A Method of Moments Estimator for Semiparametric Index Models*

by

Bas Donkers[†] and Marcia Schafgans[‡]

Contents:

The Suntory Centre
Suntory and Toyota International Centres for
Economics and Related Disciplines
London School of Economics and Political Science
Houghton Street
London WC2A 2AE
Tel.: 020 7955 6679

Discussion Paper
No.EM/05/493
July 2005

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

[†] Econometric Institute and Department of Business Economics, 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, LondonWC2A 2AE, United Kingdom. E-mail: m.schafgans@lse.ac.uk. Part 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.

Abstract

We propose an easy to use derivative based two-step estimation procedure for semiparametric index models. In the first step various functionals involving the derivatives of the unknown function are estimated using nonparametric kernel estimators. The functionals used provide moment conditions for the parameters of interest, which are used in the second step within a method-of-moments framework to estimate the parameters of interest. The estimator is shown to be root N consistent and asymptotically normal. We extend the procedure to multiple equation models. Our identification conditions and estimation framework provide natural tests for the number of indices in the model. In addition we discuss tests of separability, additivity, and linearity of the influence of the indices.

Keywords: Semiparametric estimation, multiple index models, average derivative functionals, generalized methods of moments estimator, rank testing.

JEL No.: C14, C31, C52.

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 $y \in R^S$ and explanatory variables $x \in R^L$. H is an unknown, but sufficiently smooth function, and $B = (\beta_1, ..., \beta_P)$ is the matrix containing all unknown parameters. Many econometric models can be regarded this way. With $y \in R$, the model generalizes the usual linear regression model, but it also encompasses binary choice models, disequilibrium models, duration models with competing risks and sample selection models. In the more general case of $y \in R^S$, the model includes multivariate and multinomial choice models, with y containing indicator functions for each possible alternative, and the sample selection model with y containing the selection variable as well as an outcome variable. Various root N consistent asymptotically normal estimators of B for the multiple index model have been proposed, among others, by Ichimura and Lee (1991) for the case of $y \in R$, Lee (1995) for the multinomial choice model, and Picone and Butler (2000) for the $y \in R^S$ case.

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. Much of the computational burden arises from the non-parametric regressions that have to be performed at each iteration in the optimization process for the objective function. The advantage of our method is that it only involves a single nonparametric step: computation of nonparametric estimates of various functionals involving the derivatives of the unknown function. Another advantage of the proposed estimator is that it provides a natural framework to test for the number of indices. This advances the literature on semiparametric index models as the number of indices, so far, has been assumed known. Lastly, our estimator provides for a simple test for the additivity and/or linearity of the influence of an index formulated as simple parameter restrictions.

While our exposition mainly focusses on the case of $y \in R$, we extend our results to the multiple equation setting. Unless indicated otherwise, our notation is therefore based on y being a scalar variable.[1] Let $g(x) = E(y|x)$. 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, $p = 1, .., P$,

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

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.

The estimator we propose uses a larger set of average derivative functionals to remedy this identification problem. Each of these average derivative functionals (e.g., $E(g')$, $E(g'g'^T)$, $E(g'')$, or $Var(g')$) provides information on the parameters of interest. Consider, for example, the average of the outer product of derivatives (gradient). It provides the following (additional) moment conditions for the parameters of interest

$$E(g'(x)g'(x)^T) = \sum_{p,q=1}^{P} E\left[\frac{\partial H}{\partial (x^T \beta_p)} \frac{\partial H}{\partial (x^T \beta_q)} \right] \beta_p \beta_q^T. \tag{3}$$

By including a larger set of moment conditions, our estimator provides a more efficient estimate for single index models than the regular ADE. More important, though, is the result that with these additional average derivative functionals the parameters in multiple index models can be estimated as well.

We develop the asymptotic theory for the parameter estimates for B using the results of Samarov (1993), which indicate that root N consistent, asymptotically normal, estimates of a large set of average derivative functionals can be obtained by kernel regression meth-

[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 matrix.

ods. These average derivative functionals provide moment conditions for the parameters of interest. We therefore incorporate Samarov's results within the generalized methods of moments framework (Newey and McFadden, 1994). We show that the parameter estimates for B are root N consistent and asymptotically normal. This method of moments framework facilitates the use of information from multiple equations, as it is straightforward to combine the moment conditions that result from each equation. It also facilitates the use of cross-equation restrictions, as the parameters in all equations are estimated simultaneously.

An important issue we address, preceding the estimation of the parameters, concerns parameter identification in semiparametric muliple index models. The traditional set of identifying restrictions, see Ichimura and Lee (1991), are a combination of normalization and exclusion restrictions. We allow for a more general set of identification restrictions, based on the idea that only the space spanned by the indices is identified.[2] We discuss parameter identification in detail articulating the identification conditions in terms of the parameters and the link function.

The identification conditions reveal the equivalence of the number of indices in the model and the rank of the outerproduct of the gradient. Consequently, existing tests on the rank of a matrix (e.g. Cragg and Donald 1996, 1997, Robin and Smith, 2000) can be used to test for the number of indices. This idea carries over to the multiple equation setting, where the number of indices equals the rank of the sum of the outerproduct of the gradient for each equation. Alternatively, restrictions on the number of indices can be tested using overidentifying restrictions tests within the method of moments framework.

We note that the multiple index model is also treated in the statistics 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 directions can be estimated at the parametric root N rate. However, they do not develop the asymptotic theory for the estimated parameters, nor do they provide a test on the number

[2] For related results, see the literature on factor models (Nunnally and Bernstein, 1994, and Philips, 1994), dimension reduction (Xia et al, 2002, and Li, 1991), and cointegration (Johansen, 1988).

of indices.

The paper is organized as follows. In Section 2, we discuss root N consistent, asymptotically normal nonparametric estimators of functionals of derivatives. In Section 3, we show how these nonparametrically estimated functionals can be used within a GMM framework to yield a root N consistent, asymptotically normal estimator for the parameters of interest. In Section 4, we discuss testing procedures for the number of indices and provide testing procedures for the separability, additivity and linearity of the influence of an index. In addition to theoretical results, we illustrate the estimator's usefulness in Section 5 with a simulation study. In particular, we consider the estimation of the parameters in a multinomial choice model. Section 6 concludes. An Appendix contains the assumptions we make to ensure root N consistency and asymptotic normality of the nonparametric estimators of functionals of derivatives adapted from Samarov (1993).

2 Estimating average derivative functionals

Our initial aim is to obtain root N consistent estimates of average derivative functionals. These average derivative functionals include, among numerous other possibilities, the average first order derivative (AD), the average hessian (AH) and the average outerproduct of the gradient (AOPG). Other functionals can be used, where one can think of higher order derivatives or higher order products of the first order derivatives. In this paper, we restrict ourselves to the aforesaid set of average derivative functionals for two reasons. First, these functionals guarantee identification and are informative about most properties of $H(\cdot)$ that are of interest, such as additivity, linearity and separability, see Samarov (1993). Second, the conditions that need to be satisfied to obtain root N consistent estimates of the additional quantities, in general, are stronger and include, for example, the use of higher order kernels than currently required.

Before we discuss the asymptotic distribution of the joint set of moments we introduce some notation. 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 L

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

$$g'(x) = \frac{G'(x)}{f(x)} - \frac{G(x)f'(x)}{f(x)^2}. \tag{4}$$

Similarly an expression for the second order derivatives can be obtained. Estimates of such derivatives can be obtained after estimation of its constituents, $f(x)$, $f'(x)$, $f''(x)$, $G(x), G'(x)$, and $G''(x)$. We use the Nadaraya-Watson leave-one-out kernel estimators given by

$$\begin{array}{ll} \hat{f}(x_i) = \frac{1}{(N-1)h^L} \sum\limits_{j=1,j\neq i}^{N} K\left(\frac{x_i-x_j}{h}\right); & \hat{f}'(x_i) = \frac{1}{(N-1)h^{L+1}} \sum\limits_{j=1,j\neq i}^{N} K'\left(\frac{x_i-x_j}{h}\right); \\ \hat{f}''(x_i) = \frac{1}{(N-1)h^{L+2}} \sum\limits_{j=1,j\neq i}^{N} K''\left(\frac{x_i-x_j}{h}\right); & \hat{G}(x) = \frac{1}{(N-1)h^L} \sum\limits_{j=1,j\neq i}^{N} K\left(\frac{x_i-x_j}{h}\right) y_j; \\ \hat{G}'(x_i) = \frac{1}{(N-1)h^{L+1}} \sum\limits_{j=1,j\neq i}^{N} K'\left(\frac{x_i-x_j}{h}\right) y_j; & \hat{G}''(x_i) = \frac{1}{(N-1)h^{L+2}} \sum\limits_{j=1,j\neq i}^{N} K''\left(\frac{x_i-x_j}{h}\right) y_j, \end{array} \tag{5}$$

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

To avoid very imprecise contributions to the average derivative functionals for observations with low densities we need to introduce some trimming. The need for this is highlighted by the presence of an estimate of the density in the denominator of these functionals. In our particular setting, the presence of this trimming function does not bias the parameter estimates for B, it only affects estimates of auxiliary parameters, denoted by Γ in the sequel. While various trimming functions have been considered in the literature, for the present purpose we have decided to apply the fixed trimming (or weighting) function $w(x)$, where $w(x)$ is supported on a subset of the support of $f(x)$ on which $f(x)$ is bounded away from 0. To deal meticulously with stochastic trimming[3] would detract from the main

[3] Hardle ad Stoker (1989), amongst others, consider stochastic trimming on the basis of the density, say $w_N(x) = 1(f(x) > b_N)$ where $b_N \to 0$ (or a smoothed version thereof). This typically requires imposing conditions, needed to ensure that the bias vanishes sufficiently fast in the tails of $f(x)$, that are particularly strong (see also Laverne and Vuong (1996) and Lewbel (1997)). Quantile trimming is considered by Lee (1995), where the quantile is strictly bounded away from zero to account for the difficulty to control the rate of divergence of $1/\hat{f}_N(x)$. See also Donkers and Schafgans (2003).

