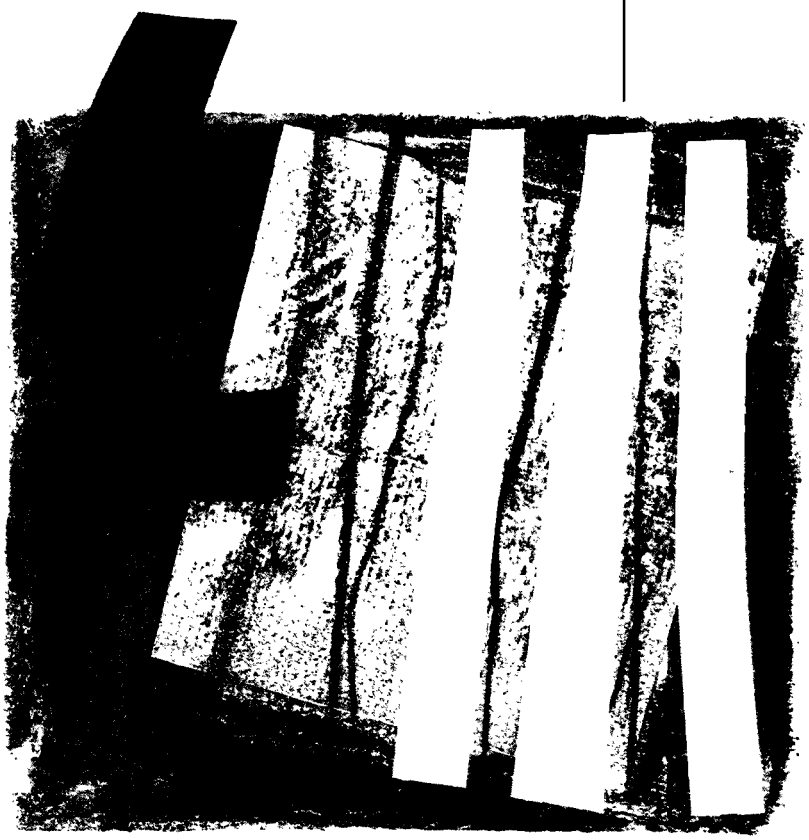
SEMIPARAMETRIC ESTIMATION OF WEAK AND STRONG SEPARABLE MODELS

Juan M. Rodríguez-Póo,
Stefan Sperlich and Philippe Vieu

00-69



WORKING PAPERS

Universidad Carlos III de Madrid

Working Paper 00-69
Statistics and Econometrics Series 32
October 2000

Departamento de Estadística y Econometría
Universidad Carlos III de Madrid
Calle Madrid, 126
28903 Getafe (Spain)
Fax (34) 91 624-98-49

SEMIPARAMETRIC ESTIMATION OF WEAK AND STRONG SEPARABLE MODELS

Juan M. Rodríguez-Póo, Stefan Sperlich and Philippe Vieu *

Abstract

In this paper we introduce a general method for estimating semiparametrically the different components in weak or strong separable models. The family of separable models is quite popular in economic theory and empirical research as this structure offers clear interpretation, has straight forward economic consequences and often is justified by theory. As will be seen in this article they are also of statistical interest since they allow to estimate semiparametrically high dimensional complexity without running in the so called curse of dimensionality. Generalized additive models and generalized partial linear models are special cases in this family of models. The idea of the new method is mainly based on a combination of local likelihood and efficient estimators in non or semiparametric models. Although this imposes some hypothesis on the error distribution this yields a very general usable method with little computational costs and high exactness even for small samples. E.g. it enables us to include models for censored and truncated variables which are quite common in quantitative economics. We give the estimation procedures and provide asymptotic theory for them. Implementation is discussed, simulations and an application demonstrate its feasibility in finite sample behavior.

Keywords: Separable models; Local likelihood; Nonparametric regression.

*Rodríguez-Póo, J.M., Departamento de Economía, Universidad de Cantabria; Sperlich, Departamento de Estadística y Econometría, Universidad Carlos III de Madrid, C/ Madrid 126 28903 Getafe, Madrid, Spain, e-mail: stefan@est-econ.uc3m.es; Vieu, Laboratoire de Statistique et Probabilités, Université Paul Sabatier. This research was financially supported by the Dirección General de Enseñanza Superior del Ministerio de Educación y Ciencia under research grants PB96-1469-C05-03 and PB98-0025, the Dirección General de Enseñanza Superior del Ministerio de Educación y Ciencia under the Subprograma de estancias de investigadores españoles en centros de investigación españoles y extranjeros, ref. PR2000-0096 and the Danish Social Science Research fund. We also thank M. Delgado and O. Linton for helpful discussion.

1 Introduction

Separability plays an extremely important rule in economics and econometrics. Already Leontief (1947a,b) introduced and discussed in detail definition, interpretation and consequences for different levels of separability (local, global, weak, strong). It is characterized by the independence of the marginal rate of substitution between a pair of inputs from changes in the level of another input, i.e. $\frac{\partial}{\partial x_k}\frac{G_i}{G_j} = 0$ or $G_j G_{ik} - G_i G_{jk} = 0$ with $G_i = \frac{\partial G}{\partial x_i}$, $G_{ik} = \frac{\partial^2 G}{\partial x_i \partial x_k}$, where $G : I\!R^d \to I\!R$ can be e.g. a production function and x_i, x_j, x_k inputs from (different) pairs, $i, j, k = 1, \ldots, d$. We speak of *weak* separability when x_i, x_j are from the same subset of inputs, but x_k from a different one. Strong separability is given when x_i, x_j can also to be from different subsets. The subsets are thought to be mutually exclusive and exhaustive. Imagine we have chosen r such subsets of inputs. Regarding the consequences for the functional form of G, well known results (see Goldman and Uzawa, 1964) are that weak separability is equivalent to $G(x) = F(\eta_1, \eta_2, \ldots, \eta_p)$, where η_s is a function of the elements x_k, $k = 1, \ldots, d_s$ of subset $s, s = 1, \ldots, p$ only. They further proved that strong separability is equivalent to (partial) additivity, i.e $G(x) = F(\eta_1 + \eta_2 + \cdots + \eta_p)$.

There exists an enormous amount of papers, discussing separability for production functions, e.g. Denny and Fuss (1977), Fuss, McFadden, Mundlak (1978), or in general for demand and utility functions, see Deaton and Muellbauer (1980). Pretty often it is considered in the context of problems of aggregation and substitution (Berndt and Christensen, 1973) and much more, especially also for the specification of flexible functional forms and separated testing. Testing separability in nonparametric context is still in development, see Sperlich, Tjøstheim, Yang (1999) who did this nonparametrically in the context of interaction analysis. Separability enables econometric analysis in terms of subsets of all possible inputs, stages or with aggregates of them. Consequently we can aggregate inputs into indices. It allows thus decentralization in analysis, optimization and decision-making. For more references see also Blundell and Robin (2000) which have extended the discussion to latent separability, i.e. grouping goods even without having weak separability or in other words, allowing some goods to be in different groups.

From the statistical point of view, Stone (1985,1986) mentioned (partly we can say, added) the points flexibility, dimensionality and interpretability. So he proved that additive modeling can circumvent the curse of dimensionality what in nonparametrics is of fundamental importance and at all makes these methods feasible for higher dimensional problems. This actually carries over to the more general case when the impact function can be decomposed in lower dimensional function at all. Flexibility we already discussed before, and the advantage of interpretability is obvious since interpreting directly a higher dimensional function without the chance of separated (i.e. component wise) considerations is hardly possible.

For nonparametric estimation in separable models, i.e. combinig the above favorable properties and avoiding the curse of dimensionality, mainly two different methods are known though for both exist various modifications: the backfitting (see e.g. Mammen, Linton and Nielsen, 1999; Hastie and Tibshirani, 1990) and the marginal integration estimator (see

e.g. Tjøstheim and Auestadt, 1994; Linton, Nielsen, 1995). Their advantages, disadvantages, different and common features are investigated and discussed in detail by Sperlich, Linton and Härdle (1999) and Nielsen, Linton (1997) but only one special version of each and only for the more simple additive model case. This is since e.g. for the backfitting little theory has been provided and still is lacking for generalized additive models. Nevertheless it is believable that the found and analyzed properties and behavior carries over for more complex situations if not even worse. Additionally there exist some articles based on series estimators, see e.g. Andrews and Whang (1990) or Newey (1995), but all of pure theoretical nature without discussion of feasibility, application or simulations. Recently, Horowitz (2000) presented a conditional moment estimator for generalized additive models with unknown link function and discusses an extension to a trivial case of weak separability. As almost no structure is assumed, this is a nice approach for pure exploratory data analysis but this certainly pays with numerical deficiencies in performance. In practice, empirical researcher often prefer to impose at least some structure as well as procedures that allow for partly modeling.

