# Endogenous semiparametric binary choice models with heteroscedasticity 

Stefan Hoderlein

The Institute for Fiscal Studies Department of Economics, UCL<br>cemmap working paper CWP34/09

# Endogenous Semiparametric Binary Choice Models with Heteroscedasticity 

Stefan Hoderlein*<br>Brown University<br>First Draft: February 15, 2006<br>This Draft: June 4, 2009


#### Abstract

In this paper we consider endogenous regressors in the binary choice model under a weak median exclusion restriction, but without further specification of the distribution of the unobserved random components. Our reduced form specification with heteroscedastic residuals covers various heterogeneous structural binary choice models. As a particularly relevant example of a structural model where no semiparametric estimator has of yet been analyzed, we consider the binary random utility model with endogenous regressors and heterogeneous parameters. We employ a control function IV assumption to establish identification of a slope parameter $\beta$ by the mean ratio of derivatives of two functions of the instruments. We propose an estimator based on direct sample counterparts, and discuss the large sample behavior of this estimator. In particular, we show $\sqrt{n}$ consistency and derive the asymptotic distribution. In the same framework, we propose tests for heteroscedasticity, overidentification and endogeneity. We analyze the small sample performance through a simulation study. An application of the model to discrete choice demand data concludes this paper.


Keywords: Semiparametric, Binary Choice, Endogeneity, Average Derivative, Control Function, Random Coefficients.

[^0]
## 1 Introduction

The Model: The binary choice model constitutes a workhorse of modern microeconometrics and has found a great many applications throughout applied economics. It is commonly treated in a latent variable formulation, i.e.

$$
\begin{align*}
& Y^{*}=X^{\prime} \beta+U  \tag{1.1}\\
& Y=\mathbb{I}\left\{Y^{*}>0\right\}
\end{align*}
$$

where $Y^{*}$ is an unobserved continuously distributed random variable, in the classical choice literature often utility or differences in utility, $X$ is a random $K$-vector of regressors, $\beta$ is a $K$-vector of fixed coefficients, and $\mathbb{I}\{\cdot\}$ denotes the indicator of an event. Throughout much of the literature, and indeed in this paper, interest centers on the coefficient $\beta$ which summarizes the effect of a set of regressors $X$ on the dependent variable. If $U$ is assumed independent of $X$, and $U$ follows a certain parametric distribution then $\mathbb{E}[Y \mid X]=F_{U}\left(X^{\prime} \beta\right)$, where $F_{U}$ is the known parametric cdf of $U$, and estimation is straightforward via ML. Both assumptions are restrictive in many economic applications and have therefore come under critique. In particular, invoking these assumptions rules out that model (1.1) is the reduced form of individual behavior in a heterogeneous population, where parameters vary across the population in an unrestricted fashion, and it rules out endogeneity.

This paper aims at weakening these critical assumptions, while retaining an estimator with a simple and interpretable structure. In particular, it establishes interpretation and constructive identification of $\beta$ under assumptions that are compatible with, e.g., a heterogeneous population characterized by an unrestricted distribution of random utility parameters. The weakening of assumptions is twofold: First, we do not want to place restrictive parametricity or full independence assumptions on the distribution of the unobservables (or indeed any random variable in this model), and employ instead relatively weak median exclusion restrictions. Second, due to its paramount importance in applications we want to handle the case of endogenous regressors, e.g., we want to allow for $X$ to be correlated with $U$. The estimator we propose has a simple, "direct" structure, akin to average derivative estimator (ADE). A characteristic feature of this class of estimators is that they use a control function instrumental variables approach for identification. The identification result is constructive in the sense that it can be employed to yield a $\sqrt{n}$ consistent semiparametric estimator for $\beta$.

Main Identification Idea: Throughout this paper, we will be concerned with model (1.1). However, we view model (1.1) as a reduced form of a structural model in a heterogeneous population. As a consequence, we will also be concerned with the providing an example for the interpretation of $\beta$ when employing sensible independence restrictions in the heterogeneous structural model.

The independence restriction we are invoking in model (1.1) is a conditional median exclusion restriction. Specifically, we introduce a $L$ random vector of instruments, denoted $Z$, and assume that they are related to $X$ via

$$
\begin{equation*}
X=\vartheta(Z)+V \tag{1.2}
\end{equation*}
$$

where $\vartheta$ is a smooth, but unknown function of $Z$. For instance, in the special case where out of $K$ continuously distributed regressors $\left(X^{1}, . ., X^{K}\right)$ only $X^{1}$ is endogenous and there is exactly one additional instrument denoted by $Z^{1}$ (so that $Z=\left(Z^{1}, X^{2} . ., X^{K}\right)^{\prime}$ ), $\vartheta$ could be a continuous version of the mean regression $m_{X}(z)=\mathbb{E}[X \mid Z=z]=\left(\mathbb{E}\left[X^{1} \mid Z=z\right], X^{2}, \ldots, X^{K}\right)^{\prime}$, in which case $V$ would be the mean regression residuals in the first equation. Alternatively, it could also be a vector containing the conditional $\alpha$ quantile of $X^{1}$ conditional on $Z$ as first element, provided this is considered a more plausible specification for the application at hand ${ }^{1}$.

If we let the conditional median of $U$ given $Z=z$ and $V=v$ be denoted by $k_{U \mid Z V}^{0.5}(z, v)$, then we may formalize our identifying assumption as

$$
k_{U \mid Z V}^{0.5}(z, v)=g(v)
$$

for all $(z, v)$ in its support. What does this assumption mean in economic terms, and why is it a sensible assumption if we think of a heterogeneous population? In section 2 we show that this assumption is implied by a heterogeneous random utility model with endogeneity arising from omitted variables. In this case, the median exclusion restriction is implied if instruments are (jointly) independent of omitted variables and of $V$, but it holds also under weaker restrictions.

What economic interpretation of $\beta$ is implied by our assumptions? Taking the binary choice model with random utility parameters as an example, in the second section we establish the following: 1. If we are willing to assume conditionally symmetric random parameters, we obtain that $\beta$ has the interpretation of a mean of marginal effects. 2 . In the absence of symmetry we show that $\beta$ has the interpretation of a local average structural derivative (see Hoderlein (2005, 2009) and Hoderlein and Mammen (2007)).

Given that we have devised a sensible identification restriction and defined an interesting structural parameter, the question that remains to be answered is how to actually identify and estimate this parameter. To answer this question, we introduce the following notation: Let $\bar{Y}=k_{Y \mid Z V}^{0.5}(Z, V)=\mathbb{I}\{\mathbb{P}[Y=0 \mid Z, V]<0.5\}$ denote the conditional median of $Y$ given $Z$ and $V$, and assume that $\vartheta(z)=\mathbb{E}[X \mid Z=z]$. Then, under assumptions to be detailed below,

$$
\begin{equation*}
\beta=\mathbb{E}\left[\left[D_{z} \mathbb{E}[X \mid Z]^{\prime}\right]^{-1} \nabla_{z} \mathbb{E}[\bar{Y} \mid Z] B(Z)\right], \tag{1.3}
\end{equation*}
$$

[^1]where $\nabla_{z}$ and $D_{z}$ denote gradient and Jacobian, and $B(z)$ denotes a bounded weighting function to be defined below. Intuitively, the identification follows by a combination of arguments employed to identify average derivatives (see Powell, Stock and Stoker (1989), PSS, for short), and the chain rule, and is only up to scale. This identification principle is constructive, and yields in a straightforward fashion a sample counterparts estimator, see equation (4.3) below. Because of its direct structure, the estimator shares all advantages of direct estimators. In particular, the estimator is robust to misspecification and avoids computationally difficult optimization problems involving nonconvex objective functions. Moreover, the estimator is $\sqrt{n}$ consistent, and has a standard limiting distribution.

Additional Contributions: Beyond constructing an estimator for a sensible parameter in a heterogeneous population, the flexibility of the model enables us to check the specification for several issues that have not been considered exhaustively, if at all, in the literature on this type of models. For instance, we propose powerful tests for endogeneity and heteroscedasticity. Another important issue we discuss is overidentification. As will turn out, in a general nonseparable setup overidentification is markedly different from the issue in the linear framework. In addition to clarifying the concept, we propose a Hausman type test for overidentification. We develop a semiparametric notion of weakness of the instruments, and establish how our approach allows to mitigate the problem of weak instruments. Finally, we show that our approach allows to handle discrete and continuous endogenous regressors.

Literature: The binary choice model (1.1) with exogenous regressors has been analyzed extensively in the semiparametric literature, most often via single index models. Since this paper employs a direct estimator, our approach is related to contributions by PSS (1989), Hristache, Juditsky and Spokoiny (2001) and Chaudhuri, Doksum and Samarov (1997), to mention just a few. The main alternative are "optimization", or $M$-, estimators for $\beta$, including semiparametric LS (Ichimura (1993)), semiparametric ML (Klein and Spady (1993)), and general $M$-estimators (Delecroix and Hristache (1997). None of these of estimators can handle general forms of heteroscedasticity even in the exogenous setting, and to do so one has to employ maximum score type estimators, see Manski (1975). But these estimators have a slow convergence rate and a nonstandard limiting distribution, and only the estimator of Horowitz (1992) achieves almost $\sqrt{n}$ convergence.

In spite of the wealth of literature about model (1.1) in the exogenous case, and the importance of the concepts of endogeneity and instruments throughout econometrics, the research on model (1.1) with endogenous regressors has been relatively limited. However, there are important contributions that deserve mentioning. For the parametric case, we refer to Blundell and Smith (1986) and Rivers and Voung (1988). For the semiparametric case, Lewbel proposes the concept of special regressors, i.e. one of the regressors is required to have infinite support,
which is essential for identification (Lewbel (1998)). Our approach is more closely related to the work of Blundell and Powell (2004, BP, for short), and Petrin and Train (2006). Like BP, we use a control function assumption to identify the model, but as already mentioned in a different fashion. This makes our approach also weakly related to other control function models in the semiparametric literature, most notably Newey, Powell and Vella (1998) and Das, Newey and Vella (2003). Related is also Matzkin (1992, 2005), who was the first to consider nonparametric identification in the exogenous binary choice model, and Bajari, Fox, Kim and Ryan (2008), who propose a nonparametric estimator of the distribution of random coefficients. Finally, our work is also related to Ai and Chen (2001), Vytlacil and Yildiz (2007), and in particular the "Local Instruments" approach of Heckman and Vytlacil (2005) and Florens, Heckman, Meghir and Vytlacil (2008) for analyzing treatment effects.

Organization of Paper: We start out in the next section by considering the case of a linear random utility model with heterogeneous parameters. We derive the median exclusion restriction formally, and establish the interpretation of $\beta$ stated above. In section three we state formally the assumptions required for identification of $\beta$ and provide a discussion. Moreover, we establish identification both in the heteroscedastic as well as the homoscedastic case (we require the latter among other things to test the random utility parameters specification). This identification principle is constructive in the sense that it yields direct estimators through sample counterparts. The asymptotic distribution of these estimators is in the focus of the fourth section. Specifically, we establish $\sqrt{n}$ consistency to a standard limiting distribution ${ }^{2}$. Beyond suggesting a $\sqrt{n}$ consistent estimator, the general identification principle is fruitful in the sense that it allows to construct tests for endogeneity, heteroscedasticity and overidentification, and this will be our concern in the fifth section. A simulation study underscores the importance of correcting for endogeneity and heterogeneity and will be discussed in the sixth section. In the seventh section, we apply our methods to a real world discrete choice demand application: We consider the decision to subscribe to cable TV, using data similar to Goolsbee and Petrin (2004). Finally, this paper ends with a conclusion and an outlook.

[^2]
## 2 A Structural Example: Binary Demand Decisions in a Heterogeneous Population

The question that should be answered for any reduced form microeconometric model is how it can be derived from individual behavior in a heterogeneous population. To answer this identification question for the one defined through (1.1), we start out with a general nonseparable model of a heterogeneous population as in Hoderlein (2005, 2009) or Hoderlein and Mammen (2007). The most general version of (1.1) has the structural unobservables (e.g., preferences) influencing the latent variable in a nonseparable fashion, i.e. $\quad Y^{*}=\phi(X, A)$, where $A \in \mathfrak{A}$ denotes the unobservables. Here $\mathfrak{A}$ is a Borel space, e.g., the space of piecewise continuous utility functions. Note that $A$ may include objects like preferences, but also other omitted determinants. In our example, we denote the former by $A_{1}$, while the remainder of $A$ is denoted by $A_{2}$. In discrete choice demand analysis for instance, $A_{2}$ are often omitted characteristics of the product.

Specific Model: While we could proceed to discuss the model on this level of generality, in this paper we restrict ourselves to linear models on individual level, largely because linear models are the dominating class of models in economic applications, and leave the most general case for a companion paper. A linear heterogeneous population with omitted variables $A_{2}$ may then be formalized through a random coefficient model, i.e.,

$$
\begin{align*}
Y^{*} & =X^{\prime} \beta\left(A_{1}\right)+A_{2}^{\prime} \gamma\left(A_{1}\right)  \tag{2.1}\\
Y & =\mathbb{I}\left\{Y^{*}>0\right\}
\end{align*}
$$

where $\theta\left(A_{1}\right)=\left(\beta\left(A_{1}\right)^{\prime}, \gamma\left(A_{1}\right)^{\prime}\right)^{\prime}$ is a mapping from the space of unobservables (say, preferences) $\mathfrak{A}_{1} \subseteq \mathfrak{A}$ into $\mathbb{R}^{K}$. Since we assume the random elements $A_{1}$ (in our example, preferences) to vary across the population, $\theta\left(A_{1}\right)$ varies across the population, too. This model admits a reduced form representation as (1.1). What are now plausible stochastic conditions that we would like to impose on the reduced form (1.1) to identify $\beta$, and how can they be derived from restrictions in the structural model (2.1), and in particular on the random coefficients $\theta\left(A_{1}\right)$ ?

We answer this question under the assumption that the endogeneity arises from potential correlation of $X$ and $A_{2}$ only, and that $A_{1}$, e.g., the unobservable preferences that determine the parameters, are independent from all economic variables in the system, i.e. $A_{1} \perp\left(X, Z, A_{2}\right)$. In discrete choice demand analysis for instance, this endogenous regressor is the own price of the good, which is assumed to be correlated with its omitted unobserved characteristics contained in $A_{2}$, see Goolsbee and Petrin (2004).

An unnecessarily strong, but economically plausible identifying independence restriction is the independence of instruments $Z$ from all unobservables in the system, i.e., $Z \perp(A, V)$. Con-
tinuing our discrete choice example, the discrete choice literature suggests to use the wholesale price, franchise fees, or other regional supply side characteristics of a market as instruments. It is plausible that these instruments are independent of individual preferences and omitted characteristics of the product. This assumption implies that $X \perp A_{2} \mid V$, and recall that our maintained hypothesis is that $A_{1} \perp\left(X, V, A_{2}\right)$ which is implied by $A_{1} \perp\left(X, Z, A_{2}\right)$, see also Petrin and Train (2006) for similar arguments, and discussions on when an additive IV equation arises.

Interpretation of $\beta$ : The following result states that these independence conditions imply an exclusion restriction that defines a sensible centrality parameter of the distribution of random coefficients in (2.1). For the result, we require the notation $\mathfrak{B}=\beta\left(A_{1}\right), \mathfrak{C}=A_{2}^{\prime} \gamma\left(A_{1}\right)$, $U=X^{\prime}\left(\beta\left(A_{1}\right)-\beta\right)+A_{2}^{\prime} \gamma\left(A_{1}\right)$, and let $\mathbb{E}_{X V}[\cdot]$ denotes integration over the distribution of $(X, V)$. Since this section is motivational, the statement of the theorem is informal (i.e., excluding regularity conditions)

Theorem 1. Let the model defined by equations (1.1) and (1.2) be the reduced form of the structural model defined in equations (1.2) and (2.1). Suppose that $A_{1} \perp\left(X, Z, A_{2}\right)$ and $Z \perp$ $(A, V)$ hold. Assume further that, conditional on $(X, V): 1$. $(\mathfrak{B}, \mathfrak{C})$ are jointly symmetrically distributed about $(\beta, \mathbb{E}[\mathfrak{C} \mid V])$, and 2. $U$ is absolutely continuously distributed with respect to Lebesque measure. Finally, assume that regularity conditions hold such that all objects exist and are well defined. Then follows that $k_{U \mid Z V}^{0.5}(Z, V)=k_{U \mid X V}^{0.5}(X, V)=g(V)$ and

$$
\beta=\mathbb{E}\left[\beta\left(A_{1}\right)\right] .
$$

If we dispense with the conditional symmetry assumption, and start out with the median exclusion restriction $k_{U \mid Z V}^{0.5}(Z, V)=g(V)$, then we obtain that $k_{U \mid X V}^{0.5}(X, V)=g(V)$ and

$$
\begin{equation*}
\beta=\mathbb{E}_{X V}\left[\mathbb{E}\left[\beta\left(A_{1}\right) \mid X, V, Y^{*}=k_{Y^{*} \mid X V}^{0.5}(X, V)\right]\right] \tag{2.2}
\end{equation*}
$$

Therefore, we conclude that our economically plausible independence assumptions together with symmetry of the distribution of random components imply a median exclusion restriction and define a parameter $\beta$ that is the mean of the distribution of random coefficients of the observable regressors. More generally, if we just assume the median exclusion restriction $k_{U \mid Z V}^{0.5}(Z, V)=g(V)$ but dispense with the symmetry assumption, we obtain that the coefficient $\beta$ has the interpretation of a local average structural derivative, i.e., $\beta=$ $\mathbb{E}\left[\beta\left(A_{1}\right) \mid X=x, V=v, Y^{*}=k_{Y^{*} \mid X V}^{0.5}(x, v)\right]$, for all $(x, v) \in \operatorname{supp}(X) \times \operatorname{supp}(V)$. Due to the linear random coefficient structure with exogenous $A_{1}$, this quantity is invariant to changes in $x, v$, and hence we may integrate over $x, v$, keeping the quantile of the unobservable latent variable fixed at the median, i.e., at the center of the conditional distribution. If we identify this center of the distribution with a type of individuals (the "median" person), then we may
speak of $\beta$ as an average structural effect for this type. Another more statistical interpretation of (2.2) is that of a best approximation to the underlying heterogeneous coefficient $\beta\left(A_{1}\right)$, conditioning on all the information that we have to our disposal in the data ${ }^{3}$. In what follows, we will treat our model under the assumption that $k_{U \mid Z, V}^{\alpha}(Z, V)=g(V)$, with probability one, so that the latter interpretation is the most adequate. The role of the symmetry assumption is to point out that under stronger assumptions $\beta$ reduces to a completely standard object.

Average Structural Function: In the exogenous binary choice model, it is often derivatives of the conditional probabilities (sometimes called "elasticities") that are of interest, because they show the nonlinearity in the causal effect. Blundell and Powell (2004) extend this notion to the endogenous setup without heteroscedasticity (e.g., random coefficients). They propose to integrate the conditional probability (conditioning is with respect to $X, V$ ) over $V$. In a heterogeneous population, integration can be seen as a step where an average is formed over the unobserved heterogeneity. As shown in theorem 1, under the weak median exclusion restriction only the structural effect of the median person is identified, but the behavior for individuals in the tails of the distribution of unobservables is left unidentified. Hence integrating over the entire distribution of unobserved heterogeneity is not possible.

However, what is possible is to integrate out the control function residuals $V$, which are accounting for the endogenous components of $U$. A natural analog to Blundell and Powell's (2004) average structural function is the following average median structural function:

$$
A M S F(x)=\int k_{Y \mid X V}^{0.5}(x, v) F_{V}(d v)=\int \mathbb{I}\left\{x^{\prime} \beta+g(V)>0\right\} F_{V}(d v)
$$

The derivative of this quantity has the interpretation of a partial integral over the local average structural derivative (in the general case with nonsymmetric random coefficients), integrating over $v$, but keeping $x$ fixed. In light of the above theorem this could be interpreted as the nonlinear structural effect on the median person, but we leave a detailed discussion of this quantity for future research, and focus on $\beta$ in this paper.

