Endogeneity in Semiparametric Binary Random Coefficient Models

Stefan Hoderlein*

Brown University

First Draft: February 15, 2006

This Draft: October 14, 2008

Abstract

In this paper we consider the case of 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. As a particularly relevant example for a model where no semiparametric estimator has of yet been analyzed, we consider the binary random coefficients model with endogenous regressors. However, many of the arguments we make hold more generally in all endogenous binary choice models with heteroscedasticity. We focus on the estimation of a centrality parameter β, because even in random coefficient models usually an average effect and not the entire distribution of coefficients is of interest. We use a control function IV assumption to identify a centrality parameter that has the interpretation of a local average structural effect of the regressor on the latent variable, and establish identification based on the mean ratio of derivatives of two functions of the instruments. We propose an estimator based on sample counterparts, and discuss the large sample behavior. 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 IO demand data concludes this paper.

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

*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, Arthur Lewbel, Amil Petrin and from seminar participants in Brown, Georgetown, Iowa, Maryland, NYU, UPenn, Tilburg, Toulouse and ESEM Milano. I am particularly grateful to Amil Petrin for comments and support with the data, and to Martin Hackmann for research assistance.

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.

$$\begin{aligned} Y^* &= X'\beta + U \\ Y &= \mathbb{I}\{Y^* > 0\}, \end{aligned} \tag{1.1}$$

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, β 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 β 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 $\mathbb{E}[Y|X] = F_U(X'\beta)$, 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 some 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 relaxing these critical assumptions, while retaining a simple and interpretable structure. In particular, it establishes interpretation and constructive identification of a centrality parameter β under assumptions that are compatible with a heterogeneous population characterized by an unrestricted distribution of random coefficients. The identification is constructive in the sense that it can indeed be employed to yield a $\sqrt{n}$ consistent semi-parametric estimator for this centrality parameter. 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.

Main Identification Idea: Throughout this paper, we will be concerned with model (1.1). However, we will 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 interpretation of β when employing a sensible independence restrictions.

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

$$X = \vartheta(Z) + V, \tag{1.2}$$

where ϑ is a smooth, but unknown function of Z. For instance, ϑ could be a continuous version of the mean regression $\mathbb{E}[X|Z = z] = m_{X|Z}(z)$, in which case V would be the mean regression residuals, or it could also be a vector containing the conditional α quantiles of X^1 conditional on Z as first element (and $X^2, .., X^K$ as remainder elements of the vector)[1]. Now, if we let the conditional median of U given $Z = z$ and $V = v$ be denoted by $k^{0.5}_{U|ZV}(z, v)$, then our identifying assumption can be formulated as

$$k^{0.5}_{U|ZV}(z, v) = g(v),$$

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

What economic interpretation of β is implied by our assumptions? Taking the binary choice random coefficients model as an example, in the second section we establish the following: 1. If we are willing to assume conditionally symmetric random coefficients, we obtain that β has the interpretation of an average coefficient. 2. In the absence of symmetry we show that β has the interpretation of a local average structural derivative (see Hoderlein (2005, 2008) 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 we actually identify and estimate this parameter. To answer this question in a particularly simple example, consider the special case where out of K continuously distributed regressors $X^1, .., X^K$, only X^1 is endogenous and there is exactly one additional instrument denoted by Z^1 (and $Z = \left(Z^1, X^2.., X^K\right)'$). Finally, let $\bar{Y} = k^{0.5}_{Y|ZV}(Z, V) = \mathbb{I}\{\mathbb{P}[Y = 0|Z, V] < 0.5\}$ denote the conditional median of Y given Z and V, and assume that $\vartheta(z) = \mathbb{E}[X|Z = z]$. Then, under assumptions to be detailed below,

$$\beta = \mathbb{E}\left[\left[D_z\mathbb{E}[X|Z]'\right]^{-1} \nabla_z\mathbb{E}\left[\bar{Y}|Z\right] B(Z)\right], \tag{1.3}$$

[1] While the identification analysis proceeds on this level of generality, for the large sample theory we specify ϑ to be the mean regression.

where ∇_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.

Additional Contributions: The flexibility in 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 an average derivative type of 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. These direct estimators have several important advantages: they are transparent in structure, easy to compute and implement, and are robust to some forms of misspecification. While they have the drawback that they are not fully efficient, they can be taken as departure point for so-called "one-step efficient estimators". "Optimization", or M-, estimators for β, including semiparametric LS (Ichimura (1993)), semiparametric ML (Klein and Spady (1993)), and general M-estimators (Delecroix and Hristache (1997)) are usually efficient, but are rarely implemented in practise, because they lead to hard optimization problems in high-dimensional spaces. Neither class of estimators can handle 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 to a more standard limiting distribution.

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). 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. 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). 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: In the next section, we establish the economic foundations of our models in a heterogeneous population. In the case of a linear random coefficients model, we derive the median exclusion restriction formally, and establish the interpretation of β stated above. We then consider a stylized version of models in empirical industrial organization, and discuss the restrictive features and assumptions in this literature that our method allows to dispense with. In section three we state formally the assumptions required for identification of β 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 coefficients specification). This identification principle is constructive in the sense that it yields direct estimator 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 will occupy the sixth section. In the seventh section, we will apply our methods to a real world discrete choice demand application: We consider the decision to subscribe to cable TV, using data similar to those in Goolsbee and Petrin (2004). Finally, this paper ends with a conclusion and an outlook.

[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 $U|X$ is only a function of the index $X'\beta$), 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.

2 An Example: Binary Demand Decisions in a Heterogeneous Population

2.1 What is a Parameter of Interest 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, 2008) 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. $Y^* = \phi(X, A)$, 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 shall be denoted by A_2. In empirical IO for instance, A_2 are often omitted characteristics of the product.

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. A linear heterogeneous population with omitted variables A_2 may then be formalized through a random coefficient model, i.e.,

$$\begin{aligned} Y^* &= X'\beta(A_1) + A_2'\gamma(A_1) \\ Y &= \mathbb{I}\left\{Y^* > 0\right\}, \end{aligned} \tag{2.1}$$

where $\theta(A_1) = (\beta(A_1)', \gamma(A_1)')'$ 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, then so does $\theta(A_1)$. 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 β, and how can they be derived from restrictions in the structural model (2.1), and in particular on the random coefficients $\theta(A_1)$?

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. $A_1 \perp (X, Z, A_2)$. In discrete choice demand analysis in empirical industrial organization 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.

An unnecessarily strong, but economically plausible identifying independence restriction is the independence of instruments Z from all unobservables in the system, i.e., $Z \perp (A, V)$.

Continuing our discrete choice example, the empirical IO 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 $X \perp A_2|V$, and recall that our maintained hypothesis is that $A_1 \perp (X, V, A_2)$ which is implied by $A_1 \perp (X, Z, A_2)$.

The following result states that under these independence conditions, we can derive 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 $\mathfrak{B} = \beta(A_1)$, $\mathfrak{C} = A_2'\gamma(A_1)$ and $U = X'(\beta(A_1) - \beta) + A_2'\gamma(A_1)$. Since this section is motivational, the statement of the theorem is informal.

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 $A_1 \perp (X, Z, A_2)$ and $Z \perp (A, V)$ hold. Assume further that, conditional on (X, V): 1. $(\mathfrak{B}, \mathfrak{C})$ are jointly symmetrically distributed about $(\beta, \mathbb{E}[\mathfrak{C}|V])$, and 2. U is absolutely continuous distributed with respect to Lebesque measure. Finally, assume that regularity conditions hold such that all objects exist and are well defined. Then follows that $k^{0.5}_{U|ZV}(Z, V) = k^{0.5}_{U|XV}(X, V) = g(V)$ and*

$$\beta = \mathbb{E}[\beta(A_1)]$$

If we dispense with the conditional symmetry assumption, and start out with the median exclusion restriction $k^{0.5}_{U|ZV}(Z, V) = g(V)$, then we obtain that $k^{0.5}_{U|XV}(X, V) = g(V)$ and

$$\beta = \mathbb{E}\left[\beta(A_1)|X = x, V = v, Y^* = k^{0.5}_{Y^*|XV}(x, v)\right], \tag{2.2}$$

for all $(x, v) \in supp(X) \times supp(V)$.

Therefore, we conclude that our economically plausible independence assumption together with symmetry in the distribution of random components imply a median exclusion restriction and define a parameter β that is the mean of the distribution of random coefficients of the observable regressors. More generally, if we just assume the median exclusion restriction $k^{0.5}_{U|ZV}(Z, V) = g(V)$ but dispense with the symmetry assumption, we obtain that the coefficient β has the interpretation of a local average structural derivative (which is invariant to changes in x, v due to the linear random coefficient structure with exogenous A_1). Since we may integrate (2.2) over x, v, this implies that $\beta = \mathbb{E}_{XV}\left[\mathbb{E}\left[\beta(A_1)|X, V, Y^* = k^{0.5}_{Y^*|XV}(X, V)\right]\right]$, where $\mathbb{E}_{XV}[\cdot]$ denotes expectation over the distribution of (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 "average" person), then we may speak of β 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

$\beta(A_1)$, 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 $k^{\alpha}_{U|Z,V}(Z,V) = g(V)$, 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 β reduces to a completely standard object.

2.2 Implications of our Assumptions: The Example of a Binary Demand Decision in Empirical IO

In this subsection, we relate our model to the literature on random coefficient models in empirical industrial organization, because it serves as a nice example for the importance of accounting appropriately for unobserved heterogeneity. The scenario typically considered in this literature is often more complex than ours (for a comprehensive overview see Ackerberg, Benkard, Berry and Pakes (2006), ABBP for short). Most notably, it involves multivariate choice model, while we focus only on the binary case. Moreover, the specific model varies depending on the data at hand, which is often a combination of micro - and macrodata. Hence, our model can only be seen as a stylized version of models in that field, where the emphasis is on the way heterogeneity and endogeneity is treated in discrete choice models.

For this subsection only we adopt the notation of the discrete choice demand literature, to make the ties absolutely clear. In this literature it is commonly assumed that utility $\mathcal{U}_{ij}$ of individual consumer i derived by consumption of product j (out of sets $i = 1, .., n$ of consumers, and $j = 1, .., J$ of products) is given by

$$\mathcal{U}_{ij} = \upsilon(\mathcal{X}_{ij}, \xi_{ij}, \zeta_i, \nu_i, \beta),$$

where $\mathcal{X}_{ij}$, and ξ_{ij} represent observed and unobserved characteristics of the product, ζ_i and ν_i denote observed and unobserved characteristics of the consumer, and β is a finite dimensional nonrandom parameter characterizing all individuals' preference orderings, see ABBP. Again, this setup is stylized in the sense that the product characteristics may vary across i, so typical direct applications of individually varying unobserved product characteristics ξ_{ij} include advertisement or retail activities. However, the arguments are easily extended to setups where the quality of the goods varies across markets or time.

We will not discuss the model on this level of generality. Instead, we follow the route typically assumed in this literature, namely to consider a linear specification for υ on individual

[3]See Hoderlein and Mammen (2008) 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 $k^{0.5}_{U|XV}(X, V) = l(V)$, but due to the lack of relevance for applications we desist from discussing this more general case here.

level (but see Berry and Haile (2008) for nonparametric identification if υ is nonlinear). The model finally considered has the form

$$\mathcal{U}_{ij} = \mathcal{X}_{ij}'\theta_i + \xi_{ij}'\lambda_i + \varepsilon_{ij}, \tag{2.3}$$

where for the individual specific marginal effect of the k-th regressor $\mathcal{X}_{ijk}$, θ_{ik}, we have $\theta_{ik} = \beta_k^1 + \beta_k^{2\prime}\zeta_i + \beta_k^{3\prime}\nu_i$, where $\beta_k^1, \beta_k^2, \beta_k^3$ are subvectors of β_k. Typically, it is assumed that ξ_{ij} is only a random scalar, and utility is normalized by setting $\lambda_i = 1$. Moreover, ε_{ij} is assumed to be an error that is iid across alternatives j and individuals i, and is typically assumed to be extreme value distributed (i.e., the utility differences are logistically distributed). Endogeneity is believed to be an issue because there is correlation between the unobserved product characteristics ξ_{ij} and the own price which is contained in $\mathcal{X}_{ij}$. Summarizing, for $J = 2$, this model fits exactly into the framework put forward in the previous subsection: the random coefficients can be seen as function of unobservables but are generally believed to be exogenous, while there is a important (set of) omitted variables, in particular unobserved product characteristics that cause the own price to be endogenous. In our notation, ν_i corresponds to one of the factors in A_{1i} (entering the specification in a restrictive fashion), while ξ_{ij} corresponds to A_{2i}.

From an economic theory point of view, the parametric specification in model (2.3) has several shortcomings that our method allows to overcome: First, the assumption of the very existence of ε_{ij} contradicts the spirit of the characteristics based approach (see Berry and Pakes (2005) for a lucid discussion). Second, the usually invoked assumption of extreme value distributed errors has implausible economic implications, most notably it violates the independence of irrelevant alternatives principle in cases where there is no other source of randomness e.g., no random coefficients. Third, the assumption that there be a single unobservable product characteristic whose effect is similar across individuals appears very restrictive. Fourth, we do not require any parametric distribution assumption on the density of random coefficients like the commonly assumed (symmetric!) normal distribution. Fifth, our approach is direct, avoids computationally difficult and costly numerical inversions, and connects to recent developments in microeconometrics.

