On differentiability of implicitly defined function in semi-parametric profile likelihood estimation

BY YUICHI HIROSE

School of Mathematics, Statistics and Operations Research, Victoria University of Wellington, New Zealand

Yuichi.Hirose@msor.vuw.ac.nz

AND JEAN-MARIE AUBRY

Laboratoire d'Analyse et de Mathématiques Appliquées, UMR CNRS 8050, Université de Paris Est à Créteil, France.

jmaubry@math.cnrs.fr

SUMMARY

In this paper we study the differentiability of implicitly defined functions which we encounter in the profile likelihood estimation of parameters in semi-parametric models. Scott and Wild (1997, 2001) and Murphy and van der Vaart (2000) developed methodologies that can avoid dealing with such implicitly defined functions by reparametrizing parameters in the profile likelihood and using an approximate least favorable submodel in semi-parametric models. Our result shows applicability of an alternative approach presented in Hirose (2010) which uses the direct expansion of the profile likelihood.

Some key words: Efficiency; Efficient information bound; Efficient score; Implicitly defined function; Profile likelihood; Semi-parametric model.

1. INTRODUCTION

Consider a general semi-parametric model

$$\mathcal{P} = \{p_{\theta,\eta}(x) : \theta \in \Theta, \eta \in H\}$$

where $p_{\theta,\eta}(x)$ is a density function on the sample space $\mathcal{X}$ which depends on a finite dimensional parameter θ and an infinite dimensional parameter η. We assume that the set Θ of the parameter θ is an open subset of $\mathbb{R}^d$ and the set H is infinite dimensional.

Once observations $X_1, \ldots, X_n$ are generated from the model, the log-likelihood is given by

$$n^{-1}\ell_n(\theta,\eta) = n^{-1}\sum_{i=1}^{n} \log p_{\theta,\eta}(X_i) = \int \log p_{\theta,\eta}(x) dF_n(x) \tag{1}$$

where F_n is the empirical cdf based on the observations. In the profile likelihood approach, we find a function $\eta_{\theta,F}$ of the parameter θ and a cdf F as the maximizer of the log-likelihood given θ:

$$\eta_{\theta,F_n} = \mathrm{argmax}_\eta \int \log p_{\theta,\eta}(x) dF_n(x) \tag{2}$$

Then the profile (log)-likelihood is given by

$$\int \log p_{\theta,\eta_{\theta,F_n}}(x) dF_n(x) \tag{3}$$

In this paper we consider the situation when the function $\eta_{\theta,F}$ is given as the solution to the operator equation of the form

$$\eta = \Psi_{\theta,F}(\eta). \tag{4}$$

Murphy, Rossini and van der Vaart (1997) encountered this type of implicitly defined function in their maximum likelihood estimation problem in the proportional odds model. According to them, "because $\hat{H}_\beta$ is not an explicit function of β, we are unable to differentiate the profile log-likelihood explicitly in β to form an estimator of Σ" (here $\hat{H}_\beta$ is the maximizer of the log-likelihood $\ell_n(\beta, H)$ given β, H is the baseline odds of failure and Σ is the efficient information). The authors (Murphy, Rossini and van der Vaart (1997)) used a numerical approximation to the problem. In the first example (Example 1) given below, we present a modified version of the proportional odds model and give an example of implicitly defined function there.

Scott and Wild (1997, 2001) also encountered implicitly defined functions in their estimation problem with data from various outcome-dependent sampling design. They proposed a method of reparametrization of profile-likelihood so that the log-likelihood is an explicitly defined function in terms of the parameters in the reparametrized model. Their estimators turned out to be efficient and Hirose and Lee (2010) showed conditions under which reparametrization gives efficient estimation in a context of multiple-sample semi-parametric model. In Example 2, we give an example of implicitly defined function in the method of Scott and Wild.

Another way to avoid dealing with implicitly defined functions is developed by Murphy and van der Vaart (2000). The paper proved the efficiency of profile likelihood estimation by introducing an 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 function in the profile likelihood.

Hirose (2010) used direct asymptotic expansion of the profile likelihood to show the efficiency of the profile likelihood estimator. Through this approach simplified logic of asymptotic expansion of the profile likelihood, we can not avoid dealing with implicitly defined functions of the form given in (4) in many applications. The purpose of this paper is to study the properties of these function such as differentiability so that the method in Hirose (2010) is applicable to many applications. The results in Hirose (2010) are summarized in SECTION 3.

In Section 2, we give 3 examples of such implicitly defined functions. The main results are presented in Section 4. In Section 5, the main results are applied to one of the examples.

2. EXAMPLES

2·1. *Example 1 (Semi-parametric proportional odds model)*

The original asymptotic theory for maximum likelihood estimator in the semi-parametric proportional odds model is developed in Murphy, Rossini and van der Vaart (1997). We present a modified version of the model in Kosorok (2008).

In this model, we observe $X = (U, \delta, Z)$, where $U = T \wedge C$, $\delta = 1_{\{U=T\}}$, $Z \in \mathbb{R}^d$ is a covariate vector, T is a failure time and C is a right censoring time. We assume C and T are independent given Z.

The proportional odds regression model is specified by the survival function of T given Z of the form

$$S(t|Z) = \frac{1}{1 + e^{\beta' Z} A(t)},$$

where $A(t)$ is nondecreasing function on $[0, \tau]$ with $A(0) = 0$. τ is the limit of censoring distribution such that $P(C > \tau) = 0$ and $P(C = \tau) > 0$. The distribution of Z and C are uninformative of S and $\mathrm{var} Z$ is positive definite.

The density function in the proportional odds model is

$$f(t|Z) = \frac{e^{\beta' Z} a(t)}{(1 + e^{\beta' Z} A(t))^2},$$

where $a(t) = dA(t)/dt$, and the hazard function is

$$h(t|Z) = \frac{f(t|Z)}{S(t|Z)} = \frac{e^{\beta' Z} a(t)}{1 + e^{\beta' Z} A(t)}.$$

The log-likelihood for an observation (U, δ, Z) is

$$\begin{aligned} \ell(U, \delta, Z; \beta, A) &= \log \left\{ h(U|Z)^{\delta} S(U|Z) \right\} \\ &= \delta(\beta' Z + \log a(U)) - (1 + \delta) \log(1 + e^{\beta' Z} A(U)). \end{aligned}$$

Consider one-dimensional submodels for A defined by the map

$$t \to A_t(s) = \int_0^s (1 + t h_1(u)) dA(u),$$

where h_1 is an arbitrary total variation bounded cadlag function on $[0, \tau]$. By differentiating the log-likelihood function $\ell(U, \delta, Z; \beta, A_t)$ with respect to t at $t = 0$, we obtain the score operator

$$\begin{aligned} B(U, \delta, Z; \beta, A)(h_1) &= \left. \frac{d}{dt} \right|_{t=0} \ell(U, \delta, Z; \beta, A_t) \\ &= \delta h_1(U) - (1 + \delta) \frac{e^{\beta' Z} \int_0^U h_1(s) dA(s)}{1 + e^{\beta' Z} A(U)}. \end{aligned}$$

Choose $h_1(u) = 1_{\{u \le t\}}$, then

$$B(U, \delta, Z; \beta, A)(h_1) = N(t) - \int_0^t W(s; \beta, A) dA(s),$$

where $N(t) = \delta 1_{\{U \le t\}}$, $Y(t) = 1_{\{U \ge t\}}$ and

$$W(s; \beta, A) = \frac{(1 + \delta) e^{\beta' Z} Y(s)}{1 + e^{\beta' Z} A(U)}.$$

For a cdf function F and a function ϕ, write $E_F \phi = \int \phi dF$. Set $E_F B(U, \delta, Z; \beta, A)(h_1) = 0$ and we obtain

$$E_F N(t) = E_F \int_0^t W(s; \beta, A) dA(s). \tag{5}$$

It is easy to check that

$$\hat{A}_{\beta,F}(t) = \int_0^t \frac{E_F dN(s)}{E_F W(s;\beta,\hat{A}_{\beta,F})} \tag{6}$$

is a solution to Equation (5).

If we let

$$\Psi_{\beta,F}(A) = \int_0^t \frac{E_F dN(s)}{E_F W(s;\beta,A)},$$

then (6) is a solution to the operator equation $A = \Psi_{\beta,F}(A)$.

