LARGE SAMPLE THEORY OF EMPIRICAL DISTRIBUTIONS IN BIASED SAMPLING MODELS

by

Richard Gill and Jon A. Wellner

TECHNICAL REPORT No. 75
December 1985

Department of Statistics
University of Washington
Seattle, Washington 98195 USA
LARGE SAMPLE THEORY OF
EMPIRICAL DISTRIBUTIONS
IN BIASED SAMPLING MODELS

by

Richard Gill and Jon A. Wellner
Centrum voor Wiskunde and University of Washington

December 1985

AMS 1980 Subject Classifications: Primary 62G05, 60F05;
Secondary 62G30, 60G44

Key Words and Phrases: Selection bias models, Vardi's estimator,
nonparametric Maximum Likelihood,
LARGE SAMPLE THEORY OF
EMPIRICAL DISTRIBUTIONS
IN BIASED SAMPLING MODELS

by

Richard Gill and Jon A. Wellner
Centrum vor Wiskunde and University of Washington

ABSTRACT

Vardi (1985) introduced an $s$-sample model for biased sampling, gave conditions which guarantee the existence and uniqueness of the nonparametric maximum likelihood estimator $E_n$ of the common underlying distribution $G$, and discussed numerical methods for calculating the estimator.

Here we examine the large sample behavior of the NPMLE $E_n$, including results on uniform consistency of $E_n$, convergence of $\sqrt{n}(E_n - G)$ to a Gaussian process, and asymptotic efficiency of $E_n$ as an estimator of $G$. The proofs are based upon recent results for empirical processes indexed by sets and functions, properties of irreducible $M$-matrices, and the homotopy invariance theorem.

A final section discusses examples and applications to stratified sampling, "choice-based" sampling in econometrics, and "case-control" studies in biostatistics.
1. INTRODUCTION: BIASED SAMPLING MODELS AND VARDI'S NONPARAMETRIC MLE.

We begin by describing the biased sampling model studied by Vardi (1985a, 1985b) and a useful iid version of the model.

The s-sample model for biased sampling introduced by Vardi (1985a) is as follows: Let \( G \) be an unknown distribution on \((X, B)\), and let \( w_1, \ldots, w_s \) be given non-negative weight functions. Suppose that \( X_1, \ldots, X_m \) are iid \( X \)-valued rv's with distribution \( F_i \) given by

\[
F_i(A) = \frac{\int_A w_i \, dG}{\int w_i \, dG} = \frac{G(\lambda_i w_i)}{G(w_i)}
\]

for \( A \in B \) and \( i = 1, \ldots, s \). In this model we let \( n = n_1 + \cdots + n_s \) and typically assume that \( \lambda_m = n_1/n \to \lambda > 0 \) as \( \min(n_i) \to 0 \).

A convenient iid-model for biased sampling can be defined in terms of the same weight functions \( w_1, \ldots, w_s \), and distribution \( G \) by assuming that \((X_j, I_j)\), \( j = 1, \ldots, n \) are iid with distribution

\[
P(X \in A, I = i) = \lambda_i F_i(A) = \frac{\lambda_i \int_A w_i \, dG}{\int w_i \, dG}
\]

for \( A \in B \) and \( i = 1, \ldots, s \). In this version of the model the sample sizes \( n_i = \sum_{j=1}^n 1[I_j = i] \) are random and we have \( \lambda_m = n_1/n \to_p \lambda > 0 \) as \( n \to \infty \).

In both cases the problem is to estimate \( G \). Vardi (1985a) studied the nonparametric maximum likelihood estimate of \( G \), gave conditions for its existence and uniqueness, and proposed an algorithm for calculating it. Vardi (1985b) also provided a heuristic outline of the large sample theory of his estimator.

In this paper we give a thorough treatment of the asymptotic distribution theory of Vardi's estimator, including results about its efficiency using the methods developed in Begun, Hall, Huang, and Wellner (1983). Consistency is treated in section 2, lower bounds for estimation are discussed in section 3, and asymptotic normality is established in section 4. In the final section, section 5, we discuss examples, applications, and connections between biased sampling models and work in econometrics on "choice-based" sampling by Cosslett (1981), and in biostatistics on "case-control" studies by Breslow and Day (1980), Prentice and Pyke (1979), and others.
Now we introduce Vardi's (1985a) nonparametric maximum likelihood estimator of \( C \) and some notation which will be used throughout the paper.

Let \( F = (F_1, \ldots, F_s)^T \) denote the vector of distributions defined by (1), and, with \( \lambda_m = \frac{\lambda}{n}, i = 1, \ldots, s \), set
\[
F = \lambda^T F \quad \text{and} \quad F_n = \lambda_n^T F.
\]
Note that \( F_n \) is random in the iid model (2). For any vector \( u \in \mathbb{R}^s \), let \( u \) denote the \( s \times s \) diagonal matrix with entries \( u_i \) on the diagonal, and set
\[
\tilde{w}(u) = u^{-1} w, \quad \tau(u) = (\lambda^T \tilde{w}(u))^{-1}, \quad \tau_n(u) = (\lambda_n^T \tilde{w}(u))^{-1}
\]
where \( w = (w_1, \ldots, w_s)^T \) is the vector of biasing functions. We also set
\[
\bar{w} = \int w \, dG = C(w)
\]
and write
\[
\tilde{w} = \tilde{w}(\bar{w}) = \frac{w}{w_0} \quad \tau = (\lambda^T \tilde{w})^{-1}, \quad \tau_n = (\lambda_n^T \tilde{w})^{-1}.
\]
Now \( F \) and \( F_n \) are related to \( C \) by
\[
dF = (\lambda^T \tilde{w}) \, dG = r^{-1} \, dG
\]
and
\[
dF_n = (\lambda_n^T \tilde{w}) \, dG = \tau_n^{-1} \, dG
\]
so that, letting \( X^+ = \{x \in X: \tau(x)^{-1} > 0\} \) and \( X^+_n = \{x \in X: \tau_n(x)^{-1} > 0\} \), we have
\[
C(A) = \int_A \tau(x) \, dF(x) \quad \text{for } A \in B \cap X^+
\]
and
\[
C(A) = \int_A \tau_n(x) \, dF_n(x) \quad \text{for } A \in B \cap X_n^+.
\]
It is clear from (9) that \( C \) can only be estimated on the set \( X^+ \). We therefore define
\[
C^+(A) = \frac{C(A \cap X^+)}{C(X^+)} = \frac{F(A \cap X^+)}{F(X^+)} \quad \text{for } A \in B.
\]
If \( \text{support}(C) \subset X^+ \), then \( F(r) = C(X^+) = 1 \) so that \( C^+ = C \), but in general \( C(X^+) < 1 \) and \( C^+ \neq C \). Also note that the distribution of the data depends on \( C \) through \( C^+ \) only, and hence the most we can hope to estimate is \( C^+ \) and \( \bar{w}^+ = \int \bar{w} \, dC^+ \neq \bar{w} \). Therefore, to ease the notational burden, we henceforth drop the plus
Our convention throughout will be to write
\[ \int h \, dF = \int \mathbf{x} \, h \, dF \]
or, in a notation which we will use frequently,
\[ F(h) = F(h \mathbf{1}_x) . \]

Note that the right side of (11) is a homogeneous function of degree 0 in the \( W_i \)'s: using the above convention and (4) and (6)
\[ \frac{F(1_A \tau(W))}{F(\tau(W))} = \frac{F(1_A \tau(cW))}{F(\tau(cW))} \quad \text{for any } c > 0 . \]
Thus it follows in particular that, with \( c = 1/W_s \),
\[ \mathcal{G}(A) = \frac{F(1_A \tau(V))}{F(\tau(V))} \]
where \( V = W/W_s \).

Now let \( F_n \) denote the empirical measure of all the \( X_i \)'s:
\[ F_n = \begin{cases} 
  n^{-1} \sum_{j=1}^{N} \delta_{X_j} & \text{in the iid model} \\
  n^{-1} \sum_{i=1}^{n} \sum_{j=1}^{n} \delta_{X_{ij}} & \text{in the s-sample model} 
\end{cases} \]

Let \( F_{ni} \) denote the empirical measure for "sample i":
\[ F_{ni} = \begin{cases} 
  n_i^{-1} \sum_{j=1}^{N_i} \delta_{X_j} 1_{[i_j = i]} & \text{in the iid model} \\
  n_i^{-1} \sum_{j=1}^{n_i} \delta_{X_{ij}} & \text{in the s-sample model} 
\end{cases} \]

where \( \delta_z \) is the measure with mass 1 at \( z \): \( \delta_z(A) = 1_A(z) \) for \( A \in \mathcal{B} \), and define
\[ H(\mathbf{u}) = \int \tau(\mathbf{u}) \mathbf{w}(\mathbf{u}) dF = F((\tau \mathbf{w})(\mathbf{u})) \]
\[ H_n(\mathbf{u}) = \int \tau_n(\mathbf{u}) \mathbf{w}(\mathbf{u}) dF_n = F_n((\tau_n \mathbf{w})(\mathbf{u})) \]
\[ \mathcal{H}_n(\mathbf{u}) = \int \tau_n(\mathbf{u}) \mathbf{w}(\mathbf{u}) dF_n = F_n((\tau_n \mathbf{w})(\mathbf{u})) \]
In view of (7) or (9) and (8) or (10) it follows that
(18) \[ H(W) = 1 \quad \text{and} \quad H_n(W) = 1. \]
and, in fact,
(19) \[ H(cW) = 1 \quad \text{for any} \ c > 0. \]
In view of (11) and (18), estimators \( E_n \) and \( W_n \) of \( C \) and \( W = C(w) \) may be defined as the solution (provided it exists) of the equations

\[
\begin{align*}
H_n(W_n) &= 1 \\
E_n(h) &= \frac{F_n(h,r_n(W_n))}{F_n(r_n(W_n))} \quad \text{for all} \ h \\
W_n &= E_n(u).
\end{align*}
\]

Alternatively, in view of (14) and (19) we can first estimate \( V = W/W_s \) as the solution, if it exists, of

(21) \[ H_m(V_n) = 1, \quad i = 1, \ldots, s-1 \]
and then estimate \( G \) by

(22) \[ E_n(h) = \frac{F_n(h,r_n(V_n))}{F_n(r_n(V_n))}. \]

The latter approach is the one taken by Vardi (1985a, 1985b).

