The Pennsylvania State University

The Graduate School

Department of Economics

NONPARAMETRIC IDENTIFICATION AND ESTIMATION OF PRODUCTION FUNCTIONS USING CONTROL FUNCTION APPROACHES TO ENDOGENEITY

A Dissertation in

Economics

by

Jian Hong

Submitted in Partial Fulfillment

of the Requirements

for the Degree of

Doctor of Philosophy

August 2008

The thesis of Jian Hong was reviewed and approved* by the following:

Mark Roberts
Professor of Economics
Dissertation Advisor, Chair of Committee

Quang Vuong
Professor of Economics

Joris Pinkse
Associate Professor of Economics

Zhibiao Zhao
Assistant Professor of Statistics

Vijay Krishna
Professor of Economics
Director of the Graduate Program of the Department of Economics

*Signatures are on file in the Graduate School.

Abstract

Endogeneity and misspecification of models are two main concerns in structural estimation, which usually involves the optimal choices of economic agents with unobservable characteristics. In estimating production functions, input variables are endogenous because input decisions depend on unobservable productivity shocks. Economic theory rarely suggests functional forms for either production functions or the distribution of productivity.

Using control function approaches to endogeneity, nonparametric identification is established for production functions under weak conditions. The distribution of productivity is also recovered nonparametrically. Instead of "inverting out" productivity shocks, the control functions "smooth out" the unobserved shocks. Controls are constructed using lagged levels of inputs as instruments, and the control function condition is justified by a Markov property of productivity shocks along with interim uncertainty of productivity faced by firms.

Nonparametric estimation of production functions then closely follows the identification strategy without imposing extra modeling assumptions. A kernel estimator is proposed for nonparametric regressions with endogeneity. If the preliminary estimators of controls converge sufficiently fast, the estimator achieves the optimal rate of uniform convergence and the asymptotic variance is unaffected by preliminary estimators.

The same strategy also applies to parametric identification. When the Cobb-Douglas production function is considered, a partial linear model arises, where the parametric part represents the production function and the nonparametric part is the control function to account for the endogeneity of input variables. An density-weighted estimator is proposed for the partially linear model with constructed controls, and $\sqrt{n}$-consistency is established under the given conditions.

The finite sample performances of the proposed estimators are illustrated by extensive Monte-Carlo experiments. The application to the Chilean panel shows the empirical relevance of the identification strategy and estimation procedure proposed in this thesis. The resulting estimates are reasonable and show that some parametric specifications may be restrictive.

Table of Contents

Acknowledgments

I am grateful to my thesis advisor, Quang Vuong and Mark Roberts, for their invaluable advice and guidance. I am indebted to Joris Pinkse for many constructive questions and suggestions. I thank my outside committee members from statistics, Zhibiao Zhao. I also think Mark Roberts for providing the data set used in this thesis.

I would like to thank Neil Wallace, John Riew, Kala Krishna, Vijay Krishna and Isabelle Perrigne for their strong support during my Ph. D. program at Penn State.

Dedication

I would like to dedicate this thesis to my wife Xiaohua Chen and my son Eric Hong for their love and support.

I would also like to dedicate this thesis to my father Guangchai Hong and my mother Qinying Huang.

Chapter 1

Introduction

1.1 Endogeneity and Misspecification

The estimation of production functions has challenged empirical researchers and econometricians for decades, despite the fact that firm production is a pillar of economics. The theoretical models of firm production have been well studied where the optimal input choices must take into account all the information available to firms when decisions are made. Input decisions generally depend on productivity shocks and prices faced by firms. Therefore, if observed, they should be incorporated into the estimation to control for their effects on production. However, firm-level prices are seldom reported and productivity shocks are difficult to measure. So we have to include productivity shocks into errors and input variables become correlated to the error term. This missing-data/omitted-variable problem can also be seen as a simultaneity problem, where not only the output but also inputs are determined simultaneously when firms solve their optimization problems. This issue applies to general structural models involving the optimal choices of economic agents. Econometrically, input variables are endogenous due to their potential correlation with the error term and the traditional OLS estimates are inconsistent. Thus, how to account for firm heterogeneity and control for idiosyncratic productivity shocks is a central issue in the identification of production functions.

Misspecification is another concern in estimating production functions. Economic theories of firm production rarely suggest functional forms for production functions or the distribution of productivity. Imposing ad hoc model specification may lead to false identification, and misspecification usually results in inconsistent estimates and misleading policy implication. Nonparametric methods do not assume functional forms for either structural relationships or the distribution of the data generating processes.[1] Therefore, nonparametric estimates are more robust to misspecification than their parametric counterparts. The downside is that nonparametric estimators converge slower than the parametric rate and are more demanding of data. With wider access to large datasets and computing resources, nonparametric and semiparametric modeling and estimation attract much more attention than before, especially in fields such as empirical auction and structural labor econometrics.[2]

Many alternatives have been proposed and two lines of literature are related to this paper. The first one is the instrumental variables (IV) methods in dynamic panel models, where exogenous variations of instruments are exploited to form moment conditions. See Arellano and Bond (1991), Arellano and Bover (1995), and Blundell and Bond (2000) among others. The second one begins with Olley and Pakes (1996) and Levinsohn and Petrin (2003), where endogenous variations

[1] See Pagan and Ullah (1999), and Li and Racine (2006) for comprehensive coverage of nonparametric and semiparametric methods.

[2] See Athey and Haile (2007) for an extensive survey of nonparametric approaches to auctions; see Heckman and Vytacil (2007) for some discussion of nonparametric and semiparametric approaches to econometric evaluation of treatment effects.

of proxies are used to control for unobserved productivity shocks. Heckman and Vytlacil (2006) call this approach the method of replacement functions since the productivity shock is replaced by the inversion of observable input decisions. In some sense, the method of replacement functions can be seen as a special case of control function approaches, which are general ways to handle endogeneity problems. I will review the literature of control function (CF) approaches in Section 1.3, before which I go over some IV methods used to estimate production functions.

1.2 Instrumental Variables Methods

As an early solution, the instrumental variables (IV) method with fixed-effects tries to address the firm heterogeneity by a firm-specific, time-invariant scalar. This is a strong assumption preventing dynamic effects on production, and the resulting estimates are discouraging (Griliches and Mairesse, 1998). The dynamic panel (DP) models then extend the fixed-effect models by introducing richer structures of the unobserved productivity shock into the error term. A typical DP model is as follows:

$$y_{it} = \beta_k k_{it} + \beta_l l_{it} + \varepsilon_{it}, \text{ where } \varepsilon_{it} = \alpha_i + \delta_t + a_{it} + \epsilon_{it}. \tag{1.1}$$

The output y_{it} is mainly determined by inputs (k_{it}, l_{it}), productivity a_{it}, and technology (summarized by β_k and β_l). The error term ε_{it} is decomposed into four components: the time-invariant fixed-effect α_i, the common time effect δ_t (e.g. general technological advance or macro shocks),[3] the *i.i.d.* noise ϵ_{it} (e.g. measurement errors), and the serially-correlated, idiosyncratic shock a_{it}. Assumptions on

[3]Notice that if (β_k, β_l) are time-invariant, δ_t just picks up the location changes due to time-varying shocks common to all firms. With δ_t, a_{it} is usually normalized by $E(a_{it}) = 0$.

the evolution of $(\alpha_i, , \delta_t, a_{it}, \epsilon_{it})$ and their relationships with (k_{it}, l_{it}) are imposed to form the moment conditions to estimate (β_k, β_l). Since α_i is time-invariant, some differencing is necessary to take α_i out to form the moment conditions. Much of useful variation is lost during differencing, which leads to weak instruments (Blundell and Bond, 2000). Additionally, DP models rely heavily on linear structures of ε_{it}.

The Cobb-Douglas production function in (1.1) summarizes the technology of an industry by two coefficients only, which is restrictive in many empirical applications. Relax the functional restriction on firm production to

$$y_{it} = g\left(x_{it}\right) + u_{it},$$

where inputs x_{it} are correlated to the error u_{it}. Hence, $g\left(x\right)$ cannot be recovered as the conditional mean of y_{it} given $\left(x_{it} = x\right)$

$$E\left(y_{it}|x_{it} = x\right) = g\left(x\right) + E\left(u_{it}|x_{it} = x\right) \neq g\left(x\right)$$

for $E\left(u_{it}|x_{it}\right) \neq 0$. The IV approach can be extended to nonparametric case given there exist instruments z such that $E\left(u_{it}|z_{it} = z\right) = 0$. So $g(x)$ can be recovered by solving the functional equation:

$$E\left(y_{it}|z_{it} = z\right) = E[g(x_{it})|z_{it} = z] = \int g(x)dF_{x|z},$$

where $F_{x|z}$ is the conditional cumulative distribution of x_{it} given $z_{it} = z$. The estimator of $g(x)$ can then be derived by plugging in their sample analogs, i.e, $\widehat{g}(x)$ solves $\widehat{E}\left(y_{it}|z_{it} = z\right) = \int g(x)d\widehat{F}_{x|z}$. Although it seems straightforward, except in the case with finite support,[4] this method suffers from the Ill-Posed Inverse problem. The problem implies that the consistency of $\widehat{E}\left(y_{it}|z_{it}=z\right)$ and $\widehat{F}_{x|z}$ does not imply the consistency of $\widehat{g}(x)$.[5] In order to get a consistent estimator of $g(x)$, some "regularization" has to be applied, see Darolles, Florens, and Renault (2002), Newey and Powell (2003), and Hall and Horowitz (2005). Although the ill-posed inverse problem is avoided and consistency is established in these papers, unclear is the implication of those technical restrictions on applications.[6] We then turn to control function approaches to endogeneity, which have been extended to nonparametric cases.

1.3 Control Function Approaches

As a generalization of control variables and proxy variables, control function approaches (CFAs) to endogeneity have been extensively used in studies of treatment effects, where the selection bias is a fundamental issue with non-experimental samples. CFAs also apply to various selectivity models, censored or truncated models, and Roy models. See Heckman (1976, 1978, 1979), Heckman and Robb (1985), and Heckman and Hotz 1989) among many others. Let's consider a simple bivariate model to illustrate how the endogeneity caused by sample selection can be *controlled* for by a function representing the selection process. We will see how the distributional and functional assumptions can be relaxed, during which we go from parametric to semiparametric, and then to nonparametric control function

[4]For instance, see Florens and Malavolti (2002), where the explanatory variable is binary and there is no ill-posed inverse problem.

[5]See Florens (2003) for details about inverse problems in instrumental variables estimation in nonparametric regressions.

[6]For a recent application, see Blundell, Chen and Kristensen (2003), where they develop sieves estimators in nonparametric IV framework to estimate Engel curves.

approaches.

Parametric Cases

For Type-2 Tobit models as in Heckman (1979), the latent variable equations are

$$y_{1i}^* = x_{1i}'\beta_1 + u_{1i}, \text{ and } y_{2i}^* = x_{2i}'\beta_2 + u_{2i}. \tag{1.2}$$

Note that linear functions are specified for both y_{1i}^* and y_{2i}^*. The outcome of interest y_{2i}^* is observed if $y_{1i}^* > 0$. An example from labor economics is that y_{1i}^* determines to work or not, and y_{2i}^* represents hours on job. So the observed variable equations can be written as

$$\begin{aligned} y_{1i} &= 1\left(y_{1i}^* > 0\right), \text{ (sample selection) and} \\ y_{2i} &= y_{2i}^* \cdot 1\left(y_{1i} = 1\right), \text{ (outcome equation).} \end{aligned}$$

Since the selection depends on observable x_1, this is a model with *selection on observables*. However, *selection on observables* is also possible in many empirical applications, in which cases instruments are often required for identification and estimation.[7]

In the spirit of Tobin (1958), β can be estimated by maximal likelihood methods; see Amemiya (1985). Although MLE does not fall into the category of control function approaches, I begin with MLE to show the difference among alternative methods in the distributional and functional restrictions required for estimation. Besides the parametric (linear) specification for (y_{1i}^*, y_{2i}^*), to derive the likelihood

[7] In Chapter 2, we will see a similar situation in estimating production functions, where we construct variables from observable instruments to control for the endogeneity of input variables.

function, a joint normal distribution is imposed on the errors:

$$\begin{pmatrix} u_1 \\ u_2 \end{pmatrix} \sim N\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & \sigma_{12} \\ \sigma_{12} & \sigma_2^2 \end{pmatrix}\right).$$

Although being efficient, the MLE is subject to misspecification. MLE is inconsistent if the errors are either non-normal or heteroskedastic.

With this in mind, Heckman (1979) proposes a two-step procedure,[8] where the joint distributional assumption on (u_1, u_2) is relaxed to a standard normal distribution on u_1 and a relationship between u_1 and u_2: $u_2 = \sigma_{12}u_1 + \epsilon$.[9] The conditional mean of y_2 given $x = (x_1, x_2)$ and $y_1^* > 0$ is

$$\begin{aligned} E\left(y_2|x, y_1^* > 0\right) &= x_2'\beta_2 + E\left(u_2|u_1 > -x_1'\beta_1\right) \\ &= x_2'\beta_2 + \sigma_{12}E\left(u_1|u_1 > -x_1'\beta_1\right) \\ &= x_2'\beta_2 + \sigma_{12}\lambda\left(x_1'\beta_1\right), \end{aligned}$$

where $\lambda(t) \equiv \phi(t) \diagup \Phi(t)$ with $\phi(t)$ and $\Phi(t)$ respectively being the PDF and CDF of standard normal distribution. Heckman's two-step procedure estimates β_2 by applying OLS to the following augmented model:

$$y_{2i} = x_{2i}'\beta_2 + \sigma_{12}\lambda\left(x_1'\widehat{\beta}_1\right) + \varepsilon_i,$$

where ε_i is an error term, and $\widehat{\beta}_1$ is estimated by probit regression of y_1 on x_1 at the first step. Notice that $\sigma_{12}\lambda(x_1'\beta_1)$ is the term to correct the selection bias and can be viewed as a control function. Although weaker than the MLE, the Heckit

[8] This two-step procedure is also called the Heckit estimator.

[9] It is assumed that ϵ is independent of u_1. Distributions other than normal may be specified for u_1.

estimator still relies heavily on distributional assumptions (on u_1).

Note that $y_{1i} = 1\left(y_{1i}^* > 0\right)$ is a discrete variable and u_1 is normally distributed. This leads to $\lambda\left(x_1'\widehat{\beta}_1\right)$ by probit at the first step, followed by OLS at the second step. In Rivers and Vuong (1988), the situation is reversed: the control is constructed by OLS residual at the first step and parameters of interest are estimated by probit at the second step.[10] Their model can be rewritten as follows:

$$\begin{aligned} y_{1i} &= x_i'\theta + v_i, \text{ (reduced form regression) and} \\ y_{2i}^* &= \alpha y_{1i} + x_{2i}'\beta_2 + u_{2i}, \text{ (outcome equation)} \end{aligned}$$

where y_{1i} is always observed while $y_{2i} = y_{2i}^*$ is observed if $y_{2i}^* > 0$ and $y_{2i} = 0$ otherwise. y_{1i} is endogenous in the outcome equation of interest given $u_{2i} = \sigma_{12}v_i + \epsilon_i$. The augmented model becomes

$$y_{2i}^* = \alpha y_{1i} + x_{2i}'\beta_2 + \sigma_{12}v_i + \epsilon_i. \tag{1.3}$$

The parameters in (1.3) can be estimated by probit as ϵ_i is independent and normally distributed, before which v_i has to be estimated as the residuals from OLS regression of y_1 on x, i.e,

$$\widehat{v_i} = y_{1i} - x_i'\widehat{\theta}. \tag{1.4}$$

In this example, the control function $\sigma_{12}v_i$ is linear and v_i is a constructed variable. In Heckman (1979) and Rivers and Vuong (1988), the model is complicated by selection, censoring or truncation, but control function approaches still work in these models.

[10] See also Smith and Blundell (1986) for the case with Tobit at the second step.

Semiparametric Cases

When no distributional assumption is imposed on the errors, a semiparametric Type-2 Tobit model arises as a partially linear model:

$$y_{2i} = x'_{2i}\beta_2 + c\left(x'_{1i}\beta_1\right) + \epsilon_{2i}, \text{ where}$$

$$c\left(x'_{1i}\beta_1\right) \equiv E\left(u_{2i}|x_i, y_{1i} = 1\right) = E\left(u_{2i}|x_i, u_{1i} > -x'_{1i}\beta_1\right). \tag{1.5}$$

No functional form is specified for the errors and $c\left(x'_1\beta_1\right)$ is the nonparametric counterpart of $\sigma_{12}\lambda\left(x'_1\beta_1\right)$. Therefore, the estimates are more robust than those obtained from the parametric methods described above. In order to estimate the parameter of interest β_2, β_1 has to be estimated and plugged into $c\left(x'_1\beta_1\right)$. For the estimators of β in similar settings and their asymptotic properties, see Powell (1987), Ichimura and Lee (1991), Ai (1997) and Li and Wooldridge (2002). This partially linear model is slightly different from the one studied by Robinson (1988) in that the conditioning variable $x'_1\beta_1$ is a constructed one. Here $x'_1\beta_1$ comes from the parametric specification in (1.2), which may be a poor approximation.

For the Cobb-Douglas production function $y_{it} = x'_{it}\beta + a_{it} + \epsilon_{it}$, if there exists a control v_{it} such that $E\left(a_{it}|x_{it}, v_{it}\right) = E\left(a_{it}|v_{it}\right) \equiv c\left(v_{it}\right)$, then we have a partially linear model

$$y_{it} = x'_{it}\beta + c\left(v_{it}\right) + \varepsilon_{it}, \text{ where } \varepsilon_{it} \equiv a_{it} + \epsilon_{it} - c\left(v_{it}\right). \tag{1.6}$$

The parametric part is the Cobb-Douglas production function while the nonparametric part is the control function to control for endogeneity. In Chapter 4, I consider a partially linear model where the nonparametric part is a control function with the control being *nonparametrically* constructed from lagged levels of inputs as instruments.

Nonparametric Cases

When both the outcome equation and control function are relaxed to be nonparametric, the model is more robust to the misspecification of underlying data generating processes. Consider a nonparametric model, $Y_i = g(X_i) + U_i$ where $E(U_i|X_i) \neq 0$. Now suppose that there is a *control* V such that

$$E(U|X,V) = E(U|V) \equiv c(V), \tag{1.7}$$

which is called *the control function assumption*. This is essentially an exclusion restriction, implying that X becomes conditionally mean-independent of U given V.[11] Under this assumption, a generalized additive model arises

$$E(Y|(X,V) = (x,v)) = g(x) + c(v), \tag{1.8}$$

and the augmented regression goes as follows

$$Y_i = g(X_i) + c(V_i) + \varepsilon_i, \text{ where } \varepsilon_i = U_i - c(V_i). \tag{1.9}$$

Intuitively, the new error term ε_i is formed by taking away the part correlated to X (i.e, the endogenous part $E(U|X,V) = c(V)$) from the old error term U_i. Therefore, after $c(V_i)$ is introduced to control for the endogeneity of X_i, ε_i is orthogonal to (X_i, V_i) by construction.

If the control is observable as in the case with selection on observables, $g(x)$ can

[11] Certainly, whether this is a strong assumption depends on the choice of V. See Chapter 2 for the case with production functions.

be estimated using standard methods of generalized additive models.[12] However, the CFA with observable controls may not apply widely in empirical applications due to several reasons. First, X and V should have no common elements so that the function of interest $g(x)$ can be nonparametrically distinguished from the control function $c(v)$.[13] Second, as in the case with selection on unobservables, controls are often not readily observed but latent.

Alternatively, the control can be constructed from instruments Z:

$$V_i = X_i - r(Z_i), \text{ where } r(Z_i) \equiv E(X_i|Z_i). \tag{1.10}$$

Comparing (1.10) to (1.4), we see that V_i is constructed similarly to Rivers and Vuong (1988) except that (1.10) is now the residual of nonparametric regression. Here X_i is decomposed into two parts: $r(Z_i) = E(X_i|Z_i)$ is the one predicted by Z_i and V_i is the residual. We see that the CF estimation and IV estimation both use instruments, but in different ways. For the IV estimation as in dynamic panel models, the exogenous variations of instruments are used directly to form moment conditions for estimation. In contrast, in the CF estimation, the instruments are used to "purge" exogenous variations away from X so that only endogenous variations in X (i.e, $X_i - E(X_i|Z_i)$) are left, which then serve as the controls.

To find the conditions under which V is a valid control, note that

$$E(U|X,V) = E(U|r(Z)+V,V) = E(U|Z,V),$$

[12] See Hastie and Tibshirani (1990) for the iterative backfitting method; and see Newey 1994, Linton and Nielsen 1995, or Chen et al 1996 for the marginal integration method.

[13] In the labor example, some factors determining job participation decision also tend to affect how much to work. Therefore, the exclusion restriction is not satisfied.

where the second equality holds if $r(\cdot)$ is strictly monotone. We also need

$$E(U|Z,V) = E(U|V) \tag{1.11}$$

to get $E(U|X,V) = E(U|V)$. The condition (1.11) is weaker than the independence between V and (U,V) and allows for heteroskedasticity. In addition, (1.11) is neither more nor less general than the identifying assumption $E(U|Z) = 0$ in the nonparametric IV estimation.[14]

Identification in Nonparametric Control Function Models

Recently, the control function approach with *constructed controls* has been extended to nonparametric regression cases. Newey, Powell, and Vella (1999) consider nonparametric control function approaches (NPCFAs) to endogeneity in the context of triangular simultaneous equations models. They give the conditions for identification, consistency and asymptotic normality. Pinkse (2000) extends the asymptotic analysis to time series cases. Das, Newy and Vella (2003) add a selection mechanism (propensity score) upon the model considered by Newey, Powell, and Vella (1999).

The control function assumptions ((1.7) or (1.11)) do not guarantee the identification of $g(\cdot)$. Newey, Powell, and Vella (1999) give a sufficient and necessary condition for the identification, see also Matzkin (2006): *Both $g(x)$ and $c(v)$ are identified up to a location if and only if $m(x,v) = 0$ implies that $g(x)$ is a constant.* To see this, note that $m(x,v) \equiv E(Y|(X,V) = (x,v))$ is identified and uniquely determined by a random sample of (X,V) so that $g(x) + c(v) = m(x,v)$. Suppose that there are other real functions $g'(x)$ and $c'(v)$ such that $g'(x) + c'(v) = m(x,v)$,

[14] See Blundell and Powell (2003) for an extensive review of alternative approaches to endogeneity, including NPCFAs.

we have

$$[g(x) - g'(x)] + [c(v) - c'(v)] = 0.$$

Then, $[g(x) - g'(x)] = \bar{c}$ and $[c(v) - c'(v)] = -\bar{c}$, where $\bar{c}$ is a constant. The identification of $g(\cdot)$ essentially comes from the additivity structure of $g(\cdot)$ and $c(\cdot)$, which in turn comes from the additivity of the error term. For the case with constructed controls satisfying (1.11), *a sufficient condition is that the rank of* $(\partial r(z) \diagup \partial z)$ *equals the dimension of* x (along with some regularity conditions).[15] This is a nonparametric version of the usual rank condition and is usually satisfied unless Z affects X in some special ways. The intuition is that $g(x)$ is identified as long as Z generates sufficient exogenous variations in X so that the conditioning information set for the control function is different from that for the function of interest.

In the production function case, all current inputs contain some information about the productivity shock. It is crucial to find controls such that the control function assumption (1.7) can be justified and the identification conditions hold.

1.4 Main Contributions

Nonparametric Identification of Production Functions

In Chapter 2, I establish the nonparametric identification of production functions using the control constructed from instruments. The control is essentially the residual of nonparametric regression of current input levels against lagged input levels, i.e, $v_{it} = x_{it} - E(x_{it}|x_{i,t-1})$. With this choice of control, the restrictions

[15] The regularity conditions include the differentiability of g, c and r, and zero probability on the boundary of support. Another sufficient condition to identify $g(x)$ is that no functional relationship exists between X and V. See Newey, Powell, and Vella (1999) and Matzkin (2006) for the proof.

imposed to obtain nonparametric identification are mild. For the productivity shock a_{it}, it is assumed that a_{it} follows an exogenous Markov process, on which a firm i has some uncertainty. Specifically, in order to makes input decisions for x_{it}, the firm has to predict a_{it} based on $a_{i,t-1}$. So the input decision can be written as $x_{it} = x_t(x_{i,t-1}, a_{i,t-1})$, which is required to be increasing in $x_{i,t-1}$ given $a_{i,t-1}$, and to satisfy a rank condition. I show that production functions can be nonparametrically identified using this identification strategy.

These results are important for several reasons. First, both the production function $g(x)$ and the distribution of productivity are identified nonparametrically. Therefore, policy suggestions based on the nonparametric procedures proposed in this thesis are robust to misspecification of underlying data generating processes (DGPs). This becomes more crucial when economic theories of firm production do not suggest functional forms for $g(x)$ or for the distribution of productivity. Second, the assumptions to make for identification can be relatively easily justified in empirical applications, and the choice of controls/instruments is flexible. For instance, either $x_{i,t-1}$ or $x_{i,t-2}$ can be used as the instrument to construct the control. Third, nonparametric identification implies parametric identification. The proposed identification strategy also works for parametric models, such as the Cobb-Douglas production function.

As imposed in Olley and Pakes (1996), Levinsohn and Petrin (2003) and Ackerberg, Caves and Frazer (2006), the assumption that the productivity shock can be "inverted out" from input decisions appears to be restrictive for many empirical applications. Instead, we only need to "smooth out" the productivity shocks, which overcomes several limitations inherited in the "invert-out" assumption. For instance, multi-dimensional shocks, unobserved prices and random measurement errors in input variables are now allowed for in this framework.

More generally, as a general way to handle endogeneity, control function approaches apply to many empirical applications other than production functions. Control functions may be derived, in a structural way, from the institutional knowledge of endogeneity issues (e.g., how a person self-selects into a program). Alternatively, they may come from statistical properties of control variables or instrumental variables.

Nonparametric Control Function Estimators

The nonparametric estimation procedures proposed in Chapter 3 closely follows the identification strategy without imposing extra modeling assumptions. Thus the connection from the function of interest to the sample analog, and then to the actual estimate is clear and precise. Newey, Powell, and Vella (1999) and Pinkse (2000) study the asymptotic properties of series-based NPCF estimators. As issue with the proposed series estimators is that the optimal rate of uniform convergence is not achieved and the asymptotic variance is affected by preliminary estimators of controls.

I propose two kernel-based NPCF estimators, $\widehat{g}(x)$ and $\widetilde{g}(x)$. When the control V is observed, the estimator $\widetilde{g}(x)$ proceeds as in a generalized additive model. I establish the asymptotic normality for $\widetilde{g}(x)$ using a second-order U-statistic, where the asymptotic variance is derived naturally. More importantly, the extension to a third-order U-statistic allows me to establish the asymptotic normality for the estimator $\widehat{g}(x)$ when the control V is unobservable but preliminarily constructed.

Basically, $\widehat{g}(x)$ is a kernel-based alternative to its series counterparts proposed in Newey, Powell, and Vella (1999) and Pinkse (2000). $\widehat{g}(x)$ has some nice asymptotic properties. When the preliminary estimator $\widehat{V}$ of V converges fast enough, $\widehat{g}(x)$ asymptotically behaves as if the controls were observed. The optimal rate of uniform convergence can be achieved, and the asymptotic variance of $\widehat{g}(x)$ is free from the effect of preliminary estimators. As by-products of asymptotic analysis of $\widehat{g}(x)$ and $\widetilde{g}(x)$, better rates of uniform convergence (as compared to Ahn (1995)) are obtained for multiple-step kernel estimation (Proposition 3.1); I also extend generalized additive models to the case with constructed variables, and show that the asymptotic properties remain unaffected if the constructed variables converge sufficiently fast.

Semiparametric Control Function Estimators

As mentioned in Section 1.2, when we consider the Cobb-Douglas production function, a partial linear model like (1.6) arises, where the parametric part represents the production function and the nonparametric part is the control function to "smooth out" unobserved shocks. The identification strategy also work, where the control is nonparametrically constructed from lagged levels of inputs as instruments. This extends partially linear models to the case with preliminary kernel estimators.

In Chapter 4, I propose an estimator $\widehat{\beta}$ for β, which can be viewed as a density-weighted and preliminarily estimated version of Robinson (1988) or as a preliminarily estimated version of Li (1996). I give the conditions under which $\widehat{\beta}$ is still $\sqrt{n}$-consistent despite that the variables in the nonparametric part are constructed ones. Since Olley and Pakes (1996), Levinsohn and Petrin (2003) and Ackerberg, Caves and Frazer (2006) all consider the Cobb-Douglas production function, $\widehat{\beta}$ can be seen as a "smooth-out" extension of these "invert-out" counterparts.

The rest of thesis is organized as follows. In Chapter 5, a set of Monte-Carlo experiments indicates that the NPCF and SPCF estimators proposed in this thesis perform well in finite samples. In Chapter 6, I apply the identification strategy and estimation procedures to a Chilean panel data set to demonstrate the empirical relevance. Respectively, Appendices A, B and C collect the proofs and technical details for the uniform consistency and asymptotic normality of NPCF estimators, and the $\sqrt{n}$-consistency of the SPCF estimator. The tables and figures of Monte-Carlo simulation results are reported in Appendix D. The empirical results are collected in Appendix E.

Chapter 2

Nonparametric Identification of Production Functions

In this chapter, I generalize the method of replacement functions to control function approaches by "smoothing out," instead of "inverting out," the productivity shock. Under the given conditions, I establish the *nonparametric* identification of production functions, using controls constructed from instruments. I then propose a method to nonparametrically recover the distribution of productivity shocks.

2.1 The Challenge and Solutions to Estimating Production Functions

2.1.1 Methods of Replacement Functions

In contrast to the IV estimation, where the exogenous variations of instruments are exploited, Olley and Pakes (1996) and Levinsohn and Petrin (2003) suggest using the exogenous variations of proxies. OP propose using the *observed* investment decision $i_t(k_{it}, a_{it})$ to proxy the *unobserved* productivity shock a_{it}. If $i_t(k_{it}, a_{it})$ is strictly increasing in a_{it}, a_{it} can be inverted out as $a_{it} = i_t^{-1}(k_{it}, i_{it})$ and the production function can be rewritten as:

$$y_{it} = \beta_k k_{it} + \beta_l l_{it} + a_{it} + \epsilon_{it} \equiv \beta_l l_{it} + \Phi(k_{it}, i_{it}) + \epsilon_{it}, \tag{2.1}$$

where $\Phi(k_{it}, i_{it}) = \beta_k k_{it} + i_t^{-1}(k_{it}, i_{it})$ can be seen as the control function for β_l (but apparently not for β_k). β_l can be estimated using Robinson's (1988) partial linear model or by the OLS with $\Phi(k_{it}, i_{it})$ being nonparametrically approximated by polynomials. With $\widehat{\beta}_l$, $\Phi(k_{it}, i_{it})$ can be estimated by $y_{it} - \widehat{\beta}_l l_{it}$ and a_{it} can be computed from

$$a_{it} = i_t^{-1}(k_{it}, i_{it}) = \widehat{\Phi}(k_{it}, i_{it}) - \beta_k k_{it}, \tag{2.2}$$

which depends on β_k. In OP, a_{it} is assumed to follow a first-order Markov process so that a_{it} can be decomposed into two parts: the predicted part $E(a_{it}|a_{i,t-1})$ and the innovation part $a_{it} - E(a_{it}|a_{i,t-1}) \equiv \xi_{it}$. A key assumption is that k_{it} is actually determined at $t-1$ so that k_{it} is orthogonal to the innovation of a_{it}. This assumption gives the moment condition to identify and estimate β_k.[1]

One potential issue with OP method is data-driven: investments are often reported zeros in many datasets, in which cases the assumption on the strict monotonicity of $i_t(k_{it}, a_{it})$ is likely to be violated. One can just use those observations with positive investments which, however, could incur efficiency loss and potential selection bias. LP instead suggest using the intermediate input decision $w_t(k_{it}, a_{it})$ as the proxy: $a_{it} = w_t^{-1}(k_{it}, w_{it})$, given $w_t(k_{it}, a_{it})$ is also strictly increasing in a_{it}. Intermediate inputs (such as materials, fuel and electricity) are

[1] Alternatively, similar to the estimation of β_l, β_k can also be estimated by applying Robinson's (1988) method to $y_{it} - \widehat{\beta}_l l_{it} = \beta_k k_{it} + a_{it} + \epsilon_{it} = \beta_k k_{it} + E(a_{it}|a_{i,t-1}) + (\xi_{it} + \epsilon_{it})$.

seldom zeros if reported at all.[2] The estimation procedure of (β_k, β_l) in LP goes the same as in OP.[3]

2.1.2 Identification Issues and Alternatives

The replacement function method advocated by OP and LP is influential and has stimulated many empirical applications. Ackerberg, Caves and Frazer (2006, henceforth ACF), however, question the identification of β_l in the first step in OP/LP estimation procedure. The intuition is that k_{it}, l_{it}, i_{it} and m_{it} are optimal decisions of firm i so that l_{it} is collinear with $\Phi_t(k_{it}, i_{it})$ in (2.2) as long as all input decisions use the same conditioning information set (e.g. k_{it} and a_{it}). To see this, the optimal choice of labor stock is

$$l_{it} = l_t(k_{it}, a_{it}) = l_t\left(k_{it}, i_t^{-1}(k_{it}, i_{it})\right) \equiv \widetilde{l}_t(k_{it}, i_{it}), \qquad (2.3)$$

which is a function of (k_{it}, i_{it}) too. They search for DGPs that maintain the identification of (β_k, β_l) under the framework of OP/LP, and find that the candidate DGPs entail strong assumptions.[4] The main reason for such a discouraging conclusion is the assumption that a_{it} is the only scalar unobservable affecting input decisions so that a_{it} can be inverted out from input decisions (i.e. $i_t(k_{it}, a_{it})$ in OP and $w_t(k_{it}, a_{it})$ in LP). In this paper, I relax this "invert-out" requirement to an "expect-out" one, which allows multiple shocks and flexible timing structures.

