SECTION 14

Biased Sampling

Vardi (1985) introduced a far-reaching extension of the classical model for length-biased sampling. He solved the problem of estimating a distribution function based on several independent samples, each subject to a different form of selection bias. Using empirical process theory, Gill, Vardi and Wellner (1988) developed the asymptotic theory for generalizations of Vardi's method to abstract settings. They showed that the general model includes many interesting examples as special cases. This section presents a reworking of the ideas in those two papers. It is part of a study carried out by me in collaboration with Robert Sherman of Yale University.

The general problem is to estimate a distribution P on some set S using independent samples of sizes n_{i+} from distributions Q_i, for $i = 1, \ldots, s$, where the Q_i are related to P by means of known nonnegative weight functions $W_i(\cdot)$ on S:

$$\frac{dQ_i}{dP} = \pi_i W_i(\cdot) \qquad \text{where } \pi_i = 1/PW_i.$$

Of course the normalizing constants π_i, which we must assume to be finite and strictly positive, are unknown. For example, the W_i might be indicator functions of various subdomains of S. The problem is then one of combining the different samples in order to form an estimate of P over the whole of S. The difficulty lies in deciding how to combine the information from samples whose subdomains overlap.

For the general problem, to ensure that we get information about P over the whole domain, we must assume that the union of the sets $\{W_i > 0\}$ covers S.

Vardi suggested that a so-called nonparametric maximum likelihood estimator $\widehat{P}_n$ be used. This is a discrete probability measure that concentrates on the combined observations $x_1, x_2, \ldots$ from all s samples. If x_j appears a total of n_{ij} times in the i^{th} sample, the combined empirical measure $\widehat{Q}_n$ puts mass n_{+j}/n at x_j, where

$$n_{+j} = \sum_i n_{ij} \qquad \text{and} \qquad n = \sum_{i,j} n_{ij}.$$

The estimator $\widehat{P}_n$ modifies $\widehat{Q}_n$, putting at x_j the mass $\widehat{p}_j$ defined by maximization

of a pseudo log-likelihood function: maximize

$$\sum_{i,j} n_{ij} \left[\log p_j - \log\left(\sum_k W_i(x_k) p_k \right) \right],$$

subject to the constraints

$$p_j > 0 \quad \text{for each } j, \qquad \text{and} \quad \sum_j p_j = 1.$$

In this form the estimation problem involves parameters whose number increases with the sample sizes. The first part of the analysis will show how to transform the problem into an equivalent maximization involving only a fixed number of unknown parameters.

Simplify the notation by writing W_{ik} for $W_i(x_k)$. Reparametrize by substituting $\exp(\beta_j)$ for p_j. Then we need to maximize the function

$$L_n(\boldsymbol{\beta}) = \sum_j n_{+j}\beta_j - \sum_i n_{i+} \log\left(\sum_k W_{ik} \exp(\beta_k) \right)$$

over all real $\{\beta_j\}$, subject to the constraint

$$\sum_j \exp(\beta_j) = 1.$$

Let $\mathbf{1}$ denote a vector of ones. The criterion function L_n is constant along the lines $\{\boldsymbol{\beta} + t\mathbf{1} : t \in \mathbb{R}\}$; the constraint serves to locate a unique point on each such line.

Simple calculus shows that L_n is a concave function. Indeed, for each fixed $\boldsymbol{\beta}$ and $\boldsymbol{\delta}$ the function $L_n(\boldsymbol{\beta} + t\boldsymbol{\delta})$ has derivative

$$\sum_j n_{+j}\delta_j - \sum_i n_{i+} \left(\frac{\sum_k W_{ik}\delta_k \exp(\beta_k)}{\sum_k W_{ik} \exp(\beta_k)} \right) \qquad \text{at } t = 0, \tag{14.1}$$

and second derivative

$$-\sum_{i,k} n_{i+} B_{ik} (\delta_k - \bar{\delta}_i)^2 \qquad \text{at } t = 0,$$

