# Identification and Estimation of a Triangular Model with Multiple Endogenous Variables and Insufficiently Many Instrumental Variables*

Liquan Huang† Umair Khalil‡ Neşe Yıldız§

First draft: April 2013. This version: Dec. 30, 2016.

**Abstract**

We develop a novel identification method for a partially linear model with multiple endogenous variables of interest but a single instrumental variable, which could even be binary. We present an easy-to-implement consistent estimator for the parametric part. This estimator retains $\sqrt{n}$-convergence rate and asymptotic normality even though we have a generated regressor in our setup. The nonparametric part of the model is also identified. We also outline how our identification strategy can be extended to a fully non-parametric model. We use our methods to assess the impact of smoking during pregnancy on birth weight.

*JEL Classifications*: *C31, C36, C13, C14*

*Key words:* Identification, multiple endogenous variables, control function approach.

*This paper was presented at the internal faculty lunch at University of Rochester, Camp Econometrics NY VIII, Second Seattle-Vancouver Econometrics Conference, and research seminars at Boston College and UC Davis. We would like to thank participants of those seminars, as well as Christoph Rothe for their valuable comments. Any remaining errors are our own.

†Department of Economics, University of Rochester, 231 Harkness Hall, Rochester, NY 14627; Email: l.huang@rochester.edu.

‡Department of Economics, Box 6025, West Virginia University, Morgantown, WV 26506-6025. Email: umair.khalil@mail.wvu.edu.

§Corresponding author: Department of Economics, University of Rochester, 231 Harkness Hall, Rochester, NY 14627; Email: nese.yildiz@rochester.edu; Phone: 585-275-5782; Fax: 585-256-2309.

# 1 Introduction

This paper considers models with multiple sources of endogeneity in which there is a single instrumental variable, which could possibly be binary. We show how parameters of interest in such models can be identified. Our baseline model is partially linear. We provide a new estimator that is easy to implement for the coefficient of the endogenous regressor for which no instrumental variables are available. By applying empirical process methods, we show that the estimator retains $\sqrt{n}$-convergence rate and is asymptotically normal even though we have a generated regressor. The nonparametric part of the baseline partially linear model is also identified up to a constant, and can be estimated with the standard nonparametric convergence rate. We also show how the identification strategy used to identify the baseline model could be extended for identification in a nonparametric model.

To see why methods developed in this paper might be useful, consider the problem of studying the effect of maternal smoking during pregnancy on birth weight. Medical literature has long established a link between maternal smoking and adverse birth outcomes, primarily measured by the effect on average birth weights. However, the primary concern involved in estimating the causal effect of maternal smoking on birth weights is the lack of an exogenous source of variation.[1] Using the methods proposed in this paper, we can potentially solve this issue by using an available IV for a second endogenous variable of interest in our structural equation. For instance, Currie and Moretti (2003) (henceforth CM) studies the effect of maternal education on birth outcomes. Given the former is clearly endogenous, they use the number of colleges in the county where the mother was resident at age 17 as their instrumental variable.

To formulate ideas, suppose $Y$ denotes birth weight of the baby, $X$ denotes mother's schooling, $Z$ represents the CM IV, and $D$ denotes our key endogenous variable, maternal smoking during pregnancy. Finally, $\varepsilon$ represents all remaining unobservables. The model we study is

$$Y = \lambda(X) + D^{\intercal}\gamma + \varepsilon, \tag{1}$$

$$X = \pi(Z) + V. \tag{2}$$

By running a first stage non-parametric regression of the endogenous regressor $X$ on the instrument we could identify $V$. The crucial assumption we impose is that the residual from this regression will be a control function for the endogenous variable for which an IV is available. That is, we assume $\mathbb{E}(\varepsilon|X,V) = \mathbb{E}(\varepsilon|V) =: \rho(V)$. Note that inclusion of $V$ into the outcome equation as a regressor will only control for endogeneity of $X$, but not of $D$. This means that we could write $\varepsilon = \rho(V) + \varsigma$, where $\mathbb{E}(\varsigma|X,V) = 0$ by construction. In particular, $X$ and $V$ will be additively separable and exogenous in the outcome equation written with $\varsigma$ as the only unobserved

[1] The most comprehensive study on the question, Almond et al. (2005) also uses a selection-on-observables framework. It is particularly hard to find a source of variation that is correlated with smoking during pregnancy but uncorrelated with unobserved mothering ability

term. However, $E(\varsigma|D)$ is not required to be 0. Given this setup, the cross partial derivative of $\mathbb{E}(Y|X,V)$ with respect to $X$ and $V$ must be equal to $\gamma$ (the coefficient vector on the endogenous regressors $D$) times the cross partial derivative of $\mathbb{E}(D|X,V)$ with respect to $X$ and $V$.

If the vector of cross partial derivatives of $\mathbb{E}(D|X,V)$ with respect to $X$ and $V$ are linearly independent, which is a testable assumption, then our coefficient of interest $\gamma$ is identified. Once $\gamma$ is identified, we can subtract $D^{\intercal}\gamma$ from the outcome and identify $\lambda(X)$ as well as $\rho(V)$ up to some location normalization. Note that if the instrument $Z$ is binary we could only identify the continuous unobservable (but identifiable) control function $V$ at two points. In that case, the coefficient vector $\gamma$ can still be identified by replacing partial derivatives with differencing, so that cross partial derivatives are replaced by differences in differences. This is our baseline model.

Under this setup, we estimate the causal effect of smoking on birth weight in Section 4. However, our methods can be applied to a wide range of applications. Consider the empirical problem of estimating the dynamic evolution of crime. That is, suppose we are interested in estimating how crime in the last period affects crime in the current period. Given unobserved criminogenic factors at the neighborhood level a naive OLS would yield biased estimates. Jacob, Lefgren and Moretti (2007) use last period's weather conditions as an IV for last period's criminal activity. However, part of this estimated effect might be due to learning-by-doing related channels on the part of the criminals. Similarly, police response to criminal behavior in a given neighborhood, might also respond to last period's criminal activity making inference difficult. Moreover, we might also be interested in evaluating the causal effect of last period's criminal activity through such additional channels whose own endogeneity can create further problems. The methods developed in this paper can be used to also evaluate such effects, as the scenario again incorporates insufficiently many instrumental variables compared to endogenous regressors.

Based on our identification strategy, we propose an easy-to-compute estimator which achieves $\sqrt{n}$-convergence rate for $\gamma$. Asymptotic normality of this estimator is also derived. We should point out that our approach consists of a multi-step estimation and applies the control function approach which leaves us with a generated regressor. Consistency and $\sqrt{n}$-convergence results are non-trivial when generated regressors are involved.

In Section 2, we outline how our identification strategy for the baseline model could be extended to a non-parametric model. This extension highlights that separability of $X$ and $D$ is not necessary for our methodology. On the other hand, while our identification strategy allows us to make inferences about effects of multiple endogenous variables with few or even a single instrument, it does rely on some crucial assumptions. As in applying any new method, researchers should make sure that our identifying assumptions are suitable for their particular application when applying our method.

Our baseline model is partially linear. The partially linear model has been well established in econometric theory since the seminal works of Robinson (1988) and Speckman (1988). Identification and estimation of partially linear models with endogeneity in either the parametric part or

the nonparametric part have been discussed in the literature by e.g. Ai and Chen (2003), Chen and Pouzo (2009), Florens (2003). In a recent study Florens et al. (2012) also propose a $\sqrt{n}$-consistent estimator for the parametric part, when endogeneity exists in both the parametric and nonparametric parts. However their method relies on the availability and strength of sufficiently many instrumental variables. Our method, on the other hand, only requires the availability of one instrumental variable.

Both in our baseline model and in non-parametric extension 1 in Section 2, $V$ is a control function (though as explained in detail in Section 2, $V$ could also be an observed covariate). As a result, this paper is also related to the control function approach (see Newey, Powell and Vella (1999)[2], Blundell and Powell (2003), Imbens and Newey (2009) among others). In our case, however, the control function is not required to account for all sources of endogeneity; its inclusion into the outcome equation is only required to control for endogeneity in $X$, but not in $D$. We use a first stage nonparametric regression of $X$ on $Z$ to obtain this control function.

Identification of marginal effects of multiple endogenous variables with a single instrument was first established Huang and Yıldız (2013). In Section 2 we outline two non-parametric extensions with one of them incorporating $V$ as a control function. As such our work is also related to Torgovitsky (2015) and D'Haultfoeuille and Fevrier (2015). Using very different methods than ours, in a model where both the outcome and first stage equations are strictly monotone in their corresponding scalar unobservable variable both of these papers demonstrate that identification of structural functions are possible even with a discrete instrument when there is a single continuous endogenous variable. In both of these papers the outcome equation is non-separable in the endogenous variable $X$ and the unonservable variable. In our related non-parametric extension the unobservable $\varepsilon$ enters the outcome equation additively separably from $(X, D)$, which are the endogenous variables. On the other hand, we demonstrate that identification of marginal effects of multiple endogenous variables is possible with a possibly binary, single instrument. This feature of our paper is also shared by Caetano and Escanciano (2015, 2016), which uses a related but different identification strategy to achieve identification of marginal effects of multiple endogenous variables with a possibly single, binary instrument in the context of the non-parametric IV modesl.

We discuss the estimation of only our baseline model in this paper. Since estimation of $\lambda$ and $\rho$ (up to a constant) would follow almost immediately from Newey, Powell and Vella (1999). Once a $\sqrt{n}$-normal estimator for $\gamma$ is at hand, we focus on estimation of $\gamma$ only. Based on our identification strategy, our proposed estimator for $\gamma$ is of the ratio form, where both the numerator and the denominator are averages of cross partial derivatives of nonparametric regressions with generated regressors. The literature on estimation problems with generated regressors has grown significantly in recent years, see e.g. Mammen et al. (2012a, 2012b), Hahn and Ridder (2013), and Escanciano et al. (2014). The way we handle the generated regressor in our problem is

[2]Our baseline model is particularly related to Newey, Powell and Vella (1999).

mostly related to Mammen et al. (2012a). They use local linear regression to simplify technical arguments and focus on estimation of conditional mean with a generated regressor. We consider averages of cross partial derivatives of nonparametric regressions, and apply local polynomial regression to obtain estimates of these cross-partial derivatives. Lee (2013) also studies averages of such regression by using the kernel estimation. However, we have a different focus, and our main contribution is $\sqrt{n}$-consistency and asymptotic normality result for the parametric estimator in a partially linear model with a single instrument. A by-product gained in our estimation is that, as an intermediate step, the estimator of the average of the second order derivatives is shown to be $\sqrt{n}$-consistent. Li et al. (2003) show that the estimator of the average of first order derivatives converges at parametric rate. Our paper extends their results to the second order derivatives. Instead of using U-statistics, we apply empirical process methods to address in analyzing the asymptotic distribution of our estimator.

The paper is organized as follows. In Section 2, we introduce the model and provide conditions under which this model is identified. Our identification conditions are novel. For this reason we include a detailed discussion of our assumptions. We also discuss how our identification strategy can be extended to non-parametric models. Section 3 proposes the estimators and derives the asymptotic behavior of the estimator for $\gamma$. Section 4 provides an empirical study which illustrates how our method can be applied. We conclude in Section 5. The proof of the main results is deferred to the mathematical appendix.

# 2 The Model and the Identification

We start this section by discussing our basic identification strategy. We present our identification strategy in the context of our baseline model, first. At the end of this section we discuss how this strategy could be extended to a nonparametric model. The baseline model we study is given by

$$Y = \lambda(X) + \gamma D + \varepsilon\,, \tag{3}$$

$$X = \pi(Z) + V\,, \tag{4}$$

where $(Y, X, D, Z)$ is observed and $(\varepsilon, V)$ is unobserved. $Y$ denotes the outcome of interest. $X$ and $D$ denote the endogenous covariates, and $Z$ is an instrumental variable. $\varepsilon$ and $V$ are structural unobservables in the outcome and first stage equations, respectively. For ease of exposition, all of these random variables are assumed to be scalar. Note that without $\gamma D$ this model would be the same as the one considered in Newey, Powell and Vella (1999), and if $D$ was exogenous then it would be a special case of the model in Newey, Powell and Vella (1999). The presence of an additional endogenous regressor, namely $D$, makes the identification of this model harder.

Now we can state the first of our two main identification assumptions:

**Assumption 1** *Suppose* $\mathbb{E}(\varepsilon|X, V) = \mathbb{E}(\varepsilon|V)$.

Assumption 1 is commonly seen in control function approach literature and also imposed in Newey, Powell and Vella (1999). This assumption states that $V$ is a control function for $X$. In other words, this assumption says that controlling for $V$ controls for the endogeneity in $X$. Note, however, that controlling for $V$ does not necessarily control for the endogeneity in $D$. Under Assumption 1, we could write

$$\varepsilon = \rho(V) + \varsigma\,, \tag{5}$$

where $\rho(V) = \mathbb{E}(\varepsilon|V)$. Then, $\mathbb{E}(\varsigma|X,V) = 0$ by construction, and

$$Y = \lambda(X) + D\gamma + \rho(V) + \varsigma, \tag{6}$$

so that

$$Y - \mathbb{E}(Y|X,V) = \gamma[D - \mathbb{E}(D|X,V)] + \varsigma.$$

Since $D$ is still endogenous, we cannot use the method of Robinson (1988) to identify $\gamma$ if instrumental variables for $D - \mathbb{E}(D|X,V)$ exist. Therefore, we propose a new identification strategy. In particular, we first note that by Assumption 1 we have

$$\mathbb{E}(Y|X,V) = \lambda(X) + \gamma\mathbb{E}(D|X,V) + \rho(V),$$

which means that the only way $X$ and $V$ can interact is through $\gamma\mathbb{E}(D|X,V)$ suggesting that one could use the effect of this interaction on the outcome to identify $\gamma$, provided that $\mathbb{E}(D|X,V)$ has a term in which $X$ and $V$ interact. In particular, note that for $(x_i, v_j) \in \text{Supp}(X,V)$ for $i,j \in \{1.2\}$ we get

$$\begin{aligned}
&\mathbb{E}(Y|X=x_2,V=v_2) - \mathbb{E}(Y|X=x_1,V=v_2,) - \big[\mathbb{E}(Y|X=x_2,V=v_1) - \mathbb{E}(Y|X=x_1,V=v_1)\big] \\
&= \gamma\Big\{\big[\mathbb{E}(D|X=x_2,V=v_2) - \mathbb{E}(D|X=x_1,V=v_2)\big] - \big[\mathbb{E}(D|X=x_2,V=v_1) - \mathbb{E}(D|X=x_1,V=v_1)\big]\Big\}.
\end{aligned}$$

