Dong, Chaohua; Gao, Jiti; Linton, Oliver

Working Paper

High dimensional semiparametric moment restriction models

cemmap working paper, No. CWP04/18

Provided in Cooperation with:
Institute for Fiscal Studies (IFS), London

Suggested Citation: Dong, Chaohua; Gao, Jiti; Linton, Oliver (2018) : High dimensional semiparametric moment restriction models, cemmap working paper, No. CWP04/18, Centre for Microdata Methods and Practice (cemmap), London, https://doi.org/10.1920/wp.cem.2018.0418

cemmap
centre for microdata methods and practice

High dimensional semiparametric moment restriction models

Chaohua Dong
Jiti Gao
Oliver Linton

The Institute for Fiscal Studies
Department of Economics, UCL

cemmap working paper CWP04/18

E·S·R·C ECONOMIC & SOCIAL RESEARCH COUNCIL
An ESRC Research Centre

High Dimensional Semiparametric Moment Restriction Models

Chaohua Dong

Southwestern University of Finance and Economics, China

Jiti Gao

Monash University, Australia

Oliver Linton

University of Cambridge, UK

November 24, 2017

Abstract

Moment restriction semiparametric models, where both the dimension of parameter and the number of restrictions are divergent and an unknown function is involved, are studied using the generalized method of moments (GMM) and sieve method dealing with the nonparametric parameter. The consistency and normality for the GMM estimators are established. Meanwhile, a new test statistic is proposed for over-identification issue, which also is workable for the traditional moment restriction models. In addition, the potential sparsity under our setting is investigated via the combination of GMM methodology and penalty function approach. Numerical examples are used to verify the established theory.

Key works: Generalized method of moments, high dimensional models, moment restriction, over-identification, sieve method, sparsity

JEL classification: C12, C14, C22, C30

1 Introduction and examples

We consider a class of moment restriction models where there are many Euclidean valued parameters as well as unknown infinite dimensional functional parameters. The setting includes

as a special case the partial linear regression model, Robinson [33], except in our case the number of covariates in the linear part may be large, i.e., increase to infinity with sample size. For example, there are often many binary covariates whose effect can be restricted to be linear without a great loss of generality. Our model framework specifies which variables affect the outcome in a linear fashion and which variables affect the outcome nonlinearly. Nonparametric and "parametric" components can both be of interest in applications but present different statistical issues. Efficiency bounds and their achievability are quite different between the two cases. Inference procedures also differ substantially. In our framework, the model components are known beforehand and are clearly demarcated, and we can compare our results for the two different components with existing results in the relevant literatures. However, the parametric component itself is growing in complexity, which raises some new issues. We will use 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 and recent work has mainly focussed on the extension to either many moment conditions (Newey and Windmeijer [28], for example) or to the case where the number of Euclidean parameters is finite but there are unknown function-valued parameters (see, for example, Chen and Liao [11]; Chen et al. [12]). We will provide inference techniques for the parametric and nonparametric components of our model.

Suppose that

$$\mathbb{E}[m(V_i, \alpha^\intercal X_i, g(Z_i))] = 0, \tag{1.1}$$

for $i = 1, \ldots, n$, where m is a known vector of functions whose dimension q is large, i.e., $q = q(n) \to \infty$ as $n \to \infty$. Here, α is an unknown Euclidean valued parameter whose dimension $p = p(n) \to \infty$ as $n \to \infty$, while g is an unknown smooth function. The observed vector variable V_i typically represents a dependent variable and possible instrumental variables, while the observed vectors X_i and Z_i are explanatory variables, where Z_i, V_i are of finite dimension, but the dimension of X_i may diverge. We will consider the case where the parameter dimension p grows to infinity but is smaller than n, similar to Portnoy [30], Portnoy [31] and Mammen [25]. This is the case in many applications. The moment restriction model (1.1) features high dimensionality in two folds: a high dimensional Euclidean parameter (α) that shows up in a single-index form, and a zero-mean function $m(\cdot)$ with divergent dimension that usually represents an error term. In addition, it includes an infinite dimensional unknown function $g(\cdot)$. Together this represents a new framework in the literature.

We suppose that a sample $(V_i, X_i^\intercal, Z_i^\intercal)_{i=1}^n$ is observed. We shall simultaneously estimate α and g in parameter spaces defined below. As the function g can be regarded as an element in some function space, which is infinite dimensional, all parameters are of high dimension. Moreover, we are also interested in transformations of α and functionals of g for which we

have plug-in estimators once we obtain the estimates of α and g. Chen et al. [12] study a fixed-dimensional moment restriction model containing an unknown function. The estimation strategy can be two step or profiled two-steps depending on the context. A similar approach is used again in Chen and Liao [11]. Kernel estimation techniques generally require an additional (albeit related) estimating equation and either two-step or profile methods are common, see, for example, Powell [32].

To illustrate the proposal of model (1.1), we give the following examples.

Example 1.1 (Conditional moment restrictions): Let W_i be a sub-vector of $(X_i^\intercal, Z_i^\intercal)^\intercal$ and $\rho(Y_i, \alpha^\intercal X_i, g(Z_i))$ be a known J-dimensional vector of generalized residual function. Then, (α, g) is determined by a conditional moment restriction

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

Let $\Phi_k(w) = (h_1(w), \ldots, h_k(w))$ be a vector of functions that can approximate any square integrable function of W in some sense arbitrarily as $k \to \infty$. Then, the conditional restriction implies

$$\mathbb{E}[\rho(Y_i, \alpha^\intercal X_i, g(Z_i)) \otimes \Phi_k(W_i)] = 0.$$

Denote $m(V_i, \alpha^\intercal X_i, g(Z_i)) = \rho(Y_i, \alpha^\intercal X_i, g(Z_i)) \otimes \Phi_k(W_i)$ where $V_i = (Y_i, W_i^\intercal)^\intercal$. Notice that the dimension of m function is Jk which increases with k. Therefore, the pair (α, g) can be solved from the unconditional moment equation $\mathbb{E}[m(V_i, \alpha^\intercal X_i, g(Z_i))] = 0$. For example, suppose that $\Lambda(Z_i) = \alpha^\intercal X_i + \varepsilon_i$, where $\Lambda(\cdot)$ is an unknown monotone function and (V_i, X_i) are observed. Under the conditional moment restriction $\mathbb{E}(\varepsilon|W) = 0$ for some vector of instrumental variables W we may obtain unconditional moment restrictions like (1.1) with Λ being the unknown function of interest.

Example 1.2 (High dimensional partially linear endogenous model): 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, Z_i) is endogenous in the sense that $\mathbb{E}[e_i|X_i, Z_i] \neq 0$. In the case where the dimensionality of α is fixed, there are various results available in the literature (see, for example, Robinson [33]; Gao and Liang [19]; Gao and Shi [20]; Härdle et al. [24]). To deal with the endogeneity, let W_i be instrumental variable and define a set of valid instruments $\lambda_i = \lambda(W_i)$ with dimension q and $q > p$.

Denote $m(V_i, \alpha^\intercal X_i, g(Z_i)) = (Y_i - \alpha^\intercal X_i - g(Z_i))\lambda_i(W_i)$ 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 nonparametric function $g(\cdot)$.

Example 1.3 (Discrete maximum likelihood): Suppose that Y_i assumes either 0 or 1, and

$$P(Y_i = 1|X_i, Z_i) = F(\alpha^\intercal X_i + g(Z_i)),$$

for $i = 1, \ldots, n$, where $\alpha, X_i \in \mathbb{R}^p$ and $Z_i \in \mathbb{R}$. The log likelihood function is

$$\ln \prod_{i=1}^{n} F^{Y_i}(\alpha^\intercal X_i + g(Z_i))[1 - F(\alpha^\intercal X_i + g(Z_i))]^{1-Y_i}.$$

A sieve method can be used to estimate the unknown $g(\cdot)$, along with the estimate of α. Suppose that the function $g(\cdot)$ can be approximated arbitrarily in some sense by a linear combination of k known functions wrapped into a column vector $\Phi_k(\cdot)$, i.e. $g(z) - \beta^\intercal \Phi_k(z)$ is approaching zero in some sense as $k \to \infty$. Thus, the estimate of (α, g) can be obtained through maximizing

$$Q_n(\alpha, \beta) := \ln \prod_{i=1}^{n} F^{Y_i}(\alpha^\intercal X_i + \beta^\intercal \Phi_k(Z_i))[1 - F(\alpha^\intercal X_i + \beta^\intercal \Phi_k(Z_i))]^{1-Y_i},$$

to have $\widehat{\alpha}$ and $\widehat{\beta}$ (hence naturally $\widehat{g}(z) = \widehat{\beta}^\intercal \Phi_k(z)$). The first order condition gives

$$\begin{aligned}
\frac{\partial Q_n}{\partial \alpha} =& \sum_{i=1}^{n} \frac{[Y_i - F(\alpha^\intercal X_i + \beta^\intercal \Phi_k(Z_i))]F(\alpha^\intercal X_i + \beta^\intercal \Phi_k(Z_i))}{F(\alpha^\intercal X_i + \beta^\intercal \Phi_k(Z_i))[1 - F(\alpha^\intercal X_i + \beta^\intercal \Phi_k(Z_i))]} X_i = 0, \\
\frac{\partial Q_n}{\partial \beta} =& \sum_{i=1}^{n} \frac{[Y_i - F(\alpha^\intercal X_i + \beta^\intercal \Phi_k(Z_i))]F(\alpha^\intercal X_i + \beta^\intercal \Phi_k(Z_i))}{F(\alpha^\intercal X_i + \beta^\intercal \Phi_k(Z_i))[1 - F(\alpha^\intercal X_i + \beta^\intercal \Phi_k(Z_i))]} \Phi_k(Z_i) = 0,
\end{aligned}$$

for $i = 1, \ldots, n$, which can be viewed as a sample version of moment condition, with $m(\cdot)$ defined properly, $\mathbb{E}[m(Y_i, X_i, \Phi_k(Z_i), \alpha^\intercal X_i, \beta^\intercal \Phi_k(Z_i))] = 0$.

Our strategy for dealing with the specification of model (1.1) is simple. Suppose that $g(\cdot)$ belongs to a suitable Hilbert space. We then expand the function $g(\cdot)$ into an infinite orthogonal series in terms of a basis in the Hilbert space, $\{\varphi_j(z)\}$, say. As a result, $g(z)$ can be approximated by the partial sum $\sum_{j=0}^{k-1} \beta_j \varphi_j(z)$ in the norm of the space. In this way, the unknown function is completely parameterized, which enables us to estimate the parameter vector α and the function $g(\cdot)$ in model (1.1) simultaneously. This procedure also avoids high level assumptions in our study. By contrast, some high level conditions are engaged in the relevant literature, such as Chen et al. [12] and Han and Phillips [22]. In addition, our approach can be classified as a one-step GMM method, in contrast with the two-step GMM study in the literature that engages an initial estimator. See, for example, Chen and Liao [11].

In addition to the estimation of model (1.1), we also propose a new test statistic, to the best of our knowledge, in order to tackle over-identification issue. Moreover, given the divergence of the number of both regressors and moment restriction, it is desirable to consider the situation for model (1.1) that possesses sparsity. That is, $p > n$ but α contains plenty of zeros except that some so-called important coefficients are nonzero. To estimate the parameters of interest under sparsity, often a penalty function should be combined with the

objective function. It can be seen in the sequel that the variable selection and estimation can be done simultaneously.

The rest of the paper is organized as follows. Section 2 gives the estimation procedure; Section 3 provides the asymptotic theory for the estimator proposed in the preceding section; Section 4 studies the over-identification issue. The sparsity in our model is investigated in Section 5, followed by numerical evidence in Section 6; and the last section then 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

The unknown function $g(z)$ can be a vector of functions or a multivariate function. Both of these contexts are useful in practice and they may be dealt with similarly using sieve method. For the sake of easy exposition, however, we suppose in this paper that $g(z)$ is a single multivariate function defined on $\mathbb{Z} \subset \mathbb{R}^d$. Let $g(z) \in L^2(\mathbb{Z}, \pi(z)) = \{f(z) : \int_{\mathbb{Z}} f^2(z)\pi(z)dz < \infty\}$ a Hilbert function space, where $\pi(z)$ is a user-chosen density function on $\mathbb{Z}$. The choice of the density $\pi(z)$ relates to how large the Hilbert space should be, 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 $f_1(z), f_2(z), f(z) \in L^2(\mathbb{Z}, \pi(z))$. Two functions $f_1(z), f_2(z) \in L^2(\mathbb{Z}, \pi(z))$ are called orthogonal if $\langle f_1, f_2\rangle = 0$, and further are orthonormal if $\|f_1\| = 1$ and $\|f_2\| = 1$.

The parameter space for model (1.1) is defined as, $\Theta = \{(\mathbf{a}, f) : \mathbf{a} \in \mathbb{R}^p, f \in L^2(\mathbb{Z}, \pi(z))\}$, which contains the true parameter (α, g) as an interior point.

Assumption 2.1 *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.*

Recall that any Hilbert space has a complete orthogonal sequence [see Theorem 5.4.7 in 15, p. 169]. In our setting, although $g(\cdot)$ is multivariate, the orthonormal sequence $\{\varphi_j(\cdot)\}$ can be constructed from the tensor product of univariate orthogonal sequences. Thus, we hereby briefly introduce some existing univariate orthonormal sequences only.

Generally speaking, an orthonormal sequence depends on its support on which it is defined and the density by which the orthogonality is defined. Hermite polynomials form a complete orthogonal sequence on $\mathbb{R}$ with respect to the density e^{-u^2}; Laguerre polynomials are a complete orthogonal sequence on $[0, \infty)$ with density e^{-u}; Legendre polynomials and also

orthogonal trigonometric polynomials are complete orthogonal sequence on $[0,1]$ with the uniform density; Chebyshev polynomials are complete orthogonal on $[-1,1]$ with density $1/\sqrt{1-u^2}$. See, e.g. Chapter One of Gautschi [21].

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

$$g(z) = \sum_{j=0}^{\infty} \beta_j \varphi_j(z), \quad \text{where } \beta_j = \langle g, \varphi_j \rangle. \tag{2.1}$$

The convergence of (2.1) normally can be understood in the sense of the norm in the space, whereas in the situation where $g(z)$ is smooth, the pointwise sense may hold. 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 after truncation. Then, $g_k(z) \to g(z)$ as $k \to \infty$ in some sense. Note that $g_k(z)$ is a parameterized version of $g(z)$ in terms of the basis $\{\varphi_j(z)\}$ where only the coefficients remain unknown. This is the main advantage of the sieve method. In addition, the Parseval equality gives $\sum_{j=0}^{\infty} \beta_j^2 = \|g\|^2 < \infty$, implying the attenuation of the coefficients. For better exposition, denote $\Phi_k(z) = (\varphi_0(z), \ldots, \varphi_{k-1}(z))^\intercal$ and $\beta = (\beta_0, \ldots, \beta_{k-1})^\intercal$ two k-vectors. Thus, $g_k(z) = \beta^\intercal \Phi_k(z)$.

Our primary goal is to estimate $(\alpha, g(\cdot))$, and the consistency studied below will be defined in terms of a norm given by

$$\|(\mathbf{a}, f(\cdot))\| = \left(\|\mathbf{a}\|_E^2 + \|f\|_{L^2}^2\right)^{1/2}, \tag{2.2}$$

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 there is no ambiguity incurred.

As usual, in order to facilitate an implementation of nonlinear optimization, α should be confined in a compact subset of $\mathbb{R}^p$ and the truncated series $g_k(z) = \beta^\intercal \Phi_k(z)$ of g function should be included in an expanding finite dimensional bounded subsets of $L^2(\mathbb{Z}, \pi(z))$. It is noteworthy that in an infinite dimensional space, a bounded subset may not necessarily be a compact set. A detailed discussion on the relationship for the compactness in infinite dimensional space can be found in Chen and Pouzo [13]. Nevertheless, in the case that the function m is linear in the second and the third arguments, such restrictions are not necessary (we shall discuss this in Section 5 using an example).

Assumption 2.2 *Suppose that B_{1n} and B_{2n} are positive reals 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 $\Theta_{2n} := \{\mathbf{b}^\intercal \Phi_k(z) : \|\mathbf{b}\| \le B_{2n}\}$.*

Here, unlike in a general 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. It is also a convention on computation that the true parameter is assumed to be contained

within a bounded set [see 27, p. 1569], whereas the difference is that in this paper we allow the bounds of α to diverge with the sample size since the dimensionality of α grows to infinity.

Meanwhile, since $\|g_k(z)\| = \|\beta\| \leq \|g\|$ it is clear that there exists an integer n_0 such that $g_k(z) \in \Theta_{2n}$ for all $n \geq n_0$. Similar to the orthogonal expansion in (2.1), for any $f(z) \in L^2(\mathbb{Z}, \pi(z))$, $f(z)$ can be approximated by $\sum_{j=0}^{k-1} b_j \varphi_j(z) = \mathbf{b}^\intercal \Phi_k(z)$ arbitrarily in the sense of norm, where b_j and $\mathbf{b}$ are defined similarly to β_j and β, respectively. This means that Θ_{2n} is approximating the function space with the increase of the sample size. Thus, the parametric space can be approximated by $\Theta_n = \Theta_{1n} \otimes \Theta_{2n}$ as $n \to \infty$. In the literature, Θ_{2n} is the so-called linear sieve space. More importantly, Θ_n is bounded and compact. The above setting is similar to but broader than that in Newey and Powell [27].

We rewrite the moment condition (1.1) as

$$\mathbb{E}[m(V_i, \alpha^\intercal X_i, \beta^\intercal \Phi_k(Z_i) + \gamma_k(Z_i))] = 0. \tag{2.3}$$

where $\gamma_k(\cdot)$ is negligible for large k. We estimate α and β by

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

Here, the involvement of q in $M_n(\mathbf{a}, \mathbf{b})$ takes into account the divergence of the dimension of the m function in order to avoid that $\|M_n(\mathbf{a}, \mathbf{b})\|$ could be large even if each element is small when we had not put q into $M_n(\mathbf{a}, \mathbf{b})$. However, this issue does not matter when the vector–valued m function has a fixed dimension. In addition, from the proof of the theorems below, the term $\sqrt{q}$ in $M_n(\mathbf{a}, \mathbf{b})$ can be replaced by some appropriately chosen function of q, which normalizes the divergence of the norm of $M_n(\mathbf{a}, \mathbf{b})$. For the sake of simplicity, we take $\sqrt{q}$ in this paper. Then, naturally define

$$\widehat{g}(z) = \widehat{\beta}^\intercal \Phi_k(z) \tag{2.5}$$

for any $z \in \mathbb{Z}$ as an estimator of $g(z)$. In the next section we establish consistency of this estimator in the sense that $\|(\widehat{\theta} - \theta, \widehat{g}(z) - g(z))\| \to_P 0$ as $n \to \infty$ where the norm is defined in (2.2).

3 Asymptotic theory

3.1 Consistency

Before starting our asymptotic theory, we need to state some necessary assumptions.

Assumption 3.1 *Suppose that: (a) for each n, $\{V_i, X_i^\intercal, Z_i^\intercal\}_1^n$ is an independent and identically distributed sequence; (b) for the density $f_Z(z)$ of Z_1, $f_Z(z) \leq C\pi(z)$ on the support $\mathbb{Z}$ of Z_1 for some constant C; (c) the function $m(\cdot,\cdot,\cdot)$ 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 a dependent data setting. About Assumption 3.1.b, the relation between the densities of the variable Z_1 and the function space is widely used in the literature. See, e.g. Condition A.2 and Proposition 2.1 of Belloni et al. [4, p. 347]. For Assumption 3.1.c, the continuity of the m function in the arguments where the parameters are involved is weak and typically used functions satisfy it. 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 $g(\cdot) \in L^2(\mathbb{Z}, \pi(z))$ and for each n there is a unique vector $\alpha \in \mathbb{R}^p$ such that model (1.1) is satisfied. In other words, for any $\delta > 0$, there is an $\epsilon > 0$ such that*

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

It is quite standard in the literature to assume such a uniqueness condition. Here again, 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, for which we do not mention repeatedly in what follows whenever the norm is scaled.

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 $(a_1, f_1), (a_2, f_2) \in \Theta$, where (V, X, Z) is any realization of (V_i, X_i, Z_i) and the function A satisfies that $\max_{i\geq 1} \mathbb{E}[A^2(V_i, X_i, Z_i)] < \infty$ uniformly in n.

This assumption is a kind of Lipschitz condition. 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 vector $\mathbf{a}$ and 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 $m(V_i, \alpha^\intercal X_i, \beta^\intercal \Phi_k(Z_i))$ to $m(V_i, \alpha^\intercal X_i, g(Z_i))$, because

$$\|m(V_i, \alpha^\intercal X_i, \beta^\intercal \Phi_k(Z_i)) - m(V_i, \alpha^\intercal X_i, g(Z_i))\|$$

$$\leq A(V_i, X_i, Z_i)\|g(Z_i) - \beta^\intercal \Phi_k(Z_i)\| = O_P(1)\|\gamma_k(z)\| = o_P(1)$$

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,

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

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

Theorem 3.1 (Consistency)**.** *In addition to Assumptions 2.1-2.2 and 3.1-3.3, suppose that* $B_{1n}^2 + B_{2n}^2 = o(n)$*. Then, we have* $\|(\widehat{\alpha} - \alpha, \widehat{g}(z) - g(z))\| \to_P 0$ *as* $n \to \infty$*.*

The proof is given in Appendix B.

3.2 Limit distributions of the estimator

As the dimension of α diverges, we may not be able to establish a limit distribution for $\widehat{\alpha} - \alpha$. Instead, we shall aim at some finite dimensional transformations of α and functionals of $g(z)$, for which plug-in estimators are used.

Let $\mathscr{L}$ be a linear transformation from $\mathbb{R}^p \mapsto \mathbb{R}^r$ with $r \geq 1$ fixed, and $\mathscr{F} = (\mathscr{F}_1, \ldots, \mathscr{F}_s)^\intercal$ with fixed s be a vector of functionals on $L^2(\mathbb{Z}, \pi(z))$. Normally, the transformation $\mathscr{L}$ can be understood as an $r \times p$ matrix with rank r, while in the literature one usually takes $r = 1$. See, e.g. Theorem 4.2 in Belloni et al. [4, p. 352] and several results such as Theorems 2 and 6 in Chang et al. [9]. Moreover, the elements of $\mathscr{F}$ can be, as described in Newey [26, p.151], the integral of $\ln[g(z)]$ on some interval which stands for consumer's surplus in microeconomics, for example. Other examples include partial derivative function, average partial derivative function and conditional partial derivative.

Thus, we shall consider the limit distributions of $\mathscr{L}(\widehat{\alpha}) - \mathscr{L}(\alpha)$ and $\mathscr{F}(\widehat{g}) - \mathscr{F}(g)$.

