On Endogeneity and Shape Invariance in Extended Partially Linear Single Index Models

Jiti Gao[1]
Monash University, Australia

Namhyun Kim [2]
University of Exeter Business School, United Kingdom

Patrick W. Saart
Cardiff Business School, United Kingdom

Abstract

This paper elaborates the usefulness of the extended generalized partially linear single-index (EGPLSI) model introduced by Xia et al. (1999) in its ability to model a flexible shape-invariant specification. More importantly, a control function approach is proposed to address endogeneity in the EGPLSI model to enhance its applicability to empirical studies. Furthermore, it is shown that the attractive asymptotic features of the single-index type of a semiparametric model are still valid given intrinsic generated covariates. Our proposed method is then illustrated by applying to address the endogeneity of expenditure in the semiparametric analysis of empirical Engel curves with the British data.

JEL Classification: C14, C18, C51.

Keyword: Extended generalized partially linear single-index, control function approach, endogeneity, semiparametric regression models.

[1]This author acknowledges the Australian Rsearch council Discovery Grants Program for its support inder grant numbers:DP150101012 & DP170104421.

[2]Namhyun, Kim, University of Exeter Business School, UK, email: n.kim@exeter.ac.uk

The authors would like to thank the editor and two anonymous referees for their kind and thoughtful suggestions and comments on the paper. All remaining errors are our own.

1. Introduction

Xia et al. (1999) introduced the extended generalized partially linear single-index (EGPLSI) model of the form

$$Y_i = X_i'\beta_0 + g(X_i'\alpha_0) + \epsilon_i, \tag{1.1}$$

where (i) (X, Y) is a set of $\mathbb{R}^q \times \mathbb{R}$-valued observable random vectors; (ii) β_0 and α_0 are unknown parameters vectors such that $\beta_0 \perp \alpha_0$ with $||\alpha_0|| = 1$; (iii) $g(\cdot)$ is an unknown link function such that $g(\cdot) : \mathbb{R} \to \mathbb{R}$ and $g''(\cdot) \neq 0$; and (iv) $E(\epsilon|X) = 0$ suggesting that $E(\epsilon|V_0) = 0$ with $V_0 = X'\alpha_0$. In fact, the EGPLSI model is the extended version of the generalized partially linear single-index (GPLSI) model of Carroll et al. (1997) and Xia and Härdle (2006) and hence a number of non- and semiparametric models are special cases of the EGPLSI model. More importantly, the EGPLSI model is useful for modelling a flexible shape-invariant specification in pooling nonparametric regression curves (see Härdle and Marron (1990), and Robinson and Pinkse (1995) for examples) to model an aggregate structural relationship incorporating the individual heterogeneity (see Blundell and Stoker (2007) for examples). The EGPLSI model allows this type of shape-invariant specification with a functional flexibility because both scale and shift parameters can be incorporated in the model. Therefore, the paper aims to address endogeneity in the EGPLSI model causing an identification problem, to enhance its applicability to empirical studies.

Recently, a number of methods have been discussed in the literature on how endogeneity can be best addressed in non- and semiparametric models. Among these, two of the most popular alternatives are the nonparametric instrumental variable estimation (NPIV) and the control function (CF) approach (see Blundell and Powell (2003) for an excellent review). The NPIV approach relies on different stochastic assumptions to the CF one and there are a few well-known difficulties that are intrinsic to the NPIV, particularly the so-called ill-posed inverse problem (see Ai and Chen (2003), and Blundell et al. (2007) for details). On the other hand, the CF approach alternatively allows the specification of endogeneity, which is based on an intuitive triangular structure of a model (see Blundell et al. (1998), and Blundell and Powell (2003) for details).

This paper particularly aims to develop the CF approach. Although the generated covariates issue is intrinsic in the development of the CF approach, similar to the study of Mammen et al. (2016), the proposed method maintains the attractive features of the single-index (SI) model with relatively mild conditions in the literature and shows an accessible extension to strictly stationary and α-mixing process. In a SI model, Härdle et al. (1993) showed that the optimal bandwidth for estimating a link function can be used for the $\sqrt{n}$-consistent estimation of the index coefficients. The current paper shows that this attractive feature is still valid with the CF approach and under-smoothing for estimating a first-stage reduced-form equation is not required in order to archive $\sqrt{n}$-consistency. These results are developed in details with the simplest data structure, namely IID random sample, then extended to a strictly stationary and α-mixing case. Furthermore, the convenient applicability of our proposed CF approach is explored by analyzing the empirical Engel curves based on the British data.

The structure of the rest of the paper is as follows. In Section 2, the usefulness of the EGPLSI model for modelling a flexible shape-invariant specification is elaborated. In addition, the development of the CF approach in the EGPLSI model and a Monte Carlo exercise assessing the finite-sample performances of the proposed estimators are also presented. In Section 3, the implementation of the empirical study of the cross sectional relationships between specific goods and the level of total expenditure are investigated. Finally, Section 4 concludes the paper with a summary of the main findings and the further issues to be investigated. All mathematical proofs of the main theoretical results are presented in the supplemental document.

2. EGPLSI Model, Shape-Invariant Specification and Endogeneity

In this section, the usefulness of the EGPLSI model introduced by Xia et al. (1999) is elaborated for specifying a flexible shape-invariant specification. This section then introduces endogeneity into the EGPLSI model, establishes the CF approach to address endogeneity and presents the asymptotic properties and finite sample performances from a Monte Carlo simulation exercise for the estimators.

2.1. Shape-Invariant Specification within EGPLSI Model Framework

Let us discuss a flexible shape-invariant specification within the EGPLSI model framework by considering the two sets of observations. The first set of observations, $(X_1, Y_1), \ldots, (X_n, Y_n)$, is assumed to follow the data generating process shown below

$$Y_i = m_1(X_i) + \varepsilon_i, \quad i = 1, \ldots, n, \tag{2.1}$$

where ε is assumed to be independent with mean 0 and the common variance σ^2. Suppose the second set of observations, $(X'_1, Y'_1), \ldots, (X'_n, Y'_n)$, is from the following nonparametric regression model

$$Y'_i = m_2(X'_i) + \varepsilon'_i, \tag{2.2}$$

where ε' is independent from ε, but otherwise has the same stochastic structure as ε and has the common variance σ'^2. The main interest here is to model the curves whose parametric nature is modelled by [3]

$$m_2(X') = S^{-1}_{\theta_0}(m_1(T^{-1}_{\theta_0}(X'))), \tag{2.3}$$

where T_θ and S_θ are invertible transformations, particularly scalings and shifts of the axes indexed by parameters $\theta \in \Theta \subseteq R^d$, and θ_0 is the vector of true values of the parameters. A good estimate of θ_0 is provided by θ for which the curve $m_1(X)$ is closely approximated by

$$m(X, \theta) = S_\theta(m_2(T_\theta(X))). \tag{2.4}$$

For the sake of illustration, the simple models are considered as follows

$$m_1(X) = (X - 0.4)^2 \quad \text{and} \quad m_2(X') = (X' - 0.5)^2 - 0.2, \tag{2.5}$$

which fit in the framework described by (2.1) to (2.4) by defining the following

$$\begin{aligned} T_\theta(X) &= \theta^{(1)}X + \theta^{(2)} \\ m_2(T_\theta(X)) &= \left(\theta^{(1)}X + \theta^{(2)} - 0.5\right)^2 - 0.2 \\ S_\theta(m_2(T_\theta(X))) &= \left(\theta^{(1)}X + \theta^{(2)} - 0.5\right)^2 - 0.2 + \theta^{(3)}X + \theta^{(4)}, \end{aligned}$$

[3] The case of (2.3) is available on the request from the author

where $\theta_0 = \left(\theta_0^{(1)}, \theta_0^{(2)}, \theta_0^{(3)}, \theta_0^{(4)}\right) = (1, 0.1, 0, 0.2)$.

When a curve comparison problem with a similar parametric nature to (2.3) is considered, Härdle and Marron (1990) suggested an estimation procedure by which separated kernel smoothers are used in order to compute the estimates of $m_1(\cdot)$ and $m_2(\cdot)$. The estimator of θ_0 is then found by minimizing a L^2-norm objective function of kernel estimates of $m_1(\cdot)$ and $m_2(\cdot)$, and the approximation in (2.4). Alternatively, pooling the two sets of observations is more desirable. Modelling the data within the EGPLSI model framework enables this type of pooling nonparametric regression. The shift and scaling of the axes illustrated in the example above fit in the EGPLSI framework, shown below

$$m_3(X_1, X_2) = [\beta_{01}X_1 + \beta_{02}X_2] + \left\{([\alpha_{01}X_1 + \alpha_{02}X_2] - 0.5)^2 - 0.2\right\}, \tag{2.6}$$

where $X_1 = \begin{cases} X \\ X' \end{cases}$ and $X_2 = \begin{cases} 1 & \text{if } X_1 = X \\ 0 & \text{if } X_1 = X' \end{cases}$. The model examples in (2.5) can be obtained by defining

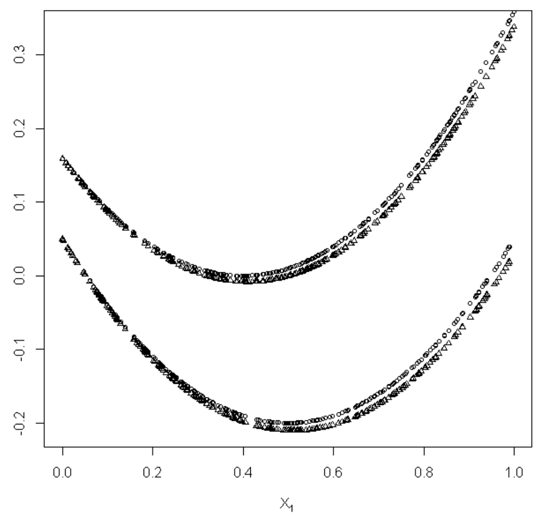
$$(\beta_{01}, \beta_{02}, \alpha_{01}, \alpha_{02}) = (0, 0.2, 1, 0.1). \tag{2.7}$$

