A derivative based estimator for semiparametric index models

Bas Donkers* Marcia Schafgans†

Erasmus University Rotterdam London School of Economics

Econometric Institute Report 2003-08

Abstract

This paper proposes a semiparametric estimator for single- and multiple index models. It provides an extension of the average derivative estimator to the multiple index model setting. The estimator uses the average of the outer product of derivatives and is shown to be root-N consistent and asymptotically normal. Unlike the average derivative estimator, our estimator still works in the single-index setting when the expected derivative is zero (symmetry). Compared to other estimators for multiple index models, the proposed estimator has the advantage of ease of computation. While many econometric models can be regarded as multiple index models with known number of indices, our estimator in addition provides for a natural framework within which to test for the number of indices required.

KEYWORDS: Semiparametric estimation, index models, average derivatives, outer product of derivatives, rank testing.

*Econometric Institute and Department of Marketing, Erasmus University Rotterdam, P.O. Box 1738, NL-3000 DR Rotterdam, The Netherlands. Bas Donkers enjoys the hospitality of CentER as Extramural Fellow.

†Department of Economics, London School of Economics, Houghton Street, London WC2A 2AE, United Kingdom. Most of the work was done during Marcia Schafgans' sabbatical leave at CentER, and she would like to thank CentER and Tilburg University for their hospitality.

For insightful discussions and helpful comments we would like to thank Oliver Linton, Bertrand Melenberg, Peter Robinson and Arthur van Soest.

1 Introduction

We consider the multiple index mean regression model

$$E(y|x) = H(x^T\beta_1, .., x^T\beta_P), \tag{1}$$

with dependent variable y and explanatory variables x. H is an unknown, but sufficiently smooth function, and $\beta = (\beta_1^T, ..., \beta_P^T)^T$ is the vector of unknown parameters. Many econometric models can be regarded this way, for example, binomial and multinomial choice model, sample selection model, and disequilibrium model. Various $\sqrt{N}$ consistent asymptotically normal estimators of β for the multiple index model have been proposed, among others, by Ichimura and Lee (1991), Lee (1995), and Picone and Butler (2000). Since the multiple index model provides a general and flexible modelling strategy, one would expect to see numerous applications of the multiple index model given the existence of these estimation methods. A simple explanation for the absence of these applications could be that these advantages are offset by the computational complexity of the proposed methods. The advantage of our estimator is the ease of computation. Another advantage of the proposed estimator is that it provides a natural framework within which to test for the number of indices required.

Let $g(x) = E(y|x)$.[1] The derivative of this unknown function, $g'(x)$, by application of the chain rule of differentiation, is a weighted average of the true coefficients β_p

$$\frac{\partial g(x)}{\partial x} = \sum_{p=1}^{P} \left(\frac{\partial H}{\partial (x^T\beta_p)} \right) \beta_p. \tag{2}$$

For single index models this property is sufficient to identify the parameters "up-to-scale". Properties of the average derivative estimator (henceforth ADE) are given by Powell, Stock and Stoker (1989), Robinson (1989), Härdle and Stoker (1989) and Stoker (1991). For multiple index models the average derivative does not provide enough information to identify β_p, $p = 1, .., P$ "up-to-scale" unless the indices have no variables in common.

[1] To clarify our use of notation, $'$ denotes the derivative of a function with respect to its argument, typically the vector x, $''$ denotes the matrix of second order derivatives, and ^{T} denotes the transposed of a vector.

The estimator we propose uses the average of the outer product of derivatives instead to remedy this identification problem. The average of the outer product of derivatives (gradient) is of practical interest because it provides us with the following moment conditions for the parameters of interest

$$E(g'(x)g'(x)^T) = \sum_{p,q} E\left[\frac{\partial H}{\partial(x^T\beta_p)}\frac{\partial H}{\partial(x^T\beta_q)}\right]\beta_p\beta_q^T. \tag{3}$$

To facilitate the derivation of the asymptotic distribution for the outer product of the gradient we use a weighted version of these moment conditions instead

$$\begin{aligned} E(\omega(x)g'(x)g'(x)^T) &= \sum_{p,q} E\left[\omega(x)\frac{\partial H}{\partial(x^T\beta_p)}\frac{\partial H}{\partial(x^T\beta_q)}\right]\beta_p\beta_q^T \\ &= \sum_{p,q}\gamma_{\omega,pq}\beta_p\beta_q^T \text{ with } \gamma_{\omega,pq} = \gamma_{\omega,qp}, \end{aligned} \tag{4}$$

where we define $\gamma_{\omega,pq} = E[\omega(x)\frac{\partial H}{\partial(x^T\beta_p)}\frac{\partial H}{\partial(x^T\beta_q)}]$. These moment conditions are used within a GMM framework to estimate β_p, $p = 1,..,P$ and the auxiliary parameters γ_ω given typical exclusion and normalization restrictions. Under these restrictions, the parameters β_p, $p = 1,..,P$ are identified "up-to-scale". An alternative set of restrictions on H is proposed as well such that exclusion restrictions can be avoided.

The estimator has the added benefit that, unlike the ADE estimator, it also works for single index models when the expected derivative of the unknown function is zero (say, g is an even function and X is symmetrically distributed). Also in less extreme situations the ADE estimator does not use all available information, since the contribution of observations with negative derivative of the unknown function is (at least partially) cancelled out by the contribution of observations with a positive derivative. Our estimator is based on the squared derivative, hence this averaging out of the local derivatives does not occur. To enhance the efficiency of our estimator, nonetheless, one could add the moment restrictions that are used by ADE. This would guarantee a more efficient estimator than the ADE. As this extension is conceptually straightforward, we do not pursue this extension in detail.

To estimate the average outerproduct of the gradient in (4), we implement a kernel based nonparametric estimator of $g'(x)$ and replace the expectation with the sample average.

After imposing identifying restrictions, estimates of β and γ_ω are obtained by applying the GMM approach to the moment conditions that hold for the average outer product of the gradient. The asymptotic theory for our estimator of β ($\hat{\beta}$) and γ_ω ($\hat{\gamma}_\omega$) is then obtained as follows. First, we show that the nonparametric estimator of the outer product of the gradient is $\sqrt{N}$ consistent and asymptotically normal. Second, we apply the generalized method of moments (GMM) framework to show that $\hat{\beta}$ and $\hat{\gamma}_\omega$ are also $\sqrt{N}$ consistent and asymptotically normal.

The multiple index model is also treated in the statistical literature, where it is interpreted as a regression-type model for dimension reduction that can be used to overcome the "curse of dimensionality" (P is smaller than the dimension of X). Xia et al. (2002) and Hristache et al. (2001) show that the effective dimension reduction (EDR) directions can be estimated at the parametric rate $\sqrt{N}$, using a computationally demanding estimation procedure. However, they do not develop the asymptotic theory for the estimated parameters. Moreover, no statistical test is provided to determine the appropriate number of EDR directions. In this paper, we show that such a test can be constructed by using a test of the rank of the outer product of derivatives. For this, we can make use of existing tests for the rank of a matrix by Cragg and Donald (1996, 1997) and Robin and Smith (2000). Alternatively, a test for the number of indices required can be based on a test for overidentifying restrictions within the GMM framework that is used to estimate β.

The paper is organized as follows. In Section 2, we give a $\sqrt{N}$ consistent, asymptotically normal estimator for a weighted version of the outer product of the gradient. In Section 3, we show how GMM yields a $\sqrt{N}$ consistent, asymptotically normal estimator for the parameters of interest on the basis of the estimated outer product of the gradient. In Section 4, we discuss a test for the number of indices (EDR directions). In addition to theoretical results, we provide some simulations to illustrate the estimator's usefulness in Section 5. Section 6 concludes. An Appendix contains proofs of the technical lemmas stated in Section 2 and the proof of consistency of the estimator of the covariance matrix.

2 Kernel estimation: outer product of derivatives

Let the observed data $z_i = (y_i, x_i^T)^T$ $i = 1, ..., N$ constitute a random sample from a distribution with density $f^*(y, x)$, y is an endogenous variable and x is a k dimensional vector of explanatory variables. Let $f(x)$ denote the marginal density of x, and $f'(x)$ its derivative. Let $G(x)$ denote the function $\int y f^*(y, x) dy$, then $g(x) = G(x)/f(x)$. The regression derivative, $g'(x)$, can be expressed as

$$g'(x) = \frac{G'(x)}{f(x)} - \frac{G(x)f'(x)}{f(x)^2}. \tag{5}$$

Our interest is in estimating the average outer product of derivatives $E(g'(x)g'(x)^T)$ (or a weighted version thereof).

A natural estimator for the average outer product of derivatives is given by its sample analogue, which uses nonparametric kernel regression estimates of the density of x, its derivative $f'(x)$, $G(x)$ and $G'(x)$:

$$\frac{1}{N}\sum_{i=1}^{N}\left(\frac{\hat{G}'(x_i)}{\hat{f}(x_i)} - \frac{\hat{G}(x_i)\hat{f}'(x_i)}{\hat{f}^2(x_i)}\right)\left(\frac{\hat{G}'(x_i)}{\hat{f}(x_i)} - \frac{\hat{G}(x_i)\hat{f}'(x_i)}{\hat{f}^2(x_i)}\right)^T. \tag{6}$$

Specifically, we use the Nadaraya-Watson leave-one-out kernel estimators

$$\begin{aligned} \hat{f}(x_i) &= \tfrac{1}{(N-1)h^k}\sum_{j=1,j\neq i}^{N} K\left(\tfrac{x_i - x_j}{h}\right); \qquad \hat{f}'(x_i) = \tfrac{1}{(N-1)h^{k+1}}\sum_{j=1,j\neq i}^{N} K'\left(\tfrac{x_i - x_j}{h}\right); \\ \hat{G}(x_i) &= \tfrac{1}{(N-1)h^k}\sum_{j=1,j\neq i}^{N} K\left(\tfrac{x_i - x_j}{h}\right) y_j; \quad \hat{G}'(x) = \tfrac{1}{(N-1)h^{k+1}}\sum_{j=1,j\neq i}^{N} K'\left(\tfrac{x_i - x_j}{h}\right) y_j, \end{aligned} \tag{7}$$

where $K(\cdot)$ is a kernel function, h is the bandwidth parameter, and $h \to 0$ as $N \to \infty$.

Obviously, we will need to introduce some trimming function to down-weight observations for which $\hat{f}(x)$ is very small. Rather than using an indicator function $1(f(x) > b_N)$ where $b_N \to 0$ as $N \to \infty$, that is used in the proof of the asymptotics of the ADE in Härdle and Stoker (1989) and Stoker (1991),[2] we introduce a smoothed indicator function

[2]Powell, Stock and Stoker (1989) consider the density weighted ADE instead of using a trimming function. We decided against a similar approach due to the conditions we would need to impose when using the asymptotics of higher order $U-$statistics, see, e.g., Hoeffding (1948).

$s(f(x)-b)$ along the lines of Horowitz (1992), or

$$\omega(x) = s(f(x)-b) \equiv s_x. \tag{8}$$

We assume b is fixed with $b > 0$. Here b could be allowed to go to zero as $N \to \infty$ at the cost of imposing additional moment conditions. The need for a smoothed trimming function instead of an indicator trimming function is explained in detail in the Appendix directly following the proof of Lemma 2.

Define the weighted outer product matrix of derivatives as

$$M = E\left\{g'(x)g'(x)^T s_x\right\} \tag{9}$$

and its kernel based estimator

$$\hat{M} = \frac{1}{N}\sum_{i=1}^{N} \hat{g}'(x_i)\hat{g}'(x_i)^T \hat{s}_{x_i}, \tag{10}$$

with $\hat{s}_{x_i} = s(\hat{f}(x_i)-b)$. Let $\operatorname{vech} X$ stack the columns of the matrix X under each other to form a single column, where only that part of each column is included in $\operatorname{vech} X$ which is on or below the diagonal of X (Henderson and Searle (1979)). The property of this transformation is that it stacks all unique elements of a symmetric matrix, such as M.

We now state the assumptions required for our main result. The first assumption formalizes our use of independent and identically distributed observations and continuity of the regressors.

Assumption 1 *Let $z_i = (y_i, x_i^T)^T$, $i = 1,..,N$ be a random sample drawn from $f^*(y,x)$, with $f^*(y,x)$ the density of (y,x). The underlying measure of (y,x) can be written as $v_y \times v_x$, where v_x is Lebesque measure. Let $f(x)$ denote the density of x. The support Ω of f is a compact, convex, possibly unbounded subset of R^k with nonempty interior.*

Assumptions 2 and 3 formalize a number of continuity and differentiability requirements and Assumption 4 describes the kernel used.

Assumption 2 *$g(x) = E(y|x)$ is twice continuously differentiable and $M_2(x) = E(y^2|x)$ is continuous in x on Ω.*

Assumption 3 *Let $p = k + 3$. All partial derivatives of $f(x)$ and $G(x) = g(x)f(x)$ of order $p + 1$ exist on Ω.*

Assumption 4 *The kernel function $K(u)$ has bounded support $\{u : |u| \leq 1\}$, is symmetric, has p moments, and vanishes at the boundary. $K(u)$ is of order p, so with $(l_1, .., l_k)$ an index set*

$$\begin{aligned}
&\int K(u)du = 1, \\
&\int u_1^{l_1}...u_k^{l_k}K(u)du = 0 \qquad l_1 + ... + l_k < p \\
&\int u_1^{l_1}...u_k^{l_k}K(u)du \neq 0 \qquad l_1 + ... + l_k = p.
\end{aligned}$$

The next assumption is an assumption on the smoothness of the trimming function $s(\cdot)$

Assumption 5 *$s(\tau) : R \rightarrow R$ is a continuous, twice differentiable function with bounded second derivatives, such that*

$$\begin{aligned}
s(\tau) &= 0, \ \textit{for } \tau \leq 0 \\
&> 0, \ \textit{for } \tau > 0.
\end{aligned}$$

The function s is used to obtain a smooth trimming of the observations with low densities. Trimming will be based on the density of x, such that observations are weighted by $\hat{s}_x = s(\hat{f}(x) - b)$, $b > 0$. This guarantees that only a compact set of x with $\hat{f}(x) \geq b$ is considered for the estimation of M. In the sequel, superscript K (L) denotes the derivative with respect to the K-th (L-th) element of x, for instance $G^K(x) = \frac{\partial G(x)}{\partial x_K}$. Moreover, s'_x denotes $\partial s(\tau)/\partial\tau$ evaluated at $v = f(x) - b$ and s_x^K denotes $\partial s(f(x) - b)/\partial x_K = s'_x f^K(x)$.

Finally, we make two technical assumptions. Assumption 6 assures that U-statistics theory can be applied to each of the elements of the linearized version of M, while Assumption 7, in conjunction with the use of a higher order kernel, assures that our estimator is asymptotically unbiased. These assumptions are given next.

Assumption 6 *The functions $f(x)$ and $g(x)$, both $R^k \to R$ obey local Lipschitz conditions: For $v \in R^k$ in an open neighborhood of 0, there exist functions $\psi_f(x)$, $\psi_{f^K}(x)$, $\psi_g(x)$, $\psi_{g^K}(x)$ and $\psi_{g^{KL}}(x)$, $R^k \to R$, such that for $1 \leq K, L \leq k$*

$$|f(x+v) - f(x)| < \psi_f(x)\,|v|\,, \qquad \left|f^K(x+v) - f^K(x)\right| < \psi_{f^K}(x)\,|v|\,,$$
$$|g(x+v) - g(x)| < \psi_g(x)\,|v|\,, \qquad \left|g^K(x+v) - g^K(x)\right| < \psi_{g^K}(x)\,|v|$$
$$\left|g^{KL}(x+v) - g^{KL}(x)\right| < \psi_{g^{KL}}(x)\,|v|$$

By continuity one can derive the bounding $\psi-$functions for products of these functions, for example, $|gf(x+v) - gf(x)| < \psi_{gf}(x)\,|v|$ with $\psi_{gf} = \psi_g f + g\psi_f + \psi_g\psi_f$.

Assumption 7 *Let ι denote an index set $(l_1, \ldots, l_k)$, with $l_1 + ... + l_k = p$. For a vector $u = (u_1, \ldots, u_k)$ define $u^\iota = u_1^{l_1} \cdots u_k^{l_k}$. Then $f_\iota^{(p)} = \partial^p f/(\partial u)^\iota$ denotes the $p-$th order partial derivative of f. Similar definitions hold for $G_\iota^{K(p)}$, $G_\iota^{(p)}$, and $f_\iota^{K(p)}$. Local Hölder continuity holds for $G_\iota^{(p)}$, $f_\iota^{(p)}$, $G_\iota^{K(p)}$, and $f_\iota^{K(p)}$, so there exists $\delta > 0$ and functions $c_G(x)$, $c_f(x)$, $c_{G^K}(x)$ and $c_{f^K}(x)$, $R^k \to R$, such that for all $v \in R^k$ in an open neighborhood of 0,*

$$\left|G_\iota^{(p)}(x+v) - G_\iota^{(p)}(x)\right| \leq c_G(x)\,|v|^\delta$$
$$\left|f_\iota^{(p)}(x+v) - f_\iota^{(p)}(x)\right| \leq c_f(x)\,|v|^\delta$$
$$\left|G_\iota^{K(p)}(x+v) - G_\iota^{K(p)}(x)\right| \leq c_{G^K}(x)\,|v|^\delta$$
$$\left|f_\iota^{K(p)}(x+v) - f_\iota^{K(p)}(x)\right| \leq c_{f^K}(x)\,|v|^\delta$$

$p + \delta$ moments of $K(\cdot)$ exist.

The main result is given next,

Theorem 1 *Given Assumptions 1–7.*
(i) $N \to \infty, h \to 0$
(ii) for some $\varepsilon > 0$, $N^{1-\varepsilon}h^{2k+4} \to \infty$
(iii) $Nh^{2(k+3)} \to 0$
then $\sqrt{N}(\text{vech}\widehat{M} - \text{vech}M))$ has a limiting normal distribution with mean 0 and variance Σ_M, where

$$\Sigma_M = E(\text{vech}\, R(z_i)\,(\text{vech}\, R(z_i))^T) - (\text{vech}\, M)(\text{vech}\, M)^T$$

is the variance-covariance matrix of vech $R(z_i)$, *and*

$$\begin{aligned} R(z_i) = & g'(x_i)g'(x_i)^T s_{x_i} + 2(g(x_i) - y_i)g''(x_i)s_{x_i} \\ & + (g(x_i) - y_i)\frac{f'(x_i)g'(x_i)^T + g'(x_i)f'(x_i)^T}{f(x_i)} s_{x_i} \\ & + (g(x_i) - y_i)(f'(x_i)g'(x_i)^T + g'(x_i)f'(x_i)^T)s'_{x_i} \\ & + g'(x_i)g'(x_i)f(x_i)s'_{x_i} - E(g'(x)g'(x)^T f(x)s'_x). \end{aligned}$$

As is typically necessary for $\sqrt{N}$ convergence of the average of nonparametric estimators, condition (*iii*) of Theorem 1 implies that the nonparametric kernel estimates must be (asymptotically) undersmoothed.