In this equation $V$ is varied from $v_1$ to $v_2$, and $x_1$ is varied from $x_1$ to $x_2$. Using this equation, we can immediately see that $\gamma$ will be identified if the following additional assumption holds:

**Assumption 2** *There exists $\mathcal{X}_1, \mathcal{X}_2, \mathcal{V}_1, \mathcal{V}_2$ such that $\mathcal{X}_1 \cap \mathcal{X}_2 = \emptyset$, $\mathcal{V}_1 \cap \mathcal{V}_2 = \emptyset$, and $\mathcal{X}_j \times \mathcal{V}_k \subseteq \text{Supp}(X,V)$, and $\mathbb{P}_{XV}(\mathcal{X}_j \times \mathcal{V}_k) > 0$ for each $j,k \in \{1,2\}$. Moreover,*

$$\begin{aligned}
&\mathbb{E}(D|X=x_2,V=v_2) - \mathbb{E}(D|X=x_1,V=v_2) \\
&\qquad\qquad \neq \mathbb{E}(D|X_1=x_2,V=v_1) - \mathbb{E}(D|X_1=x_1,V=v_1),
\end{aligned}$$

*whenever $x_j \in \mathcal{X}_j$ for $j=1,2$ and $v_k \in \mathcal{V}_k$ for $k=1,2$.*

This assumptions says that the change in the change in $\mathbb{E}(D|X=x,V=v)$ when we first change $x$ and then change $v$ (or vice versa) should be nonzero with positive probability. When this happens we can identify $\gamma$ as

$$\gamma = \frac{\mathbb{E}\Big[\Delta_{21}^x \Delta_{21}^v Y 1\{x_j \in \mathcal{X}_j, v_k \in \mathcal{V}_k, j=1,2, k=1,2\}\Big]}{\mathbb{E}\Big[\Delta_{21}^x \Delta_{21}^v D 1\{x_j \in \mathcal{X}_j, v_k \in \mathcal{V}_k, j=1,2, k=1,2\}\Big]}, \tag{7}$$

where

$$\begin{aligned}
\Delta_{21}^x \Delta_{21}^v Y &:= \mathbb{E}(Y|X=x_2,V=v_2) - \mathbb{E}(Y|X=x_1,V=v_2) \\
&- \big[\mathbb{E}(Y|X=x_2,V=v_1) - \mathbb{E}(Y|X=x_1,V=v_1)\big], \\
\Delta_{21}^x \Delta_{21}^v D &:= \mathbb{E}(D|X=x_2,V=v_2) - \mathbb{E}(D|X=x_1,V=v_2) \\
&- \big[\mathbb{E}(D|X=x_2,V=v_1) - \mathbb{E}(D|X=x_1,V=v_1)\big].
\end{aligned}$$

Once $\gamma$ is identified, we can identify $\lambda$ and $\rho$ up to constant term by a nonparametric regression of $Y - \gamma D$ on $X$ and $V$ imposing the restriction that $X$ and $V$ are additively separable.

The identification and estimation of the nonparametric part are quite standard and not the focus of our paper.

**Remark 1** *It is easy to see that $D$ could be a vector with dimension $J > 1$. With continuously distributed $X$ and $V$, if Assumption 2 is changed to*[3]*: $\mathbb{E}(\omega_D \omega_D^\intercal)$ has rank $J$, where $\omega_D = \left(\frac{\partial^2 \mathbb{E}(D_1|X,V)}{\partial X_1 \partial V}, ..., \frac{\partial^2 \mathbb{E}(D_J|X,V)}{\partial X_1 \partial V}\right)^\intercal$, then $\gamma$, which is now $J \times 1$ is identified.*

**Remark 2** *The same identification strategy goes through without any change if we were interested in the model given by equation (6) instead of the model given by equations (3) and (4), and $V$ is an observed covariate.*

Assumption 2 is a rank condition that replaces the relevance condition in the standard IV method. Like the relevance condition in the standard IV method, this assumption is testable. This condition requires that exogenous variations in the instrument should lead to variations in the endogenous variable in the model, which also leads to variations in the outcome. Here, we do not assume that there is a valid IV for $D$. Given the structure of the model, however, when we change $X$ and then $V$ the resulting change in $\mathbb{E}(Y|X,V)$ must be due to the change in $\mathbb{E}(D|X,V)$. Thus, we can only identify the coefficient on $D$ if the change in the conditional average of $D$ given $X$ and $V$ when $X$ and $V$ change sequentially is not zero.

In one particular situation it may not be reasonable to expect Assumption 1 and Assumption 2 to hold at the same time. This situation occurs if the model of interest is the one given by

[3]Here we assume that $(X,V)$ are continuously distributed only to be able to state the corresponding rank condition in a concise way.

equations (3) and (4) and the structural relationship between $X$ and $D$ is of the form

$$D = \mu(X) + \eta. \tag{8}$$

Here, $\eta$ is the structural unobservable that generated $D$. In empirical applications to argue that it is reasonable to expect Assumption 1 to hold one would try to argue that $Z$ is independent of $(\varepsilon, V)$, which is not a necessary, but a sufficient condition for Assumption 1. This is because arguing $E(\varepsilon|X,V) = E(\varepsilon|V)$ without $Z$ being independent of $\varepsilon$ and $V$ is hard. In contrast, to argue that Assumption 2 is reasonably expected to hold in an empirical context one has to argue that $\eta$ is not independent of $X$ even after we condition on $V$, and to make that argument one essentially has to argue that $\eta$ is not independent of $Z$. This means that to believe both Assumption 1 and Assumption 2 in an empirical context one would essentially believe that $Z$ is jointly independent of $\varepsilon$ and $V$, but not independent of $\eta$ which might be hard to believe.

In many other contexts this tension between Assumption 1 and Assumption 2 disappears. Before we explain why this is the case, we should point out that in the next subsection we discuss how our identification strategy can be extended to nonparametric settings, which allow the structural function relating $X$ and $D$ to $Y$ to be nonseparable in $(X, D)$. Thus, additive separability of $X$ and $D$ is not crucial to our identification strategy, and hence our method might be applicable to a wider range of situations than the baseline model suggests.

The tension between our two identifying assumptions disappears even if the model of interest is the one given in equations (3) and (4) in many empirical applications, like the ones we discuss in the Introduction[4]. In those empirical examples there is a strong reason to believe that the data generating process for $D$ is a non-separable function of $X$ and its data generating unobservable variable. For example, smoking during pregnancy has a positive mass at 0, but takes positive values as well (see Caetano (2015)). Similarly, Caetano and Maheshri (2016) provide strong evidence that police response is a non-linear function of last period's crime and cannot have an additively separable structural unobservable. On the other hand, if $D$ is binary, then the structural relationship between $X$ and $D$ has to be non-linear in $X$ in such a way that $\mathbb{E}(D|X,V)$ will have a term in which $X$ and $V$ interact as long as the structural unobservable in this relationship is not independent of $V$. In such cases, the tension between our two identifying assumptions disappears. To see this suppose that the structural function relating $X$ to $D$ relationship is of the form

$$D = h(X, \eta), \tag{9}$$

where $\eta$ denotes the structural unboservables as before. Suppose, also just for illustration purposes, that $X, \eta$ and $V$ are all continuous random variables and $h$ is continuously differentiable and the conditional density function of $\eta$ given $V$ is continuously differentiable in $V$. Then even if $X$ is independent of $\eta$ conditional on $V$, we have

$$\mathbb{E}(D|X = x, V = v) = \int h(x, e) f_{\eta|V}(e|v) \mathrm{d}e. \tag{10}$$

[4]We implement one of them in this paper.

Assuming we can interchange the order of differentiation and integration twice we would have

$$\frac{\partial^2 \mathbb{E}(D|X=x,V=v)}{\partial x \partial v} = \int \frac{\partial h(x,e)}{\partial x} \frac{\partial f_{\eta|V}(e|v)}{\partial v} \mathrm{d}e, \tag{11}$$

which will be different from 0 except in knife edge situations as long as $\frac{\partial h(x,e)}{\partial x}\frac{\partial f_{\eta|V}(e|v)}{\partial v} \neq 0$.[5]

In addition, suppose $Y = \theta(X, D) + \varepsilon$ and $D = h(X) + \eta$, with $Z$ independent of $(\varepsilon, V, \eta)$. In this case in any expansion of $\theta(X, D)$ we could still identify the coefficients of the terms in which $X$ and $D$ interact, and those might still be of interest.

Hoderlein, Su, White and Yang (2014) (HSWY), Lu and White (2014) (LW) and Lewbel, Lu and Su (2015) (LLS) all present new tests under the assumption that the unobservable variable(s) in the outcome equation are independent of all the covariates conditional on an observable variable that is excluded from the structural equation. Nothing in our method says that $V$ has to be unobserved. We can take

$$Y = \lambda(X) + \gamma D + \varepsilon,$$

as our structural equation. Then the structural function of interest is $\lambda(X) + \gamma D$, which does not depend on $V$.[6] With our notation, the HSWY, LW and LLS assumption says $\varepsilon$ is independent of $(X, D)$ conditional on $V$. As noted in HSWY, this assumption could be interpreted as: (i) an unconfoundedness assumption for the case in which $(X, D^\intercal)^\intercal$ is a vector of treatment variables; (ii) a proxy variable assumption, where $V$ is a proxy for $\varepsilon$. We make the weaker assumption $E(\varepsilon|V, X) = \rho(V)$, which says that the observable control only eliminates the endogeneity of $X$, but not of $D$.

If we interpret this as an unconfoundedness condition for $X$, then our Assumption 2 would hold if the equation for the other treatment variable has a term in which treatment variable $X$ and the observable control $V$ interact. For instance, suppose one is interested in the effects of Women Infants and Children (WIC) program participation as well as the effect of smoking during pregnancy on baby's birthweight.[7] If mother's income were observable, controlling for mother's income might eliminate the endogeneity in WIC participation, but it may not control for the unobserved heterogeneity in proneness to addiction. On the other hand, WIC participation may have a causal effect on smoking behavior of mothers during pregnancy. In this case Assumption 2 would be satisfied if the difference in the conditional probability (given WIC participation and maternal income) that a WIC participant pregnant woman smokes and a non-WIC participant pregnant woman smokes varies across different maternal income levels. If we interpret Assumption 1 as a proxy variable assumption, on the other hand, then Assumption 1 says that the observed

[5]The tension between our two identifying assumptions also disappears if $D = \mu(X, V) + \eta$ and $\mathbb{E}(\eta|X, V) = \mathbb{E}(\eta|V)$. However, convincingly arguing that the equation for $D$ is of this form might be hard outside a structural model.

[6]In Section 2.1.2 we outline identification of a non-parametric model in which the structural funtion depends on $V$.

[7]WIC is a federal program that provides aid to low income pregnant women and their young children.

variable $V$ is an imperfect proxy for $X$ (see, for example Wooldridge (2010) p.69). For example, when estimating a wage equation one might use a measure of cognitive skills, like an IQ score or an AFQT score (this would be the observable $V$), as a proxy for cognitive ability, which might deal with endogeneity in education/schooling. But if one of the variables whose effect is an indicator variable for being black ($D$), then because of possible discrimination in the job market, controlling for this particular $V$ may not eliminate endogeneity in $D$. If the fraction of blacks conditional on schooling, and, say AFQT score, varies differentially in AFQT score as we vary the schooling level, which is testable, then our Assumption 2 will be satisfied, and effects of both schooling and race will be identified.

## 2.1 Nonparametric Extensions:

### 2.1.1 Extension 1:

In this section we illustrate that separability of $X$ and $D$ is not crucial to our identification strategy. In particular, we outline how our identification strategy can be extended to the following nonparametric model:

$$Y = \theta(X, D) + \varepsilon, \tag{12}$$

$$X = m(Z, V), \tag{13}$$

where $V$ is scalar, continuous and normalized to be $Unif[0, 1]$ and $m$ is strictly increasing in $V$.[8] $Z$ is assumed to be independent of $(\varepsilon, V)$, so that $E(\varepsilon|X, V) = E(\varepsilon|V)$. In this model $X$ and $D$ can flexibly interact. Note that using Imbens and Newey (2009) we can identify $V$ as $F_{X|Z}(X|Z)$. First, suppose that

$$\theta(X, D) = \sum_{k=0}^{K} \alpha_k \psi_k(X, D),$$

where $K < \infty$ and $\psi_k$ are known functions of $X$ and $D$, but $\alpha_k$ are unknown parameters. Then

$$E(Y|X = x, V = v) = \sum_{j=0}^{K} \alpha_k \int \psi_k(x, d) \mathrm{d}F_{D|X,V}(d|x, v) + \rho(v),$$

and letting $m_Y(x, v)$ be this version of $E(Y|X = x, V = v)$ and assuming $m_Y(x, v)$ is twice continuously differentiable we have

$$\frac{\partial^2 m_Y(x, v)}{\partial x \partial v} = \sum_{j=0}^{K} \alpha_k \frac{\partial^2 \int \psi_k(x, d) \mathrm{d}F_{D|X,V}(d|x, v)}{\partial x \partial v}.$$

Then if $\mathbb{E}(\omega\omega^\intercal)$ is full rank, where

$$\omega_K := \left( \frac{\partial^2 \int \psi_1(x, d) \mathrm{d}F_{D|X,V}(d|x, v)}{\partial x \partial v}, \ldots, \frac{\partial^2 \int \psi_K(X, d) \mathrm{d}F_{D|X,V}(d|x, v)}{\partial x \partial v} \right)^\intercal$$

[8] The equation for $X$ could be extended to $X = h(Z, V)1\{h(Z, V) \geq 0\}$, or $X = h(Z, V)1\{c \geq h(Z, V) \geq 0\}$. See Caetano, Rothe and Yıldız (2016).

then $\alpha := (\alpha_1, ..., \alpha_K)^\intercal$ is identified as

$$\alpha = \left[\mathbb{E}\left(\omega_K \omega_K^\intercal\right)\right]^{-1} \mathbb{E}\left(\omega_K \frac{\partial^2 m_Y(X,V)}{\partial x \partial v}\right).$$