[2] In the Chilean panel data used in Levinsohn and Petrin (2003), Ackergerg, Caves and Frazer (2006) and this paper, about 50% observations see zero investments. On the contrary, postive levels of intermediate inputs are reported at over 90% observations.

[3] There is, however, an important difference between OP and LP methods: $i_t(k_{it}, a_{it})$ is a dynamic choice affecting production in future while $w_t(k_{it}, a_{it})$ is usually a static/interim choice affecting current production. As a result, the conditioning information set for $i_t(k_{it}, a_{it})$ is likely to be different from that for $w_t(k_{it}, a_{it})$.

[4] See ACF for a detailed and enlightening discussion on how the timing of events affects the identification.

ACF propose a new procedure based on the idea of OP/LP with the spirit of dynamic panels. They give up estimating β_l in the first step in OP/LP procedures, given that it is not identified. Using the intermediate input decision $m_{it} = m_t(k_{it}, l_{it}, a_{it})$ to proxy the productivity shock a_{it} by $w_t^{-1}(k_{it}, l_{it}, w_{it})$,[5] ACF first "net out" the non-transmitted error ϵ_{it} by a nonparametric regression of y_{it} on (k_{it}, l_{it}, w_{it}):

$$y_{it} = \beta_k k_{it} + \beta_l l_{it} + w_t^{-1}(k_{it}, l_{it}, w_{it}) + \epsilon_{it} \equiv \Phi_t(k_{it}, l_{it}, w_{it}) + \epsilon_{it}; \tag{2.4}$$

The second step then "isolates out" a_{it} from the composite term $\Phi_t(k_{it}, l_{it}, w_{it})$ by

$$a_{it} = \Phi_t(k_{it}, l_{it}, w_{it}) - \beta_k k_{it} - \beta_l l_{it}, \tag{2.5}$$

which depends on $\beta \equiv (\beta_k, \beta_l)$.[6] The assumption on the timing implies that the innovation of a_{it} is orthogonal to both k_{it} and $l_{i,t-1}$, i.e. $\xi_{it} \perp (k_{it}, l_{i,t-1})$, which provides the moment conditions to identify and estimate β. They thus provide a clever way to construct moment conditions to identify and estimate β without imposing strong assumptions on the structure of error terms. Wooldridge (2005) also suggests estimating both β_k and β_l simultaneously by GMM, where both current state variables (capital) and lagged inputs are used as instruments.

The $\sqrt{n}$-consistency and asymptotic normality have not been established in OP, LP and ACF, and the estimation is computation-intensive.[7] Furthermore,

[5] The investment decision $i_{it} = i_t(k_{it}, l_{it}, a_{it})$ can also be used to invert a_{it} out and the estimation proceeds similarly. Notice that these two proxies may entail different timing assumptions. Also notice that in ACF m_{it} depends on l_{it} as well as k_{it}.

[6] The difference between OP/LP and ACF procedures can be seen by comparing (2.2)-(2.3) to (2.5)-(2.6).

[7] Note that one needs to search over the parameter space of (β_k, β_l) with multi-step nonparametric estimation of $\Phi(k_{it}, l_{it}, w_{it})$, a_{it} and ξ_{it} at each iteration.

it is well-known that nonparametric regression is inconsistent near the boundary of the support. Thus, it is necessary to apply some trimming to the estimation procedures, which is not explicitly addressed in these papers.

One further concern is the potential misspecification of production functions. Restrictions imposed in estimation should reflect the industry of interest, and misspecification may give misleading estimates and policy suggestions. All the works discussed above assume the Cobb-Douglas production function. This specification implies that the production technology of an industry can be summarized by two parameters, β_k and β_l, which may not be a good approximation to many industries.[8] In particular, Bond and Söderbom (2005) shows that parameterizations like (2.1) or (2.4) are subject to misspecification, especially for the case with adjustment costs expressed in the form of lost output. Unfortunately, as mentioned in Chapter 1, it is difficult to extend the IV method or the method of replacement function to nonlinear or nonparametric cases, in terms either of asymptotic analysis or of empirical implementation.

2.2 Nonparametric Identification Using Control Function Approaches

In this section, I develop a strategy to nonparametrically identify production functions. The identification strategy consists of two elements: the control function to "smooth out" (or to "expect out") the unobservable shocks, and firms' uncertainty about productivity shocks.

[8] For instance, the technology applied by large firms may be quite different from the one by small firms. We will see such a case in Chapter 6, where the food industry in Chile in 1980's is examined.

2.2.1 Productivity Shocks: To *Invert Out*, or To *Smooth Out*?

The idea of OP, LP and ACF relies on the availability of *perfect* proxies. They all assume that productivity shocks can be perfectly proxied by "inverting out" a_{it} from observable input decisions.[9] This assumption has several important implications. First, the input decision must be strictly increasing in a_{it} to invert a_{it} out, which could be difficult to justify in empirical applications as mentioned in the case with $i_t(k_{it}, a_{it})$. Second, the unobserved shock can only be a scalar, which is reasonable only if all relevant shocks can be summarized by a single index. Third, there are no unobserved firm-level prices, which are important determinants of firms' production behavior but are often absent in many datasets. Fourth, there are no measurement errors in inputs, which could be prevalent in datasets collected from surveys (Angrist and Krueger, 1998). The last three items add additional unobservables to the input decisions and make the inversions impossible. In sum, *perfect* proxies ask too much from data and actually prevent some identification strategies.

In fact, to handle the endogeneity problem in regression models, we don't need such a strong assumption: we only need to "expect out" rather than to "invert out" (or to "solve out" as in Heckman and Vytlacil, 2006) the unobserved shocks. To see this, note that $E(y_{it}|k_{it}, l_{it}) = \beta_k k_{it} + \beta_l l_{it} + E(a_{it}|k_{it}, l_{it})$, where $E(a_{it}|k_{it}, l_{it}) \neq 0$ so that we need to control for $E(a_{it}|k_{it}, l_{it})$. In the method of replacement function, restrictive structures are imposed to model a_{it} directly, i.e, $a_{it} = i_t^{-1}(k_{it}, i_{it})$ in OP

[9] If a_{it} cannot be inverted out, neither of (2.1), (2.2), (2.4), and (2.5) is well defined.

and $a_{it} = m_t^{-1}(k_{it}, m_{it})$ in LP. Instead, if we can find a control v_{it} such that

$$E(a_{it}|k_{it}, l_{it}, v_{it}) = E(a_{it}|v_{it}) \equiv c_t(v_{it}), \tag{2.6}$$

we have an augmented regression function with v_{it} as an additional regressor:

$$y_{it} = g_t(k_{it}, l_{it}) + c_t(v_{it}) + \varepsilon_{it}, \text{ where } \varepsilon_{it} = a_{it} + \epsilon_{it} - c_t(v_{it}) \tag{2.7}$$

With the control function $c_t(v_{it})$, the regressors now become exogenous to the new error term ϵ_t. The control function assumption (2.6) means that the control function $c_t(v_{it})$ is sufficient in evaluating the conditional mean $E(a_{it}|\cdot)$ so that other regressors provide no extra information. Since it is not required to invert a_{it} out, multi-dimensional shocks, unobserved prices or measurement errors in inputs are allowed, as long as (2.6) holds.[10]

Now the issue is that $c_t(v)$ is unknown. As mentioned in the introduction, we can treat $c_t(v)$ nonparametrically. Furthermore, we can also treat the production function $g_t(k, l)$ nonparametrically. Once we establish the nonparametric identification of production functions, the parametric identification follows. Given $g_t(k, l)$ is identified, $g_t(k, l)$ can be estimated as a partial mean of $E(y_{it}|k_{it}, l_{it}, v_{it}) \equiv m_t(k_{it}, l_{it}, v_{it})$:

$$g_t(k, l) = E[m_t(k_{it}, l_{it}, v_{it}) \,|\, (k_{it}, l_{it}) = (k, l)], \tag{2.8}$$

where $m_t(k, l, v)$ can be consistently estimated from a sample $\{(k_{it}, l_{it}, v_{it})\}_{i=1}^n$ and the location normalization $E[c_t(v_{it})] = 0$ is imposed. By the law of iterated

[10]For the input decisions to be informative about a_{it}, it is still necessary for input decisions to be increasing in a_{it}. This monotonicity, however, can be weak as long as no information is lost in predicting a_{it}.

expectation, $E[c_t(v_{it})] = E[a_{it}] = 0$, which is also the location normalization for the distribution of productivity.

This "smooth out" strategy alone, however, does not fix the endogeneity problem caused by functional relationships among inputs, prices and shocks. In fact, something can always be computed from (2.8), and $E[m_t(k_{it}, l_{it}, v_{it}) \mid (k_{it}, l_{it}) = (k, l)]$ does not necessarily correspond to $g_t(k, l)$. This is an identification issue, essentially a nonparametric version of the collinearity problem arising in estimating β_l in OP/LP. As mentioned in Chapter 1, the identification depends on the choice of v. Although it is tempting to put either i_{it} or w_{it} into v_{it}, as do in OP/LP, it is difficult to justify the control function condition (2.6), and to obtain the identification at the same time. It is apparent that x_{it} and v_{it} cannot have common elements so that any element in x_{it} (e.g., k_{it} or l_{it}) cannot be include into v_{it}. On the other hand, $v_{it} = i_{it}$ (or $v_{it} = w_{it}$) alone is not sufficient to be a control.[11] An alternative is to use $v_{it} = (i_{it}, x_{i,t-1})$ (or $v_{it} = (w_{it}, x_{i,t-1})$) instead. However, it is also difficult to justify (2.6) because w_{it} (or i_{it}) typically is correlated to (k_{it}, l_{it}). Thus, the key is to find a control v to nonparametrically identify production functions under the control function assumption (2.6).

2.2.2 Firm Production with Interim Uncertainty of Productivity

The endogeneity problem arises when some shocks are observed by firms but not by researchers. However, firms themselves often face uncertainty and only get noise-ridden signals of shocks. Thus, firms have to take uncertainty into account

[11] The control function condition (2.6) means that controls move along with a_{it} such that conditioning on them best predicts a_{it}. The same value of i_{it} may mean high a_{it} for small firms, but low a_{it} for large firms. We need some benchmark along with i_{it} to better predict a_{it}.

when they make decisions. Interestingly, uncertainty faced by firms may actually help the identification of production functions.

The firm production under uncertainty has been explored in the literature to study firm turnover and industry evolution. In Jovanovic (1982), a firm i does not know its own cost parameter c_i but each period draws a noisy signal c_{it} about c_i from some known distribution.[12] The industry dynamics are then driven by each firm's entry/exit decisions based on the inference of c_i by $\frac{1}{T}\sum_{t=1}^{T} c_{it}$. But it takes too long (the whole lifetime) for the firm to learn c_i, which makes the firm's decisions depend on its entire history. Hopenhayn (1992) instead assumes that firms directly observe the productivity shock a_{it} before production and the uncertainty faced by the firm becomes the need to predict $a_{i,t+1}$ given a_{it} to make entry/exit decisions. It is assumed that a high a_{it} means $a_{i,t+1}$ tends to be high too, which makes exit less likely for the firm i as an incumbent (or entry more likely as a potential entrant), vice versa. With industry evolution in mind, the dynamic decisions focused by Hopenhayn (1992) are the exit or entry of firms under intertemporal uncertainty. I introduce interim uncertainty faced by firms when they make input decisions.

Consider the following firm production with uncertainty about productivity. At the beginning of period t, firm i observes the vector of state variables $(x_{i,t-1}, a_{i,t-1})$ predetermined at $t-1$. The firm faces some *uncertainty* in the sense that it cannot directly observe nor perfectly predict the productivity shock a_{it}. Since a_{it} is not observed, to make input decisions, the firm predicts a_{it} by $E(a_{it}|a_{i,t-1})$ as a_{it} follows an exogenous first-order Markov process. Denote the input decision at period t as $x_{it} = x_t(x_{i,t-1}, a_{i,t-1})$. The firm adjusts the input from $x_{i,t-1}$ to x_{it} and

[12] It is c_i that affects the firm's cost (not c_{it}, the signal of c_i). By the duality, c_i corresponds to a_{it}. Notice the difference bewteen c_i and a_{it}: c_i is time-invariant while a_{it} is not.

begins the production. The firm learns about the true value of a_{it} at the end of period t.

2.2.3 Nonparametric Identification with Constructed Controls

We need to find *primitive* conditions that are sufficient for the identification of the production function $g_t(x)$. The non-transmitted error ϵ_{it} is dismissed by Assumption 2.1. I impose a Markov structure on the productivity shock a_{it} in Assumption 2.2. In Assumption 2.3, some restrictions are imposed on the input decision $x_{it} = x_t(x_{i,t-1}, a_{i,t-1})$ and the distribution of a_{it}. Assumptions 2.2 and 2.3 are both consistent with the model of firm production with interim uncertainty of productivity as described in Section 2.2.2.

Assumption 2.1: *For all* (i,t), $E(\epsilon_{it}|x_{it}, x_{i,t-1}, \cdots, x_{i1}) = 0.$[13]

Assumption 2.2: *For all* (i,t), $\mathcal{F}_{it}$ *is the information set before production and*

$$\mathcal{F}_{it} \equiv \{x_{it}, x_{i,t-1}, a_{i,t-1}, ..., x_{i1}, a_{i1}\};$$

The productivity a_{it} *follows an exogenous First-order Markov process such that*

$$E(a_{it}|\mathcal{F}_{it}) = E(a_{it}|a_{i,t-1}).$$

Assumption 2.3: *For all* (i,t), *the input decision* $x_{it} = x_t(x_{i,t-1}, a_{i,t-1})$ *and its partial derivative* $\partial x_t(x,a)/\partial x$ *are both continuous over the compact support of*

[13] Only $E(\epsilon_{it}|x_{it}, x_{i,t-1}) = 0$ is necessary. Wooldridge (2005) argues that it is *ad hoc* to assume conditional mean independence given outcomes at t and $t-1$, without also assuming $E(\epsilon_{it}|x_{it}, x_{i,t-1}, \cdots, x_{i1}) = 0$.

(x,a); and $\partial x_t(x,a) \diagup \partial x$ and the distribution of productivity $f(a)$ satisfies

(i). $\int \frac{\partial}{\partial x} x_t(x,a) f(a)\, da > 0$ *and*

(ii). *The rank of $\int \frac{\partial}{\partial x} x_t(x,a) f(a)\, da$ equals the dimension of x.*

Assumption 2.1 is standard in the literature, see OP, LP, ACF and Wooldridge (2005) for instance. Assumption 2.1 says nothing about the dependence structure among $\{\epsilon_{it}\}_t$ and actually allow for serial dependence in $\{\epsilon_{it}\}_t$ as neither y_{it}'s nor ϵ_{it}'s appear in the conditioning set $\mathcal{F}_{it}$.

Assumption 2.2 means that the firm's expectation about a_{it} depends only on $a_{i,t-1}$ as long as a_{it} has not been learnt. Although it is more restrictive than $E(a_{it}|a_{i,t-1}, a_{i,t-2}, ..., a_{i1}) = E(a_{it}|a_{i,t-1})$, it is reasonable given the uncertainty faced by the firm. Indeed, OP assume that investments take one period to complete so that k_{it} is determined at $(t-1)$ and part of $\mathcal{F}_{it}$. Similarly, LP, ACF and Wooldridge also assume that dynamic inputs in x_{it} belong to $\mathcal{F}_{it}$. However, they assume non-dynamic inputs do not belong to $\mathcal{F}_{it}$. There is discrepancy about what should be treated as dynamic inputs. For instance, k_{it} is a dynamic input but l_{it} is not in OP while both are dynamic inputs in ACF and Wooldridge (2005). Unless we have some institutional knowledge of the industry of interest, it is difficult to make such a call. Here, I resort to firms' interim uncertainty of productivity.

Assumption 2.2 is also compatible with firm production with *adjustment costs* of inputs. Suppose that a firm i first solves its dynamic programming problem to set $x_{it} = x_t(x_{i,t-1}, a_{i,t-1})$. After incurring adjustment costs, inputs adjust from $x_{i,t-1}$ to x_{it} and the firm begins production, during which a_{it} realizes sequentially within the period t. Even if the firm knows a_{it} and finds that x_{it} are not at the optimal levels (under usual marginal productivity conditions) given the realization of a_{it}, the adjustment costs prevent the firm from changing the levels of x_{it} at will.

In the uncertainty story, the firm learns about a_{it} after period t production has finished. In the adjustment cost story, however, a_{it} can be learnt right after x_{it} are set. These two arguments can be combined together to allow for more flexible data generating processes.

Assumption 2.3 imposes restrictions on the input decision $x_{it} = x_t(x_{i,t-1}, a_{i,t-1})$ and the distribution of productivity $f(a)$. A sufficient condition for Assumption 2.3.(i) is that $x_t(x, a)$ *is strictly increasing in* x *given any value of* a, which is stronger than Assumption 2.3.(i) but still a reasonable condition. All other things (especially the productivity) being equal, a larger firm tends to have higher levels of inputs like capital and labor. Note that the rank condition on $\int \frac{\partial}{\partial x} x_t(x, a) f(a)\, da$ is not implied by the one on $\frac{\partial}{\partial x} x_t(x, a)$.

Given Assumptions 2.1-2.3, I choose $z_{it} = x_{i,t-1}$ as the instrument and construct the control v_{it} as follows:

$$v_{it} \equiv x_{it} - r(x_{i,t-1}), \text{ where } r(x_{i,t-1}) = E(x_{it} | x_{i,t-1}). \tag{2.9}$$

$r(x_{i,t-1})$ is the projection of x_{it} into $x_{i,t-1}$, showing the effect of previous input levels on the choice of current input level. v_{it} can be interpreted as the response of input decision x_{it} to the firm's prediction of a_{it}, which is based on $a_{i,t-1}$. Lagged input levels other than $x_{i,t-1}$, say $x_{i,t-2}$, can also serve as the instrument.

The restrictions imposed on the input decision $x_{it} = x_t(x_{i,t-1}, a_{i,t-1})$ and the distribution of productivity $f(a)$ imply that $r(x)$ is strictly increasing in x and the rank of $dr(x)\diagup dx$ equals the dimension of x. To see this, by definition,

$$r(x) = E(x_t(x_{i,t-1}, a_{i,t-1}) | x_{i,t-1} = x) = E(x_t(x, a_{i,t-1})),$$

where the expectation is with respect to $a_{i,t-1}$. Since $x_t(x,a)$ and $\partial x_t(x,a) \diagup \partial x$ are both continuous over the compact support of (x,a), by the Leibniz's rule

$$\frac{dr(x)}{dx} = \frac{d}{dx}E\left(x_t\left(x,a_{i,t-1}\right)\right) = \int \frac{\partial}{\partial x}x_t(x,a)\, f(a)\, da.$$

Therefore, $r(x)$ is strictly increasing in x by Assumption 2.3.(i) and, by Assumption 2.3.(ii), the rank of $dr(x) \diagup dx$ equals the dimension of x. I summarize these results in Lemma 2.1 below.

Lemma 2.1. *For all* (i,t), *under Assumption 2.3, for* $r(x) \equiv E\left(x_t\left(x,a_{it}\right)\right)$

(i). $r(x)$ *is strictly increasing in* x*;*

(ii). *The rank of* $dr(x) \diagup dx$ *equals the dimension of* x*.*

Adopting the control function approach to endogeneity, I establish the nonparametric identification of the production function $g_t(x)$ under Assumptions 2.1-2.3, using lag levels of inputs as the instrument to construct the control.

Proposition 2.1. *Under Assumptions 2.1-2.3, the production function* $g_t(x)$ *is nonparametrically identified with the control* v *being constructed by (2.9).*

Proof: Given Assumption 2.1, the endogeneity is caused by the correlation between x_{it} and a_{it}. By Lemma 2.1.(ii), the rank $dr(x) \diagup dx$ equals the dimension of x so that the rank condition is satisfied. It remains to check the control function condition (2.6).

By the law of iterated expectation, we have

$$E\left(a_{it}|x_{it}, x_{i,t-1}, a_{i,t-1}\right) = E\left(a_{it}|a_{i,t-1}\right);$$

Notice that v_{it} is a function of $(x_{it}, x_{i,t-1})$, by the law of iterated expectation,

$$E(a_{it}|v_{it}, a_{i,t-1}) = E(a_{it}|a_{i,t-1}).$$

Together, $E(a_{it}|x_{it}, x_{i,t-1}, a_{i,t-1}) = E(a_{it}|v_{it}, a_{i,t-1})$, which implies

$$E(a_{it}|x_{it}, x_{i,t-1}) = E(a_{it}|v_{it}). \tag{2.10}$$

The control function condition (2.6) then holds for $z_{it} = x_{i,t-1}$:

$$\begin{aligned} E(a_{it}|x_{it}, v_{it}) &= E(a_{it}|x_{it}, x_{it} - r(x_{i,t-1})) \\ &= E(a_{it}|x_{it}, x_{i,t-1}) \\ &= E(a_{it}|v_{it}), \end{aligned}$$

where the first equality follows by the definition of v_{it}, the second one by the strict monotonicity of $r(x)$ from Lemma 2.1.(i), and last one by (2.10). Therefore, (2.6) holds under Assumption 2.2 with the control being defined by (2.9).

Applying Theorem 2.3 in Newey, Powell, and Vella (1999),[14] the production function $g_t(x)$ is nonparametrically identified with the control v being constructed by (2.9) under Assumptions 2.1-2.3. $\square$

Proposition 2.1 establishes the nonparametric identification of $g_t(x)$ under Assumptions 2.1-2.3, which can be consistently estimated by the kernel-base nonparametric control function estimator $\widehat{g}(x)$ proposed in Chapter 3. The consistency and asymptotic normality of $\widehat{g}(x)$ are established under the given conditions in Chapter 3. Since nonparametric identification implies parametric identification, if

[14] See also Theorem 4.5 in Matzkin (2007).

we consider the Cobb-Douglas production function, $g_t(x;\beta) = x'\beta$, β can be identified using the same strategy. In Chapter 4, the semiparametric control function estimator $\widehat{\beta}$ is proposed for β, and the $\sqrt{n}$-consistency of $\widehat{\beta}$ is established under the given conditions.

The identification strategy and estimation procedure proposed in this paper has several advantages relative to the methods in the literature. First, as a result of "smooth-out" strategy, the restrictions associated with "invert-out" assumption are not required. In particular, the main assumption about the productivity shock a_{it} is the Markov property as specified in Assumption 2.2. Second, it is robust to the misspecification of underlying DGPs, not only because both identification and estimation are nonparametric, but also because the restrictions imposed on the DGPs are flexible. The within-period uncertainty is sufficient to justify Assumption 2.2. Third, it is not demanding of data as it only requires a panel of x_{it} for two periods. No other observed variables such as investments and intermediate inputs are necessary. Finally, the identification strategy applies to both the gross production function where $x_{it} = (k_{it}, l_{it}, w_{it})$ and the value-added production function where $x_{it} = (k_{it}, l_{it})$, as long as Assumptions 2.1-2.3 hold.

2.3 Recover the Distribution of Productivity

As the industry average of firm outputs, $g_t(\cdot)$ is identical across firms at the same period, and the heterogeneity of firms is mainly captured by the productivity shock a_{it}. It is a_{it} that drives the turnover of firms and evolution of industry. Thus, besides $g_t(\cdot)$, it is desirable to recover a_{it} from data as well.[15] Since OP/LP use

[15] In fact, besides the endogeneity caused by the simultaneity of inputs, another issue addressed by Olley and Pakes (1996) is the industry evolution induced by entry and exit decisions of firms. Levinsohn and Petrin (1999) also consider both issues using the Chilean panel data.

the invert-out method, it is not surprising that a_{it} can be estimated, say by (2.2).

With the "smooth-out" method proposed in this paper, nevertheless, a_{it} can still be recovered as follows. Under Assumption 2.1, the non-transmitted error ϵ_{it} can be isolated by: $\widehat{\epsilon}_{it} = y_{it} - \overline{g}_t(x_{it})$, where $\overline{g}_t(x)$ is the estimate of $E(y_{it}|x_{it} = x)$. Note that $\overline{g}_t(x)$ is not a consistent estimator of $g_t(x)$, but actually estimates $g_t(x) + E(a_{it}|x_{it} = x)$. After $g_t(x)$ is consistently estimated by $\widehat{g}(x)$ using the control function approach, the composite error $u_{it} \equiv a_{it} + \epsilon_{it}$ can also be recovered as $\widehat{u}_{it} = y_{it} - \widehat{g}_t(x_{it})$. The key to nonparametrically recover idiosyncratic productivity shocks relies on the availability of a consistent estimator of $g_t(x)$. The idiosyncratic productivity shock a_{it} is then estimated by

$$\widehat{a}_{it} = \widehat{u}_{it} - \widehat{\epsilon}_{it} = \overline{g}_t(x_{it}) - \widehat{g}_t(x_{it}). \tag{2.11}$$

Using the estimates $\widehat{a}_{it}$'s as pseudo-values of a_{it}, the empirical distribution of a_{it} can be estimated.[16]

Therefore, without imposing any functional form either on $g_t(\cdot)$ or on the true distribution of a_{it}, we can recover the distribution of productivity *nonparametrically*. This is desirable because, typically, little is known about the distribution of productivity and economic theory gives no clue about its functional form either. In addition, the conditional mean of a_{it} given $x_{it} = x$ can also be recovered, in two ways. One is to regress $\widehat{a}_{it}$ on x_{it}, and the other is by $\widehat{E}(a_{it}|x) = \overline{g}_t(x) - \widehat{g}(x)$. $E(a_{it}|x)$ reveals how a_{it} is correlated to x_{it}, and sheds light on the endogeneity issue of inputs. This is of interest theoretically and practically. In Chapter 6, I estimate $E(a_{it}|x)$ for the food industry using a Chilean panel data set.

[16] See Chapter 4 and Guerre, Perrigne and Vuong (2000) for the estimation of distribution from preliminary estimates.

Chapter 3

Nonparametric Control Function Estimation

The nonparametric estimation of production functions closely follows the identification strategy developed in Chapter 2. I propose a kernel estimator $\widehat{g}(x)$ where controls are constructed from instruments as in (2.9). A kernel estimator $\widetilde{g}(x)$ is also proposed for the case with observed controls, which facilitates asymptotic analysis of $\widehat{g}(x)$. The consistency and asymptotic normality are established for both $\widetilde{g}(x)$ and $\widehat{g}(x)$ under the given conditions.

3.1 Nonparametric Models and Control Function Estimators

First, let's summarize the model of nonparametric regressions with endogeneity using general notation.[1]

[1] In this section, random variables (or vectors) are denoted by capital letters with their realizations by corresponding small letters.

Assumption M (Models): *Suppose we observe a representative random sample of size n, either $\{Y_i, X_i, V_i\}_{i=1}^n$ when the control V is observed, or $\{Y_i, X_i, Z_i\}_{i=1}^n$ when the instrument Z is observed. $Y \in \mathbb{R}$, $X = (X_1, X_2) \in \mathbb{R}^{d_1} \times \mathbb{R}^{d_2}$, $V \in \mathbb{R}^{d_2}$ and $Z \in \mathbb{R}^{d_Z}$, where $d = d_1 + d_2 \geqslant 1$ and $d_Z \geqslant 1$.*

(i) $Y = g(X) + U$, where $E(U|X) \neq 0$ and $E(U) = 0$;

(ii) The control V satisfies the control function assumption: $E[U|X,V] = E[U|V]$. When V is unobservable (to researchers), V can be estimated from the instrument Z for the endogenous variable X_2 by $X_2 = r(Z)+V$, where $E[V|Z] = 0$ and $E[U|X_1, X_2, V] = E[U|X_1, r(Z) + V, V] = E[U|V]$; $r(z)$ is continuous and strictly monotone in z and the rank of the Jacobian matrix of $r(z)$ equals d_2.

Under Assumption M, the conditional mean $m(x, v)$ of Y given $(X, V) = (x, v)$ satisfies

$$m(x, v) = g(x) + c(v), \tag{3.1}$$

where $c(v) \equiv E[U|V = v]$. For series-based estimation, both $g(x)$ and $c(v)$ can be simultaneously estimated from (3.1) by imposing additivity of x and v on base functions. For kernel-based estimation, no such restriction can be imposed on kernels to estimate $g(x)$ directly. Instead, $m(x, v)$ has to be estimated first as an augmented regression of Y on (X, V) from either $\{Y_i, X_i, V_i\}_{i=1}^n$ or $\{Y_i, X_i, Z_i\}_{i=1}^n$, which is consistent after introducing the control V under Assumption M.

Newey, Powell, and Vella (1999) consider a similar model with unobservable controls in a simultaneous equations setting, where X_1 is a part of Z so that typically $d_Z \geqslant d$. They give the conditions under which $g(x)$ is identified, and establish the consistency and asymptotic normality of $\widehat{g}(x)$. However, the uniform convergence rate of $\widehat{g}(x)$ is always affected by the preliminary estimator $\widehat{V}$ of V.

As a result, it is impossible to achieve the optimal rate of uniform convergence as derived in Stone (1982). Additionally, the asymptotic variance of $\widehat{g}(x)$ is always affected by $\widehat{V}$. Pinkse (2000) also considers a similar model with unobserved controls, where he derives the optimal rate of pointwise convergence for the *i.i.d.* case and establishes the uniform consistency for stationary time series.

Härdle (1994) argues that series estimators are asymptotically equivalent to their kernel counterparts in standard nonparametric regressions. This is no longer true due to two complications involved here: the partial mean of kernel estimates, and the preliminary nonparametric estimates as nuisance parameters. These two complications bring challenges to asymptotic analysis of $\widehat{g}(x)$.

The first complication arises from the fact that $g(x)$ has to be estimated as a partial mean of $m(x, v)$ given x, as in a generalized additive model:

$$g(x) = E[m(x, V)]. \tag{3.2}$$

This may be done either by the iterative backfitting method (see Hastie and Tibshirani 1990, among others) or by the marginal integration method (see Newey 1994, Linton and Nielsen 1995, or Chen et al 1996, among others). Generalized additive models (GAMs) are originally motivated as a method of dimension reduction to alleviate the curse of dimensionality in nonparametric estimation. Here the additivity structure comes from the additive error terms. As a simpler method, the marginal integration method is adopted in this paper.

When V is observed, the sample analog of (3.2) is

$$\widetilde{g}(x) = n^{-1} \textstyle\sum_{i=1}^{n} \widetilde{m}(x, V_i), \tag{3.3}$$

where $\widetilde{m}(x, V_i)$ is the kernel estimator of $m(x, v)$ evaluated at (x, V_i). The asymptotic properties of $\widetilde{g}(x)$ have been well studied in the literature. Nevertheless, I propose an alternative way to establish the asymptotic normality of $\widetilde{g}(x)$ using a U-statistic. As shown in Proposition 3.2, $\widetilde{g}(x)$ can be expressed as some form of a sample mean, from which the asymptotic variance of $\widetilde{g}(x)$ is derived naturally. Moreover, it allows me to extend the GAM to the case with generated regressors.

The second complication is common to both the series and kernel estimators. Regressions with generated regressors in parametric models have been considered by Pagan (1984) among others. Extensions to kernel regressions with generated variables and additive error terms (e.g., $\widehat{m}(x, v)$ in (3.4) below) can be found in Ahn (1995) and Rilstone (1996).[2] There is no endogeneity issue with regressors in both papers, and the optimal uniform convergence rate as derived in Stone (1982) is not obtained because the approximation is not sharp enough.

When V is unobserved and has to be estimated first, the sample analog of (3.2) becomes

$$\widehat{g}(x) = n^{-1} \textstyle\sum_{i=1}^{n} \widehat{m}\left(x, \widehat{V}_i\right), \tag{3.4}$$

where $\widehat{m}\left(x, \widehat{V}_i\right)$ is the kernel estimator of $m(x, v)$ with preliminary estimates $\widehat{V}_i$'s, which are also kernel estimates based on $V_i = X_{2i} - r(Z_i)$. Thus, $\widehat{g}(x)$ is a 3-step estimator with kernel estimators $\widehat{m}$ and $\widehat{V}$ as nuisance parameters. Upon $\widetilde{g}(x)$, $\widehat{V}$ adds an additional layer of difficulty in analyzing the asymptotic properties of $\widehat{g}(x)$. Nonetheless, the consistency and asymptotic normality of $\widehat{g}(x)$ are established and, in particular, the optimal rates of uniform or pointwise convergence are possible.

[2] In Ahn (1995), the generated variable represents the expected return of schooling, which is used in the second step to evaluate the conditional choice probabilities of schooling decisions of high school graduates under uncertainty. In Rilstone (1996), the generated variable mainly acts as a dimension reduction tool by collapsing the information contained in several variables into the generated one. Ahn (1995) establishes both uniform consistency and asymptotic normality while Rilstone (1996) only considers the latter.

However, one more challenge is coming. It is well known that kernel estimators are inconsistent near the boundary of supports, called the boundary effects. Thus, both (3.3) and (3.4) are inconsistent without controlling for the boundary effects of preliminary kernel estimators (i.e., $\widetilde{m}$ and $\left(\widehat{m}, \widehat{V}\right)$ respectively). We need to add some trimming both to (3.3) and to (3.4), and the asymptotic properties of the estimators are examined within inner compact subsets of their supports.

3.2 Estimation Procedures

Let S_{XV}, S_{XZ}, S_X, S_V, and S_Z be the supports of (X,V), (X,Z), X, V and Z respectively. Let's consider an inner compact subset C_X of S_X for $g(x)$. As $g(x)$ is a partial mean of $m\left(x,v\right)$, I study $m\left(x,v\right)$ for (x,v) belonging to an inner compact subset C_{XV} of S_{XV} such that $\{x \in S_X : (x,v) \in C_{XV}\} = C_X$. Note that for $(x,v) \in C_{XV}$, the estimators of $m\left(x,v\right)$ use at most the observations in C'_{XV}, where $C'_{XV} \subsetneq S_{XV}$ is the set containing all hypercubes of size ϵ (small enough) centered at a point $(x,v) \in C_{XV}$ so that $C_{XV} \subsetneq C'_{XV}$. For $x \in C_X$, define $C_V^x \equiv \{v \in S_V : (x,v) \in C_{XV}\}$ and $C_Z^x \equiv \{z \in S_Z : z = r^{-1}\left(x - v\right), v \in C_V^x\}$.

