The Annals of Statistics
2000, Vol. 28, No. 1, 151–194

LARGE SAMPLE THEORY OF MAXIMUM LIKELIHOOD ESTIMATES IN SEMIPARAMETRIC BIASED SAMPLING MODELS[1]

BY PETER B. GILBERT

Harvard School of Public Health

Vardi [*Ann. Statist.* **13** 178–203 (1985)] introduced an s-sample biased sampling model with known selection weight functions, gave a condition under which the common underlying probability distribution G is uniquely estimable and developed simple procedure for computing the nonparametric maximum likelihood estimator (NPMLE) $\mathbb{G}_n$ of G. Gill, Vardi and Wellner thoroughly described the large sample properties of Vardi's NPMLE, giving results on uniform consistency, convergence of $\sqrt{n}\ (\mathbb{G}_n - G)$ to a Gaussian process and asymptotic efficiency of $\mathbb{G}_n$. Gilbert, Lele and Vardi considered the class of semiparametric s-sample biased sampling models formed by allowing the weight functions to depend on an unknown finite-dimensional parameter θ. They extended Vardi's estimation approach by developing a simple two-step estimation procedure in which $\hat{\theta}_n$ is obtained by maximizing a profile partial likelihood and $\mathbb{G}_n \equiv \mathbb{G}_n(\hat{\theta}_n)$ is obtained by evaluating Vardi's NPMLE at $\hat{\theta}_n$. Here we examine the large sample behavior of the resulting joint MLE $(\hat{\theta}_n, \mathbb{G}_n)$, characterizing conditions on the selection weight functions and data in order that $(\hat{\theta}_n, \mathbb{G}_n)$ is uniformly consistent, asymptotically Gaussian and efficient.

Examples illustrated here include clinical trials (especially HIV vaccine efficacy trials), choice-based sampling in econometrics and case-control studies in biostatistics.

1. Introduction: semiparametric biased sampling models and the MLE. Vardi (1985) developed methodology for the s-sample biased sampling model with known selection bias weight functions. Gilbert, Lele and Vardi (1999) considered an extension of this model which allows the weight functions to depend on an unknown finite-dimensional parameter θ. This model comprises three components: a probability measure G defined on a sample space $\mathbf{Y}$ with σ-field of subsets $\mathbf{B}$, a set of nonnegative (measurable) "stratum" weight functions $w_1, \ldots, w_s$ defined on $\mathbf{Y} \times \Theta$ and selection probabilities λ_i, $i = 1, \ldots, s$, with $\sum_{i=1}^{s} \lambda_i = 1$. The parameter θ is common to all s weight functions $w_i(\cdot, \theta)$, which are assumed to be nonnegative and of a known parametric form. Let $g \equiv dG/d\mu$ for some measure μ dominating G. The data are assumed to be an i.i.d. sample $X_k = (I_k, Y_k)$, $k = 1, \ldots, n$, from the semiparametric biased sampling model $\mathbf{P} \equiv \{P_{(\theta, G)}\colon \theta \in \Theta, G \in \mathbf{G}\}$, defined on

Received January 1998; revised June 1999.

[1]Research supported by NIH Grants 5-U01-AI38855, 5-U01-AI28076 and AI24643.

AMS 1991 *subject classifications*. Primary 60G05, 62F05; secondary 62G20, 62G30.

Key words and phrases. Asymptotic theory, choice-based sampling, clinical trials, empirical processes, generalized logistic regression, HIV vaccine trial, nonparametric maximum likelihood, selection bias models, Vardi's estimator.

$\mathbf{X} \equiv \{1, \ldots, s\} \times \mathbf{Y}$, with density given by

$$p(x, \theta, G) = p(i, y, \theta, G) = \lambda_i \frac{w_i(y, \theta)}{W_i(\theta, G)} g(y). \tag{1.1}$$

Here $W_i(\theta, G)$ is the ith normalizing function given by

$$W_i(\theta, G) \equiv \int_{\mathbf{Y}} w_i(u, \theta)\, dG(u), \tag{1.2}$$

assumed to be positive and finite for all θ in the parameter space Θ. The random variable $I \in \{1, \ldots, s\}$ denotes the stratum, selected with probability λ_i, and, conditional on $I = i$, the probability measure F_i of Y under the biased sampling model satisfies

$$F_i(A, \theta, G) \equiv W_i^{-1}(\theta, G) \int_A w_i(u, \theta)\, dG(u), \quad A \in \mathbf{B}, \quad i = 1, \ldots, s. \tag{1.3}$$

In Vardi's (1985) original treatment, the weight functions were assumed completely known (independent of θ), but there are many practical situations in which a complete specification of the w_i's is too restrictive, but the weight functions can realistically be assumed to belong to a parametric family. We illustrate with three examples.

EXAMPLE 1.1 (Univariate generalized logistic regression model). As described in the introduction of Gilbert, Lele and Vardi (1999), the generalized logistic regression (GLR) model is useful for assessing from viral data taken from breakthrough infections in preventive human immunodeficiency virus (HIV) vaccine efficacy trials how vaccine protection against infection varies by characteristics of challenge HIV. Suppose viruses are ordered by some "distance," quantifiable in a variety of ways, for instance by the percent nucleotide or amino acid mismatch in a particular gene. Let $Y \in [0, \infty)$ be a random variable denoting the distance between a virus that has been isolated from an infected trial participant and the virus strain represented in the vaccine. The GLR model is an s-sample semiparametric biased sampling model, where s is the number of intervention arms, G is the distribution of Y in the baseline placebo group (group s, say) and θ describes how strongly vaccine protection depends on strain distance. For instance, in the two-sample GLR model, the groups are vaccine and placebo. Letting F_v denote the distribution of Y in the vaccine group, the GLR model relates G and F_v in the following way:

$$F_v(y) = \frac{\int_0^y \exp\{h(u, \theta)\}\, dG(u)}{\int_0^\infty \exp\{h(u, \theta)\}\, dG(u)}, \qquad y \in [0, \infty), \tag{1.4}$$

where $h(y, \theta)$ is a given function (most simply taken to be linear, $h(y, \theta) = y\theta$) and θ is an unknown d-dimensional "differential vaccine efficacy" parameter. This is a biased sampling model (1.1)–(1.3), with $\mathbf{Y} = [0, \infty)$, $F_1 = F_v$, $F_2 = G$, $w_1(y, \theta) = \exp\{h(y, \theta)\}$, $w_2(y, \theta) = 1$ and $\lambda_i \equiv n_i/n$, $i = 1, 2$, the sampling fraction of infections in the trial from group i. After reparametrization, the

ordinary logistic regression model with case-control sampling of covariates y can be written in the form of a two-sample GLR model with $h(y, \theta) = \theta_1 + \theta_2^T y$, where G is the distribution of y for controls.

The general s-sample GLR model is formed by supposing that observations from the ith sample are distributed i.i.d. F_i with

$$F_i(y) = \frac{\int_0^y \exp\{h_i(u, \theta)\}\, dG(u)}{\int_0^\infty \exp\{h_i(u, \theta)\}\, dG(u)}, \qquad y \in [0, \infty), \tag{1.5}$$

where G is the (baseline) distribution for the sth sample and $h_i(y, \theta)$, $i = 1, \ldots, s$, are given functions. For identifiability and unique estimability, we take $h_s \equiv 0$, so that $F_s = G$. A practical choice for the functions $\{h_i\}$ is $h_i(y, \theta) = \sum_{k=1}^d h_{ik}(y)\theta_k$, where the h_{ik} are given functions of y independent of θ.

The GLR model can be used to assess how protection of a vaccine against any heterogeneous pathogen (not just HIV) depends on features of the pathogen, and more generally for investigating how the relative efficacy of two treatments varies by some random variable measured only on disease cases. Among many applications are the study of the relationship between a measure of genetic evolution of HIV in an infected person and the relative efficacy of two antiretroviral regimens, and the study of how the efficacy of a new treatment relative to a control treatment changes over time. The broad utility of the GLR model, especially for analyzing clinical trial data, is discussed more extensively in Gilbert, Self and Ashby (1998) and Gilbert, Lele and Vardi (1999).

EXAMPLE 1.2 (Multivariate generalized logistic regression model). Since HIV vaccine protection may vary with variations in several viral features, it is of practical interest to assess differential vaccine efficacy by modeling a multidimensional distance $Y \equiv (Y_1, \ldots, Y_k)^T$ according to a GLR model. This is easily done in the same manner as the univariate model in Example 1.1. For instance, a simple two-sample GLR model with a bivariate distance is specified by

$$F_v(y_1, y_2) = \frac{\int_0^{y_1}\int_0^{y_2} \exp\{u_1\theta_1 + u_2\theta_2 + u_1u_2\theta_3\}\, dG(u_1, u_2)}{\int_0^{\infty}\int_0^{\infty} \exp\{u_1\theta_1 + u_2\theta_2 + u_1u_2\theta_3\}\, dG(u_1, u_2)} \tag{1.6}$$

for $y_1, y_2 \in [0, \infty)$. This is a two-sample biased sampling model with $\mathbf{Y} = [0, \infty) \times [0, \infty)$, $d = 3$, $w_1(y, \theta) = \exp\{y_1\theta_1 + y_2\theta_2 + y_1y_2\theta_3\}$ and $w_2(y, \theta) \equiv 1$. The interaction parameter θ_3 will be zero if and only if the sum of the marginal vaccine effects against a strain with $Y_1 = y_1$ and a strain with $Y_2 = y_2$ equals the joint vaccine effect against a strain with $Y_1 = y_1$ and $Y_2 = y_2$. Setting $y_1 = 0$ ($y_2 = 0$) recovers the marginal univariate GLR model for Y_2 (Y_1).

The full k-variate GLR model applies to random variables defined on $\mathbf{Y} = [0, \infty)^k$, and can be laid out similarly to (1.6), with $\binom{k}{2}$ two-way interaction

parameters, $\binom{k}{3}$ three-way interaction parameters and so on. Parsimonious versions can be considered by setting various interaction parameters to zero.

EXAMPLE 1.3 (Choice-based sampling in econometrics; case-control studies in biostatistics). This example is adapted from Example 4.4 of Gill, Vardi and Wellner (1988). Suppose that $X = (J, Y)$, where J is discrete with values on $\{1, \ldots, M\}$ and $Y \sim G$ with density g with respect to μ is a covariate vector with values in $\mathbf{Y} \subset$ some R^d. The prospective, unbiased model H has density $h(j, y) = p_\theta(j|y)g(y)$, so that

$$H(\{j\} \times A) = \int_A p_\theta(j|y)\, dG(y)$$

for $j = 1, \ldots, M$ and $A \in \mathbf{B}(R^d)$, where $p_\theta(j|y) = P_\theta(J = j|Y = y)$ is a parametric model. The multinomial logistic regression model is a common choice of p_θ, given by

$$p_\theta(j|y) = \frac{\exp(\alpha_j + \beta_j^T y)}{\sum_{j'=1}^M \exp(\alpha_{j'} + \beta_{j'}^T y)}, \tag{1.7}$$

with $\theta = (\alpha, \beta) \in R^{(d+1)M}$, $\alpha_M = \beta_M = 0$.

The retrospective, biased sampling model F is obtained from H by the biasing functions $w_i(y) = 1_{D_i}(y)$, where $D_i \subset \{1, \ldots, M\}$ for $i = 1, \ldots, s$. For instance, consider the "pure choice-based sampling model" as described in Cosslett (1981), where the strata D_i are defined by $D_i \equiv \{i\}$, $i = 1, \ldots, s \equiv M$. As in Gill, Vardi and Wellner [(1988), page 1094], Vardi's (1985) necessary and sufficient condition for identifiability of G fails. Manski and Lerman (1977) avoid this problem by assuming that the "aggregate shares"

$$H(j, \theta, G) \equiv H(\{j\} \times \mathbf{Y}, \theta, G) = \int_{\mathbf{Y}} p_\theta(j|y)\, dG(y), \quad j = 1, \ldots, M, \tag{1.8}$$

are known. For this biasing system, one can view F as a biased distribution derived from G with new weight functions $w_j^\star(y, \theta) = p_\theta(j|y)$, $j = 1, \ldots, M$. Then G in model (1.1)–(1.3) is usually identifiable, as all that is required is $\int 1_{[p_\theta(j|y)>0]} p_\theta(j|y)\, dG(y) > 0$, $j = 1, \ldots, M$ (see Proposition 1.1 of Gill, Vardi and Wellner (1988)). If θ is known, the methods of Vardi (1985) and Gill, Vardi and Wellner (1988) apply. However, in many practical applications θ will be unknown. Since all M weight functions depend on θ, the resulting model is not a member of the class of biased sampling models primarily studied here, in which one weight function is independent of θ. However, the extra assumption that the normalizing constants $H(j, \theta, G)$ are known implies that the estimation procedure considered here still produces the maximum likelihood estimator (MLE) of (θ, G), with limiting behavior characterized by results given here.

The pure choice-based sampling design is also frequently used in case-control studies in biostatistics, where the j's often represent different disease categories. In the biostatistics application, the aim is to estimate the odds ratios $\{\beta_j\}$ of (1.7). Given known $H(j, \theta, G)$ as in (1.8), the asymptotic behavior

of these odds ratios and G are described by results developed here. The results also apply to choice-based models with general D_i's for which the biased distribution F has density

$$f(i, j, y) = \lambda_i \frac{1_{D_i}(j) p_\theta(j|y) g(y)}{\int \sum_{j'=1}^{M} 1_{D_i}(j') p_\theta(j'|y') g(y')\, d\mu(y')} \equiv \lambda_i \frac{w_i^{\star\star}(j, y, \theta) g(y)}{\int \sum_{j'=1}^{M} w_i^{\star\star}(j', y', \theta)\, dG(y')}. \tag{1.9}$$

Other examples and applications of semiparametric biased sampling models are described in Gilbert (1996) and Gilbert, Lele and Vardi (1999). Before presenting methodology for estimation of θ and G in these models, we note that Qin (1998) considered the closely related semiparametric two-sample density ratio model. Our approach resembles that taken by Qin (1998) in its utilization of a profile likelihood, but substantially differs in that it explicitly extends the original approach of Vardi (1985) and Gill, Vardi and Wellner (1988), thereby including estimation of the normalizing constants as a central component. In addition, the treatment here is more general in that s samples rather than two are handled, and arbitrary selection weight functions are allowed.

The first matter at hand is identifiability. As Gill, Vardi and Wellner (1988) is repeatedly referenced, henceforth it is abbreviated GVW.

1.1. *Identifiability.* In the case of known θ, GVW characterized necessary and sufficient conditions for G to be identifiable, namely, that the sample space $\mathbf{Y}$ equals $[y\colon w_i(y, \theta) > 0$ for some $i = 1, \ldots, s]$ and a certain graph is connected. For fixed $\theta \in \Theta$, define a graph $\mathbf{G}^\star(\theta)$ on the s vertices $i = 1, \ldots, s$ by identifying vertex i with k if and only if

$$\int 1_{[w_i(y,\theta)>0]} 1_{[w_k(y,\theta)>0]}\, dG(y) > 0.$$

The graph $\mathbf{G}^\star(\theta)$ is *connected* if every pair of vertices is connected by a path. In what follows, assume these conditions hold for all θ in some neighborhood $\Theta_0 \subset \Theta$ of the true θ_0. Notice that these two conditions hold automatically if all of the weight functions are strictly positive.

Now consider identifiability of unknown θ and G. When there is only one sample, the model is rarely identifiable. To our knowledge, the only class of one-sample models (with no restrictions on G) known to be identifiable are those with weight function whose domain depends on θ [see Gilbert, Lele and Vardi (1999), Theorem 1]. For the s-sample model, with $s \geq 2$, a large class of models is identifiable. When one of the weight functions is independent of θ, the class of identifiable models can be characterized by a simple condition on the weight functions. The following result is proved in Gilbert, Lele and Vardi (1999).

THEOREM 1.1. *Let $s \geq 2$, with w_s independent of θ. Suppose the sample space $\mathbf{Y}$ equals $\bigcup_{i=1}^{s}[y\colon w_i(y,\theta) > 0$ for all $\theta \in \Theta]$ and $\mathbf{G}^{\star}(\theta)$ is connected for all θ in some neighborhood $\Theta_0 \subset \Theta$ of the true θ_0. Then the s-sample biased sampling model is identifiable if and only if the following condition holds: for all $\tilde{\theta}, \theta \in \Theta_0$ with $\tilde{\theta} \neq \theta$, there is at least one weight function w_i, $i \in \{1, \ldots, s-1\}$, such that $w_i(y, \tilde{\theta})$ and $w_i(y, \theta)$ are linearly independent as functions of y.*

Gilbert, Lele and Vardi (1999) describe a large class of models that satisfy the conditions of Theorem 1.1. For instance, all interesting GLR models are identifiable by this theorem.

Henceforth, assume the identifiability conditions of Theorem 1.1, so that we consider identifiable biased sampling models (1.1)–(1.3) with multiple samples, in which at least one of the weight functions is independent of θ. In fact, throughout we require that the sth weight function is constant, which is used to ensure unique estimability by the employed estimation procedure.

1.2. *Maximum likelihood estimation.* The MLE that we study is computed by a procedure introduced by Gilbert, Lele and Vardi (1999). We sketch it here. Following their notation, denote the size of the ith sample by n_i, the total sample size by $n = \sum_{i=1}^{s} n_i$, and the ith sampling fraction by $\lambda_{ni} = n_i/n$. Let $t_1, \ldots, t_h$ be the distinct observed Y values, with multiplicities $r_1, \ldots, r_h$. Let n_{ij}, $i = 1, \ldots, s$, $j = 1, \ldots, h$, be the number of observations from the ith group with value t_j. Notice that h is random. Then the likelihood of the data observed according to (1.1)–(1.3) is

$$L_n(\theta, G|\underline{x}) = \prod_{i=1}^{s}\prod_{j=1}^{h}\left[\frac{w_i(t_j,\theta)G\{t_j\}}{W_i(\theta,G)}\right]^{n_{ij}}, \tag{1.10}$$

where $\underline{x}$ represents dependency on the observed data $\{X_k = (I_k, Y_k), k = 1, \ldots, n\}$, equivalent to $\{t_j, n_{ij}\colon i = 1, \ldots, s, j = 1, \ldots, h\}$. Define $w_{ij}(\theta) = w_i(t_j, \theta)$. Consider a partial likelihood defined by

$$L_{n1}(\theta, \underline{V}|\underline{x}) = \prod_{i=1}^{s}\prod_{j=1}^{h}\left[\frac{w_{ij}(\theta)V_i^{-1}}{\sum_{k=1}^{s}\lambda_{nk}w_{kj}(\theta)V_k^{-1}}\right]^{n_{ij}}, \tag{1.11}$$

with $\underline{V} = (V_1, \ldots, V_s)^T$ given by $V_i = W_i(\theta, G)/W_s(\theta, G)$, $i = 1, \ldots, s$, where the dependency of V_i on θ and G is suppressed. Notice that L_{n1} depends on G only through the normalizing constants. When w_s is a constant function (or when w_s depends on θ but $W_s(\theta, G)$ is a known constant), the following maximum partial likelihood estimation procedure yields the estimate $(\hat{\theta}_n, \mathbb{G}_n)$ which maximizes the full likelihood:

1 Maximize L_{n1} over θ and $\underline{V}$, subject to $V_1 > 0, V_2 > 0, \ldots, V_{s-1} > 0, V_s = 1$ to obtain $(\hat{\theta}_n, \underline{\mathbb{V}}_n)$.
2 Compute Vardi's NPMLE $\mathbb{G}_n \equiv \mathbb{G}_n(\hat{\theta}_n)$ from data with "known" weight functions $w_i(\cdot, \hat{\theta}_n)$.

3 Estimate W_i by $\mathbb{W}_{ni} \equiv \mathbb{W}_{ni}(\hat{\theta}_n) = \int w_i(u, \hat{\theta}_n)\, d\mathbb{G}_n(u)$, $i = 1, \ldots, s$.

The key assumption for this procedure to produce the unique maximum is that a graphical condition holds, defined in Vardi (1985) and restated here. For fixed, known $\theta \in \Theta$, consider the graph $\mathbf{G}(\theta)$ on the s vertices $i = 1, \ldots, s$ defined as follows. Define a directed edge from a vertex i to a vertex k, $i \leftrightarrow k$, if and only if $\sum_{j=1}^{h} w_{ij}(\theta) n_{kj} > 0$. The graph $\mathbf{G}(\theta)$ is *strongly connected* if, for every pair (i, k), there exists a directed path from i to k and a directed path from k to i.

Theorem 1.2, proved in Gilbert, Lele and Vardi (1999), asserts that the above estimation procedure yields the MLE of (θ, G).

THEOREM 1.2. *Suppose $s \geq 2$ with w_s a constant function, and the identifiability conditions of Theorem* 1.1 *hold. Further suppose the graph $\mathbf{G}(\theta)$ is strongly connected for all θ in some neighborhood $\Theta_0 \subset \Theta$ of θ_0, and that $(\hat{\theta}_n, \underline{\mathbb{V}}_n, \mathbb{G}_n)$ is obtained from procedure* 1–3. *If $(\hat{\theta}_n, \underline{\mathbb{V}}_n)$ uniquely maximizes the partial likelihood* (1.11), *then $(\hat{\theta}_n, \mathbb{G}_n)$ uniquely maximizes the full likelihood $L_n(\theta, G|\underline{x})$ of* (1.10).