Note that if the coefficients of $\psi_k$ in which $X$ and $D$ interact and the coefficients of $\psi_k$ that only depend on $D$ are identified, then the sum of such $\alpha_k \psi_k$ can be subtracted from $Y$, after which coefficients on $\psi_k$ which only depend on $X$ as well as $\rho(v)$ can also be identified as in Newey, Powell and Vella (1999). Finally, this discussion suggests that our identification method could be extended to the case where $\theta(X, D)$ has an expansion of the form

$$\theta(X, D) = \sum_{k=0}^{\infty} \alpha_k \psi_k(X, D).$$

### 2.1.2 Extension 2:

In this section we discuss a non-parametric model in which an observable variable $V$ controls for endogeneity in $X$, but not in $D$, and $V$ itself has a structural effect on the outcome. We outline how parameters of interest in this model can be identified. In particular, it is possible to identify the interactive effects of $V$ with $X$ and/or $D$. Suppose

$$Y = \theta(X, D, V) + \tilde{\rho}_1(V) + \varepsilon, \tag{14}$$

with $E(\varepsilon|X, V) = \tilde{\rho}_2(V)$, and $(X, D, V)$ is observed. Here we assume that we are interested in the structural effects of $X$ and $D$. Then letting $\rho(V) = \tilde{\rho}_1(V) + \tilde{\rho}_2(V)$, we get

$$Y = \theta(X, D, V) + \rho(V) + \zeta. \tag{15}$$

As in the previous extension, suppose first that

$$\theta(X, D, V) = \sum_{k=0}^{K} \alpha_k \psi_k(X, D, V),$$

where $K < \infty$ and $\psi_k$ are known functions of $X$, $D$ and $V$, but $\alpha_k$ are unknown parameters. Then

$$E(Y|X = x, V = v) = \sum_{j=0}^{K} \alpha_k \int \psi_k(x, d, v) \mathrm{d}F_{D|X,V}(d|x, v) + \rho(v),$$

and letting $m_Y(x, v)$ be this version of $E(Y|X = x, V = v)$ and assuming $m_Y(x, v)$ is twice continuously differentiable we have

$$\frac{\partial^2 m_Y(x,v)}{\partial x \partial v} = \sum_{j=0}^{K} \alpha_k \frac{\partial^2 \int \psi_k(x, d, v) \mathrm{d}F_{D|X,V}(d|x, v)}{\partial x \partial v}.$$

Then if $\mathbb{E}\left(\omega \omega^\intercal\right)$ is full rank, where

$$\tilde{\omega}_K := \left(\frac{\partial^2 \int \psi_1(x, d, v) \mathrm{d}F_{D|X,V}(d|x, v)}{\partial x \partial v}, \ldots, \frac{\partial^2 \int \psi_K(x, d, v) \mathrm{d}F_{D|X,V}(d|x, v)}{\partial x \partial v}\right)^\intercal$$

then $\alpha := (\alpha_1, ..., \alpha_K)^\intercal$ is identified as

$$\alpha = \left[\mathbb{E}\left(\tilde{\omega}_K \tilde{\omega}_K^\intercal\right)\right]^{-1} \mathbb{E}\left(\tilde{\omega}_K \frac{\partial^2 m_Y(X,V)}{\partial x \partial v}\right).$$

This discussion suggests that that our identification method could be extended to the case where $\theta(X, D, V)$ has an expansion of the form

$$\theta(X, D, V) = \sum_{k=0}^{\infty} \alpha_k \psi_k(X, D, V).$$

# 3 Estimation

Let $m_Y(x, v)$ be a version of $\mathbb{E}(Y|X = x, V = v)$ and $m_D(x, v)$ be a version of $\mathbb{E}(D|X = x, V = v)$. $\frac{\partial^2 \widehat{m}_Y(x,v)}{\partial x \partial v}$ and $\frac{\partial^2 \widehat{m}_D(x,v)}{\partial x \partial v}$ will respectively denote estimators of cross partial derivative of $m_Y(x, v)$ and $m_D(x, v)$ with respect to $X$ and $V$ which use the values of $V$, and as such these will be infeasible estimators. In contrast, $\frac{\partial^2 \widehat{m}_Y(x,\hat{v})}{\partial x \partial v}$ and $\frac{\partial^2 \widehat{m}_D(x,\hat{v})}{\partial x \partial v}$ will respectively denote estimators of the same functions which use the values of $\hat{V}$, so these are the feasible estimators that we use in the estimation of $\gamma$. In this paper, these will be local polynomial estimators. Also for a vector of random variables $\mathbf{W}$, let $\mathcal{S}_{\mathbf{W}}$ denote the support of $\mathbf{W}$. In this section we assume that $(X, Z)$ is jointly distributed with joint density $f_{XZ}(x, z)$. For estimation purposes we strengthen Assumption 2 to

**Assumption 3** *Suppose $X$ and $V$ are continuously distributed, $m_D(x, v)$ is twice continuously differentiable and $\mathbb{E}\left[\frac{\partial^2 m_D(X,V)}{\partial x \partial v}\right] \neq 0$.*

We should note that if there is a set of $X, V$ values, say $E$, that has strictly positive probability and such that $\frac{\partial^2 m_D(x,v)}{\partial x \partial v} \neq 0$ on $E$, then we could instead estimate $\gamma$ by using values of $(x, v)$ that are in $E$ only.

The estimation procedure we propose consists of three stages,

1. We run local linear regression of $X$ on $Z$ for the first stage regression (4) to get $\hat{\pi}(\cdot)$. Subtracting $\hat{\pi}(Z)$ from $X$, we obtain the residual $\hat{V}$,

   $$\hat{V} = X - \hat{\pi}(Z),$$

   where $\hat{\pi}(Z) = \hat{\mathbb{E}}(X|Z) = \hat{\alpha}_X$ is defined later by (17).

2. We use local polynomial regression to obtain the estimators $\frac{\partial^2 \widehat{m}_Y(X,\hat{V})}{\partial x \partial v}$ and $\frac{\partial^2 \widehat{m}_D(X,\hat{V})}{\partial x \partial v}$, which are later defined by (18) and (19). These estimate the partial derivatives $\frac{\partial^2 m_Y(X,V)}{\partial x \partial v}$ and $\frac{\partial^2 m_D(X,V)}{\partial x \partial v}$ with the generated regressor $\hat{V}$.

3. Finally, we calculate the sample average for the unconditional moment using the estimators from step two,

$$\hat{\gamma} = \frac{\hat{\mathbb{E}}\left[\frac{\partial^2 \widehat{m}_Y(X,\hat{V})}{\partial x \partial v}\right]}{\hat{\mathbb{E}}\left[\frac{\partial^2 \widehat{m}_D(x,\hat{V})}{\partial x \partial v}\right]}, \tag{16}$$

where $\hat{\mathbb{E}}[\cdot]$ is the sample average.

As we can see, the major challenge here is to solve how the generated regressor influences our final estimator. Before we formally define the local linear estimator and local polynomial estimators listed above, we first introduce the assumptions imposed. We impose the following regularity conditions:

**Assumption 4** *$(Y_i, X_i, Z_i)$ are independent and identically (i.i.d.) distributed as $(Y, X, Z)$, where $Y, X$ and $Z$ are all scalar random variables. Suppose $(X, Z)$ has compact support $\mathcal{S}_{XZ} \subset \mathcal{R}^2$.*

**Assumption 5** *(i) $m_Y(X, V)$ and $m_D(X, V)$ have continuous derivatives of total order $p + 1$. Let $f_{X,V}(x, v)$ denote the density function of $(X, V)$. $f_{X,V}(\cdot)$ also has continuous derivatives of total order 3, and $\inf_{(x,v)\in\mathcal{S}_{XV}} f_{X,V}(x, v) \geq \eta$ for some $\eta > 0$; (ii) $\pi(Z)$ is twice continuously differentiable.*

**Assumption 6** *The kernel function $K(\cdot)$ is a non-negative, twice continuously differentiable function with a compact support and satisfies $\int K(u)\,du = 1$ and $\int uK(u)\,du = 0$. For any multivariate vector $\nu \in \mathcal{R}^{d_\nu}$, define $K(\nu) = K(\nu_1)K(\nu_2)\cdots K(\nu_{d_\nu})$.*

Assumption 4 is very standard in the literature, and the i.i.d assumption can be relaxed to weakly dependence, but that would require more complicated derivations. Compactness of the support of $(X, Z)$ is not strictly required. We are simply doing estimation over a compact subset of this support. Assumption 5 assumes some smoothness condition on both the conditional expectation and the joint density function of $(X, V)$; such smoothness requirements are widely used in local polynomial regression literature. Assumption 5 also requires the joint density of $X$ and $V$ to be bounded away from 0. Given that $\pi$ is assumed to be smooth, compact subsets of the support of $(X, Z)$ are mapped to compact subsets of the support of $(X, V)$. Then the assumption that the joint density of $(X, V)$ is bounded away from 0 will follow if we assume that both the joint density of $(X, Z)$ and the Jacobian of $\pi$ are bounded away from 0 over the compact and subset of the support of $(X, Z)$ over which the estimation is done. On the other hand, since $X$ is observed and $V$ is identified, the assumption that the joint density of $(X, V)$ is bounded away from 0 is testable. On the other hand, one could use trimming functions as in Klein and Spady (1993) or as in Heckman, Ichimura and Todd (1998) to avoid this assumption.

Assumption 6 is very standard in nonparametric estimation. Here we state the assumption for any kernel used in the remainder of the paper. We use product kernels whenever there

are multiple (continuous) conditioning variables. There are several widely used kernel functions satisfying Assumption 6, such as bi-weight kernel.

Here and in the sequel, denote $e_t$ as a $\mathbf{N}$-dimensional[9] unit vector with 1 at the $t^{th}$ argument. Let $\frac{\partial^2 \widehat{m}_Y(x,\hat{v})}{\partial x \partial v} = e_5^\intercal \hat{\beta}_Y$ and $\frac{\partial^2 \widehat{m}_D(x,\hat{v})}{\partial x \partial v} = e_5^\intercal \hat{\beta}_D$. We have $\hat{\alpha}_i$ and $\hat{\beta}_i$ $(i = X, D, Y)$ solve the following optimization

$$(\hat{\alpha}_X, \hat{\beta}_X) = argmin_{\alpha,\beta} \sum_{i=1}^{n} (X_{1i} - \alpha - \beta(Z_i - z))^2 K_g(Z_i - z)\,, \tag{17}$$

$$(\hat{\alpha}_Y, \hat{\beta}_Y) = argmin_{\alpha,\beta} \sum_{i=1}^{n} (Y_i - \alpha - \sum_{0 \le |u| \le p} \beta^\intercal (X_i - x, \hat{V}_i - v)^u)^2 K_h((X_i - x, \hat{V}_i - v))\,, \tag{18}$$

$$(\hat{\alpha}_D, \hat{\beta}_D) = argmin_{\alpha,\beta} \sum_{i=1}^{n} (D_i - \alpha - \sum_{0 \le |u| \le p} \beta^\intercal (X_i - x, \hat{V}_i - v)^u)^2 K_h((X_i - x, \hat{V}_i - v))\,. \tag{19}$$

respectively, where $\sum_{0 \le |u| \le p}$ denotes the summation over all nonnegative integer vector $u = (u_1, \cdots, u_{d_u})$ with $|u| = \sum_{l=1}^{d_u} u_l$. Also, for any vector $w = (w_1, \cdots, w_{d_u})$, $w^u = (w_1^{u_1}, \cdots, w_{d_u}^{u_{d_u}})$, $K_g(w) = g^{-d_u} K(w/g)$, $K_h(w) = h^{-d_u} K(w/h)$. Here, $g$ and $h$ are bandwidths in first and second stage estimation.

By applying (7) and direct calculation, we expand $\hat{\gamma} - \gamma$ as below,

$$\frac{\left\{\hat{\mathbb{E}}\left[\frac{\partial^2 \widehat{m}_Y(X,\hat{V})}{\partial x \partial v}\right] - \mathbb{E}\left[\frac{\partial^2 m_Y(X,V)}{\partial x \partial v}\right]\right\} \mathbb{E}\left[\frac{\partial^2 m_D(X,V)}{\partial x \partial v}\right] - \left\{\hat{\mathbb{E}}\left[\frac{\partial^2 \widehat{m}_D(X,\hat{V})}{\partial x \partial v}\right] - \mathbb{E}\left[\frac{\partial^2 m_D(X,V)}{\partial x \partial v}\right]\right\} \mathbb{E}\left[\frac{\partial^2 m_Y(X,V)}{\partial x \partial v}\right]}{\hat{\mathbb{E}}\left[\frac{\partial^2 \widehat{m}_D(X,\hat{V})}{\partial x \partial v}\right] \mathbb{E}\left[\frac{\partial^2 m_D(X,V)}{\partial x \partial v}\right]}\,. \tag{20}$$

To show that $\hat{\gamma}$ converges to $\gamma$ at the parametric rate, the major step is to demonstrate that

$$\sqrt{n}\left\{\hat{\mathbb{E}}\left[\frac{\partial^2 \widehat{m}_Y(X,\hat{V})}{\partial x \partial v}\right] - \mathbb{E}\left[\frac{\partial^2 m_Y(X,V)}{\partial x \partial v}\right]\right\} = O_P(1)$$

and also $\sqrt{n}\left\{\hat{\mathbb{E}}\left[\frac{\partial^2 \widehat{m}_D(X,\hat{V})}{\partial x \partial v}\right] - \mathbb{E}\left[\frac{\partial^2 m_D(X,V)}{\partial x \partial v}\right]\right\} = O_P(1)$. As it is easy to see that the two identities can be proved by the same method, to save space, we only state the result for the first one here in Proposition 1. The result for the other is presented in the Appendix as Corollary A.1.

To apply empirical process methods, we need to restrict the complexity of the smooth class that the cross-partial derivative of the expectation belongs to. We use the smooth class defined in van der Vaart and Wellner (1996).

**Definition** $C_M^\alpha(\mathcal{S})$**:** For any vector $s = (s_1, \cdots, s_{d_s})$ in a compact set $\mathcal{S} \in \mathcal{R}^{d_s}$ and any integer vector $q = (q_1, \cdots, q_d)$, let $D^q$ denote the differential operator $D^q = \frac{\partial^{|q|}}{\partial}$, where $|q| = \sum_{l=1}^{d} q_l$. Let $\underline{\alpha}$ be the greatest integer smaller than $\alpha$. Define

$$\|g\|_\alpha = \max_{|q| \le \underline{\alpha}} \sup |D^q g(s)| + \max_{|q| \le \underline{\alpha}} \sup_{s \ne s'} |D^q g(s) - D^q g(s')| / \|s - s'\|^{\alpha - \underline{\alpha}}\,,$$

[9] $\mathbf{N}$ is given by (32) in the Appendix.

where the supremum is taken over the interior of $\mathcal{S}$. Then $C_M^{\alpha}(\mathcal{S})$ is the set of all continuous functions $g : \mathcal{S} \mapsto \mathcal{R}$ with $\|g\|_{\alpha} \leq M$.

Before we state our results, we impose the following assumptions.

