Local Linear GMM Estimation of Functional Coefficient IV Models with Application to the Estimation of Rate of Return to Schooling*

Liangjun Su,[a] Irina Murtazashvili,[b] Aman Ullah[c]

[a] School of Economics, Singapore Management University, Singapore

[b] Department of Economics, University of Pittsburgh, PA, USA

[c] Department of Economics, University of California at Riverside, CA, USA

June 2, 2011

Abstract

We consider the local linear GMM estimation of functional coefficient models with a mix of discrete and continuous data and in the presence of endogenous regressors. We establish the asymptotic normality of the estimator and propose a nonparametric test for the constancy of the varying coefficients. Simulations are conducted to evaluate both the estimator and test. Applications to the 1985 Australian Longitudinal Survey data indicate a clear rejection of the null hypothesis of the constant rate of return to education, and that the returns to education obtained in earlier studies tend to be overestimated for all the work experience.

JEL Classifications: C12, C13, C14

Key Words: Discrete variables; Endogeneity; Functional coefficient; Heterogeneity; Instrumental variable (IV); Local linear GMM estimation; Schooling.

*Correspondence: Liangjun Su, School of Economics, Singapore Management University, 90 Stamford Road, Singapore, 178903, E-mail: ljsu@smu.edu.sg; Phone: +65 6828- 0386. Irina Murtazashvili, Dept of Economics, University of Pittsburgh, PA, 15260; E-mail: irinam@pitt.edu; Phone: 412-648-1762. Aman Ullah, Department of Economics, UCR, Riverside, CA 92521-0427; E-mail: aman.ullah@ucr.edu; Phone: 951-827-1591. The third author gratefully acknowledges the financial support from the Academic Senate, UCR.

1 Introduction

In classical econometrics literature, in both applied and theory, econometric models are often studied in a linear parametric regression form with its coefficients (derivatives or marginal changes) assumed to be constant over time or across cross section units. In practice this may not be true, e.g., it may be hard to believe that the marginal propensity to save or to consume would be the same for a younger as for an older group of individuals in a given cross section data set, or that the elasticity of wages with respect to schooling or the returns to schooling would be the same for individuals with less experience compared to those with more experience. In the case of nonlinear parametric regression models, the coefficients have been taken as constant but derivatives do vary depending on the specification of models, e.g., the translog production function has constant coefficients, and the elasticities (derivatives) based on this function vary linearly with inputs. Realizing the fact that some or all of the coefficients in a regression may be varying, the traditional econometrics literature has tried to consider various forms of the parametric specifications of the varying coefficients. See, e.g., the papers of Hildreth and Houck (1968), Swamy (1970), Singh and Ullah (1974), and Granger and Teräsvirta (1999), and the books by Swamy (1971), Raj and Ullah (1981), and Granger and Teräsvirta (1993). However, it is now well known that the specifications of the constant or parametric varying coefficient models may often be misspecified, and therefore this may lead to inconsistent estimation and testing procedures and hence misleading empirical analysis and policy evaluations.

In view of the above issues, in recent years, the nonparametric varying/functional coefficient models have been considered by various authors, including Cleveland, Grosse, and Shyu (1992), Chen and Tsay (1993), Hastie and Tibshirani (1993), Fan and Zhang (1999), and Cai, Fan, and Yao (2000), among others. The coefficients in these models have an unknown function of the observed variables which vary with the data, and the estimation of such models have been carried out by using nonparametric procedures; see Pagan and Ullah (1999) and Racine and Li (2007) for details on such procedures. An additional advantage of the functional coefficient model is that it also considers the unknown functional form of the interacting variables which in empirical parametric models is often misspecified to be linear. Most of the above works on functional coefficient models are focused on models with exogenous regressors. Recently Das (2005), Cai, Das, Xiong, and Wu (2006, CDXW hereafter), Lewbel (2007), Cai and Li (2008), Tran and Tsionas (2010), and Su (2011) among others, have considered the semiparametric models with endogenous variables and they suggest a nonparametric/semiparametric generalized method of moments (GMM) instrumental variable (IV) approach to estimate them. In particular, CDXW (2006), Cai and Li (2008), and Tran and Tsionas (2010) focus on functional coefficient models with endogenous regressors.

CDXW (2006) propose a two-stage local linear estimation procedure to estimate the functional coefficient models, which unfortunately requires one to first estimate a high-dimension nonparametric model and then to estimate the functional coefficients using the first-stage nonparametric estimates as generated regressors. In contrast, Cai and Li (2008) suggest a one-step local linear GMM estimator which corresponds to our local linear GMM estimator with an identity weight matrix. Tran and Tsionas (2010) provide a local constant two-step GMM estimator with a specified weighting matrix that can be chosen to minimize the asymptotic variances in the class of GMM estimators. However, the local constant estimation procedure, as is now well known, is less desirable than the local linear estimation procedure, especially at the boundaries. In addition, all of these papers consider varying coefficients with continuous variables only. On the other hand, Su, Chen, and Ullah (2009, SCU hereafter) consider both continuous and categorical variables in functional coefficient regression models and show that the consideration for the categorical variables is extremely important for empirical analysis, and it improves on the specifications of the traditional linear parametric dummy-variable models. But they do no consider the endogeneity issue which prevails in economics.

In addition, in the estimation context, the advantage of using the traditional constant coefficient models rest on their validity. Nevertheless, to the best of our knowledge, there is no nonparametric hypothesis testing procedure available for this when endogeneity is present, although there are some tests (e.g., Fan, Zhang, and Zhang (2001) and Hong and Lee (2009)) in the absence of endogeneity. In view of the above deficiencies in the existing literature we first focus on further improvement in the estimation area, and then provide a consistent test for the constancy of functional coefficient models. If we fail to reject the null of constancy, then we can continue to rely on the traditional constant coefficient models. Otherwise we may have to consider the functional coefficients with unknown form.

In this paper, we develop local linear GMM estimation of functional coefficient IV models with a general weight matrix. A varying coefficient model is considered in which some or all the regressors are endogenous and their coefficients are varying with respect to continuous and/or categorical variables. An optimal nonparametric GMM estimator is proposed where the weight matrix is obtained by minimizing the asymptotic variance of the GMM estimator. The consistency and asymptotic normality of the nonparametric GMM estimators are established. Then we develop a new test statistic for testing the hypothesis that the functional coefficients are constant. It is argued that the test based on the Lagrangian multiplier (LM) principle needs restricted estimation and may suffer from the curse of dimensionality, and similarly the test using the likelihood ratio (LR) method also requires both unrestricted and involved restricted estimation. For these reasons a simpler Wald type of test is proposed which is based on the unrestricted estimation. The consistency, asymptotic null distribution, and asymptotic local power of the proposed test are established. It is well known that nonparametric tests based on the critical values of their asymptotic normal distributions may perform poorly in finite samples. In view of this, we also provide a bootstrap procedure to approximate the asymptotic null distribution of our test statistic. To assess the finite sample properties of the proposed local linear GMM estimator and the new test statistic, we conduct a small set of simulations. The results show that our local linear GMM estimator may not perform as well as the estimator of SCU in the absence of endogeneity, but when endogeneity is present, our estimator dominates the SCU's estimator in both mean squared error (MSE) and mean absolute deviation (MAD). Also, the simulations show that our test performs well in finite samples.

Another important objective of this paper is to employ our proposed nonparametric GMM estimator to study the empirical relationship between earnings and education using the 1985 Australian Longitudinal Survey. Labor economists have long studied two major problems arising when estimating the wage equation: endogeneity of education and heterogeneity of returns to education, see Card (2001) for detailed stimulating discussions. Our nonparametric estimator is able to deal with both problems in a flexible way. Specifically, in contrast to other existing estimators, our estimator allows the returns to education to depend on both continuous (experience) and discrete (marital status, union membership, etc.) characteristics of individuals while controlling for endogeneity of education. Further, we use our proposed new nonparametric test to check for constancy of functional coefficients in the wage equation.

Our findings are unambiguous: the returns to education do depend on both, experience and the categorical variables we use, in a non-linear manner. Additionally, we find that the returns to education tend to be overestimated for all of the observed work experience when the categorical explanatory variables are not accounted for in functional coefficients as in CDXW (2006) and Cai and Xiong (2010). These results are also important since our proposed tests show the absence of the constancy of the return to education, which is often assumed in most of the parametric empirical studies in labor economics.

The paper is structured as follows. In Section 2 we introduce our functional coefficient IV model and propose a local linear GMM procedure to estimate the functional coefficients and their first order derivatives. The asymptotic properties of these estimators are studied in Section 3. We propose a specification test for our model in Section 4. We conduct a small set of Monte Carlo studies to check the relative performance of the proposed estimator and test in Section 5. Section 6 provides empirical data analysis. Final remarks are contained in Section 7. All technical details are relegated to the Appendix.

For natural numbers a and b, we use I_a to denote an $a \times a$ identity matrix, and $\mathbf{0}_{a\times b}$ an $a \times b$ matrix of zeros. Let $\otimes$ and $\odot$ denote the Kronecker and Hadamard products, respectively. If $\mathbf{c}$ and $\mathbf{d}$ are vectors of the same dimension, $\mathbf{c}/\mathbf{d}$ denotes the vector of elementwise divisions. For a matrix $\mathbf{M}$, $\mathbf{M}'$ means the transpose of $\mathbf{M}$, and $\|\mathbf{M}\| = \sqrt{\mathrm{tr}\left(\mathbf{MM}'\right)}$. We use $1\{\cdot\}$ to denote the usual indicator function which takes value 1 if the condition inside the curly bracket holds and 0 otherwise, and C to signify a generic constant whose exact value may vary from case to case. We use $\overset{d}{\to}$ and $\overset{P}{\to}$ to denote convergence in distribution and probability, respectively.

2 Functional Coefficient Estimation with Mixed Data and Estimated Covariate

In this section we first introduce a functional coefficient IV model where the coefficient function may depend on both continuous and discrete exogenous regressors and the endogenous regressors enter the model only linearly. Then we propose local linear estimates for the functional coefficients.

2.1 Functional coefficient representation

We consider the following functional coefficient IV model

$$\begin{cases} Y_i = \mathbf{g}\left(\mathbf{U}_i^c, \mathbf{U}_i^d\right)' \mathbf{X}_i + \varepsilon_i = \sum_{j=1}^d g_j\left(\mathbf{U}_i^c, \mathbf{U}_i^d\right) X_{i,j} + \varepsilon_i \\ E\left(\varepsilon_i | \mathbf{Z}_i, \mathbf{U}_i\right) = 0 \text{ almost surely (a.s.)} \end{cases} \tag{2.1}$$

where for $i = 1, \cdots, n$, Y_i is an observed scalar random variable, $\mathbf{g} = (g_1, \cdots, g_d)'$, $\{g_j\}_{j=1}^d$ are the unknown structural functions of interest, $X_{i,1} = 1$, $\mathbf{X}_i = (X_{i,1}, \cdots, X_{i,d})'$ is a $d \times 1$ vector consisting of $d-1$ endogenous regressors, $\mathbf{U}_i = (\mathbf{U}_i^{c\prime}, \mathbf{U}_i^{d\prime})'$, $\mathbf{U}_i^c$ and $\mathbf{U}_i^d$ denote a $p_c \times 1$ vector of continuous exogenous regressors and a $p_d \times 1$ vector of discrete exogenous regressors, respectively, and $\mathbf{Z}_i$ is a $q_z \times 1$ vector of instrumental variables. Let $p = p_c + p_d$.

In the absence of $\mathbf{U}_i^d$, (2.1) reduces to the model of CDXW (2006). If none of the variables in $\mathbf{X}_i$ is endogenous, the model becomes that of SCU (2009). As the latter authors demonstrate through the estimation of earnings function, it is important to allow the variables in the functional coefficients to include both continuous and discrete variables, where the discrete variables may represent race, profession, or region, etc. In practice, it may be also important to allow some of the continuous variables in the functional coefficient to be unobservable, which complicates the asymptotic analysis. When some variables in $\mathbf{U}_i^c$ are not observed and estimated from the data, we conjecture that the theory developed in this paper continue to hold as long as the estimates of these regressors converge to their truth at sufficiently fast rate. For simplicity, we focus on the case where $Y_i, \mathbf{X}_i, \mathbf{Z}_i, \mathbf{U}_i$, are all observed.

2.2 Local linear GMM estimation

The orthogonality condition

$$E\left(\varepsilon_i | \mathbf{Z}_i, \mathbf{U}_i\right) = 0 \text{ a.s.} \tag{2.2}$$

suggests that we can estimate the unknown functional coefficients via the principle of nonparametric generalized method of moments (NPGMM), which is similar to the GMM of Hansen (1982) for parametric models. Let $\mathbf{V}_i = \left(\mathbf{Z}_i', \mathbf{U}_i\right)'$. (2.2) indicates that for any $k \times 1$ vector function $\mathbf{Q}\left(\mathbf{V}_i\right)$, we have

$$E\left[\mathbf{Q}\left(\mathbf{V}_i\right) \varepsilon_i | \mathbf{V}_i\right] = E\left[\mathbf{Q}\left(\mathbf{V}_i\right)\left\{Y_i - \sum_{j=1}^{d} g_j\left(\mathbf{U}_i^c, \mathbf{U}_i^d\right) X_{i,j}\right\} | \mathbf{U}_i\right] = 0 \tag{2.3}$$

by the law of iterated expectations. Following Cai and Li (2008), we propose an estimation procedure to combine the orthogonality conditions in (2.3) with the idea of local linear fitting in the nonparametrics literature to estimate the unknown functional coefficients.

Like Racine and Li (2004), we use $\mathbf{U}_{i,t}^d$ to denote the tth component of $\mathbf{U}_i^d$. $\mathbf{U}_{i,t}^c$ is similarly defined. Analogously, we let u_t^d and u_t^c to denote the tth component of $\mathbf{u}^d$ and $\mathbf{u}^c$, respectively, i.e., $\mathbf{u}^d = \left(u_1^d, \cdots, u_{p_d}^d\right)'$ and $\mathbf{u}^c = \left(u_1^c, \cdots, u_{p_c}^c\right)'$. We assume that $\mathbf{U}_{i,t}^d$ can take $c_t \geq 2$ different values, i.e., $U_{i,t}^d \in \{0, 1, \cdots, c_t - 1\}$ for $t = 1, \cdots, p_d$. Let $\mathbf{u} = \left(\mathbf{u}^c, \mathbf{u}^d\right) \in \mathbb{R}^{p_c} \times \mathbb{R}^{p_d}$. We use $f_{\mathbf{U}}\left(\mathbf{u}\right) = f_{\mathbf{U}}\left(\mathbf{u}^c, \mathbf{u}^d\right)$ to denote the joint density function of $\mathbf{U}_i^c$ and $\mathbf{U}_i^d$ and $\mathcal{D} = \Pi_{t=1}^{p_d}\{0, 1, \cdots, c_t - 1\}$ to denote the range assumed by $\mathbf{U}_i^d$. To define the kernel weight function, we focus on the case for which there is no natural ordering in $\mathbf{U}_i^d$. Define

$$l\left(U_{i,t}^d, u_t^d, \lambda_t\right) = \begin{cases} 1 & \text{if } \mathbf{U}_{i,t}^d = u_t^d, \\ \lambda_t & \text{if } \mathbf{U}_{i,t}^d \neq u_t^d, \end{cases} \tag{2.4}$$

where λ_t is a bandwidth that lies on the interval $[0, 1]$. Clearly, when $\lambda_t = 0$, $l\left(U_{i,t}^d, u_t^d, 0\right)$ becomes an indicator function, and $\lambda_t = 1$, $l\left(U_{i,t}^d, u_t^d, 1\right)$ becomes an uniform weight function. We define the product kernel for the discrete random variables by

$$L\left(\mathbf{U}_i^d, \mathbf{u}^d, \boldsymbol{\lambda}\right) = L_{\boldsymbol{\lambda}}\left(\mathbf{U}_i^d - \mathbf{u}^d\right) = \prod_{t=1}^{p_d} l\left(U_{i,t}^d, u_t^d, \lambda_t\right). \tag{2.5}$$

For the continuous random variables, we use $w\left(\cdot\right)$ to denote a univariate kernel function and define the product kernel function by $W_{\mathbf{h},i\mathbf{u}^c} = W_{\mathbf{h}}\left(\mathbf{U}_i^c - \mathbf{u}^c\right) = \Pi_{t=1}^{p_c} h_t^{-1} w\left(\left(U_{i,t}^c - u_t^c\right)/h_t\right)$, where $\mathbf{h} = \left(h_1, \cdots, h_{p_c}\right)$ denotes the p_c-vector of smoothing parameters. We then define the kernel weight function $K_{\mathbf{h}\boldsymbol{\lambda},i\mathbf{u}}$ by

$$K_{\mathbf{h}\boldsymbol{\lambda},i\mathbf{u}} = W_{\mathbf{h},i\mathbf{u}^c} L_{\boldsymbol{\lambda},i\mathbf{u}^d} \tag{2.6}$$

where $L_{\boldsymbol{\lambda},i\mathbf{u}^d} = L\left(\mathbf{U}_i^d, \mathbf{u}^d, \boldsymbol{\lambda}\right)$.

To estimate the unknown functional coefficients in model (2.1) via the local linear regression technique, we assume that $\left\{g_j\left(\mathbf{u}^c, \mathbf{u}^d\right), j=1, \cdots, d\right\}$ are twice continuously differentiable. Denote by $\dot{g}_j\left(\mathbf{u}^c, \mathbf{u}^d\right) = \partial g_j\left(\mathbf{u}^c, \mathbf{u}^d\right)/\partial \mathbf{u}^c$ the $p_c \times 1$ vector of first order derivatives of g_j with respect to its continuous-valued argument $\mathbf{u}^c$. Denote by $\ddot{g}_j\left(\mathbf{u}^c, \mathbf{u}^d\right) = \partial^2 g_j\left(\mathbf{u}^c, \mathbf{u}^d\right)/\left(\partial \mathbf{u}^c \partial \mathbf{u}^c\right)$ the $p_c \times p_c$ matrix of second order derivatives of g_j with respect to $\mathbf{u}^c$. We use $g_{j,ss}\left(\mathbf{u}^c, \mathbf{u}^d\right)$ to denote the sth diagonal element of $\ddot{g}_j\left(\mathbf{u}^c, \mathbf{u}^d\right)$. For any given $\mathbf{u}^c$ and $\mathbf{U}_i^c$ in a neighborhood of $\mathbf{u}^c$, it follows from a first order Taylor expansion that

$$g_j\left(\mathbf{U}_i^c, \mathbf{u}^d\right) \approx g_j\left(\mathbf{u}^c, \mathbf{u}^d\right) + \dot{g}_j\left(\mathbf{u}^c, \mathbf{u}^d\right)'\left(\mathbf{U}_i^c - \mathbf{u}^c\right). \tag{2.7}$$

Let $\mathbf{v} = (\mathbf{z}, \mathbf{u})$. We can approximate $E\left[\mathbf{Q}\left(\mathbf{V}_i\right)\left\{Y_i - \sum_{j=1}^d g_j\left(\mathbf{U}_i^c, \mathbf{U}_i^d\right)X_{i,j}\right\}|\mathbf{U}_i = \mathbf{u}\right]$ by its sample analogue

$$\begin{aligned}
&\frac{1}{n}\sum_{i=1}^n \mathbf{Q}\left(\mathbf{V}_i\right)\left\{Y_i - \sum_{j=1}^d \left[g_j\left(\mathbf{u}^c, \mathbf{u}^d\right) + \dot{g}_j\left(\mathbf{u}^c, \mathbf{u}^d\right)'\left(\mathbf{U}_i^c - \mathbf{u}^c\right)\right]X_{i,j}\right\}K_{\mathbf{h}\boldsymbol{\lambda},i\mathbf{u}} \\
= \;& \frac{1}{n}\mathbf{Q}_{\mathbf{h}}\left(\mathbf{u}\right)'\mathbf{K}_{\mathbf{h}\boldsymbol{\lambda}}\left(\mathbf{u}\right)\left[\mathbf{Y} - \boldsymbol{\xi}\left(\mathbf{u}\right)\boldsymbol{\alpha}\right]
\end{aligned} \tag{2.8}$$

where $\mathbf{Y} = \left(Y_1, \cdots, Y_n\right)'$, $\boldsymbol{\xi}\left(\mathbf{u}\right) = \left(\boldsymbol{\xi}_{1,\mathbf{u}}, \cdots, \boldsymbol{\xi}_{n,\mathbf{u}}\right)'$, $\boldsymbol{\xi}_{i,\mathbf{u}} = \begin{pmatrix} \mathbf{X}_i \\ \mathbf{X}_i \otimes \left(\mathbf{U}_i^c - \mathbf{u}^c\right)\end{pmatrix}$ and $\boldsymbol{\alpha} = (g_1, \cdots, g_d, \dot{g}_1', \cdots, \dot{g}_d')'$ are both $d\left(p_c+1\right) \times 1$ vectors, $\mathbf{K}_{\mathbf{h}\boldsymbol{\lambda}}\left(\mathbf{u}\right) = \mathrm{diag}(K_{\mathbf{h}\boldsymbol{\lambda},1\mathbf{u}}, \cdots, K_{\mathbf{h}\boldsymbol{\lambda},n\mathbf{u}})$, $\mathbf{Q}_{\mathbf{h}}\left(\mathbf{u}\right) = (\mathbf{Q}_{\mathbf{h},1\mathbf{u}}, \cdots, \mathbf{Q}_{\mathbf{h},n\mathbf{u}})'$ and $\mathbf{Q}_{\mathbf{h},i\mathbf{u}} = \mathbf{Q}\left(\mathbf{V}_i\right)$. To obtain estimates of g_j and $\dot{g}_j$, we can choose $\boldsymbol{\alpha}$ to minimize the following local linear GMM criterion function

$$\frac{1}{n}\left[\mathbf{Q}_{\mathbf{h}}\left(\mathbf{u}\right)'\mathbf{K}_{\mathbf{h}\boldsymbol{\lambda}}\left(\mathbf{u}\right)\left(\mathbf{Y} - \boldsymbol{\xi}\left(\mathbf{u}\right)\boldsymbol{\alpha}\right)\right]'\boldsymbol{\Psi}_n\left(\mathbf{u}\right)^{-1}\left[\mathbf{Q}_{\mathbf{h}}\left(\mathbf{u}\right)'\mathbf{K}_{\mathbf{h}\boldsymbol{\lambda}}\left(\mathbf{u}\right)\left(\mathbf{Y} - \boldsymbol{\xi}\left(\mathbf{u}\right)\boldsymbol{\alpha}\right)\right], \tag{2.9}$$

where $\boldsymbol{\Psi}_n\left(\mathbf{u}\right)$ is a symmetric $k \times k$ weight matrix that is positive definite for large n. Clearly, the solution to the above minimization problem is given by

$$\begin{aligned}
\widehat{\boldsymbol{\alpha}}_{\boldsymbol{\Psi}_n}\left(\mathbf{u}\right) \;=\;& \left[\boldsymbol{\xi}\left(\mathbf{u}\right)'\mathbf{K}_{\mathbf{h}\boldsymbol{\lambda}}\left(\mathbf{u}\right)\mathbf{Q}_{\mathbf{h}}\left(\mathbf{u}\right)\boldsymbol{\Psi}_{\mathbf{n}}\left(\mathbf{u}\right)^{-\mathbf{1}}\mathbf{Q}_{\mathbf{h}}\left(\mathbf{u}\right)'\mathbf{K}_{\mathbf{h}\boldsymbol{\lambda}}\left(\mathbf{u}\right)\boldsymbol{\xi}\left(\mathbf{u}\right)\right]^{-1} \\
&\times\boldsymbol{\xi}\left(\mathbf{u}\right)'\mathbf{K}_{\mathbf{h}\boldsymbol{\lambda}}\left(\mathbf{u}\right)\mathbf{Q}_{\mathbf{h}}\left(\mathbf{u}\right)\boldsymbol{\Psi}_{\mathbf{n}}\left(\mathbf{u}\right)^{-\mathbf{1}}\mathbf{Q}_{\mathbf{h}}\left(\mathbf{u}\right)'\mathbf{K}_{\mathbf{h}\boldsymbol{\lambda}}\left(\mathbf{u}\right)\mathbf{Y}.
\end{aligned} \tag{2.10}$$

Let $\mathbf{e}_{j,d(1+p_c)}$ denote the $d\left(1+p_c\right) \times 1$ unit vector with 1 at the jth position and 0 elsewhere. Let $\widetilde{\mathbf{e}}_{j,p_c,d(1+p_c)}$ denote the $p_c \times d\left(1+p_c\right)$ selection matrix such that $\widetilde{\mathbf{e}}_{j,p_c,d(1+p_c)}\boldsymbol{\alpha} = \dot{g}_j\left(\mathbf{u}\right)$. Then the local linear GMM estimator of $g_j\left(\mathbf{u}\right)$ and $\dot{g}_j\left(\mathbf{u}\right)$ are respectively given by

$$\widehat{g}_j\left(\mathbf{u}\right) = \mathbf{e}_{j,d(1+p_c)}'\widehat{\boldsymbol{\alpha}}_{\boldsymbol{\Psi}_n}\left(\mathbf{u}\right) \text{ and } \widehat{\dot{g}}_j\left(\mathbf{u}\right) = \widetilde{\mathbf{e}}_{j,p_c,d(1+p_c)}\widehat{\boldsymbol{\alpha}}_{\boldsymbol{\Psi}_n}\left(\mathbf{u}\right) \text{ for } j = 1, \cdots, d. \tag{2.11}$$

We will study the asymptotic properties of $\widehat{\boldsymbol{\alpha}}_{\boldsymbol{\Psi}_n}\left(\mathbf{u}\right)$ in the next section.