Procedure 1–3 can be carried out as follows. Step 1 can be accomplished via profile likelihood. For fixed $\theta \in \Theta$, let $\underline{\mathbb{V}}_n(\theta) = (\mathbb{V}_{n1}(\theta), \ldots, \mathbb{V}_{ns-1}(\theta), 1)^T$ be the unique solution of

$$\mathbb{H}_{ni}(V_1(\theta), \ldots, V_s(\theta)) \equiv V_i^{-1}(\theta) \sum_{j=1}^{h} \frac{r_j w_{ij}(\theta)}{\sum_{k=1}^{s} n_k w_{kj}(\theta)\, V_k^{-1}(\theta)} = 1, \tag{1.12}$$

$i = 1, \ldots, s-1$, in the region $V_1(\theta) > 0, \ldots, V_{s-1}(\theta) > 0$, $V_s = 1$. Vardi (1985) proved that (1.12) has a unique solution if and only if the graph $\mathbf{G}(\theta)$ is strongly connected. The estimator $\hat{\theta}_n$ is the argument which maximizes the profile partial likelihood $L_{n1\,pro}$ defined by

$$\begin{aligned} L_{n1\,pro}(\theta|\underline{x}) &= L_{n1\,pro}(\theta, \underline{\mathbb{V}}_n(\theta)|\underline{x}) \\ &= \prod_{i=1}^{s} \prod_{j=1}^{h} \left[\frac{w_{ij}(\theta)\mathbb{V}_{ni}^{-1}(\theta)}{\sum_{k=1}^{s} \lambda_{nk} w_{kj}(\theta)\mathbb{V}_{nk}^{-1}(\theta)} \right]^{n_{ij}}. \end{aligned} \tag{1.13}$$

Set $\underline{\mathbb{V}}_n = \underline{\mathbb{V}}_n(\hat{\theta}_n)$. Then Step 2 proceeds by setting $\hat{p} = p(\hat{\theta}_n)$, where

$$p_j(\theta) \propto r_j \Big/ \sum_{k=1}^{s} n_k\, w_{kj}\,(\theta)\, \mathbb{V}_{nk}^{-1}(\theta), \qquad j = 1, \ldots, h.$$

Hence, $\mathbb{G}_n(A) \equiv \mathbb{G}_n(A, \hat{\theta}_n) = \frac{1}{n} \sum_{j=1}^{h} 1_A(t_j) \hat{p}_j$ is Vardi's NPMLE evaluated at $\hat{\theta}_n$, which exists uniquely if $\mathbf{G}(\hat{\theta}_n)$ is strongly connected. The semiparametric MLE $\mathbb{G}_n$ is written explicitly as

$$\mathbb{G}_n(A) = \frac{\frac{1}{n} \sum_{j=1}^{h} 1_A(t_j) r_j \left[\sum_{k=1}^{s} \lambda_{nk}\, w_{kj}\,(\hat{\theta}_n)\, \mathbb{V}_{nk}^{-1}\,(\hat{\theta}_n) \right]^{-1}}{\frac{1}{n} \sum_{j=1}^{h} r_j \left[\sum_{k=1}^{s} \lambda_{nk}\, w_{kj}\,(\hat{\theta}_n)\, \mathbb{V}_{nk}^{-1}\,(\hat{\theta}_n) \right]^{-1}}.$$

The semiparametric MLE can also be written in terms of the empirical measure $\mathbb{F}_n$, defined by

$$\mathbb{F}_n(A) = \frac{1}{n}\sum_{i=1}^{s}\sum_{j=1}^{n_i} 1_A(Y_{ij}) = \sum_{i=1}^{s}\lambda_{ni}\frac{1}{n_i}\sum_{j=1}^{n_i} 1_A(Y_{ij}) = \sum_{i=1}^{s}\lambda_{ni}\mathbb{F}_{ni}(A) \tag{1.14}$$

for $A \in \mathbf{B}$, $i = 1, \ldots, s$, where Y_{ij} is the jth observation from sample i. We have

$$\mathbb{G}_n(A) = \frac{\int_A \left[\sum_{k=1}^{s}\lambda_{nk} w_k(y, \hat{\theta}_n)\mathbb{V}_{nk}^{-1}(\hat{\theta}_n)\right]^{-1} d\mathbb{F}_n(y)}{\int_{\mathbf{Y}} \left[\sum_{k=1}^{s}\lambda_{nk} w_k(y, \hat{\theta}_n)\mathbb{V}_{nk}^{-1}(\hat{\theta}_n)\right]^{-1} d\mathbb{F}_n(y)}. \tag{1.15}$$

The above procedure 1–3 is computationally attractive because it only requires calculation of Vardi's NPMLE $\mathbb{G}_n$ one time; once $(\hat{\theta}_n, \underline{\mathbb{V}}_n)$ is obtained, $\hat{p}$ is obtained through substitution only. Thus, in essence, the procedure only requires maximizing a function over a finite-dimensional parameter space and solving a system of equations of fixed $(s-1)$ dimension.

This paper is organized as follows. For identifiable s-sample semiparametric biased sampling models, $s \geq 2$, with w_s a constant function, large sample properties of the MLE are developed in Sections 2 through 6. We begin in Section 2 by discussing conditions for unique estimability of the model in the limit with probability 1. Section 3 provides heuristic discussion of the approach taken and defines notation. In Section 4, information bounds for estimation of θ and G are calculated. Then results parallel to GVW's results on consistency, asymptotic normality and efficiency are given in Section 5. An asymptotically consistent estimator of the limiting covariance process is constructed by substitution of the MLE into the inverse generalized Fisher information. In Section 6, the theorems derived in this paper are applied to the examples introduced in Section 1. Proofs are presented in Section 7. Concluding comments and open problems are discussed in Section 8.

2. Estimability and uniqueness. Under what conditions does the likelihood (1.10) have a unique maximum in (θ, G) with probability 1 as $n \to \infty$? As stated in Theorem 1.2, the problem of maximizing the likelihood is equivalent to maximizing the partial likelihood (1.11), which can be accomplished by maximizing the profile partial likelihood (1.13). Thus our approach is to identify mild conditions under which the profile partial likelihood (1.13), and hence the full likelihood (1.10), has a unique maximum with probability 1.

For known θ, Vardi [(1985), Lemma, page 197] showed that if the graph $\mathbf{G}^{\star}(\theta)$ is connected, then a unique solution $\underline{\mathbb{V}}_n(\theta)$ of (1.12) exists with probability 1 as $n \to \infty$. Therefore, assuming the graph $\mathbf{G}^{\star}(\theta)$ is connected for all θ in some neighborhood $\Theta_0 \subset \Theta$ of θ_0, the first step in maximizing the profile partial likelihood always works when the sample size is large. Then the profile partial likelihood will have a unique maximum in θ with probability 1 as n grows large if the limiting log profile partial likelihood is strictly concave on Θ. Thus we have proved the following theorem.

THEOREM 2.1 (Unique estimability via maximum partial likelihood estimation). *Suppose* $s \geq 2$, w_s *is a constant function*, *the identifiability conditions of Theorem* 1.1 *hold*, *and for each* $i = 1, \ldots, s$, $n_i \to \infty$ *and* $\lambda_{ni} \to \lambda_i > 0$. *Further suppose the limiting log profile partial likelihood is strictly concave on* Θ. *Then with probability converging to* 1, *the likelihood* (1.10) *has a unique maximum, which can be obtained by the maximum partial likelihood estimation procedure* 1–3.

Henceforth assume the data and weights satisfy the hypotheses of Theorem 2.1, so that the s-sample biased sampling model is identifiable and uniquely estimable asymptotically. For this class of biased sampling models, we establish the large sample theory.

3. Preliminaries: asymptotic theory for the MLE. GVW characterized minimal assumptions on the known selection weight functions and the data under which Vardi's NPMLE $\mathbb{G}_n$ is uniformly $\sqrt{n}$-consistent, asymptotically normal and efficient, where the uniformity is over a Donsker class. Essentially, the hypotheses needed for these results are connectedness of the graph $\mathbf{G}^\star$, $n_i \to \infty$ with $\lambda_{ni} \to \lambda_i > 0$, $i = 1, \ldots, s$, the normalizing constants $W_i(G)$ are finite and at least one weight function is bounded away from zero. Under these conditions (assuming that $\mathbf{G}^\star(\theta)$ is connected for all θ in some neighborhood $\Theta_o \subset \Theta$ of θ_0), we give sufficient additional integrability and smoothness hypotheses on the weight functions in order that the joint MLE $(\hat{\theta}_n, \mathbb{G}_n)$ retains these asymptotic properties. The basic extra conditions are twice-differentiability of each weight function $w_i(\cdot, \theta)$ in θ at θ_0 and square-integrability of each "score for w_i" $\partial/\partial\theta \log\{w_i(\cdot, \theta)\}$ in a neighborhood of θ_0.

Consistency is derived by following GVW's approach and utilizing special features of the maximum profile partial likelihood estimation procedure. At first thought, since the dimension of the parameter space of the semiparametric estimate based on (1.10) grows with the sample size, one might expect that the estimate of θ suffers from inconsistency due to infinitely many incidental parameters; see, for example, Kiefer and Wolfowitz (1956). However, the equivalence of maximizing the full likelihood over an infinite-dimensional parameter space with maximizing the partial likelihood over a finite-dimensional parameter space (Theorem 1.2) implies that this is not the case. Asymptotic normality of $(\hat{\theta}_n, \mathbb{G}_n)$ is proved as follows. First, asymptotic normality of $\hat{\theta}_n$ is established through application of a Master theorem [see, e.g., Bickel, Klaassen, Ritov and Wellner (1993), Theorem 1, page 312] to the finite collection of estimating equations formed by setting derivatives of the log profile partial likelihood (1.13.) to zero. The key stochastic, approximation hypothesis of the Master theorem is verified using tools in empirical process theory. Second, follow Gill, Vardi and Wellner's (1988) (henceforth GVW) proof of asymptotic normality of $\mathbb{G}_n$, adding the necessary hypotheses for it to go through for $\mathbb{G}_n$ evaluated at $\hat{\theta}_n$. Efficiency is established by comparison of the limiting covariance structure of the MLE to the information bounds. The needed semi-

parametric efficiency theory is largely taken from Bickel, Klaassen, Ritov and Wellner (1993), henceforth referred to as BKRW.

An alternative method for establishing asymptotic normality and efficiency of infinite-dimensional maximum likelihood estimators, developed by Van der Vaart (1995), does not require calculation of the information bounds. However, the hypotheses of Van der Vaart's powerful approach are difficult to verify for this problem. We have successfully applied Van der Vaart's method for the case $\mathbf{Y} = R$, but have not yet succeeded for a general space $\mathbf{Y}$ [Gilbert (1996)]. The difficulty is in identifying a set of functions on which the score operator acts and a metric space for this set for which Van der Vaart's Fréchet differentiability and continuous invertibility hypotheses can both be verified. To obtain maximum generality with minimal assumptions, we directly extend the approach of GVW, exploiting the elegant structure in the maximum profile partial likelihood estimation procedure.

3.1. *Notation.* We give some notation, which corresponds to that used by GVW. Let $(\theta_0, \underline{V}_0, G)$ denote the "true" parameter value, and let $F_{i\theta}(A) \equiv F_i(A, \theta, G)$, $A \in \mathbf{B}$. Define

$$\tilde{w}_i(y, \theta) \equiv w_i(y, \theta)/W_i(\theta, G) \quad \text{and} \quad r(y, \theta) \equiv \left[\sum_{i=1}^{s} \lambda_i \tilde{w}_i(y, \theta)\right]^{-1}.$$

Let $\underline{r} = (r_1, \ldots, r_s)^T$, with $r_i(y, \theta) \equiv \lambda_i \tilde{w}_i(y, \theta)/\sum_{k=1}^{s} \lambda_k \tilde{w}_k(y, \theta) = \lambda_i \tilde{w}_i(y, \theta)$ $r(y, \theta)$. Notice that $\sum_{i=1}^{s} r_i(y, \theta) = 1$ for all y and θ. Put $\phi(y, \theta) = \sum_{i=1}^{s} r_i(y, \theta)$ $(\dot{w}_i(y, \theta)/w_i(y, \theta))$, and let $\underline{w} \equiv (w_1, \ldots, w_s)^T$, $\underline{\tilde{w}} \equiv (\tilde{w}_1, \ldots, \tilde{w}_s)^T$, $\underline{W}_0 \equiv (W_1(\theta_0, G), \ldots, W_s(\theta_0, G))^T$, $\underline{\mathbb{W}}_n \equiv (\mathbb{W}_{n1}, \ldots, \mathbb{W}_{ns})^T$, $\underline{V}(\theta) = (V_1(\theta), \ldots, V_s(\theta))^T$ and $\underline{\underline{\lambda}}$ be the diagonal matrix with diagonal $(\lambda_1, \ldots, \lambda_s)^T$. For $\theta \in \Theta$, define an $s \times s$ matrix $M(\theta)$ by

$$\begin{aligned} M(\theta) &= \underline{\underline{\lambda}}^{-1} - \int r(y, \theta)\underline{\tilde{w}}(y, \theta)\underline{\tilde{w}}^T(y, \theta)\, dG(y) \\ &= \underline{\underline{\lambda}}^{-1} - G\big(r(\theta)\underline{\tilde{w}}(\theta)\underline{\tilde{w}}^T(\theta)\big). \end{aligned} \tag{3.1}$$

As in GVW (page 1083), the matrix $M(\theta)$ is singular and has rank $s - 1$. Let $M^-(\theta)$ be a $\{1, 2\}$-generalized inverse of $M(\theta)$. The needed facts about $M^-(\theta)$ are given in Lemmas 5.1 and 5.2 of GVW, which hold for each fixed $\theta \in \Theta$. Let r_n, r_{ni}, ϕ_n, $\underline{\underline{\lambda}}_n$ and $M_n(\theta)$ equal r, r_i, ϕ, $\underline{\underline{\lambda}}$ and $M(\theta)$, respectively, with the λ_i's replaced with λ_{ni}'s and the W_i's replaced with $\mathbb{W}_{ni}$'s.

Let $\overline{F}_\theta(y) \equiv \sum_{i=1}^{s} \lambda_i F_{i\theta}(y)$ be the average of the biased sampling distributions (1.3), and $\overline{F}_{n\theta}(y) = \sum_{i=1}^{s} \lambda_{ni} F_{i\theta}(y)$. Notice from (1.3) that $\overline{F}_\theta$ and G are related by

$$d\overline{F}_\theta(y) = r^{-1}(y, \theta)\, dG(y) \quad \text{and} \quad dG(y) = r(y, \theta)\, d\overline{F}_\theta(y). \tag{3.2}$$

With the observations viewed as i.i.d. random variables $X_k = (I_k, Y_k)$, $k = 1, \ldots, n$, $\overline{F}_\theta(h(Y))$ is the expectation of h under the law of X_1. For a function h depending on $x = (i, y)$, define

$$P_{(\theta,G)}\big[h(X)\big] = E_{(\theta,G)}\big[h(I, Y)\big] = \sum_{i=1}^{s} \lambda_i F_{i\theta}\big(h(i, Y)\big) = \sum_{i=1}^{s} \lambda_i G\big(h(i, Y)\tilde{w}_i(Y, \theta)\big).$$

(Evidently, $P_{(\theta,G)}[h(Y)] = \overline{F}_\theta[h(Y)]$.) Similarly define the biased sampling empirical measure by $\mathbb{P}_n[h(X)] = \sum_{i=1}^{s} \lambda_{ni} \mathbb{F}_{ni}[h(i, Y)]$. Let $P_0 = P_{(\theta_0,G)}$ be the "true" semiparametric biased sampling distribution. Let $\|\cdot\|_1$ denote the L_1 norm $\|\underline{x}\|_1 = \sum_{i=1}^{d} |x_i|$ for $\underline{x} \in R^d$.

4. Information bounds. We compute expressions for the limiting covariance process of an efficient estimator of (θ_0, G) in the semiparametric biased sampling model. The notion according to which it is efficient is that of a best regular estimator sequence; see, for example, BKRW [page 21]. An estimator sequence is considered to be asymptotically efficient relative to the tangent set $\dot{\mathbf{P}}_0^0$ if, among the class of regular estimators, its limiting variance has the least dispersion. The limiting process with "least dispersion" is defined by the information bounds $I^{-1}(\theta_0)$ for θ_0 and I_G^{-1} for G.

Let $\dot{l}_\theta$ be the score for θ, the derivative of the log biased sampling density p of (1.1) with respect to θ, given by

$$\begin{aligned} \dot{l}_\theta(x) &= \frac{\dot{w}_i(y, \theta)}{w_i(y, \theta)} - E\left[\frac{\dot{w}_i(Y, \theta)}{w_i(Y, \theta)} | I = i\right] \\ &= \frac{\dot{w}_i(y, \theta)}{w_i(y, \theta)} - G\left(\frac{\dot{w}_i(Y, \theta)}{w_i(Y, \theta)}\tilde{w}_i(Y, \theta)\right), \end{aligned} \tag{4.1}$$

where $\dot{w}_i(y, \theta) \equiv (\partial/\partial\theta')w_i(y, \theta')|_{\theta'=\theta}$. We refer to $\dot{w}_i(\cdot, \theta)/w_i(\cdot, \theta) = (\partial/\partial\theta')\log\{w_i(\cdot, \theta')\}|_{\theta'=\theta}$ as the *score for* w_i, and denote the kth component of $\dot{w}_i$, $k = 1, \ldots, d$, by $\dot{w}_{ik}(y, \theta) \equiv (\partial/\partial\theta'_k)w_i(y, \theta')|_{\theta'=\theta}$. For a tangent $h \equiv (\partial/\partial\eta)\log g_\eta|_{\eta=0} \in L_2^0(G)$, where $L_2^0(G)$ is the space of functions with $\int h^2\, dG < \infty$ and $\int h\, dG = 0$, the *score (operator) for* g equals

$$\begin{aligned} \dot{l}_g h(x) &= \frac{\partial}{\partial\eta} \log dP_{(\theta,G_\eta)}|_{\eta=0} \\ &= h(y) - E[h(Y)|I = i] \\ &= h(y) - G\big(h(Y)\tilde{w}_i(Y, \theta)\big). \end{aligned} \tag{4.2}$$

Define a weighted average score for θ over the s samples by $\phi^\star(y, \theta) = \sum_{i=1}^{s} r_i(y, \theta)\dot{l}_\theta(i, y)$.

4.1. *Information bound for* θ_0.

THEOREM 4.1 (Information bound for θ_0). *Suppose each weight function* $w_i(y,\theta)$ *is differentiable in* θ *at* θ_0. *Further suppose*:

(I) *For some* $i \in \{1,\dots,s\}$, *there does not exist a vector* $\underline{b} \in R^d$, $b \neq 0$, *such that*

$$\underline{b}^T\Bigg(\frac{\dot{w}_i(y,\theta_0)}{w_i(y,\theta_0)} - \Big[\phi^\star(y,\theta_0) - G\big(\phi^\star(\theta_0)\big)\Big] - G\big(\phi^\star(\theta_0)\underline{\tilde{w}}^T(\theta_0)\big)\big(M^-(\theta_0)\big)^T\Big[\underline{\tilde{w}}(y,\theta_0)r(y,\theta_0) - G\big(\underline{\tilde{w}}(\theta_0)r(\theta_0)\big)\Big]\Bigg) \tag{4.3}$$

$$\equiv b^T\Bigg(\frac{\dot{w}_i(y,\theta_0)}{w_i(y,\theta_0)} - a_\star(y,\theta_0)\Bigg)$$

is a.s. constant with respect to the law F_{i0}. *Then*:

A. *The unique efficient score function* $\dot{l}^\star_{\theta_0}$ *for* θ_0 *is*

$$\dot{l}^\star_{\theta_0}(x) = \frac{\dot{w}_i(y,\theta_0)}{w_i(y,\theta_0)} - a_\star(y,\theta_0) - E\Big[\frac{\dot{w}_i(Y,\theta_0)}{w_i(Y,\theta_0)} - a_\star(Y,\theta_0)|I=i\Big],$$

where

$$\begin{aligned} a_\star(y,\theta_0) \equiv{}& \phi^\star(y,\theta_0) - G\big(\phi^\star(\theta_0)\big) \\ &+ G\big(\phi^\star(\theta_0)\underline{\tilde{w}}^T(\theta_0)\big)\big(M^-(\theta_0)\big)^T \\ &\times\Big[\underline{\tilde{w}}(y,\theta_0)r(y,\theta_0) - G\big(\underline{\tilde{w}}(\theta_0)r(\theta_0)\big)\Big]. \end{aligned}$$

B. *The efficient information matrix* $I(\theta_0)$ *for* θ_0 *is nonsingular, and is given by*

$$\begin{aligned} P_0\Big[\dot{l}^\star_{\theta_0}(I,Y)\dot{l}^{\star T}_{\theta_0}(I,Y)\Big] = P_0\Bigg\{&\frac{\dot{w}_I(Y,\theta_0)}{w_I(Y,\theta_0)}\frac{\dot{w}_I^T(Y,\theta_0)}{w_I(Y,\theta_0)} \\ &- \phi(Y,\theta_0)\phi(Y,\theta_0)^T\Bigg\} \\ &- A(\theta_0)M^-(\theta_0)A(\theta_0)^T, \end{aligned} \tag{4.4}$$

where $M^-(\theta_0)$ *is a* $\{1,2\}$*-generalized inverse of* $M(\theta_0)$ *and*

$$\begin{aligned} A(\theta_0) = G\Bigg\{&\Bigg(\frac{\dot{w}_1(Y,\theta_0)}{w_1(Y,\theta_0)} - \phi(Y,\theta_0)\Bigg)\tilde{w}_1(Y,\theta_0), \\ &\dots, \Bigg(\frac{\dot{w}_s(Y,\theta_0)}{w_s(Y,\theta_0)} - \phi(Y,\theta_0)\Bigg)\tilde{w}_s(Y,\theta_0)\Bigg\}. \end{aligned} \tag{4.5}$$

Expression (4.3) *and the efficient score and information are independent of the choice of* $\{1, 2\}$*-generalized inverse* $M^-(\theta_0)$.

When $s = 2$, *the inverse efficient information* (*covariance*) *matrix* $I^{-1}(\theta_0)$ *equals*

$$\frac{1}{\lambda_1\lambda_2}\Bigg[G\Big(\frac{\dot w_1(\theta_0)}{w_1(\theta_0)}\frac{\dot w_1^T(\theta_0)}{w_1(\theta_0)}\tilde w_1(\theta_0)r(\theta_0)\Big) - \frac{G\Big(\frac{\dot w_1(\theta_0)}{w_1(\theta_0)}\tilde w_1(\theta_0)r(\theta_0)\Big)G\Big(\frac{\dot w_1^T(\theta_0)}{w_1(\theta_0)}\tilde w_1(\theta_0)r(\theta_0)\Big)}{G\big(\tilde w_1(\theta_0)r(\theta_0)\big)}\Bigg]^{-1}. \tag{4.6}$$

REMARK 4.1. The reason $\dot l^\star_{\theta_0}$ and $I(\theta_0)$ do not depend on the choice of $\{1, 2\}$-generalized inverse $M^-(\theta_0)$ is because each row of $G(\phi^\star(\theta_0)\underline{\tilde w}^T(\theta_0))$, and the vector $[r(y, \theta_0)\underline{\tilde w}(y, \theta_0) - G(r(\theta_0)\underline{\tilde w}(\theta_0))]$ is in the range of M, equal to Range $(M) = \{\underline{x}\colon \lambda^T\underline{x} = 0\}$. See Lemma 5.2 (iv) of GVW. In applications, a convenient choice of $M^-(\theta_0)$ is

$$M^-(\theta_0) = \begin{pmatrix} M_{11}^{-1}(\theta_0) & 0 \\ 0 & 0 \end{pmatrix}, \tag{4.7}$$

where $M_{11}(\theta_0)$ is the upper-left $s - 1 \times s - 1$ submatrix of $M(\theta_0)$. GVW [page 1080] proved that $M_{11}(\theta_0)$ is nonsingular under connectivity of $\mathbf{G}^\star(\theta_0)$.

