WORKING PAPERS

N° TSE-654

May 2016

"Estimation of a semiparametric transformation model in the presence of endogeneity"

Anne Vanhems and Ingrid Van Keilegom

Toulouse School of Economics

Estimation of a semiparametric transformation model in the presence of endogeneity

Anne Vanhems * § Ingrid Van Keilegom * ¶

May 24, 2016

Abstract

We consider a semiparametric transformation model, in which the regression function has an additive nonparametric structure and the transformation of the response is assumed to belong to some parametric family. We suppose that endogeneity is present in the explanatory variables. Using a control function approach, we show that the proposed model is identified under suitable assumptions, and propose a profile estimation method for the transformation. The proposed estimator is shown to be asymptotically normal under certain regularity conditions. A simulation study shows that the estimator behaves well in practice. Finally, we give an empirical example using the U.K. Family Expenditure Survey.

Key Words: Causal inference; Semiparametric regression; Transformation models; Profiling; Endogeneity; Instrumental variable; Control function; Additive models.

*We deeply thank Ying-Ying Lee for pointing out some incoherencies in a previous version of our paper and for most stimulating discussions on the impact of a generated covariate on the asymptotic variance of our estimator.

§Université de Toulouse, Toulouse Business School and Toulouse School of Economics, a.vanhems@tbs-education.fr

¶Institute of Statistics, Biostatistics and Actuarial Sciences, Université catholique de Louvain, Voie du Roman Pays 20, B 1348 Louvain-la-Neuve, Belgium. E-mail address: ingrid.vankeilegom@uclouvain.be.
This research was supported by the European Research Council under the European Community's Seventh Framework Programme (FP7/2007-2013) / ERC Grant agreement No. 203650 and 295298, by IAP research network grant nr. P7/06 of the Belgian government (Belgian Science Policy) and by the contract "Projet d'Actions de Recherche Concertées" (ARC) 11/16-039 of the "Communauté française de Belgique" (granted by the "Académie universitaire Louvain").

1 Introduction

In this paper we consider the problem of estimating a semiparametric transformation model, when some explanatory variables in the model are endogenous. Endogeneity is an important issue in statistics, which is however often ignored in practice. It arises naturally in observational studies, like e.g. in medicine, economics, social sciences, psychology, education, etc. It occurs when some of the independent variables in the model are related to the error term. The formal meaning of 'being related to the error term' depends on the model, like e.g. it could mean that the conditional expectation of the error term is non-zero, or that the error term and the independent variables are not independent. Endogeneity can happen e.g. when relevant explanatory variables are omitted from the model, when certain variables are measured with error, when confounding factors are present, or when simultaneous equations are in place. On the other hand, covariates that are not related to the error term are called exogenous. We refer to the textbooks by Hayashi (2000), Wooldridge (2008) and Imbens and Rubin (2015) for excellent introductions into the problem of endogeneity and how to cope with it in identification, estimation or testing problems.

We are interested in studying the issue of endogeneity in the context of semiparametric transformation models of the following form :

$$\Lambda_\theta(Y) = \phi(X, Z) + \epsilon. \tag{1.1}$$

Here, the response Y is one-dimensional, X takes values in $\mathbb{R}^{d_x}$, and Z in $\mathbb{R}^{d_z}$, with $d_x \geq 1$ and $d_z \geq 0$. The class $\{\Lambda_\theta : \theta \in \Theta\}$ is a parametric family of strictly increasing functions, and the true regression function $\phi_0(\cdot, \cdot)$ has an additive structure given by

$$\phi_0(x, z) = c + \sum_{\alpha=1}^{d_x} \phi_{x0}^\alpha(x_\alpha) + \sum_{\alpha=1}^{d_z} \phi_{z0}^\alpha(z_\alpha), \tag{1.2}$$

with $\mathrm{E}[\phi_{x0}^\alpha(X_\alpha)] = 0$ for $\alpha = 1, \ldots, d_x$ and $\mathrm{E}[\phi_{z0}^\alpha(Z_\alpha)] = 0$ for $\alpha = 1, \ldots, d_z$. We assume moreover that X is endogenous, while Z represents a vector of exogenous random variables, meaning that (X, Z) and ϵ are not independent. Our objective is to identify the structure $(\Lambda_\theta(\cdot), \phi(\cdot, \cdot), F_\epsilon(\cdot) = \Pr(\epsilon \leq \cdot))$, to estimate θ and ϕ given a sample of observations and to do inference on these estimators.

When endogeneity is present, ordinary regression techniques produce biased and inconsistent estimators. There exist several approaches to cope with this issue. The technique we use in this paper is based on so-called 'control variables'. A control variable is such that the error term in the model is conditionally unrelated to the explanatory variables given this

control variable, whereas without conditioning on this variable the explanatory variables (or at least the endogenous ones) would be related to the error term. So, in a sense the control variable re-establishes in a sense the desirable property that the covariates and the error term are not related, which is crucial to do correct inference. The control function approach has been detailed in several papers, see e.g. Newey, Powell and Vella (1999) or Imbens and Newey (2009).

A legitimate question is how to find an appropriate control function in practice. As we will see further in this paper, a control variable can be constructed once we have so-called 'instrumental variables' at our disposal. These are variables that are not part of the original model, they are depending on the endogenous variables conditional on the other covariates, and they are unrelated to the error term in the model (i.e. the instruments do not suffer from the same problem as the original explanatory variables). In other words, the instrumental variable does not have a direct effect on the response, other than through the endogenous variables.

We illustrate the concept of instrumental variable by means of the following textbook example : let X be the price of an agricultural good, and let Y be the demand for the good. This is a case where endogeneity could be present, since the price of a good influences the demand, and vice versa (so we have so-called 'simultaneous equations'). A possible instrument W in this case could be a certain measure of favorable growing conditions, since it could be believed that W is related to X and does not influence Y in a direct way, other than through X.

Many other examples can be found in the literature, see e.g. Angrist and Krueger (2001), Johannes, Van Bellegem and Vanhems (2013) and Manzi, San Martin and Van Bellegem (2014). Detecting sources of endogeneity and finding appropriate instrumental variables is a difficult empirical problem. The aim of this paper is not to propose solutions to this problem. Researchers doing applied work are in a much better position for answering this delicate question. Instead, our goal is to study the interesting statistical challenges encountered when endogeneity arises in the semiparametric transformation model defined in (1.1) with a given instrumental variable W.

Transformation models lie at the heart of many problems in statistics, since they aid interpretability, they lead to approximately additive regression functions, they stabilize the variance of the error, and they help to obtain errors that are approximately normal. A seminal paper in the literature on transformations is the one by Box and Cox (1964), who proposed a parametric family of power transformations that includes as special cases the logarithm and the identity. Other transformations have been proposed in the literature, like

for example, the Zellner and Revankar (1969) transform and the Bickel and Doksum (1981) transform. See also the book by Carroll and Ruppert (1988) and the review paper by Sakia (1992) for more details and references on this topic.

Various papers have studied transformation models under different sets of assumptions. In a fully exogenous setting, some papers have considered nonparametric forms for Λ and ϕ, like Horowitz (2001) or Jacho-Chavez, Lewbel and Linton (2010). Other papers have analyzed semiparametric transformation models by either assuming a parametric form for ϕ, like in Horowitz (1996) or Moon (2013), or a parametric form for Λ, as in Linton, Sperlich and Van Keilegom (2008). The latter model has also been studied by Colling, Heuchenne, Samb and Van Keilegom (2015) and Heuchenne, Samb and Van Keilegom (2014), who studied nonparametric estimators of the density and of the distribution function of the error term, and by Colling and Van Keilegom (2014) and Neumeyer, Noh and Van Keilegom (2014), who developed tests under this model. Our work extends the latter model by considering a vector X of endogenous variables, and we focus on the problem of estimating the different components of the model.

The issue of endogeneity has already been investigated in the setting of transformation models. Chiappori, Komunjer and Kristensen (2010) consider a fully nonparametric setting and, with a little stronger assumption of conditional independence between ϵ and one coordinate of X, are able to identify the model and recover a parametric rate of convergence for the estimated transformation operator. On the other hand, Florens and Sokullu (2012) and Fève and Florens (2010) consider a semiparametric form for the function ϕ and identify and estimate the model using an instrument W and by imposing very few technical assumptions (like conditional mean independence) in the line of ill-posed inverse problems theory. In our case, the parametric assumption concerns the operator Λ and we identify the model using a control function approach.

We also note that there exists a limited literature on other semiparametric regression models with endogenous variables. We refer to Chen and Pouzo (2009) for semiparametric inference with nonsmooth residuals, Florens, Johannes and Van Bellegem (2012) for instrumental regression in partially linear models, and Birke, Van Bellegem and Van Keilegom (2014) for instrumental regression in semiparametric single index models.

At last, one could also relate our work to the semiparametric analysis with generated covariates developed in Mammen, Rothe and Schienle (2012) since the control function needs to be estimated in a first step. However, we also need to take into account the estimation of the density of the error term ϵ in the estimation process, and our estimation procedure is therefore, from a structural point of view, quite different from theirs. We will detail more

explicitly the differences with the existing literature further on in this paper.

The paper is organized as follows. Section 2 is devoted to the identification of the model. In Section 3 we explain in detail our estimation procedure. Section 4 states the consistency and asymptotic normality of the estimators of θ and ϕ. A finite sample study is presented in Section 5, including some simulations and an application to real data, and we also propose a bootstrap procedure to estimate the distribution of $\widehat{\theta}$ in practice. Some general conclusions are given in Section 6, and finally all the proofs are collected in the Appendix.

2 Identification

Consider model (1.1) with, as explained earlier, a vector of endogenous variables X and a vector of exogenous variables Z. We use a control function approach to treat the endogeneity and we assume that there exists a control variable V such that :

(A.1) (X, Z) and ϵ are independent conditional on V

(A.2) The support of V conditional on (X, Z) equals the support of V.

These assumptions are standard in the literature on nonseparable models (see Imbens and Newey 2009) and will allow to identify the functions ϕ and F_ϵ. The result we present below allows to identify the fully nonparametric structure $(\Lambda, \phi, F_\epsilon)$, i.e. ϕ is not necessarily additive but can take any functional form, and Λ can be any monotone transformation that does not necessarily belong to a parametric family. Therefore, in this section, we omit the index θ for the operator Λ and the functions ϕ and F_ϵ.

To stick to a general setting, we suppose there exists an unknown function r and an instrumental variable W such that $V \equiv r(X, Z, W)$ satisfies assumptions $(A.1)$ and $(A.2)$ and r is identified. In Remark 2.2 below, we will give some classical examples of this function r. Moreover, we assume that the random vector (X, Z, W, Y) is absolutely continuous with density $f_{X,Z,W,Y}$, whose support is $R_{X,Z,W,Y} \subset \mathbb{R}^{d_x+d_z+d_w+1}$.

We also need to identify Λ and based on Chiappori, Komunjer and Kristensen (2010) and Linton, Sperlich and Van Keilegom (2008), we impose the following additional assumptions:

(A.3) Λ is a continuously differentiable and strictly increasing function defined on the support R_Y of Y.

(A.4) For almost all $(x, z) \in R_{X,Z}$ (the support of (X, Z)), the density $f_{\epsilon|X,Z}(\cdot|x, z)$ exists, is strictly positive and continuously differentiable.

(A.5) The derivative of ϕ with respect to x_1 (the first coordinate of x) exists and the set $\{(x,z) \in R_{X,Z} : \frac{\partial}{\partial x_1}\phi(x,z) \neq 0\}$ has a nonempty interior.

(A.6) $\mathrm{E}(\Lambda(Y)) = 1$, $\Lambda(0) = 0$, and $\mathrm{E}(\epsilon) = 0$.

Our result is based on the equality:

$$\begin{aligned} F_{Y|X,Z,V}(y|x,z,v) &= \Pr[\Lambda(Y) \leq \Lambda(y)|X = x,\ Z = z,\ V = v] \\ &= \Pr[\epsilon \leq \Lambda(y) - \phi(X,Z)|X = x,\ Z = z,\ V = v] \\ &= \Pr[\epsilon \leq \Lambda(y) - \phi(x,z)|V = v], \end{aligned}$$

where the first equality comes from the monotonicity Assumption $(A.3)$, and the third one follows from Assumption $(A.1)$. Then, following Imbens and Newey (2009) we have:

$$\int F_{Y|X,Z,V}(y|x,z,v)F_V(dv) = F_\epsilon(\Lambda(y) - \phi(x,z)). \tag{2.1}$$

Proposition 2.1. *Under Assumptions $(A.1) - (A.6)$, the structure $(\Lambda, \phi, F_\epsilon)$ is identified.*

The proof is given in the Appendix.

Remark 2.1. *1. Note that Chiappori, Komunjer and Kristensen (2010) suggest a slightly different independence assumption, instead of $(A.1)$: ϵ is independent of X_1 conditional on (X_{-1}, Z, V) (where $X = (X_1, X_{-1})$). Although an equivalent identification result could be derived with their set of assumptions, the estimation of the parameter θ would become more tricky since the distribution of ϵ would remain conditional on (X_{-1}, Z).*

2. Note also that Proposition 2.1 only gives sufficient conditions to identify the structure $(\Lambda, \phi, F_\epsilon)$. In particular, Assumption $(A.2)$ could be weakened using a separability assumption as proposed in Newey, Powell and Vella (1999). Indeed, once Λ is identified using Assumptions $(A.1)$, $(A.3) - (A.6)$, we get:

$$\mathrm{E}\left(\Lambda(Y)|X = x, Z = z, V = v\right) = \phi(x,z) + \lambda(v),$$

where $\lambda(v) = \mathrm{E}\left[\epsilon|V = v\right]$. Then, using Theorem 2.2 in Newey, Powell and Vella (1999) and the normalization assumption $(A.6)$, we conclude that if there is no functional relationship between (X, Z) and V, then ϕ is identified.

Remark 2.2. *Note that different candidates can be proposed to characterize the control variable V. In the line of Newey, Powell and Vella (1999), V can be defined as the error of the following (separable) nonparametric model :*

$$X = \psi(Z, W) + V, \tag{2.2}$$

where W is a vector of instrumental variables taking values in $\mathbb{R}^{d_w}$ such that (ϵ, V) and (Z, W) are independent, in order to satisfy Assumption (A.1).

A second option would be to consider a nonseparable model and a single endogenous variable X defined by:

$$X = \psi(Z, W, \eta), \tag{2.3}$$

where ψ is strictly monotone in η. Then, $V = F_{X|Z,W}(X|Z, W) = F_\eta(\eta)$ is a uniformly distributed control variable under the following conditions: (i) (ϵ, η) and (Z, W) are independent, and (ii) η is a continuously distributed random variable with strictly increasing distribution function on the support of η and $\psi(Z, W, t)$ is strictly monotone in t with probability 1 (see Imbens and Newey 2009 for more details).

A natural extension of model (2.3) when X is multidimensional, consists in considering the set of one-dimensional independent models:

$$\begin{cases} X_1 &= \psi_1(Z, W, \eta_1) \\ &\vdots \\ X_{d_x} &= \psi_{d_x}(Z, W, \eta_{d_x}), \end{cases} \tag{2.4}$$

and $\eta = (\eta_1, ..., \eta_{d_x})$.

3 Estimation

Although a fully nonparametric approach is possible, we return now (and for the rest of the paper) to model (1.1), which assumes that the transformation Λ is parametric and that the true regression function ϕ_0 has the additive structure given in (1.2). Hence, we assume that $\Lambda(\cdot) \equiv \Lambda_\theta(\cdot)$, for some parametric family $\{\Lambda_\theta(\cdot) : \theta \in \Theta\}$, where we suppose that Θ is compact. Indeed, considering a parametric transformation can lead to easier interpretation, like for the family of power transformations proposed by Box and Cox (1964), and the Bickel and Doksum (1981) class of transformations.