Remark 1 (Choice of instruments) The choice of $\mathbf{Q}\left(\mathbf{V}_i\right)$ is important in establishing our asymptotic theory. Motivated by the idea of local linear fitting, a simple choice of $\mathbf{Q}\left(\mathbf{V}_i\right)$ is

$$\mathbf{Q}_{\mathbf{h},i\mathbf{u}} = \mathbf{Q}\left(\mathbf{V}_i\right) = \begin{pmatrix} \mathbf{Z}_i^a \\ \mathbf{Z}_i^a \otimes \left(\mathbf{U}_i^c - \mathbf{u}^c\right)/\mathbf{h}\end{pmatrix}, \tag{2.12}$$

where $\mathbf{Z}_i^a$ is a $q \times 1$ vector of "global" instruments so that the dimension of $\mathbf{Q}_{\mathbf{h},i\mathbf{u}}$ is $k \times 1$ with $k = q\left(p_c+1\right)$. Note that we allow the "local" instrument $\mathbf{Q}_{\mathbf{h},i\mathbf{u}}$ to depend on both the continuous point of evaluation $\mathbf{u}^c$ and the bandwidth parameter for the continuous regressor $(\mathbf{U}_i^c)$ in the functional coefficient. The global instrument may be chosen from the union of $\mathbf{Z}_i$ and $\mathbf{U}_i^d$ such that certain identification condition is satisfied. In particular, in order for the local linear GMM estimator $\widehat{\boldsymbol{\alpha}}$ to be well defined, a necessary condition is $q \geq d$, which ensures that the dimension of $\mathbf{Q}_{\mathbf{h},i\mathbf{u}}$ is not smaller than the dimension of $\boldsymbol{\alpha}$. Note that the choice of $\mathbf{Q}(\mathbf{V}_i)$ in (2.12) is computationally simple but it may not be optimal in the sense of minimizing the asymptotic variance for the class of local linear GMM estimators given the orthogonal condition in (2.2). It is not clear how to construct the optimal instruments by extending the work of Newey (1990) and Ai and Chen (2003) to our framework. Therefore in this paper we focus only on the simple case where $\mathbf{Q}(\mathbf{V}_i)$ has the form given in (2.12).

Remark 2 (Local linear versus local constant GMM estimators) An alternative to the local linear GMM estimator is the local constant GMM estimator, see, for example, Lewbel (2007) and Tran and Tsionas (2010). In this case, the parameter of interest $\boldsymbol{\alpha}$ contains only the set of functional coefficients g_j, $j = 1, \cdots, d$, evaluated at $\mathbf{u} = \left(\mathbf{u}^{c\prime}, \mathbf{u}^{d\prime}\right)'$, but not their first order derivatives with respect to the continuous arguments, and a simple choice of $\mathbf{Q}(\mathbf{V}_i)$ is

$$\mathbf{Q}_{\mathbf{h},i\mathbf{u}} = \mathbf{Q}(\mathbf{V}_i) = \mathbf{Z}_i^a, \tag{2.13}$$

where $\mathbf{Z}_i^a$ is defined as above. In this case, $k = q$ and one also requires $q \geq d$ for the local identification of the functional coefficients. In addition, our local linear GMM estimator in (2.10) reduces to that of Cai and Li (2008) by setting $\boldsymbol{\Psi}_{\mathbf{n}}(\mathbf{u})$ to be the identity matrix and choosing $q = d$ global instruments. The latter condition is necessary for the model to be locally *just identified*.

3 Asymptotic Properties of the Local Linear GMM Estimator

In this section, we first give a set of assumptions and then study the asymptotic properties of the local linear GMM estimator.

3.1 Assumptions

To facilitate the presentation, define

$$\boldsymbol{\Omega}(\mathbf{u}) = E\left(\mathbf{Z}_i^a \mathbf{X}_i' | \mathbf{U}_i = \mathbf{u}\right) \text{ and } \boldsymbol{\Omega}^*(\mathbf{u}) = E\left[\mathbf{Z}_i^a \mathbf{Z}_i^{a\prime} \sigma^2\left(\mathbf{Z}_i^a, \mathbf{U}_i\right) | \mathbf{U}_i = \mathbf{u}\right]$$

where $\sigma^2(\mathbf{z}^a, \mathbf{u}) = E\left[\varepsilon_i^2 | \mathbf{Z}_i^a = \mathbf{z}^a, \mathbf{U}_i = \mathbf{u}\right]$. Let $f_{\mathbf{U}}(\mathbf{u}) = f_{\mathbf{U}}\left(\mathbf{u}^c, \mathbf{u}^d\right)$ denote the joint density of $\mathbf{U}_i^c$ and $\mathbf{U}_i^d$ and $p\left(\mathbf{u}^d\right)$ be the marginal probability mass of $\mathbf{U}_i^d$ at $\mathbf{u}^d$.

We now list the assumptions that will be used to establish the asymptotic distribution of our estimator.

Assumption A1. $(Y_i, \mathbf{X}_i, \mathbf{Z}_i, \mathbf{U}_i), i = 1, \cdots, n$, are independent and identically distributed (IID).

Assumption A2. $E|\varepsilon_i|^{2+\delta} < \infty$ for some $\delta > 0$. $E\|\mathbf{Z}_i^a \mathbf{X}_i'\|^2 < \infty$.

Assumption A3. (i) The functions $f_{\mathbf{U}}\left(\cdot, \widetilde{\mathbf{u}}^d\right)$, $\boldsymbol{\Omega}\left(\cdot, \widetilde{\mathbf{u}}^d\right)$, and $\boldsymbol{\Omega}^*\left(\cdot, \widetilde{\mathbf{u}}^d\right)$ are continuously differentiable for all $\widetilde{\mathbf{u}}^d \in \mathcal{D}$. $0 < f_{\mathbf{U}}\left(\mathbf{u}^c, \mathbf{u}^d\right) \leq C$ for some $C < \infty$. (ii) The functions $g_j\left(\cdot, \widetilde{\mathbf{u}}^d\right)$, $j = 1, \cdots, d$, are second order continuously differentiable for all $\widetilde{\mathbf{u}}^d \in \mathcal{D}$.

Assumption A4. (i) rank$(\boldsymbol{\Omega}(\mathbf{u})) = d \leq q$, and the $q \times q$ matrix $\boldsymbol{\Omega}^*(\mathbf{u})$ is positive definite. (ii) $\boldsymbol{\Psi}_n(\mathbf{u}) = \boldsymbol{\Psi}(\mathbf{u}) + o_P(1)$, where $\boldsymbol{\Psi}(\mathbf{u})$ is symmetric and positive definite.

Assumption A5. The kernel function $w(\cdot)$ is a probability density function that is symmetric, bounded, and compactly supported.

Assumption A6. As $n \to 0$, the bandwidth sequences $\mathbf{h} = (h_1, \cdots, h_{p_c})'$ and $\boldsymbol{\lambda} = (\lambda_1, \cdots, \lambda_{p_d})'$ satisfy (i) $n\mathbf{h}! \to \infty$, and (ii) $(n\mathbf{h}!)^{1/2}(\|\mathbf{h}\|^2 + \|\boldsymbol{\lambda}\|) = O(1)$, where $\mathbf{h}! = h_1 \cdots h_{p_c}$.

A1 requires IID observations. Following Cai and Li (2008) and SCU (2009), this assumption can be relaxed to allow for time series observations. A2 and A3 impose some moment and smoothness conditions, respectively. A4(i) imposes rank conditions for the identification of the functional coefficients and their first order derivatives and A4(ii) is weak in that it allows the random weight matrix $\boldsymbol{\Psi}_n$ to be consistently estimated from the data. As Hall, Wolf, and Yao (1999) remark, the requirement in A5 that $w(\cdot)$ is compactly supported can be removed at the cost of lengthier arguments used in the proofs, and in particular, Gaussian kernel is allowed. A6 is standard for nonparametric regression with mixed data; see, e.g., Li and Racine (2008).

3.2 Asymptotic theory for the local linear estimator

Let $\mu_{s,t} = \int_{\mathbb{R}} v^s w(v)^t \, dv$, $s, t = 0, 1, 2$. Define

$$\boldsymbol{\Phi}(\mathbf{u}) = f_{\mathbf{U}}(\mathbf{u}) \begin{pmatrix} \boldsymbol{\Omega}(\mathbf{u}) & \mathbf{0}_{q \times dp_c} \\ \mathbf{0}_{qp_c \times d} & \mu_{2,1}\boldsymbol{\Omega}(\mathbf{u}) \otimes \mathbf{I}_{p_c} \end{pmatrix}, \text{ and} \tag{3.1}$$

$$\boldsymbol{\Upsilon}(\mathbf{u}) = f_{\mathbf{U}}(\mathbf{u}) \begin{pmatrix} \mu_{0,2}^{p_c}\boldsymbol{\Omega}^*(\mathbf{u}) & \mathbf{0}_{q \times qp_c} \\ \mathbf{0}_{qp_c \times q} & \mu_{2,2}\boldsymbol{\Omega}^*(\mathbf{u}) \otimes \mathbf{I}_{p_c} \end{pmatrix}. \tag{3.2}$$

Clearly, $\boldsymbol{\Phi}(\mathbf{u})$ is a $q(1+p_c) \times d(1+p_c)$ matrix and $\boldsymbol{\Upsilon}(\mathbf{u})$ is $q(1+p_c) \times q(1+p_c)$ matrix. To describe the leading bias term associated with the discrete random variables, we define an indicator function $I_s(\cdot, \cdot)$ by $I_s(\mathbf{u}^d, \tilde{\mathbf{u}}^d) = 1\{\mathbf{u}^d \neq \tilde{\mathbf{u}}_s^d\} \prod_{t \neq s}^{p_d} 1\{\mathbf{u}^d = \tilde{\mathbf{u}}_t^d\}$. That is, $I_s(\mathbf{u}^d, \tilde{\mathbf{u}}^d)$ is one if and only $\mathbf{u}^d$ and $\tilde{\mathbf{u}}^d$ differ only in the sth component and is zero otherwise. Let

$$\mathbf{B}(\mathbf{u}; \mathbf{h}, \boldsymbol{\lambda}) = \left\{ \begin{pmatrix} \frac{1}{2}\mu_{2,1} f_{\mathbf{U}}(\mathbf{u}) \boldsymbol{\Omega}(\mathbf{u}) \mathbf{A}(\mathbf{u}) \\ \mathbf{0}_{dp_c \times 1} \end{pmatrix} \right.$$
$$\left. + \sum_{\tilde{\mathbf{u}}^d \in \mathcal{D}} \sum_{s=1}^{p_d} \lambda_s I_s(\mathbf{u}^d, \tilde{\mathbf{u}}^d) f_{\mathbf{U}}(\mathbf{u}^c, \tilde{\mathbf{u}}^d) \begin{pmatrix} \boldsymbol{\Omega}(\mathbf{u}^c, \tilde{\mathbf{u}}^d)(\mathbf{g}(\mathbf{u}^c, \tilde{\mathbf{u}}^d) - \mathbf{g}(\mathbf{u}^c, \mathbf{u}^d)) \\ -\mu_{2,1}(\boldsymbol{\Omega}(\mathbf{u}^c, \tilde{\mathbf{u}}^d) \otimes \mathbf{I}_{p_c}) \dot{\mathbf{g}}(\mathbf{u}^c, \mathbf{u}^d) \end{pmatrix} \right\}, \tag{3.3}$$

where $\mathbf{A}(\mathbf{u}) = (\sum_{s=1}^{p_c} h_s^2 g_{1,ss}(\mathbf{u}), \cdots, \sum_{s=1}^{p_c} h_s^2 g_{d,ss}(\mathbf{u}))'$, $\mathbf{g}(\mathbf{u}) = (g_1(\mathbf{u}), \cdots., g_d(\mathbf{u}))'$, and $\dot{\mathbf{g}}(\mathbf{u}) = (\dot{g}_1(\mathbf{u})', \cdots, \dot{g}_d(\mathbf{u})')'$.

Now we state our first main theorem.

Theorem 3.1 *Suppose that Assumptions A1-A6 hold. Then* $\sqrt{n\mathbf{h}!}\{\mathbf{H}[\widehat{\boldsymbol{\alpha}}_{\boldsymbol{\Psi}_n}(\mathbf{u}) - \boldsymbol{\alpha}(\mathbf{u})] - (\boldsymbol{\Phi}'\boldsymbol{\Psi}^{-1}\boldsymbol{\Phi})^{-1} \boldsymbol{\Phi}'\boldsymbol{\Psi}^{-1}\mathbf{B}(\mathbf{u}; \mathbf{h}, \boldsymbol{\lambda})\} \xrightarrow{d} N(0, (\boldsymbol{\Phi}'\boldsymbol{\Psi}^{-1}\boldsymbol{\Phi})^{-1} \boldsymbol{\Phi}'\boldsymbol{\Psi}^{-1}\boldsymbol{\Upsilon}\Psi^{-1}\boldsymbol{\Phi}(\boldsymbol{\Phi}'\boldsymbol{\Psi}^{-1}\boldsymbol{\Phi})^{-1})$, *where we have suppressed the dependence of* $\boldsymbol{\Phi}$, $\boldsymbol{\Psi}$, *and* $\boldsymbol{\Upsilon}$ *on* $\mathbf{u}$, *and* $\mathbf{H} = diag(1, \cdots, 1, \mathbf{h}', \cdots, \mathbf{h}')$ *is a* $d(p_c+1) \times d(p_c+1)$ *diagonal matrix with both* 1 *and* $\mathbf{h}$ *appearing* d *times.*

Remark 3 (Optimal choice of weight matrix) To minimize the asymptotic variance-covariance matrix of $\widehat{\boldsymbol{\alpha}}_{\boldsymbol{\Psi}_n}$, we can choose $\boldsymbol{\Psi}_n(\mathbf{u})$ as a consistent estimate of $\Upsilon(\mathbf{u})$. Let $\widetilde{\boldsymbol{\alpha}}(\mathbf{u})$ be a preliminary estimate of $\boldsymbol{\alpha}(\mathbf{u})$ by setting $\boldsymbol{\Psi}_n(\mathbf{u}) = \mathbf{I}_{q(p_c+1)}$. Define the local residual $\widetilde{\varepsilon}_i(\mathbf{u}) = Y_i - \sum_{j=1}^{d} \widetilde{g}_j(\mathbf{u}) X_{i,j}$, where $\widetilde{g}_j(\mathbf{u})$ is the jth component of $\widetilde{\boldsymbol{\alpha}}(\mathbf{u})$. Let

$$\widehat{\boldsymbol{\Upsilon}}(\mathbf{u}) = \frac{\mathbf{h}!}{n} \sum_{i=1}^{n} \begin{pmatrix} \mathbf{Z}_i^a \mathbf{Z}_i^{a\prime} \widetilde{\varepsilon}_i(\mathbf{u})^2 & \mathbf{Z}_i^a \left(\mathbf{Z}_i^{a\prime} \otimes \boldsymbol{\eta}_i(\mathbf{u}^c)'\right) \widetilde{\varepsilon}_i(\mathbf{u})^2 \\ (\mathbf{Z}_i^a \otimes \boldsymbol{\eta}_i(\mathbf{u}^c)) \mathbf{Z}_i^{a\prime} \widetilde{\varepsilon}_i(\mathbf{u})^2 & (\mathbf{Z}_i^a \mathbf{Z}_i^{a\prime}) \otimes \left(\boldsymbol{\eta}_i(\mathbf{u}^c) \boldsymbol{\eta}_i(\mathbf{u}^c)'\right) \widetilde{\varepsilon}_i(\mathbf{u})^2 \end{pmatrix} K^2_{\mathbf{h}\boldsymbol{\lambda},i\mathbf{u}}$$

where $\boldsymbol{\eta}_i(\mathbf{u}^c) = (\mathbf{U}_i^c - \mathbf{u}^c)/\mathbf{h}$. It is easy to show that under Assumptions A1-A6 $\widehat{\boldsymbol{\Upsilon}}(\mathbf{u}) = \boldsymbol{\Upsilon}(\mathbf{u}) + o_P(1)$.[1] By choosing $\boldsymbol{\Psi}_n(\mathbf{u}) = \widehat{\boldsymbol{\Upsilon}}(\mathbf{u})$, we denote the resulting local linear GMM estimator of $\boldsymbol{\alpha}(\mathbf{u})$ as $\widehat{\boldsymbol{\alpha}}_{\widehat{\boldsymbol{\Upsilon}}}(\mathbf{u})$. We summarize the asymptotic properties of this estimator in the following corollary, whose proof is straightforward.

Corollary 3.2 *Suppose that Assumptions A1-A6 hold. Then* $\sqrt{n\mathbf{h}!}\{\mathbf{H}[\widehat{\boldsymbol{\alpha}}_{\widehat{\boldsymbol{\Upsilon}}}(\mathbf{u}) - \boldsymbol{\alpha}(\mathbf{u})] - \left(\boldsymbol{\Phi}'\boldsymbol{\Upsilon}^{-1}\boldsymbol{\Phi}\right)^{-1} \boldsymbol{\Phi}'\boldsymbol{\Upsilon}^{-1}\mathbf{B}(\mathbf{u};\mathbf{h},\boldsymbol{\lambda})\} \xrightarrow{d} N(0, \left(\boldsymbol{\Phi}'\boldsymbol{\Upsilon}^{-1}\boldsymbol{\Phi}\right)^{-1})$. *In particular,* $\sqrt{n\mathbf{h}!}\{\widehat{\mathbf{g}}_{\widehat{\boldsymbol{\Upsilon}}}(\mathbf{u}) - \mathbf{g}(\mathbf{u}) - f_{\mathbf{U}}(\mathbf{u})^{-1}[\boldsymbol{\Omega}'(\mathbf{u})\boldsymbol{\Omega}^*(\mathbf{u})^{-1} \boldsymbol{\Omega}(\mathbf{u})]^{-1}\ \boldsymbol{\Omega}'(\mathbf{u})\boldsymbol{\Omega}^*(\mathbf{u})^{-1}\mathbf{B}_0(\mathbf{u};\mathbf{h},\boldsymbol{\lambda})\} \xrightarrow{d} N(0,\ \mu_{0,2}^{p_c} f_{\mathbf{U}}(\mathbf{u})^{-1}[\boldsymbol{\Omega}'(\mathbf{u})\boldsymbol{\Omega}^*(\mathbf{u})^{-1}\boldsymbol{\Omega}(\mathbf{u})]^{-1})$, *where* $\widehat{\mathbf{g}}_{\widehat{\boldsymbol{\Upsilon}}}(\mathbf{u})$ *and* $\mathbf{B}_0(\mathbf{u};\mathbf{h},\boldsymbol{\lambda})$ *denote the first* d *elements of* $\widehat{\boldsymbol{\alpha}}_{\widehat{\boldsymbol{\Upsilon}}}(\mathbf{u})$ *and* $\mathbf{B}(\mathbf{u};\mathbf{h},\boldsymbol{\lambda})$*, respectively.*

Remark 4 (Asymptotic independence of the estimators of functional coefficients and their first order derivatives) Theorem 3.1 indicates the for the general choice of $\boldsymbol{\Psi}_n$ that may not be block diagonal, the estimators of the functional coefficients and those of their first order derivatives may not be asymptotically independent. Nevertheless, if one chooses $\boldsymbol{\Psi}_n$ as an asymptotically block diagonal matrix (i.e., the limit of $\boldsymbol{\Psi}_n$ is block diagonal) as in Corollary 3.2, then we have asymptotic independence between the estimator for $\mathbf{g}(\mathbf{u})$ and that for $\dot{\mathbf{g}}(\mathbf{u})$. If further $q = d$, then the formulae for the asymptotic bias and variance of $\widehat{\mathbf{g}}_{\widehat{\boldsymbol{\Upsilon}}}(\mathbf{u})$ can be simplified to $\boldsymbol{\Omega}(\mathbf{u})^{-1}\mathbf{B}_0(\mathbf{u};\mathbf{h},\boldsymbol{\lambda})/f_{\mathbf{U}}(\mathbf{u})$ and $\mu_{0,2}^{p_c}\boldsymbol{\Omega}(\mathbf{u})^{-1}\boldsymbol{\Omega}^*(\mathbf{u})(\boldsymbol{\Omega}(\mathbf{u})^{-1})'/f_{\mathbf{U}}(\mathbf{u})$, respectively.

3.3 Selection of smoothing parameters

By Corollary 3.2, we can define the following aggregate mean squared error (AMSE) of $\widehat{\mathbf{g}}_{\widehat{\boldsymbol{\Upsilon}}}(\mathbf{u})$ is

$$\begin{aligned} AMSE\left(\widehat{\mathbf{g}}_{\widehat{\boldsymbol{\Upsilon}}}(\mathbf{u})\right) &= \left\| f_{\mathbf{U}}(\mathbf{u})^{-1}\left(\boldsymbol{\Omega}'(\mathbf{u})\boldsymbol{\Omega}^*(\mathbf{u})^{-1}\boldsymbol{\Omega}(\mathbf{u})\right)^{-1}\boldsymbol{\Omega}'(\mathbf{u})\boldsymbol{\Omega}^*(\mathbf{u})^{-1}\mathbf{B}_0(\mathbf{u};\mathbf{h},\boldsymbol{\lambda})\right\|^2 \\ &\quad + \frac{1}{n\mathbf{h}!}\mathrm{tr}\left(\mu_{0,2}^{p_c} f_{\mathbf{U}}(\mathbf{u})^{-1}\left(\boldsymbol{\Omega}'(\mathbf{u})\boldsymbol{\Omega}^*(\mathbf{u})^{-1}\boldsymbol{\Omega}(\mathbf{u})\right)^{-1}\right). \end{aligned}$$

where terms of smaller order are ignored. By symmetry, all h_j, $j = 1, \cdots, p_c$, should have the same order and all λ_l, $l = 1, \cdots, p_d$, should also have the same order but with $\lambda_l \propto h_j^2$. By an argument similar to Li and Racine (2008), it is easy to obtain the optimal rates of bandwidths in terms of minimizing a weighted integrated version of $AMSE(\widehat{\mathbf{g}}_{\widehat{\boldsymbol{\Upsilon}}}(\mathbf{u}))$ as follows: $h_j \propto n^{-1/(4+p_c)}$ and $\lambda_l \propto n^{-2/(4+p_c)}$ for $j = 1, \cdots, p_c$ and $l = 1, \cdots, p_d$. Nevertheless, the exact formula for the optimal smoothing parameters

[1] Alternatively, we can obtain the estimates $\widetilde{\boldsymbol{\alpha}}(u)$ and thus $\widetilde{g}_j(u)$ for $u = U_i$, $i = 1, \cdots, n$, and then we can define the global residual $\widetilde{\varepsilon}_i = Y_i - \sum_{j=1}^{d}\widetilde{g}_j(U_i)X_{i,j}$. Replacing $\widetilde{\varepsilon}_i(u)$ in the definition of $\widehat{\boldsymbol{\Gamma}}(u)$ by $\widetilde{\varepsilon}_i$ also yield a consistent estimate of $\boldsymbol{\Gamma}(u)$, but this needs preliminary estimation of the functional coefficients at all data points and thus is much more computationally expensive.

is difficult to obtain except for the simplest cases (e.g., $p_c = 1$ and $p_d = 0$ or 1). This also suggests that it is infeasible to use the plug-in bandwidth in applied setting since the plug-in method would first require the formula for each smoothing parameter and then pilot estimates for some unknown functions in the formula.

