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

Liangjun Su,[a] Irina Murtazashvili,[b] Aman Ullah[c]
[a] School of Economics, Singapore Management University (ljsu@smu.edu.sg)
[b] Department of Economics, Drexel University (im99@drexel.edu)
[c] Department of Economics, University of California, Riverside (aman.ullah@ucr.edu)

November 17, 2012

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 derive the optimal instrumental variable that minimizes the asymptotic variance-covariance matrix among the class of all local linear GMM estimators. Data-dependent bandwidth sequences are also allowed for. We propose a nonparametric test for the constancy of the functional coefficients, study its asymptotic properties under the null hypothesis as well as a sequence of local alternatives and global alternatives, and propose a bootstrap version for it. 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; Heterogeneity; Functional coefficient; Local linear GMM estimation; Optimal instrumental variable; Schooling.

1 Introduction

In the classical econometrics literature, an econometric model is 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 rate of return 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 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 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 are modeled as unknown functions of the observed variables which can be estimated nonparametrically. 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 (2012), 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 coefficients 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 not consider the endogeneity issue which prevails in economics.

In addition, in the estimation context, the advantage of using the traditional constant coefficient models rests 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 coefficients. 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 exogenous continuous and categorical variables. For given IVs an optimal local linear GMM estimator is proposed where the weight matrix is obtained by minimizing the asymptotic variance-covariance matrix (AVC) of the GMM estimator. We also consider the choice of optimal IVs to minimize the AVC among the class of all local linear GMM estimators and establish the asymptotic normality of the local linear GMM estimator for a data-dependent bandwidth sequence. Then we develop a new test statistic for testing the hypothesis that a subvector of the functional coefficients is 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 and justify its asymptotic validity. 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 performs well in comparison with some existing estimators in the literature and our test has correct size and good power properties 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 schooling 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 $\|\mathbf{M}\| = \sqrt{\text{tr}\left(\mathbf{M}\mathbf{M}'\right)}$. 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 $\xrightarrow{d}$ and $\xrightarrow{P}$ 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 linearly. Then we propose local linear GMM estimates for the functional coefficients.

2.1 Functional coefficient representation

We consider the following functional coefficient IV model

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

where Y_i is a 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, $\mathbf{Z}_i$ is a $q_z \times 1$ vector of instrumental variables, and a.s. abbreviates almost surely. We assume that a random sample $\{Y_i, \mathbf{X}_i, \mathbf{Z}_i, \mathbf{U}_i\}_{i=1}^n$ is observed.

In the absence of $\mathbf{U}_i^d$, (2.1) reduces to the model of CDXW (2006). If none of the variables in $\mathbf{X}_i$ are 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, region, etc.

2.2 Local linear GMM estimation

The orthogonality condition in (2.1) 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 $\mathbf{V}_i = \left(\mathbf{Z}_i', \mathbf{U}_i'\right)'$. It indicates that for any $k \times 1$ vector function $\mathbf{Q}\left(\mathbf{V}_i\right)$, we have

$$E\left[\mathbf{Q}\left(\mathbf{V}_i\right)\varepsilon_i | \mathbf{V}_i\right] = E\left[\mathbf{Q}\left(\mathbf{V}_i\right)\left\{Y_i - \sum_{j=1}^{d} g_j\left(\mathbf{U}_i^c, \mathbf{U}_i^d\right) X_{i,j}\right\} | \mathbf{V}_i\right] = 0. \tag{2.2}$$

Following Cai and Li (2008), we propose an estimation procedure to combine the orthogonality condition in (2.2) 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 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 $\mathbf{u} = \left(\mathbf{u}^c, \mathbf{u}^d\right) \in \mathbb{R}^{p_c} \times \mathbb{R}^{p_d}$. To define the kernel weight function, we focus on the case for which there is no natural ordering in $\mathbf{U}_i^d$. Define

$$l\left(U_{i,t}^d, u_t^d, \lambda_t\right) = \begin{cases} 1 & \text{if } \mathbf{U}_{i,t}^d = u_t^d, \\ \lambda_t & \text{if } \mathbf{U}_{i,t}^d \neq u_t^d, \end{cases} \tag{2.3}$$

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 a uniform weight function. We define the product kernel for the discrete random variables by

$$L\left(\mathbf{U}_i^d, \mathbf{u}^d, \lambda\right) = L_\lambda\left(\mathbf{U}_i^d - \mathbf{u}^d\right) = \prod_{t=1}^{p_d} l\left(U_{i,t}^d, u_t^d, \lambda_t\right). \tag{2.4}$$

For the continuous random variables, we use $w(\cdot)$ 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 $\mathbf{h} = (h_1, \cdots, h_{p_c})'$ denotes the p_c-vector of smoothing parameters. We then define the kernel weight function $K_{\mathbf{h}\lambda,i\mathbf{u}}$ by

$$K_{\mathbf{h}\lambda,i\mathbf{u}} = W_{\mathbf{h},i\mathbf{u}^c} L_{\lambda,i\mathbf{u}^d} \tag{2.5}$$

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

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 with respect to $\mathbf{u}^c$. 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 $\mathbf{u}^c$. Denote by $\ddot{g}_j\left(\mathbf{u}^c, \mathbf{u}^d\right) = \partial^2 g_j\left(\mathbf{u}^c, \mathbf{u}^d\right)/\left(\partial \mathbf{u}^c \partial \mathbf{u}^{c\prime}\right)$ 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 of $g_j\left(\mathbf{U}_i^c, \mathbf{u}^d\right)$ around $\left(\mathbf{u}^c, \mathbf{u}^d\right)$ that

$$\sum_{j=1}^{d} g_j\left(\mathbf{U}_i^c, \mathbf{u}^d\right) X_{i,j} \approx \sum_{j=1}^{d} \left[g_j\left(\mathbf{u}^c, \mathbf{u}^d\right) + \dot{g}_j\left(\mathbf{u}^c, \mathbf{u}^d\right)'\left(\mathbf{U}_i^c - \mathbf{u}^c\right)\right] X_{i,j} = \alpha\left(\mathbf{u}\right)' \xi_{i,\mathbf{u}} \tag{2.6}$$

where $\alpha(\mathbf{u}) = \left(g_1\left(\mathbf{u}\right), \cdots, g_d\left(\mathbf{u}\right), \dot{g}_1(\mathbf{u})', \cdots, \dot{g}_d\left(\mathbf{u}\right)'\right)'$ and $\xi_{i,\mathbf{u}} = \begin{pmatrix} \mathbf{X}_i \\ \mathbf{X}_i \otimes \left(\mathbf{U}_i^c - \mathbf{u}^c\right) \end{pmatrix}$ are both $d\left(p_c + 1\right) \times 1$ vectors.

Motivated by the idea of local linear fitting, for the "global" instrument $\mathbf{Q}\left(\mathbf{V}_i\right)$ we define its associated "local" version as

$$\mathbf{Q}_{\mathbf{h},i\mathbf{u}} = \begin{pmatrix} \mathbf{Q}\left(\mathbf{V}_i\right) \\ \mathbf{Q}\left(\mathbf{V}_i\right) \otimes \left(\mathbf{U}_i^c - \mathbf{u}^c\right)/\mathbf{h} \end{pmatrix}. \tag{2.7}$$

Clearly, the dimension of $\mathbf{Q}_{\mathbf{h},i\mathbf{u}}$ is $k\left(p_c + 1\right)$ as $\mathbf{Q}\left(\mathbf{V}_i\right)$ is a $k \times 1$ vector. In view of the fact that the orthogonality condition in (2.2) continues to hold when we replace $\left(\mathbf{Q}\left(\mathbf{V}_i\right), \mathbf{V}_i\right)$ by $\left(\mathbf{Q}_{\mathbf{h},i\mathbf{u}}, \mathbf{U}_i\right)$, we approximate $E[\mathbf{Q}_{\mathbf{h},i\mathbf{u}}\{Y_i - \sum_{j=1}^{d} g_j\left(\mathbf{U}_i^c, \mathbf{U}_i^d\right) X_{i,j}\}|\mathbf{U}_i = \mathbf{u}]$ by its sample analog

$$\frac{1}{n}\sum_{i=1}^{n} \mathbf{Q}_{\mathbf{h},i\mathbf{u}}\left[Y_i - \alpha\left(\mathbf{u}\right)'\xi_{i,\mathbf{u}}\right] K_{\mathbf{h}\lambda,i\mathbf{u}} = \frac{1}{n}\mathbf{Q}_{\mathbf{h}}\left(\mathbf{u}\right)'\mathbf{K}_{\mathbf{h}\lambda}\left(\mathbf{u}\right)\left[\mathbf{Y} - \xi\left(\mathbf{u}\right)\alpha\right]$$

where $\mathbf{Y} = \left(Y_1, \cdots, Y_n\right)'$, $\xi\left(\mathbf{u}\right) = \left(\xi_{1,\mathbf{u}}, \cdots, \xi_{n,\mathbf{u}}\right)'$, $\alpha = \alpha\left(\mathbf{u}\right)$, $\mathbf{K}_{\mathbf{h}\lambda}\left(\mathbf{u}\right) = \text{diag}(K_{\mathbf{h}\lambda,1\mathbf{u}}, \cdots, K_{\mathbf{h}\lambda,n\mathbf{u}})$, and $\mathbf{Q}_{\mathbf{h}}\left(\mathbf{u}\right) = (\mathbf{Q}_{\mathbf{h},1\mathbf{u}}, \cdots, \mathbf{Q}_{\mathbf{h},n\mathbf{u}})'$. To obtain estimates of g_j and $\dot{g}_j$, we can choose α to minimize the following local linear GMM criterion function

$$\frac{1}{n}\left[\mathbf{Q}_{\mathbf{h}}\left(\mathbf{u}\right)'\mathbf{K}_{\mathbf{h}\lambda}\left(\mathbf{u}\right)\left(\mathbf{Y} - \xi\left(\mathbf{u}\right)\alpha\right)\right]'\mathbf{\Psi}_n\left(\mathbf{u}\right)^{-1}\left[\mathbf{Q}_{\mathbf{h}}\left(\mathbf{u}\right)'\mathbf{K}_{\mathbf{h}\lambda}\left(\mathbf{u}\right)\left(\mathbf{Y} - \xi\left(\mathbf{u}\right)\alpha\right)\right], \tag{2.8}$$

where $\mathbf{\Psi}_n(\mathbf{u})$ is a symmetric $k(p_c+1)\times k(p_c+1)$ weight matrix that is positive definite for large n. Clearly, the solution to the above minimization problem is given by

$$\begin{aligned}\widehat{\alpha}_{\mathbf{\Psi}_n}(\mathbf{u};\mathbf{h},\lambda) &= \left[\xi(\mathbf{u})'\mathbf{K}_{\mathbf{h}\lambda}(\mathbf{u})\mathbf{Q}_{\mathbf{h}}(\mathbf{u})\mathbf{\Psi}_{\mathbf{n}}(\mathbf{u})^{-1}\mathbf{Q}_{\mathbf{h}}(\mathbf{u})'\mathbf{K}_{\mathbf{h}\lambda}(\mathbf{u})\xi(\mathbf{u})\right]^{-1} \\ &\quad\times\xi(\mathbf{u})'\mathbf{K}_{\mathbf{h}\lambda}(\mathbf{u})\mathbf{Q}_{\mathbf{h}}(\mathbf{u})\mathbf{\Psi}_{\mathbf{n}}(\mathbf{u})^{-1}\mathbf{Q}_{\mathbf{h}}(\mathbf{u})'\mathbf{K}_{\mathbf{h}\lambda}(\mathbf{u})\mathbf{Y}. \qquad (2.9)\end{aligned}$$

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

$$\widehat{g}_j(\mathbf{u};\mathbf{h},\lambda)=\mathbf{e}'_{j,d(1+p_c)}\widehat{\alpha}_{\mathbf{\Psi}_n}(\mathbf{u};\mathbf{h},\lambda) \text{ and } \widehat{\dot{g}}_j(\mathbf{u};\mathbf{h},\lambda)=\widetilde{\mathbf{e}}_{j,p_c,d(1+p_c)}\widehat{\alpha}_{\mathbf{\Psi}_n}(\mathbf{u};\mathbf{h},\lambda) \text{ for } j=1,\cdots,d. \qquad (2.10)$$

We will study the asymptotic properties of $\widehat{\alpha}_{\mathbf{\Psi}_n}(\mathbf{u};\mathbf{h},\lambda)$ in the next section.

Remark 1 (Choice of IVs) The choice of $\mathbf{Q}(\mathbf{V}_i)$ is important in applications. One can choose it from the union of $\mathbf{Z}_i$ and $\mathbf{U}_i$ (e.g., $\mathbf{Q}(\mathbf{V}_i)=\mathbf{V}_i$) such that a certain identification condition is satisfied. A necessary identification condition is $k\geq d$, which ensures that the dimension of $\mathbf{Q}_{\mathbf{h},i\mathbf{u}}$ is not smaller than the dimension of $\alpha(\mathbf{u})$. Below we will consider the optimal choice of $\mathbf{Q}(\mathbf{V}_i)$ where optimality is in the sense of minimizing the asymptotic variance-covariance (AVC) matrix for the class of local linear GMM estimators given the orthogonal condition in (2.1). We do so by extending the work of Newey (1990, 1993), Baltagi and Li (2002), and Ai and Chen (2003) to our framework, but the latter authors only consider optimal IVs for GMM estimates of *finite* dimensional parameters based on conditional moment conditions.

Remark 2 (Local linear versus local constant GMM estimators) An alternative to the local linear GMM estimator is the local constant GMM estimator; see, e.g., Lewbel (2007) and Tran and Tsionas (2010). In this case, the parameter of interest α 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. As a result, one can set $\mathbf{Q}_{\mathbf{h},i\mathbf{u}}=\mathbf{Q}(\mathbf{V}_i)$ so that there is no distinction between local and global instruments. In addition, our local linear GMM estimator in (2.9) reduces to that of Cai and Li (2008) by setting $\mathbf{\Psi_n}(\mathbf{u})$ to be the identity matrix and choosing $k=d$ 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

$$\mathbf{\Omega}_1(\mathbf{u})=E(\mathbf{Q}(\mathbf{V}_i)\mathbf{X}'_i|\mathbf{U}_i=\mathbf{u}) \text{ and } \mathbf{\Omega}_2(\mathbf{u})=E\left[\mathbf{Q}(\mathbf{V}_i)\mathbf{Q}(\mathbf{V}_i)'\sigma^2(\mathbf{V}_i)|\mathbf{U}_i=\mathbf{u}\right]$$

where $\sigma^2(\mathbf{v})\equiv E\left[\varepsilon_i^2|\mathbf{V}_i=\mathbf{v}\right]$. Let $f_{\mathbf{U}}(\mathbf{u})\equiv f_{\mathbf{U}}\left(\mathbf{u}^c,\mathbf{u}^d\right)$ 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 use $\mathcal{U}^c$ and $\mathcal{U}^d=\Pi_{t=1}^{p_d}\{0,1,\cdots,c_t-1\}$ to denote the support of $\mathbf{U}_i^c$ and $\mathbf{U}_i^d$, respectively.

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 $\delta > 0$. $E\|\mathbf{Q}(\mathbf{V}_i)\mathbf{X}_i'\|^2 < \infty$.

Assumption A3. (i) $\mathcal{U}^c$ is compact. (ii) The functions $f_{\mathbf{U}}(\cdot, \widetilde{\mathbf{u}}^d)$, $\mathbf{\Omega}_1(\cdot, \widetilde{\mathbf{u}}^d)$, and $\mathbf{\Omega}_2(\cdot, \widetilde{\mathbf{u}}^d)$ are continuously differentiable on $\mathcal{U}^c$ for all $\widetilde{\mathbf{u}}^d \in \mathcal{U}^d$. $0 < f_{\mathbf{U}}(\mathbf{u}^c, \mathbf{u}^d) \leq C$ for some $C < \infty$. (iii) The functions $g_j(\cdot, \widetilde{\mathbf{u}}^d)$, $j = 1, \cdots, d$, are second order continuously differentiable on $\mathcal{U}^c$ for all $\widetilde{\mathbf{u}}^d \in \mathcal{U}^d$.

Assumption A4. (i) $\text{rank}(\mathbf{\Omega}_1(\mathbf{u})) = d$, and the $k \times k$ matrix $\mathbf{\Omega}_2(\mathbf{u})$ is positive definite. (ii) $\mathbf{\Psi}_n(\mathbf{u}) = \mathbf{\Psi}(\mathbf{u}) + o_P(1)$, where $\mathbf{\Psi}(\mathbf{u})$ is symmetric and positive definite.

Assumption A5. The kernel function $w(\cdot)$ is a probability density function (PDF) that is symmetric, bounded, and has compact support $[-c_w, c_w]$. It satisfies the Lipschitz condition $|w(v_1) - w(v_2)| \leq C_w|v_1 - v_2|$ for all $v_1, v_2 \in [-c_w, c_w]$.

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

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 $\mathbf{\Psi}_n$ 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, the 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

$$\mathbf{\Phi}(\mathbf{u}) = f_{\mathbf{U}}(\mathbf{u}) \begin{pmatrix} \mathbf{\Omega}_1(\mathbf{u}) & \mathbf{0}_{k \times dp_c} \\ \mathbf{0}_{kp_c \times d} & \mu_{2,1}\mathbf{\Omega}_1(\mathbf{u}) \otimes \mathbf{I}_{p_c} \end{pmatrix}, \text{ and} \tag{3.1}$$

$$\mathbf{\Upsilon}(\mathbf{u}) = f_{\mathbf{U}}(\mathbf{u}) \begin{pmatrix} \mu_{0,2}^{p_c}\mathbf{\Omega}_2(\mathbf{u}) & \mathbf{0}_{k \times kp_c} \\ \mathbf{0}_{kp_c \times k} & \mu_{2,2}\mathbf{\Omega}_2(\mathbf{u}) \otimes \mathbf{I}_{p_c} \end{pmatrix}. \tag{3.2}$$

Clearly, $\mathbf{\Phi}(\mathbf{u})$ is a $k(1+p_c) \times d(1+p_c)$ matrix and $\mathbf{\Upsilon}(\mathbf{u})$ is $k(1+p_c) \times k(1+p_c)$ matrix.