Moment	**Kernel estimator**
$AD = E\{w(x)\frac{\partial g(x)}{\partial x}\}$	$\frac{1}{N}\sum_{i=1}^{N} w(x_i)\left(\frac{\hat{G}'(x_i)}{\hat{f}(x_i)} - \frac{\hat{G}(x_i)\hat{f}'(x_i)}{\hat{f}^2(x_i)}\right)$
$AOPG = E\{w(x)\frac{\partial g(x)}{\partial x}\frac{\partial g(x)}{\partial x}^T\}$	$\frac{1}{N}\sum_{i=1}^{N} w(x_i)\left(\frac{\hat{G}'(x_i)}{\hat{f}(x_i)} - \frac{\hat{G}(x_i)\hat{f}'(x_i)}{\hat{f}^2(x_i)}\right)\left(\frac{\hat{G}'(x_i)}{\hat{f}(x_i)} - \frac{\hat{G}(x_i)\hat{f}'(x_i)}{\hat{f}^2(x_i)}\right)^T.$
$AH = E\{w(x)\frac{\partial g(x)}{\partial x \partial x^T}\}$	$\frac{1}{N}\sum_{i=1}^{N} w(x_i)\left(\frac{\hat{G}''(x_i)}{\hat{f}(x_i)} - \frac{\hat{G}(x_i)\hat{f}''(x_i)+\hat{G}'(x_i)\hat{f}'(x_i)^T+\hat{f}'(x_i)\hat{G}'(x_i)^T}{\hat{f}^2(x_i)} + \frac{2\hat{G}(x_i)\hat{f}'(x_i)\hat{f}'(x_i)^T}{\hat{f}^3(x_i)}\right)$

Table 1: Moments of interest and their kernel estimates.

contribution of this paper and is left to a separate paper. Density weighting has been suggested as an alternative to trimming (e.g., Powell, Stock, and Stoker, 1989). The higher order moments in our case would require weighting with the density to the fourth power, however. This results in large variation in the weights given to each observation, with most weight being given to a small fraction of the data thereby harming efficiency.

The (weighted) moments of interest and their kernel estimates are summarized in Table 1. Under conditions given by Samarov (1993) root N consistent estimates of each element of these moments can be obtained. In general, the assumptions concern smoothness properties of the densities and the conditional expectations, boundedness of a number of variances, and the use of higher order kernels in combination with restrictions on the bandwidth used. The details of the conditions required for each moment are presented in the Appendix. It should be noted that when a kernel of order $L+3$ is used in combination with $h = c \times N^{-\frac{1}{2L+5}}$, the assumptions on both the rate of convergence and the order of the kernel are satisfied for all moments.

When combining all moments used to estimate the parameters of interest, we need the joint distribution of these moment estimators. For this, Theorem 1 of Samarov (1993) is of interest, as it states that each estimated element of the moments can be written as a sample average. More specifically, each typical element can be written as

$$\hat{m}_N = \frac{1}{N}\sum_{i=1}^{N} \phi_m(x_i, \hat{f}(x_i), \hat{f}^k(x_i), \hat{f}^l(x_i), \hat{f}^{kl}(x_i), \hat{g}(x_i), \hat{g}^k(x_i), \hat{g}^l(x_i), \hat{g}^{kl}(x_i)) \qquad (6)$$

with ϕ_m satisfying necessary smoothness properties, see Corollary 1 in the Appendix. Here, and below, the superscripts k and l denote the derivatives with respect to the k^{th} and l^{th} element of the L dimensional x vector, respectively. Samarov's theorem then states that as $N \to \infty$, the following asymptotic expansion holds

$$\hat{m}_N - m = \frac{1}{N}\sum_{i=1}^{N} D\phi_m(x_i, y_i) - E\{D\phi_m(x_i, y_i)\} + o_p(N^{-1/2}), \tag{7}$$

with m the particular moment under consideration and $D\phi_m$ its influence function. Similar results have been obtained and used by Härdle and Stoker (1989) and Powell, Stock and Stoker (1989), see also Donkers and Schafgans (2003).

Let $\widehat{M}$ denote the vector containing all nonparametric estimates of the (unique) average derivative functionals of interest, so $\widehat{M} = (\widehat{AD}^T$,vech$(\widehat{AOPG})^T$,vech$(\widehat{AH})^T)^T$. We can now conclude that

$$\left(\widehat{M} - M\right) = \frac{1}{N}\sum_{i=1}^{N} R(x_i, y_i) - E\{R(x_i, y_i)\} + o_p(N^{-1/2}), \tag{8}$$

with $R(x_i, y_i)$ the vector with all stacked influence functions, $D\phi_m$. Applying the Lindeberg Levy central limit theorem yields the root N consistency and asymptotic normality of $\widehat{M}$ with

$$\sqrt{N}\left(\widehat{M} - M\right) \xrightarrow{d} N(0, \Sigma_M) \tag{9}$$

and $\Sigma_M = \text{Var}(R(x_i, y_i))$. The influence functions of the (weighted) average derivative, average outerproduct of gradient, and average hessian have been derived using second order $U-$statistics theory (Hoeffding, 1948, Powell, Stock and Stoker, 1989) and are given in Table 2. A consistent estimator for Σ_M, $\hat{\Sigma}_M$, can be obtained by taking the sample variance of $\hat{R}(x_i, y_i)$, where $\hat{R}(x_i, y_i)$ is the nonparametric estimator of $R(x_i, y_i)$. Alternatively, a bootstrap estimator for Σ_M can be used (as in Samarov, 1993, and Buchinsky, 1994), which is obtained as

$$\frac{1}{B}\sum_{b=1}^{B}\left(\widehat{M}_b - \widehat{M}\right)\left(\widehat{M}_b - \widehat{M}\right)', \tag{10}$$

where for each of the B bootstrapped samples an estimator $\widehat{M}_b$ is obtained.

With $y \in R^S$, let $g(x) = (g_1(x), ..., g_S(x))^T$. Upon extending our estimator to the multiple equation setting, one should realize that the set of moment conditions may include

Moment	**Influence Function**
AD	$R^{AD}(x,y) = w(x)\left\{g'(x) - (y - g(x))\frac{f'(x)}{f(x)}\right\}$
$AOPG$	$R^{AOPG}(x,y) = w(x)\left\{g'(x)g'(x)^T - (y - g(x))\left[\frac{f'(x)g'(x)^T}{f(x)} + \frac{g'(x)f'(x)^T}{f(x)} + 2g''(x)\right]\right\}$
AH	$R^{AH}(x,y) = w(x)\left\{g''(x) + (y - g(x))\frac{f''(x)}{f(x)}\right\}$

Table 2: Moments of interest and their influence functions.

all moments of each equation separately. In addition, cross-equation moments such as $AOPG_{s,t} = E\{w(x)\frac{\partial g_s(x)}{\partial x}\frac{\partial g_t(x)}{\partial x}^T\}$, $t \neq s = 1, ..., S$ can be included. Root N consistent, asymptotically normal, estimators for this functional can be defined in a similar way. The associated influence function is given by

$$\begin{aligned} R^{AOPG}_{s,t}(x,y) &= w(x)\left\{g'_s(x)g'_t(x)^T - (y_s - g_s(x))\left[\frac{f'(x)g'_t(x)^T}{f(x)} + g''_t(x)\right]\right. \qquad (11) \\ &\left. -(y_t - g_t(x))\left[\frac{f'(x)g'_s(x)^T}{f(x)} + g''_s(x)\right]\right\}. \end{aligned}$$

The main asymptotic result for the joint distribution of the stacked nonparametric estimators of the (unique) average derivative functions holds suitably augmented.

3 Parameter identification and estimation

The previous section discussed how various moments involving average derivative functionals of the unknown regression function can be estimated root N consistently using nonparametric kernel estimation methods. This section deals with the ultimate objective of estimating the parameters in the multiple index model with the use of these asymptotically normal estimates. The method of moments estimator provides a natural framework for this, as each average derivative functional (moment) contains particular information about the parameters of interest. Table 3 presents this information, which can be derived by application of the chain rule of differentiation. The expectations of the relevant (weighted) properties of the link function $H(\cdot)$ are incorporated in a vector or matrix of auxiliary parameters that we refer to as Γ.

Moment	**Parameter information**
AD	$\sum_{p=1}^{P} E\left[w(x)\frac{\partial H(x^T\beta_1,..,x^T\beta_P)}{\partial(x^T\beta_p)}\right]\beta_p \equiv B\Gamma^D$
$AOPG$	$\sum_{p,q} E\left[w(x)\frac{\partial H(x^T\beta_1,..,x^T\beta_P)}{\partial(x^T\beta_p)}\frac{\partial H(x^T\beta_1,..,x^T\beta_P)}{\partial(x^T\beta_q)}\right]\beta_p\beta_q^T \equiv B\Gamma^{OPG}B^T$
AH	$\sum_{p,q} E\left[w(x)\frac{\partial^2 H(x^T\beta_1,..,x^T\beta_P)}{\partial(x^T\beta_p)\partial(x^T\beta_q)}\right]\beta_p\beta_q^T \equiv B\Gamma^H B^T$

Table 3: Average derivative functional moments and the model parameters.

Clearly, there is a close link between the parameters of interest and each of the moments based on the average derivative functionals. The relationships in Table 3 can be used to construct the moment conditions, e.g., $AOPG - B\Gamma^{OPG}B^T = 0$, or in general $M - M_{B,\Gamma} = 0$. The method of moment estimator selects those parameters (B and Γ) for which the sample analogue of these moments holds as closely as possible.

