The Canadian Journal of Statistics
Vol. 29, No. 2, 2001, Pages 269–288
La revue canadienne de statistique

Consistency of semiparametric maximum likelihood estimators for two-phase sampling

Aad van der VAART and Jon A. WELLNER

Key words and phrases: Consistency; design; empirical processes; Glivenko–Cantelli theorem; identifiability; maximum likelihood; missing data; mixture; outcome dependence; stratified sampling; two-phase sampling.

MSC 2000: Primary: 60F05, 60F17; secondary 60J65, 60J70.

Abstract: Semiparametric maximum likelihood estimators have recently been proposed for a class of two-phase, outcome-dependent sampling models. All of them were "restricted" maximum likelihood estimators, in the sense that the maximization is carried out only over distributions concentrated on the observed values of the covariate vectors. In this paper, the authors give conditions for consistency of these restricted maximum likelihood estimators. They also consider the corresponding unrestricted maximization problems, in which the "absolute" maximum likelihood estimators may then have support on additional points in the covariate space. Their main consistency result also covers these unrestricted maximum likelihood estimators, when they exist for all sample sizes.

Convergence des estimateurs du maximum de vraisemblance semiparamétrique dans le cadre d'échantillonnage à deux phases

Résumé : Des estimateurs du maximum de vraisemblance semiparamétrique ont récemment été proposés dans le cadre de modèles pour plans d'échantillonnage doubles à probabilités de sélection dépendant de covariables. Il s'agissait dans tous les cas d'estimateurs à vraisemblance maximale restreinte, en ce sens que la maximisation n'était effectuée que sur les lois ayant pour support l'ensemble des valeurs observées des vecteurs de covariables. Dans cet article, les auteurs donnent des conditions assurant la convergence de ces estimateurs à vraisemblance maximale restreinte. Ils considèrent en outre les problèmes de maximisation non-restreinte, dans lesquels les estimateurs à vraisemblance maximale "absolus" peuvent dépendre de points additionnels de l'espace des covariables. Leur principal résultat de convergence s'applique à ces estimateurs à vraisemblance maximale non-restreinte, lorsque ceux-ci existent pour toute taille d'échantillon.

1. INTRODUCTION

Outcome-dependent two-phase sampling designs have been the subject of considerable work in the recent statistical literature. Here is a simple example of the type of problem we have in mind.

Example 1 (Binary response logistic regression). Suppose that $Y \mid X = x \sim$ Bernoulli $\{f_\theta(1 \mid x)\}$ and $X \sim G$ on $\mathcal{X}$, where

$$f_\theta(1 \mid x) = \frac{\exp(\theta' x)}{1 + \exp(\theta' x)} = 1 - f_\theta(0 \mid x)$$

for $\theta \in \Theta \subset R^d$ and $x \in \mathcal{X} \subset R^d$. Let $q_{\theta,g}(y, x) \equiv f_\theta(y \mid x) g(x)$ denote the resulting joint density of (Y, X), where g is a density of G with respect to some dominating measure μ. Suppose that $\mathcal{X}$ and $\mathcal{Y} = \{0, 1\}$ are partitioned as

$$\mathcal{X} = \bigcup_{l=1}^{L} \mathcal{A}_l \qquad \text{and} \qquad \mathcal{Y} = \{0\} \cup \{1\} \equiv \mathcal{B}_0 \cup \mathcal{B}_1$$

and the resulting partition of $\mathcal{Y} \times \mathcal{X}$ is the product partition

$$\mathcal{X} \times \mathcal{Y} = \bigcup_{l=1}^{L} \bigcup_{k=0}^{1} \mathcal{A}_l \times \mathcal{B}_k \equiv \bigcup_{j=1}^{J} \mathcal{S}_j$$

into $J = 2L$ sets. If "complete" or "full" data were available, then we would observe $(Y_1, X_1), \ldots, (Y_n, X_n)$ independent and identically distributed (i.i.d.) with density $q_{\theta,g}$. Often however, exact measurement of all the components of the X_i's is expensive, while knowledge of which category $\mathcal{S}_j$ into which the data fall is much less expensive. Let $S_i = j \in \{1, \ldots, J\}$ when $(Y_i, X_i) \in \mathcal{S}_j$. Then the "incomplete" or "two-phase" sampling data is (R_i, T_i) where

$$T_i \equiv \begin{cases} (Y_i, X_i), & \text{if } R_i = 1, \\ S_i, & \text{if } R_i = 0, \end{cases}$$

and the probability of observing a "complete observation" for the ith individual is

$$\mathrm{P}(R_i = 1 \mid Y_i, X_i) = \sum_{j=1}^{J} p_j 1\{(Y_i, X_i) \in \mathcal{S}_j\}.$$

Here the sampling probabilities p_j for the cells $\mathcal{S}_j$, $j = 1, \ldots, J$, can be chosen by the investigator. In this particular example, the p_js would typically be chosen to be 1 for "rare cells" $\mathcal{S}_j$, and much smaller for the more frequently occurring (and/or expensive) cells in the design.

Example 1 is exactly the model used by Breslow & Chatterjee (1999), and studied in considerable detail by Breslow & Holubkov (1997). Our general framework, which includes Example 1, is given in Section 2; for further examples, see Section 4.

Although a variety of ad-hoc, inefficient estimation methods have been used frequently in the past, recent work by Breslow & Holubkov (1997), Lawless, Wild & Kalbfleisch (1999), and Scott & Wild (1998) has focused on nonparametric maximum likelihood estimation for natural semiparametric regression models connected with these designs. Breslow & Holubkov (1997) showed how to obtain maximum likelihood estimators for a logistic semiparametric regression model with two-phase outcome dependent sampling which generalized the classical logistic regression case-control model studied by Prentice & Pyke (1979). Lawless, Wild & Kalbfleisch (1999) and Scott & Wild (1998) generalized the approach and results of Breslow & Holubkov (1997) and Scott & Wild (1997), and demonstrate how the computation of the maximum likelihood estimators is possible in a wide range of regression models and two-phase designs. Lawless, Wild & Kalbfleisch (1999) carried out a simulation study to compare the semiparametric maximum likelihood estimators with estimators based on the "complete data likelihood" and various "estimated" and "weighted" pseudo-likelihoods. Their Monte-Carlo evidence suggests that the maximum likelihood estimator is fully efficient, while the various alternative methods can be quite inefficient, sometimes severely so.

Our goal in this paper is to establish consistency of the MLEs for the Bernoulli sampling (or i.i.d.) version of these outcome-dependent, two-phase sampling designs. Breslow, McNeney, and Wellner (2000) have established information lower bounds for estimation in these models; these authors have also shown that the semiparametric MLEs studied by Breslow & Holubkov (1997), Scott & Wild (1998), and Lawless, Wild & Kalbfleisch (1999) are asymptotically normal and efficient: they achieve the information lower bounds.

The models considered here are closely related to the biased sampling models considered by Gill, Vardi & Wellner (1988), the semiparametric generalizations thereof treated by Gilbert (1998), and to the "partially censored" mixture models studied by van der Vaart & Wellner (1992) and van der Vaart (1994). We will comment further on some of the similarities and differences in Sections 3 and 5. For a discussion of some of the practical aspects of these designs and an interesting Monte-Carlo study, we refer the interested reader to Breslow & Chatterjee (1999).

In Section 2, we present the model and a derivation of the distribution of the data under several of the two-phase models studied here. In Section 3, we state our consistency result for the i.i.d. special case (Bernoulli sampling or VPS1) of the model. Section 4 gives several examples, while Section 5 contains discussion and further problems. Proofs for Section 3 are given in the Appendix. For information bounds, results on asymptotic normality and efficiency of the estimators, we refer the interested reader to Breslow, McNeney & Wellner (2000).

2. THE MODEL AND MAXIMUM LIKELIHOOD ESTIMATION

Suppose that (Y, X) has density $f(y \mid x; \theta)g(x) = f_\theta(y \mid x)g(x)$ with respect to a dominating measure $\nu \times \mu$ on $\mathcal{Y} \times \mathcal{X}$. Here $\theta \in \Theta$, where Θ is a compact metric space (typically a compact subset of R^d for some d), and g is a completely unknown density for the covariate vector X with values in R^d. (In fact the distribution G of X need not have a density g in general, and a density will not be assumed in Sections 3 to 6.) We suppose throughout that the true values of the parameters (θ, G) are (θ_0, G_0). Both X and Y may be multivariate.

Let $\mathcal{Y} \times \mathcal{X} = \bigcup_{j=1}^J \mathcal{S}_j$ for a partition $\{\mathcal{S}_j\}$ into J strata; here $\mathcal{S}_j \cap \mathcal{S}_{j'} = \emptyset$ for $j \neq j'$. Suppose that

$$Q_j(\theta, G) = \mathrm{P}\{(Y, X) \in \mathcal{S}_j\}, \qquad Q_j^*(x, \theta) \equiv \mathrm{P}\{(Y, x) \in \mathcal{S}_j \mid X = x\} 1_{\mathcal{S}_j^*}(x),$$

for $j = 1, \ldots, J$, where $\mathcal{S}_j^* \equiv \{x \in \mathcal{X} : \text{ for some } (y, x) \in \mathcal{S}_j\}$. Thus

$$Q_j^*(x, \theta) = \int_{\{y:(y,x)\in\mathcal{S}_j\}} f(y \mid x; \theta)\, d\nu(y),$$

and

$$Q_j(\theta, G) = \int Q_j^*(x, \theta)\, dG(x),$$

for $j = 1, \ldots, J$. Note that the $\mathcal{S}_j^*$s do not, in general form a partition of $\mathcal{X}$, and may in fact intersect in quite arbitrary ways. We will however, often be interested in the case in which the partition $\{\mathcal{S}_j\}$ is formed as the natural product partition in terms of partitions $\{\mathcal{X}_m\}$ of $\mathcal{X}$ and $\{\mathcal{Y}_{m'}\}$ of $\mathcal{Y}$, and then $\mathcal{S}_j^* = \mathcal{X}_m$ for some m.