REMARK 4.2. Note that by the Cauchy–Schwarz inequality, the information for θ_0 when $s = 2$ and $d = 1$ is positive if and only if $\dot w_1(\cdot, \theta_0)/w_1(\cdot, \theta_0)$ is nondegenerate, which is true if and only if hypothesis (I) of Theorem 4.1 holds. This sheds light on the necessity of hypothesis (I).

REMARK 4.3. The information for estimation of the regression parameter θ_0 in the Cox proportional hazards model takes a similar form. Let Z denote a covariate and T a failure time, with no censoring. Under the proportional hazards model $\lambda(t|z) = \lambda(t|0)\exp\{\theta_0^T z\}$, the information for θ_0 is $I(\theta_0) = E\,\mathrm{Var}[Z|T]$. This is derived in BKRW (pages 80–82) by computation of the efficient score via orthogonality calculations. This same approach is used in the proof of Theorem 4.1, and $I(\theta_0)$ of (4.4) has the same structure as $I(\theta_0)$ in the Cox model, equal to $P_0\,\mathrm{Var}[[\dot w_I(Y, \theta_0)/w_I(Y, \theta_0)] - \alpha_\star(Y, \theta_0)|I]$.

4.2. *Information bound for* G. Since G is an infinite-dimensional parameter, the information bound is an *inverse information covariance functional* $I^{-1}(P_0|G, \mathbf{P}) \equiv I_G^{-1}\colon \mathbf{H} \times \mathbf{H} \to R$. It is defined, as in BKRW [page 184], by $I_G^{-1}(h_1, h_2) \equiv E[\tilde l_G(\pi_{h_1})\tilde l_G(\pi_{h_2})]$, where $\tilde l_G(\pi_h)$ is the efficient influence function for estimation of $G(h)$. Here π_h is the projection map from $l^\infty(\mathbf{H})$ to R, defined by $\pi_h(G) = G(h) = \int h\,dG$, where we now view probability distributions G as elements of $l^\infty(\mathbf{H})$, the Banach space of all bounded functions $z\colon \mathbf{H} \to R$, which is equipped with the supremum norm.

THEOREM 4.2 (Information bound for G). *Suppose the conditions of Theorem* 4.1 *hold. Additionally suppose* $r(y, \theta_0)$ *and* $r^{-1}(y, \theta_0)$ *are bounded in* y. *Then the inverse information covariance functional* I_G^{-1} *for estimation of* G *is*

$$
\begin{aligned}
I_G^{-1}(h_1, h_2) &= G\big([h_1 - G(h_1)][h_2 - G(h_2)]r(\theta_0)\big) \\
&\quad + G\big([h_1 - G(h_1)]r(\theta_0)\underline{\tilde{w}}^T(\theta_0)\big)M^-(\theta_0) \\
&\quad \times G\big([h_2 - G(h_2)]r(\theta_0)\underline{\tilde{w}}(\theta_0)\big) \\
&\quad + G\big(h_1 a_\star^T(\theta_0)\big) * I^{-1}(\theta_0) * G\big(h_2 a_\star(\theta_0)\big) \\
&= I_{G^0}^{-1}(h_1, h_2) \\
&\quad + G\big(h_1 a_\star^T(\theta_0)\big) * I^{-1}(\theta_0) * G\big(h_2 a_\star(\theta_0)\big)
\end{aligned}
\tag{4.8}
$$

For $h_1, h_2 \in \mathbf{H}$. The first three lines of (4.8) equal $I_{G^0}^{-1}$, the inverse information covariance functional for estimation of G when θ_0 is known. This functional was derived by GVW [pages 1089–1091] through computation of the inverse information operator for g, $(\dot{l}_g^T \dot{l}_g)^{-1}$. The fourth line of (4.8) is the inflation of the covariance functional when θ_0 is unknown.

Since $G([h-G(h)]r(\theta_0)\underline{\tilde{w}}(\theta_0))$ and the rows of $G(\phi(\theta_0)\underline{\tilde{w}}^T(\theta_0))$ are in Range (M), the information bound for G is independent of the choice of generalized inverse $M^-(\theta_0)$.

REMARK 4.4. If θ_0 is known and there is no biasing in the sampling, r and all the weights are unity, so that I_G^{-1} reduces to $I_G^{-1}(h_1, h_2) = G(h_1 h_2) - G(h_1)G(h_2)$. This is the information bound for estimation of a probability measure G from an independent, identically distributed sample from G. It is well known that the empirical measure is regular and has limiting covariance given by $G(h_1 h_2) - G(h_1)G(h_2)$, so that it is efficient. Therefore, the inverse information covariance functional I_G^{-1} matches this known special case, as it must.

REMARK 4.5. The assumption that $r^{-1}(y, \theta_0)$ is bounded in y, which is equivalent to each of the weight functions $w_i(y, \theta_0)$, $i = 1, \ldots, s$, being bounded in y, is only used for calculation of the information bound for G. This technically unpleasant assumption is needed to ensure existence of the inverse information operator $(\dot{l}_g^T \dot{l}_g)^{-1}$.

We write down I_G^{-1} for two-sample semiparametric biased sampling models. We calculate $M^-(\theta_0)$ of (4.7) to equal the 2×2 matrix with $\lambda_1[\lambda_2 G(\tilde{w}_1(\theta_0) r(\theta_0))]^{-1}$ as the upper-left element and zeros elsewhere. Then direct calculation gives

$$
a^\star(y, \theta) = r_1(y, \theta)\left[\frac{\dot{w}_1(y, \theta)}{w_1(y, \theta)} - G\left(\frac{\dot{w}_1(\theta)}{w_1(\theta)}\tilde{w}_1(\theta)r(\theta)\right) \Big/ G\Big(\tilde{w}_1(\theta)r(\theta)\Big)\right]
$$

and, with all functions evaluated at θ_0,

$$
\begin{aligned}
I_G^{-1}(h_1,h_2) = {} & G\big([h_1 - G(h_1)][h_2 - G(h_2)]r\big) \\
& + G\big([h_1 - G(h_1)]r\tilde{w}_1\big)\frac{\lambda_1}{\lambda_2 G(r\tilde{w}_1)} G\big([h_2 - G(h_2)]r\tilde{w}_1\big) \\
& + \left\{ \lambda_1 \left[G\left(h_1 \frac{\dot{w}_1}{w_1}\tilde{w}_1 r \right) G(\tilde{w}_1 r) - G\left(\frac{\dot{w}_1}{w_1}\tilde{w}_1 r \right) G(h_1\tilde{w}_1 r) \right] \right. \\
& \qquad \left. \times \left[G\left(h_2 \frac{\dot{w}_1^T}{w_1}\tilde{w}_1 r \right) G(\tilde{w}_1 r) - G\left(\frac{\dot{w}_1^T}{w_1}\tilde{w}_1 r \right) G(h_2\tilde{w}_1 r) \right] \right\} \\
& \times \left\{ \lambda_2 G(\tilde{w}_1 r) \left[G\left(\frac{\dot{w}_1}{w_1}\frac{\dot{w}_1^T}{w_1}\tilde{w}_1 r \right) G(\tilde{w}_1 r) \right. \right. \\
& \qquad \left. \left. - G\left(\frac{\dot{w}_1}{w_1}\tilde{w}_1 r \right) G\left(\frac{\dot{w}_1^T}{w_1}\tilde{w}_1 r \right) \right] \right\}^{-1}.
\end{aligned} \tag{4.9}
$$

The consistent plug-in estimate of $\mathrm{Cov}(\mathbb{G}_n(h_1), \mathbb{G}_n(h_2))$ is given by the empirical version of (4.9).

To work out the asymptotic covariance between $\sqrt{n}(\hat{\theta}_n - \theta_0)$ and $\sqrt{n}(\hat{G}_n(h) - G(h))$ for an efficient estimator $(\hat{\theta}_n, \hat{G}_n)$, suppose $\hat{\theta}_n$ and $\hat{G}_n$ are each asymptotically linear, with efficient influence functions $\tilde{l}_{\theta_0} = I^{-1}(\theta_0)\dot{l}^{\star}_{\theta_0}$ and $\tilde{l}_G$, respectively. Set $Z_{\theta_0} \equiv \lim_n \sqrt{n}(\hat{\theta}_n - \theta_0)$ and $Z(h) = \lim_n \sqrt{n}(\hat{G}_n(h) - G(h))$. As in BKRW (panel 5.5.27, page 216), since $\tilde{l}_G(\pi_h) = \dot{l}_g(\dot{l}_g^T \dot{l}_g)^{-1}(\pi_h) - l^{\star T}_{\theta_0} I^{-1}(\theta_0) G([h - G(h)]a_{\star}(\theta_0))$, where $l^{\star}_{\theta_0}$ is orthogonal to Range $(\dot{l}_g)$,

$$
\begin{aligned}
\mathrm{Cov}(Z_{\theta_0}, Z(h)) &= P_0\big(\tilde{l}_{\theta_0}(X)\tilde{l}_G(\pi_h)(X)\big) \\
&= P_0\left(\left[I^{-1}(\theta_0) l^{\star}_{\theta_0} \right] \left[-l^{\star T}_{\theta_0} I^{-1}(\theta_0) G\big([h - G(h)]a_{\star}(\theta_0)\big) \right] \right) \\
&= -I^{-1}(\theta_0) G\big([h - G(h)]a_{\star}(\theta_0)\big) \\
&= -I^{-1}(\theta_0) G\big(h a_{\star}(\theta_0)\big).
\end{aligned} \tag{4.10}
$$

5. Asymptotic behavior of $(\hat{\theta}_n, \mathbb{G}_n)$.

5.1. *Consistency.* Consistency of $(\hat{\theta}_n, \mathbb{G}_n)$ is established in two steps. First, consistency of $\hat{\theta}_n$ is proved under the hypotheses that the log limiting profile partial likelihood $l_{1\,pro}(\theta)$ and the log profile partial likelihood $l_{n1\,pro}(\theta|\underline{x}) \equiv \log L_{n1\,pro}(\theta|\underline{x})$ for all n large enough are strictly concave on Θ. Since $\mathbb{G}_n \equiv \mathbb{G}_n(\hat{\theta}_n)$ is Vardi's NPMLE evaluated at $\hat{\theta}_n$, consistency of $\mathbb{G}_n$ essentially follows by consistency of $\hat{\theta}_n$ and GVWs proof that Vardi's NPMLE is consistent.

From (1.13), the log profile partial likelihood $l_{n1\,pro}$ is given by

$$\frac{1}{n} l_{n1\,pro}(\theta|\underline{x}) = \sum_{i=1}^{s} \lambda_{ni} \frac{1}{n_i} \sum_{j=1}^{n_i} \log\left\{ \frac{w_{ij}(\theta)\mathbb{V}_{ni}^{-1}(\theta)}{\sum_{k=1}^{s} \lambda_{nk} w_{kj}(\theta)\mathbb{V}_{nk}^{-1}(\theta)} \right\} \tag{5.1}$$

$$= \mathbb{P}_n \log\{\lambda_{nI}^{-1} r_{nI}(Y, \theta)\}.$$

THEOREM 5.1 (Consistency of $\hat{\theta}_n$). *Suppose that*:

(i) *The parameter space $\Theta \subset R^d$ is an open convex set.*

(ii) *Each weight function w_i is continuously differentiable in θ in a neighborhood of θ_0.*

(iii) *For all n larger than some fixed N, $l_{n1\,pro}(\theta|\underline{x})$ is strictly concave on Θ.*

Then as each $n_i \to \infty$ with $\lambda_{ni} \to \lambda_i > 0$,

$$\frac{1}{n} l_{n1\,pro}(\theta|\underline{x}) \to_{\text{a.s.}} l_{1\,pro}(\theta)$$

$$= \sum_{i=1}^{s} \lambda_i F_{i\theta} \log\left\{ \frac{w_i(Y, \theta)V^{-1}(\theta)}{\sum_{k=1}^{s} \lambda_k w_k(Y, \theta) V_k^{-1}(\theta)} \right\} \tag{5.2}$$

$$= P_{(\theta,G)} \log\{\lambda_I^{-1} r_I(Y, \theta)\}$$

for each fixed $\theta \in \Theta$. By strict concavity of $l_{n1\,pro}(\theta|\underline{x})$ on Θ for all $n > N$ and strict concavity of $l_{1\,pro}(\theta)$ on Θ, which is implied by (iii) *and* (5.2), *it follows that $l_{n1\,pro}(\theta|\underline{x})$ has a unique maximum at $\hat{\theta}_n$ for every $n > N$ and $l_{1\,pro}$ has a unique maximum at θ_0, so that $\hat{\theta}_n \to_{\text{a.s.}} \theta_0$ as $n \to \infty$.*

REMARK 5.1. Through theoretical results and simulations, Gilbert (1996) and Gilbert, Lele and Vardi (1999) (see especially Figure 2 and Theorems 5 and 6) showed that $l_{n1\,pro}(\theta|\underline{x})$ and $l_{1\,pro}(\theta)$ are strictly concave on $\Theta \subset R^d$ for a large class of weight functions and data sets. A sufficient general condition for the concavity hypothesis (iii) to hold is that condition (I) of Theorem 4.1 holds for the function of (4.3) evaluated at every $\theta \in \Theta$. In the proof of Theorem 5.3, we show that the second derivative matrix of $l_{1\,pro}(\theta)$ equals $-I(\theta)$, so that positive definiteness of $I(\theta)$ for all $\theta \in \Theta$ implied by (I) for all $\theta \in \Theta$ yields strict concavity of $l_{1\,pro}(\theta)$ on Θ. Now, the second derivative matrix of $l_{n1\,pro}(\theta|\underline{x})$ is asymptotically equivalent to minus the observed information matrix $-I_n(\theta)$ by regularity, and $I_n(\theta)$ converges in probability to $I(\theta)$ under the hypotheses of Proposition 5.1. Thus, under these hypotheses and condition (I) for all $\theta \in \Theta$, hypothesis (iii) holds. Condition (I) for all θ in Θ will usually hold if the following easy-to-verify condition holds: for some $i \in \{1, \ldots, s\}$ and all $l \in \{1, \ldots, d\}$ and every $\theta \in \Theta$, $\dot{w}_{il}(y, \theta)/w_i(y, \theta)$ and $w_i(y, \theta)$ are not a.s. constant with respect to the law $F_{i\theta}$.

The first task in establishing consistency of $\mathbb{G}_n$ is establishing consistency of $\underline{\mathbb{V}}_n(\hat{\theta}_n)$, which is defined as the solution of (1.12) evaluated at $\hat{\theta}_n$, or equivalently as the maximizer jointly with $\hat{\theta}_n$ of the partial likelihood (1.11).

PROPOSITION 5.1 [Consistency of $\underline{\mathbb{V}}_n(\hat{\theta}_n)$ and $\underline{\mathbb{W}}_n(\hat{\theta}_n)$]. *Suppose conditions* (i)–(ii) *of Theorem* 5.1 *hold, and its conclusion* $\hat{\theta}_n \to_{\text{a.s.}} \theta_0$. *In addition, suppose there exists a* $\delta > 0$ *such that*

(iv) $\sup_{\|\theta-\theta_0\|_1<\delta} G\big(|\dot{w}_{ik}(\theta)/w_i(\theta)|\tilde{w}_i(\theta)\big) < \infty$ *for all* $i = 1, \ldots, s$, $k = 1, \ldots, d$.

Then the equations (1.12) *have, with probability* 1 *as* $n \to \infty$ *with* $\lambda_{ni} \to \lambda_i > 0$, *the unique solution* $\underline{\mathbb{V}}_n(\hat{\theta}_n) = (\mathbb{V}_{n1}(\hat{\theta}_n), \ldots, \mathbb{V}_{ns-1}(\hat{\theta}_n), 1)^T$ *that satisfies*

$$\underline{\mathbb{V}}_n(\hat{\theta}_n) \to_{\text{a.s.}} \underline{V}(\theta_0) \equiv \underline{V}_0.$$

Moreover, $\mathbb{W}_n(\hat{\theta}_n) \to_{\text{a.s.}} \underline{W}(\theta_0)$.

As in GVW [page 1081], let h_e be a fixed nonnegative G-integrable function, let $\mathbf{C}$ be a Vapnik–Chervonenkis class of subsets of the probability space $\mathbf{Y}$, and consider the collection of functions

$$\mathbf{H} \equiv \{h_e 1_C : C \in \mathbf{C}\}. \tag{5.3}$$

We give conditions under which $\mathbb{G}_n$ is a consistent estimator of G uniformly over $\mathbf{H}$.

THEOREM 5.2 (Consistency of $\mathbb{G}_n$). *Suppose the conditions of Proposition* 5.1 *hold, and that* $\mathbf{H}$ *is defined as in* (5.3), *with* $h_e \in L_1(G)$. *Then as* $n \to \infty$ *with* $\lambda_{ni} \to \lambda_i > 0$,

$$\|\mathbb{G}_n - G\|_{\mathbf{H}} \equiv \sup_{h\in\mathbf{H}}\big\{|\mathbb{G}_n(h) - G(h)|\big\} \to_{\text{a.s.}} 0. \tag{5.4}$$

Theorem 5.2 has two immediate corollaries.

COROLLARY 5.1. *Assume that* $\mathbf{Y} = R^k$ *and* $\mathbf{C}$ *is a Vapnik–Chervonenkis class of subsets of* $\mathbf{Y}$, *and the conditions of Proposition* 5.1 *hold. Then as* $n \to \infty$ *with* $\lambda_{ni} \to \lambda_i > 0$,

$$\|\mathbb{G}_n - G\|_{\mathbf{C}} \equiv \sup_{C\in\mathbf{C}} |\mathbb{G}_n(C) - G(C)| \to_{\text{a.s.}} 0. \tag{5.5}$$

COROLLARY 5.2. *Suppose* $G(|h|) < \infty$, *and the conditions of Proposition* 5.1 *hold. Then* $\mathbb{G}_n(h) \to_{\text{a.s.}} G(h)$ *as* $n \to \infty$ *with* $\lambda_{ni} \to \lambda_i > 0$.

5.2. *Asymptotic distributions.* We formulate the asymptotic distribution theory for $\sqrt{n}(\mathbb{G}_n - G)$ in terms of the s-sample Y-marginal empirical process $\mathbb{X}_n$, defined by

$$\mathbb{X}_n \equiv \sqrt{n}(\mathbb{F}_n - \overline{F}_{n0}) = \sum_{i=1}^{s} \sqrt{\lambda_{ni}}\sqrt{n_i}(\mathbb{F}_{ni} - F_{i0}),$$

where $\mathbb{F}_n$ and $\mathbb{F}_{ni}$ are as in (1.14). If $\mathbf{F}$ is a Donsker class for each F_{i0}, $i = 1, \ldots, s$, then $\{\mathbb{X}_n(f)\colon f \in \mathbf{F}\}$ converges in distribution in $l^\infty(\mathbf{F})$ to the mean zero Gaussian process $\{\mathbb{X}(f)\colon f \in \mathbf{F}\}$ with covariance

$$\begin{aligned}\operatorname{Cov}\big(\mathbb{X}(h_1), \mathbb{X}(h_2)\big) &= \sum_{i=1}^{s} \lambda_i \big\{F_{i0}(h_1h_2) - F_{i0}(h_1)F_{i0}(h_2)\big\} \\ &= G(r^{-1}(\theta_0)h_1h_2) - G\big(h_1\underline{\tilde{w}}^T(\theta_0)\big)\underline{\underline{\lambda}}G\big(h_2\underline{\tilde{w}}(\theta_0)\big).\end{aligned}$$

To avoid problems with measurability, convergence in distribution is defined using outer expectations as in Dudley (1985).

For fixed $\theta \in \Theta$, define the *biased sampling empirical process* by $\mathbb{Z}_n(\cdot, \theta) \equiv \sqrt{n}(\mathbb{G}_n(\theta) - G)$, regarded as a process indexed by a collection of functions $\mathbf{H} \subset L_2(G)$. Thus, for $h \in \mathbf{H}$,

$$\mathbb{Z}_n(h, \theta) = \int h\, d\mathbb{Z}_n(\theta) = \sqrt{n}\int h\, d(\mathbb{G}_n(\theta) - G).$$

We formulate a limit theorem for $\mathbb{Z}_n(\cdot, \hat{\theta}_n)$ jointly with the $\sqrt{n}(\hat{\theta}_n - \theta_0)$ process. This joint process, indexed by $\Theta \times \mathbf{H}$, is defined by

$$\mathbb{Z}'_n(h, \hat{\theta}_n) = \Big(\sqrt{n}(\hat{\theta}_n - \theta_0), \mathbb{Z}_n(h, \hat{\theta}_n)\Big)^T.$$

The appropriate limiting Gaussian distribution is given by

$$Z'(h) = \Big(Z_{\theta_0}, - G\big(ha_\star^T(\theta_0)\big) * Z_{\theta_0} + Z(h)\Big)^T, \tag{5.6}$$

where $Z_{\theta_0} \equiv \lim_n \sqrt{n}(\hat{\theta}_n - \theta_0)$ and Z is a mean zero Gaussian process defined in (2.18) of GVW, with covariance function equal to $I_{G^0}^{-1}$, displayed as the first three lines of expression (4.8). The process Z' has covariance function

$$\begin{aligned}&\operatorname{Cov}\big(Z'(h_1), Z'(h_2)\big) \\ &= \begin{bmatrix} I^{-1}(\theta_0) & -I^{-1}(\theta_0) * G\big(h_2a_\star(\theta_0)\big) \\ -G\big(h_1a_\star^T(\theta_0)\big) * I^{-1}(\theta_0) & I_G^{-1}(h_1, h_2)\end{bmatrix}\end{aligned} \tag{5.7}$$

for $h_1, h_2 \in \mathbf{H}$, where $a_\star$ and $I^{-1}(\theta_0)$ are as in Theorem 4.1 and I_G^{-1} is as in Theorem 4.2.

