UNIVERSITY OF CAMBRIDGE
Faculty of Economics

Cambridge Working Papers in Economics

Cambridge Working Papers in Economics: 1881

HIGH DIMENSIONAL SEMIPARAMETRIC MOMENT RESTRICTION MODELS

Chaohua Dong Jiti Gao Oliver Linton

29 November 2018

We consider nonlinear moment restriction semiparametric models where both the dimension of the parameter vector and the number of restrictions are divergent with sample size and an unknown smooth function is involved. We propose an estimation method based on the sieve generalized method of moments (sieve GMM). We establish consistency and asymptotic normality for the estimated quantities when the number of parameters increases modestly with sample size. We also consider the case where the number of potential parameters/covariates is very large, i.e., increases rapidly with sample size, but the true model exhibits sparsity. We use a penalized sieve GMM approach to select the relevant variables, and establish the oracle property of our method in this case. We also provide new results for inference. We propose several new test statistics for the over-identification and establish their large sample properties. We provide a simulation study that shows the performance of our methodology. We also provide an application to modelling the effect of schooling on wages using data from the NLSY79 used by Carneiro et al. [17].

High Dimensional Semiparametric Moment Restriction Models

Chaohua Dong

Zhongnan University of Economics and Law, China

Jiti Gao

Monash University, Australia

Oliver Linton*

University of Cambridge, UK

November 29, 2018

Abstract

We consider nonlinear moment restriction semiparametric models where both the dimension of the parameter vector and the number of restrictions are divergent with sample size and an unknown smooth function is involved. We propose an estimation method based on the sieve generalized method of moments (sieve GMM). We establish consistency and asymptotic normality for the estimated quantities when the number of parameters increases modestly with sample size. We also consider the case where the number of potential parameters/covariates is very large, i.e., increases rapidly with sample size, but the true model exhibits sparsity. We use a penalized sieve GMM approach to select the relevant variables, and establish the oracle property of our method in this case. We also provide new results for inference. We propose several new test statistics for the over-identification and establish their large sample properties. We provide a simulation study that shows the performance of our methodology. We also provide an application to modelling the effect of schooling on wages using data from the NLSY79 used by Carneiro et al. [17].

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

JEL classification: C12, C14, C22, C30

*Corresponding author: Oliver Linton, Faculty of Economics, University of Cambridge, obl20@com.ac.uk

1 Introduction and examples

Large models are the focus of much current methodological research. As pronounced by Athey et al. [5]: "There is a large literature on semiparametric estimation of average treatment effects under unconfounded treatment assignment in settings with a fixed number of covariates. More recently attention has focused on settings with a large number of covariates". Belloni et al. [9] review a number of approaches to estimation and selection in large models defined through linear moment restrictions. We consider a class of nonlinear 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 with some weak instruments and endogeneity, Robinson [60], except in our case the number of covariates in the linear part may be large, i.e., increase to infinity with sample size. There are sometimes many binary covariates whose effect can be restricted to be linear, perhaps after a transformation of response, but other continuous covariates whose effect is thought to be nonlinear. In panel data, one may wish to allow for many fixed effects or interactive effects in an essentially parametric fashion, but capture the potential nonlinear effect of a critical covariate or a continuous treatment variable nonparametrically. If both the cross-section and time series dimension are large then these quantities are all estimable. See for example Connor et al. [27].

We 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 following Hansen [41]. There is a general theory available for non-smooth objective functions of finite dimensional parameters (e.g., Pakes and Pollard [55] and Newey and McFadden [50, Section 7]). Some recent work has focused on the extension to the case where there are many moment conditions but some conditions are more informative than others, the so-called weak instrument case, see Newey and Windmeijer [53] and Han and Phillips [40]. There is a large literature on semiparametric estimation problems with smooth objective functions of both finite and infinite dimensional parameters (e.g., Bickel et al. [14], Andrews [2], Newey [48], Newey and McFadden [50, Section 8], Pakes and Olley [54], Chen and Shen [25] and Ai and Chen [1]). Chen et al. [23] extended this theory to allow for non-smooth moment functions. Other work has sharpened and broadened the applicability of the semiparametric case where the number of Euclidean parameters is finite but there are unknown function-valued parameters and endogeneity (see, for example Chen and Liao [22]). Our work extends the semiparametric theory to the case where the parametric component is growing in complexity, which is of particular relevance for modern big data settings.

We suppose that

$$\mathbb{E}[m(V, \alpha^{\intercal} X, g(Z))] = 0, \tag{1.1}$$

where m is a known vector of functions whose dimension q is large. Here, α is an unknown Euclidean-valued parameter whose dimension p is large, while g is a vector of unknown smooth functions. The random variable V typically represents a dependent variable and possible instrumental variables, while the vectors X and Z are explanatory variables. We suppose that Z is of finite dimension, but the dimension of X (and V) may be large, i.e., diverge. We suppose that a random sample $\{V_i, X_i, Z_i,\ i = 1, \ldots, n\}$ is observed and that $p = p(n) \to \infty$ and $q = q(n) \to \infty$ as $n \to \infty$ with $q > p$. For our main inference results we consider the case where (at least) $p/n \to 0$, similar to Portnoy [57], Portnoy [58] and Mammen [47]. The moment restriction model (1.1) features high dimensionality in two ways: a high dimensional Euclidean parameter (α) (that shows up in a single-index form), and an infinite dimensional unknown function $g(\cdot)$. The number of moment conditions necessarily increases to infinity. Together this represents a new framework in the literature.

The parameters of interest are particular functionals of α and g. In many applications both types of quantities are of interest. For example, the weighted average MTE parameter in Carneiro et al. [17] depends on both α and g. In financial econometrics a leading example is the conditional value at risk, which depends on the parameters of the dynamic mean and variance model and on the quantile of the error distribution. We simultaneously estimate α and g in the parameter spaces defined below and then compute plug-in estimators from the estimates of α and g. Chen et al. [23] study a fixed-dimensional moment restriction model containing an unknown function. They consider both two step and profiled two-step methods. A similar approach is used in Chen and Liao [22]. Kernel estimation techniques in particular require an additional (albeit related) estimating equation for the function valued part, and either two-step or profile methods are common, see, for example, Powell [59]. We use the sieve methodology (see Chen [20] for a review) to estimate the model (1.1) in one step. Suppose that $g(\cdot)$ belongs to a suitable Hilbert space. We 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. Simultaneous estimation should lead to small sample efficiency gains; it also allows us to impose "cross-species" restrictions that link α and g. This approach also avoids high level assumptions, such as in Chen et al. [23] and Han and Phillips [40]. We establish the consistency and (self–normalized) asymptotic normality of the parameters of interest (which are general functionals of (α, g)) and provide a feasible central limit theorem (CLT) that allows normal based inference about the parameters of interest. We also propose some new test statistics to address the over-identification issue, and establish their large sample properties.

We then consider the ultra-high dimensional case where the number of potential X variables is extremely large, i.e., much larger than the sample size, but only a smaller subset of them are relevant, i.e., the parametric part of the model possesses sparsity. That is, we suppose that $p >> n$ but α contains many zero elements, although we do not know a priori the location of these zeros. This case has been considered by a number of recent studies in econometrics, Belloni et al. [12], and is the focus of much research in statistics. To address this issue we combine the GMM objective function with a specific penalty function, a folded concave penalty function (see Fan and Li [33]). We show that variable selection and estimation can be done simultaneously and our method achieves the oracle property, like Fan and Liao [34]. We also provide a result on post model selection inference, which allows us to use the distribution theory obtained in the first part of the paper. In a series of innovative papers by Belloni et al. [6], Belloni and Chernozhukov [7], Belloni et al. [11], Belloni et al. [9], and Belloni et al. [10], the authors develop the approximate linear model (ALM) framework. In that setting there is no formal distinction between parametric and nonparametric components and the methodology is built around the selection tools. Our more traditional semiparametric approach is explicit about the model components and their relative complexity. In particular, we specify that g is nonparametric and has to be estimated simultaneously with the parametric part.

We close with a discussion of applications. A common genesis for the unconditional moment restrictions (1.1) is conditional moment restrictions perhaps from some economic model (Hansen [41]). Let W_i be a sub-vector of $(X_i^\intercal, Z_i^\intercal)^\intercal$ and let $\rho(Y_i, \alpha^\intercal X_i, g(Z_i))$ be a known J-dimensional vector residual. Then, suppose that (α, g) is determined by the conditional moment restriction

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

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

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

Define $m(V_i, \alpha^\intercal X_i, g(Z_i)) = \rho(Y_i, \alpha^\intercal X_i, g(Z_i)) \otimes \Phi_K(W_i)$, where $V_i = (Y_i, W_i^\intercal)^\intercal$ and "$\otimes$" denotes the Kronecker product. Notice that the dimension of the function m is $q = JK$, which increases with K. Therefore, the pair (α, g) can be solved from the unconditional moment equation $\mathbb{E}[m(V_i, \alpha^\intercal X_i, g(Z_i))] = 0$. A specific example is a high dimensional partially linear model with endogenous covariates. Let $Y_i = \alpha^\intercal X_i + g(Z_i) + e_i$, $i = 1, \ldots, n$, where $\alpha \in \mathbb{R}^p$ and e_i is an error term such that $\mathbb{E}[e_i] = 0$ for all i. Here, X_i is endogenous in the sense that $\mathbb{E}[e_i|X_i] \neq 0$. In the case where the dimensionality of α is fixed, there are various results available in the literature (see, for example, Robinson [60]; Gao and Liang [36]; Gao and Shi [37]; Härdle et al. [42]).

To deal with the endogeneity, let W_i be a vector of instrumental variables and define a set of valid instruments $\lambda_i = \lambda(Z_i, W_i)$ with dimension q ($q > p$). Denote

$$m(V_i, \alpha^\intercal X_i, g(Z_i)) = (Y_i - \alpha^\intercal X_i - g(Z_i))\lambda(Z_i, W_i)$$

with $V_i = (Y_i, W_i^\intercal)^\intercal$. Then, we have the moment condition $\mathbb{E}[m(Y_i, W_i, \alpha^\intercal X_i, g(Z_i))] = 0$, which can be used to identify the parameter α and the nonparametric function $g(\cdot)$. Motivated by Robinson [60] and Belloni et al. [6] an alternative moment condition in this case is $m(V_i, \alpha^\intercal X_i, g(Z_i)) = \left(Y_i - g_Y(Z_i) - \alpha^\intercal \left(X_i - g_X(Z_i)\right), Y_i - g_Y(Z_i), (X_i - g_X(Z_i))^\intercal\right) \lambda(Z_i, W_i)$, where $g_Y(Z_i) = E(Y_i|Z_i)$ and $g_X(Z_i) = E(X_i|Z_i)$. Essentially this is the efficient score function for α in a special case, Bickel et al. [14]. One can jointly estimate α, g_Y, g_X from this moment condition and then obtain $g(Z) = g_Y(Z) - \alpha^\intercal g_X(Z)$. See Chernozhukov et al. [26] for a more general discussion of the advantages of certain moment functions over others in a general semiparametric moment condition setting. A slightly more complex model appears in Carneiro et al. [17] who consider the following in their equation (9):

$$\begin{aligned} &\mathbb{E}[Y - X^\intercal \delta - P(Z)X^\intercal \alpha - R(Z)|X, Z] = 0, \\ &\mathbb{E}[\mathbb{I}(S = 1) - P(Z)|Z] = 0, \end{aligned} \tag{1.2}$$

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

The rest of the paper is organized as follows. Section 2 gives the estimation procedure. Section 3 establishes the large sample theory for the estimator. In Section 4 we provide two methods for testing over-identification. In Section 5 we propose and analyze procedures for selecting covariates/parameters under sparsity. In Section 6 we evaluate the performance of our procedures using simulations. In Section 7 we apply our method to investigate the effect of schooling on earnings using the model and data of Carneiro et al. [17]. The last section concludes.

Throughout, $\|\cdot\|$ can be either Euclidean norm for vector or Frobenius norm for matrix, or the norm of functions in function space depending on 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

We can allow multiple indexes in m but for simplicity of notation we suppose that α is a vector rather than a matrix. The unknown function $g(\cdot)$ 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 the sieve method. For the sake of easy exposition, however, we suppose in this paper that g is a single multivariate function defined on $\mathbb{Z} \subset \mathbb{R}^d$. Let $g \in L^2(\mathbb{Z}, \pi) = \{f : \int_{\mathbb{Z}} f^2(z)\pi(z)dz < \infty\}$ a Hilbert function space, where $\pi(\cdot)$ is a user-chosen density function on $\mathbb{Z}$. The choice of the density π relates to how large the Hilbert space is chosen, since the thinner the tail of the density is, the larger the space is. For example, $L^2(\mathbb{R}, 1/(1+z^2)) \subset L^2(\mathbb{R}, \exp(-z^2))$. An inner product in the Hilbert space is given by $\langle f_1, f_2 \rangle = \int_{\mathbb{Z}} f_1(z) f_2(z) \pi(z) dz$, and hence the induced norm $\|f\| = \sqrt{\langle f, f \rangle}$ for any $f_1(z), f_2(z), f(z) \in L^2(\mathbb{Z}, \pi)$. Two functions $f_1, f_2 \in L^2(\mathbb{Z}, \pi)$ are called orthogonal if $\langle f_1, f_2 \rangle = 0$, and further are orthonormal if $\|f_1\| = 1$ and $\|f_2\| = 1$.

The parameter space for model (1.1) is defined as, $\Theta = \{(\mathbf{a}, f) : \mathbf{a} \in \mathbb{R}^p, f \in L^2(\mathbb{Z}, \pi)\}$, which contains the true parameter (α, g) as an interior point by the measure defined below in (2.2).

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

Recall that any Hilbert space has a complete orthogonal sequence (see Theorem 5.4.7 in Dudley [31, p. 169]). In our setting, although g 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 well known univariate orthonormal sequences.

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 [38], and Chen [20] for a more recent exposition.

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

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

The convergence of (2.1) normally can be understood in the sense of the norm in the space, whereas in the situation where g is smooth, the convergence in 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 the unknown parameters (α, g) and functionals thereof. The consistency studied below is defined in terms of a norm given by

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

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 no ambiguity is incurred.

In order to facilitate the implementation of nonlinear optimization, α should be confined to a compact subset of $\mathbb{R}^p$ and the truncated series $g_K(z) = \beta^\intercal \Phi_K(z)$ of the function g should be included in an expanding finite dimensional bounded subset of $L^2(\mathbb{Z}, \pi)$. It is noteworthy that in an infinite dimensional space, a bounded subset may not necessarily be compact. A detailed discussion for the compactness in infinite dimensional space can be found in Chen and Pouzo [24]. 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 6 using an example).

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

It is a common convention that the true parameter is assumed to be contained within a bounded set (Newey and Powell [51, p. 1569]); in this paper we allow the bounds for α to diverge with the sample size since the dimensionality of α grows to infinity.[1] Furthermore, since $\|g_K\| = \|\beta\| \leq \|g\|$ 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), any $f(z) \in L^2(\mathbb{Z}, \pi)$ 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 for each n. The above setting is similar to but broader than that in Newey and Powell [51].

We estimate α and β by

$$\begin{aligned} (\widehat{\alpha}, \widehat{\beta}) = \underset{\mathbf{a} \in \mathbb{R}^p, \mathbf{b} \in \mathbb{R}^K}{\operatorname{argmin}} \|M_n(\mathbf{a}, \mathbf{b})\|^2, \quad \text{subject to } \|\mathbf{a}\| \leq B_{1n} \text{ and } \|\mathbf{b}\| \leq B_{2n}, \\ \text{where } M_n(\mathbf{a}, \mathbf{b}) = \frac{1}{\sqrt{q}} \frac{1}{n} \sum_{i=1}^{n} m(V_i, \mathbf{a}^\intercal X_i, \mathbf{b}^\intercal \Phi_K(Z_i)). \end{aligned} \tag{2.3}$$

[1] 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.

Here, the involvement of q in $M_n(\mathbf{a}, \mathbf{b})$ takes into account the divergent dimensions of the vector m in order to avoid the issue that $\|M_n(\mathbf{a}, \mathbf{b})\|$ could be large even if each element is small that would arise if we had not put q into $M_n(\mathbf{a}, \mathbf{b})$. This issue does not arise when the vector–valued m function has fixed dimension. If there are additional restrictions on the model, say $f(\alpha, g(.)) = 0$, these can be incorporated into the optimization problem.

Define for any $z \in \mathbb{Z}$

$$\widehat{g}(z) = \widehat{\beta}^{\intercal} \Phi_K(z), \tag{2.4}$$

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

3 Asymptotic theory

3.1 Consistency

Before establishing our asymptotic theory, we state some assumptions that we rely on in the sequel.

