Bernoulli **18**(2), 2012, 586–605
DOI: 10.3150/10-BEJ342

Reparametrization of the least favorable submodel in semi-parametric multisample models

YUICHI HIROSE[1] and ALAN LEE[2]

[1]*School of Mathematics, Statistics and Operations Research, Victoria University of Wellington, New Zealand. E-mail: Yuichi.Hirose@msor.vuw.ac.nz*
[2]*Department of Statistics, University of Auckland, New Zealand. E-mail: lee@stat.auckland.ac.nz*

The method of estimation in Scott and Wild (*Biometrika* **84** (1997) 57–71 and *J. Statist. Plann. Inference* **96** (2001) 3–27) uses a reparametrization of the profile likelihood that often reduces the computation times dramatically. Showing the efficiency of estimators for this method has been a challenging problem. In this paper, we try to solve the problem by investigating conditions under which the efficient score function and the efficient information matrix can be expressed in terms of the parameters in the reparametrized model.

Keywords: efficiency; efficient information bound; efficient score; multisample; profile likelihood; semi-parametric model

1. Introduction

In a series of papers, Scott and Wild [12,13] developed methods of reparametrization of profile likelihood that can be applied to a variety of response-selective sampling designs. The advantage of the methods is that they often give us computationally efficient estimators. The (statistical) efficiency of these methods has been demonstrated in special cases by several authors. For example, Breslow, Robins and Wellner [3] considered case-control sampling where either a case or control is selected by a randomization device with known selection probabilities, and the covariates of the resulting case or control are measured. In the case of two-phase, outcome-dependent sampling, Breslow, McNeney and Wellner [2] applied the missing value theory of Robins, Rotnitzky and Zhao [11] and Robins, Hsieh and Newey [10]. Here, individuals in the population are selected at random and their status (e.g., case or control) is determined. Then, with a probability depending on their status, the covariates are measured. The unobserved covariates are treated as missing data. Lee and Hirose [8] used the profile likelihood method to derive a semi-parametric efficiency bound, and then showed that this bound coincides with the asymptotic variance of the Scott–Wild estimator, hence demonstrating the efficiency of the estimator.

In Lee and Hirose [8], it was demonstrated that, in the case of the Scott–Wild estimator, it is possible to reparametrize the least favorable submodel so that the efficient score function and the efficient information matrix can be expressed in terms of the parameters in the reparametrized model.

The aim of this paper is to investigate conditions under which a reparametrization of the least favorable submodel yields an efficient estimation.

We consider an S-vector of semi-parametric models $(\mathcal{P}_1, \ldots, \mathcal{P}_S)$ where, for each $s = 1, \ldots, S$,

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

is a probability model on the sample space $\mathcal{X}_s$ with the parameter of interest β, an m-dimensional parameter, and the nuisance parameter η, which may be an infinite-dimensional parameter. 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 R^m, and Θ_η is a convex set containing η_0 in a Banach space $\mathcal{B}$. We refer to the S-vector of semi-parametric models $(\mathcal{P}_1, \ldots, \mathcal{P}_S)$ as the multisample model.

Under the model, we observe S independent samples $X_{s1}, \ldots, X_{sn_s}$ $(s = 1, \ldots, S)$, where $X_{s1}, \ldots, X_{sn_s}$ are independently and identically distributed (i.i.d.) according to the model $\mathcal{P}_s$. Let $n = \sum_{s=1}^S n_s$. We assume the sample size proportions $(n_1/n, \ldots, n_S/n)$ converge to weight probabilities $(w_1, \ldots, w_S)$:

$$\left(\frac{n_1}{n}, \ldots, \frac{n_S}{n}\right) \to (w_1, \ldots, w_S), \tag{1.1}$$

where $w_s > 0$ and $\sum_{s=1}^S w_s = 1$.

The log-likelihood for the multisample data is

$$\ell_n(\beta, \eta) = \sum_{s=1}^S \sum_{i=1}^{n_s} \log p_s(X_{si}; \beta, \eta). \tag{1.2}$$

The paper is organized as follows: In the rest of Section 1, we give examples of semi-parametric multisample models. In Section 2, we introduce the least favorable submodel in multisample models and in Section 3, we present the main result of conditions under which reparametrization gives efficient estimators in multisample models. In Section 4, we give a numerical example and use the result developed in the paper to show that the estimators in the example are efficient.

1.1. Examples

The idea of multisample data is familiar from elementary statistics; for example, the well-known two-sample t-test and the one-way ANOVA for comparing several means both involve multiple samples. Following are several semi-parametric examples.

Example 1 (Biased sampling model). Vardi [14] developed the method of estimation in the S-sample biased sampling model with known selection bias weight functions. The following setup and notation are from [6].

Suppose that non-negative weight functions $w_1(x), \ldots, w_S(x)$ are given and let $G(x)$ be an unknown distribution function on a sample space $\mathcal{X}$. Define the corresponding biased sampling model by

$$p_s(x; G) = \frac{w_s(x)g(x)}{W_s(G)} \qquad (s = 1, \ldots, S),$$

where $g(x) = \mathrm{d}G(x)/\mathrm{d}\mu$ with respect to Lebesgue measure μ and $W_s(G) = \int_{\mathcal{X}} w_s(x)\,\mathrm{d}G(x)$. The S-sample biased sampling model generates S independent samples

$$X_{s1}, \ldots, X_{sn_s} \sim p_s(x; G) \qquad (s = 1, \ldots, S).$$

Gilbert, Lele and Vardi [5] considered an extension of this model that allows the weight function to depend on an unknown finite-dimensional parameter θ.

Suppose a set of non-negative weight functions $w_1(x, \theta), \ldots, w_S(x, \theta)$ depend on θ. The semi-parametric biased sampling model is defined by

$$p_s(x; \theta, G) = \frac{w_s(x, \theta)g(x)}{W_S(\theta, G)} \qquad (s = 1, \ldots, S),$$

where $W_s(\theta, G) = \int_{\mathcal{X}} w_s(x, \theta)\,\mathrm{d}G(x)$. Gilbert [4] provides a large sample theory of this example.

The following examples are semi-parametric multisample models that all have the same underlying data-generating process on the sample space $\mathcal{Y} \times \mathcal{X}$, called the full data model,

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

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

Example 2 (Case-control study). We assume that Y takes values in $\{1, \ldots, S\}$. In a case-control study, due to the design, we do not observe a random sample from the full data model $\mathcal{Q}$. Instead, for each $s = 1, \ldots, S$, we observe n_s-samples from the conditional distribution $P(X|Y = s)$. By Bayes' theorem, the density of $P(X|Y = s)$ is

$$\frac{f(s|x; \theta)g(x)}{\int f(s|x; \theta)\,\mathrm{d}G(x)}.$$

The case-control study is a special case of the semi-parametric biased sampling model of Example 1 with weight functions $w_s(x, \theta) = f(s|x; \theta)$ $(s = 1, \ldots, S)$.

