Efficient Estimation of Semiparametric Models by Smoothed Maximum Likelihood

Stephen R. Cosslett

Department of Economics

Ohio State University

December 2005

Abstract

A smoothed likelihood function is used to construct efficient estimators for a class of semiparametric models, in the case where the stochastic error terms and the regressors are independent. Smoothing the likelihood makes maximization with respect to the unknown density functions more tractable. The method is used to show the efficiency gains from knowledge of population shares in three cases: (1) binary choice; (2) binary choice when only one outcome is sampled, supplemented by random sampling of the explanatory variables; and (3) linear regression, where the shares are defined by a threshold value of the dependent variable. Semiparametric efficiency is achieved both for parametric components and for a class of functionals of the error density.

Keywords: semiparametric estimation, asymptotic efficiency, endogenous stratification, binary choice, contaminated sample.

JEL codes: C14, C24, C25

1. Introduction

This paper shows how a smoothed likelihood function can be used to construct efficient semiparametric estimators for a class of models containing unknown density functions together with parametric index functions, in the case where the stochastic error terms and the regressors are independent. The method is illustrated in two relatively simple cases, linear regression and binary choice, where efficient semiparametric estimators for the standard case are well known.

The main focus here is on the use of aggregate constraints to improve the efficiency of the estimators. For binary choice, the additional information consists of the population shares of the two choices, while for the linear model it consists of the population shares of two strata defined according to whether the dependent variable is above or below a given threshold value.

A second focus is on estimation of a binary choice model from a choice-based sample where only one outcome is sampled but there is a supplementary random sample of observations on the explanatory variables, both with and without knowledge of the population shares. (An example would be a consumer-response survey plus census data, where the census did not ask about the binary response; alternatively, it can be viewed as a contaminated choice-based sample, the random sample being the contaminated stratum.)

These models and sample designs were considered in early work on choice-based sampling, including Manski and McFadden (1981), Cosslett (1981), and McFadden (1979). The estimators derived in those papers are partly semiparametric in the sense that, while the distribution of the error terms is assumed to have a specified functional form, the distribution of the explanatory variables remains unspecified. The present paper provides one method of extending these to the fully semiparametric case, where both distributions are unspecified, while achieving the semiparametric efficiency bounds. The smoothed likelihood method can also be used to obtain

efficient semiparametric estimators for the tobit model (single-equation censored regression), for truncated regression, and for endogenously stratified regression model with two strata.[1]

The maximum likelihood approach leads to estimators of the unknown error distribution functions (for binary choice) or density functions (for the linear model), as well as the finite-dimensional parameters. Semiparametric efficiency for these functions is considered here in a limited sense, where we consider the asymptotic variances of a class of functionals of the distribution that converge at the parametric rate.

The basic idea here is to use kernel smoothing to make functional maximization of the likelihood more tractable, as opposed to its more usual application as a technique for direct estimation of a density or conditional density from the data. This is evidently somewhat counter-intuitive, since efficiency implies making best use of the available information, whereas the first step here is, in effect, to smear out the exact locations of the data points. Nevertheless, it results in a clear improvement over attempts at direct nonparametric maximization of the likelihood. While the maximum likelihood principle is used to find a suitable candidate for an efficient estimator, the estimators themselves are implemented by solving a set of score-like moment equations (rather than direct maximization of an objective function), and in that regard they are similar to the semiparametric maximum likelihood estimator of Ai (1997).

In some cases the smoothed likelihood approach just leads to a unified method of deriving estimators that are already known; in other cases, it shows how to construct efficient estimators when solution of the functional maximization problem is relatively complicated (requiring more than an application of Jensen's inequality) but still tractable; in yet other cases, where there is no explicit solution of the maximization problem, it remains to be seen whether any progress can be made by using only numerical solutions.

---

[1] The tobit estimator was obtained by a different (but closely related) method involving a smoothed self-consistency equation in Cosslett (2004). Results for truncated regression and for endogenously stratified regression will be presented elsewhere.

2. Smoothed Maximum Likelihood

2.1. Likelihood function. Consider a log likelihood function for observation i ($i = 1,...,n$) of the form

$$\log \ell(x_i, y_i, \theta, f) = \sum_j d_{j,i} \log p_j(v_j(x_i, y_i, \theta), f) + \log h(x_i). \tag{2.1}$$

The observed data consists of dependent variables y_i, explanatory variables x_i, and indicators $d_{j,i} \in \{0,1\}$ representing different strata. Each component of the vector $v_j(x, y, \theta)$ may consist of a residual of the form $u_{j,k}(x_i, y_i, \theta) = y_i - x_{i,k}\theta_k$ or an index function of the form $u_{j,k}(x_i, y_i, \theta) = x_{i,k}\theta_k$, where $x_{i,k}$ is a subvector of x_i and θ_k is a subvector of θ. The task is to estimate the unknown finite-dimensional parameter vector θ and one or more unknown density functions represented by *f*.

The smoothing technique consists of replacing each term of the form $\log p_j(v_j, f)$ by

$$\int du \frac{1}{h_n^{r_j}} K\left(\frac{v_j - u}{h_n}\right) \log p_j(u, f) \tag{2.2}$$

where r_j is the dimension of v_j, K is an r_j-dimensional kernel function, and the bandwidth h_n shrinks at a suitable rate as $n \to \infty$.

The smoothed log likelihood has the form

$$n^{-1} \log \widetilde{L}(\theta, f) = \sum_j \int du\ \widetilde{g}_j(u \mid \theta) \log p_j(u, f) \tag{2.3}$$

where

$$\widetilde{g}_j(u \mid \theta) = \frac{1}{n h_n^{r_j}} \sum_{i=1}^{n} d_{j,i} K\left(\frac{v_j(y_i, x_i, \theta) - u}{h_n}\right). \tag{2.4}$$

Note that in equation (2.3), all dependence on the data is contained in the functions $\widetilde{g}_j$, while all dependence on the unknown function *f* is contained in the functions p_j.

In the cases considered here, *f* will be the density function of a scalar error term, there will be at most two regimes *j*, and each component p_j of the likelihood will depend on only a single index or residual. This will allow a conventional univariate kernel to be used for smoothing, with $K(u) \geq 0$, $K(u) = K(-u)$, $\int du\, K(u) = 1$, $\int du\, u^2 K(u) < \infty$, and $\int du [K(u)]^2 < \infty$.

The unknown density $h(x)$, which has to be taken into account when there are aggregate constraints (and, more generally, in the case of endogenous stratification), typically has a high dimension that makes it unsuitable for kernel estimation. Instead, it is estimated (for given θ and *f*) by its empirical likelihood subject to the aggregate constraints, as in Cosslett (1981).

2.2. Variational equation. The next step is to maximize the smoothed log likelihood (2.3) with respect to *f* (or equivalently with respect to $\sqrt{f}$), subject to the normalization $\int du\, f(u) = 1$ and the condition $f(u) \geq 0$. In some simple examples this can be done directly, using Jensen's inequality. In general, the maximization problem involves solving a variational equation. Define the score operator $s_{f,j}(u, f)$ by

$$s_{f,j}(u, f)[\varphi] = 2 p_j^{-1/2}(u, f)\, D_f\, p_j^{1/2}(u, f)[\varphi] \tag{2.5}$$

where $D_f a(f)[\varphi]$ represents the functional (Hadamard) derivative of $a(f)$ with respect to $\sqrt{f}$ in the direction φ. Because *f* is a density function, φ is a function such that

$$\int dv \sqrt{f(v)}\, \varphi(v) = 0, \qquad \int dv\, \varphi(v)^2 < \infty. \tag{2.6}$$

The variational equation for *f* is

$$\sum\nolimits_j \int du\, \tilde{g}_j(u \mid \theta)\, s_{f,j}(u, f)[\varphi] = 0 \tag{2.7}$$

for all φ satisfying (2.6).

On the other hand, the transformation $(u, f) \mapsto p_j$ in typical semiparametric models should involve nothing worse than differentiable functions, integration, and possibly also differentiation. In that case, the functional derivative can be constructed by straightforward functional differentiation based on $\delta f(u) / \delta f(v) = \delta(u - v)$, which is a legitimate operation as long as *u* is an integration variable and the rest of the integrand is continuous. (A Lagrange multiplier will take care of the restriction $\int du\, f(u) = 1$.) This is of course the motivation for bringing in the smoothing operation (2.2) before attempting the functional maximization.

Let the solution of (2.7), when it exists, be $\tilde{f}(v \mid \theta) = f(v, \tilde{g}(\cdot \mid \theta))$, where $\tilde{g}(\cdot \mid \theta)$ is a vector with components $\tilde{g}_j(\cdot \mid \theta)$. The concentrated log likelihood is then

(2.8) $$n^{-1}\log\widetilde{L}(\theta)=\sum_j\int du\;\widetilde{g}_j(u\,|\,\theta)\log p_j(u,\widetilde{f}(\cdot\,|\,\theta))\,.$$

2.3. Score function. The semiparametric estimator $\hat{\theta}$ could in principle be defined as the maximum likelihood estimator corresponding to (2.8). In practice, however, it will be the solution of a suitably trimmed version of the corresponding score equation. (This is because, except in a few special cases, existing techniques for deriving asymptotic properties rely on convergence of a trimmed score function rather than on convergence of the log likelihood itself.) Before constructing the score function, we take a step back: substitute the expression (2.4) for $\widetilde{g}_j(u\,|\,\theta)$ in (2.8), and change the integration variable to $\eta=(v_j(y_i,x_i,\theta)-u)/h_n$, which gives

(2.9) $$\log\widetilde{L}(\theta)=\sum_{i=1}^{n}\sum_j\int d\eta\,K(\eta)\log p_j(v_j(y_i,x_i,\theta)-h_n\eta,\;\widetilde{f}(\cdot\,|\,\theta))\,.$$

The score function is $\widetilde{S}(\theta)=\partial\log\widetilde{L}(\theta)/\partial\theta$. However, the smoothing over η does not play any substantive role in the asymptotic properties of the score function and can be dropped without loss. We can therefore avoid an additional layer of computational complexity by instead working with the simplified score function

(2.10) $$\widetilde{S}(\theta)=\sum_{i=1}^{n}\sum_j\frac{d}{d\theta}\log p_j(v_j(y_i,x_i,\theta),\;\widetilde{f}(\cdot\,|\,\theta))\,.$$

The derivative has been written as $d/d\theta$ to emphasize that it operates on all occurrences of θ, both in the index function v and in the estimated density $\widetilde{f}$, which depends on θ indirectly through its dependence on the functions $\widetilde{g}(\cdot\,|\,\theta)$.