2·2. *Example 2(Stratified sampling)*

The method of Scott and Wild (1997, 2001) transform the profile likelihood with an implicitly defined function of the form (4) into a likelihood with explicitly defined function by reparametrization. This example is one of the situation when we can apply their method.

Suppose the underlying data generating process on the sample space $\mathcal{Y} \times \mathcal{X}$ is a model

$$\mathcal{Q} = \{p(y,x;\theta) = f(y|x;\theta)g(x) : \theta \in \Theta,\ g \in \mathcal{G}\}. \tag{7}$$

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 $\Theta \subset \mathbb{R}^d$ is an open set containing a neighborhood of the true value θ_0 and $\mathcal{G}$ is the set of density function of x containing the true value $g_0(x)$. The variable Y may be a discrete or continuous variable or combination of both in Euclidean spaces.

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

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

and let

$$Q_s(\theta,g) = \int Q_{s|X}(x;\theta)g(x)\, dx$$

be the probability of (Y,X) belonging to stratum $\mathcal{S}_s$.

In standard stratified sampling, for each $s = 1,\dots,S$, a random sample of size n_s, is taken from the conditional distribution

$$p_s(y,x;\theta,g) = \frac{f(y|x;\theta)g(x)1_{(y,x)\in\mathcal{S}_s}}{Q_s(\theta,g)} \tag{8}$$

of (Y,X) given stratum $\mathcal{S}_s$.

For each $s = 1,\dots,S$, let F_{s0} be the cumulative distribution function (cdf) for the density $p_s(y,x;\theta_0,g_0)$ at the true value (θ_0,g_0). Let w_s, $s = 1,\dots,S$, be the weight probabilities, i.e., $w_s > 0$ for all s and $\sum_s w_s = 1$. The log likelihood with the weight probabilities w_s and the cdfs F_{s0} is

$$\sum_{s=1}^S w_s \int \log p_s(y,x;\theta,g) dF_{s0} = \sum_{s=1}^S w_s \left[\int \{\log f(y|x;\theta) + \log g(x)\}\, dF_{s0} - \log Q_s(\theta,g)\right]$$

For each θ, we find a maximizer $\hat{g}_\theta(x)$ of log-likelihood under the assumption that the support of the distribution of X is finite: i.e. $\text{SUPP}(X) = \{v_1,\dots,v_K\}$. 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 the expected log-likelihood at θ, differentiate it with respect to g_k and set the derivative equal to zero,

$$\frac{\partial}{\partial g_k} \sum_{s=1}^S w_s \int \log p_s(y, x; \theta, g) dF_{s0} = \sum_{s=1}^S w_s \left\{ \frac{\int 1_{x=v_k} dF_{s0}}{g_k} - \frac{Q_{s|X}(v_k;\theta)}{Q_s(\theta, g)} \right\} = 0.$$

The solution g_k to the equation is

$$\hat{g}_\theta(v_k) = g_k = \frac{\sum_{s=1}^S w_s \int 1_{x=v_k} dF_{s0}}{\sum_{s=1}^S w_s \frac{Q_{s|X}(v_k;\theta)}{Q_s(\theta,g)}}.$$

The form of the function motivates us to work with an implicit function $\hat{g}_\theta(x)$ given by the solution of $g = \Psi_\theta(g)$ where

$$\Psi_\theta(g) = \frac{f_0^*(x)}{\sum_{s=1}^S w_s \frac{Q_{s|X}(x;\theta)}{Q_s(\theta,g)}}$$

and

$$f_0^*(x) = \sum_{s=1}^S w_s \frac{Q_{s|X}(x;\theta_0)g_0(x)}{Q_s(\theta_0, g_0)}.$$

For further development of this example, see Hirose and Lee (2010).

2·3. *Example 3(Continuous outcome with missing data)*

This example is studied in Weaver and Zhou (2005) and Song, Zhou and Kosorok (2009). As in Example 2 we assume the underlying data generating process on the sample space $\mathcal{Y} \times \mathcal{X}$ is given by the model (7) where we assume the variable Y is continuous.

We consider a situation when there are samples for which we observe complete observation (Y, X) and for which we observe only Y. Let R_i be the indicator variable for the ith observation defined by

$$R_i = \begin{cases} 1 \text{ if } X_i \text{ is observed} \\ 0 \text{ if } X_i \text{ is not observed.} \end{cases}$$

Then the index set for the complete observations is $V = \{R_i = 1\}$ and the index set for the incomplete observations is $\overline{V} = \{R_i = 0\}$. Let $n_V = |V|$, $n_{\overline{V}} = |\overline{V}|$ be the total number of complete observations and incomplete observations, respectively.

Weaver and Zhou (2005) and Song, Zhou and Kosorok (2009) consider the likelihood of the form

$$L(\theta, g) = \prod_{i \in V} \{f(Y_i|X_i;\theta)g(X_i)\} \prod_{i \in \overline{V}} f_Y(Y_i;\theta, g) \tag{9}$$

where

$$f_Y(y;\theta, g) = \int_{\mathcal{X}} f(y|x;\theta)g(x)dx. \tag{10}$$

The density function that corresponds to the likelihood (9) is

$$p(s, z;\theta, g) = 1_{\{s=1\}} f(y|x;\theta)g(x) + 1_{\{s=2\}} f_Y(y;\theta, g) \tag{11}$$

where $z = (y, x)$ when $s = 1$, $z = y$ when $s = 2$.

The log-likelihood, the $1/n$ times log of (9), is

$$\frac{1}{n}\log L(\theta, g) = \frac{n_V}{n}\frac{1}{n_V}\sum_{i\in V}\{\log f(y_i|x_i;\theta) + \log g(x_i)\} + \frac{n_{\overline{V}}}{n}\frac{1}{n_{\overline{V}}}\sum_{i\in\overline{V}}\log f_Y(y_i;\theta,g).$$

Let F_{1n} and F_{2n} be the empirical cdfs based on the samples in V and $\overline{V}$, respectively. Denote $w_{1n} = n_V/n$, $w_{2n} = n_{\overline{V}}/n$ and let $F_n = \sum_{s=1}^{2} w_{sn}F_{sn}$.

Then the log-likelihood can be expressed as

$$\begin{aligned}\frac{1}{n}\log L(\theta, g) &= \int \log p(s, z;\theta, g)dF_n \\ &= w_{1n}\int \{\log f(y_i|x_i;\theta) + \log g(x_i)\}\, dF_{1n} + w_{2n}\int \log f_Y(y_i;\theta,g)dF_{2n}.\end{aligned}$$

First we assume that the support of X is $\{v_1, \ldots, v_L\}$. Let $(g_1, \ldots, g_L) = (g(v_1), \ldots, g(v_L))$, then

$$\log g(z) = \sum_{l=1}^{L} 1_{\{x=v_l\}}\log g_l, \quad \text{and} \quad f_Y(y;\theta,g) = \sum_{l=1}^{L} f(y|v_l;\theta)g_l \tag{12}$$

For fixed θ, we find the maximizer $(g_1, \ldots, g_K)$ of the log-likelihood $\int \log p(s, x;\theta, g)dF_n$.

Using (12), the derivative of the log-likelihood with respect to g_k is

$$\frac{\partial}{\partial g_l}\int \log p(s, x;\theta, g)dF_n = w_{1n}\frac{\int 1_{\{x=v_l\}}dF_{1n}}{g_l} + w_{2n}\int \frac{f(y|v_l;\theta)}{f_Y(y;\theta,g)}dF_{2n}.$$

Let η be a Lagrange multiplier to account for $\sum_l g_l = 1$. Set $\frac{\partial}{\partial g_l}\int \log p(s, x;\theta, g)dF_n + \eta = 0$. Multiply by g_l and sum over $l = 1, \ldots, L$ to get $w_{1n} + w_{2n} + \eta = 0$. Therefore $\eta = -(w_{1n} + w_{2n}) = -1$ and $\frac{\partial}{\partial g_l}\int \log p(s, x;\theta, g)dF - 1 = 0$. By rearranging this equation, we obtain

$$g_l = \frac{w_{1n}\int 1_{\{x=v_l\}}dF_{1n}}{1 - w_{2n}\int \frac{f(y|v_l;\theta)}{f_Y(y;\theta,g)}dF_{2n}}.$$

This gives us a candidate function

$$g_{\theta,F_n}(x) = \frac{w_{1n}\int \frac{dF_{1n}}{dx}}{1 - w_{2n}\int \frac{f(y|x;\theta)}{f_Y(y;\theta,g_{\theta,F_n})}dF_{2n}}. \tag{13}$$

This is a solution to the equation $g = \Psi_{\theta,F_n}(g)$ with

$$\Psi_{\theta,F}(g) = \frac{w_1\int \frac{dF_1}{dx}}{1 - w_2\int \frac{f(y|x;\theta)}{f_Y(y;\theta,g)}dF_2}$$

We continue this example in SECTION 5.