From now on, since we focus on a binary setting we drop the j subscript. A stylized specification that overcomes these shortcomings is

$$\mathcal{U}_i = X_i'\beta + \underbrace{\xi_i'\lambda_i + Q_i}_{U_i},$$

where $X_i = (\mathcal{X}' \otimes \zeta_i)'$, $Q_i = \sum_{k=1,\dots,K} Q_{ik}$, β isz for $k = 1, ..., K$, the collection of β_k^1 and β_k^2, and $Q_{ik} = \beta_k^{3\prime}\nu_i\mathcal{X}_k$. The random error U_i has now two components. An exogenous "random coefficient interaction term" part Q_i, and an endogenous "random coefficient in unobservables characteristics" part $\xi_i'\lambda_i$. A stylized binary version of an empirical IO model would hence have the reduced form

$$Y_i = \mathbb{I}\left\{X_i'\beta + U_i > 0\right\},$$

where obviously the composite error term U_i is heteroscedastic and endogenous in exactly the (complicated) fashion outlined in the previous section. As we have seen there, a median exclusion restriction arises naturally out of this structural decision model, and gives rise to a desired interpretation of β as average structural derivative. We are now going to establish how to actually identify and estimate β under these rather weak assumptions.

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 $m_{Y|ZV}(z, v)$ a continuous version of $\mathbb{E}\left[Y|Z = z, V = v\right]$, and let $f_A(a), f_{AB}(a, b)$ and $f_{A|B}(a; b)$ be the marginal, joint and conditional Radon-Nikodym density of the random vectors A and B with respect to some underlying measure μ, which may be the Lebesgue or the counting measure, (i.e., A may be discretely or continuously distributed). Define the nonparametric score $Q_z\left(v, z\right) = \nabla_z \log f_{V|Z}(v; z)$. Let $k^{\alpha}_{S|Z}(z)$ denote the conditional α-quantile of a random variable S given $Z = z$, i.e. for $\alpha \in (0, 1)$ $k^{\alpha}_{S|Z}(z)$ is defined by $\mathbb{P}(Y \leq k^{\alpha}_{S|Z}(z)|Z = z) = \alpha$. Let G^- denote the Moore-Penrose pseudo-inverse of a matrix G. Finally, let $c_k, k = 1, 2, ...$ denote generic constants, and note that we suppress the arguments of the functions whenever if it is convenient.

As already discussed in the introduction, the main idea is now that instead of running a regression using Y, we employ $\bar{Y} = k^{0.5}_{Y|ZV}(Z, V)$, 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 β. 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* $(Y_i, X_i, Z_i), i = 1, \dots, n$ *are independent and identically distributed such that* $(Y_i, X_i, Z_i) \sim (Y, X, Z) \in \mathcal{Y} \times \mathcal{X} \times \mathcal{Z} \subset \mathbb{R}^{1+K+L}$ *The joint distribution of* (Y, X, Z) *is absolutely continuous with respect to a* σ*-finite measure* μ *on* $\mathcal{Y} \times \mathcal{X} \times \mathcal{Z}$ *with Radon-Nikodym density* $f_{YXZ}(y, x, z)$*. The underlying measure* μ *can be written as* $\mu = \mu_{YX} \times \mu_Z$*, where* μ_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* $z \in Int(\mathcal{B})$*.* $[D_z \vartheta(z)']^-$ *exists and every element is bounded from below for all* $z \in \mathcal{B}$*.* $[D_z \vartheta(Z)']^-$ *is square integrable on* $\mathcal{B}$*.*

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

Assumption 5. $\mathbb{E}\left[\bar{Y}|Z = z\right] = m_{\bar{Y}|Z}(z)$ *is continuously differentiable in the components of* z *for all* $z \in Int(\mathcal{B})$*.* $D_z m_{\bar{Y}|Z}(Z)$ *is square integrable on* $\mathcal{B}$*.*

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.* $k^{0.5}_{U|ZV}(Z, V) = k^{0.5}_{U|XV}(X, V) = g(V)$*.*

2. Either of the following hold:

a. Let $\tilde{V} = l(V) = -(g(V) + V'\beta)$*. Then assume that* $\tilde{V}$ *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,* $f_{\tilde{V}}(D_z \vartheta(Z)'\beta)$ *is absolutely integrable on* $\mathcal{B}$*.*

b. There is one endogenous regressor X^k*, and* l *is a continuous piecewise invertible function. Moreover,* $f_{V|Z}(v, z)$ *and its partial derivatives wrt the components of* z *are bounded on* $\mathcal{B}$ *from below and above, i.e.* $c_1 > \sup_{(v,z) \in supp(V) \times \mathcal{B}} f_{V|Z}(v, z) \geq \inf_{(v,z) \in supp(V) \times \mathcal{B}} f_{V|Z}(v, z) = c_2 > 0$*, and* $c_3 > \sup_{(v,z) \in supp(V) \times \mathcal{B}} \left\| \nabla_z f_{V|Z}(v, z) \right\| \geq \inf_{(v,z) \in supp(V) \times \mathcal{B}} \left\| \nabla_z f_{V|Z}(v, z) \right\| =$

$c_4 > 0$. *Finally, let* $Q_z(V, Z)$ *be absolutely integrable on supp*$(V) \times \mathcal{B}$ *, and let* $\tau(z) = \mathbb{E}\left[\bar{Y} Q_z(V, Z) | Z = z\right]$ *be square integrable on* $\mathcal{B}$.

c. Let 8.3b 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* $g(v) = v'\gamma$, *with* $\gamma \in \mathbb{R}^{K_1}$.

Remark 3.1 - Discussion of Assumptions: Starting with assumption 1, while we may allow for discrete endogenous regressors we assume to possess continuously distributed instruments. Strictly speaking, we do not even require continuous instruments, but an estimator akin to Horowitz and Haerdle (1998) in the exogenous setting is beyond the scope of this paper. The *iid* assumption is inessential and may be relaxed to allow for some time series dependence. For 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 $[D_z\vartheta(z)']^-$ exists and is bounded on $\mathcal{B}$ (cf. assumption 3), and hence we choose $B(z) = \mathbb{I}\{z \in I_z\}\mathbb{I}\{\det|D_z\vartheta(z)D_z\vartheta(z)'| \geq b\}$, 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 $\det|D_z\vartheta(z)D_z\vartheta(z)'| \geq b$ 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 β 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 $U \perp X|V$) 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 8.2b–8.2c allow for arbitrary dependence between V and Z at the expense of placing some structure on l. In the case of a single endogenous regressor this structure is very general: indeed, any continuous and piecewise

[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 $[D_z\vartheta(Z)']^*$ that is analogous to, say Ridge regression. However, we do leave the explicit behavior of such a model for future research.

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 8.2b allows for l being a piecewise invertible function of an index).

3.1.2 Essential Arguments in the Identification of β in the Heteroscedastic Case

To see how assumptions 1–8 help in identifying β, rewrite the model as follows

$$Y = \mathbb{I}\left\{(\vartheta(Z) + V)'\beta + U > 0\right\}. \tag{3.1}$$

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

$$\bar{Y} = \mathbb{I}\left\{\vartheta(Z)'\beta + k^{0.5}_{U|ZV}(Z,V) + V'\beta > 0\right\} = \mathbb{I}\left\{\vartheta(Z)'\beta > l(V)\right\}, \tag{3.2}$$

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 β 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,

$$m_{\bar{Y}|Z}(z) = \mathbb{E}\left[\bar{Y}|Z = z\right] = \mathbb{P}\left\{\vartheta(z)'\beta > l(V)\right\} \tag{3.3}$$

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 $m_{\bar{Y}|Z}(z)$ and $\vartheta(z)$ are continuously differentiable in all components of Z, for all $z \in \mathcal{B}$, we obtain by the chain rule

$$\nabla_z m_{\bar{Y}|Z}(Z) = f_{l(V)}\left(\vartheta(Z)'\beta\right) D_z\vartheta(Z)'\beta, \tag{3.4}$$

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 $[D_z\vartheta(Z)']^-$, which exists on $\mathcal{B}$ due to assumption 3, and the weighting function $B(z)$ to obtain

$$[D_z\vartheta(Z)']^- \nabla_z m_{\bar{Y}|Z}(Z)B(Z) = \beta f_{l(V)}\left(\vartheta(Z)'\beta\right) B(Z), \tag{3.5}$$

or, upon taking expectations,

$$\beta c_1 = \mathbb{E}\left[[D_z\vartheta(Z)']^- \nabla_z \mathbb{E}\left[\bar{Y}|Z\right] B(Z)\right], \tag{3.6}$$

where $c = \mathbb{E}\left[f_{l(V)}\left(\vartheta(Z)'\beta\right) B(Z)\right]$. 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 β 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:

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* $\mathbb{E}\left[f_{\bar{V}|Z}\left(\vartheta(Z)'\beta;Z\right)B(Z)\right]=1$*. Then* β *is identified by relationship:*

$$\beta=\mathbb{E}\left[\left[D_z\vartheta(Z)'\right]^{-}\nabla_z\mathbb{E}\left[\bar{Y}|Z\right]B(Z)\right]. \tag{3.7}$$

(ii) If instead of assumption 8.2a either of assumptions 8.2b – 8.2c hold, then we obtain that β *is identified up to scale by*

$$\beta=\mathbb{E}\left[\left[D_z\vartheta(Z)'\right]^{-}\left\{\nabla_z\mathbb{E}\left[\bar{Y}|Z\right]-\mathbb{E}\left[\bar{Y}Q_z\left(V,Z\right)|Z\right]\right\}B(Z)\right], \tag{3.8}$$

where $Q_z\left(V,Z\right)$ *denotes the nonparametric score* $\nabla_z\log f_{V|Z}(V;Z)$*.*

(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), β *is (up to scale) identified by*

$$\mathbb{E}\left[\left[D_z\vartheta(Z)'\right]^{-}\left\{\nabla_z\mathbb{E}\left[Y|Z\right]-\mathbb{E}\left[YQ_z(V,Z)|Z\right]\right\}B(Z)\right], \tag{3.9}$$

as well as

$$\mathbb{E}\left[\left[D_z\vartheta(Z)'\right]^{-}\left\{\nabla_z\mathbb{E}\left[\breve{Y}|Z\right]-\mathbb{E}\left[\breve{Y}Q_z(V,Z)|Z\right]\right\}B(Z)\right], \tag{3.10}$$

where $\breve{Y}=\mathbb{E}\left[Y|Z,V\right]$*.*

Remark 3.2 - Interpretation of Theorem 2: First, consider the scenario where V and Z are independent which gives rise to (3.7). β is identified by a weighted average ratio of derivatives, involving the derivatives of the function ϑ, 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 ϑ. Allowing ϑ 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 by nonparametric estimators and the expectation by the average. While we always obtain a term of the form $\mathbb{E}\left[\left[D_z\vartheta(Z)'\right]^{-}\nabla_z\mathbb{E}\left[\bar{Y}|Z\right]B(Z)\right]$, note that in the more general case where V and Z are allowed to be dependent we obtain an additional correction term, i.e. $\mathbb{E}\left[\left[D_z\vartheta(Z)'\right]^{-}\mathbb{E}\left[\bar{Y}Q_z\left(V,Z\right)|Z\right]B(Z)\right]$, which accounts for the higher order dependence in the IV equation. The same applies in the full independence scenario, i.e., when $U \perp V|Z$.

Note that we have not ruled out discrete endogenous regressors by any assumption in this section. Indeed, all derivations in this section go through if for the endogenous regressor X_1, $X_1 = \mathbb{I}\{Z'\delta > W\}$ and $Z \perp W$, $\vartheta(Z) = \mathbb{E}[X|Z] = (F_W(Z'\delta), X_2, .., X_K)'$. In this case, we may think of $\vartheta(Z)$ as smoothed and exogenous version of X_1. Consequently, we may derive an estimator with the structure of (3.7), which allows for both discrete endogenous regressors and heteroscedasticity of U.

It is instructive to compare the heteroscedastic case with the case when $U \perp V|Z$. Observe that the independence assumption 7 implies assumption 8, so that equation (3.8) remains valid. But we obtain in addition that β 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 8.2a holds (i.e., $\tilde{V} \perp Z$), we could base an optimization estimator on equation (3.2), i.e.

$$\bar{Y} = \mathbb{I}\left\{\vartheta(Z)'\beta > \tilde{V})\right\}. \tag{4.1}$$

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.

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 something like 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 ϑ. Using results in Newey (1994), it is straightforward to see that for a $\sqrt{n}$ consistent estimator of β we require that $\mathbb{E}\left[\hat{\vartheta}\right] - \vartheta = o_p\left(n^{-1/2}\right)$ for the "no bias condition" to hold. In the kernel case this is, however, only possible under smoothness assumptions on ϑ, 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 β

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 $\mathbb{E}\left[\left[D_z\vartheta(Z)'\right]^{-}\mathbb{E}\left[\bar{Y}Q_z\left(V,Z\right)|Z\right]B(Z)\right]$ for future research. Moreover, throughout this section, we focus on the case when $\vartheta(z) = m_{X|Z}(z)$, i.e., ϑ is the nonparametric mean regression, and we leave the quantile regression for future work.

