Additive Nonparametric Regression in the Presence of Endogenous Regressors

Deniz Ozabaci*
Department of Economics
State University of New York at Binghamton

Daniel J. Henderson†
Department of Economics, Finance and Legal Studies
University of Alabama

Liangjun Su‡
School of Economics
Singapore Management University

June 17, 2013

Abstract

In this paper we consider nonparametric estimation of a structural equation model under full additivity constraint. We propose estimators for both the conditional mean and gradient which are consistent, asymptotically normal, oracle efficient and free from curse of dimensionality. Monte Carlo simulations support the asymptotic developments. We employ a partially linear extension of our model to study the relationship between child care and cognitive outcomes. Some of our (average) results are consistent with the literature (e.g., negative returns to child care when mothers have higher levels of education). However, as our estimators allow for heterogeneity both across and within groups, we are able to contradict many findings in the literature (e.g., we do not find any significant differences in returns between boys and girls or for formal versus informal child care).

Keywords: Additive Regression, Endogeneity, Generated Regressors, Oracle Estimation, Nonparametric, Structural Equation
JEL Classification Codes: C14, C36, I21, J13

*Deniz Ozabaci, Department of Economics, State University of New York, Binghamton, NY 13902-6000, (607) 777-2572, Fax: (607) 777-2681, e-mail: dozabac1@binghamton.edu.

†Daniel J. Henderson, Department of Economics, Finance and Legal Studies, University of Alabama, Tuscaloosa, AL 35487-0224, (205) 348-8991, Fax: (205) 348-0186, e-mail: djhender@cba.ua.edu.

‡Liangjun Su, School of Economics, Singapore Management University, 90 Stamford Road, Singapore, 178903; Tel: (65) 6828-0386; e-mail: ljsu@smu.edu.sg.

1 Introduction

Nonparametric estimation of structural equation models is becoming increasingly popular in the literature (e.g., Newey and Powell 2003; Newey et al., 1999; Pinske, 2000; Roehrig, 1988; Su and Ullah, 2008; Vella, 1991). In this paper, we are interested in improving efficiency by imposing a full additivity constraint on each equation. Our starting point is the triangular system in Newey et al. (1999). While the assumptions of their model are relatively restrictive as compared to other examples in the literature, their estimator is typically easier to implement, which is useful for applied work. While many existing estimators allow for full flexibility, they also suffer from the curse of dimensionality.

To combat the curse, we impose an additivity constraint on each stage and propose a three step estimation procedure for our additively separable nonparametric structural equation model. We employ series estimators for our first two-stages. The first-stage involves separate (additive) regressions of each endogenous regressor on each of the exogenous regressors in order to obtain consistent estimates of the residuals. These residuals are used in our second-stage regression where we perform a single (additive) regression of our response variable on each of the endogenous regressors (not their predictions), the "included" exogenous regressors and each of the residuals from the first-stage regressions. Our final-step (one stage backfitting) involves (univariate) local-linear kernel regressions to estimate the conditional mean and gradient of each of our additive components. This process allows our final-stage estimators to be free from the curse of dimensionality. Further, our estimators have the oracle property.

We prove that our conditional mean and gradient estimates are consistent and asymptotically normal. We provide the uniform convergence rate for the additive components and their gradients. Our theoretical findings show that our final-stage estimates have an asymptotic bias and variance equivalent to a single dimension nonparametric local-linear least-square regression.

We further propose a partially linear extension of our model. We argue that the parametric components can be estimated at the parametric root-n rate and conclude that our estimates of the additive components and associated gradients remain unaffected in the asymptotic sense. Finite sample results for each of our proposed estimators are analyzed via a set of Monte Carlo simulations and support the asymptotic developments.

To showcase our estimators with empirical data, we consider a proper application relating child care use to cognitive outcomes for children (controlling for likely endogeneity). Specifically, we use the data in Bernal and Keane (2011) to examine the relationship between child test scores (our cognitive outcome) from single mothers and cumulative child care (both formal and informal). The extensive set of instrumental variables in the data set allows us to have a stronger set of instruments than what is typically used in the literature and our more flexible (partially linear) estimator leads to more insights as we can exploit the heterogeneity present both between and within groups (e.g., male versus female children).

Our empirical results show both similarities and differences from the existing literature. When we look at the average values of our estimates, we find similar results to those in Bernal and Keane (2011). On average we find mostly positive returns (to test scores) from marginal changes in income, mother's education and AFQT score. However, the mean is but one point estimate. When we check the distribution of the estimated returns, we see that the main reason behind lower returns to child care use is the *amount* of cumulative child care rather than the type. Specifically, we show that as the amount of child care use increases, additional units of child care lead to even lower returns. Bernal and Keane (2011) argue that those who use informal child care (versus formal) and girls (versus boys) receive lower returns. Our distributions of returns show no significant differences between these groups. We also find evidence of both positive and negative returns to child care use. When we analyze the characteristics of the children in each group (positive versus negative returns to child care), we find that children with negative returns are those whose mothers have higher levels education, experience and AFQT scores. Conversely, those children with positive returns typically have mother's with lower levels of education, experience and AFQT scores.

The remainder of our paper is organized as follows: Section 2 describes our methodology whereas the third section presents the asymptotic results. Section 4 considers an extension to a partially linear model and Section 5 examines the finite sample performance of our estimators via Monte Carlos simulations. The sixth section gives the empirical application and the final section concludes.

2 Methodology

In this section, we describe our estimators in detail. We propose a three-step estimation procedure which is a combination of both series and kernel methods. We provide a detailed discussion of our assumptions and theoretical results. We provide two main theorems that discuss the asymptotic properties of our second and third-step estimates. We show that the second-step bias from the series estimator consists of an asymptotically non-negligible factor from the first-stage. This terms arises because we need to use estimates from the first-step as regressors in our second-step. We further show that the estimate of the additive components in our second step is $O_p\left[\varsigma_{0\kappa}\left(\nu_n+\nu_{1n}\right)\right]$, where ν_{1n} and ν_n represent the first-step's and second-step's regression effects, respectively, and $\varsigma_{0\kappa}$ signifies the sup-norm of the basis functions used in the first two steps' estimation. In the third-stage where we turn to kernel regression via back fitting, we demonstrate that the first and second-stage regressions' effects vanish because of the extra kernel smoothing. As a result, the asymptotic properties of the conditional mean and the gradient of each additive component in the third-stage are equivalent to those of a single dimension local-linear regression. We conclude that these estimators are oracle efficient.

To proceed, we adopt the following notation. For a real matrix A, we denote its transpose as A', its Frobenius norm as $\|A\|$ $(\equiv [\mathrm{tr}(AA')]^{1/2})$, its spectral norm as $\|A\|_{\mathrm{sp}}$ $(\equiv \sqrt{\lambda_{\max}(A'A)})$, where $\mathrm{tr}(\cdot)$ is the trace operator, $\equiv$ means "is defined as" and $\lambda_{\max}(\cdot)$ denotes the largest eigenvalue of a real symmetric matrix (similarly, $\lambda_{\min}(\cdot)$ denotes the smallest eigenvalue of a real symmetric matrix). Note that the two norms are equal when A is a vector and they can be used interchangeably. For any function $q(\cdot)$ defined on the real line, we use $\dot{q}(\cdot)$ and $\ddot{q}(\cdot)$ to denote its first and second derivatives, respectively. We use $\overset{D}{\to}$ and $\overset{P}{\to}$ to denote convergence in distribution and probability, respectively.

2.1 Model

We start with the basic set-up of Newey et al. (1999). They consider a triangular system of the following form

$$\begin{cases} Y = g\left(\mathbf{X}, \mathbf{Z}_1\right) + \varepsilon, \\ \mathbf{X} = m\left(\mathbf{Z}_1, \mathbf{Z}_2\right) + \mathbf{U},\ E\left(\mathbf{U}|\mathbf{Z}_1, \mathbf{Z}_2\right) = 0,\ E\left(\varepsilon|\mathbf{Z}_1, \mathbf{Z}_2, \mathbf{U}\right) = E\left(\varepsilon|\mathbf{U}\right), \end{cases} \tag{2.1}$$

where $\mathbf{X}$ is a $d_x \times 1$ vector of endogenous regressors, $\mathbf{Z}_1 = (Z_{11}, ..., Z_{1d_1})'$ is a $d_1 \times 1$ vector of "included" exogenous regressors, $\mathbf{Z}_2 \equiv (Z_{21}, ..., Z_{2d_2})'$ is a $d_2 \times 1$ vector of "excluded" exogenous regressors, $g(\cdot, \cdot)$ denotes the true unknown structural function of interest, $m \equiv (m_1, ...m_{d_x})'$ is a $d_x \times 1$ vector of smooth functions of the instruments $\mathbf{Z}_1$ and $\mathbf{Z}_2$ and ε and $\mathbf{U} \equiv (U_1, ..., U_{d_x})'$ are error terms. Newey et al. (1999) are interested in estimating $g(\cdot, \cdot)$ consistently.

Newey et al. (1999) show that $g(\cdot, \cdot)$ can be identified up to an additive constant under the key identification conditions that $E(\mathbf{U}|\mathbf{Z}_1, \mathbf{Z}_2) = 0$ and $E(\varepsilon|\mathbf{Z}_1, \mathbf{Z}_2, \mathbf{U}) = E(\varepsilon|\mathbf{U})$. If these conditions hold, then

$$\begin{aligned} E(Y|\mathbf{X}, \mathbf{Z}_1, \mathbf{Z}_2, \mathbf{U}) &= g(\mathbf{X}, \mathbf{Z}_1) + E(\varepsilon|\mathbf{X}, \mathbf{Z}_1, \mathbf{Z}_2, \mathbf{U}) \\ &= g(\mathbf{X}, \mathbf{Z}_1) + E(\varepsilon|\mathbf{Z}_1, \mathbf{Z}_2, \mathbf{U}) \\ &= g(\mathbf{X}, \mathbf{Z}_1) + E(\varepsilon|\mathbf{U}). \qquad (2.2) \end{aligned}$$

If $\mathbf{U}$ is observed, this is a standard additive nonparametric regression model. However, in practice, $\mathbf{U}$ is not observed and it needs to be replaced by a consistent estimate. This motivates Su and Ullah (2008) to consider a three-stage procedure to obtain consistent estimates of $g(\cdot, \cdot)$ via the technique of local-polynomial regression. In the first-stage, they regress $\mathbf{X}$ on $(\mathbf{Z}_1, \mathbf{Z}_2)$ via local-polynomial regression and obtain the residuals $\widehat{\mathbf{U}}$ from this first-stage reduced-form regression. In the second-stage, they estimate $E(Y|\mathbf{X}, \mathbf{Z}, \mathbf{U})$ via another local-polynomial regression by regressing $\mathbf{Y}$ on $\mathbf{X}$, $\mathbf{Z}_1$ and $\widehat{\mathbf{U}}$. In the third-stage, they obtain the estimates of $g(\mathbf{x}, \mathbf{z}_1)$ via the method of marginal integration. Unlike previous works in the literature, including Newey et al. (1999), Pinkse (2000) and Newey and Powell (2003) that are based upon two-stage series approximations and only establish mean square and uniform convergence, they establish the asymptotic distribution for their three step local-polynomial estimator.

There are two drawbacks associated with the estimator of Su and Ullah (2008). First, it is subject to the notorious "curse of dimensionality". Without any extra restriction, the convergence rate of their second and third-stage estimators depend on $2d_x + d_1$ and $d_x + d_1$, respectively, which can be quite slow if either d_x or d_1 is not small. As a result, their estimates may perform badly even for moderately large sample sizes when $d_x + d_1 \geq 3$. Second, their estimator does not have the oracle property which an optimal estimator of the additive component in a nonparametric regression model should exhibit. In this paper we try to address both issues.

To alleviate the curse of dimensionality problem, we propose to impose some amount of structure on $g(\mathbf{X}, \mathbf{Z}_1)$, $E(\varepsilon|\mathbf{U})$ and $m_l(\mathbf{Z}_1, \mathbf{Z}_2)$, where $l = 1, ..., d_x$. Specifically, we assume that $E(\varepsilon) = 0$ and the above nonparametric objects have additive forms:

$$\begin{aligned} g(\mathbf{X}, \mathbf{Z}_1) &= \mu_g + g_1(X_1) + ... + g_{d_x}(X_{d_x}) + g_{d_x+1}(Z_{11}) + ... + g_{d_x+d_1}(Z_{1d_1}), \\ E(\varepsilon|\mathbf{U}) &= \mu_\varepsilon + g_{d_x+d_1+1}(U_1) + ... + g_{2d_x+d_1}(U_{d_x}) \text{ and} \\ m_l(\mathbf{Z}_1, \mathbf{Z}_2) &= \mu_l + m_{l,1}(Z_{11}) + ... + m_{l,d_1}(Z_{1d_1}) + m_{l,d_1+1}(Z_{21}) + ... + m_{l,d}(Z_{2d_2}), \ l = 1, ..., d_x, \end{aligned}$$

where $d = d_1 + d_2$. Consequently, we have

$$\begin{aligned} E(Y|\mathbf{X}, \mathbf{Z}_1, \mathbf{Z}_2, \mathbf{U}) &= \mu + g_1(X_1) + ... + g_{d_x}(X_{d_x}) + g_{d_x+1}(Z_{11}) + ... + g_{d_x+d_1}(Z_{1d_1}) \\ &\quad + g_{d_x+d_1+1}(U_1) + ... + g_{2d_x+d_1}(U_{d_x}) \equiv \bar{g}(\mathbf{X}, \mathbf{Z}_1, \mathbf{U}), \qquad (2.3) \end{aligned}$$

where $\mu = \mu_g + \mu_\varepsilon$. Note that the $g_j(\cdot)$'s are not fully identified without further restriction. Depending on the method that is used to estimate the additive components, different identification conditions can be imposed. For example, for the method of marginal integration, a convenient set of identification conditions would be $E\left[g_l\left(X_l\right)\right]=0$ for $l=1,...,d_x$, $E\left[g_{d_x+k}\left(Z_{1k}\right)\right]=0$ for $k=1,...,d_1$, $E\left[g_{d_x+d_1+l}\left(U_{1l}\right)\right]=0$ for $l=1,...,d_x$, $E\left[m_{l,k}\left(Z_{1k}\right)\right]=0$ for $k=1,...,d_1$ and $l=1,....,d_x$, and $E\left[m_{l,d_1+j}\left(Z_{2j}\right)\right]=0$ for $j=1,...,d_2$ and $l=1,....,d_x$.

Horowitz (2013) reviews methods for estimating nonparametric additive models, including the backfitting method, the marginal integration method, the series method and the mixture of a series method and a backfitting method to obtain oracle efficiency. It is well known that it is more difficult to study the asymptotic property of the backfitting estimator than the marginal integration estimator, but the latter has a curse of dimensionality problem if additivity is not imposed at the outset of estimation as in conventional kernel methods. Other problems that are associated with the marginal integration estimator include its lack of oracle property and its heavy computational burden. Kim et al. (1999) try to address the latter two problems by proposing a fast instrumental variable (IV) pilot estimator. However, they *cannot* avoid the curse of dimensionality problem. In fact, their IV pilot estimator depends on the estimation of the density function of the regressors at all data points. In addition, their paper ignores the notorious *boundary bias* problem for kernel density estimates and because their IV pilot estimate is not uniformly consistent on the full support, they have to use a *trimming* scheme to obtain the second-stage oracle estimator. To overcome the curse of dimensionality problem, Horowitz and Mammen (2004) propose a two-step estimation procedure with series estimation of the nonparametric additive components followed by a backfitting step that turns the series estimates into kernel estimates that are both oracle efficient and free of the curse of dimensionality.

Below we follow the lead of Horowitz and Mammen (2004) and propose a three-stage estimation procedure that is computationally efficient, oracle efficient and fully overcomes the curse of dimensionality. We shall adopt the following identification restrictions

$$\begin{aligned} g_l\left(0\right) &= g_l\left(x_l\right)\big|_{x_l=0}=0 \text{ for } l=1,...,2d_x+d_1 \text{ and} \\ m_{l,k}\left(0\right) &= 0 \text{ for } l=1,...,d_x \text{ and } k=1,2,...,d. \end{aligned}$$

Similar identification conditions are also adopted in Li (2000). The difference between our models and theirs is that their model does not allow for endogenous regressors. This complicates our problem relative to theirs as the endogeneity requires us to replace the unobserved errors in the second-stage with residuals. Hence, we need to take care of the additional bias factor from the first-step. Further, we additionally analyze the gradients of the additive nonparametric components.

2.2 Estimation

Given a random sample of n observations $\{Y_i,\mathbf{X}_i,\mathbf{Z}_{1i},\mathbf{Z}_{2i}\}_{i=1}^n$ where $\mathbf{X}_i=(X_{1i},...,X_{d_xi})'$, $\mathbf{Z}_{1i}=(Z_{11i},...,Z_{1d_1i})'$ and $\mathbf{Z}_{2i}=(Z_{21i},...,Z_{2d_2i})'$, we propose the following three-stage estimation procedure:

1. For $l=1,...,d_x$, let $\widetilde{\mu}_l$, $\{\widetilde{m}_{l,k}\left(Z_{1k,i}\right), k=1,...,d_1\}$ and $\{\widetilde{m}_{l,d_1+j}\left(Z_{2j,i}\right), j=1,...,d_2\}$, denote the series estimates of $\widetilde{\mu}_l$, $\{m_{l,k}\left(Z_{1k,i}\right), k=1,...,d_1\}$ and $\{m_{l,d_1+j}\left(Z_{2j,i}\right), j=1,...,d_2\}$ in the nonparametric additive regression

   $$X_{li}=\mu_l+m_{l,1}\left(Z_{11,i}\right)+...+m_{l,d_1}\left(Z_{1d_1,i}\right)+m_{l,d_1+1}\left(Z_{21,i}\right)+...+m_{l,d}\left(Z_{2d_2,i}\right)+U_{1i}.$$

   Let $\widetilde{U}_{li}\equiv X_{li}-\widetilde{\mu}_l-\widetilde{m}_{l,1}\left(Z_{11,i}\right)-...-\widetilde{m}_{l,d_1}\left(Z_{1d_1,i}\right)-\widetilde{m}_{l,d_1+1}\left(Z_{21,i}\right)-...-\widetilde{m}_{l,d}\left(Z_{2d_2,i}\right)$ for $l=1,...,d_x$ and $i=1,...,n$.

2. Estimate μ, $\{g_l(X_{li}), l=1,...,d_x\}$, $\{g_{d_x+j}(Z_{1j,i}), j=1,...,d_1\}$, $\{g_{d_x+d_1+k}(\widetilde{U}_{ki}), k=1,...,d_x\}$, in the following additive regression model

$$\begin{aligned} Y_i &= \mu + g_1(X_{1i}) + ... + g_{d_x}(X_{d_x i}) + g_{d_x+1}(Z_{11,i}) + ... + g_{d_x+d_1}(Z_{1d_1,i}) \\ &\quad + g_{d_x+d_1+1}(\widetilde{U}_{1i}) + ... + g_{2d_x+d_1}(\widetilde{U}_{d_x i}) + \epsilon_i \end{aligned}$$

by the series method. Here $\epsilon_i = \varepsilon_i + g_{d_x+d_1+1}(U_{1i}) + ... + g_{2d_x+d_1}(U_{d_x i}) - g_{d_x+d_1+1}(\widetilde{U}_{1i}) - ... - g_{2d_x+d_1}(\widetilde{U}_{d_x i})$ denotes the new error term. Denote the estimates as $\widetilde{\mu}$, $\{\widetilde{g}_l(X_{li}), l=1,...,d_x\}$, $\{\widetilde{g}_{d_x+j}(Z_{1j,i}), j=1,...,d_1\}$ and $\{\widetilde{g}_{d_x+d_1+k}(\widetilde{U}_{ki}), k=1,...,d_x\}$.

3. Estimate $g_1(x_1)$ and its first-order derivative by the local-linear regression of $\widetilde{Y}_{1i} = Y_i - \widetilde{\mu} - \widetilde{g}_2(X_{2i}) - ... - \widetilde{g}_{d_x}(X_{d_x i}) - \widetilde{g}_{d_x+1}(Z_{11,i}) - ... - \widetilde{g}_{d_x+d_1}(Z_{1d_1,i}) - \widetilde{g}_{d_x+d_1+1}(\widetilde{U}_{1i}) - ... - \widetilde{g}_{2d_x+d_1}(\widetilde{U}_{d_x i})$ on X_{1i}. Estimates of the other additive components in (2.3) and their first-order derivatives are obtained analogously.

In relation to Horowitz and Mammen (2004), the above first-stage is new as we have to replace the unobservable U_{li} by their consistent estimates in the second-stage. In addition, Horowitz and Mammen (2004) are only interested in estimation of the nonparametric additive components themselves, while we are also interested in estimating the first-order derivatives (gradients).

Alternatively, we could follow Kim et al. (1999) and use the kernel estimator in the first two-stages. The oracle estimator of Kim et al. (1999) has gained popularity in recent years. For example, Ozabaci and Henderson (2012) obtain the gradients of their estimator for the local-constant case and Martins-Filho and Yang (2007) consider the local-linear version of the oracle estimator, both assuming strictly exogenous regressors. However, as mentioned above, using the kernel estimators in the first two-stages here has several disadvantages and does not avoid the curse of dimensionality problem.

For notational simplicity, let $\mathbf{W} = (\mathbf{X}', \mathbf{Z}_1', \mathbf{U}')'$ and $\mathbf{w} = (\mathbf{x}', \mathbf{z}_1', \mathbf{u})'$, where, e.g., $\mathbf{u} = (u_1, ..., u_{d_x})'$ denotes a realization of $\mathbf{U}$. We shall use $\mathcal{Z} \equiv \mathcal{Z}_1 \times \mathcal{Z}_2$ and $\mathcal{W} \equiv \mathcal{X} \times \mathcal{Z}_1 \times \mathcal{U}$ to denote the support of $(\mathbf{Z}_1, \mathbf{Z}_2)$ and $\mathbf{W}$, respectively. Let $\{p_l(\cdot), l = 1, 2, ...\}$ denote a sequence of basis functions that can approximate any square-integrable function very well (to be precise later). Let $\kappa_1 = \kappa_1(n)$ and $\kappa = \kappa(n)$ be some integers such that $\kappa_1, \kappa \to \infty$ as $n \to \infty$. Let $p^{\kappa_1}(v) \equiv [p_1(v), ...p_{\kappa_1}(v)]'$. Define

$$\begin{aligned} P^{\kappa_1}(\mathbf{z}_1, \mathbf{z}_2) &\equiv \left[1,\ p^{\kappa_1}(z_{11})', ..., p^{\kappa_1}(z_{1d_1})', p^{\kappa_1}(z_{21})', ..., p^{\kappa_1}(z_{2d_2})'\right]', \\ \Phi^{\kappa}(\mathbf{w}) &\equiv \left[1,\ p^{\kappa}(x_1)', ..., p^{\kappa}(x_{d_x})', p^{\kappa}(z_{11})', ..., p^{\kappa_1}(z_{1d_1})', p^{\kappa}(u_1)', ..., p^{\kappa}(u_{d_x})'\right]'. \end{aligned}$$

For each $(\mathbf{z}_1, \mathbf{z}_2) \in \mathcal{Z}$, we approximate $m_l(\mathbf{z}_1, \mathbf{z}_2)$ and $\bar{g}(\mathbf{w})$ by $P^{\kappa_1}(\mathbf{z}_1, \mathbf{z}_2)' \boldsymbol{\alpha}_l$ and $\Phi^{\kappa}(\mathbf{w})' \boldsymbol{\beta}$, respectively, for $l = 1, ..., d_x$, where $\boldsymbol{\alpha}_l \equiv (\mu_l, \boldsymbol{\alpha}_{l,1}', ..., \boldsymbol{\alpha}_{l,d}')'$ and $\boldsymbol{\beta} = (\mu, \boldsymbol{\beta}_1', ..., \boldsymbol{\beta}_{2d_x+d_1}')'$ are $(1 + d\kappa_1) \times 1$ and $(1 + (2d_x + d_1)\kappa) \times 1$ vectors of unknown parameters to be estimated. Here, each $\boldsymbol{\alpha}_{l,k}$, $k = 1, ..., d$, is a $\kappa_1 \times 1$ vector and each $\boldsymbol{\beta}_j$, $j = 1, ..., 2d_x + d_1$, is a $\kappa \times 1$ vector. Let $\mathbb{S}_{1k}$ and $\mathbb{S}_k$ denote $\kappa_1 \times (1 + d\kappa_1)$ and $\kappa \times (1 + (2d_x + d_1)\kappa)$ selection matrices, respectively, such that $\mathbb{S}_{1k}\boldsymbol{\alpha}_l = \boldsymbol{\alpha}_{l,k}$ and $\mathbb{S}_k\boldsymbol{\beta}_l = \boldsymbol{\beta}_l$.

