Non and semi-parametric estimation in models with unknown smoothness*

Yulia Kotlyarova
Dalhousie University

Victoria Zinde-Walsh
Department of Economics, McGill University
855 Sherbrooke Street West,
Montreal, Quebec, Canada, H3A 2T7
tel. (514) 398 4834; fax (514) 398 4839
victoria.zinde-walsh@mcgill.ca

April 28, 2006

Abstract

Many asymptotic results for kernel-based estimators were established under some smoothness assumption on density. For cases where smoothness assumptions that are used to derive unbiasedness or asymptotic rate may not hold we propose a combined estimator that could lead to the best available rate without knowledge of density smoothness. A Monte Carlo example confirms good performance of the combined estimator.

JEL code C14

Key words: nonparametric estimation, combined estimator

*Support of Social Sciences and Humanities Research Council of Canada (SSHRC) and of *Fonds québecois de la recherche sur la société et la culture* (FRQSC) is acknowledged. We thank Marcia Schafgans for valuable comments.

1 Introduction

Asymptotic results for various kernel-based estimators are established under smoothness assumptions on density which may be excessively strong; for example, density of disposable income with lump-sum transfers or changes in tax rates is discontinuous. For situations when smoothness assumptions that provide optimal rate may not hold we propose an estimator that alleviates the negative consequences of incorrect assumptions about smoothness.

The approach utilizes the joint asymptotic distribution of several estimators. The joint distribution indicates that complimentary information may be supplied by estimators for different bandwidth/kernel combinations, e.g. different estimators could be asymptotically independent. A linear combination may have a smaller mean squared error (MSE) than an individual estimator; this may be particularly useful when sharp asymptotic results based on smoothness assumptions are not available leading to uncertainty about bandwidth rate. "Combined estimator" minimizes the estimated asymptotic MSE (AMSE) over linear combinations and may achieve the best rate automatically without knowledge of smoothness. Kotlyarova, Zinde-Walsh (2004) obtained the joint asymptotic distribution of Smoothed Maximum Score (SMS) estimators in the binary choice model and provided estimators for biases and variances; similar results were obtained for kernel density estimation (Kotlyarova, 2005).

While consistent estimators of biases and variances help obtain an optimal combined estimator, they may be difficult to construct. We indicate ways of constructing estimators of the asymptotic biases and variances that do not rely on knowledge of density smoothness. Examples demonstrate that even when we cannot estimate all the biases consistently the combined estimator may be better in terms of convergence rates and finite sample performance than an estimator based on incorrect assumptions about smoothness. Monte Carlo experiments demonstrate good efficiency /robustness of the combined estimator under various models; here we produce simulation results for the binary choice model.

2 Linear combinations of estimators

2.1 Notation and assumptions

We consider several kernel estimators for a $k \times 1$ parameter vector β, based on different bandwidths h_{nt}, $t = 1, ...m$, and kernel functions $K_s, s = 1, ..., l$. Denote each estimator $b_n(t,s)$ (n - sample size). Depending on the combination of bandwidth and kernel convergence rates may differ and may lead to asymptotic bias (abias) or to Gaussian limit process with or without abias.

Assumption 1. For $b_n(t,s)$ A1(1), A1(2) or A1(3) below holds as $n \to \infty$.

A1(1). There exist a sequence of positive constants $r_n(t,s) \to \infty$, constant $k \times 1$ non-zero vector $A(t,s)$ such that

$$r_n(t,s)(b_n(t,s) - \beta) - A(t,s) \to_p 0;$$

A1(2). There exist a sequence of positive constants $r_n(t,s) \to \infty$, constant $k \times 1$ non-zero vector $A(t,s)$ and constant non-zero $k \times k$ matrix $\Gamma(t,s)$ (non-negative definite) such that for some $k \times 1$ vectors $\tilde{A}_n(t,s)$ and $k \times k$ (positive definite) $\tilde{\Gamma}_n(t,s)$

$$\begin{aligned} r_n(t,s)^2 \tilde{\Gamma}_n(t,s) &\to {}_p\Gamma(t,s); \\ r_n(t,s)\tilde{A}_n(t,s) &\to {}_pA(t,s); \\ \tilde{\Gamma}_n(t,s)^{-1/2}(b_n(t,s) - \beta - \tilde{A}_n(t,s)) &\to {}_dN(0,I) \end{aligned}$$

thus

$$r_n(t,s)(b_n(t,s) - \beta) \to_d N(A(t,s), \Gamma(t,s));$$

A1(3). A1(2) holds with additionally $\tilde{A}_n(t,s) = 0$.

