (T7) 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 quite general semi-parametric models. By introducing the function $\hat{\eta}(\beta, F)$, we improve on the seminal work of Murphy and van der Vaart (2000): our improvement establishes the efficiency of the estimator through the direct quadratic expansion of the profile likelihood, which does not require the assumptions (10) and (11) in Murphy and van der Vaart (2000). In some cases, such as the two-phase outcome-dependent sampling example in this paper, when the function $\hat{\eta}(\beta, F)$ in Eq. 1 is implicitly defined, the verification of the differentiability condition (R1) in Theorem 1 may need some work which the approach in Murphy and van der Vaart (2000) does not require (provided they can construct an approximate least favorable sub-model introduced in their paper). The approach in this paper should work on examples such as the biased sampling model in Gilbert (2000), the Cox regression model with right-censored data, and the many examples in Scott and Wild (1997, 2001) and Lawless et al. (1999).

As Murphy and van der Vaart (2000) commented in "Discussion", there are examples for which it is unclear how one can verify Condition (11) in their paper. For future work, we will modify Theorem 1 in the paper for situations in which the function $\hat{\eta}(\beta, F)$ is implicitly defined and apply this method to those examples with which Murphy and van der Vaart (2000) had difficulty.

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 $\int \log p(s, z; \theta, \hat{g}(\theta, F_0))\mathrm{d}F_0(s, z)$ satisfies Conditions (11) and (12) in THEOREM 2 so that the claim follows from 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 $Q_s(\theta, g) = \int Q_{s|X}(x;\theta)g(x)\mathrm{d}x = \sum_{k=1}^K Q_{s|X}(v_k;\theta)g_k$.

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

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

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

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

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

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

This function is of the form in Eq. 18.

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 $\alpha(t) = \alpha(x, t) = g(x, t) - g_0(x)$ and write $\alpha'(x, 0) = \frac{\partial}{\partial t}|_{t=0} \alpha(x, t)$. Then

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

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

where we used

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

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 (Eq. 18) 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 $(\theta, F) \in \Theta \times \mathcal{C}_\rho$,

$$
\delta = \delta \sum_{s=1}^{S} Q_{s|X}(x;\theta) \leq B(x, \theta, F) \leq \frac{1}{\delta} \sum_{s=1}^{S} Q_{s|X}(x;\theta) = \frac{1}{\delta}. \tag{30}
$$

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

$$
\dot{B}_\theta(x, \theta, F) = \frac{\partial}{\partial\theta} B(x, \theta, F) = \sum_{s=1}^{S} \left\{ \dot{a}_{s,\theta}^* \frac{Q_{s|X}}{\hat{Q}_s} + a_s^* \frac{\dot{Q}_{s|X}\hat{Q}_s - Q_{s|X}\dot{\hat{Q}}_{s,\theta}}{\hat{Q}_s^2} \right\} \tag{31}
$$

and

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

Verifying the class $\{\phi(s,z,\theta,F) : (\theta,F)\in\Theta\times\mathcal{C}_\rho\}$ is Donsker:
By assumptions (T1) and (T7), 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 Eq. 30 and Example 2.10.9, p. 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 (T7)) and bounded in $L_1(P_{\theta_0,g_0})$ (assumption (T5)), and the functions $\frac{\dot{\hat{Q}}_{i,\theta}}{\hat{Q}_i}(\theta,F)$, $i=1,\ldots,S$, are bounded non-random continuous functions on the set $\theta\times\mathcal{C}_\rho$ (see (a) below), by Eq. 28 and Example 2.10.7, p. 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 Equations 28 and 29, 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 (T6) 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 $(T6)$, we have

$$\|\dot{a}^*_{s,\theta}(\theta, F)\| \leq \|\dot{\hat{Q}}_{s,\theta}(\theta, F)\| \leq L. \tag{32}$$

Combine assumptions (T1), (T6), Eqs. (31), (32) and $\sum_{s=1}^S Q_{s|X}(x;\theta) = 1$ to get

$$\begin{aligned}
\|\dot{B}_\theta(x,\theta,F)\| &\leq \sum_{s=1}^S \left\{\left\|\dot{a}^*_{s,\theta}\right\| \frac{Q_{s|X}}{\hat{Q}_s} + a^*_s \frac{\|\dot{Q}_{s|X}\|\hat{Q}_s + Q_{s|X}\|\dot{\hat{Q}}_{s,\theta}\|}{\hat{Q}_s^2}\right\} \\
&\leq \sum_{s=1}^S \left\{L\frac{Q_{s|X}}{\delta} + 1\cdot\frac{\|\dot{Q}_{s|X}\|\cdot 1 + Q_{s|X}L}{\delta^2}\right\} \\
&\leq \frac{L}{\delta} + \frac{\int\|\dot{f}(y|x;\theta)\|\mathrm{d}y + L}{\delta^2} \\
&= c_1\int\|\dot{f}(y|x;\theta)\|\mathrm{d}y + c_2
\end{aligned} \tag{33}$$

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,

$$\|\ddot{B}_\theta(x,\theta,F)\| \leq c_1\int\|\ddot{f}(y|x;\theta)\|\mathrm{d}y + c_2\int\|\dot{f}(y|x;\theta)\|\mathrm{d}y + c_3. \tag{34}$$

Since $B(x,\theta,F)\geq\delta > 0$, Eqs. (33) and (34) imply that, for some positive constants c_1, c_2, c_3, c_4,

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

$$\leq c_1 \int \|\ddot{f}(y|x;\theta)\|\mathrm{d}y + c_2 \left(\int \|\dot{f}(y|x;\theta)\|\mathrm{d}y \right)^2$$
$$+c_3 \int \|\dot{f}(y|x;\theta)\|\mathrm{d}y + c_4,$$

and

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

(by the Cauchy–Schwarz inequality). By assumption (T5), we have condition (b).

