Econometric Theory, **24**, 2008, 1584–1606. Printed in the United States of America.
doi:10.1017/S0266466608080626

SPECIFICATION AND ESTIMATION OF SEMIPARAMETRIC MULTIPLE-INDEX MODELS

BAS DONKERS
Erasmus University Rotterdam

MARCIA SCHAFGANS
London School of Economics

We propose an easy to use derivative-based two-step estimation procedure for semiparametric index models, where the number of indexes is not known a priori. In the first step various functionals involving the derivatives of the unknown function are estimated using nonparametric kernel estimators, in particular the average outer product of the gradient (AOPG). By testing the rank of the AOPG we determine the required number of indexes. Subsequently, we estimate the index parameters in a method of moments framework, with moment conditions constructed using the estimated average derivative functionals. The estimator readily extends to multiple equation models and is shown to be root-N-consistent and asymptotically normal.

1. INTRODUCTION

We consider the multiple-index mean regression model

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

with dependent variable $y \in R^S$ and explanatory variables $x \in R^L$. Here H is an unknown but sufficiently smooth function, and $B = (\beta_1, \ldots, \beta_P)$ is the matrix containing all unknown parameters. Many econometric models can be regarded this way, including the usual linear regression model, duration models, sample selection models, and multivariate and multinomial choice models. Various root-N-consistent asymptotically normal estimators of B for the multiple-index model have been proposed by, among others, Ichimura and Lee (1991) for $y \in R$,

For insightful discussions and helpful comments we thank Herman van Dijk, Hidehiko Ichimura, Oliver Linton, Bertrand Melenberg, Peter Robinson, Arthur van Soest, and an anonymous referee. Bas Donkers enjoys the hospitality of CentER as an Extramural Fellow. Part of the work was done during Marcia Schafgans' sabbatical leave at CentER, and she thanks CentER and Tilburg University for their hospitality. Address correspondence to Bas Donkers, Econometric Institute and Department of Business Economics, Erasmus University Rotterdam, P.O. Box 1738, NL-3000 DR Rotterdam, The Netherlands; e-mail: donkers@few.eur.nl.

 © 2008 Cambridge University Press 0266-4666/08 $15.00

https://doi.org/10.1017/S0266466608080626 Published online by Cambridge University Press

Lee (1995) for the multinomial choice model, and Picone and Butler (2000) for $y \in R^S$.

Because the multiple-index model provides a general and flexible modeling strategy, one would expect to see numerous applications of the multiple-index model given the existence of these estimation methods. We propose to remedy two problems that might have contributed to the absence of these applications. First, the number of indexes needs to be specified by the researcher in semiparametric index models. Second, the existing semiparametric methods are computationally demanding, which arises from the many nonparametric regressions that have to be performed to optimize the objective function. Given the absence of global concavity of the objective function typically encountered, this constitutes a serious drawback in any attempt to achieve the global maximum.

Specifically, we propose a test for the number of indexes and an estimator for the parameters that involve only a single nonparametric step. We achieve all this through the use of various functionals involving the derivatives of the unknown link function. A noteworthy feature of these functionals is that they provide us with information on both the number of indexes and the parameters in the model. These average derivative (AD) functionals have to be estimated fully nonparametrically, as we do not impose assumptions on the number of indexes a priori, making dimension reduction infeasible.

Our exposition focuses on the case of $y \in R$, but we extend our results to the multiple equation setting.[1] Let $g(x) = \mathrm{E}(y|x)$. The derivative of this unknown function, $g'(x)$, by application of the chain rule of differentiation, is a weighted average of the true index coefficients β_p, $p = 1, \ldots, P$,

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

For single-index models this property is sufficient to identify the parameters "up-to-scale," which inspired the AD estimator (ADE) (see Powell, Stock, and Stoker, 1989; Robinson, 1989; Härdle and Stoker, 1989; Stoker, 1991). For multiple-index models the AD does not provide enough information to identify β_p, $p = 1, \ldots, P$ up-to-scale without further assumptions.

By using a larger set of AD functionals, we solve this identification problem. In particular, we include the average outer product of the gradient (AOPG) $\mathrm{E}(g'g'^T)$, which provides—as we will show—enough information to identify all parameters in a multiple-index model. Following Samarov (1993), we use kernel regression methods to obtain root-N consistent and asymptotically normal estimates of average derivative functionals.

Before the index parameters can be estimated, the number of indexes needs to be specified. In the existing literature, the number of indexes has always been assumed known.[2] We present identification conditions in terms of the index parameters and the link function that reveal the equivalence of the number of indexes in the model and the rank of the outer product of the gradient. Hence,

https://doi.org/10.1017/S0266466608080626 Published online by Cambridge University Press

the number of indexes in index models can be determined by testing the rank of the outer product of the gradient matrix, for which a root-N-consistent asymptotically normal estimate is available. Most existing rank tests require nonsingularity of the covariance matrix of its stacked elements. As recognized by Donald, Fortuna, and Pipiras (2007), the covariance matrix of a positive definite estimate of the outer product of the gradient is of reduced rank, making these tests unsuitable for our purposes. Only Robin and Smith (2000) offer a test that does permit singularity of the covariance matrix, but their weaker rank condition is shown to be violated in our case. Consequently, we develop a new test for the rank of the AOPG.

Once the number of indexes has been determined, we use the estimated AD functionals to construct moment conditions for the index parameters. These parameters are then estimated within the generalized method of moments (GMM) framework (Newey and McFadden, 1994). The parameter estimates for B are shown to be root-N-consistent and asymptotically normal. This extends work in the statistical literature where the multiple index model is interpreted as a regression-type model for dimension reduction. Xia, Tong, Li, and Zhu (2002) and Hristache, Juditsky, Polzehl, and Spokoiny (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 of indexes.

The paper is organized as follows. In Section 2, we discuss root-N-consistent, asymptotically normal, nonparametric estimators of average functionals of derivatives. In Section 3, we discuss identification of the parameters based on the AOPG and present the test for the number of indexes in a multiple-index model. In Section 4, we show how the AD functionals can be used within a GMM framework to yield a root-N-consistent, asymptotically normal estimator for the parameters of interest. In addition to providing theoretical results, we illustrate the estimator's usefulness in Section 5 with a simulation study on a multinomial choice model. Section 6 concludes with a general discussion. Proofs of the lemmas presented in the paper are given in the Appendix.

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 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 aforementioned set of AD functionals for two reasons. First, these functionals guarantee identification and are informative about most properties of $H(\cdot)$ that are of interest, such as additive separability and linearity (see Samarov, 1993). Second, the conditions that need to be

https://doi.org/10.1017/S0266466608080626 Published online by Cambridge University Press

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. Upon extending our estimator to the multiple equation setting, the set of moment conditions may include all moments for each equation, possibly extended by cross-equation moments.

Let the observed data $z_i = (y_i, x_i^T)^T$ $i = 1, \ldots, N$ constitute a random sample from a distribution with density $f^*(y, x)$. 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 $g'(x) = G'(x)/f(x) - G(x)f'(x)/f(x)^2$. Similarly an expression for the second-order derivatives can be obtained. Estimates of such derivatives can be obtained after estimation of their constituents, $f(x), f'(x), f''(x)$, $G(x), G'(x)$, and $G''(x)$. We use the Nadaraya–Watson leave-one-out kernel estimators, with kernel function $K(\cdot)$ and bandwidth parameter h, such that $h \to 0$ as $N \to \infty$ (cf. Härdle and Stoker, 1989). To avoid very imprecise contributions to the AD functionals for observations with low densities we introduce a 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. The presence of this trimming function, we point out, does not bias the parameter estimates for B.

The moments of interest and their kernel estimates are summarized in Table 1. Under conditions given by Samarov (1993), root-N-consistent, asymptotic normal estimates 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 use of a kernel of order $L + 3$ in combination with $h = c \cdot N^{-1/(2L+5)}$, would satisfy the assumptions on both the rate of convergence and the order of the kernel for all moments.[3]

Let $\hat{M}$ denote the vector containing all nonparametric estimates of the (unique) AD functionals considered, so that $\hat{M} = (\widehat{AD}^T, \text{vech}(\widehat{AOPG})^T, \text{vech}(\widehat{AH})^T)^T$. Using Samarov (1993), we have

$$\sqrt{N}(\hat{M} - M) \xrightarrow{d} N(0, \Sigma_M), \tag{3}$$

with $\Sigma_M = \text{Var}(R(x_i, y_i))$ where $R(x_i, y_i)$ is the vector with all stacked influence functions. The influence functions of the (weighted) AD, AOPG, and AH have been derived using second-order U-statistics theory (Hoeffding, 1948; Powell et al., 1989) and are given in Table 2.