Assumption 3.1 *Suppose that*

(a) *For each n, $\{(V_i, X_i^{\intercal}, Z_i^{\intercal}), i = 1, \ldots, n\}$ is an independent and identically distributed (i.i.d.) sequence (although the distribution depends on n, which we suppress notationally in the sequel)*;

(b) *For the density f_Z of Z, there exist two constants, $0 < c < C < \infty$, such that $c\pi(z) \leq f_Z(z) \leq C\pi(z)$ on the support $\mathbb{Z}$ of Z, where $\pi(z)$ is given in the preceding section*;

(c) *Each moment function $m_j(\cdot, \cdot, \cdot)$, $j = 1, \ldots, q$, is continuous in the second and third arguments*;

(d) $q(n) - p(n) \geq K$.

The i.i.d. property in Assumption 3.1(a) simplifies the presentation and some of the calculations, although it is possible to relax it to a weakly dependent data setting. Regarding Assumption 3.1(b), the relation between the densities of the variable Z and the function space is widely used in the literature. See, e.g. Condition A.2 and Proposition 2.1 of Belloni et al. [8, p.347]. This condition is used to bound the eigenvalues of the Gram matrix for the sieve method. When the support is compact, researchers simply impose that the density $f_Z(z)$ bounded away from zero and above from infinity that is a special case where $\pi(z) \equiv 1$ in our setting. Our theory allows for unbounded support for Z provided the density π is chosen appropriately.

Regarding Assumption 3.1(c), the continuity of the m function is weak, and commonly used moment functions satisfy this. In Assumption 3.1(d) we allow for possible overidentification of the parameter vector in the moment conditions, and we shall discuss this issue further in the next section.

Assumption 3.2 *Suppose that there is a unique function $g(\cdot) \in L^2(\mathbb{Z}, \pi)$ and for each n there is a unique vector $\alpha \in \mathbb{R}^p$ such that model (1.1) is satisfied. In other words, for any $\delta > 0$, there is a sufficiently small constant $\epsilon > 0$ such that*

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

This type of condition is quite standard in the parametric and semiparametric literature, see Pakes and Pollard [55] and Chen et al. [23]. 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.

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

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

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

This is a kind of Lipschitz condition. We note that this condition can be substituted by some high level condition such as stochastic equicontinuity, in order to derive the large sample behavior of the estimator. See, for instance, Pakes and Pollard [55] and Chen et al. [23]. As argued in Chen et al. [23, p.1597], when the moment function is Lipschitz continuous the *covering number with bracketing* is bounded above by the *covering number* for the parametric space, and hence a stochastic equicontinuity condition holds. Among others, Chen and Shen [25] used this approach. We would like to keep the low level condition because additionally it facilitates calculation in some situations.

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

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

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

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

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

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

The proof is given in Appendix B.

3.2 Limit distributions of the estimators

Since the dimension of α diverges, we cannot establish a limit distribution for $\widehat{\alpha} - \alpha$ itself. Instead, we shall consider some finite dimensional transformations of α, for which plug-in estimators are used. Likewise, we consider functionals of $g(\cdot)$. Let $\mathscr{L}$ be a linear transformation from $\mathbb{R}^p \mapsto \mathbb{R}^r$ with $r \geq 1$ fixed, and let $\mathscr{F} = (\mathscr{F}_1, \ldots, \mathscr{F}_s)^\intercal$ with fixed s be a vector of functionals on $L^2(\mathbb{Z}, \pi)$. 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. [8, p. 352] and several results such as Theorems 2 and 6 in Chang et al. [19]. The elements of $\mathscr{F}$ can be, for example, as described in Newey [49, p.151], the integral of $\ln[g(z)]$ on some interval, which stands for consumer's surplus in microeconomics. Other examples include: the partial derivative function, the average partial derivative, and the conditional partial derivative. Thus, we shall consider the limit distributions of $\mathscr{L}(\widehat{\alpha}) - \mathscr{L}(\alpha)$ and $\mathscr{F}(\widehat{g}) - \mathscr{F}(g)$. Towards this end, we need the following assumptions.

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

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

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

The Lipschitz condition for the components of the m function enables us to approximate the Hessian matrix within a neighbourhood of the true parameter, which in turn facilitates the

derivation of the limit theory. We will suppose that the function g is smooth; this is implicitly spelt out in Assumption 3.6 below.

Assumption 3.5 *Suppose that*

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

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

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

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

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

Assumption 3.6 *Suppose that*

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

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

Assumption 3.6 stipulates the relation between the truncation parameter K, the diverging dimension p of the regressor, and the sample size. Normally, $\|\gamma_K\|^2 = O(K^{-a})$, where $a > 0$ is related to the smoothness order of the function g. See, for example, Newey [49]. Thus, the assumption implicitly puts some conditions on the smoothness. Notice that the combination of 3.6(a) and (b) implies that $\|\gamma_K\|^2 pK = o(1)$ and $n^{-1}pK = o(1)$, which are used in the proof of the lemmas in the supplemental material.

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

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

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

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

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

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

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

$$\begin{aligned}\Sigma_n^2 :=&\Gamma_n[\Psi_n\Psi_n^\intercal]^{-1}\Psi_n\Xi_n\Psi_n^\intercal[\Psi_n\Psi_n^\intercal]^{-1}\Gamma_n^\intercal, \quad \text{in which} \\ \Gamma_n :=& \begin{pmatrix}\mathscr{L} & 0\\ 0 & \mathscr{F}'(g)\circ {\Phi_K}^\intercal\end{pmatrix}_{(r+s)\times(p+K)}, \\ \Xi_n :=& \mathbb{E}[m(V_1,\alpha^\intercal X_1,g(Z_1))m(V_1,\alpha^\intercal X_1,g(Z_1))^\intercal]_{q\times q}, \\ \Psi_n :=& \mathbb{E}\begin{pmatrix}\frac{\partial}{\partial u}m(V_1,\alpha^\intercal X_1,g(Z_1))^\intercal\otimes X_1\\ \frac{\partial}{\partial w}m(V_1,\alpha^\intercal X_1,g(Z_1))^\intercal\otimes\Phi_K(Z_1)\end{pmatrix}_{(p+K)\times q},\end{aligned} \tag{3.1}$$

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

Theorem 3.2 (Normality)**.** *Let Assumptions 2.1-2.2, 3.1-3.7 hold. Suppose that $\Psi_n\Psi_n^\intercal$ is invertible for large enough n. Suppose also that $B_{1n}^2+B_{2n}^2=o(n)$. Then as $n\to\infty$*

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

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

The proof of the theorem is given in Appendix B. Note that the conditions in the theorem imply the consistency of the estimator in Theorem 3.1. 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, for example Chang et al. [19] and Belloni et al. [8]. Apart from the diverging dimensions of Ψ_n and Ξ_n and the use of the transformation $\mathscr{L}$ and the functional $\mathscr{F}$, the form of the covariance matrices Σ_n^2 is the same as in the standard semiparametric literature such as Hansen [41], Pakes and Pollard [55] and Chen et al. [23].

In general the convergence order of $\mathscr{F}(\widehat{g}) - \mathscr{F}(g)$ is proportional to $(\mathscr{F}'(g) \circ \Phi_K(z)^\intercal \mathscr{F}'^\intercal \circ \Phi_K(z))^{1/2} n^{-1/2}$, which is similar to the result in Theorem 2 of Newey [49]. Here, the matrix in the front of $n^{-1/2}$ is of dimension $s \times s$ and is associated with the derivative of the 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 nonparametric rates of convergence in the literature. See, for example, Dong and Linton [30].

In general the convergence order of $\mathscr{L}(\widehat{\alpha} - \alpha)$ is $n^{-1/2}$; however, Theorem 3.2 does not rule out the mildly *weak instrument case* where the matrix Σ_n is close to singular, i.e., $\det(\Sigma_n) \neq 0$ but $\det(\Sigma_n) \to 0$ with n at a certain rate; this would reduce the convergence rate of the estimators but the self-normalized distribution theory we have presented continues to hold under our conditions. However, we do rule out the more extreme cases considered in Han and Phillips [40], which would change the limiting distribution.

The requirement that $\sqrt{n}\Sigma_n^{-1}(\mathbf{0}_r^\intercal, (\mathscr{F}'(g) \circ \gamma_K)^\intercal)^\intercal = o(1)$ is an "undersmoothing" condition, playing a similar role to, for example, the condition $\sqrt{n} V_K^{-1} K^{-p/d} = o(1)$ in Corollary 3.1 of Chen and Christensen [21, p. 454] and Comment 4.3 of Belloni et al. [8]. The precise form of the condition may vary according to the parameters of interest and the underlying model; it reflects the bias variance trade-off that is relevant for estimation of those quantities in the particular model.[2] In the large dimensional α case, the bias variance trade-off can be different from usual since the parametric part can contribute a large variance; the presence of weak instruments may also affect the bias variance trade-off for certain parameters. For inference results about $g(z)$ it is quite common practice to undersmooth/overfit to avoid the bias term. Some recent research advocates using extreme undersmoothing for better inference about finite dimensional parameters in semiparametric models. See for example Cattaneo et al. [18]. Cattaneo et al. [18] 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

[2]Linton [46], Donald and Newey [28], and Ichimura and Linton [43] considered the issue of tuning parameter choice in semiparametric models. The optimal tuning parameter depends on the model and the parameter of interest as well as on the estimating equations. In some cases the optimal rates for parametric components are the same as the optimal rates for the infinite dimensional components, specifically in adaptive cases, but even then the constants in the smoothing parameters will differ. In other cases, the optimal rates for the bandwidth in estimating the parametric part are different. This is often called "undersmoothing", although it seems a little prejudicial, perhaps "different smoothing" would be more accurate.

parameters in the case where essentially p is fixed but $K(n) \propto n$. In this case, the function $g(\cdot)$ is not consistently estimated. In our methodology we pay equal attention to the function g, which itself can be of interest. See for example, Engle et al. [32]; Robinson [60]; Gao and Liang [36]; Gao and Shi [37] and Härdle et al. [42]. Our methodology is also robust to conditional heteroskedasticity.

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

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

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

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

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

Then, the feasible version of the CLT (3.2), with $\widehat{\Sigma}_n$ replacing Σ_n, follows by similar arguments to those in the proof of Theorem 3.2. This allows the construction of simultaneous confidence intervals and consistent hypothesis tests about $\mathscr{L}(\alpha), \mathscr{F}(g)$. In practice it may be necessary to regularize the large dimensional covariance matrix $\widehat{\Psi}_n\widehat{\Psi}_n^{\intercal}$, which can be done by a variety of methods, see, for example, Bickel and Levina [15].

We may improve efficiency by using a weight matrix. Let W_n be a $q \times q$ positive definite matrix that may depend on the sample data. Then, $\|M_n(\mathbf{a}, \mathbf{b})\|^2$, which measures the metric of $M_n(\mathbf{a}, \mathbf{b})$ from zero, can be substituted by $M_n(\mathbf{a}, \mathbf{b})^{\intercal}W_nM_n(\mathbf{a}, \mathbf{b})$ in the minimization of (2.3), which is also a measure of the metric for the vector $M_n(\mathbf{a}, \mathbf{b})$ from zero but in terms of the weight matrix W_n. Meanwhile, $\|M_n(\mathbf{a}, \mathbf{b})\|^2$ can be viewed as a special case that W_n is the identity matrix. We require the matrix W_n to be not too close to singular to prevent the possibility that $M_n(\mathbf{a}, \mathbf{b})^{\intercal}W_nM_n(\mathbf{a}, \mathbf{b})$ may be close to zero when $(\mathbf{a}, \mathbf{b})$ is far from (α, β).

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

$$\Gamma_n[\Psi_nW^*\Psi_n^{\intercal}]^{-1}\Psi_nW^*\Xi_nW^*\Psi_n^{\intercal}[\Psi_nW^*\Psi_n^{\intercal}]^{-1}\Gamma_n^{\intercal}.$$

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

There is an alternative way that achieves efficiency in one-step estimation, viz., the continuous updating estimator (CUE)[3] Define $W_n(\mathbf{a}, \mathbf{b}) = [\Xi_n(\mathbf{a}, \mathbf{b})]^{-1}$, where

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

Then, $(\widetilde{\alpha}, \widetilde{g}(\cdot))$ can be estimated by minimizing $M_n(\mathbf{a}, \mathbf{b})^\intercal W_n(\mathbf{a}, \mathbf{b}) M_n(\mathbf{a}, \mathbf{b})$ over $(\mathbf{a}, \mathbf{b})$.

3.3 Semiparametric single-index structure

The multivariate function $g(Z)$ could make the model (1.1) suffer from the so-called "curse of dimensionality" when the dimension of Z is moderately large, Stone [61] and Chernozhukov et al. [26]. This feature would limit the use of the model in practice. One way to tackle the curse of dimensionality is to adopt a semiparametric single-index structure so that, as argued in Dong et al. [29], the model still enjoys some nonparametric flexibility but circumvents the curse of dimensionality. Let us consider

$$\mathbb{E}[m(V_i, \alpha^\intercal X_i, g(\theta_0^\intercal Z_i))] = 0, \tag{3.5}$$

where the notation involved is the same as in model (1.1) except that the unknown function $g(\cdot)$ is defined on $\mathbb{R}$, and the single-index vector has true parameter $\theta_0 \in \mathbb{R}^d$ and $\|\theta_0\| = 1$ with the first element being positive for the sake of identification.

The model of Carneiro et al. [17] is of this form. In their case, the marginal treatment effect (MTE) is $MTE(x, p) = x^\intercal \alpha + g'(p)$ and the parameter of interest is the weighted average

[3]The empirical likelihood method considered in Imbens [44], Newey and Smith [52] and Chang et al. [19] can also be developed here.

MTE, $\Delta = \int_0^1 MTE(x,p)h(x,p)dp$ for some known weighting function h.[4] The parameter θ_0 can be estimated from the moment equation derived from the second conditional moment in (1.2), $\mathbb{E}\left[(\mathbb{I}(S=1) - \Lambda(\theta_0^\intercal Z))\Psi_q(Z)\right] = 0$, with or without the specification of the function Λ, using the conventional technique for dealing with single-index models, such as Ai and Chen [1] and Dong et al. [29].

Although θ_0 can be estimated by the second equation of (1.2), in order to derive asymptotic distributions for the estimators of α and g defined later, it is convenient if $\widehat{\theta}$, the estimate of θ_0, is independent of the data used to estimate α and g by the first equation. This is possible and one way to do is as follows. Let us split the observations $\{V_i, X_i, Z_i,\, i = 1, \ldots, n\}$ into two subsamples randomly, $\text{Sub}_1 := \{(V_i, X_i, Z_i), i = 1, \ldots, n'\}$ and $\text{Sub}_2 := \{V_i, X_i, Z_i,$ $i = n'+1, \ldots, n\}$, with $n' = [n/2]$. The ordering in both subsamples in general is not the same as in the original sample but we keep using subscript i after partition. The first subsample Sub_1 can be used to estimate θ_0 by an additional moment restriction (say), resulting in $\widehat{\theta}$, and the second Sub_2 is used to estimate the parameter α and function g. Here, due to the i.i.d. property of the sample, the independence property holds naturally. Additionally, $\sqrt{n}(\widehat{\theta} - \theta_0) = O_P(1)$ (e.g. Yu and Ruppert [63]). The data-splitting technique is used in the literature, such as Bickel [13] and Belloni et al. [6]. The independence property is important for our theoretical development and thus we recommend the use of the data-splitting method in the rest of this section. Due to this reason, we make the following assumption.

Assumption 3.8 *For θ_0 in* (3.5)*, there exists an estimator $\widehat{\theta}$ such that $\sqrt{n}(\widehat{\theta} - \theta_0) = O_P(1)$ as $n \to \infty$ and assume that $\widehat{\theta}$ is independent of observations used in minimization* (3.6) *below.*

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

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

$$
\begin{aligned}
(\widehat{\alpha}, \widehat{\beta}) = \underset{\mathbf{a}\in\mathbb{R}^p, \mathbf{b}\in\mathbb{R}^K}{\operatorname{argmin}} \|\widetilde{M}_n(\mathbf{a}, \mathbf{b})\|^2, \quad \text{subject to } \|\mathbf{a}\| \le B_{1n} \text{ and } \|\mathbf{b}\| \le B_{2n}, \\
\text{where } \widetilde{M}_n(\mathbf{a}, \mathbf{b}) = \frac{1}{\sqrt{q}} \frac{1}{n - n'} \sum_{i=n'+1}^{n} m(V_i, \mathbf{a}^\intercal X_i, \mathbf{b}^\intercal \Phi_K(\widehat{\theta}^{\,\intercal} Z_i)),
\end{aligned}
\tag{3.6}
$$

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

[4] We may be interested in estimating MTE or in estimating the parameters subject to the cross parameter restriction that MTE= 0.