[^3]
## 3 Details of the Estimator in the Endogenous Binary Choice Random Coefficients Model

### 3.1 Identification via Median Restriction on $U$

Throughout this section, and indeed through much of the paper, we require the following notation: Let the $K \times L$ matrix of derivatives of a $K$-vector valued Borel function $g(z)$ be denoted by $D_{z} g(z)$, and let $\nabla_{z} g(z)$ denote the gradient of a scalar valued function. Denote by $m_{Y \mid Z V}(z, v)$ a continuous version of $\mathbb{E}[Y \mid Z=z, V=v]$, and let $f_{A}(a), f_{A B}(a, b)$ and $f_{A \mid B}(a ; b)$ be the marginal, joint and conditional Radon-Nikodym density of the random vectors $A$ and $B$ with respect to some underlying measure $\mu$, which may be the counting or the Lebesgue measure, (i.e., $A$ may be discretely or continuously distributed). Define the nonparametric score $Q_{z}(v, z)=\nabla_{z} \log f_{V \mid Z}(v ; z)$. Let $k_{S \mid Z}^{\alpha}(z)$ denote the conditional $\alpha$-quantile of a random variable $S$ given $Z=z$, i.e. for $\alpha \in(0,1), k_{S \mid Z}^{\alpha}(z)$ is defined by $\mathbb{P}\left(Y \leq k_{S \mid Z}^{\alpha}(z) \mid Z=z\right)=\alpha$. Let $G^{-}$denote the Moore-Penrose pseudo-inverse of a matrix $G$. Finally, let $c_{k}, k=1,2, \ldots$ denote generic constants, and we suppress the arguments of the functions whenever convenient.

As already discussed in the introduction, the main idea is now that instead of running a regression using $Y$, we employ $\bar{Y}=k_{Y \mid Z V}^{0.5}(Z, V)$, i.e. the conditional median of $Y$ given $Z$ and $V$ (which is the $L_{1}$-projection of $Y$ on $\mathcal{Z} \times \mathcal{V}$ ), and consider the $L_{2}$-projection of $\bar{Y}$ on $\mathcal{Z}$. Consequently, we consider weighting functions defined on $\mathcal{Z}$ only. In the following two subsections we first list and discuss all assumptions that specify the true population distribution and the DGP, and then establish the role they play in identifying $\beta$. Readers less interested in the econometric details of this model may skip these subsections, and proceed directly to the main result (theorem 2).

### 3.1.1 Assumptions

Assumption 1. The data $\left(Y_{i}, X_{i}, Z_{i}\right), i=1, \ldots, n$ are independent and identically distributed such that $\left(Y_{i}, X_{i}, Z_{i}\right) \sim(Y, X, Z) \in \mathcal{Y} \times \mathcal{X} \times \mathcal{Z} \subset \mathbb{R}^{1+K+L}$ The joint distribution of $(Y, X, Z)$ is absolutely continuous with respect to a $\sigma$-finite measure $\mu$ on $\mathcal{Y} \times \mathcal{X} \times \mathcal{Z}$ with Radon-Nikodym density $f_{Y X Z}(y, x, z)$. The underlying measure $\mu$ can be written as $\mu=\mu_{Y} \times \mu_{X Z}$, where $\mu_{X Z}$ is the Lebesgue measure.

Assumption 2. The weighting function $B(z)$ is nonzero and bounded with compact support $\mathcal{B} \subset \mathcal{Z}$, where usually $\mathcal{Z}=\mathbb{R}^{L}$.

Assumption 3. $\vartheta(z)$ is continuously differentiable in the components of $z$ for all $z \in \operatorname{Int}(\mathcal{B})$. $\left[D_{z} \vartheta(z)^{\prime}\right]^{-}$exists and every element is bounded from below for all $z \in \mathcal{B} .\left[D_{z} \vartheta(Z)^{\prime}\right]^{-}$is square integrable on $\mathcal{B}$.

Assumption 4. $\mathbb{E}[Y \mid Z=z]=m_{Y \mid Z}(z)$ is continuously differentiable in the components of $z$ for all $z \in \operatorname{Int}(\mathcal{B}) . D_{z} m_{Y \mid Z}(Z)$ is square integrable on $\mathcal{B} . g(z, v)=F_{U \mid V}\left(\vartheta(z)^{\prime} \beta+v^{\prime} \beta ; v\right) f_{V \mid Z}(v ; z)$ is bounded in absolute value by a nonnegative integrable function $q(z)$, for all $z \in \mathcal{B}$.

Assumption 5. $\mathbb{E}[\bar{Y} \mid Z=z]=m_{\bar{Y} \mid Z}(z)$ is continuously differentiable in the components of $z$ for all $z \in \operatorname{Int}(\mathcal{B}) . D_{z} m_{\bar{Y} \mid Z}(Z)$ is square integrable on $\mathcal{B}$. Moreover $0<\mathbb{P}[\bar{Y}=1 \mid Z=z]<1$ for all $z \in \mathcal{B}$.

For the stochastic terms $U$ and $V$, the following holds:
Assumption 6. $U$ and $V$ are jointly continuously distributed.
In addition, either of the following hold:
Assumption 7. $U$ is independent of $Z$ given $V$.
Assumption 8. 1. $k_{U \mid Z V}^{0.5}(Z, V)=g(V)$.
2. Let $\tilde{V}=l(V)=-\left(g(V)+V^{\prime} \beta\right)$. Either of the following hold:
a. $\tilde{V}$ is independent of $Z$. Moreover, $\tilde{V}$ is absolutely continuously with respect to Lebesgue measure, with Radon-Nikodym density $f_{\tilde{V}}$. $f_{\tilde{V}}(\varpi)$ is differentiable for all $\varpi \in$ im $(l)$. Finally, $f_{\tilde{V}}\left(D_{z} \vartheta(Z)^{\prime} \beta\right)$ is absolutely integrable on $\mathcal{B}$.
b. There is one endogenous regressor $X^{k}$, and $l$ is a continuous piecewise invertible function. Moreover, $f_{V \mid Z}(v, z)$ and its partial derivatives wrt the components of $z$ are bounded on $\mathcal{B}$ from below and above, i.e. $c_{1}>\sup _{(v, z) \in \operatorname{supp}(V) \times \mathcal{B}} f_{V \mid Z}(v, z) \geq \inf _{(v, z) \in \operatorname{supp}(V) \times \mathcal{B}} f_{V \mid Z}(v, z)=$ $c_{2}>0$, and $c_{3}>\sup _{(v, z) \in \operatorname{supp}(V) \times \mathcal{B}}\left\|\nabla_{z} f_{V \mid Z}(v, z)\right\| \geq \inf _{(v, z) \in \operatorname{supp}(V) \times \mathcal{B}}\left\|\nabla_{z} f_{V \mid Z}(v, z)\right\|=$ $c_{4}>0$. Finally, let $Q_{z}(V, Z)$ be absolutely integrable on $\operatorname{supp}(V) \times \mathcal{B}$, and let $\tau(z)=$ $\mathbb{E}\left[\bar{Y} Q_{z}(V, Z) \mid Z=z\right]$ be square integrable on $\mathcal{B}$.
c. Let 8.36 hold, but instead of one endogenous regressor, assume there are many endogenous regressors $X^{1}, . ., X^{K_{1}}, K_{1} \leq K$, and in addition $g(v)=v^{\prime} \gamma$, with $\gamma \in \mathbb{R}^{K_{1}}$.

Remark 3.1-Discussion of Assumptions: Starting with assumption 1, we assume to possess continuously distributed instruments and regressors. Strictly speaking, we do not even require continuous instruments for identification, but only for the specific direct estimator we propose. Indeed, we conjecture that a direct estimator akin to Horowitz and Haerdle (1998) in the exogenous single index model may be devised, but this is beyond the scope of this paper. The iid assumption is inessential and may be relaxed to allow for some forms of stationary time series dependence. Note that unlike Blundell and Powell (2004), we do not require the support of $V \mid X$ to be invariant in $X$, which is why endogenous binary regressors are ruled out in their
case. However, note that the assumption that $l$ be invertible or single index.in assumption 8.2 or 8.3 effectively rules out binary endogenous regressors.

Regarding the choice of weighting function $B$, due to assumption 2 we delete all observations outside a fixed multivariate interval $I_{z}$. As such, the weighting is unrestrictive and merely serves as a devise to simplify already involved derivations below. It could be abandoned at the price of a vanishing trimming procedure. In addition we require that $\left[D_{z} \vartheta(z)^{\prime}\right]^{-}$exists and is bounded on $\mathcal{B}$ (cf. assumption 3), and hence we choose $B(z)=\mathbb{I}\left\{z \in I_{z}\right\} \mathbb{I}\left\{\operatorname{det}\left|D_{z} \vartheta(z) D_{z} \vartheta(z)^{\prime}\right| \geq b\right\}$, with $b>0$. By choosing the weighting function and the region $\mathcal{B}$ appropriately we may ensure that the instruments are not weak in the sense that $\operatorname{det}\left|D_{z} \vartheta(z) D_{z} \vartheta(z)^{\prime}\right| \geq b$ for some subset of $\mathcal{Z}$ with positive measure. If we view the derivative in a linear regression of $X$ on $Z$ as an average derivative, it may be the case that instruments are on average not strongly related to endogenous regressors, but are quite informative for $\beta$ in certain areas of $\mathcal{Z}$ space. We consider it to be an advantage of our nonparametric approach that we can concentrate on those areas, and hence suggest that a similar weighting be performed in applications. However, in applications $\mathcal{B}$ is usually not known, implying that a threshold $b$ be chosen and $D_{z} \vartheta(z)$ be pre-estimated ${ }^{4}$.

Particularly novel is assumption 8.1. Instead of the full independence of $U$ and $Z$ conditional on $V$ assumed in assumption 7 (and implying the Blundell and Powell (2004) assumption $U \perp X \mid V)$ this assumption (only) imposes a conditional location restriction. Hence it allows for all other quantiles of $U$ than the median to depend on $Z$ and $V$, and thus on $X$, in an arbitrary fashion, which as we have seen in the introduction is sensible when unobserved heterogeneity is modelled. Assumption 8.2a covers the case when $\tilde{V}$ is independent of $Z$, in which case the function $l$ need not be restricted at all. The other assumptions $8.2 \mathrm{~b}-8.2 \mathrm{c}$ allow for arbitrary dependence between $V$ and $Z$ at the expense of placing some structure on $l$. Note, In the case of a single endogenous regressor this structure is very general: indeed, any continuous and piecewise invertible function will do. If there are many endogenous regressors, we still obtain identification in the important examples when $l$ is of single index form (a combination of this assumption with assumption 8.2 b allows for $l$ being a piecewise invertible function of an index).

[^4]
### 3.1.2 Essential Arguments in the Identification of $\beta$ in the Heteroscedastic Case

To see how assumptions 1-8 help in identifying $\beta$, rewrite the model as follows

$$
\begin{equation*}
Y=\mathbb{I}\left\{(\vartheta(Z)+V)^{\prime} \beta+U>0\right\} \tag{3.1}
\end{equation*}
$$

Note first that under assumption 8.3 the conditional median $\bar{Y}$ becomes

$$
\begin{equation*}
\bar{Y}=\mathbb{I}\left\{\vartheta(Z)^{\prime} \beta+k_{U \mid Z V}^{0.5}(Z, V)+V^{\prime} \beta>0\right\}=\mathbb{I}\left\{\vartheta(Z)^{\prime} \beta>l(V)\right\}, \tag{3.2}
\end{equation*}
$$

as $\mathbb{I}$ is a monotonic function. This very much resembles the standard model, but with $\bar{Y}$ instead of $Y$. However, note two complications: first $\tilde{V}=l(V)$ may not be fully independent of $Z$, second, $l$ is unknown. We now establish that $\beta$ is nevertheless constructively identified in this setup.

To do so, we start with the case when $\tilde{V}$ is fully independent of $Z$, i.e., the scenario is as given in 8.3a. Then,

$$
\begin{equation*}
m_{\bar{Y} \mid Z}(z)=\mathbb{E}[\bar{Y} \mid Z=z]=\mathbb{P}\left\{\vartheta(z)^{\prime} \beta>l(V)\right\} \tag{3.3}
\end{equation*}
$$

due to standard arguments. To focus now on the essential arguments, we consider only a compact set $\mathcal{B} \subset \mathcal{Z}$ and a nonzero and bounded weighting function $B(z)$ with support $\mathcal{B}$, see assumption 2. Since $m_{\bar{Y} \mid Z}(z)$ and $\vartheta(z)$ are continuously differentiable in all components of $Z$, for all $z \in \mathcal{B}$, we obtain by the chain rule

$$
\begin{equation*}
\nabla_{z} m_{\bar{Y} \mid Z}(Z)=f_{l(V)}\left(\vartheta(Z)^{\prime} \beta\right) D_{z} \vartheta(Z)^{\prime} \beta \tag{3.4}
\end{equation*}
$$

with probability one. This steps rules out that $X$ contains a constant. Moreover, note that $f_{l(V)}$ is a scalar valued function. Next, we premultiply equation (3.4) by the generalized inverse $\left[D_{z} \vartheta(Z)^{\prime}\right]^{-}$, which exists on $\mathcal{B}$ due to assumption 3, and the weighting function $B(z)$ to obtain

$$
\begin{equation*}
\left[D_{z} \vartheta(Z)^{\prime}\right]^{-} \nabla_{z} m_{\bar{Y} \mid Z}(Z) B(Z)=\beta f_{l(V)}\left(\vartheta(Z)^{\prime} \beta\right) B(Z) \tag{3.5}
\end{equation*}
$$

or, upon taking expectations,

$$
\begin{equation*}
\beta c_{1}=\mathbb{E}\left[\left[D_{z} \vartheta(Z)^{\prime}\right]^{-} \nabla_{z} \mathbb{E}[\bar{Y} \mid Z] B(Z)\right], \tag{3.6}
\end{equation*}
$$

where $c=\mathbb{E}\left[f_{l(V)}\left(\vartheta(Z)^{\prime} \beta\right) B(Z)\right]$. From now on, we will tacitly suppress this constant, so that identification is only up to scale. This last step is warranted, because the elementwise square integrability of all functions on $\mathcal{B}$ (assumption 5), together with Cauchy-Schwarz ensures that the expectations exist. The identification of $\beta$ in the case when $V$ and $Z$ are not fully independent (i.e., equation (3.8) below) is harder to show, and left to the appendix.

### 3.1.3 Main Identification Results

The following theorem summarizes the discussion in the previous section and in the appendix. We use th notation $\breve{Y}=\mathbb{E}[Y \mid Z, V]$.

Theorem 2. (i) Let the true model be as defined in 1.1 and 1.2, and suppose that assumptions 1-3, 5-6 and 8.1-8.2a hold. Assume further that $\mathbb{E}\left[f_{\bar{V} \mid Z}\left(\vartheta(Z)^{\prime} \beta ; Z\right) B(Z)\right]=1$. Then $\beta$ is identified by relationship:

$$
\begin{equation*}
\beta=\mathbb{E}\left[\left[D_{z} \vartheta(Z)^{\prime}\right]^{-} \nabla_{z} \mathbb{E}[\bar{Y} \mid Z] B(Z)\right] . \tag{3.7}
\end{equation*}
$$

(ii) If instead of assumption $8.2 a$ either of assumptions $8.26-8.2 c$ hold, then we obtain that $\beta$ is identified up to scale by

$$
\begin{equation*}
\beta=\mathbb{E}\left[\left[D_{z} \vartheta(Z)^{\prime}\right]^{-}\left\{\nabla_{z} \mathbb{E}[\bar{Y} \mid Z]-\mathbb{E}\left[\bar{Y} Q_{z}(V, Z) \mid Z\right]\right\} B(Z)\right], \tag{3.8}
\end{equation*}
$$

where $Q_{z}(V, Z)$ denotes the nonparametric score $\nabla_{z} \log f_{V \mid Z}(V ; Z)$.
(iii) If we strengthen the conditional median independence assumption 8 to the full independence assumption 7 and assume that assumption 4 holds, we obtain that in addition to (3.8), $\beta$ is (up to scale) identified by

$$
\begin{equation*}
\mathbb{E}\left[\left[D_{z} \vartheta(Z)^{\prime}\right]^{-}\left\{\nabla_{z} \mathbb{E}[Y \mid Z]-\mathbb{E}\left[Y Q_{z}(V, Z) \mid Z\right]\right\} B(Z)\right] \tag{3.9}
\end{equation*}
$$

as well as by

$$
\begin{equation*}
\mathbb{E}\left[\left[D_{z} \vartheta(Z)^{\prime}\right]^{-}\left\{\nabla_{z} \mathbb{E}[\breve{Y} \mid Z]-\mathbb{E}\left[\breve{Y} Q_{z}(V, Z) \mid Z\right]\right\} B(Z)\right] . \tag{3.10}
\end{equation*}
$$

Remark 3.2-Interpretation of Theorem 2: First, consider the scenario where $V$ and $Z$ are independent which gives rise to (3.7). $\beta$ is identified by a weighted average ratio of derivatives, involving the derivatives of the function $\vartheta$, and of the mean regression of $\bar{Y}$ (i.e., the conditional median given $Z$ and $V$ ), on $Z$ alone. Note that the control residuals $V$ do not appear in this equation, however, the model relies on correct specification of the conditional median restriction and of $\vartheta$. Allowing $\vartheta$ to be a conditional mean or a quantile enables the applied researcher to choose between various specifications of the IV equation, in order to select the one with the best economic interpretation (or one that works if the endogenous regressors do not have moments).

This identification result is constructive in the sense that it suggests in a straightforward fashion a sample counterpart estimator by replacing all functions with nonparametric estimators and the expectation by the average. While we always obtain a term of the form $\mathbb{E}\left[\left[D_{z} \vartheta(Z)^{\prime}\right]^{-} \nabla_{z} \mathbb{E}[\bar{Y} \mid Z] B(Z)\right]$, note that in the more general case where $V$ and $Z$ are allowed to be dependent we obtain an additional correction term, i.e. $\mathbb{E}\left[\left[D_{z} \vartheta(Z)^{\prime}\right]^{-} \mathbb{E}\left[\bar{Y} Q_{z}(V, Z) \mid Z\right] B(Z)\right]$,
which accounts for the higher order dependence in the IV equation. The same applies in the full independence scenario, i.e., when $U \perp V \mid Z$.

It is instructive to compare the heteroscedastic case with the case when $U \perp V \mid Z$. Observe that the independence assumption 7 implies assumption 8, so that equation (3.8) remains valid. But we obtain in addition that $\beta$ is identified up to scale by (3.9) and (3.10). Under full independence, we have thus a battery of potential estimating equations, where we could either use directly an $L_{2}$-projection of $Y$ on $Z$, or use a two projection strategy, where we use $L_{1}$, respectively, $L_{2}$-projections of $Y$ on $(Z, V)$ in the first stage, and then use a $L_{2}$-projection in the second stage. As shown below, we are able to obtain a powerful test for heteroscedasticity out of a comparison.

## 4 A Sample Counterpart Estimator: Asymptotic Distribution and Conditions for $\sqrt{n}$ Consistency

### 4.1 The Case for Direct Estimation

As mentioned above, the identification principle does not necessarily imply that we have to use a direct estimator. Indeed, in the case where assumption 8.2 a holds (i.e., $\tilde{V} \perp Z$ ), we could base an optimization estimator on equation (3.2), i.e.

$$
\begin{equation*}
\left.\bar{Y}=\mathbb{I}\left\{\vartheta(Z)^{\prime} \beta>\tilde{V}\right)\right\} . \tag{4.1}
\end{equation*}
$$

However, there are a number of reasons to use direct estimators here. Several have already been mentioned: First, they are natural because they build upon sample counterparts of the identification result. Consequently, their mechanics is easily understood, which makes them accessible to applied people. Moreover, several related issues (like overidentification) can be discussed straightforwardly. Second, they are robust to certain forms of misspecification. Third, they avoid the optimization of a highly nonlinear function, which both may not lead to global maxima (sometimes not even to well defined ones, if the semiparametric likelihood is flat), and may be computationally very expensive. Finally, and perhaps most importantly, in the case where $\tilde{V}$ is not fully independent of $Z$, it is not even clear how to construct an optimization estimator.

There are, however, also reasons that speak against the use of kernel based direct estimators. One of the theoretical arguments against them is that they require higher order smoothness assumptions, as will be obvious below. Note, however, that in the general setup with unrestricted (nonparametric) IV equation $X=\vartheta(Z)+V$, there is a "diminished smoothness gap". Any optimization estimator depends on an estimator $\hat{V}$ of $V$ as a regressor. In the general
nonparametric setup, this is a function of a nonparametric estimator for $\vartheta$. Using results in Newey (1994), it is straightforward to see that for a $\sqrt{n}$ consistent estimator of $\beta$ we require that $\mathbb{E}[\hat{\vartheta}]-\vartheta=o_{p}\left(n^{-1 / 2}\right)$ for the "no bias condition" to hold. In the kernel case this is, however, only possible under smoothness assumptions on $\vartheta$ which are very similar to the ones we require to hold for our direct estimator, in particular undersmoothing.

The second main drawback of direct estimators is the lack of efficiency compared to optimization estimators. Improving the efficiency, however, is possible, as we show in a companion paper (Hoderlein (2008)) for the full independence case, where we advocate so called one step efficient estimators. Alternatively, as in Newey and Stoker's (1994) analysis of the weighted average derivative estimators, we can define optimal weights.