A second point is that it is not clear so far how to estimate both, parameter and the lower dimensional nonparametric components in non additive but weak separable models. Especially for frequent problems as Tobit models, e.g. for censored or truncated variables or in simultaneous equation systems, nonparametric weak separable models can not yet be estimated. Apart from Horowitz (2000), also Pinske (2000) considers a special (trivial) case of weak separability but both do not consider Tobit models. For some approaches to non- or semiparametric estimation of special Tobit models we refer to Newey, Powell and Vella (1999), Lewbel, Linton (2000), or Ai and Chen (1999) from which only the latter one considers semiparametric separable models but only estimates the parametric part and is more of theoretical nature. We are looking for a computational not intensive procedure that can handle these problems and allows to provide asymptotic theory. Therefore we have chosen the (conditional) weighted or local Likelihood approach developed in Staniswalis (1989). A problem when considering conditional moment estimators for simultaneous equation systems is usually the identification, e.g. if having two simultaneous equations, maybe both with selection bias problems and maybe even nested etc. For more discussion and examples about those typical problems of non identifiability, see Rodriguez-Póo, Sperlich, Fernández (1999) or compare Ai and Chen (1999), Newey, Powell and Vella (1999). Certainly, maximum or quasi likelihood procedures need more distribution assumptions than the conditional expectation estimators but can be more generally used and usually perform better in finite sample estimation. We developed our method for both, likelihood estimation under strong distribution assumptions and the quasi likelihood approach. For semi- or nonparametric ways to relax the conditions in likelihood context, see among others Severini and Stanisvalis (1994) and Heckman and Singer (1984). Actually, often errors in likelihood estimation caused by violation of distribution assumptions often are less serious as sometimes believed. Instead, for models typical in economic research, the errors due to a misspecification of the Link is negligible in comparison to misspecification in the index functions, see e.g. Fernández and Rodríguez-Poó (1997). So usually, switching from a generalized linear models (GLM)

to a single index model (SIM) with unknown link does not really change the final results whereas modeling the index flexible does a lot. For this reason testing functional forms of the index, often revealing nonlinearities, is becoming more and more a topic. Among other reasons this explains the popularity of generalized additive models (GAM), compare also Härdle, Huet, Mammen, Sperlich (1999), Burda, Härdle, Müller, Werwatz (1998) or Härdle, Sperlich, Spokoiny (1997).

The organization of the rest of the paper is as follows. In Section 2 we introduce and motivate the model. We also present some examples for typical applications in empirical economics. The estimator and its asymptotic properties are given in Section 3. In Section 4 we discuss implementation and illustrate the finite sample behavior by simulations and applications to different models and different real data sets. Section 5 concludes and gives further discussion, e.g. how to test the correct choice of the link or likelihood function. The assumptions, its discussion, proofs and computational details are postponed to the Appendices.

2 The statistical model and motivation

In order to introduce our estimation procedure, let us establish formally some statistical framework. Suppose we observe random variables $Y \in I\!R$, $X \in \mathcal{X}$, $T \in \mathcal{T}$, being $\mathcal{X}$ and $\mathcal{T}$ compact sets $\mathcal{X} \subset I\!R^d$, $\mathcal{T} \subset I\!R^k$, such that the conditional density of Y given X and T is of the form

$$\ell\left(\bullet, T; \eta_1, \cdots, \eta_p; \theta\right)$$

where θ is a parameter vector and $\eta_1, \cdots, \eta_p$ are also parameters that depend on X. The η parameters might be nonparametric functions that depend on subsets of X, and it is allowed also to consider a parametric component, θ. In order to simplify our work, we will assume that the conditional density of T given X does not depend neither on the parameter vector θ nor on the parameters $\eta_1, \cdots, \eta_p$. Therefore, the log-likelihood for a single observation can be written as

$$\log \ell\left(Y, T; \eta_1, \cdots, \eta_p; \theta\right) + \log \ell\left(T|X\right) + \log \ell\left(X\right),$$

and we can concentrate our analysis in the first term. Being more precise about this expression let

$$\left\{\ell\left(\bullet, T; \eta_1, \cdots, \eta_p; \theta\right) : \eta_1 \in H_1, \cdots, \eta_p \in H_p, \theta \in \Theta\right\}$$

denote a family of conditional density functions. Assume that Θ is a compact subset of $I\!R^k$. Moreover, assume that $H_1, \cdots, H_p$ are respectively compact subsets in $I\!R$. The parameters $\eta_1, \eta_2, \cdots, \eta_p$ are functions of x, i.e. $\eta_1 = \eta_1(x^1), \eta_2 = \eta_2(x^2), \cdots, \eta_p = \eta_p(x^p)$, and the vectors $x^i \in \mathcal{X}_{d_i}$, $i = 1, \cdots, p$, are subsets of x such that $\mathcal{X} = \mathcal{X}_{d_1} \times \cdots \mathcal{X}_{d_p}$ and $\sum_{j=1}^{p} d_j = d$. Finally, η's are assumed to be unknown smooth functions $\eta_j : \mathcal{X}_{d_j} \to H_j$, that take values in a set Γ_j

$$\Gamma_j = \left\{\phi \in C^2\left(\mathcal{X}_{d_j}\right) : \phi(x^j) \subset H_j \quad \text{for} \quad \text{all} \quad x^j \in \mathcal{X}_{d_j}\right\}.$$

For the ease of presentation we always will speak of densities of X, T and treat X in the future as a continuous variable on a compact support. Note that this is by no means necessary;

we could also include discrete or even dummy variables, especially for T, but replacing the densities by point measures respectively probabilities. For the variable X which will enter in the nonparametric estimation part we refer to Delgado and Mora (1995). They showed that the impact of discrete variables can be handled nonparametrically in the same way and do even not affect the rate of convergence.

In order to motivate more our considerations, let us present two standard examples from economics.

Example 1 *Gronau (1973) considered the housewife's decision to work or not and how much to work ending up with a so called Tobit 2 model (see Amemiya, 1985). Let w^0 be the offered wage given to each housewife independently of hours worked, w^r the reservation wage, and w the actual wage. With x, z being properly chosen explanatory variables we observe*

$$(1) \qquad \begin{aligned} w_i^0 = x_i^T\gamma + u_i \quad &, \quad w_i^r = z_i^T\alpha + v_i \\ w_i = w_i^0 \quad &if \quad w_i^0 > w_i^r \ , \\ w_i = 0 \quad &if \quad w_i^0 \leq w_i^r \ , \ i = 1, 2, \ldots, N. \end{aligned}$$

To be able to estimate such a complex structure, the random errors (u_i, v_i) are assumed to be i.i.d. bivariate normal with mean zero, variances σ_u^2, σ_v^2. Often, they are additionally assumed to be independent what actually is not necessary to identify the system. Then it is recommended to estimate the parameters of interest through maximum likelihood techniques with the likelihood

$$(2) \qquad L \ = \ \prod_0 \left[1 - F\left\{\frac{x_i^T\gamma - z_i^T\alpha}{\sqrt{\sigma_u^2 + \sigma_v^2}}\right\}\right] \times \prod_1 F\left\{\frac{w_i - z_i^T\alpha}{\sigma_v}\right\} \ \sigma_u^{-1} f\left\{\frac{w_i - x_i^T\gamma}{\sigma_u}\right\},$$

where F is the cumulated standard normal distribution function and f the corresponding density. Here, $\prod_0$ is the product over all observations without job, $\prod_1$ the product over all having one.

By incorporating a set of alternative assumptions, it is also possible to estimate the parameters of interest by using a two step method proposed by Heckman (1979).

