IMPROVING THE EFFICIENCY AND ROBUSTNESS OF THE SMOOTHED MAXIMUM SCORE ESTIMATOR

By Yulia Kotlyarova and Victoria Zinde-Walsh*
Department of Economics,
McGill University

March 16, 2004

Abstract

The binary-response maximum score (MS) estimator is a robust estimator, which can accommodate heteroskedasticity of an unknown form; J. Horowitz (1992) defined a smoothed maximum score estimator (SMS) and demonstrated that this improves the convergence rate for sufficiently smooth conditional error densities. In this paper we relax Horowitz's smoothness assumptions of the model and extend his asymptotic results. We also derive a joint limiting distribution of estimators with different bandwidths and smoothing kernels. We construct an estimator that combines SMS estimators for different bandwidths and kernels to overcome the uncertainty over choice of bandwidth when the degree of smoothnes of error distribution is unknown. A Monte Carlo study demonstrates the gains in efficiency and robustness.

*The support of the Social Sciences and Humanities Research Council of Canada (SSHRC), the FONDS QUEBECOIS DE LA RECHERCHE SUR LA SOCIETE ET LA CULTURE (FRQSC) and a travel grant from Wilfrid Laurier University are gratefully acknowledged. We thank the participants of the Canadian Econometric Study Group meeting and of the Econometric Society Australasian meeting. We also thank Don Andrews, Joel Horowitz, Robert De Jong, Chuck Manski, Murray Smith and two anonymous referees for very helpful discussion, comments and suggestions.

Corresponding author:
Victoria Zinde-Walsh
Department of Economics
McGill University,
855 Sherbrooke Street West,
Montreal, Quebec H3A 2T7
Canada
Email: victoria.zinde-walsh@mcgill.ca

1 Introduction

The maximum score (MS) estimator was introduced by Charles Manski (1975, 1985) as a robust alternative to traditional discrete-response estimators such as logit and probit. It allows for arbitrary dependence between the regressors and error term and does not impose restrictive distributional assumptions. The price of robustness is slow convergence rate ($n^{-1/3}$) and non-standard asymptotics (Kim and Pollard 1990).

J. Horowitz (1992) has proposed a smoothed version of the MS estimator. The original objective step-function was modified so that it became continuous and differentiable, and could be analyzed through the Taylor series approximation. The smoothed estimator has a rate of convergence that can under assumptions of sufficiently smooth cumulative distribution functions (CDF) be made arbitrarily close to $n^{-1/2}$ and also a normal asymptotic distribution; at the same time it preserves the robust qualities of the original estimator. Horowitz's results show that with smooth CDFs a higher-order smoothing function leads to a reduction in the MSE. However, unless the CDF smoothness assumptions hold, the rate improvement over the Manski estimator may be only marginal. Some plug-in methods were proposed by Horowitz to determine an optimal bandwidth, which minimizes the mean squared error (MSE), and to correct asymptotic bias. The process is not fully automatic and the MSE may vary substantially, as was shown by Monte Carlo experiments (Horowitz 1992). Most importantly, the estimates of the optimal bandwidth and of the asymptotic bias rely heavily on the assumption that the smoothness of CDFs is known. The incorrectly determined smoothness of the model may lead to oversmoothing or undersmoothing. Oversmoothing, which is caused by assuming the level of smoothness higher than the actual one, makes the estimator concentrate around the wrong value. Undersmoothing yields a consistent estimator but increases the mean squared error. Thus, the estimator which was intended to be a robust alternative to parametric techniques turns out to be sensitive to the smoothness properties of the model[1].

[1]Recently other binary-response estimators that allow for heteroskedasticity have been introduced. Assuming that the error distribution is conditionally independent of one of the regressors, Lewbel (2000) developed an asymptotically normal estimator which converges at the parametric rate. Khan (2001) proposed a heteroskedastic probit with incorporated sieve approximation; the estimator imposes stronger conditions on the smoothness of the model than the maximum score estimator.

Here, similarly to Horowitz, we consider a smoothed maximum score (SMS) estimator. There are two extensions of the asymptotic results for the SMS that we offer. First we extend the results of Horowitz to a wider class of models where the derivatives of the conditional CDF of the error term need not be smooth (we require only a uniform continuity condition); we also correct some problems that the proof in Horowitz (1992) had and thus confirm the validity of his results[2]. Second, similarly to Zinde-Walsh's (2002) results for the least median of squares estimator, we derive the joint limit process for SMS estimators with different bandwidths and kernel functions.

Additionally, here we propose a new estimation strategy that is robust to the degree of model smoothness. We consider a set of SMS estimators corresponding to different bandwidths (the set of bandwidths has to include undersmoothing and a Horowitz-optimal bandwidth) and, possibly, different functions (e.g. kernels of different order). We select a linear combination that minimizes the estimated mean squared error; we name the resulting estimator the "combined estimator". If Horowitz's smoothness conditions are satisfied, the combined estimator in comparison with Horowitz-optimal may lose some efficiency as a result of overparametrization, but since Horowitz-optimal estimator will always be considered as a candidate for combined estimator the loss cannot be too large. On the other hand, if the smoothness conditions do not hold, Horowitz-optimal estimator will have a large asymptotic bias caused by oversmoothing and thus will have a sub-optimal rate, but the combined estimator which always includes undersmoothed estimators among others could be asymptotically unbiased and achieve a better convergence rate. The results of our Monte Carlo experiments support these conclusions.

We find the loss of efficiency of the combined estimator relative to the best individual estimator to be small, and the performance to be uniformly good for combinations involving various sets of smoothing functions. In contrast, no individual Horowitz-optimal estimator delivers uniformly good performance over models with CDFs of varying degrees of smoothness: each one that has a low MSE in some case gives extremely bad results in some other cases.

The paper is organized as follows. Section 2 provides the definitions

[2]The problems in Horowitz's proofs were pointed out to us by D. Andrews and also by R. De Jong; they both questioned whether Horowitz's assumptions were sufficient for the results; we discuss additional assumptions.

and assumptions for the MS estimator and the SMS estimator; Horowitz's smoothness assumptions are discussed and generalized to require continuity rather than smoothness. We introduce alternative additional assumptions that permit us to fix the proof. Section 3 provides asymptotic results under our assumptions for the SMS estimator, as well as for the joint limit process for several SMS estimators. The new combined estimator is defined in Section 4, where we discuss how to construct it (selection of bandwidths, smoothing kernels, estimation of the MSE of a linear combination) and evaluate its performance in a Monte Carlo experiment.

Appendix A provides the proofs of the results in Section 3 and Appendix B provides the polynomial smoothing kernels that were used in our estimation.

2 Definitions, notation, assumptions

2.1 The binary choice model and Manski maximum score estimator

Consider the binary response model

$$y_i = sgn(x_i'\beta + u_i), i = 1, ..., n,$$

where $sgn(z) = \begin{cases} 1 & \text{if } z \geq 0 \\ -1 & \text{otherwise} \end{cases}$, $x_i \in R^k$ is a random vector of explanatory variables and u_i is a scalar error term.

Assumption 1 (Median Regression). For almost every x_i $med(u_i|x_i) = 0$.

Assumption 1 implies the same property for any scalar multiple of u_i; then β can be identified only up to scale. Consider β such that $\beta'\beta = 1$.

To estimate β from a sample of data (x_i, y_i) Manski (1975) proposed the maximum score (MS) estimator that solves the problem

$$\max_b \frac{1}{n} \sum y_i \cdot sgn(x_i'b) \text{ subject to normalization } b'b = 1, \tag{1}$$

where $\frac{1}{n}\sum y_i \cdot sgn(x_i'b)$ is called a *score function*[3]. The estimator matches up as many responses as possible. The formula (1) can be written in several equivalent forms as in Manski (1985).

[3] We utilize the sign function here rather than the indicator function; the two forms are equivalent.

The identification (even up to scale) is almost certain to fail whenever the support of X is finite or whenever one of the responses is a rare event. The next assumption ensures identifiability of $\widehat{b}$.

Let F_x be the k-variate marginal distribution of x.

Assumption 2.

(a) The support of F_x is not contained in any proper linear subspace of R^k.

(b) $0 < \Pr[y \geq 0|x] < 1$, for almost every x.

(c) The distribution of at least one of the regressors, x_j, conditional on $(x_1, ..., x_{j-1}, x_{j+1}, ..., x_k)$ has everywhere positive Lebesgue density. The corresponding coefficient $\beta_j \neq 0$.

(d) $\beta_0 = \beta / \|\beta\|$ is uniquely defined in the model with Assumption 1.

Assumption 3. (y_i, x_i), $i = 1, ..., n$, is a random sample of (y, x).

Under assumptions similar to these Manski (1975) demonstrated consistency of the MS estimator. Kim and Pollard (1990) make additional assumptions which we summarize as follows:

Assumptions KP.

(a) x_i has a continuous density $f_x(x)$;

(b) density $f_x(x)$ has compact support;

(c) $f_x(x)$ is continuously differentiable;

(d) The function $P_{|x}[(I(x'\beta + u \geq 0) - I(x'\beta + u < 0)]$ is continuously differentiable.

Under these assumptions, Kim and Pollard derived the limit process for the MS estimator: the estimator converges at rate $n^{-\frac{1}{3}}$ to the maximizer of a Gaussian process. The Assumptions KP can be partially relaxed.

2.2 Smoothed MS estimator

Horowitz (1992) considered a smoothed version of the problem (1):

$$\widehat{b} = \arg\max \frac{1}{n} \sum y_i \cdot K(\frac{x_i'b}{\sigma_n}) \text{ subject to normalization } b_1 = 1, \qquad (2)$$

where K is a smoothing kernel (similar to a distribution function). The normalization assumes that it is known which continuous variable appears with a non-zero coefficient (β_1) and requires the assumption of compactness of the parameter space.

Our smoothed version of (1) differs from (2) in a few minor details. First, since the results in Horowitz utilize only derivatives of K, we introduce smoothing via a function ψ such that $K'(\frac{x_i'b}{\sigma_n}) = \psi(\frac{x_i'b}{\sigma_n})$, where ψ is the usual kernel function. Then $\int sgn(x_i'b - \sigma_n w)\psi(w)dw = 2K(\frac{x_i'b}{\sigma_n}) - 1$.

Assumption 4.

(a) The smoothing function ψ is a continuously differentiable function with support in $[-1, 1]$;

(b) $\int \psi(w)dw = 1$;

(c) ψ is a kernel function of order h: $\int w^i\psi(w)dw = 0$ if $0 < i < h$, $h \geq 2$;

(d) The bandwidth parameter $\sigma_n \to 0$ and $\sigma_n n^{\frac{1}{3}} \to \infty$.

Second, we use Manski's normalization $b'b = 1$. This normalization allows a slightly less constrained model in not having to indicate which of the continuous components of x enters with a non-zero coefficient and automatically provides a compact parameter space. Thus, we solve

$$\widehat{b} = \arg\max_{b'b=1} n^{-1} \sum y_i \int sgn(x_i'b - \sigma_n w)\psi(w)dw. \tag{3}$$

Denoting Horowitz's estimator by b_H with $b_{H1} = 1$, b_H is in a compact space, we have that our $b = \dfrac{b_H}{||b_H||}$.