As in GVW, take $\mathbf{H}$ to be a collection of functions with G-integrable envelope function h_e such that $G(h_e^2 r(\theta_0)) < \infty$ and the class of functions $\mathbf{F} =$

$\{hr(\theta_0)\colon h \in \mathbf{H}\}$ is Donsker for each F_{i0}. We also require that the weight functions (in a neighborhood of θ_0) be included in $\mathbf{H}$, that is, for some $\delta > 0$,

$$\mathbf{W}_\delta \equiv \bigcup_{\|\theta-\theta_0\|_1<\delta} \{w_i(\theta)r(\theta), \ldots, w_s(\theta)r(\theta)\} \tag{5.8}$$

is Donsker for each F_{i0}. Additionally, assume that the finite collection of score functions

$$\dot{\mathbf{W}} \equiv \left\{ \frac{\dot{w}_{1k}(y, \theta_0)}{w_1(y, \theta_0)}, \ldots, \frac{\dot{w}_{sk}(y, \theta_0)}{w_s(y, \theta_0)}\colon k = 1, \ldots, d \right\} \tag{5.9}$$

is Donsker for each F_{i0}.

Before stating our limit theorem for $\mathbb{Z}'_n$, we establish asymptotic normality (and efficiency) of $\hat{\theta}_n$. We apply the Master theorem, originating from Huber (1967), which applies to estimators which approximately zero out a collection of score or estimating equations. In our case, the estimating equations arise from setting the derivative of the log profile partial likelihood (5.1) equal to zero. Recall that $n^{-1}l_{n1\,pro}(\theta|\underline{x}) = \mathbb{P}_n\big[\log\{\lambda_{nI}^{-1}r_{nI}(Y, \theta, \underline{\mathbb{V}}_n(\theta))\}\big]$. Let $\psi_\theta(x) \equiv \psi(x, \theta, \underline{V}(\theta))$ equal $(\partial/\partial\theta')\log\{\lambda_i^{-1}r_i(y, \theta', \underline{V}(\theta'))\}|_\theta$. By the chain rule, we calculate

$$\begin{aligned}
\psi_\theta(x) &= \frac{\partial}{\partial\theta'}\log\{\lambda_t^{-1}r_i(y, \theta', \underline{V})\}\bigg|_{(\theta, \underline{V}(\theta))} \\
&\quad + \frac{\partial}{\partial\theta'}\underline{V}^T(\theta')\bigg|_\theta \frac{\partial}{\partial\underline{V}}\log\{\lambda_i^{-1}r_i(y, \theta, \underline{V})\}\bigg|_{(\theta, \underline{V}(\theta))} \\
&= \left[\frac{\dot{w}_i(y, \theta)}{w_i(y, \theta)} - \phi(y, \theta)\right] + \frac{\partial}{\partial\theta'}\underline{V}^T(\theta')\bigg|_\theta * \underline{\underline{V}}^{-1}(\theta)\big[-\underline{e}_i + \underline{r}(y, \theta)\big],
\end{aligned} \tag{5.10}$$

where $\underline{e}_i$ is the s-column vector with a 1 at position i and zeros elsewhere.

The matrix $(\partial/\partial\theta')\underline{V}^T(\theta')|_\theta$ can be calculated by differentiating both sides of the profile likelihood equations defining $\underline{V}(\theta)$, $P[w_i(\theta)V_i^{-1}(\theta)/\sum_{k=1}^s \lambda_k w_k(\theta) V_k^{-1}(\theta)] = 1$, $i = 1, \ldots, s-1$. Setting $\underline{d}_i(\theta) = (\partial/\partial\theta')V_i(\theta')|_\theta/V_i(\theta)$, $i = 1, \ldots, s$ (with $\underline{d}_s(\theta) = \underline{0}$), this gives

$$\begin{aligned}
\underline{d}_i(\theta) &= G\left[\left(\frac{\dot{w}_i(\theta)}{w_i(\theta)} - \phi(\theta)\right)\tilde{w}_i(\theta)\right] \\
&\quad + \sum_{k=1}^s \lambda_k \underline{d}_k(\theta) G\big(r(\theta)\tilde{w}_i(\theta)\tilde{w}_k(\theta)\big), \qquad i = 1, \ldots, s.
\end{aligned}$$

In matrix notation this is written, with $\underline{\underline{d}}(\theta) = [\underline{d}_1(\theta), \ldots, \underline{d}_s(\theta)]$, as

$$\underline{\underline{d}}(\theta) = A(\theta) + \underline{\underline{d}}(\theta)\underline{\underline{\lambda}}G\big(r(\theta)\underline{\tilde{w}}(\theta)\underline{\tilde{w}}^T(\theta)\big).$$

By definition of the matrix M, we find that the solution $\underline{\underline{d}}(\theta)$ is given by $A(\theta)M^{-}(\theta)\underline{\underline{\lambda}}^{-1}$. Substituting this into (5.10) gives

$$\psi_\theta(x) = \frac{\dot{w}_i(y,\theta)}{w_i(y,\theta)} - \phi(y,\theta) + A(\theta)M^{-}(\theta)\underline{\underline{\lambda}}^{-1}[-\underline{e}_i + \underline{r}(y,\theta)]. \tag{5.11}$$

Note that, if the profile partial likelihood estimator $\hat{\theta}_n$ is efficient, the function ψ_θ will equal the efficient score function $\dot{l}^{\star}_{\theta}$ defined in Theorem 4.1.

Letting $\psi_{n\theta}(x) = \psi_n(x, \theta, \underline{\mathbb{V}}_n(\theta))$ be the empirical version of $\psi_\theta(x)$, with the λ_i, r_i and V_i replaced with λ_{ni}, r_{ni} and $\mathbb{V}_{ni}$, respectively, it follows that

$$\begin{aligned}
\Psi_n(\theta|\underline{x}) &\equiv \frac{\partial}{\partial\theta'}\frac{1}{n} l_{n1\,pro}(\theta'|\underline{x})|_\theta \\
&= \mathbb{P}_n\left[\frac{\partial}{\partial\theta'}\left[\lambda_{nI}^{-1} r_{nI}\left(Y, \theta', \underline{\mathbb{V}}_n(\theta')\right)\right]\Big|_\theta\right] \\
&= \mathbb{P}_n\left[\psi_n\left(X, \theta, \underline{\mathbb{V}}_n(\theta)\right)\right] \\
&= \mathbb{P}_n\left[\psi_\theta(X)\right] + \underline{o}_p(n^{-1/2}),
\end{aligned} \tag{5.12}$$

where the last equality will be verified in the proof of Theorem 5.3.

By the law of large numbers, $\mathbb{P}_n\psi_\theta(X) \to_p P_{(\theta,G)}\psi_\theta(X)$. Let $\dot{\Psi}_0 = P_0\nabla_\theta\psi_\theta(X)|_{\theta_0}$, the derivative matrix of $P_0\psi_\theta(X)$ evaluated at θ_0.

THEOREM 5.3 (Asymptotic normality of $\hat{\theta}_n$). *Suppose there exists a $\delta > 0$ such that $\mathbf{W}_\delta$ and $\dot{\mathbf{W}}$ of* (5.8) *and* (5.9) *are Donsker for each F_{i0}. Further assume hypothesis* (*I*) *of Theorem* 4.1 *and the hypotheses of Proposition* 5.1. *Additionally suppose*:

(v) *Each weight function is twice-differentiable in θ at θ_0.*

(vi) *There exists a $\delta > 0$ such that $\sup_{\|\theta-\theta_0\|_1<\delta} r(y,\theta)$ is bounded in y.*

(vii) *There exists a $\delta > 0$ such that $\sup_{\|\theta-\theta_0\|_1<\delta} G\big(|\dot{w}_{ik}(\theta)/w_i(\theta)|^2\tilde{w}_i(\theta_0)\big) < \infty$ for $i = 1, \ldots, s$, $k = 1, \ldots, d$.*

(viii) *For $i = 1, \ldots, s-1$, $k = 1, \ldots, d$, there exists a constant K, a $\delta > 0$ and an $\alpha > d/2$ such that*

$$\sup_{y\in\mathbf{Y}}\left[\left|\frac{\dot{w}_{ik}(y,\theta_1)}{w_i(y,\theta_1)} - \frac{\dot{w}_{ik}(y,\theta_2)}{w_i(y,\theta_2)}\right|\right] \le K\|\theta_1 - \theta_2\|_1^{\alpha}$$

for all $\theta_1, \theta_2 \in \Theta_\delta \equiv \{\theta \in \Theta: \|\theta - \theta_0\|_1 < \delta\}$.

Then $\dot{\Psi}_0$ is nonsingular and $\hat{\theta}_n$, the unique maximizer of the profile likelihood (1.13), *satisfies*

$$\begin{aligned}
\sqrt{n}\left(\hat{\theta}_n - \theta_0\right) &= -\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\dot{\Psi}_0^{-1}\psi_0(X_i) + o_P(1) \\
&\to_d N_d(\underline{0}, \Sigma),
\end{aligned}$$

where $\Sigma = \dot{\Psi}_0^{-1} P_0[\psi_0(X)\psi_0^T(X)](\dot{\Psi}_0^{-1})^T$. *Here* $\dot{\Psi}_0 = -P_0[\psi_0(X)\psi_0^T(X)]$, *so that* $\Sigma = -\dot{\Psi}_0^{-1}$. *Furthermore*, $-\dot{\Psi}_0 = I(\theta_0)$, *the efficient information matrix* $I(\theta_0)$ *defined in* (4.4), *so that*

$$\sqrt{n}(\hat{\theta}_n - \theta_0) \to_d Z_{\theta_0} \sim N_d(\underline{0}, I^{-1}(\theta_0)).$$

REMARK 5.2. To establish asymptotic normality of $\hat{\theta}_n$, the present techniques of proof rely upon square F_{i0}-integrability of each score $\dot{w}_i(\theta)/w_i(\theta)$ uniformly in θ in a neighborhood of θ_0. This condition is equivalent to $\sup_{\|\theta-\theta_0\|_1<\delta} P_0[\dot{l}_\theta^T(X)\dot{l}_\theta(X)] < \infty$, where $\dot{l}_\theta$ is the score for θ defined in (4.1).

REMARK 5.3. Hypothesis (I) is used to establish invertibility of $\dot{\Psi}_0$. Empirical process theory is used to verify the stochastic, approximation condition of the Master theorem. It is implied if a certain class of functions is Donsker [see Van der Vaart (1995), Lemma 1], and is why the classes $\mathbf{W}_\delta$ and $\dot{\mathbf{W}}$ are assumed to be Donsker. The uniform Lipschitz condition (viii) is used in verifying that classes of differences of scores for weights in a neighborhood of θ_0 are Donsker.

Given asymptotic normality of $\hat{\theta}_n$, asymptotic normality of the joint process $\mathbb{Z}'_n$ can be established.

THEOREM 5.4 (Asymptotic normality of $(\hat{\theta}_n, \mathbb{G}_n)$). *Let* $\mathbf{H}$ *be a collection of functions with envelope* h_e *such that* $\mathbf{F} = \{hr(\theta_0)\colon h \in \mathbf{H}\}$ *is Donsker for each* F_{i0}, $G(h_e^2 r(\theta_0)) = \overline{F}_0(h_e^2 r^2(\theta_0)) < \infty$ *and there exists* $\delta > 0$ *such that* $\sup_{\|\theta-\theta_0\|_1<\delta} G(h_e|\dot{w}_{ik}(\theta)/w_i(\theta)|^2 \tilde{w}_i(\theta_0)) < \infty$ *for each* $i = 1, \ldots, s$, $k = 1, \ldots, d$. *Further assume the hypotheses of Theorem* 5.3.

Then $(\hat{\theta}_n, \mathbb{G}_n)$, *the solution of procedure* (1)–(3), *satisfies*

$$\begin{aligned}\mathbb{Z}'_n(\cdot, \hat{\theta}_n) &= \left(\sqrt{n}(\hat{\theta}_n - \theta_0), \mathbb{Z}_n(\cdot, \hat{\theta}_n)\right)^T \\ &= \frac{1}{\sqrt{n}} \sum_{i=1}^n \left(\tilde{l}_{\theta_0}(X_i), \tilde{l}_G(X_i)\right)^T + o_{P^*}(1) \\ &\Rightarrow Z' \quad \text{in } \Theta \times l^\infty(\mathbf{H}),\end{aligned}$$

where Z' *is a tight mean zero Gaussian process in* $\Theta \times l^\infty(\mathbf{H})$ *with covariance process given by* (5.7).

Conditions for a class $\mathbf{F}$ to be Donsker are well characterized; for example, see Pollard (1984) and Van der Vaart and Wellner (1996). A special case is given in the following corollary, with important application $\mathbf{Y} = R^k$, $h_e = 1$, and $\mathbf{H}$ the collection of all indicator functions of lower-left orthants, or of all rectangles, or of all balls.

COROLLARY 5.3. *Suppose* $\mathbf{H} = \{h_e 1_C\colon C \in \mathbf{C}\}$, *where* $\mathbf{C}$ *is a Vapnik–Chervonenkis class of sets. If* $G(h_e^2 r(\theta_0)) < \infty$ *and there exists* $\delta > 0$ *such*

that $\sup_{\|\theta-\theta_0\|_1<\delta} G(h_e|\dot{w}_{ik}(\theta)/w_i(\theta)|^2\tilde{w}_i(\theta_0)) < \infty$, $i = 1, \ldots, s$, $k = 1, \ldots, d$, *then the result of Theorem* 5.4 *follows from the hypotheses of Theorem* 5.3.

Notice that $\mathbf{H}$ of the corollary equals $\mathbf{H}$ of (5.3), the space over which the semiparametric MLE $\mathbb{G}_n$ was proved to be uniformly consistent. For the distributional result, square G-integrability of the envelope h_e is required, while only G-integrability is needed for consistency.

Since the limiting covariance process coincides with the information bounds for θ and G, with $\text{AsymCov}(Z_{\theta_0}, Z(h))$ equal to covariance expression (4.10), the MLE $(\hat{\theta}_n, \mathbb{G}_n)$ is asymptotically efficient relative to a tangent set $\dot{\mathbf{P}}_0^0$. It is efficient in the following sense: for a given, fixed biased sampling problem as in (1.1)–(1.3) (the number of samples $s \geq 2$, the parametric form for the weight functions $w_1, \ldots, w_s$ and the sampling fractions $\lambda_1, \ldots, \lambda_s$ are all known), among the class of regular estimators, its limiting process has the least dispersion. See Remark 3.1 and Problem 6.1 in GVW for a discussion of designing the biased sampling in order to optimize some criteria.

REMARK 5.4. Note that $G(w_i(\theta)r(\theta)) < \infty$ is automatically satisfied if $W_i(\theta, G) < \infty$ and $\lambda_i > 0$, since $G(w_i(\theta)r(\theta)) = [W_i(\theta, G)/\lambda_i]G(r_i(\theta)) \leq W_i(\theta, G)/\lambda_i$. This, coupled with the hypothesis $G(h_e^2 r(\theta_0)) < \infty$, implies that the union $\mathbf{F}_\delta \equiv \mathbf{F} \cup \mathbf{W}_\delta$ is Donsker because $\mathbf{F}$ and $\mathbf{W}_\delta$ are; see, for example, Van der Vaart and Wellner [(1996), Example 2.10.7]. This fact is used in the proof of Theorem 5.4.

REMARK 5.5 (Estimation of variability). The form of the covariance process (5.7) suggests a plug-in procedure for estimating it. Simply construct the sample analogue of (5.7), with θ_0, G and λ_i replaced by $\hat{\theta}_n$, $\mathbb{G}_n$ and λ_{ni} throughout. This estimator converges in probability and almost surely to (5.7) under natural conditions. This covariance estimator can also be obtained from the full likelihood (1.10) as follows. Form the Hessian of $l_n = \log L_n$ by taking second derivatives with respect to the mass points of G (the p_j) and θ. Invert the Hessian, which is an $(h+d)\times(h+d)$ matrix. For $\underline{c} \in R^d$ and $h \in \mathbf{H}$, form the vector $(c_1, \ldots, c_d, h(u_1), \ldots, h(u_h))$, where u_j, $j = 1, \ldots, h$, is the location of the mass point p_j. Pre- and post-multiply the inverted Hessian to form the estimator of the asymptotic variance of $\underline{c}^T\sqrt{n}(\hat{\theta}_n - \theta_0) + \mathbb{Z}_n(h, \hat{\theta}_n)$. The latter method has been implemented in a computer program, and consistency corroborated in simulations [Gilbert, Lele and Vardi (1999)].

REMARK 5.6 (Bootstrap covariance estimator). We conjecture that the bootstrap covariance estimator is asymptotically first-order correct. We have verified this [Gilbert (1996)] for the case $\mathbf{Y} = R$ by utilizing the fact that bootstrapped Z-estimators are first-order asymptotically correct if they satisfy a mild integrability condition and the hypotheses of Van der Vaart's (1995) Z-theorem [Wellner and Zhan (1998)]. Bootstrap confidence limits for $(\hat{\theta}_n, \mathbb{G}_n)$ are studied in simulations by Gilbert, Lele and Vardi (1999).

REMARK 5.7 (Hypothesis tests of H_0: $\theta_0 = 0$). Gilbert, Lele and Vardi (1999) developed and investigated finite-sample properties of Wald, efficient score and likelihood ratio tests of the hypothesis H_0: $\theta_0 = 0$. It is straightforward to verify that the statistics are asymptotically χ^2_d under H_0 if the hypotheses of Theorem 5.4 hold. In addition, the profile likelihood-based confidence set for θ_0 defined by $\{\theta \in \Theta | 2[l_{n1\,pro}(\hat{\theta}_n|\underline{x}) - l_{n1\,pro}(\theta|\underline{x})] \le \chi^2_{d,\,1-\alpha}\}$ has correct converge probability $1-\alpha$ in the limit.

PROPOSITION 5.2 (Asymptotic normality of $\underline{\mathbb{W}}_n$). *Suppose the hypotheses of Theorem* 5.4 *hold. Then*

$$
\begin{aligned}
\sqrt{n}(\underline{\mathbb{W}}_n - \underline{W}_0) \to_d\ & \mathbb{X}\big(r(\theta_0)(\underline{w}(\theta_0) - \underline{W}_0)\big) \\
& + G\big(r(\theta_0)(\underline{w}(\theta_0) - \underline{W}_0)\underline{\tilde{w}}^T(\theta_0)\big) M^-(\theta_0) \mathbb{X}\big(r(\theta_0)\underline{\tilde{w}}(\theta_0)\big) \\
& + \underline{\underline{W}}_0 \Bigg[G\big(\underline{\tilde{w}}(\theta_0) a^T_\star(\theta_0)\big) \\
& \qquad + G\left(\frac{\dot{w}_1(\theta_0)}{w_1(\theta_0)}\tilde{w}_1(\theta_0), \ldots, \frac{\dot{w}_s(\theta_0)}{w_s(\theta_0)}\tilde{w}_s(\theta_0)\right)^T \Bigg] * Z_{\theta_0} \\
\sim\ & N_s(\underline{0}, \Sigma),
\end{aligned}
\tag{5.13}
$$

where

$$
\begin{aligned}
\Sigma =\ & G\Big(r(\theta_0)\big(\underline{w}(\theta_0) - \underline{W}_0\big)\big(\underline{w}(\theta_0) - \underline{W}_0\big)^T\Big) \\
& + G\big(r(\theta_0)(\underline{w}(\theta_0) - \underline{W}_0)\underline{\tilde{w}}^T(\theta_0)\big) M^-(\theta_0) \\
& \quad \times G\big(r(\theta_0)\underline{\tilde{w}}(\theta_0)(\underline{w}(\theta_0) - \underline{W}_0)^T\big) \\
& + \underline{\underline{W}}_0 \Bigg[G\big(\underline{\tilde{w}}(\theta_0) a^T_\star(\theta_0)\big) \\
& \qquad + G\left(\frac{\dot{w}_1(\theta_0)}{w_1(\theta_0)}\tilde{w}_1(\theta_0), \ldots, \frac{\dot{w}_s(\theta_0)}{w_s(\theta_0)}\tilde{w}_s(\theta_0)\right)^T \Bigg] * I^{-1}(\theta_0) \\
& \quad \times \Bigg[G\big(\underline{\tilde{w}}(\theta_0) a^T_\star(\theta_0)\big) \\
& \qquad + G\left(\frac{\dot{w}_1(\theta_0)}{w_1(\theta_0)}\tilde{w}_1(\theta_0), \ldots, \frac{\dot{w}_s(\theta_0)}{w_s(\theta_0)}\tilde{w}_s(\theta_0)\right)^T \Bigg]^T \underline{\underline{W}}_0.
\end{aligned}
\tag{5.14}
$$

Notice that when w_s is a constant function (say $w_s = c$), the sth row and column of the covariance matrix Σ are filled with zeros. Thus, the above limiting expression captures the degeneracy in the sth weight function. Note that since $\mathbb{V}_{ni} = \mathbb{W}_{ni}/\mathbb{W}_{ns} = c^{-1}\mathbb{W}_{ni}$, it follows automatically that $\sqrt{n}(\underline{\mathbb{V}}_n - \underline{V}_0) \to_d c^{-1} N_s(0, \Sigma)$, where Σ is as in (5.14).

