ARTICLE IN PRESS

Journal of Econometrics xxx (xxxx) xxx

Contents lists available at ScienceDirect

Journal of Econometrics

journal homepage: www.elsevier.com/locate/jeconom

High dimensional semiparametric moment restriction models

Chaohua Dong [a], Jiti Gao [b], Oliver Linton [c,*]

[a] Zhongnan University of Economics and Law, China
[b] Monash University, Australia
[c] University of Cambridge, UK

ARTICLE INFO

Article history:
Received 15 October 2019
Received in revised form 3 February 2021
Accepted 17 July 2021
Available online xxxx

JEL classification:
C12
C14
C22
C30

Keywords:
Generalized method of moments
High dimensional models
Moment restriction
Over-identification
Penalization
Sieve method
Sparsity

ABSTRACT

We consider nonlinear moment restriction semiparametric models where both the dimension of the parameter vector and the number of restrictions are divergent with sample size and an unknown smooth function is involved. We propose an estimation method based on the sieve generalized method of moments (sieve-GMM). We establish consistency and asymptotic normality for the estimated quantities when the number of parameters increases modestly with sample size. We also consider the case where the number of potential parameters/covariates is very large, i.e., increases rapidly with sample size, but the true model exhibits sparsity. We use a penalized sieve GMM approach to select the relevant variables, and establish the oracle property of our method in this case. We also provide new results for inference. We propose several new test statistics for the over-identification and establish their large sample properties. We provide a simulation study and an application to data from the NLSY79 used by Carneiro et al. (2011).

1. Introduction

We consider a class of semiparametric models with Euclidean valued parameters and unknown function valued parameters, where we allow the number of covariates and hence Euclidean valued parameters to be large, i.e., to increase to infinity. In the first part of the paper we consider the case where the dimensionality is moderate, and in the second part of the paper we consider the case where the number of possible covariates is extremely large. Large models are the focus of much current research, see Athey et al. (2017), Belloni et al. (2014c). We suppose that the model is defined through a set of unconditional moment conditions:

$$\mathbb{E}[m(V, \alpha_1^\intercal X_1, \ldots, \alpha_r^\intercal X_r, g_1(\theta_1^\intercal Z_1), \ldots, g_s(\theta_s^\intercal Z_s))] = 0, \tag{1.1}$$

where m is a known vector of functions whose dimension q is large, $\alpha_1, \ldots, \alpha_r$ are unknown Euclidean-valued parameters whose respective dimensions p_j may be large, while $g = (g_1, \ldots, g_s)$ is a vector of unknown smooth functions and the index vectors $\theta_1, \ldots, \theta_s$ are unit vectors with positive first elements satisfying usual identification condition for single-index model. Both r and s are fixed. The random variable V typically represents a dependent variable and possible instrumental variables, while the vectors $X_j (1 \le j \le r)$ and $Z_\ell (1 \le \ell \le s)$ are explanatory variables. We suppose that

* Correspondence to: Faculty of Economics, University of Cambridge, UK.
E-mail address: obl20@com.ac.uk (O. Linton).

https://doi.org/10.1016/j.jeconom.2021.07.004
0304-4076/

Z_j is of fixed finite dimension, but the dimension of X and V) may be large, i.e., diverge. We suppose that a random sample $\{V_i, X_{ji}, Z_{\ell i}, 1 \le j \le r, 1 \le \ell \le s, i = 1, \ldots, n\}$ is observed and that $p = p(n) \to \infty$ and $q = q(n) \to \infty$ as $n \to \infty$ with $q > p \equiv p_1 + \cdots + p_r$ where p_j is the dimension of X_j. For our main inference results we consider the case where (at least) $p/n \to 0$, similar to Portnoy (1984, 1985), Mammen (1989). The moment restriction model (1.1) features high dimensionality in two ways: high dimensional Euclidean parameters (α_j) and an unknown function vector $g(\cdot)$ with infinite dimensional parameter elements (g_ℓ); the number of moment conditions necessarily increases to infinity. The setting includes as a special case the partial linear regression model with some weak instruments and endogeneity, Robinson (1988), except in our case the number of covariates in the linear part may be large, i.e., increase to infinity with sample size. There are sometimes many binary covariates whose effect can be restricted to be linear, perhaps after a transformation of response, but other continuous covariates whose effect is thought to be nonlinear. In panel data, one may wish to allow for many fixed effects in an essentially linear fashion, but capture the potential nonlinear effect of a critical covariate or a continuous treatment variable. If both the cross-sectional and time series dimension are large then these quantities are all estimable. See for example Connor et al. (2012).

We propose using the Generalized Method of Moments (GMM) to deliver simultaneous estimation of all unknown quantities from a large dimensional moment vector. There is a considerable literature on GMM in parametric cases following Hansen (1982). There is a general theory available for non-smooth objective functions of finite dimensional parameters (e.g., Pakes and Pollard (1989) and Newey and McFadden (1994, Section 7)). Some recent work has focused on the extension to the case where there are many moment conditions but some conditions are more informative than others, the so-called weak instrument case, see Newey and Windmeijer (2009), Han and Phillips (2006). There is a large literature on semiparametric estimation problems with smooth objective functions of both finite and infinite dimensional parameters (e.g., Bickel et al. (1993), Andrews (1994), Newey (1994), Newey and McFadden (1994, Section 8), Pakes and Olley (1995), Chen and Shen (1998), Ai and Chen (2003)). Chen et al. (2003) extended this theory to allow for non-smooth moment functions. Other work has sharpened and broadened the applicability of the semiparametric case where the number of Euclidean parameters is finite but there are unknown function-valued parameters and endogeneity (see, for example Chen and Liao (2015)). Our work extends the semiparametric theory to the case where the parametric component is growing in complexity, which is of particular relevance for modern big data settings.

We simultaneously estimate α_j, g_ℓ and θ_ℓ in the parameter spaces defined below. The parameters of interest are particular functionals of α_j and g_ℓ for which we have plug-in estimators once we obtain the estimates of α_j and g_ℓ. Chen et al. (2003) study a fixed-dimensional moment restriction model containing an unknown function. They consider both two-step and profiled two-step methods. A similar approach is used in Chen and Liao (2015). Kernel estimation techniques in particular require an additional (albeit related) estimating equation for the function valued part, and either two-step or profile methods are common, see, for example, Powell (1984). We use the sieve methodology (see Chen (2007) for a review) to estimate the model (1.1) in one step. By the method of sieve, unknown function is completely parameterized, which enables us to estimate the parameter vectors α_j, the functions $g_\ell(\cdot)$ and the index vectors θ_ℓ in model (1.1) simultaneously. This approach also avoids high level assumptions, such as in Chen et al. (2003) and Han and Phillips (2006). We establish the consistency and (self-normalized) asymptotic normality of the parameters of interest which are general functionals of α_j and g_ℓ) and provide a feasible CLT that allows normal based inference about the parameters of interest. We also propose some new test statistics to address the over-identification issue, and establish their large sample properties.

Even though the sieve method can parameterize unknown functions, the estimates of θ_ℓ are still challenging. Note that the importance of the involvement of single-index structure in conditional moment restriction model has been mentioned in Ai and Chen (2003, p. 1796), but there is no explicit treatment for the estimation of index vector, as far as we are aware, in the literature on moment or conditional moment restriction models. The reason might be, in our opinion, that the commonly used profile method dealing with single-index structure in regression models is not applicable because the moment function in general is no longer linear in its components. See, for example, Dong et al. (2016) and the reference therein for the use of profile method in single-index regression. We offer a solution for this situation.

It is clear that when all vectors Z_ℓ are reduced to be scalar, the single-index structure in model (1.1) is reduced to nonparametric function $g_\ell(Z_\ell)$, and hence a relatively simpler model is

$$\mathbb{E}[m(V, \alpha_1^{\top} X_1, \ldots, \alpha_r^{\top} X_r, g_1(Z_1), \ldots, g_s(Z_s))] = 0. \tag{1.2}$$

However, this model also has wide applications because the number of unknown functions can be any fixed integer. For some discussions we just consider the case where $r = 1$ and $s = 1$.

In the second part of the paper we consider the ultra-high dimensional case where the number of potential X variables is extremely large, i.e., much larger than the sample size, but only a smaller subset of them are relevant, i.e., the parametric part of the model possesses sparsity. That is, we suppose that $p \gg n$ but α contains many zero elements, although we do not know a priori the location of these zeros. This case has been considered by a number of recent studies in econometrics, such as Belloni et al. (2016b), and is the focus of recent research in statistics. To address this issue we combine the GMM objective function with a specific penalty function, a folded concave penalty function (see Fan and Li (2001)). We show that variable selection and estimation can be done simultaneously and our method achieves the oracle property, like Fan and Liao (2014). We also provide a result on post model selection inference, which allows us to use the distribution theory obtained in the first part of the paper. An alternative framework here is the approximate linear model (ALM)

framework considered in inter alia, Belloni et al. (2016b). In that setting there is no formal distinction between parametric and nonparametric components in the ALM and the methodology is built around the selection tools. Our more traditional semiparametric approach is explicit about the model components and their relative complexity. In particular, we specify that g is nonparametric and has to be estimated simultaneously with the parametric part. We are consequently able to give inference results for a wider range of parameters.

A common genesis for the unconditional moment restrictions (1.1) is conditional moment restrictions perhaps from some economic model (Hansen, 1982). Let W_i be a sub-vector of $(X_i^\intercal, Z_i^\intercal)^\intercal$ and let $\rho(Y_i, \alpha^\intercal X_i, g(Z_i))$ be a known J-dimensional vector residual. Then, suppose that (α, g) is determined by the conditional moment restriction

$$\mathbb{E}[\rho(Y_i, \alpha^\intercal X_i, g(Z_i))|W_i] = 0, \quad \text{almost surely},$$

which then implies $\mathbb{E}[m(V_i, \alpha^\intercal X_i, g(Z_i))] = 0$ for any function m composed of the residual vector ρ multiplied by instruments taken from the space generated by W_i. In fact, if $m(V_i, \alpha^\intercal X_i, g(Z_i)) = \rho(Y_i, \alpha^\intercal X_i, g(Z_i)) \otimes \Phi_K(W_i)$, where $\Phi_K(w) = (h_1(w), \ldots, h_K(w))^\intercal$ is a vector of basis functions in some function space, $V_i = (Y_i, W_i^\intercal)^\intercal$ and "$\otimes$" denotes the Kronecker product, then this can deliver semiparametrically efficient estimation of α in the finite dimensional case. Notice that the dimension of the function m is $q = JK$, which increases with K. Therefore, the pair (α, g) can be solved from the unconditional moment equation $\mathbb{E}[m(V_i, \alpha^\intercal X_i, g(Z_i))] = 0$.

Before we conclude this section we give some important examples. One is the partial linear model with many endogenous covariates. Let $Y_i = \alpha^\intercal X_i + g(Z_i) + e_i$, $i = 1, \ldots, n$, where $\alpha \in \mathbb{R}^p$ and e_i is an error term such that $\mathbb{E}[e_i] = 0$ for all i. Here, X_i is endogenous in the sense that $\mathbb{E}[e_i|X_i] \neq 0$. To deal with the endogeneity, let W_i be a vector of instrumental variables and define a set of valid instruments $\lambda_i = \lambda(Z_i, W_i)$ with dimension q ($q > p$). Denote $m(V_i, \alpha^\intercal X_i, g(Z_i)) = (Y_i - \alpha^\intercal X_i - g(Z_i))\lambda(Z_i, W_i)$ with $V_i = (Y_i, W_i^\intercal)^\intercal$. Then, we have the moment condition $\mathbb{E}[m(Y_i, W_i, \alpha^\intercal X_i, g(Z_i))] = 0$, which can be used to identify the parameter α and the nonparametric function $g(\cdot)$. Motivated by Robinson (1988) and Belloni et al. (2012) an alternative moment condition in this case is

$$m(V_i, \alpha^\intercal X_i, g(Z_i)) = \left(Y_i - g_Y(Z_i) - \alpha^\intercal \left(X_i - g_X(Z_i)\right), Y_i - g_Y(Z_i), \left(X_i - g_X(Z_i)\right)^\intercal\right) \otimes \lambda(Z_i, W_i),$$

where $g_Y(Z_i) = E(Y_i|Z_i)$ and $g_X(Z_i) = E(X_i|Z_i)$. Essentially this is the efficient score function for α in a special case, Bickel et al. (1993). One can jointly estimate α, g_Y, g_X from this moment condition and then obtain $g(Z) = g_Y(Z) - \alpha^\intercal g_X(Z)$.

Another example is the model studied in Carneiro et al. (2011) where the authors consider the following in their equation (9):

$$\begin{aligned} &\mathbb{E}[Y - X^\intercal \delta - P(Z)X^\intercal \alpha - R(Z)|X, Z] = 0, \\ &\mathbb{E}[\mathbb{I}(S = 1) - P(Z)|Z] = 0, \end{aligned} \tag{1.3}$$

where $P(\cdot)$ and $R(\cdot)$ are nonparametric, $\mathbb{I}(\cdot)$ is the indicator function, and S is the selection indicator. The outcome variable is the log wage, and X, Z are observed individual characteristics. Here, because the dimension of Z in general is greater than three, a single-index structure is adopted for the nonparametric function $P(Z)$, i.e., $P(Z) := \Lambda(\theta_0^\intercal Z)$. Furthermore, the function $R(z) = g(P(z))$, where g is unknown. The dimension of X may be large. We consider this model in our application.

The rest of the paper is organized as follows. Section 2 develops an estimation procedure. Section 3 establishes the large sample theory for the proposed estimators. In Section 4, we provide two methods for testing over-identification. Section 5 proposes and studies selection procedures for choosing covariates/parameters under sparsity. In Section 6, we evaluate the finite sample performance of the proposed estimation procedures using simulations. In Section 7, we apply our method to investigate the effect of schooling on earnings using the model and data of Carneiro et al. (2011). The last section concludes.

Throughout, $\|\cdot\|$ can be either Euclidean norm for vector or Frobenius norm for matrix, or the norm of functions in function space that would not arise any ambiguity in the context; $\otimes$ denotes Kronecker product for matrices or vectors; $:=$ means equal by definition; I_r is the identity matrix of dimension r.

2. Estimation procedure

2.1. Moment restriction without single-index structure

We start with model (1.2). Since sieve method is used to tackle the nonparametric functions, as can be seen in the sequel sections the general case of model (1.2) with $r \geq 1$ and $s \geq 1$ is theoretically equivalent to its special case where $r = s = 1$. The only price we pay for $r > 1$ and/or $s > 1$ is the complexity of notation. This is the same for model (1.1), that is, the theory on the special case with $r = s = 1$ of model (1.1) can be extended to the general case straightforwardly. Our asymptotic theory and inference then mainly focus on the special cases for both (1.1) and (1.2) where $r = s = 1$, but for completion both the estimation procedure and the associated theory for the general model setting are given in Appendix E of the supplementary file of this paper.

Consider

$$\mathbb{E}[m(V, \alpha^\intercal X, g(Z))] = 0. \tag{2.1}$$

Here, we suppose that $g \in L^2(\mathbb{Z}, \pi) = \{f : \int_{\mathbb{Z}} f^2(z)\pi(z)dz < \infty\}$ a Hilbert function space, $\mathbb{Z} \subset \mathbb{R}^d$, where $\pi(\cdot)$ is a user-chosen density function on $\mathbb{Z}$. The choice of the density π relates to how large the Hilbert space is expected, since the thinner the tail of the density is, the larger the space is. For example, $L^2(\mathbb{R}, 1/(1+z^2)) \subset L^2(\mathbb{R}, \exp(-z^2))$. An inner product in the Hilbert space is given by $\langle f_1, f_2\rangle = \int_{\mathbb{Z}} f_1(z) f_2(z)\pi(z)dz$, and hence the induced norm $\|f\| = \sqrt{\langle f, f\rangle}$ for any $f_1(z), f_2(z), f(z) \in L^2(\mathbb{Z}, \pi)$. Two functions $f_1, f_2 \in L^2(\mathbb{Z}, \pi)$ are called orthogonal if $\langle f_1, f_2\rangle = 0$, and further are orthonormal if $\|f_1\| = 1$ and $\|f_2\| = 1$.

Assumption 2.1. Suppose that $\{\varphi_j(\cdot)\}$ is a complete orthonormal function sequence in $L^2(\mathbb{Z}, \pi)$, that is, $\langle \varphi_i(\cdot), \varphi_j(\cdot)\rangle = \delta_{ij}$ the Kronecker delta.

Recall that any Hilbert space has a complete orthogonal sequence (see Theorem 5.4.7 in Dudley (2003, p. 169)). For the multivariate function setting, the orthonormal sequence $\{\varphi_j(\cdot)\}$ can be constructed from the tensor product of univariate orthogonal sequences. See, e.g. Chapter one of Gautschi (2004), Chen (2007) for more discussion.

For the function $g(z) \in L^2(\mathbb{Z}, \pi)$, we may have an infinite orthogonal series expansion

$$g(z) = \sum_{j=0}^{\infty} \beta_j \varphi_j(z), \quad \text{where } \beta_j = \langle g, \varphi_j\rangle, \tag{2.2}$$

where the convergence is in the norm sense in the space. Moreover, if g is smooth, establishing pointwise convergence is possible. See Dong et al. (2016). For positive integer K, define $g_K(z) = \sum_{j=0}^{K-1} \beta_j \varphi_j(z)$ as a truncated series and $\gamma_K(z) = \sum_{j=K}^{\infty} \beta_j \varphi_j(z)$ the residue. Then, $g_K(z) \to g(z)$ as $K \to \infty$. For better exposition, denote $\Phi_K(z) = (\varphi_0(z), \ldots, \varphi_{K-1}(z))^\tau$ and $\beta = (\beta_0, \ldots, \beta_{K-1})^\tau$ two K-vectors. Thus, $g_K(z) = \beta^\tau \Phi_K(z)$.

Our primary goal is to estimate the unknown parameters (α, g) and functionals thereof. Define $\Theta = \{(\mathbf{a}, f) : \mathbf{a} \in \mathbb{R}^p, f \in L^2(\mathbb{Z}, \pi)\}$, the parameter space for model (2.1) and

$$\|(\mathbf{a}, f)\| = \|\mathbf{a}\|_E + \|f\|_{L^2}, \tag{2.3}$$

where $\|\cdot\|_E$ denotes the Euclidean norm on $\mathbb{R}^p$ and $\|f\|_{L^2}$ signifies the norm on the Hilbert space, of which the subscript may be suppressed whenever no ambiguity is incurred. The consistency studied below is defined in terms of this topology.

In order to facilitate the implementation of nonlinear optimization, α should be confined to a compact subset of $\mathbb{R}^p$ and the truncated series $g_K(z)$ should be included in an expanding finite dimensional bounded subset of $L^2(\mathbb{Z}, \pi)$. It is noteworthy that in an infinite dimensional space, a bounded set may not necessarily be compact. See Chen and Pouzo (2012) for detailed discussion on the compactness. The following assumption ensures that our optimization is implemented over a compact set, so that our estimation does not suffer from the *ill-posededness issue* that is encountered in the literature, such as Chen (2007), Blundell et al. (2007).

Assumption 2.2. Suppose that B_{1n} and B_{2n} are positive real numbers diverging with n such that α in model (1.1) is included in $\Theta_{1n} := \{\mathbf{a} \in \mathbb{R}^p : \|\mathbf{a}\| \le B_{1n}\}$ and for sufficient large n, $g_K(z)$ is included in $\Theta_{2n} := \{\mathbf{b}^\tau \Phi_K(z) : \|\mathbf{b}\| \le B_{2n}\}$. Define $\Theta_n = \Theta_{1n} \otimes \Theta_{2n}$.

Here, unlike in a standard single-index model, we do not require $\|\alpha\| = 1$ for identification. This is because the function $m(\cdot)$ is known and hence we are able to identify any scaling for α. Assumption 2.2 allows for the bounds of α to diverge with the sample size that entertains the divergence of its dimensionality. Furthermore, since $\|g_K\| \le \|g\|$ it is clear that there exists an integer n_0 such that $g_K(z) \in \Theta_{2n}$ for all $n \ge n_0$. On the other hand, Θ_{2n}, the so-called linear sieve space in the literature, can approximate the entire function space with the increase of the sample size, because any $f(z) \in L^2(\mathbb{Z}, \pi)$ can be approximated by a combination of this type, $\mathbf{b}^\tau \Phi_K(z)$, arbitrarily in the sense of norm. Thus, Θ can be approximated by Θ_n as $n \to \infty$. More importantly, our setting is similar to but broader than that discussed in Newey and Powell (2003).

We estimate α and β by

$$\begin{aligned}(\widehat{\alpha}, \widehat{\beta}) &= \underset{\mathbf{a}\in\mathbb{R}^p, \mathbf{b}\in\mathbb{R}^K}{\operatorname{argmin}} \|M_n(\mathbf{a}, \mathbf{b})\|^2, \quad \text{subject to } \|\mathbf{a}\| \le B_{1n} \text{ and } \|\mathbf{b}\| \le B_{2n}, \\ &\text{where } M_n(\mathbf{a}, \mathbf{b}) = \frac{1}{\sqrt{q}} \frac{1}{n} \sum_{i=1}^{n} m(V_i, \mathbf{a}^\tau X_i, \mathbf{b}^\tau \Phi_K(Z_i)).\end{aligned} \tag{2.4}$$

Here, the involvement of q in $M_n(\mathbf{a}, \mathbf{b})$ takes into account the divergent dimensions of the vector m in order to avoid the issue that $\|M_n(\mathbf{a}, \mathbf{b})\|$ could be large even if each element is small that would arise if we had not put q into $M_n(\mathbf{a}, \mathbf{b})$. This issue does not arise when the vector-valued m function has fixed dimension. Define for any $z \in \mathbb{Z}$,

$$\widehat{g}(z) = \widehat{\beta}^\tau \Phi_K(z), \tag{2.5}$$

which is our estimator of $g(z)$. In the next section we establish the asymptotic consistency of this estimator in the sense that $\|(\widehat{\alpha} - \alpha, \widehat{g} - g)\| \to_p 0$ as $n \to \infty$, where the norm is defined in (2.3).

