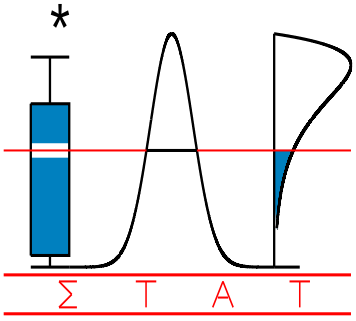
T E C H N I C A L

R E P O R T

0616

ESTIMATION OF A SEMIPARAMETRIC TRANSFORMATION MODEL

LINTON O., SPERLICH S. AND I. VAN KEILEGOM



I A P S T A T I S T I C S

N E T W O R K

INTERUNIVERSITY ATTRACTION POLE

http://www.stat.ucl.ac.be/IAP

Estimation of a Semiparametric Transformation Model*

Oliver Linton†
London School of Economics

Stefan Sperlich‡
Georg August Universität

Ingrid Van Keilegom§
Université catholique de Louvain

May 11, 2006

Abstract

This paper proposes consistent estimators for transformation parameters in semiparametric models. The problem is to find the optimal transformation into the space of models with a predetermined regression structure like additive or multiplicative separability. We give results for the estimation of the transformation when the rest of the model is estimated non- or semi-parametrically and fulfills some consistency conditions. We propose two methods for the estimation of the transformation parameter: maximizing a profile likelihood function or minimizing the mean squared distance from independence. First the problem of identification of such models is discussed. We then state asymptotic results for a general class of nonparametric estimators. Finally, we give some particular examples of nonparametric estimators of transformed separable models. The theoretical results as well as the small sample performance are studied by several simulation exercises.

Keywords: Transformation models, Generalized Structured Models, Semiparametric models, Separability.

*This research was supported by the Spanish "Dirección General de Investigación del Ministerio de Ciencia y Tecnología", number SEJ2004-04583/ECON. Linton would like to acknowledge financial support from the ESRC. Van Keilegom gratefully acknowledges financial support from the IAP research network nr. P5/24 of the Belgian government (Belgian Science Policy).

†Department of Economics, London School of Economics, Houghton Street, London WC2A 2AE, United Kingdom. E-mail address: o.linton@lse.ac.uk.

‡Institut für Statistik und Ökonometrie, Georg August Universität, Platz der Göttinger Sieben 5, D37073 Göttingen, Germany.

§Institut de Statistique, Université catholique de Louvain, Voie du Roman Pays 20, B 1348 Louvain-la-Neuve, Belgium.

1 Introduction

Taking transformations of the data has been an integral part of statistical practice for many years. Transformations have been used to aid interpretability as well as to improve statistical performance. An important contribution to this methodology was made by Box and Cox (1964) who proposed a parametric power family of transformations that nested the logarithm and the level. They suggested that the power transformation, when applied to the dependent variable in a linear regression setting, might induce normality, error variance homogeneity, and additivity of effects. They proposed estimation methods for the regression and transformation parameters. Carroll and Ruppert (1984) applied this and other transformations to both dependent and independent variables. A number of other dependent variable transformations have been suggested, for example the Zellner-Revankar transform, see Zellner and Revankar (1969). The transformation methodology has been quite successful and a large literature exists on this subject for parametric models, see Carroll and Ruppert (1988). There are also a number of applications to economics data: see Zarembka (1968), Zellner and Revankar (1969), Heckman and Polachek (1974), Ehrlich (1977), Hulten and Wykoff (1981).

In this work we concentrate on transformations in a regression setting. For many data, linearity of covariate effect after transformation may be too strong. For example, a respected study of the effects of schooling and experience on earnings (Heckman and Polachek (1974, p350)) found that while their data supported the logarithmic transformation of their dependent variable (earnings), it was "somewhat less clear on the functional form for the independent variables." We consider a rather general specification, allowing for nonparametric covariate effects without specifying their separability structure explicitly. Let X be a d-dimensional random vector and Y be a random variable, and let $\{(X_i, Y_i)\}_{i=1}^n$ be an i.i.d. sample from this population. Consider the estimation of the regression function $m(x) = E(Y \mid X = x)$. Stone (1980, 1982) and Ibragimov and Hasminskii (1980) showed that the optimal rate for estimating m is $n^{-\ell/(2\ell+d)}$, with ℓ a measure of the smoothness of m. This rate of convergence can be very slow for large dimensions d. One way of achieving better rates of convergence is making use of dimension reducing separability structures imposed e.g. by economic theory. A most typical example is additive or multiplicative modeling. An additive structure for m for example is a regression function of the form $m(x) = c + \sum_{\alpha=1}^d m_\alpha(x_\alpha)$, where $x = (x_1, \ldots, x_d)^\top$ are the d-dimensional predictor variables and m_α are one-dimensional nonparametric functions with $E[m_\alpha(X_\alpha)] = 0$. Stone (1986) showed that for such regression curves the optimal rate for estimating m is the one-dimensional rate of convergence $n^{-\ell/(2\ell+1)}$. Thus one speaks of dimensionality reduction through additive modeling.

We examine a semiparametric model that combines a parametric transformation with the flexibility of an additive nonparametric regression function. Suppose that

$$\Lambda(Y) = G(m_1(X_1), \ldots, m_d(X_d)) + \varepsilon, \tag{1}$$

where ε is independent of X, while G is a known function and Λ is a monotonic function. Special cases

of G are $G(z) = H(\sum_{\alpha=1}^{d} z_\alpha)$ and $G(z) = H(\prod_{\alpha=1}^{d} z_\alpha)$ for some strictly monotonic known function H. The general model in which Λ is monotonic and $G(z) = \sum_{\alpha=1}^{d} z_\alpha$ was previously addressed in Breiman and Friedman (1985) who suggested estimation procedures based on the iterative backfitting method, which they called ACE. However, they did not provide many results about the statistical properties of their procedures. See also Hastie and Tibshirani (1990). Linton, Chen, Wang, and Härdle (1997) considered the model with $\Lambda = \Lambda_\theta$ parametric and additive G, $G(z) = \sum_{\alpha=1}^{d} z_\alpha$. They proposed to estimate the parameters of the transformation Λ by either an instrumental variable method or a pseudo-likelihood method based on Gaussian ε. They assumed that identification held and did not provide justification for this from primitive conditions. Unfortunately, our simulation evidence suggests that both methods work poorly in practice and may even be inconsistent for many parameter configurations. To estimate the unknown functions m_α they used the marginal integration method of Linton and Nielsen (1995) and consequently their method can not achieve the semiparametric efficiency bound even in the few cases where Gaussian errors are well defined and their method is consistent.

We establish the nonparametric identification of the model (1) using results of Roehrig (1988). For practical reasons we propose estimation procedures only for the parametric transformation case where $\Lambda(y) = \Lambda_{\theta_o}(y)$ for some parametric family $\{\Lambda_\theta(.),\, \theta \in \Theta\}$ of transformations where $\Theta \subset \mathbb{R}^k$. In many studies on generalized linear models the misspecification of the index has turned out to be much more serious than the misspecification of the link. We suspect that something similar holds for transformation models: i.e. that a parametrisation of the transformation is less crucial than a parametric specification of the index. To estimate the transformation parameters we use two approaches. First, a semiparametric profile likelihood estimator (PL) that involves nonparametric estimation of the density of ε, and second a mean squared distance from independence method (MD) based on estimated c.d.f.'s of (X, ε). Both methods use a profiled estimate of the (separable) nonparametric components of m_θ. We use both the integration method and the smooth backfitting method of Mammen, Linton and Nielsen (1999) to estimate these components. The MD estimator involves discontinuous functions of nonparametric estimators and we use the theory of Chen, Linton and Van Keilegom (2003) to obtain its asymptotic properties. We derive the asymptotic distributions of our estimators under standard regularity conditions, and we show that the estimators of θ_o are root-n consistent.

The rest of the paper is organized as follows. In the next section we clarify identification issues. In Section 3 we introduce the two estimators for the transformation parameter. In Section 4 two alternative methods of additive modelling are discussed. Section 5 contains the asymptotic theory of the two estimators of the transformation parameter. Additionally, we discuss tools like bootstrap for possible inference on the transformation parameter. Finally, in Section 6 we study the finite sample performance of all methods presented and compare the different estimators of the transformation parameter as well as the different estimators of the additive components in this context. A special

emphasis is also given to the question of bandwidth choice. All proofs are deferred to Section 7 (Appendix A) and Section 8 (Appendix B).

2 Nonparametric Identification

The first question is the identification of model (1). We shall establish identification in the fully nonparametric model where $\Lambda(Y) = m(X) + \varepsilon$ and Λ and m are unknown functions under additional restrictions on the function m, while ε is independent of X. We show that additive and multiplicative separability are sufficient under normalization conditions. These restrictions are quite natural in economics applications, see e.g. Deaton and Muellbauer (1980), Blundell and Robin (2000), Rodriguez-Póo, Sperlich and Vieu (2003) or Stone (1986).

Breiman and Friedman (1985) defined $\Lambda, m_1, \ldots, m_d$ for general random variables $Y, X_1, \ldots, X_d$ as minimizers of the least squares objective function

$$e^2(\Lambda, m_1, \ldots, m_d) = \frac{E\left[\left\{\Lambda(Y) - \sum_{\alpha=1}^{d} m_\alpha(X_\alpha)\right\}^2\right]}{E\left[\Lambda^2(Y)\right]}.$$

They show the existence of minimizers and show that the set of minimizers forms a finite dimensional linear subspace under additional conditions. These conditions were that: (i) $\Lambda(Y) - \sum_{\alpha=1}^{d} m_\alpha(X_\alpha) = 0$ a.s. implies that $\Lambda(Y), m_\alpha(X_\alpha) = 0$ a.s., $\alpha = 1, \ldots, d$; (ii) $E[\Lambda(Y)] = 0, E[m_\alpha(X_\alpha)] = 0$, $E[\Lambda^2(Y)] < \infty$, and $E[m_\alpha^2(X_\alpha)] < \infty$; (iii) The conditional expectation operators $E[\Lambda(Y)|X_\alpha]$, $E[m_\alpha(X_\alpha)|Y]$, $\alpha = 1, \ldots, d$ are compact.

We establish unique identification under different conditions as we assume throughout that $\Lambda(Y) = m(X) + \varepsilon$ where ε is independent of X. We take the approach to nonparametric identification of Roehrig (1988).[1] Let us define

$$f(X, Y, \varepsilon) := \Lambda(Y) - m(X) - \varepsilon = 0. \tag{2}$$

Assume that f is a continuously differentiable function in all its arguments and that the distribution of (X, ε) is absolutely continuous with positive density on the set of interest. Further, let $f^*(X, Y, \varepsilon) := \Lambda^*(Y) - m^*(X) - \varepsilon = 0$ denote a function observationally equivalent to $f(\cdot, \cdot, \cdot)$ (Roehrig (1988), pp.435) and

$$N_\alpha = \begin{pmatrix} \partial f^* / \partial(x_\alpha, y) \\ \partial f / \partial(x_\alpha, y) \end{pmatrix}, \quad \alpha = 1, \ldots, d\ .$$

Then, Theorem 1 and Condition 3.2 of Roehrig (1988) tell us that model (2) is uniquely identified if

[1] The problems Benkard and Berry (2004) found in the article of Roehrig (1988) are not related to the results we are using here, since we restrict in our article to a single equation problem.

and only if: $|N_\alpha| = 0$ for $\alpha = 1, \ldots, d$ implies that $f^* = f$. Here, $|\cdot|$ means the determinant.[2] We will give two examples. In particular we show that additive (and likewise multiplicative) separability of the exogenous variables is sufficient to identify our model.

Without any structure on $m(\cdot)$, the model is not identified. We have

$$N_\alpha = \begin{pmatrix} -\partial m^*/\partial x_\alpha & \partial \Lambda^*/\partial y \\ -\partial m/\partial x_\alpha & \partial \Lambda/\partial y \end{pmatrix}, \quad \alpha = 1, \ldots, d,$$

whence

$$\frac{\partial m}{\partial x_\alpha}(x)\frac{\partial \Lambda^*}{\partial y}(y) = \frac{\partial m^*}{\partial x_\alpha}(x)\frac{\partial \Lambda}{\partial y}(y), \quad \alpha = 1, \ldots, d. \tag{3}$$

For all points x for which $\partial m(x)/\partial x_\beta \neq 0$, we have $\frac{\partial \Lambda^*}{\partial y} = \frac{\partial m^*}{\partial x_\beta}\frac{\partial \Lambda}{\partial y}\left(\frac{\partial m}{\partial x_\beta}\right)^{-1}$. Now, for a strictly monotonic Λ we get for any $\alpha \neq \beta$,

$$\frac{\partial m}{\partial x_\alpha}(x)\frac{\partial m^*}{\partial x_\beta}(x) = \frac{\partial m^*}{\partial x_\alpha}(x)\frac{\partial m}{\partial x_\beta}(x). \tag{4}$$

This has many solutions, as for example $m^* = m^a$ for any a. Therefore, even a parametrization of the transformation function Λ, respectively Λ^* does not automatically imply $\Lambda = \Lambda^*$ and $m = m^*$.

Suppose that we have an additive structure $m(x) = \sum_{\alpha=1}^d m_\alpha(x_\alpha)$. Then (4) becomes

$$\frac{\partial m_\beta}{\partial x_\beta}(x_\beta)\frac{\partial m_\alpha^*}{\partial x_\alpha}(x_\alpha) = \frac{\partial m_\beta^*}{\partial x_\beta}(x_\beta)\frac{\partial m_\alpha}{\partial x_\alpha}(x_\alpha).$$

Then, if X_α is not a function of X_β and Λ is nonlinear, we get

$$\frac{\partial m_\alpha^*}{\partial x_\alpha}(x_\alpha) = \frac{\partial m_\alpha}{\partial x_\alpha}(x_\alpha), \ \alpha = 1, \ldots, d; \ \frac{\partial \Lambda^*}{\partial y}(y) = \frac{\partial \Lambda}{\partial y}(y).$$

Therefore, m_α are identified up to a constant, which can be set by a location normalization on m_α, e.g., $E[m_\alpha(X_\alpha)] = 0$. Similarly, Λ is identified up to a constant, which is set by a location normalization on ε, like $E(\varepsilon) = 0$ or $q_\alpha(\varepsilon) = 0$, where q_α denotes the α quantile.

This identification result holds more generally. In particular, the pure multiplicative case is similar. Also, the cases where $G(z) = H(\sum_{\alpha=1}^d z_\alpha)$ or $G(z) = H(\prod_{\alpha=1}^d z_\alpha)$ for some strictly monotonic known function H are automatically identified by the above reasoning.

As we have seen, for identification it is not even necessary to parameterize Λ, but one would need several restrictions on it, and interpretation becomes difficult. Apart from that, having nonparametric functionals on both parts of model (2) renders the estimation problem impractical even for relatively large samples. On the other hand, for functional $m(\cdot)$ some structural assumptions, usually provided either by economic or biometric theory, are sufficient to identify our regression problem.

[2] We have applied here Condition 3.2 instead of Condition 3.1 because we restrict here for the ease of presentation to models where $\partial f^*/\partial \varepsilon^* = 0$ is always fulfilled in

$$BB^* = \{x, y, \varepsilon, \varepsilon^* : \ f(x, y, u) = 0, \ f^*(x, y, \varepsilon^*) = 0, (x, \varepsilon) \in D\},$$

with D being the joint support of X and ε. Notice that then our conclusions are clearly not affected by the criticism of Benkard and Berry (2004).