The first impression from looking at

$$\beta = \mathbb{E}\left[\left[D_z m_{X|Z}(Z)'\right]^{-}\nabla_z\mathbb{E}\left[\bar{Y}|Z\right]B(Z)\right]. \tag{4.2}$$

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,

$$\bar{Y} = k^{0.5}_{Y|ZV}(Z,V) = \mathbb{I}\left\{\mathbb{P}\left[Y=0|Z,V\right] < 0.5\right\},$$

and consequently, $\mathbb{E}\left[\bar{Y}|Z\right] = \mathbb{P}\left[\mathbb{P}\left[Y=0|Z,V\right] < 0.5|Z\right]$ This suggests estimating $\nabla_z\mathbb{E}\left[\bar{Y}|Z=z\right]$ 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., $\left[\sum_j \mathcal{K}_{hj}(z)\right]^{-1} \mathcal{K}_{hj}(z), \mathcal{K}_{hj}(z) = h^{-L}\mathcal{K}((Z_j - z)/h)$ and $\mathcal{K}((Z_j - z)/h) = \Pi_{l=1,..,L} K\left(\left(Z_j^l - z^l\right)/h\right)$ is a standard L-variate product kernel with standard univariate kernel function K. Moreover, $\widehat{P}_j$ denotes an estimator of $P_j = p(Z_j, V_j) = \mathbb{P}[Y_j = 0|Z_j, V_j]$, However, 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 version thereof. Specifically, let $\mathbb{K}(\xi) = \int_\xi^\infty K(t)dt$. Then, a straightforward sample counterpart estimator to $\beta = \mathbb{E}\left[\left[D_z m_{X|Z}(Z)'\right]^- \nabla_z \mathbb{E}\left[\bar{Y}|Z\right] B(Z)\right]$, looks as follows:

$$\hat{\beta}_H = n^{-1} \sum_i \left[D_z \hat{m}_{X|Z}(Z_i)'\right]^- \sum_{j\neq i} \nabla_z W_j(Z_i) \mathbb{K}\left\{\left(\widehat{P}_j - 0.5\right)/h\right\} B(Z_j), \tag{4.3}$$

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

$$\sqrt{n}\left(\hat{\beta}_H - \beta\right) \xrightarrow{\mathcal{D}} \mathcal{N}(0, \Sigma_H),$$

where Σ_H is defined as $\Sigma_H = \mathbb{E}\left(\sum_{k=1}^3 \sigma_k \sigma_k'\right) + 2\mathbb{E}(\sigma_2 \sigma_3') - \beta\beta'$, and

$$\begin{aligned} \sigma_1 &= \left[D_z m_{X|Z}(Z_i)'\right]^- \nabla_z m_{\bar{Y}|Z}(Z_i) B(Z_i), \\ \sigma_2 &= \left[D_z m_{X|Z}(Z_i)'\right]^- f_Z(Z_i)^{-1} \nabla_z f_Z(Z_i) V_i' \left[D_z m_{X|Z}(Z_i)'\right]^- \nabla_z m_{\bar{Y}|Z}(Z_i) B(Z_i), \\ \sigma_3 &= \left[D_z m_{X|Z}(Z_i)'\right]^- f_Z(Z_i)^{-1} \nabla_z f_Z(Z_i)\left(\bar{Y}_i - m_{\bar{Y}|Z}(Z_i)\right) B(Z_i). \end{aligned} \tag{4.4}$$

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

$$\begin{aligned} T_{1n} &= n^{-1} \sum_i \left[D_z \widehat{m}_{X|Z}(Z_i)'\right]^- \sum_{j\neq i} \nabla_z W_j(Z_i) \bar{Y}_j B(Z_j), \\ T_{2n} &= n^{-1} \sum_i \left[D_z \widehat{m}_{X|Z}(Z_i)'\right]^- \sum_{j\neq i} \nabla_z W_j(Z_i) \left[\mathbb{K}\{(P_j - 0.5)/h\} - \mathbb{I}\{P_j < 0.5\}\right] B(Z_j), \\ T_{3n} &= n^{-1} \sum_i \left[D_z \widehat{m}_{X|Z}(Z_i)'\right]^- \sum_{j\neq i} \nabla_z W_j(Z_i) \left[\mathbb{K}\left\{(\widehat{P}_j - 0.5)/h\right\} - \mathbb{K}\{(P_j - 0.5)/h\}\right] B(Z_j).. \end{aligned} \tag{4.5}$$

In this decomposition, T_{1n} 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

$$\hat{\beta}_1 = \frac{1}{n} \sum_i \left[D_z \widehat{m}_{X|Z}(Z_i)'\right]^- \nabla_z \widehat{m}_{Y|Z}(Z_i) B(Z_i). \tag{4.6}$$

Hence, we will first discuss the independence case. Then, we will give assumptions under which T_{2n} and T_{3n} will tend to zero faster than the leading term. Essentially, these conditions are higher order smoothness conditions on the conditional cdf $F_{P|Z}$ 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 β 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

$$\mu_k = \int u^k K(u) du \qquad \text{and} \qquad \kappa_k^2 = \int u^k K(u)^2 du.$$

In principle, we also have two bandwidths to consider, one in estimating $m_{X|Z}$, and one in estimating $m_{Y|Z}$. However, since the estimation problems are symmetric, Since (i.e., in particular both mean regressions share the same regressors and have thus the same dimensionality), we assume 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 $\mathbb{E}[X|Z=z]$, $\mathbb{E}[Y|Z=z]$ and $f_Z(z)$ of order $r+1$ exist for all $z \in \mathcal{B}$. Moreover, the expectations of $\left[D_z m_{X|Z}(Z)'\right]^- BY_l(Z)$ and $\left[D_z m_{X|Z}(Z)'\right]^- BX_l(Z) \left[D_z m_{X|Z}(Z)'\right]^- \nabla_z m_{Y|Z}(Z)$ exist for all $l = 1, ..., r,$, where BY_l (resp., BX_l) contains sums of products of all partial derivatives of $m_{Y|Z}$ and f_Z (resp. $m_{X|Z}$ 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. $\mu_k = \int u^k K(u) du = 0$ for all $k < r$ and $\int u^r K(u) du < \infty$).*

Assumption 11. *As $n \to \infty$, $h \to 0$, $nh^{L+2} \to \infty$ and $nh^{2r} \to 0$.*

The following theorem summarizes the results when is $\vartheta = m_{X|Z}$. 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* $\mathbb{E}\left[f_{U|V}\left(X'\beta; V\right) B(Z)\right] = 1$. *Then,*

$$\sqrt{n}\left(\hat{\beta}_1 - \beta\right) \xrightarrow{\mathcal{D}} \mathcal{N}(0, \Sigma_1)$$

where

$$\Sigma_1 = \mathbb{E}\left(\sum_{k=1}^{3} \sigma_k \sigma_k'\right) + 2\mathbb{E}\left(\sigma_2 \sigma_3'\right) - \beta\beta'$$

and

$$\begin{aligned}
\sigma_1 &= \left[D_z m_{X|Z}(Z_i)'\right]^- \nabla_z m_{Y|Z}(Z_i) B(Z_i) \\
\sigma_2 &= \left[D_z m_{X|Z}(Z_i)'\right]^- f_Z(Z_i)^{-1} \nabla_z f_Z(Z_i) V_i' \left[D_z m_{X|Z}(Z_i)'\right]^- \nabla_z m_{Y|Z}(Z_i) B(Z_i) \\
\sigma_3 &= \left[D_z m_{X|Z}(Z_i)'\right]^- f_Z(Z_i)^{-1} \nabla_z f_Z(Z_i) \left(Y_i - m_{Y|Z}(Z_i)\right) B(Z_i)
\end{aligned}$$

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 β, 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 Σ_1: Estimation of the variance components is straightforward by sample counterparts. For instance, an estimator for $\Sigma_1^{23} = \mathbb{E}\left(\sigma_2 \sigma_3'\right)$ is given by

$$\begin{aligned}
\widehat{\Sigma}_1^{23} &= n^{-1} \sum \hat{f}_Z(Z_i)^{-2} \left[D_z \hat{m}_{X|Z}(Z_i)'\right]^- \nabla_z \hat{f}_Z(Z_i) \left(Y_i - \hat{m}_{Y|Z}(Z_i)\right) \\
&\quad \times \left\{ \left[D_z \hat{m}_{X|Z}(Z_i)'\right]^- \nabla_z \hat{f}_Z(Z_i) \left(X_i - \hat{m}_{X|Z}(Z_i)\right)' \left[D_z \hat{m}_{X|Z}(Z_i)'\right]^- \nabla_z \hat{m}_{Y|Z}(Z_i) \right\}' B(Z_i).
\end{aligned}$$

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

$$\hat{\beta}_H = n^{-1} \sum_i \left[D_z \widehat{m}_{X|Z}(Z_i)'\right]^- \sum_{j \neq i} \nabla_z W_j(Z_i) \mathbb{K}\left\{\left(\widehat{P}_j - 0.5\right)/h\right\} B(Z_j),$$

Recall the decomposition $\hat{\beta}_H = T_{1n} + T_{2n} + T_{3n}$ in (4.5). The first term T_{1n} 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_{2n} and T_{3n} 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 $P_j = p(Z_j, V_j)$. To denote 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. $\int u^k K(u)du = 0$ for all $k < r$ and $\int u^r K(u)du < \infty$).*

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

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

Assumption 15. *As $n \to \infty$, $h_1, h_2 \to 0$, $nh_1, nh_2^{L+\dim(V)+2} \to \infty$ and $nh_1^{2r_1}, nh_2^{L+\dim(V)+4} \to 0$.*

Note that we require higher order smoothness conditions on $F_{P|Z}$ and f_Z that in connection with higher order kernels ensure that the bias terms $\sqrt{n}T_{2n}$ and $\sqrt{n}T_{3n}$ 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 $\mathbb{E}\left[f_{\bar{V}|Z}\left(m_{X|Z}(Z)'\beta; Z\right) B(Z)\right] = 1$ holds. Then, $\sqrt{n}T_{2n} = o_p(1)$, $\sqrt{n}T_{3n} = o_p(1)$, and $\sqrt{n}\left(\hat{\beta}_H - \beta\right) \xrightarrow{\mathcal{D}} \mathcal{N}(0, \Sigma_H)$, where Σ_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 Testing for Endogeneity

The first question that we can analyze within our framework is whether regressors are endogenous. There are a variety of options. As in Hoderlein (2005, 2008) and Hoderlein and Mammen

(2008), we may compare the regression $\mathbb{E}[Y|X]$ with the regression $\mathbb{E}[Y|X,V]$. 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 $\beta = \mathbb{E}[\nabla_x \mathbb{E}[Y|X] C(X)]$ (C is a again a bounded weighting function), and an estimator based on our identification principle (say, $\beta = \mathbb{E}\left[\left[D_z m_{X|Z}(Z)'\right]^- \nabla_z \mathbb{E}[Y|Z] B(Z)\right]$), 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 $\hat{\beta}_{Ex}$ denote a sample counterpart estimator to $\mathbb{E}[\nabla_x \mathbb{E}[Y|X] C(X)]$ like the PSS ADE, $\hat{\beta}_{End}$ any of the sample counterpart estimator to $\mathbb{E}\left[\left[D_z m_{X|Z}(Z)'\right]^- \nabla_z \mathbb{E}[Y|Z] B(Z)\right]$ defined below, $\widehat{\mathcal{B}} = (\hat{\beta}'_{Ex}, \hat{\beta}'_{End})$, and $G = (I, -I)$. Next, rewrite $H_0 : G'\mathcal{B} = 0$, and use the fact that $\widehat{\mathcal{B}} \xrightarrow{d} \mathcal{N}(0, \Sigma_E)$, where Σ_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, $\widehat{\Gamma}_1 = \left(G'\widehat{\mathcal{B}}\right)' \left[G\widehat{\Sigma}_E G'\right]^{-1} \left(G'\widehat{\mathcal{B}}\right)$ behaves asymptotically as follows:

$$\widehat{\Gamma}_1 = \left(G'\widehat{\mathcal{B}}\right)' \left[G\widehat{\Sigma}_E G'\right]^{-1} \left(G'\widehat{\mathcal{B}}\right) \xrightarrow{d} \chi^2_K. \tag{5.1}$$

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.