Assumption 3.4 *Suppose that the m function is differentiable with respect to its second and third arguments up to the third order. Let the g function be smooth such that Assumption A.2 in Appendix A is satisfied.*

The differentiability of the m function up to the third order makes the derivation of the asymptotic distribution below much simpler than in some papers in the literature since it enables us to expand the score function where the terms with higher order than the Hessian matrix can be ignored. Certainly, this condition can be relaxed to have the derivatives of up to the second order but for simplicity we retain it. It is well known that certain smoothness order of the g function is required to get rid of the truncation residues. Such a requirement is implicitly spelt out by Assumption A.2.

To investigate the asymptotics, denote the Score and Hessian functions

$$S_n(\mathbf{a},\mathbf{b}) := \begin{pmatrix} \frac{\partial}{\partial \mathbf{a}} \\ \frac{\partial}{\partial \mathbf{b}} \end{pmatrix} \|M_n(\mathbf{a},\mathbf{b})\|^2, \qquad H_n(\mathbf{a},\mathbf{b}) := \begin{pmatrix} \frac{\partial^2}{\partial \mathbf{a}\partial \mathbf{a}^\intercal} & \frac{\partial^2}{\partial \mathbf{a}\partial \mathbf{b}^\intercal} \\ \frac{\partial^2}{\partial \mathbf{b}\partial \mathbf{a}^\intercal} & \frac{\partial^2}{\partial \mathbf{b}\partial \mathbf{b}^\intercal} \end{pmatrix} \|M_n(\mathbf{a},\mathbf{b})\|^2.$$

Under certain conditions, the asymptotic behavior of $H_n(\alpha,\beta)$ and $S_n(\alpha,\beta)$ is given by Lemmas A.2 and A.3 in Appendix A.

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

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

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

Theorem 3.2 (Normality)**.** *Let Assumptions 2.1-2.2, 3.1-3.4 and A.1-A.3 (given in Appendix A) hold. Then*

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

as $n \to \infty$ provided that $\sqrt{n}\Sigma_n^{-1}(0_r^\intercal, \mathscr{F}'(g)\gamma_k(z)^\intercal)^\intercal = o(1)$, where Σ_n is given by the square root of

$$\begin{aligned}
\Sigma_n^2 :=& \Gamma_n[\Psi_n\Psi_n^\intercal]^{-1}\Psi_n\Xi_n\Psi_n^\intercal[\Psi_n\Psi_n^\intercal]^{-1}\Gamma_n^\intercal, \quad \textit{in which} \\
\Gamma_n :=& \begin{pmatrix} \mathscr{L} & 0 \\ 0 & \mathscr{F}'(g)\Phi_k(\cdot)^\intercal \end{pmatrix}_{(r+s)\times(p+k)}, \\
\Xi_n :=& \mathbb{E}[m(V_1, \alpha^\intercal X_1, g(Z_1))m(V_1, \alpha^\intercal X_1, g(Z_1))^\intercal]_{q\times q}, \\
\Psi_n :=& \mathbb{E}\begin{pmatrix} \frac{\partial}{\partial u}m(V_1, \alpha^\intercal X_1, g(Z_1))^\intercal \otimes X_1 \\ \frac{\partial}{\partial w}m(V_1, \alpha^\intercal X_1, g(Z_1))^\intercal \otimes \Phi_k(Z_1) \end{pmatrix}_{(p+k)\times q},
\end{aligned}$$

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

The proof is given in Appendix B. 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 exactly the same as in the literature such as Hansen [23], Pakes and Pollard [29] and Chen et al. [12].

The requirement of $\sqrt{n}\Sigma_n^{-1}(\mathbf{0}_r^\intercal, \mathscr{F}'(g)\gamma_k(z)^\intercal)^\intercal = o(1)$ is an undersmoothing condition, playing a similar role to its counterpart in the literature, see, for example, the condition

$\sqrt{n}V_K^{-1}K^{-p/d} = o(1)$ in Corollary 3.1 of Chen and Christensen [10, p. 454] and Comment 4.3 of Belloni et al. [4]. If $r = 1$, the transformation $\mathscr{L}$ will transform the vector α into a scalar, $\mathscr{L}(\alpha) = a_0^\intercal\alpha$, for some $a_0 \in \mathbb{R}^p$ and $a_0 \neq 0$. This is the case commonly encountered in the literature. See, e.g. Chang et al. [9] and Belloni et al. [4].

It is clear that the convergence order of $\mathscr{L}(\widehat{\alpha} - \alpha)$ is $n^{-1/2}$, while that of $\mathscr{F}(\widehat{g}) - \mathscr{F}(g)$ is proportional to $(\mathscr{F}'(g)\Phi_k(z)^\intercal\mathscr{F}'^\intercal\Phi_k(z))^{1/2}n^{-1/2}$, which is similar to the result in Theorem 2 of Newey [26]. Here, the matrix in the front of $n^{-1/2}$ is of dimension $s \times s$ associated with the derivative of functional $\mathscr{F}$. To understand how it affects the rate, consider a special case that $s = 1$ 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 rates of convergence in the literature. See, for example, Dong and Linton [14].

The result in above theorem does not rule out the 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. However, this would reduce the convergence rate.

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 consistent estimator, $\widehat{\Sigma}_n$, of Σ_n can be obtained if 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. 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 $\mathscr{F}'(g)$ by $\mathscr{F}'(\widehat{g})$ and

$$\widehat{\Xi}_n := \frac{1}{n}\sum_{i=1}^n [m(V_i, \widehat{\alpha}^\intercal X_i, \widehat{g}(Z_i))m(V_i, \widehat{\alpha}^\intercal X_i, \widehat{g}(Z_i))^\intercal], \tag{3.1}$$

$$\widehat{\Psi}_n := \frac{1}{n}\sum_{i=1}^n \begin{pmatrix} \frac{\partial}{\partial u}m(V_i, \widehat{\alpha}^\intercal X_i, \widehat{g}(Z_i))^\intercal \otimes X_i \\ \frac{\partial}{\partial w}m(V_i, \widehat{\alpha}^\intercal X_i, \widehat{g}(Z_i))^\intercal \otimes \Phi_k(Z_i) \end{pmatrix}. \tag{3.2}$$

It is readily seen that $\widehat{\Sigma}_n - \Sigma_n \to_P 0$ as $n \to \infty$.

If a weight matrix is used in the minimization, the efficiency of the limit theorem may be improved. Let $W_n = W_n(\alpha, \beta)$ be a $q \times q$ positive definite matrix depending on the parameter and the data used in M_n. 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 $M_n(\mathbf{a}, \mathbf{b})^\intercal W_n(\mathbf{a}, \mathbf{b})M_n(\mathbf{a}, \mathbf{b})$ 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. Definitely, W_n can not be close to singular to eschew the possibility that $M_n(\mathbf{a}, \mathbf{b})^\intercal W_n(\mathbf{a}, \mathbf{b})M_n(\mathbf{a}, \mathbf{b})$ 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 that $\sup_{\|(\mathbf{a}-\alpha,\mathbf{b}-\beta)\|<\delta_n} \|W_n(\mathbf{a},\mathbf{b}) - W_n\| = o_P(1)$ with $\delta_n = o(1)$ when n large. Let $(\widetilde{\alpha},\widetilde{\beta})$ be the minimizer of $M_n(\mathbf{a},\mathbf{b})^\intercal W_n(\mathbf{a},\mathbf{b}) M_n(\mathbf{a},\mathbf{b})$ and define $\widetilde{g}(z) = \Phi_k(z)^\intercal \widetilde{\beta}$.*

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

$$\Gamma_n[\Psi_n W_n \Psi_n^\intercal]^{-1}\Psi_n W_n \Xi_n W_n \Psi_n^\intercal [\Psi_n W_n \Psi_n^\intercal]^{-1}\Gamma_n^\intercal.$$

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

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

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

for all W satisfying the conditions in the proposition. In practice, both Ξ_n and Ψ_n can be replaced by their sample versions of (3.1) and (3.2), so that the optimal covariance matrices are easily estimable. Nonetheless, in order to obtain an optimal estimator 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 $\widehat{W_n}$; then at the second step one may minimize $M_n(\mathbf{a},\mathbf{b})^\intercal \widehat{W}_n M_n(\mathbf{a},\mathbf{b})$ to have a pair of optimal estimators, $(\widetilde{\alpha},\widetilde{g}(\cdot))$.

In addition to their earlier work by Cattaneo et al. [7] on a partially linear model, Cattaneo et al. [8] 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. We interpret the differencing approach proposed by Yatchew [34] and Yatchew [35] for the partially linear model as being similar to this, except that Cattaneo et al. [8] allow for heteroskedasticity and for a more complex type of nuisance component. In our methodology we pay equal attention to the function g, which itself can be of interest, see for example, Engle et al. [16]; Robinson [33]; Gao and Liang [19]; Gao and Shi [20] and Härdle et al. [24]. Our methodology is also robust to conditional heteroskedasticity.

4 Statistical inference

4.1 Test of over-identification

Hansen [23] 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} distribution. In the case where an unknown infinite dimensional parameter is involved but both p and q are still fixed with $q > p$, Chen and Liao [11] establish a statistic for over-identification testing that has an F-distribution in large samples. As far as we are aware, the test statistic proposed below seems a new one in the literature.

Because we also face an over-identification situation where $q(n) - p(n) \to \infty$, it is crucial to test whether the moment restrictions are valid by investigating the following hypotheses:

$$H_0 : \quad \mathbb{E}[m(V_i, \alpha^\intercal X_i, g(Z_i))] = 0 \ \text{ for some } (\alpha, g) \in \Theta,$$

versus

$$H_1 : \quad \mathbb{E}[m(V_i, \mathbf{a}^\intercal X_i, h(Z_i))] \neq 0 \ \text{ for any } (\mathbf{a}, h) \in \Theta,$$

where Θ is defined in Section 2.

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

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

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

Under the null, by the procedure in Section 2 and Assumptions in Theorem 3.1, we have the consistent estimator $(\widehat{\alpha}, \widehat{g})$. 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 below. It is noteworthy that this test, as clearly indicated from the proof, is also workable for the conventional moment restriction models with fixed p and q. Before showing the asymptotic distribution under the null and the consistency under the alternative for the test statistic, we introduce some necessary assumptions.

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

Assumption 4.2 Suppose that (i) $q^2 p = o(n)$ and $q^2 k = o(n)$; 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 $E[m(V_i, \alpha^\intercal X_i, g(Z_i))] = 0$, Assumption 4.1 requires that $E[m(V_i, \mathbf{a}^\intercal X_i, f(Z_i))]$ drops to zero very quickly when $(\mathbf{a}, f)$ approaches (α, g). This in spirit is the same as Assumption 3.2 but here it is a sample version and the decay of the expectation needs a certain rate. Similar assumption is also imposed by equation (4.9) of Andrews [1, p. 58] and equation (5.2) of Belloni et al. [5, p. 774]. Assumption 4.2 (i) stipulates the relationships for p, q, k with 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 particularly is satisfied for the case where z is located in some compact set in many situations 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, under H_0 and the conditions in Theorems 3.1 and 3.2 remain true. For any $\kappa \in \mathbb{R}^q$ such that $\|\kappa\| = 1$,*

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

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

Notice that if there is a zero function in m, $\kappa^\intercal m$ can be a zero function for some particular κ. Thus, the requirement on the nonzero function is trivial. The theorem shows the normality of the proposed statistic that enables us to make statistical inference.

Theorem 4.2. *Suppose that the eigenvalues of $E[m(V_1, \mathbf{a}^\intercal X_1, h(Z_1))m(V_1, \mathbf{a}^\intercal X_1, h(Z_1))^\intercal]$ are uniformly bounded away from zero and infinity in n and $(\mathbf{a}, h) \in \Theta$. Under H_1, suppose further that there exists a positive sequence δ_n such that $\inf_{(\mathbf{a},h)\in\Theta} \|E[m(V_i, \mathbf{a}^\intercal X_i, h(Z_i))]\| \geq \delta_n$ and $\liminf_{n\to\infty} \sqrt{n}\delta_n = \infty$. Then, for any vector $\mathbf{a}$ 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 broadly adopted in the literature, see, e.g. Chang et al. [9] and Belloni et al. [4]. 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. [9, 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 Wald test

The normality in Theorem 3.2 may be used for Wald test to detect, for example, $\mathcal{H}_{10} : \mathscr{L}\alpha = R$ against $\mathcal{H}_{11} : \mathscr{L}\alpha \neq R$ for some transformation $\mathscr{L}$ of $r \times p$ matrix and r-vector R; $\mathcal{H}_{20} : \mathscr{F}(g) = S$ against $\mathcal{H}_{21} : \mathscr{F}(g) \neq S$ for some s-vector functional $\mathscr{F}$ and s-vector S.

Let $\widehat{\Sigma}^2_{1n} = (I_r \;\; \mathbf{0}_{r\times s})\widehat{\Sigma}^2_n(I_r \;\; \mathbf{0}_{r\times s})^\intercal$ and $\widehat{\Sigma}^2_{2n} = (\mathbf{0}_{s\times r} \;\; I_s)\widehat{\Sigma}^2_n(\mathbf{0}_{s\times r} \;\; I_s)^\intercal$. Then, Theorem 3.2 implies that, under $\mathcal{H}_{10}$ and $\mathcal{H}_{20}$, respectively, we have

$$n(\mathscr{L}\widehat{\alpha} - R)^\intercal \widehat{\Sigma}^{-1}_{1n}(\mathscr{L}\widehat{\alpha} - R) \xrightarrow{d} \chi^2(r), \text{ and } \; n(\mathscr{F}(\widehat{g}) - S)^\intercal \widehat{\Sigma}^{-1}_{2n}(\mathscr{F}(\widehat{g}) - S) \xrightarrow{d} \chi^2(s),$$

respectively, when $n \to \infty$.

5 Penalised GMM for model selection

In the high dimensional situation, the parameter α most likely have sparsity, particularly in the ultra high dimensional case (i.e. $p = e^{n^a}$ with $0 < a < 1$). That is, there are plenty of zeros in α while only a number of elements are nonzero. In addition, the coefficient vector β may also possess sparsity, since some of them may be zero, as argued in Belloni et al. [5, p. 761], let alone the fact of their attenuation such that coefficients are negligible statistically when complexity is increasing. Hence, this section is devoted to the estimate of (α, g) under the sparsity.

In the literature, there are some studies on the variable selection in the frame work of GMM and sparsity. Belloni et al. [5] propose the combination of least squares and L_1 type lasso approach to select variables. While in a high dimensional conditional moment restriction model, Fan and Liao [18] propose to use folded concave penalty function combined with instrument variables. Caner [6] uses the same approach but a particular class of penalty functions to select variables. As Caner [6, p. 271] argued, Lasso-type GMM estimator selects correct model much more often compared with GMM-BIC and "downward testing" method proposed by Andrews and Lu [2], we shall tackle the selection issue by the penalty function in our GMM framework.

For convenience in this section, denote $v_0 = (\alpha^\intercal, \beta^\intercal)^\intercal \in \mathbb{R}^{p+k}$ the true parameter whose dimension varies with the sample size. Trivially, we may write $\alpha = (\alpha_{0S}^\intercal, \alpha_{0N}^\intercal)^\intercal$ and $\beta = (\beta_{0S}^\intercal, \beta_{0N}^\intercal)^\intercal$, where the vectors α_{0S} and β_{0S} contain all "important coefficients" from α and β, respectively, as referred in the literature such as Fan and Liao [18], while α_{0N} and β_{0N} are identical to zero. In addition, $v_{0S} = (\alpha_{0S}^\intercal, \beta_{0S}^\intercal)^\intercal$ is referred to as an oracle model. Define $t = |v_{0S}|$ 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,

$$\widehat{v} = (\widehat{\alpha}^\intercal, \widehat{\beta}^\intercal)^\intercal = \underset{v=(\mathbf{a}^\intercal, \mathbf{b}^\intercal)^\intercal \in \mathbb{R}^{p+k}}{\operatorname{argmin}} Q_n(v) := \|M_n(v)\|^2 + \sum_{j=1}^{p+k} P_n(|v_j|), \quad \text{subject to } \|\mathbf{a}\| \le B_{1n} \text{ and } \|\mathbf{b}\| \le B_{2n}. \tag{5.1}$$

where $M_n(v) = M_n(\mathbf{a}, \mathbf{b})$, B_{1n} and B_{2n} are the same as in Section 2 and $P_n(\cdot)$ is a penalty function discussed later.

5.1 Oracle Property

Let T be the support of v_0, i.e. $T = \{j : 1 \le j \le p+k, v_{0j} \neq 0\}$. We may equivalently say 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 j-th element is that of v for $j \in T$ and zero otherwise. Also, define v_S as the short

version of v_T after eliminating all zeros in the position T^c (complement set of T) from v_T. In the literature, the subspace $\mathcal{V} = \{v_T, v \in \mathbb{R}^{p+k}\}$ is called "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 vector consisting of all nonzero elements in v_T). Hence, the vector $S_{nT}(v_S)$ has dimension $t = |T| = |v_S|$. Also, define in a similar fashion $H_{nT}(v_S)$ the Hessian matrix of $t \times t$ for $\|M_n(v)\|^2$.

Suppose that $P_n(\cdot)$ belongs to the class of folded concave penalty functions in Fan and Li [17]. For any $v = (v_1, \cdots, v_t)^\intercal \in \mathbb{R}^t$ with $v_j \neq 0$, $\forall j$, define

$$\phi(v) = \limsup_{\epsilon \to 0^+} \max_{j \leq t} \sup_{(u_1,u_2) \in O(|v_j|,\epsilon)} -\frac{P_n'(u_2) - P_n'(u_1)}{u_2 - u_1},$$

where $O(\cdot,\cdot)$ is the neighbourhood with specified center and radius, respectively, implying that $\phi(v) = \max_{j \leq t} -P_n''(|v_j|)$ if P_n'' is continuous. Also, for the true parameter v_0, let

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

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)$\sqrt{t}\, P_n'(d_n) = o(d_n)$; (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 properly chosen tuning parameter, the L_r penalty ($0 < r \leq 1$), hard-thresholding (Antoniadis [3]), SCAD (Fan and Li [17]) and MCP (Zhang [36]) satisfy the requirements.

Denote 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 the decomposition $t = p_1 + k_1$ for their cardinalities.

Assumption 5.2 *Suppose that Assumptions A.1-A.3 hold with p being replaced by p_1 and k by k_1.*

The assumption is the counterpart of Assumptions A.1-A.3 under sparsity. We first show an oracle asympotic property about $\widehat{v}$ in the minimization of (5.1).

Lemma 5.1. *In addition to Assumptions 5.1-5.2, 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 $\|\widehat{v} - v_0\| = O_P(a_n + \sqrt{t}\, P_n'(d_n))$. *Moreover, for any arbitrary* $\epsilon > 0$, *the local minimizer* $\widehat{v}$ *is strict with probability at least* $1 - \epsilon$ *for all large* n.

The proof and the verification on the conditions of the lemma are relegated to Appendix B. It is worth noting that, under additional condition stated below, we show in Appendix B that $\|S_{nT}(v_{0S})\| = O_P(\sqrt{t\,\log(q)/n})$ and therefore we have $\|\widehat{v} - v_0\| = O_P(\sqrt{t\,\log(q)/n} + \sqrt{t}\, P_n'(d_n))$.

The oracle consistency in Lemma 5.1 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 5.2. *Addition to the conditions in Lemma 5.1, suppose that with probability approaching one, for* $\widehat{v} \in \mathcal{V}$ *in Lemma 5.1, 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*

$$\|M_n(v_T)\|^2 - \|M_n(v)\|^2 < \sum_{j \notin T} P_n(|v_j|). \tag{5.2}$$

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* $\widehat{v} \in \mathcal{V}$ *is strict with probability at least* $1 - \epsilon$ *for all large* n.

The proof and the verification on the conditions of the lemma are relegated to Appendix B.

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

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

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

This assumption is often encountered in the literature such as Assumption 4.3 in Fan and Liao [18]. It is known that there are many classes of distributions satisfying this condition, e.g. normal distribution and exponential distribution and so on.

For simplicity, denote ∂m the partial derivative of m; and $F_{iS} = \mathrm{diag}(X_{iS}, \Phi_{kS}(Z_i))$ a $t \times 2$ 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 $\lambda_{\min}(E\partial m^\intercal(V_i, v_{0S}^\intercal F_{iS}) \otimes F_{iS})(E\partial m^\intercal(V_i, v_{0S}^\intercal F_{iS}) \otimes F_{iS})^\intercal) > C_1$ and $\lambda_{\max}(E\partial m^\intercal(V_i, v_{0S}^\intercal F_{iS}) \otimes F_{iS})(E\partial m^\intercal(V_i, v_{0S}^\intercal F_{iS}) \otimes F_{iS})^\intercal) < C_2$; (ii) $P_n'(d_n) = o(n^{-1/2})$ and $\max_{\|v_S - v_{0S}\| < d_n/4} \phi(v_S) = o((t\log(q))^{-1/2})$; (iii) $t^{3/2}\log(q) = o(n)$, $t^{3/2}P_n'(d_n)^2 = o(1)$, $t\max_{j\in T} P_n(|v_{0j}|) = o(1)$.

All these are technical requirements on the Hessian matrix, the penalty function, the relationship among the dimensions of important coefficients, the sparsity and the sample size. 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 $t\sqrt{\log(q)/n} + t^{3/2}\log(q)/n \ll \lambda_n \ll d_n$.

Theorem 5.1. *Under Assumptions 2.1, 2.2, 3.1, 3.3 and 5.1-5.4, 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 $\widehat{T} = \{j : 1 \le j \le p + k, \widehat{v}_j \neq 0\}$. *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* $\mathscr{F}$*,*

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

as $n \to \infty$ *provided that* $\sqrt{n}\Sigma_{nT}^{-1}(0_r^\intercal, \mathscr{F}'(g)\gamma_k(z)^\intercal)^\intercal = o(1)$, *where* Σ_{nT} *is given by the square root of*

$$\begin{aligned}
\Sigma_{nT}^2 :=& \Gamma_n[\Psi_{nT}\Psi_{nT}^\intercal]^{-1}\Psi_{nT}\Xi_{nT}\Psi_{nT}^\intercal[\Psi_{nT}\Psi_{nT}^\intercal]^{-1}\Gamma_n^\intercal, \quad \textit{in which} \\
\Gamma_n :=& \begin{pmatrix} \mathscr{L} & 0 \\ 0 & \mathscr{F}'(g)\Phi_{kT}(\cdot)^\intercal \end{pmatrix}_{(r+s)\times(p_1+k_1)}, \\
\Xi_{nT} :=& \mathbb{E}[m(V_1, \alpha_{0S}^\intercal X_{1S}, g(Z_1))m(V_1, \alpha_{0S}^\intercal X_{1S}, g(Z_1))^\intercal]_{q\times q}, \\
\Psi_{nT} :=& \mathbb{E}\begin{pmatrix} \frac{\partial}{\partial u}m(V_1, \alpha_{0S}^\intercal X_{1S}, g(Z_1))^\intercal \otimes X_{1S} \\ \frac{\partial}{\partial w}m(V_1, \alpha_{0S}^\intercal X_{1S}, g(Z_1))^\intercal \otimes \Phi_{kT}(Z_1) \end{pmatrix}_{(p_1+k_1)\times q},
\end{aligned}$$

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.*

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 shall 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 which excludes all the other points to the minimizer of the objective function in population sense.