Theorem 3.3. *(1) Under Assumptions 2.1*, 2.2, 3.1*, 3.2*, 3.3* and 3.8, the consistency in Theorems 3.1 are satisfied by the $\widehat{\alpha}$ and $\widehat{g}(z)$ defined in this subsection.*

(2) Let Assumptions 2.1, 2.2, 3.1*-3.3*, 3.4, 3.5*, 3.6, 3.7* and 3.8 hold. Then, the normality in Theorem 3.2 is valid for the $\widehat{\alpha}$ and $\widehat{g}(z)$ defined in this subsection with replacement of Ξ_n and Ψ_n respectively by*

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

Using Lemmas A.4-A.6 in Appendix A, the theorem is proven in the supplemental material of the paper. The estimation of the covariance matrix can be obtained similarly to that in Theorem 3.2 and we omit this for brevity.

The above procedure can be repeated as many times as we wish (with different subsamples) and the subsamples can be exchanged for the estimations of θ_0 and (α, g). Then, we can average these estimates that would improve the accuracy.

4 Statistical inference

4.1 Test of over-identification

Hansen [41] proposes the J–test for over-identification in the situation where both p and q are fixed but $q > p$. This J-test has an asymptotic χ^2_{q-p} null distribution. In the case where an unknown infinite dimensional parameter is involved, and both p and q are still fixed with $q > p$, Chen and Liao [22] establish a statistic for over-identification testing that has an F distribution in large samples. We propose an alternative test below, which as far as we are aware, is new.

We consider the following hypotheses:

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

where Θ is defined in Section 2.

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

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

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 hypothesis, by the procedure in Section 2 and the conditions of Theorem 3.1, the estimator $(\widehat{\alpha}, \widehat{g})$ is consistent. The statistic $L_n(\widehat{\alpha}, \widehat{\beta}; \kappa)$ can be used to detect H_0 against

H_1, as shown in Theorems 4.1 and 4.2 below. This test also works for the conventional moment restriction models with fixed p and q. Before establishing the asymptotic distribution under the null and the consistency under the alternative, we introduce some assumptions.

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

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

These are technical requirements. Noting $\mathbb{E}[m(V, \alpha^\intercal X, g(Z))] = 0$, Assumption 4.1 requires that $\mathbb{E}[m(V, \mathbf{a}^\intercal X, f(Z))]$ drops to zero very quickly when $(\mathbf{a}, f)$ approaches (α, g). This is the same, in spirit, as Assumption 3.2, but here it is a sample version and the decay of the expectation needs a certain rate. A similar assumption is also imposed by equation (4.9) of Andrews [2, p.58] and equation (5.2) of Belloni et al. [11, 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 cases where z is located in some compact set or $g(z)$ is integrable on the real line, given that the g function is sufficiently smooth.

Theorem 4.1. *Let Assumptions 4.1-4.2 hold, and the conditions of Theorems 3.1 and 3.2 remain true. For any* $\kappa \in \mathbb{R}^q$ *such that* $\|\kappa\| = 1$, *under* H_0,

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

as $n \to \infty$, *where* $(\widehat{\alpha}, \widehat{\beta})$ *is the estimator given by* (2.3).

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

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

The condition on the eigenvalues is commonly adopted in the literature, see, e.g. Chang et al. [19] and Belloni et al. [8]. 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. [19, p.290]. However, we allow for $\delta_n \to 0$ with a rate slower

than $n^{-1/2}$. This means that the strongest signal ($\delta_n = \delta$) can be weakened ($\delta_n \to 0$) when our test statistic is used.

4.2 Student t test

We next propose an alternative test for model (1.1) under H_0. Define $\widehat{e} = (\widehat{e}_1, \ldots, \widehat{e}_q)^\intercal$ and $\widehat{\sigma}^2 = (\widehat{\sigma}^2(i,j))_{q\times q}$, where

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

in which for simplicity $\widehat{m}(i) := m(V_i, \widehat{\alpha}^\intercal X_i, \widehat{g}(Z_i))$ and correspondingly, for later use define $m(i) := m(V_i, \alpha^\intercal X_i, g(Z_i))$. Here, $\widehat{e}$ and $\widehat{\sigma}^2$ may be understood as the estimated mean and covariance matrix of the error vector, respectively. Define

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

The statistic is constructed from $\sqrt{n}\widehat{e}_j/\widehat{\sigma}_n(j,j)$, which is somewhat like the traditional t-test. Pesaran and Yamagata [56] proposed a similar statistic in the context of testing linear asset pricing models.

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

The proof is given in Appendix B. The requirement on $\mathbb{E}[m(i)m(i)^\intercal]$ to be a diagonal matrix implies the orthogonality between the errors. This is not stringent because, if not so, we may make a transformation $\tilde{m}(i) = (\mathbb{E}[m(i)m(i)^\intercal])^{-1/2}m(i)$ and then $\tilde{m}(i)$ would meet the requirement. Moreover, in many situations it is satisfied naturally. For instance, in Example 1.1 of Section 1, $m(i)$ is consisting of orthogonal functions of the conditional variable. This requirement is also used in some other papers, such as Gao and Anh [35]. These moment requirements are commonly used in the literature since $m_j(i)$ are generalized error terms, so we do not explain them in detail. In addition, the behaviour of T_n is like $\chi^2(q)$ but with diverging q. Therefore, after normalization we have asymptotic normal distribution for T_n.

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

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

Define also

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

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

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

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

5 Penalised GMM under sparsity

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

There are some existing papers on the variable selection under sparsity. Belloni et al. [11] propose the combination of least squares and L_1 type lasso approach to select coefficients of the sieve in nonparametric regression. Also, Su et al. [62] use L_1 type lasso approach to study continuous treatment in nonseparable models with high dimensional data. In a high dimensional conditional moment restriction model, Fan and Liao [34] propose to use a folded concave penalty function combined with instrumental variables to select the important coefficients. Caner [16] uses the same approach with a particular class of penalty functions to select variables. As Caner [16, p.271] argued, the Lasso-type GMM estimator selects the correct model much more often than GMM-BIC and the "downward testing" method proposed by Andrews and Lu [3]. We shall tackle the selection issue by the combination of a penalty function and our GMM approach.

We partition the parameter vectors as $\alpha = (\alpha_{0S}^\intercal, \alpha_{0N}^\intercal)^\intercal$ 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 (i.e. nonzero coefficients), respectively, as referred in the literature such as Fan and Liao [34], while α_{0N} and β_{0N} are zero.

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. In addition, $v_{0S} = (\alpha_{0S}^\intercal, \beta_{0S}^\intercal)^\intercal$ is referred to as an oracle model. Define $t_n = |v_{0S}|$ the dimension of v_{0S}, which may diverge with n.

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

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

where $M_n(v) = M_n(\mathbf{a}, \mathbf{b})$ is as defined in Section 2, and $P_n(\cdot)$ is a penalty function discussed later. Our framework also accommodates the case where some components of α, β are entered without selection, as in Belloni et al. [10], although we do not inscribe this in the notation for simplicity.

5.1 Oracle Property

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

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

Suppose that $P_n(\cdot)$ belongs to the class of folded concave penalty functions (see Fan and Li [33]). For any generic vector $v = (v_1, \ldots, v_{t_n})^\intercal \in \mathbb{R}^{t_n}$ with $v_j \neq 0$, for all j, define

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

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

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

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

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

There are many classes of functions that satisfy these conditions. For example, with properly chosen tuning parameter, the L_r penalty ($0 < r \leq 1$), hard-thresholding (Antoniadis [4]), SCAD (Fan and Li [33]) and MCP (Zhang [64]) satisfy the requirements.

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

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

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

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

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

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

This assumption is often encountered in the literature, such as Assumption 4.3 in Fan and Liao [34]. It is known that there are many classes of distributions satisfying this condition, e.g., a continuous distribution with compact support, a normal distribution, and an exponential distribution and so on. The thin tail of the distribution postulated in the assumption enables us to bound the score function.

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

Assumption 5.4 (i) *There are constants $C_1, C_2 > 0$ such that $\lambda_{\min}(\mathbb{E}\partial m^\intercal(V_i, v_{0S}^\intercal F_{iS}) \otimes F_{iS})(\mathbb{E}\partial m^\intercal(V_i, v_{0S}^\intercal F_{iS}) \otimes F_{iS})^\intercal) > C_1$ and $\lambda_{\max}(\mathbb{E}\partial m^\intercal(V_i, v_{0S}^\intercal F_{iS}) \otimes F_{iS})(\mathbb{E}\partial m^\intercal(V_i, v_{0S}^\intercal F_{iS}) \otimes F_{iS})^\intercal) < C_2$; (ii) $P_n'(d_n) = o(n^{-1/2})$ and $\max_{\|v_S - v_{0S}\| < d_n/4} \phi(v_S) = o((t_n \log(q))^{-1/2})$; (iii) $t_n^{3/2} \log(q) = o(n)$, $t_n^{3/2} P_n'(d_n)^2 = o(1)$, $t_n \max_{j \in T} P_n(|v_{0j}|) = o(1)$.*

All these are technical requirements on the Hessian matrix, the penalty function, the relationship among the dimensions of the important coefficients, the sparsity and the sample size. These conditions are commonly used in the literature, for example, Assumptions 4.5-4.6 in Fan and Liao [34] among others. There are several penalty functions that satisfy these conditions, for example, SCAD and MCP with tuning parameter $\lambda_n = o(d_n)$. Thence, the conditions (ii) and (iii) are satisfied if $t_n\sqrt{\log(q)/n} + t_n^{3/2} \log(q)/n \ll \lambda_n \ll d_n$. However, noting that the exact

identification is allowed, the total number of parameters $p+K$ of α and β to be estimated can be as large as $\exp(n^a)$ for some $0<a<1$, an implication of the restriction on q.

To state the following theorem, define:

$$\begin{aligned}\Sigma_{nT}^2 :=& \Gamma_n[\Psi_{nT}\Psi_{nT}^\intercal]^{-1}\Psi_{nT}\Xi_{nT}\Psi_{nT}^\intercal[\Psi_{nT}\Psi_{nT}^\intercal]^{-1}\Gamma_n^\intercal, \quad \text{in which} \\ \Gamma_n :=& \begin{pmatrix} \mathscr{L} & 0 \\ 0 & \mathscr{F}'(g)\Phi_{KT}{}^\intercal \end{pmatrix}_{(r+s)\times(p_1+K_1)}, \\ \Xi_{nT} :=& \mathbb{E}[m(V_1,\alpha_{0S}^\intercal X_{1S}, g(Z_1))m(V_1,\alpha_{0S}^\intercal X_{1S}, g(Z_1))^\intercal]_{q\times q}, \\ \Psi_{nT} :=& \mathbb{E}\begin{pmatrix} \frac{\partial}{\partial u}m(V_1,\alpha_{0S}^\intercal X_{1S}, g(Z_1))^\intercal \otimes X_{1S} \\ \frac{\partial}{\partial w}m(V_1,\alpha_{0S}^\intercal X_{1S}, g(Z_1))^\intercal \otimes \Phi_{KT}(Z_1) \end{pmatrix}_{(p_1+K_1)\times q},\end{aligned} \tag{5.2}$$

provided that $\Psi_{nT}\Psi_{nT}^\intercal$ is invertible, in which u and w stand for the second and the third arguments of the vector function $m(v,u,w)$, respectively; and the transformation $\mathscr{L}_{r\times p_1}$ and s-vector functional $\mathscr{F}$ are defined similarly in Section 3.

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

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

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

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

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

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

The proof is given in Appendix B. We remark that the post selection version of the standard errors defined in (3.3) and (3.4) can be shown to be consistent in this case thereby allowing consistent confidence intervals for the selected parameters. Furthermore, post selection versions of Theorems 4.2 and 4.3 can be shown to hold.

The estimators in this theorem are all local. This is why we exclude the identification condition in Assumption 3.2 currently, while in the next theorem we shall discuss the global

property of a local minimizer. The results (i) and (ii) indicate that under these conditions in the theorem we are able to recover the sparsity in the model; meanwhile, the discussion on the result (iii) of the theorem is similar to Theorem 3.2.

5.2 Global Property

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

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

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

The theorem says that the local minimizer of the oracle space in Theorem 5.1 is also with high probability a global minimizer in $\mathbb{R}^{p+K}$. Note that by Theorems 5.1 and 5.2, the minimization in equation (5.1) enables one to recover the sparsity in the ultra high dimensional case since $q \geq p + K$, where q can be as large as e^{n^ϵ} for some $\epsilon > 0$. This is a bit different from Fan and Liao [34] where there is no nonparametric function involved and $q = p$ (the number of IV is the same as that of regressors). Note that, given the consistency of the sparsity, the inference can be done in a similar way to Theorem 3.2.

6 Simulation experiments

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

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

Suppose that $\mathbb{E}[Y_i - \alpha^\intercal X_i - g(Z_i)|W_i] = 0$ 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 $j \geq 1$, $\varphi_j(r) = \sqrt{2}\cos(\pi j r)$. Then, $\{\varphi_j(r)\}$ is an orthonormal basis in the Hilbert space $L^2[0,1]$. In the experiment, put $\alpha = (0.4, 0.1, 0, \ldots, 0)^\intercal \in \mathbb{R}^p$ 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 $\Phi_q(\cdot) = (\varphi_0(\cdot), \ldots, \varphi_{q-1}(\cdot))^\intercal$. Notice that the dimension of m function is q which increases

with the sample size n. Thus, (α, g) can be solved from unconditional moment equations $\mathbb{E}[m(V_i, \alpha^\intercal X_i, g(Z_i))] = 0$ for $i = 1, \ldots, n$.

According to the estimation procedure in Section 2, define $(\widehat{\alpha}, \widehat{\beta}) = \underset{\mathbf{a}\in\mathbb{R}^p, \mathbf{b}\in\mathbb{R}^K}{\operatorname{argmin}} \|M_n(\mathbf{a}, \mathbf{b})\|^2$, where $M_n(\mathbf{a}, \mathbf{b}) = \frac{1}{\sqrt{q}}\frac{1}{n}\sum_{i=1}^n m(V_i, \mathbf{a}^\intercal X_i, \mathbf{b}^\intercal \Phi_K(Z_i))$. 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 as OLS. This means that in any similar situation the optimization in Section 2 does not need the compactness 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 sequel) 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,

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

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

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

It can be seen that all of the statistical quantities about the estimate of g are reasonably attenuated with the increase of both the sample size and ν that provides more information for the parameters being estimated. For the quantities about the estimate of α, we observe that they normally do not decrease with the sample size. This is because, as mentioned before,

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

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

the dimension of α is increasing with the sample size; and hence it does not make sense to compare them among different sample sizes. However, we find that, given the sample size, both quantities related to the estimate of α decrease with the increase of ν that gives more moment restrictions.

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

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

Example 6.2. We consider the binary choice model where Y_i is 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

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

Let the distribution function $F(u) = \exp(u)/[1 + \exp(u)]$. 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 $\alpha = (0.5, 0.3, 0, \ldots, 0)^\intercal \in \mathbb{R}^p$ 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 $\Phi_K(z) = (p_0(z), \ldots, p_{K-1}(z))^\intercal$ and 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 $(\widehat{\alpha}, \widehat{\beta}) = \underset{\mathbf{a}\in\mathbb{R}^p, \mathbf{b}\in\mathbb{R}^K}{\operatorname{argmin}} \|M_n(\mathbf{a}, \mathbf{b})\|^2$ 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 report the bias $B_g(n)$, standard deviation $\pi_g(n)$ and RMSE $\Pi_g(n)$ for the estimate of g and $B_\alpha(n)$ and $M_\alpha(n)$ for the estimate of α defined in the above example.

Table 2: Simulation results for Example 6.2

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

In this experiment 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 estimate of the g function converges a bit slower than that in 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 estimates.

Example 6.3. This example is to verify the proposed schedule for variable selection and parameter estimation under sparsity studied in Section 5. The model is almost the same one

in Example 6.1 but the conditional variables are different. Suppose that

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

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

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

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

In the simulation, we use SCAD of Fan and Li [33] with predetermined tuning parameters of λ as the penalty function. Therefore, the objective function is $\|M_n(v)\|^2 + \sum_{j=1}^{p+K} P_n(|v_j|)$, where $v = (\alpha^\intercal, \beta^\intercal)^\intercal$ a $(p+K)$-dimensional vector and $M_n(v) = \frac{1}{q_1 n}\sum_{i=1}^n (Y_i - \alpha^\intercal X_i - \beta^\intercal \Phi_K(Z_i))\Psi_q(W_i)$.