where

$$B_{ik} = W_{ik} \exp(\beta_k) / \sum_j W_{ij} \exp(\beta_j)$$

and $\bar{\delta}_i$ is the weighted average

$$\bar{\delta}_i = \sum_k B_{ik}\delta_k.$$

Clearly the second derivative is always nonpositive; the function L_n is concave along every line. The second derivative can equal zero only if δ_k is constant over each of the subsets

$$K(i) = \{k : W_{ik} > 0\} \qquad \text{for } i = 1, \ldots, s.$$

Under mild connectedness assumptions about the regions $\{W_i > 0\}$, it can be shown (almost surely as the n_{i+} tend to infinity) that constancy over each $K(i)$ forces $\boldsymbol{\delta}$ to

be a multiple of **1**. That is, L_n is strictly concave along all directions except the **1** direction. Moreover, the connectedness assumption also forces the derivative to be strictly negative for t large enough. It follows that the constrained maximization problem eventually has a unique solution $\widehat{\boldsymbol{\beta}}$. (Clearly $\widehat{\boldsymbol{\beta}}$ depends on n, but I will omit the subscript n to avoid notational clutter.)

For a precise statement of the necessary connectedness property, see pages 1071-1072 of Gill et al. (1988). Let us assume such a property to hold from now on.

Transformation to an equivalent problem. The function L_n must have all its directional derivatives equal to zero at its maximizing point. Putting $\boldsymbol{\delta}$ equal to a vector with δ_j as its only nonzero component, we get from (14.1) that

$$\exp(\widehat{\beta}_j) = \frac{n_{+j}}{\sum_i (n_{i+} W_{ij} / \sum_k W_{ik} \exp(\widehat{\beta}_k))} \qquad \text{for each } j. \tag{14.2}$$

Notice that the s linear combinations of the $\exp(\widehat{\beta}_k)$ values on the right-hand side determine all the $\widehat{\beta}_j$ values. That is why we will be able to reduce the problem to one involving only s unknown parameters.

Introduce new parameters $\alpha_1, \ldots, \alpha_s$. Trivially, the constrained maximization of L_n is equivalent to the problem: maximize

$$\sum_j n_{+j} \beta_j + \sum_i n_{i+} \alpha_i$$

subject to the constraints

$$\sum_j \exp(\beta_j) = 1,$$

$$\exp(-\alpha_i) = \sum_j W_{ij} \exp(\beta_j) \qquad \text{for each } i.$$

Equality (14.2) translates into a set of relations that the maximizing $\widehat{\boldsymbol{\alpha}}$ and $\widehat{\boldsymbol{\beta}}$ must satisfy; the maximization problem is unaffected if we add another constraint,

$$\exp(\beta_j) = \frac{n_{+j}}{\sum_i n_{i+} W_{ij} \exp(\alpha_i)} \qquad \text{for each } j,$$

to the list. This allows us to eliminate the $\{\beta_j\}$ from the problem altogether, leaving a constrained maximization over the $\{\alpha_i\}$: maximize

$$\sum_i n_{i+} \alpha_i - \sum_j n_{+j} \log\Big(\sum_k n_{k+} W_{kj} \exp(\alpha_k) \Big),$$

subject to the constraints

$$\sum_j \frac{n_{+j}}{\sum_i n_{i+} W_{ij} \exp(\alpha_i)} = 1,$$

$$\exp(-\alpha_i) = \sum_j \frac{W_{ij} n_{+j}}{\sum_k n_{k+} W_{kj} \exp(\alpha_k)} \qquad \text{for each } i.$$

Just as the addition of an extra constraint did not affect the previous maximization, so will the elimination of a constraint not affect this maximization. By marvellous good luck (What is going on here?) the last set of equations corresponds exactly to the requirement that the directional derivatives of the criterion function all equal zero at its global maximizing value $\widehat{\boldsymbol{\alpha}}$; it can be discarded without changing the problem. The remaining constraint then serves only to locate a unique point along the lines of constancy of the criterion function.