To obtain the first-stage estimators of the $m_l(.)$'s, let $\widetilde{\boldsymbol{\alpha}}_l \equiv (\widetilde{\mu}_l, \widetilde{\boldsymbol{\alpha}}_{l,1}', ..., \widetilde{\boldsymbol{\alpha}}_{l,d}')'$ be the solution to $\min_{\boldsymbol{\alpha}_l} n^{-1} \times \sum_{i=1}^{n} \left[X_{li} - P^{\kappa_1}\left(\mathbf{Z}_{1i}, \mathbf{Z}_{2i}\right)' \boldsymbol{\alpha}_l\right]^2$. The series estimator of $m_l\left(\mathbf{z}\right)$ is given by

$$\begin{aligned}
\widetilde{m}_l\left(\mathbf{z}_1, \mathbf{z}_2\right) &= P^{\kappa_1}\left(\mathbf{z}_1, \mathbf{z}_2\right)' \widetilde{\boldsymbol{\alpha}}_l \\
&= P^{\kappa_1}\left(\mathbf{z}_1, \mathbf{z}_2\right) \left[n^{-1} \sum_{i=1}^{n} P^{\kappa_1}\left(\mathbf{Z}_{1i}, \mathbf{Z}_{2i}\right) P^{\kappa_1}\left(\mathbf{Z}_{1i}, \mathbf{Z}_{2i}\right)'\right]^{-} n^{-1} \sum_{i=1}^{n} P^{\kappa_1}\left(\mathbf{Z}_{1i}, \mathbf{Z}_{2i}\right) X_{li} \\
&= \widetilde{\mu}_l + \sum_{k=1}^{d_1} \widetilde{m}_{l,k}\left(z_{1k}\right) + \sum_{j=1}^{d_2} \widetilde{m}_{l,d_1+j}\left(z_{2j}\right)
\end{aligned}$$

where A^{-} denotes the Moore-Penrose generalized inverse of A, $\widetilde{m}_{l,k}\left(z_{1k}\right) = p^{\kappa_1}\left(z_{1k}\right)' \widetilde{\boldsymbol{\alpha}}_{l,k}$ is a series estimator of $m_{l,k}\left(z_{1k}\right)$ for $k = 1, ..., d_1$ and $\widetilde{m}_{l,d_1+j}\left(z_{2j}\right) = p^{\kappa_1}\left(z_{2j}\right)' \widetilde{\boldsymbol{\alpha}}_{l,d_1+j}$ is a series estimator of $m_{l,d_1+j}\left(z_{2j}\right)$ for $j = 1, ..., d_2$.

To obtain the second-stage estimators of the $g_l(.)$'s, let $\widetilde{\boldsymbol{\beta}} \equiv (\widetilde{\mu}, \widetilde{\boldsymbol{\beta}}_1', ..., \widetilde{\boldsymbol{\beta}}_{2d_x+d_1}')'$ be a solution to $\min_{\boldsymbol{\beta}} n^{-1} \sum_{i=1}^{n} \left[Y_i - P^{\kappa}\left(\widetilde{\mathbf{W}}_i\right)' \boldsymbol{\beta}\right]^2$, where $\widetilde{\mathbf{W}}_i = \left(\mathbf{X}_i', \mathbf{Z}_{1i}', \widetilde{\mathbf{U}}_i'\right)'$ and $\widetilde{\mathbf{U}}_i = (\widetilde{U}_{1i}, ..., \widetilde{U}_{d_x i})'$. The series estimator of $\bar{g}\left(\mathbf{w}\right)$ is given by

$$\widetilde{\bar{g}}\left(\mathbf{w}\right) = P^{\kappa}\left(\mathbf{w}\right)' \widetilde{\boldsymbol{\beta}} = \widetilde{\mu} + \sum_{l=1}^{d_x} \widetilde{g}_l\left(x_l\right) + \sum_{k=1}^{d_1} \widetilde{g}_{d_x+k}\left(z_{1k}\right) + \sum_{j=1}^{d_x} \widetilde{g}_{d_x+d_1+k}(u_j).$$

Let $\gamma_1\left(x_1\right) \equiv \left[g_1\left(x_1\right), \dot{g}_1\left(x_1\right)\right]'$. We use $\widehat{\gamma}_1\left(x_1\right) \equiv \left[\widehat{g}_1\left(x_1\right), \widehat{\dot{g}}_1\left(x_1\right)\right]'$ to denote the local-linear estimate of $\gamma_1\left(x_1\right)$ in the third-stage by using the kernel function $K\left(\cdot\right)$ and bandwidth h. Let $\widetilde{\mathbf{Y}}_1 \equiv (\widetilde{Y}_{11}, ..., \widetilde{Y}_{1n})'$, $X_{1i}^{*}\left(x_1\right) \equiv \left[1, X_{1i} - x_1\right]$, $\mathbb{X}_1\left(x_1\right) \equiv \left[X_{11}^{*}\left(x_1\right), ..., X_{1n}^{*}\left(x_1\right)\right]'$ and $\mathbb{K}_{x_1} \equiv$diag$(K_{1x_1}, ..., K_{nx_1})$ where $K_{ix_1} \equiv K_h\left(X_{1i} - x_1\right)$ and $K_h\left(u\right) \equiv K\left(u/h\right)/h$. Then,

$$\widehat{\gamma}_1\left(x_1\right) = \left[\mathbb{X}_1\left(x_1\right)' \mathbb{K}_{x_1} \mathbb{X}_1\left(x_1\right)\right]^{-1} \mathbb{X}_1\left(x_1\right)' \mathbb{K}_{x_1} \widetilde{\mathbf{Y}}_1.$$

Below we will study the asymptotic properties of $\widehat{\beta}_1\left(x_1\right)$ via the study of the asymptotic expansion for $\widetilde{\boldsymbol{\beta}}$.

3 Asymptotic properties

In this section we first provide assumptions that are used to prove the main results and then study the asymptotic properties of the proposed estimators.

3.1 Assumptions

A real-valued function $q(\cdot)$ on the real line is said to satisfy a Hölder condition with exponent $r \in [0, 1]$ if there is c_q such that $|q(v) - q(\widetilde{v})| \leq c_q |v - \widetilde{v}|^r$ for all v and $\widetilde{v}$ on the support of $q(\cdot)$. $q(\cdot)$ is said to be γ-smooth, $\gamma = r + m$, if it is m-times continuously differentiable on $\mathcal{U}$ and its mth derivative, $\partial^m q(\cdot)$, satisfies a Hölder condition with exponent r. The γ-smooth class of functions are popular in econometrics because a γ-smooth function can be approximated well by various linear sieves; see, e.g., Chen (2007). For any scalar function $q(\cdot)$ on the real line that has r derivatives and support $\mathcal{S}$, let $\left|q(\cdot)\right|_r \equiv \max_{s \leq r} \sup_{v \in \mathcal{S}}$ $\left|\partial^s q\left(v\right)\right|$.

Let $\mathcal{X}_l$ and $\mathcal{U}_l$ denote the support of X_l and U_l, respectively, for $l = 1, ..., d_x$. Let $\mathcal{Z}_{sk}$ denote the support of Z_{sk} for $k = 1, ..., d_s$ and $s = 1, 2$. We shall use Y_i, $\mathbf{W}_i \equiv (\mathbf{X}_i', \mathbf{Z}_{1i}, \mathbf{U}_i')'$, $\mathbf{Z}_{2i}$ and U_{li} to denote the ith random observation of Y, $\mathbf{W}$, $\mathbf{Z}_2$ and U_l, respectively. Let $Q_{PP} \equiv E[P^{\kappa_1}(\mathbf{Z}_1, \mathbf{Z}_2) P^{\kappa_1}(\mathbf{Z}_1, \mathbf{Z}_2)']$, $Q_{\Phi\Phi} \equiv E[\Phi^{\kappa}(\mathbf{W}) \Phi^{\kappa}(\mathbf{W})']$ and $Q_{PP,U_l} = E[P^{\kappa_1}(\mathbf{Z}_1, \mathbf{Z}_2) P^{\kappa_1}(\mathbf{Z}_1, \mathbf{Z}_2)' U_l^2]$ for $l = 1, ..., d_x$. We make the following set of basic assumptions.

Assumption A1. (i) $\{(Y_i, \mathbf{X}_i, \mathbf{Z}_{1i}, \mathbf{Z}_{2i}), i = 1, ..., n\}$ are an IID random sample.

(ii) The supports $\mathcal{W}$ and $\mathcal{Z}$ of $\mathbf{W}_i$ and $(\mathbf{Z}_{1i}, \mathbf{Z}_{2i})$ are compact.

(iii) The distributions of $\mathbf{W}_i$ and $(\mathbf{Z}_{1i}, \mathbf{Z}_{2i})$ are absolutely continuous with respect to the Lebesgue measure.

Assumption A2.(i) For every κ_1 that is sufficiently large, there exist $\underline{c}_1$ and $\bar{c}_1$ such that $0 < \underline{c}_1 \leq \lambda_{\min}(Q_{PP}) \leq \lambda_{\max}(Q_{PP}) \leq \bar{c}_1 < \infty$ and $\lambda_{\max}(Q_{PP,U_l}) \leq \bar{c}_1 < \infty$ for $l = 1, ..., d_x$.

(ii) For every κ that is sufficiently large, there exist $\underline{c}_2$ and $\bar{c}_2$ such that $0 < \underline{c}_2 \leq \lambda_{\min}(Q_{\Phi\Phi}) \leq \lambda_{\max}(Q_{\Phi\Phi}) \leq \bar{c}_2 < \infty$.

(iii) The functions $\{m_{l,k}(\cdot),\ l = 1, ..., d, k = 1, ..., d\}$ and $\{g_j(\cdot),\ j = 2d_x + d_1\}$ belong to the class of γ-smooth functions with $\gamma \geq 2$.

(iv) There exist $\boldsymbol{\alpha}_{l,k}$'s such that $\sup_{z \in \mathcal{Z}_{1k}} |m_{l,k}(z) - p^{\kappa_1}(z)'\boldsymbol{\alpha}_{l,k}| = O(\kappa_1^{-\gamma})$ for $l = 1, ..., d_x$ and $k = 1, ..., d_1$, $\sup_{z \in \mathcal{Z}_{2k}} |m_{l,d_1+k}(z) - p^{\kappa_1}(z)'\boldsymbol{\alpha}_{l,d_1+k}| = O(\kappa_1^{-\gamma})$ for $l = 1, ..., d_x$ and $k = 1, ..., d_2$.

(v) There exist $\boldsymbol{\beta}_l$'s such that $\sup_{x \in \mathcal{X}_l} |g_l(x) - p^{\kappa}(x)'\boldsymbol{\beta}_l| = O(\kappa^{-\gamma})$ for $l = 1, ..., d_x$, $\sup_{z \in \mathcal{Z}_{1l}} |g_{d_z+k}(\cdot) - p^{\kappa}(z)'\boldsymbol{\beta}_{d_z+k}| = O(\kappa^{-\gamma})$ for $k = 1, ..., d_1$ and $\left|g_{d_x+d_1+l}(\cdot) - p^{\kappa}(\cdot)'\boldsymbol{\beta}_{d_x+d_1+l}\right|_1 = O(\kappa^{-\gamma})$ for $l = 1, ..., d_x$.

(vi) The set of basis functions, $\{p_j(\cdot), j = 1, 2, ...\}$, are twice continuously differentiable everywhere on the support of U_{li} for $l = 1, ..., d_x$. $\max_{1 \leq l \leq d_x} \max_{0 \leq s \leq r} \sup_{u_l \in \mathcal{U}_l} \|\partial^s p^{\kappa}(u_l)\| \leq \varsigma_{r\kappa}$ for $r = 0, 1, 2$.

Assumption A3. (i) The probability density functions (PDFs) of any two elements in $\mathbf{W}_i$ are bounded, bounded away from zero and twice continuously differentiable.

(ii) Let $e_i \equiv Y_i - \bar{g}(\mathbf{X}_i, \mathbf{Z}_{1i}, \mathbf{U}_i)$ and $\sigma_i^2 \equiv \sigma^2(\mathbf{X}_i, \mathbf{Z}_{1i}, \mathbf{Z}_{2i}, \mathbf{U}_i) \equiv E\left(e_i^2 | \mathbf{X}_i, \mathbf{Z}_{1i}, \mathbf{Z}_{2i}, \mathbf{U}_i\right)$. Let $Q_{sk,pp} \equiv E[p^{\kappa_1}(Z_{sk,i}) p^{\kappa_1}(Z_{sk,i})' \sigma_i^2]$ for $k = 1, ..., d_s$ and $s = 1, 2$. The largest eigenvalue of $Q_{sk,pp}$ is bounded uniformly in κ_1.

Assumption A4. The kernel function $K(\cdot)$ is a PDF that is symmetric, bounded and has compact support $[-c_K, c_K]$. It satisfies the Lipschitz condition $|K(v_1) - K(v_2)| \leq C_K |v_1 - v_2|$ for all $v_1, v_2 \in [-c_K, c_K]$.

Assumption A5.

(i) $\kappa_1 \leq \kappa$. As $n \to \infty$, $\kappa_1 \to \infty$, $\kappa^3/n \to 0$ and $\tau_n \to c_1 \in [0, \infty)$, where $\tau_n \equiv \left(\kappa^{1/2}\varsigma_{0\kappa} + \varsigma_{1\kappa}\right)\nu_{1n} + \varsigma_{0\kappa}\varsigma_{2\kappa}\nu_{1n}^2$, $\nu_{1n} \equiv \kappa_1^{1/2}/n^{1/2} + \kappa_1^{-\gamma}$ and $\nu_n \equiv \kappa^{1/2}/n^{1/2} + \kappa^{-\gamma}$.

(ii) As $n \to \infty$, $h \to 0$, $nh^3 \log n \to \infty$, $nh\kappa^{-2\gamma} \to 0$, $\tau_n \nu_{1n} = o(n^{-1/2}h^{-1/2})$ and $[h^{1/2}\varsigma_{1\kappa}(1 + n^{1/2}\kappa_1^{-\gamma}) + \varsigma_{2\kappa} n^{1/2} h^{1/2} \nu_{1n}^2](\nu_n + \nu_{1n}) \to 0$.

Assumptions A1(i)-(ii) impose IID sampling and compactness on the support of the exogenous independent variables. Either assumption can be relaxed at lengthy arguments (see, e.g., Su and Jin, 2012 who allow for both weakly dependent data and infinite support for their regressors. A1(iii) requires that the variables in $\mathbf{W}_i$ and $(\mathbf{Z}_{1i}, \mathbf{Z}_{2i})$ be continuously valued, which is standard in the literature on sieve estimation. The extension to allow for both continuous and discrete variables is possible but will not be pursued in this paper.

Assumption A2(i)-(ii) ensure the existence and non-singularity of the covariance matrix of the asymptotic form of the first two-stage estimators. They are standard in the literature; see, e.g., Newey (1997), Li (2000) and Horowitz and Mammen (2004). Note that all of these authors assume that the conditional variances of the error terms given the exogenous regressors are uniformly bounded, in which case the second part of A1(i) becomes redundant. A2(iii) imposes smoothness conditions on the relevant functions and A2(iv)-(v) quantifies the approximation error for γ-smooth functions. These conditions are satisfied, for example, for polynomials, splines and wavelets. A2(vi) is needed for the application of Taylor expansions. It is well known that $\varsigma_{r\kappa} = O\left(\kappa^{r+1/2}\right)$ and $O\left(\kappa^{2r+1}\right)$ for B-splines and power series, respectively (see Newey, 1997). The rate at which splines uniformly approximate a function is the same as that for power series, so that the uniform convergence rate for splines is faster than power series. In addition, the low multicollinearity of B-splines and recursive formula for calculation also leads to computational advantages (see Chapter 19 of Powell, 1981 and Chapter 4 of Schumaker, 2007). For these reasons, B-splines are widely used in the literature. We will use them in our simulations and application as well.

Assumptions A3(i)-(ii) and A4 are needed for the establishment of the asymptotic property of the third-stage estimators. A3(ii) is redundant under Assumption A2(i) if one assumes that the conditional variances of e_i's given $(\mathbf{X}_i, \mathbf{Z}_{1i}, \mathbf{Z}_{2i}, \mathbf{U}_i)$ are uniformly bounded. A4 is standard for local-linear regression (see Fan and Gijbels, 1996 and Masry, 1996). The compact support condition is convenient for the demonstration of the uniform convergence rate in Theorem 3.2 below. It can be removed at the cost of some lengthy arguments (see, e.g., Hansen, 2008). In particular, the Gaussian kernel can be applied. Assumptions A5(i)-(ii) specify conditions on κ_1, κ and h. Note that we allow the use of different series approximation terms in the first and second-stage estimation, which allows us to see the effect of the first-stage estimates on the second-stage estimates. The first condition (namely, $\kappa_1 \leq \kappa$) in A5(i) is needed for the proof of a technical lemma (see Lemma A.5(iii)) in the appendix and it can be removed at the cost of some additional assumptions on the basis functions. The terms that are associated with ν_{1n} arise because of the use of the nonparametrically generated regressors in the second-stage series estimation. The appearance of $\log n$ arises in order to establish uniform consistency results in Theorem 3.2 below and it can be replaced by 1 if we are only interested in the pointwise result. In the case where $\varsigma_{r\kappa} = O\left(\kappa^{r+1/2}\right)$ in Assumption A2(vi), $\tau_n = O\left(\kappa^{3/2}\nu_{1n} + \kappa^3\nu_{1n}^2\right)$. In practice, we recommend setting $\kappa_1 = \kappa$. These restrictions, in conjunction with the condition $\gamma \geq 2$, imply that the conditions in Assumption A5 can be greatly simplified as follows:

Assumption A5*.

(i) As $n \to \infty$, $\kappa \to \infty$, $\kappa^4/n \to c_1 \in [0, \infty)$.

(ii) As $n \to \infty$, $h \to 0$, $nh^3 \log n \to \infty$, $nh\kappa^{-2\gamma} \to 0$ and $n^{-1}h\kappa^5 \to 0$.

3.2 Theorems

In this section we state two theorems that give the main results of the paper. Even though several results are available in the literature on nonparametric or semiparametric regressions with nonparametrically generated regressors (see, e.g., Mammen et al., 2012 and Hahn and Ridder, 2013 for recent contributions), none of them can be directly applied to our framework. In particular, Hahn and Ridder (2013) study the asymptotic distribution of three-step estimators of a *finite-dimensional* parameter vector where the second-step consists of one or more nonparametric generated regressions on a regressor that is estimated in the first-step. In sharp contrast, our third-stage estimator is also a nonparametric estimator. Under fairly general conditions, Mammen et al. (2012) focus on two-stage nonparametric regression where the first-stage can be kernel or series estimation while the second-stage is local-linear estimation. In principle, we can treat our second and third-stage estimation as their first and second-stage estimation, respectively and then apply their results to our case. However, their results are built upon high-level assumptions and are usually not optimal. For this reason, we derive the asymptotic properties of our three-stage estimators under some primitive conditions specified in the preceding section.

The asymptotic properties of the second-stage series estimator $\widetilde{\boldsymbol{\beta}}$ are reported in the following theorem.

Theorem 3.1 *Suppose that Assumptions A.1-A.5(i) hold. Then*

(i) $\widetilde{\boldsymbol{\beta}}-\boldsymbol{\beta}=Q_{\Phi\Phi}^{-1}n^{-1}\sum_{i=1}^{n}\Phi_i e_i+Q_{\Phi\Phi}^{-1}n^{-1}\sum_{i=1}^{n}\Phi_i\left[\bar{g}\left(X_i,Z_{1i},U_i\right)-\Phi_i'\boldsymbol{\beta}\right]-Q_{\Phi\Phi}^{-1}n^{-1}\sum_{i=1}^{n}\Phi_i\sum_{l=1}^{d_x}\dot{g}_{d_x+d_1+l}\left(U_{li}\right)\left(\widetilde{U}_{li}-U_{li}\right)+\mathbf{R}_{n,\beta}$;

(ii) $\left\|\widetilde{\boldsymbol{\beta}}-\boldsymbol{\beta}\right\|=O_P\left(\nu_n+\nu_{1n}\right)$;

(iii) $\sup_{\mathbf{w}\in\mathcal{W}}\left|\widetilde{\bar{g}}\left(\mathbf{w}\right)-\bar{g}\left(\mathbf{w}\right)\right|=O_P\left[\varsigma_{0\kappa}\left(\nu_n+\nu_{1n}\right)\right]$;

where $\left\|\mathbf{R}_{n,\beta}\right\|=\tau_n O_P\left(\nu_n+\nu_{1n}\right)$ *and* ν_{1n}, ν_n *and* τ_n *are defined in Assumption A.5*(i).

To appreciate the effect of the first-stage series estimation on the second-stage series estimation, let $\bar{\boldsymbol{\beta}}$ denote a series estimator of $\boldsymbol{\beta}$ by using $\mathbf{U}_i$ together as $(\mathbf{X}_i,\mathbf{Z}_{1i})$ as the regressors. Then it is standard to show that

$$\begin{aligned}\bar{\boldsymbol{\beta}}-\boldsymbol{\beta} &= Q_{\Phi\Phi}^{-1}n^{-1}\sum_{i=1}^{n}\Phi_i e_i+Q_{\Phi\Phi}^{-1}n^{-1}\sum_{i=1}^{n}\Phi_i\left[\bar{g}\left(X_i,Z_{1i},U_i\right)-\Phi_i'\boldsymbol{\beta}\right]+\bar{\mathbf{R}}_{n,\beta}\text{ and}\\ \left\|\bar{\boldsymbol{\beta}}-\boldsymbol{\beta}\right\| &= O_P\left(\nu_n\right)\end{aligned}$$

where $\left\|\bar{\mathbf{R}}_{n,\beta}\right\|=O_P(\kappa n^{-1/2}\nu_n)=o\left(\nu_n\right)$. The third term on the right hand side of the expression in Theorem 3.1(i) signifies the asymptotically non-negligible dominant effect of the first-stage estimation on the second-stage estimation.

With Theorem 3.1, it is straightforward to show the asymptotic distribution of our three-stage estimator of $g_1\left(x_1\right)$ and its gradient.

Theorem 3.2 *Let* $H\equiv$*diag*$(1,h)$. *Suppose that Assumptions A.1-A.5 hold. Then*

(i) (Normality)

$$\sqrt{nh}H\left[\widehat{\gamma}_1\left(x_1\right)-\gamma_1\left(x_1\right)-b_1\left(x_1\right)\right]\overset{D}{\to}N\left(0,\Omega_1\left(x_1\right)\right),$$

where $b_1\left(x_1\right)\equiv\begin{pmatrix}\frac{\upsilon_{21}}{2}h^2\ddot{g}_1\left(x_1\right)\\ 0\end{pmatrix}$, $\Omega_1\left(x_1\right)\equiv\begin{pmatrix}\sigma^2(x_1)/f_{X_1}\left(x_1\right) & 0\\ 0 & \upsilon_{22}\sigma^2(x_1)/\left[\upsilon_{21}^2 f_{X_1}\left(x_1\right)\right]\end{pmatrix}$, $\sigma^2(x_1)$ $\equiv E\left(e_i^2|X_{11,i}=x_1\right)$ *and* $\upsilon_{ij}\equiv\int v^i k(v)^j dv,\ i,\ j=0,1,2.$

(ii) (Uniform consistency) Suppose that $Q_{\Phi\Phi,e}\equiv E\left(\Phi_i\Phi_i' e_i^2\right)$ *has bounded maximum eigenvalue. Then*

$$\sup_{x_1\in\mathcal{X}_1}\left\|H\left[\widehat{\gamma}_1\left(x_1\right)-\gamma_1\left(x_1\right)\right]\right\|=O_P\left(\left(nh\log n\right)^{-1/2}+h^2\right).$$

Theorem 3.2(i) indicates that our three-step estimator of $\gamma_1\left(x_1\right)=\left[g_1\left(x_1\right),\dot{g}_1\left(x_1\right)\right]'$ has the asymptotic oracle property. Asymptotically, the asymptotic distribution of local-linear estimator of $\gamma_1\left(x_1\right)$ is not affected by random sampling errors in the first two-stage estimators. In fact, the three-step estimator of $\gamma_1\left(x_1\right)$ has the same asymptotic distribution that we would have if the other components in $\bar{g}\left(x,z_1,u\right)$

were known and a local-linear procedure is used to estimate $\gamma_1(x_1)$. Theorem 3.2(ii) gives the uniform convergence rate for $\widehat{\gamma}_1(x_1)$. Similar properties can be established for the local-linear estimators of other components of $\bar{g}(x, z_1, u)$. In addition, following the standard exercise in the nonparametric kernel literature, we can also demonstrate that these estimators are asymptotically independently distributed.