The proof of $\sqrt{N}$ consistency and asymptotic normality of vech $\hat{M}$ follows a similar strategy as Härdle and Stoker (1989) and Stoker (1991). The asymptotic properties of vech $\hat{M}$ are derived element wise, where the $(K, L)^{th}$ element of $\hat{M}$, $\hat{M}_{KL}$, is given by

$$\begin{aligned} \hat{M}_{KL} = & \ \tfrac{1}{N}\textstyle\sum_{i=1}^{N}\left[\hat{g}'(x_i)\hat{g}'(x_i)^T\right]_{KL}\hat{s}_{x_i} \\ = & \ \tfrac{1}{N}\textstyle\sum_{i=1}^{N}\frac{\hat{G}^K(x_i)\hat{G}^L(x_i)\hat{f}^2(x_i)}{\hat{f}^4(x_i)}\hat{s}_{x_i} - \frac{\hat{G}(x_i)\hat{f}^K(x_i)\hat{G}^L(x_i)\hat{f}(x_i)}{\hat{f}^4(x_i)}\hat{s}_{x_i} \\ & - \frac{\hat{G}^K(x_i)\hat{G}(x_i)\hat{f}^L(x_i)\hat{f}(x_i)}{\hat{f}^4(x_i)}\hat{s}_{x_i} + \frac{\hat{G}^2(x_i)\hat{f}^K(x_i)\hat{f}^L(x_i)}{\hat{f}^4(x_i)}\hat{s}_{x_i}. \end{aligned} \tag{11}$$

We recall that derivatives with respect to the K-th (L-th) element of x are indicated using superscript K (L), so, e.g., $G^K(x) = \frac{\partial G(x)}{\partial x_K}$,where $\hat{G}^K$ indicates its nonparametric estimator.

To prove Theorem 1, we make use of three technical lemmas. The first lemma shows that in deriving the asymptotic distribution of $\hat{M}$ we can concentrate on deriving the limiting distribution of a linearized version of the estimator. The linearization of $\hat{M}_{KL}$ along the

lines of Stoker (1991), denoted by $\tilde{M}_{KL}$, is given by

$$\begin{aligned}\tilde{M}_{KL} = \; & \tfrac{1}{N}\sum_{i=1}^{N}[\,\left[g'(x_i)g'(x_i)^T\right]_{KL} s_{x_i} \\ & + \left[\hat{G}^K(x_i) - G^K(x_i)\right]\tfrac{g^L(x_i)}{f(x_i)}s_{x_i} - \left[\hat{f}^K(x_i) - f^K(x_i)\right]\tfrac{g(x_i)g^L(x_i)}{f(x_i)}s_{x_i} \\ & + \left[\hat{G}^L(x_i) - G^L(x_i)\right]\tfrac{g^K(x_i)}{f(x_i)}s_{x_i} - \left[\hat{f}^L(x_i) - f^L(x_i)\right]\tfrac{g(x_i)g^K(x_i)}{f(x_i)}s_{x_i} \\ & - \left[\hat{G}(x_i) - G(x_i)\right]\tfrac{f^K(x_i)g^L(x_i)+f^L(x_i)g^K(x_i)}{f(x_i)^2}s_{x_i} \\ & + \left[\hat{f}(x_i) - f(x_i)\right]\tfrac{-2g^L(x_i)g^K(x_i)f(x_i)+g(x_i)(f^K(x_i)g^L(x_i)+f^L(x_i)g^K(x_i))}{f(x_i)^2}s_{x_i} \\ & + [\hat{s}_{x_i} - s_{x_i}]\left[g'(x_i)g'(x_i)^T\right]_{KL}\,].\end{aligned} \tag{12}$$

Lemma 1 *Given Assumptions 1–5, as*

(i) $N \to \infty, h \to 0$

(ii) for some $\varepsilon > 0$, $N^{1-\varepsilon}h^{2k+4} \to \infty$

$\sqrt{N}\left(\hat{M}_{KL} - \tilde{M}_{KL}\right) = o_p(1)$.

The second lemma gives an asymptotic linear expansion for the linearized estimator. Since the linearized estimator can be represented as the sum of "average kernel estimators", second order $U-$statistics theory can be used to provide this asymptotic expansion (see also Hoeffding (1948) and Lemma 3.1 in Powell, Stock and Stoker (1989)).

Lemma 2 *Given Assumptions 1–6*

$$\sqrt{N}(\tilde{M}_{KL} - E(\tilde{M}_{KL})) = N^{-1/2}(\sum_{i=1}^{N} r(z_i) - E(r(z)))) + o_p(1),$$

where

$$\begin{aligned} r(z) = {} & g^L(x)g^K(x)s_x + 2(g(x) - y)g^{KL}(x)s_x + \\ & (g(x) - y)\frac{f^K(x)g^L(x) + f^L(x)g^K(x)}{f(x)}s_x + \\ & (g(x) - y)(f^K(x)g^L(x) + f^L(x)g^K(x))s'_x + \\ & g^K(x)g^L(x)f(x)s'_x - E(g^K(x)g^L(x)f(x)s'_x). \end{aligned}$$

The last lemma shows that $\tilde{M}_{KL}$ (and consequently $\hat{M}_{KL}$) are consistent estimators of M_{KL}. Here the use of higher order kernels in combination with undersmoothing is needed to ensure that possible biases vanish sufficiently fast asymptotically.

Lemma 3 *Given Assumptions 1–5, and 7 as*

(iii) $Nh^{2(k+3)} \to 0$

$E\left\{\tilde{M}_{KL}\right\} - M_{KL} = o(N^{-1/2})$.

The proofs of the lemmas are given in the Appendix.

We now turn to the proof of our main result.

Proof of Theorem 1

Lemmas 1–3 give

$$\sqrt{N}(\hat{M} - M) = N^{-1/2}\sum(R(z_i) - E(R)) + o_p(1), \tag{13}$$

with

$$\begin{aligned} R(z_i) = {} & g'(x_i)g'(x_i)^T s_{x_i} + 2(g(x_i) - y_i)g''(x_i)s_{x_i} \\ & + (g(x_i) - y_i)\frac{f'(x_i)g'(x_i)^T + g'(x_i)f'(x_i)^T}{f(x_i)}s_{x_i} \\ & + (g(x_i) - y_i)(f'(x_i)g'(x_i)^T + g'(x_i)f'(x_i)^T)s'_{x_i} \\ & + g'(x_i)g'(x_i)^T f(x_i)s'_{x_i} - E(g'(x)g'(x)^T f(x)s'_x). \end{aligned} \tag{14}$$

Clearly $E(R(z)) = E(g'(x)g'(x)^T s_x) = M$.

Using the linear vech operator, this gives

$$\sqrt{N}(\text{vech}\hat{M} - \text{vech}M)) = N^{-1/2}\sum(\text{vech}R(z_i) - E(\text{vech}R)) + o_p(1). \tag{15}$$

Application of the Lindeberg-Levy Central Limit theorem then yields the result. □

The covariance matrix can be consistently estimated as

$$\widehat{\Sigma}_M = \frac{1}{N}\sum_{i=1}^{N}\left(\left(\text{vech}\,\widehat{R}(z_i)\right)\left(\text{vech}\,\widehat{R}(z_i)\right)^T\right) - \left(\text{vech}\,\widehat{M}\right)\left(\text{vech}\,\widehat{M}\right)^T, \tag{16}$$

with

$$\begin{aligned} \widehat{R(z_i)} = {} & \left(\widehat{g}'_i\widehat{g}'^T_i + 2(\widehat{g}_i - y_i)\widehat{g}''_i + (\widehat{g}_i - y_i)\frac{\widehat{f}'_i\widehat{g}'^T_i + \widehat{g}'_i\widehat{f}'^T_i}{\widehat{f}_i}\right)\hat{s}_{x_i} + \\ & (\widehat{g}_i - y_i)(\widehat{f}'_i\widehat{g}'^T_i + \widehat{g}'_i\widehat{f}'^T_i)\hat{s}'_{x_i} + \widehat{g}_i{}'\widehat{g}_i{}'^T\widehat{f}_i\hat{s}'_{x_i} - \overline{\widehat{g}'\widehat{g}'^T\widehat{f}\hat{s}'_x}, \end{aligned} \tag{17}$$

where nonparametric kernel regression estimates, its derivatives, and kernel derivative density estimates are used.

Theorem 2 *Under the conditions of Theorem 1,* $\widehat{\Sigma}_M$ *is a consistent estimator of* Σ_M.

The proof is given in the Appendix. An alternative consistent estimator of Σ_M, in analogy to Härdle and Stoker (1989), can be derived from the U-statistics decomposition described in the Appendix.

3 Parameter identification and estimation

We now turn to the GMM framework we use to estimate the parameters of interest β_p, $p = 1, .., P$ and the auxiliary parameters γ_ω. To estimate these parameters, identifying assumptions, such as the exclusion and normalization restrictions used by, for example, Ichimura and Lee (1991) have to be imposed. In making these assumptions, implicitly, an assumption about the number of indices to be estimated is made as well. When we discuss the issue of identification, we will be more explicit about the implications of such an assumption. In Section 4, we will also present a test on the validity of the assumption about the number of indices to be estimated, something which has not been considered before.

Let us recall the moment conditions for the parameters of interest

$$\text{vech}\left[E(g'(x)g'(x)^T s_x) - \sum_{1\leq p,q\leq P} \gamma_{pq}\beta_p\beta_q^T\right] = 0, \tag{18}$$

where $\gamma_{pq} = E\left\{s_x \frac{\partial H}{\partial(x^T\beta_p)}\frac{\partial H}{\partial(x^T\beta_q)}\right\} = \gamma_{qp}$. Define $\Gamma = [\gamma_{pq}]_{p,q=1,..,P}$ and $\gamma = \text{vech}(\Gamma)$. The moment conditions depend on more parameters than can be uniquely identified in the model, where one can think, for example, about the identification of β "up-to-scale". While we will turn to the issue of identification later, let θ_0 denote the vector of identified parameters, which is an element of a compact parameter space Θ.[3]

[3]When identification is obtained by imposing the usual exclusion and normalisation restriction, we will

Assumption 8 $\theta_0 \in \Theta$, *where* Θ *is compact.*

The moment conditions given above are a function of the identified parameters and will be denoted as

$$m(\theta_0) = 0. \tag{19}$$

Let $\hat{m}(\theta_0)$ denote the estimated sample analogue of $m(\theta_0)$

$$\begin{aligned} \hat{m}(\theta_0) &\equiv \text{vech}\left[\frac{1}{N}\sum_{i=1}^{N}(\widehat{g}'(x_i)\widehat{g}'(x_i)^T\hat{s}_{x_i}) - \sum_{1\leq p,q\leq P}\gamma_{pq}\beta_p\beta_q^T\right] \\ &= \text{vech}\left[\widehat{M} - \sum_{1\leq p,q\leq P}\gamma_{pq}\beta_p\beta_q^T\right]. \end{aligned} \tag{20}$$

From Theorem 1, we note that $\sqrt{N}(\text{vech}(\hat{M}) - \text{vech}(E(g'(x)g'(x)^T s_x))) = \sqrt{N}(\hat{m}(\theta_0) - m(\theta_0)) \equiv \sqrt{N}\hat{m}(\theta_0)$ has a limiting normal distribution with mean 0 and variance-covariance matrix Σ_M. The proposed generalized method of moments estimator (or minimum distance estimator) for estimating θ_0, therefore, is given by

$$\widehat{\theta} = \arg\min_{\theta\in\Theta} \hat{m}(\theta)^T \left[\Sigma_M\right]^{-1} \hat{m}(\theta), \tag{21}$$

where we assume Σ_M to be positive definite. To implement the GMM estimator for θ_0 we can use the consistent estimator for Σ_M, $\widehat{\Sigma}_M$, given in the previous section.

To prove consistency of our parameter estimates, $\widehat{\theta}$, we need to show that the regularity conditions ensuring identification and uniform convergence are satisfied, see, for example, Theorem 2.6 in Newey and McFadden (1994).

In order to establish conditions for identification of the parameters, let us first turn to the number of indices to be estimated. This is an important determinant of the number of parameters in the model. The implications of an assumption on the number of indices in the model should therefore include information about the number of parameters that can be estimated. Recall, that $g(x)$ can be written in the "multiple index" form $g(x) = H(x^T\beta_1, .., x^T\beta_P)$. The assumption that P is the minimum number of indices required to

see that θ_0 contains all γ parameters and the parameters of β that are not restricted by the identifying restrictions.

appropriately model $E(y|x)$ as $H(x^T\beta_1, .., x^T\beta_P)$ can be formalized with two assumptions, that have to hold simultaneously. The first assumption is on the indices $B \equiv (\beta_1, .., \beta_P)$ and the second is on the shape of the function $H(.)$.

Assumption 9 *Rank*(B) =*rank*$((\beta_1, .., \beta_P)) = P$.

Assumption 10 *The function* $H(.)$ *satisfies*

$$rank\left(E\left\{\left[\frac{\partial H}{\partial(x^T\beta_p)}\frac{\partial H}{\partial(x^T\beta_q)}\right]_{1\leq p,q\leq P}\right\}\right) = P.$$

Assumption 9 assures that no fewer than P indices are needed by ruling out multicollinearity of the indices. The exclusion restrictions usually applied in semi-parametric multi-index models, see, among others, Ichimura and Lee (1991), are sufficient for this assumption to hold, but other restrictions are also possible. At the same time Assumption 10 asserts that each of the indices provides unique information on the shape of $H(.)$, that is, the derivatives of $H(.)$ with respect to each of the indices are not linearly dependent, almost everywhere. Since our estimator of the outer product of derivatives is based on an estimator that uses trimming, we strengthen Assumption 10 to ensure that the number of indices is not affected by the trimming function

Assumption 10′ *When* P *indices are estimated and the trimming function used to estimate* M *is* $s(f(x) - b) \equiv s_x$*, the function* $H(.)$ *satisfies*

$$rank\left(E\left\{\left[s_x\frac{\partial H}{\partial(x^T\beta_p)}\frac{\partial H}{\partial(x^T\beta_q)}\right]_{1\leq p,q\leq P}\right\}\right) = P$$

With Assumptions 9 and 10 we ensure that indeed P indices have to be estimated.[4]

The vector of all parameters in the model is given by $[\beta_1^T, .., \beta_P^T, \gamma^T]^T$ and includes a total of $kP + P(P+1)/2$ parameters. Let us consider the identification of these parameters

[4]Assumptions 10 and 10′ are used interchangeably.

based on the matrix M under the identifying assumptions stated in Assumptions 9 and 10, with

$$M = E(g'(x)g'(x)^T s_x) = \sum_{p,q} E\left[\frac{\partial H}{\partial(x^T\beta_p)}\frac{\partial H}{\partial(x^T\beta_q)}s_x\right]\beta_p\beta_q^T \tag{22}$$
$$= \sum_{p,q}\gamma_{pq}\beta_p\beta_q^T = B\Gamma B^T,$$

a symmetric matrix. The rank of M equals P, which follows from the fact that B has rank P and Γ is of full rank. Consequently, there are $kP - P(P-1)/2$ free elements in M. So, at most $kP - P(P-1)/2$ parameters can be estimated. In other words, P^2 identifying restrictions will be needed. Indeed, the restrictions that are usually applied in this type of models are a total of P^2 normalization and exclusion restrictions on β. A more general set of identifying assumptions that could be proposed is to impose orthonormality on the βs (Xia et al. (2002)), so $B^TB = I$, in combination with the assumption that $\Gamma = D$, a positive definite diagonal matrix. These restrictions do not require us to specify exclusion restrictions. However, interpretation of the resulting parameter estimates is more cumbersome, and we have decided to take the approach generally accepted in the econometrics literature.

The exclusion and normalization restrictions are summarized in Assumption 11 and guarantee that Assumption 9 is satisfied.

Assumption 11 *Each index $x^T\beta_p$, $p = 1, .., P$, contains one explanatory variable which does not enter the other $P-1$ indices. In the equations where these variables do occur, the parameters on these variables are normalized to equal* 1.

Assumption 11 leads to $P(P-1)$ exclusion restrictions and P normalization restrictions, resulting in a total of P^2 restricted parameters. The vector of identified parameters, θ_0, therefore contains the $P(k-P)$ free parameters in B and the parameters in γ.

The restrictions on the parameters in Assumption 11 have already been discussed by Ichimura and Lee (1991) and Lee (1995). Necessity of these restrictions is clear, but it has not yet been shown that these restrictions also provide sufficient information for identification of β. In fact, one needs an assumption on the shape of $H(\cdot)$ as well, such as Assumption

10. Otherwise, one could have, for example, $H(\cdot) = 0$, which makes identification impossible. We now set out to prove the following lemma.

Lemma 4 *With Σ_M positive definite, Assumptions 10 and 11 ensure that, for $\theta \in \Theta$, $m(\theta)^T [\Sigma_M]^{-1} m(\theta)$ is uniquely minimized at $\theta_0 \in \Theta$.*

Proof of Lemma 4

Assumption 11 enables us to write without loss of generality

$$B = \begin{pmatrix} 1 & 0 & \cdots & 0 & \beta_{1,P+1} & \beta_{1,P+2} & \cdots & \beta_{1,k} \\ 0 & 1 & \cdots & 0 & \beta_{2,P+1} & \beta_{2,P+2} & \cdots & \beta_{2,k} \\ \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 1 & \beta_{P,P+1} & \beta_{P,P+2} & \cdots & \beta_{P,k} \end{pmatrix}^T .$$

With this ordering of the indices and x's, it is straightforward to show that the first column of M is equal to $\sum_{1 \leq p \leq P} \gamma_{p1} \beta_p$, since $M = B\Gamma B^T$. Similar expressions can be derived for the first P columns of M. The first P columns of M therefore equal $B\Gamma$. Since the first P rows of B constitute the identity matrix, Γ is identified from the upper left $P \times P$ block of M, which is straightforward to see by decomposing M as

$$M = B\Gamma B^T = \begin{pmatrix} \Gamma & \Gamma \tilde{B}^T \\ \tilde{B}\Gamma & \tilde{B}\Gamma\tilde{B}^T \end{pmatrix} \tag{23}$$

where $\tilde{B}$ is such that $B = (I : \tilde{B}^T)^T$, based on the decomposition shown above.