The Vardi procedure takes a much neater form when expressed in empirical process notation. Write λ_{ni} for the proportion n_{i+}/n of observations that belong to the i^{th} sample, and $h_n(\cdot, \boldsymbol{\alpha})$ for the function $\left(\sum_i \lambda_{ni} \exp(\alpha_i) W_i(\cdot)\right)^{-1}$. Then the Vardi estimator is determined by: maximize

$$M_n(\boldsymbol{\alpha}) = \boldsymbol{\lambda}_n' \boldsymbol{\alpha} + \widehat{Q}_n \log h_n(\cdot, \boldsymbol{\alpha})$$

subject to the constraint

$$\widehat{Q}_n h_n(\cdot, \boldsymbol{\alpha}) = 1.$$

Under the connectedness assumptions mentioned earlier, the function M_n is (almost surely, with increasing sample sizes) strictly concave along all directions except those parallel to $\mathbf{1}$, along which it is constant. [Recycled notation.] The constraint locates the unique maximizing $\widehat{\boldsymbol{\alpha}}$ along a line of constancy. The measure $\widehat{P}_n$ is determined by putting mass

$$\widehat{p}_j = \exp(\widehat{\beta}_j) = \frac{n_{+j}}{n} h_n(x_j, \widehat{\boldsymbol{\alpha}}) \qquad \text{at } x_j.$$

That is, $\widehat{P}_n$ has density $h_n(\cdot, \widehat{\boldsymbol{\alpha}})$ with respect to the empirical measure $\widehat{Q}_n$.

Heuristics. The estimator $\widehat{P}_n$ is partly parametric and partly nonparametric. The $\widehat{\boldsymbol{\alpha}}$ is determined by a finite-dimensional, parametric maximization problem. It determines the density of $\widehat{P}_n$ with respect to the nonparametric estimator $\widehat{Q}_n$. Limit theorems for $\widehat{P}_n$ will follow from the parametric limit theory for $\widehat{\boldsymbol{\alpha}}$ and the nonparametric limit theory for $\widehat{Q}_n$.

To simplify the analysis let us assume that the proportions are well behaved, in the sense that $\lambda_{ni} \to \lambda_i > 0$ as $n \to \infty$, for each i. This assumption could be relaxed. Let $\widehat{Q}_{ni}$ denote the empirical measure for the i^{th} sample (mass n_{ij}/n_{i+} on each observation from Q_i). We should then have

$$\widehat{Q}_n = \sum_i \lambda_{ni} \widehat{Q}_{ni} \to \sum_i \lambda_i Q_i$$

for some mode of convergence. Call the limit measure Q. For each integrable function f,

$$Qf = \sum_i \pi_i \lambda_i P(f W_i);$$

the measure Q has density $G(\cdot) = \sum_i \pi_i \lambda_i W_i(\cdot)$ with respect to P. The function $h_n(\cdot, \boldsymbol{\alpha})$ converges pointwise to

$$h(\cdot, \boldsymbol{\alpha}) = \left(\sum_i \lambda_i \exp(\alpha_i) W_i(\cdot)\right)^{-1}.$$

Notice that $G(\cdot) = 1/h(\cdot, \boldsymbol{\alpha}^*)$, where $\boldsymbol{\alpha}^*$ is determined by

$$\exp(\alpha_i^*) = \pi_i \qquad \text{for } i = 1, \ldots, s.$$

It would seem reasonable that the limiting behavior of $\widehat{\boldsymbol{\alpha}}$ should be obtained by solving the limiting form of the constrained maximization problem. That is, $\widehat{\boldsymbol{\alpha}}$ should converge to the $\boldsymbol{\alpha}$ that maximizes

$$M(\boldsymbol{\alpha}) = \boldsymbol{\lambda}'\boldsymbol{\alpha} + Q \log\bigl[G(\cdot)h(\cdot, \boldsymbol{\alpha})\bigr],$$

subject to the constraint

$$Qh(\cdot, \boldsymbol{\alpha}) = 1.$$

The extra factor G contributes a centering of the log term; each product $G(\cdot)h(\cdot, \boldsymbol{\alpha})$ is bounded away from zero and infinity. This ensures that $M(\boldsymbol{\alpha})$ is well defined for every $\boldsymbol{\alpha}$, without affecting the location of the maximizing value.