In our notation, Vardi's (1985a) condition for existence of a unique nonparametric maximum likelihood estimator (i.e. a solution of (20) or (21)) is:

(23) \[ n \sum_{i \in B} \int_{[x: w_i(x) > 0]} dF_n(x) > \sum_{i \in B} n_i \quad \text{for every proper subset} \ B \text{ of } \{1, \ldots, s\}. \]

Vardi (1985a) shows that (23) has several equivalent forms, one of which is in terms of a directed graph \( M \) on \( s \) vertices defined as follows: a directed edge connects vertex \( i \) to vertex \( i' \), \( i \to i' \), if and only if \( \int u_i dF_{n_i} > 0 \). Then we say that \( M \) is strongly connected if for any two vertices \( x \) and \( x' \) there exists a directed path from \( x \) to \( x' \) and a directed path from \( x' \) to \( x \). Vardi (1985a) shows that (23) is equivalent to:

(24) \[ \text{the graph } M \text{ is strongly connected}. \]

Note that \( M \) is strongly connected if and only if the matrix with
THEOREM 1.1. (Vardi) If (23) or equivalently (24) holds, then the equations (21) have a unique solution and \( L_n \) given in (22) is the nonparametric maximum likelihood estimate of \( C \). Equivalently, if (23) or (24) holds, then the system of equations (20) has a unique solution \( L_n, W_n \) which is the nonparametric maximum likelihood estimate of \( C, W \). Conversely, if (23) or (24) fail, then the equations (20) do not have a unique solution \( L_n, W_n \).
2. CONSISTENCY OF $E_n$.

Our first task is to establish the consistency of the nonparametric maximum likelihood estimator $E_n$ of $G$ introduced in section 1 under both the iid and $s$-sample models. Suppose that

(1) $G(w_i^2) < \infty$ for $i = 1, \ldots, s$.

Then it follows from the strong law of large numbers (and the assumption $n_i/n \to \lambda_i$ for $i = 1, \ldots, s$ in the $s$-sample case) that for any $i, j = 1, \ldots, s$,

(2) $F_n(w_i) \to a.s. F_j(w_i) = \frac{1}{W_j} G(w_i w_j)$ as $n \to \infty$.

Thus we can define a graph $M^*$ with $s$ vertices and an edge from $i$ to $j$, $i \sim j$, if and only if

(3) $G(w_i w_j) > 0$.

In terms of the graph $M^*$, an asymptotic version of the condition (1.23) is:

(4) $M^*$ is a connected graph.

The main result of this section can now be stated as:

**THEOREM 2.1.** (Consistency of $E_n$). Suppose that (1) and (4) hold, and that $\mathcal{H} = \{h \in C \subset G \}$ where the envelope function $h_e$ satisfies $G(h_e) = F(r h_e) < \infty$ and $G$ is a Vapnik-Chervonenkis class of subsets of $X$. Then

(5) $\|E_n - G\|_H = \sup h \in \mathcal{H} \{\|E_n(h) - G(h)\|: h \in \mathcal{H}\} \to a.s. 0$ as $n \to \infty$.

**COROLLARY 2.1.** If $X = R^d$ and $\mathcal{H}$ is the class of all indicator functions of lower left orthants, of all rectangles, or of all balls, or the class of all half-spaces, then (1) and (4) imply that (5) holds.

**REMARK 2.1.** Several variants of theorem 2.1 are possible, depending on which basic Glivenko-Cantelli theorem is used. For some other possible formulations, see e.g. Dudley (1985) theorems 6.1.5 and 11.1.6 or Pollard (1984).

**REMARK 2.2.** What can be said if (4) fails? If $r \leq s$ is the number of components of the graph $M^*$, let $C_1, \ldots, C_r$ be a partition of the indices $\{1, \ldots, s\}$ corresponding to the $r$...
components, and define \( x_j = \{ x \in X : \sum_{i \in C_j} \lambda_i u_i(x) > 0 \} \) for some \( i \in C_j \). (Thus when \( r = 1 \), \( x_i = x^r \).) Then we can estimate \( C_j = \mathcal{G}(1x_j)/\mathcal{G}(1x_j) \) for each \( j \) : note that Vardi's estimator works for each component of the graph separately, that \( M \) is reducible, \( \text{rank}(M) = s - r \), and hence, after a relabeling of the samples, \( M \) is a block-diagonal matrix with blocks \( M_{jj} \) which are irreducible and \( \text{rank}(M_{jj}) = \dim(M_{jj}) - 1 \).

An important part of our argument, which we use repeatedly in the following sections, concerns the derivative matrix \( \nabla H^T \) of the vector of functions \( H(u) \) defined in (1.17). By straightforward calculation (justified by the monotone convergence theorem)

\[
(\nabla H)(u) = (\nabla H^T(u))^T
\]

\[
(6) \quad = \begin{bmatrix} H(u) \lambda^{-1} - F(r^2(u) \tilde{w}(u) \tilde{w}^T(u)) \lambda \mu^{-1} \\
\lambda^{-1} - F(r^2 \tilde{w} \tilde{w}^T) \end{bmatrix} \lambda \mu^{-1}
\]

\[
(7) \quad = \begin{bmatrix} \lambda^{-1} - F(r^2 \tilde{w} \tilde{w}^T) \end{bmatrix} \lambda \mu^{-1}
\]

when \( u = \alpha \bar{W} \) so \( H(u) = 1 \).

\( = -M \lambda \mu^{-1} \)

**PROPOSITION 2.1.** If the graph \( M^* \) is connected, then the matrix

\[
M = \lambda^{-1} - F(r^2 \tilde{w} \tilde{w}^T)
\]

defined in (7) has rank \( s-1 \). In fact, every principal proper submatrix of \( M \) is nonsingular, and the same holds for \( (\nabla H)(u) \). In particular, if the \( i \)th row and column are deleted, a matrix of full rank \( (s-1) \) results.

To prove the preceding assertions, it will be convenient to work with \( N_n = (N_{n1}, \ldots, N_{ns-1}, 1)^T \) defined by (1.21) and \( \bar{V} = (V_1, \ldots, V_{s-1}, 1)^T = \bar{W}/\bar{W}_s \).

**PROPOSITION 2.2.** If the graph \( M^* \) is connected, then the unique solution \( N_n \) of (1.21) satisfies

\[
(8) \quad N_n \to_{a.s.} \bar{V} = \bar{W}/\bar{W}_s \quad \text{as} \quad n \to \infty
\]

where \( \bar{V} \) is the unique solution of

\[
(9) \quad H_i(\bar{V}) = 1, \quad i = 1, \ldots, s-1.
\]

This proposition will be proved by means of the following lemmas.
LEMMA 2.1. \((\nabla H)(u) \cdot u = 0\) for all \(u \in \mathbb{R}^s\).

PROOF. Since each \(H_i\) is homogeneous of degree zero, \(H(cu) = H(u)\) for all \(c > 0\). Hence

(a) \(0 = \frac{\partial}{\partial c} H(cu) = (\nabla H)(cu) \cdot u\) for all \(c > 0\)

and in particular for \(c = 1\). \(\square\)

LEMMA 2.2. \((- (\nabla H)(u))_{ij} \leq 0\) for \(i \neq j\), and hence the upper left \((s-1) \times (s-1)\) submatrix of \(- (\nabla H)\) is in the collection \(Z^{(s-1)\times(s-1)}\) of Berman and Plemmons (1979) page 132.

PROOF. Note that \(F(r^2(u) \tilde{w}_i(u) \tilde{w}_j(u) \geq 0\) in (6). \(\square\)

PROOF of Proposition 2.1. By Berman and Plemmons (1979) exercise 6.4.14 page 155, and the facts that \(M \in \mathbb{R}^{s \times s}, \lambda \gg 0, \) and \(M = 0, \) \(M\) is a singular \(M\)-matrix with "property c". Since the graph \(M^*\) is connected, \(M\) is also irreducible; see theorem 2.7, Berman and Plemmons (1979), page 30. Hence by theorem 6.4.16(4), Berman and Plemmons (1979), every principal proper submatrix of \(M\) is nonsingular. The same argument applies to \((\nabla H)(u)\) starting from the observation that \((\nabla H)(u) \cdot u = 0\). \(\square\)

LEMMA 2.3. If \(M^*\) is connected (i.e. (4) holds), then every solution of (9) is an isolated solution.