While each of the moment conditions can provide information on the indices, for AD and AH this is not necessarily the case as both $E\left[w(x)\frac{\partial H}{\partial(x^T\beta_p)}\right]$ and $E\left[w(x)\frac{\partial^2 H}{\partial(x^T\beta_p)\partial(x^T\beta_q)}\right]$ can equal zero. In particular, this occurs in the situation of a symmetric function in combination with x being distributed symmetrically around zero for the AD moments and for a linear contribution of an index for the AH moment conditions. We will therefore prove identification of the parameters of interest focusing on the $AOPG$ moments, as these moments conditions provide identification under a minimal set of identifying restrictions. In many situations, one might be willing to make additional assumptions and use restrictions on the AD or AH moments to gain identification. For instance, a possible set of assumptions one could consider is $\Gamma^D = (1,..,1)^T$, which imposes a normalization ensuring that all indices have equal marginal effect on y (see also Stoker, 1991). Once identification is established, we continue with a discussion of the GMM framework to estimate the parameters.

3.1 Identification

In order to establish conditions for identification of the parameters, we first turn to the number of indices to be estimated, P. This is an important determinant of the number of parameters in the model. Any assumption on the number of indices in the model should therefore contain 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 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 parameters B and the second is on the shape of the function $H(\cdot)$.

Assumption 1 *$Rank(B) \equiv rank((\beta_1, .., \beta_P)) = P$.*

Assumption 2 *The function $H(\cdot)$ 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 1 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. Assumption 2 asserts that each of the indices provides unique information on the shape of $H(\cdot)$, that is, the derivatives of $H(\cdot)$ with respect to each of the indices are not linearly dependent, almost everywhere, see also Ichimura and Lee (1991), Lemma 3, condition 3.

To ensure that the number of indices is not affected by $w(x)$, the trimming (weighting) function used in our kernel based nonparametric estimators, we strengthen Assumption 2 :

Assumption 2′ *The function $H(\cdot)$ satisfies*

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

Assumptions 1 and 2 ensure that only P indices have to be estimated.[4] However, without further assumptions on either B or $H(\cdot)$ the parameters in B are still not identified. To illustrate the well known scaling and rotation problems present in semiparametric index models, we rewrite $H(x^T\beta_1, .., x^T\beta_P)$ as $H(x^TB)$. Let B^* denote a given $L \times P$ matrix

[4] Assumptions 2 and 2′ are used interchangeably.

with columns spanning the column space of B and define $H^*(\cdot)$ as the unique function that satisfies $g(x) = H^*(x^T B^*)$. The identification problem stems from the fact that for any $P \times P$ matrix Λ of full rank it holds that $H^*(x^T B^*) = H_\Lambda(x^T B^* \Lambda)$ with $H_\Lambda(z) = H^*(z\Lambda^{-1})$. It is therefore not clear whether one estimates B^* in combination with the properties of H^* or $B^*\Lambda$ in combination with the properties of H_Λ; they are observationally equivalent. However, once Λ is fixed, both $B^*\Lambda$ and H_Λ are uniquely determined and can be estimated. Consequently, P^2 restrictions need to be made to fully identify the remaining parameters. The traditionally used exclusion and normalization restrictions discussed in Ichimura and Lee (1991) and Lee (1995), in accordance, impose P normalization and $P(P-1)$ exclusion restrictions.

For reasons mentioned before, we focus on the $AOPG$ moments in discussing identification. In particular, $AOPG = \sum_{p,q} E\left[w(x)\frac{\partial H}{\partial(x^T\beta_p)}\frac{\partial H}{\partial(x^T\beta_q)}\right]\beta_p\beta_q^T = B\Gamma^{OPG}B^T$, with $\Gamma^{OPG} = E\left\{\left[w(x)\frac{\partial H}{\partial(x^T\beta_p)}\frac{\partial H}{\partial(x^T\beta_q)}\right]_{1\leq p,q\leq P}\right\}$ a matrix of auxiliary parameters. How does the lack of identifying information affect this moment condition? Consider the $AOPG$ moments for the general case with $E\{y|x\} = H_\Lambda(x^T B^*\Lambda)$, which then read as $AOPG = B^*\Lambda\Gamma_\Lambda^{OPG}\Lambda^T B^{*T}$, with $\Gamma_\Lambda^{OPG} = \Lambda^{-1}\Gamma^{*OPG}(\Lambda^{-1})^T$. The exclusion and normalization restrictions from Ichimura and Lee (1991) and Lee (1995) impose the identifying restrictions on Λ as follows. Let, without loss of generality, the first P variables in x denote the variables used for the exclusion and normalization restriction, and correspondingly partition $B^T = [B_1^T|B_2^T]$, with B_1 a square matrix of dimension P. The exclusion and normalization restrictions impose $B_1 = I_P$. This indeed achieves identification as it fixes $\Lambda = (B_1^*)^{-1}$ and subsequently we estimate $B = B^*\Lambda$ and $\Gamma^{OPG} = \Gamma_\Lambda^{OPG}$. However, it is possible to incorrectly specify an exclusion or normalization restriction – leading to B_1^* being singular – with the result that identification will not be achieved. This is illustrated by the following simple example: With $P = 2$, let $\beta_1 = (1\ 1\ 0\ 0)^T$ and $\beta_2 = (1\ 1\ 1\ 0)^T$. In this case one cannot use x_1 and x_2 for the normalization and exclusion restrictions. Another problem arises when x_4 is used for exclusion or normalization. It should be noted that Assumption 1 guarantees that a valid set of exclusion and normalization restrictions does exist.

Given the possible existence of these problems, it is desirable to estimate the model using a more general set of identifying restrictions, i.e. a set of restrictions that does not impose more structure than what is imposed by the identification assumptions 1 and 2. Such a set of identifying restrictions is given by the orthonormality of the β's (Xia et al., 2002), so $B^T B = I$, and diagonality of Γ^{OPG}, i.e., $\Gamma^{OPG} = D$, a matrix with positive elements on its diagonal and zeroes everywhere else. To fix the ordering of the indices, an additional assumption could be that the elements on D are ordered in decreasing order. Taking into account the symmetry of $B^T B$ and Γ^{OPG}, this again imposes exactly P^2 restrictions. Once the matrix B is estimated satisfying this set of identifying restrictions it is always possible, ex-post, to estimate B with the normalization and exclusion restrictions imposed by computing $(\hat{B}_1)^{-1}\hat{B}_2$, where the validity can be tested beforehand.[5]

We now verify that this set of identifying restrictions fixes Λ. Let $\Lambda = \left(B^{*T}B^*\right)^{-1/2}\Lambda^{\perp}$, where $\Lambda^{\perp}$ denotes any orthonormal $P \times P$ matrix. This choice, with $B = B^*\Lambda$, ensures $B^T B = I$ is satisfied. Clearly, orthonormality of the β's puts some restrictions on the model, but not enough to fix Λ completely. This is achieved by applying the restriction that $\Gamma^{OPG} = D$. With $\Gamma^{OPG} = \Gamma^{OPG}_{\Lambda} = \Lambda^{-1}\Gamma^{*OPG}(\Lambda^{-1})^T$ and $\Lambda^{-1} = \Lambda^{\perp T}\left(B^{*T}B^*\right)^{1/2}$ it follows that $\Lambda^{\perp}$ has to contain the orthonormal eigenvectors of $\left(B^{*T}B^*\right)^{1/2}\Gamma^{*OPG}(\left(B^{*T}B^*\right)^{1/2})^T$ that correspond to the eigenvalues that appear on the diagonal of D. The uniqueness of the eigenvalue decomposition (Magnus and Neudecker, 1988) then assures that $\Lambda^{\perp}$ and therefore Λ is uniquely defined.

The results in Phillips (1994) suggest that when knowledge on exclusion and normalization restrictions is available, they should be used in light of the superior small sample performance when using these identifying restrictions over the orthonormality assumption. Asymptotically, though, they are equivalent.

We summarize these identification restrictions in the following assumption.

Assumption 3 *When P indices are estimated with the use of at least the AOPG moments,*

[5] These ideas are similar to the ideas in Johansen (1988, Theorem 1) on cointegrating vectors, where maximum likelihood estimation of the space spanned by β is considered.

either

(i) Each index $x^T\beta_p$, $p = 1, .., P$, contains one explanatory variable which does not enter the other $P - 1$ indices. The parameters on these variables are normalized to equal 1, *or*

(ii) The β's are orthonormal, i.e., $B^T B = I$, and Γ^{OPG} is a diagonal matrix.

3.2 Extension to multiple equations

In the multiple equation setting with $y \in R^S$, let $H(x^T B) = (H_1(x^T B), .., H_S(x^T B))^T$. One set of identifying assumptions, imposed, for example, by Picone and Butler (2000), is to impose Assumption $2'$ for each link function $H_s(x^T B)$, $s = 1, .., S$. This might be much too strong, as one index could play a role in one equation, but not in another one. The requirement instead is that each index plays a role in at least one equation and that the contributions of the indices in each equation, $\frac{\partial H_s}{\partial(x^T\beta_p)}$, $s = 1, .., S$, do not have the same linear dependencies in all equations. Formally, we require that the columns in $\begin{pmatrix} \frac{\partial H}{\partial(x^T\beta_1)} & \cdots & \frac{\partial H}{\partial(x^T\beta_P)} \end{pmatrix} \equiv \begin{pmatrix} \frac{\partial H_1}{\partial(x^T\beta_1)} & \cdots & \frac{\partial H_1}{\partial(x^T\beta_P)} \\ \vdots & \ddots & \vdots \\ \frac{\partial H_S}{\partial(x^T\beta_1)} & \cdots & \frac{\partial H_S}{\partial(x^T\beta_P)} \end{pmatrix}$ have no linear dependence relationships almost everywhere. We ensure this by imposing

Assumption $2''$ *The function $H(\cdot)$ satisfies*

$$\mathit{rank}\left(\sum_{s=1}^{S} E\left\{\left[\frac{\partial H_s}{\partial(x^T\beta_p)}\frac{\partial H_s}{\partial(x^T\beta_q)}\right]_{1\le p,q\le P}\right\}\right) = P.$$

Clearly a much weaker assumption than imposing Assumption 2 for each equation.