### 4.2 A Sample Counterpart Estimator for $\beta$

In this section, we discuss the behavior of a sample counterpart estimator to (3.7) under independence of $V$ from $Z$, and we leave the more involved analysis that includes the correction term $\mathbb{E}\left[\left[D_{z} \vartheta(Z)^{\prime}\right]^{-} \mathbb{E}\left[\bar{Y} Q_{z}(V, Z) \mid Z\right] B(Z)\right]$ for future research. Moreover, throughout this section, we focus on the case when $\vartheta(z)=m_{X \mid Z}(z)$, i.e., $\vartheta$ is the nonparametric mean regression, and we leave the quantile regression for future work. Finally, we take $V$ as given and do not treat the effect of pre-estimation of $V$. However, given previous work on nonparametric regression with generated regressors, this is innocuous under appropriate smoothness assumption on $\vartheta$ which we are invoking anyway.

The first impression from looking at

$$
\begin{equation*}
\beta=\mathbb{E}\left[\left[D_{z} m_{X \mid Z}(Z)^{\prime}\right]^{-} \nabla_{z} \mathbb{E}[\bar{Y} \mid Z] B(Z)\right] \tag{4.2}
\end{equation*}
$$

is that due to the non-smoothness in $\bar{Y}$ no fast enough first step estimator can be devised for an average derivative type estimator to become root $n$ estimable. However, this is not the case. To see how the estimator is constructed, and understand why it is $\sqrt{n}$ consistent, note first that since $Y$ is binary,

$$
\bar{Y}=k_{Y \mid Z V}^{0.5}(Z, V)=\mathbb{I}\{\mathbb{P}[Y=0 \mid Z, V]<0.5\},
$$

and consequently, $\mathbb{E}[\bar{Y} \mid Z]=\mathbb{P}[\mathbb{P}[Y=0 \mid Z, V]<0.5 \mid Z]$ This suggests estimating $\nabla_{z} \mathbb{E}[\bar{Y} \mid Z=z]$ via

$$
\sum_{j} \nabla_{z} W_{j}(z) \mathbb{I}\left\{\widehat{P}_{j}<0.5\right\}
$$

where $W_{j}(z)$ are appropriate Kernel weights, e.g., $\left[\sum_{j} \mathcal{K}_{h j}(z)\right]^{-1} \mathcal{K}_{h j}(z), \mathcal{K}_{h j}(z)=h^{-L} \mathcal{K}\left(\left(Z_{j}-z\right) / h\right)$ and $\mathcal{K}\left(\left(Z_{j}-z\right) / h\right)=\Pi_{l=1, ., L} K\left(\left(Z_{j}^{l}-z^{l}\right) / h\right)$ is a standard $L$-variate product kernel with
standard univariate kernel function $K$. Moreover, $\widehat{P}_{j}$ denotes an estimator of $P_{j}=p\left(Z_{j}, V_{j}\right)=$ $\mathbb{P}\left[Y_{j}=0 \mid Z_{j}, V_{j}\right]$. The problem with this estimator is that the pre-estimator $\widehat{P}_{j}$ appears within the nondifferentiable indicator, resulting in a potentially very difficult pre-estimation analysis. To improve upon the tractability of the problem, we replace the indicator $\mathbb{I}$ by a smooth indicator, i.e., $\mathbb{K}(\xi)=\int_{-\infty}^{\xi} K(t) d t$. Then, a straightforward sample counterpart estimator to $\beta=\mathbb{E}\left[\left[D_{z} m_{X \mid Z}(Z)^{\prime}\right]^{-} \nabla_{z} \mathbb{E}[\bar{Y} \mid Z] B(Z)\right]$, looks as follows:

$$
\begin{equation*}
\hat{\beta}_{H}=n^{-1} \sum_{i}\left[D_{z} \hat{m}_{X \mid Z}\left(Z_{i}\right)^{\prime}\right]^{-} \sum_{j \neq i} \nabla_{z} W_{j}\left(Z_{i}\right) \mathbb{K}\left\{\left(\widehat{P}_{j}-0.5\right) / h\right\} B\left(Z_{j}\right) \tag{4.3}
\end{equation*}
$$

where the subscript $H$ indicates "heterogeneity". As is shown formally in theorems 3 and 4 below, the main result of this section is that under appropriate assumptions

$$
\sqrt{n}\left(\hat{\beta}_{H}-\beta\right) \xrightarrow{\mathcal{D}} \mathcal{N}\left(0, \Sigma_{H}\right),
$$

where $\Sigma_{H}$ is defined as $\Sigma_{H}=\mathbb{E}\left(\sum_{k=1}^{3} \sigma_{k} \sigma_{k}^{\prime}\right)+2 \mathbb{E}\left(\sigma_{2} \sigma_{3}^{\prime}\right)-\beta \beta^{\prime}$, and

$$
\begin{align*}
\sigma_{1} & =\left[D_{z} m_{X \mid Z}\left(Z_{i}\right)^{\prime}\right]^{-} \nabla_{z} m_{\bar{Y} \mid Z}\left(Z_{i}\right) B\left(Z_{i}\right), \\
\sigma_{2} & =\left[D_{z} m_{X \mid Z}\left(Z_{i}\right)^{\prime}\right]^{-} f_{Z}\left(Z_{i}\right)^{-1} \nabla_{z} f_{Z}\left(Z_{i}\right) V_{i}^{\prime}\left[D_{z} m_{X \mid Z}\left(Z_{i}\right)^{\prime}\right]^{-} \nabla_{z} m_{\bar{Y} \mid Z}\left(Z_{i}\right) B\left(Z_{i}\right),  \tag{4.4}\\
\sigma_{3} & =\left[D_{z} m_{X \mid Z}\left(Z_{i}\right)^{\prime}\right]^{-} f_{Z}\left(Z_{i}\right)^{-1} \nabla_{z} f_{Z}\left(Z_{i}\right)\left(\bar{Y}_{i}-m_{\bar{Y} \mid Z}\left(Z_{i}\right)\right) B\left(Z_{i}\right) .
\end{align*}
$$

To understand the large sample behavior of this estimator, rewrite $\hat{\beta}_{H}$ as $\hat{\beta}_{H}=T_{1 n}+T_{2 n}+T_{3 n}$, where

$$
\begin{align*}
& T_{1 n}=n^{-1} \sum_{i}\left[D_{z} \widehat{m}_{X \mid Z}\left(Z_{i}\right)^{\prime}\right]^{-} \sum_{j \neq i} \nabla_{z} W_{j}\left(Z_{i}\right) \bar{Y}_{j} B\left(Z_{j}\right), \\
& T_{2 n}=n^{-1} \sum_{i}\left[D_{z} \widehat{m}_{X \mid Z}\left(Z_{i}\right)^{\prime}\right]^{-} \sum_{j \neq i} \nabla_{z} W_{j}\left(Z_{i}\right)\left[\mathbb{K}\left\{\left(P_{j}-0.5\right) / h\right\}-\mathbb{I}\left\{P_{j}<0.5\right\}\right] B\left(Z_{j}\right),  \tag{4.5}\\
& T_{3 n}=n^{-1} \sum_{i}\left[D_{z} \widehat{m}_{X \mid Z}\left(Z_{i}\right)^{\prime}\right]^{-} \sum_{j \neq i} \nabla_{z} W_{j}\left(Z_{i}\right)\left[\mathbb{K}\left\{\left(\widehat{P}_{j}-0.5\right) / h\right\}-\mathbb{K}\left\{\left(P_{j}-0.5\right) / h\right\}\right] B\left(Z_{j}\right) . .
\end{align*}
$$

In this decomposition, $T_{1 n}$ is the leading term. It will dominate the asymptotic distribution. Its large sample behavior can be established using theorem 3, which also covers the sample counterparts estimator in the full independence case, which is defined as

$$
\begin{equation*}
\hat{\beta}_{1}=\frac{1}{n} \sum_{i}\left[D_{z} \widehat{m}_{X \mid Z}\left(Z_{i}\right)^{\prime}\right]^{-} \nabla_{z} \widehat{m}_{Y \mid Z}\left(Z_{i}\right) B\left(Z_{i}\right) . \tag{4.6}
\end{equation*}
$$

Hence, we will first discuss the independence case. Then, we will give assumptions under which $T_{2 n}$ and $T_{3 n}$ will tend to zero faster than the leading term. Essentially, these conditions are higher order smoothness conditions on the conditional cdf $F_{P \mid Z}$ and on $f_{Z}$, as well as the corresponding restrictions on the kernel (i.e., to be of higher order), so that fast enough rates of convergence are obtained.

### 4.3 The Large Sample Behavior of $\hat{\beta}_{1}$

When discussing the estimation of $\beta$ using any regression it is important to clarify the properties of details of the estimator (4.6). This concerns in particular the kernel and bandwidth. As mentioned above we use a product kernel in all regressions. Therefore we formulate our assumptions for the one-dimensional kernel functions $K$. To simplify things further, instead of a bandwidth vector $\mathbf{h} \in \mathbb{R}^{L}$ we assume that we have only one single bandwidth for each regression, denoted $h$. We shall make use of the following notation: Define kernel constants

$$
\mu_{k}=\int u^{k} K(u) d u \quad \text { and } \quad \kappa_{k}^{2}=\int u^{k} K(u)^{2} d u
$$

In principle, we also have two bandwidths to consider, one in estimating $m_{X \mid Z}$, and one in estimating $m_{Y \mid Z}$. However, since the estimation problems are symmetric (i.e., in particular both mean regressions share the same regressors and have thus the same dimensionality), we assume for ease of exposition the same kernel and the same bandwidth, denoted by $K$ and $h$, in both regressions. Our assumptions regarding kernel and bandwidth are standard (cf. PSS):

Assumption 9. Let $r=(L+4) / 2$ if $L$ is even and $r=(L+3) / 2$ if $L$ is odd. All partial derivatives of $\mathbb{E}[X \mid Z=z], \mathbb{E}[Y \mid Z=z]$ and $f_{Z}(z)$ of order $r+1$ exist for all $z \in \mathcal{B}$. Moreover, the expectations of $\left[D_{z} m_{X \mid Z}(Z)^{\prime}\right]^{-} B Y_{l}(Z)$ and $\left[D_{z} m_{X \mid Z}(Z)^{\prime}\right]^{-} B X_{l}(Z)\left[D_{z} m_{X \mid Z}(Z)^{\prime}\right]^{-} \nabla_{z} m_{Y \mid Z}(Z)$ exist for all $l=1, \ldots, r$, , where $B Y_{l}$ (resp., $B X_{l}$ ) contains sums of products of all partial derivatives of $m_{Y \mid Z}$ and $f_{Z}$ (resp. $m_{X \mid Z}$ and $f_{Z}$ ) such that the combined order of derivatives of the product is at most $l+1$.

Assumption 10. The one-dimensional kernel is Lipschitz continuous, bounded, has compact support, is symmetric around 0 and of order r (i.e. $\mu_{k}=\int u^{k} K(u) d u=0$ for all $k<r$ and $\left.\int u^{r} K(u) d u<\infty\right)$.

Assumption 11. As $n \rightarrow \infty, h \rightarrow 0, n h^{L+2} \rightarrow \infty$ and $n h^{2 r} \rightarrow 0$.
The following theorem summarizes the results when is $\vartheta=m_{X \mid Z}$. In particular, it establishes asymptotic normality of the appropriate sample counterpart estimators

Theorem 3. Let the true model be as defined in 1.1 and 1.2. Suppose assumptions 1 4, 6-7, and 9- 11 hold. Assume further that $V \perp Z$. For scale normalization, assume $\mathbb{E}\left[f_{U \mid V}\left(X^{\prime} \beta ; V\right) B(Z)\right]=1$. Then,

$$
\sqrt{n}\left(\hat{\beta}_{1}-\beta\right) \xrightarrow{\mathcal{D}} \mathcal{N}\left(0, \Sigma_{1}\right)
$$

where

$$
\Sigma_{1}=\mathbb{E}\left(\sum_{k=1}^{3} \sigma_{k} \sigma_{k}^{\prime}\right)+2 \mathbb{E}\left(\sigma_{2} \sigma_{3}^{\prime}\right)-\beta \beta^{\prime}
$$

and

$$
\begin{aligned}
& \sigma_{1}=\left[D_{z} m_{X \mid Z}\left(Z_{i}\right)^{\prime}\right]^{-} \nabla_{z} m_{Y \mid Z}\left(Z_{i}\right) B\left(Z_{i}\right) \\
& \sigma_{2}=\left[D_{z} m_{X \mid Z}\left(Z_{i}\right)^{\prime}\right]^{-} f_{Z}\left(Z_{i}\right)^{-1} \nabla_{z} f_{Z}\left(Z_{i}\right) V_{i}^{\prime}\left[D_{z} m_{X \mid Z}\left(Z_{i}\right)^{\prime}\right]^{-} \nabla_{z} m_{Y \mid Z}\left(Z_{i}\right) B\left(Z_{i}\right) \\
& \sigma_{3}=\left[D_{z} m_{X \mid Z}\left(Z_{i}\right)^{\prime}\right]^{-} f_{Z}\left(Z_{i}\right)^{-1} \nabla_{z} f_{Z}\left(Z_{i}\right)\left(Y_{i}-m_{Y \mid Z}\left(Z_{i}\right)\right) B\left(Z_{i}\right)
\end{aligned}
$$

Remark 4.1 - Discussion of Theorem 3: The first results shows a number of parallels to PSS in the case of exogenous ADEs. Similar to PSS, we obtain root- $n$ consistency of our estimator for $\beta$, and we may be able to eliminate the bias under similar assumptions on the rate of convergence as detailed in assumptions 9 and 11. The variance in term is a more complicated expression, but shares similar features, in particular in the first two terms, with the PSS result. This will be our baseline result. discuss the conditions under which we may include first stage projections of $Y$, like the median regression that is required to deal with heteroscedasticity.
Remark 4.2 - Estimating $\Sigma_{1}$ : Estimation of the variance components is straightforward by sample counterparts. For instance, an estimator for $\Sigma_{1}^{23}=\mathbb{E}\left(\sigma_{2} \sigma_{3}^{\prime}\right)$ is given by

$$
\begin{aligned}
\widehat{\Sigma}_{1}^{23}= & n^{-1} \sum \hat{f}_{Z}\left(Z_{i}\right)^{-2}\left[D_{z} \hat{m}_{X \mid Z}\left(Z_{i}\right)^{\prime}\right]^{-} \nabla_{z} \hat{f}_{Z}\left(Z_{i}\right)\left(Y_{i}-\hat{m}_{Y \mid Z}\left(Z_{i}\right)\right) \\
& \times\left\{\left[D_{z} \hat{m}_{X \mid Z}\left(Z_{i}\right)^{\prime}\right]^{-} \nabla_{z} \hat{f}_{Z}\left(Z_{i}\right)\left(X_{i}-\hat{m}_{X \mid Z}\left(Z_{i}\right)\right)^{\prime}\left[D_{z} \hat{m}_{X \mid Z}\left(Z_{i}\right)^{\prime}\right]^{-} \nabla_{z} \hat{m}_{Y \mid Z}\left(Z_{i}\right)\right\}^{\prime} B\left(Z_{i}\right) .
\end{aligned}
$$

Consistency of this estimator can essentially be shown by appealing to a law of large numbers, but this analysis is beyond the scope of this paper.

### 4.4 The Large Sample Behavior of $\hat{\beta}_{H}$

We now extend theorem 3 to the heteroscedastic case. To treat heteroscedasticity, we have introduced the two projection estimator

$$
\hat{\beta}_{H}=n^{-1} \sum_{i}\left[D_{z} \widehat{m}_{X \mid Z}\left(Z_{i}\right)^{\prime}\right]^{-} \sum_{j \neq i} \nabla_{z} W_{j}\left(Z_{i}\right) \mathbb{K}\left\{\left(\widehat{P}_{j}-0.5\right) / h\right\} B\left(Z_{j}\right)
$$

Recall the decomposition $\hat{\beta}_{H}=T_{1 n}+T_{2 n}+T_{3 n}$ in (4.5). The first term $T_{1 n}$ can be handled along exactly the same lines as the estimator in theorem 3, using some minor modifications in assumptions. It remains to be shown that the terms $T_{2 n}$ and $T_{3 n}$ tend to zero faster. To this end, we have to be precise about details of the estimator $\hat{\beta}_{H}$. First, there are several bandwidths: There is a bandwidth associated with the $\mathbb{K}\{\cdot\}$ function, as well as smoothness parameters when estimating $P_{j}=p\left(Z_{j}, V_{j}\right)$. To distinguish between the different kernels and bandwidths, we call the derivative of $\mathbb{K}\{\cdot\} K_{1}$, a kernel with bandwidth $h_{1}$ and order $r_{1}$, and the univariate elements of a product kernel employed in the estimation of $p$ as $K_{2}$, with bandwidth $h_{2}$ and order $r_{2}$.

Assumption 12. $K_{1}$ and $K_{2}$ are continuous, bounded, compactly supported, and symmetric functions of order $r_{1}, r_{2}$ (i.e. $\int u^{k} K(u) d u=0$ for all $k<r$ and $\left.\int u^{r} K(u) d u<\infty\right)$.

Assumption 13. Let $r=(L+4) / 2$ if $L$ is even and $r=(L+3) / 2$ if $L$ is odd. All partial derivatives of $F_{P \mid Z}$ and $f_{Z}(z)$ of order $r+1$ exist for all $z \in \mathcal{B}$. Moreover, the expectations of $\left[D_{z} m_{X \mid Z}(Z)^{\prime}\right]^{-} B F_{l}(Z)$ and $\left[D_{z} m_{X \mid Z}(Z)^{\prime}\right]^{-} B F_{l}(Z)\left[D_{z} m_{X \mid Z}(Z)^{\prime}\right]^{-} \nabla_{z} m_{Y \mid Z}(Z)$ exist for all $l=1, \ldots, r$, where $B F_{l}$ contains sums of products of all partial derivatives of $F_{P \mid Z}$ and $f_{Z}$ such that the combined order of derivatives of the product is at most $l+1$.

Assumption 14. $f_{Z V}$ is bounded and has bounded first partial derivatives with respect to all components of $z$, for all $z \in \mathcal{B}$.

Assumption 15. As $n \rightarrow \infty, h_{1}, h_{2} \rightarrow 0, n h_{1}, n h_{2}^{L+\operatorname{dim}(V)+2} \rightarrow \infty$ and $n h_{1}^{2 r_{1}}, n h_{2}^{L+\operatorname{dim}(V)+4} \rightarrow 0$.
Note that we require higher order smoothness conditions on $F_{P \mid Z}$ and $f_{Z}$ that in connection with higher order kernels ensure that the bias terms $\sqrt{n} T_{2 n}$ and $\sqrt{n} T_{3 n}$ are $o_{p}(1)$.

Theorem 4. Let the true model be as defined in 1.1 and 1.2, and suppose that assumptions 1-3, 5-6, 8.1-8.2a, 9-15 are true. Assume further $\mathbb{E}\left[f_{\bar{V} \mid Z}\left(m_{X \mid Z}(Z)^{\prime} \beta ; Z\right) B(Z)\right]=1$ holds. Then, $\sqrt{n} T_{2 n}=o_{p}(1), \sqrt{n} T_{3 n}=o_{p}(1)$, and $\sqrt{n}\left(\hat{\beta}_{H}-\beta\right) \xrightarrow{\mathcal{D}} \mathcal{N}\left(0, \Sigma_{H}\right)$, where $\Sigma_{H}$ is defined in equation (4.4).

This theorem characterizes the large sample behavior of our estimator. Under the smoothness and higher order bias reduction assumptions, it essentially behaves like the independence case estimator $\hat{\beta}_{1}$, with $Y$ replaced by $\bar{Y}$, i.e., with known conditional median.

## 5 Specification Testing

### 5.1 Overidentification: Issue and Test

The first question that we can analyze within our framework is how to treat overidentification if we have more instruments than regressors. In the linear model, overidentification allows to delete instruments and recover $\beta$ by various different estimators that always only use a subset of instruments. In the ( $X, V$ ) projection of the Blundell and Powell (2003) approach, as already noted by the authors a similar feature is missing. In our setup it may be introduced, and the linear model result may be better understood. We discuss in the following the full independence case, but all arguments may be trivially extended to the heteroscedastic case random coefficients case.