2.2. Moment restriction with single-index structure

Two approaches are introduced in this subsection to deal with the moment restriction models with a single-index setting, one is direct and the other is indirect, but they are used for different situations.

Similar to the preceding section, we only focus on the estimation procedure for model (1.1) in the case where $r = s = 1$,

$$\mathbb{E}[m(V, \alpha^{\intercal}X, g(\theta_0^{\intercal}Z))] = 0. \tag{2.6}$$

Our goal is to estimate α, θ_0 and g where θ_0 satisfies identification condition: $\|\theta_0\| = 1$ and the first element is positive. We suppose that the function g has support $\mathbb{R}$ and $g \in L^2(\mathbb{R}, \exp(-w^2))$. Note that the Hilbert space $L^2(\mathbb{R}, \exp(-w^2))$ is sufficiently large that includes all polynomials, powers and bounded functions; the normalized Hermite polynomials $h_j(w) = (\sqrt{\pi}2^j j!)^{-1/2}H_j(w)$ form an orthonormal basis in $L^2(\mathbb{R}, \exp(-w^2))$, where $H_j(w)$ is the jth Hermite polynomial, and $\int h_i(w)h_j(w)\exp(-w^2)dw = \delta_{ij}$; meanwhile, as a particular case of (2.2), $g(w)$ also admits an orthogonal expansion, $g(w) = \sum_{j=0}^{\infty} c_j h_j(w)$. Hence, by virtue of a property of Hermite polynomials given in Lemma A.4 and $\|\theta_0\| = 1$ with $\theta_0 = (\theta_{01}, \ldots, \theta_{0d})^{\intercal}$, for any $k \geq 1$, we have

$$g(\theta_0^{\intercal}Z) = \sum_{j=0}^{k-1} c_j h_j(\theta_0^{\intercal}Z) + \gamma_k(\theta_0^{\intercal}Z) = \sum_{j=0}^{k-1}\sum_{|u|=j} a_{ju}(\theta_0)\mathcal{H}_u(Z) + \gamma_k(\theta_0^{\intercal}Z),$$

where $\gamma_k(\cdot)$ is the truncation residue, u is a multi-index, $u = (u_1, \ldots, u_d)$, $|u| = u_1 + \cdots + u_d$ and

$$a_{ju}(\theta_0) = \frac{\pi^{d/4}2^{j/2}j!}{\sqrt{u_1! \cdots u_d!}} c_j \theta_0^u, \quad \theta_0^u = \prod_{j=1}^{d} \theta_{0j}^{u_j}, \quad \mathcal{H}_u(Z) = \prod_{j=1}^{d} h_{u_j}(Z_j).$$

This means that $g(\theta_0^{\intercal}Z)$ can be approximated by a combination of $\beta^{\intercal}\Phi_K(Z)$ where $\Phi_K(Z)$ is a vector consisting of all $\mathcal{H}_u(Z)$ for $|u| = j$ and $0 \leq j \leq k-1$, and similarly β consists of all $a_{ju}(\theta_0)$ in the same ordering as $\Phi_K(Z)$, that is, $g(\theta^{\intercal}Z) = \beta^{\intercal}\Phi_K(Z) + \gamma_k(\theta_0^{\intercal}Z)$ and $\gamma_k(\cdot) = o(1)$ as $k \to \infty$ in a sense.

Therefore, similar to (2.4), we may estimate α and β by

$$\begin{aligned} &(\widehat{\alpha}, \widehat{\beta}) = \underset{\mathbf{a}\in\mathbb{R}^p, \mathbf{b}\in\mathbb{R}^K}{\operatorname{argmin}} \|M_n(\mathbf{a}, \mathbf{b})\|^2, \quad \text{subject to } \|\mathbf{a}\| \leq B_{1n} \text{ and } \|\mathbf{b}\| \leq B_{2n}, \\ &\text{where } M_n(\mathbf{a}, \mathbf{b}) = \frac{1}{\sqrt{q}}\frac{1}{n}\sum_{i=1}^{n} m(V_i, \mathbf{a}^{\intercal}X_i, \mathbf{b}^{\intercal}\Phi_K(Z_i))\phi(\|Z_i\|) \text{ and } \phi(v) = \exp(-v^2/2). \end{aligned} \tag{2.7}$$

The involvement of $\phi(\cdot)$ in $M_n(\mathbf{a}, \mathbf{b})$ is to deal with the rapid divergence of the fourth moment of $h_j(\cdot)$. This technique is also used in Dong et al. (2021). Notice that from $\widehat{\beta}$, along with the identification condition on θ_0 we can derive the estimators of $g(\cdot)$ and θ_0, as shown in the next section.

A straightforward algebra yields that the length of $\Phi_K(Z)$ is about $O(k^d)$, that limits the dimension of Z in a narrow range. If d is relatively large, this method would fail to work since k^d grows extremely fast. Thus, the direct method is applicable to the case where d is small. This is the reason why we are going to introduce an indirect method as another approach.

Indeed, as far as we are aware, in some economic theory though a single-index structure is involved, one may be able to estimate the unknown index vector by another equation. With the estimate of θ_0 at hand, model (2.6) is reduced to model (2.1), so that α and g can be estimated by the procedure (2.4). We shall give a detailed description on the estimation of θ_0 in an economic context in the next section.

3. Asymptotic theory

3.1. Consistency

Before establishing our asymptotic theory for model (2.1), we state with some assumptions that we rely on in the sequel.

Assumption 3.1. Suppose that

(a) For each n, $\{(V_i, X_i^{\intercal}, Z_i^{\intercal}), i = 1, \ldots, n\}$ is an independent and identically distributed (i.i.d.) sequence (although the distribution depends on n, which we suppress notationally in the sequel) from (2.1);
(b) For the density f_Z of Z, there exist two constants, $0 < c < C < \infty$, such that $c\pi(z) \leq f_Z(z) \leq C\pi(z)$ on the support $\mathbb{Z}$ of Z, where $\pi(z)$ is given in the preceding section;
(c) Each moment function $m_j(\cdot, \cdot, \cdot)$, $j = 1, \ldots, q$, is continuous in the second and third arguments;
(d) $q(n) - p(n) \geq K$.

The i.i.d. property in Assumption 3.1(a) simplifies the presentation and some of the calculations, although it is possible to relax it to a weakly dependent data setting. Regarding Assumption 3.1(b), the relation between the density of the variable Z and the function space is widely used in the literature. See, e.g. Condition A.2 and Proposition 2.1 of Belloni et al. (2015, p. 347). This condition is used to bound the eigenvalues of the Gram matrix for the sieve method. When the support is compact, the existing literature simply imposes that the density $f_Z(z)$ is bounded away from zero and above from infinity that is a special case where $\pi(z) \equiv 1$ in our setting. Our theory allows for unbounded support for Z provided the density π is chosen appropriately. In the unbounded support case this assumption amounts to having an upper and lower bound on the tails of the covariate density. There is a large literature concerned with estimation of tail thickness parameters in statistics and financial econometrics, see for example Embrechts et al. (1999), which could be adapted to provide guidance on suitable choices of π. Regarding Assumption 3.1(c), the continuity of the m function is weak, and commonly used moment functions satisfy this, including those of the quantile-type. In Assumption 3.1(d) we allow for possible overidentification of the parameter vector in the moment conditions, and we shall discuss this issue further in the next section.

Assumption 3.2. Suppose that there is a unique function $g(\cdot) \in L^2(\mathbb{Z}, \pi)$ and for each n there is a unique vector $\alpha \in \mathbb{R}^p$ such that for any $\delta > 0$, there is a sufficiently small constant $\epsilon_n \equiv \epsilon_n(\delta) > 0$ such that

$$\inf_{\substack{(\mathbf{a},f)\in\Theta \\ \|(\mathbf{a}-\alpha, f-g)\|\geq\delta}} q^{-1}\|\mathbb{E}m(V_i, \mathbf{a}^\intercal X_i, f(Z_i))\|^2 > \epsilon_n,$$

and possibly $\epsilon_n \to 0$ as $n \to \infty$ but with a rate slower than $\max(\|\gamma_K\|, n^{-1})$.

The squared norm is scaled down by its dimension due to the same reason as in the formulation of M_n in the last section. This type of conditions and global identifications, is commonly made in the conventional literature with $\epsilon_n = \epsilon > 0$ independent of n, such as Pakes and Pollard (1989, p. 1308) and Chen et al. (2003, p. 1593). It guarantees the uniqueness of the true parameter in the parameter space satisfying the moment condition and hence ensures the consistency. However, our assumption is much weaker than the conventional one by allowing $\epsilon_n \to 0$ with some rate. Such ϵ_n enables us to identify the parameter because in a neighbourhood of the true parameter the criterion function attenuates to zero at rate $\max(\|\gamma_K\|, n^{-1})$ shown in Lemma A.1.

Assumption 3.3. Suppose that for each n , there is a measurable positive function $A(V, X, Z)$ such that

$$q^{-1/2}\|m(V, \mathbf{a}_1^\intercal X, f_1(Z)) - m(V, \mathbf{a}_2^\intercal X, f_2(Z))\| \leq A(V, X, Z)[\|\mathbf{a}_1 - \mathbf{a}_2\| + |f_1(Z) - f_2(Z)|]$$

for any $(a_1, f_1), (a_2, f_2) \in \Theta_n$, where (V, X, Z) is a realization of (V_i, X_i, Z_i) and $A(\cdot, \cdot, \cdot)$ satisfies that $\mathbb{E}[A^2(V_i, X_i, Z_i)] < \infty$.

This is a kind of Lipschitz condition. We note that this condition can be substituted by some high level condition such as stochastic equicontinuity, in order to study the large sample behaviour of the estimators. See, for instance, Pakes and Pollard (1989), Chen et al. (2003). As argued in Chen et al. (2003, p. 1597), when the moment function is Lipschitz continuous, the *covering number with bracketing* is bounded above by the *covering number* for the parametric space, so stochastic equicontinuity condition holds. Among others, Chen and Shen (1998) used this approach. We would like to keep this low level condition because additionally it facilitates calculation in some situations.

The positive function $A(V, X, Z)$ may be viewed as the upper bound of the norm of the partial derivatives of $q^{-1/2}m(V, \mathbf{a}^\intercal X, w)$ with respect to the vector $\mathbf{a}$ and the scalar w, respectively, and thus the condition is fulfilled if the second moment of $A(V, X, Z)$ is bounded. The assumption guarantees the approximation of $m(V_i, \alpha^\intercal X_i, \beta^\intercal \Phi_K(Z_i))$ to $m(V_i, \alpha^\intercal X_i, g(Z_i))$, because

$$\begin{aligned}&\|m(V_i, \alpha^\intercal X_i, \beta^\intercal \Phi_K(Z_i)) - m(V_i, \alpha^\intercal X_i, g(Z_i))\| \\ \leq& A(V_i, X_i, Z_i)\|g(Z_i) - \beta^\intercal \Phi_K(Z_i)\| = O_P(1)\|\gamma_K\| = o_P(1)\end{aligned}$$

by virtue of Assumption 3.1(b). Also, it ensures that $\|\mathbb{E}m(V_i, \alpha^\intercal X_i, \beta^\intercal \Phi_K(Z_i))\| = o(1)$, since $\mathbb{E}m(V_i, \alpha^\intercal X_i, g(Z_i)) = 0$. More importantly,

$$\begin{aligned}&q^{-1}\mathbb{E}\|m(V_i, \mathbf{a}^\intercal X_i, f(Z_i))\|^2 \\ \leq& 2q^{-1}\mathbb{E}\|m(V_i, 0, 0)\|^2 + 2\mathbb{E}[A(V_i, X_i, Z_i)^2][\|\mathbf{a}\|^2 + \mathbb{E}f(Z_i)^2] = O(B_{1n}^2 + B_{2n}^2)\end{aligned}$$

uniformly on $(\mathbf{a}, f) \in \Theta_n$.

Theorem 3.1 (*Consistency*). *Suppose that Assumptions 2.1–2.2 and 3.1–3.3 hold, and that $B_{1n}^2 + B_{2n}^2 = o(n)$. Then, we have $\|(\widehat{\alpha} - \alpha, \widehat{g} - g)\| \to_P 0$ as $n \to \infty$ for $(\widehat{\alpha}, \widehat{g})$ given by (2.4).*

The proof is given in Appendix B.

3.2. Limit distributions of the estimators

Since the dimension of α diverges in model (2.1), we cannot establish a limit distribution for $\widehat{\alpha}-\alpha$ itself. Instead, we shall consider some finite dimensional transformations of α, for which plug-in estimators are used. Likewise, we consider functionals of $g(\cdot)$. In many applications both types of quantities are of interest. For example, the weighted average marginal treatment effect (MTE) parameter in Carneiro et al. (2011) depends on both α and g. In financial econometrics a leading example is the conditional value at risk parameter, which depends on the parameters of the dynamic mean and variance model and on the quantile of the error distribution.

Let $\mathscr{L}$ be a transformation from $\mathbb{R}^p \mapsto \mathbb{R}^\mu$ with $\mu \geq 1$ fixed, and let $\mathscr{F}=(\mathscr{F}_1,\ldots,\mathscr{F}_\nu)^\intercal$ with fixed ν be a vector of functionals on $L^2(\mathbb{Z},\pi)$. Though in the literature one usually takes linear $\mathscr{L}$ with $\mu=1$ (see, e.g. Theorem 4.2 in Belloni et al. (2015, p. 352) and several results such as Theorems 2 and 6 in Chang et al. (2015)), we allow that $\mathscr{L}$ may have either linear or nonlinear components or both with $\mu \geq 1$. The elements of $\mathscr{F}$ can be, for example, as described in Newey (1997, p. 151), the integral of $\ln[g(z)]$ on some interval, which stands for consumer's surplus in microeconomics. Other examples include: the partial derivative function, the average partial derivative, and the conditional partial derivative. Thus, we shall consider the limit distributions of $\mathscr{L}(\widehat{\alpha})-\mathscr{L}(\alpha)$ and $\mathscr{F}(\widehat{g})-\mathscr{F}(g)$. Towards this end, we need the following assumptions.

Assumption 3.4. (a). Suppose that each element function m_j of the m function is differentiable with respect to its second and third arguments up to the second order; the second derivative functions satisfy a Lipschitz condition in a neighbourhood of the (α, g):

$$|\partial^{(u)}m_j(V,\alpha^\intercal X,g(Z))-\partial^{(u)}m_j(V,\mathbf{a}^\intercal X,f(Z))| \leq B_j(V,\alpha^\intercal X,g(Z))(\|\mathbf{a}-\alpha\|+\|g-f\|)^\tau$$

for some $\tau \in (0,1]$, where u is two-dimensional multiple index with $|u|=2$, $\partial^{(u)}$ stands for the partial derivative of the function with respect to the second and third arguments and B_j are positive functions such that $\max_{1\leq j\leq q}\mathbb{E}[B_j(V,\alpha^\intercal X, g(Z))^2]<\infty$.

(b). Let the g function be smooth with the smoothness order required being spelt out later.

The Lipschitz condition for the components of the m function enables us to approximate the Hessian matrix within a neighbourhood of the true parameter, which in turn facilitates the derivation of the limit theory. It is well known that a certain smoothness order of the g function is required to get rid of the truncation residues. Such a requirement is implicitly spelt out in Assumption 3.6.

Assumption 3.5. Suppose that

(a) $\mathbb{E}\left\|m(V,\alpha^\intercal X,g(Z))\right\|^2=O(q)$, $\mathbb{E}\|X\|^2=O(p)$ and $\mathbb{E}\|\Phi_K(Z)\|^2=O(K)$;

(b) $\mathbb{E}\left\|\frac{\partial}{\partial u}m(V,\alpha^\intercal X,g(Z))\right\|^2=O(q)$, and $\mathbb{E}\left\|\frac{\partial}{\partial w}m(V,\alpha^\intercal X,g(Z))\right\|^2=O(q)$;

(c) $\mathbb{E}\left\|\frac{\partial}{\partial u}m(V,\alpha^\intercal X,g(Z))\otimes X\right\|^2=O(pq)$, and
$\mathbb{E}\left\|\frac{\partial}{\partial w}m(V,\alpha^\intercal X,g(Z))\otimes \Phi_K(Z)\right\|^2=O(Kq)$;

(d) $\mathbb{E}\left\|\frac{\partial^2}{\partial u^2}m(V,\alpha^\intercal X,g(Z))\otimes XX^\intercal\right\|^2=O(p^2q)$, and
$\mathbb{E}\left\|\frac{\partial^2}{\partial w^2}m(V,\alpha^\intercal X,g(Z))\otimes \Phi_K(Z)\Phi_K(Z)^\intercal\right\|^2=O(K^2q)$.

We have the following comments. It is not necessary that all elements of the m vector have uniformly bounded second moments to satisfy the first supposition in 3.5(a). Because the dimension p of X diverges with n, in 3.5(a) we allow that the second moment $\mathbb{E}\|X\|^2$ diverges too; moreover, $\mathbb{E}\|\Phi_K(Z)\|^2=O(K)$ can be true for many orthogonal sequences given the relation between the densities of Z and the L^2 space in Assumption 3.1. In 3.5(b) we impose a similar condition for the norm of the function's first partial derivatives, while in 3.5(c) and (d) we stipulate moment conditions for the norms of the tensor product for regressor and the partial derivatives (the first and second, respectively) of the m function. These hold similarly as (a) and (b) but with larger dimensions, particularly when the m function is linear in its arguments.

Assumption 3.6. Suppose that

(a) $\|\gamma_K\|p=o(1)$, $\quad n^{-1}p^2=o(1)$;
(b) $\|\gamma_K\|K=o(1)$, $\quad n^{-1}K^2=o(1)$.

Assumption 3.6 stipulates the relation between the truncation parameter K, the diverging dimension p of the regressor, and the sample size. Normally, $\|\gamma_K\|^2=O(K^{-a})$, where $a>0$ is related to the smoothness order of the function g. See, for example, Newey (1997). Thus, the assumption implicitly puts some conditions on the smoothness. Notice that the combination of 3.6(a) and (b) implies that $\|\gamma_K\|^2pK=o(1)$ and $n^{-1}pK=o(1)$, which are used in the proof of the lemmas in the supplemental material.

Assumption 3.7. The partial derivatives of $m(v,u,w)$ satisfy

(a) $q^{-1/2}\left\|\frac{\partial}{\partial u}m(V,\mathbf{a}_1^\intercal X,f_1(Z))-\frac{\partial}{\partial u}m(V,\mathbf{a}_2^\intercal X,f_2(Z))\right\|\le A_1(V,X,Z)[\|\mathbf{a}_1-\mathbf{a}_2\|+|f_1(Z)-f_2(Z)|]$, where $\mathbb{E}[A_1(V,X,Z)^2]<\infty$ and $\mathbb{E}[A_1(V,X,Z)^2\|X\|^2]=O(p)$.

(b) $q^{-1/2}\left\|\frac{\partial}{\partial w}m(V,\mathbf{a}_1^\intercal X,f_1(Z))-\frac{\partial}{\partial w}m(V,\mathbf{a}_2^\intercal X,f_2(Z))\right\|\le A_2(V,X,Z)[\|\mathbf{a}_1-\mathbf{a}_2\|+|f_1(Z)-f_2(Z)|]$, where $\mathbb{E}[A_2(V,X,Z)^2]<\infty$ and $\mathbb{E}[A_2(V,X,Z)^2\|\Phi_K(Z)\|^2]=O(K)$.

The assumption is similar to Assumption 3.3 but is stipulated for the partial derivatives with extra requirements that $\mathbb{E}[A_1(V,X,Z)^2\|X\|^2]=O(p)$ and $\mathbb{E}[A_2(V,X,Z)^2\|\Phi_K(Z)\|^2]=O(K)$. This is due to the divergence of the dimensions and the argument in Assumption 3.5

Recall the Fréchet derivative operator for an operator from one Banach space to another. It is a bounded linear operator. The Fréchet derivative of $\mathscr{F}$ at $g(\cdot)$ is an s-vector of functionals, denoted by $\mathscr{F}'(g)$, such that

$$\mathscr{F}(\widehat{g})-\mathscr{F}(g)=\mathscr{F}'(g)\circ(\widehat{g}-g)+\lambda(g,\widehat{g}-g),$$

where $\lambda(g,\widehat{g}-g)=o(\|\widehat{g}-g\|)$.

Assumption 3.8. (a) The transformation $\mathscr{L}$ possesses continuous second partial derivatives and the Hessian matrix of each component $\mathscr{L}_j$ of $\mathscr{L}$ has uniformly bounded eigenvalues in a neighbourhood of α, i.e. $\lambda_{\max}(\partial^2\mathscr{L}_j)<\infty$ over n; moreover, the first partial derivative of $\mathscr{L}$ at α, $\partial\mathscr{L}(\alpha)$, has full rank. (b) The functional $\mathscr{F}$ possesses Fréchet derivative at $g(\cdot)$.