Suppose that $(Y_1, X_1), \ldots, (Y_n, X_n)$ are i.i.d. as (Y, X). Define

$$S_i = j \quad \text{if and only if} \quad (Y_i, X_i) \in \mathcal{S}_j, \quad i = 1, \ldots, n,\ j = 1, \ldots, J,$$

$$N_j = \sum_{i=1}^n 1(S_i = j), \quad j = 1, \ldots, J.$$

Then, with $Q_j \equiv Q_j(\theta, G)$, $\underline{N} \sim \mathrm{Mult}_J(n, \underline{Q})$. Furthermore, set

$$R_i = \begin{cases} 1, & \text{if } (Y_i, X_i) \text{ is fully observed,} \\ 0, & \text{if only } S_i \text{ is observed,} \end{cases} \qquad i = 1, \ldots, n.$$

The two observational schemes studied in Lawless, Wild & Kalbfleisch (1999) are:

a) ***Basic Stratified Sampling*** (BSS): In this scheme, there are n i.i.d. units (Y_i, X_i), $i = 1, \ldots, n$, from the density

$$q(y, x; \theta, g) = f(y \mid x; \theta)g(x) \tag{1}$$

in the background available for observation. The sampling proceeds in two stages:

> Stage 1: At this stage the counts N_j, $j = 1, \ldots, J$ are observed (and the (Y_i, X_i) pairs are categorized by strata); thus we observe $\underline{N} \sim \mathrm{Mult}_J(n, \underline{Q})$.

Stage 2: For each strata $\mathcal{S}_j$, we choose n_j units from the N_j units available in the jth strata to observe fully.

b) *Variable Probability Sampling* (VPS): Units are inspected sequentially as they arise from the density (1). When $(Y_i, X_i) \in \mathcal{S}_j$, the ith unit is selected for full observation ($R_i = 1$) with specified probability p_j; thus

$$\pi(Y_i, X_i) \equiv \mathrm{P}(R_i = 1 \mid Y_i, X_i) = \sum_{j=1}^{J} p_j 1\{(Y_i, X_i) \in \mathcal{S}_j\}.$$

This will lead to a random number n_j of completely observed units in stratum $\mathcal{S}_j$. There are several variants of this sampling plan, depending on how the sampling is terminated:

VPS1: Inspect a pre-specified number n of units.

VPS2: Inspect units until a total of k have been selected for full observation.

PROPOSITION 1. *For any of the above designs, the full semiparametric likelihood is of the form*

$$L_n(\theta, G) = \prod_{j=1}^{J} \left\{ \prod_{i:R_i=1,\, S_i=j} f(Y_i \mid X_i; \theta) g(X_i) \right\} Q_j(\theta, G)^{N_j - n_j}. \tag{2}$$

Proposition 1 is proved in Appendix B of Scott & Wild (1998); see also Lawless, Wild & Kablfleisch (1999), Section 3.1.1. For the observations we will often write $(R_1, T_1), \ldots, (R_n, T_n)$, where

$$T_i \equiv \begin{cases} (Y_i, X_i), & R_i = 1, \\ S_i, & R_i = 0. \end{cases}$$

Note that in all the above sampling schemes (BSS, VPS1, VPS2), the distribution of $\underline{R} \equiv (R_1, \ldots, R_n)$ depends on $(\underline{Y}, \underline{X}) \equiv ((Y_1, X_1), \ldots, (Y_n, X_n))$ only through the vector $\underline{S} \equiv (S_1, \ldots, S_n)$. When the sampling is carried out according to VPS1 (Bernoulli sampling), the (R_i, T_i), $i = 1, \ldots, n$, are independent and identically distributed since then the R_is are independent with conditional distributions depending only on (Y_i, X_i) through S_i for $i = 1, \ldots, n$. The resulting density of $(R_1, T_1), \ldots, (R_n, T_n)$ is $p(\underline{r}, \underline{t}; \theta, G) = \prod_{i=1}^{n} p(r_i, t_i; \theta, G)$, where

$$\begin{aligned} p(r, t; \theta, G) &= \left[\prod_{j=1}^{J} \left\{ \frac{f(y \mid x; \theta) g(x) 1_{\mathcal{S}_j}(y, x)}{Q_j(\theta, G)} \right\}^{\delta_j} \right]^r \prod_{j=1}^{J} Q_j(\theta, G)^{\delta_j} \prod_{j=1}^{J} (p_j^r q_j^{1-r})^{\delta_j} \\ &\equiv \left\{ \prod_{j=1}^{J} h_j(x, y; \theta, G)^{\delta_j} \right\}^r \prod_{j=1}^{J} Q_j(\theta, G)^{\delta_j} \prod_{j=1}^{J} (p_j^r q_j^{1-r})^{\delta_j}; \end{aligned} \tag{3}$$

here the densities h_j are the biased-sampling densities corresponding to the jth strata $\mathcal{S}_j$, $j = 1, \ldots, J$, and $\delta_j \equiv 1(S = j)$.

To obtain Maximum Likelihood Estimators (MLEs) of (θ, G), we begin with an "empirical" version of the log of the likelihood in (2), meaning that we simply replace the density term $g(X_i)$ by its point-mass equivalent $G\{X_i\}$:

$$\begin{aligned} & n^{-1} \log L_n(\theta, G) \\ &= \frac{1}{n} \sum_{j=1}^{J} \left[\sum_{i:R_i=1,\, S_i=j} \{\log f(Y_i \mid X_i, \theta) + \log G\{X_i\}\} + (N_j - n_j) \log Q_j(\theta, G) \right] \\ &= \mathbb{P}_n \{ R \log f_\theta(Y \mid X) + R \log G\{X\} + (1 - R) \log Q_S(\theta, G) \}. \end{aligned} \tag{4}$$

Here, $\mathbb{P}_n$ is the empirical measure of the observations (R_i, T_i), and

$$\mathbb{P}_n f(R,T) \equiv n^{-1} \sum_{i=1}^{n} f(R_i, T_i).$$

In this paper, we consider both the absolute maximizer $(\hat{\theta}^a, \hat{G}^a)$ of the likelihood and the restricted MLE $(\hat{\theta}, \hat{G})$, defined through maximizing the likelihood over all pairs (θ, G) such that G concentrates on the set $\{X_i : R_i = 1\}$. The restricted MLE always exists under the continuity conditions imposed in Section 3, but the absolute maximizer $(\hat{\theta}^a, \hat{G}^a)$ (the "true MLE") may not, as shown in Example 2 below. Proposition 2 gives sufficient conditions for existence of the "true MLE" $(\hat{\theta}^a, \hat{G}^a)$. It is also shown below (see the paragraph before Lemma A.1) that the true MLE $\hat{G}^a$, if it exists, concentrates on the set $\{X_i : R_i = 1\} \cup \{x : \hat{s}_n(x) = 0\}$, where the functions $\hat{s}_n$ are defined in (8) and converge uniformly to a strictly positive function. Thus the two types of estimators are the same asymptotically, when they both exist.

Example 2. Suppose that $Y \in \{0,1\}$ with $f(y \mid x) \equiv \mathrm{P}(Y = 1 \mid X = x) \equiv p(x)$ decreasing in $x \in R^+$. Let the partition consist of the sets $\mathcal{S}_{ij} = \{(x,y) : x \in \mathcal{A}_i,\ y = j\}$ for $i, j \in \{0,1\}$, $\mathcal{A}_0 = (0,1]$, $\mathcal{A}_1 = (1,\infty)$. Suppose that we observe the following sample of size $n = 2$: $(R_1, T_1) = (1, 1, 1/2, \mathcal{S}_{0,1})$ and $(R_2, T_2) = (0, \mathcal{S}_{1,1})$. Then the log-likelihood becomes

$$\log p(1/2) + \log G\{1/2\} + \log \int_{\mathcal{A}_1} p(x)\, dG(x).$$

It is clear that the maximizer $\hat{G}^a$ must put positive mass at $x = 1/2$ and also on the interval $\mathcal{A}_1 = (1,\infty)$. Because p is decreasing, the best location in $\mathcal{A}_1$ is "leftmost as possible", but $\mathcal{A}_1 = (1,\infty)$ does not contain its left endpoint. Thus the supremum over G is not attained. On the other hand, the restricted MLE $\hat{G}$ of G clearly assigns mass 1 to $x = 1/2$. If the partition is changed to $\mathcal{A}_0 = (0,1)$ and $\mathcal{A}_1 = [1,\infty)$, then the absolute MLE would be $\hat{G}^a\{1/2\} = \hat{G}^a\{1\} = 1/2$ for the given observations, but a similar problem would arise for other observations. The monotonicity of p together with the strata not containing their boundary are the main causes of the problem here. □

To formulate our result concerning the absolute maximizer $(\hat{\theta}^a, \hat{G}^a)$ on the positive side, suppose that we reorder the data so that $R_{(1)} = \cdots = R_{(m)} = 1$, $R_{(m+1)} = \cdots = R_{(n)} = 0$, with $S_{(i)}$, $X_{(i)}$ denoting the corresponding S's and X's. Then maximization of the log-likelihood over distributions G can be carried out in two steps:

STEP 1. For fixed $\gamma = (\gamma_1, \ldots, \gamma_m)$ with $\gamma_i \geq 0$ and $\sum \gamma_i \leq 1$, and fixed θ, maximize

$$\mathbb{P}_n(1-R)\log Q_S(\theta, G) = \frac{1}{n} \sum_{i=m+1}^{n} \log Q_{S_{(i)}}(\theta, G) \tag{5}$$

over sub-probability measures G satisfying $G\{X_{(i)}\} \geq \gamma_i$, $i = 1, \ldots, m$. (A sub-probablity measure is just a measure with total mass less than or equal to one.) Let the resulting maximum be denoted by $m(\gamma, \theta)$. Setting $M_j \equiv \sum_{i=m+1}^{n} 1(S_{(i)} = j)$, this problem is equivalent to maximizing $n^{-1} \sum_{j=1}^{J} M_j \log v_j$ over $(v_1, \ldots, v_J) \in \mathcal{V}_\gamma$ with

$$\mathcal{V}_\gamma \equiv \left(1(M_1 > 0) \sum_{i=1}^{m} \gamma_i Q_1^*(X_{(i)}, \theta), \ldots, 1(M_J > 0) \sum_{i=1}^{m} \gamma_i Q_J^*(X_{(i)}, \theta)\right) + \left(1 - \sum_{i=1}^{m} \gamma_i\right) \mathcal{W}$$

and $\mathcal{W} = \{(Q_1(\theta, G)1(M_1 > 0), \ldots, Q_J(\theta, G)1(M_J > 0)) : G$ is a sub-probability measure on $\mathcal{X}\}$.