Five hundred simulated observations of the model are represented by circles in Figure 2.1, where X_{1i} on the x-axis is a uniform random variable on $[0, 1]$ for $i = 1, \ldots, 500$. The two sets of observations are determined by X_2, which is a Bernoulli random variable with the parameter $p = 0.5$. It should be noted, however, that the set of values of the parameters in (2.7) do not satisfy the identification conditions which require that $\beta_0 \perp \alpha_0$ with $||\alpha_0|| = 1$. An approximate model that satisfies these identification conditions is obtained by first setting $\beta_{02} = 0.2$ and $\alpha_{02} = 0.1$, so that $\beta_{01} = -0.02$ and $\alpha_{01} = 0.99$ can be derived. Five hundred simulated observations of this type of a model are represented by triangles in Figure 2.1. In practice, when there is enough reason to believe (perhaps based on economic theory) that $\beta_{01} = 0$ and $\alpha_{01} = 1$, then such a model can be obtained by scaling and shifting, respectively, as follows

$$X_2 = v_{01} - \beta_{01}X_1 \text{ and } X_1 + \frac{\alpha_{02}}{\alpha_{01}}X_2 = \frac{v_{02}}{\alpha_{01}},$$

where $\beta_{01}X_1 + \beta_{02}X_2 = v_{01}$ and $\alpha_{01}X_1 + \alpha_{02}X_2 = v_{02}$. This method is illustrated in the empirical analysis in Section 3.

Figure 2.1. *500-simulated observations based on* $m_3(\cdot,\cdot)$.



2.2. Endogeneity and Newly Proposed Estimation Procedure

Despite its ability to model a flexible shape-invariant specification, the applicability of EGPLSI model to an empirical study is limited because of its shortfalls in addressing endogeneity. There are two potential sources of endogeneity in the model, namely endogeneity in the parametric and in the nonparametric components. If it is present, endogeneity in the parametric component is relatively easy to deal with.[4] Hence, to simplify the argument, the parametric covariates are assumed to belong to a subset $X_1 \subseteq \mathbb{R}^{q_1}$, for $q_1 < q$, of X such that $E(\epsilon|X_1) = 0$, namely the parametric covariates are exogenous, without loss of generality. In this case, endogeneity in the nonparametric component exists when $E(\epsilon|X) \neq 0$, which implies that $E(\epsilon|V_0) \neq 0$. An unanticipated property from the SI type of semiparametric models is that estimators of the index coefficients are still $\sqrt{n}$-consistent even with the presence of endogeneity because of the partialling-out process in estimating the index coefficients (see Ichimura (1993), Härdle et al. (1993), and Xia and Härdle (2006) for

[4] A comprehensive discussion on the presence of endogeneity in the parametric component can be found in Li and Racine (2007).

details). Nonetheless, the link function in the EGPLSI model is unidentifiable by using the conditional expectation relationship in the presence of endogeneity.

In the following, let us present the development of the CF approach in the EGPLSI model. For the sake of the notational simplicity, the simplest case is considered, namely the presence of an endogenous nonparametric covariate denoted by X_2.[5] Hereafter, let Z denote a vector of valid instruments for X_2 as follows

$$X_{2i} = g_x(Z_i) + \eta_i, \tag{2.8}$$

where $E(\eta|Z) = 0$, and $E(\epsilon|X_2) = E(\epsilon|Z,\eta) = E(\epsilon|\eta) \equiv \iota(\eta)$ with (X_2, Z) is a set of $\mathbb{R} \times \mathbb{R}^{q_z}$-valued observable random vectors, and $g_x(Z)$ and $\iota(\eta)$ are unknown real functions such that $g_x(\cdot) : \mathbb{R}^{q_z} \to \mathbb{R}$ and $\iota(\cdot) : \mathbb{R} \to \mathbb{R}$, respectively. The above stochastic assumption on ϵ is standard in the CF literature suggesting the exogeneity condition of Z, particularly $E(\epsilon|Z,\eta) = E(\epsilon|\eta)$ (see Newey et al. (1999), Blundell and Powell (2004), and Su and Ullah (2008) for examples). Furthermore, the necessary identification condition for the link function as discussed in Newey et al. (1999) is non-existence of a linear functional relationship between X_2 and η.

By imposing the structure of (2.8), the EGPLSI model in (1.1) in the presence of endogeneity is rewritten as

$$Y_i = X_i'\beta_0 + m(V_{0i}, \eta_i) + e_i, \tag{2.9}$$

where $m(v_0, \eta) \equiv g(v_0) + \iota(\eta)$ with $\iota(\eta) \neq 0$ being the endogeneity control function, and $E(e|X) = 0$. The conditional expectation relationship, based on (2.9), is obtained as follows

$$m_y(v_0, \eta) \equiv m(v_0, \eta) + m_x(v_0, \eta)'\beta_0, \tag{2.10}$$

where $m_y(v_0, \eta) \equiv E(y|V_0, \eta)$ and $m_x(v_0, \eta) \equiv E(x|V_0, \eta)$.

In the following, the performance of the CF approach in the EGPLSI model based on (2.8) to (2.10) is discussed. The identification issue is first presented as

[5] The generalized version, namely more than one endogenous nonparametric covariates, is available by a request to the author.

follows. Given α and β, let

$$J(\alpha,\beta) = E\left[Y - E(Y|V,\eta) - \{X - E(X|V,\eta)\}'\beta\right]^2$$

$$\mathcal{V} = E(\{X - E(X|V,\eta)\}\{X - E(X|V,\eta)\}'); \mathcal{W} = E(\{X - E(X|V,\eta)\}\{Y - E(Y|V,\eta)\}),$$

where $V = X'\alpha$. Suppose that $g(\cdot)$ is twice differentiable and that X has a positive density function on a union of a finite number of open convex subset in $\mathbb{R}^q$. The minimum point of $J(\alpha,\beta)$ with $\alpha \perp \beta$ is then unique at α_0 and $\beta_{\alpha_0} = \{\mathcal{V}(\alpha_0)\}^+\mathcal{W}(\alpha_0)$, where $\{\mathcal{V}(\alpha_0)\}^+$ is the Moore-Penrose inverse.

Before we discuss the optimization procedure, the necessary notation is defined for the sake of convenience. We assume that the random sample $\{(X_i', Z_i', Y_i); i = 1,\ldots,n\}$ is IID. Let $f_x(x)$ and $f_z(z)$ denote the joint density functions of X' and Z', respectively. Let us also denote $f_\alpha(v)$ as the density function of $V = X'\alpha$. We assume that $\mathcal{A}_j \subset \mathbb{R}^k$ is the union of a finite number of open sets such that $f_j(s) > C$ on $\mathcal{A}_j$, where $k = q$ or q_z and $j = x$ or z for some constant $C > 0$. Hereafter, this region is considered to avoid the boundary points. Because the region is not known in practice, Xia and Härdle (2006) suggested using the weight function such that $I_n(s) = 1$ if $\frac{1}{n}\sum_{i=1}^n K_{j,i}(s) > C$ and 0 otherwise, where K_j is a corresponding kernel function. In this paper, $I_n(s)$ is omitted for the notational simplicity. In addition, C, C' and C'' denote generic constants varying from one place to another.

The conditional expectations, namely $E(Y|V,\eta)$ and $E(X|V,\eta)$, are then estimated with the leave-one-out nonparametric estimation as follows

$$\hat{E}_i(Y_i|V_i,\eta_i) = \frac{\sum_{j\neq i} L_{h_v h_\eta}(V_j - V_i, \eta_j - \eta_i)Y_j}{\sum_{j\neq i} L_{h_v h_\eta}(V_j - V_i, \eta_j - \eta_i)} \tag{2.11}$$

$$\hat{E}_i(X_i|V_i,\eta_i) = \frac{\sum_{j\neq i} L_{h_v h_\eta}(V_j - V_i, \eta_j - \eta_i)X_j}{\sum_{j\neq i} L_{h_v h_\eta}(V_j - V_i, \eta_j - \eta_i)}, \tag{2.12}$$

where $L_{h_v h_\eta}$ is a product kernel function constructed from the product of univariate kernel functions of $k_{h_v}(\cdot) \times k_{h_\eta}(\cdot)$ with the relevant bandwidth parameters, h_v and h_η. Furthermore, the first stage leave-one-out nonparametric estimation of the reduced equation in (2.8) used to estimate η_i is as follows

$$\hat{\eta}_i = X_i - \hat{g}_{x,i}(Z_i), \tag{2.13}$$

where $\hat{g}_{x,i}(Z_i) = \frac{\sum_{j\neq i} K_{h_z}(Z_j - Z_i)X_j}{\sum_{j\neq i} K_{h_z}(Z_j - Z_i)}$ with $K_{h_z}(\cdot)$ being the product kernel function constructed from $k_{h_{z_1}}(\cdot) \times \cdots \times k_{h_{z q_z}}(\cdot)$, and h_{z_j}, for $j = 1, \ldots, q_z$, is the relevant bandwidth parameter. The LS estimates of the unknown parametric coefficients are then computed, given the initial values of the index coefficients denoted by α, as follows

$$\beta = \left(S_{\hat{U}_2}\right)^{-} S_{\hat{U}_2 \hat{W}_2}, \tag{2.14}$$

where $S_{AB} = \frac{1}{n}\sum_{i=1}^{n} A_i B_i'$, $S_A = S_{AA}$, $(S_A)^{-}$ is a generalized inverse of (S_A), $\hat{W}_{2i} \equiv Y_i - \hat{E}_i(Y_i|V_i, \hat{\eta}_i)$ and $\hat{U}_{2i} \equiv X_i - \hat{E}_i(X_i|V_i, \hat{\eta}_i)$. Next, based on $\beta \in B_n$, $\hat{\alpha}$, $\hat{h}_v$ and $\hat{h}_{\hat{\eta}}$ are computed by minimizing the objective function as follows