These well known estimation procedures present as a main drawback that in order to obtain consistent estimators for the parameters of interest it is necessary to assume that both the conditional distribution is known, and the index function falls within the class of a known parametric function.

If we relax the model assumption of known index in both equations of (1) and thus allow for nonlinear relations

$$w_i^0 = \eta_1(x_i) + u_i \ , \ w_i^r = \eta_2(z_i) + v_i$$

with η_1, η_2 being arbitrary smooth functions, and set $\theta = (\sigma_u, \sigma_v)^T$ we would replace the corresponding expressions in the likelihood equation (2). We can model the η's even to be constructs of lower dimensional components, compare Example 2.

This estimation problem is by far not trivial and to our knowledge so far unsolved. Certainly, in the much more simple when the model can be written (and identified) as a conditional expectation, and even having a generalized additive model, i.e. $E[Y|X] = G\left\{\sum_j^d \eta_j(X_j)\right\}$, several approaches are done. Under them, the probably most general is due to Horowitz (1999) who proposes a kernel smoothing method to estimate both nonparametrically, the link $G(\cdot)$ as well as the index functions $\eta_j(\cdot)$. Huet, Härdle, Mammen and Sperlich (1999) proved bootstrap and constructed tests on the $\eta_j(\cdot)$ and $G(\cdot)$ when Y is taken from an exponential family. Furthermore, Ai and Chen (1999) propose a (more theoretical) semi-parametric estimator based on conditional moment restrictions but concentrate only on the estimation of the parametric part. The assumption of weak separability is much weaker than the additivity. Mainly it allows for combinations and thus interaction terms between the components and many other nonlinear relationships between the different groups of variables, recall the discussion in Section 1. As indicated in the introduction, making the strong assumption of knowing the likelihood function is due to the aim of estimating also Tobit models what the abovementioned methods can not. As discussed before, there exists an increasing amount of articles that are concerning about censored and truncated models, but as indicated are pure nonparametric approaches without parametric part nor allowing for any structure as separability. Finally, our new method yields reasonable performance for small (/real) data sets.

Example 2 *Let us consider a typical Tobit 1 model with truncated variables. Imagine we are interested in a labor supply model looking on the hours of work y. Then we observe only*

$$y_i = \begin{cases} h(x_i, t_i) + u_i & \text{if } h(x_i, t_i) + u_i > 0 \\ 0 & \text{otherwise} \end{cases}$$

where $t_i \in \mathbb{R}^d$ are some dummy and $x_i \in \mathbb{R}^3$ other explanatory variables. Further, u is a normal distributed error with variance σ^2. We could model $h(\cdot)$ e.g. in the following form

$$h(x,t) = t^T\gamma + \eta_1(x_1) + \eta_2(x_2) + \eta_3(x_3)$$

what would be additivity, i.e. strong separability, or alternatively

$$h(x,t) = t^T\gamma + \eta_1(x_1) + \eta_2(x_2) + \eta_2(x_2)\eta_3(x_3)$$

being thus a weak separable model. In both, the η_j, $j = 1,2,3$ are arbitrary smooth functions with the only restrictions $E[\eta_1] = E[\eta_3] = 0$ for identification. The Likelihood to maximize would be

$$L = \prod_1 F\left(\frac{h(x_i,t_i)}{\sigma}\right)^{-1} \sigma^{-1} f\left(\frac{y_i - h(x_i,t_i)}{\sigma}\right) .$$

Setting $\theta = (\gamma^T, \sigma)^T$ we again could apply directly our method. For this, compare also our application on Section 4.

Note again, that whereas the other above mentioned methods can not be applied here, ours can.

3 The estimator

Suppose we have a sample of N independent replicates $\{(Y_i, X_i, T_i)\}_{i=1,\cdots,N}$; our goal is to estimate the sequence of parameters $\eta_1, \cdots, \eta_p, \theta$ from the sample information. We remind that the η's are unknown functions evaluated locally, i.e. at point $x_0 = \left(x_0^1, \cdots, x_0^d\right)$, respectively $\eta_1^0 = \eta_1(x_0^1), \cdots, \eta_p^0 = \eta_p(x_0^p)$. The estimation procedure is based in the weighted local likelihood approach developed in Staniswalis (1989). The proposed method consists in approximating the likelihood function locally. However, under certain hypothesis on the likelihood function we also develop an estimator which is based on maximizing a local quasi-likelihood function (See McCullagh and Nelder, 1989 and Severini and Staniswalis, 1994). The main advantage of the method based on the quasi-likelihood function is that there is no need to assume the knowledge of the conditional density function. However, its drawback is that the underlying conditional density must belong to the family of exponential functions. This rules out the possibility of considering some econometric problems that are typical in standard microeconometric analysis. On these grounds we present both alternative estimating approaches.

Let us denote the estimators of the different curves at point x_0 by $\hat{\eta}_1 = \hat{\eta}_1(x_0^1), \cdots, \hat{\eta}_p = \hat{\eta}_p(x_0^p)$. Then, the estimation is implemented through a three step procedure. The steps are the following:

1. For a given value $x_0 = \left(x_0^1, \cdots, x_0^d\right)$ and fixed θ, we estimate $\eta_1, \eta_2, \cdots, \eta_p$ as the solution to the problem

$$(\hat{\eta}_{1,\theta}, \hat{\eta}_{2,\theta}, \cdots, \hat{\eta}_{p,\theta}) = \sup_{\eta_1 \in H_1, \cdots, \eta_p \in H_p} W(\eta_1, \cdots, \eta_p, \theta),$$

where the weighted likelihood is

$$W(\eta_1, \cdots, \eta_p, \theta) = \sum_{i=1}^{N} K\left(\frac{x_0 - X_i}{h}\right) \log \ell(Y_i, T_i; \eta_1, \cdots, \eta_p, \theta),$$

where $K(\cdot)$ d-variate kernel function and h is the corresponding bandwidth. Note also that all estimators depend on θ.

2. Given the previous estimates for the nonparametric part, we perform a simple likelihood for estimating θ, i.e.

$$\hat{\theta}_N = \sup_{\theta\in\Theta} \sum_{i=1}^{N} \log \ell \left(Y_i, T_i; \hat{\eta}_{1,\theta}(X_i^1), \cdots, \hat{\eta}_{p,\theta}(X_i^p), \theta\right) \tag{3}$$

and set $\hat{\eta}_j = \hat{\eta}_{j,\hat{\theta}_N}$ for all $j = 1, 2, \ldots, p$.

3. Now, with the estimators obtained in steps 1 and 2, we re-estimate the nonparametric part as follows

$$\widehat{\eta_j}(x_0) = \sup_{\eta_j\in H_j} \sum_{i=1}^{N} K_j \left(\frac{x_0^j - X_i^j}{h_j}\right) \log \ell \left(Y_i, T_i; \hat{\eta}_1(X_i^1), \cdots, \eta_j, \cdots, \hat{\eta}_p(X_i^p), \hat{\theta}_N\right)$$

for all $j = 1, \cdots, p$.

In this method, the first two step are derived from the profile likelihood approach proposed in Severini and Wong (1992) and extended according our aims, e.g. to a vector structure for the nonparametric part. However, we must emphasize that this two steps provide a root-N consistent semiparametric efficient estimator of θ_0, but the nonparametric components are estimated with the problem of curse of dimensionality. In order to avoid this, and to take advantage of the weak separable structure we introduce additionally the third step. We will see that the resulting estimators form the third step avoid the problem mentioned above and also yield fully efficient estimators. For the nonparametric part we speak of efficient in the sense of being equivalent to an estimator based on knowing the other components of the regression function. More details and discussion about the procedure can be found in the next section whereas in this section we focus more on the theoretical part .