3. ASYMPTOTIC NORMALITY OF PROFILE LIKELIHOOD ESTIMATOR

Hirose (2010) showed the efficiency of the estimator based on the profile likelihood in semiparametric models using the direct asymptotic expansion of the profile likelihood. The method gives alternative to the one proposed by Murphy and van der Vaart (2000) which uses an asymptotic expansion of approximate profile likelihood. We summarize the results from the paper. To

be able to use the results to the examples with implicitly defined function of the form (4), we must establish the differentiability of implicitly defined functions. This is the motivation of this paper and we present the differentiability of such functions in the next section.

Suppose we have a function $\eta_{\theta,F}$ that depends on (θ, F) such that $\tilde{\ell}_0(x) \equiv \tilde{\ell}_{\theta_0,F_0}(x)$ is the efficient score function, where

$$\tilde{\ell}_{\theta,F}(x) \equiv \frac{\partial}{\partial\theta} \log p(x;\theta,\eta_{\theta,F}). \tag{14}$$

The theorem below show that if the solution $\hat{\theta}_n$ to the estimating equation

$$\int \tilde{\ell}_{\hat{\theta}_n,F_n}(x) dF_n = 0 \tag{15}$$

is consistent then it is asymptotically linear with the efficient influence function $\tilde{I}_0^{-1}\tilde{\ell}_0(x)$ so that

$$n^{-1/2}(\hat{\theta}_n - \theta_0) = \int \tilde{I}_0^{-1}\tilde{\ell}_0(s,x) d\{n^{-1/2}(F_n - F_0)\} + o_P(1) \xrightarrow{d} N(0, \tilde{I}_0^{-1}), \tag{16}$$

where $N(0, \tilde{I}_0^{-1})$ is a normal distribution with mean zero and variance $\tilde{I}_0^{-1}$. Since $\tilde{I}_0 = E_0(\tilde{\ell}_0\tilde{\ell}_0^T)$ is the efficient information matrix, this demonstrates that the estimator $\hat{\theta}_n$ is efficient.

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

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

THEOREM 1. *(Hirose (2010)) Assumptions:*

(R0) The function $g_{\theta,F}$ satisfies $g_{\theta_0,F_0} = g_0$ and the function

$$\tilde{\ell}_0(x) = \tilde{\ell}_{\theta_0,F_0}(x)$$

is the efficient score function where $\tilde{\ell}_{\theta,F}(x)$ is given by (14).

(R1) The empirical process F_n is $n^{1/2}$-consistent, i.e., $n^{1/2}\|F_n - F_0\| = O_P(1)$, and there exists a $\rho > 0$ and a neighborhood Θ of θ_0 such that for each $(\theta, F) \in \Theta \times \mathcal{C}_\rho$, the log-likelihood function $\log p(x;\theta,\hat{g}_{\theta,F})$ is twice continuously differentiable with respect to θ and Hadamard differentiable with respect to F for all x.

(R2) The efficient information matrix $\tilde{I}_0 = E_0(\tilde{\ell}_0\tilde{\ell}_0^T)$ is invertible.

(R3) There exists a $\rho > 0$ and a neighborhood Θ of θ_0 such that the class of functions $\{\tilde{\ell}_{\theta,F}(x) : (\theta, F) \in \Theta \times \mathcal{C}_\rho\}$ is Donsker with square integrable envelope function, and that the class of functions $\{\frac{\partial}{\partial\theta}\tilde{\ell}_{\theta,F}(x) : (\theta, F) \in \Theta \times \mathcal{C}_\rho\}$ is Glivenko-Cantelli with integrable envelope function.

Under the assumptions $\{(R0), (R1), (R2), (R3)\}$, for a consistent solution $\hat{\theta}_n$ to the estimating equation (15), the equation (16) holds.

4. MAIN RESULTS

In this section we show the differentiability of implicitly defined function which is given as a solution to the operator equation (4). First, we state the 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

$$t^{-1}\{\psi(x_t) - \psi(x)\} \to d\psi(x)(h) \quad \text{as} \quad t \downarrow 0$$

for any map $t \to x_t$ with $x_{t=0} = x$ and $t^{-1}(x_t - x) \to h \in B_1$ (as $t \downarrow 0$).

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).)

We denote the second derivative of ψ in the sense of Hadamard by $d^2\psi(x)$. The usual first and second derivative of a parametric function ψ are denoted by $\dot{\psi}$ and $\ddot{\psi}$.

As we stated in Introduction, we consider a general semi-parametric model

$$\mathcal{P} = \{p_{\theta,\eta}(x) : \theta \in \Theta, \eta \in H\}$$

where $p_{\theta,\eta}(x)$ is a density function on the sample space $\mathcal{X}$ which depends on a finite dimensional parameter θ and an infinite dimensional parameter η. We assume that the set Θ of the parameter θ is an open subset of $\mathbb{R}^d$ and the set H is a convex set in a Banach space $\mathcal{B}$ which we may assume the closed linear span of H.

THEOREM 2. *Suppose the function $\Psi_{\theta,F}(\eta)$ is*

(A1) two times continuously differentiable with respect to θ and two times Hadamard differentiable with respect to η and Hadamard differentiable with respect to F so that the derivatives $\dot{\Psi}_{\theta,F}(\eta)$, $\ddot{\Psi}_{\theta,F}(\eta)$, $d_\eta\Psi_{\theta,F}(\eta)$, $d^2_\eta\Psi_{\theta,F}(\eta)$, $d_\eta\dot{\Psi}_{\theta,F}(\eta)$ and $d_F\Psi_{\theta,F}(\eta)$ exist (where, for example, $\dot{\Psi}_{\theta,F}(\eta)$ is the first derivative with respect to θ, and $d_\eta\Psi_{\theta,F}(\eta)$ is the first derivative with respect to η in the sense of Hadamard. Similarly, the rest is defined).

(A2) the true values (θ_0, η_0, F_0) satisfy $\eta_0 = \Psi_{\theta_0,F_0}(\eta_0)$.

(A3) the linear operator $d_\eta\Psi_{\theta_0,F_0}(\eta_0) : \mathcal{B} \to \mathcal{B}$ has the operator norm $\|d_\eta\Psi_{\theta_0,F_0}(\eta_0)\| < 1$.