$$\min_{\alpha \in A_n, h_v, h_{\hat{\eta}} \in \mathcal{H}_n} \hat{J}(\alpha, h_v, h_{\hat{\eta}}) \equiv \min_{\alpha \in A_n, h_v, h_{\hat{\eta}} \in \mathcal{H}_n} \frac{1}{n}\sum_{i=1}^{n} (\hat{W}_{2i} - \hat{U}_{2i}'\beta)^2, \tag{2.15}$$

where $A_n = \{\alpha : ||\alpha - \alpha_0|| \leq Cn^{-1/2}\}$, $B_n = \{\beta : ||\beta - \beta_0|| \leq Cn^{-1/2}\}$ and $\mathcal{H}_n = \left\{h_z, h_v, h_\eta : Cn^{-1/5} \leq h_z, h_v, h_\eta \leq C'n^{-1/5}\right\}$ for $0 < C < C' < \infty$. Finally, re-estimate β_0 by using $\hat{\alpha}$, $\hat{h}_{\hat{v}}$ and $\hat{h}_{\hat{\eta}}$ as follows

$$\hat{\beta} = \left(S_{\hat{U}_3}\right)^{-} S_{\hat{U}_3 \hat{W}_3}, \tag{2.16}$$

where $\hat{W}_{3i} \equiv Y_i - \hat{E}_i(Y_i|\hat{V}_i, \hat{\eta}_i)$ and $\hat{U}_{3i} \equiv X_i - \hat{E}_i(X_i|\hat{V}_i, \hat{\eta}_i)$ with $\hat{V}_i = X_i'\hat{\alpha}$.

Remark 2.1. *The conditions for α and β below (2.15) are not as restrictive as it seemed because $\hat{\alpha}$ and $\hat{\beta}$ are $\sqrt{n}$-consistent. Furthermore, $\sqrt{n}$-consistency is achieved without under-smoothing in the first-stage of the proposed estimation procedure (i.e. estimation of the reduced-form equation in (2.8)). In general, under-smoothing in the first-stage of the estimation procedure is not required when $q_z < 3$ and $q - q_1 < 3/2$.*

The remaining task is then to identify the unknown link function. It is plausible to apply the marginal integration technique of Linton and Nielsen (1995), and Tjøstheim and Auestad (1994) to identify each of the functions because of the additive specification of the conditional expectation relation (see below (2.9)). The standard identification condition in the literature is assuming that $E(g(V_0)) = E(\iota(\eta)) = 0$ (see Hastie and Tibshirani (1990), Gao et al. (2006) and Gao (2007) for details). Hence, the marginal integration technique identifies $g(\cdot)$ and $\iota(\cdot)$ functions up to some constant values as follows

$$m_1(V_0) \equiv \int m(V_0, \eta)\, dQ(\eta) = g(V_0) + C \text{ and } m_2(\eta) \equiv \int m(V_0, \eta)\, dQ(V_0) = \iota(\eta) + C',$$

where $C \equiv \int \iota(\eta) dQ(\eta)$, $C' \equiv \int g(V_0) dQ(V_0)$ and Q is a probability measure in $\mathbb{R}$ with $\int dQ(\eta) = \int dQ(V_0) = 1$. The estimate of the link function can therefore be obtained by

$$\hat{m}_1(\hat{V}) = \frac{1}{n}\sum_{i=1}^{n} \hat{m}(\hat{V}, \hat{\eta}_i) \text{ and } \hat{g}(\hat{V}) = \hat{m}_1(\hat{V}) - \hat{C}, \tag{2.17}$$

where $\hat{m}(\hat{V}, \hat{\eta}_i) = \hat{E}(Y|\hat{V}, \hat{\eta}_i) - \hat{E}(X|\hat{V}, \hat{\eta}_i)'\hat{\beta}$, $\hat{C} = \frac{1}{n}\sum_{i=1}^{n} \hat{m}_1(\hat{V}_i)$, and $\hat{m}_1(\hat{V})$ is estimated by keeping $\hat{V}_i$ at $\hat{V}$ while taking average over $\hat{\eta}_i$.

Before discussing the main theoretical results of the estimators proposed above, the estimation procedure is briefly summarized as follows.

Step 2.1: Estimate the endogeneity control covariate, $\hat{\eta}$, as in (2.13).

Step 2.2: Estimate β as in (2.14) with $\hat{\eta}_i$ from Step 2.1 and α.

Step 2.3: Estimate $\hat{\alpha}$ and $\hat{\beta}$ as in (2.16) and (2.18), respectively.

Step 2.4: Estimate $\hat{m}(\hat{V}_i, \hat{\eta}_i)$ by using (2.10) with $\hat{\alpha}$ and $\hat{\beta}$ from Step 2.3, then perform the marginal integration technique to estimate $\hat{g}(\hat{V})$ as in (2.17).

2.3. Asymptotic Properties of Proposed Estimators

In this subsection, the asymptotic properties of the estimators are discussed as follows. The required necessary conditions are presented first. Given ρ, let $\mathcal{A}_{j'}^{\rho}$ denote the set of all points in $\mathbb{R}^{k'}$, where $k' = q$ or 1, at a distance no greater than ρ from $\mathcal{A}_{j'}$ for $j' = x, \eta$. Let $\mathcal{U} = \{(V_0, \eta) : X \in \mathcal{A}_x^{\rho} \text{ and } \eta \in \mathcal{A}_\eta^{\rho}\}$ and $f(V_0, \eta)$ denote the joint density function of (V_0, η) with random arguments of X' and η. The necessary regularity conditions are then as follows.

Assumption 2.1. *The vector of instrumental variables $\{Z_i : i \geq 1\}$ satisfy (2.8).*

Assumption 2.2. *The joint density functions of $f_z(Z)$ and $f(V, \eta)$ are bounded and are bounded away from zero with bounded and continuous second derivatives on $\mathcal{A}_z$ and $\mathcal{U}$ for all values of $\alpha \in A_n$, respectively.*

Assumption 2.3. *Assume that* $g_x(Z)$*, and* $m(V,\eta)$*,* $m_y(V,\eta)$ *and* $m_x(V,\eta)$ *have bounded and continuous second derivatives on* $\mathcal{A}_z$ *and* $\mathcal{U}$ *for all values of* $\alpha \in A_n$*.*

Assumption 2.4. *Assume that a univariate kernel function* $k(\cdot)$ *and its first derivative* $k^{(1)}(\cdot)$ *are supported on the interval* $(-1,1)$ *and* $k(\cdot)$ *is a symmetric density function. Furthermore, both* $k(\cdot)$ *and* $k^{(1)}(\cdot)$ *satisfy the Lipschitz conditions.*

Assumption 2.5. *Let* $E(\eta|Z) = 0$ *and* $E(\eta^2|Z) = \sigma_1^2(Z)$*,* $E(e|X,\eta) = 0$ *and* $E(e^2|X,\eta) = \sigma^2(X,\eta)$*,* $E(u|X,\eta) = 0$ *and* $E(u^2|X,\eta) = \sigma_2^2(X,\eta)$*, and the functions* σ^2*,* σ_1^2 *and* σ_2^2 *are bounded and continuous. In addition,* $\sup_i E||X_i||^l < \infty$*,* $\sup_i E|Y_i|^l < \infty$ *and* $\sup_i E||Z_i||^l < \infty$ *for some large enough* $l > 2$*.*

Assumption 2.2 is necessary to avoid the random denominator problem. Assumptions 2.2 and 2.3 ensure that the kernel function in Assumption 2.4 leads to a second-order bias in kernel smoothing. A higher-order bias can be achieved by imposing more restrictive conditions on the smoothness of the functions (see Robinson (1988) for details). The condition on the first derivative of the kernel function in Assumption 2.4 permits the use of the Taylor expansion argument to address the generated covariate, $\hat{\eta}_i$ (a similar condition on the derivatives of the kernel function can be found in Hansen (2008)). The Lipschitz conditions for both the kernel function and its derivative provide the convenience for the proof of the uniform convergence. Finally, Assumption 2.5 grants the use of the Chebyshev inequality.