Calculation of first and second directional derivatives, in much the same way as before, shows that M is concave. The connectedness assumption implies strict concavity, except along the $\mathbf{1}$ direction, along which it is constant. Modulo $\mathbf{1}$, it has a unique maximizing value, determined by setting all the partial derivatives

$$\begin{aligned}\frac{\partial M}{\partial \alpha_i} &= \lambda_i - Q\left(\frac{\lambda_i \exp(\alpha_i) W_i}{\sum_k \lambda_k \exp(\alpha_k) W_k}\right) \\ &= \lambda_i - \lambda_i \exp(\alpha_i) P(W_i G h(\cdot, \boldsymbol{\alpha}))\end{aligned}$$

to zero. Since $1/G(\cdot) = h(\cdot, \boldsymbol{\alpha}^*)$, these derivatives are zero at $\boldsymbol{\alpha} = \boldsymbol{\alpha}^*$, and the constraint is satisfied:

$$Qh(\cdot, \boldsymbol{\alpha}^*) = P(G/G) = 1.$$

It follows that $\boldsymbol{\alpha}^*$ uniquely solves the limiting constrained maximization problem.

If $\widehat{\boldsymbol{\alpha}}$ does converge to $\boldsymbol{\alpha}^*$ then the density $h_n(\cdot, \widehat{\boldsymbol{\alpha}})$ of $\widehat{P}_n$ with respect to $\widehat{Q}_n$ will converge pointwise to $h(\cdot, \boldsymbol{\alpha}^*) = 1/G(\cdot)$. For a fixed integrable f we should then have

$$\widehat{P}_n f = \widehat{Q}_n\bigl(f(\cdot)h(\cdot, \widehat{\boldsymbol{\alpha}})\bigr) \approx Q(f/G) = Pf.$$

A precise formulation of these heuristic approximations will establish a central limit theorem for $\widehat{P}_n$ as an estimator for P.

Asymptotic behavior of $\widehat{\boldsymbol{\alpha}}$. Decompose $\widehat{\boldsymbol{\alpha}}$ into a sum $\boldsymbol{\alpha}^* + \widehat{\boldsymbol{\delta}}/\sqrt{n} + \widehat{\epsilon}\,\mathbf{1}$, where the random vector $\widehat{\boldsymbol{\delta}}$ lies in the subspace $\mathcal{D}$ of vectors in $\mathbb{R}^s$ that are orthogonal to $\mathbf{1}$. Constancy of M_n along the $\mathbf{1}$ directions lets us ignore the $\widehat{\epsilon}$ in the maximization; the vector $\widehat{\boldsymbol{\delta}}$ maximizes the concave function

$$H_n(\boldsymbol{\delta}) = n\bigl(M_n(\boldsymbol{\alpha}^* + \boldsymbol{\delta}/\sqrt{n}) - M_n(\boldsymbol{\alpha}^*)\bigr)$$

over $\boldsymbol{\delta}$ in $\mathcal{D}$. The constraint may be written as

$$\exp(\widehat{\epsilon}) = \widehat{Q}_n h_n(\cdot, \boldsymbol{\alpha}^* + \widehat{\boldsymbol{\delta}}/\sqrt{n}). \tag{14.3}$$

Equivalently, for each integrable f,

$$\widehat{P}_n f = \frac{\widehat{Q}_n f h_n(\cdot, \boldsymbol{\alpha}^* + \widehat{\boldsymbol{\delta}}/\sqrt{n})}{\widehat{Q}_n h_n(\cdot, \boldsymbol{\alpha}^* + \widehat{\boldsymbol{\delta}}/\sqrt{n})}. \tag{14.4}$$

The denominator has the interpretation of a normalization factor needed to make $\widehat{P}_n$ a probability measure.