In practice, we propose to use least squares cross validation (LSCV) to choose the smoothing parameters. We choose $(\mathbf{h}, \boldsymbol{\lambda})$ to minimize the following least squares cross validation criterion function

$$CV(\mathbf{h}, \boldsymbol{\lambda}) = \frac{1}{n} \sum_{i=1}^{n} \left(Y_i - \sum_{j=1}^{d} \widehat{g}_j^{(-i)} (\mathbf{U}_i, \mathbf{h}, \boldsymbol{\lambda}) X_{i,j} \right)^2 a(\mathbf{U}_i),$$

where $\widehat{g}_j^{(-i)} (\mathbf{U}_i; \mathbf{h}, \boldsymbol{\lambda})$ is the leave-one-out functional coefficient estimator of $g_j(\mathbf{U}_i)$ by using bandwidth $(\mathbf{h}, \boldsymbol{\lambda})$, and $a(\mathbf{U}_i)$ is a weight function that serves to avoid division by zero and perform trimming in areas of sparse support. In practice and the following numerical study we will set $a(\mathbf{U}_i) = \Pi_{j=1}^{p_c} 1\{|U_{i,j}^c - \overline{U}_j^c| \leq 2s_{U_j^c}\}$, where $\overline{U}_j^c$ and $s_{U_j^c}$ denote the sample mean and standard deviation of $\{U_{i,j}^c, 1 \leq i \leq n\}$, respectively. To implement, one can use grid search for $(\mathbf{h}, \boldsymbol{\lambda})$ when the dimensions of $\mathbf{U}_i^c$ and $\mathbf{U}_j^d$ are both small. Alternatively, one can apply the minimization function built in various software; but multiple starting values are recommended to reduce the chance of local solutions.

4 A Specification Test

In this section, we derive test statistics for the hypothesis that some of the functional coefficients are constant over the regressors in the functional coefficients. The test can be applied to any nonempty subset of the full set of functional coefficients.

4.1 Hypotheses and test statistic

We first split up the set of regressors in $\mathbf{X}_i$ and the set of functional coefficients in $\mathbf{g}(\mathbf{u})$ into two components (after possibly rearranging the regressors): $\mathbf{X}_{1i} = (X_{i,1}, \cdots, X_{i,d_1})'$ associated with $\mathbf{g}_1(\mathbf{u}) = (g_1(\mathbf{u}), \cdots, g_{d_1}(\mathbf{u}))'$, and $\mathbf{X}_{2i} = (X_{i,d_1+1}, \cdots, X_{i,d})'$, associated with $\mathbf{g}_2(\mathbf{u}) = (g_{d_1+1}(\mathbf{u}), \cdots, g_d(\mathbf{u}))'$, where $X_{i,1}$ may not denote the constant term (1) in this section. Then we can rewrite the model (2.1) as

$$\begin{cases} Y_i = \mathbf{g}_1(\mathbf{U}_i)' \mathbf{X}_{1i} + \mathbf{g}_2(\mathbf{U}_i)' \mathbf{X}_{2i} + \varepsilon_i \\ E(\varepsilon_i | \mathbf{Z}_i, \mathbf{U}_i) = 0 \text{ a.s.} \end{cases}. \tag{4.1}$$

Suppose that we want to test for the constancy of functional coefficients for a subset of the regressors $\mathbf{X}_{1i}$ and maintain the assumption that the functional coefficients of $\mathbf{X}_{2i}$ may depend on the set of exogenous regressors $\mathbf{U}_i$. Then the null hypothesis is

$$\mathbb{H}_0 : \mathbf{g}_1(\mathbf{U}_i) = \boldsymbol{\theta}_1 \text{ a.s. for some parameter } \boldsymbol{\theta}_1 \in \mathbb{R}^{d_1}, \tag{4.2}$$

and the alternative hypothesis $\mathbb{H}_1$ denotes the negation of $\mathbb{H}_0$. Under $\mathbb{H}_0$, d_1 of the d functional coefficients are constant whereas under $\mathbb{H}_1$, at least one of the functional coefficients in $\mathbf{g}_1$ is not constant.

There are many ways to test the null hypothesis in (4.2). One way is to estimate the following restricted semiparametric functional coefficient IV model[2]

$$Y_i = \boldsymbol{\theta}_1' \mathbf{X}_{1i} + \mathbf{g}_2(\mathbf{U}_i)' \mathbf{X}_{2i} + \varepsilon_i^{(r)} \tag{4.3}$$

where $\varepsilon_i^{(r)}$ is the restricted error term defined by (4.3) such that $E(\varepsilon_i^{(r)}|\mathbf{Z}_i, \mathbf{U}_i) = 0$ a.s. under the null, and propose a Lagrangian multiplier (LM) type of test based on the estimation of this restricted model only, say by considering the test statistic based on the sample analog of $E[\varepsilon_i^{(r)} E[\varepsilon_i^{(r)}|\mathbf{V}_i] f_{\mathbf{V}}(\mathbf{V}_i)]$ where $f_{\mathbf{V}}$ is the PDF of $\mathbf{V}_i = (\mathbf{Z}_i', \mathbf{U}_i')'$.[3] The second way is to adopt the likelihood ratio (LR) principle to estimate both the unrestricted and restricted models and construct various test statistics, say, by comparing the estimates of either $\mathbf{g}_1$ or $\mathbf{g} = (\mathbf{g}_1', \mathbf{g}_2')'$ in both models through certain distance measure (e.g., Hong and Lee, 2009), or by extending the generalized likelihood ratio (GLR) test of Fan, Zhang, and Zhang (2001) to our framework where endogeneity is present. Clearly tests based the LM principle (and $E[\varepsilon_i^{(r)} E[\varepsilon_i^{(r)}|\mathbf{V}_i] f_{\mathbf{V}}(\mathbf{V}_i)]$ in particular) may suffer from the problem of curse of dimensionality because the dimension of the continuous variables in $\mathbf{V}_i$ is typically larger than the dimension p_c of $\mathbf{U}_i^c$. Tests based on the LR principle requires nonparametric/semiparametric estimation under both the null and alternative, and unless $d_1 = d$, the estimation of the restricted model (4.3) is more involved than the estimation of the unrestricted model.[4]

For this reason, we propose a Wald type of test statistic that requires only consistent estimation of the unrestricted model. Let $\widehat{\mathbf{g}}_{\boldsymbol{\Psi}_n}(\mathbf{u})$ denote the first d element of $\widehat{\boldsymbol{\alpha}}_{\boldsymbol{\Psi}_n}(\mathbf{u})$. It is the estimator of $\mathbf{g}(\mathbf{u}) = (\mathbf{g}_1(\mathbf{u})', \mathbf{g}_2(\mathbf{u}))'$. Split $\widehat{\mathbf{g}}_{\boldsymbol{\Psi}_n}(\mathbf{u})$ as $\widehat{\mathbf{g}}_1(\mathbf{u}) = \widehat{\mathbf{g}}_{1,\boldsymbol{\Psi}_n}(\mathbf{u})$ and $\widehat{\mathbf{g}}_2(\mathbf{u}) = \widehat{\mathbf{g}}_{2,\boldsymbol{\Psi}_n}(\mathbf{u})$ so that $\widehat{\mathbf{g}}_l(\mathbf{u})$ estimates $\mathbf{g}_l(\mathbf{u})$, $l = 1, 2$. Our proposed test statistic is

$$T_n = (\mathbf{h}!)^{1/2} \sum_{i=1}^{n} \left\| \widehat{\mathbf{g}}_1(\mathbf{U}_i) - \overline{\widehat{\mathbf{g}}}_1 \right\|^2 \tag{4.4}$$

where $\overline{\widehat{\mathbf{g}}}_1 = n^{-1} \sum_{i=1}^{n} \widehat{\mathbf{g}}_1(\mathbf{U}_i)$. In the next subsection, we show that after being suitably normalized, T_n is asymptotically distributed as $N(0,1)$ under the null hypothesis and diverges to infinity under the alternative.

4.2 Asymptotic distributions of the test statistic

Let $\boldsymbol{\Phi}_n(\mathbf{u}) = n^{-1} \mathbf{Q}_{\mathbf{h}}(\mathbf{u})' \mathbf{K}_{\mathbf{h}\boldsymbol{\lambda}}(\mathbf{u}) \boldsymbol{\xi}(\mathbf{u}) \mathbf{H}^{-1}$. Define

$$\begin{aligned} \boldsymbol{\Gamma}_{n1}(\mathbf{u}) &= \mathbb{S}_1 \left[\boldsymbol{\Phi}_n(\mathbf{u})' \boldsymbol{\Psi}_n(\mathbf{u})^{-1} \boldsymbol{\Phi}_n(\mathbf{u}) \right]^{-1} \boldsymbol{\Phi}_n(\mathbf{u})' \boldsymbol{\Psi}_n(\mathbf{u})^{-1}, \text{ and} \\ \overline{\boldsymbol{\Gamma}}_1(\mathbf{u}) &= \mathbb{S}_1 \left[\boldsymbol{\Phi}(\mathbf{u})' \boldsymbol{\Psi}(\mathbf{u})^{-1} \boldsymbol{\Phi}(\mathbf{u}) \right]^{-1} \boldsymbol{\Phi}(\mathbf{u})' \boldsymbol{\Psi}(\mathbf{u})^{-1}, \end{aligned} \tag{4.5}$$

[2] (4.3) is similar to the partially varying coefficient IV model studied by Cai and Xiong (2006).

[3] Note that $E[\varepsilon_i^{(r)} E[\varepsilon_i^{(r)}|\mathbf{V}_i] f_{\mathbf{V}}(\mathbf{V}_i)] = 0$ under $\mathbb{H}_0$ and is strictly positive under $\mathbb{H}_1$.

[4] There are two ways to estimate $\boldsymbol{\theta}_1$ and $\mathbf{g}_2$ in (4.3). One is through the estimation of the unrestricted model in (4.1), i.e., by estimating $\mathbf{g}_1(\mathbf{u})$ and $\mathbf{g}_2(\mathbf{u})$ in (4.1), say by $\widehat{\mathbf{g}}_1(\mathbf{u})$ and $\widehat{\mathbf{g}}_2(\mathbf{u})$, then averaging $\widehat{\mathbf{g}}_1(\mathbf{U}_i)$ to obtain the estimate of $\boldsymbol{\theta}_1$, and using $\widehat{\mathbf{g}}_2(\mathbf{u})$ as the estimate of $\mathbf{g}_2(\mathbf{u})$ in (4.3). The other is to estimate the restricted model in (4.3) directly by using the idea of profile least squares. The latter requires two steps: 1) approximate $\mathbf{g}_2(\cdot)$ locally, minimize some kernel weighted GMM objective function, and express $\mathbf{g}_2(\mathbf{u})$ as a function of $\boldsymbol{\theta}_1$ for $\mathbf{u} = \mathbf{U}_i$, $i = 1, ..., n$, by treating $\boldsymbol{\theta}_1$ as if it were known, 2) plug in the above expression of $\mathbf{g}_2(\mathbf{U}_i)$ into a global GMM objective function and minimize it with respect to $\boldsymbol{\theta}_1$ to obtain the estimate of $\boldsymbol{\theta}_1$, and then recover the estimate of $\mathbf{g}_2(\mathbf{u})$ by using the functional relationship between $\mathbf{g}_2(\mathbf{u})$ and $\boldsymbol{\theta}_1$ obtained in step 1). This latter method is more involved than the former one unless $d = d_1$ so that $\mathbf{g}_2$ is absent in (4.3).

where $\mathbb{S}_1 = \left(\mathbf{I}_{d_1}, \mathbf{0}_{d_1 \times (d_1 p_c + d_2(p_c+1))}\right)$ is a selection matrix. We add the following assumptions.

Assumption A7. (i) $\boldsymbol{\Psi}_n(\mathbf{u}) = \boldsymbol{\Psi}(\mathbf{u}) + O_P(\nu_n)$ uniformly in $\mathbf{u}$, where $\boldsymbol{\Psi}(\mathbf{u})$ is symmetric and positive definite for each $\mathbf{u}$ and $\nu_n \to 0$ as $n \to \infty$. (ii) $\sup_{\mathbf{u}} \left|\overline{\boldsymbol{\Gamma}}_1(\mathbf{u})\right| < C < \infty$.

Assumption A8. As $n \to \infty$, (i) $n^{1/2}(\|\mathbf{h}\|^2 + \|\boldsymbol{\lambda}\|)\nu_n \to 0$, (ii) $(\|\mathbf{h}\|^2 + \|\boldsymbol{\lambda}\|)\left(\mathbf{h}!^{-1/2}\right)\sqrt{\log n} \to 0$, and (iii) $n(\mathbf{h}!)^{1/2}\left(\|\mathbf{h}\|^4 + \|\boldsymbol{\lambda}\|^2\right) \to 0$.

A7 strengthens A4(ii). It is satisfied if one chooses $\boldsymbol{\Psi}_n(\mathbf{u})$ as the identity matrix $\mathbf{I}_{q(p_c+1)}$ for all $\mathbf{u}$, in which case optimal weight matrix is not used to construct the test statistic and $\nu_n = 0$. Alternatively, if one chooses $\boldsymbol{\Psi}_n(\mathbf{u}) = \widehat{\boldsymbol{\Upsilon}}(\mathbf{u})$, then one can verify that A7(i) is satisfied with $\nu_n = \|\mathbf{h}\|^2 + \|\boldsymbol{\lambda}\| + (n\mathbf{h}!/\log n)^{-1/2}$. A7(ii) is weak given the compact support of the continuous regressor $\mathbf{U}_i^c$. A8(i) can easily be satisfied whereas A8(ii) requires that $p_c \leq 3$, which is not restrictive due to the notorious "curse of dimensionality" in the nonparametric literature. A8(iii) requires that undersmoothing bandwidth must be used in order to remove the effect of asymptotic bias of our nonparametric estimators.

To proceed, we first consider the consistent estimation of $\boldsymbol{\theta}_1$ under $\mathbb{H}_0$. We estimate it by

$$\widehat{\boldsymbol{\theta}}_1 = \overline{\widehat{\mathbf{g}}}_1 = n^{-1}\sum_{i=1}^{n}\widehat{\mathbf{g}}_1(\mathbf{U}_i). \tag{4.6}$$

By (A.1) in the appendix, we have the following usual bias and variance decomposition for $\widehat{\mathbf{g}}_1(\mathbf{U}_i)$:

$$\widehat{\mathbf{g}}_1(\mathbf{U}_i) - \mathbf{g}_1(\mathbf{U}_i) = \boldsymbol{\Gamma}_{n1}(\mathbf{U}_i)\mathcal{B}_n(\mathbf{U}_i) + \boldsymbol{\Gamma}_{n1}(\mathbf{U}_i)\mathcal{V}_n(\mathbf{U}_i). \tag{4.7}$$

where $\boldsymbol{\Gamma}_{n1}(\mathbf{u})$ is defined in (4.5), the bias term $\mathcal{B}_n(\mathbf{U}_i)$ and the variance term $\mathcal{V}_n(\mathbf{U}_i)$ are defined in the line after (A.1). Under $\mathbb{H}_0$,

$$\sqrt{n}\left(\widehat{\boldsymbol{\theta}}_1 - \boldsymbol{\theta}_1\right) = n^{-1/2}\sum_{i=1}^{n}\boldsymbol{\Gamma}_{n1}(\mathbf{U}_i)\mathcal{B}_n(\mathbf{U}_i) + n^{-1/2}\sum_{i=1}^{n}\boldsymbol{\Gamma}_{n1}(\mathbf{U}_i)\mathcal{V}_n(\mathbf{U}_i). \tag{4.8}$$

We shall show that the first term (bias) on the right hand side of (4.8) is asymptotically negligible under some extra condition on the bandwidth sequence, whereas the second term contributes to the asymptotic variance-covariance (VC) of $\widehat{\boldsymbol{\theta}}_1$. To characterize the asymptotic VC matrix of $\widehat{\boldsymbol{\theta}}_1$, let $\boldsymbol{\zeta}_i = (\mathbf{U}_i', \mathbf{Z}_i^{a\prime}, \varepsilon_i)'$,

$$\boldsymbol{\varphi}(\boldsymbol{\zeta}_i, \boldsymbol{\zeta}_j) = \overline{\boldsymbol{\Gamma}}_1(\mathbf{U}_i)\begin{pmatrix} \mathbf{Z}_j^a\varepsilon_j \\ (\mathbf{Z}_j^a\varepsilon_j)\otimes\boldsymbol{\eta}_j(\mathbf{U}_i^c)\end{pmatrix}K_{\mathbf{h}\boldsymbol{\lambda},j\mathbf{U}_i}, \text{ and } \overline{\boldsymbol{\varphi}}(\boldsymbol{\zeta}_i) = \int\boldsymbol{\varphi}(\boldsymbol{\zeta}, \boldsymbol{\zeta}_i)dF_{\boldsymbol{\zeta}}(\boldsymbol{\zeta}), \tag{4.9}$$

where $F_{\boldsymbol{\zeta}}$ denotes the CDF of $\boldsymbol{\zeta}_i$. Let $\boldsymbol{\Sigma}_{\boldsymbol{\theta}_1} = \lim_{n\to\infty} E\left[\overline{\boldsymbol{\varphi}}(\boldsymbol{\zeta}_i)\overline{\boldsymbol{\varphi}}(\boldsymbol{\zeta}_i)'\right]$. Straightforward but tedious calculations show that

$$\boldsymbol{\Sigma}_{\boldsymbol{\theta}_1} = \left(\int\int w(t)\,w(t-s)dtds\right)^{p_c}\int\overline{\boldsymbol{\Gamma}}_1(\mathbf{u})\begin{pmatrix}\boldsymbol{\Omega}^*(\mathbf{u}) & \mathbf{0}_{q\times p_c q} \\ \mathbf{0}_{p_c q\times q} & \mathbf{0}_{p_c q\times p_c q}\end{pmatrix}\overline{\boldsymbol{\Gamma}}_1(\mathbf{u})' f_{\mathbf{U}}(\mathbf{u})^2\, dF_{\mathbf{U}}(\mathbf{u}). \tag{4.10}$$

The following theorem establishes the $\sqrt{n}$-consistency and asymptotic normality of $\widehat{\boldsymbol{\theta}}_1$ under $\mathbb{H}_0$.

Theorem 4.1 *Suppose Assumptions A1-A4(i) and A5-A8 hold. Suppose that* $n^{1/2}(\|\mathbf{h}\|^2 + \|\boldsymbol{\lambda}\|) = o(1)$, $\nu_n(\mathbf{h}!)^{-1/2} = o(1)$, *and* $n(\mathbf{h}!)^2/\log n \to \infty$ *as* $n \to \infty$. *Then under* $\mathbb{H}_0$, $\sqrt{n}(\widehat{\boldsymbol{\theta}}_1 - \boldsymbol{\theta}_1) \xrightarrow{d} N(\mathbf{0}_{d_1\times 1}, \boldsymbol{\Sigma}_{\boldsymbol{\theta}_1})$.

Clearly Theorem 4.1 says that under $\mathbb{H}_0$, $\widehat{\boldsymbol{\theta}}_1$ can consistently estimate $\boldsymbol{\theta}_1$ at the usual $\sqrt{n}$-rate. The extra conditions on the bandwidth in the above theorem ensures that the bias term in (4.7) vanishes asymptotically and the replacement of $\boldsymbol{\Gamma}_{n1}(\mathbf{U}_i)$ in (4.7) by $\overline{\boldsymbol{\Gamma}}_1(\mathbf{U}_i)$ has asymptotically negligible effect on the asymptotic normality of $\widehat{\boldsymbol{\theta}}_1$. If all functional coefficients are constant under $\mathbb{H}_0$, then $\mathcal{B}_n(\mathbf{U}_i) = 0$ a.s. so that we do not need the first the extra condition on the bandwidth in the theorem.

Let $B_n \equiv n^{-2}(\mathbf{h}!)^{1/2}\sum_{i=1}^n\sum_{j=1}^n ||\boldsymbol{\varphi}(\boldsymbol{\zeta}_i,\boldsymbol{\zeta}_j)||^2$ and $\sigma_0^2 \equiv \lim_{n\to\infty} 2\mathbf{h}! E_j E_l[\int \boldsymbol{\varphi}(\boldsymbol{\zeta},\boldsymbol{\zeta}_j)'\boldsymbol{\varphi}(\boldsymbol{\zeta},\boldsymbol{\zeta}_l)dF_{\boldsymbol{\zeta}}(\boldsymbol{\zeta})]^2$, where E_j denotes the expectation with respect to $\boldsymbol{\zeta}_j$. We can now describe the asymptotic distribution of T_n under $\mathbb{H}_0$ as $n\to\infty$.

Theorem 4.2 *Suppose Assumptions A1-A4(i) and A5-A8 hold. Then under* $\mathbb{H}_0$, $T_n - B_n \xrightarrow{d} N(0,\sigma_0^2)$.

Following the last remark after Theorem 4.1, Assumption A8(iii) is not needed for the above theorem if we are testing the constancy of all functional coefficients.

To implement the test, we consistently estimate B_n and σ_0^2 using

$$\widehat{B}_n \equiv \frac{(\mathbf{h}!)^{1/2}}{n^2}\sum_{i=1}^n\sum_{j=1}^n \left\|\widehat{\boldsymbol{\varphi}}_{ij}\right\|^2 \text{ and } \widehat{\sigma}_n^2 = \frac{2\mathbf{h}!}{n(n-1)}\sum_{j=1}^n\sum_{l\neq j}^n\left[\frac{1}{n}\sum_{i=1}^n \widehat{\boldsymbol{\varphi}}_{ij}'\widehat{\boldsymbol{\varphi}}_{il}\right]^2,$$

where $\widehat{\boldsymbol{\varphi}}_{ij} = \boldsymbol{\Gamma}_{n1}(\mathbf{U}_i)\begin{pmatrix}\mathbf{Z}_j^a\widehat{\varepsilon}_j\\ (\mathbf{Z}_j^a\widehat{\varepsilon}_j)\otimes\boldsymbol{\eta}_j(\mathbf{U}_i^c)\end{pmatrix}K_{\mathbf{h}\boldsymbol{\lambda},j\mathbf{U}_i}$, and $\widehat{\varepsilon}_i = Y_i - \widehat{\mathbf{g}}_{\Psi_n}(\mathbf{U}_i)'\mathbf{X}_i$. It is straightforward to show that $\widehat{B}_n - B_n = o_P(1)$ and $\widehat{\sigma}_n^2 - \sigma_0^2 = o_P(1)$. Then we have

$$J_n \equiv \frac{T_n - \widehat{B}_n}{\sqrt{\widehat{\sigma}_n^2}} \xrightarrow{d} N(0,1) \text{ under } \mathbb{H}_0.$$

When n is sufficiently large, we can compare J_n to the one-sided critical value z_α, the upper α percentile from the $N(0,1)$ distribution, and reject the null at asymptotic level α if $J_n > z_\alpha$.