Example 3 (Missing data). Instead of observing full data (Y, X) from the full data model $\mathcal{Q}$ for all individuals, we observe (Y, X) for n_0-samples and observe Y for n_1-samples. The result is the multisample data

$$(x_{01}, y_{01}), \ldots, (x_{0n_0}, y_{0n_0}), y_{11}, \ldots, y_{1n_1}$$

from a multisample model with densities

$$p_0(y, x; \theta, g) = f(y|x; \theta)g(x)$$

and

$$p_1(y;\theta,g)=\int f(y|x;\theta)g(x)\,\mathrm{d}x.$$

This example is not a special case of Example 1.

Example 4 (Standard stratified sampling and two-phase, outcome-dependent sampling). For a partition of the sample space $\mathcal{Y}\times\mathcal{X}=\bigcup_{s=1}^{S}\mathcal{S}_s$, let

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

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

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

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

of (Y,X) given stratum $\mathcal{S}_s$. This is a more general version of the semi-parametric biased sampling model of Example 1 with weight functions $w_s(y,x,\theta)=f(y|x;\theta)1_{(y,x)\in\mathcal{S}_s}$ $(s=1,\ldots,S)$.

Lawless, Kalbfleisch and Wild [7] discussed variations of the two-phase, outcome-dependent sampling design (the variable probability sampling designs (VPS1, VPS2) and the basic stratified sampling design (BSS)). For all sampling schemes (VPS1, VPS2 and BSS), we have m_s fully observed units and n_s-m_s subjects where the only information retained is the identity of the stratum, $s=1,\ldots,S$. The corresponding likelihood is

$$L(\theta,G)=\left\{\prod_{s=1}^{S}\prod_{i=1}^{m_s}f(y_{si}|x_{si};\theta)g(x_{si})\right\}\left\{\prod_{s=1}^{S}Q_s(\theta,G)^{n_s-m_s}\right\}. \tag{1.3}$$

We interpret the observed data from two-phase, outcome-dependent sampling as data from a multisample model with densities

$$p_1(y,x;\theta,G)=f(y|x;\theta)g(x)$$

and

$$p_2(s;\theta,G)=Q_s(\theta,G).$$

This example is not a special case of Example 1.

2. The least favorable submodel

The *log-likelihood function* for a single observation in the multisample model is

$$\ell(s,x;\beta,\eta)=\log p_s(x;\beta,\eta)\qquad(x\in\mathcal{X}_s,\ s=1,\ldots,S). \tag{2.1}$$

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

We assume that there is a differentiable function $\beta \to \hat{\eta}_\beta$ such that

$$\hat{\eta}_{\beta_0} = \eta_0 \tag{2.2}$$

and

$$\dot{\ell}^*(s,x) = \frac{\partial}{\partial\beta}\bigg|_{\beta=\beta_0} \ell(s,x,\beta;\hat{\eta}_\beta) \tag{2.3}$$

is the efficient score function (definition of the efficient score function in the multisample model is given in Appendix A). We call the model

$$p_s(x;\beta,\hat{\eta}_\beta) \qquad (\beta \in \Theta_\beta, s = 1,\dots,S),$$

the least favorable submodel for the multisample model $(\mathcal{P}_1,\dots,\mathcal{P}_S)$.

Remark 2.1. Under mild regularity conditions with the assumption that

$$\hat{\eta}_\beta = \arg\max_{\eta\in\Theta_\eta} \sum_{s=1}^{S} w_s E_{s,\beta_0,\eta_0}\{\log p_s(X;\beta,\eta)\}$$

exists for all β in some neighborhood of β_0, (2.3) is the efficient score function due to [9]. The definition of the least favorable submodel given above includes this as a special case but we do not limit our consideration only in this case.

Our approach uses the method in Scott and Wild [12,13] to find a candidate function $\hat{\eta}_\beta$ as well as Theorem A.2 in Appendix A to verify that (2.3) with the candidate function gives the efficient score function. In the next example we illustrate this procedure.

2.1. Example: Stratified sampling (continued)

Stratified sampling was introduced in Example 4.

Let

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

For each $s = 1,\dots,S$, let F_{s0} be the cumulative distribution function for the density $p_s(y,x;\theta_0,g_0)$ at the true value (θ_0,g_0). The expected likelihood in the model is

$$\sum_{s=1}^{S} w_s E_{s,0}\{\log p_s(y,x;\theta,g)\} = \sum_{s=1}^{S} w_s \int \log p_s(y,x;\theta,g)\,\mathrm{d}F_{s0}(y,x).$$

For each θ, the method in Scott and Wild [12,13] finds a maximizer $\hat{g}_\theta(x)$ of log-likelihood under the assumption that the support of the distribution of X is finite; that is,

$\text{SUPP}(X) = \{v_1, \ldots, v_K\}$. 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 the expected log-likelihood

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

at θ, differentiate this expression with respect to g_k and set the derivative equal to zero,

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

The solution g_k to the equation is

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

The form of the function motivates us to prove the following result.

Lemma 2.1 (The least favorable submodel). *For $\theta \in \Theta$, let*

$$\hat{g}_\theta(x) = \frac{f_0^*(x)}{\sum_{s=1}^{S} w_s Q_{s|X}(x;\theta)/\hat{Q}_s(\theta)}, \tag{2.4}$$

where

$$f_0^*(x) = \sum_{s=1}^{S} w_s \frac{Q_{s|X}(x;\theta_0) g_0(x)}{Q_s(\theta_0, g_0)}, \tag{2.5}$$

and

$$\hat{Q}_s(\theta) = \int Q_{s|X}(x;\theta)\hat{g}_\theta(x)\,\mathrm{d}x \qquad (s = 1, \ldots, S). \tag{2.6}$$

Then the efficient score function is given by

$$\dot{\ell}^*(s, y, x) = \frac{\partial}{\partial\theta}\bigg|_{\theta=\theta_0} \log p_s(y, x; \theta, \hat{g}_\theta). \tag{2.7}$$

Proof. In Appendix B, we show that $\sum_{s=1}^{S} w_s \int \log p_s(y, x; \theta, \hat{g}_\theta)\,\mathrm{d}F_{s0}$ satisfies conditions (A.1) and (A.2) in Theorem A.2 in Appendix A so that the claim follows from this theorem. □

Remark 2.2. Note that equations (2.4) and (2.6) are consistent at $\theta = \theta_0$: (2.4) and (2.5) imply that $\hat{g}_{\theta_0}(x) = g_0(x)$ if $\hat{Q}_s(\theta_0) = Q_s(\theta_0, g_0)$. On the other hand, if $\hat{g}_{\theta_0}(x) = g_0(x)$, we have $\hat{Q}_s(\theta_0) = \int Q_{s|X}(x;\theta_0) g_0(x)\,\mathrm{d}x = Q_s(\theta_0, g_0)$ by (2.6).