Third, we partition the vectors b and x_i in a different way from Horowitz, who projects x onto $z = x\beta$ and onto $\widetilde{x} = (x_2, ..., x_k)$. Consider the projector onto the space spanned by β, $P_\beta = \frac{\beta\beta'}{\beta'\beta} = \beta\beta'$ and orthogonal $M_\beta = I - P_\beta$; denote $x_i'\beta$ by z_i; then $x_i = P_\beta x_i + M_\beta x_i = \beta\beta' x_i + M_\beta x_i = \beta z_i + V_i$, where $V_i = M_\beta x_i$. Denote $M_\beta b$ by g, $\beta' b$ by b_β. This provides $b = \beta\beta' b + M_\beta b = b_\beta\beta + g$, $b'b = b_\beta^2 + g'g$, and $x_i'b = z_i b_\beta + V_i' g$[4].

Denote the density of z_i conditional on V_i by $f_{|V}(z)$ and the cumulative distribution of u_i conditional on z_i and V_i by $F_{|z,V}(u) \equiv F(u|z, V)$. For any integer $i > 0$ define $F^{(i)}_{|z,V}(-z) = \dfrac{\partial^i}{\partial z^i} F_{|z,V}(-z)$. Thus, $F^{(1)}_{|z,V}(-z) = -\dfrac{\partial}{\partial u} F(u|z, V)|_{u=-z} + \dfrac{\partial}{\partial z} F(u|z, V)|_{u=-z}$. Its smoothness depends on the

[4]The limit processes for this case are very similar in form to Horowitz's, and either bootstrap or the same methods as in Horowitz will provide estimates for the limiting moments. We thus do not focus on the differences and in referring to Horowitz's assumptions and proofs consider them applied to our (Manski's) normalization with appropriate modifications.

shape of the conditional density of u, $\frac{\partial}{\partial u}F(u|z,V)$, and the form of heteroskedasticity, $\frac{\partial}{\partial z}F(u|z,V)$.

We extend the results of Horowitz to cases of non-differentiable derivatives of the CDF of the error. In order to represent such results we need to distinguish between the degree of smoothness of the derivatives in the model and the order of kernel, denoted h. Pollard (1993) denoted by s the degree of smoothness of the conditional density $\frac{\partial}{\partial u}F(u|z,V)$ in some neighbourhood of $z=0$ for almost every V; he extended some of the results to fractional s including the range $1<s<1+\alpha$, $0<\alpha<1$, where $\frac{\partial}{\partial u}F(u|z,V)$ satisfied an α-order Lipschitz condition in the neighbourhood of zero (in Horowitz (1992) integer $s=h$, $h\geq 2$, so the smoothness of the derivatives is the same as the order of the kernel). We focus here on situations where we do not assume anything beyond continuity of $F^{(1)}_{|z,V}(-z)$; we denote this degree of smoothness by $s\geq 1_+$; this includes the cases considered by Pollard.

Assumption 5.

(a) For z in some neighbourhood of zero $N(0)$ and almost all V, the conditional density $f_{z|V}(z)$ exists, satisfies $0<|f_{z|V}(z)|<M<\infty$ and satisfies a Lipschitz condition at 0; also $f_{z|V}(z)$ exists and is bounded by M a.e.

(b) For z in some neighbourhood of zero $N(0)$, for the conditional distribution $F_{u|z,V}(u)$ its derivative $F^{(1)}_{|z,V}(-z)$ exists, satisfies $0<|F^{(1)}_{|z,V}(-z)|<M<\infty$ and is uniformly continuous at $z=0$ a.e.;

(c) The components of V and of the matrices VV' and $VV'VV'$ have finite first absolute moments.

If Assumption 8 of Horowitz is satisfied (or its analogue for this normalization), our Assumption 5(a) follows; if Assumption 9 holds, our Assumption 5(b) follows; thus our Assumptions 5(a,b) relax those of Horowitz. Note however that Kim and Pollard have more stringent Assumptions KP on the regressors and that in the absence of those or similar assumptions the rate and limit process for the Manski MS estimator have not been established. We also find that we cannot correct the error in Horowitz's proof without some additional restriction[5]. Adding the KP assumptions would be sufficient.

[5]De Jong and Woutersen also provide additional to Horowitz's assumptions in their working paper "Dynamic time series binary choice" (2003).

Another alternative is provided in the following assumption.

Assumption 6.

(a) $F^{(1)}_{|z,V}(-z)$ exists and is bounded by M a.e.; and

(b) $f_{z|V}(z)$ satisfies a Lipschitz condition $|f_{z|V}(z+\alpha) - f_{z|V}(z)| < M\alpha$ a.e.

Define the scalar constants $\delta_\psi \equiv \int \psi^2(w)dw$ and $\alpha_\psi = \int \psi(w)dw$; they determine the dependence of the asymptotic variance of the smoothed estimator on the smoothing function.

As in Horowitz (1992), we introduce matrices D and Q, which will characterize the asymptotic distribution (note that by Assumptions 5 the moments exist):

$D \equiv E\left[f_{z|V}(0)VV'\right]$ and

$Q \equiv 2E[F^{(1)}_{|z,V}(0)f_{z|V}(0)VV']$.

Recall that $V = M_\beta X$ thus for any vector α such that $M_\beta \alpha = 0$ both $Q\alpha = D\alpha = 0$. Denote the subspace onto which M_β projects by $R^{k-1}(M_\beta)$.

Assumption 7. The matrix Q has rank $k-1$ and is negative definite on the space $R^{k-1}(M_\beta)$.

3 Asymptotic results for the smoothed estimator

Under Horowitz's (1992) smoothness conditions on the derivatives of the conditional distribution, which correspond to integer s and require continuous differentiability of the derivatives in the neighbourhood of zero up to degree $h = s$, using any hth order kernel K' produces an optimal rate of $n^{-\frac{h}{2h+1}}$ for the estimator of β. The resulting distribution has an asymptotic bias that can be eliminated either by subtracting the estimate of the bias or by undersmoothing, in which case the bandwidth sequence approaches zero faster than at the optimal rate.

Here in subsection 3.1 we derive the limit process for the smoothed estimator $\hat{b}$ when the degree of smoothness is $s \geq 1_+$ (continuity). The resulting distribution is similar to Horowitz's but in non-smooth cases may have a slow (marginally better than $n^{-\frac{1}{3}}$) convergence rate.

In Section 3.2 we provide the joint distribution of smoothed MS estimators based on several bandwidths and smoothing functions; the joint distribution

implies that there may be efficiency gains from considering several estimators jointly.

3.1 Asymptotic results for the smoothed estimator with degree of smoothness $s \geq 1_+$ (continuity)

Without differentiability of the first derivative of the CDF the sharp conditions on the rate of the estimator stated in Horowitz's Theorem 2 do not hold. To express the conditions under which we can state asymptotic results we define

$$\begin{aligned} \xi(\sigma_n w, V) &= \left[1 - 2F_{u|z=\sigma_n w,V}(-\sigma_n w)\right] \cdot f_{z|V}(\sigma_n w) \\ &\quad +2\sigma_n w F^{(1)}_{|z=0,V}(0) f_{z|V}(0) \end{aligned} \tag{4}$$

and define

$$A(\sigma_n) = \frac{1}{\sigma_n} E(V \int \xi(\sigma_n w, V)\psi(w)dw). \tag{5}$$

Under assumptions 1-7 $A(\sigma_n)$ converges to 0 (see Appendix A, Lemma 1). Under Horowitz's assumptions a sharp rate for $A(\sigma_n)$ can be determined.

Theorem 1. *Under Assumptions 1 - 7, if σ_n is such that as $n \to \infty$*

(a) $n^{1/2}\sigma_n^{3/2} A(\sigma_n) \to 0$
then $n^{1/2}\sigma^{1/2}(b - \beta) \xrightarrow{d} N(0, \delta Q^{-1} D Q^{-1})$;
more specifically, $n^{1/2}\sigma^{1/2} M_\beta (b - \beta) \xrightarrow{d} N(0, \delta Q^{-1} D Q^{-1})$
and $P_\beta(b - \beta) = o_p(n^{-1}\sigma_n^{-1})$;

(b) $n^{1/2}\sigma_n^{3/2} A(\sigma_n) \to A$, *where* $0 < ||A|| < \infty$
then $n^{1/2}\sigma^{1/2}(b - \beta) \xrightarrow{d} N(-Q^{-1}A, \delta Q^{-1} D Q^{-1})$
and $P_\beta(b - \beta) = O_p(n^{-1}\sigma^{-1})$;

(c) $n^{1/2}\sigma_n^{3/2} A(\sigma_n) \to \infty$
then $\sigma_n^{-1}||A(\sigma_n)||^{-1}(b - \beta) + Q^{-1}||A(\sigma_n)||^{-1} A(\sigma_n) \xrightarrow{p} o_p(1)$.

Proof in Appendix A.

Thus for case (a) (undersmoothing) we obtain a limit normal distribution and for (b) and (c) the estimator is asymptotically biased. Without knowing the specific $s \geq 1_+$ all that is known is that for some rate of $\sigma_n \to 0$

there is undersmoothing: no asymptotic bias and a limiting Gaussian distribution, and for some slower convergence rate of σ_n there is oversmoothing: the estimator is not consistent. Existence of an optimal rate depends on convergence properties of $A(\sigma)$ that cannot be asserted without strengthening our assumptions[6].

3.2 The joint limit process for smoothed MS estimators

Assume that $b(\sigma_n, \psi)$ represents the smoothed MS estimator when the function ψ and bandwidth σ_n are utilized, and consider a number of values of $\sigma_n : \sigma_{n1} < \sigma_{n2} < ... < \sigma_{nm}$. Assume that σ_{ni} for $i \leq m'$ corresponds to undersmoothing (part (a) of Theorem 1) while σ_{ni} for i such that $m' < m'' < i \leq m$ corresponds to oversmoothing (part (c) of Theorem 1). If $m'' > m' + 1$ then σ_{ni} with $m' + 1 \leq i \leq m''$ corresponds to the optimal rate $O(n^{\frac{h}{2h+1}})$ in Horowitz if integer $s = h$.