The above conditions are quite natural and standard. Given the conditions in Assumption 3.8(a), $\mathscr{L}(\widehat{\alpha})$ can be approximated well by the linear form $\mathscr{L}(\alpha)+\partial\mathscr{L}(\alpha)^\intercal(\widehat{\alpha}-\alpha)$. The condition is fulfilled in particular when $\mathscr{L}$ is a linear or quadratic transform. Indeed, if $\mathscr{L}(\mathbf{a})=A\mathbf{a}$, then $\partial\mathscr{L}(\mathbf{a})\equiv A$ a constant matrix and $\partial^2\mathscr{L}(\mathbf{a})\equiv 0$ for any vector $\mathbf{a}$; especially if $r=1$, a linear transform $\mathscr{L}$ will map a vector into a scalar, $\mathscr{L}(\mathbf{a})=a_0^\intercal\mathbf{a}$, with some $a_0\in\mathbb{R}^p$ and $a_0\neq 0$. This is the case commonly encountered in the literature. See, for example Chang et al. (2015), Belloni et al. (2015). When $\mathscr{L}$ is quadratic, such as $\mathscr{L}(\mathbf{a})=\|\mathbf{a}\|^2$, we then have $\partial\mathscr{L}(\mathbf{a})=2\mathbf{a}$ and $\partial^2\mathscr{L}(\mathbf{a})\equiv 2I_p$ for any $\mathbf{a}$.

We are now ready to establish an asymptotic normality result. Define

$$\begin{aligned}\Sigma_n^2 &:= \Gamma_n[\Psi_n\Psi_n^\intercal]^{-1}\Psi_n\Xi_n\Psi_n^\intercal[\Psi_n\Psi_n^\intercal]^{-1}\Gamma_n^\intercal, \quad \text{in which} \\ \Gamma_n &:= \begin{pmatrix}\partial\mathscr{L}(\alpha)^\intercal & 0\\ 0 & \mathscr{F}'(g)\circ\Phi_K{}^\intercal\end{pmatrix}_{(\mu+\nu)\times(p+K)}, \\ \Xi_n &:= \mathbb{E}[m(V,\alpha^\intercal X,g(Z))m(V,\alpha^\intercal X,g(Z))^\intercal]_{q\times q}, \\ \Psi_n &:= \mathbb{E}\begin{pmatrix}\frac{\partial}{\partial u}m(V,\alpha^\intercal X,g(Z))^\intercal\otimes X\\ \frac{\partial}{\partial w}m(V,\alpha^\intercal X,g(Z))^\intercal\otimes\Phi_K(Z)\end{pmatrix}_{(p+K)\times q},\end{aligned} \tag{3.1}$$

provided that $\Psi_n\Psi_n^\intercal$ is invertible; here u and w stand for the second and the third arguments of the vector function $m(v,u,w)$, respectively.

Theorem 3.2 (*Normality*). *Let Assumptions 2.1–2.2, 3.1–3.8 hold. Suppose also that $B_{1n}^2+B_{2n}^2=o(n)$. Then for $(\widehat{\alpha},\widehat{g})$ given by (2.4), as $n\to\infty$*

$$\sqrt{n}\Sigma_n^{-1}\begin{pmatrix}\mathscr{L}(\widehat{\alpha})-\mathscr{L}(\alpha)\\ \mathscr{F}(\widehat{g})-\mathscr{F}(g)\end{pmatrix}\xrightarrow{d}N(0,I_{\mu+\nu}), \tag{3.2}$$

provided that $\sqrt{n}\Sigma_n^{-1}(0_\mu^\intercal,(\mathscr{F}'(g)\circ\gamma_K)^\intercal)^\intercal=o(1)$, where Σ_n is given by the square root of Σ_n^2 defined in (3.1).

The proof of the theorem is given in Appendix B. Note that the conditions in the theorem imply the consistency of the estimator in Theorem 3.1. Apart from the diverging dimensions of Ψ_n and Ξ_n and the use of the transformation $\mathscr{L}$ and the functional $\mathscr{F}$, the form of the covariance matrices Σ_n^2 is the same as in the standard semiparametric literature, such as Hansen (1982), Pakes and Pollard (1989), Chen et al. (2003).

In general the convergence order of $\mathscr{F}(\widehat{g})-\mathscr{F}(g)$ is proportional to $(\mathscr{F}'(g)\circ\Phi_K(z)^\intercal\mathscr{F}'^\intercal\circ\Phi_K(z))^{1/2}n^{-1/2}$, which is similar to the result in Theorem 2 of Newey (1997). Here, the matrix in the front of $n^{-1/2}$ is of dimension $\nu\times\nu$ and is associated with the derivative of the functional $\mathscr{F}$. To understand how it affects the rate, consider a special case that $\nu=1$ and $\mathscr{F}(g)=g(z)$ for some particular z, implying $\mathscr{F}(\widehat{g})-\mathscr{F}(g)=\widehat{g}(z)-g(z)$ and $\mathscr{F}'(g)\equiv 1$. Then, the matrix is a scalar and the rate becomes $\|\Phi_K(z)\|n^{-1/2}$, which coincides with the conventional nonparametric rate of convergence established in the literature. See, for example, Dong and Linton (2018).

In general, the convergence order of $\mathscr{L}(\widehat{\alpha})-\mathscr{L}(\alpha)$ is $n^{-1/2}$; however, Theorem 3.2 does not rule out the mildly *weak instrument case* where the matrix Σ_n is close to singular, i.e., $|\Sigma_n|\neq 0$ but $|\Sigma_n|\to 0$ with n at a certain rate; this would reduce the convergence rate of the estimators but the self-normalized distribution theory we have presented continues to hold under our conditions. However, we do rule out the more extreme cases considered in Han and Phillips (2006), which would change the limiting distribution.

The requirement that $\sqrt{n}\Sigma_n^{-1}(0_\mu^\intercal,(\mathscr{F}'(g)\circ\gamma_K)^\intercal)^\intercal=o(1)$ is an "undersmoothing" condition, playing a similar role to, for example, the condition $\sqrt{n}V_K^{-1}K^{-p/d}=o(1)$ in Corollary 3.1 of Chen and Christensen (2015, p. 454) and Comment

4.3 of Belloni et al. (2015). The precise form of the condition may vary according to the parameters of interest and the underlying model; it reflects the bias variance trade-off that is relevant for estimation of those quantities in the particular model. In the large dimensional α case, the bias variance trade-off can be different from usual since the parametric part can contribute a large variance; the presence of weak instruments may also affect the bias variance trade-off for certain parameters. For inference results about $g(z)$ it is a common practice to undersmooth/overfit to avoid the bias term. Some recent research advocates using extreme undersmoothing for better inference about finite dimensional parameters in semiparametric models. See, for example Cattaneo et al. (2018). Cattaneo et al. (2018) recently develop heteroskedasticity robust inference methods for the finite dimensional parameters of a linear model in the presence of a large number of linearly estimated nuisance parameters in the case where essentially p is fixed but $K(n) \propto n$. In this case, the function $g(\cdot)$ is not consistently estimated. In our methodology we pay equal attention to the function g, which itself can be of interest. Our methodology is also robust to conditional heteroskedasticity.

Example 3.1. Suppose that $Y = \alpha^{\intercal} X + g(Z) + \varepsilon$, where $\mathbb{E}[\varepsilon|X, Z] = 0$ and the dimension of X is p. By the Robinson (1988) transformation, $\tilde{Y} = \alpha^{\intercal}\tilde{X} + \varepsilon$, where $\tilde{Y} = Y - \mathbb{E}[Y|Z]$ and $\tilde{X} = X - \mathbb{E}[X|Z]$. The Robinson estimator of α is asymptotically normal with asymptotic variance equal to (under homoskedasticity) $\sigma^2[\mathbb{E}(\tilde{X}\tilde{X}^{\intercal})]^{-1}$ when p is fixed, where $\sigma^2 = \mathrm{var}(\varepsilon|X, Z)$. If ε is i.i.d. Gaussian, this is the semiparametric efficiency bound. See Ai and Chen (2003), Chen et al. (2003), Chen (2007), Chen and Pouzo (2009) for discussion of this model under a range of different assumptions.

Our approach considers an approximated version of the model, that is, $Y = \alpha^{\intercal} X + \beta^{\intercal}\Phi_K(Z) + e$, where $\Phi_K(Z)$ is a K-vector of orthonormal basis functions on Z, $e = \delta_K(Z) + \varepsilon$ and $\delta_K(Z) = g(Z) - \beta^{\intercal}\Phi_K(Z)$ that in some sense is negligible under our conditions. Write $Y = \theta^{\intercal}\lambda + e$, where $\theta = (\alpha^{\intercal}, \beta^{\intercal})^{\intercal}$ and $\lambda = (X^{\intercal}, \Phi_K(Z)^{\intercal})^{\intercal}$. Using the moment conditions $\mathbb{E}[\varepsilon\,\lambda] = 0$, our approach gives $\widehat{\theta} = (\sum_{i=1}^n \lambda_i\lambda_i^{\intercal})^{-1}\sum_{i=1}^n Y_i\lambda_i$, where $\lambda_i = (X_i^{\intercal}, \Phi_K(Z_i)^{\intercal})^{\intercal}$. This estimator has finite sample covariance matrix conditional on X_i, Z_i, $i = 1, \ldots, n$) equal to $\sigma^2(\sum_{i=1}^n \lambda_i\lambda_i^{\intercal})^{-1}$. Making use of the block form of λ and the orthonormality of $\Phi_K(Z)$, we have the asymptotic covariance matrix for $\widehat{\alpha}$: $\lim_{K\to\infty}[\mathbb{E}(XX^{\intercal}) - \mathbb{E}(X\Phi_K(Z)^{\intercal})\mathbb{E}(\Phi_K(Z)X^{\intercal})]^{-1}$. Here, $\mathbb{E}(X\Phi_K(Z)^{\intercal}) = \mathbb{E}(\mathbb{E}(X|Z)\Phi_K(Z)^{\intercal}) = \mathbb{E}(h(Z)\Phi_K(Z)^{\intercal})$ are the coefficients of the expansion of $h(Z) := \mathbb{E}(X|Z)$ in terms of the orthogonal basis $\Phi_K(Z)$, hence $\mathbb{E}(X\Phi_K(Z)^{\intercal})\Phi_K(Z)$ converges to $\mathbb{E}(X|Z)$ in some sense as $K \to \infty$. Finally, $\mathbb{E}(X\Phi_K(Z)^{\intercal})\mathbb{E}(\Phi_K(Z)X^{\intercal}) \to \mathbb{E}(\mathbb{E}(X|Z)\mathbb{E}(X|Z)^{\intercal})$ as $K \to \infty$, which gives that the covariance of $\widehat{\alpha}$ converges to $(\mathbb{E}[XX^{\intercal}] - \mathbb{E}(\mathbb{E}(X|Z)\mathbb{E}(X|Z)^{\intercal}))^{-1} = (\mathbb{E}[\tilde{X}\tilde{X}^{\intercal}])^{-1}$, the same as Robinson's. We now consider the case where $p \to \infty$. We partition $\alpha = (\alpha_1, \alpha_2^{\intercal})^{\intercal}$, where α_1 is a scalar parameter of interest and α_2 is of dimension $p-1$. It follows from Theorem 3.2 that our estimator of α_1 is square root-n consistent (under our conditions) and has asymptotic variance given by σ^2/ω (provided $\omega > 0$), where $\omega = \lim_{p\to\infty}\mathbb{E}\left[\left\{\tilde{X}_1 - \mathbb{E}(\tilde{X}_1\tilde{X}_2^{\intercal})(\mathbb{E}[\tilde{X}_2\tilde{X}_2^{\intercal}])^{-1}\tilde{X}_2\right\}^2\right]$, $\tilde{X}_1 = X_1 - \mathbb{E}(X_1|Z)$ and $\tilde{X}_2 = X_2 - \mathbb{E}(X_2|Z)$.

We expect this to be the semiparametric efficiency bound of α_1 under Gaussian errors, although there is very little work on efficiency bounds in the case where the parametric part is large. Under similar conditions, the Robinson estimator achieves the same efficiency. However, one important difference between our method and Robinson's method is that his requires each function $E(Y|Z)$ and $E(X_j|Z)$, $j = 1, \ldots, p$ to satisfy smoothness conditions, whereas we only need to assume smoothness directly on the single function g. More generally, our regularity conditions 2.1–2.2 and 3.1–3.8 can be verified under primitive conditions similar to Robinson (1988). □

The limiting normal distribution involves unknown parameters in the matrix Σ_n. In practice one would need a consistent estimator for this matrix. It is easily seen that the estimator, $\widehat{\Sigma}_n$, in which we replace α and $g(\cdot)$ in Σ_n by $\widehat{\alpha}$ and $\widehat{g}(\cdot)$, as well as the expectations in Ξ_n and Ψ_n by their sample versions, is consistent. More precisely, let

$$\widehat{\Sigma}_n^2 = \widehat{\Gamma}_n[\widehat{\Psi}_n\widehat{\Psi}_n^{\intercal}]^{-1}\widehat{\Psi}_n\widehat{\Xi}_n\widehat{\Psi}_n^{\intercal}[\widehat{\Psi}_n\widehat{\Psi}_n^{\intercal}]^{-1}\widehat{\Gamma}_n^{\intercal},$$

where $\widehat{\Gamma}_n$ is Γ_n with replacement of $\partial\mathscr{L}(\alpha)$ by $\partial\mathscr{L}(\widehat{\alpha})$ and of $\mathscr{F}'(g)$ by $\mathscr{F}'(\widehat{g})$, and

$$\widehat{\Xi}_n := \frac{1}{n}\sum_{i=1}^n [m(V_i, \widehat{\alpha}^{\intercal}X_i, \widehat{g}(Z_i))m(V_i, \widehat{\alpha}^{\intercal}X_i, \widehat{g}(Z_i))^{\intercal}], \tag{3.3}$$

$$\widehat{\Psi}_n := \frac{1}{n}\sum_{i=1}^n \begin{pmatrix} \frac{\partial}{\partial u}m(V_i, \widehat{\alpha}^{\intercal}X_i, \widehat{g}(Z_i))^{\intercal} \otimes X_i \\ \frac{\partial}{\partial w}m(V_i, \widehat{\alpha}^{\intercal}X_i, \widehat{g}(Z_i))^{\intercal} \otimes \Phi_K(Z_i) \end{pmatrix}. \tag{3.4}$$

Then, the feasible version of the CLT (3.2), with $\widehat{\Sigma}_n$ replacing Σ_n, follows by similar arguments to those in the proof of Theorem 3.2. This allows the construction of simultaneous confidence intervals and consistent hypothesis tests about $\mathscr{L}(\alpha)$ and $\mathscr{F}(g)$.

We may improve efficiency by using a weight matrix. Let W_n be a $q \times q$ positive definite matrix that may depend on the sample data. Then, $\|M_n(\mathbf{a}, \mathbf{b})\|^2$, which measures the metric of $M_n(\mathbf{a}, \mathbf{b})$ from zero, can be substituted by $M_n(\mathbf{a}, \mathbf{b})^{\intercal}W_nM_n(\mathbf{a}, \mathbf{b})$ in the minimization of (2.4), which is also a measure of the metric for the vector $M_n(\mathbf{a}, \mathbf{b})$ from zero but in terms of the weight matrix W_n. Meanwhile, $\|M_n(\mathbf{a}, \mathbf{b})\|^2$ can be viewed as a special case that W_n is the identity matrix. We require the matrix W_n to be not too close to be singular to prevent the possibility that $M_n(\mathbf{a}, \mathbf{b})^{\intercal}W_nM_n(\mathbf{a}, \mathbf{b})$ may be close to zero when $(\mathbf{a}, \mathbf{b})$ is far from (α, β).

Proposition 3.1. *Suppose that the eigenvalues of W_n are bounded away from zero and above from infinity uniformly in n, and there exists a deterministic matrix W^* such that $\|W_n - W^*\| = o_P(1)$ as $n \to \infty$. Let $(\widetilde{\alpha}, \widetilde{\beta})$ be the minimizer of $M_n(\mathbf{a}, \mathbf{b})^\tau W_n M_n(\mathbf{a}, \mathbf{b})$ and define $\widetilde{g}(z) = \Phi_K(z)^\tau \widetilde{\beta}$.*

Then, (1) Under the same conditions in Theorem 3.1, the consistency of the weighted estimator holds; (2) Under the same conditions the normality for the weighted estimator in Theorem 3.2 holds with Σ_n^2 replaced by

$$\Gamma_n[\Psi_n W^* \Psi_n^\tau]^{-1}\Psi_n W^* \Xi_n W^* \Psi_n^\tau [\Psi_n W^* \Psi_n^\tau]^{-1}\Gamma_n^\tau.$$

(3) If $W^* = \Xi_n^{-1}$, the optimal covariance matrices is obtained, $\Gamma_n[\Psi_n \Xi_n^{-1}\Psi_n^\tau]^{-1}\Gamma_n^\tau$.

The proof is given in Appendix B. Here, the optimal covariance is in the sense that

$$\Gamma_n[\Psi_n W \Psi_n^\tau]^{-1}\Psi_n W \Xi_n W \Psi_n^\tau [\Psi_n W \Psi_n^\tau]^{-1}\Gamma_n^\tau \geq \Gamma_n[\Psi_n \Xi_n^{-1}\Psi_n^\tau]^{-1}\Gamma_n^\tau,$$

for all W satisfying the conditions in the proposition. Though $W_n = \Xi_n^{-1}$ might make the estimator efficient, it is not feasible since Ξ_n involves the true parameters. In practice, both Ξ_n and Ψ_n can be replaced by their sample versions of (3.3) and (3.4), so that the optimal covariance matrices are easily estimable. To do so, one will need to implement a two-step estimation method, as has normally been done in the literature, that is, at the first step minimizing $\|M_n(\mathbf{a}, \mathbf{b})\|^2$ to have $\widehat{\alpha}$ and $\widehat{g}(\cdot)$ that are used to construct $\widehat{W}_n = \widehat{\Xi}_n^{-1}$; then at the second step one may minimize $M_n(\mathbf{a}, \mathbf{b})^\tau \widehat{W}_n M_n(\mathbf{a}, \mathbf{b})$ to have a pair of optimal estimators, $(\widetilde{\alpha}, \widetilde{g}(\cdot))$.

There is an alternative way that achieves efficiency in one-step estimation, viz., the continuous updating estimator (CUE) and generalized empirical likelihood estimator; see Newey and Smith (2004), Chang et al. (2015). Define $W_n(\mathbf{a}, \mathbf{b}) = [\Xi_n(\mathbf{a}, \mathbf{b})]^{-1}$, where

$$\Xi_n(\mathbf{a}, \mathbf{b}) := \frac{1}{n}\sum_{i=1}^n [m(V_i, \mathbf{a}^\tau X_i, \mathbf{b}^\tau \Phi_K(Z_i))m(V_i, \mathbf{a}^\tau X_i, \mathbf{b}^\tau \Phi_K(Z_i))^\tau].$$

Then, $(\widetilde{\alpha}, \widetilde{g}(\cdot))$ can be estimated by minimizing $M_n(\mathbf{a}, \mathbf{b})^\tau W_n(\mathbf{a}, \mathbf{b}) M_n(\mathbf{a}, \mathbf{b})$ over $(\mathbf{a}, \mathbf{b})$. We do not pursue this direction here, but refer the reader to Hansen et al. (1996) .

3.3. Single-index structure

When the dimension of variable Z is relatively small, model (2.6) can be estimated by the procedure (2.7) which essentially is the same as (2.4). Thus, $(\widehat{\alpha}, \widehat{\beta})$ is consistent under similar conditions as in Theorem 3.1. When $\widehat{\beta}$ is obtained we need to detangle to have the estimates of c_j and θ_0, from which we can construct the estimator of $g(z)$.

If $c_1 \neq 0$, by the relationship $a_{1u}(\theta_0) = \sqrt{2}\pi^{d/4} c_1 \theta_0^u$ for all $|u| = 1$, along with the identification condition on θ_0, the estimate $\widehat{\beta}$ gives

$$\widehat{c}_1 = \operatorname{sgn}(\widehat{a}_{1u_0}) \frac{1}{\sqrt{2}\pi^{d/4}} \left(\sum_{|u|=1} \widehat{a}_{1u}^2\right)^{1/2}, \quad \text{where } u_0 = (1, 0, \ldots, 0).$$

Because $\widehat{\beta}$ is consistent, $\widehat{c}_1 \neq 0$ with probability approaching one. Let

$$\widehat{\theta} = \frac{1}{\widehat{c}_1} Q \widehat{\beta},$$

where $Q = (0_{d\times p}, I_d, 0_{d\times(K-d)})$ that chooses the corresponding estimates in $\widehat{\beta}$ of all a_{1u} with $|u| = 1$. As the estimate of θ_0, $\widehat{\theta}$ is consistent by the consistency of $\widehat{\beta}$.

If $c_1 = 0$, without loss of generality suppose that there exists some j_0, $1 \leq j_0 \leq k-1$, such that $c_{j_0} \neq 0$ (this can be almost guaranteed since k diverges). Then, the estimate of θ_0 can be recovered by all estimates of $a_{j_0 u}(\theta_0)$ in β, which is similar to but a bit complicated than the case of $c_1 \neq 0$. We omit this as it is the same as Dong et al. (2015, p. 304). It follows that we obtain the estimate $\widehat{\theta}$ of θ_0, along with that of c_j, from $\widehat{\beta}$.

Now we turn to consider another situation where the index vector θ_0 in model (2.6) satisfies one extra equation that helps to estimate the vector.