Then the solution $\eta_{\theta,F}$ to the equation

$$\eta = \Psi_{\theta,F}(\eta) \tag{17}$$

exists in an neighborhood of (θ_0, F_0) and it is two times continuously differentiable with respect to θ and Hadamard differentiable with respect to F in the neighborhood. Moreover, the derivatives are given by

$$\dot{\eta}_{\theta,F} = [I - d_\eta\Psi_{\theta,F}(\eta_{\theta,F})]^{-1}\dot{\Psi}_{\theta,F}(\eta_{\theta,F}), \tag{18}$$

$$\begin{aligned}\ddot{\eta}_{\theta,F} = [I - d_\eta\Psi_{\theta,F}(\eta_{\theta,F})]^{-1} \Big[&\ddot{\Psi}_{\theta,F}(\eta_{\theta,F}) + d_\eta\dot{\Psi}_{\theta,F}(\eta_{\theta,F})\dot{\eta}^T_{\theta,F} \\ &+ d_\eta\dot{\Psi}^T_{\theta,F}(\eta_{\theta,F})\dot{\eta}_{\theta,F} + d^2_\eta\Psi_{\theta,F}(\eta_{\theta,F})\dot{\eta}_{\theta,F}\dot{\eta}^T_{\theta,F}\Big],\end{aligned} \tag{19}$$

and

$$d_F\eta_{\theta,F} = [I - d_\eta\Psi_{\theta,F}(\eta_{\theta,F})]^{-1}d_F\Psi_{\theta,F}(\eta_{\theta,F}). \tag{20}$$

4·1. *Proof of* THEOREM 2

Existence and invertibility:

We assumed the derivative $d_\eta\Psi_{\theta_0,F_0}(\eta_0)$ exists and $\|d_\eta\Psi_{\theta_0,F_0}(\eta_0)\| < 1$. By the continuity with respect to the parameters (θ, η, F), there is a neighborhood of (θ_0, η_0, F_0) such that $\|d_\eta\Psi_{\theta,F}(\eta)\| < 1$ for all (θ, η, F) in the neighborhood. Let $I : \mathcal{B} \to \mathcal{B}$ be the identity operator on the space $\mathcal{B}$. In the neighborhood, the map $(I - d_\eta\Psi_{\theta,F}(\eta)) : \mathcal{B} \to \mathcal{B}$ has the inverse

$(I - d_\eta \Psi_{\theta,F}(\eta))^{-1}$, which is also a bounded linear map (cf. Kolmogorov and Fomin (1975), Theorem 4, p 231). It also follows that there is a neighborhood of (θ_0, η_0, F_0) such that, for each (θ, F), the map $\eta \to \Psi_{\theta,F}(\eta)$ is a contraction mapping in the neighborhood. By Banach's contraction principle (cf. Agarwal, O'Regan and Sahu (2009), Theorem 4.1.5, p178), the solution to the equation (17) exists uniquely in the neighborhood.

Differentiability with respect to F:
Fix h_1 and h_2 in appropriate spaces and let F_t and η_t be maps such that $F_{t=0} = F$, $\eta_{t=0} = \eta$, $t^{-1}\{F_t - F\} \to h_1$ and $t^{-1}\{\eta_t - \eta\} \to h_2$ as $t \downarrow 0$. Then, $F_t \to F$, $\eta_t \to \eta$ (as $t \downarrow 0$), and by condition (A1), as $t \downarrow 0$,

$$t^{-1}\{\Psi_{\theta,F_t}(\eta) - \Psi_{\theta,F}(\eta)\} \to d_F \Psi_{\theta,F}(\eta) h_1$$

and

$$t^{-1}\{\Psi_{\theta,F}(\eta_t) - \Psi_{\theta,F}(\eta)\} \to d_\eta \Psi_{\theta,F}(\eta) h_2.$$

Therefore, as $t \downarrow 0$,

$$\begin{aligned} & t^{-1}\{\eta_{\theta,F_t} - \eta_{\theta,F}\} = t^{-1}\{\Psi_{\theta,F_t}(\eta_{\theta,F_t}) - \Psi_{\theta,F}(\eta_{\theta,F})\} \\ &= t^{-1}\{\Psi_{\theta,F_t}(\eta_{\theta,F_t}) - \Psi_{\theta,F}(\eta_{\theta,F_t})\} + t^{-1}\{\Psi_{\theta,F}(\eta_{\theta,F_t}) - \Psi_{\theta,F}(\eta_{\theta,F})\} \\ &= d_F \Psi_{\theta,F}(\eta_{\theta,F}) h_1 + d_\eta \Psi_{\theta,F}(\eta_{\theta,F}) t^{-1}\{\eta_{\theta,F_t} - \eta_{\theta,F}\} + o(1). \end{aligned}$$

It follows that

$$[I - d_\eta \Psi_{\theta,F}(\eta_{\theta,F})] t^{-1}\{\eta_{\theta,F_t} - \eta_{\theta,F}\} = d_F \Psi_{\theta,F}(\eta_{\theta,F}) h_1 + o(1)$$

and

$$t^{-1}\{\eta_{\theta,F_t} - \eta_{\theta,F}\} \to [I - d_\eta \Psi_{\theta,F}(\eta_{\theta,F})]^{-1} d_F \Psi_{\theta,F}(\eta_{\theta,F}) h_1$$

as $t \downarrow 0$. Since the map $[I - d_\eta \Psi_{\theta,F}(\eta_{\theta,F})]^{-1} d_F \Psi_{\theta,F}(\eta_{\theta,F})$ is bounded and linear, the function $\eta_{\theta,F}(x)$ is Hadamard differentiable with respect to F.

Differentiability with respect to θ:
Similar to the case for differentiability with respect to F, for $t^{-1}(\theta_t - \theta) \to a \in \mathbb{R}^d$ as $t \downarrow 0$, we have

$$t^{-1}\{\eta_{\theta_t,F} - \eta_{\theta,F}\} = a^T \dot{\Psi}_{\theta,F}(\eta_{\theta,F}) + d_\eta \Psi_{\theta,F}(\eta_{\theta,F}) t^{-1}\{\eta_{\theta_t,F} - \eta_{\theta,F}\} + o(1).$$

It follows that the first derivative $\dot{\eta}_{\theta,F}$ of $\eta_{\theta,F}(x)$ with respect to θ is given by

$$a^T \dot{\eta}_{\theta,F} = [I - d_\eta \Psi_{\theta,F}(\eta_{\theta,F})]^{-1} a^T \dot{\Psi}_{\theta,F}(\eta_{\theta,F}). \tag{21}$$

From (21), we have

$$a^T \dot{\eta}_{\theta,F} = a^T \dot{\Psi}_{\theta,F}(\eta_{\theta,F}) + d_\eta \Psi_{\theta,F}(\eta_{\theta,F})(a^T \dot{\eta}_{\theta,F}).$$

Using this equation, for $t^{-1}(\theta_t - \theta) \to b \in \mathbb{R}^d$ as $t \downarrow 0$, we get $\theta_t \to \theta$ and hence

$$\begin{aligned} & t^{-1}\{a^T \dot{\eta}_{\theta_t,F} - a^T \dot{\eta}_{\theta,F}\} \\ &= t^{-1}\{a^T \dot{\Psi}_{\theta_t,F}(\eta_{\theta_t,F}) - a^T \dot{\Psi}_{\theta,F}(\eta_{\theta,F})\} + t^{-1}\{d_\eta \Psi_{\theta_t,F}(\eta_{\theta_t,F})(a^T \dot{\eta}_{\theta_t,F}) - d_\eta \Psi_{\theta,F}(\eta_{\theta,F})(a^T \dot{\eta}_{\theta,F})\} \\ &= a^T \ddot{\Psi}_{\theta,F}(\eta_{\theta,F}) b + a^T d_\eta \dot{\Psi}_{\theta,F}(\eta_{\theta,F})(\dot{\eta}_{\theta,F}^T b) + \{d_\eta \dot{\Psi}_{\theta,F}(\eta_{\theta,F})(a^T \dot{\eta}_{\theta,F})\}^T b \\ &\quad + d_\eta^2 \Psi_{\theta,F}(\eta_{\theta,F})(a^T \dot{\eta}_{\theta,F})(\dot{\eta}_{\theta,F}^T b) + d_\eta \Psi_{\theta,F}(\eta_{\theta,F}) t^{-1}\{a^T \dot{\eta}_{\theta_t,F} - a^T \dot{\eta}_{\theta,F}\} + o(1). \end{aligned}$$

By rearranging this we obtain

$$\begin{aligned}&[I - d_\eta \Psi_{\theta,F}(\eta_{\theta,F})]t^{-1}\{a^T\dot{\eta}_{\theta_t,F} - a^T\dot{\eta}_{\theta,F}\}\\ &= a^T\ddot{\Psi}_{\theta,F}(\eta_{\theta,F})b + a^T d_\eta \dot{\Psi}_{\theta,F}(\eta_{\theta,F})(\dot{\eta}_{\theta,F}^T b) + \{d_\eta \dot{\Psi}_{\theta,F}(\eta_{\theta,F})(a^T\dot{\eta}_{\theta,F})\}^T b\\ &\quad + d_\eta^2 \Psi_{\theta,F}(\eta_{\theta,F})(a^T\dot{\eta}_{\theta,F})(\dot{\eta}_{\theta,F}^T b) + o(1),\end{aligned}$$

and hence, as $t \downarrow 0$,

$$t^{-1}\{a^T\dot{\eta}_{\theta_t,F} - a^T\dot{\eta}_{\theta,F}\} \to a^T\ddot{\eta}_{\theta,F}b$$

where

$$\begin{aligned}a^T\ddot{\eta}_{\theta,F}b = [I - d_\eta\Psi_{\theta,F}(\eta_{\theta,F})]^{-1}\Big[&a^T\ddot{\Psi}_{\theta,F}(\eta_{\theta,F})b + a^T d_\eta\dot{\Psi}_{\theta,F}(\eta_{\theta,F})(\dot{\eta}_{\theta,F}^T b)\\ &+\{d_\eta\dot{\Psi}_{\theta,F}(\eta_{\theta,F})(a^T\dot{\eta}_{\theta,F})\}^T b + d_\eta^2\Psi_{\theta,F}(\eta_{\theta,F})(a^T\dot{\eta}_{\theta,F})(\dot{\eta}_{\theta,F}^T b)\Big].\end{aligned}$$

Therefore $\dot{\eta}_{\theta,F}$ is differentiable with respect to θ with derivative $\ddot{\eta}_{\theta,F}$.