We combine each σ_{ni} with each smoothing function ψ_j from some set of functions that satisfy Assumption 4, $j = 1, ..., l$. Denote by $A(\sigma_i, \psi_j)$ the function $A(\sigma)$ from (5) for the function $\psi = \psi_j$, and similarly $A(\psi_j)$ for the A in part (b) of Theorem 1. Define

$$\eta(\sigma_i, \psi_j) = \begin{cases} n^{1/2}\sigma_i^{1/2}(b(\sigma_i, \psi_j) - \beta) \text{ for } i = 1, ..., m' \\ n^{1/2}\sigma_i^{1/2}(b(\sigma_i, \psi_j) - \beta + Q^{-1}A(\psi_i)) \text{ for } i = m' + 1, ..., m'' \\ ||A(\frac{\sigma_i}{b_\beta}, \psi_j)||^{-1} \left[\sigma_n^{-1}(b(\sigma_i, \psi_j) - \beta) + Q^{-1}A(\frac{\sigma_i}{b_\beta}, \psi_j)\right] \\ \text{for } i = m'' + 1, ..., m. \end{cases}$$

Let $\delta = (\frac{\delta_{\psi_1}}{\alpha_{\psi_1}^2}, ..., \frac{\delta_{\psi_l}}{\alpha_{\psi_l}^2})$ and $\tau_{\psi_i \psi j} \equiv \frac{\int \psi_i(w)\psi_j(w)dw}{\alpha_{\psi_i}\alpha_{\psi j}}$. Note that δ and $\tau_{\psi_i \psi j}$ are invariant with respect to positive scale changes in the functions ψ therefore we can assume that $\alpha_{\psi_i} = 1$ for all i and then have $\delta = (\delta_{\psi_1}, ...\delta_{\psi_l})$ and $\tau_{\psi_i \psi_j} = \int \psi_i(w)\psi_j(w)dw$.

Theorem 2. *Suppose that Assumptions 1 - 7 hold for each bandwidth* $\sigma_{ni}, 1 \leq i \leq m$, *and for each* ψ_j, $1 \leq j \leq l$ *and that the functions* $\{\psi_j\}_{j=1}^l$ *form a linearly independent set.*

[6]Note that if the s is unbounded (infinite differentiability) choosing a higher h is always preferable asymptotically; therefore in this case as well one cannot find an optimal rate and weighting function that will ensure the lowest MSE.

(a) If each $\sigma_1, ..., \sigma_{m'}$ $(m' \leq m)$ *satisfies condition (a) of Theorem 1 then*

$$(\eta(\sigma_1, \psi_1)', ..., \eta(\sigma_1, \psi_l)', ..., \eta(\sigma_{m'}, \psi_1)', ..., \eta(\sigma_{m'}, \psi_l)')' \xrightarrow{d} N(0, \Psi \otimes Q^{-1}DQ^{-1})$$

where the $lm' \times lm'$ *matrix* Ψ *has elements*

$$\{\Psi\}_{ij} = \begin{cases} \tau_{\psi_i\psi_j} \text{ if } \sigma_i = \sigma_j, \\ \sqrt{d}\int \psi_i(w)\,\psi_j(dw)\,dw \quad \text{if} \quad \sigma_i/\sigma_j = d < \infty, \\ 0 \quad \text{if} \quad \sigma_i/\sigma_j \to 0 \quad \text{or} \quad \sigma_i/\sigma_j \to \infty; \end{cases}$$

(b) If each $\sigma_{m'+1}, ..., \sigma_{m''}$ $(m' \leq m'' \leq m)$ *satisfies condition (b) of Theorem 1 then*

$$(\eta(\sigma_{m'+1}, \psi_1)', ..., \eta(\sigma_{m'+1}, \psi_l)', ..., \eta(\sigma_{m''}, \psi_1)', ..., \eta(\sigma_{m''}, \psi_l)')' \xrightarrow{d} N(0, \Psi \otimes Q^{-1}DQ^{-1})$$

where the $lm' \times lm'$ *matrix* Ψ *has elements* $\{\Psi\}_{ij} = \sqrt{d}\int \psi_i(w)\,\psi_j(dw)\,dw$, *with* $\sigma_i/\sigma_j = d < \infty$;

(c) If each $\sigma_{m''+1}, ..., \sigma_m$ $(m'' \leq m)$ *satisfies condition (c) of Theorem 1 then*

$$(\eta(\sigma_{m''+1}, \psi_1)', ..., \eta(\sigma_{m''+1}, \psi_l)', ..., \eta(\sigma_m, \psi_1)', ..., \eta(\sigma_m, \psi_l)')' \xrightarrow{p} 0$$

(d) $Cov(\eta(\sigma_{i_1}, \psi_{j_1}), \eta(\sigma_{i_2}, \psi_{j_2})) \to 0$ *for* $1 \leq i_1 \leq m''$ *and* $m'' + 1 \leq i_2 \leq m$, *and any* j_1, j_2.

Thus, if the bandwidths approach 0 at different rates or $\int \psi_i(w)\psi_j(w)dw = 0$, the corresponding estimators $b(\sigma, \psi)$ are asymptotically independent. This is a consequence of the fact that only a small fraction of observations have any effect on the estimator, therefore reweighting observations with different kernel functions can produce estimators with independent limit processes.

4 The combined estimator

As the results in Section 3 show, an optimal rate for an SMS estimator may be problematic. Here we use the results of Theorem 2 to construct a new combined estimator that optimally combines several bandwidths and smoothing functions in the sample instead of focussing on a single bandwidth.

Although efficiency may suffer in straightforward cases when an optimal rate can be found, the Monte Carlo experiments show that the combined estimator provides remarkably robust performance over a variety of cases. Section 4.1 defines the combined estimator. Section 4.2 addresses practical issues of construction of the combined estimator. Section 4.3 discusses performance in a Monte Carlo experiment.

4.1 Definition of the combined estimator.

Suppose that bandwidths $\sigma_{n1} < \sigma_{n2} < ... < \sigma_{nm}$ represent sequences of rates where σ_{n1} corresponds to undersmoothing and σ_{nm} to oversmoothing; some optimal rate may or may not exist. For a set of smoothing functions $\psi_1, ..., \psi_l$, Theorem 2 indicates the structure of the joint limit distribution of $b(\sigma_{ni}, \psi_j)$.

Consider a linear combination $b(\{a_{ij}\}) = \sum a_{ij} b(\sigma_{ni}, \psi_j)$, $\sum a_{ij} = 1$. Assume that the biases, variances and covariances for all $b(\sigma_{ni}, \psi_j)$ are known. Then one could find weights $\{a_{ij}\}$ that minimize the mean squared error $MSE(b(\{a_{ij}\}))$. Each individual $b(\sigma_{ni}, \psi_j)$ is included, thus the minimized MSE cannot be above the MSE for individual (σ_{ni}, ψ_j).

To determine the weights in practice we need to estimate the biases and covariances of all $b(\sigma_{ni}, \psi_j)$.

Denote estimated biases and covariances by "hats".

Then $\widehat{MSE}(b(\{a_{ij}\})) = tr \sum a_{i_1 j_1} a_{i_2 j_2} \{\widehat{bias}(b(\sigma_{i_1}, \psi_{j_1}))\widehat{bias}(b(\sigma_{i_2}, \psi_{j_2}))$
$+\widehat{Cov}(b(\sigma_{i_1}, \psi_{j_1}), b(\sigma_{i_2}, \psi_{j_2}))\}$,
and the combined estimator is

$$\widehat{b_c} = b(\{\widehat{a}_{ij}\}), \text{ where } \{\widehat{a}_{ij}\} = \arg\min \widehat{MSE}(b(\{a_{ij}\})). \tag{6}$$

4.2 Construction of the combined estimator

4.2.1 Estimation of variances and biases

Consistent estimators for biases and covariances can be obtained by various procedures, e.g. by the bootstrap. Note that for $i = 1$ and $j = 1, ..., l$ all $b(\sigma_{ni}, \psi_j)$ are "undersmoothed" and thus asymptotically unbiased: $Eb(\sigma_{n1}, \psi_j) = \beta$; we can write that

$bias(b(\sigma_{ni}, \psi_j)) = E(b(\sigma_{ni}, \psi_j)) - E(b(\sigma_{n1}, \psi_j))$.

Then by bootstrap

$\widehat{bias}(b(\sigma_{ni},\psi_j)) = B^{-1}\sum_{s=1}^{B} b_s(\sigma_{ni},\psi_j) - l^{-1}B^{-1}\sum_{j=1}^{l}\sum_{s=1}^{B} b_s(\sigma_{n1},\psi_j)$,
where B is the number of bootstrap samples.
Similarly, an estimator of covariance, $\widehat{Cov}(b(\sigma_{i_1},\psi_{j_1}), b(\sigma_{i_2},\psi_{j_2}))$
$= B^{-1}\sum_{s=1}^{B}(b_s(\sigma_{i_1},\psi_{j_1}) - B^{-1}\sum b_s(\sigma_{i_1},\psi_{j_1}))$
$\times(b_s(\sigma_{i_2},\psi_{j_2}) - B^{-1}\sum b_s(\sigma_{i_2},\psi_{j_2}))$.

In our Monte Carlo experiment we used less computationally intensive estimators. We estimate variances at the highest bandwidth using the Horowitz formula (1992, Theorem 3) and then approximate variances for other bandwidths $Var(\psi,\sigma_j) = \frac{\sigma_t}{\sigma_j}Var(\psi,\sigma_t)$ and covariances $Cov(b(\psi_i,\sigma_j), b(\psi_s,\sigma_t))$ by $\sqrt{\frac{d_{jt}Var(\psi_i,\sigma_j)Var(\psi_s,\sigma_t)}{\delta_i\delta_s}}\cdot\int\psi_i(w)\psi_s(d_{jt}w)dw$, where $d_{jt} = \sigma_j/\sigma_t$. This result follows from Theorem 2[7].

The estimators with the smallest bandwidth (undersmoothing) are asymptotically unbiased. To find individual biases, we can subtract the average of estimators with the smallest bandwidth from actual estimators: $Bias(\psi,\sigma_j) = b(\psi,\sigma_j) - \overline{b}(.,\sigma_1)$.

4.2.2 Selection of functions and bandwidths

In our Monte Carlo experiment we use polynomial functions that satisfy Assumption 4 for $h \geq 2$. We also consider sets of functions that satisfy conditions leading to opposite asymptotic biases in estimators and to asymptotically independent estimators. The functions and their anticipated properties in the combined estimator are described in Appendix B.

The largest bandwidth in the set is the maximum of Horowitz-optimal bandwidths for individual estimators. If it belongs to a truly optimal function/bandwidth combination then as sample size increases it should yield the fastest convergence rate. Otherwise, it will correspond to oversmoothing.

The lowest bandwidth should represent undersmoothing. It is chosen on the basis of the estimated empirical distribution of $|x\beta|$. When the bandwidth is equal to some low quantile of $|x\beta|$, only a small fraction of observations is in the smoothing area; it leads to an asymptotically unbiased estimator. In

[7] One can go even further and calculate the variance of just one estimator, $Var(\psi_s,\sigma_t)$. Then other variances are related to the first one as $Var(\psi_i,\sigma_j) = \frac{\delta_i\sigma_t}{\delta_s\sigma_j}Var(\psi_s,\sigma_t)$.

our experiments we use the original Manski estimator b^{MS} to estimate the distribution of $|x\beta|$. For the sample size of 2000, the lowest bandwidth is set to be equal to the 25th percentile of that distribution. For other sample sizes we shrink the lowest bandwidth at the rate $n^{-1/3}$, that is, we determine it as the 25th percentile times $(size/2000)^{-1/3}$. The intermediate bandwidths are spread evenly in terms of the quantiles of $|xb^{MS}|$. Alternatively, we can find for the largest Horowitz-optimal bandwidth a corresponding quantile α from the distribution of $|xb^{MS}|$ and obtain other bandwidths as $\frac{i}{m}\alpha$th quantiles of $|xb^{MS}|$, $i = 1, ..., m$.