Theorem 5.2. *Under Assumption 3.2 and those of Theorem 5.1, the local minimizer* $\widehat{v}$ *satisfies that, for any* $\delta > 0$*, there exists* $\eta > 0$ *such that*

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

Therefore, as indicated by Theorems 5.1 and 5.2, the minimization of (5.1) enables to recover the sparsity in ultra high dimensional moment restriction models since $q > p + k$ where q can be as large as $e^{n^{\epsilon}}$ for some $\epsilon > 0$. This is a bit different from Fan and Liao [18] where there is no nonparametric function involved and $q = p$ (the number of IV is the same as that of regressors). The consistency of the sparsity is given, and more importantly, the inference can be done similar to the relative lower dimensional models (Theorem 3.2).

6 Simulation experiments

In this section, we are about to investigate the performance of the proposed estimators in finite sample situation.

Example 6.1. This experiment uses the model in Example 1.1 of Section 1. Let vector $X_i = (X_{1i}, X_{2i}^{\intercal})^{\intercal}$ where X_{1i} assumes 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 in continuous variables. Let Z_i be uniformly distributed on $(0,1)$.

Suppose that $\mathbb{E}[Y_i - \alpha^{\intercal}X_i - g(Z_i)|W_i] = 0$ 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 $j \geq 1$, $\varphi_j(r) = \sqrt{2}\cos(\pi jr)$. Then, $\{\varphi_j(r)\}$ is an orthonormal basis in the Hilbert space $L^2[0,1]$. In the experiment, put $\alpha = (0.4, 0.1, 0, \cdots, 0)^{\intercal} \in \mathbb{R}^p$ and $g(z) = z^2 + \sin(z)$.

Denote $m(V_i, \alpha^\intercal X_i, g(Z_i)) = (Y_i - \alpha^\intercal X_i - g(Z_i))\Phi_q(Z_i)$ where $V_i = (Y_i, W_i)$, $W_i = Z_i$ and $\Phi_q(\cdot) = (\varphi_0(\cdot), \cdots, \varphi_{q-1}(\cdot))^\intercal$. Notice that the dimension of m function is q which increases with the sample size n. Thus, the parameter (α, g) can be solved from unconditional moment equations $\mathbb{E}[m(V_i, \alpha^\intercal X_i, g(Z_i))] = 0$ for $i = 1, \cdots, n$.

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 $M_n(\mathbf{a}, \mathbf{b}) = \frac{1}{\sqrt{q}}\frac{1}{n}\sum_{i=1}^n m(V_i, \mathbf{a}^\intercal X_i, \mathbf{b}^\intercal \Phi_k(Z_i))$. Thus, $\widehat{\alpha}$ and $\widehat{g}(\cdot) := \widehat{\beta}^\intercal \Phi_k(\cdot)$ are the estimates of $(\alpha, g(\cdot))$.

Here, we emphasize that since the m function is linear in both $\alpha^\intercal X_i$ and $g(Z_i)$, $M_n(\mathbf{a}, \mathbf{b})$ actually has a linear relationship with $\mathbf{a}$ and $\mathbf{b}$,

$$
\begin{aligned}
M_n(\mathbf{a}, \mathbf{b}) =& \frac{1}{\sqrt{q}}\frac{1}{n}\sum_{i=1}^n (Y_i - \mathbf{a}^\intercal X_i - \mathbf{b}^\intercal \Phi_k(Z_i))\Phi_q(Z_i) \\
=& \frac{1}{\sqrt{q}}\frac{1}{n}\sum_{i=1}^n Y_i \Phi_q(Z_i) - \left(\frac{1}{\sqrt{q}}\frac{1}{n}\sum_{i=1}^n \Phi_q(Z_i) X_i^\intercal\right)\mathbf{a} - \left(\frac{1}{\sqrt{q}}\frac{1}{n}\sum_{i=1}^n \Phi_q(Z_i)\Phi_k(Z_i)^\intercal\right)\mathbf{b}.
\end{aligned}
$$

Accordingly, $(\widehat{\alpha}, \widehat{\beta})$ has an explicit expression simply by OLS. This means that in any similar situation the optimization in Section 2 does not need the compact restrictions.

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 experiments) satisfy Assumption 3.1. The replication number of the experiment is $M = 1000$. We shall report for the estimate of the g function the bias (denoted by $B_g(n)$), standard deviation (denoted by $\pi_g(n)$) and RMSE (denoted by $\Pi_g(n)$), that is,

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

where the superscript ℓ indicates the ℓ-th replication, $\overline{\widehat{g}}(\cdot) = \Phi_k(\cdot)^\intercal \overline{\widehat{\beta}}$ is the average of $\widehat{g}^\ell(\cdot)$ over Monto Carlo replications $\ell = 1, \cdots, M$, $g^\ell(Z_i)$ means the value of g evaluated for the Z_i in the ℓ-th replication.

Regarding of 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 Monto 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.

Table 1: Simulation of Example 5.1, $q = p + k + \nu$

	$\nu = 2$				$\nu = 4$		
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

It can be seen that all of the statistical quantities about the estimator of g are reasonably attenuated with the increase of both the sample size and ν that provides more information for the parameters to be estimated. For the quantities about the estimator 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 both quantities decrease with the increase of ν that gives more moment restrictions.

This is understandable. Because the conditional moment $\mathbb{E}[Y_i - \alpha^\intercal X_i - g(Z_i)|Z_i]$ determines a function $U(z) := \mathbb{E}[Y_i - \alpha^\intercal X_i - g(Z_i)|Z_i = z]$ and $\{\varphi_j(z)\}$ is an orthornomal 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 $U(z)$.

Example 6.2. This experiment uses the model in Example 1.3 in Section 1. Let the distribution function $F(u) = \exp(u)/[1 + \exp(u)]$. Y_i assumes either 0 or 1, and

$$P(Y_i = 1|X_i, Z_i) = F(\alpha^\intercal X_i + g(Z_i)),$$

for $i = 1, \cdots, n$, where $\alpha, X_i \in \mathbb{R}^p$ and $Z_i \in \mathbb{R}$. Here, let $X_i \sim N(0, \Sigma_x)$, where $\Sigma_x = (\sigma_{i,j})_{p\times p}$

with $\sigma_{i,i} = 1$, $\sigma_{i,j} = 0.5$ for $|i-j| = 1$ and $\sigma_{i,j} = 0$ for $|i-j| > 1$, and $Z_i \sim N(0,1)$. In this experiment, put $\alpha = (0.5, 0.3, 0, \cdots, 0)^\intercal \in \mathbb{R}^p$ and $g(z) = z^2 + \sin(z)$. The Hilbert space that contains $g(\cdot)$ is $L^2(\mathbb{R}, \exp(-z^2))$. Let $\{p_j(z), j \geq 0\}$ be the sequence of Hermite polynomials that forms an orthonormal basis in $L^2(\mathbb{R}, \exp(-z^2))$.

Denote $\Phi_k(z) = (p_0(z), \cdots, p_{k-1}(z))^\intercal$ and similarly to Example 1.3, define

$$Q_n(\alpha, \beta) := \ln \prod_{i=1}^{n} F^{Y_i}(\alpha^\intercal X_i + \beta^\intercal \Phi_k(Z_i))[1 - F(\alpha^\intercal X_i + \beta^\intercal \Phi_k(Z_i))]^{1-Y_i},$$
$$M_n(\alpha, \beta) := \left(\frac{\partial Q_n}{\partial \alpha^\intercal}, \frac{\partial Q_n}{\partial \beta^\intercal}\right)^\intercal$$

and we have $(\widehat{\alpha}, \widehat{\beta}) = \underset{\mathbf{a}\in\mathbb{R}^p, \mathbf{b}\in\mathbb{R}^k}{\operatorname{argmin}} \|M_n(\mathbf{a}, \mathbf{b})\|^2$ and naturally $\widehat{g}(\cdot) := \widehat{\beta}^\intercal \Phi_k(\cdot)$ is the estimate of $g(\cdot)$.

For $n = 200, 500$ and 1000, let $k = [C_1 n^{\tau_1}]$ and $p = [C_2 n^{\tau_2}]$ where C_i and τ_i, $i = 1, 2$, take the same values as in the preceding example. The replication number of the experiment is $M = 1000$. We shall report the bias $B_g(n)$, standard deviation $\pi_g(n)$ and RMSE $\Pi_g(n)$ for the estimator of g and $B_\alpha(n)$ and $M_\alpha(n)$ for the estimator of α defined in the above example.

Table 2: Simulation results for Example 5.2

n	300	600	1000	n	300	600	1000
$B_\alpha(n)$	0.0130	0.0105	0.0065	$B_g(n)$	-0.0100	0.0059	0.0037
$M_\alpha(n)$	0.0125	0.0103	0.0075	$\pi_g(n)$	0.3608	0.3128	0.2315
				$\Pi_g(n)$	0.3320	0.2323	0.1732

The moment restriction model is exactly identified, since it is formulated from the partial derivatives that imply $q = p + k$. All results in Table 2 converge satisfactorily, though it seems in this example the estimator of the g function converge a bit slower than the last example. This might be because in the last example there is an explicit solution while this example needs a minimization of the nonlinear distribution function to have the estimators.

7 Conclusion

A class of high dimensional semiparametric moment restriction models have been studied using the GMM and sieve method. The consistency and normality of the proposed estimators

have obtained. A new test statistic has been proposed for over-identification testing where the strong signal can be weakened when our test statistic is used. In addition, the potential sparsity of the model has been dealt with via the combination of GMM methodology and penalty function approach. The theoretical results are verified through finite sample experiments. We find that the more the number of the moment restrictions, the accurate the estimates, although it may not be interesting to compare the estimates of the high–dimensional parameters for different sample sizes.

8 Acknowledgement

We acknowledge Professor Xiaohong Chen for her insightful suggestion and pointing out some relevant references. The first author thanks the financial support from National Natural Science Foundation of China under grant No. 71671143. The second author is supported by the Australian Research Council Discovery Grants Program for its support under Grant numbers: DP150101012 & DP170104421.

A Lemmas

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

1. $\|M_n(\alpha,\beta)\|^2 = O_P(\|\gamma_k(z)\|^2) + O_P(n^{-1})$.
2. *Given* $B_{1n}^2 + B_{2n}^2 = o(n)$, $\sup_{\substack{\|\mathbf{a}\|\le B_{1n}, \|\mathbf{b}\|\le B_2 \\ \|(\mathbf{a}-\alpha,\mathbf{b}-\beta)\|>\delta}} \|M_n(\mathbf{a},\mathbf{b})\|^{-2} = O_P(1)$ *for each* $\delta > 0$, *when* n *is large.*

Proof. 1. Observe that

$$\begin{aligned}\|M_n(\alpha,\beta)\|^2 =& \left\| \frac{1}{\sqrt{q}} \frac{1}{n} \sum_{i=1}^n m(V_i, \alpha^\intercal X_i, \beta^\intercal \Phi_k(Z_i)) \right\|^2 \\ =& \frac{1}{q} \sum_{\ell=1}^q \left[\frac{1}{n} \sum_{i=1}^n m_\ell(V_i, \alpha^\intercal X_i, \beta^\intercal \Phi_k(Z_i)) \right]^2,\end{aligned}$$

where we denote $m(\cdots) = (m_1(\cdots), \cdots, m_q(\cdots))^\intercal$. Moreover,

$$\begin{aligned}&\frac{1}{q} \sum_{\ell=1}^q \mathbb{E} \left[\frac{1}{n} \sum_{i=1}^n m_\ell(V_i, \alpha^\intercal X_i, \beta^\intercal \Phi_k(Z_i)) \right]^2 \\ =& \frac{1}{q} \sum_{\ell=1}^q \left[\frac{1}{n} \sum_{i=1}^n \mathbb{E} m_\ell(V_i, \alpha^\intercal X_i, \beta^\intercal \Phi_k(Z_i)) \right]^2 + \frac{1}{q} \sum_{\ell=1}^q \mathrm{Var} \left[\frac{1}{n} \sum_{i=1}^n m_\ell(V_i, \alpha^\intercal X_i, \beta^\intercal \Phi_k(Z_i)) \right] \\ =& \frac{1}{q} \sum_{\ell=1}^q \left[\mathbb{E} m_\ell(V_1, \alpha^\intercal X_1, \beta^\intercal \Phi_k(Z_1)) \right]^2 + \frac{1}{q} \frac{1}{n^2} \sum_{\ell=1}^q \sum_{i=1}^n \mathrm{Var} \left[m_\ell(V_i, \alpha^\intercal X_i, \beta^\intercal \Phi_k(Z_i)) \right]\end{aligned}$$

$$
\begin{aligned}
=&\frac{1}{q}\left\|\mathbb{E}m(V_1,\alpha^\intercal X_1,\beta^\intercal\Phi_k(Z_1))\right\|^2+\frac{1}{q}\frac{1}{n}\sum_{\ell=1}^{q}\mathrm{Var}(m_\ell(V_1,\alpha^\intercal X_1,\beta^\intercal\Phi_k(Z_1)))\\
\le&\frac{1}{q}\left\|\mathbb{E}m(V_1,\alpha^\intercal X_1,\beta^\intercal\Phi_k(Z_1))\right\|^2+\frac{1}{n}\frac{1}{q}\sum_{\ell=1}^{q}\mathbb{E}(m_\ell^2(V_1,\alpha^\intercal X_1,\beta^\intercal\Phi_k(Z_1)))\\
=&\frac{1}{q}\left\|\mathbb{E}m(V_1,\alpha^\intercal X_1,\beta^\intercal\Phi_k(Z_1))\right\|^2+\frac{1}{n}\frac{1}{q}\mathbb{E}\|m(V_1,\alpha^\intercal X_1,\beta^\intercal\Phi_k(Z_1))\|^2
\end{aligned}
$$

due to the property of the i.i.d. sequence.

Since $\mathbb{E}[m(V_1,\alpha^\intercal X_1,g(Z_1))=0$, it follows from Assumption 3.3 that

$$
\begin{aligned}
&\frac{1}{q}\left\|\mathbb{E}m(V_1,\alpha^\intercal X_1,\beta^\intercal\Phi_k(Z_1))\right\|^2\\
=&\frac{1}{q}\left\|\mathbb{E}[m(V_1,\alpha^\intercal X_1,\beta^\intercal\Phi_k(Z_1))-m(V_1,\alpha^\intercal X_1,\beta^\intercal\Phi_k(Z_1)+\gamma_k(Z_1))]\right\|^2\\
\le&\{\mathbb{E}[A(V_1,X_1,Z_1)]|\gamma_k(Z_1)|\}^2\le\mathbb{E}[A(V_1,X_1,Z_1)]^2\mathbb{E}|\gamma_k(Z_1)|^2\\
\le&C\|\gamma_k(z)\|^2=o(1),
\end{aligned}
$$

by virtue of Assumption 3.1(b), and for the second term,

$$
\begin{aligned}
&\frac{1}{q}\mathbb{E}\|m(V_1,\alpha^\intercal X_1,\beta^\intercal\Phi_k(Z_1))\|^2\\
\le&2\frac{1}{q}\mathbb{E}\|m(V_1,\alpha^\intercal X_1,g(Z_1))\|^2+2\frac{1}{q}\mathbb{E}\|m(V_1,\alpha^\intercal X_1,\beta^\intercal\Phi_k(Z_1))-m(V_1,\alpha^\intercal X_1,g(Z_1))\|^2\\
=&O(1)+\mathbb{E}[A^2(V_1,X_1,Z_1)|\gamma_k(Z_1)|^2]=O(1)
\end{aligned}
$$

by the dominated convergence theorem, implying the second term is $O(n^{-1})$.

2. First, note that

$$
\begin{aligned}
&M_n(\mathbf{a},\mathbf{b})-\frac{1}{\sqrt{q}}\mathbb{E}m(V_1,\mathbf{a}^\intercal X_1,\mathbf{b}^\intercal\Phi_k(Z_1))\\
=&\frac{1}{\sqrt{q}}\frac{1}{n}\sum_{i=1}^{n}[m(V_i,\mathbf{a}^\intercal X_i,\mathbf{b}^\intercal\Phi_k(Z_i))-\mathbb{E}m(V_i,\mathbf{a}^\intercal X_i,\mathbf{b}^\intercal\Phi_k(Z_i))].
\end{aligned}
$$

It follows from the property of i.i.d. sequence and Assumption 3.3 that

$$
\begin{aligned}
&\mathbb{E}\left\|M_n(\mathbf{a},\mathbf{b})-\frac{1}{\sqrt{q}}\mathbb{E}m(V_1,\mathbf{a}^\intercal X_1,\mathbf{b}^\intercal\Phi_k(Z_1))\right\|^2\\
=&\frac{1}{n^2}\sum_{i=1}^{n}\frac{1}{q}\mathbb{E}\|m(V_i,\mathbf{a}^\intercal X_i,\mathbf{b}^\intercal\Phi_k(Z_i))-\mathbb{E}m(V_i,\mathbf{a}^\intercal X_i,\mathbf{b}^\intercal\Phi_k(Z_i))\|^2\\
\le&\frac{1}{n}\frac{1}{q}\mathbb{E}\|m(V_1,\mathbf{a}^\intercal X_1,\mathbf{b}^\intercal\Phi_k(Z_1))\|^2=O(n^{-1}(B_{1n}^2+B_{2n}^2)),
\end{aligned}
$$

uniformly in $(\mathbf{a},\mathbf{b}^\intercal\Phi_k(z))\in\Theta_n$ by Assumption 3.3, which implies by the triangle inequality that

$$
\left|\|M_n(\mathbf{a},\mathbf{b})\|-\frac{1}{\sqrt{q}}\|\mathbb{E}m(V_1,\mathbf{a}^\intercal X_1,\mathbf{b}^\intercal\Phi_k(Z_1))\|\right|
$$

$$\leq \left\| M_n(\mathbf{a},\mathbf{b}) - \frac{1}{\sqrt{q}}\mathbb{E}m(V_1, \mathbf{a}^\intercal X_1, \mathbf{b}^\intercal \Phi_k(Z_1)) \right\| = O_P(n^{-1/2}(B_{1n}+B_{2n})),$$

that is, $\|M_n(\mathbf{a},\mathbf{b})\| = \frac{1}{\sqrt{q}}\|\mathbb{E}m(V_1, \mathbf{a}^\intercal X_1, \mathbf{b}^\intercal \Phi_k(Z_1))\| + O_P(n^{-1/2}(B_{1n}+B_{2n}))$ where the last term is independent of $(\mathbf{a},\mathbf{b})$. This is equivalent to $\|M_n(\mathbf{a},\mathbf{b})\|^2 = \frac{1}{q}\|\mathbb{E}m(V_1, \mathbf{a}^\intercal X_1, \mathbf{b}^\intercal \Phi_k(Z_1))\|^2 + O_P(n^{-1}(B_{1n}^2 + B_{2n}^2))$ by basic algebra.

Second, for any $\|\mathbf{b}\|^2 \leq B_{2n}$, we have $\mathbf{b}^\intercal \Phi_k(z) \in \Theta_{2n}$. Also, $\|\mathbf{b}^\intercal \Phi_k(z) - g(z)\|^2 = \|\mathbf{b} - \beta\|^2 + \|\gamma_k(z)\|^2$ by the orthogonality of the basis sequence.

For any $\delta > 0$, let n be large (so k large) such that $\delta > \|\gamma_k(z)\|$. Moreover, by Assumption 3.2, regarding of this $\delta > 0$ there exists an $\epsilon > 0$ such that

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

Notice further that

$$\begin{aligned}
&\inf_{\substack{\|\mathbf{a}\|\leq B_{1n}, \|\mathbf{b}\|\leq B_{2n} \\ \|(\mathbf{a}-\alpha, \mathbf{b}-\beta)\|\geq\delta}} \frac{1}{q}\|\mathbb{E}m(V_i, \mathbf{a}^\intercal X_i, \mathbf{b}^\intercal \Phi_k(Z_i))\|^2 \\
=& \inf_{\substack{\|\mathbf{a}\|\leq B_{1n}, \|\mathbf{b}\|\leq B_{2n} \\ \|\mathbf{a}-\alpha\|^2 + \|\mathbf{b}-\beta\|^2\geq\delta^2}} \frac{1}{q}\|\mathbb{E}m(V_i, \mathbf{a}^\intercal X_i, \mathbf{b}^\intercal \Phi_k(Z_i))\|^2 \\
\geq& \inf_{\substack{\|\mathbf{a}\|\leq B_{1n}, \|\mathbf{b}\|\leq B_{2n} \\ \|\mathbf{a}-\alpha\|^2 + \|\mathbf{b}-\beta\|^2\geq\delta^2 - \|\gamma_k(z)\|^2}} \frac{1}{q}\|\mathbb{E}m(V_i, \mathbf{a}^\intercal X_i, \mathbf{b}^\intercal \Phi_k(Z_i))\|^2 \\
\geq& \inf_{\substack{(\mathbf{a}, \mathbf{b}^\intercal \Phi_k(z))\in\Theta_n \\ \|\mathbf{a}-\alpha\|^2 + \|\mathbf{b}^\intercal \Phi_k(z) - g(z)\|^2\geq\delta^2}} \frac{1}{q}\|\mathbb{E}m(V_i, \mathbf{a}^\intercal X_i, \mathbf{b}^\intercal \Phi_k(Z_i))\|^2 \\
\geq& \inf_{\substack{(\mathbf{a},f)\in\Theta \\ \|\mathbf{a}-\alpha\|^2 + \|f-g\|^2\geq\delta^2}} \frac{1}{q}\|\mathbb{E}m(V_i, \mathbf{a}^\intercal X_i, f(Z_i))\|^2 \\
\geq& \inf_{\substack{(\mathbf{a},f)\in\Theta \\ \|(\mathbf{a}-\alpha, f-g)\|\geq\delta}} \frac{1}{q}\|\mathbb{E}m(V_i, \mathbf{a}^\intercal X_i, f(Z_i))\|^2 > \epsilon,
\end{aligned}$$

due to $\Theta_n \subset \Theta$, which, along with the approximation in the first part, is tantamount to the assertion.

□

Denote $m(v,u,w) = (m_1(v,u,w), \cdots, m_q(v,u,w))^\intercal$.