Now let us introduce a few necessary notations used in the main theoretical results below. Let $\mathcal{K}_{z,2} = \int z^2 K_{h_z}(z)dz$, $\mathcal{K}_{\eta,2} = \int \eta^2 k_{h_\eta}(\eta)d\eta$ and $\mathcal{K}_{v,2} = \int v_0^2 k_{h_v}(v_0)dv_0$. Furthermore, let $\mathcal{K}_z = \int k_{h_{z,j}}(z)^2 dz$ and $\mathcal{K} = \mathcal{K}_v\mathcal{K}_\eta$, where $\mathcal{K}_v = \int k_{h_v}(v_0)^2 dv_0$ and $\mathcal{K}_\eta = \int k_{h_\eta}(\eta)^2 d\eta$. Let $f_{z,j}^{(r)}$ be the r-th derivatives of $f_z(z)$ with respect to Z_j, for $j = 1, \cdots, q_z$, and let $f_{v_0}^{(r)}(v_0,\eta)$ and $f_\eta^{(r)}(v_0,\eta)$ be the r-th partial derivatives of $f(v_0,\eta)$ with respect to V_0 and η, respectively. Moreover, let $g_{x,j}^{(r)}(z)$ be the r-th partial derivatives of $g_x(z)$ with respect to Z_j, and let $m_{v_0}^{(r)}(V_0,\eta)$ and $m_\eta^{(r)}(v_0,\eta)$ be that of $m(v_0,\eta)$ with respect to V_0 and η, respectively. Then, let

$$B_z(z) \equiv \tfrac{\mathcal{K}_{z,2}}{2f(z)}\left\{2f_{z,j}^{(1)}(z)g_{x,j}^{(1)}(z) + f_z(z)g_{x,j}^{(2)}(z)\right\}$$
$$B_v(v_0,\eta) \equiv \tfrac{\mathcal{K}_{v,2}}{2f(v_0,\eta)}\left\{2f_{v_0}^{(1)}(v_0,\eta)m_{v_0}^{(1)}(v_0,\eta) + f(v_0,\eta)m_{v_0}^{(2)}(v_0,\eta)\right\}$$

$$B_\eta(v_0,\eta) \equiv \frac{\mathcal{K}_{\eta,2}}{2f(v_0,\eta)} \left\{2f_\eta^{(1)}(v_0,\eta)m_\eta^{(1)}(v_0,\eta) + f(v_0,\eta)m_\eta^{(2)}(v_0,\eta)\right\}.$$

In addition, let

$$IMSE_1(h_z) \asymp \int \left\{ \left[\sum_{j=1}^{q_z} B_{z,j}(z)h_{z,j}^2\right]^2 + \frac{\mathcal{K}_z^{q_z}}{nh_{z,1}\dots h_{z,q_2}}\frac{\sigma_1^2(z)}{f_z(z)} \right\} f(z)dz$$

$$IMSE_2(h_v,h_\eta) \asymp \int \left\{ \left[B_v(v_0,\eta)h_v^2 + B_\eta(v_0,\eta)h_\eta^2\right]^2 + \frac{\mathcal{K}}{nh_vh_\eta}\frac{\sigma^2(V_0,\eta)}{f(v_0,\eta)} \right\} f(x,\eta)dxd\eta,$$

where $\asymp$ means that the quotient of the two sides tends to 1 as $n \to \infty$.

Theorem 2.1. *Under Assumptions 2.1 to 2.5, the minimizing objective function in (2.15) is rewritten as follows*

$$\hat{J}(\alpha,h_v,h_{\hat{\eta}}) = \tilde{J}(\alpha) + T_1(h_z) + T_2(h_v,h_\eta) + R_1(\alpha,h_v,h_\eta) + R_2(\alpha,h_v,h_\eta,h_z), \quad (2.18)$$

where $T_1(h_z) \equiv \frac{1}{n}\sum_{i=1}^n \{\hat{g}_{x,i}(Z_i) - g_x(Z_i)\}^2 = IMSE_1(h_z) + R_3(h_z)$, $T_2(h_v,h_\eta) \equiv \frac{1}{n}\sum_{i=1}^n \{\hat{m}_i(V_{0i},\eta_i) - m(V_{0i},\eta_i)\}^2 = IMSE_2(h_v,h_\eta) + R_4(h_v,h_\eta)$, $\sup_{\alpha\in A_n,h_v,h_\eta\in\mathcal{H}_n} |R_1(\alpha,h_v,h_\eta)| = o_p(n^{-1/2})$, *and* $\sup_{\alpha\in A_n,h_v,h_\eta,h_z\in\mathcal{H}_n} |R_2(\alpha,h_v,h_\eta,h_z)| = o_p(n^{-1/2})$ *with* $\hat{m}_i(\cdot)$ *and* $\hat{g}_{x,i}(\cdot)$ *being the leave-one-out local constant estimators of* $m(\cdot)$ *and* $g_x(\cdot)$, *respectively. More importantly*

$$\tilde{J}(\alpha) = \frac{1}{n}\sum_{i=1}^n \{W_i - U_i'\beta\}^2,$$

where $W_i \equiv Y_i - E(Y_i|V_i,\eta_i)$ *and* $U_i \equiv X_i - E(X_i|V_i,\eta_i)$. *Furthermore,* $\sup_{h_z\in\mathcal{H}_n} |R_3(h_z)| = o_p(n^{1/5})$ *and* $\sup_{h_v,h_\eta\in\mathcal{H}_n} |R_4(h_v,h_\eta)| = o_p(n^{1/5})$ *because they do not depend on* α.

The results of Theorem 2.1 show the attractive properties of our proposed CF approach. Similar to the results of Härdle et al. (1993) and Xia et al. (1999), Theorem 2.1 shows that the properties of the bandwidth parameter estimators can be studied while assuming α_0 is known. Moreover, the asymptotically optimal bandwidth parameters for estimating $m(\cdot)$ function are assumed to be used for the $\sqrt{n}$-consistent estimation of α_0. In addition, under-smoothing is not required in estimating the first-stage reduced-form equation, as already stated in Remark 2.1. In particular, Theorem 2.1 suggests that minimizing $\hat{J}(\alpha,h_v,h_{\hat{\eta}})$ simultaneously with respect to α,

h_v and $h_{\hat{\eta}}$, is asymptotically equivalent to separately minimizing $\tilde{J}(\alpha)$ with respect to α, $T_1(h_z)$ with respect to h_z, and $T_2(h_v, h_\eta)$ with respect to h_v and h_η, assuming that α_0 and η are known. This is because the remainder terms, namely $R_1(\alpha, h_v, h_\eta)$ and $R_2(\alpha, h_z, h_v, h_\eta)$, are shown to be asymptotically negligible.

Next, the asymptotic properties of $\hat{\alpha}$ and $\hat{\beta}$ are shown as a corollary of Theorem 2.1 given that $\Phi_{U_0} = [\{X - E(X|V_0,\eta)\}\{X - E(X|V_0,\eta)\}']$.

Corollary 2.1. *Under the assumptions of Theorem 2.1, the asymptotic properties of $\hat{\alpha}$ and $\hat{\beta}$ are as follows*

$$\sqrt{n}(\hat{\beta} - \beta_0) \to_D N(0, Var_1), \tag{2.19}$$

where $Var_1 = \sigma^2 \left[\Phi_{U_0}^- - \left(m_0^{(1)} \Phi_{U_0} \right)^- \Phi_{U_0} \left\{ m_0^{(1)} \right\}^2 \left(m_0^{(1)} \Phi_{U_0} \right)^- \right]$, *and*

$$\sqrt{n}(\hat{\alpha} - \alpha_0) \to_D N(0, Var_2), \tag{2.20}$$

where $Var_2 = \sigma^2 \left[\left\{ \left(m_0^{(1)} \right)^2 \Phi_{U_0} \right\}^- - \left\{ m_0^{(1)} \Phi_{U_0} \right\}^- \Phi_{U_0} \left\{ m_0^{(1)} \Phi_{U_0} \right\}^- \right]$.

Finally, the asymptotic properties of $\hat{g}(\hat{v})$ are presented in Theorem 2.2 below.

Theorem 2.2. *Under the assumptions of Theorem 2.1, and* $\inf_{z \in \mathcal{A}_z} f_z(z) > 0$ *and* $\inf_{x,\eta \in \mathcal{U}} f(v_0, \eta) > 0$, *the asymptotic results of $\hat{g}(\hat{v})$ are as follows*

$$\sqrt{nh_v}\left(\hat{g}(\hat{v}) - g(v_0) - Bias\right) \to_D N(0, Var),$$

where $Bias = h_v^2 B_v(v_0, \eta) + h_\eta^2 B_\eta(v_0, \eta)$ *and* $Var = f_\alpha(v_0)\mathcal{K}_v \int \frac{\sigma^2(V_0,\eta) f_\eta^2(\eta)}{f^2(v_0,\eta)} dQ(\eta)$ *with* $f_\alpha(v_0)$ *and* $f_\eta(\eta)$ *denoting the density functions of* V_0 *and* η, *respectively.*

Remark 2.2. *In these results, it is clear the first stage nonparametric estimation does not contribute to the asymptotic variance of the estimators in the final stage. This characteristic is common among multi-stage nonparametric estimation procedures (see Su and Ullah (2008) for an example). However, this differs from the work of Li and Wooldridge (2002) which considers parametrically generated covariates in a PL semiparametric regression model. Li and Wooldridge (2002) showed that the variance of the first stage estimation is not asymptotically negligible instead contributes to the variances of the estimators of the finite-dimensional parameters in the final stage.*

Remark 2.3. *It is also interesting to explore the case of performing the CF approach without the presence of endogeneity. The essential stochastic assumption of the CF approach below (2.8) implies no existence of any endogeneity control function and, hence there is no identification problem in estimating the link function. Therefore, performing the CF approach without the presence of endogeneity causes an unnecessary multi-stage nonparametric estimation and the presence of redundant covariates in estimating the link function. However, the theoretical results of the proposed estimators particularly Theorems 2.1 and 2.2 and Corollary 2.1 are still valid with minor modifications, especially in terms of* $IMSE_2(h_v, h_\eta)$, Var_1 *and* Var_2*, and the bias and the variance of* $\hat{g}(\hat{v})$*. The minor modifications are as follows*

$$IMSE_2(h_v, h_\eta)^* \asymp \int \left\{ \left[B_v^*(v_0, \eta) h_v^2 h_\eta^2\right]^2 + \frac{\mathcal{K}}{n h_v h_\eta} \frac{\sigma^{*2}(V_0, \eta)}{f(v_0, \eta)} \right\} f(x, \eta) dx d\eta$$

$$Bias^* = h_v^2 B_v^*(v_0, \eta) \text{ and } Var^* = f_\alpha(v_0) \mathcal{K}_v \int \frac{\sigma^{*2}(V_0) f_\eta^2}{f^2(v_0, \eta)} dQ(\eta),$$

where $B_v^*(v_0, \eta) = \frac{\mathcal{K}_{v,2}}{2f(v_0,\eta)} \left\{ 2 f_{v_0}(v_0, \eta) g^{(1)}(v_0) + f(v_0, \eta) g_{v_0}^{(2)}(v_0, \eta) \right\}$ *and* $\sigma^{*2} = E(\epsilon^2 | X, \eta)$ $= E(\epsilon^2|X)$*, and* Var_1^* *and* Var_2^* *are obtained by replacing* $m_0^{(1)}$ *with* $g_0^{(1)}$ *in (2.19) and (2.20) with* $g_0^{(1)}$ *being the first derivative of* $g(v_0)$ *with respect to* V_0*.*

Remark 2.4. *Our results can also be extended to more general data structure where a random sample* $\{(X_t', Z_t', Y_t);\ t = 1, \ldots, n\}$ *is a strictly stationary and* α*-mixing process under Assumptions 2.6 and 2.7 below in addition to 2.1 to 2.5 above.*

In the rest of this section, we discuss about how to extend these established theoretical results to stationary time series data as in Remark 2.4. First, let $\xi_t \equiv (X_t'\alpha_0, \eta_t)$ and $f_\xi(\xi)$ denote the joint density function of $X'\alpha_0$ and η. The necessary regularity conditions for the strictly stationary and α-mixing case are then as follows.