Four performance measures are reported. The first measure is the mean standard error (MSE_S) of the important regressors, that is, the average of $\|\widehat{\alpha}_S - \alpha_S\|$ and that of $\|\widehat{\beta}_S - \beta_S\|$ over Monte Carlo replications. The second measure is the mean standard error (MSE_N) of the unimportant regressors for α and β, respectively. The third measure, denoted by TP_S, is the number of correctly selected nonzero coefficients, and the fourth, TP_N, the number of correctly selected unimportant coefficients for α and β, respectively. The initial value for v in the simulation is taken as $(0, \ldots, 0)$. The results are reported in Tables 3 and 4 with different parameters.

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

Table 3: Simulation results of Example 6.3($n = 100$)

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

7 Empirical illustration

There are many papers dealing with the marginal treatment effect (MTE) of a selection process. For example, Carneiro et al. [17, CHV, hereafter] study MTE for schooling, while most recently Su et al. [62] study continuous MTE in nonseparable models. Economists would like to know, on average, how the marginal return to schooling changes as the number of years of education increases, and would also like to be able to evaluate policies that change the probability of attaining a certain level of schooling. Let Y_1 be the potential log wage if the individual were to attend college and Y_0 be the potential log wage if the individual were not to attend college. Define potential outcome equations:

$$Y_1 = \mu_1(X) + U_1, \qquad \text{and} \qquad Y_0 = \mu_0(X) + U_0,$$

where X is a vector of relevant variables, $\mu_1(x) = \mathbb{E}(Y_1|X = x)$ and $\mu_0(x) = \mathbb{E}(Y_0|X = x)$.

Then, a selection process can be described as follows:

$$S = \begin{cases} 1, & I_S > 0, \\ 0, & \text{otherwise}, \end{cases} \qquad \text{where } I_S = \mu_S(Z) - V,$$

here I_S stands for the net benefit of attending college, $\mu_S(Z)$ is defined in CHV, in which Z is observable and V is unobservable, so that $S = 1$ means that the agent goes to college while $S = 0$ means that he/she does not. Let $Y = SY_1 + (1 - S)Y_0$ be the earnings of an individual.

Table 4: Simulation results of Example 6.3 ($n = 150$)

	$p = 15$, $K = 10$, $q = 150$				$p = 20$, $K = 10$, $q = 200$		
$\lambda =$	0.4	0.2	0.05	$\lambda =$	0.4	0.2	0.05
$\text{MSE}_S(\alpha)$	0.2068	0.2130	0.1848	$\text{MSE}_S(\alpha)$	0.2212	0.2228	0.1530
$\text{MSE}_S(\beta)$	0.1485	0.0868	0.0475	$\text{MSE}_S(\beta)$	0.1327	0.0937	0.0482
$\text{MSE}_N(\alpha)$	0.0000	0.0000	0.0014	$\text{MSE}_N(\alpha)$	0.0008	0.0001	0.0007
$\text{MSE}_N(\beta)$	0.0000	0.0000	0.0006	$\text{MSE}_N(\beta)$	0.0000	0.0000	0.0006
$\text{TP}_S(\alpha)$	4	4	4	$\text{TP}_S(\alpha)$	4	4	4
$\text{TP}_S(\beta)$	2	2	2	$\text{TP}_S(\beta)$	2	2	2
$\text{TP}_N(\alpha)$	10.36	10.2	9.40	$\text{TP}_N(\alpha)$	14.88	14.00	13.28
$\text{TP}_N(\beta)$	7.48	7.50	6.90	$\text{TP}_N(\beta)$	7.44	7.15	6.84

CHV analyse the marginal treatment effect for schooling, defined by the derivative of $\mathbb{E}(Y|X = x, P(Z) = p)$ with respect to p, denoted by $\text{MTE}(x, p)$. The dataset constructed by CHV is available at `www.aeaweb.org/articles?id=10.1257/aer.101.6.2754`. Specifically, the data comes from the 1979 National Longitudinal Survey of Youth (NLSY79), which surveys individuals born in 1957-1964 and includes basic demographic, economic and educational information for each individual. It also includes a well-known proxy for ability of earning that is thought of beyond schooling and work experience: the Armed Forces Qualification Test (AFQT), which gives a measure usually understood as a proxy for the "intrinsic ability" of the respondent. This data has been used repeatedly to either control for or estimate the effects of ability in empirical studies in economics and other disciplines. See CHV for further details and references.

We shall use exactly the variables X and Z in CHV but with our proposed methodology to estimate parameters and test hypotheses of interest.[5]

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

7.1 Estimation of MTE

We note that equation (9) of CHV implies that

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

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

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

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

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

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

With $\widehat{\gamma}$ at hand, we first calculate the average derivative of each variable in the choice model (7.1), that is, for each individual we compute the effect of increasing each variable by one unit (keeping all the others constant) on the probability of enrolling in college and then we average across all individuals. The results are reported in Table 5.

Table 5: Average marginal derivatives in decision model

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

The marginal derivatives reflect the changes in probability of attending a college when some policy was implemented to increase the relevant variable by one unit. For example, the

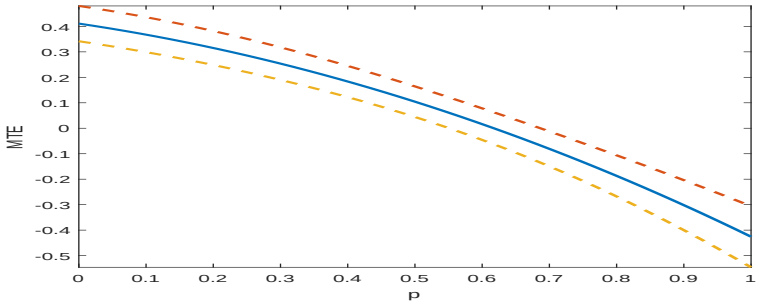


Figure 1: Estimated MTE calculated at $x = \bar{X}$ and the 95% Confidence Interval

marginal derivative of "Permanent local log earnings of 17", -0.0265, means that when the earnings increases 100 dollars, the probability on average of attending a college would decrease 2.65%. By contrast, this derivative in CHV is 0.1820, meaning that a 100 dollar increase in the labor market would result in an increase of 18.20% enrolling in a college. This seems contradictory with intuition.

Moreover, equation (7.2), along with $\widehat{\gamma}$, allows us to estimate θ_0 and $g(\cdot)$ by transforming it to unconditional moments. The estimation procedure and asymptotic theory for this semiparametric single-index structure has been established in Section 3.3. Since the function $g(\cdot)$ is defined on $[0,1]$, a power series $\{p^j, j \geq 1\}$ in $L^2[0,1]$ is employed to approximate the unknown $g(\cdot)$, and the same procedure as in Example 6.1 gives $\widehat{\theta}$ and $\widehat{g}(p)$. Hence, we have the estimate of MTE, $\widehat{\text{MTE}}(x,p) = x^\intercal \widehat{\theta} + \widehat{g}'(p)$, where $\widehat{\theta}$ is given in Table 6 and $\widehat{g}'(p) = 0.6462 - 0.3898p - 0.4470p^2$. The plot of $\widehat{\text{MTE}}(x,p)$ with $x = \bar{X}$, along with the upper and lower 95% significance bounds, is given in Figure 1. It can be seen that with the increase of the probability of attending college, the MTE decreases. The plot is quite similar to Figure 4 in CHV(p. 20).

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

Furthermore, with regard to testing whether $g(p)$ is a constant function, in CHV this test is implemented through specifying $g(p)$ as polynomials of order 2-5, respectively, and then test whether their coefficients are jointly zero. Nonetheless, we actually have done this in the estimate of $\widehat{g}(p)$ without any specification, because we treat $g(p)$ as a nonparametrically unknown function, and two coefficients in $\widehat{g}(p)$ are found to be significant. Thus, we think the test would not be necessary.

Table 6: Estimated coefficients of θ_0 and $\widehat{g}(p)$ in MTE

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

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

7.2 Nonlinearity of AFQT

We realize that an individual's ability of earning (AFQT) may affect the wage in a complicated way, instead of in linear or quadratic form in CHV. The pattern of this affect is possibly different in different groups of people. To evaluate this issue, we split the sample constructed by CHV into two subsamples: the first one for high school dropouts or graduated students (Subsample H, hereafter), while the second includes college dropouts, graduates and postgraduates (Subsample C, hereafter). The sample sizes are $n_1 = 882$ and $n_2 = 865$, respectively.

Let Y be the log wage of individual, U be the AFQT, X_{-1} be the vector consisting of all variables in X except U. Consider conditional moment model $\mathbb{E}[(Y - \alpha^\intercal X_{-1} - f(U))|W] = 0$, where W is the instrument. Then we have unconditional moment equations $\mathbb{E}[(Y - \alpha^\intercal X_{-1} - f(U))\Psi_q(W)] = 0$, where $\Psi_q(W)$ is a q-vector of basis functions on the instrument W, $q = \Pi_{j=1}^4 q_j$ and $q_j = 3$, meaning that the conditional moment function is developed using the same number of basis functions in all directions of coordinates. The model will be fitted by Subsamples H and C, respectively.

Since the score of AFQT has been standardized, we use the Hermite polynomial sequence for the development of the $f(U)$. We choose the truncation parameter based on the estimation for different truncation parameters, and the optimal parameter is the one that the estimated variance $\widehat{\sigma} = \widehat{\sigma}(K)$, using the procedure in Section 2, is the minimum among chosen K's. Denote by $\widehat{\sigma}_1(K)$ and $\widehat{\sigma}_2(K)$ the variances calculated using the two subsamples, respectively.

Table 7: Estimated standard deviation ($\times 10^4$)

	Truncation parameter					
	K =1	K = 2	K =3	K =4	K =5	K =6
$\widehat{\sigma}_1(K)$	10.1768	7.0663	9.3926	8.6882	8.2505	7.7928
$\widehat{\sigma}_2(K)$	4.1264	4.7496	4.9697	3.7082	3.8777	3.8558

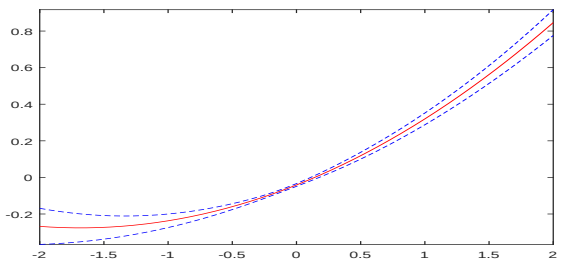
It can be seen from Table 7 that the optimal choices of the truncation parameters are $\widehat{K}_1 = 2$ and $\widehat{K}_2 = 4$ for f_1 in Subsample H and f_2 in Subsample C, respectively. Accordingly, the estimated functions are

$$\widehat{f}_1(u) = 0.2622h_1(u) + 0.0778h_2(u), \tag{7.4}$$

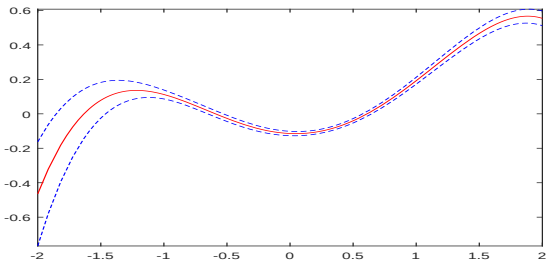
$$\widehat{f}_2(u) = 0.0713h_1(u) + 0.1086h_2(u) + 0.0826h_3(u) - 0.1233h_4(u), \tag{7.5}$$

where $h_j(u) = H_j(u)/\sqrt{\sqrt{\pi}2^j j!}$ and $H_j(u)$ are Hermite polynomials. Notice that there is no constant term in the estimated function as the constant is not identifiable from the intercept of the equations. Since we mainly focus on the estimate of nonparametric function, all estimated coefficients of X_{-1} by the two subsamples are given in the supplementary material of the paper.

Figure 2: The plots of $\widehat{f}_1(u)$ and $\widehat{f}_2(u)$



(a) The plot of $\widehat{f}_1(u)$



(b) The plot of $\widehat{f}_2(u)$

Figure 2 shows the plots of two functions $\widehat{f}_1(u)$ and $\widehat{f}_2(u)$ estimated by the subsamples H and C, respectively, and the 95% confidence upper and lower bounds on the main support of the dataset. As can be seen from Figure 2a, for the high school dropouts and graduates their corrected scores of AFQT contribute to their earnings by an increasing function. Thus, the higher the score is, the higher the earning is. On the other hand, from Figure 2b we see that the estimated function is mainly increasing as well, except a small sub-interval where it is a bit downward. This means that for college graduates and postgraduates the contribution of the

AFQT is somewhat complicated; specifically, with AFQT greater than the mean (i.e. zero), individuals' income is increasing as AFQT increases, whereas with AFQT less than the mean, individuals' income firstly increases and then decreases with their AFQTs. This phenomenon motivates that some interactive terms might be included. We however do not pursue this issue here since it is beyond the scope of our theoretical setting. Note that the negative values of the functions do not imply anything since the score has been corrected to have mean zero and unit variance, and we fail to identify their intercepts. Here, we only emphasize their forms.

8 Conclusion

We provided estimation and inference tools for a class of high dimensional semiparametric moment restriction models based on the sieve GMM method and the penalized sieve GMM method. Our approach is based on simultaneous selection of and estimation of the unknown quantities. The theoretical results are verified through finite sample experiments. We found that the more the number of moment restrictions, the more accurate the estimates. In addition, in our empirical study we also found our results to be more reasonable in some respects than the existing literature. The framework we have considered is quite general but can be generalized in a number of ways. First, we may allow explicitly for panel data and allow for weak dependent sampling schemes. Second, we may allow for a large number of nonparametric functions to enter the moment condition provided they are each defined on low dimensional spaces. Another question of interest here is efficiency; Jankova and Geer [45] develop some results about efficiency in the large linear model framework.

9 Acknowledgement

We thank Professor Xiaohong Chen for her insightful suggestions and for providing us with some relevant references. We also thank the audience of the seminar in Monash University and the Fifth China Meeting of Econometric Society 2018 in Shanghai. 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

This section gives all technical lemmas, additional assumptions and some notation used for the theoretical derivations, while the proofs of these lemmas are postponed to the supplementary material of the paper.

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

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

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

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

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

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

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

and

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

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

The unimportant constant shall be ignored in what follows.

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

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

and define

$$h_{11}(\alpha,g) := \frac{1}{q}\sum_{\ell=1}^{q}\left(\mathbb{E}\frac{\partial}{\partial u}m_\ell(V_1,\alpha^\intercal X_1,g(Z_1))X_1\right)\left(\mathbb{E}\frac{\partial}{\partial u}m_\ell(V_1,\alpha^\intercal X_1,g(Z_1))X_1\right)^\intercal$$
$$=\frac{1}{q}\left[\mathbb{E}\left(\frac{\partial}{\partial u}m(V_1,\alpha^\intercal X_1,g(Z_1))\right)^\intercal\otimes X_1\right]\left[\mathbb{E}\left(\frac{\partial}{\partial u}m(V_1,\alpha^\intercal X_1,g(Z_1))\right)^\intercal\otimes X_1\right]^\intercal,$$
$$h_{12}(\alpha,g) := \frac{1}{q}\sum_{\ell=1}^{q}\left(\mathbb{E}\frac{\partial}{\partial u}m_\ell(V_1,\alpha^\intercal X_1,g(Z_1))X_1\right)\left(\mathbb{E}\frac{\partial}{\partial w}m_\ell(V_1,\alpha^\intercal X_1,g(Z_1))\Phi_K(Z_1)\right)^\intercal$$
$$=\frac{1}{q}\left[\mathbb{E}\left(\frac{\partial}{\partial u}m(V_1,\alpha^\intercal X_1,g(Z_1))\right)^\intercal\otimes X_1\right]\left[\mathbb{E}\left(\frac{\partial}{\partial w}m(V_1,\alpha^\intercal X_1,g(Z_1))\right)^\intercal\otimes \Phi_K(Z_1)\right]^\intercal,$$
$$h_{21}(\alpha,g) := h_{12}(\alpha,g)^\intercal,$$
$$h_{22}(\alpha,g) := \frac{1}{q}\sum_{\ell=1}^{q}\left(\mathbb{E}\frac{\partial}{\partial w}m_\ell(V_1,\alpha^\intercal X_1,g(Z_1))\Phi_K(Z_1)\right)\left(\mathbb{E}\frac{\partial}{\partial w}m_\ell(V_1,\alpha^\intercal X_1,g(Z_1))\Phi_K(Z_1)\right)^\intercal$$
$$=\frac{1}{q}\left[\mathbb{E}\left(\frac{\partial}{\partial w}m(V_1,\alpha^\intercal X_1,g(Z_1))\right)^\intercal\otimes \Phi_K(Z_1)\right]\left[\mathbb{E}\left(\frac{\partial}{\partial w}m(V_1,\alpha^\intercal X_1,g(Z_1))\right)^\intercal\otimes \Phi_K(Z_1)\right]^\intercal.$$

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

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

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

Denote $S_n(\mathbf{a},\mathbf{b}) = (S_{1n}(\mathbf{a},\mathbf{b})^\intercal, S_{2n}(\mathbf{a},\mathbf{b})^\intercal)^\intercal$. We now focus on $S_n(\alpha,\beta)$ with sub-vectors $S_{1n}(\alpha,\beta)$ and $S_{2n}(\alpha,\beta)$. Define

$$s_{1n}(\alpha,g) = \frac{1}{qn}\sum_{\ell=1}^{q}\sum_{i=1}^{n} m_\ell(V_i,\alpha^\intercal X_i,g(Z_i))\mathbb{E}\frac{\partial}{\partial u}m_\ell(V_1,\alpha^\intercal X_1,g(Z_1))X_1,$$

$$= \left[\frac{1}{q}\mathbb{E}\left(\frac{\partial}{\partial u}m(V_1, \alpha^\intercal X_1, g(Z_1))^\intercal \otimes X_1\right)\right]\frac{1}{n}\sum_{i=1}^{n} m(V_i, \alpha^\intercal X_i, g(Z_i)),$$

$$s_{2n}(\alpha, g) = \frac{1}{qn}\sum_{\ell=1}^{q}\sum_{i=1}^{n} m_\ell(V_i, \alpha^\intercal X_i, g(Z_i))\mathbb{E}\frac{\partial}{\partial w}m_\ell(V_1, \alpha^\intercal X_1, g(Z_1))\Phi_K(Z_1)$$

$$= \left[\frac{1}{q}\mathbb{E}\left(\frac{\partial}{\partial w}m(V_1, \alpha^\intercal X_1, g(Z_1))^\intercal \otimes \Phi_K(Z_1)\right)\right]\frac{1}{n}\sum_{i=1}^{n} m(V_i, \alpha^\intercal X_i, g(Z_i)),$$

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*

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

The following lemmas A.4-A.6 are used to prove the results in Subsection 3.3.