Given that Γ is identified and nonsingular by Assumption 10, $\tilde{B}$ (and B) is identified from the remaining $k - P$ rows of the first P columns of M, which equal $\tilde{B}\Gamma$. As there is only a single θ that satisfies the moment conditions for the first P columns of M, there is at most one θ that satisfies all moment conditions, which is θ_0. $\square$

Lemma 4 shows that we are able to uniquely identify $\tilde{B}$ and Γ, so we define $\theta_0 = [\text{vec}(\tilde{B})^T, \text{vech}(\Gamma)^T]^T$ as the vector of identified parameters under assumptions 9 and 10. This uniqueness result in combination with the continuity of $m(\theta)$ on Θ and the compactness Assumption 8 ensures identification of our estimator $\hat{\theta}$. Uniform weak convergence is

ensured by the consistency of $\hat{M}$ and $\widehat{\Sigma}_M$, which was proved in the previous section, and Assumption 8.

Provided we assume that θ_0 lies in the interior of Θ, the only additional condition that needs to be considered to ensure that all regularity conditions required for our asymptotic normality result of $\hat{\theta}$ are satisfied (see Theorem 3.2 in Newey and McFadden (1994)), is that $m'(\theta_0)^T \Sigma_M^{-1} m'(\theta_0)$ is nonsingular, where

$$m'(\theta_0) = \frac{\partial m(\theta_0)}{\partial \theta^T} = \frac{\partial}{\partial \theta^T} \operatorname{vech}\left(B \Gamma B^T\right). \tag{24}$$

In other words, given the nonsingularity of Σ_M, we need to show that $m'(\theta_0)$ has full column rank. Using the decomposition given in (23), it is sufficient to show that the submatrix $\left(\left(\frac{\partial}{\partial \theta^T} \operatorname{vech}(\Gamma)\right)^T, \left(\frac{\partial}{\partial \theta^T} \operatorname{vec}(\tilde{B}\Gamma)\right)^T\right)^T$ has full column rank. The reason for this is that reducing the number of rows and changing the order of the rows can only lead to a reduction in the column rank of a matrix.

We start with analyzing $\frac{\partial}{\partial \theta^T} \operatorname{vech}(\Gamma)$. With $\frac{\partial}{\partial \operatorname{vec}(\tilde{B})^T} \operatorname{vech}(\Gamma) = 0$ and $\frac{\partial}{\partial \operatorname{vech}(\Gamma)^T} \operatorname{vech}(\Gamma) = I_{P(P+1)/2}$, we obtain $\frac{\partial}{\partial \theta^T} \operatorname{vech}(\Gamma) = \left(0 : I_{P(P+1)/2}\right)$. To analyze $\frac{\partial}{\partial \theta^T} \operatorname{vec}(\tilde{B}\Gamma)$ we make use of results and notation from Magnus and Neudecker (1988).[5,6] In particular, using the equalities $\operatorname{vec}(\tilde{B}\Gamma) = (\Gamma \otimes I_{k-P}) \operatorname{vec}(\tilde{B})$ by the symmetry of Γ and $\operatorname{vec}(\tilde{B}\Gamma) = (I_P \otimes \tilde{B}) \operatorname{vec}(\Gamma)$, we obtain $\frac{\partial}{\partial \operatorname{vec}(\tilde{B})^T} \operatorname{vech}(\tilde{B}\Gamma) = (\Gamma \otimes I_{k-P})$ and $\frac{\partial}{\partial \operatorname{vech}(\Gamma)^T} \operatorname{vech}(\tilde{B}\Gamma) = (I_P \otimes \tilde{B}) D_P$, where D_P is the duplication matrix, which satisfies $D_P \operatorname{vech}(\Gamma) = \operatorname{vec}(\Gamma)$. In other words, we obtain $\frac{\partial}{\partial \theta^T} \operatorname{vec}(\tilde{B}\Gamma) = \left((\Gamma \otimes I_{k-P}) : (I_P \otimes \tilde{B}) D_P\right)$.

Combining these results we get

$$\left(\left(\frac{\partial}{\partial \theta^T} \operatorname{vech}(\Gamma)\right)^T, \left(\frac{\partial}{\partial \theta^T} \operatorname{vec}(\tilde{B}\Gamma)\right)^T\right)^T = \begin{pmatrix} 0 & I_{P(P+1)/2} \\ \Gamma \otimes I_{k-P} & (I_P \otimes \tilde{B}) D_P \end{pmatrix}. \tag{25}$$

Given the structure of this matrix, especially the block of zeroes in the top left corner, the rank of this matrix is at least as large as $\operatorname{rank}(I_{P(P+1)/2}) + \operatorname{rank}(\Gamma \otimes I_{k-P})$, which

[5] When A, B, and C are three matrices such that the matrix product ABC is defined, $\operatorname{vec}(ABC) = (C^T \otimes A) \operatorname{vec}(B)$ (Magnus and Neudekker (1988)).

[6] With A an $n \times n$ matrix, the duplication matrix D_n (unique), transforms vech into vec, or $D_n \operatorname{vech}(A) = \operatorname{vec}(A)$ (Magnus and Neudekker (1988)).

equals $\frac{1}{2}P(P+1)+P(k-P)$ using Assumption 10. As this equals the number of columns, we have found that $\left(\left(\frac{\partial}{\partial\theta^T}\text{vech}(\Gamma)\right)^T, \left(\frac{\partial}{\partial\theta^T}\text{vec}(\tilde{B}\Gamma)\right)^T\right)^T$, and therefore $m'(\theta_0)$, has full column rank.

By satisfying all regularity conditions of GMM estimators (minimum distance estimators), our final result is given by

Theorem 3 *Under the conditions of Theorem 1, with Σ_M positive definite, θ_0 in the interior of Θ, Assumptions 8, 10′ and 11*

$$\sqrt{N}(\hat{\theta}-\theta_0) \sim N(0,\Omega)$$

with $\Omega = \left[m'(\theta_0)^T\Sigma_M^{-1}m'(\theta_0)\right]^{-1}$.

Since $\frac{\partial m(\theta)}{\partial\theta}$ is continuous in θ and θ is consistenly estimated by $\hat{\theta}$, Ω can be consistently estimated by $\left[m'(\hat{\theta})^T\hat{\Sigma}_M^{-1}m'(\hat{\theta})\right]^{-1}$.

4 Testing for the number of indices

Estimation of β can only be performed conditional on the number of indices to be estimated. So far, the number of indices has been imposed, either by economic theory, or by the researcher. However, we note that it is possible to test for the number of indices. Assumptions 9 and 10 imply that P indices have to be estimated, but, at the same time, they also imply that $M = B\Gamma B^T$ has rank P. Each index in the multiple index model reduces the null space of M with one dimension. The number of indices therefore can be tested by testing the rank of the estimated average outer product of the gradient. This is similar to determining the number of effective dimension reduction (EDR) directions in Xia et al. (2002). No statistical procedure for testing the number of dimensions is provided by these authors.

In this section we discuss testing for the number of indices in the multiple index model. We start with tests that are based on the property that rank M equals P. These tests do not require an estimate of β itself. Another type of test for the number of indices

is provided by the GMM framework used to estimate β. The validity of the assumption concerning the number of indices can be tested by a test of the validity of the overidentifying restrictions. One word of caution should be made here. The overidentifying restrictions test is a general test for misspecification. When identification is obtained by assuming $B^TB = I$ and $\Gamma = D$, a diagonal matrix, the only possible misspecification is too few indices in the model. When exclusion and normalization restrictions are used, this imposes a structure on the model, which will be tested by the overidentifying restrictions test as well. The test for overidentifying restrictions within the GMM framework is well known. In the remainder of this section we focus on tests for the number of indices based on the rank of M.

In recent years, there has been a resurgence of interest in the development of tests of the rank of a matrix. Cragg and Donald (1996, 1997), Gill and Lewbel (1992), and Robin and Smith (2000) develop tests for the rank of a matrix that is unobserved but for which a $\sqrt{N}$ consistent asymptotically normal estimator is available. Gill and Lewbel (1992), the first authors to consider this problem, base their test on a Gaussian elimination Lower-Diagonal-Upper triangular (LDU) decomposition. For any symmetric matrix M of full rank, there exists a unique decomposition of the form $QMQ^T = LDL^T$, where Q is a permutation matrix (row permutations) derived from complete pivoting, L is a lower triangular matrix and D is a diagonal matrix with the diagonal elements decreasing in absolute size. In case M is not of full rank, this decomposition can still be made, but it is not unique. The rank of M equals the number of nonzero elements in the diagonal "pivot" matrix D in this decomposition (see also Golub and van Loan (1983)). Focussing on a test for the rank of M being P, Gill and Lewbel partition the LDL^T decomposition as

$$LDL^T = \begin{bmatrix} L_{11} & 0 \\ L_{21} & L_{22} \end{bmatrix} \begin{bmatrix} D_1 & 0 \\ 0 & D_2 \end{bmatrix} \begin{bmatrix} L_{11}^T & L_{21}^T \\ 0 & L_{22}^T \end{bmatrix}. \tag{26}$$

L_{11} and L_{22} are a lower triangular matrices of dimension P and $(k-P)$ respectively, L_{21} is a $(k-P) \times P$ matrix, and D_1 and D_2 are diagonal matrices, also of dimension P and $(k-P)$ respectively. When the matrix has rank P, the diagonal elements in D_2 will be

zero, resulting in the non-uniqueness of the decomposition in L_{22}, which complicated the development of the asymptotic theory in Gill and Lewbel. In fact, Cragg and Donald (1996) show that the asymptotic theory only holds for $k - P = 1$.

Instead of full pivoting, proposed by Gill and Lewbel (1992), Cragg and Donald (1996) propose not to "sweep" the smallest $k - P$ pivots and to partition QMQ^T as

$$QMQ^T = \begin{bmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{bmatrix}.$$

Here A_{11} is a $P \times P$ matrix of assumed full rank, A_{12} is a $P \times (k - P)$ matrix, $A_{21} = A_{12}^T$, and A_{22} is a matrix of dimension $k-P$. Cragg and Donald consider a test statistic based on $\Omega_{22} = A_{22} - A_{21}A_{11}^{-1}A_{12}$. This is what remains in the bottom-right $(k-P) \times (k-P)$ corner after sweeping the first P columns. The test is based on the realization that $\text{rank}(\Omega_{22}) = 0$ if and only if M has rank equal to P. Specifically, the test statistic for symmetric matrices is given by $N \,\text{vech}(\hat{\Omega}_{22})^T \hat{W}^{-1} \,\text{vech}(\hat{\Omega}_{22})$, where

$$\begin{aligned} \hat{W} &= R\hat{H}[\widehat{\text{Avar}}\left(\text{vec}(QMQ^T)\right)]\hat{H}^T R^T, \\ \hat{H} &= \left[-\hat{A}_{21}\hat{A}_{11}^{-1} : I_{k-P}\right] \otimes \left[-\hat{A}_{12}^T\hat{A}_{11}^{-1} : I_{k-P}\right]. \end{aligned} \tag{27}$$

R is the Moore-Penrose inverse (unique) of the duplication matrix D_{k-P}, D_{k-P}^+, which converts the vec of Ω_{22} into $\text{vech}(\Omega_{22})$ (see also Magnus and Neudecker, 1988). The estimated asymptotic variance of $\text{vec}(QMQ^T)$ is given by $(\hat{Q} \otimes \hat{Q})D_{kP}\Sigma_M D_{kP}^T(\hat{Q}^T \otimes \hat{Q}^T)$, where D_{kP} is again a duplication matrix.

The gaussian elimination based test is asymptotically equivalent to the minimum chi-squared approach presented in more detail in Cragg and Donald (1997)

$$\min_M \{N(\text{vech}(\hat{M} - M)^T \Sigma_M^{-1} \,\text{vech}(\hat{M} - M)) : rank(M) = P\}.$$

The resulting test statistic is distributed asymptotically as chi-squared with $(k-P)(k-P+1)/2$ degrees of freedom. This test is equivalent to the test for overidentifying restrictions in the GMM framework. An interpretation of the degrees of freedom, therefore, easily follows from the number of overidentifying restrictions.

The Robin and Smith (2000) test for the rank of a matrix involves the characteristic roots of a quadratic form in M given by $\Sigma M \Psi M^T$, with Σ and Ψ positive definite. We set Σ and Ψ equal to the identity matrix. Again, when $\text{rank}(M) = P$, the smallest $k - P$ eigenvalues of $\Sigma \hat{M} \Psi \hat{M}^T$ (and similarly of $\hat{M}$) converge to 0 in probability. Robin and Smith show that N times the sum of the $k - P$ smallest eigenvalues of $\Sigma \hat{M} \Psi \hat{M}^T$ converges to a weighted sum of independent χ^2_1 distributed variables. The advantage of this test is that the variance-covariance matrix of $\text{vec}(\hat{M})$ is not required to be positive definite, which circumvents the difficulties that arise from symmetry of the matrix.

A caveat which we see with these tests, in finite samples, is that they do not take into account the precision with which the elements in the M matrix are estimated. This holds in particular for the Gaussian elimination procedure in Cragg and Donald (1996) and the selection of the $k - P$ smallest eigenvalues in Robin and Smith (2000). We partially solve this problem by performing both weighted and unweighted variants of these tests. Instead of testing the rank of M we consider testing the rank of the weighted variant $\Lambda M \Lambda^T$, where Λ is a diagonal matrix of full rank which ensures that the diagonal elements of $\Lambda M \Lambda^T$ are estimated with equal precision. Since Λ is of full rank, the rank of M equals that of $\Lambda M \Lambda^T$. This weighting scheme only corrects for differences in the estimation precision of the diagonal elements. One would, however, prefer to correct also for the differences in estimation precision of the non-diagonal elements and the correlations between the estimates. As the test for overidentifying restrictions uses full information on the estimation uncertainty of each element, theoretical considerations would favor the use of this test. In the simulation study we compare the performance of each of these tests in practice.

It should be noted that the tests by Cragg and Donald (1996, 1997) and Robin and Smith (2000) tests do not deal specifically with the positive semidefiniteness of the M matrix unlike Gill and Lewbel (1992). Checking the validity of the tests for positive semi definite matrices and developing the relevant extensions (when required) lies beyond the scope of the present paper.

A sequential procedure for obtaining a weakly consistent estimator for the rank of a

matrix involves testing sequentially whether the rank of a matrix equals r against the alternative that the rank exceeds r, $r = 0, 1, .., k-1$, and halting at the first value for r for which the statistics indicates nonrejection of the null $\text{rank}(M) = r$. Appropriate adjustments dependent on the sample size to the asymptotic size α_P of the test at each stage of the sequential procedure, is required. In particular, we require $\alpha_{rN} = o(1)$ and $-N \ln \alpha_{rN} = o(1)$ (see also Cragg and Donald (1997) and Robin and Smith (2000)).

5 Simulation

In order to illustrate the estimator's usefulness, we perform various simulations. We consider two models each with two indices, one multiplicative model and one additive model

$$\begin{aligned} y &= \left(X^T\beta_1\right)\left(X^T\beta_2\right) + 0.5\varepsilon \qquad \text{(Model 1)} \\ y &= \left(X^T\beta_1\right) + \left(X^T\beta_2\right)^2 + 0.5\varepsilon \qquad \text{(Model 2).} \end{aligned} \tag{28}$$

From Assumption 11, we note that for our two index models ($P = 2$) at least three explanatory variables are required. We assume that $X \sim N(0, I_k)$ and $\varepsilon \sim N(0, 1)$ are independent random variables, where k indicates the number of explanatory variables set equal to three. The multivariate kernel function $K(\cdot)$ (on R^3) is chosen as the product of three univariate kernel functions. We let $\beta_1 = (1, 0, 1)^T$ and $\beta_2 = (0, 1, 1)^T$. The sample size is set at 1000 and 500 replications are drawn in each case.

With the number of explanatory variables equal to three, our theoretical results imply that we use a sixth order kernel, $p = k + 3$. We consider

$$K_6(x) = \frac{35}{256}(-99x^6 + 189x^4 - 105x^2 + 15)1(|x| \leq 1) \tag{29}$$

as proposed by Gasser et al. (1985). Besides using this higher order kernel ("bias-corrected" kernel) we consider using the second order quartic kernel ("not bias-corrected" kernel) as well, because of its easier implementation. Both are bounded, symmetric kernels. A bandwidth sequence $\{h_n\}$ satisfying the assumptions is given by $h_n = cn^{-1/(2k+5)}[= cn^{-1/11}]$, where c is a constant factor independent of n, which we allow to vary for each explanatory

variable. Using cross-validation, we determine for each explanatory variable the constant, where it is noted that the cross-validation bandwidth is given by the optimal bandwidth sequence $h_n^{gcv} = cn^{-1/(2p+3)}$ (see Stone (1982)) with p equalling the order of the kernel. Keeping the constant fixed, we proceed by using the following three increasingly undersmoothed bandwidths, h_n^{gcv}, $h_n^{gcv} \cdot n^{-\frac{1}{2}[1/(2k+5)-1/(2p+3)]}$, and $h_n^{gcv} \cdot n^{-[1/(2k+5)-1/(2p+3)]}$. The latter bandwidth sequence is an undersmoothed bandwidth sequence in accordance with the theoretical requirements, i.e., $h_n = c^* n^{-1/(2k+5)}$, while the middle bandwidth sequence considers some intermediate sequence.

It should be noted, however, that this approach only leads to increasingly undersmoothed bandwidths as long as $p > k+1$. When we use a second order kernel the approach would lead to oversmoothing instead. Our theoretical analysis therefore does not provide guidelines with respect to the amount of undersmoothing required when using a second order kernel. In that case, we apply the same amount of undersmoothing as would be required when using the sixth order kernel – the order theoretically required. Therefore, the three bandwidths considered are given by $h_n^{gcv}, h_n^{gcv} n^{-\frac{1}{2}[1/11-1/15]}$, and $h_n^{gcv} n^{-[1/11-1/15]}$ whether we use a sixth or a second order kernel.

Finally, we consider the following smoothed trimming function

$$s(f(x)-b) = \begin{cases} [\frac{1}{3}d^2(f(x)-b)^3 - \frac{1}{2}d(f(x)-b)^4 + & \\ \qquad \frac{1}{5}(f(x)-b)^5] * 30 * d^{-5} & \text{for} \quad f(x) \in (b, b+d) \\ 0 & \text{for} \quad f(x) \leq b \\ 1 & \text{for} \quad f(x) \geq b+d, \end{cases} \tag{30}$$

where we let $b = 0.0025$ and $d = 0.0025$. This trimming is such that about 9 percent of the observations receive no weight and about 7 percent of the observations receive a weight between zero and one. It is noted though that this depends on the kernel and bandwidths used as well. We also consider the sensitivity to the smoothed trimming function by comparing our results to those we obtain when using the indicator function instead, as suggested in Härdle and Stoker (1989), among others. Here we ignore the influence the indicator trimming function has on the asymptotic variance.

Table 1 presents the results on various rank tests for the two models using the smoothed trimming function. In Table A.1 in the Appendix, the results are presented using the indicator trimming function. The results are presented for second order kernels (not bias corrected) and sixth order kernels (bias corrected) for different bandwidths. The cross-validated bandwidth is obtained using the Nelder Meade optimization algorithm.[7] For Model 1, the second order kernel gives, on average, rise to the cross-validated bandwidth $(0.861, 0.865, 0.565)^T$, while the sixth order kernel gives $(3.091, 3.073, 2.185)^T$. For model 2, the cross-validated bandwidths are, on average, $(1.249, 0.580, 0.559)^T$ and $(3.902, 2.238, 2.202)^T$ for second and higher order kernels respectively.