4.2.3 Estimation procedure for the combined estimator

The entire procedure for a combined estimator includes the following steps: (i) for each smoothing function, find the SMS estimator using a fixed bandwidth $n^{-1/(2h+1)}$, estimate the "optimal" bandwidths, choose their maximum as the highest bandwidth; (ii) find the original MS estimator, determine the 25th percentile of xb^{MS} and the smallest bandwidth; (iii) find the SMS estimators for all smoothing functions and bandwidths; (iv) estimate the biases and the covariance matrix; and (v) find the optimal weights for the linear combination and compute (6).

4.3 Performance of the combined estimator

If a Horowitz optimal function/bandwidth pair is included among the (σ_{ni}, ψ_j) and all the biases and variances are consistently estimated, the true MSE of the combined estimator at its worst will be approaching the MSE of the Horowitz-optimal estimator; it will eventually be smaller when the Horowitz procedure actually selects an inappropriately large σ. Our Monte Carlo study provides the finite sample confirmation of these relations.

All individual SMS estimators are evaluated at the bandwidths determined by the Horowitz's procedure. The weights and bandwidths in linear combinations of the estimators are chosen as described in 4.2.

4.3.1 DGP and estimation of the SMS and combined estimator

We consider four different data-generating processes. The first two models have infinitely differentiable derivatives of the conditional distributions; thus, the SMS estimator evaluated at the "optimal" bandwidth using some

high-order smoothing function should be a good choice. The same estimator, however, may be badly biased in two other models, in which the first derivative of the CDF of the error term is continuous but not differentiable. We also expect that the advantage from using a combination of estimators will be obvious in these non-smooth cases.

Similarly to Horowitz (1992), we work with the model

$$y = \begin{cases} 1 & \text{if } \beta_1 x_1 + \beta_2 x_2 + u \geq 0, \\ -1 & \text{otherwise.} \end{cases}$$

The true value of β is $(\frac{1}{\sqrt{2}}, \frac{1}{\sqrt{2}})$, $x_1 \sim N(1,1)$, and $x_2 \sim N(0,1)$. Four conditional distributions of the error term u are considered.

Distribution S (smooth homoskedastic): $u \sim$ logistic with median 0 and variance 1;

Distribution SH (smooth heteroskedastic): $u = 0.25(1+2z^2+z^4)v$, where $z = x_1 + x_2$ and $v \sim$ S;

Distribution NS (non-smooth homoskedastic):

$$pdf(u) = \begin{cases} 0.5 \text{ if } u \in [0,1], \\ 0.5 + 5u \text{ if } u \in [-0.1, 0), \\ -\frac{1}{38} - \frac{10}{38}u \text{ if } u \in [-2; -0.1), \\ 0 \text{ otherwise;} \end{cases}$$

Distribution NSH (non-smooth heteroskedastic): $u = 0.25(1+2z^2+z^4)v$, where $z = x_1 + x_2$ and $v \sim$ NS[8].

The sample sizes used in the experiments are $N = 2000, 4000$, and 8000. We deliberately have chosen relatively large samples in order to reduce some known small-sample-size effects. These effects include considerable underestimation of a true variance while using the Horowitz's formula (see Horowitz 1992) as well as sensitivity of the "optimal" bandwidth to the choice of the initial bandwidth. To achieve stable results, 1000 replications per experiment have been performed.

The procedure and formulas for finding SMS estimators at the "optimal" bandwidth are given in Horowitz (1992). Here we will only report the values of auxiliary parameters, which were not precisely determined there. The initial estimators are obtained using the bandwidth $\sigma_n = n^{-\frac{1}{2h+1}}$, where n is a sample size and h is the kernel order of a smoothing function. When calculating an estimator of the bias, we switch to the bandwidth $\sigma_n^{0.1}$ if $\sigma < .637$ and to the bandwidth $1.5\sigma_n$ otherwise. Having estimated the "optimal"

[8]The smoothness of distributions NS and NSH corresponds to $s = 2_-$ (Lipschitz condition).

bandwidth, we redo our maximization of the objective function. The results are then corrected using an estimator of the asymptotic bias. Similarly to Horowitz, we perform a grid search instead of a global optimization procedures appropriate for multivariate cases. The search is performed over $2000 \div 4000$ points on a unitary circle (since $b'b = 1$). The original MS estimators are found using the algorithm from Manski and Thompson (1986).

The combined estimator is constructed as described in Section 3.2. The functions are provided in Appendix F. For the combined estimator we use (a) the 4th order kernel $f4$ (combined at four bandwidths) providing the estimator *comb*4, (b) the set $\{f2, f4\}$ of a 2d and 4th order kernels at four bandwidths giving *comb*24, (c) the set $\{f3a, f3b\}$ of two orthogonal 3d order kernels at four bandwidths giving *comb*33, and (d) the set $\{g3a, g3b, g4\}$ of two 3d and one 4th order orthogonal kernels at four bandwidths yielding *comb*334.

4.3.2 Summary of the results

The results are summarized in Tables I-IV. Each table corresponds to a different data-generating process. We report the bias, the variance and the MSE of each estimator.

Smooth homoskedastic model (S).

The smoothness of the model corresponds to $s = \infty$, that is why the convergence rate of the estimators is determined by the kernel order of corresponding smoothing functions. The $f4$ estimator is the most accurate. Another estimator with the same convergence rate, $g4$, has a twice larger MSE which is explained by higher values of δ and $\int x^4\psi(x)dx$[9]. All combined estimators and the $f2$ estimator are strictly worse than $f4$ but better than $g4$. Simulation results confirm that the combined estimators behave well even when individual estimators have extremely large MSE (e.g., asymmetric third-order kernels $g3a$ and $g3b$).

Smooth heteroskedastic model (SH).

Although the model is also infinitely smooth, heteroskedasticity of the error term changes in a peculiar way the ranking of the estimators. The $f4$ estimator is dominated by the lower-order kernel $f2$. Their combined estimator *comb*24 is more precise than any of the individual estimators.

[9]If the model is smooth enough, the MSE at the "optimal rate", with the optimal bandwidth, is increasing in the following characteristics of a smoothing function ψ of order h: $\delta = \int \psi^2(x)dx$ and $\int x^h\psi(x)dx$.

Moreover, the $g4$ and $f4$ estimators have similar MSEs, although the former should have asymptotically an advantage. All individual estimators with symmetric kernels as well as all combined estimators yield good results: $MSE < 2MSE_{comb24}$, whereas the individual estimators with asymmetric kernels are heavily biased with $MSE > 4MSE_{comb24}$. Note that in the smooth homoskedastic model the large MSE of estimators with asymmetric third-order functions was caused mainly by their large variance. Since we correct individual estimators for asymptotic biases, the presence of substantial finite-sample biases is another indicator that under heteroskedasticity the finite-sample behaviour of the estimator may differ markedly from the limiting process even at $n = 8000$.

Non-smooth homoskedastic model (NS).

The best estimator is $f3b$; note that its shape resembles the conditional density function of the error term. At the same time, the estimator $f3a$ whose kernel is a mirror image of $f3b$ has the MSE 4-5 times larger. The performance of the $g3b$ estimator is even worse because the kernel is more erratic than $f3a$, yet in combinations these adverse effects all but disappear. The combined estimators and the estimators with traditional symmetric kernels $f2$ and $f4$ yield only slightly worse results than the most accurate $f3b$ estimator.

Non-smooth heteroskedastic model (NSH).

We observe that the bias of the $f4$ estimator does not diminish with the sample size. The Horowitz bias-correction procedure is not helpful since the real bias is of a lower order (<2) than the correction itself. Moreover, the magnitude of the bias is not consistent with the optimal ratio of the variance to the squared bias. In a sufficiently smooth one-dimensional model, this ratio is $2h$, where h is the order of the kernel[10]. Thus, at the optimal bandwidth the variance should be more than two times larger than the squared bias, whereas here the $f4$ estimator has squared bias 3-7 times larger than the variance. The bias of the combined estimators also contributes more than a half of the MSE but its fraction of the MSE diminishes with an increase in the sample size while the bias of the $f2$ and $f4$ estimators becomes relatively larger. The symmetric $g4$ estimator is the most accurate. Since the model

[10]From Horowitz's (1992) Theorem 2c the MSE-minimizing $\lambda = (\delta_\psi D)/(2h\alpha_\psi^2 A^2)$; $Var = n^{-1}\sigma^{-1}\delta_\psi D/Q^2$, and $Bias^2 = n^{-1}\sigma^{-1}\lambda\alpha_\psi^2 A^2/Q^2$. The ratio $\dfrac{Var}{Bias^2} = \dfrac{\delta_\psi D}{\lambda\alpha_\psi^2 A^2} = 2h$.

corresponds to $s = 2\_$, this estimator evaluated at the Horowitz-optimal rate should also be suboptimal. Possibly, its bias will start to dominate the variance at much larger sample size.

We summarize the performance of the estimators in the following table:

MSE	S	SH	NS	NSH
best	$f4$	$comb24$	$f3b$	$g4$
good	$f2, comb4,$ $comb24,$ $comb33,$ $comb334$	$f2, f4, g4,$ $comb4,$ $comb33,$ $comb334$	$f2, f4, comb4,$ $g4, comb24,$ $g3a, comb33,$ $comb334$	$comb4, f3b,$ $comb24,$ $comb33,$ $comb334$
fair	$g4$			$g3a$
bad	$f3a, f3b,$ $g3a, g3b$	$f3a, f3b,$ $g3a, g3b$	$f3a, g3b$	$f2, f4,$ $f3a, g3b$

The worst MSE in the "good" category would be less than two times the "best" MSE; in the "bad" category the best would be more than 2 times the MSE of the worst "good" estimator (in the NSH model for sizes 2000 and 4000 the MSE are somewhat closer).

The performance of individual smoothed MS estimators depends on the underlying data-generating process. The symmetric fourth-order kernel $f4$ and second-order kernel $f2$ are not appropriate for the heteroskedastic non-smooth model; the asymmetric third-order kernels are highly sensitive to the shape of the derivatives of the error term CDF: a match produces very good results but a mismatch is disastrous. Any of the combined estimators, on the contrary, yield stable results under all four specifications.

Appendix A. Proofs of the Theorems

Rewrite the smoothed score function from (3) using $x_i'b = z_i b_\beta + V_i'g$; then

$$\widehat{b} = \arg\max \frac{1}{n} \sum y_i \int sgn(b_\beta z_i + V_i'g - \sigma_n w)\psi(w)dw$$

subject to $b_\beta^2 + g'g - 1 = 0$.

The Lagrangian for this problem is
$L = n^{-1} \sum y_i \int sgn(b_\beta z_i + V_i'g - \sigma_n w)\psi(w)dw + \lambda(b_\beta^2 + g'g - 1)$.