The model of Carneiro et al. (2011) is in this situation. In their case, the marginal treatment effect (MTE) is $MTE(x, p) = x^\tau \alpha + g'(p)$ and the parameter of interest is the weighted average MTE, $\Delta = \int_0^1 MTE(x, p)h(x, p)dp$ for some known weighting function h. The parameter θ_0 can be estimated from the moment equation derived from the second conditional moment in (1.3), $\mathbb{E}\left[(\mathbb{I}(S = 1) - \Lambda(\theta_0^\tau Z))\Psi_q(Z)\right] = 0$, with or without the specification of the function Λ, using the conventional technique for dealing with single-index models, such as Ai and Chen (2003), Dong et al. (2016).

Although θ_0 can be estimated by the second equation of (1.3), in order to derive asymptotic distributions for the estimators of α and g defined later, it is convenient if $\widehat{\theta}$, the estimate of θ_0, is independent of the data used to estimate α and g by the first equation. This is possible and one way to do is as follows. Let us split the observations $\{V_i, X_i, Z_i, i = 1, \ldots, n\}$ into two subsamples randomly, $\text{Sub}_1 := \{(V_i, X_i, Z_i), i = 1, \ldots, n'\}$ and $\text{Sub}_2 := \{V_i, X_i, Z_i, i = n'+1, \ldots, n\}$,

with $n' = [n/2]$. The ordering in both subsamples in general is not the same as in the original sample but we keep using subscript i after partition. The first subsample Sub_1 can be used to estimate θ_0 by an additional moment restriction (say), resulting in $\widehat{\theta}$, and the second Sub_2 is used to estimate the parameter α and function g. Here, due to the i.i.d. property of the sample, the independence property holds naturally. Additionally, $\sqrt{n}(\widehat{\theta} - \theta_0) = O_P(1)$ (e.g. Yu and Ruppert (2002)). The data-splitting technique is used in the literature, such as Bickel (1982) and Belloni et al. (2012). The independence property is important for our theoretical development and thus we recommend the use of the data-splitting method in the rest of this section. Due to this reason, we make the following assumption.

Assumption 3.9. For θ_0 in Eq. (2.6), there exists an estimator $\widehat{\theta}$ such that $\sqrt{n}(\widehat{\theta} - \theta_0) = O_P(1)$ as $n \to \infty$ and assume that $\widehat{\theta}$ is independent of observations used in minimization Eq. (3.5).

With the single-index structure, the nonparametric function is defined on the real line. Therefore, for the establishment of our theory, we need some corresponding assumptions that are counterparts of Assumptions 2.1, 3.1–3.3, 3.5 and 3.7, denoted by Assumptions 2.1*, 3.1*–3.3*, 3.5* and 3.7*, respectively, and are given in Appendix A for brevity.

Under Assumption 2.1* we have the expansion of $g(z)$ and hence $g(z)$ can be approximated by the partial sum, that is, $g(z) = \sum_{j=0}^{K-1} b_j \varphi_j(z) + \gamma_K(z)$ with $\gamma_K(z) \to 0$ in some sense. Hence, we can estimate $\beta = (b_0, \ldots, b_{K-1})^\tau$, together with α, by

$$
\begin{aligned}
(\widehat{\alpha}, \widehat{\beta}) &= \underset{\mathbf{a}\in\mathbb{R}^p, \mathbf{b}\in\mathbb{R}^K}{\operatorname{argmin}} \|\widetilde{M}_n(\mathbf{a}, \mathbf{b})\|^2, \quad \text{subject to } \|\mathbf{a}\| \le B_{1n} \text{ and } \|\mathbf{b}\| \le B_{2n}, \\
&\text{where } \widetilde{M}_n(\mathbf{a}, \mathbf{b}) = \frac{1}{\sqrt{q}} \frac{1}{n - n'} \sum_{i=n'+1}^{n} m(V_i, \mathbf{a}^\tau X_i, \mathbf{b}^\tau \Phi_K(\widehat{\theta}^{\,\tau} Z_i)),
\end{aligned} \tag{3.5}
$$

where $\Phi_K(z)$ is the vector of the basis functions. With this $\widehat{\beta}$, we can define similarly $\widehat{g}(z) = \widehat{\beta}^\tau \Phi_K(z)$.

Theorem 3.3 (*1*)**.** *Under Assumptions 2.1*, 2.2, 3.1*, 3.2*, 3.3*, and 3.9, the consistency in Theorem 3.1 are satisfied by the $\widehat{\alpha}$ and $\widehat{g}(z)$ defined in this subsection.*

(2) Let Assumptions 2.1, 2.2, 3.1*–3.3*, 3.4, 3.5*, 3.6, 3.7*, and 3.9 hold. Then, the normality in Theorem 3.2 is valid for the $\widehat{\alpha}$ and $\widehat{g}(z)$ defined in this subsection with replacement of Ξ_n and Ψ_n respectively by*

$$
\begin{aligned}
\widetilde{\Xi}_n &:= \mathbb{E}[m(V, \alpha^\tau X, g(\theta_0^\tau Z)) m(V, \alpha^\tau X, g(\theta_0^\tau Z))^\tau]_{q\times q}, \\
\widetilde{\Psi}_n &:= \mathbb{E}\left(\begin{array}{c} \frac{\partial}{\partial u} m(V, \alpha^\tau X, g(\theta_0^\tau Z))^\tau \otimes X \\ \frac{\partial}{\partial w} m(V, \alpha^\tau X, g(\theta_0^\tau Z))^\tau \otimes \Phi_K(\theta_0^\tau Z) \end{array} \right)_{(p+K)\times q}.
\end{aligned}
$$

Using Lemmas A.5–A.7 in Appendix A, the theorem is proven in the supplemental material of the paper. The estimation of the covariance matrix can be obtained similarly to that in Theorem 3.2 and we omit this for brevity.

4. Statistical inference

4.1. Test of over-identification

Hansen (1982) proposes the J-test for over-identification in the situation where both p and q are fixed but $q > p$. This J-test has an asymptotic χ^2_{q-p} null distribution. In the case where an unknown infinite dimensional parameter is involved, and both p and q are still fixed with $q > p$, Chen and Liao (2015) establish a statistic for over-identification testing that has an F distribution in large samples. We propose an alternative statistic below, which as far as we are aware, appears to be new.

We consider the following hypotheses in model (2.1):

$$
\begin{aligned}
H_0: &\quad \mathbb{E}[m(V_i, \alpha^\tau X_i, g(Z_i))] = 0 \quad \text{for some } (\alpha, g) \in \Theta, \\
H_1: &\quad \mathbb{E}[m(V_i, \mathbf{a}^\tau X_i, h(Z_i))] \neq 0 \quad \text{for any } (\mathbf{a}, h) \in \Theta,
\end{aligned}
$$

where Θ is defined in Section 2.

Define, for $\mathbf{a} \in \mathbb{R}^p$, $\mathbf{b} \in \mathbb{R}^K$ and any given $\kappa \in \mathbb{R}^q$ such that $\|\kappa\| = 1$,

$$
L_n(\mathbf{a}, \mathbf{b}; \kappa) = \frac{1}{D_n(\mathbf{a}, \mathbf{b}; \kappa)} \sum_{i=1}^{n} \kappa^\tau m(V_i, \mathbf{a}^\tau X_i, \mathbf{b}^\tau \Phi_K(Z_i)),
$$

where $D_n(\mathbf{a}, \mathbf{b}; \kappa) = \left(\sum_{i=1}^n [\kappa^\tau m(V_i, \mathbf{a}^\tau X_i, \mathbf{b}^\tau \Phi_K(Z_i))]^2\right)^{1/2}$.

Under the null hypothesis, by the procedure in Section 2 and the conditions of Theorem 3.1, the estimator $(\widehat{\alpha}, \widehat{g})$ is consistent. The statistic $L_n(\widehat{\alpha}, \widehat{\beta}; \kappa)$ can be used to detect H_0 against H_1, as shown in Theorems 4.1 and 4.2. This test also works for the conventional moment restriction models with fixed p and q. Before establishing an asymptotic distribution under the null and asymptotic consistency under the alternative, we introduce some assumptions.

Assumption 4.1. Let $\overline{m}_n^*(\widehat{\alpha}, \widehat{g}; \kappa) = o_P(1)$ when $n \to \infty$, where we denote $\overline{m}_n^*(a, f; \kappa) = n^{-1/2} \sum_{i=1}^n \mathbb{E}[\kappa^\intercal m(V_i, a^\intercal X_i, f(Z_i))]$ for $(a, f) \in \Theta$ and κ such that $\|\kappa\| = 1$.

Assumption 4.2. Suppose that (i) $qp^2 = o(n)$ and $qK^2 = o(n)$; and (ii) $\sup_z \gamma_K^2(z) = o(q^{-1})$ as, along with $n \to \infty$, $K, p, q \to \infty$.

These are technical requirements. Noting $\mathbb{E}[m(V, \alpha^\intercal X, g(Z))] = 0$, Assumption 4.1 requires that $\mathbb{E}[m(V, \mathbf{a}^\intercal X, f(Z))]$ drops to zero very quickly when $(\mathbf{a}, f)$ approaches (α, g). This is the same, in spirit, as Assumption 3.2, but here it is a sample version and the decay of the expectation needs a certain rate. A similar assumption is also imposed by (4.9) of Andrews (1994, p. 58) and (5.40) of Belloni et al. (2014b, p. 634). Assumption 4.2 (i) stipulates the relationship between (p, q, K) and n when they are diverging, while Assumption 4.2(ii) imposes a decay rate for the residue $\gamma_K^2(z)$ uniformly for all z not slower than $o(q^{-1})$. This is trivially satisfied for the case where either z is located in some compact set or $g(z)$ is integrable on the real line, given that the g function is sufficiently smooth.

Theorem 4.1. *Suppose that there is no zero function in the vector m of functions. Let Assumptions 4.1–4.2 hold, and the conditions of Theorems 3.1 and 3.2 remain true. For any $\kappa \in \mathbb{R}^q$ such that $\|\kappa\| = 1$, under H_0,*

$$L_n(\widehat{\alpha}, \widehat{\beta}; \kappa) \to_D N(0, 1),$$

as $n \to \infty$, where $(\widehat{\alpha}, \widehat{\beta})$ is the estimator given by (2.4).

Notice that if there are zero functions in m, the product $\kappa^\intercal m$ can be a zero function for some particular choice of κ. Thus, excluding zero functions is necessary. The theorem establishes the normality of the proposed statistic under the null that enables us to make statistical inference.

Theorem 4.2. *Suppose that the eigenvalues of $\mathbb{E}[m(V, \mathbf{a}^\intercal X, h(Z))m(V, \mathbf{a}^\intercal X, h(Z))^\intercal]$ are bounded away from zero and infinity uniformly in n and $(\mathbf{a}, h) \in \Theta$. Under H_1, suppose further that there exists a positive sequence δ_n such that $\inf_{(\mathbf{a},h)\in\Theta} \|E[m(V, \mathbf{a}^\intercal X, h(Z))]\| \geq \delta_n$ and $\liminf_{n\to\infty} \sqrt{n}\delta_n = \infty$. Then, for any vectors $\mathbf{a}$ and $\mathbf{b}$, there exists some $\kappa^* \in \mathbb{R}^q$ such that $\|\kappa^*\| = 1$ and $L_n(\mathbf{a}, \mathbf{b}; \kappa^*) \to_P \infty$, as $n \to \infty$.*

The condition on the eigenvalues is commonly adopted in the literature, see, e.g. Chang et al. (2015), Belloni et al. (2015). The expression of κ^* shown in the proof of the theorem is $\kappa^* = \mathbb{E}[m(V_i, \mathbf{a}^\top X_i, \mathbf{b}^\top \Phi_K(Z_i))]/\|\mathbb{E}[m(V_i, \mathbf{a}^\top X_i, \mathbf{b}^\top \Phi_K(Z_i))]\|$, where the denominator does not vanish ensured by H_1. Moreover, in the special case where $\delta_n = \delta$, the condition that $\liminf_{n\to\infty} \sqrt{n}\delta_n = \infty$ is satisfied automatically, and this is the most commonly used assumption in the literature, see, equation (24) of Chang et al. (2015, p. 290). However, we allow for $\delta_n \to 0$ with a rate slower than $n^{-1/2}$. This means that the strongest signal ($\delta_n = \delta$) can be weakened ($\delta_n \to 0$) when our test statistic is used.

4.2. Student t test

We next propose an alternative test for model (2.1) under H_0. Define $\widehat{m}(i) := m(V_i, \widehat{\alpha}^\intercal X_i, \widehat{g}(Z_i))$ for simplicity and correspondingly, for later use define $m(i) := m(V_i, \alpha^\intercal X_i, g(Z_i))$. Let $\widehat{e} = (\widehat{e}_1, \ldots, \widehat{e}_q)^\intercal$ and $\widehat{\sigma}^2 = (\widehat{\sigma}^2(i, j))_{q\times q}$, where

$$\widehat{e} := \frac{1}{n}\sum_{i=1}^n \widehat{m}(i), \quad \text{and} \quad \widehat{\sigma}^2 := \frac{1}{n}\sum_{i=1}^n \widehat{m}(i)\widehat{m}(i)^\intercal,$$

Here, $\widehat{e}$ and $\widehat{\sigma}^2$ may be understood as the estimated mean and covariance matrix of the error vector, respectively. Define

$$T_n := \frac{1}{q}\sum_{j=1}^q \left(\frac{\sqrt{n}\widehat{e}_j}{\widehat{\sigma}(j, j)}\right)^2.$$

The statistic is constructed from $\sqrt{n}\widehat{e}_j/\widehat{\sigma}_n(j, j)$, which is somewhat like the traditional t-test. Pesaran and Yamagata (2017) proposed a similar statistic.

Theorem 4.3. *Let the conditions of Theorems 3.1–3.2 hold. Let also Assumptions 4.1–4.2 hold under H_0. Suppose that $\mathbb{E}[m(i)m(i)^\intercal]$ is a diagonal matrix with $\min_{1\leq j\leq q} \mathbb{E}[m_j(i)^2] > c > 0$ and $\sup_{1\leq j\leq q} \mathbb{E}[m_j(i)^4] < C < \infty$ for some constants c and C. Then, $\sqrt{q/2}(T_n - 1) \to_D N(0, 1)$ as $n \to \infty$.*

The proof is given in Appendix C of the supplement. The requirement on $\mathbb{E}[m(i)m(i)^\intercal]$ to be a diagonal matrix implies the orthogonality between the errors. This is not stringent because, if not so, we may make a transformation $\tilde{m}(i) = (\mathbb{E}[m(i)m(i)^\intercal])^{-1/2}m(i)$ and then $\tilde{m}(i)$ would meet the requirement. Moreover, in many situations it is satisfied naturally. For instance, in Example 1.1 of Section 1, $m(i)$ is consisting of orthogonal functions of the conditional variable. These moment requirements are commonly used in the literature since $m_j(i)$ are generalized error terms, so we do not explain them in detail. In addition, the behaviour of T_n is like $\chi^2(q)$ but with diverging q. Therefore, after normalization we have asymptotic normal distribution for T_n.

Next, consider the consistency of T_n. For any vector $\mathbf{a} \in \mathbb{R}^p$ and function $h(\cdot)$, define $\widetilde{m}(i) \equiv \widetilde{m}(i; \mathbf{a}, h) = m(V_i, \mathbf{a}^\tau X_i, h(Z_i))$, $\widetilde{e} = (\widetilde{e}_1, \ldots, \widetilde{e}_q)^\tau$ and $\widetilde{\sigma} = (\widetilde{\sigma}_{ij})_{q\times q}$, where

$$\widetilde{e} = \frac{1}{n}\sum_{i=1}^{n} \widetilde{m}(i), \quad \text{and} \quad \widetilde{\sigma} = \frac{1}{n}\sum_{i=1}^{n} \widetilde{m}(i)\widetilde{m}(i)^\tau.$$

Define also

$$\widetilde{T}_n := \frac{1}{q}\sum_{j=1}^{q}\left(\frac{\sqrt{n}\widetilde{e}_j}{\widetilde{\sigma}(j,j)}\right)^2.$$

Note that if H_0 is true, $\widetilde{T}_n$ would become T_n when $\mathbf{a}$ and $h(\cdot)$ are substituted by $\widehat{\alpha}$ and $\widehat{g}$, respectively, while if H_1 is true, $\widetilde{T}_n$ would diverge as shown in the following theorem.

Theorem 4.4. *Suppose that* $\max_{1\le j\le q}\sup_{\mathbf{a},h}\mathbb{E}[\widetilde{m}_j(i)^2] \le C < \infty$ *for some constant* C. *Then, under the conditions in Theorem 4.2 and* H_1, *for any vector* $\mathbf{a} \in \mathbb{R}^p$ *and function* $h(\cdot)$, *as* $n \to \infty$, $\widetilde{T}_n \to_P \infty$ *provided that* $\sqrt{n/q}\delta_n \to \infty$.

The proof is given in Appendix C of the supplemental material. Notice that in terms of statistical inference in practice it is impossible to distinguish T_n from $\widetilde{T}_n$. Instead, one may use our estimation procedure to obtain the "estimates" of the parameters, then construct $\widetilde{T}_n$ and finally make an inference according to Theorem 4.3. The uniform boundedness of the second moment is reasonable in the i.i.d. setting. Comparing with Theorem 4.2, the attenuation of δ_n is slowed down as we require $\sqrt{n/q}\delta_n \to \infty$. This is because of the difference in the constructions of T_n and $L_n(\mathbf{a}, \mathbf{b}; \kappa)$.

5. Penalized GMM under sparsity

We now consider the ultra-high dimensional situation of model (2.1) where the potential number of covariates is much larger than the sample size (i.e., $p = e^{n^a}$ with $0 < a < 1$), but the parameter vector α has sparsity. That is, there are many zeros in α and only a number of elements are nonzero, but the identity of the non-zero elements is not known a priori. In addition, the coefficient vector β in the partial sum of the expansion of the nonparametric function may also possess sparsity in two potential scenarios: (a) its elements may be zero if the unknown function is located in a subspace that has small dimensionality (e.g. the simulation below), and (b) its elements are attenuated as the number of terms increases, so that many of them are negligible statistically. Hence, this section is devoted to estimate (α, g) under the sparsity condition. This "big-data" context is becoming increasingly relevant in applications.

There are some existing papers on the variable selection under sparsity. Belloni et al. (2014a) propose the combination of least squares and L_1 type lasso approach to select coefficients of the sieve in nonparametric regression. Also, Su et al. (2018) use L_1 type lasso approach to study continuous treatment in nonseparable models with high dimensional data. In a high dimensional conditional moment restriction model, Fan and Liao (2014) propose to use a folded concave penalty function combined with instrumental variables to select the important coefficients. Caner (2009) uses the same approach with a particular class of penalty functions to select variables. As Caner (2009, p. 271) argued, the Lasso-type GMM estimator selects the correct model much more often than GMM-BIC and the "downward testing" method proposed by Andrews and Lu (2001). We shall tackle the selection issue by the combination of a penalty function and our GMM approach.

Regarding of GMM approach, to reduce the risk of misspecification Andrews (1999) defines moment selection criterion (MSC) using J-test statistics and shows that the consistent moment selection can be achieved by choosing the selection vector minimizing the MSC. There are also other papers studying the selection issue using generalized empirical likelihood statistic. However, we mention that all of these methods are nonetheless subject to pretest bias and post-model selection inferential problems (Leeb and Pötscher (2005)).

We partition the parameter vectors as $\alpha = (\alpha_{0S}^\tau, \alpha_{0N}^\tau)^\tau$ and $\beta = (\beta_{0S}^\tau, \beta_{0N}^\tau)^\tau$, where the vectors α_{0S} and β_{0S} contain all "important coefficients" from α and β (i.e. nonzero coefficients), respectively, as referred in the literature such as Fan and Liao (2014), while α_{0N} and β_{0N} are zero.

For convenience in this section, denote $v_0 = (\alpha^\tau, \beta^\tau)^\tau \in \mathbb{R}^{p+K}$ the true parameter whose dimension varies with the sample size. In addition, $v_{0S} = (\alpha_{0S}^\tau, \beta_{0S}^\tau)^\tau$ is referred to as an oracle model. Define $t_n = |v_{0S}|$ the dimension of v_{0S}, which may diverge with n.

Let $\widehat{v} \in \mathbb{R}^{p+K}$ be the estimated parameter of v_0 by the penalized GMM, which solves:

$$\widehat{v} = (\widehat{\alpha}^\tau, \widehat{\beta}^\tau)^\tau = \underset{v=(\mathbf{a}^\tau,\mathbf{b}^\tau)^\tau\in\mathbb{R}^{p+K}}{\operatorname{argmin}} Q_n(v) := \|M_n(v)\|^2 + \sum_{j=1}^{p+K} P_n(|v_j|), \tag{5.1}$$

where $M_n(v) = M_n(\mathbf{a}, \mathbf{b})$ is as defined in Section 2, and $P_n(\cdot)$ is a penalty function discussed later. Our framework also accommodates the case where some components of α, β are entered without selection, as in Belloni et al. (2016a), although we do not inscribe this in the notation for simplicity.

5.1. Oracle property

Let T be the support of v_0, the indexes of the nonzero components, i.e., $T = \{j : 1 \le j \le p+K, v_{0j} \neq 0\}$. We may equivalently say that T is the oracle model. Moreover, for a generic vector $v \in \mathbb{R}^{p+K}$, denote by v_T the vector in $\mathbb{R}^{p+K}$ whose jth element equals v_j if $j \in T$ and zero otherwise. Also, define v_S as the short version of v_T after eliminating all zeros in the position T^c (the complement set of T) from v_T. In the literature, the subspace $\mathcal{V} = \{v_T, v \in \mathbb{R}^{p+K}\}$ is called the "oracle space" of $\mathbb{R}^{p+K}$. Certainly, $v_0 \in \mathcal{V}$.