3. PARAMETER IDENTIFICATION AND THE NUMBER OF INDEXES

The AD functionals provide information on the parameters of interest. For example, it is straightforward to show that

https://doi.org/10.1017/S0266466608080626 Published online by Cambridge University Press

https://doi.org/10.1017/S0266466608080626 Published online by Cambridge University Press

Table 1. Moments of interest and their kernel estimates

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

Table 2. Moments of interest and their influence functions

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	$R^{AH}(x,y) = w(x)\left\{g''(x) + (y - g(x))\frac{f''(x)}{f(x)}\right\}$

$$AOPG = \sum_{p,q} \mathrm{E}\left[w(x) \frac{\partial H(x^T\beta_1, \ldots, x^T\beta_P)}{\partial (x^T\beta_p)} \frac{\partial H(x^T\beta_1, \ldots, x^T\beta_P)}{\partial (x^T\beta_q)} \right] \beta_p \beta_q^T$$
$$\equiv B \Gamma^{OPG} B^T.$$

Here Γ^{OPG} is a $P \times P$ matrix of auxiliary parameters denoting the expectation of the relevant properties of the link function $H(\cdot)$.

The identification conditions reveal an equivalence between the number of indexes required and the rank of AOPG. By introducing a test for the rank of AOPG, we develop a test for the number of indexes required in a multiple-index model. Because estimation of the index parameters can only be performed conditional on the number of indexes to be estimated this advances the literature on semiparametric multiple-index models, where, so far, the number of indexes has been assumed known.

3.1. Identification

To establish conditions for identification of the parameters, we only focus on the AOPG.[4] We first turn to the number of indexes to be estimated, P, as this is an important determinant of the number of parameters in the model. Any assumption on the number of indexes in the model should therefore be informative about the number of parameters that can be identified. Recall that $g(x)$ can be written in the "multiple-index" form $g(x) = H(x^T\beta_1, \ldots, x^T\beta_P)$. The assumption that P is the minimum number of indexes required to appropriately model $\mathrm{E}(y|x)$ as $H(x^T\beta_1, \ldots, x^T\beta_P)$ can be formalized with two assumptions, which 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. $\mathrm{Rank}(B) \equiv \mathrm{rank}(\beta_1, \ldots, \beta_P) = P$.

Assumption 2. The function $H(\cdot)$ satisfies

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

Assumption 1 ensures that no fewer than P indexes are needed by ruling out multicollinearity of the indexes. The exclusion restrictions usually applied in semiparametric multi-index models (see, among others, Ichimura and Lee, 1991; Lee, 1995) are sufficient for this assumption to hold, but other restrictions are also possible. Assumption 2 asserts that each of the indexes provides unique information on the shape of $H(\cdot)$, that is, the derivatives of $H(\cdot)$ with respect to each of the indexes are not linearly dependent, almost everywhere; see also Ichimura and Lee (1991, Lem. 3, cond. 3).

Assumptions 1 and 2 ensure that only P indexes have to be estimated. However, without further assumptions on either B or $H(\cdot)$, the parameters in B are

https://doi.org/10.1017/S0266466608080626 Published online by Cambridge University Press

still not identified. To illustrate the well-known scaling and rotation problems present in semiparametric index models, we rewrite $H(x^T\beta_1, \ldots, x^T\beta_P)$ as $H(x^TB)$. Let B^* denote a given $L \times P$ matrix with columns spanning the column space of B and define $H^*(\cdot)$ as the unique function that satisfies $g(x) = H^*(x^TB^*)$. The identification problem stems from the fact that for any $P \times P$ matrix Λ of full rank $H^*(x^TB^*) = H_\Lambda(x^TB^*\Lambda)$ with $H_\Lambda(z) = H^*(z\Lambda^{-1})$. To attain identification Λ needs to be fixed, resulting in both $B^*\Lambda$ and H_Λ being uniquely determined. Consequently, P^2 restrictions need to be made to fully identify the remaining parameters. In accordance, the traditionally used exclusion and normalization restrictions impose P normalization and $P(P-1)$ exclusion restrictions.

Even though Assumption 1 guarantees the existence of a valid set of exclusion and normalization restrictions, exclusion restrictions are potentially incorrect, and identification might not be achieved. To avoid the use of possibly invalid restrictions, we obtain identification by combining Assumptions 1 and 2, which fix the number of indexes, with an assumption that fixes the scaling and rotation of the parameters. Such a set of identifying restrictions is given by orthonormality of the β's (Xia et al., 2002), so that $B^TB = I$, and diagonality of Γ^{OPG}, that is, $\Gamma^{OPG} = D$, a matrix with positive elements on its diagonal and zeros everywhere else. Taking into account the symmetry of B^TB and Γ^{OPG}, this again imposes exactly P^2 restrictions.[5] Note that these restrictions in combination with $AOPG = B\Gamma^{OPG}B^T$ correspond to the singular value decomposition of AOPG after ignoring the zero singular values, with Γ^{OPG} a diagonal matrix containing the nonzero singular values and B the corresponding eigenvectors.[6] As is well known, this decomposition is unique when the nonzero singular values are all distinct and are put in decreasing order (Magnus and Neudecker, 1988). Hence, to ensure uniqueness, the singular values corresponding to the β's have to be different. Validity of this assumption in applications is an empirical question that, importantly, can also be tested; the singular values are uniquely defined even if they are identical—only B will be underidentified in that case. Hence, inequality of the singular values results in a testable restriction, unlike the normalization and exclusion restrictions.

We summarize these identification restrictions in the following assumption.

Assumption 3. When P indexes are estimated with the use of at least the AOPG moments, one of the following two conditions holds.

(a) Each index $x^T\beta_p$, $p = 1, \ldots, P$, contains one explanatory variable that does not enter the other $P-1$ indexes. The parameters on these variables are normalized to equal 1.
(b) The β's are orthonormal; that is, $B^TB = I$, Γ^{OPG} is a diagonal matrix, and the nonzero elements in Γ^{OPG} are unique and sorted in descending order.

Extension to multiple equations. In the multiple equation setting with $y \in R^S$, let $H(x^TB) = (H_1(x^TB), \ldots, H_S(x^TB))^T$. One set of identifying assumptions,

https://doi.org/10.1017/S0266466608080626 Published online by Cambridge University Press

imposed, for example, by Picone and Butler (2000), is to impose Assumption 2 for each link function $H_s(x^TB)$, $s = 1, \ldots, S$. This might be too strong, as one index could play a role in one equation but not in another one; think of the exclusion restrictions in sample selection models. A much weaker requirement instead is that each index plays a role in at least one equation and that the contributions of the indexes in the equations, $\partial H_s/\partial(x^T\beta_p)$, $s = 1, \ldots, S$, do not have the same linear dependencies in all equations. By redefining $\Gamma^{OPG} = \sum_{s=1}^{S} \Gamma_s^{OPG}$ with $\Gamma_s^{OPG} = \mathrm{E}\{[w(x)(\partial H_s/\partial(x^T\beta_p))(\partial H_s/\partial(x^T\beta_q))]_{1\le p,q\le P}\}$, this is ensured by imposing Assumption 2. Assumption 3 also directly extends to the multiple equation setting with this redefinition of Γ^{OPG}. Conformably, we redefine $AOPG = \sum_{s=1}^{S} AOPG_s$ with $AOPG_s$ the AOPG of the sth equation.

3.2. A Test for the Number of Indexes

The identification assumptions, Assumptions 1 and 2, imply that, whenever P indexes have to be estimated, $AOPG = B\Gamma^{OPG}B^T$ has rank P. The number of indexes to be estimated can thus be inferred from the rank of AOPG, or an estimate thereof.