4 Partially linear additive models

In this section we consider a slight extension of the model in (2.1) to the following partially linear functional coefficient model

$$\begin{cases} Y = g(\mathbf{X}, \mathbf{Z}_1) + \theta'\mathbf{V} + \varepsilon, \\ \mathbf{X} = m(\mathbf{Z}_1, \mathbf{Z}_2) + \Psi\mathbf{V} + \mathbf{U},\ E(\mathbf{U}|\mathbf{Z}_1, \mathbf{Z}_2, \mathbf{V}) = 0,\ E(\varepsilon|\mathbf{Z}_1, \mathbf{Z}_2, \mathbf{U}, \mathbf{V}) = E(\varepsilon|\mathbf{U}), E(\varepsilon) = 0, \end{cases} \tag{4.1}$$

where Y, $\mathbf{X}$, $\mathbf{Z}_1$, $\mathbf{Z}_2$, $\mathbf{Z}$, and ε are defined as above, $\mathbf{V}$ is a $k \times 1$ vector of exogenous variables, θ is a $k \times 1$ parameter vector and $\Psi = [\psi_1', ..., \psi_{d_x}']'$ is a $d_x \times k$ matrix of parameters in the reduced form regression for $\mathbf{X}$. To avoid the curse of dimensionality, we continue to assume that $m(\mathbf{Z}_1, \mathbf{Z}_2)$, $g(\mathbf{X}, \mathbf{Z}_1)$ and $E(\varepsilon|\mathbf{U})$ have the additive forms given in Section 2.1.

We remark that the results developed in previous sections extend straightforwardly to the model specified in (4.1). Note that

$$\begin{aligned} E(Y|\mathbf{X}, \mathbf{Z}_1, \mathbf{Z}_2, \mathbf{U}, \mathbf{V}) &= g(\mathbf{X}, \mathbf{Z}_1) + E(\varepsilon|\mathbf{U}) + \theta'\mathbf{V} = \bar{g}(\mathbf{X}, \mathbf{Z_1}, \mathbf{U}) + \boldsymbol{\theta}'\mathbf{V} \text{ and} && (4.2) \\ E(\mathbf{X}|\mathbf{Z}_1, \mathbf{Z}_2, \mathbf{V}) &= m(\mathbf{Z}_1, \mathbf{Z}_2) + \Psi\mathbf{V}. && (4.3) \end{aligned}$$

Given a random sample $\{(Y_i, \mathbf{X}_i, \mathbf{Z}_{1i}, \mathbf{Z}_{2i}, \mathbf{V}_i), i = 1, ..., n\}$, we can continue to adopt the three-step procedure outlined in Section 2.2 to estimate the above model. First, we choose $(\boldsymbol{\alpha}_l,\ \psi_l)$ to minimize $n^{-1}\sum_{i=1}^n \left[X_{li} - P^{\kappa_1}(\mathbf{Z}_{1i}, \mathbf{Z}_{2i})'\boldsymbol{\alpha}_l - \mathbf{V}_i'\psi_l\right]^2$. Let $(\widetilde{\boldsymbol{\alpha}}_l, \widetilde{\psi}_l)$ denote the solution. The series estimator of $m_l(\mathbf{z}_1, \mathbf{z}_2)$ is given by $\widetilde{m}_l(\mathbf{z}_1, \mathbf{z}_2) = P^{\kappa_1}(\mathbf{z}_1, \mathbf{z}_2)'\widetilde{\boldsymbol{\alpha}}_l$. Define the residuals $\widetilde{U}_{li} = X_{li} - \widetilde{m}_l(\mathbf{Z}_{1i}, \mathbf{Z}_{2i}) - \widetilde{\psi}_l'\mathbf{V}_i$. Let $\widetilde{\mathbf{U}}_i = (\widetilde{U}_{1i}, ..., \widetilde{U}_{d_x i})'$, $\widetilde{\mathbf{W}}_i = \left(\mathbf{X}_i', \mathbf{Z}_{1i}', \widetilde{\mathbf{U}}_i'\right)'$ and $P^{\kappa}\left(\widetilde{\mathbf{W}}_i\right)$ be defined as before. Second, we choose $(\boldsymbol{\beta},\ \theta)$ to minimize $n^{-1}\sum_{i=1}^n \left[Y_i - P^{\kappa}\left(\widetilde{\mathbf{W}}_i\right)'\boldsymbol{\beta} - \mathbf{V}_i'\theta\right]^2$. Let $\widetilde{\boldsymbol{\beta}} \equiv (\widetilde{\mu}, \widetilde{\boldsymbol{\beta}}_1', ..., \widetilde{\boldsymbol{\beta}}_{2d_x+d_1}')'$ and $\widetilde{\theta}$ denote the solution. Define $\widetilde{Y}_{1i} = Y_i - \widetilde{\boldsymbol{\beta}}_{(-1)}'P^{\kappa}\left(\widetilde{\mathbf{W}}_i\right) - \widetilde{\theta}'\mathbf{V}_i$ where $\widetilde{\boldsymbol{\beta}}_{(-1)}$ is defined as $\widetilde{\boldsymbol{\beta}}$ with its component $\widetilde{\boldsymbol{\beta}}_1$ being replaced by a $\kappa \times 1$ vector of zeros. Third, we estimate $g_1(x_1)$ and its first-order derivative by regressing $\widetilde{Y}_{1i}$ on X_{1i} via the local-linear procedure. Let $\widehat{\gamma}_1(x_1)$ denote the estimate of $\gamma_1(x_1)$ via local-linear fitting.

It is well known that the finite dimensional parameter vectors ψ_l's and θ can be estimated at the parametric $\sqrt{n}$-rate and the appearance of the linear components in (4.1) won't affect the asymptotic properties of $\widetilde{\boldsymbol{\beta}}$ and $\widehat{\gamma}_1(x_1)$. To conserve space, we do not repeat the arguments here.

5 Finite sample properties

In this section we discuss the finite sample properties of our estimators using a set of Monte Carlo simulations. We report average bias, variance and root mean square error (RMSE) for both the final-stage conditional mean and gradients. We perform 1000 Monte Carlo simulations for three different sample sizes: 100, 200 and 400. We consider four different data generating processes (DGPs) of structural equations between Y, X, Z (once for V), ε and u. Unless we state otherwise, Z_1 and Z_2 are independently distributed $U[0,1]$ and ε and u are independently distributed as $N(0,1)$ and are mutually independent of one another and of X and (Z_1, Z_2). For the first three DGPs, each error distribution is assumed to be homoskedastic.

Our first DGP is our baseline model and is given as

$$\begin{aligned} Y &= \sin(X) + \sin(Z_1) + \varepsilon \\ X &= \sin(Z_1) + \sin(Z_2) + u. \end{aligned}$$

Our second DGP considers a slightly more complicated first-stage as Z_2 enters the reduced form of X via the PDF of a logistic distribution

$$\begin{aligned} Y &= 0.25X^2 + 0.5Z_1^2 + \varepsilon \\ X &= 0.20Z_1^2 + \frac{e^{-Z_2}}{(1+e^{-Z_2})^2} + u. \end{aligned}$$

Our third DGP considers the partially linear extension where V_1 and V_2 are distributed via a binomial and Gaussian distribution, respectively

$$\begin{aligned} Y &= \sin(X) + \sin(Z_1) + 0.5V_1 + V_2 + \varepsilon \\ X &= \sin(Z_1) + \sin(Z_2) + u. \end{aligned}$$

Finally, our fourth DGP is similar to the first, but allows for heteroskedasticity (note that our theory allows for heteroskedasticity). Specifically, we allow the variance of ε to be a function of Z_1 and Z_2 via $0.1 + 0.5Z_1^2 + 0.5Z_2^2$. All other variables are generated as above. For completeness, we restate the DGP as

$$\begin{aligned} Y &= \sin(X) + \sin(Z_1) + \varepsilon \\ X &= \sin(Z_1) + \sin(Z_2) + u. \end{aligned}$$

We estimate each of the DGPs in three steps. In the first two steps we use smoothing splines and in the third step we use kernel regression. For the spline estimates, (in order to minimize computing time) we use a fixed number (six) of knots. Using the same logic, for our basis functions, we fix the degree at three (i.e., we use cubic splines). In our final step (kernel regression), we employ a local-linear estimator, use a rule-of-thumb bandwidth and a Gaussian kernel (in our application we will use cross-validation methods in each step). The results of our simulations are discussed below.

All of the simulation results for the final-stage regressions can be found in Table 1. Each of the results are as expected. The average RMSE of each estimator decreases with the sample size. The conditional mean is estimated more precisely than its gradients. The estimators in the homoskedastic DGP outperform those from the heteroskedastic DGP (DGP 1 versus 4). We are confident in the performance of our estimators and now turn our attention to empirical data.

6 Application: Child care use and test scores

It is generally accepted in the literature that early childhood achievement is a strong predictor for success (better labor market outcomes) later in life (Keane and Wolpin, 1997, 2001; Bernal and Keane, 2011;

Table 1: Monte Carlo simulations for the final-stage conditional mean and gradient estimates. Each element in the table is the average of the reported criteria over the 1000 runs.

	$\widehat{g}(\cdot,\cdot)$	$\frac{\partial\widehat{g}(\cdot,\cdot)}{\partial x}$	$\frac{\partial\widehat{g}(\cdot,\cdot)}{\partial z_1}$	$\widehat{g}(\cdot,\cdot)$	$\frac{\partial\widehat{g}(\cdot,\cdot)}{\partial x}$	$\frac{\partial\widehat{g}(\cdot,\cdot)}{\partial z_1}$	$\widehat{g}(\cdot,\cdot)$	$\frac{\partial\widehat{g}(\cdot,\cdot)}{\partial x}$	$\frac{\partial\widehat{g}(\cdot,\cdot)}{\partial z_1}$
		$n = 100$			$n = 200$			$n = 400$	
Bias									
DGP 1	0.0522	-0.0611	0.2847	0.0419	-0.0586	0.0092	0.0297	-0.0358	-0.0078
DGP 2	-0.0649	0.0525	-0.0346	-0.0621	0.0178	-0.0483	-0.0434	0.0577	-0.0013
DGP 3	0.0526	-0.0544	-0.0019	0.0371	-0.0575	-0.0024	0.0278	-0.0502	-0.0015
DGP 4	0.0515	-0.0341	0.0542	0.0412	-0.0556	0.0166	0.0285	-0.0347	-0.0138
Variance									
DGP 1	0.0526	0.0879	0.3017	0.0316	0.0593	0.1672	0.0195	0.0419	0.1120
DGP 2	0.1234	0.2485	0.7299	0.1211	0.2352	0.6851	0.0508	0.2242	0.3233
DGP 3	0.1170	0.1530	0.6956	0.0703	0.1054	0.4047	0.0443	0.0686	0.2732
DGP 4	0.1182	0.1520	0.6670	0.0692	0.1016	0.3773	0.0436	0.0714	0.2562
RMSE									
DGP 1	0.2468	0.3699	0.6608	0.1870	0.2938	0.4880	0.1465	0.2405	0.3785
DGP 2	0.3708	0.6714	0.9763	0.3701	0.6684	0.9310	0.2365	0.6382	0.6197
DGP 3	0.3589	0.4947	0.9839	0.2779	0.4143	0.7609	0.2199	0.3235	0.5865
DGP 4	0.3659	0.5004	0.9740	0.2755	0.3974	0.7161	0.2163	0.3171	0.5669

Cameron and Heckman, 1998). Thus, researchers have focused on the determinants of childhood achievement. Various models have been developed and most focus on cognitive ability as the outcome measure. In the present context, we are concerned whether or not child care improves or hurts a particular measure of cognitive ability, test scores. Although this is an interesting question, previously there were serious data limitations.

The two major limitations associated with cognitive ability production functions in this context are sample selection bias and endogeneity. Sample selection bias occurs when only mothers' labor force participation is used in the analysis. This variable implicitly assumes that it is a direct indicator of child care use. The main problem here is that working mothers and non-working mothers may differ substantially in the cognitive ability production process and if only labor force participation is used, the analysis is going to rule out "non-working" mothers. Adding actual child care use can help take care of the selectivity problem (see Bernal, 2008 and Bernal and Keane, 2011).

The second issue is potential endogeneity of the child care use variable. To the best of our knowledge, there are relatively few papers in this literature that use instrumental variable estimation to solve the endogeneity problem. Two possible reasons for this are the use of restrictive methods (those that likely hide the existing heterogeneity of mothers hinder the sources of potential endogeneity) and data limitations. The three papers that we are aware of which use IV regressions are Blau and Grossberg (1992), James-Burdumy (2005) and Bernal and Keane (2011).

Blau and Grossberg (1992) use maternal labor supply as an indicator of child care use and analyze children's cognitive development. They define endogeneity via the participation decision of mothers. They define it as a comparison between in home and market production. They state that the employed and unemployed mothers' differences may create differences in child quality production. Hence they focus on the endogeneity of the mothers. They instrument for maternal labor supply and conclude that there is no statistical heterogeneity between employed and unemployed mothers (and reject the IV model). An issue with their paper (which they point out), is the possibility of weak instruments. They also do not have a detailed control for child care. James-Burgundy (2005) focuses on the same problem and uses labor market conditions for her fixed effects IV model. Potentially weak instruments are again blamed for rejection of the IV model.

In response to the issues mentioned above, Bernal and Keane (2011) obtain data on actual child care use (which helps correct for selectivity issues) as well as an extensive number of instrumental variables (which helps correct for the weak instrumental variable issue). Further, they choose a larger age range (compared to existing studies) for children in their application (previous studies found stronger correlations for their target ages). Also, they focus only on single mothers which arguably fits their set of instruments better. They conclude that their IV regression perform well.

We start our analysis with Bernal and Keane's (2011) data, but consider a more flexible cognitive ability production function ($g(\cdot)$). Consider the following cognitive ability production function

$$test = g(age, care, inc, nmchild, char, iabil) + \varepsilon \tag{6.1}$$

where $test$ is the logarithm of child test scores (our measure of cognitive ability), age represents the child's age, $care$ (the primary variable of interest) is cumulative child care, inc is the logarithm of cumulative income since childbirth, $char$ is a vector of group characteristics of the child and the mother (e.g., mother's AFQT score), $nmchild$ is the number of children and $iabil$ measures initial ability (e.g., birth weight). Equation (6.1) is the baseline cognitive ability production equation (minus any functional form assumptions) in Bernal and Keane (2011, pp. 474).

A primary concern of Equation (6.1) is that $care$, inc and $nmchild$ may be correlated with the error term. Hence we use instruments to correct for this potential endogeneity. Following Bernal and Keane (2011), we use local demand conditions and welfare rules as instruments.

Our contribution here is to provide a more flexible version of the cognitive ability production function proposed by Bernal and Keane (2011). This allows us to obtain the effects of child care for each child. Standard least-squares estimation methods are best suited to data near the mean. Looking solely at the mean may be misleading. Further, it is arguable that we are more interested in the upper or lower tails of the distribution of returns to child care. Our approach allows us to observe the overall variation.

Here we will be using the partially linear additive nonparametric specification. In our first-stage we estimate three separate regressions: one for each endogenous regressor ($care$, inc and $nmchild$). In Equation (4.1), these are given as the regressions of X on Z_1, Z_2 and V. Specifically, our first-stage equations are written as

$$\begin{aligned} care &= m_{c_1}(mafqt) + m_{c_2}(med) + m_{c_3}(mex) + m_{c_4}(mage) + \theta_c'\mathbf{V} + u_c \\ inc &= m_{i_1}(mafqt) + m_{i_2}(med) + m_{i_3}(mex) + m_{i_4}(mage) + \theta_i'\mathbf{V} + u_i \\ nmchild &= m_{n_1}(mafqt) + m_{n_2}(med) + m_{n_3}(mex) + m_{n_4}(mage) + \theta_n'\mathbf{V} + u_n \end{aligned} \tag{6.2}$$

where we allow the control variables of mother's AFQT score ($mafqt$), mother's education (med), mother's experience (mex) and mother's age ($mage$) to enter nonlinearly. The remaining control variables as well as each of the instruments are contained in $\mathbf{V}$. Note that $\mathbf{V}$ includes interactions for each instrument with $mafqt$ and med. This results in a total of 99 regressors in each first-stage regression (clearly indicating the need for a partially linear model).

After obtaining consistent estimates of each of the residual vectors from Equation (6.2), we run the second-stage model via a nonparametric additively separable partially linear regression of log test scores ($test$) on the endogenous variables (not their predicted values), each of the residuals from the first-stage and the remaining control variables as

$$\begin{aligned} test &= g_1(mafqt) + g_2(med) + g_3(mex) + g_4(mage) \\ &\quad + g_5(care) + g_6(inc) + g_7(nmchild) \\ &\quad + g_8(\widehat{u}_c) + g_9(\widehat{u}_i) + g_{10}(\widehat{u}_n) \\ &\quad + \Psi\widetilde{\mathbf{V}} + \epsilon, \end{aligned} \tag{6.3}$$

where $\widetilde{\mathbf{V}}$ is the same (twenty) control variables included in $\mathbf{V}$ (and does not include any of the instruments Z_2). Estimation (of the additive components and their gradients) in the final-stage follows from Section 2.2.

Regarding implementation, note that in the first two stages we use cross-validation techniques to choose the number of knots for our B-splines. In our final-stage estimates we use cross-validation techniques to determine the bandwidth in our kernel function. R code which can be used to estimate our model is available from the authors upon request.

6.1 Data

Our data comes directly from Bernal and Keane (2011). The data are extensive and we will attempt to summarize them in a concise manner. For those interested in the specific details, we refer you to the excellent description in the aforementioned paper. Their primary data source is the National Longitudinal Survey of Youth of 1979 (NLSY79). The exact instruments and control variables can be found in Tables 1 and 2 in Bernal and Keane (2011).

As noted in the introduction, the data set consists of single mothers. Although this may seem like a data restriction at first, it leads to stronger instruments. The main reason behind this choice, as explained by Bernal and Keane (2011), is that single mothers fit their set of instruments better. The primary instruments used here are welfare rules, which (as claimed by Bernal and Keane, 2011) give exogenous variation for single mothers. The (1990s) welfare policy changes resulted in increased employment rates for single mothers, hence higher child care use. We describe our variables for the 2454 observations in our sample in more detail below.

6.1.1 Endogenous regressors

We consider three potentially endogenous variables (X): cumulative child care, cumulative income and number of children. These are the left-hand-side variables in our first-stage equations. They are modeled in an additively separable nonparametric fashion in the second-stage regression.

6.1.2 Instruments

We group our instruments (Z_2) into four categories: time limits, work requirements, earning disregards, and other policy variables and local demand conditions. We briefly explain each of these categories and refer the reader to a more in depth description of the instruments in Bernal and Keane (2011, pp. 466-469).

Time limits We consider (time limits for) two programs which aid in helping families with children: Aid to Families with Dependent Children (AFDC) and Waivers and Temporary Aid for Needy Families (TANF). Under AFDC, single mothers with children under age 18 may be eligible to get help from the program as long as they fit certain criteria that are typically set by the state and program regulations. TANF on the other hand, enables the states to set certain time limits on the benefits they provide to eligible individuals. AFDC provides the benefits and TANF creates the variability because states can set their own limits. The limits are important for benefit receivers because an eligible female may become ineligible by hitting the limit and she may choose to save some of the eligibility for later use. We include each of the eight (time limit) instruments proposed by Bernal and Keane (2011).

Work requirements TANF requires eligible females to return to work after a certain time, as set by the state, to be able remain eligible. These rules are state dependent. While the main required length for females is to start working within two years, several states prefer to choose shorter time limits. Some states lift this requirement for females with young children. Besides the variation amongst states, even within states there exists variation. Here we include each of the nine (work requirement) instruments.

Earning disregards The AFDC and TANF benefits are adjusted by states depending upon the number of children and earnings of the eligible females. While more children may lead to greater benefits, more earnings may lower them. States set the level for AFDC grants and adjust the amount of reduction in benefits via TANF. Specifically, our first-stage regressions include both the "flat amount of earnings disregarded in calculating the benefit amount" and the "benefit reduction rate" instrumental variables.

Other policy variables and local demand conditions Our remaining instruments are grouped in one generic category: other policy variables and local demand conditions. Here we consider two additional programs for families with young children. These programs are Child Support Enforcement (CSE) and the Child Care Development Fund (CCDF). Bernal and Keane (2011) report CSE as a significant source of income for single mothers via the 2002 Current Population Survey. CSE's goal is to find absent parents and establish relationships with their children. CCDF on the other hand, is a subsidy based program which provides child care for low-income families. States are independent in designing their own programs and hence variation is present.

In addition to the policy variables, earned income tax credit (EITC), the unemployment rate and hourly wage rate are listed as instruments. EITC is a wage support program for low-income families. This is a subsidy based program and the subsidy amount varies with family size. The benefit levels are not conditioned on actual family size since family size is endogenous. Our first-stage regressions include six instruments from this category.

6.1.3 Control variables

In addition to the instrumental variables, we have twenty-four control variables (Z_1) which show up in each stage (four nonlinearly and twenty linearly). These variables primarily represent characteristics of the mothers and children. These are each assumed to be exogenous. The four variables that enter nonlinearly are the mother's AFQT score, education level, work experience and age. We treat these nonparametrically as we believe their impacts vary with their levels and we are unaware of an economic theory which states exactly how they should enter. In order to understand the intuition behind this specification, consider a model where these variables enter linearly. In that case, having linear schooling in a model would imply that each additional year of schooling a mother gets will lead to the same percentage change on the child's test score. The same will hold true for a linearly modeled AFQT score, age or experience. There is no reason to assume that this will be true.

6.2 Results

There are a vast number of results that can be presented from this procedure and data. We plan to limit our discussion to a few important issues. First, we want to determine the strength of our instruments. To determine how well the instruments predict the endogenous regressors, we propose a F-type (wild) bootstrap based test for joint significance of the instruments. Second, we are interested in whether or not endogeneity exists. We check for this by testing for joint significance of $g_8(\cdot)$, $g_9(\cdot)$ and $g_{10}(\cdot)$ in Equation (6.3). Finally, we are interested in potential heterogeneity in the return to child care use. We accomplish this by separating our observation specific estimates up amongst different pre-specified groups.

Our final-stage results we be given in three separate tables. The first two tables will give the 10th, 25th, 50th, 75th and 90th percentiles of the estimated returns and their related (wild) bootstrapped standard errors. The first of those tables (Table 2) will look at the returns to (gradients of) the final stage estimates of each of the (excluding the residuals) nonlinear variables from the second-stage regressions. The remaining tables (Tables 3 and 4) will decompose the gradients for the child care use variable.

We also provide several figures of estimated densities and distribution functions. Figures 1-3 give a set of density plots for the estimated returns to child care use. This allows us to see the overall distribution. We believe this is a more informative type of analysis as compared to solely showing results at the mean or percentiles. Figure 4 looks at empirical cumulative distribution functions (ECDF) for both positive and negative gradients with respect to the amount of child care.

6.2.1 First and second-stage estimation

For our first-stage regressions, our main focus is the performance of our instruments. In order to analyze this, we check the significance of our instruments in each of our three first-stage regressions. Noting that we have 99 regressors in each first-stage, the percentage of significant instruments in each first-stage regression is roughly one-half. This type of analysis, of course, is informal. Rather than relying on univariate-type significances, we prefer to perform formal tests to check for joint significance.

Here we perform a nonparametric F-type test, originally proposed by Ullah (1985). The test involves comparing the residual sum of squares between a restricted and unrestricted model. Our restricted model assumes that each of our instruments have coefficients equal to zero. The asymptotic distribution can be obtained by multiplying the statistic by a constant, but it is well known that using the asymptotic distribution is problematic in practice. Instead, we use a wild bootstrap to determine the conclusion of our tests. For each first-stage regression we perform a test where the null is that each instrument is irrelevant. In each case our p-value is zero to at least four decimal places. Hence, we argue (as did Bernal and Keane, 2011) that our instruments are relevant in our prediction of our endogenous regressors.