The asymptotic behavior of $\widehat{\delta}$ will be controlled by a quadratic approximation to H_n. To develop the approximation we decompose the empirical measures into deterministic parts (for which Taylor expansion to quadratic terms is appropriate) plus smaller perturbations due to the empirical processes. Define

$$\nu_{ni} = \sqrt{n_{i+}}(\widehat{Q}_{ni} - Q_i) \qquad \text{for } i = 1, \ldots, s,$$
$$\nu_n = \sqrt{n}(\widehat{Q}_n - \mathbb{P}\widehat{Q}_n) = \sum_i \sqrt{\lambda_{ni}}\,\nu_{ni}.$$

Here $\mathbb{P}\widehat{Q}_n$ represents the measure $\sum_i \lambda_{ni} Q_i$, which has density

$$G_n(\cdot) = \sum_i \pi_i \lambda_{ni} W_i(\cdot) = 1/h_n(\cdot, \boldsymbol{\alpha}^*)$$

with respect to P. For each f we have a decomposition

$$\widehat{Q}_n f = P(G_n f) + \frac{1}{\sqrt{n}}\nu_n f. \tag{14.5}$$

If $P(Gf^2) < \infty$, the random component has an asymptotic normal distribution,

$$\nu_n f = \sum_i \sqrt{\lambda_{ni}}\,\nu_{ni} f \rightsquigarrow N(0, \sigma^2(f)), \tag{14.6}$$

where

$$\sigma^2(f) = \sum_i \lambda_i \left(Q_i f^2 - (Q_i f)^2\right) = P(Gf^2) - \sum_i \lambda_i \pi_i^2 (PW_i f)^2.$$

A similar multivariate central limit theorem would hold for each vector-valued function $\mathbf{f}$ with $P(G|\mathbf{f}|^2)$ finite.

Substituting for $\widehat{Q}_n$ in the definition of M_n, using (14.5), we get

$$\begin{aligned} H_n(\boldsymbol{\delta}) = &\left[\sqrt{n}\boldsymbol{\lambda}_n'\boldsymbol{\delta} + nPG_n \log\left(G_n h_n(\cdot, \boldsymbol{\alpha}^* + \boldsymbol{\delta}/\sqrt{n})\right)\right] \\ &+ \sqrt{n}\nu_n \log\left(G_n h_n(\cdot, \boldsymbol{\alpha}^* + \boldsymbol{\delta}/\sqrt{n})\right) \end{aligned} \tag{14.7}$$

Fix $\boldsymbol{\delta}$. Calculation of first and second derivatives, in much the same way as for L_n, shows that the deterministic contribution (the first term on the right-hand side) is of the form $-\frac{1}{2}\boldsymbol{\delta}' V \boldsymbol{\delta} + o(1)$ as $n \to \infty$, where V equals $\operatorname{diag}(\lambda_i)$ minus the $s \times s$ matrix whose $(i,j)^{th}$ element is $\pi_i \pi_j \lambda_i \lambda_j\, P(W_i W_j / G)$. Of course $V\mathbf{1} = \mathbf{0}$, but the connectedness assumption ensures that V acts as a positive definite linear transformation on the subspace $\mathcal{D}$.

The term linear in $\boldsymbol{\delta}$ is contributed by the random perturbation (the second term on the right-hand side of (14.7)). Again a Taylor expansion gives

$$\log\left(G_n h_n(\cdot, \boldsymbol{\alpha}^* + \boldsymbol{\delta}/\sqrt{n})\right) = \frac{1}{\sqrt{n}}\boldsymbol{\delta}' \mathbf{D}_n(\cdot) + \rho_n(\cdot),$$

where $\mathbf{D}_n$ is an $s \times 1$ vector of uniformly bounded functions,

$$D_{ni}(\cdot) = \frac{\pi_i \lambda_{ni} W_i(\cdot)}{G_n(\cdot)},$$

and ρ_n is a remainder function less than $|\delta|^2/n$ in absolute value. For fixed δ, the contribution to $H_n(\delta)$ from ρ_n converges in probability to zero, because

$$\begin{aligned}\operatorname{var}(\nu_n\rho_n) &= \operatorname{var}\Big(\sum_i \sqrt{\lambda_{ni}}\nu_{ni}\rho_n\Big)\\ &\le \sum_i \lambda_{ni}Q_i(\rho_n^2)\\ &\le |\delta|^4/n^2.\end{aligned}$$

The remainder term $\sqrt{n}\nu_n\rho_n$ is actually of order $O_p(1/\sqrt{n})$.