**Assumption 7** *For bandwidth $g$ and $h$, as $n \to \infty$, we have (i)$\frac{\log n}{\sqrt{n}g} \to 0$ and $\sqrt{n}g^4 \to 0$; (ii) $nh^{2p} \to 0$ and $\frac{nh^6}{\log n} \to \infty$. (iii) $ngh^8 \to \infty$.*

**Assumption 8** *$\frac{\partial^2 m_Y(\cdot,\cdot)}{\partial x \partial v}, \frac{\partial^2 m_D(\cdot,\cdot)}{\partial x \partial v} \in C_M^{\alpha}(\mathcal{S}_{XV})$, where $\underline{\alpha} = 2$.*

**Assumption 9** *Let $u := D - \mathbb{E}(D|X,V)$, and $\xi := \varsigma + \gamma u$. $\mathbb{E}[\exp(l|\xi|)|X,V] \leq C$ almost surely for a constant $C > 0$ and $l > 0$ small enough.*

Assumption 7 provides the conditions on bandwidths $h$ and $g$. The second part of this assumption is comparable to the corresponding assumption in Li et al. (2003). The first and third part are due to reasons similar to the ones in Corollary 6 in Mammen et al. (2012a). Assumption 7(i) is about the first stage local linear estimation, $\sqrt{n}g^4 \to 0$ part can be relaxed if a higher order kernel is used[10]. Assumption 8 and Assumption 9 are also used by Mammen et al. (2012a) to apply empirical process theory. Assumption 8 is for the stochastic equicontinuity argument in empirical process theory. It implies $\frac{\partial^2 m_Y(X,V)}{\partial x \partial v}$ and $\frac{\partial^2 m_D(X,V)}{\partial x \partial v}$ belong to a Donsker class by Example 19.9 in van der Vaart (2000). The smooth class is not limited to the one we choose, and this assumption can be replaced by imposing a condition on the bracket entropy of the class directly. Assumption 7 and 8 together will ensure the cccuracy and complexity assumptions in Mammen et al. (2012a) are satisfied, which makes some of their results are applicable to our model.

For any matrix $A$, $[A]_{i,j}$ represents the $(i,j)$ entry of matrix $A$. We have the following result with the generated regressor $\hat{V}$,

**Proposition 1:** Under Assumptions 1-9, we have

$$\sqrt{n}\left\{\hat{\mathbb{E}}\left[\frac{\partial^2 \widehat{m}_Y(X,\hat{V})}{\partial x \partial v}\right] - \mathbb{E}\left[\frac{\partial^2 m_Y(X,V)}{\partial x \partial v}\right] - h^{p-1}\text{Bias}_Y\right\} \xrightarrow{d} N(0,\Lambda)\,,$$

where

$$\begin{aligned}\Lambda = {} & 4\mathbb{E}\{\sigma_{\xi}^2[R(X,V)]_{5,1}^2\} + \left(\mathbb{E}\left[\frac{\partial^2 m_Y(X,V)}{\partial x \partial v}\right]\right)^2 + \mathbb{E}\left\{V^2\mathbb{E}_{X|Z}\left[\frac{\partial^3 m_Y(X,V)}{\partial x \partial v^2}\right]^2\right\} \\ & - 2\mathbb{E}\left\{V\mathbb{E}_{X|Z}\left[\frac{\partial^3 m_Y(X,V)}{\partial x \partial v^2}\right]\frac{\partial^2 m_Y(X,V)}{\partial x \partial v}\right\}, \\ \text{Bias}_Y = {} & 2e_5^{\intercal}M^{-1}B\mathbb{E}[m_Y^{(p+1)}(X,V)]\,,\end{aligned}$$

[10]Suppose $g = n^{-a}$ and $h = n^{-b}$. Then Assumption 7(i) requires $1/8 < a < 1/2$. The first part of Assumption 7(ii) and Assumption 7(iii) together require that $1/(2p) < b < (1-a)/8$.

$M$, $B$, $R(x,v)$ and $m_Y^{(p+1)}(x,v)$ are defined by (33), (34), (36) and the line above (36) respectively in the Appendix. The bias term comes from estimation of $\frac{\partial^2 m_Y(X,V)}{\partial x \partial v}$. Thus, even if $V$ was observed as opposed to generated we would still have this bias term. Similarly, even if $V$ was observed we would have the first two terms in $\Lambda$. The third term in $\Lambda$ appears because $V$ is estimated and we have to differentiate $\frac{\partial^2 m_Y(X,V)}{\partial x \partial v}$ with respect to $V$ to deal with this fact. The last term in $\Lambda$ appears because of the covariance of this additional term in the expansion and $\frac{\partial^2 m_Y(X_i,V_i)}{\partial x \partial v}$.

As we can see, this result is about the estimation of the average of second order derivatives. In Li et al. (2003), they show that the estimator of the average of first order derivatives converges at parametric rate. In Proposition 1, we obtain the $\sqrt{n}$-consistent result for the estimator of the average of second order derivatives as a by-product, which further extends the existing literature. This propositon may also be useful in estimation of Slutsky matrices or marginal rates of transformation in estimation of production functions[11].

In the Appendix, we state a result that is analogous to the one in Proposition 1 describing the asymptotic behavior of $\hat{\mathbb{E}}\left[\frac{\partial^2 \widehat{m}_D(X,\hat{V})}{\partial x \partial v}\right]$. These two results combined with the specific form of the estimator gives us the theorem below. This is our main result and states that $\hat{\gamma}$ converges at the parametric rate.

**Theorem 1:** Under Assumptions 1-9, we have

$$\sqrt{n}(\hat{\gamma}-\gamma) \xrightarrow{d} N\left(0, \frac{1}{\left(\mathbb{E}\left[\frac{\partial^2 m_D(X,V)}{\partial x \partial v}\right]\right)^2} 4\sigma_\varsigma^2 \mathbb{E}\{([R(X,V)]_{5,1})^2\}\right), \tag{21}$$

where $\sigma_\varsigma^2$ is the variance of $\varsigma$. Recall that $\varsigma = \varepsilon - \mathbb{E}(\varepsilon|X,V) = \varepsilon - \rho(V)$. $[R(x,v)]_{5,1}$ is the $(5,1)$ entity in the matrix $R(x,v)$ defined by equation (36) in the Appendix. The $R(x,v)$ matrix depends on the kernel function as well as $f_{X,V}(x,v)$ and its first and second order derivatives.

Each piece in the expression for the asymptotic variance of our estimator can be estimated by corresponding sample quantities. For example, $\mathbb{E}\left[\frac{\partial^2 m_D(X,V)}{\partial x \partial v}\right]$ can be estimated by $\sum_{i=1}^n \frac{1}{n}\frac{\partial^2 \hat{m}_D(X_i,\hat{V}_i)}{\partial x \partial v}$. Similarly, $\mathbb{E}\{([R(X,V)]_{5,1})^2\}$ can be estimated by $\frac{1}{n}\sum_{i=1}^n \left([\hat{R}(X_i,\hat{V}_i)]_{(5,1)}\right)^2$, where $\hat{R}(X_i,\hat{V}_i)$ is a matrix defined analogously to $R(X,V)$ in which $f_{X,V}(x,v)$ and its first and second order derivatives are replaced by their corresponding consistent estimators, which, of course use the generated regressor values $\hat{V}_i$. Finally, $\sigma_\varsigma^2$ can be estimated by $\frac{1}{n}\sum_{i=1}^n \hat{\varsigma}_i^2$, where $\hat{\varsigma}_i$ are the residuals obtained from a non-parametric regression of $Y - D\hat{\gamma}$ on $X$ and $\hat{V}$.

The result indicates that our estimator has no bias terms from the local polynomial regression. This desirable property comes from the special ratio form of the estimator we proposed and also the equality (7), as the bias terms from the estimation of the denominator and numerator exactly cancel each other.

[11] We would like to thank an anonymous referee for this suggestion.

$\lambda(\cdot)$ and $\rho(\cdot)$ can be estimated by regressing $Y - \hat{\gamma}D$ on $X, \hat{V}$ as in Mammen et al. (2012a). As $\hat{\gamma}$ converges at parametric rate, the result in Mammen et al. (2012a) holds. Therefore, we can apply Mammen et al. (2012a) directly.

# 4 Empirical Example

## 4.1 Background

Smoking during pregnancy and its causal effect on birth outcomes has been a long standing issue in both medical and policy circles. Medical literature has established a link between maternal smoking and adverse birth outcomes, primarily measured by the effect on average birth weights. The most comprehensive treatment of the question is provided by Almond et al. (2005), under a selection-on-observables framework, and estimates a significant difference in birth weights of around 200 grams between smokers and non-smokers. However, the primary concerns involved in estimating the causal effect of maternal smoking on birth weights is the lack of an exogenous source of variation to help deal with potential selection concerns between smokers and non-smokers. It is particularly hard to find a source of variation that is correlated with smoking during pregnancy but uncorrelated with unobserved 'mothering ability'. As Caetano (2015) argues even after using the most comprehensive set of controls, as in Almond et. al. (2005), there is significant left over selection that hinders meaningful causal inference. The framework developed in this paper can potentially help contribute to this persistent problem in the literature.

To formulate the empirical setup more precisely, consider the question explored by Currie and Moretti: the effect of maternal education on birth outcomes. Their setup can be expressed in terms of the standard case of one endogenous variable, mother's education attainment, and one instrumental variable, the number of colleges in the county where the mother was resident at age 17.[12] Applying this setup to our question we can have two potentially endogenous variables of interest in our structural equation: maternal education and maternal smoking, and an IV for the former. Thus using the framework developed so far in this paper, we can use the artificially generated exogenous source of variation to estimate the causal effect of smoking on birth weight.

## 4.2 Data

We use Vital Statistics data from 1993-2002 and closely follow the sample restrictions used by Currie and Moretti. In particular, we restrict our sample to all singleton births to White mothers aged between 24 and 45 years. We also drop mothers that reside in counties with less than 100,000 population and where we lack information on the instrument. This leaves us with a sample size

[12] In their paper, CM actually include separate IVs for the number of four-year colleges and the number of two-year colleges in the county of the mother at age 17.

of close to 2.2 million births. The key variables of interest for us are the two endogenous variables in our framework, the number of years of schooling of the mother, $X$, and the reported number of cigarettes smoked by the mother during pregnancy, $D$. Following the previous literature, our key birth outcome of interest is the birth weight of an infant in grams, $Y$.[13] Our instrument, $Z$, is the total number of four-year and two-year colleges in the county where the mother resides at age 17.[14]

## 4.3 Estimation

The first step of the estimation calculates the control function $V$ using a local linear regression with maternal education, $X$ regressed on the instrument $Z$.[15] In the next step, we estimate conditional expectations of both $Y$ and $D$ given $X$ and $V$ using a degree three bivariate local polynomial regression.[16] Using the estimation results we can construct, $\frac{\partial^2 E(Y|X,V)}{\partial X \partial V}$ and $\frac{\partial^2 E(D|X,V)}{\partial X \partial V}$, respectively.[17] The final step then takes the ratio of the two sample averages as given in equation (16) to calculate $\hat{\gamma}$, the effect of maternal smoking on birth weight.

## 4.4 Results

This subsection presents the results from the above outlined procedure to estimate the causal effect of smoking during pregnancy on birth weight. The first column of table 1 reports OLS results using the basic set of controls described above and estimates an effect of -17.61 grams per cigarette smoked during pregnancy.[18] Specification - II then uses a much more detailed set of covariates including demographics of the parents, pregnancy characteristics, pregnancy history, and various interactions between them. This specification closely follows the most comprehensive

[13] In our empirical setting we follow the baseline model given by equation (3) and (4), and remain in a partially linear setup. Given that the effect of smoking during pregnancy on birth weight is primarily physiological, we do not expect an interaction between smoking and education in our structural equation. We thank an anonymous referee for stressing on this point.

[14] We are extremely grateful to Janet Currie and Enrico Moretti for graciously providing us with their novel dataset on college openings in the US.

[15] In a preliminary step maternal education $X$ is regressed on maternal age dummies, 10-year birth of cohort dummies and mother's county-at-age-17*year of birth fixed effects, and the residual from this regression is used in the local linear regression to calculate $V$. In addition, we also restrict to mothers with education levels between 7 and 17 years of schooling to hone on to the sample for which the openings of new colleges is likely to be the most relevant.

[16] This birth weight variable is cleaned off the variation solely due to maternal age dummies, 10-year birth cohort dummies and mother's county-at-age-17*year of birth fixed effects.

[17] Given the massive size of our dataset we have 227,880 unique combinations of $X$ and $V$ at which the local polynomial regressions have to be evaluated. We therefore, run the local polynomial regression at 10% of the total number of unique combinations. Results are robust even when we increase this to 20% for select combinations of bandwidths of $X$ and $V$.

[18] We follow Currie and Moretti closely for this specification.

Table 1: Effect of Maternal Smoking on Birth Weight

| | OLS | | Local Polynomial | |
|---|---|---|---|---|
| | Spec - I | Spec - II | $h_X = 3; h_{\hat{V}} = 1$ | $h_X = 5; h_{\hat{V}} = 3$ |
| Cigarettes Smoked | $-17.61^{**}$ | $-16.08^{**}$ | $-26.22^{**}$ | $-24.47^{**}$ |
| | (0.198) | (0.183) | (1.302) | (0.050) |
| Years of Schooling | $18.20^{**}$ | $12.27^{**}$ | – | – |
| | (0.631) | (0.644) | | |
| Number of Observations | 2,257,460 | 2,257,460 | 2,257,460 | 2,257,460 |

**,* Indicates significance at 1, and 5 percent, respectively. OLS specification - I includes non-parametric controls for mother's age, dummies for 10-year birth cohort and mother's county-at-age-17*year of birth fixed effect. Specification - II controls for a more elaborate set of controls with various interactions and closely follows the one used by Almond et. al. The last two columns then present estimates using our methodology and employ a bi-variate local polynomial of degree 3. Standard errors are calculated as per Theorem 1.

one used in the literature so far by Almond et. al. We see a slight reduction of 1.5 grams in the effect of each cigarette smoked on birth weight using the full set of controls.

The last two columns finally present results from the methodology developed in this paper, implementing the estimation details given in the above section. Using a degree 3 bivariate local polynomial estimator, we find an effect size of $-26.22$ grams per cigarette smoked, which is significantly larger in magnitude compared to the OLS estimates. We use different combinations for the bandwidth of $X$ and $V$ and report two of those in table 1. Table 2 then documents nine different combinations of the two bandwidths. These reported results are also robust to running a degree 5 polynomial instead of 3 and by using different bandwidths for the first stage estimation of the control function $V$.

One might expect the OLS estimates to actually decrease in magnitude given that mother's who smoke are more likely to be selected negatively on unobservables. However, our estimation methodology is inherently instrument variable based, and hence recovers only a local average treatment effect (LATE). Using terminology from the treatment effects literature, the group of 'compliers' for our IV are mothers who are more likely to go to college as a result of more college openings in their county of residence. These mothers, in turn, are also more likely to be positively selected on other unobservable dimensions which could be positively correlated with mothering ability. This, therefore, can account for the increase in magnitude in effect sizes that we document using our methodology compared to the OLS estimates.