STEP 2. Maximize the function $n^{-1}\sum_{i=1}^{m}\log\gamma_i + m(\gamma,\theta)$ over all sub-probability vectors $\gamma = (\gamma_1,\ldots,\gamma_m)$. Let the resulting maximum be denoted by

$$n^{-1}\sum_{i=1}^{m}\log\gamma_i(\theta) + m\{\gamma(\theta),\theta\}.$$

STEP 3. Maximize the function $r_n(\theta) \equiv \mathbb{P}_n\{R\log f_\theta(Y\mid X)\} + n^{-1}\sum_{i=1}^{m}\log\gamma_i(\theta) + m\{\gamma(\theta),\theta\}$ over $\theta\in\Theta$.

PROPOSITION 2. *If $\mathcal{V}_\gamma$ (equivalently $\mathcal{W}$) is compact, then Step 1 has a solution, and the unrestricted maximum of* (4) *over distributions G on $\mathcal{X}$ exists (for each fixed θ). If the continuity hypothesis C2 of Section 3 holds, then the unrestricted maximum of* (4) *over pairs (θ,G) with $\theta\in\Theta$ and G a distribution on $\mathcal{X}$ exists.*

Proof. That Step 1 has a solution if $\mathcal{V}_\gamma$ is compact follows from the continuity of the right side of (5) in $(v_1,\ldots,v_J)$. Since $\mathcal{V}_\gamma$ depends continuously on γ, it follows that $m(\gamma,\theta)$ is a continuous function of γ (on the set where it is $>-\infty$). Thus Step 2 has a solution, and we conclude that the unrestricted maximum over G exists for each fixed θ. Finally, under the continuity hypothesis C2 the function $\theta\to r_n(\theta)$ defined in Step 3, which is equal to the supremum of the likelihood over G, is a supremum of upper semi-continuous functions of θ. Hence it is upper semi-continuous and attains its maximum on the compact set Θ. □

Remark 1. Note that compactness of $\mathcal{W}$ is implied by compactness of

$$\mathcal{W}_0 \equiv \{\alpha(1(M_1>0)Q_1^*(x,\theta),\ldots,1(M_J>0)Q_J^*(x,\theta)) : x\in\mathcal{X},\ 0\le\alpha\le 1\}.$$

Note that compactness of $\mathcal{W}$ and $\mathcal{W}_0$ fails in Example 2. The compactness hypothesis of Proposition 2 will typically fail for general configurations of the data with strata based on continuous covariates because of the same type of difficulties at the boundaries exhibited there, and the "true" (or unrestricted) MLE may fail to exist for problems with continuous covariates. On the other hand the compactness hypothesis will typically hold when the strata are based on discrete covariates.

Failure of the unrestricted MLE to exist, such as in Example 2, is clearly a problem of strata not containing their boundary. This problem is not resolved given more observations. A compromise between restricted and unrestricted maximization would be to maximize over all discrete G supported on the observed and suitably chosen points close to the boundaries of the strata. This type of restricted MLE will typically exist and if it does, our proof in the Appendix shows that it is consistent.

To examine the restricted MLE $(\hat\theta,\widehat{G})$ where $\widehat{G}$ concentrates on $\{X_i : R_i=1,\ i=1,\ldots,n\}$, let $\mathbb{H}_n$ be the empirical measure of the observed X_i's, viz.

$$\mathbb{H}_n \equiv \frac{\sum_{i=1}^{n} R_i\delta_{X_i}}{\sum_{i=1}^{n} R_i}.$$

Note that $\mathbb{H}_n$ estimates the measure $G_0(\cdot\mid R=1)$: for bounded measurable functions h,

$$\mathbb{H}_n(h) \xrightarrow[\text{a.s.}]{} \frac{G_0(s_0h)}{G_0(s_0)} = G_0(h\mid R=1),$$

where $s_0(x) = \sum_{j=1}^{J} p_jQ_j^*(x,\theta_0)$. By straightforward maximization of $L_n(\theta,G)$ for fixed θ subject to the constraint that the assigned masses add to one, it is easily seen that the restricted

MLE $\widehat{G}$ of G is of the following form: for any measurable set A,

$$\widehat{G}(A) = \frac{\displaystyle\int_A \left[\sum_{j=1}^{J}\left\{1 - \frac{N_j - n_j}{nQ_j(\hat{\theta},\widehat{G})}\right\}Q_j^*(x,\hat{\theta})\right]^{-1} d\mathbb{H}_n(x)}{\displaystyle\int_{\mathcal{X}} \left[\sum_{j=1}^{J}\left\{1 - \frac{N_j - n_j}{nQ_j(\hat{\theta},\widehat{G})}\right\}Q_j^*(x,\hat{\theta})\right]^{-1} d\mathbb{H}_n(x)} \equiv \frac{\mathbb{H}_n(1_A \hat{s}_n^{-1})}{\mathbb{H}_n(\hat{s}_n^{-1})}$$

$$= \mathbb{G}_n(1_A \hat{s}_n^{-1}) = \mathbb{P}_n(R 1_A \hat{s}_n^{-1}) \tag{6}$$

since $\mathbb{H}_n(\hat{s}_n^{-1}) = 1/\overline{R}$. Here $\mathbb{G}_n \equiv n^{-1}\sum_{i=1}^{n} R_i \delta_{X_i} \to_{\text{a.s.}} G_0(s_0 \cdot)$.

Here is a brief summary of Section 2 and what will be addressed in Section 3:

a) The unrestricted or absolute MLE $(\hat{\theta}^a, \widehat{G}^a)$ of (θ, G) exists under a (data dependent) compactness hypothesis which, however, is easily violated for strata defined by continuous covariates, but which will often hold for the case of discrete covariates.

b) The restricted MLE $(\hat{\theta}, \widehat{G})$ with $\widehat{G}$ concentrated on the set $\{X_i : R_i = 1,\ i = 1, \ldots, n\}$ exists under the continuity hypotheses imposed in Section 3, and—as shown by Breslow & Holubkov (1997), Scott & Wild (1998), Lawless, Wild & Kalbfleisch (1999)—can be computed via profile likelihood methods.

c) In Section 3, we will show that both the restricted MLE and the absolute MLE (when the latter exists for all n) are consistent under some reasonable continuity and compactness hypotheses.

3. MAIN RESULTS: IDENTIFIABILITY AND CONSISTENCY

In this section, we state our main results: first, that identifiability of (θ, G) is preserved in the two-phase design if (θ, G) is identifiable in the original regression model and the design sampling satisfies two natural conditions, namely that the strata formed have positive probability of occurring and that items are sampled in each strata with positive probability. Our second main result here is that the semiparametric maximum likelihood estimators studied by Lawless, Wild & Kalbfleisch (1999) and Scott & Wild (1998) are consistent.

We say that (θ, G) are *identifiable in the model* $\mathcal{P}$ (with $\mathcal{P} \equiv \{P_{\theta,G} : \theta \in \Theta,\ G \in \mathcal{G}\}$) if $P_{\theta,G} = P_{\theta',G'}$ implies $\theta = \theta'$ and $G = G'$. Here are the conditions for identifiability:

A1. $p_j > 0$ for $j = 1, \ldots, J$.

A2. The pair of parameters (θ_0, G_0) is identifiable in the model

$$\mathcal{Q} = \{Q_{\theta,G} : dQ_{\theta,G}/d(\nu \times \mu) = q(\cdot\,; \theta, G),\ \theta \in \Theta,\ G \in \mathcal{G}\},$$

where $q(y, x; \theta, G) = f(y \mid x; \theta) g(x)$ as in (1).

Without loss of generality, we may assume that

A3. $Q_j(\theta_0, G_0) \in (0, 1)$ for $j = 1, \ldots, J$.

THEOREM 1 (Identifiability). *Suppose that A1 and A2 hold. Then* (θ, G) *is identifiable in the model*

$$\mathcal{P} = \{P_{\theta,G} : dP_{\theta,G}/d(\nu \times \mu) = p(\cdot\,; \theta, G),\ \theta \in \Theta,\ G \in \mathcal{G}\},$$

where $p(\cdot\,; \theta, G)$ *is given by* (3).

Our goal is to establish consistency of the semiparametric MLEs under natural conditions. Although the models we are considering are quite closely related to those treated by van der Vaart & Wellner (1992) (they are exactly the same if θ is known), the sufficient conditions for consistency given there fail completely for our situation. In particular, requirement (3.3) on p. 138 of van der Vaart & Wellner (1992) fails in our current setting. However, a different approach will yield consistency in the present case.

Here are our assumptions:

C1. $\mathcal{X}$ is a semi-metric space that has a completion that is compact and contains $\mathcal{X}$ as a Borel set.

C2. The maps $(\theta, x) \mapsto Q_j^*(x, \theta)$ are uniformly continuous, and $\theta \mapsto f_\theta(y \mid x)$ are upper semicontinuous for all $(y, x) \in \mathcal{Y} \times \mathcal{X}$.

C3. Θ is a compact metric space.

C4. $P_0\{\sup_{\theta \in \Theta} \log(f_\theta / f_{\theta_0})(y \mid x)\} < \infty$.

C5. A1–A3 hold; it follows that (θ_0, G_0) is identifiable in the two-phase sampling model.

THEOREM 2 (Consistency of $(\hat{\theta}_n, \widehat{G}_n)$). ***Suppose that C1–C5 hold. Then*** $\hat{\theta}_n \to_{a.s.} \theta_0$. ***Furthermore,*** $\sup_{h \in \mathcal{H}} |(\widehat{G}_n - G_0)h| \to_{a.s.} 0$ ***for every GC-class*** $\mathcal{H}$ ***that is bounded in*** $L_1(G_0)$.