2.4. Trimming correction. As is well known, a complication of kernel-based estimators is the need for trimming. A typical term in the score function is $\partial\log\widetilde{g}/\partial\theta=\widetilde{g}^{-1}\partial\widetilde{g}/\partial\theta$. Because the denominator $\widetilde{g}$ is not bounded away from zero, it has to be trimmed in order to get the uniform convergence in y_i, x_i, and θ that is needed to derive asymptotic properties. Define a general trimming function t such that $t(u)=1$ for $u\geq 1$ and $t(u)=0$ for $u\leq 0$, with a smooth polynomial interpolation for $0\leq u\leq 1$ such that the second derivative is continuous (see, for example, Ai 1997). Then define the trimming function for $\widetilde{g}$ by

$$\tau(\tilde{g}) = t([\tilde{g} - b_n]/b_n)$$

where b_n is a shrinking lower bound on $\tilde{g}$, with the understanding that $g^{-1}\tau(g)$ is zero if $g = 0$. (The dependence of τ on b_n will not be shown explicitly.) A trimmed version of $\partial \log \tilde{g}/\partial\theta$ is then $\tau(\tilde{g})\partial \log \tilde{g}/\partial\theta$. More generally, if the score function contains several denominator terms, an overall trimming factor of the form $\tau(\tilde{g}_1 \tilde{g}_2)$ can be used to trim the common denominator.

The existing literature is somewhat unclear about how the argument for asymptotic efficiency of the semiparametric maximum likelihood estimator, as given by Newey (1994), survives this type of trimming. A key condition for semiparametric efficiency is

$$\mathrm{E}[D_f\, s(x, y, \theta_0, f(\cdot\,|\,\theta_0))] = 0 \tag{2.11}$$

where D_f denotes the functional derivative with respect to f (see Newey, 1994, in particular equations 3.12–3.16 and the associated discussion). In effect, this says that variations in the likelihood due to $\tilde{f}$ are asymptotically orthogonal to variations due to $\hat{\theta}$. Since $f(\varepsilon\,|\,\theta) = f(\varepsilon, g(\cdot\,|\,\theta))$ depends on $g(\cdot\,|\,\theta)$, and in general the components of $g(\cdot\,|\,\theta)$ can vary independently, we can replace (2.11) by the equivalent condition

$$\mathrm{E}[D_g\, s(x, y, \theta_0, f(\cdot\,|\,\theta_0))] = 0\,. \tag{2.12}$$

However, (2.12) does not necessarily hold for the trimmed score $s^*(x, y, \theta_0, f(\cdot|\theta_0))$, and to overcome the resulting asymptotic bias one may have to use higher-order (bias-reducing) kernels or a more complex trimming scheme.

Lemmas A1 and A2 (in the appendix) provides a method for correcting this asymptotic bias. They are applicable if the efficient score (or, more specifically, the part of it that requires trimming) satisfies $\mathrm{E}[s(\theta_0)\,|\,y_d, v(\theta_0)] = 0$ and $\mathrm{E}[s(\theta_0)\,|\,x] = 0$ (where y_d represents the discrete dependent variables, if any), while the corresponding trimmed score satisfies $\mathrm{E}[s^*(\theta_0)\,|\,y_d, v(\theta_0)] = 0$. These conditions hold for the models considered in this paper. The

corrected trimmed score function in the regression context, where $v(\theta_0) = \varepsilon$ (the error term in the regression), is

$$\tilde{s}^{**}(x, y, \theta, \tilde{f}(\cdot \mid \theta)) = \tilde{s}^{*}(x, y, \theta, \tilde{f}(\cdot \mid \theta)) - \tilde{s}^{c}(x, \theta, \tilde{f}(\cdot \mid \theta))$$

with the correction term

$$\tilde{s}^{c}(x, \theta, \tilde{f}(\cdot \mid \theta)) = \int dy\, \tilde{f}(y - x\theta \mid \theta)\, \tilde{s}^{*}(x, y, \theta, \tilde{f}(\cdot \mid \theta)). \tag{2.13}$$

In the case of discrete dependent variables, $\tilde{f}(y - x\theta \mid \theta)$ is replaced by the estimated discrete probability $\tilde{P}(y_d \mid v(\theta), \tilde{f}(\cdot \mid \theta))$ and the integral is replaced by a sum over y_d. In fact, (2.13) can be extended in a straightforward way to models with both discrete and continuous variables. The corrected trimmed score then satisfies the required orthogonality property, $\mathrm{E}[D_g\, s^{**}(x, y, \theta_0, f(\cdot \mid \theta_0))] = 0$.

2.5. Asymptotic properties. Why would we expect the solution of the score equations (or a suitably trimmed version of them) to result in efficient estimators? Under some standard regularity conditions, the kernel estimators $\tilde{g}_j(u \mid \theta)$ converge, uniformly in u and θ, to asymptotic limits $g_j(u \mid \theta)$ at a rate depending on the rate at which $h_n \to 0$, and similarly for derivatives of $\tilde{g}_j(u \mid \theta)$ (see appendix A.2 for further details). The limiting functions $g_j(u \mid \theta)$ involve the conditional expectations of the functions $p_j(u, f)$, and it follows that

$$n^{-1} \log \bar{L}(\theta, f) \equiv \sum_j \int du\, g_j(u \mid \theta) \log p_j(u, f) = \mathrm{E}[\log \ell(x, y, \theta, f)]. \tag{2.14}$$

The function $f = f(\varepsilon \mid \theta) = f(\varepsilon, g(\cdot \mid \theta))$ therefore maximizes $\mathrm{E}[\log \ell(x, y, \theta, f)]$. It follows (see Newey, 1994) that the score function

$$s(x, y, \theta, f(\cdot \mid \theta)) = \frac{\partial}{\partial \theta} \ell(x, y, \theta, f(\cdot \mid \theta))$$

is the (semiparametric) efficient score. This implies that $\bar{\theta}$, the artificial estimator obtained from the moment conditions $\sum_i s(x_i, y_i, \theta, f(\cdot \mid \theta)) = 0$ (artificial because $f(\cdot \mid \theta)$ is unknown), meets the semiparametric efficiency bound.

The essential step is then to show that uniform convergence of $\tilde{g}$ to g implies uniform convergence of a suitably trimmed version of the score function when expressed in terms of the solution of (2.7), i.e.,

$$s(x,y,\theta,f(\cdot,\tilde{g}(\cdot|\theta))) \to s(x,y,\theta,f(\cdot,g(\cdot|\theta))). \tag{2.15}$$

Specifically, if we denote the score functions corresponding to the left and right hand sides of (2.15) by $\tilde{S}(\theta)=\sum_i \tilde{s}_i(\theta)$ and $S(\theta)=\sum_i s_i(\theta)$, then a standard classical argument shows that $\hat{\theta}$ will have the same asymptotic distribution as $\bar{\theta}$ (and so will be asymptotically efficient) if (*i*) $n^{-1}\tilde{S}(\theta)$ and $n^{-1}\partial\tilde{S}(\theta)/\partial\theta$ converge to $n^{-1}S(\theta)$ and $n^{-1}\partial S(\theta)/\partial\theta$ in probability uniformly in θ and (*ii*) $n^{-1/2}\tilde{S}(\theta_0)$ converges in probability to $n^{-1/2}S(\theta_0)$.

It is difficult to find general conditions on the likelihood function such that these convergence conditions hold. Instead, we proceed by first deriving an explicit expression for the estimated score function and then verifying that it does indeed converge to the efficient score. The convergence rate of the trimmed score and its derivative can be found by methods developed in the literature on kernel-based semiparametric estimators by, among others, Ichimura and Lee (1991), Klein and Spady (1993), and Ai (1997). A summary is given in appendixA.2—under standard regularity conditions, there is a range of convergence rates for the window width h_n and the trimming parameter b_n that is sufficient for asymptotic efficiency of $\hat{\theta}$. These rates depend only on the convergence rates of $\tilde{g}$ and its derivatives, and are the same for each of the single-index models considered in this paper.

2.6. Estimated distribution function. Semiparametric maximum likelihood also delivers an estimator $\tilde{F}$ of the distribution function, and this may improve when there is additional information such as knowledge of aggregate shares. For binary choice models, we consider efficiency of $\tilde{F}$ in the limited sense of efficient estimation of the functional $\psi(a)=\int du\, a(u)F(u)$, for a suitable class of bounded integrable functions $a(u)$. Estimates of $\psi(a)$ converge at the parametric rate $n^{-1/2}$, so its asymptotic variance can be compared with a

conventional semiparametric efficiency bound. One can also estimate functionals of the form $\psi(a) = \int du\, a(u)\, f(u)$ in regression models, although this is probably less useful.

The estimator of $\psi(a)$ is $\hat{\psi}(a) = \int du\, a(u) \widetilde{F}(u \mid \hat{\theta})$, where now $\widetilde{F}(\cdot \mid \theta)$ is re-estimated using a faster-shrinking bandwidth h_n in order to control asymptotic bias. This is an easier strategy than maximizing $\log \widetilde{L}(\theta, F)$ with respect to F subject to $\int du\, a(u) F(u) = \psi$ and then maximizing the concentrated log likelihood with respect to the parameters θ and ψ. The derivation of the asymptotic variance is summarized in the appendix, and with some restrictions on $a(u)$ we find that $\hat{\psi}(a)$ achieves the efficiency bound.