Case A1(1) corresponds to abias, A1(2) to limit Gaussian process with abias, A1(3) to limit Gaussian process without abias.

The Assumption is satisfied by many estimators even though most of the results in the literature focus on bandwidths that get rid of abias (case A1(3)) or on optimal bandwidths (case A1(2)). See, e.g. Pagan, Ullah (1999) for various estimators; Horowitz (1992) discusses all A1(1-3) cases for the SMS estimator.

Next, define a partial order on rates. For pairs (t,s) if as $n \to \infty$ the ratio $\frac{r_n(t_1,s_1)}{r_n(t_2,s_2)} \to \infty$, write $r_n(t_1,s_1) \succ r_n(t_2,s_2)$ ($r_n(t_1,s_1) \not\cong r_n(t_2,s_2)$); if for constants d, D the ratio satisfies $0 < d < \frac{r_n(t_1,s_1)}{r_n(t_2,s_2)} < D < \infty$ for all n, write $r_n(t_1, s_1) \cong r_n(t_2, s_2)$; write $r_n(t_1, s_1) \succeq r_n(t_2, s_2)$ if either $r_n(t_1, s_1) \succ r_n(t_2, s_2)$ or $r_n(t_1, s_1) \cong r_n(t_2, s_2)$. Let $j = 1, 2, ..., J$; $J = l \times m$ correspond to an ordering of the rates:

$$r_n(t_1, s_1) \succeq r_n(t_2, s_2) \succeq ... \succeq r_n(t_J, s_J);$$

since some rates may be equivalent this ordering is not necessarily unique.

We refer to $r_n(t_j, s_j)$ as r_j, to $b_n(t_j, s_j)$ as b_j, and similarly A_j (which is zero in case A1(3)) and Γ_{jj} (which is zero in case A1(1)).

Consider the joint distribution of estimators b_j.

Assumption 2. In addition to Assumption 1 the joint limit process for vector $\begin{pmatrix} r_1(b_1 - \beta) \\ \vdots \\ r_J(b_J - \beta) \end{pmatrix}$ is Gaussian

$$N\left(\begin{pmatrix} A_1 \\ \vdots \\ A_J \end{pmatrix}, \begin{pmatrix} \Gamma_{11} & \cdots & \Gamma_{1J} \\ \vdots & \ddots & \vdots \\ \Gamma_{1J} & \cdots & \Gamma_{JJ} \end{pmatrix}\right),$$

with Γ_{ij} the limit matrix of covariances for $r_i(b_i - \beta)$,$r_j(b_j - \beta)$; if $r_i \ncong r_j$ or if for at least one of i or j the case A1(1) holds the corresponding $\Gamma_{ij} = 0$; the limit distribution may be degenerate.

The joint limit distribution of Assumption 2 is given for kernel estimators of continuous density (not necessarily smooth) in Kotlyarova (2005), for SMS estimator in the binary choice model with less restrictive assumptions on smoothness than in Horowitz, 1992 (Hölder continuous rather than twice differentiable conditional density) in Kotlyarova, Zinde-Walsh (2004). Zinde-Walsh (2002) derived joint limit process for smoothed least median of squares estimator (LMS) for different kernels (same bandwidth).

2.2 Asymptotic distribution of a linear combination of estimators.

Denote by b_n the stacked vector of estimators $(b_1'(n), ..., b_J'(n))$. Denote by a_n a $J \times 1$ weights vector that satisfies $a_n' \iota_J = 1$; ι_J is a $J \times 1$ vector of ones. Consider the linear combination $\Sigma a_{nj} b_j(n)$ of estimators corresponding to different bandwidth/kernel pairs; the sum of the weights is one (the weights are not necessarily non-negative).