From equation (2.1) we obtain:

$$\int f_{Y|X,Z,V}(y|x, z, v) dF_V(v) = f_{\epsilon(\theta_0)}\left(\Lambda_0(y) - \phi_0(x, z)\right) . \Lambda_0'(y), \tag{3.1}$$

where $f_{\epsilon(\theta_0)}$ and $f_{Y|X,Z,V}$ are the probability density functions of ϵ and of Y given (X, Z, V), respectively, and where θ_0 is the true value of θ and $\Lambda_0 \equiv \Lambda_{\theta_0}$.

Consider now a randomly drawn i.i.d. sample (X_i, Z_i, W_i, Y_i), $i = 1, ..., n$ from the random vector (X, Z, W, Y). Then, the criterion function is derived from equation (3.1) by:

$$\sum_{i=1}^{n} \Big\{ \log[f_{\epsilon(\theta_0)} (\Lambda_0(Y_i) - \phi_0(X_i, Z_i))] + \log[\Lambda_0'(Y_i)] \Big\}. \tag{3.2}$$

This criterion function depends on the unknown functions $f_{\epsilon(\theta_0)}$, r and ϕ_0. The idea is now to estimate θ be replacing all unknown quantities in the above criterion function by nonparametric estimators for a fixed value of θ, and to maximize the so-obtained expression with respect to the unknown parameter θ. In what follows, we denote $H(\theta, f, \phi) = \mathrm{E}\big\{\log[f (\Lambda_\theta(Y) - \phi(X, Z))] + \log[\Lambda_\theta'(Y)]\big\}$ and we let $H_n(\theta, f, \phi)$ be its empirical counterpart.

Let us first of all consider the estimation of the function ϕ_0. Consider, for $\theta \in \Theta$, the functions

$$m_\theta(x, z, v) = \mathrm{E}\left(\Lambda_\theta(Y) | X = x, Z = z, V = v\right)$$

and

$$\phi_\theta^{add}(x, z) := c_\theta + \sum_{\alpha=1}^{d_x} \phi_{x\theta}^{\alpha}(x_\alpha) + \sum_{\alpha=1}^{d_z} \phi_{z\theta}^{\alpha}(z_\alpha)$$

with $\phi_{x\theta}^{\alpha}(x_\alpha) = \mathrm{E}(m_\theta(x_\alpha, X_{-\alpha}, Z, V)) - c_\theta$, where $X = (X_\alpha, X_{-\alpha})$, $\phi_{z\theta}^{\alpha}(z_\alpha) = \mathrm{E}(m_\theta(X, z_\alpha, Z_{-\alpha}, V)) - c_\theta$, where $Z = (Z_\alpha, Z_{-\alpha})$, and $c_\theta = \mathrm{E}[\Lambda_\theta(Y)]$. Hence, for estimating $\phi_\theta^{add}(x, z)$, we first need to estimate $m_\theta(x, z, v)$.

Remark 3.1. *Note that for $\phi_\theta(x, z) := \mathrm{E}[m_\theta(x, z, V)]$ we have in general that $\phi_\theta^{add}(x, z) \neq \phi_\theta(x, z)$ except if $\theta = \theta_0$, since the additive structure of $m_\theta(x, z, v)$ only holds for $\theta = \theta_0$.*

Denoting $m_0 \equiv m_{\theta_0}$, we have that

$$m_0(x, z, v) \;=\; \phi_0(x, z) + \lambda(v), \tag{3.3}$$

where $\lambda(v) = \mathrm{E}\left[\epsilon | V = v\right]$ using assumption $(A.1)$. Note that, under Assumption $(A.6)$ we have:

$$\mathrm{E}\left[\lambda(V)\right] = \mathrm{E}\left[\mathrm{E}\left(\epsilon |\ V\right)\right] = \mathrm{E}\epsilon \;=\; 0.$$

We assume in what follows that we dispose of a nonparametric estimator of $V_i = r(X_i, Z_i, W_i)$, denoted by $\widehat{V}_i = \widehat{r}(X_i, Z_i, W_i)$ $(i = 1, \ldots, n)$. For instance, consider the nonseparable equation (2.3). A nonparametric estimator of V_i is then given by

$$\begin{aligned} \widehat{V}_i &= \widehat{F}_{X|Z,W}(X_i | Z_i, W_i) \\ &= \frac{\sum_{j=1}^{n} 1(X_j \leq X_i) K_h(Z_i - Z_j) K_h(W_i - W_j)}{\sum_{j=1}^{n} K_h(Z_i - Z_j) K_h(W_i - W_j)}, \end{aligned} \tag{3.4}$$

where K is a d-dimensional product kernel of the form $K(u_1,\ldots,u_d)=\prod_{j=1}^{d}k_1(u_j)$, with $d=d_z$ or d_w and k_1 is a univariate kernel function. As usual, h is a bandwidth converging to zero when n tends to infinity, $k_{1h}(\cdot)=k_1(\cdot/h)/h$ and $K_h(u_1,\ldots,u_d)=\prod_{j=1}^{d}k_{1h}(u_j)$. Later in the paper we will develop conditions on $\widehat{V}_i-V_i$ that are needed for the asymptotic theory.

We first estimate the function $m_\theta(x,z,v)$ by using a nonparametric kernel estimator based on $(X_i,Z_i,\widehat{V}_i,Y_i)$ $(i=1,\ldots,n)$:

$$\begin{aligned}\widehat{m}_\theta(x,z,v) &= \widehat{\mathrm{E}}\left[\Lambda_\theta(Y)|X=x,Z=z,\widehat{V}=v\right]\\ &= \frac{\sum_{i=1}^{n}\Lambda_\theta(Y_i)K_h(x-X_i)K_h(z-Z_i)K_h(v-\widehat{V}_i)}{\sum_{i=1}^{n}K_h(x-X_i)K_h(z-Z_i)K_h(v-\widehat{V}_i)}.\end{aligned}$$

For simplifying the presentation, we work with the same bandwidth for all variables.

In what follows, we use marginal integration techniques (see e.g. Linton and Nielsen 1995). Note that other methods could have been used like smooth backfitting techniques (see Mammen, Linton and Nielsen 1999). We briefly comment on this in Section 4. Consider

$$\begin{aligned}\widehat{\phi}^{\alpha}_{x\theta}(x_\alpha) &= \frac{1}{n}\sum_{i=1}^{n}\widehat{m}_\theta(x_\alpha,X_{-\alpha i},Z_i,\widehat{V}_i)-\widehat{c}_\theta \qquad (\alpha=1,\ldots,d_x)\\ \widehat{\phi}^{\alpha}_{z\theta}(z_\alpha) &= \frac{1}{n}\sum_{i=1}^{n}\widehat{m}_\theta(X_i,z_\alpha,Z_{-\alpha i},\widehat{V}_i)-\widehat{c}_\theta \qquad (\alpha=1,\ldots,d_z),\end{aligned}$$

where $\widehat{c}_\theta=n^{-1}\sum_{i=1}^{n}\Lambda_\theta(Y_i)$. The nonparametric estimator of $\phi^{add}_\theta(x,z)$ is now given by:

$$\widehat{\phi}^{add}_\theta(x,z)=\widehat{c}_\theta+\sum_{\alpha=1}^{d_x}\widehat{\phi}^{\alpha}_{x\theta}(x_\alpha)+\sum_{\alpha=1}^{d_z}\widehat{\phi}^{\alpha}_{z\theta}(z_\alpha). \tag{3.5}$$

Using the estimator of $\phi^{add}_\theta(x,z)$ we can now estimate the error density $f_{\epsilon(\theta)}$ of the variable $\epsilon(\theta)=\Lambda_\theta(Y)-\phi^{add}_\theta(X,Z)$ for a fixed value of θ:

$$\widehat{f}_{\epsilon(\theta)}(e)=\frac{1}{n}\sum_{i=1}^{n}k_{2g}\left(e-\widehat{\epsilon}_i(\theta)\right), \tag{3.6}$$

where $\widehat{\epsilon}_i(\theta)=\Lambda_\theta(Y_i)-\widehat{\phi}^{add}_\theta(X_i,Z_i)$, k_2 is a univariate kernel, and g is a bandwidth parameter.

Finally, we are in position to estimate the transformation parameter θ, by plugging-in the estimators of all unknown quantities in the criterion function given in (3.2):

$$\begin{aligned}\widehat{\theta} &= \arg\max_{\theta\in\Theta}H_n\left(\theta,\widehat{f}_{\epsilon(\theta)},\widehat{\phi}^{add}_\theta\right) \qquad (3.7)\\ &= \arg\max_{\theta}\sum_{i=1}^{n}\Big\{\log[\widehat{f}_{\epsilon(\theta)}(\Lambda_\theta(Y_i)-\widehat{\phi}^{add}_\theta(X_i,Z_i))]+\log[\Lambda'_\theta(Y_i)]\Big\}.\end{aligned}$$

Once θ is estimated we can estimate the unknown regression function $\phi_0(x, y)$. This gives

$$\widehat{\phi}^{add}(x, z) = \widehat{\phi}^{add}_{\widehat{\theta}}(x, z)$$

for any x and z.

4 Large sample properties

In this section we present the consistency and the asymptotic normality of our estimators. Our consistency result will be proved using the paper by Delsol and Van Keilegom (2014), which considers general semi-parametric M-estimation problems when the criterion function is not necessarily smooth and is allowed to have several local maxima. This framework is appropriate in our context, since the criterion function defined in (3.7) depends in a complicated way on θ, and so the existence of a unique (local) maximizer is not guaranteed.

The regularity conditions (C.1)–(C.10) under which the results below are valid, are given in the Appendix.

Theorem 4.1. *Assume (A.1)–(A.6) and (C.1)–(C.9). Then,*

$$\widehat{\theta} - \theta_0 \xrightarrow{P} 0.$$

Given that we now know that $\widehat{\theta}$ is a consistent estimator of θ_0, we can from now on maximize the criterion function with respect to a shrinking neighborhood around θ_0. In this shrinking neighborhood the criterion function will have a unique local maximum (namely $\widehat{\theta}$) and hence we can from now on consider $\widehat{\theta}$ as the solution of the derivative of the criterion function H with respect to θ over this shrinking neighborhood, and prove the asymptotic normality using the general framework considered in Chen, Linton and Van Keilegom (2003) on semiparametric Z-estimation. (Note that Delsol and Van Keilegom (2014) also propose some asymptotic distribution theory of their estimator, but in a much more general setting since their criterion function may not be differentiable, which is not our case.)

We now denote Θ for a shrinking neighborhood of the finite dimensional parameter set around θ_0 (and we will implicitly consider the associated shrinking neighborhood for the infinite dimensional parameter space). We define a non-random measurable vector-valued function G by the derivative of the function H with respect to θ:

$$G(\theta, \gamma^{add}_{\theta}) = \mathrm{E}\{M(\theta, \gamma^{add}_{\theta}, X, Z, W, Y)\}.$$

Here, γ_θ^{add} is a vector of nuisance parameters defined by

$$\gamma_\theta^{add} = (\phi_\theta^{add}, \dot{\phi}_\theta^{add}, f_{\epsilon(\theta)}, f'_{\epsilon(\theta)}, \dot{f}_{\epsilon(\theta)})^t,$$

where $\dot{\phi}_\theta^{add}$ (respectively $\dot{f}_{\epsilon(\theta)}$) denotes the vector of partial derivatives of ϕ_θ^{add} (respectively $f_{\epsilon(\theta)}$) with respect to the components of θ and $f'_{\epsilon(\theta)}(y)$ denotes the derivative of $f_{\epsilon(\theta)}(y)$ with respect to y, and the function M is defined as follows:

$$\begin{aligned} & M(\theta, \gamma_\theta^{add}, X, Z, W, Y) \\ & = \frac{1}{f_{\epsilon(\theta)}(\epsilon(\theta))}\Big[f'_{\epsilon(\theta)}(\epsilon(\theta))\big\{\dot{\Lambda}_\theta(Y) - \dot{\phi}_\theta^{add}(X, Z)\big\} + \dot{f}_{\epsilon(\theta)}(\epsilon(\theta))\Big] + \frac{\dot{\Lambda}'_\theta(Y)}{\Lambda'_\theta(Y)}. \end{aligned} \tag{4.1}$$

Let $M_n(\theta, \gamma_\theta^{add}) = n^{-1}\sum_{i=1}^n M(\theta, \gamma_\theta^{add}, X_i, Z_i, W_i, Y_i)$. Then, $M_n(\theta, \widehat{\gamma}_\theta^{add})$ is the derivative (up to the multiplicative factor n^{-1}) of the criterion function defined in equation (3.7) with respect to θ, where $\widehat{\gamma}_\theta^{add} = (\widehat{\phi}_\theta^{add}, \dot{\widehat{\phi}}_\theta^{add}, \widehat{f}_{\epsilon(\theta)}, \widehat{f}'_{\epsilon(\theta)}, \dot{\widehat{f}}_{\epsilon(\theta)})^t$.

Remark 4.1. *Note that, by construction, $G(\theta, \gamma_\theta^{add}) = 0$ at $\theta = \theta_0 \in \Theta$ where $\theta_0 \in \Theta$ and $\gamma_0^{add} \equiv \gamma_{\theta_0}^{add}$ are the true unknown finite and infinite dimensional parameters. Note also that $\|M_n(\theta, \widehat{\gamma}_\theta^{add})\|$ takes its minimum at $\widehat{\theta}$, where $\|\cdot\|$ denotes the Euclidean norm.*

We denote by Γ the matrix of partial derivatives of $G(\theta, \gamma_\theta^{add})$ with respect to θ:

$$\Gamma = \dot{G}(\theta, \gamma_\theta^{add})\Big|_{\theta=\theta_0} \tag{4.2}$$

We also need to introduce the matrix

$$\Sigma \;=\; \mathrm{Var}\big\{A(T)\big\}, \tag{4.3}$$

where

$$A(T) \;=\; M(\theta_0, \gamma_0^{add}, T) + \sum_{\alpha=1}^{d_x+d_z} D_1^\alpha(T) + D_2(T), \tag{4.4}$$

$T = (X, Z, W, Y)$, and the functions D_1^α and D_2 are given in (7.6) and (7.7) in the Appendix.

We are now ready to state the asymptotic normality result :

Theorem 4.2. *Assume (A.1)–(A.6) and (C.1)–(C.10). Then,*

$$n^{1/2}(\widehat{\theta} - \theta_0) \xrightarrow{d} N(0, \Omega),$$

where

$$\Omega = \Gamma^{-1}\Sigma(\Gamma^t)^{-1},$$

and where Γ and Σ are defined in (4.2) and (4.3).

The following corollary is a by-product of the main result:

Corollary 4.1. *Assume (A.1)–(A.6) and (C.1)–(C.10). Consider the notation* $S = (X, Z)$ *and* $d_s = d_x + d_z$. *Then, for any* $s = (x, z) \in R_{X,Z}$,

$$(nh)^{1/2}\Big(\widehat{\phi}^{add}(s) - \phi_0(s)\Big) \xrightarrow{d} N(0, \sigma^2(s)),$$

where

$$\sigma^2(s) = \int k_1^2(u)du \sum_{\alpha=1}^{d_s} f_{S_\alpha}(s_\alpha)\, Var\Big\{\Big[\Lambda_0(Y) - m_{\theta_0}(S,V)\Big] f^{-1}_{S_\alpha|S_{-\alpha},V}(s_\alpha|S_{-\alpha},V)\Big|S_\alpha = s_\alpha\Big\}.$$

Let us comment on these asymptotic results:

1. It can be seen from the proof of Theorem 4.2 that the extra terms in the formula of Σ come from the estimation of the nuisance functions ϕ_0, ϕ_0^{add}, $f_{\epsilon(\theta_0)}$, $f'_{\epsilon(\theta_0)}$ and $\dot{f}_{\epsilon(\theta_0)}$. Note that these terms would be equal to zero if (X, Z) and ϵ would be independent, which is the case in the exogenous model considered by Linton, Sperlich and Van Keilegom (2008). Another difference between the variance in the endogenous and the exogenous case lies in the formula of $\phi_\theta^{add}(x, z)$ (denoted by $m_\theta(x)$ in their paper). Even for $\theta = \theta_0$, the function $\phi_0(x, z)$ is different in the two cases, namely in the exogenous case it equals $\mathrm{E}[\Lambda_0(Y)|X = x, Z = z]$, whereas in the endogenous case it is given by $\int \mathrm{E}[\Lambda_0(Y)|X = x, Z = z, V = v]dF_V(v)$.