3. Binary choice

This provides a relatively simple example to illustrate the basic approach, without having to solve a variational equation or to estimate the density of the explanatory variables. Not surprisingly, the resulting estimator in this case is essentially the same as the efficient semiparametric estimator of Klein and Spady (1993). The present approach allows the use of standard (as opposed to bias-reducing) kernels, although it is based on the solution of the score equations rather than maximization of a bona fide objective function (the estimated log likelihood in the Klein-Spady estimator). In the following, F is unrestricted, so there is no intercept in the index function $x\theta$ and there is an (unspecified) scale normalization for θ.

For the standard binary choice model (with random sampling and unknown aggregate shares) the log likelihood is

$$\log L(\theta, F) = \sum_{i=1}^{n} \{1(y_i = 1) \log F(x_i \theta) + 1(y_i = 0) \log[1 - F(x_i \theta)]\}$$

and the smoothed version is

$$n^{-1} \log \widetilde{L}(\theta, F) = \int du\, \{\widetilde{g}_1(u \mid \theta) \log F(u) + \widetilde{g}_0(u \mid \theta) \log[1 - F(u)]\} \tag{3.1}$$

where

$$\widetilde{g}_j(u \mid \theta) = \frac{1}{n h_n} \sum_{i=1}^{n} 1(y_i = j)\, K\left(\frac{x_i \theta - u}{h_n} \right).$$

By Jensen's inequality, (3.1) is maximized with respect to F at

$$\widetilde{F}(u \mid \theta) = \widetilde{g}_1(u \mid \theta) / \widetilde{g}(u \mid \theta) \tag{3.2}$$

where $\widetilde{g}(u \mid \theta) = \widetilde{g}_1(u \mid \theta) + \widetilde{g}_0(u \mid \theta)$, and the concentrated likelihood is

$$n^{-1}\widetilde{L}(\theta) = \int du \{\widetilde{g}_1(u \mid \theta) \log \widetilde{g}_1(u \mid \theta) + \widetilde{g}_0(u \mid \theta) \log \widetilde{g}_0(u \mid \theta) - \widetilde{g}(u \mid \theta) \log \widetilde{g}(u \mid \theta)\}.$$

The score function corresponding to (2.10) (after dropping the "outer" smoothing, as discussed in Section 2.3) is then

$$\widetilde{S}(\theta) = \sum_{i=1}^{n} \widetilde{s}_i(\theta) = \sum_{i=1}^{n} \left\{ 1(y_i = 1) \frac{d}{d\theta} \log \frac{\widetilde{g}_1(x_i\theta \mid \theta)}{\widetilde{g}(x_i\theta \mid \theta)} + 1(y_i = 0) \frac{d}{d\theta} \log \frac{\widetilde{g}_0(x_i\theta \mid \theta)}{\widetilde{g}(x_i\theta \mid \theta)} \right\} \tag{3.3}$$

which is the same as the score for the Klein and Spady (1993) estimator.

One way of ensuring that $\mathrm{E}[D_g \, s^*(\theta_0)] = 0$ is to multiply all terms in $\widetilde{s}_i(\theta)$ by a common trimming factor $\widetilde{\tau}_i(\theta)$:

$$\widetilde{s}_i^*(\theta) = 1(y_i = 1) \widetilde{\tau}_i(\theta) \frac{d}{d\theta} \log \frac{\widetilde{g}_1(x_i\theta|\theta)}{\widetilde{g}(x_i\theta|\theta)} + 1(y_i = 0) \widetilde{\tau}_i(\theta) \frac{d}{d\theta} \log \frac{\widetilde{g}_0(x_i\theta|\theta)}{\widetilde{g}(x_i\theta|\theta)}. \tag{3.4}$$

A suitable term is $\widetilde{\tau}_i(\theta) = \tau[\widetilde{g}_1(x_i\theta|\theta)] \cdot \tau[\widetilde{g}_0(x_i\theta|\theta)]$. (There is no need to trim $\widetilde{g}$ separately because $\widetilde{g} \geq \max\{\widetilde{g}_1, \widetilde{g}_2\}$, with non-negative kernels.) The asymptotic limit of (3.4) at θ_0 is

$$s_i^*(\theta_0) = -\big(x_i - \mathrm{E}[X \mid X\theta_0 = x_i\theta_0]\big)\tau_i(\theta_0) \left\{ 1(y_i = 1) \frac{f(x_i\theta_0)}{F(x_i\theta_0)} - 1(y_i = 0) \frac{f(x_i\theta_0)}{1 - F(x_i\theta_0)} \right\} \tag{3.5}$$

where $\tau_i(\theta_0) = \tau[h_*(x_i\theta_0) F(x_i\theta_0)] \cdot \tau[h_*(x_i\theta_0)(1 - F(x_i\theta_0))]$ and $h_*(\cdot)$ is the marginal density of $x\theta_0$. Then $\mathrm{E}[s^*(\theta_0) \mid y, x\theta_0] = 0$ and $\mathrm{E}[s(\theta_0) \mid x] = 0$, so Lemma A2 is applicable. Substituting (3.4) in equation (A.3) gives $\widetilde{s}^c(\theta) = 0$, and therefore no trimming correction is needed when a common trimming factor is used.

Finally, $\hat{\theta}$ is computed by solving the trimmed score equation $\widetilde{S}^*(\theta) = 0$. Asymptotic efficiency can be demonstrated by the method discussed in Section 2.5, and the rate-of-convergence calculations are summarized in appendix A.2. We note that efficiency of $\hat{\theta}$ can be

achieved using a standard kernel function if the bandwidth shrinks at a rate between $n^{-1/5}$ and $n^{-1/8}$. The asymptotic variance of $\hat{\theta}$ is then given by the efficiency bound

$$V_\theta^{-1} = \int du\, h_*(u) \frac{f(u)^2}{F(u)[1-F(u)]} \operatorname{var}[X \mid X\theta_0 = u]. \tag{3.6}$$

To estimate the functional $\psi(a) = \int du\, a(u) F(u)$, we have to trim the denominator in $\widetilde{F}$:

$$\hat{\psi}(a) = \int du\, a(u) \tau[\widetilde{g}(u \mid \hat{\theta})] \widetilde{F}(u \mid \hat{\theta}). \tag{3.7}$$

The kernel bandwidth used to compute (3.6) has to shrink at a faster rate, between $n^{-1/2}$ and $n^{-1/4}$, to avoid asymptotic bias in $\sqrt{n}(\hat{\psi} - \psi)$. There is no practical way (within the present approach) to avoid trimming bias in the tails of the integrand, so we require $a(u)$ to have bounded support and that $h_*(u) > 0$. The derivation of the asymptotic variance is summarized in the appendix. It is again equal to the semiparametric efficiency bound, which in this case is

$$V_\psi = V_1 + V_2' V_\theta V_2 \tag{3.8}$$

where

$$V_1 = \int du\, \frac{a(u)^2}{h_*(u)} F(u)[1-F(u)] \tag{3.9}$$

$$V_2 = \int du\, a(u)\, f(u) \mathrm{E}[x \mid x\theta_0 = u] \tag{3.10}$$

provided the integral in (3.9) exists.

4. Binary choice with known shares

Let the population shares of the two outcomes be Q_0 and Q_1, and let the sample shares be $H_0 = n_0 / n$ and $H_1 = n_1 / n$. The new information is given by the constraint equation

$$\int dx\, h(x) F(x\theta) = Q_1 \tag{4.1}$$

where Q_1, the population share of outcome 1, is known. This shows that the unknown density $h(x)$ of the explanatory variables can not be ignored in maximizing the likelihood

$$\log L(\theta, F, h) = \sum_{i=1}^{n} \{1(y_i = 1) \log F(x_i\theta) + 1(y_i = 0) \log[1 - F(x_i\theta)] + \log h(x_i)\}.$$

As in the case of endogenously stratified sampling (Cosslett 1981), construct the nonparametric maximum likelihood estimator of $h(x)$ subject to (4.1) for given F and θ. This assigns a mass point w_i to each observation,

$$w_i = w_i(\theta, F) = \frac{1}{n} \frac{1}{1+\lambda[Q_1 - F(x_i\theta)]}$$

where the Lagrange multiplier $\lambda = \lambda(\theta, F)$ is determined by the aggregate share constraint,

$$\frac{1}{n}\sum_{i=1}^{n} \frac{F(x_i\theta)}{1+\lambda[Q_1 - F(x_i\theta)]} = Q_1 . \tag{4.2}$$

(This equation has a spurious solution at $\lambda = -1/Q_1$, but this will not be a problem because $\lambda \to 1$ in the asymptotic limit.) Substituting back in $L(\theta, F, h)$ gives the partially concentrated log likelihood

$$\log L(\theta, F) = \sum_{i=1}^{n} \left\{ 1(d_i = 1) \log F(x_i\theta) + 1(d_i = 0) \log[1 - F(x_i\theta)] - \log(1 + \lambda[Q_1 - F(x_i\theta)]) \right\}$$

(apart from constants). Smoothing over $x_i\theta$ then gives

$$n^{-1} \log \tilde{L}(\theta, F) = \int du \left\{ \tilde{g}_1(u \mid \theta) \log F(u) + \tilde{g}_0(u \mid \theta) \log[1 - F(u)] - \tilde{g}(u \mid \theta) \log(1 + \tilde{\lambda}[Q_1 - F(u)]) \right\}$$

with $\tilde{\lambda} = \tilde{\lambda}(\theta, F)$ determined by the smoothed version of (4.2),

$$\int du\, \tilde{g}(u|\theta) \frac{F(u)}{1+\tilde{\lambda}[Q_1 - F(u)]} = Q_1 , \tag{4.3}$$

and $\tilde{g}_1(u|\theta)$, $\tilde{g}_0(u|\theta)$ and $\tilde{g}(u|\theta)$ defined as in the previous section.