To examine the asymptotic local power of the J_n test, we consider the sequence of Pitman local alternatives

$$\mathbb{H}_1(r_n): \mathbf{g}_1(\mathbf{U}_i) = \boldsymbol{\theta}_1 + r_n\boldsymbol{\delta}_n(\mathbf{U}_i) \text{ a.s.},$$

where $r_n \to 0$ as $n\to\infty$ and the $\boldsymbol{\delta}_n$'s are a sequence of real continuous vector-valued functions such that $\mu_0 \equiv \lim_{n\to\infty} E[\|\boldsymbol{\delta}_n(\mathbf{U}_i) - E[\boldsymbol{\delta}_n(\mathbf{U}_i)]\|^2] < \infty$. The following theorem establishes the asymptotic local power of the J_n test.

Theorem 4.3 *Suppose Assumptions A1-A4(i) and A5-A8 hold. Then under* $\mathbb{H}_1(r_n)$ *with* $r_n = n^{-1/2}(\mathbf{h}!)^{-1/4}$, $J_n \xrightarrow{d} N(\mu_0/\sigma_0, 1)$.

Theorem 4.3 shows that the J_n test has nontrivial power against Pitman local alternatives that converge to zero at rate $n^{-1/2}(\mathbf{h}!)^{-1/4}$. The asymptotic local power function is given by $\lim_{n\to\infty} P(J_n \geq z \mid \mathbb{H}_1(r_n)) = 1 - \Phi(z - \mu_0/\sigma_0)$, where Φ is the standard normal CDF.

The next theorem establishes the consistency of the test.

Theorem 4.4 *Suppose Assumptions A1-A4(i) and A5-A8 hold. Then under* $\mathbb{H}_1$, $n^{-1}(\mathbf{h}!)^{-1/2}J_n = \mu_A/\sigma_0 + o_P(1)$, *where* $\mu_A \equiv E[\mathbf{g}_1(\mathbf{U}_i) - \boldsymbol{\theta}_1]^2$, *so that* $P(J_n > c_n) \to 1$ *under* $\mathbb{H}_1$ *for any nonstochastic sequence* $c_n = o(n(\mathbf{h}!)^{1/2})$.

4.3 A bootstrap version of our test

It is well known that a nonparametric test based on its asymptotic normal null distribution may perform poorly in finite samples. This is especially true for our test when we have discrete conditioning variables in $\mathbf{U}_i$. So we suggest using a bootstrap method to obtain the bootstrap p-values. We propose to generate the bootstrap data $\{Y_i^*, \mathbf{U}_i^*, \mathbf{X}_i^*, \mathbf{Z}_i^*, \mathbf{Z}_i^{a*}\}$ based on the following procedure:

1. Obtain the local linear GMM estimates $\widehat{\mathbf{g}}_1(\mathbf{U}_i)$ and $\widehat{\mathbf{g}}_2(\mathbf{U}_i)$ by using the optimal weight matrix and the bandwidth $(\mathbf{h}, \boldsymbol{\lambda})$, and calculate the residuals $\widehat{\varepsilon}_i^r = Y_i - \widehat{\boldsymbol{\theta}}_1' \mathbf{X}_{1i} - \widehat{\mathbf{g}}_2(\mathbf{U}_i)' \mathbf{X}_{2i}$ with $\widehat{\boldsymbol{\theta}}_1 \equiv n^{-1} \sum_{i=1}^n \widehat{\mathbf{g}}_1(\mathbf{U}_i)$.
2. Draw $\{(\mathbf{X}_i^*, \mathbf{Z}_i^*, \mathbf{Z}_i^{a*}, \mathbf{U}_i^*, \varepsilon_i^*)\}$ randomly from $\{(\mathbf{X}_i, \mathbf{Z}_i, \mathbf{Z}_i^a, \mathbf{U}_i, \widehat{\varepsilon}_i^r), i = 1, ..., n\}$.
3. Generate $Y_i^* = \tilde{\boldsymbol{\theta}}_1' \mathbf{X}_{1i}^* + \tilde{\mathbf{g}}_2(\mathbf{U}_i^*) \mathbf{X}_{2i}^* + \varepsilon_i^*$, where $\tilde{\boldsymbol{\theta}}_1 \equiv n^{-1} \sum_{i=1}^n \tilde{\mathbf{g}}_1(\mathbf{U}_i)$, and $\tilde{\mathbf{g}}_1(\mathbf{U}_i)$ and $\tilde{\mathbf{g}}_2(\mathbf{U}_i)$ are the local linear GMM estimates by using the optimal weight matrix and the bandwidth $(\tilde{\mathbf{h}}, \tilde{\boldsymbol{\lambda}})$ which is larger than $(\mathbf{h}, \boldsymbol{\lambda})$ elementwise.
4. Compute the bootstrap test statistic J_n^* in the same way as J_n by using $\{Y_i^*, \mathbf{U}_i^*, \mathbf{X}_i^*, \mathbf{Z}_i^*, \mathbf{Z}_i^{a*}\}$ instead.
5. Repeat Steps 1-4 B times to obtain B bootstrap test statistic $\{J_{nj}^*\}_{j=1}^B$. Calculate the bootstrap p-values $p^* \equiv B^{-1} \sum_{j=1}^B 1\{J_{nj}^* \geq J_n\}$ and reject the null hypothesis $\mathbb{H}_0 : \mathbf{g}_1(\mathbf{U}_i) = \boldsymbol{\theta}_1$ a.s. if p^* is smaller than the prescribed nominal level of significance.

Note that we do not use the simple wild bootstrap method in the above procedure because it does not take into account the endogeneity issue in the data. Instead, we draw $(\mathbf{X}_i^*, \mathbf{Z}_i^*, \mathbf{Z}_i^{a*}, \mathbf{U}_i^*, \varepsilon_i^*)$ as a whole randomly from $\{(\mathbf{X}_i, \mathbf{Z}_i, \mathbf{Z}_i^a, \mathbf{U}_i, \widehat{\varepsilon}_i), i = 1, ..., n\}$ which helps maintain the dependence structure among these variables. In addition, we impose the null hypothesis in Step 3 and obtain estimates $\tilde{\boldsymbol{\theta}}_1$ and $\tilde{\mathbf{g}}_2$ by using larger values of bandwidths. See Härdle and Marron (1991) for the justification of this.

5 Monte Carlo Simulations

In this section, we conduct a small set of Monte Carlo experiments to illustrate the finite sample performance of our local linear GMM estimator of functional coefficients and that of our test for the constancy of some functional coefficients.

5.1 Evaluation of the local linear GMM estimates

To evaluate the local linear GMM estimates, we consider the following data generating process (DGP):

$$Y_i = (0.5 + 0.25U_i^{c2} + 0.5U_i^c U_{i,1}^d + U_{i,2}^d) + (1 + U_{i,1}^d e^{0.1U_i^c} + U_{i,2}^d U_i^c) X_i + s\varepsilon_i, \tag{5.1}$$

where $U_i^c \backsim N(0,1)$ truncated at ± 2, $U_{i,1}^d$ and $U_{i,2}^d$ are both Bernoulli random variables taking value 1 with probability 0.5, $\varepsilon_i \backsim N(0,1)$, $X_i = (Z_i + \tau\varepsilon_i) / \sqrt{1+\tau^2}$, and $(Z_i, \varepsilon_i)' \backsim N(\mathbf{0}_{2\times 1}, \mathbf{I}_2)$. Here we use τ to control the degree of endogeneity and choose s to ensure the signal-noise ratio to be 1 when we generate observations on Y_i.[5] We assume that $\{Y_i, U_i^c, U_{i,1}^d, U_{i,2}^d, X_i, Z_i\}_{i=1}^n$ are observed and are interested in the

[5] We set s as the sample standard deviation of $m_i \equiv (0.5 + 0.25U_i^{c2} + 0.5U_i^c U_{i,1}^d + U_{i,2}^d) + (1 + U_{i,1}^d e^{0.1U_i^c} + U_{i,2}^d U_i^c) X_i$ over that of ε_i.

Table 1: Finite Sample Comparison of Various GMM Estimators for the DGP in (5.1)

		ROT bandwidth				LSCV bandwidth			
n	τ	MAD		MSE		MAD		MSE	
		$g_1(u)$	$g_2(u)$	$g_1(u)$	$g_2(u)$	$g_1(u)$	$g_2(u)$	$g_1(u)$	$g_2(u)$
100	0	0.553	0.339	0.649	0.325	0.578	0.301	0.692	0.321
		0.671	0.350	0.770	0.320	0.694	0.320	0.819	0.327
		0.311	0.300	0.920	0.447	0.303	0.131	0.602	0.188
	0.32	0.545	0.342	0.627	0.334	0.568	0.313	0.668	0.334
		0.659	0.354	0.746	0.331	0.680	0.329	0.792	0.340
		0.616	0.398	1.132	0.527	0.634	0.299	0.861	0.259
	0.75	0.509	0.358	0.558	0.375	0.525	0.351	0.585	0.389
		0.607	0.372	0.652	0.375	0.621	0.365	0.683	0.391
		0.908	0.593	1.317	0.716	0.908	0.521	1.122	0.440
	2.07	0.420	0.484	0.424	0.651	0.453	0.526	0.449	0.673
		0.464	0.529	0.468	0.690	0.481	0.536	0.464	0.659
		0.737	0.832	0.734	1.020	0.743	0.750	0.669	0.747
400	0	0.378	0.200	0.376	0.118	0.399	0.153	0.418	0.100
		0.553	0.246	0.554	0.138	0.579	0.223	0.593	0.120
		0.123	0.143	0.385	0.134	0.107	0.036	0.264	0.058
	0.32	0.367	0.201	0.356	0.121	0.390	0.155	0.398	0.106
		0.540	0.248	0.531	0.144	0.567	0.226	0.572	0.126
		0.570	0.312	0.654	0.219	0.570	0.306	0.547	0.147
	0.75	0.337	0.212	0.307	0.145	0.357	0.173	0.345	0.136
		0.487	0.257	0.445	0.168	0.509	0.239	0.484	0.155
		0.933	0.589	1.071	0.459	0.924	0.588	0.989	0.390
	2.07	0.261	0.332	0.216	0.354	0.281	0.356	0.229	0.362
		0.333	0.385	0.262	0.379	0.347	0.394	0.276	0.372
		0.764	0.886	0.636	0.867	0.752	0.876	0.595	0.796

Note: Columns 3-6 and 7-10 report the results for estimation based on the rule of thumb (ROT) bandwidth and least squares cross-validated (LSCV) bandwidth, respectively. For each combination of n and τ, the three rows report the results for our estimators with optimal and identity weight matrices, and the SCU estimators, respectively. The values are averages over 500 replications.

estimation of two functional coefficients: $g_1(\mathbf{u}) = g_1\left(u^c, u_1^d, u_2^d\right) = 0.5 + 0.25u^{c2} + 0.5u^c u_1^d + u_2^d$, and $g_2(\mathbf{u}) = g_2\left(u^c, u_1^d, u_2^d\right) = 1 + u_1^d e^{0.1u^c} + u_2^d u^c$.

We consider three nonparametric estimator for $g_1(\mathbf{u})$ and $g_2(\mathbf{u})$. The first two are obtained as our local linear GMM functional coefficient estimators when the weight matrix Ψ_n is chosen to be the identity matrix and optimal weight matrix, respectively, and the instrument Z_i^a is also set to be Z_i. The third estimator is local linear estimator of SCU (2009) where the endogeneity of X_i is neglected. For all estimators, we use the standardized Epanechnikov kernel $k(u) = \frac{3}{4\sqrt{5}}(1-\frac{1}{5}u^2)1\{|u| \leq \sqrt{5}\}$, and consider two choices of smoothing parameters $(\mathbf{h}, \boldsymbol{\lambda}) = (h, \lambda_1, \lambda_2)$ for the conditioning variables $U_i^c, U_{i,1}^d, U_{i,2}^d$ in the functional coefficients. We consider choosing the smoothing parameters $(\mathbf{h}, \boldsymbol{\lambda})$ by both the least squares cross-validation (LSCV) method discussed in Section 3.3, and the simple rule of thumb (ROT) method. In the latter case, we follow the literature and set $h = s_{U^c} n^{-1/5}$, and $\lambda_1 = \lambda_2 = s_{U^c} n^{-2/5}$.

To evaluate the finite sample performance of different estimators, we consider evaluating the estimates of the functional coefficients at some prescribed points. For each of the four possible values that $(U_{i,1}^d, U_{i,2}^d)$ can take, we choose 25 equally-spaced points on the interval $[-2, 2]$ for the continuous variable U_i^c, which gives us 100 evaluation points $\mathbf{u}$ for the variables in the functional coefficients. We estimate both $g_1(\mathbf{u})$ and $g_2(\mathbf{u})$ at these 100 points by using the above three estimation method and two sequences of bandwidths, and calculate the mean absolute deviation (MAD) and mean squared error (MSE) of each estimator based on these 100 evaluations. For each estimator, by averaging these MADs and MSEs across 500 replications, we obtain the final performance measure as follows:

$$MAD_l = \frac{1}{100 \times 500} \sum_{r=1}^{500} \sum_{j=1}^{100} \left| \widehat{g}_l^{(r)}(\mathbf{u}_j) - g_l(\mathbf{u}_j) \right| \text{ and } MSE_l = \frac{1}{100 \times 500} \sum_{r=1}^{500} \sum_{j=1}^{100} \left[\widehat{g}_l^{(r)}(\mathbf{u}_j) - g_l(\mathbf{u}_j) \right]^2$$

for $l = 1, 2$, where $\widehat{g}_l^{(r)}(\mathbf{u}_j)$ is an estimator of $g_l(\mathbf{u}_j)$ in the rth replication by using any one of the above estimation methods or bandwidth sequences.

We consider two sample sizes: $n = 100$ and 400. We choose four different values of τ, namely, 0, 0.32, 0.75, 2.07, to ensure the correlations between X_i and ε_i are 0, 0.3, 0.6, 0.9, respectively.

Table 1 reports the results. We summarize some important findings from Table 1. First, in the absence of endogeneity $(\tau = 0)$, our local linear GMM estimators with optimal weight or identity weight matrices may not outperform the SCU estimators in terms of either MAD or MSE, which is expected. Second, in almost all cases, the local linear GMM estimators with optimal weight matrix tends to outperform the local linear GMM estimators with identity weight matrix. Third, in the case of endogeneity, both local linear GMM estimators generally perform better than the SCU estimators in terms of either MAD or MSE. The stronger endogeneity, the more gains to apply our local linear GMM estimation method.

5.2 Tests for the constancy of functional coefficients

We now consider the finite sample performance of our test. To this goal, we modify (5.1) to

$$Y_i = [0.5 + \delta_1(0.25U_i^{c2} + 0.5U_i^c U_{i,1}^d + U_{i,2}^d)] + [1 + \delta_2\left(U_{i,1}^d e^{0.1U_i^c} + U_{i,2}^d U_i^c\right)]X_i + s\varepsilon_i, \tag{5.2}$$

where all variables are generated as in the above subsection, s is chosen to ensure the signal and noise ratio to be 1, and we allow δ_1 and δ_2 to take different values to evaluate both the size and power properties of our test. Here $g_1(\mathbf{u}) = g_1(u^c, u_1^d, u_2^d) = 0.5 + \delta_1\left(0.25u^{c2} + 0.5u^c u_1^d + u_2^d\right)$, and $g_2(\mathbf{u}) = g_2(u^c, u_1^d, u_2^d) = 1 + \delta_2\left(u_1^d e^{0.1u^c} + u_2^d u^c\right)$.And when $\delta_1 = \delta_2 = 1$, (5.2) reduces to (5.1).

We consider the following three null hypotheses:

$$\begin{aligned} \mathbb{H}_{0,1} &: \quad g_1(\mathbf{U}_i) = \theta_1 \text{ a.s.}, \\ \mathbb{H}_{0,2} &: \quad g_2(\mathbf{U}_i) = \theta_2 \text{ a.s.}, \\ \mathbb{H}_{0,12} &: \quad (g_1(\mathbf{U}_i), g_2(\mathbf{U}_i)) = (\theta_1, \theta_2) \text{ a.s.}, \end{aligned}$$

for some unknown parameters θ_1 and θ_2.

To construct the test statistic, we need to choose both the kernel and the bandwidth. As in the previous section, we choose the standardized Epanechnikov kernel and consider the use bootstrap to approximate the asymptotic null distribution of our test statistics. Section 4.3 suggests that we need to choose two sets of bandwidths $(\mathbf{h}, \boldsymbol{\lambda}) \equiv (h, \lambda_1, \lambda_2)$ and $(\tilde{\mathbf{h}}, \tilde{\boldsymbol{\lambda}}) \equiv (\tilde{h}, \tilde{\lambda}_1, \tilde{\lambda}_2)$: the former is used to construct

Table 2: Rejection Frequency of Nonparametric Tests for the Constancy of Functional Coefficients

n	δ	τ	1%			5%			10%		
			$H_{0,1}$	$H_{0,2}$	$H_{0,12}$	$H_{0,1}$	$H_{0,2}$	$H_{0,12}$	$H_{0,1}$	$H_{0,2}$	$H_{0,12}$
100	0	0.32	0.008	0.004	0.010	0.054	0.048	0.050	0.100	0.092	0.104
		0.75	0.010	0.008	0.010	0.044	0.046	0.056	0.094	0.112	0.104
		2.07	0.004	0.020	0.012	0.046	0.070	0.064	0.096	0.166	0.128
	0.5	0.32	0.316	0.750	0.756	0.538	0.906	0.900	0.654	0.934	0.934
		0.75	0.318	0.574	0.606	0.578	0.780	0.812	0.730	0.874	0.878
		2.07	0.054	0.214	0.174	0.224	0.398	0.356	0.380	0.530	0.500
	1	0.32	0.558	0.962	0.952	0.764	0.986	0.984	0.838	0.994	0.994
		0.75	0.622	0.898	0.902	0.838	0.970	0.968	0.908	0.982	0.978
		2.07	0.282	0.496	0.450	0.548	0.676	0.696	0.690	0.792	0.792
200	0	0.32	0.008	0.008	0.014	0.052	0.042	0.044	0.102	0.094	0.092
		0.75	0.010	0.008	0.016	0.050	0.050	0.048	0.082	0.104	0.096
		2.07	0.006	0.020	0.006	0.042	0.066	0.050	0.084	0.134	0.110
	0.5	0.32	0.574	0.990	0.996	0.776	0.998	0.998	0.876	1.000	1.000
		0.75	0.706	0.946	0.958	0.900	0.986	0.998	0.952	0.994	0.998
		2.07	0.246	0.334	0.402	0.554	0.552	0.654	0.706	0.664	0.754
	1	0.32	0.848	1.000	1.000	0.954	1.000	1.000	0.982	1.000	1.000
		0.75	0.948	1.000	1.000	0.994	1.000	1.000	1.000	1.000	1.000
		2.07	0.778	0.762	0.852	0.938	0.892	0.954	0.966	0.928	0.974

Note: The first three column contains information about the sample size n, the parameters δ and τ, respectively. The remaining nine columns report the tests for the three null hypotheses at 1%, 5%, and 10% levels, respectively. 500 replications and 200 bootstraps are used to obtain the results. The rule of thumb method is used to choose the bandwidth.

the test statistic based on the observed data and to yield the restricted residuals, and the latter is used to obtain $\widetilde{\boldsymbol{\theta}}_1(\mathbf{U}_i)$ and $\tilde{\mathbf{g}}_2(\mathbf{U}_i)$ in order to construct the bootstrap dependent variable by imposing the null hypothesis. We set $h = cs_{U^c}n^{-1/(p_c+3)}$, $\lambda_1 = \lambda_2 = cs_{U^c}n^{-2/(p_c+3)}$, $\tilde{h} = cs_{U^c}n^{-1/(p_c+5)}$, $\tilde{\lambda}_1 = \tilde{\lambda}_2 = cs_{U^c}n^{-2/(p_c+5)}$ for different values of c to check the sensitivity of our test to the choice of bandwidth. Note that we apply undersmoothing bandwidth sequences $(h, \lambda_1, \lambda_2)$ and purposely set $(\tilde{h}, \tilde{\lambda}_1, \tilde{\lambda}_2)$ to be larger than $(h, \lambda_1\lambda_2)$ elementwise because they are required for the asymptotic theory of our test. We have tried three values for c : 0.5, 1 and 2 and found that our test is not sensitive to the choice of c. To save space, we only focus on the case where $c = 1$ in the following analysis.

We consider two sample sizes $n = 100$ and 200, set $\delta_1 = \delta_2 = \delta$, and consider three values of δ : 0, 0.5, and 1. We choose three different values of τ, namely, 0.32, 0.75, and 2.07, which correspond to low, moderate and extremely high degrees of endogeneity, respectively. For each case, we consider 500 replications and 200 bootstrap resamples for each replication.

The results are reported in Table 2. We summarize some important findings from Table 2. First the level of our test generally behaves very well for all three values of τ, all three null hypotheses, and all three nominal levels (1%, 5%, and 10%) under investigation. The only exception occurs for the null hypothesis $\mathbb{H}_{0,2}$ when the degree of endogeneity is extremely high, in which case the test is moderately oversized. Second, the power of our test is reasonably good. As either δ or the sample size increases, we observe an increase of the empirical power. Third, endogeneity does affect the level and power behavior of our test. When endogeneity is very strong ($\tau = 2.07$), our test for $\mathbb{H}_{0,2}$ tends to be oversized in small samples. The power of our test for all three null hypotheses are also adversely affected by such strong endogeneity.

6 An Empirical Example: Estimating the Wage Equation

Labor economists have been devoting a tremendous amount of effort to investigating the causal effect of education on labor market earnings. As Card (2001 p. 1127) suggests, the endogeneity of education in the wage equation might partially explain the continuing interest "in this very difficult task of uncovering the causal effect of education in labor market outcomes." The classical framework of the human capital earnings function due to Mincer (1974) assumes additivity of education and work experience that used as explanatory variables. However, recent studies have questioned the appropriateness of this assumption. In particular, Card (2001) approaches the matter of non-additivity of the explanatory variables by arguing that the returns to education are heterogeneous since the economic benefits of schooling are individual-specific. Becker and Chiswick (1966) are among the authors who maintain that variation in returns to education can partially account for variation over time in aggregate inequality. Card's (2001) claim suggests that a more general functional form of heterogeneity in the returns to education would make the empirical relation between earnings and education even more realistic. Indeed, if, for example, work experience is valued by employers, then one can expect earnings to be increasing in experience for any given level of education. Further, the returns to education may also differ substantially among different groups defined by some individual-specific characteristics, say, marital status. Therefore, we estimate the causal effect of education on earnings in the following functional coefficient model:

$$\log(Y) = g_1(\mathbf{U}) + g_2(\mathbf{U})S + \varepsilon, \tag{6.1}$$

where Y is a measure of individual earnings, S is years of education, and $\mathbf{U}$ is a vector of mixed (both continuous and discrete) variables. Equation (6.1) allows studying not only the direct effects of variables in $\mathbf{U}$ on wage in a flexible way but also the effects of these variables on the return to education. Existing literature has already provided support for a nonlinear relation between wage and work experience (see, for example, Murphy and Welch (1990) and Ullah (1985)). In addition, Card and Lemieux (2001) emphasize that the rising return to education has been more profound in the younger cohorts than in the older ones since the 1980s.

Our goal is to study the empirical relation between earnings and education as presented in (6.1) using our proposed estimator from the previous sections. For this purpose, we use the Australian Longitudinal Survey (ALS) conducted annually since 1984. Specifically, we employ the 1985 wave of the ALS, and consider young Australian women, who reported working and were aged 16 to 25 in 1985. Our sample is constructed using the guidelines from Vella (1994), who is among the first researchers extensively working with this dataset. We follow the empirical analysis from Vella (1994) and choose $\mathbf{U}$ to include a continuous variable – work experience, and four categorical variables for martial status, union membership, government employment, and whether a person is born in Australia.

We follow CDXW (2006) and Das, Newey, and Vella (2003), who rely on findings from Vella (1994), and use an index of labor market attitudes as the instrumental variable for the schooling levels.[6] The

[6] Here, we do not question the credibility of the instrument but take its validity as a maintained assumption in order to illustrate the proposed estimation method.