If we return to the theorem 2 and the associated assumptions, we see that $\beta$ would be identified by taking the derivatives w.r.t. any subset of instruments $Z_{1}$ such that $Z=\left(Z_{1}^{\prime}, Z_{-1}^{\prime}\right)^{\prime}$
and $\left[D_{z_{1}} m_{X \mid Z}(z) D_{z_{1}} m_{X \mid Z}(z)^{\prime}\right]$ would be nonsingular for all $z \in \mathcal{B}$. by similar arguments as in theorem 2 , the following result holds:

$$
\begin{equation*}
\beta=\mathbb{E}\left[\left[D_{z_{1}} m_{X \mid Z}(Z)^{\prime}\right]^{-} \nabla_{z_{1}} m_{Y \mid Z}(Z) B(Z)\right] . \tag{5.1}
\end{equation*}
$$

Consequently, the question of overidentification is not about exclusion of instruments in the regression. Instead the question of overidentification is about exclusion of derivatives of instruments, while the instruments should always be included in the regressions. Indeed, one can show that otherwise a nonvanishing bias term of the form $\mathbb{E}\left[\mathbb{E}[Y \mid Z] Q_{z_{1}} \mid Z_{1}, V\right]$, where $Q_{z_{1}}=\nabla_{z_{1}} \log f_{Z_{-1} \mid Z_{1} V}\left(Z_{-} ; Z_{1}, V\right)$, is obtained. Excluding instruments is only possible if they can be excluded from both equation (using, say, a standard omission-of-variables test).

An overidentification test is straightforwardly constructed as in Hausman (1978): Suppose $M$ such partition of $Z=\left(Z_{1}^{\prime}, Z_{-1}^{\prime}\right)^{\prime}$ exist s.th. $\beta$ is identified, which may be obtained by successively deleting one or more derivatives in constructing the estimator, then we simply compare their distance using some metric. The test would consider $H_{0}: \beta^{(1)}=\beta^{(2)}=\ldots=\beta^{(M)}$. To this end, we determine the joint distribution of $\mathcal{B}=\left(\beta^{(1) \prime}, \beta^{(2) \prime}, \ldots, \beta^{(M)^{\prime}}\right)^{\prime}$. As a corollary from the large sample theory of this paper, $R^{\prime} \mathcal{B}=0, \widehat{\mathcal{B}} \xrightarrow{d} \mathcal{N}\left(0, \Sigma_{I}\right)$, where $\Sigma_{I}$ is a covariance matrix with typical element $\Sigma_{j k}$. This element is given by

$$
\Sigma_{j k}=\mathbb{E}\left(\sum_{l=1}^{3} \sigma_{k}^{j} \sigma_{k}^{k \prime}\right)+2 \mathbb{E}\left(\sigma_{2}^{j} \sigma_{3}^{k \prime}\right)-\beta \beta^{\prime}
$$

where for $h=j, k$.

$$
\begin{aligned}
\sigma_{1}^{h} & =\left[D_{z_{h}} m_{X \mid Z}\left(Z_{i}\right)^{\prime}\right]^{-} \nabla_{z_{h}} m_{Y \mid Z}\left(Z_{i}\right) B\left(Z_{i}\right) \\
\sigma_{2}^{h} & =\left[D_{z_{h}} m_{X \mid Z}\left(Z_{i}\right)^{\prime}\right]^{-} f_{Z}\left(Z_{i}\right)^{-1} \nabla_{z_{h}} f_{Z}\left(Z_{i}\right) V_{i}^{\prime}\left[D_{z_{h}} m_{X \mid Z}\left(Z_{i}\right)^{\prime}\right]^{-} \nabla_{z_{h}} m_{Y \mid Z}\left(Z_{i}\right) B\left(Z_{i}\right) \\
\sigma_{3}^{h} & =\left[D_{z_{h}} m_{X \mid Z}\left(Z_{i}\right)^{\prime}\right]^{-} f_{Z}\left(Z_{i}\right)^{-1} \nabla_{z_{h}} f_{Z}\left(Z_{i}\right)\left(Y_{i}-m_{Y \mid Z}\left(Z_{i}\right)\right) B\left(Z_{i}\right)
\end{aligned}
$$

Then,

$$
\widehat{\Gamma}_{O v I d}=\left(R^{\prime} \widehat{\mathcal{B}}\right)^{\prime}\left[R \widehat{\Sigma}_{I} R^{\prime}\right]^{-1}\left(R^{\prime} \widehat{\mathcal{B}}\right) \xrightarrow{\mathcal{D}} \chi_{M-1}^{2}
$$

by standard arguments.

### 5.2 Testing for Heterogeneity under the Assumption of Endogeneity

The principle of comparing different coefficients as means for testing a hypothesis under our specification can be maintained more generally. If we assume to be in the scenario with endogenous regressors, we can test whether we have a heteroscedastic error or not. To illustrate the main idea, suppose that $V \perp Z$, and hence, in the case of heteroscedasticity we know that a sample counterpart to

$$
\begin{equation*}
\beta=\mathbb{E}\left[\left[D_{z} m_{X \mid Z}(Z)^{\prime}\right]^{-} \nabla_{z} m_{\bar{Y} \mid Z}(Z) B(Z)\right], \tag{5.2}
\end{equation*}
$$

where $\bar{Y}=k_{Y \mid Z, V}^{0.5}(Z, V)$ produces a root $n$ consistent, asymptotically normal estimator regardless of heteroscedasticity of $U$, while a sample counterpart estimator based on

$$
\beta=\mathbb{E}\left[\left[D_{z} m_{X \mid Z}(Z)^{\prime}\right]^{-} \nabla_{z} m_{Y \mid Z}(Z) B(Z)\right]
$$

will be inconsistent under heteroscedasticity. However, under $H_{0}$ of homoscedasticity, we have again that both estimators should vary only by sampling error. Using the notation $\mathbb{E}[\bar{Y}-Y \mid Z]=m_{\bar{Y}-Y \mid Z}(Z)$, a straightforward test statistic is therefor suggested by the following reformulation of $H_{0}$ :

$$
0=\mathbb{E}\left[\left[D_{z} m_{X \mid Z}(Z)^{\prime}\right]^{-} \nabla_{z} m_{\bar{Y}-Y \mid Z} B(Z)\right]=\delta
$$

The theory of the obvious sample counterpart $\hat{\delta}=n^{-1} \sum\left[D_{z} \widehat{m}_{X \mid Z}\left(Z_{i}\right)^{\prime}\right]^{-} \nabla_{z} \widehat{m}_{\bar{Y}-Y \mid Z}\left(Z_{i}\right) B\left(Z_{i}\right)$ is a corollary to theorem 4 . Specifically, $\sqrt{n} \hat{\delta} \xrightarrow{\mathcal{D}} \mathcal{N}\left(0, \Sigma_{\delta}\right)$, where $\Sigma_{\delta}$ is defined as in equation (4.4), safe for the fact that $\bar{Y}$ is replaced by $\bar{Y}-Y$. A test statistic for heteroscedasticity is then simply a Wald test of whether $\delta$ is greater than zero, i.e.

$$
\widehat{\Gamma}_{h e t}=\hat{\delta}^{\prime} \hat{\Sigma}_{\delta}^{-1} \hat{\delta} \xrightarrow{\mathcal{D}} \chi_{K}^{2},
$$

where $\hat{\Sigma}_{\delta}$ is an estimator for $\Sigma_{\delta}$, and this test statistic may be used to assess whether our model is truly heteroscedastic.

### 5.3 Testing for Endogeneity

Finally, consider analyzing whether regressors are endogenous. There are a variety of options. As in Hoderlein $(2005,2008$ a) and Hoderlein and Mammen (2008), we may compare the regression $\mathbb{E}[Y \mid X]$ with the regression $\mathbb{E}[Y \mid X, V]$. Under the null of exogeneity, the two functions should be the same, and hence we use a standard nonparametric omission of variables test, with the only added difficulty that $V$ is now a generated regressor. This test would be consistent regardless of whether the single index specification on the regressors is correct or not, and would deliver nonparametric test statistics that have local power against Pitman alternatives converging at a certain rate. This procedure can be seen as a nonparametric generalization of Hausman's (1978) second test for the inclusion of control functions as test of exogeneity in a linear model.

However, if we believe the index specification to be correct, than there are other, and in some instances better, options. Note that, under the null of exogeneity, a sample counterpart estimators to the average derivative identification principle $\beta=\mathbb{E}\left[\nabla_{x} \mathbb{E}[Y \mid X] C(X)\right](C$ is a again a bounded weighting function), and an estimator based on our identification principle (say, $\beta=\mathbb{E}\left[\left[D_{z} m_{X \mid Z}(Z)^{\prime}\right]^{-} \nabla_{z} \mathbb{E}[Y \mid Z] B(Z)\right]$ ), should yield estimators that vary only
be sample randomness, while under the alternative they should differ significantly. Hence, a similar test as the original test in Hausman (1978) may be performed. Let $\hat{\beta}_{E x}$ denote a sample counterpart estimator to $\mathbb{E}\left[\nabla_{x} \mathbb{E}[Y \mid X] C(X)\right]$ like the PSS ADE, $\hat{\beta}_{E n d}$ any of the sample counterpart estimator to $\mathbb{E}\left[\left[D_{z} m_{X \mid Z}(Z)^{\prime}\right]^{-} \nabla_{z} \mathbb{E}[Y \mid Z] B(Z)\right]$ defined below, $\widehat{\mathcal{B}}=\left(\hat{\beta}_{E x}^{\prime}, \hat{\beta}_{E n d}^{\prime}\right)$, and $G=(I,-I)$. Next, rewrite $H_{0}: G^{\prime} \mathcal{B}=0$, and use the fact that $\widehat{\mathcal{B}} \xrightarrow{d} \mathcal{N}\left(0, \Sigma_{E}\right)$, where $\Sigma_{E}$ is a variance covariance matrix that is straightforwardly derived from the theory below, in particular theorem 3 (the subscript $E$ is meant to denote endogeneity). Then, a Hausman-type test statistic for $H_{0}, \widehat{\Gamma}_{1}=\left(G^{\prime} \widehat{\mathcal{B}}\right)^{\prime}\left[G \widehat{\Sigma}_{E} G^{\prime}\right]^{-1}\left(G^{\prime} \widehat{\mathcal{B}}\right)$ behaves asymptotically as follows:

$$
\begin{equation*}
\widehat{\Gamma}_{1}=\left(G^{\prime} \widehat{\mathcal{B}}\right)^{\prime}\left[G \widehat{\Sigma}_{E} G^{\prime}\right]^{-1}\left(G^{\prime} \widehat{\mathcal{B}}\right) \xrightarrow{d} \chi_{K}^{2} . \tag{5.3}
\end{equation*}
$$

What would be the advantage of such a specification test? First, it has more power against certain alternatives. Indeed, because of the parametric rate of all estimators, we may detect local alternatives in the parameter vector that converge to $H_{0}$ at root $n$. Therefore this test will be superior, provided the misspecification due to endogeneity affects the index.

## 6 Simulation

The finite sample performance of the estimators we propose is best analyzed by a Monte Carlo simulation study. In this section, we are chiefly concerned with analyzing the behavior of $\hat{\beta}_{H}$. The main scenario we consider involves an asymmetric error distribution, such that conditional mean and median differ. Moreover, we assume that $V$ in the IV equation is fully independent of $Z$, in which case there is no correction term, and the estimator takes the convenient ratio-of-coefficients form as in (4.3).

To obtain an idea of the behavior of our estimator, we analyze the performance of our estimator at different data sizes. We find that our estimator performs well for moderate data sizes, and as theory predicts, we find that the mean square error reduces as the sample size increases, but we observe a small bias even in relatively large samples. However, we establish that our estimator is superior to parametric and semiparametric estimators that do not account for heterogeneity. As examples for estimators that do not account for heterogeneity we consider the parametric estimator of River and Voung (1988) $\hat{\beta}_{R V}$, and the full independence estimator $\hat{\beta}_{1}$. Moreover, we show that even an infeasible oracle estimator that uses some prior knowledge not available to the econometrician shows slow convergence behavior in this setup.

We consider the case of one endogenous regressor, w.l.o.g. $X_{1 i}$, and denote the set of regressors by $X_{i}=\left(X_{1 i}, X_{2 i}, . ., X_{5 i}\right)^{\prime}$, and the set of all instruments $Z_{i}=\left(Z_{1 i}, X_{2 i}, \ldots, X_{5 i}\right)^{\prime}$. For
the purpose of concreteness, we specify the DGP as the following 5-dimensional regression:

$$
\begin{aligned}
Y_{i} & =\mathbb{I}\left\{\beta_{1} X_{1 i}+\beta_{2} X_{2 i}+\beta_{3} X_{3 i}+\beta_{4} X_{4 i}+\beta_{5} X_{5 i}+U_{i}>0\right\}, \\
X_{1 i} & =Z_{1 i}+V_{i}, \quad i=1, . ., n,
\end{aligned}
$$

where $\beta=(1,0.5,0.5,0.5,0.5)^{\prime}$ and the data $\left(Y_{i}, X_{i}, Z_{i}, U_{i}\right), i=1, . ., n$, are iid draws from the following distribution: For the error $U_{i}$, we assume that there is an omitted determinants called $W_{i}$ such that

$$
\begin{aligned}
\left(\log \left(W_{i}\right), Z_{i}\right)^{\prime} & \backsim \mathcal{N}(\mu, \Sigma), \\
\log \left(V_{i}\right)^{\prime} & \backsim \mathcal{N}(0,1),
\end{aligned}
$$

where

$$
\mu=0, \quad \Sigma=\left[\begin{array}{cccccc}
2 & 1.5 & 0 & 0 & 0 & 0 \\
1.5 & 2 & 1 & 1 & 1 & 1 \\
0 & 1 & 2 & 1 & 1 & 1 \\
0 & 1 & 1 & 2 & 1 & 1 \\
0 & 1 & 1 & 1 & 2 & 1 \\
0 & 1 & 1 & 1 & 1 & 2
\end{array}\right]
$$

and $V_{i}$ is independent of $\left(\log \left(W_{i}\right), Z_{i}\right)^{\prime}$. Observe that the $W_{i}$ are in particular correlated with $Z_{1 i}$. Next, the error $U_{i}$ is defined through:

$$
U_{i}=W_{i}-k_{W \mid Z}^{0.5}\left(Z_{i}\right)+V_{i},
$$

so that $k_{U \mid Z V}^{0.5}\left(Z_{i}, V_{i}\right)=V_{i}$. Hence, as we require the error $U_{i}$ obeys the conditional median exclusion restriction, but depends on $Z_{i}$. As baseline, our estimator (4.3) is defined as a local quartic polynomial estimator, with Epanechnikov kernels. Moreover, the "smooth indicator" is defined as the integral of the Epanechnikov kernel over the positive areas. The conditional probability is also estimated using a local quartic polynomial estimator, with Epanechnikov kernel. The independence estimator $\hat{\beta}_{1}$ is defined similarly, with the exception that no "smooth indicator" is required. The oracle estimator is obtained by using fitted values $Y_{1 i}$ instead of either the conditional median $k_{Y \mid Z V}^{0.5}\left(Z_{i}, V_{i}\right)$ (as is the case in the $\hat{\beta}_{H}$ ) or the $Y_{i}$ (as is the case in $\hat{\beta}_{1}$ ). The fitted values $Y_{1 i}$ are obtained in the following way: We assume that the oracle has knowledge of the $Y_{i}^{*}$, and compute the conditional median $k_{Y^{*} \mid Z V}^{0.5}\left(Z_{i}, V_{i}\right)$, and set $Y_{1 i}=\mathbb{I}\left\{k_{Y * \mid Z V}^{0.5}\left(Z_{i}, V_{i}\right)>0\right\}$. Bandwidths for all estimators are obtained by doing a grid search for the bandwidth that minimizes the MSE in 100 repetitions. Finally, $\hat{\beta}_{R V}$ is obtained by estimating a probit model with control function residuals as additional regressors.

The result of applying our methods can be found in figures $1-3$ in the graphs in the appendix. For each $j=1, . ., J, J=500$, a new sample ( $X_{i}, Z_{i}, W_{i}, U_{i}$ ) of size $n$ is drawn from
the distribution specified above. To illustrate the behavior of the estimator, in fig. 1 we plot the density of the estimated four elements of $\beta$ for $n=2500$ as solid line. Note that the first coefficient is normalized to one. The vertical line in all of these plots is at the true value of 0.5 . The closer this distribution is to a spike centered at this value, the better the performance of the estimator. We compare the distribution of $\hat{\beta}_{H}$ with that of the parametric estimator $\hat{\beta}_{R V}$ (first dotted line), the independence estimator $\hat{\beta}_{1}$ (second dotted line), and the oracle estimator $\hat{\beta}_{O}$ (solid line).

First, the most obvious feature of the result is the clear ordering in terms of the performance of the estimators, regardless of data size. Obviously, in terms of the bias, $\hat{\beta}_{H}$ is less biased than any of the two alternative feasible estimators $\hat{\beta}_{1}$ and $\hat{\beta}_{R V}$, and only the infeasible oracle estimator show less bias. In terms of variance, all three estimators nonparametric estimators are approximately similar. The fact that the variance is not significantly affected by the first step estimation of the conditional median arises because the median estimator still uses all observations. While the parametric estimator $\hat{\beta}_{R V}$ shows less variance than any of the semiparametric estimators, its much larger bias results in a greatly increased MSE compared to any semiparametric estimator, a result of the double misspecification of $\hat{\beta}_{R V}$ : First, the estimator erroneously imposes a parametric structure, second it erroneously assumes full conditional independence. Indeed, going from $\hat{\beta}_{R V}$ over $\hat{\beta}_{1}$ to $\hat{\beta}_{H}$ may be seen as first removing the parametricity assumption, and then correcting for heteroscedasticity. As we see, the first step reduces the bias at the expense of increasing the variance somewhat. Still, the MSE is almost halved. The second step reduces again the bias while only affecting the variance marginally, and the MSE is again reduced significantly.

It is interesting to see how the estimator behaves as $n$ varies. The heteroscedasticity robust estimator $\hat{\beta}_{H}$ significantly outperforms $\hat{\beta}_{1}$ and $\hat{\beta}_{R V}$ at moderate sample sizes $(n=2500)$; for smaller sample sizes the advantage in particular over the semiparametric competitor $\hat{\beta}_{1}$ becomes less pronounced. As such we find the familiar results in other simulation studies on the binary choice case (e.g., Frölich (2005)), namely that in binary choice models semiparametric methods require a significant amount of data to outperform misspecified models. Once, however, we have a significant amount of data, the advantages become obvious, see fig. 2 and 3 , who show the behavior with $n=7500$ and $n=15000$ observations. The bias of $\hat{\beta}_{H}$ starts to vanish, clearly visible in the bottom right panel of fig. 3. The same result is also obtained from the
tables, cf tab 1-4 below, which provide the specific numerical results. We obtain for $\hat{\beta}_{H}$ :

| Coefficient | 2 | 3 | 4 | 5 |
| :--- | :--- | :--- | :--- | :--- |
| $n=2500$ | 0.017288 | 0.015739 | 0.016421 | 0.016928 |
| $n=7500$ | 0.010301 | 0.009498 | 0.008755 | 0.008411 |
| $n=15000$ | 0.009207 | 0.008299 | 0.007690 | 0.007135 |

Table 1: MSE of $\hat{\beta}_{H}$ at Different Data Sizes
The reduction of the MSE with increasing sample size is obvious. Note also that due to the largely symmetric setup, all four coefficients are equally affected. A more detailed analysis shows a reduction in both bias and variance, as is also evident from the graphs, see fig. 13. Note, however, that the reduction in bias is quite slow. It is instructive to compare the estimator with the other estimators. For $\hat{\beta}_{1}$, we obtain the following result:

| Coefficient | 2 | 3 | 4 | 5 |
| :--- | :--- | :--- | :--- | :--- |
| $n=2500$ | 0.020972 | 0.020581 | 0.020385 | 0.021524 |
| $n=7500$ | 0.017027 | 0.016416 | 0.015744 | 0.015766 |
| $n=15000$ | 0.016466 | 0.016306 | 0.015324 | 0.014847 |

Table 2: MSE of $\hat{\beta}_{I}$ at Different Data Sizes
This result is clearly worse than the heteroscedasticity robust estimator $\hat{\beta}_{H}$, with an increase in MSE of roughly 25-100 \%. In contrast, as was to be expected, the (infeasible) oracle estimator $\hat{\beta}_{O}$ outperforms both estimators:

| Coefficient | 2 | 3 | 4 | 5 |
| :--- | :--- | :--- | :--- | :--- |
| $n=2500$ | 0.012412 | 0.009253 | 0.011056 | 0.009996 |
| $n=7500$ | 0.004054 | 0.004245 | 0.005386 | 0.005296 |
| $n=15000$ | 0.002875 | 0.002583 | 0.003155 | 0.002967 |

Table 3: MSE of $\hat{\beta}_{O}$ at Different Data Sizes
When decomposing the MSE, we find that the variance remains very comparable across all semiparametric estimators given the data size, while it is the bias that causes the differences. While the oracle estimator starts out relatively unbiased, and remains so, the independence estimator contains a nonvanishing bias component. The heteroscedastic estimator starts out with a bias that diminishes with increasing sample size. Note that the difference between $\hat{\beta}_{H}$ and $\hat{\beta}_{O}$ can be seen as a measure of the degree of information loss associated with the indicator function. Viewing the indicator as a filter, we conclude that the information loss is quite severe, and that significant data sizes are required to distinguish between different structure within the indicator.