3. Main result

Suppose there is a finite-dimensional, vector-valued function $\beta \to q_\beta$ such that the density for the least favorable submodel is of the form

$$p_s(x;\beta,\hat{\eta}_\beta) = p_s^*(x;\beta,q_\beta) \qquad \text{for all } \beta \in \Theta_\beta \ (s=1,\dots,S), \tag{3.1}$$

where the function $p_s^*(x;\beta,q)$ is twice continuously differentiable with respect to (β,q) and q is a finite-dimensional parameter. Further, suppose

$$\sum_{s=1}^{S} w_s \int p_s^*(x;\beta,q)\,\mathrm{d}x = 1 \qquad \text{for all } (\beta,q) \in \Theta_\beta \times D_q, \tag{3.2}$$

where Θ_β and D_q are neighborhoods of β_0 and q_{β_0}, respectively. Then the model

$$p_s^*(x;\beta,q) \qquad (\beta \in \Theta_\beta, q \in D_q, s=1,\dots,S),$$

is called a *reparametrized model* for the least favorable submodel. The score functions for β and q in the reparametrized model are denoted by $\dot{\ell}_1(s,x;\beta,q) = (\partial/\partial\beta)\log p_s^*(x;\beta,q)$ and $\dot{\ell}_2(s,x;\beta,q) = (\partial/\partial q)\log p_s^*(x;\beta,q)$, respectively.

Remark 3.1. In general, we may not have the condition

$$\int p_s^*(x;\beta,q)\,\mathrm{d}x = 1 \qquad \text{for all } (\beta,q) \in \Theta_\beta \times D_q \ (s=1,\dots,S).$$

Therefore, there is no guarantee that each $p_s^*(x;\beta,q)$ is a probability model. However, (3.2) ensures that the linear combination $\sum_{s=1}^{S} w_s p_s^*(x;\beta,q)$ acts like a probability model. This looks like a mixture model. The main differences between the multisample model and the mixture model are data and asymptotics. For example, the log-likelihood and the information matrix in the mixture model are, respectively,

$$\sum_{i=1}^{n} \log\left\{\sum_{s=1}^{s} w_s p_s(x_i;\beta,q)\right\}$$

and

$$\int \left(\frac{(\partial/\partial(\beta,q))\sum_{s=1}^{S} w_s p_s(x;\beta,q)}{\sum_{s=1}^{S} w_s p_s(x;\beta,q)}\right)^{\otimes 2} \sum_{s} w_s p_s(x;\beta,q)\,\mathrm{d}x,$$

while the log-likelihood and the information matrix in the multisample model are given by, respectively, (1.2) and

$$\sum_{s=1}^{S} w_s \int \left(\frac{(\partial/\partial(\beta,q))p_s(x;\beta,q)}{p_s(x;\beta,q)}\right)^{\otimes 2} p_s(x;\beta,q)\,\mathrm{d}x.$$

Remark 3.2. Note that, since $q_{\beta_0} = \hat{\eta}_{\beta_0} = \eta_0$, we have $p_s(x; \beta_0, \eta_0) = p_s^*(x; \beta_0, q_{\beta_0})$ $(s = 1, \ldots, S)$. Therefore, for the reparametrized model, the notation $E_{s,0}$, $s = 1, \ldots, S$ is used for the expectations at the true value (β_0, q_{β_0}).

For a measurable function $f(s, x; \beta, q)$, define the *centering* of $f(s, x; \beta, q)$ by

$$f^c(s, x; \beta, q) = f(s, x; \beta, q) - E_{s,0}\{f(s, x; \beta_0, q_{\beta_0})\}.$$

The function $f^c(s, x; \beta, q)$ is called the *centered* $f(s, x; \beta, q)$.

Theorem 3.1 (Efficiency in a reparametrized model). *We assume that the least favorable submodel and the corresponding reparametrized model are as in* (2.2), (2.3), (3.1) *and* (3.2). *Further, assume that*

$$\left.\frac{\partial}{\partial q}\right|_{q=q_\beta} \sum_{s=1}^{S} w_s E_{s,0}\{\log p_s^*(x; \beta, q)\} = 0 \qquad \text{for } \beta \in \Theta_\beta \tag{3.3}$$

and $\sum_{s=1}^{S} w_s E_{s,0}(\dot{\ell}_2^c \dot{\ell}_2^{cT})$ *is non-singular. Then the efficient score function and the efficient information matrix in the original multisample model* $(\mathcal{P}_1, \ldots, \mathcal{P}_s)$ *are given by*

$$\dot{\ell}^*(s, x) = \dot{\ell}_1^c - \left\{\sum_{s=1}^{S} w_s E_{s,0}(\dot{\ell}_1^c \dot{\ell}_2^{cT})\right\}\left\{\sum_{s=1}^{S} w_s E_{s,0}(\dot{\ell}_2^c \dot{\ell}_2^{cT})\right\}^{-1} \dot{\ell}_2^c \tag{3.4}$$

and

$$\begin{aligned} I^* = &\sum_{s=1}^{S} w_s E_{s,0}(\dot{\ell}_1^c \dot{\ell}_1^{cT}) \\ &- \left\{\sum_{s=1}^{S} w_s E_{s,0}(\dot{\ell}_1^c \dot{\ell}_2^{cT})\right\}\left\{\sum_{s=1}^{S} w_s E_{s,0}(\dot{\ell}_2^c \dot{\ell}_2^{cT})\right\}^{-1}\left\{\sum_{s=1}^{S} w_s E_{s,0}(\dot{\ell}_2^c \dot{\ell}_1^{cT})\right\}, \end{aligned} \tag{3.5}$$

where $\dot{\ell}_1^c(s, x; \beta, q)$ *and* $\dot{\ell}_2^c(s, x; \beta, q)$ *are the centered score functions for* β *and* q *in the reparametrized model, respectively.*

Proof. By (2.3) and (3.1), the efficient score function is given by

$$\dot{\ell}^*(s, x) = \left.\frac{\partial}{\partial \beta}\right|_{\beta=\beta_0} \log p_s^*(x; \beta, q_\beta) = \dot{\ell}_1(s, x; \beta_0, q_{\beta_0}) + \dot{q}_{\beta_0}^T \dot{\ell}_2(s, x; \beta_0, q_{\beta_0}). \tag{3.6}$$

Since $E_{s,\beta_0\eta_0}\{\dot{\ell}^*(s, X)\} = 0$ $(s = 1, \ldots, S)$, we have

$$E_{s,\beta_0\eta_0}\{\dot{\ell}_1(s, x; \beta_0, q_{\beta_0})\} + \dot{q}_{\beta_0}^T E_{s,\beta_0\eta_0}\{\dot{\ell}_2(s, x; \beta_0, q_{\beta_0})\} = 0 \qquad (s = 1, \ldots, S). \tag{3.7}$$

Therefore, (3.6) and (3.7) imply

$$\dot{\ell}^*(s, x) = \dot{\ell}_1^c(s, x; \beta_0, q_{\beta_0}) + \dot{q}_{\beta_0}^T \dot{\ell}_2^c(s, x; \beta_0, q_{\beta_0}). \tag{3.8}$$