Let $\tilde{F}(u|\theta)$ be the distribution function that maximizes $\log \tilde{L}(\theta, F)$. The variational equation (first-order condition for F) gives

$$\frac{\tilde{g}_1(u|\theta)}{\tilde{F}(u|\theta)} - \frac{\tilde{g}_0(u|\theta)}{1-\tilde{F}(u|\theta)} + \tilde{\lambda} \frac{\tilde{g}(u|\theta)}{1+\tilde{\lambda}[Q_1 - \tilde{F}(u|\theta)]} = 0 . \tag{4.4}$$

Note that (4.3) is the implicit definition of $\tilde{\lambda}(\theta, F)$, not a constraint equation, and that $\partial \log \tilde{L} / \partial \tilde{\lambda} = 0$. From (4.4), the solution is

$$\widetilde{F}(u|\theta) = \frac{(1+\widetilde{\lambda}_1 Q_1)\,\widetilde{g}_1(u|\theta)}{(1+\widetilde{\lambda}_1 Q_1)\,\widetilde{g}_1(u|\theta) + (1-\widetilde{\lambda}_1 Q_0)\,\widetilde{g}_0(u|\theta)},$$

although it is not guaranteed to be a proper distribution function. Substituting $F(u) = \widetilde{F}(u\,|\,\theta)$ in (4.3) gives

$$\int du\, \widetilde{g}_1(u|\theta) / [1 - \widetilde{\lambda} Q_0] = Q_1.$$

But $\int du\, \widetilde{g}_1(u|\theta) = H_1$, and therefore $\widetilde{\lambda}(\theta, \widetilde{F}(\cdot|\,\theta))$ does not depend on θ:

$$\widetilde{\lambda} = (H_0 / Q_0) - (H_1 / Q_1).$$

Note that $H_1 \to Q_1$ and $H_0 \to Q_0$ (under random sampling) as $n \to \infty$ and therefore $\widetilde{\lambda} \to 0$. The solution of the maximization problem can now be written as

$$\widetilde{F}(u|\theta) = \frac{Q_1}{H_1}\widetilde{g}_1(u|\theta)\left(\frac{Q_1}{H_1}\widetilde{g}_1(u|\theta) + \frac{Q_0}{H_0}\widetilde{g}_0(u|\theta)\right)^{-1}. \tag{4.5}$$

Substituting for F and $\widetilde{\lambda}$ in $\log \widetilde{L}(\theta, F)$, and using the identities $\int du\, \widetilde{g}_1(u|\theta) = H_1$ and $\int du\, \widetilde{g}_0(u|\theta) = H_0$ to eliminate constant terms, the concentrated log likelihood can be written as

$$n^{-1} \log \widetilde{L}(\theta) = \int du \left\{ \widetilde{g}_1(u|\theta) \log \frac{\widetilde{g}_1(u|\theta)}{\widetilde{g}(u|\theta)} + \widetilde{g}_0(u|\theta) \log \frac{\widetilde{g}_0(u|\theta)}{\widetilde{g}(u|\theta)} \right\}.$$

This is the same as in the case of unknown aggregate shares, so knowledge of the aggregate shares does not affect estimation of θ (which is not surprising because the semiparametric efficiency bounds are the same in the two cases).

The additional information is, however, responsible for the improved estimator (4.5) of the distribution function. Let $\widetilde{g}_d(u\,|\,\theta)$ denote the denominator term in (4.5). Then the analog of (3.7) is

$$\hat{\psi}(a) = \int du\, a(u)\, \tau[\widetilde{g}_d(u\,|\,\hat{\theta})]\, \widetilde{F}(u\,|\,\hat{\theta}) \tag{4.6}$$

with $\widetilde{F}(u\,|\,\theta)$ now given by (4.5). The asymptotic variance again has the form $V_\psi = V_1 + V_2' V_\theta V_2$, with V_2 given by (3.10) and V_θ by (3.6), but with an extra term in V_1,

(4.7) $$V_1 = \int du\, \frac{a(u)^2}{h_*(u)}\, F(u)[1-F(u)] - \frac{1}{Q_1 Q_0}\left(\int du\, a(u)\, F(u)[1-F(u)]\right)^2$$

where the second integral represents the improvement due to knowledge of Q.

5. Binary choice with a "contaminated" choice-based sample

In this case the data consists of two samples of x, one drawn from the stratum with $y = 1$ and the other from the whole population. In the context of endogenous stratification, this can be viewed as a truncated sample, where only one of the strata is sampled and a supplementary random sample of x is then needed to identify the model. Alternatively, it can be viewed as a contaminated choice-based sample, where the random sample is "contaminated" by cases with $y = 1$ instead of being drawn entirely from the stratum with $y = 0$. There do not seem to be any previous results on fully semiparametric estimation of this model, where $F(\varepsilon)$ is unknown as well as $h(x)$.

Define stratum indicators by $s = 1$ for the n_1 choice-based observations with $y = 1$, and $s = 2$ for the n_2 observations from the a random sample on x. Let $H_1 = n_1 / n$ and $H_2 = n_2 / n$, where $n = n_1 + n_2$. Note that the likelihood for an observation with $s = 1$ is $\Pr\{x \mid y = 1\}$. The overall log likelihood is then

(5.1) $$n^{-1} \log L(\theta, F, h) = \frac{1}{n} \sum_{i=1}^{n} \left\{ 1(s_i = 1) \log[F(x_i\theta) / Q_1] + \log h(x_i) \right\}$$

where, as before,

(5.2) $$Q_1 = \int dx\, h(x)\, F(x\theta).$$

If Q_1 is unknown, then (5.2) defines Q_1 in (5.1) as a function of θ and F. If Q_1 is known, then (5.2) is a constraint to be taken into account when maximizing the likelihood. In either case we have to deal with the unknown function $h(x)$.

5.1. Unknown shares. The first step is nonparametric maximum likelihood estimation of $h(x)$ for given θ and F. This is the same as in the case of a parametric distribution function F (Cosslett, 1981). The discrete weights are

$$w_i = \frac{1}{n}\frac{1}{(H_1/q_1)F(x_i\theta)+H_2}$$

where q_1 is defined as the solution of

$$\frac{1}{n}\sum_{i=1}^{n}\frac{1}{(H_1/q_1)F(x_i\theta)+H_2}=1. \tag{5.3}$$

The partially concentrated likelihood is

$$\log L(\theta, f)=\sum_{i=1}^{n}\left\{1(s_i=1)\log F(x_i\theta)-\log[(H_1/q_1)F(x_i\theta)+H_2]\right\}-H_1\log q_1.$$

The smoothed version is

$$n^{-1}\log\tilde{L}(\theta, f)=\int du\left\{\tilde{g}_1(u\,|\,\theta)\log F(u)-\tilde{g}(u\,|\,\theta)\log[(H_1/\tilde{q}_1)F(u)+H_2]\right\}-H_1\log\tilde{q}_1$$

where

$$\tilde{g}_s(u\,|\,\theta)=\frac{1}{nh_n}\sum_{i=1}^{n}1(s_i=s)K\left(\frac{x_i\theta-u}{h_n}\right),\qquad s=1,2 \tag{5.4}$$
$$\tilde{g}(u\,|\,\theta)=\tilde{g}_1(u\,|\,\theta)+\tilde{g}_2(u\,|\,\theta)$$

and $\tilde{q}_1$ is the solution of the smoothed version of (5.3),

$$\int du\,\tilde{g}(u\,|\,\theta)\frac{1}{(H_1/\tilde{q}_1)F(u)+H_2}=1. \tag{5.5}$$

The variational equation for maximization of $\tilde{L}(\theta, F)$ with respect to F is

$$\frac{\tilde{g}_1(u|\theta)}{\tilde{F}(u|\theta)}-\frac{H_1}{\tilde{q}_1}\frac{\tilde{g}(u|\theta)}{(H_1/\tilde{q}_1)\tilde{F}(u|\theta)+H_2}=0$$

(the derivative of $\tilde{L}(\theta, F)$ with respect to $\tilde{q}_1$ is zero, so the functional derivative of $\tilde{q}_1$ with respect to F is not needed), and therefore

$$\tilde{F}(u\,|\,\theta)=\tilde{q}_1\frac{H_2}{H_1}\frac{\tilde{g}_1(u\,|\,\theta)}{\tilde{g}_2(u\,|\,\theta)}. \tag{5.6}$$

Substituting into $\tilde{L}(\theta, F)$ finally gives the concentrated smoothed log likelihood,

$$n^{-1}\log\tilde{L}(\theta)=\int du\left\{\tilde{g}_1(u\,|\,\theta)\log\frac{\tilde{g}_1(u\,|\,\theta)}{\tilde{g}(u\,|\,\theta)}+\tilde{g}_2(u\,|\,\theta)\log\frac{\tilde{g}_2(u\,|\,\theta)}{\tilde{g}(u\,|\,\theta)}\right\}. \tag{5.7}$$

This has the same form as binary choice between the two strata—we can just assign $y = 0$ to all cases in the supplementary (or contaminated) sample and proceed as if the pooled data had been randomly sampled.

The same trimmed score can be used as in Section 3—Lemma A2 is still valid if $\Pr\{y \mid x\theta_0\}$ is replaced by $\Pr\{s \mid x\theta_0\}$, the probability that a randomly drawn observation from the pooled sample is in stratum s conditional on the value of $x\theta_0$. Although the score has the same functional form as for conventional binary choice, its asymptotic limit is different:

$$s(\theta_0) = -(x - \mathrm{E}[X \mid x\theta_0])\ f(x\theta_0) \left\{ 1(s=1) \frac{1}{F(x\theta_0)} - \frac{H_1}{Q_1} \left(\frac{H_1}{Q_1} F(x\theta_0) + H_2 \right)^{-1} \right\}$$

is the same as the efficient score, and the asymptotic variance for $\hat{\theta}$ is equal to the semiparametric efficiency bound

$$V_\theta^{-1} = \frac{H_1 H_2}{Q_1} \int du\, h_0(u) \frac{f(u)^2}{F(u)\,[(H_1/Q_1)F(u) + H_2]} \mathrm{var}[X \mid X\theta_0 = u]. \tag{5.8}$$

Substituting $F(u) = \widetilde{F}(u \mid \theta)$ in (5.5) results in an identity, so $\widetilde{q}_1$ is not identified, and (5.6) determines $\widetilde{F}(u \mid \theta)$ only up to a constant factor. One could attempt "identification at infinity" by setting $\widetilde{F}(x\hat{\theta} \mid \hat{\theta}) = 1$ at the largest observed value of $x\hat{\theta}$, but the information bound for $\psi(a)$ is zero and thus there is no $\sqrt{n}$-consistent estimator of $\psi(a)$.