In the second-stage regression we are concerned with the joint significance of each of the residuals from the first-stage regressions. We perform a similar Ullah (1985) type test as above and reject the null that the three sets of residuals are jointly insignificant with a p-value that is zero to four decimal places. We conclude that endogeneity is likely present and thus justify the use of our procedure.

6.2.2 Final-stage estimation

Here we are interested in comparing our gradient estimate results to those of Bernal and Keane (2011). Our average gradient estimates (Table 2) for the nonlinear variables are often similar in magnitude and sign to their results. We find mostly positive effects for mother's income, education and AFQT scores. For our primary variable of interest, the gradient on child care is also similar at the median (noting that our median result is insignificant). To put this in perspective, the median coefficient of -0.0005 is equivalent to a 0.2% decrease in test scores for an additional year of child care use (or 0.05% for an additional quarter). That being said, each of these statements ignore the heterogeneity in the gradient estimates allowed by our procedure.

Table 2: Final-stage gradient estimates for each of the nonlinear variables at various percentiles (10%, 25%, 50%, 75% and 90%) with corresponding wild bootstrapped standard errors

	0.10	0.25	0.50	0.75	0.90
care	-0.0089	-0.0061	-0.0005	0.0049	0.0079
	0.0034	0.0038	0.0042	0.0027	0.0054
inc	-0.0415	-0.0045	0.0250	0.0463	0.0741
	0.0225	0.0258	0.0249	0.0329	0.0304
nmchild	-0.0157	-0.0157	-0.0117	-0.0021	-0.0021
	0.0077	0.0077	0.0046	0.0065	0.0065
med	-0.0058	-0.0058	-0.0046	0.0037	0.0128
	0.0062	0.0062	0.0061	0.0091	0.0080
mex	-0.0034	-0.0022	-0.0016	0.0003	0.0031
	0.0031	0.0031	0.0025	0.0038	0.0038
mafqt	-0.0012	0.0000	0.0015	0.0034	0.0046
	0.0008	0.0009	0.0009	0.0013	0.0019
mage	-0.0051	-0.0017	0.0035	0.0051	0.0063
	0.0028	0.0024	0.0024	0.0020	0.0037

When we look at the percentiles, we see both positive and negative estimates. This is not possible with standard linear models. We therefore want to determine the story behind these variable returns. To do so we break the results up for different child care type, amount of child care and between gender.

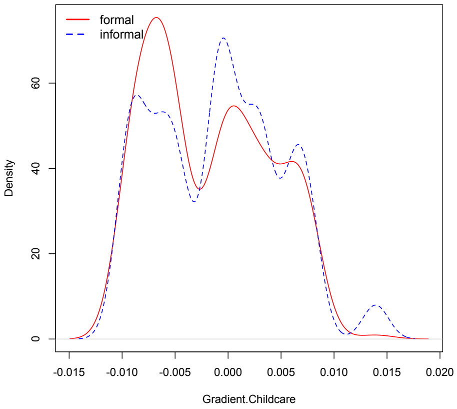

Figure 1: Density of estimated returns to child care for those with only formal versus those with only informal child care

We also examine whether heterogeneity is present amongst mothers (level of education, experience and age). Finally, we try to determine what attributes are common with those receiving positive or negative returns to child care use.

Disaggregating the child care gradient In the first few rows of Table 3, we analyze the child care gradient with respect to child care type, child gender and amount of child care use. In this table we report the percentiles for the estimated gradients for the related groups and associated standard errors. We also provide Figures 1-3 which show the overall variation for the (selected) chosen pairs. Before we get into the details for different groups, we want to point out that many of the results are insignificant. In fact, only 634 of the 2454 estimates are statistically significant at the five-percent level. What this implies is that for a large portion of the sample, an additional unit of child care will have no impact on test scores. That being said, we find many cases where it does matter and we will highlight the results below.

Bernal and Keane (2011) found that only informal child care (e.g., a grandparent) had significantly negative effects. Specifically, they found that an additional year of informal child care led to a 2.6% reduction in test scores. To compare our results, we separated the gradients on child care use between those who received only formal or only informal child care. Although there are some differences in our percentile estimates, Figure 1 shows essentially no difference in the densities of estimates between those who received only formal (e.g., a licensed child care facility) versus only informal child care. Bernal and Keane (2011) also find differences in returns to child care between genders. Both Table 3 and Figure 2 show essentially no difference between these two groups.

Where we do find a difference, is with respect to the amount of child care. In Figure 3, we look at the estimated gradients of child care use for those children who get below and above the median total child care. We find that more child care use leads to lower returns. In fact, we can see negative returns for those children receiving relatively more child care which suggests decreasing returns to child care use. We consider this finding important since this shows us evidence to believe that the lower returns may be

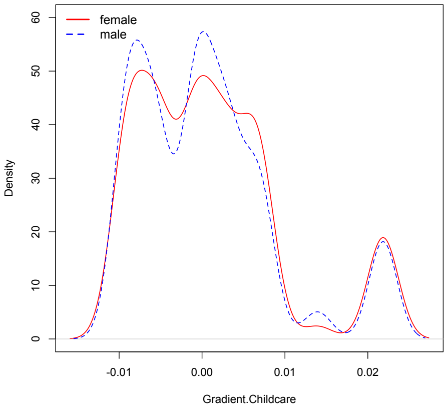

Figure 2: Density of estimated returns to child care for male and female children

associated with the amount of child care use rather than type of child care or gender of the child.

In the remaining rows of Table 3, we separate our estimated gradients on child care based on attributes of the mothers. Specifically, we analyze the estimated returns for mothers of different levels of education, experience and age. We see variation in our estimates for mothers of different age groups. As mothers get older, we see lower returns to child care use (more negative and significant estimates) as compared to those of younger mothers.

We also see variation for different experience levels. Here we find more negative (and significant) estimates for mother's with more experience. On the other hand, for mother's with no experience, we see much larger (and significant) returns to child care.

Similarly, mother's with more education appear to have more negative returns. Many of the percentiles are negative for mother's with twelve or more years of education. For mother's without a high school diploma, we see positive effects for all but the lowest percentile. In other words, for less educated mothers, more child care may actually improve their child's test scores.

All these results shows us two important things. First, there is substantial heterogeneity in our returns to child care use that cannot be captured with a single parameter estimate. Second, the returns tend to be related to the amount of child care and the quality of maternal time. For those mothers who have more education and experience, child care tends to hurt their child's test score. On the other hand, for those mothers with less education and experience, our results tend to suggest that their children may be better off with more child care.

Positive and negative returns For most of the reported estimates, we tend to see both positive and negative returns. This is perhaps a more important result than the lower and higher estimated returns. What this finding suggests is that there are some children who benefit from additional child care and there are some who are harmed. We hope to uncover who these children are and hopefully, the drivers of such returns.

Table 4 separates the partial effects on child care by those which are positive and negative. Perhaps the first point of interest is that slightly more than half of the gradients are negative. If we were to run a

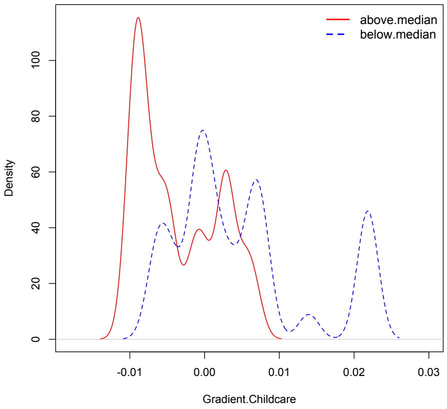

Figure 3: Density of estimated returns to child care for those with above and below the median units of child care

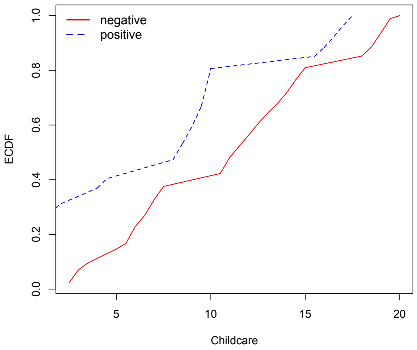

Figure 4: Empirical cummulative distribution functions for the amount of child care for those with both positive and negative returns to child care

simple ordinary least squares regression, this would (back of the envelope) suggest a negative coefficient. This is what is typically found in the literature.

As for the remaining values in the table, the rows represent characteristics of interest and each number represents the median value for that characteristic for both children with positive and negative returns. Many values are the same. For example, mother's education, child's age, quarters of formal child care and number of children is the same at the median in each group. However, we see that for negative returns that, (median) mothers have more experience and are older. It is true that they have more informal child care (which likely represents the result of Bernal and Keane, 2011), but it is also true that they have far more quarters of child care at the median.

Again, these are only point estimates. If we were to plot the ECDFs of types of child care for the groups with negative and positive child care gradient, we would see that the amount of both formal and informal child care use are higher for those with negative returns. For example, Figure 4 plots the distribution functions with respect to total child care use between those with negative and those with positive returns to child care. Those children with negative returns to child care, receive more child care overall. We fail to reject the null of first-order dominance via a Kolmogorov-Smirnov test with a p-value near unity. We also find this same level of dominance when we look at formal versus informal or male versus female. This is strong evidence that it is the amount of child care and not necessarily the type that matters.

7 Conclusion

In this paper we develop an oracle efficient additive nonparametric structural equation model. We show that our estimators of the conditional mean and gradient are consistent, asymptotically normal and free from the curse of dimensionality. The finite sample results support the asymptotic development. We also consider a partially linear extension of our model which we use in an empirical application relating child care to test scores. In our application we find that the amount of child care use and not the type, is primarily responsible for the sign of the returns. Given that our nonparametric procedure will give us observation specific estimates, we are able to uncover, that in addition to the amount of child care, what attributes of mothers are related to different returns. We find evidence that more educated, more experienced mothers with higher test scores (themselves) are associated with lower returns to child care for their children. On the other hand, less educated, less experienced mothers with lower test scores' (for themselves) children often have positive returns to child care.

Appendix

A Proof of the Results in Section 3

For notational simplicity, let $\mathbf{Z}_i \equiv (\mathbf{Z}'_{1i}, \mathbf{Z}'_{2i})'$, $P_i \equiv P^{\kappa_1}(\mathbf{Z}_i)$, $\Phi_i = \Phi^{\kappa}(\mathbf{W}_i)$ and $\widetilde{\Phi}_i = \Phi^{\kappa}(\widetilde{\mathbf{W}}_i)$. Then $Q_{PP} \equiv E(P_iP_i')$ and $Q_{\Phi\Phi} \equiv E(\Phi_i\Phi_i')$. Let $Q_{n,PP} \equiv n^{-1}\sum_{i=1}^n P_iP_i'$, $Q_{n,\Phi\Phi} \equiv n^{-1}\sum_{i=1}^n \Phi_i\Phi_i'$ and $\widetilde{Q}_{n,\Phi\Phi} \equiv n^{-1}\sum_{i=1}^n \widetilde{\Phi}_i\widetilde{\Phi}_i'$. By Lemmas A.1$(ii)$ and(v) and Lemma A.4(iv) below, $Q_{n,PP}$, $Q_{n,\Phi\Phi}$ and $\widetilde{Q}_{n,\Phi\Phi}$ are invertible with probability approaching 1 (w.p.a.1) so that in large samples we can replace the generalized inverse $Q_{n,PP}^-$, $Q_{n,\Phi\Phi}^-$ and $\widetilde{Q}_{n,\Phi\Phi}^-$ by $Q_{n,PP}^{-1}$, $Q_{n,\Phi\Phi}^{-1}$ and $\widetilde{Q}_{n,\Phi\Phi}^{-1}$, respectively. Recall $\nu_{1n} \equiv \kappa_1^{1/2}/n^{1/2} + \kappa_1^{-\gamma}$ and $\nu_n \equiv \kappa^{1/2}/n^{1/2} + \kappa^{-\gamma}$.

Lemma A.1 *Suppose that Assumptions A1 and A2(i)-(ii) and (vi) hold. Then*
(i) $\|Q_{n,PP} - Q_{PP}\|^2 = O_P(\kappa_1^2/n)$;
(ii) $\lambda_{\min}(Q_{n,pp}) = \lambda_{\min}(Q_{PP}) + o_P(1)$ *and* $\lambda_{\max}(Q_{n,pp}) = \lambda_{\max}(Q_{PP}) + o_P(1)$;
(iii) $\left\|Q_{n,PP}^{-1} - Q_{PP}^{-1}\right\|_{sp} = O_P(\kappa_1/n^{1/2})$;
(iv) $\|Q_{n,\Phi\Phi} - Q_{\Phi\Phi}\|^2 = O_P(\kappa^2/n)$;
(v) $\lambda_{\min}(Q_{n,\Phi\Phi}) = \lambda_{\min}(Q_{\Phi\Phi}) + o_P(1)$ *and* $\lambda_{\max}(Q_{n,\Phi\Phi}) = \lambda_{\max}(Q_{\Phi\Phi}) + o_P(1)$.

Proof. By straightforward moment calculations, we can show that $E\|Q_{n,PP} - Q_{PP}\|^2 = O(\kappa_1^2/n)$ under Assumption A1(i)-(ii) and A2(vi). Then (i) follows from Markov inequality. By Weyl inequality [e.g., Bernstein (2005, Theorem 8.4.11)] and the fact that $\lambda_{\max}(A) \le \|A\|$ for any symmetric matrix A (as $|\lambda_{\max}(A)|^2 = \lambda_{\max}(AA) \le \|A\|^2$), we have

$$\begin{aligned}\lambda_{\min}(Q_{n,PP}) &\le \lambda_{\min}(Q_{PP}) + \lambda_{\max}(Q_{n,PP} - Q_{PP}) \\ &\le \lambda_{\min}(Q_{PP}) + \|Q_{n,PP} - Q_{PP}\| = \lambda_{\min}(Q_{PP}) + o_P(1).\end{aligned}$$

Similarly,

$$\begin{aligned}\lambda_{\min}(Q_{n,PP}) &\ge \lambda_{\min}(Q_{PP}) + \lambda_{\min}(Q_{n,PP} - Q_{PP}) \\ &\ge \lambda_{\min}(Q_{PP}) - \|Q_{n,PP} - Q_{PP}\| = \lambda_{\min}(Q_{\kappa_1}) - o_P(1).\end{aligned}$$

Analogously, we can prove the second part of (ii). Thus (ii) follows. By the submultiplicative property of the spectral norm, (i)-(ii) and Assumption A2(i)

$$\begin{aligned}\left\|Q_{n,PP}^{-1} - Q_{PP}^{-1}\right\|_{\text{sp}} &= \left\|Q_{n,PP}^{-1}(Q_{PP} - Q_{n,PP})Q_{PP}^{-1}\right\|_{\text{sp}} \le \left\|Q_{n,PP}^{-1}\right\|_{\text{sp}} \|Q_{PP} - Q_{n,PP}\|_{\text{sp}} \left\|Q_{PP}^{-1}\right\|_{\text{sp}} \\ &= O_P(1)\, O_P\left(\kappa_1/n^{1/2}\right) O_P(1) = O_P\left(\kappa_1/n^{1/2}\right),\end{aligned}$$

where we use the fact that $\left\|Q_{n,PP}^{-1}\right\|_{\text{sp}} = [\lambda_{\min}(Q_{n,PP})]^{-1} = [\lambda_{\min}(Q_{PP}) + o_P(1)]^{-1} = O_P(1)$ by (ii) and Assumption A2(i). Then (iii) follows. The proof of (iv)-(v) is analogous to that of (i)-(ii) and thus omitted. ■

Lemma A.2 *Let $\xi_{nl} \equiv n^{-1}\sum_{i=1}^n P_iU_{li}$ and $\zeta_{nl} \equiv n^{-1}\sum_{i=1}^n P_i[m_l(\mathbf{Z}_i) - P_i'\boldsymbol{\alpha}_l]$ for $l = 1, ..., d_x$. Suppose that Assumptions A1-A2 hold. Then*
(i) $\|\xi_{nl}\|^2 = O_P(\kappa_1/n)$;
(iii) $\|\zeta_{nl}\|^2 = O_P(\kappa_1^{-2\gamma})$;
(iii) $\widetilde{\boldsymbol{\alpha}}_l - \boldsymbol{\alpha}_l = Q_{\kappa_1}^{-1}n^{-1}\sum_{i=1}^n P_iU_{li} + Q_{\kappa_1}^{-1}n^{-1}\sum_{i=1}^n P_i[m_l(\mathbf{Z}_i) - P_i'\boldsymbol{\alpha}_l] + r_{nl}$;
where $\|r_{nl}\| = O_P(\kappa_1/n + \kappa_1^{-\gamma+1/2}/n^{1/2})$ *and* $l = 1, ..., d_x$.

Proof. (*i*) By Assumption A1(*i*) and A2(*i*), $E\left\|\xi_{nl}\right\|^2 = n^{-2}\text{tr}\{\sum_{i=1}^n E(P_iP_i'U_{li}^2)\} \leq n^{-1}\left(1+d\kappa_1\right) \lambda_{\max}\left(Q_{PP,U_l}\right) = O(\kappa_1/n)$. Then $\left\|\xi_{nl}\right\|^2 = O_P(\kappa_1/n)$ by Markov inequality.

(*ii*) By the facts that $\left\|a\right\|_{\text{sp}}^2 = \left\|a\right\|^2$ for any vector a, $|a'b| \leq \|a\|\,\|b\|$ for any two conformable vectors a and b and that $\varkappa' A\varkappa \leq \lambda_{\max}\left(A\right)\left\|\varkappa\right\|^2$ for any p.s.d. matrix A and conformable vector $\varkappa$, Cauchy-Schwarz inequality, Lemma A.1(*ii*) and Assumptions A2(*iv*), we have

$$\begin{aligned}
\left\|\zeta_{nl}\right\|^2 &= \left\|\zeta_{nl}\right\|_{\text{sp}}^2 = \lambda_{\max}\left(\zeta_{nl}\zeta_{nl}'\right) = \max_{\|\varkappa\|=1} n^{-2}\sum_{i=1}^n\sum_{j=1}^n \varkappa' P_iP_j'\varkappa\left[m_l\left(\mathbf{Z}_i\right) - P_i'\boldsymbol{\alpha}_l\right]\left[m_l\left(\mathbf{Z}_j\right) - P_j'\boldsymbol{\alpha}_l\right] \\
&\leq \max_{\|\varkappa\|=1}\left\{n^{-1}\sum_{i=1}^n\left\{\varkappa' P_iP_i'\varkappa\left[m_l\left(\mathbf{Z}_i\right) - P_i'\boldsymbol{\alpha}_l\right]^2\right\}^{1/2}\right\}^2 \\
&\leq O_P(\kappa_1^{-2\gamma})\max_{\|\varkappa\|=1}\left\{n^{-1}\sum_{i=1}^n \varkappa' P_iP_i'\varkappa\right\} \leq O_P(\kappa_1^{-2\gamma})\lambda_{\max}\left(Q_{n,PP}\right) = O_P(\kappa_1^{-2\gamma}).
\end{aligned}$$

(*iii*) Noting that $X_{li} = m_l\left(\mathbf{Z}_i\right) + U_{li} = P_i'\boldsymbol{\alpha}_l + U_{li} + \left[m_l\left(\mathbf{Z}_i\right) - P_i'\boldsymbol{\alpha}_l\right]$, by Lemma A.1(*ii*), w.p.a.1 we have

$$\begin{aligned}
\widetilde{\boldsymbol{\alpha}}_l - \boldsymbol{\alpha}_l &= \left(\sum_{i=1}^n P_iP_i'\right)^{-}\sum_{i=1}^n P_iX_{li} - \boldsymbol{\alpha}_l \\
&= Q_{n,PP}^{-1}n^{-1}\sum_{i=1}^n P_iU_{li} + Q_{n,PP}^{-1}n^{-1}\sum_{i=1}^n P_i\left[m_l\left(\mathbf{Z}_i\right) - P_i'\boldsymbol{\alpha}_l\right] \\
&= Q_{n,PP}^{-1}\xi_{nl} + Q_{n,PP}^{-1}\zeta_{nl} \equiv a_{1l} + a_{2l},\ \text{say.} \qquad \text{(A.1)}
\end{aligned}$$

Note that $a_{1l} = Q_{\kappa_1}^{-1}\xi_{nl} + r_{1nl}$ where $r_{1nl} = \left(Q_{n,PP}^{-1} - Q_{PP}^{-1}\right)\xi_{nl}$ satisfies that

$$\begin{aligned}
\left\|r_{1l}\right\| &\leq = \left\{\text{tr}\left[\left(Q_{n,PP}^{-1} - Q_{PP}^{-1}\right)\xi_{nl}\xi_{nl}'\left(Q_{n,PP}^{-1} - Q_{PP}^{-1}\right)\right]\right\}^{1/2} \\
&\leq \left\|\xi_{nl}\right\|_{\text{sp}}\left\|Q_{n,PP}^{-1} - Q_{PP}^{-1}\right\| = O_P(\kappa_1^{1/2}/n^{1/2})O_P(\kappa_1^{1/2}/n^{1/2}) = O_P(\kappa_1/n)
\end{aligned}$$

by Lemmas A.1(*iii*) and A.2(*i*). For a_{2l}, we have $a_{2l} = Q_{\kappa_1}^{-1}\zeta_{nl} + r_{2nl}$ where $r_{2nl} = \left(Q_{n\kappa_1}^{-1} - Q_{n\kappa_1}^{-1}\right)\zeta_{nl}$ satisfies that

$$\left\|r_{2l}\right\| \leq \left\|\zeta_{nl}\right\|_{\text{sp}}\left\|Q_{n,PP}^{-1} - Q_{PP}^{-1}\right\| = O_P(\kappa_1^{-\gamma})O_P(\kappa_1^{1/2}/n^{1/2}) = O_P(\kappa_1^{-\gamma+1/2}/n^{1/2})$$

by Lemmas A.1(*iii*) and A.2(*ii*). The result follows. ■

Lemma A.3 *Suppose that Assumptions A1-A3 hold. Then for* $l = 1, ..., d_x$,

(*i*) $n^{-1}\sum_{i=1}^n\left(\widetilde{U}_{li} - U_{li}\right)^2\left[\sigma_i^2\right]^r = O_P\left(\nu_{1n}^2\right)$ *for* $r = 0, 1$;

(*ii*) $n^{-1}\sum_{i=1}^n\left(\widetilde{U}_{li} - U_{li}\right)^2\left\|\Phi_i\right\|^r = O_P\left(\varsigma_{0\kappa}^r\nu_{1n}^2\right)$ *for* $r = 1, 2$;

(*iii*) $n^{-1}\sum_{i=1}^n\left\|p^\kappa\left(\widetilde{U}_{li}\right) - p^\kappa\left(U_{li}\right)\right\|^2 = O_P\left(\varsigma_{1\kappa}^2\nu_{1n}^2\right)$;

(*iv*) $\left\|n^{-1}\sum_{i=1}^n\left[p^\kappa\left(\widetilde{U}_{li}\right) - p^\kappa\left(U_{li}\right)\right]\Phi_i'\right\| = O_P\left(\kappa^{1/2}\varsigma_{0\kappa}\nu_{1n} + \varsigma_{0\kappa}\varsigma_{2\kappa}\nu_{1n}^2\right)$;

(*v*) $\left\|n^{-1}\sum_{i=1}^n\left[p^\kappa\left(\widetilde{U}_{li}\right) - p^\kappa\left(U_{li}\right)\right]e_i\right\| = O_P\left(n^{-1/2}\varsigma_{1\kappa}\nu_{1n}\right)$.