In this theorem, $\widehat{G}_n$ may be both the unrestricted MLE (if this exists for all n) or a restricted MLE. The proof of this result proceeds via a series of lemmas; these are given in the Appendix. The main idea is to use the likelihood equations to find an alternative expression for the point masses in the likelihood. Next, we carry through a Wald-type proof after substituting this alternative expression in the log-likelihood. A main technical problem is to induce enough integrability so that a uniform law of large numbers can be applied. For this, we will work not with the log-likelihood itself, but a transformed likelihood.

Because the functions $x \mapsto Q_j^*(x, \theta)$ vanish outside the projections $\mathcal{S}_j^*$ of the partitioning sets, the continuity of these maps assumed in (C2) may appear problematic: we would not want to assume that the probabilities $Q_j^*(x, \theta)$ tend to zero at the boundaries of the partitioning sets. This problem disappears by a proper choice of the metric on the set $\mathcal{X}$. In fact, because we are free to choose this metric, the continuity condition may be considered as just a convenient language to express a regularity condition. In many cases, a "natural" metric on $\mathcal{X}$ should not be the metric of choice for the application of our theorem. We illustrate this with two examples.

Example 3. Functions on $\mathcal{X} = R$ that are right continuous with left limits and that jump at at most finitely many fixed points $a_1 < \cdots < a_m$ will be uniformly continuous relatively to a metric of the form $\rho(x_1, x_2) = |F(x_1) - F(x_2)|$, where F is a cumulative distribution function with a positive jump at every of the points $a_1, \ldots, a_m$ that is strictly increasing and continuous on each of the intervals $(-\infty, a_1), [a_1, a_2), \ldots$.

In fact, under ρ the real line is topologically identical to the union of the disconnected sets $(-\infty, a_1), [a_1, a_2), \ldots$, which are at positive distances:

$$\rho\{[a_{i-1}, a_i), [a_i, a_{i+1})\} = |F(a_i-) - F(a_i)| > 0.$$

Uniform continuity relative to ρ means uniform continuity on each of the intervals $[-\infty, a_1), [a_1, a_2), \ldots$ relative to the Euclidean topology. A function that is càdlàg with jumps only at $a_1, a_2, \ldots, a_m$ is ρ-uniformly continuous. This follows because the function z_i defined by

$$z_i(t) = z(t), \qquad a_i \le t < a_{i+1}, \qquad z_i(a_{i+1}) = z(a_{i+1}-)$$

is continuous on the compact $[a_i, a_{i+1}]$ relative to the Euclidean topology and hence uniformly continuous by continuity of F. The completion of R under ρ is topologically identical to the union of the disconnected compact intervals $[a_i, a_{i+1}]$, and hence compact. It follows that condition (C1) and the first part of (C2) are satisfied relative to ρ as soon as the maps $(\theta, x) \mapsto Q_j^*(x, \theta)$ are uniformly continuous relative to the natural metrics on each of the sets $\Theta \times [a_i, a_{i+1})$. Jumps at the cell boundaries are permitted. □

We present our second example of a choice of a metric to satisfy C1–C2 as a lemma. It applies to the situation that the elements $\mathcal{S}_j$ of the partition are product sets in $\mathcal{Y} \times \mathcal{X}$. Recall that a Polish topological space is a separable topological space that can be metrized by a complete metric. Examples include open subsets, closed subsets or half-open intervals in Euclidean space.

LEMMA 1. *Suppose that the elements $\mathcal{S}_j$ of the partition are product sets $\mathcal{Y}_{m'} \times \mathcal{X}_m$ for sets $\mathcal{X}_m$ that are Polish topological spaces. Furthermore, assume that the maps $(\theta, x) \to f(y \mid x; \theta) = f_\theta(y \mid x)$ are uniformly continuous for all $y \in \mathcal{Y}$ and that the densities $f(\cdot \mid x; \theta)$ for $(x, \theta) \in \mathcal{X} \times \Theta$ are equi-integrable in $L_1(\nu)$. Then C1–C2 are satisfied.*

Proof. The set $\mathcal{X}$ is the finite union $\mathcal{X} = \bigcup_m \mathcal{X}_m$ of the partitioning sets $\mathcal{X}_m$. For each $\mathcal{X}_m$ there exists a metric d_m under which $\mathcal{X}_m$ is totally bounded. (See, e.g., Dudley 1989, Theorem 2.8.2.) We can assume without loss of generality that $d_m \leq 1$ for every m. Then defining $d(x, y) = 1$ if x, y do not belong to the same $\mathcal{X}_m$, and $d(x, y) = d_m(x, y)$ if $x, y \in \mathcal{X}_m$, we obtain a metric on $\mathcal{X}$, under which $\mathcal{X}$ is totally bounded. Every $(\mathcal{X}_m, d_m)$ possesses a compact completion $(\overline{\mathcal{X}}_m, \bar{d}_m)$ in which $\mathcal{X}_m$ is contained as a G_δ set and hence is a Borel set. (See, e.g., Dudley 1989, Theorem 2.5.4.) The completion of $(\mathcal{X}, d)$ is the union $\overline{\mathcal{X}} = \bigcup_m \overline{\mathcal{X}}_m$ equipped with the metric $\bar{d}$ that is defined from the metrics $\bar{d}_m$ as d is defined from the metrics d_m. Thus we have verified C1.

The maps $\theta \to f_\theta(y \mid x)$ are continuous by assumption. Thus for C2 it suffices to show that the maps $(\theta, x) \to Q_j^*(x, \theta)$ are uniformly continuous. Here, for $\mathcal{S}_j = \mathcal{Y}_{m'} \times \mathcal{X}_m$ we have

$$Q_j^*(x, \theta) = \int_{\mathcal{Y}_{m'}} f_\theta(y \mid x)\, d\nu(y) 1_{\mathcal{X}_m}(x).$$

Fix $\varepsilon \in (0, 1)$. If $d(x_1, x_2) < \varepsilon$, then x_1 and x_2 belong to the same partitioning set. So either $Q_j^*(x_i, \theta) = 0$ for $i = 1, 2$, or $Q_j^*(x_i, \theta) = \int_{\mathcal{Y}_{m'}} f_\theta(y \mid x_i)\, d\nu(y)$ for $i = 1, 2$. Thus it suffices to show that the maps $(\theta, x) \to \int_{\mathcal{Y}_{m'}} f_\theta(y \mid x)\, d\nu(y)$ are uniformly continuous. If $d\{(\theta, x), (\theta', x')\} \to 0$, then by assumption $f_\theta(y \mid x) - f_{\theta'}(y \mid x') \to 0$ for every y, and hence $\int |f_\theta(y \mid x) - f_{\theta'}(y \mid x')|\, d\nu(y) \to 0$ by equi-integrability. □

4. EXAMPLES

Here are a few concrete examples to which our results apply.

Example 4 (Binary response logistic regression). This is as in Example 1 of the introduction. Note that when there are just two cells $\mathcal{S}_j$, $j = 1, 2$, corresponding to $Y = 0$ and $Y = 1$ respectively, then this example is (random sample size version of) the classical "case-control" design often used in epidemiological studies; see, e.g., Prentice & Pyke (1979).

Example 5 ("Polychotomous" logistic regression). Suppose that

$$Y \mid X = x \sim \text{Multinomial}_K\{1, (p_1(x, \theta), \ldots, p_K(x, \theta))\}$$

where

$$p_k(\theta, x) = \frac{\exp(\mu_k + \nu_k' x)}{\sum_{k'=1}^K \exp(\mu_{k'} + \nu_{k'}' x)}, \qquad k = 1, \ldots, K;$$

and where we take $\mu_K = 0$ and $\nu_K = 0$ for identifiability. Here $\theta = (\nu, \mu) \in R^K \times R^{dK}$, and $x \in \mathcal{X} \subset R^d$. This is exactly the parametrization used by Prentice & Pyke (1979) in their discussion of the case-control version of this model. One natural choice of strata is given by

$$\mathcal{X} \times \mathcal{Y} = \bigcup_{l=1}^{L} \bigcup_{k=1}^{K} \mathcal{A}_l \times \mathcal{B}_k \equiv \bigcup_{j=1}^{J} \mathcal{S}_j$$

into $J = LK$ sets where $\mathcal{X} = \bigcup_{l=1}^{L} \mathcal{A}_l$ as in Example 1, and $\mathcal{Y} = \{1\} \cup \cdots \cup \{K\} \equiv \mathcal{B}_1 \cup \cdots \cup \mathcal{B}_K$ is shorthand for the K-coordinate vectors $(1, 0, \ldots, 0), \ldots, (0, \ldots, 0, 1)$.

Example 6 (Bivariate logistic or Palmgren model). Suppose that

$$Y \mid X = x = (Y_1, Y_2) \mid X = x \sim \text{Bivariate Logistic}\,(\pi_{kl}(x),\ k, l \in \{0, 1\})$$

where $\pi_{1\cdot} = \pi_{11} + \pi_{10}$, $\pi_{\cdot 1} = \pi_{11} + \pi_{01}$, $\psi = (\pi_{11}\pi_{00})/(\pi_{10}\pi_{01})$ are modelled as follows:

$$\text{logit}\{\pi_{1\cdot}(x)\} = \theta_1 x_{(1)}, \quad \text{logit}\{\pi_{\cdot 1}(x)\} = \theta_2 x_{(2)}, \quad \log\{\psi(x)\} = \theta_3 x_{(3)},$$

where the coordinates of $x_{(j)}$ are a subset of the coordinates of $x \in R^p$ for $j = 1, 2, 3$. Let $d_j = \text{dimension}\,(x_{(j)})$, $d = d_1 + d_2 + d_3$, $\theta = (\theta_1, \theta_2, \theta_3)$. This is a model introduced by Palmgren (1989). A natural choice of the strata would be

$$\mathcal{X} \times \mathcal{Y} = \bigcup_{l=1}^{L} \bigcup_{k=1}^{4} \mathcal{A}_l \times \mathcal{B}_k \equiv \bigcup_{j=1}^{J} \mathcal{S}_j$$

into $J = 4L$ sets and $\mathcal{Y} = \{(1,1)\} \cup \{(1,0)\} \cup \{(0,1)\} \cup \{(0,0)\} \equiv \mathcal{B}_1 \cup \cdots \cup \mathcal{B}_4$.