As it was previously remarked, the knowledge of the conditional likelihood function in some situations is a strong assumption than necessary. It is possible to relax this by taking into account the following setting:

The conditional density $\log \ell\,(Y, T; \eta_1, \cdots, \eta_p; \theta)$ is an exponential family distribution.i.e.

$$\ell\,(Y, T; \eta_1, \cdots, \eta_p; \theta) = \exp\{Y\delta - b(\delta) + c(Y)\}$$

where $\delta = (\eta_1, \cdots, \eta_p, \theta)$. In this case, by the properties of the exponential density function

$$\begin{aligned} E\,(Y|X = x, T = t) &= g(t, \eta_1(x^1), \cdots, \eta_p(x^p), \theta_0) \\ V\,(Y|X = x, T = t) &= \sigma^2 V\left(g(t, \eta_1(x^1), \cdots, \eta_p(x^p), \theta_0)\right), \end{aligned}$$

where $g(\cdot)$ and $V(\cdot)$ are known functions. Note that in both cases heteroscedastic models are included. However, to estimate the functional effect of heteroscedasticity is often a question of identification, moreover than a question of the algorithm.

In this case, it is possible to substitute in steps 1, 2 and 3 the log-likelihood $\log \ell(\cdot)$ by the quasi-likelihood function $r\,(\cdot, g(t, \eta_1(x^1), \cdots, \eta_p(x^p), \theta_0))$ that is defined as

$$r\,(y, g) = \int_g^y \frac{(s - y)}{V(s)} ds.$$

This quasi-likelihood function has been motivated by many ways; one is to interpret it as a weighted least squares as $r(y,\mu)$ is equal to $-.5(\mu-y)^2v^{-1}$ where v^{-1} is a weighted average of $1/V(s)$. We will study first the asymptotic behavior of the local maximum likelihood estimators and later the quasi-maximum likelihood estimators, but before to do so we introduce some notation and terminology that will be used in the remainder of the paper. Let us denote by $p(X)$ the marginal density of $X=(X^1,\cdots,X^p)$. Furthermore $p_j(X^j)$ is the marginal density of X^j. $\sigma^2(x)=E[Y^2|X=x]$ and $\sigma_j^2(x^j)=E[Y^2|X^j=x^j]$.

$$\varphi(y,t;\eta_1,\cdots,\eta_p,\theta) = \ln\ell(y,t;\eta_1,\cdots,\eta_p,\theta) \tag{4}$$

$$F_j^{(l)}(y,t;\eta_1,\cdots,\eta_p,\theta) = \frac{\partial}{\partial\eta_j^l}F(y,t;\eta_1,\cdots,\eta_p,\theta) \quad j=1,\cdots,p; \tag{5}$$

Where $F_j(\cdot)$ can be respectively $\ell_j(\cdot)$, $\varphi_j(\cdot)$ or $r_j(\cdot)$. Then, if $\hat{\theta}_N$ is the solution to the optimization problem (3), in Step 2, the following result is proved in Appendix I

Theorem 1 *Under assumptions (A.1)-(A.2), (B.1)-(B.3), (K.1) and (H.1), stated in the Appendix, then as N tends to infinity*

$$\sqrt{N}\left(\hat{\theta}_N-\theta_0\right)\to_d N\left(0,I_\theta^{-1}(\eta_1,\cdots,\eta_p,\theta)\right),$$

where

$$I_\theta(\eta_1,\eta_2,\cdots,\eta_p,\theta)$$

$$=E\left[\tfrac{\partial}{\partial\theta}\varphi(Y,T;\eta,\theta)\,\tfrac{\partial}{\partial\theta^T}\varphi(Y,T;\eta,\theta)\right]-E\left[\tfrac{\partial}{\partial\theta}\varphi(Y,T;\eta,\theta)\,\tfrac{\partial}{\partial\eta^T}\varphi(Y,T;\eta,\theta)\right]$$

$$\times E\left[\tfrac{\partial}{\partial\eta}\varphi(Y,T;\eta,\theta)\,\tfrac{\partial}{\partial\eta^T}\varphi(Y,T;\eta,\theta)\right]^{-1}E\left[\tfrac{\partial}{\partial\eta}\tfrac{\partial}{\partial\theta^T}\varphi(Y,T;\eta,\theta)\right],$$

and

$$\frac{\partial}{\partial\theta}\varphi(Y,T;\eta,\theta)=\left(\frac{\partial}{\partial\theta_1}\varphi(Y,T;\eta_1,\cdots,\eta_p,\theta),\cdots,\frac{\partial}{\partial\theta_k}\varphi(Y,T;\eta_1,\cdots,\eta_p,\theta)\right)^T$$

$$\frac{\partial}{\partial\eta}\varphi(Y,T;\eta,\theta)=\left(\frac{\partial}{\partial\eta_1}\varphi(Y,T;\eta_1,\cdots,\eta_p,\theta),\cdots,\frac{\partial}{\partial\eta_p}\varphi(Y,T;\eta_1,\cdots,\eta_p,\theta)\right)^T$$

As it can be observed from this result, the semiparametric estimator achieves the semiparametric efficiency bound (see Chamberlain, 1992 or Newey, 1990, 1995). Further note that the asymptotic variance can be approximated with the aid of the Hessian matrix, something we get automatically out from the procedure, e.g. when using the Newton-Raphson algorithm.

In order to show the asymptotic behavior of the nonparametric estimators obtained in Step 3, we need some further assumptions on bandwidths, kernel functions and identification of the η_j, $j=1,\ldots,p$.

(C.1) $Nh_j^{d_j} \to \infty$ and $Nh_j^{d_j+4} \to 0$, for $j = 1, \cdots, p$, as N tends to infinity.

(C.2) The support of kernel K_j is compact and $\int tK_j(t)dt = 0$, for any j.

(C.3) For all $j = 1, \cdots, p$

$$E\left[\varphi_j^{(1)}\left(Y, T; \eta_1(X^1), \cdots, \eta_j(X^j), \cdots, \eta_p(X^p)\right) | X^j = x^j\right] = 0$$

All previous assumptions are standard in nonparametric regression literature. For example, condition (C.1) makes variance and bias tend to zero when the sample size increases. Condition (C.3) is an identification condition that is similar, in additive models under gaussian errors to the backfiting algorithm (see Hastie and Tibshirani, 1990). Then, it it shown in Appendix I that

Theorem 2 *Under the conditions of Theorem 1 and assumptions (C.1) to (C.3) we have for any j*
i)

$$\frac{\sqrt{Nh_j^{d_j}}\left(\widehat{\eta}_j(x_0^j) - \eta_j(x_0^j)\right)}{V_j^{1/2}\left(\widehat{\eta}_j(x_0^j), \hat{\theta}_N\right)} \longrightarrow_d N(0,1),$$

ii)

$$\sup_{x_0^j \in \mathcal{X}_{d_j}} \left|\widehat{\eta}_j(x_0^j) - \eta_j(x_0^j)\right| = O_p\left(\sqrt{\frac{\log N}{Nh_j^{d_j}}}\right),$$

where

$$(6) \qquad V_j\left(\eta_j, \theta_0\right) = \frac{\int K_j^2(t)dt}{p_j(x_0^j) I_j\left(\eta_j\right)},$$

$$(7) \qquad I_j\left(\eta_j, \theta_0\right) = E\left[\varphi_j^{(1)}\left(Y, T; \eta_1(X^1), \cdots, \eta_j(X^j), \cdots, \eta_p(X^p), \theta_0\right)^2 \middle| X^j = x_0^j\right]$$

as N tends to infinity.

As it can be remarked from this theorem, all nonparametric components are estimated at the minimum possible rate overcoming the curse of dimensionality. This achievement agrees with the results found by Stone (1986) for additive models, but we remark that the same result is now achieved with a weaker restriction as the weak separability. Also, as indicated before, we reach the same asymptotics as if the other components in the model would have been known. Finally, again the (pointwise) asymptotic expressions can be drawn out from the algorithm and not much calculation, or even plug-in estimation, is necessary.

Quasi-likelihood estimation

If we replace in all steps of our estimation procedure the log-likelihood function by the quasi-likelihood function, then it is shown in Appendix I that the previous results still hold, although as it could be expected there is an efficiency loss if the specification is not equivalent to the real distribution.