Proof. (i) We only prove the $r=1$ case as the proof of the other case is almost identical. By the definition of $\widetilde{U}_{li}$ and (A.1), we can decompose $\widetilde{U}_{li}-U_{li}=[X_{li}-\widetilde{m}_l(\mathbf{Z}_i)]-U_{li}$ as follows

$$\begin{aligned}\widetilde{U}_{li}-U_{li} &= (\mu_l-\widetilde{\mu}_l)+\sum_{k=1}^{d_1}\left[m_{l,k}(Z_{1k,i})-\widetilde{m}_{l,k}(Z_{1k,i})\right]+\sum_{k=1}^{d_2}\left[m_{l,d_1+k}(Z_{2k,i})-\widetilde{m}_{l,d_1+k}(Z_{2k,i})\right]\\ &= -(\widetilde{\mu}_l-\mu_l)-\sum_{k=1}^{d_1}p^{\kappa_1}(Z_{1k,i})'\mathbb{S}_{1k}a_{1l}-\sum_{k=1}^{d_2}p^{\kappa_1}(Z_{2k,i})'\mathbb{S}_{1,d_1+k}a_{1l}\\ &\quad -\sum_{k=1}^{d_1}p^{\kappa_1}(Z_{1k,i})'\mathbb{S}_{1k}a_{2l}-\sum_{k=1}^{d_2}p^{\kappa_1}(Z_{2k,i})'\mathbb{S}_{1,d_1+k}a_{2l}\\ &\equiv -u_{1l,i}-u_{2l,i}-u_{3l,i}-u_{4l,i}-u_{5l,i},\text{ say.}\end{aligned}\tag{A.2}$$

Then by Cauchy-Schwarz inequality, $n^{-1}\sum_{i=1}^n(\widetilde{U}_{li}-U_{li})^2\sigma_i^2\leq 5\sum_{s=1}^5 n^{-1}\sum_{i=1}^n u_{sl,i}^2\sigma_i^2\equiv 5\sum_{s=1}^5 V_{nl,s}$, say. Apparently, $V_{nl,1}=O_P\left(n^{-1}\right)$ as $\widetilde{\mu}_l-\mu_l=O_P\left(n^{-1/2}\right)$.

$$\begin{aligned}V_{nl,2} &= n^{-1}\sum_{i=1}^n\left(\sum_{k=1}^{d_1}p^{\kappa_1}(Z_{1k,i})'\mathbb{S}_{1k}a_{1l}\right)^2\sigma_i^2\\ &\leq d_1\sum_{k=1}^{d_1}n^{-1}\sum_{i=1}^n\left(p^{\kappa_1}(Z_{1k,i})'\mathbb{S}_{1k}a_{1l}\right)^2\sigma_i^2=d_1\sum_{k=1}^{d_1}\operatorname{tr}\left(\mathbb{S}_{1k}a_{1l}a_{1l}'\mathbb{S}_{1k}'Q_{n1k,pp}\right)\\ &\leq d_1\sum_{k=1}^{d_1}\lambda_{\max}(Q_{n1k,pp})\operatorname{tr}\left(a_{1l}a_{1l}'\mathbb{S}_{1k}'\mathbb{S}_{1k}\right)\leq d_1\sum_{k=1}^{d_1}\lambda_{\max}(Q_{n1k,pp})\left\|\mathbb{S}_{1k}\right\|_{\rm sp}^2\left\|a_{1l}\right\|^2.\end{aligned}$$

where $Q_{n1k,pp}=n^{-1}\sum_{i=1}^n p^{\kappa_1}(Z_{1k,i})p^{\kappa_1}(Z_{1k,i})'\sigma_i^2$ such that $\lambda_{\max}(Q_{n1k,pp})=O_P(1)$ by Assumption A3(ii) and arguments analogous to those used in the proof of Lemma A.1(ii). In addition, $\left\|\mathbb{S}_{1k}\right\|_{\rm sp}^2=\lambda_{\max}(\mathbb{S}_{1k}\mathbb{S}_{1k}')=1$ and $\|a_{1l}\|^2\leq\left\|Q_{n,PP}^{-1}\right\|_{\rm sp}^2\|\xi_{nl}\|^2=O_P(1)\,O_P(\kappa_1/n)=O_P(\kappa_1/n)$ by Lemma A.1(iii) and A.2(i) and Assumption A2(i). It follows that $V_{nl,2}=O_P(1)\times 1\times O_P(\kappa_1/n)=O_P(\kappa_1/n)$. Similarly, using the fact that $\|a_{2l}\|^2\leq\left\|Q_{n,PP}^{-1}\right\|_{\rm sp}^2\|\zeta_{nl}\|^2=O_P(1)\,O_P(\kappa_1^{-2\gamma})$, we have

$$\begin{aligned}V_{nl,4} &= n^{-1}\sum_{i=1}^n\left(\sum_{k=1}^{d_1}p^{\kappa_1}(Z_{1k,i})'\mathbb{S}_{1k}a_{2l}\right)^2\sigma_i^2\leq d_1\sum_{k=1}^{d_1}\lambda_{\max}(Q_{n1k,pp})\left\|\mathbb{S}_{1k}\right\|_{\rm sp}^2\operatorname{tr}\left(a_{2l}a_{2l}'\right)\\ &= O_P(1)\times 1\times O_P(\kappa_1^{-2\gamma})=O_P(\kappa_1^{-2\gamma}).\end{aligned}$$

By the same token, $V_{nl,3}=O_P(\kappa_1 n^{-1})$ and $V_{nl,5}=O_P(\kappa_1^{-2\gamma})$.

(ii) The result follows from (i) and the fact that $\max_{1\leq i\leq n}\|\Phi_i\|=O_P(\varsigma_{0\kappa})$ under Assumption A2(vi).

(iii) By Assumption A2(vi), Taylor expansion and (i),

$$\begin{aligned}n^{-1}\sum_{i=1}^n\left\|p^{\kappa}\left(\widetilde{U}_{li}\right)-p^{\kappa}(U_{li})\right\|^2 &= n^{-1}\sum_{i=1}^n\left\|\dot{p}^{\kappa}\left(U_{li}^{\dagger}\right)\left(\widetilde{U}_{li}-U_{li}\right)\right\|^2\\ &\leq O\left(\varsigma_{1\kappa}^2\right)n^{-1}\sum_{i=1}^n\left(\widetilde{U}_{li}-U_{li}\right)^2=O_P\left(\varsigma_{1\kappa}^2\nu_{1n}^2\right),\end{aligned}$$

where $U_{li}^{\dagger}$ lies between $\widetilde{U}_{li}$ and U_{li}.

(*iv*) By Assumption A2(*vi*) , Taylor expansion and triangle inequality, $\left\| n^{-1} \sum_{i=1}^{n} \left[p^{\kappa}\left(\widetilde{U}_{li}\right) - p^{\kappa}\left(U_{li}\right)\right] \Phi_i' \right\|_{\text{sp}}$ is bounded by

$$\left\| n^{-1} \sum_{i=1}^{n} \dot{p}^{\kappa}\left(U_{li}\right) \Phi_i' \left(\widetilde{U}_{li} - U_{li}\right) \right\|_{\text{sp}} + \frac{1}{2} \left\| n^{-1} \sum_{i=1}^{n} \ddot{p}^{\kappa}\left(U_{li}^{\ddagger}\right) \Phi_i' \left(\widetilde{U}_{li} - U_{li}\right)^2 \right\|_{\text{sp}} \equiv T_{nl,1} + T_{nl,2},$$

where $U_{li}^{\ddagger}$ lies between $\widetilde{U}_{li}$ and U_{li}. By triangle and Cauchy-Schwarz inequalities and (*ii*) ,

$$\begin{aligned} T_{nl,1} &\leq n^{-1} \sum_{i=1}^{n} \left\| \dot{p}^{\kappa}\left(U_{li}\right) \right\|_{\text{sp}} \left\| \Phi_i' \left(\widetilde{U}_{li} - U_{li}\right) \right\|_{\text{sp}} \\ &\leq \left\{ n^{-1} \sum_{i=1}^{n} \left\| \dot{p}^{\kappa}\left(U_{li}\right) \right\|^2 \right\}^{1/2} \left\{ n^{-1} \sum_{i=1}^{n} \left\| \Phi_i \right\|^2 \left| \widetilde{U}_{li} - U_{li} \right|^2 \right\}^{1/2} \\ &= O_P\left(\kappa^{1/2}\right) O_P\left(\varsigma_{0\kappa} \nu_{1n}\right) = O_P\left(\kappa^{1/2} \varsigma_{0\kappa} \nu_{1n}\right). \end{aligned}$$

By triangle inequality and (*i*) , $T_{nl,2} \leq O\left(\varsigma_{0\kappa}\varsigma_{2\kappa}\right) n^{-1} \sum_{i=1}^{n} (\widetilde{U}_{li} - U_{li})^2 = O_P(\varsigma_{0\kappa}\varsigma_{2\kappa}\nu_{1n}^2)$. Then (*iv*) follows.

(*v*) Let $\Gamma_{nl} \equiv [p^{\kappa}(\widetilde{U}_{l1}) - p^{\kappa}\left(U_{l1}\right), ..., [p^{\kappa}(\widetilde{U}_{ln}) - p^{\kappa}\left(U_{ln}\right)]]'$ and $\mathbf{e} = (e_1, ..., e_n)'$. Then we can write $n^{-1} \sum_{i=1}^{n} [p^{\kappa}(\widetilde{U}_{li}) - p^{\kappa}\left(U_{li}\right)] e_i$ as $n^{-1} \Gamma_{nl}' \mathbf{e}$. Let $\mathbb{D}_n \equiv \{(\mathbf{X}_i, \mathbf{Z}_i, \mathbf{U}_i)\}_{i=1}^{n}$. By the law of iterated expectations, Taylor expansion, Assumptions A1(*i*) , A3(*ii*) and A2(*vi*) and (*i*)

$$\begin{aligned} E\left\{ \left\| n^{-1} \Gamma_{nl}' \mathbf{e} \right\|^2 | \mathbb{D}_n \right\} &= n^{-2} E\left[\operatorname{tr}\left(\Gamma_n' \mathbf{e}\mathbf{e}' \Gamma_n \right) \right] = n^{-2} E\left[\operatorname{tr}\left(\Gamma_n' E\left(\mathbf{e}\mathbf{e}' | \mathbb{D}_n \right) \Gamma_n \right) \right] \\ &= n^{-2} \sum_{i=1}^{n} [p^{\kappa}(\widetilde{U}_{li}) - p^{\kappa}\left(U_{li}\right)]^2 \sigma_i^2 \\ &\leq O_P\left(\varsigma_{1\kappa}\right) n^{-2} \sum_{i=1}^{n} \left(\widetilde{U}_{li} - U_{li} \right)^2 \sigma_i^2 = O_P\left(n^{-1} \varsigma_{1\kappa}^2 \nu_{1n}^2 \right). \end{aligned}$$

It follows that $\left\| n^{-1} \Gamma_{nl}' \mathbf{e} \right\| = O_P\left(n^{-1/2} \varsigma_{1\kappa} \nu_{1n} \right)$ by the conditional Chebyshev inequality. ■

Lemma A.4 *Suppose Assumptions A1-A3 hold. Then*

(*i*) $n^{-1} \sum_{i=1}^{n} \left\| \widetilde{\Phi}_i - \Phi_i \right\|^2 = O_P\left(\varsigma_{1\kappa}^2 \nu_{1n}^2 \right)$;

(*ii*) $\left\| n^{-1} \sum_{i=1}^{n} \left(\widetilde{\Phi}_i - \Phi_i \right) \Phi_i' \right\|_{sp} = O_P\left(\kappa^{1/2} \varsigma_{1\kappa} \nu_{1n} \right)$;

(*iii*) $\left\| \widetilde{Q}_{n,\Phi\Phi} - Q_{n,\Phi\Phi} \right\|_{sp} = O_P\left(\kappa^{1/2} \varsigma_{0\kappa} \nu_{1n} + \varsigma_{0\kappa} \varsigma_{2\kappa} \nu_{1n}^2 \right)$;

(*iv*) $\left\| \widetilde{Q}_{n,\Phi\Phi}^{-1} - Q_{\Phi\Phi}^{-1} \right\|_{sp} = O_P\left(\kappa^{1/2} \varsigma_{0\kappa} \nu_{1n} + \varsigma_{0\kappa} \varsigma_{2\kappa} \nu_{1n}^2 \right)$;

(*v*) $\left\| n^{-1} \sum_{i=1}^{n} \left(\widetilde{\Phi}_i - \Phi_i \right) e_i \right\| = O_P\left(n^{-1/2} \varsigma_{1\kappa} \nu_{1n} \right)$;

(*vi*) $n^{-1} \sum_{i=1}^{n} \left(\widetilde{\Phi}_i - \Phi_i \right) \left[\bar{g}\left(X_i, Z_{1i}, U_i \right) - \Phi_i' \boldsymbol{\beta} \right] = O_P\left(\kappa_1^{-\gamma} \varsigma_{1\kappa} \nu_{1n} \right)$.

Proof. (*i*) Noting that $n^{-1} \sum_{i=1}^{n} \left\| \widetilde{\Phi}_i - \Phi_i \right\|^2 = \sum_{l=1}^{d_x} n^{-1} \sum_{i=1}^{n} \left\| p^{\kappa}\left(\widetilde{U}_{li} \right) - p^{\kappa}\left(U_{li} \right) \right\|^2$, the result follows from Lemma A.3(*iii*).

(*ii*) Noting that $\left\| n^{-1} \sum_{i=1}^{n} \left(\widetilde{\Phi}_i - \Phi_i \right) \Phi_i' \right\|^2 = \sum_{l=1}^{d_x} \left\| n^{-1} \sum_{i=1}^{n} \left[p^{\kappa}\left(\widetilde{U}_{li} \right) - p^{\kappa}\left(U_{li} \right) \right] \Phi_i' \right\|^2$, the result follows from Lemma A.3(*iv*).

(*iii*) Noting that $\widetilde{Q}_{n,\Phi\Phi} - Q_{n,\Phi\Phi} = n^{-1} \sum_{i=1}^{n} (\widetilde{\Phi}_i \widetilde{\Phi}_i' - \Phi_i \Phi_i') = n^{-1} \sum_{i=1}^{n} (\widetilde{\Phi}_i - \Phi_i)(\widetilde{\Phi}_i - \Phi_i)' + n^{-1} \sum_{i=1}^{n} (\widetilde{\Phi}_i - \Phi_i) \Phi_i' + n^{-1} \sum_{i=1}^{n} \Phi_i (\widetilde{\Phi}_i - \Phi_i)'$, the result follows from (*i*)-(*ii*) and the triangle inequality.

(iv) By the triangle inequality $\left\|\widetilde{Q}_{n,\Phi\Phi}^{-1}-Q_{\Phi\Phi}^{-1}\right\|_{\text{sp}} \leq \left\|\widetilde{Q}_{n,\Phi\Phi}^{-1}-Q_{n,\Phi\Phi}^{-1}\right\|_{\text{sp}}+\left\|Q_{n,\Phi\Phi}^{-1}-Q_{\Phi\Phi}^{-1}\right\|_{\text{sp}}$. Arguments like those used in the proof of Lemma A.1(ii) show that $\left\|\widetilde{Q}_{n,\Phi\Phi}^{-1}\right\|_{\text{sp}}=\left[\lambda_{\min}\left(\widetilde{Q}_{n,\Phi\Phi}\right)\right]^{-1}=\left[\lambda_{\min}\left(Q_{\Phi\Phi}\right)+o_P\left(1\right)\right]^{-1}=O_P\left(1\right)$ where the second equality follows from (iii) and Lemma A.1(ii). By the submultiplicative property of the spectral norm and (iii),

$$\begin{aligned}\left\|\widetilde{Q}_{n,\Phi\Phi}^{-1}-Q_{n,\Phi\Phi}^{-1}\right\|_{\text{sp}} &= \left\|\widetilde{Q}_{n,\Phi\Phi}^{-1}\left(\widetilde{Q}_{n,\Phi\Phi}-\widetilde{Q}_{n,\Phi\Phi}\right)Q_{n,\Phi\Phi}^{-1}\right\|_{\text{sp}} \\ &\leq \left\|\widetilde{Q}_{n,\Phi\Phi}^{-1}\right\|_{\text{sp}}\left\|\widetilde{Q}_{n,\Phi\Phi}-\widetilde{Q}_{n,\Phi\Phi}\right\|_{\text{sp}}\left\|Q_{n,\Phi\Phi}^{-1}\right\|_{\text{sp}} \\ &= O_P\left(\kappa^{1/2}\varsigma_{0\kappa}\nu_{1n}+\varsigma_{0\kappa}\varsigma_{2\kappa}\nu_{1n}^2\right).\end{aligned}$$

Similarly, $\left\|Q_{n,\Phi\Phi}^{-1}-Q_{\Phi\Phi}^{-1}\right\|_{\text{sp}}=O_P\left(\kappa/n^{1/2}\right)$ by Lemma A.1(iii). It follows that $\left\|\widetilde{Q}_{n,\Phi\Phi}^{-1}-Q_{\Phi\Phi}^{-1}\right\|_{\text{sp}}=O_P\left(\kappa^{1/2}\varsigma_{0\kappa}\nu_{1n}+\varsigma_{0\kappa}\varsigma_{2\kappa}\nu_{1n}^2\right)$.

(v) Noting that $\left\|n^{-1}\sum_{i=1}^n\left(\widetilde{\Phi}_i-\Phi_i\right)e_i\right\|^2=\sum_{l=1}^{d_x}\left\|n^{-1}\sum_{i=1}^n\left[p^\kappa\left(\widetilde{U}_{li}\right)-p^\kappa\left(U_{li}\right)\right]e_i\right\|^2$, the result follows from Lemma A.3(v).

(vi) Let $\delta_i\equiv\bar{g}\left(X_i,Z_{1i},U_i\right)-\Phi_i'\boldsymbol{\beta}$. By triangle inequality, Assumption A2(v), Jensen inequality and (i), we have $\left\|n^{-1}\sum_{i=1}^n\left(\widetilde{\Phi}_i-\Phi_i\right)\delta_i\right\|\leq O_P\left(\kappa^{-\gamma}\right)n^{-1}\sum_{i=1}^n\left\|\widetilde{\Phi}_i-\Phi_i\right\|=O_P\left(\kappa^{-\gamma}\right)O_P\left(\varsigma_{1\kappa}\nu_{1n}\right)=O_P\left(\kappa^{-\gamma}\varsigma_{1\kappa}\nu_{1n}\right)$. ■

Lemma A.5 *Let* $\xi_n\equiv n^{-1}\sum_{i=1}^n\Phi_ie_i$ *and* $\zeta_n\equiv n^{-1}\sum_{i=1}^n\Phi_i\left[\bar{g}\left(X_i,Z_{1i},U_i\right)-\Phi_i'\boldsymbol{\beta}\right]$. *Suppose Assumptions A1-A3 hold. Then*

(i) $\|\xi_n\|=O_P(\kappa^{1/2}/n^{1/2})$;

(iii) $\|\zeta_n\|=O_P(\kappa^{-\gamma})$;

(iii) $\left\|Q_{\Phi\Phi}^{-1}n^{-1}\sum_{i=1}^n\Phi_i\sum_{l=1}^{d_x}\dot{p}^\kappa\left(U_{li}\right)'\boldsymbol{\beta}_{d_x+d_1+l}\left(\widetilde{U}_{li}-U_{li}\right)\right\|=O_P\left(\nu_{1n}\right)$ *for* $l=1,...,d_x$.

Proof. The proof of (i)-(ii) is analogous to that of Lemma A.2 (i)-(ii), respectively. Noting that $\left\|Q_{\Phi\Phi}^{-1}\right\|_{\text{sp}}=O\left(1\right)$ by Assumption A2(ii), we can prove (iii) by showing that $\|T_{nl}\|=O_P(\nu_{1n})$, where $T_{nl}=n^{-1}\sum_{i=1}^n\Phi_i\delta_{li}(\widetilde{U}_{li}-U_{li})$ where $\delta_{li}=\dot{p}^\kappa\left(U_{li}\right)'\boldsymbol{\beta}_{d_x+d_1+l}$. By triangle inequality and Assumptions A1(ii) and A2(iii) and (v)

$$\begin{aligned}c_{\delta_l} &\equiv \max_{1\leq i\leq n}\|\delta_{li}\|\leq\sup_{u_l\in\mathcal{U}_l}\left\|\dot{g}_{d_x+d_1+l}\left(u_l\right)-\dot{p}^\kappa\left(u_l\right)'\boldsymbol{\beta}_{d_x+d_1+l}\right\|+\sup_{u_l\in\mathcal{U}_l}\left\|\dot{g}_{d_x+d_1+l}\left(u_l\right)\right\| \\ &= O\left(\kappa^{-\gamma}\right)+O\left(1\right)=O\left(1\right).\end{aligned}$$

By (A.2), $T_{nl}=n^{-1}\sum_{i=1}^n\Phi_i\delta_{li}(\widetilde{U}_{li}-U_{li})=\sum_{s=1}^5n^{-1}\sum_{i=1}^n\Phi_i\delta_{li}u_{sl,i}=\sum_{s=1}^5T_{nl,s}$, say.

Let $\eta_{nlk}\equiv n^{-1}\sum_{i=1}^n\delta_{li}\Phi_ip^{\kappa_1}\left(Z_{1k,i}\right)'$ and $\bar{\eta}_{lk}=E\left(\eta_{nlk}\right)$. Then $\|\eta_{nlk}-\bar{\eta}_{lk}\|=O_P((\kappa\kappa_1/n)^{1/2})$ by Chebyshev inequality and

$$\|\bar{\eta}_k\|_{\text{sp}}^2=\left\|E\left[\delta_{li}\Phi_ip^{\kappa_1}\left(Z_{1k,i}\right)'\right]\right\|_{\text{sp}}^2\leq c_\delta^2\lambda_{\max}\left(M\right)=O\left(1\right),$$

where $M\equiv E\left[\Phi_ip^{\kappa_1}\left(Z_{1k,i}\right)'\right]E\left[p^{\kappa_1}\left(Z_{1k,i}\right)\Phi_i'\right]$ and we use the fact that M has bounded largest eigenvalue. To see the last point, first note that for $\kappa_1\leq\kappa$, $E\left[\Phi_ip^{\kappa_1}\left(Z_{1k,i}\right)'\right]$ is a submatrix of $A\equiv E\left(\Phi_i\Phi_i'\right)$ which has bounded largest eigenvalue. Partition A as follows

$$A=\begin{bmatrix}A_{11} & A_{12} & A_{13}\\ A_{21} & A_{22} & A_{23}\\ A_{31} & A_{32} & A_{33}\end{bmatrix}$$

where $A_{ij} = A'_{ji}$ for i, $j = 1, 2, 3$ and $E\left[\Phi_i p^{\kappa_1}(Z_{1k,i})'\right] = \left[\begin{array}{ccc} A'_{12} & A_{22} & A'_{32} \end{array}\right]'$. Then

$$M = \left[\begin{array}{ccc} A_{12}A'_{12} & A_{12}A_{22} & A_{12}A'_{32} \\ A_{22}A'_{12} & A_{22}A_{22} & A_{22}A'_{32} \\ A_{32}A'_{12} & A_{32}A_{22} & A_{32}A'_{32} \end{array}\right].$$

By Thompson and Freede (1970, Theorem 2), $\lambda_{\max}(M) \leq \lambda_{\max}(A_{12}A'_{12}) + \lambda_{\max}(A_{22}A'_{22}) + \lambda_{\max}(A_{32}A'_{32})$. By Fact 8.9.3 in Bernstein (2005), the positive definiteness of A ensures that both $A_{12}A'_{12}$ and $A_{32}A'_{32}$ have finite maximum eigenvalues as both $\left[\begin{array}{cc} A_{11} & A_{12} \\ A_{21} & A_{22} \end{array}\right]$ and $\left[\begin{array}{cc} A_{22} & A_{23} \\ A_{32} & A_{33} \end{array}\right]$ are also positive definite. In addition, $\lambda_{\max}(A_{22}A_{22}) = [\lambda_{\max}(A_{22})]^2$ is finite as A has bounded maximum eigenvalue. It follows that $\lambda_{\max}(M) = O(1)$. Consequently, $\|\eta_{nlk}\| = O_P(1 + (\kappa\kappa_1/n)^{1/2}) = O_P(1)$.

Analogously, noting that 1 is the first element of Φ_i, we can show that $\left\|n^{-1}\sum_{i=1}^n \Phi_i\delta_i\right\|_{\text{sp}} = O_P(1 + (\kappa/n)^{1/2}) = O_P(1)$. It follows that

$$\begin{aligned} \|T_{nl,1}\| &= \left\|n^{-1}\sum_{i=1}^n \Phi_i\delta_{li}\right\|_{\text{sp}} |\widetilde{\mu}_l - \mu_l| = O_P(1)\, O_P\left(n^{-1/2}\right) = O_P\left(n^{-1/2}\right), \\ \|T_{nl,2} + T_{nl,4}\| &\leq \sum_{k=1}^{d_1} \|\eta_{nlk}\| \, \|\mathbb{S}_{1k}\|_{\text{sp}} (\|a_{1l}\| + \|a_{2l}\|) = O_P(1)\, O(1)\, O(\nu_{1n}) = O(\nu_{1n}), \end{aligned}$$

and $\|T_{nl,3} + T_{nl,5}\| = O(\nu_{1n})$ by the same token. Thus we have shown that $\|T_{nl}\| = O_P(\nu_{1n})$. ■

Lemma A.6 *Let* $c \equiv (c_1, c_2)'$ *be an arbitrary* 2×1 *nonrandom vector such that* $\|c\| = 1$. *Suppose that Assumptions A1-A5 hold. Then for* $l = 2, ..., d_x$

(*i*) $S_{2nl}(x_1) \equiv n^{-1/2}h^{1/2}\sum_{i=1}^n K_{ix_1}c'H^{-1}X^*_{1i}(x_1)(\widetilde{U}_{li} - U_{li})\dot{p}^{\kappa}(U_{li}) = h^{1/2}\varsigma_{1\kappa}O_P(1 + n^{1/2}\kappa_1^{-\gamma})$ *uniformly in* x_1;

(*ii*) $S_{2nl}(x_1) \equiv n^{-1/2}h^{1/2}\sum_{i=1}^n K_{ix_1}\left|c'H^{-1}X^*_{1i}(x_1)\right|(\widetilde{U}_{li} - U_{ki})^2 = n^{1/2}h^{1/2}O_P(\nu_{1n}^2)$ *uniformly in* x_1.