Since $\|M_n(\mathbf{a},\mathbf{b})\|^2 = \frac{1}{qn^2}\sum_{\ell=1}^q \left(\sum_{i=1}^n m_\ell(V_i, \mathbf{a}^\intercal X_i, \mathbf{b}^\intercal \Phi_k(Z_i))\right)^2$ we have

$$\begin{aligned}
\frac{\partial}{\partial \mathbf{a}}\|M_n(\mathbf{a},\mathbf{b})\|^2 =& 2\frac{1}{qn^2}\sum_{\ell=1}^q \sum_{i=1}^n m_\ell(V_i, \mathbf{a}^\intercal X_i, \mathbf{b}^\intercal \Phi_k(Z_i)) \\
&\times \sum_{j=1}^n \frac{\partial}{\partial u} m_\ell(V_j, \mathbf{a}^\intercal X_j, \mathbf{b}^\intercal \Phi_k(Z_j)) X_j
\end{aligned}$$

$$\frac{\partial}{\partial \mathbf{b}}\|M_n(\mathbf{a},\mathbf{b})\|^2 =2\frac{1}{qn^2}\sum_{\ell=1}^{q}\sum_{i=1}^{n} m_\ell(V_i,\mathbf{a}^\intercal X_i,\mathbf{b}^\intercal\Phi_k(Z_i)) \times\sum_{j=1}^{n}\frac{\partial}{\partial w}m_\ell(V_j,\mathbf{a}^\intercal X_j,\mathbf{b}^\intercal\Phi_k(Z_j))\Phi_k(Z_j),$$

and

$$\begin{aligned}\frac{\partial^2}{\partial \mathbf{a}\partial \mathbf{a}^\intercal}\|M_n(\mathbf{a},\mathbf{b})\|^2 =&2\frac{1}{qn^2}\sum_{\ell=1}^{q}\sum_{i=1}^{n}\sum_{j=1}^{n}\frac{\partial}{\partial u}m_\ell(V_i,\mathbf{a}^\intercal X_i,\mathbf{b}^\intercal\Phi_k(Z_i)) \times\frac{\partial}{\partial u}m_\ell(V_j,\mathbf{a}^\intercal X_j,\mathbf{b}^\intercal\Phi_k(Z_j))X_jX_i^\intercal\\ &+2\frac{1}{qn^2}\sum_{\ell=1}^{q}\sum_{i=1}^{n}\sum_{j=1}^{n} m_\ell(V_i,\mathbf{a}^\intercal X_i,\mathbf{b}^\intercal\Phi_k(Z_i)) \times\frac{\partial^2}{\partial u^2}m_\ell(V_j,\mathbf{a}^\intercal X_j,\mathbf{b}^\intercal\Phi_k(Z_j))X_jX_j^\intercal\\ \frac{\partial^2}{\partial \mathbf{a}\partial \mathbf{b}^\intercal}\|M_n(\mathbf{a},\mathbf{b})\|^2 =&2\frac{1}{qn^2}\sum_{\ell=1}^{q}\sum_{i=1}^{n}\sum_{j=1}^{n}\frac{\partial}{\partial w}m_\ell(V_i,\mathbf{a}^\intercal X_i,\mathbf{b}^\intercal\Phi_k(Z_i)) \times\frac{\partial}{\partial u}m_\ell(V_j,\mathbf{a}^\intercal X_j,\mathbf{b}^\intercal\Phi_k(Z_j))X_j\Phi_k(Z_i)^\intercal\\ &+2\frac{1}{qn^2}\sum_{\ell=1}^{q}\sum_{i=1}^{n}\sum_{j=1}^{n} m_\ell(V_i,\mathbf{a}^\intercal X_i,\mathbf{b}^\intercal\Phi_k(Z_i)) \times\frac{\partial^2}{\partial u\partial w}m_\ell(V_j,\mathbf{a}^\intercal X_j,\mathbf{b}^\intercal\Phi_k(Z_j))X_j\Phi_k(Z_j)^\intercal\\ \frac{\partial^2}{\partial \mathbf{b}\partial \mathbf{b}^\intercal}\|M_n(\mathbf{a},\mathbf{b})\|^2 =&2\frac{1}{qn^2}\sum_{\ell=1}^{q}\sum_{i=1}^{n}\sum_{j=1}^{n}\frac{\partial}{\partial w}m_\ell(V_i,\mathbf{a}^\intercal X_i,\mathbf{b}^\intercal\Phi_k(Z_i)) \times\frac{\partial}{\partial w}m_\ell(V_j,\mathbf{a}^\intercal X_j,\mathbf{b}^\intercal\Phi_k(Z_j))\Phi_k(Z_j)\Phi_k(Z_i)^\intercal\\ &+2\frac{1}{qn^2}\sum_{\ell=1}^{q}\sum_{i=1}^{n}\sum_{j=1}^{n} m_\ell(V_i,\mathbf{a}^\intercal X_i,\mathbf{b}^\intercal\Phi_k(Z_i)) \times\frac{\partial^2}{\partial w^2}m_\ell(V_j,\mathbf{a}^\intercal X_j,\mathbf{b}^\intercal\Phi_k(Z_j))\Phi_k(Z_j)\Phi_k(Z_j)^\intercal.\end{aligned}$$

The unimportant constant shall be ignored in what follows.

Denote each block of $H_n(\mathbf{a},\mathbf{b})$ by

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

and define

$$h_{11}(\alpha,g) :=\frac{1}{q}\sum_{\ell=1}^{q}\left(\mathbb{E}\frac{\partial}{\partial u}m_\ell(V_1,\alpha^\intercal X_1,g(Z_1))X_1\right)\left(\mathbb{E}\frac{\partial}{\partial u}m_\ell(V_1,\alpha^\intercal X_1,g(Z_1))X_1\right)^\intercal,$$

$$
\begin{aligned}
=&\frac{1}{q}\left[\mathbb{E}\left(\frac{\partial}{\partial u}m(V_1,\alpha^\intercal X_1,g(Z_1))\right)^\intercal\otimes X_1\right]\left[\mathbb{E}\left(\frac{\partial}{\partial u}m(V_1,\alpha^\intercal X_1,g(Z_1))\right)^\intercal\otimes X_1\right]^\intercal\\
h_{12}(\alpha,g):=&\frac{1}{q}\sum_{\ell=1}^{q}\left(\mathbb{E}\frac{\partial}{\partial u}m_\ell(V_1,\alpha^\intercal X_1,g(Z_1))X_1\right)\left(\mathbb{E}\frac{\partial}{\partial w}m_\ell(V_1,\alpha^\intercal X_1,g(Z_1))\Phi_k(Z_1)\right)^\intercal,\\
=&\frac{1}{q}\left[\mathbb{E}\left(\frac{\partial}{\partial u}m(V_1,\alpha^\intercal X_1,g(Z_1))\right)^\intercal\otimes X_1\right]\left[\mathbb{E}\left(\frac{\partial}{\partial w}m(V_1,\alpha^\intercal X_1,g(Z_1))\right)^\intercal\otimes \Phi_k(Z_1)\right]^\intercal\\
h_{21}(\alpha,g):=&h_{12}(\alpha,g)^\intercal,\\
h_{22}(\alpha,g):=&\frac{1}{q}\sum_{\ell=1}^{q}\left(\mathbb{E}\frac{\partial}{\partial w}m_\ell(V_1,\alpha^\intercal X_1,g(Z_1))\Phi_k(Z_1)\right)\left(\mathbb{E}\frac{\partial}{\partial w}m_\ell(V_1,\alpha^\intercal X_1,g(Z_1))\Phi_k(Z_1)\right)^\intercal\\
=&\frac{1}{q}\left[\mathbb{E}\left(\frac{\partial}{\partial w}m(V_1,\alpha^\intercal X_1,g(Z_1))\right)^\intercal\otimes\Phi_k(Z_1)\right]\left[\mathbb{E}\left(\frac{\partial}{\partial w}m(V_1,\alpha^\intercal X_1,g(Z_1))\right)^\intercal\otimes\Phi_k(Z_1)\right]^\intercal.
\end{aligned}
$$

Denote

$$
h_n(\alpha,g)=\begin{pmatrix}h_{11}(\alpha,g) & h_{12}(\alpha,g)\\ h_{21}(\alpha,g) & h_{22}(\alpha,g)\end{pmatrix}=\frac{1}{q}\Psi_n\Psi_n^\intercal, \tag{A.1}
$$

where

$$
\Psi_n=\mathbb{E}\begin{pmatrix}\frac{\partial}{\partial u}m(V_1,\alpha^\intercal X_1,g(Z_1))^\intercal\otimes X_1\\ \frac{\partial}{\partial w}m(V_1,\alpha^\intercal X_1,g(Z_1))^\intercal\otimes\Phi_k(Z_1)\end{pmatrix}_{(p+k)\times q}.
$$

Assumption A.1. When sample size is n, suppose that

(i) $\mathbb{E}\left\|m(V_1,\alpha^\intercal X_1,g(Z_1))\right\|^2=O(q)$, $E\|X_1\|^2=O(p)$ and $E\|\Phi_k(Z_1)\|^2=O(k)$;

(ii) $\mathbb{E}\left\|\frac{\partial}{\partial u}m(V_1,\alpha^\intercal X_1,g(Z_1))\right\|^2=O(q)$, and $\mathbb{E}\left\|\frac{\partial}{\partial w}m(V_1,\alpha^\intercal X_1,g(Z_1))\right\|^2=O(q)$;

(iii) $\mathbb{E}\left\|\frac{\partial}{\partial u}m(V_1,\alpha^\intercal X_1,g(Z_1))\otimes X_1\right\|^2=O(pq)$, and
$\mathbb{E}\left\|\frac{\partial}{\partial w}m(V_1,\alpha^\intercal X_1,g(Z_1))\otimes\Phi_k(Z_1)\right\|^2=O(kq)$;

(iv) $\mathbb{E}\left\|\frac{\partial^2}{\partial u^2}m(V_1,\alpha^\intercal X_1,g(Z_1))\otimes X_1X_1^\intercal\right\|^2=O(p^2q)$, and
$\mathbb{E}\left\|\frac{\partial^2}{\partial w^2}m(V_1,\alpha^\intercal X_1,g(Z_1))\otimes\Phi_k(Z_1)\Phi_k(Z_1)^\intercal\right\|^2=O(k^2q)$.

We have the following comments on the assumption. It is nature to alow that each element of m function has the same second moment that suffices the first supposition in A.1(i). Because the dimension p of X_1 diverges with n, in A.1(i) we allow the second moment $E\|X_1\|^2$ diverges too, but, as can be seen in the proof of the following lemma, $E\|X_1\|^2=O(p)$ may be substituted with some appropriate increasing function of p; moreover, $E\|\Phi_k(Z_1)\|^2=O(k)$

can be true for many orthogonal sequences given the relation between the densities of Z_1 and the L^2 space. In A.1(ii) we impose similar condition for the norm of the function's first partial derivatives. A.1(iii) and (iv) 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.

Assumption A.2.

(i) $\|\gamma_k(z)\|^2 p^2 = o(1), \qquad n^{-1}p^2 = o(1);$

(ii) $\|\gamma_k(z)\|^2 k^2 = o(1), \qquad n^{-1}k^2 = o(1).$

Assumption A.2 stipulates the relations among truncation parameter k, the diverging dimension p of the regressor and the sample size. Normally, $\|\gamma_k(z)\|^2 = O(k^{-\tau})$ where τ is related with the smoothness order of the function g. See, for example, Newey [26]. Thus, the assumption implicitly puts some conditions on the smoothness. Notice that the combination of A.2(i) and (ii) implies $\|\gamma_k(z)\|^2 pk = o(1)$ and $n^{-1}pk = o(1)$ which are used in the proof of the following lemma.

Assumption A.3 The partial derivatives of $m(v, u, w)$ satisfy

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

(ii) $q^{-1/2}\left\|\frac{\partial}{\partial w}m(V, \mathbf{a}_1^\intercal X, f_1(Z)) - \frac{\partial}{\partial w}m(V, \mathbf{a}_2^\intercal X, f_2(Z))\right\| \leq A_2(V, X, Z)[\|\mathbf{a}_1 - \mathbf{a}_2\| + |f_1(Z) - f_2(Z)|]$ 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 stipulated for the partial derivatives with additional 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 the consequence of the partial derivatives and is reasonably diverging with the related dimensions.

Lemma A.2. *Under Assumptions 2.1-2.2 and A.1-A.3, (1) $H_n(\alpha, \beta)$ is asymptotically almost surely positive definite; (2) let $h_n(\alpha, g)$ be defined in* (A.1)*, and we then have $\|H_n(\alpha, \beta) - h_n(\alpha, g)\| = o_P(1)$ as $n \to \infty$.*

Proof of Lemma A.2. (1) Split the matrix $H_n(\alpha, \beta) := \tilde{H}_n(\alpha, \beta) + \Delta_n(\alpha, \beta)$ where $\tilde{H}_n(\alpha, \beta)$ is a symmetric 2-by-2 block matrix with blocks

$$\tilde{H}_{11}(\alpha, \beta) = \frac{1}{q}\sum_{\ell=1}^{q}\left(\frac{1}{n}\sum_{j=1}^{n}\frac{\partial}{\partial u}m_\ell(V_j, \alpha^\intercal X_j, \beta^\intercal \Phi_k(Z_j))X_j\right)$$

$$\times \left(\frac{1}{n}\sum_{i=1}^{n}\frac{\partial}{\partial u}m_\ell(V_i,\alpha^\intercal X_i,\beta^\intercal\Phi_k(Z_i))X_i\right)^\intercal,$$

$$\tilde{H}_{12}(\alpha,\beta) = \frac{1}{q}\sum_{\ell=1}^{q}\left(\frac{1}{n}\sum_{j=1}^{n}\frac{\partial}{\partial u}m_\ell(V_j,\alpha^\intercal X_j,\beta^\intercal\Phi_k(Z_j))X_j\right)$$
$$\times \left(\frac{1}{n}\sum_{i=1}^{n}\frac{\partial}{\partial w}m_\ell(V_i,\alpha^\intercal X_i,\beta^\intercal\Phi_k(Z_i))\Phi_k(Z_i)\right)^\intercal,$$

$$\tilde{H}_{22}(\alpha,\beta) = \frac{1}{q}\sum_{\ell=1}^{q}\left(\frac{1}{n}\sum_{j=1}^{n}\frac{\partial}{\partial w}m_\ell(V_j,\mathbf{a}^\intercal X_j,\mathbf{b}^\intercal\Phi_k(Z_j))\Phi_k(Z_j)\right)$$
$$\left(\frac{1}{n}\sum_{i=1}^{n}\frac{\partial}{\partial w}m_\ell(V_i,\mathbf{a}^\intercal X_i,\mathbf{b}^\intercal\Phi_k(Z_i))\Phi_k(Z_i)\right)^\intercal,$$

and $\tilde{H}_{21}(\alpha,\beta) = \tilde{H}_{12}(\alpha,\beta)^\intercal$, and $\Delta_n(\alpha,\beta)$ has blocks

$$\Delta_{11}(\alpha,\beta) = \frac{1}{q}\sum_{\ell=1}^{q}\left(\frac{1}{n}\sum_{i=1}^{n}m_\ell(V_i,\alpha^\intercal X_i,\beta^\intercal\Phi_k(Z_i))\right)$$
$$\times\left(\frac{1}{n}\sum_{j=1}^{n}\frac{\partial^2}{\partial u^2}m_\ell(V_j,\alpha^\intercal X_j,\beta^\intercal\Phi_k(Z_j))X_jX_j^\intercal\right),$$

$$\Delta_{12}(\alpha,\beta) = \frac{1}{q}\sum_{\ell=1}^{q}\left(\frac{1}{n}\sum_{i=1}^{n}m_\ell(V_i,\alpha^\intercal X_i,\beta^\intercal\Phi_k(Z_i))\right)$$
$$\times\left(\frac{1}{n}\sum_{j=1}^{n}\frac{\partial^2}{\partial u\partial w}m_\ell(V_j,\alpha^\intercal X_j,\beta^\intercal\Phi_k(Z_j))X_j\Phi_k(Z_j)^\intercal\right),$$

$$\Delta_{22}(\alpha,\beta) = \frac{1}{q}\sum_{\ell=1}^{q}\left(\frac{1}{n}\sum_{i=1}^{n}m_\ell(V_i,\alpha^\intercal X_i,\beta^\intercal\Phi_k(Z_i))\right)$$
$$\times\left(\frac{1}{n}\sum_{j=1}^{n}\frac{\partial^2}{\partial w^2}m_\ell(V_j,\alpha^\intercal X_j,\beta^\intercal\Phi_k(Z_j))\Phi_k(Z_j)\Phi_k(Z_j)^\intercal\right),$$

and $\Delta_{21}(\alpha,\beta) = \Delta_{12}(\alpha,\beta)^\intercal$. To fulfill the assertion, we shall show

(i) $\tilde{H}_n(\alpha,\beta)$ is almost surely positive definite and

(ii) $\|\Delta_n(\alpha,\beta)\| = o_P(1)$.

Firstly, for any vectors $\mathbf{a}\in\mathbb{R}^p$ and $\mathbf{b}\in\mathbb{R}^k$ where either $\mathbf{a}\neq 0$ or $\mathbf{b}\neq 0$, we have

$$(\mathbf{a}^\intercal,\mathbf{b}^\intercal)\tilde{H}_n(\alpha,\beta)(\mathbf{a}^\intercal,\mathbf{b}^\intercal)^\intercal$$
$$= \frac{1}{q}\sum_{\ell=1}^{q}\left(\frac{1}{n}\sum_{j=1}^{n}\frac{\partial}{\partial u}m_\ell(V_j,\alpha^\intercal X_j,\beta^\intercal\Phi_k(Z_j))\mathbf{a}^\intercal X_j\right)^2$$
$$+ 2\frac{1}{q}\sum_{\ell=1}^{q}\left(\frac{1}{n}\sum_{j=1}^{n}\frac{\partial}{\partial u}m_\ell(V_j,\alpha^\intercal X_j,\beta^\intercal\Phi_k(Z_j))\mathbf{a}^\intercal X_j\right)$$

$$
\begin{aligned}
&\times \left(\frac{1}{n} \sum_{i=1}^{n} \frac{\partial}{\partial w} m_\ell(V_i, \alpha^\intercal X_i, \beta^\intercal \Phi_k(Z_i)) \mathbf{b}^\intercal \Phi_k(Z_i) \right) \\
&+ \frac{1}{q} \sum_{\ell=1}^{q} \left(\frac{1}{n} \sum_{j=1}^{n} \frac{\partial}{\partial w} m_\ell(V_j, \mathbf{a}^\intercal X_j, \mathbf{b}^\intercal \Phi_k(Z_j)) \mathbf{b}^\intercal \Phi_k(Z_j) \right)^2 \\
=&\frac{1}{q} \sum_{\ell=1}^{q} \left[\frac{1}{n} \sum_{j=1}^{n} \left(\frac{\partial}{\partial u} m_\ell(V_j, \alpha^\intercal X_j, \beta^\intercal \Phi_k(Z_j)) \mathbf{a}^\intercal X_j + \frac{\partial}{\partial w} m_\ell(V_j, \mathbf{a}^\intercal X_j, \mathbf{b}^\intercal \Phi_k(Z_j)) \mathbf{b}^\intercal \Phi_k(Z_j) \right) \right]^2,
\end{aligned}
$$

which is almost surely positive. Hence, $\tilde{H}_n(\alpha, \beta)$ is almost surely positive definite.

Secondly, to show $\|\Delta_n(\alpha, \beta)\| = o_P(1)$, it suffices to prove the result for each block. Indeed, appealing to the triangle inequality and Cauchy-Schwarz inequality,

$$
\begin{aligned}
\|\Delta_{11}(\alpha, \beta)\|^2 \leq& \frac{1}{q} \sum_{\ell=1}^{q} \left(\frac{1}{n} \sum_{i=1}^{n} m_\ell(V_i, \alpha^\intercal X_i, \beta^\intercal \Phi_k(Z_i)) \right)^2 \\
&\times \frac{1}{q} \sum_{\ell=1}^{q} \left\| \frac{1}{n} \sum_{j=1}^{n} \frac{\partial^2}{\partial u^2} m_\ell(V_j, \alpha^\intercal X_j, \beta^\intercal \Phi_k(Z_j)) X_j X_j^\intercal \right\|^2 \\
=& \|M_n(\alpha, \beta)\|^2 \frac{1}{q} \sum_{\ell=1}^{q} \left\| \frac{1}{n} \sum_{j=1}^{n} \frac{\partial^2}{\partial u^2} m_\ell(V_j, \alpha^\intercal X_j, \beta^\intercal \Phi_k(Z_j)) X_j X_j^\intercal \right\|^2.
\end{aligned}
$$

Because $\|M_n(\alpha, \beta)\|^2 = O_P(\|\gamma_k(z)\|^2) + O_P(n^{-1})$ by Lemma A.1, we only need to deal with the second factor. Note that

$$
\begin{aligned}
&\frac{1}{q} \sum_{\ell=1}^{q} \left\| \frac{1}{n} \sum_{j=1}^{n} \frac{\partial^2}{\partial u^2} m_\ell(V_j, \alpha^\intercal X_j, \beta^\intercal \Phi_k(Z_j)) X_j X_j^\intercal \right\|^2 \\
\leq& \frac{2}{q} \sum_{\ell=1}^{q} \left\| \mathbb{E} \frac{\partial^2}{\partial u^2} m_\ell(V_1, \alpha^\intercal X_1, \beta^\intercal \Phi_k(Z_1)) X_1 X_1^\intercal \right\|^2 \\
&+ \frac{2}{q} \sum_{\ell=1}^{q} \left\| \frac{1}{n} \sum_{j=1}^{n} \left(\frac{\partial^2}{\partial u^2} m_\ell(V_j, \alpha^\intercal X_j, \beta^\intercal \Phi_k(Z_j)) X_j X_j^\intercal \right. \right. \\
&\qquad \left. \left. - \mathbb{E} \frac{\partial^2}{\partial u^2} m_\ell(V_j, \alpha^\intercal X_j, \beta^\intercal \Phi_k(Z_j)) X_j X_j^\intercal \right) \right\|^2
\end{aligned}
$$

where by Assumption A.1 the first term is $O(p^2)$, while by the iid property for the second we have

$$
\begin{aligned}
&\frac{1}{q} \sum_{\ell=1}^{q} \mathbb{E} \left\| \frac{1}{n} \sum_{j=1}^{n} \left(\frac{\partial^2}{\partial u^2} m_\ell(V_j, \alpha^\intercal X_j, \beta^\intercal \Phi_k(Z_j)) X_j X_j^\intercal \right. \right. \\
&\qquad \left. \left. - \mathbb{E} \frac{\partial^2}{\partial u^2} m_\ell(V_j, \alpha^\intercal X_j, \beta^\intercal \Phi_k(Z_j)) X_j X_j^\intercal \right) \right\|^2 \\
=& \frac{1}{n^2} \frac{1}{q} \sum_{\ell=1}^{q} \sum_{j=1}^{n} \mathbb{E} \left\| \frac{\partial^2}{\partial u^2} m_\ell(V_j, \alpha^\intercal X_j, \beta^\intercal \Phi_k(Z_j)) X_j X_j^\intercal - \mathbb{E} \frac{\partial^2}{\partial u^2} m_\ell(V_j, \alpha^\intercal X_j, \beta^\intercal \Phi_k(Z_j)) X_j X_j^\intercal \right\|^2
\end{aligned}
$$

$$
\begin{aligned}
=&\frac{1}{n}\frac{1}{q}\sum_{\ell=1}^{q}\mathbb{E}\left\|\frac{\partial^2}{\partial u^2}m_\ell(V_1,\alpha^\intercal X_1,\beta^\intercal\Phi_k(Z_1))X_1X_1^\intercal-\mathbb{E}\frac{\partial^2}{\partial u^2}m_\ell(V_1,\alpha^\intercal X_1,\beta^\intercal\Phi_k(Z_1))X_1X_1^\intercal\right\|^2\\
\leq&\frac{1}{n}\frac{1}{q}\sum_{\ell=1}^{q}\mathbb{E}\left\|\frac{\partial^2}{\partial u^2}m_\ell(V_1,\alpha^\intercal X_1,\beta^\intercal\Phi_k(Z_1))X_1X_1^\intercal\right\|^2\\
=&\frac{1}{n}\frac{1}{q}\mathbb{E}\left\|\frac{\partial^2}{\partial u^2}m(V_1,\alpha^\intercal X_1,\beta^\intercal\Phi_k(Z_1))\otimes X_1X_1^\intercal\right\|^2\\
=&O(n^{-1}p^2),
\end{aligned}
$$

by Assumption A.1, from which $\|\Delta_{11}(\alpha,\beta)\|^2=O_P(\|\gamma_k(z)\|^2p^2)+O_P(n^{-1}p^2)=o_P(1)$.