Tests for the rank of a matrix that is unobserved, but for which a root-N-consistent asymptotically normal estimator is available, have been developed by Cragg and Donald (1997), Gill and Lewbel (1992), Kleibergen and Paap (2006), and Robin and Smith (2000). Unfortunately, most tests require a full rank covariance matrix and are unsuitable for our purpose. As shown by Donald et al. (2007), the covariance matrix of the stacked elements of a positive semidefinite estimate of AOPG (AOPG is positive semidefinite and has a reduced rank in case $P < L$) cannot be of full rank. Singularity of Σ_M in (3) results, in particular in the rows and columns related to the AOPG moments. The reduced rank of Σ_M^{OPG}, the submatrix of Σ_M associated with the AOPG moments, is provided by Lemma 1, which follows. The proofs of all lemmas can be found in the Appendix.

LEMMA 1. *Given Assumptions 1 and 2, Σ_M^{OPG}, the submatrix of Σ_M associated with* AOPG*, has rank $PL - P(P-1)/2$.*

Only the test for the rank of a matrix that is presented by Robin and Smith (2000) imposes a weaker rank condition on the covariance matrix. The particular structure of our estimation problem construes a link between the eigenvectors of AOPG and the covariance matrix of its elements that renders the Robin and Smith test inappropriate, as stated in the next lemma.

LEMMA 2. *The (weaker) rank condition in Assumption 2.4 of Robin and Smith (2000) is not satisfied when the null hypothesis in the rank test is* $\mathrm{rank}(AOPG) = P$.

https://doi.org/10.1017/S0266466608080626 Published online by Cambridge University Press

Consequently, we set out to develop a test for the rank of AOPG that builds on the particular structure of our problem. A specific problem arises in the case that $P = 0$ and not a single variable affects the conditional expectation of the dependent variable. As this situation is unlikely to occur in practice, we will assume that at least one of the explanatory variables has some influence and hence $P \geq 1$ and $\text{rank}(\Sigma_M^{OPG}) \neq 0$.

To find the appropriate number of indexes, we sequentially test whether $\text{rank}(AOPG)$ equals r against the alternative that it exceeds r, $r = 1, 2, \ldots, L-1$, halting at the first value of r for which the statistic indicates nonrejection of the null $\text{rank}(AOPG) = r$, as suggested by Cragg and Donald (1997) and Robin and Smith (2000). To achieve a weakly consistent estimator of the number of indexes, α_N need to satisfy the conditions $\alpha_N \to 0$ and $-\log(\alpha_N)/N \to 0$.

We suggest testing $H_0 : \text{rank}(AOPG) = r$ versus $H_1 : \text{rank}(AOPG) > r$ by

$$\min_{(B,\Gamma^{OPG}) \text{ st rank} = \text{r}} [\sqrt{N}\ \text{vech}(\widehat{AOPG} - B\Gamma^{OPG}B^T)]^T \times [\sqrt{N}\ \text{vech}(\widehat{AOPG} - B\Gamma^{OPG}B^T)], \quad (4)$$

with critical values generated by drawing from $N(0, \hat{\Sigma}_M^{OPG})$, taking the sum of squares of the elements, and evaluating the $1 - \alpha_N$ quantile of these simulated values. Such a Monte-Carlo-based test, where we use a consistent estimate of Σ_M^{OPG}, is asymptotically valid by smoothness arguments (cf. Dufour, 2006). To improve the power of this test, one might use the knowledge on the number of zero eigenvalues in Σ_M^{OPG}. In particular, one could consistently estimate Σ_M^{OPG} using an approximation based on its singular value decomposition with the smallest $r^2 - Lr + r(r-1)/2$ eigenvalues replaced by zeros as suggested by Lütkepohl and Burda (1997, Prop. 2).

We note that a generalized Wald test approach, that is, considering $\min_{(B,\Gamma^{OPG}) \text{ st rank}=\text{r}} [\sqrt{N}\ \text{vech}(\widehat{AOPG} - B\Gamma^{OPG}B^T)]^T [\tilde{\Sigma}_M^{OPG}]^+ [\sqrt{N}\ \text{vech}(\widehat{AOPG} - B\Gamma^{OPG}B^T)]$, with $[\tilde{\Sigma}_M^{OPG}]^+$ a consistent estimator of the generalized inverse $[\Sigma_M^{OPG}]^+$, would not work in our setting as there are no overidentifying restrictions (cf. Lemma 1). This is exactly the reason why the test of Robin and Smith (2000) and the minimum χ^2 procedure suggested by Cragg and Donald (1997) do not work in our setting.

4. ESTIMATION

With the number of indexes in the model determined using the estimated AOPG, we turn to the method of moments framework to estimate the parameters of interest. Just as the AOPG provides information on the parameters of interest, so do the other AD functionals. In particular, $AD = B\Gamma^D$ and $AH = B\Gamma^H B^T$, where Γ^D and Γ^H denote matrixes of auxiliary parameters similar to Γ^{OPG}. These relationships provide a set of unique moment conditions that we use for esti-

https://doi.org/10.1017/S0266466608080626 Published online by Cambridge University Press

mation, denoted by $m(\theta_0) \equiv M - M_{B,\Gamma} = 0$, where θ_0 denotes the vector of all free parameters in B and Γ (identifiability). We make the following assumption about θ_0.

Assumption 4. θ_0 is in the interior of Θ with Θ compact.

Let $\hat{m}(\cdot)$ denote the estimated sample analogue of $m(\cdot)$, based on nonparametric kernel estimates, and note, using Section 2, that for $\theta = \theta_0$ and Σ_M positive semidefinite

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

The information on the rank of Σ_M provided by Lemma 1 has important consequences for a GMM estimator. It requires us to be careful when efficiently weighting the moment conditions for two reasons. First, Andrews (1987) has shown that replacing Σ_M by a consistent estimate does not necessarily ensure consistency of its inverse, and so a consistent estimate of the weighting matrix might not be available. Second, the reduced rank of Σ_M implies that certain (linear combinations of) moment conditions will receive zero weight, potentially resulting in a loss of identification. We therefore start with the unweighted GMM framework. Using the resulting estimate of B, we are then able to construct a consistent estimator for $[\Sigma_M]^+$, the pseudo-inverse of Σ_M. We then show that identification is not lost by efficiently weighting the moments.

The unweighted GMM estimator for θ_0 is given by

$$\hat{\theta}_{unw} = \arg\min_{\theta \in \Theta} \hat{m}(\theta)^T \hat{m}(\theta). \tag{5}$$

With consistency of $\hat{m}(\theta)$, $\hat{\theta}$ will be a consistent estimate of θ_0 when θ_0 uniquely minimizes $m(\theta)^T m(\theta)$, which is stated by the following lemma.

LEMMA 3. *Given Assumptions 1–3, $m(\theta)^T m(\theta)$ has a unique minimum at $\theta = \theta_0$.*

Following Newey and McFadden (1994), we obtain the following result.

THEOREM 1. *Given Assumptions 1–4, and $\sqrt{N}(\hat{M} - M) \xrightarrow{d} N(0, \Sigma_M)$,*

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

with $\Omega_{unw} = [m'(\theta_0)^T m'(\theta_0)]^{-1}[m'(\theta_0)^T \Sigma_M m'(\theta_0)][m'(\theta_0)^T m'(\theta_0)]^{-1}$.

To implement efficiently weighted GMM, a consistent estimate of $[\Sigma_M]^+$ is needed. As Andrews (1987) shows, $[\hat{\Sigma}_M]^+$ does not necessarily provide a consistent estimate of the weighting matrix $[\Sigma_M]^+$. Because the problem only concerns the AOPG moment conditions, we focus here on Σ_M^{OPG}. The inconsis-

https://doi.org/10.1017/S0266466608080626 Published online by Cambridge University Press

tency of $[\hat{\Sigma}_M^{OPG}]^+$ for $[\Sigma_M^{OPG}]^+$ stems from the fact that $L(L+1)/2 - LP + P(P-1)/2$ eigenvalues of $\hat{\Sigma}_M^{OPG}$ converge to zero as N goes to infinity, but they never reach zero in any finite sample. This affects the pseudo-inverse as is clear from the expressions for the pseudo-inverse based on the singular value decomposition. Let $A = D\Lambda D^T$; then $A^+ = D\Lambda^+ D^T$, with Λ^+ a diagonal matrix with the inverse eigenvalues of A, whenever these are nonzero, and zeros when the eigenvalues are zero. For nonzero eigenvalues, consistency of the eigenvalue ensures consistency of its inverse. For the zero eigenvalues, the estimated counterparts converge to zero, causing their inverses to explode, whereas they should converge to zero for consistency of the weighting matrix. This problem was also highlighted by Samarov (1993, Rmk. 3.2) and its relevance shown in his simulation study, where large t-values were obtained for a parameter known to be zero.