Table 2: Effect of Maternal Smoking on Birth Weight -

| | Bandwidth ($h_X$) | | |
|---|---|---|---|
| Bandwidth ($h_{\hat{V}}$) | 3 | 4 | 5 |
| 1 | $-26.22^{**}$ | $-26.26^{**}$ | $-25.96^{**}$ |
| | (1.302) | (1.615) | (0.093) |
| 2 | $-27.09^{**}$ | $-26.12^{**}$ | $-25.97^{**}$ |
| | (0.396) | (0.238) | (0.054) |
| 3 | $-26.00^{**}$ | $-24.85$ | $-24.47^{**}$ |
| | (0.304) | (0.119) | (0.050) |

** Indicates significance at the 1% level. Each estimate of $\gamma$ is calculated using a degree bivariate local polynomial specification. Given the large sample size of our data we evaluate the local polynomial regressions at a 10% random sample of the total number of unique combinations of $X$ and $V$. These leave us with 22,786 evaluations each for $\frac{\partial^2 E(Y|X,V)}{\partial X \partial V}$ and $\frac{\partial^2 E(D|X,V)}{\partial X \partial V}$, respectively. Standard errors are calculated as per Theorem 1.

# 5 Conclusion

In this paper, we propose a novel identification method which delivers identification of marginal effects of multiple endogenous variables, with a single instrument that could even be binary. Our baseline model is partially linear. We also outline how our identification method can be extended to show identification in two non-parametric models. These extensions further highlight the identifying power of the assumptions we make. While these assumptions are able to deliver identification of parameters of interest in models with multiple endogenous variables with few, even a single instrument, in applying our method, one has to discuss carefully the plausibility of our identifying assumptions in the context of the application at hand. We would like to stress that our proposed methodology does not provide a panacea for the lack of an available instrumental variable, but instead provides an additional methodology to the toolkit of empirical researchers. For our baseline model we develop a new and simple estimator for the coefficient of the endogenous regressor(s), $D$, for which no instrument is available. We provide the asymptotic behavior of this estimator for the case when $D$ is scalar, and show that the proposed estimator is $\sqrt{n}$-normal. A by-product of our method is an $\sqrt{n}$-consistent estimator of the average of second order derivatives, which is also a new result in the literature to our knowledge. In the empirical section, we use the methods proposed in this paper to assess the causal impact of mother's smoking during pregnancy on the infant's birth weight. Our results seem to be in line with findings of the previous literature. We leave detailed analysis of the non-parametric extensions we outlined for future research.

# References


[1] Ai, C. and X. Chen (2003). Efficient Estimation of Models with Conditional Moment Restrictions Containing Unknown Functions. *Econometrica* **71** (6), 1795-1843.

[2] Almond Douglas, Kenneth Y. Chay and David S. Lee (2005). The Costs of Low Birth Weight. *Quarterly Journal of Economics* **120** (3), 1031-1083.

[3] Blundell, R. W. and J.L. Powell (2003). Endogeneity in Nonparametric and Semiparametric Regression Models. *Advances in Economics and Econometrics: Theory and Applications*, Eighth World Congress, Vol. II. Cambridge: Cambridge University Press.

[4] Blundell, R. and J. Powell (2004). Endogeneity in Semiparametric Binary Response Models. *Review of Economic Studies* **71**, 655-679.

[5] Caetano, C. (2015). A Test of Endogeneity without Instrumental Variables in Models with Bunching. *Econometrica* **83**, 1581-1600.

[6] Caetano, C. and J. C. Escanciano (2015). Identifying Multiple Marginal Effects with a Single Binary Instrument or by Regression Discontinuity. *Unpublished Manuscript*.

[7] Caetano, C. and J. C. Escanciano (2016). Identifying Marginal Effects Using Covariates. *Unpublished Manuscript*.

[8] Caetano, G. and V. Maheshri (2016). Identifying Dynamic Spillovers of Crime with a Causal Approach to Model Selection. *Forthcoming, Quantitative Economics*.

[9] Chen, X. and D. Pouzo (2009). Efficient Estimation of Semiparametric Conditional Moment Models with Possibly Nonsmooth Residuals. *Journal of Econometrics* **152**, 46-60.

[10] Currie, Janet and Enrico Moretti (2003). Mother's Education and the Intergenerational Transmission of Human Capital: Evidence from College Openings. *Quarterly Journal of Economics* **118** (4), 1495-1532.

[11] D'Haultfoeuille, X. and P. Février (2015). Identification of Nonseparable Triangular Models With Discrete Instruments. *Econometrica* **83**, 1199-1210.

[12] Escanciano, J.C., D.T. Jacho-Chávez and A. Lewbel (2012). Identification and Estimation of Semiparametric Two Step Models. *Unpublished manuscript*.

[13] Escanciano, J.C., D.T. Jacho-Chávez, and A. Lewbel (2014). Uniform Convergence of Weighted Sums of Non and Semiparametric Residuals for Estimation and Testing. *Journal of Econometrics* **178** (P3), pages 426-443.

[14] Florens, J.-P. (2003). Inverse Problems and Structural Econometrics: the Example of Instrumental Variables. *Advances in Economics and Econometrics: Theory and Applications*, Eight World Congress, Econometric Society Monograph Series, ESM 36, Volume II, 284-312. Cambridge: Cambridge University Press.

[15] Florens, J.-P., J. Johannes and S. Van Bellegem (2012). Instrumental Regression in Partially Linear Models. *The Econometrics Journal* **15** (2), 304-324.

[16] Hahn, J. and G. Ridder (2013). The Asymptotic Variance of Semi-parametric Estimators with Generated Regressors. *Econometrica* **81**, 315-340.

[17] Heckman, J. J., H. Ichimura, and P. Todd (1998). Matching as an Econometric Evaluation Estimator. *Review of Economic Studies* **65**, 261-294.

[18] Hoderlein, S., L. Su, H. White and T. Yang (2014). Testing for Monotonicity in Unobservables under Unconfoundedness. *Working Paper*, Dept. of Economics, Boston College.

[19] Huang, L. and N. Yıldız (2013). Identification of Causal Effects in a Triangular Model with Endogenous Regressors. *Camp Econometrics NY VIII Poster*.

[20] Imbens, G. W. and W.K. Newey (2009). Identification and Estimation of Triangular Simultaneous Equations Models Without Additivity. *Econometrica* **77**, 1481-1512.

[21] Jacob, B., L. Lefgren and E. Moretti (2007). The Dynamics of Criminal Behavior: Evidence from Weather Shocks *The Journal of Human Resources* **42**, 489-527.

[22] Klein, R. W. and R. H. Spady (2013). An Efficient Semiparametric Estimator for Binary Response Models. *Econometrica* **61**, 387-421.

[23] Kramer (1987). Determinants of Low Birth Weight: Methodological Assessment and Meta-analysis. *Bull World Health Organ* **65**, 663-737.

[24] Lee (2013). Partial Mean Processes with Generated Regressors: Continuous Treatment Effects and Nonseparable Models. *Working paper*.

[25] Lewbel, A., X. Lu and L. Su (2015). Specification Testing for Transformation Models with an Application to Generalized Accelerated Failure-Time Models. *Journal of Econometrics* **184**, 81-96.

[26] Li, Q., Lu and Ullah (2003). Multivariate Local Polynomial Regression for Estimating Average Derivatives. *Journal of Nonparametric Statistics* 15:4-5, 607-624.

[27] Lu, X. and H. White (2014). Testing for Separability in Structural Equations. *Journal of Econometrics* **182**, 14-26.

[28] Mammen, E., C. Rothe, and M. Schienle (2012a). Nonparametric Regression with Nonparametrically Generated Covariates. *Annals of Statistics* **40**, 1132-1170.

[29] Mammen, E., C. Rothe, and M. Schienle (2012b). Semiparametric Estimation with Generated Covariates, forthcoming in *Econometric Theory*.

[30] Masry, E. (1996a). Multivariate Regression Estimation Local Polynomial Fitting for Time Series. *Stochastic Processes and Their Applications* **65**, 81-101.

[31] Masry, E. (1996b). Multivariate Local Polynomial Regression for Time Series: Uniform Strong Consistency and Rates. *Journal of Time Series Analysis* **17**, 571-599.

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

[33] Robinson, P.M. (1988). Root-N-consistent Semiparametric Regression. *Econometrica* **56** (4), 931-954.

[34] Speckman, P. (1988). Kernel Smoothing in Partial Linear Models. *Journal of the Royal Statistical Society. Series B (Methodological)*, 413-436.

[35] Tominey, Emma (2007). Maternal Smoking during Pregnancy and Early Child Outcomes. *CEP Discussion Papers*. Centre for Economic Performance, LSE.

[36] Torgovitsky, T. (2015). Identification of Nonseparable Models Using Instruments with Small Support. *Econometrica* **83**, 1185-1197.

[37] van der Vaart, A. (2000). *Asymptotic Statistics*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press.

[38] van der Vaart, A., and J. Wellner (1996). *Weak Convergence and Empirical Processes: with Applications to Statistics*. Springer Verlag.

[39] Wooldridge, J. M. (2010). *Econometric Analysis of Cross Section and Panle Data*, 2nd ed., The MIT Press.

# A Mathematical Proof

We first introduce the definition of empirical measure that is commonly used in empirical process theory. For any $X_1, \cdots, X_n$ that are i.i.d. random variables with distribution $P$ and for any measurable function $\phi$, the empirical measure $G_n$ is defined as

$$\phi \mapsto G_n\phi = \sqrt{n}(P_n - P)\phi = \sqrt{n}\left[\frac{1}{n}\sum_{i=1}^{n}\phi(X_i) - \mathbb{E}\phi\right]. \tag{22}$$

We first prove Proposition 1, which indicates that $\sqrt{n}\left\{\hat{\mathbb{E}}\left[\frac{\partial^2 \widehat{m}_Y(X,\hat{V})}{\partial x \partial v}\right] - \mathbb{E}\left[\frac{\partial^2 m_Y(X,V)}{\partial x \partial v}\right]\right\} = O_P(1)$. Similarly, we have Corollary A.1 which states that $\sqrt{n}\left\{\hat{\mathbb{E}}\left[\frac{\partial^2 \widehat{m}_D(X,\hat{V})}{\partial x \partial v}\right] - \mathbb{E}\left[\frac{\partial^2 m_D(X,V)}{\partial x \partial v}\right]\right\} = O_P(1)$. Finally, combining the results of these two results and rearranging (20) we obtain the results of Theorem 1.

**Proof of Proposition 1:** We use the directional derivative to decompose the influences of the generated regressor. Following the proof of Corollary 6 in Mammen et al. (2012a), let

$$\bar{\phi}(x,z) = (\bar{\phi}_1, \bar{\phi}_2) = \left(\frac{\partial^2 m_Y(x, \bar{\phi}_2(x,z))}{\partial x \partial v}, x_1 - \pi(z)\right),$$
$$\hat{\phi}(x,z) = (\hat{\phi}_1, \hat{\phi}_2) = \left(\frac{\partial^2 \widehat{m}_Y(x, \hat{\phi}_2(x,z))}{\partial x \partial v}, x_1 - \hat{\pi}(z)\right),$$

where $\bar{\phi}$ represents the true regression function and $\hat{\phi}$ is the estimator.

For any $\phi = (\phi_1, \phi_2)$, define

$$S_n(\phi) := \frac{1}{n}\sum_{i=1}^{n}\phi_1(X_i, \phi_2(X_i, Z_i)) - \mathbb{E}\left[\frac{\partial^2 m_Y(X,V)}{\partial x \partial v}\right].$$

Then the directional derivative can be written as

$$\begin{aligned}
\dot{S}_n(\bar{\phi})[\phi - \bar{\phi}] &= \lim_{s\to 0}\frac{1}{s}[S_n(\bar{\phi} + s(\phi - \bar{\phi})) - S_n(\bar{\phi})] \\
&= \lim_{s\to 0}\frac{1}{s}\frac{1}{n}\sum_{i=1}^{n}\{[\bar{\phi}_1 + s(\phi_1 - \bar{\phi}_1)](X_i, \bar{\phi}_2 + s(\phi_2 - \bar{\phi}_2)) - \bar{\phi}_1(X_i, \bar{\phi}_2 + s(\phi_2 - \bar{\phi}_2)) \\
&\quad + \bar{\phi}_1(X_i, \bar{\phi}_2 + s(\phi_2 - \bar{\phi}_2)) - \bar{\phi}_1(X_i, \bar{\phi}_2)\} \\
&= \frac{1}{n}\sum_{i=1}^{n}[\phi_1 - \bar{\phi}_1](X_i, \bar{\phi}_2) + \frac{1}{n}\sum_{i=1}^{n}\frac{\partial \bar{\phi}_1}{\partial v}(X_i, \bar{\phi}_2)\cdot[\phi_2 - \bar{\phi}_2](X_i, Z_i) \\
&=: T_{1,n}(\phi) + T_{2,n}(\phi).
\end{aligned} \tag{23}$$

For any $\phi = (\phi_1, \phi_2)$ with bounded second derivatives, we have that

$$\begin{aligned}
&\|S_n(\phi) - S_n(\bar{\phi}) - \dot{S}_n(\bar{\phi})[\phi - \bar{\phi}]\|_\infty \\
&\qquad = O(\|\phi_2 - \bar{\phi}_2\|_\infty^2) + O\left(\|\phi_2 - \bar{\phi}_2\|_\infty \left\|\frac{\partial \phi_1}{\partial v} - \frac{\bar{\phi}_1}{\partial v}\right\|_\infty\right) =: \text{s.o.}(\phi).
\end{aligned} \tag{24}$$

where s.o.($\phi$) defined above is later shown to be of small order. Replacing $\phi$ in (24) by the estimator $\hat{\phi}$, we have

$$
\begin{aligned}
S_n(\hat{\phi}) &= \hat{\mathbb{E}}\left[\frac{\partial^2 \widehat{m}_Y(X,\hat{V})}{\partial x \partial v}\right] - \mathbb{E}\left[\frac{\partial^2 m_Y(X,V)}{\partial x \partial v}\right] \qquad (25)\\
&= S_n(\bar{\phi}) + T_{1,n}(\hat{\phi}) + T_{2,n}(\hat{\phi}) + \text{s.o.}(\hat{\phi})\\
&= \frac{1}{n}\sum_{i=1}^{n} \bar{\phi}_1(X_i, \bar{\phi}_2) - \mathbb{E}\left[\frac{\partial^2 m_Y(X,V)}{\partial x \partial v}\right] + \frac{1}{n}\sum_{i=1}^{n}[\hat{\phi}_1 - \bar{\phi}_1](X_i, \bar{\phi}_2) + T_{2,n}(\hat{\phi}) + \text{s.o.}(\hat{\phi})\\
&= \frac{1}{n}\sum_{i=1}^{n} \hat{\phi}_1(X_i, \bar{\phi}_2) - \mathbb{E}\left[\frac{\partial^2 m_Y(X,V)}{\partial x \partial v}\right] + T_{2,n}(\hat{\phi}) + \text{s.o.}(\hat{\phi})\,, \qquad (26)
\end{aligned}
$$

where the first term is the averaged estimator for the cross-partial derivative of the conditional expectation with true regressor instead of generated regressor. The rest of proof consists of three parts. Firstly, we use Lemma A.1-A.3 to prove asymptotic normality for the averaged estimator of the cross-partial derivative of the conditional expectation with true regressor $V$, and we obtain that the estimator for the average second order cross-partial derivatives of the conditional expectation converges at parametric rate as a by-product. Secondly, we show that s.o.($\hat{\phi}$) = $o_P(n^{-1/2})$. Thirdly, we apply Lemma A.4 to show $\sqrt{n}T_{2,n}(\hat{\phi}) = O_p(1)$. Lastly, we combine the results to finish the proof.