Now I describe the estimation procedure of the nonparametric control function (NPCF) estimators $\widetilde{g}(x)$ and $\widehat{g}(x)$ *with trimming.*

Step 1: Generation of the Control Variable $\widehat{V}$

When V is observed, this step is unnecessary. If V is unobserved, it can be estimated by $\widehat{V} = X_2 - \widetilde{r}(Z)$, where $\widetilde{r}(Z)$ is the kernel estimator of $r(z)$:

$$\widetilde{r}(z) = \widehat{E}[X_2|Z = z] = \frac{1}{n}\sum_{l=1}^{n} X_{2l}K_h\left(z - Z_l\right) \big/ \widetilde{f}_z(z), \tag{3.5}$$

where $\widetilde{f}_z(z) \equiv \frac{1}{n}\sum_{l=1}^{n} K_h(z - Z_l)$ is the density estimator, $K_h(z - Z_l) \equiv \frac{1}{h_1^{d_Z}}\kappa\left(\frac{z-Z_l}{h_1}\right)$ is the kernel and h_1 is the bandwidth.[3] Note that $\widetilde{r}(z)$ is inconsistent for z near the boundary. The estimated control variable is defined as

$$\widehat{V}_j = \begin{cases} X_{2j} - \widetilde{r}(Z_j), \text{ if } Z_j \in C_Z \subsetneq S_Z; \\ \infty, \text{ otherwise.} \end{cases} \tag{3.6}$$

Notice that for $\widehat{V}_j \neq \infty$, $\left(\widehat{V}_j - V_j\right) = [r(Z_j) - \widetilde{r}(Z_j)]$, so that the former has the same asymptotic behavior as the latter. $\widehat{V}_j$ is a consistent estimate of V_j if $\widehat{V}_j \neq \infty$, because $\widetilde{r}(z)$ is consistent for $z \in C_Z$ under $E[V|Z] = 0$.

Step 2: Nonparametric Estimation of the Augmented Regression Function $m(x, v)$

With the preliminary estimates $\widehat{V}_j$'s, for $(x, v) \in C_{XV}$, $\widehat{m}(x, v)$ is defined as follows:

$$\widehat{m}(x, v) = \widehat{E}\left[Y|(X, \widehat{V}) = (x, v)\right] \equiv \widehat{q}(x, v) / \widehat{f}(x, v), \tag{3.7}$$

where
$$\widehat{q}(x, v) \equiv \frac{1}{n}\sum_{j=1}^{n} K_h(x - X_j) K_h\left(v - \widehat{V}_j\right) \mathbf{1}_{C'_{XV}}\left(X_j, \widehat{V}_j\right) Y_j,$$

$$\widehat{f}(x, v) \equiv \frac{1}{n}\sum_{j=1}^{n} K_h(x - X_j) K_h\left(v - \widehat{V}_j\right) \mathbf{1}_{C'_{XV}}\left(X_j, \widehat{V}_j\right).$$

Here $K_h(x - X_j) \equiv \frac{1}{h^d}\kappa\left(\frac{x-X_j}{h}\right)$ and $K_h\left(v - \widehat{V}_j\right) \equiv \frac{1}{h^{d_2}}\kappa\left(\frac{v-\widehat{V}_j}{h}\right)$ are kernels, and h is the bandwidth.[4] The trimming $\mathbf{1}_{C'_{XV}}\left(X_j, \widehat{V}_j\right)$ is the indicator function such that $\mathbf{1}_{C'_{XV}}\left(X_j, \widehat{V}_j\right)$ equals 1 if $\left(X_j, \widehat{V}_j\right) \in C'_{XV}$ and zero otherwise. To see how the trimming ensures that $\widehat{V}_j$ is a consistent estimate for V_j, note that $\left(X_j, \widehat{V}_j\right) \in$

[3] $\widetilde{r}(z)$ is a standard kernel estimator except that $r(z)$ may be a vector function when $d_2 \geqslant 2$. Here each component in V uses the same set of instruments, which can be relaxed to allow each component to use different sets of instruments. The convergence rate of $\widehat{V}$ is then determined by the slowest one among the convergence rates of the component in $\widehat{V}$. Also, V can contain some observed elements, which do not affect the rate of convergence of Step 1.

[4] To keep the notation compact, the kernel $K_h(\cdot)$ is distinguished by their arguments and so does $\kappa(\cdot)$. Also implicit is the dependence of bandwidths on the sample size.

$C'_{XV} \subsetneq S_{XV}$ implies that $Z_j \in C_Z \subsetneq S_Z$ so that $\widetilde{r}(Z_j)$ and $\widehat{V}_j$ are consistent. While $Z_j \in C_Z$ is sufficient for $\widehat{V}_j$ to be consistent, $\mathbf{1}_{C'_{XV}}\left(X_j, \widehat{V}_j\right)$ facilitates the asymptotic analysis and incurs no efficiency loss. The trimming is unnecessary if kernels with bounded supports are used in (3.7), because those inconsistent estimates $\widehat{V}_j$'s go to infinity by (3.6) and thus have zero weights.[5]

When the control V is observed, $m(x, v)$ can be estimated by:

$$\widetilde{m}(x, v) = \widetilde{E}[Y|(X, V) = (x, v)] \equiv \widetilde{q}(x, v) \diagup \widetilde{f}(x, v), \tag{3.8}$$

where $\widetilde{q}(x, v)$ and $\widetilde{f}(x, v)$ are defined similarly to (3.7), with $\widehat{V}_i$ being replaced by V_i and no trimming being required. By construction, both $\widehat{m}\,(x, v)$ and $\widetilde{m}(x, v)$ are consistent due to the effect of the control V under Assumption M.

Step 3: Estimation of the Structural Function $g(x)$

For $x \in C_X$, when V is unobservable, the trimmed version of (3.4) is

$$\widehat{g}(x) = n^{-1} \textstyle\sum_{i=1}^{n} \widehat{m}(x, \widehat{V}_i) \mathbf{1}^{p_x}_{C_{XV}}\left(x, \widehat{V}_i\right); \tag{3.9}$$

When V is observed, the trimmed version of (3.3) is

$$\widetilde{g}(x) = n^{-1} \textstyle\sum_{i=1}^{N} \widetilde{m}(x, V_i) \mathbf{1}^{p_x}_{C_{XV}}\left(x, V_i\right), \tag{3.10}$$

where $\mathbf{1}^{p_x}_{C_{XV}}(x, v) \equiv \mathbf{1}_{C_{XV}}(x, v) \diagup p_x$, and $p_x \equiv \Pr\left(V \in C^x_V\right)$. $(1 \diagup p_x)$ is introduced to correct the bias caused by the trimming function $\mathbf{1}_{C_{XV}}(x, v)$. Lemma A.2 shows

[5] This technique has been used by Guerre, Perrigne and Vuong (2000) in estimating the distribution of private values of bidders in first-price auctions. The pseudo private values of bidders are estimated from observed bids and defined similarly to (3.6). The empirical distribution of private values is then estimated from these pseudo values using kernels with compact supports.

that p_x can be consistently estimated either by $n^{-1}\sum_{i=1}^{n}\mathbf{1}_{C_{XV}}\left(x,\widehat{V}_i\right)$ for (3.9) or by $n^{-1}\sum_{i=1}^{n}\mathbf{1}_{C_{XV}}\left(x,V_i\right)$ for (3.10).

The trimming makes sure that the preliminary estimates are consistent. The estimates of $m(x,v)$ near the boundary of the support of (X,V) are trimmed away by the trimming functions $\mathbf{1}_{C_{XV}}\left(x,\widehat{V}_i\right)$ in (3.9), or $\mathbf{1}_{C_{XV}}\left(x,V_i\right)$ in (3.10). Especially, for (3.9), $\mathbf{1}_{C_{XV}}\left(x,\widehat{V}_i\right)$ *simultaneously* guarantees the consistency of preliminary kernel estimates, not only for $\widehat{m}(\cdot,\cdot)$ but for $\widehat{V}_i$ also. To see this, note that $\widehat{m}(x,v)$ is consistent for any $(x,v)\in C_{XV}$; and by (3.6), $\mathbf{1}_{C_{XV}}\left(x,\widehat{V}_i\right)=1$ means that $\widehat{V}_i\neq\infty$, which in turn implies $Z_i\in C_Z$ so that $\widehat{V}_i=X_i-\widetilde{r}(Z_i)$ is consistent.

3.3 Regularity Conditions

For the analysis of asymptotic properties of $\widehat{g}\,(x)$ and $\widetilde{g}\,(x)$, some regularity assumptions are imposed on key objects in kernel estimation: the kernel functions, bandwidths, and underlying data generating processes (DGPs) in each step.

Assumption K (Kernels):

(i). *Let $K(s)$ be the class of Borel measurable, bounded, real-valued functions $k(\psi)$ with compact support such that $\int k(\psi)d\psi=1$, $\int k^2(\psi)d\psi<\infty$, and $\int\psi^j k(\psi)d\psi=0$ for all $j<s$;*[6]

(ii). $k(\cdot)$ *has continuous bounded derivatives up to the second order.*

(iii). *For step 1, $\kappa\left(\frac{z-Z}{h_1}\right)=\prod_{p=1}^{d_Z}k\left(\frac{z_p-Z_p}{h_1}\right)$ with $k\in K(s_1)$; for step 2, $\kappa\left(\frac{x-X}{h}\right)=$*

[6] The method described in Bierens (1987) can be used to construct higher order kernels from univariate kernels as the base kernel, such as the Epanechnikov's kernel. The compactness of the support can be replaced by the Parzen-Rosenblatt condition requiring that the kernel $k(\psi)$ satisfies $|\psi|k(\psi)\to 0$ as $|\psi|\to\infty$.

$\prod_{p=1}^{d} k\left(\frac{x_p - X_p}{h}\right)$ *and* $\kappa\left(\frac{v-V}{h}\right) = \prod_{p=1}^{d_2} k\left(\frac{v_p - V_p}{h}\right)$ *with* $k \in K(s)$, *where both* s *and* s_1 *are strictly positive integers.*[7]

Assumption B (Bandwidths): *For uniform consistency, let the bandwidths be*

$$h = \lambda\left(\frac{log(n)}{n}\right)^{(1+\sigma)/(2s+d+d_2)} \text{ and } h_1 = \lambda_1\left(\frac{log(n)}{n}\right)^{(1+\sigma_1)/(2s_1+d_Z)};$$

For pointwise consistency and asymptotic normality, let the bandwidths be

$$h = \lambda(1/n)^{(1+\sigma)/(2s+d+d_2)} \text{ and } h_1 = \lambda_1(1/n)^{(1+\sigma_1)/(2s_1+d_Z)},$$

where λ*'s are strictly positive constants while* σ*'s can be small positive or negative constants.*

Assumption D (DGPs):

(i). *The densities* $f_{x_0}(x)$, $f_{v_0}(v)$, $f_{z_0}(z)$ *and* $f_0(x, v)$ *are all bounded away from zero within their compact supports* S_X, S_V, S_Z *and* S_{XV} *respectively; however, they equal zero at the boundary of their respective supports.*

(ii). *Within their respective supports, both* $r_0(z)$, $f_{v_0}(v)$ *and* $f_{z_0}(z)$ *are continuously differentiable with bounded derivatives up to the* s_1*-th order and all of* $g_0(x)$, $m_0(x, v)$ *and* $f_0(x, v)$ *are continuously differentiable with bounded derivatives up to the* s*-th order.* $E[|Y|^p] < \infty$ *for some* $p > 2$.

(iii). *The following holds: For* $x \in C''_X \subsetneq S_X$

$\frac{1}{nh^d}\sum_{j=1}^{n}\left|\kappa\left(\frac{x-X_j}{h}\right)\right| \xrightarrow{a.s.} f_{X_0}(x)\int|\kappa(t)|\,dt$, *and*

$\frac{1}{nh^d}\sum_{j=1}^{n}\left|\kappa\left(\frac{x-X_j}{h}\right)\right||Y_i| \xrightarrow{a.s.} E[|Y| \mid X = x]\, f_{X_0}(x)\int|\kappa(t)|\,dt;$

[7] For convenience of exposition, the product kernels are used for the multivariate conditioning variable case. The general multivariate kernels, however, can be used without physical effect on the main results of this paper. Additionally, step 1 may use other kernels than step 2, but the subscript is surpressed.

both of which are bounded almost surely within C_X; For $(x,v) \in C''_{XV} \subsetneq S_{XV}$,

$$\frac{1}{nh^{d+d_2}} \sum_{j=1}^{n} \iota^{\intercal} \left|\kappa'\left(\frac{v-V_j}{h}\right)\right| \left|\kappa\left(\frac{x-X_j}{h}\right)\right| \xrightarrow{a.s.} f_0(x,v) \cdot \int \iota^{\intercal} \left|\kappa'(\omega)\right| \left|\kappa(t)\right| dt d\omega, \text{ and}$$

$$\frac{1}{nh^{d+d_2}} \sum_{j=1}^{n} \iota^{\intercal} \left|\kappa'\left(\frac{v-V_j}{h}\right)\right| \left|\kappa\left(\frac{x-X_j}{h}\right)\right| |Y_i| \xrightarrow{a.s.} E\left[|Y| \mid x,v\right] \cdot f_0(x,v) \cdot \int \iota^{\intercal} \left|\kappa'(\omega)\right| \left|\kappa(t)\right| dt d\omega,$$

both of which are bounded almost surely within C''_{XV}. Here $\iota \equiv (1,1,\cdots,1)^{\intercal} \in \mathbb{R}^{d_2}$.

The assumptions on kernels are quite standard in nonparametric econometrics. In particular, we use higher order kernels to reduce the asymptotic biases and impose the boundedness of (partial) derivatives of kernel functions up to the second order. Compared to bandwidths, the choice of kernels has less impact on the asymptotic behavior of kernel estimators.

Bandwidths are critical parameters in kernel estimation. The assumptions on the bandwidths ensure that, as the sample size $n \to \infty$, the h's go to zero while nh^d and $nh_1^{d_z}$ approach infinity. When the σ's are zero, h's are the optimal bandwidths while $\sigma > 0$ implies undersmoothing and $\sigma < 0$ implies oversmoothing (see Stone, 1982). There is a trade-off between the convergence rates and asymptotic biases in the analysis of asymptotic normality. When the optimal bandwidth is adopted, there exists an asymptotic bias. On the other hand, the asymptotic bias can be taken away by undersmoothing, which lowers the convergence speed.[8] For uniform consistency, define the uniform convergence rates as follow

$$\begin{aligned}
\gamma &\equiv (\log(n) \diagup n)^{s \diagup (2s+d)}, \\
\gamma_1 &\equiv (\log(n) \diagup n)^{s_1 \diagup (2s_1+d_Z)}, \text{ and} \\
\gamma_2 &\equiv (\log(n) \diagup n)^{s \diagup (2s+d+d_2)}.
\end{aligned}$$

[8] When the asymptotic bias is difficult to estimate, it may be desirable to undersmooth a little bit by setting σ's to be small positive numbers.

For pointwise consistency and asymptotic normality, redefine them as

$$
\begin{aligned}
\gamma &\equiv (1/n)^{s/(2s+d)}, \\
\gamma_1 &\equiv (1/n)^{s_1/(2s_1+d_Z)}, \text{ and} \\
\gamma_2 &\equiv (1/n)^{s/(2s+d+d_2)}.
\end{aligned}
$$

Assumption D places restrictions on the smoothness of the underlying DGPs to facilitate the derivation of uniform consistency and asymptotic normality. As shown in Härdle (1994) among others, the optimal convergence rates of nonparametric estimators are determined by their relative smoothness conditions, which are (s_1/d_Z) for step 1, $(s/(d+d_2))$ for step 2, and (s/d) for step 3. The last part of Assumption D is the boundedness restriction used to get sharper approximation results in Proposition 3.1 and Theorem 3.1.

3.4 Uniform Consistency

Consistency and asymptotic normality are main properties of estimators. This subsection studies uniform consistency while pointwise consistency will be established along with asymptotic normality in Section 3.5.

As standard results of nonparametric econometrics, within inner compact subsets of their respective supports, the optimal rates of uniform convergence for $\overline{g}(x)$, $\widetilde{r}(z)$ and $\widetilde{m}(x,v)$ are $(1/\gamma)$, $(1/\gamma_1)$ and $(1/\gamma_2)$ respectively. Here $\overline{g}(x)$ is the kernel estimator of the conditional mean of Y given $X = x$. Due to the extra dimension from the control V, the convergence rate of $\widetilde{m}(x,v)$ slows down to $(1/\gamma_2)$, which is improved back to $(1/\gamma)$ for $\widetilde{g}(x)$ by averaging $\widetilde{m}(x,V_j)$ over V_j's, see Stone (1982). Thus $\widetilde{g}(x)$ corrects the endogeneity and, at the same time, maintains the same convergence rate as $\overline{g}\,(x)$.

Now the question is how the preliminary estimator of the control V affects the uniform convergence rate of $\widehat{g}\,(x)$ when V is unobserved but can be estimated. Since $\widehat{g}\,(x)$ is the partial mean of $\widehat{m}(x,v)$, I first establish the uniform consistency with the rates of convergence of $\widehat{m}(x,v)$ in Proposition 3.1.

Proposition 3.1. *Let Assumptions M, K, B and D hold, let* $\sigma=\sigma_1=0$*, and suppose* $\frac{s_1}{(2s_1+d_Z)}-\frac{1+d_2}{(2s+d+d_2)}>0$*, then*

(*a*) $\sup\limits_{C_{XV}}|\widehat{f}(x,v)-f_0(x,v)|=O\left(\gamma_2+\gamma_1/h\right)$*, a.s.;*

(*b*) $\sup\limits_{C_{XV}}|\widehat{m}(x,v)-m_0(x,v)|=O\left(\gamma_2+\gamma_1/h\right)$*, a.s..*

Proof: See Appendix A.

The most important feature of Proposition 3.1 is that $\widehat{m}(x,v)$ can achieve the same optimal uniform rate $(1/\gamma_2)$ as $\widetilde{m}(x,v)$ when the preliminary estimator $\widehat{V}$ converges fast enough compared to $\widetilde{m}(x,v)$ in the sense that $((\gamma_1/h)/\gamma_2)=o\,(1)$, *i.e.*, $\frac{1+s}{(2s+d+d_2)}\leqslant\frac{s_1}{(2s_1+d_Z)}$. As a result, the effect of $\widehat{V}$ on $\widehat{m}(x,v)$ is negligible asymptotically. On the other hand, when $\widehat{V}$ does not converge so fast, the uniform convergence rate of $\widehat{m}(x,v)$ will be dominated by $\widehat{V}$ and only the suboptimal rate (h/γ_1) is possible. This is even slower than $(1/\gamma_1)$ and depends on step-2 bandwidth h. A larger h implies faster convergence of $\widehat{m}(x,v)$.[9]

Proposition 3.1 is of interest beyond the nonparametric control function approach to endogeneity as the control function assumption (Assumption M.(ii)) is not used except in the third step. Thus it applies to more general 2-step kernel estimators with preliminary kernel estimates, including kernel estimators of densities

[9] The intuition is that a larger h effectively includes more observations of $\left(X_j,\widehat{V}_j\right)$ to estimate $\widehat{m}(x,v)$. Since the kernel estimators are local averages, more observations help cancel out the noise caused by $\widehat{V}$.

as indicated in Proposition 3.1.(a). In particular, Proposition 3.1 is an improvement upon Ahn (1995), where the lower bound of uniform convergence rate for his 2-step kernel estimator is $\left(\gamma_1 / h^{d+d_2+1}\right)$. The optimal rate of uniform convergence is impossible and the reason is that the approximation (i.e., Lemma A.3 in Ahn (1995)) is too conservative. I extend a technique in Guerre, Perrigne and Vuong (2000) to the case of general kernel regressions to achieve better rates given in Proposition 3.1.

The uniform convergence rate of $\widehat{g}(x)$ exhibits a structure similar to that of $\widehat{m}(x,v)$, as indicated in Theorem 3.1 below.

Theorem 3.1. *Let Assumptions M, K, B,and D hold, let* $\sigma_1 = 0$, *also let* $\sigma < 0$ *such that* $h = \lambda(\frac{log(n)}{n})^{1/(2s+d)}$, *and suppose* $\frac{s_1}{(2s_1+d_Z)} - \frac{1+d_2}{(2s+d+d_2)} > 0$, *then*

$$\sup_{C_X} |\widehat{g}(x) - g_0(x)| = O\left(\gamma + \gamma_1 / h\right) \text{ a.s. for } x \in C_X \subsetneq S_X.$$

Proof: It is a standard result that $\sup_{C_X} |\widetilde{g}(x) - g_0(x)| = O(\gamma)$ a.s. Proposition A.1 in Appendix A shows that

$$\sup_{C_X} |\widehat{g}(x) - \widetilde{g}(x)| = O\left(\gamma + \gamma_1 / h\right) \text{ a.s.}$$

By the triangle inequality,

$$\begin{aligned}\sup_{C_X} |\widehat{g}(x) - g_0(x)| &\leqslant \sup_{C_X} |\widehat{g}(x) - \widetilde{g}(x)| + \sup_{C_X} |\widetilde{g}(x) - g_0(x)| \\ &= O\left(\gamma + \gamma_1 / h\right). \qquad \square\end{aligned}$$

Theorem 3.1 shows that $\widehat{g}(x)$ is able to achieve the same optimal rate of uniform convergence as $\overline{g}(x)$ and $\widetilde{g}(x)$ if the preliminary estimator $\widehat{V}$ converges fast enough in the sense that $O\left(\left(\gamma_1 / h\right) / \gamma\right) = o(1)$, i.e, $\frac{s+1}{(2s+d)} \leqslant \frac{s_1}{(2s_1+d_Z)}$. Thus, the unobservability of the control variable V does not affect the uniform consistency of $\widehat{g}(x)$ if V can be estimated fast enough. Notice that the step-2 bandwidth $h = \lambda(\frac{log(n)}{n})^{1/(2s+d)}$ undersmoothes $\widehat{m}(x,v)$ to get the optimal uniform convergence rates for $\widehat{g}(x)$.

This result complements and extends Newey, Powell, and Vella (1999) by allowing for the optimal and suboptimal rates of uniform convergence under less restrictive conditions. To see this, denote the relative smoothness $\alpha_1 = s_1/d_Z$ for step-1 estimation and $\alpha_2 = s/d$ for step-2 estimation (or step 3 in this paper as kernel estimation does not have the same additive feature as series estimation). The uniform convergence rates derived in Newey, Powell, and Vella (1999) are $O_p\left(h^d/\gamma + h^d/\gamma_1\right)$ for power series under the relative smoothness condition $\alpha_2 \geqslant \frac{3+5\alpha_1}{2\alpha_1}$, and $O_p\left(h^{d/2}/\gamma + h^{d/2}/\gamma_1\right)$ for splines under $\alpha_2 \geqslant \frac{3+3\alpha_1}{2(\alpha_1-1)}$. The optimal rate $(1/\gamma)$ cannot be achieved for those estimators due to the terms h^d (for power series) or $h^{d/2}$ (for splines).

In contrast, $\widehat{g}\,(x)$ can achieve the optimal rate of uniform convergence under less restrictive conditions. Under the relative smoothness condition $\alpha_2 \geqslant \frac{1+d_2/d}{2\alpha_1+1+d_2/d}$, the optimal rate $(1/\gamma)$ can be achieved if $\frac{\alpha_2+1/d}{2\alpha_2+1} \leqslant \frac{\alpha_1}{2\alpha_1+1}$; otherwise the suboptimal rate (h/γ_1) can be achieved.[10] For instance, when $d=3$, $d_2=1$, $s=2$, $d_z=1$, and $s_1=3$, $\widehat{g}(x)$ achieve the optimal rate of uniform convergence $(\log(n)/n)^{3/7}$ almost surely, which is impossible for the series estimators proposed in Newey, Powell, and Vella (1999).

Theorem 3.1 also extends the literature of generalized additive models by al-

[10] The condition for the uniform consistency in this paper is less restrictive: the relative smoothness condition for $\widehat{g}\,(x)$ is $\alpha_2 \geqslant \frac{1+d_2/d}{2\alpha_1+1+d_2/d}$, which is less restrictive than either $\alpha_2 \geqslant \frac{3+5\alpha_1}{2\alpha_1}$ or $\alpha_2 \geqslant \frac{3+3\alpha_1}{2(\alpha_1-1)}$ in Newey, Powell and Vella (1999). To see this, notice that $\frac{1+d_2/d}{2\alpha_1+1+d_2/d} < \frac{3+5\alpha_1}{2\alpha_1}$ and $\frac{1+d_2/d}{2\alpha_1+1+d_2/d} < \frac{3+3\alpha_1}{2(\alpha_1-1)}$ as $d_2/d \leqslant 1$. Thus, for the same α_2, the uniform convergence rate of $\widehat{g}\,(x)$ given in Theorem 3.1 applies for more values of α_1.

lowing regressors to be estimated preliminarily, and maintaining the optimal rate of uniform convergence.

3.5 Pointwise Consistency and Asymptotic Normality

Asymptotic normality is also an important property of estimators, upon which the statistical inference of confidence intervals can be made. In Proposition 3.2, an alternative way based on U-statistic is proposed to establish the asymptotic normality of $\widetilde{g}(x)$,[11] which also facilitates the derivation of the asymptotic normality of $\widehat{g}(x)$.

Proposition 3.2. *Under Assumptions M, K, B, and D, let* $a_i' \equiv \mathbf{1}^{p_x}_{C'_{XV}}(x, V_i)$, $\widetilde{g}(x) - g_0(x)$ *can be expressed as*

$$\widetilde{g}(x) - g_0(x) = \frac{1}{n}\sum_{i=1}^{n} K_h(x - X_i)(Y_i - m(x, V_i)) \frac{a_i' f_{v0}(V_i)}{f_0(x, V_i)} + O_p(h^s). \tag{3.11}$$

Thus $\widetilde{g}(x) - g_0(x) = O_p\left(h^s + \frac{1}{\sqrt{nh^d}}\right)$, *and*

(i) If $\lim \sqrt{nh^d} h^s = c \geqslant 0$, *then* $\sqrt{nh^d}(\widetilde{g}(x) - g_0(x)) \xrightarrow{d} N(cB_g(x), V_g(x))$,

where the asymptotic bias $B_g(x)$ *is given by* $(B.2)$ *in Appendix B,*[12] *and the asymptotic variance is*

$$V_g(x) = \frac{1}{nh^d} \int Var(Y|x, v) \frac{(a_i')^2 f_{v0}^2(v)}{f_0(x, v)} dv \cdot \int \kappa^2(t)\, dt. \tag{3.12}$$

[11] For the asymptotic normality of the estimator $\widehat{m}(x, v)$, see Ahn (1995) and Rilstone (1996).

[12] The asymptotic bias $B_g(x)$ includes additional biases introduced by averaging $\widetilde{m}(x, V_i)$'s over V_i (the second term of the RHS of (3.11)), as well as the bias from the first term of the RHS of (3.11).

(ii) If $\lim \sqrt{nh^d}h^s = \infty \geqslant 0$, *then* $\frac{1}{h^s}\left(\widetilde{g}(x) - g_0(x)\right) \xrightarrow{p} B_g(x)$.

Proof: See Appendix B.

Proposition 3.2 shows that $\widetilde{g}(x)$ can achieve the optimal rate of pointwise convergence when $O_p(h^s) = O_p\left(1/\sqrt{nh^d}\right)$, that is, the optimal rate of $\widetilde{g}(x)$ is $n^{s/(2s+d)}$ when the bandwidth h is of exact order of $n^{-1/(2s+d)}$. Note that this bandwidth is the optimal bandwidth for the simple kernel estimator $\overline{g}(x)$. It also undersmooths $\widetilde{m}(x,v)$ (i.e., $\sigma < 0$) so that $\widetilde{m}(x,v)$ is asymptotically unbiased. The consistent estimator for the asymptotic variance $V_g(x)$ can be obtained by plugging consistent estimators of the components of $V_g(x)$ into (3.12), see Newey (1994) for instance.

Although kernel estimators of partial means have been studied in the literature of generalized additive models, by the way of U-statistics, we can express $\widetilde{g}(x) - g_0(x)$ as a sample average (the first term of the RHS of (3.11)) where the asymptotic variance naturally arises as in (3.12). Moreover, the effect of the preliminary estimator $\widehat{V}$ can be analyzed relatively easily by extending the second-order U-statistic for $\widetilde{g}(x)$ to a third-order U-statistic for $\widehat{g}(x)$. The asymptotic properties of $\widetilde{g}(x)$ as derived in Proposition 3.2 form the basis for the asymptotic normality of $\widehat{g}(x)$ as indicated in Theorem 3.2 below.

Theorem 3.2. *Under Assumptions M, K, B, and D,*

$$\widehat{g}(x) - g_0(x) = \left(\widehat{g}(x) - \widetilde{g}(x)\right) + \left(\widetilde{g}(x) - g_0(x)\right) = O_p(\gamma_1 + \gamma).$$

Thus, if $O(\gamma_1/\gamma) = o(1)$ (i.e.,$\frac{s}{d} < \frac{s_1}{d_Z}$), *then* $\widehat{g}(x) - g_0(x) = O_p(\gamma)$ *and*

$$\sqrt{nh^d}\left(\widehat{g}(x) - g_0(x)\right) \xrightarrow{d} N(cB_g(x), V_g(x)),$$

where the asymptotic bias and asymptotic variance are given in Proposition 3.2.

Proof: The asymptotic properties of $(\widetilde{g}(x) - g(x))$ is derived in Proposition 3.2 and it remains to study that of $\widehat{g}(x) - \widetilde{g}(x)$. In Appendix B, Proposition B.1 shows that

$$(\widehat{g}(x) - \widetilde{g}(x)) = O_p(\gamma_1), \text{ so that}$$

$$\widehat{g}(x) - g_0(x) = (\widehat{g}(x) - \widetilde{g}(x)) + (\widetilde{g}(x) - g_0(x)) = O_p(\gamma_1) + O_p(\gamma).$$

If $O(\gamma_1 / \gamma) = o(1)$, $O\left(\sqrt{nh^d}\right) O_p(\gamma_1) = o_p(1)$ so that

$$\begin{aligned} \sqrt{nh^d}\,(\widehat{g}(x) - g_0(x)) &= \sqrt{nh^d}\,(\widetilde{g}(x) - g_0(x)) + o_p(1) \\ &\xrightarrow{d} N(cB_g(x), V_g(x)). \qquad \square \end{aligned}$$

Two points are worthwhile to mention. First, despite the fact that there are endogenous variables in X and that the control V has to be estimated preliminarily by $\widehat{V}$, $\widehat{g}(x)$ may still achieve the optimal rate of pointwise convergence if $\widehat{V}$ converges faster than $\overline{g}(x)$. Second, also in this case, the asymptotic variance of $\widehat{g}(x)$ is unaffected by $\widehat{V}$. As long as the unobserved control V can be estimated fast enough, the estimator $\widehat{g}(x)$ behaves as if V were actually observed. In contrast, the asymptotic variance of the series estimators in Newey, Powell, and Vella (1999) is always affected by the preliminary estimators.

Chapter 4

Semiparametric Control Function Estimation

4.1 Partially Linear Models

Partially linear models are capable of capturing nonlinear relationships while mitigating the curse of dimension. As the result, partially linear models have been extensively used in empirical studies. In a pioneer empirical application of partially linear models, Engle, Granger, Rice and Weiss (1986) study the relationship between electricity sales and temperature, which is typically nonlinear as both heating in low temperatures and air-conditioning in high temperatures increases electricity consumption. More examples include household gasoline consumption in the United States (Schmalensee and Stoker, 1999), Engle curves (Blundell, Duncan and Pendakur, 1998), the production frontier of US banking industry (Adams, Berger and Sickles, 1999), just to name a few. For an extensive treatment of the theory and applications of partially linear models, see Hardle, Liang and Gao (2000).

Partially linear models can also be motivated as a way to control for endogeneity.[1] In sample selection models, the endogeneity is caused by the selection bias, which can be corrected by a function representing the selection process. As mentioned in Chapter 1, the semiparametric Type-2 Tobit model arise as a partially linear model:

$$Y_{2i} = X_{2i}'\beta_2 + c\left(X_{1i}'\beta_1\right) + \epsilon_{2i}, \text{ where}$$

$$c\left(X_{1i}'\beta_1\right) \equiv E\left(u_{2i}|X_i, Y_{1i} = 1\right) = E\left(u_{2i}|X_i, u_{1i} > -X_{1i}'\beta_1\right).$$

Notice that no functional form is specified for $g\left(\cdot\right)$, which makes the model semiparametric. Also notice that, in order to estimate the parameter of interest β_2, β_1 has to be estimated first so that the conditioning variable $X_{1i}'\beta_1$ becomes a constructed one. For the $\sqrt{n}$-consistency of $\left(\widetilde{\beta}_1, \widetilde{\beta}_2\right)$, see Powell (1987), Ichimura and Lee (1991), Ai (1997) and Li and Wooldridge (2002). $\widetilde{\beta}_1$ (and hence $X_{1i}'\widetilde{\beta}_1$) converges at the parametric rate.