The table reports the percentage of times we accept that the rank equals 0, 1, or 2 at the five percent significance level given that the true rank equals 2, using the sequential procedure described above. W stands for the Cragg and Donald Wald test based on the LDU decomposition, CRT gives the Robin and Smith test, while OI gives the overidentification test. The overidentification test is the only test that requires estimation of β and γ consistent with the assumed rank of M and the identifying restrictions. It is a minimum Chi-squared test, where subject to the assumption that the true rank equals 0, 1, or 2, the test is computed as

$$N(\text{vech}(\hat{M} - \tilde{M})^T \hat{\Sigma}_M^{-1} \text{vech}(\hat{M} - \tilde{M})) : rank(\tilde{M}) = 0, 1, 2,$$

where $\tilde{M} = \hat{B}\hat{\Gamma}\hat{B}^T$. As we use the normalization and exclusion restrictions in estimating the parameters, these assumptions are also tested. In our simulation setting, we know these assumptions are satisfied. The tests for the rank of a matrix do not rely on the parameter estimates of β or γ. However, these tests do not account for the estimation uncertainty in $\hat{M}$. Therefore, we perform an unweighted and a weighted variant of these tests. The weighted version of the test is based on $\Lambda M \Lambda^T$, such that the estimated variances of the diagonal elements are equal.

[7] To ensure convergence to the global optimum, we start the procedure with different starting values, and consider our problem optimized once no improved cross-validation objective is found four times in a row.

Table 1: Testing the Rank of a Matrix

Model 1: $y = (X^T\beta_1)(X^T\beta_2) + \varepsilon$

Not Bias Corrected, $h_n^{gcv} = (0.861, 0.865, 0.565)^T$ *Bias Corrected,* $h_n^{gcv} = (3.091, 3.073, 2.185)^T$

Rank	W	W_w	CRT	CRT_w	OI	W	W_w	CRT	CRT_w	OI
	$h_n = h_n^{gcv}$									
0	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
1	0.000	0.000	0.000	0.000	0.000	0.080	0.056	0.098	0.024	0.082
2	0.082	0.348	0.504	0.604	0.642	0.782	0.924	0.894	0.970	0.912
	$h_n = h_n^{gcv} n^{-\frac{1}{2}[1/11-1/15]}$									
0	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
1	0.000	0.002	0.000	0.000	0.006	0.046	0.030	0.096	0.014	0.052
2	0.038	0.230	0.336	0.430	0.486	0.762	0.956	0.894	0.982	0.944
	$h_n = h_n^{gcv} n^{-[1/11-1/15]}$									
0	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
1	0.002	0.004	0.000	0.000	0.010	0.032	0.022	0.100	0.006	0.046
2	0.028	0.156	0.220	0.324	0.368	0.618	0.952	0.880	0.990	0.954

Model 2: $y = (X^T\beta_1) + (X^T\beta_2)^2 + \varepsilon$

Not Bias Corrected, $h_n^{gcv} = (1.249, 0.580, 0.559)^T$ *Bias Corrected,* $h_n^{gcv} = (3.902, 2.238, 2.202)^T$

Rank	W	W_w	CRT	CRT_w	OI	W	W_w	CRT	CRT_w	OI
	$h_n = h_n^{gcv}$									
0	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
1	0.000	0.000	0.000	0.000	0.000	0.002	0.002	0.182	0.006	0.002
2	0.002	0.334	0.100	0.454	0.484	0.210	0.980	0.674	0.980	0.984
	$h_n = h_n^{gcv} n^{-[1/11-1/15]}$									
0	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
1	0.000	0.000	0.002	0.000	0.002	0.002	0.000	0.278	0.002	0.002
2	0.000	0.238	0.046	0.330	0.358	0.134	0.990	0.542	0.990	0.994
	$h_n = h_n^{gcv} n^{-[1/11-1/15]}$									
0	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
1	0.000	0.000	0.002	0.000	0.000	0.004	0.004	0.330	0.002	0.004
2	0.000	0.164	0.010	0.248	0.276	0.060	0.966	0.416	0.982	0.984

Focussing on the tests that only require estimation of the M matrix, we see a noticeable difference between the weighted and unweighted variants of the tests. By using the weighted M matrix, we improve our ability to accept the null that the rank equals the true rank both with the Cragg and Donald test and the Robin and Smith test. Moreover, the desire to use a higher order kernel in accordance with the theoretical results over a second order kernel is highlighted. While our ability to find the true rank is still limited when using a second order kernel, our ability to find the true rank is greatly enhanced when using the higher order kernel. In model 1, for instance, using a second order kernel with cross-validated bandwidth we only find support for the true rank in 35 percent of our simulations using the Cragg and Donald test and 60 percent using the Robin and Smith test. Using the higher order kernel instead, we find support for the true rank in respectively 92 and 97 percent of our simulations. Our ability to find the true rank further improves when undersmoothing is used, i.e., when the theoretically correct bandwidth is used. The empirical size of the test of the null that the rank equals the true rank, nevertheless, is lower than the nominal size of the test when using the bias reducing kernel with bandwidth corresponding to our theoretical result, 0.026 and 0.004 for the Cragg and Donald and Robin and Smith test respectively (cannot be obtained directly from our table). This difference between the empirical and nominal size could have arisen from our estimator of the covariance matrix, giving rise to somewhat larger theoretical standard errors than the empirical ones. The power against accepting the null of too few number of indices is better for the Robin and Smith test than the Cragg and Donald test, 0.994 versus 0.978.

The test results for model 2 are similar, be it with a slight improvement in our ability to observe the true rank when using the bias correcting kernel. However, the results of the intermediate level of undersmoothing are better than the results obtained with the theoretically correct bandwidth. While the power is almost identical for the two tests for model 2, 0.996–0.998, the empirical size of the Cragg and Donald test is slightly closer to the nominal size than the Robin and Smith test, 0.03 versus 0.016 using the bandwidth corresponding to our theoretical result. Overall, the weighted Robin and Smith test gives the best performance among the rank based tests in terms of identifying the true rank,

however, when comparing the size and power of the hypothesis that the rank equals the true rank preference is given to the Cragg and Donald test.

The test of overidentifying restrictions confirms the result that our models using two indices are particularly well specified when estimated using a higher order kernel, i.e., the distinct $k(k+1)/2$ moments are jointly zero. When using less indices than needed, the overidentified moments generally are significantly different from zero. Using a second order kernel does not lead to a similar support for the specification being appropriate when the true number of indices are considered. We argue therefore for the need of using a higher order kernel in accordance with theoretical requirements, in contrast with the usual practice where second order kernels are more common. Importantly, the overidentification test results correspond closely to the weighted test results and are thereby supportive or our modification of the rank tests, not considered elsewhere in the literature. Comparing the performance of the tests for determining the number indices, we finally note that while the weighted Robin and Smith test outperforms the overidentification test for model 1, the reverse is true for model 2. The weighted Robin and Smith test for testing the null that the rank equals the true rank has similar empirical size as the overidentification test for both models.

When we use the indicator trimming instead, see Table A.1, we can draw exactly the same conclusions for the bias corrected kernel. The only difference is a slight improvement in the test results for model 2, against a slight worsening in the test results for model 1. For the second order kernel, the performance of the estimator with indicator trimming gets even worse. The right model is selected in only about 10 percent of the cases, compared to more than 90 percent for the bias corrected estimates.

Tables 2 and 3 present an analysis of the parameter estimates for $[\tilde{\beta}_1^T, .., \tilde{\beta}_P^T]^T$ for the two models, where the number of indices, P, equals the true number of indices (which in turn is equal to the true rank of M). Because we are primarily interested in the β parameter estimates, tables giving the analysis of the γ parameter estimates are relayed to the Appendix, see Tables A.2 and A.3. Given the smoothed trimming function used ($b = 0.0025$, $d = 0.0025$), the true parameter vector $\theta = [\beta_{1,3}, \beta_{2,3}, \gamma^T]^T$ of the two models

Table 2: Model 1, β parameter estimates

Model 1: $y = (X^T\beta_1)(X^T\beta_2) + \varepsilon$

Not Bias Corrected, $h_n^{gcv} = (0.861, 0.865, 0.565)^T$

	True	MEAN	Var *sample*	Var *theory*	MSE	LQ	Median	UQ	MAE
				$h_n = h_n^{gcv}$					
$\beta_{1,3}$	1.000	1.047	0.025	0.058	0.027	0.949	1.040	1.141	0.124
$\beta_{2,3}$	1.000	1.059	0.024	0.058	0.028	0.963	1.047	1.149	0.125
				$h_n = h_n^{gcv} n^{-\frac{1}{2}[1/11-1/15]}$					
$\beta_{1,3}$	1.000	1.018	0.048	0.078	0.049	0.901	1.022	1.122	0.156
$\beta_{2,3}$	1.000	1.066	0.054	0.079	0.058	0.931	1.045	1.173	0.170
				$h_n = h_n^{gcv} n^{-[1/11-1/15]}$					
$\beta_{1,3}$	1.000	1.010	0.093	0.108	0.093	0.833	0.996	1.138	0.214
$\beta_{2,3}$	1.000	1.066	0.112	0.115	0.117	0.882	1.031	1.186	0.228

Bias Corrected, $h_n^{gcv} = (3.091, 3.073, 2.185)^T$

	True	MEAN	Var *sample*	Var *theory*	MSE	LQ	Median	UQ	MAE
				$h_n = h_n^{gcv}$					
$\beta_{1,3}$	1.000	1.056	0.019	0.064	0.022	0.972	1.055	1.127	0.110
$\beta_{2,3}$	1.000	1.054	0.030	0.076	0.033	0.966	1.038	1.122	0.110
				$h_n = h_n^{gcv} n^{-\frac{1}{2}[1/11-1/15]}$					
$\beta_{1,3}$	1.000	1.018	0.015	0.059	0.015	0.938	1.023	1.094	0.095
$\beta_{2,3}$	1.000	1.007	0.017	0.063	0.017	0.927	1.006	1.088	0.099
				$h_n = h_n^{gcv} n^{-[1/11-1/15]}$					
$\beta_{1,3}$	1.000	0.978	0.017	0.072	0.017	0.896	0.980	1.053	0.100
$\beta_{2,3}$	1.000	0.976	0.019	0.076	0.019	0.900	0.971	1.054	0.102

Table 3: Model 2, β parameter estimates

Model 2: $y = (X^T\beta_1) + (X^T\beta_2)^2 + \varepsilon$

Not Bias Corrected, $h_n^{gcv} = (1.249, 0.580, 0.559)^T$

	True	MEAN	Var *sample*	Var *theory*	MSE	LQ	Median	UQ	MAE
				$h_n = h_n^{gcv}$					
$\beta_{1,3}$	1.000	1.171	0.018	0.054	0.047	1.080	1.168	1.245	0.181
$\beta_{2,3}$	1.000	1.011	0.007	0.011	0.008	0.962	1.001	1.059	0.063
				$h_n = h_n^{gcv} n^{-\frac{1}{2}[1/11-1/15]}$					
$\beta_{1,3}$	1.000	1.121	0.028	0.071	0.043	1.015	1.118	1.211	0.161
$\beta_{2,3}$	1.000	1.014	0.013	0.016	0.013	0.952	1.003	1.066	0.079
				$h_n = h_n^{gcv} n^{-[1/11-1/15]}$					
$\beta_{1,3}$	1.000	1.082	0.046	0.099	0.053	0.940	1.062	1.190	0.167
$\beta_{2,3}$	1.000	1.015	0.033	0.029	0.033	0.918	0.995	1.073	0.112

Bias Corrected, $h_n^{gcv} = (3.902, 2.238, 2.202)^T$

	True	MEAN	Var *sample*	Var *theory*	MSE	LQ	Median	UQ	MAE
				$h_n = h_n^{gcv}$					
$\beta_{1,3}$	1.000	0.956	0.031	0.085	0.033	0.861	0.943	1.032	0.122
$\beta_{2,3}$	1.000	1.016	0.005	0.016	0.006	0.971	1.011	1.055	0.056
				$h_n = h_n^{gcv} n^{-\frac{1}{2}[1/11-1/15]}$					
$\beta_{1,3}$	1.000	0.892	0.018	0.068	0.029	0.807	0.891	0.966	0.139
$\beta_{2,3}$	1.000	1.018	0.006	0.023	0.006	0.974	1.018	1.059	0.056
				$h_n = h_n^{gcv} n^{-[1/11-1/15]}$					
$\beta_{1,3}$	1.000	0.846	0.018	0.081	0.042	0.765	0.848	0.931	0.170
$\beta_{2,3}$	1.000	1.016	0.008	0.031	0.009	0.963	1.016	1.064	0.070

is $(1, 1, 1.325, 0.646, 1.325)$ and $(1, 1, 0.870, 0, 5.294)$ respectively, where the true parameter values for the gamma parameters (the last three or $P(P+1)/2$ parameters) are based on a simulation with 10,000 draws of x.

The tables present the following summary statistics for the 500 replications: the sample mean, the sample variance, the average of the theoretical variance, mean squared error (MSE), lower quartile (LQ), median, upper quartile (UQ), and mean absolute error (MAE). Each are reported for the different bandwidths considered, h_n^{gcv}, $h_n^{gcv} \cdot n^{-\frac{1}{2}[1/11-1/15]}$, and $h_n^{gcv} \cdot n^{-[1/11-1/15]}$. While $h_n^{gcv} \cdot n^{-[1/11-1/15]}$ would provide a bandwidth sequence in accordance with our theory, when we use the sixth order kernel, our chosen values for c, c^*, are not necessarily optimal.[8] For the optimization with respect to β and γ two sets of starting values were considered, the true values (as can only be done in the simulation setting) and starting values obtained directly from our estimated M matrix, with $\hat{\Gamma} = \hat{M}^{P,P}$ and $\hat{B} = \hat{M}_{k-P,P}\ \hat{\Gamma}^{-1}$, where $\hat{M}^{P,P}$ is the upper left $P \times P$ block of $\hat{M}$ and $\hat{M}_{k-P,P}$ is the lower left $(k-P) \times P$ block of $\hat{M}$. Of the two sets of parameter estimates thus obtained, we select the estimate which yield the best GMM objective function. It should be noted that for the bandwidths considered the parameter estimates obtained using these two starting values were by and large identical. In Tables A.4 and A.5 in the Appendix, an analysis of the parameter estimates is presented using the indicator trimming function.

From Table 2, we notice quite reasonable parameter estimates for $\beta_{1,3}$ and $\beta_{2,3}$. Choosing the bandwidth parameter smaller (undersmoothing) causes a reduction in bias while increasing the variance. The selection of a bias-correcting kernel further reduces the bias of our parameter estimates without significantly affecting the theoretical variance of our parameter estimates. The latter is due to the larger bandwidths used (based on cross-validation again) when applying the bias-correcting kernel. In contrast to Powell, Stock, and Stoker (1989) we do not keep c constant when changing from second to higher order kernels. Our theoretical variance typically exceeds that of the sample variance, which might be due to our inability in small samples to observe events that happen only with

[8]The theory for optimally chosing c, when undersmoothing is required, is not well developed.

small, though non-negligible, probability. The MSE of the higher order kernel generally is lower than the second order kernel. Both in terms of MSE and MAE our results give preference to the higher order kernel. This corresponds to the results obtained from our overidentification test, which lend strong support of our specification revealing all unique $k(k+1)/2$ moments to be jointly zero. The median of our parameter estimates closely follows the pattern of the mean of our estimates. When the bias-corrected kernel is used with the undersmoothed bandwidth sequences $h_n^{gcv} \cdot n^{-\frac{1}{2}[1/11-1/15]}$ and $h_n^{gcv} \cdot n^{-[1/11-1/15]}$ the bias of the β parameters is insignificant at the 5 percent level of significance, while the bias is highly significant for the second order kernel estimates. Using the indicator function instead as our trimming function, see Table A.4, as suggested e.g., in Härdle and Stoker (1989), leads as expected to a lower theoretical variance as it ignores the influence our smoothing function has on the asymptotic variance. The sample variance, on the other hand, increases due to the discontinuity of the trimming function. The bias of the parameter estimates using the bias-corrected kernel with the undersmoothed bandwidth sequences is somewhat larger, though not significantly so.

When considering Table 3, we notice that the parameter estimate of $\beta_{2,3}$ is estimated more precisely than $\beta_{1,3}$. Not only does $\hat{\beta}_{2,3}$ have a lower variance, it has been estimated with less bias as well. The stronger signal of the β_2 parameters, which enter through the index $X^T\beta_2$ quadratically, is reflected in the large value for $\gamma_{2,2}$ and provides an explanation for this finding. Contrary to the good estimates for $\beta_{2,3}$, not significantly different from one at the 1 percent level of significance using both bias corrected and not bias corrected kernels, we notice that the bias of $\beta_{1,3}$ is significantly different from zero at this level of significance. While the p-value improves when we use a higher order kernel instead of a second order kernel, the bias of $\beta_{1,3}$ remains significantly different from zero. The improvement in parameter estimates for $\beta_{1,3}$ when using the higher order kernel should be seen in the light of the less biased parameter estimates of γ obtained when using higher order kernels as well. Using the higher order kernel about halves the MAE associated with $\hat{\gamma}_{2,2}$ for instance. As in model 1, the higher order kernel is associated with a lower MSE, lending support to its adoption in addition to the overidentification test results which revealed the

appropriateness of our model when using the bias-corrected kernel. The bias of $\beta_{1,3}$ worsens when rather than using the smoothed trimming, we apply the indicator trimming. While the interquartile range contains the true value when we use the cross-validated bandwidth with smoothed trimming, the true value does not lie in the interquartile range when using indicator trimming instead. So, $\beta_{1,3}$ is strongly underestimated, in combination with the overestimation of $\gamma_{1,1}$.

6 Concluding Remarks

In this paper we consider the estimation of semiparametric multi-index models. Although estimation methods for these models are available for quite some time, these methods all are rather computationally intensive. The advantage of our method is that it only involves a single non-parametric step, which is the computation of the average outer product of the derivatives and its covariance matrix. Parameter estimation is then based on the nonparametrically estimated matrix using a GMM approach. This step involves a simple minimization problem, but importantly, no additional kernel based calculations are required. The estimator is shown to be root-N consistent and asymptotically normal.

Parameter estimation in multi-index models is only feasible when the number of indices is given. So far, the number of indices has been imposed, either by economic theory, or by the researcher – no data driven procedures were considered to determine this. We provide such a procedure, as it can be shown that the rank of the outer product of derivatives equals the number of indices required in the semiparametric model. Application of existing tests for the rank of a matrix then provides the desired testing procedure. The GMM framework used for estimating the parameters of interest provides an alternative way to test the appropriateness of the number of indices chosen through the overidentifying restrictions test.