Example 7 (Continuous response variable Y). Suppose that $Y \mid X = x \sim \text{N}(\nu' x, \sigma^2)$ so that $\theta = (\nu, \sigma^2) \in R^{d+1}$ and

$$f_\theta(y \mid x) = (2\pi)^{-1/2} \exp\left\{-(y - \nu' x)^2 / 2\sigma^2\right\}.$$

One simple possibility for defining strata is

$$\mathcal{S}_1 = \{(y, x) : y \le c(x)\}, \qquad \mathcal{S}_2 = \{(y, x) : y > c(x)\},$$

where $c\colon R^d \to R$ is a fixed (and known) function of x. Note that this gives an example in which the response variable Y is continuous rather than discrete, and in which the strata are not necessarily product sets. This is the type of model used by Scott & Wild (1998) in their Example 3 involving analysis of low birthweights: in their example Y is the birthweight, the first component X_1 of X is "age", and $c(X) = d(X_1)$ for a known function d.

Example 8 (Censored data). Suppose that Y represents a survival time, so the densities $f_\theta(y \mid x)$ are all concentrated on $R^+ = [0, \infty)$. Suppose that C is a censoring variable, and Y and C are conditionally independent given X. Suppose that we consider $Z = (C, X)$ as a new "augmented" covariate vector, and two strata are defined by

$$\mathcal{S}_1 = \{(y, c, x) = (y, z) : y \le c\}, \qquad \mathcal{S}_2 = \{(y, c, x) = (y, z) : y > c\}.$$

Frequently in this type of model the "cases", or observations falling in stratum 1 are relatively rare so we would take $p_1 = 1$ and $p_2 < 1$; i.e., obtain complete data for the "true failures", and sample a fraction p_2 of the censored individuals. This is Example 3 of Lawless, Wild & Kalbfleisch (1999). Our results do apply in this model (assuming that $f_\theta(y \mid x)$ satisfies our regularity hypotheses). It is closely related to "case-cohort" sampling in biostatistics (see Prentice 1986; Self & Prentice 1988), but in the typical case-cohort model the parametric model $f_\theta(y \mid x)$ would be replaced by a semiparametric model, usually the Cox proportional hazards model $f_{\theta,\lambda}(y \mid x)$ depending on an additional infinite-dimensional parameter λ, the baseline hazard rate, and hence the results obtained here do not apply to estimation in the usual model for case-cohort data.

5. DISCUSSION AND FURTHER PROBLEMS

In this paper, we have discussed consistency only under the Bernoulli sampling scheme (VPS1). Consistency, asymptotic normality, and efficiency of the estimators under the important (and often used) stratified sampling plan (BSS) remains to be studied.

An interesting related model is the biased sampling model studied by Gilbert (2000). His model is very similar to ours, but a key difference is that the sampling is only from J different biased samples from G where the biasing kernels (our functions $f(y \mid x;\theta)1\{(y,x) \in \mathcal{S}_j\}$) are allowed to depend on θ. This yields a density given by just the first term of the first line of (3), and hence the information about the weights (strata probabilities) given by the δ's (or S's) is missing in Gilbert's problem. This makes his problem different and somewhat more difficult. Nevertheless, some of the methods used here might be useful for alternative proofs of consistency of maximum likelihood estimators in Gilbert's model.

As mentioned briefly in Example 8, the models covered here necessarily involve a parametric regression model: the conditional density $f_\theta(y \mid x)$ is assumed to be parameterized by a finite-dimensional parameter θ. It would be of considerable interest to extend the present results to cases in which the conditional density for the regression model is also allowed to depend on an infinite-dimensional parameter λ.

APPENDIX: PROOFS FOR SECTION 3

Proof of Theorem 1. One easy proof of this is via consistency of the Horovitz–Thompson estimators. Ignoring the δ's (or S's) and writing $Z = (Y, X)$, then for bounded measurable functions h it follows from A1 that

$$\frac{1}{n}\sum_{i=1}^{n} \frac{h(Z_i)}{\pi(Z_i)} R_i \to \iint h(z) f_\theta(y \mid x)\, dG(x)\, d\nu(y) \qquad \text{almost surely } P_{\theta,G}.$$

Thus if $P_{\theta_1,G_1} = P_{\theta_2,G_2}$, we conclude that

$$\iint h(z) f_{\theta_1}(y \mid x)\, dG_1(x)\, d\nu(y) = \iint h(z) f_{\theta_2}(y \mid x)\, dG_2(x)\, d\nu(y)$$

for all bounded measurable functions h. By A2 this implies that $(\theta_1, G_1) = (\theta_2, G_2)$. □

The proof of Theorem 2 proceeds via a series of lemmas. The likelihood equations for G were given previously in (6). We can rewrite them in the form

$$\widehat{G}\{h1(\hat{s}_n > 0)\} = \mathbb{P}_n\left(\frac{R}{\hat{s}_n} h\right) \tag{7}$$

for the function

$$\hat{s}_n(x) = 1 - \mathbb{P}_n^{R,S}\left\{\frac{(1-R)Q_S^*(x,\hat{\theta})}{Q_S(\hat{\theta},\widehat{G})}\right\}. \tag{8}$$

Here $\mathbb{P}_n^{R,S} f(R,S,x)$ means expectation of f over (R,S) with x fixed. An alternative way to obtain (7) is to insert $\hat{\theta}$ and the probability measures G_t defined by $dG_t = \{1 + t(h - \widehat{G}h)\}\, d\widehat{G}$ into the likelihood. Then it is maximized with respect to t at $t = 0$. Setting the derivative at $t = 0$ equal to zero and rearranging the equation yields $\widehat{G}(h\hat{s}_n) = \mathbb{P}_n Rh$. We deduce (7) in view of the following lemma.

LEMMA A.1. $\hat{s}_n(X_i) \geq 1/n > 0$ *for every* X_i *with* $R_i = 1$.

Proof. The equation $\widehat{G}(h\hat{s}_n) = \mathbb{P}_n Rh$ evaluated at $h = 1_{\{X_i\}}$ yields $\widehat{G}\{X_i\}\hat{s}_n(X_i) = 1/n$ if $R_i = 1$. Of course $\widehat{G}\{X_i\} \leq 1$. □

It appears to be unclear if $\hat{s}_n(x)$ is nonnegative for every x, but in view of Lemma A.1, we can replace $\hat{s}_n$ by $\hat{s}_n^+$ in (7) without changing the likelihood equations. We shall do this later on without further comment.

LEMMA A.2. $\min_{1\le j\le J: p_j<1} Q_j(\hat{\theta}, \widehat{G})$ *is bounded away from zero a.s. as* $n \to \infty$.

Proof. Let $\mathcal{J}_+ \equiv \{j \in \{1, \ldots, J\} : p_j < 1\}$; note that the cells $\mathcal{S}_j$ corresponding to $p_j = 1$ do not enter in the "first phase" contributions to the log-likelihood represented by the third term in (4). Consider

$$\begin{aligned}
&\sup_{(\theta,G):\min_{j\in\mathcal{J}_+} Q_j(\theta,G)<\varepsilon} n^{-1}\log L_n(\theta,G)\\
&\quad\le \sup_\theta \mathbb{P}_n\{R\log f(Y\mid X;\theta)\} + \sup_G \mathbb{P}_n(R\log G\{X\})\\
&\qquad + \sup_{(\theta,G):\min_{j\in\mathcal{J}_+} Q_j(\theta,G)<\varepsilon}\left[\sum_{j=1}^J \mathbb{P}_n\left\{(1-R)1_{[S=j]}\right\}\log Q_j(\theta,G)\right]\\
&\quad\le \sup_\theta \mathbb{P}_n\{R\log f(Y\mid X;\theta)\} - (\mathbb{P}_nR)\log(n\mathbb{P}_nR) + \min_{j\in\mathcal{J}_+}\mathbb{P}_n\left\{(1-R)1_{[S=j]}\right\}\log\varepsilon.
\end{aligned}$$

For $(\theta_0, \mathbb{P}_n(R\,\cdot)/\mathbb{P}_n(R))$, the log-likelihood is

$$\begin{aligned}
n^{-1}\log L_n(\theta_0, \mathbb{P}_n(R\,\cdot)/\mathbb{P}_n(R)) &= \mathbb{P}_n\{R\log f(Y\mid X;\theta_0)\} - (\mathbb{P}_nR)\log(n\mathbb{P}_nR)\\
&\quad + \left[\sum_{j=1}^J \mathbb{P}_n\left\{(1-R)1_{[S=j]}\right\}\log\left\{\frac{\mathbb{P}_nRQ_j^*(\,\cdot\,,\theta_0)}{\mathbb{P}_nR}\right\}\right].
\end{aligned}$$

The difference is bounded above by

$$\begin{aligned}
&\sup_\theta \mathbb{P}_n\left\{R\log\frac{f(Y\mid X;\theta)}{f(Y\mid X;\theta_0)}\right\} + \min_{j\in\mathcal{J}_+}\mathbb{P}_n\left\{(1-R)1_{[S=j]}\right\}\log\varepsilon\\
&\qquad - \sum_{j=1}^J\left[\left\{\mathbb{P}_n(1-R)1_{[S=j]}\right\}\log\left(\frac{\mathbb{P}_nRQ_j^*(\,\cdot\,,\theta_0)}{\mathbb{P}_nR}\right)\right],
\end{aligned}$$

where the third term converges to a finite quantity not depending on ε. Since the first term does not depend on ε and is $O(1)$ a.s. by C4, and

$$\min_{j\in\mathcal{J}_+}\mathbb{P}_n\left\{(1-R)1_{[S=j]}\right\} \underset{\text{a.s.}}{\longrightarrow} \min_{j\in\mathcal{J}_+}(1-p_j)Q_j(\theta_0,G_0) > 0,$$

it follows that for sufficiently small $\varepsilon > 0$ the difference is strictly negative eventually, almost surely. Hence the maximum likelihood estimator cannot remain inside the set of (θ, G) such that $\min_{j\in J_+} Q_j(\theta, G) < \varepsilon$ for such small ε. □

Assume from now on without loss of generality that $\min_{1\le j\le J} Q_j(\hat{\theta}_n, \widehat{G}_n) \ge \varepsilon > 0$ for some ε. Define

$$s(x,\theta,G) = 1 - \sum_{j=1}^J (1-p_j)\frac{Q_j(\theta_0,G_0)}{Q_j(\theta,G)}\,Q_j^*(x,\theta).$$

LEMMA A.3. $\hat{s}_n(x) = s(x, \hat{\theta}_n, \widehat{G}_n) + o(1)$ ***a.s. uniformly in*** x.