A solution would have to satisfy the first-order conditions:

$$\frac{\partial L}{\partial b_\beta} = 2n^{-1}\sigma_n^{-1}\sum y_i\psi(\frac{b_\beta z_i + V_i'g}{\sigma_n})z_i + 2\lambda b_\beta = 0, \quad (7)$$

$$\frac{\partial L}{\partial g} = 2n^{-1}\sigma_n^{-1}\sum y_i\psi(\frac{b_\beta z_i + V_i'g}{\sigma_n})V_i + 2\lambda g = 0, \text{ and} \quad (8)$$

$$\frac{\partial L}{\partial \lambda} = b_\beta^2 + g'g - 1 = 0. \quad (9)$$

We introduce the notation:

$$C_i(b_\beta, g) = y_i\psi(\frac{b_\beta z_i + V_i'g}{\sigma_n}), \text{ so } C_i(b_\beta, 0) = y_i\psi(\frac{b_\beta z_i}{\sigma_n}); \quad (10)$$

$$B_i(b_\beta, g) = y_i\psi'(\frac{b_\beta z_i + V_i'g}{\sigma_n}).$$

Then (7) can be rewritten as

$$n^{-1}\sigma_n^{-1}\sum C_i(b_\beta, g)z_i + \lambda b_\beta = 0; \quad (11)$$

and (8) can be expanded as a function of g in some neighbourhood around $g = 0$ and in notation (10) written as

$$n^{-1}\sigma_n^{-1}\sum C_i(b_\beta, 0)V_i + [n^{-1}\sigma_n^{-2}\sum B_i(\widetilde{g})V_iV_i' + \lambda I]g = 0, \quad (12)$$

where $\widetilde{g} = \alpha g$ for some $0 \leq \alpha \leq 1$.

To prove Theorem 1 we provide limits for the terms of (12) in Lemma 1.

Lemma 1. Under the conditions of Theorem 1, for any sequence $\varepsilon \to 0$ as $n \to \infty$:

$$\sup_{|b_\beta - 1| < \varepsilon; ||g|| < \varepsilon} ||n^{-1}\sigma_n^{-2}\sum B_i(b_\beta, g)V_iV_i' - Q|| \xrightarrow{p} 0; \quad (13)$$

$$\sup_{|b_\beta - 1| < \varepsilon; ||g|| < \varepsilon} |n^{-1}\sigma_n^{-1}\sum C_i(b_\beta, g)z_i| \xrightarrow{p} 0; \quad (14)$$

and

$$n^{-1/2}\sum(\sigma^{-1/2}C_i(b_\beta,0)V_i - \frac{\sigma^{3/2}}{b_\beta^2}A(\frac{\sigma_n}{b_\beta})) \xrightarrow{d} N(0,\delta D) \tag{15}$$

as $n\to\infty,\ b_\beta - 1 \to 0$

Proof of (13).

First, consider the (l,m)th elements of the $k\times k$ matrices in (13) and show that for any sequence ε such that $\varepsilon\to 0$ as $n\to\infty$

$$\sup_{|b_\beta-1|<\varepsilon;|g||<\varepsilon} |n^{-1}\sigma_n^{-2}\sum B_i(b_\beta,g)V_{il}V_{im} - Q_{lm}| \xrightarrow{p} 0. \tag{16}$$

Denote $\zeta_i = \sigma_n^{-2}B_i(b_\beta,g)V_{il}V_{im},\ \ i=1,...,n$. This is an i.i.d. sequence for any g. We first show its uniform convergence to its mean: for any $\nu>0$, $a>0$, any $b: |b-1|<\tilde{\varepsilon}<1$ and any g there is N such that for any $n>N$

$$\Pr(\sup_{b,g} |n^{-1}\sum \zeta_i - E\zeta_i| > a) \le \nu \tag{17}$$

Indeed by Chebyshev's inequality

$\Pr(|n^{-1}\sum\zeta_i - E\zeta_i| > a) \le \dfrac{E(\zeta-E\zeta)^2}{na^2} = \dfrac{E\zeta^2-(E\zeta)^2}{na^2} \le \dfrac{E\zeta^2}{na^2}.$

For $E\zeta_i^2$ using the transformation $w = \frac{b_\beta z_i + V_i'g}{\sigma_n}$ and applying Assumptions 4(a) (implies that $\max|\psi'(\cdot)|$ is bounded) and 5(a,c) we obtain

$$\begin{aligned}
E\zeta_i^2 &= E\sigma_n^{-4}B_i^2(b_\beta,g)(V_{il}V_{im})^2 = \sigma_n^{-4}E(V_{il}V_{im})^2E_{|V}B_i(b_\beta,g)^2 \\
&= \sigma_n^{-4}E(V_{il}V_{im})^2\int\psi'(\frac{b_\beta z_i+V_i'g}{\sigma_n})^2 f_{z|V}(z)dz \\
&= \frac{1}{\sigma_n^3 b_\beta}E(V_{il}V_{im})^2\int f_{z|V}(\frac{w\sigma_n}{b_\beta}-\frac{V_i'g}{b_\beta})\psi'(w)^2dw \\
&\le \frac{2}{\sigma_n^3 b_\beta}E(V_{il}V_{im})^2 M\max|\psi'(\cdot)|^2,
\end{aligned}$$

and for $|b_\beta-1|<\tilde{\varepsilon}$

$$\Pr(\sup_{b,g}|n^{-1}\sum\zeta_i - E\zeta_i| > a) \le n^{-1}\sigma_n^{-3}\frac{2}{a^2(1-\tilde{\varepsilon})}\max_{l,m}E(V_{il}V_{im})^2M\max|\psi'(\cdot)|^2.$$

By Assumption 4(d) $n^{-1}\sigma_n^{-3}\to 0$ and thus there exists N such that for $n>N$

$$n^{-1}\sigma_n^{-3} < \frac{\nu a^2}{2\max_{l,m}E(V_{il}V_{im})^2M\max|\psi'(\cdot)|^2}$$

which proves (17).

Now find $E\zeta_i = E\sigma_n^{-2}B_iV_{il}V_{im}$.

Since the derivative $F^{(1)}_{|z,V}(-z)$ exists a.e. under Assumption 6(a)[11]:

$$\begin{aligned}
& E\sigma_n^{-2}B_iV_{il}V_{im} = E\sigma_n^{-2}V_{il}V_{im} \\
& \cdot \int\int sgn(z_i+u_i)\psi'(\frac{b_\beta z_i + V_i'g}{\sigma_n})\frac{\partial}{\partial u}F(u|z,V)du f_{z|V}(z)dz \\
= \; & E\sigma_n^{-2}V_{il}V_{im}\int\left[\left[1-2F_{u|z,V}(-z)\right]\psi'(\frac{b_\beta z_i + V_i'g}{\sigma_n})f_{z|V}(z)dz\right] \\
= \; & E\frac{1}{\sigma_n b_\beta}V_{il}V_{im}\int\left[1-2F_{u\left|z=\frac{w\sigma_n}{b_\beta}-\frac{V_i'g}{b_\beta},V\right.}(-\frac{\sigma_n}{b_\beta}w+\frac{V_i'g}{b_\beta})\right] \\
& \cdot f_{z|V}(\frac{w\sigma_n}{b_\beta}-\frac{V_i'g}{b_\beta})\psi'(w)dw
\end{aligned}$$

by using the substitution $w = \frac{b_\beta z_i + V_i'g}{\sigma_n}$. By Assumption 6(a), this can be written for some measurable function $\tilde{w} = \tilde{w}(w)$ with $|\tilde{w}| < |w|$ as

$$\begin{aligned}
& E\frac{1}{\sigma_n b_\beta}V_{il}V_{im}\int\left[1-2F_{u\left|z=-\frac{V_i'g}{b_\beta},V\right.}(\frac{V_i'g}{b_\beta})-2F^{(1)}_{|z,V}(-\frac{\sigma_n}{b_\beta}\widetilde{w}+\frac{V_i'g}{b_\beta})\frac{w\sigma_n}{b_\beta}\right] \\
& \cdot\left[f_{z|V}(-\frac{V_i'g}{b_\beta})+\xi_{z|V}(-\frac{V_i'g}{b_\beta},\frac{w\sigma_n}{b_\beta})\right]\psi'(w)dw \\
= \; & E(I_1+I_2+I_3+I_4),
\end{aligned}$$

where $\xi_{z|V}(-\frac{V_i'g}{b_\beta},\frac{w\sigma_n}{b_\beta}) = f_{z|V}(\frac{w\sigma_n}{b_\beta}-\frac{V_i'g}{b_\beta}) - f_{z|V}(-\frac{V_i'g}{b_\beta})$ and

$$I_1 = \frac{1}{\sigma_n b_\beta}V_{il}V_{im}(1-2F_{u\left|z=-\frac{V_i'g}{b_\beta},V\right.}(\frac{V_i'g}{b_\beta}))f_{z|V}(-\frac{V_i'g}{b_\beta})\int\psi'(w)dw,$$

[11] The role of Assumption 6(a) is to permit consideration of bounded derivative $F^{(1)}_{|z,V}(-z)$ a.e., which provides existence of suitable moments and permits truncation of $||V||$ in the moments, to obtain $\frac{V'g}{b_\beta}$ small enough (in $N(0)$) for appropriate ε, since $||g|| < \varepsilon$. If instead Assumtpions KP hold then $||V||$ is bounded and similarly $\frac{V_i'g}{b_\beta}$ is in $N(0)$. Then local properties (Assumption 5) can be used directly.

$$
\begin{aligned}
I_2 &= \frac{1}{\sigma_n b_\beta} V_{il} V_{im} \int (1 - 2F_{u\left|z=-\frac{V_i' g}{b_\beta}, V\right.}(\frac{V_i' g}{b_\beta})) \xi_{z|V}(-\frac{V_i' g}{b_\beta}, \frac{w\sigma_n}{b_\beta}) \psi'(w) dw, \\
I_3 &= -2\frac{1}{\sigma_n b_\beta} V_{il} V_{im} \int f_{z|V}(-\frac{V_i' g}{b_\beta}) F^{(1)}_{|z,V}\left(-\frac{\sigma_n}{b_\beta}\widetilde{w} + \frac{V_i' g}{b_\beta}\right) \frac{w\sigma_n}{b_\beta} \psi'(w) dw, \\
I_4 &= -2\frac{1}{\sigma_n b_\beta} V_{il} V_{im} \int F^{(1)}_{|z,V}\left(-\frac{\sigma_n}{b_\beta}\widetilde{w} + \frac{V_i' g}{b_\beta}\right) \\
&\quad \cdot \xi_{z|V}(-\frac{V_i' g}{b_\beta}, \frac{w\sigma_n}{b_\beta}) \frac{w\sigma_n}{b_\beta} \psi'(w) dw.
\end{aligned}
$$

Term $EI_1 = 0$ because $\int \psi'(w)dw = 0$ by Assumption 4(c).