In a simulation study, we have examined the performance of our model. Our results clearly indicate that using a second order kernel, which is done often in practice, leads to worse results than using a bias reducing kernel, especially for the test on the number

of indices required. Similarly we investigate the effects of using the indicator function as a weighting function, as was proposed by Härdle and Stoker (1989). As argued in the theoretical section, this does not affect the bias of the estimates, but it increases the variance.

While our estimator of β uses moment conditions for the outerproduct of the gradient only, it is straightforward to use additional moment conditions, such as the gradient itself, within our framework in an attempt to improve the efficiency of our estimator. We see various other areas in which this research could be extended. As our method is derivative based, parameter estimation is only feasible for continuous variables. We would like to consider an extension of the work by Horowitz and Härdle (1996) for the ADE framework to deal with discrete variables. Furthermore, the estimator could be extended to the estimation of multi-equation models. The selection of the bandwidth is another area where further research is warranted. In this paper, we used an approach to select the required undersmoothed bandwidth based upon an adjustment of the rate of convergence upon the bandwidth attained under generalized cross validation. An advantage of this approach is that the bandwidth depends not only on the distribution of the explanatory variables, but also on the amount of noise in the dependent variable, which is not the case for many rule-of-thumb procedures. Without further theoretical developments on this topic, this procedure is nothing more than a rule-of-thumb, be it that it is a bit more sophisticated.

Appendix

This appendix gives the proofs of Lemmas 1–3 and Theorem 2 of Section 2. The need for a smoothed trimming function in place of the indicator function arises in the proof of Lemma 2. As a consequence of this, we discuss this issue in more deatil at the end of the proof of Lemma 2. As indicated, the proofs bear close resemblance to that of Härdle and Stoker (1989), Powell, Stock and Stoker (1989), and Stoker (1991).

For notational convenience we do not explicitly mention the dependence of each function on x_i, where no ambiguity exists, so in the sequel $G = G(x_i)$, $f = f(x_i)$, $s = s_{x_i}$. We recall, superscript K (L) denotes the derivative with respect to the K-th (L-th) element of x, and

$\hat{}$ indicates a nonparametrically estimated quantity.

Proof of Lemma 1

To prove that $\sqrt{N}\left(\hat{M}_{KL} - \tilde{M}_{KL}\right) = o_p(1)$ we first rewrite $\hat{M}_{KL}$. From (11)

$$\hat{M}_{KL} = \frac{1}{N}\sum\nolimits_{i=1}^{N}\left(\frac{\hat{G}^K\hat{f}}{\hat{f}^2} - \frac{\hat{G}\hat{f}^K}{\hat{f}^2}\right)\left(\frac{\hat{G}^L\hat{f}}{\hat{f}^2} - \frac{\hat{G}\hat{f}^L}{\hat{f}^2}\right)\hat{s}. \tag{A.1}$$

Taking out $\hat{f}^{-4}$ and replacing $\hat{G}^K$ by $\hat{G}^K - G^K + G^K$ and $\hat{f}^K$ by $\hat{f}^K - f^K + f^K$ gives

$$\begin{aligned}\hat{M}_{KL} \;&= \tfrac{1}{N}\textstyle\sum_{i=1}^{N}\hat{f}^{-4}[(\hat{G}^K - G^K)\hat{f}\left(\hat{G}^L\hat{f} - \hat{G}\hat{f}^L\right)\hat{s} + G^K\hat{f}\left(\hat{G}^L\hat{f} - \hat{G}\hat{f}^L\right)\hat{s} \\ &\quad -(\hat{f}^K - f^K)\hat{G}\left(\hat{G}^L\hat{f} - \hat{G}\hat{f}^L\right)\hat{s} - f^K\hat{G}\left(\hat{G}^L\hat{f} - \hat{G}\hat{f}^L\right)\hat{s}].\end{aligned} \tag{A.2}$$

Similarly, replacing $\hat{G}^L$ by $\hat{G}^L - G^L + G^L$ and $\hat{f}^L$ by $\hat{f}^L - f^L + f^L$, results in

$$\begin{aligned}\hat{M}_{KL} = \frac{1}{N}\textstyle\sum_{i=1}^{N}\hat{f}^{-4}\Big[&(\hat{G}^K - G^K)\hat{f}^2\left(\hat{G}^L - G^L\right)\hat{s} + (\hat{G}^K - G^K)\hat{f}^2G^L\hat{s} \\ &- (\hat{G}^K - G^K)\hat{f}\hat{G}(\hat{f}^L - f^L)\hat{s} - (\hat{G}^K - G^K)\hat{f}\hat{G}f^L\hat{s} \\ &+ G^K\hat{f}^2(\hat{G}^L - G^L)\hat{s} + G^K\hat{f}^2G^L\hat{s} - G^K\hat{f}\hat{G}(\hat{f}^L - f^L)\hat{s} - G^K\hat{f}\hat{G}f^L\hat{s} \\ &- (\hat{f}^K - f^K)\hat{G}\hat{f}(\hat{G}^L - G^L)\hat{s} - (\hat{f}^K - f^K)\hat{G}\hat{f}G^L\hat{s} \\ &+ (\hat{f}^K - f^K)\hat{G}^2(\hat{f}^L - f^L)\hat{s} + (\hat{f}^K - f^K)\hat{G}^2f^L\hat{s} \\ &- f^K\hat{G}\hat{f}(\hat{G}^L - G^L)\hat{s} - f^K\hat{G}\hat{f}G^L\hat{s} + f^K\hat{G}^2(\hat{f}^L - f^L)\hat{s} + f^K\hat{G}^2f^L\hat{s}\Big].\end{aligned} \tag{A.3}$$

Substituting $\hat{G} = \hat{G} - G + G$ and $\hat{f} = \hat{f} - f + f$ in the numerator and using

$$\hat{f}^{-4} = f^{-4}[1 - 4\left(\tfrac{\hat{f}-f}{\hat{f}}\right) + 6\left(\tfrac{\hat{f}-f}{\hat{f}}\right)^2 - 4\left(\tfrac{\hat{f}-f}{\hat{f}}\right)^3 + \left(\tfrac{\hat{f}-f}{\hat{f}}\right)^4] \tag{A.4}$$

we obtain

$$\hat{M}_{KL} = \frac{1}{N}\sum_{i=1}^{N} f^{-4}[1 - 4\left(\tfrac{\hat{f}-f}{\hat{f}}\right) + 6\left(\tfrac{\hat{f}-f}{\hat{f}}\right)^2 - 4\left(\tfrac{\hat{f}-f}{\hat{f}}\right)^3 + \left(\tfrac{\hat{f}-f}{\hat{f}}\right)^4]\times \tag{A.5}$$

$$\begin{aligned}
&[(\hat{G}^K - G^K)(\hat{f}-f)^2(\hat{G}^L - G^L)\hat{s} + (\hat{G}^K - G^K)2(\hat{f}-f)f(\hat{G}^L - G^L)\hat{s} \\
&+ (\hat{G}^K - G^K)f^2(\hat{G}^L - G^L)\hat{s} + (\hat{G}^K - G^K)(\hat{f}-f)^2 G^L\hat{s} \\
&+ (\hat{G}^K - G^K)2(\hat{f}-f)fG^L\hat{s} + (\hat{G}^K - G^K)f^2G^L\hat{s} \\
&- (\hat{G}^K - G^K)(\hat{f}-f)(\hat{G}-G)(\hat{f}^L - f^L)\hat{s} - (\hat{G}^K - G^K)f(\hat{G}-G)(\hat{f}^L - f^L)\hat{s} \\
&- (\hat{G}^K - G^K)(\hat{f}-f)G(\hat{f}^L - f^L)\hat{s} - (\hat{G}^K - G^K)fG(\hat{f}^L - f^L)\hat{s} \\
&- (\hat{G}^K - G^K)(\hat{f}-f)(\hat{G}-G)f^L\hat{s} - (\hat{G}^K - G^K)f(\hat{G}-G)f^L\hat{s} \\
&- (\hat{G}^K - G^K)(\hat{f}-f)Gf^L\hat{s} - (\hat{G}^K - G^K)fGf^L\hat{s} + G^K(\hat{f}-f)^2(\hat{G}^L - G^L)\hat{s} \\
&+ G^K 2(\hat{f}-f)f(\hat{G}^L - G^L)\hat{s} + G^K f^2(\hat{G}^L - G^L)\hat{s} + G^K(\hat{f}-f)^2G^L\hat{s} \\
&+ G^K 2(\hat{f}-f)fG^L\hat{s} + G^K f^2G^L\hat{s} - G^K(\hat{f}-f)(\hat{G}-G)(\hat{f}^L - f^L)\hat{s} \\
&- G^K f(\hat{G}-G)(\hat{f}^L - f^L)\hat{s} - G^K(\hat{f}-f)G(\hat{f}^L - f^L)\hat{s} - G^K fG(\hat{f}^L - f^L)\hat{s} \\
&- G^K(\hat{f}-f)(\hat{G}-G)f^L\hat{s} - G^K f(\hat{G}-G)f^L\hat{s} - G^K(\hat{f}-f)Gf^L\hat{s} - G^K fGf^L\hat{s} \\
&- (\hat{f}^K - f^K)(\hat{G}-G)(\hat{f}-f)(\hat{G}^L - G^L)\hat{s} - (\hat{f}^K - f^K)(\hat{G}-G)f(\hat{G}^L - G^L)\hat{s} \\
&- (\hat{f}^K - f^K)G(\hat{f}-f)(\hat{G}^L - G^L)\hat{s} - (\hat{f}^K - f^K)Gf(\hat{G}^L - G^L)\hat{s} \\
&- (\hat{f}^K - f^K)(\hat{G}-G)(\hat{f}-f)G^L\hat{s} - (\hat{f}^K - f^K)(\hat{G}-G)fG^L\hat{s} \\
&- (\hat{f}^K - f^K)G(\hat{f}-f)G^L\hat{s} - (\hat{f}^K - f^K)GfG^L\hat{s} + (\hat{f}^K - f^K)(\hat{G}-G)^2(\hat{f}^L - f^L)\hat{s} \\
&+ (\hat{f}^K - f^K)2(\hat{G}-G)G(\hat{f}^L - f^L)\hat{s} + (\hat{f}^K - f^K)G^2(\hat{f}^L - f^L)\hat{s} \\
&+ (\hat{f}^K - f^K)(\hat{G}-G)^2f^L\hat{s} + (\hat{f}^K - f^K)2(\hat{G}-G)Gf^L\hat{s} + (\hat{f}^K - f^K)G^2f^L\hat{s} \\
&- f^K(\hat{G}-G)(\hat{f}-f)(\hat{G}^L - G^L)\hat{s} - f^K(\hat{G}-G)f(\hat{G}^L - G^L)\hat{s} \\
&- f^KG(\hat{f}-f)(\hat{G}^L - G^L)\hat{s} - f^KGf(\hat{G}^L - G^L)\hat{s} \\
&- f^K(\hat{G}-G)(\hat{f}-f)G^L\hat{s} - f^K(\hat{G}-G)fG^L\hat{s} - f^KG(\hat{f}-f)G^L\hat{s} - f^KGfG^L\hat{s} \\
&+ f^K(\hat{G}-G)^2(\hat{f}^L - f^L)\hat{s} + f^K 2(\hat{G}-G)G(\hat{f}^L - f^L)\hat{s} + f^KG^2(\hat{f}^L - f^L)\hat{s} \\
&+ f^K(\hat{G}-G)^2f^L\hat{s} + f^K 2(\hat{G}-G)Gf^L\hat{s} + f^KG^2f^L\hat{s}].
\end{aligned}$$

Finally, we substitute $\hat{s} = \hat{s} - s + s \equiv (\hat{s} - s + s)I^*$, where $I^* = I(\hat{f}(x) \geq b \text{ or } f(x) \geq b)$.

The result is given below in (A.7), where the following notation (similar to Stoker (1991)) is used

$$\begin{aligned}
&\zeta_f = f^{-1}\left[\hat{f} - f\right] I^*;\ \zeta_{GK} = f^{-1}\left[\hat{G}^K - G^K\right] I^*; \zeta_{fK} = f^{-1}\left[\hat{f}^K - f^K\right] I^*; \\
&\zeta_G = f^{-1}\left[\hat{G} - G\right] I^*;\ \zeta_{\hat{f}} = \hat{f}^{-1}\left[\hat{f} - f\right] I^*; \\
&\zeta_{\hat{f}\hat{f}} = \left[6(\tfrac{\hat{f}-f}{\hat{f}})^2 - 4(\tfrac{\hat{f}-f}{\hat{f}})^3 + (\tfrac{\hat{f}-f}{\hat{f}})^4\right] I^*;\ \text{and} \\
&\zeta_s = [\hat{s} - s]\, I^*.
\end{aligned} \tag{A.6}$$

The final rewritten version of $\hat{M}_{KL}$, thus obtained, is then given by

$$\begin{aligned}
\hat{M}_{KL} = \frac{1}{N}\textstyle\sum_{i=1}^{N} &\left[1 - 4\zeta_{\hat{f}} + \zeta_{\hat{f}\hat{f}}\right][\zeta_s + s] \times \\
&\big[\zeta_{GK}\zeta_f^2\zeta_{GL} + 2\zeta_{GK}\zeta_f\zeta_{GL} + \zeta_{GK}\zeta_{GL} + \zeta_{GK}\zeta_f^2 G^L/f + 2\zeta_{GK}\zeta_f G^L/f + \zeta_{GK}G^L/f \\
&- \zeta_{GK}\zeta_f\zeta_G\zeta_{fL} - \zeta_{GK}\zeta_G\zeta_{fL} - \zeta_{GK}\zeta_f\zeta_{fL}G/f - \zeta_{GK}\zeta_{fL}G/f - \zeta_{GK}\zeta_f\zeta_G f^L/f \\
&- \zeta_{GK}\zeta_G f^L/f - \zeta_{GK}\zeta_f G f^L/f^2 - \zeta_{GK}G f^L/f^2 + \zeta_f^2\zeta_{GL}G^K/f + 2\zeta_f\zeta_{GL}G^K/f \\
&+ \zeta_{GL}G^K/f + \zeta_f^2 G^K G^L/f^2 + 2\zeta_f G^K G^L/f^2 + G^K G^L/f^2 - \zeta_f\zeta_G\zeta_{fL}G^K/f - \zeta_G\zeta_{fL}G^K/f \\
&- \zeta_f\zeta_{fL}G^K G/f^2 - \zeta_{fL}G^K G/f^2 - \zeta_f\zeta_G G^K f^L/f^2 - \zeta_G G^K f^L/f^2 - \zeta_f G^K G f^L/f^3 \\
&- G^K G f^L/f^3 - \zeta_{fK}\zeta_G\zeta_f\zeta_{GL} - \zeta_{fK}\zeta_G\zeta_{GL} - \zeta_{fK}\zeta_f\zeta_{GL}G/f - \zeta_{fK}\zeta_{GL}G/f \\
&- \zeta_{fK}\zeta_G\zeta_f G^L/f - \zeta_{fK}\zeta_G G^L/f - \zeta_{fK}\zeta_f G G^L/f^2 - \zeta_{fK}G G^L/f^2 + \zeta_{fK}\zeta_G^2\zeta_{fL} \\
&+ 2\zeta_{fK}\zeta_G\zeta_{fL}G/f + \zeta_{fK}\zeta_{fL}G^2/f^2 + \zeta_{fK}\zeta_G^2 f^L/f + 2\zeta_{fK}\zeta_G G f^L/f^2 + \zeta_{fK}G^2 f^L/f^3 \\
&- \zeta_G\zeta_f\zeta_{GL}f^K/f - \zeta_G\zeta_{GL}f^K/f - \zeta_f\zeta_{GL}f^K G/f^2 - \zeta_{GL}f^K G/f^2 - \zeta_G\zeta_f f^K G^L/f^2 \\
&- \zeta_G f^K G^L/f^2 - \zeta_f f^K G G^L/f^3 - f^K G G^L/f^3 + \zeta_G^2\zeta_{fL}f^K/f + 2\zeta_G\zeta_{fL}f^K G/f^2 \\
&+ \zeta_{fL}f^K G^2/f^3 + \zeta_G^2 f^K f^L/f^2 + 2\zeta_G f^K G f^L/f^3 + f^K G^2 f^L/f^4\big].
\end{aligned} \tag{A.7}$$

The linearized version, $\tilde{M}_{KL}$, see equation (12), can similarly be rewritten as

$$\begin{aligned}\tilde{M}_{KL} =& \frac{1}{N}\sum_{i=1}^{N}[s+\zeta_s]\times[G^KG^L/f^2 - Gf^KG^L/f^3 - G^KGf^L/f^3 + G^2f^Kf^L/f^4] \\ &+ s\times[\zeta_{GK}G^L/f - \zeta_{GK}Gf^L/f^2 - \zeta_{fK}GG^L/f^2 + \zeta_{fK}G^2f^L/f^3 \\ &- \zeta_{fL}GG^K/f^2 + \zeta_{fL}G^2f^K/f^3 + \zeta_{GL}G^K/f - \zeta_{GL}Gf^K/f^2 \\ &- \zeta_Gf^KG^L/f^2 + \zeta_Gf^KGf^L/f^3 - \zeta_Gf^LG^K/f^2 + \zeta_Gf^LGf^K/f^3 \\ &+ \zeta_fG^KG^L/f^2 - \zeta_fG^KGf^L/f^3 + \zeta_fG^LG^K/f^2 - \zeta_fG^LGf^K/f^3] \\ &- 4s\zeta_f\times[G^KG^L/f^2 - Gf^KG^L/f^3 - G^KGf^L/f^3 + G^2f^Kf^L/f^4].\end{aligned} \tag{A.8}$$

The first five lines of this expression also appear in the expression for $\hat{M}_{KL}$ in (A.7). The elements in the last line also appear there, except that ζ_f is replaced by $\zeta_{\hat{f}}$. Since $\zeta_f - \zeta_{\hat{f}} = \zeta_f\zeta_{\hat{f}}$ this results in $\sqrt{N}\left(\hat{M}_{KL} - \tilde{M}_{KL}\right)$ equalling second and higher order polynomial terms in the ζ's only.