Similarly,

$$
\|\Delta_{12}(\alpha,\beta)\|^2\leq\|M_n(\alpha,\beta)\|^2\frac{1}{q}\sum_{\ell=1}^{q}\left\|\frac{1}{n}\sum_{j=1}^{n}\frac{\partial^2}{\partial u\partial w}m_\ell(V_j,\alpha^\intercal X_j,\beta^\intercal\Phi_k(Z_j))X_j\Phi_k(Z_j)\right\|^2
$$

and for the second factor using again the iid property, we have

$$
\begin{aligned}
&\frac{1}{q}\sum_{\ell=1}^{q}\mathbb{E}\left\|\frac{1}{n}\sum_{j=1}^{n}\frac{\partial^2}{\partial u\partial w}m_\ell(V_j,\alpha^\intercal X_j,\beta^\intercal\Phi_k(Z_j))X_j\Phi_k(Z_j)^\intercal\right\|^2\\
\leq&2\frac{1}{q}\sum_{\ell=1}^{q}\left\|\mathbb{E}\frac{\partial^2}{\partial u\partial w}m_\ell(V_1,\alpha^\intercal X_1,\beta^\intercal\Phi_k(Z_1))X_1\Phi_k(Z_1)^\intercal\right\|^2\\
&+2\frac{1}{n}\frac{1}{q}\sum_{\ell=1}^{q}\mathbb{E}\left\|\frac{\partial^2}{\partial u\partial w}m_\ell(V_1,\alpha^\intercal X_1,\beta^\intercal\Phi_k(Z_1))X_1\Phi_k(Z_1)^\intercal\right.\\
&\qquad\left.-\mathbb{E}\frac{\partial^2}{\partial u\partial w}m_\ell(V_1,\alpha^\intercal X_1,\beta^\intercal\Phi_k(Z_1))X_1\Phi_k(Z_1)^\intercal\right\|^2\\
\leq&2\frac{1}{q}\sum_{\ell=1}^{q}\left\|\mathbb{E}\frac{\partial^2}{\partial u\partial w}m_\ell(V_1,\alpha^\intercal X_1,\beta^\intercal\Phi_k(Z_1))X_1\Phi_k(Z_1)^\intercal\right\|^2\\
&+2\frac{1}{n}\frac{1}{q}\sum_{\ell=1}^{q}\mathbb{E}\left\|\frac{\partial^2}{\partial u\partial w}m_\ell(V_1,\alpha^\intercal X_1,\beta^\intercal\Phi_k(Z_1))X_1\Phi_k(Z_1)^\intercal\right\|^2\\
=&2\frac{1}{q}\left\|\mathbb{E}\frac{\partial^2}{\partial u\partial w}m(V_1,\alpha^\intercal X_1,\beta^\intercal\Phi_k(Z_1))\otimes X_1\Phi_k(Z_1)^\intercal\right\|^2\\
&+2\frac{1}{n}\frac{1}{q}\mathbb{E}\left\|\frac{\partial^2}{\partial u\partial w}m(V_1,\alpha^\intercal X_1,\beta^\intercal\Phi_k(Z_1))\otimes X_1\Phi_k(Z_1)^\intercal\right\|^2\\
=&O(pk)+O(n^{-1}pk),
\end{aligned}
$$

which implies $\|\Delta_{12}(\alpha,\beta)\|^2=O_P(\|\gamma_k(z)\|^2pk)+O_P(n^{-1}pk)=o_P(1)$.

Furthermore,

$$
\|\Delta_{22}(\alpha,\beta)\|^2\leq\|M_n(\alpha,\beta)\|^2\frac{1}{q}\sum_{\ell=1}^{q}\left\|\frac{1}{n}\sum_{j=1}^{n}\frac{\partial^2}{\partial w^2}m_\ell(V_j,\alpha^\intercal X_j,\beta^\intercal\Phi_k(Z_j))\Phi_k(Z_j)\Phi_k(Z_j)^\intercal\right\|^2
$$

where the second factor can be derived similarly

$$
\begin{aligned}
&\frac{1}{q}\sum_{\ell=1}^{q}\mathbb{E}\left\|\frac{1}{n}\sum_{j=1}^{n}\frac{\partial^2}{\partial w^2}m_\ell(V_j,\alpha^\intercal X_j,\beta^\intercal\Phi_k(Z_j))\Phi_k(Z_j)\Phi_k(Z_j)^\intercal\right\|^2\\
\leq& 2\frac{1}{q}\left\|\mathbb{E}\frac{\partial^2}{\partial w^2}m(V_1,\alpha^\intercal X_1,\beta^\intercal\Phi_k(Z_1))\otimes\Phi_k(Z_1)\Phi_k(Z_1)^\intercal\right\|^2\\
&+2\frac{1}{n}\frac{1}{q}\mathbb{E}\left\|\frac{\partial^2}{\partial w^2}m(V_1,\alpha^\intercal X_1,\beta^\intercal\Phi_k(Z_1))\otimes\Phi_k(Z_1)\Phi_k(Z_1)^\intercal\right\|^2\\
=&O(k^2)+O(n^{-1}k^2),
\end{aligned}
$$

giving that $\|\Delta_{22}(\alpha,\beta)\|^2=O_P(\|\gamma_k(z)\|^2k^2)+O_P(n^{-1}k^2)=o_P(1)$. This finishes the assertion (i).

Now, we show (ii). Because $\|H_n(\alpha,\beta)-h_n(\alpha,g)\|\leq\|\Delta_n(\alpha,\beta)\|+\|\tilde{H}_n(\alpha,\beta)-h_n(\alpha,g)\|=o_P(1)+\|\tilde{H}_n(\alpha,\beta)-h_n(\alpha,g)\|$, what we need to show is $\|\tilde{H}_n(\alpha,\beta)-h_n(\alpha,g)\|=o_P(1)$. It is sufficient to show the result in block-sense. Indeed,

$$
\begin{aligned}
&\tilde{H}_{11}(\alpha,\beta)-h_{11}(\alpha,g)\\
=&\frac{1}{q}\sum_{\ell=1}^{q}\left(\frac{1}{n}\sum_{j=1}^{n}\frac{\partial}{\partial u}m_\ell(V_j,\alpha^\intercal X_j,\beta^\intercal\Phi_k(Z_j))X_j\right)\left(\frac{1}{n}\sum_{i=1}^{n}\frac{\partial}{\partial u}m_\ell(V_i,\alpha^\intercal X_i,\beta^\intercal\Phi_k(Z_i))X_i\right)^\intercal\\
&-\frac{1}{q}\sum_{\ell=1}^{q}\left(\mathbb{E}\frac{\partial}{\partial u}m_\ell(V_1,\alpha^\intercal X_1,g(Z_1))X_1\right)\left(\mathbb{E}\frac{\partial}{\partial u}m_\ell(V_1,\alpha^\intercal X_1,g(Z_1))X_1\right)^\intercal\\
=&\frac{1}{q}\sum_{\ell=1}^{q}\frac{1}{n}\sum_{j=1}^{n}\left(\frac{\partial}{\partial u}m_\ell(V_j,\alpha^\intercal X_j,\beta^\intercal\Phi_k(Z_j))X_j-\mathbb{E}\frac{\partial}{\partial u}m_\ell(V_j,\alpha^\intercal X_j,g(Z_j))X_j\right)\\
&\qquad\times\left(\frac{1}{n}\sum_{i=1}^{n}\frac{\partial}{\partial u}m_\ell(V_i,\alpha^\intercal X_i,\beta^\intercal\Phi_k(Z_i))X_i\right)^\intercal\\
&+\frac{1}{q}\sum_{\ell=1}^{q}\left(\mathbb{E}\frac{\partial}{\partial u}m_\ell(V_1,\alpha^\intercal X_1,g(Z_1))X_1\right)\\
&\qquad\times\frac{1}{n}\sum_{i=1}^{n}\left(\frac{\partial}{\partial u}m_\ell(V_i,\alpha^\intercal X_i,\beta^\intercal\Phi_k(Z_i))X_i-\mathbb{E}\frac{\partial}{\partial u}m_\ell(V_i,\alpha^\intercal X_i,g(Z_i))X_i\right)^\intercal\\
:=&I_1+I_2,\quad\text{say.}
\end{aligned}
$$

Notice further that

$$
\begin{aligned}
I_1=&\frac{1}{q}\sum_{\ell=1}^{q}\frac{1}{n}\sum_{j=1}^{n}\left(\frac{\partial}{\partial u}m_\ell(V_j,\alpha^\intercal X_j,\beta^\intercal\Phi_k(Z_j))X_j-\frac{\partial}{\partial u}m_\ell(V_j,\alpha^\intercal X_j,g(Z_j))X_j\right)\\
&\qquad\times\left(\frac{1}{n}\sum_{i=1}^{n}\frac{\partial}{\partial u}m_\ell(V_i,\alpha^\intercal X_i,\beta^\intercal\Phi_k(Z_i))X_i\right)^\intercal\\
&+\frac{1}{q}\sum_{\ell=1}^{q}\frac{1}{n}\sum_{j=1}^{n}\left(\frac{\partial}{\partial u}m_\ell(V_j,\alpha^\intercal X_j,g(Z_j))X_j-\mathbb{E}\frac{\partial}{\partial u}m_\ell(V_j,\alpha^\intercal X_j,g(Z_j))X_j\right)
\end{aligned}
$$

$$\times \left(\frac{1}{n} \sum_{i=1}^{n} \frac{\partial}{\partial u} m_\ell(V_i, \alpha^\intercal X_i, \beta^\intercal \Phi_k(Z_i)) X_i \right)^\intercal .$$

Hence, using Cauchy-Schwarz inequality,

$$\begin{aligned}
\|I_1\|^2 \leq & \frac{1}{q} \sum_{\ell=1}^{q} \left\| \frac{1}{n} \sum_{j=1}^{n} \left(\frac{\partial}{\partial u} m_\ell(V_j, \alpha^\intercal X_j, \beta^\intercal \Phi_k(Z_j)) - \frac{\partial}{\partial u} m_\ell(V_j, \alpha^\intercal X_j, g(Z_j)) \right) X_j \right\|^2 \\
& \times \frac{1}{q} \sum_{\ell=1}^{q} \left\| \frac{1}{n} \sum_{i=1}^{n} \frac{\partial}{\partial u} m_\ell(V_i, \alpha^\intercal X_i, \beta^\intercal \Phi_k(Z_i)) X_i \right\|^2 \\
& + \frac{1}{q} \sum_{\ell=1}^{q} \left\| \frac{1}{n} \sum_{j=1}^{n} \left(\frac{\partial}{\partial u} m_\ell(V_j, \alpha^\intercal X_j, g(Z_j)) X_j - \mathbb{E} \frac{\partial}{\partial u} m_\ell(V_j, \alpha^\intercal X_j, g(Z_j)) X_j \right) \right\|^2 \\
& \times \frac{1}{q} \sum_{\ell=1}^{q} \left\| \frac{1}{n} \sum_{i=1}^{n} \frac{\partial}{\partial u} m_\ell(V_i, \alpha^\intercal X_i, \beta^\intercal \Phi_k(Z_i)) X_i \right\|^2 \\
:= & I_{11} \times I_{13} + I_{12} \times I_{13}, \quad \text{say.}
\end{aligned}$$

Due to the i.i.d. property and the Law of Large Number (LLN, hereafter), I_{11} has the same order in probability as

$$\begin{aligned}
& \frac{1}{q} \sum_{\ell=1}^{q} \left\| \mathbb{E} \left(\frac{\partial}{\partial u} m_\ell(V_1, \alpha^\intercal X_1, \beta^\intercal \Phi_k(Z_1)) - \frac{\partial}{\partial u} m_\ell(V_1, \alpha^\intercal X_1, g(Z_1)) \right) X_1 \right\|^2 \\
= & \frac{1}{q} \left\| \mathbb{E} \left(\frac{\partial}{\partial u} m(V_1, \alpha^\intercal X_1, \beta^\intercal \Phi_k(Z_1)) - \frac{\partial}{\partial u} m(V_1, \alpha^\intercal X_1, g(Z_1)) \right) \otimes X_1 \right\|^2 \\
\leq & \mathbb{E}[A_1(V_1, X_1, Z_1)^2 \|X_1\|^2] \mathbb{E}[\gamma_k(Z_1)^2] = O(\|\gamma_k(z)\|^2 p),
\end{aligned}$$

while for I_{12}, by the iid property,

$$\begin{aligned}
\mathbb{E}[I_{12}] = & \frac{1}{n^2} \frac{1}{q} \sum_{\ell=1}^{q} \sum_{j=1}^{n} \mathbb{E} \left\| \frac{\partial}{\partial u} m_\ell(V_j, \alpha^\intercal X_j, g(Z_j)) X_j - \mathbb{E} \frac{\partial}{\partial u} m_\ell(V_j, \alpha^\intercal X_j, g(Z_j)) X_j \right\|^2 \\
= & \frac{1}{n} \frac{1}{q} \sum_{\ell=1}^{q} \mathbb{E} \left\| \frac{\partial}{\partial u} m_\ell(V_1, \alpha^\intercal X_1, g(Z_1)) X_1 - \mathbb{E} \frac{\partial}{\partial u} m_\ell(V_1, \alpha^\intercal X_1, g(Z_1)) X_1 \right\|^2 \\
\leq & \frac{1}{n} \frac{1}{q} \sum_{\ell=1}^{q} \mathbb{E} \left\| \frac{\partial}{\partial u} m_\ell(V_1, \alpha^\intercal X_1, g(Z_1)) X_1 \right\|^2 \\
\leq & \frac{1}{n} \frac{1}{q} \mathbb{E} \left\| \frac{\partial}{\partial u} m(V_1, \alpha^\intercal X_1, g(Z_1)) \otimes X_1 \right\|^2 = O(n^{-1} p)
\end{aligned}$$

by Assumption A.1. Moreover, by virtue of the iid property and the LLN, I_{13} has the same order in probability as

$$\frac{1}{q} \sum_{\ell=1}^{q} \left\| \mathbb{E} \frac{\partial}{\partial u} m_\ell(V_i, \alpha^\intercal X_i, g(Z_i)) X_i \right\|^2$$

$$
\begin{aligned}
&+\frac{1}{q}\sum_{\ell=1}^{q}\left\|\frac{1}{n}\sum_{i=1}^{n}\left[\frac{\partial}{\partial u}m_\ell(V_i,\alpha^\intercal X_i,\beta^\intercal\Phi_k(Z_i))-\frac{\partial}{\partial u}m_\ell(V_i,\alpha^\intercal X_i,g(Z_i))\right]X_i\right\|^2\\
=&\frac{1}{q}\left\|\mathbb{E}\frac{\partial}{\partial u}m(V_1,\alpha^\intercal X_1,g(Z_1))\otimes X_1\right\|^2\\
&+\frac{1}{q}\sum_{\ell=1}^{q}\left\|\mathbb{E}\left[\frac{\partial}{\partial u}m_\ell(V_1,\alpha^\intercal X_1,\beta^\intercal\Phi_k(Z_1))-\frac{\partial}{\partial u}m_\ell(V_1,\alpha^\intercal X_1,g(Z_1))\right]X_1\right\|^2\\
=&O(p)+\frac{1}{q}\left\|\mathbb{E}\left[\frac{\partial}{\partial u}m(V_1,\alpha^\intercal X_1,\beta^\intercal\Phi_k(Z_1))-\frac{\partial}{\partial u}m(V_1,\alpha^\intercal X_1,g(Z_1))\right]\otimes X_1\right\|^2\\
\le&O(p)+(\mathbb{E}[A_1(V_1,X_1,Z_1)|\gamma_k(Z_1)|\|X_1\|])^2\le O(p)+O(\|\gamma_k(z)\|^2p)
\end{aligned}
$$

due to Assumptions A.1 and A.3, implying that $\|I_1\|^2=O_P(n^{-1}p^2)+O_P(\|\gamma_k(z)\|^2p^2)=o_P(1)$ by Assumption A.2.

Now, we consider I_2. Note that

$$
\begin{aligned}
\|I_2\|^2\le&\frac{1}{q}\sum_{\ell=1}^{q}\left\|\mathbb{E}\frac{\partial}{\partial u}m_\ell(V_1,\alpha^\intercal X_1,g(Z_1))X_1\right\|^2\\
&\times\frac{1}{q}\sum_{\ell=1}^{q}\left\|\frac{1}{n}\sum_{i=1}^{n}\left(\frac{\partial}{\partial u}m_\ell(V_i,\alpha^\intercal X_i,\beta^\intercal\Phi_k(Z_i))X_i-\mathbb{E}\frac{\partial}{\partial u}m_\ell(V_i,\alpha^\intercal X_i,g(Z_i))X_i\right)\right\|^2\\
\le&2\frac{1}{q}\left\|\mathbb{E}\frac{\partial}{\partial u}m(V_1,\alpha^\intercal X_1,g(Z_1))\otimes X_1\right\|^2\\
&\times\frac{1}{q}\left\|\frac{1}{n}\sum_{i=1}^{n}\left(\frac{\partial}{\partial u}m(V_i,\alpha^\intercal X_i,\beta^\intercal\Phi_k(Z_i))-\frac{\partial}{\partial u}m(V_i,\alpha^\intercal X_i,g(Z_i))\right)\otimes X_j\right\|^2\\
&+2\frac{1}{q}\left\|\mathbb{E}\frac{\partial}{\partial u}m(V_1,\alpha^\intercal X_1,g(Z_1))\otimes X_1\right\|^2\\
&\times\frac{1}{q}\sum_{\ell=1}^{q}\left\|\frac{1}{n}\sum_{i=1}^{n}\left(\frac{\partial}{\partial u}m_\ell(V_i,\alpha^\intercal X_i,g(Z_i))X_i-\mathbb{E}\frac{\partial}{\partial u}m_\ell(V_i,\alpha^\intercal X_i,g(Z_i))X_i\right)\right\|^2\\
:=&2I_{21}(I_{22}+I_{23}),\quad\text{say.}
\end{aligned}
$$

By Assumption A.1, $I_{21}=O(p)$. In addition, by the LLN I_{22} has the same order in probability as

$$
\begin{aligned}
&\frac{1}{q}\left\|\mathbb{E}\left(\frac{\partial}{\partial u}m(V_1,\alpha^\intercal X_1,\beta^\intercal\Phi_k(Z_1))-\frac{\partial}{\partial u}m(V_1,\alpha^\intercal X_1,g(Z_1))\right)\otimes X_1\right\|^2\\
\le&(\mathbb{E}[A_1(V_1,X_1,Z_1)|\gamma_k(Z_1)|\|X_1\|])^2\le O(p)\|\gamma_k(z)\|^2
\end{aligned}
$$

using Assumption A.3; meanwhile, by the iid property,

$$
\mathbb{E}[I_{23}]=\frac{1}{n^2}\frac{1}{q}\sum_{\ell=1}^{q}\sum_{i=1}^{n}\mathbb{E}\left\|\frac{\partial}{\partial u}m_\ell(V_i,\alpha^\intercal X_i,g(Z_i))X_i-\mathbb{E}\frac{\partial}{\partial u}m_\ell(V_i,\alpha^\intercal X_i,g(Z_i))X_i\right\|^2
$$

$$
\begin{aligned}
=&\frac{1}{n}\frac{1}{q}\sum_{\ell=1}^{q}\mathbb{E}\left\|\frac{\partial}{\partial u}m_\ell(V_1,\alpha^\intercal X_1,g(Z_1))X_1-\mathbb{E}\frac{\partial}{\partial u}m_\ell(V_1,\alpha^\intercal X_1,g(Z_1))X_1\right\|^2\\
\le&\frac{1}{n}\frac{1}{q}\sum_{\ell=1}^{q}\mathbb{E}\left\|\frac{\partial}{\partial u}m_\ell(V_1,\alpha^\intercal X_1,g(Z_1))X_1\right\|^2\\
=&\frac{1}{n}\frac{1}{q}\mathbb{E}\left\|\frac{\partial}{\partial u}m(V_1,\alpha^\intercal X_1,g(Z_1))\otimes X_1\right\|^2=O(n^{-1}p)
\end{aligned}
$$

by Assumption A.1. Hence, $\|I_2\|^2=O_P(n^{-1}p^2)+O_P(\|\gamma_k(z)\|^2p^2)=o_P(1)$. Thus, $\|\tilde{H}_{11}(\alpha,\beta)-h_{11}(\alpha,\beta)\|^2=O_P(1)$.

Moreover,

$$
\begin{aligned}
&\tilde{H}_{12}(\alpha,\beta)-h_{12}(\alpha,g)\\
=&\frac{1}{q}\sum_{\ell=1}^{q}\left(\frac{1}{n}\sum_{j=1}^{n}\frac{\partial}{\partial u}m_\ell(V_j,\alpha^\intercal X_j,\beta^\intercal\Phi_k(Z_j))X_j\right)\left(\frac{1}{n}\sum_{i=1}^{n}\frac{\partial}{\partial w}m_\ell(V_i,\alpha^\intercal X_i,\beta^\intercal\Phi_k(Z_i))\Phi_k(Z_i)\right)^\intercal\\
&-\frac{1}{q}\sum_{\ell=1}^{q}\left(\mathbb{E}\frac{\partial}{\partial u}m_\ell(V_1,\alpha^\intercal X_1,g(Z_1))X_1\right)\left(\mathbb{E}\frac{\partial}{\partial u}m_\ell(V_1,\alpha^\intercal X_1,g(Z_1))\Phi_k(Z_1)\right)^\intercal\\
=&\frac{1}{q}\sum_{\ell=1}^{q}\left(\frac{1}{n}\sum_{j=1}^{n}\frac{\partial}{\partial u}m_\ell(V_j,\alpha^\intercal X_j,\beta^\intercal\Phi_k(Z_j))X_j-\mathbb{E}\frac{\partial}{\partial u}m_\ell(V_1,\alpha^\intercal X_1,g(Z_1))X_1\right)\\
&\quad\times\left(\frac{1}{n}\sum_{i=1}^{n}\frac{\partial}{\partial w}m_\ell(V_i,\alpha^\intercal X_i,\beta^\intercal\Phi_k(Z_i))\Phi_k(Z_i)\right)^\intercal\\
&+\frac{1}{q}\sum_{\ell=1}^{q}\left(\mathbb{E}\frac{\partial}{\partial u}m_\ell(V_1,\alpha^\intercal X_1,g(Z_1))X_1\right)\\
&\quad\times\left(\frac{1}{n}\sum_{i=1}^{n}\frac{\partial}{\partial w}m_\ell(V_i,\alpha^\intercal X_i,\beta^\intercal\Phi_k(Z_i))\Phi_k(Z_i)-\mathbb{E}\frac{\partial}{\partial u}m_\ell(V_1,\alpha^\intercal X_1,g(Z_1))\Phi_k(Z_1)\right)^\intercal\\
:=&I_3+I_4,\quad\text{say.}
\end{aligned}
$$

Similar to I_1, $\|I_3\|^2=O_P(n^{-1}pk)+O_P(\|\gamma_k(z)\|^2pk)=o_P(1)$ by Assumption A.2; and similar to I_2, we may have $\|I_4\|^2=O_P(n^{-1}pk)+O_P(\|\gamma_k(z)\|^2pk)=o_P(1)$. We then have $\|\tilde{H}_{12}(\alpha,\beta)-h_{12}(\alpha,\beta)\|^2=o_P(1)$.