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

$$\beta = \mathbb{E}\left[\left[D_z m_{X|Z}(Z)'\right]^- \nabla_z \mathbb{E}\left[\bar{Y}|Z\right] B(Z)\right], \tag{5.2}$$

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

$$\beta = \mathbb{E}\left[\left[D_z m_{X|Z}(Z)'\right]^{-} \nabla_z \mathbb{E}\left[Y|Z\right] B(Z)\right],$$

will be inconsistent under heteroscedasticity. However, under H_0 of homoscedasticity, we have again that both estimators should vary only by sampling error. A straightforward test statistic is suggested by the following reformulation of H_0 :

$$0 = \mathbb{E}\left[\left[D_z m_{X|Z}(Z)'\right]^{-} \nabla_z \left(\mathbb{E}\left[\bar{Y} - Y|Z\right]\right) B(Z)\right] = \delta.$$

The theory of the obvious sample counterpart $\hat{\delta} = n^{-1} \sum \left[D_z \widehat{m}_{X|Z}(Z_i)'\right]^{-} \nabla_z \widehat{m}_{\bar{Y}-Y|Z}\left(Z_i\right) B(Z_i)$ is a corollary to theorem 4. Specifically, $\sqrt{n}\hat{\delta} \xrightarrow{\mathcal{D}} \mathcal{N}(0, \Sigma_\delta)$, where Σ_δ 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 δ is greater than zero, i.e.

$$\widehat{\Gamma}_{het} = \hat{\delta}'\hat{\Sigma}_\delta^{-1}\hat{\delta} \xrightarrow{\mathcal{D}} \chi^2_K,$$

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

5.3 Overidentification: Issue and Test

In the linear model, overidentification allows to delete instruments and recover β 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 β would be identified by taking the derivatives w.r.t. any subset of instruments Z_1 such that $Z = (Z_1', Z_{-1}')'$ and $\left[D_{z_1} m_{X|Z}(z) D_{z_1} m_{X|Z}(z)'\right]$ would be nonsingular for all $z \in \mathcal{B}$. by similar arguments as in theorem 2, the following result holds:

$$\beta = \mathbb{E}\left[\left[D_{z_1} m_{X|Z}(Z)'\right]^{-} \nabla_{z_1} \mathbb{E}\left[Y|Z\right] B(Z)\right]. \tag{5.3}$$

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 $\mathbb{E}\left[\mathbb{E}\left[Y|Z\right] Q_{z_1}|Z_1, V\right]$, where

$Q_{z_1} = \nabla_{z_1} \log f_{Z_{-1}|Z_1 V}(Z_-; Z_1, V)$, 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 $Z = (Z_1', Z_{-1}')'$ exist s.th. β 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 $H_0 : \beta^{(1)} = \beta^{(2)} = ... = \beta^{(M)}$. To this end, we determine the joint distribution of $\mathcal{B} = \left(\beta^{(1)\prime}, \beta^{(2)\prime}, ..., \beta^{(M)\prime}\right)'$. As a corollary from the large sample theory of this paper, $R'\mathcal{B} = 0$, $\widehat{\mathcal{B}} \xrightarrow{d} \mathcal{N}(0, \Sigma_I)$, where Σ_I is a covariance matrix with typical element Σ_{jk}. This element is given by

$$\Sigma_{jk} = \mathbb{E}\left(\sum_{l=1}^{3} \sigma_k^j \sigma_k^{k\prime}\right) + 2\mathbb{E}\left(\sigma_2^j \sigma_3^{k\prime}\right) - \beta\beta'$$

where for $h = j, k$.

$$\begin{aligned}
\sigma_1^h &= \left[D_{z_h} m_{X|Z}(Z_i)'\right]^- \nabla_{z_h} m_{Y|Z}(Z_i) B(Z_i) \\
\sigma_2^h &= \left[D_{z_h} m_{X|Z}(Z_i)'\right]^- f_Z(Z_i)^{-1} \nabla_{z_h} f_Z(Z_i) V_i' \left[D_{z_h} m_{X|Z}(Z_i)'\right]^- \nabla_{z_h} m_{Y|Z}(Z_i) B(Z_i) \\
\sigma_3^h &= \left[D_{z_h} m_{X|Z}(Z_i)'\right]^- f_Z(Z_i)^{-1} \nabla_{z_h} f_Z(Z_i) \left(Y_i - m_{Y|Z}(Z_i)\right) B(Z_i)
\end{aligned}$$

Then,

$$\widehat{\Gamma}_{OvId} = \left(R'\widehat{\mathcal{B}}\right)' \left[R\widehat{\Sigma}_I R'\right]^{-1} \left(R'\widehat{\mathcal{B}}\right) \xrightarrow{\mathcal{D}} \chi^2_{M-1},$$

by standard arguments.

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 even modest data sizes, and as theory predicts, we find that the mean square error reduces as the sample size increases, but we observe a moderate bias even with quite large sample sizes. However, we establish that our estimator is superior to semiparametric estimators that do not account for heterogeneity. As an example for an estimator that does not account for heterogeneity we consider 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, too.

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

$$\begin{aligned} Y_i &= \mathbb{I}\left\{\beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 X_{3i} + \beta_4 X_{4i} + \beta_5 X_{5i} + U_i > 0\right\}, \\ X_{1i} &= Z_{1i} + V_i, \quad i = 1, .., n, \end{aligned}$$

where $\beta = (1, 0.5, 0.5, 0.5, 0.5)'$ and the data (Y_i, X_i, Z_i, U_i), $i = 1, .., n$, 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

$$\begin{aligned} (\log(W_i), Z_i)' &\backsim \mathcal{N}(\mu, \Sigma), \\ \log(V_i)' &\backsim \mathcal{N}(0, 1), \end{aligned}$$

where

$$\mu = 0, \quad \Sigma = \begin{bmatrix} 2 & 1.5 & 0 & 0 & 0 & 0 \\ 1.5 & 2 & 1 & 1 & 1 & 1 \\ 0 & 1 & 2 & 1 & 1 & 1 \\ 0 & 1 & 1 & 2 & 1 & 1 \\ 0 & 1 & 1 & 1 & 2 & 1 \\ 0 & 1 & 1 & 1 & 1 & 2 \end{bmatrix},$$

and V_i is independent of $(\log(W_i), Z_i)'$. Observe that the W_i are in particular correlated with Z_{1i}. Next, the error U_i is defined through:

$$U_i = W_i - k^{0.5}_{W|Z}(Z_i) + V_i,$$

so that $k^{0.5}_{U|ZV}(Z_i, V_i) = V_i$. 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_{1i} instead of either the conditional median $k^{0.5}_{Y|ZV}(Z_i, V_i)$ (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_{1i} are obtained in the following way: We assume that the oracle has knowledge of the (unobservable) Y_i^*, and compute the conditional median $k^{0.5}_{Y^*|ZV}(Z_i, V_i)$, and set $Y_{1i} = \mathbb{I}\left\{k^{0.5}_{Y^*|ZV}(Z_i, V_i) > 0\right\}$. Bandwidths for all estimators are obtained by doing a grid search for the bandwidth that minimizes the MSE in 100 repetitions.

The result of applying our methods can be found in figures 1 - 3 in the graph 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 four estimated β (which in this case could have the local average structural effect interpretation, as mentioned in the introduction), 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 it with the distribution of the estimated β if we erroneously use the independence estimator (dotted line) and the oracle one (solid line).

Obviously, $\hat{\beta}_H$ is less biased than $\hat{\beta}_1$, but more so than the oracle estimator, denoted $\hat{\beta}_O$, while in terms of variance all three estimators are the same, see also the corresponding table 1 below. The fact that the variance is not significantly affected arises because due to the discreteness of the problem, the median estimator in both cases still uses all observations.

It is interesting to see how the estimator behaves as n varies. The heteroscedasticity robust estimator $\hat{\beta}_H$ significantly outperforms $\hat{\beta}_1$ at moderate sample sizes ($n = 2500$); for smaller sample sizes the advantage 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.3 and 4, who show the behavior with $n = 7500$ and $n = 15000$ observations. The bias of $\hat{\beta}_H$ starts to vanish, quite nicely visible in the bottom right panel of fig. 4. The same result is also obtained from the tables, cf tab 3-5 below.

The tables replicate the result from the figures. The heteroscedasiticy robust estimator outperforms the independence estimator, the difference becomes more pronounced with increasing sample size. This difference in reduction is due to the vanishing bias. The variance remains roughly comparable between all three estimators. The specific numerical results are the following: First, 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. 1-3. Note, however, that the reduction in bias is quite slow. It is instructive to compare the estimator with the Independence estimator, $\hat{\beta}_1$, and the Oracle estimator $\hat{\beta}_O$. For the former, 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 $\hat{\beta}_H$, the MSE is roughly 25 - 40 % above that of the heteroscedasticity one. In contrast, as was to be expected, the (infeasible) Oracle estimator $\hat{\beta}_O$.outperforms both estimators:

Coefficient	2	3	4	5
$n = 2500$	0.010615	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 estimators given the data size, while it is the bias that causes the differences. While the oracle estimator starts out 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. Hence we tentatively conclude that correcting for endogeneity may have a larger effect than modelling the heteroscedasticity structure in an unrestrictive fashion. Our application below, however, will make the importance of being less restrictive in this part of the model quite clear.

7 Application to Discrete Consumer Choice

7.1 Description of Data and Variables

As our motivation came in parts from the way heterogeneity is modelled in empirical industrial organization, 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). 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 for cable. Compared to cable, DBS provides a greater variety of channels and more pay per view options but bares the potential for signal interference and also 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 absorbed into an unidentified constant.

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

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

[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.

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).

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 bases 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.

7.2 Empirical Results

The focus in our empirical analysis is on the (endogenous) own price effect, and how the result is altered by introducing our method. The effect of household covariates is not of interested, and these variables act merely as controls. Hence we use 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 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 4: Linear Probability Model - 2SLS

[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.

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.

	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 5: 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 do obtain that the income effect is positive (and actually of as small 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:

[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, which is a common devise to mitigate small sample bias in the construction of pointwise confidence bands in nonparametric regression.

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

Here we see that the own price effect is significantly negative (again this effect is negative, and the income effect is positive in the non normalized data) At first glance, the results appears to be slightly different from Goolsbee and Petrin (2004), who find a relatively low 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 considers the consumption of cars). However, given that price of cable is a significant instrument also 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 heterogeneous individuals, in particular if the unobserved determinants and omitted variables are believed to be correlated with observables.

The most important class of such models that have been employed in applied work on discrete choices are random coefficient models. Most often, interest centers on average effects.

In this paper, we analyze the random coefficient model under a median exclusion restriction that defines such a (local) average effect. We show how to nonparametrically identify such an effect, and we propose a $\sqrt{n}$ consistent, asymptotically normal 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, which are also frequently encountered in practise, 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**, 933–964.

[2] Blundell, R. and J. Powell. 2004. Endogeneity in semiparametric binary response models, *Review of Economic Studies* **71**, 655–679.

[3] Blundell, R., and R. Smith. 1986. An Exogeneity Test for a Simultaneous Tobit Model, *Econometrica* **54**, 679—685.´

[4] Chaudhuri, P., K. Doksum and A. Samarov. 1997. On average derivative quantile regression, *Ann. Statist.* **25**, 715—744.

[5] Das, M., Newey, W. and F. Vella. 2003. Nonparametric estimation of sample selection models, *Review of Economic Studies* **70**, 33–58.

[6] Delecroix, M., and M. Hristache. 1999. M-estimateurs semi-paramétriques dans les modèles à direction révélatrice unique, *Bull. Belg. Maths. Soc.* **6**, 161–185.

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

[8] 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.

[9] Goolsbee, A. and A. Petrin, 2004. The Consumer Gains from Direct Broadcast Satellites and the Competition with Cable TV, *Econometrica*, **72**, 351-381.

[10] Heckman, J. and E. Vytlacil, 2005. Structural Equations, Treatment Effects, and Econometric Policy Evaluation, *Econometrica*, **73**, 669-738

[11] Horowitz, J. L. 1992, A smoothed maximum score estimator for the binary response model, *Econometrica* **60**, 505–531.

[12] Horowitz, J., and W., Härdle. 1996. Direct Semiparametric Estimation of Single Index Models with Discrete Covariates, *JASA* **91**, 1632–1640.

[13] 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.*

[14] Hristache, M., A. Juditsky, J. Polzehl, and V. Spokoiny. 2001. Structure Adaptive Approach for Dimension Reduction, *Ann. Statist.* **29**, 1537–1566.

[15] Ichimura, H. 1993, Semiparametric Least Squares (SLS) and Weighted SLS Estimation of Single-Index Models, *Journal of Econometrics* **58**, 71–120.

[16] Klein, R.W. and R.S. Spady. 1993, An Efficient Semiparametric Estimator of the Binary Response Model, *Econometrica* **61**, 387–422.

[17] Lewbel, A. 1998. Semiparametric Latent Variable Model Estimation with Endogenous or Mismeasured Regressors, *Econometrica*, **66**, 105-122,.

[18] Manski, C. F. 1975. Maximum score estimation of the stochastic utility model of choice, *Journal of Econometrics* **3**, 205–228.

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

[20] Powell, J. L., J. H. Stock, and T. M. Stoker. 1989. Semiparametric estimation of index coefficients, *Econometrica* **57**, 1403–1430.

[21] Stoker, T. M. 1986. Consistent Estimation of Scaled Coefficients, *Econometrica* **54**, 1461–1481.