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

$$\mathbf{B}(\mathbf{u}; \mathbf{h}, \lambda) = \left\{ \begin{pmatrix} \frac{1}{2}\mu_{2,1} f_{\mathbf{U}}(\mathbf{u}) \mathbf{\Omega}_1(\mathbf{u}) \mathbf{A}(\mathbf{u}; \mathbf{h}) \\ \mathbf{0}_{kp_c \times 1} \end{pmatrix} \right.$$
$$\left. + \sum_{\widetilde{\mathbf{u}}^d \in \mathcal{U}^d} \sum_{s=1}^{p_d} \lambda_s I_s(\mathbf{u}^d, \widetilde{\mathbf{u}}^d) f_{\mathbf{U}}(\mathbf{u}^c, \widetilde{\mathbf{u}}^d) \begin{pmatrix} \mathbf{\Omega}_1(\mathbf{u}^c, \widetilde{\mathbf{u}}^d)(\mathbf{g}(\mathbf{u}^c, \widetilde{\mathbf{u}}^d) - \mathbf{g}(\mathbf{u}^c, \mathbf{u}^d)) \\ -\mu_{2,1}(\mathbf{\Omega}_1(\mathbf{u}^c, \widetilde{\mathbf{u}}^d) \otimes \mathbf{I}_{p_c}) \dot{\mathbf{g}}(\mathbf{u}^c, \mathbf{u}^d) \end{pmatrix} \right\}, \tag{3.3}$$

where $\mathbf{A}(\mathbf{u}; \mathbf{h}) = (\sum_{s=1}^{p_c} h_s^2 g_{1,ss}(\mathbf{u}), \cdots, \sum_{s=1}^{p_c} h_s^2 g_{d,ss}(\mathbf{u}))'$, $\mathbf{g}(\mathbf{u}) = (g_1(\mathbf{u}), \cdots ., g_d(\mathbf{u}))'$, 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* $\sqrt{n\mathbf{h}!}\{\mathbf{H}[\widehat{\alpha}_{\mathbf{\Psi}_n}(\mathbf{u};\mathbf{h},\lambda)-\alpha(\mathbf{u})]-\left(\mathbf{\Phi}'\mathbf{\Psi}^{-1}\mathbf{\Phi}\right)^{-1}\mathbf{\Phi}'\mathbf{\Psi}^{-1}\mathbf{B}(\mathbf{u};\mathbf{h},\lambda)\} \xrightarrow{d} N(0,\left(\mathbf{\Phi}'\mathbf{\Psi}^{-1}\mathbf{\Phi}\right)^{-1}\mathbf{\Phi}'\mathbf{\Psi}^{-1}\mathbf{\Upsilon}\mathbf{\Psi}^{-1}\mathbf{\Phi}\left(\mathbf{\Phi}'\mathbf{\Psi}^{-1}\mathbf{\Phi}\right)^{-1})$, *where we have suppressed the dependence of* $\mathbf{\Phi}$, $\mathbf{\Psi}$, *and* $\mathbf{\Upsilon}$ *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 the weight matrix) To minimize the AVC matrix of $\widehat{\alpha}_{\mathbf{\Psi}_n}$, we can choose $\mathbf{\Psi}_n(\mathbf{u})$ as a consistent estimate of $\Upsilon(\mathbf{u})$, say $\widehat{\mathbf{\Upsilon}}(\mathbf{u})$. Then the AVC matrix of $\widehat{\alpha}_{\widehat{\mathbf{\Upsilon}}}(\mathbf{u};\mathbf{h},\lambda)$ is given by $\mathbf{\Sigma}(\mathbf{u})=[\mathbf{\Phi}(\mathbf{u})'\mathbf{\Upsilon}(\mathbf{u})^{-1}\mathbf{\Phi}(\mathbf{u})]^{-1}$, which is the minimum AVC matrix conditional on the choice of the global instruments $\mathbf{Q}(\mathbf{V}_i)$. Let $\widetilde{\alpha}(\mathbf{u})$ be a preliminary estimate of $\alpha(\mathbf{u})$ by setting $\mathbf{\Psi}_n(\mathbf{u})=\mathbf{I}_{k(p_c+1)}$. 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 $\widetilde{\alpha}(\mathbf{u})$. Let

$$\widehat{\mathbf{\Upsilon}}(\mathbf{u})=\frac{\mathbf{h}!}{n}\sum_{i=1}^{n}\begin{pmatrix}\mathbf{Q}_i\mathbf{Q}_i'\widetilde{\varepsilon}_i(\mathbf{u})^2 & (\mathbf{Q}_i\mathbf{Q}_i')\otimes\eta_i(\mathbf{u}^c)'\widetilde{\varepsilon}_i(\mathbf{u})^2\\ (\mathbf{Q}_i\mathbf{Q}_i')\otimes\eta_i(\mathbf{u}^c)\widetilde{\varepsilon}_i(\mathbf{u})^2 & (\mathbf{Q}_i\mathbf{Q}_i')\otimes[\eta_i(\mathbf{u}^c)\eta_i(\mathbf{u}^c)']\widetilde{\varepsilon}_i(\mathbf{u})^2\end{pmatrix}K^2_{\mathbf{h}\lambda,i\mathbf{u}}$$

where $\mathbf{Q}_i\equiv\mathbf{Q}(\mathbf{V}_i)$ and $\eta_i(\mathbf{u}^c)\equiv(\mathbf{U}_i^c-\mathbf{u}^c)/\mathbf{h}$. It is easy to show that under Assumptions A1-A6 $\widehat{\mathbf{\Upsilon}}(\mathbf{u})=\mathbf{\Upsilon}(\mathbf{u})+o_P(1)$. Alternatively, we can obtain the estimates $\widetilde{\alpha}(u)$ 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 $\widehat{\mathbf{\Upsilon}}(\mathbf{u})$ by $\widetilde{\varepsilon}_i$ also yields a consistent estimate of $\mathbf{\Upsilon}(\mathbf{u})$, but this needs preliminary estimation of the functional coefficients at all data points and thus is much more computationally expensive. By choosing $\mathbf{\Psi}_n(\mathbf{u})=\widehat{\mathbf{\Upsilon}}(\mathbf{u})$, we denote the resulting local linear GMM estimator of $\alpha(\mathbf{u})$ as $\widehat{\alpha}_{\widehat{\mathbf{\Upsilon}}}(\mathbf{u};\mathbf{h},\lambda)$. We summarize the asymptotic properties of this estimator in the following corollary, whose proof is straightforward.

Corollary 3.2 *Suppose that Assumptions A1-A4(i) and A5-A6 hold. Then* $\sqrt{n\mathbf{h}!}\{\mathbf{H}[\widehat{\alpha}_{\widehat{\mathbf{\Upsilon}}}(\mathbf{u};\mathbf{h},\lambda)-\alpha(\mathbf{u})]-\left(\mathbf{\Phi}'\mathbf{\Upsilon}^{-1}\mathbf{\Phi}\right)^{-1}\mathbf{\Phi}'\mathbf{\Upsilon}^{-1}\mathbf{B}(\mathbf{u};\mathbf{h},\lambda)\} \xrightarrow{d} N(0,\left(\mathbf{\Phi}'\mathbf{\Upsilon}^{-1}\mathbf{\Phi}\right)^{-1})$. *In particular,* $\sqrt{n\mathbf{h}!}\{\widehat{\mathbf{g}}_{\widehat{\mathbf{\Upsilon}}}(\mathbf{u};\mathbf{h},\lambda)-\mathbf{g}(\mathbf{u})-f_{\mathbf{U}}(\mathbf{u})^{-1}[\mathbf{\Omega}_1'(\mathbf{u})\mathbf{\Omega}_2(\mathbf{u})^{-1}\mathbf{\Omega}_1(\mathbf{u})]^{-1}\mathbf{\Omega}_1'(\mathbf{u})\mathbf{\Omega}_2(\mathbf{u})^{-1}\mathbf{B}_0(\mathbf{u};\mathbf{h},\lambda)\} \xrightarrow{d} N(0,\mu_{0,2}^{p_c}f_{\mathbf{U}}(\mathbf{u})^{-1}[\mathbf{\Omega}_1'(\mathbf{u})\mathbf{\Omega}_2(\mathbf{u})^{-1}\mathbf{\Omega}_1(\mathbf{u})]^{-1})$, *where* $\widehat{\mathbf{g}}_{\widehat{\mathbf{\Upsilon}}}(\mathbf{u};\mathbf{h},\lambda)$ *and* $\mathbf{B}_0(\mathbf{u};\mathbf{h},\lambda)$ *denote the first* d *elements of* $\widehat{\alpha}_{\widehat{\mathbf{\Upsilon}}}(\mathbf{u};\mathbf{h},\lambda)$ *and* $\mathbf{B}(\mathbf{u};\mathbf{h},\lambda)$, *respectively.*

Remark 4 (Asymptotic independence between estimates of functional coefficients and their first order derivatives) Theorem 3.1 indicates that, for the general choice of $\mathbf{\Psi}_n$ 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 $\mathbf{\Psi}_n$ as an asymptotically block diagonal matrix (i.e., the limit of $\mathbf{\Psi}_n$ is block diagonal) as in Corollary 3.2, then we have asymptotic independence between the estimates of $\mathbf{g}(\mathbf{u})$ and $\dot{\mathbf{g}}(\mathbf{u})$. If further $k=d$, then the formulae for the asymptotic bias and variance of $\widehat{\mathbf{g}}_{\widehat{\mathbf{\Upsilon}}}(\mathbf{u})$ can be simplified to $\mathbf{\Omega}_1(\mathbf{u})^{-1}\mathbf{B}_0(\mathbf{u};\mathbf{h},\lambda)/f_{\mathbf{U}}(\mathbf{u})$ and $\mu_{0,2}^{p_c}\mathbf{\Omega}_1(\mathbf{u})^{-1}\mathbf{\Omega}_2(\mathbf{u})(\mathbf{\Omega}_1(\mathbf{u})^{-1})'/f_{\mathbf{U}}(\mathbf{u})$, respectively.

3.3 Optimal choice of global instruments

To derive the optimal global instruments for the estimation of $\alpha(\mathbf{u})$ based on the conditional moment restriction given in (2.1), define

$$\mathbf{Q}^*(\mathbf{V}_i)=\mathbf{C}E(\mathbf{X}_i|\mathbf{V}_i)/\sigma^2(\mathbf{V}_i) \tag{3.4}$$

where $\mathbf{C}$ is any nonsingular nonrandom $d\times d$ matrix. As $\mathbf{Q}^*(\mathbf{V}_i)$ is a $d\times 1$ vector, the weight matrix $\mathbf{\Psi}_n$ does not play a role. It is easy to verify that the local linear GMM estimator corresponding to this

choice of IV has the following AVC matrix

$$\begin{aligned}\boldsymbol{\Sigma}^{*}(\mathbf{u}) &= f_{\mathbf{U}}^{-1}(\mathbf{u})\begin{pmatrix}\mu_{0,2}^{p_c}\boldsymbol{\Omega}^{*}(\mathbf{u})^{-1} & \mathbf{0}_{d\times dp_c}\\ \mathbf{0}_{dp_c\times d} & \left(\mu_{2,2}/\mu_{2,1}^{2}\right)\boldsymbol{\Omega}^{*}(\mathbf{u})^{-1}\otimes\mathbf{I}_{p_c}\end{pmatrix}\\ &= f_{\mathbf{U}}^{-1}(\mathbf{u})\,\mathcal{K}\begin{pmatrix}\mu_{0,2}^{p_c}\boldsymbol{\Omega}^{*}(\mathbf{u}) & \mathbf{0}_{d\times dp_c}\\ \mathbf{0}_{dp_c\times d} & \mu_{2,2}\boldsymbol{\Omega}^{*}(\mathbf{u})\otimes\mathbf{I}_{p_c}\end{pmatrix}^{-1}\mathcal{K},\end{aligned}$$

where $\boldsymbol{\Omega}^{*}(\mathbf{u})\equiv E[E(\mathbf{X}_i|\mathbf{V}_i)E(\mathbf{X}_i|\mathbf{V}_i)'\sigma^{-2}(\mathbf{V}_i)|\mathbf{U}_i=\mathbf{u}]$ and $\mathcal{K}\equiv\begin{pmatrix}\mu_{0,2}^{p_c}\mathbf{I}_d & \mathbf{0}_{d\times dp_c}\\ \mathbf{0}_{dp_c\times d} & (\mu_{2,2}/\mu_{2,1})\mathbf{I}_{dp_c}\end{pmatrix}$. Noting that $\boldsymbol{\Sigma}^{*}(\mathbf{u})$ is free of the choice of $\mathbf{C}$, hereafter we simply take $\mathbf{C}=\mathbf{I}_d$ and continue to use $\mathbf{Q}^{*}(\mathbf{V}_i)$ to denote $E(\mathbf{X}_i|\mathbf{V}_i)/\sigma^{2}(\mathbf{V}_i)$. We now follow Newey (1993) and argue that $\mathbf{Q}^{*}(\mathbf{V}_i)$ is the optimal IV in the sense of minimizing the AVC matrix of our local linear GMM estimator of $\alpha(\mathbf{u})$ among the class of all local linear GMM estimators.

Let $\mathbf{Q}_i^{*}\equiv\mathbf{Q}^{*}(\mathbf{V}_i)$. Define $m_{i,\mathbf{Q}}\equiv\boldsymbol{\Phi}(\mathbf{u})'\boldsymbol{\Psi}(\mathbf{u})^{-1}\begin{pmatrix}\mathbf{Q}_i\\ \mathbf{Q}_i\otimes\eta_i(\mathbf{u}^c)\end{pmatrix}\varepsilon_i K_{\mathbf{h}\lambda,i\mathbf{u}}(\mathbf{h}!)^{1/2}$ and $m_{i,\mathbf{Q}^{*}}\equiv\mathcal{K}^{-1}\begin{pmatrix}\mathbf{Q}_i^{*}\\ \mathbf{Q}_i^{*}\otimes\eta_i(\mathbf{u}^c)\end{pmatrix}\varepsilon_i K_{\mathbf{h}\lambda,i\mathbf{u}}(\mathbf{h}!)^{1/2}$. By the law of iterated expectations and moment calculations,

$$\begin{aligned}E\left(m_{i,\mathbf{Q}^{*}}m_{i,\mathbf{Q}^{*}}'\right) &= \mathcal{K}^{-1}E\left[\begin{pmatrix}\mathbf{Q}_i^{*}\mathbf{Q}_i^{*\prime} & (\mathbf{Q}_i^{*}\mathbf{Q}_i^{*\prime})\otimes\eta_i(\mathbf{u}^c)'\\ (\mathbf{Q}_i^{*}\mathbf{Q}_i^{*\prime})\otimes\eta_i(\mathbf{u}^c) & (\mathbf{Q}_i^{*}\mathbf{Q}_i^{*\prime})\otimes[\eta_i(\mathbf{u}^c)\eta_i(\mathbf{u}^c)']\end{pmatrix}\sigma^{2}(\mathbf{V}_i)K_{\mathbf{h}\lambda,i\mathbf{u}}^{2}\mathbf{h}!\right]\mathcal{K}^{-1}\\ &= \mathcal{K}^{-1}E\left[\begin{pmatrix}\boldsymbol{\Omega}^{*}(\mathbf{U}_i) & \boldsymbol{\Omega}^{*}(\mathbf{U}_i)\otimes\eta_i(\mathbf{u}^c)'\\ \boldsymbol{\Omega}^{*}(\mathbf{U}_i)\otimes\eta_i(\mathbf{u}^c) & \boldsymbol{\Omega}^{*}(\mathbf{U}_i)\otimes[\eta_i(\mathbf{u}^c)\eta_i(\mathbf{u}^c)']\end{pmatrix}K_{\mathbf{h}\lambda,i\mathbf{u}}^{2}\mathbf{h}!\right]\mathcal{K}^{-1}\\ &= f_{\mathbf{U}}(\mathbf{u})\,\mathcal{K}^{-1}\begin{pmatrix}\mu_{0,2}^{p_c}\boldsymbol{\Omega}^{*}(\mathbf{u}) & \mathbf{0}_{d\times dp_c}\\ \mathbf{0}_{dp_c\times d} & \mu_{2,2}\boldsymbol{\Omega}^{*}(\mathbf{u})\otimes\mathbf{I}_{p_c}\end{pmatrix}\mathcal{K}^{-1}+o(1)\\ &= \left[\boldsymbol{\Sigma}^{*}(\mathbf{u})\right]^{-1}+o(1).\end{aligned}$$

Similarly, $E(m_{i,\mathbf{Q}}m_{i,\mathbf{Q}}')=\boldsymbol{\Phi}(\mathbf{u})'\boldsymbol{\Psi}(\mathbf{u})^{-1}\boldsymbol{\Upsilon}(\mathbf{u})\boldsymbol{\Psi}(\mathbf{u})^{-1}\boldsymbol{\Phi}(\mathbf{u})+o(1)$ and $E(m_{i,\mathbf{Q}}m_{i,\mathbf{Q}^{*}}')=\boldsymbol{\Phi}(\mathbf{u})'\boldsymbol{\Psi}(\mathbf{u})^{-1}\boldsymbol{\Phi}(\mathbf{u})+o(1)$. It follows that

$$\begin{aligned}&\left(\boldsymbol{\Phi}'\boldsymbol{\Psi}^{-1}\boldsymbol{\Phi}\right)^{-1}\boldsymbol{\Phi}'\boldsymbol{\Psi}^{-1}\boldsymbol{\Upsilon}\boldsymbol{\Psi}^{-1}\boldsymbol{\Phi}\left(\boldsymbol{\Phi}'\boldsymbol{\Psi}^{-1}\boldsymbol{\Phi}\right)^{-1}-\boldsymbol{\Sigma}^{*}(\mathbf{u})\\ =\;&\left[E(m_{i,\mathbf{Q}}m_{i,\mathbf{Q}^{*}}')\right]^{-1}E(m_{i,\mathbf{Q}}m_{i,\mathbf{Q}}')\left[E(m_{i,\mathbf{Q}}m_{i,\mathbf{Q}^{*}}')\right]^{-1}-\left[E(m_{i,\mathbf{Q}^{*}}m_{i,\mathbf{Q}^{*}}')\right]^{-1}+o(1)\\ =\;&\left[E(m_{i,\mathbf{Q}}m_{i,\mathbf{Q}^{*}}')\right]^{-1}\left\{E(m_{i,\mathbf{Q}}m_{i,\mathbf{Q}}')-E(m_{i,\mathbf{Q}}m_{i,\mathbf{Q}^{*}}')\left[E(m_{i,\mathbf{Q}^{*}}m_{i,\mathbf{Q}^{*}}')\right]^{-1}E(m_{i,\mathbf{Q}^{*}}m_{i,\mathbf{Q}}')\right\}\\ &\times\left[E(m_{i,\mathbf{Q}^{*}}m_{i,\mathbf{Q}}')\right]^{-1}+o(1)\\ =\;&E[R_iR_i']+o(1)\end{aligned}$$

where $R_i\equiv[E(m_{i,\mathbf{Q}}m_{i,\mathbf{Q}^{*}}')]^{-1}\{m_{i,\mathbf{Q}}-E(m_{i,\mathbf{Q}}m_{i,\mathbf{Q}^{*}}')[E(m_{i,\mathbf{Q}^{*}}m_{i,\mathbf{Q}^{*}}')]^{-1}m_{i,\mathbf{Q}^{*}}\}$. The positive semi-definiteness of $E[R_iR_i']$ implies that the local linear GMM estimator of $\alpha(\mathbf{u})$ based on $\mathbf{Q}^{*}(\mathbf{V}_i)$ is asymptotically optimal among the class of all local linear GMM estimators of $\alpha(\mathbf{u})$. In this sense, we say that $\mathbf{Q}^{*}(\mathbf{V}_i)$ is the optimal IV within the class.

Remark 5 (Comparison with the optimal IV for parametric GMM estimation) Consider a simple parametric model in which $Y_i=\beta'\mathbf{X}_i+\varepsilon_i$ and $E(\varepsilon_i|\mathbf{V}_i)=0$ a.s. The results in Newey (1993) imply that the optimal IV for the GMM estimation of β is given by $E(\mathbf{X}_i|\mathbf{V}_i)/E(\varepsilon_i^{2}|\mathbf{V}_i)$. Such an IV will minimize the AVC matrix of the GMM estimator of β among the class of all GMM estimators based on the given conditional moment restriction. The optimal IV $\mathbf{Q}^{*}(\mathbf{V}_i)$ for our functional coefficient model in

(2.1) takes the same functional form. In addition, it is worth mentioning that the squared asymptotic bias for a parametric GMM estimate is asymptotically negligible in comparison with its asymptotic variance, whereas for our nonparametric GMM estimate it is not unless one uses an undersmoothing bandwidth sequence in the estimation. Different choices of IVs yield different asymptotic bias formulae and it is extremely hard to compare them. Even if the use of the optimal IV $\mathbf{Q}^*(\mathbf{V}_i)$ minimizes the asymptotic variance of the estimate of each element in $\alpha(\mathbf{u})$, it may not minimize the asymptotic mean squared error (AMSE). We think that this is an important reason why we cannot find any application of optimal IV for *nonparametric* GMM estimation in the literature. Another reason is also essential. To apply the optimal IV, we have to estimate both $E(\mathbf{X}_i|\mathbf{V}_i)$ and $\sigma^2(\mathbf{V}_i)$ nonparametrically, and the theoretical justification is technically challenging and beyond the scope of this paper. The similar results and remarks also hold when one considers the optimal IV for local constant GMM estimation. In particular, $\mathbf{Q}^*(\mathbf{V}_i)$ is also the optimal IV for local constant estimation of $\mathbf{g}(\mathbf{u})$. We will compare local linear and local constant GMM estimates based on non-optimal and estimated optimal IVs through Monte Carlo simulations.

3.4 Data-dependent bandwidth

By Theorem 3.1 we can define the asymptotic mean integrated squared error (AMISE) of $\{\widehat{g}_j(\mathbf{u}), j = 1, ..., d\}$, and choose h_r $(r = 1, \cdots, p_c)$ and λ_s $(s = 1, \cdots, p_d)$ to minimize it. By an argument similar to Li and Racine (2008), it is easy to obtain the optimal rates of bandwidths in terms of minimizing the AMISE: $h_r \propto n^{-1/(4+p_c)}$ and $\lambda_s \propto n^{-2/(4+p_c)}$ for $r = 1, \cdots, p_c$ and $s = 1, \cdots, p_d$. Nevertheless, the exact formula for the optimal smoothing parameters 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 $(\mathbf{h}, \lambda)$ to minimize the following least squares cross validation criterion function

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

where $\widehat{g}_j^{(-i)}(\mathbf{U}_i; \mathbf{h}, \lambda)$ is the leave-one-out functional coefficient estimate of $g_j(\mathbf{U}_i)$ using bandwidth $(\mathbf{h}, \lambda)$, 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 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 $(\mathbf{h}, \lambda)$ 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.