Let's start from the first part. With true regressor $V$, we have

$$
\begin{aligned}
\sqrt{n}\left(\frac{1}{n}\sum_{i=1}^{n}\frac{\partial^2 \widehat{m}_Y(X_i,V_i)}{\partial x \partial v} - \mathbb{E}\left[\frac{\partial^2 m_Y(X,V)}{\partial x \partial v}\right]\right) &= G_n\left[\frac{\partial^2 \widehat{m}_Y(X_i,V_i)}{\partial x \partial v} - \frac{\partial^2 m_Y(X_i,V_i)}{\partial x \partial v}\right] \qquad (27)\\
&\quad + G_n\left[\frac{\partial^2 m_Y(X_i,V_i)}{\partial x \partial v}\right] \qquad (28)\\
&\quad + \sqrt{n}\mathbb{E}\left[\frac{\partial^2 \widehat{m}_Y(X_i,V_i)}{\partial x \partial v} - \frac{\partial^2 m_Y(X_i,V_i)}{\partial x \partial v}\right], \qquad (29)
\end{aligned}
$$

where $G_n$ is the empirical measure defined in (22). The first term (27) is $o_P(1)$ by the stochastic equicontinuity result in Lemma A.1. The second term (28) is $O_P(1)$ by the Donsker property of $C_M^\alpha(\mathcal{S}_{XV})$. The third term (29) contributes to the influence from estimating the cross partial derivatives of the conditional expectation which we shall see later.

**Lemma A.1 (Stochastic Equicontinuity):** $G_n\left[\frac{\partial^2 \widehat{m}_Y(X_i,V_i)}{\partial x \partial v} - \frac{\partial^2 m_Y(X_i,V_i)}{\partial x \partial v}\right] = o_P(1)$.

**Proof of Lemma A.1:** Define $Z_{ni}(\phi) = \frac{1}{\sqrt{n}}\phi(X_i,V_i)$ indexed by $\phi \in \mathcal{F} = C_M^\alpha(\mathcal{S}_{XV})$. Note that $\frac{\partial^2 \widehat{m}_Y(X,V)}{\partial x \partial v} \in C_M^\alpha(\mathcal{S}_{XV})$ with probability approaching to 1 by Masry (1996) and by the fact that we use local polynomial estimator with a twice continuously differentiable kernel that has a compact support. By Assumption 8 and Example 19.9 in van der Vaart (2000), we know that $C_M^\alpha(\mathcal{S}_{XV})$ is $P$-Donsker, and also totally bounded in $L_2(P)$ since for any $\varepsilon > 0$, the entropy $\log N_{[\,]}(\varepsilon, C_M^\alpha, L_2(P)) < \infty$. The bracketing CLT (Theorem 2.11.9 in van der Vaart and Wellner,

1996) implies that $\sum_{i=1}^n [Z_{ni}(\phi) - Z_{ni}(\tilde{\phi})]$ is asymptotically equicontinuous in $\phi$ with respect to the $L_2$ norm $\|\phi - \tilde{\phi}\|_2$, which further implies (27) is of small order. To apply this theorem we need to verify the assumptions of this theorem.

(i) We know that $|\phi(x,v)|$ is uniformly bounded as $\phi \in C_M^\alpha(\mathcal{S}_{XV})$. Therefore, for any $\eta > 0$, $1_{\|Z_{ni}\|_{\mathcal{F}} > \eta} = 0$ when $n$ is large enough. Obviously, we have $\sum_{i=1}^n \mathbb{E}^* \|Z_{ni}\|_{\mathcal{F}} 1_{\|Z_{ni}\|_{\mathcal{F}} > \eta} \to 0$ for every $\eta > 0$.

(ii) We have $\sup_{\|\phi - \phi'\|_2 < \delta_n} \sum_{i=1}^n \mathbb{E}[Z_{ni}(\phi) - Z_{ni}(\phi')]^2 = \sup_{\|\phi - \phi'\|_2 < \delta_n} \mathbb{E}[\phi(X,V) - \phi'(X,V)]^2 \to 0$ for any $\delta_n \downarrow 0$.

(iii) $\int_0^{\delta_n} \sqrt{\log N_{[]}(\varepsilon, \mathcal{F}, L_2(P))}\, d\varepsilon \to 0$ holds for every $\delta_n \to 0$ also by Assumption 8 and Example 19.9 in van der Vaart (2000) (or Corollary 2.7.2 in van der Vaart and Wellner, 1996). ■

For any integer vector $j = (j_1, \ldots, j_{d+1})$ and random vector $w = (w_1, \cdots, w_{d+1})$, we have already defined that $|j| = j_1 + \cdots + j_{d+1}$ and $w^j = (w_1^{j_1}, \cdots, w_{d+1}^{j_{d+1}})$. Let $j! = j_1! \times \cdots \times j_{d+1}!$.

For ease of notation, let $W = (X, V)$. For the rest of the proof of Proposition 1 we will use $m(w)$ to denote $m_Y(w) = m_Y(x,v)$. Also to eliminate one subscript we will use $f(w)$ to denote the joint density of $(X,V)$, $f_{XV}(x,v)$. As in Masry (1996a,b), $k!\hat{b}_k(w)$ estimates the $|k|$th order partial derivative $D^k m(w)$ (or $\frac{\partial^{|k|} m(w)}{\partial w_1^{k_1} \cdots \partial w_2^{k_2}}$), where $\hat{b}_k(w)$ minimizes the weighted least squares

$$\sum_{i=1}^n \left[ Y_i - \sum_{0 \le |k| \le p} b_k(w)(W_i - w)^k \right]^2 K\left(\frac{W_i - w}{h}\right). \tag{30}$$

Minimizing (30), the F.O.C can be formulated as

$$t_{n,j} = \sum_{0 \le |k| \le p} h^{|k|} \hat{b}_k(w) s_{n,j+k}(w), \quad 0 \le |j| \le p\,, \tag{31}$$

where

$$t_{n,j} = \frac{1}{n} \sum_{i=1}^n Y_i \left(\frac{W_i - w}{h}\right)^j K_h(W_i - w)\,,$$
$$s_{n,j} = \frac{1}{n} \sum_{i=1}^n \left(\frac{W_i - w}{h}\right)^j K_h(W_i - w)\,.$$

We write (31) in a matrix form by using a lexicographical order. Let $N_i = \binom{i+1}{1} = i + 1$ be the number of distinct pairs with $|j| = i$, which is the number of distinct derivatives with order $i$. Order these $N_i$ pairs as a sequence in the lexicographical order with the highest priority to the first position (highest order derivative w.r.t. $x$), so the sequence starts from $(i, \ldots, 0, 0)$ and ends at $(0, 0, \ldots, i)$. Let $g^{-1}$ denote this one-to-one map. Define

$$\tau_n = \left[\tau_{n,0}^\intercal, \tau_{n,1}^\intercal, \ldots, \tau_{n,p}^\intercal\right]^\intercal$$

where $\tau_{n,|j|}$ is a $N_{|j|} \times 1$ vector with $(\tau_{n,|j|})_k = t_{n,g|j|}(k)$. Note that $\tau_n$ is of dimension $\mathbf{N} \times 1$ with

$$\mathbf{N} = \sum_{i=0}^{p} N_i \,. \tag{32}$$

Similarly we arrange $h^{|k|}\hat{b}_k$ in the same order to get

$$\hat{\beta}_n = \begin{bmatrix} \hat{\beta}_{n,0}^{\intercal} & \hat{\beta}_{n,1}^{\intercal} & \cdots & \hat{\beta}_{n,p}^{\intercal} \end{bmatrix}^{\intercal} .$$

We also arrange the possible values of $s_{n,j+k}$ by a matrix $S_{n,|j|,|k|}$ in a lexicographical order with the $(l,m)$ element $[S_{n,|j|,|k|}]_{l,m} = s_{n,g_{|j|}(l)+g_{|k|}(m)}$. Define the $\mathbf{N} \times \mathbf{N}$ matrix $S_n$

$$S_n = \begin{bmatrix} S_{n,0,0} & S_{n,0,1} & \cdots & S_{n,0,p} \\ S_{n,1,0} & S_{n,1,1} & \cdots & S_{n,1,p} \\ \vdots & \vdots & \ddots & \vdots \\ S_{n,p,0} & S_{n,0,1} & \cdots & S_{n,p,p} \end{bmatrix} .$$

Before we proceed, we need to introduce some notations following Masry (1996a,b) and Li et al. (2003). For each $j$ with $0 \le |j| \le p$, define

$$\mu_j = \int \upsilon_1^{j_1} \upsilon_2^{j_2} K(\upsilon)\, d\upsilon \,, \quad \nu_{1,j} = \int (\upsilon_1^{j_1+1}, \upsilon_2^{j_2}) K(\upsilon)\, d\upsilon, \quad \nu_{2,j} = \int \upsilon_1^{j_1} \upsilon_2^{j_2+1} K(\upsilon)\, d\upsilon,$$
$$\kappa_{r,j} = \int \upsilon_1^{r_1+j_1} \upsilon_2^{r_2+j_2} K(\upsilon)\, d\upsilon, \quad |r| = 2 \,,$$

where $\upsilon = (\upsilon_1, \upsilon_2)$. Then define $\mathbf{N} \times \mathbf{N}$ matrices $M$, $U_s$ $(s = 1, 2)$, $H_r$ $(|r| = 2)$, $U(w)$ and $H(w)$ by

$$M = \begin{bmatrix} M_{0,0} & M_{0,1} & \cdots & M_{0,p} \\ M_{1,0} & M_{1,1} & \cdots & M_{1,p} \\ \vdots & \vdots & \ddots & \vdots \\ M_{p,0} & M_{p,1} & \cdots & M_{p,p} \end{bmatrix}, U_s = \begin{bmatrix} U_{s,0,0} & U_{s,0,1} & \cdots & U_{s,0,p} \\ U_{s,1,0} & U_{s,1,1} & \cdots & U_{s,1,p} \\ \vdots & \vdots & \ddots & \vdots \\ U_{s,p,0} & U_{s,p,1} & \cdots & U_{s,p,p} \end{bmatrix}, H_r = \begin{bmatrix} H_{r,0,0} & H_{r,0,1} & \cdots & H_{r,0,p} \\ H_{r,1,0} & H_{r,1,1} & \cdots & H_{r,1,p} \\ \vdots & \vdots & \ddots & \vdots \\ H_{r,p,0} & H_{r,p,1} & \cdots & H_{r,p,p} \end{bmatrix} \tag{33}$$

and

$$U(w) = \frac{\partial f(w)}{\partial w_1} U_1 + \frac{\partial f(w)}{\partial w_2} U_2 \,, \quad H(w) = \frac{\partial^2 f(w)}{\partial w_1^2} H_{(2,0)} + \frac{\partial^2 f(w)}{\partial w_1 \partial w_2} H_{(1,1)} + \frac{\partial^2 f(w)}{\partial w_2^2} H_{(0,2)}$$

where $M_{i,j}$, $U_{s,i,j}$ and $H_{r,i,j}$ are $N_i \times N_j$ dimensional matrices whose $(l,m)$ element are $\mu_{g_i(l)+g_j(m)}$, $\nu_{s,g_i(l)+g_j(m)}$ and $\kappa_{r,g_i(l)+g_j(m)}$ respectively. Define the centered $t_{n,j}(w)$ as

$$\begin{aligned} t_{n,j}^*(w) &= \frac{1}{n} \sum_{i=1}^{n} [Y_i - m(W_i)] \left( \frac{W_i - w}{h} \right)^j K_h(W_i - w) \\ &= \frac{1}{n} \sum_{i=1}^{n} \xi_i \left( \frac{W_i - w}{h} \right)^j K_h(W_i - w) \,. \end{aligned}$$

Then the $\mathbf{N} \times 1$ matrix $\tau_n^*(w)$ are defined the same way as $\tau_n(w)$ but with $t_{n,j}(w)$ replaced with $t_{n,j}^*(w)$.

Use the same lexicographical order as before, we define a column vector $m^{(p+1)}(w)$ as the $N_{p+1}$ elements of derivatives $1/j!(D^j m)(w)$ with $|j| = p+1$. Define the $\mathbf{N} \times N_{p+1}$ matrices $B_n(w)$ and $B$ by

$$B_n(w) = \begin{bmatrix} S_{n,0,p+1} \\ S_{n,1,p+1} \\ \vdots \\ S_{n,p,p+1} \end{bmatrix} \quad \text{and} \quad B = \begin{bmatrix} M_{0,p+1} \\ M_{1,p+1} \\ \vdots \\ M_{p,p+1} \end{bmatrix}. \tag{34}$$

Similarly as (A.9) in Li et al. (2003), applying Theorem 4 in Masry (1996b), we have (29) becomes

$$\begin{aligned} & e_t^\intercal \frac{1}{h^2} \mathbb{E}[\hat{\beta}_n(W) - \beta(W)] \\ &= e_t^\intercal \frac{1}{h^2} \left[ \int S_n^{-1}(w) \tau_n^*(w)\, dF(w) + h^{p+1} \int S_n^{-1}(w) B_n(w) m^{(p+1)}(w)\, dF(w) \right] + O_P(\sqrt{\frac{\log n}{nh^2}} h^{p-1} + h^p), \end{aligned} \tag{35}$$

where $t = 5$. For any $k^{th}$ order derivative of the expectation, we shall have $t \in \{1 + N_1 + \cdots + N_{k-1} + 1, \ldots, 1 + N_1 + \cdots + N_k\}$. By Assumption 7(ii), $O_P(\sqrt{\frac{\log n}{nh^2}} h^{p-1} + h^p) = o_P(n^{-1/2})$.