Finally, consider estimation of the biased distributions $F_{10}, \ldots, F_{s0}$. The semiparametric MLE $\hat{\mathbb{F}}_{ni}$ of F_{i0} is given by

$$\hat{\mathbb{F}}_{ni}(h, \hat{\theta}_n) = \mathbb{G}_n\Big(hw_i(\hat{\theta}_n)\Big)/\mathbb{G}_n\Big(w_i(\hat{\theta}_n)\Big)$$

for $h \in \mathbf{H}$, $i = 1, \ldots, s$. Set $\mathbb{Y}_{ni} = \sqrt{n}(\hat{\mathbb{F}}_{ni}(\cdot, \hat{\theta}_n) - F_{i0})$, $i = 1, \ldots, s$, and let $\underline{\mathbb{Y}}_n = (\mathbb{Y}_{n1}, \ldots, \mathbb{Y}_{ns})^T$, which is a vector of processes indexed by a collection of functions $\mathbf{H} \subset L_2(\overline{F}_0)$. Define a vector of Gaussian processes $\underline{Y}$: $\mathbf{H} \to R^s$ by

$$\begin{aligned}\underline{Y}(h) = {} & \mathbb{X}\big(hr(\theta_0)\underline{\tilde{w}}(\theta_0)\big) \\ & - \Big[G\big(h\underline{\tilde{w}}(\theta_0)\big)\underline{\underline{\lambda}}^{-1} - \big(hr(\theta_0)\underline{\tilde{w}}(\theta_0)\underline{\tilde{w}}^T(\theta_0)\big)\Big]M^-(\theta_0)\mathbb{X}\big(r(\theta_0)\underline{\tilde{w}}(\theta_0)\big) \\ & + \left[-G\big(h\underline{\tilde{w}}(\theta_0)a_\star^T(\theta_0)\big) + G\Big(h\Big[\dot{l}_{\theta_0}(i, Y), \ldots, \dot{l}_{\theta_0}(s, Y)\Big]^T\Big)\right] * Z_{\theta_0}\end{aligned} \tag{5.15}$$

PROPOSITION 5.3 (Asymptotic normality of $\underline{\mathbb{Y}}_n$). *Let $\delta > 0$ and $\mathbf{H}$ be a collection of functions such that $\mathbf{F} \equiv \bigcup_{i=1}^s \{hr(\theta_0)w_i(\theta_0)$: $h \in \mathbf{H}\}$ is Donsker and $G(h_e^2 r(\theta)) < \infty$ for all $\|\theta - \theta_0\|_1 < \delta$. Suppose the hypotheses of Theorem* 5.4 *hold. Then $\underline{\mathbb{Y}}_n \Rightarrow \underline{Y}$ in $l^\infty(\mathbf{H})^s$ as $n \to \infty$ with $\lambda_{ni} \to \lambda_i > 0$.*

6. Examples and applications.

EXAMPLE 6.1 (Examples 1.1 and 1.2 continued; generalized logistic regression model). For which GLR models is the MLE $(\hat{\theta}_n, \mathbb{G}_n)$ consistent, asymptotically normal and efficient? For a large class of useful ones, as we now assert. First consider the univariate s-sample model (1.5) with $\mathbf{Y} = R$, $\mathbf{H} = \{1_{[0,t]}$: $t \in [0, \infty)\}$. Take $h_i(y, \theta) = \sum_{k=1}^d h_{ik}(y)\theta_k$ for known nonnegative functions h_{ik} with $h_{ik}(0) = 0$, $i = 1, \ldots, s-1$, and set $h_s \equiv 0$. In this case,

$$\frac{\dot{w}_{ik}(y, \theta)}{w_i(y, \theta)} = h_{ik}(y)$$

and

$$w_i(y, \theta)r(y, \theta) = \frac{\exp\{\sum_{k=1}^d h_{ik}(y)\theta_k\}}{\sum_{l=1}^s \lambda_l\Big[\exp\{\sum_{k=1}^d h_{lk}(y)\theta_k\}/W_l(\theta, G)\Big]}.$$

Choose the functions h_{ik} such that the classes

$$\begin{aligned}&\{h_{ik}\text{: } i = 1, \ldots, s-1, k = 1, \ldots, d\} \quad \text{and} \\ &\{w_i(\cdot, \theta)r(\cdot, \theta)\text{: } \|\theta - \theta_0\|_1 < \delta, i = 1, \ldots, s\}\end{aligned} \tag{6.1}$$

are F_{l0}-Donsker for some $\delta > 0$, for each $l = 1, \ldots, s$. A sufficient condition for this to hold for the first class of (6.1) is that each h_{ik} is monotone; see, for example, Van der Vaart and Wellner [(1996), Theorem 2.7.5], which is a natural condition for the vaccine trial application. A sufficient condition for the second class of (6.1) to be Donsker is that each $w_i r$, or equivalently each

r_i, is Lipschitz $\alpha > 1/2$ in y, uniformly in $\|\theta - \theta_0\|_1 < \delta$, and $\overline{F}_0(|Y|^{2+\delta}) < \infty$ for some $\delta > 0$. Here we are using Van der Vaart's (1994) theorem to verify the hypothesis of Ossiander's (1987) bracketing entropy central limit theorem for classes of smooth functions.

THEOREM 6.1 (Asymptotic properties of the univariate GLR model). *Suppose $n_i \to \infty$ and $\lambda_{ni} \to \lambda_i > 0$, $i = 1, \ldots, s$, there exists an $i \in \{1, \ldots, s-1\}$ such that the set of functions $\{h_{i1}, \ldots, h_{id}\}$ in* (6.1) *is linearly independent and the limiting log profile partial likelihood is strictly concave on Θ. Further suppose there exists a $\delta > 0$ such that, for each $l = 1, \ldots, s$,*

(A) $G(h_{ik}^2 \tilde{w}_i(\theta_0)) < \infty$ *for each $i = 1, \ldots, s-1$, and the classes of* (6.1) *are F_{l0}-Donsker for each $l = 1, \ldots, s$.*

Then the MLE $(\hat{\theta}_n, \mathbb{G}_n)$ of (θ_0, G) in the GLR model exists uniquely, can be computed by the maximum partial likelihood estimation procedure and is consistent uniformly over $\mathbf{H} \equiv \{1_{[0,t]}\colon t \in [0, \infty)\}$, asymptotically Gaussian in $R^d \times l^\infty(\mathbf{H})$ and efficient at $P_0 = P_{(\theta_0, G)}$.

The hypothesis $\lambda_{ni} \to \lambda_i > 0$, $i = 1, \ldots, s$, states that the fraction of trial participants infected during follow-up does not vanish in the limit for any treatment group. For a two-arm vaccine trial, it simply says that the vaccine is not 100% protective. As stated by Theorem 5 in Gilbert, Lele and Vardi (1999), the hypothesis of a strictly concave limiting log profile partial likelihood holds for all two-sample GLR models. Moreover, an easy extension of the proof of this theorem shows strict concavity in the s-sample special case with $h_{1k} = \cdots = h_{s-1k}$, $k = 1, \ldots, d$. Thus we have two corollaries.

COROLLARY 6.1 (Two-sample GLR model). *Suppose $s = 2$, $n_i \to \infty$ with $\lambda_{ni} \to \lambda_i > 0$, $i = 1, 2$ and* (*A*) *holds. Then the conclusion of Theorem* 6.1 *holds.*

COROLLARY 6.2 (Special case s-sample GLR model). *Suppose $s \geq 2$, $n_i \to \infty$ with $\lambda_{ni} \to \lambda_i > 0$, $i = 1, \ldots, s$ and $h_{1k} = \cdots = h_{s-1k}$, $k = 1, \ldots, d$. Then under hypothesis* (*A*), *the conclusion of Theorem* 6.1 *holds.*

Now consider the multivariate GLR model, in which $\mathbf{Y} = R^k$ and $\mathbf{H}$ is the collection of indicators of lower-left orthants. Since there are many possibilities for forms of weight functions, we do not state a theorem. Rather, we note that, for a model of interest, the conditions of Theorems 1.1, 2.1, 5.2 and 5.4 can be checked systematically. Essentially, the multivariate situation offers no added difficulties; identifiability and unique estimability in the limit follow under the same conditions, and although the classes (6.1) will contain more functions, they often have entropy bounds under the same hypotheses as for the univariate case. In conclusion, the large sample properties of the MLE established here hold for a broad collection of GLR models.

For two-sample univariate GLR, we explicitly write the limiting variance-covariance of $(\hat{\theta}_n, \mathbb{G}_n)$. From (4.6), the limiting variance $I^{-1}(\theta_0)$ of $\sqrt{n}(\hat{\theta}_n - \theta_0)$ is

$$\left\{ \frac{\lambda_1 \lambda_2}{\int \exp\{h(\theta_0)\}\, dG} \left[G\Big(\dot{h}^2(\theta_0) * \exp\{h(\theta_0)\} * r(\theta_0)\Big) \right.\right. \tag{6.2}$$
$$\left.\left. - \frac{G\Big[\big(\dot{h}(\theta_0) * \exp\{h(\theta_0)\} * r(\theta_0)\big)\Big]^2}{G\big(\exp\{h(\theta_0)\} * r(\theta_0)\big)} \right] \right\}^{-1},$$

with $\dot{h}(\theta_0) = \frac{\partial}{\partial \theta} h(\cdot, \theta)|_{\theta_0}$ and $r^{-1}(y, \theta_0) = \lambda_1\big(\exp\{h(y, \theta_0)\}/\int \exp\{h(\theta_0)\}\, dG\big) + \lambda_2$. The consistent plug-in estimate of variance is obtained by replacing λ_i and (θ_0, G) in (6.2) by λ_{ni} and the MLE $(\hat{\theta}_n, \mathbb{G}_n)$. The limiting covariance process $\{\text{AsymCov}(\mathbb{G}_n(s), \mathbb{G}_n(t))\colon s, t \in [0, \infty)\}$ and its estimate are obtained by selecting $h_1 = 1_{[0, s]}$, $h_2 = 1_{[0, t]}$, and substituting w_1, $\dot{w}_1$ and r into (4.9).

EXAMPLE 6.2 (Example 1.3 continued; choice-based sampling in econometrics). In the pure choice-based sampling design of Example 1.3, the M weight functions in the biased sampling model are $w_j(y, \theta) = p_\theta(j|y)$, $j = 1, \ldots, M$, $y \in \mathbf{Y}$. In this application, the sample space $\mathbf{Y} \subset R^d$, and $\mathbf{H}$ is the set of indicators of lower-left orthants times the indicators of $\mathbf{Y}$. Since every weight function depends on θ, Theorem 1.1 cannot be used to establish identifiability of the model. However, Theorem 3 of Gilbert, Lele and Vardi (1999) gives a sufficient condition for identifiability in this case. It yields, for instance, that the semiparametric biased sampling model is identifiable when the parametric model p_θ is a baseline category logit model [Cox and Snell (1989); Fienberg (1980)] given by

$$p_\theta(j|y) = \frac{\exp\big(\beta_j^T y\big)}{\sum_{j'=1}^M \exp\big(\beta_{j'}^T y\big)}, \qquad j = 1, \ldots, M,$$

or an adjacent categories linear logit model [Agresti (1984); Fienberg (1980)] given by

$$p_\theta(j|y) = \frac{\exp\big(\alpha_j + j\beta^T y\big)}{\sum_{j'=1}^M \exp\big(\alpha_{j'} + j'\beta^T y\big)}, \qquad j = 1, \ldots, M.$$

When p_θ is the full multinomial logistic regression model (1.7), the biased sampling model can also be shown to be identifiable, arguing as in the proof of Theorem 3 in Gilbert, Lele and Vardi (1999), utilizing the fact that the aggregate shares are known. Moreover, the maximum partial likelihood estimation procedure applies because the aggregate shares are known.

For this problem, $\dot{w}_{jk}(y, \theta)/w_j(y, \theta) = \dot{l}_{\theta_k}(j|y)$, the kth component of the score for the parametric model $p_\theta(j|y)$. Suppose there exists a $\delta > 0$ such that $\sup_{\|\theta - \theta_0\|_1 < \delta} G(|\dot{l}_{\theta_k}(j|Y)|^2 p_{\theta_0}(j|Y)) < \infty$, which usually holds, for instance, if the covariates are assumed bounded. This immediately implies condition (vii) $\sup_{\|\theta - \theta_0\|_1 < \delta} G(|\dot{w}_{kl}(\theta)/w_k(\theta)|^2 \tilde{w}_k(\theta_0)) < \infty$ of Theorem 5.3, and consistency

of the MLE $(\hat{\theta}_n, \mathbb{G}_n)$ also follows if each $\dot{l}_{\theta_k}(j|\cdot)$ is continuous in θ at θ_0 and $l_{1\,pro}(\theta)$ is strictly concave on Θ. As described in Remark 5.1, $l_{1\,pro}(\theta)$ is strictly concave on Θ and (iii) holds if the information $I(\theta)$ for θ in the biased sampling model is positive definite for all $\theta \in \Theta$. For most parametric models p_θ defining the biasing functions, this is the case, which can be directly verified by checking condition (I).

The uniform Lipschitz condition (viii) of Theorem 5.3 takes the form

$$\left|\dot{l}_{\theta_{1k}}(j|y) - \dot{l}_{\theta_{2k}}(j|y)\right| \le K\|\theta_1 - \theta_2\|_1^\alpha$$

for all $\theta_1, \theta_2 \in \{\theta \in \Theta\colon \|\theta - \theta_0\|_1 < \delta\}$. For the case in which p_θ has form (1.7), $\theta = [\alpha_1, \ldots, \alpha_M, \beta_1, \ldots, \beta_M]^T$, and straightforward calculation shows that, for $k = 1, \ldots, M$ (i.e., $\theta_1 = \alpha_1, \ldots, \theta_M = \alpha_M$),

$$\begin{aligned}\left[\dot{l}_{\theta_{1k}}(j|y) - \dot{l}_{\theta_{2k}}(j|y)\right] &= [p_{\theta_2}(k|y) - p_{\theta_1}(k|y)] \\ &= \left[\frac{\exp(\alpha_{2k} + \beta_{2k}^T y)}{\sum_{j'=1}^M \exp(\alpha_{2j'} + \beta_{2j'}^T y)} - \frac{\exp(\alpha_{1k} + \beta_{1k}^T y)}{\sum_{j'=1}^M \exp(\alpha_{1j'} + \beta_{1j'}^T y)}\right]. \qquad (6.3)\end{aligned}$$

For $k = M+1, \ldots, 2M$ (i.e., $\theta_{M+1} = \beta_1, \ldots, \theta_{2M} = \beta_M$),

$$\left[\dot{l}_{\theta_{1k}}(j|y) - \dot{l}_{\theta_{2k}}(j|y)\right] = y[p_{\theta_2}(k-M|y) - p_{\theta_1}(k-M|y)].$$

Thus, the uniform Lipschitz condition will hold if the covariate space $\mathbf{Y}$ is bounded and, for each k, (6.3) is bounded by a constant times $\|\theta_1 - \theta_2\|_1^\alpha$ for some $\alpha > d/2$.

Then asymptotic normality and efficiency of $(\hat{\theta}_n, \mathbb{G}_n)$ follow as long as the two classes

$$\begin{aligned}&\{\dot{l}_{\theta_0 k}(j|\cdot)\colon j = 1, \ldots, M, k = 1, \ldots, d\}, \\ &\left\{\frac{p_\theta(j|\cdot)}{\sum_{j'=1}^M \lambda_{j'}\left[p_\theta(j'|\cdot)/G(p_\theta(j'|\cdot))\right]}\colon \|\theta - \theta_0\|_1 < \delta, j = 1, \ldots, M\right\},\end{aligned}$$

are F_{l0}-Donsker for each $l = 1, \ldots, s$, for some $\delta > 0$. The limiting covariance structure of $(\hat{\theta}_n, \mathbb{G}_n)$ can be explicitly written down by substituting the chosen model p_{θ_0} into the information bound formulae.

7. Proofs.

PROOF OF THEOREM 4.1. The efficient score function for θ in $\mathbf{P}$ can be calculated by identifying a tangent $a' \in L_2^0(G)$ satisfying $E[(\dot{l}_\theta - \dot{l}_g a')(\dot{l}_g a)] = 0$ for all $a \in L_2^0(G)$, and verifying that the tangent space of G, $\dot{\mathbf{P}}_2$, equals $\overline{\mathbf{R}(\dot{l}_g)}$,

the closure of the range of the score for g. Using (4.1) and (4.2), we calculate

$$
\begin{aligned}
0 &= E\Big[\{\dot{l}_\theta(I,Y) - \dot{l}_g a'(I,Y)\}\{\dot{l}_g a(I,Y)\}\Big] \\
&= E\left[\left[\frac{\dot{w}_I(Y,\theta)}{w_I(Y,\theta)} - a'(Y)\right]a(Y)\right] \\
&\quad - E\left[E\left[\frac{\dot{w}_I(Y,\theta)}{w_I(Y,\theta)} - a'(Y)|I\right]E\big[a(Y)|I\big]\right] \qquad (7.1)\\
&= \int\left\{\sum_{i=1}^{s}\lambda_i\left[\frac{\dot{w}_i(y,\theta)}{w_i(y,\theta)} - a'(y)\right.\right. \\
&\qquad \left.\left. - E\left[\frac{\dot{w}_i(Y,\theta)}{w_i(Y,\theta)} - a'(Y)|I=i\right]\right]\tilde{w}_i(y,\theta)\right\}a(y)\,dG(y).
\end{aligned}
$$

Since this equality holds for all $a \in L_2^0(G)$, it follows that the quantity in braces must vanish almost everywhere in $L_2^0(G)$, which leads to a Fredholm integral equation of the second kind:

$$
a'(y) - \int a'(z)K(y,z)\,dG(z) = \phi^\star(y), \qquad (7.2)
$$

with kernel K and function $\phi^\star$ given by $K(y,z) = \sum_{i=1}^{s} r_i(y,\theta)\tilde{w}_i(z,\theta)$ and $\phi^\star(y) = \sum_{i=1}^{s} r_i(y,\theta)\dot{l}_\theta(i,y)$.

We suppress the dependence of a', K and $\phi^\star$ on θ and G. As in Tricomi (1957), since the kernel is Pincherle–Goursat (i.e., it factors in y and z), the Fredholm equation (7.2) can be equivalently expressed as an algebraic system of s equations in s unknowns $A\underline{z} = B$, where $\underline{z} = (z_1,\ldots,z_s)^T$ are unknowns, $A = I_s - (a_{ij}) = I_s - G(\underline{r}\underline{\tilde{w}}^T)$ and $B = [\underline{b}_1^T,\ldots,\underline{b}_s^T]^T$, with

$$
\begin{aligned}
a_{ij} &= \int r_j(y)\tilde{w}_i(y)\,dG(y) = \lambda_j G(r\tilde{w}_i\tilde{w}_j), \\
\underline{b}_j &= \int \phi^\star(y)\tilde{w}_j(y)\,dG(y) = \sum_{i=1}^{s}\lambda_i G\big(\dot{l}_\theta(i,Y)r\tilde{w}_i\tilde{w}_j\big).
\end{aligned} \qquad (7.3)
$$

It is easily seen that $A = M\underline{\underline{\lambda}}$, where M is the matrix defined in (3.1), so that A has rank $s-1$. Letting M^- be a generalized inverse of M, we solve $\underline{z} = \underline{\underline{\lambda}}^{-1}M^-B$, which implies

$$
a'(y) = \phi^\star(y) + B^T(M^-)^T\underline{\underline{\lambda}}^{-1}r = \phi^\star(y) + G\big(\phi^\star\underline{\tilde{w}}^T\big)(M^-)^T r\underline{\tilde{w}}.
$$

Set $\dot{l}_\theta^\star = \dot{l}_\theta - \dot{l}_g a'$, which is unique since $(\dot{l}_g a')(x) = a'(y) - G(a'\tilde{w}_i)$ is independent of the choice of M^-. This follows by Lemma 5.2(iv) of GVW, since $r\underline{\tilde{w}} - G(r\underline{\tilde{w}}\tilde{w}_i) \in \mathrm{Range}(M)$. Since $G(a')$ is not necessarily equal to zero, we center by taking $a_\star = a' - G(a') \in L_2^0(G)$, and set $\dot{l}_\theta^\star = \dot{l}_\theta - \dot{l}_g a_\star = \dot{l}_\theta - \dot{l}_g a'$, which is unique and orthogonal to $\overline{\mathbf{R}(\dot{l}_g)}$. A slight adaptation of Theorem 3 of

BKRW (page 123) proves that if the graph $\mathbf{G}^\star(\theta)$ is connected for all θ in some neighborhood of θ_0, then $\dot{\mathbf{P}}_2 = \{a - E(a|I)\colon a \in L_2^0(G)\} = \overline{\mathbf{R}(\dot{l}_g)}$, so that $\dot{l}_\theta^\star$ is orthogonal to $\dot{\mathbf{P}}_2$ and is therefore the efficient score function.

Nonsingularity of $I(\theta_0)$ is shown in the proof of Theorem 5.3. Straightforward calculation, aided with the identity $\sum_{i=1}^s \lambda_i G(\tilde{w}_i r) = 1$, yields the form (4.6) of the inverse information when $s = 2$.

PROOF OF THEOREM 4.2. Consider the parameter G as a functional of the model: $\nu(P_{(\theta,G)}) = \chi(G) = G$. Let $\dot{\chi}$ denote the pathwise derivative of χ, and $\dot{\chi}^T$ its adjoint. Let $(\dot{l}_g^T \dot{l}_g)^\star$ be the information bound operator for estimation of G, viewed as a map from $L_2(G)$ and $L_2(G)$. We apply Corollary 3 of Begun, Hall, Huang and Wellner (1983), proved in BKRW (pages 218–220), which states that if ν is pathwise differentiable at P_0, $\dot{\mathbf{P}} = \overline{\mathbf{R}(\dot{l}_g)} + \overline{\mathbf{R}(\dot{l}_\theta)}$, $\mathbf{R}(\dot{\chi}^T) \subset \mathbf{R}(\dot{l}_g^T \dot{l}_g)$, and $(\dot{l}_g^T \dot{l}_g)^{\star -1}$ and $I^{-1}(\theta)$ exist, then the inverse information covariance functional for ν equals

$$
\begin{aligned}
I_\nu^{-1}(h_1, h_2) = {} & E\Big[\dot{\chi}^T \pi_{h_1} (\dot{l}_g^T \dot{l}_g)^{\star -1} \dot{\chi}^T \pi_{h_2}\Big] \\
& + E\big[\dot{\chi}^T \pi_{h_1} a_\star\big]^T I^{-1}(\theta) E\big[\dot{\chi}^T \pi_{h_2} a_\star\big],
\end{aligned} \tag{7.4}
$$

where $a_\star$ is the unique tangent in $L_2^0(G)$ satisfying $E[(\dot{l}_\theta - \dot{l}_g a_\star)(\dot{l}_g a)] = 0$ for all $a \in L_2^0(G)$.