In the following we argue that the result in Theorem 3.1 continues to hold when the nonstochastic bandwidth $(\mathbf{h}, \lambda)$ is replaced by some data-dependent stochastic bandwidth, say, $\widehat{\mathbf{h}} \equiv (\widehat{h}_1, ..., \widehat{h}_{p_c})'$ and $\widehat{\lambda} \equiv (\widehat{\lambda}_1, ..., \widehat{\lambda}_{p_d})'$. Following Li and Li (2010) we assume that $(\widehat{h}_r - h_r^0)/h_r^0 = o_P(1)$ and $(\widehat{\lambda}_s - \lambda_s^0)/\lambda_s^0 = o_P(1)$ for $r = 1, \cdots, p_c$ and $s = 1, \cdots, p_d$, where $\mathbf{h}^0 \equiv (h_1^0, ..., h_{p_c}^0)'$ and $\lambda^0 \equiv (\lambda_1^0, ..., \lambda_{p_d}^0)'$ denotes nonstochastic bandwidth sequences for $\mathbf{U}_i^c$ and $\mathbf{U}_i^d$, respectively. For example, $(\widehat{\mathbf{h}}, \widehat{\lambda})$ could be the LSCV bandwidth. If so, one can follow Hall, Racine and Li (2004) and Hall, Li and Racine (2007) and show that the above requirement holds for $(\mathbf{h}^0, \lambda^0)$ which is optimal in minimizing a weighted version of the AMISE of $\{\widehat{g}_j(\mathbf{u}), j = 1, ..., d\}$. The following theorem summarizes the key result.

Theorem 3.3 *Suppose that Assumptions A1-A5 hold. Suppose that* $(\widehat{h}_r - h_r^0)/h_r^0 = o_P(1)$ *and* $(\widehat{\lambda}_s - \lambda_s^0)/\lambda_s^0 = o_P(1)$ *for* $r = 1, \cdots, p_c$ *and* $s = 1, \cdots, p_d$, *where* $\mathbf{h}^0$ *and* λ^0 *satisfy A6. Then* $\sqrt{n\widehat{\mathbf{h}}!}\{\widehat{\mathbf{H}}[\widehat{\alpha}_{\mathbf{\Psi}_n}(\mathbf{u}; \widehat{\mathbf{h}}, \widehat{\lambda}) - \alpha(\mathbf{u})] - \left(\mathbf{\Phi}'\mathbf{\Psi}^{-1}\mathbf{\Phi}\right)^{-1}\mathbf{\Phi}'\mathbf{\Psi}^{-1}\mathbf{B}\left(\mathbf{u}; \mathbf{h}^0, \lambda^0\right)\} \xrightarrow{d} N(0, \left(\mathbf{\Phi}'\mathbf{\Psi}^{-1}\mathbf{\Phi}\right)^{-1}\mathbf{\Phi}'\mathbf{\Psi}^{-1}\,\mathbf{\Upsilon}\mathbf{\Psi}^{-1}\mathbf{\Phi}\left(\mathbf{\Phi}'\mathbf{\Psi}^{-1}\mathbf{\Phi}\right)^{-1})$, *where* $\widehat{\mathbf{H}}$ *is analogously defined as* $\mathbf{H}$ *with* $\mathbf{h}$ *being replaced by* $\hat{\mathbf{h}}$.

4 A Specification Test

In this section, we consider testing the hypothesis that some of the functional coefficients are constant. 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 in this section. Then we can rewrite the model in (2.1) as

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

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

$$\mathbb{H}_0 : \mathbf{g}_1(\mathbf{U}_i) = \theta_1 \text{ a.s. for some parameter } \theta_1 \in \mathbb{R}^{d_1}, \tag{4.2}$$

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

$$Y_i = \theta_1'\mathbf{X}_{1i} + \mathbf{g}_2(\mathbf{U}_i)'\mathbf{X}_{2i} + \varepsilon_i^{(r)} \tag{4.3}$$

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. Then one can 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 $\mathbf{V}_i = (\mathbf{Z}_i', \mathbf{U}_i')'$. 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.

For this reason, we propose a Wald-type statistic that requires only consistent estimation of the unrestricted model. Let $\widehat{\mathbf{g}}_{\mathbf{\Psi}_n}(\mathbf{u})$ denote the first d element of $\widehat{\alpha}_{\mathbf{\Psi}_n}(\mathbf{u}) \equiv \widehat{\alpha}_{\mathbf{\Psi}_n}(\mathbf{u}; \mathbf{h}, \lambda)$. It is the estimator

of $\mathbf{g}(\mathbf{u}) = (\mathbf{g}_1(\mathbf{u})', \mathbf{g}_2(\mathbf{u}))'$. Split $\widehat{\mathbf{g}}_{\mathbf{\Psi}_n}(\mathbf{u})$ as $\widehat{\mathbf{g}}_1(\mathbf{u}) = \widehat{\mathbf{g}}_{1,\mathbf{\Psi}_n}(\mathbf{u})$ and $\widehat{\mathbf{g}}_2(\mathbf{u}) = \widehat{\mathbf{g}}_{2,\mathbf{\Psi}_n}(\mathbf{u})$ so that $\widehat{\mathbf{g}}_l(\mathbf{u})$ estimates $\mathbf{g}_l(\mathbf{u})$ for $l = 1, 2$. 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 $\overline{\widehat{\mathbf{g}}}_1 \equiv n^{-1} \sum_{i=1}^{n} \widehat{\mathbf{g}}_1(\mathbf{U}_i)$. In the next subsection, we show that after being suitably normalized, T_n is asymptotically distributed as $N(0,1)$ under $\mathbb{H}_0$ and diverges to infinity under $\mathbb{H}_1$.

4.2 Asymptotic distribution of the test statistic

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

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

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

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

A7 strengthens A4(ii). It is satisfied if one chooses $\mathbf{\Psi}_n(\mathbf{u})$ as the identity matrix $\mathbf{I}_{k(p_c+1)}$ for all $\mathbf{u}$, in which case $\nu_n = 0$. Alternatively, if one chooses $\mathbf{\Psi}_n(\mathbf{u}) = \widehat{\mathbf{\Upsilon}}(\mathbf{u})$, then one can verify that A7(i) is satisfied with $\nu_n = \|\mathbf{h}\|^2 + \|\lambda\| + (n\mathbf{h}!/\log n)^{-1/2}$. 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 $p_c \leq 3$; one can use higher order local polynomial estimation if $p_c > 3$. A8(iii) requires that undersmoothing bandwidth must be used in order to remove the effect of asymptotic bias of our nonparametric estimators. Without loss of generality, we consider the choice of $\mathbf{\Psi}_n(\mathbf{u})$ as $\widehat{\mathbf{\Upsilon}}(\mathbf{u})$ and set $h_s \propto n^{-1/\delta}$ for $s = 1, ..., p_c$ and $\lambda_t \propto n^{-2/\delta}$ for $t = 1, ..., p_d$; i.e., $h_1, ..., h_{p_c}$ pass to 0 at the same rate and similarly for $\lambda_1, ..., \lambda_{p_d}$. Then the conditions in A8 are all satisfied by setting $\delta \in (1, 4.5)$ for $p_c = 1$, $\delta \in (2, 5)$ for $p_c = 2$, and $\delta \in (3, 5.5)$ for $p_c = 3$.

To proceed, we first consider the consistent estimation of θ_1 under $\mathbb{H}_0$. We estimate it by

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

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

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

where $\mathbf{\Gamma}_{n1}(\mathbf{u})$ is defined in (4.5), and 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 $\mathbb{H}_0$,

$$\sqrt{n}\left(\widehat{\theta}_1 - \theta_1\right) = n^{-1/2} \sum_{i=1}^{n} \mathbf{\Gamma}_{n1}(\mathbf{U}_i) \mathcal{B}_n(\mathbf{U}_i) + n^{-1/2} \sum_{i=1}^{n} \mathbf{\Gamma}_{n1}(\mathbf{U}_i) \mathcal{V}_n(\mathbf{U}_i). \tag{4.8}$$

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 AVC of $\widehat{\theta}_1$.

To characterize the AVC matrix of $\widehat{\theta}_1$, let $\zeta_i \equiv (\mathbf{U}_i', \mathbf{Q}_i', \varepsilon_i)'$, and

$$\varphi(\zeta_i, \zeta_j) \equiv \overline{\mathbf{\Gamma}}_1(\mathbf{U}_i) \begin{pmatrix} \mathbf{Q}_j \varepsilon_j \\ (\mathbf{Q}_j \varepsilon_j) \otimes \eta_j(\mathbf{U}_i^c) \end{pmatrix} K_{\mathbf{h}\lambda, j\mathbf{U}_i}, \text{ and } \overline{\varphi}(\zeta_i) = \int \varphi(\zeta, \zeta_i) dF_\zeta(\zeta), \tag{4.9}$$

where F_ζ denotes the CDF of ζ_i. Let $\mathbf{\Sigma}_{\theta_1} = \lim_{n\to\infty} E[\overline{\varphi}(\zeta_i)\overline{\varphi}(\zeta_i)']$. Straightforward but tedious calculations show that

$$\mathbf{\Sigma}_{\theta_1} = \left(\int\int w(t) w(t-s) dt ds \right)^{p_c} \int \overline{\mathbf{\Gamma}}_1(\mathbf{u}) \begin{pmatrix} \mathbf{\Omega}_2(\mathbf{u}) & \mathbf{0}_{k \times p_c k} \\ \mathbf{0}_{p_c k \times k} & \mathbf{0}_{p_c k \times p_c k} \end{pmatrix} \overline{\mathbf{\Gamma}}_1(\mathbf{u})' f_{\mathbf{U}}(\mathbf{u})^2 dF_{\mathbf{U}}(\mathbf{u}). \tag{4.10}$$

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

Theorem 4.1 *Suppose Assumptions A1-A4(i) and A5-A8 hold. Suppose that* $n^{1/2}(\|\mathbf{h}\|^2 + \|\lambda\|) = o(1)$, $\nu_n(\mathbf{h}!)^{-1/2} = o(1)$, *and* $n(\mathbf{h}!)^2 / \log n \to \infty$ *as* $n \to \infty$. *Then under* $\mathbb{H}_0$, $\sqrt{n}(\widehat{\theta}_1 - \theta_1) \xrightarrow{d} N(0, \mathbf{\Sigma}_{\theta_1})$.

Clearly Theorem 4.1 says that under $\mathbb{H}_0$, $\widehat{\theta}_1$ can consistently estimate θ_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 $\mathbf{\Gamma}_{n1}(\mathbf{U}_i)$ in (4.7) by $\overline{\mathbf{\Gamma}}_1(\mathbf{U}_i)$ has asymptotically negligible effect on the asymptotic normality of $\widehat{\theta}_1$. 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 extra condition on the bandwidth in the theorem.

Let $B_n \equiv n^{-2}(\mathbf{h}!)^{1/2} \sum_{i=1}^n \sum_{j=1}^n ||\varphi(\zeta_i, \zeta_j)||^2$ and $\sigma_0^2 \equiv \lim_{n\to\infty} 2\mathbf{h}! E_j E_l[\int \varphi(\zeta, \zeta_j)' \varphi(\zeta, \zeta_l) dF_\zeta(\zeta)]^2$, where E_j denotes the expectation with respect to ζ_j. The next theorem studies the asymptotic distribution of T_n under $\mathbb{H}_0$.

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

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

where $\widehat{\varphi}_{ij} = \mathbf{\Gamma}_{n1}(\mathbf{U}_i) \begin{pmatrix} \mathbf{Q}_j \widehat{\varepsilon}_j \\ (\mathbf{Q}_j \widehat{\varepsilon}_j) \otimes \eta_j(\mathbf{U}_i^c) \end{pmatrix} K_{\mathbf{h}\lambda, j\mathbf{U}_i}$, 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

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

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, we consider the sequence of Pitman local alternatives

$$\mathbb{H}_1(r_n) : \mathbf{g}_1(\mathbf{U}_i) = \theta_1 + r_n \delta_n(\mathbf{U}_i) \text{ a.s.}$$

where $r_n \to 0$ as $n \to \infty$ and the δ_n's are a sequence of real continuous vector-valued functions such that $\mu_0 \equiv \lim_{n\to\infty} E[\|\delta_n(\mathbf{U}_i) - E[\delta_n(\mathbf{U}_i)]\|^2] < \infty$. 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* $\mu_A \equiv E[\mathbf{g}_1(\mathbf{U}_i) - \theta_1]^2$, *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. So we suggest using a bootstrap method to obtain the bootstrap approximation to the finite-sample distribution of our test statistic under the null. We find that it is easy to adopt the fixed-design wild bootstrap method in the spirit of Hansen (2000) in our framework; see also Su and White (2010) and Su and Ullah (2012). The great advantage of this method lies in the fact that we do not need to mimic some important features (such as dependence or endogeneity structure) in the data generating process and can still justify its asymptotic validity.

We propose to generate the bootstrap version of J_n as follows:

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 weight matrix $\mathbf{\Psi}_n$ and the bandwidth $(\mathbf{h}, \lambda)$, and calculate the unrestricted residuals $\widehat{\varepsilon}_i = Y_i - \widehat{\mathbf{g}}_1(\mathbf{U}_i)' \mathbf{X}_{1i} - \widehat{\mathbf{g}}_2(\mathbf{U}_i)' \mathbf{X}_{2i}$.

2. For $i = 1, ..., n$, generate the wild bootstrap residuals $\varepsilon_i^* = \widehat{\varepsilon}_i e_i$ where e_i's are IID $N(0,1)$.

3. For $i = 1, ..., n$, generate $Y_i^* = \widehat{\theta}_1' \mathbf{X}_{1i} + \widehat{\theta}_2' \mathbf{X}_{2i} + \varepsilon_i^*$ where $\widehat{\theta}_1 \equiv n^{-1} \sum_{i=1}^n \widehat{\mathbf{g}}_1(\mathbf{U}_i)$ and $\widehat{\theta}_2 \equiv n^{-1} \sum_{i=1}^n \widehat{\mathbf{g}}_2(\mathbf{U}_i)$ are the restricted local linear GMM estimates under the null hypothesis $\mathbb{H}_{0s}$: $\mathbf{g}(\mathbf{U}_i) = \theta$ a.s. for some parameter $\theta \in \mathbb{R}^d$.

4. Compute the bootstrap test statistic J_n^* in the same way as J_n by using $\{Y_i^*, \mathbf{U}_i, \mathbf{X}_i, \mathbf{Z}_i\}_{i=1}^n$ and the weight matrix $\mathbf{\Psi}_n^*$.

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 $\mathbb{H}_0 : \mathbf{g}_1(\mathbf{U}_i) = \theta_1$ a.s. if p^* is smaller than the prescribed nominal level of significance.

Note that in Step 3 we impose the null hypothesis $\mathbb{H}_{0s} : \mathbf{g}(\mathbf{U}_i) = \theta$ a.s., which is stronger than $\mathbb{H}_0 : \mathbf{g}_1(\mathbf{U}_i) = \theta_1$ a.s. unless $d_1 = d$. Intuitively speaking, in order to justify the asymptotic validity of the above bootstrap procedure, we need to demonstrate that the bootstrap test statistic J_n^* has the asymptotic distribution $N(0,1)$ *no matter whether the original sample is generated under the null hypothesis (*$\mathbb{H}_0$*) or not*. We will show that conditional on the original sample J_n^* is asymptotically $N(0,1)$, which implies that it is also asymptotically $N(0,1)$ unconditionally. Note that the original test statistic J_n is asymptotically $N(0,1)$ under $\mathbb{H}_0$ and our bootstrap statistic has the same asymptotic distribution. This ensures the correct asymptotic size of our bootstrap test. Further, note that the original test statistic J_n diverges to infinity at the rate $n(\mathbf{h}!)^{1/2}$ under the global alternative hypothesis $\mathbb{H}_1$ whereas the bootstrap test

statistic J_n^* remains asymptotically $N(0,1)$ in this case. This ensures the consistency of our bootstrap test.

By imposing a stronger hypothesis $\mathbb{H}_{0s}$ than the original null hypothesis of interest ($\mathbb{H}_0$), our bootstrap test has both pros and cons. The major pros lie in two aspects. First, one can easily conduct the bootstrap test for testing the constancy of various subvectors of $\mathbf{g}(\cdot)$ in a single step because we can generate the same bootstrap dependent variable once for all and the computation burden is almost identical to the case of testing the constancy of a single subvector of $\mathbf{g}(\cdot)$. Second, one can easily justify the asymptotic validity of our bootstrap method and there is no need to use oversmoothing bandwidth for first stage estimation as in Härdle and Marron (1991). Our simulations indicate this procedure does not result in a loss of power in comparison with the alternative approach by generating Y_i^* through $\widetilde{\theta}_1'\mathbf{X}_{1i} + \widetilde{\mathbf{g}}_2(\mathbf{U}_i)'\mathbf{X}_{2i} + \varepsilon_i^*$ where one imposes the exact null hypothesis to be tested, $\widetilde{\theta}_1 \equiv n^{-1}\sum_{i=1}^n \widetilde{\mathbf{g}}_1(\mathbf{U}_i)$, and $(\widetilde{\mathbf{g}}_1, \widetilde{\mathbf{g}}_2)$ is a preliminary estimate of $(\mathbf{g}_1, \mathbf{g}_2)$. But the justification for the validity of this latter approach would be much harder because one needs to show that the second order derivatives of $\widetilde{\mathbf{g}}_2(\cdot)$ are uniformly well behaved, which typically requires oversmoothing; see Härdle and Marron (1991). The major cons of our bootstrap procedure lie in the potential loss of second order efficiency. In other words, the imposition of a stronger hypothesis than necessary in the bootstrap world is expected to have a second order asymptotic effect. For parametric tests, it is often argued that a bootstrap test based on an asymptotically pivotal statistic may yield a higher order efficiency than a test based on the asymptotic normal or chi-square distributions. For nonparametric tests, it is extremely challenging to demonstrate higher order efficiency for a bootstrap test statistic. Therefore we think higher order efficiency is a less important issue than ensuring the correct asymptotic size and consistency of a bootstrap test. Its formal study is certainly beyond the scope of the current paper.

To show that the bootstrap statistic J_n^* can be used to approximate the asymptotic null distribution of J_n, we follow Li, Hsiao and Zinn (2003) and Su and Ullah (2012) and rely on the notion of *convergence in distribution in probability*, which generalizes the usual convergence in distribution to allow for conditional (random) distribution functions. The following theorem establishes the asymptotic validity of the above bootstrap procedure.

Theorem 4.5 *Suppose Assumptions A1-A4(i) and A5-A8 hold. Suppose that* $\mathbf{\Psi}_n^*(\mathbf{u}) = \mathbf{\Psi}_n(\mathbf{u}) + O_{P^*}(\nu_n)$ *uniformly in* $\mathbf{u}$, *where* P^* *is the probability measure induced by the wild bootstrap. Let* z_α^* *be the* α*-level bootstrap critical value based on* $B \to \infty$ *bootstrap resamples. Then* J_n^* *converges to* $N(0,1)$ *in distribution in probability,* $\lim_{n\to\infty} P(J_n \geq z_\alpha^*) = \alpha$ *under* $\mathbb{H}_0$, $\lim_{n\to\infty} P(J_n \geq z_\alpha^*) \to 1 - \Phi(z_\alpha - \mu_A/\sigma_0)$ *under* $\mathbb{H}_1(n^{-1/2}h^{-p/4})$, *and* $\lim_{n\to\infty} P(J_n \geq z_\alpha^*) = 1$ *under* $\mathbb{H}_1$, *where* z_α *denotes the* $100(1-\alpha)$*th percentile of the standard normal distribution.*

Theorem 4.5 shows that the bootstrap provides an asymptotic valid approximation to the limit null distribution of J_n. This holds as long as we generate the bootstrap data by imposing the null hypothesis. If the null hypothesis does not hold in the observed sample, then J_n explodes at the rate $n(\mathbf{h}!)^{1/2}$ but J_n^* is still well behaved, which intuitively explains the consistency of the bootstrap-based test J_n^*.