Finally, we derive similarly for $\tilde{H}_{22}(\alpha,\beta)-h_{22}(\alpha,\beta)$,

$$
\begin{aligned}
&\tilde{H}_{22}(\alpha,\beta)-h_{22}(\alpha,g)\\
=&\frac{1}{q}\sum_{\ell=1}^{q}\left(\frac{1}{n}\sum_{j=1}^{n}\frac{\partial}{\partial w}m_\ell(V_j,\alpha^\intercal X_j,\beta^\intercal\Phi_k(Z_j))\Phi_k(Z_j)\right)\left(\frac{1}{n}\sum_{i=1}^{n}\frac{\partial}{\partial w}m_\ell(V_i,\alpha^\intercal X_i,\beta^\intercal\Phi_k(Z_i))\Phi_k(Z_i)\right)^\intercal\\
&-\frac{1}{q}\sum_{\ell=1}^{q}\left(\mathbb{E}\frac{\partial}{\partial w}m_\ell(V_1,\alpha^\intercal X_1,g(Z_1))\Phi_k(Z_1)\right)\left(\mathbb{E}\frac{\partial}{\partial w}m_\ell(V_1,\alpha^\intercal X_1,g(Z_1))\Phi_k(Z_1)\right)^\intercal\\
=&\frac{1}{q}\sum_{\ell=1}^{q}\left(\frac{1}{n}\sum_{j=1}^{n}\frac{\partial}{\partial w}m_\ell(V_j,\alpha^\intercal X_j,\beta^\intercal\Phi_k(Z_j))\Phi_k(Z_j)-\mathbb{E}\frac{\partial}{\partial w}m_\ell(V_1,\alpha^\intercal X_1,g(Z_1))\Phi_k(Z_1)\right)
\end{aligned}
$$

$$\times \left(\frac{1}{n} \sum_{i=1}^{n} \frac{\partial}{\partial w} m_\ell(V_i, \alpha^\intercal X_i, \beta^\intercal \Phi_k(Z_i)) \Phi_k(Z_i) \right)^\intercal$$

$$+ \frac{1}{q} \sum_{\ell=1}^{q} \left(\mathbb{E} \frac{\partial}{\partial w} m_\ell(V_1, \alpha^\intercal X_1, g(Z_1)) \Phi_k(Z_1) \right)$$

$$\times \left(\frac{1}{n} \sum_{i=1}^{n} \frac{\partial}{\partial w} m_\ell(V_i, \alpha^\intercal X_i, \beta^\intercal \Phi_k(Z_i)) \Phi_k(Z_i) - \mathbb{E} \frac{\partial}{\partial w} m_\ell(V_1, \alpha^\intercal X_1, g) \Phi_k(Z_1) \right)^\intercal$$

$:= I_5 + I_6$, say.

Using the same approach we have $\|I_5\|^2 = O_P(n^{-1}k^2) + O_P(\|\gamma_k(z)\|^2 k^2) = o_P(1)$ and $\|I_6\|^2 = O_P(n^{-1}k^2) + O_P(\|\gamma_k(z)\|^2 k^2) = o_P(1)$ by Assumption A.2. The whole proof is complete. □

Denote $S_n(\mathbf{a}, \mathbf{b}) = (S_{1n}(\mathbf{a}, \mathbf{b})^\intercal, S_{2n}(\mathbf{a}, \mathbf{b})^\intercal)^\intercal$, where

$$S_{1n}(\mathbf{a}, \mathbf{b}) = \frac{\partial}{\partial \mathbf{a}} \|M_n(\mathbf{a}, \mathbf{b})\|^2 = \frac{1}{qn^2} \sum_{\ell=1}^{q} \sum_{i=1}^{n} m_\ell(V_i, \mathbf{a}^\intercal X_i, \mathbf{b}^\intercal \Phi_k(Z_i)) \times \sum_{j=1}^{n} \frac{\partial}{\partial u} m_\ell(V_j, \mathbf{a}^\intercal X_j, \mathbf{b}^\intercal \Phi_k(Z_j)) X_j$$

$$S_{2n}(\mathbf{a}, \mathbf{b}) = \frac{\partial}{\partial \mathbf{b}} \|M_n(\mathbf{a}, \mathbf{b})\|^2 = \frac{1}{qn^2} \sum_{\ell=1}^{q} \sum_{i=1}^{n} m_\ell(V_i, \mathbf{a}^\intercal X_i, \mathbf{b}^\intercal \Phi_k(Z_i)) \times \sum_{j=1}^{n} \frac{\partial}{\partial w} m_\ell(V_j, \mathbf{a}^\intercal X_j, \mathbf{b}^\intercal \Phi_k(Z_j)) \Phi_k(Z_j).$$

We now focus on $S_n(\alpha, \beta)$ with sub-vectors $S_{1n}(\alpha, \beta)$ and $S_{2n}(\alpha, \beta)$. Define

$$\begin{aligned} s_{1n}(\alpha, g) &= \frac{1}{qn} \sum_{\ell=1}^{q} \sum_{i=1}^{n} m_\ell(V_i, \alpha^\intercal X_i, g(Z_i)) \mathbb{E} \frac{\partial}{\partial u} m_\ell(V_1, \alpha^\intercal X_1, g(Z_1)) X_1, \\ &= \left[\frac{1}{q} \mathbb{E} \left(\frac{\partial}{\partial u} m(V_1, \alpha^\intercal X_1, g(Z_1))^\intercal \otimes X_1 \right) \right] \frac{1}{n} \sum_{i=1}^{n} m(V_i, \alpha^\intercal X_i, g(Z_i)), \\ s_{2n}(\alpha, g) &= \frac{1}{qn} \sum_{\ell=1}^{q} \sum_{i=1}^{n} m_\ell(V_i, \alpha^\intercal X_i, g(Z_i)) \mathbb{E} \frac{\partial}{\partial w} m_\ell(V_1, \alpha^\intercal X_1, g(Z_1)) \Phi_k(Z_1) \\ &= \left[\frac{1}{q} \mathbb{E} \left(\frac{\partial}{\partial w} m(V_1, \alpha^\intercal X_1, g(Z_1))^\intercal \otimes \Phi_k(Z_1) \right) \right] \frac{1}{n} \sum_{i=1}^{n} m(V_i, \alpha^\intercal X_i, g(Z_i)), \end{aligned}$$

and hence

$$s_n(\alpha, g) = (s_{1n}(\alpha, g)^\intercal, s_{2n}(\alpha, g)^\intercal)^\intercal = \frac{1}{q} \Psi_n \frac{1}{n} \sum_{i=1}^{n} m(V_i, \alpha^\intercal X_i, g(Z_i)), \tag{A.2}$$

where Ψ_n is given by (A.1).

Lemma A.3. *Under Assumptions 2.1-2.2, 3.1, 3.3, A.1-A.3, as $n \to \infty$ we have*

$$\|S_n(\alpha, \beta) - s_n(\alpha, g)\| = o_P(1).$$

Proof. It is sufficient to show that $\|S_{1n}(\alpha, \beta) - s_{1n}(\alpha, g)\| = o_P(1)$ and $\|S_{2n}(\alpha, \beta) - s_{2n}(\alpha, g)\| = o_P(1)$. Observe that

$$\begin{aligned}
&S_{1n}(\alpha, \beta) - s_{1n}(\alpha, g) \\
=&\frac{1}{q}\sum_{\ell=1}^{q}\frac{1}{n}\sum_{i=1}^{n}[m_\ell(V_i, \alpha^\intercal X_i, \beta^\intercal \Phi_k(Z_i)) - m_\ell(V_i, \alpha^\intercal X_i, g(Z_i))] \\
&\quad \times \frac{1}{n}\sum_{j=1}^{n}\frac{\partial}{\partial u}m_\ell(V_j, \alpha^\intercal X_j, \beta^\intercal \Phi_k(Z_j))X_j \\
&+ \frac{1}{q}\sum_{\ell=1}^{q}\frac{1}{n}\sum_{i=1}^{n} m_\ell(V_i, \alpha^\intercal X_i, g(Z_i)) \\
&\quad \times \frac{1}{n}\sum_{j=1}^{n}\left(\frac{\partial}{\partial u}m_\ell(V_j, \alpha^\intercal X_j, \beta^\intercal \Phi_k(Z_j)) - \frac{\partial}{\partial u}m_\ell(V_j, \alpha^\intercal X_j, g(Z_j))\right) X_j \\
&+ \frac{1}{q}\sum_{\ell=1}^{q}\frac{1}{n}\sum_{i=1}^{n} m_\ell(V_i, \alpha^\intercal X_i, g(Z_i)) \\
&\quad \times \frac{1}{n}\sum_{j=1}^{n}\left(\frac{\partial}{\partial u}m_\ell(V_j, \alpha^\intercal X_j, g(Z_j))X_j - \mathbb{E}\frac{\partial}{\partial u}m_\ell(V_j, \alpha^\intercal X_j, g(Z_j))X_j\right) \\
:=&I_1 + I_2 + I_3, \quad \text{say.}
\end{aligned}$$

Then, using Cauchy-Schwarz inequality gives

$$\begin{aligned}
\|I_1\|^2 \leq&\frac{1}{q}\sum_{\ell=1}^{q}\left(\frac{1}{n}\sum_{i=1}^{n}[m_\ell(V_i, \alpha^\intercal X_i, \beta^\intercal \Phi_k(Z_i)) - m_\ell(V_i, \alpha^\intercal X_i, g(Z_i))]\right)^2 \\
&\quad \times \frac{1}{q}\sum_{\ell=1}^{q}\left\|\frac{1}{n}\sum_{j=1}^{n}\frac{\partial}{\partial u}m_\ell(V_j, \alpha^\intercal X_j, \beta^\intercal \Phi_k(Z_j))X_j\right\|^2 \\
:=&I_{11} \times I_{12}, \quad \text{say.}
\end{aligned}$$

Observe further that

$$\begin{aligned}
\mathbb{E}[I_{11}] =&\frac{1}{q}\sum_{\ell=1}^{q}\mathbb{E}\left(\frac{1}{n}\sum_{i=1}^{n}[m_\ell(V_i, \alpha^\intercal X_i, \beta^\intercal \Phi_k(Z_i)) - m_\ell(V_i, \alpha^\intercal X_i, g(Z_i))]\right)^2 \\
=&\frac{1}{q}\sum_{\ell=1}^{q}\mathrm{Var}\left(\frac{1}{n}\sum_{i=1}^{n}[m_\ell(V_i, \alpha^\intercal X_i, \beta^\intercal \Phi_k(Z_i)) - m_\ell(V_i, \alpha^\intercal X_i, g(Z_i))]\right) \\
&+ \frac{1}{q}\sum_{\ell=1}^{q}\left(\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}[m_\ell(V_i, \alpha^\intercal X_i, \beta^\intercal \Phi_k(Z_i)) - m_\ell(V_i, \alpha^\intercal X_i, g(Z_i))]\right)^2
\end{aligned}$$

$$
\begin{aligned}
=&\frac{1}{q}\sum_{\ell=1}^{q}\frac{1}{n^2}\sum_{i=1}^{n}\mathrm{Var}[m_\ell(V_i,\alpha^\intercal X_i,\beta^\intercal\Phi_k(Z_i))-m_\ell(V_i,\alpha^\intercal X_i,g(Z_i))]\\
&+\frac{1}{q}\sum_{\ell=1}^{q}\left(\mathbb{E}m_\ell(V_1,\alpha^\intercal X_1,\beta^\intercal\Phi_k(Z_1))\right)^2\\
\le&\frac{1}{n}\frac{1}{q}\sum_{\ell=1}^{q}\mathrm{Var}[m_\ell(V_1,\alpha^\intercal X_1,\beta^\intercal\Phi_k(Z_1))-m_\ell(V_1,\alpha^\intercal X_1,g(Z_1))]\\
&+\frac{1}{q}\left\|\mathbb{E}m(V_1,\alpha^\intercal X_1,\beta^\intercal\Phi_k(Z_1))\right\|^2\\
\le&\frac{1}{n}\frac{1}{q}\sum_{\ell=1}^{q}\mathbb{E}[m_\ell(V_1,\alpha^\intercal X_1,\beta^\intercal\Phi_k(Z_1))-m_\ell(V_1,\alpha^\intercal X_1,g(Z_1))]^2\\
&+\frac{1}{q}\left\|\mathbb{E}m(V_1,\alpha^\intercal X_1,\beta^\intercal\Phi_k(Z_1))\right\|^2\\
\le&\frac{1}{n}\frac{1}{q}\mathbb{E}\|m(V_1,\alpha^\intercal X_1,\beta^\intercal\Phi_k(Z_1))-m(V_1,\alpha^\intercal X_1,g(Z_1))\|^2\\
&+\frac{1}{q}\left\|\mathbb{E}[m(V_1,\alpha^\intercal X_1,\beta^\intercal\Phi_k(Z_1))-m(V_1,\alpha^\intercal X_1,g(Z_1))]\right\|^2\\
\le&\frac{1}{n}\mathbb{E}|A(V_1,X_1,Z_1)\gamma_k(Z_1)|^2+\mathbb{E}|A(V_1,X_1,Z_1)|^2)\|\gamma_k(z)\|^2\\
=&o(n^{-1})+O(\|\gamma_k(z)\|^2)
\end{aligned}
$$

by Assumptions 3.1 and 3.3, the dominated convergence theorem and Cauchy-Schwarz inequality. Moreover, it is clear by Assumptions 3.3 and A.1 that

$$
\mathbb{E}[I_{12}]\le\frac{1}{q}\mathbb{E}\left\|\frac{\partial}{\partial u}m(V_1,\alpha^\intercal X_1,g(Z_1))\otimes X_1\right\|^2=O(p).
$$

Hence, $I_1=o_P(1)$ by Assumption A.2.

For I_2, by Cauchy-Schwarz inequality again,

$$
\begin{aligned}
\|I_2\|^2\le&\frac{1}{q}\sum_{\ell=1}^{q}\left(\frac{1}{n}\sum_{i=1}^{n}m_\ell(V_i,\alpha^\intercal X_i,g(Z_i))\right)^2\\
&\times\frac{1}{q}\sum_{\ell=1}^{q}\left\|\frac{1}{n}\sum_{j=1}^{n}\frac{\partial}{\partial u}m_\ell(V_j,\alpha^\intercal X_j,\beta^\intercal\Phi_k(Z_j))X_j-\frac{\partial}{\partial u}m_\ell(V_j,\alpha^\intercal X_j,g(Z_j))X_j\right\|^2\\
:=&I_{21}\times I_{22},\quad\text{say.}
\end{aligned}
$$

By virtue of the iid property and Assumption A.1,

$$
\begin{aligned}
\mathbb{E}[I_{21}]=&\frac{1}{n^2}\frac{1}{q}\sum_{\ell=1}^{q}\sum_{i=1}^{n}\mathbb{E}m_\ell(V_i,\alpha^\intercal X_i,g(Z_i))^2\\
=&\frac{1}{n}\frac{1}{q}\sum_{\ell=1}^{q}\mathbb{E}m_\ell(V_1,\alpha^\intercal X_1,g(Z_1))^2=\frac{1}{n}\frac{1}{q}\mathbb{E}\|m(V_1,\alpha^\intercal X_1,g(Z_1))\|^2\\
=&O(n^{-1}).
\end{aligned}
$$

Meanwhile, invoking of the LLN I_{22} has the same order in probability as

$$\begin{aligned}
&\frac{1}{q}\sum_{\ell=1}^{q}\left\|\mathbb{E}\left[\frac{\partial}{\partial u}m_\ell(V_1,\alpha^\intercal X_1,\beta^\intercal\Phi_k(Z_1))X_1-\frac{\partial}{\partial u}m_\ell(V_1,\alpha^\intercal X_1,g(Z_1))X_1\right]\right\|^2\\
=&\frac{1}{q}\left\|\mathbb{E}\left[\frac{\partial}{\partial u}m(V_1,\alpha^\intercal X_1,\beta^\intercal\Phi_k(Z_1))\otimes X_1-\frac{\partial}{\partial u}m_\ell(V_1,\alpha^\intercal X_1,g(Z_1))\otimes X_1\right]\right\|^2\\
\le&|\mathbb{E}[A_1(V_1,X_1,Z_1)|\gamma_k(Z_1)|\|X_1\|]|^2\le O(\|\gamma_k(z)\|^2p)\\
=&o(1)
\end{aligned}$$

due to Assumption A.3 and Cauchy-Schwarz inequality, implying $I_2=o_P(1)$.

Again, using Cauchy-Schwarz inequality gives

$$\begin{aligned}
\|I_3\|^2\le&\frac{1}{q}\sum_{\ell=1}^{q}\left(\frac{1}{n}\sum_{i=1}^{n}m_\ell(V_i,\alpha^\intercal X_i,g(Z_i))\right)^2\\
&\times\frac{1}{q}\sum_{\ell=1}^{q}\left\|\frac{1}{n}\sum_{j=1}^{n}\left(\frac{\partial}{\partial u}m_\ell(V_j,\alpha^\intercal X_j,g(Z_j))X_j-\mathbb{E}\frac{\partial}{\partial u}m_\ell(V_j,\alpha^\intercal X_j,g(Z_j))X_j\right)\right\|^2\\
=&O_P(n^{-1})O_P(p)=O_P(n^{-1}p)=o_P(1)
\end{aligned}$$

due to the iid property and Assumption A.1. This finishes the proof of $\|S_{1n}(\alpha,\beta)-s_{1n}(\alpha,g)\|=o_P(1)$.

Now, we are to show $\|S_{2n}(\alpha,\beta)-s_{2n}(\alpha,g)\|=o_P(1)$. Note that

$$\begin{aligned}
&S_{2n}(\alpha,\beta)-s_{2n}(\alpha,g)\\
=&\frac{1}{qn^2}\sum_{\ell=1}^{q}\sum_{i=1}^{n}m_\ell(V_i,\alpha^\intercal X_i,\beta^\intercal\Phi_k(Z_i))\\
&\quad\times\sum_{j=1}^{n}\frac{\partial}{\partial w}m_\ell(V_j,\alpha^\intercal X_j,\beta^\intercal\Phi_k(Z_j))\Phi_k(Z_j)\\
&-\frac{1}{qn}\sum_{\ell=1}^{q}\sum_{i=1}^{n}m_\ell(V_i,\alpha^\intercal X_i,g(Z_i))\mathbb{E}\frac{\partial}{\partial w}m_\ell(V_1,\alpha^\intercal X_1,g(Z_1))\Phi_k(Z_1)\\
=&\frac{1}{qn^2}\sum_{\ell=1}^{q}\sum_{i=1}^{n}[m_\ell(V_i,\alpha^\intercal X_i,\beta^\intercal\Phi_k(Z_i))-m_\ell(V_i,\alpha^\intercal X_i,g(Z_i))]\\
&\quad\times\sum_{j=1}^{n}\frac{\partial}{\partial w}m_\ell(V_j,\alpha^\intercal X_j,\beta^\intercal\Phi_k(Z_j))\Phi_k(Z_j)\\
&+\frac{1}{qn^2}\sum_{\ell=1}^{q}\sum_{i=1}^{n}m_\ell(V_i,\alpha^\intercal X_i,g(Z_i))\\
&\quad\times\sum_{j=1}^{n}\left(\frac{\partial}{\partial w}m_\ell(V_j,\alpha^\intercal X_j,\beta^\intercal\Phi_k(Z_j))-\frac{\partial}{\partial w}m_\ell(V_j,\alpha^\intercal X_j,g(Z_j))\right)\Phi_k(Z_j)\\
&+\frac{1}{qn}\sum_{\ell=1}^{q}\sum_{i=1}^{n}m_\ell(V_i,\alpha^\intercal X_i,g(Z_i))
\end{aligned}$$

$$\times \left(\frac{1}{n} \sum_{j=1}^{n} \frac{\partial}{\partial w} m_\ell(V_j, \alpha^\intercal X_j, g(Z_j)) \Phi_k(Z_j) - \mathbb{E} \frac{\partial}{\partial w} m_\ell(V_1, \alpha^\intercal X_1, g(Z_1)) \Phi_k(Z_1) \right)$$

$:= I_4 + I_5 + I_6, \quad$ say.