In this thesis, the type of endogeneity, the estimator and its asymptotic properties are all different. For the Cobb-Douglas production function $g(X;\beta) = X'\beta$, the (capital and labor) coefficients β cannot be consistently estimated by OLS due to the endogeneity of input X. A partial linear model arises naturally where the parametric part represents the production function and the nonparametric part is the control function to "smooth out" unobserved shocks. Given that there exists a control V satisfying the control function assumption

$$E\left(a_i|X_i, V_i\right) = E\left(a_i|V_i\right) \equiv c\left(V_i\right), \tag{4.1}$$

[1] Other motivations include the presence of heteroskedasticity of unknown form and rational expecation in macroeconomic models; see Pagan and Ullah (1999).

the augmented regression becomes

$$Y_i = X_i'\beta + c(V_i) + \varepsilon_i, \text{ where } \varepsilon_i \equiv a_i + \epsilon_i - c(V_i)$$

By construction, the regressors (X_i, V_i) in the augmented regression are exogenous: $E(\varepsilon_i | X_i, V_i) = 0$.[2]

Were the observable control available, β can be consistently estimated by Robinson's (1988) method:

$$\widetilde{\beta}_R \equiv \left[\sum_i \left(X_i - \widetilde{X}_i\right)\left(X_i - \widetilde{X}_i\right)'\right]^{-1} \sum_i \left(X_i - \widetilde{X}_i\right)\left(Y_i - \widetilde{Y}_i\right) 1_i,$$

where $\widetilde{W}_i \equiv \frac{1}{n}\sum_j W_i K_h(V_i - V_j) \big/ \widetilde{f}_i$ for $W_i = X_i$ or $W_i = Y_i$ is the kernel regressor and $\widetilde{f}_i \equiv \frac{1}{n}\sum_j K_h(V_i - V_j)$ is the kernel density estimator. The indicator function $1_i \equiv 1\left(\widetilde{f}_i \geqslant b\right)$ is a trimming function to handle the random denominator problem in estimating $\widetilde{X}_i$ and $\widetilde{Y}_i$. The trimming complicates the asymptotic analysis and, besides the bandwidth h, the trimming parameter b needs to be specified too. Li (1996) proposes a density-weighted version, where the trimming is not needed:

$$\widetilde{\beta}_L \equiv \left[\sum_i \left(X_i - \widetilde{X}_i\right)\widetilde{f}_i^2\left(X_i - \widetilde{X}_i\right)'\right]^{-1} \sum_i \left(X_i - \widetilde{X}_i\right)\widetilde{f}_i\left(Y_i - \widetilde{Y}_i\right)\widetilde{f}_i.$$

$\sqrt{n}$-consistency is established both for $\widetilde{\beta}_R$ and for $\widetilde{\beta}_L$, so that, under some regularity conditions, they still converge at the parametric rate in spite of the presence of preliminary kernel estimators.[3]

[2] The subscript t is surpressed for notation simplicity.

[3] For the asymptotic analysis, see also Speckman (1988), Stock (1989), and Andrew (1994). Both Robinson's (1988) and Li (1996) adopt kernels to estimate $\widetilde{x}$ and $\widetilde{y}$; for partially linear models using series methods, see Donald and Newey (1994).

As mentioned in Chapter 2, it is difficult to find such observables that satisfy the control function assumption and maintain the identification of production functions. The controls are constructed from the instruments, such as the lagged levels of inputs.

$$V_i = X_i - r\left(Z_i\right), \text{ where } r\left(Z_i\right) \equiv E\left(X_i|Z_i\right). \tag{4.2}$$

Given $E\left(U|Z,V\right) = E\left(U|V\right)$ and $r\left(Z\right)$ is strictly monotone in Z, the control function condition (4.1) holds for V as constructed by (4.2). Since the constructed control $\widehat{V}$ is estimated nonparametrically and converges slower than the parametric rate, the asymptotic analysis of $\widehat{\beta}$ with $c\left(\widehat{V}\right)$ will be different from $\widetilde{\beta}_R$ and $\widetilde{\beta}_L$.[4]

For the partially linear model with constructed variables in the nonparametric part, I propose an kernel-based estimator of β:

$$\widehat{\beta} \equiv \left[\sum_i \left(X_i - \widehat{X}_i\right) \widehat{f}_i^{\,2} \left(X_i - \widehat{X}_i\right)'\right]^{-1} \sum_i \left(X_i - \widehat{X}_i\right) \widehat{f}_i \left(Y_i - \widehat{Y}_i\right) \widehat{f}_i, \tag{4.3}$$

where $\widehat{X}_i$, $\widehat{Y}_i$ and $\widehat{f}_i$ are to be defined below. It can be viewed as a density-weighted and preliminarily estimated version of $\widetilde{\beta}_R$ (Robinson, 1988) or a preliminarily estimated version of $\widetilde{\beta}_L$ (Li, 1996).

4.2 Semiparametric Estimation Procedures

Now I describe the estimation procedure of the semiparametric control function (SPCF) estimator.

Step 1: Construct the Control $\widehat{V}$

[4] Stengos and Yan (2001) also consider partially linear models with contructed variables, where the contructed variables are not in the nonparametric part, but in the parametric part.

Similar to Section 3, the control $\widehat{V}_j$ can be estimated by

$$\widehat{V}_j = \begin{cases} X_j - \widetilde{r}(Z_j), \text{ if } Z_j \in C_Z \subsetneq S_Z; \\ \infty, \text{ otherwise}; \end{cases} \tag{4.4}$$

where C_Z is an inner subset of the support S_Z of Z and $\widetilde{r}(Z)$ is the kernel estimator of $r(z)$:

$$\widetilde{r}(z) = \widehat{E}[X|Z = z] = \frac{1}{n}\sum_{l=1}^{n} X_l K_h(z - Z_l) \diagup \widetilde{f}_z(z),$$

Again, $\widetilde{f}_z(z) \equiv \frac{1}{n}\sum_{l=1}^{n} K_h(z - Z_l)$ is the density estimator; $K_h(z - Z_l)$ and h_1 are the kernel and bandwidth respectively. Notice that for $\widehat{V}_j \neq \infty$, $\widehat{V}_j$ is a consistent estimator of V_j and $\left(\widehat{V}_j - V_j\right) = [r(Z_j) - \widetilde{r}(Z_j)]$.

Step 2: Nonparametric Estimation of $\widehat{X}$ *and* $\widehat{Y}$

With the preliminary estimates $\widehat{V}_j$'s, $\widehat{W}_i \equiv \widehat{E}\left[W_i|\widehat{V}_i\right]$ ($W = X$ or $W = Y$) are defined as follows:

$$\widehat{W}_i \equiv \frac{1}{n}\sum_{j\neq i} W_i K_h\left(\widehat{V}_i - \widehat{V}_j\right) \diagup \widehat{f}_i,$$

where $\widehat{f}_i \equiv \widehat{f}\left(\widehat{V}_i\right) = \frac{1}{n}\sum_{j\neq i} K_h\left(\widehat{V}_i - \widehat{V}_j\right)$ is the kernel density estimator, $K_h\left(v - \widehat{V}_j\right) \equiv \frac{1}{h^d}\kappa\left(\frac{v-\widehat{V}_j}{h}\right)$ is the kernel, and h is the bandwidth. To simplify the asymptotic analysis, kernels with bounded supports are used so that the trimming is unnecessary because those inconsistent estimates $\widehat{V}_j$'s go to infinity and have zero weights. Also, note that $\widehat{W}_i$ is a leave-one-out kernel estimator, which also facilitate the asymptotic analysis.

Step 3: Estimation of β

Estimate β as an OLS estimator for the regression

$$(Y_i - E(Y_i|V_i)) f_i = [X_i - E(X_i|V_i)]' \beta f_i + u_i f_i,$$

where $\widehat{X}_i$ is plugged in for $E(X_i|V_i)$, $\widehat{Y}_i$ for $E(Y_i|V_i)$ and $\widehat{f}_i$ for f_i. The formula of $\widehat{\beta}$ is (4.3). Being density-weighted, $\widehat{\beta}$ is free from the problem of random denominators for there is no denominator in $\widehat{W_i}\widehat{f}_i \equiv \frac{1}{n}\sum_{j\neq i} W_i K_h(V_i - V_j)$. Furthermore, given that the density $f(x,v)$ is zero near the boundary of the support of (X,V), inconsistent estimates of $\left(\widehat{X},\widehat{Y}\right)$ due to the boundary effect are offset by the density weighting. Therefore, we obtain the consistent estimator $\widehat{\beta}$ of β without trimming.

As we can see, the estimation procedure of $\widehat{\beta}$ is straightforward and no iterative algorithm is required. Since no trimming being involved, the only parameters we need to decide are the bandwidths (h, h_1). Usually, the rule of thumb is used to choose (h, h_1) in practice. There are, however, some asymptotic restrictions imposed on (h, h_1) in order to achieve the $\sqrt{n}$-consistency of $\widehat{\beta}$, as indicated in the conditions for Theorem 4.1.

4.3 Regularity Conditions

To establish the $\sqrt{n}$-consistency of $\widehat{\beta}$, some regularity assumptions are imposed on key objects in kernel estimation: the kernel functions, bandwidths, and underlying data generating processes.

Assumption SP.M (Model)

(i). $\{Y_i, X_i, Z_i\}_{i=1}^n$ *is an i.i.d. sample where* $Y \in \mathbb{R}$, $X, V \in \mathbb{R}^d$, $Z \in \mathbb{R}^{d_Z}$, *and* $d, d_Z \geqslant 1$

(ii). *$Y = g(X;\beta) + U = X'\beta + U$, where $E(U|X) \neq 0$ and $E(U) = 0$;*

(iii). *$X = r(Z) + V$, where $E[V|Z] = 0$ and $E[U|X,V] = E[U|r(Z)+V,V] = E[U|V]$; $r(z)$ is continuous and strictly monotone in z and the rank of the Jacobian matrix of $r(z)$ equals d.*

Assumption SP.D (DGPs)

(i). *The densities $f_{x_0}(x)$, $f_{v_0}(v)$, and $f_{z_0}(z)$ are all bounded away from zero within their compact supports S_X, S_V, and S_Z respectively; however, they equal zero at the boundary of their respective supports.*

(ii). *Within their respective supports, both $r_0(z)$, $f_{v_0}(v)$ and $f_{z_0}(z)$ are continuously differentiable with bounded derivatives up to the s_1-th order and all of $E[X_i|V_i]$, $E[Y_i|V_i]$ and $f_0(x,v)$ are continuously differentiable with bounded derivatives up to the s-th order. $E[|Y|^p] < \infty$ for some $p > 2$.*

(iii). *$E(\varepsilon^2|x,v) = \sigma^2(x,v)$ is continuous in (x,v); both ε and X have finite fourth moments.*

Assumption SP.K (Kernels)

(i). *Let $K(s)$ be the class of Borel measurable, bounded, real-valued functions $k(\psi)$ with compact support such that $\int k(\psi)d\psi = 1$, $\int k^2(\psi)d\psi < \infty$, and $\int \psi^j k(\psi)d\psi = 0$ for all $j < s$;*

(ii). *Moreover, $k(\cdot)$ has continuous bounded derivatives up to the second order.*

(iii). *For step 1, $\kappa\left(\frac{z-Z}{h_1}\right) = \prod_{p=1}^{d_Z} k\left(\frac{z_p-Z_p}{h_1}\right)$ with $k \in K(s_1)$; for step 2, $\kappa\left(\frac{v-V}{h}\right) = \prod_{p=1}^{d} k\left(\frac{v_p-V_p}{h}\right)$ with $k \in K(s)$, where both s and s_1 are strictly positive integers.*

Assumption SP.B (Bandwidths):

As $n \to \infty$, $nh^{2d} \to \infty$, $nh^{4s} \to 0$, and $nh_1^{4s_1} \to 0$.

Comparable to the assumptions made by Robinson (1988) and Li (1996), Assumptions SP made above are quite standard in nonparametric econometrics. The main departure is the restriction on the bandwidth h_1 for estimating the control $\widehat{V}$, where $nh_1^{4s_1} \to 0$ is imposed. When $nh^{2d} \to \infty$ is not binding, $nh_1^{4s_1} \to 0$ is symmetric to $nh^{4s} \to 0$ in some sense.

4.4 $\sqrt{n}$-Consistency

As we can see from the estimation procedure, $\widehat{\beta}$ depends on $\left(\widehat{X}, \widehat{Y}\right)$, which in turn depend on $\widehat{V}$. So we need to take into account the fact that the conditioning variables $\widehat{V}$ are preliminary kernel estimators. Compared to $\widetilde{\beta}_R$ and $\widetilde{\beta}_L$, the asymptotic analysis of $\widehat{\beta}$ is further complicated by $\widehat{V}$. Nevertheless, the $\sqrt{n}$-consistency of $\widehat{\beta}$ is established in Theorem 4.1.

Theorem 4.1. *Under Assumptions SP.M, SP.D, SP.K, and SP.B,*

$$\sqrt{n}\left(\widehat{\beta} - \beta_0\right) \xrightarrow{d} N\left(0, \Phi_f^{-1}\Psi_f\Phi_f^{-1}\right),$$

where the asymptotic variance is determined by

$$\begin{aligned}
\Phi_f &\equiv E\left[\left(X_i - E\left(X_i|V_i\right)\right)\left(X_i - E\left(X_i|V_i\right)\right)' f_i^2\right] \text{ and} \\
\Psi_f &\equiv E\left[\sigma^2\left(X_i, V_i\right)\left(X_i - E\left(X_i|V_i\right)\right)\left(X_i - E\left(X_i|V_i\right)\right)' f_i^4\right].
\end{aligned}$$

Proof: Note that $Y_i - \widehat{Y}_i = \left(X_i - \widehat{X}_i\right)' \beta + (g_i - \widehat{g}_i + \varepsilon_i - \widehat{\varepsilon}_i)$ so that we have

$$\begin{aligned}
\widehat{\beta} &= S^{-1}_{\left(X-\widehat{X}\right)\widehat{f}} S_{\left(X-\widehat{X}\right)\widehat{f},\left(Y-\widehat{Y}\right)\widehat{f}} \\
&= S^{-1}_{\left(X-\widehat{X}\right)\widehat{f}} S_{\left(X-\widehat{X}\right)\widehat{f},\left(X-\widehat{X}\right)'\beta+(g-\widehat{g}+\varepsilon-\widehat{\varepsilon})\widehat{f}} \\
&= \beta_0 + S^{-1}_{\left(X-\widehat{X}\right)\widehat{f}} S_{\left(X-\widehat{X}\right)\widehat{f},(g-\widehat{g}+\varepsilon-\widehat{\varepsilon})\widehat{f}}.
\end{aligned}$$

With normalization by $\sqrt{n}$, we have

$$\begin{aligned}
\sqrt{n}\left(\widehat{\beta} - \beta_0\right) &= S^{-1}_{\left(X-\widehat{X}\right)\widehat{f}} \sqrt{n} S_{\left(X-\widehat{X}\right)\widehat{f},(g-\widehat{g}+\varepsilon-\widehat{\varepsilon})\widehat{f}} \\
&= S^{-1}_{\left(X-\widehat{X}\right)\widehat{f}} \sqrt{n}\left(S_{\left(X-\widehat{X}\right)\widehat{f},(g-\widehat{g})\widehat{f}} + S_{\left(X-\widehat{X}\right)\widehat{f},\varepsilon\widehat{f}} - S_{\left(X-\widehat{X}\right)\widehat{f},\widehat{\varepsilon}\widehat{f}}\right).
\end{aligned}$$

Respectively, in Appendix C, Propositions C.1 and C.2 establish that

$$\begin{aligned}
S_{\left(X-\widehat{X}\right)\widehat{f}} &\xrightarrow{p} \Phi_f, \text{ and} \\
\sqrt{n} S_{\left(X-\widehat{X}\right)\widehat{f},\varepsilon\widehat{f}} &= \sqrt{n} S_{\eta f,\varepsilon f} + o_p(1) \xrightarrow{d} N(0, \Psi_f).
\end{aligned}$$

The remaining two terms, $S_{\left(X-\widehat{X}\right)\widehat{f},(g-\widehat{g})\widehat{f}}$ and $S_{\left(X-\widehat{X}\right)\widehat{f},\widehat{\varepsilon}\widehat{f}}$, are asymptotically negligible for both are $o_p\left(n^{-1/2}\right)$ as indicated by Proposition C.3 in Appendix C. Therefore, we have

$$\begin{aligned}
\sqrt{n}\left(\widehat{\beta} - \beta_0\right) &= (\Phi_f + o_p(1))^{-1}\left[o_p(1) + \left(\sqrt{n} S_{\eta f,\varepsilon f} + o_p(1)\right) + o_p(1)\right] \\
\xrightarrow{d} \Phi_f^{-1} N(0, \Psi_f) &= N\left(0, \Phi_f^{-1}\Psi_f\Phi_f^{-1}\right). \qquad \square
\end{aligned}$$

Theorem 4.1 is an analog to the Theorem in Robinson (1988) or Theorems 1 and 2 in Li (1996). It shows that the effect of the preliminary kernel estimator $\widehat{V}$ is asymptotically negligible as long as $\widehat{V}$ converges sufficiently fast. The effect of

$\widehat{V}$ is mute in the proof of Theorem 4.1, but apparent in Lemma C.1 in Appendix C, which shows that

$$E\left[\left(c\left(V_i\right)-c\left(V_1\right)\right)K_h\left(\widehat{V}_i-\widehat{V}_1\right)|V_1\right]=O\left(h^s+h_1^{s_1}\right). \tag{4.5}$$

To make the comparison, note that if V were observed, then

$$E\left[\left(c\left(V_i\right)-c\left(V_1\right)\right)K_h\left(V_i-V_1\right)|V_1\right]=O\left(h^s\right).$$[5]

We see that $h_1^{s_1}$ in (4.5) is due to the fact that the control V is unobservable and has to be preliminarily estimated. (4.5) also leads to the condition imposed for h_1 in Assumption SP.B for the $\sqrt{n}$-consistency of $\widehat{\beta}$.

The consistent estimator of the asymptotic variance $\Phi_f^{-1}\Psi_f\Phi_f^{-1}$ can be obtained by plugging in consistent estimators of Φ_f and Ψ_f. The consistent estimators for Φ_f and Ψ_f respectively are

$$\begin{aligned}\widehat{\Phi}_f &\equiv \frac{1}{n}\sum\nolimits_i\left(X_i-\widehat{X}_i\right)\widehat{f}_i^2\left(X_i-\widehat{X}_i\right)' \text{ and}\\ \widehat{\Psi}_f &\equiv \frac{1}{n}\sum\nolimits_i\left(X_i-\widehat{X}_i\right)\widehat{f}_i\left(\widehat{\varepsilon_i f_i}\right)^2\widehat{f}_i\left(X_i-\widehat{X}_i\right)',\end{aligned}$$

where $\widehat{\varepsilon_i f_i}\equiv\left(Y_i-\widehat{Y}_i\right)\widehat{f}_i-\left(X_i-\widehat{X}_i\right)'\widehat{\beta}\widehat{f}_i$ is a consistent estimator for the density-weighted error $\varepsilon_i f_i$.

[5] See Lemma 5 in Robinson (1988) or Lemma 1 in Li (1996).

Chapter 5

Monte Carlo Experiments

5.1 Experiment Design

To illustrate the estimation procedure and to check the finite sample performance of $\widetilde{g}(x)$ and $\widehat{g}(x)$ proposed in Chapter 3, I conduct a set of Monte Carlo simulations. Set the sample size $n = 1000$ and the number of replications $R = 100$. To show and compare the true and estimated functions graphically, we set $d = d_2 = 1$. Four specifications for the true function $g_0(x)$ are considered: Linear: $Y_i = 1 + X_i + U_i$; Quadratic: $Y_i = 1 + .2X_i^2 + U_i$; Cubic: $Y_i = 1 + .2X_i^3 + U_i$; and Exponential: $Y_i = e^{.5X_i} + U_i$.

The (pseudo) random variables are generated as follows. (U_i, V_i)'s are *i.i.d.* random draws from joint Normal distribution with zero mean, unit variance, and correlation coefficient ρ. Z_i's are *i.i.d.* random draws from uniform distribution on $[-5, +5]$. X is then generated by $X_i = 1 + 0.5Z_i + V_i$.[1] Therefore, X is exogenous if $\rho = 0$ and (severely) endogenous if $\rho = 1$. Except for Figures 5-8, set $\rho = .5$ and the correlation coefficient of X and U is .35. Next, Y is generated by $Y_i = g_0(X_i) + U_i$,

[1] For figue 9 and 10, the dimension of Z is 3, where for each component of Z, random draws are generated from the uniform distribution on $[-2, +2]$; the $X_i = Z_{1i} + Z_{2i} + Z_{3i} + V_i$.

where $g_0(x)$ is specified as above.

For each specification, I run R times of simulations. At each replication, the random draws of Z and (U, V) are generated as described above. The estimation follows the procedure described in Section 3.2. When V is unobserved, $\widetilde{r}(z)$ is estimated by (3.5), $\widehat{V}$ by (3.6), $\widehat{m}(x, v)$ by (3.7), and $\widehat{g}(x)$ by (3.9). When V is observed, $\widetilde{m}(x, v)$ is estimated by (3.8), and $\widetilde{g}(x)$ by (3.10). The bandwidths are chosen by the rule of thumb selector: $h_1 = 1.06\widehat{\sigma}_Z \cdot n^{-1/5}$ if $d_Z = 1$ and $h_1 = 1.06\widehat{\sigma}_Z \cdot n^{-1/7}$ if $d_Z = 3$; $h = 1.06\widehat{\sigma}_{XV} \cdot n^{-1/6}$. The Epanechnikov's kernel is used, which satisfies Assumption K with $s = 2$. For comparison purpose, $\overline{g}(x)$ is also estimated by kernel regression of Y on X using bandwidth $1.06\widehat{\sigma}_X \cdot n^{-1/5}$. Trimming is slightly more complicated than the procedure in Section 3.2 because, besides boundary effects, we need to account for the random denominator problem in kernel regressions when the data are sparse in some region. Therefore, to handle both problems in the estimation, I apply the trimming suggested by Robinson (1988) by specifying the estimated density to be bounded from bottom by n^{-1}.

The estimation is carried out on a grid with 100 equally-spaced points for x and for v. On each point of the grid, I show the mean, the 5% percentile and 95% percentile of the 100 estimates of $\widehat{g}(x)$, which gives us the pointwise 90% confidence interval for $g_0(x)$. I also compare $\widehat{g}(x)$ to $\widetilde{g}(x)$ and $\overline{g}(x)$. In the figures, the estimates between 5% and 95% percentile of x are shown. The solid black line is the true function $g_0(x)$; the red line of plus sign is the mean of $\widehat{g}(x)$ estimates, 90% of which are contained between the two red doted lines; the blue dashed line is $\widetilde{g}(x)$; and the blue dash-doted line is $\overline{g}(x)$. For Figures E.7-10, only $g_0(x)$ and the 90% confidence intervals of $\widehat{g}(x)$ estimates and of $\widetilde{g}(x)$ estimates are reported as there is no significant difference between their means.

For the semiparametric control function (SPCF) estimator $\widehat{\beta}$ proposed in Chapter 4, let $Y_i = X_i'\beta + U_i$ as we are interested in the Cobb-Douglas production function. Here, $X_i = (X_{1i}, X_{2i})' \in \mathbb{R}^2$ and $\beta = (0.3, 0.7)'$. Set the sample size $n = 100$, $n = 500$, and $n = 1000$, and the number of replications $R = 100$. At each replication, U_i's are random draws from $N(0,1)$. (V_{1i}, V_{2i})'s are generated by $V_{pi} = \rho U_i + T_{pi}$ $(p = 1, 2)$, where T_{pi}'s are random draws from uniform distribution on $[-1, 1]$. Then X_{pi}'s are generated by $X_{pi} = 0.9Z_{pi} + V_{pi}$, where Z_{pi}'s are random draws from uniform distribution on $[0, 5]$. So X is endogenous if $\rho \neq 0$, and I set $\rho = 0.5$. Then the control $V_i = (V_{1i}, V_{2i})$ is estimated by (4.2) and β by (4.3). The bandwidths and kernels are the same as above.

5.2 Simulation Results

I first report the performance of $\widehat{g}(x)$ compared to $\overline{g}\,(x)$ in these four model specifications. In the first set of figures, from Figures E.1 to E.4, the true function $g_0(x)$, $\widehat{g}(x)$ with its 90% confidence interval, and $\overline{g}\,(x)$ are displayed for each specification. In every case, $\widehat{g}(x)$ is consistent with the true function $g_0(x)$ contained in the 90% confidence interval of $\widehat{g}(x)$. At the same time, $\overline{g}\,(x)$ is inconsistent for all four specifications, which is expected as it does not account for the potential endogeneity of X.

The second set of figures shows how the correlation coefficient ρ affects $\widehat{g}(x)$, $\widetilde{g}(x)$ and $\overline{g}\,(x)$.[2] When $\rho = 0$, X is exogenous but becomes severely endogenous when $\rho = 1$, which means that U and V are perfectly correlated. Figure E.5 shows that when there is no endogeneity problem, $\overline{g}\,(x)$ actually outperforms $\widehat{g}(x)$ as the latter has noises from the first and second step estimation. Figure E.6 is the

[2] From now on, I only report the results for the quadratic specification to save space and the results for other specifications are similar.

opposite case where $\overline{g}\,(x)$ is inconsistent with large biases and $\widehat{g}(x)$ is consistent except in some small area near the boundary. This is because the high (positive) correlation between X and V means that the data points are sparse in the area in the southeast and northwest corners of the support of (x, v). Thus the boundary effects are further aggravated by the correlation between X and V.

It is also interesting to see how $\widetilde{g}(x)$ changes with ρ, compared to $\widehat{g}(x)$. When $\rho = 0$, V contains no information about U so that there is no significant difference between $\widehat{g}(x)$ and $\widetilde{g}(x)$ as shown in Figure E.7. Also note that there the boundary effects are absent in Figure E.7 as the data points are evenly distributed over the support of (x, v). However, when $\rho = 1$, V is a very good control for U and $\widetilde{g}(x)$ should and does perform very well, as shown in Figure E.8, which indicates control function approaches really work. In this case, as indicated by tighter confidence intervals, $\widehat{g}(x)$ is dominated by $\widetilde{g}(x)$ due to the fact that the control V has to be estimated for $\widehat{g}(x)$.

As indicated in Theorems 3.1 and 3.2, higher dimension d_z of the instrument Z slows down the convergence rate of step-1 estimation and eventually affects $\widehat{g}(x)$ if it converges slower than $\overline{g}(v)$. However, it has no effect on $\widetilde{g}(x)$ because V is observed and not estimated from Z. In Figures E.9 and E.10, d_z increases to 3, as expected by Theorems 3.1 and 3.2, $\widetilde{g}(x)$ outperforms $\widehat{g}(x)$.

Finally, since bandwidths are critical parameters in kernel estimation, I check whether the bandwidth selector used here is a good one and whether the estimation is sensitive to bandwidths. As mentioned above, the rule of thumb selector is used and $h = 1.06\widehat{\sigma}_X n^{-1/5}$. The step-2 bandwidth is set to $h/2$ in Figure E.11 and $2 * h$ in Figure E.12. These two figures show the trade-off between the bias and variance of the estimator. Due to the undersmoothing in Figure E.11, $\widehat{g}(x)$ has larger variances but smaller biases than that in Figure E.12, where the oversmoothing results in smaller variances and larger biases. This fairly comprehensive set of Monte Carlo simulations shows that both $\widehat{g}(x)$ and $\widetilde{g}(x)$ perform well in finite samples under various scenarios.

As for the SPCF estimator $\widehat{\beta}$, the simulation results in Table D.1 in Appendix D indicate that $\widehat{\beta}$ perform better as sample size increases. Although $\widehat{\beta}$ is $\sqrt{n}$-consistent, the nonparametric preliminary estimators, i.e, $\widehat{V}$, $\widehat{X}$ and $\widehat{Y}$, converge slower than the parametric rate and are more sensitive to sample sizes and dimensions.

Chapter 6

Empirical Example

There are two purposes in this empirical example. The first is to illustrate the empirical relevance of the identification strategy and the nonparametric and semi-parametric estimators proposed in this thesis. After designing and testing the procedures (as in Chapters 2-4 and 5), we want to see how it works in practice, which is related to the second purpose. We want to make comparison to alternative methods, especially the methods by Olley and Pakes (1996)/Levinsohn and Petrin (2003), and by Ackerberg, Caves and Frazer (2006).

6.1 The Dataset and Estimators

For the empirical example, I use the same Chilean data set as Levinsohn and Petrin (2003) and Ackerberg, Caves and Frazer (2006). This Chilean panel is representative of many firm/plant level panels, in which investments and many intermediate inputs are reported along with capital and labor.[1] The focus here is how the endogeneity of inputs is addressed using control function approaches. Since the interest is in demonstrating the empirical relevance of the proposed identification strategy

[1] Details about this dataset can be found in Roberts and Tybout (1996).

and kernel estimators, I only report the results for food industry (CIIU 311) in 1986.[2] This industry is suitable for nonparametric estimation as it has about 800 observations each year, the largest one among surveyed industries. See Table E.1 in Appendix E for summary statistics of the subsample used in the estimation.

Similar to ACF, I estimate *value-added* production functions ($x_{it} = (k_{it}, l_{it})$) rather than *gross-revenue* production functions ($x_{it} = (k_{it}, l_{it}, w_{it})$). In the Chilean dataset, w_{it} includes materials, electricity, and fuels. For $\widehat{g}(x)$, the curse of dimension will make the estimation imprecise given the sample size. The dimension of $x_{it} = (k_{it}, l_{it}, w_{it})$ is 5 while that of $x_{it} = (k_{it}, l_{it})$ is 2, which is also suitable for graphic display of the estimates of $g(x)$. For $\widehat{\beta}$, one concern with estimating a gross-revenue production function is that the elements in w_{it} are highly collinear with each other (and with k_{it} and l_{it} as well). This multicollinearity will make the estimates of β instable.

As discussed in Chapter 2, I estimate the production function with controls estimated from lagged levels of capital and labor as instruments. Both $(k_{i,t-1}, l_{i,t-1})$ and $(k_{i,t-2}, l_{i,t-2})$ are used as the instruments and, respectively, denote the NPCF estimators as $\widehat{g}_t^{k_1 l_1}$and $\widehat{g}_t^{k_2 l_2}$, and the SPCF estimators as $\widehat{\beta}_t^{k_1 l_1}$and $\widehat{\beta}_t^{k_2 l_2}$. This shows the flexibility in the choice of instruments and, as indicated in Section 6.2, the estimates are not sensitive to the choice of instruments.

The estimation then follows from the procedures proposed in Chapters 3 and 4. The control V is constructed by (3.6). Similar to Chapter 5, the second-order Epanechnikov kernel is used, the bandwidths are chosen by the rule of thumb, exogenous trimming suggested by Robinson (1988) is adopted.[3] Both $\widehat{g}_t^{k_1 l_1}$and $\widehat{g}_t^{k_2 l_2}$ are estimated by (3.9), and $\widehat{\beta}_t^{k_1 l_1}$and $\widehat{\beta}_t^{k_2 l_2}$ by (4.3). $g_t(k, l)$ is estimated on a

[2]More estimation results on other industries/years are available from the author.

[3]Since $\widehat{\beta}$ is density weighted, no trimming is needed in the second and third steps.

30×20 grid of (k, l), which covers the region between 10% and 90% percentiles of k and of l respectively. Since $\widehat{g}(x)$ and $\widehat{\beta}$ are complex kernel estimators, I resort to the basic idea of bootstrap, treat the sample as population, and directly resample from the data. See Horowitz (2001) for an extensive review of bootstrap.

In order to make the comparison to alternative methods, I also applying LP and ACF methods to the sample used for $\widehat{\beta}$, denoting the LP estimator as β^{LP} and the ACF estimator as β^{ACF}.[4] For β^{LP} and β^{ACF}, I use either material or electricity as the proxy. To highlight the endogeneity issue, I also compute the standard OLS and fixed-effects estimators, denoted as β^{OLS} and β^{FE} respectively.

Nonparametric identification and estimation are robust to misspecification of underlying data generating processes (DGPs). This is desirable because little is actually known about the true DGPs for the surveyed industries in Chile. However, in order to make a comparison to β estimated using the methods of SPCF, LP and ACF, I compute the density-weighted mean coefficients of capital and labor. For production function estimator $\widehat{g}(k, l)$, the capital and labor coefficients estimates $\beta^{NP} \equiv \left(\overline{\beta}_k, \overline{\beta}_l\right)$ are computed as follows

$$\overline{\beta}_k \equiv \sum\nolimits_{i,j} \widehat{g}_k(k_i, l_j)\,\omega(k_i, l_j) \text{ and } \overline{\beta}_l \equiv \sum\nolimits_{i,j} \widehat{g}_l(k_i, l_j)\,\omega(k_i, l_j),$$

where both $\widehat{g}_k(k_i, l_j)$ and $\widehat{g}_l(k_i, l_j)$ are partial derivative defined as follows

$$\widehat{g}_k(k_i, l_j) \equiv \frac{\widehat{g}(k_{i+1}, l_j) - \widehat{g}(k_i, l_j)}{k_{i+1} - k_i} \text{ and } \widehat{g}_l(k_i, l_j) \equiv \frac{\widehat{g}(k_i, l_{j+1}) - \widehat{g}(k_i, l_j)}{l_{j+1} - l_j},$$

The weight $\omega(k_i, l_j) \equiv \widehat{f}(k_i, l_j) \big/ \sum_{t,s} \widehat{f}(k_t, l_s)$, where $\widehat{f}(k_i, l_j)$ is the density esti-

[4] See LP and ACF for details of their estimation procedures. Here, third-degree polynomials are used.

mated at (k_i, l_j).[5]

6.2 Estimation Results

Since the Cobb-Douglas production summarizes the industry in a succinct way, let's consider the estimates of β. All the estimates are reported in Table E.2 in Appendix E with the bootstrapped standard errors in parentheses. Notice that both $\widehat{\beta}$ and $\overline{\beta}$ are not sensitive to the choice of controls/instruments. Switching the instrument from $z_{it} = (k_{i,t-1}, l_{i,t-1})$ to $z_{it} = (k_{i,t-2}, l_{i,t-2})$, the estimates are not significantly different. It changes from 0.297 to 0.303 for $\overline{\beta}_k$, from 0.807 to 0.814 for $\overline{\beta}_l$; from 0.369 to 0.372 for $\widehat{\beta}_k$, and from 0.765 to 0.770 for $\widehat{\beta}_l$.

First, let's compare $\widehat{\beta}$ and $\overline{\beta}$ to the LP estimates β^{LP} and ACF estimates β^{ACF}. For the capital coefficient β_k, $\widehat{\beta}_k$ and $\overline{\beta}_k$ are significantly smaller than either β_k^{LP} or β_k^{ACF}, no matter which set of instruments is used. On the other hand, $\widehat{\beta}_l$ and $\overline{\beta}_l$ is smaller than β_l^{ACF} but larger than β_l^{LP}. Together, the return to scale parameter is around 1.1 using the methods proposed in this paper. Given that the major portion of observations came from bakery in Chile in 1980's, it is reasonable to describe the industry as labor-intensive with slightly increasing returns to scale. Therefore, it is reasonable to believe that $\widehat{\beta}$ and $\overline{\beta}$ strike a better balance. Without controlling for the endogeneity, the OLS estimator β^{OLS} overestimates β_l and the return to scale parameter while the FE estimator β^{FE} underestimates β_k and the return to scale parameter.