3 Estimating the optimal Transformation

In the sequel we consider the model

$$\Lambda_{\theta_o}(Y) = m(X) + \varepsilon, \tag{5}$$

where $\{\Lambda_\theta : \theta \in \Theta\}$ is a parametric family of strictly increasing functions, while the function $m(.)$ is of unknown form but with a certain predetermined structure that is sufficient to yield identification. We assume that the error term ε is independent of X and has distribution F. The covariate X is d-dimensional and has compact support $\mathcal{X} = \prod_{\alpha=1}^{d} R_{X_\alpha}$. Among the many transformations of interest, the following ones are used most commonly: (Box-Cox) $\Lambda_\theta(y) = \frac{y^\theta - 1}{\theta}$ $(\theta \neq 0)$ and $\Lambda_\theta(y) = \log(y)$ $(\theta = 0)$; (Zellner-Revankar) $\Lambda_\theta(y) = \ln y + \theta y^2$; (Arcsinh) $\Lambda_\theta(y) = \sinh^{-1}(\theta y)/\theta$. The arcsinh transform is discussed in Johnson (1949) and more recently in Robinson (1991). The main advantage of the arcsinh transform is that it works for y taking any value, while the Box-Cox and the Zellner-Revankar transforms are only defined if y is positive. For these transformations, the error term cannot be normally distributed except for a few isolated parameters, and so the Gaussian likelihood is misspecified. In fact, as Amemiya and Powell (1981) point out, the resulting estimators (in the parametric case) are inconsistent when only $n \to \infty$.

We let Θ denote a finite dimensional parameter set (a compact subset of $\mathbb{R}^k$) and $\mathcal{M}$ an infinite dimensional parameter set. We assume that $\mathcal{M}$ is a vector space of functions endowed with metric $\|\cdot\|_{\mathcal{M}} = \|\cdot\|_\infty$. We denote $\theta_o \in \Theta$ and $m_o \in \mathcal{M}$ as the true unknown finite and infinite dimensional parameters. Define the regression function

$$m_\theta(x) = E[\Lambda_\theta(Y) \,|X = x]$$

for each $\theta \in \Theta$. Note that $m_{\theta_o}(\cdot) \equiv m_o(\cdot)$.

We suppose that we have a randomly drawn sample $Z_i = (X_i, Y_i)$, $i = 1, \ldots, n$, from model (5). Define, for $\theta \in \Theta$ and $m \in \mathcal{M}$,

$$\varepsilon(\theta, m) = \Lambda_\theta(Y) - m(X),$$

and let $\varepsilon_\theta = \varepsilon(\theta) = \varepsilon(\theta, m_\theta)$, and $\varepsilon_o = \varepsilon_{\theta_o}$. When there is no ambiguity, we also use the notations ε and m to indicate ε_o and m_o. Moreover, let $\Lambda_o = \Lambda_{\theta_o}$.

3.1 The Profile Likelihood (PL) Estimator

The method of profile likelihood has already been applied to many different semiparametric estimation problems. The basic idea is simply to replace all unknown expressions of the likelihood function by their nonparametric (kernel) estimates. We consider $\Lambda_\theta(Y) = m_\theta(X) + \varepsilon_\theta$ for any $\theta \in \Theta$. Then, the cumulative distribution function is

$$\Pr[Y \leq y|X] = \Pr[\Lambda_\theta(Y) \leq \Lambda_\theta(y)|X] = \Pr[\varepsilon_\theta \leq \Lambda_\theta(y) - m_\theta(X)|X] = F_{\varepsilon(\theta)}(\Lambda_\theta(y) - m_\theta(X)),$$

where $F_{\varepsilon(\theta)}(e) = F_{\varepsilon(\theta,m_\theta)}(e)$ and $F_{\varepsilon(\theta,m)} = P(\varepsilon(\theta,m) \leq e)$, and so

$$f_{Y|X}(y|x) = f_{\varepsilon(\theta)}(\Lambda_\theta(y) - m_\theta(x))\Lambda'_\theta(y)$$

where $f_{\varepsilon(\theta)}$ and $f_{Y|X}$ are the probability density functions of $\varepsilon(\theta)$ and of Y given X. Then, the log likelihood function is

$$\sum_{i=1}^{n} \Big\{ \log f_{\varepsilon(\theta)}(\Lambda_\theta(Y_i) - m_\theta(X_i)) + \log \Lambda'_\theta(Y_i) \Big\}.$$

Let $\widehat{m}_\theta(.)$ be one of our estimators (see Section 4 below) of $m_\theta(.)$, and let

$$\widehat{f}_{\varepsilon(\theta)}(e) := \frac{1}{ng} \sum_{i=1}^{n} K_2 \left(\frac{e - \widehat{\varepsilon}_i(\theta)}{g} \right), \tag{6}$$

with $\widehat{\varepsilon}_i(\theta) = \widehat{\varepsilon}_i(\theta, m_\theta)$ and $\widehat{\varepsilon}_i(\theta, m) = \varepsilon_i(\theta, \widehat{m}) = \Lambda_\theta(Y_i) - \widehat{m}(X_i)$. Here, K_2 is a scalar kernel and g is a bandwidth sequence. Then, define the profile likelihood estimator of θ_o by

$$\widehat{\theta}_{PL} = \arg\max_{\theta\in\Theta} \sum_{i=1}^{n} \Big[\log \widehat{f}_{\varepsilon(\theta)}(\Lambda_\theta(Y_i) - \widehat{m}_\theta(X_i)) + \log \Lambda'_\theta(Y_i) \Big]. \tag{7}$$

The computation of $\widehat{\theta}_{PL}$ can be done by grid search in the scalar case and using derivative-based algorithms in higher dimensions, assuming that the kernels are suitably smooth.

3.2 Mean Square Distance from Independence (MD) Estimator

Although the profile likelihood approach yields an efficient estimator for the transformation under certain conditions, there are four good reasons why it is worth providing alternatives when it comes to practical work. First, as we will see in Section 6, the profile likelihood method is computationally quite expensive. In particular, so far we have not found a reasonable implementation for the recentered bootstrap.Second, for that approach we do not only face the typical question of bandwidth choice for the nonparametric part m_θ, we additionally face a bandwidth for the density estimation, see equation (6). Third, there are some transformation models Λ_θ for which the support of Y depends on the parameter θ and so are non-regular. Finally, although the estimator we get from the profile likelihood is under certain conditions efficient in the asymptotic sense, Severini and Wong (1992), this tells us little about its finite sample performance, neither in absolute terms nor in comparison with competitors.

One possible and computationally attractive competitor is the minimization of the mean square distance from independence. Why it is computationally more attractive will be explained in Section 6. This method we will introduce here has been reviewed in Koul (2001) for other problems.

Define, for each $\theta \in \Theta$ and $m \in \mathcal{M}$, the empirical distribution functions

$$\widehat{F}_X(x) = \frac{1}{n} \sum_{i=1}^{n} 1(X_i \leq x) \; ; \; \widehat{F}_{\varepsilon(\theta)}(e) = \frac{1}{n} \sum_{i=1}^{n} 1(\widehat{\varepsilon}_i(\theta) \leq e) \; ;$$

$$\widehat{F}_{X,\varepsilon(\theta)}(x,e) = \frac{1}{n}\sum_{i=1}^{n} 1(X_i \leq x)1(\widehat{\varepsilon}_i(\theta) \leq e),$$

the moment function

$$G_{nMD}(\theta, \widehat{m}_\theta)(x,e) = \widehat{F}_{X,\varepsilon(\theta)}(x,e) - \widehat{F}_X(x)\widehat{F}_{\varepsilon(\theta)}(e)$$

and the criterion function

$$\|G_{nMD}(\theta, \widehat{m}_\theta)\|_2^2 = \int \left[G_{nMD}(\theta, \widehat{m}_\theta)(x,e)\right]^2 d\mu(x,e) \tag{8}$$

for some probability measure μ. We define an estimator of θ, denoted $\widehat{\theta}_{MD}$, as any approximate minimizer of $\|G_{nMD}(\theta, \widehat{m}_\theta)\|_2^2$ over Θ. To be precise let

$$\|G_{nMD}(\widehat{\theta}_{MD}, \widehat{m}_{\widehat{\theta}})\|_2 = \inf_{\theta\in\Theta} \|G_{nMD}(\theta, \widehat{m}_\theta)\|_2 + o_p(1/\sqrt{n}).$$

There are many algorithms available for computing the optimum of general non-smooth functions, e.g., the Nelder-Mead, and the more recent genetic and evolutionary algorithms.

We can use in (8) the empirical measure $d\mu_n$ of $\{X_i, \widehat{\varepsilon}_i(\theta)\}_{i=1}^n$, which results in a criterion function

$$Q_n(\theta) = \frac{1}{n}\sum_{i=1}^{n} \left[G_{nMD}(\theta, \widehat{m}_\theta)(X_i, \widehat{\varepsilon}_i(\theta))\right]^2. \tag{9}$$

4 Estimating the Nonparametric Index

We here discuss how to estimate the function m_θ imposing the structure we have assumed. We only discuss here the additive case

$$m(x) = c + \sum_{\alpha=1}^{d} m_\alpha(x_\alpha),$$

where $E[m_\alpha(X_\alpha)] = 0$. We start with the marginal integration [MI in the sequel] estimator. For each $\alpha = 1, \ldots, d$, partition $x = (x_\alpha, x_{\underline{\alpha}})$, where x_α is a one-dimensional direction of interest and $x_{\underline{\alpha}}$ is a $(d-1)$-dimensional nuisance direction, likewise with $X = (X_\alpha, X_{\underline{\alpha}})$ and $X_i = (X_{\alpha i}, X_{\underline{\alpha} i})$. Let f_X be the covariate density and f_{X_α} be its marginals, and let $f_{X_{\underline{\alpha}}}$ be the joint density of $X_{\underline{\alpha}}$.

For each θ, we first estimate $m_\theta(x)$ by local linear regression. That is, let $(\widehat{a}, \widehat{b})$ minimize the following localized least squares criterion

$$\sum_{i=1}^{n} K\left(\frac{x - X_i}{h}\right)\left[\Lambda_\theta(Y_i) - a - b^\top(X_i - x)\right]^2, \tag{10}$$

where $K(t) = \prod_{j=1}^{d} k(t_j)$ and k is a univariate kernel function, while $h = h(n)$ is a bandwidth. Then let $\widehat{m}_\theta(x) = \widehat{a}$. Now define

$$\widehat{\gamma}_\alpha(x_\alpha; \theta) = \frac{1}{n}\sum_{i=1}^{n} \widehat{m}_\theta(x_\alpha, X_{\underline{\alpha} i}). \tag{11}$$

This estimator goes back to Newey (1994), Tjøstheim and Auestad (1994), and Linton and Nielsen (1995). We will use here the improved version of Kim, Linton and Hengartner (1999) and Hengartner and Sperlich (2005). Now, let

$$\widehat{m}_\theta^{MI}(x) = \sum_{\alpha=1}^{d} \widehat{\gamma}_\alpha(x_\alpha;\theta) - (d-1)\widehat{c}_\theta \ , \tag{12}$$

where $\widehat{c}_\theta = n^{-1}\sum_{i=1}^n \Lambda_\theta(Y_i)$.

A second estimator of $m_\theta(x)$ is the so called smooth backfitting [BF in the sequel] estimator which we denote by $\widehat{m}_\theta^{BF}(x)$. It has been introduced by Mammen, Linton and Nielsen (1999). Here, we just give a brief definition, see Nielsen and Sperlich (2005) for implementation, finite sample performance, bandwidth choice and further explanation of the method. We define the 'empirical projection' estimates $\{\widehat{m}_\alpha^{BF}(\cdot), \alpha = 1,\ldots,d, \widehat{m}_0^{BF}\}$ as the minimizers of the following criterion

$$\int \left[\widehat{m}_\theta(x) - \bar{m}_0 - \bar{m}_1(x_1) - \ldots - \bar{m}_d(x_d)\right]^2 \widehat{f}(x)dx, \tag{13}$$

where the minimization runs over all functions $\overline{m}(x) = \bar{m}_0 + \sum_\alpha \bar{m}_\alpha(x_\alpha)$, with $\int \bar{m}_\alpha(x_\alpha)\widehat{f}_\alpha(x_\alpha)dx_\alpha = 0$, where $\widehat{f}_\alpha(x_\alpha) = \int \widehat{f}(x)dx_{\underline{\alpha}}$ is the marginal of the density estimate $\widehat{f}(x) = n^{-1}h^{-d}\sum_{i=1}^n K\left(\frac{x-X_i}{h}\right)$. This is the one-dimensional kernel density estimate $\widehat{f}_\alpha(x_\alpha) = n^{-1}\sum_{i=1}^n K_h(x_\alpha - X_{\alpha i})$. A minimizer of (13) exists if the density estimate $\widehat{f}$ is non-negative.

It has been shown that $\widehat{\gamma}_\alpha(x_\alpha;\theta)$ consistently estimates the population quantity $\gamma_\alpha(x_\alpha;\theta) = \int m_\theta(x_\alpha, x_{\underline{\alpha}}) f_{X_{\underline{\alpha}}}(x_{\underline{\alpha}})dx_{\underline{\alpha}}$. Under the additive model, $\gamma_\alpha(x_\alpha;\theta_o) = c + m_\alpha(x_\alpha)$. Then, $\widehat{m}_\theta^{MI}(x)$ estimates consistently

$$m_\theta^{MI}(x) = \sum_{\alpha=1}^{d} \gamma_\alpha(x_\alpha;\theta) - (d-1)c_\theta,$$

where $c_\theta = E[\Lambda_\theta(Y)]$. Note that $m_{\theta_o}^{MI}(x) = m_{\theta_o}(x)$, i.e. combining the covariate effects in the θ_o scale gives us the regression function. However, $m_\theta^{MI}(x) \neq m_\theta(x)$ for $\theta \neq \theta_o$. This information is used to identify θ_o. Regarding the backfitting estimator, $\widehat{m}_\theta^{BF}(x)$ consistently estimates a function $m_\theta^{BF}(x)$, where $m_{\theta_o}^{BF}(x) = m_{\theta_o}(x)$, but $m_\theta^{BF}(x) \neq m_\theta(x)$ for $\theta \neq \theta_o$. We give an interpretation to the functions $m_\theta^{MI}(x)$ and $m_\theta^{BF}(x)$. Define the subspace of additive functions

$$\mathcal{M}_{add} = \{m : \ m(x) = \sum_{\alpha=1}^{d} m_\alpha(x_\alpha) \quad \text{for some} \quad m_1(.),\ldots,m_d(.)\},$$

and define the two additive approximations

$$m_{\theta,add}(.) = \arg\min_{m\in\mathcal{M}_{add}} \int \left[(m_\theta(X) - m(X))^2\right] f_X(X)dX$$

$$m_{\theta,addprod}(.) = \arg\min_{m\in\mathcal{M}_{add}} \int \left[(m_\theta(X) - m(X))^2\right] \prod_{\alpha=1}^{d} f_{X_\alpha}(X_\alpha)dX_\alpha.$$

Nielsen and Linton (1998) showed that $m_\theta^{MI} = m_{\theta,addprod}$. Mammen, Linton and Nielsen (1999) show that $m_\theta^{BF} = m_{\theta,add}$. In general, these will be different functions, and will have different derivatives at θ_0. Define

$$\frac{\partial m_\theta^{MI}}{\partial \theta}(\cdot) = \arg\min_{m \in \mathcal{M}_{add}} \int \left[\left(\frac{\partial m_\theta}{\partial \theta}(X) - m(X) \right)^2 \right] \prod_{\alpha=1}^{d} f_{X_\alpha}(X_\alpha) dX_\alpha \tag{14}$$

$$\frac{\partial m_\theta^{BF}}{\partial \theta}(\cdot) = \arg\min_{m \in \mathcal{M}_{add}} \int \left[\left(\frac{\partial m_\theta}{\partial \theta}(X) - m(X) \right)^2 \right] f_X(X) dX. \tag{15}$$

These functions play an important part in the limiting distributions below.

In the sequel we will denote m_θ to indicate either the function $E[\Lambda_\theta(Y)|X = \cdot]$ or the functions m_θ^{BF} and m_θ^{MI} defined above. It will be clear from the context which function it represents.

5 Asymptotic Properties

We now discuss the asymptotic properties of our procedures. Note that although nonparametric density estimation with non- or semiparametrically constructed variables has already been considered in Van Keilegom and Veraverbeke (2002) and in Sperlich (2005), their results cannot be applied directly to our problem. The first ones treated the more complex problem of censored regression models but have no additional parameter like our θ. Nevertheless, as they consider density estimation with nonparametrically estimated residuals, their results come much closer to our needs than the second paper. Neither offer results on derivative estimation. As we will see now, this we need when we translate our estimation problem into the estimation framework of Chen, Linton and Van Keilegom (2003) [CLV in the sequel].