2. Note that the asymptotic distribution of $\widehat{\phi}^{add}(x, z)$ in Corollary 4.1 is the same as that of $\widehat{\phi}_0^{add}(x, z)$, i.e. the asymptotic distribution is as if the parameter θ_0 were known.

3. Instead of using the marginal integration method to estimate $\phi_0(x, z)$, we could as well use other estimation procedures, like e.g. the smooth backfitting method (see e.g. Mammen, Linton and Nielsen, 1999, and Mammen and Park, 2005). However, the proofs are considerably more complicated in that case. For the smooth backfitting, we expect that the asymptotic distribution of $\widehat{\theta}$ will be the same as for the marginal integration method, except that $\phi_\theta^{add}(x, z)$ is now given by the components depending on x and z of the function $m_\theta^{add}(x, z, v)$ defined as:

$$m_\theta^{add}(x, z, v) = \mathrm{argmin}_{m \in \mathcal{M}_{add}} \int \Big[m_\theta(x, z, v) - m(x, z, v)\Big]^2 dF_{X,Z,V}(x, z, v),$$

where

$$\mathcal{M}_{add} = \Bigg\{ m : m(x,z,v) = \sum_{\alpha=1}^{d_x} m_{x_\alpha}(x_\alpha) + \sum_{\alpha=1}^{d_z} m_{z_\alpha}(z_\alpha) + m_v(v) \\ \text{for some } m_{x_1}, \ldots, m_{x_{d_x}}, m_{z_1}, \ldots, m_{z_{d_z}}, m_v \Big\} .$$

Proving the asymptotic properties of this type of estimator is however not at all an easy task. We therefore restrict attention in this paper to the marginal integration estimator. The refinement of our method to smooth backfitting methods (or other methods to estimate an additive regression function) is left as a topic of future research.

4. The asymptotic results of this section can be compared with some related papers. First of all, the paper by Mammen, Rothe and Schienle (2012) considers also a general class of semiparametric optimization estimators with infinite-dimensional nuisance parameters that include a conditional expectation function estimated nonparametrically using generated covariates. In our model, the generated covariate V affects the function ϕ_θ^{add}, its derivative with respect to θ, the residual density function $f_{\epsilon(\theta)}$ as well as its derivatives with respect to the principal argument and to θ. This structural difference between both models has of course an impact, not only on the estimation step, but also on the inference.

   Second, our model extends the setup considered in Linton, Sperlich and Van Keilegom (2008), which includes no endogenous variable X, and therefore no generated covariate V. As it has just been stressed, the estimation of V appears in each step and thus affects all the nuisance functions. In addition, the assumption of endogeneity implies that (X, Z) and ϵ are not independent anymore, which complicates a lot the derivation of the asymptotic variance in Theorem 3.2. This second main difference is stressed in the first comment above, as well as in the proof where more lemmas are required to derive the asymptotic normality (Lemmas 7.2 and 7.3).

   Third, our framework is also very different from Imbens and Newey (2009) although the identification proof is partly based on their arguments. From a structural point of view, we need to identify two functions namely Λ and ϕ whereas they only consider the identification of ϕ. Moreover, we consider a semiparametric model and our estimation procedure includes the estimation of the parameter θ (whereas they consider a fully nonparametric setting). As we have stressed above, the estimation of θ in an endogenous setting complicates a lot the estimation step, since our model also requires the

estimation of the function ϕ and the density f_ϵ of the error, as well as the derivatives of ϕ and f_ϵ.

5 Finite sample study

5.1 Simulations

We consider the following data generating process:

$$\Lambda_\theta(Y) = b_0 + b_1 X + \epsilon,$$

where Λ_θ is the Box-Cox transformation, that is $\Lambda_\theta(y) = \frac{y^\theta - 1}{\theta}$ $(\theta \neq 0)$, $\Lambda_\theta(y) = \log(y)$ $(\theta = 0)$, and ϵ is drawn from $N(0, \sigma_e^2)$. In this setting, we omit the exogenous variable Z. The variable X is generated from the following generating process:

$$X = a_0 + a_1 W + a_2 \epsilon + U,$$

where W, ϵ and U are mutually independent, W is drawn from $N(0, \sigma_w^2)$ and U from $N(0, \sigma_u^2)$. The regressor X is then correlated with the error term ϵ and the instrumental variable W is correlated with X but not with ϵ in order to correct for this endogeneity issue. The control function V is identified as the residual of the regression of X on W. We present here the results for the case where $b_0 = 1, b_1 = 0.25, a_0 = 1, a_1 = -0.5, a_2 = 2$, $\sigma_w = 1, \sigma_e = 0.25$ and $\sigma_u = 0.2$.

The parameter θ_0 is set equal to 1, 2 and 3 and is estimated using the package "optimize" in R. We use the gaussian kernel and fix the bandwidth parameters as follows: $h_X = h_W = 0.1$, $h_V = 0.04$ and $h_\epsilon = 0.05$. Note that optimizing the bandwidth parameters in order to minimize the mean squared error should give better results but we believe this is beyond the scope of this paper. The Monte Carlo study has been performed with $mc = 500$ replications and a sample size $n = 100$. We provide each time the mean, the standard deviation and the mean squared error (mse hereafter) of $\widehat{\theta}$. We also provide the bias, the standard deviation and mse for the nonparametric estimator $\widehat{\phi}(x)$ evaluated at the median value of X. Moreover we also present the same results when the true value of V is used. The results are summarized in Table 1 and show that the method works well for reasonable sample size, that is the bias and variance are relatively small.

θ_0	mean($\widehat{\theta}$)	sd($\widehat{\theta}$)	mse($\widehat{\theta}$)	bias($\widehat{\phi}$)	sd($\widehat{\phi}$)	mse($\widehat{\phi}$)
1	0.94 (0.96)	0.69 (0.64)	0.48 (0.41)	0.09 (0.09)	0.44 (0.42)	0.20 (0.18)
2	1.91 (1.95)	0.76 (0.74)	0.58 (0.54)	0.06 (0.06)	0.36 (0.38)	0.14 (0.14)
3	2.89 (2.93)	0.81 (0.79)	0.66 (0.63)	0.05 (0.05)	0.33 (0.34)	0.11 (0.11)

Table 1: Simulation results for θ_0 and $\phi_0(x)$ evaluated at the median of X. The numbers between parentheses correspond to the values computed using the true control function V.

5.2 Bootstrap

Note that although the asymptotic limit of $n^{1/2}(\widehat{\theta} - \theta_0)$ is explicitly defined and has a simple normal distribution, it cannot be directly applied in practice, since the covariance matrix contains a number of unknown quantities, namely the parameter vector θ_0, the error density $f_{\epsilon(\theta_0)}$, its derivative $f'_{\epsilon(\theta_0)}$, the function ϕ_0 and the derivatives of these functions with respect to θ. Each of these functions can be estimated by a kernel estimator, by taking the appropriate derivative of the kernel estimator of ϕ_0 and of $f_{\epsilon(\theta_0)}$ given in (3.5) and (3.6). This approach leads (under suitable conditions on the bandwidths) to a consistent estimator of the asymptotic variance, by using similar results as in Lemma 7.1 (for ϕ_0 and its derivatives) and Lemma 7.2 (for $f_{\epsilon(\theta_0)}$ and its derivatives). However, we do not recommend to follow this approach in practice since some of these unknown quantities are hard to estimate and require the introduction of new smoothing parameters.

An alternative approach consists in approximating the variance, or even the whole distribution, of $\widehat{\theta}$ by means of a bootstrap procedure. The use of bootstrap techniques in the context of semiparametric inference has received a lot of attention in recent years. Chen, Linton and Van Keilegom (2003) propose a naive bootstrap procedure and give primitive conditions under which the bootstrap estimator converges to the same limit as the original estimator. Our estimator, which is a two-step semiparametric Z-estimator whose nuisance function depends on θ, is a special case of the general estimator considered in their setting. In a closely related context of one-step semiparametric M-estimation whose nuisance function is independent of θ, Cheng and Huang (2010), respectively Cheng (2015), proposed an exchangeable bootstrap scheme for approximating the distribution, respectively the moments, of $\widehat{\theta}$, whereas Cheng and Pillai (2012) proposed a model based bootstrap procedure. Finally, instead of using a bootstrap procedure, one could also make use of Bayesian inference techniques to approximate the distribution of a semiparametric estimator. We refer to Cheng

and Kosorok (2008) for more details.

Let us now focus on how the naive bootstrap proposed in Chen, Linton and Van Keilegom (2003) can be applied in our setting. Let $(X_i^*, Z_i^*, W_i^*, Y_i^*)$, $i = 1, \ldots, n$, be drawn randomly with replacement from the original data (X_i, Z_i, W_i, Y_i), $i = 1, \ldots, n$, and for any θ let $\widehat{\gamma}^{add,*} = (\widehat{\phi}_\theta^{add,*}, \dot{\widehat{\phi}}_\theta^{add,*}, \widehat{f}^*_{\epsilon(\theta)}, \widehat{f}'^*_{\epsilon(\theta)}, \dot{\widehat{f}}^*_{\epsilon(\theta)})^t$ be the same estimator as $\widehat{\gamma}_\theta^{add}$ but based on the bootstrap data. For each (θ, γ), define

$$M_n^*(\theta, \gamma) = n^{-1} \sum_{i=1}^n M(\theta, \gamma, X_i^*, Z_i^*, W_i^*, Y_i^*)$$

and define

$$\widehat{\theta}^* = \text{argmin}_{\theta \in \Theta} \left\| M_n^*(\theta, \widehat{\gamma}_\theta^{add,*}) \right\|.$$

Theorem B in Chen, Linton and Van Keilegom (2003) shows that under certain regularity conditions $n^{1/2}(\widehat{\theta}^* - \widehat{\theta})$ and $n^{1/2}(\widehat{\theta} - \theta_0)$ converge in distribution to the same normal limit. More precisely, using similar techniques as in the proof of Theorem 4.2, we conjecture that

$$n^{1/2}(\widehat{\theta}^* - \widehat{\theta}) = -\Gamma^{-1} n^{-1/2} \sum_{i=1}^n \left[A(X_i^*, Z_i^*, W_i^*, Y_i^*) - A(X_i, Z_i, W_i, Y_i) \right] + o_{P^*}(1), \quad (5.1)$$

where the function $A(X, Z, W, Y)$ is defined in (4.4) and where the $o_{P^*}(1)$-term goes to zero in probability, conditionally on the original data (X_i, Z_i, W_i, Y_i), $i = 1, \ldots, n$. From this claim together with the central limit theorem and Theorem 4.2 the result would follow. However, a detailed proof of (5.1) is beyond the scope of this paper, since it requires elaborate, lengthy and sophisticated calculations which are too space consuming. Instead we will check the validity of the proposed bootstrap procedure by means of a simulation study.

We continue to use the same model as in Subsection 5.1. For each sample of observations $(X_i, Y_i, W_i)_{i=1,\ldots,n}$ of size $n = 100$, we generate $B = 100$ bootstrapped samples $(X_i^*, Y_i^*, W_i^*)_{i=1,\ldots,n}$ of the same size, drawn randomly with replacement from the original data. Then, from these bootstrapped samples, B estimators $(\widehat{\theta}^{b,*})_{b=1,\ldots,B}$ are computed as well as the mean and the variance of these B bootstrapped estimators. We simulate $mc = 100$ initial samples $(X_i, Y_i, W_i)_{i=1,\ldots,n}$ in order to obtain a total of mc bootstrapped means and bootstrapped variances. At last, we provide the histograms of these bootstrapped means and bootstrapped variances for different values of θ_0 (see Figures 1, 2 and 3). In order to provide an empirical proof of the validity of our bootstrap procedure, we check that each histogram is centered around the mean and the variance of the 100 estimated values of θ_0. This is indeed the case for each of the 6 figures which therefore suggests that the proposed bootstrap procedure works well in practice.

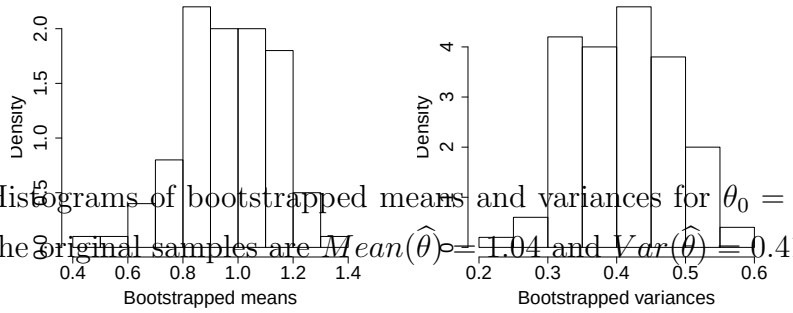

Figure 1: Histograms of bootstrapped means and variances for $\theta_0 = 1$. The corresponding values for the original samples are $Mean(\widehat{\theta}) = 1.04$ and $Var(\widehat{\theta}) = 0.41$.

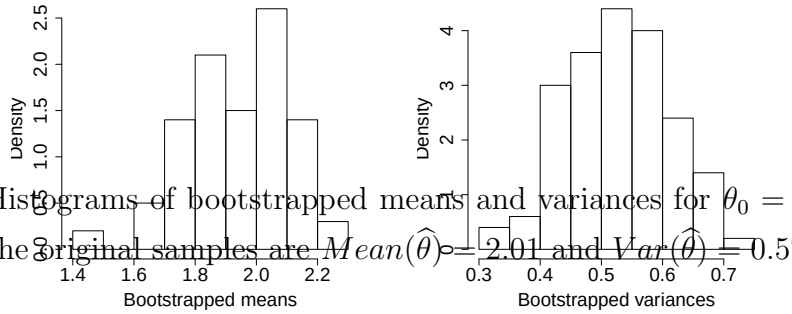

Figure 2: Histograms of bootstrapped means and variances for $\theta_0 = 2$. The corresponding values for the original samples are $Mean(\widehat{\theta}) = 2.01$ and $Var(\widehat{\theta}) = 0.57$.

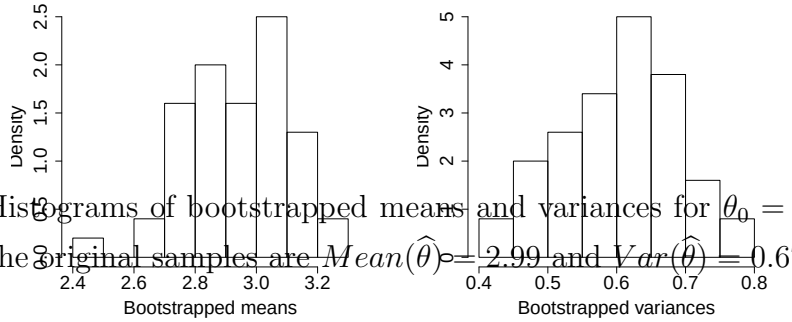

Figure 3: Histograms of bootstrapped means and variances for $\theta_0 = 3$. The corresponding values for the original samples are $Mean(\widehat{\theta}) = 2.99$ and $Var(\widehat{\theta}) = 0.67$.