Recall that the score vector $S_n(\cdot)$ denotes the partial derivative of $\|M_n(\cdot)\|^2$ defined in Section 3. Now, denote $S_{nT}(v_S)$ the partial derivative of $\|M_n(v)\|^2$ with respect to v_j for $j \in T$, at v_T (bearing in mind that v_S is the short version of v_T). Hence, the vector $S_{nT}(v_S)$ has dimension $t_n = |T| = |v_S|$. Here and hereafter, for set T, $|T|$ stands for its cardinality, while for a vector v, $|v|$ stands for its dimension. Also, define in a similar fashion $H_{nT}(v_S)$ the $t_n \times t_n$ Hessian matrix for $\|M_n(v)\|^2$.

Suppose that $P_n(\cdot)$ belongs to the class of folded concave penalty functions (see Fan and Li (2001)). For any generic vector $v = (v_1, \ldots, v_{t_n})^\tau \in \mathbb{R}^{t_n}$ with $v_j \neq 0$, for all j, define

$$\phi(v) = \limsup_{\epsilon \to 0^+} \max_{j \le t_n} \sup_{u_1 < u_2, (u_1,u_2) \subset O(|v_j|,\epsilon)} -\frac{P_n'(u_2) - P_n'(u_1)}{u_2 - u_1},$$

where $O(\cdot,\cdot)$ is the neighbourhood with specified centre and radius, respectively, implying that $\phi(v) = \max_{j \le t_n} -P_n''(|v_j|)$ if P_n'' is continuous. Also, for the true parameter v_0, let

$$d_n = \frac{1}{2}\min\{|v_{0j}| : v_{0j} \neq 0, j = 0, \ldots, p+K\}$$

represent the strength of the signal. The following assumption is about the penalty function.

Assumption 5.1. The penalty function $P_n(u)$ satisfies (i) $P_n(0) = 0$; (ii) $P_n(u)$ is concave, nondecreasing on $[0, \infty)$, and has a continuous derivative $P_n'(u)$ for $u > 0$; $(iii)\sqrt{t_n}P_n'(d_n) = o(d_n)$; (iv) There exists $c > 0$ such that $\sup_{v \in O(v_{0S}, cd_n)} \phi(v) = o(1)$.

There are many classes of functions that satisfy these conditions. For example, with a properly chosen tuning parameter in each case, the L_r penalty ($0 < r \le 1$), hard-thresholding (Antoniadis (1996)), SCAD (Fan and Li (2001)) and MCP (Zhang (2010)) all satisfy the requirements.

Denoting the oracle model $T = T_1 \cup T_2$, where T_1 is the set of indices of nonzero elements in α and T_2 that of β, accordingly, we have $t_n = p_1 + K_1$ for the corresponding cardinalities.

Assumption 5.2. Let Assumptions 3.5–3.7 hold with p being replaced by p_1 and K by K_1.

The assumption is a counterpart of Assumptions 3.5–3.7 under sparsity.

Assumption 5.3. There exist $b_1, b_2 > 0$ such that (i) for any $\ell \le q$ and $u > 0$,

$$P(|m_\ell(V, \alpha^\tau X, \beta^\tau \Phi_K(Z))| > u) \le \exp(-(u/b_1)^{-b_2});$$

and (ii) $Var(m_\ell(V, \alpha^\tau X, \beta^\tau \Phi_K(Z)))$ are bounded away from zero and above from infinity uniformly for all ℓ.

This assumption is often encountered in the literature, such as Assumption 4.3 in Fan and Liao (2014). It is known that there are many classes of distributions satisfying this condition, e.g., a continuous distribution with compact support, a normal distribution, and an exponential distribution and so on. The thin tail of the distribution postulated in the assumption enables us to bound the score function.

For simplicity, denote ∂m the partial derivative of m; and $F_{iS} = \mathrm{diag}(X_{iS}, \Phi_{KS}(Z_i))$ a $t_n \times 2$ matrix where X_{iS} is the sub-vector of X_i consisting of all X_{ij} for $j \in T_1$; $\Phi_{KS}(Z_i)$ is the sub-vector of $\Phi_K(Z_i)$ consisting of all $\varphi_j(Z_i)$ for $j \in T_2$.

Assumption 5.4. (i) There are constants $C_1, C_2 > 0$ such that

$$\lambda_{\min}(\mathbb{E}\partial m^\tau(V_i, v_{0S}^\tau F_{iS}) \otimes F_{iS})(\mathbb{E}\partial m^\tau(V_i, v_{0S}^\tau F_{iS}) \otimes F_{iS})^\tau > C_1, \quad and$$
$$\lambda_{\max}(\mathbb{E}\partial m^\tau(V_i, v_{0S}^\tau F_{iS}) \otimes F_{iS})(\mathbb{E}\partial m^\tau(V_i, v_{0S}^\tau F_{iS}) \otimes F_{iS})^\tau < C_2;$$

(ii) $P_n'(d_n) = o(n^{-1/2})$ and $\max_{\|v_S - v_{0S}\| < d_n/4} \phi(v_S) = o((t_n \log(q))^{-1/2})$; (iii) $t_n^{3/2}\log(q) = o(n)$, $t_n^{3/2}P_n'(d_n)^2 = o(1)$ and $t_n \max_{j \in T} P_n(|v_{0j}|) = o(1)$.

All these are technical requirements on the Hessian matrix, the penalty function, the relationship among the dimensions of the important coefficients, the sparsity and the sample size. These conditions are commonly used in the literature, see, for example, Assumptions 4.5-4.6 in Fan and Liao (2014) among others. There are several penalty functions that satisfy these conditions, for example, SCAD and MCP with tuning parameter $\lambda_n = o(d_n)$. Thence, the conditions (ii) and (iii) are satisfied if $t_n\sqrt{\log(q)/n} + t_n^{3/2}\log(q)/n \ll \lambda_n \ll d_n$. However, noting that the exact identification is allowed, the total number of parameters $p + K$ of α and β to be estimated can be as large as $\exp(n^a)$ for some $0 < a < 1$, an implication of the restriction on q.

To state the following theorem, define:

$$\Sigma_{nT}^2 := \Gamma_n[\Psi_{nT}\Psi_{nT}^\intercal]^{-1}\Psi_{nT}\Xi_{nT}\Psi_{nT}^\intercal[\Psi_{nT}\Psi_{nT}^\intercal]^{-1}\Gamma_n^\intercal, \quad \text{in which} \tag{5.2}$$

$$\Gamma_n := \begin{pmatrix} \partial\mathscr{L}(\alpha_{0S}) & 0 \\ 0 & \mathscr{F}'(g)\Phi_{KT}{}^\intercal \end{pmatrix}_{(\mu+\nu)\times(p_1+K_1)},$$

$$\Xi_{nT} := \mathbb{E}[m(V_1, \alpha_{0S}^\intercal X_{1S}, g(Z_1))m(V_1, \alpha_{0S}^\intercal X_{1S}, g(Z_1))^\intercal]_{q\times q},$$

$$\Psi_{nT} := \mathbb{E}\begin{pmatrix} \frac{\partial}{\partial u}m(V_1, \alpha_{0S}^\intercal X_{1S}, g(Z_1))^\intercal \otimes X_{1S} \\ \frac{\partial}{\partial w}m(V_1, \alpha_{0S}^\intercal X_{1S}, g(Z_1))^\intercal \otimes \Phi_{KT}(Z_1) \end{pmatrix}_{(p_1+K_1)\times q},$$

provided that $\Psi_{nT}\Psi_{nT}^\intercal$ is invertible, in which u and w stand for the second and the third arguments of the vector function $m(v, u, w)$, respectively; and the transformation $\mathscr{L}$ and vector functional $\mathscr{F}$ are defined in Section 3.

Theorem 5.1. *Let Assumptions 2.1, 2.2, 3.1, 3.3 and 5.1–5.4 hold. Then, there exists a local minimizer* $\widehat{v} = ((\widehat{\alpha}_S^\intercal, \widehat{\alpha}_N^\intercal)^\intercal, (\widehat{\beta}_S^\intercal, \widehat{\beta}_N^\intercal)^\intercal)$*, for which we have (i)*

$$\lim_{n\to\infty} P(\widehat{\alpha}_N = 0, \widehat{\beta}_N = 0) = 1.$$

In addition, the local minimizer $\widehat{v}$ *is strict with probability arbitrarily close to one for all large n.*

(ii) Let $\widehat{T} = \{j : 1 \le j \le p + K, \widehat{v}_j \ne 0\}$*. Then,*

$$\lim_{n\to\infty} P(\widehat{T} = T) = 1.$$

(iii) Meanwhile, for the transformation $\mathscr{L}_{r\times p_1}$ *and s -vector functional* $\mathscr{F}$*,*

$$\sqrt{n}\Sigma_{nT}^{-1}\begin{pmatrix} \mathscr{L}(\widehat{\alpha}_S) - \mathscr{L}(\alpha_{0S}) \\ \mathscr{F}(\widehat{g}) - \mathscr{F}(g) \end{pmatrix} \xrightarrow{d} N(0, I_{\mu+\nu}),$$

as $n \to \infty$ *provided that* $\sqrt{n}\Sigma_{nT}^{-1}(0_\mu^\intercal, \mathscr{F}'(g)\gamma_K^\intercal)^\intercal = o(1)$*, where* Σ_{nT} *is given by the square root of* Σ_{nT}^2 *defined in (5.2).*

The proof is given in Appendix B. Note that the undersmoothing condition can be satisfied if Σ_{nT} that has finite dimensionality has minimal eigenvalue greater than zero and γ_K decays to zero sufficiently fast. We remark that, due to the asymptotic theory in Section 3, the post selection version of the standard errors defined in (3.3) and (3.4) can be shown to be consistent in this case thereby allowing consistent confidence intervals for the selected parameters.

The estimators in this theorem are all local. This is why we exclude the identification condition in Assumption 3.2 currently, while in the next theorem we shall discuss the global property of a local minimizer. The results (i) and (ii) indicate that under these conditions in the theorem we are able to recover the sparsity in the model; meanwhile, the discussion on the result (iii) of the theorem is similar to Theorem 3.2.

5.2. Global property

In this section, we show that under Assumption 3.2, the local minimizer in Theorem 5.1 is nearly global. Recall that Assumption 3.2 is an identification condition that excludes all the other points to be the minimizer of the objective function in the population sense.

Theorem 5.2. *In addition to the conditions of Theorem 5.1, suppose Assumption 3.2 holds. Then, the local minimizer* $\widehat{v}$ *satisfies that, for any* $\delta > 0$*, there exists* $\eta > 0$ *such that*

$$\lim_{n\to\infty} P\left(Q_n(\widehat{v}) + \eta < \inf_{\|v - v_0\| \ge \delta} Q_n(v)\right) = 1.$$

It is proved in Appendix B. The theorem says that the local minimizer of the oracle space in Theorem 5.1 is also with high probability a global minimizer in $\mathbb{R}^{p+K}$. Note that by Theorems 5.1 and 5.2, the minimization in Eq. (5.1) enables one to recover the sparsity in the ultra high dimensional case since $q \ge p + K$, where q can be as large as e^{n^ϵ} for some $\epsilon > 0$. This is a bit different from Fan and Liao (2014) where there is no nonparametric function involved and $q = p$ (the number of IV is the same as that of regressors). Note that, given the consistency of the sparsity, the inference can be done in a similar way to Theorem 3.2.

6. Simulation experiments

In this section we investigate the performance of the proposed estimators in finite sample situations.

Table 1
Simulation results of Example 6.1, $q = p + K + \nu$.

	$\nu = 2$				$\nu = 4$		
n	300	600	1000	n	300	600	1000
$B_g(n)$	0.0046	−0.0040	−0.0026	$B_g(n)$	−0.0023	−0.0019	0.0006
$\pi_g(n)$	0.3533	0.1965	0.1948	$\pi_g(n)$	0.1660	0.1530	0.1520
$\Pi_g(n)$	0.3401	0.1700	0.1682	$\Pi_g(n)$	0.1356	0.1217	0.1176
$B_\alpha(n)$	0.0700	0.0410	0.0684	$B_\alpha(n)$	0.0281	0.0271	0.0501
$M_\alpha(n)$	0.0355	0.0282	0.0665	$M_\alpha(n)$	0.0259	0.0244	0.0319
	$\nu = 6$				$\nu = 8$		
n	300	600	1000	n	300	600	1000
$B_g(n)$	0.0023	0.0019	−0.0000	$B_g(n)$	0.0009	0.0011	−0.0000
$\pi_g(n)$	0.1544	0.1445	0.1444	$\pi_g(n)$	0.1482	0.1370	0.1359
$\Pi_g(n)$	0.1218	0.1092	0.1031	$\Pi_g(n)$	0.1176	0.1015	0.0945
$B_\alpha(n)$	0.0124	0.0267	0.0265	$B_\alpha(n)$	0.0078	0.0048	0.0250
$M_\alpha(n)$	0.0254	0.0154	0.0464	$M_\alpha(n)$	0.0117	0.0098	0.0306

Example 6.1. This experiment uses the partial linear model with endogenous covariates considered in the introduction. Let vector $X_i = (X_{1i}, X_{2i}^\top)^\top$, where X_{1i} takes values 1 and -1 with probability 1/2, respectively, $X_{2i} \sim N(0, \Sigma_{p-1})$, where $\Sigma_{p-1} = (\sigma_{i,j})_{(p-1)\times(p-1)}$ with $\sigma_{i,i} = 1$, $\sigma_{i,j} = 0.3$ for $|i-j| = 1$ and $\sigma_{i,j} = 0$ for $|i-j| > 1$. Here, the first component of X_i is a discrete variable with which we intend to show that our theoretical results do not confine application to continuous variables only. Let Z_i be uniformly distributed on $(0, 1)$.

Suppose that $\mathbb{E}[Y_i - \alpha^\top X_i - g(Z_i)|W_i] = 0$ with $W_i = Z_i$, and $g(\cdot) \in L^2[0,1] = \{u(r) : \int_0^1 u^2(r)dr < \infty\}$. Let $\varphi_0(r) \equiv 1$, and for $j \geq 1$, $\varphi_j(r) = \sqrt{2}\cos(\pi j r)$. Then, $\{\varphi_j(r)\}$ is an orthonormal basis in the Hilbert space $L^2[0,1]$. In the experiment, put $\alpha = (0.4, 0.1, 0, \ldots, 0)^\top \in \mathbb{R}^p$ and $g(z) = z^2 + \sin(z)$.

Denote $m(V_i, \alpha^\top X_i, g(Z_i)) = (Y_i - \alpha^\top X_i - g(Z_i))\Phi_q(Z_i)$ where $V_i = (Y_i, W_i)$, $W_i = Z_i$ and $\Phi_q(\cdot) = (\varphi_0(\cdot), \ldots, \varphi_{q-1}(\cdot))^\top$. We have $\mathbb{E}[m(V_i, \alpha^\top X_i, g(Z_i))] = 0$ for $i = 1, \ldots, n$.

According to the estimation procedure in Section 2, define $(\widehat{\alpha}, \widehat{\beta}) = \underset{\mathbf{a}\in\mathbb{R}^p, \mathbf{b}\in\mathbb{R}^K}{\operatorname{argmin}} \|M_n(\mathbf{a}, \mathbf{b})\|^2$, where $M_n(\mathbf{a}, \mathbf{b}) = \frac{1}{\sqrt{q}}\frac{1}{n}\sum_{i=1}^n m(V_i, \mathbf{a}^\top X_i, \mathbf{b}^\top \Phi_K(Z_i))$. Thus, $\widehat{\alpha}$ and $\widehat{g}(\cdot) := \widehat{\beta}^\top \Phi_K(\cdot)$ are the estimates of $(\alpha, g(\cdot))$.

For $n = 200, 500$ and 1000, let $K = [C_1 n^{\tau_1}]$ with $C_1 = 1$ and $\tau_1 = 1/4$, and $p = [C_2 n^{\tau_2}]$ with $C_2 = 1$ and $\tau_2 = 1/5$. Also, let $q = p + K + \nu$ ($\nu \geq 0$ specified in the sequel) satisfy Assumption 3.1. The replication number of the experiment is $M = 1000$. We shall report the bias (denoted by $B_g(n)$), standard deviation (denoted by $\pi_g(n)$) and RMSE (denoted by $\Pi_g(n)$) of the estimate of the g function, that is,

$$B_g(n) := \frac{1}{Mn}\sum_{\ell=1}^M\sum_{i=1}^n [\widehat{g}^\ell(Z_i) - g^\ell(Z_i)], \quad \pi_g(n) := \left(\frac{1}{Mn}\sum_{\ell=1}^M\sum_{i=1}^n [\widehat{g}^\ell(Z_i) - \overline{\widehat{g}}(Z_i)]^2\right)^{1/2},$$

$$\Pi_g(n) := \left(\frac{1}{Mn}\sum_{\ell=1}^M\sum_{i=1}^n [\widehat{g}^\ell(Z_i) - g^\ell(Z_i)]^2\right)^{1/2},$$

where the superscript ℓ indicates the ℓ-th replication, $\overline{\widehat{g}}(\cdot)$ is the average of $\widehat{g}^\ell(\cdot)$ over Monte Carlo replications $\ell = 1, \ldots, M$, and $g^\ell(\cdot)$ means the value of g in the ℓ-th replication.

Regarding the parameter α, we report the following quantities, $B_\alpha(n) := \|\alpha - \overline{\widehat{\alpha}}\|$ and $M_\alpha(n) := \text{median}(\|\alpha - \widehat{\alpha}\|)$, where $\overline{\widehat{\alpha}}$ is the average of $\widehat{\alpha}^\ell$ and median$(\cdots)$ is the median of the sequence over Monte Carlo replications. Notice that, due to the divergence of the dimension, it might not make any sense to compare the estimated results for different sample sizes (see Table 1).

It can also be seen from Table 1 that all of the statistical quantities about the estimate of g are reasonably attenuated with the increase of both the sample size and ν that provides more information for the parameters being estimated. For the quantities about the estimate of α, we observe that they normally do not decrease with the sample size. This is because, as mentioned before, the dimension of α is increasing with the sample size; and hence it does not make sense to compare them among different sample sizes. However, we find that, given the sample size, both quantities related to the estimate of α decrease with the increase of ν that gives more moment restrictions.

This is understandable. Because the conditional moment $\mathbb{E}[Y_i - \alpha^\top X_i - g(Z_i)|Z_i]$ determines a function $U(z) := \mathbb{E}[Y_i - \alpha^\top X_i - g(Z_i)|Z_i = z]$ and $\{\varphi_j(z)\}$ is an orthonormal sequence in the space that contains $U(z)$, the greater the ν is, the more axes in the space we use to explain the unknown function $U(z)$.

Additionally, the involvement of the discrete variable X_{1i} does not affect the performance of all measures. This might suggest for the practitioner that in this setting discrete variables are as tractable as continuous variables.

Table 2
Simulation results of Example 6.2 ($n = 100$).

	$p = 8, K = 6, q = 100$				$p = 12, K = 6, q = 120$	
λ	0.4	0.2	0.08	λ	0.4	0.2	0.08
$\text{MSE}_S(\alpha)$	0.2017	0.2811	0.1915	$\text{MSE}_S(\alpha)$	0.3065	0.2322	0.1970
$\text{MSE}_S(\beta)$	0.1288	0.1009	0.0789	$\text{MSE}_S(\beta)$	0.1900	0.0837	0.0624
$\text{MSE}_N(\alpha)$	0.0001	0.0026	0.0031	$\text{MSE}_N(\alpha)$	0.0015	0.0039	0.0016
$\text{MSE}_N(\beta)$	0.0000	0.0004	0.0001	$\text{MSE}_N(\beta)$	0.0000	0.0000	0.0008
$\text{TP}_S(\alpha)$	4	4	4	$\text{TP}_S(\alpha)$	4	4	4
$\text{TP}_S(\beta)$	2	2	2	$\text{TP}_S(\beta)$	2	2	2
$\text{TP}_N(\alpha)$	3.48	3.24	3.55	$\text{TP}_N(\alpha)$	6.88	6.72	5.90
$\text{TP}_N(\beta)$	3.28	3.40	2.96	$\text{TP}_N(\beta)$	3.46	3.36	2.92

Example 6.2. This example is to verify the proposed schedule for variable selection and parameter estimation under sparsity studied in Section 5. The model is almost the same one in Example 6.1 but the conditional variables are different. Suppose that

$$\mathbb{E}[Y_i - \alpha^\top X_i - g(Z_i)|W_i] = 0$$

where $(\alpha_1, \ldots, \alpha_4) = (2, -4, 3, 5)$, $\alpha_j = 0$ for $5 \le j \le p$. Here, $W_i = (X_{1i}, X_{2i})^\top$ and $g(\cdot) \in L^2[0, 1]$. The conditional moment gives the function $H(W) \equiv 0$, where $H(W) = \mathbb{E}[Y_i - \alpha^\top X_i - g(Z_i)|W_i = W]$. Thus, the instrument variable should be $\Psi_q(W_i)$, a basis vector of bivariate functions.