PROOF. This follows from Proposition 2.1 and the implicit function theorem; see e.g. Apostol (1957) or Ortega and Rheinboldt (1970). \(\square\)

LEMMA 2.4. (Vardi) If \(M^*\) is connected, then with probability one for \(n \geq 0\) some \(\eta_n\), there is a unique solution \(\eta_n\) of the equations (1.21).

PROOF. This follows from Vardi (1985a), theorems 1 and 2. \(\square\)

LEMMA 2.5. Let \(K\) be a compact set in \(\mathbb{R}^s\) with all coordinates bounded away from 0. Then

\[
\max_{1 \leq i \leq s} \sup_{u \in K} |H_1(u) - H_i(u)| \to_{\text{as}} 0 \quad \text{as} \quad n \to \infty .
\]

PROOF. Note that the functions

(a) \(h_i(x, u) = r_n(u) \tilde{w}_i(u)(x), \quad i = 1, \ldots, s, \quad u \in K\)

are uniformly bounded (by \(\max_{1 \leq i \leq s} (1/\lambda_i)\)). Since \(K\) is compact and since each \(h_i\) is a monotone decreasing function of each coordinate of \(u\), the supremum in (10) may easily be reduced to a maximum over a finite set of \(u\)'s in \(K\) and then (10) follows by
the strong law of large numbers. Alternatively, (10) follows from a
Glivenko - Cantelli theorem for \( F_n \) indexed by the collection of
functions in (a): see e.g. Pollard (1982), (1984), or Dudley (1985).

\[ \square \]

**PROOF of proposition 2.2.** If (4) holds, then by lemma 2.3,
\( \bar{y} = \frac{w}{w_s} \) is an isolated solution of (9). Let \( K \) be a small com-
 pact set containing \( \bar{y} \) in its relative interior \( K^0 \) in \( R^{s-1} \) and no
other solution. Since \( H(v_1, \ldots, v_{s-1}) \) is continuous,
\( \inf \{ \max_{u \in \partial K} |H_i(u) - 1| \} = \varepsilon > 0 \).

Furthermore, for \( 0 < \delta < \varepsilon \),
\[ \max_{1 \leq i \leq s-1} \sup_{u \in \partial K} |H_n(u) - H_i(u)| < \frac{\delta}{\varepsilon} \]
by lemma 2.5. Hence by the homotopy invariance theorem (or
6.1.6 page 152) of Ortega and Rheinboldt (1970), for \( n \geq N_\varepsilon \)
\[ \deg(H_n, K^0, 1) = \deg(H, K^0, 1) \]
where the right side is \( \pm 1 \) since the upper left \( (s-1) \times (s-1) \)
submatrix of \( \nabla H^T \) is nonsingular everywhere by proposition and since
\( H \) has only one solution of (9) in \( K^0 \). Thus the unique solution
\( \bar{y} \) of (1.21) guaranteed by lemma 2.4 is also in \( K^0 \). Since \( K \) can
be chosen to be arbitrarily small, this implies (8).

Now this same argument applies to any other solution \( \bar{y} \) of
(9), since any other such solution is isolated by lemma 2.3. Hence
\( \bar{y} = \bar{y} \); i.e. (9) has only one solution, namely \( \bar{y} \).

\[ \square \]

**PROOF of theorem 2.1.** For a fixed function \( h \) we write
\( |F_n(h r_n(N_n)) - F(hr)| \)
\( \leq |F_n(h(r_n(N_n) - r))| + |F_n(hr) - F(hr)| \)
\( \leq \| \frac{r_n(N_n)}{r} - 1 \|_\infty |F_n(hr)| + |F_n(hr) - F(hr)| \)
\( \xrightarrow{n \to \infty} 0 \)
for any fixed function \( h \) with \( F(hr) = C(h) < \infty \) by the strong
law of large numbers, and by continuity and boundedness of
\( r(u)/r \) (as a function of \( x \)) together with proposition 2.2. Thus
by Pollard's Glivenko - Cantelli theorem (see Dudley (1985),
theorems 11.1.2 and 11.1.6), for \( \mathcal{H} \) of the form hypothesized,
\( \sup_{h \in \mathcal{H}} |F_n(h r_n(N_n) - F(hr)| \)
\[ \| r_n(N_n) \|_r - 1 = \sup_{h \in \mathcal{M}} | F_n(hr) | + \sup_{h \in \mathcal{N}} | F_n(hr) - F(hr) | \]

\[ \rightarrow_{as} 0 \sup_{h \in \mathcal{M}} | F(hr) | + 0 = 0. \]

Thus, since

\[ | C_n(h) - C(h) | = \left| \frac{F_n(hx_n(N_n))}{F_n(r_n(N_n))} - \frac{F(hx)}{F(x)} \right| \]

\[ \leq \left| \frac{F_n(hx_n(N_n)) - F(hx)}{F_n(r_n(N_n))} \right| + \left| \frac{F(hx)}{F(x)} \right| \left| \frac{F_n(r_n(N_n)) - F(x)}{F_n(r_n(N_n))} \right|. \]

(5) follows from (a) with \( h = 1 \) and (b).

Before ending this section, we record some useful facts concerning generalized inverses of the matrix \( M \) which will be used repeatedly in sections 2-5.

**PROPOSITION 2.3.** If the graph \( M^* \) is connected, then the matrix \( M \) has a \( \{1,2\} \)-inverse \( M^- \). Thus \( M^- \) satisfies both

(11) \( MM^-M = M \)

and

(12) \( M^-MM^- = M^- \).

Any such \( \{1,2\} \)-inverse \( M^- \) also satisfies:

(13) \( u = Mx \) implies \( x = M^-u + c \lambda \) for some \( c \)

where \( \lambda \) is the unique eigenvector of \( M \) with 0 eigenvalue. If, in addition, \( M^- \) is the \( \{1,2,3,4\} \) or Moore-Penrose generalized inverse of \( M \), then

(14) \( MM^- = 1 - \theta \theta^T \)

where \( \theta = \sqrt{\lambda} \).

**PROOF.** By proposition 2.2, deleting any row and column, in particular the last row and column, from \( M \) yields an \((s-1) \times (s-1)\) matrix \( M_{11} \) of rank \( s-1 \). Thus a \( \{1,2\} \) inverse \( M^- \) of \( M \) is given by

(15) \( M^- = \begin{bmatrix} M_{11}^{-1} & 0 \\ 0 & 0 \end{bmatrix} \)

see e.g. Seber (1977) page 76 and also Berman and Plemmons (1979) page 117.

Since \( M \) is symmetric, another way to get a \( \{1,2\} \)-inverse \( M^- \) satisfying (11) and (12) is to use the decomposition
\[ M = PDP^T \] where the columns of \( P \) are the normalized eigenvectors of \( M \) and \( D \) is the diagonal matrix of the eigenvalues \( d_1 \geq d_2 \geq \cdots \geq d_{s-1} > d_s = 0 \) of \( M \). Then

\[(16) \quad M^{-} = PDP^T \]

where \( D^{-} \) is the diagonal matrix with \( D_{ii}^{-} = 1/d_i \) for \( i = 1, \ldots, s-1 \) and \( D_{ss}^{-} = 0 \), is in fact the Moore - Penrose (or \{1,2,3,4\} generalized inverse of \( M^{-} \); see Berman and Plemmons (1979) page 117.

Now for any \{1,2\} - inverse \( M^{-} \) (an inverse satisfying both (11) and (12)), \( M^{-}M \) is the projector on \( R(M^{-}) \) along \( N(M) \); see Berman and Plemmons (1979) page 118. Thus if \( z = z_0 + c\lambda \) and \( y = Mz \), it follows that \( M^{-}y = M^{-}Mz = z_0 = z - c\lambda \) and hence (13) holds. The proof of (14) proceeds by direct computation using \( M = PDP^T \) and (16). \( \square \)
3. A LOWER BOUND FOR ESTIMATION OF G.

Our goal here will be to derive convolution theorem for regular estimates of G. Our computations will follow the approach of Begun, Hall, Huang, and Wellner (1983). An alternative derivation could be based on the results of Millar (1985).