Lemma A.4. *Under Assumptions 2.1*-2.2, 3.1*-3.3*, we have*

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

The following assumptions are imposed for the case of single-index structure in Section 3. Their discussions are similar to their counterparts and hence are omitted.

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

Assumption 3.1* *Assumptions (a), (c) and (d) in Assumption 3.1 remain the same but (b) is replaced by*:

(b) for the density $f_\theta(z)$ of $\theta^\intercal Z_1$, there exists two constants $0 < c < C < \infty$* such that *$c\pi(z) \le f_\theta(z) \le C\pi(z)$ on the support $\mathbb{Z}$* of *$\theta^\intercal Z_1$ for θ in some neighbourhood of θ_0.*

Assumption 3.2* *Suppose that there is a unique function $g(\cdot) \in L^2(\mathbb{Z}, \pi)$ and for each n there is a unique vector $\alpha \in \mathbb{R}^p$ such that model (3.5) is satisfied. In other words, for any $\delta > 0$, there is an $\epsilon > 0$ such that*

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

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

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

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

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

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

Similar to $H_n(\mathbf{a}, \mathbf{b})$, we define $\widetilde{H}_n(\mathbf{a}, \mathbf{b})$ as the Hessian matrix of $\|\widetilde{M}_n(\mathbf{a}, \mathbf{b})\|^2$, which has the following blocks:

$$\widetilde{H}_{11}(\mathbf{a}, \mathbf{b}) := \frac{\partial^2}{\partial \mathbf{a} \partial \mathbf{a}^\intercal} \|\widetilde{M}_n(\mathbf{a}, \mathbf{b})\|^2, \qquad \widetilde{H}_{12}(\mathbf{a}, \mathbf{b}) := \frac{\partial^2}{\partial \mathbf{a} \partial \mathbf{b}^\intercal} \|\widetilde{M}_n(\mathbf{a}, \mathbf{b})\|^2$$
$$\widetilde{H}_{22}(\mathbf{a}, \mathbf{b}) := \frac{\partial^2}{\partial \mathbf{b} \partial \mathbf{b}^\intercal} \|\widetilde{M}_n(\mathbf{a}, \mathbf{b})\|^2 \qquad \widetilde{H}_{21}(\mathbf{a}, \mathbf{b}) = \widetilde{H}_{12}(\mathbf{a}, \mathbf{b})^\intercal.$$

Meanwhile, define $\widetilde{h}_n(\alpha, g)$ in the same way as $h_n(\alpha, g)$ given by (A.1) with Z_1 being replaced by $\theta_0^\intercal Z_1$.

Lemma A.5. *Let Assumptions 2.1*-2.2 and 3.5*, 3.6 and 3.7* hold. Then (1)* $\widetilde{H}_n(\alpha, \beta)$ *is asymptotically almost surely positive definite; and (2) we have* $\|\widetilde{H}_n(\alpha, \beta) - \widetilde{h}_n(\alpha, g)\| = o_P(1)$ *as* $n \to \infty$.

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

$$\widetilde{s}_n(\alpha, g) = (\widetilde{s}_{1n}(\alpha, g)^\intercal, \widetilde{s}_{2n}(\alpha, g)^\intercal)^\intercal = \frac{1}{q} \widetilde{\Psi}_n \frac{1}{n} \sum_{i=1}^{n} m(V_i, \alpha^\intercal X_i, g(\theta_0^\intercal Z_i)). \tag{A.3}$$

Lemma A.6. *Under Assumptions 2.1*-2.2, 3.1*, 3.3*, 3.5*, 3.6, 3.7*, as* $n \to \infty$ *we have*

$$\|\widetilde{S}_n(\alpha, \beta) - \widetilde{s}_n(\alpha, g)\| = o_P(1).$$

The following two lemmas are made for the proofs of the theorems in Section 5.

Lemma A.7. *Let Assumptions 5.1-5.2 hold. Suppose that (i) There exists a positive sequence* $a_n = o(d_n)$ *such that* $\|S_{nT}(v_{0S})\| = O_P(a_n)$*; (ii) For any* $\epsilon > 0$*, there exists a constant* $C = C(\epsilon) > 0$ *such that for all large* n, $P(\lambda_{\min}(H_{nT}(v_{0S})) > C) > 1 - \epsilon$*; (iii) For any* $\epsilon > 0$, $\delta > 0$ *and any nonnegative sequence* $\eta_n = o(d_n)$*, there is an* $N > 0$ *such that whenever* $n > N$,

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

Then there exists a local minimizer $\widehat{v} \in \mathcal{V}$ *of*

$$Q_n(v_T) = \|M_n(v_T)\|^2 + \sum_{j \in T} P_n(|v_j|),$$

such that $\|\widehat{v} - v_0\| = O_P(a_n + \sqrt{t_n} P_n'(d_n))$*. Moreover, for any arbitrary* $\epsilon > 0$*, the local minimizer* $\widehat{v}$ *is strict with probability at least* $1 - \epsilon$ *for all large* n.

The proof and the verification of the conditions of the lemma are relegated to Appendix C. It is worth noting that, under an additional condition stated below, we show in Appendix C that $\|S_{nT}(v_{0S})\| = O_P(\sqrt{t_n \log(q)/n})$ and therefore we have $\|\widehat{v} - v_0\| = O_P(\sqrt{t_n \log(q)/n} + \sqrt{t_n} P'_n(d_n))$.

The oracle consistency in Lemma A.7 is derived based on the knowledge of T, the support of v_0. To make the result useful, it is desirable to show that the local minimizer of Q_n restricted on $\mathcal{V}$ is also a minimizer of Q_n on $\mathbb{R}^{p+K}$.

Lemma A.8. *Let the conditions in Lemma A.7 hold. Suppose that with probability approaching one, for $\widehat{v} \in \mathcal{V}$ in Lemma A.7, there exists a neighbourhood $O_1 \subset \mathbb{R}^{p+K}$ of $\widehat{v}$ such that for all $v \in O_1$ but $v \notin \mathcal{V}$, we have*

$$\|M_n(v_T)\|^2 - \|M_n(v)\|^2 < \sum_{j \notin T} P_n(|v_j|). \tag{A.4}$$

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

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

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

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

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

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

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

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,

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

Hence, the inclusion $(\widehat{\alpha},\widehat{\beta}) \in \{(\mathbf{a},\mathbf{b}) : \ \|\mathbf{a}\| \leq B_{1n}, \|\mathbf{b}\| \leq B_{2n}, \|(\mathbf{a}-\alpha,\mathbf{b}-\beta)\| > \delta\}$ holds with probability at most 2ϵ,

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

As ϵ and δ are arbitrarily chosen, we then have $\|(\widehat{\alpha}-\alpha,\widehat{\beta}-\beta)\| \to_P 0$. Notice further that

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

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

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

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

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

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

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

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

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,

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

Recall $h_n(\alpha, g) = \frac{1}{q}\Psi_n\Psi_n^\intercal$ and $s_n(\alpha, g) = \frac{1}{q}\Psi_n \frac{1}{n}\sum_{i=1}^n m(V_i, \alpha^\intercal X_i, g(Z_i))$ by (A.1) 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 [55]. 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

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

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

(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 rates 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 $\|\widehat{\alpha} - \alpha\|^2 = O_P(p/n)$ and $\|\widehat{\beta} - \beta\|^2 = O_P(K/n)$.

Moreover, by the first order Taylor expansion,

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

$$
\begin{aligned}
&+\|\widehat{\beta}-\beta\|^2\frac{4}{n}\sum_{i=1}^{n}\left\|\frac{\partial m(V_i,\alpha^\intercal X_i,g(Z_i))}{\partial w}\otimes\Phi_K(Z_i)\right\|^2\\
&+\frac{4}{n}\sum_{i=1}^{n}\left|\kappa^\intercal\frac{\partial m(V_i,\alpha^\intercal X_i,g(Z_i))}{\partial w}\right|^2\gamma_K^2(Z_i)\\
&+2\|\widehat{\alpha}-\alpha\|\left(\frac{1}{n}\sum_{i=1}^{n}\left\|\frac{\partial m(V_i,\alpha^\intercal X_i,g(Z_i))}{\partial u}\otimes X_i\right\|^2\right)^{1/2}\\
&\quad\times\left(\kappa^\intercal\frac{1}{n}\sum_{i=1}^{n}m(V_i,\alpha^\intercal X_i,g(Z_i))m(V_i,\alpha^\intercal X_i,g(Z_i))^\intercal\kappa\right)^{1/2}\\
&+2\left(\frac{1}{n}\sum_{i=1}^{n}\left|\kappa^\intercal\frac{\partial m(V_i,\alpha^\intercal X_i,g(Z_i))}{\partial w}\right|^2(\widehat{g}(Z_i)-g(Z_i))^2\right)^{1/2}\\
&\quad\times\left(\kappa^\intercal\frac{1}{n}\sum_{i=1}^{n}m(V_i,\alpha^\intercal X_i,g(Z_i))m(V_i,\alpha^\intercal X_i,g(Z_i))^\intercal\kappa\right)^{1/2}\\
=&\|\widehat{\alpha}-\alpha\|^2O_P(pq)+\|\widehat{\beta}-\beta\|^2O_P(Kq)+O_P(q)\sup_z\gamma_K^2(z)\\
&+\|\widehat{\alpha}-\alpha\|O_P(\sqrt{pq})+\|\widehat{\beta}-\beta\|O_P(\sqrt{Kq})+O_P(\sqrt{q})\sup_z|\gamma_K(z)|\\
=&o_P(1)
\end{aligned}
$$

by Assumptions 3.5 and 4.2. 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

$$
\begin{aligned}
&m(V_i,\mathbf{a}^\intercal X_i,f(Z_i))-m(V_i,\alpha^\intercal X_i,g(Z_i))\\
&=\frac{\partial m(V_i,\alpha^\intercal X_i,g(Z_i))}{\partial u}(\mathbf{a}-\alpha)^\intercal X_i+\frac{\partial m(V_i,\alpha^\intercal X_i,g(Z_i))}{\partial w}(f(Z_i)-g(Z_i)),
\end{aligned}\tag{B.3}
$$

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

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

$$
\begin{aligned}
&+P\left(\sup_{\|(\mathbf{a},f)-(\alpha,g)\|<\delta}\left|\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\left[\kappa^{\intercal}\frac{\partial m}{\partial w}\gamma_K(Z_i)-E\kappa^{\intercal}\frac{\partial m}{\partial w}\gamma_K(Z_i)\right]\right|>\eta/4\right)\\
\leq& P\left(\sup_{\|(\mathbf{a},f)-(\alpha,g)\|<\delta}\left\|\frac{1}{\sqrt{np}}\sum_{i=1}^{n}\left[\kappa^{\intercal}\frac{\partial m}{\partial u}X_i-E\kappa^{\intercal}\frac{\partial m}{\partial u}X_i\right]\right\|\|\sqrt{p}(\mathbf{a}-\alpha)\|>\eta/2\right)\\
&+P\left(\sup_{\|(\mathbf{a},f)-(\alpha,g)\|<\delta}\left\|\frac{1}{\sqrt{nK}}\sum_{i=1}^{n}\left[\kappa^{\intercal}\frac{\partial m}{\partial w}\Phi_K(Z_i)-E\kappa^{\intercal}\frac{\partial m}{\partial w}\Phi_K(Z_i)\right]\right\|\|\sqrt{K}(\mathbf{b}-\beta)\|>\eta/4\right)\\
&+P\left(\sup_{\|(\mathbf{a},f)-(\alpha,g)\|<\delta}\left|\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\left[\kappa^{\intercal}\frac{\partial m}{\partial w}\gamma_K(Z_i)-E\kappa^{\intercal}\frac{\partial m}{\partial w}\gamma_K(Z_i)\right]\right|>\eta/4\right)\\
:=&I_{1n}+I_{2n}+I_{3n},\quad \text{say.}
\end{aligned}
$$

Observe by the i.i.d. property that

$$
\frac{1}{\sqrt{np}}\sum_{i=1}^{n}\left[\kappa^{\intercal}\frac{\partial m}{\partial u}X_i-E\kappa^{\intercal}\frac{\partial m}{\partial u}X_i\right]=O_P(1), \tag{B.4}
$$

$$
\frac{1}{\sqrt{nK}}\sum_{i=1}^{n}\left[\kappa^{\intercal}\frac{\partial m}{\partial w}\Phi_K(Z_i)-E\kappa^{\intercal}\frac{\partial m}{\partial w}\Phi_K(Z_i)\right]=O_P(1). \tag{B.5}
$$

It follows that if $\|\sqrt{p}(\mathbf{a}-\alpha)\|$ and $\|\sqrt{K}(\mathbf{b}-\beta)\|$ are sufficiently small, $I_{1n}<\varepsilon/3$ and $I_{2n}<\varepsilon/3$. Meanwhile, using the condition that $\sqrt{q}\sup_z|\gamma_K(z)|=o(1)$ 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

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

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 $\|\kappa\|=1$,

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

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

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

as $n\to\infty$ for any $(\mathbf{a},\mathbf{b})\in\mathbb{R}^{p+K}$.

Note by the Law of Large Numbers that

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

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,

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

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

Proof of Theorem 4.3. Note that

$$\begin{aligned}\widehat{\sigma}^2 =& \frac{1}{n}\sum_{i=1}^n m(i)m(i)^\intercal + \frac{1}{n}\sum_{i=1}^n [\widehat{m}(i)\widehat{m}(i)^\intercal - m(i)m(i)^\intercal] \\ :=& \mathbb{E}[m(i)m(i)^\intercal](1 + O_P(qn^{-1/2})) + \Delta_{\sigma,n}\end{aligned}$$

by the Law of Large Numbers, where $m(i) := m(V_i, \alpha^\intercal X_i, g(Z_i))$ for simplicity, and it follows from Assumption 3.3 that

$$\begin{aligned}\|\Delta_{\sigma,n}\| \leq& \frac{1}{n}\sum_{i=1}^n \|\widehat{m}(i)\widehat{m}(i)^\intercal - m(i)m(i)^\intercal\| \\ \leq& \frac{1}{n}\sum_{i=1}^n \|\widehat{m}(i) - m(i)\|^2 + 2\frac{1}{n}\sum_{i=1}^n \|m(i)\|\|(\widehat{m}(i) - m(i))\| \\ =& \sqrt{q}\, O_P(\|\widehat{\alpha} - \alpha\| + \|\widehat{g} - g\|) = o_P(1).\end{aligned}$$

This gives $\Lambda_n = M_q^2 + o_P(1)$ where $\Lambda_n := \mathrm{Diag}(\widehat{\sigma}(j,j)^2, j = 1, \cdots, q)$ and $M_q^2 := \mathrm{Diag}(\mathbb{E}[m_j(i)^2], j = 1, \cdots, q)$. Notice also that

$$\widehat{e} = \frac{1}{n}\sum_{i=1}^n m(i) + \frac{1}{n}\sum_{i=1}^n [\widehat{m}(i) - m(i)] := e_n + \Delta_{e,n},$$

where $\sqrt{n}\kappa^\intercal \Delta_{e,n} \to_P 0$ has been proven by Theorem 4.1, implying $\widehat{e} = e_n + o_P(1)$ as $n \to \infty$.