By Assumption 6(b) $|\xi_{z|V}(-\frac{V_i' g}{b_\beta}, \frac{w\sigma_n}{b_\beta})| < \left|\frac{w\sigma_n}{b_\beta}\right| M$, by 5(c) moments $E|V_{il}V_{im}|$ exist, by 4(a) support of ψ is $[-1, 1]$ and $|\psi'|$ is bounded, thus we can bound

$$
|EI_2| \leq \frac{1}{b_\beta^2} E|V_{il}V_{im}| M \int |w| |\psi'(w)| dw \leq \frac{1}{b_\beta^2} \max_{l,m} E|V_{il}V_{im}| M \cdot 2 \max|\psi'(w)|;
$$

similarly

$$
|EI_3| \leq \frac{2}{b_\beta^2} \max_{l,m} E|V_{il}V_{im}| M^2 2 \max|\psi'(w)|;
$$

and

$$
|EI_4| \leq \frac{2}{b_\beta^2} \max_{l,m} E|V_{il}V_{im}| M^2 2 \max|\psi'(w)|.
$$

From the existence of moments it follows that for any ε_1 there is $\Gamma(\varepsilon_1) < \infty$ such that $|EI_j I(||V|| > \Gamma(\varepsilon_1))| < \varepsilon_1, j = 2, 3, 4$.

Now consider for each EI_j, $j = 2, 3, 4$, $\overline{EI_j} = EI_j I(||V|| \leq \Gamma(\varepsilon_1))$, then $\max_{l,m} |EI_j - \overline{EI_j}| < \varepsilon_1$.

Next, consider $|\overline{EI_2} + \overline{EI_3} + \overline{EI_4} - Q_{lm}| \leq |\overline{EI_2}| + |\overline{EI_3} - Q_{lm}| + |\overline{EI_4}|$. We establish that each term on the right-hand side goes to zero.

For any $\gamma > 0$ define some $\varepsilon(\gamma)$ that satisfies $\frac{\varepsilon(\gamma)\Gamma(\varepsilon_1)}{1 - \tilde{\varepsilon}} < \gamma$; then for $||g|| < \varepsilon(\gamma), |b_\beta - 1| < \tilde{\varepsilon}, ||V|| \leq \Gamma(\varepsilon_1)$ we get $\left|\frac{V_i' g}{b_\beta}\right| < \gamma$.

For $\left|\overline{EI_2}\right|$ select $\gamma_2 > 0$ for which

$$
\gamma_2 < \frac{\varepsilon_1 (1 - \tilde{\varepsilon})^2}{4M^2 \Gamma(\varepsilon_1)^2 \max|\psi'(w)|}.
$$

For $||g|| < \varepsilon(\gamma_2), |b_\beta - 1| < \tilde{\varepsilon}$, $||V|| \leq \Gamma(\varepsilon_1)$ we have $\left|\frac{V_i' g}{b_\beta}\right| < \gamma_2$ then

$$\left|1 - 2F_{u\left|z=-\frac{V_i' g}{b_\beta},V\right.}\left(\frac{V_i' g}{b_\beta}\right)\right| = \left|2f_{u|z=-\frac{V_i'\tilde{g}}{b_\beta},V}\left(\frac{V_i'\tilde{g}}{b_\beta}\right)\frac{V_i' g}{b_\beta}\right| \leq 2M\left|\frac{V_i'\tilde{g}}{b_\beta}\right| < 2M\gamma_2.$$

Then

$|\overline{EI_2}| \leq 2(1-\tilde{\varepsilon})^{-2}\Gamma(\varepsilon_1)^2 M^2 \gamma_2 \cdot 2\max|\psi'| < \varepsilon_1$.

Similarly by Assumption 4(d) we can find N_1 such that for $n > N_1$ we get that

$$\sigma_n < \frac{\varepsilon_1(1-\tilde{\varepsilon})^3}{4M^2\Gamma(\varepsilon_1)^2 \max|\psi'(w)|}.$$

Then $|\overline{EI_4}| \leq 2(1-\tilde{\varepsilon})^{-3}\Gamma(\varepsilon_1)^2 M^2 \sigma_n \cdot 2\max|\psi'| < \varepsilon_1$.

Next consider $|\overline{EI_3} - Q_{lm}|$.

By Assumption 5(a,b) (uniform continuity in $N(0)$) there is some γ_3 such that all $z : |z| < 2\gamma_3$ are in $N(0)$ and

$$\sup_{|z|<\gamma_3;|\tilde{z}|<2\gamma_3} |f_{z|V}(z)F^{(1)}_{|z,V}(-\tilde{z}) - F^{(1)}_{|z,V}(0)f_{z|V}(0)| < \frac{\varepsilon_1(1-\tilde{\varepsilon})^2}{4\Gamma(\varepsilon_1)^2\max|\psi'|}$$

For γ_3 find $\varepsilon(\gamma_3)$ such that for $||g|| < \varepsilon(\gamma_3)$ we have $|\frac{V_i' g}{b_\beta}| \leq \frac{\Gamma(\varepsilon_1)\varepsilon(\gamma_3)}{1-\tilde{\varepsilon}} < \gamma_3$ and find N_2 such that for $n > N_2$ we have $\frac{\sigma_n}{1-\tilde{\varepsilon}} < \gamma_3$; then

$$\begin{aligned} &\sup_{n>N;||g||<\varepsilon(\gamma_3)} \left| f_{z|V}(-\frac{V_i' g}{b_\beta})F^{(1)}_{|z,V}\left(-\frac{\sigma_n}{b_\beta}\widetilde{w} + \frac{V_i' g}{b_\beta}\right) - F^{(1)}_{|z,V}(0)\, f_{z|V}(0)\right| \\ < \;& \frac{\varepsilon_1(1-\tilde{\varepsilon})^2}{4\Gamma(\varepsilon_1)^2\max|\psi'|}. \end{aligned}$$

Then $|\overline{EI_3} - \frac{1}{b_\beta^2}Q_{lm}| < \varepsilon_1$. If $\tilde{\varepsilon} = \tilde{\varepsilon}(\varepsilon_1)$ is such that for $|b_\beta - 1| < \tilde{\varepsilon}(\varepsilon_1)$ we have $\max_{l,m}\left|\frac{1}{b_\beta^2}Q_{lm} - Q_{lm}\right| < \varepsilon_1$ then $\max_{l,m}|\overline{EI_3} - Q_{lm}| < 2\varepsilon_1$.

Combining we get that for any ε_1 we can find $\Gamma(\varepsilon_1)$ and then $N(\varepsilon_1) = \max\{N_1, N_2\}$ and $\varepsilon(\varepsilon_1) = \min\{\varepsilon(\gamma_2), \varepsilon(\gamma_3), \tilde{\varepsilon}(\varepsilon_1)\}$ so that

$$\sup_{|b_\beta-1|<\varepsilon(\varepsilon_1);||g||<\varepsilon(\varepsilon_1);n>N(\varepsilon_1)} |E\sigma_n^{-2}B_i V_{il}V_{im} - Q_{lm}| < 7\varepsilon_1. \tag{18}$$

Combining (17) with (18) we obtain (13).

Proof of (14).

Using Chebyshev's inequality for arbitrary g,

$$\Pr\left(\left|n^{-1}\sum C_i(b_\beta,g)\frac{z_i}{\sigma_n}\right|>a\right)\leq\frac{n^{-1}E(C_i(g)\frac{z_i}{\sigma_n})^2}{a^2}.$$

Compute

$E\left(C_i(b_\beta,g)\frac{z_i}{\sigma_n}\right)^2=E\int\psi^2(\frac{b_\beta z_i+V_i'g}{\sigma_n})\left(\frac{z_i}{\sigma_n}\right)^2 f_{z|V}(z)dz;$

by substituting $w=\frac{b_\beta z_i}{\sigma_n}$ this is

$E\frac{\sigma_n}{b_\beta}\int\psi^2(w+\frac{V_i'g}{\sigma_n})\left(\frac{w}{b_\beta}\right)^2 f_{z|V}(\frac{\sigma_n}{b_\beta}w)dw,$

where for any g, and any $b_\beta:|1-b_\beta|<\tilde{\varepsilon}<1$

$$\left|E\frac{\sigma_n}{b_\beta}\int\psi^2(w+\frac{V_i'g}{\sigma_n})\left(\frac{w}{b_\beta}\right)^2 f_{z|V}(\frac{\sigma_n}{b_\beta}w)dw\right|<\frac{\sigma_n}{(1-\tilde{\varepsilon})^3}2\max\psi^2\cdot M.$$

Since $\sigma_n\to 0$ by Assumption 4(d) this provides (14).

Proof of (15).

Consider first $A(\sigma_n)=\frac{1}{\sigma_n}E(V\int\xi(\sigma_n w,V)\psi(w)dw)$

$=\frac{1}{\sigma_n}EV\int\{\left[1-2F_{u|z=\sigma_n w,V}(-\sigma_n w)\right]\cdot f_{z|V}(\sigma_n w)$

$+2\sigma_n wF^{(1)}_{|z=0,V}(0)f_{z|V}(0)\}\psi(w)dw.$

Support of $\psi(w)$ is $|w|<1$, therefore $|\sigma_n w|<\sigma_n$ and since $\sigma_n\to 0$ we have $\sigma_n w\in N(0)$ for large enough n.

Next using

(i) $F_{u|z=0,V}(0)=\frac{1}{2}$:

$1-2F_{u|z=w\sigma_n,V}(-w\sigma_n)=-2F^{(1)}_{|z=\sigma_n\widetilde{w},V}(-\sigma_n\widetilde{w})\sigma_n w,\ \ 0<\widetilde{w}<w;$

(ii) Assumption 5 for $F^{(1)}_{|z,V}(-z)$ and the Lipschitz condition for $f_{z|V}(z)$, we have that $\sigma^{-1}\xi(\sigma_n w,V)\to 0$ uniformly a.e., thus $A(\sigma_n)\to 0$.

Consider $\eta_i=\sigma_n^{-1/2}C_i(b_\beta,0)V_i-\frac{\sigma_n^{3/2}}{b_\beta^2}A(\frac{\sigma_n}{b_\beta})$; since $M_\beta\eta_i=\eta_i$ we can consider the vectors restricted to $R^{k-1}(M_\beta)$.

Conditional expectation $E_{|V}\left(\sigma_n^{-1/2}C_i(b_\beta,0)\right)$ can be easily shown by direct computation to equal $\frac{\sigma_n^{1/2}}{b_\beta}\int\xi(\frac{\sigma_n}{b_\beta}w,V)\psi(w)dw-2\frac{\sigma_n^{3/2}}{b_\beta^2}F^{(1)}_{|z,V}(0)f_{z|V}(0)\int w\psi(w)dw;$ the second term is zero by Assumption 4(c); by definition of $A(\frac{\sigma_n}{b_\beta})$ (see (5)) it follows that $E(\eta_i)=0$.