Note further by Cauchy-Schwarz inequality that

$$\begin{aligned}
\|I_4\|^2 \leq & \frac{1}{q} \sum_{\ell=1}^{q} \left(\frac{1}{n} \sum_{i=1}^{n} [m_\ell(V_i, \alpha^\intercal X_i, \beta^\intercal \Phi_k(Z_i)) - m_\ell(V_i, \alpha^\intercal X_i, g(Z_i))] \right)^2 \\
& \times \frac{1}{q} \sum_{\ell=1}^{q} \left\| \frac{1}{n} \sum_{j=1}^{n} \frac{\partial}{\partial w} m_\ell(V_j, \alpha^\intercal X_j, \beta^\intercal \Phi_k(Z_j)) \Phi_k(Z_j) \right\|^2 \\
\leq & 2 \frac{1}{q} \sum_{\ell=1}^{q} \left(\frac{1}{n} \sum_{i=1}^{n} [m_\ell(V_i, \alpha^\intercal X_i, \beta^\intercal \Phi_k(Z_i)) - m_\ell(V_i, \alpha^\intercal X_i, g(Z_i))] \right)^2 \\
& \times \frac{1}{q} \sum_{\ell=1}^{q} \left\| \frac{1}{n} \sum_{j=1}^{n} \left[\frac{\partial}{\partial w} m_\ell(V_j, \alpha^\intercal X_j, \beta^\intercal \Phi_k(Z_j)) - \frac{\partial}{\partial w} m_\ell(V_j, \alpha^\intercal X_j, g(Z_j)) \right] \Phi_k(Z_j) \right\|^2 \\
& + 2 \frac{1}{q} \sum_{\ell=1}^{q} \left(\frac{1}{n} \sum_{i=1}^{n} [m_\ell(V_i, \alpha^\intercal X_i, \beta^\intercal \Phi_k(Z_i)) - m_\ell(V_i, \alpha^\intercal X_i, g(Z_i))] \right)^2 \\
& \times \frac{1}{q} \sum_{\ell=1}^{q} \left\| \frac{1}{n} \sum_{j=1}^{n} \frac{\partial}{\partial w} m_\ell(V_j, \alpha^\intercal X_j, g(Z_j)) \Phi_k(Z_j) \right\|^2,
\end{aligned}$$

where due to Assumption A.3 the second term is the leading one, which by the LLN has the same order as

$$\begin{aligned}
& \frac{1}{q} \sum_{\ell=1}^{q} \left(\mathbb{E}[m_\ell(V_1, \alpha^\intercal X_1, \beta^\intercal \Phi_k(Z_1)) - m_\ell(V_1, \alpha^\intercal X_1, g(Z_1))] \right)^2 \\
& \qquad \times \frac{1}{q} \sum_{\ell=1}^{q} \left\| \mathbb{E} \frac{\partial}{\partial w} m_\ell(V_1, \alpha^\intercal X_1, g(Z_1)) \Phi_k(Z_1) \right\|^2 \\
= & \frac{1}{q} \left\| \mathbb{E}[m(V_1, \alpha^\intercal X_1, \beta^\intercal \Phi_k(Z_1)) - m(V_1, \alpha^\intercal X_1, g(Z_1))] \right\|^2 \\
& \times \frac{1}{q} \left\| \mathbb{E} \frac{\partial}{\partial w} m(V_1, \alpha^\intercal X_1, g(Z_1)) \otimes \Phi_k(Z_1) \right\|^2 \\
\leq & \left| \mathbb{E}[A(V_1, X_1, Z_1) \gamma_k(Z_1)] \right|^2 O(k) \leq O(\|\gamma_k(z)\|^2 k) = o(1)
\end{aligned}$$

in probability by Assumption A.2 as $n \to \infty$.

Moreover, invoking Assumptions A.2-A.3, $I_5 = o_P(1)$. Finally,

$$\begin{aligned}
\|I_6\|^2 \leq & \frac{1}{q} \sum_{\ell=1}^{q} \left(\frac{1}{n} \sum_{i=1}^{n} m_\ell(V_i, \alpha^\intercal X_i, g(Z_i)) \right)^2 \\
& \times \frac{1}{q} \sum_{\ell=1}^{q} \left\| \frac{1}{n} \sum_{j=1}^{n} \left[\frac{\partial}{\partial w} m_\ell(V_j, \alpha^\intercal X_j, g(Z_j)) \Phi_k(Z_j) - \mathbb{E} \frac{\partial}{\partial w} m_\ell(V_i, \alpha^\intercal X_i, g(Z_i)) \Phi_k(Z_i) \right] \right\|^2
\end{aligned}$$

$:= I_{61} \times I_{62}$, say.

Here, $I_{61} = I_{21}$ and thus $\mathbb{E}[I_{61}] = O(n^{-1})$. Meanwhile,

$$\begin{aligned}
\mathbb{E}[I_{62}] =& \frac{1}{q}\frac{1}{n^2}\sum_{\ell=1}^{q}\sum_{j=1}^{n}\mathbb{E}\left\|\frac{\partial}{\partial w}m_\ell(V_j, \alpha^\intercal X_j, g(Z_j))\Phi_k(Z_j) - \mathbb{E}\frac{\partial}{\partial w}m_\ell(V_i, \alpha^\intercal X_i, g(Z_i))\Phi_k(Z_i)\right\|^2 \\
=& \frac{1}{q}\frac{1}{n}\sum_{\ell=1}^{q}\mathbb{E}\left\|\frac{\partial}{\partial w}m_\ell(V_1, \alpha^\intercal X_1, g(Z_1))\Phi_k(Z_1) - \mathbb{E}\frac{\partial}{\partial w}m_\ell(V_1, \alpha^\intercal X_1, g(Z_1))\Phi_k(Z_1)\right\|^2 \\
=& \frac{1}{q}\frac{1}{n}\mathbb{E}\left\|\frac{\partial}{\partial w}m(V_1, \alpha^\intercal X_1, g(Z_1)) \otimes \Phi_k(Z_1) - \mathbb{E}\frac{\partial}{\partial w}m(V_1, \alpha^\intercal X_1, g(Z_1)) \otimes \Phi_k(Z_1)\right\|^2 \\
\leq& \frac{1}{q}\frac{1}{n}\mathbb{E}\left\|\frac{\partial}{\partial w}m(V_1, \alpha^\intercal X_1, g(Z_1)) \otimes \Phi_k(Z_1)\right\|^2 = O(n^{-1}k) = o(1)
\end{aligned}$$

appealing to Assumptions A.1-A.2, implying $\|I_6\|^2 = o_P(n^{-1}k) = o_P(1)$. The proof is complete. □

B Proofs of the main results

Proof of Theorem 3.1. In Lemma A.1, we have shown that

(i) $\|M_n(\alpha, \beta)\|^2 = o_P(1)$,

(ii) $\sup_{\substack{\|\mathbf{a}\| \leq B_{1n}, \|\mathbf{b}\| \leq B_{2n} \\ \|(\mathbf{a}-\alpha, \mathbf{b}-\beta)\| > \delta}} \|M_n(\mathbf{a}, \mathbf{b})\|^{-2} = O_P(1)$ for each $\delta > 0$.

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

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

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

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

which gives

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

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

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

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,

$$\limsup P\left(\|(\widehat{\alpha}-\alpha,\widehat{\beta}-\beta)\| > \delta\right) \leq 2\epsilon.$$

As ϵ and δ are arbitrarily chosen, we 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. By the first order condition $S_n(\widehat{\alpha},\widehat{\beta}) = 0$, consistency and Taylor expansion, we have expansion

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

where higher order term is omitted. As shown in Lemmas A.2-A.3, under Assumptions 2.1-2.2, 3.1, 3.3 and A.1-A.3 in Appendix A, $H_n(\alpha,\beta)$ 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) and (A.2)), respectively, that is, $\|H_n(\alpha,\beta)-h_n(\alpha,g)\| = o_P(1)$ and $\|S_n(\alpha,\beta)-s_n(\alpha,g)\| = o_P(1)$. Thence,

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

Noting that $\widehat{g}(z)-g(z) = \Phi_k(z)^\intercal(\widehat{\beta}-\beta) - \gamma_k(z)$, the linearity of Fréchet derivative and ignoring the higher order term in the definition of Fréchet derivative,

$$\begin{aligned}
&\begin{pmatrix}\mathscr{L}(\widehat{\alpha})-\mathscr{L}(\alpha) \\ \mathscr{F}(\widehat{g})-\mathscr{F}(g)\end{pmatrix} = \begin{pmatrix}\mathscr{L}(\widehat{\alpha}-\alpha) \\ \mathscr{F}'(g)(\widehat{g}(z)-g(z))\end{pmatrix} \\
&= \begin{pmatrix}\mathscr{L}(\widehat{\alpha}-\alpha) \\ \mathscr{F}'(g)\Phi_k(z)^\intercal(\widehat{\beta}-\beta)\end{pmatrix} - \begin{pmatrix}0 \\ \mathscr{F}'(g)\gamma_k(z)\end{pmatrix} \\
&= \begin{pmatrix}\mathscr{L} & 0 \\ 0 & \mathscr{F}'(g)\Phi_k(z)^\intercal\end{pmatrix}\begin{pmatrix}\widehat{\alpha}-\alpha \\ \widehat{\beta}-\beta\end{pmatrix} - \begin{pmatrix}0 \\ \mathscr{F}'(g)\gamma_k(z)\end{pmatrix}
\end{aligned}$$

$$= - \begin{pmatrix} \mathscr{L} & 0 \\ 0 & \mathscr{F}'(g)\Phi_k(z)^\intercal \end{pmatrix} h_n(\alpha, g)^{-1} s_n(\alpha, g) - \begin{pmatrix} 0 \\ \mathscr{F}'(g)\gamma_k(z) \end{pmatrix}$$

$$:= \Lambda_{1n} + \Lambda_{2n}, \quad \text{say.}$$

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) and (A.2). Hence, $\Lambda_{1n} = \frac{1}{n}\Gamma_n(\Psi_n\Psi_n^\intercal)^{-1}\Psi_n \sum_{i=1}^n m(V_i, \alpha^\intercal X_i, g(Z_i))$ where

$$\Gamma_n = - \begin{pmatrix} \mathscr{L} & 0 \\ 0 & \mathscr{F}'(g(z))\Phi_k(z)^\intercal \end{pmatrix}.$$

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 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 [29]. 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. By the conventional 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

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

Toward this end, we shall show

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

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

(1). Notice that

$$\begin{aligned}\frac{1}{n}D_n(\widehat{\alpha},\widehat{\beta};\kappa)^2=&\frac{1}{n}\sum_{i=1}^{n}[\kappa^{\intercal}m(V_i,\widehat{\alpha}^{\intercal}X_i,\widehat{\beta}^{\intercal}\Phi_k(Z_i))]^2\\=&\frac{1}{n}\sum_{i=1}^{n}[\kappa^{\intercal}m(V_i,\alpha^{\intercal}X_i,g(Z_i))]^2\\&+\frac{1}{n}\sum_{i=1}^{n}\{[\kappa^{\intercal}m(V_i,\widehat{\alpha}^{\intercal}X_i,\widehat{g}(Z_i))]^2-[\kappa^{\intercal}m(V_i,\alpha^{\intercal}X_i,g(Z_i))]^2\}\end{aligned}$$

and we shall show that the second term is $o_P(1)$. First of all, we need the convergence rate of $\|\widehat{\alpha}-\alpha\|^2$ and $\|\widehat{\beta}-\beta\|^2$. It follows from (B.1) in the proof of Theorem 3.2 that $((\widehat{\alpha}-\alpha)^{\intercal},(\widehat{\beta}-\beta)^{\intercal})$ has leading term $h_n(\alpha,g)^{-1}s_n(\alpha,g)$. Then, by the expressions of $h_n(\alpha,g)$ and $s_n(\alpha,g)$ it is readily seen that $\|\widehat{\alpha}-\alpha\|^2=O_P(q/n)$ and $\|\widehat{\beta}-\beta\|^2=O_P(q/n)$.

Moreover, by the first order Taylor expansion,

$$\begin{aligned}&\frac{1}{n}\sum_{i=1}^{n}|[\kappa^{\intercal}m(V_i,\widehat{\alpha}^{\intercal}X_i,\widehat{g}(Z_i))]^2-[\kappa^{\intercal}m(V_i,\alpha^{\intercal}X_i,g(Z_i))]^2|\\\le&\frac{1}{n}\sum_{i=1}^{n}|\kappa^{\intercal}[m(V_i,\widehat{\alpha}^{\intercal}X_i,\widehat{g}(Z_i))-m(V_i,\alpha^{\intercal}X_i,g(Z_i))]|^2\\&+2\frac{1}{n}\sum_{i=1}^{n}|\kappa^{\intercal}[m(V_i,\widehat{\alpha}^{\intercal}X_i,\widehat{g}(Z_i))-m(V_i,\alpha^{\intercal}X_i,g(Z_i))]||\kappa^{\intercal}m(V_i,\alpha^{\intercal}X_i,g(Z_i))|\\\le&\frac{2}{n}\sum_{i=1}^{n}\left|\kappa^{\intercal}\frac{\partial m(V_i,\alpha^{\intercal}X_i,g(Z_i))}{\partial u}(\widehat{\alpha}-\alpha)^{\intercal}X_i\right|^2\\&+\frac{2}{n}\sum_{i=1}^{n}\left|\kappa^{\intercal}\frac{\partial m(V_i,\alpha^{\intercal}X_i,g(Z_i))}{\partial w}(\widehat{g}(Z_i)-g(Z_i))\right|^2\\&+\frac{2}{n}\sum_{i=1}^{n}\left|\kappa^{\intercal}\frac{\partial m(V_i,\alpha^{\intercal}X_i,g(Z_i))}{\partial u}(\widehat{\alpha}-\alpha)^{\intercal}X_i\right||\kappa^{\intercal}m(V_i,\alpha^{\intercal}X_i,g(Z_i))|\\&+\frac{2}{n}\sum_{i=1}^{n}\left|\kappa^{\intercal}\frac{\partial m(V_i,\alpha^{\intercal}X_i,g(Z_i))}{\partial w}(\widehat{g}(Z_i)-g(Z_i))\right||\kappa^{\intercal}m(V_i,\alpha^{\intercal}X_i,g(Z_i))|\\\le&\|\widehat{\alpha}-\alpha\|^2\frac{2}{n}\sum_{i=1}^{n}\left\|\frac{\partial m(V_i,\alpha^{\intercal}X_i,g(Z_i))}{\partial u}\otimes X_i\right\|^2\\&+\|\widehat{\beta}-\beta\|^2\frac{4}{n}\sum_{i=1}^{n}\left\|\frac{\partial m(V_i,\alpha^{\intercal}X_i,g(Z_i))}{\partial w}\otimes\Phi_k(Z_i)\right\|^2\\&+\frac{4}{n}\sum_{i=1}^{n}\left\|\frac{\partial m(V_i,\alpha^{\intercal}X_i,g(Z_i))}{\partial w}\right\|^2\gamma_k^2(Z_i)\\&+\|\widehat{\alpha}-\alpha\|\frac{2}{n}\left(\sum_{i=1}^{n}\left\|\frac{\partial m(V_i,\alpha^{\intercal}X_i,g(Z_i))}{\partial u}\otimes X_i\right\|^2\right)^{1/2}\left(\sum_{i=1}^{n}\|m(V_i,\alpha^{\intercal}X_i,g(Z_i))\|^2\right)^{1/2}\end{aligned}$$

$$+\frac{2}{n}\left(\sum_{i=1}^{n}\left\|\frac{\partial m(V_i,\alpha^\intercal X_i,g(Z_i))}{\partial w}(\widehat{g}(Z_i)-g(Z_i))\right\|^2\right)^{1/2}\left(\sum_{i=1}^{n}\|m(V_i,\alpha^\intercal X_i,g(Z_i))\|^2\right)^{1/2}$$
$$=\|\widehat{\alpha}-\alpha\|^2 O_P(qp)+\|\widehat{\beta}-\beta\|^2 O_P(qk)+\sup_z \gamma_k^2(z)O_P(q)=o_P(1)$$

by Assumptions A.1 and 4.2, where Cauchy-Schwarz inequality is used to show the last two sums are of smaller order. Thus, the assertion of (1) holds.

(2). We first consider

$$\nu_n(\mathbf{a},f;\kappa)=\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\kappa^\intercal(m(V_i,\mathbf{a}^\intercal X_i,f(Z_i))-E[m(V_i,\mathbf{a}^\intercal X_i,f(Z_i))]), \tag{B.2}$$

for any $\kappa\in\mathbb{R}^q$ such that $\|\kappa\|=1$ and $(\mathbf{a},f)\in\Theta$. Because of the convergence in Theorem 3.2, we eventually will show $\nu_n(\widehat{\alpha},\widehat{g};\kappa)-\nu_n(\alpha,g;\kappa)=o_P(1)$.

Notice by the first order Taylor expansion that

$$\begin{aligned}&m(V_i,\mathbf{a}^\intercal X_i,f(Z_i))-m(V_i,\alpha^\intercal X_i,g(Z_i))\\ =&\frac{\partial m(V_i,\alpha^\intercal X_i,g(Z_i))}{\partial u}(\mathbf{a}-\alpha)^\intercal X_i+\frac{\partial m(V_i,\alpha^\intercal X_i,g(Z_i))}{\partial w}(f(Z_i)-g(Z_i)),\end{aligned}$$

for all $(\mathbf{a},f)$ in the neighbourhood of (α,g), where f has the form $\mathbf{b}^\intercal\Phi_k(\cdot)$. Thus,

$$\begin{aligned}&P\left(\sup_{\|(\mathbf{a},f)-(\alpha,g)\|<\delta}|\nu_n(\mathbf{a},f;\kappa)-\nu_n(\alpha,g;\kappa)|>\eta\right)\\ \le&P\left(\sup_{\|(\mathbf{a},f)-(\alpha,g)\|<\delta}\left|\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\kappa^\intercal[\frac{\partial m}{\partial u}(\mathbf{a}-\alpha)^\intercal X_i-E\frac{\partial m}{\partial u}(\mathbf{a}-\alpha)^\intercal X_i]\right|>\eta/2\right)\\ &+P\left(\sup_{\|(\mathbf{a},f)-(\alpha,g)\|<\delta}\left|\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\kappa^\intercal[\frac{\partial m}{\partial w}(f(Z_i)-g(Z_i))-E\frac{\partial m}{\partial w}(f(Z_i)-g(Z_i))]\right|>\eta/2\right)\\ \le&P\left(\sup_{\|(\mathbf{a},f)-(\alpha,g)\|<\delta}\left|\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\left[\kappa^\intercal\frac{\partial m}{\partial u}X_i-E\kappa^\intercal\frac{\partial m}{\partial u}X_i\right]^\intercal(\mathbf{a}-\alpha)\right|>\eta/2\right)\\ &+P\left(\sup_{\|(\mathbf{a},f)-(\alpha,g)\|<\delta}\left|\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\left[\kappa^\intercal\frac{\partial m}{\partial w}\Phi_k(Z_i)-E\kappa^\intercal\frac{\partial m}{\partial w}\Phi_k(Z_i)\right]^\intercal(\mathbf{b}-\beta)\right|>\eta/4\right)\\ &+P\left(\sup_{\|(\mathbf{a},f)-(\alpha,g)\|<\delta}\left|\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\left[\kappa^\intercal\frac{\partial m}{\partial w}\gamma_k(Z_i)-E\kappa^\intercal\frac{\partial m}{\partial w}\gamma_k(Z_i)\right]\right|>\eta/4\right)\\ \le&P\left(\sup_{\|(\mathbf{a},f)-(\alpha,g)\|<\delta}\left\|\frac{1}{\sqrt{np}}\sum_{i=1}^{n}\left[\kappa^\intercal\frac{\partial m}{\partial u}X_i-E\kappa^\intercal\frac{\partial m}{\partial u}X_i\right]\right\|\|\sqrt{p}(\mathbf{a}-\alpha)\|>\eta/2\right)\\ &+P\left(\sup_{\|(\mathbf{a},f)-(\alpha,g)\|<\delta}\left\|\frac{1}{\sqrt{nk}}\sum_{i=1}^{n}\left[\kappa^\intercal\frac{\partial m}{\partial w}\Phi_k(Z_i)-E\kappa^\intercal\frac{\partial m}{\partial w}\Phi_k(Z_i)\right]\right\|\|\sqrt{k}(\mathbf{b}-\beta)\|>\eta/4\right)\\ &+P\left(\sup_{\|(\mathbf{a},f)-(\alpha,g)\|<\delta}\left|\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\left[\kappa^\intercal\frac{\partial m}{\partial w}\gamma_k(Z_i)-E\kappa^\intercal\frac{\partial m}{\partial w}\gamma_k(Z_i)\right]\right|>\eta/4\right)\end{aligned}$$

$:= I_{1n} + I_{2n} + I_{3n}$, say.

Observe by the classical CLT that

$$\frac{1}{\sqrt{np}} \sum_{i=1}^{n} \left[\kappa^{\intercal} \frac{\partial m}{\partial u} X_i - E\kappa^{\intercal} \frac{\partial m}{\partial u} X_i \right] = O_P(1),$$
$$\frac{1}{\sqrt{nk}} \sum_{i=1}^{n} \left[\kappa^{\intercal} \frac{\partial m}{\partial w} \Phi_k(Z_i) - E\kappa^{\intercal} \frac{\partial m}{\partial w} \Phi_k(Z_i) \right] = O_P(1).$$

It follows that if $\|\sqrt{p}(\mathbf{a}-\alpha)\|$ and $\|\sqrt{k}(\mathbf{b}-\beta)\|$ are sufficient small, $I_{1n} < \varepsilon/3$ and $I_{2n} < \varepsilon/3$. Meanwhile, using the condition that $\sqrt{n}\sup_z \|\gamma_k(z)\| = o(1)$ we have $I_{3n} < \varepsilon/3$. This shows that, in view of Theorem 3.2, when n is large, $P\left(|\nu_n(\widehat{\alpha}, \widehat{g}; \kappa) - \nu_n(\alpha, g; \kappa)| > \eta\right) < \varepsilon$ for any given $\varepsilon, \eta > 0$.

Furthermore, since

$$\frac{1}{\sqrt{n}} \sum_{i=1}^{n} \kappa^{\intercal} [m(V_i, \widehat{\alpha}^{\intercal} X_i, \widehat{\beta}^{\intercal} \Phi_k(Z_i)) - m(V_i, \alpha^{\intercal} X_i, g(Z_i))]$$
$$= \nu_n(\widehat{\alpha}, \widehat{g}; \kappa) - \nu_n(\alpha, g; \kappa) + \sqrt{n}\overline{m}_n^*(\widehat{\alpha}, \widehat{g}; \kappa),$$

the assertion of (2) holds by virtue of Assumption 4.1. This finishes the proof. □

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

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

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 Number that

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

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

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

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

Proof of Lemma 5.1. Define $\rho_n = a_n + \sqrt{t}\, P_n'(d_n)$ and then $\rho_n = o(1)$ by Assumption 5.1. Denote $\mathcal{N}_\tau = \{v \in \mathbb{R}^{p+k} : \ \|v_T - v_0\| \le \rho_n \tau\}$ for $\tau > 0$. Let $\partial\mathcal{N}_\tau$ be the boundary of $\mathcal{N}_\tau$. Also, define an event

$$A_n(\tau) = \left\{ Q_n(v_0) < \inf_{v \in \partial\mathcal{N}_\tau} Q_n(v_T) \right\}.$$

On the event $A_n(\tau)$, by the continuity of $Q_n(v)$ with respect to v_j for $j \in T$, there exists a local minimizer of $Q_n(v_T)$ inside $\mathcal{N}_\tau$. That is, there exists a local minimizer $\widehat{v} \in \mathcal{V}$ of $Q_n(v_T)$ such that $\|\widehat{v} - v_0\| < \tau\rho_n$. Therefore, it suffices to show that for $\forall \epsilon > 0$, there exists a $\tau > 0$ such that $P(A_n(\tau)) \ge 1 - \epsilon$ for all large n.

For any $v \in \partial\mathcal{N}_\tau$, viz. $\|v_T - v_0\| = \tau\rho_n$, there is an v^* lying on the segment joining v and v_0 such that by the mean value theorem,

$$\begin{aligned} Q_n(v_T) - Q_n(v_0) =& (v_S - v_{0S})^\intercal S_{nT}(v_{0S}) + \frac{1}{2}(v_S - v_{0S})^\intercal H_{nT}(v_S^*)(v_S - v_{0S}) \\ &+ \sum_{j \in T} [P_n(|v_{Sj}|) - P_n(|v_{0S,j}|)], \end{aligned}$$

where v_{0S} and v_S are defined before, so is v_S^*.