Denote $\tilde{A}(n)$ the $Jk \times 1$ vector $\tilde{A} = (\tilde{A}_1', ..., \tilde{A}_J')'$ with $k \times 1$ subvector $\tilde{A}_j$

$$\tilde{A}_j = r_j^{-1} A_j$$

Denote $\tilde{\Gamma}(n)$ the matrix

$$\tilde{\Gamma}(n) = \begin{pmatrix} \tilde{\Gamma}_{11} & \cdots & \tilde{\Gamma}_{1J} \\ \vdots & \ddots & \vdots \\ \tilde{\Gamma}_{1J} & \cdots & \tilde{\Gamma}_{JJ} \end{pmatrix},$$

with $k \times k$ components

$$\tilde{\Gamma}_{ij} = \begin{cases} (r_j)^{-2} \Gamma_{jj} & \text{if } i = j; \\ (r_i)^{-1} (r_j)^{-1} \Gamma_{ij} & \text{otherwise.} \end{cases}$$

Consider for $\overline{a} : \overline{a}' \iota_J = 1$ a vector ζ_n distributed as $N(\Sigma \overline{a}_j \tilde{A}_j, \Sigma_{i,j} \overline{a}_i \overline{a}_j \tilde{\Gamma}_{ij})$.

Theorem 1 *Under Assumptions 1, 2 for any (possibly random) sequence $a_n \to \overline{a}$ (possibly $a_n \to_p \overline{a}$), $\overline{a}$ a constant vector, the limit process for $\Sigma a_{nj}(b_j(n) - \beta)$ is the same as for ζ_n.*

Proof. For $b_j(n)$ by convergence of a_n

$$(a_{nj} - \overline{a}_j)(b_j(n) - \beta) = o_p(r_j^{-1}),$$

thus $a_{nj} b_j(n)$ and $\overline{a}_j b_j(n)$ have the same limit process; $\Sigma a_{nj}(b_j(n) - \beta)$ has the same limit process as $\Sigma \overline{a}_j (b_j(n) - \beta)$: Gaussian process with the same asymptotic mean and covariance as ζ_n.■

Asymptotic mean square error for the linear combination is

$$AMSE(\Sigma a_{nj} b_j(n)) = \sum_{i,j} a_i a_j (\tilde{A}_i' \tilde{A}_j + tr \tilde{\Gamma}_{ij}) = a' D a; \quad (1)$$

$$\{D\}_{ij} = \tilde{A}_i' \tilde{A}_j + tr \tilde{\Gamma}_{ij}. \quad (2)$$

From Assumption 2

$$AMSE(\Sigma a_{nj} b_j(n)) = AMSE^I + AMSE^{II} = a_I' D^I a_I + a_{II}' D^{II} a_{II} \quad (3)$$

where $a = (a_I', a_{II}')'$, D^{II} corresponds to estimators of A1(1) (abias); $rank D^{II} = 1$, $\dim D^{II} \geq 2$; while D^I is invertible (and may contain at most one A1(1) estimator).

Under uncertainty about model smoothness (thus about the best rate/bandwidth) combining estimators can at least reduce the worst possible AMSE from an incorrect choice. Indeed, suppose that an estimator b_1 has associated rate r_1; b_2 has a slower rate r_2 ($r_1 \succ r_2$). If this were known then we would use b_1. If there is uncertainty about the degree of smoothness, a linear combination, e.g. $a'b = \frac{1}{2}b_1 + \frac{1}{2}b_2$ while converging slower than b_1 at least halves the AMSE associated with the incorrect choice, b_2.

Next we consider optimizing the choice of weights.

3 Combined estimator and its asymptotic distribution

First, we find a that minimizes the $AMSE$ in (1) subject to $a'\iota = 1$ for known $\tilde{A}$ and $\tilde{\Gamma}$.

Recall that while the rates, r_j, are ordered from fastest to slowest, some may be strictly faster than others but some may be equivalent; correspondingly, partition $\{1, 2, ..., J\}$ as $\{E_1, ..., E_V\}$ with $E_v = \{j_{v-1} + 1, ..., j_v\}$, $v = 1, ...V$: such that $r_{j_{\nu-1}+1} \cong r_{j_\nu}$ and if $j_1 \in E_{v_1}, j_2 \in E_{v_2}$ for $v_1 > v_2$ then $r_{j_i} \succ r_{j_2}$. Partition D correspondingly and denote by D_{11} the top left submatrix of D corresponding to E_1 (fastest rate).