5.2. Known shares. The first step, nonparametric estimation of $h(x)$, is the same as for standard binary choice with known shares, in Section 4. The partially concentrated smoothed log likelihood is

$$n^{-1} \log \widetilde{L}(\theta, F) = \int du \left\{ \widetilde{g}_1(u \mid \theta) \log F(u) - \widetilde{g}(u \mid \theta) \log\left(1 + \widetilde{\lambda}[Q_1 - F(u)]\right) \right\} \tag{5.9}$$

where $\widetilde{g}_1$ and $\widetilde{g}$ are defined by (5.4) and $\widetilde{\lambda} = \widetilde{\lambda}(\theta, F)$ is determined as the solution of (4.3). The solution of the variational problem for maximizing $\widetilde{L}(\theta, F)$ with respect to F is

$$\widetilde{F}(u \mid \theta) = -[(1 + \widetilde{\lambda} Q_1)/\widetilde{\lambda}] \cdot \widetilde{g}_1(u \mid \theta) / \widetilde{g}_2(u \mid \theta). \tag{5.10}$$

Substituting for F in (4.3) gives

$$\int du\, \tilde{g}_2(u|\theta)/[1+\tilde{\lambda}Q_1] = H_2/[1+\tilde{\lambda}Q_1] = 1$$

and therefore $\tilde{\lambda} = \lambda(\theta, \tilde{F}) = -H_1/Q_1$.

The resulting expressions for $\tilde{F}(u \mid \theta)$ and $\tilde{L}(\theta)$ are the same as for the case of unknown shares, (5.6) and (5.7), except that $\tilde{q}_1$ is replaced by Q_1. The estimator of θ is unchanged, which again is not surprising because the semiparametric efficiency bound is also the same. On the other hand, $\tilde{F}(u \mid \theta)$ is now properly identified because Q_1 is known.

The functional $\psi(a) = \int du\, a(u) F(u)$ can be estimated by the analog of (3.7), which is

$$\hat{\psi}(a) = \int du\, a(u)\, \tau[\tilde{g}_2(u \mid \hat{\theta})]\, \tilde{F}(u \mid \hat{\theta}). \tag{5.11}$$

The asymptotic variance again has the form $V_\psi = V_1 + V_2' V_\theta V_2$, with V_2 given by (3.10), but now V_θ is given by (5.8) and there is a new expression for V_1,

$$V_1 = \int du\, \frac{a(u)^2}{h_*(u)} F(u)\left(\frac{1}{H_2} F(u) + \frac{Q_1}{H_1} \right) - \frac{1}{H_1 H_2}\left(\int du\, a(u) F(u) \right)^2. \tag{5.12}$$

6. Linear regression

The adaptive estimator for linear regression (Bickel, 1982) is by now a classic example of efficient semiparametric estimation. All the same, linear regression provides a nice illustration of smoothed maximum likelihood and trimming bias correction, before considering the more complicated case of known population shares. In the following, f is unrestricted, so there is no intercept in the regression. The smoothed log likelihood is

$$n^{-1} \log \tilde{L}(\theta, f) = \int du\, \tilde{g}(u \mid \theta) \log f(u) \tag{6.1}$$

with

$$\tilde{g}(u \mid \theta) = \frac{1}{n h_n} \sum_{i=1}^{n} K\left(\frac{y_i - x_i\theta - u}{h_n} \right). \tag{6.2}$$

By Jensen's inequality, (6.1) is maximized at $f = \tilde{f}(u \mid \theta) = \tilde{g}(u \mid \theta)$. The concentrated log likelihood function is

$$n^{-1} \log \tilde{L}(\theta) = \int du\, \tilde{g}(u \mid \theta) \log \tilde{g}(u \mid \theta)$$

and the trimmed score function corresponding to (2.10) is

$$\widetilde{S}^*(\theta) = \sum_{i=1}^{n} \widetilde{s}_i^*(\theta) = \sum_{i=1}^{n} \tau[\widetilde{g}(y_i - x_i\theta \mid \theta)] \frac{d}{d\theta} \log \widetilde{g}(y_i - x_i\theta \mid \theta) . \tag{6.3}$$

This converges at $\theta = \theta_0$ to

$$s_i^*(\theta_0) = -(x_i - \mathrm{E}[X])\tau[f(\varepsilon_i)]\, f'(\varepsilon_i) / f(\varepsilon_i)$$

(where $\varepsilon_i = y_i - x_i\theta_0$), which is a trimmed version of the efficient score. The conditions for Lemma A1 hold, and the trimming bias correction is given by (A.1). The final expression for the trimmed, bias-corrected score is[2]

$$\widetilde{s}_i^*(\theta) = \tau[\widetilde{g}(y_i - x_i\theta \mid \theta)] \frac{d}{d\theta} \log \widetilde{g}(y_i - x_i\theta \mid \theta) - \int dy\, \tau[\widetilde{g}(y - x_i\theta)] \frac{d}{d\theta} \widetilde{g}(y - x_i\theta \mid \theta) . \tag{6.4}$$

Then $\hat{\theta}$ is the solution of $\widetilde{S}^*(\theta) = 0$ (or if there are multiple solutions, the one closest to the least squares estimator).

As in the case of binary choice, asymptotic efficiency can be demonstrated by the method discussed in Section 2.5, and the rate-of-convergence calculations summarized in the appendix also apply here, with h_n shrinking at a rate between $n^{-1/5}$ and $n^{-1/8}$. Convergence of the second term in (6.4) is discussed in appendix A.2, and can be achieved without trimming of the range of integration. The asymptotic variance of $\hat{\theta}$ is equal to the usual semiparametric efficiency bound, $V_\theta^{-1} = \mathrm{var}[X] \int d\varepsilon\, f'(\varepsilon)^2 / f(\varepsilon)$.

7. Linear regression with known stratum shares

Without loss, let the two strata be $\{y < 0\}$ and $\{y \geq 0\}$. The known shares are Q_0 and $Q_1 = 1 - Q_0$ where

$$Q_0 = \Pr\{y_i < 0\} = \int dx\, h(x) F(-x\theta) .$$

[2] The integral does increase the computational complexity—it has to evaluated numerically over the range where $0 < \tau[\widetilde{g}(y - x\theta)] < 1$, although this should be small and not require great precision.

As in the other cases of known shares (Sections 4 and 5.2), we start by nonparametric maximum likelihood estimation of the density $h(x)$. The only difference is that, according to the usual conventions, the labeling of the strata is reversed and the sign of the argument of F is changed. The resulting smoothed log likelihood is given by

$$n^{-1}\log \widetilde{L}(\theta, f) = \int du \left\{ \widetilde{g}_1(u \mid \theta) \log f(u) - \widetilde{g}_2(u \mid \theta) \log \overline{F}(-u, \lambda) \right\} \tag{7.1}$$

where

$$\widetilde{g}_1(u \mid \theta) = \frac{1}{nh_n} \sum_{i=1}^{n} K\left(\frac{y_i - x_i\theta - u}{h_n} \right), \qquad \widetilde{g}_2(u \mid \theta) = \frac{1}{nh_n} \sum_{i=1}^{n} K\left(\frac{x_i\theta - u}{h_n} \right),$$

$$\overline{F}(u, \lambda) \equiv 1 + \lambda[Q_0 - F(u)],$$

and $\lambda = \lambda(\theta, f)$ is the solution of

$$\int du\, \widetilde{g}_2(u|\theta) \frac{F(-u)}{\overline{F}(-u, \lambda)} = Q_0 . \tag{7.2}$$

Optimization. Maximize $\widetilde{L}(\theta, f)$ subject to $\int du\, f(u) = 1$. Let $f = \widetilde{f}(u \mid \theta)$ be the solution, with distribution function $\widetilde{F}(u|\theta)$, and let $\widetilde{\overline{F}}(u, \lambda|\theta)$ be the corresponding solution for $\overline{F}(u, \lambda)$. The first-order condition is

$$\frac{\widetilde{g}_1(v|\theta)}{\widetilde{f}(v|\theta)} + \int_{-\infty}^{-v} du\, \widetilde{g}_2(u|\theta) \frac{\lambda}{\widetilde{\overline{F}}(-u, \lambda|\theta)} + \mu = 0 \tag{7.3}$$

where μ is a Lagrange multiplier. This can be rewritten as

$$\widetilde{g}_1(v|\theta) - \widetilde{g}_2(-v|\theta) - \frac{d}{dv}\left\{ \widetilde{\overline{F}}(v, \lambda|\theta) \int_{-\infty}^{-v} du\, \widetilde{g}_2(u|\theta) \frac{1}{\widetilde{\overline{F}}(-u, \lambda|\theta)} \right\} + \mu\, \widetilde{f}(v|\theta) = 0 .$$