Assumptions 1 and $2''$ (suitably adjusted for trimming) indicate that P indices need to be estimated, but they do not guarantee parameter identification. The discussion of identification based on the frequently used exclusion and normalization restrictions on B translates directly to the multiple equation setting. Some more discussion is warranted for the set of normalization restrictions based on $B^T B = I$ and diagonality of Γ^{OPG}. Diagonality of Γ^{OPG} in the single equation setting was needed to fix the rotation of B. In the multiple equation situation, if one equation is known to be a function of all indices, say equation s, imposing diagonality on Γ_s^{OPG} will be a sufficient restriction on the rotation of the indices in the multiple index setting. Since one cannot guarantee the existence of an equation in which all indices should be included, it will not always be sufficient to impose diagonality on Γ_s^{OPG} for a single equation.

A general approach to fix the rotation of B in the multiple equation setting is to sequentially consider equations and fix the rotation of the indices that appear for the first time in the equation at hand. Suppose that the rotation of the first p indices has been obtained by the s equations that have been considered so far. For the next equation, only the rotation of additional indices, whose rotation has not been restricted yet, needs to be fixed. This amounts to requiring diagonality of Γ_{s+1}^{OPG}, with the exception of the top-left $p \times p$ submatrix. Suppose now that one additional index is normalized in this step. This causes p zeroes in the $(p+1)^{\text{th}}$ row (and also in the column, due to symmetry) of Γ_{s+1}^{OPG}. The subsequent rows and columns only contain zeroes as these indices do not (yet) play a role in the model. In case more than one additional index is relevant, this will result, for the j^{th} additional index, in $p+j-1$ zeroes in the $(p+j)^{\text{th}}$ row. In doing so, we have imposed for index p, $p = 1, .., P$, exactly $(p-1)$ zero restrictions in Γ_s^{OPG}, with s being the equation where index p first appeared. Once the rotation of all P indices has been fixed, the desired $\sum_{p=1}^{P}(p-1) = P(P-1)/2$ restrictions are imposed. The other $P(P+1)/2$ restrictions result from $B^T B = I$.

3.3 Estimation

We now turn to the GMM framework used to estimate the parameters of interest β_p, $p = 1, .., P$ and the auxiliary parameters, Γ. The relationship between the average derivative functional (moment) and the parameters of interest, summarized in Table 3, provide the moment conditions on which our estimator is based. The (unique) moment conditions used in estimation, in its general form are denoted by $M - M_{B,\Gamma} = 0$. Let us write the moment conditions as

$$m(\theta_0) \equiv M - M_{B,\Gamma} = 0.$$

Here θ_0 denotes the vector of all free parameters in B and Γ (identifiability), which we assume to be an element of a compact parameter space Θ.

Assumption 4 *$\theta_0 \in \Theta$, where Θ is compact.*

Given a valid set of identifying restrictions, it now holds that $m(\theta) = 0$ if and only if $\theta = \theta_0$, where $m(\theta) \equiv M - M_{B,\Gamma}$ with (B, Γ) determined by $\theta \in \Theta$. Let $\hat{m}(.)$ denote the estimated sample analogue of $m(.)$, based on nonparametric kernel estimates. Using Section 2, we note that, for $\theta = \theta_0$,

$$\sqrt{N}\left(\widehat{M} - M_{B,\Gamma}\right) = \sqrt{N}\hat{m}(\theta_0) \xrightarrow{d} N(0, \Sigma_M).$$

The efficiently weighted generalized method of moments estimator (or minimum distance estimator) for estimating θ_0, therefore, is given by

$$\widehat{\theta} = \arg\min_{\theta \in \Theta} \hat{m}(\theta)^T \left[\Sigma_M\right]^{-1} \hat{m}(\theta), \tag{12}$$

where we assume Σ_M to be positive definite. To implement the efficiently weighted GMM estimator we can use the consistent estimator for Σ_M presented 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). Under Assumptions 1 and 2 and a set of identification assumptions we showed that the parameters are uniquely determined.[6] This uniqueness result in combination with the continuity of $m(\theta)$ on Θ and the compactness Assumption 4 ensures identification of our estimator $\hat{\theta}$. Uniform weak convergence is ensured by the consistency of $\hat{M}$ and $\widehat{\Sigma}_M$ and the compactness of the parameter space (Assumption 4).

[6] The uniqueness discussion above assumed the inclusion of AOPG in the set of moments used in the estimation. Once Γ^{AOPG} and B are identified, Γ^{AD} and Γ^{AH} are uniquely identified as well: in the case of orthonormality $\Gamma^{AD} = B^T [AD]$ and $\Gamma^{AH} = B^T [AH] B$; in the case of exclusion and normalisation restrictions $\Gamma^{AD} = AD_1$ and $\Gamma^{AH} = AH_{11}$ where AD and AH are partitioned conform $B = [B_1|B_2]$ where $B_1 = I$.

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. Given the nonsingularity of Σ_M, we need to show that $m'(\theta_0)$ has full column rank. This condition, also called the rank condition (necessary) for local identification, follows directly from the uniqueness result on the parameters.

By satisfying all regularity conditions of GMM estimators, our final result is given by

Theorem 1 *Given that $\sqrt{N}\left(\hat{M} - M\right) \xrightarrow{d} N(0, \Sigma_M)$, with Σ_M positive definite, θ_0 in the interior of Θ, Assumptions 1, 2', 3, and 4*

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

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

For the moment conditions we consider, it is straightforward to see that $\frac{\partial m(\theta)}{\partial \theta}$ is continuous in θ. With $\hat{\theta}$ a consistent estimate for θ, Ω can therefore be consistently estimated by $\left[m'(\hat{\theta})^T \hat{\Sigma}_M^{-1} m'(\hat{\theta})\right]^{-1}$.

Even though our estimator uses more information than, for example, the ADE estimator, it might not attain the semiparametric efficiency bound under our conditional mean index assumption, see also Ai and Chen (2003). When strenghtening our conditional mean index assumption to a conditional distribution index assumption, our root N consistent estimator can be followed by a one-step update, as suggested by Delecroix et al. (2003). This yields a simple three step estimator that attains the semiparametric efficiency bound. Without this strenghtening of the conditional mean index assumption, our estimator will provide a very good starting point for existing iterative procedures that are asymptotically efficient (Ichimura and Lee, 1991 and Newey, 2004). Obviously, the more moment conditions we incorporate, the less of a concern this loss in efficiency should be in light of the spanning condition argument in Newey (2004).

4 Inference

In this section we discuss various model specification tests. First, we discuss a test for the number of indices in our model, a test which logically precedes the estimation of the index parameters B (and Γ). Indeed, our two-step estimation procedure provides a natural framework for this as we provide a test based on our first-step estimators which does not involve the estimation of the index parameters. Obviously, the GMM framework used in the second-step allows for the verification of the correct specification of the number of indices by using an overidentiying restriction test as well. Second, we discuss model specification tests for separability, additivity and linearity of the influence of an index, based on our root N consistent, asymptotically normal parameter estimates of the auxiliary parameters Γ.

4.1 Testing for the number of indices

Estimation of the index parameters in B 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. This section offers a test on the number of indices that can be performed prior to parameter estimation. It thereby advances the literature on semiparametric multiple index models, where so far, the number of indices has been assumed known. An exception to this is the projection pursuit model (Friedman and Stuetzle, 1981) where other restrictive assumptions are imposed.

Specifically, we propose to test for the number of indices by testing the rank conditions in Assumptions 1 and 2. Indeed, Assumptions 1 and 2 imply that (i) P indices have to be estimated and (ii) $AOPG = B\Gamma^{OPG}B^T$ has rank P.[7] The null space of $AOPG$ is reduced with one dimension for each index in the multiple index model. The number of indices therefore can be tested by testing the rank of the estimated average outer product of the gradient. This test is related to procedures in the statistical literature "estimating" the number of effective dimension reduction directions (e.g., Xia et al., 2002), where directions

[7] In the multiple equation setting, we define $AOPG = \sum_{s=1}^{S} AOPG_s$, where $AOPG_s$ is the average outer product of the gradient of the s^{th} equation. Conformably, $\Gamma^{OPG} = \sum_{s=1}^{S} \Gamma_s^{OPG}$.

with an impact below an (arbitrary) cutoff level are ignored. The advantage of the proposed testing procedure is that it quantifies how likely it is that a small effect actually is zero.

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 root 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. The rank of $AOPG$ equals the number of nonzero elements in the diagonal "pivot" matrix D in this decomposition. Consequently Gill and Lewbel tested for the number of zero elements on the diagonal of D. Recognizing that their asymptotic theory only holds for $k-P=1$, Cragg and Donald (1996) develop a modified procedure to test the rank of a matrix. This is a Wald type test on a number of elements in the matrix being zero. This test is asymptotically equivalent to the minimum chi-squared approach presented in more detail in Cragg and Donald (1997). This test, we note, is identical to a test for overidentifying restrictions in our GMM framework when only the $AOPG$ moments are used in estimation. It is distributed asymptotically as χ^2 with $(k-P)(k-P+1)/2$ degrees of freedom. An interpretation of the degrees of freedom easily follows from the number of overidentifying restrictions.