We suppose that \( \text{support}(G) \subset \text{support}(\lambda \tau \hat{w}) \) so that \( G = G^* \) and \( F(r) = 1 \), that \( r^{-1} = \lambda \tau \hat{w} \) and \( r \) are bounded, that \( G \) has density \( g \) with respect to \( \nu \), and write

\[
(1) \quad f_i = \frac{w_i g}{W_i} = \hat{w}_i g,
\]

with

\[
W_i = \int w_i g \, d\nu = G(w_i) = \langle w_i, 1 \rangle_G
\]

\[
(2) \quad r = (\sum_{i=1}^{s} \lambda_i \hat{w}_i)^{-1} = (\lambda \tau \hat{w})^{-1},
\]

\[
(3) \quad f = \sum_{i=1}^{s} \lambda_i f_i = r^{-1} g.
\]

It is easily verified that each \( f_i \) is Hellinger differentiable with respect to \( g \) with derivative

\[
(4) \quad A_i \beta = f_i^{1/2} (\frac{\beta}{g^{1/2}} - \int \hat{w}_i \beta g^{1/2} d\nu), \quad i = 1, \ldots, s
\]

so that

\[
(5) \quad A_i^T A_i \beta = \frac{f_i \beta}{g} - \frac{f_i}{g^{1/2}} \int \beta f_i g^{-1/2} d\nu, \quad i = 1, \ldots, s
\]

and hence

\[
U \beta = \sum_{i=1}^{s} \lambda_i A_i^T A_i \beta
\]

\[
= \sum_{i=1}^{s} \lambda_i \hat{w}_i \beta - \sum_{i=1}^{s} \lambda_i \hat{w}_i g^{1/2} \langle \hat{w}_i g^{1/2}, \beta \rangle_{\nu}
\]

\[
= r^{-1} \beta - \hat{w}^T \frac{\lambda}{g} g^{1/2} \langle \hat{w} g^{1/2}, \beta \rangle_{\nu}.
\]

The operator \( U \) maps \( L^2(\nu) \) to \( L^2(\nu) \) (since \( r^{-1} \) is bounded) and is precisely the operator \( A^T A \) of Begun, Hall, Huang, and Wellner (1983) in the iid case. It is often convenient to work instead with \( \tilde{U} \) mapping \( L^2(G) \) to \( L^2(G) \) defined by

\[
\tilde{U} \beta = g^{-1/2} U(g^{1/2} \beta), \quad \text{for } \beta \in L^2(G).
\]

Thus
(7) \( \tilde{\mu}^{-1} \tilde{\beta} = \tilde{\mu}^{-1} \tilde{\beta} - \tilde{\mu}^T \lambda \langle \tilde{\mu}, \tilde{\beta} \rangle_C \)

and we have
\[
\langle \tilde{\mu}, \tilde{r} \tilde{\beta} \rangle_C = (I - \langle \tilde{\mu}, \tilde{r} \tilde{\mu}^T \lambda \rangle \langle \tilde{\mu}, \tilde{\beta} \rangle_C
\]
\[
= (\lambda^{-1} - \langle \tilde{\mu}, \tilde{r} \tilde{\mu}^T \lambda \rangle \langle \tilde{\mu}, \tilde{\beta} \rangle_C
\]
\[
= M \lambda \langle \tilde{\mu}, \tilde{\beta} \rangle_C
\]

where
\[
M = \lambda^{-1} - \langle \tilde{\mu}, \tilde{r} \tilde{\mu}^T \rangle_C
\]
Thus, by (2.13), (8) can be inverted to yield
\[
(10) \quad \lambda \langle \tilde{\mu}, \tilde{\beta} \rangle_C = M^{-1} \langle \tilde{\mu}, \tilde{r} \tilde{\mu}^T \rangle_C - \alpha \lambda
\]
where the constant \( \alpha = \alpha(\tilde{\beta}) \) is still to be determined. Substitution of (10) into (7) yields
\[
(11) \quad \tilde{r} \tilde{\beta} = \tilde{\beta} - \tilde{\mu}^T M^{-1} \langle \tilde{\mu}, \tilde{r} \tilde{\beta} \rangle_C - \alpha
\]
which implies
\[
(12) \quad \tilde{\mu}^{-1} \tilde{\beta} = \tilde{\beta} + \tilde{\mu}^T M^{-1} \langle \tilde{\mu}, \tilde{r} \tilde{\beta} \rangle_C + \alpha
\]
where
\[
(13) \quad \tilde{V} \tilde{\beta} = \tilde{\beta} + \tilde{\mu}^T M^{-1} \langle \tilde{\mu}, \tilde{r} \tilde{\beta} \rangle_C
\]
Now we want \( \tilde{U}^{-1} \tilde{\beta} \perp 1 \) in \( L^2(C) \); therefore
\[
\alpha = \alpha(\tilde{\beta}) = -\langle \tilde{V} \tilde{\beta}, 1 \rangle_C
\]
and
\[
(14) \quad \tilde{V} \tilde{\beta} = \tilde{V} \tilde{\beta} - \langle \tilde{V} \tilde{\beta}, 1 \rangle_C
\]
Thus it follows that
\[
K(h, \tilde{h}) = \langle g^{1/2}(h - C(h)), \tilde{U}^{-1} g^{1/2}(\tilde{h} - C(\tilde{h})) \rangle_C
\]
\[
= \langle h - C(h), \tilde{U}^{-1}(\tilde{h} - C(\tilde{h})) \rangle_C
\]
\[
= \langle h - C(h), \tilde{V}(\tilde{h} - C(\tilde{h})) \rangle_C
\]
\[
= \langle h - C(h), \tilde{V}(\tilde{h} - C(\tilde{h})) \rangle_C
\]
where \( \tilde{V} \) is given in (13) and \( M \) is given in (9).

We say that \( \tilde{E}_n \) is a regular estimator of \( C \) if, under
\[
P_n = P_{\tilde{E}_n},
\]
(16) \( \sqrt{n}(\widehat{C}_n - C) \Rightarrow E \) independent of \( \beta \)

whenever \( \sqrt{n}(g_n^{1/2} - g^{1/2}) \rightarrow \beta \) in \( L_2(\nu) \). Then the above calculations in combination with theorem 4.1 of Begun, Hall, Huang, and Wellner (1983) yield:

**THEOREM 3.1.** Suppose that \( M^* \) is connected, that \( F(r) = 1 \) so that \( C^* = C \), and that both \( r^{-1} \) and \( r \) are bounded. Then the limit process \( E \) for any regular estimator of \( C \) in the biased sampling model (1) can be represented as

(17) \( E = ZZ + W \)

where \( W \) is independent of the mean 0 Gaussian process \( ZZ \) with covariance function given by (15).

A local asymptotic minimax lower bound can also be stated on the basis of the preceding calculations, but we forego this here.
4. ASYMPTOTIC NORMALITY OF THE NONPARAMETRIC MLE \( \hat{\theta}_n \).

We now study the asymptotic behavior of the process

\[
Z_n = \sqrt{n} (\hat{\theta}_n - \theta)
\]

regarded as a process indexed by functions \( h \in L_2(G) \). We do this in both the iid case and the s-sample case simultaneously (getting the same result in each case). To accomplish this, we define

\[
X_n^* = \sqrt{n} (\hat{\theta}_n - \theta)
\]

\[
\lambda_{ni} = n_i/n, \quad i = 1, \ldots, s \quad \text{and}
\]

\[
F_n = \sum_{i=1}^s \lambda_{ni} F_i \neq \sum_{i=1}^s \lambda_i F_i = F.
\]

as in (1.3). Here the \( n_i \)'s are random sample sizes in the iid sampling model (and hence the \( \lambda_{ni} \)'s are random too), and deterministic sample sizes in the s-sample model, and we have

\[
\lambda_{ni} \to_p \lambda_i \quad \text{as} \quad n \to \infty \quad \text{in the iid model}
\]

\[
\lambda_{ni} \to \lambda_i \quad \text{as} \quad n \to \infty \quad \text{in the s-sample model}
\]

by the weak law of large numbers in the first case, and by assumption in the second case. Centering by \( F_n \) rather than \( F \) amounts to conditioning on the \( \lambda_{ni} \)'s in the iid case. In either case we can write

\[
X_n^* = \sum_{i=1}^s \sqrt{\lambda_{ni}} \sqrt{n_i} (\hat{\theta}_n - F_i).
\]

Then, if \( \mathcal{F} \) is a Donsker class for \( F \) (or if \( \mathcal{F} \) is a Donsker class for all \( F_i, \quad i = 1, \ldots, s \) in the s-sample case), for a special construction as in Dudley and Philipp (1983),

\[
\|X_n^* - X^*\|_{\mathcal{F}}^2 = \sup_{f \in \mathcal{F}} |X_n^*(f) - X^*(f)| \to_p 0 \quad \text{as} \quad n \to \infty
\]

where \( X^* \) is (a sequence of) mean 0 Gaussian processes with covariance function

\[
\text{Cov}(X^*(h),X^*(\tilde{h})) = \sum_{i=1}^s \lambda_i \{F_i(h\tilde{h}) - F_i(h)F_i(\tilde{h})\}
\]

\[
= F(h\tilde{h}) - F(h)^T \Sigma F(\tilde{h})
\]

\[
= C(r^{-1}h\tilde{h}) - C(h\tilde{h})^T \Sigma C(\tilde{h})
\]

This follows in the iid case by a random sample size version of the classical theorems which can easily be deduced from the results of
Dudley and Philipp (1983), and in the s-sample case by straightforward calculation, also using the Dudley-Philipp results or other central limit theorems for the empirical process in the one-sample case such as Pollard (1982) or Ossiander (1985).

Note that in the iid case $X_n^*$ and $X^*$ are related to the more familiar empirical process $X_n = \sqrt{n}(F_n - F)$ and its limit $X$ by

$$X_n = X_n^* + \sqrt{n}(\lambda_n - \lambda)^T F$$

and

$$X^* = X^* + Z_F$$

where $X_n^*$ and $\sqrt{n}(\lambda_n - \lambda) \to_d Z_F \equiv N_s(0, \lambda \lambda^T)$ (by the multivariate CLT for a multinomial random vector) are asymptotically independent by a straightforward covariance calculation. Thus

$$\text{Cov}(X(h), X(h)) = \text{Cov}(X_n(h), X_n^*(h)) + F^T(h)(\lambda - \lambda \lambda^T)F(h)$$

The first step in studying $Z_n$ is to establish asymptotic normality of $\sqrt{n}(W_n - W)$ where $W_n$ results from solving (1.20).