Second, comparing $\widehat{\beta}$ to $\overline{\beta}$, $\widehat{\beta}$ gives higher estimates of the capital coefficient β_k but lower estimates of the labor coefficient β_l. However, the estimates of return to scale are not significantly different from each other. A possible explanation is

[5] See Pagan and Ullah (1999) for nonparametric estimation of derivatives and their average.

different weighting schemes involved in $\widehat{\beta}$ and $\overline{\beta}$. For $\widehat{\beta}$, we impose that all firms (plants) have the same coefficients. In contrast, firms with different capital and labor stocks may have different values of β and $\overline{\beta}$ is an (density-weighted) average of those values. Indeed, large firms tend to have higher values of β than small firms, which is also apparent in nonparametric estimates of $g(k,l)$. Therefore, the Cobb-Douglas production function may not be a good approximation to the industry of interest.

Nonparametric estimation is more flexible than parametric specification using the Cobb-Douglas production function, and enables us to learn more about the industry beyond two coefficients (β_k, β_l). All nonparametric estimates are reported in Appendix E. In Figures E.1 and E.2, I first present $\widehat{g}_t^{k_1 l_1}$and $\widehat{g}_t^{k_2 l_2}$, the estimates using instruments $z_{it} = (k_{i,t-1}, l_{i,t-1})$ and $z_{it} = (k_{i,t-2}, l_{i,t-2})$ respectively. Both figures give us a similar big picture of the industry. The production function $g(k,l)$ is increasing in k and l. In addition, the estimates of derivatives of $g(k,l)$ (i.e, the β's) with respect to k and to l are not constant across (k,l). Thus, it seems restrictive to assume β to be constant as in the case of Cobb-Douglas production. Figures E.3 shows the difference between $\widehat{g}_t^{k_1 l_1}$and $\widehat{g}_t^{k_2 l_2}$, which is not significantly different from zero except on some corner regions.[6] The stability of estimates of g_t indicates that the proposed estimator $\widehat{g}$ is not sensitive to the choice of instruments.

With consistent estimates of g_t, we can recover the idiosyncratic productivity shock using the method proposed in Section 2.3. a_{it} can be consistently estimated by (2.11), and its empirical distribution $\widehat{f}(a)$ is then estimated from $\widehat{a}_{it}$, as shown in Figure E.4. It is clear that $\widehat{f}(a)$ is not symmetric and the normal distribution

[6]The regions where $\widehat{g}_t^{k_1 l_1}$and $\widehat{g}_t^{k_2 l_2}$ do not agree with each other are points (k,l)'s with large k and small l, or small k and large l. The data on these regions are sparse due to the proportion between k and l in the food insdustry. Due to the difficulty in 3-dimensional display, the bootstrapped (pointwise) confidence intervals are not shown in the figures.

may not be a good approximation to $f(a)$. Note that the mean of $\widehat{a}_{it}$ is not significant different from zero, which is consistent with the location normalization in Chapters 2 and 3.

Let's see how the endogeneity of inputs in production estimation is addressed using control function approaches. For comparison purpose, Figure E.5 shows the estimate of $\overline{g}_t(k,l)$, the conditional expectation of y given (k,l). Figure E.6 then shows the difference between $\overline{g}_t(k,l)$ and $\widehat{g}_t^{k_1 l_1}(k,l)$, which is also the conditional mean of productivity shock a_{it} given (k,l). It appears that a_{it} is positively correlated to capital and labor. An interpretation is that higher a_{it} induces firms to have higher levels of capital.

Besides the function $g(k,l)$ of interest, the control function $c(v)$ is also estimated. Figures E.7 shows the control function $\widehat{c}(v)$, where controls $v_t = \left(v_t^k, v_t^l\right)$ are estimated from instruments $z_{it} = (k_{i,t-1}, l_{i,t-1})$. Except for some values of v near the two corners, $\widehat{c}(v)$ is increasing in both v_t^k and v_t^l.[7] This is essence of control function approaches: the control v_t moves along with the productivity so that v_t can be used to control for the unobserved productivity.

A firm's output is determined by a_{it} and $g_t(x_{it})$, where $g_t(x)$ is the same for every firm at t. Levinsohn and Petrin (1999) ask the following question. What makes an industry more productive: relocation of resources from less productive firm to more productive ones, or progress of all firms? If the answer is the latter, we see $g_t(x)$ increases with t. This question can be better answered by the difference between $\widehat{g}_{86}^{k_1 l_1}$ and $\widehat{g}_{85}^{k_1 l_1}$, shown in Figure E.8. We see that most firms (around 70%) becomes significantly more productive from 1985 to 1986.[8] Certainly, macroeconomic shocks are incorporated into $g_t(x)$ so that the conclusion is

[7] Figure E.7 can be compared to Figure 1 in LP, where controls are observed.

[8] Again, the estimates on the two corners are not precise due to sparseness of data on those two corners.

not without qualification. The purpose is to show the potential of nonparametric control function approaches in production function estimation.

Bibliography

1. Ackerberg, D., Caves, K., and G. Frazer (2006). Structural Identification of Production Functions, Working Paper, UCLA.

2. Adams, R.M., A.N. Berger and R.C. Sickles (1999. Semiparametric Approaches to Stochastic Panel Frontiers with Applications in the Banking Industry, Journal of Business and Economic Statistics, 17, 349-58.

3. Ahn, H. (1995). Nonparametric Two-stage Estimation of Conditional Choice Probabilities in a Binary Choice Model under Uncertainty, Journal of Econometrics, 67: 337-378.

4. Ahn, H. and J. Powell (1993). Semiparametric Estimation of Censored Selection Models with a Nonparametric Selection Mechanism, Journal of Econometrics, 58: 3-29.

5. Ai, C. (1997) A Semiparametric Maximum Likelihood Estimator Econometrica, 65: 933-964 .

6. Amemiya, T. (1985). Advanced Econometrics, Cambridge, MA, Harvard University Press.

7. Angrist, J. and A. Krueger (1999). Empirical Strategies in Labor Economics, in Handbook of Labor Economics, 3A, ed. by Ashenfelter and Card, Elsevier Science, Amsterdam.

8. Arellano, M. and S. Bond (1991). Some Tests of Specification for Panel Data: Monte Carlo Evidence and an Application to Employment Equations, Review of Economic Studies 58: 277-297.

9. Arellano, M. and O. Bover (1995). Another Look at the Instrumental Variable Estimation of Error Components Models, Journal of Econometrics, 68: 29-51.

10. Athey, S, and P. A. Haile (2007). Nonparametric Approaches to Auctions, in: J.J. Heckman & E.E. Leamer (ed.), Handbook of Econometrics, Vol. 6, Chapter 60, Elsevier.

11. Bierens, H.J. (1987). Kernel Estimators of Regression Functions, in Advances in Econometrics: Fifth World Congress, Vol.I, Cambridge University Press, London.

12. Blundell, R. and S. Bond (2000). GMM Estimation with Persistent Panel Data: An Application to Production Functions, Econometric Review, 19: 321-340.

13. Blundell R., Duncan, A. and K. Pendakur (1998). Semiparametric estimation and consumer demand, Journal of Applied Econometrics 13, 453-461.

14. Blundell, R.W. and J.L. Powell (2003). Endogeneity in Semiparametric and Nonparametric Regression Models, in Advances in Econometrics: Eighth World Congress, Vol.II, Cambridge University Press, London.

15. Blundell, R.W. and J.L. Powell (2004). Endogeneity in Single Index Models, Review of Economic Studies, 71: 655-679.

16. Bond, S. and M. Söderbom (2005). Adjustment Costs and the Identification of Cobb-Douglas Production Functions, The Institute for Fiscal Studies, Working Paper Series No. 05/04.

17. Chen, R., Härdle, W., Linton, O. B., and E. Severance-Lossin (1996). Nonparametric estimation of additive separable regression models, in Statistical Theory and Computational Aspects of Smoothing, ed. by W. Härdle and M. Schimek, 247-254, Physica-Verlag.

18. Darolles, S., J.-P. Florens, and E. Renault (2002). Nonparametric Instrumental Regression, working paper, GREMAQ, University of Social Science, Toulouse.

19. Das, M., W. K. Newey, and F. Vella (2003). Nonparametric Estimation of Sample Selection Models, Review of Economic Studies, 70, 33–58.

20. Donald, S. G., and W. K. Newey (2001). Choosing the Number of Instruments, Econometrica, 69, 1161–1191.

21. Engle, R.F., Grange, C.W.J., Rice, J., Weiss, A., 1986. Semiparametric Estimates of Time Relationship Between Weather and Electricity Sales. Journal of the American Statistical Association 81, 310–320.

22. Florens, J.-P. (2003). Inverse Problems in Structural Econometrics: The Example of Instrumental Variables, in Advances in Economics and Econometrics, Vol. 2, eds by M. Dewatripont, L.P. Hansen and S.J. Turnovsky, Cambridge University Press, 284-311.

23. Florens, J.-P. and L. Malavolti (2002). Instrumental Regression with Discrete Variables, Mimeo, University of Toulouse.

24. Griliches, Z. and J. Mairesse (1998). Production Functions: The Search for Identification, in Econometrics and Economic Theory in the Twentieth Century, ed. by S. Strøm, Cambridge University Press, London.

25. Guerre, E., I. Perrigne, and Q. Vuong (2000). Optimal Nonparametric Estimation of First-Price Auctions, Econometrica, 68: 525-574.

26. Hall, P. and J.L. Horowitz (2005). Nonparametric Methods for Inference in the Presence of Instrumental Variables, Annals of Statistics,33, 2904-2929.

27. Härdle, W. (1994). Applied Nonparametric Regression, Cambridge University Press, London.

28. Härdle, W., H. Liang and J. Gao (2000). Partially Linear Models. Springer-Verlag.

29. Hastie, T., and R. Tibshirani. (1990). Generalized additive models, Chapman & Hall, London.

30. Heckman, J. J. (1976). The Common Structure of Statistical Models of Truncation, Sample Selection and Limited Dependent Variables and a Simple Estimator for Such Models, Annals of Economic and Social Measurement, 5, 475–492.

31. Heckman, J. and R. Robb (1985). Alternative Methods for Estimating The Impact of Interventions, in Longitudinal Analysis of Labor Market Data, ed. by J. Heckman and B. Singer, Cambridge University Press, London.

32. Heckman, J. J. (1978). Dummy Endogenous Variables in a Simultaneous Equations System, Econometrica, 46, 931–960.

33. Heckman, J. J. (1979). Sample Selection as a Specification Error, Econometrica, 47, 153–161.

34. Heckman, J. J., and V. Hotz (1989). Choosing among Alternative Nonexperimental Methods for Evaluating the Impact of Social Programs, Journal of the American Statistical Association, 84, 862–880.

35. Heckman, J. J., and R. Robb (1985). Alternative Methods for Evaluating the Impact of Interventions, in Longitudinal Analysis of Labor Market Data, J. J. Heckman and B. Singer (Eds.), Cambridge, UK, Cambridge University Press.

36. Heckman, J. and E. J. Vytacil (2007). Econometric Evaluation of Social Programs (Part I &II), forthcoming in Handbook of Econometrics, Vol. 6, ed. by J. Heckman and E. Leamer, Elsevier Science, Amsterdam.

37. Hopenhayn, A. (1992). Exit, Selection and the Value of Firms", Journal of Economic Dynamics and Control, 10, 621-653.

38. Horowitz, J. L. (2001). The Bootstrap, in Handbook of Econometrics, Vol. 5, ed. by J.J. Heckman and E.E. Leamer, Elsevier Science, Amsterdam.

39. Ichimura, H. and L.-F. Lee (1991). Semiparametric least squares estimation of multiple index models: single equation estimation, in International Symposia in Economic Theory and Econometrics, edited by William A. Barnett, James Powell, and George Tauchen. Cambridge University Press, 3-49.

40. Jovanovic, B. (1982). Selection and the Evolution of Industry. Econometrica, 50, 649-670.

41. Levinsohn, J. and A. Petrin (1999). When Industries Become More Productive, Do Firms?: Investigating Productivity Dynamics, Working Paper, University of Michigan.

42. Levinsohn, J. and A. Petrin (2003). Estimating Production Functions Using Inputs To Control For Unobservables, Review of Economic Studies, 70: 317-341.

43. Li, Q. (1996). On the root-N-consistent semiparametric estimation of partially linear models, Economics Letters, 1996, 51, 277-285.

44. Li, Q. and J.S. Racine (2006). Nonparametric Econometrics, Princeton University Press.

45. Li, Q. and J. M. Wooldridge (2002). Semiparametric Estimation Of Partially Linear Models For Dependent Data With Generated Regressors, Econometric Theory, 18 (03), 625-645.

46. Linton, O. B. and Nielsen, J. P. (1995). A kernel method of estimating structured nonparametric regression based on marginal integration. Biometrika 82 93-100.

47. Matzkin, R. L. (2006). Nonparametric Identification, forthcoming in Handbook of Econometrics, Vol. 6, ed. by J. Heckman and E. Leamer, Elsevier Science, Amsterdam.

48. Newey, W. K. (1994). Kernel Estimation of Partial Means and a General Variance Estimator, Econometric Theory, 10: 233-253.

49. Newey, W.K. and J.L. Powell (2003). Instrumental Variable Estimation of Nonparametric Models, Econometrica, 71, 1565-1578.

50. Newey, W.K., J.L. Powell, and F. Vella (1999). Nonparametric Estimation of Triangular Simultaneous Equations Models, Econometrica, 67: 565-603.

51. Olley, S. and A. Pakes (1996). The Dynamics of Productivity in the Telecommunications Equipment Industry, Econometrica, 64: 1263–98.

52. Pagan, A.(1984). Econometric Issues in the Analysis of Regressions with Generated Regressors, International Economic Review, 25: 221-47.

53. Pagan, A. and A. Ullah (1999). Nonparametric Econometrics, Cambridge University Press.

54. Pinkse, J. (2000). Nonparametric Two-step Regression Estimation when Regressors and Error are Dependent, Canadian Journal of Statistics, 28: 289-300.

55. Powell, J. L. (1986). Symmetrically Trimmed Least Squares Estimation for Tobit Models, Econometrica, 54, 1435–1460.

56. Powell, J., J. Stock and T. Stoker (1989). Semiparametric Estimation of Index Coefficients, Econometrica, 57: 1403-30.

57. Rilstone, P. (1996). Nonparametric Estimation of Models with Generated Regressors, International Economic Review, 37: 299-313.

58. Rivers, D. and Q. H. Vuong (1988). Limited Information Estimators and Exogeneity Tests for Simultaneous Probit Models, Journal of Econometrics, 39: 347-366.

59. Roberts, M. and J. R. Tybout (1996). Industrial Evolution in Developing Countries: Micro Patterns of Turnover, Productivity, and Market Structure, Oxford University Press, London.

60. Robinson, P.M. (1988b) $\sqrt{N}$-consistent semiparametric regression, Econometrica 56: 931-954.

61. Schmalensee, R. and Stoker, T. M. (1999). Household Gasoline Demand in the United States. Econometrica 67(3): 645-662.

62. Smith, R. and R. Blundell (1986). An Exogeneity Test for a Simultaneous Equation Tobit Model with an Application to Labor Supply, Econometrica, 54: 679-685.

63. Speckman, P. (1988). Kernel Smoothing in Partial Linear Models, Journal of the Royal Statistical Society, B, 50, 413–446.

64. Stengos, T. and B. Yan, (2001). Double Kernel Nonparametric Estimation in Semiparametric Econometric Models, Nonparametric Statistics 13, 883-906.

65. Stock, J.H., (1989). Nonparametric Policy Analysis. Journal of American Statistical Association 84, 567–575.

66. Stone, C.J. (1982). Optimal Global Rates of Convergence for Nonparametric Regression, Annal of Statistics, 10: 1040-53.

67. Tobin, J. (1958). Estimation of Relationships for Limited Dependent Variables, Econometrica, 26, 24–36.

68. Wooldridge, J. (2005). On Estimating Firm-level Production Functions Using Proxy Variables to Control for Unobservables, mimeo, Michigan State University.

Appendix A

Uniform Consistency

Lemma A.1. *Let Assumptions K, B, and D hold, for $C'_{XV} \subsetneq S_{XV}$ defined in Section 3,*

$$\sup_i 1\left(\left(X_i, \widehat{V}_i\right) \in C_{XV}\right)\left|\widehat{V}_i - V_i\right| = O\left(\gamma_1\right) \ a.s.$$

Proof: By the definition of $\widehat{V}_i$ in (3.6), $\left(X_i, \widehat{V}_i\right) \in C_{XV}$ implies $Z_i \in C_Z$, where C_Z is the inner subset of the support S_X of Z. Since $\widehat{V}_i - V_i = r_0(Z_i) - \widetilde{r}(Z_i)$, if $\widehat{V}_i \neq \infty$,

$$\begin{aligned} \sup_i 1\left(\left(X_i, \widehat{V}_i\right) \in C_{XV}\right)\left|\widehat{V}_i - V_i\right| &= \sup_i 1\left(Z_i \in C_Z\right)\left|\widetilde{r}(Z_i) - r_0(Z_i)\right| \\ &\leqslant \sup_{C_Z}\left|\widetilde{r}(z) - r_0(z)\right| = O\left(\gamma_1\right) a.s., \end{aligned}$$

where $\sup_{C_Z}\left|\widetilde{r}(z) - r_0(z)\right| = O\left(\gamma_1\right)$ a.s. is a standard result in nonparametric econometrics. $\square$

Lemma A.2. *For any $x \in C_X$, let $p_x \equiv \Pr\left(V \in C_V^x\right)$, then*

(i) $n^{-1}\sum_{i=1}^n 1_{C_{XV}}\left(x, V_i\right) \xrightarrow{a.s.} p_x$; *and*

(ii) $n^{-1}\sum_{i=1}^n 1_{C_{XV}}\left(x, \widehat{V}_i\right) \xrightarrow{a.s.} p_x$, where $C_V^x \equiv \{v \in S_V : (x, v) \in C_{XV}\}$ *and* $C_Z^x \equiv \{z \in S_Z : z = r^{-1}(x - v), v \in C_V^x\}$.

Proof: (i) Notice that $\mathbf{1}_{C_{XV}}(x, V_i) = \mathbf{1}(V_i \in C_V^x)$, so that by the Law of Large Numbers,

$$n^{-1}\sum_{i=1}^n \mathbf{1}_{C_{XV}}(x, V_i) = n^{-1}\sum_{i=1}^n \mathbf{1}(V_i \in C_V^x) \xrightarrow{a.s.} \Pr(V \in C_V^x) = p_x.$$

(ii) Notice that $p_x = \Pr(V \in C_V^x) = \Pr(x - r(Z) \in C_V^x) = \Pr(Z \in C_Z^x)$. Now define $C_Z^{x\prime} \equiv \{z \in S_Z : z = \widetilde{r}^{-1}(x - v), v \in C_V^x\}$. Since $\widehat{V}$ is a uniformly consistent estimator of V for $Z \in C_Z$, $C_Z^x \subseteq C_Z^{x\prime}$. As $n \to \infty$, $\Pr(Z \in C_Z^{x\prime} \backslash C_Z^x) = 0$ so that $\Pr(Z \in C_Z^{x\prime}) = \Pr(Z \in C_Z^x)$ and

$$\begin{aligned} & n^{-1}\sum_{i=1}^n \mathbf{1}_{C_{XV}}\left(x, \widehat{V}_i\right) \\ = \ & n^{-1}\sum_{i=1}^n \mathbf{1}(Z \in C_Z^{x\prime}) \xrightarrow{a.s.} \Pr(Z \in C_Z^{x\prime}) = \Pr(Z \in C_Z^x) = p_x.\square \end{aligned}$$

Proof of Proposition 3.1: Part (a).

Note that $\widehat{f}(x,v) - f_0(x,v) = \left[\widehat{f}(x,v) - \widetilde{f}(x,v)\right] + \left[\widetilde{f}(x,v) - f_0(x,v)\right]$

so that by the triangle inequality,

$$\sup_{C_{XV}}|\widehat{f}(x,v) - f_0(x,v)| \leqslant \sup_{C_{XV}}|\widehat{f}(x,v) - \widetilde{f}(x,v)| + \sup_{C_{XV}}|\widetilde{f}(x,v) - f_0(x,v)|.$$

As a standard result of nonparametric econometrics,

$$\sup_{C_{XV}}|\widetilde{f}(x,v) - f_0(x,v)| = O(\gamma_2),$$

and it remains to find out the order of $\sup_{C_{XV}}|\widehat{f}(x,v) - \widetilde{f}(x,v)|$.[1]

[1]It is worthwhile to note that the same bandwidth and kernel should be used for $\widetilde{f}(x,v)$

Define C'_{XV} as an inner closed subset of S_{XV} containing all hypercubes of size δ (small enough) centered at a point (x,v) in C_{XV}; also define C''_{XV} similarly with respect to C'_{XV}. Thus $C_{XV} \subsetneq C'_{XV} \subsetneq C''_{XV} \subsetneq S_{XV}$. Note that for $(x,v) \in C_{XV}$ with n large enough, $\widehat{f}(x,v)$ uses at most observations $\left(X_j, \widehat{V}_j\right)$ in C'_{XV} with the corresponding point (X_j, V_j) in C''_{XV} because $\widehat{V}_j \xrightarrow{a.s.} V_j$ uniformly within C'_{XV} for all $\left(X_j, \widehat{V}_j\right)$. Also note that $\widetilde{f}(x,v)$ uses at most observations (X_j, V_j) in C''_{XV}. So almost surely for n large enough, for $(x,v) \in C_{XV}$

$$\widehat{f}(x,v)-\widetilde{f}(x,v) = \frac{1}{nh^{d+d_2}} \sum_{j=1}^{n} 1_{C''_{XV}}(X_j, V_j) \left(\kappa\left(\tfrac{x-X_j}{h}\right)\kappa\left(\tfrac{v-\widehat{V}_j}{h}\right) - \kappa\left(\tfrac{x-X_j}{h}\right)\kappa\left(\tfrac{v-V_j}{h}\right)\right),$$

where $\overline{V}_j$ is between $\left(\frac{v-\widehat{V}_j}{h}\right)$ and $\left(\frac{v-V_j}{h}\right)$, and $\iota = [1,1,...,1]^\intercal \in \mathbb{R}^{d_2}$. Then by a second order Taylor expansion, for $(x,v) \in C_{XV}$

$$\begin{aligned}
&\left|\widehat{f}(x,v) - \widetilde{f}(x,v)\right| \\
\leqslant\ & \frac{1}{nh^{d+d_2+1}} \sum_{j=1}^{n} \mathbf{1}_{C''_{XV}}(X_j, V_j) \left|\left(\widehat{V}_j - V_j\right)^\intercal \kappa'\left(\tfrac{v-V_j}{h}\right)\kappa\left(\tfrac{x-X_j}{h}\right)\right| \\
&+\frac{1}{2nh^{d+d_2+2}} \sum_{j=1}^{n} \mathbf{1}_{C''_{XV}}(X_j, V_j) \left|\left(\widehat{V}_j - V_j\right)^\intercal \kappa''\left(\overline{V}_j\right)\left(\widehat{V}_j - V_j\right)\kappa\left(\tfrac{x-X_j}{h}\right)\right| \\
\leqslant\ & \frac{1}{nh^{d+d_2+1}} \sum_{j=1}^{n} \mathbf{1}_{C''_{XV}}(X_j, V_j) \left|\widehat{V}_j - V_j\right|^\intercal \left|\kappa'\left(\tfrac{v-V_j}{h}\right)\right| \left|\kappa\left(\tfrac{x-X_j}{h}\right)\right| \\
&+\frac{1}{nh^{d+d_2+2}} \sum_{j=1}^{n} \mathbf{1}_{C''_{XV}}(X_j, V_j) \left|\widehat{V}_j - V_j\right|^\intercal \left|\kappa''\left(\overline{V}_j\right)\right| \left|\widehat{V}_j - V_j\right| \left|\kappa\left(\tfrac{x-X_j}{h}\right)\right| \\
\leqslant\ & O\left(\gamma_1 / h\right) \cdot \frac{1}{nh^{d+d_2}} \sum_{j=1}^{n} \iota^\intercal \left|\kappa'\left(\tfrac{v-V_j}{h}\right)\right| \left|\kappa\left(\tfrac{x-X_j}{h}\right)\right| \\
&+O\left(\gamma_1^2 / h^{2+d_2}\right) \cdot \sup_{v} \iota^\intercal \left|\kappa''(v)\right| \iota \cdot \frac{1}{nh^{d}} \sum_{j=1}^{n} \left|\kappa\left(\tfrac{x-X_j}{h}\right)\right|
\end{aligned}$$

where both the first and the second inequalities follow from the triangular in-

in $\left[\widehat{f}(x,v) - \widetilde{f}(x,v)\right]$ as for $\widetilde{f}(x,v)$ in $\left[\widetilde{f}(x,v) - f(x,v)\right]$ when we study the rates of uniform convergence.

equality, and the third one from Lemma A.1. Both $e^{\intercal}\left|\kappa'\left(\frac{v-V_j}{h}\right)\right|\left|\kappa\left(\frac{x-X_j}{h}\right)\right|$ and $\left|\kappa\left(\frac{x-X_j}{h}\right)\right|$ can be viewed as kernels except that they do not necessarily integrate to 1. By Assumption D,

$$\begin{aligned}
&\frac{1}{nh^{d+d_2}}\sum_{j=1}^{n}\iota^{\intercal}\left|\kappa'\left(\tfrac{v-V_j}{h}\right)\right|\left|\kappa\left(\tfrac{x-X_j}{h}\right)\right| \xrightarrow{a.s.} f_0(x,v)\cdot\int\iota^{\intercal}\left|\kappa'(\omega)\right|\left|\kappa(t)\right|dtd\omega,\\
&\frac{1}{nh^{d}}\sum_{j=1}^{n}\left|\kappa\left(\tfrac{x-X_j}{h}\right)\right| \xrightarrow{a.s.} f_{X_0}(x)\cdot\int\left|\kappa(t)\right|dt,
\end{aligned}$$

both of which are bounded almost surely. Thus we have,

$$\sup_{C_{XV}}|\widehat{f}(x,v)-\widetilde{f}(x,v)| = O\left(\gamma_1/h+\gamma_1^2/h^{2+d_2}\right).$$

Note that $\frac{s_1}{(2s_1+d_Z)}-\frac{1+d_2}{(2s+d+d_2)}>0$ implies that $O\left(\gamma_1/h_2\right)=o(1)$ and that

$$O\left(\frac{\gamma_1^2/h^{2+d_2}}{\gamma_1/h}\right)=O\left(\gamma_1/h^{1+d_2}\right)=\left(LogN/N\right)^{\frac{s_1}{(2s_1+d_Z)}-\frac{1+d_2}{(2s+d+d_2)}}=o\left(1\right),$$

so that $O\left(\gamma_1/h\right)>O\left(\gamma_1^2/h^{d_2+2}\right)$. Therefore we have

$$\sup_{C_{XV}}|\widehat{f}(x,v)-\widetilde{f}(x,v)| = O\left(\gamma_1/h\right)a.s.$$

Collecting the results, we have $\sup_{C_{XV}}|\widehat{f}(x,v)-f_0(x,v)|=O\left(\gamma_2+\gamma_1/h\right)$.

Part (b). First prove that $\sup_{C_{XV}}|\widehat{q}(x,v)-q_0(x,v)| = O\left(\gamma_2+\gamma_1/h\right)$ almost surely. By the triangle inequality,

$$\sup_{C_{XV}}|\widehat{q}(x,v)-q_0(x,v)| \leqslant \sup_{C_{XV}}|\widehat{q}(x,v)-\widetilde{q}(x,v)|+\sup_{C_{XV}}|\widetilde{q}(x,v)-q_0(x,v)|.$$

Again we know that $\sup_{C_{XV}}|\widetilde{q}(x,v)-q_0(x,v)|=O\left(\gamma_2\right)$, and it remains to find out the order of $\sup_{C_{XV}} |\widehat{q}(x,v) - \widetilde{q}(x,v)|$. Similar to Part (a), for $(x,v) \in C_{XV}$

$$
\begin{aligned}
& |\widehat{q}(x,v) - \widetilde{q}(x,v)| \\
\leqslant\; & \frac{1}{nh^{d+d_2+1}} \sum_{j=1}^{n} \mathbf{1}_{C''_{XV}}(X_j, V_j) \left|\widehat{V}_j - V_j\right|^{\mathsf{T}} \left|\kappa'\left(\tfrac{v-V_j}{h}\right)\right| . \left|\kappa\left(\tfrac{x-X_j}{h}\right)\right| . |Y_j| \\
& + \frac{1}{nh^{d+d_2+2}} \sum_{j=1}^{n} \mathbf{1}_{C''_{XV}}(X_j, V_j) \left|\widehat{V}_j - V_j\right|^{\mathsf{T}} \left|\kappa''\left(\overline{V}_j\right)\right| \left|\widehat{V}_j - V_j\right| . \left|\kappa\left(\tfrac{x-X_j}{h}\right)\right| . |Y_j| \\
\leqslant\; & O\left(\gamma_1 / h\right) \cdot \frac{1}{nh_2^{d+d_2}} \sum_{j=1}^{n} \iota^{\mathsf{T}} \left|\kappa'\left(\tfrac{v-V_j}{h}\right)\right| \left|\kappa\left(\tfrac{x-X_j}{h}\right)\right| |Y_j| \\
& + O\left(\gamma_1^2 / h^{2+d_2}\right) \cdot \sup_{v} \iota^{\mathsf{T}} \left|\kappa''(v)\right| \iota \cdot \frac{1}{nh^{d}} \sum_{j=1}^{n} \left|\kappa\left(\tfrac{x-X_j}{h}\right)\right| |Y_j| ,
\end{aligned}
$$

where $\overline{V}_j$ is between $\left(\frac{v-\widehat{V}_j}{h}\right)$ and $\left(\frac{v-V_j}{h}\right)$. By Assumption D, for $(x,v) \in C_{XV}$

$$
\begin{aligned}
& \frac{1}{nh^{d+d_2}} \sum_{j=1}^{n} \iota^{\mathsf{T}} \left|\kappa'\left(\tfrac{v-V_j}{h}\right)\right| \left|\kappa\left(\tfrac{x-X_j}{h}\right)\right| |Y_i| \\
& \xrightarrow{a.s.} E\left[|Y| \mid x, v\right] \cdot f_0(x,v) \cdot \textstyle\int \iota^{\mathsf{T}} \left|\kappa'(\omega)\right| \left|\kappa(t)\right| dt d\omega, \text{ and} \\
& \frac{1}{nh^{d}} \sum_{j=1}^{n} \left|\kappa\left(\tfrac{x-X_j}{h}\right)\right| |Y_i| \xrightarrow{a.s.} E\left[|Y| \mid X = x\right] \cdot f_{X_0}(x) \cdot \textstyle\int \left|\kappa(t)\right| dt,
\end{aligned}
$$

both of which are bounded almost surely. Thus, similar to Part (a), we have

$$
\sup_{C_{XV}} |\widehat{q}(x,v) - \widetilde{q}(x,v)| = O\left(\gamma_1 / h + \gamma_1^2 / h^{d_2+2}\right) = O\left(\gamma_1 / h\right) \text{ a.s,}
$$

so that almost surely

$$
\sup_{C_{XV}} |\widehat{q}(x,v) - q_0(x,v)| = O\left(\gamma_2 + \gamma_1 / h\right).
$$

Notice that

$$
\widehat{m}(x,v) - m_0(x,v) = \widehat{f}(x,v)^{-1} \left\{ \left[\widehat{q}(x,v) - q_0(x,v)\right] - m_0(x,v) \left[\widehat{f}(x,v) - f_0(x,v)\right] \right\},
$$

and that within C_{XV}, $m(x,v)$ is bounded and $\widehat{f}(x,v)$ is bounded away from zero a.s.. Thus for all $(x,v) \in C_{XV}$, almost surely we have

$$\begin{aligned}
& \sup_{C_{XV}} \left|\widehat{m}(x,v) - m_0(x,v)\right| \\
\leqslant \; & \sup_{C_{XV}} \left|\widehat{f}(x,v)^{-1}\right| \cdot \left\{ \sup_{C_{XV}} \left|\widehat{q}(x,v) - q_0(x,v)\right| + \sup_{C_{XV}} \left|m_0(x,v)\right| \cdot \sup_{C_{XV}} \left|\widehat{f}(x,v) - f_0(x,v)\right| \right\}.
\end{aligned}$$

Part (b) then follows from Part (a) and the results above. □

Proposition A.1: For $x \in C_X$, $\sup_{C_X} \left|\widehat{g}(x) - \widetilde{g}(x)\right| = O\left(\gamma_1 / h\right)$ a.s.
Proof: We need to check the order of $\sup_{C_X} \left|\widehat{g}(x) - \widetilde{g}(x)\right|$, i.e. the uniform convergence rate of

$$n^{-1} \sum_{i=1}^{n} \left[\mathbf{1}_{C_{XV}}^{p_x} \left(x, \widehat{V}_i\right) \widehat{m}(x, \widehat{V}_i) - \mathbf{1}_{C_{XV}}^{p_x} \left(x, V_i\right) \widetilde{m}(x, V_i) \right].$$

First, study the rate of uniform convergence of $\widehat{f}(x, \widehat{V}_i)$ to $\widetilde{f}(x, V_i)$ within C_{XV}:

$$\begin{aligned}
& \mathbf{1}_{C_{XV}}^{p_x} \left(x, \widehat{V}_i\right) \widehat{f}(x, \widehat{V}_i) - \mathbf{1}_{C_{XV}}^{p_x} \left(x, V_i\right) \widetilde{f}(x, V_i) \\
= \; & \mathbf{1}_{C_{XV}}^{p_x} \left(x, \widehat{V}_i\right) \left(\widehat{f}(x, \widehat{V}_i) - \widetilde{f}(x, \widehat{V}_i)\right) \\
& + \left(\mathbf{1}_{C_{XV}}^{p_x} \left(x, \widehat{V}_i\right) \widetilde{f}(x, \widehat{V}_i) - \mathbf{1}_{C_{XV}}^{p_x} \left(x, V_i\right) \widetilde{f}(x, V_i) \right). \qquad \text{(A.1)}
\end{aligned}$$