Assumption 2.6. *(i) The conditional densities satisfy the following conditions*

$$f_{\xi_1, \xi_l | X_1, X_l}(\xi_1, \xi_l) \le C < \infty;\ f_{\xi_1, \xi_l | Y_1, Y_l}(\xi_1, \xi_l) \le C' < \infty;\ f_{Z_1, Z_l | X_1, X_l}(Z_1, Z_l) \le C'' < \infty$$

for some constants $C, C', C'' > 0$ *and for all* $l \ge 1$*. (ii) The mixing and moment conditions are as follows*

$$\sum_l l^a [\alpha(l)]^{1-2/l} < \infty,\ E||X_0||^l < \infty \text{ and } f_{\xi_1|X_1}(\xi|X) \le C < \infty;$$

$$\sum_l l^{a'}[\alpha(l)]^{1-2/l} < \infty,\ E|Y_0|^l < \infty \text{ and } f_{\xi_1|Y_1}(\xi|Y) \le C' < \infty;$$
$$\sum_l l^{a''}[\alpha(l)]^{1-2/l} < \infty,\ E||Z_0||^l < \infty \text{ and } f_{Z_1|X_1}(z|X) \le C'' < \infty,$$

where $l > 2$ and $a, a', a'' > 1 - 2/l$. (iii) There is a sequence of positive integer s_T, which satisfies $s_T \to \infty$ and $s_T = o\left\{(nh_{z,T}^{q_z})^{1/2}\right\}$, such that $(n/h_{z,T}^{q_z})^{1/2}\alpha(s_T) \to 0$ as $T \to \infty$.

Assumption 2.7. *(i) Let the density functions $f_z(z)$ and $f(v_0, \eta)$ satisfy $\inf_{z\in\mathcal{A}_z} f_z(z) > 0$ and $\inf_{x,\eta\in\mathcal{U}} f(v_0, \eta) > 0$. (ii) In addition, we require the following moments conditions*

$$E||X||^s < \infty,\ \sup_{\xi\in\mathcal{U}} \int ||X||^s f(x,\xi)dx,\ E|Y|^s < \infty, \sup_{\xi\in\mathcal{U}} \int |Y|^s(y,\xi)dy;\ \int_{x\in\mathcal{A}_z} ||X||^s f(x,z)dx,$$

for some $s > 2$. (iii) The bandwidth sequences, h_v, h_η and h_z, tend to zero as $T \to \infty$ and satisfy, for some $\delta > 0$,

$$T^{1-2s^{-1}-2\delta} h_z^{q_z} \to \infty;\ T^{1-2s^{-1}-2\delta} h_v h_\eta \to \infty; T^{1-2s^{-1}-2\delta}\left(h_z^{q_z} h_v h_\eta^3\right)^{1/2} \to \infty.$$

In the proof of the $\sqrt{n}$-consistency of $\hat\alpha$ and $\hat\beta$ in the case of Remark 2.4, Propositions A.1 to A.15 in the supplementary document encompass the extra covariance terms caused by the serial dependences in the sample. Under Assumptions 2.1 to 2.5 and 2.6(i)(ii), those covariance terms can be shown to be $o_p(n^{-1/2})$. For instance, the extra covariance term in Proposition A.1 might be derived as $\sum_{l=1}^{n-1}(1 - t/n)\text{Cov}(\hat\varphi_1, \hat\varphi_{l+1}) = o(h_v h_\eta)$. However the consistency of $\hat g(\hat v)$ requires stronger conditions than the case of $\hat\alpha$ and $\hat\beta$, namely the uniform convergence of $\hat f(v_0, \eta)$, which requires the uniform convergences of Q_j, where $j = 1, \cdots, 5$ in (B.1) in the supplementary document. Under Assumptions 2.1 to 2.5, 2.6(i)-(ii) and 2.7, Q_j are shown to be $o_p(1)$ as follows

$$\sup_{\xi\in\mathcal{U}, z\in A_z} |Q_{2i}| = \sup_{\xi\in\mathcal{U}, z\in A_z} |Q_{5i}| = O_p\left\{\left(\frac{(\ln n)^2}{n^2 h_z^{q_z} h_v h_\eta^3}\right)^{1/2} + h_z^2(h_v^2 + h_\eta^2)\right\}.$$

Furthermore, the asymptotic normality of $\hat g(\hat v)$ is then obtained by applying Assumption 2.6 (iii) for the standard nonparametric small-block and large-block arguments. Nonetheless, the asymptotic normalities of $\hat\alpha$ and $\hat\beta$ are obtained by applying the part of Assumption 2.6 (ii), namely $\sum_l l^a[\alpha(l)]^{1-2/l} < \infty$, $E||X_0||^l < \infty$,

$\sum_l l^{a'}[\alpha(l)]^{1-2/l} < \infty$ and $E|Y_0|^l < \infty$, to (A.6) and (A.10) for the small-block and large-block arguments of a standard strictly stationary and α-mixing process.

2.4. Simulation Studies

In this section[6], the finite-sample performance of the proposed estimator is investigated by making a comparison between the performances of the estimation method introduced in Xia et al. (1999) referred as the XTL procedure and the CF approach established in Section 2.2 as the KS procedure in the presence of endogeneity. Throughout this section, optimization is implemented by using a limited memory Broyden-Fletcher-Goldfarb-Shanno algorithm for the bound constrained optimization of Byrd et al. (1995). All simulation exercises are conducted in R with the Gaussian kernel function and the number of replications $Q = 200$. To compare and evaluate the finite sample performances of the procedures, the mean and mean absolute errors of the estimates of both coefficients, α_0 and β_0, across Q replications are computed in Tables 2.1 and 2.2. The averaged absolute error of the estimates of the unknown structural function is also computed as follows

$$\text{ae}_{\hat{g}} = \frac{1}{n}\sum_{i=1}^{n}\left|\hat{g}(\hat{V}_i) - g(V_{0i})\right|,$$

where n is the number of samples.

In the analysis that follows, an example model of the following form is considered

$$Y_i = \beta_{01}X_{1i} + \beta_{02}X_{2i} + \beta_{03}X_{3i} + g(V_{0i}) + \epsilon_i, \tag{2.21}$$

where $V_0 = \alpha_{01}X_1 + \alpha_{02}X_2 + \alpha_{03}X_3$, $g(V_0) = \exp\left\{-2(\alpha_{01}X_1 + \alpha_{02}X_2 + \alpha_{03}X_3)^2\right\}$, and X_j is independently and uniformly distributed on $[-1, 1]$ for $j = 1, 2$. It is required that $\beta_0 \perp \alpha_0$ with $\| \alpha_0 \| = 1$. In order for these conditions to be satisfied, define $\beta_{02} = 0.4$, $\beta_{03} = 0$, $\alpha_{01} = 0.7$, $\alpha_{02} = -0.6$, then β_{01} and α_{03} are defined as follows

$$\alpha_{03} = \sqrt{1 - \alpha_{01}^2 - \alpha_{02}^2} \text{ and } \beta_{01} = -\frac{\beta_{02}\alpha_{02}}{\alpha_{01}}.$$

[6] The results of extensive simulation exercises for GPLSI model are available by a request to the author.

In this example, endogeneity is introduced by letting $X_3 = Z + \eta$, where Z and η are independently and uniformly distributed on $[-0.5, 0.5]$ and $[-1, 1]$, respectively, and $\epsilon = \eta + e$ with e is independent and standard normally distributed. Tables 2.1 and 2.2 present the results based on the XTL and KS procedures, respectively.

The simulation results in Table 2.1 show the strong evidence against the use of XTL procedure in the presence of endogeneity. This evidence is clear when the averaged absolute errors, $\mathrm{ae}_{\hat{g}}$, in Table 2.1 are considered. On the other hand, the simulation results in Table 2.2 suggest that the KS procedure is able to identify the link function, namely $g(\cdot)$ function, in the presence of endogeneity.