A solution is to determine what eigenvalues of $\hat{\Sigma}_M^{OPG}$ are estimates of zero and subsequently replace them by zero. In the proof of Lemma 1 in the Appendix we show $\Sigma_M^{OPG} = GVG^T$ with V of full rank, from which it follows that the zero eigenvalues of Σ_M^{OPG} correspond to eigenvectors in the null space of G. Let $\tilde{\Sigma}_M^{OPG}$ denote the matrix that results from $\hat{\Sigma}_M^{OPG}$ after projecting the rows and columns on the spaces spanned by G and G^T, respectively; then $\tilde{\Sigma}_M^{OPG}$ is a consistent estimate for Σ_M^{OPG}. Here G is unknown but depends only on the index vectors in B, and so we can obtain a consistent estimate of G using the unweighted GMM estimate of B. The major difference between $\tilde{\Sigma}_M^{OPG}$ and $\hat{\Sigma}_M^{OPG}$ is that $\tilde{\Sigma}_M^{OPG}$ has reduced rank. In particular, $L(L+1)/2 - LP + P(P-1)/2$ eigenvalues are zero instead of small but nonzero estimates of zero. The consequence is that $[\tilde{\Sigma}_M^{OPG}]^+$ is a consistent estimate for $[\Sigma_M^{OPG}]^+$, unlike $[\hat{\Sigma}_M^{OPG}]^+$. The argument readily extends to the full covariance matrix Σ_M: from $\hat{\Sigma}_M$ we project out the null space of G for all rows and columns corresponding to AOPG and subsequently obtain its pseudo-inverse $[\tilde{\Sigma}_M]^+$ to yield a consistent estimate of $[\Sigma_M]^+$. The reason that this covariance matrix does not suffer from the problem discussed by Andrews (1987) is that the space corresponding to the zero eigenvalues is projected out.

Our efficiently weighted GMM estimator of θ_0 is given by

$$\hat{\theta}_{eff} = \arg\min_{\theta \in \Theta} \hat{m}(\theta)^T [\tilde{\Sigma}_M]^+ \hat{m}(\theta). \tag{6}$$

To show consistency of the estimator, we need to reassess our identification result as we are ignoring information in some of the moments. This is formalized in the next lemma.

LEMMA 4. *Given Assumptions 1–3, $m(\theta)^T[\Sigma_M]^+ m(\theta)$ has a unique minimum at $\theta = \theta_0$.*

In line with Theorem 1 and Newey and McFadden (1994) we obtain our final result.

https://doi.org/10.1017/S0266466608080626 Published online by Cambridge University Press

THEOREM 2. *Given Assumptions 1–4, and* $\sqrt{N}(\hat{M} - M) \xrightarrow{d} N(0, \Sigma_M)$,

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

with $\Omega_{eff} = (m'(\theta_0)^T[\Sigma_M]^+ m'(\theta_0))^{-1}$.

Borrowing ideas from the continuously updated GMM estimator, where the covariance matrix is updated during the iterations in the optimization process, we propose to update $\widehat{\Sigma}_M$ also with the most recent estimate of B using the procedure described earlier.

5. SIMULATION STUDY

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

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

$$y_j^* = z_j^T \gamma_j - u_j,$$

where y_j^* is the latent variable associated with the random "utility" of each alternative. Because the observed choices are only informative on the differences between "utilities" and not on the "utilities" themselves, it is commonplace to define the multinomial choice model in differenced form. Define $x^T\beta_1 = z_1^T\gamma_1 - z_3^T\gamma_3$ and $x^T\beta_2 = z_2^T\gamma_2 - z_3^T\gamma_3$, 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_1 and e_2 only depend on x through the indexes $(x^T\beta_1, x^T\beta_2)$, the following multiple-index mean regression representation holds:

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

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

We let $X \sim N(0, I_3)$ and $\beta_1 = (1,0,1)^T$ and $\beta_2 = (0,1,1)^T$. Here $(e_1, e_2)^T$ is distributed independently of X, with bivariate normal distribution with unit variances and correlation $\rho_{e_1 e_2} = 0.5$ (which conforms to the assumptions of the multinomial probit model with independent and identically distributed [i.i.d.]

https://doi.org/10.1017/S0266466608080626 Published online by Cambridge University Press

errors on the utilities). With Φ_2 denoting the cumulative distribution function (c.d.f.) of a standard bivariate normal distribution, this yields, for example, $H_1(x^T\beta_1, x^T\beta_2) = \Phi_2(x^T\beta_1, (x^T\beta_1 - x^T\beta_2), 0.5)$. The sample size N is set at 1,000, and 500 replications are drawn.

With the number of explanatory variables equal to three ($L = 3$), our theoretical results imply the use of a sixth-order kernel, $r = L + 3$. The multivariate kernel function $K(\cdot)$ (on R^3) is chosen as the product of three univariate kernel functions given by $K(x) = ((45{,}045/2{,}048)x^8 - (31{,}185/512)x^6 + (59{,}535/1{,}024)x^4 - (11{,}025/512)x^2 + (4{,}725/2{,}048))1(|x| \leq 1)$. This kernel is symmetric and bounded and satisfies the assumption that the kernel and its derivative vanish at the boundary.

An appropriate bandwidth sequence $\{h_N\}$ is given by $h_N = cN^{-1/(2L+5)}$ $[= cN^{-1/11}]$, where c is a constant factor independent of N. The particular parameterization chosen for our multinomial choice 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 our knowledge of the true data generating process, we optimally select the bandwidth by minimizing the mean squared error of the nonparametrically estimated moments. Alternatively, one could use cross-validation, suitably scaled to obtain the proper rate of convergence, or apply a a "plug-in" estimator for the optimal bandwidth as discussed by Powell and Stoker (1996). Using a grid search algorithm and 100 simulations, we arrived at the bandwidths $(3.0, 3.0, 3.0)$.[7] No evidence for a need for distinct bandwidths for x_1 (x_2) and x_3 was found.

We use an indicator weighting function, $w(x_i) = 1(f(x_i) > b)$, which trims away the observations with small values of the density. With $b = 0.0025$, this yields a trimming of a bit less than 10% of the observations. Finally, we use the bootstrap estimator (with 500 bootstraps) to estimate the covariance matrix for the nonparametrically estimated moments, as in Samarov (1993). It is obtained as $(1/B)\sum_{b=1}^{B}(\hat{M}_b - \hat{M})(\hat{M}_b - \hat{M})^T$, where for each of the B bootstrapped samples an estimate $\hat{M}_b$ is obtained. The resulting estimate of the variance is comparable to the empirical variance for all moments.

In Figures 1(a) and 1(b) we present results from our sequential procedure on testing the rank of AOPG$(= \sum_s AOPG_s)$ matrix using different levels of significance for our multinomial choice model where the true rank (number of indexes) equals two. The figure shows the percentage of simulations (cumulative) for which the selected number of indexes in our multinomial choice model equals 1, 2, or 3. We compare the two proposed methods of generating critical values. Although both are based on Σ_M^{OPG}, one method improves power by ignoring the smallest eigenvalues, which are known to be zero under the null hypothesis.