Proof. We have

$$\hat{s}_n(x) = 1 - \sum_{j=1}^{J} \frac{\mathbb{P}_n(1-R)1(S=j)}{Q_j(\hat{\theta}, \widehat{G})} Q_j^*(x, \hat{\theta}_n).$$

Because $Q_j(\hat{\theta}_n, \widehat{G}_n)$ is bounded away from 0 by Lemma A.2, Lemma A.3 is immediate from the strong law and $P_0(1-R)1(S=j) = (1-p_j)Q_j(\theta_0, G_0)$. □

Suppose for the moment that we can prove that, in an almost sure sense, $\hat{\theta} \to \theta_0$ and $Q_j(\hat{\theta}, G) \to Q_j(\theta_0, G_0)$ for $j = 1, \ldots, J$. Then by the preceding lemma and C2 we have that

$$\begin{aligned} \hat{s}_n(x) &= s(x, \hat{\theta}_n, \widehat{G}_n) + o(1) \\ &\equiv 1 - \sum_{j=1}^{J} (1-p_j) \frac{Q_j(\theta_0, G_0)}{Q_j(\hat{\theta}, \widehat{G})} Q_j^*(x, \hat{\theta}) + o(1) \\ &\to s_0(x) \equiv s(x, \theta_0, G_0) = \sum_{j=1}^{J} p_j Q_j^*(x, \theta), \end{aligned}$$

because $\sum_j Q_j^*(x, \theta) = 1$. The convergence in this display is uniform in x. Then we can conclude the proof of Theorem 2 by deducing the consistency of $\widehat{G}$ from the likelihood equations, as follows. By (7),

$$\widehat{G}h = \mathbb{P}_n \frac{R}{\hat{s}_n} h = \mathbb{P}_n \frac{R}{s_0} h + o(1) \to P_0 \frac{R}{s_0} h = G_0 h,$$

uniformly in h in a Glivenko–Cantelli class such that $P_0(R/s_0)|h| = G_0|h|$ is uniformly bounded in h.

Thus the proof of Theorem 2 is complete once we have proved that $\hat{\theta}$ and the probabilities $Q_j(\hat{\theta}, \widehat{G})$ are consistent estimators. This turns out to be the hard part of the proof, and is the subject of the remainder of the paper.

By C5 and the fact that $\sum_j Q_j^*(x, \theta) = 1$, the function $s_0(x) \equiv s(x, \theta_0, G_0)$ is bounded away from 0. Thus for $0 < \lambda < 1$ we can define a probability measure $\widetilde{G}_\lambda$ by

$$\widetilde{G}_\lambda h = \frac{\mathbb{P}_n \left\{ R \left(\frac{\lambda}{\hat{s}_n} + \frac{1-\lambda}{s_0} \right) h \right\}}{\mathbb{P}_n \left\{ R \left(\frac{\lambda}{\hat{s}_n} + \frac{1-\lambda}{s_0} \right) \right\}}.$$

In view of (7) and the fact that $\mathbb{P}_n R(1/\hat{s}_n) = 1$,

$$\frac{d\widehat{G}}{d\widetilde{G}_\lambda} = \frac{1/\hat{s}_n}{\left(\frac{\lambda}{\hat{s}_n} + \frac{1-\lambda}{s_0} \right) / \left\{ \lambda + (1-\lambda)\mathbb{P}_n \left(R \frac{1}{s_0} \right) \right\}} = \frac{\lambda + (1-\lambda)\mathbb{P}_n(R/s_0)}{\lambda + (1-\lambda)\frac{\hat{s}_n}{s_0}},$$

and, by the definition of $(\hat{\theta}, \widehat{G})$,

$$\mathbb{P}_n \left[R \log \frac{f_{\hat{\theta}}}{f_{\theta_0}}(y \mid x) + R \log \left\{ \frac{\lambda + (1-\lambda)\mathbb{P}_n(R/s_0)}{\lambda + (1-\lambda)\hat{s}_n/s_0} \right\} + (1-R) \log \frac{Q_S(\hat{\theta}, \widehat{G})}{Q_S(\theta_0, \widetilde{G})} \right] \geq 0. \quad (9)$$

For fixed $\lambda > 0$, the measures $\widehat{G}_\lambda$ are convenient because the densities $d\widehat{G}/d\widehat{G}_\lambda$ exist and are bounded. Later in the proof, it will be necessary to let $\lambda \to 0$ to conclude consistency.

LEMMA A.4. *Under C2–C3 the class of functions* $\{Q_j^*(\,\cdot\,,\theta) : \theta \in \Theta\}$ *is Glivenko–Cantelli.*

Proof. This follows from finiteness of the L_1-bracketing numbers. See van der Vaart (1998), Example 19.8, p. 272. □

LEMMA A.5. *Suppose that C2–C4 hold. Then, almost surely,*

$$P_0\left[R\log\frac{f_{\hat\theta}}{f_{\theta_0}}(y\,|\,x) + R\log\left\{1\Big/\left(\lambda + (1-\lambda)\frac{s(\,\cdot\,,\hat\theta,\widehat G)^+}{s(\,\cdot\,,\theta_0,G_0)}\right)\right\} + (1-R)\log\frac{Q_S(\hat\theta,\widehat G)}{Q_S(\theta_0,\widetilde{G_\lambda})}\right] \geq -o(1). \tag{10}$$

Proof. The idea is to replace the empirical measure in (9) by P_0 and $\hat s_n$ by $s(\,\cdot\,,\hat\theta,\widehat G)^+$. The latter is permitted in view of Lemma A.3.

By C2-C4 the functions $\{\log(f_\theta/f_{\theta_0}) : \theta \in \Theta\}$ form a one-sided Glivenko–Cantelli class with an integrable upper envelope. See the one-sided Glivenko–Cantelli theorem of Le Cam (1953) as given by Ferguson (1996). Thus we can replace $\mathbb{P}_n$ by P_0 in the first term of (9).

The functions $s(\,\cdot\,,\theta,G)$ are multiples of convex combinations of the functions $Q_j^*(\,\cdot\,,\theta)$, and hence form a GC-class. The transformation to

$$R\log\left\{1\Big/\left(\lambda + (1-\lambda)\frac{s(\,\cdot\,,\theta,G)^+}{s(\,\cdot\,,\theta_0,G_0)}\right)\right\}$$

is continuous and these functions are uniformly bounded. By the GC-preservation theorem of van der Vaart & Wellner (2000), this is also a GC-class. Furthermore $\mathbb{P}_n(R/s_0) \to_{a.s.} P_0(R/s_0) = 1$. Thus we can also replace $\mathbb{P}_n$ by P_0 in the second term of (9).

The third term can be handled by just the ordinary law of large numbers applied to the averages $\mathbb{P}_n(1-R)1_{\mathcal{S}_j}(Y,X)$. The lemma follows. □

Because the functions $\{Q_j^*(\,\cdot\,,\theta) : \theta \in \Theta\}$ form a GC-class, so do the functions

$$\left\{f(R,x;\theta_1,\theta_2,\underline{\lambda}) \equiv R\,\frac{1}{\left(\sum_{j=1}^J \lambda_j Q_j^*(x,\theta_1)\right)^+ + \varepsilon}\,Q_j^*(x,\theta_2) : \theta_1,\theta_2 \in \Theta, \sum_j |\lambda_j| \leq 1,\ \varepsilon \geq \varepsilon_0\right\}$$

for every fixed $\varepsilon_0 > 0$, by the GC-preservation theorem of van der Vaart & Wellner (2000). Thus we have that, almost surely, for every $\varepsilon > 0$,

$$\sup_{\theta_1,\theta_2,\sum|\lambda_j|\leq 1}\left|(\mathbb{P}_n - P)\left[R\frac{Q_j^*(x,\theta_1)}{\left\{\sum_j \lambda_j Q_j^*(x,\theta_1)\right\}^+ + \varepsilon}\right]\right| \to 0.$$

Fix some ω for which this is true, for every $\varepsilon > 0$, and such that (10) is true. By the compactness of Θ and the unit simplex, every subsequence of n has a further subsequence along which, for the chosen ω, and some $Q_{j,\infty}$, $Q_{j,\infty}^0$, and θ_∞, $\hat\theta_n \to \theta_\infty$,

$$Q_j(\hat\theta_n,\widehat G_n) \to Q_{j,\infty} \qquad \text{and} \qquad Q_j(\theta_0,\widehat G_n) \to Q_{j,\infty}^0, \quad j = 1,\ldots,J.$$

LEMMA A.6. *Assume C2–C3 and let $Q_{j,\infty}$, $Q^0_{j,\infty}$, and θ_∞ be constructed as in the preceding discussion. Then*

(i) $P_0(R/s^+_\infty) \le 1$.

(ii) $Q_{j,\infty} \ge Q_j(\theta_\infty, G_\infty)$, $j = 1, \ldots, J$.

(iii) $Q^0_{j,\infty} \ge Q_j(\theta_0, G_\infty)$, $j = 1, \ldots, J$ *for the sub-probability measure G_∞ defined by*:

$$G_\infty h = P_0\left(Rh\frac{1}{s^+_\infty}\right),$$

where

$$s_\infty(x) = 1 - \sum_{j=1}^{J}(1 - p_j)\,\frac{Q_j(\theta_0, G_0)}{Q_{j,\infty}}\,Q^*_j(x, \theta_\infty).$$

Proof. It is clear that, for every x, $\hat{s}_n(x) = s(x, \hat{\theta}, \widehat{G}) + o(1) \to s_\infty(x)$. For every $\varepsilon > 0$, we have

$$1 = \widehat{G}1 \ge \mathbb{P}_n R\frac{1}{\hat{s}^+_n} \ge \mathbb{P}_n R\frac{1}{\hat{s}^+_n + \varepsilon} = P_0\left(R\frac{1}{\hat{s}^+_n + \varepsilon}\right) + o(1) \to P_0\left(R\frac{1}{s^+_\infty + \varepsilon}\right),$$

where the last step follows by dominated convergence, and the equality in the next to last step follows because the functions $\hat{s}_n$ are of the form $\sum_j \lambda_j Q^*_j(x, \theta)$. Letting $\varepsilon \downarrow 0$ yields (i), and hence G_∞ is a well-defined sub-probability measure.