Integrate from $-\infty$ to v:

$$\widetilde{G}(v|\theta) - \frac{1}{\lambda}(1 + \lambda Q_0)(1 - \lambda Q_1) - \widetilde{\overline{F}}(v, \lambda|\theta) \left\{ \int_{-\infty}^{-v} du\, \widetilde{g}_2(u|\theta) \frac{1}{\widetilde{\overline{F}}(-u, \lambda|\theta)} - \frac{1}{\lambda}(1 + \lambda Q_0) \right\} = 0$$

where we define

$$\widetilde{G}(v|\theta) = \int_{-\infty}^{v} du \left\{ \widetilde{g}_1(u|\theta) - \widetilde{g}_2(-u|\theta) \right\} \tag{7.4}$$

and use the boundary condition $\widetilde{G}(\infty|\theta)=0$ to eliminate the Lagrange multiplier μ. Finally, after dividing by $\widetilde{\overline{F}}(v,\lambda)$ and differentiating with respect to *v*, the first-order condition becomes a linear differential equation

$$\frac{d\widetilde{\overline{F}}(v,\lambda|\theta)}{dv}\left\{\widetilde{G}(v|\theta)-\lambda^{-1}(1+\lambda Q_0)(1-\lambda Q_1)\right\}-\widetilde{\overline{F}}(v,\lambda|\theta)\,\widetilde{g}_1(v|\theta)=0\,.$$

Taking into account the boundary condition $\overline{F}(-\infty,\lambda)=1+\lambda Q_0$, and defining

$$\widetilde{m}_1(v|\theta,\lambda)=\frac{\widetilde{g}_1(v|\theta)}{(1+\lambda Q_0)(1-\lambda Q_1)-\lambda\widetilde{G}(v|\theta)}, \tag{7.5}$$

the solution is

$$\widetilde{\overline{F}}(u,\lambda|\theta)=(1+\lambda Q_0)\exp\left\{-\lambda\int_{-\infty}^{u}dv\,\widetilde{m}_1(v|\theta,\lambda)\right\}. \tag{7.6}$$

The upper boundary condition $\overline{F}(\infty,\lambda)=1-\lambda Q_1$ then gives the restriction

$$\lambda\int dv\,\widetilde{m}_1(v|\theta,\lambda)-\log\frac{1+\lambda Q_0}{1-\lambda Q_1}=0\,. \tag{7.7}$$

Given (7.6), the restrictions (7.2) and (7.7) are equivalent.

The resulting concentrated log likelihood is

$$\log\widetilde{L}(\theta)=\sum_{i=1}^{n}\log\widetilde{m}_1(y_i-x_i\theta\,|\,\theta,\lambda)-\sum_{i=1}^{n}\lambda\int_0^{y_i}dv\,\widetilde{m}_1(v-x_i\theta\,|\,\theta,\lambda) \tag{7.8}$$

with the understanding that $\int_0^y dv$ is to be interpreted as $-\int_y^0 dv$ when $y<0$.

Score function. Note that the Lagrange multiplier λ is not a free parameter, but is determined implicitly as a function of θ by the subsidiary condition (7.7). Since $\lambda\to 0$ in the asymptotic limit, we can simplify the restriction. The expansion of (7.7) in powers of λ has the form $\widetilde{r}_1(\theta)\lambda^2+\widetilde{r}_2(\theta)\lambda^3+\cdots$, and without loss we can use the asymptotic approximation[3]

$$d\lambda/d\theta=-\widetilde{r}_1'(\theta)/\widetilde{r}_2(\theta)\,. \tag{7.9}$$

---

[3] From (7.13) below, $\mathrm{E}[s_\lambda(\theta_0)]=0$; otherwise, we would need to expand $d\lambda/d\theta$ to first order in λ in order to get the correct expression for the asymptotic Hessian matrix $dS(\theta_0)/d\theta$.

The score function is

$$\widetilde{S}(\theta)=\sum_{i=1}^{n}\left\{\widetilde{s}_{i,\theta}(\theta)-[\widetilde{r}_1'(\theta)/\widetilde{r}_2(\theta)]\,\widetilde{s}_{i,\lambda}(\theta)\right\} \tag{7.10}$$

where

$$\widetilde{s}_{i,\theta}(\theta)=\frac{d}{d\theta}\left(\log\widetilde{m}_1(y_i-x_i\theta\,|\,\theta,\lambda)\right)-\lambda\int_0^{y_i}dv\,\frac{d}{d\theta}\widetilde{m}_1(v-x_i\theta\,|\,\theta,\lambda)$$

$$\widetilde{s}_{i,\lambda}(\theta)=\frac{\partial}{\partial\lambda}\left(\log\widetilde{m}_1(y_i-x_i\theta\,|\,\theta,\lambda)\right)-\int_0^{y_i}dv\,\frac{\partial}{\partial\lambda}\left(\lambda\,\widetilde{m}_1(v-x_i\theta\,|\,\theta,\lambda)\right).$$

Trimming. Because $|\widetilde{G}(u|\theta)|\le 1$, the denominator term $(1+\lambda Q_0)(1-\lambda Q_1)-\lambda\widetilde{G}(u|\theta)$ is bounded away from zero, provided that $|\lambda|<\frac{1}{2}$. (That is not a substantive restriction because the known asymptotic limit of λ is 0.) The term $\widetilde{r}_2(\theta)$ is not necessarily positive in finite samples, but it converges in probability to a strictly positive limit. It follows that the only term in the score function that needs to be trimmed (in order to derive asymptotic results) is the same as the score for the unrestricted linear model:

$$\widetilde{s}_U(\theta)\equiv\frac{d}{d\theta}\log\widetilde{g}_1(y_i-x_i\theta|\theta).$$

Apply the same trimming scheme as in (6.4), with the correction term, and leave the rest of the score function unchanged.[4] Then $s^{**}(\theta)-s(\theta)=s_U^{**}(\theta)-s_U(\theta)$. Because the untrimmed scores satisfy the orthogonality condition (2.12), and because $\mathrm{E}[D_g s_U^{**}(\theta_0)]=0$ by Lemma A1, it follows that the orthogonality condition $\mathrm{E}[D_g s^{**}(\theta_0)]=0$ also holds for the restricted model.

Asymptotics. The asymptotic score function at $\theta=\theta_0$ is now more complicated,

$$s_i(\theta_0)=s_{i,\theta}(\theta_0)-[r_1'(\theta_0)/r_2(\theta_0)]\,s_{i,\lambda}(\theta_0), \tag{7.11}$$

with the following components:

$$s_{i,\theta}(\theta_0)=\left(\mathrm{E}[X]-x_i\right)f'(\varepsilon_i)/f(\varepsilon_i) \tag{7.12}$$

[4] A simplification is to use $\widetilde{g}(u\,|\,\theta)$ instead of $\widetilde{f}(u\,|\,\theta)$ in the correction term. Both converge to $f(u)$ at the same rate, and there is no need for efficiency in estimating the bias.

(7.13) $$s_{i,\lambda}(\theta_0) = F(-x_i\theta) + H_0(-\varepsilon_i) - 2Q_0$$

(7.14) $$r_1'(\theta_0) = \int dv\, f(v) h_0(-v) \left(\mathrm{E}[X] - \mathrm{E}[X \mid X\theta_0 = -v] \right)$$

(7.15) $$r_2(\theta_0) = \int dv [h_0(-v) F(v)^2 + f(v) H_0(-v)^2] - 2Q_0^2 .$$

Asymptotic efficiency can again be demonstrated by the method discussed in Section 2.5, using the rate-of-convergence calculations summarized in the appendix. The only new feature to be taken into account is that, conditional on θ, (7.7) defines an implicit functional dependence of λ on $\tilde{g}_1$ and $\tilde{G}$. The analog of (7.9) is

$$D_g\lambda = -D_g[\tilde{r}_1(\theta)] / \tilde{r}_2(\theta),$$

which is then used in carrying out the linear expansion of the score in $(\tilde{g}_1 - g)$ and $(\tilde{G} - G)$.

The asymptotic variance of $\hat{\theta}$ is given by

(7.16) $$V_\theta^{-1} = \int d\varepsilon \frac{f'(\varepsilon)^2}{f(\varepsilon)} \mathrm{var}[X] + \frac{r_1'(\theta_0) r_1'(\theta_0)^T}{r_2(\theta_0)} .$$

The second term represents the improvement due to knowledge of the population shares. As expected, (7.11) is the efficient score and (7.16) the semiparametric efficiency bound.

8. Conclusion

Smoothing the log likelihood makes maximization with respect to unknown functions tractable in a number of cases, without compromising the efficiency of estimation from the concentrated likelihood. In fact, comparison with early attempts at semiparametric estimation using nonparametric maximum likelihood to estimate unknown density functions shows that a smoothed log likelihood does better than unsmoothed as far as efficiency is concerned. In some cases, the method provides a unifying perspective on existing efficient semiparametric estimators, while in other cases it can lead to new and reasonably tractable efficient estimators both for the parametric component and for functionals of the unknown distribution function.

Appendix

A.1. Orthogonality condition for the score function

The variational equation for maximizing $\mathrm{E}[\log \ell(x, y, \theta, f)]$ with respect to f is

$$\int dx\, dy\, \ell(x, y, \theta_0, f_0)\, D_f \log \ell(x, y, \theta, f(\cdot|\theta)) = 0$$

(see Newey, 1994, for an explanation of why this leads to the efficient score). Differentiating with respect to θ gives

$$\int dx\, dy\, \ell(x, y, \theta_0, f_0) \frac{d}{d\theta} D_f \log \ell(x, y, \theta, f(\cdot|\theta)) .$$

Define the (untrimmed) score function as the total derivative with respect to θ,

$$s(x, y, \theta, f(\cdot|\theta)) = \frac{\partial}{\partial\theta} \log \ell(x, y, \theta, f(\cdot|\theta)) + D_f \log \ell(x, y, \theta, f(\cdot|\theta))[\partial f(\cdot|\theta)/\partial\theta] .$$

Then

$$\int dx\, dy\, \ell(x, y, \theta_0, f_0)\, D_f s(x, y, \theta, f(\cdot|\theta)) = \mathrm{E}[D_f s(x, y, \theta, f(\cdot|\theta))] = 0 .$$

If $f(\varepsilon|\theta)$ has the form $f(\varepsilon, g(\cdot|\theta))$, and if $D_g f$ is a bounded linear operator, then also $\mathrm{E}[D_g s(x, y, \theta, f(\cdot|g(\cdot|\theta)))] = 0$.