**Theorem 4.1.** (CLT for $W_n$ ). If $M$ is connected, and $W_n$ is the solution of (1.20), then, for the same special construction as in (6),

$$\sqrt{n}(W_n - W) \to_d X^*(r\tilde{u}) + Z_nW$$

$$= K^{-1}X^*(r\tilde{u})$$

$$+ \left[ X^*(r/F(r)) - \{(1 - F(r))r + C(r\tilde{u}^T)M^{-1}X^*(r\tilde{u})\}W/F(r) \right]$$

$$= K^{-1}X^*(r\tilde{u}) - X^*(r + r\tilde{u}^TM^{-1}C(r\tilde{u}))W$$

if $F(r) = 1$

$$= X^*(K^{-1}\tilde{u} + (1 + \tilde{u}^TM^{-1}C(r\tilde{u}))W)$$

where both

$$K = M^{-1}W$$

and

$$K^{-1} = W^{-1}M^{-1}$$

have rank $s-1$. Recall that $M = \lambda^{-1} - F(r^2\tilde{u}\tilde{u}^T)$, and that $M^-$ is a $\{1, 2\}$-generalized inverse as defined in proposition 2.3.

**Proof of Theorem 4.1.** Since

$$1 = H_n(W_n) = H_n(W)$$

$$= H_n(W) + \nabla H_n(W_n)^*(W_n - W)$$
where $W^*_n$ lies on the line segment between $W_n$ and $W$, it follows that
(b) $-\mathcal{H}_n(W^*_n)\sqrt{n}(W_n - W) = \sqrt{n}(\mathcal{H}_n(W) - \mathcal{H}(W)) = X^*_n(r\tilde{w})$
and hence, letting $n \to \infty$, and writing $Z^*_W = \lim_n \sqrt{n}(W_n - W)$.
(c) $X^*(r\tilde{w}) = -\mathcal{H}(W)Z^*_W = M^*_W X^*(r\tilde{w})$
or, by (2.13) in proposition 2.3
$$Z^*_W = W^{-1} M^*_W X^*(r\tilde{w}) + W^{-1} Z^*_W$$
(d) $= K^* X^*(r\tilde{w}) + Z^*_W$.

This completes the proof of (10). We postpone identification of $Z_a$ as given in (11) since this step requires the convergence of $Z^*_W$, which will be established in theorem 4.2. In proving theorem 4.2 we will use only (10), and not (11).

Now define a Gaussian process $Z^*$ by
(14) $Z^*(h) = X^*(\tilde{v}(h - C(h)))$
where, for $\beta$ with $\beta\sqrt{r} \in L_2(C)$,
(15) $\tilde{v}\beta = \frac{r}{F(r)} \beta + r\tilde{w}^T M^*-G(r\tilde{w}\beta)$$
= \tilde{v}\beta$ of (3.13) if $F(r) = 1$.

**THEOREM 4.2.** (CLT for $Z_n^*$). Suppose that (2.1) and (2.4) hold (so the graph $M^*$ is connected), and that $\mathcal{M}$ is a collection of functions containing the constant function 1 such that (6) holds for the class $\mathcal{F} = \{h_r: h \in \mathcal{M}\}$. We also assume that $C(h^2_r) = F(h^2_r \theta^2) < \infty$ where $h_\theta$ is an envelope function for $\mathcal{M}$; $|h| \leq h_\theta$ for all $h \in \mathcal{M}$. Then,
(17) $\|Z_n - Z^*\|_{\mathcal{M}} = \sup_{h \in \mathcal{M}} |Z_n(h) - Z^*(h)| \to 0$ as $n \to \infty$
where $Z$ is the (sequence of) mean 0 Gaussian processes (14) with covariance function
$$K^*(h, h) = Cov(Z^*(h), Z^*(h))$$
(18) $= C((h - C(h))\tilde{v}(h - C(h)))$
$= C(r^*(h - \overline{h})(h - \overline{h})) + C((h - \overline{h})r\tilde{w}^T)M^*G((h - \overline{h})r\tilde{w})$
with $r^* = r/F(r)$, and $\overline{h} = C(h)$.
(19) \[ = C((h - G(h))\tilde{V}(\tilde{h} - G(\tilde{h}))) \text{ if } F(r) = 1. \]

where \( \tilde{V} \) is given by (15).

There are many sufficient conditions which imply that (6) holds for \( \mathcal{F} = \{hr: h \in \mathcal{H}\} \); see e.g. Dudley and Philipp (1983), Pollard (1984), or Ossiander (1985). The important special case of \( \mathcal{H} = \{C: C \in \mathcal{C}\} \) where \( \mathcal{C} \) is a Vapnik-Chervonenkis class of subsets of \( X \), which follows as a consequence of Pollard's (1982) theorem, is singled out in the following corollary.

**COROLLARY 4.2.** Suppose that \( X = \mathbb{R}^d \) and \( \mathcal{H} \) is the collection of all indicator functions of lower left orthants, or of all rectangles, or of all balls, or the class of all half spaces. If \( C(r) = F(r^2) < \infty \) and (2.1) and (2.4) hold, then \( Z_n \) satisfies (17).

**PROOF of theorem 4.2.** Now for any fixed \( h \) with \( F(h^2 r^2) = C(h^2 r) < \infty \)

(a) \[ Z_n(h) = \sqrt{n} (L_n(h) - G(h)) \]
\[ = \sqrt{n} \left[ \frac{F_n(hr_n(W_n))}{F_n(r_n(W_n))} - \frac{F_n(hr_n(W_n))}{F_n(r_n(W_n))} \right] \]
\[ = \{W_n^*(hr_n - r_n) \}
\[ + F_n((h - r_n) \frac{F_n(hr_n)}{F_n(r_n)}) \sqrt{n} (r_n(W_n) - r_n))/F_n(r_n(W_n)) \]

(b) \[ = \{W_n^*(hr_n - r_n) \}
\[ + F_n((h - r_n) \frac{F_n(hr_n)}{F_n(r_n)}) r_n^2(W_n^*) \left( \frac{1}{\lambda} \tilde{w}(W_n^*) \right)^T \]
\[ \times \sqrt{n} (W_n - W)/F_n(r_n(W_n)) \]

where \( W_n^* \) is on the line segment between \( W_n \) and \( W \)

since \( \nabla u r_n(u) = r_n^2(u) \lambda u^{-1} w(u) \)

(c) \[ \gamma_p \{W_n^*(hr - \frac{F(hr)}{F(r)}) \]
\[ + F((h - r_n) \frac{F_n(hr)}{F_n(r_n)}) r_n^2(\lambda \tilde{w})^T Z_n \}/F(r) \]

by (6), (4), proposition 2.2, and theorem 4.1.
\[ = \{X^\star(hr) - r \frac{F(hr)}{F(r)}\} \]
\[ + \{F(hr^2 \tilde{u}^T) - \frac{F(hr)}{F(r)} F(r^2 \tilde{u}^T)\} M - X^\star(r \tilde{u}) \}/F(r) \]
\[ = \{X^\star(hr) - r \frac{F(hr)}{F(r)}\} \]
\[ + \{F(hr^2 \tilde{u}^T) - \frac{F(hr)}{F(r)} F(r^2 \tilde{u}^T)\} M - X^\star(r \tilde{u}) \}/F(r) \]
\[ \text{since } X^\star(r \tilde{u}) = M \lambda W^{-1} Z \implies \]
\[ \frac{\lambda W^{-1} Z}{Z} = M - X^\star(r \tilde{u}) + Z \lambda , \text{ and} \]
\[ \{F(hr^2 \tilde{u}^T) - \frac{F(hr)}{F(r)} F(r^2 \tilde{u}^T)\} \lambda = F(hr) - \frac{F(hr)}{F(r)} F(r) = 0 \]
\[ = X^\star\left[ \frac{r}{F(r)} (h - G(h)) + r \tilde{u}^T M - \left[ G(hr \tilde{u}) - G(h) G(r \tilde{u}) \right] \right] \]
\[ = X^\star(\tilde{v}^\star(h - G(h))) = Z^\star(h) \]

where \( \tilde{v}^\beta \) is defined in (15).

It remains only to establish that the convergence in (c) holds uniformly in \( h \in \mathcal{H} \). By comparison of (b) and (c) and by theorem 4.1, it clearly suffices to show

\[ (d) \sup_{h \in \mathcal{H}} \{X^\star(hr_n) - r_n \frac{F_n(hr_n)}{F_n(r_n)}\} - X^\star(hr - \frac{F(hr)}{F(r)}) \to_p 0 \]

and

\[ (e) \sup_{h \in \mathcal{H}} F_n\left( (h - r_n \frac{F_n(hr_n)}{F_n(r_n)}) r_n^2 (W_n^\star(\lambda W^{-1} \tilde{u})(W_n^\star))^T \right) \]
\[ - F\left( (h - \frac{F(hr)}{F(r)}) r^2 (\lambda W^{-1} \tilde{u})^T \right) \to_p 0. \]

Now, since \( ||W_n^\star - \tilde{W}|| \leq ||W_n - \tilde{W}|| \to_p 0 \) by proposition 2.2 and \( F(h_n^2) < \infty \), (e) follows from a Glivenko - Cantelli theorem for \( F_n \) as in the proof of theorem 2.2.