Since the set $\{x|\hat{f}(x) \geq b \text{ or } f(x) \geq b\}$ is compact and $h \to 0$, following the arguments of Collomb and Härdle (1986) or Silverman (1978), we can assert as Härdle and Stoker (1989)

$$\begin{aligned}\sup|\zeta_f| &= O_p\left[\left(N^{1-\varepsilon/2}h^k\right)^{-1/2}\right] \\ \sup|\zeta_{GK}| &= O_p\left[\left(N^{1-\varepsilon/2}h^{k+2}\right)^{-1/2}\right] \\ \sup|\zeta_{fK}| &= O_p\left[\left(N^{1-\varepsilon/2}h^{k+2}\right)^{-1/2}\right] \\ \sup|\zeta_G| &= O_p\left[\left(N^{1-\varepsilon/2}h^k\right)^{-1/2}\right],\end{aligned} \tag{A.9}$$

and, using $N^{1-\varepsilon/2}h^k \to \infty$,

$$\begin{aligned}\sup\left|\zeta_{\hat{f}}\right| &= O_p\left[\left(N^{1-\varepsilon/2}h^k\right)^{-1/2}\right] \\ \sup\left|\zeta_{\hat{f}\hat{f}}\right| &= O_p\left[\left(N^{1-\varepsilon/2}h^k\right)^{-1}\right].\end{aligned} \tag{A.10}$$

By continuity of $s(\cdot)$, and given the boundedness of s' and s'' a second order Taylor expansion of $\hat{s}$ around f reveals

$$\sup|\zeta_s| = O_p\left[\left(N^{1-\varepsilon/2}h^k\right)^{-1/2}\right] \tag{A.11}$$

since $\sup|\zeta_s| \leq \sup|s'| \sup\left|\hat{f}-f\right| + \sup|s''| \sup\left|\hat{f}-f\right|^2$. These results allow us to derive the rate of convergence of $\sqrt{N}\left(\hat{M}_{KL}-\tilde{M}_{KL}\right)$ by analyzing each term in $\sqrt{N}\left(\hat{M}_{KL}-\tilde{M}_{KL}\right)$ individually. Take, for example, $\frac{1}{\sqrt{N}}\sum \zeta_{fK}\zeta_{fL}G^2/f^2 \equiv \frac{1}{\sqrt{N}}\sum \zeta_{fK}\zeta_{fL}IG^2/f^2$. It follows that $\left|\frac{1}{\sqrt{N}}\sum \zeta_{fK}\zeta_{fL}G^2/f^2\right| \leq \sqrt{N}\sup|\zeta_{fK}|\sup|\zeta_{fL}|\,\frac{1}{N}\sum|I^*G^2/f^2| = O_p(N^{1/2}N^{-1+\varepsilon/2}h^{-(k+2)}) = o_p(1)$, since $\frac{1}{N}\sum|I^*G^2/f^2|$ is bounded by Chebychev's inequality and $N^{1-\varepsilon}h^{2k+4} \to \infty$ by condition (ii).

The other terms in $\sqrt{N}\left(\hat{M}_{KL}-\tilde{M}_{KL}\right)$ are analyzed similarly, allowing us to conclude $\sqrt{N}\left|\hat{M}_{KL}-\tilde{M}_{KL}\right| = O_p(N^{1/2}N^{-1+\varepsilon/2}h^{-(k+2)}) = O_p(N^{-1/2+\varepsilon/2}h^{-(k+2)}) = o_p(1)$. $\square$

Proof of Lemma 2

To prove that $\sqrt{N}(\tilde{M}_{KL}-E(\tilde{M}_{KL})) = N^{-1/2}(\sum_{i=1}^N r(z_i)-E(r(z)))+o_p(1)$ we rewrite the linearized version of $\hat{M}_{KL}$, $\tilde{M}_{KL}$, as the sum of "average kernel estimators"

$$\tilde{M}_{KL} = \tilde{\delta}_0 + \tilde{\delta}_1 - \tilde{\delta}_2 + \tilde{\delta}_3 - \tilde{\delta}_4 - \tilde{\delta}_5 + \tilde{\delta}_6 + \tilde{\delta}_7, \tag{A.12}$$

with

$$\begin{aligned}
\tilde{\delta}_0 &= \frac{1}{N}\sum_{i=1}^N \left[g'g'^T\right]_{KL} s \\
\tilde{\delta}_1 &= \frac{1}{N}\sum_{i=1}^N \hat{G}^K \frac{g^L}{f} s;\ \tilde{\delta}_2 = \frac{1}{N}\sum_{i=1}^N \hat{f}^K \frac{gg^L}{f} s \\
\tilde{\delta}_3 &= \frac{1}{N}\sum_{i=1}^N \hat{G}^L \frac{g^K}{f} s;\ \tilde{\delta}_4 = \frac{1}{N}\sum_{i=1}^N \hat{f}^L \frac{gg^K}{f} s \\
\tilde{\delta}_5 &= \frac{1}{N}\sum_{i=1}^N \hat{G} \frac{f^K g^L + f^L g^K}{f^2} s;\ \tilde{\delta}_6 = \frac{1}{N}\sum_{i=1}^N \hat{f} \frac{-2g^L g^K f + g(f^K g^L + f^L g^K)}{f^2} s \\
\tilde{\delta}_7 &= \frac{1}{N}\sum_{i=1}^N (\hat{s}-s)\left[g'g'^T\right]_{KL}.
\end{aligned} \tag{A.13}$$

Notice that in rewriting many of the non-estimated terms in (12) cancel out. U-statistic theory, see Hoeffding (1948), can then be used to show that each of the $\widetilde{\delta}$'s is $\sqrt{N}$ equivalent to an ordinary sample average.

For $\tilde{\delta}_0$ this follows straightforward, since

$$\sqrt{N}(\tilde{\delta}_0 - E(\tilde{\delta}_0)) = N^{-1/2}(\sum_{i=1}^N r_0(x_i) - E(r_0(x))), \tag{A.14}$$

where $r_0(x) = \left[g'(x)g'(x)^T\right]_{KL} s_x$.

The analysis for $\tilde{\delta}_1$ through $\tilde{\delta}_7$ is fairly similar. We focus our discussion primarily on $\tilde{\delta}_1$, while pointing out the more interesting modifications for the other $\tilde{\delta}$'s.

Using a leave-one-out nonparametric estimator for $\hat{G}_i^K$, we note that $\tilde{\delta}_1$ can be written as the following second order U-statistic

$$U_1 = \binom{N}{2}^{-1} \sum_{i=1}^{N-1} \sum_{j=i+1}^{N} p_{1N}(z_i, z_j), \tag{A.15}$$

where $z_i = (y_i, x_i^T)^T$ and

$$p_{1N}(z_i, z_j) = \frac{1}{2}\left(\frac{1}{h}\right)^{k+1} K^K\left(\frac{x_i - x_j}{h}\right)\left(y_j \frac{g_i^L}{f_i} s_i - y_i \frac{g_j^L}{f_j} s_j\right). \tag{A.16}$$

K^K denotes the first derivative of the kernel K with respect to the K^{th} argument. We assume a symmetric kernel, hence $K^K(-u) = -K^K(u)$. Using Lemma 3.1 from Powell, Stock and Stoker (1989),

$$\sqrt{N}(\tilde{\delta}_1 - E(\tilde{\delta}_1)) = N^{-1/2}\left(\sum_{i=1}^{N} r_{1N}(z_i) - E(r_{1N}(z))\right) + o_p(1), \tag{A.17}$$

where $r_{1N}(z) = 2E(p_{1N}(z, z_j)|z)$, provided $E\left(|p_{1N}(z_i, z_j)|^2\right) = o(N)$. This condition is verified next. Let $M_2(x_i) = E(y_i^2|x_i)$, then

$$\begin{aligned}
&E\left(|p_{1N}(z_i, z_j)|^2\right) \\
&\leq \frac{1}{4b^2 h^{2(k+1)}} \int \left|K^K\left(\frac{x_i - x_j}{h}\right)\right|^2 [M_2(x_j)(g^L(x_i)s_{x_i})^2 + M_2(x_i)(g^L(x_j)s_{x_j})^2 \\
&\quad - 2g(x_j)g(x_i)(g^L(x_i)s_{x_i})(g^L(x_j)s_{x_j})]f(x_i)f(x_j)dx_i dx_j \qquad \text{(A.18)} \\
&= \frac{1}{4b^2 h^{k+2}} \int \left|K^K(u)\right|^2 [M_2(x_i + uh)(g^L(x_i)s_{x_i})^2 + M_2(x_i)(g^L(x_i + uh)s_{x_i+uh})^2 \\
&\quad - 2g(x_i + uh)g(x_i)(g^L(x_i)s_{x_i})(g^L(x_i + uh)s_{x_i+uh})]f(x_i)f(x_i + uh)dx_i du \\
&= O(h^{-(k+2)}) = O(N(Nh^{k+2})^{-1}) = o(N)
\end{aligned}$$

by continuity and $Nh^{k+2} \to \infty$ is implied by condition (ii).

Using (A.16), we find

$$\begin{aligned} r_{1N}(z_i) = & \left(\tfrac{1}{h}\right)^{k+1} \int K^K\left(\tfrac{x_i-x}{h}\right)\left(g(x)\tfrac{g_i^L}{f_i}s_i - y_i\tfrac{g^L(x)}{f(x)}s_x\right) f(x)dx \\ = & \left(\tfrac{1}{h}\right)^{k+1}\left[\tfrac{g_i^L}{f_i}s_i\int K^K\left(\tfrac{x_i-x}{h}\right)(gf)(x)dx - y_i\int K^K\left(\tfrac{x_i-x}{h}\right)g^L(x)s_x dx\right] \\ = & -\left(\tfrac{1}{h}\right)\left[\tfrac{g_i^L}{f_i}s_i\int K^K(u)(gf)(x_i+uh)du - y_i\int K^K(u)\,g^L(x_i+uh)s_{x_i+uh}du\right] \\ = & \tfrac{g_i^L}{f_i}s_i\int K(u)(gf)^K(x_i+uh)du \\ & -y_i\int K(u)\left[g^{LK}(x_i+uh)s_{x_i+uh} + g^L(x_i+uh)s^K_{x_i+uh}\right]du. \end{aligned} \tag{A.19}$$

The third equality applies a change of variable $x = x_i + uh$ and the symmetry of $K(\cdot)$, the last equality integrates by parts the first two terms making use of the fact that the kernel vanishes at the boundary. Note $s^K_{x_i} = s'_{x_i}f^K(x_i)$, where $s'_x = \partial s(\tau)/\partial\tau$, evaluated at $\tau = f(x) - b$. We now separate $r_{1N}(z_i)$ into two parts: $r_1(z_i)$, independent of N, and $t_{1N}(z_i)$, with

$$\begin{aligned} r_1(z_i) &= \left(\frac{g^L(x_i)}{f(x_i)}(gf)^K(x_i) - y_i g^{LK}(x_i)\right)s_{x_i} - y_i(g^L f^K)(x_i)s'_{x_i} \\ &= \left((g^L g^K)(x_i) + (gg^L f^K)(x_i)/f(x_i) - y_i g^{LK}(x_i)\right)s_{x_i} - y_i(g^L f^K)(x_i)s'_{x_i} \end{aligned} \tag{A.20}$$

$$\begin{aligned} t_{1N}(z_i) = & \tfrac{g^L(x_i)}{f(x_i)}s_i\int K(u)\left[(gf)^K(x_i+uh) - (gf)^K(x_i)\right]du + \\ & -y_i\int K(u)\left[g^{LK}(x_i+uh)s_{x_i+uh} - g^{LK}(x_i)s_{x_i}\right]du + \\ & -y_i\int K(u)\left[(g^L f^K)(x_i+uh)s'_{x_i+uh} - (g^L f^K)(x_i)s'_{x_i}\right]du. \end{aligned} \tag{A.21}$$

As $r_{1N}(z_i) = r_1(z_i) + t_{1N}(z_i)$,

$$\begin{aligned} N^{-1/2}\left(\textstyle\sum_{i=1}^N r_{1N}(z_i) - E(r_{1N}(z))\right) = & N^{-1/2}\left(\textstyle\sum_{i=1}^N r_1(z_i) - E(r_1(z))\right) \\ & + N^{-1/2}\left(\textstyle\sum_{i=1}^N t_{1N}(z_i) - E(t_{1N}(z))\right). \end{aligned} \tag{A.22}$$

To show that the last term converges in probability to zero, we show that its variance is bounded by $E(t_{1N}(z)^2) = o(1)$. Lipschitz conditions on $(gf)^K$, $g^{LK}s$ and $g^L f^K s'$ (implied by Lipschitz conditions on g^{LK}, g^K, and f^K, and Assumption 5) guarantee that the second moments of the three terms in (A.21) vanish in probability. For example, by Assumption 6, the second moment of the first term on the right hand side of (A.21) is bounded

by $h^2\left(\int |u|\, K(u)du\right)^2 E(g^L(x)^2 s_x^2 \psi^2_{(gf)^K}(x))$, which, by our assumptions, is $O(h^2) = o(1)$. To allow $b \to 0$ as $N \to \infty$ in the trimming function, an assumption on the existence of $E(g^L(x)^2\psi^2_{(gf)^K}(x))$ will be required, as in Härdle and Stoker (1989). To conclude, $E(t_{1N}(z)^2) = o(1)$, so

$$\sqrt{N}(\tilde{\delta}_1 - E(\tilde{\delta}_1)) = N^{-1/2}(\sum_{i=1}^{N} r_1(z_i) - E(r_1(z))) + o_p(1), \tag{A.23}$$

where $r_1(z) = \left((g^L g^K)(x) + (gg^L f^K)(x)/f(x) - yg^{LK}(x)\right) s_x - y(g^L f^K)(x)s'_x$.

Using a leave-one-out nonparametric estimator for $\widehat{f}_i^K$, $\tilde{\delta}_2$ can be written as the following second order U-statistic

$$U_2 = \binom{N}{2}^{-1} \sum_{i=1}^{N-1} \sum_{j=i+1}^{N} p_{2N}(z_i, z_j), \tag{A.24}$$

and

$$p_{2N}(z_i, z_j) = \frac{1}{2}\left(\frac{1}{h}\right)^{k+1} K^K\left(\frac{x_i - x_j}{h}\right)\left(\frac{g_i g_i^L}{f_i}s_i - \frac{g_j g_j^L}{f_j}s_j\right). \tag{A.25}$$

A similar analysis as for $\tilde{\delta}_1$ reveals that, under the appropriate assumptions, such as Lipschitz conditions on f^K, $(gg^L)^K s$, and $gg^L f^K s'$ (implied by Lipschitz conditions on g, f^K, g^K, and g^{KL} and Assumption 5),

$$\sqrt{N}(\tilde{\delta}_2 - E(\tilde{\delta}_2)) = N^{-1/2}(\sum_{i=1}^{N} r_2(z_i) - E(r_2(z))) + o_p(1), \tag{A.26}$$

where $r_2(z) = (\left(gg^L f^K)(x)/f(x) - (g^K g^L)(x) - (gg^{LK})(x)\right) s_x - (gg^L f^K)(x)s'_x$.

By symmetry

$$\sqrt{N}(\tilde{\delta}_k - E(\tilde{\delta}_k)) = N^{-1/2}(\sum_{i=1}^{N} r_k(z_i) - E(r_k(z))) + o_p(1),\ k = 3, 4 \tag{A.27}$$

with

$$r_3(z) = \left((g^K g^L)(x) + (gg^K f^L)(x)/f(x) - yg^{KL}(x)\right) s_x - y(g^K f^L)(x)s'_x \tag{A.28}$$

$$r_4(z) = \left((gg^K f^L)(x)/f(x) - (g^K g^L)(x) - (gg^{LK})(x)\right) s_x - (gg^K f^L)(x)s'_x. \tag{A.29}$$

The analysis for $\tilde{\delta}_5$ and $\tilde{\delta}_6$ is very similar again. Using leave-one-out nonparametric estimators for $\hat{G}_i$ and $\widehat{f}_i$, $\tilde{\delta}_5$ and $\tilde{\delta}_6$ can be written as the following second order U-statistics

$$U_k = \binom{N}{2}^{-1} \sum_{i=1}^{N-1} \sum_{j=i+1}^{N} p_{kN}(z_i, z_j),\ k = 5, 6 \tag{A.30}$$

with

$$p_{5N}(z_i, z_j) = \frac{1}{2}\left(\frac{1}{h}\right)^k K\left(\frac{x_i - x_j}{h}\right)[y_j \frac{f_i^K g_i^L + f_i^L g_i^K}{f_i^2} s_i + y_i \frac{f_j^K g_j^L + f_j^L g_j^K}{f_j^2} s_j] \tag{A.31}$$

$$\begin{aligned} p_{6N}(z_i, z_j) = \frac{1}{2}\left(\frac{1}{h}\right)^k K\left(\frac{x_i - x_j}{h}\right)[&- 2\frac{g_i^L g_i^K}{f_i} s_i + \frac{f_i^K g_i g_i^L}{f_i^2} s_i + \frac{f_i^L g_i g_i^K}{f_i^2} s_i \\ &- 2\frac{g_j^L g_j^K}{f_j} s_j + \frac{f_j^K g_j g_j^L}{f_j^2} s_j + \frac{f_j^L g_j g_j^K}{f_j^2} s_j]. \end{aligned} \tag{A.32}$$

Under our assumptions, we obtain

$$\sqrt{N}(\tilde{\delta}_k - E(\tilde{\delta}_k)) = N^{-1/2}(\sum_{i=1}^N r_k(z_i) - E(r_k(z))) + o_p(1),\ k = 5, 6, \tag{A.33}$$

where

$$r_5(z) = [g(x) + y][(f^K(x)g^L(x) + f^L(x)g^K(x))/f(x)]s_x \tag{A.34}$$

$$r_6(z) = 2[-2g^L(x)g^K(x) + g(x)(f^K(x)g^L(x) + f^L(x)g^K(x))/f(x)]s_x. \tag{A.35}$$

Finally, we consider $\tilde{\delta}_7 = \frac{1}{N}\sum_{i=1}^N [\hat{s} - s]\, g^K g^L$. A second order Taylor expansion of $\hat{s}$ around $f(x)$, gives

$$\tilde{\delta}_7 = \frac{1}{N}\sum_{i=1}^N \left(s'(\hat{f} - f) + \tfrac{1}{2}s''(f^*)(\hat{f} - f)^2\right) g^K g^L, \tag{A.36}$$

where f^* lies between f and $\hat{f}$. Given s'' is bounded, (A.9) gives

$$\tilde{\delta}_7^* \equiv \frac{1}{N}\sum_{i=1}^N s'(\hat{f} - f) g^K g^L = \tilde{\delta}_7 + o_p(N^{-1/2}). \tag{A.37}$$

We now note that $\tilde{\delta}_7^*$, using a leave-one-out nonparametric estimator for $\widehat{f_i}$, can be written as the following second order U-statistic

$$U_7 = \binom{N}{2}^{-1} \sum_{i=1}^{N-1} \sum_{j=i+1}^{N} p_{7N}(z_i, z_j), \tag{A.38}$$

and

$$\begin{aligned} p_{7N}(z_i, z_j) \ &= \tfrac{1}{2}\left[\left(\frac{1}{h}\right)^k K\left(\frac{x_i - x_j}{h}\right) - f_i\right] g_i^L g_i^K s_i' + \\ &\quad \tfrac{1}{2}\left[\left(\frac{1}{h}\right)^k K\left(\frac{x_i - x_j}{h}\right) - f_j\right] g_j^L g_j^K s_j'. \end{aligned} \tag{A.39}$$

Using a similar approach as before, it can be shown that

$$\sqrt{N}(\tilde{\delta}_7^* - E(\tilde{\delta}_7^*)) = N^{-1/2}(\sum_{i=1}^{N} r_7(z_i) - E(r_7(z))) + o_p(1), \tag{A.40}$$

where

$$r_7(z) = (g^K g^L f)(x)s'_x - E((g^K g^L f)(x)s'_x). \tag{A.41}$$

In other words, we argue that the difference between trimming on the basis of the estimated density relative to the true density is not negligible, unlike Härdle and Stoker (1989) and Stoker (1991).