[22] Vytlacil, E. and N. Yildiz, 2007. "Dummy Endogenous Variables in Weakly Separable Models," *Econometrica*, **75**, 757-779,

Appendix 1: Technical Proofs

Proof of the Identification Theorems

Proof of Theorem 1

To see the first statement, rewrite

$$Y = \mathbb{I}\left\{X'\beta(A_1) + A_2'\gamma(A_1) > 0\right\} = \mathbb{I}\{X'\beta + \underbrace{X'\left(\beta(A_1) - \beta\right) + A_2'\gamma(A_1)}_{U} > 0\}. \tag{8.1}$$

Then we obtain that

$$k^{0.5}_{Y|XV}(X,V) = \mathbb{I}\left\{k^{0.5}_{Y^*|XV}(X,V) > 0\right\} = \mathbb{I}\{X'\beta + k^{0.5}_{U|XV}(X,V) > 0\}$$

Next, note that due to $Z \perp (A,V) \implies X \perp A|V$, and thus $\mathbb{E}\left[\mathfrak{C}\,|X,V\right] = \mathbb{E}\left[\mathfrak{C}\,|V\right] = g(V)$ and $\mathbb{E}\left[\mathfrak{B}|X,V\right] = \mathbb{E}\left[\mathfrak{B}|V\right]$. Since $(X,V,A_2) \perp A_1 \implies V \perp \mathfrak{B}$, it moreover holds that $\mathbb{E}\left[\mathfrak{B}|V\right] = \mathbb{E}\left[\mathfrak{B}\right]$, and the condition that $(\mathfrak{B},\mathfrak{C})$ are jointly symmetrically distributed about $(\beta, \mathbb{E}\left[\mathfrak{C}|V\right])$ conditional on (V,X) holds also conditioning on V only. Consequently, $\beta = \mathbb{E}\left[\mathfrak{B}|V\right] = \mathbb{E}\left[\beta(A_1)\right]$, and β is the mean of the distribution of random coefficients. Moreover,

$$k^{0.5}_{U|XV}(X,V) = \mathbb{E}\left[X'\left(\mathfrak{B} - \beta\right) + \mathfrak{C}\,|X,V\right] = X'\mathbb{E}\left[\mathfrak{B} - \beta|V\right] + \mathbb{E}\left[\mathfrak{C}\,|V\right] = g(V),$$

and a very similar argument holds to show that $k^{0.5}_{U|ZV}(Z,V) = g(V)$ as well.

To see equation (2.2), observe first that $k^{0.5}_{U|ZV}(Z,V) = g(V)$ implies that $k^{0.5}_{U|XV}(X,V) = g(V)$. Start by using the definition of the median to obtain

$$\mathbb{P}(U \leq k^{\alpha}_{U|Z,V}(Z,V)|Z,V) = 0.5 = \mathbb{P}(U \leq k^{\alpha}_{U|X,V}(X,V)|X,V).$$

Taking conditional expectations with respect to (X,V) on both sides produces

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

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

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

implying that $k^{\alpha}_{U|X,V}(X,V) = g(V)$, provided U is continuously distributed.

Hence, if we assume the median exclusion restriction $k^{0.5}_{U|ZV}(Z,V) = g(V)$, we obtain that $\nabla_x k^{0.5}_{Y^*|XV}(X,V) = \beta$. Since $Y^* = X'\beta(A_1) + A_2'\gamma(A_1) = \phi(X,A)$, and $X \perp A|V$, we can apply Hoderlein and Mammen's (2007) theorem to obtain that the (constant) derivative has the following interpretation:

$$\beta = \mathbb{E}\left[\beta(A_1)|X = x, V = v, Y^* = k^{0.5}_{Y^*|XV}(x,v)\right], \tag{8.2}$$

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

Proof of Theorem 2

*Ad (i).*The case when $\tilde{V}$ is independent of Z is already discussed in the main text.

Ad (ii). Next, consider the case defined by assumptions 8.2b: 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 $\vartheta(z)'\beta \leq \max_{v \in S_1} l(v) = l(a)$. Then,

$$\begin{aligned} m_{\bar{Y}|Z}(z) &= \mathbb{P}\left[\vartheta(Z)'\beta > l(V)|Z = z\right] \\ &= \mathbb{P}\left[\vartheta(Z)'\beta > l(V), V \in S_1|Z = z\right] + \mathbb{P}\left[\vartheta(Z)'\beta > l(V), V \in S_2|Z = z\right] \\ &= \mathbb{P}\left[l_1\left(\vartheta(Z)'\beta\right) > V \wedge a|Z = z\right] + \mathbb{P}\left[l_2\left(\vartheta(Z)'\beta\right) > V \vee a|Z = z\right] \\ &= \int_{-\infty}^{l_1(\vartheta(z)'\beta)} f_{V|Z}(v|z)dv + \int_{l_2(\vartheta(z)'\beta)}^{\infty} f_{V|Z}(v|z)dv \end{aligned} \tag{8.3}$$

Taking derivatives by applying Leibnitz' rule produces

$$\begin{aligned} \nabla_z m_{\bar{Y}|Z}(z) &= D_z\vartheta(z)'\beta\left[\frac{\partial l_1}{\partial s}\left(\vartheta(z)'\beta\right) f_{V|Z}(l_1(\vartheta(z)'\beta)|z) - \frac{\partial l_2}{\partial s}\left(\vartheta(z)'\beta\right) f_{V|Z}(l_2(\vartheta(z)'\beta)|z)\right] \\ &\quad + \int_{-\infty}^{l_1(\vartheta(z)'\beta)} \nabla_z\left[\log f_{V|Z}(v|z)\right] f_{V|Z}(v|z)dv \\ &\quad + \int_{l_2(\vartheta(z)'\beta)}^{\infty} \nabla_z\left[\log f_{V|Z}(v|z)\right] f_{V|Z}(v|z)dv \\ &= D_z\vartheta(z)'\beta\left[\cdots\right] + \mathbb{E}\left\{\tilde{Y}Q_z\left(V;Z\right)|Z = z\right\}. \end{aligned} \tag{8.4}$$

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

Finally, consider the case defined by assumptions 8.2c. For simplicity, consider the two dimensional case: $V = (V_1, V_2)$, i.e., $l(v) = av_1 + bv_2$. Then,

$$\begin{aligned} m_{\bar{Y}|Z}(z) &= \mathbb{P}\left[\vartheta(Z)'\beta > aV_1 + bV_2|Z = z\right] \\ &= \mathbb{P}\left[V_1 < a^{-1}\left(\vartheta(Z)'\beta - bV_2\right)|Z = z\right] \\ &= \int_{-\infty}^{b^{-1}\vartheta(z)'\beta} \int_{-\infty}^{a^{-1}(\vartheta(z)'\beta - bv_2)} f_{V|Z}(v;z)dv_1dv_2 \end{aligned} \tag{8.5}$$

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*

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

Then,

$$\partial_x F(x) = \int_{-\infty}^{\alpha(x)} \int_{-\infty}^{\beta(x,v)} \partial_x f(u,v;x) du dv + \partial_x \alpha(x) \int_{-\infty}^{\beta(x,\alpha(x))} f(u,\alpha(x);x) du + \int_{-\infty}^{\alpha(x)} \partial_x \beta(x,v) f(\beta(x,v),v;x) dv.$$

Proof.

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

$$\begin{aligned} &= \int_{-\infty}^{\alpha(x)} \int_{-\infty}^{\beta(x,v)} \left[f + \partial_x f h \right] du dv + \int_{-\infty}^{\alpha(x)} \int_{\beta(x,v)}^{\beta(x,v)+\partial_x\beta(x,v)h} \left[f + \partial_x f h \right] du dv \\ &\quad + \int_{\alpha(x)}^{\alpha(x)+\partial_x\alpha(x)h} \int_{-\infty}^{\beta(x,v)} \left[f + \partial_x f h \right] du dv \\ &\quad + \int_{\alpha(x)}^{\alpha(x)+\partial_x\alpha(x)h} \int_{\beta(x,v)}^{\beta(x,v)+\partial_x\beta(x,v)h} \left[f + \partial_x f h \right] du dv + O(h^2) \\ &= F(x) + h \int_{-\infty}^{\alpha(x)} \int_{-\infty}^{\beta(x,v)} \partial_x f du dv + \int_{-\infty}^{\alpha(x)} \int_{\beta(x,v)}^{\beta(x,v)+\partial_x\beta(x,v)h} f du dv \\ &\quad + \int_{\alpha(x)}^{\alpha(x)+\partial_x\alpha(x)h} \int_{-\infty}^{\beta(x,v)} f du dv + O(h^2). \end{aligned}$$

Since

$$\lim_{h\to 0} h^{-1} \int_{\alpha(x)}^{\alpha(x)+\partial_x\alpha(x)h} \underbrace{\int_{-\infty}^{\beta(x,v)} f(u,v;x) du dv}_{B(x,v)} = \partial_x \alpha(x) B(x,\alpha(x))$$

and

$$\begin{aligned}
&\lim_{h\to 0} h^{-1}\int_{-\infty}^{\alpha(x)}\int_{\beta(x,v)}^{\beta(x,v)+\partial_x\beta(x,v)h} f(u,v;x)dudv \\
= &\int_{-\infty}^{\alpha(x)}\lim_{h\to 0} h^{-1}\int_{\beta(x,v)}^{\beta(x,v)+\partial_x\beta(x,v)h} f(u,v;x)dudv \\
= &\int_{-\infty}^{\alpha(x)} \partial_x\beta(x,v)f(\beta(x,v),v;x)dv,
\end{aligned}$$

the assertion follows. *Q.E.D.*

Adapting this result to our scenario produces

$$\begin{aligned}
\nabla_z m_{\bar{Y}|Z}(z) &= a^{-1}\beta\left[\int_{-\infty}^{b^{-1}\vartheta(z)'\beta} f_{V|Z}(a^{-1}\left(\vartheta(z)'\beta - bv_2\right), v_2; z)dv_2\right] \\
&+\int_{-\infty}^{b^{-1}\vartheta(z)'\beta}\int_{-\infty}^{a^{-1}(\vartheta(z)'\beta-bv_2)} \underbrace{\nabla_z\left[\ln f_{V|Z}(v_1,v_2;z)\right]}_{Q_z(v_1,v_2;z)} f_{V|Z}(v_1,v_2;z)dv_1dv_2
\end{aligned}$$

Finally, the 2nd term may be written

$$\begin{aligned}
&\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} 1_{\{-\infty\le v_1\le a^{-1}(\vartheta(z)'\beta-bv_2),-\infty\le v_2\le b^{-1}\vartheta(z)'\beta\}} Q_z(v_1,v_2;z)f_{V|Z}(v_1,v_2;z)dv_1dv_2 \quad (8.6)\\
= &\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} 1_{\{av_1+bv_2\le\vartheta(z)'\beta\}} Q_z(v_1,v_2;z)f_{V|Z}(v_1,v_2;z)dv_1dv_2 \\
= &\,\mathbb{E}\left[YQ_z(V;Z)|Z=z\right],
\end{aligned}$$

which shows the statement.

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

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

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 $\mathbb{E}\left[\breve{Y}|Z\right] = \mathbb{E}\left[Y|Z\right]$. Hence, $\mathbb{E}\left[Y|Z\right] = \mathbb{E}\left[F_{U|V}(\vartheta(Z)'\beta + V'\beta, V)|Z\right]$, and

$$\nabla_z m_{Y|Z}(z) = \mathbb{E}\left[f_{U|V}(\vartheta(Z)'\beta + V'\beta, V)|Z=z\right] D_z\vartheta(z)'\beta + \mathbb{E}\left[YQ_z(V,Z)|Z=z\right],$$

where $Q_z(v,z) = \nabla_z \log f_{V|Z}(v;z)$, 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 $G_i = D_z m_{X|Z}(Z_i)'$, $B_i = \nabla_z m_{Y|Z}(Z_i)$, $\hat{G}_i = D_z \hat{m}_{X|Z}(Z_i)'$ and $\hat{B}_i = \nabla_z \widehat{m}_{Y|Z}(Z_i)$. Tedious, but straightforward manipulations lead to

$$
\begin{aligned}
\hat{G}_i^- \hat{B}_i &= G_i^- B_i + G_i^- \left[(G_i - \hat{G}_i) G_i^- B_i + (\hat{B}_i - B_i) \right] \\
&\quad + G_i^- (G_i - \hat{G}_i) \left(G_i^- - \hat{G}_i^- \right) B_i \\
&\quad + G_i^- (G_i - \hat{G}_i) G_i^- (\hat{B}_i - B_i) \\
&\quad + G_i^- (G_i - \hat{G}_i) \left(G_i^- - \hat{G}_i^- \right) (\hat{B}_i - B_i).
\end{aligned} \tag{8.7}
$$

Now, in (8.7) 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 ϑ is a mean regression. Specifically, we show in **Step 1a** that

$$
\tau_{1n} = n^{-1} \sum_i G_i^- B_i + G_i^- \left[(G_i - \hat{G}_i) G_i^- B_i + (\hat{B}_i - B_i) \right] = S_{1n} + S_{2n}
$$

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_{1n} may be handled using projection arguments coming from U-statistic theory, while in **Step 1c** we show that the bias term S_{2n}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 (8.7), i.e., the behavior of terms three to five. In **Step 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

$$
\begin{aligned}
\tau_{1n} &= n^{-1} \sum_i \left\{ G_i^- B_i + G_i^- (G_i - \hat{G}_i) G_i^- B_i + G_i^- (\hat{B}_i - B_i) \right\} && (8.8) \\
&= n^{-1} \sum_i G_i^- B_i + n^{-1} \sum_i G_i^- (G_i - \hat{G}_i) G_i^- B_i + n^{-1} \sum_i G_i^- (\hat{B}_i - B_i). && (8.9)
\end{aligned}
$$