Theorem 3 *Under assumptions (B.1')-(B.3'), (K.1), (H.1) and (Q.1)-(Q.3), stated in the Appendix, then as N tends to infinity*

$$\sqrt{N}\left(\hat{\theta}_N - \theta_0\right) \to_d N\left(0, J_\theta^{-1}\left(\eta_1, \cdots, \eta_p, \theta\right)\right),$$

where

$$J_\theta\left(\eta_1, \eta_2, \cdots, \eta_p, \theta\right)$$

$$= E\left[\tfrac{\partial}{\partial\theta} g(T;\eta,\theta_0)\tfrac{\partial}{\partial\theta^T} g(T;\eta,\theta_0)\right] - E\left[\tfrac{\partial}{\partial\theta} g(T;\eta,\theta_0)\tfrac{\partial}{\partial\eta^T} g(T;\eta,\theta_0)\right]$$

$$\times E\left[\tfrac{\partial}{\partial\eta} g(T;\eta,\theta_0)\tfrac{\partial}{\partial\eta^T} g(T;\eta,\theta_0)\right]^{-1} E\left[\tfrac{\partial}{\partial\eta}\tfrac{\partial}{\partial\theta^T} g(T;\eta,\theta_0)\right],$$

where

$$\frac{\partial}{\partial\theta} g(T;\eta,\theta) = \left(\frac{\partial}{\partial\theta_1} g(T;\eta_1,\cdots,\eta_p,\theta),\cdots,\frac{\partial}{\partial\theta_k} g(T;\eta_1,\cdots,\eta_p,\theta_0)\right)^T$$

and

$$\frac{\partial}{\partial\eta} g(T;\eta,\theta) = \left(\frac{\partial}{\partial\eta_1} g(T;\eta_1,\cdots,\eta_p,\theta),\cdots,\frac{\partial}{\partial\eta_k} g(T;\eta_1,\cdots,\eta_p,\theta_0)\right)^T.$$

As it can be expected the quasi-maximum likelihood estimator of θ is not always efficient. However, if the model is correctly specified both criterion function coincides and therefore we obtain the efficient estimator that was shown in Theorem 1.

Before we give the next result we need to incorporate the following assumption

(C.3') For all $j = 1, \cdots, p$

$$E\left[\frac{Y - g(T;\eta_1(X^1),\cdots,\eta_j(X^j),\cdots,\eta_p(X^p),\theta_0)}{V\left(g(T;\eta_1(X^1),\cdots,\eta_j(X^j),\cdots,\eta_p(X^p),\theta_0)\right)}\right.$$

$$\left.\times\frac{\partial}{\partial\eta_j} g(T;\eta_1(X^1),\cdots,\eta_j(X^j),\cdots,\eta_p(X^p),\theta_0)|X^j = x^j\right] = 0$$

what is the same as the identification condition (C.3) but now in terms of the quasi likelihood function. In the next result we also show that the nonparametric estimators of the additive components avoid also the curse of dimensionality.

Theorem 4 *Under the conditions of Theorem 3 and assumptions (C.1), (C.2) and (C.3') we have for any j*

i)

$$\frac{\sqrt{Nh_j^{d_j}}\left(\widehat{\eta}_j(x_0^j) - \eta_j(x_0^j)\right)}{V_j^{1/2}\left(\widehat{\eta}_j(x_0^j), \hat{\theta}_N\right)} \longrightarrow_d N(0,1),$$

ii)

$$\sup_{x_0^j \in \mathcal{X}_{d_j}} \left|\widehat{\eta}_j(x_0^j) - \eta_j(x_0^j)\right| = O_p\left(\sqrt{\frac{\log N}{Nh_j^{d_j}}}\right),$$

where

$$V_j\left(\eta_j, \theta_0\right) = \frac{\int K_j^2(t)dt}{p_j(x_0^j) J_j\left(\eta_j\right)},$$

$$J_j\left(\eta_j, \theta_0\right) = E\left[\frac{1}{V\left(g(T; \eta_1(X^1), \cdots, \eta_j(X^j), \cdots, \eta_p(X^p), \theta_0)\right)} \times \frac{\partial}{\partial \eta_j} g(T; \eta_1(X^1), \cdots, \eta_j(X^j), \cdots, \eta_p(X^p), \theta_0) \middle| X^j = x_0^j\right]$$

as N tends to infinity.

For discussion of efficiency and other remarks we refer to the detailed statements above. A careful check of the proof reveals that the same statements can be done for time series data with some strong mixing conditions. However for transparency of the ideas, especially in the proof we have restricted ourselves on the independent case and instead refer to Vieu (1991).

4 The Procedure, Simulations and Applications

In this section we first discuss some questions in practical application including computational remarks and give some simulation results. Secondly we present a real data example reflecting the aforementioned items and demonstrating the feasibility and performance of our method in empirical economic research.

The implementation of the procedure can be done in various ways. Among them, slight modifications yield the same numerical results but could be more attractive from the computational point of view. For the maximization of the (log-) Likelihood the Newton Raphson or Fisher Scoring are popular methods but need the Hessian matrix. This, if $\vec{\eta} := (\eta_1, \ldots, \eta_p)^T$ is a vector of not additively combined functions, can be very tedious what can be easily seen

in Appendix. Further, often you want to separate the functions more detailed than it is possible in the first step without running in problems of identification. An obvious modification is to let η a lower dimensional vector then $\overrightarrow{\eta}$ or even be the multidimensional function $\mathbb{R}^q \mapsto \mathbb{R}$, replacing then $\overrightarrow{\eta}$ by η and first get this way $\hat{\theta}$ and $\hat{\eta}$. In the third step, with appropriate initials for $\eta_1, \ldots, \eta_p$, given $\hat{\theta}$, $\hat{\eta}$ we can estimate $\overrightarrow{\eta}$ as suggested in Theorem 2. In the case of separating, say η_1 from the first step into $\eta_{1,1} + \eta_{1,2}$ in the third step. This demands a solution of two lower dimensional maximization equation, i.e. maximizing $\forall\, l = 1, \ldots, n$

$$\widehat{\eta_{1,1}}(X_l^{1,1}) = \sup_{\eta_{1,1} \in H_{1,1}} \sum_{i=1}^{N} K\left(\frac{X_l^{1,1} - X_i^{1,1}}{h}\right) \log \ell\left(Y_i, T_i; \eta_{1,1}, \hat{\eta}_{1,2}(X_i^{1,2}), \cdots, \hat{\eta}_p(X_i^p), \hat{\theta}_N\right)$$

$$\widehat{\eta_{1,2}}(X_l^{1,2}) = \sup_{\eta_{1,2} \in H_{1,2}} \sum_{i=1}^{N} K\left(\frac{X_l^{1,2} - X_i^{1,2}}{h}\right) \log \ell\left(Y_i, T_i; \hat{\eta}_{1,1}(X_i^{1,1}), \eta_{1,2}, \cdots, \hat{\eta}_p(X_i^p), \hat{\theta}_N\right)$$

An iteration over this second step can be performed to improve numerically the final result. Further advantages of this implementation are that the result of $\hat{\theta}$ does not depend on the modeling of the nonparametric part and consequently, when comparing different combinations for $\overrightarrow{\eta}$, only the second step has to be modified and repeated. For more details see also Appendix.

The problem of bandwidth choice is not that problematic in practice as could be expected. On the one hand the optimal bandwidth could be estimated with plug-in methods as we give explicit expressions for the asymptotics of the estimators of θ and $\overrightarrow{\eta}$. The necessary rate is given in condition (H.1). On the other hand, in practice all we need is smoothing the nonparametric part sufficiently to reach convergence for the Newton Raphson algorithm. A small simulation study confirmed this strongly. In general we chose the smallest bandwidth that yield convergence. For questions of weighting or trimming we refer to the application part as for simulated data the estimator worked perfectly without any trimming.

When considering weak separability but not additivity and allowing for arbitrary smooth functions, a rather sophisticated problem in practice can be the proper model specification that identifies uniquely the components of $\overrightarrow{\eta}$. Consider e.g. Example 2. The nonparametric part is equivalent to

$$\eta(x_1, x_2, x_3) = \eta_1(x_1) + \eta_2(x_2)\{1 + \eta_3(x_3)\}\ . \tag{8}$$

Note that the restrictions

a) $E[\eta_1(x_1)] = E[\eta_2(x_2)] = 0$, η_3 arbitrary,

b) $E[\eta_1(x_1)] = E[\eta_3(x_3)] = 0$, η_2 arbitrary,

c) $E[\eta_2(x_2)] = E[\eta_3(x_3)] = 0$, η_1 arbitrary,

can lead to different results, not only "different up to a constant". Thus, whereas it is a minor problem from the mathematical or statistical point of view, this is a much harder point for the practitioner, especially as for him this problem so long never appeared in this form.