Finally, Robin and Smith (2000) develop a test for the rank of a matrix that involves the characteristic roots of a quadratic form in $AOPG$. Again, when $\text{rank}(AOPG)=P$, the smallest $k-P$ eigenvalues of $\widehat{AOPG}$ (and the quadratic form) converge to 0 in probability. Robin and Smith show that N times the sum of the $k-P$ smallest eigenvalues converges to a weighted sum of independent χ^2_1 distributed variables. The advantage of this test is that the variance-covariance matrix of $\text{vec}(\widehat{AOPG})$ 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 $AOPG$ 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 $AOPG$ we consider testing the rank of the weighted variant $\Psi(AOPG)\Psi^T$, where Ψ is a diagonal matrix of full rank which ensures that the diagonal elements of $\Psi(AOPG)\Psi^T$ are estimated with equal precision. Since Ψ is of full rank, the rank of $AOPG$ equals that of $\Psi(AOPG)\Psi^T$.

Obviously, our GMM framework itself also provides a natural testing procedure for the correct specification of the number of indices through the overidentifying restriction test. Note that the overidentifying restrictions test is a general test for misspecification. When identification is obtained by assuming $B^T B = I$ and $\Gamma^{AOPG} = D$, a diagonal matrix, the only possible misspecification is too few indices in the model. When exclusion and normalization restrictions are used, additional structure is imposed on the model, which will be tested by the overidentifying restrictions test as well.

Theoretical considerations that would favour the use of the overidentifying restriction test are (i) the tests by Cragg and Donald (1996) and Robin and Smith (2000) do not deal specifically with the positive semidefiniteness of the $AOPG$ matrix and (ii) even with our proposed weighting scheme the Cragg and Donald (1996) and the Robin and Smith (2000) tests only account for differences in the estimation precision of the diagonal elements whereas one would prefer to correct for differences in estimation precision of all elements and for the correlations between the estimates. In the simulation study we compare the performance of these tests in practice.

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 rank$(AOPG) = r$. This requires at each stage of the sequential procedure an adjustment to the asymptotic size α_P of the test that depends on the sample size. 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).

4.2 Other model specification tests

In the literature, various tests for additivity and or linearity of the contribution of a single explanatory variable have been proposed (Härdle, Sperlich and Spokoiny, 2001; Samarov, 1993; Stoker 1989). Our consistent, asymptotic normal, parameter estimator of B and Γ allow for similar model specification tests in the context of semiparametric index models. In particular, we focus here on testing of the additivity (separability) and/or linearity of the influence of an index (instead of a single explanatory variable). These model specification hypotheses can all be reformulated as linear restrictions on our auxilliary parameters, Γ. As such, they can all be carried out using the standard Wald test, yielding asymptotic chi-squared tests.

First, we consider a test of additivity (separability) of the influence of an index. That is, we are interested in testing whether our conditional mean can be represented as

$$H(x^T\beta_1, ..., x^T\beta_P) \equiv H^1(x^T\beta_1) + H^2(x^T\beta_2, ..., x^T\beta_P). \tag{13}$$

Under such an additivity restriction, it holds that

$$\Gamma^{AH} = \begin{bmatrix} E\left(w(x)\frac{\partial^2 H^1(x^T\beta_1)}{\partial^2(x^T\beta_1)}\right) & 0 \\ 0 & E\left(w(x)\frac{\partial^2 H^2(x^T\beta_{-1})}{\partial(x^T\beta_{-1})\partial(x^T\beta_{-1})^T}\right) \end{bmatrix} \equiv \begin{bmatrix} \Gamma^{AH}_{11} & \Gamma^{AH}_{12} \\ \Gamma^{AH}_{21} & \Gamma^{AH}_{22} \end{bmatrix},$$

where $(x^T\beta_{-1}) \equiv (x^T\beta_2, ..., x^T\beta_P)$. The additive separability can therefore be tested by verifying whether the relevant off-diagonal elements of Γ^{AH} are statistically significant, $\Gamma^{AH}_{21} = 0$ $(= \Gamma^{AH}_{12}$ by symmetry). This test generalizes to the separability of groups of indices. Testing whether all indices are additively separable (projection pursuit model), an extreme example of the above, can similary be described as testing whether all off-diagonal elements of Γ^{AH} are zero, see also Härdle, Sperlich and Spokoiny (2001).

Next, we consider a test of linearity of the influence of an index. We can formulate our conditional mean subject to such a linearity restriction as

$$H(x^T\beta_1, ..., x^T\beta_P) \equiv (x^T\beta_1)H^1(x^T\beta_2, ..., x^T\beta_P) + H^2(x^T\beta_2, ..., x^T\beta_P).$$

As this linear influence model gives rise to $\Gamma^{AH} = \begin{bmatrix} 0 & \Gamma_{12}^{AH} \\ \Gamma_{21}^{AH} & \Gamma_{22}^{AH} \end{bmatrix}$, linearity of the influence of an index can be tested on the basis of the diagonal elements of Γ^{AH}.

Jointly testing the additivity and linearity of the influence of an index, constitutes a test of the partial linear regression model providing an adequate representation of the conditional mean. The partial linear regression model, specified as

$$H(x^T\beta_1, ..., x^T\beta_P) \equiv c \cdot x^T\beta_1 + H^2(x^T\beta_2, ..., x^T\beta_P),$$

has frequently been used in the sample selection literature and results in $\Gamma^{AH} = \begin{bmatrix} 0 & 0 \\ 0 & \Gamma_{22}^{AH} \end{bmatrix}$. This restriction can be tested by jointly testing the significance of the relevant diagonal and off-diagonal elements of Γ^{AH}.

One caveat of these tests has to be mentioned. As with the average derivative estimator, one could postulate particular specifications of our multiple index model which in combination with, say, a symmetric distribution of X, yield zero elements of the AH matrix for reasons other than those described above. In these very special settings, the power of these tests would be negligible. To gain power against such alternatives, one may consider combining differently weighted versions of AH, following the ideas in Newey and Stoker (1993), section 5.3.

5 Simulation study

In order to illustrate the estimator's usefulness, we provide simulation results based on the multinomial choice model with three choice alternatives. The latent variable representation of this model is given by

$$\begin{aligned} y_j &= \begin{cases} 1 \text{ if } y_j^* = \arg\max(y_1^*, y_2^*, y_3^*) \\ 0 \text{ otherwise} \end{cases}, \text{ for } j = 1, 2, 3 \\ y_j^* &= z^T\gamma_j - u_j, \end{aligned}$$

where y_j^* is the latent variable associated with the random "utilities" associated with each alternative. Since the observed choices are only informative on the difference of these "utilities" and not on the "utilities" themselves, it is commonplace to define the multinomial choice model in differenced form. Define $x^T\beta_1 = z^T\gamma_1 - z^T\gamma_3$ and $x^T\beta_2 = z^T\gamma_2 - z^T\gamma_3$, where x denotes the vector of all distinct exogenous variables, and $e_1 = u_1 - u_3$, $e_2 = u_2 - u_3$. Under the assumption that the differenced errors e_j, $j = 1, 2$, only depend on the data x through the indices $(x^T\beta_1, x^T\beta_2)$, the choice probabilities conditional on the x's form the following multiple index mean regression representation:

$$\begin{aligned} g_1 \equiv E(y_1|x) &= \Pr(x^T\beta_1 \geq e_1, x^T\beta_1 - x^T\beta_2 \geq e_1 - e_2|x) \equiv H_1(x^T\beta_1, x^T\beta_2) \\ g_2 \equiv E(y_2|x) &= \Pr(x^T\beta_2 \geq e_2, x^T\beta_2 - x^T\beta_1 \geq e_2 - e_1|x) \equiv H_2(x^T\beta_1, x^T\beta_2). \end{aligned}$$

Since the choices are mutually exclusive and exhaustive, the choice probability associated with the last alternative, the reference category, is ignored, since it does not add any information to the model.

In the multinomial choice model, our conditional mean index assumption coincides with a conditional index assumption on the distribution of the dependent variable. We can therefore perform a one-step efficient update to convert our computational efficient estimator into an estimator that achieves the semiparametric efficiency bound for the multinomial choice model as given in Lee (1993). We do not pursue this extension in our simulation, but refer to Delecroix et al. (2003) for more details.

From Assumption 3, suitably adjusted for the multiple equation setting (as discussed in Section 3.2), we note that for our two index model ($P = 2$) at least three explanatory variables are required. We assume that $X \sim N(0, I_3)$ and let $\beta_1 = (1, 0, 1)^T$ and $\beta_2 = (0, 1, 1)^T$. We assume that (e_1, e_2) are distributed independently of X, with bivariate normal distribution with unit variances and correlation $\rho_{e_1e_2} = 0.5$ (which conforms the assumptions of the multinomial probit model with i.i.d. errors on the utilities). With Φ_2

denoting the CDF of a standard bivariate normal distribution, this yields[8]

$$H(x^T\beta_1, x^T\beta_2) = \left(\begin{array}{c} H_1(x^T\beta_1, x^T\beta_2) \\ H_2(x^T\beta_1, x^T\beta_2) \end{array} \right) = \left(\begin{array}{c} \Phi_2(x^T\beta_1, \left(x^T\beta_1 - x^T\beta_2\right), 0.5) \\ \Phi_2(x^T\beta_2, \left(x^T\beta_2 - x^T\beta_1\right), 0.5) \end{array} \right).$$

The multivariate kernel function $K(\cdot)$ (on R^3) is chosen as the product of three univariate kernel functions. The sample size is set at 1000 and 500 replications are drawn in each case.

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

$$K_6(x) = \left(\tfrac{45\,045}{2048}x^8 - \tfrac{31\,185}{512}x^6 + \tfrac{59\,535}{1024}x^4 - \tfrac{11\,025}{512}x^2 + \tfrac{4725}{2048}\right) 1(|x| \leq 1). \tag{14}$$

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, which satisfy our assumption that the kernel and its derivative vanish at the boundary.