Table 3: The ALS Sample Characteristics

Variable	Source	Mean	St. Dev.	Min	Max
Born in Australia	B.3	.862	.345	0	1
Married	A.7	.183	.386	0	1
Union Member	G.11	.425	.494	0	1
Government Employee	G.10	.286	.452	0	1
Age	A.4	20.718	2.622	15	25
Years of Education	E.3, 5, 8, 12, 14, 21, 23	11.736	1.529	3	16
Years of Experience	F.3-4, 7-10, 31-3, G.23-5	1.489	1.997	0	14
Hourly Wage ($)	G.3-5 and 7-8	6.662	2.579	.375	47.5
Attitudes Index	O.1	1.969	.351	.7	2.8

Notes: The sample is based on the 1985 wave of the Australian Longitudinal Survey (ALS). The sample size is 2049 observations. Column Source provides information about the questions from the ALS, which were used to obtain the variables. Hourly wage is in 1985 dollars. Attitudes Index is constructed using only six out of seven equations about work, social roles and school attitudes of individuals toward working women. Specifically, we exclude question (iii).

ALS includes seven questions about work, social roles and school attitudes of individuals toward working women. Individuals respond to these questions with "(1) strongly agree; (2) agree; (3) don't know; (4) disagree; and (5) strongly disagree". The wording of the questions implies that a response with a higher score indicates more positive attitude towards the schooling of women and their role in the labor market. We use only six out of seven available questions to construct our attitudes index, since questions (ii) and (iii) seem to be very similar to each other and might be repetitive.[7] We sum the responses to the questions we pick, and divide the total by 10. This way our attitudes index can range from 0.6 to 3.0, similar to CDXW (2006).[8] We exclude two observations with reported wage being more than $200 per hour as extreme outliers.[9] The resulting sample consists of 2049 observations. Table 3 reports summary statistics for our sample. Figure 1 plots wage against work experience and years of education. The right panel of Figure 1 suggests that there is a positive relationship between wage and years of education. The left panel describing the relation between wage and work experience is not that straightforward. However, both figures also provide some evidence of a nonlinear nature of the two relationships they present. The peculiar relation between wage and experience is actually not surprising as our sample consists of young adults being 15 to 25 years old.

Without accounting for the endogeneity of education in the wage equation, SCU (2009) estimate the returns to education using the same specification – (6.1), while also allowing for mixed covariates in the model. CDXW (2006) employ the same data set but use a somewhat different model specification:

$$\log(Y) = \mathbf{Z}\boldsymbol{\delta} + g_1(U) + g_2(U)S + \varepsilon, \tag{6.2}$$

where U contains work experience only, and $\mathbf{Z}$ includes the four categorical variables, i.e., indicators for marital status, union membership, government employment, and whether a person is born in Australia. CDXW (2006) exploit a two-step nonparametric procedure to estimate the returns to education in

[7] Questions (ii) and (iii) read "A woman is really fulfilled only when she becomes a mother,"and "Whatever career a women may have, her most important role in life is that of becoming a mother?", respectively. Source: Question O.1 of the 1985 wave of the ALS. We choose Question (ii) over Question (iii).

[8] We point out that we do not know if our index is based on the same questions as the index of CDXW (2006), since we do not know which six questions were used to build the index from CDXW (2006).

[9] The highest hourly wage in the sample after the exclusion of the two outliers is $47.5 per hour.

the context of model (6.2). Cai and Xiong (2010) consider the same data set and model specification as CDXW (2006). However, they use a three-step nonparametric method to estimate this model. We compare our estimates of the return to education with the estimates based on all there existing approaches – the ones from SCU (2009), CDXW (2006) and Cai and Xiong (2010). When doing so, we mainly concentrate on work experience below 8 years. The main reason for our decision is that our sample contains only 8 observations (out of 2049 available) with experience being more or equal to 9 years. We also suspect that the sample used by CDXW (2006) and Cai and Xiong (2010) excludes observations at the high levels of the observed years of experience in our sample, as their sample contains 1996 observations only.[10] Thus, for comparative purposes, we primarily focus on work experience being less than 8 years.

For the ease of presentation of the regression results of model (6.1) we plot wage-experience profiles of different cells defined by a discrete characteristic averaged over other categorical regressors. We use the second order Epanechnikov kernel in our estimation, and choose the bandwidth by both the rule of thumb and LSCV methods discussed in Section 3.3.

Figure 2 reports the estimated $g_1(Experience, :)$ and $g_2(Experience, :)$ of model (6.1) depending on whether a woman is married or not, a union member or not, employed by the government or not, and born in Australia or not averaged over all other categorical variables. We use the rule of thumb bandwidth to obtain Figure 2. Following SCU (2009), we will view $g_1(Experience,\ Individual\ Characteristic, :)$ as the direct effects of experience on wage for a particular characteristic of a woman (averaged over all other categorical variables). At the same time, we can think that $g_2(Experience,\ Individual\ Characteristic, :)$ represents the return to education as a function of experience for a particular individual characteristic. In both profiles – with and without controlling for endogeneity, i.e., profiles using the SCU method and our approach with optimal weight matrix, respectively, we find that the range of $\hat{g}_2$ is positive and nonlinear for all values of experience in our sample. However, the apparent differences between correcting and not correcting for endogeneity are in the magnitude and shape of $\hat{g}_2$. When correcting for endogeneity, the returns to education, on average, are predicted to be higher for most of the observed years of work experience. Also, when correcting for endogeneity, $\hat{g}_2$ is mostly concave, while it is convex for low levels of experience (below about 5 years) and concave for high levels of experience (above 5 years) when we do not correct for endogeneity. Further, we observe that the returns to education are smaller for non-unionized women than for the unionized ones. We also note that the profile of $\hat{g}_1$ when correcting for endogeneity is almost constant, while it is quite nonlinear without the correction. Specifically, the estimated direct effects of the four categorical individual characteristics we are able to control for seem to be close to zero for most of the interval of the observed work experience, when controlling for endogeneity. Without correcting for endogeneity, we do observe some differences both across and within the categories of the four individual characteristics.

Figure 3 plots the estimated $g_1(Experience, :)$ and $g_2(Experience, :)$ of model (6.1) averaged over all categorical variables. We use the rule of thumb bandwidth to obtain Figure 3. Similarly to Figure 2, we notice that the direct effect of work experience on wage is almost constant and close to zero for high levels of experience when controlling for endogeneity. At the same time, when correcting for endogeneity, the derivative of return to education as a function of experience changes over its range,

[10] Since there is no descriptive statistics provided by either CDXW (2006) or Cai and Xiong (2010) we cannot verify our guess.

being negative at high levels of experience (above about 8 years) and positive at low and (most of) middle levels of experience (below 8 years). In other words, while the marginal returns to education are positive, the returns themselves decline in experience for high levels of the observed years of experience. To the contrary, when we do not correct for endogeneity, it is the other way around: the returns to education decrease in experience for low levels of experience (below about 5 years) and increase for high levels of experience (above 5 years).

While we do not observe overly drastic distinctions in the results based on our approach and approaches by CDXW (2006) and Cai and Xiong (2010), we do see some notable differences across the three approaches. First, findings by Cai and Xiong's (2010) and CDXW (2006)[11] indicate that the returns to education may vary from (roughly) 15 to 22% and 16.5 to 30%, respectively. Our findings reported in Figure 3 suggest that the returns to education may vary from about 12 to 18%.[12] Clearly, our range is tighter than the other two ranges suggested, and the middle point for the range obtained using our approach is (at least) 2.5 percentage points smaller than the middle points for the other two intervals. Second, the shapes of the estimated g_2 from the three methods being compared – our approach, CDXW (2006) and Cai and Xiong (2010) – are somewhat different, as well. Contrary to CDXW (2006), our approach suggests that the returns to education start declining after (about) 8 years of experience, which would be more compatible with the shape of the estimated g_2 from Cai and Xiong (2010) for the high levels of observed experience. However, in sharp contrast to Cai and Xiong (2010), our results predict a different behavior of the estimated g_2 for the low levels of observed experience. Cai and Xiong's (2010) results are indicative of the sharply declining returns to education for experience below (about) 3 years. We suggest that the returns to education are increasing for that interval of observed work experience.

Figures 4 and 5 provide the same information as Figures 2 and 3, respectively, when the LSCV method is used instead of the rule of thumb to obtain the bandwidths for findings in Figures 4 and 5. The LSCV method provides very similar results to the ones based on the rule of thumb approach to the choice of bandwidth.

Finally, using the specification tests introduced in Section 4, we test the hypothesis that g_1, g_2 or both are constant over the four categorical variables and experience. We calculate the normalized test statistics for the following three null hypotheses:

$$
\begin{aligned}
\mathbb{H}_{0,1} &: \quad g_1(\mathbf{U}_i) = \theta_1 \text{ a.s.}, \\
\mathbb{H}_{0,2} &: \quad g_2(\mathbf{U}_i) = \theta_2 \text{ a.s.}, \\
\mathbb{H}_{0,12} &: \quad (g_1(\mathbf{U}_i), g_2(\mathbf{U}_i)) = (\theta_1, \theta_2) \text{ a.s.},
\end{aligned}
$$

for some unknown parameters θ_1 and θ_2, where $\mathbf{U}$ contains work experience and four categorical variables for martial status, union membership, government employment, and whether a person is born in Australia. Using the rule of thumb approach and 500 replications, the obtained p-values for the three considered hull hypotheses are 0.180, 0.006, and 0.006, respectively. These results imply that we cannot reject $\mathbb{H}_{0,1}$ at any conventional level. Clearly, this finding is not surprising, given that $\hat{g}_1$ obtained when correcting for endogeneity seems almost constant and close to zero for a large domain of work experience

[11] See Figure 1 in Cai and Xiong (2010) and Figure 3(a) in CDXW (2006), respectively.

[12] Such high returns to education are actually not surprising as we observe young women being between 15 and 25 years old.

in Figures 2-5. More importantly, both $\mathbb{H}_{0,2}$ and $\mathbb{H}_{0,2}$ can be rejected in favor of a one-sided alternative at 1% level. Therefore, our empirical findings strongly support the discussion of the nonlinear nature of the effect of education on wages from Card (2001).

7 Concluding Remarks

This paper proposes a local linear GMM estimation procedure for functional coefficient IV models where endogenous regressors enter the model linearly, and the functional coefficients contain both continuous and discrete exogenous regressors. We establish the asymptotic normality of the local linear GMM estimator. We also propose tests for the constancy of the functional coefficients. Simulations indicate that our tests perform reasonably well in finite samples. Applications to an Australian Longitudinal Survey data indicate the importance of our estimation and testing procedure in empirical research. Although we have not shown in this paper, we conjecture that the theory of our paper continue to hold when the coefficients are the function of some of the continuous variables which are unobserved, but as long as their estimated values converge to their truth at sufficiently fast rate. Also, it will be a subject of the future study to develop results for the case where the coefficients are functions of endogenous variables.

Appendix

A Proof of the Main Results

A.1 Proof of Theorem 3.1

For notational simplicity, in this proof we suppress the dependence of $\boldsymbol{\xi}$, $\mathbf{K}_{\mathbf{h}\boldsymbol{\lambda}}$, $\mathbf{Q}_{\mathbf{h}}$, and $\boldsymbol{\Psi}_n$ on $\mathbf{u}$. Let $d_{\mathbf{U}_i^d\mathbf{u}^d} = \sum_{t=1}^{p_d} 1\{U_{i,t}^d \neq u_t^d\}$, indicating the number of disagreeing components between $\mathbf{U}_i^d = (U_{i,1}^d, \cdots, U_{i,p_d}^d)'$ and $\mathbf{u}^d = (u_1^d, \cdots, u_{p_d}^d)'$. Let $G_{i,j} = G_{i,j}(\mathbf{u}) = [g_j(\mathbf{U}_i^c, \mathbf{U}_i^d) - g_j(\mathbf{u}^c, \mathbf{u}^d) - \dot{g}_j(\mathbf{u}^c, \mathbf{u}^d)'(\mathbf{U}_i^c - \mathbf{u}^c)]X_{i,j}$, and $\mathbf{G}_i = \mathbf{G}_i(\mathbf{u}) = (G_{i,1}(\mathbf{u}), \cdots, G_{i,d}(\mathbf{u}))'$. Let $R_i = R_i(\mathbf{u}) = \mathbf{G}_i(\mathbf{u})'\mathbf{X}_i = \sum_{j=1}^d G_{i,j}(\mathbf{u}) X_{i,j}$. Then $Y_i = \sum_{j=1}^d [g_j(\mathbf{u}^c, \mathbf{u}^d) - \dot{g}_j(\mathbf{u}^c, \mathbf{u}^d)'(\mathbf{U}_i^c - \mathbf{u}^c)]X_{i,j} + \varepsilon_i + R_i = \boldsymbol{\xi}_{i,\mathbf{u}}'\boldsymbol{\alpha} + \varepsilon_i + R_i$, where $\boldsymbol{\xi}_{i,\mathbf{u}}$ and $\boldsymbol{\alpha} = (g_1, \cdots, g_d, \dot{g}_1', \cdots, \dot{g}_d')'$ are defined after eq. (2.8). Let $\boldsymbol{\varepsilon} = (\varepsilon_1, \cdots, \varepsilon_n)'$ and $\mathbf{R} = \mathbf{R}(\mathbf{u}) = (R_1(\mathbf{u}), \cdots, R_n(\mathbf{u}))'$. Then we have the following bias-variance decomposition:

$$\begin{aligned}
\mathbf{H}\left(\widehat{\boldsymbol{\alpha}}_{\boldsymbol{\Psi}_n}(\mathbf{u}) - \boldsymbol{\alpha}(\mathbf{u})\right) &= \left(\mathbf{H}^{-1}\boldsymbol{\xi}'\mathbf{K}_{\mathbf{h}\boldsymbol{\lambda}}\mathbf{Q}_{\mathbf{h}}\boldsymbol{\Psi}_n^{-1}\mathbf{Q}_{\mathbf{h}}'\mathbf{K}_{\mathbf{h}\boldsymbol{\lambda}}\boldsymbol{\xi}\mathbf{H}^{-1}\right)^{-1}\mathbf{H}^{-1}\boldsymbol{\xi}'\mathbf{K}_{\mathbf{h}\boldsymbol{\lambda}}\mathbf{Q}_{\mathbf{h}}\boldsymbol{\Psi}_n^{-1}\mathbf{Q}_{\mathbf{h}}'\mathbf{K}_{\mathbf{h}\boldsymbol{\lambda}}\mathbf{R} \\
&\quad + \left(\mathbf{H}^{-1}\boldsymbol{\xi}'\mathbf{K}_{\mathbf{h}\boldsymbol{\lambda}}\mathbf{Q}_{\mathbf{h}}\boldsymbol{\Psi}_n^{-1}\mathbf{Q}_{\mathbf{h}}'\mathbf{K}_{\mathbf{h}\boldsymbol{\lambda}}\boldsymbol{\xi}\mathbf{H}^{-1}\right)^{-1}\mathbf{H}^{-1}\boldsymbol{\xi}'\mathbf{K}_{\mathbf{h}\boldsymbol{\lambda}}\mathbf{Q}_{\mathbf{h}}\boldsymbol{\Psi}_n^{-1}\mathbf{Q}_{\mathbf{h}}'\mathbf{K}_{\mathbf{h}\boldsymbol{\lambda}}\boldsymbol{\varepsilon} \\
&= \left(\boldsymbol{\Phi}_n'\boldsymbol{\Psi}_n^{-1}\boldsymbol{\Phi}_n\right)^{-1}\boldsymbol{\Phi}_n'\boldsymbol{\Psi}_n^{-1}\mathcal{B}_n + \left(\boldsymbol{\Phi}_n'\boldsymbol{\Psi}_n^{-1}\boldsymbol{\Phi}_n\right)^{-1}\boldsymbol{\Phi}_n'\boldsymbol{\Psi}_n^{-1}\mathcal{V}_n, \qquad \text{(A.1)}
\end{aligned}$$

where $\boldsymbol{\Phi}_n = \boldsymbol{\Phi}_n(\mathbf{u}) = n^{-1}\mathbf{Q}_{\mathbf{h}}'\mathbf{K}_{\mathbf{h}\boldsymbol{\lambda}}\boldsymbol{\xi}\mathbf{H}^{-1}$, $\mathcal{B}_n = \mathcal{B}_n(\mathbf{u}) = n^{-1}\mathbf{Q}_{\mathbf{h}}'\mathbf{K}_{\mathbf{h}\boldsymbol{\lambda}}\mathbf{R}$, and $\mathcal{V}_n = \mathcal{V}_n(\mathbf{u}) = n^{-1}\mathbf{Q}_{\mathbf{h}}'\mathbf{K}_{\mathbf{h}\boldsymbol{\lambda}}\boldsymbol{\varepsilon}$. We prove Theorem 3.1 by proving the following three lemmata.

Lemma A.1 $\boldsymbol{\Phi}_n(\mathbf{u}) = \boldsymbol{\Phi}(\mathbf{u}) + o_P(1)$, *where* $\boldsymbol{\Phi}(\mathbf{u})$ *is defined in (3.1).*

Proof. Noting that $\mathbf{Q}_{\mathbf{h},i\mathbf{u}} = \begin{pmatrix} \mathbf{Z}_i^a \\ \mathbf{Z}_i^a \otimes \boldsymbol{\eta}_i(\mathbf{u}^c) \end{pmatrix}$ with $\boldsymbol{\eta}_i(\mathbf{u}^c) = (\mathbf{U}_i^c - \mathbf{u}^c)/\mathbf{h}$, we have

$$\begin{aligned}
\boldsymbol{\Phi}_n(\mathbf{u}) &= \frac{1}{n}\sum_{i=1}^n K_{\mathbf{h}\boldsymbol{\lambda},i\mathbf{u}}\mathbf{Q}_{\mathbf{h},i\mathbf{u}}\boldsymbol{\xi}_{i,\mathbf{u}}'\mathbf{H}^{-1} \\
&= \frac{1}{n}\sum_{i=1}^n K_{\mathbf{h}\boldsymbol{\lambda},i\mathbf{u}}\begin{pmatrix} \mathbf{Z}_i^a \\ \mathbf{Z}_i^a \otimes \boldsymbol{\eta}_i(\mathbf{u}^c) \end{pmatrix}\left(\mathbf{X}_i' \quad \mathbf{X}_i' \otimes (\mathbf{U}_i^c - \mathbf{u}^c)'\right)\mathbf{H}^{-1} = \begin{pmatrix} \underset{q\times d}{\boldsymbol{\Phi}_{n,11}} & \underset{q\times p_c d}{\boldsymbol{\Phi}_{n,12}} \\ \underset{qp_c\times d}{\boldsymbol{\Phi}_{n,21}} & \underset{qp_c\times p_c d}{\boldsymbol{\Phi}_{n,22}} \end{pmatrix},
\end{aligned}$$

where $\boldsymbol{\Phi}_{n,11} = \frac{1}{n}\sum_{i=1}^n K_{\mathbf{h}\boldsymbol{\lambda},i\mathbf{u}}\mathbf{Z}_i^a\mathbf{X}_i'$, $\boldsymbol{\Phi}_{n,12} = \frac{1}{n}\sum_{i=1}^n K_{\mathbf{h}\boldsymbol{\lambda},i\mathbf{u}}\mathbf{Z}_i^a[\mathbf{X}_i' \otimes \boldsymbol{\eta}_i(\mathbf{u}^c)']$, $\boldsymbol{\Phi}_{n,21} = \frac{1}{n}\sum_{i=1}^n K_{\mathbf{h}\boldsymbol{\lambda},i\mathbf{u}}[\mathbf{Z}_i^a \otimes \boldsymbol{\eta}_i(\mathbf{u}^c)]\mathbf{X}_i'$, and $\boldsymbol{\Phi}_{n,22} = \frac{1}{n}\sum_{i=1}^n K_{\mathbf{h}\boldsymbol{\lambda},i\mathbf{u}}\mathbf{Z}_i^a\mathbf{X}_i' \otimes [\boldsymbol{\eta}_i(\mathbf{u}^c)\boldsymbol{\eta}_i(\mathbf{u}^c)']$. It suffices to show that $\boldsymbol{\Phi}_{n,lj} = \boldsymbol{\Phi}_{lj}(\mathbf{u}) + o_p(1)$ for $l, j = 1, 2$, where $\boldsymbol{\Phi}_{lj}(\mathbf{u})$ denotes the (l, j) block of the block diagonal matrix $\boldsymbol{\Phi}(\mathbf{u})$.

By Assumptions A1-A3,

$$\begin{aligned}
E[\boldsymbol{\Phi}_{n,11}] &= E[\mathbf{Z}_i^a\mathbf{X}_i' K_{\mathbf{h}\boldsymbol{\lambda},i\mathbf{u}}] \\
&= E[\mathbf{Z}_i^a\mathbf{X}_i' W_{\mathbf{h},i\mathbf{u}^c} | d_{\mathbf{U}_i^d\mathbf{u}^d} = 0]p(\mathbf{u}^d) + \sum_{s=1}^{p_d} E[\mathbf{Z}_i^a\mathbf{X}_i' W_{\mathbf{h},i\mathbf{u}^c} L_{\boldsymbol{\lambda},i\mathbf{u}^d} | d_{\mathbf{U}_i^d\mathbf{u}^d} = s]p(d_{\mathbf{U}_i^d\mathbf{u}^d} = s) \\
&= E[\boldsymbol{\Omega}(\mathbf{U}_i^c, \mathbf{U}_i^d) W_{\mathbf{h},i\mathbf{u}^c} | d_{\mathbf{U}_i^d\mathbf{u}^d} = 0]p(\mathbf{u}^d) + O(\|\boldsymbol{\lambda}\|) \\
&= \int \boldsymbol{\Omega}(\mathbf{u}^c + \mathbf{h}\odot\mathbf{t}, \mathbf{u}^d) f_{\mathbf{U}}(\mathbf{u}^c + \mathbf{h}\odot\mathbf{t}, \mathbf{u}^d) W(\mathbf{t}) d\mathbf{t} + O(\|\boldsymbol{\lambda}\|) \\
&= \boldsymbol{\Omega}(\mathbf{u}) f_{\mathbf{U}}(\mathbf{u}) + O\left(\|\mathbf{h}\|^2 + \|\boldsymbol{\lambda}\|\right). \qquad \text{(A.2)}
\end{aligned}$$

Define two column vectors $\omega_1 \in \mathbb{R}^q$ and $\omega_2 \in \mathbb{R}^d$ such that $\|\omega_l\| = 1$ for $l = 1, 2$. Then it is easy to show that $\mathrm{Var}(\omega_1' \mathbf{\Phi}_{n,11} \omega_2) = \frac{1}{n} \mathrm{Var}(\omega_1' \mathbf{Z}_i^a \mathbf{X}_i' \omega_2 K_{\mathbf{h}\boldsymbol{\lambda},i\mathbf{u}}) = O((n\mathbf{h}!)^{-1}) = o(1)$. It follows by the Chebyshev's inequality that $\mathbf{\Phi}_{n,11} = \mathbf{\Omega}(\mathbf{u}) f_{\mathbf{U}}(\mathbf{u}) + o_P(1)$. Similarly,

$$
\begin{aligned}
\mathbf{\Phi}_{n,22} &= E[\mathbf{\Phi}_{n,22}] + O_P\left((n\mathbf{h}!)^{-1/2}\right) \\
&= E\left[\mathbf{Z}_i^a \mathbf{X}_i' \otimes ((\mathbf{U}_i^c - \mathbf{u}^c)/\mathbf{h})((\mathbf{U}_i^c - \mathbf{u}^c)/\mathbf{h})' K_{\mathbf{h}\boldsymbol{\lambda},i\mathbf{u}}\right] + O_P\left((n\mathbf{h}!)^{-1/2}\right) \\
&= \int \left[\mathbf{\Omega}\left(\mathbf{u}^c + \mathbf{h} \odot \mathbf{t}, \mathbf{u}^d\right) \otimes \mathbf{t}\mathbf{t}'\right] f_{\mathbf{U}}\left(\mathbf{u}^c + \mathbf{h} \odot \mathbf{t}, \mathbf{u}^d\right) W(\mathbf{t})\, d\mathbf{t} + O_P\left(\|\boldsymbol{\lambda}\| + (n\mathbf{h}!)^{-1/2}\right) \\
&= \mu_{2,1}\left[\mathbf{\Omega}(\mathbf{u}) \otimes \mathbf{I}_{p_c}\right] f_{\mathbf{U}}(\mathbf{u}) + o_P(1).
\end{aligned}
$$

By the same token, $\mathbf{\Phi}_{n,12} = o_P(1)$, and $\mathbf{\Phi}_{n,21} = o_P(1)$. This completes the proof. ■

Lemma A.2 $\sqrt{n\mathbf{h}!}\mathcal{B}_n(\mathbf{u}) = \sqrt{n\mathbf{h}!}\mathbf{H}\mathbf{B}(\mathbf{u}; \mathbf{h}, \boldsymbol{\lambda}) + o_P(1)$, *where* $\mathbf{B}(\mathbf{u}; \mathbf{h}, \boldsymbol{\lambda})$ *is defined in (3.3).*

Proof. Write $\sqrt{n\mathbf{h}!}\mathcal{B}_n(\mathbf{u}) = \frac{1}{n}\sum_{i=1}^n \sqrt{n\mathbf{h}!} K_{\mathbf{h}\boldsymbol{\lambda},i\mathbf{u}} \mathbf{Q}_{\mathbf{h},i\mathbf{u}} \mathbf{R}_i = \frac{1}{n}\sum_{i=1}^n \boldsymbol{\varsigma}_i$, where

$$
\begin{aligned}
\boldsymbol{\varsigma}_i &= \sqrt{n\mathbf{h}!} \sum_{j=1}^d \left[g_j(\mathbf{U}_i^c, \mathbf{U}_i^d) - g_j(\mathbf{u}^c, \mathbf{u}^d) - \dot{g}_j(\mathbf{u}^c, \mathbf{u}^d)'(\mathbf{U}_i^c - \mathbf{u}^c)\right] \begin{pmatrix} \mathbf{Z}_i^a X_{i,j} \\ \mathbf{Z}_i^a X_{i,j} \otimes \boldsymbol{\eta}_i(\mathbf{u}^c) \end{pmatrix} K_{\mathbf{h}\boldsymbol{\lambda},i\mathbf{u}} \\
&= \sqrt{n\mathbf{h}!} \begin{pmatrix} \mathbf{Z}_i^a \mathbf{X}_i' \mathbf{G}_i \\ (\mathbf{Z}_i^a \mathbf{X}_i' \mathbf{G}_i) \otimes \boldsymbol{\eta}_i(\mathbf{u}^c) \end{pmatrix} K_{\mathbf{h}\boldsymbol{\lambda},i\mathbf{u}}.
\end{aligned}
$$

It follows that $\sqrt{n\mathbf{h}!}E[\mathcal{B}_n(\mathbf{u})] = E(\boldsymbol{\varsigma}_i) = E(\boldsymbol{\varsigma}_i | d_{\mathbf{U}_i^d \mathbf{u}^d} = 0) p(\mathbf{u}^d) + E(\boldsymbol{\varsigma}_i | d_{\mathbf{U}_i^d \mathbf{u}^d} = 1)\ P(d_{\mathbf{U}_i^d \mathbf{u}^d} = 1) + O(\sqrt{n\mathbf{h}!}\,\|\boldsymbol{\lambda}\|^2) \equiv \mathbf{b}_{n,1} + \mathbf{b}_{n,2} + o(1)$.