To demonstrate the performance we present two simulation results for Example 2: Consider a typical Tobit 1 model with truncated variables. We simulated the model

$$y_i = \begin{cases} h(x_i, t_i) + u_i & \text{if } h(x_i, t_i) + u_i > 0 \\ 0 & \text{otherwise} \end{cases} \tag{9}$$

with $t_i \sim U[0,2]$, $x_i \sim U[-1,1]^3$, first additively

$$h(x,t) = t^T\gamma + \eta_1(x_1) + \eta_2(x_2) + \eta_3(x_3)$$

but then

$$h(x,t) = t^T\gamma + \eta_1(x_1) + \eta_2(x_2) + \eta_2(x_2)\eta_3(x_3).$$

We set in both cases $\beta^T = (-1.5, 2)$, $\eta_1(x_1) = 2\sin(2x_1)$, $\eta_2(x_2) = 4x_2^2$, and $\eta_3(x_3) = 2x_3 + c$, where $c = 0$ in the additive model and $c = 1$ in the other one. The error distribution was standard normal. We draw about 600 observations to end up always with 400 uncensored observations. For the estimation we used the identification condition $E[\eta_1] = 0$, $E[\eta_3] = 1$. It turned out that in our simulations for the non additive model we had a slight identification problem, probably due to the small sample size, and often overestimated both, the slopes of the functions and the variance (see also Figure 1 right side), what cancels in the index. Nevertheless, for large samples these numerical effects vanish. For the additive model we got after 150 replications for $\hat{\theta} = (\beta_1, \beta_2, \sigma)^T$ in the mean $(-1.51, 1.99, 1.59)^T$ with the standard deviation $(0.155, 0.156, 0.056)^T$. In Figure 1 we give the bands yield after the 150 replications for all functions in both models only skipping the two worst estimates. This corresponds approximately to 99% confidence bands. The dashed lines are the data generating functions.

Figure 1 about here.

We used for estimating θ bandwidth $h = 2.25$ for all directions. This was the smallest possible without running to often into numerical problems as we did no trimming! Because of the construction of the second model it is recommended to smooth the last component more than the other ones. For a better comparison we did this also in the additive model. Thus, in step 3 we used $h_1 = h_2 = 0.75$, but $h_3 = 1.25$. It can be seen that in both cases the estimator works pretty well for such a complex structure.

4.1 Female labor supply in West and East Germany

We aim to apply our method on a typical limited dependent variable problem. Consider female labor supply measured in real hours of work for married woman. Note that this variable accounts for the number of hours per week that the woman has declared to work and not the number of contracted hours. The hours are assumed to be generated by equation (9), Example 2. The difference to Example 2 will be that now $x \in I\!R^6$ and we try more than only

$$h_{weak}(t,x) = t^T\gamma + \eta_1(x_1) + \cdots + \eta_5(x_5) + \eta_5(x_5)\eta_6(x_6) \tag{10}$$

but will also consider the additive case

$$h_{strong}(t,x) = t^T\gamma + \eta_1(x_1) + \cdots + \eta_5(x_5) + \eta_6(x_6)\ . \tag{11}$$

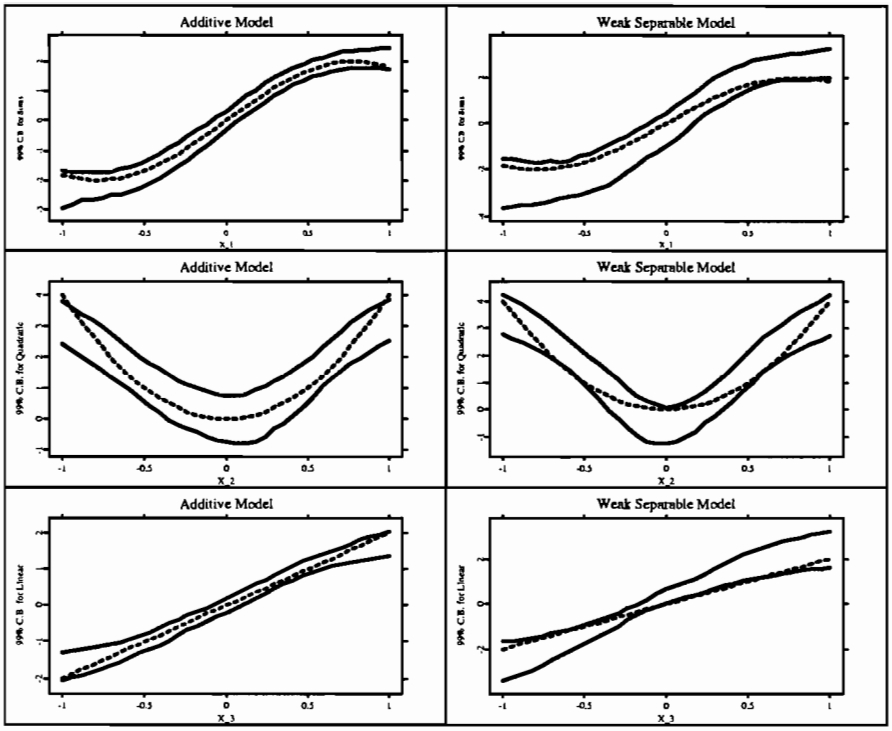


Figure 1: 99% Confidence Bands for all functions in both models based on 150 replications. Dashed lines are the data generating functions.

To make them comparable we restrict in both cases $E[\eta_j(x_j)] = 0$, $j = 1, 2, 3, 4, 6$ but nothing for η_5. If by this separability assumption the model is well identified and specified, x_5, x_6 are more or less independent, we should get out the same estimates for both specifications, up to a multiplying constant $c = E[\eta_5(x_5)]$ for η_6. Here, equation (10) is a trial to model possible interaction.

Many different (parametric) specifications have been tried to model the hours function in this context. A most famous one is the study about the sensitivity against economic and statistic assumptions by Mroz (1987). We concentrate on a comparison of specifications (11) and (10) as well as of possibly different behavior of married woman in East and West Germany three years after unification, i.e. in 1993. Those comparisons became quite popular as, due to completely different political, economic and social systems before 1990, the levels of employment of woman where quite different also; in 1993 in the East still about 65%, in the West about 54%. The studies consequently concentrated on participation at all. Among them, Kempe (1997) tried a semiparametric analysis and found clear differences in behavior, not only in slopes but even in functional forms of the (additive) impact functions. For more motivation and discussion see also Holst, Schupp (1991,1994) or Merz (1990).

We use the same data as Kempe (1997), taken from the Social Economic Panel of Germany, wave 1993, cleaned for persons with missing values in the relevant questions and skipping

East Germans living in West, West Germans living in East. As the Likelihood runs only over the employed (married) woman, we have just 681 observations for West and 611 for East Germany. We chose the explaining variables along the aforementioned articles and took the number of children (Ch1= $\mathbb{1}\{one\ child\}$, Ch2= $\mathbb{1}\{more\ children\}$), education (Edu1= $\mathbb{1}\{high\ school\}$, Edu2= $\mathbb{1}\{academic\ degree\}$) and unemployment rate of the country the person lives in (Urate) for the linear part ($t^T\gamma$). Note, that in East Germany there are only 5 countries. For the nonparamertic part $\eta(x)$ we have age of woman (Age), net wage per real hours (Wage), prestige index of their job (PI) and number of years of interruption of professional career (off). For further main income and expenditures we included also the net income of partner per month (Income), and the expenditures for flat minus net income from letting flats (R & L = rent-let). As indicated before, by modeling (10) we want to allow for some interaction between Income and rent-let.

Table 1 about here.

	West		East	
bounds:	lower	upper	lower	upper
Wage	-	50.0	-	30.1
PI	0.0	70.1	0.0	75.0
off	-	36.6	-	12.5
Income	-	11.0	-	5.45
R & L	-12.5	-	-1.4	1.4

Table 1: Trimming for x to calculate $\hat{\sigma}_{x_j}$, $j = 1, \ldots, 6$ and the convergence criteria. Age is skipped here as it was not trimmed at all. x_5, x_6 in 1000 DM.