Table 2.1. *EGPLSI model with endogeneity and the XTL's procedure.*

n	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\alpha}_1$	$\hat{\alpha}_2$	$\hat{\alpha}_3$
50	0.3130	0.4332	0.8884	-0.7748	0.5597
150	0.3088	0.4340	0.8993	-0.7671	0.5279
300	0.3142	0.4264	0.8988	-0.7674	0.5225
500	0.3135	0.4288	0.8960	-0.7653	0.5179

n	$\lvert\hat{\beta}_1 - \beta_{01}\rvert$	$\lvert\hat{\beta}_2 - \beta_{02}\rvert$	$\lvert\hat{\alpha}_1 - \alpha_{01}\rvert$	$\lvert\hat{\alpha}_2 - \alpha_{02}\rvert$	$\lvert\hat{\alpha}_3 - \alpha_{03}\rvert$	$\mathrm{ae}_{\hat{g}}$
50	0.0656	0.0714	0.1691	0.1253	0.1586	0.0905
150	0.0428	0.04572	0.0859	0.0559	0.0910	0.0891
300	0.0331	0.03377	0.0629	0.0548	0.0426	0.0895
500	0.0306	0.0319	0.0229	0.0156	0.0181	0.0906

Table 2.2. *EGPLSI model with endogeneity and the KS procedure.*

n	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\alpha}_1$	$\hat{\alpha}_2$	$\hat{\alpha}_3$
50	0.2645	0.4652	0.9638	-0.8249	0.5483
150	0.3260	0.4135	0.8975	-0.7852	0.4756
300	0.3486	0.3945	0.8090	-0.6997	0.4382
500	0.3555	0.3891	0.7353	-0.6295	0.3992

n	$\lvert\hat{\beta}_1 - \beta_{01}\rvert$	$\lvert\hat{\beta}_2 - \beta_{02}\rvert$	$\lvert\hat{\alpha}_1 - \alpha_{01}\rvert$	$\lvert\hat{\alpha}_2 - \alpha_{02}\rvert$	$\lvert\hat{\alpha}_3 - \alpha_{03}\rvert$	$\mathrm{ae}_{\hat{g}}$
50	0.0816	0.0684	0.1678	0.1389	0.1195	0.0632
150	0.0307	0.0264	0.1244	0.0962	0.0769	0.0265
300	0.0213	0.0183	0.0446	0.0327	0.0285	0.0160
500	0.0189	0.0159	0.0416	0.0319	0.0263	0.0124

3. Semiparametric CF approach to Shape-Invariant Empirical Engel Curves

In this section, a flexible shape-invariant Engel curves system is analyzed within the framework of the EGPLSI model with the proposed CF approach above. The consumer optimization theory suggests to include a scale and a shift parameters within a flexible shape-invariant empirical Engel curve for the individual household heterogeneity (see Pendakur (1999), Blundell and Powell (2003) and Blundell et al. (2007) for examples). In addition, the endogeneity of total expenditure is well-known which is caused by the two-stage budgeting model (see Blundell et al. (1998) and Blundell et al. (2007) for details). Hence, it is natural to study a shape-invariant Engel curves system within the framework of the EGPLSI model with the newly developed CF approach.

3.1. The Empirical Model and Estimation

Hereafter, let $\{Y_{il}, X_{1i}, X_{2i}\}_{i=1}^{n}$ represent an IID sequence of n household observations on the budget share Y_{il} of good $l = 1, \ldots, L \geq 1$ for each household i facing the same relative prices, the log of total expenditure X_{1i}, and a vector of household composition variables X_{2i}. For each commodity l, budget shares and total outlay are related by a general stochastic Engel curve, namely $Y_l = G_l(X_1) + \epsilon_l$, where $G_l(\cdot)$ is an unknown function that can be estimated by using a standard nonparametric method under the exogeneity assumption of total expenditure (i.e. $E(\epsilon_l|X_1) = 0$.) Nonetheless, a number of previous studies have reported that household expenditures typically display great variation with demographic composition. A simple approach for estimating the model is to stratify by each distinct discrete outcome of X_2 and then carry out our estimation with nonparametric smoothing within each cell. At some point, however, it may be useful to pool the Engel curves across different household demographic types and to allow X_1 to enter each Engel curve semiparametrically. This idea leads to the specification below

$$Y_{il} = \beta_{0l}' X_{2i} + g_l(X_{1i} - \phi(\gamma_0' X_{2i})) + \epsilon_{il}, \tag{3.1}$$

where $g_l(\cdot)$ is an unknown function and $\phi(\gamma_0' X_{2i})$ is a known function up to a finite set of unknown parameters γ_0, which can be interpreted as the log of general equivalence scales for household i. In the current paper, $\phi(\gamma_0' X_{2i}) = \gamma_0' X_{2i}$ is chosen so that (3.1) is specified as follows

$$Y_{il} = \beta_{0l}' X_{2i} + g_l(X_{1i} - \gamma_0' X_{2i}) + \epsilon_{il}. \tag{3.2}$$

In this application, total expenditure is allowed to be endogenous and a measure of earning of the head of each household is used as an instrument.

Following the CF approach discussed above, the empirical model to be estimated is the following form below

$$Y_{il} = \beta_{01,l}X_{1i} + \beta'_{0l}X_{2i} + g_l(\alpha_{01}X_{1i} + \alpha'_{02}X_{2i}) + \epsilon_{il} \quad (3.3)$$

$$X_{1i} = m_{X1}(Z_i) + \eta_i, \text{ where } E(\eta|Z) = 0 \text{ and } E(\epsilon_l|Z,\eta) = E(\epsilon_l|\eta) \neq 0, \quad (3.4)$$

with $m_{X1}(Z) = E(X_1|Z)$ and $\{Z_i\}_{i=1}^n$ represents an IID sequence of the measure of earning of n heads of households and (3.3) is a semiparametric model that satisfies all the identification conditions required in the construction of the EGPLSI model. The theoretically consistent model in (3.1) can then be solved based on (3.3). To this end, a similar scaling transformation to that explained in Section 2 is used. In the remainder of this section, some specific details about the estimation procedure are discussed. Rather than basing the discussion on (3.3) to (3.4), it is statistically more equivalent to do so based on as follows

$$Y_{il} = \beta'_{0l}X_{2i} + g_l(X_{1i} - \gamma'_0X_{2i}) + \epsilon_{il} \quad (3.5)$$

$$X_{1i} = m_{X1}(Z_i) + \eta_i, \text{ where } E(\eta|Z) = 0 \text{ and } E(\epsilon_l|Z,\eta) = E(\epsilon_l|\eta) \neq 0. \quad (3.6)$$

These models suggest the conditional expectation relationship shown below

$$E\left(Y_l|(X_1 - \gamma'_0X_2),\eta\right) - \beta'_{0l}E\left(X_2|(X_1 - \gamma'_0X_2),\eta\right) = g_l(X_1 - \gamma'_0X_2) + \iota_l(\eta), \quad (3.7)$$

where $E\left(\epsilon_l|(X_1 - \gamma'_0X_2),\eta\right) = E\left(\epsilon_l|\eta\right) \equiv \iota_l(\eta) \neq 0$, which immediately leads to

$$Y_{il} = \beta'_{0l}X_{2i} + g_l(X_{1i} - \gamma'_0X_{2i}) + \iota_l(\eta_i) + e_{il}, \quad (3.8)$$

$$X_{1i} = m_{X1}(Z_i) + \eta_i, \quad (3.9)$$

where $E(e_l|X_1,X_2,\eta) = 0$. Let $m_l\left(\{X_{1i} - \gamma'_0X_{2i}\},\eta_i\right) = g_l(X_{1i} - \gamma'_0X_{2i}) + \iota_l(\eta_i)$. In order to use (3.8), it is important to note that

$$m_{1,l}(X_1 - \gamma'_0X_2) = \int m_l(\{X_1 - \gamma'_0X_2\},\eta)\, d\eta \text{ and } g_l(X_1 - \gamma'_0X_2) = m_{1,l}(X_1 - \gamma'_0X_2) - C, \quad (3.10)$$

where $C = \int \iota(\eta)dQ(\eta)$ and $E(g_l(\cdot)) = 0$.

If a linear specification is imposed on $\iota_l(\cdot)$, (3.8) would be similar to the extended partially linear (EPL) model discussed in Blundell et al. (1998). In this case, Blundell et al. (1998) showed that a test of the endogeneity null can be constructed by testing $H_0 : \iota_l = 0$, where ι_l is an unknown parameter. The current paper, however, suggests more flexible functional form for testing the endogeneity null by constructing the variability bands for $\iota(\cdot)$. To do so, the following procedure is employed.

Step 3.1.1: Obtain an empirical estimate of $g_l(X_1 - \gamma_0' X_2)$ in (3.10).

Step 3.1.2: Regress (3.8) using the estimates in Step 3.1.1 to obtain the nonparametric estimates of $\iota_l(\cdot)$.

Step 3.1.3: Compute the bias-corrected confidence bands for the nonparametric smoothing using the procedure introduced by Xia (1998). Finally, the Bonferroni-type variability bands are obtained by using a similar procedure discussed in Eubank and Speckman (1993).

To perform Step 3.1.1, the estimation procedure introduced in Section 2 is used. However, some modifications are required to take the vector of index coefficients, γ_0 (a general equivalence scale for household i), into account. In this case, the objective function (2.15) is only used for a particular commodity l. The new objective function, $\min_{\gamma \in A_n, h_{v,l}, h_{\hat{\eta},l} \in \mathcal{H}_n} \hat{J}(\gamma, h_{v,l}, h_{\hat{\eta},l})$, is the summation of these individual functions that is minimized with respect to γ and 14 smoothing parameters, particularly two for each commodity. Finally, the estimation procedure is completed by using $\hat{\gamma}$ as well as $\hat{h}_{\hat{v},l}$ and $\hat{h}_{\hat{\eta},l}$.

In addition, the model in (3.8) can also be re-stated as

$$Y_{il}^* = g_l(X_{1i} - \gamma_0' X_{2i}) + e_{il}, \tag{3.11}$$

where $Y_l^* \equiv Y_l - \beta_{0l}' X_2 - \iota_l(\eta)$. The use of (3.11) relies on

$$m_{2,l}(\eta) \;=\; \int m_l(v, \eta)\, dv = \iota_l(\eta) + C' \text{ and } \iota_l(\eta) = m_{2,l}(\eta) - C', \tag{3.12}$$

where $V = X_1 - \gamma' X_2$, $C' = \int g(v) dQ(v)$ and $E(\iota_l(\cdot)) = 0$, which corresponds to (3.10) above. Hence, the model in (3.11) suggests that the estimates of the shape-invariant Engel curves and the related confidence bands are obtained as follows.