Proof. (*i*) Let $\eta_{nl}(x_1) \equiv n^{-1}\sum_{i=1}^n K_{ix}c'H^{-1}X^*_{1i}(x_1)\, p^{\kappa}(X_{li})$ and $\bar{\eta}_l(x_1) \equiv E[\eta_{nl}(x_1)]$. By straightforward moment calculations and Chebyshev inequality, we have $\eta_{nl}(x_1) = \bar{\eta}_l(x_1) + r_{\eta l}(x_1)$ where $\|r_{\eta l}(x_1)\| = O_P(\kappa^{1/2}n^{-1/2}h^{-1/2})$. In fact, $\sup_{x_1 \in \mathcal{X}_1}\|r_{\eta l}(x_1)\| = O_P(\kappa^{1/2}(nh/\log n)^{-1/2})$ with a simple application of Bernstein inequality for independent observations [see, e.g., Serfling (1980, p. 95)]. As an aside, note that the proof of Lemma 7 in Horowitz and Mammen (2004) contains various errors as they ignore the fact that κ is diverging to infinity as $n \to \infty$. Note that for $l = 2, ..., d_x$,

$$\begin{aligned} \bar{\eta}_l(x_1) &= E\left[K_h(X_{1i} - x_1)\, c'H^{-1}X^*_{1i}(x_1)\, p^{\kappa}(X_{li})\right] \\ &= \int K(v)(c_1 + c_2 v)\, p^{\kappa}(x_l)\, f_{1l}\left(x_1 + h^{1/2}v, x_l\right) dv dx_l \\ &= c_1\int f_{1l}(x_1, x_l)\, p^{\kappa}(x_l)\, dx_l + c_1\int K(v)\left[f_{1l}(x_1 + hv, x_l) - f_{1l}(x_1, x_l)\right] p^{\kappa}(x_l)\, dx_l \\ &\quad + c_2\int K(v)\, v\left[f_{1l}(x_1 + hv, x_l) - f_{1l}(x_1, x_l)\right] dv p^{\kappa}(x_l)\, dx_l \\ &\equiv c_1\bar{\eta}_{1l}(x_1) + c_1\bar{\eta}_{2l}(x_1) + c_2\bar{\eta}_{3l}(x_1). \end{aligned}$$

As in Horowitz and Mammen (2004, p. 2435), in view of the fact that the components of $\bar{\eta}_{1l}(x_1)$ are the Fourier coefficients of a function that is bounded uniformly over $\mathcal{X}_1$, we have $\sup_{x_1 \in \mathcal{X}_1}\|\bar{\eta}_{1l}(x_1)\|^2 = O(1)$. In addition, using Assumptions A2(v) and A3(i), we can readily show that $\sup_{x_1 \in \mathcal{X}_1}\|\bar{\eta}_{2l}(x_1)\| = O_P\left(\kappa^{1/2}h^2\right)$ and $\sup_{x_1 \in \mathcal{X}_1}\|\bar{\eta}_{3l}(x_1)\| = O_P\left(\kappa^{1/2}h\right)$. It follows that $\sup_{x_1 \in \mathcal{X}_1}\|\bar{\eta}_l(x_1)\| = O_P\left(1 + \kappa^{1/2}h\right) = O_P(1)$ under Assumption A5(ii) and $\sup_{x_1 \in \mathcal{X}_1}\|\eta_{nl}(x_1)\| = O_P(1)$.

By (A.2), $S_{1nl}(x_1) = -\sum_{s=1}^{5} n^{-1/2}h^{1/2}\sum_{i=1}^{n} K_{ix_1} c'H^{-1}X_{1i}^{*}(x_1)\,\dot{p}^{\kappa}(U_{li})\,u_{sl,i} \equiv -\sum_{s=1}^{5} S_{1nl,s}(x_1)$, say. Noting that $S_{1nl,1}(x_1) = n^{1/2}h^{1/2}\eta_{nl}(x_1)(\widetilde{\mu}_l - \mu_l)$, we have

$$\sup_{x_1\in\mathcal{X}_1}\|S_{1nl,1}(x_1)\| \le n^{1/2}h^{1/2}\sup_{x_1\in\mathcal{X}_1}\|\eta_{nl}(x_1)\| \ \ |\widetilde{\mu}_l-\mu_l| = n^{1/2}h^{1/2}O_P(1)\,O_P\left(n^{-1/2}\right) = o_P(1).$$

Next, note that $S_{1nl,2}(x_1) = \sum_{k=1}^{d_1} S_{1nl,2k}(x_1)$ where $S_{1nl,2k}(x_1) = n^{-1/2}h^{1/2}\sum_{i=1}^{n} K_{ix_1}c'H^{-1}X_{1i}^{*}(x_1)$ $\dot{p}^{\kappa}(U_{li})\,p^{\kappa_1}(Z_{1k,i})'\,\mathbb{S}_{1k}a_{1l}$. We decompose $S_{1nl,2k}$ as follows:

$$\begin{aligned}
S_{1nl,2k}(x_1) &= n^{-1/2}h^{1/2}\sum_{i=1}^{n} K_{ix_1}c'H^{-1}X_{1i}^{*}\dot{p}^{\kappa}(U_{li})\,p^{\kappa_1}(Z_{1k,i})'\,\mathbb{S}_{1k}Q_{n,PP}^{-1}\xi_{nl}\\
&= n^{1/2}h^{1/2}\psi_{nkl}(x_1)\,\mathbb{S}_{1k}Q_{PP}^{-1}\xi_{nl} + n^{1/2}h^{1/2}\psi_{nkl}(x_1)\,\mathbb{S}_{1k}\left(Q_{n,PP}^{-1}-Q_{PP}^{-1}\right)\xi_{nl}\\
&\equiv S_{1nl,2k1}(x_1) + S_{1nl,2k2}(x_1)\,,\ \text{say.}
\end{aligned}$$

where $\psi_{nkl}(x_1) \equiv n^{-1}\sum_{i=1}^{n} K_{ix_1}c'H^{-1}X_{1i}^{*}(x_1)\,\dot{p}^{\kappa}(U_{li})\,p^{\kappa_1}(Z_{1k,i})'$. Let $\bar{\psi}_{kl}(x_1) \equiv E[\psi_{nkl}(x_1)]$. As in the analysis of $\eta_{nl}(x_1)$, we can show that $\sup_{x_1\in\mathcal{X}_1}\left\|\bar{\psi}_{kl}(x_1)\right\|_{\mathrm{sp}} = O_P(\varsigma_{1\kappa})$ and $\sup_{x_1\in\mathcal{X}_1}\left\|\psi_{nkl}(x_1)-\bar{\psi}_{kl}(x_1)\right\|_{\mathrm{sp}}$ $\equiv O_P((\kappa_1\kappa\log n/n)^{-1/2})$. It follows that $\sup_{x_1\in\mathcal{X}_1}\|\psi_{nkl}(x_1)\|_{\mathrm{sp}} = O_P(\varsigma_{1\kappa}+(\kappa_1\kappa\log n/n)^{-1/2}) = O_P(\varsigma_{1\kappa})$ under Assumption A5(i). Then following the analysis of $B_{nl,1}(x_1)$ in the proof of Theorem 3.2, we can show that $\|S_{1nl,2k1}(x_1)\| = O_P\left(h^{1/2}\varsigma_{1\kappa}\right)$ uniformly in x_1. In addition,

$$\begin{aligned}
\sup_{x_1\in\mathcal{X}_1}\|S_{1nl,2k2}(x_1)\| &\le n^{1/2}h^{1/2}\sup_{x_1\in\mathcal{X}_1}\|\psi_{nkl}(x_1)\|_{\mathrm{sp}}\|\mathbb{S}_{1k}\|_{\mathrm{sp}}\left\|Q_{n,PP}^{-1}-Q_{PP}^{-1}\right\|_{\mathrm{sp}}\|\xi_{nl}\|\\
&= n^{1/2}h^{1/2}O_P(\varsigma_{1\kappa})\,O(1)\,O_P(\kappa_1 n^{-1/2})O_P(\kappa_1^{1/2}n^{-1/2}) = O_P(\varsigma_{1\kappa}\kappa_1^{3/2}n^{-1/2}h^{1/2}).
\end{aligned}$$

It follows that $\sup_{x_1\in\mathcal{X}_1}\|S_{1nl,2k}(x_1)\| = O_P(h^{1/2}\varsigma_{1\kappa}) + O_P(\varsigma_{1\kappa}\kappa_1^{3/2}n^{-1/2}h^{1/2}) = O_P(h^{1/2}\varsigma_{1\kappa})$ under Assumption A5(i). Analogously,

$$\begin{aligned}
\sup_{x_1\in\mathcal{X}_1}\|S_{1nl,4}(x_1)\| &\le \sum_{k=1}^{d_1} n^{1/2}h^{1/2}\sup_{x_1\in\mathcal{X}_1}\|\psi_{nkl}(x_1)\|\,\|\mathbb{S}_{1k}\|_{\mathrm{sp}}\|a_{2l}\|\\
&= n^{1/2}h^{1/2}O_P(\varsigma_{1\kappa})O_P(\kappa_1^{-\gamma}) = O_P(n^{1/2}h^{1/2}\varsigma_{1\kappa}\kappa_1^{-\gamma}).
\end{aligned}$$

By the same token, we can show that $\sup_{x_1\in\mathcal{X}_1}\|S_{1nl,3}(x_1)\| = O_P\left(h^{1/2}\varsigma_{1\kappa}\right)$ and $\sup_{x_1\in\mathcal{X}_1}\|S_{1nl,5}(x_1)\| = O_P(n^{1/2}h^{1/2}\varsigma_{1\kappa}\kappa_1^{-\gamma})$. It follows that $\sup_{x_1\in\mathcal{X}_1}\|S_{1n}(x_1)\| = h^{1/2}\varsigma_{1\kappa}O_P(1+n^{1/2}\kappa_1^{-\gamma})$.

(ii) By (A.2) and Cauchy-Schwarz inequality

$$S_{2nl}(x_1) \le 5\sum_{s=1}^{5} n^{-1/2}h^{1/2}\sum_{i=1}^{n} K_{ix_1}\left|c'H^{-1}X_{1i}^{*}(x_1)\right|u_{sl,i}^{2} \equiv 5\sum_{s=1}^{5} S_{2nl,s},\ \text{say.}$$

It is easy to show that $\sup_{x_1\in\mathcal{X}_1} S_{2nl,1}(x_1) = O_P(n^{-1/2}h^{1/2})$. Note that $S_{2nl,2}(x_1) = n^{-1/2}h^{1/2}\sum_{i=1}^{n} K_{ix_1}$ $\times\left|c'H^{-1}X_{1i}^{*}(x_1)\right|u_{2l,i}^{2} \le d_1\sum_{k=1}^{d_1} S_{2nl,2k}(x_1)$, where

$$\begin{aligned}
S_{2nl,2k}(x_1) &= n^{-1/2}h^{1/2}\sum_{i=1}^{n} K_{ix_1}\left|c'H^{-1}X_{1i}^{*}(x_1)\right|p^{\kappa_1}(Z_{1k,i})'\,\mathbb{S}_{1k}a_{1l}a_{1l}'\mathbb{S}_{1k}'p^{\kappa_1}(Z_{1k,i})\\
&= n^{1/2}h^{1/2}\mathrm{tr}\left(\mathbb{S}_{1k}a_{1l}a_{1l}'\mathbb{S}_{1k}'v_{nlk}(x_1)\right)
\end{aligned}$$

and $v_{nlk}(x_1) \equiv n^{-1}\sum_{i=1}^{n} K_{ix_1}\left|c'H^{-1}X_{1i}^{*}\right|p^{\kappa_1}(Z_{1k,i})\,p^{\kappa_1}(Z_{1k,i})'$. As in the analysis of $\eta_{nl}(x_1)$, we can show that $\sup_{x_1\in\mathcal{X}_1}\|v_{nlk}(x_1)\|_{\mathrm{sp}} = O_P(1)$. By the fact $\mathrm{tr}(AB) \le \lambda_{\max}(A)\mathrm{tr}(B)$ and $\|B\|_{\mathrm{sp}} = \lambda_{\max}(B)$ for any symmetric matrix A and conformable positive-semidefinite matrix B.

$$\begin{aligned}
S_{2nl,2k}(x_1) &\le n^{1/2}h^{1/2}\mathrm{tr}\left(\mathbb{S}_{1k}a_{1l}a_{1l}'\mathbb{S}_{1k}'\right)\|v_{nlk}(x_1)\|_{\mathrm{sp}} = n^{1/2}h^{1/2}\,\|\mathbb{S}_{1k}a_{1l}\|_{\mathrm{sp}}^{2}\,\|v_{nlk}(x_1)\|_{\mathrm{sp}}\\
&\le n^{1/2}h^{1/2}\,\|\mathbb{S}_{1k}\|_{\mathrm{sp}}^{2}\,\|a_{1l}\|_{\mathrm{sp}}^{2}\,\|v_{nlk}(x_1)\|_{\mathrm{sp}}\\
&= n^{1/2}h^{1/2}O(1)\,O_P\left(\kappa_1 n^{-1}\right)O_P(1) = O_P\left(\kappa_1 n^{-1/2}h^{1/2}\right)\ \text{uniformly in } x_1.
\end{aligned}$$

It follows that $\sup_{x_1\in\mathcal{X}_1} S_{2nl,2}(x_1) = O_P\left(\kappa_1 n^{-1/2}h^{1/2}\right)$. Similarly, uniformly in x_1

$$\begin{aligned} S_{2nl,4}(x_1) &\leq n^{1/2}h^{1/2}\left\|\mathbb{S}_{1k}\right\|_{\text{sp}}^2 \left\|a_{2l}\right\|_{\text{sp}}^2 \left\|v_{nlk}(x_1)\right\|_{\text{sp}} \\ &= n^{1/2}h^{1/2}O(1)\,O_P(\kappa_1^{-2\gamma})O_P(1) = O_P\left(n^{1/2}\kappa_1^{-2\gamma}h^{1/2}\right). \end{aligned}$$

By the same token, $S_{2nl,3}(x_1) = O_P(\kappa_1 n^{-1/2}h^{1/2})$ and $S_{2nl,5}(x_1) = O_P(n^{1/2}\kappa_1^{-2\gamma}h^{1/2})$ uniformly in x_1. Consequently, $\sup_{x_1\in\mathcal{X}_1} S_{2nl}(x_1) = n^{1/2}h^{1/2}O_P(\nu_{1n}^2)$. ■

Proof of Theorem 3.1. (i) Noting that $Y_i = \bar{g}(X_i, Z_{1i}, U_i) + e_i = \widetilde{\Phi}_i'\boldsymbol{\beta} + e_i + \left[\bar{g}(X_i, Z_{1i}, U_i) - \widetilde{\Phi}_i'\boldsymbol{\beta}\right]$, we have

$$\begin{aligned} \widetilde{\boldsymbol{\beta}} - \boldsymbol{\beta} &= \widetilde{Q}_{n,\Phi\Phi}^{-1} n^{-1}\sum_{i=1}^n \widetilde{\Phi}_i Y_i - \boldsymbol{\beta} = \widetilde{Q}_{n,\Phi\Phi}^{-1} n^{-1}\sum_{i=1}^n \widetilde{\Phi}_i e_i + \widetilde{Q}_{n,\Phi\Phi}^{-1} n^{-1}\sum_{i=1}^n \widetilde{\Phi}_i \left[\bar{g}(X_i, Z_{1i}, U_i) - \widetilde{\Phi}_i'\boldsymbol{\beta}\right] \\ &= \widetilde{Q}_{n,\Phi\Phi}^{-1}\xi_n + \widetilde{Q}_{n,\Phi\Phi}^{-1}\zeta_n + \widetilde{Q}_{n,\Phi\Phi}^{-1} n^{-1}\sum_{i=1}^n \Phi_i\left(\Phi_i - \widetilde{\Phi}_i\right)'\boldsymbol{\beta} + \widetilde{Q}_{n,\Phi\Phi}^{-1} n^{-1}\sum_{i=1}^n \left(\widetilde{\Phi}_i - \Phi_i\right)e_i \\ &\quad + \widetilde{Q}_{n,\Phi\Phi}^{-1} n^{-1}\sum_{i=1}^n \left(\widetilde{\Phi}_i - \Phi_i\right)\left[\bar{g}(X_i, Z_{1i}, U_i) - \Phi_i'\boldsymbol{\beta}\right] - \widetilde{Q}_{n,\Phi\Phi}^{-1} n^{-1}\sum_{i=1}^n \left(\widetilde{\Phi}_i - \Phi_i\right)\left(\widetilde{\Phi}_i - \Phi_i\right)'\boldsymbol{\beta} \\ &\equiv b_{1n} + b_{2n} + b_{3n} + b_{4n} + b_{5n} - b_{6n}, \text{ say.} \end{aligned}$$

Note that $b_{1n} = Q_{\Phi\Phi}^{-1}\xi_n + r_{1n}$, where $r_{1n} = \left(\widetilde{Q}_{n,\Phi\Phi}^{-1} - Q_{\Phi\Phi}^{-1}\right)\xi_n$ satisfies $\|r_{1n}\| \leq \left\|\widetilde{Q}_{n,\Phi\Phi}^{-1} - Q_{\Phi\Phi}^{-1}\right\|_{\text{sp}}\|\xi_n\|_{\text{sp}} = O_P\left[(\kappa^{1/2}\varsigma_{0\kappa}\nu_{1n} + \varsigma_{0\kappa}\varsigma_{2\kappa}\nu_{1n}^2)\kappa^{1/2}n^{-1/2}\right]$ by Lemmas A.4(iv) and A.5(i). Similarly, $b_{2n} = Q_{\Phi\Phi}^{-1}\zeta_n + r_{2n}$, where $r_{2n} = \left(\widetilde{Q}_{n,\Phi\Phi}^{-1} - Q_{\Phi\Phi}^{-1}\right)\zeta_n$ satisfies $\|r_{2n}\| \leq \left\|\widetilde{Q}_{n,\Phi\Phi}^{-1} - Q_{\Phi\Phi}^{-1}\right\|\|\zeta_n\|_{\text{sp}} = O_P\left[(\kappa^{1/2}\varsigma_{0\kappa}\nu_{1n} + \varsigma_{0\kappa}\varsigma_{2\kappa}\nu_{1n}^2)\kappa^{-\gamma}\right]$ by Lemmas A.4(iv) and A.5(ii). Next, we decompose b_{3n} as follows:

$$b_{3n} = Q_{\Phi\Phi}^{-1} n^{-1}\sum_{i=1}^n \Phi_i\left(\Phi_i - \widetilde{\Phi}_i\right)'\boldsymbol{\beta} + \left(\widetilde{Q}_{n,\Phi\Phi}^{-1} - Q_{\Phi\Phi}^{-1}\right)n^{-1}\sum_{i=1}^n \Phi_i\left(\Phi_i - \widetilde{\Phi}_i\right)'\boldsymbol{\beta} \equiv b_{3n,1} + b_{3n,2}, \text{ say.}$$

We further decompose $b_{3n,1}$ as follows:

$$\begin{aligned} b_{3n,1} &= -Q_{\Phi\Phi}^{-1} n^{-1}\sum_{i=1}^n \Phi_i \sum_{l=1}^{d_x}\left[p^\kappa\left(\widetilde{U}_{li}\right) - p^\kappa(U_{li})\right]'\boldsymbol{\beta}_{d_x+d_1+l} \\ &= \sum_{l=1}^{d_x} Q_{\Phi\Phi}^{-1} n^{-1}\sum_{i=1}^n \Phi_i \dot{p}^\kappa\left(U_{li}^\dagger\right)'\boldsymbol{\beta}_{d_x+d_1+l}\left(U_{li} - \widetilde{U}_{li}\right) \\ &= \sum_{l=1}^{d_x} Q_{\Phi\Phi}^{-1} n^{-1}\sum_{i=1}^n \Phi_i \dot{g}_{d_x+d_1+l}(U_{li})\left(U_{li} - \widetilde{U}_{li}\right) \\ &\quad + \sum_{l=1}^{d_x} Q_{\Phi\Phi}^{-1} n^{-1}\sum_{i=1}^n \Phi_i\left[\dot{g}_{d_x+d_1+l}\left(U_{li}^\dagger\right) - \dot{g}_{d_x+d_1+l}(U_{li})\right]\left(U_{li} - \widetilde{U}_{li}\right) \\ &\quad + \sum_{l=1}^{d_x} Q_{\Phi\Phi}^{-1} n^{-1}\sum_{i=1}^n \Phi_i\left[\dot{p}^\kappa\left(U_{li}^\dagger\right)'\boldsymbol{\beta}_{d_x+d_1+l} - \dot{g}_{d_x+d_1+l}\left(U_{li}^\dagger\right)\right]\left(U_{li} - \widetilde{U}_{li}\right) \\ &\equiv \sum_{l=1}^{d_x} b_{3n,11l} + \sum_{l=1}^{d_x} b_{3n,12l} + \sum_{l=1}^{d_x} b_{3n,13l}, \text{ say,} \end{aligned}$$

where $U_{li}^{\dagger}$ lies between $\widetilde{U}_{li}$ and U_{li}. Noting that $|\dot{g}_{d_x+d_1+l}\left(U_{li}^{\dagger}\right)-\dot{g}_{d_x+d_1+l}\left(U_{li}\right)|\leq c_{\dot{g}}|\widetilde{U}_{li}-U_{li}|$ where $c_{\dot{g}}=\max_{1\leq l\leq d_x}\max_{u_l\in\mathcal{U}_l}|\dot{g}_{d_x+d_1+l}\left(u_l\right)|=O\left(1\right)$ by Assumptions A1(ii) and A2(iii),

$$\|b_{3n,12l}\|\leq c_{\dot{g}}n^{-1}\sum_{i=1}^{n}\|\Phi_i\|\left(U_{li}-\widetilde{U}_{li}\right)^2=\varsigma_{0\kappa}O_P\left(\nu_{1n}^2\right)$$

by Lemma A.3(i). By Assumption A2(ii), Cauchy-Schwarz inequality and Lemma A.3(i)

$$\begin{aligned}\|b_{3n,13l}\| &\leq O\left(\kappa^{-\gamma}\right)\left\|Q_{\Phi\Phi}^{-1}\right\|_{\text{sp}}n^{-1}\sum_{i=1}^{n}\|\Phi_i\|\left|U_{li}-\widetilde{U}_{li}\right| \\ &\leq O\left(\kappa^{-\gamma}\right)\left\|Q_{\Phi\Phi}^{-1}\right\|_{\text{sp}}\left\{n^{-1}\sum_{i=1}^{n}\|\Phi_i\|^2\right\}^{1/2}\left\{n^{-1}\sum_{i=1}^{n}\left(U_{li}-\widetilde{U}_{li}\right)^2\right\}^{1/2} \\ &= O\left(\kappa^{-\gamma}\right)O\left(1\right)O\left(\kappa^{1/2}\right)O_P\left(\nu_{1n}\right)=\kappa^{-\gamma+1/2}O_P\left(\nu_{1n}\right).\end{aligned}$$