Because the difference of using $\Delta_{\sigma,n}$ and $\Delta_{e,n}$ is negligible in probability, as shown in the above, we may consider, a bit loosely use of the notation,

$$\begin{aligned}T_n =& \frac{1}{q}\sum_{j=1}^q \frac{1}{\mathbb{E}[m_j(i)^2]}\left(\frac{1}{\sqrt{n}}\sum_{i=1}^n m_j(i)\right)^2 = \frac{1}{q}\sum_{j=1}^q \frac{1}{\mathbb{E}[m_j(i)^2]}\frac{1}{n}\sum_{i=1}^n\sum_{i'=1}^n m_j(i)m_j(i') \\ =& \frac{1}{qn}\sum_{i=1}^n \left(\sum_{j=1}^q \frac{1}{\mathbb{E}[m_j(i)^2]} m_j(i)^2\right) + \frac{2}{qn}\sum_{i=2}^n \left(\sum_{j=1}^q \frac{1}{\mathbb{E}[m_j(i)^2]} m_j(i)\sum_{i'=1}^{i-1} m_j(i')\right) \\ :=& T_{n1} + T_{n2}, \quad \text{say.}\end{aligned} \tag{B.6}$$

We first consider the second term T_{n2}. It is obvious that, given $\mathcal{F}_{i-1}$ the information up to $i-1$,

$$\xi_{ni} := \frac{2}{qn}\sum_{j=1}^q \frac{1}{\mathbb{E}[m_j(i)^2]} m_j(i)\sum_{i'=1}^{i-1} m_j(i')$$

is a martingale difference sequence, so that $T_{n2} = \sum_{i=2}^n \xi_{ni}$ becomes a martingale. The conditional variance is

$$D_n^2 = \sum_{i=2}^n \mathbb{E}[\xi_{ni}^2 | \mathcal{F}_{i-1}]$$

$$
\begin{aligned}
&= \sum_{i=2}^{n} \mathbb{E}\left[\left(\frac{2}{qn} \sum_{j=1}^{q} \frac{1}{\mathbb{E}[m_j(i)^2]} m_j(i) \sum_{i'=1}^{i-1} m_j(i') \right)^2 \Big| \mathcal{F}_{i-1} \right] \\
&= \frac{4}{q^2 n^2} \sum_{i=2}^{n} \sum_{j=1}^{q} \left(\sum_{i'=1}^{i-1} \frac{m_j(i')}{\sqrt{\mathbb{E}[m_j(i)^2]}} \right)^2 \\
&\quad + \frac{4}{q^2 n^2} \sum_{i=2}^{n} \sum_{j=1}^{q} \sum_{j'=1, \neq j}^{q} \frac{\mathbb{E}[m_j(i) m_{j'}(i)]}{\mathbb{E}[m_j(i)^2]} \sum_{i'=1}^{i-1} m_j(i') \sum_{i'=1}^{i-1} m_{j'}(i') \\
&= \frac{4}{q^2 n^2} \sum_{i=2}^{n} \sum_{j=1}^{q} \left(\sum_{i'=1}^{i-1} \frac{m_j(i')^2}{\mathbb{E}[m_j(i)^2]} + \sum_{i_1=1}^{i-1} \sum_{i_2=1, \neq i_1}^{i-1} \frac{m_j(i_1) m_j(i_2)}{\mathbb{E}[m_j(i)^2]} \right),
\end{aligned}
$$

due to $\mathbb{E}[m_j(i) m_{j'}(i)] = 0$ for $j \neq j'$. It follows that

$$
\mathbb{E}[D_n^2] = \frac{4}{qn^2} \sum_{i=2}^{n} (i-1) = \frac{4}{qn^2} \frac{n(n-1)}{2} = 2 \frac{n-1}{qn}.
$$

In addition,

$$
\begin{aligned}
&\mathbb{E}[(D_n^2 - \mathbb{E}[D_n^2])^2] \\
=&\mathbb{E}\left[\frac{4}{q^2 n^2} \sum_{i=2}^{n} \sum_{j=1}^{q} \left(\sum_{i'=1}^{i-1} \frac{m_j(i')^2}{\mathbb{E}[m_j(i)^2]} + \sum_{i_1=1}^{i-1} \sum_{i_2=1, \neq i_1}^{i-1} \frac{m_j(i_1) m_j(i_2)}{\mathbb{E}[m_j(i)^2]} \right) - \frac{4}{qn^2} \sum_{i=2}^{n} (i-1) \right]^2 \\
=&\frac{16}{q^4 n^4} \mathbb{E}\left[\sum_{i=2}^{n} \sum_{j=1}^{q} \left(\sum_{i'=1}^{i-1} \frac{m_j(i')^2 - \mathbb{E}[m_j(i)^2]}{\mathbb{E}[m_j(i)^2]} + \sum_{i_1=1}^{i-1} \sum_{i_2=1, \neq i_1}^{i-1} \frac{m_j(i_1) m_j(i_2)}{\mathbb{E}[m_j(i)^2]} \right) \right]^2 \\
\leq&\frac{32}{q^4 n^4} \mathbb{E}\left(\sum_{i=2}^{n} \sum_{j=1}^{q} \sum_{i'=1}^{i-1} \frac{m_j(i')^2 - \mathbb{E}[m_j(i)^2]}{\mathbb{E}[m_j(i)^2]} \right)^2 + \frac{32}{q^4 n^4} \mathbb{E}\left(\sum_{i=2}^{n} \sum_{j=1}^{q} \sum_{i_1=1}^{i-1} \sum_{i_2=1, \neq i_1}^{i-1} \frac{m_j(i_1) m_j(i_2)}{\mathbb{E}[m_j(i)^2]} \right)^2 \\
:=&I_1 + I_2, \quad \text{say.}
\end{aligned}
$$

Moreover,

$$
\begin{aligned}
I_1 =& \frac{32}{q^4 n^4} \mathbb{E}\left(\sum_{i=2}^{n} \sum_{j=1}^{q} \sum_{i'=1}^{i-1} \frac{m_j(i')^2 - \mathbb{E}[m_j(i')^2]}{\mathbb{E}[m_j(i')^2]} \right)^2 \\
=& \frac{32}{q^4 n^4} \sum_{i=2}^{n} \mathbb{E}\left(\sum_{j=1}^{q} \sum_{i'=1}^{i-1} \frac{m_j(i')^2 - \mathbb{E}[m_j(i')^2]}{\mathbb{E}[m_j(i')^2]} \right)^2 \\
&+ \frac{64}{q^4 n^4} \sum_{i_3=3}^{n} \sum_{i_4=2}^{i_3-1} \mathbb{E}\left[\left(\sum_{j=1}^{q} \sum_{i'=1}^{i_3-1} \frac{m_j(i')^2 - \mathbb{E}[m_j(i')^2]}{\mathbb{E}[m_j(i')^2]} \right) \left(\sum_{j=1}^{q} \sum_{i'=1}^{i_4-1} \frac{m_j(i')^2 - \mathbb{E}[m_j(i')^2]}{\mathbb{E}[m_j(i')^2]} \right) \right] \\
=& \frac{32}{q^4 n^4} \sum_{i=2}^{n} \sum_{i'=1}^{i-1} \mathbb{E}\left(\sum_{j=1}^{q} \frac{m_j(i')^2 - \mathbb{E}[m_j(i')^2]}{\mathbb{E}[m_j(i')^2]} \right)^2 \\
&+ \frac{64}{q^4 n^4} \sum_{i_3=3}^{n} \sum_{i_4=2}^{i_3-1} \mathbb{E}\left(\sum_{j=1}^{q} \sum_{i'=1}^{i_4-1} \frac{m_j(i')^2 - \mathbb{E}[m_j(i')^2]}{\mathbb{E}[m_j(i')^2]} \right)^2
\end{aligned}
$$

$$
\begin{aligned}
=&\frac{32}{q^4n^4}\sum_{i=2}^{n}\sum_{i'=1}^{i-1}\sum_{j=1}^{q}\mathbb{E}\left(\frac{m_j(i')^2-\mathbb{E}[m_j(i')^2]}{\mathbb{E}[m_j(i')^2]}\right)^2\\
&+\frac{64}{q^4n^4}\sum_{i=2}^{n}\sum_{i'=1}^{i-1}\sum_{j_1=2}^{q}\sum_{j_2=1}^{j_1-1}\mathbb{E}\left(\frac{m_{j_1}(i')^2-\mathbb{E}[m_{j_1}(i')^2]}{\mathbb{E}[m_{j_1}(i')^2]}\frac{m_{j_2}(i')^2-\mathbb{E}[m_{j_2}(i')^2]}{\mathbb{E}[m_{j_2}(i')^2]}\right)\\
&+\frac{64}{q^4n^4}\sum_{i_3=3}^{n}\sum_{i_4=2}^{i_3-1}\sum_{i'=1}^{i_4-1}\sum_{j=1}^{q}\mathbb{E}\left(\frac{m_j(i')^2-\mathbb{E}[m_j(i')^2]}{\mathbb{E}[m_j(i')^2]}\right)^2\\
&+\frac{64}{q^4n^4}\sum_{i_3=3}^{n}\sum_{i_4=2}^{i_3-1}\sum_{i'=1}^{i_4-1}\sum_{j_1=2}^{q}\sum_{j_1=2}^{q}\mathbb{E}\left(\frac{m_{j_1}(i')^2-\mathbb{E}[m_{j_1}(i')^2]}{\mathbb{E}[m_{j_1}(i')^2]}\frac{m_{j_2}(i')^2-\mathbb{E}[m_{j_2}(i')^2]}{\mathbb{E}[m_{j_2}(i')^2]}\right)\\
\le&C\frac{1}{q^2n},
\end{aligned}
$$

and

$$
\begin{aligned}
I_2=&\frac{64}{q^4n^4}\mathbb{E}\left(\sum_{i=3}^{n}\sum_{j=1}^{q}\sum_{i_1=2}^{i-1}\sum_{i_2=1}^{i_1-1}\frac{m_j(i_1)m_j(i_2)}{\mathbb{E}[m_j(i)^2]}\right)^2\\
=&\frac{64}{q^4n^4}\sum_{i=3}^{n}\mathbb{E}\left(\sum_{i_1=2}^{i-1}\sum_{i_2=1}^{i_1-1}\sum_{j=1}^{q}\frac{m_j(i_1)m_j(i_2)}{\mathbb{E}[m_j(i)^2]}\right)^2\\
&+\frac{128}{q^4n^4}\sum_{i_5=4}^{n}\sum_{i_6=3}^{i_5-1}\mathbb{E}\left(\sum_{i_1=2}^{i_5-1}\sum_{i_2=1}^{i_1-1}\sum_{j=1}^{q}\frac{m_j(i_1)m_j(i_2)}{\mathbb{E}[m_j(i)^2]}\right)\left(\sum_{j=1}^{q}\sum_{i_1=2}^{i_6-1}\sum_{i_2=1}^{i_1-1}\frac{m_j(i_1)m_j(i_2)}{\mathbb{E}[m_j(i)^2]}\right)\\
=&\frac{64}{q^4n^4}\sum_{i=3}^{n}\sum_{i_1=2}^{i-1}\mathbb{E}\left(\sum_{i_2=1}^{i_1-1}\sum_{j=1}^{q}\frac{m_j(i_1)m_j(i_2)}{\mathbb{E}[m_j(i)^2]}\right)^2\\
&+\frac{128}{q^4n^4}\sum_{i=3}^{n}\sum_{i_1=3}^{i-1}\sum_{i_7=2}^{i_1-1}\mathbb{E}\left(\sum_{i_2=1}^{i_1-1}\sum_{j=1}^{q}\frac{m_j(i_1)m_j(i_2)}{\mathbb{E}[m_j(i)^2]}\right)\left(\sum_{i_2=1}^{i_7-1}\sum_{j=1}^{q}\frac{m_j(i_7)m_j(i_2)}{\mathbb{E}[m_j(i)^2]}\right)\\
&+\frac{128}{q^4n^4}\sum_{i_5=4}^{n}\sum_{i_6=3}^{i_5-1}\mathbb{E}\left(\sum_{i_1=2}^{i_6-1}\sum_{i_2=1}^{i_1-1}\sum_{j=1}^{q}\frac{m_j(i_1)m_j(i_2)}{\mathbb{E}[m_j(i)^2]}\right)^2\\
=&\frac{64}{q^4n^4}\sum_{i=3}^{n}\sum_{i_1=2}^{i-1}\sum_{i_2=1}^{i_1-1}\mathbb{E}\left(\sum_{j=1}^{q}\frac{m_j(i_1)m_j(i_2)}{\mathbb{E}[m_j(i)^2]}\right)^2\\
&+\frac{128}{q^4n^4}\sum_{i_5=4}^{n}\sum_{i_6=3}^{i_5-1}\sum_{i_1=2}^{i_6-1}\sum_{i_2=1}^{i_1-1}\mathbb{E}\left(\sum_{j=1}^{q}\frac{m_j(i_1)m_j(i_2)}{\mathbb{E}[m_j(i)^2]}\right)^2\\
\le&C\frac{1}{q^3n}.
\end{aligned}
$$

Thus, $D_n^2-\mathbb{E}[D_n^2]=O_P(n^{-1/2}q^{-1})$. Also note that

$$
\left(\frac{D_n^2}{\mathbb{E}[D_n^2]}-1\right)^2=\frac{(D_n^2-\mathbb{E}[D_n^2])^2}{(\mathbb{E}[D_n^2])^2}=O_P\left(n^{-1}\right)=o_P(1).
$$

To show the asymptotic normality of $T_{n2}=\sum_{i=2}^{n}\xi_{ni}$, according to Corollary 3.1 of Hall and Heyde [39], we need to check whether for any $\eta>0$,

$$
\sum_{i=2}^{n}\mathbb{E}[\xi_{ni}^2\mathbb{I}(|\xi_{ni}|>\eta)|\mathcal{F}_{t-1}]\to_P 0.
$$