Masry (1996a,b) has shown that $\sup_{x \in \mathcal{D}} |[S_n(w)]^{-1} - [f(w)M]^{-1}| = o(1)$, and Li et al. (2003) have demonstrated that $\sup_{w \in \mathcal{D}} |[S_n(w)]^{-1} - \{[f(w)M]^{-1} - hG(w)\}| = o(h)$ *a.s.*, where $G(w) = M^{-1}U(w)M^{-1}/f^2(w)$, and $\mathcal{D} \subseteq \mathcal{S}_{X,V}$. We prove that higher order result also holds.

**Lemma A.2:** Under the assumptions of Proposition 1, we have

$$\sup_{w \in \mathcal{D}} |S_n(w) - f(w)M - hU(w) - h^2 H(w)| = O\left( h^3 + \left( \frac{\log n}{nh^2} \right)^{1/2} \right) = o(h^2).$$

Also, the inverse holds that

$$\sup_{w \in \mathcal{D}} |S_n^{-1}(w) - \{[f(w)M]^{-1} - hG(w) + h^2 Q(w)\}| = o(h^2),$$

where $Q(w) = [f(w)M]^{-1}U(w)[f(w)M]^{-1}U(w)[f(w)M]^{-1} - [f(w)M]^{-1}H(w)[f(w)M]^{-1}$.

**Proof of Lemma A.2:** It suffices to show that for each $j$ with $0 \leq |j| \leq \mathbf{N}$, $\sup_{w \in \mathcal{D}} |s_{n,j}(w) - f(w)\mu_j - h\sum_{s=1}^2 f_s^{(1)}(w)\nu_{s,j} - h^2 \sum_{|r|=2} f_r^{(2)}(w)\kappa_{r,j}| = O(h^3 + [\log n/(nh^2)]^{1/2})$. Note that $s_{n,j}(w) = n^{-1}\sum_{i=1}^n ((W_i - w)/h)^j K_h(W_i - w)$, we have

$$\begin{aligned} \mathbb{E}[s_{n,j}(w)] &= \int (w_i - w)^j / h^j K_h(w_i - w) f(w_i)\, dw_i \\ &= \int \upsilon^j K(\upsilon) f(w + h\upsilon)\, d\upsilon + f(w) \int \upsilon^j K(\upsilon)\, d\upsilon + h \sum_{s=1}^2 f_s^{(1)}(w) \int \upsilon_s \upsilon^j K(\upsilon)\, d\upsilon \\ &\quad + h^2 \sum_{|r|=2} f_r^{(2)}(w) \int \upsilon_r \upsilon^j K(\upsilon)\, d\upsilon + O(h^3), \end{aligned}$$

uniformly in $w \in \mathcal{D}$. The first result follows from this together with Theorem 2 in Masry (1996b). Thus, we have

$$\begin{aligned} S_n(w) &= f(w)M + hU(w) + h^2H(w) + O\left(h^3 + \left(\frac{\log n}{nh^2}\right)^{1/2}\right) \\ &= f(w)M\{I_n + h[f(w)M]^{-1}U(w) + h^2[f(w)M]^{-1}H(w) + o(h^2)\} \quad a.s.\,, \end{aligned}$$

uniformly in $\mathcal{D}$. The second equality is by Assumption 7(ii). It is easy to see that

$$\begin{aligned} &\{I_n + h[f(w)M]^{-1}U(w) + h^2[f(w)M]^{-1}H(w) + o(h^2)\}^{-1} \\ &= I_n - h[f(w)M]^{-1}U(w) + h^2\{[f(w)M]^{-1}U(w)[f(w)M]^{-1}U(w) - [f(w)M]^{-1}H(w)\} + o(h^2)\,, \end{aligned}$$

uniformly in $\mathcal{D}$ by Taylor expansion, as the eigenvalues of the matrix

$$h[f(w)M]^{-1}U(w) + h^2[f(w)M]^{-1}H(w) + o(h^2)$$

are of small order. Then we have

$$S_n^{-1}(w) = [f(w)M]^{-1} - hG(w) + h^2Q(w) + o(h^2)\,.$$

where

$$\begin{aligned} G(w) &= [f(w)M]^{-1}U(w)[f(w)M]^{-1} \\ Q(w) &= [f(w)M]^{-1}U(w)[f(w)M]^{-1}U(w)[f(w)M]^{-1} - [f(w)M]^{-1}H(w)[f(w)M]^{-1}\,. \blacksquare \end{aligned}$$

Applying Lemma A.2 on (35), we shall show that the leading term is indeed determined by higher order terms, as those from the first and second terms of Taylor expansion are degenerate. We have

$$\begin{aligned} e_t^\intercal \frac{1}{h^2}\mathbb{E}[\hat{\beta}_n(W) - \beta(W)] &= e_t^\intercal \frac{1}{h^2}\bigg\{\int [f(w)M]^{-1}\tau_n^*(w)\,dF(w) \\ &\quad - \int hG(w)\tau_n^*(w)\,dF(w) + \int h^2Q(w)\tau_n^*(w)\,dF(w) \\ &\quad + h^{p+1}\int S_n^{-1}(w)B_n(w)m^{(p+1)}(w)dF(w)\bigg\} + o_P(n^{-1/2}) \\ &= A_1 - A_2 + A_3 + A_4 + o_P(n^{-1/2})\,. \end{aligned}$$

We shall show that $A_2$ and $A_3$ contribute to the variance, $A_4$ is the bias term and the rest terms are negligible. Similar to the proof of Lemma A.3 in Li et al. (2003)

$$\begin{aligned} A_1 &= \int f^{-1}(w)\frac{1}{nh^2}\sum_{i=1}^n \sum_{0\le|j|\le p} [M^{-1}]_{5,j}\xi_i\left(\frac{W_i - w}{h}\right)^j K_h(W_i - w)\,dF(w) \\ &= \frac{1}{nh^2}\sum_{i=1}^n \xi_i \sum_{0\le|j|\le p}[M^{-1}]_{t,j}\int u^j K(u)\,du \\ &= \frac{1}{nh^2}\sum_{i=1}^n \xi_i \sum_{0\le|j|\le p}[M^{-1}]_{t,j}[M]_{j,1} = \frac{1}{nh^2}\sum_{i=1}^n \xi_i[M^{-1}M]_{5,1} = 0\,. \end{aligned}$$

We shall show $A_2$ is $O_P(n^{-1/2})$,

$$
\begin{aligned}
A_2 &= \int \frac{1}{nh}\sum_{i=1}^{n}\sum_{0\le|j|\le p}[G(w)]_{5,j}\xi_i\left(\frac{W_i-w}{h}\right)^j K_h(W_i-w)\,dF(w)\\
&= \frac{1}{nh}\sum_{i=1}^{n}\xi_i\int\sum_{0\le|j|\le p}[G(w)]_{5,j}\left(\frac{W_i-w}{h}\right)^j K_h(W_i-w)\,dF(w)\\
&= \frac{1}{nh}\sum_{i=1}^{n}\xi_i\sum_{0\le|j|\le p}\int f^{-1}(W_i-uh)[M^{-1}U(W_i-uh)M^{-1}]_{5,j}u^jK(u)\,du\\
&= \frac{1}{nh}\sum_{i=1}^{n}\xi_i f^{-1}(W_i)\sum_l[M^{-1}U(W_i)]_{5,l}\sum_{0\le|j|\le p}[M^{-1}]_{l,j}\int u^jK(u)\,du\\
&\quad -\frac{1}{n}\sum_{i=1}^{n}\xi_i\int\sum_{s'=1}^{2}(f^{-1})^{(1)}_{s'}(W_i)\sum_{0\le|j|\le p}[M^{-1}U(W_i)M^{-1}]_{5,j}u_{s'}u^jK(u)\,du\\
&\quad -\frac{1}{n}\sum_{i=1}^{n}\xi_i\sum_{0\le|j|\le p}\int f^{-1}(W_i)\sum_{s'=1}^{2}[M^{-1}(U(W_i))^{(1)}_{s'}M^{-1}]_{5,j}u_{s'}u^jK(u)\,du+o_P(n^{-1/2}),
\end{aligned}
$$

where we have used Taylor expansion for $f^{-1}(W_i-uh)$ and $U(W_i-uh)$. By the definition of $M$ and $U_s(W_i)$, we have

$$
\begin{aligned}
A_2 &= \frac{1}{nh}\sum_{i=1}^{n}\xi_i f^{-1}(W_i)\sum_k[M^{-1}]_{5,k}\sum_{s=1}^{2}f^{(1)}_s(W_i)\sum_l[U_s]_{k,l}I_{l,1}\\
&\quad -\frac{1}{n}\sum_{i=1}^{n}\xi_i\int\sum_{s'=1}^{2}\sum_{s=1}^{2}(f^{-1})^{(1)}_{s'}(W_i)f^{(1)}_s(W_i)\sum_{0\le|j|\le p}[M^{-1}U_sM^{-1}]_{5,j}u_{s'}u^jK(u)\,du\\
&\quad -\frac{1}{n}\sum_{i=1}^{n}\xi_i\sum_{0\le|j|\le p}\int f^{-1}(W_i)\sum_{s=1}^{2}\sum_{s'=1}^{2}f^{(2)}_{ss'}(W_i)[M^{-1}U_sM^{-1}]_{5,j}u_{s'}u^jK(u)\,du+o_P(n^{-1/2})\\
&= \frac{1}{nh}\sum_{i=1}^{n}\xi_i f^{-1}(W_i)\sum_{s=1}^{2}f^{(1)}_s(W_i)\sum_k[M^{-1}]_{5,k}[U_s]_{k,1}\\
&\quad -\frac{1}{n}\sum_{i=1}^{n}\xi_i\sum_{s=1}^{2}\sum_{s'=1}^{2}[(f^{-1})^{(1)}_{s'}(W_i)f^{(1)}_s(W_i)+f^{-1}(W_i)f^{(2)}_{ss'}(W_i)]\sum_{0\le|j|\le p}[M^{-1}U_sM^{-1}]_{5,j}[U_{s'}]_{j,1}\\
&\quad +o_P(n^{-1/2})\\
&= A_{21}-A_{22}+o_P(n^{-1/2}).
\end{aligned}
$$

The first term $A_{21}$ is degenerate, as for any $k$, we can show $[M^{-1}]_{5,k}[U_s]_{k,1}=0$. It is sufficient to show $[M^{-1}]_{5,k}=0$, whenever $[U_s]_{k,1}\neq 0$. We have that for $s=1,2$,

$$
\begin{aligned}
U_s[\cdot,1] &= [0\ \square_{1\times N_1}\ 0_{1\times N_2}\ \square_{1\times N_3}\ 0_{1\times N_4}\ \dots\ 0_{1\times N_p}]^\intercal, \quad \text{if p is even},\\
U_s[\cdot,1] &= [0\ \square_{1\times N_1}\ 0_{1\times N_2}\ \square_{1\times N_3}\ 0_{1\times N_4}\ \dots\ \square_{1\times N_p}]^\intercal, \quad \text{if p is odd},
\end{aligned}
$$

where $\square_{1\times l}$ represents a $1\times l$ nonzero row vector. Note that $N_0 = 1$. As we use product kernel, it is easy to see that when $p$ is odd, $M$ is a block matrix with zero blocks and nonzero blocks appear alternatively, i.e.

$$M = \begin{bmatrix} \square_{N_0\times N_0} & 0_{N_0\times N_1} & \square_{N_0\times N_2} & \dots & 0_{N_0\times N_p} \\ 0_{N_1\times N_0} & \square_{N_1\times N_1} & 0_{N_1\times N_2} & \dots & \square_{N_1\times N_p} \\ \vdots & \vdots & \vdots & \dots & \vdots \\ 0_{N_p\times N_0} & \square_{N_p\times N_1} & 0_{N_p\times N_2} & \dots & 0_{N_p\times N_p} \end{bmatrix}.$$

When $p$ is even, it is clear to see $M$ has the same pattern. We use the result of the lemma below.

**Lemma A.3:** $M^{-1}$ has the same zero blocks as $M$.

**Proof of Lemma A.3:** We only prove the result when $p$ is odd. When $p$ is even, the derivation is similar. Let $p = 2q-1$, $N_e = N_0 + N_2 + \cdots + N_{p-1}$ and $N_e = N_1 + N_3 + \cdots + N_p$. We first apply the row switching block elementary matrix $T_{i,j}$ on $M$, then we have

$$T_{N_1,N_{p-1}}T_{N_3,N_{p-3}}\cdots T_{N_{q-1},N_q}MT_{N_{q-1},N_q}\cdots T_{N_3,N_{p-3}}T_{N_1,N_{p-1}} = \begin{bmatrix} A_{N_e\times N_e} & 0_{N_e\times N_o} \\ 0_{N_o\times N_e} & B_{N_o\times N_o} \end{bmatrix},$$

where $A$ and $B$ are invertible matrices. Then, we have

$$M^{-1} = (T_{N_1,N_{p-1}}T_{N_3,N_{p-3}}\cdots T_{N_{q-1},N_q})^{-1}\begin{bmatrix} A^{-1}_{N_e\times N_e} & 0_{N_e\times N_o} \\ 0_{N_o\times N'_e} & B^{-1}_{N_o\times N_o} \end{bmatrix}(T_{N_{q-1},N_q}\cdots T_{N_3,N_{p-3}}T_{N_1,N_{p-1}})^{-1}. \blacksquare$$

For the $t$th row of $M$, it belongs to the $N_0+N_1+1$ to $N_0+N_1+N_2$ block, so we have

$$[M]_{t,\cdot} = \begin{bmatrix} \square_{1\times N_0} & 0_{1\times N_1} & \square_{1\times N_2}\dots \end{bmatrix}.$$

Thus, it is obvious that $A_{21}$ is degenerate. For $A_{22}$, we have

$$\begin{aligned} A_{22} &= \frac{1}{n}\sum_{i=1}^{n}\xi_i\sum_{s=1}^{2}\sum_{s'=1}^{2}[(f^{-1})^{(1)}_{s'}(W_i)f^{(1)}_s(W_i) + f^{-1}(W_i)f^{(2)}_{ss'}(W_i)][M^{-1}U_sM^{-1}U_{s'}]_{5,1} \\ &= \frac{1}{n}\sum_{i=1}^{n}\xi_i[R_1(W_i)]_{5,1}, \end{aligned}$$

where $R_1(w) = \sum_{s=1}^{2}\sum_{s'=1}^{2}[(f^{-1})^{(1)}_{s'}(w)f^{(1)}_s(w) + f^{-1}(w)f^{(2)}_{ss'}(w)]M^{-1}U_sM^{-1}U_{s'}$.