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 two data generating processes (DGPs):
DGP 1: $Y_i = (1 + 0.25U_i^{c2} + 0.5U_{i,1}^d + 0.25U_{i,2}^d) + [1 + U_{i,1}^c + \varphi(U_i^c) - 0.5U_{i,1}^d + 0.5U_{i,2}^d]X_i + \sigma_i\varepsilon_i$,
DGP 2: $Y_i = (1 + e^{-U_i^c}) + [1 + 2\sin(U_i^c)]X_i + \sigma_i\varepsilon_i$,
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, $X_i = (Z_i + \tau\varepsilon_i)/\sqrt{1+\tau^2}$, $(Z_i, \varepsilon_i)' \backsim N(0, \mathbf{I}_2)$, and $\varphi(\cdot)$ is the standard normal PDF. Note that there is no discrete random variable in DGP 2. Here we use τ to control the degree of endogeneity; e.g., τ =0.32 and 0.75 indicates that the correlations between X_i and ε_i are 0.3 and 0.6, respectively. We consider both conditionally homoskedastic and heteroskedastic errors. For the homoskedastic case, $\sigma_i = 1$ in both DGPs 1 and 2; for the heteroskedastic case, we specify σ_i as follows

$$\sigma_i = \sqrt{0.1 + 0.5(Z_i^2 + U_i^{c2} + 0.5U_{i,1}^d + U_{i,2}^d)} \text{ in DGP 1 and } \sigma_i = \sqrt{0.1 + 0.5(Z_i^2 + U_i^{c2})} \text{ in DGP 2.}$$

We assume that we observe $\{Y_i, U_i^c, U_{i,1}^d, U_{i,2}^d, X_i, Z_i\}_{i=1}^n$ and $\{Y_i, U_i^c, X_i, Z_i\}_{i=1}^n$ in DGP 1 and DGP 2, respectively. The definitions of the functional coefficients, $g_1(\mathbf{u})$ and $g_2(\mathbf{u})$, in each DGP are self-evident.

We consider six nonparametric estimates for $g_1(\mathbf{u})$ and $g_2(\mathbf{u})$. The first estimate is the local linear estimate of SCU (2009) where the endogeneity of X_i is neglected. The second and third estimates are obtained as our local linear GMM functional coefficient estimators by choosing the global IV respectively as $\mathbf{Q}(\mathbf{V}_i) = [1, Z_i]'$ and local linear estimate of $\mathbf{Q}^*(\mathbf{V}_i) = [1, E(X_i|\mathbf{V}_i)]'/\sigma^2(\mathbf{V}_i)$,respectively, where $\mathbf{V}_i = (Z_i, U_i^c, U_{i,1}^d, U_{i,2}^d)'$ and $\mathbf{V}_i = (Z_i, U_i^c)'$ in DGPs 1 and 2, respectively. Since the dimension of $\mathbf{Q}(\mathbf{V}_i)$ is the same as that of $[1, X_i]'$, the weight matrix Ψ_n does not affect the local linear GMM estimate so that we can simply use the identity weight (IW) matrix as the weight matrix for our second estimate, which also reduces to the estimate of Cai and Li (2008). Similarly, the third estimate is the optimal IV (OIV) estimate which is not influenced by the choice of weight matrix. The fourth and fifth estimates are the local constant analogues of the second and third estimates, respectively; they are also the estimates of Tran and Tsionas (2010, TT) when the IV is chosen to be $\mathbf{Q}(\mathbf{V}_i)$ and the local constant estimate of $\mathbf{Q}^*(\mathbf{V}_i)$, respectively. The sixth estimate is the two-stage local linear estimate of CDXW. Below we will denote these six estimates as SCU, IW$_{ll}$, OIV$_{ll}$, IW$_{lc}$, OIV$_{lc}$, and CDXW in order.

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 [$(\mathbf{h}, \lambda) = (h, \lambda_1, \lambda_2)$ for the conditioning variables $U_i^c, U_{i,1}^d, U_{i,2}^d$ in the functional coefficients in DGP 1 and $\mathbf{h} = h$ for the conditioning variable U_i^c in the functional coefficients in DGP 2]; one is obtained by the LSCV method discussed in Section 3.4, and the other by the simple rule of thumb (ROT): $h = s_{U^c} n^{-1/5}$ in both DGPs 1 and 2, and $\lambda_1 = \lambda_2 = n^{-2/5}$ in DGP 1. Here s_A denotes the sample standard deviation of $\{A_i\}_{i=1}^n$. To estimate the optimal IVs, we need to estimate both $E(X_i|\mathbf{V}_i)$ and $\sigma^2(\mathbf{V}_i)$ by the local linear or local constant method. For both cases, we use the standardized Epanechnikov kernel and undersmoothing ROT bandwidths by specifying $\tilde{\mathbf{h}} = [s_Z n^{-1/5}\ s_{U^c} n^{-1/5}]$ in both DGPs 1 and 2 and $\tilde{\lambda}_1 = \tilde{\lambda}_2 = n^{-2/5}$ in DGP 2 when we regress either X_i or $\widehat{\varepsilon}_i^2$ on $\mathbf{V}_i$. The use of undersmoothing bandwidths helps to eliminate the effect of early stage estimates' bias on the final estimate; see Mammen, Rothe and Schienle (2010). To obtain the CDXW estimate, we need first to obtain the local linear estimate of $E(X_i|\mathbf{V}_i)$ by specifying a similar undersmoothing bandwidth. In addition, we find that the LSCV and ROT choices of $(\mathbf{h}, \lambda)$ yield qualitatively similar results. So we focus on the ROT bandwidth below for brevity.

To evaluate the finite sample performance of different functional coefficient estimates, we calculate both the mean absolute deviation (MAD) and mean squared error (MSE) for each estimate evaluated at

all n data points:

$$MAD_l^{(r)} = \frac{1}{n}\sum_{i=1}^{n}\left|\widehat{g}_l^{(r)}(\mathbf{U}_i) - g_l(\mathbf{U}_i)\right| \text{ and } MSE_l = \frac{1}{n}\sum_{i=1}^{n}\left[\widehat{g}_l^{(r)}(\mathbf{U}_i) - g_l(\mathbf{U}_i)\right]^2$$

where for $l = 1, 2$, $\widehat{g}_l^{(r)}(\cdot)$ is an estimator of $g_l(\cdot)$ in the rth replication by using any one of the above estimation methods. We consider two sample sizes: $n = 100$ and 400.

Table 1 reports the results where the MADs and MSEs are averages over 500 replications for each functional coefficient. We summarize some important findings from Table 1. First, in both homoskedastic and heteroskedastic cases, the SCU estimate without taking into account the endogeneity issue is generally the worst estimate among all six estimates. Exceptions may occur when n is small or no heteroskedasticity is present. Second, in the case of homoskedastic errors, the local GMM estimates obtained by using the estimated optimal IVs for either our local linear method or TT's local constant method may or may not outperform the one using simple IV with identity weight matrix. This is similar to the findings in Altonji and Segal (1996) who show that the use of optimal weights in the GMM estimation may be dominated by the one-step equally weighted GMM estimation in finite samples. Third, in the case of heteroskedastic errors, we observe substantial gain by using the estimated optimal IVs in the local GMM estimation procedure; this is true for both our local linear GMM estimates and TT's local constant estimates. Fourth, for both DGPs under investigation, the local linear method tends to outperform the local constant method. We conjecture this is due to the notorious boundary bias issue associated with the local constant estimates when the support of U_i^c is compact. Fifth, the CDXW estimate generally is outperformed in DGP 1 by both the local linear and local constant estimates with or without using the estimated optimal IVs. For DGP 2, the CDXW may outperform the local constant GMM estimates but not the local linear GMM estimates.

5.2 Tests for the constancy of functional coefficients

We now consider the finite sample performance of our test. To this goal, we modify DGPs 1-2 as follows
DGP 1′: $Y_i = \left[1 + \delta\left(0.25U_i^{c2} + 0.5U_{i,1}^d + 0.25U_{i,2}^d\right)\right] + [1 + \delta(U_{i,1}^c + \varphi(U_i^c) - 0.5U_{i,1}^d + 0.5U_{i,2}^d)]X_i + \sigma_i\varepsilon_i$,
DGP 2′: $Y_i = (1 + \delta e^{-U_i^c}) + [1 + 2\delta\sin(U_i^c)]X_i + \sigma_i\varepsilon_i$,
where all variables are generated as in the above subsection, and we allow δ to take different values to evaluate both the size and power properties of our test. When $\delta = 1$, DGPs 1′ and 2′ reduce to DGPs 1 and 2, respectively. For both DGPs, we consider the following three null hypotheses:

$$\begin{aligned}
\mathbb{H}_{0,1} &: g_1(\mathbf{U}_i) = \theta_1 \text{ a.s.}, \\
\mathbb{H}_{0,2} &: g_2(\mathbf{U}_i) = \theta_2 \text{ a.s.}, \\
\mathbb{H}_{0,12} &: (g_1(\mathbf{U}_i), g_2(\mathbf{U}_i)) = (\theta_1, \theta_2) \text{ a.s.},
\end{aligned} \tag{5.1}$$

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 of the bootstrap to approximate the asymptotic null distribution of our test statistics. Assumption A8(iii) suggests that we need to choose undersmoothing bandwidth sequences. We set $h = cs_{U^c}n^{-1/(p_c+3)}$, and $\lambda_1 = \lambda_2 = cn^{-2/(p_c+3)}$ for DGP 1 and $h = cs_{U^c}n^{-1/(p_c+3)}$ in DGP 2 for different values of c to check the sensitivity of our test to the choice of bandwidth. 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, four values for $\delta : 0$, 0.2, 0.4, and 0.6, and two values for $\tau : 0.32$ and 0.75. For each scenario, we consider 500 replications and 200 bootstrap resamples for each replication.

The results for our bootstrap-based test at 5% nominal level are reported in Table 2. We summarize some important findings from Table 2. First, the empirical levels of our test (corresponding to $\delta = 0$ in the table) generally behave very well for all values of τ and all three null hypotheses, and under both conditional homoskedasticity and heteroskedasticity. The only exception occurs for testing $\mathbb{H}_{0,1}$ in DGP 2 when $n = 100$, in which case the test is moderately undersized. Second, the power of our test is reasonably good in almost all cases– the relatively low power in testing $\mathbb{H}_{0,1}$ in DGP 2 simply reflects the difficulty in testing exponential alternatives of the form $\delta e^{-U_i^c}$. As either δ or the sample size increases, we observe a fast increase of the empirical power. Third, the degree of endogeneity has some effect on the level and power behavior of our test but the direction is not obvious.

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 are 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, a person's 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. The 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 was 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 marital 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. 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. The 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. We choose question (ii) over question (iii). 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). We exclude two observations with reported wage being more than \$200 per hour as extreme outliers. The highest hourly wage in the sample after the exclusion of the two outliers is \$47.5 per hour. 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 we do – the ALS – but use a somewhat different model specification:

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

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 the context of model (6.2). Cai and Xiong (2010) consider the same data set and model specification as in 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 or equal to 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. Thus, for comparative purposes, we primarily focus on work experience being less than 9 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 3.4.

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, 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) 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%. 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.

A potential concern regarding our model specification (6.1) is that some of the categorical variables in $\mathbf{U}$ are endogenous. For example, Lee (2005, p. 431) maintains that "for women, marital status is endogenous and jointly determined with labor supply decisions." Further, Duncan and Leigh (1985) reject the hypothesis of the exogeneity of union status in their wage equation. However, the empirical evidence on endogeneity of these individual characteristics is actually mixed. In particular, Korenman and Neumark (1992) find that female marital status is neither endogenous nor significant in the standard wage equation. Nevertheless, we attempt to address this potential concern by checking the sensitivity of our findings to the choice of $\mathbf{U}$. First, we re-estimate model (6.1) for all four combinations of our four categorical variables in the original $\mathbf{U}$ taken three at a time. Then, we also consider the case when $\mathbf{U}$ contains only indicators for government employment and whether a person is born in Australia, since these two indicators are more likely to be viewed exogenous. When using these alternative choices of $\mathbf{U}$, our original findings are mainly unchanged (results are available upon request). Specifically, the estimated returns to education vary mainly between 12 and 19% except for the observed years of education of 8 years and more. When education exceeds 8 years, the returns to education are estimated to be less than 12%. However, given our above discussion of the latter case, we do not find this result surprising. A slight increase in the average of the estimated returns to education is also anticipated, since fewer controls are included in the functional coefficient of our model similar to CDXW (2006) and Cai and Xiong (2010). We conclude that our findings based on the original $\mathbf{U}$ are sufficiently robust to our choice of categorical variables.

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 three null hypotheses in (5.1) where $\mathbf{U}$ contains work experience and four categorical variables for marital 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 0.012, 0.000, and 0.008, respectively. These results imply that we can reject $\mathbb{H}_{0,1}$ at the 5% level only. 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 in Figures 2-5. More importantly, both $\mathbb{H}_{0,2}$ and $\mathbb{H}_{0,12}$ can be rejected in favor of a one-sided alternative at the 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

and propose a test for the constancy of a subvector of the functional coefficients. Simulations indicate that our estimator and test 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.

Some extensions are possible. First, as we have discussed in Section 3.3, we follow the parametric literature and consider the choice of optimal IVs in the sense of minimizing the AVC matrix among the class of local linear GMM estimators. It seems worth considering the choice of IVs that is optimal in the sense of minimizing certain AMSE or AMISE criterion function. Second, the optimal IVs depend on two conditional expectation objects which are typically to be estimated nonparametrically. It is worthwhile to develop the asymptotic theory by allowing for nonparametrically estimated optimal IVs. We conjecture that research along this line can be done by extending either Newey's (1993) sieve and nearest-neighbor estimates or Mammen, Rothe and Schienle's (2010) kernel approach to our framework. Third, it is also interesting to estimate a restricted functional coefficient model where some functional coefficients are constant while others are not. To this goal, one can follow the idea of profile least squares or likelihood (e.g., Fan and Huang (2005), Su and Jin (2010)) and extend it to our local linear GMM framework. Fourth, it is also possible to allow the variables in the functional coefficients to be endogenous. This is associated with the well-known ill-posed inverse problem as in typical nonparametric IV regression and several regularization techniques can be called upon. Fifth, one can consider the optimal choice of data-driven bandwidth for the testing problem. We leave these for future research.

Appendix

A Proof of the Results in Section 3

Proof of Theorem 3.1. For notational simplicity, in this proof we suppress the dependence of ξ, $\mathbf{K}_{\mathbf{h}\lambda}$, $\mathbf{Q}_{\mathbf{h}}$, and $\mathbf{\Psi}_n$ on $\mathbf{u}$. Let $d_{\mathbf{U}_i^d\mathbf{u}^d} \equiv \sum_{t=1}^{p_d} 1\{U_{i,t}^d \neq u_t^d\}$, 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 $G_{i,j} \equiv G_{i,j}(\mathbf{u}) = [g_j(\mathbf{U}_i^c, \mathbf{U}_i^d) - g_j(\mathbf{u}^c, \mathbf{u}^d) - \dot{g}_j(\mathbf{u}^c, \mathbf{u}^d)'(\mathbf{U}_i^c - \mathbf{u}^c)]$, and $\mathbf{G}_i \equiv \mathbf{G}_i(\mathbf{u}) = (G_{i,1}(\mathbf{u}), \cdots, G_{i,d}(\mathbf{u}))'$. Let $R_i \equiv R_i(\mathbf{u}) = \mathbf{G}_i(\mathbf{u})'\mathbf{X}_i$. Then $Y_i = \sum_{j=1}^d [g_j(\mathbf{u}^c, \mathbf{u}^d) - \dot{g}_j(\mathbf{u}^c, \mathbf{u}^d)'(\mathbf{U}_i^c - \mathbf{u}^c)]X_{i,j} + \varepsilon_i + R_i = \xi_{i,\mathbf{u}}'\alpha(\mathbf{u}) + \varepsilon_i + R_i$, where $\alpha(\mathbf{u})$ and $\xi_{i,\mathbf{u}}$ are defined after eq. (2.6). Let $\varepsilon \equiv (\varepsilon_1, \cdots, \varepsilon_n)'$ and $\mathbf{R} \equiv \mathbf{R}(\mathbf{u}) = (R_1(\mathbf{u}), \cdots, R_n(\mathbf{u}))'$. Then we have the following bias-variance decomposition:

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

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

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

Proof. Recall that $\eta_i(\mathbf{u}^c) \equiv (\mathbf{U}_i^c - \mathbf{u}^c)/\mathbf{h}$ and $\mathbf{Q}_i \equiv \mathbf{Q}(\mathbf{V}_\mathbf{i})$. By the definition of $\mathbf{Q}_{\mathbf{h},i\mathbf{u}}$ in (2.7) we have

$$\mathbf{\Phi}_n(\mathbf{u}) = \frac{1}{n}\sum_{i=1}^n K_{\mathbf{h}\lambda,i\mathbf{u}}\begin{pmatrix} \mathbf{Q}_i \\ \mathbf{Q}_i \otimes \eta_i(\mathbf{u}^c) \end{pmatrix}\left(\mathbf{X}_i' \;\; \mathbf{X}_i' \otimes (\mathbf{U}_i^c - \mathbf{u}^c)'\right)\mathbf{H}^{-1} = \begin{pmatrix} \underset{k\times d}{\mathbf{\Phi}_{n,11}} & \underset{k\times p_c d}{\mathbf{\Phi}_{n,12}} \\ \underset{kp_c\times d}{\mathbf{\Phi}_{n,21}} & \underset{kp_c\times p_c d}{\mathbf{\Phi}_{n,22}} \end{pmatrix},$$

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

By Assumptions A1-A3,

$$\begin{aligned}
E[\mathbf{\Phi}_{n,11}] &= E[\mathbf{Q}_i\mathbf{X}_i^{'}K_{\mathbf{h}\lambda,i\mathbf{u}}] \\
&= E[\mathbf{Q}_i\mathbf{X}_i^{'}W_{\mathbf{h},i\mathbf{u}^c}|d_{\mathbf{U}_i^d\mathbf{u}^d}=0]p(\mathbf{u}^d) + \sum_{s=1}^{p_d} E[\mathbf{Q}_i\mathbf{X}_i^{'}W_{\mathbf{h},i\mathbf{u}^c}L_{\lambda,i\mathbf{u}^d}|d_{\mathbf{U}_i^d\mathbf{u}^d}=s]p(d_{\mathbf{U}_i^d\mathbf{u}^d}=s) \\
&= E[\mathbf{\Omega}_1(\mathbf{U}_i^c,\mathbf{U}_i^d)W_{\mathbf{h},i\mathbf{u}^c}|d_{\mathbf{U}_i^d\mathbf{u}^d}=0]p(\mathbf{u}^d) + O(\|\lambda\|) \\
&= \int \mathbf{\Omega}_1(\mathbf{u}^c+\mathbf{h}\odot\mathbf{t},\mathbf{u}^d) f_{\mathbf{U}}(\mathbf{u}^c+\mathbf{h}\odot\mathbf{t},\mathbf{u}^d) W(\mathbf{t})\,d\mathbf{t} + O(\|\lambda\|) \\
&= \mathbf{\Omega}_1(\mathbf{u}) f_{\mathbf{U}}(\mathbf{u}) + O\left(\|\mathbf{h}\|^2 + \|\lambda\|\right). \qquad \text{(A.2)}
\end{aligned}$$

Define two column vectors $\omega_1 \in \mathbb{R}^k$ and $\omega_2 \in \mathbb{R}^d$ such that $\|\omega_l\| = 1$ for $l = 1,2$. Then it is easy to show that $\mathrm{Var}(\omega_1'\mathbf{\Phi}_{n,11}\omega_2) = \frac{1}{n}\mathrm{Var}(\omega_1'\mathbf{Q}_i\mathbf{X}_i'\omega_2 K_{\mathbf{h}\lambda,i\mathbf{u}}) = O((n\mathbf{h}!)^{-1}) = o(1)$. It follows by Chebyshev's inequality that $\mathbf{\Phi}_{n,11} = \mathbf{\Omega}_1(\mathbf{u}) f_{\mathbf{U}}(\mathbf{u}) + o_P(1)$. Similarly,

$$\begin{aligned}
\mathbf{\Phi}_{n,22} &= E[\mathbf{\Phi}_{n,22}] + O_P\left((n\mathbf{h}!)^{-1/2}\right) = E\left[(\mathbf{Q}_i\mathbf{X}_i') \otimes (\eta_i(\mathbf{u}^c)\eta_i(\mathbf{u}^c)')K_{\mathbf{h}\lambda,i\mathbf{u}}\right] + O_P\left((n\mathbf{h}!)^{-1/2}\right) \\
&= \int \left[\mathbf{\Omega}_1(\mathbf{u}^c+\mathbf{h}\odot\mathbf{t},\mathbf{u}^d) \otimes \mathbf{t}\mathbf{t}'\right] f_{\mathbf{U}}(\mathbf{u}^c+\mathbf{h}\odot\mathbf{t},\mathbf{u}^d) W(\mathbf{t})\,d\mathbf{t} + O_P\left(\|\lambda\| + (n\mathbf{h}!)^{-1/2}\right) \\
&= \mu_{2,1}\left[\mathbf{\Omega}_1(\mathbf{u}) \otimes \mathbf{I}_{p_c}\right] f_{\mathbf{U}}(\mathbf{u}) + o_P(1).
\end{aligned}$$