On the set $\left\{\mathbf{U}_i^d = \mathbf{u}^d,\ W_{\mathbf{h},i\mathbf{u}^c} > 0\right\}$, $g_j(\mathbf{U}_i^c, \mathbf{U}_i^d) - g_j(\mathbf{u}^c, \mathbf{u}^d) - \dot{g}_j(\mathbf{u}^c, \mathbf{u}^d)'(\mathbf{U}_i^c - \mathbf{u}^c) = \frac{1}{2} A_{i,j}(\mathbf{u}) + o(\|\mathbf{h}\|^2)$, where $A_{i,j}(\mathbf{u}) = (\mathbf{U}_i^c - \mathbf{u}^c)' \ddot{g}_j(\mathbf{u})(\mathbf{U}_i^c - \mathbf{u}^c)$ and $\ddot{g}_j(\mathbf{u}) = \partial \dot{g}_j(\mathbf{u}) / \partial \mathbf{u}^{c\prime}$. Let $\mathbf{A}_i(\mathbf{u}) = (A_{i,1}(\mathbf{u}), \cdots, A_{i,d}(\mathbf{u}))'$. Recall $\mathbf{A}(\mathbf{u}) = (\sum_{s=1}^p h_s^2 g_{1,ss}(\mathbf{u}), \cdots, \sum_{s=1}^p h_s^2 g_{d,ss}(\mathbf{u}))'$ and $\dot{\mathbf{g}}(\mathbf{u}) = (\dot{g}_1(\mathbf{u})', \cdots, \dot{g}_d(\mathbf{u})')'$. Then we have

$$
\begin{aligned}
\mathbf{b}_{n,1} &= \frac{1}{2}\sqrt{n\mathbf{h}!}\ E_0\left[\begin{pmatrix} \mathbf{Z}_i^a \mathbf{X}_i' \mathbf{A}_i(\mathbf{u}) \\ (\mathbf{Z}_i^a \mathbf{X}_i' A_i(\mathbf{u})) \otimes \boldsymbol{\eta}_i(\mathbf{u}^c) \end{pmatrix} W_{\mathbf{h},i\mathbf{u}^c}\right] p(\mathbf{u}^d) + o(\sqrt{n\mathbf{h}!}\,\|\mathbf{h}\|^2) \\
&= \frac{1}{2}\sqrt{n\mathbf{h}!}\ E_0\left[\begin{pmatrix} \mathbf{\Omega}(\mathbf{U}_i) \mathbf{A}_i(\mathbf{u}) \\ (\mathbf{\Omega}(\mathbf{U}_i) A_i(\mathbf{u})) \otimes \boldsymbol{\eta}_i(\mathbf{u}^c) \end{pmatrix} W_{\mathbf{h},i\mathbf{u}^c}\right] p(\mathbf{u}^d) + o(1) \\
&= \frac{\sqrt{n\mathbf{h}!}\mu_{2,1}}{2} \begin{pmatrix} f_{\mathbf{U}}(\mathbf{u}) \mathbf{\Omega}(\mathbf{u}) \mathbf{A}(\mathbf{u}) \\ 0 \end{pmatrix} + o(1),
\end{aligned}
$$

and

$$
\begin{aligned}
\mathbf{b}_{n,2} &= \sqrt{n\mathbf{h}!} E_1 \left\{ \sum_{j=1}^d [g_j(\mathbf{U}_i) - g_j(\mathbf{u}) - \dot{g}_j(\mathbf{u})'(\mathbf{U}_i^c - \mathbf{u}^c)] \begin{pmatrix} \mathbf{Z}_i^a X_{i,j} \\ \mathbf{Z}_i^a X_{i,j} \otimes \boldsymbol{\eta}_i(\mathbf{u}^c) \end{pmatrix} K_{\mathbf{h}\boldsymbol{\lambda},i\mathbf{u}} \right\} p_1 \\
&= \sqrt{n\mathbf{h}!} E_1 \left[\sum_{j=1}^d \begin{pmatrix} \mathbf{Z}_i^a \mathbf{X}_i' \mathbf{G}_i \\ (\mathbf{Z}_i^a \mathbf{X}_i' \mathbf{G}_i) \otimes \boldsymbol{\eta}_i(\mathbf{u}^c) \end{pmatrix} K_{\mathbf{h}\boldsymbol{\lambda},i\mathbf{u}} \right] p_1
\end{aligned}
$$

$$
\begin{aligned}
&= \sqrt{n\mathbf{h}!}E_1\left[\left(\begin{array}{c}\mathbf{\Omega}(\mathbf{U}_i)\left[\mathbf{g}(\mathbf{U}_i)-\mathbf{g}(\mathbf{u})\right]-\left(\mathbf{\Omega}(\mathbf{U}_i)\otimes\boldsymbol{\eta}_i(\mathbf{u}^c)'\right)\dot{\mathbf{g}}(\mathbf{u})\\ \left(\mathbf{\Omega}(\mathbf{U}_i)\left[\mathbf{g}(\mathbf{U}_i)-\mathbf{g}(\mathbf{u})\right]\right)\otimes\boldsymbol{\eta}_i(\mathbf{u}^c)-\left(\mathbf{\Omega}(\mathbf{U}_i)\otimes\left[\boldsymbol{\eta}_i(\mathbf{u}^c)\boldsymbol{\eta}_i(\mathbf{u}^c)'\right]\right)\dot{\mathbf{g}}(\mathbf{u})\end{array}\right)K_{\mathbf{h}\boldsymbol{\lambda},i\mathbf{u}}\right]p_1\\
&\quad +o(1)\\
&= \sqrt{n\mathbf{h}!}\sum_{\widetilde{\mathbf{u}}^d\in\mathcal{D}}\sum_{s=1}^{p_d}\lambda_s I_s\left(\mathbf{u}^d,\widetilde{\mathbf{u}}^d\right)f_{\mathbf{U}}\left(\mathbf{u}^c,\widetilde{\mathbf{u}}^d\right)\left(\begin{array}{c}\mathbf{\Omega}\left(\mathbf{u}^c,\widetilde{\mathbf{u}}^d\right)\left(\mathbf{g}\left(\mathbf{u}^c,\widetilde{\mathbf{u}}^d\right)-\mathbf{g}\left(\mathbf{u}^c,\mathbf{u}^d\right)\right)\\ -\mu_{2,1}\left(\mathbf{\Omega}\left(\mathbf{u}^c,\widetilde{\mathbf{u}}^d\right)\otimes\mathbf{I}_{p_c}\right)\dot{\mathbf{g}}\left(\mathbf{u}^c,\mathbf{u}^d\right)\end{array}\right)+o(1),
\end{aligned}
$$

where $E_l\{\cdot\} = E\{\cdot|d_{\mathbf{U}_i^d\mathbf{u}^d} = l\}$ for $l = 0$ and 1, and $p_1 = P(d_{\mathbf{U}_i^d\mathbf{u}^d} = 1)$.

Consequently, $\sqrt{n\mathbf{h}!}E[\mathcal{B}_n(\mathbf{u})] = \sqrt{n\mathbf{h}!}\mathbf{B}(\mathbf{u};\mathbf{h},\boldsymbol{\lambda}) + o(1)$, where $\mathbf{B}(\mathbf{u};\mathbf{h},\boldsymbol{\lambda})$ is defined in (3.3). By straightforward calculations, we can show that $\mathrm{Var}(\sqrt{n\mathbf{h}!}\mathcal{B}_n(\mathbf{u})) = O(\|\mathbf{h}\|^2 + \|\boldsymbol{\lambda}\|) = o(1)$. The conclusion then follows by the Chebyshev's inequality. ■

Lemma A.3 $\sqrt{n\mathbf{h}!}\mathcal{V}_n(\mathbf{u}) = n^{-1/2}(\mathbf{h}!)^{1/2}\sum_{i=1}^n\left(\begin{array}{c}\mathbf{Z}_i^a\varepsilon_i\\ (\mathbf{Z}_i^a\varepsilon_i)\otimes((\mathbf{U}_i^c-\mathbf{u}^c)/\mathbf{h})\end{array}\right)K_{\mathbf{h}\boldsymbol{\lambda},i\mathbf{u}} \xrightarrow{d} N(0,\boldsymbol{\Upsilon}(\mathbf{u}))$, *where* $\boldsymbol{\Upsilon}(\mathbf{u})$ *is defined in (3.2).*

Proof. Let $\mathbf{c}$ be a unit vector on $\mathbb{R}^{q(p_c+1)}$. Let $\zeta_i = (\mathbf{h}!)^{1/2}\mathbf{c}'\left(\begin{array}{c}\mathbf{Z}_i^a\varepsilon_i\\ (\mathbf{Z}_i^a\varepsilon_i)\otimes\boldsymbol{\eta}_i(\mathbf{u}^c)\end{array}\right)K_{\mathbf{h}\boldsymbol{\lambda},i\mathbf{u}}$. By the Cramér-Wold device, it suffices to prove $\sqrt{n\mathbf{h}!}\mathcal{V}_n(\mathbf{u}) = n^{-1/2}\sum_{i=1}^n\zeta_i \xrightarrow{d} N(0,\mathbf{c}'\boldsymbol{\Upsilon}\mathbf{c})$. By the law of iterated expectations, $E(\zeta_i) = 0$. Now by arguments similar to those used in the proof of Lemma A.1,

$$
\begin{aligned}
&\mathrm{Var}\left(\sqrt{n\mathbf{h}!}\mathcal{V}_n(\mathbf{u})\right) = \mathrm{Var}(\zeta_1)\\
&= \mathbf{h}!\mathbf{c}'E\left\{\left(\begin{array}{cc}\mathbf{Z}_i^a\mathbf{Z}_i^{a\prime}\varepsilon_i^2 & \mathbf{Z}_i^a\left(\mathbf{Z}_i^{a\prime}\otimes\boldsymbol{\eta}_i(\mathbf{u}^c)'\right)\varepsilon_i^2\\ \left(\mathbf{Z}_i^a\otimes\boldsymbol{\eta}_i(\mathbf{u}^c)\right)\mathbf{Z}_i^{a\prime}\varepsilon_i^2 & \left(\mathbf{Z}_i^a\mathbf{Z}_i^{a\prime}\right)\otimes\left(\boldsymbol{\eta}_i(\mathbf{u}^c)\boldsymbol{\eta}_i(\mathbf{u}^c)'\right)\varepsilon_i^2\end{array}\right)K_{\mathbf{h}\boldsymbol{\lambda},i\mathbf{u}}^2\right\}\mathbf{c}\\
&= \mathbf{h}!\mathbf{c}'E\left\{\left(\begin{array}{cc}\mathbf{Z}_i^a\mathbf{Z}_i^{a\prime}\sigma^2(\mathbf{Z}_i^a,\mathbf{U}_i) & \mathbf{Z}_i^a\left(\mathbf{Z}_i^{a\prime}\otimes\boldsymbol{\eta}_i(\mathbf{u}^c)'\right)\sigma^2(\mathbf{Z}_i^a,\mathbf{U}_i)\\ \left(\mathbf{Z}_i^a\otimes\boldsymbol{\eta}_i(\mathbf{u}^c)\right)\mathbf{Z}_i^{a\prime}\sigma^2(\mathbf{Z}_i^a,\mathbf{U}_i) & \left(\mathbf{Z}_i^a\mathbf{Z}_i^{a\prime}\right)\otimes\left(\boldsymbol{\eta}_i(\mathbf{u}^c)\boldsymbol{\eta}_i(\mathbf{u}^c)'\right)\sigma^2(\mathbf{Z}_i^a,\mathbf{U}_i)\end{array}\right)K_{\mathbf{h}\boldsymbol{\lambda},i\mathbf{u}}^2\right\}\mathbf{c}\\
&= \mathbf{h}!\mathbf{c}'E\left\{\left(\begin{array}{cc}\mathbf{\Omega}^*(\mathbf{U}_i) & \mathbf{\Omega}^*(\mathbf{U}_i)\otimes\boldsymbol{\eta}_i(\mathbf{u}^c)'\\ \mathbf{\Omega}^*(\mathbf{U}_i)\otimes\boldsymbol{\eta}_i(\mathbf{u}^c) & \mathbf{\Omega}^*(\mathbf{U}_i)\otimes\left(\boldsymbol{\eta}_i(\mathbf{u}^c)\boldsymbol{\eta}_i(\mathbf{u}^c)'\right)\end{array}\right)K_{\mathbf{h}\boldsymbol{\lambda},i\mathbf{u}}^2\right\}\mathbf{c}\\
&= \mathbf{c}'\boldsymbol{\Upsilon}\mathbf{c} + o(1).
\end{aligned}
$$

Thus $\mathrm{Var}(\sqrt{n\mathbf{h}!}\mathcal{V}_n(\mathbf{u})) \to \mathbf{c}'\boldsymbol{\Upsilon}\mathbf{c}$ as $n\to\infty$. It is standard to check the Liapounov condition holds (see, e.g., Li and Racine (2007)). ■

By Lemmas A.1-A.3 and the Slutsky lemma,

$$
\begin{aligned}
&\sqrt{n\mathbf{h}!}\left(\mathbf{H}\left(\widehat{\boldsymbol{\alpha}}_{\boldsymbol{\Psi}_n}(\mathbf{u})-\boldsymbol{\alpha}(\mathbf{u})\right)-\left(\mathbf{\Phi}'\boldsymbol{\Psi}^{-1}\mathbf{\Phi}\right)^{-1}\mathbf{\Phi}'\boldsymbol{\Psi}^{-1}\mathbf{B}(\mathbf{u};\mathbf{h},\boldsymbol{\lambda})\right)\\
= &\left(\mathbf{\Phi}_n'\boldsymbol{\Psi}_n^{-1}\mathbf{\Phi}_n\right)^{-1}\mathbf{\Phi}_n'\boldsymbol{\Psi}_n^{-1}\sqrt{n\mathbf{h}!}\mathcal{V}_n(\mathbf{u}) + \left(\mathbf{\Phi}_n'\boldsymbol{\Psi}_n^{-1}\mathbf{\Phi}_n\right)^{-1}\mathbf{\Phi}_n'\boldsymbol{\Psi}_n^{-1}\sqrt{n\mathbf{h}!}\mathcal{B}_n(\mathbf{u})\\
&-\left(\mathbf{\Phi}'\boldsymbol{\Psi}^{-1}\mathbf{\Phi}\right)^{-1}\mathbf{\Phi}'\boldsymbol{\Psi}^{-1}\sqrt{n\mathbf{h}!}\mathbf{B}(\mathbf{u};\mathbf{h},\boldsymbol{\lambda})\\
&\xrightarrow{d} N\left(0,\left(\mathbf{\Phi}'\boldsymbol{\Psi}^{-1}\mathbf{\Phi}\right)^{-1}\mathbf{\Phi}'\boldsymbol{\Psi}^{-1}\boldsymbol{\Upsilon}\boldsymbol{\Psi}^{-1}\mathbf{\Phi}\left(\mathbf{\Phi}'\boldsymbol{\Psi}^{-1}\mathbf{\Phi}\right)^{-1}\right).
\end{aligned}
$$

This completes the proof.

A.2 Proof of the results in Section 4

Proof of Theorems 4.1

Under $\mathbb{H}_0$, we have $\sqrt{n}(\widehat{\boldsymbol{\theta}}_1 - \boldsymbol{\theta}_1) = n^{-1/2}\sum_{i=1}^n \boldsymbol{\Gamma}_{n1}(\mathbf{U}_i)\mathcal{B}_n(\mathbf{U}_i) + n^{-1/2}\sum_{i=1}^n \boldsymbol{\Gamma}_{n1}(\mathbf{U}_i)\mathcal{V}_n(\mathbf{U}_i)$. Noting that $\sup_{\mathbf{u}}\|\boldsymbol{\Phi}_n(\mathbf{u}) - \boldsymbol{\Phi}(\mathbf{u})\| = O_P(\|\mathbf{h}\|^2 + \|\boldsymbol{\lambda}\| + (n\mathbf{h}!/\log n)^{-1/2})$ by strengthening the result in Lemma A.1, following the same lines of proof as in Masry (1996) we can readily show that

$$\sup_{\mathbf{u}}\|\boldsymbol{\Gamma}_{n1}(\mathbf{u})\| = O_P(1) \text{ and } \sup_{\mathbf{u}}\left\|\boldsymbol{\Gamma}_{n1}(\mathbf{u}) - \overline{\boldsymbol{\Gamma}}_1(\mathbf{u})\right\| = O_P\left(\nu_n + \|\mathbf{h}\|^2 + \|\boldsymbol{\lambda}\| + (n\mathbf{h}!/\log n)^{-1/2}\right) \tag{A.3}$$

by Assumption A7, where $\overline{\boldsymbol{\Gamma}}_1(\mathbf{u})$ is defined in (4.5). It is standard to show that

$$\sup_{\mathbf{u}}\|\mathcal{B}_n(\mathbf{u})\| = O_P(\|\mathbf{h}\|^2 + \|\boldsymbol{\lambda}\|) \text{ and } \sup_{\mathbf{u}}\|\mathcal{V}_n(\mathbf{u})\| = O_P((n\mathbf{h}!/\log n)^{-1/2}). \tag{A.4}$$

It follows that $n^{-1/2}\sum_{i=1}^n \boldsymbol{\Gamma}_{n1}(\mathbf{U}_i)\mathcal{B}_n(\mathbf{U}_i) = n^{1/2}O_P(\|\mathbf{h}\|^2 + \|\boldsymbol{\lambda}\|) = o(1)$, and

$$\begin{aligned}
n^{-1/2}\sum_{i=1}^n \boldsymbol{\Gamma}_{n1}(\mathbf{U}_i)\mathcal{V}_n(\mathbf{U}_i) &= n^{-1/2}\sum_{i=1}^n \overline{\boldsymbol{\Gamma}}_1(\mathbf{U}_i)\mathcal{V}_n(\mathbf{U}_i) + n^{-1/2}\sum_{i=1}^n\left[\boldsymbol{\Gamma}_{n1}(\mathbf{U}_i) - \overline{\boldsymbol{\Gamma}}_1(\mathbf{U}_i)\right]\mathcal{V}_n(\mathbf{U}_i) \\
&= A_n + n^{1/2}O_P\left(\left(\nu_n + \|\mathbf{h}\|^2 + \|\boldsymbol{\lambda}\| + (n\mathbf{h}!/\log n)^{-1/2}\right)(n\mathbf{h}!/\log n)^{-1/2}\right) \\
&= A_n + o_P(1)
\end{aligned}$$

where $A_n = n^{-1/2}\sum_{i=1}^n \overline{\boldsymbol{\Gamma}}_1(\mathbf{U}_i)\mathcal{V}_n(\mathbf{U}_i)$. Next, using the notation in (4.9), we have $A_n = n^{-3/2}\sum_{i=1}^n\sum_{j=1}^n \boldsymbol{\varphi}(\boldsymbol{\zeta}_i, \boldsymbol{\zeta}_j) = n^{-1/2}\sum_{i=1}^n \overline{\boldsymbol{\varphi}}(\boldsymbol{\zeta}_i) + o_P(1)$. By direct calculations, $E[\overline{\boldsymbol{\varphi}}(\boldsymbol{\zeta}_i)] = 0$, and

$$\begin{aligned}
E[\overline{\boldsymbol{\varphi}}(\boldsymbol{\zeta}_i)\overline{\boldsymbol{\varphi}}(\boldsymbol{\zeta}_i)'] &= E\left[\int\int \overline{\boldsymbol{\Gamma}}_1(\mathbf{U})\begin{pmatrix} \mathbf{Z}_j^a\mathbf{Z}_j^{a\prime}\varepsilon_j^2 & \mathbf{Z}_j^a\varepsilon_j^2\left((\mathbf{Z}_j^{a\prime})\otimes\boldsymbol{\eta}_j(\widetilde{\mathbf{U}}^c)'\right) \\ \left(\mathbf{Z}_j^a\otimes\boldsymbol{\eta}_j(\mathbf{U}^c)\right)\mathbf{Z}_j^{a\prime}\varepsilon_j^2 & (\mathbf{Z}_j^a\mathbf{Z}_j^{a\prime}\varepsilon_j^2)\otimes\left(\boldsymbol{\eta}_j(\mathbf{U}^c)\boldsymbol{\eta}_j(\widetilde{\mathbf{U}}^c)'\right)\end{pmatrix}\right. \\
&\quad \left.\times\,\overline{\boldsymbol{\Gamma}}_1(\widetilde{\mathbf{U}})'K_{\mathbf{h}\boldsymbol{\lambda},j\mathbf{U}}K_{\mathbf{h}\boldsymbol{\lambda},j\widetilde{\mathbf{U}}}dF_{\boldsymbol{\zeta}}(\boldsymbol{\zeta})\,dF_{\boldsymbol{\zeta}}(\widetilde{\boldsymbol{\zeta}})\right] \\
&= \int\int\int \overline{\boldsymbol{\Gamma}}_1(\mathbf{U})\begin{pmatrix}\boldsymbol{\Omega}^*(\mathbf{U}_j) & \mathbf{0}_{q\times p_cq} \\ \mathbf{0}_{p_cq\times q} & \mathbf{0}_{p_cq\times p_cq}\end{pmatrix}\overline{\boldsymbol{\Gamma}}_1(\widetilde{\mathbf{U}})'W_{\mathbf{h}}\left(\mathbf{U}_j^c - \mathbf{U}^c\right)W_{\mathbf{h}}(\mathbf{U}_j^c - \widetilde{\mathbf{U}}^c) \\
&\quad \times L_{\boldsymbol{\lambda}}\left(\mathbf{U}_j^d - \mathbf{U}^d\right)L_{\boldsymbol{\lambda}}(\mathbf{U}_j^d - \widetilde{\mathbf{U}}^d)dF_{\mathbf{U}}(\mathbf{U})\,dF_{\mathbf{U}}(\widetilde{\mathbf{U}})dF_{\mathbf{U}}(\mathbf{U}_j) + o(1) \\
&= \int \overline{\boldsymbol{\Gamma}}_1(\mathbf{u})\begin{pmatrix}\boldsymbol{\Omega}^*(\mathbf{u}) & \mathbf{0}_{q\times p_cq} \\ \mathbf{0}_{p_cq\times q} & \mathbf{0}_{p_cq\times p_cq}\end{pmatrix}\overline{\boldsymbol{\Gamma}}_1(\mathbf{u})'f_{\mathbf{U}}(\mathbf{u})^2\,dF_{\mathbf{U}}(\mathbf{u})\left(\int\int w(t)\,w(t-s)dtds\right)^{p_c} \\
&= \boldsymbol{\Sigma}_{\boldsymbol{\theta}_1} + o(1).
\end{aligned}$$

One can also verify the Liapounov condition and conclude $A_n \xrightarrow{d} N(0, \boldsymbol{\Sigma}_{\boldsymbol{\theta}_1})$. This completes the proof. ■

Proof of Theorems 4.2 and 4.3

We only prove Theorem 4.3, as the proof of Theorem 4.2 is a special case. Decompose T_n as follows

$$T_n = (\mathbf{h}!)^{1/2}\sum_{i=1}^n\left[\widehat{\mathbf{g}}_1(\mathbf{U}_i) - \overline{\mathbf{g}}_1 + \overline{\mathbf{g}}_1 - \overline{\widehat{\mathbf{g}}}_1\right]'\left[\widehat{\mathbf{g}}_1(\mathbf{U}_i) - \overline{\mathbf{g}}_1 + \overline{\mathbf{g}}_1 - \overline{\widehat{\mathbf{g}}}_1\right] = T_{n1} - T_{n2},$$

where $T_{n1} = (\mathbf{h}!)^{1/2}\sum_{i=1}^n[\widehat{\mathbf{g}}_1(\mathbf{U}_i) - \overline{\mathbf{g}}_1]'[\widehat{\mathbf{g}}_1(\mathbf{U}_i) - \overline{\mathbf{g}}_1]$, and $T_{n2} = n(\mathbf{h}!)^{1/2}[\overline{\mathbf{g}}_1 - \overline{\widehat{\mathbf{g}}}_1]'[\overline{\mathbf{g}}_1 - \overline{\widehat{\mathbf{g}}}_1]$. It suffices to show that under $\mathbb{H}_1(r_n)$, (i) $T_{n1} - B_n - \mu_0 \to N\left(0, \sigma_0^2\right)$ and (ii) $T_{n2} = o_P(1)$.