Finally, we have

$$\begin{aligned} A_3 &= \frac{1}{n}\sum_{i=1}^{n}\xi_i\sum_{0\le|j|\le p}\int[Q(w)]_{5,j}\left(\frac{W_i-w}{h}\right)^j K_h(W_i-w)\,dF(w) \\ &= \frac{1}{n}\sum_{i=1}^{n}\xi_i\sum_{0\le|j|\le p}\int[Q(W_i-uh)]_{5,j}f(W_i-uh)u^jK(u)\,du \\ &= \frac{1}{n}\sum_{i=1}^{n}\xi_i\sum_{0\le|j|\le p}[Q(W_i)]_{5,j}\mu_jf(W_i) + o_P(n^{-1/2}) = \frac{1}{n}\sum_{i=1}^{n}\xi_i[R_2(W_i)]_{5,1} + o_P(n^{-1/2}), \end{aligned}$$

where $R_2(w) = Q(w)f(w)M$. Combining $A_{22}$ and $A_3$, let

$$R(w) = R_1(w)+R_2(w) = \sum_{s=1}^{2}\sum_{s'=1}^{2}[(f^{-1})^{(1)}_{s'}(w)f^{(1)}_{s}(w)+f^{-1}(w)f^{(2)}_{ss'}(w)]M^{-1}U_sM^{-1}U_{s'}+Q(w)f(w)M\,. \tag{36}$$

Finally, applying the uniform results in Masry (1996b), we have

$$\begin{aligned} A_4 &= e_t^{\intercal}h^{p-1}\int S_n^{-1}(w)B_n(w)m^{(p+1)}(w)\,dF(w) \\ &= e_t^{\intercal}h^{p-1}M^{-1}B\mathbb{E}[m^{(p+1)}(w)] + O_P(h^p)\,. \end{aligned}$$

Altogether we have

$$\sqrt{n}\mathbb{E}\left[\frac{\partial^2\widehat{m}(W)}{\partial w_1\partial w_2} - \frac{\partial^2 m(W)}{\partial w_1\partial w_2} - e_5^{\intercal}h^{p-1}M^{-1}B\mathbb{E}[m^{(p+1)}(w)]\right] = \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\xi_i[R(W_i)]_{5,1}\,,$$

then we can apply CLT directly for the righthand side.

Now, let's check for the order of s.o.$(\hat{\phi})$. From Masry (1996a,b), we have

$$\|\hat{\pi}-\pi\|_\infty = O_P\left([\log n/(ng)]^{1/2}+g^2\right)\,.$$

Therefore, $O(\|\hat{\phi}_2-\bar{\phi}_2\|_\infty^2) = o_P(n^{-1/2})$ by Assumption 7(i). By a similar argument as in Masry (1996a,b) or Mammen et al. (2012a), we can show that we have $\|\hat{\phi}_2-\bar{\phi}_2\|_\infty\|\hat{\phi}_1^{(v)}-\bar{\phi}_1^{(v)}\|_\infty = O_P(\log n/(n^2gh^8)^{1/2}) = o_P(n^{-1/2})$ by Assumption 7(iii). Therefore, s.o.$(\hat{\phi}) = o_P(n^{-1/2})$.

To show $T_{2,n}(\hat{\phi}) = O_P(n^{-1/2})$, recall that

$$T_{2,n}(\hat{\phi}) = \frac{1}{n}\sum_{i=1}^{n}\frac{\partial^3 m_Y(X_i,V_i)}{\partial x\partial v^2}(\hat{V}_i - V_i)\,.$$

By Lemma A.4 below, we can show that

$$T_{2,n}(\hat{f}) = \mathbb{E}\left[(\hat{V}-V)\frac{\partial^3 m_Y(X,V)}{\partial x\partial v^2}\right] + o_P(n^{-1/2})\,.$$

**Lemma A.4:** Under the assumptions of Proposition 1, we have

$$\sup_{V_1,V_2\in\bar{\mathcal{M}}_n}\left|\frac{1}{n}\sum_{i=1}^{n}[V_1(X_i,Z_i)-V_2(X_i,Z_i)]\frac{\partial^3 m_Y(X_i,V_i)}{\partial x\partial v^2} - \mathbb{E}\left[(V_1-V_2)\frac{\partial^3 m_Y(X,V)}{\partial x\partial v^2}\right]\right| = o_P(n^{-1/2})\,,$$

**Proof of Lemma A.4:** Define $\Delta_i(V_1,V_2) = [V_1(X_i,Z_i)-V_2(X_i,Z_i)]\frac{\partial^3\mathbb{E}(Y_i|X_i,V_i)}{\partial x\partial v^2} - \mathbb{E}[(V_1-V_2)\frac{\partial^3\mathbb{E}(Y|X,V)}{\partial x\partial v^2}]$. Then the proof follows closely of Lemma A.5 in Lee (2013), which modifies that of Lemma 1 in Mammen et al. (2012a). We need to verify that Assumption 2 (Accuracy) and Assumption 3 (Complexity) in Mammen et al. (2012a) hold, which are

- (Accuracy) $\sup_z|\hat{\pi}(z)-\pi(z)| = o_P(n^{-\iota})$ for some $\iota$ such that $n^{-\iota} = o(h)$.

- (Complexity) There exists sequences of sets $\mathcal{M}_n$, such that

  1. $\hat{\pi}(\cdot) \in \mathcal{M}_n$.
  2. For a constant $C_M > 0$ and a function $\pi_n$ with $\|\pi_n - \pi\|_\infty = o(n^{-\iota})$, the set $\bar{\mathcal{M}}_n = \mathcal{M}_n \cap \{\pi^\star : \|\pi^\star - \pi_n\|_\infty \leq n^{-\iota}\}$ can be covered by at most $C_M \exp(c_\lambda^{-\psi} n^\zeta)$ balls with $\|\cdot\|_\infty$-radius $c_\lambda$ for all $c_\lambda \leq n^{-\iota}$, where $0 < \psi \leq 2$, $\zeta \in \mathcal{R}$.

and also $\iota - \frac{1}{2}(\iota\psi + \zeta) > 0$, which indicates $\kappa_1 > 1/2$ in Lemma A.3 in Lee (2013). As we use local linear regression for $\hat{\pi}(z)$, we know $n^{-\iota} = o(h)$ by Assumption 7 and we can let $\mathcal{M}_n$ be the set of functions defined on the compact support of $Z$ with up to $\alpha$ order partial derivatives and uniformly bounded by some multiple of $n^{\zeta^\star}$ with $\zeta^\star \geq 0$. Then, by Corollary 2.7.2 in van der Vaart and Wellner (1996), we have $\psi = \frac{1}{\alpha}$ and $\zeta = \zeta^\star \psi$. By choosing $\zeta^\star$ sufficiently small and $1 < 2\alpha$, the result holds. ■

We know that $\hat{V} - V = -[\hat{\pi}(Z) - \pi(Z)]$, then

$$\sqrt{n} T_{2,n}(\hat{f}) = -\frac{1}{\sqrt{n}} \sum_{i=1}^n V_i \mathbb{E}_{X|Z_i} \left\{ \frac{\partial^3 m_Y(X, X - \pi(Z_i))}{\partial x \partial v^2} \right\} + o_P(1) \,.$$

Altogether, we have by (26),

$$\begin{aligned} \sqrt{n}\{S_n(\hat{f}) - 2e_t' h^{p-1} M^{-1} B \mathbb{E}[m_Y^{(p+1)}(w)]\} &= \frac{2}{\sqrt{n}} \sum_{i=1}^n \xi_i [R(W_i)]_{5,1} \\ &\quad - \frac{1}{\sqrt{n}} \sum_{i=1}^n V_i \mathbb{E}_{X|Z_i} \left\{ \frac{\partial^3 m_Y(X, X - \pi(Z_i))}{\partial x \partial v^2} \right\} \\ &\quad + \frac{1}{\sqrt{n}} \sum_{i=1}^n \left\{ \frac{\partial^2 m_Y(X_i, V_i)}{\partial x \partial v} - \mathbb{E}\left[ \frac{\partial^2 m_Y(X, V)}{\partial x \partial v} \right] \right\} + o_P(1) \,. \end{aligned}$$

Thus, the desired result follows by the CLT. ■

Let $m_D(w) = \mathbb{E}(D|W = w)$, the following can be proved in the same way as Proposition 1.

**Corollary A.1:** Under assumptions of Theorem 1, we have

$$\begin{aligned} \sqrt{n} \left\{ \hat{\mathbb{E}} \left[ \frac{\partial^2 \widehat{m}_D(X, \hat{V})}{\partial x \partial v} \right] - \mathbb{E}\left[ \frac{\partial^2 m_D(X, V)}{\partial x \partial v} \right] - 2e_5' h^{p-1} M^{-1} B \mathbb{E}[m_D^{p+1}(X, V)] \right\} &= \frac{2}{\sqrt{n}} \sum_{i=1}^n u_i [R(X_i, V_i)]_{5,1} \\ - \frac{1}{\sqrt{n}} \sum_{i=1}^n V_i \mathbb{E}_{X|Z_i} \left\{ \frac{\partial^3 m_D(X, X - \pi(Z_i))}{\partial x \partial v^2} \right\} + \frac{1}{\sqrt{n}} \sum_{i=1}^n \left\{ \frac{\partial^2 m_D(X_i, V_i)}{\partial x \partial v} - \mathbb{E}\left[ \frac{\partial^2 m_D(X, V)}{\partial x \partial v} \right] \right\} + o_P(1) \,. \end{aligned}$$

**Proof of Theorem 1:** Note that

$$\begin{aligned} \sqrt{n}(\hat{\gamma} - \gamma) = \frac{1}{\mathbb{E}\left[ \frac{\partial^2 m_D(X,V)}{\partial x \partial v} \right] \frac{1}{n} \sum_{i=1}^n \frac{\partial^2 \hat{m}_D(X_i, V_i)}{\partial x \partial v}} \\ \times \left\{ \mathbb{E}\left[ \frac{\partial^2 m_D(X, V)}{\partial x \partial v} \right] \frac{1}{\sqrt{n}} \sum_{i=1}^n \left( \frac{\partial^2 \hat{m}_Y(X_i, V_i)}{\partial x \partial v} - \mathbb{E}\left[ \frac{\partial^2 m_Y(X, V)}{\partial x \partial v} \right] \right) \right. \\ \left. - \mathbb{E}\left[ \frac{\partial^2 m_Y(X, V)}{\partial x \partial v} \right] \frac{1}{\sqrt{n}} \sum_{i=1}^n \left( \frac{\partial^2 \hat{m}_D(X_i, V_i)}{\partial x \partial v} - \mathbb{E}\left[ \frac{\partial^2 m_D(X, V)}{\partial x \partial v} \right] \right) \right\}. \end{aligned}$$

Also note that $\frac{\partial^2 m_Y(X,V)}{\partial x \partial v} = \frac{\partial^2 m_D(X,V)}{\partial x \partial v}$, $\mathbb{E}\left[\frac{\partial^2 m_Y(X,V)}{\partial x \partial v}\right] = \mathbb{E}\left[\frac{\partial^2 m_D(X,V)}{\partial x \partial v}\right]$, and $\frac{\partial^3 m_Y(X,V)}{\partial x \partial v^2} = \frac{\partial^3 m_D(X,V)}{\partial x \partial v^2}$. So it is easy to see that

$$\begin{aligned}
&\mathbb{E}\left[\frac{\partial^2 m_D(X,V)}{\partial x \partial v}\right] \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \left( \frac{\partial^2 \hat{m}_Y(X_i,V_i)}{\partial x \partial v} - \mathbb{E}\left[\frac{\partial^2 m_Y(X,V)}{\partial x \partial v}\right] \right) \\
-&\mathbb{E}\left[\frac{\partial^2 m_Y(X,V)}{\partial x \partial v}\right] \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \left( \frac{\partial^2 \hat{m}_D(X_i,V_i)}{\partial x \partial v} - \mathbb{E}\left[\frac{\partial^2 m_D(X,V)}{\partial x \partial v}\right] \right) = 0, \\
-&\mathbb{E}\left[\frac{\partial^2 m_D(X,V)}{\partial x \partial v}\right] \frac{1}{\sqrt{n}} \sum_{i=1}^{n} V_i \mathbb{E}_{X|Z_i}\left\{ \frac{\partial^3 m_D(X, X-\pi(Z_i))}{\partial x \partial v^2} \right\} \\
+&\mathbb{E}\left[\frac{\partial^2 m_Y(X,V)}{\partial x \partial v}\right] \frac{1}{\sqrt{n}} \sum_{i=1}^{n} V_i \mathbb{E}_{X|Z_i}\left\{ \frac{\partial^3 m_Y(X, X-\pi(Z_i))}{\partial x \partial v^2} \right\} = 0.
\end{aligned}$$

Similarly, using $\xi = \varsigma + \gamma u$, we have

$$\begin{aligned}
&\mathbb{E}\left[\frac{\partial^2 m_D(X,V)}{\partial x \partial v}\right] \xi_i [R(X_i,V_i)]_{(5,1)} - \mathbb{E}\left[\frac{\partial^2 m_Y(X,V)}{\partial x \partial v}\right] u_i [R(X_i,V_i)]_{(5,1)} && (37) \\
=&\mathbb{E}\left[\frac{\partial^2 m_D(X,V)}{\partial x \partial v}\right] \varsigma_i [R(X_i,V_i)]_{(5,1)}. && (38)
\end{aligned}$$

Finally, the bias terms cancel each other also because of the form of $M$ and $B$. By combining and the results of Proposition 1 and Corollary A.1, and rearranging terms we get

$$\sqrt{n}(\hat{\gamma} - \gamma) = \frac{1}{\mathbb{E}\left[\frac{\partial^2 m_D(X,V)}{\partial x \partial v}\right]} \frac{2}{\sqrt{n}} \sum_{i=1}^{n} \varsigma_i [R(X_i,V_i)]_{5,1} + o_P(1)\,.$$

Thus by CLT, we have the desired result. ■

**Proof of Corollary 1:** By the proof of Proposition 1 and the fact

$$\frac{\partial^2 m_Y(X,V)}{\partial x \partial v} = \gamma \left[\frac{\partial^2 m_D(X,V)}{\partial x \partial v}\right] \quad \text{and} \quad \frac{\partial^3 m_Y(X,V)}{\partial x \partial v^2} = \gamma \left[\frac{\partial^3 m_D(X,V)}{\partial x \partial v^2}\right],$$

it is easy to see that any additional term caused by generated regressor is cancelled. Therefore, we have the identical result as the one in Theorem 1. ■