Variance of η_i can be computed as

$\Omega = \frac{1}{b_\beta}\delta D + \frac{1}{b_\beta}EVV' \int [f_{z|V}(\frac{\sigma w}{b_\beta}) - f_{z|V}(0)] \cdot \psi(w)^2 dw - \frac{\sigma_n^3}{b_\beta^4}\left(A(\frac{\sigma_n}{b_\beta})\right)^2$.

The second term by Lipschitz condition on $f_{z|V}(.)$ is of order $O(\sigma_n)$, the third $o(\sigma_n^3)$.

For any ε_1 consider N such that σ_n for $n > N$ is small enough for any $b : |b_\beta - 1| < \tilde{\varepsilon}$ that $||\Omega - \delta D|| < \varepsilon_1$. Note that D is positive definite on the subspace considered. Then by Lindeberg-Levy Theorem for the iid vectors η_i: $n^{-1/2}(\delta D)^{-1/2} \sum \eta_i \xrightarrow{d} N(0, I)$. Statement (15) follows. ■

Proof of Theorem 1.

From (14) of Lemma 1 and (11) it follows that

$$\lambda = o_p(1) \text{ for any } g \text{ and any } 0 < 1 - \tilde{\varepsilon} < b_\beta. \tag{19}$$

Note that the estimator in the Manski's normalization used here ($b : ||b|| = 1$) is related to Horowitz's b_H with $b_{H1} = 1$ by $b = \dfrac{b_H}{||b_H||}$. Then by Horowitz (1992) Theorem 1 $b_H \xrightarrow{a.s.} \beta$; consequently, $||b_H|| \xrightarrow{a.s.} ||\beta||$ and then $b \xrightarrow{a.s.} \beta$ (with $||\beta|| = 1$) and $(g, b_\beta) \xrightarrow{a.s.} (0, 1)$. This means that for any ε a large enough N exists such that $|b_\beta - 1| < \varepsilon$ a.s. and $||g|| < \varepsilon$ a.s. Select $\tilde{\varepsilon} > 0$ such that the expansion (11) is valid and select an arbitrary sequence $\varepsilon \to 0$, $\tilde{\varepsilon} > \varepsilon$. Choose $N(\varepsilon)$ such that $|b_\beta - 1| < \varepsilon$ a.s.

Thus by (13) and (19)

$$\begin{aligned} \sqrt{n\sigma}g &= -Q^{-1}\left[n^{-1/2}\sum\left(\sigma^{-1/2}C_i(0)V_i - \frac{\sigma^{3/2}}{b_\beta^2}A(\frac{\sigma_n}{b_\beta})\right)\right] \\ &\quad -Q^{-1}n^{1/2}\frac{\sigma^{3/2}}{b_\beta^2}A(\frac{\sigma_n}{b_\beta}) + o_p(1). \end{aligned} \tag{20}$$

If (a) holds for σ_n then for $b_\beta \to 1$ it also holds if σ_n is replaced by $\frac{\sigma_n}{b_\beta}$. So $n^{1/2}\frac{\sigma^{3/2}}{b_\beta^2}A(\frac{\sigma_n}{b_\beta}) = o(1)$ as $n \to \infty$; then $\sqrt{n\sigma}g \xrightarrow{d} N(0, \delta Q^{-1}DQ^{-1})$. By (5) $|b_\beta - 1| = |\sqrt{1 - g'g} - 1| = \frac{1}{2}g'g + O_p(g'g)^2 = O_p(n^{-1}\sigma^{-1})$.

If (b): $n^{1/2}\sigma_n^{3/2}A(\frac{\sigma_n}{b_\beta}) \to A = const$, then $\sqrt{n\sigma}g \xrightarrow{d} N(-Q^{-1}A, \delta Q^{-1}DQ^{-1})$.

If (c) holds: $n^{1/2}\sigma_n^{3/2}A(\frac{\sigma_n}{b_\beta}) \to \infty$, then

$\sigma_n^{-1}||A(\frac{\sigma_n}{b_\beta})||^{-1}g + Q^{-1}||A(\frac{\sigma_n}{b_\beta})||^{-1}A(\frac{\sigma_n}{b_\beta}) \xrightarrow{p} o_p(1)$ and $|b_\beta - 1| = o_p(\sigma_n^2 A(\sigma_n)^2)$, thus the conclusion of (c) holds.

It follows from condition (c) that for large enough n $A(\sigma_n) \neq 0$, thus $||A(\sigma_n)||^{-1}A(\sigma_n)$ is a unit length vector and $Q^{-1}||A(\sigma_n)||^{-1}A(\sigma_n) = O_p(1)$ and is bounded away from zero.

Proof of Theorem 2.

To prove Theorem 2 all that is required in addition to the results in Theorem 1 is to consider covariances between $\eta(\sigma, \psi)$. Denote by $C(b_\beta, 0, (\sigma, \psi))$ the $C_i(b_\beta, 0)$ that corresponds to the pair (σ, ψ) (for observation i). By independence it still follows that the terms appearing in the covariances are non-zero only for the same i. Here for entries in the covariance matrix

$$E(C(b_\beta, 0, (\sigma_{i_1}, \psi_{j_1}))C(b_\beta, 0, (\sigma_{i_2}, \psi_{j_2}))) = E\left(\psi_{j_1}(\frac{b_\beta z}{\sigma_{i_1}})\psi_{j_2}(\frac{b_\beta z}{\sigma_{i_2}})VV'\right)$$

$$= \int \left(\psi_{j_1}(\frac{b_\beta z}{\sigma_{i_1}})\psi_{j_2}(\frac{b_\beta z}{\sigma_{i_2}})f_{z|V}(z)dz\right) VV'dF(V)$$

$$= \frac{\sigma_{i_1}}{b_\beta}\int\left(\int \psi_{j_1}(w)\psi_{j_2}(w\frac{\sigma_{i_1}}{\sigma_{i_2}})f_{z|V}(\frac{\sigma_{i_1}}{b_\beta}w)dw\right) VV'dF(V)$$

$$= \frac{\sigma_{i_1}}{b_\beta}\int \psi_{j_1}(w)\psi_{j_2}(\frac{\sigma_{i_1}}{\sigma_{i_2}}w)dw \cdot E\left[f_{z|V}(0)VV'\right] + o(\sigma_{i_1})$$

$$= \begin{cases} \frac{\sigma_{i_1}}{b_\beta}\tau_{ij}D + o(\sigma_{i_1}) \text{ if } \sigma_{i_1} = \sigma_{i_2}, \\ \frac{\sigma_{i_1}}{b_\beta}\int \psi_{j_1}(w)\,\psi_{j_2}(dw)\,dw \cdot D + o(\sigma_{i_1}) \text{ if } \sigma_{i_1}/\sigma_{i_2} = d < \infty, \\ \frac{\sigma_{i_1}}{b_\beta}\psi_{j_2}(0)D + o(\sigma_{i_1}) \text{ if } \sigma_{i_1}/\sigma_{i_2} \to 0 \\ \frac{\sigma_{i_2}}{b_\beta}\psi_{j_1}(0)D + o(\sigma_{i_2}) \text{ if } \sigma_{i_1}/\sigma_{i_2} \to \infty \end{cases}$$

Then the covariance matrix of the limiting joint distribution includes

$$(\sigma_{i_1}\sigma_{i_2})^{-1/2}E\left(\psi_{j_1}(\tfrac{z}{\sigma_{i_1}})\psi_{j_2}(\tfrac{z}{\sigma_{i_2}})VV'\right)$$

$$\to \begin{cases} \tau_{ij}D \text{ if } \sigma_{i_1} = \sigma_{i_2}, \\ \sqrt{d}\int \psi_{j_1}(w)\,\psi_{j_2}(dw)\,dw \cdot D \text{ if } \sigma_{i_1}/\sigma_{i_2} = d < \infty, \\ \sqrt{\frac{\sigma_{i_1}}{\sigma_{i_2}}}\psi_{j_2}(0)D = o(1)D \text{ if } \sigma_{i_1}/\sigma_{i_2} \to 0\ , \\ \sqrt{\frac{\sigma_{i_2}}{\sigma_{i_1}}}\psi_{j_1}(0)D = o(1)D \text{ if } \sigma_{i_1}/\sigma_{i_2} \to \infty. \end{cases}$$

So (a,b) follow; (c) follows from (c) of Theorem 1. For (d) the covariances are zero because $\sigma_{i_1}/\sigma_{i_2} \to 0$.■

5 Appendix B. Smoothing functions and subsets for use in combined estimators.

We provide seven smoothing functions and four different combinations of estimators. The smoothing functions are selected to be polynomials that satisfy Assumption 4 (a,b,c).

Consider an n-degree polynomial, $\sum_{i=0}^{n} a_i x^i$.

1. Assumption 4(a) corresponds to the following restrictions, imposed on the coefficients of the polynomial:

$\sum_{i=0}^{n} a_i(-1)^i = 0; \quad \sum_{i=0}^{n} a_i = 0; \quad \sum_{i=0}^{n} i a_i(-1)^{i-1} = 0; \quad \sum_{i=0}^{n} i a_i = 0.$

2. Assumption 4(b) requires $\sum_{i=0}^{n} \frac{a_i}{i}(1-(-1)^{i+1}) = 1$.

3. Assumption 4(c)corresponds to $\sum_{i=0}^{n} \frac{a_i}{i+1}(1-(-1)^{i+2}) = 0, etc.$

A simple second-order kernel is

$f2 = \frac{15}{16}\left(1-x^2\right)^2$.

A standard fourth-order kernel (used also by Horowitz 1992) is

$f4 = \frac{105}{64}\left(1-3x^2\right)\left(1-x^2\right)^2$.

From Theorem 2 it follows that there may be benefits from using orthogonal polynomials in a combined estimator since they lead to asymptotically independent SMS estimators. The orthogonality condition for two such distinct polynomials ψ_i, ψ_j, $i \neq j$, is

4. $\int \psi_i(x)\psi_j(x)dx = 0$.

The biases of SMS estimators based on a pair (ψ_i, ψ_j) may offset each other for non-symmetric function if

5. $\psi_i(x) = \psi_j(-x)$.

Finally, asymptotic variance of SMS is proportionate to $\int \psi^2$, thus when combining estimators for different functions $\psi_1, ..., \psi_l$ one may wish to impose

6. $\int \psi_i^2(x)dx = const$ for $i = 1, ..., l$.

We construct two kernels of third order, $f3a$ and $f3$, that satisfy conditions 1-6.

$f3a(x) = \frac{105}{64}\left(1-3x^2\right)\left(1+\sqrt{23}x\right)\left(1-x^2\right)^2$ and

$f3b(x) = \frac{105}{64}\left(1-3x^2\right)\left(1-\sqrt{23}x\right)\left(1-x^2\right)^2$.

In fact these two polynomials are the smallest order (seven) that permits solving the equations for the coefficients that are imposed by conditions 1-5;

condition 6 is satisfied automatically here.

Three orthogonal polynomials of degree 8 (two 3rd-order kernels and one 4th order kernel) are constructed so that the conditions 1-4, 6 are satisfied; condition 5 is satisfied for the two distinct 3rd order kernels and the 4th order kernels is a symmetric function.