Step 3.2.1: Obtain empirical estimates of $\iota_l(\eta)$ in (3.12).

Step 3.2.2: Regress (3.11) using the estimates in Step 3.2.1 to obtain the nonparametric estimates of $g_l(\cdot)$.

Step 3.2.3: Compute the bias-corrected confidence bands about the nonparametric estimator in Step 3.2.2 using the procedure introduced by Xia (1998).

3.2. The Engel Curve Data

In our application, the data set is drawn from the British Family Expenditure Survey (FES) 1995-96. The seven broad categories of goods are considered as follows: (1) fuel, light and power (fuel hereafter); (2) fares, other travel costs and running of motor vehicles (fares); (3) food; (4) alcoholic drink and tobacco (alcohol); (5) leisure goods & services (leisure goods); (6) clothing and footwear (clothing); (7) personal goods & services (personal goods).

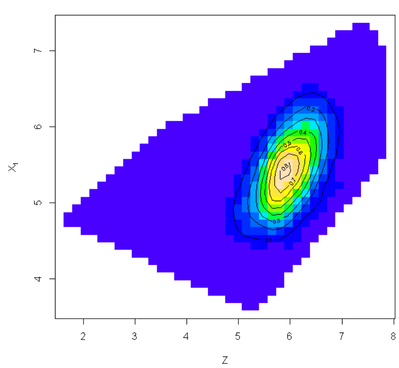
Table 3.1. *Descriptive statistics.*

	Couples with 1 or 2 children		**Couples without children**	
	Mean	**Std. Dev**	**Mean**	**Std. Dev**
Budget shares:				
Fuel	0.0692	0.0011	0.0618	0.0012
Fares	0.1537	0.0025	0.1715	0.0031
Food	0.3235	0.0028	0.2768	0.0031
Alcohol	0.0844	0.0022	0.1144	0.0031
Leisure goods	0.2155	0.0038	0.2298	0.0045
Clothing	0.0926	0.0024	0.0872	0.0029
Personal goods	0.0606	0.0016	0.0581	0.0019
Expenditure and income:				
log (total expenditure)	5.4374	0.0130	5.4524	0.0161
log (income)	5.9205	0.0153	6.0397	0.0166
Sample size	1072		1278	

To maintain some demographic homogeneity, a subset of married or cohabiting couples are selected from the FES, particularly categories 1 and 3 of variable *ms* in table *adult*. In addition, those where the head of household is aged between 20 and 55 (i.e. variable *age* in table *adult*) and in work (i.e. excluding the category 1 of the variable *fted* in the table *adult* and category 6 of the variable *a093* in the table *set8*) are considered. Finally, all households with three or more children are excluded. Our demographic variable, X_2, is a binary dummy variable that reflects whether a couple has 1 or 2 children (where $X_2 = 1$) or no children (where $X_2 = 0$). Overall, there are 2350 observations, 1278 are couples with one or two children. Table 3.1 shows larger expenditure shares for fuel, food, clothing and personal goods for the households with children as expected. Also as expected, households without children are able to spend higher proportions of their total expenditure on alcohol and leisure goods. Overall, there are clear differences in the consumption patterns between the two demographic groups. The estimates of the scale and the shift coefficients are expected to reflect these differences.

Furthermore, the log of total expenditure on the nondurables and services is our measure of the continuous endogenous explanatory variable, X_1. In our analysis that follows, the log of normal weekly disposable head of household income, specifically variable *p389* of the table *set3*, is used as an instrument. The two variables show strongly-positive correlation with the correlation coefficients of 0.5660 and 0.5954 for couples with and without children, respectively. Figures 3.1 and 3.2 present plots of the kernel estimates of the joint density for these variables. Finally, in the empirical application the instrument variable $Z = \Phi(\text{log earnings})$ is taken, similar to Blundell et al. (2007).

Figure 3.1. *Kernel joint density estimates for log total expenditure and log weekly income – couples with 1 or 2 children.*



3.3. Empirical Findings

The important empirical findings are now presented and summarized in Table 3.2. Although exact definitions of the data are not given in Blundell et al. (1998), Blundell et al. (1998) estimated the shape-invariant Engel curves for four broad categories of nondurables and services by using the FES data, namely fuel, fares, alcohol and leisure similar to this paper. The empirical estimate, $\hat{\gamma}$, of 0.36355 reported in the first column is very close to 0.3698 as found in Blundell et al. (1998). Furthermore, the signs of the parameter estimates, $\hat{\beta}_l$, for the four broad categories are all consistent with those of Blundell et al. (1998); specifically they are positive for food and leisure, but negative for alcohol, fares and fuel.

Figure 3.2. *Kernel joint density estimates for log total expenditure and log weekly income – couples without children.*

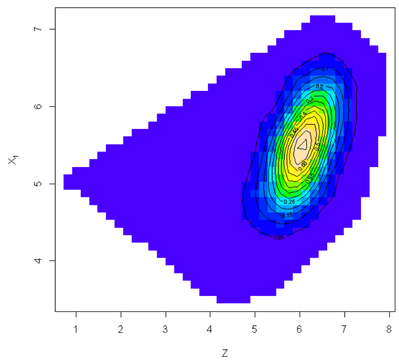
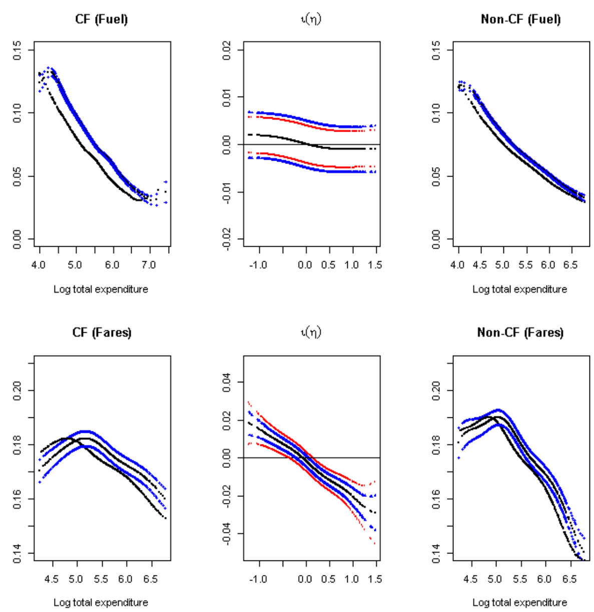


Table 3.2. *Empirical results*

$\hat{\gamma}$	Categories of goods	$\hat{\beta}_l$	$\hat{h}_{v,l}$	$\hat{h}_{\hat{\eta},l}$
0.36355	Fuel, light and power	-0.01401	0.14021	0.93631
	Fares, other travel costs and running of motor vehicles	-0.02027	0.19545	0.26831
	Food	0.00537	0.15120	0.25826
	Alcoholic Drink and Tobacco	-0.05205	0.30802	0.22569
	Leisure goods and services	0.05077	0.14663	0.40277
	Clothing and footwear	0.02079	0.14846	0.27234
	Personal goods and services	0.00738	0.49331	0.49335

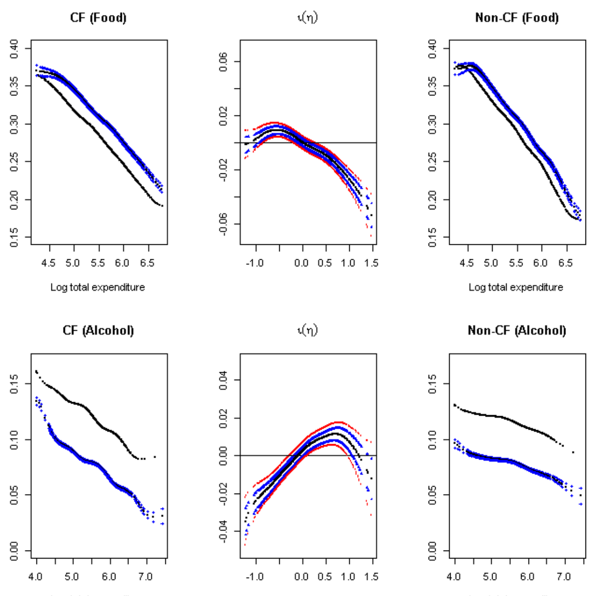
The first columns of Figures 3.3 to 3.6 present the empirical estimates of the Engel curves for seven of the goods in our system based on the CF approach discussed in Section 3.1. For these plots, the smoothing parameters presented in the fourth and fifth columns of Table 3.2 are used. Furthermore, the third columns of these figures show the empirical estimates of the Engel curves computed from the Xia et al. (1999)'s procedure by which the exogeneity assumption is imposed on the total expenditure. Together with the estimated Engel curves, their 90% point-wise confidence bands are also reported. The bands are obtained by using the procedure discussed in Section 3.1. Let us now concentrate on the first columns. For fuel, food and alcohol, the Engel curves appear to demonstrate that the Working-Leser linear logarithmic formulation may provide a reasonable approximation. Nonetheless, for other shares, especially for fares, a nonlinear relationship between the shares and the log expenditure is evident. A detailed investigation of the data shows that on average, up to 70% of fares belongs to running of motor vehicles. Hence, motor vehicles seemed to be a necessity good for a household for which the log of total expenditure is more than around 5.3 for those with children, for those without children, it is up to around 4.8. It seemed that motor vehicles are a superior good for those household where the log of total expenditure, is below these levels. The estimated shares for the couples with children are higher than those for couples without children, except extremely lower quantile of the log of total expenditure. This could lead to the nonlinear relationship witnessed in Figure 3.3.