Finally, consider the estimator from Rivers and Voung (1988). Due to the significant bias, this estimator performs worst, see the following table 4.

| Coefficient | 2 | 3 | 4 | 5 |
| :--- | :--- | :--- | :--- | :--- |
| $n=2500$ | 0.036005 | 0.037955 | 0.036398 | 0.036653 |
| $n=7500$ | 0.035706 | 0.034574 | 0.034498 | 0.035680 |
| $n=15000$ | 0.034166 | 0.034278 | 0.034125 | 0.034720 |

Table 4: MSE of $\hat{\beta}_{R V}$ at Different Data Sizes
Indeed, by closer inspection we find that the MSE is almost entirely due to the squared bias, and the variance contribution is quite small. Thus, the performance of the estimator shows only very little improvement with increasing sample size. Since all estimators correct for endogeneity, we can summarize the finding by saying that in this scenario correcting for heterogeneity and adopting a semiparametric procedure produces significantly better results. We conclude that the interaction between the various sources of misspecification makes at least in this setup semiparametric estimators quite attractive. And with a sufficient amount of data, it is also evident that allowing for a heterogeneous error structure improves the result significantly, and leads to a performance which is not much worse than that of an infeasible oracle estimator. Our application below will make the importance of being less restrictive in this part of the model also for real world data apparent.

## $7 \quad$ Application to Discrete Consumer Choice

### 7.1 Description of Data and Variables

As an example for an application of our method to a structural model in a heterogeneous population, we use data that is very similar to the one employed in Goolsbee and Petrin (2004) about the choice of television transmission mode, see Table A. 1 for an overview of all variables. The data comes from two data sources. First, from December 2000 until January 2001 NFO Worldwide fielded a household survey on television choices sponsored by Forrester Research as part of their Technographics 2001 program ${ }^{5}$. These households were randomly drawn from the NFO mail panel that is designed to be nationally representative.

The households that were surveyed basically have the choice between four different ways to receive television programming: local antenna, direct broadcast satellite (DBS), as well as basic and expanded cable, which we group into cable versus non-cable (satellite dish/local antenna).

[^5]Local antenna reception is free but only carries the local broadcast stations ${ }^{6}$. DBS systems are national companies that deliver many of the cable channels that usually priced uniformly across the whole country (in 2001 the two leading companies DirectTV and DISH Network (Echostar) charged $\$ 30$ and $\$ 32$ per month respectively). Hence, there is almost no price variation in the alternative. Compared to cable, DBS provides a greater variety of channels and more pay per view options but bares the potential for signal interference and charges a higher price. The fair amount of regional variation in cable prices permits us to estimate own price effects, while the cross price effects are constant, and hence neglectable.

Other than the choices people make, the survey also provides information on various socioeconomic household characteristics e.g. household income, household composition, education of the head of household and if applicable of the respective partner. Dropping observations with missing values in their choices or doubtful values in several household characteristics and removing outliers (recall that we also have to compactify our support) reduces the sample to approximately 15.900 observations. Table A. 2 in the appendix provides summary statistics for the sub sample including renter status and whether households live in single unit dwellings. Both characteristics are known to influence the ability to receive satellite.

We also make use of a second source of data, which provides us with information on cable prices and cable franchise characteristics each household faces (within a specific cable franchise area). The data come from Warren Publishing's 2002 Television and Cable Factbook, and provides detailed information on the cable industry, which is divided into geographically separated cable systems. From this data source, we use the channel capacity of the cable system, whether pay per view is available, the price of basic plus expanded basic service, the price for premium channels (here we use the price for HBO ) and the number of over-the-air channels (this corresponds to the number of local channels carried by the cable system).

As source of endogeneity, we follow Goolsbee and Petrin (2004), and assume that prices are correlated with unobserved characteristics like advertisement. To deal with endogeneity, we use variation in city franchise tax/fee to instrument cable prices (recall that the own price might be correlated with unobserved cable characteristics e.g. advertising or quality). Table A. 3 presents summary statistic for the respective variables. Technically, we can match both data sources using Warren's ICA system identification number, which is based on zip code information. Hence, we can assign a specific household to the adequate local cable company ${ }^{7}$ even though these individuals might not subscribe to cable.

[^6]
### 7.2 Empirical Results

The focus in our empirical analysis is on the own price effect, and how the result is altered by the introduction of our method. The effect of household covariates is not of interest, and we use these variables merely as controls. Since we are not interested in their effect, we employ principal components to reduce them to some three orthogonal approximately continuous variables, mainly because we require continuous covariates for nonparametric estimation. While this has some additional advantages, it is arguably ad hoc. However, we performed some robustness checks like alternating the components or adding parametric indices to the regressions, and the results do not change in an appreciable fashion (nor is the remaining variation statistically significant).

To show the performance of our estimator, it is instructive to start out with the standard practise of estimating a linear probability model and using 2SLS. We obtain the following result:

|  | Estimate | Std. Error | t value | p value |
| :--- | :---: | :--- | :--- | :--- |
| Intercept | 0.697908 | 0.008805 | 79.266 | 0 |
| Own Price | 0.228026 | 0.020040 | 11.379 | 0 |
| Income | 0.028096 | 0.002513 | 11.181 | 0 |
| PrinComp 1 | -0.025945 | 0.008904 | -2.914 | 0.003575 |
| PrinComp 2 | 0.014143 | 0.004033 | 3.507 | 0.000454 |
| PrinComp 3 | -0.018363 | 0.002663 | -6.895 | 0 |

Table 5: Linear Probability Model - 2SLS
There are two things noteworthy: First, quite in contrast to Economic theory, the model predicts that higher own price is associated with higher demand. Second, the income effect is positive, but small in absolute size. Due to the large sample size of $n=15.918$ all variables are highly significant, with $p$-values of virtually zero. This holds true even for the - in absolute size - small income effect. This finding remains stable across specifications, however, the own price effect becomes progressively more plausible as we move to less obviously misspecified specifications.

The following tables show the behavior of the full independence estimator $\hat{\beta}_{1}$. Specifically, it shows the point estimate, as well as the 2.5 and 97.5 quantile of the bootstrap distribution ${ }^{8}$ instead of the asymptotic distribution which is cumbersome to estimate. In this procedure, a coefficient is statistically not significant from zero if the confidence interval contains zero.

[^7]|  | Estimate | BS 0.025 value | BS 0.975 value |
| :--- | :---: | :--- | :---: |
| Own Price | -2.10788 | -5.94305 | 1.76096 |
| Income | 1 | 1 | 1 |
| PrinComp 1 | -3.31908 | -4.32975 | -2.49948 |
| PrinComp 2 | 1.70843 | 1.23556 | 2.35323 |
| PrinComp 3 | -0.35490 | -1.15326 | 0.48962 |

Table 6: Coefficients of $\hat{\beta}_{1}$ (Relative to Income) with Bootstrap Confidence Intervals

As we see from the results, this is the case for the own price effect, which is in absolute value only twice as strong as the income effect. Compared to the income effect, the estimate points in the opposite direction. And if we look at the non normalized results we also obtain that the income effect is positive (and actually of as small an order of magnitude as in the linear probability model), while the price effect is negative as it should be, but as mentioned insignificant. The first two principal components are significant, however not the third, and have generally the same sign and relative order of magnitude as in the linear probability model.

Finally, the heteroscedasticity robust estimator $\hat{\beta}_{H}$ produces the most sensible results:

|  | Estimate | BS 0.025 value | BS 0.975 value |
| :--- | ---: | :--- | :--- |
| Own Price | -8.02943 | -12.91400 | -2.90706 |
| Income | 1 | 1 | 1 |
| PrinComp 1 | -0.12521 | -1.04786 | 0.52044 |
| PrinComp 2 | 1.38809 | 0.93442 | 2.01614 |
| PrinComp 3 | -0.88721 | -2.02577 | -0.08665 |

Table 7: Coefficients of $\hat{\beta}_{H}$ (Relative to Income) with Bootstrap Confidence Intervals

Here we see that the own price effect is significantly negative. At first glance, the results appears to be slightly different from Goolsbee and Petrin (2004), who find a relatively low own price elasticity. However, as the income effect is rather weak (it is again of the same order of magnitude as in the linear probability model in the non normalized version. but recall that identification is only up to scale), this is not necessarily a contradiction. With respect to the application, we conclude that the likelihood that the average person in this population chooses cable reacts only modestly to an increase in income, which given the small fraction of total expenditures seems plausible (and is perhaps very different if one were to consider the demand for cars). However, given that price of cable is a significant variable in the marketing of this good, the average consumer seems to react more strongly to price incentives, and as theory
predicts, a price increase reduces the probability of buying cable.
With respect to the performance of various different estimators, we conclude that avoiding the misspecification associated with the linear probability model, as well as allowing for heterogeneous preferences (compared to the full independence estimator $\hat{\beta}_{1}$ ) substantially alters the result, and provides us with more plausible estimates for the (centrality) parameter of interest.

## 8 Summary and Outlook

The notion that we do not observe important determinants of individual behavior even in data sets with large cross section variation becomes more and more influential across microeconometrics. Indeed, it is widely believed now that unobserved tastes and preferences account for much more of the variation than observable characteristics. Hence, it is imperative to devise models that account for heterogeneity on individual level, in particular if the unobserved determinants and omitted variables are believed to be correlated with observables.

In these heterogeneous models, most often interest centers on average effects. In this paper, we analyze the binary choice model with random utility parameters under a median exclusion restriction that defines such a (local) average effect. It is moreover established that this effect coincides with the parameter $\beta$ in the reduced form binary choice model with heteroscedastic errors under a median exclusion restriction. We show how to nonparametrically identify this parameter $\beta$, and we propose a $\sqrt{n}$ consistent, asymptotically normal sample counterparts estimator. Moreover, based on our theory, we propose tests for overidentification, endogeneity as well as heterogeneity. Therefore we can provide means to check the specification, in addition to provide the first estimator for this parameter in this class of models.

In a Monte Carlo study we show that our estimator performs superior to an estimator which does not exploit the heterogeneity structure of the model. In an application, we show that our estimator uses significantly weaker assumptions than those employed in the literature, and through its use we may be able to reveal new and interesting features. How to extend this type of semiparametric approach from binary choice data to multinomial choice data and more complicated settings including simultaneity remains an interesting direction for future research. Our conjecture is that a similar estimation principle may be applicable to a large class of models.

## References

[1] Ai, Ch. 1997. A Semiparametric Maximum Likelihood Estimator, Econometrica 65, 933964.
[2] Bajari, P., Fox, J., Kim K. and S. Ryan, 2008. Discrete Choice Models with a Nonparametric Distribution of Random Coefficients, Working Paper.
[3] Blundell, R. and J. Powell. 2004. Endogeneity in semiparametric binary response models, Review of Economic Studies 71, 655-679.
[4] Blundell, R., and R. Smith. 1986. An Exogeneity Test for a Simultaneous Tobit Model, Econometrica 54, 679-685.'
[5] Chaudhuri, P., K. Doksum and A. Samarov. 1997. On average derivative quantile regression, Ann. Statist. 25, 715-744.
[6] Das, M., Newey, W. and F. Vella. 2003. Nonparametric estimation of sample selection models, Review of Economic Studies 70, 33-58.
[7] Delecroix, M., and M. Hristache. 1999. M-estimateurs semi-paramétriques dans les modèles à direction révélatrice unique, Bull. Belg. Maths. Soc. 6, 161-185.
[8] Fan, J. and I. Gijbels. 1996. Local Polynomial Modelling and Its Applications, Chapman and Hall, London.
[9] Florens, J.P., Heckman J, Meghir, C., and E. Vytlacil, 2008. Identification of Treatment Effects Using Control Functions in Models with Continuous, Endogenous Treatment and Heterogeneous Effects, Econometrica, forthcoming.
[10] Goolsbee, A. and A. Petrin, 2004. The Consumer Gains from Direct Broadcast Satellites and the Competition with Cable TV, Econometrica, 72, 351-381.
[11] Heckman, J. and E. Vytlacil, 2005. Structural Equations, Treatment Effects, and Econometric Policy Evaluation, Econometrica, 73, 669-738.
[12] Hoderlein, S., 2005. Nonparametric Demand Systems, Instrumental Variables and a Heterogeneous Population, Working Paper, Brown University.
[13] Hoderlein, S., 2009. How many Consumers are Rational?, Working Paper, Brown University.
[14] Hoderlein, S. and E. Mammen, 2007. Identification of marginal effects in nonseparable models without monotonicity, Econometrica 75, 1513-1518.
[15] Horowitz, J. L. 1992, A smoothed maximum score estimator for the binary response model, Econometrica 60, 505-531.
[16] Horowitz, J., and W., Härdle. 1996. Direct Semiparametric Estimation of Single Index Models with Discrete Covariates, JASA 91, 1632-1640.
[17] Horowitz, J., 1997, Bootstrap Methods in Econometrics: Theory and Numerical Performance, In Kreps, D.M. and Wallis, K.F. (eds), Advances in Economics and Econometrics: Theory and Applications, pp. 188-222, Cambridge University Press.
[18] Hristache, M., A. Juditsky, J. Polzehl, and V. Spokoiny. 2001. Structure Adaptive Approach for Dimension Reduction, Ann. Statist. 29, 1537-1566.
[19] Ichimura, H. 1993, Semiparametric Least Squares (SLS) and Weighted SLS Estimation of Single-Index Models, Journal of Econometrics 58, 71-120.
[20] Klein, R.W. and R.S. Spady. 1993, An Efficient Semiparametric Estimator of the Binary Response Model, Econometrica 61, 387-422.
[21] Lewbel, A. 1998. Semiparametric Latent Variable Model Estimation with Endogenous or Mismeasured Regressors, Econometrica, 66, 105-122.
[22] Manski, C. F. 1975, Maximum score estimation of the stochastic utility model of choice, Journal of Econometrics 3, 205-228.
[23] Matzkin, R. 1992. Nonparametric and Distribution-Free Estimation of the Binary Threshold Crossing and the Binary Choice Models, Econometrica, 62, 239-270.
[24] Matzkin, R. 2005. Heterogeneous Choice, for Advances in Economics and Econometrics, edited by Richard Blundell, Whitney Newey, and Torsten Persson, Cambridge University Press; presented at the Invited Symposium on Modeling Heterogeneity, World Congress of the Econometric Society, London, U.K.
[25] Newey, W.K., Powell, J.L. and Vella, F. 1999. Nonparametric Estimation of Triangular Simultaneous Equations Models, Econometrica 67, 565-603.
[26] Petrin, A. and K. Train. Control Function Corrections for Omitted Attributes in Differentiated Product Markets. 2006 Working paper.
[27] Powell, J. L., J. H. Stock, and T. M. Stoker. 1989. Semiparametric estimation of index coefficients, Econometrica 57, 1403-1430.
[28] Rivers, D., and Q Vuong. 1988. Limited Information Estimators and Exogeneity Tests for Simultaneous Probit Models, Journal of Econometrics, 39, 347-364.
[29] Stoker, T. M. 1986. Consistent Estimation of Scaled Coefficients, Econometrica 54, 14611481.
[30] Vytlacil, E. and N. Yildiz, 2007. "Dummy Endogenous Variables in Weakly Separable Models," Econometrica, 75, 757-779.

## 9 Appendix 1: Technical Proofs

## Proof of the Identification Theorems

## Proof of Theorem 1

To see the first statement, rewrite

$$
\begin{equation*}
Y=\mathbb{I}\left\{X^{\prime} \beta\left(A_{1}\right)+A_{2}^{\prime} \gamma\left(A_{1}\right)>0\right\}=\mathbb{I}\{X^{\prime} \beta+\underbrace{X^{\prime}\left(\beta\left(A_{1}\right)-\beta\right)+A_{2}^{\prime} \gamma\left(A_{1}\right)}_{U}>0\} . \tag{9.1}
\end{equation*}
$$

Then we obtain that

$$
k_{Y \mid X V}^{0.5}(X, V)=\mathbb{I}\left\{k_{Y^{*} \mid X V}^{0.5}(X, V)>0\right\}=\mathbb{I}\left\{X^{\prime} \beta+k_{U \mid X V}^{0.5}(X, V)>0\right\}
$$

Next, note that due to $Z \perp(A, V) \Longrightarrow X \perp A \mid V$, and thus $\mathbb{E}[\mathfrak{C} \mid X, V]=\mathbb{E}[\mathfrak{C} \mid V]=g(V)$ and $\mathbb{E}[\mathfrak{B} \mid X, V]=\mathbb{E}[\mathfrak{B} \mid V]$. Since $\left(X, V, A_{2}\right) \perp A_{1} \Longrightarrow V \perp \mathfrak{B}$, it moreover holds that $\mathbb{E}[\mathfrak{B} \mid V]=\mathbb{E}[\mathfrak{B}]$, and the condition that $(\mathfrak{B}, \mathfrak{C})$ are jointly symmetrically distributed about $(\beta, \mathbb{E}[\mathfrak{C} \mid V])$ conditional on $(V, X)$ holds also conditioning on $V$ only. Consequently, $\beta=$ $\mathbb{E}[\mathfrak{B} \mid V]=\mathbb{E}\left[\beta\left(A_{1}\right)\right]$, and $\beta$ is the mean of the distribution of random coefficients. Moreover,

$$
k_{U \mid X V}^{0.5}(X, V)=\mathbb{E}\left[X^{\prime}(\mathfrak{B}-\beta)+\mathfrak{C} \mid X, V\right]=X^{\prime} \mathbb{E}[\mathfrak{B}-\beta \mid V]+\mathbb{E}[\mathfrak{C} \mid V]=g(V),
$$

and a very similar argument holds to show that $k_{U \mid Z V}^{0.5}(Z, V)=g(V)$ as well.
To see equation (2.2), observe first that $k_{U \mid Z V}^{0.5}(Z, V)=g(V)$ implies that $k_{U \mid X V}^{0.5}(X, V)=$ $g(V)$. Start by using the definition of the $\alpha$-quantile to obtain

$$
\mathbb{P}\left(U \leq k_{U \mid Z, V}^{\alpha}(Z, V) \mid Z, V\right)=\alpha=\mathbb{P}\left(U \leq k_{U \mid X, V}^{\alpha}(X, V) \mid X, V\right),
$$

for any $\alpha \in(0,1)$. Taking conditional expectations with respect to $(X, V)$ on both sides produces

$$
\mathbb{E}\left[\mathbb{E}\left\{\mathbb{I}\left(U \leq k_{U \mid Z, V}^{\alpha}(Z, V)\right) \mid Z, V\right\} \mid X, V\right]=\mathbb{P}\left(U \leq k_{U \mid X, V}^{\alpha}(X, V) \mid X, V\right) .
$$

But due to $k_{U \mid Z, V}^{\alpha}(Z, V)=g(V)$, and the law of iterated expectations, we have that

$$
\mathbb{E}[\mathbb{I}(U \leq g(V)) \mid X, V]=\mathbb{P}\left(U \leq k_{U \mid X, V}^{\alpha}(X, V) \mid X, V\right),
$$

implying that $k_{U \mid X, V}^{\alpha}(X, V)=g(V)$, provided $U$ is continuously distributed.
Hence, if we assume the median exclusion restriction $k_{U \mid Z V}^{0.5}(Z, V)=g(V)$, we obtain that $\nabla_{x} k_{Y^{*} \mid X V}^{0.5}(X, V)=\beta$. Since $Y^{*}=X^{\prime} \beta\left(A_{1}\right)+A_{2}^{\prime} \gamma\left(A_{1}\right)=\phi(X, A)$, and $X \perp A \mid V$, we can apply Hoderlein and Mammen's (2007) theorem to obtain that the (constant) derivative has the following interpretation:

$$
\begin{equation*}
\beta=\mathbb{E}\left[\beta\left(A_{1}\right) \mid X=x, V=v, Y^{*}=k_{Y^{*} \mid X V}^{0.5}(x, v)\right] \tag{9.2}
\end{equation*}
$$

for all $(x, v) \in \operatorname{supp}(X) \times \operatorname{supp}(V)$.
Q.E.D.