5.3 Real data analysis

We conclude this finite sample study by considering the estimation of Engel curves based on the UK Family Expenditure Survey as in Blundell, Chen and Kristensen (2007). The Engel

curve relationship describes the expansion path for commodity demands as the household's budget increases. The motivation for a control function approach derives from the endogeneity of the total budget variable. As total expenditure is endogenous for individual commodity demands, we use gross earnings of the household head as an instrument (see Blundell, Chen and Kristensen 2007 for a detailed discussion). In this application, we consider a single year of study, 1995, and 3 broad categories of nondurables and services: (1) leisure goods and services, (2) travel and (3) household goods and services. To preserve some demographic homogeneity, we consider couples where the head of household is aged between 20 and 55 and at work and among them select a subset of couples with 3 children. We first present some descriptive statistics for this subsample in Table 2.

	Mean	Sd.
Leisure goods	0.129	0.105
Travel	0.190	0.098
Household goods	0.114	0.085
log nondurable expenditure	5.810	0.637
log gross earnings	5.769	0.644
Sample size	294	

Table 2: Data descriptives

The objective is to estimate the model defined in (1.1) where Y represents a budget share (leisure, travel or household) and X the log of nondurable expenditure. There is no exogenous variable Z in the application. The instrumental variable W used to identify and estimate the model is the log of gross earnings. The operator Λ_θ is chosen as the Box-Cox transformation. The control variable V is identified as the conditional distribution of X given W and the bandwidth parameters are fixed as follows: $h_X = h_W = 0.5$, $h_V = 0.02$ and $h_\epsilon = 0.3$. The same remark as in Subsection 5.1 applies, that is optimizing the bandwidth parameters in order to minimize the mse should give better results but this is beyond the scope of this paper. Figure 4 presents the estimated curves of ϕ_0 for the three goods and the corresponding 95% pointwise confidence bands obtained using the naive bootstrap described above and based on 100 resamples. The results for the estimation of θ_0 are presented in Table 3 with the values of the mean and the standard deviation obtained by the same bootstrap procedure. The results show small standard deviations and relatively small confidence intervals.

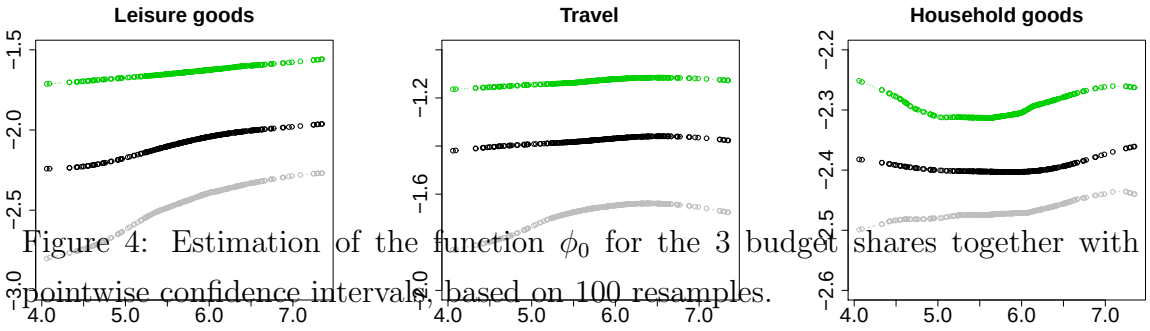

Figure 4: Estimation of the function ϕ_0 for the 3 budget shares together with the 95% pointwise confidence intervals, based on 100 resamples.

	$\widehat{\theta}$	Mean$(\widehat{\theta}^*)$	Sd$(\widehat{\theta}^*)$
Leisure goods	0.120	0.107	0.087
Travel	0.303	0.314	0.126
Household goods	0.001	0.003	0.012

Table 3: Estimation of θ_0 for the 3 budget shares together with bootstrapped means and standard deviations based on 100 resamples.

6 Conclusion

In this work we have studied a semiparametric transformation model with a parametric transformation operator Λ_θ, a nonparametric regression function ϕ and some endogenous explanatory variables. We use a control function approach to identify the nonparametric structure $(\Lambda, \phi, F_\epsilon)$. A profiling method is proposed to estimate the parametric transformation, and by imposing an additive structure on the function ϕ, we showed the asymptotic normality of the proposed estimator with $\sqrt{n}$ rate of convergence. Some finite sample simulations confirm the validity of our method. Finally, we illustrated our method using data from the UK Family Expenditure Survey.

7 Appendix : Proofs

In this Appendix we first prove in Subsection 7.1 the identification of the model stated in Proposition 2.1. Next, in Subsection 7.2 we state the conditions under which the asymptotic

results of Section 4 are valid. To prove these asymptotic results we need a number of lemmas, which are proved in Subsection 7.3. Finally, Subsections 7.4 and 7.5 contain the proofs of the consistency and asymptotic normality results, respectively.

7.1 Identification

Proof of Proposition 2.1. To prove identification of the structure $(\Lambda, \phi, F_\epsilon)$, we proceed in two steps: we first establish identification of Λ and then prove that ϕ and F_ϵ are identified.

1. *Identification of* Λ. This first step is inspired by the proof of Chiappori, Komunjer and Kristensen (2010). Under the regularity assumptions $(A.3)$ and $(A.4)$, we can differentiate equation (2.1) with respect to y and x_1 (the first coordinate of x) to obtain:

$$\begin{aligned}\frac{\partial}{\partial y}\int F_{Y|X,Z,V}(y|x,z,v)F_V(dv) &= f_\epsilon(\Lambda(y)-\phi(x,z)).\Lambda'(y)\\ \frac{\partial}{\partial x_1}\int F_{Y|X,Z,V}(y|x,z,v)F_V(dv) &= -f_\epsilon\left(\Lambda(y)-\phi\left(x,z\right)\right).\frac{\partial}{\partial x_1}\phi(x,z).\end{aligned}$$

Let $A = \{(x,z) \in R_{X,Z} : \frac{\partial}{\partial x_1}\int F_{Y|X,Z,V}(y|x,z,v)F_V(dv) \neq 0 \text{ for every } y \in R_Y\}$. Under Assumptions $(A.4)$ and $(A.5)$, the set A has a nonempty interior. Then, for any point $(x,z) \in A$ and for every $y \in R_Y$, we have:

$$-\frac{\Lambda'(y)}{\frac{\partial}{\partial x_1}\phi(x,z)} = s(y,x,z),$$

where $s(y,x,z) = \frac{\frac{\partial}{\partial y}\int F_{Y|X,Z,V}(y|x,z,v)F_V(dv)}{\frac{\partial}{\partial x_1}\int F_{Y|X,Z,V}(y|x,z,v)F_V(dv)}$. Note that $s(y,x,z)$ is non zero and keeps a constant sign for all $y \in R_Y$. Integrating from 0 to y and under Assumption $(A.6)$ we get:

$$\Lambda(y) = -\frac{\partial}{\partial x_1}\phi(x,z).S(y,x,z),$$

where $S(y,x,z) = \int_0^y s(t,x,z)dt$. Again, $S(y,x,z)$ is nonzero and keeps a constant sign for all $y \in R_Y$. Hence, $\mathrm{E}[S(Y,x,z)] \neq 0$. Using again Assumption $(A.6)$ we get:

$$\frac{\partial}{\partial x_1}\phi(x,z) = -\frac{1}{\mathrm{E}[S(Y,x,z)]},$$

and finally we obtain that:

$$\Lambda(y) = \frac{S(y,x,z)}{\mathrm{E}[S(Y,x,z)]}.$$

Hence, Λ is identified.

2. *Identification of ϕ and F_ϵ.* The identification of ϕ is a direct consequence of Assumptions $(A.1)$ and $(A.2)$ following Imbens and Newey (2009). Identification of F_ϵ eventually follows from equation (2.1). This finishes the proof. □

7.2 Assumptions

For any $\ell \geq 1$ we let $\frac{\partial}{\partial e_\ell}$ denote the derivative with respect to the ℓth argument of a vector e, ∇_e denotes the gradient with respect to the vector e, and ∇_e^t is its transpose. At last, we denote by $\dot{m}_\theta(x,z,v)$ the vector of partial derivatives of $m_\theta(x,z,v)$ with respect to the components of θ. The following regularity conditions are required for the asymptotic results:

(C.1) For $j = 1, 2$, k_j is a symmetric kernel of order $q_j \geq 4$, i.e. $\int u^m k_j(u)\,du = 0$ for $m = 1, \ldots, q_j - 1$ and $\int u^{q_j} k_j(u)\,du \neq 0$. Moreover, k_j has compact support and is twice continuously differentiable, and q_1 satisfies $q_1 > 2d_z + d_w + d_x + d_v + 1$.

(C.2) $nh^{4d_z+2d_w+2d_x+2d_v+2} \to \infty$, $nh^{2q_1} \to 0$, $ng^6(\log g^{-1})^{-2} \to \infty$ and $ng^{2q_2} \to 0$, where q_1 and q_2 are defined in condition (C.1).

(C.3) The density $f_{X,Z,V}$ exists and is bounded away from zero and infinity. Moreover, $f_{X,Z,V}$ is Lipschitz continuous and has a compact support $R_{X,Z,V}$.

(C.4) $m_\theta(x,z,v)$, $\dot{m}_\theta(x,z,v)$ and $\nabla_v m_\theta(x,z,v)$ exist and are q_1 times continuously differentiable with respect to the components of x, z and v on $R_{X,Z,V} \times \Theta$. In addition, all derivatives up to order q_1 are bounded, uniformly in (x,z,v,θ) in $R_{X,Z,V} \times \Theta$.

(C.5) $f_{Z,W}(z,w)$ and $F_{X|Z,W}(x|z,w)$ exist and are q_1 times continuously differentiable with respect to the components of z and w on $R_{Z,W}$. In addition, all derivatives up to order q_1 are bounded, uniformly in $(x,z,w) \in R_{X,Z,W}$, and $f_{Z,W}(z,w)$ is bounded away from zero, uniformly in z and w.

(C.6) $\Lambda_\theta(y)$ is three times continuously differentiable with respect to y and θ, and there exists a $\delta > 0$ such that

$$\mathrm{E}\Big[\sup_{\|\theta'-\theta\|\leq\delta}\Big|\frac{\partial^{k+l}}{\partial y^k \partial\theta_1^{l_1}\ldots\partial\theta_p^{l_p}}\Lambda_{\theta'}(Y)\Big|\Big] < \infty$$

for all θ in Θ and for all k and l such that $0 \leq k + l \leq 3$, where $l = l_1 + \ldots + l_p$ and $\theta = (\theta_1, \ldots, \theta_p)^t$. Moreover,

$$\sup_{\theta\in\Theta}\mathrm{E}\Big\|\dot{\Lambda}_\theta(Y)\Big\|^2 < \infty.$$

(C.7) $F_{\epsilon(\theta)}(y)$ is three times continuously differentiable with respect to y and θ, and

$$\sup_{\theta,y} \Big| \frac{\partial^{k+l}}{\partial y^k \partial \theta_1^{l_1} ... \partial \theta_p^{l_p}} F_{\epsilon(\theta)}(y) \Big| < \infty$$

for all k and l such that $0 \leq k + l \leq 2$, where $l = l_1 + ... + l_p$ and $\theta = (\theta_1, ..., \theta_p)^t$.

(C.8) $\forall \epsilon > 0,\ \exists \delta(\epsilon) > 0$ such that $\| \theta - \theta_0 \| > \epsilon$ implies

$$H(\theta_0, f_{\epsilon(\theta_0)}, \phi_0) - H(\theta, f_{\epsilon(\theta)}, \phi_\theta^{add}) > \delta(\epsilon)$$

(C.9) The control function V_i and its estimate $\widehat{V}_i$ satisfy

$$\widehat{V}_i - V_i = \Big(n^{-1} \sum_{k=1}^{n} B_{ik} \Big)(1 + R_i),$$

where $(B_{ik})_{k=1,...,n}$ have the same dimension as V_i and

$$B_{ik} = Q(X_i, X_k, Z_i, W_i) K_h(Z_i - Z_k) K_h(W_i - W_k)$$

for some bounded function Q,

$$\max_{1 \leq i,k \leq n} \Big\| \mathrm{E}(B_{ik} | Z_k, W_k, X_i, Z_i, W_i) \Big\| = O_P(h^{q_1}),$$

and R_i is the residual term of dimension 1 such that $\max_{1 \leq i \leq n} |R_i| = o_P(1)$.

(C.10) The matrix Γ is of full rank.

Remark 7.1. *Conditions (C.1)–(C.7) are quite similar to the assumptions in Linton, Sperlich and Van Keilegom (2008). Condition (C.8) is needed to identify the true parameter θ_0. It is taken from the paper of Delsol and Van Keilegom (2014) on which our consistency proof is based. Also note that, contrary to other papers in the literature, we explicitly show the consistency of $\widehat{\theta}$. At last, condition (C.9) gives high level conditions for the convergence of the generated regressor $\widehat{V}$ to V, which is required to prove the consistency and the rate of convergence of $\widehat{\theta}$.*

Let us check briefly that condition (C.9) is satisfied for the estimator $\widehat{V}_i$ defined in (3.4).

We have:

$$
\begin{aligned}
&\widehat{V}_i - V_i \\
&= \frac{\sum_{k=1}^n \Big[1(X_k \le X_i) - F_{X|ZW}(X_i|Z_i,W_i)\Big] K_h(Z_i - Z_k) K_h(W_i - W_k)}{\sum_{k=1}^n K_h(Z_i - Z_k) K_h(W_i - W_k)} \\
&= \frac{n^{-1}\sum_{k=1}^n \Big[1(X_k \le X_i) - F_{X|ZW}(X_i|Z_i,W_i)\Big] K_h(Z_i - Z_k) K_h(W_i - W_k)}{f_{ZW}(Z_i,W_i)} \\
&\quad + O_P((nh^{d_z+d_w})^{-1}) + O(h^{2q_1}) \\
&:= \Big(n^{-1}\sum_{k=1}^n B_{ik}\Big)(1 + o_P(1)).
\end{aligned}
$$

It can be shown that $\mathrm{E}(B_{ik}|Z_k, W_k, X_i, Z_i, W_i) = O_P(h^{q_1})$ *uniformly in* i *and* k *which proves the result.*

7.3 Some useful lemmas

We first start this subsection by presenting a few lemmas that will be useful to prove both the consistency and the asymptotic normality result.

From now on, in order to simplify the notations, we consider $S = (X, Z)$ and $d_s = d_x + d_z$. The following lemma gives an i.i.d. representation of the estimators $\widehat{\phi}_\theta^{add}(s)$ and $\widehat{\dot{\phi}}_\theta^{add}(s)$, uniformly in θ and s, and will be a key ingredient for obtaining the asymptotic limit of our estimator $\widehat{\theta}$.