Since χ is just the identity, it is straightforward to show [see BKRW, Example 1, page 191 and BKRW, panel (5), page 58] that $\dot{\chi}^T \pi_h(y) = h(y) - G(h)$ for a projection map π_h, $h \in \mathbf{H}$. The operator $(\dot{l}_g^T \dot{l}_g)^{\star -1}$ is calculated by GVW (pages 1089–1091), and requires that $r(y, \theta_0)$ and $r^{-1}(y, \theta_0)$ are bounded on $\mathbf{Y}$. Substituting it and $\dot{\chi}^T \pi_h(y) = h(y) - G(h)$ into (7.4) and calculation shows that $I_\nu^{-1}(h_1, h_2)$ equals expression (4.8) with $a_\star$ as in the proof of Theorem 4.1. □

PROOF OF THEOREM 5.1. For fixed θ, let $\underline{\mathbb{V}}_n(\theta) = (\mathbb{V}_{n1}(\theta), \ldots, \mathbb{V}_{ns-1}(\theta), 1)^T$ be the solution of the system (1.12). Since θ is fixed, $\underline{\mathbb{V}}_n(\theta) \to_{\text{a.s.}} \underline{V}(\theta)$ by Proposition 2.1 of GVW [page 1081]. Since each $\lambda_{ni}^{-1} r_{ni}(\cdot, \theta)$ is a bounded function of y, if the processes

$$
\left\{ C_{ni}(v) \equiv \frac{1}{n_i} \sum_{j=1}^{n_i} \log\left\{ \frac{w_{ij}(\theta) v_i^{-1}}{\sum_{k=1}^s \lambda_{nk} w_{kj}(\theta) v_k^{-1}} \right\} \colon v = (v_1, \ldots, v_s)^T \in V \right\},
$$

$$
V \equiv [0, \infty)^{s-1} \times 1 \tag{7.5}
$$

converge uniformly on V, then the log profile partial likelihood $l_{n1\,pro}$ satisfies, for each fixed $\theta \in \Theta$

$$
\frac{1}{n} l_{n1\,pro}(\theta|\underline{x}) \to_{\text{a.s.}} \sum_{i=1}^s \lambda_i \int \log\left\{ \frac{w_i(y, \theta) V_i^{-1}(\theta)}{\sum_{k=1}^s \lambda_k w_k(y, \theta) V_k^{-1}(\theta)} \right\} \tilde{w}_i(y, \theta_0)\, dG(y)
$$

$$= l_{1\,pro}(\theta).$$

Since $l_{1\,pro}(\theta)$ and $l_{n1\,pro}(\theta|\underline{x})$ for $n > N$ are strictly concave on Θ, this implies $\hat{\theta}_n \to_{\text{a.s.}} \theta_0$ as $n \to \infty$, applying Theorem II.1 in Appendix II of Andersen and Gill (1982).

Uniform convergence of the processes in (7.5) can be established by the bracketing entropy Glivenko–Cantelli Theorem 2.4.1 of Van der Vaart and Wellner (1996). □

To prove consistency of $\mathbb{G}_n$, some additional notation is needed. Where possible, it matches the notation used by GVW. Define sets of $s-1$ equations, which correspond to (1.12), by

$$1 = \mathbb{H}_{ni}(\mathbb{V}_1, \ldots, \mathbb{V}_{s-1}, 1, \theta), \qquad 1 = H_{ni}(V_1, \ldots, V_{s-1}, 1, \theta) \tag{7.6}$$

for $i = 1, \ldots, s-1$, $\theta \in \Theta$, where

$$\mathbb{H}_{ni}(u_1, \ldots, u_s, \theta) \equiv \frac{1}{u_i} \int \frac{w_i(y, \theta)}{\sum_{j=1}^{s} (\lambda_{nj} w_j(y, \theta))/u_j} \, d\mathbb{F}_n(y)$$

and H_{ni} equals $\mathbb{H}_{ni}$ with $\mathbb{F}_n$ replaced by $\overline{F}_{n0}$. Reparametrize by setting $z_j = \log\{\lambda_{nj}/u_j\}$, and define $\mathbb{K}_{ni}(z_1, \ldots, z_s, \theta) = \mathbb{H}_{ni}(\lambda_{n1}e^{-z_1}, \ldots, \lambda_{ns}e^{-z_s}, \theta) - \lambda_{ni}$ and $K_{ni}(z_1, \ldots, z_s, \theta) = H_{ni}(\lambda_{n1}e^{-z_1}, \ldots, \lambda_{ns}e^{-z_s}, \theta) - \lambda_{ni}$. Then (7.6) become

$$0 = \mathbb{K}_{ni}(\mathbb{Z}_1, \ldots, \mathbb{Z}_{s-1}, \log\{\lambda_{ns}\}\theta), \qquad 0 = K_{ni}(Z_1, \ldots, Z_{s-1}, \log\{\lambda_{ns}\}, \theta)$$

for $i = 1, \ldots, s-1$, $\theta \in \Theta$. Set $\underline{\mathbb{H}}_n = (\mathbb{H}_{n1}, \ldots, \mathbb{H}_{ns})^T$, and similarly defined $\underline{H}_n$, $\underline{\mathbb{K}}_n$ and $\underline{K}_n$. Observe that $\underline{\mathbb{K}}_n(z, \theta) = \nabla \mathbb{D}_n(\underline{z}, \theta)$, where $\mathbb{D}_n$: $R^s \times \Theta \to R^1$ is given by

$$\mathbb{D}_n(\underline{z}, \theta) = \int \log\left[\sum_{i=1}^{s} e^{z_i} w_i(y, \theta)\right] d\mathbb{F}_n(y) - \sum_{i=1}^{s} \lambda_{ni} z_i.$$

Similarly, we can write $\underline{K}_n(z, \theta) = \nabla D_n(\underline{z}, \theta)$, where D_n equals $\mathbb{D}_n$ with $\mathbb{F}_n$ replaced by $\overline{F}_{n0}$.

As proved in GVW (pages 1079–1080), if the graph $\mathbf{G}^\star(\theta)$ is connected, then $\mathbb{D}_n(\underline{z}, \theta)$ and $D_n(\underline{z}, \theta)$ are convex functions of $\underline{z}$ (θ fixed). Thus the reparametrization allows us to use convexity theory to conclude consistency.

For an s-vector $\underline{u} \in R^s$, define

$$\underline{\tilde{w}}(u) \equiv \operatorname{diag}(u^{-1})\underline{w}, \qquad r(\underline{u}) \equiv (\underline{\lambda}^T \underline{\tilde{w}}(\underline{u}))^{-1},$$
$$r_n(\underline{u}) \equiv (\underline{\lambda}_n^T \underline{\tilde{w}}(\underline{u}))^{-1}, \qquad r_{ni}(\underline{u}) \equiv (\lambda_{ni} \tilde{w}_i r_n(\underline{u})).$$

Therefore, $\underline{\tilde{w}}$, r, r_n and r_{ni} are given by

$$\underline{\tilde{w}} = \underline{\tilde{w}}(\underline{W}) = \operatorname{diag}(\underline{W}^{-1})\underline{w}, \qquad r \equiv (\underline{\lambda}^T \underline{\tilde{w}})^{-1},$$
$$r_n \equiv (\underline{\lambda}_n^T \underline{\tilde{w}})^{-1}, \qquad r_{ni} \equiv (\lambda_{ni} \tilde{w}_i r_n).$$

We can rewrite the functions $\underline{\mathbb{H}}_n$, $\underline{H}_n$, $\underline{\mathbb{D}}_n$ and D_n in this notation:

$$\underline{\mathbb{H}}_n(\underline{u}, \theta) = \int r_n(\underline{u}, \theta)\underline{\tilde{w}}(\underline{u}, \theta)\, d\mathbb{F}_n,$$

$$\underline{H}_n(\underline{u}, \theta) = \int r_n(\underline{u}, \theta)\underline{\tilde{w}}(\underline{u}, \theta)\, d\overline{F}_{n0},$$

$$\mathbb{D}_n(\underline{z}, \theta) = \int \log\left[\sum_{i=1}^{s} e^{z_i} w_i(\theta)\right] d\mathbb{F}_n - \sum_{i=1}^{s} \lambda_{ni} z_i,$$

$$D_n(\underline{z}, \theta) = \int \log\left[\sum_{i=1}^{s} e^{z_i} w_i(\theta)\right] d\overline{F}_{n0} - \sum_{i=1}^{s} \lambda_{ni} z_i.$$

Introduce functions $\underline{H}$ and $\underline{D}$ equal to $\underline{H}_n$ and $\underline{D}_n$, respectively, with λ_{ni}, r_n and $\overline{F}_{n0}$ replaced with λ_i, r and $\overline{F}_0$. Finally, define $\tilde{\mathbb{D}}_n(\underline{z}, \theta) = \mathbb{D}_n(\underline{z}, \theta) - \mathbb{D}_n(\underline{Z}, \theta)$ and $\tilde{D}(\underline{z}) = D(\underline{z}) - D(\underline{Z})$, where $\underline{Z} \equiv (Z_1, \ldots, Z_s)^T$, $Z_i \equiv \log\{\lambda_i / V_{i0}\}$. Since $\mathbb{D}_n(\underline{z}, \theta)$ and $D(\underline{z}, \theta)$ are convex functions of $\underline{z}$ for each $\theta \in \Theta$, so are $\tilde{\mathbb{D}}_n$ and $\tilde{D}$.

Using this notation, we establish consistency, which depends on the uniqueness of $\underline{\mathbb{V}}_n(\hat{\theta}_n)$ and $\underline{V}(\theta_0)$ as solutions of the systems of equations (7.6), together with convexity of $\tilde{\mathbb{D}}_n(\cdot, \hat{\theta}_n)$ and $\tilde{D}(\cdot, \theta_0)$.

The following lemma generalizes Lemma 5.3 in GVW, allowing the function $\tilde{\mathbb{D}}_n$ to depend on θ.

LEMMA 7.1. *Assume hypotheses* (*ii*) *of Theorem* 5.1 *and* (*iv*) *of Proposition* 5.1. *Then*

$$\tilde{\mathbb{D}}_n(\underline{z}, \hat{\theta}_n) \to_{\text{a.s.}} \tilde{D}(\underline{z}, \theta_0) \tag{7.7}$$

as $n \to \infty$ *with* $\lambda_{ni} \to \lambda_i > 0$ *for each fixed* $\underline{z} \in R^s$. *Since* $\tilde{\mathbb{D}}_n$ *is convex, it also holds that*

$$\sup_{\underline{z} \in C} \left|\tilde{\mathbb{D}}_n(\underline{z}, \hat{\theta}_n) - \tilde{D}(\underline{z}, \theta_0)\right| \to_{\text{a.s.}} 0 \tag{7.8}$$

as $n \to \infty$ *with* $\lambda_{ni} \to \lambda_i > 0$ *for any compact subset* $C \subset R^s$.

PROOF OF LEMMA 7.1. Define

$$q(\underline{z}, y, \theta) \equiv \log\left\{\frac{\sum_{i=1}^{s} e^{z_i} w_i(y, \theta)}{\sum_{i=1}^{s} e^{Z_i} w_i(y, \theta)}\right\}.$$

Observe that $q(\underline{z}, y, \theta)$ is a bounded function of y and θ. Let Θ_0 be a neighborhood of θ_0 for which (iv) of Proposition 5.1 holds. Towards showing (7.7), we have

$$\tilde{\mathbb{D}}_n(\underline{z}, \hat{\theta}_n) = \sum_{j=1}^{s} \lambda_{nj} \mathbb{F}_{nj}\left[q(\underline{z}, Y, \hat{\theta}_n)\right] - \sum_{i=1}^{s} \lambda_{ni}(\underline{z}_i - \underline{Z}_i).$$

Expand q about θ_0, giving, for fixed $\underline{z}$,

$$\mathbb{F}_{nj}\Big[q(\underline{z}, Y, \hat{\theta}_n)\Big] = \mathbb{F}_{nj}\big[q(\underline{z}, Y, \theta_0)\big] + (\hat{\theta}_n - \theta_0)^T * \mathbb{F}_{nj}\left[\frac{\sum_{i=1}^s e^{z_i}\dot{w}_i(Y, \theta^\star)}{\sum_{i=1}^s e^{z_i} w_i(Y, \theta^\star)} - \frac{\sum_{i=1}^s e^{Z_i}\dot{w}_i(Y, \theta^\star)}{\sum_{i=1}^s e^{Z_i} w_i(Y, \theta^\star)}\right]$$

for a $\theta^\star$ between θ_0 and $\hat{\theta}_n$. The first piece converges almost surely to $\int q(\underline{z}, y, \theta_0)\tilde{w}_j(y, \theta_0)\, dG(y)$ by the strong law of large numbers (as in Lemma 5.3 of GVW), while the second piece is $o_p(1)$ because $\hat{\theta}_n - \theta_0 \to_p \underline{0}$ and the $\mathbb{F}_{nj}[\cdot]$ term is $O_p(1)$ for each $j = 1, \ldots, s$ and $k = 1, \ldots, d$. Here continuity of each w_i and $\dot{w}_i/w_i$ in θ at θ_0 and hypothesis (iv) of Proposition 5.1 are used. Thus (7.7) holds.

Since $\tilde{\mathbb{D}}_n$ is convex, (7.8) follows from (7.7), using Theorem II.1 in Appendix II of Andersen and Gill (1982) [also see Exercise 3.2.4 of Van der Vaart and Wellner (1996)]. □

PROOF OF PROPOSITION 5.1. Our proof closely follows the proof of Proposition 2.1 in GVW. Choose N so that, for $n > N$, $\hat{\theta}_n$ is in a neighborhood Θ_0 of θ_0 for which (iv) of Proposition 5.1 holds. Fix $n > N$. Since the graph $\mathbf{G}^\star(\hat{\theta}_n)$ is connected, Lemma 1 of Vardi [(1985), page 197] implies that $\underline{V}(\hat{\theta}_n)$ is the unique solution of the right equation system in (7.6). Therefore, $\underline{Z}_n = (Z_{n1}, \ldots, Z_{ns})^T$, with $Z_{ni} = \log\{\lambda_i/V_i(\hat{\theta}_n)\}$, is the unique minimizer of $D(\underline{z}, \hat{\theta}_n)$, and thus also of $\tilde{D}(\underline{z}, \hat{\theta}_n)$, subject to $z_s = \log\{\lambda_s\}$. Let $\underline{\mathbb{Z}}_n$ be the corresponding minimizer of $D_n(\underline{z}, \hat{\theta}_n)$, which also minimizes $\tilde{D}_n(\underline{z}, \hat{\theta}_n)$ subject to $z_s = \log\{\lambda_{ns}\}$. Also let $\underline{Z}$ be the unique minimizer of $\tilde{D}(\underline{z}, \theta_0)$. Then for any compact set $C \in R^s$ with $\underline{Z}$ in the interior of C, it follows from (7.7) and (7.8) of Lemma 7.1, and the definition of $\tilde{D}(\underline{z}, \theta)$ and $\underline{Z}$, that $\inf_{\underline{z}\in\partial C} \tilde{\mathbb{D}}_n(\underline{z}, \hat{\theta}_n) \to_{\text{a.s.}} \inf_{\underline{z}\in\partial C} \tilde{D}_n(\underline{z}, \theta_0) > 0$, while $\tilde{\mathbb{D}}_n(\underline{\mathbb{Z}}_n, \hat{\theta}_n) \to_{\text{a.s.}} \tilde{D}(\underline{Z}, \theta_0) = 0$. The convexity of $\tilde{\mathbb{D}}_n(\underline{z}, \theta)$ in $\underline{z}$ for all $\theta \in \Theta_0$ implies that $\underline{\mathbb{Z}}_n \in C$ for all n greater than some N with probability 1. Since C can be made arbitrarily small, a simple $\varepsilon - \delta$ argument yields $\underline{\mathbb{Z}}_n \to_{\text{a.s.}} \underline{Z}$. Since $\underline{\mathbb{Z}}_n = (\mathbb{Z}_{n1}, \ldots, \mathbb{Z}_{ns})^T$ with $\mathbb{Z}_{ni} = \log\{\lambda_{ni}/\mathbb{V}_{ni}(\hat{\theta}_n)\}$, where $\underline{\mathbb{V}}_n(\hat{\theta}_n) = (\mathbb{V}_{n1}(\hat{\theta}_n), \ldots, \mathbb{V}_{ns-1}(\hat{\theta}_n), 1)^T$ is the solution of the left equation system in (7.6), it follows that $\underline{\mathbb{V}}_n(\hat{\theta}_n) \to_{\text{a.s.}} \underline{V}(\theta_0)$. The proof that $\underline{\mathbb{W}}_n \to_{\text{a.s.}} \underline{W}(\theta_0)$ is similar and is omitted. □

PROOF OF THEOREM 5.2. This proof closely follows the proof of Theorem 2.1 in GVW. Fix $h \in \mathbf{H}$, with $\mathbf{H}$ defined in (5.3), which satisfies $G(h) < \infty$. Write

$$\begin{aligned}(7.9)\qquad &\sup_{h\in\mathbf{H}}\big|\mathbb{F}_n(hr_n(\underline{\mathbb{V}}_n(\hat{\theta}_n))) - \overline{F}_0(hr(\theta_0))\big| \\ &\quad\le \sup_{h\in\mathbf{H}}\Big|\mathbb{F}_n\Big(h\Big[r_n(\underline{\mathbb{V}}_n(\hat{\theta}_n)) - r(\hat{\theta}_n)\Big]\Big)\Big| \\ &\qquad + \sup_{h\in\mathbf{H}}\big|\mathbb{F}_n(hr(\hat{\theta}_n)) - \overline{F}_0(hr(\theta_0))\big|\end{aligned}$$

$$\leq \left\| \frac{r_n(\underline{\mathbb{V}}_n(\hat{\theta}_n))}{r(\hat{\theta}_n)} - 1 \right\|_\infty |\mathbb{F}_n(h_e r(\hat{\theta}_n))| + \sup_{h\in\mathbf{H}} \left| \mathbb{F}_n(hr(\hat{\theta}_n)) - \overline{F}_0(hr(\theta_0)) \right|. \tag{7.10}$$

Since $\hat{\theta}_n \to_{\text{a.s.}} \theta_0$ and $\underline{\mathbb{V}}_n(\hat{\theta}_n) \to_{\text{a.s.}} \underline{V}(\theta_0)$ by Theorem 5.1 and Proposition 5.1, and $r_n(\underline{u})/r$ is continuous and bounded in y and θ, the $\|\cdot\|_\infty$ term converges almost surely to zero. The term $|\mathbb{F}_n(h_e r(\hat{\theta}_n))|$ is stochastically bounded. Next, we show that

$$\sup_{h\in\mathbf{H}} \left| \mathbb{F}_n(hr_n(\hat{\theta}_n)) - \overline{F}_0(hr(\theta_0)) \right| \to_{\text{a.s.}} 0. \tag{7.11}$$

Expanding about θ_0 by differentiating $r_n(y,\theta)$ with respect to θ, (7.11) is bounded above by

$$\sup_{h\in\mathbf{H}} \left| \mathbb{F}_n(hr_n(\theta_0)) - \overline{F}_0(hr(\theta_0)) \right| \tag{7.12}$$

$$+ \sum_{i=1}^{s} \lambda_{ni} \sup_{h\in\mathbf{H}} \left| (\hat{\theta}_n - \theta_0)^T * \int h(y) \Bigg(- \sum_{k=1}^{s} \lambda_{nk} \Bigg[\frac{\dot{w}_k(y,\theta^\star)}{w_k(y,\theta^\star)} - \int \frac{\dot{w}_k(z,\theta^\star)}{w_k(z,\theta^\star)} \tilde{w}_k(z,\theta^\star,\mathbb{F}_n)\, d\mathbb{F}_n(z) \Bigg] \times \tilde{w}_k(y,\theta^\star,\mathbb{F}_n) r_n^2(y,\theta^\star) \Bigg) d\mathbb{F}_n(y) \right|, \tag{7.13}$$

where $\tilde{w}_k(z,\theta,\mathbb{F}_n) \equiv w_k(z,\theta)/W_k(\theta,\mathbb{F}_n)$. Exactly as in GVW, piece (7.12) converges to zero almost surely by Pollard's Glivenko–Cantelli theorem [see Dudley (1984), Theorems 11.1.2 and 11.1.6] for $\mathbf{H}$ as in (5.3). By (iv), (vi) and (3.2), and since $h \in L_1(G)$, each supremum in (7.13) is $o_p(1)$, so that (7.11) holds. It follows from (7.11) and (7.10) that piece (7.9) converges almost surely to zero. Then

$$\begin{aligned} |\mathbb{G}_n(h) - G(h)| &= \left| \frac{\mathbb{F}_n(hr_n(\underline{\mathbb{V}}_n(\hat{\theta}_n)))}{\mathbb{F}_n(r_n(\underline{\mathbb{V}}_n(\hat{\theta}_n)))} - \frac{\overline{F}_0(hr(\theta_0))}{\overline{F}_0(r(\theta_0))} \right| \\ &\leq \frac{|\mathbb{F}_n(hr_n(\underline{\mathbb{V}}_n(\hat{\theta}_n))) - \overline{F}_0(hr(\theta_0))|}{\mathbb{F}_n(r_n(\underline{\mathbb{V}}_n(\hat{\theta}_n)))} \\ &\quad + \frac{\overline{F}_0(hr(\theta_0))}{\overline{F}_0(r(\theta_0))} \frac{|\mathbb{F}_n(r_n(\underline{\mathbb{V}}_n(\hat{\theta}_n))) - \overline{F}_0(r(\theta_0))|}{\mathbb{F}_n(r_n(\underline{\mathbb{V}}_n(\hat{\theta}_n)))}, \end{aligned}$$

so (5.4) follows from (7.11) with $h = 1$. □

PROOF OF THEOREM 5.3. We apply the Master theorem; see, for example, BKRW (Theorem 1, page 312). First we calculate the matrix $\dot{\Psi}_0$. Differentia-

tion of $P_0\psi_0(X)$ gives

$$\dot{\Psi}_0 = P_0\nabla_\theta\left[\frac{\dot{w}_I(Y,\theta)}{w_I(Y,\theta)} - \phi(Y,\theta)\right]\Big|_{\theta_0} + \nabla_\theta \underline{\underline{d}}(\theta)\big|_{\theta_0} P_0[-\underline{e}_I + \underline{r}(Y,\theta_0)] + \underline{\underline{d}}(\theta_0)P_0\nabla_\theta[\underline{r}(Y,\theta)]|_{\theta_0}. \tag{7.14}$$

The term $P_0[-\underline{e}_I + \underline{r}(Y,\theta_0)] = 0$, so that it is not necessary to compute $\nabla_\theta \underline{\underline{d}}(\theta)|_{\theta_0}$. Direct differentiation using the chain rule and the property $M^-MM^- = M^-$ of a $\{1,2\}$-generalized inverse shows that the rightmost term of (7.14) is also the zero-vector. Thus $\dot{\Psi}_0$ equals the first term on the right-hand side of (7.14). Then, differentiation (again using the chain rule) and tedious calculation (which can be readily checked for the case $s = 2$) shows that $\dot{\Psi}_0$ equals $-I(\theta_0)$ as displayed in (4.4).