Figure 1 reveals that for most levels of significance our test selects the true number of indexes. Choosing a very low level of significance might lead to inappropriately selecting a single-index specification, whereas choosing a very large significance level (in excess of 60%) might lead to accepting a fully non-

https://doi.org/10.1017/S0266466608080626 Published online by Cambridge University Press

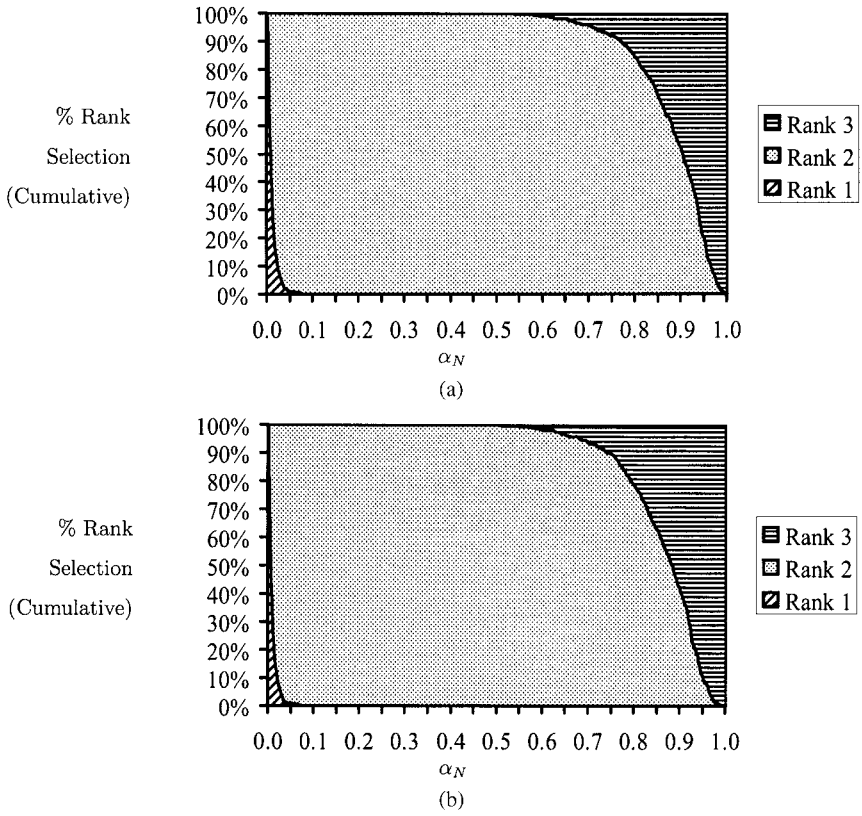


FIGURE 1. Test of rank of AOPG using (a) Σ_M and (b) Σ_M excluding small eigenvalues.

parametric specification (where no dimension reduction is deemed acceptable). Ignoring the smallest eigenvalues in generating the critical values yields a slight improvement in power against selecting too few indexes at all levels of significance. To ensure the selection of sufficient indexes for our multinomial choice model, we advocate, in agreement with Robin and Smith (2000), not choosing the significance level, α_N, too conservatively (small). As expected our test provides an empirical level below the nominal level as it is based on a bounded critical value.

In Table 3, we present the numerical results for the 5 and 10% significance levels. Clearly, at these levels of significance, our test, based on testing the rank of AOPG, is able to find the correct number of indexes for our multinomial choice model.

Table 4 presents an analysis of the index parameter estimates for the unweighted and efficiently weighted method of moments estimator. When efficient weighting is applied, the weights are updated with the most recent value

https://doi.org/10.1017/S0266466608080626 Published online by Cambridge University Press

TABLE 3. Testing for the number of indexes

Significance level (α_N)	Critical value simulation procedure	Number of indexes selected by test		
		1	2	3
0.05	Using Σ_M	1.0%	99.0%	0.0%
	Using Σ_M excluding smallest eigenvalues	0.8%	99.2%	0.0%
0.10	Using Σ_M	0.0%	100.0%	0.0%
	Using Σ_M excluding smallest eigenvalues	0.0%	100.0%	0.0%

TABLE 4. Parameter estimates, multinomial choice model

	True	Mean	Variance (sample)	Variance (theory)	LQ	Median	UQ	MAE
GMM unweighted (AOPG only)								
$\beta_{1,3}$	1.000	0.985	0.025	0.031	0.862	0.975	1.090	0.129
$\beta_{2,3}$	1.000	0.988	0.023	0.031	0.874	0.975	1.091	0.121
GMM unweighted (AOPG/AD/AH)								
$\beta_{1,3}$	1.000	0.979	0.022	0.024	0.868	0.972	1.076	0.120
$\beta_{2,3}$	1.000	0.985	0.019	0.024	0.895	0.983	1.065	0.108
GMM efficiently weighted (AOPG only)								
$\beta_{1,3}$	1.000	0.982	0.021	0.025	0.871	0.977	1.082	0.119
$\beta_{2,3}$	1.000	0.986	0.019	0.026	0.882	0.979	1.081	0.111
GMM efficiently weighted (AOPG/AD/AH)								
$\beta_{1,3}$	1.000	1.007	0.014	0.012	0.928	0.993	1.081	0.091
$\beta_{2,3}$	1.000	1.011	0.012	0.012	0.933	1.006	1.084	0.087
Maximum likelihood estimation (MLE) 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

of B that is available during the iterations. To illustrate the benefits of using more moments, we present the results when only AOPG is used and those when AD and AH are included also. We only estimate the parameters for $P = 2$, the true number of indexes, where we impose normalization and exclusion restrictions and estimate $\beta_{1,3}$ and $\beta_{2,3}$.[8]

For comparison, the multinomial probit (MNP) parameter estimates are reported also, where the same exclusion restrictions have been imposed. To ensure comparability of MNP with the semiparametric estimates, where $\beta_{1,1}$

https://doi.org/10.1017/S0266466608080626 Published online by Cambridge University Press

and $\beta_{2,1}$ are normalized to 1, we report $\hat{\beta}^{(p)}_{1,3}/\hat{\beta}^{(p)}_{1,1}$ and $\hat{\beta}^{(p)}_{2,3}/\hat{\beta}^{(p)}_{2,1}$, where $\hat{\beta}^{(p)}$ are the MNP parameter estimates.

The table presents the following summary statistics: sample mean, sample variance, average of the theoretical variance, lower quartile (LQ), median, upper quartile (UQ), and mean absolute error (MAE). From Table 4 it is clear that the parameters can be obtained from the nonparametrically estimated moment conditions quite accurately.

As expected, adding more moments and making use of efficient weighting in the GMM procedure increases efficiency as indicated by the reductions in both the theoretical and the sample variances. The total efficiency gain is almost 50%, which is quite substantial. The theoretical values of the variances are slightly larger than the actual variation in estimates that we observe, except for the situation of efficiently weighting all moments. The loss in efficiency arising from not knowing the distribution of the disturbances is within reason; the variance of the semiparametric estimates using efficient weighting of all available moments is less than double that of the parametric ones.

6. DISCUSSION

We consider the estimation of semiparametric multi-index models for single and multiple equation models. Existing estimation methods have assumed the number of indexes to be given, either by economic theory or by the researcher—no data driven procedures were considered to determine this. We advance the literature by providing a procedure to determine the number of indexes nonparametrically. For single equation models, we show that the rank of the outer product of the gradient equals the number of indexes required in the semiparametric model. For multiple-index models this generalizes such that the number of indexes equals the rank of the sum of the outer product of the gradient of the separate equations. A test for the rank of the outer product of the gradient then provides the desired testing procedure. Nonparametrically estimated AD functionals are then used to estimate the index parameters within a GMM framework. This step involves a simple minimization problem where, importantly, no additional kernel-based calculations are required. The resulting estimator is root-N-consistent and asymptotically normal.

In closing we point out three fruitful areas for further research addressing limitations and interesting extensions to our approach.

First, derivative-based methods have the obvious drawback relating to the estimation of parameters associated with discrete explanatory variables. There are various ways that discrete explanatory variables could be incorporated into our analysis. First of all, our approach readily extends to the situation where the discrete explanatory variables are treated fully nonparametrically, that is, where the conditional mean is given by $\mathrm{E}(y|x_c, x_d) = H(x_c^T\beta_1, \ldots, x_c^T\beta_P, x_d)$, with x_c the continuous and x_d the discrete explanatory variables. Alternatively,

https://doi.org/10.1017/S0266466608080626 Published online by Cambridge University Press

we can follow Fan, Härdle, and Mammen (1998), who estimated an additive partially linear model where the discrete variables enter the model in a linear fashion, so that $\mathrm{E}(y|x_c, x_d) = H(x_c^T\beta_1, \ldots, x_c^T\beta_P) + x_d^T\alpha$. Finally, under the assumption that each discrete variable enters in only a single known index, the work of Horowitz and Härdle (1996) easily generalizes.