By differentiating (3.2) with respect to q, for all $(\beta, q) \in \Theta_\beta \times D_q$, we have

$$\sum_{s=1}^{S} w_s \int \dot{\ell}_2(s, x; \beta, q) p_s^*(x; \beta, q)\, \mathrm{d}x = 0.$$

In particular, for all $\beta \in \Theta_\beta$,

$$\sum_{s=1}^{S} w_s \int \dot{\ell}_2(s, x; \beta, q_\beta) p_s^*(x; \beta, q_\beta)\, \mathrm{d}x = 0.$$

By differentiating with respect to β at β_0,

$$\begin{aligned}
&\sum_{s=1}^{S} w_s \int \left(\frac{\partial}{\partial \beta}\bigg|_{\beta=\beta_0} \dot{\ell}_2(s, x; \beta, q_\beta) \right) p_s^*(x; \beta_0, q_{\beta_0})\, \mathrm{d}x \\
&\quad = -\sum_{s=1}^{S} w_s \int \dot{\ell}_2(s, x; \beta_0, q_{\beta_0}) \left(\frac{\partial}{\partial \beta}\bigg|_{\beta=\beta_0} p_s^*(x; \beta, q_\beta) \right) \mathrm{d}x.
\end{aligned}$$

By the first equality in (3.6), this equation is equivalent to

$$\sum_{s=1}^{S} w_s E_{s,0}\left\{ \frac{\partial}{\partial \beta}\bigg|_{\beta=\beta_0} \dot{\ell}_2(s, x; \beta, q_\beta) \right\} = -\sum_{s=1}^{S} w_s E_{s,0}(\dot{\ell}_2 \dot{\ell}^{*T}). \tag{3.9}$$

By differentiating (3.3) with respect to β at β_0, we get

$$\begin{aligned}
0 &= \frac{\partial}{\partial \beta}\bigg|_{\beta=\beta_0} \frac{\partial}{\partial q}\bigg|_{q=q_\beta} \sum_{s=1}^{S} w_s E_{s,0}\{\log p_s^*(x; \beta, q)\} = \sum_{s=1}^{S} w_s E_{s,0}\left\{ \frac{\partial}{\partial \beta}\bigg|_{\beta=\beta_0} \dot{\ell}_2(s, x, \beta, q_\beta) \right\} \\
&= -\sum_{s=1}^{S} w_s E_{s,0}(\dot{\ell}_2 \dot{\ell}^{*T}) = -\sum_{s=1}^{S} w_s E_{s,0}(\dot{\ell}_2^c \dot{\ell}^{*T}),
\end{aligned}$$

where we used (3.9) and $E_{s,0}\{\dot{\ell}^*(s, X)\} = 0$ $(s = 1, \ldots, S)$.

Therefore, the centered score function $\dot{\ell}_2^c(s, x; \beta_0, q_{\beta_0})$ and the efficient score function $\dot{\ell}^*(s, x)$ are uncorrelated. Since $\dot{\ell}^* = \dot{\ell}_1^c + \dot{q}_{\beta_0}^T \dot{\ell}_2^c$ (cf. (3.8)), by the projection theorem (Theorem A.1 in Appendix A), we have

$$\dot{q}_{\beta_0}^T \dot{\ell}_2^c = -\left\{ \sum_{s=1}^{S} w_s E_{s,0}(\dot{\ell}_1^c \dot{\ell}_2^{cT}) \right\} \left\{ \sum_{s=1}^{S} w_s E_{s,0}(\dot{\ell}_2^c \dot{\ell}_2^{cT}) \right\}^{-1} \dot{\ell}_2^c.$$

The rest of the claims follow by substituting this expression into (3.8). □

Remark 3.3. Under the usual regularity conditions, the solution $(\hat{\beta}_n, \hat{q}_n)$ to the system of the score equations,

$$\begin{cases} \sum_{s=1}^{S} \sum_{i=1}^{n_i} \dot{\ell}_1(s, X_{si}; \hat{\beta}_n, \hat{q}_n) = 0, \\ \sum_{s=1}^{S} \sum_{i=1}^{n_i} \dot{\ell}_2(s, X_{si}; \hat{\beta}_n, \hat{q}_n) = 0, \end{cases}$$

is asymptotically distributed as

$$\begin{Bmatrix} n^{1/2}(\hat{\beta}_n - \beta_0) \\ n^{1/2}(\hat{q}_n - q_0) \end{Bmatrix} \sim N\left\{ \begin{pmatrix} 0 \\ 0 \end{pmatrix}, \Sigma^{-1} \right\},$$

where

$$\Sigma = \begin{Bmatrix} \sum_{s=1}^{S} w_s E_{s,0}(\dot{\ell}_1^c \dot{\ell}_1^{cT}), \sum_{s=1}^{S} w_s E_{s,0}(\dot{\ell}_1^c \dot{\ell}_2^{cT}) \\ \sum_{s=1}^{S} w_s E_{s,0}(\dot{\ell}_2^c \dot{\ell}_1^{cT}), \sum_{s=1}^{S} w_s E_{s,0}(\dot{\ell}_2^c \dot{\ell}_2^{cT}) \end{Bmatrix}.$$

Then the asymptotic variance of $n^{1/2}(\hat{\beta}_n - \beta_0)$ is given by $(I^*)^{-1}$, where I^* is the efficient information for β given by (3.5) (cf. Bickel *et al.* [1], page 28). In this case, the estimator $\hat{\beta}_n$ is efficient. This efficiency of the estimator based on the reparametrization is demonstrated in a numerical example given in Section 4.

3.1. Example: Stratified sampling (continued)

In this section, we illustrate the use of Theorem 3.1 to derive the expressions of the efficient score function and the efficient information bound in terms of the parameters in a reparametrized form of the least favorable submodel in the stratified sampling example.

Lemma 2.1 gives the least favorable submodel with densities

$$p_s(y, x; \theta, \hat{g}_\theta) = \frac{f(y|x; \theta) 1_{(y,s) \in \mathcal{S}_s} \hat{g}_\theta(x)}{\hat{Q}_s(\theta)} \qquad (s = 1, \ldots, S),$$

where $\hat{g}_\theta$ is given by (2.4). By replacing $\hat{Q}(\theta) = (\hat{Q}_1(\theta), \ldots, \hat{Q}_{S-1}(\theta), \hat{Q}_S(\theta))$ with $q = (q_1, \ldots, q_{S-1}, 1)$, we consider a reparametrized model of the form

$$p_s^*(y, x; \theta, q) = \frac{f(y|x; \theta) 1_{(y,s) \in \mathcal{S}_s} \hat{g}_{\theta,q}(x)}{q_s} \qquad (s = 1, \ldots, S), \tag{3.10}$$

where

$$\hat{g}_{\theta,q}(x) = \frac{f_0^*(x)}{\sum_{s=1}^{S} w_s Q_{s|X}(x; \theta)/q_s} \tag{3.11}$$

with $f_0^*(x)$ given by (2.5).