The same basis as in Example 6.1 is used for the orthogonal expansion of $g(z)$, viz., $\varphi_0(r) \equiv 1$, and for $j \ge 1$, $\varphi_j(r) = \sqrt{2}\cos(\pi j r)$. Here, put $g(z) = 1 + \sqrt{2}\cos(\pi z)$. Thus, the expansion of $g(z)$ has coefficients $\beta_i = 1$, $i = 0, 1$, while $\beta_i = 0$ for all $i \ge 2$, implying the sparsity of the coefficient vector β (equivalently, the sparse nonparametric function $g(z)$).

Suppose that p-vector X_i are i.i.d. N(0, I_p) and Z_i are i.i.d. U(0, 1). Given the normal distribution of X_i, we use Hermite polynomial sequence to form $\Psi_q(W_i)$, that is, $\Psi_q(W_i) = (h_{j_1-1}(X_{1i})h_{j_2-1}(X_{2i}), j_1, j_2 = 1, \ldots, q_1)$, where $q_1 = [\sqrt{q+1}]$ and $\{h_j(\cdot)\}$ is the Hermite polynomial sequence. The rationale behind the formulation of $\Psi_q(w_1, w_2)$ is that the tensor product $\{h_{j_1}(w_1)h_{j_2}(w_2)\}$ is an orthogonal basis system to expand $H(w_1, w_2)$.

In the simulation, we use SCAD of Fan and Li (2001) with predetermined tuning parameters of λ as the penalty function. Therefore, the objective function is $\|M_n(v)\|^2 + \sum_{j=1}^{p+K} P_n(|v_j|)$, where $v = (\alpha^\top, \beta^\top)^\top$ a $(p + K)$-dimensional vector and $M_n(v) = \frac{1}{q_1 n}\sum_{i=1}^n (Y_i - \alpha^\top X_i - \beta^\top \Phi_K(Z_i))\Psi_q(W_i)$.

Four performance measures are reported. The first measure is the mean standard error (MSE_S) of the important regressors, that is, the average of $\|\widehat{\alpha}_S - \alpha_S\|$ and that of $\|\widehat{\beta}_S - \beta_S\|$ over Monte Carlo replications. The second measure is the mean standard error (MSE_N) of the unimportant regressors for α and β, respectively. The third measure, denoted by TP_S, is the number of correctly selected nonzero coefficients, and the fourth, TP_N, the number of correctly selected unimportant coefficients for α and β, respectively. The initial value for v in the simulation is taken as $(0, \ldots, 0)$. The results are reported in Table 2 with different parameters, and more results can be found in the supplemental document of this paper.

It can be seen from the tables that all MSE's perform reasonably and particularly those for α_N and β_N are really well. They also seem to be smaller when both n and q become larger. Although the dimensions of α and β increase and $q \ge n$, the scheme can always correctly choose all the important coefficients. This is perhaps because all important coefficients in absolute are significantly greater than zero, as suggested by the literature that we do not pursue here. By contrast, some unimportant coefficients may be chosen as important ones, implying the scheme may not necessarily lead to parsimonious models.

7. Empirical illustration

There are many papers dealing with the marginal treatment effect (MTE) of a selection process. For example, Carneiro et al. (2011, CHV, hereafter) study MTE for schooling, while most recently Su et al. (2018) study continuous MTE in nonseparable models. Economists would like to know, on average, how the marginal return to schooling changes as the number of years of education increases, and would also like to be able to evaluate policies that change the probability of attaining a certain level of schooling. Let Y_1 be the potential log wage if the individual were to attend college and Y_0 be the potential log wage if the individual were not to attend college. Define potential outcome equations: $Y_1 = \mu_1(X) + U_1$ and $Y_0 = \mu_0(X) + U_0$, where X is a vector of relevant variables, $\mu_1(x) = \mathbb{E}(Y_1|X = x)$ and $\mu_0(x) = \mathbb{E}(Y_0|X = x)$.

Then, a selection process can be described as follows: $S = 1$ if $I_S = \mu_S(Z) - V > 0$ and $S = 0$ otherwise. Here, I_S stands for the net benefit of attending college, $\mu_S(Z)$ is defined in CHV, in which Z is observable and V is unobservable, so that $S = 1$ means that the agent goes to college while $S = 0$ means that he/she does not. Let $Y = SY_1 + (1 - S)Y_0$ be the earnings of an individual.

CHV analyse the marginal treatment effect for schooling, defined by the derivative of $\mathbb{E}(Y|X = x, P(Z) = p)$ with respect to p, denoted by MTE(x, p). The dataset constructed by CHV is available at www.aeaweb.org/articles?id=10.1257/aer.101.

Table 3
Average marginal derivatives in decision model.

AFQT	0.2073
Mother's years of schooling	0.0400
Number of siblings	−0.0209
Urban residence at 14	0.0028
Permanent local log earnings of 17	−0.0265
Permanent state unemployment rate at 17	0.0013
Presence of a college at 14	0.0190
Local log earnings at 17	−0.0250
Local unemployment rate at 17	0.0092
Tuition in 4 year public college at 17	−0.0017

6.2754 which comes from the 1979 National Longitudinal Survey of Youth (NLSY79) including a well-known proxy for ability of earning that is thought of beyond schooling and work experience: the Armed Forces Qualification Test (AFQT). See CHV for further details and references.

We shall use exactly the same variables X and Z as in CHV but with our proposed methodology to estimate parameters and test hypotheses of interest.[1] Note that equation (9) of CHV implies that

$$Y = X^{\top}\delta_0 + P(Z)X^{\top}\theta_0 + g(P(Z)) + \varepsilon, \tag{7.1}$$

$$\Pr(S = 1|Z) = P(Z) = \Lambda(Z^{\top}\gamma_0), \qquad \mathbb{E}(\varepsilon|X, Z) = 0, \tag{7.2}$$

where $P(Z)$ stands for the probability of attending college for the individual with characteristic Z, which is specified in the form of $\Lambda(Z^{\top}\gamma_0)$. In this case, $\mathrm{MTE}(x, p) = x^{\top}\theta_0 + g'(p)$. Eqs. (7.1) and (7.2) motivate an alternative way to estimate MTE. Precisely, Eq. (7.2) implies

$$\mathbb{E}[\left(\mathbb{I}(S = 1) - \Lambda(Z^{\top}\gamma_0)\right)\Phi_q(Z)] = 0, \tag{7.3}$$

where $\Lambda(z) = \exp(z)/[1 + \exp(z)]$ and $\Phi_q(\cdot)$ is a q -vector consisting of basis functions.

Note that in CHV the vector Z has dimension 34 which is relatively large. Hence, our theoretical result in Section 5 enables us to estimate γ_0 utilizing the moment condition (7.3) coupled with a penalty function (we use SCAD).

With $\widehat{\gamma}$ at hand, we first calculate the average derivative of each variable in the choice model (7.1), that is, for each individual we compute the effect of increasing each variable by one unit (keeping all the others constant) on the probability of enrolling in college and then we average across all individuals. The results are reported in Table 3.

The marginal derivatives reflect the changes in probability of attending a college when some policy was implemented to increase the relevant variable by one unit. For example, the marginal derivative of "Permanent local log earnings of 17" , −0.0265, means that when the earnings increases 100 dollars, the probability on average of attending a college would decrease 2.65%. By contrast, this derivative in CHV is 0.1820, meaning that a 100 dollar increase in the labour market would result in an increase of 18.20% enrolling in a college. This seems contradictory with intuition.

Moreover, Eq. (7.2), along with $\widehat{\gamma}$, allows us to estimate θ_0 and $g(\cdot)$ by transforming it to unconditional moments. The estimation procedure and asymptotic theory for this semiparametric single-index structure has been established in Section 3.3. Since the function $g(\cdot)$ is defined on [0, 1], a power series $\{p^j, j \geq 1\}$ in $L^2[0, 1]$ is employed to approximate the unknown $g(\cdot)$, and the same procedure as in Example 6.1 gives $\widehat{\theta}$ and $\widehat{g}(p)$. Hence, we have the estimate of MTE, $\widehat{\mathrm{MTE}}(x, p) = x^{\top}\widehat{\theta} + \widehat{g}'(p)$, where $\widehat{\theta}$ is given in Table 4 and $\widehat{g}'(p) = 0.6462 - 0.3898p - 0.4470p^2$. The plot of $\widehat{\mathrm{MTE}}(x, p)$ with $x = \bar{X}$, along with the upper and lower 95% significance bounds, is given in Fig. 1. It can be seen that with the increase of the probability of attending college, the MTE decreases. The plot is quite similar to Figure 4 in CHV(p. 20).

For the implementation of the estimation above, we emphasize that in order to coincide with the theoretical procedure described in Section 3.3, we use a subsample with size 874 drawn randomly to estimate γ_0 to obtain $\widehat{\gamma}$, then the rest of the sample with size 873 is used to estimate θ_0 and $g(\cdot)$, obtaining $\widehat{\theta}$ and $\widehat{g}(p)$. The number of basis functions used is selected by the minimum MSE criterion over a candidate set. To have the standard deviations of the coefficients in $\widehat{\theta}$ and $\widehat{g}(p)$, a bootstrap method is employed with 250 replications. The standard deviations of the coefficients in $\widehat{g}(p)$ are 0.5319, 0.0919 and 0.0738, implying that the last two coefficients are significant at the 95% level.

Furthermore, with regard to testing whether $g(p)$ is a constant function, in CHV this test is implemented through specifying $g(p)$ as polynomials of order 2–5, respectively, and then test whether their coefficients are jointly zero. Nonetheless, we actually have done this in the estimate of $\widehat{g}(p)$ without any specification, because we treat $g(p)$ as a nonparametrically unknown function, and two coefficients in $\widehat{g}(p)$ are found to be significant. Therefore, this offers some

[1] The vector X consists of the year of mother's education, number of siblings, average of log earnings 1979–2000 in county of residence at 17, average of unemployment 1979–2000 in state of residence at 17, urban residence at 14, cohort dummies, years of experience in 1991, average of local log earnings in 1991, local unemployment in 1991, while Z contains some variables in X, as well as instruments, that is, presence of a College at Age 14 (Card 1993, Cameron and Taber 2004), local earnings at 17 (Cameron and Heckman 1998, Cameron and Taber 2004), local unemployment at 17 (Cameron and Heckman 1998), local tuition in public 4 year colleges at 17 (Kane and Rouse 1995). These papers in parentheses are such papers that previously used these instruments. See CHV for details and their explanation.

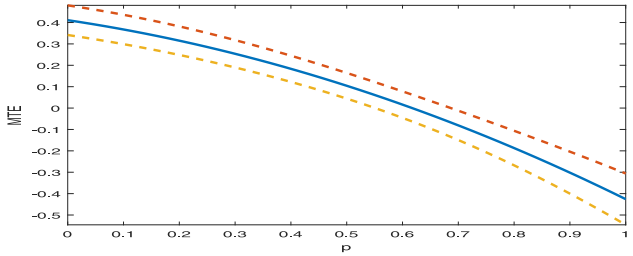

Fig. 1. Estimated MTE calculated at $x = \bar{X}$ and the 95% Confidence Interval.

Table 4
Estimated coefficients of θ_0 and $\widehat{g}(p)$ in MTE.

Estimated coefficients of θ_0							
−0.2852	−0.2089	0.2382	−0.1296	−0.3728	−0.0458	0.4915	0.8161
(0.2840)	(0.1530)	(0.1611)	(0.2420)	(0.1612)	(0.0108)	(0.3908)	(0.7419)
0.0454	0.1059	0.0115	−0.7552	1.1762	0.2706	0.3666	−1.1519
(0.0924)	(0.1372)	(0.0167)	(0.4263)	(0.6864)	(0.5630)	(0.3185)	(0.4768)
−0.2508	−0.0428	−0.9744	−0.2847	−1.3112	−0.0417		
(0.2811)	(0.0653)	(0.4925)	(0.3183)	(0.5518)	(0.0159)		
Estimated coefficients in $\widehat{g}(p)$							
	0.6462		−0.1949		−0.1490		
	(0.5319)		(0.0919)**		(0.0738)**		

** indicates that they are significant at the 95% level.

strong evidence to support a non-constant functional form for $g(p)$, which would be equivalent to rejecting the null hypothesis that the functional form of $g(p)$ is constant. In addition, some detailed justification about the nonlinearity of AFQT is available at Dong et al. (2018) for the interested reader.

8. Conclusion

We have provided estimation and inference tools for a class of high dimensional semiparametric moment restriction models based on the sieve GMM method and the penalized sieve GMM method. Our approach is based on simultaneous selection and estimation of the unknown quantities. The theoretical results are verified through finite sample experiments. We have found that the more the number of moment restrictions, the more accurate the estimates. In addition, in our empirical study we have also found our results to be more reasonable in some respects than those reported in the existing literature. The framework we have considered is quite general but can be generalized in a number of ways. First, we may allow explicitly for panel data and allow for weak dependent sampling schemes. Second we may allow for a large number of nonparametric functions to enter the moment condition provided they are each defined on low dimensional spaces. Another question of interest here is efficiency; Jankova and Geer (2018) develop some results about efficiency in their large linear model framework.

Acknowledgements

We thank Professor Xiaohong Chen for her insightful suggestions and for providing us with some relevant references. We also thank the audience of the seminar in Monash University, the Fifth China Meeting of Econometric Society 2018 in Shanghai and the 2019 Asia Meeting of the Econometric Society in Xiamen. The first author thanks the financial support from National Natural Science Foundation of China under grants Nos. 72073143 & 71671143. The second author is supported by the Australian Research Council Discovery Grants Program for its support under Grant numbers: DP150101012 & DP170104421.

Appendix A. Lemmas

This section gives all technical lemmas, additional assumptions and some notation used for the theoretical derivations, while the proofs of these lemmas are postponed to the supplementary material of the paper or the working paper version Dong et al. (2018).

Lemma A.1. *Under Assumptions 2.1–2.2 and 3.1–3.3, we have*

1. $\|M_n(\alpha,\beta)\|^2 = O_P(\|\gamma_K\|^2) + O_P(n^{-1})$.

2. Given $B_{1n}^2 + B_{2n}^2 = o(n)$, $\sup_{\substack{\|\mathbf{a}\|\le B_{1n},\|\mathbf{b}\|\le B_{2n}\\ \|(\mathbf{a}-\alpha,\mathbf{b}-\beta)\|>\delta}} \|M_n(\mathbf{a},\mathbf{b})\|^{-2} = O_P(1/\epsilon_n)$ *for each* $\delta > 0$, *when n is large and where* $\epsilon_n \equiv \epsilon_n(\delta)$ *stipulated in Assumption 3.2.*

Denote $m(v,u,w) = (m_1(v,u,w),\ldots,m_q(v,u,w))^\intercal$. To investigate the asymptotics, denote the Score and Hessian functions of $\|M_n(\mathbf{a},\mathbf{b})\|^2$ as

$$S_n(\mathbf{a},\mathbf{b}) = \begin{pmatrix} S_{1n}(\mathbf{a},\mathbf{b}) \\ S_{2n}(\mathbf{a},\mathbf{b}) \end{pmatrix} := \begin{pmatrix} \frac{\partial}{\partial \mathbf{a}} \\ \frac{\partial}{\partial \mathbf{b}} \end{pmatrix} \|M_n(\mathbf{a},\mathbf{b})\|^2,$$

$$H_n(\mathbf{a},\mathbf{b}) = \begin{pmatrix} H_{11}(\mathbf{a},\mathbf{b}) & H_{12}(\mathbf{a},\mathbf{b}) \\ H_{21}(\mathbf{a},\mathbf{b}) & H_{22}(\mathbf{a},\mathbf{b}) \end{pmatrix} := \begin{pmatrix} \frac{\partial^2}{\partial \mathbf{a}\partial \mathbf{a}^\intercal} & \frac{\partial^2}{\partial \mathbf{a}\partial \mathbf{b}^\intercal} \\ \frac{\partial^2}{\partial \mathbf{b}\partial \mathbf{a}^\intercal} & \frac{\partial^2}{\partial \mathbf{b}\partial \mathbf{b}^\intercal} \end{pmatrix} \|M_n(\mathbf{a},\mathbf{b})\|^2.$$

Moreover, define

$$h_n(\alpha,g) = \frac{1}{q}\Psi_n\Psi_n^\intercal, \quad \text{and} \quad s_n(\alpha,g) = \frac{1}{q}\Psi_n\frac{1}{n}\sum_{i=1}^n m(V_i,\alpha^\intercal X_i, g(Z_i)), \tag{A.1}$$

where

$$\Psi_n = \mathbb{E}\begin{pmatrix} \frac{\partial}{\partial u} m(V,\alpha^\intercal X, g(Z))^\intercal \otimes X \\ \frac{\partial}{\partial w} m(V,\alpha^\intercal X, g(Z))^\intercal \otimes \Phi_K(Z) \end{pmatrix}_{(p+K)\times q}.$$

Lemma A.2. *Let Assumptions 2.1–2.2 and 3.1, 3.3–3.7 hold. Then, (1)* $H_n(\alpha,\beta)$ *is asymptotically positive definite with probability one; (2)* $\|H_n(\alpha,\beta) - h_n(\alpha,g)\| = o_P(1)$ *as* $n\to\infty$.

Lemma A.3. *Under Assumptions 2.1–2.2, 3.1, 3.3–3.7, as* $n\to\infty$, $\|S_n(\alpha,\beta) - s_n(\alpha,g)\| = o_P(1)$.

Lemma A.4. *Suppose that* $\theta = (\theta_1,\ldots,\theta_d)', Z = (Z_1,\ldots,Z_d)' \in \mathbb{R}^d$ *and* $\|\theta\| = 1$. *Then for any* $m \ge 1$ *and Hermite polynomial* H_m,

$$H_m(\theta^\intercal Z) = \sum_{|u|=m}\binom{m}{u}\prod_{j=1}^d H_{u_j}(Z_j)\prod_{j=1}^d \theta_j^{u_j},$$

where $u = (u_1,\ldots,u_d)$ *is multi-index,* $|u| = u_1+\cdots+u_d$ *and* $\binom{m}{u} = \frac{m!}{\prod_{j=1}^d u_j!}$.

Assumption 2.1*. Let $\mathbb{Z}$ be the support of $\theta_0^\intercal Z_i$. Suppose that $\{\varphi_j(\cdot)\}$ is a complete orthonormal function sequence in $L^2(\mathbb{Z},\pi(\cdot))$, that is, $\langle\varphi_i(\cdot),\varphi_j(\cdot)\rangle = \delta_{ij}$ the Kronecker delta.

Assumption 3.1*. Assumptions (a), (c) and (d) in Assumption 3.1 remain the same but (b) is replaced by: (b*) for the density $f_\theta(z)$ of $\theta^\intercal Z_1$, there exist two constants $0 < c < C < \infty$ such that $c\pi(z) \le f_\theta(z) \le C\pi(z)$ on the support $\mathbb{Z}$ of $\theta^\intercal Z_1$ for θ in some neighbourhood of θ_0.

Assumption 3.2*. Suppose that there is a unique function $g(\cdot)\in L^2(\mathbb{Z},\pi)$ and for each n there is a unique vector $\alpha\in\mathbb{R}^p$ such that model (2.6) is satisfied. In other words, for any $\delta > 0$, there is some $\epsilon_n > 0$ such that

$$\inf_{\substack{(\mathbf{a},f)\in\Theta\\ \|(\mathbf{a}-\alpha,f-g)\|\ge\delta}} q^{-1}\|\mathbb{E}m(V_i,\mathbf{a}^\intercal X_i, f(\theta_0^\intercal Z_i))\|^2 > \epsilon_n,$$

and possibly $\epsilon_n\to 0$ as $n\to\infty$ but with a rate slower than $\max(\|\gamma_K(\cdot)\|, n^{-1})$.

Assumption 3.3*. Suppose that for each n, there is a measurable positive function $A(V,X,Z)$ such that

$$q^{-1/2}\|m(V,\mathbf{a}_1^\intercal X, f_1(\theta^\intercal Z)) - m(V,\mathbf{a}_2^\intercal X, f_2(\theta^\intercal Z))\| \le A(V,X,Z)[\|\mathbf{a}_1-\mathbf{a}_2\| + |f_1(\theta^\intercal Z) - f_2(\theta^\intercal Z)|]$$

for any $(a_1,f_1),(a_2,f_2)\in\Theta$ and for θ in some neighbourhood of θ_0, where (V,X,Z) is any realization of (V_i,X_i,Z_i) and the function A satisfies that $\mathbb{E}[A^2(V,X,Z)] < \infty$ uniformly in n.

Assumption 3.5*. All statements in Assumption 3.5 are true when Z_1 is replaced by $\theta_0^\intercal Z_1$.

Assumption 3.7*. The partial derivatives of $m(v,u,w)$ satisfy those inequalities in Assumption 3.7 when Z is replaced by $\theta_0^\intercal Z$.

Similar to $H_n(\mathbf{a},\mathbf{b})$, we define block matrix $\widetilde{H}_n(\mathbf{a},\mathbf{b}) = (\widetilde{H}_{ij}(\mathbf{a},\mathbf{b}))_{i,j=1,2}$ as the Hessian matrix of $\|\widetilde{M}_n(\mathbf{a},\mathbf{b})\|^2$. Meanwhile, define $\widetilde{\Psi}_n$ and $\widetilde{h}_n(\alpha,g)$ in the same way as Ψ_n and $h_n(\alpha,g)$ given by (A.1) with Z being replaced by $\theta_0^\intercal Z$.