5. EXAMPLE 3 CONTINUED

Let θ_0, g_0 and F_0 be the true values of θ, g and F at which data are generated.
For $\theta \in \mathbb{R}^d$, $F = \sum_s w_s F_s$ and function $g(x)$, define

$$\Psi_{\theta,F}(g) = \frac{\int \frac{\pi_1(dF)}{dx}}{A(x;\theta,g,F)}, \tag{22}$$

where $\pi_s : F = \sum_{s'} w_{s'}F_{s'} \to w_s F_s$, $s = 1, 2$, are projections, and

$$A(x;\theta,g,F) = 1 - \int \frac{f(y|x;\theta)}{f_Y(y;\theta,g)}\pi_2(dF). \tag{23}$$

Then the function $g_{\theta,F_n}(x)$ given by (13) is the solution to the operator equation

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

with $F = F_n$.

First, we prove the following lemma which we need in the subsequent verifications.

LEMMA 1. *At (θ_0, F_0), $g_0(x)$ is a solution to the operator equation (24).*

Proof. Since $\int \frac{dF_{10}}{dx} = \int f(y|x;\theta_0)g_0(x)dy = g_0(x)$, and $\frac{dF_{20}(y)}{dy} = f_Y(y;\theta_0,g_0)$, $w_{10} + w_{20} = 1$, we have

$$\Psi_{\theta_0,F_0}(g_0)(x) = \frac{w_{10}\int\frac{dF_{10}}{dx}}{1 - w_{20}\int\frac{f(y|x;\theta_0)}{f_Y(y;\theta_0,g_0)}dF_{20}} = \frac{w_{10}g_0(x)}{1 - w_{20}\int\frac{f(y|x;\theta_0)}{f_Y(y;\theta_0,g_0)}f_Y(y;\theta_0,g_0)dy} = g_0(x)$$

where we used $\int f(y|x;\theta)dy = 1$ for each x. □

We show the differentiability of the solution $g_{\theta,F}(x)$ to the equation (24) with respect to θ and F.

THEOREM 3. *We assume the function $f(y|x;\theta)$ is twice continuously differentiable with respect to θ and $\frac{w_{20}}{w_{10}} < 1$.*

The solution $g_{\theta,F}(x)$ to the operator equation (24) exists in an neighborhood of (θ_0, F_0) and it is two times continuously differentiable with respect to θ and Hadamard differentiable with respect to F in the neighborhood.

We verify conditions (A1), (A2) and (A3) in Theorem 1 so that the results in the theorem follows from Theorem 1.

We denote $f = f(y|x;\theta)$, $f_Y = f_Y(y;\theta,g)$, $A = A(x;\theta,g,F)$, $\dot{f} = \frac{\partial}{\partial\theta} f(y|x;\theta)$, $\ddot{f} = \frac{\partial^2}{\partial\theta\,\partial\theta^T} f(y|x;\theta)$, $\dot{f}_Y = \int \dot{f}(y|x;\theta)g(x)dx$, and $\ddot{f}_Y = \int \ddot{f}(y|x;\theta)g(x)dx$.

Verification of condition (A1): We show that the map $\Psi_{\theta,F}(g)$ is differentiable with respect to θ, F and g.

(a) (The derivative of $\Psi_{\theta,F}(g)$ with respect to F)
Suppose a map $t \to F_t$ satisfies $t^{-1}(F_t - F) \to h$ as $t \downarrow 0$.
Then

$$t^{-1}\left\{\Psi_{\theta,F_t}(g) - \Psi_{\theta,F}(g)\right\} = t^{-1}\left\{\frac{\int \frac{\pi_1(dF_t)}{dx}}{A(x;\theta,g,F_t)} - \frac{\int \frac{\pi_1(dF)}{dx}}{A(x;\theta,g,F)}\right\} \to d_F\Psi_{\theta,F}(g)h$$

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

$$d_F\Psi_{\theta,F}(g)h = \frac{\int \frac{\pi_1(dh)}{dx} A(x;\theta,g,F) - \int \frac{\pi_1(dF)}{dx}\left\{d_F A(x;\theta,g,F)h\right\}}{\{A(x;\theta,g,F)\}^2} \tag{25}$$

and

$$d_F A(x;\theta,g,F)h = -\int \frac{f(y|x;\theta)}{f_Y(y;\theta,g)}\pi_2(dh).$$

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 $d_F\Psi_{\theta,F}(g)$).

(b) (The derivative of $\Psi_{\theta,F}(g)$ with respect to g)
Now, suppose a map $t \to g_t$ satisfies $t^{-1}(g_t - g) \to h^*$ as $t \downarrow 0$. Then, as $t \downarrow 0$,

$$t^{-1}\left\{\Psi_{\theta,F}(g_t) - \Psi_{\theta,F}(g)\right\} = t^{-1}\left\{\frac{\int \frac{\pi_1(dF)}{dx}}{A(x;\theta,g_t,F)} - \frac{\int \frac{\pi_1(dF)}{dx}}{A(x;\theta,g,F)}\right\} \to d_g\Psi_{\theta,F}(g)h^*$$

where

$$d_g\Psi_{\theta,F}(g)h^* = \frac{-\int \frac{\pi_1(dF)}{dx}\left\{d_g A(x;\theta,g,F)h^*\right\}}{\{A(x;\theta,g,F)\}^2}, \tag{26}$$

and

$$d_g A(x;\theta,g,F)h^* = \int f(y|x;\theta)\frac{\int f(y|x;\theta)h^*(x)dx}{\{f_Y(y;\theta,g)\}^2}\pi_2(dF). \tag{27}$$

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. In "Verification of condition (A3)", we show the boundedness of the derivative $d_g\Psi_{\theta,F}(g)$.

(c) (The second derivative of $\Psi_{\theta,F}(g)$ with respect to g)
Suppose a map $t \to g_t$ satisfies $t^{-1}(g_t - g) \to h_2$ as $t \downarrow 0$. Then, as $t \downarrow 0$,

$$\begin{aligned} & t^{-1}\{d_g\Psi_{\theta,F}(g_t)h_1 - d_g\Psi_{\theta,F}(g)h_1\} \\ &= t^{-1}\left[\frac{-\int \frac{\pi_1(dF)}{dx}\{d_gA(x;\theta,g_t,F)h_1\}}{\{A(x;\theta,g_t,F)\}^2} - \frac{-\int \frac{\pi_1(dF)}{dx}\{d_gA(x;\theta,g,F)h_1\}}{\{A(x;\theta,g,F)\}^2}\right] \\ &\to d_g^2\Psi_{\theta,F}(g)h_1h_2 \end{aligned}$$

where

$$\begin{aligned} & d_g^2\Psi_{\theta,F}(g)h_1h_2 \\ &= \int \frac{\pi_1(dF)}{dx}\left[-\frac{d_g^2A(x;\theta,g,F)h_1h_2}{\{A(x;\theta,g_t,F)\}^2} + \frac{2\{d_gA(x;\theta,g,F)h_1\}\{d_gA(x;\theta,g,F)h_2\}}{\{A(x;\theta,g,F)\}^3}\right], \end{aligned} \tag{28}$$

and

$$d_g^2A(x;\theta,g_t,F)h_1h_2 = -2\int f(y|x;\theta)\frac{\left\{\int f(y|x;\theta)h_1(x)dx\right\}\left\{\int f(y|x;\theta)h_2(x)dx\right\}}{\{f_Y(y;\theta,g)\}^3}\pi_2(dF).$$

(Again, we omit the proof of boundedness of the derivative.)