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

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

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

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

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 $A_{i,j}(\mathbf{u}) \equiv (\mathbf{U}_i^c-\mathbf{u}^c)'\ddot{g}_j(\mathbf{u})(\mathbf{U}_i^c-\mathbf{u}^c)$ and $\ddot{g}_j(\mathbf{u}) \equiv \partial\dot{g}_j(\mathbf{u})/\partial\mathbf{u}^{c\prime}$. Let $\mathbf{A}_i(\mathbf{u}) \equiv (A_{i,1}(\mathbf{u}), \cdots, A_{i,d}(\mathbf{u}))'$. Then we have

$$\begin{aligned}
\mathbf{b}_{n,1} &= \frac{1}{2}\sqrt{n\mathbf{h}!}\,E_0\left[\begin{pmatrix}\mathbf{Q}_i\mathbf{X}_i'\mathbf{A}_i(\mathbf{u}) \\ (\mathbf{Q}_i\mathbf{X}_i'\mathbf{A}_i(\mathbf{u}))\otimes\eta_i(\mathbf{u}^c)\end{pmatrix}W_{\mathbf{h},i\mathbf{u}^c}\right]p(\mathbf{u}^d) + o(\sqrt{n\mathbf{h}!}\,\|\mathbf{h}\|^2) \\
&= \frac{1}{2}\sqrt{n\mathbf{h}!}\,E_0\left[\begin{pmatrix}\mathbf{\Omega}_1(\mathbf{U}_i)\mathbf{A}_i(\mathbf{u}) \\ (\mathbf{\Omega}_1(\mathbf{U}_i)\mathbf{A}_i(\mathbf{u}))\otimes\eta_i(\mathbf{u}^c)\end{pmatrix}W_{\mathbf{h},i\mathbf{u}^c}\right]p(\mathbf{u}^d) + o(1) \\
&= \frac{\sqrt{n\mathbf{h}!}\mu_{2,1}}{2}\begin{pmatrix}f_{\mathbf{U}}(\mathbf{u})\mathbf{\Omega}_1(\mathbf{u})\mathbf{A}(\mathbf{u};\mathbf{h}) \\ \mathbf{0}_{kp_c\times 1}\end{pmatrix} + o(1), \text{ and}
\end{aligned}$$

$$
\begin{aligned}
\mathbf{b}_{n,2} &= \sqrt{n\mathbf{h}!}E_1\left\{\sum_{j=1}^{d}[g_j(\mathbf{U}_i)-g_j(\mathbf{u})-\dot{g}_j(\mathbf{u})'(\mathbf{U}_i^c-\mathbf{u}^c)]\begin{pmatrix}\mathbf{Q}_iX_{i,j}\\ \mathbf{Q}_iX_{i,j}\otimes\eta_i(\mathbf{u}^c)\end{pmatrix}K_{\mathbf{h}\lambda,i\mathbf{u}}\right\}p_1\\
&= \sqrt{n\mathbf{h}!}E_1\left[\sum_{j=1}^{d}\begin{pmatrix}\mathbf{Q}_i\mathbf{X}_i'\mathbf{G}_i\\ (\mathbf{Q}_i\mathbf{X}_i'\mathbf{G}_i)\otimes\eta_i(\mathbf{u}^c)\end{pmatrix}K_{\mathbf{h}\lambda,i\mathbf{u}}\right]p_1\\
&= \sqrt{n\mathbf{h}!}E_1\left[\begin{pmatrix}\mathbf{\Omega}_1(\mathbf{U}_i)[\mathbf{g}(\mathbf{U}_i)-\mathbf{g}(\mathbf{u})]-(\mathbf{\Omega}_1(\mathbf{U}_i)\otimes\eta_i(\mathbf{u}^c)')\dot{\mathbf{g}}(\mathbf{u})\\ (\mathbf{\Omega}_1(\mathbf{U}_i)[\mathbf{g}(\mathbf{U}_i)-\mathbf{g}(\mathbf{u})])\otimes\eta_i(\mathbf{u}^c)-(\mathbf{\Omega}_1(\mathbf{U}_i)\otimes[\eta_i(\mathbf{u}^c)\eta_i(\mathbf{u}^c)'])\dot{\mathbf{g}}(\mathbf{u})\end{pmatrix}K_{\mathbf{h}\lambda,i\mathbf{u}}\right]p_1\\
&\quad +o(1)\\
&= \sqrt{n\mathbf{h}!}\sum_{\widetilde{\mathbf{u}}^d\in\mathcal{U}^d}\sum_{s=1}^{p_d}\lambda_sI_s(\mathbf{u}^d,\widetilde{\mathbf{u}}^d)f_{\mathbf{U}}(\mathbf{u}^c,\widetilde{\mathbf{u}}^d)\begin{pmatrix}\mathbf{\Omega}_1(\mathbf{u}^c,\widetilde{\mathbf{u}}^d)(\mathbf{g}(\mathbf{u}^c,\widetilde{\mathbf{u}}^d)-\mathbf{g}(\mathbf{u}^c,\mathbf{u}^d))\\ -\mu_{2,1}(\mathbf{\Omega}_1(\mathbf{u}^c,\widetilde{\mathbf{u}}^d)\otimes\mathbf{I}_{p_c})\dot{\mathbf{g}}(\mathbf{u}^c,\mathbf{u}^d)\end{pmatrix}+o(1),
\end{aligned}
$$

where $\mathbf{A}(\mathbf{u};\mathbf{h})$ and $\dot{\mathbf{g}}(\mathbf{u})$ are defined in Section 3.2, $E_l\{\cdot\}=E\{\cdot|d_{\mathbf{U}_i^d\mathbf{u}^d}=l\}$ for $l=0$ and 1, and $p_1=P(d_{\mathbf{U}_i^d\mathbf{u}^d}=1)$. Consequently, $\sqrt{n\mathbf{h}!}E[\mathcal{B}_n(\mathbf{u})]=\sqrt{n\mathbf{h}!}\mathbf{B}(\mathbf{u};\mathbf{h},\lambda)+o(1)$. Noting that $\mathrm{Var}(\sqrt{n\mathbf{h}!}\mathcal{B}_n(\mathbf{u}))=O(\|\mathbf{h}\|^2+\|\lambda\|)=o(1)$, the conclusion then follows by Chebyshev's inequality. ■

Lemma A.3 $\sqrt{n\mathbf{h}!}\mathcal{V}_n(\mathbf{u}) = n^{-1/2}(\mathbf{h}!)^{1/2}\sum_{i=1}^{n}\begin{pmatrix}\mathbf{Q}_i\varepsilon_i\\ (\mathbf{Q}_i\varepsilon_i)\otimes((\mathbf{U}_i^c-\mathbf{u}^c)/\mathbf{h})\end{pmatrix}K_{\mathbf{h}\lambda,i\mathbf{u}} \xrightarrow{d} N(0,\mathbf{\Upsilon}(\mathbf{u}))$,

where $\mathbf{\Upsilon}(\mathbf{u})$ *is defined in (3.2).*

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

$$
\begin{aligned}
&\mathrm{Var}\left(\sqrt{n\mathbf{h}!}\mathbf{c}'\mathcal{V}_n(\mathbf{u})\right)=\mathrm{Var}(\zeta_1)\\
&= \mathbf{h}!\mathbf{c}'E\left\{\begin{pmatrix}\mathbf{Q}_i\mathbf{Q}_i'\varepsilon_i^2 & (\mathbf{Q}_i\mathbf{Q}_i')\otimes\eta_i(\mathbf{u}^c)'\varepsilon_i^2\\ (\mathbf{Q}_i\mathbf{Q}_i')\otimes\eta_i(\mathbf{u}^c)\varepsilon_i^2 & (\mathbf{Q}_i\mathbf{Q}_i')\otimes[\eta_i(\mathbf{u}^c)\eta_i(\mathbf{u}^c)']\varepsilon_i^2\end{pmatrix}K_{\mathbf{h}\lambda,i\mathbf{u}}^2\right\}\mathbf{c}\\
&= \mathbf{h}!\mathbf{c}'E\left\{\begin{pmatrix}\mathbf{Q}_i\mathbf{Q}_i'\sigma^2(\mathbf{V}_i) & (\mathbf{Q}_i\mathbf{Q}_i')\otimes\eta_i(\mathbf{u}^c)'\sigma^2(\mathbf{V}_i)\\ (\mathbf{Q}_i\mathbf{Q}_i')\otimes\eta_i(\mathbf{u}^c)\sigma^2(\mathbf{V}_i) & (\mathbf{Q}_i\mathbf{Q}_i')\otimes[\eta_i(\mathbf{u}^c)\eta_i(\mathbf{u}^c)']\sigma^2(\mathbf{V}_i)\end{pmatrix}K_{\mathbf{h}\lambda,i\mathbf{u}}^2\right\}\mathbf{c}\\
&= \mathbf{h}!\mathbf{c}'E\left\{\begin{pmatrix}\mathbf{\Omega}_2(\mathbf{U}_i) & \mathbf{\Omega}_2(\mathbf{U}_i)\otimes\eta_i(\mathbf{u}^c)'\\ \mathbf{\Omega}_2(\mathbf{U}_i)\otimes\eta_i(\mathbf{u}^c) & \mathbf{\Omega}_2(\mathbf{U}_i)\otimes[\eta_i(\mathbf{u}^c)\eta_i(\mathbf{u}^c)']\end{pmatrix}K_{\mathbf{h}\lambda,i\mathbf{u}}^2\right\}\mathbf{c}=\mathbf{c}'\mathbf{\Upsilon}\mathbf{c}+o(1).
\end{aligned}
$$

The result follows as it is standard to check the Liapounov condition; see, e.g., Li and Racine (2007). ■

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

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

where dependence of $\mathbf{\Phi}$, $\mathbf{\Psi}$ and $\mathbf{\Upsilon}$ on $\mathbf{u}$ is suppressed. This completes the proof of Theorem 3.1.

Proof of Theorems 3.3. Let $\mathbf{\Phi}_n(\mathbf{u};\mathbf{h},\lambda)$, $\mathcal{B}_n(\mathbf{u};\mathbf{h},\lambda)$ and $\mathcal{V}_n(\mathbf{u};\mathbf{h},\lambda)$ be as defined after (A.1). Let $\bar{\mathcal{B}}_n(\mathbf{u};\mathbf{h},\lambda)\equiv\mathcal{B}_n(\mathbf{u};\mathbf{h},\lambda)-\mathbf{B}(\mathbf{u};\mathbf{h},\lambda)$. Let $J_{1n}\equiv\mathbf{\Phi}_n(\mathbf{u};\widehat{\mathbf{h}},\widehat{\lambda})-\mathbf{\Phi}_n(\mathbf{u};\mathbf{h},\lambda)$, $J_{2n}\equiv\sqrt{n\mathbf{h}!}[\mathcal{V}_n(\mathbf{u};\widehat{\mathbf{h}},\widehat{\lambda})-$

$\mathcal{V}_n(\mathbf{u};\mathbf{h},\lambda)]$, and $J_{3n} \equiv \sqrt{n\mathbf{h}!}[\bar{\mathcal{B}}_n(\mathbf{u};\widehat{\mathbf{h}},\widehat{\lambda}) - \bar{\mathcal{B}}_n(\mathbf{u};\mathbf{h},\lambda)]$. By the result in Theorem 3.1 and the expansion in (A.1), it suffices to show that (i) $J_{1n} = o_P(1)$, (ii) $J_{2n} = o_P(1)$, and (iii) $J_{3n} = o_P(1)$.

For notational simplicity, for the moment we assume that $p_c = p_d = 1$ so that we can write the bandwidth $(\mathbf{h},\lambda)$ simply as (h,λ). Similarly, we write $(\mathbf{U}_i^c, \mathbf{u}^c)$ and $\left(\mathbf{U}_i^d, \mathbf{u}^d\right)$ as (U_i^c, u^c) and $\left(U_i^d, u^d\right)$, respectively. Let $h = bn^{-\delta}$ and $\lambda = rn^{-\sigma}$ for some $b \in \left[\underline{b}, \bar{b}\right]$, $r \in [\underline{r}, \bar{r}]$, $\delta > 0$, and $\sigma > 0$. Note that when $p_c = p_d = 1$, we can write $\mathbf{h}!K_{\mathbf{h}\lambda,i\mathbf{u}}$ as

$$hK_{h\lambda,i\mathbf{u}} = w\left(\frac{U_i^c - u^c}{h}\right)\lambda^{1\left\{U_i^d \neq u^d\right\}} = w\left(\frac{U_i^c - u^c}{bn^{-\delta}}\right)\left(rn^{-\sigma}\right)^{1\left\{U_i^d \neq u^d\right\}} \equiv \overline{K}_{br,i\mathbf{u}}.$$

For any nonnegative random variable ς_i, define $m_\zeta(\mathbf{u}) = E(\varsigma_i|\mathbf{U}_i = \mathbf{u})$. m_ζ is usually continuous and uniformly bounded below. Then by the C_r inequality, for any $\gamma > 0$,

$$\begin{aligned} E\left[\left|\overline{K}_{b'r',i\mathbf{u}} - \overline{K}_{br,i\mathbf{u}}\right|_i^\gamma \varsigma_i\right] &= E\left[\left|h'K_{h'\lambda',i\mathbf{u}} - hK_{h\lambda,i\mathbf{u}}\right|^\gamma m_\varsigma(\mathbf{U}_i)\right] \\ &\leq c_\gamma\left\{E\left[\left|h'K_{h'\lambda',i\mathbf{u}} - hK_{h\lambda,i\mathbf{u}}\right|^\gamma m_\zeta(U_i)\right] + E\left[h\left|K_{h\lambda',i\mathbf{u}} - K_{h\lambda,i\mathbf{u}}\right|^\gamma m_\zeta(\mathbf{U}_i)\right]\right\} \\ &\equiv c_\gamma\left\{K_1 + K_2\right\}, \text{ say,} \end{aligned}$$

where $c_\gamma = 1$ if $\gamma \in (0,1]$ and $c_\gamma = 2^{\gamma-1}$ if $\gamma > 1$. Here and in the remainder of this proof prime does not denote transpose. Let $c_b = \bar{b}/\underline{b}$. By the fact that $\lambda' \in (0,1]$ and Assumption A5, for any b, $b' \in \left[\underline{b}, \bar{b}\right]$,

$$\begin{aligned} K_1 &= \sum_{u_i^d \in \mathcal{U}^d}\int\left|\left[w\left(\frac{u_i^c - u^c}{h'}\right) - w\left(\frac{u_i^c - u^c}{h}\right)\right](\lambda')^{1\left\{u_i^d \neq u^d\right\}}\right|^\gamma m_\varsigma\left(u_i^c, u_i^d\right) f\left(u_i^c, u_i^d\right) du_i^c \\ &\leq h\sum_{u_i^d \in \mathcal{U}^d}\int_{-c_w c_b}^{c_w c_b}\left|w(vh/h') - w(v)\right|^\gamma m_\varsigma\left(u^c + hv, u_i^d\right) f\left(u^c + hv, u_i^d\right) dv \\ &\leq C_{1\varsigma}C_w^\gamma h\left|1 - h/h'\right|^\gamma\int_{-c_w c_b}^{c_w c_b}|v|^\gamma dv = C_{1\varsigma}C_w^\gamma h\left|(b'-b)/b'\right|^\gamma\int_{-c_w c_b}^{c_w c_b}|v|^\gamma dv \leq C_{2\varsigma}h\left|b'-b\right|^\gamma, \end{aligned}$$

where $C_{s\varsigma}$ is a finite constant that depends on ς_i; e.g., $C_{1\varsigma} \equiv \sup_{u^c \in \mathcal{U}^c}\sum_{u_i^d \in \mathcal{U}^d} m_\varsigma\left(u^c + hv, u_i^d\right) f\left(u^c + hv, u_i^d\right) dv$ $< \infty$. Similarly,

$$\begin{aligned} K_2 &= \sum_{u_i^d \in \mathcal{U}^d, u_i^d \neq u^d}\int\left|w\left(\frac{u_i^c - u^c}{h}\right)\right|^\gamma\left|\lambda' - \lambda\right|^\gamma m_\varsigma\left(u_i^c, u_i^d\right) f\left(u_i^c, u_i^d\right) du_i^c \\ &= h\left|\lambda' - \lambda\right|^\gamma\sum_{u_i^d \in \mathcal{U}^d}\int_{-c_w c_b}^{c_w c_b} w(v)^\gamma m_\varsigma\left(u^c + hv, u_i^d\right) f\left(u^c + hv, u_i^d\right) dv \\ &\leq C_{3\varsigma}h\left|\lambda' - \lambda\right|^\gamma \leq C_{3\varsigma}hn^{-\gamma\sigma}\left|r' - r\right|^\gamma. \end{aligned}$$

It follows that

$$E\left[\left|\overline{K}_{b'r',i\mathbf{u}} - \overline{K}_{br,i\mathbf{u}}\right|^\gamma \varsigma_i\right] \leq c_\gamma\left(C_{2\varsigma} \vee C_{3\varsigma}\right)h\left(\left|b'-b\right|^\gamma + \left|r'-r\right|^\gamma\right), \tag{A.3}$$

where $a \vee b = \max(a,b)$. Then by the C_r inequality

$$\begin{aligned} E\left[\left|\overline{K}_{b'r',i\mathbf{u}}\right|^\gamma \varsigma_i\right] &\leq c_\gamma E\left[\left|h'K_{h'\lambda',i\mathbf{u}} - hK_{h\lambda,i\mathbf{u}}\right|^\gamma \varsigma_i\right] + c_\gamma E\left[\left|hK_{h\lambda,i\mathbf{u}}\right|^\gamma \varsigma_i\right] \\ &\leq c_\gamma\left(C_{2\varsigma} \vee C_{3\varsigma}\right)h\left(\left|b'-b\right|^\gamma + \left|\lambda'-\lambda\right|^\gamma\right) + C_{4\varsigma}h \leq C_{5\varsigma}h. \end{aligned} \tag{A.4}$$

Note that

$$\left|(h')^\alpha - h^\alpha\right| = \left|(b')^\alpha - b^\alpha\right|n^{-\alpha\delta} = \alpha n^{-\alpha\delta}(b^*)^{\alpha-1}\left|b'-b\right| \leq C_6 h^\alpha\left|b'-b\right| \tag{A.5}$$

where the second equality follows from the intermediate value theory and b^* lies between b' and b. Then by the C_r inequality, and (A.3)-(A.5), we have that for any $\alpha > 0$ and $\gamma > 0$,

$$\begin{aligned}
& E\left[\left|(h')^{-\alpha}\overline{K}_{b'r',i\mathbf{u}} - h^{\alpha}\overline{K}_{br,i\mathbf{u}}\right|^{\gamma}\varsigma_i\right] \\
\leq\ & c_\gamma E\left[\left|h^{-\alpha}\left(\overline{K}_{b'r',i\mathbf{u}} - \overline{K}_{br,i\mathbf{u}}\right)\right|^{\gamma}\varsigma_i\right] + c_\gamma E\left[\left|\left[(h')^{-\alpha} - h^{-\alpha}\right]\overline{K}_{b'r',i\mathbf{u}}\right|^{\gamma}\varsigma_i\right] \\
\leq\ & c_\gamma h^{-\alpha\gamma} E\left[\left|\overline{K}_{b'r',i\mathbf{u}} - \overline{K}_{br,i\mathbf{u}}\right|^{\gamma}\varsigma_i\right] + c_\gamma\left|(h')^{-\alpha} - h^{-\alpha}\right|^{\gamma} E\left[\left|\overline{K}_{b'r',i\mathbf{u}}\right|^{\gamma}\varsigma_i\right] \\
\leq\ & c_\gamma^2 h^{1-\alpha\gamma}\left(|b'-b|^{\gamma} + |\lambda'-\lambda|^{\gamma}\right) + c_\gamma C_6 h^{-\alpha\gamma}|b'-b| C_{5\varsigma} h \leq C_{7\varsigma} h^{1-\alpha\gamma}\left(|b'-b|^{\gamma} + |\lambda'-\lambda|^{\gamma}\right).
\end{aligned}$$

In the general case where $p_c > 1$ or $p_d > 1$, with a little bit abuse of notation we use row vectors to denote the bandwidth parameters. We can write $h_s = b_s n^{-\delta_s}$ and $\lambda_t = r_t n^{-\sigma_t}$ for some b_s, r_t, δ_s, $\sigma_t > 0$, $s = 1, ..., p_c$, and $t = 1, ..., p_d$. Let $\mathbf{b} = (b_1, ..., b_{p_c})$ and $\mathbf{r} = (r_1, ..., r_{p_d})$. Similarly, $(\mathbf{h}', \lambda') = (h'_1, ..., h'_{p_c}, \lambda'_1, ..., \lambda'_{p_d})$ and $(\mathbf{b}', \mathbf{r}') = (b'_1, ..., b'_{p_c}, r'_1, ..., r'_{p_d})$ are connected through $h'_s = b'_s n^{-\delta_s}$ and $\lambda'_t = r'_t n^{-\sigma_t}$ for b'_s, $r'_t > 0$. Then using the fact that our multivariate kernel function is a product of univariate kernel functions, we can follow the above arguments and readily show that

$$E\left[\left|(\mathbf{h}'!)^{-\alpha}\overline{K}_{\mathbf{b'r'},i\mathbf{u}} - (\mathbf{h}!)^{\alpha}\overline{K}_{\mathbf{br},i\mathbf{u}}\right|^{\gamma}\varsigma_i\right] \leq C_{8\varsigma} h^{1-\alpha\gamma}\left(\|\mathbf{b}'-\mathbf{b}\|^{\gamma} + \|\lambda'-\lambda\|^{\gamma}\right) \tag{A.6}$$

where $C_{8\varsigma}$ is a finite constant depending on ς_i. Below we use C to denote a generic constant that can vary from equation to equation.