Lemma A.5. *Under Assumptions 2.1*–2.2, 3.1*–3.3*, we have*
1. $\|\widetilde{M}_n(\alpha,\beta)\|^2 = O_P(\|\gamma_K\|^2) + O_P(n^{-1})$.
2. Given $B_{1n}^2 + B_{2n}^2 = o(n)$, $\sup_{\substack{\|\mathbf{a}\|\le B_{1n},\|\mathbf{b}\|\le B_2 \\ \|(\mathbf{a}-\alpha,\mathbf{b}-\beta)\|>\delta}} \|\widetilde{M}_n(\mathbf{a},\mathbf{b})\|^{-2} = O_P(1/\epsilon_n)$ *for each* $\delta > 0$ *when* n *is large, where* ϵ_n *is given by Assumption 3.2*.*

Lemma A.6. *Let Assumptions 2.1*–2.2, and 3.1*, 3.3*, 3.5*, 3.6, and 3.7* hold. Then (1)* $\widetilde{H}_n(\alpha,\beta)$ *is asymptotically positive definite with probability one; and (2) we have* $\|\widetilde{H}_n(\alpha,\beta) - \widetilde{h}_n(\alpha,g)\| = o_P(1)$ *as* $n \to \infty$.

Similarly to $S_n(\mathbf{a},\mathbf{b})$, we define $\widetilde{S}_n(\mathbf{a},\mathbf{b}) = (\widetilde{S}_{1n}(\mathbf{a},\mathbf{b})^\intercal, \widetilde{S}_{2n}(\mathbf{a},\mathbf{b})^\intercal)^\intercal$ as the Score function of $\widetilde{M}_n(\mathbf{a},\mathbf{b})$ and define $\widetilde{s}_n(\alpha,g) := (\widetilde{s}_{1n}(\alpha,g)^\intercal, \widetilde{s}_{2n}(\alpha,g)^\intercal)^\intercal$, which is the same as $s_n(\alpha,g)$ but with Z_i being replaced by $\theta_0^\intercal Z_i$, i.e.

$$\widetilde{s}_n(\alpha,g) = (\widetilde{s}_{1n}(\alpha,g)^\intercal, \widetilde{s}_{2n}(\alpha,g)^\intercal)^\intercal = \frac{1}{q}\widetilde{\Psi}_n \frac{1}{n}\sum_{i=1}^n m(V_i, \alpha^\intercal X_i, g(\theta_0^\intercal Z_i)). \tag{A.2}$$

Lemma A.7. *Under the same conditions as Lemma A.6,* $\|\widetilde{S}_n(\alpha,\beta) - \widetilde{s}_n(\alpha,g)\| = o_P(1)$ *as* $n \to \infty$.

Lemma A.8. *Let Assumptions 5.1–5.2 hold. Suppose that (i) There exists a positive sequence* $a_n = o(d_n)$ *such that* $\|S_{nT}(v_{0S})\| = O_P(a_n)$*; (ii) For any* $\epsilon > 0$*, there exists a constant* $C = C(\epsilon) > 0$ *such that for all large* n, $P(\lambda_{\min}(H_{nT}(v_{0S})) > C) > 1-\epsilon$*; (iii) For any* $\epsilon > 0$, $\delta > 0$ *and any nonnegative sequence* $\eta_n = o(d_n)$*, there is an* $N > 0$ *such that whenever* $n > N$,

$$P\left(\sup_{\|v_T - v_0\| \le \eta_n} \|H_{nT}(v_T) - H_{nT}(v_0)\| \le \delta\right) > 1-\epsilon.$$

Then there exists a local minimizer $\widehat{v} \in \mathcal{V}$ *of* $Q_n(v_T) = \|M_n(v_T)\|^2 + \sum_{j\in T} P_n(|v_j|)$*, such that* $\|\widehat{v} - v_0\| = O_P(a_n + \sqrt{t_n}P_n'(d_n))$*. Moreover, for any arbitrary* $\epsilon > 0$*, the local minimizer* $\widehat{v}$ *is strict with probability at least* $1-\epsilon$ *for all large* n.

It is worth noting that we show in Appendix C $\|S_{nT}(v_{0S})\| = O_P(\sqrt{t_n\ \log(q)/n})$ under an additional condition stated below, and therefore we have $\|\widehat{v} - v_0\| = O_P(\sqrt{t_n\ \log(q)/n} + \sqrt{t_n}P_n'(d_n))$.

The oracle consistency in Lemma A.8 is derived based on the knowledge of T, the support of v_0. To make the result useful, it is desirable to show that the local minimizer of Q_n restricted on $\mathcal{V}$ is also a minimizer of Q_n on $\mathbb{R}^{p+K}$.

Lemma A.9. *Let the conditions in Lemma A.8 hold. Suppose that with probability approaching one, for* $\widehat{v} \in \mathcal{V}$ *in Lemma A.8, there exists a neighbourhood* $O_1 \subset \mathbb{R}^{p+K}$ *of* $\widehat{v}$ *such that for all* $v \in O_1$ *but* $v \notin \mathcal{V}$*, we have*

$$\|M_n(v_T)\|^2 - \|M_n(v)\|^2 < \sum_{j\notin T} P_n(|v_j|). \tag{A.3}$$

Then, (i) With probability close to unity arbitrarily, the $\widehat{v} \in \mathcal{V}$ *is a local minimizer in* $\mathbb{R}^{p+K}$ *of* $Q_n(v) = \|M_n(v)\|^2 + \sum_{j=1}^{p+K} P_n(|v_j|)$*; (ii) For* $\forall \epsilon > 0$*, the local minimizer* $\widehat{v}$ *is strict with probability at least* $1-\epsilon$ *for all large* n.

Appendix B. Proofs of the main results

Proof of Theorem 3.1. In Lemma A.1, we have shown (i) $\|M_n(\alpha,\beta)\|^2 = O_P(u_n)$ with $u_n = \max(\|\gamma_K\|^2, n^{-1})$; and (ii) $\sup_{\substack{\|\mathbf{a}\|\le B_{1n},\|\mathbf{b}\|\le B_{2n} \\ \|(\mathbf{a}-\alpha,\mathbf{b}-\beta)\|>\delta}} \|M_n(\mathbf{a},\mathbf{b})\|^{-2} = O_P(1/\epsilon_n)$ for each $\delta > 0$.

Fix $\varepsilon > 0$ and $\delta > 0$. Assertion (ii) means that there exists a large but fixed M for which

$$\limsup P\left(\epsilon_n \sup_{\substack{\|\mathbf{a}\|\le B_{1n},\|\mathbf{b}\|\le B_{2n} \\ \|(\mathbf{a}-\alpha,\mathbf{b}-\beta)\|>\delta}} \|M_n(\mathbf{a},\mathbf{b})\|^{-2} > M\right) < \varepsilon.$$

Meanwhile, by the definition of the estimator and (i), we have

$$\|M_n(\widehat{\alpha},\widehat{\beta})\|^2 = \inf_{\|\mathbf{a}\|\le B_{1n},\|\mathbf{b}\|\le B_{2n}} \|M_n(\mathbf{a},\mathbf{b})\|^2 \le \|M_n(\alpha,\beta)\|^2 = O_P(u_n),$$

which gives $\epsilon_n\|M_n(\widehat{\alpha},\widehat{\beta})\|^{-2} = O_P(\epsilon_n/u_n) \to_P \infty$ by Assumption 3.2 and hence

$$P\left(\epsilon_n\|M_n(\widehat{\alpha},\widehat{\beta})\|^{-2} > M\right) \to 1.$$

It follows that, with probability of at least $1-2\varepsilon$ for all n large enough,

$$\epsilon_n\|M_n(\widehat{\alpha},\widehat{\beta})\|^{-2} > M \geq \epsilon_n \sup_{\substack{\|\mathbf{a}\|\leq B_{1n},\|\mathbf{b}\|\leq B_{2n}\\ \|(\mathbf{a}-\alpha,\mathbf{b}-\beta)\|>\delta}} \|M_n(\mathbf{a},\mathbf{b})\|^{-2}.$$

Hence, the inclusion $(\widehat{\alpha},\widehat{\beta}) \in \{(\mathbf{a},\mathbf{b}) : \|\mathbf{a}\| \leq B_{1n}, \|\mathbf{b}\| \leq B_{2n}, \|(\mathbf{a}-\alpha,\mathbf{b}-\beta)\| > \delta\}$ holds with probability at most 2ε, $P\left(\|(\widehat{\alpha}-\alpha,\widehat{\beta}-\beta)\| > \delta\right) \leq 2\varepsilon$. As ε and δ are arbitrarily chosen, we then have $\|(\widehat{\alpha}-\alpha,\widehat{\beta}-\beta)\| \to_P 0$. Notice further that

$$\begin{aligned}
&\|(\widehat{\alpha}-\alpha,\widehat{g}(z)-g(z))\|^2 = \|\widehat{\alpha}-\alpha\|^2 + \int [\widehat{g}(z)-g(z)]^2\pi(z)dz\\
=&\|\widehat{\alpha}-\alpha\|^2 + \int [(\widehat{\beta}-\beta)^\intercal \Phi_K(z)-\gamma_K(z)]^2\pi(z)dz = \|\widehat{\alpha}-\alpha\|^2 + \|\widehat{\beta}-\beta\|^2 + \|\gamma_K(z)\|^2\\
=&\|(\widehat{\alpha}-\alpha,\widehat{\beta}-\beta)\|^2 + \|\gamma_K(z)\|^2 \to_P 0,
\end{aligned}$$

as $n, K \to \infty$, by the orthogonality of the basis sequence, which then completes the proof. □

Proof of Theorem 3.2. Notice that the conditions of the theorem imply the consistency of the estimator that is used in the sequel. By the first order condition $S_n(\widehat{\alpha},\widehat{\beta}) = 0$, consistency and Taylor expansion, we have expansion

$$\begin{aligned}
0 = S_n(\widehat{\alpha},\widehat{\beta}) =& S_n(\alpha,\beta) + H_n(\bar{\alpha},\bar{\beta})\begin{pmatrix}\widehat{\alpha}-\alpha\\ \widehat{\beta}-\beta\end{pmatrix}\\
=& S_n(\alpha,\beta) + H_n(\alpha,\beta)\begin{pmatrix}\widehat{\alpha}-\alpha\\ \widehat{\beta}-\beta\end{pmatrix} + [H_n(\bar{\alpha},\bar{\beta}) - H_n(\alpha,\beta)]\begin{pmatrix}\widehat{\alpha}-\alpha\\ \widehat{\beta}-\beta\end{pmatrix},
\end{aligned}$$

where $(\bar{\alpha},\bar{\beta})$ is some point on the joint line between $(\widehat{\alpha},\widehat{\beta})$ and (α,β). Notice that the last term is of smaller order in probability comparing to the second term. Indeed, by the Lipschitz condition in Assumption 3.4, the last term in norm is bounded by $O_P(p+K)[\|\widehat{\alpha}-\alpha\| + \|\widehat{\beta}-\beta\|]^{1+\tau}$, while the second term is $O_P(p+K)[\|\widehat{\alpha}-\alpha\| + \|\widehat{\beta}-\beta\|]$. Thus, we may write

$$0 = S_n(\widehat{\alpha},\widehat{\beta}) = S_n(\alpha,\beta) + H_n(\alpha,\beta)\begin{pmatrix}\widehat{\alpha}-\alpha\\ \widehat{\beta}-\beta\end{pmatrix}(1+o_P(1)),$$

in view of the consistency and for simplicity we shall ignore the term $o_P(1)$ in the sequel. As shown in Lemmas A.2–A.3, under Assumptions 2.1–2.2, 3.1, 3.3 and 3.5–3.7 in Section 3, $H_n(\alpha,\beta)$ is asymptotically positive definite, and $H_n(\alpha,\beta)$ and $S_n(\alpha,\beta)$ are approximated by $h_n(\alpha,g)$ and $s_n(\alpha,g)$ (defined in (A.1)), respectively, that is, $\|H_n(\alpha,\beta) - h_n(\alpha,g)\| = o_P(1)$ and $\|S_n(\alpha,\beta) - s_n(\alpha,g)\| = o_P(1)$. Hence, for large n,

$$\begin{pmatrix}\widehat{\alpha}-\alpha\\ \widehat{\beta}-\beta\end{pmatrix} = -H_n(\alpha,\beta)^{-1}S_n(\alpha,\beta) = -h_n(\alpha,g)^{-1}s_n(\alpha,g)(1+o_P(1)). \tag{B.1}$$

Note that $\mathscr{L}(\widehat{\alpha}) - \mathscr{L}(\alpha) = \partial\mathscr{L}(\alpha)^\intercal(\widehat{\alpha}-\alpha) + (\widehat{\alpha}-\alpha)^\intercal \otimes (\partial^2\mathscr{L}_1(\bar{\alpha}),\ldots,\partial^2\mathscr{L}_r(\bar{\alpha}))^\intercal \otimes (\widehat{\alpha}-\alpha)$ where $\mathscr{L}_j$ is the component of the transformation $\mathscr{L}$ and $\bar{\alpha}$ is on the segment joining α and $\widehat{\alpha}$, and by Assumption 3.8 the second term is negligible; $\widehat{g}(z) - g(z) = \Phi_K(z)^\intercal(\widehat{\beta}-\beta) - \gamma_K(z)$. By the linearity of Fréchet derivative and ignoring the higher order term in the definition of Fréchet derivative, we have

$$\begin{aligned}
&\begin{pmatrix}\mathscr{L}(\widehat{\alpha})-\mathscr{L}(\alpha)\\ \mathscr{F}(\widehat{g})-\mathscr{F}(g)\end{pmatrix} = \begin{pmatrix}\partial\mathscr{L}(\alpha)^\intercal(\widehat{\alpha}-\alpha)\\ \mathscr{F}'(g)(\widehat{g}(z)-g(z))\end{pmatrix} = \begin{pmatrix}\partial\mathscr{L}(\alpha)^\intercal & 0\\ 0 & \mathscr{F}'(g)\Phi_K(z)^\intercal\end{pmatrix}\begin{pmatrix}\widehat{\alpha}-\alpha\\ \widehat{\beta}-\beta\end{pmatrix}\\
&\quad - \begin{pmatrix}0\\ \mathscr{F}'(g)\gamma_K(z)\end{pmatrix} = \Gamma_n h_n(\alpha,g)^{-1}s_n(\alpha,g) - \begin{pmatrix}0\\ \mathscr{F}'(g)\gamma_K(z)\end{pmatrix} := \Lambda_{1n} + \Lambda_{2n}, \quad \text{say.}
\end{aligned}$$

Recall $h_n(\alpha,g) = \frac{1}{q}\Psi_n\Psi_n^\intercal$ and $s_n(\alpha,g) = \frac{1}{q}\Psi_n\frac{1}{n}\sum_{i=1}^n m(V_i,\alpha^\intercal X_i, g(Z_i))$ by (A.1).
Hence, $\Lambda_{1n} = \frac{1}{n}\Gamma_n(\Psi_n\Psi_n^\intercal)^{-1}\Psi_n\sum_{i=1}^n m(V_i,\alpha^\intercal X_i,g(Z_i))$. Then, the covariance matrix of $\sqrt{n}\Lambda_{1n}$ is

$$\Sigma_n^2 := \Gamma_n(\Psi_n\Psi_n^\intercal)^{-1}\Psi_n\Xi_n\Psi_n^\intercal(\Psi_n\Psi_n^\intercal)^{-1}\Gamma_n^\intercal,$$

in which $\Xi_n := \mathbb{E}[m(V_1,\alpha^\intercal X_1,g(Z_1))m(V_1,\alpha^\intercal X_1,g(Z_1))^\intercal]$. It follows from the standard central limit theorem (i.i.d. innovations) that $\sqrt{n}\Sigma_n^{-1}\Lambda_{1n} \to_D N(0,I_{r+s})$ as $n\to\infty$. Then the assertion follows because of $\sqrt{n}\Sigma_n^{-1}(\mathbf{0}_r^\intercal, \mathscr{F}'(g)\gamma_K(z)^\intercal)^\intercal = o(1)$, yielding $\sqrt{n}\Lambda_{2n} = o(1)$. □

Proof of Proposition 3.1. The assertions (1) and (2) can be shown similarly to Lemmas 3.4 and 3.5 in Pakes and Pollard (1989). For brevity we omit the proof. For (3), factor $\Xi_n = C_nC_n^\intercal$ and denote $\Omega_n = [\Psi_nW\Psi_n^\intercal]^{-1}\Psi_nWC_n$ and $T_n = \Omega_n - [\Psi_n\Xi_n^{-1}\Psi_n^\intercal]^{-1}\Psi_n(C_n^{-1})^\intercal$. It follows that $T_nT_n^\intercal = \Omega_n\Omega_n^\intercal - [\Psi_n\Xi_n^{-1}\Psi_n^\intercal]^{-1}$, from which

$$\Gamma_n[\Psi_nW\Psi_n^\intercal]^{-1}\Psi_nW\Xi_nW\Psi_n^\intercal[\Psi_nW\Psi_n^\intercal]^{-1}\Gamma_n^\intercal \geq \Gamma_n[\Psi_n\Xi_n^{-1}\Psi_n^\intercal]^{-1}\Gamma_n^\intercal,$$

for all W satisfying the conditions, in view of the nonnegative definiteness of $T_nT_n^\intercal$. □

Proof of Theorem 4.1. In view of the condition about m and the i.i.d. observations, by the standard central limit theorem

$$\left(\sum_{i=1}^{n}[\kappa^{\intercal}m(V_i,\alpha^{\intercal}X_i,g(Z_i))]^2\right)^{-1/2}\sum_{i=1}^{n}\kappa^{\intercal}m(V_i,\alpha^{\intercal}X_i,g(Z_i))\to_D N(0,1),$$

as $n\to\infty$ for any $\kappa\in\mathbb{R}^q$ such that $\|\kappa\|=1$.

Thus, the result follows immediately if we show

$$L_n(\widehat{\alpha},\widehat{\beta};\kappa)=\left(\sum_{i=1}^{n}[\kappa^{\intercal}m(V_i,\alpha^{\intercal}X_i,g(Z_i))]^2\right)^{-1/2}\sum_{i=1}^{n}\kappa^{\intercal}m(V_i,\alpha^{\intercal}X_i,g(Z_i))+o_P(1).$$

Towards this end, we shall show

(1). $\frac{1}{n}D_n(\widehat{\alpha},\widehat{\beta};\kappa)^2-\frac{1}{n}\sum_{i=1}^{n}[\kappa^{\intercal}m(V_i,\alpha^{\intercal}X_i,g(Z_i))]^2=o_P(1)$; and

(2). $\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\kappa^{\intercal}m(V_i,\widehat{\alpha}^{\intercal}X_i,\widehat{\beta}^{\intercal}\Phi_K(Z_i))-\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\kappa^{\intercal}m(V_i,\alpha^{\intercal}X_i,g(Z_i))=o_P(1)$.

However, the proof is lengthy, so we refer the interested reader to the working paper version of Dong et al. (2018, p43-45). This finishes the proof. □

Proof of Theorem 4.2. Because for any $(\mathbf{a},\mathbf{b})$ and κ with $\|\kappa\|=1$,

$$\begin{aligned}\frac{1}{\sqrt{n}}D_n(\mathbf{a},\mathbf{b};\kappa)&=\left(\mathbb{E}[\kappa^{\intercal}m(V_1,\mathbf{a}^{\intercal}X_1,\mathbf{b}^{\intercal}\Phi_K(Z_1))]^2\right)^{1/2}+o_P(1)\\&=\left(\kappa^{\intercal}\mathbb{E}[m(V_1,\mathbf{a}^{\intercal}X_1,\mathbf{b}^{\intercal}\Phi_K(Z_1))m(V_1,\mathbf{a}^{\intercal}X_1,\mathbf{b}^{\intercal}\Phi_K(Z_1))^{\intercal}]\kappa\right)^{1/2}+o_P(1),\end{aligned}$$

which is bounded away from zero and infinity in probability, it suffices to show that there is some κ^* with $\|\kappa^*\|=1$ such that

$$\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\kappa^{*\intercal}m(V_i,\mathbf{a}^{\intercal}X_i,\mathbf{b}^{\intercal}\Phi_K(Z_i))\to_P\infty$$

as $n\to\infty$ for any $(\mathbf{a},\mathbf{b})\in\mathbb{R}^{p+K}$. Note by the Law of Large Numbers that

$$\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\kappa^{\intercal}m(V_i,\mathbf{a}^{\intercal}X_i,\mathbf{b}^{\intercal}\Phi_K(Z_i))=\sqrt{n}\{\mathbb{E}[\kappa^{\intercal}m(V_i,\mathbf{a}^{\intercal}X_i,\mathbf{b}^{\intercal}\Phi_K(Z_i))]+o_P(1)\}.$$

Let $\kappa^*=\mathbb{E}[m(V_i,\mathbf{a}^{\intercal}X_i,\mathbf{b}^{\intercal}\Phi_K(Z_i))]/\|\mathbb{E}[m(V_i,\mathbf{a}^{\intercal}X_i,\mathbf{b}^{\intercal}\Phi_K(Z_i))]\|$. Then,

$$\begin{aligned}&\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\kappa^{*\intercal}m(V_i,\mathbf{a}^{\intercal}X_i,\mathbf{b}^{\intercal}\Phi_K(Z_i))=\sqrt{n}\{\|\mathbb{E}[m(V_i,\mathbf{a}^{\intercal}X_i,\mathbf{b}^{\intercal}\Phi_K(Z_i))]\|+o_P(1)\}\\&\geq\sqrt{n}\{\inf_{(\mathbf{a},h)\in\Theta}\|\mathbb{E}[m(V_i,\mathbf{a}^{\intercal}X_i,h(Z_i))]\|+o_P(1)\}\geq\sqrt{n}(\delta_n+o_P(1))\to_P\infty,\end{aligned}$$

as $n\to\infty$, which finishes the proof. □

Proof of Theorem 5.1. (i) and (ii). As shown in Lemma A.9, if $Q_n(v)$ has a local minimizer $\widehat{v}=(\widehat{v}_S^{\intercal},\widehat{v}_N^{\intercal})^{\intercal}$, then $\widehat{v}_N=0$ with probability arbitrarily close to one for large n, which implies the assertion (i) and $P(\widehat{T}\subset T)\to 1$.