## Proof of Theorem 2

$A d$ (i).The case when $\tilde{V}$ is independent of $Z$ is already discussed in the main text.
$A d$ (ii). Next, consider the case defined by assumptions 8.2 b : W.l.o.g, we consider two subsets of the support of $V$, denoted $S_{1}$ and $S_{2}$. Then, let $l \nearrow$ on $S_{1}=(-\infty, a)$, $\searrow$ on $S_{2}=(a, \infty)$ with inverses $l_{1}, l_{2}$. Let $\vartheta(z)^{\prime} \beta \leq \max _{v \in S_{1}} l(v)=l(a)$. Then,

$$
\begin{align*}
m_{\bar{Y} \mid Z}(z) & =\mathbb{P}\left[\vartheta(Z)^{\prime} \beta>l(V) \mid Z=z\right] \\
& =\mathbb{P}\left[\vartheta(Z)^{\prime} \beta>l(V), V \in S_{1} \mid Z=z\right]+\mathbb{P}\left[\vartheta(Z)^{\prime} \beta>l(V), V \in S_{2} \mid Z=z\right] \\
& =\mathbb{P}\left[l_{1}\left(\vartheta(Z)^{\prime} \beta\right)>V \wedge a \mid Z=z\right]+\mathbb{P}\left[l_{2}\left(\vartheta(Z)^{\prime} \beta\right)>V \vee a \mid Z=z\right] \\
& =\int_{-\infty}^{l_{1}\left(\vartheta(z)^{\prime} \beta\right)} f_{V \mid Z}(v \mid z) d v+\int_{l_{2}\left(\vartheta(z)^{\prime} \beta\right)}^{\infty} f_{V \mid Z}(v \mid z) d v \tag{9.3}
\end{align*}
$$

Taking derivatives by applying Leibnitz' rule produces

$$
\begin{align*}
\nabla_{z} m_{\bar{Y} \mid Z}(z)= & D_{z} \vartheta(z)^{\prime} \beta\left[\frac{\partial l_{1}}{\partial s}\left(\vartheta(z)^{\prime} \beta\right) f_{V \mid Z}\left(l_{1}\left(\vartheta(z)^{\prime} \beta\right) \mid z\right)-\frac{\partial l_{2}}{\partial s}\left(\vartheta(z)^{\prime} \beta\right) f_{V \mid Z}\left(l_{2}\left(\vartheta(z)^{\prime} \beta\right) \mid z\right)\right] \\
& +\int_{-\infty}^{l_{1}\left(\vartheta(z)^{\prime} \beta\right)} \nabla_{z}\left[\log f_{V \mid Z}(v \mid z)\right] f_{V \mid Z}(v \mid z) d v \\
& +\int_{l_{2}\left(\vartheta(z)^{\prime} \beta\right)}^{\infty} \nabla_{z}\left[\log f_{V \mid Z}(v \mid z)\right] f_{V \mid Z}(v \mid z) d v \\
= & D_{z} \vartheta(z)^{\prime} \beta[\cdots]+\mathbb{E}\left\{\tilde{Y} Q_{z}(V ; Z) \mid Z=z\right\} \tag{9.4}
\end{align*}
$$

where $Q_{z}(V ; Z)=\nabla_{z}\left[\log f_{V \mid Z}(V ; Z)\right]$, and all the integrals on the right hand side of the first and second equality exist by assumption 8.3 b

Finally, consider the case defined by assumptions 8.2c. For simplicity, consider the two dimensional case: $V=\left(V_{1}, V_{2}\right)$, i.e., $l(v)=a v_{1}+b v_{2}$. Then,

$$
\begin{align*}
m_{\bar{Y} \mid Z}(z) & =\mathbb{P}\left[\vartheta(Z)^{\prime} \beta>a V_{1}+b V_{2} \mid Z=z\right] \\
& =\mathbb{P}\left[V_{1}<a^{-1}\left(\vartheta(Z)^{\prime} \beta-b V_{2}\right) \mid Z=z\right] \\
& =\int_{-\infty}^{b^{-1} \vartheta(z)^{\prime} \beta} \int_{-\infty}^{a^{-1}\left(\vartheta(z)^{\prime} \beta-b v_{2}\right)} f_{V \mid Z}(v ; z) d v_{1} d v_{2} \tag{9.5}
\end{align*}
$$

To handle this expression, we need the following auxiliary lemma. Observe that ( $U, V, X$ ) are any random variables here:

Lemma A.1: Let $(U, V, X)$ (for simplicity) be random variables. Let the conditional density of $(U, V)$ given $X$ be denoted by $f(u, v ; x)$. Let

$$
F(x)=\int_{-\infty}^{\alpha(x)} \int_{-\infty}^{\beta(x, v)} f(u, v ; x) d u d v
$$

Then,

$$
\begin{gathered}
\partial_{x} F(x)=\int_{-\infty}^{\alpha(x)} \int_{-\infty}^{\beta(x, v)} \partial_{x} f(u, v ; x) d u d v+\partial_{x} \alpha(x) \int_{-\infty}^{\beta(x, \alpha(x))} f(u, \alpha(x) ; x) d u \\
+\int_{-\infty}^{\alpha(x)} \partial_{x} \beta(x, v) f(\beta(x, v), v ; x) d v
\end{gathered}
$$

Proof.

$$
\begin{aligned}
& F(x+h)=\int_{-\infty}^{\alpha(x+h)} \int_{-\infty}^{\beta(x+h, v)} f(u, v ; x+h) d u d v \\
& =\int_{-\infty}^{\alpha(x)+\partial_{x} \alpha(x) h} \int_{-\infty}^{\beta(x, v)+\partial_{x} \beta(x, v) h}\left[f(u, v ; x)+\partial_{x} f(u, v ; x) h\right] d u d v \\
& +O\left(h^{2}\right) \\
& =\left(\int_{-\infty}^{\alpha(x)}+\int_{\alpha(x)}^{\alpha(x)+\partial_{x} \alpha(x) h}\right)\left(\int_{-\infty}^{\beta(x, v)}+\int_{\beta(x, v)}^{\beta(x, v)+\partial_{x} \beta(x, v) h}\right)\left[f+\partial_{x} f h\right] d u d v \\
& +O\left(h^{2}\right) \\
& =\int_{-\infty}^{\alpha(x)} \int_{-\infty}^{\beta(x, v)}\left[f+\partial_{x} f h\right] d u d v+\int_{-\infty}^{\alpha(x)} \int_{\beta(x, v)}^{\beta(x, v)+\partial_{x} \beta(x, v) h}\left[f+\partial_{x} f h\right] d u d v \\
& +\int_{\alpha(x)}^{\alpha(x)+\partial_{x} \alpha(x) h} \int_{-\infty}^{\beta(x, v)}\left[f+\partial_{x} f h\right] d u d v \\
& +\int_{\alpha(x)}^{\alpha(x)+\partial_{x} \alpha(x) h} \int_{\beta(x, v)}^{\beta(x, v)+\partial_{x} \beta(x, v) h}\left[f+\partial_{x} f h\right] d u d v+O\left(h^{2}\right) \\
& =F(x)+h \int_{-\infty}^{\alpha(x)} \int_{-\infty}^{\beta(x, v)} \partial_{x} f d u d v+\int_{-\infty}^{\alpha(x)} \int_{\beta(x, v)}^{\beta(x, v)+\partial_{x} \beta(x, v) h} f d u d v \\
& +\int_{\alpha(x)}^{\alpha(x)+\partial_{x} \alpha(x) h} \int_{-\infty}^{\beta(x, v)} f d u d v+O\left(h^{2}\right) .
\end{aligned}
$$

Since

$$
\lim _{h \rightarrow 0} h^{-1} \int_{\alpha(x)}^{\alpha(x)+\partial_{x} \alpha(x) h} \underbrace{\int_{-\infty}^{\beta(x, v)} f(u, v ; x) d u d v}_{B(x, v)}=\partial_{x} \alpha(x) B(x, \alpha(x))
$$

and

$$
\begin{aligned}
& \lim _{h \rightarrow 0} h^{-1} \int_{-\infty}^{\alpha(x)} \int_{\beta(x, v)}^{\beta(x, v)+\partial_{x} \beta(x, v) h} f(u, v ; x) d u d v \\
= & \int_{-\infty}^{\alpha(x)} \lim _{h \rightarrow 0} h^{-1} \int_{\beta(x, v)}^{\beta(x, v)+\partial_{x} \beta(x, v) h} f(u, v ; x) d u d v \\
= & \int_{-\infty}^{\alpha(x)} \partial_{x} \beta(x, v) f(\beta(x, v), v ; x) d v,
\end{aligned}
$$

the assertion follows.
Q.E.D.

Adapting this result to our scenario produces

$$
\begin{aligned}
\nabla_{z} m_{\bar{Y} \mid Z}(z)= & a^{-1} \beta\left[\int_{-\infty}^{b^{-1} \vartheta(z)^{\prime} \beta} f_{V \mid Z}\left(a^{-1}\left(\vartheta(z)^{\prime} \beta-b v_{2}\right), v_{2} ; z\right) d v_{2}\right] \\
& +\int_{-\infty}^{b^{-1} \vartheta(z)^{\prime} \beta} \int_{-\infty}^{a^{-1}\left(\vartheta(z)^{\prime} \beta-b v_{2}\right)} \underbrace{\nabla_{z}\left[\ln f_{V \mid Z}\left(v_{1}, v_{2} ; z\right)\right]}_{Q_{z}\left(v_{1}, v_{2} ; z\right)} f_{V \mid Z}\left(v_{1}, v_{2} ; z\right) d v_{1} d v_{2}
\end{aligned}
$$

Finally, the 2nd term may be written

$$
\begin{aligned}
& \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} 1_{\left\{-\infty \leq v_{1} \leq a^{-1}\left(\vartheta(z)^{\prime} \beta-b v_{2}\right),-\infty \leq v_{2} \leq b^{-1} \vartheta(z)^{\prime} \beta\right\}} Q_{z}\left(v_{1}, v_{2} ; z\right) f_{V \mid Z}\left(v_{1}, v_{2} ; z\right) d v_{1} d v_{2} \text { (9.6) } \\
= & \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} 1_{\left\{a v_{1}+b v_{2} \leq \vartheta(z)^{\prime} \beta\right\}} Q_{z}\left(v_{1}, v_{2} ; z\right) f_{V \mid Z}\left(v_{1}, v_{2} ; z\right) d v_{1} d v_{2} \\
= & \mathbb{E}\left[Y Q_{z}(V ; Z) \mid Z=z\right],
\end{aligned}
$$

which shows the statement.

Ad iii. In the case of $U \perp V \mid Z$,

$$
\breve{Y}=\mathbb{E}[Y \mid Z, V]=F_{U \mid V}\left(\vartheta(Z)^{\prime} \beta+V^{\prime} \beta, V\right),
$$

provided assumption 6 holds. Next, we apply a similar logic as in the previous subsection, with $\breve{Y}$ in place of $\bar{Y}$. However, due to the law of iterated expectations, we have that $\mathbb{E}[\breve{Y} \mid Z]=$ $\mathbb{E}[Y \mid Z]$. Hence, $\mathbb{E}[Y \mid Z]=\mathbb{E}\left[F_{U \mid V}\left(\vartheta(Z)^{\prime} \beta+V^{\prime} \beta, V\right) \mid Z\right]$, and

$$
\nabla_{z} m_{Y \mid Z}(z)=\mathbb{E}\left[f_{U \mid V}\left(\vartheta(Z)^{\prime} \beta+V^{\prime} \beta, V\right) \mid Z=z\right] D_{z} \vartheta(z)^{\prime} \beta+\mathbb{E}\left[Y Q_{z}(V, Z) \mid Z=z\right]
$$

where $Q_{z}(v, z)=\nabla_{z} \log f_{V \mid Z}(v ; z)$, for all $z \in \mathcal{B}$, due to differentiability and domination assumptions 4. Rearranging terms, and premultiplying with $B(Z)$ and taking expectations produces (3.9) up to a constant of scale. This expectation exists again under the elementwise square integrability of all functions on $\mathcal{B}$ (assumption 4). Note that the right hand side of (3.9) may be rewritten as (3.10), using the law of iterated expectations.
Q.E.D.

## The Proof of Theorem 3

## The Structure of the Proof

Neglect for a moment the weighting function $A$. Rewrite (4.6) as $\frac{1}{n} \sum_{i} \hat{G}_{i}^{-} \hat{B}_{i}$, where $G_{i}=$ $D_{z} m_{X \mid Z}\left(Z_{i}\right)^{\prime}, B_{i}=\nabla_{z} m_{Y \mid Z}\left(Z_{i}\right), \hat{G}_{i}=D_{z} \hat{m}_{X \mid Z}\left(Z_{i}\right)^{\prime}$ and $\hat{B}_{i}=\nabla_{z} \hat{m}_{Y \mid Z}\left(Z_{i}\right)$. Tedious, but straightforward manipulations lead to

$$
\begin{align*}
\hat{G}_{i}^{-} \hat{B}_{i}= & G_{i}^{-} B_{i}+G_{i}^{-}\left[\left(G_{i}-\hat{G}_{i}\right) G_{i}^{-} B_{i}+\left(\hat{B}_{i}-B_{i}\right)\right] \\
& +G_{i}^{-}\left(G_{i}-\hat{G}_{i}\right)\left(G_{i}^{-}-\hat{G}_{i}^{-}\right) B_{i} \\
& +G_{i}^{-}\left(G_{i}-\hat{G}_{i}\right) G_{i}^{-}\left(\hat{B}_{i}-B_{i}\right)  \tag{9.7}\\
& +G_{i}^{-}\left(G_{i}-\hat{G}_{i}\right)\left(G_{i}^{-}-\hat{G}_{i}^{-}\right)\left(\hat{B}_{i}-B_{i}\right) .
\end{align*}
$$

Now, in (9.7) the first two terms on the right hand side will provide us with the asymptotic distribution, while the terms from three to five will prove asymptotically negligible. In Step 1, we treat the behavior of the first two summands first in the case where $\vartheta$ is a mean regression. Specifically, we show in Step 1a that

$$
\tau_{1 n}=n^{-1} \sum_{i} G_{i}^{-} B_{i}+G_{i}^{-}\left[\left(G_{i}-\hat{G}_{i}\right) G_{i}^{-} B_{i}+\left(\hat{B}_{i}-B_{i}\right)\right]=S_{1 n}+S_{2 n}
$$

i.e., the sum can be decomposed into two terms, the first of which provides us with the asymptotic distribution, while the second one produces the bias. In Step 1b, we establish that the large sample theory of $S_{1 n}$ may be handled using projection arguments coming from $U$-statistic theory, while in Step 1c we show that the bias term $S_{2 n}$ will vanish under appropriate conditions on the bandwidths, as in PSS. Finally, in Step 1d we derive the asymptotic distribution. In Step 2, we discuss the behavior of the higher order terms in (9.7), i.e., the behavior of terms three to five. In Step $\mathbf{3}$ we establish under which conditions generated dependent variables do not matter for the asymptotic distribution of the estimator.

## Step 1: The General Proof

Step 1a: Consider

$$
\begin{align*}
\tau_{1 n} & =n^{-1} \sum_{i}\left\{G_{i}^{-} B_{i}+G_{i}^{-}\left(G_{i}-\hat{G}_{i}\right) G_{i}^{-} B_{i}+G_{i}^{-}\left(\hat{B}_{i}-B_{i}\right)\right\}  \tag{9.8}\\
& =n^{-1} \sum_{i} G_{i}^{-} B_{i}+n^{-1} \sum_{i} G_{i}^{-}\left(G_{i}-\hat{G}_{i}\right) G_{i}^{-} B_{i}+n^{-1} \sum_{i} G_{i}^{-}\left(\hat{B}_{i}-B_{i}\right) \tag{9.9}
\end{align*}
$$

Since the first term has a trivial structure, and the second and third terms are similar, we start by considering the second term on the right hand side of (9.8) first. In the case where $\hat{G}_{i}$ is a nonparametric Nadaraya Watson derivative estimator, it rewrites as

$$
\begin{equation*}
\left(\sum_{j \neq i} \mathcal{K}_{h j}\left(Z_{i}\right)\right)^{-1}\left[\sum_{j \neq i} \nabla_{z} \mathcal{K}_{h j}\left(Z_{i}\right) X_{j}^{\prime}-\left(\sum_{j \neq i} \mathcal{K}_{h j}\left(Z_{i}\right)\right)^{-1} \sum_{j \neq i} \nabla_{z} \mathcal{K}_{h j}\left(Z_{i}\right) \sum_{j \neq i} \mathcal{K}_{h j}\left(Z_{i}\right) X_{j}^{\prime}\right] . \tag{9.10}
\end{equation*}
$$

Hence, $G_{i}-\hat{G}_{i}$ has a representation as

$$
D_{z} m_{X \mid Z}\left(Z_{i}\right)-\left(\sum_{j \neq i} \mathcal{K}_{h j}\left(Z_{i}\right)\right)^{-1}\left[\sum_{j \neq i}\left[\nabla_{z} \mathcal{K}_{h j}\left(Z_{i}\right)-W_{n}\left(Z_{i}\right) \mathcal{K}_{h j}\left(Z_{i}\right)\right]\left[V_{j}^{\prime}+m_{X \mid Z}\left(Z_{j}\right)^{\prime}\right]\right]
$$

where

$$
W_{n}\left(Z_{i}\right)=\left(\sum_{s \neq i} \mathcal{K}_{h s}\left(Z_{i}\right)\right)^{-1} \sum_{j \neq i} \nabla_{z} \mathcal{K}_{h s}\left(Z_{i}\right)
$$

Separate this expressions into the two parts, where

$$
\begin{align*}
-P_{1 i} & =\left(\sum_{j \neq i} \mathcal{K}_{h j}\left(Z_{i}\right)\right)^{-1}\left[\sum_{j \neq i}\left[\nabla_{z} \mathcal{K}_{h j}\left(Z_{i}\right)-W_{n}\left(Z_{i}\right) \mathcal{K}_{h j}\left(Z_{i}\right)\right] V_{j}^{\prime}\right]  \tag{9.11}\\
& =(n-1)^{-1} \sum_{j \neq i} \mathcal{W}_{j n}\left(Z_{i}\right) V_{j}^{\prime}
\end{align*}
$$

where $\mathcal{W}_{j n}\left(Z_{i}\right)=\left((n-1)^{-1} \sum_{j \neq i} \mathcal{K}_{h j}\left(Z_{i}\right)\right)^{-1}\left[\nabla_{z} \mathcal{K}_{h j}\left(Z_{i}\right)-W_{n}\left(Z_{i}\right) \mathcal{K}_{h j}\left(Z_{i}\right)\right]$ and

$$
\begin{aligned}
P_{2 i} & =D_{z} m_{X \mid Z}\left(Z_{i}\right)^{\prime}-\left(\sum_{j \neq i} \mathcal{K}_{h j}\left(Z_{i}\right)\right)^{-1}\left[\sum_{j \neq i}\left[\nabla_{z} \mathcal{K}_{h j}\left(Z_{i}\right)-W_{n}\left(Z_{i}\right) \mathcal{K}_{h j}\left(Z_{i}\right)\right] m_{X \mid Z}\left(Z_{j}\right)^{\prime} 9.12\right) \\
& =D_{z} m_{X \mid Z}\left(Z_{i}\right)^{\prime}-(n-1)^{-1} \sum_{j \neq i} \mathcal{W}_{j n}\left(Z_{i}\right) m_{X \mid Z}\left(Z_{j}\right)^{\prime}
\end{aligned}
$$

Note that $\mathcal{W}_{j n}\left(Z_{i}\right)=-\mathcal{W}_{i n}\left(Z_{j}\right)$ by the symmetry of the kernel. The first part, (9.11), will contribute to the variance of the estimators, whereas the second will be produce the leading
bias term for which we shall give conditions under which it vanishes. Rewriting

$$
\begin{aligned}
n^{-1} \sum_{i} G_{i}^{-}\left(G_{i}-\hat{G}_{i}\right) G_{i}^{-} B_{i}= & -(n(n-1))^{-1} \sum_{i} \sum_{j \neq i} G_{i}^{-} \mathcal{W}_{j n}\left(Z_{i}\right) V_{j}^{\prime} G_{i}^{-} B_{i} \\
& +(n(n-1))^{-1} \sum_{i} G_{i}^{-} P_{2 i} G_{i}^{-} B_{i} \\
= & S_{1 n}^{2}+S_{2 n}^{2},
\end{aligned}
$$

where the superscript 2 denotes the second term in the expression (9.8). A similar decomposition may be performed on $n^{-1} \sum_{i} G_{i}^{-}\left(\hat{B}_{i}-B_{i}\right)=(n(n-1))^{-1} \sum_{i} \sum_{j \neq i} G_{i}^{-} \mathcal{W}_{j n}\left(Z_{i}\right) Q_{j}+$ $(n(n-1))^{-1} \sum_{i} G_{i}^{-} P_{4 i}=S_{1 n}^{3}+S_{2 n}^{3}$, where $Q_{i}=Y_{i}-m_{Y \mid Z}\left(Z_{i}\right)$ and $P_{4 i}$ denotes again bias terms in the regression of $Y$ on $Z$. In total, we obtain that

$$
\begin{equation*}
\tau_{1 n}=n^{-1} \sum_{i} G_{i}^{-} B_{i}+S_{1 n}^{2}+S_{1 n}^{3}+S_{2 n}^{2}+S_{2 n}^{3}=S_{1 n}+S_{2 n} \tag{9.13}
\end{equation*}
$$

where $S_{1 n}=n^{-1} \sum_{i} G_{i}^{-} B_{i}+S_{1 n}^{2}+S_{1 n}^{3}$ collects all terms that affect the asymptotic distribution, while $S_{2 n}=S_{2 n}^{2}+S_{2 n}^{3}$ are all bias terms that vanish under appropriate conditions.