Now we prove (i). Let $J_{1n,st}(\mathbf{b}, \mathbf{r}) = \mathbf{\Phi}_{n,st}(\mathbf{u}; \mathbf{h}, \lambda)$ for s, $t = 1, 2$, where $\mathbf{\Phi}_{n,st}$'s are defined in the proof of Lemma A.1. By Theorem 3.1 in Li and Li (2010), it suffices to show that for any $(\mathbf{b}', \mathbf{r}')$ and $(\mathbf{b}, \mathbf{r})$ that lie in a compact set (e.g., $[\underline{b}, \bar{b}] \times [\underline{r}, \bar{r}]$ for the case $p_c = p_d = 1$), there exist $\alpha > 0$ and $\gamma > 1$ such that

$$E\left\{\|J_{1n,st}(\mathbf{b}', \mathbf{r}') - J_{1n,st}(\mathbf{b}, \mathbf{r})\|^{\alpha}\right\} \leq C\left\{\|\mathbf{b}'-\mathbf{b}\|^{\gamma} + \|\lambda'-\lambda\|^{\gamma}\right\} \text{ for some } C < \infty. \tag{A.7}$$

We only show (A.7) for the case $(s, t) = (1, 1)$ as the other cases are similar. Let $\varsigma_{i,jl}$ denote the (j, l)th element of $\mathbf{Q}_i \mathbf{X}'_i$ for $j = 1, ..., k$ and $l = 1, ..., d$. Let $J^{(j,l)}_{1n,st}(b, r)$ denote the (j, l)th element of $J_{1n,st}(b, r)$. Then $E|J^{(j,l)}_{1n,11}(\mathbf{b}', \mathbf{r}') - J^{(j,l)}_{1n,11}(\mathbf{b}, \mathbf{r})|^2$ is bounded above by

$$\begin{aligned}
& 2E\left|\frac{1}{n}\sum_{i=1}^{n}\left\{\left[(\mathbf{h}'!)^{-1}\overline{K}_{\mathbf{b'r'},i\mathbf{u}} - (\mathbf{h}!)^{-1}\overline{K}_{\mathbf{br},i\mathbf{u}}\right]\varsigma_{i,jl} - E\left\{\left[(\mathbf{h}'!)^{-1}\overline{K}_{\mathbf{b'r'},i\mathbf{u}} - (\mathbf{h}!)^{-1}\overline{K}_{\mathbf{br},i\mathbf{u}}\right]\varsigma_{i,jl}\right\}\right\}\right|^2 \\
& +2\left|E\left\{\left[(\mathbf{h}'!)^{-1}\overline{K}_{\mathbf{b'r'},i\mathbf{u}} - (\mathbf{h}!)^{-1}\overline{K}_{\mathbf{br},i\mathbf{u}}\right]\varsigma_{i,jl}\right\}\right|^2 \equiv 2S_1 + 2S_2, \text{ say.}
\end{aligned}$$

By Assumption A1, Jensen's inequality, and (A.6), for sufficiently large n

$$\begin{aligned}
S_1 &= \frac{1}{n^2}\sum_{i=1}^{n}\mathrm{Var}\left(\left[(\mathbf{h}'!)^{-1}\overline{K}_{\mathbf{b'r'},i\mathbf{u}} - (\mathbf{h}!)^{-1}\overline{K}_{\mathbf{br},i\mathbf{u}}\right]\varsigma_{i,jl}\right) \\
&\leq n^{-1}E\left\{\left[(\mathbf{h}'!)^{-1}\overline{K}_{\mathbf{b'r'},i\mathbf{u}} - (\mathbf{h}!)^{-1}\overline{K}_{\mathbf{br},i\mathbf{u}}\right]\varsigma_{i,jl}\right\}^2 \\
&\leq n^{-1}Ch^{1-2}\left(\|\mathbf{b}'-\mathbf{b}\|^2 + \|\lambda'-\lambda\|^2\right) \leq C\left(\|\mathbf{b}'-\mathbf{b}\|^2 + \|\lambda'-\lambda\|^2\right)
\end{aligned}$$

as $nh \to \infty$ implies that for sufficiently large n, $n^{-1}h^{-1}$ can be bounded by the constant 1. By (A.6),

$$\begin{aligned}
S_2 &= \left|E\left\{\left[(\mathbf{h}'!)^{-1}\overline{K}_{\mathbf{b'r'},i\mathbf{u}} - (\mathbf{h}!)^{-1}\overline{K}_{\mathbf{br},i\mathbf{u}}\right]\varsigma_{i,jl}\right\}\right|^2 \\
&\leq \left[C\left(\|\mathbf{b}'-\mathbf{b}\| + \|\lambda'-\lambda\|\right)\right]^2 \leq C\left(\|\mathbf{b}'-\mathbf{b}\|^2 + \|\lambda'-\lambda\|^2\right).
\end{aligned}$$

It follows that $E\{|J^{(j,l)}_{1n,11}(\mathbf{b}', \mathbf{r}') - J^{(j,l)}_{1n,11}(\mathbf{b}, \mathbf{r})|^2\} \leq C(\|\mathbf{b}'-\mathbf{b}\|^2 + \|\lambda'-\lambda\|^2)$ and (A.7) holds for $\alpha = \gamma = 2$ and $(s, t) = (1, 1)$. Analogously, we can show that it also holds for $\alpha = \gamma = 2$ and other values of s and t. This completes the proof of (i).

Next we prove (ii). Let $J_{2n}(\mathbf{b},\mathbf{r}) \equiv \sqrt{n\mathbf{h}!}\mathcal{V}_n(\mathbf{u};\mathbf{h},\lambda)$. By Theorem 3.1 in Li and Li (2010), it suffices to show that for any $(\mathbf{b}',\mathbf{r}')$ and $(\mathbf{b},\mathbf{r})$ that lie in a compact set, there exist $\alpha > 0$ and $\gamma > 1$ such that

$$E\left\{\|J_{2n}(\mathbf{b}',\mathbf{r}') - J_{2n}(\mathbf{b},\mathbf{r})\|^{\alpha}\right\} \leq C\left\{\|\mathbf{b}'-\mathbf{b}\|^{\gamma} + \|\lambda'-\lambda\|^{\gamma}\right\} \text{ for some } C < \infty. \tag{A.8}$$

Let $\mathbf{e}_i \equiv (\mathbf{Q}_i', \mathbf{Q}_i' \otimes ((\mathbf{U}_i^c - \mathbf{u}^c)'/\mathbf{h})'$. By (A.6)

$$\begin{aligned} E\left\{\|J_{2n}(\mathbf{b}',\mathbf{r}') - J_{2n}(\mathbf{b},\mathbf{r})\|^2\right\} &= E\left\{\left\|n^{-1/2}\sum_{i=1}^n\left[(\mathbf{h}'!)^{-1/2}\overline{K}_{\mathbf{b}'\mathbf{r}',i\mathbf{u}} - (\mathbf{h}!)^{-1/2}\overline{K}_{\mathbf{br},i\mathbf{u}}\right]\mathbf{e}_i\varepsilon_i\right\|^2\right\} \\ &= E\left\{\left|(\mathbf{h}'!)^{-1/2}\overline{K}_{\mathbf{b}'\mathbf{r}',i\mathbf{u}} - (\mathbf{h}!)^{-1/2}\overline{K}_{\mathbf{br},i\mathbf{u}}\right|^2\mathbf{e}_i'\mathbf{e}_i\varepsilon_i^2\right\} \\ &\leq C\left(\|\mathbf{b}'-\mathbf{b}\|^2 + \|\lambda'-\lambda\|^2\right). \end{aligned}$$

Thus (ii) follows.

Next, we prove (iii). Decompose $\bar{\mathcal{B}}_n(\mathbf{u};\mathbf{h},\lambda) = \bar{\mathcal{B}}_{1n}(\mathbf{u};\mathbf{h},\lambda) + \bar{\mathcal{B}}_{2n}(\mathbf{u};\mathbf{h},\lambda)$, where $\bar{\mathcal{B}}_{1n}(\mathbf{u};\mathbf{h},\lambda) \equiv \frac{1}{n}\sum_{i=1}^n\{K_{\mathbf{h}\lambda,i\mathbf{u}}\mathbf{Q}_{\mathbf{h},i\mathbf{u}}R_i(\mathbf{u}) - E[K_{\mathbf{h}\lambda,i\mathbf{u}}\mathbf{Q}_{\mathbf{h},i\mathbf{u}}R_i(\mathbf{u})]\}$, and $\bar{\mathcal{B}}_{2n}(\mathbf{u};\mathbf{h},\lambda) \equiv E[K_{\mathbf{h}\lambda,i\mathbf{u}}\mathbf{Q}_{\mathbf{h},i\mathbf{u}}R_i(\mathbf{u})] - \mathbf{B}(\mathbf{u};\mathbf{h},\lambda)$. Then $J_{3n} = J_{3n,1} + J_{3n,2}$, where $J_{3n,1} \equiv \sqrt{n\mathbf{h}!}[\bar{\mathcal{B}}_{1n}(\mathbf{u};\widehat{\mathbf{h}},\widehat{\lambda}) - \bar{\mathcal{B}}_{1n}(\mathbf{u};\mathbf{h},\lambda)]$ and $J_{3n,2} \equiv \sqrt{n\mathbf{h}!}[\bar{\mathcal{B}}_{2n}(\mathbf{u};\widehat{\mathbf{h}},\widehat{\lambda}) - \bar{\mathcal{B}}_{2n}(\mathbf{u};\mathbf{h},\lambda)]$. It suffices to prove (iii) by showing that $(iii1)$ $J_{3n,1} = o_P(1)$ and $(iii2)$ $J_{3n,2} = o(1)$. By Taylor expansion and Assumptions A3 and A6, $\sqrt{n\mathbf{h}!}\bar{\mathcal{B}}_{2n}(\mathbf{u};\mathbf{h},\lambda) = \sqrt{n\mathbf{h}!}o\left(|\mathbf{h}|^2 + |\lambda|\right) = o(1)$ uniformly in $(\mathbf{h},\lambda)$, which implies $(iii2)$ by Corollary 3.1 in Li and Li (2010). The proof of $(iii1)$ is analogous to that of (ii) and thus omitted. ■

B Proof of the Results in Section 4

Proof of Theorem 4.1. Observe that $\sqrt{n}(\widehat{\theta}_1 - \theta_1) = n^{-1/2}\sum_{i=1}^n \mathbf{\Gamma}_{n1}(\mathbf{U}_i)\mathcal{B}_n(\mathbf{U}_i) + n^{-1/2}\sum_{i=1}^n \mathbf{\Gamma}_{n1}(\mathbf{U}_i)\mathcal{V}_n(\mathbf{U}_i)$. Noting that $\sup_{\mathbf{u}}\|\mathbf{\Phi}_n(\mathbf{u}) - \mathbf{\Phi}(\mathbf{u})\| = O_P(\|\mathbf{h}\|^2 + \|\lambda\| + (n\mathbf{h}!/\log n)^{-1/2})$ by strengthening the result in Lemma A.1, following the same lines of proof as in Masry (1996) we can readily show that

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

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

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

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

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

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

$$\begin{aligned} E[\overline{\varphi}(\zeta_i)\overline{\varphi}(\zeta_i)'] &= E\left[\int\int \overline{\mathbf{\Gamma}}_1(\mathbf{U})\begin{pmatrix} \mathbf{Q}_j\mathbf{Q}_j'\varepsilon_j^2 & (\mathbf{Q}_j\mathbf{Q}_j'\varepsilon_j^2)\otimes\eta_j(\widetilde{\mathbf{U}}^c)' \\ (\mathbf{Q}_j\mathbf{Q}_j'\varepsilon_j^2)\otimes\eta_j(\mathbf{U}^c) & (\mathbf{Q}_j\mathbf{Q}_j'\varepsilon_j^2)\otimes(\eta_j(\mathbf{U}^c)\eta_j(\widetilde{\mathbf{U}}^c)' \end{pmatrix}\right. \\ &\quad \left.\times\,\overline{\mathbf{\Gamma}}_1(\widetilde{\mathbf{U}})'K_{\mathbf{h}\lambda,j\mathbf{U}}K_{\mathbf{h}\lambda,j\widetilde{\mathbf{U}}}dF_\zeta(\zeta)\,dF_\zeta(\widetilde{\zeta})\right] \\ &= \mathbf{\Sigma}_{\theta_1} + o(1), \end{aligned}$$

where $\mathbf{\Sigma}_{\theta_1}$ is defined in (4.10). One can verify the Liapounov condition and conclude $A_n \xrightarrow{d} N\left(0, \mathbf{\Sigma}_{\theta_1}\right)$. 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

$$T_n = (\mathbf{h}!)^{1/2} \sum_{i=1}^{n} \left[\widehat{\mathbf{g}}_1(\mathbf{U}_i) - \overline{\mathbf{g}}_1 + \overline{\mathbf{g}}_1 - \overline{\widehat{\mathbf{g}}}_1\right]' \left[\widehat{\mathbf{g}}_1(\mathbf{U}_i) - \overline{\mathbf{g}}_1 + \overline{\mathbf{g}}_1 - \overline{\widehat{\mathbf{g}}}_1\right] = T_{n1} - T_{n2}, \tag{B.3}$$

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:

$$\begin{aligned} T_{n1} &= (\mathbf{h}!)^{1/2} \sum_{i=1}^{n} [\widehat{\mathbf{g}}_1(\mathbf{U}_i) - \mathbf{g}_1(\mathbf{U}_i)]' [\widehat{\mathbf{g}}_1(\mathbf{U}_i) - \mathbf{g}_1(\mathbf{U}_i)] + (\mathbf{h}!)^{1/2} \sum_{i=1}^{n} [\mathbf{g}_1(\mathbf{U}_i) - \overline{\mathbf{g}}]' [\mathbf{g}_1(\mathbf{U}_i) - \overline{\mathbf{g}}] \\ &\quad + 2(\mathbf{h}!)^{1/2} \sum_{i=1}^{n} [\widehat{\mathbf{g}}_1(\mathbf{U}_i) - \mathbf{g}_1(\mathbf{U}_i)]' [\mathbf{g}_1(\mathbf{U}_i) - \overline{\mathbf{g}}_1] \\ &\equiv T_{n11} + T_{n12} + 2T_{n13}, \text{ say.} \end{aligned} \tag{B.4}$$

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

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

First, $T_{n11}^{(1)} = O_P((n(\mathbf{h}!)^{1/2}(\|\mathbf{h}\|^2 + \|\lambda\|)^2) = o_P(1)$ by (B.1), (B.2), and Assumption A8. Applying (B.1), (B.2), 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 $\overline{\mathcal{B}}(\mathbf{u}) = E[\mathcal{B}_n(\mathbf{u})] = O(\|\mathbf{h}\|^2 + \|\lambda\|)$ and $\mathcal{B}_n(\mathbf{u}) - E[\mathcal{B}_n(\mathbf{u})] = O_P((n\mathbf{h}!/\log n)^{-1/2}(\|\mathbf{h}\|^2 + \|\lambda\|))$ uniformly in $\mathbf{u}$, we can readily show that $T_{n11}^{(2)} = o_P(1)$. It follows that $T_{n11} = \overline{T}_{n11} + o_P(1)$. Now, write

$$\begin{aligned} \overline{T}_{n11} &= (\mathbf{h}!)^{1/2} \sum_{i=1}^{n} \mathcal{V}_n(\mathbf{U}_i)' \overline{\mathbf{\Gamma}}_1(\mathbf{U}_i)' \overline{\mathbf{\Gamma}}_1(\mathbf{U}_i) \mathcal{V}_n(\mathbf{U}_i) \\ &\quad + (\mathbf{h}!)^{1/2} \sum_{i=1}^{n} \mathcal{V}_n(\mathbf{U}_i)' \left[\mathbf{\Gamma}_{n1}(\mathbf{U}_i) - \overline{\mathbf{\Gamma}}_1(\mathbf{U}_i)\right]' \left[\mathbf{\Gamma}_{n1}(\mathbf{U}_i) - \overline{\mathbf{\Gamma}}_1(\mathbf{U}_i)\right] \mathcal{V}_n(\mathbf{U}_i) \\ &\quad + 2(\mathbf{h}!)^{1/2} \sum_{i=1}^{n} \mathcal{V}_n(\mathbf{U}_i)' \left[\mathbf{\Gamma}_{n1}(\mathbf{U}_i) - \overline{\mathbf{\Gamma}}_1(\mathbf{U}_i)\right]' \mathcal{V}_n(\mathbf{U}_i) \overline{\mathbf{\Gamma}}_1(\mathbf{U}_i) \mathcal{V}_n(\mathbf{U}_i) \\ &\equiv \overline{T}_{11a} + \overline{T}_{11b} + 2\overline{T}_{11c}, \text{ say.} \end{aligned} \tag{B.5}$$

Using the definitions of ζ_i and $\varphi(\zeta_i, \zeta_j)$ in (4.9) we have

$$\begin{aligned} \overline{T}_{n11a} &= \frac{(\mathbf{h}!)^{1/2}}{n^2} \sum_{i=1}^{n} \sum_{j=1}^{n} \sum_{l=1}^{n} \varphi(\zeta_i, \zeta_j)' \varphi(\zeta_i, \zeta_l) \\ &= \frac{(\mathbf{h}!)^{1/2}}{n^2} \sum_{i=1}^{n} \sum_{j=1}^{n} \|\varphi(\zeta_i, \zeta_j)\|^2 + \frac{(\mathbf{h}!)^{1/2}}{n^2} \sum_{i=1}^{n} \sum_{j\neq i}^{n} \sum_{l\neq j,i}^{n} \varphi(\zeta_i, \zeta_j)' \varphi(\zeta_i, \zeta_l) \\ &\quad + \frac{2(\mathbf{h}!)^{1/2}}{n^2} \sum_{i=1}^{n} \sum_{j\neq i}^{n} \varphi(\zeta_i, \zeta_j)' \varphi(\zeta_i, \zeta_i) \equiv B_n + V_{n1} + R_{n1}, \text{ say.} \end{aligned} \tag{B.6}$$