Second, our setup precludes functionally related regressors. Clearly, adding deterministic transformations of the regressors, for example, interactions, to the set of regressors might induce a lower requirement on the number of indexes. Accordingly, the number of indexes can no longer be derived from the AOPG once functionally related regressors are allowed for. Indeed, more restrictions are needed on $H(\cdot)$ and the functionally related regressors to distinguish between, for example, the single-index model $\mathrm{E}\{y|x\} = H_1(x_1 + x_1^2\gamma + x_2^T\beta)$ and the multi-index model $\mathrm{E}\{y|x\} = H_2(x_1, x_2^T\beta)$ with $H_2(z_1, z_2) = H_1(z_1 + z_1^2\gamma + z_2)$. The determination of the required number of indexes in combination with the necessary additional model assumptions is beyond the scope of the current paper.

Finally, even though our estimator is clearly more efficient than the ADE estimator, further efficiency gains might be achieved. In light of the spanning condition argument in Newey (2004), adding more moment conditions will likely increase the efficiency of the estimator. Alternatively, strengthening the conditional mean index assumption to a conditional distribution index assumption permits a one-step efficient update, as suggested by Delecroix, Härdle, and Hristache (2003).

NOTES

1. To clarify our 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 transpose of a matrix.

2. An exception is the projection pursuit model (Friedman and Stuetzle, 1981) where the restrictive assumption that all indexes enter additively separable in the model is imposed.

3. Our definition of higher order kernels accords with Härdle and Stoker (1989) and Powell et al. (1989). Donkers and Schafgans (2005) extend the results of Samarov (1993) to enable the use of such higher order kernels.

4. We focus on the AOPG, as it permits identification under a minimal set of identifying restrictions. There exist situations where the AD and AH do not provide information on the index parameters, leaving them unidentified (Xia et al., 2002).

5. Once the matrix B is estimated satisfying this set of identifying restrictions it is always possible, retroactively, 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. The results in Phillips (1994), where the space spanned by cointegrating vectors is considered, suggest that when knowledge on exclusion and normalization restrictions is available, they should be used in light of the superior small-sample performance.

6. This result suggests that applying the singular value decomposition to AOPG will also provide estimates of B and Γ^{OPG}.

7. Cross-validation yielded bandwidths, after adjusting for the rate, that are similar in magnitude to those obtained by minimizing the mean squared error of the nonparametrically estimated moments.

https://doi.org/10.1017/S0266466608080626 Published online by Cambridge University Press

8. We also estimated the parameters using moments estimated with a second-order (quartic) kernel. The results were comparable, and the small differences were often in favor of the (inappropriate) lower order kernel. These results and estimation results for the numerous auxiliary parameters can be obtained from the authors.

REFERENCES

Andrews, D.W.K. (1987) Asymptotic results for generalized Wald tests. *Econometric Theory* 3, 348–358.

Cragg, J.G. & S.G. Donald (1997) Inferring the rank of a matrix. *Journal of Econometrics* 76, 223–250.

Delecroix, M., W. Härdle, & M. Hristache (2003) Efficient estimation in conditional single-index regression. *Journal of Multivariate Analysis* 86, 213–226.

Donald, S.G., N. Fortuna, & V. Pipiras (2007) On rank estimation in symmetric matrices: The case of indefinite matrix estimators. *Econometric Theory* 23, 1217–1232.

Donkers, B. & M. Schafgans (2005) A Method of Moments Estimator for Semiparametric Index Models. Sticerd Working paper EM/2005/493, London School of Economics.

Dufour, J.-M. (2006) Monte Carlo tests with nuisance parameters: A general approach to finite-sample inference and nonstandard asymptotics. *Journal of Econometrics* 133, 443–477.

Fan, J., W. Härdle, & E. Mammen (1998) Direct estimation of low-dimensional components in additive models. *Annals of Statistics* 26, 943–971.

Friedman, J.H. & W. Stuetzle (1981) Projection pursuit regression. *Journal of the American Statistical Association* 76, 817–823.

Gill, L. & 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.

Härdle, W. & T.M. Stoker (1989) Investigating smooth multiple regression by the method of average derivatives. *Journal of the American Statistical Association* 84, 986–995.

Hoeffding, W. (1948) A class of statistics with asymptotically normal distribution. *Annals of Mathematical Statistics* 19, 293–325.

Horowitz, J.L. & W. Härdle (1996) Direct semiparametric estimation of single-index models with discrete covariates. *Journal of the American Statistical Association* 91, 1632–1640.

Hristache, M., A. Juditsky, J. Polzehl, & V. Spokoiny (2001) Structure adaptive approach for dimension reduction. *Annals of Statistics* 29, 1537–1566.

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

Kleibergen, F. & R. Paap (2006) Generalized reduced rank tests using the singular value decomposition. *Journal of Econometrics* 133, 97–126.

Lee, L.-F. (1995) Semiparametric maximum likelihood estimation of polychotomous and sequential choice models. *Journal of Econometrics* 65, 381–428.

Lütkepohl, H. & M.M. Burda (1997) Modified Wald tests under nonregular conditions. *Journal of Econometrics* 78, 315–332.

Magnus, J.R. & H. Neudecker (1988) *Matrix Differential Calculus with Applications in Statistics and Econometrics.* Wiley.

Newey, W.K. (2004) Efficient semiparametric estimation via moment restrictions. *Econometrica* 72, 1877–1998.

Newey, W.K. & D.L. McFadden (1994) Large sample estimation and hypothesis testing. In R.F. Engle & D.L. McFadden (eds.), *Handbook of Econometrics*, vol. 4, pp. 2111–2245. North-Holland.

Phillips, P.C.B. (1994) Some exact distribution theory for maximum likelihood estimators of cointegrating coefficients in error correction models. *Econometrica* 62, 73–93.

https://doi.org/10.1017/S0266466608080626 Published online by Cambridge University Press

Picone, G.A. & J.S. Butler (2000) Semiparametric estimation of multiple equation models. *Econometric Theory* 16, 551–575.
Powell, J.L., J.H. Stock, & T.M. Stoker (1989) Semiparametric estimation of weighted average derivatives. *Econometrica* 57, 1403–1430.
Powell, J.L. & T.M. Stoker (1996) Optimal bandwidth choice for density-weighted averages. *Journal of Econometrics* 75, 291–316.
Robin, J.-M. & R.J. Smith (2000) Tests of rank. *Econometric Theory* 16, 151–175.
Robinson, P.M. (1989) Hypothesis testing in semiparametric and nonparametric models for econometric time series. *Review of Economic Studies* 56, 511–534.
Samarov, A.M. (1993) Exploring regression structure using nonparametric functional estimation. *Journal of the American Statistical Association* 88, 836–847.
Stoker, T. M. (1991) Equivalence of direct, indirect, and slope estimators of average derivatives. In W.A. Barnett, J. Powell, & G.E. Tauchen (eds.), *Nonparametric and Semiparametric Estimation Methods in Econometrics and Statistics*, pp. 99–118. Cambridge University Press.
Xia, Y., H. Tong, W.K. Li, & L.-X. Zhu (2002) An adaptive estimation of dimension reduction space. *Journal of the Royal Statistical Society, Series B* 64, 363–388.

APPENDIX: Proofs

Proof of Lemma 1. We are interested in evaluating the rank of $\Sigma_M \equiv \mathrm{Var}(\mathrm{vech}(R(z_i)))$, where $R(z_i) = g'(x_i)g'(x_i)^T - (y_i - g(x_i))[(f'(x_i)g'(x_i)^T/f(x_i)) + (g'(x_i)f'(x_i)^T/f(x_i)) + 2g''(x_i)]$. Using $g(x_i) = H(x_i^T B) \equiv H_i$ and denoting $y_i - g(x_i)$ by ε_i, we rewrite $R(z_i)$ as

$$R(z_i) = Z_i B^T + BZ_i^T,$$

with $Z_i = B(H_i' H_i'^T/2 - \varepsilon_i H_i'') - \varepsilon_i (f_i'/f_i) H_i'^T$. Here Z_i is an $L \times P$ matrix; H' and H'' are the first two derivatives of H with respect to $x^T B$.