References

Andersen, P. K., Borgan, O., Gill, R. D., Keiding, N. (1993). *Statistical models based on counting processes.* New York: Springer.

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

Bickel, P. J., Klaassen, C. A. J., Ritov, Y., Wellner, J. A. (1993). *Efficient and adaptive estimation for semiparametric models.* Baltimore: Johns Hopkins University Press.

Breslow, N. E., Chatterjee, N. (1999). Design and analysis of two-phase studies with binary outcome applied to Wilms tumor prognosis. *Journal of the Royal Statistical Society: Series C, 48*, 457–468.

Breslow, N. E., Holubkov, R. (1997). Maximum likelihood estimation of logistic regression parameters under two-phase outcome-dependent sampling. *Journal of the Royal Statistical Society: Series B, 59*, 447–461.

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

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

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

Gilbert, P. G. (2000). Large sample theory of maximum likelihood estimation in semiparametric selection biased sampling models. *The Annals of Statistics, 28*, 151–194.

Gilbert, P. G., Lele, S. R., Vardi, Y. (1999). Maximum likelihood estimation in semiparametric selection bias models with application to AIDS vaccine trials. *Biometrika, 86*, 27–43.

Gill, R. D. (1989). Non- and semi-parametric maximum likelihood estimators and the von Mises method (part 1). *Scandinavian Journal of Statistics, 16*, 97–128.

Gill, R. D., Vardi, Y., Wellner, J. A. (1988). Large sample theory of empirical distribution in biased sampling models. *The Annals of Statistics, 3*, 1069–1112.

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

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

Kiefer, J., Wolfowitz, J. (1956). Consistency of the maximum likelihood estimator in the presence of infinitely many incidental parameters. *Annals of Mathematical Statistics, 27*, 887–906.

Kolmogorov, A. N., Fomin, S. V. (1975). *Introductory real analysis.* New York: Dover.

Lawless, J. L., Kalbfleish, J. D., Wild, C. J. (1999). Semiparametric methods for response-selective and missing data problems in regression. *Journal of the Royal Statistical Society: Series B, 61*, 413–438.

Lee, A. J., Hirose, Y. (2008). Semi-parametric efficiency bounds for regression models under case–control sampling: The profile likelihood approach. *Annals of the Institute of Statistical Mathematics*. doi:10.1007/s10463-008-0205-1.

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

Murphy, S. A., van der Vaart, A. W. (2000). On profile likelihood (with discussion). *Journal of the American Statistical Association, 95*, 449–485.

Newey, W. K. (1990). Semi-parametric efficiency bounds. *Journal of Applied Economics, 5*, 99–135.

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

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

Robins, J. M., Rotnitzky, A., Zhao, L. P. (1994). Estimation of regression coefficients when some regressors are not always observed. *Journal of the American Statistical Association, 89*, 846–866.

Robins, J. M., Hsieh, F., Newey, W. K. (1995). Semiparametric efficient estimation of a conditional density with missing or mismeasured covariates. *Journal of the Royal Statistical Society: Series B, 57*, 409–424.

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

Scott, A. J., Wild, C. J. (2001). Maximum likelihood for generalised case–control studies. *Journal of Statistical Planning and Inference, 96*, 3–27.

Seber, G. A. F., Lee, A. J. (2003). *Linear regression analysis* (2nd ed.). New York: Wiley.

Shapiro, A. (1990). On concepts of directional differentiability. *Journal of Optimization Theory and Applications, 66*, 477–487.

Tsiatis, A. B. (2006). *Semiparametric theory and missing data*. New York: Springer.

van de Geer, S. A. (2000). *Empirical processes in M-estimation*. Cambridge: Cambridge University of Press.

van der Vaart, A. W. (1991). On differentiable functionals. *The Annals of Statistics, 19*, 178–204.

van der Vaart, A. W. (1998). *Asymptotic statistics*. Cambridge: Cambridge University of Press.

van der Vaart, A. W., Wellner, J. A. (1996). *Weak convergence and empirical processes*. New York: Springer.

van der Vaart, A. W., Wellner, J. A. (2000). Preservation theorems for Glivenko–Cantelli and uniform Glivenko–Cantelli class. In E. Gine, D. M. Mason, J. A. Wellner (Eds.), *High dimensional probability* (Vol. II, pp. 115–134). Boston: Birkhäuser.

van der Vaart, A. W., Wellner, J. A. (2001). Consistency of semiparametric maximum likelihood estimators for two-phase sampling. *Canadian Journal of Statistics, 29*, 269–288.