Consider the first term on the RHS of $(A.1)$. By Proposition 3.1.(a), for given $(x, \widehat{V}_i) \in C_{XV}$ and n large enough, we have

$$\begin{aligned}
& \sup_i \left| \mathbf{1}_{C_{XV}}^{p_x} \left(x, \widehat{V}_i\right) \left(\widehat{f}(x, \widehat{V}_i) - \widetilde{f}(x, \widehat{V}_i)\right) \right| \\
= \; & \sup_{\left(x, \widehat{V}_i\right) \in C_{XV}} \left| \left(\widehat{f}(x, \widehat{V}_i) - \widetilde{f}(x, \widehat{V}_i)\right) \right| = O\left(\gamma_1 / h\right). \qquad \text{(A.2)}
\end{aligned}$$

Now turn to the second term on the RHS of $(A.1)$. Similar to the proof of

Proposition 3.1, define C'_{XV} as an inner closed subset of S_{XV} containing all hypercubes of size δ centered at a point (x, v) in C_{XV} so that $C_{XV} \subsetneq C'_{XV} \subsetneq S_{XV}$. For $\left(x, \widehat{V}_i\right) \in C_{XV}$ and n large enough, $\widetilde{f}(x, \widehat{V}_i)$ uses at most observations (X_j, V_j) in C'_{XV} and so does $\widetilde{f}(x, V_i)$ for $(x, V_i) \in C_{XV}$. Therefore,

$$\begin{aligned}
& \mathbf{1}^{p_x}_{C_{XV}}\left(x, \widehat{V}_i\right) \widetilde{f}(x, \widehat{V}_i) - \mathbf{1}^{p_x}_{C_{XV}}(x, V_i)\, \widetilde{f}(x, V_i) \\
= \;& \frac{1}{nh^{d+d_2}} \sum_{i=1}^{n} \mathbf{1}^{p_x}_{C'_{XV}}(X_j, V_j) \left[\kappa\left(\tfrac{x-X_j}{h}\right)\kappa\left(\tfrac{\widehat{V}_i - V_j}{h}\right) - \kappa\left(\tfrac{x-X_j}{h}\right)\kappa\left(\tfrac{V_i - V_j}{h}\right)\right].
\end{aligned}$$

By the second order Taylor expansion, for $(x, \widehat{V}_i) \in C_{XV}$ and $(x, V_i) \in C_{XV}$

$$\begin{aligned}
& \left|\widetilde{f}(x, \widehat{V}_i) - \widetilde{f}(x, V_i)\right| \\
\leqslant \;& \frac{1}{nh^{d+d_2+1}} \sum_{j=1}^{n} \mathbf{1}^{p_x}_{C'_{XV}}(X_j, V_j) \left|\widehat{V}_i - V_i\right|^{\intercal} \left|\kappa'\left(\tfrac{V_i - V_j}{h}\right)\right| \left|\kappa\left(\tfrac{x-X_j}{h}\right)\right| \\
& + \frac{1}{nh^{d+d_2+2}} \sum_{j=1}^{n} \mathbf{1}^{p_x}_{C'_{XV}}(X_j, V_j) \left|\widehat{V}_i - V_i\right|^{\intercal} \left|\kappa''\left(\overline{V}_j\right)\right| \left|\widehat{V}_i - V_i\right| \left|\kappa\left(\tfrac{x-X_j}{h}\right)\right| \\
\leqslant \;& O\left(\gamma_1 \big/ h\right) \cdot \frac{1}{nh^{d+d_2}} \sum_{j=1}^{n} \iota^{\intercal} \left|\kappa'\left(\tfrac{V_i - V_j}{h}\right)\right| \left|\kappa\left(\tfrac{x-X_j}{h}\right)\right| \\
& + O\left(\gamma_1^2 \big/ h^{2+d_2}\right) \cdot \sup_{v} \iota^{\intercal} \left|\kappa''(v)\right| \iota \cdot \frac{1}{nh^{d}} \sum_{j=1}^{n} \left|\kappa\left(\tfrac{x-X_j}{h}\right)\right| \\
= \;& O\left(\gamma_1 \big/ h\right),
\end{aligned}$$

where $\overline{V}_j$ is between $\left(\frac{V_i - V_j}{h}\right)$ and $\left(\frac{\widehat{V}_i - V_j}{h}\right)$, the second inequality follows from Lemma A.1 and the last equality holds as in the proof of Proposition 3.1.(a). Thus with the condition $\frac{s_1}{(2s_1+d_Z)} - \frac{1+d_2}{(2s+d+d_2)} > 0$, we have

$$\sup_{C_{XV}} \left|\mathbf{1}^{p_x}_{C_{XV}}\left(x, \widehat{V}_i\right) \widetilde{f}(x, \widehat{V}_i) - \mathbf{1}^{p_x}_{C_{XV}}(x, V_i)\, \widetilde{f}(x, V_i)\right| = O\left(\gamma_1 \big/ h\right) \text{ a.s.} \tag{A.3}$$

Combine $(A.2)$ and $(A.3)$, for $(x,\widehat{V}_l)\in C_{XV}$ and $(x,V_l)\in C_{XV}$ with n large enough,

$$\sup_i \left|\mathbf{1}^{p_x}_{C_{XV}}\left(x,\widehat{V}_i\right)\widehat{f}(x,\widehat{V}_i)-\mathbf{1}^{p_x}_{C_{XV}}\left(x,V_i\right)\widetilde{f}(x,V_i)\right| = O\left(\gamma_1 / h\right) \text{ a.s.} \tag{A.4}$$

Similar to the proof above and of Proposition 3.1.(b), for $(x,\widehat{V}_i)\in C_{XV}$ and $(x,V_i)\in C_{XV}$ with n large enough, we have

$$\sup_i \left|\mathbf{1}^{p_x}_{C_{XV}}\left(x,\widehat{V}_i\right)\widehat{q}(x,\widehat{V}_i)-\mathbf{1}^{p_x}_{C_{XV}}\left(x,V_i\right)\widetilde{q}(x,V_i)\right| = O\left(\gamma_1 / h\right) a.s. \tag{A.5}$$

Notice that

$$\begin{aligned}
& \widehat{f}(x,\widehat{V}_i)^{-1}\left\{\left[\mathbf{1}^{p_x}_{C_{XV}}\left(x,\widehat{V}_i\right)\widehat{q}(x,\widehat{V}_i)-\mathbf{1}^{p_x}_{C_{XV}}\left(x,V_i\right)\widetilde{q}(x,V_i)\right]\right. \\
& \left. -m_0(x,v)\left[\mathbf{1}^{p_x}_{C_{XV}}\left(x,\widehat{V}_i\right)\widehat{f}(x,\widehat{V}_i)-\mathbf{1}^{p_x}_{C_{XV}}\left(x,V_i\right)\widetilde{q}(x,V_i)\right]\right\} \\
= \;& \mathbf{1}^{p_x}_{C_{XV}}\left(x,\widehat{V}_i\right)\left[\mathbf{1}^{p_x}_{C_{XV}}\left(x,\widehat{V}_i\right)\widehat{m}(x,\widehat{V}_i)-\mathbf{1}^{p_x}_{C_{XV}}\left(x,V_i\right)\widetilde{m}(x,V_i)\right].
\end{aligned}$$

Because exactly those data points in C_{XV} are used for the estimation,

$$\begin{aligned}
& \left[\mathbf{1}^{p_x}_{C_{XV}}\left(x,\widehat{V}_i\right)\widehat{m}(x,\widehat{V}_i)-\mathbf{1}^{p_x}_{C_{XV}}\left(x,V_i\right)\widetilde{m}(x,V_i)\right] \\
= \;& \mathbf{1}^{p_x}_{C_{XV}}\left(x,\widehat{V}_i\right)\left[\mathbf{1}^{p_x}_{C_{XV}}\left(x,\widehat{V}_i\right)\widehat{m}(x,\widehat{V}_i)-\mathbf{1}^{p_x}_{C_{XV}}\left(x,V_i\right)\widetilde{m}(x,V_i)\right].
\end{aligned}$$

Again, as $\widetilde{m}(x,v)$ is bounded and $\widehat{f}(x,v)$ is bounded away from zero a.s. for $(x,v)\in C_{XV}$, similar to Proposition 3.1.(b), from $(A.4)$ and $(A.5)$ we get

$$\sup_i \left|\mathbf{1}^{p_x}_{C_{XV}}\left(x,\widehat{V}_i\right)\widehat{m}(x,\widehat{V}_i)-\mathbf{1}^{p_x}_{C_{XV}}\left(x,V_i\right)\widetilde{m}(x,V_i)\right| = O\left(\gamma_1 / h\right) \text{ a.s.}$$

So almost surely we have

$$
\begin{aligned}
& \sup_{C_X} |\widehat{g}(x) - \widetilde{g}(x)| \\
= \; & \left| n^{-1} \textstyle\sum_{i=1}^{n} \mathbf{1}_{C_{XV}}^{p_x} \left(x, \widehat{V}_i\right) \widehat{m}(x, \widehat{V}_i) - n^{-1} \textstyle\sum_{i=1}^{n} \mathbf{1}_{C_{XV}}^{p_x} \left(x, V_i\right) \widetilde{m}(x, V_i) \right| \\
\leqslant \; & n^{-1} \sup_i \left| \mathbf{1}_{C_{XV}}^{p_x} \left(x, \widehat{V}_i\right) \widehat{m}(x, \widehat{V}_i) - \mathbf{1}_{C_{XV}}^{p_x} \left(x, V_i\right) \widetilde{m}(x, V_i) \right| \\
= \; & O\left(\gamma_1 \diagup h\right).
\end{aligned}
$$

Thus for $x \in C_X$ and n large enough, $\sup_{C_X} |\widehat{g}(x) - \widetilde{g}(x)| = O\left(\gamma_1 \diagup h\right)$ a.s. $\square$

Appendix B

Aysmptotic Normality

Some useful lemmas are collected here, which are extensions of standard results in the literature, tailored for this paper.[1]

Lemma B.1 (Linearization): For $(x,v) \in C_{XV} \subsetneq S_{XV}$*,*

(i) $\widetilde{m}(x,v) - m(x,v) = \left[\widetilde{q}(x,v) - \widetilde{f}(x,v)\, m(x,v)\right] \big/ f(x,v) + O_p\left(\gamma_2^2\right)$

if $\widetilde{m}(x,v) - m(x,v) = O_p(\gamma_2)$*;*

(ii) $\widehat{m}(x,\widehat{v}) - m(x,v) = \left[\widehat{q}(x,\widehat{v}) - \widehat{f}(x,\widehat{v})\, m(x,v)\right] \big/ f(x,v) + O_p\left(\gamma_m^2\right)$

if $\widehat{m}(x,\widehat{v}) - m(x,v) = O_p(\gamma_m)$*;*

where $\widetilde{m}(\cdot,\cdot)$ *and* $\widehat{m}(\cdot,\cdot)$ *are the kernel estimators of conditional mean as defined in* (3.7) *and* (3.8)*, and the density* $f(x,v)$ *is bounded away from zero.*

For (i), this kind of linearization of kernel estimators has been studied in Ahn and Powell (1993) among others, usually in semiparametric settings with the reminder terms simply represented as $O_p\left(n^{-1/2}\right)$. (ii) extends the linearization to the

[1] In this section, to make notation compact, the subscript 0 to indicate the true underlying function is suppressed. Also suppressed are the subscripts for the density functions, which are distinguished by the arguments. For instance, $f(x)$ is the density for X and $f(z)$ is for Z.

case with preliminary kernel estimators as the conditioning variables. The proof of (ii), however, is basically the same as that of (i). $\widehat{m}(x, \widehat{v})$ is slightly different from $\widehat{m}(x, v)$ as the former is evaluated at $(x, \widehat{v})$.

Lemma B.2 (Projection of U-statistics): The projection of a U-statistic U_n on the basic observation ξ_i is $\widehat{U}_n = \theta_n + \frac{c}{n}\sum_{i=1}^{n}[r_n(\xi_i) - \theta_n]$ where $r_n(\xi_i) \equiv E[P_n(\cdot)|\xi_i]$, $\theta_n \equiv E[r_n(\xi_i)] = E[P_n(\cdot)]$ and $P_n(\cdot)$ is the symmetric kernel of U_n. The constant $c = 2$ for the second-order U-statistics, and $c = 3$ for the second-order U-statistics. If $E\left[\|P_n(\cdot)\|^2\right] = o(n^2\gamma^2)$, then $U_n = \widehat{U}_n + o_p(\gamma)$.

For the projection of second-order U-statistics, Ahn (1995) extends Lemma 3.1 in Powell, Stock and Stoker (1989), relaxing the condition $E\left[\|P_n(\cdot)\|^2\right] = o(n)$ to $E\left[\|P_n(\cdot)\|^2\right] = o(n^2\gamma^2)$ to allow for wider choices of the bandwidth.[2] The extension to the third-order U-statistic is straightforward using the same reasoning in these two papers.

Lemma B.3 (Extended Bochner's Lemma): Suppose that both $m(x)$ and $f(x)$ are functions from $\mathbb{R}^d$ to $\mathbb{R}$ with the s-th order derivatives that are uniformly continuous and bounded within their supports. Also $k(\cdot)$ is a s-th order kernel with bounded support. Using the change of variable $t = \frac{X-x}{h}$, as the bandwidth h goes to zero,

$$\text{(i)} \int \frac{1}{h^d} k\left(\frac{X-x}{h}\right)[m(X) - m(x)]\,dX = \int k(t)[m(x+ht) - m(x)]\,dt$$

$$= h^s k_s m^{(s)}(x) + o(h^s),$$

where $k_s \equiv \frac{1}{s!}\int k(t)t^s dt$ and $m^{(s)}(x)$ is the s-th order derivative of $m(x)$;

[2] Note that the nonparametric convergence rate $1/\gamma$ is slower than the parametric rate $\sqrt{n}$ so that $o(n^2\gamma^2) = o\left(n(\sqrt{n}/(1/\gamma))^2\right) > o(n)$. If $1/\gamma = \sqrt{n}$, the condition $o(n^2\gamma^2) = o(n)$ goes back to the original case of Powell, Stock and Stoker (1989), where they consider the $\sqrt{n}$-consistency of a semiparametric estimator for a single-index model.

(ii) $\int \frac{1}{h^d} k\left(\frac{X-x}{h}\right) f(X)\, dX = \int k(t) f(x+ht)\, dt$

$= f(x) + h^s k_s f^{(s)}(x) + o(h^s)\,;$

(iii) $\int \frac{1}{h^d} k\left(\frac{X-x}{h}\right) f(X)\left[m(X) - m(x)\right] dX$
$= \int k(t) f(x+ht)\left[m(x+ht) - m(x)\right] dt$

$= h^s k_s \left[(m\cdot f)^{(s)}(x) - m(x) f^{(s)}(x)\right] + o(h^s),$

where $(mf)^{(s)}(x)$ *is the* s*-th order derivative of* $(m(x) f(x))$*.*

Bochner's Lemma is extensively applied in the literature of kernel estimation, usually with $O(h^s)$ to denote the reminder term. Lemma B.3 is a special case of Bochner's Lemma for differentiable functions, where the *s-th* order Taylor expansion is used to derive the explicit expression of the limits.

Lemma B.4: Suppose both $m(x)$ *and* $f(x)$ *be functions from* $\mathbb{R}^d$ *to* $\mathbb{R}$ *with up to the* s*-th order derivatives that are uniformly continuous and bounded within their supports. Additionally,* $f(x) = 0$ *for* x *at the boundary of the support of* $f(x)$*. Also* $k(\cdot)$ *is a* s*-th order kernel with bounded support. As the bandwidth* h *goes to zero,*

(i) $.\frac{1}{h}\int \frac{1}{h^d} k'\left(\frac{x-X}{h}\right) f(X)\, dX = f'(x) + O(h^s)\,;$

(ii) $.\frac{1}{h}\int \frac{1}{h^d} k'\left(\frac{x-X}{h}\right) f(X)\, m(X)\, dX = f'(x)\, m(x) + f(x)\, m'(x) + O(h^s)\,.$

Proof: Using the change of variables $\frac{X-x}{h}=h$

$$\begin{aligned}
\frac{1}{h}\int\frac{1}{h^d}k'\left(\tfrac{x-X}{h}\right)f\left(X\right)dX &= \frac{1}{h}\int k'\left(-t\right)f\left(x+ht\right)dt \\
&= -\int f\left(x+ht\right)dk\left(t\right) \\
&= -f\left(x+ht\right)dk\left(t\right)|_{\underline{t}}^{\overline{t}}+\int f'\left(x+ht\right)k\left(t\right)dt \\
&= f'\left(x\right)+O\left(h^s\right),
\end{aligned}$$

where $f\left(x+ht\right)k\left(t\right)|_{\underline{t}}^{\overline{t}}=f\left(X\right)k\left(\frac{x-X}{h}\right)|_{\underline{X}}^{\overline{X}}=0$ as $f\left(x\right)=0$ at the boundary. The last equality follows from Lemma B.3.(ii), where $f'\left(x\right)$ admits derivatives up to (s-1)-th order only so that the reminder terms are all zero due to the s-th order kernel. The proof of (ii) is similar to that of (i). $\square$

Linearization of Kernel Estimators with Preliminary Estimates

By linearization, the stochastic denominator problem in kernel regressions is avoided and the estimators can be expressed as U-statistics more easily. By Lemma B.1, the linearization of $\widehat{m}(x,v)$ gives

$$\widehat{m}(x,v)=\frac{1}{f\left(x,v\right)}\left[\widehat{q}\left(x,v\right)-\widehat{f}\left(x,v\right)m\left(x,v\right)\right]+m\left(x,v\right)+o_p(\gamma).$$

To introduce the trimming into the linearization of $\widehat{m}(x,\widehat{V}_i)$, note that $\widehat{V}_i$ converges to V_i uniformly within C_{XV} so that $1_{C_{XV}}\left(x,\widehat{V}_i\right)=1_{C'_{XV}}\left(x,V_i\right)$ and $1_{C'_{XV}}\left(x,\widehat{V}_j\right)=1_{C''_{XV}}\left(x,V_j\right)$ where $C_{XV}\subsetneq C'_{XV}\subsetneq C''_{XV}$. Let $a_i\equiv 1^{p_x}_{C_{XV}}\left(x,\widehat{V}_i\right)$, $a'_i\equiv 1^{p_x}_{C'_{XV}}\left(x,V_i\right)$ and $a''_i\equiv 1^{p_x}_{C''_{XV}}\left(x,V_i\right)$ and let $a'_j\equiv 1^{p_x}_{C'_{XV}}\left(X_j,V_j\right)$, $a''_j\equiv 1^{p_x}_{C''_{XV}}\left(X_j,V_j\right)$. Therefore,

$$a_i\widehat{m}(x,\widehat{V}_i)=\frac{1}{n}\sum_{j=1}^{n}K_h\left(x-X_j\right)K_h\left(\widehat{V}_i-\widehat{V}_j\right)a'_ia''_j\frac{\left(Y_j-m\left(x,V_i\right)\right)}{f\left(x,V_i\right)}+a_im(x,V_i)+o_p(\gamma).$$

Thus $\widehat{g}(x)-g\left(x\right)$ can be written as

$$\begin{aligned}
\widehat{g}(x)-g\left(x\right) &= n^{-1}\textstyle\sum_{i=1}^{n}a_i\widehat{m}(x,\widehat{V}_i) \\
&= \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}K_h\left(x-X_j\right)K_h\left(\widehat{V}_i-\widehat{V}_j\right)a_i'a_j''\frac{\left(Y_j-m\left(x,V_i\right)\right)}{f\left(x,V_i\right)} \\
&\quad+\left[\frac{1}{n}\sum_{i=1}^{n}a_i'm\left(x,V_i\right)-g\left(x\right)\right]+o_p(\gamma) \\
&= \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}K_h\left(x-X_j\right)K_h\left(\widehat{V}_i-\widehat{V}_j\right)a_i'a_j''\frac{\left(Y_j-m\left(x,V_i\right)\right)}{f\left(x,V_i\right)}+o_p(\gamma),
\end{aligned}$$

where the third equality follows from the Central Limit Theorem. A Taylor expansion of $K_h\left(\widehat{V}_i-\widehat{V}_j\right)$ around $K_h\left(V_i-V_j\right)$ yields

$$\begin{aligned}
K_h\left(\widehat{V}_i-\widehat{V}_j\right)-K_h\left(V_i-V_j\right) &= \frac{1}{h^{d_2+1}}\kappa_h'\left(\tfrac{V_i-V_j}{h}\right)\left(\left(\widehat{V}_i-V_i\right)-\left(\widehat{V}_j-V_j\right)\right) \\
&\quad+\frac{1}{h^{d_2+2}}\kappa_h''\left(\overline{V}\right)\left(\left(\widehat{V}_i-V_i\right)-\left(\widehat{V}_j-V_j\right)\right)^2,
\end{aligned}$$

where $\overline{V}$ is between $\left(\frac{\widehat{V}_i-\widehat{V}_j}{h}\right)$ and $\left(\frac{V_i-V_j}{h}\right)$. Since the second-order term is asymptotically negligible, we only need to consider the first-order term of the expansion.[3] Note that for $\widehat{V}_i\neq\infty$, $\widehat{V}_i-V_i=r\left(Z_i\right)-\widetilde{r}\left(Z_i\right)$ and the linearization of $\widetilde{r}\left(Z_i\right)$ yields

$$\widetilde{r}\left(Z_i\right)-r\left(Z_i\right)=\frac{1}{n}\sum_{l=1}^{n}K_h\left(Z_i-Z_l\right)\frac{\left(r\left(Z_i\right)-X_{2l}\right)}{f\left(Z_i\right)}.$$

[3] To see this, note that
$E\left\|\left(\widehat{V}_i-V_i\right)-\left(\widehat{V_j}-V_j\right)\right\|^2\leqslant E\left\|\widehat{V}_i-V_i\right\|^2+E\left\|\widehat{V_j}-V_j\right\|^2=2O\left(\gamma_1^2\right)=o\left(\gamma_1\right).$

Therefore $\widehat{g}(x) - g(x)$ can be rewritten as

$$
\begin{aligned}
& \widehat{g}(x) - g(x) \\
= & \left(\widehat{g}(x) - \widetilde{g}(x)\right) + \left(\widetilde{g}(x) - g(x)\right) \qquad \text{(B.1)} \\
= & \binom{n}{2}^{-1} \frac{1}{2} \sum_{i=1}^{n} \sum_{j \neq i}^{n} K_{h_x}(x - X_j) K_h(V_i - V_j) a_i' a_j'' \frac{(Y_j - m(x, V_i))}{f(x, V_i)} \\
& + \binom{n}{3}^{-1} \frac{1}{6} \sum_{i=1}^{n} \sum_{j \neq i} \sum_{l \neq i, l \neq j} K_h(x - X_j) \frac{1}{h} K_h'(V_i - V_j) a_i' a_j'' \tfrac{(Y_j - m(x, V_i))}{f(x, V_i)} K_h(Z_i - Z_l) \tfrac{(r(Z_i) - X_{2l})}{f(Z_i)} \\
& + \binom{n}{3}^{-1} \frac{1}{6} \sum_{i=1}^{n} \sum_{j \neq i} \sum_{l \neq i, l \neq j} K_h(x - X_j) \frac{1}{h} K_h'(V_i - V_j) a_i' a_j'' \tfrac{(Y_j - m(x, V_i))}{f(x, V_i)} K_h(Z_j - Z_l) \tfrac{(r(Z_j) - X_{2l})}{f(Z_j)} \\
& + o_p(\gamma),
\end{aligned}
$$

where the second equality follows from the fact that the terms with $i = j$, $j = l$, or $l = i$ are asymptotically negligible.[4] To derive the asymptotic normality of $\widehat{g}(x) - g(x)$, I study the asymptotic properties of $\left(\widetilde{g}(x) - g(x)\right)$ (the first term of $(B.1)$) in Proposition 3.2, and those of $\left(\widehat{g}(x) - \widetilde{g}(x)\right)$ (the second and third terms of $(B.1)$) in Theorem 3.2.

Proof of Proposition 3.2: Note that we can express $\widetilde{g}(x) - g(x)$ as a U-statistic:

$$
\widetilde{g}(x) - g(x) = \binom{n}{2}^{-1} \sum_{i=1}^{n} \sum_{j > i} P_n\left(\xi_i, \xi_j\right) + o_p(\gamma) \equiv U_n + o_p(\gamma),
$$

where $P_n\left(\xi_j, \xi_j\right)$ is the kernel of the U-statistic U_n and

$$
\begin{aligned}
P_n\left(\xi_i, \xi_j\right) \equiv & \frac{1}{2} K_h(x - X_j) K_h(V_i - V_j)(Y_j - m(x, V_i)) a_i' a_j'' / f(x, V_i) \\
& + K_h(x - X_i) K_h(V_j - V_i)(Y_i - m(x, V_j)) a_i' a_j'' / f(x, V_j) \\
\equiv & \frac{1}{2}\left[P_{n1}\left(\xi_i, \xi_j\right) + P_{n2}\left(\xi_i, \xi_j\right)\right].
\end{aligned}
$$

[4]See a similar result in the proof of Theorem 3 in Ahn and Powell (1993).

Next, I project U_n onto the basic observation ξ_i. The projection of U_n is

$$\widehat{U}_n = \theta_n + \frac{2}{n}\sum_{i=1}^{n}\left[r_n\left(\xi_i\right) - \theta_n\right],$$

where $r_n\left(\xi_i\right) \equiv E\left[P_n\left(\xi_i, \xi_j\right)|\xi_i\right]$ and $\theta_n \equiv E\left[r_n\left(\xi_i\right)\right] = E\left[P_n\left(\xi_i, \xi_j\right)\right]$. By Lemma B.2, if $E\left[\left\|P_n\left(\xi_i, \xi_j\right)\right\|^2\right] = o\left(n^2\gamma^2\right)$, then $U_n = \widehat{U}_n + o_p(\gamma)$, where $\widehat{U}_n$ is the projection of U_n. It is easy to show that $E\left[\left\|P_n\left(\xi_i, \xi_j\right)\right\|^2\right] = O\left(1 / h^{d+d_2}\right) = o\left(n^2\gamma^2\right)$ if and only if $\left(nh^d\gamma\right)\left(nh^{d_2}\gamma\right) \longrightarrow \infty$, which is implied by Assumption B.

Consider the projection of $P_{n1}\left(\xi_i, \xi_j\right)$ on ξ_i. Let $\frac{x_j - x}{h} = t_1$ and $\frac{v_j - V_i}{h} = t_2$,

$$\begin{aligned}
& E\left[P_{n1}\left(\xi_i, \xi_j\right)|\xi_i\right] \\
= & \int K_h\left(x - x_j\right) K_h\left(V_i - v_j\right) a_i' a_j'' \left[m\left(x_j, v_j\right) - m\left(x, V_i\right)\right] \frac{f\left(x_j, v_j\right)}{f\left(x, V_i\right)} dx_j dv_j \\
= & \frac{a_i' a_i''}{f\left(x, V_i\right)} \int \kappa\left(-t_1\right) \kappa\left(-t_2\right) a_j \left[m\left(x + ht_1, V_i + ht_2\right) - m\left(x, V_i\right)\right] \\
& \qquad a_j'' f\left(x + ht_1, V_i + ht_2\right) dt_1 dt_2 \\
= & h^s \frac{k_s a_i'}{f\left(x, V_i\right)} \left[(mf)^{(s)}\left(x, V_i\right) - m\left(x, V_i\right) f^{(s)}\left(x, V_i\right)\right] + o_p\left(h^s\right) \\
\equiv & h^s a_i' B_1\left(x, V_i\right) + o_p\left(h^s\right),
\end{aligned}$$

where $a_i' a_i'' = a_i'$ and the third equality follows from Lemma B.3.[5]

[5] Note that by Bochner's lemma,
$\int \kappa\left(t_1\right)\kappa\left(t_2\right) 1_{C''_{XV}}\left(x + h_x t_1, V_i + h_v t_2\right) dt_1 dt_2$
$= 1_{C''_{XV}}\left(x, V_i\right) \int \kappa\left(t_1\right)\kappa\left(t_2\right) dt_1 dt_2 = 1_{C''_{XV}}\left(x, V_i\right) = a_i''$.
Also note that $a_i' a_i'' \equiv \mathbf{1}_{C'_{XV}}\left(x, V_i\right) \cdot \mathbf{1}_{C''_{XV}}\left(x, V_i\right) = \mathbf{1}_{C'_{XV}}\left(x, V_i\right) = a_i'$.

For the projection of $P_{n2}\left(\xi_i,\xi_j\right)$ on ξ_i, let $\frac{v_j-V_i}{h}=t$,

$$
\begin{aligned}
& E\left[P_{n2}\left(\xi_i,\xi_j\right)|\xi_i\right] \\
= \; & K_h\left(x-X_i\right)a_i'\int K_h\left(v_j-V_i\right)\left(Y_i-m\left(x,v_j\right)\right)a_j''\frac{f\left(v_j\right)}{f\left(x,v_j\right)}dv_j \\
= \; & K_h\left(x-X_i\right)a_i'a_i''\int \kappa\left(t\right)\left(Y_i-m\left(x,V_i+ht\right)\right)\frac{f\left(V_i+ht\right)}{f\left(x,V_i+ht\right)}dt \\
= \; & K_h\left(x-X_i\right)a_i'\left(Y_i-m\left(x,V_i\right)\right)\int \kappa\left(t\right)\tfrac{f\left(V_i+ht\right)}{f\left(x,V_i+ht\right)}dt \\
& -K_h\left(x-X_i\right)a_i'\int \kappa\left(t\right)\left[m\left(x,V_i+ht\right)-m\left(x,V_i\right)\right]\frac{f\left(V_i+ht\right)}{f\left(x,V_i+ht\right)}dt \\
= \; & K_h\left(x-X_i\right)\left(Y_i-m\left(x,V_i\right)\right)\frac{a_i'f\left(V_i\right)}{f\left(x,V_i\right)} \\
& +h^s k_s K_h\left(x-X_i\right)\left(Y_i-m\left(x,V_i\right)\right)a_i'\left(\frac{f_V}{f_{XV}}\right)_v^{(s)}\left(x,V_i\right) \\
& -h^s k_s K_h\left(x-X_i\right)a_i'\left[\left(m\frac{f_V}{f_{XV}}\right)_v^{(s)}\left(x,V_i\right)-m\left(x,V_i\right)\cdot\left(\frac{f_V}{f_{XV}}\right)_v^{(s)}\left(x,V_i\right)\right]
\end{aligned}
$$

where $a_i'a_i''=a_i'$ and the fourth equality follows from Lemma B.3.[6]

Now consider θ_n, which is actually the leading term of the bias of $(\widetilde{g}(x)-g\left(x\right))$.

$$
\begin{aligned}
\theta_n \equiv \; & E\left[r_n\left(\xi_i\right)\right]=E\left[P_n\left(\xi_i,\xi_j\right)\right] \qquad \text{(B.2)} \\
= \; & \frac{1}{2}h^s k_s\int\left[\left(m\left(x,v\right)f\left(x,v\right)\right)_x^{(s)}-m\left(x,v\right)f_x^{(s)}\left(x,v\right)\right]\frac{a_i'f\left(v\right)}{f\left(x,v\right)}dv \\
& +\frac{1}{2}h^s\int a_i'\left[\left(m\frac{f_V}{f_{XV}}\right)_v^{(s)}\left(x,v\right)-m\left(x,v\right)\left(\frac{f_V}{f_{XV}}\right)_v^{(s)}\left(x,v\right)\right]f\left(x,v\right)dv \\
& +\frac{1}{2}h^s\int a_i'B_1\left(x,v\right)f\left(v\right)dv+o\left(h^s\right) \\
\equiv \; & h^s\frac{1}{2}\left[B_0\left(x\right)+B_1\left(x\right)+B_2\left(x\right)\right]+o\left(h^s\right) \\
\equiv \; & h^s B_g\left(x\right)+o\left(h^s\right).
\end{aligned}
$$

[6] Here $\left(f_V\diagup f_{XV}\right)_v^{(s)}\left(x,V_i\right)$ is the s-th order partial derivative of $\left(f\left(v\right)\diagup f\left(x,v\right)\right)$ w.r.t. v evaluated at $\left(x,V_i\right)$, and $\left(mf_V\diagup f_{XV}\right)_v^{(s)}\left(x,V_i\right)$ is the s-th order partial derivative of $\left(m\left(x,v\right)f\left(v\right)\diagup f\left(x,v\right)\right)$ w.r.t. v evaluated at $\left(x,V_i\right)$.

Note that the bias is at the order of $O\left(h^{s}\right)$ as $B_{g}\left(x\right)$ is bounded for $x\in C_{X}$.

Put together,

$$\begin{aligned}\widehat{U}_{n}-\theta_{n} &= \frac{1}{n}\sum_{i=1}^{n}K_{h}\left(x-X_{i}\right)\left(Y_{i}-m\left(x,V_{i}\right)\right)\frac{a_{i}^{\prime}f\left(V_{i}\right)}{f\left(x,V_{i}\right)}+o\left(h^{s}\right)\text{ and}\\ \widetilde{g}(x)-g\left(x\right) &= \left(\widehat{U}_{n}-\theta_{n}\right)+\left(\theta_{n}-0\right)\\ &= \frac{1}{n}\sum_{i=1}^{n}K_{h}\left(x-X_{i}\right)\left(Y_{i}-m\left(x,V_{i}\right)\right)\frac{a_{i}^{\prime}f\left(V_{i}\right)}{f\left(x,V_{i}\right)}+O_{p}\left(h^{s}\right),\end{aligned}$$

By the Liapunov's Central Limit Theorem,

$$\frac{1}{n}\sum_{i=1}^{n}K_{h}\left(x-X_{i}\right)\left(Y_{i}-m\left(x,V_{i}\right)\right)\frac{a_{i}f\left(V_{i}\right)}{f\left(x,V_{i}\right)}\xrightarrow{d}N\left(B_{0}\left(x\right),V_{g}\left(x\right)\right),\text{ where}$$

$$V_{g}\left(x\right)=\frac{1}{nh^{d}}\int Var\left(Y|x,v\right)\frac{\left(a_{i}^{\prime}\right)^{2}f^{2}(v)}{f(x,v)}dv\cdot\int\kappa^{2}\left(t\right)dt.$$

Therefore the variance is at the order of $O\left(\frac{1}{nh^{d}}\right)$.