Let $\overline{\varphi}(\zeta_i,\zeta_j,\zeta_l) \equiv [\varphi(\zeta_i,\zeta_j)'\varphi(\zeta_i,\zeta_l) + \varphi(\zeta_j,\zeta_i)'\varphi(\zeta_j,\zeta_l) + \varphi(\zeta_l,\zeta_i)'\varphi(\zeta_l,\zeta_j)]/3$. Then $V_{n1} = 6\left(\mathbf{h}!\right)^{1/2} n^{-2} \sum_{1\le i<j<l\le n} \overline{\varphi}(\zeta_i,\zeta_j,\zeta_l) = \frac{(n-1)(n-2)}{n}\overline{V}_{n1}$, where $\overline{V}_{n1} \equiv \frac{6(\mathbf{h}!)^{1/2}}{n(n-1)(n-2)} \sum_{1\le i<j<l\le n} \overline{\varphi}(\zeta_i,\zeta_j,\zeta_l)$. Note that for all $i \neq j \neq l$, $\theta \equiv E[\overline{\varphi}(\zeta_i,\zeta_j,\zeta_l)] = 0$, $\overline{\varphi}_1(\mathbf{a}) \equiv E[\overline{\varphi}(\mathbf{a},\zeta_j,\zeta_l)] = 0$, and $\overline{\varphi}_2(\mathbf{a},\widetilde{\mathbf{a}}) \equiv E[\overline{\varphi}(\mathbf{a},\widetilde{\mathbf{a}},\zeta_l)] = \frac{1}{3}E[\varphi(\zeta_l,\mathbf{a})'\varphi(\zeta_l,\widetilde{\mathbf{a}})]$, where $\mathbf{a}$ and $\widetilde{\mathbf{a}}$ are nonrandom. Let $\overline{\varphi}_3(\mathbf{a},\widetilde{\mathbf{a}},\overline{\mathbf{a}}) \equiv \overline{\varphi}(\mathbf{a},\widetilde{\mathbf{a}},\mathbf{a}) - \overline{\varphi}_2(\mathbf{a},\widetilde{\mathbf{a}}) - \overline{\varphi}_2(\mathbf{a},\overline{\mathbf{a}}) - \overline{\varphi}_2(\widetilde{\mathbf{a}},\mathbf{a})$. By the Hoeffding decomposition, $\overline{V}_{n1} = 3H_n^{(2)} + H_n^{(3)}$ where $H_n^{(2)} \equiv \frac{2(h!)^{1/2}}{n(n-1)}\sum_{1\le i<j\le n}\overline{\varphi}_2(\zeta_i,\zeta_j)$ and $H_n^{(3)} \equiv \frac{6(h!)^{1/2}}{n(n-1)(n-2)}\sum_{1\le i<j<l\le n}\overline{\varphi}_3(\zeta_i,\zeta_j,\zeta_l)$. Noting that $E[\overline{\varphi}_3(\mathbf{a},\widetilde{\mathbf{a}},\zeta_i)] = 0$ and that $\overline{\varphi}_3$ is symmetric in its arguments by construction, it is straightforward to show that $E[H_n^{(3)}] = 0$ and $E[H_n^{(3)}]^2 = O(n^{-3}\left(\mathbf{h}!\right)^{-1})$. Hence, $H_n^{(3)} = O_P(n^{-3/2}\left(\mathbf{h}!\right)^{-1/2}) = o_P\left(n^{-1}\right)$ by Chebyshev's 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

$$\mathcal{H}_n \equiv \frac{2\left(\mathbf{h}!\right)^{1/2}}{n}\sum_{1\le i<j\le n} 3\overline{\varphi}_2(\zeta_i,\zeta_j) = \frac{2\left(\mathbf{h}!\right)^{1/2}}{n}\sum_{1\le i<j\le n}\int \varphi(\mathbf{a},\zeta_i)'\varphi(\mathbf{a},\zeta_j)dF_\zeta(\mathbf{a}).$$

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 $\mathcal{H}_n : \mathcal{H}_n \xrightarrow{d} N\left(0,\sigma_0^2\right)$, where the asymptotic variance of $\mathcal{H}_n$ is given by $\sigma_0^2 \equiv \lim_{n\to\infty}\sigma_n^2$ and $\sigma_n^2 \equiv 2\mathbf{h}!E_jE_l[\int\varphi(\zeta,\zeta_j)'\varphi(\zeta,\zeta_l)F_\zeta(d\zeta)]^2$. Consequently $V_{n1} \xrightarrow{d} N\left(0,\sigma_0^2\right)$. For R_{n1}, it is easy to verify that $E(R_{n1}) = 0$ and $E\left(R_{n1}^2\right) = O\left((n\mathbf{h}!)^{-1}\right) = o(1)$. So $R_{n1} = o_P(1)$ by Chebyshev's inequality and

$$\overline{T}_{n11a} - B_n \xrightarrow{d} N\left(0,\sigma_0^2\right). \tag{B.7}$$

By (B.1)-(B.2) and Assumption A8, we have

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

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

Under $\mathbb{H}_1(r_n)$, $\overline{\mathbf{g}}_1 = n^{-1}\sum_{i=1}^n \mathbf{g}_1(\mathbf{U}_i) = \theta_1 + r_n\overline{\delta}_n$, where $\overline{\delta}_n = n^{-1}\sum_{i=1}^n \delta_n(\mathbf{U}_i) = E[\delta_n(\mathbf{U}_i)] + O_P\left(n^{-1/2}\right)$. So $T_{n12} = n^{-1}\sum_{i=1}^n\left\|\delta_n(\mathbf{U}_i) - \overline{\delta}_n\right\|^2 = \lim_{n\to\infty} E\left[\left\|\delta_n(\mathbf{U}_i) - E[\delta_n(\mathbf{U}_i)]\right\|^2\right] = \mu_0$, and

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

where $\overline{T}_{n13} = \frac{r_n(\mathbf{h}!)^{1/2}}{n}\sum_{i=1}^n\sum_{j\neq i}^n\varphi(\zeta_i,\zeta_j)\{\delta_n(\mathbf{U}_i) - E[\delta_n(\mathbf{U}_i)]\}$. 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 Chebyshev's inequality. Thus $T_{n13} = o_P(1)$. In sum, we have $T_{n1} - B_n - \mu_0 \xrightarrow{d} N\left(0,\sigma_0^2\right)$.

Now we show (ii). Noting that $\overline{\widehat{\mathbf{g}}}_1 - \overline{\mathbf{g}}_1 = \frac{1}{n}\sum_{i=1}^n[\mathbf{\Gamma}_{n1}(\mathbf{U}_i)\mathcal{B}_n(\mathbf{U}_i) + \mathbf{\Gamma}_{n1}(\mathbf{U}_i)\mathcal{V}_n(\mathbf{U}_i)]$, we have $T_{n2} = \frac{(\mathbf{h}!)^{1/2}}{n}\sum_{i=1}^n\sum_{j=1}^n\mathcal{B}_n(\mathbf{U}_i)'\mathbf{\Gamma}_{n1}(\mathbf{U}_i)'\mathbf{\Gamma}_{n1}(\mathbf{U}_j)\mathcal{B}_n(\mathbf{U}_j) + \frac{(\mathbf{h}!)^{1/2}}{n}\sum_{i=1}^n\sum_{j=1}^n\mathcal{V}_n(\mathbf{U}_i)\mathbf{\Gamma}_{n1}(\mathbf{U}_i)'\mathbf{\Gamma}_{n1}(\mathbf{U}_j)\mathcal{V}_n(\mathbf{U}_j) + \frac{2(\mathbf{h}!)^{1/2}}{n}\sum_{i=1}^n\sum_{j=1}^n\mathcal{B}_n(\mathbf{U}_i)'\mathbf{\Gamma}_{n1}(\mathbf{U}_i)'\mathbf{\Gamma}_{n1}(\mathbf{U}_j)\mathcal{V}_n(\mathbf{U}_j) \equiv T_{n21} + T_{n22} + 2T_{n23}$, say. For T_{n21}, we have $T_{n21} \le \sup_{\mathbf{u}}\|\mathcal{B}_n(\mathbf{u})\|^2\frac{(\mathbf{h}!)^{1/2}}{n}\sum_{i=1}^n\sum_{j=1}^n\mathrm{tr}(\mathbf{\Gamma}_{n1}(\mathbf{U}_i)'\mathbf{\Gamma}_{n1}(\mathbf{U}_j)) = n\left(\mathbf{h}!\right)^{1/2}\sup_{\mathbf{u}}\|\mathcal{B}_n(\mathbf{u})\|^2\mathrm{tr}(\overline{\mathbf{\Gamma}}_{n1}'\overline{\mathbf{\Gamma}}_{n1}) = n\left(\mathbf{h}!\right)^{1/2}O_P((\|\mathbf{h}\|^2 + \|\lambda\|)^2) = o_P(1)$, 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 $E\left|\overline{T}_{n2,2}\right| = E\left[\overline{T}_{n2,2}\right] = O((\mathbf{h}!)^{1/2} + n^{-1}(\mathbf{h}!)^{-1})$, we have $T_{n22} = o_P(1)$ by Markov's inequality. Then by Cauchy-Schwarz's inequality, $T_{n23} \leq \{T_{n21}\}^{1/2} \{T_{n22}\}^{1/2} = o_P(1)$. 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 $n^{-1}(\mathbf{h}!)^{-1/2} T_{n2} = o_P(1)$ and

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

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. ■

Proof of Theorems 4.5. Let P^* denote the probability measure induced by the wild bootstrap conditional on the original sample $\mathcal{W}_n$ and E^* and Var^* denote the expectation and variance with respect to P^*. Let $O_{P^*}(\cdot)$ and $o_{P^*}(\cdot)$ denote the probability order under P^*; e.g., $b_n = o_{P^*}(1)$ if for any $\epsilon > 0$, $P^*(\|b_n\| > \epsilon) = o_P(1)$. Note that $b_n = o_P(1)$ implies that $b_n = o_{P^*}(1)$. The proof follows closely from that of Theorem 4.3.

Let $\widehat{\mathbf{g}}_1^*$, $\overline{\widehat{\mathbf{g}}}_1^*$, T_n^*, Γ_{n1}^*, $\mathcal{B}_n^*$, $\mathcal{V}_n^*$, B_n^*, $\widehat{B}_n^*$ and $\widehat{\sigma}_n^{*2}$ denote the bootstrap analogue of $\widehat{\mathbf{g}}_1$, $\overline{\widehat{\mathbf{g}}}_1$, T_n, Γ_{n1}, $\mathcal{B}_n$, $\mathcal{V}_n$, B_n, $\widehat{B}_n$ and $\widehat{\sigma}_n^2$, respectively. For example, $\mathbf{\Gamma}_{n1}^*(\mathbf{u}) = \mathbb{S}_1[\mathbf{\Phi}_n(\mathbf{u})' \mathbf{\Psi}_n^*(\mathbf{u})^{-1} \mathbf{\Phi}_n(\mathbf{u})]^{-1} \mathbf{\Phi}_n(\mathbf{u})' \mathbf{\Psi}_n^*(\mathbf{u})^{-1}$ as $\mathbf{\Phi}_n(\mathbf{u})$ is the same in both the real data and bootstrap worlds. Note that $\widehat{\theta}_1 = n^{-1} \sum_{i=1}^n \widehat{\mathbf{g}}_1(\mathbf{U}_i)$ in the bootstrap world plays the role of $\mathbf{g}_1(\cdot)$ in the real data world. The decomposition of T_n in (B.3) continues to hold for T_n^* in the bootstrap world: $T_n^* = (\mathbf{h}!)^{1/2} \sum_{i=1}^n [\widehat{\mathbf{g}}_1^*(\mathbf{U}_i^*) - \widehat{\theta}_1 + \widehat{\theta}_1 - \overline{\widehat{\mathbf{g}}}_1^*]' [\widehat{\mathbf{g}}_1^*(\mathbf{U}_i^*) - \widehat{\theta}_1 + \widehat{\theta}_1 - \overline{\widehat{\mathbf{g}}}_1^*] = T_{n1}^* - T_{n2}^*$, where $T_{n1}^* = (\mathbf{h}!)^{1/2} \sum_{i=1}^n [\widehat{\mathbf{g}}_1^*(\mathbf{U}_i) - \widehat{\theta}_1]' [\widehat{\mathbf{g}}_1^*(\mathbf{U}_i) - \widehat{\theta}_1]$, and $T_{n2} = n(\mathbf{h}!)^{1/2} [\widehat{\theta}_1 - \overline{\widehat{\mathbf{g}}}_1^*]' [\widehat{\theta}_1 - \overline{\widehat{\mathbf{g}}}_1^*]$. We prove the first part of the theorem by showing that (i) $(T_{n1}^* - B_n^*)/\sigma_n^{*2} \to N(0,1)$ in distribution in probability, (ii) $T_{n2}^* = o_{P^*}(1)$, (iii) $\widehat{B}_n^* - B_n^* = o_{P^*}(1)$, and (iv) $\widehat{\sigma}_n^{*2} - \sigma_n^{*2} = o_{P^*}(1)$, where σ_n^{*2} is defined below.

Note that $Y_i^* = \widehat{\theta}_1' \mathbf{X}_{1i} + \widehat{\theta}_2' \mathbf{X}_{2i} + \varepsilon_i^*$. As $\widehat{\theta}_1$ and $\widehat{\theta}_2$, when treated as functions of $\mathbf{u}$, have zero derivatives up to the infinite order, $\mathcal{B}_n^*(\mathbf{U}_i) = 0$ for all i and (4.7) now takes the following form in the bootstrap world

$$\widehat{\mathbf{g}}_1^*(\mathbf{U}_i) - \widehat{\theta}_1 = \mathbf{\Gamma}_{n1}^*(\mathbf{U}_i)\, \mathcal{V}_n^*(\mathbf{U}_i) \tag{B.8}$$

By the definition of $\mathbf{\Gamma}_{n1}^*(\mathbf{u})$ and the extra condition in the theorem

$$\mathbf{\Gamma}_{n1}^*(\mathbf{u}) = \mathbf{\Gamma}_{n1}(\mathbf{u}) + O_{P^*}(\nu_n) \text{ uniformly in } \mathbf{u}. \tag{B.9}$$

Let $\zeta_i^*(\mathbf{u}) \equiv \begin{pmatrix} \mathbf{Q}_i \varepsilon_i^* \\ (\mathbf{Q}_i \varepsilon_i^*) \otimes \eta_i(\mathbf{u}^c) \end{pmatrix} K_{\mathbf{h}\lambda, i\mathbf{u}}$. Then $\mathcal{V}_n^*(\mathbf{u}) = n^{-1} \sum_{i=1}^n \zeta_i^*$. Observing that $E^*[\mathcal{V}_n^*(\mathbf{u})] = 0$ and $\mathrm{Var}^*[\mathcal{V}_n^*(\mathbf{u})] = n^{-2} \sum_{i=1}^n E^*[\zeta_i^*(\mathbf{u}) \zeta_i^*(\mathbf{u})'] = O_P((n\mathbf{h}!)^{-1})$, $\mathcal{V}_n^*(\mathbf{u}) = O_{P^*}((n\mathbf{h}!)^{-1/2})$ by Chebyshev's inequality. By the use of the exponential inequality, this result can be strengthened to

$$\mathcal{V}_n^*(\mathbf{u}) = O_{P^*}((n\mathbf{h}!)^{-1/2} \sqrt{\log n}) \text{ uniformly in } \mathbf{u}. \tag{B.10}$$

To show (i), observe that by (B.8)-(B.10) and Assumption A8, $T_{n1}^* = \overline{T}_{n1}^* + o_{P^*}(1)$, where $\overline{T}_{n1}^* = (\mathbf{h}!)^{1/2} \sum_{i=1}^n \mathcal{V}_n^*(\mathbf{U}_i)' \mathbf{\Gamma}_{n1}(\mathbf{U}_i)' \mathbf{\Gamma}_{n1}(\mathbf{U}_i) \mathcal{V}_n^*(\mathbf{U}_i)$. Let $\varphi_{in}^*(e_j) = \Gamma_{n1}(\mathbf{U}_i) \zeta_j^*(\mathbf{U}_i)$ and $\bar{\varphi}_n^*(e_i, e_j) = n^{-1} \sum_{s=1}^n \varphi_{sn}^*(e_i)' \varphi_{sn}^*(e_j)$. Then

$$\overline{T}_{n1}^* = \frac{(\mathbf{h}!)^{1/2}}{n} \sum_{i=1}^n \bar{\varphi}_n^*(e_i, e_i) + \frac{2(\mathbf{h}!)^{1/2}}{n} \sum_{1 \leq i < j \leq n} \bar{\varphi}_n^*(e_i, e_j) \equiv B_n^* + V_{n1}^*, \text{ say.}$$

As V_{n1}^* is a second order degenerate U-statistic, we can apply the CLT for second order degenerate U-statistic for independent but nonidentically distributed observations (e.g., De Jong, 1987) and conclude that $V_{n1}^*/\sigma_n^{*2} \to N(0,1)$ in distribution in probability, where $\sigma_n^{*2} \equiv 2\mathbf{h}!E^*[\overline{\varphi}_n^*(e_1,e_2)]^2$. Then (i) follows.

Now we show (ii). By (B.8)-(B.10) and Assumption A8, $T_{n2} = \frac{(\mathbf{h}!)^{1/2}}{n}\sum_{i=1}^n\sum_{j=1}^n \mathcal{V}_n^*(\mathbf{U}_i)\,\mathbf{\Gamma}_{n1}^*(\mathbf{U}_i)'$ $\mathbf{\Gamma}_{n1}^*(\mathbf{U}_j)\,\mathcal{V}_n^*(\mathbf{U}_j) = \overline{T}_{n2}^* + o_P(1)$, where $\overline{T}_{n2}^* = \frac{(\mathbf{h}!)^{1/2}}{n}\sum_{i=1}^n\sum_{j=1}^n \mathcal{V}_n^*(\mathbf{U}_i)\;\Gamma_{n1}(\mathbf{U}_i)'\,\Gamma_{n1}(\mathbf{U}_j)\,\mathcal{V}_n^*(\mathbf{U}_j)$. Then (ii) follows by the fact that $E^*|\overline{T}_{n2}^*| = E^*[\overline{T}_{n2}^*] = o_P(1)$ and Markov's inequality. For (iii), noting that $B_n^* = n^{-2}(\mathbf{h}!)^{1/2}\sum_{i=1}^n\sum_{j=1}^n ||\varphi_{jn}^*(e_i)||^2$ and the bootstrap analogue of $\widehat{B}_n$ is given by $\widehat{B}_n^* = n^{-2}(\mathbf{h}!)^{1/2}\sum_{i=1}^n\sum_{j=1}^n ||\widehat{\varphi}_{ij}^*||^2$ where $\widehat{\varphi}_{ij}^* = \mathbf{\Gamma}_{n1}(\mathbf{U}_i)\begin{pmatrix}\mathbf{Q}_j\widehat{\varepsilon}_j^* \\ (\mathbf{Q}_j\widehat{\varepsilon}_j^*)\otimes\eta_j(\mathbf{U}_i^c)\end{pmatrix}K_{\mathbf{h}\lambda,j\mathbf{U}_i}$ and $\widehat{\varepsilon}_j^* = Y_j^* - \widehat{\mathbf{g}}_1^*(\mathbf{U}_j)'\mathbf{X}_{1j} - \widehat{\mathbf{g}}_2^*(\mathbf{U}_j)'\mathbf{X}_{2j}$, it is standard to show that $\widehat{B}_n^* = B_n^* + o_{P^*}(1)$ by using (B.8)-(B.10), the corresponding result for $\widehat{\mathbf{g}}_2^*$, and Chebyshev's inequality. Analogously, we can prove (iv).

Let $\bar{z}_\alpha^*$ denote the $1-\alpha$ conditional quantile of J_n^* given $\mathcal{W}_n$, i.e., $P(J_n^* \geq \bar{z}_\alpha^*|\mathcal{W}_n) = \alpha$. Recall z_α^* is the $1-\alpha$ quantile of the empirical distribution of $\{J_{nj}^*\}_{j=1}^B$. By choosing B sufficiently large, the approximation error of z_α^* to $\bar{z}_\alpha^*$ can be made arbitrarily small and negligible. By the first part of the theorem, $\bar{z}_\alpha^* \to z_\alpha$ in probability. Then in view of Theorem 4.2 and the remark after it, $\lim_{n\to\infty} P(J_n \geq z_\alpha^*) = \lim_{n\to\infty} P(J_n \geq z_\alpha) = \alpha$ under $\mathbb{H}_0$. By Theorem 4.3 and the fact that $\widehat{B}_n = B_n + o_P(1)$ and $\widehat{\sigma}_n^2 = \sigma_0^2 + o_P(1)$ under $\mathbb{H}_1(n^{-1/2}(\mathbf{h}!)^{-1/4})$, we have $\lim_{n\to\infty} P(J_n \geq z_\alpha^*) = \lim_{n\to\infty} P(J_n \geq z_\alpha) = 1-\Phi(z_\alpha - \mu_0/\sigma_0)$ under $\mathbb{H}_1(n^{-1/2}(\mathbf{h}!)^{-1/4})$. By Theorem 4.4 $\lim_{n\to\infty} P(J_n \geq z_\alpha^*) = \lim_{n\to\infty} P(J_n \geq z_\alpha) = 1$ under $\mathbb{H}_1$.■

ACKNOWLEDGMENTS

The authors sincerely thank Keisuke Hirano, the associate editor, and two anonymous referees for their many insightful comments and suggestions that lead to a substantial improvement of the presentation. They are also thankful to David Card, Todd Sorensen, and seminar participants at the University of Queensland, City University of Hong Kong, and the conference in honor of M. Hashem Pesaran at Cambridge University. The third author gratefully acknowledges the financial support from the Academic Senate, UCR.