Invoking the condition $\|S_{nT}(v_{0S})\| = O_P(a_n)$, for $\forall \epsilon > 0$, there exists a $C_1 > 0$ such that the event A_1 given below satisfies $P(A_1) > 1 - \epsilon/4$ for all large n, where

$$A_1 = \{(v_S - v_{0S})^\intercal S_{nT}(v_{0S}) \ge -C_1 a_n \|v_S - v_{0S}\|\}.$$

Also, by Condition (ii) and for this ϵ, there exists a C_2 such that $P(A_2) > 1 - \epsilon/4$ for all large n, where

$$A_2 = \{(v_S - v_{0S})^\intercal H_{nT}(v_{0S})(v_S - v_{0S}) \ge C_2 \|v_S - v_{0S}\|^2\}.$$

Meanwhile, define event $A_3 = \{\|H_{nT}(v_{0S}) - H_{nT}(v_S^*)\| \ge C_2/4\}$. By Condition (iii) and $\|v_T - v_0\| = \|v_S - v_{0S}\| = \tau\rho_n$, for any τ, $P(A_3) \ge 1 - \epsilon/4$ for all large n. Hence, $A_4 \subset A_2 \cap A_3$ where

$$A_4 = \{(v_S - v_{0S})^\intercal H_{nT}(v_S^*)(v_S - v_{0S}) > \frac{3}{4} C_2 \|v_S - v_{0S}\|^2\}.$$

On the other hand, it follows from Lemma B.1 in Fan and Liao [18] that $\sum_{j \in T}[P_n(|v_{Sj}|) - P_n(|v_{0S,j}|)] \ge -\sqrt{t}\, P_n'(d_n)\|v_S - v_{0S}\|$. Whence, for any $v \in \partial\mathcal{N}_\tau$, on $A_1 \cap A_4$,

$$Q_n(v_T) - Q_n(v_0) \ge \rho_n \tau \left(\frac{3}{8} \rho_n \tau C_2 - C_1 a_n - \sqrt{t}\, P_n'(d_n) \right).$$

For $\rho_n = a_n + \sqrt{t}\, P_n'(d_n)$, $C_1 a_n + \sqrt{t}\, P_n'(d_n) \le (C_1 + 1)\rho_n$. Thus, choosing $\tau > 8(C_1 + 1)/3C_2$ yields that $Q_n(v_T) - Q_n(v_0) > 0$ uniformly on $v \in \partial\mathcal{N}_\tau$. It follows that for all large n, with $\tau > 8(C_1 + 1)/3C_2$, $P(A_n(\tau)) > P(A_1 \cap A_4) \ge 1 - \epsilon$.

We next show that the local minimizer, denoted by $\widehat{v} \in \mathcal{V}$, is strict with a probability arbitrarily close to one. For each $h \neq 0$, define

$$\psi(h) = \limsup_{\epsilon \to 0^+} \sup_{(u_1,u_2) \in O(|h|,\epsilon)} -\frac{P_n'(u_2) - P_n'(u_1)}{u_2 - u_1}.$$

By the concavity, $\psi(\cdot) \geq 0$. For any $v \in \mathcal{N}_\tau$, let $\Omega(v) = H_{nT}(v_S) - \mathrm{diag}(\psi(v_{S1}), \cdots, \psi(v_{St}))$. It suffices to show that $\Omega(\widehat{v})$ is positive definite with probability arbitrarily close to unity. On the event $A_5 = \{\phi(\widehat{v}_S) \leq \sup_{v_S \in O(v_{0S}, cd_n)} \phi(v_S)\}$ where $\widehat{v}_S$ is the t-vector consisting of nonzero elements of $\widehat{v}$, and c is the same in (iv) of Assumption 5.1, we have

$$\max_{j \leq t} \psi(\widehat{v}_{Sj}) \leq \phi(\widehat{v}_S) \leq \sup_{v_S \in O(v_{0S}, cd_n)} \phi(v_S).$$

Let $A_6 = \{\|H_{nT}(\widehat{v}_S) - H_{nT}(v_{0S})\| < C_2/4\}$ and $A_7 = \{\lambda_{\min}(H_{nT}(v_{0S})) > C_2\}$. Then, for any $u \in \mathbb{R}^t$ with $\|u\| = 1$, it follows from (iv) of Assumption 5.1 that

$$\begin{aligned} u^\intercal \Omega(\widehat{v}) u =& u^\intercal H_{nT}(\widehat{v}_S) u - u^\intercal \mathrm{diag}(\psi(\widehat{v}_{S1}), \cdots, \psi(\widehat{v}_{St})) u \\ \geq& u^\intercal H_{nT}(v_{0S}) u - |u^\intercal [H_{nT}(\widehat{v}_S) - H_{nT}(v_{0S})] u| - \max_{j \leq s} \psi(\widehat{v}_{Sj}) \\ \geq& 3C_2/4 - \sup_{v_S \in O(v_{0S}, cd_n)} \phi(v_S) \geq C_2/4 \end{aligned}$$

on the event $A_5 \cap A_6 \cap A_7$ for all large n.

Finally, we are about to show that $P(A_5 \cap A_6 \cap A_7) \geq 1 - \epsilon$. As $P(A_7) \geq 1 - \epsilon$, it suffices to show $P(A_5 \cap A_6) \geq 1 - \epsilon$ for $\forall \epsilon > 0$. Indeed, due to $\rho_n = o(d_n)$, $P(A_5) \geq P(\widehat{v}_S \in O(v_{0S}, cd_n)) \geq 1 - \epsilon/2$ for all large n. Also,

$$\begin{aligned} P(A_6^c) \leq& P(A_6^c, \|\widehat{v} - v_0\| \leq \rho_n) + P(\|\widehat{v} - v_0\| > \rho_n) \\ \leq& P\left(\sup_{v_S \in O(v_{0S}, cd_n)} \|H_{nT}(v_S) - H_{nT}(v_{0S})\| \geq C_2/4 \right) + \epsilon/4 \leq \epsilon/2. \end{aligned}$$

□

Proof of Lemma 5.2. Recall that $\widehat{v} \in \mathcal{V}$ is a local minimizer of $Q_n(v_T)$. Hence, there is a small neighbourhood O_1 of $\widehat{v}$ such that for any $v \in O_1$ with $v \notin \mathcal{V}$ we have $Q_n(\widehat{v}) \leq Q_n(v_T)$. However, by the condition of (5.2),

$$Q_n(v_T) - Q_n(v) = \|M_n(v_T)\|^2 - \|M_n(v)\|^2 - \sum_{j \notin T} P_n(|v_j|) < 0. \tag{B.3}$$

This means $Q_n(\widehat{v}) < Q_n(v)$, yielding the first assertion, while, from which and the last statement of Lemma 5.1, the second assertion is also implied. □

Verification of Conditions in Lemma 5.1 Condition (i): Notice that $S_{nT}(v_{0S}) = \partial_{v_{0S}}\|M_n(v_0)\|^2 = 2A_n(v_{0S})M_n(v_0)$, where

$$A_n(v_{0S}) = \frac{1}{\sqrt{q}n}\sum_{i=1}^{n}\partial m^\intercal(V_i, v_{0S}^\intercal F_{iS})\otimes F_{iS}.$$

By Assumption 5.2, $\|A_n(v_{0S})\| = O_P(\sqrt{t})$. Meanwhile, due to $Em(\cdot) = 0$ at the true parameter, by virtue of Assumption 5.3, Bernstein inequality and Bonferroni inequality, there exist $C > 0$, for any $u > 0$,

$$\begin{aligned}
&P\left(\max_{\ell\le q}\left|\frac{1}{n}\sum_{i=1}^{n} m_\ell(V_i, v_{0S}^\intercal F_{iS})\right| > u\right)\\
\le& q\max_{\ell\le q} P\left(\left|\frac{1}{n}\sum_{i=1}^{n} m_\ell(V_i, v_{0S}^\intercal F_{iS})\right| > u\right)\\
\le& \exp(\log q - Cu^2/n).
\end{aligned}$$

Hence, $\max_{\ell\le q}\left|\frac{1}{n}\sum_{i=1}^{n} m_\ell(V_i, \alpha_{0S}^\intercal X_{iS}, \beta_{0S}^\intercal\Phi_{kS}(Z_i))\right| = O_P(\sqrt{\log(q)/n})$, which then gives

$$\|M_n(v_0)\| = \left\|\frac{1}{\sqrt{q}n}\sum_{i=1}^{n} m(V_i, \alpha_{0S}^\intercal X_{iS}, \beta_{0S}^\intercal\Phi_{kS}(Z_i))\right\| = O_P(\sqrt{\log(q)/n}). \tag{B.4}$$

Accordingly, $\|S_{nT}(v_{0S})\| = O_P(\sqrt{t\ \log(q)/n})$.

Condition (ii): It is clear that $H_{nT}(v_S) = 2A_n(v_S)A_n(v_S)^\intercal + 2A_{1n}(v_S)M_n(v_T)$ where

$$A_{1n}(v_S) = \frac{1}{\sqrt{q}n}\sum_{i=1}^{n}\partial^2 m(V_i, v_{0S}^\intercal F_{iS})\otimes F_{iS}F_{iS}^\intercal.$$

Here, $\partial^2 m$ stands for the second order partial derivative of m with respect to its arguments where the parameter is involved.

As shown in Lemma A.2 that $A_n(v_S)A_n(v_S)^\intercal$ is almost surely positive definite, while similar to the verification of Condition (i), the second term is $o_P(1)$. Thus, using Assumption 5.4, the condition can be verified using arguments similar to Fan and Liao [18].

Condition (iii): Observe that

$$\begin{aligned}
&H_{nT}(v_S) - H_{nT}(v_{0S})\\
=&2[A_n(v_S)A_n(v_S)^\intercal - A_n(v_{0S})A_n(v_{0S})^\intercal] + 2A_{1n}(v_S)M_n(v_T) + 2A_{1n}(v_{0S})M_n(v_0)\\
=&2[A_n(v_S) - A_n(v_{0S})]A_n(v_S)^\intercal] + 2A_n(v_{0S})[A_n(v_S) - A_n(v_{0S})]^\intercal]\\
&+ 2A_{1n}(v_S)M_n(v_T) + 2A_{1n}(v_{0S})M_n(v_0),
\end{aligned}$$

and each term is $o_P(1)$, from which the condition follows.

Verification of the condition in Lemma 5.2: Let $\widehat{v} \in \mathcal{V}$ be the minimizer of Q_n. We shall show that there is a neighbourhood of $\widehat{v}$ in which for any $v \notin \mathcal{V}$, the condition of

(5.2) holds, that is, $\|M_n(v_T)\|^2 - \|M_n(v)\|^2 < \sum_{j\notin T} P_n(|v_j|)$. This is tantamount to showing $Q_n(v_T) < Q_n(v)$.

Using the mean value theorem, there exists a v_* on the segment joining v_T and v such that

$$\|M_n(v_T)\|^2 - \|M_n(v)\|^2 = S_n(v_*)^\intercal(v_T - v) = S_n(v_*)^\intercal v_{T^c},$$

where T^c is the complement set of T w.r.t. $\{1, \cdots, p+k\}$ and noting $v = v_T + v_{T^c}$ for any v.

Here, we know $\|S_n(v_{0S})\| = O_P(\sqrt{t\log(q)/n})$, $\|\widehat{v} - v_0\| = O_P(\sqrt{t\log(q)/n} + \sqrt{t}\, P_n'(d_n))$. In a small neighbourhood of $\widehat{v}$, $O(\widehat{v}, r_n/(p+k))$ say, where r_n is a sufficient small number, $\|S_n(v)\| = O_P(\sqrt{t\log(q)/n})$ uniformly holds in v and $\sup_{v\in O}\|v - \widehat{v}\|_1 \le r_n$.

On the other hand, for some $\mu \in (0,1)$, $\sum_{j\notin T} P_n(|v_j|) = \sum_{j\notin T, v_j\neq 0} |v_j| P_n'(\mu|v_j|) \ge \sum_{j\notin T, v_j\neq 0} |v_j| P_n'(r_n)$ by the nonincreasingness of $P_n'(u)$. Let r_n so small that $P_n'(r_n) \ge P_n'(0^+)/2$. Hence, $\sum_{j\notin S} P_n(|\beta_j|) \ge C r_n$ in probability.

Then, by virtue of Assumption 5.4 and following a similar argument as Fan and Liao [18], the condition is verified.

Proof of Theorem 5.1. (i) and (ii). As shown in Lemma 5.2, 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,

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

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 $P_n'(|\widehat{v}_S|) := (P_n'(|\widehat{v}_{S1}|), \cdots, P_n'(|\widehat{v}_{St}|))^\intercal$ and $\mathrm{sgn}(\widehat{v}_S) := (\mathrm{sgn}(\widehat{v}_{S1}), \cdots, \mathrm{sgn}(\widehat{v}_{St}))^\intercal$.

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, which further implies

$$\widehat{v}_S - v_{0S} = H_{nT}(v_{0S})^{-1}[S_{nT}(\widehat{v}_S) - S_{nT}(v_{0S})]$$

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

under the condition for $t = p_1 + k_1$ 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 $\widehat{g}(z) := \Phi_{kT}(z)^\intercal \widehat{\beta}_S$,

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

Notice that the structure

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

is standard, so that invoking classical central limit theorem 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 [18] we may show that

$$
\|P_n'(|\widehat{v}_S|) \diamond \operatorname{sgn}(\widehat{v}_S)\| = O_P(\max_{\|v_S - v_{0S}\| \le d_n/4} \phi(v_S) \sqrt{t \log(q)/n} + P_n'(d_n)).
$$

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 $\|M_n(\widehat{v}_T)\|^2 = O_P(t^{3/2} \log(q)/n + t^{3/2} P_n'(d_n)^2 + t\sqrt{\log(q)/n} P_n'(d_n))$. 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,

$$
|S_{nT}(v_{0S})^\intercal (\widehat{v}_S - v_{0S})| \le \|S_{nT}(v_{0S})\| \|\widehat{v}_S - v_{0S}\| = O_P(t \log(q)/n + t\sqrt{\log(q)/n} P_n'(d_n))
$$

due to $\|S_{nT}(v_{0S})\| = O_P(\sqrt{t\,\log(q)/n})$ and $\|\widehat{v}_S - v_{0S}\| = O_P(\sqrt{t\,\log(q)/n} + \sqrt{t}P'_n(d_n))$. Meanwhile, it follows from Assumption 5.2 that

$$\begin{aligned}
&|[S_{nT}(v_S^*) - S_{nT}(v_{0S})]^\intercal(\widehat{v}_S - v_{0S})| \le \|S_{nT}(v_S^*) - S_{nT}(v_{0S}\|\|\widehat{v}_S - v_{0S}\| \\
\le& O_P(\sqrt{t})\|v_S^* - v_{0S}\|\|\widehat{v}_S - v_{0S}\| \le O_P(\sqrt{t})\|\widehat{v}_S - v_{0S}\|^2 \\
=& O_P(t^{3/2}\log(q)/n + t^{3/2}P'_n(d_n)^2).
\end{aligned}$$

The assertion then follows by noting from (B.4) that $\|M_n(v_0)\|^2 = \log(q)/n$.

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

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

from which the assertion follows.

Now, for any $\delta > 0$,

$$\begin{aligned}
&\inf_{\|v-v_0\|\ge\delta}Q_n(v) \ge \inf_{\|v-v_0\|\ge\delta}\|M_n(v)\|^2 \\
=& \inf_{\|v-v_0\|\ge\delta}\frac{1}{q}\left\|\frac{1}{n}\sum_{i=1}^n m(V_i, v^\intercal F_i)\right\|^2 \\
\ge& \inf_{\|v-v_0\|\ge\delta}\frac{1}{2q}\|Em(V_1, v^\intercal F_1)\|^2 - \inf_{\|v-v_0\|\ge\delta}\frac{1}{q}\left\|\frac{1}{n}\sum_{i=1}^n m(V_i, v^\intercal F_i) - Em(V_1, v^\intercal F_1)\right\|^2 \\
=& \inf_{\|v-v_0\|\ge\delta}\frac{1}{2q}\|Em(V_1, v^\intercal F_1)\| + o_P(n^{-1/2}) \\
=& \inf_{\|(\mathbf{a}-\alpha, f-g)\|\ge\delta+\|\gamma_k(z)\|}\frac{1}{q}\|Em(V_1, \mathbf{a}^\intercal X_1, f(Z_1))\| + o_P(n^{-1/2}),
\end{aligned}$$

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

Taking $0 < \eta < \epsilon$,

$$\begin{aligned}
&P(Q_n(\widehat{v}) + \eta > \inf_{\|v-v_0\|\ge\delta}Q_n(v)) \\
=& P(Q_n(\widehat{v}_T) + \eta > \inf_{\|v-v_0\|\ge\delta}Q_n(v)) + o(1) \\
\le& P(Q_n(\widehat{v}_T) + \eta > \epsilon) + P(\inf_{\|v-v_0\|\ge\delta}Q_n(v) < \epsilon) + o(1) \\
\le& P(Q_n(\widehat{v}_T) > \epsilon - \eta) + o(1) = o(1)
\end{aligned}$$

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

References

[1] Andrews, D. (1994). Asymptotics for semiparametric econometric models via stochastic equicontinuity. *Econometrica*, 62:43–72.

[2] Andrews, D. and Lu, B. (2001). Consistent model and moment selection procedures for GMM estimation with application to dynamic panel data models. *Journal of Econometrics*, 101:123–165.

[3] Antoniadis, A. (1996). Smoothing noisy data with tapered coiflets series. *Scandinavian Journal of Statistics*, 23:313–330.

[4] Belloni, A., Chernozhukov, V., Chetverikov, D., and Kato, K. (2015). Some new asymptotic theory for least squares series: Pointwise and uniform results. *Journal of Econometrics*, 186:345–366.

[5] Belloni, A., Chernozhukov, V., and Wang, L. (2014). Pivitor estimation via square-root Lasso in non-parametric regression. *Annals of Statistics*, 42:757–788.

[6] Caner, M. (2009). Lasso-type GMM estimator. *Econometric Theory*, 25:270–290.

[7] Cattaneo, M. D., Jansson, M., and Newey, W. K. (2016). Alternative asymptotics and the partially linear model with many regressors. *Econometric Theory*, forthcoming:1–25.

[8] Cattaneo, M. D., Jansson, M., and Newey, W. K. (2017). Inference in linear regression models with many covariates and heteroscedasticity. *Working paper*.

[9] Chang, J., Chen, S., and Chen, X. (2015). High dimensional generalized empirical likelihood for moment restrictions with dependent data. *Journal of Econometrics*, 185:283–304.

[10] Chen, X. and Christensen, T. (2015). Optimal uniform convergence rates and asymptotic normality for series estimators under weak dependence and weak conditions. *Journal of Econometrics*, 188:447–465.

[11] Chen, X. and Liao, Z. (2015). Sieve semiparametric two-step GMM under weak dependence. *Journal of Econometrics*, 189:163–186.

[12] Chen, X., Linton, O., and Keilegom, I. V. (2003). Estimation for semiparametric models when the criterion function is not smooth. *Econometrica*, pages 1591–1608.

[13] Chen, X. and Pouzo, D. (2012). Estimation of nonparametric conditional moment models with possibly nonsmooth generalized residuals. *Econometrica*, 80:277–321.

[14] Dong, C. and Linton, O. (2016). Additive nonparametric models with time variable and both stationary and nonstationary regressors. https://papers.ssrn.com/sol3/papers.cfm? abstract id=2847681.

[15] Dudley, R. M. (2003). *Real Analysis and Probability*. Cambridge studies in advanced mathematics 74. Cambridge University Press, Cambridge, U.K.

[16] Engle, R., Granger, C. W. J., Rice, J., and Weiss, A. (1986). Semiparametric Estimates of the Relation Between Weather and Electricity Sales. *Journal of the American Statistical Association*, 81:310–320.

[17] Fan, J. and Li, R. (2001). Variable selection via nonconcave penalized likelihood and its Oracle properties. *Journal of the American Statistical Association*, 96:1348–1360.

[18] Fan, J. and Liao, Y. (2014). Endogeneity in high dimensions. *Annals of Statistics*, 42:872–917.

[19] Gao, J. and Liang, H. (1997). Statistical inference in single–index and partially nonlinear models. *Annals of the Institute of Statistical Mathematics*, 49(3):493–517.

[20] Gao, J. and Shi, P. (1997). M-type smoothing splines in non– and semi–parametric regression models. *Statistica Sinica*, 7(3):1155–1169.

[21] Gautschi, W. (2004). *Orthogonal Polynomials: Computation and Approximation*. Numerical Mathematics and Scientific Computation. Oxford University Press, Oxford.

[22] Han, C. and Phillips, P. C. B. (2006). GMM with many moment conditions. *Econometrica*, 18(74):147–192.

[23] Hansen, L. P. (1982). Large sample properties of generalized method of moments estimators. *Econometrica*, 50:1029–1054.

[24] Härdle, W., Liang, H., and Gao, J. (2000). *Partially Linear Models*. Springer–Verlag, New York.

[25] Mammen, E. (1989). Asymptotics with increasing dimension for robust regression with applications to the Bootstrap. *Annals of Statistics*, 17:382–400.

[26] Newey, W. K. (1997). Convergence rates and asymptotic normality for series estimators. *Journal of Econometrics*, 79:147–168.

[27] Newey, W. K. and Powell, J. L. (2003). Instrumental variable estimation of nonparametric models. *Econometrica*, 71:1565–1578.

[28] Newey, W. K. and Windmeijer, F. (2009). Generalized method of moments with many weak moment conditions. *Econometrica*, 77:687–719.

[29] Pakes, A. and Pollard, D. (1989). Simulation and the asymptotics of optimization estimators. *Econometrica*, 57:1027–1057.

[30] Portnoy, S. (1984). Asymptotic behaviour of M–estimators of p regression parameters when p^2/n is large. I: Consistency. *Annals of Statistics*, 12(4):1298–1309.

[31] Portnoy, S. (1985). Asymptotic behaviour of M–estimators of p regression parameters when p^2/n is large. II: Normal approximation. *Annals of Statistics*, 13(4):1403–1417.

[32] Powell, J. L. (1984). *Estimation of Semiparametric Models*. Handbook of Econometrics IV. Edited by R. F. Engle and D. L. McFadden. Elsevier, New York.

[33] Robinson, P. M. (1988). Root-N-consistent semiparametric regression. *Econometrica*, 56(2):931–954.

[34] Yatchew, A. (1997). An elementary estimator of the partial linear model. *Economic Letters*, 57(2):135–143.

[35] Yatchew, A. (2003). *Semiparametric Regression for the Applied Econometrician*. Cambridge University Press, New York.

[36] Zhang, C. H. (2010). Nearly unbiased variable selection under minmax concave penalty. *Annals of Statistics*, 38:894–942.