Four versions of our estimator are considered, which differ according to the moments we include, with a view to illustrate the potential to increase efficiency in a finite sample setting. The first version (Estimator 1) includes all AOPG moments, inclusive of the cross-equation AOPG moment, ensuring identification of our parameters. The second and third version add to this moment, either the AD moment (Estimator 2) or the AH moment (Estimator 3). The final version incorporates all three moments (Estimator 4) and should provide the most efficient estimates.

A bandwidth sequence $\{h_n\}$ satisfying the assumptions for these versions of our estimator is given by $h_n = cn^{-1/(2L+5)}[= cn^{-1/11}]$, where c is a constant factor independent of n, which we allow to vary for each explanatory variable. The particular parameterization chosen for our multinomial selection model, allows us to treat the bandwidth choice for x_1 and x_2 symmetrically, which reduces our selection of bandwidth parameters to two. In light of

[8] Note that $\sigma^2_{e_1-e_2} = \sigma^2_{e_2-e_1} = \sigma^2_{e_1} + \sigma^2_{e_2} - 2\sigma_{e_1e_2} = 1$, $\rho_{e_1,e_1-e_2} = \sigma^2_{e_1} - \sigma_{e_1e_2} = 0.5$, and $\rho_{e_2,e_2-e_1} = \sigma^2_{e_2} - \sigma_{e_1e_2} = 0.5$

our knowledge of the true data generating process, we decided to optimally select the bandwidth by minimizing the mean squared error of the nonparametrically estimated moments. Using a gridsearch algorithm, and 100 simulations, we arrived at the following bandwidths for the second and sixth order kernel respectively $(1.0, 1.0, 1.0)$ and $(3.0, 3.0, 3.0)$. There was no evidence in the simulation that distinct bandwidths for x_1 (x_2) and x_3 are needed. Alternatively, one could use cross-validation[9] or apply a 'plug-in' estimator for the optimal bandwidth as discussed by Powell and Stoker (1996).

In place of the weight function, we used the indicator function $w(x_i) = 1(\hat{f}(x_i) > b)$, which trims away the observations with small values of the density estimator (as in Härdle and Stoker, 1989 and Samarov, 1993). With $b = 0.0025$, this yielded a trimming of a bit less than 10 percent of the observations.

As estimator of the covariance matrix of these nonparametrically estimated moments, $\hat{\Sigma}_M$, we applied the bootstrap estimator (with 500 bootstraps) defined in (10). It provided an estimate of the variance comparable to the empirical variance for all moments. The theoretical estimate of the variance for the AOPG was remarkably similar to the empirical one, but for the AD and AH, the empirical variances were considerably underestimated when using the bias corrected kernel. For AD and AH, the theoretical estimator of the variance, based on first-order asymptotics, apparently suffers from a non-ignorable contribution of higher order terms in finite samples. An exploration of the higher order terms in the asymptotic expansion could resolve this issue, see Heckman, Ichimura and Todd (1997).

Table 4 presents the results of various tests for the number of indices for our multinomial choice model with two indices. The results are presented using both second order kernels (not bias corrected) and sixth order kernels (bias corrected). The table reports the fraction of the simulations for which we accept that the true number of indices for our multinomial choice model equals 0, 1 or 2 at the five percent significance level, using the sequential procedure described above.

The first four columns report the results based on testing the rank of $AOPG$ (=

[9] Cross-validation yielded bandwidths similar in magnitude as those obtained by minimizing the mean squared error of the nonparametrically estimated moments.

Kernel	# Indices	CD_u	CD_w	RS_u	RS_w	OI_0	OI_1	OI_2	OI_3	OI_4
Not Bias Corrected	0	0.001	0.000	0.000	0.000	0.000 [6]	0.000 [21]	0.000 [27]	0.000 [33]	0.000 [39]
	1	0.000	0.000	0.000	0.000	0.000 [3]	0.000 [15]	0.000 [19]	0.000 [25]	0.000 [29]
	2	0.848	0.840	0.998	0.998	1.000 [1]	0.990 [7]	0.998 [9]	0.972 [13]	0.996 [15]
Bias Corrected	0	0.000	0.000	0.000	0.000	0.000 [6]	0.000 [21]	0.000 [27]	0.000 [33]	0.000 [39]
	1	0.000	0.000	0.000	0.000	0.000 [3]	0.000 [15]	0.000 [19]	0.000 [25]	0.000 [29]
	2	0.880	0.944	0.992	0.992	0.994 [1]	0.998 [7]	1.000 [9]	0.994 [13]	0.998 [15]

Table 4: Testing for the number of indices

$\sum_s AOPG_s$). Since the unweighted versions of the Cragg and Donald and the Robin and Smith tests do not account for the estimation uncertainty in $\hat{M}$ in determining the smallest pivots and eigenvalues, we perform a weighted variant of these tests as well (see section 4). The columns labelled CD_u and CD_w relate to the Cragg and Donald test based on the LDU decomposition (unweighted and weighted) and the columns labelled RS_u and RS_w relate to the Robin and Smith test (unweighted and weighted). The remaining columns provide the results of a number of tests on overidentification restrictions. The column labelled OI_0, is also based on $AOPG$, which provides the minimal identifying set of moment conditions, just as the Cragg and Donald and Robin and Smith tests. The columns, labelled OI_1, OI_2, OI_3 and OI_4, are based on the moment conditions used in estimation conform the 4 sets of moments conditions we use for model estimation. The degrees of freedom of the OI tests are listed in square brackets.

The notable distinction between these two sets of tests is that the tests for the rank of a matrix developed by Cragg and Donald and by Robin and Smith do not rely on parameter estimates of B and Γ. The overidentification tests, on the other hand, are all based on estimation of B and Γ consistent with the assumed number of indices in the model and the identifying assumptions. So with $\tilde{B}$ and $\tilde{\Gamma}$ having the appropriate dimensions, the overidentifying restrictions (or minimum Chi-squared) tests are computed as

$$\min_{\tilde{B},\tilde{\Gamma}} \left(N \left[(\hat{M} - M_{\tilde{B},\tilde{\Gamma}})^T \hat{\Sigma}_M^{-1} (\hat{M} - M_{\tilde{B},\tilde{\Gamma}}) \right] \right).$$

In particular, for estimator 1 (including only the AOPG as moments) $\hat{M} = \Big[vech(\widehat{AOP}G_1)^T, vech(\widehat{AOP}G_2)^T, vec(\widehat{AOP}G_{1,2})^T\Big]^T$ and $M_{\tilde{B},\tilde{\Gamma}} = \Big[vech(\tilde{B}\tilde{\Gamma}_1^{OPG}\tilde{B}^T)^T, vech(\tilde{B}\tilde{\Gamma}_2^{OPG}\tilde{B}^T)^T, vec(\tilde{B}\tilde{\Gamma}_{1,2}^{OPG}\tilde{B}^T)^T\Big]^T$. Here the subscripts 1 and 2 point to the particular equation it relates to and $AOPG_{1,2}$ is the cross-equation average outer product of the gradients.

The first test on overidentifying restrictions, OI_0, is based on the minimal identifying set of moment conditions and has the smallest degrees of freedom. Clearly the degrees of freedom of the overidentifying restrictions tests increases as we add more unique moment conditions to our estimator, and decreases as we increase the hypothesized number of indices. In determining the degrees of freedom of these overidentifying restrictions we take account of the symmetry of the $AOPG$ and AH matrices. In estimating the parameters, we use a set of normalization and exclusion restrictions. With the overidentifying restrictions tests we also test these assumptions but in our simulation setting we know these assumptions are satisfied.

All tests clearly demonstrate their ability to find the true number of indices for our multinomial choice model. The power against accepting the null of too few indices is close or equal to one in all cases. Without bias correction, all except for the Cragg and Donald tests accept that the true number of indices equals 2 in excess of the 95% level of significance at which we performed these tests. With bias correction, the results of the (weighted) Cragg and Donald test improve, where for the weighted version in 95% of the cases the correct number of indices is accepted in accordance with the size of the test. Without the weighting to take account of the estimation uncertainty in $\hat{M}$ we still accept in most of the cases (88%) the true number of indices with the Cragg and Donald test. Without bias correction, the testing results using the theoretical variance (not reported) of

$\hat{M}$ are comparable to those above using the bootstrapped variance. With bias correction, the testing results based on the theoretical variance are at odds with our knowledge of the true number of indices, clearly a consequence of the underestimated variance. The efficiently weighted GMM approach used in estimation of the parameters of interest uses the bootstrap estimator of Σ_M for this reason.

Table 5 presents an analysis of the parameter estimates for $[\tilde{\beta}_1^T, .., \tilde{\beta}_P^T]^T$ for the four versions of our estimator, where the number of indices, P, equals the true number of indices. With our normalisation and exclusion restrictions we have $\tilde{\beta}_1^T = \beta_{1,3}$ and $\tilde{\beta}_2^T = \beta_{2,3}$.[10] For comparison, the multinomial probit parameter estimates are reported as well, where the same exclusion restrictions have been imposed. To ensure comparability with the semiparametric estimates, where $\beta_{1,1}$ and $\beta_{2,1}$ are standardized to 1, we report $\hat{\beta}_{1,3}^{(p)}/\hat{\beta}_{1,1}^{(p)}$ and $\hat{\beta}_{2,3}^{(p)}/\hat{\beta}_{2,1}^{(p)}$ for the MNP regressions, where $\hat{\beta}^{(p)}$ are the MNP parameter estimates.

The tables present the following summary statistics for the 500 replications: the sample mean, the sample variance, the average of the theoretical variance (based on the bootstrapped variance of $\hat{M}$), lower quartile (LQ), median, upper quartile (UQ), and mean absolute error (MAE). From Table 5 it is clear that the parameters can be obtained from the nonparametrically estimated moment conditions quite accurately. The resulting parameter estimates for $\beta_{1,3}$ and $\beta_{2,3}$ are close to their true values with a small upward bias, comparable to the parametric estimates.