To prove (d), note that the left side is bounded by

\[ (f) \sup_{h \in \mathcal{H}} \{X^\star(hr_n) - r_n \frac{F_n(hr_n)}{F_n(r_n)}\} - X^\star(hr_n - \frac{r_n F_n(hr_n)}{F_n(r_n)}) \]
\[ + \sup_{h \in \mathcal{H}} |W^*(hr_n - r_n \frac{F_n(hr_n)}{F_n(r_n)}) - W^*(hr - r \frac{F(hr)}{F(r)})| \]

\[ = I + II. \]

Then, since

\[ II \leq \sup_{h \in \mathcal{H}} |W^*(h(r_n - r))| + \sup_{h \in \mathcal{H}} \frac{F_n(hr_n)}{F_n(r_n)} |W^*(r_n) - \frac{F(hr)}{F(r)} W^*(r)| \]

where

(g) \[ ||h(r_n - r)||_{L^2(F)} \leq \frac{r_n}{r} - 1 ||w^2(h \sigma^2)1/2 \to 0 \] uniformly in \( h \in \mathcal{H} \)

and

(h) \[ |F_n(hr_n) - F(hr)| \leq |F_n(hr_n - hr)| + |(F_n - F)(hr)| \]

\[ \leq \frac{r_n}{r} - 1 ||w^2(h \sigma^2)1/2 \sum_{i=1}^{\infty} |\lambda_m - \lambda_i| F_i(h \sigma^2) \]

\[ \to 0 \] uniformly in \( h \in \mathcal{H} \),

and both (g) and (h) hold with \( h = 1 \), \( II \to_p 0 \) by uniform continuity of \( W^* \). To handle \( I \), note that

\[ I \leq \sup_{h \in \mathcal{H}} |W_n^*(hr_n) - W^*(hr_n)| + \sup_{h \in \mathcal{H}} \frac{F_n(hr_n)}{F_n(r_n)} |W_n^*(r_n) - W^*(r_n)| \]

\[ = A + B \]

where

\[ A \leq \sup_{h \in \mathcal{H}} |W_n^*(hr(\frac{r_n}{r} - 1))| + \sup_{h \in \mathcal{H}} |W_n^*(hr) - W^*(hr)| \]

\[ + \sup_{h \in \mathcal{H}} |W_n^*(hr(1 - \frac{r_n}{r}))| \]

\[ \leq \frac{r_n}{r} - 1 ||w^2(h \sigma^2)1/2 \sup_{h \in \mathcal{H}} |W_n^*(hr)| + ||W_n^*(r) - W^*(r)|| \]

\[ + \frac{r_n}{r} - 1 ||w^2(h \sigma^2)1/2 \sup_{h \in \mathcal{H}} |W^*(hr)| \]

(i) \( \to_p 0 \) by (6).

Since \( 1 \in \mathcal{H} \), \( B \to_p 0 \) by (h) and (i). Thus \( I \to_p 0 \), and hence (d) holds, which completes the proof of (17).

Now we establish the covariance formula (18). First suppose that \( C(h) = C(\tilde{h}) = 0 \). Now from (8) and (15) it follows that
(j) \[ \text{Cov}(Z^*(h), Z^*(\tilde{h})) = G(r^{-1}\tilde{V}(h)\tilde{V}^*(\tilde{h})) - G(\tilde{\omega}^T\tilde{V}(h))\lambda G(\tilde{\omega}\tilde{V}(\tilde{h})). \]

but

\[ G(\tilde{\omega}\tilde{V}(h)) = <\tilde{\omega}, rh>F(r) + <\tilde{\omega}, r\tilde{\omega}^T>\gamma M^{-1}C(hr\tilde{\omega}) \]

(k)

\[ = \left[ I + <\tilde{\omega}, r\tilde{\omega}^T>\gamma M^{-1} \right] G(hr\tilde{\omega}). \]

Now if \( M^- \) is the Moore-Penrose generalized inverse of \( M \), it follows from (2.14) of proposition 2.3 that \( MM^- = I - \theta \theta^T \) where \( \theta \) is the (only) normalized eigenvector of \( M \) with eigenvalue 0. Since \( M\lambda = 0 \) it follows that

\[ MM^- = \left[ \lambda^{-1} - <\tilde{\omega}, rh>\gamma M^{-1} \right] M^- = I - \lambda \lambda^T / \lambda^T \lambda \]

and hence

\[ G(\tilde{\omega}\tilde{V}(h)) = (\lambda^{-1}M^- + \lambda \lambda^T / \lambda^T \lambda)G(hr\tilde{\omega}) \]

(m)

\[ = \lambda^{-1}M^-G(hr\tilde{\omega}) \]

since \( \lambda^T G(hr\tilde{\omega}) = G(hr^{-1}) = G(h) = 0 \). Also,

\[ G(r^{-1}\tilde{V}(h)\tilde{V}(\tilde{h})) = G((\frac{1}{F(r)})h + \tilde{\omega}^T M^{-1}C(\tilde{\omega}\tilde{V}(\tilde{h})) \tilde{V}(\tilde{h}) \]

(n)

\[ = G(h\tilde{V}(\tilde{h})) + G(hr\tilde{\omega}^T)M^{-1}C(\tilde{\omega}\tilde{V}(\tilde{h})). \]

Thus

\[ \text{Cov}[Z^*(h), Z^*(\tilde{h})] = G(h\tilde{V}(\tilde{h})) + G(hr\tilde{\omega}^T)M^{-1}C(\tilde{\omega}\tilde{V}(\tilde{h}))) \]

\[ - G(hr\tilde{\omega}^T)M^{-1} \lambda^{-1} \lambda G(\tilde{\omega}\tilde{V}(\tilde{h}))) \]

(o)

\[ = G(h\tilde{V}(\tilde{h})). \]

Now for arbitrary \( h \), since \( Z^*(1) = 0 \),

\[ Z^*(h) = Z^*(h - G(h)) = Z^* (\tilde{V}(h - G(h))) \]

so from (o) it follows that

\[ \text{Cov}[Z^*(h), Z^*(\tilde{h})] = <h - G(h), \tilde{V}(\tilde{h} - G(\tilde{h}))>\gamma. \]

and hence (16) holds. \( \Box \)

**Proof of (11) of theorem 4.1:** (Identification of \( ZZ_a \)) From (1.20) we have \( W_n = E_n(\omega) \), and since \( \bar{W} = G(\omega) \) by definition, it follows that

\[ \sqrt{n}(W_n - \bar{W}) = ZZ_a(\omega) \]

where \( \max_{1 \leq s \leq s} ||w||_\infty \leq \max_{1 \leq s \leq s}(W_l / \lambda) < \infty \). Thus, on the one
hand by theorem 4.2
\[ \sqrt{n}(\hat{\lambda}_n - \lambda) \rightarrow_p \mathbb{Z}^*(\lambda) \]
\[ = \{\mathbb{X}^*(\lambda) - F(\lambda)\} + \left[ F(r^2\hat{\lambda}^T) - F(r) \right] M^{-\mathbb{X}^*}(\hat{\lambda}^T) \]
(a) \[ = \{\mathbb{X}^*(\lambda') - F(\lambda')\} + \left[ G((\lambda' - F(\lambda')r\hat{\lambda}^T)) \right] M^{-\mathbb{X}^*}(\hat{\lambda}^T) \]
while on the other hand, by (10)
(b) \[ \sqrt{n}(\hat{\lambda}_n - \lambda) \rightarrow_p \frac{\lambda^{-1}M^{-\mathbb{X}^*}(\hat{\lambda}^T)}{\lambda} = \mathbb{Z}_a\lambda \]
Thus the expressions on the right sides in (a) and (b) must be equal, and, upon multiplying across by \( \lambda^{-1}M^{-\mathbb{X}^*}(\hat{\lambda}^T) \), this yields
\[ \{\mathbb{X}^*(\lambda(r^{-1} - 1/F(\tau))) + G(\lambda(r^{-1} - 1/F(\tau))\hat{\lambda}^T) \}
\[ M^{-\mathbb{X}^*}(\hat{\lambda}^T) \] / F(\tau) \]
\[ = 1^T M^{-\mathbb{X}^*}(\hat{\lambda}^T) + \mathbb{Z}_a \]
or, since \( \mathbb{X}^*(\lambda) = 0 \),
\[ \{\mathbb{X}^*(\lambda/F(\tau)) + \left[ 1^T - G(\lambda\hat{\lambda}^T) \right] \}
\[ M^{-\mathbb{X}^*}(\hat{\lambda}^T) \] / F(\tau) = 1^T M^{-\mathbb{X}^*}(\hat{\lambda}^T) + \mathbb{Z}_a \]
so that
\[ \mathbb{Z}_a = \left[ \mathbb{X}^*(\lambda/F(\tau)) + \left[ 1^T - G(\lambda\hat{\lambda}^T) \right] \right] / F(\tau) - 1^T M^{-\mathbb{X}^*}(\hat{\lambda}^T) \]
\[ = \left[ \mathbb{X}^*(\lambda/F(\tau)) + \left[ (1 - F(\lambda))1^T - G(\lambda\hat{\lambda}^T) \right] \right] / F(\tau) \]
which proves (11). When \( F(\tau) = 1 \) (11) reduces to (12).