$g3a(x) = -\frac{9\left(\sqrt{17}+6\sqrt{2}\right)}{5632}(605\sqrt{17}x^4 + (48\sqrt{714} - 576\sqrt{21})x^3 + (336\sqrt{2}$
$-386\sqrt{17})x^2 + (192\sqrt{21} - 16\sqrt{714})x - 112\sqrt{2} + 37\sqrt{17})\left(1-x^2\right)^2;$
$g3b(x) = -\frac{9\left(\sqrt{17}+6\sqrt{2}\right)}{5632}(605\sqrt{17}x^4 - (48\sqrt{714} - 576\sqrt{21})x^3 + (336\sqrt{2}$
$-386\sqrt{17})x^2 - (192\sqrt{21} - 16\sqrt{714})x - 112\sqrt{2} + 37\sqrt{17})\left(1-x^2\right)^2;$
$g4(x) = \frac{45\left(-\sqrt{17}+12\sqrt{2}\right)}{138752}\left(271x^2 - 79 + 8\sqrt{34}\right)\left(11\sqrt{17}x^2 - 3\sqrt{17} - 8\sqrt{2}\right)$
$\times\left(1-x^2\right)^2.$

The $f4$ estimator will serve as a benchmark when we evaluate the performance of various combinations in the two smooth models. The Horowitz-optimal bandwidths, however, should lead to oversmoothing in non-smooth cases. Therefore, a weighted average of the SMS estimators with the same $f4$ smoothing function and different bandwidths, *comb*4, should improve upon the individual $f4$ estimator for distributions NS and NSH.

When the model is infinitely smooth, the convergence rate of the $f2$ estimator should be slower than that of the first estimator. For that reason, it is expected to be more reliable in non-smooth models. So, we evaluate separately the $f2$ estimator at the "optimal" rate and also the combination *comb*24 of estimators with kernels $f2$ and $f4$. It is expected that the combination will have a more robust behaviour in general than any individual estimator.

The choice of smoothing functions may have important effects on the efficiency of the estimator. Individually, kernels $g3a$, $g3b$ and $g4$ are not as good as kernels $f3a$, $f3b$ and $f4$. They oscillate more, resulting in higher values for the integral of squared functions, and consequently, in larger variances of corresponding estimators since the variance is proportional to δ. Indeed, while $\delta_{f4} = 1.4$ and $\delta_{f3a} = \delta_{f3b} = 2.8$, the value of δ for functions $g3a$, $g3b$ and $g4$ is 3.8. Also, asymptotic biases in smooth models, which are proportional to $\int x^3\psi(x)dx$ with third-order kernels and to $\int x^4\psi(x)dx$ with fourth-order kernels, are larger for the last three functions. By comparing their combination, *comb*334, to other combined estimators we wish to evaluate robustness of combined estimators against suboptimal (on its own) choice of smoothing functions.

References

[1] Horowitz, J. L. (1992) A Smoothed Maximum Score Estimator for the Binary Response Model. *Econometrica* 60, 505-531.

[2] Horowitz, J. L. (2002) Bootstrap Critical Values for Tests Based on the Smoothed Maximum Score Estimator. *Journal of Econometrics* 111, 141-167.

[3] Khan, S. (2001) Estimation of Semiparametric Binary Choice Models Using Probit Criterion Function. University of Rochester, working paper.

[4] Kim, J. and D. Pollard (1990) Cube Root Asymptotics. *The Annals of Statistics* 18, 191-219.

[5] Lewbel A. (2000) Semiparametric qualitative response model estimation with unknown heteroskedasticity or instrumental variables. *Journal of Econometrics* 97, 145-177.

[6] Manski, C. F. (1975) Maximum Score Estimation of the Stochastic Utility Model of Choice. *Journal of Econometrics* 3, 205-228.

[7] Manski, C. F. (1985) Semiparametric Analysis of Discrete Response: Asymptotic Properties of the Maximum Score Estimator. *Journal of Econometrics* 27, 313-333.

[8] Manski, C. F. and T. S. Thompson (1986) Operational Characteristics of Maximum Score Estimation. *Journal of Econometrics* 32, 85-108.

[9] Pollard D. (1993) The Asymptotics of a Binary Choice Model. Yale University, working paper.

[10] Zinde-Walsh, V. (2002) Asymptotic Theory for some High Breakdown Point Estimators. *Econometric Theory* 18, 1172-1196.

Table I. Smooth homoskedastic model.

size	estimator	bias	variance	MSE
2000	f4	-0.001	0.0015	0.0015
	comb4	-0.001	0.0018	0.0018
	f2	-0.001	0.0016	0.0016
	comb24	-0.003	0.0016	0.0016
	f3a	-0.003	0.0038	0.0038
	f3b	0.005	0.0038	0.0039
	comb33	-0.001	0.0020	0.0020
	g3a	-0.002	0.0059	0.0059
	g3b	-0.005	0.0056	0.0056
	g4	0.004	0.0033	0.0033
	comb334	-0.001	0.0019	0.0019
4000	f4	-0.001	0.0008	0.0008
	comb4	-0.001	0.0012	0.0012
	f2	-0.002	0.0013	0.0013
	comb24	-0.002	0.0011	0.0011
	f3a	-0.003	0.0025	0.0025
	f3b	0.005	0.0026	0.0026
	comb33	0.000	0.0014	0.0014
	g3a	-0.003	0.0075	0.0075
	g3b	-0.004	0.0035	0.0036
	g4	0.003	0.0018	0.0018
	comb334	0.001	0.0013	0.0013
8000	f4	-0.0014	0.00041	0.00041
	comb4	-0.0010	0.00068	0.00068
	f2	-0.0014	0.00057	0.00057
	comb24	-0.0014	0.00062	0.00062
	f3a	-0.0029	0.00138	0.00139
	f3b	0.0044	0.00121	0.00123
	comb33	-0.0005	0.00077	0.00077
	g3a	0.0008	0.00143	0.00143
	g3b	-0.0042	0.00234	0.00235
	g4	0.0017	0.00104	0.00104
	comb334	0.0002	0.00074	0.00074

Table II. Smooth heteroskedastic model.

size	estimator	bias	variance	MSE
2000	f4	-0.005	0.00024	0.00026
	comb4	-0.003	0.00020	0.00021
	f2	-0.004	0.00027	0.00028
	comb24	-0.004	0.00019	0.00020
	f3a	-0.020	0.00038	0.00077
	f3b	0.023	0.00040	0.00094
	comb33	0.001	0.00023	0.00023
	g3a	0.032	0.00093	0.00196
	g3b	-0.028	0.00076	0.00157
	g4	0.002	0.00028	0.00029
	comb334	-0.001	0.00023	0.00023
4000	f4	-0.005	0.00019	0.00022
	comb4	-0.003	0.00012	0.00012
	f2	-0.004	0.00010	0.00011
	comb24	-0.003	0.00010	0.00011
	f3a	-0.017	0.00020	0.00048
	f3b	0.020	0.00021	0.00061
	comb33	0.001	0.00014	0.00014
	g3a	0.027	0.00045	0.00118
	g3b	-0.023	0.00039	0.00093
	g4	0.002	0.00016	0.00016
	comb334	-0.000	0.00013	0.00013
8000	f4	-0.003	6.3e-5	7.4e-5
	comb4	-0.002	6.0e-5	6.4e-5
	f2	-0.003	4.9e-5	5.7e-5
	comb24	-0.002	5.4e-5	5.7e-5
	f3a	-0.015	10.9e-5	32.3e-5
	f3b	0.017	10.8e-5	38.9e-5
	comb33	-0.000	7.7e-5	7.7e-5
	g3a	0.024	18.8e-5	74.1e-5
	g3b	-0.019	20.1e-5	57.7e-5
	g4	0.002	6.9e-5	7.2e-5
	comb334	-0.000	7.3e-5	7.3e-5

Table III. Non-smooth homoskedastic model.

size	estimator	bias	variance	MSE
2000	f4	0.048	0.0059	0.0082
	comb4	0.043	0.0063	0.0082
	f2	0.053	0.0057	0.0085
	comb24	0.043	0.0055	0.0074
	f3a	0.079	0.0168	0.0231
	f3b	0.030	0.0056	0.0065
	comb33	0.046	0.0063	0.0084
	g3a	0.040	0.0089	0.0105
	g3b	0.109	0.0177	0.0296
	g4	0.033	0.0096	0.0107
	comb334	0.045	0.0058	0.0078
4000	f4	0.037	0.0033	0.0046
	comb4	0.034	0.0040	0.0052
	f2	0.045	0.0030	0.0051
	comb24	0.033	0.0036	0.0047
	f3a	0.076	0.0128	0.0186
	f3b	0.023	0.0031	0.0037
	comb33	0.035	0.0038	0.0050
	g3a	0.034	0.0033	0.0045
	g3b	0.111	0.0126	0.0250
	g4	0.020	0.0058	0.0062
	comb334	0.034	0.0037	0.0048
8000	f4	0.030	0.0019	0.0028
	comb4	0.024	0.0024	0.0030
	f2	0.037	0.0018	0.0032
	comb24	0.023	0.0022	0.0027
	f3a	0.067	0.0097	0.0142
	f3b	0.023	0.0022	0.0027
	comb33	0.025	0.0023	0.0029
	g3a	0.031	0.0037	0.0046
	g3b	0.104	0.0098	0.0205
	g4	0.012	0.0033	0.0035
	comb334	0.025	0.0023	0.0029

Table IV. Non-smooth heteroskedastic model.

size	estimator	bias	variance	MSE
2000	f4	0.034	0.00039	0.00156
	comb4	0.028	0.00047	0.00127
	f2	0.037	0.00045	0.00180
	comb24	0.030	0.00040	0.00132
	f3a	0.008	0.00216	0.00223
	f3b	0.031	0.00058	0.00155
	comb33	0.023	0.00075	0.00129
	g3a	0.041	0.00065	0.00235
	g3b	0.017	0.00339	0.00368
	g4	0.019	0.00076	0.00110
	comb334	0.027	0.00059	0.00131
4000	f4	0.033	0.00024	0.00130
	comb4	0.024	0.00033	0.00090
	f2	0.034	0.00024	0.00141
	comb24	0.026	0.00025	0.00091
	f3a	0.012	0.00133	0.00148
	f3b	0.025	0.00026	0.00087
	comb33	0.020	0.00048	0.00089
	g3a	0.033	0.00026	0.00136
	g3b	0.029	0.00177	0.00260
	g4	0.016	0.00041	0.00068
	comb334	0.023	0.00034	0.00088
8000	f4	0.030	0.00012	0.00103
	comb4	0.018	0.00023	0.00054
	f2	0.031	0.00013	0.00111
	comb24	0.020	0.00016	0.00057
	f3a	0.016	0.00086	0.00110
	f3b	0.019	0.00014	0.00050
	comb33	0.017	0.00031	0.00061
	g3a	0.028	0.00012	0.00089
	g3b	0.037	0.00097	0.00235
	g4	0.012	0.00023	0.00037
	comb334	0.019	0.00022	0.00057