Collecting together these contributions to H_n we get, for each fixed δ,

$$H_n(\delta) - \delta'\nu_n\mathbf{D}_n \to -\tfrac{1}{2}\delta'V\delta \qquad \text{in probability.}$$

The stochastic process on the left-hand side is concave in δ. A simple modification (see Section 6 of Pollard 1990, for example) of a standard result from convex analysis (Theorem 10.8 of Rockafellar 1970) shows that such convergence automatically holds in a stronger sense:

$$H_n(\delta) = \delta'\nu_n\mathbf{D}_n - \tfrac{1}{2}\delta'V\delta + o_p(1) \qquad \text{uniformly on compacta.} \tag{14.8}$$

The $o_p(1)$ term is a random function of δ and n whose supremum over bounded sets of δ converges in probability to zero.

Singularity of V slightly complicates the argument leading from (14.8) to an asymptotic expression for $\widehat{\delta}$. A reparametrization will solve the problem. Let J be an $s \times (s-1)$ matrix whose columns span $\mathcal{D}$. Then for θ ranging over $\mathbb{R}^{s-1}$,

$$H_n(J\theta) = \theta'J'\nu_n\mathbf{D}_n - \tfrac{1}{2}\theta'J'VJ\theta + o_p(1) \qquad \text{uniformly on compacta.}$$

The $(s-1)\times(s-1)$ matrix $J'VJ$ is nonsingular. A small concavity argument (as in Pollard 1990) shows that the $\widehat{\theta}$ that maximizes $H_n(J\theta)$ over $\mathbb{R}^{s-1}$ must lie close to the value that maximizes the quadratic approximation, that is,

$$\widehat{\theta} = (J'VJ)^{-1}J'\nu_n\mathbf{D}_n + o_p(1).$$

Hence

$$\widehat{\delta} = J(J'VJ)^{-1}J'\nu_n\mathbf{D}_n + o_p(1). \tag{14.9}$$

Let us denote by V^- the matrix multiplying $\nu_n\mathbf{D}_n$; it is a generalized inverse of V.

For each i, the functions D_{ni} converge uniformly to

$$D_i(\cdot) = \frac{\pi_i\lambda_i W_i(\cdot)}{G(\cdot)}.$$

This allows us to invoke a multivariate analogue of (14.6) to show that

$$\nu_n\mathbf{D}_n = \nu_n\mathbf{D} + o_p(1) \rightsquigarrow N\Big(\mathbf{0},\, P(G\mathbf{D}\mathbf{D}') - \sum_i \lambda_i(Q_i\mathbf{D})(Q_i\mathbf{D})'\Big). \tag{14.10}$$

It follows that $\widehat{\delta}$ also has an asymptotic normal distribution.

It is possible to solve (14.3) to get a similar asymptotic expression for $\widehat{\epsilon}$, and hence for $\widehat{\alpha}$. That would lead to an asymptotic normal distribution for $\sqrt{n}(\widehat{\alpha}-\alpha^*)$. Such

a calculation will be implicit in the next stage of the argument, which will apply the so-called delta method to (14.4) to derive a central limit theorem for $\widehat{P}_n$.

Asymptotic behavior of $\widehat{P}_n$. Yet another Taylor expansion gives an approximation that lets us capture the effect of $\widehat{\delta}$ on $\widehat{P}_n f$.

$$h_n(x, \alpha^* + \delta/\sqrt{n}) = \frac{1}{G_n(x)} - \frac{\delta' \mathbf{D}_n(x)}{\sqrt{n} G_n(x)} + \frac{|\delta|^2}{n G_n} R_n(x, \delta).$$

The remainder function R_n is uniformly bounded on compact sets of δ, in the sense that for each compact K there is a constant C_K such that

$$|R_n(x, \delta)| \le C_K \qquad \text{for all } x, \text{ all } n, \text{ all } \delta \text{ in } K.$$

If f is P-integrable, the contribution from the remainder term can be ignored because

$$\frac{|\delta|^2}{n} \mathbb{P}\widehat{Q}_n \left| \frac{f R_n}{G_n} \right| \le \frac{|\delta|^2}{n} P|f R_n| \le C_K \frac{|\delta|^2}{n} P|f|. \tag{14.11}$$

Since $|\widehat{\delta}| = O_p(1)$, the remainder terms will contribute only a $O_p(1/n)$ to $\widehat{P}_n f$.