To be able to apply the results of CLV for proving the asymptotics of the profile likelihood, we need an objective function that takes its minimum at θ_o. Therefore we introduce some notation. For any function φ we define $\dot{\varphi} := \partial\varphi/\partial\theta$ and $\dot{\widehat{\varphi}} := \partial\widehat{\varphi}/\partial\theta$ respectively. Similarly we define for any function φ: $\varphi'(u) := \partial\varphi(u)/\partial u$ and $\widehat{\varphi}'(u) := \partial\widehat{\varphi}(u)/\partial u$ respectively. The same holds for any combination of primes and dots.

We use the abbreviated notation $s = (m, r, f, g, h)$, $s_\theta = (m_\theta, \dot{m}_\theta, f_{\varepsilon(\theta)}, f'_{\varepsilon(\theta)}, \dot{f}_{\varepsilon(\theta)})$, $s_o = s_{\theta_o}$, and $\widehat{s}_\theta = (\widehat{m}_\theta, \dot{\widehat{m}}_\theta, \widehat{f}_{\varepsilon(\theta)}, \widehat{f}'_{\varepsilon(\theta)}, \dot{\widehat{f}}_{\varepsilon(\theta)})$.

Then, define for any $s = (m, r, f, g, h)$,

$$\begin{aligned} &G_{nPL}(\theta, s) \\ &= n^{-1} \sum_{i=1}^{n} \left\{ \frac{1}{f\{\varepsilon_i(\theta, m)\}} [g\{\varepsilon_i(\theta, m)\}\{\dot{\Lambda}_\theta(Y_i) - r(X_i)\} + h\{\varepsilon_i(\theta, m)\}] + \frac{\dot{\Lambda}'_\theta(Y_i)}{\Lambda'_\theta(Y_i)} \right\}, \end{aligned} \tag{16}$$

and let $G_{PL}(\theta, s) = E[G_{nPL}(\theta, s)]$, and $\Gamma_{1PL} = \frac{\partial}{\partial\theta} G_{PL}(\theta, s_\theta)\big\downarrow_{\theta=\theta_o}$.

Note that $\|G_{PL}(\theta, s_\theta)\|$ and $\|G_{nPL}(\theta, \widehat{s}_\theta)\|$ take their minimum at θ_o and $\widehat{\theta}_{PL}$ respectively. We assume in the appendix that the estimators $\widehat{m}_{MI}$ and $\widehat{m}_{BF}$ obey a certain asymptotic expansion. The proof of such expansions can be found in Lemmas 6.1 and 6.2 of Mammen and Park (2005) for backfitting and in Linton, Chen, Wang and Härdle (1997) for marginal integration. In consequence we obtain expansions for $\widehat{f}_{\varepsilon(\theta)}, \widehat{f}'_{\varepsilon(\theta)}, \dot{\widehat{f}}_{\varepsilon(\theta)}$.

Theorem 1. *Under assumptions A.1–A.8 given in Appendix A, we have*

$$\widehat{\theta}_{PL} - \theta_o = -\Gamma_{1PL}^{-1} G_{nPL}(\theta_o, s_o) + o_p(n^{-1/2}),$$

and hence

$$\sqrt{n}(\widehat{\theta}_{PL} - \theta_o) \Longrightarrow N(0, \Omega_{PL}),$$

where

$$\Omega_{PL} = \Gamma_{1PL}^{-1} \mathrm{Var}\{G_{1PL}(\theta_o, s_o)\}(\Gamma_{1PL}^T)^{-1}.$$

Note that the variance of $\widehat{\theta}_{PL}$ equals the variance of the estimator of θ_o that is based on the true (unknown) values of the nuisance functions $m_o, \dot{m}_o, f_\varepsilon, f'_\varepsilon$ and $\dot{f}_\varepsilon$. When $m_o = m_{\theta_o}^{BF}$, we expect that the profile likelihood estimator is semiparametrically efficient following Severini and Wong (1992), see also Linton and Mammen (2005).

We obtain the asymptotic distribution of $\widehat{\theta}_{MD}$ using a modification of Theorems 1 and 2 of Chen, Linton and Van Keilegom (2003). That result applied to the case where the norm in (8) was finite dimensional, although their Theorem 1 is true as stated with the more general norm. Regarding their Theorem 2, we need to modify only Condition 2.5 to take account of the fact that $G_{nMD}(\theta, m_\theta)$ is a stochastic process in (x, e). Let $\lambda_\theta(y) = \dot{\Lambda}_\theta(y) = \partial \Lambda_\theta(y)/\partial\theta$ and let $\lambda_o = \lambda_{\theta_o}$. We also note that

$$\frac{\partial}{\partial \theta} E[\Lambda_\theta(Y)|X]\Big|_{\theta=\theta_o} = \int \lambda_o(\Lambda_o^{-1}(m_o(X) + e)) f_\varepsilon(e) de.$$

Define the matrix

$$\Gamma_{1MD}(x, e) = f_\varepsilon(e) E\left[(1(X \le x) - F_X(x))\left(\lambda_o(\Lambda_o^{-1}(m_o(X) + e)) + \dot{m}_o(X)\right)\right],$$

and the i.i.d. mean zero and finite variance random variables

$$\begin{aligned}\overline{U}_i = &\int [1(X_i \le x) - F_X(x)][1(\varepsilon_i \le e) - F_\varepsilon(e)]\Gamma_{1MD}(x, e) d\mu(x, e) \\ &+ f_X(X_i) \sum_{\alpha=1}^{d} v_{o1\alpha}(X_{\alpha i}, \varepsilon_i) \int f_\varepsilon(e) \left(1(X_i \le x) - F_X(x)\right) \Gamma_{1MD}(x, e) d\mu(x, e).\end{aligned}$$

Let $V_{1MD} = E[\overline{U}_i \overline{U}_i^\top]$ and $\overline{\Gamma}_{1MD} = \int \Gamma_{1MD}(x, e)\Gamma_{1MD}^T(x, e) d\mu(x, e)$.

Theorem 2. *Under the assumptions B.1–B.8 given in Appendix B, we have*

$$\widehat{\theta}_{MD} - \theta_o = -\overline{\Gamma}_{1MD}^{-1}\overline{U_i} + o_p(n^{-1/2}),$$

and hence,

$$\sqrt{n}(\widehat{\theta}_{MD} - \theta_o) \Longrightarrow N(0, \Omega_{MD}),$$

where

$$\Omega_{MD} = \overline{\Gamma}_{1MD}^{-1} V_{1MD} \overline{\Gamma}_{1MD}^{-1}.$$

REMARKS.

1. Bootstrap standard errors. CLV proposes and justifies the use of the ordinary bootstrap. Let $\{Z_i^*\}_{i=1}^n$ be drawn randomly with replacement from $\{Z_i\}_{i=1}^n$, let

$$G_{nMD}^*(\theta, m)(x, e) = \widehat{F}_{X\varepsilon(\theta)}^*(x, e) - \widehat{F}_X^*(x)\widehat{F}_{\varepsilon(\theta)}^*(e),$$

where $\widehat{F}_{X\varepsilon(\theta)}^*, \widehat{F}_X^*(x)$, and $\widehat{F}_{\varepsilon(\theta)}^*$ are computed from the bootstrap data. Let also $\widehat{m}_\theta^*(\cdot)$ (for each θ) be the same estimator as $\widehat{m}_\theta(\cdot)$ but based on the bootstrap data. Following Hall and Horowitz (1996, p897) it is necessary to recenter the moment condition, at least in the overidentified case. Thus, define the bootstrap estimator $\widehat{\theta}_{MD}^*$ to be any sequence that satisfies

$$\|G_{nMD}^*(\widehat{\theta}_{MD}^*, \widehat{m}_{\widehat{\theta}_{MD}^*}^*) - G_{nMD}(\widehat{\theta}_{MD}, \widehat{m}_{\widehat{\theta}_{MD}})\| = \inf_{\theta\in\Theta} \|G_{nMD}^*(\theta, \widehat{m}_\theta^*) - G_{nMD}(\widehat{\theta}_{MD}, \widehat{m}_{\widehat{\theta}_{MD}})\| + o_{p^*}(n^{-1/2}), \tag{17}$$

where superscript $*$ denotes a probability or moment computed under the bootstrap distribution conditional on the original data set $\{Z_i\}_{i=1}^n$. The resulting bootstrap distribution of $\sqrt{n}(\widehat{\theta}_{MD}^* - \widehat{\theta}_{MD})$ can be shown to be asymptotically the same as the distribution of $\sqrt{n}(\widehat{\theta}_{MD} - \theta_o)$, by following the same arguments as in the proof of Theorem B in CLV .

2. Estimated weights. Suppose that we have estimated weights $\mu_n(x, e)$ that satisfy $\sup_{x,e} |\mu_n(x, e) - \mu(x, e)| = o_p(1)$. Then the estimator computed with the estimated weights $\mu_n(x, e)$ has the same distribution theory as the estimator that used the limiting weights $\mu(x, e)$.

3. Note that the asymptotic distributions in Theorem 1 and 2 do not depend on the details of the estimators $\widehat{m}_\theta^{MI}(x)$ and $\widehat{m}_\theta^{BF}(x)$ only on their population interpretations through (14) and (15).

6 Simulations

In this section our interest is directed to the performance of our methods for distinct models, error variances, and sample sizes. But we are also interested in practical questions like bandwidth choice and computational expense.

We will work with the following data generating process:

$$\Lambda_\theta(Y) = b_0 + b_1 X_1^2 + b_2 \sin(\pi X_2) + \varepsilon\sigma_e, \tag{18}$$

where Λ_θ is the Box–Cox transformation, $\varepsilon \sim N(0,1)$ but restricted on $[-3,3]$ and $X_1, X_2 \sim U[-0.5,0.5]^2$. We study three different models setting $b_0 = 3.0\sigma_e + b_2$ and b_1, b_2, σ_e as follows: for model1 $b_1 = 5.0$, $b_2 = 2.0$, $\sigma_e = 1.5$; for model2 $b_1 = 3.5$, $b_2 = 1.5$, $\sigma_e = 1.0$; and for model3 $b_1 = 2.5$, $b_2 = 1.0$, $\sigma_e = 0.5$. Note that the setting of all parameter and error distribution has been chosen such that the variable $\Lambda_\theta(Y)$ is positive to avoid problems when generating the Y for arbitrary $\theta \in [-0.5, 1.5]$ in our simulations.

We have done simulations for the cases when the real data generating parameter θ_o was set to $0.0, 0.5$ or 1.0. The estimate was taken from a grid of step length 0.0625 on the interval $[-0.5, 1.5]$. The additive model has been estimated by the two above mentioned approaches: by marginal integration (MI) and by smooth backfitting (BF). We used the quartic kernel $K(u) = \frac{15}{16}(1-u^2)_+^2$ throughout. We chose $h_1 = h_2 = n^{-1/5}h_0$ for a large range of h_0 – values. Further, for the MI, where it is allowed or even recommended to choose larger bandwidths (let us call them g_α) in the nuisance directions, we set $g_1 = g_2 = 2 \cdot h_1$ due to our experiences from Hengartner and Sperlich (2005). Certainly, neither setting $h_1 = h_2$ nor $g_1 = g_2 = 2h_1$ is always optimal, especially not when the additive components have rather different smoothness or the variables X_α differ a lot in distribution. However, both cases hardly meet here so we think our choice is fair enough for our purposes. For the density estimator of the estimated residuals in the PL we used the Silverman's rule of thumb bandwidth in each iteration of the maximization: $1.06n^{-1/5}\frac{\widehat{IR}}{1.34}$ where $\widehat{IR}$ denotes the estimated interquartile range of the $\widehat{\varepsilon}_i(\theta)$, $i = 1, \ldots, n$. This is a quite reasonable choice in practice as long as the residuals follow a somehow bell-shaped distribution.

6.1 Comparing PL vs MD and MI vs BF

First, we do some basic considerations like general performance in comparison (MD and PL, MI and BF) and robustness against bandwidths choice, all for different θ_o. To this end, we generated 500 samples of size $n = 100$ for each combination of estimator. Tables 1 and 2 give the means and standard deviations calculated from these 500 replications for each data generating θ_o and different bandwidth $h_0 n^{-1/5}$. Notice that due to the fact that we always set $\Theta = [-0.5, 1.5]$, the simulation results for $\theta_o = 0.0$, respectively 1.0, are biased towards the interior of the interval. Note further that there is also a relation between bandwidth and θ (the estimated one as well as the real one), that is the smoothness of the model. As we use local constant smoothers, the estimates will have more bias for larger derivatives ("steeper functionals"). On the other hand, both a smaller θ and a larger h_0 make the data "smoother" and the other way around. Thus, some unwanted interaction, even if asymptotically vanishing, would not be surprising.

Table 1 gives the results for any combination of model, bandwidth and method, always keeping the same nonparametric smoother (here MI) to estimate the additive components. The effect of using different smoothers (MI versus BF) can be seen when comparing later Table 1 with Table 2 where the same results for BF are presented.

Both Methods when using MI

	$\theta_o =$	.00	.50	1.0	.00	.50	1.0	.00	.50	1.0
	$h_0 =$		0.4			0.5			0.6	
MD	model1	.0393	.6163	.8692	.0394	.6210	.8890	.0368	.6269	.8995
		.1640	.5767	.6578	.1453	.5611	.6604	.1431	.5547	.6471
		.0284	.3461	.4499	.0227	.3295	.4485	.0218	.3238	.4287
	model2	.0642	.6210	.8656	.0625	.6389	.9015	.0566	.6363	.9067
		.2178	.6026	.6855	.1909	.5662	.6390	.1926	.5684	.6505
		.0516	.3778	.4879	.0403	.3399	.4181	.0403	.3417	.4318
	model3	.1434	.6744	.8825	.1241	.6875	.9519	.1169	.6911	.9611
		.3468	.6336	.6766	.2965	.5847	.6265	.2877	.5835	.6256
		.1408	.4319	.4717	.1033	.3770	.3948	.0964	.3770	.3929
PL	model1	.0013	.4209	.8083	.0004	.4372	.8153	.0049	.4392	.8111
		.0811	.3351	.5213	.0856	.3501	.5342	.0877	.3512	.5277
		.0066	.1186	.3085	.0073	.1265	.3195	.0077	.1270	.3142
	model2	-.002	.4436	.8290	.0027	.4471	.8302	.0056	.4450	.8114
		.1123	.3496	.5197	.1118	.3570	.5260	.1116	.3646	.5280
		.0126	.1254	.2993	.0125	.1303	.3055	.0125	.1360	.3144
	model3	.0076	.4799	.8698	.0078	.4778	.8581	.0125	.4777	.8586
		.1731	.3867	.5215	.1675	.3873	.5104	.1730	.3957	.5119
		.0300	.1499	.2889	.0281	.1505	.2806	.0301	.1571	.2820

Table 1: *Performance of MD and PL method with MI estimator: Means (first line), standard deviations (second line), and means squared error (third line) of the* $\widehat{\theta}$ *for different* θ_o*, models [see (18)], and bandwidths* $h_\alpha = h_0 n^{-1/5}$*,* $\alpha = 1, 2$*, for sample size* $n = 100$*. All numbers calculated from 500 replications.*

It is clear by its definition that if the error distribution is small compared to estimation error, then the MD is expected to do worse. Indeed, even though model3 is the smoothest model and therefore the easiest estimation problem, for the smallest error ($\sigma_e = 0.5$) the MD does worse. In those cases the PL estimator should perform better and so it does. It might be surprising that θ mostly gets better estimated in model1 than in model2 and model3, where the nonparametric functionals are much easier to estimate, but notice that for the quality of $\widehat{\theta}$ the relation between estimation error and model error is more important. This holds also true for the PL method. Nevertheless, at least for small samples none of the estimators seems to outperform uniformly the other: so the PL has mostly smaller variance whereas MD has mostly smaller bias. We later compare the methods for n

increasing to $n = 1000$. Note that this finding does not depend on the particular smoother MI or BF (compare Table 2). As expected, for very small samples the results depend on the bandwidth. For this reason, and due to its importance in practice we will investigate this problem more in detail below.