The most prominent finding is that adding the AD and the AH moments to the AOPG does not improve the efficiency of the estimates very much. In fact, in our simulation adding the AH moment slightly worsens the estimation precision. This holds both for the bias corrected and the non-bias corrected estimates. A loss in efficiency arising from not knowing the distribution of the disturbances occurs as expected, but is within reason; the variance of the semiparametric estimates is less than double that of the parametric ones. A comparison of the theoretical variance and the sample variance of the estimates reveals a slight underestimation for the theoretical variance.

[10] A table containing the numerous auxilliary parameter estimates for the four version of our estimator can be obtained from the authors.

Parametric Estimation

	True	MEAN	Var *sample*	Var *theory*	LQ	Median	UQ	MAE
Multinomial probit								
$\beta_{1,3}$	1.000	1.000	0.008	0.007	0.937	0.996	1.055	0.070
$\beta_{2,3}$	1.000	1.009	0.007	0.007	0.955	1.005	1.061	0.064

Semiparametric Estimation: Not bias corrected

	True	MEAN	Var *sample*	Var *theory*	LQ	Median	UQ	MAE
Estimator 1 (AOPG)								
$\beta_{1,3}$	1.000	1.007	0.014	0.011	0.917	0.995	1.077	0.096
$\beta_{2,3}$	1.000	1.015	0.013	0.011	0.939	1.010	1.084	0.089
Estimator 2 (AOPG/AD)								
$\beta_{1,3}$	1.000	1.007	0.014	0.011	0.920	0.997	1.078	0.096
$\beta_{2,3}$	1.000	1.015	0.013	0.011	0.938	1.007	1.088	0.089
Estimator 3 (AOPG/AH)								
$\beta_{1,3}$	1.000	1.007	0.014	0.010	0.922	1.000	1.079	0.096
$\beta_{2,3}$	1.000	1.015	0.013	0.011	0.938	1.004	1.091	0.090
Estimator 4 (AOPG/AD/AH)								
$\beta_{1,3}$	1.000	1.007	0.015	0.010	0.923	1.005	1.079	0.096
$\beta_{2,3}$	1.000	1.015	0.013	0.010	0.936	1.004	1.088	0.091

Table 5: Parameter estimates multinomial selection model

Semiparametric Estimation: bias corrected								
	True	MEAN	Var *sample*	Var *theory*	LQ	Median	UQ	MAE
Estimator 1 (AOPG)								
$\beta_{1,3}$	1.000	1.006	0.015	0.013	0.921	1.004	1.081	0.098
$\beta_{2,3}$	1.000	1.013	0.013	0.013	0.928	1.014	1.090	0.092
Estimator 2 (AOPG/AD)								
$\beta_{1,3}$	1.000	1.006	0.014	0.012	0.922	0.998	1.079	0.096
$\beta_{2,3}$	1.000	1.012	0.013	0.012	0.934	1.010	1.080	0.089
Estimator 3 (AOPG/AH)								
$\beta_{1,3}$	1.000	1.006	0.016	0.013	0.924	1.003	1.074	0.098
$\beta_{2,3}$	1.000	1.013	0.014	0.013	0.930	1.012	1.090	0.093
Estimator 4 (AOPG/AD/AH)								
$\beta_{1,3}$	1.000	1.006	0.015	0.012	0.923	1.001	1.078	0.096
$\beta_{2,3}$	1.000	1.012	0.013	0.012	0.931	1.012	1.086	0.092

Table 5: Parameter estimates multinomial selection model (continued)

The differences between the bias corrected and not bias corrected estimates are as expected. The bias in the estimates when the higher order kernel is used is smaller than that of the not bias corrected estimates. Clearly, this comes at the cost of a higher variance, as is well know. Looking at the total error using MAE (or MSE), the estimators without bias correction have the best performance.

Finally Table 6, reports the results of the specification tests on the linearity and additivity of the indices in our multinomial choice model. The tests are performed for the version of our estimator which includes all moments (Estimator 4). For each equation, the table reports the fraction of the simulations for which we accept the null of linearity or additivity of the index in addition to the parameter estimates on which these tests are based. The tests are all performed equation by equation and for the system of the two equations together.

As discussed, these tests can be described as linear restrictions on the auxiliary parameters Γ_1^{AH} and Γ_2^{AH}. The test of linearity of the first equation with respect to the first index, $x^T\beta_1$, is identical to testing whether $E\left(w(x)\frac{\partial^2 H_1(x^T\beta_1, x^T\beta_2)}{\partial^2(x^T\beta_1)}\right) \equiv \Gamma_{1,11}^{AH} = 0$. Similarly, the test of linearity of the first equation with respect to the second index, $x^T\beta_2$, is performed by testing $\Gamma_{1,22}^{AH} = 0$. Using the Wald test, all simulations gave evidence against the linearity of our first equation in the index, $x^T\beta_1$, while we rejected linearity of the first equation in the index $x^T\beta_2$ in 64 percent of the simulations. The nonlinearity of our first equation $H_1(x^T\beta_1, x^T\beta_2) = \Phi_2(x^T\beta_1, x^T\beta_2 - x^T\beta_1, 0.5)$ in terms of our index $x^T\beta_1$ is clearly easier to identify (mixed impact) than that of $x^T\beta_2$. The test results for the second equation are similar, resulting from a symmetry imposed by our specification of the simulation. In particular, comparing the expressions for the two equations, the two indices are simply interchanged. This results in, among other things, $\Gamma_{1,11}^{AH} \equiv E\left(w(x)\frac{\partial^2 H_1(x^T\beta_1, x^T\beta_2)}{\partial^2(x^T\beta_1)}\right) = E\left(w(x)\frac{\partial^2 H_2(x^T\beta_1, x^T\beta_2)}{\partial^2(x^T\beta_2)}\right) \equiv \Gamma_{2,22}^{AH}$, $\Gamma_{1,22}^{AH} = \Gamma_{2,11}^{AH}$, and $\Gamma_{1,12}^{AH} = \Gamma_{2,21}^{AH}$. Indeed, this is confirmed by the similarity of the parameter estimates in the simulations.

The test of additive separability of our two indices in the first equation is equivalent

		Semiparametric Estimation (Estimator 4)			
		Not Bias Corrected		Bias Corrected	
		Estimates [Mean,Var]	Test Rejection Rate	Estimates [Mean,Var]	Test Rejection Rate
Test for Linearity					
Index $(x^T\beta_1)$					
Equation 1	$\Gamma^{AH}_{1,11}$	0.055 $[0.133e{-}3]$	1.000	0.066 $[0.133e{-}3]$	1.000
Equation 2	$\Gamma^{AH}_{2,11}$	-0.027 $[0.138e{-}3]$	0.672	-0.033 $[0.142e{-}3]$	0.644
System			0.998		0.996
Index $(x^T\beta_2)$					
Equation 1	$\Gamma^{AH}_{1,22}$	-0.027 $[0.138e{-}3]$	0.648	-0.033 $[0.138e{-}3]$	0.630
Equation 2	$\Gamma^{AH}_{2,22}$	0.055 $[0.133e{-}3]$	0.998	0.067 $[0.131e{-}3]$	1.000
System			0.998		1.000
Test for Additivity					
Equation 1	$\Gamma^{AH}_{1,12}$	-0.028 $[0.070e{-}3]$	0.902	-0.033 $[0.104e{-}3]$	0.922
Equation 2	$\Gamma^{AH}_{2,12}$	-0.028 $[0.070e{-}3]$	0.890	-0.034 $[0.104e{-}3]$	0.930
System			1.000		1.000

Table 6: Specification tests for the multinomial selection model

to testing whether $\Gamma_{1,12}^{AH} \equiv E\left(w(x)\frac{\partial^2 H_1(x^T\beta_1, x^T\beta_2)}{\partial(x^T\beta_1)\partial(x^T\beta_2)}\right) = 0$. The evidence against additive separable indices in the multinomial choice model is strong with a 90%-93% rejection rate. The test for additive separability of the indices in both equations simultaneously is soundly rejected, independent of the type of kernel used in estimation. Given the strong evidence agains linearity and additivity, we do not consider testing the validity of the partial linear model specification.

6 Conclusions

In this paper we consider the estimation of semiparametric multi-index models for single *and* multiple equation 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 various average derivative functionals and their covariance matrix. Parameter estimation is then based on the nonparametrically estimated moments using a GMM approach. This step involves a simple minimization problem where, importantly, no additional kernel based calculations are required. The estimator is shown to be root N consistent and asymptotically normal.

Parameter estimation in multiple 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 advance the literature by providing such a procedure. For single equation models, we show that the rank of the outer product of the gradient equals the number of indices required in the semiparametric model. For multiple index models this generalizes such that the number of indices equals the rank of the sum of the outerproduct of the gradient of the separate equations. 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 our simulation study, the tests on the number of indices required were very successful in determining the appropriate dimension. The parameters of interest were also rather precisely retrieved from the moment conditions we consider. In particular the outerproduct of the gradient, which we showed to contain sufficient information for parameter identification, already produced accurate estimates. Adding the moment conditions implied by the average derivative and average hessian did not substantially improve the estimation results. Adding further moment conditions could be considered, at the cost of stronger conditions on the regression function and the kernels. Given the small efficiency gains from adding moments to the outerproduct of the gradient, adding more moments might not be worthwhile. Better ways to increase efficiency, or in fact attain efficiency, would be to use our estimator as a starting value for the iterative procedures of Ichimura and Lee (1991) or Newey (2004). More interesting is the one-step efficient update procedure of Delecroix et al. (2003), but this requires the stronger conditional index distribution assumption.