Now consider the case of a trimmed score function. Let $\widetilde{g}$ be a vector of kernel estimates used to construct the estimated density function $\widetilde{f}$, and let g be the asymptotic limit of $\widetilde{g}$. Denote the score function by $\widetilde{s}(x, y, \theta)$ and the trimmed score by $\widetilde{s}^*(x, y, \theta)$, and their asymptotic limits by $s(x, y, \theta)$ and $s^*(x, y, \theta)$. In general, the orthogonality condition $\mathrm{E}[D_g\, s^*(X, Y, \theta_0)] = 0$ may not hold. The following results provide a method that can be used in certain cases to restore the orthogonality condition for the trimmed score.

*Lemma A*1. In the case where $y = x\theta_0 + \varepsilon$, with ε and x independent, suppose that $\mathrm{E}[s^*(X, Y, \theta_0)\,|\,\varepsilon] = 0$ and $\mathrm{E}[s(x, Y, \theta_0)\,|\,x] = 0$. Let

(A.1) $$\widetilde{s}^c(x, \theta) = \int dy\, \widetilde{f}(y - x\theta\,|\,\theta)\, \widetilde{s}^*(x, y, \theta)$$

and let $\widetilde{s}^{**}(x, y, \theta) = \widetilde{s}^*(x, y, \theta) - s^c(x, \theta)$ be the corrected trimmed score. Then:

(*i*) $\mathrm{E}[D_g s^{**}(X, Y, \theta_0)] = 0$; (*ii*) $s^{**}(x, y, \theta_0) \to s(x, y, \theta_0)$ as $b_n \to 0$, where b_n is the trimming parameter; and (*iii*) $\mathrm{E}[s^{**}(X, Y, \theta_0)] = 0$.

Proof. To verify (*i*), consider the functional derivative

$$D_g s^c(x,\theta_0) = \int dy\, f(y - x\theta_0 \mid \theta_0)\, D_g s^*(x,y,\theta_0) + \int dy\, (D_g f(y - x\theta_0 \mid \theta_0))\, s^*(x,y,\theta_0)$$
$$= \mathrm{E}[D_g s^*(x,Y,\theta_0) \mid x] + \int d\varepsilon\, (D_g f(\varepsilon \mid \theta_0))\, s^*(x, x\beta_0 + \varepsilon, \theta_0)$$

with expected value

$$\mathrm{E}[D_g s^c(X,\theta_0)] = \mathrm{E}[D_g s^*(X,Y,\theta_0)] + \int d\varepsilon\, (D_g f(\varepsilon \mid \theta_0))\, \mathrm{E}[s^*(X,Y,\theta_0) \mid \varepsilon].$$

By assumption the last term is zero and therefore

$$\mathrm{E}[D_g s^*(X,Y,\beta_0)] - \mathrm{E}[D_g s^c(X,\beta_0)] = 0$$

as required. To verify (*ii*), consider the asymptotic limit of (A.1) at $\theta = \theta_0$:

(A.2) $$s^c(x,\theta_0) = \int d\varepsilon\, f(\varepsilon \mid \theta_0)\, s^*(x, x\theta_0 + \varepsilon, \theta_0) = \mathrm{E}[s^*(x,Y,\theta_0) \mid x].$$

As $b_n \to 0$, $\mathrm{E}[s^*(x,Y,\theta_0) \mid x] \to \mathrm{E}[s(x,Y,\theta_0) \mid x]$, which is zero by assumption. Therefore $s^c(x,\theta_0) \to 0$, which is the same as $s^{**}(x,y,\theta_0) \to s(x,y,\theta_0)$. To verify (*iii*), note that (A.2) implies $\mathrm{E}[s^{**}(x,Y,\theta_0) \mid x] = 0$, and the result follows. □

*Lemma A*2. In the case where y is discrete, with $p(y \mid x,\theta) = p(y \mid x\theta)$, suppose that $\mathrm{E}[s^*(X,y,\theta_0) \mid y, x\theta_0] = 0$ and $\mathrm{E}[s(x,Y,\theta_0) \mid x] = 0$. Then the results of Lemma A1 hold if we replace the correction term (A.1) by

(A.3) $$\tilde{s}^c(x,\theta) = \sum_y \tilde{p}(y \mid x\theta,\theta)\, \tilde{s}^*(x,y,\theta).$$

The proofs of Lemmas A1 and A2 are essentially the same. An obvious corollary is that, under the stated conditions, $\mathrm{E}[D_g s^*(X,Y,\theta_0)] = 0$ if (A.1) or (A.3) is zero.

A.2. Asymptotic rates of convergence

Score functions

The convergence rate of the trimmed score and its derivative can be found by methods developed in the semiparametric literature by Ichimura and Lee (1991), Klein and Spady (1993), Ai (1997), and others. A detailed account is given in Ai (1997); for a discussion more closely aligned with the present set-up, see also Cosslett (2004). Essentially the same proofs hold here with minor

variations. We therefore summarize the procedure, including enough detail to explain the role of the orthogonality condition $\mathrm{E}[D_g\, s^*(Y, X, \theta_0)] = 0$.

As discussed in Section 2.5, one needs to show that $n^{-1}[\widetilde{S}^*(\theta) - S(\theta)] \to 0$ and $n^{-1}[\partial\widetilde{S}^*(\theta)/\partial\theta - \partial S(\theta)/\partial\theta] \to 0$ in probability uniformly in θ, and that $n^{-1/2}[\widetilde{S}^*(\theta_0) - S(\theta_0)] \to 0$ converges in probability. Each of these is done in two steps, for example $n^{-1}[\widetilde{S}^*(\theta) - S^*(\theta)] \to 0$ and $n^{-1}[S^*(\theta) - S(\theta)]$, where $S^*(\theta)$ denotes the trimmed version of $S(\theta)$. The first step is the critical one; the second step requires $\mathrm{E}[S^*(\theta_0)] = 0$, but then just depends on $b_n \to 0$ at a rate which can be made as slow as needed to accommodate the first step.

(*i*) Under some standard regularity conditions, there are uniform bounds for the convergence of kernel estimators $\widetilde{g}_j(u \mid \theta)$ and their derivatives to their limiting values. Specifically,[5]

$$| d^r \widetilde{g}(u - x\theta \mid \theta)/d\theta^r - d^r g(u - x\theta \mid \theta)/d\theta^r | = (c + |x|^r)\{O(h_n^{-r-1/2}) o_p(n^{-1/2+}) + O(h_n^2)$$

(where the first bound comes from uniform convergence and the second is the kernel bias). For definiteness, consider the regression estimator. In showing that $n^{-1}[\widetilde{S}^*(\theta) - S^*(\theta)] \to 0$ and $n^{-1}[\partial\widetilde{S}^*(\theta)/\partial\theta - \partial\widetilde{S}(\theta)/\partial\theta] \to 0$, the critical term (i.e., with the slowest rate of convergence) is

$$\tau(\widetilde{g})\,\widetilde{g}^{-1}(d^2\widetilde{g}/d\theta^2) - \tau(g)\,g^{-1}(d^2 g/d\theta^2) = (c + |x|^2)\, O(b_n^{-1} h_n^{-5/2})\, o_p(n^{-1/2+})$$

(anticipating that b_n will shrink more slowly than h_n), so we must have $b_n^{-1} h_n^{-5/2} n^{-1/2+} \to 0$.

(*ii*) A different technique is needed to show that $n^{-1/2}(\widetilde{S}^*(\theta_0) - S^*(\theta_0)) \to 0$, because the kernel estimates necessarily converge at a slower rate than $n^{-1/2}$. Instead, expand the difference in powers of $(\widetilde{g} - g)$ and $(d\widetilde{g}/d\theta - dg/d\theta)$, and then deal separately with the linear terms and the higher-order terms. The second-order terms are quite complicated: the worst-case terms are $\tau(g) g^{-2}(\widetilde{g} - g)(d\widetilde{g}/d\theta - dg/d\theta)$ and $\tau'(g) g^{-1}(\widetilde{g} - g)(d\widetilde{g}/d\theta - dg/d\theta)$ with rate $b_n^{-2} h_n^{1/2} n^{-1/2+}$, and $\tau''(g) g^{-1}(\widetilde{g} - g)^2 (dg/d\theta)$ with rate $b_n^{-3} h_n^4$. We therefore need $b_n^{-2} h_n^{1/2} \to 0$ and $b_n^{-3} h_n^4\, n^{1/2} \to 0$.

(*iii*) Finally, the linear terms are disposed of by combining the sum over observations *i* contributing to the score function with the sums over observations *j* in the kernel estimates,

[5] The notation n^{p+} means any power of n greater than *p*; logarithmic terms have been dropped.

resulting in a double sum of the form $n^{-3/2}\sum_{i,j}\psi_{i,j}$ with $\psi_{i,j} = \psi(\varepsilon_i, \varepsilon_j, x_i, x_j)$ and $\mathrm{E}[\psi_{i,j}] = 0$. This can be symmetrized to form a *U*-statistic, and the projection theorem of Powell et al. (1989) can be applied to show that this term is of order $\mathrm{E}[\psi_{i,j} \mid \varepsilon_i, x_i] + \mathrm{E}[\psi_{i,j} \mid \varepsilon_j, x_j]$. The first of these conditional expectations contains the bias in the kernel estimates and is $O(b_n^{-2} h_n^2)$, which is not a limiting factor, while the second term is zero. This where the condition $\mathrm{E}[D_g\, s(Y, X, \theta_0)] = 0$ comes into play—it is the reason for $\mathrm{E}[\psi_{i,j} \mid \varepsilon_j, x_j] = 0$, while otherwise the convergence rate would depend on the tail structure of the density *f* as well as on b_n and would be difficult to control.