Concerning the different θ_o to estimate we observe for $\theta_o = 0.0$ and 1.0 what we expected, a bias towards the interior of the interval $[-0.5, 1.5]$. Note that we did also simulation studies for $\theta_o = 0.25$ and 0.75 that are not presented here. The quality of the results for these two values was similar to that we obtain for $\theta_o = 0.5$.

Let us mention also that, as already indicated before, the PL method is much more expensive to calculate than the MD.

Next we would like to see the actual convergence rate of the estimates $\widehat{\theta}$ in practice. There are three possibilities: it converges slower than the parametric $\sqrt{n}$-rate due to the necessity of first estimating the nonparametric (additive) model, it converges faster as for rather small samples there is a bias due to the nonparametric pre-step that vanishes for increasing n, or it converges more or less at rate $\sqrt{n}$. In a simulation study with bandwidth $h_\alpha = h_0 n^{-1/5}$, $\alpha = 1, 2$ we applied our methods with the marginal integration smoother and $g_\alpha = 2h_\alpha$, $\alpha = 1, 2$ on model1. In Figure 1 we give the mean squared error multiplied by $\sqrt{n}$ for data generating $\theta = 0.0$, 0.5, and 1.0 calculated from 100 simulation runs. The data generating model was model1, the bandwidth used was $h = 0.5n^{-1/5}$.

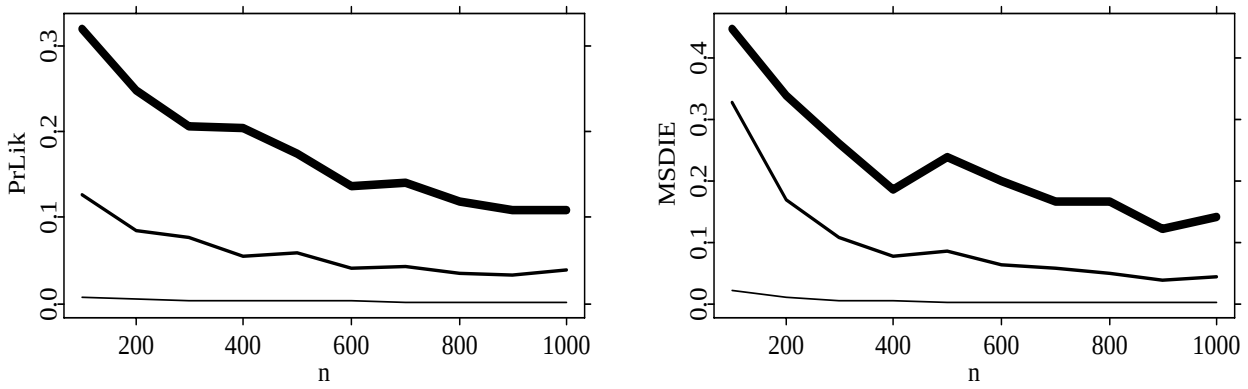


Figure 1: *Mean Squared Error times $\sqrt{n}$ as a function of sample size. Left: PL, right: MD. Data generating process was model1 with: thick line: $\theta = 1.0$, middle sized line: $\theta = 0.5$, thin line: $\theta = 0.0$. Numbers are calculated from 100 replicates for each n. Additive function estimator was MI with $h_0 = 0.5$.*

As we can see clearly, for both methods and all θ the estimate converges faster to θ than $\sqrt{n}$, i.e.

the parametric rate. This actually is not surprising as it is expected that for such small data sets higher order terms still matter a lot, in particular in the bias. So the reduction in the mean squared error is, at least at the beginning, faster than the asymptotic $\sqrt{n}$-rate. In both methods $\theta = 0.0$ seems to be estimated best. For $n = 1000$ the mean squared error is pretty close to zero.

We next focus on the question what happens when we change the smoother, in concrete what happens if we take Smooth Backfitting (BF) instead of Marginal Integration (MI). For this let us have a look at Table 2 below.

Both Methods when using BF

	$\theta_o =$	.00	.50	1.0	.00	.50	1.0	.00	.50	1.0
	$h_0 =$		0.3			0.4			0.5	
MD	model1	.0178	.5267	.9204	.0233	.5324	.9219	.0262	.5576	.9239
		.1079	.4044	.5519	.1154	.4220	.5758	.1214	.4437	.5815
		.0120	.1643	.3109	.0139	.1791	.3377	.0154	.2002	.3439
	model2	.0315	.5735	.9401	.0337	.5831	.9374	.0432	.5950	.9425
		.1497	.4354	.5556	.1563	.4594	.5692	.1591	.4698	.5839
		.0234	.1950	.3123	.0256	.2180	.3279	.0272	.2297	.3443
	model3	.0534	.5969	.9593	.0686	.6118	.9579	.0820	.6307	.9718
		.2257	.4738	.5352	.2358	.4906	.5721	.2421	.5014	.5839
		.0538	.2339	.2881	.0603	.2532	.3291	.0653	.2684	.3417
	$h_0 =$		0.2			0.3			0.4	
PL	model1	-.004	.4297	.8335	-.006	.4257	.8336	-.002	.4318	.8261
		.0711	.2763	.4369	.0755	.2870	.4655	.0758	.3055	.4916
		.0051	.0813	.2186	.0057	.0879	.2444	.0057	.0980	.2719
	model2	-.001	.4490	.8710	-.004	.4395	.8470	-.003	.4505	.8423
		.0951	.3057	.4610	.1027	.3151	.4725	.1044	.3345	.4974
		.0090	.0961	.2292	.0106	.1030	.2467	.0109	.1143	.2723
	model3	.0045	.4622	.8732	.0037	.4519	.8562	.0028	.4510	.8562
		.1516	.3398	.4590	.1570	.3612	.4842	.1566	.3570	.4929
		.0230	.1169	.2268	.0247	.1328	.2551	.0245	.1299	.2636

Table 2: *Performance of MD and PL method with BF estimator: Means (first line), standard deviations (second line), and means squared error (third line) of the $\widehat{\theta}$ for different θ_o, models [see (18)], and bandwidths $h_\alpha = h_0 n^{-1/5}$, $\alpha = 1, 2$, for sample size $n = 100$. All numbers calculated from 500 replications.*

The findings stated above (when we used MI) still hold. But, backfitting clearly does uniformly

better than MI as our (however, asymptotic) theory indicates for the MD. Nevertheless it is also worth to mention that the implementation is much simpler for MI and computationally much less expensive. This is because for each possible θ the backfitting algorithm has to be performed completely in new, since it is iterative. In contrast, the MI works with weighting matrices that only depend on the design so that for any θ all we have to do is a simple matrix multiplication to get the estimates $\widehat{m}_\alpha$.

Note further that the selected results refer to bandwidths that are slightly different. This is because marginal integration needs larger bandwidths than smooth backfitting to get final estimates of similar smoothness, see Sperlich, Linton and Härdle (1999). We see here a clear dependence between the θ and h_{opt} (the optimal bandwidth to estimate θ). So again, obviously it is worth to think somewhat more about the crucial question of bandwidth choice.

6.2 Bandwidth Discussion and Bootstrap

One might think that the probably easiest approach would be to apply plug-in bandwidths for the particular problem under consideration. For many of the regression problems and their corresponding estimators that might be true, and in particular for additive model estimators considered here, those plug-in rules can be found in the literature. However, they rely on asymptotic expressions with unknown functions and parameters that in our particular case are even more complicated to estimate. Further, in simulations (see Sperlich, Linton and Härdle, 1999, or Mammen and Park, 2005) they turned out not to work satisfactory.

Another, also quite natural approach would be to apply cross validation, i.e. the jackknife or the generalized version. For the generalized cross validation one needs to estimate the degrees of freedom of the nonparametric estimator, something that is even in the simple nonparametric regression a quite crucial point, this does not become better for our problem. On the other hand, for the smooth backfitting (see Nielsen and Sperlich, 2005) the jackknife method has been implemented successfully, and Kim et al. (1999) discussed a version for the internalized MI, i.e. the version applied here. Therefore we implemented the jackknife cv-bandwidth for BF and MI also for our context. However, for the MI we set all bandwidths of nuisance directions g_α to $g_\alpha = 2h_\alpha$. In Table 3 we give the results for minimizing the MD over $\theta \in \Theta$ choosing $h \in R^d$ by cross validation as described in Nielsen and Sperlich (2005), respectively in Kim et al (1999). Notice that we allow for different bandwidths in each direction. The simulations are executed as before but only for model1 and based on just 100 simulation runs what is enough to see the following: the results presented indicate that this method seems to work for any θ. As for this exercise we did no further investigation on what happens for n running from 100 to 1000 we have added the results for the case $n = 200$. It might surprise that the constant for "optimal" cv - bandwidths does not only change with θ but even more with n (not shown in table). Have in mind that in small samples the second order terms of bias and variance are still quite influential and thus the rate $n^{-1/5}$ is to be taken carefully; compare with the above convergence-rate study.

MD with cv-bandwidth

		BF			MI		
n	θ_o	mean$(\widehat{\theta})$	std$(\widehat{\theta})$	mse	mean$(\widehat{\theta})$	std$(\widehat{\theta})$	mse
	0.0	.0069	.1369	.0188	.0032	.1619	.0262
100	0.5	.5039	.5319	.2829	.4776	.6251	.3913
	1.0	.8345	.6087	.3979	.7176	.7211	.5998
	0.0	.0194	.0647	.0046	.0245	.0876	.0083
200	0.5	.5509	.2892	.0862	.5490	.3685	.1382
	1.0	1.017	.3695	.1368	.9689	.5104	.2615

Table 3: *Simulation results when applying BF and MI with cross validation bandwidth to minimize (9) w.r.t. θ. Numbers are calculated from 100 replications.*

A disadvantage of this cross validation procedure is that it is computationally rather expensive, and often rather hard to implement in practice as well. This gets even worse if one wants to combine the cross validation with the PL method. Therefore we additionally suggest a procedure that is relatively easy to implement, quite fast, and works reasonably well even for small data sets. The idea is to choose θ and the bandwidth simultaneously minimizing, respectively maximizing, the considered criteria function (7), respectively (9). Intuitively, this approach seems rather appealing to us as the interpretation of the results is easier when keeping the same criteria function to minimize / maximize.

MD & PL for θ and h_0 using MI estimator

		MD			PL		
n	θ_o	mean$(\widehat{\theta})$	std$(\widehat{\theta})$	mse	mean$(\widehat{\theta})$	std$(\widehat{\theta})$	mse
	0.0	.0032	.1587	.0252	-.016	.0960	.0095
100	0.5	.5108	.6020	.3625	.3589	.3954	.1763
	1.0	.7245	.7224	.5978	.6720	.5704	.4330
	0.0	.0238	.0858	.0079	-.001	.0568	.0032
200	0.5	.5596	.3463	.1235	.4183	.2711	.0802
	1.0	1.008	.4971	.2472	.8302	.4374	.2202

Table 4: *Simulation results when minimizing (9) / maximizing (7) simultaneously w.r.t. θ and the bandwidth. Numbers are calculated from 100 replications.*

In Table 4 we give the results for minimizing the MD over $\theta \in \Theta$ and h simultaneously. For computational ease, we did this only for the MI smoother. The simulations are the same as above, model1 with only 100 simulation runs but again for $n = 100$ and $n = 200$. To simplify the simulations we chose $h_1 = h_2$, $g_1 = g_2 = 2h_1$ as we did at the beginning of this section, see discussion above.

The results presented in Table 4 indicate that this method seems to work very well, too. Certainly it would be very hard to do any theory proving that the obtained bandwidths will converge to the optimal ones. For both, this method and cross validation, it is also rather tedious to derive the asymptotic properties for $\widehat{\theta}$ since then the bandwidth is random.

As discussed in the sections above, often the asymptotic expressions given in our theorems are little helpful in practice due to various reasons: they contain various unknown functions and parameters, and usually large sample sizes are needed before second order terms become really negligible. Therefore we have suggested a bootstrap procedure to estimate the distribution of $\widehat{\theta}$ in small samples whose usefulness and performance we want to investigate now. For the sake of shortness we restrict again to model1, applying the MI smoother. For this model we already know from above that the optimal bandwidth constant h_0 is about 0.5.

Bootstrap estimates of standard deviation and bias for θ

Method	n	θ_o	std$(\widehat{\theta})$	$\widehat{std}(\widehat{\theta})$	$std(\widehat{std}(\widehat{\theta}))$	$bias(\widehat{\theta})$	$\widehat{bias}(\widehat{\theta})$	$std(\widehat{bias}(\widehat{\theta}))$
		0.0	.1453	.2394	.0736	.0394	.0357	.0865
	100	0.5	.5611	.5585	.0693	.1210	.0520	.2134
MD		1.0	.6604	.6179	.0751	-.111	-.072	.2138
		0.0	.0810	.1097	.0287	.0129	.0074	.0578
	200	0.5	.3598	.4045	.0723	.0224	.0428	.2217
		1.0	.4972	.4996	.1019	-.057	-.035	.2588
		0.0	.1453	.1658	.0364	.0394	.0154	.1116
recentered	100	0.5	.5611	.4597	.0514	.1210	.0341	.3208
		1.0	.6604	.5577	.0621	-.111	-.082	.2682
		0.0	.0856	.1044	.0148	.0004	-.002	.0416
	100	0.5	.3501	.4048	.0497	-.063	.0155	.1619
PL		1.0	.5342	.5583	.0595	-.185	-.032	.2073
		0.0	.0505	.0716	.0099	.0066	-.004	.0314
	200	0.5	.2319	.2992	.0400	-.036	-.013	.1271
		1.0	.3862	.4532	.0559	-.109	-.055	.1782

Table 5: *Approximation of the distribution of $\widehat{\theta}$ by bootstrap for both methods, using MI. Here, the values for $\widehat{std}$ and $\widehat{bias}$ are averages over* 200 *simulations with* 250 *bootstrap samples. Numbers are calculated from* 200 *simulation runs.*

For our simulation study we did only 250 bootstrap replicates. In Table 5 we give the results calculated from 200 replications. We see clearly the bootstrap is doing reasonable in estimating the standard deviation but, as usually, doing less well for the bias of $\widehat{\theta}$. As one might expect, the bootstrap gives some conservative results for the standard deviation, always (slightly) overestimating

the standard deviation. So it can be recommended for statistical inference on the model.

7 Appendix A : Profile likelihood estimator

To prove the asymptotic normality of the profile likelihood estimator, we will use Theorems 1 and 2 of Chen, Linton and Van Keilegom (2003) (abbreviated by CLV in the sequel). Therefore, we need to define the space to which the nuisance function $s = (m, r, f, g, h)$ belongs. We define this space by $\mathcal{H}_{PL} = \mathcal{M}^2 \times C_1^1(\mathbb{R})^3$, where $C_a^b(R)$ ($0 < a < \infty$, $0 < b \leq 1$, $R \subset \mathbb{R}^k$ for some k) is the set of all continuous functions $f : R \to \mathbb{R}$ for which

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

and where the space $\mathcal{M}$ depends on the model at hand. For instance, when the model is additive, a good choice for $\mathcal{M}$ is $\mathcal{M} = \sum_{\alpha=1}^d C_1^1(R_{X_\alpha})$, and when the model is multiplicative $\mathcal{M} = \prod_{\alpha=1}^d C_1^1(R_{X_\alpha})$. We also need to define, according to CLV, a norm for the space $\mathcal{H}_{PL}$. Let

$$\|s\|_{PL} = \sup_{\theta \in \Theta} \max\{\|m_\theta\|_\infty, \|r_\theta\|_\infty, \|f_\theta\|_2, \|g_\theta\|_2, \|h_\theta\|_2\},$$

where $\|\cdot\|_\infty$ ($\|\cdot\|_2$) denotes the L_∞ (L_2) norm. Finally, let's denote $\|\cdot\|$ for the Euclidean norm.

We assume that the estimator $\widehat{m}_\theta$ is constructed based on a kernel function of degree q_1, which we assume of the form $K_1(u_1) \times \ldots \times K_1(u_d)$, and a bandwidth h. The required conditions on K_1, q_1 and h are mentioned in the list of regularity conditions given below.