As our method is derivative based, parameter estimation is only feasible for continuous explanatory variables. Future research could consider an extension of the work by Horowitz and Härdle (1996) for the ADE framework to deal with discrete variables in the multiple index setting. Another interesting avenue for further research is to obtain analytical expressions for an estimator of the variance of the nonparametrically estimated moment conditions that work well in small samples, i.e. that incorporate higher order terms in the asymptotic expansion.

7 Appendix

The assumptions presented in this appendix are based on Samarov (1993) and specialized for our application. In particular, the assumptions relating to the kernels have been modified, where, unlike Samarov, we define higher order kernels in line with Härdle and Stoker (1989) and Powell, Stock and Stoker (1989). For each assumption we specify its formulation for each of the three moment conditions discussed, AD, $AOPG$, and AH, in case they differ.

Let U be a convex open-bounded set in R^L. The trimming function, $w(x)$, is supported on this set, U. We denote with $D_{\nu+1}(U)$ the set of functions whose derivatives up to order ν satisfy the following Lipschitz condition on U: $\sup_{u,u+v\in U,||v||\leq v_0} |a(u) - a(u+v)|/||v|| < \infty$ for some $v_0 > 0$.

Assumption A.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.*

Assumption A.2 *Density bounded away from zero on U*
For some $\varepsilon > 0$, $\inf_{x\in U} f(x) \geq \varepsilon$, and $U_\varepsilon = \{z : ||z - x|| \leq \varepsilon,\ x \in U\} \subseteq supp(f)$.

The assumptions on the kernel and bandwidth are given in Assumptions (A.3) and (A.4). Instead of using the kernel proposed in Samarov (1993), we define our higher order kernel, in line with Härdle and Stoker (1989) and Powell, Stock and Stoker (1989) to make the assumptions comparable. The modifications needed in the proof of Samarov (1993) in particular relate to the proof of Lemma 1. In addition to standard Taylor expansion arguments (as applied in Levit (1978)), the revised proof makes use of integration by parts as in Härdle and Stoker (1989), amongst others. This modification does call for a slight strengthening of the differentiability requirement (with one order, see also Assumption (A.5)), making it in line with differentiability assumptions given in Härdle and Stoker (1989) and Donkers and Schafgans (2003). At the same time the assumptions on the order of the kernel are weakened.

Assumption A.3 *Kernel definition.*
The kernel function $K(u)$ has bounded support $\{u : |u| \leq 1\}$, is symmetric, differentiable up to the order r (to be defined in the next assumption) and has r moments. The kernel and its derivative, $K'(u)$, vanish at the boundary. $K(u)$ is of order r, so with $(l_1, .., l_k)$ an

index set

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

Assumption A.4 *Bandwidth selection.*
With $N \to \infty$, $h \to 0$ and
(i) $Nh^{2r} = o(1)$, and
(ii) for some $\varepsilon > 0$,

- AD: $N^{1-\epsilon}h^{2L+2} \to \infty$ (or $r \geq L + 2$)
- AOPG and AH: $N^{1-\epsilon}h^{2L+4} \to \infty$ (or $r \geq L + 3$)

Condition (i) on the bandwidth, $Nh^{2r} = o(1)$, ensures that the bias vanishes sufficiently fast. Condition (ii) ensures that the linearization used in the proof is sufficiently close. These conditions are derived by adapting Samarov's proof of his Lemma 1 to follow that of Härdle and Stoker (1991), see also Donkers and Schafgans (2003).[11]

Assumption A.5 *Existence and boundedness of derivatives.*
Partial derivatives of f and G up to the order $r + 1$ are bounded and $f, G \in D_{r+2}(U)$

Assumptions (A.1), (A.2), (A.3), and (A.5) ensure smoothness properties, see condition 6 in Samarov (1993), for the average derivative functionals under consideration, as indicated by the following Corrolary.

Corrolary 1 *Let the average derivative function be denoted as $\int \phi_m(x, u(x))f(x)dx$, with $\phi_m : R^L \times W \to R$, for some $W \in R^8$, and $u(x) = (f(x_i), f^k(x_i), f^l(x_i), f^{kl}(x_i), g(x_i), g^k(x_i), g^l(x_i), g^{kl}(x_i))$. ϕ_m incorporates the trimming function $w(x)$ supported on U, conform the*

[11] While the results for AD and AOPG can be found in Härdle and Stoker (1991) and Donkers and Schafgans (2003), respectively, the formal derivation of the proof for the AH will be provided upon request.

main text. Assumptions (A.1), (A.2), (A.3), and (A.5) then ensure
(i) ϕ_m *and its partial derivatives* $\phi_m^k = d\phi/du_k$, $k = 1,..,8$, *are bounded and* $\phi \in D_2(R^L \times W)$.
(ii) $\phi_m^k f \in D_{r+2}(R^L)$ *as a function for* x, *for* $k = 1,..,8$.

Finally, we make an assumption to ensure that asymptotically the variance of the estimators of the derivative functionals under consideration vanishes.

Assumption A.6 *Finite variance of the components that need to be estimated.*

- *AD: Components of the random vector* $R^{AD}(X,Y) = w(X)\{g'(X) - (Y - g(X))[\frac{f'(X)}{f(X)}]\}$ *have finite variances.*
- *AOPG: Components of the random matrix* $R^{AOPG}(X,Y) = w(X)\{g'(X)g'^T(X) - (Y - g(X))[(f'(X)g'^T(X) + g'(X)f'^T(X))/f(X) + 2g''(X)]\}$ *have finite variances.*
- *AH: Components of the random matrix* $R^{AH}(X,Y) = w(X)\{(Y - g(X))f''(X)/f(X) + g''(X)\}$ *have finite variances.*

References

[1] Ai, C. and X. Chen, 2003, Efficient Estimation of Models with Conditional Moment Restrictions Containing Unknown Functions, Econometrica 71, 1795–1844.

[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, 1301-1309.

[3] _____ 1997, Inferring the Rank of a Matrix, Journal of Econometrics 76, 223–250.

[4] Donkers, B. and M. Schafgans, 2003, A Derivative Based Estimator for Semiparametric Index Models, Econometric Institute Report 2003-08, Erasmus University, Rotterdam.

[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] 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.

[7] Heckman, J.J., H. Ichimura, P. Todd, 1997, How Details Make a Difference: Semiparametric Estimation of the Partially Linear Regression Model, unpublished.

[8] Delecroix, M., W. Härdle and M. Hristache, 2003, ,Efficient Estimation in Conditional Single-Index Regression, Journal of Multivariate Analysis 86, 213–226.

[9] Friedman, J.H. and W. Stuetzle, 1981, Projection Pursuit Regression, Journal of the American Statistical Association 76, 817–823.

[10] Härdle, W., S. Sperlich and V. Spokoiny, 2001, Structural tests in additive regression, Journal of the American Statistical Association 96, 1333-1347.

[11] Hoeffding, W., 1948, A Class of Statistics with Asymptotically Normal Distribution, Annals of Mathematical Statistics 19, 293–325.

[12] 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.

[13] Hristache, M., A. Juditsky, J. Polzehl, and V. Spokoiny, 2001, Structure Adaptive Approach for Dimension Reduction, Annals of Statistics 29, 1537-1566.

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

[15] Johansen, S., 1988, Statistical Analysis of Cointegrating Vectors, Journal of Economic Dynamics and Control 44, 215-238

[16] Laverne, P. and Q. Vuong, 1996, Nonparametric Selection of Regressors: The Nonnested Case, Econometrica 64, 207–220.

[17] Lee, L.-F., 1995, Semiparametric Maximum Likelihood Estimation of Polychotomous and Sequential Choice Models, Journal of Econometrics 65, 381–428.

[18] Levit, B., 1978, Asymptotically Efficient Estimation of Nonlinear Functionals, Problems of Information Transmission 14, 204–209.

[19] Lewbel, A., 1997, Semiparametric Estimation of Location and Other Discrete Choice Moments, Econometric Theory 13, 32–51.

[20] Li, K.-C., 1991, Sliced Inverse Regression for Dimension Reduction, Journal of the American Statistical Association 86, 316–327.

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

[22] Newey, W.K., 2004, Efficient Semiparametric Estimation via Moment Restrictions, Econometrica 72, 1877–1998.

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

[24] Newey, W.K. and T.M. Stoker, 1993, Efficiency of Weighted Average Derivative Estimators and Index Models, Econometrica 61, 1199-1223.

[25] Nunnally, Jum C. and Ira H. Bernstein, 1994, Psychometric Theory, (McGraw-Hill Inc, New York).

[26] Phillips, P.C.B., 1994, Some Exact Distribution Theory for Maximum Likelihood Estimators of Cointegrating Coefficients in Error Correction Models, Econometrica 62, 73-93.

[27] Picone, G.A. and J.S. Butler, 2000, Semiparametric Estimation of Multiple Equation Models, Econometric Theory 16, 551–575.

[28] Powell, J. L., J. H. Stock, and T. M. Stoker, 1989, Semiparametric Estimation of Weighted Average Derivatives, Econometrica 57, 1403–1430.

[29] Powell, J. L., and T. M. Stoker, 1996, Optimal Bandwidth Choice for Density-Weighted Averages, Journal of Econometrics 75, 291–316.

[30] Robin, J.-M. and R.J. Smith, 2000, Tests of Rank, Econometric Theory 16, 151–175.

[31] Robinson, P.M., 1989, Hypothesis Testing in Semiparametric and Nonparametric Models for Econometric Timeseries, Review of Economic Studies 56, 511–534.

[32] Samarov, A.M., 1993, Exploring Regression Structure Using Nonparametric Functional Estimation, Journal of the American Statistical Association 88, 836-847.

[33] Stoker, T. M., 1989, Tests of Additive Derivative Constraints, Review of Economic Studies 56, 535-552.

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

[35] 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.