We close this section with some remarks on the estimation of the asymptotic variance of the process \( \sqrt{n}(\hat{\lambda}_n - \lambda) \). From the Glivenko Cantelli theorems similar to those used in section 2 one can derive conditions for the sample analogue of (18) (with \( \lambda \) and \( \hat{\lambda} \) replaced throughout - i.e. also in \( \hat{\lambda} \), \( \hat{\lambda}^T \) and \( \hat{\lambda}_n \) - by \( \hat{\lambda}_n \) and \( \hat{\lambda}_n \)) to converge in probability to (18) as \( n \to \infty \). In fact this (co-) variance estimator is also obtained by formal likelihood calculations (inversion of the \((n-1)x(n-1)\) matrix of second derivatives of the loglikelihood) continuing the derivation of \( \hat{\lambda}_n \) itself as a maximum likelihood estimator in the model where \( \lambda \) is discrete, with mass at the actual observations only.
5. EXAMPLES AND APPLICATIONS.

EXAMPLE 5.1. (Length biased sampling, Vardi (1982)). Let $X = R^+ = [0,\infty)$ and suppose $0 < \mu = \int x \, dG(x) < \infty$. Let $w_1(x) = 1$, so that the first sample is from $G$ itself, and let $w_2(x) = x$, so that the second sample is from the "length - biased" distribution $F_2(x) = \mu^{-1} \int_0^x y \, dG(y)$. Then $X^* = X$, $G^* = G$, and (2.4) holds. Furthermore $\mathbf{w} = (1,\mu)^T$, and letting $\lambda_1 = \lambda \in (0,1)$, $\lambda_2 = 1 - \lambda = \bar{\lambda}$,

$$r(x) = (\lambda + \frac{x}{\mu})^{-1}.$$ 

Thus we have

$$M = \begin{pmatrix} 1/\lambda - F(r^2) & -F(r^2 \bar{w}_2) \\ -F(r^2 \bar{w}_2) & 1/\lambda - F(r^2 \bar{w}_2^2) \end{pmatrix} = \begin{pmatrix} \bar{\lambda}K/\lambda - K \\ -K & \lambda K/\lambda \end{pmatrix},$$

since $M\Lambda = 0$, where

(1) $K = F(r^2 \bar{w}_2) = G(r \bar{w}_2) = \int_0^\infty \frac{x}{\mu \lambda + x/\mu} dG(x)$.

Also define

(2) $K(x) = \int_0^x \bar{w}_2 r \, dG = G(1_{[0,x]} \bar{w}_2 r)$

and note that

(3) $G(1_{[0,x]}r) = \frac{1}{\lambda}(G(x) - \bar{\lambda}K(x))$.

Then by (4.18) with $K(s,t) = K(1_{[0,s]}1_{[0,t]})$,

$$K(s,t) = G(r(1_{[0,s]} - G(s))(1_{[0,t]} - G(t))$$

$$+ G(r1_{[0,s]} - G(s)) \frac{\lambda}{\lambda K} G(r1_{[0,t]} - G(t)))$$

by using $M^- = \begin{pmatrix} \frac{\lambda}{\lambda K} & 0 \\ 0 & 0 \end{pmatrix}$

as the {1,2} - inverse of $M$; recall proposition 2.3.

(4) $\frac{1}{\lambda} \left[ G(s \land t) - G(s)G(t) \right] - \frac{\bar{\lambda}}{\lambda K} \left[ \frac{K(s \land t)}{K} - \frac{K(s)K(t)}{K} \right]$,

in agreement with equation (3.6) of Vardi (1982).
Moreover \( G(r\tilde{w}) = (G(r), K)^T = (1 - \lambda K, K)^T \), and hence, using \( K^n = W_1^{-1} M^n \), and the same version of \( M^n \),

\[
\left[ K^n \tilde{w} - (1 + \tilde{w}^T M^n G(r \tilde{w})) \tilde{w} \right] \tau
\]

\[
= \begin{bmatrix}
0 \\
\frac{x}{\lambda K}
\end{bmatrix} \tau - \frac{1}{\lambda} \tau^-1 \begin{bmatrix} 1 \\
\mu
\end{bmatrix} = \begin{bmatrix}
0 \\
\frac{x}{\lambda K}
\end{bmatrix} \tau - \frac{1}{\lambda} \begin{bmatrix} 1 \\
\mu
\end{bmatrix}
\]

and hence the limit \( rv \ ZW \) of theorem 4.1 is

\[
X^* \begin{bmatrix}
0 \\
\frac{x}{\lambda K}
\end{bmatrix} = \begin{bmatrix} 0 \\
X^* \left( \frac{x}{\lambda K} \right)
\end{bmatrix}
\]

where, by (4.8),

\[
X^* \left( \frac{x}{\lambda K} \right) \equiv N(0, \frac{1}{\lambda K^2} \left[ G(x^2) - C(xr(1, x)) \lambda C(xr(1, x))^T \right])
\]

\[= N(0, \frac{\lambda^2 (1 - K)}{\lambda K}) \]

by straightforward calculation, which agrees with (3.3) of Vardi (1982).

**Example 5.2.** (\( s = 1 \); biasing with only one stratum). For a general sample space \( X \) suppose that \( s = 1 \) and that \( w_1 = w \) satisfies both \( C(w) < \infty \) and \( W_{-1} = C(w^{-1}) < \infty \). Then condition (2.4) is trivially satisfied, and, letting \( W = C(w), \ r = W/w = \tilde{w}^{-1}, \ F(r) = C(X^*) \) may be \( = 1 \) or \( < 1 \) depending on \( w \) and \( C \). Since \( M = 0 \) and \( r\tilde{w} = 1 \), theorem 4.1 yields

\[
\sqrt{n} (W_n - W) \rightarrow_d -X^* (r/F^2(r)) W \equiv N(0, \frac{W^2}{F(r)^4} (W_{-1} W - 1)).
\]

Furthermore, the process \( Z^* \) of theorem 4.2 reduces to

\[
Z^* (h) = X^* (r (h - C^*) / F(r))
\]

\[= X^* (r (h - C^*)) \quad \text{when} \ F(r) = 1 \]

with covariance function

\[
K(h, \tilde{h}) = C^* (r (h - C^*) (\tilde{h} - C^*) ) / F(r)
\]

\[= C(r (h - C^*) (\tilde{h} - C^*)) \quad \text{when} \ F(r) = 1 \]

\[
= W_{-1} W \left[ \frac{C(h \tilde{h})}{W_{-1}} - C(h) \frac{C(\tilde{h})}{W_{-1}} \right] + \left[ C(h \tilde{h}) - \frac{C(h)}{W_{-1}} C(\tilde{h}) \right].
\]
These formulas agree with the results of Vardi (1985a) section 7(ii) in the case \( F(r) = 1 \). The further special case \( u(z) = z \) (which is also the special case \( \lambda = 0 \) in example 1) was considered by Cox (1969).

**EXAMPLE 5.3.** (Truncated sampling or restricted measurement). This is a further special case of example 2. For a general sample space \( \mathbf{X} \) suppose that \( s = 1 \) and that \( w_1(x) = 1_C(x) \) where \( C \in \mathbf{B} \), \( C \neq \mathbf{X} \), and \( C(C) < 1 \). Then \( \mathbf{X}^* = C \), \( w_1 = C(C) < 1 \), \( r(x) = C(C)1_C(x) \), and \( C^* = G/G(C) \) is simply the conditional distribution \( G(\cdot | C) \). Thus

\[
C^*(A) = \frac{C(A \cap C)}{C(C)} = F(A) \quad \text{for} \quad A \in \mathbf{B} \cap C,
\]

and the estimator \( E_n^* \) of \( C^* \) is simply \( F_n \). Note that \( w_1 = G(C) \) is not identifiable in this situation.

**EXAMPLE 5.4.** ("Choice - based" sampling in econometrics; "case - control studies" in biostatistics). Suppose that \( \mathbf{X} = (Y, Z) \) where \( Y \) takes values in \( \{1, \ldots, M\} \) and \( Z \sim \mu \) with density \( h \) with respect to \( \mu \) is a covariate vector with values in \( Z \subset \mathbf{R}^p \). The basic (unbiased or prospective) model \( C \) has density

\[(9) \quad g(y, z) = p_\theta(y|z)h(z)\]

so that

\[(10) \quad C(\{y\} \times A) = \int_A p_\theta(y|z)dH(z)\]

for \( y = 1, \ldots, M \) and \( A \in \mathbf{B}(\mathbf{R}^p) \) where \( p_\theta(y|z) = P_\theta(Y = y|Z = z) \) is a parametric (finite - dimensional) model. A frequent choice is the logistic regression model

\[(11) \quad p_\theta(y|z) = \frac{\exp(\alpha_y + \beta^T_y z)}{1 + \sum_{y'} \exp(\alpha_{y'} + \beta^T_{y'} z)}\]

with \( \theta = (\alpha, \beta) \in \mathbf{R}^{(p+1)M} \).

The biased (or retrospective) sampling model \( F \) is obtained from \( G \) via the weight functions \( w_i(x) = w_i(y) = 1_{D_i}(y) \) where \( D_i \subset \{1, \ldots, M\} \) for \( i = 1, \ldots, s \). This again yields a semi-parametric submodel, since only distributions \( G \) of the form (10) are considered.