7.1 Assumptions

We assume throughout this appendix that the conditions stated below are satisfied. Condition A.1-A.7 are regularity conditions on the kernels, bandwidths, distributions F_X, F_ε, etc., whereas condition A.8 contains primitive conditions on the estimator $\widehat{m}_\theta$, that need to be checked depending on which model structure and which estimator $\widehat{m}_\theta$ one has chosen.

A.1 The probability density function K_j ($j = 1, 2$) is symmetric and has compact support, $\int u^k K_j(u)\, du = 0$ for $k = 1, \ldots, q_j - 1$, $\int u^{q_j} K_j(u)\, du \neq 0$ and K_j is twice continuously differentiable.

A.2 $nh \to \infty$, $nh^{2q_1} \to 0$, $ng^6(\log g^{-1})^{-2} \to \infty$ and $ng^{2q_2} \to 0$, where q_1 and q_2 are defined in condition A.1 and $q_1, q_2 \geq 4$.

A.3 The density f_X is bounded away from zero and infinity and is Lipschitz continuous on the compact support $\mathcal{X}$.

A.4 The functions $m_\theta(x)$ and $\dot{m}_\theta(x)$ are q_1 times continuously differentiable with respect to the components of x on $\mathcal{X} \times \mathcal{N}(\theta_o)$, and all derivatives up to order q_1 are bounded, uniformly in (x, θ) in $\mathcal{X} \times \mathcal{N}(\theta_o)$.

A.5 The transformation $\Lambda_\theta(y)$ is three times continuously differentiable in both θ and y, and there exists a $\delta > 0$ such that

$$E\left[\sup_{\|\theta'-\theta\|\leq\delta}\left|\frac{\partial^{k+l}}{\partial y^k \partial\theta^l}\Lambda_{\theta'}(Y)\right|\right] < \infty$$

for all θ in Θ and all $0 \leq k + l \leq 3$.

A.6 The distribution $F_{\varepsilon(\theta)}(y)$ is three times continuously differentiable with respect to y and θ, and

$$\sup_{\theta,y}\left|\frac{\partial^{k+l}}{\partial y^k \partial\theta^l}F_{\varepsilon(\theta)}(y)\right| < \infty$$

for all $0 \leq k + l \leq 2$.

A.7 For all $\eta > 0$, there exists $\epsilon(\eta) > 0$ such that

$$\inf_{\|\theta-\theta_o\|>\eta}\|G_{PL}(\theta, s_\theta)\| \geq \epsilon(\eta) > 0.$$

Moreover, the matrix Γ_{1PL} is of full (column) rank.

A.8 The estimators $\widehat{m}_o$ and $\dot{\widehat{m}}_o$ can be written as

$$\widehat{m}_o(x) - m_o(x) = \frac{1}{nh}\sum_{i=1}^{n}\sum_{\alpha=1}^{d} K_1\Big(\frac{x_\alpha - X_{\alpha i}}{h}\Big)v_{o1\alpha}(X_{\alpha i}, \varepsilon_i) + \frac{1}{n}\sum_{i=1}^{n} v_{o2}(\varepsilon_i) + \widehat{v}_o(x),$$

and

$$\dot{\widehat{m}}_o(x) - \dot{m}_o(x) = \frac{1}{nh}\sum_{i=1}^{n}\sum_{\alpha=1}^{d} K_1\Big(\frac{x_\alpha - X_{\alpha i}}{h}\Big)w_{o1\alpha}(X_{\alpha i}, \varepsilon_i) + \frac{1}{n}\sum_{i=1}^{n} w_{o2}(\varepsilon_i) + \widehat{w}_o(x),$$

where $\sup_x |\widehat{v}_o(x)| = o_p(n^{-1/2})$, $\sup_x |\widehat{w}_o(x)| = o_p(n^{-1/2})$, the functions $v_{o1\alpha}(x, e)$ and $w_{o1\alpha}(x, e)$ are q_1 times continuously differentiable with respect to the components of x, their derivatives up to order q_1 are bounded, uniformly in x and e, $E(v_{o2}(\varepsilon)) = 0$ and $E(w_{o2}(\varepsilon)) = 0$. Moreover, with probability tending to 1, $\widehat{m}_\theta, \dot{\widehat{m}}_\theta \in \mathcal{M}$, $\sup_{\theta\in\Theta}\|\widehat{m}_\theta - m_\theta\| = o_p(1)$, $\sup_{\theta\in\Theta}\|\dot{\widehat{m}}_\theta - \dot{m}_\theta\| = o_p(1)$, $\|\widehat{m}_\theta - m_\theta\| = o_p(n^{-1/4})$ and $\|\dot{\widehat{m}}_\theta - \dot{m}_\theta\| = o_p(n^{-1/4})$ uniformly over all θ with $\|\theta - \theta_o\| = o(1)$, and

$$\sup_x |(\dot{\widehat{m}}_\theta - \dot{m}_\theta)(x) - (\dot{\widehat{m}}_o - \dot{m}_o)(x)| = o_p(1)\|\theta - \theta_o\| + O_p(n^{-1/2})$$

for all θ with $\|\theta - \theta_o\| = o(1)$. Finally, the space $\mathcal{M}$ satisfies $\int \sqrt{\log N(\varepsilon, \mathcal{M}, \|\cdot\|_\infty)}\, d\varepsilon < \infty$.

7.2 Proof of Theorem 1

The proof consists in verifying the conditions given in Theorem 1 (regarding consistency) and 2 (regarding asymptotic normality) in CLV. In Lemmas A1-A8 below, we verify these conditions. The result then follows immediately from those lemmas, assuming that the primitive conditions on $\widehat{m}_\theta$ and the regularity conditions stated in A.1-A.8 hold true.

LEMMA A1. *Uniformly for all $\theta \in \Theta$, $G_{PL}(\theta, s)$ is continuous in s at $s = s_\theta$ in the uniform norm.*

LEMMA A2.

$$\sup_y \sup_{\theta\in\Theta} |\widehat{f}_{\varepsilon(\theta)}(y) - f_{\varepsilon(\theta)}(y)| = o_p(1), \qquad \sup_y \sup_{\theta\in\Theta} |\dot{\widehat{f}}_{\varepsilon(\theta)}(y) - \dot{f}_{\varepsilon(\theta)}(y)| = o_p(1),$$

and

$$\sup_y \sup_{\theta\in\Theta} |\widehat{f}'_{\varepsilon(\theta)}(y) - f'_{\varepsilon(\theta)}(y)| = o_p(1).$$

LEMMA A3. *For all sequences of positive numbers $\delta_n = o(1)$,*

$$\sup_{\theta\in\Theta, \|s-s_\theta\|_{PL}\le\delta_n} \|G_{nPL}(\theta, s) - G_{PL}(\theta, s)\| = o_p(1).$$

LEMMA A4. *The ordinary partial derivative in θ of $G_{PL}(\theta, s_\theta)$, denoted $\Gamma_{1PL}(\theta, s_\theta)$, exists in a neighborhood of θ_o, is continuous at $\theta = \theta_o$, and the matrix $\Gamma_{1PL} = \Gamma_{1PL}(\theta_o, s_o)$ is of full (column) rank.*

For any $\theta \in \Theta$, we say that $G_{PL}(\theta, s)$ is pathwise differentiable at s in the direction $[\overline{s} - s]$ if $\{s + \tau(\overline{s} - s) : \tau \in [0, 1]\} \subset \mathcal{H}_{PL}$ and $\lim_{\tau\to 0}[G_{PL}(\theta, s + \tau(\overline{s} - s)) - G_{PL}(\theta, s)]/\tau$ exists; we denote the limit by $\Gamma_{2PL}(\theta, s)[\overline{s} - s]$.

LEMMA A5. *The pathwise derivative $\Gamma_{2PL}(\theta, s_\theta)$ of $G_{PL}(\theta, s_\theta)$ exists in all directions $s - s_\theta$ and satisfies*:

$$(i) \quad \|G_{PL}(\theta, s) - G_{PL}(\theta, s_\theta) - \Gamma_{2PL}(\theta, s_\theta)[s - s_\theta]\| \le c\|s - s_\theta\|^2_{PL}$$

for all θ with $\|\theta - \theta_o\| = o(1)$, all s with $\|s - s_\theta\|_{PL} = o(1)$, some constant $c < \infty$;

$$(ii) \quad \|\Gamma_{2PL}(\theta, s_\theta)[\widehat{s}_\theta - s_\theta] - \Gamma_{2PL}(\theta_o, s_o)[\widehat{s}_o - s_o]\| \le c\|\theta - \theta_o\| \times o_p(1) + O_p(n^{-1/2})$$

for all θ with $\|\theta - \theta_o\| = o(1)$, where $\widehat{s} = (\widehat{m}, \dot{\widehat{m}}, \widehat{f}_\varepsilon, \dot{\widehat{f}}_\varepsilon, \widehat{f}'_\varepsilon)$.

LEMMA A6. *With probability tending to one,* $\widehat{f}_{\varepsilon}, \dot{\widehat{f}}_{\varepsilon}, \widehat{f}'_{\varepsilon} \in C_1^1(I\!R)$. Moreover,

$$\sup_y \sup_{\|\theta-\theta_o\|\le\delta_n} |\widehat{f}_{\varepsilon(\theta)}(y) - f_{\varepsilon(\theta)}(y)| = o_p(n^{-1/4}),$$
$$\sup_y \sup_{\|\theta-\theta_o\|\le\delta_n} |\dot{\widehat{f}}_{\varepsilon(\theta)}(y) - \dot{f}_{\varepsilon(\theta)}(y)| = o_p(n^{-1/4}),$$

and

$$\sup_y \sup_{\|\theta-\theta_o\|\le\delta_n} |\widehat{f}'_{\varepsilon(\theta)}(y) - f'_{\varepsilon(\theta)}(y)| = o_p(n^{-1/4}),$$

for any $\delta_n = o(1)$.

LEMMA A7. *For all sequences of positive numbers* $\{\delta_n\}$ *with* $\delta_n = o(1)$,

$$\sup_{\|\theta-\theta_o\|\le\delta_n, \|s-s_\theta\|_{PL}\le\delta_n} \|G_{nPL}(\theta,s) - G_{PL}(\theta,s) - G_{nPL}(\theta_o,s_o)\| = o_p(n^{-1/2}).$$

LEMMA A8.

$$\sqrt{n}\{G_{nPL}(\theta_o,s_o) + \Gamma_{2PL}(\theta_o,s_o)[\widehat{s}-s_o]\} \Longrightarrow N(0, \mathrm{Var}\{G_{1PL}(\theta_o,s_o)\}).$$

7.3 Proofs of Lemmas A1-A8

Before proving Lemmas A1-A8, we first need to consider some preliminary results concerning the estimator $\widehat{f}_{\varepsilon(\theta)}$ and its derivatives.

The first result states that the asymptotic behavior of the estimator $\widehat{f}_{\varepsilon(\theta)}(y)$, which is a kernel estimator based on the estimated residuals $\widehat{\varepsilon}_{i\theta} = \Lambda_\theta(Y_i) - \widehat{m}_\theta(X_i)$, is the same as that of the kernel estimator based on the (unobserved) true errors $\varepsilon_{i\theta} = \Lambda_\theta(Y_i) - m_\theta(X_i)$.

LEMMA A9. *For all* $y \in I\!R$,

$$\widehat{f}_\varepsilon(y) - f_\varepsilon(y) = n^{-1}\sum_{i=1}^n K_{2g}(\varepsilon_i - y) - f_\varepsilon(y) + f'_\varepsilon(y)\, n^{-1}\sum_{i=1}^n \Big[\sum_{\alpha=1}^d v_{o1\alpha}(X_{\alpha i},\varepsilon_i) f_{X_\alpha}(X_{\alpha i}) + v_{o2}(\varepsilon_i)\Big] + \widehat{r}_o(y),$$

where $\sup_y |\widehat{r}_o(y)| = o_p(n^{-1/2})$, *and where the functions* $v_{o1\alpha}$ *and* v_{o2} *are defined in assumption A.8. Moreover,*

$$\sup_y \sup_{\theta\in\Theta} |\widehat{f}_{\varepsilon(\theta)}(y) - f_{\varepsilon(\theta)}(y)| = o_p(1)$$

and

$$\sup_y \sup_{\|\theta-\theta_o\|\le\delta_n} |\widehat{f}_{\varepsilon(\theta)}(y) - f_{\varepsilon(\theta)}(y)| = o_p(n^{-1/4})$$

for all $\delta_n = o(1)$.

Proof. Write

$$\begin{aligned}
&\widehat{f}_\varepsilon(y) - f_\varepsilon(y) \\
&= \frac{1}{ng}\sum_{i=1}^n K'_{2g}(\varepsilon_i - y)(\widehat{\varepsilon}_i - \varepsilon_i) + \frac{1}{n}\sum_{i=1}^n K_{2g}(\varepsilon_i - y) - f_\varepsilon(y) + o_p(n^{-1/2}) \\
&= -\frac{1}{ng}\sum_{i=1}^n K'_{2g}(\varepsilon_i - y)\left\{\frac{1}{n}\sum_{k=1}^n\sum_{\alpha=1}^d K_{1h}(X_{\alpha i} - X_{\alpha k}) v_{o1\alpha}(X_{\alpha k}, \varepsilon_k) + \frac{1}{n}\sum_{k=1}^n v_{o2}(\varepsilon_k) + \widehat{v}_o(X_i)\right\} \\
&\quad + \frac{1}{n}\sum_{i=1}^n K_{2g}(\varepsilon_i - y) - f_\varepsilon(y) + o_p(n^{-1/2}) \\
&= \frac{1}{n^2}\sum_{\alpha=1}^d\sum_{i,k=1}^n v_{o1\alpha}(X_{\alpha k}, \varepsilon_k)\varphi_{nik} + f'_\varepsilon(y)\frac{1}{n}\sum_{k=1}^n v_{o2}(\varepsilon_k) \\
&\quad + \frac{1}{n}\sum_{i=1}^n K_{2g}(\varepsilon_i - y) - f_\varepsilon(y) + o_p(n^{-1/2}), \qquad (19)
\end{aligned}$$

where $\varphi_{nik} = -\frac{1}{g}K'_{2g}(\varepsilon_i - y)K_{1h}(X_{\alpha i} - X_{\alpha k})$. Since

$$E(\varphi_{nik}|X_k) = f'_\varepsilon(y) f_{X_\alpha}(X_{\alpha k}) + o_p(1),$$

it follows that (19) equals

$$\begin{aligned}
&f'_\varepsilon(y)\frac{1}{n}\sum_{k=1}^n\Big[\sum_{\alpha=1}^d v_{o1\alpha}(X_{\alpha k}, \varepsilon_k) f_{X_\alpha}(X_{\alpha k}) + v_{o2}(\varepsilon_k)\Big] \\
&+ \frac{1}{n}\sum_{i=1}^n K_{2g}(\varepsilon_i - y) - f_\varepsilon(y) + o_p(n^{-1/2}).
\end{aligned}$$

In a similar way as for Lemma A9, we can prove the following results. The proofs are omitted.

LEMMA A10. *For all* $y \in I\!R$,

$$\begin{aligned}
\dot{\widehat{f}}_\varepsilon(y) - \dot{f}_\varepsilon(y) &= (ng)^{-1}\sum_{i=1}^n K'_{2g}(\varepsilon_i - y)(\dot{\Lambda}_\theta(Y_i) - \dot{m}_\theta(X_i)) - \dot{f}_\varepsilon(y) \\
&\quad + \dot{f}'_\varepsilon(y)\, n^{-1}\sum_{i=1}^n\Big[\sum_{\alpha=1}^d v_{o1\alpha}(X_{\alpha i}, \varepsilon_i) f_{X_\alpha}(X_{\alpha i}) + v_{o2}(\varepsilon_i)\Big] \\
&\quad + f'_\varepsilon(y)\, n^{-1}\sum_{i=1}^n\Big[\sum_{\alpha=1}^d w_{o1\alpha}(X_{\alpha i}, \varepsilon_i) f_{X_\alpha}(X_{\alpha i}) + w_{o2}(\varepsilon_i)\Big] + \widehat{r}_o(y),
\end{aligned}$$

where $\sup_{\theta,y}|\widehat{r}_o(y)| = o_p(n^{-1/2})$. *Moreover,*

$$\sup_y \sup_{\theta\in\Theta} |\dot{\widehat{f}}_{\varepsilon(\theta)}(y) - \dot{f}_{\varepsilon(\theta)}(y)| = o_p(1)$$

and

$$\sup_y \sup_{\|\theta-\theta_o\|\le\delta_n} |\dot{\widehat{f}}_{\varepsilon(\theta)}(y) - \dot{f}_{\varepsilon(\theta)}(y)| = o_p(n^{-1/4})$$

for all $\delta_n = o(1)$.