From (A.14), (A.23), (A.26), (A.27), (A.33), (A.37) and (A.40), finally, we conclude that $\tilde{M}_{KL} \equiv \tilde{\delta}_0 + \tilde{\delta}_1 - \tilde{\delta}_2 + \tilde{\delta}_3 - \tilde{\delta}_4 - \tilde{\delta}_5 + \tilde{\delta}_6 + \tilde{\delta}_7$ has the representation

$$\sqrt{N}(\tilde{M}_{KL} - E(\tilde{M}_{KL})) = N^{-1/2}(\sum_{i=1}^{N} r(z_i) - E(r(z)))) + o_p(1), \tag{A.42}$$

where

$$\begin{aligned} r(z_i) &= r_0(z_i) + r_1(z_i) - r_2(z_i) + r_3(z_i) - r_4(z_i) - r_5(z_i) + r_6(z_i) + r_7(z_i) \\ &= g^L(x_i)g^K(x_i)s_{x_i} + 2(g(x_i) - y_i)g^{KL}(x_i)s_{x_i} + \\ &(g(x_i) - y_i)\frac{f^K(x_i)g^L(x_i) + f^L(x_i)g^K(x_i)}{f(x_i)}s_{x_i} + \\ &(g(x_i) - y_i)(f^K(x_i)g^L(x_i) + f^L(x_i)g^K(x_i))s'_{x_i} + \\ &g^K(x_i)g^L(x_i)f(x_i)s'_{x_i} - E(g^K(x)g^L(x)f(x)s'_x). \quad \square \end{aligned} \tag{A.43}$$

The need for a smoothed trimming function instead of an indicator trimming function is best illustrated using the proof of Lemma 2 given above, as it affects the asymptotic properties of $\tilde{\delta}_1, \tilde{\delta}_2, \tilde{\delta}_3, \tilde{\delta}_4$, and $\tilde{\delta}_7$ while leaving the asymptotic properties of $\tilde{\delta}_5$ and $\tilde{\delta}_6$ unchanged. Indeed, only s_x appears in the asymptotic properties of $\tilde{\delta}_5$ and $\tilde{\delta}_6$, which can be replaced by the indicator function, while s'_x appears in the analysis of the other terms.

Let us return to the last equality in (A.19). With the use of the indicator function, our assumptions guarantee, for example, that

$$h^{-1}y_i \int K^K(u)\, g^L(x_i + uh) 1\left(f(x_i + uh) > b\right) du \tag{A.44}$$

converges to $-y_i g^{LK}(x_i)1(f(x_i) > b)$ almost everywhere. However, it does not do so near the boundaries, where the trimming function is active. Let us be more precise and rewrite (A.44) as

$$h^{-1}y_i \int_{u:1(f(x_i+uh)>b)} K^K(u)\, g^L(x_i+uh)du. \tag{A.45}$$

Integration by parts gives that (A.45) equals

$$h^{-1}y_i \left[K(u)g^L(x_i+uh)\right]_{u:1(f(x_i+uh)=b)} - y_i \int_{u:1(f(x_i+uh)>b)} K(u)\, g^{LK}(x_i+uh)du, \tag{A.46}$$

where the second term behaves nicely for $h \to 0$.

The first expression only behaves well when a bounded kernel is used and x_i is sufficiently far away from the boundary, as this leads to $K(u) = 0$ for $u \in \{u : 1(f(x_i + uh) = b)\}$. For x_i close to the boundary, or when an unbounded kernel is used, the behavior of $h^{-1}y_i \left[K(u)g^L(x_i+uh)\right]_{u:1(f(x_i+uh)=b)}$ is dominated by $h^{-1} \to \infty$. Consequently, there exists no square integrable function which bounds $\left|h^{-1}y_i \int K^K(u)\, g^L(x_i+uh)I_{x_i+uh}du\right|$, which is needed to ensure the $\sqrt{N}$ asymptotic normality, see also Assumption 1 of Proposition 6 in Lee (1995). However, at the same time the part of the support of x, Ω, that is close enough to the boundary to be affected, shrinks at a rate h. These two opposing effects result in a non-negligible effect on the asymptotic behavior of the estimator. Using the smoothed trimming function we are able to quantify this influence, as is shown by the presence of s' in the expression for the asymptotic covariance matrix.

Proof of Lemma 3

To prove that the bias of the estimator is $o(N^{-1/2})$, we rewrite $E\left\{\tilde{M}_{KL}\right\} - M_{KL}$ as $\tau_{1N} - \tau_{2N} + \tau_{3N} - \tau_{4N} - \tau_{5N} + \tau_{6N} + \tau_{7N}$, where

$$\begin{aligned}
\tau_{1N} &= E\left\{\left(\hat{G}^K - G^K\right)\frac{g^L}{f}s\right\};\ \tau_{2N} = E\left\{\left(\hat{f}^K - f^K\right)\frac{gg^L}{f}s\right\}; \\
\tau_{3N} &= E\left\{\left(\hat{G}^L - G^L\right)\frac{g^K}{f^2}s\right\};\ \tau_{4N} = E\left\{\left(\hat{f}^L - f^L\right)\frac{G}{f}\frac{g^K}{f}s\right\}; \\
\tau_{5N} &= E\left\{\left(\hat{G} - G\right)\left[\frac{f^K}{f^2}g^L + \frac{f^L}{f^2}g^K\right]s\right\}; \\
\tau_{6N} &= E\left\{\left(\hat{f} - f\right)\frac{-2g^Lg^Kf + g(f^Kg^L + f^Lg^K)}{f^2}s\right\};\ \tau_{7N} = E\left\{[\hat{s} - s]\, g^Kg^L\right\}.
\end{aligned} \tag{A.47}$$

We can analyze the bias of the estimator by analyzing the bias components individually. We will focus our attention on the first bias component, τ_{1N}. Let ι denote an index set $(l_1, \ldots, l_k)$, with $\sum_m l_m = p$. For a vector $u = (u_1, .., u_k)$ define $u^\iota = u_1^{l_1} \cdots u_k^{l_k}$ and $f_\iota^{(p)} = \partial^p f/(\partial u)^\iota$. Similar definitions hold for $G_\iota^{K(p)}$, $G_\iota^{(p)}$, and $f_\iota^{(p)}$.

Using this notation, we note

$$\begin{aligned}
\tau_{1N} &= E\left\{\left(\hat{G}^K - G^K\right)\frac{g^L}{f}s\right\} \\
&= \int\left[\frac{1}{h^k}\int\frac{\partial K\left(\frac{x_i - x_j}{h}\right)}{\partial x_{iK}}G(x_j)dx_j - G^K(x_i)\right]g^L(x_i)s_{x_i}dx_i \\
&= \int\left[\frac{1}{h}\int\frac{\partial K\left(u\right)}{\partial u_K}G(x_i - uh)du - G^K(x_i)\right]g^L(x_i)s_{x_i}dx_i \\
&= \int\int K\left(u\right)\left[G^K(x_i - uh) - G^K(x_i)\right]du\; g^L(x_i)s_{x_i}dx_i \qquad \text{(A.48)} \\
&= \int\int K\left(u\right)h^p\sum_\iota G_\iota^{K(p)}(\xi)u^\iota du\; g^L(x_i)s_{x_i}dx_i \\
&= h^p\int g^L(x_i)s_{x_i}\sum_\iota G_\iota^{K(p)}(x_i)\int K\left(u\right)u^\iota du\; dx_i \\
&+ h^p\int g^L(x_i)s_{x_i}\sum_\iota\int K\left(u\right)(G_\iota^{K(p)}(\xi) - G^{K(p)}(x_i))u^\iota du dx_i \\
&= O(h^p).
\end{aligned}$$

The third equality applies a change of variables, the fourth equality integrates by parts, using the fact that the kernel vanishes at the boundary. The fifth equality uses a Taylor approximation, where ξ is a point between x and $x + uh$. The last equality uses our assumption 7, cf Stoker assumption 6, specifically $G_\iota^{K(p)}$ is locally Hölder continuous and the existence of $p + \delta$ moments of $K(\cdot)$. The boundedness of our integrals is ensured by the presence of our trimming function, s_x, which guarantees integration over a bounded interval only, and the existence of the functions over which we want to integrate. Bounding conditions on these integrals will be required when allowing $b \to 0$ as $N \to \infty$, see Härdle and Stoker (1989).

After a similar analysis for $\tau_2, .., \tau_7$, we obtain that given assumptions 1–5, and 7, $E\left\{\tilde{M}_{KL}\right\} - M_{KL} = O(h^p) + o(N^{-1/2}) = O(N^{-1/2}(Nh^{2p})^{1/2}) + o(N^{-1/2})$. Condition (iii)

$Nh^{2p} \to 0$ ensures $E\left\{\tilde{M}_{KL}\right\} - M_{KL} = o(N^{-1/2})$. $\square$

Proof of Theorem 2

Because $\hat{M}$ is a consistent estimator for M, we focus on establishing that the leading term $\frac{1}{N}\sum_{i=1}^{N}\left(\text{vech}\,\widehat{R}(z_i)\,\text{vech}\,\widehat{R}(z_i)^T\right)$ is consistent for $E(\text{vech}\,R(z_i)\,\text{vech}\,R(z_i)^T)$. $\hat{R}(z_i)$ is the matrix containing as $(K,L)^{th}$ element the estimated elements $\hat{r}(z_i)$, with

$$r(z_i) = r_0(z_i) + r_1(z_i) - r_2(z_i) + r_3(z_i) - r_4(z_i) - r_5(z_i) + r_6(z_i) + r_7(z_i) \tag{A.49}$$

defined in the proof of Lemma 2. We note

$$\begin{aligned}
&N^{-1}\sum_{i=1}^{N}\left(\text{vech}\,\widehat{R}(z_i)\,\text{vech}\,\widehat{R}(z_i)^T\right) - E(\text{vech}\,R(z_i)\,\text{vech}\,R(z_i)^T) \qquad \text{(A.50)}\\
&= N^{-1}\sum_{i=1}^{N}\left(\text{vech}(\widehat{R}(z_i) - R(z_i))\,\text{vech}(\widehat{R}(z_i) - R(z_i))^T\right)\\
&+ N^{-1}\sum_{i=1}^{N}\left(\text{vech}\,R(z_i)\,\text{vech}(\widehat{R}(z_i) - R(z_i))^T\right)\\
&+ N^{-1}\sum_{i=1}^{N}\left(\text{vech}(\widehat{R}(z_i) - R(z_i))\,\text{vech}\,R(z_i)^T\right)\\
&+ N^{-1}\sum_{i=1}^{N}\text{vech}\,R(z_i)\,\text{vech}\,R(z_i)^T - E(R(z_i)R(z_i)^T).
\end{aligned}$$

Because of the existence of the variance of the elements in $R\,(z_i)$, the latter term converges to zero by Khinchine's law of large numbers. Once, we show that $E\left\|(\widehat{r}(z_i) - r(z_i))^2\right\| = o(1)$, Chebychev's and Hölders inequality give us the desired result that all term vanish asymptotically. In particular, $\forall\varepsilon > 0$,

$$\begin{aligned}
&\Pr\left(\left\|\frac{1}{N}\sum_{i=1}^{N}\left(\text{vech}(\widehat{R}(z_i) - R(z_i))\,\text{vech}(\widehat{R}(z_i) - R(z_i))^T\right)\right\| > \varepsilon\right) \qquad \text{(A.51)}\\
&\leq \frac{E\left\|\text{vech}(\widehat{R}(z_i) - R(z_i))\,\text{vech}(\widehat{R}(z_i) - R(z_i))^T\right\|}{\varepsilon}\\
&\leq \frac{\sqrt{\sum_{K=1}^{k}\sum_{L=1}^{K}E\left|(\widehat{r}(z) - r(z))^2\right|}}{\varepsilon},
\end{aligned}$$

where the second inequality uses the relation $\|A\| = \sqrt{tr(A'A)}$, with A a matrix. Now,

$$
\begin{aligned}
& E\left|(\widehat{r}(z) - r(z))^2\right| \qquad (A.52)\\
& \leq E\left|\hat{g}^L\hat{g}^K\hat{s} - g^L g^K s\right|^2 \\
& + 2E\left|\hat{g}\hat{g}^{KL}\hat{s} - g g^{KL} s\right|^2 + 2Ey^2\left|\hat{g}^{KL}\hat{s} - g^{KL}s\right|^2 \\
& + E\left|\hat{g}\left(\frac{\hat{f}^K}{\hat{f}}\hat{g}^L + \frac{\hat{f}^L}{\hat{f}}\hat{g}^K\right)\hat{s} - g\left(\frac{f^K}{f}g^L + \frac{f^L}{f}g^K\right)s\right|^2 \\
& + Ey^2\left|\left(\frac{\hat{f}^K}{\hat{f}}\hat{g}^L + \frac{\hat{f}^L}{\hat{f}}\hat{g}^K\right)\hat{s} - \left(\frac{f^K}{f}g^L + \frac{f^L}{f}g^K\right)s\right|^2 \\
& + E\left|\hat{g}\left(\hat{f}^K\hat{g}^L + \hat{f}^L\hat{g}^K\right)\hat{s}' - g\left(f^K g^L + f^L g^K\right)s'\right|^2 \\
& + Ey^2\left|\left(\hat{f}^K\hat{g}^L + \hat{f}^L\hat{g}^K\right)\hat{s}' - \left(f^K g^L + f^L g^K\right)s'\right|^2 \\
& + E\left|\hat{g}^K\hat{g}^L\hat{f}\hat{s}' - g^K g^L f s'\right|^2 + E\left|\overline{\hat{g}^K\hat{g}^L\hat{f}\hat{s}' - g^K g^L f s'}\right|^2 + E\left|g^K g^L f s' - E(g^K g^L f s')\right|^2.
\end{aligned}
$$

Using equalities A.9–A.11, it is easy to derive

$$
\begin{aligned}
&\sup\|[\hat{g} - g]I^*\| = O(\left(N^{1-\frac{\varepsilon}{2}}h^k\right)^{-1/2}) \qquad (A.53)\\
&\sup\left\|[\hat{g}^K - g^K]I^*\right\| = O(\left(N^{1-\frac{\varepsilon}{2}}h^{k+2}\right)^{-1/2}),
\end{aligned}
$$

where we recall $\hat{s} = \hat{s} - s + s \equiv (\hat{s} - s + s)I^*$, where $I^* = I(\hat{f}(x) \geq b \text{ or } f(x) \geq b)$. For the second order derivatives of the nonparametric regression estimator, we obtain

$$
\sup\left\|[\hat{g}^{KL} - g^{KL}]I^*\right\| = O(\left(N^{1-\frac{\varepsilon}{2}}h^{k+4}\right)^{-1/2}). \qquad (A.54)
$$

With these results and our equalities A.9–A.11, we obtain

$$
E\left|(\widehat{r}(z) - r(z))^2\right| = O(\left(N^{1-\frac{\varepsilon}{2}}h^{k+4}\right)^{-1}),
$$

where we note the boundedness of $E(y^2 I^*)$. Provided $N^{2-\varepsilon}h^{(2k+8)} \to \infty$, our result is established, which follows from our conditions (i), (ii) and $h \to 0$. $\square$

Table A.1: Testing the Rank of a Matrix. Trimming using the indicator function

Model 1: $y = (X\beta_1)(X\beta_2) + \varepsilon$

	Not Bias Corrected					*Bias Corrected*				
Rank	W	W_w	CRT	CRT_w	OI	W	W_w	CRT	CRT_w	OI
	$h_n = h_n^{gcv}$									
0	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.002	0.000	0.000
1	0.000	0.000	0.000	0.000	0.000	0.002	0.000	0.120	0.004	0.002
2	0.008	0.088	0.170	0.218	0.238	0.632	0.966	0.830	0.980	0.986
	$h_n = h_n^{gcv} n^{-\frac{1}{2}[1/11-1/15]}$									
0	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.002	0.000	0.000
1	0.000	0.000	0.000	0.000	0.002	0.008	0.008	0.194	0.008	0.014
2	0.004	0.058	0.106	0.148	0.160	0.610	0.966	0.764	0.984	0.978
	$h_n = h_n^{gcv} n^{-[1/11-1/15]}$									
0	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.004	0.000	0.000
1	0.000	0.000	0.002	0.000	0.000	0.020	0.006	0.226	0.032	0.022
2	0.006	0.052	0.058	0.094	0.104	0.446	0.952	0.700	0.948	0.970

Model 2: $y = (X\beta_1) + (X\beta_2)^2 + \varepsilon$

	Not Bias Corrected					*Bias Corrected*				
Rank	W	W_w	CRT	CRT_w	OI	W	W_w	CRT	CRT_w	OI
	$h_n = h_n^{gcv}$									
0	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
1	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.464	0.030	0.002
2	0.000	0.072	0.024	0.132	0.140	0.126	0.972	0.350	0.952	0.984
	$h_n = h_n^{gcv} n^{-\frac{1}{2}[1/11-1/15]}$									
0	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
1	0.000	0.000	0.000	0.000	0.000	0.006	0.004	0.590	0.028	0.004
2	0.000	0.084	0.006	0.134	0.136	0.110	0.952	0.210	0.946	0.976
	$h_n = h_n^{gcv} n^{-[1/11-1/15]}$									
0	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
1	0.000	0.000	0.006	0.000	0.000	0.004	0.004	0.610	0.028	0.012
2	0.000	0.080	0.006	0.108	0.114	0.068	0.920	0.110	0.920	0.944