Step 1b: To analyze all terms that affect the distribution and are contained in $S_{1 n}$, consider $S_{1 n}^{2}$ first. Manipulating this expression produces

$$
\begin{aligned}
U_{n} & =(n(n-1))^{-1} \sum_{i} \sum_{j>i}\left\{G_{i}^{-} \mathcal{W}_{j n}\left(Z_{i}\right) V_{j}^{\prime} G_{i}^{-} B_{i}-G_{j}^{-} \mathcal{W}_{j n}\left(Z_{i}\right) V_{i}^{\prime} G_{j}^{-} B_{j}\right\} \\
& =(n(n-1))^{-1} \sum_{i} \sum_{j>i} p_{n}\left(S_{i}, S_{j}\right)
\end{aligned}
$$

where $S_{i}=\left(Y_{i}, X_{i}^{\prime}, Z_{i}^{\prime}\right)^{\prime}$, with $p_{n}$ symmetric, and we made use of $\mathcal{W}_{j n}\left(Z_{i}\right)=-\mathcal{W}_{i n}\left(Z_{j}\right)$. To apply Lemma 3.1 of PSS which yields $\sqrt{n}\left(\hat{U}_{n}-U_{n}\right)=o_{p}(1)$, where

$$
\begin{equation*}
\hat{U}_{n}=\theta+n^{-1} \sum_{i} \mathbb{E}\left[p_{n}\left(S_{i}, S_{j}\right) \mid S_{i}\right] \tag{9.14}
\end{equation*}
$$

we require that $\mathbb{E}\left(\left\|p_{n}\left(S_{i}, S_{j}\right)\right\|^{2}\right)=o(n)$. Following similar and straightforward, but more tedious arguments as in PSS, this is the case provided $n h^{L+2} \rightarrow \infty$. To analyze (9.14), note first that $\theta=\mathbb{E}\left[p_{n}\left(S_{i}, S_{j}\right)\right]=0$, and consider first $p_{n}^{*}$ which equals $p_{n}$ save that in $\mathcal{W}_{j n}\left(Z_{i}\right)$, $(n-1)^{-1} \sum_{s \neq i} \mathcal{K}_{h s}\left(Z_{i}\right)$ and $(n-1)^{-1} \sum_{s \neq i} \nabla_{z} \mathcal{K}_{h s}\left(Z_{i}\right)$ are replaced with their probability limits,
$f_{Z}\left(Z_{i}\right)$.and $\nabla_{z} f_{Z}\left(Z_{i}\right)$. Then,

$$
\begin{align*}
& \mathbb{E}\left[p_{n}^{*}\left(S_{i}, S_{j}\right) \mid S_{i}=s_{i}\right] \\
= & \int h^{-(L+1)}\left(D_{z} m_{X \mid Z}\left(z_{i}\right)^{-} f_{Z}\left(z_{i}\right)^{-1}\left(\nabla_{z} \mathcal{K}\left(\left(z_{i}-z\right) / h\right)-\nabla_{z} f_{Z}\left(z_{i}\right) f_{Z}\left(z_{i}\right)^{-1} h \mathcal{K}\left(\left(z_{i}-z\right) / h\right)\right)\right. \\
& \times v_{i}^{\prime} D_{z} m_{X \mid Z}\left(z_{i}\right)^{-} \nabla_{z} m_{Y \mid Z}\left(z_{i}\right) f_{Z}(z) d z \\
= & D_{z} m_{X \mid Z}\left(z_{i}\right)^{-} f_{Z}\left(z_{i}\right)^{-1} \int h^{-1} \nabla_{z} \mathcal{K}(\psi) f_{Z}\left(z_{i}+\psi h\right) d \psi v_{i}^{\prime} D_{z} m_{X \mid Z}\left(z_{i}\right)^{-} \nabla_{z} m_{Y \mid Z}\left(z_{i}\right)  \tag{9.15}\\
& -D_{z} m_{X \mid Z}\left(z_{i}\right)^{-} f_{Z}\left(z_{i}\right)^{-2} \nabla_{z} f_{Z}\left(z_{i}\right) \int \mathcal{K}(\psi) f_{Z}\left(z_{i}+\psi h\right) d \psi v_{i}^{\prime} D_{z} m_{X \mid Z}\left(z_{i}\right)^{-} \nabla_{z} m_{Y \mid Z}\left(z_{i}\right) \\
= & -D_{z} m_{X \mid Z}\left(z_{i}\right)^{-} f_{Z}\left(z_{i}\right)^{-1} \nabla_{z} f_{Z}\left(z_{i}\right) v_{i}^{\prime} D_{z} m_{X \mid Z}\left(z_{i}\right)^{-} \nabla_{z} m_{Y \mid Z}\left(z_{i}\right)+\eta_{2 i} \\
= & -g_{i}^{-} f_{Z}\left(z_{i}\right)^{-1} \nabla_{z} f_{Z}\left(z_{i}\right) v_{i}^{\prime} g_{i}^{-} b_{i}+\eta_{2 i},
\end{align*}
$$

where $\eta_{2 i}$ denotes higher order terms, for which, by standard arguments $n^{-1 / 2} \sum_{i} \eta_{2 i}=o_{p}(1)$ (Here we use $g_{i}^{-}, b_{i}$ to denote $G_{i}^{-}, B_{i}$ at a fixed position $z_{i}$. We will now that we may replace $p_{n}$ by $p_{n}^{*}$ at the expense of a higher order term that vanishes as well (under boundedness assumptions on the densities), i.e.,

$$
n^{-1 / 2} \sum_{i} \mathbb{E}\left[p_{n}\left(S_{i}, S_{j}\right)-p_{n}^{*}\left(S_{i}, S_{j}\right) \mid S_{i}\right]=o_{p}(1)
$$

To see this, consider a typical expression in $\mathbb{E}\left[p_{n}\left(S_{i}, S_{j}\right)-p_{n}^{*}\left(S_{i}, S_{j}\right) \mid S_{i}\right]$. Using the right hand side of the third equality in equation (9.15),

$$
\begin{aligned}
\rho_{n i} & =D_{z} m_{X \mid Z}\left(z_{i}\right)^{-}\left\{\hat{f}_{Z}\left(z_{i}\right)^{-1}-f_{Z}\left(z_{i}\right)^{-1}\right\} \nabla_{z} f_{Z}\left(z_{i}\right) v_{i}^{\prime} D_{z} m_{X \mid Z}\left(z_{i}\right)^{-} \nabla_{z} m_{Y \mid Z}\left(z_{i}\right) \\
& =\left\{f_{Z}\left(z_{i}\right)-\hat{f}_{Z}\left(z_{i}\right)\right\} \hat{f}_{Z}\left(z_{i}\right)^{-1} D_{z} m_{X \mid Z}\left(z_{i}\right)^{-} f_{Z}\left(z_{i}\right)^{-1} \nabla_{z} f_{Z}\left(z_{i}\right) v_{i}^{\prime} D_{z} m_{X \mid Z}\left(z_{i}\right)^{-} \nabla_{z} m_{Y \mid Z}\left(z_{i}\right)
\end{aligned}
$$

where $\hat{f}_{Z}\left(z_{i}\right)=(n-1)^{-1} \sum_{s \neq i} \mathcal{K}_{h s}\left(Z_{i}\right)$. Next, write

$$
\begin{equation*}
n^{-1 / 2} \sum_{i} \rho_{n i}=n^{1 / 2} \int \frac{f_{Z}(z)-\hat{f}_{Z}(z)}{\hat{f}_{Z}(z)} \chi(z, v) \hat{F}_{Z V}(d z, d v), \tag{9.16}
\end{equation*}
$$

where $\chi(z, v)=D_{z} m_{X \mid Z}(z)^{-} f_{Z}(z)^{-1} \nabla_{z} f_{Z}(z) v^{\prime} D_{z} m_{X \mid Z}(z)^{-} \nabla_{z} m_{Y \mid Z}(z)$, and $\hat{F}_{Z V}$ denotes the empirical cdf. Considering the denominator in (9.16), observe that

$$
\begin{equation*}
\frac{1}{\left|f_{Z}(z)+\hat{f}_{Z}(z)-f_{Z}(z)\right|} \leq \frac{1}{\left|f_{Z}(z)\right|-\left|\hat{f}_{Z}(z)-f_{Z}(z)\right|} \leq \frac{2}{b} \tag{9.17}
\end{equation*}
$$

since $f_{Z}(z) \geq b$ by the assumption that $Z$ is continuously distributed RV on $\mathcal{B}$, with density bounded away from zero. Moreover, $\left|\hat{f}_{Z}(z)-f_{Z}(z)\right| \leq b / 2$ with probability going to one, as $\hat{f}_{Z}(z)$ is consistent by assumptions on kernels and bandwidths. Hence, $n^{-1 / 2} \sum_{i} \rho_{n i}$ is, in absolute value, bounded by

$$
\begin{equation*}
c \sup _{z \in \mathcal{B}}\left|f_{Z}(z)-\hat{f}_{Z}(z)\right| n^{-1 / 2} \sum_{i}\left|\chi\left(Z_{i}, V_{i}\right)\right|, \tag{9.18}
\end{equation*}
$$

But since $n^{-1 / 2} \sum_{i}\left|\chi\left(Z_{i}, V_{i}\right)\right|$ converges by a standard CLT for $i i d$ random variables to a normal limit (provided the second moment are finite which we tacitly assume), and $\sup _{z \in \mathcal{B}}\left|f_{Z}(z)-\hat{f}_{Z}(z)\right|=$ $O_{p}\left(h^{2 r}+\left(n h^{L}\right)^{-1 / 2} \log n\right)=o_{p}(1)$ under general conditions, it follows that $n^{-1 / 2} \sum_{i} \rho_{n i}=$ $o_{p}(1)$. Similar arguments can be applied to any other term appearing in $\mathbb{E}\left[p_{n}\left(S_{i}, S_{j}\right)-p_{n}^{*}\left(S_{i}, S_{j}\right) \mid S_{i}\right]$, implying that the difference vanishes.

Repeating the same arguments as from the start of Step 1b, we can show that $G_{i}^{-}\left(\hat{B}_{i}-B_{i}\right)=$ $G_{i}^{-} f_{Z}\left(Z_{i}\right)^{-1} \nabla_{z} f_{Z}\left(Z_{i}\right) Q_{i}+\eta_{3 n}$, where $Q_{i}=Y_{i}-m_{Y \mid Z}\left(Z_{i}\right)$. Returning to (9.13)

$$
\begin{aligned}
S_{1 n} & =n^{-1} \sum_{i}\left\{G_{i}^{-} B_{i}+G_{i}^{-}\left[\left(G_{i}-\hat{G}_{i}\right) G_{i}^{-} B_{i}+\left(\hat{B}_{i}-B_{i}\right)\right]\right\} \\
& =n^{-1} \sum_{i}\left\{G_{i}^{-} B_{i}-G_{i}^{-} f_{Z}\left(Z_{i}\right)^{-1} \nabla_{z} f_{Z}\left(Z_{i}\right) V_{i}^{\prime} G_{i}^{-} B_{i}-G_{i}^{-} f_{Z}\left(Z_{i}\right)^{-1} \nabla_{z} f_{Z}\left(Z_{i}\right) Q_{i}\right\}+n^{-1} \sum_{i} T_{3 i},
\end{aligned}
$$

where and $T_{3 i}$ denotes all higher order terms that. Note that $\sqrt{n}\left[n^{-1} \sum_{i} T_{3 i}\right]=o_{p}(1)$, by arguments above..

Step 1c: To analyze all terms that affect the distribution and are contained in $S_{2 n}$, consider $S_{2 n}^{2}$ first. More specifically,

$$
\begin{aligned}
S_{2 n}^{2}= & n^{-1 / 2} \sum_{i} G_{i}^{-}\left\{D_{z} m_{X \mid Z}\left(Z_{i}\right)^{\prime}-(n-1)^{-1} \sum_{j \neq i} \mathcal{W}_{j n}\left(Z_{i}\right) m_{X \mid Z}\left(Z_{j}\right)^{\prime}\right\} G_{i}^{-} B_{i} \\
= & \sqrt{n} \iint\left[D_{z} m_{X \mid Z}(\zeta)^{\prime}\right]^{-} \\
& \times\left\{D_{z} m_{X \mid Z}(\zeta)^{\prime}-\left[\frac{h^{-L-1} \nabla_{z} \mathcal{K}((z-\zeta) / h)}{\hat{f}_{Z}(\zeta)}-\frac{\nabla_{z} \hat{f}_{Z}(\zeta) h^{-L} \mathcal{K}((z-\zeta) / h)}{\left(\hat{f}_{Z}(\zeta)\right)^{2}}\right] m_{X \mid Z}(z)^{\prime}\right\} \\
& \times\left[D_{z} m_{X \mid Z}(\zeta)^{\prime}\right]^{-} \nabla_{z} m_{Y \mid Z}(\zeta) \hat{F}_{Z}(d z) \hat{F}_{Z}(d \zeta) .
\end{aligned}
$$

Next, let $S_{2 n}^{2}=A^{1}+\omega_{n}$ where $A^{1}$ equals $S_{2 n}^{2}$ with the exception that we replace $\hat{F}_{Z}$ by $F_{Z}$, and we replace $\hat{f}_{Z}(\zeta)$ by $f_{Z}(\zeta)$. Hence we get a remainder term that contains expressions of the form $\hat{F}_{Z}-F_{Z}$ and $\hat{f}_{Z}(\zeta)-f_{Z}(\zeta)$. In the case of the replacement of $\hat{F}_{Z}$ by $F_{Z}$, we can appeal to Glivenko-Cantelli together with the fact that $\mathcal{B}$ is compact, and by arguments as in equations (9.16) and (9.17), we can show that $\omega_{n}=o_{p}\left(A^{1}\right)$, so that we focus on the leading term $A^{1}$. After change of variable, this is

$$
\begin{aligned}
& \sqrt{n} \int\left[D_{z} m_{X \mid Z}(\zeta)^{\prime}\right]^{-}\left\{D_{z} m_{X \mid Z}(\zeta)^{\prime}-\right. \\
& \left.\times \int \frac{h^{-1} \nabla_{\psi} \mathcal{K}(\psi)}{f_{Z}(\zeta)} m_{X \mid Z}(\psi h+\zeta)^{\prime} f_{Z}(\psi h+\zeta) d \psi-\int \frac{\nabla_{z} f_{Z}(\zeta) \mathcal{K}(\psi)}{\left(f_{Z}(\zeta)\right)^{2}} m_{X \mid Z}(\psi h+\zeta)^{\prime} f_{Z}(\psi h+\zeta) d \psi\right\} \\
& \times\left[D_{z} m_{X \mid Z}(\zeta)^{\prime}\right]^{-} \nabla_{z} m_{Y \mid Z}(\zeta) f_{Z}(\zeta) d \zeta .
\end{aligned}
$$

Then make use of partial integration and apply a standard Taylor expansion, to obtain that

$$
\begin{equation*}
A^{1}=\sqrt{n} \int\left[D_{z} m_{X \mid Z}(\zeta)^{\prime}\right]^{-} B X(\zeta)\left[D_{z} m_{X \mid Z}(\zeta)^{\prime}\right]^{-} \nabla_{z} m_{Y \mid Z}(\zeta) f_{Z}(\zeta) d \zeta+O\left(\sqrt{n} h^{r}\right) \tag{9.19}
\end{equation*}
$$

where denotes higher order bias terms, i.e. $B X(\zeta)=\sum_{l=1 ., r} \mu_{k} h^{l} B X_{l}(\zeta)$, and $B X_{l}(\zeta)$ contains sums of products of all higher order derivatives of $m_{X \mid Z}$ and $f_{Z}$, where the order of the product of derivatives combined is at most of order $l+1$. The expectations of these terms exist due to assumption 9, and provided that $r=2 L$ in connection with assumption 11. Consequently, $\sqrt{n} S_{2 n}^{2}=o_{p}(1)$. Under samilar conditions on $B Y(\zeta)$ (cf. assumption 9 ), and by similar arguments $\sqrt{n} S_{3 n}^{2}=o_{p}(1)$ and hence the bias expression proves asymptotically negligible under our assumptions.

Step 1d: Finally, the first terms provide us with the variance. Since $\mathbb{E}\left[\left(V_{i}^{\prime}, Q_{i}\right)^{\prime} \mid Z_{i}\right]=0, \sigma_{1 i}$ is uncorrelated with $\sigma_{2 i}$ and $\sigma_{3 i}$. The result follows by application of a standard central limit theorem.
Q.E.D.

## Step 2: The Behavior of Higher Order Terms

The characteristic feature of all terms in the expansion is that they involve higher powers in $G_{i}-\hat{G}_{i}$ or $\hat{B}_{i}-B_{i}$. Intuitively, what happens is that these terms will add a factor that tends to zero faster as the variance terms cancel, and the term is of the order of the squared bias terms. To fix ideas, recall that $B_{i}=\nabla_{z} m_{Y \mid Z}\left(Z_{i}\right)$ and consider

$$
\begin{aligned}
& n^{-1 / 2} \sum_{i} G_{i}^{-}\left(G_{i}-\hat{G}_{i}\right)\left(\hat{B}_{i}-B_{i}\right) \\
= & n^{1 / 2} \int G_{i}^{-}\left(m_{X \mid Z}(z)^{\prime}-D_{z} \widehat{m}_{X \mid Z}(z)^{\prime}\right)\left(\nabla_{z} \widehat{m}_{Y \mid Z}(z)-\nabla_{z} m_{Y \mid Z}(z)\right) \hat{F}_{Z}(d z) .
\end{aligned}
$$

The expression on the right hand side is in absolute value bounded by

$$
n^{1 / 2} \sup _{z \in \mathcal{B}}\left|D_{z} m_{X \mid Z}(z)^{\prime}-D_{z} \widehat{m}_{X \mid Z}(z)^{\prime}\right| \sup _{z \in \mathcal{B}}\left|\nabla_{z} \widehat{m}_{Y \mid Z}(z)-\nabla_{z} m_{Y \mid Z}(z)\right| \underbrace{n^{1 / 2} \int\left|G_{i}^{-}\right| \hat{F}_{Z}(d z)}_{C_{n}}
$$

Since $\sup _{z \in \mathcal{B}}\left|D_{z} m_{X \mid Z}(z)^{\prime}-D_{z} \widehat{m}_{X \mid Z}(Z)^{\prime}\right| \sup _{z \in \mathcal{B}}\left|\nabla_{z} \widehat{m}_{Y \mid Z}(z)-\nabla_{z} m_{Y \mid Z}(z)\right|=O_{p}\left(h^{2 r}+\left(n h^{L+2}\right)^{-1} \ln (n)\right)$ by an extensions to a theorem of Masry (1994), and $C_{n}$ converges to a nondegenerate random variable, provided that the second moment of $G_{i}^{-}$is finite (which is implied by assumption 3), this term is $o_{p}(1)$ under general conditions. Materially similar, yet more involved arguments can be used to establish the assertion for the other higher order terms, using assumptions 3 and 4.
Q.E.D.

## Step 3: Modifications with Generated Dependent Variables - Theorem 4

To have an idea why $T_{2 n}$ and $T_{3 n}$ vanish, consider first $T_{2 n}$ in Step 3a, and then $T_{3 n}$ in Step 3b.