References

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

Altonji, J. and L. Segal (1996), "Small-sample Bias in GMM Estimation of Covariance Structures," *Journal of Business & Economic Statistics* 14, 353-366.

Baltagi, B. H. and Q. Li (2002), "On Instrumental Variable Estimation of Semiparametric Dynamic Panel Data Models," *Economics Letters* 76, 1-9.

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.

De Jong, P. (1987), "A Central Limit Theorem for Generalized Quadratic Forms," *Probability Theory and Related Fields* 75, 261-277.

Duncan, G. M. and D. E. Leigh (1985), "The Endogeneity of Union Status: An Empirical Test," *Journal of Labor Economics* 3, 385-402.

Fan, J. and T. Huang (2005), "Profile Likelihood Inferences on Semiparametric Varying-Coefficient Partially Linear Models," *Bernoulli* 11, 1031-1057.

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), *Modeling 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., Q. Li, and J. Racine (2007), "Nonparametric Estimation of Regression Functions in the Presence of Irrelevant Regressors," *Review of Economic and Statistics* 89, 784-789.

Hall, P., J. Racine, and Q. Li (2004), "Cross-Validation and the Estimation of Conditional Probability Densities," *Journal of the American Statistical Association* 99, 1015-1026.

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.

Hansen, B. E. (2000), "Testing for Structural Change in Conditional Models," *Journal of Econometrics* 97, 93-115.

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.

Korenman, S. and D. Neumark (1992), "Marriage, Motherhood, and Wages," *Journal of Human Resources* 27, 233-255.

Lee, J. (2005), "Marriage, Female Labor Supply, and Asian Zodiacs," *Economics Letters* 87, 427-432.

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

Li, D., and Q. Li (2010), "Nonparametric/Semiparametric Estimation and Testing of Economic Models with Data Dependent Smoothing Parameters," *Journal of Econometrics* 157, 179-190.

Li, Q., C. Hsiao, and J. Zinn (2003), Consistent Specification Tests for Semiparametric/nonparametric Models Based on Series Estimation Method. *Journal of Econometrics* 112, 295-325.

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.

Mammen, E., C. Rothe, and M. Schienle (2010), "Nonparametric Regression with Nonparametrically Generated Covariates," *Working paper*, University of Mannheim.

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.

Newey, W. K. (1993), "Efficient Estimation of Models with Conditional Moment Restrictions," in G. S. Maddala, C. R. Rao and H. D. Vinod (eds), *Handbook of Statistics* 11, 419-454.

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. (2012), "A Semi-parametric GMM Estimation of Spatial Autoregressive Models," *Journal of Econometrics* 167, 543-560.

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

Su, L. and S. Jin (2010), "Profile Quasi-maximum Likelihood Estimation of Spatial Autoregressive Models," *Journal of Econometrics* 157, 18-33.

Su, L. and A. Ullah (2012), "A Nonparametric Goodness-of-fit-based Test for Conditional Heteroskedasticity," *Econometric Theory*, forthcoming.

Su, L. and H. White (2010), "Testing Structural Change in Partially Linear Models," *Econometric Theory* 26, 1761-1806.

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.

Table 1: Finite Sample Comparison of Various Nonparametric Estimators

DGP	n	τ	Estimates	Homoskedasticity				Heteroskedasticity			
				g_1		g_2		g_1		g_2	
				MAD	MSE	MAD	MSE	MAD	MSE	MAD	MSE
1	100	0.32	SCU	0.742	0.984	0.400	0.304	1.247	2.722	0.703	0.920
			IW_{ll}	0.497	0.433	0.264	0.124	0.754	1.091	0.468	0.400
			OIV_{ll}	0.498	0.437	0.284	0.138	0.580	0.622	0.380	0.253
			IW_{lc}	0.671	0.828	0.391	0.274	0.884	1.585	0.547	0.555
			OIV_{lc}	0.676	0.814	0.419	0.300	0.700	0.951	0.469	0.395
			CDXW	0.699	0.967	0.373	0.268	0.929	1.679	0.566	0.594
		0.75	SCU	0.982	1.251	0.628	0.515	1.744	3.899	1.100	1.579
			IW_{ll}	0.540	0.684	0.340	0.314	0.834	1.769	0.602	0.811
			OIV_{ll}	0.473	0.391	0.310	0.168	0.549	0.545	0.421	0.305
			IW_{lc}	0.680	1.018	0.456	0.469	0.916	2.029	0.658	0.938
			OIV_{lc}	0.636	0.729	0.449	0.356	0.674	0.892	0.517	0.491
			CDXW	0.845	1.470	0.522	0.566	1.074	2.301	0.751	1.092
	400	0.32	SCU	0.603	0.503	0.322	0.139	1.071	1.622	0.575	0.469
			IW_{ll}	0.323	0.192	0.164	0.049	0.525	0.572	0.316	0.196
			OIV_{ll}	0.328	0.195	0.178	0.056	0.406	0.337	0.258	0.127
			IW_{lc}	0.379	0.250	0.217	0.080	0.535	0.550	0.332	0.198
			OIV_{lc}	0.402	0.280	0.247	0.103	0.424	0.353	0.281	0.144
			CDXW	0.455	0.441	0.226	0.105	0.640	0.876	0.371	0.274
		0.75	SCU	0.946	0.988	0.602	0.395	1.697	3.261	1.060	1.253
			IW_{ll}	0.344	0.287	0.205	0.112	0.565	0.863	0.400	0.386
			OIV_{ll}	0.323	0.195	0.202	0.075	0.400	0.333	0.299	0.177
			IW_{lc}	0.379	0.287	0.247	0.122	0.549	0.686	0.396	0.322
			OIV_{lc}	0.388	0.266	0.269	0.126	0.421	0.365	0.320	0.196
			CDXW	0.573	0.770	0.332	0.258	0.754	1.284	0.502	0.534
2	100	0.32	SCU	0.626	0.692	0.664	2.882	1.006	1.636	0.611	0.673
			IW_{ll}	0.421	0.398	0.795	2.967	0.585	0.778	0.440	0.506
			OIV_{ll}	0.518	0.720	0.686	2.005	0.460	0.594	0.402	0.432
			IW_{lc}	0.546	0.701	0.929	3.141	0.648	0.946	0.526	0.639
			OIV_{lc}	0.602	0.808	0.892	2.705	0.552	0.760	0.501	0.580
			CDXW	0.523	0.615	0.826	3.012	0.662	0.976	0.467	0.559
		0.75	SCU	0.917	1.051	0.690	1.915	1.541	2.865	1.008	1.262
			IW_{ll}	0.434	0.466	0.623	1.775	0.618	1.007	0.518	0.681
			OIV_{ll}	0.494	0.689	0.566	1.261	0.468	0.636	0.444	0.451
			IW_{lc}	0.533	0.698	0.747	1.954	0.648	1.014	0.571	0.727
			OIV_{lc}	0.579	0.778	0.731	1.696	0.561	0.819	0.531	0.611
			CDXW	0.601	0.785	0.696	1.829	0.743	1.208	0.577	0.760
	400	0.32	SCU	0.533	0.343	0.322	0.395	0.912	1.010	0.508	0.332
			IW_{ll}	0.205	0.082	0.213	0.346	0.309	0.214	0.225	0.094
			OIV_{ll}	0.242	0.188	0.230	0.362	0.240	0.154	0.202	0.078
			IW_{lc}	0.277	0.176	0.292	0.404	0.343	0.257	0.273	0.130
			OIV_{lc}	0.309	0.213	0.313	0.387	0.287	0.208	0.260	0.114
			CDXW	0.257	0.133	0.228	0.364	0.353	0.276	0.240	0.110
		0.75	SCU	0.903	0.854	0.599	0.389	1.530	2.491	0.984	1.041
			IW_{ll}	0.209	0.092	0.177	0.079	0.319	0.259	0.261	0.141
			OIV_{ll}	0.241	0.199	0.195	0.119	0.245	0.177	0.226	0.107
			IW_{lc}	0.270	0.168	0.252	0.135	0.339	0.257	0.298	0.164
			OIV_{lc}	0.300	0.202	0.277	0.155	0.290	0.215	0.278	0.140
			CDXW	0.309	0.202	0.224	0.113	0.404	0.365	0.302	0.182

Note. See the text for the definition of the six estimates: SCU, IW_{ll}, OIV_{ll}, IW_{lc}, OIV_{lc}, and CDXW. The MAD and MSE are averages over 500 replications.

Table 2: Rejection Frequency of Nonparametric Tests for the Constancy of Functional Coefficients (nominal level: 0.05)

DGP	n	δ	τ	Homoskedasticity			Heteroskedasticity		
				$\mathbb{H}_{0,1}$	$\mathbb{H}_{0,2}$	$\mathbb{H}_{0,12}$	$\mathbb{H}_{0,1}$	$\mathbb{H}_{0,2}$	$\mathbb{H}_{0,12}$
1	100	0	0.32	0.042	0.058	0.046	0.056	0.058	0.048
			0.75	0.016	0.066	0.032	0.022	0.062	0.036
		0.2	0.32	0.180	0.922	0.732	0.126	0.358	0.354
			0.75	0.192	0.824	0.696	0.096	0.280	0.270
		0.4	0.32	0.800	0.996	1.000	0.560	0.890	0.886
			0.75	0.862	0.998	0.996	0.602	0.754	0.804
		0.6	0.32	0.988	1.000	1.000	0.892	0.994	0.992
			0.75	0.992	1.000	1.000	0.920	0.958	0.974
	200	0	0.32	0.058	0.048	0.058	0.068	0.036	0.056
			0.75	0.026	0.056	0.050	0.036	0.044	0.048
		0.2	0.32	0.282	0.996	0.924	0.156	0.572	0.478
			0.75	0.292	0.984	0.928	0.154	0.448	0.420
		0.4	0.32	0.904	1.000	1.000	0.648	0.992	0.990
			0.75	0.968	1.000	1.000	0.738	0.960	0.970
		0.6	0.32	1.000	1.000	1.000	0.952	1.000	1.000
			0.75	1.000	1.000	1.000	0.986	0.998	1.000
2	100	0	0.32	0.018	0.040	0.024	0.038	0.066	0.048
			0.75	0.018	0.026	0.022	0.026	0.046	0.038
		0.2	0.32	0.038	0.754	0.184	0.048	0.264	0.112
			0.75	0.038	0.618	0.196	0.040	0.190	0.092
		0.4	0.32	0.204	0.990	0.734	0.132	0.724	0.438
			0.75	0.214	0.982	0.768	0.132	0.580	0.402
		0.6	0.32	0.366	0.914	0.882	0.280	0.960	0.786
			0.75	0.424	0.994	0.980	0.318	0.866	0.738
	200	0	0.32	0.060	0.066	0.064	0.060	0.058	0.052
			0.75	0.042	0.048	0.040	0.038	0.050	0.040
		0.2	0.32	0.132	0.944	0.386	0.088	0.398	0.172
			0.75	0.126	0.864	0.374	0.076	0.288	0.156
		0.4	0.32	0.474	1.000	0.986	0.260	0.940	0.684
			0.75	0.494	1.000	0.970	0.268	0.836	0.632
		0.6	0.32	0.792	0.992	0.976	0.552	0.998	0.972
			0.75	0.848	1.000	1.000	0.572	0.986	0.936

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).

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

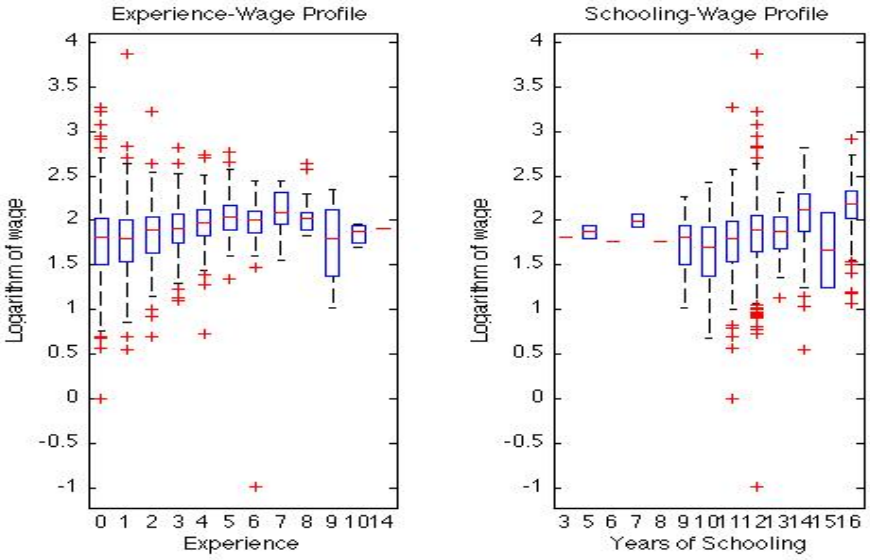
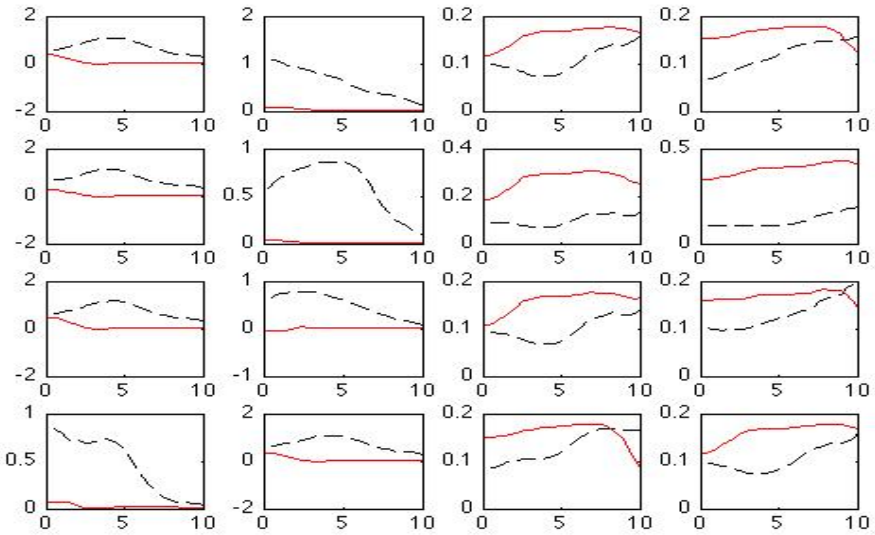
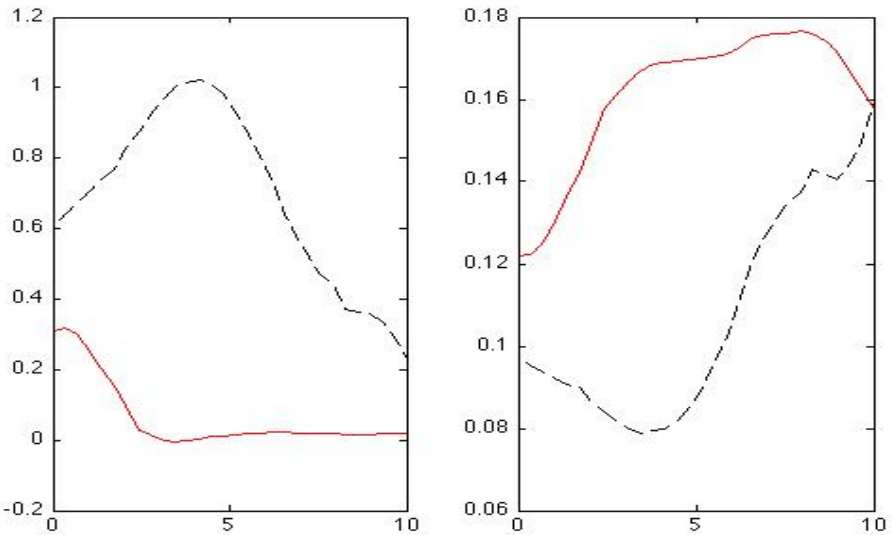


Figure 2: **Plots of** $g_1(Experience,\ Individual\ Characteristic,:)$ **and** $g_2(Experience,\ Individual\ Characteristic,:)$ **Averaged over Other Categorical Variables**



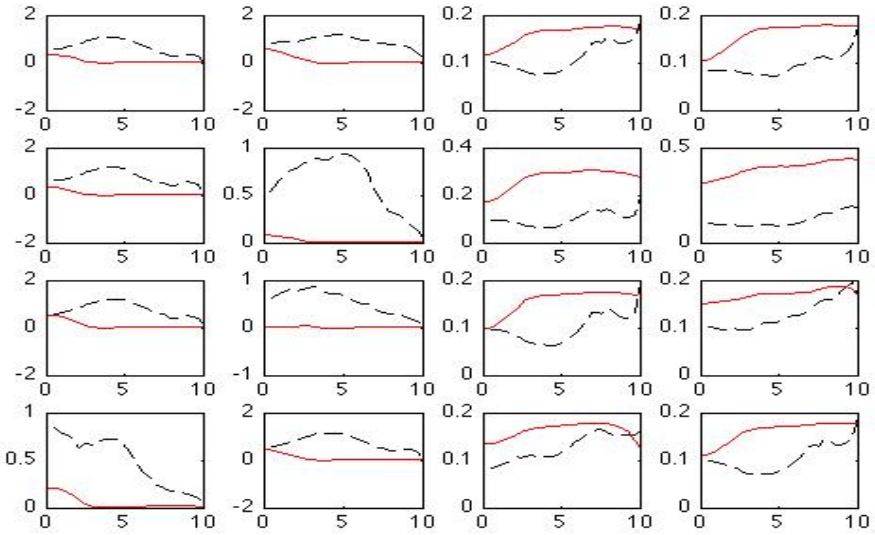
Notes: Horizontal Axis - Experience. Vertical Axis - g_1 or g_2. SCU estimate, dashed line; our proposed estimate with optimal weight matrix, solid line. The rule of thumb method is used to choose the bandwidth. The four rows correspond to *Individual Characteristic* being a binary indicator of whether 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**



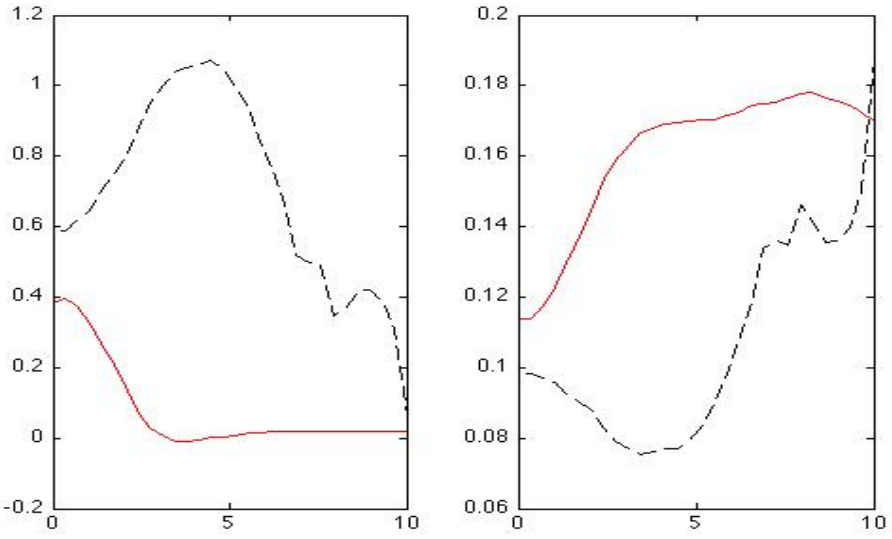
Notes: Horizontal Axis - Experience. Vertical Axis - g_1 or g_2. SCU estimate, dashed line; our proposed estimate with optimal weight matrix, solid line. The rule of thumb method is used to choose the bandwidth. The two columns from the left to the right correspond to g_1 and g_2, respectively.

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.