Define

$$a_n = \arg\min_a AMSE(\Sigma a_j b_j(n)) \text{ subject to } a'\iota = 1 \tag{4}$$

Theorem 2 *Suppose Assumptions 1, 2 hold; $n \to \infty$.*

(a) If $D \equiv D^I, a = a^I_{\lim} + o(1)$, where

$$a^I_{\lim} = \left(\left(\frac{1}{\iota' D_{11}^{-1} \iota} D_{11}^{-1} \iota \right)', 0, ...0 \right)'. \tag{5}$$

(b) If $D \equiv D^{II}$, then $\left\| a_n - a^{II}_{\lim} \right\| = o(1)$, where $\{II\}$ consists of vectors $a^{II}_{\lim}$ satisfying

$$a^{II\prime}_{\lim} D^{II} a^{II}_{\lim} = 0; a^{II\prime}_{\lim} \iota = 1. \tag{6}$$

Proof.

(a) Lagrangian for the AMSE minimization is

$$a'Da + \lambda(a'\iota_J - 1).$$

First-order conditions are

$$\begin{aligned} 2Da + \lambda \iota_J &= 0; \\ a'\iota_J - 1 &= 0. \end{aligned}$$

Then a satisfies

$$\begin{aligned} a &= -\lambda (2D)^{-1}\iota_J; \\ a'\iota_J &= -\lambda \iota_J'(2D)^{-1}\iota_J = 1, \end{aligned}$$

providing

$$a = \left(\iota_J' D^{-1}\iota_J\right)^{-1} D^{-1}\iota_J. \tag{7}$$

In the block-matrix D diagonal blocks $D_{vv} \approx O(r_{j_{v-1}+1}^{-2}) = o(r_{j_v+1}^{-2})$; first block goes to zero fastest, each successive block converges slower; off-diagonal $D_{vw} = o(r_{j_{v-1}+1}^{-1} r_{j_{w-1}+1}^{-1})$ and go to zero faster than the slower of the corresponding diagonal blocks. For simplicity partition D as

$$\begin{pmatrix} D_{11} & D_{12} \\ D_{12}' & D_{22} \end{pmatrix}.$$

Rate structure of D is

$$\begin{pmatrix} O(r_1^{-2}) & o(r_1^{-1} r_2^{-1}) \\ o(r_1^{-1} r_2^{-1}) & O(r_2^{-2}) \end{pmatrix}$$

with $r_2 = o(r_1)$.The partitioned inverse provides

$$D^{-1} = \begin{pmatrix} D_{11}^{-1} + o(r_1^2) & O(r_2 r_1) \\ O(r_2 r_1) & O(r_2^2) \end{pmatrix}$$

and $D^{-1} = \begin{pmatrix} D_{11}^{-1} & 0 \\ 0 & 0 \end{pmatrix} + o(r_1^2)$. Substitution into (7) concludes (a).

(b) The minimized value of the limit AMSE is zero in this case. Since $rankD^{II} = 1 < \dim D^{II}$ there exists a (possibly non-unique) $a_{\lim}^{II}$ such that $a_{\lim}^{II\prime} D^{II} a_{\lim}^{II} = 0; a_{\lim}^{II\prime}\iota = 1$. The minimizers of the quadratic form will get close to a vector that provides a zero limit asymptotically. ■

Remarks. In case (a) $a_{\lim}$ has non-negative components. For (b) the limit linear combination converges at a faster rate, $r_{(b)}$, than the best of the biased estimators; . If (3) holds non-trivially consider correspondingly

$$a = \alpha(a_{\lim}^{I\prime}, 0')' + (1-\alpha)(0', a_{\lim}^{II\prime})';$$

$0 \leq \alpha \leq 1$. If $r_{(b)}$ is better than the best rate for the D^I part then $\alpha = 0$, if worse, $\alpha = 1$, otherwise there could be an optimal combination involving both undersmoothed and oversmoothed estimators.

Examples.

(a) $E_1 = \{1\}$. Then $a_{\lim} = (1, 0...0)'$. In the optimal linear combination the weight on fastest convergent estimator is 1, the others get zero weights.

(b) $E_1 = \{1, 2\}$. Then if the estimators are undersmoothed $a_{\lim} = (a, 1 - a, 0, ...0)'$. For e.g. b_1 undersmoothed (no abias, possibly substantial variance), b_2 oversmoothed (zero avariance, but abias), $\lim \frac{r_2}{r_1} = d > 1$

$$r_1(a'_{\lim} b - \beta) \to_d N(\frac{(1-a)}{d} A_2, a^2 \Gamma_1).$$

Linear combination has abias reduced relative to that of b_2 and avariance smaller than for b_1.