Next we show invertibility of $\dot{\Psi}_0 = -I(\theta_0)$. Consider $I(\theta_0)$ expressed as an expected conditional variance as in Remark 4.3. For $b \in R^d$, evidently $b^T I(\theta_0)b \geq 0$, with equality if and only if

$$\mathrm{Var}\left(b^T\left[\frac{\dot{w}_I(Y,\theta_0)}{w_I(Y,\theta_0)} - a_\star(Y,\theta_0)\right]\middle| I = i\right) = 0$$

for each $i = 1,\ldots,s$, which holds if and only if $b^T[(\dot{w}_i(y,\theta_0)/w_i(y,\theta_0)) - a_\star(y,\theta_0)]$ is constant a.s. with respect to the law F_{i0}, for each $i = 1,\ldots,s$. Thus (I) implies $\dot{\Psi}_0$ and $I(\theta_0)$ are nonsingular.

Direct calculation shows that $P_0[\psi_{\theta_0}(X)\psi_{\theta_0}(X)^T]$, where the influence function ψ_{θ_0} is as defined in (5.11), equals $I(\theta_0)$. Therefore, $\Sigma = (-I^{-1}(\theta_0))I(\theta_0)(-I^{-1}(\theta_0))^T = I^{-1}(\theta_0)$.

Next we verify (5.12), that is, that the linearization condition of the Master theorem [(GM2) of BKRW Theorem 1, page 312] holds. Writing $\mathbb{P}_n[\psi_n(X,\theta,\underline{\mathbb{V}}_n(\theta,\mathbb{P}_n)) - \psi(X,\theta,\underline{V}(\theta,P))]$ as $(\mathbb{P}_n - P)[\cdot] + P[\cdot]$, the $(\mathbb{P}_n - P)[\cdot]$ term can straightforwardly be shown to be $o_p(n^{-1/2})$, so it suffices to show that

$$P\Big[\psi_n\big(X,\theta,\underline{\mathbb{V}}_n(\theta,\mathbb{P}_n)\big) - \psi\big(X,\theta,\underline{V}(\theta,P)\big)\Big]$$

$$= P\Big[\psi_n\big(X,\theta,\underline{\mathbb{V}}_n(\theta,\mathbb{P}_n)\big) - \psi_n\big(X,\theta,\underline{V}(\theta,P)\big)\Big] \tag{7.15}$$

$$+P\Big[\psi_n\big(X,\theta,\underline{V}(\theta,P)\big) - \psi\big(X,\theta,\underline{V}(\theta,P)\big)\Big] = o_p(n^{-1/2}), \tag{7.16}$$

where $\psi_n(X,\theta,\underline{V}(\theta,P))$ equals $\psi(X,\theta,\underline{V}(\theta,P))$ with the λ_i's replaced with λ_{ni}'s. To verify that (7.15) is $o_p(n^{-1/2})$, we show that $P[\psi_n(X,\theta,\underline{V}(\theta,P))]$ is Fréchet differentiable. Throughout the proof, the λ_i's in all expressions are implicitly assumed to be λ_{ni}'s. With G^P identified with P and G^Q identified with Q, direct calculation shows that the Gateaux derivative $\partial/\partial\alpha[(1-\alpha)P + \alpha Q][\psi_n(X,\theta,\underline{V}(\theta,(1-\alpha)P+\alpha Q))]|_{\alpha=0}$ of $P[\psi_n(X,\theta,\underline{V}(\theta,P))]$ equals $(Q -$

$P)\psi_n(X, \theta, \underline{V}(\theta, P))$ plus

$$G^P\left[\sum_{i=1}^{s}\lambda_{ni}\frac{\dot{w}_i(\theta)}{w_i(\theta)}\tilde{w}_i(\theta)\left[d_i-\sum_{k=1}^{s}r_k(\theta)d_k\right]\right]$$

$$-A(\theta)M^-(\theta)G^P\left\{\left[d_1-\sum_{k=1}^{s}r_k(\theta)d_k\right]\tilde{w}_1(\theta),\ldots,\right.$$

$$\left.\left[d_s-\sum_{k=1}^{s}r_k(\theta)d_k\right]\tilde{w}_s(\theta)\right\}^T, \tag{7.17}$$

where $d_i \equiv d/d\alpha V_i(\theta,(1-\alpha)P+\alpha Q)|_{\alpha=0} * V_i^{-1}(\theta, P)$. The d_i are found by solving the system of equations

$$\frac{\partial}{\partial\alpha}\left[\sum_{l=1}^{s}\lambda_{nl}\big((1-\alpha)G^P+\alpha G^Q\big)\left[\left(\frac{w_i(Y,\theta)V_i^{-1}\big(\theta,(1-\alpha)G^P+\alpha G^Q\big)}{\sum_{k=1}^{s}\lambda_{nk}w_k(Y,\theta)V_k^{-1}\big(\theta,(1-\alpha)G^P+\alpha G^Q\big)}\right)\right.\right.$$

$$\left.\left.\times\ \tilde{w}_l\big(Y,\theta,(1-\alpha)G^P+\alpha G^Q\big)\right]\right]\Bigg|_{\alpha=0}=0,$$

$$i=1,\ldots,s-1,$$

with solution $d_i = D_{[i,]}(G^Q-G^P)(\underline{\tilde{w}}(\theta, P))$, where $D-\underline{\underline{\lambda}}_n^{-1}M^-(\theta, P)M(\theta, P)\underline{\underline{\lambda}}_n$ with $[i,]$ denoting the ith row of the matrix. Substituting the d_i into (7.17) yields that the Gateaux derivative equals

$$(Q-P)\psi_n(X,\theta,\underline{V}(\theta,P))$$

$$+G_P\left[\sum_{i=1}^{s}\left(\lambda_{ni}\frac{\dot{w}_i(\theta)}{w_i(\theta)}-\big[A(\theta)M^{-1}(\theta)\big]_{[,i]}\right)\right.$$

$$\left.\times\,\tilde{w}_i(\theta)\left[D_{[i,]}-\sum_{k=1}^{s}r_k(\theta)D_{[k,]}\right]\right] \tag{7.18}$$

$$\times\,(G^O-G^P)\big(\underline{\tilde{w}}(\theta,P)\big).$$

A key ingredient in this calculation is that $P[-\underline{e}_I+\underline{r}(\theta)]=0$, which implies the complicated term $d/d\alpha(A(\theta,(1-\alpha)G^P+\alpha G^Q)M^-(\theta,(1-\alpha)G^P+\alpha G^Q))|_{\alpha=0}$ does not appear in the derivative. Now,

$$G_P\left[\sum_{i=1}^{s}\left(\lambda_{ni}\frac{\dot{w}_i(\theta)}{w_i(\theta)}-\big[A(\theta)M^-(\theta)\big]_{[,i]}\right)\right.$$

$$\left.\times\,\tilde{w}_i(\theta)\left[D_{[i,]}-\sum_{k=1}^{s}r_k(\theta)D_{[k,]}\right]\right]=0, \tag{7.19}$$

as can easily be directly verified (by calculating $A(\theta)$ and $M^-(\theta)$ using (4.7)) for the case $s=2$. Thus, the Gateaux derivative equals $(Q-P)\psi_n(X,\theta,\underline{V}(\theta,P))$.

To prove (7.15) is $o_p(n^{-1/2})$, it is then enough to show that the Gateaux derivative is also a Fréchet derivative with respect to the norm $\|Q-P\| \equiv \sup\{\sum_{i=1}^{s} | \int 1_A(y)\tilde{w}_i(y,\theta_0)\,d(G^Q - G^P)(y)|\colon A \in \mathbf{B}\}$. Thus, it suffices to show that

$$\begin{aligned}\Big|Q\Big[\psi_n\big(X,\theta,\underline{V}(\theta,Q)\big)\Big] - P\Big[\psi_n\big(X,\theta,\underline{V}(\theta,P)\big)\Big] \\ -(Q-P)\Big[\psi_n\big(X,\theta,\underline{V}(\theta,P)\big)\Big]\Big| = o(\|Q-P\|).\end{aligned} \tag{7.20}$$

The expression on the left-hand side of (7.20) equals $|Q[\psi_n(X,\theta,\underline{V}(\theta,Q)) - \psi_n(X,\theta,\underline{V}(\theta,P))]|$, and

$$\begin{aligned}&\frac{\Big|Q\Big[\psi_n\big(X,\theta,\underline{V}(\theta,Q)\big) - \psi_n\big(X,\theta,\underline{V}(\theta,P)\big)\Big]\Big|}{\|Q-P\|} \\ &\quad= \frac{\Big|Q\Big[\frac{\partial}{\partial\alpha}\psi_n\big(X,\theta,\underline{V}\big(\theta,(1-\alpha)P+\alpha Q\big)\big)\Big|_{\alpha=0}\Big]\Big|}{\|Q-P\|} * O_p(1) \\ &\quad= \Bigg|G_Q\Bigg[\sum_{i=1}^{s}\Bigg(\lambda_{ni}\frac{\dot{w}_i(\theta)}{w_i(\theta)} - \big[A(\theta)M^{-}(\theta)\big]_{[\,,i]}\Bigg) \\ &\qquad\qquad\times \tilde{w}_i(\theta)\Bigg[D_{[i,\,]} - \sum_{k=1}^{s} r_k(\theta)D_{[k,\,]}\Bigg]\Bigg]\Bigg| \\ &\qquad\times \frac{\big|(G^Q - G^P)\big(\underline{\tilde{w}}(\theta,P)\big)\big|}{\|Q-P\|} * O_p(1) \\ &\quad= o_p(1) * O_p(1) * O_p(1) = o_p(1) \quad \text{as } \|Q-P\| \to 0,\end{aligned} \tag{7.21}$$

using (7.19) to conclude that expression (7.21) is $o_p(1)$ as $\|Q-P\| \to 0$. Thus (7.15) is $o_p(n^{-1/2})$.

To show that (7.16) is $o_p(n^{-1/2})$, expand ψ_n about $\underline{\lambda} = (\lambda_1,\ldots,\lambda_s)^T$, so that

$$\begin{aligned}&P\Big[\psi_n\big(X,\theta,\underline{V}(\theta,P)\big) - \psi\big(X,\theta,\underline{V}(\theta,P)\big)\Big] \\ &\quad= (\underline{\lambda}_n - \underline{\lambda})^T * P\Bigg[\frac{\partial}{\partial\underline{\lambda}}\psi\big(X,\theta,\underline{V}(\theta,P)\big)\Big|_{\underline{\lambda}^\star}\Bigg]\end{aligned} \tag{7.22}$$

for a vector $\underline{\lambda}^\star$ between $\underline{\lambda}$ and $\underline{\lambda}_n$. Since ψ is the efficient score function for θ,

$$\begin{aligned}P\Bigg[\frac{\partial}{\partial\underline{\lambda}}\psi\big(X,\theta,\underline{V}(\theta,P)\big)\Big|_{\underline{\lambda}^\star}\Bigg] &= P\Bigg[\frac{\partial}{\partial\underline{\lambda}}\big[\dot{l}_\theta(X) - I_{\theta G}I_G^{-1}\dot{l}_g(X)\big]|_{\underline{\lambda}^\star}\Bigg] \\ &= P\Bigg[\frac{\partial}{\partial\underline{\lambda}}\dot{l}_\theta(X)|_{\underline{\lambda}^\star}\Bigg] - \frac{\partial}{\partial\underline{\lambda}}\big[I_{\theta G}I_G^{-1}\big]\big|_{\underline{\lambda}^\star} * P\Big[\dot{l}_g(X)|_{\underline{\lambda}^\star}\Big] \\ &\quad -\big[I_{\theta G}I_G^{-1}\big]\big|_{\underline{\lambda}^\star} * P\Bigg[\frac{\partial}{\partial\underline{\lambda}}\dot{l}_g(X)|_{\underline{\lambda}^\star}\Bigg].\end{aligned} \tag{7.23}$$

Now, if we view the biased sampling density p of (1.1) as a function of the parameter triplet $(\underline{\lambda}, \theta, G)$, then orthogonality of the score for $\underline{\lambda}$ and the closed linear span of $(\dot{l}_\theta, \dot{l}_g)$ implies that the first and third terms in the sum (7.23) are zero. The second term in the sum is also zero since score functions have zero expectation. Therefore, (7.16) is $o_p(n^{-1/2})$, completing the proof that (5.12) holds.

It remains to verify the stochastic condition of the Master theorem [BKRW, condition (GM1), page 312; equivalent, Van der Vaart (1995), condition (3), Theorem 1]. By Lemma 1 of Van der Vaart (1995), it suffices to show that, for each component $l = 1, \ldots, d$,

$$\{\psi_{\theta_l} - \psi_{\theta_{0l}} : \|\theta - \theta_0\|_1 < \delta\} \tag{7.24}$$

is Donsker for some $\delta > 0$ and

$$P_0(\psi_{\theta_l} - \psi_{\theta_{0l}})^2 \to 0 \quad \text{as } \theta \to \theta_0. \tag{7.25}$$

From (5.11), we calculate

$$\begin{aligned}
\psi_{\theta_l}(x) - \psi_{\theta_{0l}}(x) = &\left[\frac{\dot{w}_{il}(y,\theta)}{w_i(y,\theta)} - \frac{\dot{w}_{il}(y,\theta_0)}{w_i(y,\theta_0)}\right] \\
&- \sum_{k=1}^{s}\left\{r_k(y,\theta)\left[\frac{\dot{w}_{kl}(y,\theta)}{w_k(y,\theta)} - \frac{\dot{w}_{kl}(y,\theta_0)}{w_k(y,\theta_0)}\right]\right\} \\
&- \sum_{k=1}^{s}\left\{\frac{\dot{w}_{kl}(y,\theta_0)}{w_k(y,\theta_0)}[r_k(y,\theta) - r_k(y,\theta_0)]\right\} \\
&+ A_l(\theta)M^-(\theta)\underline{\underline{\lambda}}^{-1}[-\underline{e}_i + \underline{r}(y,\theta)] \\
&- A_l(\theta_0)M^-(\theta_0)\underline{\underline{\lambda}}^{-1}[-\underline{e}_i + \underline{r}(y,\theta_0)] \\
\equiv\ & f_{1l\theta}(x) - f_{2l\theta}(x) - f_{3l\theta}(x) + f_{4l\theta}(x) - f_{5l\theta_0}(x),
\end{aligned}$$

where the $1 \times s$ matrix $A_l(\theta)$ is the lth row of the matrix $A(\theta)$ defined in (4.5).

To show that the class $S_{1l\delta} \equiv \{f_{1l\theta} : \|\theta - \theta_0\|_1 < \delta\}$ is Donsker for some $\delta > 0$ for each $l = 1, \ldots, d$, we use the uniform Lipschitz hypothesis (viii) on the weight functions and apply the Jain–Marcus central limit theorem; see, for example, Van der Vaart and Wellner [(1996), pages 213–214]. For fixed $i \in 1, \ldots, s$, $k = 1, \ldots, d$, choose δ, K and $\alpha > d/2$ so that (viii) holds. Define $\Theta_\delta \equiv \{\theta \in \Theta : \|\theta - \theta_0\|_1 < \delta\}$ and $Z_{nil}(y,\theta) \equiv n^{-1/2}[(\dot{w}_{il}(y,\theta)/w_i(y,\theta)) - (\dot{w}_{il}(y,\theta_0)/w_i(y,\theta_0))]$, so that $S_{1l\delta} \subset \bigcup_{i=1}^{s} S_{1il\delta}$, with $S_{1il\delta} = \{n^{1/2}Z_{nil}(y,\theta) : \theta \in \Theta_\delta\}$. Set $M_{nk} = n^{-1/2}K$ for $k = 1, \ldots, n$, so that $\sum_{k=1}^{n} EM_{nk}^2 = K^2 = O(1)$. Define a semimetric ρ_α on Θ_δ by $\rho_\alpha(f_{\theta_1}, f_{\theta_2}) = \|\theta_1 - \theta_2\|_1^\alpha$ for $f_{\theta_1}, f_{\theta_2} \in S_{1il\delta}$. We require that $\alpha > d/2$ because this is necessary and sufficient for the semimetric to be Gaussian dominated, that is, to have a finite entropy integral [satisfying the third panel display on page 212 of Van der Vaart and Wellner (1996)]. Next, we verify that the triangular array of norms $\|Z_{nil}\|_{\Theta_\delta} =$

$\sup_{\theta\in\Theta_\delta} |Z_{nil}(\cdot,\theta)|$ satisfies the Lindeberg condition of Theorem 2.11.11 in Van der Vaart and Wellner (1996):

$$\sum_{i=1}^{n} E\{\|Z_{nil}\|^2_{\Theta_\delta}\}1\{\|Z_{nil}\|_{\Theta_\delta} > \eta\} \to 0 \quad \text{for every } \eta > 0.$$

This follows because (viii) implies $E\|Z_{nil}\|^2_{\Theta_\delta} < n^{-1}K^2\delta^{2\alpha} = O(n^{-1})$ and $1_{[\|Z_{nil}\|_{\Theta_\delta}>\eta]} \le 1_{[K\delta^\alpha>\eta n^{1/2}]} \to 0$ as $n\to\infty$ for every $\eta > 0$.

Since the sequence of covariance functions $\mathrm{Cov}(Z_{ni_1l}(Y_1,\theta), Z_{ni_2l}(Y_2,\theta))$ converges pointwise in $\Theta_\delta\times\Theta_\delta$ for each $i_1\neq i_2\in\{1,\dots,s\}$, the Jain–Marcus CLT yields that, for each fixed $i=1,\dots,s$, $l=1,\dots,d$, $S_{1il\delta}$ is Donsker. Thus, $S_{1l\delta} = \{f_{1l\theta}\colon \|\theta-\theta_0\|_1 < \delta\}$ is Donsker.

For each $l=1,\dots,d$, the class $\{f_{2l\theta}\colon \|\theta-\theta_0\|_1<\delta\}$ is easily seen to be Donsker, using the assumption that the class $\mathbf{W}_\delta$ of (5.8) is Donsker and the proof that $\{f_{1l\theta}\colon \|\theta-\theta_0\|_1<\delta\}$ is Donsker.

To show $S_{4l\delta}\equiv\{f_{4l\theta}\colon \|\theta-\theta_0\|_1<\delta\}$ is Donsker, take $M^-(\theta)$ as in (4.7). Continuous invertibility of $M_{11}(\theta)$ at θ_0 and $\sup_{\|\theta-\theta_0\|_1<\delta} G(|\dot w^2_{kl}(\theta)/w_k(\theta)|^2\tilde w_k(\theta_0)) < \infty$ imply that each element of the vector $A_l(\theta)M^-(\theta)$ is a bounded function uniformly on Θ_δ. Since each $\{r_i(\theta)\colon \|\theta-\theta_0\|_1<\delta\}$ is Donsker and uniformly bounded, it follows that $S_{4l\delta}$ is Donsker, as is $S_{5l\delta}\equiv\{f_{5l\theta_0}\}$.

Lastly, we show $S_{3l\delta}\equiv\{f_{3l\theta}\colon \|\theta-\theta_0\|_1<\delta\}$ is Donsker. Fix $k\in\{1,\dots,s\}$. Define $d_k(y,\theta) = r_k(y,\theta) - r_k(y,\theta_0)$. By continuity of r_k in θ at θ_0, for any fixed positive $\epsilon<1$ we can choose δ so that, for each $y\in\mathbf{Y}$, $\sup_{\|\theta-\theta_0\|_1<\delta}|d_k(y,\theta)|<\epsilon$. Write the class $S_{3l\delta}$ as the sum of the two classes

$$\left\{\frac{\dot w_{kl}(y,\theta_0)}{w_k(y,\theta_0)}1_{\left[(\dot w_{kl}(y,\theta_0)/w_k(y,\theta_0))\le\frac{2\varepsilon}{1-\varepsilon}\right]} * d_k(y,\theta)\colon \|\theta-\theta_0\|_1<\delta\right\}, \tag{7.26}$$

$$\left\{\frac{\dot w_{kl}(y,\theta_0)}{w_k(y,\theta_0)}1_{\left[(\dot w_{kl}(y,\theta_0)/w_k(y,\theta_0))>\frac{2\varepsilon}{1-\varepsilon}\right]} * d_k(y,\theta)\colon \|\theta-\theta_0\|_1<\delta\right\}. \tag{7.27}$$

Since $\dot{\mathbf{W}}$ of (5.9) is Donsker, class (7.26) is Donsker as the product of uniformly bounded Donsker classes. Class (7.27) may not be uniformly bounded. We invoke Theorem 2.10.6 of Van der Vaart and Wellner (1996), which applies to Lipschitz transformations of Donsker classes. It asserts that if

$$\begin{aligned}
&\left|\frac{\dot w_{kl}(y,\theta_0)}{w_k(y,\theta_0)}1_{\left[(\dot w_{kl}(y,\theta_0)/w_k(y,\theta_0))>\frac{2\varepsilon}{1-\varepsilon}\right]}\,[d_k(y,\theta_1)-d_k(y,\theta_2)]\right|^2\\
&\quad\le\left|\frac{\dot w_{kl}(y,\theta_0)}{w_k(y,\theta_0)}1_{\left[(\dot w_{kl}(y,\theta_0)/w_k(y,\theta_0))>\frac{2\varepsilon}{1-\varepsilon}\right]}-d_k(y,\theta_1)\right|^2\\
&\qquad+\left|\frac{\dot w_{kl}(y,\theta_0)}{w_k(y,\theta_0)}1_{\left[(\dot w_{kl}(y,\theta_0)/w_k(y,\theta_0))>\frac{2\varepsilon}{1-\varepsilon}\right]}-d_k(y,\theta_2)\right|^2
\end{aligned}$$

for all $\theta_1,\theta_2\in\Theta_\delta$, then class (7.27) is Donsker. This follows by straightforward algebra, so that the kth summand of $S_{3l\delta}$ is Donsker. Then $S_{3l\delta}$ itself is Donsker, and condition (7.24) holds.