Lemma 7.1. *Assume (A.1)–(A.6) and (C.1)–(C.9). Then, using the abbreviated notation* $S = (X, Z)$ *and* $s = (x, z)$*, we have*

$$
\begin{aligned}
&\widehat{\phi}_\theta^{add}(s) - \phi_\theta^{add}(s) \\
&= n^{-1}\sum_{i=1}^n \Bigg(\sum_{\alpha=1}^{d_s} k_{1h}(s_\alpha - S_{\alpha i}) \Big[\Lambda_\theta(Y_i) - m_\theta(S_i, V_i)\Big] f^{-1}_{S_\alpha|S_{-\alpha},V}(S_{\alpha i}|S_{-\alpha i}, V_i) \\
&\quad + \sum_{\alpha=1}^{d_x} \mathrm{E}_{X_{-\alpha}}\Big[\nabla_v^t\Big\{\frac{\mathrm{E}(\Lambda_\theta(Y)|S,W) - m_\theta(S,V)}{f_{S_\alpha|S_{-\alpha}V}(s_\alpha|S_{-\alpha},V)}\Big\} f_{S_\alpha ZW|X_{-\alpha}}(s_\alpha, Z, W|X_{-\alpha}) \\
&\qquad \times Q(X, S_i, W_i)\Big| S_\alpha = s_\alpha, Z = Z_i, W = W_i\Big] \\
&\quad + \sum_{\alpha=d_x+1}^{d_s} \mathrm{E}_X\Big[\nabla_v^t\Big\{\frac{\mathrm{E}(\Lambda_\theta(Y)|S,W) - m_\theta(S,V)}{f_{S_\alpha|S_{-\alpha}V}(s_\alpha|S_{-\alpha},V)}\Big\} f_{ZW|X}(Z,W|X) f_{S_\alpha|XZ_{-\alpha}W}(s_\alpha|X, Z_{-\alpha}, W) \\
&\qquad \times Q(X, s_\alpha, S_{-\alpha i}, W_i)\Big| S_\alpha = s_\alpha, Z_{-\alpha} = Z_{-\alpha i}, W = W_i\Big]
\end{aligned}
$$

$$
\begin{aligned}
&+\sum_{\alpha=1}^{d_s} \mathrm{E}_X\Big[\nabla_v^t m_\theta(s_\alpha, S_{-\alpha}, V) Q(X, S_i, W_i)\Big| Z = Z_i, W = W_i\Big] f_{ZW}(Z_i, W_i) \\
&+\Big[\sum_{\alpha=1}^{d_s} m_\theta(s_\alpha, S_{-\alpha i}, V_i) - (d_s - 1)\Lambda_\theta(Y_i) - \phi_\theta^{add}(s)\Big]\Big) \\
&+o_P(n^{-1/2}),
\end{aligned}
$$

uniformly in $s \in R_{X,Z}$ *and* $\theta \in \Theta$. *The i.i.d. representation for* $\dot{\widehat{\phi}}_\theta^{add}(s) - \dot{\phi}_\theta^{add}(s)$ *is obtained by replacing* Λ_θ, m_θ *and* ϕ_θ^{add} *in the above representation by respectively* $\dot{\Lambda}_\theta$, $\dot{m}_\theta$ *and* $\dot{\phi}_\theta^{add}$.

Proof of Lemma 7.1. We restrict attention to proving the first result of Lemma 7.1, since the second one can be shown in a very similar way. We first decompose $\widehat{\phi}_\theta^{add}(s) - \phi_\theta^{add}(s)$ as follows:

$$
\begin{aligned}
&\widehat{\phi}_\theta^{add}(s) - \phi_\theta^{add}(s) \\
&= \frac{1}{n}\sum_{i=1}^{n}\left[\sum_{\alpha=1}^{d_s} \widehat{m}_\theta(s_\alpha, S_{-\alpha i}, \widehat{V}_i) - (d_s - 1)\Lambda_\theta(Y_i)\right] - \phi_\theta^{add}(s) \\
&= \frac{1}{n}\sum_{i=1}^{n}\left[\sum_{\alpha=1}^{d_s} m_\theta(s_\alpha, S_{-\alpha i}, V_i) - (d_s - 1)\Lambda_\theta(Y_i) - \phi_\theta^{add}(s)\right] \\
&\quad + \sum_{\alpha=1}^{d_s}\frac{1}{n}\sum_{i=1}^{n}\Big[\widehat{m}_\theta(s_\alpha, S_{-\alpha i}, \widehat{V}_i) - m_\theta(s_\alpha, S_{-\alpha i}, V_i)\Big] \\
&= R_1(s) + \sum_{\alpha=1}^{d_s} R_2^\alpha(s).
\end{aligned}
$$

Then, using a Taylor expansion on $R_2^\alpha(s)$, we have:

$$
\begin{aligned}
R_2^\alpha(s) &= \frac{1}{n}\sum_{i=1}^{n}\left(\widehat{m}_\theta - m_\theta\right)(s_\alpha, S_{-\alpha i}, \widehat{V}_i) + \frac{1}{n}\sum_{i=1}^{n}\Big(m_\theta(s_\alpha, S_{-\alpha i}, \widehat{V}_i) - m_\theta(s_\alpha, S_{-\alpha i}, V_i)\Big) \\
&= \frac{1}{n}\sum_{i=1}^{n}\left(\widehat{m}_\theta - m_\theta\right)(s_\alpha, S_{-\alpha i}, V_i) + \frac{1}{n}\sum_{i=1}^{n}\nabla_v^t m_\theta(s_\alpha, S_{-\alpha i}, V_i)(\widehat{V}_i - V_i) \\
&\quad + \frac{1}{n}\sum_{i=1}^{n}\nabla_v^t\left(\widehat{m}_\theta - m_\theta\right)(s_\alpha, S_{-\alpha i}, \xi_i)(\widehat{V}_i - V_i) \\
&\quad + \frac{1}{2n}\sum_{i=1}^{n}(\widehat{V}_i - V_i)^t \nabla_{vv} m_\theta(s_\alpha, S_{-\alpha i}, \xi_i')(\widehat{V}_i - V_i) \\
&= R_{21}^\alpha(s) + R_{22}^\alpha(s) + R_{23}^\alpha(s) + R_{24}^\alpha(s),
\end{aligned}
$$

where $\xi_i = \lambda_i V_i + (1 - \lambda_i)\widehat{V}_i$ for some $\lambda_i \in [0, 1]$, $\xi_i' = \lambda_i' V_i + (1 - \lambda_i')\widehat{V}_i$ for $\lambda_i' \in [0, 1]$, $(\widehat{V}_i - V_i)^t$ is the transpose of the vector $\widehat{V}_i - V_i$, $\nabla_v m_\theta$ represents the gradient of m_θ, i.e. the vector

of partial derivatives of m_θ with respect to the components of v, $\nabla_v^t m_\theta$ its transpose, and $\nabla_{vv} m_\theta$ is the Hessian matrix.

In what follows we concentrate on $R_{21}^\alpha(s)$ and $R_{22}^\alpha(s)$, since it is easily seen that $R_{23}^\alpha(s)$ and $R_{24}^\alpha(s)$ are of lower order.

We start with $R_{21}^\alpha(s)$. Write

$$\begin{aligned} R_{21}^\alpha(s) &= \frac{1}{n}\sum_{i=1}^n (\widehat{m}_\theta - \widetilde{m}_\theta)(s_\alpha, S_{-\alpha i}, V_i) + \frac{1}{n}\sum_{i=1}^n (\widetilde{m}_\theta - m_\theta)(s_\alpha, S_{-\alpha i}, V_i) \\ &= R_{211}^\alpha(s) + R_{212}^\alpha(s), \end{aligned}$$

where

$$\widetilde{m}_\theta(s,v) = \frac{\sum_{i=1}^n \Lambda_\theta(Y_i) K_h(s - S_i) K_h(v - V_i)}{\sum_{i=1}^n K_h(s - S_i) K_h(v - V_i)},$$

i.e. with respect to $\widehat{m}_\theta(s,v)$ we have replaced the $\widehat{V}_i$'s by the true (but unknown) V_i's. The term $R_{212}^\alpha(s)$ can be worked out similarly as in e.g. Linton and Nielsen (1995), since this is the ordinary marginal integration estimator. Hence, this term equals

$$n^{-1}\sum_{i=1}^n \Big[\Lambda_\theta(Y_i) - m_\theta(S_i, V_i)\Big] k_{1h}(s_\alpha - S_{\alpha i}) f^{-1}_{S_\alpha | S_{-\alpha}, V}(S_{\alpha i} | S_{-\alpha i}, V_i) + o_P(n^{-1/2}).$$

Now consider

$$(\widehat{m}_\theta - \widetilde{m}_\theta)(s_\alpha, S_{-\alpha i}, V_i) = \frac{\sum_{j=1}^n \widehat{N}_{ij}}{\sum_{j=1}^n \widehat{D}_{ij}} - \frac{\sum_{j=1}^n \widetilde{N}_{ij}}{\sum_{j=1}^n \widetilde{D}_{ij}},$$

where $\widehat{N}_{ij} = \Lambda_\theta(Y_j) k_{1h}(s_\alpha - S_{\alpha j}) K_h(S_{-\alpha i} - S_{-\alpha j}) K_h(V_i - \widehat{V}_j)$, $\widehat{D}_{ij} = k_{1h}(s_\alpha - S_{\alpha j}) K_h(S_{-\alpha i} - S_{-\alpha j}) K_h(V_i - \widehat{V}_j)$, and similarly for $\widetilde{N}_{ij}$ and $\widetilde{D}_{ij}$. In analogy with these notations, we define $N_i = \mathrm{E}(\Lambda_\theta(Y) | s_\alpha, S_{-\alpha i}, V_i) f_{S,V}(s_\alpha, S_{-\alpha i}, V_i)$ and $D_i = f_{S,V}(s_\alpha, S_{-\alpha i}, V_i)$. In order to simplify the notation, we have omitted the dependence on θ and s_α, but of course it will be a crucial point in the proof. Next, write

$$\begin{aligned} R_{211}^\alpha(s) &= n^{-1}\sum_{i=1}^n\sum_{j=1}^n (\widehat{N}_{ij} - \widetilde{N}_{ij}) \frac{1}{\sum_{j=1}^n \widehat{D}_{ij}} + n^{-1}\sum_{i=1}^n\sum_{j=1}^n \widetilde{N}_{ij}\Big(\frac{1}{\sum_{j=1}^n \widehat{D}_{ij}} - \frac{1}{\sum_{j=1}^n \widetilde{D}_{ij}}\Big) \\ &= \Big[n^{-2}\sum_{i=1}^n\sum_{j=1}^n (\widehat{N}_{ij} - \widetilde{N}_{ij})\frac{1}{D_i} - n^{-2}\sum_{i=1}^n\sum_{j=1}^n (\widehat{D}_{ij} - \widetilde{D}_{ij})\frac{N_i}{D_i^2}\Big](1 + o_P(1)), \end{aligned}$$

where the $o_P(1)$ term is uniform in s_α (from assumption (C.3)) and in θ. The latter equals

$$\begin{aligned} &\Big[n^{-2}\sum_{i=1}^n D_i^{-1}\sum_{j=1}^n \Big\{\nabla_v^t \widetilde{N}_{ij} - \nabla_v^t \widetilde{D}_{ij}\frac{N_i}{D_i}\Big\}(V_j - \widehat{V}_j)\Big](1 + o_P(1)) \\ &= -\Big[n^{-3}\sum_{i=1}^n D_i^{-1}\sum_{j=1}^n\sum_{k=1}^n \Big\{\nabla_v^t \widetilde{N}_{ij} - \nabla_v^t \widetilde{D}_{ij}\frac{N_i}{D_i}\Big\} B_{jk}\Big](1 + o_P(1)), \end{aligned} \tag{7.1}$$

where the $o_P(1)$ term is again uniform in s_α and θ, and where $\nabla_v \widetilde{N}_{ij} = \Lambda_\theta(Y_j) k_{1h}(s_\alpha - S_{\alpha j}) K_h(S_{-\alpha i} - S_{-\alpha j}) h^{-1} \nabla_v K_h(V_i - V_j)$, $\nabla_v \widetilde{D}_{ij} = k_{1h}(s_\alpha - S_{\alpha j}) K_h(S_{-\alpha i} - S_{-\alpha j}) h^{-1} \nabla_v K_h(V_i - V_j)$, and

$$\widehat{V}_j - V_j = \Big(n^{-1} \sum_{k=1}^{n} B_{jk} \Big)(1 + o_P(1)),$$

uniformly in $1 \leq j \leq n$, by condition (C.9). Ignoring the factor $(1 + o_P(1))$, (7.1) is a V-process of order three depending on s_α, θ and h, which can be rewritten as:

$$n^{-3} \sum_{i=1}^{n} \sum_{j=1}^{n} \sum_{k=1}^{n} q(T_i, T_j, T_k, s_\alpha, \theta, h)$$

where $T_i = (X_i, Z_i, W_i, Y_i)^t$ and

$$q(T_i, T_j, T_k, s_\alpha, \theta, h) = -D_i^{-1} \Big\{ \nabla_v^t \widetilde{N}_{ij} - \nabla_v^t \widetilde{D}_{ij} \frac{N_i}{D_i} \Big\} B_{jk}.$$

We denote $p(T_i, T_j, T_k, s_\alpha, \theta, h) = h^{2d_z + d_w + d_x + d_v + 1} q(T_i, T_j, T_k, s_\alpha, \theta, h)$ and consider the following V-process:

$$\mathcal{V}_n(s_\alpha, \theta, h) = n^{-3} \sum_{i=1}^{n} \sum_{j=1}^{n} \sum_{k=1}^{n} p(T_i, T_j, T_k, s_\alpha, \theta, h).$$

Since a V-process can be written as a U-process plus negligible terms, following Sherman (1994), we introduce the associated U-process $\mathcal{U}_n(s_\alpha, \theta, h)$ which can be decomposed as (see equation (6) on page 449 in Sherman):

$$\begin{aligned}
&\mathcal{U}_n(s_\alpha, \theta, h) \\
&= \frac{1}{n(n-1)(n-2)} \sum_{i,j,k \neq} p(T_i, T_j, T_k, s_\alpha, \theta, h) \\
&= n^{-1} \sum_{i=1}^{n} \mathrm{E}[p(T_i, T, T', s_\alpha, \theta, h) | T_i] + n^{-1} \sum_{j=1}^{n} \mathrm{E}[p(T, T_j, T', s_\alpha, \theta, h) | T_j] \\
&\quad + n^{-1} \sum_{k=1}^{n} \mathrm{E}[p(T, T', T_k, s_\alpha, \theta, h) | T_k] - 2\mathrm{E}[p(T, T', T'', s_\alpha, \theta, h)] + \mathcal{R}_n(s_\alpha, \theta, h), \qquad (7.2)
\end{aligned}$$

where T, T', T'' are i.i.d. and have the same distribution as $T_1, ..., T_n$. The last term $\mathcal{R}_n(s_\alpha, \theta, h)$ is by construction the sum of two degenerate U-processes, one of order 2, denoted by $\mathcal{R}_{n2}(s_\alpha, \theta, h)$, and one of order 3, denoted by $\mathcal{R}_{n3}(s_\alpha, \theta, h)$. In what follows, we concentrate on $\mathcal{R}_{n2}(s_\alpha, \theta, h)$, which will be dominant. In order to control uniformly in s_α, θ and

h the term $\mathcal{R}_{n2}(s_\alpha, \theta, h)$, we will apply Corollary 4 in Sherman (1994). Let us first introduce some notations. We define the following functional class associated to the U-process $\mathcal{U}_n(s_\alpha, \theta, h)$:

$$\mathcal{F} = \{(t, t', t'') \to p(t, t', t'', s_\alpha, \theta, h) : s_\alpha \in R_{S_\alpha}, \theta \in \Theta, h > 0\}.$$