(d) (The first and second derivative of $\Psi_{\theta,F}(g)$ with respect to θ)
It is straightforward to show that the function $\Psi_{\theta,F}(g)$ defined by (22) is twice continuously differentiable with respect to θ. Let us denote the first and second derivatives by $\dot{\Psi}_{\theta,F}(g)$ and $\ddot{\Psi}_{\theta,F}(g)$, respectively. They are given by, for $a, b \in R^d$,

$$a^T\dot{\Psi}_{\theta,F}(g) = a^T\left\{\frac{\partial}{\partial\theta}\Psi_{\theta,F}(g)\right\} = -\frac{\int \frac{\pi_1(dF)}{dx}a^T\dot{A}}{A^2}, \tag{29}$$

$$a^T\ddot{\Psi}_{\theta,F}(g)b = a^T\left\{\frac{\partial^2}{\partial\theta\,\partial\theta^T}\Psi_{\theta,F}(g)\right\}b = -\frac{\int \frac{\pi_1(dF)}{dx}\{A(a^T\ddot{A}b) - 2(a^T\dot{A})(\dot{A}^Tb)\}}{A^3}, \tag{30}$$

where

$$a^T\dot{A} = a^T\left\{\frac{\partial}{\partial\theta}A(x;\theta,g,F)\right\} = -\int \frac{f_Y(a^T\dot{f}) - f(a^T\dot{f}_Y)}{f_Y^2}\pi_2(dF)$$

and

$$\begin{aligned} & a^T\ddot{A}b = a^T\left\{\frac{\partial^2}{\partial\theta\,\partial\theta^T}A(x;\theta,g,F)\right\}b \\ &= -\int \frac{f_Y^2(a^T\ddot{f}b) - ff_Y(a^T\ddot{f}_Yb) + 2f(a^T\dot{f}_Y)(\dot{f}_Y^Tb) - f_Y(a^T\dot{f})(\dot{f}_Y^Tb) - f_Y(a^T\dot{f}_Y)(\dot{f}^Tb)}{f_Y^3}\pi_2(dF). \end{aligned}$$

(e) (The derivative of $\Psi_{\theta,F}(g)$ with respect to θ and g)
Suppose a map $t \to g_t$ is such that $g_t \to g$ and $t^{-1}(g_t - g) \to h^*$ as $t \downarrow 0$. Then, as $t \downarrow 0$,

$$\begin{aligned} & t^{-1}\{a^T\dot{\Psi}_{\theta,F}(g_t) - a^T\dot{\Psi}_{\theta,F}(g)\} \\ &= -t^{-1}\left[\frac{\int \frac{\pi_1(dF)}{dx}a^T\dot{A}(x;\theta,g_t,F)}{\{A(x;\theta,g_t,F)\}^2} - \frac{\int \frac{\pi_1(dF)}{dx}a^T\dot{A}(x;\theta,g,F)}{\{A(x;\theta,g,F)\}^2}\right] \end{aligned}$$

$$\to a^T d_g \dot{\Psi}_{\theta,F}(g)h^*,$$

where

$$a^T d_g \dot{\Psi}_{\theta,F}(g)h^* = -\int \frac{\pi_1(dF)}{dx}\left[\frac{a^T d_g \dot{A}(x;\theta,g,F)h^*}{\{A(x;\theta,g,F)\}^2} - \frac{2a^T \dot{A}(x;\theta,g,F) d_g A(x;\theta,g,F)h^*}{\{A(x;\theta,g,F)\}^3}\right], \tag{31}$$

and

$$\begin{aligned} &a^T d_g \dot{A}(x;\theta,g,F)h^* \\ &= \int (a^T \dot{f})\frac{\int f h^* dx}{f_Y^2}\pi_2(dF) + \int f \frac{\int (a^T \dot{f})h^* dx}{f_Y^2}\pi_2(dF) - 2\int f (a^T \dot{f}_Y)\frac{\int f h^* dx}{f_Y^3}\pi_2(dF). \end{aligned}$$

Verification of condition (A2):
This is verified in LEMMA 1.

Verification of condition (A3):
Let L_1 be the space of all real valued measurable functions $h(x)$ with $\|h\|_1 = \int |h(x)|dx < \infty$. Then L_1 is a Banach space with the norm $\|\cdot\|_1$. The supnorm is denoted by $\|h\|_\infty = \sup_x |h(x)|$.

Since $\int \frac{\pi_1(dF_0)}{dx} = w_{10} g_0(x)$, (26) implies

$$d_g \Psi_{\theta_0,F_0}(g_0)h^* = \frac{-w_{10} g_0(x) d_g A(x;\theta_0,g_0,F_0)h^*}{\{A(x;\theta_0,g_0,F_0)\}^2}.$$

By (23), $\pi_2(dF_0) = w_{20} f_Y(y;\theta_0,g_0)dy$ and $\int f(y|x;\theta)dy = 1$, for all x, we have

$$A(x;\theta,g_0,F_0) = 1 - \int \frac{f(y|x;\theta_0)}{f_Y(y;\theta_0,g_0)}\pi_2(dF_0) = 1 - w_{20} = w_{10}.$$

These equations and (27) imply

$$d_g \Psi_{\theta_0,F_0}(g_0)h^* = -\frac{w_{20}}{w_{10}} g_0(x) \int f(y|x;\theta_0)\frac{\int f(y|x;\theta_0)h^*(x)dx}{f_Y(y;\theta_0,g_0)}dy. \tag{32}$$

The L_1 norm of (32) is

$$\begin{aligned} \|d_g \Psi_{\theta_0,F_0}(g_0)h^*\|_1 &= \int \left| \frac{w_{20}}{w_{10}} g_0(x) \int f(y|x;\theta_0)\frac{\int f(y|x;\theta_0)h^*(x)dx}{f_Y(y;\theta_0,g_0)}dy \right| dx \\ &\le \frac{w_{20}}{w_{10}} \int g_0(x) \left(\int f(y|x;\theta_0)\frac{\int f(y|x;\theta_0)|h^*(x)|dx}{f_Y(y;\theta_0,g_0)}dy \right) dx \\ &= \frac{w_{20}}{w_{10}} \int |h^*(x)|dx \quad (\text{by Fubini's theorem and } \int f(y|x;\theta_0)dy = 1) \\ &= \frac{w_{20}}{w_{10}} \|h^*\|_1 \end{aligned}$$

From the calculation above, we see that the operator $h^* \to d_g \Psi_{\theta_0,F_0}(g_0)h^*$ has the operator norm $\le \frac{w_{20}}{w_{10}}$. Since we assumed $\frac{w_{20}}{w_{10}} < 1$, we have condition (A3).

5·1. *Asymptotic normality and efficiency*

Here we apply THEOREM 1 to show efficiency of the estimator in Example 3. First, we identify the efficient score function in the example. Then, we verify conditions (R0)–(R3) in the theorem.

Efficient score function: We show that the candidate function (13) (the solution to the equation (24)) gives us the efficient score function in Example 3.

THEOREM 4 (THE EFFICIENT SCORE FUNCTION). *Suppose $g_{\theta,F}$ is the solution to the equation (24) and $p(s,x;\theta,g)$ is given by (11). Then the function*

$$\tilde{\ell}_{\theta_0,F_0}(s,x) = \frac{\partial}{\partial\theta}\bigg|_{\theta=\theta_0} \log p(s,x;\theta,g_{\theta,F_0}) \tag{33}$$

is the efficient score function in the model in Example 3.

Proof. We check conditions (38) and (39) in Theorem 5 in Appendix.

Condition (38) is checked in LEMMA 1.