Putting all the rate requirements together, the bandwidth h_n can shrink at a rate between $n^{-1/5}$ and $n^{-1/8}$ (as far as asymptotic theory is concerned).[6]

Integrals of trimmed kernel estimates

These convergence results can be extended to the case where a kernel estimate is integrated over an infinite range, such as the bias correction term in (6.4) or the constraint equation (7.7). This avoids the need to trim the range of integration, which would introduce a whole new layer of complexity. For definiteness, consider the example

$$\Delta\widetilde{M}(\theta) = \widetilde{M}(\theta) - M(\theta) = \int du \frac{\partial \widetilde{g}(u|\theta)}{\partial\theta} \tau(\widetilde{g}(u|\theta)) - \int du \frac{\partial g(u|\theta)}{\partial\theta} \tau(g(u|\theta))$$

where $\widetilde{g}(u|\theta)$ is given by (6.2). Rewrite as $\widetilde{M}(\theta) = \widetilde{M}_1(\theta) + \widetilde{M}_2(\theta)$, where

$$\widetilde{M}_1(\theta) = \int du \frac{\partial \widetilde{g}(u|\theta)}{\partial\theta} \tau(g(u|\theta)), \qquad \widetilde{M}_2(\theta) = \int du \frac{\partial \widetilde{g}(u|\theta)}{\partial\theta} [\tau(\widetilde{g}(u|\theta)) - \tau(g(u|\theta))].$$

Change the order of summation and integration,

$$\widetilde{M}_1(\theta) = -\frac{1}{n}\sum_{i=1}^{n} x_i \int du \frac{1}{h_n^2} K'\left(\frac{e_i - u}{h_n}\right) \tau(g(u|\theta))$$

where $e_i = y_i - x_i\theta$, and then integrate by parts and change the integration variable,

---

[6] The trimming factor b_n can then converge at a rate $n^{-\beta}$, where β is the smallest of $\alpha/4$, $(1-5\alpha)/2$, and $(8\alpha - 1)/6$. As with other kernel-based semiparametric estimators, these rates are very slow and are admittedly of little practical significance. Some previous simulation studies of kernel-based estimators for regression and binary choice suggest that the trimming parameter can in fact shrink faster than the rates that have been proved sufficient.

$$\widetilde{M}_1(\theta) = -\frac{1}{n}\sum_{i=1}^{n} x_i \int d\eta\, K(\eta)\, \tau'(g(e_i - h_n\eta \mid \theta))\frac{\partial g(e_i - h_n\eta \mid \theta)}{\partial \theta}.$$

This is a sum of independent terms, and $\widetilde{M}_1(\theta) - \mathrm{E}[\widetilde{M}_1(\theta)]$ is uniformly bounded (under conventional regularity conditions) by $o_p(n^{-1/2})O(b_n^{-1})$. Similarly,

$$\widetilde{M}_2(\theta) = -\frac{1}{n}\sum_{i=1}^{n} x_i \int d\eta\, K(\eta)\left(\tau'(\widetilde{g}(e_i - h_n\eta|\theta))\frac{\partial \widetilde{g}(e_i - h_n\eta|\theta)}{\partial \theta} - \tau'(g(e_i - h_n\eta|\theta))\frac{\partial g(e_i - h_n\eta|\theta)}{\partial \theta}\right).$$

Applying the uniform bounds $\partial\widetilde{g}/\partial\theta - \partial g/\partial\theta = o_p(n^{-1/2+})O(h_n^{-3/2})$ and $\tau'(\widetilde{g}) - \tau'(g) = O(h_n^2 b_n^{-2})$, we find that $\widetilde{M}_2(\theta)$ is bounded by $o_p(n^{-1/2+})O(h_n^{-3/2} b_n^{-1}) + O(h_n^2 b_n^{-2})$. The remaining term is the kernel bias $M_1(\theta) - \mathrm{E}[\widetilde{M}_1(\theta)] = O(h_n^2)$. Collecting terms, the bound on $\Delta\widetilde{M}(\theta)$ is $o_p(n^{-1/2+})O(h_n^{-3/2} b_n^{-1}) + O(h_n^2 b_n^{-2})$. This is the same as the rate of convergence of the integrand apart from a factor $O(b_n^{-1})$, which an adequate bound for present purposes.

If the integrand does not need to be trimmed, as in (7.7), the additional factor $O(b_n^{-1})$ does not arise.

Estimating a weighted integral of the distribution function

The rates of convergence associated with a weighted integral of the distribution function are different from those associated with the score function. For definiteness, consider binary choice with $\hat{\psi} = \hat{\psi}(a)$ defined by (3.7). The assumption that $A = \{u \mid a(u) > 0\}$ is bounded and that $h_*(u)$ is bounded away from zero for $u \in A$ means that eventually $\psi^* = \psi$, where ψ^* is the trimmed version of ψ, $\psi^* = \int du\, a(u) F(u) \tau[h_*(u)]$. Expand $\sqrt{n}(\hat{\psi} - \psi)$ to first order in $\widetilde{G}_1(u|\theta_0) - G_1(u|\theta_0)$, $\widetilde{G}(u|\theta_0) - G(u|\theta_0)$, and $\hat{\theta} - \theta_0$, noting that $G_1(u|\theta_0) = F(u) h_*(u)$ and $G(u|\theta_0) = h_*(u)$:

$$\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\int du \frac{a(u)}{h_*(u)}\frac{1}{h_n} K\left(\frac{x_i\theta_0 - u}{h_n}\right)[d_{1,i} - F(u)]$$

$$+\frac{1}{n}\sum_{i=1}^{n}\int du \frac{a(u)}{h_*(u)}\frac{1}{h_n^2} K'\left(\frac{x_i\theta_0 - u}{h_n}\right) x_i [d_{1,i} - F(u)]\cdot\sqrt{n}(\hat{\theta} - \theta_0) + o_p(n^{-1/2+})O(h_n^{-1} b_n^{-3})$$

where the remainder term comes from the uniform convergence rate for $\widetilde{G}_1(u|\theta)$ and $\widetilde{G}(u|\theta)$ and from expansion of the trimming factor. Under conventional regularity conditions, the

sample mean in the second sum converges in probability to $-V_2 + O(h_n^2)$, where V_2 is given by (3.10). Substitute for $\hat{\theta} - \theta_0$,

$$n^{1/2}(\hat{\theta} - \theta_0) = n^{-1/2} V_\theta \sum_{i=1}^{n} s_i(\theta_0) + o_p(1)$$

where V_θ is the asymptotic variance of $\hat{\theta}$ and $s_i(\theta_0)$ is the efficient score for the binary choice model. Suppose that $h_n = o(n^{-1/4})$, so that the asymptotic bias is negligible. Then the remaining sum

$$\frac{1}{\sqrt{n}} \sum_{i=1}^{n} \left\{ \int du \, \frac{a(u)}{h_*(u)} \frac{1}{h_n} K\left(\frac{x_i \theta_0 - u}{h_n} \right) [d_{1,i} - F(u)] - V_2 V_\theta \, s_i(\theta_0) \right\}$$

converges in distribution with asymptotic variance

$$V_\psi = \int du \, \frac{a(u)^2}{h_*(u)} F(u)[1 - F(u)] + V_2' V_\theta V_2$$

as required. Examining the bounds on the bias and remainder terms shows that we need $h_n \sim n^{-\alpha}$ with $\frac{1}{4} < \alpha < \frac{1}{2}$. (This is, of course, a different bandwidth from the one used in estimating θ.) A drawback of this approach is that slow convergence rate of the trimming parameter b_n generally does not allow $\psi^* - \psi$ to converge at the parametric rate, which is why we had to restrict $a(u)$ to functions with bounded support.[7]

References

Ai, C., "A semiparametric maximum likelihood estimator," *Econometrica* 65 (1997), 933–963.

Bickel, P. J., "On adaptive estimation," *Annals of Statistics* 10 (1982), 647–671.

Cosslett, S. R., "Efficient Estimation of Discrete Choice Models," in C. F. Manski and D. McFadden, eds., *Structural Analysis of Discrete Data with Econometric Applications,* (Cambridge, MA: MIT Press, 1981).

[7] Even with higher-order kernels, $\widetilde{G}$ necessarily converges more slowly than $n^{-1/2}$, and therefore convergence of the remainder term requires b_n to converge more slowly than $n^{-1/6}$.

Cosslett, S. R., "Efficiency bounds for distribution-free estimators from endogenously stratified samples," paper presented at the World Congress of the Econometric Society, 1985.

Cosslett, S. R., "Efficiency bounds for distribution-free estimators of the binary choice and the censored regression models," *Econometrica* 55 (1987), 559–585.

Cosslett, S. R., "Estimation from endogenously stratified samples," in *Handbook of Statistics, Vol. 11: Econometrics*. (North Holland, 1993).

Cosslett, S. R., "Efficient semiparametric estimation of censored and truncated regressions via a smoothed self-consistency equation," *Econometrica* 72 (2004), 1277–1293.

Ichimura, H., and L-F. Lee, "Semiparametric least squares estimation of multiple index models: single equation estimation," in W. A. Barnett, J. Powell, and G. E. Tauchen, eds., *Nonparametric and Semiparametric Methods in Econometrics and Statistics* (Cambridge: Cambridge University Press, 1991), 3–49.

Klein, R. W., and R. H. Spady, "An efficient semiparametric estimator of binary response models," *Econometrica* 61 (1993), 387–421.

Manski, C. and D. McFadden, "Alternative Estimators and Sample Designs for Discrete Choice Analysis," in C. F. Manski and D. McFadden, eds., *Structural Analysis of Discrete Data with Econometric Applications* (Cambridge, MA: MIT Press, 1981).

McFadden, D., "Econometric Analysis of Discrete Data," Fisher-Schultz Lecture, Athens, 1979.

Newey, W. K., "Semiparametric efficiency bounds," *Journal of Applied Econometrics* 5 (1990), 99–135.

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

Powell, J. L., J. H. Stock, and T. M. Stoker, "Semiparametric estimation of index coefficients," *Econometrica* 57 (1989), 1403–1430.