In order to apply Corollary 4, we need to check that $\mathcal{F}$ is Euclidean (see Sherman 1994 or Pakes and Pollard 1989 for a precise definition). Using conditions (C.5), (C.6) and (C.9), and Lemma 2.14 and Example 2.10 in Pakes and Pollard (1989), it follows that $\mathcal{F}$ is Euclidean and so is the class of functions associated to $\mathcal{R}_{n2}(s_\alpha, \theta, h)$ (see Lemma 6 in Sherman 1994). Then, using Corollary 4 in Sherman (1994), it follows that

$$\sup_{s_\alpha, \theta, h} |\mathcal{R}_{n2}(s_\alpha, \theta, h)| = O_P(n^{-1})$$

and hence, using Assumption (C.2),

$$\begin{aligned} h_n^{-(2d_z+d_w+d_x+d_v+1)} \sup_{s_\alpha, \theta} |\mathcal{R}_{n2}(s_\alpha, \theta, h_n)| &= O_P(n^{-1} h_n^{-(2d_z+d_w+d_x+d_v+1)}) \\ &= o_P(n^{-1/2}), \end{aligned}$$

where h_n denotes (here) the smoothing parameter associated to the sample size n (in order to make the distinction with the parameter h of the U-process). Let us now go back to the first term on the right hand side of equation (7.2) evaluated at $h = h_n$:

$$n^{-1} \sum_{i=1}^{n} \mathrm{E}[p(T_i, T, T', s_\alpha, \theta, h_n)|T_i] := n^{-1} h_n^{2d_z+d_w+d_x+d_v+1} \sum_{i=1}^{n} \mathrm{E}[q(T_i, T, T', s_\alpha, \theta, h_n)|T_i].$$

By definition, we have:

$$n^{-1} \sum_{i=1}^{n} \mathrm{E}[q(T_i, T_j, T_k, s_\alpha, \theta, h_n)|T_i] = -n^{-1} \sum_{i=1}^{n} D_i^{-1} \mathrm{E}\Big[\Big\{\nabla_v^t \widetilde{N}_{ij} - \nabla_v^t \widetilde{D}_{ij} \frac{N_i}{D_i}\Big\} B_{jk} \Big| T_i\Big].$$

From condition (C.9) we know that $\|E(B_{jk}|T_j)\| = O_P(h_n^{q_1})$ uniformly in j. Then, it easily follows that

$$n^{-1} \sum_{i=1}^{n} \mathrm{E}[q(T_i, T, T', s_\alpha, \theta, h_n)|T_i] = O_P(h_n^{q_1}) = o_P(n^{-1/2}),$$

since $n h_n^{2q_1} \to 0$ and by using assumptions (C.3), (C.5) and (C.6). In reality the order is even smaller than $O_P(h_n^{q_1})$, but it is not necessary to do a more detailed order calculation, since we reach already the required $o_P(n^{-1/2})$-rate based on this simple argument. In a similar

way we can show the order of the second term on the right hand side of (7.2) (with p replaced by q) :

$$n^{-1}\sum_{j=1}^{n}\mathrm{E}[q(T,T_j,T',s_\alpha,\theta,h_n)|T_j] = O_P(h_n^{q_1}) = o_P(n^{-1/2}).$$

For the third term more work is needed. It is easily seen that

$$\begin{aligned}
&\mathrm{E}[q(T,T_j,T_k,s_\alpha,\theta,h_n)|T_j,T_k]\\
&= \nabla_v^t\Big\{\frac{\Lambda_\theta(Y_j)-\mathrm{E}(\Lambda_\theta(Y)|s_\alpha,S_{-\alpha j},V_j)}{f_{S_\alpha|S_{-\alpha}V}(s_\alpha|S_{-\alpha j},V_j)}\Big\}k_{1h}(s_\alpha-S_{\alpha j})K_h(Z_j-Z_k)K_h(W_j-W_k)\\
&\quad\times Q(X_j,X_k,Z_j,W_j)+o_P(n^{-1/2}) \qquad (7.3)
\end{aligned}$$

uniformly in j, k, s_α and θ. The calculation of the expected value of (7.3) with respect to T_j depends on the value of α. In fact, when $\alpha = 1,\ldots,d_x$, $S_{\alpha j} = X_{\alpha j}$ and so the variables in the product $k_{1h}(s_\alpha - S_{\alpha j})K_h(Z_j - Z_k)$ appearing in (7.3) are d_z+1 different variables. However, when $\alpha = d_x+1,\ldots,d_s$, then we have only d_z different variables, one of the components of Z_j appearing in fact twice. This has an impact on the expected value. For $\alpha = 1,\ldots,d_x$, it is easily shown that

$$\begin{aligned}
&n^{-1}\sum_{k=1}^{n}\mathrm{E}[q(T,T',T_k,s_\alpha,\theta,h_n)|T_k]\\
&= E_{X_{-\alpha}}\Big[\nabla_v^t\Big\{\frac{E(\Lambda_\theta(Y)|s_\alpha,S_{-\alpha},W)-\mathrm{E}(\Lambda_\theta(Y)|s_\alpha,S_{-\alpha},V)}{f_{S_\alpha|S_{-\alpha}V}(s_\alpha|S_{-\alpha},V)}f_{S_\alpha ZW|X_{-\alpha}}(s_\alpha,Z,W|X_{-\alpha})\Big\}\\
&\quad\times Q(s_\alpha,X_{-\alpha},S_k,W_k)\Big|S_\alpha=s_\alpha,Z=Z_k,W=W_k\Big]+o_P(n^{-1/2}),
\end{aligned}$$

uniformly in s_α and θ. The derivation for $\alpha = d_x+1,\ldots,d_s$ can be done in a similar manner. Finally, it follows from the above calculations that $\mathrm{E}[q(T,T',T'',s_\alpha,\theta,h_n)] = o(n^{-1/2})$ uniformly in s_α and θ. This finishes the calculation of $R_{211}^\alpha(s)$, and hence of $R_{21}^\alpha(s)$.

Next, consider $R_{22}^\alpha(s)$. Using again Sherman (1994)'s result on degenerate U-processes, we can prove in a very similar way as for $R_{21}^\alpha(s)$ that

$$\begin{aligned}
R_{22}^\alpha(s) &= n^{-1}\sum_{i=1}^{n}\nabla_v^t m_\theta(s_\alpha,S_{-\alpha i},V_i)(\widehat{V}_i-V_i)\\
&= \Big\{n^{-2}\sum_{i=1}^{n}\sum_{k=1}^{n}\nabla_v^t m_\theta(s_\alpha,S_{-\alpha i},V_i)B_{ik}\Big\}(1+o_P(1))
\end{aligned}$$

$$
\begin{aligned}
&= \Big\{ n^{-1} \sum_{i=1}^{n} \mathrm{E}\Big[\nabla_v^t m_\theta(s_\alpha, S_{-\alpha i}, V_i) Q(X_i, X, Z_i, W_i) K_h(Z_i - Z) K_h(W_i - W) \Big| T_i \Big] \\
&\quad + n^{-1} \sum_{k=1}^{n} \mathrm{E}\Big[\nabla_v^t m_\theta(s_\alpha, S_{-\alpha}, V) Q(X, X_k, Z, W) K_h(Z - Z_k) K_h(W - W_k) \Big| T_k \Big] \\
&\quad - \mathrm{E}\Big[\nabla_v^t m_\theta(s_\alpha, S_{-\alpha 1}, V_1) S_{12} \Big] \Big\} (1 + o_P(1)) + o_P(n^{-1/2}) \\
&= n^{-1} \sum_{k=1}^{n} \mathrm{E}\Big[\nabla_v^t m_\theta(s_\alpha, S_{-\alpha}, V) Q(X, S_k, W_k) \Big| T_k, Z = Z_k, W = W_k \Big] f_{ZW}(Z_k, W_k) \\
&\quad + O(h_n^{q_1}) + o_P(n^{-1/2}),
\end{aligned}
$$

provided $nh_n^{2q_1} \to 0$. This finishes the proof. □

Next, write the result of Lemma 7.1 for $\theta = \theta_0$ using the following abbreviated notations:

$$
\widehat{\phi}_0^{add}(s) - \phi_0^{add}(s) = n^{-1} \sum_{i=1}^{n} \Big\{ \sum_{\alpha=1}^{d_s} k_{1h}(s_\alpha - S_{\alpha i}) v_1^\alpha(T_i) + v_2(s, T_i) \Big\} + o_P(n^{-1/2}), \quad (7.4)
$$

and

$$
\dot{\widehat{\phi}}_0^{add}(s) - \dot{\phi}_0^{add}(s) = n^{-1} \sum_{i=1}^{n} \Big\{ \sum_{\alpha=1}^{d_s} k_{1h}(s_\alpha - S_{\alpha i}) w_1^\alpha(T_i) + w_2(s, T_i) \Big\} + o_P(n^{-1/2}), \quad (7.5)
$$

uniformly in $s \in R_{X,Z}$, where $S_i = (X_i, Z_i)$ and $T_i = (S_i, W_i, Y_i)$ for $i = 1, \ldots, n$.

Lemma 7.2. *Assume (A.1)–(A.6) and (C.1)–(C.9). Then,*

$$
\begin{aligned}
&\widehat{f}_{\epsilon(\theta_0)}(y) - f_{\epsilon(\theta_0)}(y) \\
&\quad = n^{-1} \sum_{i=1}^{n} \Big\{ \sum_{\alpha=1}^{d_s} v_1^\alpha(T_i) \frac{\partial}{\partial y} f_{\epsilon(\theta_0), S_\alpha}(y, S_{\alpha i}) + \mathrm{E}\Big[v_2(S, T_i) f'_{\epsilon(\theta_0)|S}(y|S) \Big| T_i \Big] \Big\} \\
&\qquad + n^{-1} \sum_{i=1}^{n} k_{2g}(y - \epsilon_i(\theta_0)) - f_{\epsilon(\theta_0)}(y) + R_{n1}(y), \\
&\widehat{f}'_{\epsilon(\theta_0)}(y) - f'_{\epsilon(\theta_0)}(y) \\
&\quad = n^{-1} \sum_{i=1}^{n} \Big\{ \sum_{\alpha=1}^{d_s} v_1^\alpha(T_i) \frac{\partial^2}{\partial y^2} f_{\epsilon(\theta_0), S_\alpha}(y, S_{\alpha i}) + \mathrm{E}\Big[v_2(S, T_i) f''_{\epsilon(\theta_0)|S}(y|S) \Big| T_i \Big] \Big\} \\
&\qquad + (ng)^{-1} \sum_{i=1}^{n} k'_{2g}\big(y - \epsilon_i(\theta_0)\big) - f'_{\epsilon(\theta_0)}(y) + R_{n2}(y),
\end{aligned}
$$

$$
\begin{aligned}
&\dot{\widehat{f}}_{\epsilon(\theta_0)}(y) - \dot{f}_{\epsilon(\theta_0)}(y) \\
&\quad = n^{-1}\sum_{i=1}^{n}\Big\{\sum_{\alpha=1}^{d_s} v_1^{\alpha}(T_i)\frac{\partial}{\partial y}\dot{f}_{\epsilon(\theta_0),S_\alpha}(y,S_{\alpha i}) + \mathrm{E}\Big[v_2(S,T_i)\dot{f}'_{\epsilon(\theta_0)|S}(y|S)\Big|T_i\Big] \\
&\qquad + \sum_{\alpha=1}^{d_s} w_1^{\alpha}(T_i)\frac{\partial}{\partial y} f_{\epsilon(\theta_0),S_\alpha}(y,S_{\alpha i}) + \mathrm{E}\Big[w_2(S,T_i)f'_{\epsilon(\theta_0)|S}(y|S)\Big|T_i\Big]\Big\} \\
&\qquad + (ng)^{-1}\sum_{i=1}^{n} k'_{2g}\big(y-\epsilon_i(\theta_0)\big)\big(\dot{\Lambda}_0(Y_i) - \dot{\phi}_0^{add}(S_i)\big) - \dot{f}_{\epsilon(\theta_0)}(y) + R_{n3}(y),
\end{aligned}
$$

where $\sup_y |R_{nj}(y)| = o_P(n^{-1/2})$, $j = 1,2,3$.

The proof of Lemma 7.2 is similar to that of Lemmas A.1–A.3 in Linton, Sperlich and Van Keilegom (2008), and is therefore omitted. The only difference is that here ϵ and (X,Z) are not independent, which has an effect on the main term in the above representations.

For the next lemma, we say for any $\theta \in \Theta$ that $G(\theta,\gamma)$ is pathwise differentiable at γ in the direction $[\bar{\gamma}-\gamma]$ if the limit $\lim_{\tau\to 0}\big[G\big\{\theta,\gamma+\tau(\bar{\gamma}-\gamma)\big\} - G(\theta,\gamma)\big]/\tau$ exists. The limit is in that case denoted by $\Delta(\theta,\gamma)[\bar{\gamma}-\gamma]$. This limit places an important role in the calculation of the asymptotic variance of $\widehat{\theta}$.

Lemma 7.3. *Assume (A.1)–(A.6) and (C.1)–(C.9). Then,*

$$
\Delta(\theta_0,\gamma_0^{add})[\widehat{\gamma}_0^{add} - \gamma_0^{add}] \;=\; n^{-1}\sum_{i=1}^{n}\Big\{\sum_{\alpha=1}^{d_s} D_1^{\alpha}(T_i) + D_2(T_i)\Big\} + o_P(n^{-1/2}),
$$

where for $i = 1,\ldots,n$,

$$
\begin{aligned}
D_1^{\alpha}(T_i) \;=\; & v_1^{\alpha}(T_i)E\Big[-\frac{\partial}{\partial\epsilon}\Big(\frac{1}{f_{\epsilon(\theta_0)}(y)}\nabla_\theta\left[f_{\epsilon(\theta_0)}(\epsilon(\theta_0))\right]\Big)f_{S_\alpha|\epsilon(\theta_0)}(S_{\alpha i}|\epsilon(\theta_0)) \\
& \qquad -\frac{1}{f^2_{\epsilon(\theta_0)}(\epsilon(\theta_0))}\frac{\partial}{\partial\epsilon}f_{\epsilon(\theta_0),S_\alpha}(\epsilon(\theta_0),S_{\alpha i})\nabla_\theta\left[f_{\epsilon(\theta_0)}(\epsilon(\theta_0))\right] \\
& \qquad +\frac{1}{f_{\epsilon(\theta_0)}(\epsilon(\theta_0))}\nabla_\theta\left[\frac{\partial}{\partial\epsilon}f_{\epsilon(\theta_0),S_\alpha}(\epsilon(\theta_0),S_{\alpha i})\right]\Big|S_{\alpha i}\Big] \\
& + w_1^{\alpha}(T_i)E\Big[\frac{\partial}{\partial\epsilon}f_{S_\alpha|\epsilon(\theta_0)}(S_{\alpha i}|\epsilon(\theta_0))\Big|S_{\alpha i}\Big],
\end{aligned}
\tag{7.6}
$$

and

$$
\begin{aligned}
D_2(T_i) \;=\; & E\Big[\frac{1}{f_{\epsilon(\theta_0)}(\epsilon(\theta_0))}\Big\{\frac{f'_{\epsilon(\theta_0)}(\epsilon(\theta_0))}{f_{\epsilon(\theta_0)}(\epsilon(\theta_0))}\nabla_\theta\left[f_{\epsilon(\theta_0)}(\epsilon(\theta_0))\right]v_2(S,T_i) \\
& \qquad -\frac{\nabla_\theta\left[f_{\epsilon(\theta_0)}(\epsilon(\theta_0))\right]}{f_{\epsilon(\theta_0)}(\epsilon(\theta_0))}E\Big(f'_{\epsilon(\theta_0)|S}(\epsilon(\theta_0)|S)v_2(S,T_i)\Big|\epsilon(\theta_0),T_i\Big) \\
& \qquad -\nabla_\theta\left[f'_{\epsilon(\theta_0)}(\epsilon(\theta_0))\right]v_2(S,T_i) \\
& \qquad +\nabla_\theta\epsilon(\theta_0)E\Big(f''_{\epsilon(\theta_0)|S}(\epsilon(\theta_0)|S)v_2(S,T_i)\Big|\epsilon(\theta_0),T_i\Big) \\
& \qquad +E\Big(\dot{f}'_{\epsilon(\theta_0)|S}(\epsilon(\theta_0)|S)v_2(S,T_i)\Big|\epsilon(\theta_0),T_i\Big)\Big\}\Big|T_i\Big] \\
& +E\Big[\frac{1}{f_{\epsilon(\theta_0)}(\epsilon(\theta_0))}\Big\{-f'_{\epsilon(\theta_0)}(\epsilon(\theta_0))w_2(S,T_i) \\
& \qquad +E\Big(f'_{\epsilon(\theta_0)|S}(\epsilon(\theta_0)|S)w_2(S,T_i)\Big|\epsilon(\theta_0),T_i\Big)\Big\}\Big|T_i\Big], \qquad (7.7)
\end{aligned}
$$

and where the functions v_1^α, w_1^α, v_2 and w_2 are defined in (7.4) and (7.5), ∇_θ denotes the gradient with respect to the vector θ, and $\frac{\partial}{\partial\epsilon}$ denotes the derivative with respect to the argument ϵ.