We verify Condition (39). Note that the candidate function (13) evaluated at (θ, F_0) is

$$g_{\theta,F_0}(x) = \frac{w_{10}\int \frac{dF_{10}}{dx}}{1 - w_{20}\int \frac{f(y|x;\theta)}{f_Y(y;\theta,g_{\theta,F_0})}dF_{20}}. \tag{34}$$

Let $g_t(x)$ be a path in the space of density functions with $g_{t=0}(x) = g_0(x)$. Define $\alpha_t(x) = g_t(x) - g_0(x)$ and write $\dot{\alpha}_0(x) = \frac{\partial}{\partial t}|_{t=0}\,\alpha_t(x)$. Then

$$\begin{aligned}
&\frac{\partial}{\partial t}\bigg|_{t=0}\int \log p(s,z;\theta,g_{\theta,F_0}+\alpha_t)dF_0 \\
&= \frac{\partial}{\partial t}\bigg|_{t=0}\left[w_{10}\int\{\log f(y|x;\theta) + \log(g_{\theta,F_0}+\alpha_t)\}dF_{10} + w_{20}\int \log f_Y(y;\theta,g_{\theta,F_0}+\alpha_t)dF_{20}\right] \\
&= w_{10}\int \frac{\dot{\alpha}_0(x)}{g_{\theta,F_0}(x)}dF_{10} + w_{20}\int \frac{\int f(y|x;\theta)\dot{\alpha}_0(x)dx}{f_Y(y;\theta,g_{\theta,F_0})}dF_{20} \\
&= \int \dot{\alpha}_0(x)dx = \frac{\partial}{\partial t}\bigg|_{t=0}\int g_t(x)dx = 0 \ \text{(by (34) and since } g_t(x) \text{ is a density).}
\end{aligned}$$

□

Efficiency of the profile likelihood estimator: Let $\tilde{\ell}_{\theta,F}(s,x)$ be the score function given by (33) with θ_0 and F_0 are replaced by θ and F. LEMMA 1 shows that the score function evaluated at (θ_0, F_0) is the efficient score function in Example 3.

We verify conditions $(R0)$, $(R1)$, $(R2)$, and $(R3)$ of THEOREM 1 so that we can apply the theorem to show that the solution $\hat{\theta}_n$ to the estimating equation

$$\sum_{s=1}^{2}\sum_{i=1}^{n}\tilde{\ell}_{\hat{\theta}_n,F_n}(s,X_{si}) = 0$$

is asymptotically linear estimator with the efficient influence function, i.e., (16) holds. This shows the efficiency of the MLE based on the profile likelihood in this example.

Condition (R0): This condition is verified by LEMMA 1 and THEOREM 4.

Condition (R1): As we assumed in THEOREM 3, we assume that

(T1) For all $\theta \in \Theta$, the function $f(y|x;\theta)$ is twice continuously differentiable with respect to θ.

The maps

$$g \to \log g(x)$$

and

$$g \to f_Y(y;\theta,g) = \int_{\mathcal{X}} f(y|x;\theta)g(x)dx$$

are Hadamard differentiable (cf. Gill (1989)). It follows that the log-likelihood function

$$\log p(s,z;\theta,g) = 1_{\{s=1\}}\{\log f(y|x;\theta) + \log g(x)\} + 1_{\{s=2\}} \log f_Y(y;\theta,g)$$

is Hadamard differentiable with respect to g and, by assumption (T1), it is also twice continuously differentiable with respect to θ. We verified the function $g_{\theta,F}$ is Hadamard differentiable with respect to F and twice continuously differentiable with respect to θ. By the chain rule and product rule of Hadamard differentiable maps, the log-likelihood $\log p(s,x;\theta,g_{\theta,F})$ is Hadamard differentiable with respect to F and twice continuously differentiable with respect to θ. Therefore we verified condition (R1).

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

$$\log p(s,z;\theta,g_{\theta,F}) = 1_{\{s=1\}}\{\log f(y|x;\theta) + \log g_{\theta,F}(x)\} + 1_{\{s=2\}} \log f_Y(y;\theta,g_{\theta,F}). \quad (35)$$

The derivative of the log-likelihood with respect to θ is

$$\begin{aligned}\tilde{\ell}_{\theta,F}(s,z) &= \frac{\partial}{\partial\theta}\log p(s,z;\theta,g_{\theta,F}) \\ &= 1_{\{s=1\}}\left\{\frac{\dot{f}}{f} + \frac{\dot{g}_{\theta,F}}{g_{\theta,F}}\right\} + 1_{\{s=2\}}\frac{\dot{f}_Y + d_g f_Y(\dot{g}_{\theta,F})}{f_Y}. \end{aligned} \quad (36)$$

The second derivative of the log-likelihood function with respect to θ is

$$\begin{aligned}&\frac{\partial}{\partial\theta^T}\tilde{\ell}_{\theta,F}(s,z) = \frac{\partial^2}{\partial\theta\,\partial\theta^T}\log p(s,z;\theta,g_{\theta,F}) \\ &= 1_{\{s=1\}}\left\{\frac{\ddot{f}}{f} - \frac{\dot{f}\dot{f}^T}{f^2} + \frac{\ddot{g}_{\theta,F}}{g_{\theta,F}} - \frac{\dot{g}_{\theta,F}\dot{g}_{\theta,F}^T}{g_{\theta,F}^2}\right\} \\ &\quad +1_{\{s=2\}}\left\{\frac{\ddot{f}_Y + d_g\dot{f}_Y(\dot{g}_{\theta,F})}{f_Y} - \frac{\dot{f}_Y\dot{f}_Y^T + \dot{f}_Y d_g f_Y(\dot{g}_{\theta,F}^T)}{f_Y^2}\right. \\ &\qquad \left. +\frac{d_g\dot{f}_Y^T(\dot{g}_{\theta,F}) + d_g f_Y(\ddot{g}_{\theta,F})}{f_Y} - \frac{d_g f_Y(\dot{g}_{\theta,F})\dot{f}_Y^T + d_g f_Y(\dot{g}_{\theta,F}) d_g f_Y(\dot{g}_{\theta,F}^T)}{f_Y^2}\right\} \end{aligned} \quad (37)$$

Here we used the notations $\dot{f}_Y = \dot{f}_Y(y;\theta,g_{\theta,F})$, $\ddot{f}_Y = \ddot{f}_Y(y;\theta,g_{\theta,F})$, $d_g f_Y(g_{\theta,F}) = \int f(y|x;\theta)g_{\theta,F}(x)dx$, and $d_g\dot{f}_Y(g_{\theta,F}) = \int \dot{f}(y|x;\theta)g_{\theta,F}(x)dx$.

Condition (R2): We assume that

(T2) There is no $a \in \mathbb{R}^d$ such that $a^T\frac{\dot{f}}{f}(y|x;\theta)$ is constant in y for almost all x.

The term $\frac{\dot{g}_{\theta,F}}{g_{\theta,F}}(x,\theta_0,F_0)$ is a function of x. Therefore, by Equation (36) and assumption (T2), there is no $a \in \mathbb{R}^d$ such that $a^T\tilde{\ell}_{\theta,F}(1,z)$ is constant in y for almost all x. By THEOREM 1.4 in Seber and Lee (2003), $E_{1,\theta_0,F_0}(\tilde{\ell}_{\theta_0,F_0}\tilde{\ell}_{\theta_0,F_0}^T)$ is non-singular with the bounded inverse.

Conditions (R3): Since verification of Condition (R3) require more assumptions and it does not add anything new, we omit this. Instead, we assume:

(T3) Let $\mathcal{F}$ be the set of cdf functions and for some $\rho > 0$ define $\mathcal{C}_\rho = \{F \in \mathcal{F} : \|F - F_0\|_\infty \leq \rho\}$. The class of function

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

is P_{θ_0,g_0}-Donsker with square integrable envelope function and the class

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

is P_{θ_0,g_0}-Glivenko-Cantelli with integrable envelope function.

6. DISCUSSION

We have shown the differentiability of implicitly defined function which we encounter in the maximum likelihood estimation in semiparametric model. We assumed the implicitly defined function is the solution to the operator equation (4) and in THEOREM 2 we obtained the derivatives of the (implicitly defined) function. In some application, it may be difficult to show the condition (A3) in the theorem (that is $\|d_\eta \Psi_{\theta_0,F_0}(\eta_0)\| < 1$). The future work is to relax the condition to $\|d_\eta \Psi_{\theta_0,F_0}(\eta_0)\| < \infty$. Once the differentiability of the implicitly defined function has been established, the results in Hirose (2010) (we summarized in SECTION 3) are applicable.