The true value of (θ, q) is

$$(\theta_0, q_0) = \left(\theta_0, \left(\frac{Q_1(\theta_0, g_0)}{Q_S(\theta_0, g_0)}, \ldots, \frac{Q_{S-1}(\theta_0, g_0)}{Q_S(\theta_0, g_0)}, 1\right)\right).$$

Let D_q be some neighborhood of q_0.

We will demonstrate that the conditions in Theorem 3.1 are satisfied, so that we can apply the theorem to identify the efficient score function and the efficient information matrix in the example.

First, we will show that

$$\sum_{s=1}^{S} w_s \int p_s^*(y, x; \theta, q)\, \mathrm{d}y\, \mathrm{d}x = 1 \qquad \text{for all } (\theta, q) \in \Theta_0 \times D_q.$$

For any (θ, q), since $Q_{s|X}(x; \theta) = \int f(y|x; \theta) 1_{(y,s)\in\mathcal{S}_s}\, \mathrm{d}y$,

$$\begin{aligned}
\sum_{s=1}^{S} w_s \int p_s^*(y, x; \theta, q)\, \mathrm{d}y\, \mathrm{d}x &= \sum_{s=1}^{S} w_s \int \frac{f(y|x; \theta) 1_{(y,s)\in\mathcal{S}_s} \hat{g}_{\theta,q}(x)}{q_s}\, \mathrm{d}y\, \mathrm{d}x \\
&= \sum_{s=1}^{S} w_s \int \frac{Q_{s|X}(x; \theta) \hat{g}_{\theta,q}(x)}{q_s}\, \mathrm{d}x \\
&= \int \sum_{s=1}^{S} w_s \frac{Q_{s|X}(x; \theta)}{q_s} \hat{g}_{\theta,q}(x)\, \mathrm{d}x \\
&= \int f_0^*(x)\, \mathrm{d}x \qquad \text{(by (3.11))} \\
&= 1.
\end{aligned}$$

Second, we will show that for all $\theta \in \Theta_0$,

$$\left.\frac{\partial}{\partial q}\right|_{q=\hat{Q}(\theta)} \sum_{s=1}^{S} w_s E_{s,0}\{\log p_s(y, x; \theta, q)\} = 0. \tag{3.12}$$

For $j = 1, \ldots, S-1$, the derivative is

$$\begin{aligned}
&\frac{\partial}{\partial q_j} \sum_{s=1}^{S} w_s E_{s,0}\{\log p_s(y, x; \theta, q)\} \\
&\quad = -\frac{\partial}{\partial q_j} \sum_{s=1}^{S} w_s E_{s,0} \left\{ \log \sum_{s'=1}^{S} w_{s'} \frac{Q_{s'|X}(x; \theta)}{q_{s'}} + \log q_s \right\}
\end{aligned}$$

$$
\begin{aligned}
&= \sum_{s=1}^{S} w_s E_{s,0}\left\{ \frac{w_j Q_{j|X}(x;\theta)/q_j^2}{\sum_{s'=1}^{S} w_{s'} Q_{s'|X}(x;\theta)/q_{s'}} \right\} - \frac{w_j}{q_j} \\
&= \sum_{s=1}^{S} w_s \int \frac{w_j Q_{j|X}(x;\theta)/q_j^2}{\sum_{s'=1}^{S} w_{s'} Q_{s'|X}(x;\theta)/q_{s'}} \frac{Q_{s|X}(x;\theta_0) g_0(x)}{Q_s(\theta_0, g_0)} \,\mathrm{d}x - \frac{w_j}{q_j} \\
&= \int \frac{w_j Q_{j|X}(x;\theta)/q_j^2 f_0^*(x)}{\sum_{s'=1}^{S} w_{s'} Q_{s'|X}(x;\theta)/q_{s'}} \,\mathrm{d}x - \frac{w_j}{q_j} \qquad \text{(by (2.5))} \\
&= \frac{w_j}{q_j^2}\left(\int Q_{j|X}(x;\theta) \hat{g}_{\theta,q}(x) \,\mathrm{d}x - q_j \right).
\end{aligned}
$$

Therefore, at $q = (q_1, \ldots, q_{S-1}, 1) = (\frac{\hat{Q}_1(\theta)}{\hat{Q}_S(\theta)}, \ldots, \frac{\hat{Q}_{S-1}(\theta)}{\hat{Q}_S(\theta)}, 1)$, we have (3.12).

By Theorem 3.1, the efficient score function and the efficient information matrix in the example are calculated by (3.4) and (3.5), respectively, where the score functions are given by

$$
\dot{\ell}_1(s, y, x; \theta, q) = \frac{(\partial/\partial\theta) f(y|x;\theta)}{f(y|x;\theta)} - \frac{\sum_{s'=1}^{S} w_{s'} (\partial/\partial\theta) Q_{s'|X}(x;\theta)/q_{s'}}{\sum_{s'=1}^{S} w_{s'} Q_{s'|X}(x;\theta)/q_{s'}}
$$

and $\dot{\ell}_2(s, y, x; \theta, q) = \{\dot{\ell}_{21}(s, y, x; \theta, q), \ldots, \dot{\ell}_{2(S-1)}(s, y, x; \theta, q)\}$, where

$$
\dot{\ell}_{2j}(s, y, x; \theta, q) = \frac{w_j}{q_j^2}\left\{ \frac{Q_{j|X}(x;\theta)}{\sum_{s'=1}^{S} w_{s'} Q_{s'|X}(x;\theta)/q_{s'}} - q_j \right\} \qquad (j = 1, \ldots, S-1).
$$

Here verification of the non-singularity of $\sum_{s=1}^{S} w_s E_{s,0}(\dot{\ell}_2^c \dot{\ell}_2^{cT})$ is omitted.

4. Numerical example: Stratified sampling with logistic regression

Here we compare the maximum likelihood estimator (MLE) and estimators based on reparametrizations of the least favorable submodel, and demonstrate that the estimators based on reparametrizations are statistically as efficient as the MLE and computationally more efficient.

The data in the Table 1 were taken from Scott and Wild [12,13] and were the case-control sampling part of the study of people under 35 in Northern Malawi. Cases are those with new cases of leprosy and controls are those without leprosy. The variable "Scar" indicates the presence or absence of a BCG vaccination scar (1 = present, 0 = absent).