For every $\varepsilon > 0$, we have

$$\begin{aligned} Q_{j,\infty} &= Q_j(\hat{\theta}_n, \widehat{G}_n) + o(1) = \widehat{G}_n Q^*_j(\,\cdot\,, \hat{\theta}_n) \ge \mathbb{P}_n\left\{R\frac{1}{\hat{s}^+_n}Q^*_j(\,\cdot\,, \hat{\theta}_n)\right\} \\ &\ge \mathbb{P}_n\left\{R\frac{1}{\hat{s}^+_n + \varepsilon}Q^*_j(\,\cdot\,, \hat{\theta}_n)\right\} = P_0\left\{R\frac{1}{\hat{s}^+_n + \varepsilon}Q^*_j(\,\cdot\,, \hat{\theta}_n)\right\} + o(1) \\ &\to P_0\left\{R\frac{1}{s^+_\infty + \varepsilon}Q^*_j(\,\cdot\,, \theta_\infty)\right\}. \end{aligned}$$

Letting $\varepsilon \downarrow 0$ yields

$$Q_{j,\infty} \ge P_0\left\{R\frac{1}{s^+_\infty}Q^*_j(\,\cdot\,, \theta_\infty)\right\} = G_\infty Q^*_j(\,\cdot\,, \theta_\infty) = Q_j(\theta_\infty, G_\infty).$$

This gives (ii). Assertion (iii) follows by the same argument, but with $\hat{\theta}_n$ replaced by θ_0. □

If we had equality in (i) of Lemma A.6, then G_∞ would be a probability measure. This does not seem to be automatic. If (i) is an equality, then so are (ii)–(iii), because both left and right sides then sum up to 1. (The left sides always do.)

LEMMA A.7. *Assume C2–C4. Then in the setting as described before Lemma A.6,*

$$P_0\Bigg\{R\log\frac{f_{\theta_\infty}}{f_{\theta_0}}(y \mid x) + R\log\frac{dG_\infty}{d\{\lambda G_\infty + (1-\lambda)G_0\}} + (1-R)\log\frac{Q_{S,\infty}}{\lambda Q^0_{S,\infty} + (1-\lambda)Q_S(\theta_0, G_0)}\Bigg\} \ge 0. \tag{11}$$

Proof. We have

$$\frac{1}{\lambda + (1-\lambda)\dfrac{s(\,\cdot\,,\hat\theta,\widehat G)^+}{s(\,\cdot\,,\theta_0,G_0)}} \to \frac{1}{\lambda + (1-\lambda)\dfrac{s^+_\infty}{s_0}} = \frac{1/s^+_\infty}{\left(\dfrac{\lambda}{s^+_\infty} + \dfrac{1-\lambda}{s_0}\right)} = \frac{dG_\infty}{d(\lambda G_\infty + (1-\lambda)G_0)}$$

by the definition of $G_\infty h = P_0(Rh/s^+_\infty)$ and the fact that $G_0 h = P_0(Rh/s_0)$. Also note that $G_0(s^+_\infty = 0) = 0$ by Lemma A.6(i). These functions are uniformly bounded above and bounded away from zero. Lemma A.7 now follows from Lemma A.5, Fatou's lemma, and dominated convergence, and

$$\begin{aligned} Q_j(\theta_0, \widetilde{G_\lambda}) &= \widetilde{G_\lambda} Q^*_j(\,\cdot\,,\theta_0) = \frac{\lambda \widehat G Q^*_j(\,\cdot\,,\theta_0) + (1-\lambda)\mathbb{P}_n\{R(1/s_0)Q^*_j(\,\cdot\,,\theta_0)\}}{\lambda + (1-\lambda)\mathbb{P}_n(R/s_0)} \\ &= \frac{\lambda Q_j(\theta_0,\widehat G) + (1-\lambda)P_0\{R(1/s_0)Q^*_j(\,\cdot\,,\theta_0)\} + o(1)}{1+o(1)} \\ &\to \lambda Q^0_{j,\infty} + (1-\lambda)Q_j(\theta_0,G_0). \end{aligned}$$

□

By Lemma A.6(ii) and (iii), we have

$$Q_{j,\infty} \ge Q_j(\theta_\infty, G_\infty), \qquad \lambda Q^0_{j,\infty} + (1-\lambda)Q_j(\theta_0,G_0) \ge Q_j\{\theta_0, \lambda G_\infty + (1-\lambda)G_0\}. \quad (12)$$

These are equalities if and only if G_∞ is a probability measure. If these were equalities, then (11) would read

$$P_0\left\{ R\log\frac{f_{\theta_\infty}}{f_{\theta_0}}(y\,|\,x) + R\log\frac{dG_\infty}{d(\lambda G_\infty + (1-\lambda)G_0)} + (1-R)\log\frac{Q_S(\theta_\infty,G_\infty)}{Q_S(\theta_0,\lambda G_\infty + (1-\lambda)G_0)}\right\} \ge 0. \quad (13)$$

This does imply $(\theta_\infty, G_\infty) = (\theta_0, G_0)$ because (θ_0, G_0) is identifiable from the data, and the following lemma.

LEMMA A.8. *If, for probability densities p, q and p_0,*

$$P_0 \log\left\{\frac{p}{\lambda q + (1-\lambda)p_0)}\right\} \ge 0$$

for every $\lambda \in (0,1)$, then $p = p_0$.

Proof. Let $p_\lambda = \lambda q + (1-\lambda)p_0$ and let H denote the Hellinger distance. Because $\log x \le 2(\sqrt{x} - 1)$ for every $x \ge 0$,

$$\begin{aligned} P_0 \log\frac{p}{p_\lambda} &\le 2P_0\left(\sqrt{\frac{p}{p_\lambda}} - 1\right) \\ &= 2P_\lambda\left(\sqrt{\frac{p}{p_\lambda}} - 1\right) + 2\int (p_0 - p_\lambda)\left(\frac{\sqrt{p} - \sqrt{p_\lambda}}{\sqrt{p_\lambda}}\right) \end{aligned}$$

$$
\begin{aligned}
&= -H^2(P_\lambda, P) + 2\lambda \int (p_0 - q)\left(\frac{\sqrt{p} - \sqrt{p_\lambda}}{\sqrt{p_\lambda}}\right) \\
&\le -H^2(P_\lambda, P) + 2\lambda \left\{\int \frac{(p_0 - q)^2}{p_\lambda}\right\}^{1/2} H(P, P_\lambda) \\
&\le -H^2(P_\lambda, P) + 2\lambda \frac{\sqrt{2}}{\sqrt{\lambda(1-\lambda)}} H(P_0, Q) H(P, P_\lambda),
\end{aligned}
$$

because $(p_0 + q)/p_\lambda \le 1/(1-\lambda) + 1/\lambda$. If the left side is nonnegative, then it follows that

$$
H(P, P_\lambda) \le 2\sqrt{2} \sqrt{\frac{\lambda}{1-\lambda}}\, H(P_0, Q).
$$

As $\lambda \downarrow 0$ we have $p_\lambda \to p_0$ pointwise, and $\sqrt{p_\lambda} \le \sqrt{q} + \sqrt{p_0}$ is dominated by a square-integrable function. Conclude that $H(P, P_\lambda) \to H(P, P_0)$. Because the right side converges to 0, we have $H(P, P_0) = 0$. □

The left side of (13) is the expectation in the preceding lemma for p_0, p and q the densities of the data under (θ_0, G_0), $(\theta_\infty, G_\infty)$ and (θ_0, G_∞), respectively. Therefore, we can conclude that *if* G_∞ in Lemma A.6 is a probability measure, then $(\theta_\infty, G_\infty) = (\theta_0, G_0)$.

Finally, consider the case that G_∞ is not a probability measure. Then (13) is not valid, but we have an analogue of it under the following condition: There exists a probability measure $\overline{G}$ such that

$$
\overline{G} \ge G_\infty, \qquad Q_{j,\infty} = Q_j(\theta_\infty, \overline{G}), \qquad Q^0_{j,\infty} = Q_j(\theta_0, \overline{G}), \qquad j = 1, \ldots, J. \tag{14}
$$

Under this condition we have

$$
P_0\left\{ R \log \frac{f_{\theta_\infty}}{f_{\theta_0}}(y \mid x) + R \log \frac{d\overline{G}}{d\{\lambda \overline{G} + (1-\lambda) G_0\}} + (1-R) \log \frac{Q_S(\theta_\infty, \overline{G})}{Q_S(\theta_0, \lambda \overline{G} + (1-\lambda) G_0)} \right\} \ge 0. \tag{15}
$$

To see this, note that

$$
g \to \frac{g}{\lambda g + (1-\lambda) g_0}
$$

is increasing, so that (11) increases if G_∞ is replaced by $\overline{G}$. The resulting inequality can be rewritten in the form (15).

As before, condition (15) implies $\theta_\infty = \theta_0$ and $\overline{G} = G_0$, and hence $Q_{j,\infty} = Q^0_{j,\infty} = Q_j(\theta_0, G_0)$. Thus $\hat{\theta} \to \theta_0$ and $Q_j(\hat{\theta}, \widehat{G}) \to Q_j(\theta_0, G_0)$ for every $j = 1, \ldots, J$.

Condition (14) requires that we can "add the missing mass" to G_∞ such that the numbers $Q_j(\theta_\infty, G_\infty)$ and $Q_j(\theta_0, G_\infty)$ increase to $Q_{j,\infty}$ and $Q^0_{j,\infty}$. Under conditions C1–C5 this is possible. The proof of the following lemma shows that the "missing mass" must actually be orthogonal to G_0. As it "drifts away" and apparently does not cause trouble, we shall allow it to drift "outside $\mathcal{X}$" and next solve (14) for $\overline{G}$ a measure on the completion $\overline{\mathcal{X}}$ of $\mathcal{X}$.