To prove (i), we further decompose T_{n1} as follows:

$$\begin{aligned}
T_{n1} &= (\mathbf{h}!)^{1/2}\sum_{i=1}^{n}\left[\widehat{\mathbf{g}}_1(\mathbf{U}_i)-\mathbf{g}_1(\mathbf{U}_i)\right]'\left[\widehat{\mathbf{g}}_1(\mathbf{U}_i)-\mathbf{g}_1(\mathbf{U}_i)\right]+(\mathbf{h}!)^{1/2}\sum_{i=1}^{n}\left[\mathbf{g}_1(\mathbf{U}_i)-\overline{\mathbf{g}}\right]'\left[\mathbf{g}_1(\mathbf{U}_i)-\overline{\mathbf{g}}\right]\\
&\quad +2(\mathbf{h}!)^{1/2}\sum_{i=1}^{n}\left[\widehat{\mathbf{g}}_1(\mathbf{U}_i)-\mathbf{g}_1(\mathbf{U}_i)\right]'\left[\mathbf{g}_1(\mathbf{U}_i)-\overline{\mathbf{g}}_1\right]\\
&\equiv T_{n11}+T_{n12}+2T_{n13},\ \text{say.} \qquad \text{(A.5)}
\end{aligned}$$

We study each of the three terms on the right hand side. By (4.7), we can decompose T_{n11} as follows:

$$\begin{aligned}
T_{n11} &= (\mathbf{h}!)^{1/2}\sum_{i=1}^{n}\mathcal{V}_n(\mathbf{U}_i)'\,\boldsymbol{\Gamma}_{n1}(\mathbf{U}_i)'\,\boldsymbol{\Gamma}_{n1}(\mathbf{U}_i)\,\mathcal{V}_n(\mathbf{U}_i)+(\mathbf{h}!)^{1/2}\sum_{i=1}^{n}\mathcal{B}_n(\mathbf{U}_i)'\,\boldsymbol{\Gamma}_{n1}(\mathbf{U}_i)'\,\boldsymbol{\Gamma}_{n1}(\mathbf{U}_i)\,\mathcal{B}_n(\mathbf{U}_i)\\
&\quad +2(\mathbf{h}!)^{1/2}\sum_{i=1}^{n}\mathcal{B}_n(\mathbf{U}_i)'\,\boldsymbol{\Gamma}_{n1}(\mathbf{U}_i)'\,\boldsymbol{\Gamma}_{n1}(\mathbf{U}_i)\,\mathcal{V}_n(\mathbf{U}_i)\\
&= \overline{T}_{n11}+T_{n11}^{(1)}+2T_{n11}^{(2)},\ \text{say.}
\end{aligned}$$

First, $T_{n11}^{(1)}=O_P((n(\mathbf{h}!)^{1/2}(\|\mathbf{h}\|^2+\|\boldsymbol{\lambda}\|)^2)=o_P(1)$ by (A.3), (A.4), and Assumption A8. Applying (A.3), (A.4), the fact that $n^{-1}\sum_{i=1}^{n}\|\mathcal{V}_n(\mathbf{U}_i)\|=O_P((n\mathbf{h}!)^{-1/2})$, and the fact that $\overline{\mathcal{B}}(\mathbf{u})=E[\mathcal{B}_n(\mathbf{u})]=O(\|\mathbf{h}\|^2+\|\boldsymbol{\lambda}\|)$ and $\mathcal{B}_n(\mathbf{u})-E[\mathcal{B}_n(\mathbf{u})]=O_P((n\mathbf{h}!/\log n)^{-1/2}(\|\mathbf{h}\|^2+\|\boldsymbol{\lambda}\|))$ uniformly in $\mathbf{u}$, we can show that

$$\begin{aligned}
T_{n11}^{(2)} &= (\mathbf{h}!)^{1/2}\sum_{i=1}^{n}\overline{\mathcal{B}}(\mathbf{U}_i)'\,\overline{\boldsymbol{\Gamma}}_1(\mathbf{U}_i)'\,\overline{\boldsymbol{\Gamma}}_1(\mathbf{U}_i)\,\mathcal{V}_n(\mathbf{U}_i)\\
&\quad +(\mathbf{h}!)^{1/2}\sum_{i=1}^{n}\left[\mathcal{B}_n(\mathbf{U}_i)-\overline{\mathcal{B}}(\mathbf{U}_i)\right]'\overline{\boldsymbol{\Gamma}}_1(\mathbf{U}_i)'\,\overline{\boldsymbol{\Gamma}}_1(\mathbf{U}_i)\,\mathcal{V}_n(\mathbf{U}_i)\\
&\quad +(\mathbf{h}!)^{1/2}\sum_{i=1}^{n}\mathcal{B}_n(\mathbf{U}_i)'\left[\boldsymbol{\Gamma}_{n1}(\mathbf{U}_i)'\,\boldsymbol{\Gamma}_{n1}(\mathbf{U}_i)-\overline{\boldsymbol{\Gamma}}_1(\mathbf{U}_i)'\,\overline{\boldsymbol{\Gamma}}_1(\mathbf{U}_i)\right]\mathcal{V}_n(\mathbf{U}_i)\\
&= O_P\left(\left(\|\mathbf{h}\|^2+\|\boldsymbol{\lambda}\|\right)\left(\mathbf{h}!^{-1/2}+\sqrt{n\mathbf{h}!}\right)\right)\\
&\quad +n(\mathbf{h}!)^{1/2}O_P\left((n\mathbf{h}!/\log n)^{-1/2}(\|\mathbf{h}\|^2+\|\boldsymbol{\lambda}\|)\right)O_P((n\mathbf{h}!)^{-1/2})\\
&\quad +n(\mathbf{h}!)^{1/2}O_P\left((\|\mathbf{h}\|^2+\|\boldsymbol{\lambda}\|)\right)O_P\left(\nu_n+\|\mathbf{h}\|^2+\|\boldsymbol{\lambda}\|+(n\mathbf{h}!/\log n)^{-1/2}\right)O_P((n\mathbf{h}!)^{-1/2})\\
&= o_P(1)\ \text{by Assumption A8.}
\end{aligned}$$

It follows that

$$T_{n11}=\overline{T}_{n11}+o_P(1). \qquad \text{(A.6)}$$

Now, write

$$\begin{aligned}
\overline{T}_{n11} &= (\mathbf{h}!)^{1/2}\sum_{i=1}^{n}\mathcal{V}_n(\mathbf{U}_i)'\,\overline{\boldsymbol{\Gamma}}_1(\mathbf{U}_i)'\,\overline{\boldsymbol{\Gamma}}_1(\mathbf{U}_i)\,\mathcal{V}_n(\mathbf{U}_i)\\
&\quad +(\mathbf{h}!)^{1/2}\sum_{i=1}^{n}\mathcal{V}_n(\mathbf{U}_i)'\left[\boldsymbol{\Gamma}_{n1}(\mathbf{U}_i)-\overline{\boldsymbol{\Gamma}}_1(\mathbf{U}_i)\right]'\left[\boldsymbol{\Gamma}_{n1}(\mathbf{U}_i)-\overline{\boldsymbol{\Gamma}}_1(\mathbf{U}_i)\right]\mathcal{V}_n(\mathbf{U}_i)\\
&\quad +2(\mathbf{h}!)^{1/2}\sum_{i=1}^{n}\mathcal{V}_n(\mathbf{U}_i)'\left[\boldsymbol{\Gamma}_{n1}(\mathbf{U}_i)-\overline{\boldsymbol{\Gamma}}_1(\mathbf{U}_i)\right]'\mathcal{V}_n(\mathbf{U}_i)\,\overline{\boldsymbol{\Gamma}}_1(\mathbf{U}_i)\,\mathcal{V}_n(\mathbf{U}_i)\\
&\equiv \overline{T}_{11a}+\overline{T}_{11b}+2\overline{T}_{11c},\ \text{say.} \qquad \text{(A.7)}
\end{aligned}$$

Recall $\boldsymbol{\zeta}_i = (\mathbf{U}_i', \mathbf{Z}_i^{a\prime}, \varepsilon_i)'$ and $\boldsymbol{\varphi}(\boldsymbol{\zeta}_i, \boldsymbol{\zeta}_j) = \overline{\boldsymbol{\Gamma}}_1(\mathbf{U}_i)\left(\left(\begin{array}{c}\mathbf{Z}_j^a\varepsilon_j \\ (\mathbf{Z}_j^a\varepsilon_j)\otimes\boldsymbol{\eta}_j(\mathbf{U}_i^c)\end{array}\right)K_{\mathbf{h}\boldsymbol{\lambda},j\mathbf{U}_i}\right)$. So

$$\begin{aligned}
\overline{T}_{n11a} &= \frac{(\mathbf{h}!)^{1/2}}{n^2}\sum_{i=1}^n\sum_{j=1}^n\sum_{l=1}^n \boldsymbol{\varphi}(\boldsymbol{\zeta}_i,\boldsymbol{\zeta}_j)'\boldsymbol{\varphi}(\boldsymbol{\zeta}_i,\boldsymbol{\zeta}_l) \\
&= \frac{(\mathbf{h}!)^{1/2}}{n^2}\sum_{i=1}^n\sum_{j=1}^n \left\|\boldsymbol{\varphi}(\boldsymbol{\zeta}_i,\boldsymbol{\zeta}_j)\right\|^2 + \frac{(\mathbf{h}!)^{1/2}}{n^2}\sum_{i=1}^n\sum_{j\neq i}^n\sum_{l\neq j,i}^n \boldsymbol{\varphi}(\boldsymbol{\zeta}_i,\boldsymbol{\zeta}_j)'\boldsymbol{\varphi}(\boldsymbol{\zeta}_i,\boldsymbol{\zeta}_l) \\
&\quad + \frac{2(\mathbf{h}!)^{1/2}}{n^2}\sum_{i=1}^n\sum_{j\neq i}^n \boldsymbol{\varphi}(\boldsymbol{\zeta}_i,\boldsymbol{\zeta}_j)'\boldsymbol{\varphi}(\boldsymbol{\zeta}_i,\boldsymbol{\zeta}_i) \\
&\equiv B_n + V_{n1} + R_{n1}, \text{ say.} \qquad \text{(A.8)}
\end{aligned}$$

Let $\overline{\boldsymbol{\varphi}}(\boldsymbol{\zeta}_i,\boldsymbol{\zeta}_j,\boldsymbol{\zeta}_l) \equiv [\boldsymbol{\varphi}(\boldsymbol{\zeta}_i,\boldsymbol{\zeta}_j)'\boldsymbol{\varphi}(\boldsymbol{\zeta}_i,\boldsymbol{\zeta}_l) + \boldsymbol{\varphi}(\boldsymbol{\zeta}_j,\boldsymbol{\zeta}_i)'\boldsymbol{\varphi}(\boldsymbol{\zeta}_j,\boldsymbol{\zeta}_l) + \boldsymbol{\varphi}(\boldsymbol{\zeta}_l,\boldsymbol{\zeta}_i)'\boldsymbol{\varphi}(\boldsymbol{\zeta}_l,\boldsymbol{\zeta}_j)]/3$. Then

$$V_{n1} = \frac{6(\mathbf{h}!)^{1/2}}{n^2}\sum_{1\leq i<j<l\leq n}\overline{\boldsymbol{\varphi}}(\boldsymbol{\zeta}_i,\boldsymbol{\zeta}_j,\boldsymbol{\zeta}_l) = \frac{(n-1)(n-2)}{n}\overline{V}_{n1},$$

where $\overline{V}_{n1} \equiv \frac{6(\mathbf{h}!)^{1/2}}{n(n-1)(n-2)}\sum_{1\leq i<j<l\leq n}\overline{\boldsymbol{\varphi}}(\boldsymbol{\zeta}_i,\boldsymbol{\zeta}_j,\boldsymbol{\zeta}_l)$. Note that for all $i\neq j\neq l$, $\theta \equiv E[\overline{\boldsymbol{\varphi}}(\boldsymbol{\zeta}_i,\boldsymbol{\zeta}_j,\boldsymbol{\zeta}_l)] = 0$, $\overline{\boldsymbol{\varphi}}_1(\mathbf{a}) \equiv E[\overline{\boldsymbol{\varphi}}(\mathbf{a},\boldsymbol{\zeta}_j,\boldsymbol{\zeta}_l)] = 0$, and $\overline{\boldsymbol{\varphi}}_2(\mathbf{a},\widetilde{\mathbf{a}}) \equiv E[\overline{\boldsymbol{\varphi}}(\mathbf{a},\widetilde{\mathbf{a}},\boldsymbol{\zeta}_l)] = \frac{1}{3}E[\boldsymbol{\varphi}(\boldsymbol{\zeta}_l,\mathbf{a})'\,\boldsymbol{\varphi}(\boldsymbol{\zeta}_l,\widetilde{\mathbf{a}})]$, where $\mathbf{a}$ and $\widetilde{\mathbf{a}}$ are nonrandom. Let $\overline{\boldsymbol{\varphi}}_3(\mathbf{a},\widetilde{\mathbf{a}},\overline{\mathbf{a}}) \equiv \overline{\boldsymbol{\varphi}}(\mathbf{a},\widetilde{\mathbf{a}},\mathbf{a}) - \overline{\boldsymbol{\varphi}}_2(\mathbf{a},\widetilde{\mathbf{a}}) - \overline{\boldsymbol{\varphi}}_2(\mathbf{a},\overline{\mathbf{a}}) - \overline{\boldsymbol{\varphi}}_2(\widetilde{\mathbf{a}},\mathbf{a})$. By the Hoeffding decomposition,

$$\overline{V}_{n1} = 3H_n^{(2)} + H_n^{(3)},$$

where $H_n^{(2)} \equiv \frac{2(h!)^{1/2}}{n(n-1)}\sum_{1\leq i<j\leq n}\overline{\boldsymbol{\varphi}}_2(\boldsymbol{\zeta}_i,\boldsymbol{\zeta}_j)$ and $H_n^{(3)} \equiv \frac{6(h!)^{1/2}}{n(n-1)(n-2)}\sum_{1\leq i<j<l\leq n}\overline{\boldsymbol{\varphi}}_3(\boldsymbol{\zeta}_i,\boldsymbol{\zeta}_j,\boldsymbol{\zeta}_l)$. Noting that $E[\overline{\boldsymbol{\varphi}}_3(\mathbf{a},\widetilde{\mathbf{a}},\boldsymbol{\zeta}_i)] = 0$ and that $\overline{\boldsymbol{\varphi}}_3$ is symmetric in its arguments by construction, it is straightforward to show that $E[H_n^{(3)}] = 0$ and $E[H_n^{(3)}]^2 = O(n^{-3}(\mathbf{h}!)^{-1})$. Hence, $H_n^{(3)} = O_P(n^{-3/2}(\mathbf{h}!)^{-1/2}) = o_P(n^{-1})$ by the Chebyshev inequality. It follows that $V_{n1} = \frac{n(n-2)}{n-1}\overline{V}_{n1} = \{1+o(1)\}\mathcal{H}_n + o_P(1)$, where

$$\mathcal{H}_n \equiv \frac{2(\mathbf{h}!)^{1/2}}{n}\sum_{1\leq i\leq j\leq n}3\overline{\boldsymbol{\varphi}}_2(\boldsymbol{\zeta}_i,\boldsymbol{\zeta}_j) = \frac{2(\mathbf{h}!)^{1/2}}{n}\sum_{1\leq i<j\leq n}\int\boldsymbol{\varphi}(\mathbf{a},\boldsymbol{\zeta}_i)'\,\boldsymbol{\varphi}(\mathbf{a},\boldsymbol{\zeta}_j)dF_{\boldsymbol{\zeta}}(\mathbf{a}).$$

As $\mathcal{H}_n$ is a second order degenerate U-statistic, it is straightforward but tedious to verify that all the conditions of Theorem 1 of Hall (1984) are satisfied, implying that a central limit theorem applies to $\mathcal{H}_n$: $\mathcal{H}_n \xrightarrow{d} N(0,\sigma_0^2)$, where the asymptotic variance of $\mathcal{H}_n$ is given by $\sigma_0^2 \equiv \lim_{n\to\infty}\sigma_n^2$ and $\sigma_n^2 \equiv 2\mathbf{h}!E_jE_l[\int\boldsymbol{\varphi}(\boldsymbol{\zeta},\boldsymbol{\zeta}_j)'\boldsymbol{\varphi}(\boldsymbol{\zeta},\boldsymbol{\zeta}_l)F_{\boldsymbol{\zeta}}(d\boldsymbol{\zeta})]^2$. Consequently

$$V_{n1} \xrightarrow{d} N(0,\sigma_0^2). \qquad \text{(A.9)}$$

For R_{n1}, it is easy to verify that $E(R_{n1}) = 0$ and $E(R_{n1}^2) = O(n(\mathbf{h}!)^{-1}) = o(1)$. So $R_{n1} = o_P(1)$ by the Chebyshev inequality. Combining this with (A.8) and (A.9) yields

$$\overline{T}_{n11a} - B_n \xrightarrow{d} N(0,\sigma_0^2). \qquad \text{(A.10)}$$

By Assumption A8, we have

$$\begin{aligned}\overline{T}_{n11b} &= (\mathbf{h}!)^{1/2}\sum_{i=1}^{n}\operatorname{tr}\left\{\mathcal{V}_n(\mathbf{U}_i)'\mathcal{V}_n(\mathbf{U}_i)\left[\boldsymbol{\Gamma}_{n1}(\mathbf{U}_i)-\overline{\boldsymbol{\Gamma}}_1(\mathbf{U}_i)\right]\left[\boldsymbol{\Gamma}_{n1}(\mathbf{U}_i)-\overline{\boldsymbol{\Gamma}}_1(\mathbf{U}_i)\right]'\right\}\\ &\leq (\mathbf{h}!)^{1/2}\sup_{\mathbf{u}}\left\|\boldsymbol{\Gamma}_{n1}(\mathbf{u})-\overline{\boldsymbol{\Gamma}}_1(\mathbf{u})\right\|^2\sum_{i=1}^{n}\mathcal{V}_n(\mathbf{U}_i)'\mathcal{V}_n(\mathbf{U}_i)\\ &= (\mathbf{h}!)^{1/2}O_P\left(\left(n^{-1/2}(\mathbf{h}!)^{-1/2}\sqrt{\log n}+\|\mathbf{h}\|^2+\|\boldsymbol{\lambda}\|\right)^2\right)O_P\left((\mathbf{h}!)^{-1}\right)=o_P(1).\end{aligned}$$

Similarly, $\overline{T}_{n11c}=(\mathbf{h}!)^{1/2}O_P(((n\mathbf{h}!)^{-1/2}\sqrt{\log n}+\|\mathbf{h}\|^2+\|\boldsymbol{\lambda}\|))O_P((\mathbf{h}!)^{-1})=o_P(1)$. It follows that $\overline{T}_{n11}-B_n\xrightarrow{d}N\left(0,\sigma_0^2\right)$, and by (A.6) that

$$T_{n11}-B_n\xrightarrow{d}N\left(0,\sigma_0^2\right). \tag{A.11}$$

Under $\mathbb{H}_1(r_n)$, $\overline{\mathbf{g}}_1=n^{-1}\sum_{i=1}^n\mathbf{g}_1(\mathbf{U}_i)=\boldsymbol{\theta}_1+r_n\overline{\boldsymbol{\delta}}_n$, where $\overline{\boldsymbol{\delta}}_n=n^{-1}\sum_{i=1}^n\boldsymbol{\delta}_n(\mathbf{U}_i)=E[\boldsymbol{\delta}_n(\mathbf{U}_i)]+O_P\left(n^{-1/2}\right)$. It follows that

$$T_{n12}=n^{-1}\sum_{i=1}^{n}\left\|\boldsymbol{\delta}_n(\mathbf{U}_i)-\overline{\boldsymbol{\delta}}_n\right\|^2=\lim_{n\to\infty}E\left[\left\|\boldsymbol{\delta}_n(\mathbf{U}_i)-E[\boldsymbol{\delta}_n(\mathbf{U}_i)]\right\|^2\right]=\mu_0, \tag{A.12}$$

and

$$\begin{aligned}T_{n13} &= r_n(\mathbf{h}!)^{1/2}\sum_{i=1}^{n}\left[\boldsymbol{\Gamma}_{n1}(\mathbf{U}_i)\mathcal{B}_n(\mathbf{U}_i)+\boldsymbol{\Gamma}_{n1}(\mathbf{U}_i)\mathcal{V}_n(\mathbf{U}_i)\right]'\left[\boldsymbol{\delta}_n(\mathbf{U}_i)-\overline{\boldsymbol{\delta}}_n\right]\\ &= r_n(\mathbf{h}!)^{1/2}\sum_{i=1}^{n}\overline{\boldsymbol{\Gamma}}_1(\mathbf{U}_i)\mathcal{V}_n(\mathbf{U}_i)\left[\boldsymbol{\delta}_n(\mathbf{U}_i)-\overline{\boldsymbol{\delta}}_n\right]+o_P(1)\\ &= \overline{T}_{n13}+o_P(1),\end{aligned}$$

where $\overline{T}_{n13}=\frac{r_n(\mathbf{h}!)^{1/2}}{n}\sum_{i=1}^n\sum_{j\neq i}^n\boldsymbol{\varphi}(\boldsymbol{\zeta}_i,\boldsymbol{\zeta}_j)\{\boldsymbol{\delta}_n(\mathbf{U}_i)-E[\boldsymbol{\delta}_n(\mathbf{U}_i)]\}$. Noting that $E\left[\overline{T}_{n13}\right]=0$ and $E\left[\overline{T}_{n13}\right]^2=r_n^2\mathbf{h}!O(n+(\mathbf{h}!)^{-1})=o(1)$, we have $\overline{T}_{n13}=o_P(1)$ by the Chebyshev inequality. Consequently,

$$T_{n13}=o_P(1). \tag{A.13}$$

Combining (A.5) with (A.11)-(A.13) yields $T_{n1}-B_n-\mu_0\xrightarrow{d}N\left(0,\sigma_0^2\right)$.