Table 1. Leprosy data

	Scar = 0		Scar = 1		Total	
Age	Case	Control	Case	Control	Case	Control
2.5	1	24	1	31	2	55
7.5	11	22	14	39	25	61
12.5	28	23	22	27	50	50
17.5	16	5	28	22	44	27
22.5	20	9	19	12	39	21
27.5	36	17	11	5	47	22
32.5	47	21	6	3	53	24
Total					260	260

Let $x = (x_1, x_2)$ with $x_1 = \text{Scar}$ and $x_2 = 100(\text{Age} + 7.5)^{-2}$. We consider a stratified sampling (case-control sampling) with the logistic regression model

$$f(y|x;\alpha,\beta) = \frac{\exp\{y(\alpha + x^T\beta)\}}{1+\exp(\alpha + x^T\beta)} \qquad (y \in \{0,1\}, x \in R^2) \tag{4.1}$$

and the partition $\mathcal{Y} \times \mathcal{X} = (\{0\} \times \mathcal{X}) \cup (\{1\} \times \mathcal{X})$, where $\alpha \in R$ and $\beta \in R^2$. In this case, with $s = 0, 1$,

$$Q_s(\alpha,\beta,g) = \int f(y = s|x;\alpha,\beta)g(x)\,\mathrm{d}x$$

and

$$Q_{s|X}(x,\alpha,\beta) = f(y = s|x;\alpha,\beta).$$

From (3.10) and (3.11), a reparametrized model for the multisample model is

$$\begin{aligned} p_s^*(x;\alpha,\beta,\rho_1) &= \frac{(q_0/q_s)f(y=s|x;\theta)}{\sum_{s'=0}^1 w_{s'}(q_0/q_{s'})Q_{s'|X}(x;\alpha,\beta)} f_0^*(x) \\ &= \frac{\exp\{s(\alpha + \log\rho_1 + x^T\beta)\}}{w_0 + w_1\exp\{(\alpha + \log\rho_1 + x^T\beta)\}} f_0^*(x), \end{aligned}$$

where $\rho_0 = q_0/q_0 = 1$ and $\rho_1 = q_0/q_1$. The parameters in the model are not identifiable and the parameters α and ρ_1 cannot be estimated separately. By the proof in the stratified sampling example in Section 3.1, the efficient information bound for (α,β) is given by (3.5) in Theorem 3.1 with $\dot{\ell}_1(s,x;\alpha,\beta,\rho_1) = \{\partial/\partial(\alpha,\beta)\}\log p_s^*(x;\alpha,\beta,\rho_1)$ and $\dot{\ell}_2(s,x;\alpha,\beta,\rho_1) = \{\partial/\partial\rho_1\}\log p_s^*(x;\alpha,\beta,\rho_1)$. The estimator $(\hat{\alpha},\hat{\beta},\hat{\rho}_1)$ based on this nonidentifiable reparametrization is the maximizer of the log-likelihood $\ell_n(\alpha,\beta,\rho_1) = \sum_{s=0}^1\sum_{i=1}^{n_s}\log p_s^*(x_{si};\alpha,\beta,\rho_1)$.

To gain identifiability of the parameters, we let $\alpha^* = \alpha + \log \rho_1$, and the model is further reparametrized as

$$p_s^*(x; \alpha^*, \beta) = \frac{\exp\{s(\alpha^* + x^T\beta)\}}{w_0 + w_1 \exp\{(\alpha^* + x^T\beta)\}} f_0^*(x).$$

If we treat the parameters α and g in the original model as nuisance parameters, Theorem 3.1 gives the efficient information bound for an estimator of the parameter β: it is (3.5) in Theorem 3.1 with $\dot{\ell}_1(s, x; \alpha^*, \beta) = (\partial/\partial\beta) \log p_s^*(x; \alpha^*, \beta)$ and $\dot{\ell}_2(s, x; \alpha^*, \beta) = (\partial/\partial\alpha^*) \log p_s^*(x; \alpha^*, \beta)$. The proof is similar to the one for the stratified sampling example given above and, therefore, we omit it. The estimator $(\hat{\alpha}^*, \hat{\beta})$ based on this identifiable reparametrization is the maximizer of the log-likelihood for the data $\ell_n(\alpha^*, \beta) = \sum_{s=0}^{1} \sum_{i=1}^{n_s} \log p_s^*(x_{si}; \alpha^*, \beta)$.

If X takes values in $\{v_1, \ldots, v_K\}$, let $g_k = g(v_k)$, $k = 1, \ldots, K$. Then the log-likelihood for a single observation in the model can be written as

$$\log p_s(x; \alpha, \beta, g) = \log f(y = s|x; \alpha, \beta) + \sum_{k=1}^{K} 1_{\{x=v_k\}} \log g_k - \log \sum_{k=1}^{K} f(y = s|v_k; \alpha, \beta) g_k.$$

The MLE $(\hat{\alpha}, \hat{\beta}, \hat{g})$, where $\hat{g} = (\hat{g}_1, \ldots, \hat{g}_K)$, is the maximizer of the log-likelihood $\ell_n(\alpha, \beta, g) = \sum_{s=0}^{1} \sum_{i=1}^{n_s} \log p_s(x_{si}; \alpha, \beta, g)$.

For each case (non-identifiable reparametrization, identifiable reparametrization and maximum likelihood), let θ_1 be the parameter of interest and θ_2 be the nuisance parameter. Then an estimated variance of the estimator (of the parameter of interest) is given by the formula (3.5) except that each $\sum_s w_s E_{s,0}(\dot{\ell}_i^c \dot{\ell}_j^{cT})$ $(i, j = 1, 2)$ is replaced with the corresponding second-degree partial derivative $-n^{-1}(\partial^2/\partial\theta_i\, \partial\theta_j^T)\ell_n$.

Estimates of regression coefficients and their standard error (SE) in these models are given in Table 2. Note that in the maximum likelihood and non-identifiable reparametrization, the intercept parameter is not identifiable. Its estimates and the corresponding SE are unreliable and unstable. Therefore, we do not look at estimates of the intercept parameter in these models. The

Table 2. Model fitting results for the leprosy data

	Maximum likelihood		Reparametrization			
			Not identifiable		Identifiable	
	Coef	SE	Coef	SE	Coef	SE
Intercept	1.55720	94.52766	0.61334	8388784	–	–
Age	−0.30205	0.19737	−0.30211	0.19737	−0.30215	0.19736
Scar	−4.30992	0.57891	−4.31017	0.57892	−4.30988	0.57889
Computation time	43.61 sec		2.80 sec		2.44 sec	

Table 3. Relative efficiency with respect to the maximum likelihood

	Reparametrization	
	Not identifiable	Identifiable
Age	0.99997	0.99992
Scar	1.00005	0.99994
Computation time	0.06421	0.05595

estimated coefficients of "Age" and "Scar" and their SE are very similar to each other among these models. This is consistent with the prediction made by Theorem 3.1 that reparametrization gives the semi-parametric efficiency bound that is achieved by the MLE.

Table 3 gives the relative efficiency of estimates in non-identifiable reparametrization and identifiable reparametrization with respect to the maximum likelihood, along with the relative efficiency in computation times (which is defined as the ratio of the corresponding computation times). The table indicates that these reparametrizations are statistically as efficient as, and computationally more efficient than, the method of maximum likelihood.

5. Discussion

Theorem 3.1 gives conditions under which the efficient score function and the efficient information matrix can be expressed in terms of the parameters in the reparametrized model, namely (3.4) and (3.5), respectively. In Section 4, we demonstrated that Theorem 3.1 can be used to show the efficiency of estimators based on non-identifiable and identifiable reparametrizations in the logistic regression model, and that these estimators are computationally more efficient than the MLE. The results of the paper can be used to find a reparametrization of the least favorable submodel (or profile likelihood) that gives statistically and computationally efficient estimators in multisample models.