Collecting the results above, we have

$$\widetilde{g}\left(x\right)-g\left(x\right)=O_{p}\left(h^{s}+\tfrac{1}{\sqrt{nh^{d}}}\right). \tag{B.3}$$

This shows that the optimal rate of (pointwise) convergence of $\widetilde{g}\left(x\right)$ to $g\left(x\right)$ is achieved when h^{s} has exactly the order of $\frac{1}{\sqrt{nh^{d}}}$. That is, when h assumes the optimal bandwidth of the exact order $n^{-1/(2s+d)}$, $\widetilde{g}\left(x\right)$ obtains the optimal rate $n^{s/(2s+d)}$. To establish Proposition 3.2.(i), it suffices to multiply $(B.3)$ by $\sqrt{nh^{d}}$ and to take the limit as $n\to\infty$. Note that the bias is $B_{g}\left(x\right)=B_{0}\left(x\right)+B_{1}\left(x\right)+B_{2}\left(x\right)$, not just $B_{0}\left(x\right)$. This is because $\widetilde{g}\left(x\right)$ is estimated by averaging $\widetilde{m}\left(x,V_{i}\right)$'s over V_{i}, which introduces additional biases. To prove Proposition 3.2.(ii), it suffices to divide $(B.3)$ by h^{s} and to take the limit as $n\to\infty$. $\square$

Proposition B.1. *For* $x \in C_X$, $(\widehat{g}(x) - \widetilde{g}(x)) = O_p(\gamma_1)$.

Proof: The asymptotic properties of $(\widetilde{g}(x) - g(x))$ is derived in Proposition 3.2 and it remains to study that of $\widehat{g}(x) - \widetilde{g}(x)$. Express $\widehat{g}(x) - \widetilde{g}(x)$ as a third-order U-statistic:

$$
\begin{aligned}
& \widehat{g}(x) - \widetilde{g}(x) \\
= & \binom{n}{3}^{-1} \sum_{i=1}^{n} \sum_{j \neq i} \sum_{l \neq i, l \neq j} \frac{1}{6} K_h(x - X_j) \frac{1}{h} K_h'(V_i - V_j) a_i' a_j'' \tfrac{(Y_j - m(x,V_i))}{f(x,V_i)} K_h(Z_i - Z_l) \tfrac{(r(Z_i) - X_{2l})}{f(Z_i)} \\
& + \binom{n}{3}^{-1} \sum_{i=1}^{n} \sum_{j \neq i} \sum_{l \neq i, l \neq j} \frac{1}{6} K_h(x - X_j) \frac{1}{h} K_h'(V_i - V_j) a_i' a_j'' \tfrac{(Y_j - m(x,V_i))}{f(x,V_i)} K_h(Z_j - Z_l) \tfrac{(r(Z_j) - X_{2l})}{f(Z_j)} \\
& + o_p(\gamma) \\
\equiv & \binom{n}{3}^{-1} \sum_{i=1}^{n} \sum_{j > i} \sum_{l > j} P_{n1}(\xi_i, \xi_j, \xi_l) + \binom{n}{3}^{-1} \sum_{i=1}^{n} \sum_{j > i} \sum_{l > j} P_{n2}(\xi_i, \xi_j, \xi_l) + o_p(\gamma) \\
\equiv & U_{n1} + U_{n2} + o_p(\gamma)
\end{aligned} \tag{B.4}
$$

where both U_{n1} and U_{n2} are third-order U-statistics with the kernels $P_{n1}(\xi_i, \xi_j, \xi_l)$ and $P_{n2}(\xi_i, \xi_j, \xi_l)$ respectively.

Lemma B.5 and B.6 show that $U_{n1} = O_p(h_1^{s_1})$ and $U_{n2} = O_p(h_1^{s_1})$ respectively. Unless we oversmooth in Step 1, $O_p(h_1^{s_1}) \leqslant O_p(\gamma_1)$. Hence, for $x \in C_X$

$$
(\widehat{g}(x) - \widetilde{g}(x)) = O_p(\gamma_1). \qquad \square
$$

Lemma B.5: *For the third order U-statistic* U_{n1} *defined in* $(B.4)$, $U_{n1} = O_p(h_1^{s_1})$.

Proof: $U_{n1} = \binom{n}{3}^{-1} \sum_{i=1}^{n} \sum_{j>i} \sum_{l>j} P_{n1}(\xi_i, \xi_j, \xi_l)$, where $P_{n1}(\xi_i, \xi_j, \xi_l)$ is the kernel of U_{n1} and

$$
\begin{aligned}
P_{n1}(\xi_i, \xi_j, \xi_l) = & \frac{1}{6}[p_{n1}(\xi_i, \xi_j, \xi_l) + p_{n1}(\xi_i, \xi_l, \xi_j) + p_{n1}(\xi_j, \xi_i, \xi_l) \\
& + p_{n1}(\xi_l, \xi_i, \xi_j) + p_{n1}(\xi_j, \xi_l, \xi_i) + p_{n1}(\xi_l, \xi_j, \xi_i)],
\end{aligned}
$$

with $p_{n1}(\cdot,\cdot,\cdot)$'s to be defined below.

The asymptotic behavior of U_{n1} is studied by the projection $\widehat{U}_{n1}$ of U_{n1} onto the basic observations ξ_i's.

$$\widehat{U}_{n1} = \theta_{n1} + \frac{6}{n}\sum_{i=1}^{n}\left[r_{n1}(\xi_i) - \theta_{n1}\right],$$

where $r_{n1}(\xi_i) \equiv E\left[P_{n1}\left(\xi_i,\xi_j,\xi_l\right)|\xi_i\right]$ and $\theta_{n1} \equiv E\left[r_{n1}(\xi_i)\right] = E\left[P_{n1}\left(\xi_i,\xi_j,\xi_l\right)\right]$. By Lemma B.2, if $E\left[\left\|P_{n1}\left(\xi_i,\xi_j,\xi_l\right)\right\|^2\right] = o\left(n^2\gamma^2\right)$, then $U_{n1} = \widehat{U}_{n1} + o_p(\gamma)$. It can be shown that $E\left[\left\|P_{n1}\left(\xi_i,\xi_j,\xi_l\right)\right\|^2\right] = O\left(\frac{1}{h^{d+d_2+1}}\frac{1}{h_1^{d_z}}\right) = o\left(n^2\gamma^2\right)$ if and only if $n^2\gamma^2h^{d+d_2+1}h_1^{d_z} \longrightarrow \infty$, which is implied by Assumption B.

One by one, I examine the projection of six components of $P_{n1}\left(\xi_i,\xi_j,\xi_l\right)$ on ξ_i. Since U_{n1} is a third-order U-statistic, it is difficult to project U_{n1} on ξ_i directly. I do it in a sequential way, where the techniques are similar to but more involved than those used the proof of Proposition 3.2.

For $p_{n1}\left(\xi_i,\xi_j,\xi_l\right) = K_h\left(x - X_j\right)\frac{1}{h}K_h'\left(V_i - V_j\right)\frac{(Y_j - m(x,V_i))a_i'a_j''}{f(x,V_i)}K_h\left(Z_i - Z_l\right)\frac{(r(Z_i) - X_{2l})}{f(Z_i)}$:

Sequential projection on decreasing sets of conditioning variables yields

$$E\left[p_{n1}\left(\xi_i,\xi_j,\xi_l\right)|\xi_i,\xi_j\right] = K_h\left(x - X_j\right)\frac{1}{h}K_h'\left(V_i - V_j\right)\frac{(Y_j - m(x,V_i))a_i'a_j''}{f(x,V_i)f(Z_i)}\cdot h_1^{s_1}R_1\left(Z_i\right),$$

where $R_1\left(Z_i\right) = \left[(rf)^{(s_1)}\left(Z_i\right) - r\left(Z_i\right)f^{(s_1)}\left(Z_i\right)\right]\int\kappa(t)\,t^{s_1}dt$;

$$\begin{aligned} & E\left[p_{n1}\left(\xi_i,\xi_j,\xi_l\right)|\xi_i,V_j\right] \\ = \; & \frac{1}{h}K_h'\left(V_i - V_j\right)\frac{h_1^{s_1}R_1(Z_i)a_i'a_j''}{f(x,V_i)f(Z_i)}\left[f(x)\left(m\left(x,V_j\right) - m\left(x,V_i\right)\right) + O_p\left(h^s\right)\right]; \end{aligned}$$

$$E\left[p_{n1}\left(\xi_i,\xi_j,\xi_l\right)|\xi_i\right] = h_1^{s_1}\frac{f(x)\,f(V_i)\,a_i'}{f(x,V_i)\,f(Z_i)}m_v'\left(x,V_i\right) + o_p\left(h_1^{s_1}\right).$$

Similarly, for

$$p_{n1}\left(\xi_i,\xi_l,\xi_j\right)=K_h\left(x-X_l\right)\tfrac{1}{h}K_h'\left(V_i-V_l\right)\tfrac{\left(Y_l-m(x,V_i)\right)a_i'a_j''}{f(x,V_i)}K_h\left(Z_i-Z_j\right)\tfrac{\left(r(Z_i)-X_{2j}\right)}{f(Z_i)},$$

$$E\left[p_{n1}\left(\xi_i,\xi_l,\xi_j\right)|\xi_i\right]=h_1^{s_1}\tfrac{f(x)f(V_i)a_i'}{f(x,V_i)f(Z_i)}m_v'\left(x,V_i\right)+o_p\left(h_1^{s_1}\right).$$

Note that $E\left[p_{n1}\left(\xi_i,\cdot,\cdot\right)\right]=O_p\left(h_1^{s_1}\right)$ as $\frac{f(x)f(v)a_i'}{f(x,v)}m_v'\left(x,v\right)$ is bounded within C_{XV}'.[7]

For $p_{n1}\left(\xi_j,\xi_i,\xi_l\right)=K_h\left(x-X_i\right)\frac{1}{h}K_h'\left(V_j-V_i\right)\frac{\left(Y_i-m(x,V_j)\right)a_i'a_j''}{f(x,V_j)}K_h\left(Z_j-Z_l\right)\frac{\left(r(Z_j)-X_{2l}\right)}{f(Z_j)}$:

Sequential projection on decreasing sets of conditioning variables yields

$$E\left[p_{n1}\left(\xi_j,\xi_i,\xi_l\right)|\xi_i,\xi_j\right]=K_h\left(x-X_i\right)\frac{1}{h}K_h'\left(V_j-V_i\right)\tfrac{\left(Y_i-m(x,V_j)\right)a_i'a_j''}{f(x,V_j)f(Z_j)}\cdot h_1^{s_1}R_1\left(Z_j\right);$$

$$E\left[p_{n1}\left(\xi_j,\xi_i,\xi_l\right)|\xi_i,V_j\right]=K_h\left(x-X_i\right)\frac{1}{h}K_h'\left(V_j-V_i\right)\tfrac{\left(Y_i-m(x,V_j)\right)a_i'a_j''}{f(x,V_j)f(Z_j)}h_1^{s_1}\cdot\overline{R}_1;$$

where $\overline{R}_1=\int R_1\left(z\right)dz$;

$$\begin{aligned}&E\left[p_{n1}\left(\xi_j,\xi_i,\xi_l\right)|\xi_i\right]\\=\;&h_1^{s_1}K_h\left(x-X_i\right)\left[\left(Y_i-m\left(x,V_i\right)\right)\left(\tfrac{f_V}{f_{XV}}\right)_v'\left(x,V_i\right)-m_v'\left(x,V_i\right)\tfrac{f(V_i)}{f(x,V_i)}\right]a_i'\overline{R}_1+o_p\left(h_1^{s_1}\right).\end{aligned}$$

Similarly, for

$$p_{n1}\left(\xi_l,\xi_i,\xi_j\right)=K_h\left(x-X_i\right)\tfrac{1}{h}K_h'\left(V_l-V_i\right)\tfrac{\left(Y_i-m(x,V_l)\right)a_i'a_j''}{f(x,V_l)}K_h\left(Z_l-Z_j\right)\tfrac{\left(r(Z_l)-X_j\right)}{f(Z_l)},$$

$$\begin{aligned}&E\left[p_{n1}\left(\xi_l,\xi_i,\xi_j\right)|\xi_i\right]\\=\;&h_1^{s_1}K_h\left(x-X_i\right)\left[\left(Y_i-m\left(x,V_i\right)\right)\left(\tfrac{f_V}{f_{XV}}\right)_v'\left(x,V_i\right)-m_v'\left(x,V_i\right)\tfrac{f(V_i)}{f(x,V_i)}\right]a_i'\overline{R}_1+o_p\left(h_1^{s_1}\right).\end{aligned}$$

Note that $E\left[p_{n1}\left(\cdot,\xi_i,\cdot\right)\right]=O_p\left(h_1^{s_1}\right)$ as $m_v'\left(x,v\right)\frac{f(v)a_i'}{f(x,v)}\overline{R}_1$ is bounded within C_{XV}'.

[7] Note that $a_i'=\mathbf{1}\left(\left(x,v\right)\in C_{XV}'\right)$ but the notation a_i' is still used.

For $p_{n1}\left(\xi_j,\xi_l,\xi_i\right)=K_h\left(x-X_l\right)\frac{1}{h}K_h'\left(V_j-V_l\right)\frac{(Y_l-m(x,V_j))a_i'a_j''}{f(x,V_j)}K_h\left(Z_j-Z_i\right)\frac{(r(Z_j)-X_{2i})}{f(Z_j)}$:

Sequential projection on decreasing sets of conditioning variables yields

$$\begin{aligned}&E\left[p_{n1}\left(\xi_j,\xi_l,\xi_i\right)|\xi_i,\xi_j,V_l\right]\\=\;&\frac{1}{h}K_h'\left(V_j-V_l\right)a_i'a_j''\tfrac{f(x)[m(x,V_l)-m(x,V_j)]+O_p(h^s)}{f(x,V_j)}K_h\left(Z_j-Z_i\right)\tfrac{(r(Z_j)-X_{2i})}{f(Z_j)};\end{aligned}$$

$$\begin{aligned}&E\left[p_{n1}\left(\xi_j,\xi_l,\xi_i\right)|\xi_i,V_j,V_l\right]\\=\;&\frac{1}{h}K_h'\left(V_j-V_l\right)a_i'a_j''\tfrac{f(x)[m(x,V_l)-m(x,V_j)]+O_p(h^s)}{f(x,V_j)}\left[\left(r\left(Z_i\right)-X_{2i}\right)+O_p\left(h_1^{s_1}\right)\right];\end{aligned}$$

$$E\left[p_{n1}\left(\xi_j,\xi_l,\xi_i\right)|\xi_i,V_j\right]=\frac{f\left(x\right)a_i'a_j''}{f\left(x,V_j\right)}\left[f\left(V_j\right)m_v'\left(x,V_j\right)\right]\left[\left(r\left(Z_i\right)-X_{2i}\right)\right]+O_p\left(h_1^{s_1}\right);$$

$$E\left[p_{n1}\left(\xi_j,\xi_l,\xi_i\right)|\xi_i\right]=f\left(x\right)\left(r\left(Z_i\right)-X_{2i}\right)\int\tfrac{f^2(v)m_v'(x,v)a_i'}{f(x,v)}dv+O_p\left(h_1^{s_1}\right).$$

Similarly, for

$$p_{n1}\left(\xi_l,\xi_j,\xi_i\right)=K_h\left(x-X_j\right)\tfrac{1}{h}K_h'\left(V_l-V_j\right)\tfrac{(Y_j-m(x,V_l))a_i'a_j''}{f(x,V_l)}K_h\left(Z_l-Z_i\right)\tfrac{(r(Z_l)-X_{2i})}{f(Z_l)},$$

$$E\left[p_{n1}\left(\xi_l,\xi_j,\xi_i\right)|\xi_i\right]=f\left(x\right)\left(r\left(Z_i\right)-X_{2i}\right)\int\tfrac{f^2(v)m_v'(x,v)a_i'}{f(x,v)}dv+O_p\left(h_1^{s_1}\right).$$

Note that $E\left[p_{n1}\left(\cdot,\cdot,\xi_i\right)\right]=O_p\left(h_1^{s_1}\right)$ as $E\left[r\left(Z_i\right)-X_{2i}\right]=0$.

Put together, $\left(\widehat{U}_{n1}-\theta_{n1}\right)=\frac{6}{n}\sum_{i=1}^{n}\left[r_{n1}\left(\xi_i\right)-\theta_{n1}\right]=O_p\left(h_1^{s_1}\right)$ and $\left(\theta_{n1}-0\right)=O_p\left(h_1^{s_1}\right)$ so that $\widehat{U}_{n1}=O_p\left(h_1^{s_1}\right)$. Therefore $U_{n1}=\widehat{U}_{n1}+o_p\left(\gamma\right)=O_p\left(h_1^{s_1}\right)$. $\square$

Lemma B.6: *For the third order U-statistic U_{n2} defined in (B.4), $U_{n2}=O_p\left(h_1^{s_1}\right)$.*

Proof: $U_{n2}=\binom{n}{3}^{-1}\sum_{i=1}^{n}\sum_{j>i}\sum_{l>j}P_{n2}\left(\xi_i,\xi_j,\xi_l\right)$, where $P_{n2}\left(\xi_i,\xi_j,\xi_l\right)$ is the kernel of

U_{n2} and

$$P_{n2}\left(\xi_i,\xi_j,\xi_l\right) = \frac{1}{6}[p_{n2}\left(\xi_i,\xi_j,\xi_l\right)+p_{n2}\left(\xi_i,\xi_l,\xi_j\right)+p_{n2}\left(\xi_j,\xi_i,\xi_l\right)$$
$$+p_{n2}\left(\xi_l,\xi_i,\xi_j\right)+p_{n2}\left(\xi_j,\xi_l,\xi_i\right)+p_{n2}\left(\xi_l,\xi_j,\xi_i\right)].$$

with $p_{n2}\left(\cdot,\cdot,\cdot\right)$'s to be defined below.

Similar to U_{n1}, we have $U_{n2}=\widehat{U}_{n2}+o_p(\gamma)$, where $\widehat{U}_{n2}$ is the projection of U_{n2} onto the basic observation ξ_i.

$$\widehat{U}_{n2}=\theta_{n2}+\frac{6}{n}\sum_{i=1}^{n}\left[r_{n2}\left(\xi_i\right)-\theta_{n2}\right],$$

where $r_{n2}\left(\xi_i\right)\equiv E\left[P_{n2}\left(\xi_i,\xi_j,\xi_l\right)|\xi_i\right]$ and $\theta_{n1}\equiv E\left[r_{n2}\left(\xi_i\right)\right]=E\left[P_{n2}\left(\xi_i,\xi_j,\xi_l\right)\right]$. Since the techniques involved in the projection of U_{n2} on ξ_i is similar to those in Lemma B.5, I just report the final results of the projection of six components of $P_{n2}\left(\xi_i,\xi_j,\xi_l\right)$ on ξ_i.

For both

$$p_{n2}\left(\xi_i,\xi_j,\xi_l\right)=K_h\left(x-X_j\right)\tfrac{1}{h}K_h'\left(V_i-V_j\right)\tfrac{\left(Y_j-m(x,V_i)\right)a_i'a_j''}{f(x,V_i)}K_h\left(Z_j-Z_l\right)\tfrac{\left(X_{2l}-r(Z_j)\right)}{f(Z_j)}$$

$$p_{n2}\left(\xi_i,\xi_l,\xi_j\right)=K_h\left(x-X_l\right)\tfrac{1}{h}K_h'\left(V_i-V_l\right)\tfrac{\left(Y_l-m(x,V_i)\right)a_i'a_j''}{f(x,V_i)}K_h\left(Z_l-Z_j\right)\tfrac{\left(X_{2j}-r(Z_l)\right)}{f(Z_l)}:$$

$$E\left[p_{n2}\left(\xi_i,\cdot,\cdot\right)|\xi_i\right]=h_1^{s_1}\frac{f\left(x\right)f\left(V_i\right)a_i'}{f\left(x,V_i\right)}m_v'\left(x,V_i\right)\overline{R}_1+o_p\left(h_1^{s_1}\right),$$

and the expectation $E\left[p_{n2}\left(\xi_i,\cdot,\cdot\right)\right]=O\left(h_1^{s_1}\right)$ as $\frac{f(x)f(v)a_i'}{f(x,v)}m_v'\left(x,v\right)\overline{R}_1$ is bounded within C_{XV}'.

For both

$$p_{n2}\left(\xi_j,\xi_i,\xi_l\right)=K_h\left(x-X_i\right)\tfrac{1}{h}K_h'\left(V_j-V_i\right)\tfrac{\left(Y_i-m(x,V_j)\right)a_i'a_j''}{f(x,V_j)}K_h\left(Z_i-Z_l\right)\tfrac{\left(X_{2l}-r(Z_i)\right)}{f(Z_i)}$$

$$p_{n2}\left(\xi_l,\xi_i,\xi_j\right) = K_h\left(x-X_i\right)\tfrac{1}{h}K_h'\left(V_l-V_i\right)\tfrac{\left(Y_i-m(x,V_l)\right)a_i'a_j''}{f(x,V_l)}K_h\left(Z_i-Z_j\right)\tfrac{\left(X_j-r(Z_i)\right)}{f(Z_i)}:$$

$$\begin{aligned} & E\left[p_{n2}\left(\cdot,\xi_i,\cdot\right)|\xi_i\right] \\ = \;& h_1^{s_1}\tfrac{K_h(x-X_i)a_iR_1(Z_i)}{f(Z_i)}\left[\left(Y_i-m\left(x,V_i\right)\right)\left(\tfrac{f(V_i)}{f(x,V_i)}\right)_v' - m\left(x,V_i\right)_v'\tfrac{f(V_i)}{f(x,V_i)}\right]+o_p\left(h_1^{s_1}\right), \end{aligned}$$

and $E\left[p_{n2}\left(\cdot,\xi_i,\cdot\right)\right]=O\left(h_1^{s_1}\right)$ as $\frac{f(x)f(v)}{f(x,v)}m_v'\left(x,v\right)a_i'\overline{R}_1$ is bounded within C_{XV}'.

For both

$$p_{n2}\left(\xi_j,\xi_l,\xi_i\right) = K_h\left(x-X_l\right)\tfrac{1}{h}K_h'\left(V_j-V_l\right)\tfrac{\left(Y_l-m(x,V_j)\right)a_i'a_j''}{f(x,V_j)}K_h\left(Z_l-Z_i\right)\tfrac{\left(X_{2i}-r(Z_l)\right)}{f(Z_l)}$$

$$p_{n2}\left(\xi_l,\xi_j,\xi_i\right) = K_h\left(x-X_j\right)\tfrac{1}{h}K_h'\left(V_l-V_j\right)\tfrac{\left(Y_j-m(x,V_l)\right)a_i'a_j''}{f(x,V_l)}K_h\left(Z_j-Z_i\right)\tfrac{\left(X_{2i}-r(Z_j)\right)}{f(Z_j)}:$$

$$E\left[p_{n2}\left(\cdot,\cdot,\xi_i\right)|\xi_i\right] = \left[r\left(Z_i\right)-X_{2i}\right]f\left(x\right)\int\frac{f^2\left(v\right)m_v'\left(x,v\right)a_i'}{f\left(x,v\right)}dv+O_p\left(h_1^{s_1}\right),$$

and $E\left[p_{n1}\left(\cdot,\cdot,\xi_i\right)\right]=O\left(h_1^{s_1}\right)$ as $E\left[r\left(Z_i\right)-X_{2i}\right]=0$.

Although $\left[r\left(Z_i\right)-X_{2i}\right]f\left(x\right)\int\frac{f^2(v)m_v'(x,v)a_i'}{f(x,v)}dv$ is not of order $O_p\left(h_1^{s_1}\right)$,

$$\begin{aligned} & \frac{1}{n}\sum_{i=1}^{n}E\left[p_{n2}\left(\cdot,\cdot,\xi_i\right)|\xi_i\right] \\ = \;& f\left(x\right)\int\frac{f^2\left(v\right)m_v'\left(x,v\right)a_i'}{f\left(x,v\right)}dv\cdot\frac{1}{n}\sum_{i=1}^{n}\left[r\left(Z_i\right)-X_{2i}\right]=O_p\left(n^{-1/2}\right), \end{aligned}$$

where the variance is of order $O_p\left(n^{-1/2}\right)$ and $f\left(x\right)\int\frac{f^2(v)m_v'(x,v)a_i'}{f(x,v)}dv$ is bounded within C_{XV}'.

Put together, $\left(\widehat{U}_{n2}-\theta_{n2}\right)=\frac{6}{n}\sum_{i=1}^{n}\left[r_{n2}\left(\xi_i\right)-\theta_{n2}\right]=O_p\left(h_1^{s_1}\right)$ and $\left(\theta_{n2}-0\right)=O_p\left(h_1^{s_1}\right)$ so that $\widehat{U}_{n2}=O_p\left(h_1^{s_1}\right)$. Therefore $U_{n2}=\widehat{U}_{n2}+o_p\left(\gamma\right)=O_p\left(h_1^{s_1}\right)$. $\square$

Appendix C

Root-N-Consistency

I establish the $\sqrt{n}$-consistency of $\widehat{\beta}$, the density-weighted estimator with constructed variables in the nonparametric part of the partially linear model. The proof proceeds similarly to that in Li (1996), except that we need to take into account the fact that the conditioning variable V is a constructed one $\widehat{V}$. An analogue to Lemma 1 in Li (1996), Lemma C.1 is the key difference and shows the effect of preliminary kernel estimator $\widehat{V}$.

Define $\xi_i \equiv E(X_i|V_i)$ and $\eta_i \equiv X_i - \xi_i$ so that $X_i = \xi_i + \eta_i$ and $\widehat{X}_i = \widehat{\xi}_i + \widehat{\eta}_i$ where, $\widehat{f}_i \equiv \dfrac{1}{n}\sum_j K_h\left(\widehat{V}_i - \widehat{V}_j\right)$, and $\widehat{W}_i \equiv \dfrac{1}{n}\sum_j W_j K_h\left(\widehat{V}_i - \widehat{V}_j\right) \diagup \widehat{f}_i$ for $W_i = X_i, \xi_i, \eta_i$. Let $m(V_i) = c(V_i)$ or $m(V_i) = E(X_i|V_i) = \xi_i$, and $\mu_i = \varepsilon_i$ or $\mu_i = \eta_i$. I prove only the case that m is a scalar function $(d = 1)$. For the case with $d > 1$, the proof follows by the Cauchy inequality. Since $\frac{n}{n-1} \to 1$ as $n \to \infty$, the difference between n and $(n-1)$ is ignored.

Lemma C.1. $E\left[(m(V_i) - m(V_1)) K_h\left(\widehat{V}_i - \widehat{V}_1\right)|V_1\right] = O\left(h^s + h_1^{s_1}\right).$

Proof: As mentioned in Appendix B,

$$K_h\left(\widehat{V}_i - \widehat{V}_1\right) - K_h(V_i - V_1) = \frac{1}{h}K_h'(V_i - V_1)\left(\left(\widehat{V}_i - V_i\right) - \left(\widehat{V}_1 - V_1\right)\right) + s.o,$$

where s.o. represents that the remainder term is of smaller order than the first order term. Therefore, expand $K_h\left(\widehat{V}_i - \widehat{V}_1\right)$ and we get

$$\begin{aligned}
& E\left[\left(m\left(V_i\right)-m\left(V_1\right)\right) K_h\left(\widehat{V}_i-\widehat{V}_1\right)|V_1\right] \\
= \; & E\left[\left(m\left(V_i\right)-m\left(V_1\right)\right) K_h\left(V_i-V_1\right)|V_1\right] \\
& +E\left[\left(m\left(V_i\right)-m\left(V_1\right)\right) \frac{1}{h} K_h'\left(V_i-V_1\right)\left(\widehat{V}_i-V_i\right)|V_1\right] \\
& -E\left[\left(m\left(V_i\right)-m\left(V_1\right)\right) \frac{1}{h} K_h'\left(V_i-V_1\right)\left(\widehat{V}_1-V_1\right)|V_1\right].
\end{aligned}$$

It is a standard result that the first term $E\left[\left(m\left(V_i\right)-m\left(V_1\right)\right) K_h\left(V_i-V_1\right)|V_1\right] = O\left(h^s\right)$, see Robinson (1988) or Li (1996). The second and third terms are due to the preliminary estimator $\widehat{V}$, and I show that both terms are $O\left(h_1^{s_1}\right)$. From Appendix B, for $\widehat{V}_i \neq \infty$,

$$\widehat{V}_i - V_i = r\left(Z_i\right) - \widetilde{r}\left(Z_i\right) = \frac{1}{n}\sum_{l=1}^{n} K_h\left(Z_i - Z_l\right)\frac{\left(r\left(Z_i\right)-r\left(Z_l\right)\right)}{f\left(Z_i\right)}.$$

First, consider $E\left[\left(m\left(V_i\right)-m\left(V_1\right)\right) \frac{1}{h} K_h'\left(V_i-V_1\right)\left(\widehat{V}_1-V_1\right)|V_1\right]$. Conditioning on (V_1, Z_1, V_i),

$$\begin{aligned}
& E\left[\left(m\left(V_i\right)-m\left(V_1\right)\right) \frac{1}{h} K_h'\left(V_i-V_1\right)\left(\widehat{V}_1-V_1\right) \mid V_1, Z_1, V_i\right] \\
= \; & \left(m\left(V_i\right)-m\left(V_1\right)\right) \frac{1}{h} K_h'\left(V_i-V_1\right) E\left[\frac{1}{n}\sum_{l=1}^{n}\left(K_h\left(Z_1-Z_l\right)\frac{\left(r\left(Z_1\right)-r\left(Z_l\right)\right)}{f\left(Z_1\right)}\right) \mid Z_1\right] \\
= \; & \left(m\left(V_i\right)-m\left(V_1\right)\right) \frac{1}{h} K_h'\left(V_i-V_1\right) E\left[\left(K_h\left(Z_1-Z_l\right)\frac{\left(r\left(Z_1\right)-r\left(Z_l\right)\right)}{f\left(Z_1\right)}\right) \mid Z_1\right] \\
= \; & \left(m\left(V_i\right)-m\left(V_1\right)\right) \frac{1}{h} K_h'\left(V_i-V_1\right) \cdot h_1^{s_1} k_s r^{(s_1)}\left(Z_1\right),
\end{aligned}$$

where the last equality follows by Lemma B.3.(i). Then by Lemma B.4,

$$\begin{aligned}
& E\left[\left(m\left(V_i\right)-m\left(V_1\right)\right)\frac{1}{h}K_h'\left(V_i-V_1\right)\left(\widehat{V}_1-V_1\right)|V_1\right] \\
= \;& h_1^{s_1}k_s r^{(s_1)}\left(Z_1\right)\left[m'\left(V_1\right)f\left(V_1\right)+O\left(h^s\right)\right].
\end{aligned}$$

Since the functions $r^{(s_1)}(Z_1)$, $m'(V_1)$ and $f(V_1)$ are all bounded,

$$E\left[\left(m\left(V_i\right)-m\left(V_1\right)\right)\frac{1}{h}K_h'\left(V_i-V_1\right)\left(\widehat{V}_1-V_1\right)|V_1\right]=O\left(h_1^{s_1}\right).$$

Next, consider $E\left[\left(m\left(V_i\right)-m\left(V_1\right)\right)\frac{1}{h}K_h'\left(V_i-V_1\right)\left(\widehat{V}_i-V_i\right)|V_1\right]$. Conditioning on (V_1,V_i,Z_i),

$$\begin{aligned}
& E\left[\left(m\left(V_i\right)-m\left(V_1\right)\right)\frac{1}{h}K_h'\left(V_i-V_1\right)\left(\widehat{V}_i-V_i\right)\mid V_1,V_i,Z_i\right] \\
= \;& \left(m\left(V_i\right)-m\left(V_1\right)\right)\frac{1}{h}K_h'\left(V_i-V_1\right)E\left[\frac{1}{n}\sum_{l=1}^{n}\left(K_h\left(Z_i-Z_l\right)\frac{\left(r\left(Z_i\right)-r\left(Z_l\right)\right)}{f\left(Z_i\right)}\right)\mid Z_i\right] \\
= \;& \left(m\left(V_i\right)-m\left(V_1\right)\right)\frac{1}{h}K_h'\left(V_i-V_1\right)E\left[\left(K_h\left(Z_i-Z_l\right)\frac{\left(r\left(Z_i\right)-r\left(Z_l\right)\right)}{f\left(Z_i\right)}\right)\mid Z_i\right] \\
= \;& \left(m\left(V_i\right)-m\left(V_1\right)\right)\frac{1}{h}K_h'\left(V_i-V_1\right)\cdot h_1^{s_1}k_s r^{(s_1)}\left(Z_i\right),
\end{aligned}$$

where the last equality follows by Lemma B.3.(i). Similarly, we have,

$$E\left[\left(m\left(V_i\right)-m\left(V_1\right)\right)\frac{1}{h}K_h'\left(V_i-V_1\right)\left(\widehat{V}_i-V_i\right)|V_1\right]=O\left(h_1^{s_1}\right).$$

Together, $E\left[\left(m\left(V_i\right)-m\left(V_1\right)\right)K_h\left(\widehat{V}_i-\widehat{V}_1\right)|V_1\right]=O\left(h^s+h_1^{s_1}\right)$. □

Lemma C.2. $S_{(\widehat{m}-m)\widehat{f}}=O_p\left(h^{2s}+h_1^{2s_1}+h^2\left(nh^d\right)^{-1}\right)$**.**

Proof: Let $E_1[\cdot]\equiv E(\cdot|V_1)$ and $K_{i1}\equiv K_h\left(\widehat{V}_i-\widehat{V}_1\right)$. It can be shown that

$E\left[(m_i - m_1)^2 K_{i1}^2\right] = O\left(h^{2-d}\right)$.

$$
\begin{aligned}
& E\left[\left|S_{(\widehat{m}-m)\widehat{f}}\right|\right] \\
= \; & \frac{1}{n}\sum_i E\left[(\widehat{m}_i - m_i)^2 \widehat{f}_i^{\,2}\right] = E\left[(\widehat{m}_1 - m_1)^2 \widehat{f}_1^{\,2}\right] \\
= \; & \frac{1}{n^2}\textstyle\sum_i \sum_j E\left[(m_i - m_1) K_{i1} (m_j - m_1) K_{j1}\right] \\
= \; & \frac{1}{n^2}\sum_{i\neq 1}\left\{E\left[(m_i - m_1)^2 K_{i1}^2\right] + \sum_{j\neq 1, j\neq i} E\left(E_1\left[(m_i - m_1) K_{i1}\right] \cdot E_1 (m_j - m_1) K_{j1}\right)\right\} \\
\leqslant \; & \frac{1}{n}\left[O\left(h^{2-d}\right) + nO\left(h^s + h_1^{s_1}\right)^2\right] \\
= \; & O\left(h^{2s} + h_1^{2s_1} + h^2\left(nh^d\right)^{-1}\right),
\end{aligned}
$$

where the inequality follows from Lemma C.1. $\square$

Lemma C.3. $S_{(\widehat{m}-m)\widehat{f},\mu\widehat{f}} = o_p\left(n^{-1/2}\right)$.