For the semiparametric estimation, we used trimming for input x in two steps; when calculating standard deviations for each x_j and when calculating the convergence criteria (for the Newton Raphson). The standard deviations were used to determine the bandwidths. For West Germany we took always $h_j = 1.25\hat{\sigma}_{x_j}$, $j = 1, \ldots, 6$, for East Germany $h_j = 1.5\hat{\sigma}_{x_j}$ as we had less data. We trimmed as given in Table 1.

Table 2 about here.

We first consider the comparison of the different specifications and focus for presentation on the West German data. In Figure 2 and Table 2 (left side for West Germany) we see the results for the additive case, equation (11). In the table are given additionally the results for a pure parametric linear model (first two columns), all with its standard deviations in brackets. In the parametric model we introduced Age**2, used ln(Wage) instead of Wage and tried different models but give only results for this model which was the best. This parametric analysis was only done to compare with the the parameter estimates $\hat{\theta} = (\hat{\gamma}^T, \hat{\sigma})$ in our semiparametric model. It can be seen that, apart from Edu2 for East Germans, the coefficient estimates do hardly change but as well the error variance (σ) as well as the variances of the estimates could be reduced a lot using semiparametric methods. These findings are in accordance with those made in Rodriguez-Póo, Sperlich, Fernández (1999).

	West Germany				East Germany			
Ch1	-7.847	(1.087)	-6.913	(.7850)	-2.702	(1.054)	-2.152	(.9910)
Ch2	-11.91	(1.221)	-10.84	(.9549)	-2.313	(1.178)	-2.040	(1.130)
Edu1	-.1027	(1.777)	.5738	(1.383)	1.670	(1.300)	1.318	(1.180)
Edu2	.1403	(2.070)	2.125	(2.084)	1.575	(1.610)	4.868	(1.562)
Urate	.2003	(.2254)	.0925	(.1587)	-.5204	(.3242)	-.4256	(.2934)
Age	1.351	(.4662)	-	(-)	1.460	(.4034)	-	(-)
Age**2	-.0184	(1.E-6)	-	(-)	-.0186	(1.E-6)	-	(-)
ln(Wage)	-7.431	(1.067)	-	(-)	-4.126	(.9695)	-	(-)
PI	.2673	(.0436)	-	(-)	.0820	(.0300)	-	(-)
off	-.3485	(.0616)	-	(-)	-.7367	(.1741)	-	(-)
Income	-.1206	(.0245)	-	(-)	-.1200	(.0316)	-	(-)
R & L	.0188	(.0141)	-	(-)	.1092	(.0469)	-	(-)
σ	10.21	(.2961)	6.955	(.1145)	7.828	(.2241)	6.303	(.1803)
Const	24.06	(9.317)	32.44	(-)	30.16	(9.118)	47.25	(-)

Table 2: Results for parametric linear model (columns 1,2 and 5,6) and the semiparametric model (columns 3,4 and 7,8). The standard deviations are given in brackets. In the last line, for the semiparametric model *Const* refers to $\widehat{E}[\eta_5(X_5)] = \frac{1}{N}\sum_i \hat{\eta}_5(x_{i5})$.

We want to emphasize that we could reduce σ a lot. This is a good indicator for a real improvement in the empirical part of economic research.

Figure 2 about here.

Figure 3 about here.

Compare now figures 2 and 3. In Figure 3 are given the results for the two last component estimates from the specification (10), η_5 also given after centering to zero for a better comparison. On the bottom of all graphs are given crosses for each observation to indicate the density of the corresponding variable. Surprisingly, up to a multiplying constant c for η_6, they are all the same. For this reason the other components for model (10) are not shown as they are exactly the same as we see them in Figure 2. Moreover, $c = Const$ from Table 2. Note that further $corr(x_5, x_6) = .106$. This could be taken as a good sign that our model is very well specified and thus the results quite robust against slight modifications.

Figure 4 about here.

Now we look on a comparison between the West and the East. As said in the beginning, they come from completely different political, social and economic systems, and though in 1993 at least the political and the economic systems were the same, there were still differences in the economic and political environments. We want to mention only some specials from the East: the unemployment rate was much higher in the East (in 1993), a higher willingness and motivation of women to search a job, partly based on the lower salaries (compared to the West) of their husbands, a much wider provision of kinder gardens and other possibilities

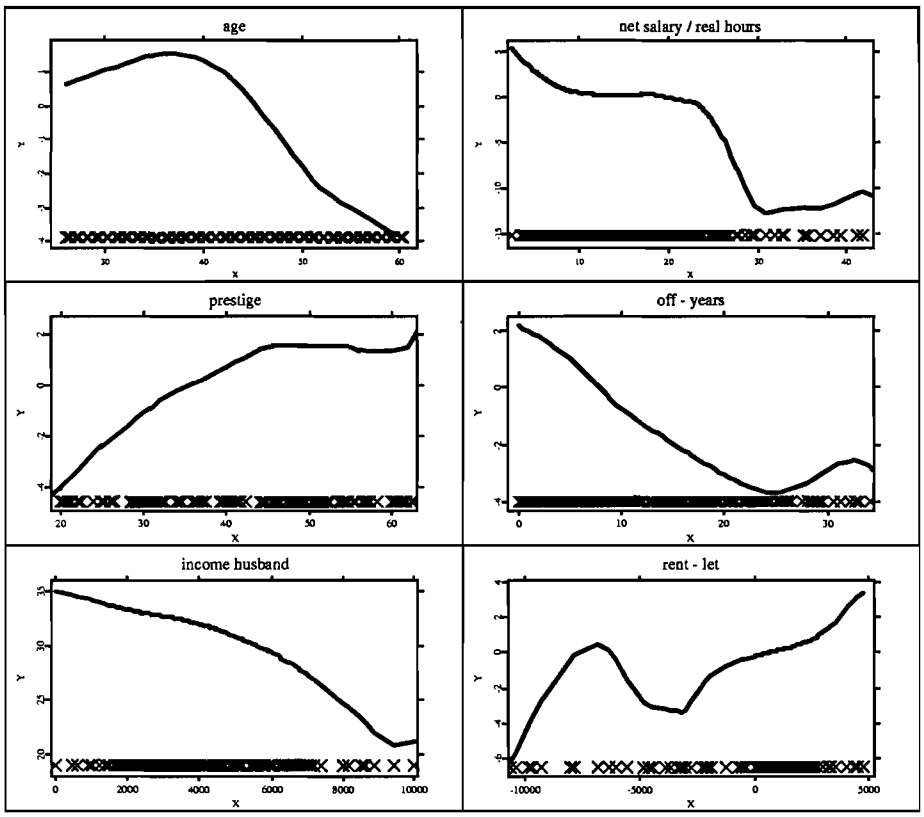


Figure 2: West German women. Results for specification (11) for the non trimmed range. Here, η_5 is centered to zero. Crosses stand for the observations to indicate the density.

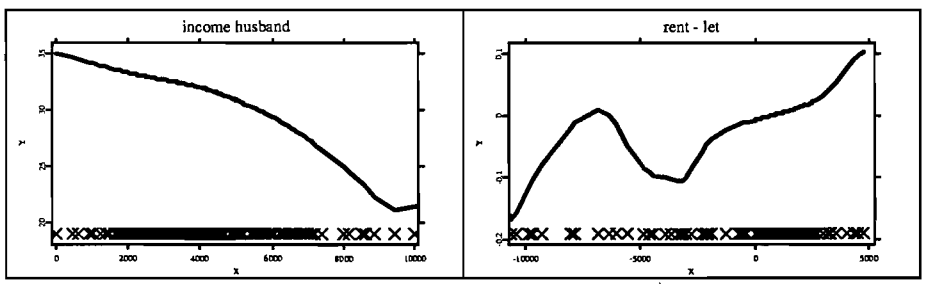


Figure 3: West German women. Results for last two components in specification (10) on non trimmed range. Here, η_5 is centered to zero. Crosses stand for the observations to indicate the density.

to leave his children in the East. The results are provided in Table 2, Figure 2 (for the West) and 4 (for the East), all based on model (11). As we concern only for the demonstration of the feasibility of our method and hereby a comparison of possible different behavior in East