Because $\mathrm{rank}(\Sigma_M) \equiv \mathrm{rank}(\mathrm{Var}(\mathrm{vec}(R(z_i)))$, we focus on properties of $\mathrm{vec}(R(z_i))$. Let $e_j^{(L)}$ denote the jth column of I_L, the L-dimensional identity matrix; then $D \equiv (I_L \otimes e_1^{(L)} | \ldots | I_L \otimes e_P^{(L)})$ is a full rank matrix with $\mathrm{vec}(Z_i^T) = D\mathrm{vec}(Z_i)$ (see Magnus and Neudecker, 1988). We then obtain

$$\mathrm{vec}(R(z_i)) = (B \otimes I_L)\mathrm{vec}(Z_i) + (I_L \otimes B)\mathrm{vec}(Z_i^T) = G \cdot \mathrm{vec}(Z_i),$$

with G, an $L^2 \times LP$ matrix, defined as

$$G = (B \otimes I_L) + (I_L \otimes B)D = [b_1 \otimes I_L + I_L \otimes b_1 | \ldots | b_P \otimes I_L + I_L \otimes b_P].$$

Denoting $V = \mathrm{Var}(\mathrm{vec}(Z_i))$, we obtain $\mathrm{rank}(\Sigma_M) \equiv \mathrm{rank}(GVG^T)$. With Z_i nonsymmetric and given nonzero variability of $(f_i'/f_i)H_i'^T$, implied by Assumption 2, and continuous differentiability of f', implied by the regularity conditions underlying Samarov (1993), V is positive definite and

$$\mathrm{rank}(\Sigma_M) = \mathrm{rank}(G).$$

To investigate the column rank of G, we study the dimension of the null space of G, that is, the space spanned by $v \in R^{PL}$ for which $Gv = 0$. Let $\Delta = [B|B^C]$ with B^C the orthogonal complement of B. The columns of Δ form a basis of R^L, or $B \oplus B^C = R^L$.

https://doi.org/10.1017/S0266466608080626 Published online by Cambridge University Press

We consider the basis $I_P \otimes \Delta$ for R^{PL} and let δ_l denote the lth column of Δ. The vectors in the R^{PL} space are then characterized by $e_j^{(P)} \otimes \delta_l$, $j = 1, \ldots, P$, $l = 1, \ldots, K$. For $j \neq l$, we observe $G(e_j^{(P)} \otimes \delta_l) = b_j \otimes \delta_l + \delta_l \otimes b_j$. Because $\delta_l = b_l$ for $l = 1, \ldots, P$, this implies that we can find $P(P-1)/2$ pairs of vectors $v = e_j^{(P)} \otimes \delta_l - e_l^{(P)} \otimes \delta_j \neq 0$ with $Gv = 0$. The column rank therefore equals $LP - P(P-1)/2$.

To investigate the row rank of G, we write G^T as

$$\begin{pmatrix} b_{11} I_L + e_1^{(L)} \otimes b_1^T & \ldots & b_{1k} I_L + e_P^{(L)} \otimes b_1^T \\ \\ b_{P1} I_L + e_1^{(L)} \otimes b_P^T & \ldots & b_{Pk} I_L + e_P^{(L)} \otimes b_P^T \end{pmatrix}.$$

Hence we study the space spanned by $v \in R^{L^2}$ for which $G^T v = 0$. We consider the basis $\Delta \otimes \Delta$ for R^{L^2}, or equivalently the space spanned by $(B \otimes B, B^C \otimes B^C, B \otimes B^C, B^C \otimes B)$. It is easy to see that $G^T[(B^C \otimes B) - (B \otimes B^C)] = 0$ and $G^T(B^C \otimes B^C) = 0$. Moreover, $G^T(b_j \otimes b_l) = e_j^{(P)} \otimes b_l + e_l^{(P)} \otimes b_j$, and so for all pairs $j, l = 1, \ldots, p, j \neq l$ $G^T[(b_j \otimes b_l) - (b_l \otimes b_j)] = 0$. Therefore the null space is of order $(L-P)P + (L-P)^2 + P(P-1)/2 = L^2 - LP + P(P-1)/2$, rendering the row rank of G equal to the column rank $LP - P(P-1)/2$. ■

Proof of Lemma 2. To test for rank$(AOPG) = P$, Assumption 2.4 of Robin and Smith requires Rank$[(D_{L-P} \otimes C_{L-P})^T \Omega (D_{L-P} \otimes C_{L-P})] > 0$, with D_{L-P} and C_{L-P} the left and right eigenvectors of AOPG and Ω the part of Σ_M that corresponds to AOPG. As AOPG is symmetric, D_{L-P} and C_{L-P} are identical, and we use C_{L-P} to denote the $L - P$ eigenvectors that correspond to the zero eigenvalues of AOPG (with P denoting the true rank, i.e., the true number of indexes). Assumption 2.4 of Robin and Smith then simplifies to Rank$[(C_{L-P} \otimes C_{L-P})^T \Omega (C_{L-P} \otimes C_{L-P})] > 0$. From $AOPG = B \Gamma^{OPG} B^T$ it follows that C_{L-P} contains the $L - P$ vectors in B^C from the orthonormal basis $\Delta = [B | B^C]$ defined in the proof of Lemma 1.

Recall from the proof of Lemma 1 that $\Omega = GVG^T$ and $G^T(B^C \otimes B^C) = 0$. Using $C_{L-P} = B^C$, Assumption 2.4 translates into rank$((B^C \otimes B^C)^T GVG^T (B^C \otimes B^C)) > 0$. Because $G^T(B^C \otimes B^C) = 0$, Assumption 2.4 of Robin and Smith (2000) is not satisfied when the rank being tested is the true rank of AOPG. ■

Proof of Lemma 3. The objective function obviously has a minimum of zero when $\theta = \theta_0$. To show that this minimum is unique, we need to show that $m'(\theta)$ has full column rank when evaluated at $\theta = \theta_0$. As the AOPG moments identify B, we focus on the column rank of $m'_{OPG}(\theta)$. It is straightforward to show that including additional moments increases the rank of $m'(\theta)$ such that the auxiliary parameters corresponding to these moment conditions are also uniquely determined.

To prove that $m'_{OPG}(\theta)|_{\theta=\theta_0}$ has full column rank, we use a slightly different set of identifying assumptions, replacing orthonormality of B with orthogonality of its columns in combination with additional restrictions on Γ^{OPG}, in particular $\Gamma^{OPG} = I$. This allows us to write $m_{OPG}(\theta_0) = \text{vec}(AOPG - BB^T) = \text{vec}(AOPG) - \sum_{s=1}^{P} (b_s \otimes b_s)$. As the change in identifying assumptions results in a one-to-one transformation of the parameters, at least for $\theta = \theta_0$, the proof holds under all possible sets of (valid) identifying restrictions. We assume that for $p = 1, \ldots, P-1$ the upper left $p \times p$ block of B is invertible. Assumption 1 indicates that this can always be realized by appropriately ordering of the x's.

https://doi.org/10.1017/S0266466608080626 Published online by Cambridge University Press

Let B_l denote the matrix consisting of the first l columns of B. Then with $B_{l,-j}$ we denote B_l after removing the first j rows, and similarly, $b_{l,-j}$ denotes b_l after removing its first j rows. In addition, $B_l^{(l)}$ denotes the upper $l \times l$ block. To illustrate the orthogonality assumptions on B we further introduce

$$B = \begin{bmatrix} b_{11} & c_{21} & c_{31} & \\ b_{12} & b_{21} & c_{32} & \\ & & b_{31} & \dots \\ b_{1L} & b_{2L} & b_{3L} & \end{bmatrix} = \left[b_1 \middle| \begin{matrix} c_2 \\ b_{2,-1} \end{matrix} \middle| \begin{matrix} c_3 \\ b_{3,-2} \end{matrix} \middle| \cdots \right] = [B_{p-1} | b_p | \cdots]$$

$$= \left[\begin{matrix} B_{p-1}^{(p-1)} \\ B_{p-1,-(p-1)} \end{matrix} \middle| \begin{matrix} c_p \\ b_{p,-(p-1)} \end{matrix} \middle| \cdots \right],$$

through which (a) the vector of identified parameters is given by $\tilde{B} = (b_1, b_{2,-1}, \dots, b_{P,-(P-1)})$ and (b) the constraints imposed by the orthogonality assumptions ($B_{s-1}^T b_s = 0$) yield

$$c_s = -(B_{s-1}^{(s-1)T})^{-1} B_{s-1,-(s-1)}^T b_{s,-(s-1)}, \qquad s = 2, \dots, P.$$

Note that the invertibility of $B_{s-1}^{(s-1)T}$ is guaranteed by the ordering of the x's.