LEMMA A11. *For all* $y \in I\!R$,

$$\widehat{f}'_\varepsilon(y) - f'_\varepsilon(y) = (ng)^{-1}\sum_{i=1}^n K'_{2g}(y-\varepsilon_i) - f'_\varepsilon(y) + f''_\varepsilon(y)\, n^{-1}\sum_{i=1}^n \Big[\sum_{\alpha=1}^d v_{o1\alpha}(X_{\alpha i},\varepsilon_i) f_{X_\alpha}(X_{\alpha i}) + v_{o2}(\varepsilon_i)\Big] + \widehat{r}_o(y),$$

where $\sup_y|\widehat{r}_o(y)| = o_p(n^{-1/2})$. *Moreover,*

$$\sup_y \sup_{\theta\in\Theta} |\widehat{f}'_{\varepsilon(\theta)}(y) - f'_{\varepsilon(\theta)}(y)| = o_p(1)$$

and

$$\sup_y \sup_{\|\theta-\theta_o\|\le\delta_n} |\widehat{f}'_{\varepsilon(\theta)}(y) - f'_{\varepsilon(\theta)}(y)| = o_p(n^{-1/4})$$

for all $\delta_n = o(1)$.

PROOF OF LEMMA A1. Note that

$$G_{PL}(\theta,s) = E\left[\frac{1}{f(\varepsilon(\theta,m))}\{g(\varepsilon(\theta,m))(\dot{\Lambda}_\theta(Y) - r(X)) + h(\varepsilon(\theta,m))\} + \frac{\dot{\Lambda}'_\theta(Y)}{\Lambda'_\theta(Y)}\right],$$

which is continuous in s at $s = s_\theta$ provided conditions A.4-A.6 are satisfied.

PROOF OF LEMMA A2. This follows from Lemmas A9-A11.

PROOF OF LEMMA A3. The proof is similar (but easier) than that of Lemma A7. We therefore omit the proof.

PROOF OF LEMMA A4. This follows from assumption A.7.

PROOF OF LEMMA A5. Some straightforward calculations show that

$$
\begin{aligned}
&\Gamma_{2PL}(\theta, s_\theta)[\widehat{s}_\theta - s_\theta] \qquad (20)\\
&= \lim_{\tau\to 0} \frac{1}{\tau}\{G_{PL}(\theta, s_\theta + \tau(\widehat{s}_\theta - s_\theta)) - G_{PL}(\theta, s_\theta)\}\\
&= E\Big[\Big\{\frac{f'_{\varepsilon(\theta)}(\varepsilon_\theta)}{f^2_{\varepsilon(\theta)}(\varepsilon_\theta)}(\widehat{m}_\theta - m_\theta)(X) - \frac{(\widehat{f}_{\varepsilon(\theta)} - f_{\varepsilon(\theta)})(\varepsilon_\theta)}{f^2_{\varepsilon(\theta)}(\varepsilon_\theta)}\Big\}\Big\{f'_{\varepsilon(\theta)}(\varepsilon_\theta)[\dot{\Lambda}_\theta(Y) - \dot{m}_\theta(X)] + \dot{f}_{\varepsilon(\theta)}(\varepsilon_\theta)\Big\}\\
&\quad + \frac{1}{f_{\varepsilon(\theta)}(\varepsilon_\theta)}\Big\{ - f''_{\varepsilon(\theta)}(\varepsilon_\theta)[\dot{\Lambda}_\theta(Y) - \dot{m}_\theta(X)](\widehat{m}_\theta - m_\theta)(X) + (\widehat{f}'_{\varepsilon(\theta)} - f'_{\varepsilon(\theta)})(\varepsilon_\theta)[\dot{\Lambda}_\theta(Y) - \dot{m}_\theta(X)]\\
&\quad - f'_{\varepsilon(\theta)}(\varepsilon_\theta)(\dot{\widehat{m}}_\theta - \dot{m}_\theta)(X) + (\dot{\widehat{f}}_{\varepsilon(\theta)} - \dot{f}_{\varepsilon(\theta)})(\varepsilon_\theta) - \dot{f}'_{\varepsilon(\theta)}(\varepsilon_\theta)(\widehat{m}_\theta - m_\theta)(X)\Big\}\Big].
\end{aligned}
$$

The first part of Lemma A5 now follows immediately. The second part follows from the uniform consistency of $\widehat{m}$, $\dot{\widehat{m}}$, $\widehat{f}_{\varepsilon(\theta)}$, $\dot{\widehat{f}}_{\varepsilon(\theta)}$ and $\widehat{f}'_{\varepsilon(\theta)}$, and from the fact that

$$
\sup_x |(\dot{\widehat{m}}_\theta - \dot{m}_\theta)(x) - (\dot{\widehat{m}}_o - \dot{m}_o)(x)| = o_p(1)\|\theta - \theta_o\| + O_p(n^{-1/2}),
$$

which follows from assumption A.8.

PROOF OF LEMMA A6. This follows from Lemmas A9-A11.

PROOF OF LEMMA A7. We will make use of Theorem 3 in Chen, Linton and Van Keilegom (2003). According to this result we need to prove that
(i)

$$
E\Big[\sup_{\|\theta'-\theta\|<\eta, \|s'-s\|_{PL}<\eta} |g_{PL}(X, Y, \theta', s') - g_{PL}(X, Y, \theta, s)|^2\Big] \leq K\eta^2,
$$

for all $(\theta, s) \in \Theta \times \mathcal{H}_{PL}$, all $\eta > 0$ and for some $K > 0$.
(ii)

$$
\int_0^\infty \sqrt{\log N(\varepsilon, \mathcal{H}_{PL}, \|\cdot\|_{PL})}d\varepsilon < \infty,
$$

where $N(\varepsilon, \mathcal{H}_{PL}, \|\cdot\|_{PL})$ is the covering number with respect to the norm $\|\cdot\|_{PL}$ of the class $\mathcal{H}_{PL}$, i.e. the minimal number of balls of $\|\cdot\|_{PL}$-radius ε needed to cover $\mathcal{H}_{PL}$.
Part (ii) follows from Corollary 2.7.4 in van der Vaart and Wellner (1996), together with assumption A.8. Part (i) follows from the mean value theorem, together with the differentiability conditions imposed on the functions of which the function g_{PL} is composed.

PROOF OF LEMMA A8. Combining the formula of $\Gamma_{2PL}(\theta_o, s_o)$ given in (20) with the represen-

tations of $\widehat{f}_{\varepsilon(\theta)}$, $\dot{\widehat{f}}_{\varepsilon(\theta)}$ and $\widehat{f}'_{\varepsilon(\theta)}$ given in Lemmas A9-A11, we obtain after some calculations:

$$
\begin{aligned}
&G_{nPL}(\theta_o, s_o) + \Gamma_{2PL}(\theta_o, s_o)[\widehat{s} - s_o] \qquad (21)\\
&= n^{-1} \sum_i \left\{ \frac{1}{f_\varepsilon(\varepsilon_i)} [f'_\varepsilon(\varepsilon_i)\{\dot{\Lambda}_o(Y_i) - \dot{m}_o(X_i)\} + \dot{f}_\varepsilon(\varepsilon_i)] + \frac{\dot{\Lambda}'_o(Y_i)}{\Lambda'_o(Y_i)} \right\} \\
&\quad + E\left[-\frac{1}{f^2_\varepsilon(\varepsilon)} \left\{ \frac{1}{ng} \sum_i K_2\Big(\frac{\varepsilon_i - \varepsilon}{g}\Big) - f_\varepsilon(\varepsilon) \right\} \{f'_\varepsilon(\varepsilon)[\dot{\Lambda}_o(Y) - \dot{m}_o(X)] + \dot{f}_\varepsilon(\varepsilon)\} \right. \\
&\quad + \frac{1}{f_\varepsilon(\varepsilon)} \left\{ -\frac{1}{ng^2} \sum_i K'_2\Big(\frac{\varepsilon_i - \varepsilon}{g}\Big) - f'_\varepsilon(\varepsilon) \right\} \{\dot{\Lambda}_o(Y) - \dot{m}_o(X)\} \\
&\quad \left. + \frac{1}{f_\varepsilon(\varepsilon)} \left\{ \frac{1}{ng^2} \sum_i K'_2\Big(\frac{\varepsilon_i - \varepsilon}{g}\Big)(\dot{\Lambda}_o(Y_i) - \dot{m}_o(X_i)) - \dot{f}_\varepsilon(\varepsilon) \right\} \right] + o_p(n^{-1/2}).
\end{aligned}
$$

We next show that

$$
E\left[\frac{\dot{f}_\varepsilon(\varepsilon)}{f_\varepsilon(\varepsilon)} \right] = 0, \qquad (22)
$$

$$
E\left[\frac{1}{f_\varepsilon(\varepsilon)} \left\{ \frac{1}{ng^2} \sum_i K'_2\Big(\frac{\varepsilon_i - \varepsilon}{g}\Big)(\dot{\Lambda}_o(Y_i) - \dot{m}_o(X_i)) - \dot{f}_\varepsilon(\varepsilon) \right\} \right] = 0, \qquad (23)
$$

and

$$
\begin{aligned}
&E\left[-\frac{1}{f^2_\varepsilon(\varepsilon)} \left\{ \frac{1}{ng} \sum_i K_2\Big(\frac{\varepsilon_i - \varepsilon}{g}\Big) \right\} \{f'_\varepsilon(\varepsilon)[\dot{\Lambda}_o(Y) - \dot{m}_o(X)] + \dot{f}_\varepsilon(\varepsilon)\} \right. \\
&\qquad \left. + \frac{1}{f_\varepsilon(\varepsilon)} \left\{ -\frac{1}{ng^2} \sum_i K'_2\Big(\frac{\varepsilon_i - \varepsilon}{g}\Big) \right\} \{\dot{\Lambda}_o(Y) - \dot{m}_o(X)\} \right] = 0. \qquad (24)
\end{aligned}
$$

It then follows that only the first term on the right hand side of (21) (i.e. the term $G_{nPL}(\theta_o, s_o)$) is non-zero, from which the result follows. We start by showing (22) :

$$
E\left[\frac{\dot{f}_\varepsilon(\varepsilon)}{f_\varepsilon(\varepsilon)} \right] = \int \dot{f}_\varepsilon(y)\, dy = \frac{\partial}{\partial \theta} \int f_{\varepsilon(\theta)}(y)\, dy \Big|_{\theta=\theta_o} = 0,
$$

since $\int f_{\varepsilon(\theta)}(y)\, dy = 1$. Next, consider (23). The left hand side equals

$$
\begin{aligned}
&\frac{1}{ng^2} \sum_i (\dot{\Lambda}_o(Y_i) - \dot{m}_o(X_i)) E\left[\frac{1}{f_\varepsilon(\varepsilon)} K'_2\Big(\frac{\varepsilon_i - \varepsilon}{g}\Big) \right] - E\left[\frac{\dot{f}_\varepsilon(\varepsilon)}{f_\varepsilon(\varepsilon)} \right] \\
&= \frac{1}{ng} \sum_i (\dot{\Lambda}_o(Y_i) - \dot{m}_o(X_i)) \int K'_2(u)\, du = 0.
\end{aligned}
$$

Finally, for (24), note that the left hand side can be written as

$$\frac{1}{ng}\sum_i E\left[\frac{1}{f_\varepsilon^2(\varepsilon)}\Big\{-K_2\Big(\frac{\varepsilon_i-\varepsilon}{g}\Big)\frac{d}{d\theta}\, f_{\varepsilon(\theta)}(\varepsilon(\theta))\big\downarrow_{\theta=\theta_o}+\frac{d}{d\theta}\, K_2\Big(\frac{\varepsilon_i-\varepsilon(\theta)}{g}\Big)\Big|_{\downarrow\theta=\theta_o} f_\varepsilon(\varepsilon)\Big\}\right]$$
$$=\frac{1}{ng}\sum_i E\left[\frac{d}{d\theta}\frac{K_2\Big(\frac{\varepsilon_i-\varepsilon(\theta)}{g}\Big)}{f_{\varepsilon(\theta)}(\varepsilon(\theta))}\Bigg|_{\downarrow\theta=\theta_o}\right]=\frac{1}{ng}\sum_i\frac{d}{d\theta}\int K_2\Big(\frac{\varepsilon_i-e}{g}\Big)\,de=0,$$

since $\int K_2\Big(\frac{\varepsilon_i-e}{g}\Big)\,de=g$. This finishes the proof.

8 Appendix B : MD estimator

8.1 Assumptions

We assume throughout this appendix that assumptions B.1–B.8 given below are valid.

B.1 The probability density function K_1 is symmetric and has compact support, $\int u^k K_1(u)\,du=0$ for $k=1,\ldots,q_1-1$, $\int u^{q_1}K_1(u)\,du\neq 0$ and K_1 is twice continuously differentiable.

B.2 $nh\to\infty$ and $nh^{2q_1}\to 0$, where q_1 is defined in condition B.1 and $q_1\geq 4$.

B.3 The density f_X is bounded away from zero and infinity and is Lipschitz continuous on the compact support $\mathcal{X}$.

B.4 The function $m_\theta(x)$ is q_1 times continuously differentiable with respect to the components of x on $\mathcal{X}\times\mathcal{N}(\theta_o)$, and all derivatives up to order q_1 are bounded, uniformly in (x,θ) in $\mathcal{X}\times\mathcal{N}(\theta_o)$.

B.5 The transformation $\Lambda_\theta(y)$ is twice continuously differentiable in both θ and y, and there exists a $\delta>0$ such that

$$E\left[\sup_{\|\theta-\theta'\|\leq\delta}|\lambda_{\theta'}(Y)|^k\right]<\infty$$

for all k and for all θ in Θ.

B.6 The distribution $F_\varepsilon(y)$ is twice continuously differentiable with respect to y, and $\sup_y|f_\varepsilon'(y)|<\infty$.

B.7 For all $\eta>0$, there exists $\epsilon(\eta)>0$ such that

$$\inf_{\|\theta-\theta_o\|>\eta}\|G_{MD}(\theta,m_\theta)\|_2\geq\epsilon(\eta)>0.$$

Moreover, the matrix $\Gamma_{1MD}(x,e)$ (defined in Section 5) is of full (column) rank for a set of positive μ-measure (x,e).

B.8 The estimator $\widehat{m}_o$ can be written as

$$\widehat{m}_o(x) - m_o(x) = \frac{1}{nh}\sum_{i=1}^{n}\sum_{\alpha=1}^{d} K_1\Big(\frac{x_\alpha - X_{\alpha i}}{h}\Big) v_{o1\alpha}(X_{\alpha i}, \varepsilon_i) + \frac{1}{n}\sum_{i=1}^{n} v_{o2}(\varepsilon_i) + \widehat{v}_o(x),$$

where $\sup_x |\widehat{v}_o(x)| = o_p(n^{-1/2})$, the function $v_{o1\alpha}(x,e)$ is q_1 times continuously differentiable with respect to the components of x, their derivatives up to order q_1 are bounded, uniformly in x and e, $E(v_{o2}(\varepsilon)) = 0$. Moreover, with probability tending to 1, $\widehat{m}_\theta \in \mathcal{M}$, $\sup_{\theta\in\Theta} \|\widehat{m}_\theta - m_\theta\| = o_p(1)$, $\|\widehat{m}_\theta - m_\theta\| = o_p(n^{-1/4})$ uniformly over all θ with $\|\theta - \theta_o\| = o(1)$, and

$$\sup_x |(\widehat{m}_\theta - m_\theta)(x) - (\widehat{m}_o - m_o)(x)| = o_p(1)\|\theta - \theta_o\| + O_p(n^{-1/2})$$

for all θ with $\|\theta - \theta_o\| = o(1)$. Finally, the space $\mathcal{M}$ satisfies $\int \sqrt{\log N(\varepsilon, \mathcal{M}, \|\cdot\|_\infty)}\, d\varepsilon < \infty$.