By Lemma A.5(iii), $\|b_{3n,11l}\|=O_P(\nu_{1n})$ which dominates both $\|b_{3n,12l}\|$ and $\|b_{3n,13l}\|$. Thus $\|b_{3n,2}\|\leq\left\|\widetilde{Q}_{n,\Phi\Phi}^{-1}-Q_{\Phi\Phi}^{-1}\right\|_{\text{sp}}O_P(\nu_{1n})=O_P[(\kappa^{1/2}\varsigma_{0\kappa}\nu_{1n}+\varsigma_{0\kappa}\varsigma_{2\kappa}\nu_{1n}^2)\nu_{1n}]$. It follows that $b_{3n}=\sum_{l=1}^{d_x}Q_{\Phi\Phi}^{-1}n^{-1}\sum_{i=1}^{n}\Phi_i$ $\times\dot{g}_{d_x+d_1+l}\left(U_{li}\right)\left(U_{li}-\widetilde{U}_{li}\right)+\bar{b}_{3n}$, where $\left\|\bar{b}_{3n}\right\|=O_P[(\kappa^{1/2}\varsigma_{0\kappa}\nu_{1n}+\varsigma_{0\kappa}\varsigma_{2\kappa}\nu_{1n}^2)\nu_{1n}]$. By Lemmas A.4$(v)$-$(vi)$, $\|b_{4n}\|=O_P\left(n^{-1/2}\varsigma_{1\kappa}\nu_{1n}\right)$ and

$$\|b_{5n}\|\leq\left\|\widetilde{Q}_{n,\Phi\Phi}^{-1}\right\|_{\text{sp}}\left\|n^{-1}\sum_{i=1}^{n}\left(\widetilde{\Phi}_i-\Phi_i\right)\left[\bar{g}\left(X_i,Z_{1i},U_i\right)-\Phi_i'\boldsymbol{\beta}\right]\right\|=O_P\left(\kappa^{-\gamma}\varsigma_{1\kappa}\nu_{1n}\right),$$

where we use the fact that $\left\|\widetilde{Q}_{n,\Phi\Phi}^{-1}\right\|_{\text{sp}}\leq\left\|\widetilde{Q}_{n,\Phi\Phi}^{-1}-Q_{\Phi\Phi}^{-1}\right\|_{\text{sp}}+\left\|Q_{\Phi\Phi}^{-1}\right\|_{\text{sp}}=o_P\left(1\right)+O\left(1\right)=O_P\left(1\right)$. For b_{6n}, we have by Taylor expansion and triangle inequality that

$$\begin{aligned}\|b_{6n}\| &\leq \sum_{l=1}^{d_x}\left\|\widetilde{Q}_{n,\Phi\Phi}^{-1}n^{-1}\sum_{i=1}^{n}\left(\widetilde{\Phi}_i-\Phi_i\right)\left[p^{\kappa}\left(\widetilde{U}_{li}\right)-p^{\kappa}\left(U_{li}\right)\right]'\boldsymbol{\beta}_{d_x+d_1+l}\right\| \\ &= \sum_{l=1}^{d_x}\left\|\widetilde{Q}_{n,\Phi\Phi}^{-1}n^{-1}\sum_{i=1}^{n}\left(\widetilde{\Phi}_i-\Phi_i\right)p^{\kappa}\left(U_{li}^{\dagger}\right)'\boldsymbol{\beta}_{d_x+d_1+l}\left(\widetilde{U}_{li}-U_{li}\right)\right\| \\ &\leq \sum_{l=1}^{d_x}\left\|\widetilde{Q}_{n,\Phi\Phi}^{-1}\right\|_{\text{sp}}\left\|n^{-1}\sum_{i=1}^{n}\left(\widetilde{\Phi}_i-\Phi_i\right)\dot{g}_{d_x+d_1+l}\left(U_{li}^{\dagger}\right)\left(\widetilde{U}_{li}-U_{li}\right)\right\| \\ &\quad +\sum_{l=1}^{d_x}\left\|\widetilde{Q}_{n,\Phi\Phi}^{-1}\right\|_{\text{sp}}\left\|n^{-1}\sum_{i=1}^{n}\left(\widetilde{\Phi}_i-\Phi_i\right)\left[p^{\kappa}\left(U_{li}^{\dagger}\right)'\boldsymbol{\beta}_{d_x+d_1+l}-\dot{g}_{d_x+d_1+l}\left(U_{li}^{\dagger}\right)\right]\left(\widetilde{U}_{li}-U_{li}\right)\right\| \\ &\equiv \sum_{l=1}^{d_x}b_{6nl,1}+\sum_{l=1}^{d_x}b_{6nl,2},\text{ say.}\end{aligned}$$

By the triangle inequality, Lemmas A.3(i) and A.4(i),

$$\begin{aligned}b_{6nl,1} &\leq c_{\dot{g}}\left\|\widetilde{Q}_{n,\Phi\Phi}^{-1}\right\|_{\text{sp}}\left\{n^{-1}\sum_{i=1}^{n}\left\|\widetilde{\Phi}_i-\Phi_i\right\|^2\right\}^{1/2}\left\{n^{-1}\sum_{i=1}^{n}\left(\widetilde{U}_{li}-U_{li}\right)^2\right\}^{1/2} \\ &= O_P(1)O_P\left(\varsigma_{1\kappa}\nu_{1n}\right)O_P\left(\nu_{1n}\right)=O_P\left(\varsigma_{1\kappa}\nu_{1n}^2\right).\end{aligned}$$

Similarly, we can show that $b_{6nl,2}=\kappa^{-\gamma}O_P\left(\varsigma_{1\kappa}\nu_{1n}^2\right)$ by Assumption A2(v) and Lemmas A.3(i) and A.4(i). It follows that $\|b_{6n}\|=O_P\left(\varsigma_{1\kappa}\nu_{1n}^2\right)$. Combining the above results yield the conclusion in (i).

(ii) Noting that $\left\|Q_{\Phi\Phi}^{-1}\xi_n\right\| \leq \left\|Q_{\Phi\Phi}^{-1}\right\|_{\rm sp}\|\xi_n\| = O_P\left(\kappa^{1/2}/n^{1/2}\right)$ and $||Q_{\Phi\Phi}^{-1}\zeta_n|| \leq \left\|Q_{\Phi\Phi}^{-1}\right\|_{\rm sp}\|\zeta_n\| = O_P\left(\kappa^{-\gamma}\right)$ by Lemmas A.5(i)-(ii), the result in part (ii) follows from part (i), Lemma A.4 and the fact that $\|\mathbf{R}_{n,\beta}\| = O_P(\nu_{1n})$ under Assumption A5(i)

(iii) By (ii) and Assumptions A2(v) , $\sup_{\mathbf{w}\in\mathcal{W}}\left|\widetilde{g}\left(\mathbf{w}\right)-\bar{g}\left(\mathbf{w}\right)\right| = \sup_{\mathbf{w}\in\mathcal{W}}|\Phi\left(\mathbf{w}\right)'(\widetilde{\boldsymbol{\beta}}-\boldsymbol{\beta}) + [\boldsymbol{\beta}'\Phi\left(\mathbf{w}\right)$ $-\bar{g}\left(\mathbf{w}\right)]| \leq \sup_{\mathbf{w}\in\mathcal{W}}\|\Phi\left(\mathbf{w}\right)\|\left\|\widetilde{\boldsymbol{\beta}}-\boldsymbol{\beta}\right\| + \sup_{w\in\mathcal{W}}\left|\boldsymbol{\beta}'\Phi\left(\mathbf{w}\right)-\bar{g}\left(\mathbf{w}\right)\right| = O_P\left[\varsigma_{0\kappa}\left(\nu_n+\nu_{1n}\right)\right]$ as the second term is $O\left(\nu_n\right)$. ■

Proof of Theorem 3.2.

Let $Y_{1i} \equiv Y_i - \mu - g_2\left(X_{2i}\right) - ... - g_{d_x}\left(X_{d_x i}\right) - g_{d_x+1}\left(Z_{11,i}\right) - ... - g_{d_x+d_1}\left(Z_{1d_1,i}\right) - g_{d_x+d_1+1}(U_{1i}) - ... - g_{2d_x+d_1}(U_{d_x i})$ and $\mathbf{Y}_1 \equiv (Y_{11},...,Y_{1n})'$. Using the notation defined at the end of section 2.2, we have

$$\begin{aligned}
H\widehat{\beta}_1\left(x_1\right) &= \left[H^{-1}\mathbb{X}_1\left(x_1\right)'\mathbb{K}_{x_1}\mathbb{X}_1\left(x_1\right)H^{-1}\right]^{-1}H^{-1}\mathbb{X}_1\left(x_1\right)'\mathbb{K}_{x_1}\mathbb{X}_1\left(x_1\right)\mathbf{Y}_1 \\
&\quad + \left[H^{-1}\mathbb{X}_1\left(x_1\right)'\mathbb{K}_{x_1}\mathbb{X}_1\left(x_1\right)H^{-1}\right]^{-1}H^{-1}\mathbb{X}_1\left(x_1\right)\mathbb{K}_{x_1}(\widetilde{\mathbf{Y}}_1-\mathbf{Y}_1) \\
&\equiv J_{1n}\left(x_1\right)+J_{2n}\left(x_1\right), \text{ say.}
\end{aligned}$$

By standard results in local-linear regressions [e.g., Masry (1996) and Hansen (2008)], $n^{-1}H^{-1}\mathbb{X}_1\left(x_1\right)'$ $\mathbb{K}_{x_1}\mathbb{X}_1\left(x_1\right)H^{-1} = f_{X_1}\left(x_1\right)\begin{pmatrix}1 & 0\\ 0 & \int u^2K\left(u\right)du\end{pmatrix} + o_P\left(1\right)$ uniformly in x_1, $n^{1/2}h^{1/2}\left[J_{1n}\left(x_1\right)-b_1\left(x_1\right)\right] \overset{D}{\to}$ $N\left(0,\Omega_1\left(x_1\right)\right)$ and $\sup_{x_1\in\mathcal{X}_1}\|J_{1n}\left(x_1\right)\| = O_P\left(\left(nh\log n\right)^{-1/2}+h^2\right)$, where $b_1\left(x_1\right)$ and $\Omega_1\left(x_1\right)$ are defined in Theorem 3.2. It suffices to prove the theorem by showing that $n^{-1/2}h^{1/2}H^{-1}\mathbb{X}_1\left(x_1\right)'\mathbb{K}_{x_1}(\widetilde{\mathbf{Y}}_1-\mathbf{Y}_1)$ $= o_P\left(1\right)$ uniformly in x_1 (for part (i) of Theorem 3.2 we only need the pointwise result to hold).

We make the following decomposition:

$$\begin{aligned}
(n/h)^{-1/2}H^{-1}\mathbb{X}_1\left(x_1\right)\mathbb{K}_{x_1}(\mathbf{Y}_1-\widetilde{\mathbf{Y}}_1) &= n^{-1/2}h^{1/2}\sum_{i=1}^{n}K_{ix_1}H^{-1}X_{1i}^{*}\left(x_1\right)\left(Y_{1i}-\widetilde{Y}_{1i}\right) \\
&= \sqrt{n}\left(\widetilde{\mu}-\mu\right)n^{-1}h^{1/2}\sum_{i=1}^{n}K_{ix_1}H^{-1}X_{1i}^{*}\left(x_1\right) \\
&\quad + \sum_{l=2}^{d_x}n^{-1/2}h^{1/2}\sum_{i=1}^{n}K_{ix_1}H^{-1}X_{1i}^{*}\left(x_1\right)\left[\widetilde{g}_l\left(X_{li}\right)-g_l\left(X_{li}\right)\right] \\
&\quad + \sum_{j=1}^{d_1}n^{-1/2}h^{1/2}\sum_{i=1}^{n}K_{ix_1}H^{-1}X_{1i}^{*}\left(x_1\right)\left[\widetilde{g}_{d_x+j}\left(Z_{1j,i}\right)-g_{d_x+j}\left(Z_{1j,i}\right)\right] \\
&\quad + \sum_{l=1}^{d_x}n^{-1/2}h^{1/2}\sum_{i=1}^{n}K_{ix_1}H^{-1}X_{1i}^{*}\left(x_1\right)\left[\widetilde{g}_{d_x+d_1+l}(\widetilde{U}_{li})-g_{d_x+d_1+l}(U_{li})\right] \\
&\equiv A_n\left(x_1\right)+\sum_{l=2}^{d_x}B_{nl}\left(x_1\right)+\sum_{j=1}^{d_1}C_{nj}\left(x_1\right)+\sum_{l=1}^{d_x}D_{nl}\left(x_1\right).
\end{aligned}$$

We prove the first part of the theorem by showing that $(i1)$ $A_n\left(x_1\right) = o_P\left(1\right)$, $(i2)$ $B_{nl}\left(x_1\right) = o_P\left(1\right)$ for $l = 2,...,d_x$, $(i3)$ $C_{nj}\left(x_1\right) = o_P\left(1\right)$ for $j = 1,...,d_1$ and $(i4)$ $D_{nl}\left(x_1\right) = o_P\left(1\right)$ for $l = 1,...,d_x$, all uniformly in x_1.

$(i1)$ holds by noticing that $\sqrt{n}\left(\widetilde{\mu}-\mu\right) = O_P\left(1\right)$ and $n^{-1}\sum_{i=1}^{n}K_{ix_1}H^{-1}X_{1i}^{*}\left(x_1\right) = O_P(1)$ uniformly in x_1. Let $c \equiv \left(c_1,c_2\right)'$ be an arbitrary 2×1 nonrandom vector such that $\|c\| = 1$. Recall that $\eta_{nl}\left(x_1\right) \equiv$

$n^{-1}\sum_{i=1}^n K_{ix}c'H^{-1}X_{1i}^*(x_1)\,p^{\kappa}(X_{li})$. For $(i2)$, we make the following decomposition

$$\begin{aligned}
c'B_{nl}(x_1) &= n^{-1/2}h^{1/2}\sum_{i=1}^n K_{ix_1}c'H^{-1}X_{1i}^*(x_1)\,p^{\kappa}(X_{li})'\,\mathbb{S}_l\left(\widetilde{\boldsymbol{\beta}}-\boldsymbol{\beta}\right)\\
&\quad +n^{-1/2}h^{1/2}\sum_{i=1}^n K_{ix_1}c'H^{-1}X_{1i}^*(x_1)\left[p^{\kappa}(X_{li})'\,\mathbb{S}_l\boldsymbol{\beta}-g_l(X_{li})\right]\\
&= n^{1/2}h^{1/2}\eta_{nl}(x_1)'\,\mathbb{S}_lQ_{\Phi\Phi}^{-1}\xi_n+n^{1/2}h^{1/2}\eta_{nl}(x_1)\,\mathbb{S}_lQ_{\Phi\Phi}^{-1}\zeta_n\\
&\quad -n^{-1/2}h^{1/2}\eta_{nl}(x_1)'\,\mathbb{S}_lQ_{\Phi\Phi}^{-1}\sum_{j=1}^n\Phi_j\sum_{k=1}^{d_x}\delta_{kj}\left(\widetilde{U}_{kj}-U_{kj}\right)+n^{-1/2}h^{1/2}\eta_{nl}(x_1)'\,\mathbb{S}_l\mathbf{R}_{n,\beta}\\
&\quad +n^{-1/2}h^{1/2}\sum_{i=1}^n K_{ix_1}c'H^{-1}X_{1i}^*(x_1)\left[p^{\kappa}(X_{li})'\,\mathbb{S}_l\boldsymbol{\beta}-g_l(X_{li})\right]\\
&\equiv B_{nl,1}(x_1)+B_{nl,2}(x_1)-B_{nl,3}(x_1)+B_{nl,4}(x_1)+B_{nl,5}(x_1),
\end{aligned}$$

where recall $\delta_{kj}\equiv\dot{p}^{\kappa}(U_{kj})'\,\boldsymbol{\beta}_{d_x+d_1+k}$, $\xi_n\equiv n^{-1}\sum_{j=1}^n\Phi_je_j$ and $\zeta_n\equiv n^{-1}\sum_{j=1}^n\Phi_j\left[\bar{g}(X_j,Z_{1j},U_j)-\boldsymbol{\beta}'\Phi_j\right]$. Let $\bar{\eta}_l(x_1)\equiv E[\eta_{nl}(x_1)]$ and $r_{\eta l}(x_1)=\eta_{nl}(x_1)-\bar{\eta}_l(x_1)$. By the proof of Lemma A.6(i), $\|r_{\eta l}(x_1)\|=O_P(\kappa^{1/2}(nh/\log n)^{-1/2})$, $\|\bar{\eta}_l(x_1)\|=O_P\left(1+\kappa^{1/2}h\right)$ and $\|\eta_{nl}(x_1)\|=O_P(1)$ uniformly in x_1. Note that

$$B_{nl,1}(x_1)=n^{1/2}h^{1/2}\bar{\eta}_l(x_1)'\,\mathbb{S}_lQ_{\Phi\Phi}^{-1}\xi_n+n^{1/2}h^{1/2}r_{\eta l}(x_1)'\,\mathbb{S}_lQ_{\Phi\Phi}^{-1}\xi_n\equiv B_{nl,11}(x_1)+B_{nl,12}(x_1),\ \text{say.}$$

Noting that

$$\begin{aligned}
E\left[B_{nl,11}^2(x_1)\right] &= h\bar{\eta}_l(x_1)'\,\mathbb{S}_lQ_{\Phi\Phi}^{-1}E\left(\Phi_j\Phi_j'e_j^2\right)P_{\Phi\Phi}^{-1}\mathbb{S}_l'\bar{\eta}_l(x_1)\\
&\leq h\lambda_{\max}\left(E\left(\Phi_j\Phi_j'e_j^2\right)\right)\left[\lambda_{\min}(P_{\Phi\Phi})\right]^{-2}\lambda_{\max}(\mathbb{S}_l\mathbb{S}_l')\,\|\bar{\eta}_l(x_1)\|^2\\
&= hO(1)\,O_P(1)\,O_P(1)=O_P(h),
\end{aligned}$$

we have $|B_{nl,11}(x_1)|=O_P\left(h^{1/2}\right)$ for each $x_1\in\mathcal{X}_1$. Let $\breve{\eta}_l(x_1)\equiv Q_{\Phi\Phi}^{-1}\mathbb{S}_l'\bar{\eta}_l(x_1)$. Then we can write $\bar{\eta}_l(x_1)'\,\mathbb{S}_lQ_{\Phi\Phi}^{-1}\xi_n$ as $n^{-1}\sum_{i=1}^n\breve{\eta}_l(x_1)'\,\Phi_ie_i$. Noting that $E[\breve{\eta}_l(x_1)'\,\Phi_ie_i]=0$ and $E\left[\breve{\eta}_l(x_1)'\,\Phi_ie_i\right]^2=\breve{\eta}_l(x_1)'\,E\left(\Phi_i\Phi_i'e_i^2\right)\breve{\eta}_l(x_1)\leq\lambda_{\max}(Q_{\Phi\Phi,e})\left\|Q_{\Phi\Phi}^{-1}\right\|_{\text{sp}}^2\sup_{x_1\in\mathcal{X}_1}\|\bar{\eta}_l(x_1)\|=O(1)$, we can readily divide $\mathcal{X}_l$ into intervals of appropriate length and apply Bernstein inequality to show that $\bar{\eta}_l(x_1)'\,\mathbb{S}_lQ_{\Phi\Phi}^{-1}\xi_n=O_P\left((n/\log n)^{-1/2}\right)$. Consequently,

$$\sup_{x_1\in\mathcal{X}_1}|B_{nl,11}(x_1)|=n^{1/2}h^{1/2}O_P\left((n/\log n)^{-1/2}\right)=O_P\left((h/\log n)^{-1/2}\right)=o_P(1).$$

For $B_{nl,12}(x_1)$, we have by Lemma A.5(i)

$$\begin{aligned}
\sup_{x_1\in\mathcal{X}_1}\|B_{nl,12}(x_1)\| &\leq n^{1/2}h^{1/2}\sup_{x_1\in\mathcal{X}_1}\|r_{\eta l}(x_1)\|\,\|\mathbb{S}_l\|_{\text{sp}}\left\|Q_{\Phi\Phi}^{-1}\right\|_{\text{sp}}\|\xi_n\|\\
&= n^{1/2}h^{1/2}O_P(\kappa^{1/2}(nh/\log n)^{-1/2})O(1)\,O_P(1)\,O_P(\kappa^{1/2}n^{-1/2})\\
&= O_P(\kappa(n/\log n)^{-1/2})=o_P(1).
\end{aligned}$$

It follows that $\sup_{x_1\in\mathcal{X}_1}|B_{nl,1}(x_1)|=o_P(1)$. By Lemma A.5$(ii)$ and Assumptions A2(ii) and (v) and A5,

$$\begin{aligned}
\sup_{x_1\in\mathcal{X}_1}|B_{nl,2}(x_1)| &= \sup_{x_1\in\mathcal{X}_1}n^{1/2}h^{1/2}\left|\eta_{nl}(x_1)'\,\mathbb{S}_lQ_{\Phi\Phi}^{-1}\zeta_n\right|\leq n^{1/2}h^{1/2}\sup_{x_1\in\mathcal{X}_1}\|\eta_{nl}(x_1)\|\left\|Q_{\Phi\Phi}^{-1}\right\|_{\text{sp}}\|\mathbb{S}_l\|_{\text{sp}}\|\zeta_n\|\\
&= n^{1/2}h^{1/2}O_P(1)\,O(1)\,O(1)O_P(\kappa^{-\gamma})=o_P(1),\\
\sup_{x_1\in\mathcal{X}_1}|B_{nl,4}(x_1)| &= \sup_{x_1\in\mathcal{X}_1}n^{1/2}h^{1/2}\left|\eta_{nl}(x_1)'\,\mathbb{S}_l\mathbf{R}_{n,\beta}\right|\leq n^{1/2}h^{1/2}\sup_{x_1\in\mathcal{X}_1}\|\eta_{nl}(x_1)\|\,\|\mathbb{S}_l\|_{\text{sp}}\|\mathbf{R}_{n,\beta}\|\\
&= n^{1/2}h^{1/2}O_P(1)\,O(1)\,o_P\left(n^{-1/2}h^{-1/2}\right)=o_P(1),
\end{aligned}$$

and

$$\sup_{x_1\in\mathcal{X}_1}\left|B_{nl,5}\left(x_1\right)\right| \leq O\left(\kappa^{-\gamma}\right)n^{1/2}h^{1/2}\left\{\sup_{x_1\in\mathcal{X}_1} n^{-1}\sum_{i=1}^{n}K_{ix_1}\left|c'H^{-1}X_{1i}^{*}\left(x_1\right)\right|\right\} = O_P\left(n^{1/2}h^{1/2}\kappa^{-\gamma}\right) = o_P\left(1\right).$$

For $B_{nl,3}\left(x_1\right)$, we have $B_{nl,3}\left(x_1\right)=\sum_{k=1}^{d_x}B_{nl,3k}\left(x_1\right)$ where $B_{nl,3k}\left(x_1\right)=n^{-1/2}h^{1/2}\eta_{nl}\left(x_1\right)'\mathbb{S}_lQ_{\Phi\Phi}^{-1}\sum_{j=1}^{n}$ $\Phi_j\delta_{kj}(\widetilde{U}_{kj}-U_{kj})$. Using (A.2), we make the following decomposition

$$B_{nl,3k}\left(x_1\right)=-\sum_{s=1}^{5}n^{-1/2}h^{1/2}\eta_{nl}\left(x_1\right)'\mathbb{S}_lQ_{\Phi\Phi}^{-1}\sum_{j=1}^{n}\Phi_j\delta_{kj}u_{sk,j}\equiv-\sum_{s=1}^{5}B_{nl,3ks}\left(x_1\right),\text{ say.}$$

First, noting that δ_{kj} is uniformly bounded, we can show $\left\|n^{-1}\sum_{j=1}^{n}\Phi_j\delta_{kj}\right\|_{\mathrm{sp}}=O_P\left(1\right)$ using arguments similar to those used in the proof of Lemma A.5(iii). It follows that

$$\begin{aligned}\sup_{x_1\in\mathcal{X}_1}\left|B_{nl,3k1}\left(x_1\right)\right| &\leq h^{1/2}\sup_{x_1\in\mathcal{X}_1}\left\|\eta_{nl}\left(x_1\right)\right\|\left\|\mathbb{S}_l\right\|_{\mathrm{sp}}\left\|Q_{\Phi\Phi}^{-1}\right\|_{\mathrm{sp}}\left\|n^{-1}\sum_{j=1}^{n}\Phi_j\delta_{kj}\right\|_{\mathrm{sp}}n^{1/2}\left|\mu_k-\widetilde{\mu}_k\right|\\ &= h^{1/2}O_P\left(1\right)O\left(1\right)O\left(1\right)O_P\left(1\right)O\left(1\right)O_P\left(1\right)=o_P\left(1\right).\end{aligned}$$