Appendix A

We define the Hilbert space, projection and the efficient score function.

A.1. Hilbert space and the projection

Let $\mathcal{H}$ be the Hilbert space of m-dimensional measurable functions with zero mean and finite variance:

$$\mathcal{H} = \left\{ \psi(s,x):\ E_{s,0}(\psi) = 0\ (s = 1,\dots,S), \sum_{s=1}^{S} w_s E_{s,0}(\psi^T \psi) < \infty \right\}.$$

The covariance of $\psi, \phi \in \mathcal{H}$ is defined by $\mathrm{cov}(\psi, \phi) = \sum_{s=1}^{S} w_s E_{s,0}(\psi\phi^T)$. We say ψ and ϕ are uncorrelated if $\mathrm{cov}(\psi, \phi) = 0$. For a set of functions $\mathcal{G}$ in $\mathcal{H}$, $\mathcal{G}^\perp$ is the set of all functions $\psi \in \mathcal{H}$ with $\mathrm{cov}(\psi, \phi) = 0$ for all $\phi \in \mathcal{G}$. The projection $\Pi(\psi|\mathcal{G})$ of $\psi \in \mathcal{H}$ onto a closed subspace $\mathcal{G}$ is characterized by

$$\Pi(\psi|\mathcal{G}) \in \mathcal{G} \quad \text{and} \quad \psi - \Pi(\psi|\mathcal{G}) \in \mathcal{G}^\perp.$$

For an arbitrary Banach space $\mathcal{B}$, let $\mathcal{B}^*$ be its dual. Let $A : \mathcal{B} \to \mathcal{H}$ be a bounded linear operator and $\psi \in \mathcal{H}$. The *adjoint operator* $A^T : \mathcal{H} \to \mathcal{B}^*$ of $A : \mathcal{B} \to \mathcal{H}$ is defined by the map

$$(A^T\psi)(b) = \langle Ab, \psi \rangle = \sum_{s=1}^{S} w_s E_{s,0}\{(Ab)\psi^T\}, \qquad b \in \mathcal{B}.$$

Suppose that $(A^T A)^{-1}$ exists and let $\psi \in \mathcal{H}$. By the projection theorem for an operator equation,

$$\Pi(\psi|\overline{A(\mathcal{B})}) = A(A^T A)^{-1} A^T \psi$$

is a projection of ψ onto the closure $\overline{A(\mathcal{B})}$ of the range of A.

A.2. The projection theorem

Theorem A.1 (The projection theorem). *Suppose $\phi(s, x)$ is an l-dimensional vector of measurable functions such that*

(1) *for $s = 1, \ldots, S$, $E_{s,0}(\phi) = 0$;*
(2) *$\sum_{s=1}^{S} w_s E_{s,0}(\phi^T\phi) < \infty$;*
(3) *$\{\sum_{s=1}^{S} w_s E_{s,0}(\phi\phi^T)\}^{-1}$ exists.*

Let $\mathcal{G} = \{A\phi\colon\ A \in R^{m\times l}\}$ be the closed subspace of $\mathcal{H}$ generated by ϕ. Then, for each $\psi \in \mathcal{H}$, the projection of ψ onto the closed subspace $\mathcal{G}$ is given by

$$\pi(\psi|\mathcal{G}) = \left\{\sum_{s=1}^{S} w_s E_{s,0}(\psi\phi^T)\right\}\left\{\sum_{s=1}^{S} w_s E_{s,0}(\phi\phi^T)\right\}^{-1}\phi.$$

Proof. The proof is similar to the one for the standard case. □

A.3. The efficient score function

Here, we give the definition of the efficient score function in a multisample model.

We assume the log-likelihood function for a single observation $\ell(s, x; \beta, \eta)$ (defined by (2.1)) is continuously differentiable with respect to β for all $\beta \in \Theta_\beta$ and Hadamard differentiable with respect to η for all $\eta \in \Theta_\eta$. The *score function* $\dot{\ell}(s, x; \beta, \eta)$ for β and the *score operator* $A(s, x; \beta, \eta)$ for η in the multisample model are the derivatives of the log-likelihood function with respect to β and η, respectively.

The tangent space for η is the closure $\overline{A(\mathcal{B})}$ of range of the score operator A for η.
The uncorrelated complement of the score function $\dot{\ell}_\beta$ with respect to the tangent space for η,

$$\dot{\ell}^* = \dot{\ell} - \Pi(\dot{\ell}|\overline{A(\mathcal{B})}),$$

is called the *efficient score function* in the multisample model $(\mathcal{P}_1, \ldots, \mathcal{P}_S)$.

A.4. Theorem to identify the efficient score function

To verify that the function given by (2.3) is the efficient score function, the following theorem may be useful.

Theorem A.2. *A path $t \to \eta_t$ is a continuously differentiable map in a neighborhood of* 0 *such that $\eta_{t=0} = \eta_0$. Define $\alpha_t = \eta_t - \eta_0$. If $\beta \to \hat{\eta}_\beta$ is a differentiable function such that*

$$\hat{\eta}_{\beta_0} = \eta_0 \tag{A.1}$$

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

$$\left.\frac{\partial}{\partial t}\right|_{t=0} \sum_{s=1}^{S} w_s E_{s,0}\{\log p_s(x; \beta, \hat{\eta}_\beta + \alpha_t)\} = 0, \tag{A.2}$$

then the function

$$\dot{\ell}^*(s, x) = \left.\frac{\partial}{\partial \beta}\right|_{\beta=\beta_0} \log p_s(x; \beta, \hat{\eta}_\beta) \tag{A.3}$$

is the efficient score function.

Proof. Condition (A.2) implies that

$$\begin{aligned} 0 &= \left.\frac{\partial}{\partial \beta}\right|_{\beta=\beta_0} \left.\frac{\partial}{\partial t}\right|_{t=0} \sum_{s=1}^{S} w_s E_{s,0}\{\log p_s(x; \beta, \hat{\eta}_\beta + \alpha_t)\} \\ &= \left.\frac{\partial}{\partial t}\right|_{t=0} \sum_{s=1}^{S} w_s E_{s,0}\left\{ \left.\frac{\partial}{\partial \beta}\right|_{\beta=\beta_0} \log p_s(x; \beta, \hat{\eta}_\beta + \alpha_t) \right\}. \end{aligned} \tag{A.4}$$

By differentiating the identity

$$\sum_{s=1}^{S} w_s \int \left\{ \frac{\partial}{\partial \beta} \log p_s(x; \beta, \hat{\eta}_\beta + \alpha_t) \right\} p_s(x; \beta, \hat{\eta}_\beta + \alpha_t)\, \mathrm{d}x = 0$$