Step 3a: Recall that

$$
T_{2 n}=n^{-1} \sum_{i}\left[D_{z} \widehat{m}_{X \mid Z}\left(Z_{i}\right)^{\prime}\right]^{-} \sum_{j \neq i} \nabla_{z} W_{j}\left(Z_{i}\right)\left[\mathbb{K}\left\{\left(P_{j}-0.5\right) / h\right\}-\mathbb{I}\left\{P_{j}<0.5\right\}\right] B\left(Z_{j}\right) .
$$

As before, we analyze this expression in several steps. We start by considering

$$
T_{2 n}^{*}=n^{-1} \sum_{i}\left[D_{z} m_{X \mid Z}\left(Z_{i}\right)^{\prime}\right]^{-} \sum_{j \neq i} \nabla_{z} W_{j}\left(Z_{i}\right)\left[\mathbb{K}\left\{\left(P_{j}-0.5\right) / h\right\}-\mathbb{I}\left\{P_{j}<0.5\right\}\right] B\left(Z_{j}\right),
$$

and note that $T_{2 n}=T_{2 n}^{*}+R_{n}$, where $R_{n}$ contains the difference $\left[D_{z} \widehat{m}_{X \mid Z}\left(Z_{i}\right)^{\prime}\right]^{-}-\left[D_{z} m_{X \mid Z}\left(Z_{i}\right)^{\prime}\right]^{-}$ instead of $\left[D_{z} m_{X \mid Z}\left(Z_{i}\right)^{\prime}\right]^{-}$. As is easy to see (given the discussion above), $R_{n}$ produces a faster vanishing higher order bias term. Quite obviously, this expression has a similar structure as the one analyzed in Step 1 b above, safe for the fact that $Y_{j}$ is replaced by $\mathbb{K}\left\{\left(P_{j}-0.5\right) / h\right\}-$ $\mathbb{I}\left\{P_{j}<0.5\right\}$. Following the same argumentation as the one in Step1c, we arrive at the crucial decomposition $T_{2 n}^{*}=T_{2 n}^{* *}+\varrho_{n}$, where $\varrho_{n}$ are terms that converge faster by Glivenko-Cantelli and compact support $\mathcal{B}$, and $T_{2 n}^{* *}$ is defined as follows:

$$
\begin{aligned}
= & \iiint h^{-(L+1)}\left(D_{z} m_{X \mid Z}^{* *}(\zeta)^{-} f_{Z}(\zeta)^{-1}\left(\nabla_{\zeta} \mathcal{K}((z-\zeta) / h)-\nabla_{z} f_{Z}(\zeta) f_{Z}(\zeta)^{-1} h \mathcal{K}((\zeta-z) / h)\right)\right. \\
& \times[\mathbb{K}\{(p-0.5) / h\}-\mathbb{I}\{p<0.5\}] D_{z} m_{X \mid Z}(\zeta)^{-} \nabla_{z} m_{Y \mid Z}(\zeta) F_{P Z}(d p, d z) F_{Z}(d \zeta) \\
= & \int g(\zeta) \iint h^{-1} \nabla_{\psi} \mathcal{K}(\psi)[\mathbb{K}\{(p-0.5) / h\}-\mathbb{I}\{p<0.5\}] f_{Z}(\zeta+\psi h) F_{P \mid Z}(d p ; \zeta+\psi h) d \psi \times \\
& D_{z} m_{X \mid Z}(\zeta)^{-} \nabla_{z} m_{Y \mid Z}(\zeta)-g(\zeta) \nabla_{z} f_{Z}(\zeta) \iint \mathcal{K}(\psi)[\mathbb{K}\{(p-0.5) / h\}-\mathbb{I}\{p<0.5\}] \times \\
& f_{Z}(\zeta+\psi h) F_{P \mid Z}(d p ; \zeta+\psi h) d \psi g(\zeta) \nabla_{z} m_{Y \mid Z}(\zeta) F_{Z}(d \zeta) \\
= & Q_{1 n}-Q_{2 n},
\end{aligned}
$$

where $g(\zeta)=D_{z} m_{X \mid Z}(\zeta)^{-} f_{Z}(\zeta)^{-1}$. Next, consider the inner integral in $Q_{1 n}$ :

$$
\begin{align*}
& h^{-1} \iint \nabla_{\psi} \mathcal{K}(\psi) \mathbb{K}\{(p-0.5) / h\} f_{Z}(\zeta+\psi h) d F_{P \mid Z}(d p ; \zeta+\psi h) d \psi \\
& -h^{-1} \iint \nabla_{\psi} \mathcal{K}(\psi) \mathbb{I}\{p<0.5\} f_{Z}(\zeta+\psi h) d F_{P \mid Z}(d p ; \zeta+\psi h) d \psi \\
= & h^{-1} \iint \nabla_{\psi} \mathcal{K}(\psi) K(\tau) F_{P \mid Z}(0.5+h \tau ; \zeta+\psi h) f_{Z}(\zeta+\psi h) d \tau d \psi  \tag{9.20}\\
& -h^{-1} \int \nabla_{\psi} \mathcal{K}(\psi) F_{P \mid Z}(0.5 ; \zeta+\psi h) f_{Z}(\zeta+\psi h) d \psi,
\end{align*}
$$

where we made use of Fubini's theorem in connection with standard arguments for integrals of kernels. Next, use integration by parts to obtain that the rhs of (9.20) equals

$$
\begin{align*}
& \int \mathcal{K}(\psi) \nabla_{\psi}\left[F_{P \mid Z}(0.5 ; \zeta+\psi h) f_{Z}(\zeta+\psi h)\right] d \psi  \tag{9.21}\\
& -\int \mathcal{K}(\psi) \int K(\tau) \nabla_{\psi}\left[F_{P \mid Z}(0.5+h \tau ; \zeta+\psi h) f_{Z}(\zeta+\psi h)\right] d \tau d \psi
\end{align*}
$$

Inserting $F_{P \mid Z}(0.5+h \tau ; \zeta+\psi h)=F_{P \mid Z}(0.5 ; \zeta+\psi h)+h \tau f_{P \mid Z}(0.5 ; \zeta+\psi h)+\ldots$ $+\left(r_{1}!\right)^{-1} h^{r_{1}} \tau^{r_{1}} \partial_{p}^{r_{1}-1} f_{P \mid Z}(0.5+\lambda h \tau ; \zeta+\psi h)$, where $\lambda \in(0,1)$, we obtain that (9.21) reduces, under the familiar assumption on all moments of the kernel up to order $r_{1}$ to be zero to

$$
-\int \mathcal{K}(\psi)\left(r_{1}!\right)^{-1} h^{r_{1}} \mu_{r_{1}} \nabla_{\psi} \partial_{p}^{r_{1}-1} f_{P Z}(0.5 ; \zeta+\psi h) d \psi
$$

plus a term of smaller order. Applying standard arguments, in particular expand $\partial_{p}^{r_{1}-1} f_{P \mid Z}(0.5 ; \zeta+\psi h)$ in $\psi$, we obtain that $\sqrt{n} T_{2 n}^{* *}=o_{p}(1)$, provided that $\sqrt{n} h^{r_{1}} h^{r}=o(1)$. The same argumentation holds for $Q_{2 n}$.

Step 3b: Next, consider

$$
T_{3 n}=n^{-1} \sum_{i}\left[D_{z} \widehat{m}_{X \mid Z}\left(Z_{i}\right)^{\prime}\right]^{-} \sum_{j \neq i} \nabla_{z} W_{j}\left(Z_{i}\right)\left[\mathbb{K}\left\{\left(\widehat{P}_{j}-0.5\right) / h\right\}-\mathbb{K}\left\{\left(P_{j}-0.5\right) / h\right\}\right] B\left(Z_{j}\right)
$$

we can rewrite the last term on the right hand side as:

$$
\begin{aligned}
T_{3 n} & =n^{-1} \sum_{i}\left[D_{z} \widehat{m}_{X \mid Z}\left(Z_{i}\right)^{\prime}\right]^{-} \sum_{j \neq i} \nabla_{z} W_{j}\left(Z_{i}\right) h^{-1} K\left\{\left(P_{j}-0.5\right) / h\right\}\left(\widehat{P}_{j}-P_{j}\right) B\left(Z_{j}\right)+R_{1 n} \\
& =T_{4 n}+R_{1 n}
\end{aligned}
$$

where $R_{1 n}$ denotes higher order terms in a mean value expansion, and $R_{n}=o_{p}\left(T_{4 n}\right)$. Using again $\left[D_{z} \widehat{m}_{X \mid Z}\left(Z_{i}\right)^{\prime}\right]^{-}=\left[D_{z} m_{X \mid Z}\left(Z_{i}\right)^{\prime}\right]^{-}+\left[\left[D_{z} \widehat{m}_{X \mid Z}\left(Z_{i}\right)^{\prime}\right]^{-}-\left[D_{z} m_{X \mid Z}\left(Z_{i}\right)^{\prime}\right]^{-}\right]$, which produces a leading term $T_{5 n}$ and again a faster converging remainder, we find that

$$
\begin{aligned}
& \sqrt{n} T_{5 n} \\
= & n^{-1 / 2} \sum_{i}\left[D_{z} m_{X \mid Z}\left(Z_{i}\right)^{\prime}\right]^{-} n^{-1} \sum_{j \neq i} \nabla_{z} \frac{h^{-L-2} \mathcal{K}\left(\left(Z_{j}-Z_{i}\right) / h\right)}{\hat{f}_{Z}\left(Z_{i}\right)} K\left(\left(P_{j}-0.5\right) / h\right)\left(\widehat{P}_{j}-P_{j}\right) \\
= & \sqrt{n} \iint\left[D_{z} m_{X \mid Z}(z)^{\prime}\right]^{-} h^{-2} \int \nabla_{\psi} \frac{\mathcal{K}(\psi)}{\hat{f}_{Z}(z)} K_{1}\left(\left(p\left(z+\psi h_{1}, \varpi\right)-0.5\right) / h\right) \\
& \times\left(\widehat{p}\left(z+\psi h_{1}, \varpi\right)-p\left(z+\psi h_{1}, \varpi\right)\right) f_{Z V}(z+\psi h, \varpi) d \psi d \varpi F_{Z}(d z) \\
= & \sqrt{n} \iint\left[D_{z} m_{X \mid Z}(z)^{\prime}\right]^{-} h^{-2} \hat{f}_{Z}(z)^{-1} \times \\
& \nabla_{z}\left[K_{1}((p(z, \varpi)-0.5) / h)(\widehat{p}(z, \varpi)-p(z, \varpi)) f_{Z V}(z, \varpi)\right] d \varpi F_{Z}(d z)+\rho_{n} \\
= & T_{6 n}+\rho_{n},
\end{aligned}
$$

where $\rho_{n}=o_{p}\left(T_{6 n}\right)$. Hence, $\sqrt{n} T_{5 n}$ is bounded in absolute value by

$$
\begin{equation*}
\left.\left.c_{1} \sup _{z, v \in \mathcal{B} \times \mathcal{V}} \mid \nabla_{z} \widehat{p}(z, v)-\nabla_{z} p(z, v)\right)\left|b_{1 n}+c_{2} \sup _{z, v \in \mathcal{B} \times \mathcal{V}}\right| \widehat{p}(z, v)-p(z, v)\right) \mid b_{2 n}, \tag{9.22}
\end{equation*}
$$

where $b_{1 n}=n^{-1 / 2} \sum_{i}\left|\left[D_{z} m_{X \mid Z}\left(Z_{i}\right)^{\prime}\right]^{-} K_{1}\left(\left(p\left(Z_{i}, V_{i}\right)-0.5\right) / h_{1}\right) f_{Z V}\left(Z_{i}, V_{i}\right)\right|$ and $b_{2 n}=n^{-1 / 2} \sum_{i}\left|\left[D_{z} m_{X \mid Z}\left(Z_{i}\right)^{\prime}\right]^{-} \nabla_{z}\left[K_{1}\left(\left(p\left(Z_{i}, V_{i}\right)-0.5\right) / h_{1}\right) f_{Z V}\left(Z_{i}, V_{i}\right)\right]\right|$ converge to nondegenerate distributions. To see this, pick $b_{1 n}=h_{1}^{1 / 2}\left(n h_{1}\right)^{-1 / 2} \sum_{i}\left|\left[D_{z} m_{X \mid Z}\left(Z_{i}\right)^{\prime}\right]^{-}\right| f_{Z V}\left(Z_{i}, V_{i}\right) \times$ $K_{1}\left(\left(P_{i}-0.5\right) / h_{1}\right)=h_{1}^{1 / 2} b_{3 n}$, where $b_{3 n}$ is a nonparametric estimator of

$$
\mathbb{E}\left[\left|\left[D_{z} m_{X \mid Z}\left(Z_{i}\right)^{\prime}\right]^{-}\right| f_{Z V}\left(Z_{i}, V_{i}\right) \mid P_{i}=0.5\right] f_{P}(0.5) .
$$

Observe that $b_{3 n}$ converges to a nondegenerate limiting distribution provided that the second moment of $\left|\left[D_{z} m_{X \mid Z}\left(Z_{i}\right)^{\prime}\right]^{-}\right| f_{Z V}\left(Z_{i}, V_{i}\right)$ exist. But this follows by elementwise square integrability in assumption 3 , together with the boundedness assumption 14. Hence,

$$
\left.c_{1} \sup _{z, v \in \mathcal{B} \times \mathcal{V}} \mid \nabla_{z} \widehat{p}(z, v)-\nabla_{z} p(z, v)\right) \mid b_{1 n}=o_{p}(1) .
$$

Similar arguments can be made for the second summand in (9.22), using the boundedness of the derivatives in assumption 14. Consequently, $\sqrt{n} T_{5 n}=o_{p}(1)$, implying that $\sqrt{n} T_{3 n}=$ $o_{p}(1)$.
Q.E.D.

## Appendix 2: Graphs and Tables

## Table A.1: Variables in Data Set

| 1 | dma | dma code, code for television market |
| :--- | :--- | :--- |
| 2 | income | household income in \$ |
| 3 | owncable | does household have cable TV |
| 4 | ownsat | does household have satellite TV |
| 5 | cableco | cable company |
| 6 | age | what range best describes your age |
| 7 | hhsize | household size |
| 8 | hhcomp | household composition |
| 9 | educ | education |
| 10 | hisp | hispanic or not |
| 11 | single | single or couple |
| 12 | state |  |
| 13 | rent | renter status (do they rent or own the house) |
| 14 | typeres | type of residence (house, apartment, condominium) |
| 15 | angle | dish angle |
| 16 | avgpbi | instrument, average price of basic cable across other cable franchises |
| 17 | avgppi | same for premium |
| 18 | tvsel1 | tv choice (1: basic cable, 2: premium cable, $0:$ nothighTV, 3 or $4:$ satellite) |
| 19 | yearst | year established (satellite dish) |
| 20 | chancap | channel capacity |
| 21 | airchan | number of over the air-channels available |
| 22 | paychan | number of pay channels available |
| 23 | othchan | other channels |
| 24 | ppv | pay per view available |
| 25 | cityff | city fixed fee (tax) |
| 26 | pricebe | price of basic cable |
| 27 | gender | gender |
| 28 | varelev | variance of the local terrain and the average elevation |

Table A.1: Variables in Data Set(cont.)

| 29 | mild | local weather index |
| :--- | :--- | :--- |
| 30 | bright | local weather index |
| 31 | stable | local weather index |
| 32 | climate | local weather index |
| 33 | twoway | cable franchise char - probably whether signals can be sent both ways |
| 34 | hboprice | HBO price |
| 35 | density | population density in an area (city density) |
| 36 | cnts | number of sampled households in that cable franchise market |
| 37 | poprank | city code (market area: necessary to merge with damachers, cable98) |

Table A. 2 Summary Statistics for Forrester Data

|  | Mean | Std. Dev. | $25 \%$ | $50 \%$ | $75 \%$ |
| :--- | ---: | ---: | ---: | ---: | ---: |
| Satellite | 0.10 | 0.30 | 0.00 | 0.00 | 0.00 |
| Cable | 0.72 | 0.45 | 0.00 | 1.00 | 1.00 |
| Household income in \$ | 57,366 | 28,642 | 32,500 | 55,000 | 87,500 |
| Rent | 0.22 | 0.42 | 0.00 | 0.00 | 0.00 |
| Single unit dwelling | 0.78 | 0.41 | 1.00 | 1.00 | 1.00 |
| Household size | 2.16 | 1.88 | 1.00 | 1.00 | 3.00 |
| Single | 0.18 | 0.38 | 0.00 | 0.00 | 0.00 |
| Age of HH | 50.59 | 15.42 | 39.00 | 49.00 | 61.00 |
| Education in years | 14.06 | 2.69 | 12.00 | 13.00 | 16.00 |

The education level corresponds to the mean education in a non-single household.

Table A. 3 Summary Statistics for Warren's Factbook Data

|  | Mean | Std. Dev | $25 \%$ | $50 \%$ | $75 \%$ |
| :--- | ---: | ---: | ---: | ---: | ---: |
| Monthly cable price in $\$$ | 25.45 | 8.39 | 20.88 | 24.43 | 29.95 |
| HBO price in $\$$ | 11.13 | 1.51 | 9.95 | 10.95 | 12.45 |
| Channel capacity | 65.36 | 17.44 | 54.00 | 62.00 | 78.00 |
| Pay-per-view available | 0.92 | 0.26 | 1.00 | 1.00 | 1.00 |
| Year franchise began | 1974.94 | 9.82 | 1971 | 1976 | 1982 |
| City franchise fee | 4.06 | 1.55 | 3.00 | 5.00 | 5.00 |
| Number of over-the-air channels | 11.46 | 3.38 | 8.00 | 12.00 | 14.00 |
| Observations | 132 |  |  |  |  |

Fig.1: Comparison of Distribution of Estimator for Centrality Parameter

> Heterogeneity Robust (Solid Line)
> Conditional Independence (Broken)
> Rivers Voung (Broken and Dotted)
> Oracle (Dotted)




$$
n=2500
$$

Kernel Density of Estimators in 400 Realizations of the DGP

Fig.2: Comparison of Distribution of Estimator for Centrality Parameter

> Heterogeneity Robust (Solid Line)
> Conditional Independence (Broken)
> Rivers Voung (Broken and Dotted)
> Oracle (Dotted)




$$
\mathrm{n}=7500
$$

Fig. 3: Comparison of Distribution of Estimators for Centrality Parameter

> Heterogeneity Robust (Solid Line)
> Conditional Independence (Broken)
> Rivers Voung (Broken and Dotted)
> Oracle (Dotted)





$$
n=15000
$$


[^0]:    *Brown University, Department of Economics, Robinson Hall \#302C, Providence, RI 02912, USA, email: stefan_hoderlein@yahoo.com. I have received helpful comments from Richard Blundell, Joel Horowitz, Simon Lee, Arthur Lewbel, Oliver Linton, Amil Petrin, Ed Vytlacil and from seminar participants in Brown, Boston College, Georgetown, Iowa, Maryland, Minnesota, NYU, Penn, Princeton, St. Gallen, Tilburg, Toulouse, and ESEM Milano. I am also grateful to Martin Hackmann for research assistance.

[^1]:    ${ }^{1}$ While the identification analysis proceeds on this level of generality, for the large sample theory we specify $\vartheta$ to be the mean regression.

[^2]:    ${ }^{2}$ This is in stark contrast to the exogenous binary choice model, where single index estimators only allow for very limited forms of heteroscedasticity (namely that the distribution of $U \mid X$ is only a function of the index $X^{\prime} \beta$ ), and only maximum score type estimators allow for heteroscedastic errors of general form (Manski (1975, 1985), Horowitz (1992)), but those do not achieve $\sqrt{n}$ rate of convergence.

[^3]:    ${ }^{3}$ See Hoderlein and Mammen (2007) for a related discussion in the case of a continuous dependent variable. As already mentioned, this result could be generalized to models of the form $Y^{*}=\phi(X, A)=m(X)+U$, with $U=\phi(X, A)-m(X)$ and $k_{U \mid X V}^{0.5}(X, V)=l(V)$, but due to the lack of relevance for applications we desist from discussing this more general case here.

[^4]:    ${ }^{4}$ The trimming becomes then dependent on estimated quantities. We skip the large sample theory of such an approach, because it adds little new insight and makes the analysis more involved. An interesting situation arises when the instruments are weak everywhere. We conjecture that we may derive a generalized inverse by some type of regularization, e.g. by constructing a matrix $\left[D_{z} \vartheta(Z)^{\prime}\right]^{*}$ that is analogous to, say Ridge regression. However, we do leave the explicit behavior of such a model for future research.

[^5]:    ${ }^{5}$ NFO was the largest custom market-research firm in the United States until it became part of the TNS Group in 2004.

[^6]:    ${ }^{6}$ Looking at households that have a TV allows to assume that local antenna forms the chosen alternative for those who neither declare to subscribe to cable nor to DBS.
    ${ }^{7}$ Typically only one cable company receives the right to serve a region as a result of a franchise agreement with a local government even though the household might not subscribe to cable.

[^7]:    ${ }^{8}$ We have performed $n=200$ bootstrap repetitions with replacement from the same data. Since the choice of bandwidth is not clear (we conjecture that a second order expansion type of analysis can be performed), we have settled for a slightly smaller bandwidth in the bootstrap replications, because this is a common devise to mitigate small sample bias in the construction of pointwise confidence bands in nonparametric regression.