Now we decompose $B_{nl,3k2}\left(x_1\right)$ as follows:

$$\begin{aligned}B_{nl,3k2}\left(x_1\right) &= \sum_{m=1}^{d_1}n^{-1/2}h^{1/2}\bar{\eta}_l\left(x_1\right)'\mathbb{S}_lQ_{\Phi\Phi}^{-1}\sum_{j=1}^{n}\delta_{kj}\Phi_jp^{\kappa_1}\left(Z_{1m,j}\right)'\mathbb{S}_{1m}a_{1k}\\ &\quad+\sum_{m=1}^{d_1}n^{-1/2}h^{1/2}r_{\eta l}\left(x_1\right)'\mathbb{S}_lQ_{\Phi\Phi}^{-1}\sum_{j=1}^{n}\delta_{kj}\Phi_jp^{\kappa_1}\left(Z_{1m,j}\right)'\mathbb{S}_{1m}a_{1k}\\ &\equiv B_{nl,3k2}^{(1)}\left(x_1\right)+B_{nl,3k2}^{(2)}\left(x_1\right),\text{ say.}\end{aligned}$$

Let $\varphi_{nlkm}\left(x_1\right)=\bar{\eta}_l\left(x_1\right)'\mathbb{S}_lQ_{\Phi\Phi}^{-1}n^{-1}\sum_{j=1}^{n}\delta_{kj}\Phi_jp^{\kappa_1}\left(Z_{1m,j}\right)$ and $\bar{\varphi}_{lkm}\left(x_1\right)=E\left[\varphi_{nlkm}\left(x_1\right)\right]$. Arguments like those used to study $\eta_{nl}\left(x_1\right)$ in the proof of Lemma A.6(i) show that $\left\|\bar{\varphi}_{lkm}\left(x_1\right)\right\|=O\left(\left\|\bar{\eta}_l\left(x_1\right)\right\|\right)=O\left(1+\kappa^{1/2}h\right)=O\left(1\right)$ under Assumption A5(ii) and $\left\|\varphi_{nlkm}\left(x_1\right)-E\left[\varphi_{nlkm}\left(x_1\right)\right]\right\|=\left\|\bar{\eta}_l\left(x_1\right)\right\|O_P((\kappa^{1/2}\log n/n)^{-1/2})=O_P((\kappa^{1/2}\log n/n)^{-1/2})$ uniformly in x_1. We further decompose $B_{nl,3k2}^{(1)}\left(x_1\right)$ as follows

$$\begin{aligned}B_{nl,3k2}^{(1)}\left(x_1\right) &= \sum_{m=1}^{d_1}n^{-1/2}h^{1/2}\bar{\eta}_l\left(x_1\right)'\mathbb{S}_lQ_{\Phi\Phi}^{-1}\sum_{j=1}^{n}\delta_{kj}\Phi_jp^{\kappa_1}\left(Z_{1m,j}\right)'\mathbb{S}_{1m}Q_{n,PP}^{-1}\xi_{nk}\\ &= \sum_{m=1}^{d_1}n^{1/2}h^{1/2}\bar{\varphi}_{lkm}\left(x_1\right)'\mathbb{S}_{1m}Q_{PP}^{-1}\xi_{nk}+\sum_{m=1}^{d_1}n^{1/2}h^{1/2}\bar{\varphi}_{lkm}\left(x_1\right)'\mathbb{S}_{1m}(Q_{n,PP}^{-1}-Q_{PP}^{-1})\xi_{nk}\\ &\quad+\sum_{m=1}^{d_1}n^{1/2}h^{1/2}r_{nlkm}\left(x_1\right)'\mathbb{S}_{1m}Q_{n,PP}^{-1}\xi_{nk}\\ &\equiv B_{nl,3k2}^{(1,1)}\left(x_1\right)+B_{nl,3k2}^{(1,2)}\left(x_1\right)+B_{nl,3k2}^{(1,3)}\left(x_1\right).\end{aligned}$$

Following the analysis of $B_{nl,11}\left(x_1\right)$, we can show that $\sup_{x_1\in\mathcal{X}_1}\left|B_{nl,3k2}^{(1,1)}\left(x_1\right)\right|=O_P\left((h/\log n)^{1/2}\right)$. In addition,

$$\begin{aligned}\sup_{x_1\in\mathcal{X}_1}\left|B_{nl,3k2}^{(1,2)}\left(x_1\right)\right| &\leq n^{1/2}h^{1/2}\sup_{x_1\in\mathcal{X}_1}\sum_{m=1}^{d_1}\left\|\bar{\varphi}_{lkm}\left(x_1\right)\right\|\left\|\mathbb{S}_{1m}\right\|_{\mathrm{sp}}\left\|Q_{n,PP}^{-1}-Q_{PP}^{-1}\right\|_{\mathrm{sp}}\left\|\xi_{nk}\right\|\\ &= n^{1/2}h^{1/2}O_P(1)O\left(1\right)O_P\left(\kappa_1n^{-1/2}\right)O_P(\kappa_1^{1/2}n^{-1/2})=o_P\left(1\right),\end{aligned}$$

and

$$\begin{aligned}\sup_{x_1\in\mathcal{X}_1}\left|B^{(1,3)}_{nl,3k2}\left(x_1\right)\right| &\le n^{1/2}h^{1/2}\sup_{x_1\in\mathcal{X}_1}\sum_{m=1}^{d_1}\left\|r_{nlkm}\left(x_1\right)\right\|\left\|\mathbb{S}_{1m}\right\|_{\mathrm{sp}}\left\|Q^{-1}_{n,PP}\right\|_{\mathrm{sp}}\left\|\xi_{nk}\right\|_{\mathrm{sp}}\\ &= n^{1/2}h^{1/2}O_P((\kappa^{1/2}\log n/n)^{-1/2})O\left(1\right)O_P\left(1\right)O_P\left(\kappa_1^{1/2}n^{-1/2}\right)=o_P\left(1\right).\end{aligned}$$

It follows that $\sup_{x_1\in\mathcal{X}_1}\left|B^{(1)}_{nl,3k2}\left(x_1\right)\right|=o_P\left(1\right)$. For $B^{(2)}_{nl,3k2}\left(x_1\right)$, we have

$$\begin{aligned}\sup_{x_1\in\mathcal{X}_1}\left|B^{(2)}_{nl,3k2}\left(x_1\right)\right| &\le n^{1/2}h^{1/2}\sup_{x_1\in\mathcal{X}_1}\left\|r_{\eta l}\left(x_1\right)\right\|\left\|\mathbb{S}_{l}\right\|_{\mathrm{sp}}\left\|Q^{-1}_{\Phi\Phi}\right\|_{\mathrm{sp}}\sum_{m=1}^{d_1}\left\|t_{nkm}\right\|_{\mathrm{sp}}\left\|\mathbb{S}_{1m}\right\|_{\mathrm{sp}}\left\|a_{1k}\right\|\\ &= n^{1/2}h^{1/2}O_P\left((\kappa^{1/2}\log n/n)^{-1/2}\right)O\left(1\right)O_P\left(1\right)O\left(1\right)O_P\left(\kappa_1^{1/2}n^{-1/2}\right)=o_P\left(1\right),\end{aligned}$$

where $t_{nkm}\equiv n^{-1}\sum_{j=1}^{n}\delta_{kj}\Phi_j p^{\kappa_1}\left(Z_{1m,j}\right)'$, we use the fact that $\left\|t_{nkm}\right\|_{\mathrm{sp}}=O_P\left(1\right)$ by following similar arguments to those used in the proof of Lemma A.5(iii) and noticing that δ_{kj} is uniformly bounded. Consequently we have shown that $\sup_{x_1\in\mathcal{X}_1}\left|B_{nl,3k2}\left(x_1\right)\right|=o_P\left(1\right)$. Analogously,

$$\begin{aligned}\sup_{x_1\in\mathcal{X}_1}\left|B_{nl,3k4}\left(x_1\right)\right| &\le n^{1/2}h^{1/2}\sup_{x_1\in\mathcal{X}_1}\left\|\eta_{nl}\left(x_1\right)\right\|_{\mathrm{sp}}\left\|\mathbb{S}_{l}\right\|_{\mathrm{sp}}\left\|Q^{-1}_{\Phi\Phi}\right\|_{\mathrm{sp}}\sum_{m=1}^{d_1}\left\|t_{nkm}\right\|_{\mathrm{sp}}\left\|\mathbb{S}_{1m}\right\|_{\mathrm{sp}}\left\|a_{2k}\right\|\\ &= n^{1/2}h^{1/2}O_P(1)O\left(1\right)O_P\left(1\right)O_P\left(1\right)O\left(1\right)O_P\left(\kappa_1^{-\gamma}\right)=o_P\left(1\right).\end{aligned}$$

By the same token, we can show that $B_{nl,3k3}\left(x_1\right)=o_P\left(1\right)$ and $B_{nl,3k}\left(x_1\right)=o_P\left(1\right)$ uniformly in x_1. It follows that $\sup_{x_1\in\mathcal{X}_1}\left\|B_{nl,3k}\left(x_1\right)\right\|=o_P\left(1\right)$ for $k=1,...,d_x$. Analogously, we can show that $(i3)$: $\sup_{x_1\in\mathcal{X}_1}\left\|C_{nj}\left(x_1\right)\right\|=o_P\left(1\right)$ for $j=1,...,d_1$.

Now we show $(i4)$. We make the following decomposition

$$\begin{aligned}c'D_{nl}\left(x_1\right) &= n^{-1/2}h^{1/2}\sum_{i=1}^{n}K_{ix_1}c'H^{-1}X^*_{1i}\left(x_1\right)\left[\widetilde{g}_{d_x+d_1+l}(\widetilde{U}_{li})-g_{d_x+d_1+l}(\widetilde{U}_{li})\right]\\ &\quad +n^{-1/2}h^{1/2}\sum_{i=1}^{n}K_{ix_1}c'H^{-1}X^*_{1i}\left(x_1\right)\left[g_{d_x+d_1+l}(\widetilde{U}_{li})-g_{d_x+d_1+k}(U_{li})\right]\\ &\equiv D_{nl,1}\left(x_1\right)+D_{nl,2}\left(x_1\right),\text{ say.}\end{aligned}$$

In view of the fact that $\widetilde{g}_{d_x+d_1+l}(\widetilde{U}_{li})-g_{d_x+d_1+l}(\widetilde{U}_{li})=p^{\kappa}(\widetilde{U}_{li})'\mathbb{S}_{d_x+d_1+k}(\widetilde{\boldsymbol{\beta}}-\boldsymbol{\beta})+\left[p^{\kappa}(\widetilde{U}_{li})'\boldsymbol{\beta}_l-g_{d_x+d_1+l}(\widetilde{U}_{li})\right]$, we continue to decompose $D_{nl,1}\left(x_1\right)$ as follows:

$$\begin{aligned}D_{nl,1}\left(x_1\right) &= n^{-1/2}h^{1/2}\sum_{i=1}^{n}K_{ix_1}c'H^{-1}X^*_{1i}\left(x_1\right)p^{\kappa}\left(U_{li}\right)'\mathbb{S}_{d_x+d_1+l}\left(\widetilde{\boldsymbol{\beta}}-\boldsymbol{\beta}\right)\\ &\quad +n^{-1/2}h^{1/2}\sum_{i=1}^{n}K_{ix_1}c'H^{-1}X^*_{1i}\left(x_1\right)\left[p^{\kappa}(\widetilde{U}_{li})-p^{\kappa}\left(U_{li}\right)\right]'\mathbb{S}_{d_x+d_1+l}\left(\widetilde{\boldsymbol{\beta}}-\boldsymbol{\beta}\right)\\ &\quad -n^{-1/2}h^{1/2}\sum_{i=1}^{n}K_{ix_1}c'H^{-1}X^*_{1i}\left(x_1\right)\left[g_{d_x+d_1+l}(\widetilde{U}_{li})-p^{\kappa}(\widetilde{U}_{li})'\boldsymbol{\beta}_l\right]\\ &\equiv D_{nl,11}\left(x_1\right)+D_{nl,12}\left(x_1\right)+D_{nl,13}\left(x_1\right),\text{ say.}\end{aligned}$$

Analogous to the analysis of $B_{nl,1}(x_1)$, we can readily show that $\sup_{x_1\in\mathcal{X}_1}|D_{nl,11}(x_1)| = o_P(1)$. For $D_{nl,12}(x_1)$, by Taylor expansion,

$$\begin{aligned} D_{nl,12}(x_1) &= n^{-1/2}h^{1/2}\sum_{i=1}^{n} K_{ix_1}c'H^{-1}X_{1i}^{*}(x_1)(\widetilde{U}_{li}-U_{li})\dot{p}^{\kappa}(U_{li})'\left(\widetilde{\boldsymbol{\beta}}_l-\boldsymbol{\beta}_l\right) \\ &\quad +\frac{1}{2}n^{-1/2}h^{1/2}\sum_{i=1}^{n} K_{ix_1}c'H^{-1}X_{1i}^{*}(x_1)(\widetilde{U}_{li}-U_{li})^2\ddot{p}^{\kappa}\left(U_{li}^{\ddagger}\right)'\left(\widetilde{\boldsymbol{\beta}}_l-\boldsymbol{\beta}_l\right) \\ &\equiv D_{nl,121}(x_1)+\frac{1}{2}D_{nl,122}(x_1), \text{ say,} \end{aligned}$$

where $U_{li}^{\ddagger}$ lies between $\widetilde{U}_{li}$ and U_{li}. By Theorem 3.1 and Lemmas A.6(i)-(ii), $\sup_{x_1\in\mathcal{X}_1}|D_{nl,121}(x_1)| = h^{1/2}\varsigma_{1\kappa}\ O_P(1+n^{1/2}\kappa_1^{-\gamma})O_P(\nu_n+\nu_{1n}) = o_P(1)$ and

$$\begin{aligned} \sup_{x_1\in\mathcal{X}_1}|D_{nl,122}(x_1)| &\leq \varsigma_{2\kappa}\sup_{x_1\in\mathcal{X}_1}\left\{n^{-1/2}h^{1/2}\sum_{i=1}^{n}K_{ix_1}c'H^{-1}X_{1i}^{*}(x_1)(\widetilde{U}_{li}-U_{li})^2\right\}\left\|\widetilde{\boldsymbol{\beta}}_l-\boldsymbol{\beta}_l\right\| \\ &= \varsigma_{2\kappa}n^{1/2}h^{1/2}O_P(\kappa_1 n^{-1}+\kappa_1^{-2\gamma})O_P(\nu_n+\nu_{1n}) = o_P(1). \end{aligned}$$

In addition, $\sup_{x_1\in\mathcal{X}_1}\|D_{nl,13}(x_1)\| \leq n^{1/2}h^{1/2}O(\kappa^{-\gamma})\sup_{x_1\in\mathcal{X}_1}n^{-1}\sum_{i=1}^{n}K_{ix_1}\left\|H^{-1}X_{1i}^{*}(x_1)\right\| = O_P(n^{1/2}$ $h^{1/2}\kappa^{-\gamma}) = o_P(1)$. It follows that $\sup_{x_1\in\mathcal{X}_1}|D_{nl,1}(x_1)| = o_P(1)$.

By Taylor expansion,

$$\begin{aligned} D_{nl,2}(x_1) &= n^{-1/2}h^{1/2}\sum_{i=1}^{n}K_{ix_1}c'H^{-1}X_{1i}^{*}(x_1)\dot{g}(U_{li})\left(\widetilde{U}_{li}-U_{li}\right) \\ &\quad +n^{-1/2}h^{1/2}\sum_{i=1}^{n}K_{ix_1}c'H^{-1}X_{1i}^{*}(x_1)\ddot{g}_{d_x+d_1+l}\left(U_{li}^{\ddagger}\right)\left(\widetilde{U}_{li}-U_{li}\right)^2 \\ &\equiv D_{nl,21}(x_1)+D_{nl,22}(x_1). \end{aligned}$$

Arguments like those used to study $B_{nl,3}(x_1)$ show that $\sup_{x_1\in\mathcal{X}_1}|D_{nl,21}(x_1)| = o_P(1)$. By Lemma A.6($ii$), $\sup_{x_1\in\mathcal{X}_1}|D_{nl,22}(x_2)| \leq c_{\ddot{g}}\sup_{x_1\in\mathcal{X}_1}\{n^{-1/2}h^{1/2}\sum_{i=1}^{n}K_{ix_1}\left|c'H^{-1}X_{1i}^{*}(x_1)\right|(\widetilde{U}_{li}-U_{ki})^2\} = n^{1/2}h^{1/2}$ $O_P(\nu_{1n}^2) = o_P(1)$, where $c_{\ddot{g}} = \sup_{u_l\in\mathcal{U}_l}\ddot{g}_{d_x+d_1+l}(u_l) = O(1)$. ■

References

Bernstein, D. S., 2005. *Matrix Mathematics: Theory, Facts and Formulas with Application to Linear Systems Theory.* Princeton University Press, Princeton.

Bernal, R. 2008. The effect of maternal employment and child care on children's cognitive development. *International Economic Review* 49, 4, 1173-209.

Bernal, R., Keane, M.P., 2011. Child care choices and children's cognitive achievement: the case of single mothers. *Journal of Labor Economics* 29, 459-12.

Blau, F. D., Grossberg, A. J. 1992. Maternal labor supply and children's cognitive development. *Review of Economics and Statistics* 74, 3, 474-1.

Cameron, S. V., Heckman, J. J. 1998. Life Cycle Schooling and Dynamic Selection Bias: Models and Evidence for Five Cohorts. *NBER Working Papers* 6385, National Bureau of Economic Research, Inc.

Chen, X., 2007. Large sample sieve estimation of semi-nonparametric models. In J. J. Heckman and E. Leamer (eds.), *Handbook of Econometrics*, 6B (Chapter 76), 5549-5632. New York: Elsevier Science.

Fan, J., Gijbels, I., 1996. *Local Polynomial Modelling and Its Applications.* Chapman & Hall, London.

Hahn, J., Ridder, G., 2013. Asymptotic variance of semiparametric estimators with generated regressors. *Econometrica* 81, 315-340.

Hansen, B. E., 2008. Uniform convergence rates for kernel estimation with dependent data. *Econometric Theory* 24, 726-748.

Horowitz, J. L., 2013. Nonparametric additive models. In A. Ullah, J. Racine and L. Su (eds.), *Handbook of Applied Nonparametric and Semiparametric Econometrics and Statistics*, forthcoming, Oxford University Press, Oxford.

Horowitz, J. L., Mammen, E., 2004. Nonparametric estimation of an additive model with a link function. *Annals of Statistics* 32, 2412-2443.

James-Burdumy, S. 2005. The effect of maternal labor force participation on child development. *Journal of Labor Economics* 23, 1, 177-11.

Keane, M. P., Wolpin, K. I. 2001. The effect of parental transfers and borrowing constraints on educational attainment. *International Economic Review* 42, 4, 1051-103.

Keane, M. P., Wolpin, K. I. 2001. Estimating welfare effects consistent with forward-looking behavior. *Journal of Human Resources* 37, 3, 600-22.

Kim, W., Linton O. B., Hengartner N. W., 1999. A computationally efficient estimator for additive nonparametric regression with bootstrap confidence intervals. *Journal of Computational and Graphical Statistics* 8, 278-297.

Li, Q., 2000. Efficient estimation of additive partially linear models. *International Economic Review* 41, 1073-1092.

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

Mammen, E., Rothe, C., Schienle, M., 2012. Nonparametric regression with nonparametrically generated covariates. *Annals of Statistics* 40, 1132-1170.

Martins-Filho, C., Yang, K., 2007. Finite sample performance of kernel-based regression methods for nonparametric additive models under common bandwidth selection criterion. *Journal of Nonparametric Statistics* 19, 23-62.

Masry, E., 1996. Multivariate local-polynomial regression for time series: uniform strong consistency rates. *Journal of Time Series Analysis,* 17, 571-599.

Newey, W. K., 1997. Convergence rates and asymptotic normality for series estimators. Journal of Econometrics 79, 147-168.

Newey, W. K., Powell, J. L., 2003. Instrumental variable estimation of nonparametric models. *Econometrica* 71, 1565-1578.

Newey, W. K., Powell, J. L., Vella, F., 1999. Nonparametric estimation of triangular simultaneous equation models. *Econometrica* 67, 565-603.

Ozabaci, D., Henderson, D. J. 2012. Gradients via oracle estimation for additive nonparametric regression with application to returns to schooling. Working paper, State University of New York at Binghamton.

Pagan, A., Ullah, A. 1999. *Nonparametric Econometrics.* Cambridge University Press, Cambridge.

Pinkse, J., 2000. Nonparametric two-step regression estimation when regressors and error are dependent. *Canadian Journal of Statistics* 28, 289-300.

Powell, M. J. D., 1981. *Approximation Theory and Methods.* Cambridge University Press, Cambridge.

Roehrig, C.S., 1988. Conditions for identification in nonparametric and parametric models. *Econometrica* 56, 433-47.

Schumaker, L. L., 2007. *Spline Functions: Basic Theory*, 3rd ed. Cambridge University Press, Cambridge.

Serfling, R. J., 1980. *Approximation Theorems of Mathematical Statistics.* John Wiley & Sons, New York.

Su, L., Jin, S., 2012. Sieve estimation of panel data models with cross section dependence. *Journal of Econometrics* 169, 34-47.

Su, L., Ullah, A. 2008. Local polynomial estimation of nonparametric simultaneous equations models. *Journal of Econometrics* 144, 193-218.

Thompson, R., Freede, 1974. Eigenvalues of partitioned Hermitian matrices. *Bulletin Australian Mathematical Society* 3, 23-37

Ullah, A., 1985. Specification analysis of econometric models. *Journal of Quantitative Economics* 1, 187-209

Vella, F. 1991. A Simple Two-Step Estimator for Approximating Unknown Functional Forms in Models with Endogenous Explanatory Variables. *Australian National University, Department of Economics* Working Paper.

Table 3: Final-stage gradient estimates for child care use for different types of child care, gender, amount and for different attributes of the mother at various percentiles (10%, 25%, 50%, 75% and 90%) for the specific group with corresponding wild bootstrapped standard errors

	0.10	0.25	0.50	0.75	0.90
Formal	-0.0087	-0.0067	-0.0016	0.0035	0.0065
	0.0052	0.0037	0.0036	0.0032	0.0031
Informal	-0.0090	-0.0061	-0.0007	0.0028	0.0067
	0.0089	0.0038	0.0036	0.0049	0.0042
Female	-0.0092	-0.0061	-0.0005	0.0057	0.0079
	0.0038	0.0038	0.0042	0.0046	0.0055
Male	-0.0089	-0.0067	-0.0005	0.0036	0.0079
	0.0034	0.0037	0.0042	0.0046	0.0055
Above median child care	-0.0092	-0.0087	-0.0054	0.0028	0.0049
	0.0038	0.0052	0.0036	0.0037	0.0037
Below median child care	-0.0061	-0.0010	0.0013	0.0076	0.0218
	0.0038	0.0035	0.0042	0.0034	0.0080
Education < 12	-0.0067	0.0014	0.0028	0.0076	0.0218
	0.0037	0.0036	0.0037	0.0034	0.0080
Education ≥ 12	-0.0090	-0.0068	-0.0014	0.0028	0.0067
	0.0089	0.0032	0.0036	0.0049	0.0042
Experience < 5	-0.0090	-0.0068	-0.0014	0.0028	0.0067
	0.0036	0.0032	0.0036	0.0049	0.0042
Experience ≥ 5	-0.0087	-0.0046	0.0011	0.0065	0.0218
	0.0052	0.0038	0.0040	0.0031	0.0080
No experience	-0.0061	-0.0007	0.0028	0.0076	0.0218
	0.0038	0.0036	0.0037	0.0034	0.0080
Age < 23	-0.0087	-0.0053	0.0007	0.0063	0.0139
	0.0052	0.0036	0.0039	0.0036	0.0066
Age $> 23, < 29$	-0.0090	-0.0061	-0.0007	0.0049	0.0079
	0.0089	0.0038	0.0053	0.0037	0.0055
Age ≥ 30	-0.0090	-0.0076	-0.0016	0.0028	0.0076
	0.0036	0.0034	0.0036	0.0037	0.0034

Table 4: Median characteristics for groups with positive versus negative negative child care use gradients

Attribute	Positve	Negative
Mother's educaton	12	12
Mother's experience	4	6
Mother's AFQT	15	19
Mother's age	22	24
Child's age	5.8	5.8
Formal child care	0	0
Informal child care	6	9
Total child care	8.5	11.5
Number of children	1	1
Sample size	1141	1313