LEMMA A.9. *Assume C1–C2. Then in the setting of Lemma A.6 there exists a probability measure* $\overline{G}$ *on the completion* $\overline{\mathcal{X}}$ *of* $\mathcal{X}$ *such that* (14) *holds, where* $Q_j(\theta, \overline{G}) = \overline{G}\overline{Q}^*_j(\,\cdot\,, \theta)$ *for* $\overline{Q}^*_j(\,\cdot\,, \theta)$ *the continuous extension to* $\overline{\mathcal{X}}$.

Proof. The measures $\widehat{G}, G_0, G_\infty$ can be viewed as Borel measures on $\overline{\mathcal{X}}$. The functions $Q_j^*(\,\cdot\,,\theta)$ and hence $\hat{s}_n$ and $s(\,\cdot\,,\theta,G)$ can be extended to $\overline{\mathcal{X}}$ by continuity. Recall that θ_∞, $Q_{j,\infty}$ and $Q_{j,\infty}^\circ$ are constructed as limits along a subsequence of $\{n\}$. By Prohorov's theorem, there exists a further subsequence along which $\widehat{G}_n \to \overline{G}$ in distribution for some probability measure $\overline{G}$ on $\overline{\mathcal{X}}$. By the equicontinuity of the functions $\theta \to Q_j^*(x,\theta)$ and Lemma A.3,

$$\hat{s}_n(x) = s(x,\hat{\theta},\widehat{G}) + o(1) = 1 - \sum_{j=1}^{J}(1-p_j)\,\frac{Q_j(\theta_0,G_0)}{Q_j(\hat{\theta},\widehat{G})}\,Q_j^*(x,\hat{\theta}) + o(1) \to s_\infty(x),$$

uniformly in $x \in \overline{\mathcal{X}}$. Thus for every continuous function $h\colon \overline{\mathcal{X}} \to R$,

$$\widehat{G}(\hat{s}_n h) = \widehat{G}(\hat{s}_n^+ h) = \widehat{G}(s_\infty^+ h) + o(1) \to \overline{G}(s_\infty^+ h),$$

by the continuity of s_∞^+. On the other hand,

$$\widehat{G}(\hat{s}_n h) = \mathbb{P}_n(Rh) \to P_0(Rh) = G_\infty(s_\infty^+ h),$$

because

$$G_\infty(s_\infty^+ h) = P_0\{(R/s_\infty^+)s_\infty^+ h\} = P_0\{Rh1(s_\infty^+ > 0)\} = P_0(Rh).$$

It follows that

$$\overline{G}(s_\infty^+ h) = G_\infty(s_\infty^+ h).$$

This implies that $\overline{G}$ and G_∞ coincide on the set $\{x \in \overline{\mathcal{X}} : s_\infty^+(x) > 0\}$. Because G_0 and G_∞ are concentrated on this set with G_0 and G_∞ mutually absolutely continuous, $\overline{G}$ is the sum of G_∞ and a measure that is orthogonal to G_0:

$$\overline{G}h = G_\infty h + \overline{G}(h1\{s_\infty^+ = 0\})$$

is the Lebesgue decomposition of $\overline{G}$ relative to G_0. Obviously $\overline{G} \geq G_\infty$. Finally by equicontinuity of $Q_j^*(x,\theta)$, we have

$$\begin{aligned} Q_{j,\infty} &= Q_j(\hat{\theta},\widehat{G}) + o(1) = \widehat{G}Q_j^*(x,\hat{\theta}) + o(1) \\ &= \widehat{G}Q_j^*(x,\theta_\infty) + o(1) \to \overline{G}Q_j^*(x,\theta_\infty) = Q_j(\theta_\infty,\overline{G}). \end{aligned}$$

The third line of (14) follows similarly. □

If $\mathcal{X}$ were compact, then $\overline{\mathcal{X}} = \mathcal{X}$ and the preceding lemma solves (14) within the original set-up. In that case the proof is complete. If $\mathcal{X}$ is not compact, then by C1 its completion is compact. We can apply our proof to this compactified $\mathcal{X}$ provided that we can "lift" the whole problem into the compactification. We shall now show that this is always possible under the conditions C1–C5.

By C2 the functions $(\theta,x) \to Q_j^*(x,\theta)$ can be extended (uniquely) to continuous functions $\overline{Q}_j^*$ on $\Theta \times \overline{\mathcal{X}}$, for $\overline{\mathcal{X}}$ the completion of $\mathcal{X}$, which is compact by C1. The next step is to define an extension $\overline{\mathcal{Y}} \supset \mathcal{Y}$ and densities $\bar{f}_\theta(\,\cdot \mid x)$ on $\overline{\mathcal{Y}}$ for $x \in \overline{\mathcal{X}} - \mathcal{X}$ and $\theta \in \Theta$ and to define sets $\overline{\mathcal{S}}_j \subset \overline{\mathcal{Y}} \times \overline{\mathcal{X}}$ in such a way that

$$\int_{(\overline{\mathcal{S}}_j)_x} \bar{f}_\theta(y \mid x)\,d\bar{\nu}(y) = \overline{Q}_j^*(x,\theta), \qquad x \in \overline{\mathcal{X}} - \mathcal{X}$$

$$\overline{\mathcal{S}}_j \cap (\mathcal{Y} \times \mathcal{X}) = \mathcal{S}_j, \qquad j = 1,\ldots,J$$

and such that the maps $\theta \to \bar{f}_\theta(y \mid x)$ are upper semicontinuous. In that case C1–C5 carry over to the extended model, with sample space $\overline{\mathcal{Y}} \times \overline{\mathcal{X}}$ and parameter set the product of Θ and the

set of all probability measures on $\overline{\mathcal{X}}$. Because under (θ_0, G_0) the observations (Y_i, X_i) fall with probability one into $\mathcal{Y} \times \mathcal{X}$, the (restricted) maximum likelihood estimator will not change, and we obtain consistency of $\hat{\theta}$ and the probabilities $Q_j(\hat{\theta}, \widehat{G})$ by the preceding argument.

Under some conditions, the extended model can be constructed from natural extensions of f_θ and the strata $\mathcal{S}_j$. However, an extension exists under minimal conditions. For instance, define $\overline{\mathcal{Y}} = \mathcal{Y} \cup \{1, \ldots, J\}$ and let $\bar{f}_\theta(\cdot \mid x)$ be the density of a discrete distribution $F_\theta(\cdot \mid x)$ on the set $\{1, 2, \ldots, J\}$ such that

$$F_\theta(\{j\} \mid x) = \overline{Q}_j^*(x, \theta), \qquad j = 1, \ldots, J, \quad x \in \overline{\mathcal{X}} - \mathcal{X}.$$

Furthermore, define $\overline{\mathcal{S}}_j = \mathcal{S}_j \cup (\{j\} \times \overline{\mathcal{X}} - \mathcal{X})$, so that $(\mathcal{S}_j)_x = \{j\}$ for $x \in \overline{\mathcal{X}} - \mathcal{X}$. To obtain a partitioning of $\overline{\mathcal{Y}} \times \overline{\mathcal{X}}$ we must also add the stratum $\overline{\mathcal{Y}} \times \overline{\mathcal{X}} - \cup \overline{\mathcal{S}}_j$, but this will have probability zero under all possible parameters.

ACKNOWLEDGEMENTS

We owe thanks to Norman Breslow for introducing us to two-phase sampling models and to Brad McNeney for several helpful discussions and suggestions. Wellner's research on this project was supported in part by National Science Foundation grant DMS-9532039, NIAID grant 2R01 AI291968-04, and the Stieltjes Institute.

REFERENCES

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

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

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

R. M. Dudley (1989). *Real Analysis and Probability*. Wadsworth & Brooks/Cole, Belmont, CA.

T. S. Ferguson (1996). *A Course in Large Sample Theory*. Chapman & Hall, London.

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

R. D. Gill, Y. Vardi & J. A. Wellner (1988). Large sample theory of empirical distributions in biased sampling models. *The Annals of Statistics*, 16, 1069–1112.

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

L. Le Cam (1953). On some asymptotic properties of maximum likelihood estimates and related estimates. *University of California Publications in Statistics*, 1, 277–330.

B. McNeney (1998). Asymptotic Efficiency in Semiparametric Models with non-i.i.d. Data. Ph. D. dissertation, University of Washington, Seattle, WA.

J. Palmgren (1989). Regression models for bivariate binary responses. *Technical Report*, 101, Department of Biostatistics, School of Public Health and Community Medicine, University of Washington, Seattle.

R. L. Prentice (1986). A case-cohort design for epidemiologic cohort studies and disease prevention trials. *Biometrika*, 73, 1–11.

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

A. J. Scott & C. J. Wild (1991). Fitting logistic models in stratified case-control studies. *Biometrics*, 47, 497–510.

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

A. J. Scott & C. J. Wild (1998). Maximum likelihood for generalised case-control studies. Preprint, Dept. of Statistics, University of Auckland.

S. G. Self & R. L. Prentice (1988). Asymptotic distribution theory and efficiency results for case-cohort studies. *The Annals of Statistics*, 16, 64–81.

A. W. van der Vaart (1994). Maximum likelihod estimation with partially censored observations. *The Annals of Statistics*, 22, 1896–1916.

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

A. W. van der Vaart & J. A. Wellner (1992). Existence and consistency of maximum likelihood in upgraded mixture models. *Journal of Multivariate Analysis*, 43, 133–146.

A. W. van der Vaart & J. A. Wellner (2000). Preservation theorems for Glivenko–Cantelli and uniform Glivenko–Cantelli classes. In *High Dimensional Probability II* (E. Giné, D. M. Mason & J. A. Wellner, eds.), Birkhäuser, Boston, pp. 113–132.

Received 11 May 2000
Accepted 18 October 2000

Aad van der VAART: aad@cs.vu.nl
Divisie Wiskunde en Informatica
Vrije Universiteit Amsterdam, De Boelelaan 1081a
1081 HV Amsterdam, The Netherlands

Jon A. WELLNER: jaw@stat.washington.edu
Statistics Department, University of Washington
Box 354322, Seattle, WA 98195-4322, USA