Table A.2: Model 1, γ parameter estimates

Not Bias Corrected	True	MEAN	Var(s)	Var(t)	MSE	LQ	Median	UQ	MAE
				$h_n = h_n^{gcv}$					
$\gamma_{1,1}$	1.325	1.025	0.018	0.045	0.108	0.938	1.023	1.109	0.302
$\gamma_{1,2}$	0.646	0.437	0.008	0.021	0.052	0.382	0.435	0.502	0.209
$\gamma_{2,2}$	1.325	1.021	0.019	0.045	0.111	0.935	1.012	1.097	0.308
				$h_n = h_n^{gcv} n^{-\frac{1}{2}[1/11-1/15]}$					
$\gamma_{1,1}$	1.325	1.105	0.026	0.063	0.074	1.000	1.098	1.201	0.235
$\gamma_{1,2}$	0.646	0.440	0.010	0.028	0.053	0.374	0.440	0.511	0.208
$\gamma_{2,2}$	1.325	1.087	0.028	0.062	0.085	0.984	1.081	1.174	0.254
				$h_n = h_n^{gcv} n^{-[1/11-1/15]}$					
$\gamma_{1,1}$	1.325	1.189	0.040	0.090	0.059	1.055	1.185	1.311	0.196
$\gamma_{1,2}$	0.646	0.436	0.014	0.036	0.058	0.367	0.436	0.510	0.214
$\gamma_{2,2}$	1.325	1.169	0.048	0.088	0.073	1.031	1.168	1.300	0.216
Bias Corrected	True	MEAN	Var(s)	Var(t)	MSE	LQ	Median	UQ	MAE
				$h_n = h_n^{gcv}$					
$\gamma_{1,1}$	1.325	1.115	0.030	0.047	0.074	1.004	1.113	1.237	0.227
$\gamma_{1,2}$	0.646	0.519	0.010	0.027	0.026	0.453	0.516	0.585	0.138
$\gamma_{2,2}$	1.325	1.122	0.034	0.051	0.075	1.012	1.118	1.237	0.227
				$h_n = h_n^{gcv} n^{-\frac{1}{2}[1/11-1/15]}$					
$\gamma_{1,1}$	1.325	1.239	0.031	0.059	0.038	1.117	1.234	1.358	0.157
$\gamma_{1,2}$	0.646	0.559	0.010	0.028	0.018	0.490	0.558	0.628	0.110
$\gamma_{2,2}$	1.325	1.246	0.032	0.064	0.038	1.139	1.243	1.363	0.156
				$h_n = h_n^{gcv} n^{-[1/11-1/15]}$					
$\gamma_{1,1}$	1.325	1.353	0.033	0.085	0.034	1.232	1.343	1.466	0.143
$\gamma_{1,2}$	0.646	0.582	0.011	0.034	0.015	0.518	0.576	0.648	0.100
$\gamma_{2,2}$	1.325	1.358	0.036	0.094	0.037	1.237	1.346	1.460	0.146

Table A.3: Model 2, γ parameter estimates

Not Bias Corrected	True	MEAN	Var(s)	Var(t)	MSE	LQ	Median	UQ	MAE
				$h_n = h_n^{gcv}$					
$\gamma_{1,1}$	0.870	0.659	0.005	0.008	0.050	0.611	0.649	0.703	0.212
$\gamma_{1,2}$	0.000	-0.044	0.015	0.033	0.017	-0.115	-0.038	0.035	0.099
$\gamma_{2,2}$	5.294	4.384	0.295	0.598	1.122	4.033	4.384	4.719	0.937
				$h_n = h_n^{gcv} n^{-\frac{1}{2}[1/11-1/15]}$					
$\gamma_{1,1}$	0.870	0.723	0.007	0.012	0.029	0.666	0.711	0.776	0.152
$\gamma_{1,2}$	0.000	-0.054	0.018	0.043	0.020	-0.129	-0.058	0.033	0.113
$\gamma_{2,2}$	5.294	4.486	0.386	0.759	1.039	4.140	4.485	4.867	0.863
				$h_n = h_n^{gcv} n^{-[1/11-1/15]}$					
$\gamma_{1,1}$	0.870	0.793	0.010	0.019	0.016	0.727	0.785	0.855	0.104
$\gamma_{1,2}$	0.000	-0.068	0.026	0.058	0.030	-0.164	-0.064	0.035	0.132
$\gamma_{2,2}$	5.294	4.587	0.607	0.967	1.106	4.169	4.567	5.080	0.846
Bias Corrected	True	MEAN	Var(s)	$Var(t)$	MSE	LQ	Median	UQ	MAE
				$h_n = h_n^{gcv}$					
$\gamma_{1,1}$	0.870	0.796	0.025	0.029	0.031	0.689	0.786	0.882	0.140
$\gamma_{1,2}$	0.000	-0.022	0.015	0.036	0.015	-0.106	-0.017	0.060	0.098
$\gamma_{2,2}$	5.294	5.061	0.416	0.900	0.470	4.715	5.090	5.415	0.511
				$h_n = h_n^{gcv} n^{-\frac{1}{2}[1/11-1/15]}$					
$\gamma_{1,1}$	0.870	0.906	0.023	0.036	0.024	0.808	0.892	0.978	0.114
$\gamma_{1,2}$	0.000	-0.029	0.016	0.047	0.017	-0.117	-0.026	0.059	0.104
$\gamma_{2,2}$	5.294	5.135	0.393	1.129	0.418	4.798	5.167	5.476	0.469
				$h_n = h_n^{gcv} n^{-[1/11-1/15]}$					
$\gamma_{1,1}$	0.870	1.006	0.024	0.048	0.043	0.913	0.983	1.079	0.155
$\gamma_{1,2}$	0.000	-0.036	0.021	0.066	0.022	-0.124	-0.039	0.057	0.118
$\gamma_{2,2}$	5.294	5.224	0.438	1.490	0.443	4.760	5.213	5.660	0.521

Table A.4: Model 1, β and γ parameter estimates: Trimming with indicator function

Model 1: $y = (X\beta_1)(X\beta_2) + \varepsilon$

Not Bias Corrected, $h_n^{gcv} = (0.861, 0.865, 0.565)$

	True	MEAN	Var *sample*	Var *theory*	MSE	LQ	Median	UQ	MAE
				$h_n = h_n^{gcv}$					
$\beta_{1,3}$	1.000	1.054	0.027	0.039	0.030	0.953	1.051	1.145	0.130
$\beta_{2,3}$	1.000	1.043	0.026	0.038	0.028	0.932	1.040	1.135	0.127
$\gamma_{1,1}$	1.471	1.173	0.023	0.025	0.112	1.063	1.165	1.273	0.210
$\gamma_{1,2}$	0.714	0.505	0.008	0.009	0.052	0.444	0.504	0.556	0.210
$\gamma_{2,2}$	1.471	1.177	0.020	0.024	0.106	1.077	1.167	1.272	0.296
				$h_n = h_n^{gcv} n^{-\frac{1}{2}[1/11-1/15]}$					
$\beta_{1,3}$	1.000	1.028	0.038	0.051	0.039	0.907	1.025	1.133	0.150
$\beta_{2,3}$	1.000	1.044	0.037	0.051	0.039	0.922	1.038	1.149	0.150
$\gamma_{1,1}$	1.471	1.267	0.030	0.038	0.071	1.149	1.260	1.379	0.207
$\gamma_{1,2}$	0.714	0.511	0.010	0.013	0.052	0.443	0.506	0.571	0.207
$\gamma_{2,2}$	1.471	1.258	0.030	0.037	0.076	1.139	1.243	1.369	0.237
				$h_n = h_n^{gcv} n^{-[1/11-1/15]}$					
$\beta_{1,3}$	1.000	1.041	0.100	0.073	0.102	0.857	1.007	1.187	0.223
$\beta_{2,3}$	1.000	1.019	0.095	0.071	0.096	0.859	1.005	1.156	0.219
$\gamma_{1,1}$	1.471	1.351	0.055	0.058	0.070	1.212	1.338	1.492	0.210
$\gamma_{1,2}$	0.714	0.511	0.014	0.019	0.056	0.436	0.505	0.584	0.210
$\gamma_{2,2}$	1.471	1.363	0.060	0.057	0.072	1.199	1.353	1.499	0.214

Model 1: $y = (X\beta_1)(X\beta_2) + \varepsilon$

Bias Corrected, $h_n^{gcv} = (3.091, 3.073, 2.185)$

	True	MEAN	Var *sample*	Var *theory*	MSE	LQ	Median	UQ	MAE
				$h_n = h_n^{gcv}$					
$\beta_{1,3}$	1.000	1.005	0.024	0.046	0.024	0.908	1.001	1.096	0.113
$\beta_{2,3}$	1.000	1.008	0.023	0.048	0.023	0.918	1.005	1.083	0.109
$\gamma_{1,1}$	1.471	1.394	0.041	0.041	0.047	1.268	1.381	1.518	0.120
$\gamma_{1,2}$	0.714	0.623	0.013	0.010	0.021	0.545	0.619	0.695	0.120
$\gamma_{2,2}$	1.471	1.397	0.042	0.046	0.048	1.269	1.384	1.515	0.172
				$h_n = h_n^{gcv} n^{-\frac{1}{2}[1/11-1/15]}$					
$\beta_{1,3}$	1.000	0.972	0.022	0.059	0.023	0.888	0.977	1.060	0.112
$\beta_{2,3}$	1.000	0.965	0.025	0.065	0.026	0.870	0.952	1.051	0.120
$\gamma_{1,1}$	1.471	1.538	0.050	0.079	0.054	1.379	1.516	1.686	0.110
$\gamma_{1,2}$	0.714	0.659	0.016	0.016	0.019	0.572	0.652	0.747	0.110
$\gamma_{2,2}$	1.471	1.546	0.052	0.085	0.057	1.392	1.537	1.680	0.184
				$h_n = h_n^{gcv} n^{-[1/11-1/15]}$					
$\beta_{1,3}$	1.000	0.934	0.032	0.080	0.037	0.831	0.933	1.037	0.143
$\beta_{2,3}$	1.000	0.928	0.038	0.084	0.043	0.820	0.912	1.008	0.156
$\gamma_{1,1}$	1.471	1.679	0.078	0.138	0.121	1.492	1.644	1.826	0.113
$\gamma_{1,2}$	0.714	0.681	0.020	0.026	0.021	0.588	0.677	0.759	0.113
$\gamma_{2,2}$	1.471	1.694	0.087	0.140	0.136	1.501	1.664	1.839	0.274

Table A.5: Model 2, β and γ parameter estimates: Trimming with indicator function

Model 2: $y = (X\beta_1) + (X\beta_2)^2 + \varepsilon$

Not Bias Corrected, $h_n^{gcv} = (1.249, 0.580, 0.559)$

	True	MEAN	Var *sample*	Var *theory*	MSE	LQ	Median	UQ	MAE
				$h_n = h_n^{gcv}$					
$\beta_{1,3}$	1.000	1.126	0.018	0.035	0.034	1.039	1.128	1.206	0.150
$\beta_{2,3}$	1.000	1.009	0.004	0.005	0.004	0.971	1.008	1.048	0.048
$\gamma_{1,1}$	0.907	0.714	0.007	0.006	0.044	0.658	0.704	0.763	0.099
$\gamma_{1,2}$	0.000	-0.042	0.013	0.016	0.015	-0.121	-0.046	0.037	0.099
$\gamma_{2,2}$	5.892	5.071	0.299	0.246	0.974	4.729	5.026	5.401	0.870
				$h_n = h_n^{gcv} n^{-\frac{1}{2}[1/11-1/15]}$					
$\beta_{1,3}$	1.000	1.082	0.023	0.046	0.029	0.982	1.087	1.180	0.137
$\beta_{2,3}$	1.000	1.010	0.006	0.007	0.006	0.960	1.003	1.059	0.061
$\gamma_{1,1}$	0.907	0.784	0.009	0.009	0.024	0.718	0.769	0.840	0.109
$\gamma_{1,2}$	0.000	-0.055	0.016	0.022	0.019	-0.141	-0.057	0.031	0.109
$\gamma_{2,2}$	5.892	5.200	0.379	0.336	0.857	4.752	5.187	5.586	0.787
				$h_n = h_n^{gcv} n^{-[1/11-1/15]}$					
$\beta_{1,3}$	1.000	1.045	0.034	0.064	0.036	0.931	1.038	1.146	0.144
$\beta_{2,3}$	1.000	1.010	0.010	0.011	0.010	0.949	1.001	1.069	0.076
$\gamma_{1,1}$	0.907	0.859	0.012	0.015	0.015	0.790	0.848	0.926	0.126
$\gamma_{1,2}$	0.000	-0.067	0.022	0.031	0.026	-0.167	-0.061	0.033	0.126
$\gamma_{2,2}$	5.892	5.303	0.536	0.480	0.883	4.815	5.259	5.754	0.782

Model 2: $y = (X\beta_1) + (X\beta_2)^2 + \varepsilon$

Bias Corrected, $h_n^{gcv} = (3.902, 2.238, 2.202)$

	True	MEAN	Var *sample*	Var *theory*	MSE	LQ	Median	UQ	MAE
				$h_n = h_n^{gcv}$					
$\beta_{1,3}$	1.000	0.819	0.080	0.102	0.113	0.710	0.792	0.886	0.238
$\beta_{2,3}$	1.000	1.024	0.008	0.021	0.008	0.969	1.017	1.076	0.067
$\gamma_{1,1}$	0.907	1.016	0.043	0.079	0.055	0.894	1.006	1.129	0.116
$\gamma_{1,2}$	0.000	-0.019	0.022	0.021	0.022	-0.107	-0.020	0.076	0.116
$\gamma_{2,2}$	5.892	5.867	0.553	0.911	0.553	5.389	5.876	6.334	0.566
				$h_n = h_n^{gcv} n^{-\frac{1}{2}[1/11-1/15]}$					
$\beta_{1,3}$	1.000	0.783	0.075	0.106	0.122	0.661	0.757	0.861	0.269
$\beta_{2,3}$	1.000	1.020	0.011	0.032	0.012	0.962	1.014	1.078	0.079
$\gamma_{1,1}$	0.907	1.147	0.059	0.110	0.117	1.007	1.123	1.252	0.132
$\gamma_{1,2}$	0.000	-0.035	0.029	0.035	0.031	-0.133	-0.023	0.082	0.132
$\gamma_{2,2}$	5.892	6.010	0.731	1.403	0.745	5.466	5.963	6.477	0.634
				$h_n = h_n^{gcv} n^{-[1/11-1/15]}$					
$\beta_{1,3}$	1.000	0.737	0.123	0.119	0.192	0.617	0.719	0.825	0.318
$\beta_{2,3}$	1.000	1.033	0.021	0.050	0.022	0.947	1.023	1.103	0.105
$\gamma_{1,1}$	0.907	1.276	0.073	0.132	0.210	1.119	1.233	1.379	0.156
$\gamma_{1,2}$	0.000	-0.029	0.041	0.057	0.042	-0.152	-0.030	0.096	0.156
$\gamma_{2,2}$	5.892	6.016	0.957	1.873	0.973	5.454	5.948	6.527	0.724

References

[1] Collomb, G. and W. Härdle (1986): "Strong Uniform Convergence Rates in Robust Nonparametric Time Series Analysis and Prediction: Kernel Regression Estimation from Dependent Observations," *Stochastic Processes and Their Applications*, 23, 77–89.

[2] Cragg, J.G. and S.G. Donald (1996): "On the Asymptotic Properties of LDU-Based Tests of the Rank of a Matrix," *Journal of the American Statistical Association*, 91, 1301-1909.

[3] _____ (1997): "Inferring the Rank of a Matrix," *Journal of Econometrics*, 67, 223–250.

[4] Gasser, T., H.-G. Müller and V. Mammemitzsch (1985): "Kernels for Nonparametric Curve Estimation," *Journal of the Royal Statistical Society, B*, 47, 238–252.

[5] Gill, L. and A. Lewbel (1992): "Testing the Rank and Definiteness of Estimated Matrices with Applications to Factor, State-Space and ARMA Models," *Journal of the American Statistical Association*, 87, 766–776.

[6] Golub, G.H. and G.F. van Loan (1983), *Matrix Computations*, Baltimore: The Johns Hopkins University Press.

[7] Härdle, W. and T.M. Stoker (1989): "Investigating Smooth Multiple Regression by the Method of Average Derivatives," *Journal of the American Statistical Association*, 84, 986–995.

[8] Henderson, H.V. and S.R. Searle (1979): "Vec and Vech Operators for Matrices, with some uses in Jacobian and Multivariate Statistics," *Canadian Journal of Statistics*

[9] Hoeffding, W. (1948): "A Class of Statistics with Asymptotically Normal Distribution," *Annals of Mathematical Statistics*, 19, 293–325.

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

[11] Horowitz, J.L. and W. Härdle (1996): "Direct Semiparametric Estimation of Single-Index Models with Discrete Covariates," *Journal of the American Statistical Association*, 91, 1632–1640.

[12] Hristache, M., A. Juditsky, J. Polzehl, and V. Spokoiny (2001): "Structure Adaptive Approach for Dimension Reduction," *Annals of Statistics*, 29, 1537-1566.

[13] Ichimura, H. and L.F. Lee (1991): "Semiparametric Least Squares Estimation of Multiple Index Models: Single Equation Estimation," in *Nonparametric and Semiparametric Estimation Methods in Econometrics and Statistics*, ed. by W.A. Barnett, J. Powell, and G.E. Tauchen. Cambridge: Cambridge University Press.

[14] Lee, L.-F. (1995): "Semiparametric Maximum Likelihood Estimation of Polychotomous and Sequential Choice Models," *Journal of Econometrics*, 65, 381–428.

[15] Magnus, J.R. and H. Neudecker (1988): *Matrix Differential Calculus with Applications in Statistics and Econometrics*, John Wiley & Sons.

[16] Newey, W.K. and D.L. McFadden (1994): "Large Sample Estimation and Hypothesis Testing," in *Handbook of Econometrics*, Vol. 4, ed. by R.F. Engle and D.L. McFadden, Elsevier: North-Holland.

[17] Picone, G.A. and J.S. Butler (2000): "Semiparametric Estimation of Multiple Equation Models," *Econometric Theory*, 16, 551–575.

[18] Powell, J. L., J. H. Stock, and T. M. Stoker (1989): "Semiparametric Estimation of Weighted Average Derivatives," *Econometrica*, 57, 1403–1430.

[19] Robin, J.-M. and R.J. Smith (2000): "Tests of Rank," *Econometric Theory*, 16, 151–175.

[20] Robinson, P.M. (1989): "Hypothesis Testing in Semiparametric and Nonparametric Models for Econometric Timeseries," *Review of Economic Studies*, 56, 511–534.

[21] Silverman, B.W. (1978): "Weak and Strong Uniform Consistency of the Kernel Estimate of a Density Function and Its Derivatives," *The Annals of Statistics*, 6, 177–184; Addendum (1980), 8, 1175–1176.

[22] Stoker, T. M. (1991): "Equivalence of Direct, Indirect, and Slope Estimators of Average Derivatives," in *Nonparametric and Semiparametric Estimation Methods in Econometrics and Statistics*, ed. by W.A. Barnett, J. Powell, and G.E. Tauchen. Cambridge: Cambridge University Press.

[23] Stone, C.J. (1982): "Optimal Global Rates of Convergence for Nonparametric Regression," *Annals of Statistics*, 10, 1040–1053.

[24] Xia, Y., H. Tong, W.K. Li, and Zhu, L.-X. (2002): "An Adaptive Estimation of Dimension Reduction Space," *Journal of the Royal Statistical Society B*, 64, 363-388.