To this end, it suffices to show $\sum_{i=2}^n \mathbb{E}[\xi_{ni}^4|\mathcal{F}_{t-1}] \to_P 0$, or to show $\sum_{i=2}^n \mathbb{E}[\xi_{ni}^4] \to 0$. Indeed,

$$
\begin{aligned}
\sum_{i=2}^n \mathbb{E}[\xi_{ni}^4] =& \frac{16}{q^4n^4}\sum_{i=2}^n \mathbb{E}\left(\sum_{j=1}^q \frac{1}{\mathbb{E}[m_j(i)^2]} m_j(i) \sum_{i'=1}^{i-1} m_j(i')\right)^4 \\
=& \frac{16}{q^4n^4}\sum_{i=2}^n \sum_{j=1}^q \mathbb{E}\left(\frac{1}{\mathbb{E}[m_j(i)^2]} m_j(i) \sum_{i'=1}^{i-1} m_j(i')\right)^4 \\
&+ \frac{96}{q^4n^4}\sum_{i=2}^n \sum_{j_1=1}^q \sum_{j_2=1,\neq j_2}^q \\
&\quad \mathbb{E}\left[\left(\frac{1}{\mathbb{E}[m_{j_1}(i)^2]} m_{j_1}(i) \sum_{i'=1}^{i-1} m_{j_1}(i')\right)^2 \left(\frac{1}{\mathbb{E}[m_{j_2}(i)^2]} m_{j_2}(i) \sum_{i'=1}^{i-1} m_{j_2}(i')\right)^2\right] \\
=& \frac{16}{q^4n^4}\sum_{i=2}^n \sum_{j=1}^q \frac{1}{(\mathbb{E}[m_j(i)^2])^4}\mathbb{E}[m_j(i)^4]\mathbb{E}\left(\sum_{i'=1}^{i-1} m_j(i')\right)^4 \\
&+ \frac{96}{q^4n^4}\sum_{i=2}^n \sum_{j_1=1}^q \sum_{j_2=1,\neq j_2}^q \frac{1}{(\mathbb{E}[m_{j_1}(i)^2])^2}\frac{1}{(\mathbb{E}[m_{j_2}(i)^2])^2} \\
&\quad \mathbb{E}[m_{j_2}(i)^2 m_{j_1}(i)^2]\mathbb{E}\left[\left(\sum_{i'=1}^{i-1} m_{j_1}(i')\right)^2 \left(\sum_{i'=1}^{i-1} m_{j_2}(i')\right)^2\right] \\
=& \frac{16}{q^4n^4}\sum_{i=2}^n \sum_{j=1}^q \frac{1}{(\mathbb{E}[m_j(i)^2])^4}\mathbb{E}[m_j(i)^4] \\
&\quad \times \left[\sum_{i'=1}^{i-1} \mathbb{E}[m_j(i')^4] + 6\sum_{i_1=1}^{i-1}\sum_{i_2=1,\neq i_1}^{i-1} \mathbb{E}[m_j(i_1)^2 m_j(i_2)^2]\right] \\
&+ \frac{96}{q^4n^4}\sum_{i=2}^n \sum_{j_1=1}^q \sum_{j_2=1,\neq j_2}^q \frac{\mathbb{E}[m_{j_2}(i)^2 m_{j_1}(i)^2]}{(\mathbb{E}[m_{j_1}(i)^2])^2(\mathbb{E}[m_{j_2}(i)^2])^2} \\
&\quad \times \mathbb{E}\left[\left(\sum_{i'=1}^{i-1} m_{j_1}(i')^2 + \sum_{i_1=1}^{i-1}\sum_{i_2=1,\neq i_1}^{i-1} m_{j_1}(i_1)m_{j_1}(i_2)\right)\right. \\
&\qquad \left.\times \left(\sum_{i'=1}^{i-1} m_{j_2}(i')^2 + \sum_{i_3=1}^{i-1}\sum_{i_4=1,\neq i_3}^{i-1} m_{j_2}(i_3)m_{j_2}(i_4)\right)\right] \\
\leq& C_1\frac{1}{q^4n^4}\sum_{i=2}^n \sum_{j=1}^q (i+i^2) \\
&+ \frac{96}{q^4n^4}\sum_{i=2}^n \sum_{j_1=1}^q \sum_{j_2=1,\neq j_2}^q \frac{\mathbb{E}[m_{j_2}(i)^2 m_{j_1}(i)^2]}{(\mathbb{E}[m_{j_1}(i)^2])^2(\mathbb{E}[m_{j_2}(i)^2])^2} \\
&\quad \times \left[\sum_{i'=1}^{i-1} \mathbb{E}[m_{j_1}(i')^4] + 2\sum_{i_1=1}^{i-1}\sum_{i_2=1,\neq i_1}^{i-1} \mathbb{E}[m_{j_1}(i_1)^2 m_{j_1}(i_2)^2]\right] \\
\leq& C\frac{1}{q^2n} \to 0.
\end{aligned}
$$

Thus, $D_n^{-1}\sum_{i=2}^n \xi_{ni} \to_D N(0,1)$ as $n \to \infty$.

On the other hand, the first term T_{n1} of T_n in (B.6) converges to 1 in probability. In fact,

$$\begin{aligned}
&\mathbb{E}\left[\frac{1}{qn}\sum_{i=1}^{n}\left(\sum_{j=1}^{q}\frac{1}{\mathbb{E}[m_j(i)^2]}m_j(i)^2\right)-1\right]^2\\
=&\mathbb{E}\left[\frac{1}{qn}\sum_{i=1}^{n}\left(\sum_{j=1}^{q}\frac{1}{\mathbb{E}[m_j(i)^2]}(m_j(i)^2-\mathbb{E}[m_j(i)^2])\right)\right]^2\\
=&\frac{1}{q^2n^2}\sum_{i=1}^{n}\mathbb{E}\left[\left(\sum_{j=1}^{q}\frac{1}{\mathbb{E}[m_j(i)^2]}(m_j(i)^2-\mathbb{E}[m_j(i)^2])\right)\right]^2\\
=&\frac{1}{q^2n^2}\sum_{i=1}^{n}\sum_{j=1}^{q}\mathbb{E}\left[\frac{1}{\mathbb{E}[m_j(i)^2]}(m_j(i)^2-\mathbb{E}[m_j(i)^2])\right]^2\\
&+\frac{1}{q^2n^2}\sum_{i=1}^{n}\sum_{j_1=1}^{q}\sum_{j_2=1,\neq j_1}^{q}\frac{\mathbb{E}\left[(m_{j_1}(i)^2-\mathbb{E}[m_{j_1}(i)^2])(m_{j_2}(i)^2-\mathbb{E}[m_{j_2}(i)^2])\right]}{\mathbb{E}[m_{j_1}(i)^2]\mathbb{E}[m_{j_2}(i)^2]}\\
\leq&C\frac{1}{qn}.
\end{aligned}$$

It follows that $T_{n1}-1=O_P((qn)^{-1/2})$. Thence,

$$\sqrt{q/2}(T_n-1)=\sqrt{q/2}O_P((qn)^{-1/2})+\frac{D_n}{\sqrt{\mathbb{E}(D_n^2)}}\frac{1}{D_n}T_{n2}\xrightarrow{d}N(0,1),$$

as $n\to\infty$. □

Proof of Theorem 4.4. Note by the i.i.d property of the data that

$$\left\|\widetilde{\sigma}-\mathbb{E}[\widetilde{m}(i)\widetilde{m}(i)^\intercal]\right\|=\left\|\frac{1}{n}\sum_{i=1}^{n}\widetilde{m}(i)\widetilde{m}(i)^\intercal-\mathbb{E}[\widetilde{m}(i)\widetilde{m}(i)^\intercal]\right\|=O_P\left(\frac{1}{\sqrt{n}}q\right)=o_P(1).$$

Moreover,

$$\begin{aligned}
\widetilde{T}_n=&\frac{1}{q}\sum_{j=1}^{q}\left(\frac{\sqrt{n}\,\widetilde{e}_j}{\widetilde{\sigma}_n(j,j)}\right)^2=(1+o_P(1))\frac{1}{q}n\sum_{j=1}^{q}\frac{1}{\mathbb{E}[\widetilde{m}_j(i)^2]}(\mathbb{E}[\widetilde{m}_j(i)])^2\\
\geq&C^{-1}(1+o_P(1))\frac{1}{q}n\sum_{j=1}^{q}(\mathbb{E}[\widetilde{m}_j(i)])^2=C^{-1}(1+o_P(1))\frac{1}{q}n\|\mathbb{E}[\widetilde{m}_j(i)]\|^2\\
\geq&C^{-1}(1+o_P(1))\frac{1}{q}n\delta_n^2\to_P\infty,
\end{aligned}$$

as $n\to\infty$. □

Proof of Theorem 5.1. (i) and (ii). As shown in Lemma A.8, if $Q_n(v)$ has a local minimizer $\widehat{v}=(\widehat{v}_S^\intercal,\widehat{v}_N^\intercal)^\intercal$, then $\widehat{v}_N=0$ with probability arbitrarily close to one for large n, which implies the assertion (i) and $P(\widehat{T}\subset T)\to 1$.

On the other hand,

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

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

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

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

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

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

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

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

Notice that the structure

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

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 $\sqrt{n}\Sigma_{nT}^{-1}P_n'(|\widehat{v}_S|) \diamond \mathrm{sgn}(\widehat{v}_S) = o_P(1)$. Similar to Lemma C.2 of Fan and Liao [34] we may show that

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

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

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

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

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

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

Notice further,

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

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

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

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

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

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

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

Notice further that

$$\begin{aligned}Q_n(v) \ge \|M_n(v)\|^2 =& \frac{1}{q}\left\|\frac{1}{n}\sum_{i=1}^n m(V_i, v^\intercal F_i)\right\|^2 \\ \ge& \frac{1}{2q}\|Em(V_1, v^\intercal F_1)\|^2 - \frac{1}{q}\left\|\frac{1}{n}\sum_{i=1}^n m(V_i, v^\intercal F_i) - Em(V_1, v^\intercal F_1)\right\|^2 \\ =& \frac{1}{2q}\|Em(V_1, v^\intercal F_1)\| + o_P(n^{-1/2}),\end{aligned}$$

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

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

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

Taking $0 < \eta < \epsilon$,

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

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

References

[1] Ai, C. and Chen, X. (2003). Efficient estimation of models with conditional moment restrictions containing unknown functions. *Econometrica*, 71:1795–1843.

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

[3] 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.

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

[5] Athey, S., Imbens, G., Pham, T., and Wager, S. (2017). Estimating average treatment effects: supplementary analysis and remaining challenges. *American Economic Review*, 107:278–281.

[6] Belloni, A., Chen, D., Chernozhukov, V., and Hansen, C. (2012). Sparse models and methods for optimal instruments with an application to eminent domain. *Econometrica*, 80:2369–2429.

[7] Belloni, A. and Chernozhukov, V. (2013). Least squares after model selection in high-dimensional sparse models. *Bernoulli*, 19:521–547.

[8] 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.

[9] Belloni, A., Chernozhukov, V., and Hansen, C. (2014a). High-dimensional methods and inference on structural and treatment effects. *Journal of Economic Perspectives*, 28:29–50.

[10] Belloni, A., Chernozhukov, V., Hansen, C., and Wei, Y. (2016a). Post–selection inference for generalized linear models with many controls. *Journal of Business and Economic Statistics*, 34:590–605.

[11] Belloni, A., Chernozhukov, V., and Wang, L. (2014b). Pivotal estimation via square-root Lasso in nonparametric regression. *Annals of Statistics*, 42:757–788.

[12] Belloni, A., Rosenbaum, M., and Tsybakov, A. B. (2016b). Linear and conic programming estimators in high–dimensional errors–in–variables models. *Electronic Journal of Statistics*, 10:1729–1750.

[13] Bickel, P. J. (1982). On adaptive estimation. *Annals of Statistics*, 10:647–671.

[14] Bickel, P. J., Klaassen, C. A., Ritov, Y., and Wellner, J. A. (1993). *Efficient and Adaptive Estimation for Semiparametric Models*. The John Hopkins University Press, Baltimore and London.

[15] Bickel, P. J. and Levina, E. (2008). Covariance regularization by thresholding. *Annals of Statistics*, 36:2577–2604.

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

[17] Carneiro, P., Heckman, J., and Vytlacil, E. (2011). Estimating marginal returns to education. *American Economic Review*, 101:2754–2781.

[18] Cattaneo, M. D., Jansson, M., and Newey, W. K. (2018). Inference in linear regression models with many covariates and heteroskedasticity. *Journal of the American Statistical Association*, 113:1350–1361.

[19] 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.

[20] Chen, X. (2007). *Handbook of Econometrics*, volume 6B, chapter Large sample sieve estimation of semi-parametric models, pages 5550–5588. Elsevier, Amsterdam: North Holland.

[21] 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.

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

[23] Chen, X., Linton, O., and Keilegom, I. V. (2003). Estimation for semiparametric models when the criterion function is not smooth. *Econometrica*, 71:1591–1608.

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

[25] Chen, X. and Shen, X. (1998). Sieve extremum estimates for weakly dependent data. *Econometrica*, 66:289–314.

[26] Chernozhukov, V., Chetverikovz, D., Demirery, M., Dufloy, E., Hansenx, C., Neweyy, W., and Robins, J. (2018). Double/debiased machine learning for treatment and structural parameters. *Econometrics Journal*, 21:1–68.

[27] Connor, G., Hagmann, M., and Linton, O. (2012). Efficient semiparametric estimation of the Fama-French model and extensions. *Econometrica*, 80:713–754.

[28] Donald, S. and Newey, W. K. (2001). Choosing the number of instruments. *Econometrica*, 69:1161–1191.

[29] Dong, C., Gao, J., and Tjøstheim, D. (2016). Estimation for single-index and partially linear single-index integrated models. *Annals of Statistics*, 44:425–453.

[30] Dong, C. and Linton, O. (2018). Additive nonparametric models with time variable and both stationary and nonstationary regressors. *Journal of Econometrics*, 207:212–236.

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

[32] 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.

[33] 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.

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

[35] Gao, J. and Anh, V. (2000). A central limit theorem for a random quadratic form of strictly stationary processes. *Statistic and Probability Letters*, 49:69–79.

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

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

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

[39] Hall, P. and Heyde, C. C. (1980). *Martingale Limit Theory and Its Application*. Academic Press, New York.

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

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

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

[43] Ichimura, H. and Linton, O. (2003). Asymptotic expansions for some semiparametric program evaluation estimators. *LSE Research Online Documents on Economics 2098*, Working paper.

[44] Imbens, G. W. (2002). Generalized method of moments and empirical likelihood. *Journal of Business and Economic Statistics*, 20:493–506.

[45] Jankova, J. and Geer, S. V. D. (2018). Semiparametric efficiency bounds for high dimensional models. *Annals of Statistics*, 46:2336–2359.

[46] Linton, O. (1995). Asymptotic expansions for some semiparametric program evaluation estimators. *Econometrica*, 63:1079–1112.

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

[48] Newey, W. K. (1994). The asymptotic variance of semiparametric estimators. *Econometrica*, 62:1349–1382.

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

[50] Newey, W. K. and McFadden, D. F. (1994). *Handbook of Econometrics*, volume IV, chapter Large sample estimation and hypothesis testing, pages 2111–2245. Elsevier, Amsterdam: North Holland.

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

[52] Newey, W. K. and Smith, R. (2004). Higher order properties of GMM and generalized empirical likelihood estimators. *Econometrica*, 72:219–255.

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

[54] Pakes, A. and Olley, S. (1995). A limit theorem for a smooth class of semiparametric estimators. *Journal of Econometrics*, 65:295–332.

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

[56] Pesaran, M. H. and Yamagata, T. (2017). Testing for alpha in linear factor pricing models with a large number of securities. CESifo Working Paper Series No. 6432, Available at SSRN: https://ssrn.com/abstract=2973079.

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

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

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

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

[61] Stone, C. J. (1985). Additive regression and other nonparametric models. *Annals of Statistics*, 13:689–705.

[62] Su, L., Ura, T., and Zhang, Y. (2018). Non-separable models with high-dimensional data. https://arxiv.org/abs/1702.04625.

[63] Yu, Y. and Ruppert, D. (2002). Penalized spline estimation for partially linear single-index models. *Journal of the American Statistical Association*, 97:1042–1054.

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

Supplementary material to "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

Abstract The proofs of technical lemmas and extra outcomes of empirical study are shown in this supplementary material.