Proof. Consider an arbitrary θ. Straightforward calculations show that

$$
\begin{aligned}
&\Delta(\theta,\gamma_\theta^{add})[\widehat{\gamma}_\theta^{add}-\gamma_\theta^{add}] \\
&= E\Big[\Big\{\frac{f'_{\epsilon(\theta)}(\epsilon(\theta))}{f^2_{\epsilon(\theta)}(\epsilon(\theta))}(\widehat{\phi}_\theta^{add}-\phi_\theta^{add})(S)-\frac{(\widehat{f}_{\epsilon(\theta)}-f_{\epsilon(\theta)})(\epsilon(\theta))}{f^2_{\epsilon(\theta)}(\epsilon(\theta))}\Big\} \\
&\quad\times\Big\{f'_{\epsilon(\theta)}(\epsilon(\theta))\big[\dot{\Lambda}_\theta(Y)-\dot{\phi}_\theta^{add}(S)\big]+\dot{f}_{\epsilon(\theta)}(\epsilon(\theta))\Big\} \\
&\quad+\frac{1}{f_{\epsilon(\theta)}(\epsilon(\theta))}\Big\{-f''_{\epsilon(\theta)}(\epsilon(\theta))\big[\dot{\Lambda}_\theta(Y)-\dot{\phi}_\theta^{add}(S)\big](\widehat{\phi}_\theta^{add}-\phi_\theta^{add})(S) \\
&\qquad+(\widehat{f}'_{\epsilon(\theta)}-f'_{\epsilon(\theta)})(\epsilon(\theta))\big[\dot{\Lambda}_\theta(Y)-\dot{\phi}_\theta^{add}(S)\big] \\
&\qquad-f'_{\epsilon(\theta)}(\epsilon(\theta))(\dot{\widehat{\phi}}_\theta^{add}-\dot{\phi}_\theta^{add})(S) \\
&\qquad+(\dot{\widehat{f}}_{\epsilon(\theta)}-\dot{f}_{\epsilon(\theta)})(\epsilon(\theta))-\dot{f}'_{\epsilon(\theta)}(\epsilon(\theta))(\widehat{\phi}_\theta^{add}-\phi_\theta^{add})(S)\Big\}\Big].
\end{aligned}
$$

In order to calculate this expression for $\theta=\theta_0$, we make use of the expansions given in (7.4) and (7.5) and of Lemma 7.2. We will develop i.i.d. expansions for the terms involving v_1^α, v_2, w_1^α and w_2.

We start with w_1^α. The terms that contribute to w_1^α are those involving $\big(\dot{\widehat{\phi}}_0^{add}-\dot{\phi}_0^{add}\big)(S)$ and $\big(\dot{\widehat{f}}_{\epsilon(\theta_0)}-\dot{f}_{\epsilon(\theta_0)}\big)(\epsilon(\theta_0))$. More precisely, from the i.i.d. representations of these expressions,

we get

$$
\begin{aligned}
&n^{-1}\sum_{i=1}^{n}\sum_{\alpha=1}^{d_s} w_1^{\alpha}(T_i)E\Big[-\frac{f'_{\epsilon(\theta_0)}(\epsilon(\theta_0))}{f_{\epsilon(\theta_0)}(\epsilon(\theta_0))}k_{1h}(S_\alpha - S_{\alpha i}) + \frac{\frac{\partial}{\partial\epsilon}f_{\epsilon(\theta_0),S_\alpha}(\epsilon(\theta_0),S_{\alpha i})}{f_{\epsilon(\theta_0)}(\epsilon(\theta_0))}\Big|S_{\alpha i}\Big] \\
&= n^{-1}\sum_{i=1}^{n}\sum_{\alpha=1}^{d_s} w_1^{\alpha}(T_i)E\Big[-\frac{f'_{\epsilon(\theta_0)}(\epsilon(\theta_0))}{f_{\epsilon(\theta_0)}(\epsilon(\theta_0))}f_{S_\alpha|\epsilon(\theta_0)}(S_{\alpha i}|\epsilon(\theta_0)) + \frac{\frac{\partial}{\partial\epsilon}f_{\epsilon(\theta_0),S_\alpha}(\epsilon(\theta_0),S_{\alpha i})}{f_{\epsilon(\theta_0)}(\epsilon(\theta_0))}\Big|S_{\alpha i}\Big] \\
&= n^{-1}\sum_{i=1}^{n}\sum_{\alpha=1}^{d_s} w_1^{\alpha}(T_i)E\Big[\frac{\partial}{\partial\epsilon}\Big(\frac{f_{\epsilon(\theta_0),S_\alpha}(\epsilon(\theta_0),S_{\alpha i})}{f_{\epsilon(\theta_0)}(\epsilon(\theta_0))}\Big)\Big|S_{\alpha i}\Big] \\
&= n^{-1}\sum_{i=1}^{n}\sum_{\alpha=1}^{d_s} w_1^{\alpha}(T_i)E\Big[\frac{\partial}{\partial\epsilon}f_{S_\alpha|\epsilon(\theta_0)}(S_{\alpha i}|\epsilon(\theta_0))\Big|S_{\alpha i}\Big]. \qquad (7.8)
\end{aligned}
$$

Note that the terms in this sum have mean zero, since $E[w_1^{\alpha}(T)|S_\alpha] = 0$.

We now consider the terms involving v_1^{α}. Note that

$$\nabla_\theta\left[f_{\epsilon(\theta)}(\epsilon(\theta))\right] = f'_{\epsilon(\theta)}(\epsilon(\theta))\left[\dot\Lambda_\theta(Y) - \dot\phi_\theta^{add}(S)\right] + \dot f_{\epsilon(\theta)}(\epsilon(\theta))$$

and that

$$\nabla_\theta\left[f_{\epsilon(\theta),S_\alpha}(\epsilon(\theta),S_\alpha)\right] = \frac{\partial}{\partial\epsilon}f_{\epsilon(\theta),S_\alpha}(\epsilon(\theta),S_\alpha)\left[\dot\Lambda_\theta(Y) - \dot\phi_\theta^{add}(S)\right] + \dot f_{\epsilon(\theta),S_\alpha}(\epsilon(\theta),S_\alpha).$$

The terms that involve v_1^{α} can hence be written as

$$
\begin{aligned}
&n^{-1}\sum_{i=1}^{n}\sum_{\alpha=1}^{d_s} v_1^{\alpha}(T_i)E\Big[\frac{f'_{\epsilon(\theta_0)}(\epsilon(\theta_0))}{f^2_{\epsilon(\theta_0)}(\epsilon(\theta_0))}k_{1h}(S_\alpha - S_{\alpha i})\nabla_\theta\left[f_{\epsilon(\theta_0)}(\epsilon(\theta_0))\right] \\
&\quad -\frac{\frac{\partial}{\partial\epsilon}f_{\epsilon(\theta_0),S_\alpha}(\epsilon(\theta_0),S_{\alpha i})}{f^2_{\epsilon(\theta_0)}(\epsilon(\theta_0))}\nabla_\theta\left[f_{\epsilon(\theta_0)}(\epsilon(\theta_0))\right] - \frac{\nabla_\theta\left[f'_{\epsilon(\theta_0)}(\epsilon(\theta_0))\right]}{f_{\epsilon(\theta_0)}(\epsilon(\theta_0))}k_{1h}(S_\alpha - S_{\alpha i}) \\
&\quad +\frac{\frac{\partial^2}{\partial\epsilon^2}f_{\epsilon(\theta_0),S_\alpha}(\epsilon(\theta_0),S_{\alpha i})}{f_{\epsilon(\theta_0)}(\epsilon(\theta_0))}\nabla_\theta\left[\epsilon(\theta_0)\right] + \frac{\frac{\partial}{\partial\epsilon}\dot f_{\epsilon(\theta_0),S_\alpha}(\epsilon(\theta_0),S_{\alpha i})}{f_{\epsilon(\theta_0)}(\epsilon(\theta_0))}\Big|S_{\alpha i}\Big] \\
&= n^{-1}\sum_{i=1}^{n}\sum_{\alpha=1}^{d_s} v_1^{\alpha}(T_i)E\Big[-\frac{\partial}{\partial\epsilon}\Big(\frac{\nabla_\theta\left[f_{\epsilon(\theta_0)}(\epsilon(\theta_0))\right]}{f_{\epsilon(\theta_0)}(\epsilon(\theta_0))}\Big)f_{S_\alpha|\epsilon(\theta_0)}(S_{\alpha i}|\epsilon(\theta_0)) \\
&\quad -\frac{\frac{\partial}{\partial\epsilon}f_{\epsilon(\theta_0),S_\alpha}(\epsilon(\theta_0),S_{\alpha i})}{f^2_{\epsilon(\theta_0)}(\epsilon(\theta_0))}\nabla_\theta\left[f_{\epsilon(\theta_0)}(\epsilon(\theta_0))\right] + \frac{\nabla_\theta\left[\frac{\partial}{\partial\epsilon}f_{\epsilon(\theta_0),S_\alpha}(\epsilon(\theta_0),S_{\alpha i})\right]}{f_{\epsilon(\theta_0)}(\epsilon(\theta_0))}\Big|S_{\alpha i}\Big]. \qquad (7.9)
\end{aligned}
$$

Again, note that the above expression has mean zero since $E[v_1^{\alpha}(T)|S_\alpha] = 0$.

We now turn to the calculation of the expressions involving w_2, which are given by

$$
\begin{aligned}
&n^{-1}\sum_{i=1}^{n} E\Big[\frac{1}{f_{\epsilon(\theta_0)}(\epsilon(\theta_0))}\Big\{-f'_{\epsilon(\theta_0)}(\epsilon(\theta_0))w_2(S,T_i) \\
&\qquad\qquad +E\Big(f'_{\epsilon(\theta_0)|S}(\epsilon(\theta_0)|S)w_2(S,T_i)\Big|\epsilon(\theta_0),T_i\Big)\Big\}\Big|T_i\Big]. \qquad (7.10)
\end{aligned}
$$

Finally, the terms involving v_2 can be calculated in a similar manner. It now suffices to combine this calculation with (7.8), (7.9) and (7.10) to get the required result. □

7.4 Consistency

Proof of Theorem 4.1. We prove the consistency of $\widehat{\theta}$ by checking the conditions of Theorem 1 in Delsol and Van Keilegom (2014) (DVK hereafter). In that paper, high level conditions are developed for the consistency of the maximizer of a fairly general semiparametric maximization problem. In our setting,

$$\widehat{\theta} = \arg\max_{\theta\in\Theta} H_n(\theta, \widehat{f}_{\epsilon(\theta)}, \widehat{\phi}_\theta^{add}),$$

and we need to check whether the functions H_n, $\widehat{f}_{\epsilon(\theta)}$ and $\widehat{\phi}_\theta^{add}$ satisfy the conditions of the above theorem. First of all, condition (A1) in DVK is satisfied by definition of the estimator $\widehat{\theta}$, and their condition (A2) is our condition (C.8). Next, define the class

$$\mathcal{H} = \mathcal{M} \times C_1^1(I\!R),$$

where $\mathcal{M} = \sum_{\alpha=1}^{d_s} C_a^1(R_{S_\alpha})$ and $C_a^b(R)$ $(0 < a < \infty,\ 0 < b \leq 1,\ R \subset I\!R^k$ for some $k)$ is the set of all continuous functions $f : R \to I\!R$ for which

$$\sup_y |f(y)| + \sup_{y,y'} \frac{|f(y) - f(y')|}{\|y - y'\|^b} \leq a.$$

We equip the space $\mathcal{M}$ with the L_2-norm $\|\cdot\|_{L_2}$. For condition (A3) in DVK we need to show that

$$\mathrm{P}((\widehat{\phi}_\theta^{add}, \widehat{f}_{\epsilon(\theta)}) \in \mathcal{H} \ \ \forall\theta \in \Theta) \to 1 \text{ as } n \to \infty,$$

and that $\sup_{\theta\in\Theta} \|\widehat{\phi}_\theta^{add} - \phi_\theta^{add}\|_{L_2} = o_P(1)$ and $\sup_{\theta\in\Theta} \|\widehat{f}_{\epsilon(\theta)} - f_{\epsilon(\theta)}\|_{L_2} = o_P(1)$. For $\widehat{\phi}_\theta^{add}$, the decomposition in Lemma 7.1 allows to uniformly bound $\widehat{\phi}_\theta^{add} - \phi_\theta^{add}$ whereas for $\widehat{f}_{\epsilon(\theta)}$ this follows from Lemma 7.2 together with Corollary 2.7.4 in Van der Vaart and Wellner (1996). For condition (A4), since $H(\theta, h_1, h_2) = \mathrm{E}\{\log[h_2(\Lambda_\theta(Y) - h_1(X, Z))] + \log[\Lambda'_\theta(Y)]\}$, it suffices to show that $\sup_{\theta\in\Theta,(h_1,h_2)\in\mathcal{H}} |H_n(\theta, h_1, h_2) - H(\theta, h_1, h_2)| = o_P(1)$, i.e. we need to show that the family $\mathcal{F} = \{(x, z, y) \to \log[h_2(\Lambda_\theta(y) - h_1(x, z))] + \log[\Lambda'_\theta(y)] :\ \theta \in \Theta, (h_1, h_2) \in \mathcal{H}\}$ is Glivenko-Cantelli. This follows easily from Corollaries 2.7.2 and 2.7.4 in Van der Vaart and Wellner (1996). At last, condition (A5) is a regularity condition on H which is automatically satisfied since H is continuously differentiable of order 1. This finishes the proof of the consistency. □

7.5 Asymptotic normality

Proof of Theorem 4.2. In order to prove the asymptotic properties of our estimator, we need to check the high level assumptions of Theorems 1 and 2 in Chen, Linton and Van Keilegom (2003). Note that our setting is very different from Linton, Sperlich and Van Keilegom (2008) due to the fact that $S = (X, Z)$ and ϵ are not independent in our case and that we also have a generated covariate $\widehat{V}$ to take into account. However the structure of our proof is somewhat similar to the structure of the proof of their Theorem 4.1.

A crucial assumption of their Theorem 4.1 is assumption A.8 given in the Appendix of their paper, which gives the properties that the estimator $\widehat{\phi}_\theta^{add}(s)$ (denoted by $\widehat{m}_\theta(x)$ in their paper) needs to satisfy. In addition, to check condition (2.6) of Theorem 2 in Chen, Linton and Van Keilegom (2003), they use the results of 11 lemmas given in their Appendix A.2.