Figure 3.3. *Fuel and fares (90% confidence bands drawn for households with children)*



As expected, the estimated shares of fuel and food for households with children are consistently above those for households without children. Couples without children spends around 3% more of their budget on fuel and food than couples with children. In addition, the estimated shares of alcohol, leisure, clothing and personal goods for households with children are consistently below those for households without children. Couples with children spend around 3%, 8% and 2% more of their budget on leisure, clothing and personal than couples with children at the same level of expenditure. In all but one case (i.e. fares), there seem to be a broadly parallel shift in the Engel curves from one demographic group to another. Our results suggest that fuel, food and alcohol may be categorized as necessity goods in the sense that the demand for these goods increases proportionally less than the increase in the total expenditure. These goods whose demand increases with the total expenditure are leisure, clothing and personal. The second column presents the nonparametric estimates of the control functions, $\iota_l(\cdot)$. With the estimated control functions, the two sets of bands, namely the 90% bias-corrected confidence bands for the nonparametric smoothing of Xia (1998) (blue) and the 90% Bonferroni-type variability bands of Eubank and Speckman (1993) (red) are also reported. Regarding fuel and personal, $\iota_l(\cdot)$ for these cases do not seem statistically significant. However, the opposite is found for fares, food, leisure and clothing. Hence, neglecting potential endogeneity in the estimation can lead to incorrect estimates of the shape of Engel curves for these goods. This can be seen by comparing the first and the third columns of the figures. For these goods it is clear that the curvature changes significantly as the presence of endogeneity is allowed.

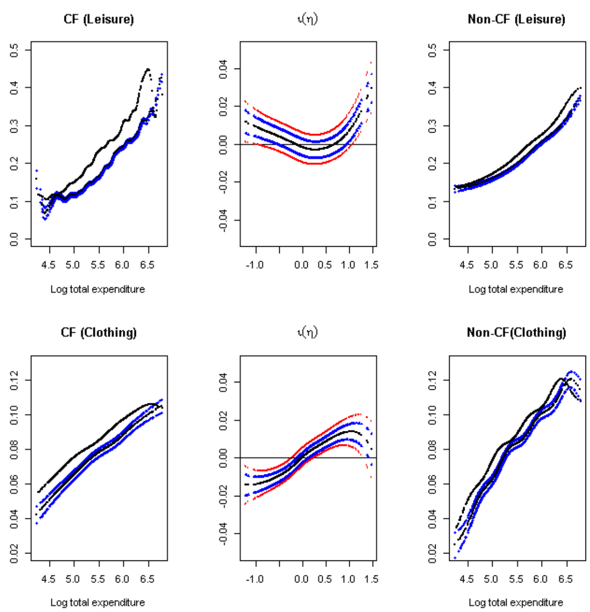
Figure 3.4. *Food and alcohol (90% confidence bands drawn for households with children)*



4. Conclusion

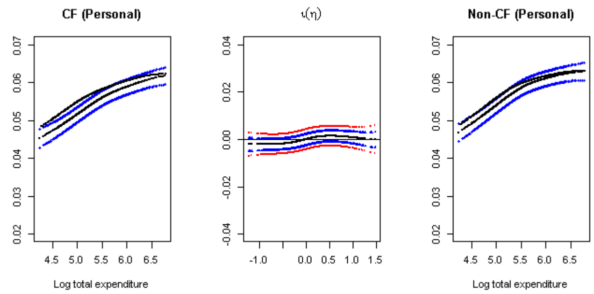
In this paper, the usefulness of the EGPLSI model in its ability to model a flexible shape-invariant specification is elaborated. A flexible shape-invariant specification is easily studied within the EGPLSI framework because both scale and shift parameters are easily incorporated in the EGPLSI model. However, the applicability of the EGPLSI model to an empirical study is limited because of its shortfalls in addressing endogeneity. Hence, the current paper develops the CF approach to address endogeneity in the EGPLSI model. The proposed CF approach inherits an intrinsic feature of the generated endogeneity control covariates and hence multi-stage nonparametric estimation procedure. This paper establishes the theoretical validity of the proposed estimation procedure and closes with the theoretical discussion by providing the straightforward extension of the results to a strictly stationary and α-mixing process. The paper also presents the satisfactory finite sample performance of proposed estimators from a Monte Carlo simulation exercise. Finally, the semi-parametric analysis of a system of shape-invariant empirical Engel curves using the FES (1995-96) data set within the framework of the EGPLSI model with our proposed CF approach is conducted. Not only are the findings interesting empirically but the accessible applicability of our proposed CF approach is also explored.

Figure 3.5. *Leisure and clothing (90% confidence bands drawn for households with children)*



Additionally, the development of the CF approach in this paper also provides the foundation for addressing the presence of weak instruments in the EGPLSI model. Han (2011) discussed how the intuitive triangular structure of the CF approach in a simple nonparametric regression model translates the difficult problem (the presence of weak instruments in a reduced-form equation) into a much simpler one, particularly the multicollinearity problem in a structural equation. Hence it is plausible to develop the current paper further to the presence of weak instruments case. However, a thorough investigation is required to examine a number of important issues, particularly examining the $\sqrt{n}$-consistent estimation of α_0 and β_0, and the properties of the smoothing parameters in each stage of an estimation procedure, and how to address the presence of weak instruments in the EGPLSI model.

Figure 3.6. *Engel curves for personal (90% confidence bands drawn for households with children)*



References

Ai, C., Chen, X., 2003. Efficient estimation of models with conditional moment restrictions containing unknown functions. Econometrica 71 (6), 1795–1843.

Blundell, R., Chen, X., Kristensen, D., 2007. Semi-nonparametric iv estimation of shape-invariant engel curves. Econometrica 75 (6), 1613–1669.

Blundell, R., Duncan, A., Pendakur, K., 1998. Semiparametric estimation and consumer demand. Journal of Applied Econometrics, 435–461.

Blundell, R., Powell, J. L., 2003. Endogeneity in nonparametric and semiparametric regression models. Econometric society monographs 36, 312–357.

Blundell, R., Stoker, T. M., 2007. Models of aggregate economic relationships that account for heterogeneity. Handbook of Econometrics 6, 4609–4666.

Blundell, R. W., Powell, J. L., 2004. Endogeneity in semiparametric binary response models. The Review of Economic Studies 71 (3), 655–679.

Byrd, R. H., Lu, P., Nocedal, J., Zhu, C., 1995. A limited memory algorithm for bound constrained optimization. SIAM Journal on Scientific Computing 16 (5), 1190–1208.

Carroll, R. J., Fan, J., Gijbels, I., Wand, M. P., 1997. Generalized partially linear single-index models. Journal of the American Statistical Association 92 (438), 477–489.

Eubank, R. L., Speckman, P. L., 1993. Confidence bands in nonparametric regression. Journal of the American Statistical Association 88 (424), 1287–1301.

Gao, J., 2007. Nonlinear Time Series: Semiparametric and Nonparametric Methods. CRC Press.

Gao, J., Lu, Z., Tjøstheim, D., et al., 2006. Estimation in semiparametric spatial regression. The Annals of Statistics 34 (3), 1395–1435.

Han, S., 2011. Nonparametric triangular simultaneous equations models with weak instruments.

Hansen, B. E., 2008. Uniform convergence rates for kernel estimation with dependent data. Econometric Theory 24 (3), 726–748.

Härdle, W., Hall, P., Ichimura, H., 1993. Optimal smoothing in single-index models. The Annals of Statistics 21 (1), 157–178.

Härdle, W., Marron, J. S., 1990. Semiparametric comparison of regression curves. The Annals of Statistics 18 (1), 63–89.

Hastie, T., Tibshirani, R., 1990. Generalized Additive Models. John Wiley & Sons, Inc.

Ichimura, H., 1993. Semiparametric least squares (sls) and weighted sls estimation of single-index models. Journal of Econometrics 58 (1-2), 71–120.

Li, Q., Racine, J. S., 2007. Nonparametric Econometrics: Theory and Practice. Princeton University Press.

Li, Q., Wooldridge, J. M., 2002. Semiparametric estimation of partially linear models for dependent data with generated regressors. Econometric Theory 18 (3), 625–645.

Linton, O., Nielsen, J. P., 1995. A kernel method of estimating structured nonparametric regression based on marginal integration. Biometrika 82 (1), 93–100.

Mammen, E., Rothe, C., Schienle, M., 2016. Semiparametric estimation with generated covariates. Econometric Theory 32 (5), 1140–1177.

Newey, W. K., Powell, J. L., Vella, F., 1999. Nonparametric estimation of triangular simultaneous equations models. Econometrica 67 (3), 565–603.

Pendakur, K., 1999. Semiparametric estimates and tests of base-independent equivalence scales. Journal of Econometrics 88 (1), 1–40.

Robinson, P., Pinkse, C., 1995. Pooling nonparametric estimates of regression functions with similar shape. In: Advances in Econometrics and Quantitative Economics. Wiley-Blackwell, 172–195.

Robinson, P. M., 1988. Root-n-consistent semiparametric regression. Econometrica 56 (4), 931–954.

Su, L., Ullah, A., 2008. Local polynomial estimation of nonparametric simultaneous equations models. Journal of Econometrics 144 (1), 193–218.

Tjøstheim, D., Auestad, B. H., 1994. Nonparametric identification of nonlinear time series: projections. Journal of the American Statistical Association 89 (428), 1398–1409.

Xia, Y., 1998. Bias-corrected confidence bands in nonparametric regression. Journal of the Royal Statistical Society: Series B (Statistical Methodology) 60 (4), 797–811.

Xia, Y., Härdle, W., 2006. Semi-parametric estimation of partially linear single-index models. Journal of Multivariate Analysis 97 (5), 1162–1184.

Xia, Y., Tong, H., Li, W., 1999. On extended partially linear single-index models. Biometrika 86 (4), 831–842.