Appendix C

Proof of Lemma A.1. 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,

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

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

$$\begin{aligned}&\frac{1}{q} \left\| \mathbb{E} m(V_1, \alpha^\intercal X_1, \beta^\intercal \Phi_K(Z_1)) \right\|^2 \\ =& \frac{1}{q} \left\| \mathbb{E}[m(V_1, \alpha^\intercal X_1, \beta^\intercal \Phi_K(Z_1)) - m(V_1, \alpha^\intercal X_1, \beta^\intercal \Phi_K(Z_1) + \gamma_K(Z_1))] \right\|^2\end{aligned}$$

$$\leq \{\mathbb{E}[A(V_1, X_1, Z_1)]|\gamma_K(Z_1)|\}^2 \leq \mathbb{E}[A(V_1, X_1, Z_1)]^2 \mathbb{E}|\gamma_K(Z_1)|^2$$
$$\leq C\|\gamma_K(z)\|^2 = o(1),$$

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

$$\frac{1}{q}\mathbb{E}\|m(V_1, \alpha^\intercal X_1, \beta^\intercal \Phi_K(Z_1))\|^2$$
$$\leq 2\frac{1}{q}\mathbb{E}\|m(V_1, \alpha^\intercal X_1, g(Z_1))\|^2 + 2\frac{1}{q}\mathbb{E}\|m(V_1, \alpha^\intercal X_1, \beta^\intercal \Phi_K(Z_1)) - m(V_1, \alpha^\intercal X_1, g(Z_1))\|^2$$
$$= O(1) + \mathbb{E}[A^2(V_1, X_1, Z_1)|\gamma_K(Z_1)|^2] = O(1)$$

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

2. First, note that

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

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

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

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

$$\inf_{\substack{\|\mathbf{a}\| \leq B_{1n}, \|\mathbf{b}\| \leq B_{2n} \\ \|(\mathbf{a}-\alpha, \mathbf{b}-\beta)\| \geq \delta}} \frac{1}{q}\|\mathbb{E}m(V_i, \mathbf{a}^\intercal X_i, \mathbf{b}^\intercal \Phi_K(Z_i))\|^2$$

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

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

□

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

$$
\begin{aligned}
\tilde{H}_{11}(\alpha,\beta) =& \frac{1}{q}\sum_{\ell=1}^{q}\left(\frac{1}{n}\sum_{j=1}^{n}\frac{\partial}{\partial u}m_\ell(V_j,\alpha^\intercal X_j,\beta^\intercal\Phi_K(Z_j))X_j\right)\\
&\times\left(\frac{1}{n}\sum_{i=1}^{n}\frac{\partial}{\partial u}m_\ell(V_i,\alpha^\intercal X_i,\beta^\intercal\Phi_K(Z_i))X_i\right)^\intercal,\\
\tilde{H}_{12}(\alpha,\beta) =& \frac{1}{q}\sum_{\ell=1}^{q}\left(\frac{1}{n}\sum_{j=1}^{n}\frac{\partial}{\partial u}m_\ell(V_j,\alpha^\intercal X_j,\beta^\intercal\Phi_K(Z_j))X_j\right)\\
&\times\left(\frac{1}{n}\sum_{i=1}^{n}\frac{\partial}{\partial w}m_\ell(V_i,\alpha^\intercal X_i,\beta^\intercal\Phi_K(Z_i))\Phi_K(Z_i)\right)^\intercal,\\
\tilde{H}_{22}(\alpha,\beta) =& \frac{1}{q}\sum_{\ell=1}^{q}\left(\frac{1}{n}\sum_{j=1}^{n}\frac{\partial}{\partial w}m_\ell(V_j,\mathbf{a}^\intercal X_j,\mathbf{b}^\intercal\Phi_K(Z_j))\Phi_K(Z_j)\right)\\
&\left(\frac{1}{n}\sum_{i=1}^{n}\frac{\partial}{\partial w}m_\ell(V_i,\mathbf{a}^\intercal X_i,\mathbf{b}^\intercal\Phi_K(Z_i))\Phi_K(Z_i)\right)^\intercal,
\end{aligned}
$$

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

$$
\begin{aligned}
\Delta_{11}(\alpha,\beta) =& \frac{1}{q}\sum_{\ell=1}^{q}\left(\frac{1}{n}\sum_{i=1}^{n}m_\ell(V_i,\alpha^\intercal X_i,\beta^\intercal\Phi_K(Z_i))\right)\\
&\times\left(\frac{1}{n}\sum_{j=1}^{n}\frac{\partial^2}{\partial u^2}m_\ell(V_j,\alpha^\intercal X_j,\beta^\intercal\Phi_K(Z_j))X_jX_j^\intercal\right),\\
\Delta_{12}(\alpha,\beta) =& \frac{1}{q}\sum_{\ell=1}^{q}\left(\frac{1}{n}\sum_{i=1}^{n}m_\ell(V_i,\alpha^\intercal X_i,\beta^\intercal\Phi_K(Z_i))\right)\\
&\times\left(\frac{1}{n}\sum_{j=1}^{n}\frac{\partial^2}{\partial u\partial w}m_\ell(V_j,\alpha^\intercal X_j,\beta^\intercal\Phi_K(Z_j))X_j\Phi_K(Z_j)^\intercal\right),
\end{aligned}
$$

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

and $\Delta_{21}(\alpha,\beta) = \Delta_{12}(\alpha,\beta)^\intercal$. To fulfil 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

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

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, applying 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 need only to deal with the second factor. Note that

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

$$
\begin{aligned}
\leq & \frac{2}{q}\sum_{\ell=1}^{q}\left\|\mathbb{E}\frac{\partial^2}{\partial u^2}m_\ell(V_1,\alpha^\intercal X_1,\beta^\intercal\Phi_K(Z_1))X_1X_1^\intercal\right\|^2 \\
& + \frac{2}{q}\sum_{\ell=1}^{q}\left\|\frac{1}{n}\sum_{j=1}^{n}\left(\frac{\partial^2}{\partial u^2}m_\ell(V_j,\alpha^\intercal X_j,\beta^\intercal\Phi_K(Z_j))X_jX_j^\intercal\right.\right. \\
& \qquad \left.\left.-\mathbb{E}\frac{\partial^2}{\partial u^2}m_\ell(V_j,\alpha^\intercal X_j,\beta^\intercal\Phi_K(Z_j))X_jX_j^\intercal\right)\right\|^2,
\end{aligned}
$$

where by Assumption 3.5 the first term is $O(p^2)$, while by the iid property for the second we have

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

by Assumption 3.5, 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

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

$$
\begin{aligned}
=&2\frac{1}{q}\left\|\mathbb{E}\frac{\partial^2}{\partial u\partial w}m(V_1,\alpha^\intercal X_1,\beta^\intercal\Phi_K(Z_1))\otimes X_1\Phi_K(Z_1)^\intercal\right\|^2\\
&+2\frac{1}{n}\frac{1}{q}\mathbb{E}\left\|\frac{\partial^2}{\partial u\partial w}m(V_1,\alpha^\intercal X_1,\beta^\intercal\Phi_K(Z_1))\otimes X_1\Phi_K(Z_1)^\intercal\right\|^2\\
=&O(pK)+O(n^{-1}pK),
\end{aligned}
$$

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

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

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 Numbers (LLN, hereafter), I_{11} has the same order in probability as

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

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 3.5. Moreover, by virtue of the i.i.d. property and the LLN, I_{13} has the same order in probability as

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

due to Assumptions 3.5 and 3.7, implying that $\|I_1\|^2=O_P(n^{-1}p^2)+O_P(\|\gamma_K(z)\|^2p^2)=o_P(1)$ by Assumption 3.6.

Now, we consider I_2. Note that

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

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 3.7; meanwhile, by the i.i.d. property,

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

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

by Assumption 3.5. 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 3.6; 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 $\tilde{H}_{22}(\alpha,\beta) - h_{22}(\alpha,\beta)$,

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

$$
\begin{aligned}
&+\frac{1}{q}\sum_{\ell=1}^{q}\left(\mathbb{E}\frac{\partial}{\partial w}m_\ell(V_1,\alpha^\intercal X_1,g(Z_1))\Phi_K(Z_1)\right)\\
&\quad\times\left(\frac{1}{n}\sum_{i=1}^{n}\frac{\partial}{\partial w}m_\ell(V_i,\alpha^\intercal X_i,\beta^\intercal\Phi_K(Z_i))\Phi_K(Z_i)-\mathbb{E}\frac{\partial}{\partial w}m_\ell(V_1,\alpha^\intercal X_1,g)\Phi_K(Z_1)\right)^\intercal\\
:=&I_5+I_6,\quad\text{say.}
\end{aligned}
$$

Using the same approach, we have $\|I_5\|^2 = O_P(n^{-1}K^2) + O_P(\|\gamma_K(z)\|^2K^2) = o_P(1)$ and $\|I_6\|^2 = O_P(n^{-1}K^2) + O_P(\|\gamma_K(z)\|^2K^2) = o_P(1)$ by Assumption 3.6. The whole proof is completed. □

Proof of Lemma A.3. 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

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

Then, using Cauchy-Schwarz inequality gives

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

Observe further that

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

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

by Assumptions 3.1 and 3.3, the dominated convergence theorem and Cauchy-Schwarz inequality. Moreover, it is clear by Assumptions 3.3 and 3.5 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 3.6.

For I_2, by Cauchy-Schwarz inequality again,

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

By virtue of the i.i.d. property and Assumption 3.5,

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

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 3.7 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 3.5. 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

$$\begin{aligned}
&S_{2n}(\alpha,\beta)-s_{2n}(\alpha,g)\\
=&\frac{1}{qn^2}\sum_{\ell=1}^{q}\sum_{i=1}^{n}m_\ell(V_i,\alpha^\intercal X_i,\beta^\intercal\Phi_K(Z_i))\\
&\quad\times\sum_{j=1}^{n}\frac{\partial}{\partial w}m_\ell(V_j,\alpha^\intercal X_j,\beta^\intercal\Phi_K(Z_j))\Phi_K(Z_j)\\
&-\frac{1}{qn}\sum_{\ell=1}^{q}\sum_{i=1}^{n}m_\ell(V_i,\alpha^\intercal X_i,g(Z_i))\mathbb{E}\frac{\partial}{\partial w}m_\ell(V_1,\alpha^\intercal X_1,g(Z_1))\Phi_K(Z_1)\\
=&\frac{1}{qn^2}\sum_{\ell=1}^{q}\sum_{i=1}^{n}[m_\ell(V_i,\alpha^\intercal X_i,\beta^\intercal\Phi_K(Z_i))-m_\ell(V_i,\alpha^\intercal X_i,g(Z_i))]\\
&\quad\times\sum_{j=1}^{n}\frac{\partial}{\partial w}m_\ell(V_j,\alpha^\intercal X_j,\beta^\intercal\Phi_K(Z_j))\Phi_K(Z_j)\\
&+\frac{1}{qn^2}\sum_{\ell=1}^{q}\sum_{i=1}^{n}m_\ell(V_i,\alpha^\intercal X_i,g(Z_i))\\
&\quad\times\sum_{j=1}^{n}\left(\frac{\partial}{\partial w}m_\ell(V_j,\alpha^\intercal X_j,\beta^\intercal\Phi_K(Z_j))-\frac{\partial}{\partial w}m_\ell(V_j,\alpha^\intercal X_j,g(Z_j))\right)\Phi_K(Z_j)\\
&+\frac{1}{qn}\sum_{\ell=1}^{q}\sum_{i=1}^{n}m_\ell(V_i,\alpha^\intercal X_i,g(Z_i))\\
&\quad\times\left(\frac{1}{n}\sum_{j=1}^{n}\frac{\partial}{\partial w}m_\ell(V_j,\alpha^\intercal X_j,g(Z_j))\Phi_K(Z_j)-\mathbb{E}\frac{\partial}{\partial w}m_\ell(V_1,\alpha^\intercal X_1,g(Z_1))\Phi_K(Z_1)\right)\\
:=&I_4+I_5+I_6,\quad\text{say.}
\end{aligned}$$

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 3.7 the second term is the leading one, which by the LLN has the same order as

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

in probability by Assumption 3.6 as $n\to\infty$.

Moreover, invoking Assumptions 3.6-3.7, $I_5=o_P(1)$. Finally,

$$\begin{aligned}
\|I_6\|^2 \leq & \frac{1}{q}\sum_{\ell=1}^{q}\left(\frac{1}{n}\sum_{i=1}^{n}m_\ell(V_i,\alpha^\intercal X_i,g(Z_i))\right)^2 \\
& \times\frac{1}{q}\sum_{\ell=1}^{q}\left\|\frac{1}{n}\sum_{j=1}^{n}\left[\frac{\partial}{\partial w}m_\ell(V_j,\alpha^\intercal X_j,g(Z_j))\Phi_K(Z_j)-\mathbb{E}\frac{\partial}{\partial w}m_\ell(V_i,\alpha^\intercal X_i,g(Z_i))\Phi_K(Z_i)\right]\right\|^2 \\
:= & I_{61}\times I_{62}, \quad \text{say.}
\end{aligned}$$

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

$$\mathbb{E}[I_{62}]=\frac{1}{q}\frac{1}{n^2}\sum_{\ell=1}^{q}\sum_{j=1}^{n}\mathbb{E}\left\|\frac{\partial}{\partial w}m_\ell(V_j,\alpha^\intercal X_j,g(Z_j))\Phi_K(Z_j)-\mathbb{E}\frac{\partial}{\partial w}m_\ell(V_i,\alpha^\intercal X_i,g(Z_i))\Phi_K(Z_i)\right\|^2$$

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

appealing to Assumptions 3.5-3.6, implying $\|I_6\|^2=o_P(n^{-1}K)=o_P(1)$. The proof is complete. □

Proof of Lemmas A.4-A.6. The proof should be the same as that of Lemmas A.1-A.3 but we have to take into account the approximation $\sqrt{n}(\widehat{\theta}-\theta_0)=O_P(1)$. Since $\widehat{\theta}$ is independent of the sample used to estimate the α and g, this is easy but lengthy so omitted. □

Proof of Theorem 3.3. Using Lemmas A.4-A.6, we may prove Theorem 3.3. Due to the same reason as above, the proof is omitted. □

Proof of Lemma A.7. Define $\rho_n=a_n+\sqrt{t_n}P_n'(d_n)$ 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 $P(A_n(\tau))\ge1-\epsilon$ 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

$$
A_1=\{(v_S-v_{0S})^\intercal S_{nT}(v_{0S})\ge-C_1a_n\|v_S-v_{0S}\|\}.
$$

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

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

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, $P(A_3)\ge1-\epsilon/4$ 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 [34] that $\sum_{j\in T}[P_n(|v_{Sj}|) - P_n(|v_{0S,j}|)] \geq -\sqrt{t_n}P_n'(d_n)\|v_S - v_{0S}\|$. Whence, for any $v \in \partial\mathcal{N}_\tau$, on $A_1 \cap A_4$,

$$Q_n(v_T) - Q_n(v_0) \geq \rho_n\tau\left(\frac{3}{8}\rho_n\tau C_2 - C_1 a_n - \sqrt{t_n}P_n'(d_n)\right).$$

For $\rho_n = a_n + \sqrt{t_n}P_n'(d_n)$, $C_1 a_n + \sqrt{t_n}P_n'(d_n) \leq (C_1+1)\rho_n$. 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 large n, with $\tau > 8(C_1+1)/3C_2$, $P(A_n(\tau)) > P(A_1 \cap A_4) \geq 1-\epsilon$.

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_n-vector consisting of nonzero elements of $\widehat{v}$, and c is the same in (iv) of Assumption 5.1, we have

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

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 $u \in \mathbb{R}^{t_n}$ 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 A.8. 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 (A.4),

$$Q_n(v_T) - Q_n(v) = \|M_n(v_T)\|^2 - \|M_n(v)\|^2 - \sum_{j\notin T}P_n(|v_j|) < 0. \tag{C.1}$$

This means $Q_n(\widehat{v}) < Q_n(v)$, yielding the first assertion, while, from which and the last statement of Lemma A.7, 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, $\|A_n(v_{0S})\| = O_P(\sqrt{t_n})$. 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

$$\|M_n(v_0)\| = \left\| \frac{1}{\sqrt{q}n} \sum_{i=1}^{n} m(V_i, \alpha_{0S}^\intercal X_{iS}, \beta_{0S}^\intercal \Phi_{KS}(Z_i)) \right\| = O_P\left(\sqrt{\log(q)/n}\right). \tag{C.2}$$

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

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 [34].

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 A.8: 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 (A.4) holds, that is, $\|M_n(v_T)\|^2 - \|M_n(v)\|^2 < \sum_{j \notin T} P_n(|v_j|)$. This is equivalent 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. $\{1, \ldots, p+K\}$ and noting $v = v_T + v_{T^c}$ for any v.

Here, we know $\|S_n(v_{0S})\| = O_P(\sqrt{t_n \log(q)/n})$, $\|\widehat{v} - v_0\| = O_P(\sqrt{t_n \log(q)/n} + \sqrt{t_n} P_n'(d_n))$. In a small neighbourhood of $\widehat{v}$, $O(\widehat{v}, r_n/(p+K))$ say, where r_n is a sufficient small number, $\|S_n(v)\| = O_P(\sqrt{t_n \log(q)/n})$ uniformly holds in v and $\sup_{v \in O} \|v - \widehat{v}\|_1 \leq r_n$.

On the other hand, for some $\mu \in (0,1)$,

$$\sum_{j \notin T} P_n(|v_j|) = \sum_{j \notin T, v_j \neq 0} |v_j| P_n'(\mu |v_j|) \geq \sum_{j \notin T, v_j \neq 0} |v_j| P_n'(r_n)$$

by the nonincreasingness of $P_n'(u)$. Let r_n so small that $P_n'(r_n) \geq P_n'(0+)/2$. Hence, $\sum_{j \notin S} P_n(|\beta_j|) \geq C r_n$ in probability.

Then, by virtue of Assumption 5.4 and following a similar argument as Fan and Liao [34], the condition is verified.

Appendix D

The estimates of some important coefficients in Section 7 are reported in this section.

Table 8: Estimated coefficients for Subsample H

Mother's Years of Schooling	0.1349
Number of Siblings	0.0215
Urban Residence at 14	0.2936
"Permanent" Local Log Earnings at 17	-0.0263
"Permanent" State Unemployment Rate at 17	-0.0745
Instruments (W):	
Local Log Earnings at 17	0.2531
State Unemployment Rate at 17	0.0097
Tuition in 4 Year Public Colleges at 17	-0.0006

Table 9: Estimated coefficients for Subsample C

Mother's Years of Schooling	0.0030
Number of Siblings	-0.0190
Urban Residence at 14	-0.0472
"Permanent" Local Log Earnings at 17	-0.0045
"Permanent" State Unemployment Rate at 17	-0.0205
Instruments (W):	
Local Log Earnings at 17	0.2092
State Unemployment Rate at 17	0.0244
Tuition in 4 Year Public Colleges at 17	-0.0075