8.2 Proof of Theorem 2

We use a generalization of Theorems 1 (about consistency) and 2 (about asymptotic normality) of Chen, Linton and Van Keilegom (2003), henceforth CLV. Below, we state the primitive conditions under which these results are valid (see Lemmas B1–B6). Their proof is given in Section 8.3.

Given these lemmas, we have the desired result. We just reprieve the last part of the argument because it is slightly different from CLV due to the different norm. Note that

$$\begin{aligned} F_{\varepsilon(\theta,m)}(e) &= \Pr\left[\Lambda_\theta(Y) - m(X) \le e\right] \\ &= \Pr\left[Y \le \Lambda_\theta^{-1}(m(X) + e)\right] \\ &= \Pr\left[\varepsilon \le \Lambda_o(\Lambda_\theta^{-1}(m(X) + e)) - m_o(X)\right] \\ &= E F_\varepsilon[\Lambda_o(\Lambda_\theta^{-1}(m(X) + e)) - m_o(X)]. \end{aligned}$$

Likewise, $F_{X,\varepsilon(\theta,m)}$ satisfies

$$\begin{aligned} F_{X,\varepsilon(\theta,m)}(x,e) &= \Pr\left[X \le x, \Lambda_\theta(Y) - m(X) \le e\right] \\ &= E \Pr\left[X \le x, \varepsilon \le \Lambda_o(\Lambda_\theta^{-1}(m(X) + e)) - m_o(X)\right] \\ &= E\left[1(X \le x) F_\varepsilon[\Lambda_o(\Lambda_\theta^{-1}(m(X) + e)) - m_o(X)]\right]. \end{aligned}$$

Define

$$G_{MD}(\theta, m)(x,e) = F_{X,\varepsilon(\theta,m)}(x,e) - F_X(x) F_{\varepsilon(\theta,m)}(e).$$

Define now the stochastic processes

$$L_n(x,e) = \sqrt{n}[\widehat{F}_{X,\varepsilon}(x,e) - F_{X,\varepsilon}(x,e)] - F_X(x)\sqrt{n}[\widehat{F}_\varepsilon(e) - F_\varepsilon(e)] - F_\varepsilon(e)\sqrt{n}[\widehat{F}_X(x) - F_X(x)]$$

and

$$\mathcal{L}_n(\theta)(x,e) = L_n(x,e) + \Gamma_{1MD}(x,e)(\theta - \theta_o) + [\Gamma_{2MD}(\theta_o, m_o)(\widehat{m} - m_o)](x,e),$$

where for any $\theta \in \Theta$ and any $m, \overline{m} \in \mathcal{M}$, $\Gamma_{2MD}(\theta, m)(\overline{m} - m)(x, e)$ is defined in the following way. We say that $G_{MD}(\theta, m)$ is pathwise differentiable at m in the direction $[\overline{m} - m]$ at (x, e) if $\{m + \tau(\overline{m} - m) : \tau \in [0, 1]\} \subset \mathcal{M}$ and $\lim_{\tau \to 0}[G_{MD}(\theta, m + \tau(\overline{m} - m))(x, e) - G_{MD}(\theta, m)(x, e)]/\tau$ exists; we denote the limit by $\Gamma_{2MD}(\theta, m)[\overline{m} - m](x, e)$.

A consequence of Lemmas B1–B6 is that

$$\sup_{\|\theta - \theta_o\| \leq \delta_n} \|G_{nMD}(\theta, \widehat{m}_\theta) - \mathcal{L}_n(\theta)\|_2^2 = o_p(n^{-1/2}),$$

which means we can effectively deal with the minimizer of $\mathcal{L}_n(\theta)$, say $\overline{\theta}$. Note that $\overline{\theta}$ has an explicit solution and indeed

$$\begin{aligned}\sqrt{n}(\overline{\theta} - \theta_o) = &- \left[\int \Gamma_{1MD}\Gamma_{1MD}^{\top}(x, e)d\mu(x, e)\right]^{-1} \\ &\times \int [L_n(x, e) + [\Gamma_{2MD}(\theta_o, m_o)(\widehat{m} - m_o)](x, e)]\Gamma_{1MD}(x, e)d\mu(x, e).\end{aligned}$$

Then apply Lemma B6 below to get the desired result.

LEMMA B1. *Uniformly for all $\theta \in \Theta$, $G_{MD}(\theta, m)$ is continuous in m at $m = m_\theta$ in the uniform norm.*

LEMMA B2. *For all sequences of positive numbers $\delta_n = o(1)$,*

$$\sup_{\theta \in \Theta, \|m - m_\theta\|_{\mathcal{M}} \leq \delta_n} \|G_{nMD}(\theta, m) - G_{MD}(\theta, m)\|_2 = o_p(1).$$

LEMMA B3. *For all (x, e), the ordinary partial derivative in θ of $G_{MD}(\theta, m_\theta)(x, e)$, denoted $\Gamma_{1MD}(\theta, m_\theta)(x, e)$, exists in a neighborhood of θ_o, is continuous at $\theta = \theta_o$, and the matrix $\Gamma_{1MD}(x, e) = \Gamma_{1MD}(\theta_o, m_o)(x, e)$ is of full (column) rank for a set of positive μ-measure (x, e).*

LEMMA B4. *For μ-all (x, e), the pathwise derivative $\Gamma_{2MD}(\theta, m_\theta)(x, e)$ of $G_{MD}(\theta, m_\theta)(x, e)$ exists in all directions $m - m_\theta$ and satisfies*:

$$(i) \quad \|G_{MD}(\theta, m) - G_{MD}(\theta, m_\theta) - \Gamma_{2MD}(\theta, m_\theta)[m - m_\theta]\|_2 \leq c\|m - m_\theta\|_{\mathcal{M}}^2$$

for all θ with $\|\theta - \theta_o\| = o(1)$, all m with $\|m - m_\theta\|_{\mathcal{M}} = o(1)$, some constant $c < \infty$;

$$(ii) \quad \|\Gamma_{2MD}(\theta, m_\theta)[\widehat{m}_\theta - m_\theta] - \Gamma_{2MD}(\theta_o, m_o)[\widehat{m} - m_o]\|_2 \leq c\|\theta - \theta_o\| \times o_p(1) + O_p(n^{-1/2})$$

for all θ with $\|\theta - \theta_o\| = o(1)$.

LEMMA B5. *For all sequences of positive numbers* $\{\delta_n\}$ *with* $\delta_n = o(1)$,

$$\sup_{\|\theta-\theta_o\|\leq\delta_n, \|m-m_\theta\|_{\mathcal{M}}\leq\delta_n} \|G_{nMD}(\theta,m) - G_{MD}(\theta,m) - G_{nMD}(\theta_o,m_o)\|_2 = o_p(n^{-1/2}).$$

LEMMA B6.

$$\sqrt{n}\int\{G_{nMD}(\theta_o,m_o) + \Gamma_{2MD}(\theta_o,m_o)[\widehat{m}-m_o]\}(x,e)\Gamma_{1MD}(x,e)d\mu(x,e) \Longrightarrow N(0,V_{1MD}).$$

8.3 Proofs of Lemmas B1–B6

PROOF OF LEMMA B1. This follows from the representation

$$G_{MD}(\theta,m_\theta)(x,e) = E\left[[1(X\leq x) - F_X(x)]F_\varepsilon[\Lambda_o(\Lambda_\theta^{-1}(m_\theta(X)+e)) - m_o(X)]\right], \tag{25}$$

and the smoothness of F_ε, Λ_o, and Λ_θ^{-1}.

PROOF OF LEMMA B2. Define the linearization :

$$G^L_{nMD}(\theta,m)(x,e) = \widehat{F}_{X,\varepsilon(\theta,m)}(x,e) - F_X(x)\widehat{F}_{\varepsilon(\theta,m)}(e) - \widehat{F}_X(x)F_{\varepsilon(\theta,m)}(e) + F_X(x)F_{\varepsilon(\theta,m)}(e).$$

By the triangle inequality we have

$$\begin{aligned}
&\sup_{\theta\in\Theta,\|m-m_\theta\|_{\mathcal{M}}\leq\delta_n} \|G_{nMD}(\theta,m) - G_{MD}(\theta,m)\|_2 \\
&\leq \sup_{\theta\in\Theta,\|m-m_\theta\|_{\mathcal{M}}\leq\delta_n} \|G^L_{nMD}(\theta,m) - G_{MD}(\theta,m)\|_2 + \sup_{\theta\in\Theta,\|m-m_\theta\|_{\mathcal{M}}\leq\delta_n} \|G_{nMD}(\theta,m) - G^L_{nMD}(\theta,m)\|_2.
\end{aligned}$$

We must show that both terms on the right hand side are $o_p(1)$. Define the stochastic processes

$$\tau_{n\varepsilon}(\theta,m,e) = \widehat{F}_{\varepsilon(\theta,m)}(e) - F_{\varepsilon(\theta,m)}(e) \text{ and } \tau_{nX\varepsilon}(\theta,m,x,e) = \widehat{F}_{X,\varepsilon(\theta,m)}(x,e) - F_{X,\varepsilon(\theta,m)}(x,e)$$

for each $\theta\in\Theta$, $m\in\mathcal{M}$, $x\in\mathbb{R}^k, e\in\mathbb{R}$. We claim that

$$\sup_{\theta\in\Theta,\|m-m_\theta\|_{\mathcal{M}}\leq\delta_n, e\in\mathbb{R}} |\tau_{n\varepsilon}(\theta,m,e)| = o_p(1) \tag{26}$$

$$\sup_{\theta\in\Theta,\|m-m_\theta\|_{\mathcal{M}}\leq\delta_n, x\in\mathbb{R}^k, e\in\mathbb{R}} |\tau_{nX\varepsilon}(\theta,m,x,e)| = o_p(1), \tag{27}$$

which implies that

$$\begin{aligned}
&\sup_{\theta\in\Theta,\|m-m_\theta\|_{\mathcal{M}}\leq\delta_n} \|G^L_{nMD}(\theta,m) - G^L_{MD}(\theta,m)\|_2 \\
&= \sup_{\theta\in\Theta,\|m-m_\theta\|_{\mathcal{M}}\leq\delta_n} \|(\widehat{F}_{X,\varepsilon(\theta,m)} - F_{X,\varepsilon(\theta,m)}) - F_X(\widehat{F}_{\varepsilon(\theta,m)} - F_{\varepsilon(\theta,m)}) - F_{\varepsilon(\theta,m)}(\widehat{F}_X - F_X)\|_2 \\
&\leq \Bigg[\sup_{\theta\in\Theta,\|m-m_\theta\|_{\mathcal{M}}\leq\delta_n, e\in\mathbb{R}} |\tau_{nX\varepsilon}(\theta,m,e)| + \sup_{\theta\in\Theta,\|m-m_\theta\|_{\mathcal{M}}\leq\delta_n, x\in\mathbb{R}^k, e\in\mathbb{R}} |\tau_{n\varepsilon}(\theta,m,x,e)| \\
&+ \sup_{x\in\mathbb{R}^k} |\widehat{F}_X(x) - F_X(x)|\Bigg] \\
&= o_p(1).
\end{aligned}$$

Similarly, $\sup_{\theta\in\Theta,\|m-m_\theta\|_{\mathcal{M}}\leq\delta_n}\|G_{nMD}(\theta,m)-G^L_{nMD}(\theta,m)\|_2=o_p(1)$. The proof of (26) and (27) is based on Theorem 3 in CLV. We omit the details because it is similar to our proof of Lemma B5.

PROOF OF LEMMA B3. Below, we calculate $\Gamma_{1MD}(x,e)=\Gamma_{1MD}(\theta_o,m_o)(x,e)$. In a similar way $\Gamma_{1MD}(\theta,m_\theta)(x,e)$ can be obtained. First, we have

$$\begin{aligned}
&\frac{\partial}{\partial\theta}F_{\varepsilon(\theta,m_\theta)}(e)\Big|_{\downarrow\theta=\theta_o}\\
&=E\,\frac{\partial}{\partial\theta}F_\varepsilon\left[\Lambda_o(\Lambda_\theta^{-1}(m_\theta(X)+e))-m_o(X)\right]\Big|_{\downarrow\theta=\theta_o}\\
&=f_\varepsilon(e)E\frac{\partial}{\partial\theta}\Lambda_o(\Lambda_\theta^{-1}(m_\theta(X)+e))\Big|_{\downarrow\theta=\theta_o}\\
&=f_\varepsilon(e)E\Lambda_o'(\Lambda_o^{-1}(m_o(X)+e))\,\frac{\partial}{\partial\theta}(\Lambda_\theta^{-1}(m_\theta(X)+e))\Big|_{\downarrow\theta=\theta_o}\\
&=f_\varepsilon(e)E\Lambda_o'(\Lambda_o^{-1}(m_o(X)+e))\left[\frac{\lambda_o(\Lambda_o^{-1}(m_o(X)+e))}{\Lambda_o'(\Lambda_o^{-1}(m_o(X)+e))}+\frac{1}{\Lambda_o'(\Lambda_o^{-1}(m_o(X)+e))}\dot m_o(X)\right]\\
&=f_\varepsilon(e)E\left[\lambda_o(\Lambda_o^{-1}(m_o(X)+e))+\dot m_o(X)\right]
\end{aligned}$$

by the chain rule. Similarly,

$$\frac{\partial}{\partial\theta}F_{X,\varepsilon(\theta,m_\theta)}(x,e)\Big|_{\downarrow\theta=\theta_o}=f_\varepsilon(e)E\left[1\left(X\leq x\right)\left\{\lambda_o(\Lambda_o^{-1}(m_o(X)+e))+\dot m_o(X)\right\}\right].$$

Therefore,

$$\begin{aligned}
\Gamma_{1MD}(x,e)=\Gamma_{1MD}(\theta_o,m_o)(x,e)&=\frac{\partial G_{MD}(\theta,m_\theta)}{\partial\theta}(x,e)\Big|_{\downarrow\theta=\theta_o}\\
&=\frac{\partial}{\partial\theta}F_{X,\varepsilon(\theta,m_\theta)}(x,e)-F_X(x)\frac{\partial}{\partial\theta}F_{\varepsilon(\theta,m_\theta)}(e)\\
&=f_\varepsilon(e)E\left[(1(X\leq x)-F_X(x))\left(\lambda_o(\Lambda_o^{-1}(m_o(X)+e))+\dot m_o(X)\right)\right].
\end{aligned}\tag{28}$$

PROOF OF LEMMA B4. By the law of iterated expectation and partial differentiation we obtain that

$$\begin{aligned}
&[\Gamma_{2MD}(\theta_o,m_o)(m-m_o)](x,e)\\
&=\frac{\partial G_{MD}(\theta_o,m_o+t(m-m_o))}{\partial t}(x,e)\Big|_{\downarrow t=0}\\
&=f_\varepsilon(e)E\left[(1(X\leq x)-F_X(x))\,(m(X)-m_o(X))\right].
\end{aligned}$$

Similarly, the formula of $[\Gamma_{2MD}(\theta, m_\theta)(m - m_\theta)](x, e)$ is given by

$$\begin{aligned}
&[\Gamma_{2MD}(\theta, m_\theta)(m - m_\theta)](x, e) \\
&= \lim_{\tau \to 0} \frac{1}{\tau} E\left[\{1(X \le x) - F_X(x)\} f_\varepsilon[\Lambda_o\{\Lambda_\theta^{-1}(m_\theta(X) + e)\} - m_o(X)]\right. \\
&\quad \left. \times[\Lambda_o\{\Lambda_\theta^{-1}(m_\theta(X) + \tau(m - m_\theta)(X) + e)\} - \Lambda_o\{\Lambda_\theta^{-1}(m_\theta(X) + e)\}]\right].
\end{aligned}$$

The two inequalities in the statement of Lemma B4 now follow easily, using the consistency of $\widehat{m}_\theta$ and the fact that $\sup_x |(\widehat{m}_\theta - m_\theta)(x) - (\widehat{m}_o - m_o)(x)| = o_p(1)\|\theta - \theta_o\| + O_p(n^{-1/2})$.