Now we show (ii). Noting that $\overline{\overline{\mathbf{g}}}_1-\overline{\mathbf{g}}_1=\frac{1}{n}\sum_{i=1}^n[\boldsymbol{\Gamma}_{n1}(\mathbf{U}_i)\mathcal{B}_n(\mathbf{U}_i)+\boldsymbol{\Gamma}_{n1}(\mathbf{U}_i)\mathcal{V}_n(\mathbf{U}_i)]$, we have

$$\begin{aligned}T_{n2} &= \frac{(\mathbf{h}!)^{1/2}}{n}\sum_{i=1}^{n}\sum_{j=1}^{n}\mathcal{B}_n(\mathbf{U}_i)'\boldsymbol{\Gamma}_{n1}(\mathbf{U}_i)'\boldsymbol{\Gamma}_{n1}(\mathbf{U}_j)\mathcal{B}_n(\mathbf{U}_j)\\ &\quad+\frac{(\mathbf{h}!)^{1/2}}{n}\sum_{i=1}^{n}\sum_{j=1}^{n}\mathcal{V}_n(\mathbf{U}_i)\boldsymbol{\Gamma}_{n1}(\mathbf{U}_i)'\boldsymbol{\Gamma}_{n1}(\mathbf{U}_j)\mathcal{V}_n(\mathbf{U}_j)\\ &\quad+\frac{2(\mathbf{h}!)^{1/2}}{n}\sum_{i=1}^{n}\sum_{j=1}^{n}\mathcal{B}_n(\mathbf{U}_i)'\boldsymbol{\Gamma}_{n1}(\mathbf{U}_i)'\boldsymbol{\Gamma}_{n1}(\mathbf{U}_j)\mathcal{V}_n(\mathbf{U}_j)\\ &\equiv T_{n21}+T_{n22}+2T_{n23},\ \text{say.}\end{aligned}$$

For T_{n21}, we have $T_{n21} \leq \sup_{\mathbf{u}} \|\mathcal{B}_n(\mathbf{u})\|^2 \frac{(\mathbf{h}!)^{1/2}}{n} \sum_{i=1}^n \sum_{j=1}^n \text{tr}(\mathbf{\Gamma}_{n1}(\mathbf{U}_i)' \mathbf{\Gamma}_{n1}(\mathbf{U}_j)) = n(\mathbf{h}!)^{1/2} \sup_{\mathbf{u}} \|\mathcal{B}_n(\mathbf{u})\|^2 \text{tr}(\overline{\mathbf{\Gamma}}_{n1}' \overline{\mathbf{\Gamma}}_{n1}) = n(\mathbf{h}!)^{1/2} O_P((\|\mathbf{h}\|^2 + \|\boldsymbol{\lambda}\|)^2) = o_P(1)$, where $\overline{\mathbf{\Gamma}}_{n1} = n^{-1} \sum_{i=1}^n \mathbf{\Gamma}_{n1}(\mathbf{U}_i) = O_P(1)$. For T_{n22}, we can show that $T_{n22} = \overline{T}_{n22} + o_P(1)$, where $\overline{T}_{n22} = \frac{(\mathbf{h}!)^{1/2}}{n} \sum_{i=1}^n \sum_{j=1}^n \mathcal{V}_n(\mathbf{U}_i)$ $\overline{\mathbf{\Gamma}}_1(\mathbf{U}_i)' \overline{\mathbf{\Gamma}}_1(\mathbf{U}_j) \mathcal{V}_n(\mathbf{U}_j)$. Noting that $E\left|\overline{T}_{n2,2}\right| = E\left[\overline{T}_{n2,2}\right] = O((\mathbf{h}!)^{1/2} + n^{-1}(\mathbf{h}!)^{-1})$ by straightforward calculations, we have $T_{n22} = o_P(1)$ by the Markov inequality. Then by the Cauchy-Schwarz inequality, $T_{n23} \leq \frac{(\mathbf{h}!)^{1/2}}{n} \sum_{i=1}^n \sum_{j=1}^n \{\mathcal{B}_n(\mathbf{U}_i)' \mathbf{\Gamma}_{n1}(\mathbf{U}_i)' \mathbf{\Gamma}_{n1}(\mathbf{U}_j) \mathcal{B}_n(\mathbf{U}_j)\}^{1/2} \{\mathcal{V}_n(\mathbf{U}_i)' \mathbf{\Gamma}_{n1}(\mathbf{U}_i)' \mathbf{\Gamma}_{n1}(\mathbf{U}_j)$ $\mathcal{V}_n(\mathbf{U}_j)\}^{1/2} \leq \{T_{n21}\}^{1/2} \{T_{n22}\}^{1/2} = o_P(1)$. So $T_{n2} = o_P(1)$. ■

Proof of Theorems 4.4

Using the notation defined in the proof of Theorem 4.3, we have $n^{-1}(\mathbf{h}!)^{-1/2} T_n = n^{-1}(\mathbf{h}!)^{-1/2}$ $(T_{n1} - T_{n2})$. Under $\mathbb{H}_1$, it is easy to show that

$$\begin{aligned} n^{-1}(\mathbf{h}!)^{-1/2} T_{n1} &= n^{-1}(\mathbf{h}!)^{-1/2} T_{n12} + o_P(1) = n^{-1} \sum_{i=1}^n [\mathbf{g}_1(\mathbf{U}_i) - \overline{\mathbf{g}}]' [\mathbf{g}_1(\mathbf{U}_i) - \overline{\mathbf{g}}] + o_P(1) \\ &= E\|\mathbf{g}_1(\mathbf{U}_i) - E[\mathbf{g}_1(\mathbf{U}_i)]\|^2 + o_P(1) = \mu_A + o_P(1), \end{aligned}$$

and $n^{-1}(\mathbf{h}!)^{-1/2} T_{n2} = o_P(1)$. On the other hand, $n^{-1}(\mathbf{h}!)^{-1/2} \widehat{B}_n = O_P\left(n^{-1}(\mathbf{h}!)^{-1}\right) = o_P(1)$ and $\widehat{\sigma}_n^2 = \sigma_0^2 + o_P(1)$. It follows that $n^{-1}(\mathbf{h}!)^{-1/2} J_n = (n^{-1}(\mathbf{h}!)^{-1/2} T_n - n^{-1}(\mathbf{h}!)^{-1/2} \widehat{B}_n)/\sqrt{\widehat{\sigma}_n^2} \xrightarrow{P} \mu_A/\sigma_0$, and the conclusion follows. ■

References

Ai, C. and X. Chen (2003), "Efficient Estimation of Models with Conditional Moment Restrictions Containing Unknown Functions," *Econometrica* 71, 1795–1843.

Becker, G. and B. Chiswick (1966), "Education and Distribution of Earnings," *American Economic Review* 56, 358-369.

Cai, Z., and H. Xiong (2010), "Efficient Estimation of Partially Varying Coefficient Instrumental Variables models," *WISE Working Paper Series WISEWP 0614*, Xiamen University.

Cai, Z., M. Das, H. Xiong, and X. Wu (2006), "Functional Coefficient Instrumental Variables Models," *Journal of Econometrics* 133, 207-241.

Cai, Z., J. Fan, and Q. Yao (2000), "Functional-coefficient Regression Models for Nonlinear Time Series," *Journal of American Statistical Association* 95, 941–956.

Cai, Z. and Li, Q. (2008), "Nonparametric Estimation of Varying Coefficient Dynamic Panel Data Models," *Econometric Theory* 24, 1321-1342.

Card, D. (2001), "Estimating the Return to Schooling: Progress on Some Persistent Econometric Problems," *Econometrica* 69, 1127-1160.

Card, D., and T. Lemieux (2001), "Can Falling Supply Explain the Rising Return to College for Younger Men? A Cohort-based Analysis," *The Quarterly Journal of Economics* 116, 705-746.

Chen, R. and R. S. Tsay (1993), “Functional-coefficient Autoregressive Models,” *Journal of the American Statistical Association* 88, 298-308.

Cleveland, W. S., E. Grosse, and W. M. Shyu (1992), “Local Regression Models,” in J. M. Chambers and T. J. Hastie (Eds.), *Statistical Models in S,* pp. 309-376. Pacific Grove, CA: Wadsworth & Brooks/Cole.

Das, M. (2005), “Instrumental Variables Estimators for Nonparametric Models with Discrete Endogenous Regressors,” *Journal of Econometrics* 124, 335-361.

Das, M., W. K. Newey and F. Vella (2003), “Nonparametric Estimation of Sample Selection Models,” *Review of Economic Studies* 80, 33-58.

Fan, J. and W. Zhang (1999), “Statistical Estimation in Varying Coefficient Models,” *The Annals of Statistics* 27, 1491–1518.

Fan, J., C. Zhang, and J. Zhang (2001), “Generalized Likelihood Ratio Statistics and Wilks Phenomenon,” *The Annals of Statistics* 29, 153-193.

Granger, C. W. J. and T. Teräsvirta (1993), *Modelling Nonlinear Economic Relationships.* Oxford: Oxford University Press.

Granger, C. W. J. and T. Teräsvirta (1999), “A Simple Nonlinear Time Series Model with Misleading Linear Properties,” *Economics Letters* 62, 161-165.

Hall, P. (1984), “Central Limit Theorem for Integrated Square Error Properties of Multivariate Nonparametric Density Estimators,” *Journal of Multivariate Analysis* 14, 1-16.

Hall, P., R. C. L. Wolf, and Q. Yao (1999), “Methods of Estimating a Conditional Distribution Function,” *Journal of the American Statistical Association 94*, 154-163.

Härdle. W. and J. S. Marron (1991), “Bootstrap Simultaneous Error Bars for Nonparametric Regression,” *The Annals of Statistics* 19, 778-796.

Hastie, T. J., and R. J. Tibshirani (1993), “Varying-coefficient Models (with Discussion),” *Journal of the Royal Statistical Society*, Series B. 55, 757-796.

Hildreth, C. and J. P. Houck (1968), “Some Estimators for a Linear Model with Random Coefficients,” *Journal of the American Statistical Association* 63, 584-595.

Hong, Y., and Y. Lee (2009), “A Loss Function Approach to Model Specification Testing and Its Relative Efficiency to the GLR Test,” *Discussion paper*, Dept. of Economics, Cornell University.

Lewbel, A. (2007), “A Local Generalized Method of Moments Estimator,” *Economics Letters* 94, 124-128.

Li, Q., and J. Racine (2007), *Nonparametric Econometrics: Theory and Practice.* Princeton University Press, Princeton and Oxford.

Li, Q., and J. Racine (2008), "Nonparametric Estimation of Conditional CDF and Quantile Functions with Mixed Categorical and Continuous Data," *Journal of Business and Economic Statistics* 26, 423-734.

Masry, E. (1996), "Multivariate Local Polynomial Regression for Time series: Uniform Strong Consistency Rates," *Journal of Time Series Analysis* 17, 571-599.

Mincer, J. (1974), *Schooling, Experience and Earnings*, New York: National Bureau of Economic Research.

Murphy, K., and F. Welch (1990), "Empirical Age-earnings Profiles," *Journal of Labor Economics* 8, 202-229.

Newey, W. K. (1990), "Efficient Instrumental Variables Estimation of Nonlinear Models," *Econometrica* 58, 809-837.

Pagan, A. and A. Ullah (1999), *Nonparametric Econometrics,* New York: Cambridge University Press.

Racine, J. and Q. Li (2004), "Nonparametric Estimation of Regression Functions with Both Categorical and Continuous Data," *Journal of Econometrics* 119, 99-130.

Raj, B. and A. Ullah (1981), *Econometrics: A Varying Coefficients Approach,* St. Martins Press.

Singh, B. and A. Ullah (1974), "Estimation of Seemingly Unrelated Regressions with Random Coefficients," *Journal of the American Statistical Association* 69, 191-195.

Su, L. (2011), "A Semi-parametric GMM Estimation of Spatial Autoregressive Models," forthcoming in *Journal of Econometrics.*

Su, L., Y. Chen, and A. Ullah (2009), "Functional Coefficient Estimation with Both Categorical and Continuous Data," *Advances in Econometrics* 25, 131-167.

Swamy, P. A. V. B. (1970), "Efficient Inference in a Random Coefficient Regression Model," *Econometrica* 38, 311-323.

Swamy, P. A. V. B. (1971), *Statistical Inference in Random Coefficient Regression Models.* Berlin-Heidelberg-New York: Springer-Verlag.

Tran, K. C., and E. G. Tsionas (2010), "Local GMM Estimation of Semiparametric Panel Data with Smooth Coefficient Models," *Econometric Reviews* 29, 39-61.

Ullah, A. (1985), "Specification Analysis of Econometric Models," *Journal of Quantitative Economics* 1, 187-209.

Vella, F. (1994), "Gender Roles and Human Capital Investment: The Relationship between Traditional Attitudes and Female Labor Market Performance," *Economica* 61, 191-211.

Figure 1: **Experience-Wage and Educaiton-Wage Profiles**

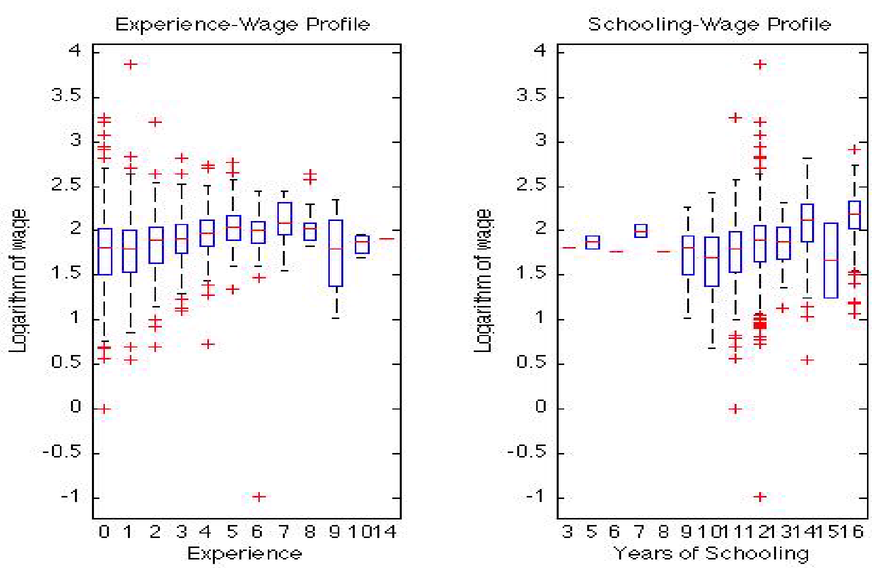
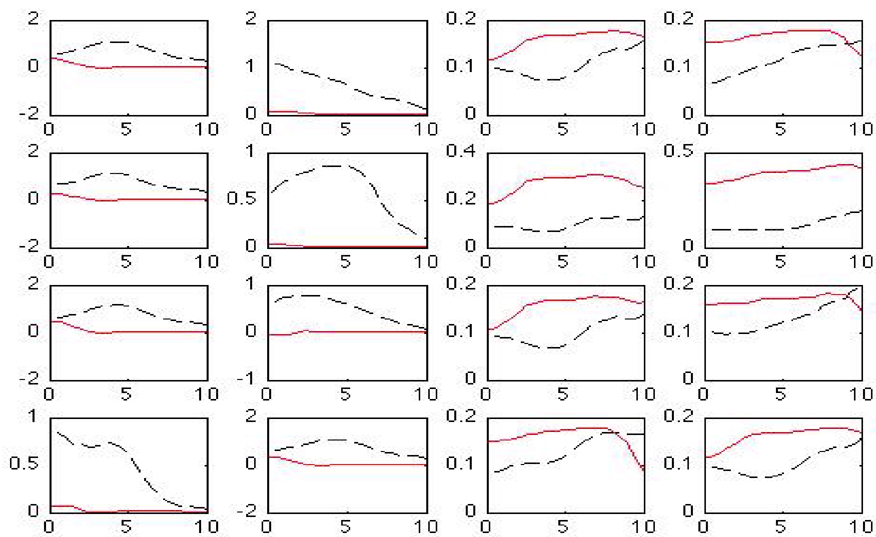
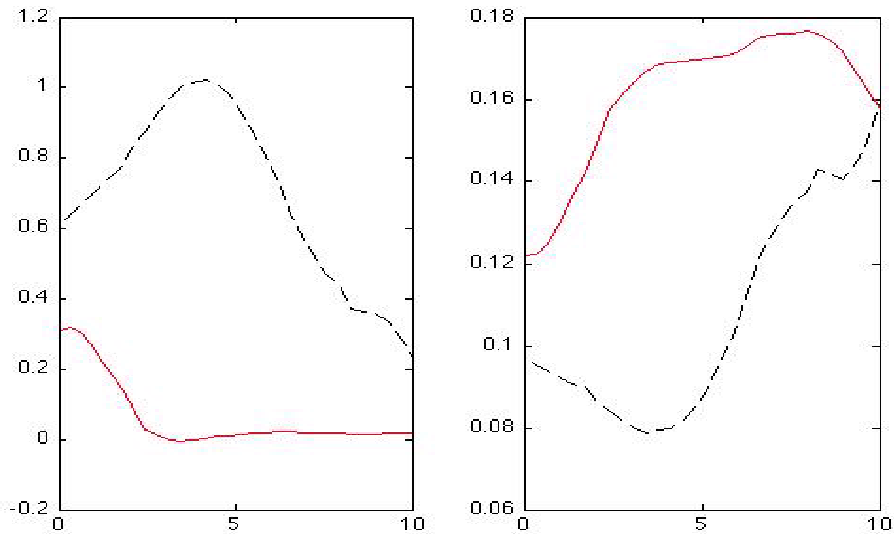


Figure 2: **Plots of** $g_1(Experience,\ Individual\ Characteristic, :)$ **and** $g_2(Experience,$ $Individual\ Characteristic, :)$ **Averaged over Other Categorical Variables**



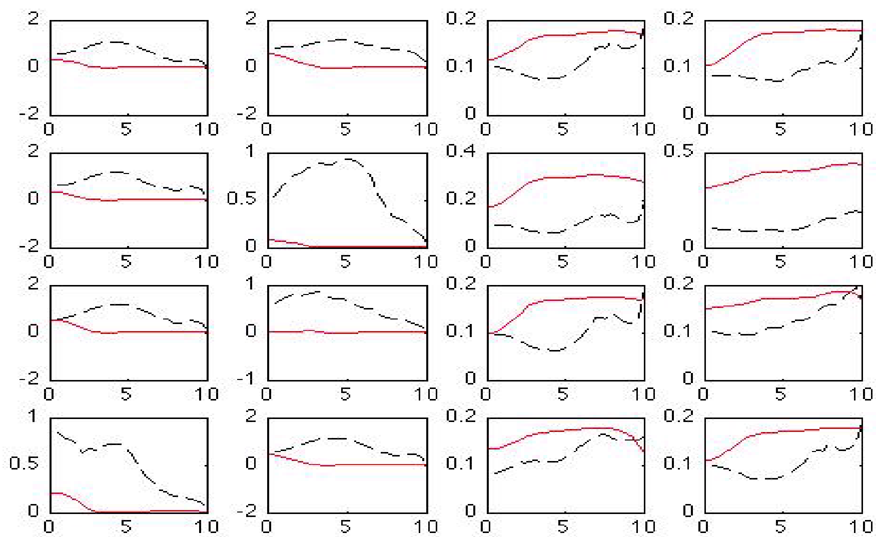
Notes: Horizontal Axis - Experience. Vertical Axis - g_1 or g_2. SCU estimate, dashed line; our proposed estimate with optimal weight matrix, solid line. The rule of thumb method is used to choose the bandwidth. The four rows correspond to *Individual Characteristic* being a binary indicator of whether a woman is married, a union member, a government employee, and born in Australia, from the top to the bottom. The four columns from the left to the right correspond to g_1 for $Individual\ Characteristic = 1$ and 0, and g_2 for $Individual\ Characteristic = 1$ and 0, respectively.

Figure 3: **Plots of** $g_1(Experience, :)$ **and** $g_2(Experience, :)$ **Averaged over All Categorical Variables**



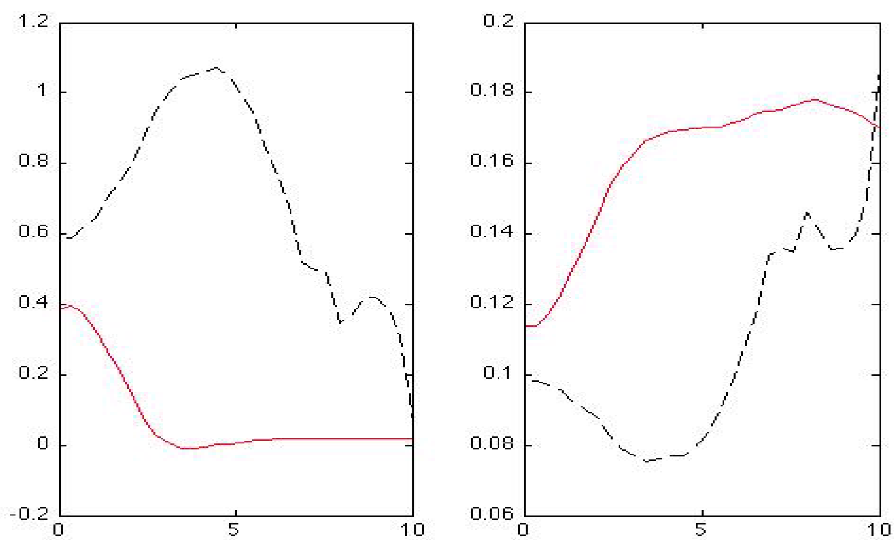
Notes: Horizontal Axis - Experience. Vertical Axis - g_1 or g_2. SCU estimate, dashed line; our proposed estimate with optimal weight matrix, solid line. The rule of thumb method is used to choose the bandwidth. The two columns from the left to the right correspond to g_1 and g_2, respectively.

Figure 4: **Plots of** $g_1(Experience,\ Individual\ Characteristic, :)$ **and** $g_2(Experience,\ Individual\ Characteristic, :)$ **Averaged over Other Categorical Variables**



Notes: Horizontal Axis - Experience. Vertical Axis - g_1 or g_2. SCU estimate, dashed line; our proposed estimate with optimal weight matrix, solid line. The LSCV method is used to choose the bandwidth.
The four rows correspond to *Individual Characteristic* being a binary indicator of whether a woman is married, a union member, a government employee, and born in Australia, from the top to the bottom. The four columns from the left to the right correspond to g_1 for *Individual Characteristic* $= 1$ and 0, and g_2 for *Individual Characteristic* $= 1$ and 0, respectively.

Figure 5: **Plots of $g_1(Experience, :)$ and $g_2(Experience, :)$ Averaged over All Categorical Variables**



Notes: Horizontal Axis - Experience. Vertical Axis - g_1 or g_2. SCU estimate, dashed line; our proposed estimate with optimal weight matrix, solid line. The LSCV method is used to choose the bandwidth. The two columns from the left to the right correspond to g_1 and g_2, respectively.