We can show that

$$m'(\theta_0) = \frac{\partial \text{vec}(AOPG - B\Gamma^{OPG}B^T)}{\partial \text{vec}(B)^T} \frac{\partial \text{vec}(B)}{\partial \text{vec}(\tilde{B})^T} = GZ,$$

where G is as defined in the proof of Lemma 1. Define $R_p = -(B_p^{(p)T})^{-1}$; then Z (a $PL \times (PL - P(P-1))$ matrix) is given by

$$Z = \begin{pmatrix} I_L & 0 & 0 & 0 & 0 \\ \begin{pmatrix} R_1 b_2^T \\ 0 \end{pmatrix} & \begin{pmatrix} R_1 B_{1,-1} \\ I_{L-1} \end{pmatrix} & 0 & 0 & 0 \\ \begin{pmatrix} R_2(e_1^{(2)} \otimes b_3^T) \\ 0 \end{pmatrix} & \begin{pmatrix} R_2(e_2^{(2)} \otimes b_{3,-1}^T) \\ 0 \end{pmatrix} & \begin{pmatrix} R_2 B_{2,-2} \\ I_{L-2} \end{pmatrix} & 0 & 0 \\ \begin{pmatrix} R_3(e_1^{(3)} \otimes b_4^T) \\ 0 \end{pmatrix} & \begin{pmatrix} R_3(e_2^{(3)} \otimes b_{4,-1}^T) \\ 0 \end{pmatrix} & \begin{pmatrix} R_3(e_3^{(3)} \otimes b_{4,-2}^T) \\ 0 \end{pmatrix} & \ddots & 0 \\ & \dots & & & \ddots \end{pmatrix}. \tag{A.1}$$

Let $l > j$. The existence of a vector $v \in R^{PL - P(P-1)/2}$ for which $Zv = e_j^{(L)} \otimes b_l - e_l^{(L)} \otimes b_j$ would provide a vector in the null space of G, leading to a reduced rank of $m'_{OPG}(\theta_0)$. Conditions under which such a vector exists, therefore, provide the conditions that result in a lack of identification. Because of the structure of Z, if such a vector exists, it has to be a $PL - P(P-1)/2$-dimensional vector consisting of P stacked vector "blocks" of diminishing dimensions L, $L-1$, till $L-P$, with zeros everywhere except

https://doi.org/10.1017/S0266466608080626 Published online by Cambridge University Press

$b_{l,-(j-1)}$ in the jth block and $-b_{j,-(l-1)}$ in the lth block, so that

$$v = (0 \quad b_{l,-(j-1)} \quad 0 \quad -b_{j,-(l)-1} \quad 0)^T. \tag{A.2}$$

The zeros represent zero row vectors of appropriate dimensions.

Let Z^1 represent the first L rows of Z, Z^2 the second set of L rows, and so on. We consider the multiplication of Z times v for each of these sets of L rows. In particular, consider the sets of equations

$$Z^t v = \begin{cases} 0 & \text{for } t \neq j, l \\ b_l & \text{for } t = j \\ -b_j & \text{for } t = l. \end{cases}$$

For our purposes, it suffices to show that there is no v with $Z^l v = -b_j$ and hence $GZv = 0$ iff $v = 0$. We only need to consider the first $j - 1$ rows, as only these rows involve multiplication with matrixes other than the identity matrix. Using (A.1) and (A.2) we get

$$R_{l-1}(e_j^{(l-1)} \otimes b_{l,-(j-1)}^T) b_{l,-(j-1)} - R_{l-1} B_{l-1,-(l-1)} b_{j,-(l-1)} = \begin{pmatrix} -b_{j1} \\ -b_{j,l-1} \end{pmatrix}, \quad \text{or}$$

$$e_j^{(l-1)} \otimes b_{l,-(j-1)}^T b_{l,-(j-1)} - \begin{pmatrix} b_{1,-(l-1)}^T b_{j,-(l-1)} \\ \vdots \\ b_{l-1,-(l-1)}^T b_{j,-(l-1)} \end{pmatrix} = -(R_{l-1})^{-1} \begin{pmatrix} b_{j1} \\ b_{j,l-1} \end{pmatrix}.$$

Rewriting allows us to obtain

$$\begin{pmatrix} -b_{1,-(l-1)}^T b_{j,-(l-1)} \\ \cdots \\ b_{l,-(j-1)}^T b_{l,-(j-1)} - b_{j,-(l-1)}^T b_{j,-(l-1)} \\ \cdots \\ -b_{l-1,-(l-1)}^T b_{j,-(l-1)} \end{pmatrix} = \begin{pmatrix} \sum_{s=1}^{l-1} b_{1s} b_{js} \\ \cdots \\ \sum_{s=1}^{l-1} b_{js}^2 \\ \cdots \\ \sum_{s=1}^{l-1} b_{(l-1)s} b_{js} \end{pmatrix}$$

$$\text{or} \quad \begin{pmatrix} -b_1^T b_j \\ \cdots \\ b_{l,-(j-1)}^T b_{l,-(j-1)} - b_j^T b_j \\ \cdots \\ -b_{l-i}^T b_j \end{pmatrix} = 0,$$

which provides the condition $b_{l,-(j-1)}^T b_{l,-(j-1)} = b_j^T b_j$ (orthogonality ensures that the other equalities are satisfied automatically). The assumption that the singular values are

https://doi.org/10.1017/S0266466608080626 Published online by Cambridge University Press

unique and sorted in descending order, however, ensures under the current normalization that

$$b_{l,-(j-1)}^T b_{l,-(j-1)} \leq b_l^T b_l < b_j^T b_j,$$

from which we deduce that there exists no vector $v \in R^{PL-P(P-1)/2}$ for which $m'_{OPG}(\theta_0)v \equiv GZv = 0$, and full column rank of $m'_{OPG}(\theta_0)$ is shown. ■

Proof of Lemma 4. We follow the proof of Lemma 3, noting that information is only lost for the AOPG moments. We exploit the structure of Σ_M^{OPG} and the fact that its nonzero eigenvalues correspond to a set of eigenvectors that spans the space that is spanned by G^T. Let $D\Lambda D^T$ denote the singular value decomposition of Σ_M^{OPG} after exclusion of the zero eigenvalues and the corresponding eigenvectors; then the space spanned by the $PL - P(P-1)/2$ columns in D is the space spanned by G. (This follows directly from $\Sigma_M^{OPG} = GVG^T = GD_V\Lambda_V D_V^T G^T$ where $D_V\Lambda_V D_V^T$ is the singular value decomposition of V and the fact that D_V has full rank.) The objective function of efficiently weighted GMM can then be rewritten as $m_{OPG}^T(\theta)D\Lambda^{-1}D^T m_{OPG}(\theta)$. To ensure uniqueness it suffices to show that $D^T m'_{OPG}(\theta)$ has full column rank, as Λ is full rank.

In the proof of Lemma 3, the full column rank of $m'_{OPG}(\theta)$ was proved by showing that $GZv = 0$ iff $v = 0$. In line with this, full column rank of $D^T m'_{OPG}(\theta)$ is equivalent to $G^T GZv = 0$ iff $v = 0$, as the null space of D^T is that of G^T. Realizing that

$$G^T G = 2\begin{pmatrix} I_L + b_1^T \otimes b_1 & b_1^T \otimes b_2 & \dots & b_1^T \otimes b_P \\ b_2^T \otimes b_1 & I_L + b_2^T \otimes b_2 & \dots & b_2^T \otimes b_P \\ \vdots & & \ddots & \vdots \\ b_P^T \otimes b_1 & \dots & \dots & I_L + b_P^T \otimes b_P \end{pmatrix},$$

it is straightforward to show that the same set of restrictions on v that had to hold for $GZv = 0$ in the proof of Lemma 3 prevails. Hence, $G^T GZv = 0$ iff $v = 0$, showing full column rank of $D^T m'_{OPG}(\theta)$. ■

https://doi.org/10.1017/S0266466608080626 Published online by Cambridge University Press