It remains to verify (7.25). Hypotheses (iv), (vii) and (viii) imply that $P_0[\sup_{\{\|\theta-\theta_0\|_1<\delta\}}(\psi_{\theta_l}-\psi_{\theta_{0l}})^2]<\infty$ for each $l=1,\ldots,d$. Thus, in view of the dominated convergence theorem, it suffices to show that $\psi_{\theta_l}\to\psi_{\theta_{0l}}$ pointwise in x, for each l. This is true by continuity of the weights and scores for weights in θ at θ_0. $\square$

PROOF OF THEOREM 5.4. For $h\in\mathbf{H}$, write

$$\mathbb{Z}_n(h,\hat\theta_n)=\sqrt{n}\big(\mathbb{G}_n\big(h,\hat\theta_n,\underline{\mathbb{V}}_n(\hat\theta_n)\big)-\mathbb{G}_n\big(h,\theta_0,\underline{\mathbb{V}}_n(\theta_0)\big)\big) \tag{7.28}$$

$$+\sqrt{n}\big(\mathbb{G}_n\big(h,\theta_0,\underline{\mathbb{V}}_n(\theta_0)\big)-\mathbb{G}_n\big(h,\theta_0,\underline{V}(\theta_0)\big)\big). \tag{7.29}$$

Since $G(h_e^2 r(\theta_0))<\infty$ and the class $\mathbf{F}$ is Donsker for each F_{i0}, piece (7.29) satisfies

$$\|\sqrt{n}(\mathbb{G}_n(h,\theta_0,\underline{\mathbb{V}}_n(\theta_0))-G(h,\theta_0,\underline{V}(\theta_0)))-Z(h)\|_{\mathbf{H}}\to_p 0$$

by Theorem 2.2 of GVW. Expanding, write piece (7.28) as

$$\begin{aligned}&\sqrt{n}\big(\mathbb{G}_n(h,\hat\theta_n,\underline{\mathbb{V}}_n(\hat\theta_n))-\mathbb{G}_n(h,\theta_0,\underline{\mathbb{V}}_n(\theta_0))\big)\\&\quad=(\nabla\mathbb{G}_n(h,\theta^\star,\underline{\mathbb{V}}_n(\theta^\star)))^T*\sqrt{n}(\hat\theta_n-\theta_0)\end{aligned} \tag{7.30}$$

for some $\theta^\star$ between θ_0 and $\hat\theta_n$, where $\nabla\mathbb{G}_n(h,\theta,\underline{\mathbb{V}}_n(\theta))\equiv((\partial/\partial\theta_1')\mathbb{G}_n(h,\theta',\underline{\mathbb{V}}_n(\theta'))|_{\theta'=\theta},\ldots,(\partial/\partial\theta_d')\mathbb{G}_n(h,\theta',\underline{\mathbb{V}}_n(\theta'))|_{\theta'=\theta})^T$. From representation (1.15) of $\mathbb{G}_n$, the gradient $\nabla\mathbb{G}_n(h,\theta,\underline{\mathbb{V}}_n(\theta))$ can be calculated by the chain rule, using $(\partial/\partial\theta')\underline{\mathbb{V}}_n(\theta')|_\theta=A_n(\theta)M_n^-(\theta)\underline{\underline{\lambda}}_n^{-1}\underline{\underline{\mathbb{V}}}_n(\theta)$, and equals

$$\begin{aligned}&\frac{-\mathbb{F}_n\big(h\sum_{i=1}^s r_{ni}(\theta)\frac{\dot w_i(Y,\theta)}{w_i(Y,\theta)}r_n(\theta)\big)}{\mathbb{F}_n(r_n(\theta))}+\frac{\mathbb{F}_n\big(\sum_{i=1}^s r_{ni}(\theta)\frac{\dot w_i(Y,\theta)}{w_i(Y,\theta)}r_n(\theta)\big)}{\mathbb{F}_n(r_n(\theta))}*\frac{\mathbb{F}_n(hr_n(\theta))}{\mathbb{F}_n(r_n(\theta))}\\&+A_n(\theta)M_n^-(\theta)\bigg[\frac{-\mathbb{F}_n(h\underline{\tilde w}(\theta)r_n^2(\theta))}{\mathbb{F}_n(r_n(\theta))}+\frac{\mathbb{F}_n(\underline{\tilde w}(\theta)r_n^2(\theta))}{\mathbb{F}_n(r_n(\theta))}*\frac{\mathbb{F}_n(hr_n(\theta))}{\mathbb{F}_n(r_n(\theta))}\bigg].\end{aligned}$$

By uniform consistency of $\mathbb{G}_n(h,\hat\theta_n,\underline{\mathbb{V}}_n(\hat\theta_n))$, $\nabla\mathbb{G}_n(h,\theta^\star,\underline{\mathbb{V}}_n(\theta^\star))$ converges in probability to

$$\begin{aligned}&-G\bigg(\sum_{k=1}^s r_k(\theta_0)\frac{\dot w_k(\theta_0)}{w_k(\theta_0)}(h-G(h))\bigg)\\&\quad-A(\theta_0)M^-(\theta_0)G\big[r(\theta_0)\underline{\tilde w}(\theta_0)(h-G(h))\big]\\&\quad=-G(ha_\star(\theta_0)).\end{aligned} \tag{7.31}$$

Thus for fixed $h\in\mathbf{H}$, piece (7.28) converges in distribution to $-G(ha_\star(\theta_0))*Z_{\theta_0}$. We then have

$$\begin{aligned}&(\nabla\mathbb{G}_n(h,\theta^\star,\underline{\mathbb{V}}_n(\theta^\star)))^T*\sqrt{n}(\hat\theta_n-\theta_0)\\&\quad=\big[\big(\nabla\mathbb{G}_n(h,\theta^\star,\underline{\mathbb{V}}_n(\theta^\star))\big)^T+G(ha_\star^T(\theta_0))\big]*\sqrt{n}(\hat\theta_n-\theta_0)\\&\quad\quad-G(ha_\star^T(\theta_0))*\sqrt{n}(\hat\theta_n-\theta_0)\end{aligned}$$

$$= o_p(1) * O_p(1) - G(ha_\star^T(\theta_0)) * \sqrt{n}(\hat{\theta}_n - \theta_0)$$
$$\Rightarrow -G(ha_\star^T(\theta_0)) * Z_{\theta_0} \quad \text{in } l^\infty(\mathbf{H}).$$

To establish the convergence of $\mathbb{Z}_n'(h, \hat{\theta}_n)$ to $Z'(h)$ of (5.6), it remains to show that $\lim_n \mathbb{Z}_n'$ has covariance (5.7). Using (7.30) and (7.31), write

$$\mathbb{Z}_n'(h, \hat{\theta}_n) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \begin{pmatrix} \tilde{l}_{\theta_0}(X_i) \\ -G(ha_\star^T(\theta_0)) * \tilde{l}_{\theta_0}(X_i) + \tilde{l}_{G^0}(\pi_h)(X_i) \end{pmatrix} + o_p(1)$$
$$\to_d N_{d+1}\left(\underline{0}, \begin{bmatrix} \Sigma_{11} & \Sigma_{12}(h) \\ \Sigma_{21}(h) & \Sigma_{22}(h, h) \end{bmatrix}\right),$$

where $\tilde{l}_\theta$ is the efficient influence function for θ and $\tilde{l}_{G^0}$ is the efficient influence function for G when $\theta = \theta_0$ is known. Evidently, $\Sigma_{11} = I^{-1}(\theta_0)$. Since the scores for θ must be orthogonal to the nuisance parameter scores, and $\tilde{l}_{G^0}(\pi_h)$ is in the space of nuisance parameter scores, it follows that $E(\tilde{l}_{\theta_0}(X)\tilde{l}_{G^0}(\pi_h)(X)) = \underline{0}$. Then elementary calculation yields $\Sigma_{12}(h) = -I^{-1}(\theta_0)G(ha_\star(\theta_0))$ and $\Sigma_{22}(h, h) = G(ha_\star^T(\theta_0))I^{-1}(\theta_0)G(ha_\star(\theta_0)) + I_{G^0}^{-1}(h, h) = I_G^{-1}(h, h)$. It follows that $\text{AsymCov}(\mathbb{Z}_n'(h_1, \hat{\theta}_n), \mathbb{Z}_n'(h_2, \hat{\theta}_n))$, for $h_1, h_2 \in \mathbf{H}$, matches the covariance formula (5.7). □

PROOF OF PROPOSITION 5.2. Adding and subtracting $\mathbb{G}_n(\underline{w}(\hat{\theta}_n), \theta_0)$ and $\mathbb{G}_n(\underline{w}(\theta_0), \theta_0)$, and expanding about θ_0, allows us to write

$$\sqrt{n}(\underline{\mathbb{W}}_n - \underline{W}_0) = \left[\nabla\mathbb{G}_n(\underline{w}(\hat{\theta}_n), \theta^\star, \underline{\mathbb{V}}_n(\theta^\star)) + \mathbb{G}_n(\underline{\dot{w}}(\theta^\star), \theta_0, \underline{\mathbb{V}}_n(\theta_0))\right]^T \times \sqrt{n}(\hat{\theta}_n - \theta_0) \tag{7.32}$$
$$+ \sqrt{n}(\underline{\mathbb{W}}_n(\theta_0) - \underline{W}_0) \tag{7.33}$$

for some $\theta^\star$ between θ_0 and $\hat{\theta}_n$. By Proposition 2.2 of GVW, the process in (7.33) converges in distribution to the first two lines of (5.13), with covariance given by the first three lines of (5.14). Slutsky's theorem and (7.31) show that expression (7.32) converges to the third and fourth lines of (5.13), with covariance given by the third and fourth lines of (5.14). Since $\underline{\mathbb{W}}_n(\theta_0)$ has influence function $\tilde{l}_{\underline{W}_0} = [\tilde{l}_{G^0}(\pi_{w_1(\theta_0)}), \ldots, \tilde{l}_{G^0}(\pi_{w_s(\theta_0)})]^T$, the asymptotic covariance of the two terms (7.32) and (7.33) equals

$$\underline{\underline{W}}_0\left[-G(\underline{\tilde{w}}(\theta_0)a_\star^T(\theta_0)) + G\left(\frac{\dot{w}_1(\theta_0)}{w_1(\theta_0)}\tilde{w}_1(\theta_0), \ldots, \frac{\dot{w}_s(\theta_0)}{w_s(\theta_0)}\tilde{w}_s(\theta_0)\right)^T\right] \times E[\tilde{l}_{\theta_0}(X)\tilde{l}_{\underline{W}_0}^T(X)].$$

Since $E(\tilde{l}_{\theta_0}(X)\tilde{l}_{G^0}(\pi_h)) = \underline{0}$ for $h \in \mathbf{H}$, and each $w_i(\cdot, \theta_0) \in \mathbf{H}$ by requirement (5.8), the last expectation term is zero. □

PROOF OF PROPOSITION 5.3. Write

$$\sqrt{n}(\hat{\mathbb{F}}_{ni}(h,\hat{\theta}_n)-F_{i0}(h))=\sqrt{n}\Bigg(\frac{\mathbb{G}_n(hw_i(\hat{\theta}_n),\hat{\theta}_n,\underline{\mathbb{V}}_n(\hat{\theta}_n))}{\mathbb{G}_n(w_i(\hat{\theta}_n),\hat{\theta}_n,\underline{\mathbb{V}}_n(\hat{\theta}_n))} \tag{7.34}$$

$$-\frac{\mathbb{G}_n(hw_i(\theta_0),\theta_0,\underline{\mathbb{V}}_n(\theta_0))}{\mathbb{G}_n(w_i(\theta_0),\theta_0,\underline{\mathbb{V}}_n(\theta_0))}\Bigg)$$

$$+\sqrt{n}\big(\hat{\mathbb{F}}_{ni}(h,\theta_0)-F_{i0}(h)\big). \tag{7.35}$$

The process of (7.35) converges in distribution to the first two lines of (5.15) by Theorem 2.3 of GVW. Using the identity $(\partial/\partial\theta')\tilde{w}_i(y,\theta')|_{\theta'=\theta}=\dot{l}_\theta(i,y)$, a Taylor expansion shows that piece (7.34) equals

$$\big[\nabla\mathbb{G}_n(h\tilde{w}_i(\hat{\theta}_n,\mathbb{F}_n),\theta^\star,\underline{\mathbb{V}}_n(\theta^\star))+\mathbb{G}_n(h\dot{l}_{\theta^\star}(i,Y),\theta_0,\underline{\mathbb{V}}_n(\theta_0))\big]^T*\sqrt{n}(\hat{\theta}_n-\theta_0)+o_p(1)$$

for some θ^* between θ_0 and $\hat{\theta}_n$, which converges in distribution to the term in the last line of (5.15). From the expansion used by GVW [page 1107], the influence function $\tilde{l}_{F_{i0}}$ of $\hat{\mathbb{F}}_{ni}$ equals

$$\tilde{l}_{F_{i0}}(\pi_h)=\tilde{l}_{G^0}(\pi_{w_i(\theta_0)h})+G(\tilde{w}_i(\theta_0)h)W_{0i}^{-1}\tilde{l}_{G^0}(\pi_{w_i(\theta_0)}).$$

The asymptotic covariance between (7.34) and (7.35) equals

$$\big[-G([h\tilde{w}_i(\theta_0)-G(h\tilde{w}_i(\theta_0))]a_\star(\theta_0))+G(h\dot{l}_{\theta_0}(i,Y))\big]^T E\big[\tilde{l}_{\theta_0}(X)\tilde{l}_{F_i(\theta_0)}(\pi_h)\big],$$

which is zero by the assumption that each $w_i(\cdot,\theta_0)$ and each $w_i(\cdot,\theta_0)h$ is in $\mathbf{H}$. □

PROOF OF THEOREM 6.1, COROLLARY 6.1 AND COROLLARY 6.2. Since the weights are positive, the connectivity of $\mathbf{G}^\star(\theta)$ for all θ in a neighborhood of θ_0 holds automatically. As discussed in Section 3 of Gilbert, Lele and Vardi (1999), the remaining conditions of identifiability Theorem 1.1 are met if $h_{ik}(0)=0$ for each $i=1,\dots s-1$, $k=1,\dots,d$, and for some $i\in\{1,\dots,s-1\}$, the set of functions $\{h_{i1},\dots,h_{id}\}$ is linearly independent. The MLE obtained by procedure (1)–(3) exists uniquely in the limit with probability 1, as verified in Example 6.1. In addition, methods for verifying the hypotheses of the theorems in Section 5 were outlined in Example 6.1. □

8. Discussion and further problems. Results from a simulation study of the MLE $(\hat{\theta}_n,\mathbb{G}_n)$ in the two- and three-sample GLR models [Gilbert (1996); Gilbert, Lele and Vardi (1999)] corroborate the large sample properties described here. The simulation design utilized a quasi-Newton scheme to compute the MLE via procedure (1)–(3). The log profile partial likelihood was found to be smooth and strictly concave for the great majority of data configurations. Asymptotic unbiasedness and normality of $(\hat{\theta}_n,\mathbb{G}_n)$ were confirmed, as were consistency of the plug-in and bootstrap variance estimates. Further, the likelihood ratio, Wald and efficient score tests were found to be consistent and possess approximate chi-squared distributions, and profile partial likelihood-based confidence sets for θ_0 had accurate coverage probabilities.

In summary, the maximum likelihood estimator in the semiparametric biased sampling model shares many properties with the maximum likelihood estimator in Cox's semiparametric proportional hazards model. The estimate $\hat{\theta}_n$ is computed conveniently by maximizing a smooth quadratic log profile partial likelihood, and then the NPMLE of the baseline distribution is computed. The asymptotic properties of the joint MLEs are comparable, each converging at rate $n^{1/2}$ and achieving semiparametric efficiency, with similar forms and derivations of information bounds. These similarities are not surprising, as the s-group proportional hazards model, defined by $\lambda(y, \theta|i) = \exp(\theta_i)\lambda(y|s)$ for $i = 1, \ldots, s$, $\theta = (\theta_1, \ldots, \theta_s)^T$, $\theta_s \equiv 0$, has the analytic form of an s-sample biased sampling model (albeit with weight functions depending on the infinite-dimensional parameter G) with $w_i(y, \theta, G) = (1 - G(y))^{\exp(\theta_i)-1}$, where G is the distribution function Y for the sth group.

PROBLEM 8.1 (All weight functions unknown). As alluded to in the Introduction, the one-sample biased sampling model (with no restrictions on G) is rarely identifiable when the weight function depends on θ. In cases that it is [see Gilbert, Lele and Vardi (1999), Theorem 1], no systematic estimation procedure exists. In multiple-sample biased sampling models in which all weight functions depend on an unknown parameter, the class of identifiable models is larger [see Gilbert, Lele and Vardi (1999), Theorem 3]. If one of the normalizing constants is known, the methodology given here applies. If not, which is commonly the case, there does not exist a systematic procedure for estimation. A thorough treatment of these situations would be of interest.

PROBLEM 8.2 (Adaptive or nonparametric estimation of the weight functions). In many problems, there may be little rationale for choosing a parametric form for the weight functions. For instance, consider the vaccine trial Examples 1.1, 1.2, and 6.1. For vaccines against some heterogeneous pathogens, including HIV, there are very few data available to guide selection of the weight functions. In this situation, it is of interest to consider more adaptive estimation of the $\{w_i\}$. In the econometrics literature, Manski (1993) and references therein consider estimation of the weight function in one biased sample via kernel methods. Sun and Woodroofe (1997) consider one- and two-sample biased sampling models in which the underlying distribution function is specified parametrically, $w_2 = 1$ if there is a second sample, and the weight function w_1 is estimated nonparametrically. They assume w_1 is a monotone density function and construct a consistent (penalized) maximum likelihood estimator $\hat{w}_1$. It would be of interest to investigate the asymptotic distributional properties of this estimator. Another approach is to estimate w_1 by smoothing splines through maximization of a roughness-penalized version of the profile partial likelihood considered here. We are pursuing research in this direction.

Adaptive or nonparametric estimation of the weight functions is of interest in its own right, or as a goodness-of-fit diagnostic for checking the parametric form of the weight functions.

Acknowledgment. I am grateful to Jon Wellner for his very helpful and kind guidance.

REFERENCES

AGRESTI, A. (1984). *Analysis of Ordinal Categorical Data*. Wiley, New York.

ANDERSEN, P. K. and GILL, R. D. (1982). Cox's regression model for counting processes: a large sample study. *Ann. Statist.* **10** 1100–1120.

BEGUN, J. M., HALL, W. J., HUANG, W-M. and WELLNER, J. A. (1983). Information and asymptotic efficiency in parametric-nonparametric models. *Ann. Statist.* **11** 432–452.

BICKEL, P. J., KLAASSEN, C. A., RITOV, Y. and WELLNER, J. A. (1993). *Efficient and Adaptive Estimation for Semiparametric Models*. Johns Hopkins Univ. Press, Baltimore, MD.

COSSLETT, S. R. (1981). Maximum likelihood estimator for choice based samples. *Econometrika* **49** 1289–1316.

COX, D. R. and SNELL, E. J. (1989). *The Analysis of Binary Data*. 2nd ed. Chapman and Hall, London.

DUDLEY, R. M. (1984). A course on empirical process. *Ecole d' Été de Probabilitiés de Saint Flour XII. Lecture Notes in Math*. **1097** 1–142. Springer, New York.

DUDLEY, R. M. (1985). An extended Wichura theorem, definitions of Donsker classes, and weighted empirical processes. *Probability in Banach Spaces V. Lecture Notes in Math.* **1153** 1306–1326. Springer, New York.

FIENBERG, S. E. (1980). *The Analysis of Cross-Classified Categorical Data*. MIT Press.

GILBERT, P. B. (1996). Sieve analysis: statistical methods for assessing differential vaccine protection against HIV types. Ph.D. dissertation, Univ. Washington.

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.

GILBERT, P. B., SELF, S. G. and ASHBY, M. A. (1998). Statistical methods for assessing differential vaccine protection against HIV types. *Biometrics* **54** 799-814.

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.

HUBER, P. J. (1967). The behavior of maximum likelihood estimates under nonstandard conditions. *Proceedings of the Fifth Berkeley Symp. Math. Statist. Probab.* **1** 221–233. Univ. California Press, Berkeley.

KIEFER, J. and WOLFOWITZ, J. (1956). Consistency of the maximum likelihood estimator in the presence of infinitely many incidental parameters. *Ann. Math. Statist.* **27** 887–906.

MANSKI, C. F. (1993). The selection problem in econometrics and statistics. In *Handbook of Statistics* **11** (G. S. Maddala, C. R. Rao and H. D. Vinod, Eds.) 73–84. North-Holland, Amsterdam.

MANSKI, C. F. and LERMAN, S. R. (1977). The estimation of choice probabilities from choice-based samples. *Econometrics* **45** 1977–1988.

OSSIANDER, M. (1987). A central limit theorem under metric entropy with L_2 bracketing. *Ann. Probab.* **15** 897–919.

POLLARD, D. (1984). *Convergence of Stochastic Processes*. Springer, New York.

QIN, J. (1998). Inferences for case-control and semiparametric two-sample density ratio models. *Biometrika* **85** 619–630.

SUN, J. and WOODROOFE, M. (1997). Semiparametric estimates for biased sampling models. *Statist. Sinica* **7** 545–575.

TRICOMI, F. G. (1957). *Integer Equations*. Interscience, New York.

VAN DER VAART, A. (1994). Bracketing smooth functions. *Stochast. Process. Appl.* **52** 93–105.

VAN DER VAART, A. (1995). Efficiency of infinite dimensional M-estimators. *Statist. Neerlandica* **49** 9–30.

VAN DER VAART, A. and WELLNER, J. A. (1996). *Weak Convergence and Empirical Processes*. Springer, New York.

VARDI, Y. (1985). Empirical distributions in selection bias models. *Ann. Statist.* **13** 178–203.
WELLNER, J. A. and ZHAN, Y. (1998). Bootstrapping Z-estimators. Technical Report 308, Dept. Statistics, Univ. Washington.

DEPARTMENT OF BIOSTATISTICS
HARVARD SCHOOL OF PUBLIC HEALTH
BOSTON, MASSACHUSETTS 02115
E-MAIL: pgilbert@hsph.harvard.edu