APPENDIX

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

THEOREM 5. *We assume the general semi-parametric model given in "Introduction" with the density $p_{\theta,\eta}(x) = p(x;\theta,\eta)$ is differentiable with respect to θ and Hadmard differentiable with respect to η. Suppose g_t is an arbitrary path such that $g_{t=0} = g_0$ and let $\alpha_t = g_t - g_0$. If $g_{\theta,F}$ is a function of (θ, F) such that*

$$g_{\theta_0,F_0} = g_0 \tag{38}$$

and, for each $\theta \in \Theta$,

$$\left.\frac{\partial}{\partial t}\right|_{t=0} E_0\left[\log p(x;\theta, g_{\theta,F_0} + \alpha_t)\right] = 0, \tag{39}$$

then the function $\tilde{\ell}_{\theta_0,F_0}(x) = \frac{\partial}{\partial\theta}|_{\theta=\theta_0} \log p(x;\theta, g_{\theta,F_0})$ is the efficient score function.

Proof. Condition (39) implies that

$$\begin{aligned} 0 &= \left.\frac{\partial}{\partial\theta}\right|_{\theta=\theta_0} \left.\frac{\partial}{\partial t}\right|_{t=0} E_0\left[\log p(x;\theta, g_{\theta,F_0} + \alpha_t)\right] \\ &= \left.\frac{\partial}{\partial t}\right|_{t=0} E_0\left[\left.\frac{\partial}{\partial\theta}\right|_{\theta=\theta_0} \log p(x;\theta, g_{\theta,F_0} + \alpha_t)\right]. \end{aligned} \tag{40}$$

By differentiating the identity

$$\int \left(\frac{\partial}{\partial \theta} \log p(x; \theta, g_{\beta,F_0} + \alpha_t) \right) p(x; \theta, g_{\beta,F_0} + \alpha_t) dx = 0$$

with respect to t at $t = 0$ and $\theta = \theta_0$, we get

$$\begin{aligned} 0 &= \left. \frac{\partial}{\partial t} \right|_{t=0,\theta=\theta_0} \int \left(\frac{\partial}{\partial \theta} \log p(x; \theta, g_{\theta,F_0} + \alpha_t) \right) p(x; \theta, g_{\theta,F_0} + \alpha_t) dx \\ &= E_0 \left[\tilde{\ell}_{\theta_0,F_0}(x) \left(\left. \frac{\partial}{\partial t} \right|_{t=0} \log p(x; \theta_0, g_t) \right) \right] \quad (\text{ by (38)}) \\ &\quad + \left. \frac{\partial}{\partial t} \right|_{t=0} E_0 \left[\left. \frac{\partial}{\partial \theta} \right|_{\theta=\theta_0} \log p(x; \theta, g_{\theta,F_0} + \alpha_t) \right] \\ &= E_0 \left[\tilde{\ell}_{\theta_0,F_0}(x) \left(\left. \frac{\partial}{\partial t} \right|_{t=0} \log p(x; \theta_0, g_t) \right) \right] \quad (\text{ by (40)}). \end{aligned} \tag{41}$$

Let $c \in R^m$ be arbitrary. Then, it follows from Equation (41) that the product $c' \tilde{\ell}_{\theta_0,F_0}(x)$ is orthogonal to the nuisance tangent space $\dot{\mathcal{P}}_g$ which is the closed linear span of score functions of the form $\frac{\partial}{\partial t}|_{t=0} \log p(x; \beta_0, g_t)$.

Using Condition (38), we have

$$\begin{aligned} \tilde{\ell}_{\theta_0,F_0}(x) &= \left. \frac{\partial}{\partial \theta} \right|_{\theta=\theta_0} \log p(x; \theta, g_0) + \left. \frac{\partial}{\partial \beta} \right|_{\theta=\theta_0} \log p(x; \theta_0, g_{\theta,F_0}) \\ &= \dot{\ell}_{\theta_0,g_0}(x) - \psi_{\theta_0,g_0}(x), \end{aligned}$$

where $\dot{\ell}_{\theta_0,g_0}(x) = \frac{\partial}{\partial \theta}|_{\theta=\theta_0} \log p(x; \theta, g_0)$ is the score function for θ and $\psi_{\theta_0,g_0}(x) = -\frac{\partial}{\partial \theta}|_{\theta=\theta_0} \log p(x; \theta_0, g_{\theta,F_0})$. Finally, $c' \tilde{\ell}_{\theta_0,F_0}(x) = c' \dot{\ell}_{\theta_0,g_0}(x) - c' \psi_{\theta_0,g_0}(x)$ is orthogonal to the nuisance tangent space $\dot{\mathcal{P}}_g$ and $c' \psi_{\theta_0,g_0}(x) \in \dot{\mathcal{P}}_g$ implies that $c' \psi_{\theta_0,g_0}(x)$ is the orthogonal projection of $c' \dot{\ell}_{\theta_0,g_0}(x)$ onto the nuisance tangent space $\dot{\mathcal{P}}_g$. Since $c \in R^m$ is arbitrary, $\tilde{\ell}_{\theta_0,F_0}(x)$ is the efficient score function.

REFERENCES

AGARWAL, R.P., O'REGAN, D. AND SAHU, D.R. (2009). *Fixed Point Theory for Lipschitzian-type Mappings with Applications.* Springer, New York.

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 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. (2000). Large sample theory for semiparametric regression models with two-phase outcome dependent sampling. Technical Report 381, Dept. Statistics, Univ. Washington.

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., ROBINS, J.M. AND WELLNER, J.A. (2000). On the semi-parametric efficiency of logistic regression under case-control sampling. *Bernoulli* **6** 447–455.

GILBERT, P.G. (2000). Large sample theory of maximum likelihood estimation in semiparametric selection biased sampling models *Ann. Statist.* **28** 151–194.

GILBERT, P.G., LELE, S.R. AND 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. AND WELLNER, J.A. (1988). Large sample theory of empirical distribution in biased sampling models *Ann. Statist.* **3** 1069–1112.

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

Hirose, Y. (2010). Efficiency of profile likelihood in semi-parametric models, *Ann. Inst. Statist. Math.* DOI 10.1007/s10463-010-0280-y.

HIROSE, Y. AND LEE, A.J. (2010). Reparametrization of the Least Favorable Submodel in Semi-Parametric Multi-Sample Models, Submitted to Bernoulli.

KOLMOGOROV, A.N. AND FOMIN, S.V. (1975). *Introductory Real Analysis.* Dover, New York.

KOLMOGOROV, M.R. (2008). *Introduction to Empirical Processes and Semiparametric Inference* Springer, New York.

LAWLESS, J.L., KALBFLEISH, J.D. AND WILD, C.J. (1999). Estimation 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. (2008). Semi-parametric efficiency bounds for regression models under case-control sampling: the profile likelihood approach, Ann. Inst. Statist. Math. DOI 10.1007/s10463-008-0205-1

MURPHY, S.A., ROSSINI, A.J. AND VAN DER VAART, A.W. (1997). Maximum likelihood estimation in the proportional odds model. *J. Amer. Statist. Assoc.* **92** 968–976.

MURPHY, S.A. AND VAN DER VAART, A.W. (2000). On profile likelihood (with discussion). *J. Amer. Statist. Assoc.* **95** 449–485.

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.

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.

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.

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

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

SONG, R., ZHOU, H. AND KOSOROK, M.R. (2009). A note on semiparametric efficient inference for two-stage outcome-dependent sampling with a continuous outcome *Biometrika* **96** 221–228.

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.

VARDI, Y. (1985). Empirical distributions in selection bisa models *Ann. Statist.* **13** 178–203.

WEAVER, M.A. AND ZHOU, H. (2005). An estimated likelihood method for continuous outcome regression models with outcome-dependent sampling. *J. Amer. Statist. Assoc.* **100**, 459–69.

[*Received November* 2009. *Revised xxxx* 20xx]