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 (8.8) first. In the case where $\hat{G}_i$ is a nonparametric Nadaraya Watson derivative estimator, it rewrites as

$$\left(\sum_{j\neq i}\mathcal{K}_{hj}(Z_i)\right)^{-1}\left[\sum_{j\neq i}\nabla_z\mathcal{K}_{hj}(Z_i)X_j' - \left(\sum_{j\neq i}\mathcal{K}_{hj}(Z_i)\right)^{-1}\sum_{j\neq i}\nabla_z\mathcal{K}_{hj}(Z_i)\sum_{j\neq i}\mathcal{K}_{hj}(Z_i)X_j'\right]. \tag{8.10}$$

Hence, $G_i - \hat{G}_i$ has a representation as

$$D_z m_{X|Z}(Z_i) - \left(\sum_{j\neq i}\mathcal{K}_{hj}(Z_i)\right)^{-1}\left[\sum_{j\neq i}\left[\nabla_z\mathcal{K}_{hj}(Z_i) - W_n(Z_i)\mathcal{K}_{hj}(Z_i)\right]\left[V_j' + m_{X|Z}(Z_j)'\right]\right]$$

where

$$W_n(Z_i) = \left(\sum_{s\neq i}\mathcal{K}_{hs}(Z_i)\right)^{-1}\sum_{j\neq i}\nabla_z\mathcal{K}_{hs}(Z_i).$$

Separate this expressions into the two parts, where

$$\begin{aligned} -P_{1i} &= \left(\sum_{j\neq i}\mathcal{K}_{hj}(Z_i)\right)^{-1}\left[\sum_{j\neq i}\left[\nabla_z\mathcal{K}_{hj}(Z_i) - W_n(Z_i)\mathcal{K}_{hj}(Z_i)\right]V_j'\right] \qquad (8.11)\\ &= (n-1)^{-1}\sum_{j\neq i}\mathcal{W}_{jn}(Z_i)V_j' \end{aligned}$$

where $\mathcal{W}_{jn}(Z_i) = \left((n-1)^{-1}\sum_{j\neq i}\mathcal{K}_{hj}(Z_i)\right)^{-1}\left[\nabla_z\mathcal{K}_{hj}(Z_i) - W_n(Z_i)\mathcal{K}_{hj}(Z_i)\right]$ and

$$\begin{aligned} P_{2i} &= D_z m_{X|Z}(Z_i)' - \left(\sum_{j\neq i}\mathcal{K}_{hj}(Z_i)\right)^{-1}\left[\sum_{j\neq i}\left[\nabla_z\mathcal{K}_{hj}(Z_i) - W_n(Z_i)\mathcal{K}_{hj}(Z_i)\right]m_{X|Z}(Z_j)'\right] \qquad (8.12)\\ &= D_z m_{X|Z}(Z_i)' - (n-1)^{-1}\sum_{j\neq i}\mathcal{W}_{jn}(Z_i)m_{X|Z}(Z_j)' \end{aligned}$$

Note that $\mathcal{W}_{jn}(Z_i) = -\mathcal{W}_{in}(Z_j)$ by the symmetry of the kernel. The first part, (8.11), 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

$$\begin{aligned} n^{-1}\sum_i G_i^-(G_i - \hat{G}_i)G_i^-B_i &= -\left(n\left(n-1\right)\right)^{-1}\sum_i\sum_{j\neq i}G_i^-\mathcal{W}_{jn}(Z_i)V_j'G_i^-B_i\\ &\quad + \left(n\left(n-1\right)\right)^{-1}\sum_i G_i^-P_{2i}G_i^-B_i\\ &= S_{1n}^2 + S_{2n}^2, \end{aligned}$$

where the superscript 2 denotes the second term in the expression (8.8). A similar decomposition may be performed on $n^{-1}\sum_i G_i^-(\hat{B}_i - B_i) = \left(n\left(n-1\right)\right)^{-1}\sum_i\sum_{j\neq i}G_i^-\mathcal{W}_{jn}(Z_i)Q_j +$ $\left(n\left(n-1\right)\right)^{-1}\sum_i G_i^-P_{4i} = S_{1n}^3 + S_{2n}^3$, where $Q_i = Y_i - m_{Y|Z}(Z_i)$ and P_{4i} denotes again bias terms in the regression of Y on Z. In total, we obtain that

$$\tau_{1n} = n^{-1}\sum_i G_i^-B_i + S_{1n}^2 + S_{1n}^3 + S_{2n}^2 + S_{2n}^3 = S_{1n} + S_{2n}, \tag{8.13}$$

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

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

$$\begin{aligned} U_n &= (n(n-1))^{-1}\sum_i\sum_{j>i}\left\{G_i^-\mathcal{W}_{jn}(Z_i)V_j' G_i^- B_i - G_j^-\mathcal{W}_{jn}(Z_i)V_i' G_j^- B_j\right\} \\ &= (n(n-1))^{-1}\sum_i\sum_{j>i} p_n(S_i,S_j), \end{aligned}$$

where $S_i = (Y_i, X_i', Z_i')'$, with p_n symmetric, and we made use of $\mathcal{W}_{jn}(Z_i) = -\mathcal{W}_{in}(Z_j)$. To apply Lemma 3.1 of PSS which yields $\sqrt{n}\left(\hat{U}_n - U_n\right) = o_p(1)$, where

$$\hat{U}_n = \theta + n^{-1}\sum_i \mathbb{E}\left[p_n(S_i,S_j)|S_i\right], \tag{8.14}$$

we require that $\mathbb{E}\left(\|p_n(S_i,S_j)\|^2\right) = o(n)$. Following similar and straightforward, but more tedious arguments as in PSS, this is the case provided $nh^{L+2}\to\infty$. To analyze (8.14), note first that $\theta = \mathbb{E}\left[p_n(S_i,S_j)\right] = 0$, and consider first p_n^* which equals p_n save that in $\mathcal{W}_{jn}(Z_i)$, $(n-1)^{-1}\sum_{s\neq i}\mathcal{K}_{hs}(Z_i)$ and $(n-1)^{-1}\sum_{s\neq i}\nabla_z\mathcal{K}_{hs}(Z_i)$ are replaced with their probability limits, $f_Z(Z_i)$.and $\nabla_z f_Z(Z_i)$. Then,

$$\begin{aligned} &\mathbb{E}\left[p_n^*(S_i,S_j)|S_i = s_i\right] \\ =& \int h^{-(L+1)}(D_z m_{X|Z}(z_i)^- f_Z(z_i)^{-1}\left(\nabla_z\mathcal{K}((z_i - z)/h) - \nabla_z f_Z(z_i) f_Z(z_i)^{-1} h\mathcal{K}((z_i-z)/h)\right) \\ &\times v_i' D_z m_{X|Z}(z_i)^-\nabla_z m_{Y|Z}(z_i) f_Z(z)dz \\ =& D_z m_{X|Z}(z_i)^- f_Z(z_i)^{-1}\int h^{-1}\nabla_z\mathcal{K}(\psi) f_Z(z_i+\psi h)d\psi v_i' D_z m_{X|Z}(z_i)^-\nabla_z m_{Y|Z}(z_i) \qquad (8.15) \\ &-D_z m_{X|Z}(z_i)^- f_Z(z_i)^{-2}\nabla_z f_Z(z_i)\int\mathcal{K}(\psi) f_Z(z_i+\psi h)d\psi v_i' D_z m_{X|Z}(z_i)^-\nabla_z m_{Y|Z}(z_i) \\ =& -D_z m_{X|Z}(z_i)^- f_Z(z_i)^{-1}\nabla_z f_Z(z_i) v_i' D_z m_{X|Z}(z_i)^-\nabla_z m_{Y|Z}(z_i) + \eta_{2i} \\ =& -g_i^- f_Z(z_i)^{-1}\nabla_z f_Z(z_i) v_i' g_i^- b_i + \eta_{2i}, \end{aligned}$$

where η_{2i} denotes higher order terms, for which, by standard arguments $n^{-1/2}\sum_i\eta_{2i} = o_p(1)$ (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.,

$$n^{-1/2}\sum_i\mathbb{E}\left[p_n(S_i,S_j) - p_n^*(S_i,S_j)|S_i\right] = o_p(1).$$

To see this, consider a typical expression in $\mathbb{E}\left[p_n(S_i, S_j) - p_n^*(S_i, S_j)|S_i\right]$. Using the right hand side of the third equality in equation (8.15),

$$\begin{aligned}\rho_{ni} &= D_z m_{X|Z}(z_i)^- \left\{\hat{f}_Z(z_i)^{-1} - f_Z(z_i)^{-1}\right\} \nabla_z f_Z(z_i) v_i' D_z m_{X|Z}(z_i)^- \nabla_z m_{Y|Z}(z_i) \\ &= \left\{f_Z(z_i) - \hat{f}_Z(z_i)\right\} \hat{f}_Z(z_i)^{-1} D_z m_{X|Z}(z_i)^- f_Z(z_i)^{-1} \nabla_z f_Z(z_i) v_i' D_z m_{X|Z}(z_i)^- \nabla_z m_{Y|Z}(z_i),\end{aligned}$$

where $\hat{f}_Z(z_i) = (n-1)^{-1} \sum_{s \neq i} \mathcal{K}_{hs}(Z_i)$. Next, write

$$n^{-1/2} \sum_i \rho_{ni} = n^{1/2} \int \frac{f_Z(z) - \hat{f}_Z(z)}{\hat{f}_Z(z)} \chi(z, v) \hat{F}_{ZV}(dz, dv), \tag{8.16}$$

where $\chi(z,v) = D_z m_{X|Z}(z)^- f_Z(z)^{-1} \nabla_z f_Z(z) v' D_z m_{X|Z}(z)^- \nabla_z m_{Y|Z}(z)$, and $\hat{F}_{ZV}$ denotes the empirical cdf. Considering the denominator in (8.16), observe that

$$\frac{1}{\left|f_Z(z) + \hat{f}_Z(z) - f_Z(z)\right|} \leq \frac{1}{|f_Z(z)| - \left|\hat{f}_Z(z) - f_Z(z)\right|} \leq \frac{2}{b}, \tag{8.17}$$

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, $n^{-1/2} \sum_i \rho_{ni}$ is, in absolute value, bounded by

$$c \sup_{z \in \mathcal{B}} \left|f_Z(z) - \hat{f}_Z(z)\right| n^{-1/2} \sum_i \left|\chi(Z_i, V_i)\right|, \tag{8.18}$$

But since $n^{-1/2} \sum_i |\chi(Z_i, V_i)|$ converges by a standard CLT for *iid* random variables to a normal limit (provided the second moment are finite which we tacitly assume), and $\sup_{z \in \mathcal{B}} \left|f_Z(z) - \hat{f}_Z(z)\right| = O_p\left(h^{2r} + \left(nh^L\right)^{-1/2} \log n\right) = o_p(1)$ under general conditions, it follows that $n^{-1/2} \sum_i \rho_{ni} = o_p(1)$. Similar arguments can be applied to any other term appearing in $\mathbb{E}\left[p_n(S_i, S_j) - p_n^*(S_i, S_j)|S_i\right]$, implying that the difference vanishes.