From (14.5), the leading term in the Taylor expansion contributes

$$\widehat{Q}_n(f/G_n) = Pf + \frac{1}{\sqrt{n}} \nu_n(f/G_n), \tag{14.12}$$

which, by (14.6), is asymptotically normal if $P(f^2/G) < \infty$.

The linear term contributes

$$-\frac{1}{\sqrt{n}} \widehat{\delta}' \Big(P(f \mathbf{D}_n) + \frac{1}{\sqrt{n}} \nu_n(f \mathbf{D}_n / G_n) \Big).$$

The ν_n part can be absorbed into the $O_p(1/n)$ term if $P(f^2/G) < \infty$, because

$$\operatorname{var} \nu_n(f D_{nj}/G_n) \le \text{const} \sum_i \lambda_{ni} Q_i(f^2/G_n^2) < \text{const}\, P(f^2/G). \tag{14.13}$$

If $P(1/G) < \infty$, similar approximations are valid for the denominator in (14.4). Consequently, if both $P(1/G) < \infty$ and $P(f^2/G) < \infty$ (which also takes care of P-integrability of f),

$$\widehat{P}_n f = \frac{Pf + \Big(\nu_n(f/G_n) - \widehat{\delta}' P(f \mathbf{D}_n) \Big)/\sqrt{n} + o_p(1/\sqrt{n})}{1 + \Big(\nu_n(1/G_n) - \widehat{\delta}' P \mathbf{D}_n \Big)/\sqrt{n} + o_p(1/\sqrt{n})}.$$

The right-hand side simplifies to

$$Pf + \frac{1}{\sqrt{n}} \Big(\big(\nu_n(f/G_n) - Pf\, \nu_n(1/G_n) \big) - \widehat{\delta}' \big(P(f \mathbf{D}_n) - Pf\, P \mathbf{D}_n \big) \Big)$$

plus terms of order $o_p(1/\sqrt{n})$. The coefficient of the linear term in $\widehat{\delta}$ might be thought of as a covariance. Substituting from (14.9) for $\widehat{\delta}$, then consolidating the lower-order terms, we get

$$\sqrt{n}(\widehat{P}_n f - Pf) = \nu_n \Big(\operatorname{cov}_P(\mathbf{D}, f)' V^- \mathbf{D} + f/G_n - (Pf)/G_n \Big) + o_p(1). \tag{14.14}$$

The right-hand side has an asymptotic normal distribution, by virtue of the multivariate central limit theorem.

Uniformity in f. The preceding calculations are easily extended to provide a functional central limit theorem for $\widehat{\nu}_n = \sqrt{n}(\widehat{P}_n - P)$ treated as a stochastic process indexed by a class of functions $\mathcal{F}$.

Let us assume that $\mathcal{F}$ has an envelope $F(\cdot)$, that is, $|f| \leq F$ for each f in $\mathcal{F}$. If F is P-integrable, the analogue of (14.11), with f replaced by F, shows that the remainder terms are of order $O_p(1/n)$ uniformly over $\mathcal{F}$.

If both $P(1/G) < \infty$ and $P(F^2/G) < \infty$, and if the processes indexed by the classes of functions that appear in (14.13) and (14.14) are manageable in the sense of Section 7, then the maximal inequalities from that section can take over the role played by (14.13). (Here the stability results from Section 5 could be applied.) The random contribution to the linear term can again be absorbed into the $o_p(1/\sqrt{n})$, this time uniformly over $\mathcal{F}$. The $o_p(1)$ remainder in (14.14) then also applies uniformly over $\mathcal{F}$, which gives the desired uniform functional central limit theorem.

REMARKS. The concavity argument leading to the central limit theorem for $\widehat{\delta}$ is adapted from similar arguments for least absolute deviations regression estimators in Pollard (1990). Almost sure convergence of $\widehat{P}_n$ could be established by an even simpler concavity argument, based on pointwise application of a strong law of large numbers, somewhat in the style of Lemma 5.3 of Gill et al (1988). Concavity also explains the success of Vardi's (1985) algorithm—his procedure climbs a concave hill by successive maximizations along coordinate directions.