One case of particular interest is that of "pure choice - based sampling" in the terminology of Coslett (1981). In this sampling scheme, the strata \( D_i \) are taken to be just \( D_i = \{i\}, \)
In this case the graph $M^*$ of section 2 is not connected: in fact $G(w_i w_j) = 0$ for all $i \neq j$ and hence there is no unique nonparametric MLE $C_n$ of $G$ for this sampling scheme. Manski and Lerman (1977) avoid this difficulty of pure choice - based sampling by assuming that the "aggregate shares" $G(y) = G((y|x) \times Z) = \int p_\theta(y|x) dH(z)$, $y = 1, \ldots, M$ are known. Note that for this biasing system we can view $F$ as a biased distribution derived from $H$ with new biasing (weight) functions $W^*_i = G((y|z))$, $y = 1, \ldots, M$, depending on the unknown parameter $\theta$. Then $W^*_i = G(y)$, typically the condition on $M^*$ for these $w^*$'s will hold, and if $\theta$ is known the methods of the preceding sections yield estimates of $H$ together with the asymptotic behavior of the estimates.

This same pure choice - based sampling design is also frequently used in "case - control studies" in biostatistics where the $y$ 's often denote different disease categories. In the biostatistics applications interest centers on odd's ratios which can be estimated from purely choice - based sampling in spite of the fact that $G$ itself cannot be estimated; see e.g. Prentice and Pyke (1979), who examine the case of (11), and Breslow and Day (1980). If the "pure choice - based" design is "enriched" by taking $s = M + 1$, $\lambda + 1 = 0$, and choosing $W_{M+1}(y) = 1_{1, \ldots, M}(y)$, then (2.4) holds and the nonparametric MLE $C_n$ of $G$ exists (a.s. for $n \geq \text{some } N_\omega$) and is unique. See example 5.

For general $D_i$'s the biased distribution $F$ has density

$$f(y,z,i) = \lambda_i \frac{1_{D_i}(y)p_\theta(y|z)h(z)}{\int \sum_{y=1}^{M} 1_{D_i}(y)p_\theta(y'|z')h(z')d\mu(z')}$$

and the condition (2.4) for existence of a unique solution is precisely Coslett's (1981) Assumption 10:

$$\left( \cup_{i \in B} D_i \right) \cap \left( \cup_{i \in B^c} D_i \right) \neq \emptyset \quad \text{for every proper subset } B \text{ of } \{1, \ldots, s\};$$

see Vardi (1985a) sections 2 and 8.

For known $\theta$, efficient estimates of $H$ and their asymptotic behavior via the preceding sections can be obtained as follows: The marginal distribution of $(z,i)$ is

$$f(z,i) = \lambda_i \frac{\left[ \sum_{y=1}^{M} 1_{D_i}(y)p_\theta(y'|z) \right] h(z)}{\int \left[ \sum_{y=1}^{M} 1_{D_i}(y)p_\theta(y'|z') \right] h(z')d\mu(z')}.$$
where the new biasing functions \( w_i'(z) = w_i'(z; \theta) \) depend on \( \theta \).
Thus if \( \theta \) is known, the methods of Vardi (1985a) and the preceding sections apply to yield efficient estimates of \( H \), which can, in turn, be used to construct efficient estimates of \( \theta \). This method is implicit in Cosslett (1981) section 4, and will be discussed in more detail in Bickel, Klaassen, Ritov, and Wellner (1986).

**EXAMPLE 5.5.** ("Enriched" stratified sampling). Let \( X \) be a general sample sample space and suppose that \( D_1, \ldots, D_s \) form a (measurable) partition of \( X \): \( D_i \cap D_j = \emptyset \) for \( i \neq j \) and \( \bigcup_{i=1}^{s-1} D_i = X \).

Let \( w_i(x) = 1_{D_i}(x) \) for \( i = 1, \ldots, s-1 \) and suppose that \( w_s(x) = 1. \) Thus the stratified sample from \( D_1, \ldots, D_{s-1} \) is "enriched" by sampling from all of \( X \) with sampling fraction \( \lambda_s > 0 \); this terminology is that of Cosslett (1981).

For this sampling scheme (2.4) holds (assuming without loss that \( G(D_i) > 0 \), \( i = 1, \ldots, s-1 \), \( F(r) = 1 \), and we have

\[
W_i = G(D_i), \quad i = 1, \ldots, s-1 \\
\rho(x) = \sum_{i=1}^{s-1} \frac{1}{\lambda_s + \lambda_i / G(D_i)} 1_{D_i}(x)
\]

so that the upper left \( s-1 \times s-1 \) submatrix of \( F(\tilde{r} \tilde{w} \tilde{w}^T) \) is diagonal with elements

\[
F(\tilde{r} \tilde{w}^2) = \frac{1}{\lambda_s G(D_i) + \lambda_i}
\]

Hence the upper left \( s-1 \times s-1 \) submatrix of \( M \) is diagonal with elements

\[
M_{ii} = \frac{1}{\lambda_i} - \frac{1}{\lambda_s G(D_i) + \lambda_i} = \frac{\lambda_s G(D_i)}{\lambda_i(\lambda_s G(D_i) + \lambda_i)}
\]

for \( i = 1, \ldots, s-1 \), and a \( \{1,2\} \) - inverse \( M^- \) of \( M \) is given by the diagonal matrix with last row and column containing all zeros and having diagonal elements \( M_{ii}^{-1} \); recall proposition 2.3 and (2.15). Thus, \( K^- = \sum \lambda_i M^- \) is also diagonal with last row and column all zero and diagonal entries

\[
K^-_{ii} = \frac{1}{\lambda_i} \left( \lambda_s G(D_i) + \lambda_i \right) = \frac{1}{\lambda_i} a_i
\]
for \( i = 1, \ldots, s - 1 \). Similarly,
\[
G(\tilde{w}_i) = \frac{G(D_i)}{a_i},
\]
and hence \( M^{-1}G(\tilde{w}) = [\lambda_1/\lambda_s, \ldots, \lambda_{s-1}/\lambda_s, 0] \).
Therefore
\[
(1 + \tilde{w}^T M^{-1}G(\tilde{w}))W = \left(1 + \sum_{j=1}^{s-1} \frac{1/a_j}{G(D_i)/\lambda_s} \right) \frac{1}{\lambda_s} r W
\]
which is constant in \( x \), and hence the limiting random vector in theorem 4.1 becomes \( K^* \mathcal{N}(\tilde{w}) \), the last element of which is 0 by the form of \( K^* \), and where the first \( s-1 \) elements of \( \mathcal{N}(\tilde{w}) \) have an \((s-1)\times(s-1)\) covariance matrix
\[
\begin{align*}
(16) \quad C &= a - (a, aG(D))\tilde{\lambda}(a, aG(D))^T \\
&= a \left[ a^{-1} - (I, \tilde{\lambda})(I, \tilde{\lambda})^T \right] a,
\end{align*}
\]
where \( p = G(D) = (G(D_1), \ldots, G(D_s))^T \). Thus by theorem 4.1, and straightforward calculation using (15) and (16), it follows that
\[
\sqrt{n}(\mathcal{N}_n(D) - G(D)) \to_d N_{s-1}(-1/\lambda_s, (G(D) - G(D)G(D)^T)).
\]
Note that this is just the covariance matrix for the usual (multinomial) estimate of \( G(D) \) from a random sample of size \( n \lambda_s \) from all of \( X \). In other words, sampling within the strata \( D_i \) does not help in estimating the strata probabilities \( G(D) \).

**EXAMPLE 5.6.** (Stratified or truncated regression). This interesting and rich family of semiparametric submodels of the general biased sampling model begin with ordinary linear regression with unknown error distribution \( G_0 \) as the basic (unbiased) model: Suppose that \( X = (Y, Z) \equiv G \) where \( Y = \theta^T Z + \varepsilon \) with \( \varepsilon \equiv G_0 \) with density \( g_0 \) with respect to Lebesgue measure and \( Z \equiv H \) independent of \( \varepsilon \) with density \( h \) with respect to \( \mu \). Thus \( G \) has density
\[
g(y, z) = g_0(y - \theta^T z)h(z).
\]
The biased sampling model is typically determined by weight functions \( w_i(x) = w_i(y) = 1_{D_i}(y), \ i = 1, \ldots, s \), where the \( D_i \)'s are disjoint subintervals of \( R^1 \). The case of \( s = 1 \) and \( D_1 = (-\infty, \gamma_0] \),
which is also a special case of example 3, has been considered by Bhattacharya, Chernoff, and Yang (1983). Jewell (1985) considers the case $s = 2$ and $D_1 = (-\infty, y_0]$, $D_2 = (y_0, \infty)$ in which condition (2.4) fails, so a unique completely nonparametric estimator of $C$ does not exist in view of Vardi's theorem 1.1. Nevertheless the parameters $\theta$, $C_0$, and $H$ are identifiable in this model, and for known $\theta$ the methods of Vardi (1985a) can be applied iteratively by first regarding $C_0$ as known and absorbing it into the biasing functions and estimating $H$, and then by treating $H$ as known and absorbing it into the biasing functions and estimating $C_0$, and so forth. This type of semiparametric submodel of the biased sampling model will be treated by Bickel, Klaassen, Ritov, and Wellner (1986).
BIBLIOGRAPHY