Repeating the same arguments as from the start of Step 1b, we can show that $G_i^-(\hat{B}_i - B_i) = G_i^- f_Z(Z_i)^{-1} \nabla_z f_Z(Z_i) Q_i + \eta_{3n}$, where $Q_i = Y_i - m_{Y|Z}(Z_i)$. Returning to (8.13)

$$\begin{aligned}S_{1n} &= n^{-1} \sum_i \left\{G_i^- B_i + G_i^- \left[(G_i - \hat{G}_i) G_i^- B_i + (\hat{B}_i - B_i)\right]\right\} \\ &= n^{-1} \sum_i \left\{G_i^- B_i - G_i^- f_Z(Z_i)^{-1} \nabla_z f_Z(Z_i) V_i' G_i^- B_i - G_i^- f_Z(Z_i)^{-1} \nabla_z f_Z(Z_i) Q_i\right\} + n^{-1} \sum_i T_{3i},\end{aligned}$$

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

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

$$\begin{aligned} S_{2n}^2 &= n^{-1/2}\sum_i G_i^- \left\{ D_z m_{X|Z}(Z_i)' - (n-1)^{-1}\sum_{j\neq i}\mathcal{W}_{jn}(Z_i)m_{X|Z}(Z_j)' \right\} G_i^- B_i \\ &= \sqrt{n}\int\int \left[D_z m_{X|Z}(\zeta)'\right]^- \\ &\quad \times \left\{ D_z m_{X|Z}(\zeta)' - \left[\frac{h^{-L-1}\nabla_z \mathcal{K}\left((z-\zeta)/h\right)}{\hat{f}_Z(\zeta)} - \frac{\nabla_z \hat{f}_Z(\zeta)h^{-L}\mathcal{K}\left((z-\zeta)/h\right)}{\left(\hat{f}_Z(\zeta)\right)^2} \right] m_{X|Z}(z)' \right\} \\ &\quad \times \left[D_z m_{X|Z}(\zeta)'\right]^- \nabla_z m_{Y|Z}(\zeta)\hat{F}_Z(dz)\hat{F}_Z(d\zeta). \end{aligned}$$

Next, let $S_{2n}^2 = A^1 + \omega_n$ where A^1 equals S_{2n}^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 (8.16) and (8.17), 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

$$\begin{aligned} &\sqrt{n}\int \left[D_z m_{X|Z}(\zeta)'\right]^- \{D_z m_{X|Z}(\zeta)' - \\ &\times \int \frac{h^{-1}\nabla_\psi \mathcal{K}(\psi)}{f_Z(\zeta)} m_{X|Z}(\psi h+\zeta)' f_Z(\psi h+\zeta)d\psi - \int \frac{\nabla_z f_Z(\zeta)\mathcal{K}(\psi)}{\left(f_Z(\zeta)\right)^2} m_{X|Z}(\psi h+\zeta)' f_Z(\psi h+\zeta)d\psi\} \\ &\times \left[D_z m_{X|Z}(\zeta)'\right]^- \nabla_z m_{Y|Z}(\zeta) f_Z(\zeta)d\zeta. \end{aligned}$$

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

$$A^1 = \sqrt{n}\int \left[D_z m_{X|Z}(\zeta)'\right]^- BX(\zeta)\left[D_z m_{X|Z}(\zeta)'\right]^- \nabla_z m_{Y|Z}(\zeta) f_Z(\zeta)d\zeta + O(\sqrt{n}h^r). \quad (8.19)$$

where denotes higher order bias terms, i.e. $BX(\zeta) = \sum_{l=1..,r}\mu_k h^l BX_l(\zeta)$, and $BX_l(\zeta)$ contains sums of products of all higher order derivatives of $m_{X|Z}$ 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 $r = 2L$ in connection with assumption 11. Consequently, $\sqrt{n}S_{2n}^2 = o_p(1)$. Under samilar conditions on $BY(\zeta)$ (cf. assumption 9), and by similar arguments $\sqrt{n}S_{3n}^2 = o_p(1)$ and hence the bias expression proves asymptotically negligible under our assumptions.

Step 1d: Finally, the first terms provide us with the variance. Since $\mathbb{E}[\,(V_i', Q_i)'\,|Z_i] = 0$, σ_{1i} is uncorrelated with σ_{2i} and σ_{3i}. 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 $B_i = \nabla_z m_{Y|Z}(Z_i)$ and consider

$$\begin{aligned} & n^{-1/2} \sum_i G_i^-(G_i - \hat{G}_i)(\hat{B}_i - B_i) \\ = \;& n^{1/2} \int G_i^- \left(m_{X|Z}(z)' - D_z \widehat{m}_{X|Z}(z)'\right) \left(\nabla_z \widehat{m}_{Y|Z}(z) - \nabla_z m_{Y|Z}(z)\right) \hat{F}_Z(dz). \end{aligned}$$

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

$$n^{1/2} \sup_{z \in \mathcal{B}} \left|D_z m_{X|Z}(z)' - D_z \widehat{m}_{X|Z}(z)'\right| \sup_{z \in \mathcal{B}} \left|\nabla_z \widehat{m}_{Y|Z}(z) - \nabla_z m_{Y|Z}(z)\right| \underbrace{n^{1/2} \int \left|G_i^-\right| \hat{F}_Z(dz)}_{C_n}$$

Since $\sup_{z \in \mathcal{B}} \left|D_z m_{X|Z}(z)' - D_z \widehat{m}_{X|Z}(Z)'\right| \sup_{z \in \mathcal{B}} \left|\nabla_z \widehat{m}_{Y|Z}(z) - \nabla_z m_{Y|Z}(z)\right| = O_p(h^{2r} + \left(nh^{L+2}\right)^{-1} \ln(n))$ 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_{2n} and T_{3n} vanish, consider first T_{2n} in Step 3a, and then T_{3n} in Step 3b.

Step 3a: Recall that

$$T_{2n} = n^{-1} \sum_i \left[D_z \widehat{m}_{X|Z}(Z_i)'\right]^- \sum_{j \neq i} \nabla_z W_j(Z_i) \left[\mathbb{K}\left\{(P_j - 0.5)/h\right\} - \mathbb{I}\left\{P_j < 0.5\right\}\right] B(Z_j).$$

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

$$T_{2n}^* = n^{-1} \sum_i \left[D_z m_{X|Z}(Z_i)'\right]^- \sum_{j \neq i} \nabla_z W_j(Z_i) \left[\mathbb{K}\left\{(P_j - 0.5)/h\right\} - \mathbb{I}\left\{P_j < 0.5\right\}\right] B(Z_j),$$

and note that $T_{2n} = T_{2n}^* + R_n$, where R_n contains the difference $\left[D_z \widehat{m}_{X|Z}(Z_i)'\right]^- - \left[D_z m_{X|Z}(Z_i)'\right]^-$ instead of $\left[D_z m_{X|Z}(Z_i)'\right]^-$. 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 1b above, safe for the fact that Y_j is replaced by $\mathbb{K}\left\{(P_j - 0.5)/h\right\} - \mathbb{I}\left\{P_j < 0.5\right\}$. Following the same argumentation as the one in Step1c, we arrive at the crucial decomposition $T_{2n}^{*}=T_{2n}^{**}+\varrho_n$, where ϱ_n are terms that converge faster by Glivenko-Cantelli and compact support $\mathcal{B}$, and T_{2n}^{**} is defined as follows:

$$
\begin{aligned}
& T_{2n}^{**} \\
= & \int\int\int h^{-(L+1)}(D_z m_{X|Z}(\zeta)^{-} f_Z(\zeta)^{-1}\left(\nabla_\zeta \mathcal{K}((z-\zeta)/h)-\nabla_z f_Z(\zeta) f_Z(\zeta)^{-1} h\mathcal{K}((\zeta-z)/h)\right) \\
& \times\left[\mathbb{K}\{(p-0.5)/h\}-\mathbb{I}\{p<0.5\}\right] D_z m_{X|Z}(\zeta)^{-}\nabla_z m_{Y|Z}(\zeta) F_{PZ}(dp,dz) F_Z(d\zeta) \\
= & \int g(\zeta)\int\int h^{-1}\nabla_\psi \mathcal{K}(\psi)\left[\mathbb{K}\{(p-0.5)/h\}-\mathbb{I}\{p<0.5\}\right] f_Z(\zeta+\psi h) F_{P|Z}(dp;\zeta+\psi h) d\psi \times \\
& D_z m_{X|Z}(\zeta)^{-}\nabla_z m_{Y|Z}(\zeta)-g(\zeta)\nabla_z f_Z(\zeta)\int\int \mathcal{K}(\psi)\left[\mathbb{K}\{(p-0.5)/h\}-\mathbb{I}\{p<0.5\}\right]\times \\
& f_Z(\zeta+\psi h) F_{P|Z}(dp;\zeta+\psi h) d\psi g(\zeta)\nabla_z m_{Y|Z}(\zeta) F_Z(d\zeta) \\
= & Q_{1n}-Q_{2n},
\end{aligned}
$$

where $g(\zeta)=D_z m_{X|Z}(\zeta)^{-} f_Z(\zeta)^{-1}$. Next, consider the inner integral in Q_{1n}:

$$
\begin{aligned}
& h^{-1}\int\int \nabla_\psi \mathcal{K}(\psi)\mathbb{K}\{(p-0.5)/h\} f_Z(\zeta+\psi h) dF_{P|Z}(dp;\zeta+\psi h) d\psi \\
& -h^{-1}\int\int \nabla_\psi \mathcal{K}(\psi)\mathbb{I}\{p<0.5\} f_Z(\zeta+\psi h) dF_{P|Z}(dp;\zeta+\psi h) d\psi \\
= & h^{-1}\int\int \nabla_\psi \mathcal{K}(\psi) K(\tau) F_{P|Z}(0.5+h\tau;\zeta+\psi h) f_Z(\zeta+\psi h) d\tau d\psi \qquad (8.20) \\
& -h^{-1}\int \nabla_\psi \mathcal{K}(\psi) F_{P|Z}(0.5;\zeta+\psi h) f_Z(\zeta+\psi h) d\psi,
\end{aligned}
$$

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 (8.20) equals

$$
\begin{aligned}
& \int \mathcal{K}(\psi)\nabla_\psi\left[F_{P|Z}(0.5;\zeta+\psi h) f_Z(\zeta+\psi h)\right] d\psi \qquad (8.21) \\
& -\int \mathcal{K}(\psi)\int K(\tau)\nabla_\psi\left[F_{P|Z}(0.5+h\tau;\zeta+\psi h) f_Z(\zeta+\psi h)\right] d\tau d\psi
\end{aligned}
$$

Inserting $F_{P|Z}(0.5+h\tau;\zeta+\psi h)=F_{P|Z}(0.5;\zeta+\psi h)+h\tau f_{P|Z}(0.5;\zeta+\psi h)+...$
$+(r_1!)^{-1} h^{r_1}\tau^{r_1}\partial_p^{r_1-1} f_{P|Z}(0.5+\lambda h\tau;\zeta+\psi h)$, where $\lambda\in(0,1)$, we obtain that (8.21) reduces, under the familiar assumption on all moments of the kernel up to order r_1 to be zero to

$$
-\int \mathcal{K}(\psi)(r_1!)^{-1} h^{r_1}\mu_{r_1}\nabla_\psi \partial_p^{r_1-1} f_{PZ}(0.5;\zeta+\psi h) d\psi,
$$

plus a term of smaller order. Applying standard arguments, in particular expand $\partial_p^{r_1-1} f_{P|Z}(0.5;\zeta+\psi h)$ in ψ, we obtain that $\sqrt{n}T_{2n}^{**}=o_p(1)$, provided that $\sqrt{n}h^{r_1}h^{r}=o(1)$. The same argumentation holds for Q_{2n}.

Step 3b: Next, consider

$$
T_{3n}=n^{-1}\sum_i\left[D_z\widehat{m}_{X|Z}(Z_i)'\right]^{-}\sum_{j\neq i}\nabla_z W_j(Z_i)\left[\mathbb{K}\left\{(\widehat{P}_j-0.5)/h\right\}-\mathbb{K}\{(P_j-0.5)/h\}\right] B(Z_j),
$$

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

$$
\begin{aligned}
T_{3n} &= n^{-1}\sum_i \left[D_z\widehat{m}_{X|Z}(Z_i)'\right]^- \sum_{j\neq i}\nabla_z W_j(Z_i)h^{-1}K\left\{(P_j-0.5)/h\right\}\left(\widehat{P}_j-P_j\right)B(Z_j)+R_{1n}\\
&= T_{4n}+R_{1n},
\end{aligned}
$$

where R_{1n} denotes higher order terms in a mean value expansion, and $R_n = o_p(T_{4n})$. Using again $\left[D_z\widehat{m}_{X|Z}(Z_i)'\right]^- = \left[D_z m_{X|Z}(Z_i)'\right]^- + \left[\left[D_z\widehat{m}_{X|Z}(Z_i)'\right]^- - \left[D_z m_{X|Z}(Z_i)'\right]^-\right]$, which produces a leading term T_{5n} and again a faster converging remainder, we find that

$$
\begin{aligned}
&\sqrt{n}T_{5n}\\
&= n^{-1/2}\sum_i \left[D_z m_{X|Z}(Z_i)'\right]^- n^{-1}\sum_{j\neq i}\nabla_z \frac{h^{-L-2}\mathcal{K}\left((Z_j-Z_i)/h\right)}{\hat{f}_Z(Z_i)}K\left((P_j-0.5)/h\right)\left(\widehat{P}_j-P_j\right)\\
&= \sqrt{n}\int\int \left[D_z m_{X|Z}(z)'\right]^- h^{-2}\int \nabla_\psi \frac{\mathcal{K}(\psi)}{\hat{f}_Z(z)}K_1((p(z+\psi h_1,\varpi)-0.5)/h)\\
&\quad \times(\widehat{p}(z+\psi h_1,\varpi)-p(z+\psi h_1,\varpi))f_{ZV}(z+\psi h,\varpi)d\psi d\varpi F_Z(dz)\\
&= \sqrt{n}\int\int \left[D_z m_{X|Z}(z)'\right]^- h^{-2}\hat{f}_Z(z)^{-1}\times\\
&\quad \nabla_z\left[K_1((p(z,\varpi)-0.5)/h)(\widehat{p}(z,\varpi)-p(z,\varpi))f_{ZV}(z,\varpi)\right]d\varpi F_Z(dz)+\rho_n\\
&= T_{6n}+\rho_n,
\end{aligned}
$$

where $\rho_n = o_p(T_{6n})$. Hence, $\sqrt{n}T_{5n}$ is bounded in absolute value by

$$
c_1\sup_{z,v\in\mathcal{B}\times\mathcal{V}}\left|\nabla_z\widehat{p}(z,v)-\nabla_z p(z,v))\right|b_{1n}+c_2\sup_{z,v\in\mathcal{B}\times\mathcal{V}}\left|\widehat{p}(z,v)-p(z,v))\right|b_{2n}, \tag{8.22}
$$

where $b_{1n} = n^{-1/2}\sum_i\left|\left[D_z m_{X|Z}(Z_i)'\right]^- K_1((p(Z_i,V_i)-0.5)/h_1)f_{ZV}(Z_i,V_i)\right|$ and $b_{2n} = n^{-1/2}\sum_i\left|\left[D_z m_{X|Z}(Z_i)'\right]^-\nabla_z\left[K_1((p(Z_i,V_i)-0.5)/h_1)f_{ZV}(Z_i,V_i)\right]\right|$ converge to nondegenerate distributions. To see this, pick $b_{1n} = h_1^{1/2}(nh_1)^{-1/2}\sum_i\left|\left[D_z m_{X|Z}(Z_i)'\right]^-\right| f_{ZV}(Z_i,V_i)\times K_1((P_i-0.5)/h_1) = h_1^{1/2}b_{3n}$, where b_{3n} is a nonparametric estimator of

$$
\mathbb{E}\left[\left|\left[D_z m_{X|Z}(Z_i)'\right]^-\right| f_{ZV}(Z_i,V_i)|P_i=0.5\right]f_P(0.5).
$$