On the other hand,

$$\begin{aligned}P(T\not\subset\widehat{T})=&P(\exists j\in T,\widehat{v}_j=0)\leq P(\exists j\in T,|v_{0j}-\widehat{v}_j|\geq|v_{0j}|)\\\leq&P(\max_j|v_{0j}-\widehat{v}_j|\geq d_n)\leq P(\|\widehat{v}-v_0\|\geq d_n)=o(1),\end{aligned}$$

implying $P(T\subset\widehat{T})\to 1$. Accordingly, $P(T=\widehat{T})\to 1$.

(iii). Let $\widehat{v}=(\widehat{v}_S^{\intercal},\widehat{v}_N^{\intercal})^{\intercal}$ be the local minimizer of $Q_n(v)$ where $\widehat{v}_N=0$ with probability arbitrarily close to one. Define $P_n'(|\widehat{v}_S|):=(P_n'(|\widehat{v}_{S1}|),\ldots,P_n'(|\widehat{v}_{St}|))^{\intercal}$ and $\mathrm{sgn}(\widehat{v}_S):=(\mathrm{sgn}(\widehat{v}_{S1}),\ldots,\mathrm{sgn}(\widehat{v}_{St}))^{\intercal}$.

By the Karush–Kuhn–Tucker (KKT) condition,

$$S_{nT}(\widehat{v}_S)=-P_n'(|\widehat{v}_S|)\diamond\mathrm{sgn}(\widehat{v}_S),$$

where the operator $\diamond$ is the product in elementwise.

It follows from Taylor theorem that $S_{nT}(\widehat{v}_S) = S_{nT}(v_{0S}) + H_{nT}(v_{0S})(\widehat{v}_S - v_{0S})$, where a higher order term is ignored because of the consistency of $\widehat{v}_S$ implied by Lemma A.8 and the order of a_n given in Appendix C, which further implies

$$\begin{aligned}\widehat{v}_S - v_{0S} =& H_{nT}(v_{0S})^{-1}[S_{nT}(\widehat{v}_S) - S_{nT}(v_{0S})]\\ =& -H_{nT}(v_{0S})^{-1}[S_{nT}(v_{0S}) + P_n'(|\widehat{v}_S|)\diamond \operatorname{sgn}(\widehat{v}_S)]\\ =& -h_{nT}(\alpha_{0S}, g)^{-1}[s_{nT}(\alpha_{0S}, g) + P_n'(|\widehat{v}_S|)\diamond \operatorname{sgn}(\widehat{v}_S)](1 + o_P(1))\end{aligned}$$

under the condition for $t_n = p_1 + K_1$ by Lemmas A.2 and A.3 where $h_{nT}(\alpha_{0S}, g)$ and $s_{nT}(\alpha_{0S}, g)$ are the counterparts of $h_n(\alpha, g)$ and $s_n(\alpha, g)$, respectively, under the oracle model T.

Similar to the proof of Theorem 3.2, by $\widehat{g}(z) := \Phi_{KT}(z)^{\intercal}\widehat{\beta}_S$,

$$\begin{pmatrix}\mathscr{L}(\widehat{\alpha}_S) - \mathscr{L}(\alpha_{0S})\\ \mathscr{F}(\widehat{g}(z)) - \mathscr{F}(g(z))\end{pmatrix} = \Gamma_n(\widehat{v}_S - v_{0S}) + \begin{pmatrix}0\\ \mathscr{F}'(g)\gamma_K(z)\end{pmatrix}$$

$$= -\Gamma_n h_{nT}(\alpha_{0S}, g)^{-1}[s_{nT}(\alpha_{0S}, g) + P_n'(|\widehat{v}_S|)\diamond \operatorname{sgn}(\widehat{v}_S)] + \begin{pmatrix}0\\ \mathscr{F}'(g)\gamma_K(z)\end{pmatrix}.$$

Notice that the structure

$$\Gamma_n h_{nT}(\alpha_{0S}, g)^{-1}s_{nT}(\alpha_{0S}, g) = \frac{1}{n}\Gamma_n(\Psi_{nT}\Psi_{nT}^{\intercal})^{-1}\Psi_{nT}\sum_{i=1}^{n} m(V_i, \alpha_{0S}^{\intercal}X_{iS}, g(Z_i)).$$

So that invoking classical central limit theorem (i.i.d. innovations) gives

$$\sqrt{n}\Sigma_{nT}^{-1}\Gamma_n h_{nT}(\alpha_{0S}, g)^{-1}s_{nT}(\alpha_{0S}, g) \xrightarrow{d} N(0, I_{r+s})$$

as $n \to \infty$. It remains to show $\sqrt{n}\Sigma_{nT}^{-1}P_n'(|\widehat{v}_S|)\diamond \operatorname{sgn}(\widehat{v}_S) = o_P(1)$. Similar to Lemma C.2 of Fan and Liao (2014) we may show that

$$\|P_n'(|\widehat{v}_S|)\diamond \operatorname{sgn}(\widehat{v}_S)\| = O_P(\max_{\|v_S - v_{0S}\|\le d_n/4}\phi(v_S)\sqrt{t_n \log(q)/n} + P_n'(d_n)).$$

Note also that Σ_{nT} has fixed dimension and its eigenvalues are bounded from zero and above. Thus, the assertion holds under Assumption 5.4. This finishes the proof. □

Proof of Theorem 5.2. Recall that $\widehat{v} = (\widehat{v}_S^{\intercal}, \widehat{v}_N^{\intercal})^{\intercal}$ and $P(\widehat{v}_N = 0) \to 1$. Also, recall the notation $\widehat{v}_T = (\widehat{\alpha}_S^{\intercal}, 0^{\intercal}, \widehat{\beta}_S^{\intercal}, 0^{\intercal})^{\intercal}$.

First, we shall show that $\|M_n(\widehat{v}_T)\|^2 = O_P(t_n^{3/2}\log(q)/n + t_n^{3/2}P_n'(d_n)^2 + t_n\sqrt{\log(q)/n}P_n'(d_n))$. Notice that $\|M_n(\widehat{v}_T)\|^2 = \|M_n(v_0)\|^2 + \|M_n(\widehat{v}_T)\|^2 - \|M_n(v_0)\|^2$ and by the mean value theorem,

$$\begin{aligned}&\|M_n(\widehat{v}_T)\|^2 - \|M_n(v_0)\|^2 = S_{nT}(v_S^*)^{\intercal}(\widehat{v}_S - v_{0S})\\ =& S_{nT}(v_{0S})^{\intercal}(\widehat{v}_S - v_{0S}) + [S_{nT}(v_S^*) - S_{nT}(v_{0S})]^{\intercal}(\widehat{v}_S - v_{0S}).\end{aligned}$$

where v_S^* is a point on the segment joining $\widehat{v}_S$ and v_{0S}.

Notice further,

$$|S_{nT}(v_{0S})^{\intercal}(\widehat{v}_S - v_{0S})| \le \|S_{nT}(v_{0S})\|\|\widehat{v}_S - v_{0S}\| = O_P(t_n\log(q)/n + t_n\sqrt{\log(q)/n}P_n'(d_n))$$

due to $\|S_{nT}(v_{0S})\| = O_P(\sqrt{t_n\log(q)/n})$ and $\|\widehat{v}_S - v_{0S}\| = O_P(\sqrt{t_n\log(q)/n} + \sqrt{t_n}P_n'(d_n))$. Meanwhile, it follows from Assumption 5.2 that

$$\begin{aligned}&|[S_{nT}(v_S^*) - S_{nT}(v_{0S})]^{\intercal}(\widehat{v}_S - v_{0S})| \le \|S_{nT}(v_S^*) - S_{nT}(v_{0S})\|\|\widehat{v}_S - v_{0S}\|\\ \le& O_P(\sqrt{t_n})\|v_S^* - v_{0S}\|\|\widehat{v}_S - v_{0S}\| \le O_P(\sqrt{t_n})\|\widehat{v}_S - v_{0S}\|^2 = O_P(t_n^{3/2}\log(q)/n + t_n^{3/2}P_n'(d_n)^2).\end{aligned}$$

The assertion then follows by noting that $\|M_n(v_0)\|^2 = O_P(\log(q)/n)$ shown by (C.3) in the supplemental material of this paper.

Second, we shall show that $Q_n(\widehat{v}_T) = O_P(t_n^{3/2}\log(q)/n + t_n^{3/2}P_n'(d_n)^2 + t_n\sqrt{\log(q)/n}P_n'(d_n) + t_n\max_{j\in T}P_n(|v_{0j}|))$. Indeed, using the mean value theorem again

$$\begin{aligned}&\sum_{j\in T}P_n(|\widehat{v}_j|) \le \sum_{j\in T}P_n(|v_{0j}|) + \sum_{j\in T}P_n'(|v_{0j}^*|)|\widehat{v}_j - v_{0j}|\\ \le& t_n\max_{j\in T}P_n(|v_{0j}|) + \sum_{j\in T}P_n'(d_n)|\widehat{v}_j - v_{0j}| \le t_n\max_{j\in T}P_n(|v_{0j}|) + \sqrt{t_n}P_n'(d_n)\|\widehat{v} - v_0\|,\end{aligned}$$

from which the assertion follows. Combining the two steps gives $Q_n(\widehat{v}_T) = o_P(1)$.

Notice further that

$$Q_n(v) \geq \|M_n(v)\|^2 = \frac{1}{q}\left\|\frac{1}{n}\sum_{i=1}^{n} m(V_i, v^\tau F_i)\right\|^2$$
$$\geq \frac{1}{2q}\|\mathbb{E}m(V_1, v^\tau F_1)\|^2 - \frac{1}{q}\left\|\frac{1}{n}\sum_{i=1}^{n} m(V_i, v^\tau F_i) - \mathbb{E}m(V_1, v^\tau F_1)\right\|^2$$
$$= \frac{1}{2q}\|\mathbb{E}m(V_1, v^\tau F_1)\| + o_P(n^{-1/2}),$$

uniformly in v. Then, for any $\delta > 0$,

$$\inf_{\|v-v_0\|\geq\delta} Q_n(v) \geq \inf_{\|v-v_0\|\geq\delta} \frac{1}{2q}\|\mathbb{E}m(V_1, v^\tau F_1)\| + o_P(n^{-1/2})$$
$$= \inf_{\|(\mathbf{a}-\alpha, f-g)\|\geq\delta+\|\gamma_K(z)\|} \frac{1}{q}\|\mathbb{E}m(V_1, \mathbf{a}^\tau X_1, f(Z_1))\| + o_P(n^{-1/2}),$$

due to by definition $\|v - v_0\| = \|\mathbf{a} - \alpha\| + \|\mathbf{b} - \beta\| = \|\mathbf{a} - \alpha\| + \|f - g\| - \|\gamma_K(z)\|$. As a result, by Assumption 3.2, there exists $\epsilon > 0$ such that $\inf_{\|v-v_0\|\geq\delta} Q_n(v) \geq \epsilon$ for sufficient large n.

Taking $0 < \eta < \epsilon$,

$$P\left(Q_n(\widehat{v}) + \eta > \inf_{\|v-v_0\|\geq\delta} Q_n(v)\right) = P\left(Q_n(\widehat{v_T}) + \eta > \inf_{\|v-v_0\|\geq\delta} Q_n(v)\right) + o(1)$$
$$\leq P\left(Q_n(\widehat{v_T}) + \eta > \epsilon\right) + P\left(\inf_{\|v-v_0\|\geq\delta} Q_n(v) < \epsilon\right) + o(1) \leq P\left(Q_n(\widehat{v_T}) > \epsilon - \eta\right) + o(1) = o(1)$$

because $Q_n(\widehat{v_T}) = o_P(1)$. □

Appendix C. Proofs of Lemmas A.1–A.9 and Theorems 4.3 and 4.4

Supplementary material related to this article can be found online at https://doi.org/10.1016/j.jeconom.2021.07.004.

References

Ai, C., Chen, X., 2003. Efficient estimation of models with conditional moment restrictions containing unknown functions. Econometrica 71, 1795–1843.
Andrews, D.W., 1994. Asymptotics for semiparametric econometric models via stochastic equicontinuity. Econometrica 62, 43–72.
Andrews, D.W., 1999. Consistent moment selection procedures for generalized method of moments estimation. Econometrica 67, 543–564.
Andrews, D.W., Lu, B., 2001. Consistent model and moment selection procedures for GMM estimation with application to dynamic panel data models. J. Econometrics 101, 123–165.
Antoniadis, A., 1996. Smoothing noisy data with tapered coiflets series. Scand. J. Stat. 23, 313–330.
Athey, S., Imbens, G., Pham, T., Wager, S., 2017. Estimating average treatment effects: Supplementary analysis and remaining challenges. Amer. Econ. Rev. 107, 278–281.
Belloni, A., Chen, D., Chernozhukov, V., Hansen, C., 2012. Sparse models and methods for optimal instruments with an application to eminent domain. Econometrica 80, 2369–2429.
Belloni, A., Chernozhukov, V., Chetverikov, D., Kato, K., 2015. Some new asymptotic theory for least squares series: Pointwise and uniform results. J. Econometrics 186, 345–366.
Belloni, A., Chernozhukov, V., Hansen, C., 2014a. High-dimensional methods and inference on structural and treatment effects. J. Econ. Perspect. 28, 29–50.
Belloni, A., Chernozhukov, V., Hansen, C., 2014b. Inference on treatment effects after selection among high-dimensional controls. Rev. Econom. Stud. 81, 608–650.
Belloni, A., Chernozhukov, V., Hansen, C., Wei, Y., 2016a. Post-selection inference for generalized linear models with many controls. J. Bus. Econom. Statist. 34, 590–605.
Belloni, A., Chernozhukov, V., Wang, L., 2014c. Pivitor estimation via square-root lasso in nonparametric regression. Ann. Statist. 42, 757–788.
Belloni, A., Rosenbaum, M., Tsybakov, A.B., 2016b. Linear and conic programming estimators in high-dimensional errors-in-variables models. Electron. J. Stat. 10, 1729–1750.
Bickel, P.J., 1982. On adaptive estimation. Ann. Statist. 10, 647–671.
Bickel, P., Klaassen, C.A., Ritov, Y., Wellner, J.A., 1993. Efficient and Adaptive Estimation for Semiparametric Models. The John Hopkins University Press, Baltimore and London.
Blundell, R., Chen, X., Christensen, D., 2007. Semi-nonparametric IV estimation of shape-invariant engel curve . Econometrica 75, 1613–1669.
Caner, M., 2009. Lasso-type GMM estimator. Econom. Theory 25, 270–290.
Carneiro, P., Heckman, J., Vytlacil, E., 2011. Estimating marginal returns to education. Amer. Econ. Rev. 101, 2754–2781.
Cattaneo, M.D., Jansson, M., Newey, W.K., 2018. Inference in linear regression models with many covariates and heteroskedasticity. J. Amer. Statist. Assoc. 113, 1350–1361.
Chang, J., Chen, S., Chen, X., 2015. High dimensional generalized empirical likelihood for moment restrictions with dependent data. J. Econometrics 185, 283–304.
Chen, X., 2007. Large sample sieve estimation of semi-parametric models. In: Engle, R.F., MacFadden, D.L. (Eds.), In: Handbook of Econometrics, Vol. 6B, Elsevier, Amsterdam: North Holland, pp. 5550–5588.

Chen, X., Christensen, T., 2015. Optimal uniform convergence rates and asymptotic normality for series estimators under weak dependence and weak conditions. J. Econometrics 188, 447–465.
Chen, X., Liao, Z., 2015. Sieve semiparametric two-step GMM under weak dependence. J. Econometrics 189, 163–186.
Chen, X., Linton, O., Keilegom, I.V., 2003. Estimation for semiparametric models when the criterion function is not smooth. Econometrica 71, 1591–1608.
Chen, X., Pouzo, D., 2009. Efficient estimation of semiparametric conditional moment models with possibly nonsmooth residuals. J. Econometrics 152, 46–60.
Chen, X., Pouzo, D., 2012. Estimation of nonparametric conditional moment models with possibly nonsmooth generalized residuals. Econometrica 80, 277–321.
Chen, X., Shen, X., 1998. Sieve extremum estimates for weakly dependent data. Econometrica 66, 289–314.
Connor, G., Hagmann, M., Linton, O., 2012. Efficient semiparametric estimation of the fama-french model and extensions. Econometrica 80, 713–754.
Dong, C., Gao, J., Linton, O., 2018. High dimensional semiparametric moment restriction models. In: Cambridge Working Papers in Economics 1881.
Dong, C., Gao, J., Peng, B., 2015. Semiparametric single-index panel data models with cross-sectional dependence. J. Econometrics 188, 301–312.
Dong, C., Gao, J., Tjøstheim, D., 2016. Estimation for single-index and partially linear single-index integrated models. Ann. Statist. 44, 425–453.
Dong, C., Linton, O., 2018. Additive nonparametric models with time variable and both stationary and nonstationary regressors. J. Econometrics 207, 212–236.
Dong, C., Linton, O., Peng, B., 2021. A weighted sieve estimator for nonparametric time series models with nonstationary variables. J. Econometrics 222, 909–932.
Dudley, R.M., 2003. Real Analysis and Probability. In: Cambridge studies in advanced mathematics 74, Cambridge University Press, Cambridge, U.K..
Embrechts, P., Klüppelberg, C., Mikosch, T., 1999. Modelling Extremal Events for Insurance and Mathematics. Springer-Verlag, Berlin.
Fan, J., Li, R., 2001. Variable selection via nonconcave penalized likelihood and its oracle properties. J. Amer. Statist. Assoc. 96, 1348–1360.
Fan, J., Liao, Y., 2014. Endogeneity in high dimensions. Ann. Statist. 42, 872–917.
Gautschi, W., 2004. Orthogonal Polynomials: Computation and Approximation. In: Numerical Mathematics and Scientific Computation, Oxford University Press, Oxford.
Han, C., Phillips, P.C.B., 2006. GMM with many moment conditions. Econometrica 74, 147–192.
Hansen, L.P., 1982. Large sample properties of generalized method of moments estimators. Econometrica 50, 1029–1054.
Hansen, L., Heaton, J., Yaron, A., 1996. Finite-sample properties of some alternative GMM estimators. J. Bus. Econom. Statist. 14, 262–280.
Jankova, J., Geer, S.V.D., 2018. Semiparametric efficiency bounds for high dimensional models. Ann. Statist. 46, 2336–2359.
Leeb, H., Pötscher, B.M., 2005. Model selection and inference: facts and fiction. Econom. Theory 21, 21–59.
Mammen, E., 1989. Asymptotics with increasing dimension for robust regression with applications to the bootstrap. Ann. Statist. 17, 382–400.
Newey, W.K., 1994. The asymptotic variance of semiparametric estimators. Econometrica 62, 1349–1382.
Newey, W.K., 1997. Convergence rates and asymptotic normality for series estimators. J. Econometrics 79, 147–168.
Newey, W.K., McFadden, D., 1994. Large sample estimation and hypothesis testing. In: Engle, R.F., MacFadden, D.L. (Eds.), In: Handbook of Econometrics, Vol. IV, Elsevier, Amsterdam: North Holland, pp. 2111–2245.
Newey, W.K., Powell, J.L., 2003. Instrumental variable estimation of nonparametric models. Econometrica 71, 1565–1578.
Newey, W.K., Smith, R.J., 2004. Higher order properties of GMM and generalized empirical likelihood estimators. Econometrica 72, 219–255.
Newey, W.K., Windmeijer, F., 2009. Generalized method of moments with many weak moment conditions. Econometrica 77, 687–719.
Pakes, A., Olley, S., 1995. A limit theorem for a smooth class of semiparametric estimators. J. Econometrics 65, 295–332.
Pakes, A., Pollard, D., 1989. Simulation and the asymptotics of optimization estimators. Econometrica 57, 1027–1057.
Pesaran, M.H., Yamagata, T., 2017. Testing for alpha in linear factor pricing models with a large number of securities. CESifo Working Paper Series No. 6432, Available at SSRN: https://ssrn.com/abstract=2973079.
Portnoy, S., 1984. Asymptotic behaviour of M–estimators of p regression parameters when p^2/n is large. I: Consistency. Ann. Statist. 12, 1298–1309.
Portnoy, S., 1985. Asymptotic behaviour of M–estimators of p regression parameters when p^2/n is large. II: Normal approximation. Ann. Statist. 13, 1403–1417.
Powell, J.L., 1984. In: Engle, R., McFadden, D. (Eds.), Estimation of Semiparametric Models. In: Handbook of Econometrics IV, Elsevier, New York, pp. 2444–2521.
Robinson, P.M., 1988. Root-N-consistent semiparametric regression. Econometrica 56, 931–954.
Su, L., Ura, T., Zhang, Y., 2018. Non-separable models with high-dimensional data. Unpublished paper at https://arxiv.org/abs/1702.04625.
Yu, Y., Ruppert, D., 2002. Penalized spline estimation for partially linear single-index models. J. Amer. Statist. Assoc. 97, 1042–1054.
Zhang, C.H., 2010. Nearly unbiased variable selection under minmax concave penalty. Ann. Statist. 38, 894–942.