Proof of Lemma B5. Define the stochastic processes

$$\nu_{n\varepsilon}(\theta, m, e) = \sqrt{n}[\widehat{F}_{\varepsilon(\theta,m)}(e) - F_{\varepsilon(\theta,m)}(e)] \text{ and } \nu_{nX\varepsilon}(\theta, m, x, e) = \sqrt{n}[\widehat{F}_{X,\ \varepsilon(\theta,m)}(x, e) - F_{X,\varepsilon(\theta,m)}(x, e)]$$

for each $\theta : \|\theta - \theta_o\| \le \delta_n$ and $m : \|m - m_\theta\|_{\mathcal{M}} \le \delta_n$, $x \in \mathbb{R}^k, e \in \mathbb{R}$. We claim that

$$\sup_{\|\theta-\theta_o\| \le \delta_n, \|m-m_\theta\|_{\mathcal{M}} \le \delta_n, e \in \mathbb{R}} |\nu_{n\varepsilon}(\theta, m, e)| = o_p(1) \tag{29}$$

$$\sup_{\|\theta-\theta_o\| \le \delta_n, \|m-m_\theta\|_{\mathcal{M}} \le \delta_n, x \in \mathbb{R}^d, e \in \mathbb{R}} |\nu_{nX\varepsilon}(\theta, m, x, e)| = o_p(1). \tag{30}$$

The proof of these results are based on Theorem 3 in CLV. We have to show that their condition (3.2) is satisfied, which requires in our case [with $g(Z, \theta, m) = 1(\varepsilon(\theta, m) \le e) - E1(\varepsilon(\theta, m) \le e)$ and $g(Z, \theta, m) = 1(X \le x)1(\varepsilon(\theta, m) \le e) - E1(X \le x)1(\varepsilon(\theta, m) \le e)$] that

$$\left(E\left[\sup_{(\theta', m'): \|\theta' - \theta\| < \delta, \|m' - m\|_{\mathcal{M}} < \delta} |g(Z, \theta', m') - g(Z, \theta, m)|^r\right]\right)^{1/r} \le K\delta^s$$

for all $(\theta, m) \in \Theta \times \mathcal{M}$, all small positive value $\delta = o(1)$, and for some constants $s \in (0, 1]$, $K > 0$, and that the bound holds for μ-almost all (x, e). We have

$$\begin{aligned}
|g(Z, \theta', m') - g(Z, \theta, m)| \le\ & |1(\varepsilon(\theta, m) \le e) - 1(\varepsilon(\theta', m') \le e)| \\
& + |E1(\varepsilon(\theta, m) \le e) - E1(\varepsilon(\theta', m') \le e)|,
\end{aligned}$$

and

$$\begin{aligned}
|1(\varepsilon(\theta, m) \le e) - 1(\varepsilon(\theta', m') \le e)| &= |1(\Lambda_\theta(Y) - m(X) \le e) - 1(\Lambda_{\theta'}(Y) - m'(X) \le e)| \\
&\le |1(\Lambda_\theta(Y) - m(X) \le e) - 1(\Lambda_\theta(Y) - m'(X) \le e)| \\
&\quad + |1(\Lambda_\theta(Y) - m'(X) \le e) - 1(\Lambda_{\theta'}(Y) - m'(X) \le e)|.
\end{aligned}$$

For all $m' \in \mathcal{M}$ with $\|m' - m\|_{\mathcal{M}} \le \delta \le 1$, we have for all Y, X, e:

$$\begin{aligned}
&\sup_{\|m'-m\|_{\mathcal{M}} \le \delta} |1(m'(X) \ge \Lambda_\theta(Y) - e) - 1(m(X) \ge \Lambda_\theta(Y) - e)| \\
&\qquad \le 1(m(X) + \delta \ge \Lambda_\theta(Y) - e) - 1(m(X) - \delta \ge \Lambda_\theta(Y) - e).
\end{aligned}$$

The preceding term is either one or zero and its expectation is the probability that $m(X) + \delta \geq \Lambda_\theta(Y) - e \geq m(X) - \delta$, which is the probability that $e + \delta \geq \Lambda_\theta(Y) - m(X) \geq e - \delta$, which is

$$\begin{aligned} F_{\varepsilon(\theta,m)}(e+\delta) - F_{\varepsilon(\theta,m)}(e-\delta) &= EF_\varepsilon[\Lambda_o(\Lambda_\theta^{-1}(m(X)+e+\delta)) - m_o(X)] \\ &\quad - EF_\varepsilon[\Lambda_o(\Lambda_\theta^{-1}(m(X)+e-\delta)) - m_o(X)]. \end{aligned}$$

We then apply the smoothness conditions on F_ε, Λ_o, and Λ_θ^{-1} to bound the right hand side by $K\delta$ for small enough δ and constant $K < \infty$.

Next, by the Mean Value Theorem, we have

$$\Lambda_\theta(Y) - \Lambda_{\theta'}(Y) = \lambda_{\theta^*}(Y) \times (\theta - \theta'),$$

where θ^* is an intermediate value between θ and θ'. For all $\alpha > 0$, by the Bonferroni and Markov inequalities,

$$\begin{aligned} &\Pr\left[\max_{1\leq i\leq n} \sup_{\|\theta-\theta'\|\leq\delta} |\lambda_{\theta'}(Y_i)| > c \times n^\alpha\right] \\ &\leq n \times \Pr\left[\sup_{\|\theta-\theta'\|\leq\delta} |\lambda_{\theta'}(Y)| > c \times n^\alpha\right] \\ &\leq n \times \frac{E\left[\sup_{\|\theta-\theta'\|\leq\delta} |\lambda_{\theta'}(Y)|^k\right]}{c^k n^{k\alpha}} = o(1), \end{aligned}$$

provided $k > \alpha^{-1}$.

Therefore, we can safely assume that there is some upper bound c such that $\sup_{\|\theta-\theta'\|\leq\delta} |\Lambda_\theta(Y) - \Lambda_{\theta'}(Y)| \leq c \times \delta$. Therefore, on this set

$$\begin{aligned} &\sup_{\|\theta'-\theta\|\leq\delta} |1(\Lambda_\theta(Y) - m'(X) \leq e) - 1(\Lambda_{\theta'}(Y) - m'(X) \leq e)| \\ &\leq 1(\Lambda_\theta(Y) + c\delta - m'(X) \leq e) - 1(\Lambda_\theta(Y) - c\delta - m'(X) \leq e)|, \end{aligned}$$

which has probability bounded by $K\delta$ for some $K > 0$.

Therefore, condition (3.2) of Theorem 3 in CLV is satisfied with $r = 2$ and $s = 1/2$, and condition (3.3) of Theorem 3 is satisfied by the condition on the bracketing number of the class $\mathcal{M}$, stated in assumption B.8.

Proof of Lemma B6. We show below that

$$\begin{aligned} &[\Gamma_{2MD}(\theta_o, m_o)(\widehat{m} - m_o)](x, e) \\ &= f_\varepsilon(e)\sqrt{n} \int \left[(1(X \leq x) - F_X(x))\left(\widehat{m}(X) - m_o(X)\right)\right] f_X(X) dX \\ &= f_\varepsilon(e)\frac{1}{\sqrt{n}} \sum_{i=1}^n (1(X_i \leq x) - F_X(x)) f_X(X_i) \sum_{\alpha=1}^d v_{o1\alpha}(X_{\alpha i}, \varepsilon_i) + o_p(1). \end{aligned} \tag{31}$$

Therefore,

$$[L_n(x,e) + [\Gamma_{2MD}(\theta_o, m_o)(\widehat{m} - m_o)](x,e)] = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} U_i(x,e) + o_p(1),$$

where

$$U_i(x,e) = [1(X_i \leq x)1(\varepsilon_i \leq e) - F_{X,\varepsilon}(x,e)] - F_X(x)[1(\varepsilon_i \leq e) - F_\varepsilon(e)] - F_\varepsilon(e)[1(X_i \leq x) - F_X(x)]$$
$$+ f_X(X_i) \sum_{\alpha=1}^{d} v_{o1\alpha}(X_{\alpha i}, \varepsilon_i) f_\varepsilon(e) \left(1(X_i \leq x) - F_X(x)\right),$$

and where $E[U_i(x,e)] = 0$ for all x, e. Because $F_{X,\varepsilon}(x,e) = F_X(x)F_\varepsilon(e)$ we have

$$U_i(x,e) = [1(X_i \leq x) - F_X(x)][1(\varepsilon_i \leq e) - F_\varepsilon(e)] + f_X(X_i) \sum_{\alpha=1}^{d} v_{o1\alpha}(X_{\alpha i}, \varepsilon_i) f_\varepsilon(e) \left(1(X_i \leq x) - F_X(x)\right).$$

Now integrating $U_i(x,e)$ with respect to $\Gamma_{1MD}(x,e)d\mu(x,e)$ gives the answer.

Proof of (31). Write

$$\widehat{m}(X) - m_o(X) = \frac{1}{nh} \sum_{i=1}^{n} \sum_{\alpha=1}^{d} K_1\Big(\frac{X_\alpha - X_{\alpha i}}{h}\Big) v_{o1\alpha}(X_{\alpha i}, \varepsilon_i) + \frac{1}{n} \sum_{i=1}^{n} v_{o2}(\varepsilon_i) + o_p(n^{-1/2}).$$

Then, provided $nh^{2q_1} \to 0$,

$$\sqrt{n} \int \left[\left(1(X \leq x) - F_X(x)\right)\left(\widehat{m}(X) - m_o(X)\right)\right] f_X(X) dX$$
$$= \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \sum_{\alpha=1}^{d} v_{o1\alpha}(X_{\alpha i}, \varepsilon_i) \int \left[\left(1(X \leq x) - F_X(x)\right) \frac{1}{h} K_1\left(\frac{X_\alpha - X_{\alpha i}}{h}\right)\right] f_X(X) dX$$
$$+ \frac{1}{\sqrt{n}} \sum_{i=1}^{n} v_{o2}(\varepsilon_i) \int \left[\left(1(X \leq x) - F_X(x)\right)\right] f_X(X) dX + o_p(1)$$
$$= \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \sum_{\alpha=1}^{d} v_{o1\alpha}(X_{\alpha i}, \varepsilon_i) \int \left[\left(1(X_i + uh \leq x) - F_X(x)\right) K_1(u_\alpha)\right] f_X(X_i + uh) du + o_p(1)$$
$$= \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \sum_{\alpha=1}^{d} v_{o1\alpha}(X_{\alpha i}, \varepsilon_i) \left(1(X_i \leq x) - F_X(x)\right) f_X(X_i) + o_p(1).$$

We also have to substitute $\frac{\partial m_\theta}{\partial \theta}(x)\big\downarrow_{\theta=\theta_o}$ into the formula for Γ_{1MD}.

References

Amemiya, T. and J.L. Powell, (1981), A comparison of the Box-Cox maximum likelihood estimator and the non-linear two-stage least squares estimator, *Journal of Econometrics*, **17**, 351-381.

Benkard, C.L. and S. Berry (2004), On the nonparametric identification of nonlinear simultaneous equation models: Comment on B. Brown (1983) and Roehrig (1988), *Preprint.*

Blundell, R. and J.-M. Robin, (2000), Latent separability: grouping goods without weak separability. *Econometrica*, **68**, 53-84.

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

Breiman, L. and J.H. Friedman, (1985), Estimating optimal transformations for multiple regression and correlation (with discussion), *Journal of the American Statistical Association,* **80**, 580-619.

Carroll, R.J. and D. Ruppert, (1984), Power transformation when fitting theoretical models to data, *Journal of the American Statistical Association,* **79**, 321-328.

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

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

Deaton, A. and J. Muellbauer, (1980), *Economics and Consumer Behavior.* Cambridge University Press, New York.

Ehrlich, I., (1977), Capital punishment and deterrence: some further thoughts and additional evidence, *Journal of Political Economy,* **85**, 741-788.

Hall, P. and J.L. Horowitz, (1996), Bootstrap critical values for tests based on generalized-method-of-moments estimators, *Econometrica*, **64**, 891-916.

Hastie, T.J. and R.J. Tibshirani, (1990), *Generalized Additive Models.* Chapman and Hall, London.

Heckman, J.J. and S. Polachek, (1974), Empirical evidence on the functional form of the earnings-schooling relationship, *Journal of the American Statistical Association*, **69**, 350-354.

Hengartner, N.W. and S. Sperlich, (2005), Rate optimal estimation with the integration method in the presence of many covariates, *Journal of Multivariate Analysis*, **95**, 246-272.

Hulten, C.R. and F.C. Wykoff, (1981), The estimation of economic depreciation using vintage asset prices: an application of the Box-Cox power transformation, *Journal of Econometrics*, **15**, 367-396.

Ibragimov, I.A. and R.Z. Hasminskii, (1980), On nonparametric estimation of regression, *Soviet Math. Dokl.,* **21**, 810-814.

Johnson, N.L., (1949), Systems of frequency curves generated by methods of translation, *Biometrika,* **36**, 149-176.

Kim, W., O.B. Linton and N. Hengartner, (1999), A computationally efficient oracle estimator of additive nonparametric regression with bootstrap confidence intervals, *The Journal of Computational and Graphical Statistics,* **8**, 278-297.

Koul, H. L., (2001), *Weighted Empirical Processes in Regression and Autoregression Models*, Springer-Verlag, New York.

Linton, O.B., R. Chen, N. Wang and W. Härdle, (1997), An analysis of transformations for additive nonparametric regression, *Journal of the American Statistical Association,* **92**, 1512-1521.

Linton, O. and E. Mammen (2005), Estimating semiparametric ARCH(∞) models by kernel smoothing, Econometrica 73, 771-836.

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

Mammen, E., O.B. Linton and J.P. Nielsen, (1999), The existence and asymptotic properties of a backfitting projection algorithm under weak conditions, *Annals of Statistics,* **27**, 1443-1490.

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

Newey, W.K., (1994), Kernel estimation of partial means, *Econometric Theory,* **10**, 233-253.

Nielsen, J.P., and O. Linton (1998). An optimization interpretation of integration and backfitting estimators for separable nonparametric models *Journal of The Royal Statistical Society, Series B* (1998), **60**, 217-22.

Nielsen, J.P. and S. Sperlich, (2005), Smooth backfitting in practice, *Journal of the Royal Statistical Society, Series B*, **61**, 43-61.

Robinson, P.M., (1991), Best nonlinear three-stage least squares estimation of certain econometric models, *Econometrica,* **59**, 755-786.

Rodríguez-Poó, J.M., S. Sperlich and P. Vieu, (2003), Semiparametric estimation of weak and strong separable models, *Econometric Theory*, **19**, 1008-1039.

Roehrig, C., (1988), Conditions for identification in nonparametric and parametric models, *Econometrica*, **56**, 433-447.

Severini, T.A., and W.H. Wong (1992), Profile likelihood and conditionally parametric models, *Annals of Statistics* 20, 1768-1802.

Sperlich, S., (2005), A note on nonparametric estimation with constructed variables and generated regressors, *Preprint, University Carlos III de Madrid, Spain.*

Sperlich, S., O.B. Linton and W. Härdle, (1999), Integration and backfitting methods in additive models: finite sample properties and comparison, *Test*, **8**, 419-458.

Stone, C.J., (1980), Optimal rates of convergence for nonparametric estimators, *Annals of Statistics,* **8**, 1348-1360.

Stone, C.J., (1982), Optimal global rates of convergence for nonparametric regression, *Annals of Statistics,* **8**, 1040-1053.

Stone, C.J., (1986), The dimensionality reduction principle for generalized additive models, *Annals of Statistics*, **14**, 592-606.

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

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

Van Keilegom, I. and N. Veraverbeke, (2002), Density and hazard estimation in censored regression models, *Bernoulli*, **8**, 607-625.

Zarembka, P., (1968), Functional form in the demand for money, *Journal of the American Statistical Association*, **63**, 502-511.

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