In our setting, using the conditions $(C.1) - (C.9)$ which already include their assumptions A.1.-A.7., everything boils down to checking an analogue of their assumption A.8. and an analogue of their lemmas. Let's start with the analogue of their assumption A.8, which in our case corresponds to the following:

(i) The estimator $\widehat{\phi}_0^{add}$ can be written as

$$\begin{aligned}&\widehat{\phi}_0^{add}(s) - \phi_0^{add}(s)\\ &= n^{-1}\sum_{i=1}^{n}\sum_{\alpha=1}^{d_s} k_{1h}(s_\alpha - S_{\alpha i}) v_{01\alpha}(s_\alpha, T_i) + n^{-1}\sum_{i=1}^{n} v_{02}(s, T_i) + \widehat{v}_0(s),\end{aligned}$$

where $T_i = (X_i, Z_i, W_i, Y_i)^t$, $\sup_s |\widehat{v}_0(s)| = o_P(n^{-1/2})$, $E(v_{01\alpha}(s_\alpha, T)|S_\alpha = s_\alpha) = 0$ and $E(v_{02}(s, T)) = 0$. Moreover, a similar expansion holds for the estimator $\dot{\widehat{\phi}}_0^{add}$.

(ii) Consider the space $\mathcal{M}$ defined in the proof of the consistency, Theorem 4.1. Then, $P(\widehat{\phi}_\theta^{add}, \dot{\widehat{\phi}}_\theta^{add} \in \mathcal{M} \text{ for all } \theta \in \Theta) \to 1$ as $n \to \infty$.

(iii) The space $\mathcal{M}$ satisfies $\int \sqrt{\log N(\lambda, \mathcal{M}, \|\cdot\|_{L_2})}\, d\lambda < \infty$, where $N(\lambda, \mathcal{M}, \|\cdot\|_{L_2})$ is the covering number with respect to the norm $\|\cdot\|_{L_2}$ of the class $\mathcal{M}$, i.e. the minimal number of balls of $\|\cdot\|_{L_2}$-radius λ needed to cover $\mathcal{M}$.

(iv) $\sup_{\theta\in\Theta} \|\widehat{\phi}_\theta^{add} - \phi_\theta^{add}\|_{L_2} = o_P(1)$, $\sup_{\theta\in\Theta} \|\dot{\widehat{\phi}}_\theta^{add} - \dot{\phi}_\theta^{add}\|_{L_2} = o_P(1)$.

(v) Uniformly over all θ with $\|\theta - \theta_0\| = o(1)$, $\|\widehat{\phi}_\theta^{add} - \phi_\theta^{add}\|_{L_2} = o_P(n^{-1/4})$ and $\|\dot{\widehat{\phi}}_\theta^{add} - \dot{\phi}_\theta^{add}\|_{L_2} = o_P(n^{-1/4})$.

(vi) For all θ with $\|\theta - \theta_0\| = o(1)$,

$$\sup_s \left| (\dot{\widehat{\phi}}_\theta^{add} - \dot{\phi}_\theta^{add})(s) - (\dot{\widehat{\phi}}_0^{add} - \dot{\phi}_0^{add})(s) \right| = o_P(1)\|\theta - \theta_0\| + O_P(n^{-1/2}).$$

First, for point (i), note that the i.i.d. representations for $\widehat{\phi}_0^{add}(s) - \phi_0^{add}(s)$ and $\dot{\widehat{\phi}}_0^{add}(s) - \dot{\phi}_0^{add}(s)$ are given in Lemma 7.1.

Next, let us check that $P(\widehat{\phi}_\theta^{add}, \dot{\widehat{\phi}}_\theta^{add} \in \mathcal{M} \text{ for all } \theta \in \Theta) \to 1$ as $n \to \infty$. We have to prove that $\widehat{\phi}_\theta^{add}$ and $\dot{\widehat{\phi}}_\theta^{add}$ are uniformly bounded in s and θ as well as their first derivatives with respect to the components of s. Using condition (C.2), the decomposition in Lemma 7.1 allows to uniformly bound $\widehat{\phi}_\theta^{add} - \phi_\theta^{add}$ and $\dot{\widehat{\phi}}_\theta^{add} - \dot{\phi}_\theta^{add}$. As for the first derivatives of these estimators, it suffices to show that they converge in probability to the true functions, uniformly in s and θ. The proof for these derivatives is somewhat similar in structure to the proof of Lemma 7.1, and we therefore restrict to explaining the main differences. In fact, the proof is even much simpler than that of Lemma 7.1, since the remainder terms are only required to be $o_P(1)$, instead of the much sharper bound $o_P(n^{-1/2})$ that is required in the aforementioned proof. In particular, contrary to the proof of Lemma 7.1, we do not need to develop expansions of U-processes and we do not need to perform detailed order calculations. Hence, the uniform boundedness of these derivatives follows, which shows point (ii) above.

For point (iii), note that the covering number $N(\lambda, \mathcal{M}, \|\cdot\|_{L_2})$ satisfies $\log N(\lambda, \mathcal{M}, \|\cdot\|_{L_2}) \leq K\lambda^{-1}$ (see Corollary 2.7.2 in Van der Vaart and Wellner, 1996), and hence

$$\int_0^\infty \sqrt{\log N(\lambda, \mathcal{M}, \|\cdot\|_{L_2})}\, d\lambda < \infty.$$

Next, using Lemma 7.1 it is easy to show that $\sup_{\theta\in\Theta} \|\widehat{\phi}_\theta^{add} - \phi_\theta^{add}\|_{L_2} = O_P((nh^{1/2})^{-1/2} + h^{q_1}) = o_P(n^{-1/4})$ (the uniformity in θ can be shown using standard arguments based on partitioning the compact set Θ in small subsets, and the rate of the L_2-distance can be proved following e.g. the method of proof in Härdle and Mammen, 1993). In a similar way we can show that $\sup_{\theta\in\Theta} \|\dot{\widehat{\phi}}_\theta^{add} - \dot{\phi}_\theta^{add}\|_{L_2} = o_P(n^{-1/4})$. This shows (iv) and (v).

Finally, for point (vi), note that (again using the second part of Lemma 7.1) it suffices to control (for all i)

$$\left\| \dot{\Lambda}_\theta(Y_i) - \dot{m}_\theta(S_i, V_i) - \dot{\Lambda}_0(Y_i) + \dot{m}_0(S_i, V_i) \right\|,$$

and this is bounded by

$$\left\| \ddot{\Lambda}_0(Y_i) - \ddot{m}_0(S_i, V_i) \right\| \|\theta - \theta_0\|(1 + o_P(1)) = o_P(1)\|\theta - \theta_0\|,$$

which is of the required order, and where $\ddot{\Lambda}_0$ represents the Hessian matrix with respect to θ_0. This finishes the proof of results (i)-(vi).

The next step is to present the analogues of the 11 lemmas given in Linton, Sperlich and Van Keilegom (2008). Their lemmas A.1-A-3, A.5 and A.9 concern results about the density estimation of the error ϵ and its derivatives and correspond to our Lemma 7.2. Their lemmas A.4, A.6-A.8, A.10-A.11 concern results about the functions M, M_n and their derivatives and correspond to our Lemma 7.3.

Conditions (C.1)-(C.10), the results (i)-(vi) stated above and these last two lemmas allow us to conclude. In particular, Lemma 7.3 is crucial for calculating the asymptotic variance of $\widehat{\theta}$, which is equal to the asymptotic variance of $\Gamma^{-1}\{M_n(\theta_0, \gamma_0^{add}) + \Delta(\theta_0, \gamma_0^{add})[\widehat{\gamma_0}^{add} - \gamma_0^{add}]\}$, with $\Delta(\theta_0, \gamma_0^{add})[\widehat{\gamma_0}^{add} - \gamma_0^{add}]$ defined in the paragraph above Lemma 7.3 (see condition (2.6) in Theorem 2 in Chen, Linton and Van Keilegom 2003). The asymptotic normality of $\widehat{\theta}$ then follows. □

Proof of Corollary 4.1. Write

$$\widehat{\phi}^{add}(s) - \phi_0(s) = \Big[\widehat{\phi}^{add}_{\widehat{\theta}}(s) - \widehat{\phi}^{add}_0(s)\Big] + \Big[\widehat{\phi}^{add}_0(s) - \phi^{add}_0(s)\Big]. \tag{7.11}$$

The first term on the right hand side equals $(\dot{\widehat{\phi}}^{add}_\theta(s)|_{\theta=\xi})^t(\widehat{\theta} - \theta_0)$ for some ξ on the line segment between $\widehat{\theta}$ and θ_0. From the proof of Theorem 4.2 it follows that

$$\sup_{\theta\in\Theta} \|\dot{\widehat{\phi}}^{add}_\theta(s)\| \leq \sup_{\theta\in\Theta} \|\dot{\widehat{\phi}}^{add}_\theta(s) - \dot{\phi}^{add}_\theta(s)\| + \sup_{\theta\in\Theta} \|\dot{\phi}^{add}_\theta(s)\| = O_P(1),$$

and hence the first term of (7.11) is $O_P(n^{-1/2}) = o_P((nh)^{-1/2})$ by Theorem 4.2. For the second term of (7.11) we apply Lemma 7.1, which yields that

$$\begin{aligned}
&\widehat{\phi}^{add}_0(s) - \phi^{add}_0(s) \\
&= n^{-1}\sum_{i=1}^{n}\sum_{\alpha=1}^{d_s} k_{1h}(s_\alpha - S_{\alpha i})\Big[\Lambda_0(Y_i) - m_0(S_i, V_i)\Big] f^{-1}_{S_\alpha|S_{-\alpha},V}(S_{\alpha i}|S_{-\alpha i}, V_i) + o_P((nh)^{-1/2}).
\end{aligned}$$

The result now follows from e.g. Lindeberg's central limit theorem, together with standard variance calculations. □

References

Angrist, J.D. and Krueger, A.B. (2001). Instrumental variables and the search for identification: from supply and demand to natural experiments. *Journal of Economic Perspectives*, **15**, 69-85.

Bickel, P.J. and Doksum, K. (1981). An analysis of transformations revisited. *Journal of the American Statistical Association*, **76**, 296-311.

Birke, M., Van Bellegem, S. and Van Keilegom, I. (2014). Semi-parametric estimation in a single-index model with endogenous variables. Technical report (http://www.uclouvain.be/en-369695.html, DP2014/43).

Blundell, R., Chen, X. and Kristensen, D. (2007). Semi-nonparametric IV estimation of shape-invariant Engel curves. *Econometrica*, **75**, 1613-1669.

Box, G.E.P. and Cox, D.R. (1964). An analysis of transformations. *Journal of the Royal Statistical Society - Series B*, **26**, 211-252.

Carroll, R.J. and Ruppert, D. (1988). *Transformation and Weighting in Regression*. Chapman and Hall, New York.

Chen, X., Linton, O.B. and Van Keilegom, I. (2003). Estimation of semiparametric models when the criterion function is not smooth. *Econometrica*, **71**, 1591-1608.

Chen, X. and Pouzo, D. (2009). Efficient estimation of semiparametric conditional moment models with possibly nonsmooth residuals. *Journal of Econometrics*, **152**, 46-60.

Cheng, G. (2015). Moment consistency of the exchangeably weighted bootstrap for semiparametric M-estimation. *Scandinavian Journal of Statistics* (to appear).

Cheng, G. and Huang, J.Z. (2010). Bootstrap consistency for general semiparametric M-estimation. *Annals of Statistics*, **38**, 2884–2915.

Cheng, G. and Kosorok, M.R. (2008). General frequentist properties of the posterior profile distribution. *Annals of Statistics*, **36**, 1819–1853.

Cheng, G. and Pillai, N. (2012). Semiparametric model based bootstrap. Working paper.

Chiappori, P.-A., Komunjer, I. and Kristensen, D. (2010). Nonparametric identification and estimation of transformation models. Technical report.

Colling, B., Heuchenne, C., Samb, R. and Van Keilegom, I. (2015). Estimation of the error density in a semiparametric transformation model. *Annals of the Institute of Statistical Mathematics*, **67**, 1-18.

Colling, B. and Van Keilegom, I. (2014). Goodness-of-fit tests in semiparametric transformation models. Technical report (http://www.uclouvain.be/en-369695.html, DP2014/17).

Delsol, L. and Van Keilegom, I. (2014). Semiparametric M-estimation with non-smooth criterion functions. Technical report (http://www.uclouvain.be/en-369695.html, DP2011/41).

Fève, F. and Florens, J.P. (2010). The practice of nonparametric estimation by solving inverse problems: the example of transformation models. *Econometrics Journal*, **13**, S1-S27.

Florens, J.-P., Johannes, J. and Van Bellegem, S. (2012). Instrumental regression in partially linear models. *Econometrics Journal*, **15**, 304-324.

Florens, J.P. and Sokullu, S. (2012). Nonparametric estimation of semiparametric transformation models. Technical report.

Härdle, W. and Mammen, E. (1993). Comparing nonparametric versus parametric regression fits. *Annals of Statistics*, **21**, 1926-1947.

Hayashi, F. (2000). *Econometrics*. Princeton University Press.

Heuchenne, C., Samb, R. and Van Keilegom, I. (2014). Estimating the residual distribution in semiparametric transformation models. Technical report (http://www.uclouvain.be/en-369695.html, DP2014/11).

Horowitz, J.L. (1996). Semiparametric estimation of a regression model with an unknown transformation of the dependent variable. *Econometrica*, **64**, 103-137.

Horowitz, J.L. (2001). Nonparametric estimation of a generalized additive model with an unknown link function. *Econometrica*, **69**, 499-513.

Imbens, G. and Newey, W. (2009). Identification and estimation of triangular simultaneous equations models without additivity. *Econometrica*, **77**, 1481-1512.

Imbens, G.W. and Rubin, D.B. (2015). *Causal Inference in Statistics, Social, and Biomedical Sciences: An Introduction*. Cambridge University Press, New York.

Jacho-Chavez, D., Lewbel, A. and Linton, O. (2010). Identification and nonparametric estimation of a transformed additively separable model. *Journal of Econometrics*, **156**, 392-407.

Johannes, J., Van Bellegem, S. and Vanhems, A. (2013). Iterative regularization in nonparametric instrumental regression. *Journal of Statistical Planning and Inference*, **143**, 24-39.

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

Linton, O., Sperlich, S. and Van Keilegom, I. (2008). Estimation on a semiparametric transformation model. *Annals of Statistics*, **36**, 686-718.

Mammen, E., Linton, O.B. and Nielsen, J.P. (1999). The existence and asymptotic properties of a backfitting projection algorithm under weak conditions. *Annals of Statistics*, **27**,

1443-1490.

Mammen, E. and Park, B.U. (2005). Bandwidth selection for smooth backfitting in additive models. *Annals of Statistics*, **33**, 1260-1294.

Mammen, E., Rothe, C. and Schienle, M. (2012). Semiparametric estimation with generated covariates. Technical report.

Manzi, J., San Martín, E. and Van Bellegem, S. (2014). School system evaluation by value-added analysis under endogeneity. *Psychometrika*, **79**, 130-153.

Moon, J.M. (2013). Sieve extremum estimation of transformation models. Technical report (UCSD, working papers).

Neumeyer, N., Noh, H. and Van Keilegom, I. (2014). Heteroscedastic semiparametric transformation models: estimation and testing for validity. Technical report (http://www.uclouvain.be/en-369695.html, DP2014/47).

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

Pakes, A. and Pollard, D. (1989). Simulation and the asymptotics of optimization estimators. *Econometrica*, **57**, 1027-1057.

Sakia, R.M. (1992). The Box-Cox transformation technique : a review. *The Statistician*, **41**, 169-178.

Sherman, R. (1994). Maximal inequalities for degenerate U-processes with applications to optimization estimators. *Annals of Statistics*, **22**, 439-459.

Sokullu, S. (2011). Nonparametric analysis of two-sided markets. Technical report.

Van der Vaart, A.W. and Wellner, J.A. (1996). *Weak Convergence and Empirical Processes*. Springer-Verlag, New York.

Wooldridge, J. (2008). *Introductory Econometrics: A Modern Approach*. South-Western College Pub.

Zellner, A. and Revankar, N.S. (1969). Generalized production functions. *Review of Economic Studies*, **36**, 241-250.