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

$$
\begin{aligned}
0 &= \frac{\partial}{\partial t}\Big|_{t=0,\beta=\beta_0} \sum_{s=1}^{S} w_s \int \left(\frac{\partial}{\partial \beta}\log p_s(x;\beta,\hat{\eta}_\beta+\alpha_t)\right) p(x;\beta,\hat{\eta}_\beta+\alpha_t)\,\mathrm{d}x \\
&= \sum_{s=1}^{S} w_s E_{s,0}\left[\dot{\ell}^*(s,x)\left\{\frac{\partial}{\partial t}\Big|_{t=0}\log p_s(x;\beta_0,\eta_t)\right\}\right] \qquad \text{(we used (A.3) and} \\
&\qquad\qquad \hat{\eta}_{\beta_0}+\alpha_t=\eta_t \text{ by (A.1))} \qquad \text{(A.5)} \\
&\quad + \frac{\partial}{\partial t}\Big|_{t=0} \sum_{s=1}^{S} w_s E_{s,0}\left\{\frac{\partial}{\partial \beta}\Big|_{\beta=\beta_0}\log p_s(x;\beta,\hat{\eta}_\beta+\alpha_t)\right\} \\
&= \sum_{s=1}^{S} w_s E_{s,0}\left[\dot{\ell}^*(s,x)\left\{\frac{\partial}{\partial t}\Big|_{t=0}\log p_s(x;\beta_0,\eta_t)\right\}\right] \qquad \text{(by (A.4))}.
\end{aligned}
$$

Let $c\in R^m$ be arbitrary. Then, it follows from (A.5) that the product $c'\dot{\ell}^*(s,x)$ is orthogonal to the nuisance tangent space $\dot{\mathcal{P}}_\eta$, which is the closed linear span of score functions of the form $\phi(s,x)=\frac{\partial}{\partial t}\big|_{t=0}\log p_s(x;\beta_0,\eta_t)$. By (A.3) with (A.1) , we have

$$
\begin{aligned}
\dot{\ell}^*(s,x) &= \frac{\partial}{\partial \beta}\Big|_{\beta=\beta_0}\log p_s(x;\beta,\eta_0)+\frac{\partial}{\partial \beta}\Big|_{\beta=\beta_0}\log p_s(x;\beta_0,\hat{\eta}_\beta) \\
&= \dot{\ell}_\beta(s,x)-\psi(s,x),
\end{aligned}
$$

where $\dot{\ell}_\beta(s,x)=\frac{\partial}{\partial\beta}\big|_{\beta=\beta_0}\log p_s(x;\beta,\eta_0)$ and $\psi(s,x)=-\frac{\partial}{\partial\beta}\big|_{\beta=\beta_0}\log p_s(x;\beta_0,\hat{\eta}_\beta)$. Finally, $c'\dot{\ell}^*(s,x)=c'\dot{\ell}_\beta(s,x)-c'\psi(s,x)$ is orthogonal to the nuisance tangent space $\dot{\mathcal{P}}_\eta$ and $c'\psi(s,x)\in\dot{\mathcal{P}}_\eta$ implies that $c'\psi(s,x)$ is the orthogonal projection of $c'\dot{\ell}_\beta(s,x)$ onto the nuisance tangent space $\dot{\mathcal{P}}_\eta$. Since $c\in R^m$ is arbitrary, the function $\dot{\ell}^*(s,x)$ given by (A.3) is the efficient score function. □

Appendix B

B.1. Proof of Lemma 2.1

Proof. We show that $\sum_{s=1}^{S} w_s\int \log p_s(y,x;\theta,\hat{g}_\theta)\,\mathrm{d}F_{s0}$ satisfies conditions (A.1) and (A.2) in Theorem A.2 in Appendix A so that the claim follows from this theorem.

Condition (A.1) is verified in Remark 2.2. Now we verify (A.2). 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 $\alpha_0'(x)=$

$(\mathrm{d}/\mathrm{d}t)|_{t=0}\alpha_t(x)$. Then

$$\begin{aligned}
&\frac{\partial}{\partial t}\bigg|_{t=0}\sum_{s=1}^{S} w_s \int \log p_s(y,x;\theta,\hat{g}_\theta+\alpha_t)\,\mathrm{d}F_{s0} \\
&\quad=\frac{\partial}{\partial t}\bigg|_{t=0}\sum_{s=1}^{S} w_s\left[\int \log\{\hat{g}_\theta(x)+\alpha_t(x)\}\,\mathrm{d}F_{s,0}-\log Q_s(\theta,\hat{g}_\theta+\alpha_t)\right] \\
&\quad=\frac{\partial}{\partial t}\bigg|_{t=0}\left[\int \log\{\hat{g}_\theta(x)+\alpha_t(x)\}f_0^*(x)\,\mathrm{d}x-\sum_{s=1}^{S} w_s\log Q_s(\theta,\hat{g}_\theta+\alpha_t)\right] \\
&\quad=\int\frac{\alpha_0'(x)}{\hat{g}_\theta(x)}f_0^*(x)\,\mathrm{d}x-\sum_{s=1}^{S} w_s\frac{\int Q_{s|X}(x;\theta)\alpha_0'(x)\,\mathrm{d}x}{\hat{Q}_s(\theta)}=0
\end{aligned}$$

by (2.4) and (2.5). □

References

[1] Bickel, P.J., Klaassen, C.A.J., Ritov, Y. and Wellner, J.A. (1993). *Efficient and Adaptive Estimation for Semiparametric Models*. Baltimore, MD: Johns Hopkins Univ. Press. MR1245941

[2] Breslow, N., 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. MR2001644

[3] 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. MR1762555

[4] Gilbert, P.B. (2000). Large sample theory of maximum likelihood estimates in semiparametric biased sampling models. *Ann. Statist.* **28** 151–194. MR1762907

[5] Gilbert, P.B., 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. MR1688069

[6] Gill, R.D., Vardi, Y. and Wellner, J.A. (1988). Large sample theory of empirical distributions in biased sampling models. *Ann. Statist.* **16** 1069–1112. MR0959189

[7] Lawless, J.F., Kalbfleisch, J.D. and Wild, C.J. (1999). Semiparametric methods for response-selective and missing data problems in regression. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **61** 413–438. MR1680310

[8] 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.* **62** 1023–1052.

[9] Newey, W.K. (1994). The asymptotic variance of semiparametric estimators. *Econometrica* **62** 1349–1382. MR1303237

[10] Robins, J.M., Hsieh, F.S. and Newey, W. (1995). Semiparametric efficient estimation of a conditional density with missing or mismeasured covariates. *J. Roy. Statist. Soc. Ser. B* **57** 409–424. MR1323347

[11] 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. MR1294730

[12] Scott, A.J. and Wild, C.J. (1997). Fitting regression models to case-control data by maximum likelihood. *Biometrika* **84** 57–71. MR1450191

[13] Scott, A.J. and Wild, C.J. (2001). Maximum likelihood for generalised case-control studies. *J. Statist. Plann. Inference* **96** 3–27. MR1843447

[14] Vardi, Y. (1985). Empirical distributions in selection bias models (with discussion). *Ann. Statist.* **13** 178–205. MR0773161

Received April 2010 and revised October 2010