Proof: Notice that $S_{(\widehat{m}-m)\widehat{f},\mu\left(\widehat{f}-f\right)}$ is of smaller order than $S_{(\widehat{m}-m)\widehat{f},\mu f}$ since $\left(\widehat{f} - f\right) = o_p(1)$. So $S_{(\widehat{m}-m)\widehat{f},\mu\widehat{f}}$ is of the same order as $S_{(\widehat{m}-m)\widehat{f},\mu f}$, which is derived below.

$$
\begin{aligned}
E\left[S^2_{(\widehat{m}-m)\widehat{f},\mu f}\right] &= \frac{1}{n^2}\sum_i E\left[(\widehat{m}_i - m_i)^2 \widehat{f}_i^{\,2} \mu_i^2 f_i^2\right] \\
&= \frac{1}{n}E\left[(\widehat{m}_1 - m_1)^2 \widehat{f}_1^{\,2} \sigma_\mu^2 (X_1, V_1) f_i^2\right] \\
&\leqslant \frac{B_\sigma}{n}E\left[(\widehat{m}_1 - m_1)^2 \widehat{f}_1^{\,2}\right] \\
&= \frac{B_\sigma}{n}E\left[S_{(\widehat{m}-m)\widehat{f}}\right] = n^{-1}o(1) = o\left(n^{-1}\right),
\end{aligned}
$$

where the inequality follows from the boundedness of the functions of (X_1, V_1) and the fourth equality from Lemma C.2. Therefore, $S_{(\widehat{m}-m)\widehat{f},\mu\widehat{f}} = o_p\left(n^{-1/2}\right)$. $\square$

Lemma C.4. $S_{(\widehat{m}-m)\widehat{f},\widehat{\mu}\widehat{f}} = o_p\left(n^{-1/2}\right)$.

Proof: By the Cauchy inequality,

$$\begin{aligned}
\left|S_{(\widehat{m}-m)\widehat{f},\widehat{\mu}\widehat{f}}\right| &\leqslant \left[\left|S_{(\widehat{m}-m)\widehat{f}}\right|\left|S_{\widehat{\mu}\widehat{f}}\right|\right]^{1/2} \\
&= \left[O_p\left(h^{2s}+h_1^{2s_1}+h^2\left(nh^d\right)^{-1}\right)O_p\left(\left(nh^d\right)^{-1}\right)\right]^{1/2} \\
&= O_p\left(h^s\left(nh^d\right)^{-1/2}+h_1^{s_1}\left(nh^d\right)^{-1/2}+h\left(nh^d\right)^{-1}\right) \\
&= o_p\left(n^{-1/2}\right). \qquad \square
\end{aligned}$$

Lemma C.5.

(i) . $S_{\varepsilon\widehat{f},\widehat{\eta}\widehat{f}} = o_p\left(n^{-1/2}\right)$; (ii). $S_{\widehat{\varepsilon}\widehat{f},\eta\widehat{f}} = o_p\left(n^{-1/2}\right)$; (iii). $S_{\widehat{\varepsilon}\widehat{f},\widehat{\eta}\widehat{f}} = o_p\left(n^{-1/2}\right)$.

Proof of (i): Again notice that $S_{\varepsilon\left(\widehat{f}-f\right),\widehat{\eta}\widehat{f}}$ is of smaller order than $S_{\varepsilon f,\widehat{\eta}\widehat{f}}$ since $\left(\widehat{f}-f\right) = o_p(1)$. So $S_{\varepsilon\widehat{f},\widehat{\eta}\widehat{f}}$ is of the same order as $S_{\varepsilon f,\widehat{\eta}\widehat{f}}$, which is derived as follows.

$$\begin{aligned}
E\left[S^2_{\varepsilon f,\widehat{\eta}\widehat{f}}\right] &= \frac{1}{n^2}\sum_i E\left[\mu_i^2 f_i^2\left(\widehat{\eta}_i\widehat{f}_i\right)^2\right] \\
&= \frac{1}{n}E\left[\sigma_\mu^2\left(X_1,V_1\right)f_i^2\left(\widehat{\eta}_i\widehat{f}_i\right)^2\right] \\
&\leqslant \frac{B_\sigma}{n}E\left[\left(\widehat{\eta}_i\widehat{f}_i\right)^2\right] = O\left(n^{-1}\left(nh^d\right)^{-1}\right),
\end{aligned}$$

where the inequality follows from the boundedness of the functions of (X_1,V_1) and the last equality from Lemma C.2. Therefore, $S_{\varepsilon\widehat{f},\widehat{\eta}\widehat{f}} = O\left(\left(nh^{d/2}\right)^{-1}\right) = o_p\left(n^{-1/2}\right)$.

Proof of (ii): The same as (i).

Proof of (iii): By the Cauchy inequality,

$$\begin{aligned}
\left|S_{\widehat{\varepsilon}\widehat{f},\widehat{\eta}\widehat{f}}\right| &\leqslant \left[\left|S_{\widehat{\varepsilon}\widehat{f}}\right|\left|S_{\widehat{\eta}\widehat{f}}\right|\right]^{1/2} \\
&= \left[O_p\left(\left(nh^d\right)^{-1}\right)O_p\left(\left(nh^d\right)^{-1}\right)\right]^{1/2} \\
&= O_p\left(\left(nh^d\right)^{-1}\right) = o_p\left(n^{-1/2}\right). \qquad \square
\end{aligned}$$

Lemma C.6. (i). $S_{\widehat{\mu}\widehat{f}} = o_p(1)$; (ii). $S_{\widehat{\mu}\widehat{f},\mu\widehat{f}} = o_p(1)$.

Proof of (i):

$$\begin{aligned}
E\left[\left|S_{\widehat{\mu}\widehat{f}}\right|\right] &= \frac{1}{n}\sum_i E\left[\widehat{\mu}_i^2\widehat{f}_i^2\right] = E\left[\widehat{\mu}_1\widehat{f}_1\right] \\
&= \frac{1}{n^2}\sum_i\sum_j E\left[\mu_i\mu_j K_{i1}K_{j1}\right] \\
&= \frac{1}{n^2}\sum_i E\left[\mu_i^2 K_{i1}^2\right] \\
&= \frac{1}{n}E\left[\sigma_\mu^2\left(X_1,V_1\right)K_{i1}^2\right] \\
&\leqslant \frac{B_\sigma}{n}E\left[K_{i1}^2\right] = O\left(\left(nh^d\right)^{-1}\right) = o(1).
\end{aligned}$$

Proof of (ii): Follows from Lemma C.5.(i). $\square$

Proposition C.1. $S_{\left(X-\widehat{X}\right)\widehat{f}} \xrightarrow{p} \Phi_f$.

Proof: Note that $X_i = \xi_i + \eta_i$ and $\widehat{X}_i = \widehat{\xi}_i + \widehat{\eta}_i$ so that $\left(X-\widehat{X}\right)\widehat{f} = \left(\xi - \widehat{\xi}\right)\widehat{f} + \eta\widehat{f} - \widehat{\eta}\widehat{f}$. We have

$$\begin{aligned}
S_{\left(X-\widehat{X}\right)\widehat{f}} &= S_{\left(\xi-\widehat{\xi}\right)\widehat{f}+\eta\widehat{f}-\widehat{\eta}\widehat{f}} \\
&= S_{\left(\xi-\widehat{\xi}\right)\widehat{f}} + S_{\eta\widehat{f}} + S_{\widehat{\eta}\widehat{f}} + 2S_{\left(\xi-\widehat{\xi}\right)\widehat{f},\eta\widehat{f}} - 2S_{\left(\xi-\widehat{\xi}\right),\widehat{\eta}\widehat{f}} - 2S_{\eta\widehat{f},\widehat{\eta}\widehat{f}} \\
&= S_{\eta\widehat{f}} + o_p(1) = S_{\eta f} + o_p(1) \\
&= \frac{1}{n}\sum_i \eta_i\eta_i' f_i + o_p(1) \xrightarrow{p} E\left(\eta_1\eta_1' f_1\right) \equiv \Phi_f,
\end{aligned}$$

by Lemmas C.2 - C.6 and the Law of Large Numbers. □

Proposition C.2. $\sqrt{n}S_{\left(X-\widehat{X}\right)\widehat{f},\varepsilon\widehat{f}} \xrightarrow{d} N\left(0,\Psi_f\right).$

Proof: Since $\left(X-\widehat{X}\right)\widehat{f} = \left(\xi-\widehat{\xi}\right)\widehat{f} + \eta\widehat{f} - \widehat{\eta}\widehat{f},$

$$\begin{aligned}
\sqrt{n}S_{\left(X-\widehat{X}\right)\widehat{f},\varepsilon\widehat{f}} &= \sqrt{n}\left(S_{\left(\xi-\widehat{\xi}\right)\widehat{f},\varepsilon\widehat{f}} + S_{\eta\widehat{f},\varepsilon\widehat{f}} + S_{\widehat{\eta}\widehat{f},\varepsilon\widehat{f}}\right) \\
&= \sqrt{n}S_{\eta\widehat{f},\varepsilon\widehat{f}} + o_p\left(1\right) = \sqrt{n}S_{\eta f,\varepsilon f} + o_p\left(1\right) \\
&= \frac{1}{\sqrt{n}}\sum_i \eta_i\varepsilon_i f_i^2 \xrightarrow{d} N\left(0,\Psi_f\right),
\end{aligned}$$

by Lemmas C.2 - C.6 and the Central Limit Theorem. □

Proposition C.3. (i) $S_{\left(X-\widehat{X}\right)\widehat{f},(g-\widehat{g})\widehat{f}} = o_p\left(n^{-1/2}\right)$, and (ii) $S_{\left(X-\widehat{X}\right)\widehat{f},\widehat{\varepsilon}\widehat{f}} = o_p\left(n^{-1/2}\right).$

Proof: Since $\left(X-\widehat{X}\right)\widehat{f} = \left(\xi-\widehat{\xi}\right)\widehat{f} + \eta\widehat{f} - \widehat{\eta}\widehat{f}$, we have (i):

$$\begin{aligned}
S_{\left(X-\widehat{X}\right)\widehat{f},(g-\widehat{g})\widehat{f}} &= S_{\left(\xi-\widehat{\xi}\right)\widehat{f}+\eta\widehat{f}-\widehat{\eta}\widehat{f},(g-\widehat{g})\widehat{f}} \\
&= S_{\left(\xi-\widehat{\xi}\right)\widehat{f},(g-\widehat{g})\widehat{f}} + S_{\eta\widehat{f},(g-\widehat{g})\widehat{f}} - S_{\widehat{\eta}\widehat{f},(g-\widehat{g})\widehat{f}} \\
&= o_p\left(n^{-1/2}\right);
\end{aligned}$$

and (ii):

$$\begin{aligned}
S_{\left(X-\widehat{X}\right)\widehat{f},\widehat{\varepsilon}\widehat{f}} &= S_{\left(\xi-\widehat{\xi}\right)\widehat{f},\widehat{\varepsilon}\widehat{f}} + S_{\eta\widehat{f},\widehat{\varepsilon}\widehat{f}} - S_{\widehat{\eta}\widehat{f},\widehat{\varepsilon}\widehat{f}} \\
&= o_p\left(n^{-1/2}\right).
\end{aligned}$$

by Lemmas C.2 - C.6. □

Appendix D

Monte Carlo Simulations

Semiparametric Control Function Estimation:

$Y = X'\beta + U$, *where* $E(U|X) \neq 0$ *and* $E(U) = 0$*;*

$V = X - E(X|Z)$, *where* $E[U|X, V] = E[U|V]$.

True values of parameters: $\beta_1 = 0.3$ and $\beta_2 = 0.7$.

Sample Size n	β_1	**S.E.**	β_2	**S.E.**
$n = 100$	0.290	(0.031)	0.711	(0.072)
$n = 500$	0.293	(0.026)	0.706	(0.086)
$n = 1000$	0.297	(0.020)	0.702	(0.059)

Table D.1 Estimates of β

Nonparametric Control Function Estimation

Four Specifications

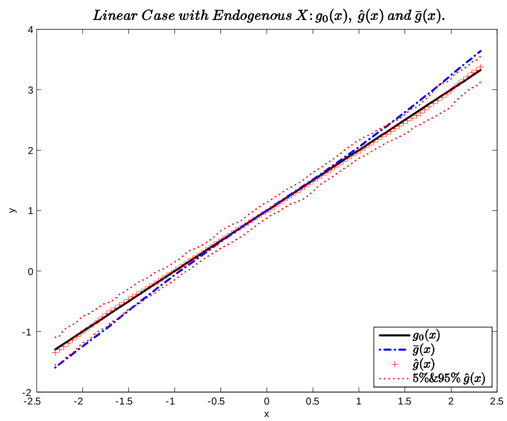

Figure D.1

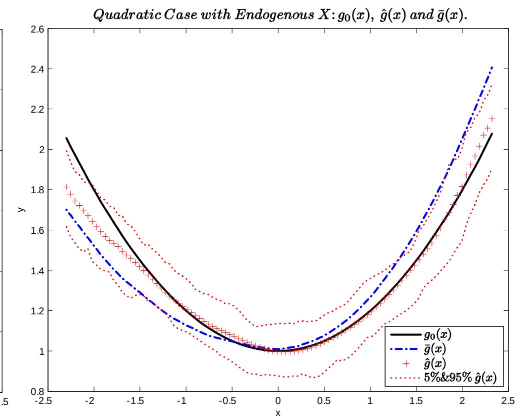

Figure D.2

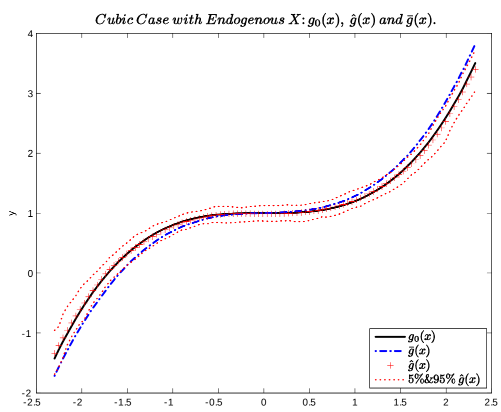

Figure D.3

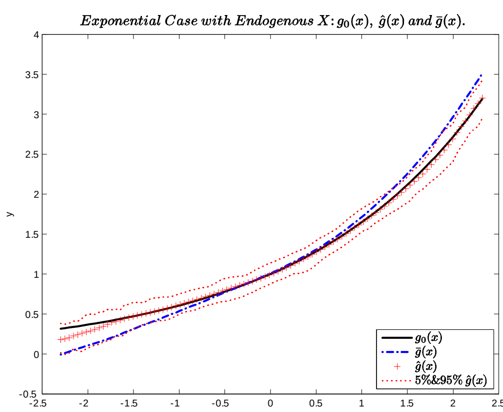

Figure D.4

Exogenous v.s Endogenous Regressors

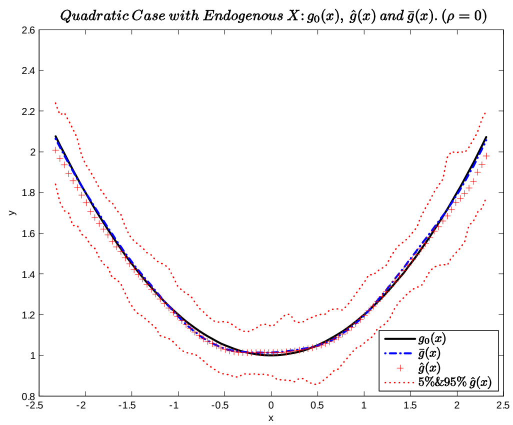

Figure D.5

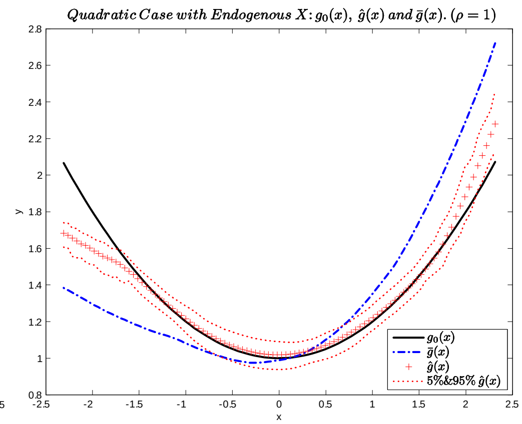

Figure D.6

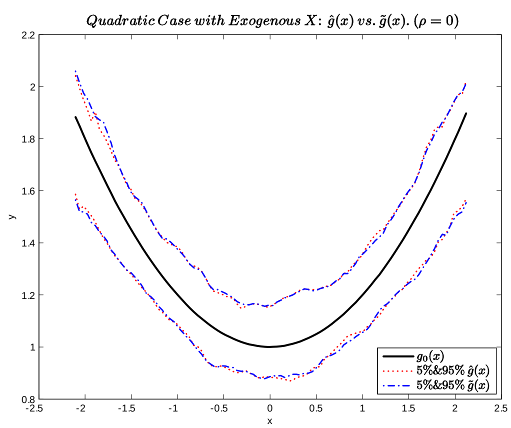

Figure D.7

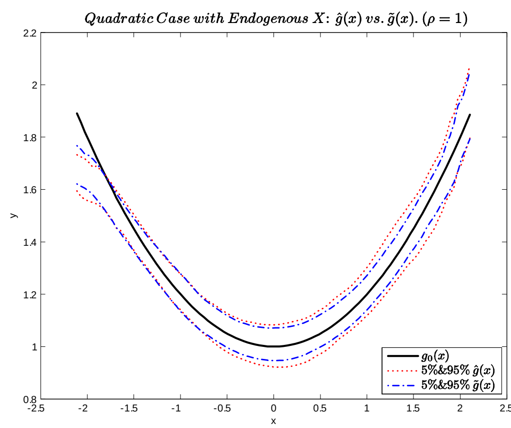

Figure D.8

The Effect of The Dimension of Instruments

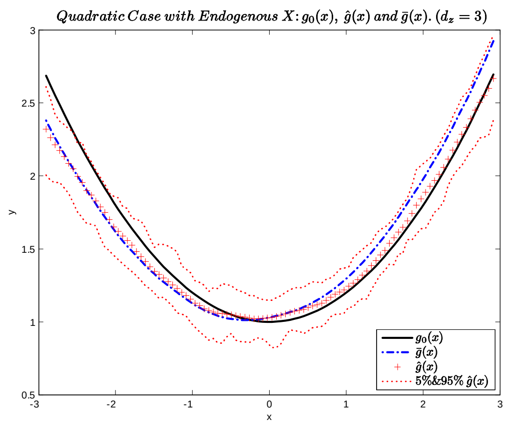

Figure D.9

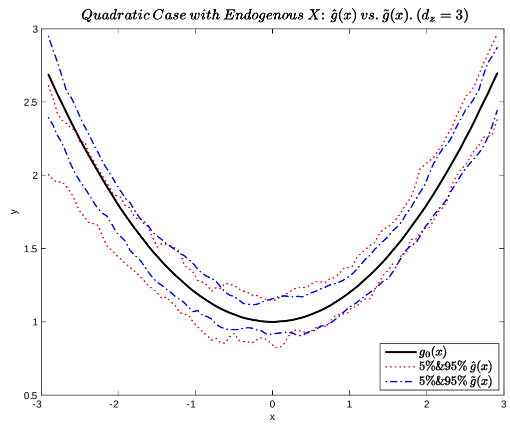

Figure D.10

The Effect of Bandwidths

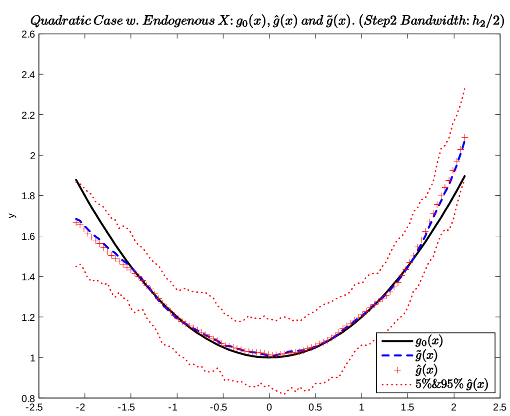

Figure D.11

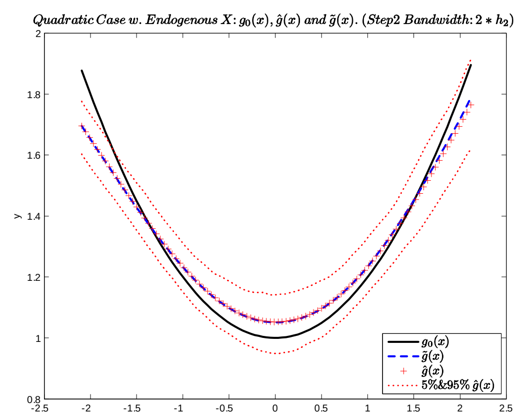

Figure D.12

Appendix E

Empirical Results

Chilean Panel Dataset - Food Industry - 1986

$t =$ **1986**	**Obs**	**Mean**	**S.D.**	**Min**	**Max**
Log Value Added y_{it}	787	9.018	1.686	5.034	14.192
Log Capital k_{it}	787	8.047	2.061	-0.659	14.472
Log Capital $k_{i,t-1}$	787	8.061	2.033	-0.554	14.482
Log Capital $k_{i,t-2}$	787	8.094	2.003	-0.448	14.520
Log Labor l_{it}	787	3.392	0.935	2.303	6.731
Log Labor $l_{i,t-1}$	787	3.377	0.902	2.303	6.575
Log Labor $l_{i,t-2}$	787	3.343	0.878	2.303	6.548

Table E.1. Summary Statistics

Chilean Panel Dataset - Food Industry - 1986

Food Industry	β_k	**S.E.**	β_l	**S.E.**
NPCF: $z_{it} = (k_{i,t-1}, l_{i,t-1})$	0.297	(0.054)	0.807	(0.072)
NPCF: $z_{it} = (k_{i,t-2}, l_{i,t-2})$	0.303	(0.059)	0.814	(0.086)
SPCF: $z_{it} = (k_{i,t-1}, l_{i,t-1})$	0.369	(0.043)	0.765	(0.059)
SPCF: $z_{it} = (k_{i,t-2}, l_{i,t-2})$	0.372	(0.041)	0.770	(0.061)
ACF - m_{it}	0.383	(0.042)	0.832	(0.054)
ACF - e_{it}	0.386	(0.040)	0.867	(0.051)
LP - m_{it}	0.458	(0.039)	0.681	(0.039)
LP - e_{it}	0.451	(0.037)	0.764	(0.046)
OLS	0.344	(0.029)	0.937	(0.034)
FE	0.165	(0.052)	0.704	(0.060)

Table E.2 Estimates of Capital and Labor Coefficients

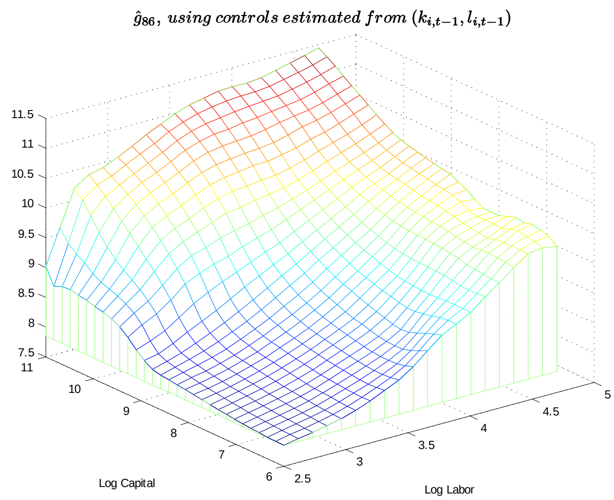

Figure E.1

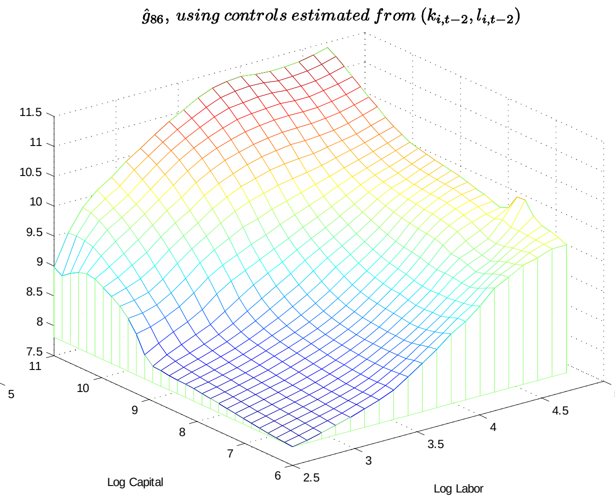

Figure E.2

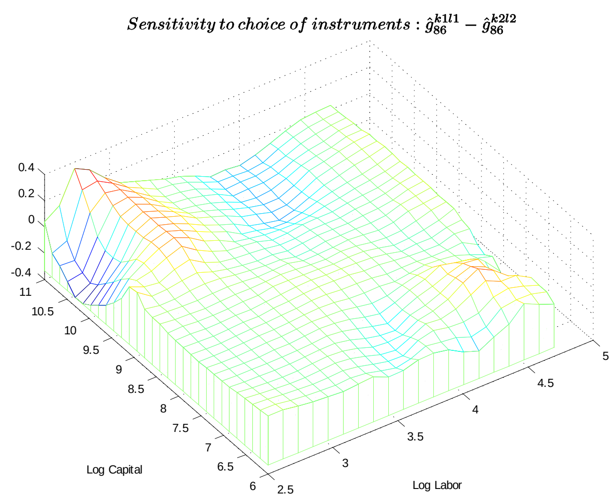

Figure E.3

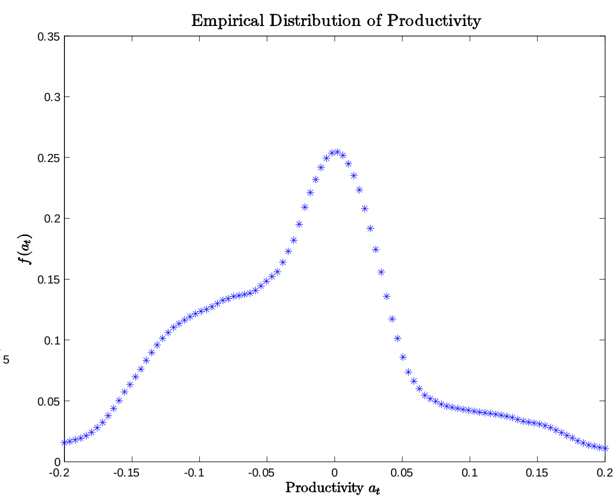

Figure E.4

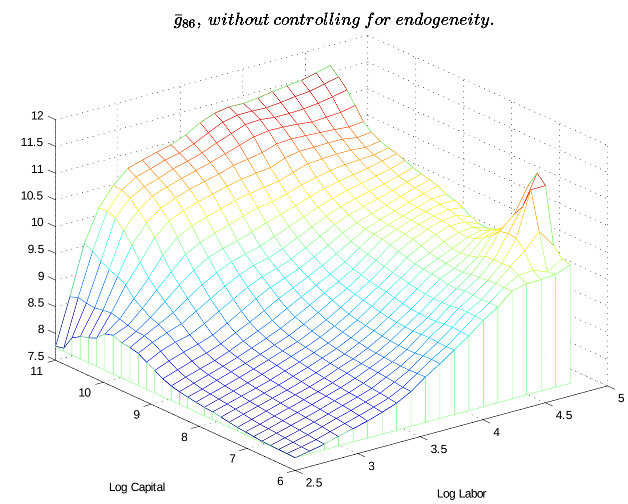

Figure E.5

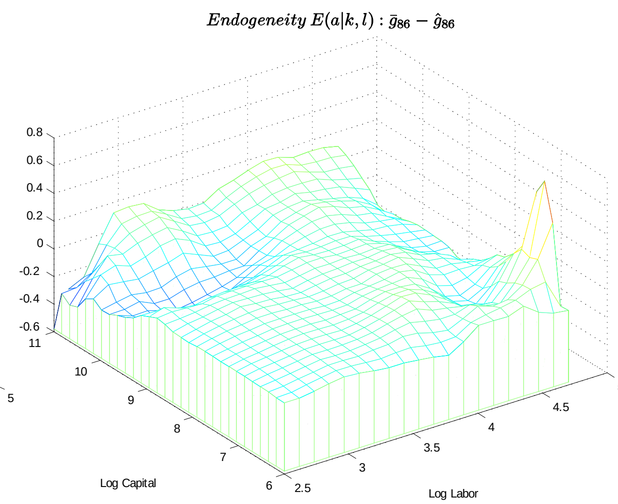

Figure E.6

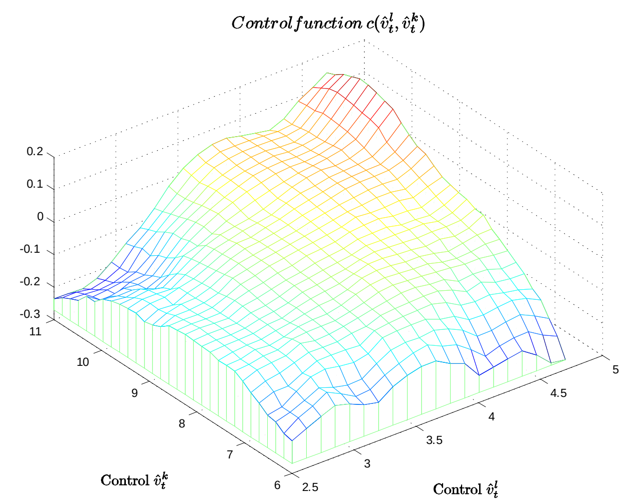

Figure E.7

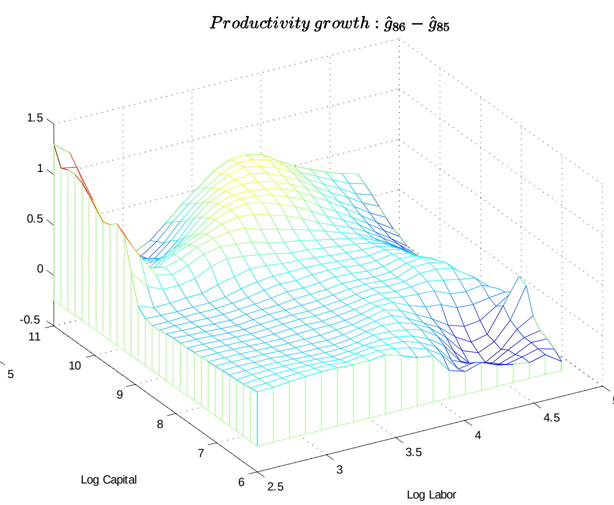

Figure E.8

Jian Hong

Address:	Department of Economics 608 Kern Graduate Building The Pennsylvania State University University Park, PA 16802	Telephone: E-mail: Website:	Office: (814) 865-1108 Cell: (814) 360-2934 jhong@psu.edu http://www.personal.psu.edu/jzh109/

Curriculum Vitae

CITIZENSHIP:
- China (F-1 visa)

EDUCATION:
- Ph.D., Economics, Penn State University, expected Summer 2008
- M.A., Economics, Peking University, China, 2001
- B.S., Information Technology, Beijing Normal University, China, 1997

PH.D. THESIS:
- Nonparametric Estimation and Testing Using Control Function Approaches
- **Thesis Advisor**: Professor Quang Vuong

FIELDS:
- **Primary**: Applied and Theoretical Econometrics
- **Secondary**: Industrial Organization, Applied Microeconomics

PAPERS:
- "Nonparametric Identification and Estimation of Production Functions Using Control Function Approaches to Endogeneity," 2007 (Job Market Paper)
- "Semiparametric Identification and Estimation of Production Functions," 2007
- "A Nonparametric Hausman Test of Exogeneity," 2007 (in progress)
- "Residential Electricity Consumption under Real-time Pricing," 2007
- "Identification in Empirical Auctions," 2004

TEACHING EXPERIENCE:
- Instructor:
  Intermediate Macroeconomics
- Teaching Assistant:
  Intermediate Microeconomics, Money and Banking, Principles of Economics, Principles of Macroeconomics

RESEARCH EXPERIENCE:
- Research Assistant, Summer 2002, 2004 for Professor Susanna Esteban
- Research Assistant, Summer 2003 for Professor Gustavo Ventura

PRESENTATIONS:
- Applied Microeconomics Workshop, Penn State, Summer 2005

AWARDS:
- Graduate Scholar Award, Penn State, 2002-2003

REFERENCES:
- Professor Quang Vuong, Penn State, (qxv1@psu.edu)
- Professor Mark Roberts, Penn State, (mroberts@psu.edu)
- Professor Isabelle Perrigne, Penn State, (iup2@psu.edu)
- Professor Joris Pinkse, Penn State, (joris@psu.edu)