Observe that b_{3n} converges to a nondegenerate limiting distribution provided that the second moment of $\left|\left[D_z m_{X|Z}(Z_i)'\right]^-\right| f_{ZV}(Z_i,V_i)$ exist. But this follows by elementwise square integrability in assumption 3, together with the boundedness assumption 14. Hence,

$$
c_1\sup_{z,v\in\mathcal{B}\times\mathcal{V}}\left|\nabla_z\widehat{p}(z,v)-\nabla_z p(z,v))\right|b_{1n} = o_p(1).
$$

Similar arguments can be made for the second summand in (8.22), using the boundedness of the derivatives in assumption 14. Consequently, $\sqrt{n}T_{5n} = o_p(1)$, implying that $\sqrt{n}T_{3n} = o_p(1)$. *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

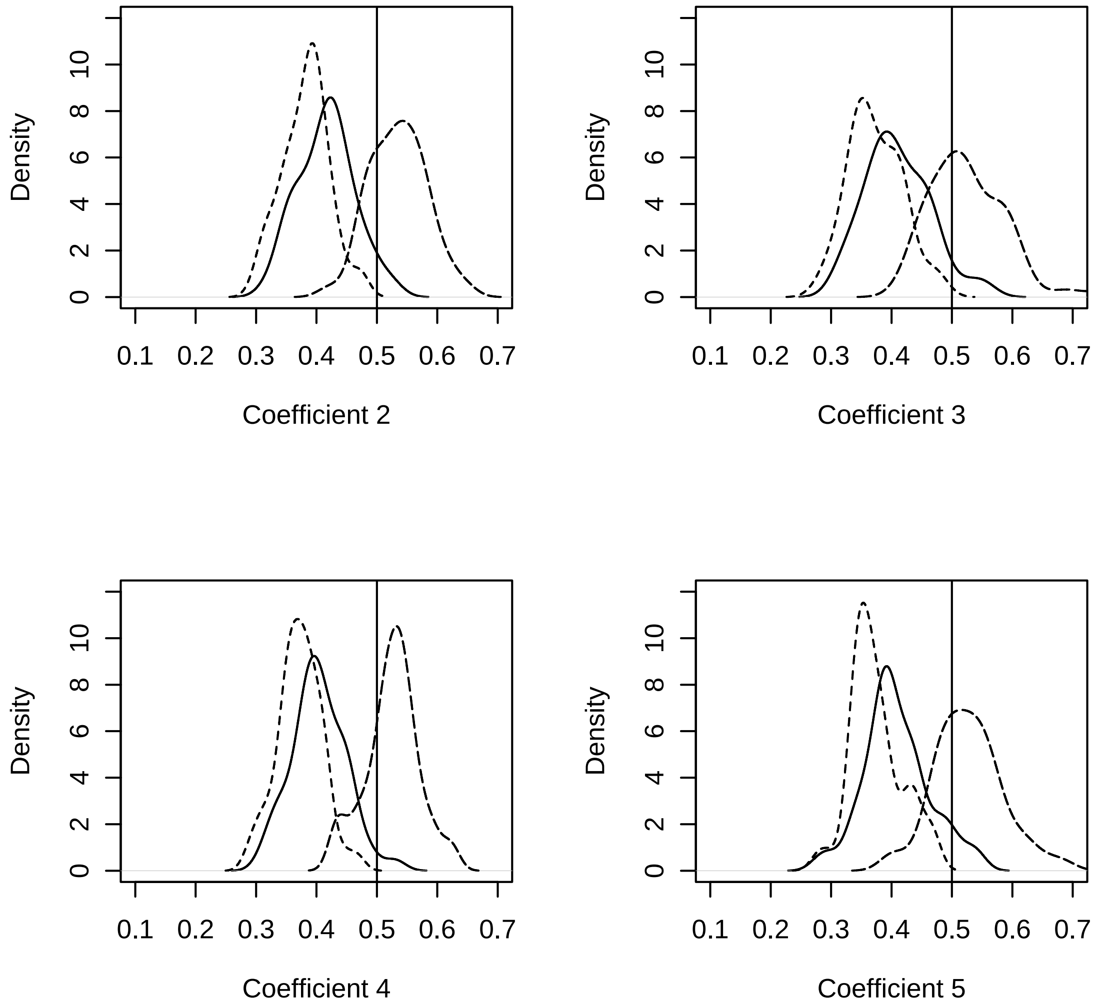
The education level corresponds to the mean education in a non-single household.

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 Estimators for Centrality Parameter

Heteroscedasticity Robust (Solid Line)
Independence (Dotted Line)
Oracle (Broken Line)

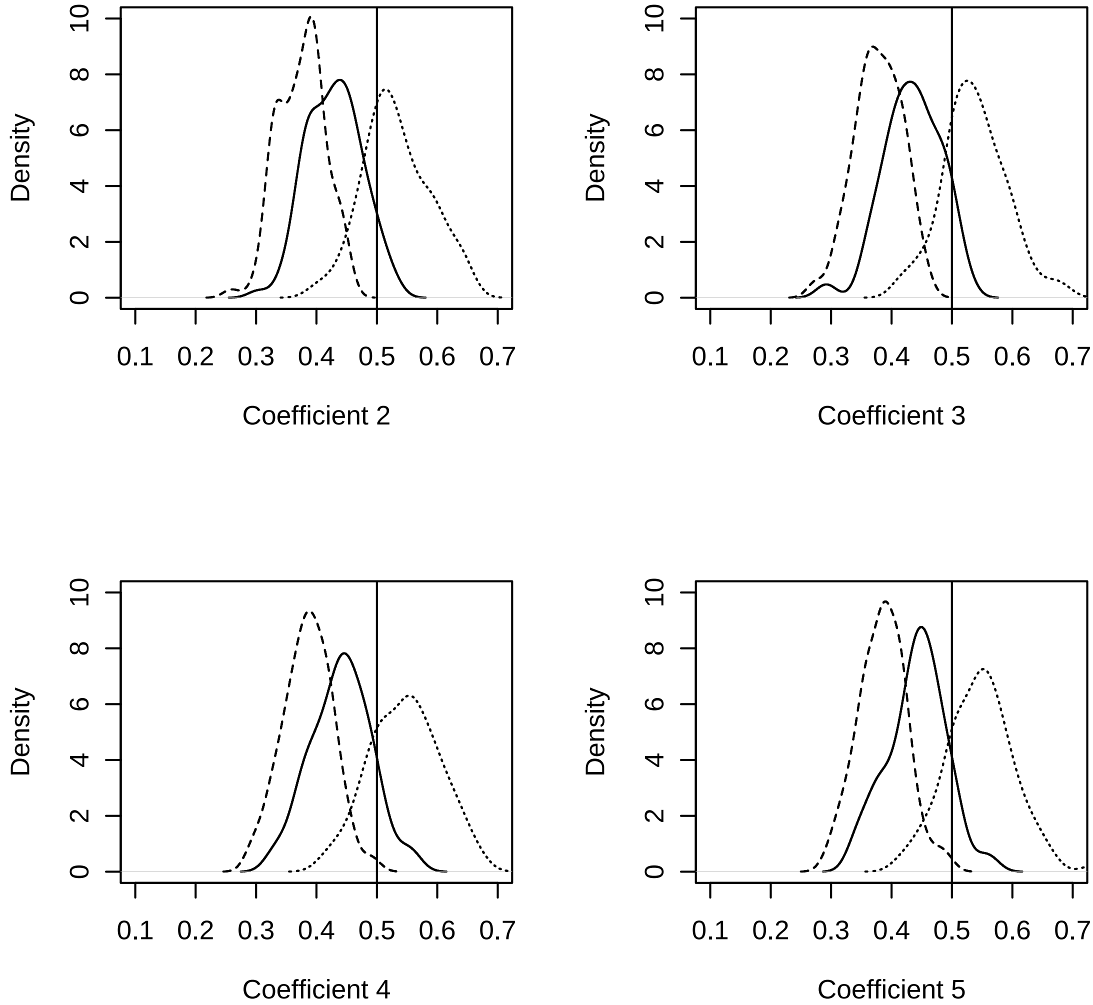


Kernel Estimate of Density of Estimator

n = 2500

Fig. 3: Comparison of Estimators for Centrality Parameter

Heteroscedasticity Robust (Solid Line)
Independence (Dotted Line)
Oracle (Broken Line)

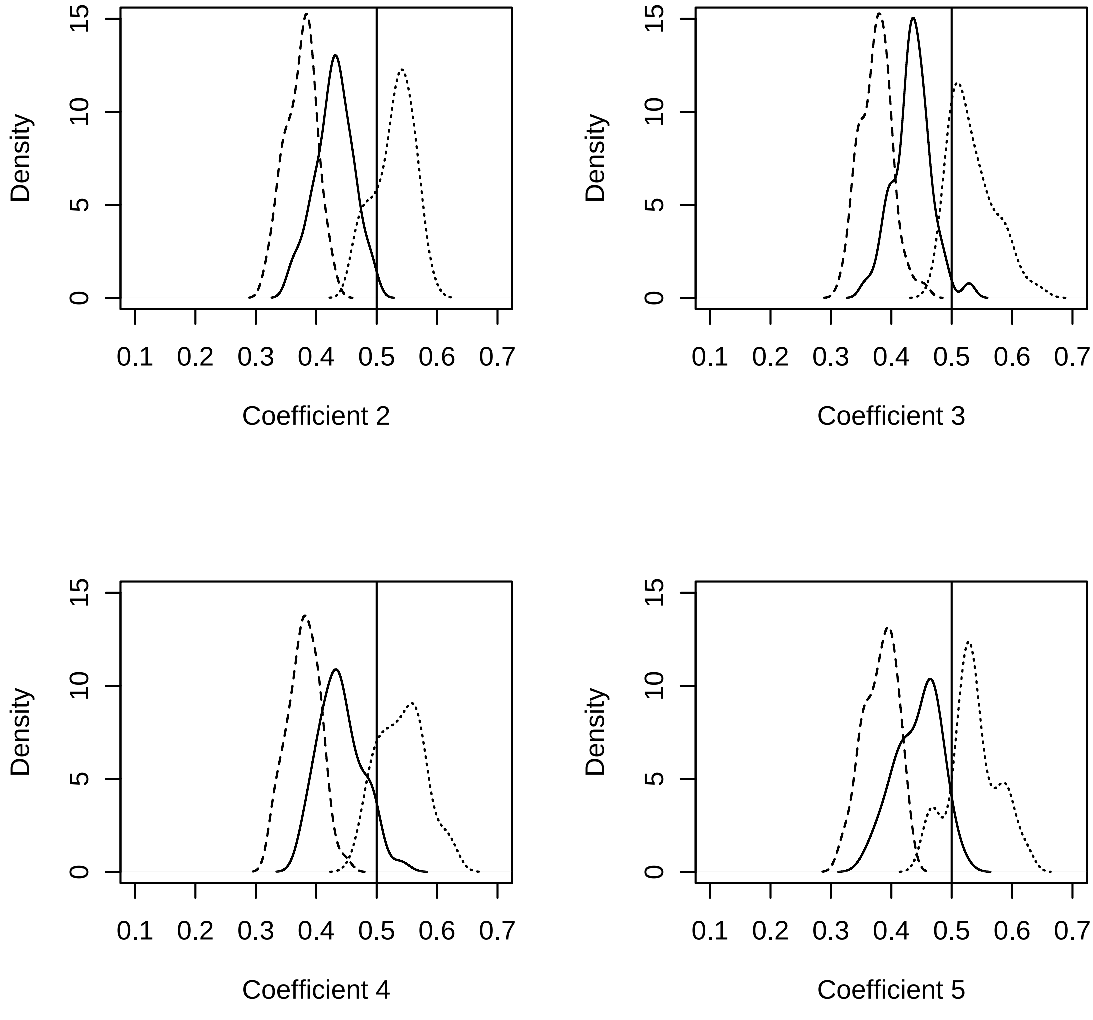


Kernel Estimate of Density of Estimator

n = 7500

Fig. 4: Comparison of Estimators for Centrality Parameter

Heteroscedasticity Robust (Solid Line)
Independence (Dotted Line)
Oracle (Broken Line)



Kernel Estimate of Density of Estimator

n = 15000