Typically we do not know the abiases and variances; estimated biases and variances could be substituted to obtain estimated AMSE for linear combination of estimators: $\widehat{AMSE(a'b_n)}$. Call $b_{comb} = \Sigma a_{combj} b_j(n)$ with

$$a_{comb} = \arg\min_{a_n} \widehat{AMSE(\Sigma a_{nj}} b_j(n))$$

the combined estimator.

Assumption 3. Consistent estimators $\hat{A}(n)$ and $\hat{\Gamma}(n)$ for $\tilde{A}(n)$ and $\tilde{\Gamma}(n)$ are available; so for $\left\{\hat{D}\right\}_{ij} = \hat{A}'_i \hat{A}_j + tr\hat{\Gamma}_{ij}$ and $W = diag(r_1, ..., r_J)$

$$W(\hat{D} - D)W \to_p 0.$$

Theorem 3 *Under the conditions of Theorem 2 and Assumption 3 for* $\widehat{AMSE(a'b_n)} = a'\hat{D}a$

$$a_{comb} = a_{\lim} + o(1).$$

Proof. Follows by modifying the proof of Theorem 2.

The Theorem states that a combined estimator that minimizes AMSE based on consistent estimators of biases and variances has the same limit weights as the combination that minimizes the (true) AMSE.

4 Estimation issues and performance

When the degree of smoothness is not known consistent estimators for abias and avariance that do not rely on such knowledge are required. It is not difficult to estimate the asymptotic variance consistently, e.g. by bootstrap.

Estimation of bias is more difficult. For various estimators as bandwidth increases variance declines, but bias increases. If b_i is an undersmoothed estimator its abias is zero and

$$b_i - \beta = O_p(r_i^{-1}).$$

For some b_j that corresponds to a higher (possibly oversmoothed) bandwidth

$$b_j - \beta - \tilde{A}_j = o_p(r_i^{-1}),$$

because $\text{avar}\left(b_j - \beta - \tilde{A}_j\right) < avar(b_i - \beta)$. So

$$b_j - b_i - \tilde{A}_j = O_p(r_i^{-1}).$$

Then if $r_j \prec r_i$ a consistent estimator of $\tilde{A}_j$ (abias of b_j) is $b_j - b_i$. To improve this estimator consider L undersmoothed estimators and estimate $\tilde{A}_j$ by $b_j - \frac{1}{L}\sum_{i=1}^{L} b_i$ (bootstrap could be applied to this estimator). This method was used by Kotlyarova (2005) in combined estimator for density and for SMS.

Consider the model from Horowitz (1992):

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

$x_1 \sim N(1,1); x_2 \sim N(0,1)$; with heteroscedastic $u = .25[1 + (x_1 + x_2)^2]^2 v$; and v logistic (median=0; variance=0.5). This is the smooth case; we add a non-smooth case where v is piece-wise linear. We normalize the estimator $b = (b_1, b_2)'$ by $\|b\| = 1$; $\beta_1 = \beta_2 = \frac{1}{\sqrt{2}}$. We use the Horowitz-optimal kernel/bandwidth f_4/h_0 (with bias correction) and consider several combined estimators (no correction). For selection of bandwidths a starting point is h_0 optimal for the smooth case; in non-smooth cases it is oversmoothing, thus other bandwidths are percentiles of h_0. Polynomial kernels are used in combinations: *comb*4 uses fourth order f_4; *comb*24 uses f_4 and second-order f_2; *comb*334 combines f_4 with two (orthogonal) third order kernels; each kernel at four bandwidths.

Monte Carlo results for MSE of these estimators in the following table demonstrate good performance of the combined estimator.

MSE of estimator.

n	Estimator	Smooth model	Non-smooth
2000	f_4/h_0	.00026	.00156
	$comb4$	.00021	.00127
	$comb24$	.00020	.00132
	$comb334$	.00023	.00131
4000	f_4/h_0	.00022	.00130
	$comb4$	.00012	.00090
	$comb24$	.00011	.00091
	$comb334$	.00013	.00088

References

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

[2] Kotlyarova, Y. and V. Zinde-Walsh (2004) Improving the Efficiency and Robustness of the Smoothed Maximum Score Estimator, McGill University, Department of Economics working paper

[3] Kotlyarova, Y. (2005) Kernel Estimators: Testing and Bandwidth Selection in Models of Unknown Smoothness, Ph.D. thesis, McGill University.

[4] Pagan, A. and A. Ullah (1999). *Nonparametric Econometrics*, Cambridge University Press.

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