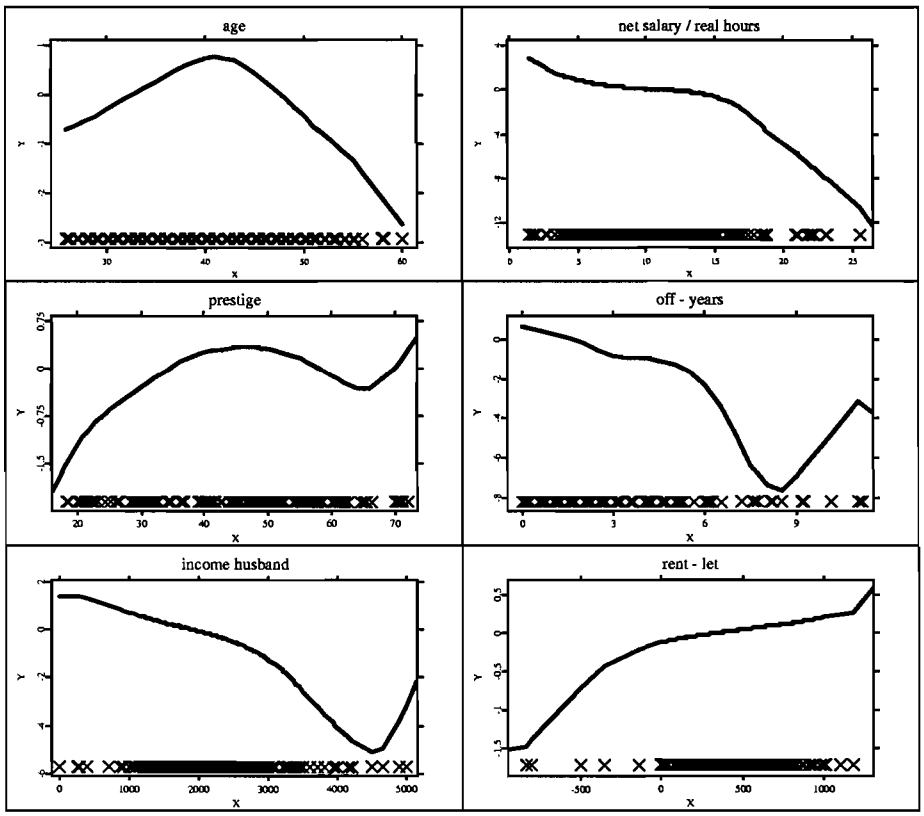


Figure 4: East German women. Results for specification (11) for the non trimmed range. Here, η_5 is centered to zero. Crosses stand for the observations to indicate the density.

and West, we will not discuss the particular coefficients neither the shapes of the estimated nonparametric components. Apart from the fact that our method obviously works also fine for this different, smaller data set (N=611), we find that behavior for labor supply measured in real hours of work is pretty the same in the East and the West, except for education and number of children. The latter outcome was expected for aforementioned reasons. Comparing with e.g. Kempe (1997) who used the same sample, this is a little bit surprising as he found big differences in behavior when looking on participation at all and was thus in accordance with e.g. Holst and Schupp (1994).

5 Conclusion and Discussion

In this article we have presented a new method for estimating weak and strong separable models typical in economics. Assuming the knowledge of the conditional distribution of the random errors we are able to include complex regression systems as Tobit models. This, to our knowledge is so far only possible for purely nonparametric models without the opportu-

nity of imposing structure or a parametric part. Aside of the theoretical consequences, the use of maximum likelihood makes the estimator feasible in small data sets typical in empirical research. This has been demonstrated in Section 4 together with a detailed discussion of problems related to application.

For the future it will be necessary to spend also more attention to the distribution assumption. In the case of SIM, i.e. models of the form $E[Y|X] = G\{\beta^T t + \eta(x)\}$, $\eta(\cdot)$ being nonparametric, you can test the specification of $G(\cdot)$ as proposed by Härdle, Mammen and Proenca (2000), as done in the aforementioned paper by Härdle, Huet, Mammen, Sperlich (1999). Often, also approximate χ^2 tests could be applied on the residuals.

Finally, a detailed simulation study comparing the different methods could reveal information, if there is the alternative of different methods, which one does best. Nevertheless, often the alternative is simply not given, as e.g. when having truncated variables.

Appendix I

Proof of the main results

In order to show the main results we need to introduce the following definitions and assumptions:

Let μ_x denote a k-vector of nonnegative integer constants. For such vector define
(i) $|\mu_x| = \sum_{j=1}^{k} \mu_j$, where $\mu_x = (\mu_1, \cdots, \mu_k)^T$,
(ii) For any function $a(x)$ on $\mathbb{R}^k$.

$$D^{\mu_x} a(x) = \frac{\partial^{|\mu_x|}}{\partial x_1^{\mu_1} \partial x_2^{\mu_2} \cdots \partial x_k^{\mu_k}} a(x)$$

Assumptions

(A.1) For fixed but arbitrary $\theta_1, \eta_1^+, \cdots, \eta_p^+$, where $\theta_1 \in \Theta$, and $\eta_1^+ \in H_1, \cdots, \eta_p^+ \in H_p$, let

$$\rho(\eta_1, \eta_2, \cdots, \eta_p, \theta) = \int \varphi(y, t; \eta_1, \cdots, \eta_p, \theta)\, \ell\left(y, t; \eta_1^+, \cdots, \eta_p^+, \theta_1\right) dy,$$

$$\theta \in \Theta, \eta_1 \in H_1, \cdots, \eta_p \in H_p$$

If $\theta \neq \theta_1$, then

$$\rho(\eta_1, \eta_2, \cdots, \eta_p, \theta) < \rho\left(\eta_1^+, \eta_2^+, \cdots, \eta_p^+, \theta_1\right)$$

Let $I_\theta(\eta_1, \eta_2, \cdots, \eta_p, \theta)$ denote the marginal Fisher information for θ in the parametric model, that is

$$I_\theta(\eta_1, \eta_2, \cdots, \eta_p, \theta)$$

$$= E\left[\frac{\partial}{\partial\theta}\varphi(Y,T;\eta,\theta)\frac{\partial}{\partial\theta^T}\varphi(Y,T;\eta,\theta)\right] - E\left[\frac{\partial}{\partial\theta}\varphi(Y,T;\eta,\theta)\frac{\partial}{\partial\eta^T}\varphi(Y,T;\eta,\theta)\right]$$

$$\times E\left[\frac{\partial}{\partial\eta}\varphi(Y,T;\eta,\theta)\frac{\partial}{\partial\eta^T}\varphi(Y,T;\eta,\theta)\right]^{-1} E\left[\frac{\partial}{\partial\eta}\frac{\partial}{\partial\theta^T}\varphi(Y,T;\eta,\theta)\right]$$

where

$$\frac{\partial}{\partial\theta}\varphi(Y,T;\eta,\theta) = \left(\frac{\partial}{\partial\theta_1}\varphi(Y,T;\eta_1,\cdots,\eta_p,\theta),\cdots,\frac{\partial}{\partial\theta_k}\varphi(Y,T;\eta_1,\cdots,\eta_p,\theta)\right)^T$$

and

$$\frac{\partial}{\partial\eta}\varphi(Y,T;\eta,\theta) = \left(\frac{\partial}{\partial\eta_1}\varphi(Y,T;\eta_1,\cdots,\eta_p,\theta),\cdots,\frac{\partial}{\partial\eta_p}\varphi(Y,T;\eta_1,\cdots,\eta_p,\theta)\right)^T$$

Then assume that the matrix $I_\theta(\eta_1, \eta_2, \cdots, \eta_p, \theta)$ is positive definite for all $\theta \in \Theta$ and $\eta_1 \in H_1, \cdots, \eta_p \in H_p$.

(A.2) Assume that for vectors $|r_\eta| \leq 4$ and $|s_\theta| \leq 4$ such that $|r_\eta| + |s_\theta| \leq 4$ the function

$$D^{r_\eta} D^{s_\theta}\varphi(Y,T;\eta_1,\cdots,\eta_p,\theta)$$

exists for almost all Y and T and that

$$E\left\{\sup_\theta \sup_\eta |D^{r_\eta} D^{s_\theta}\varphi(Y,T;\eta_1,\cdots,\eta_p,\theta)|^2\right\} < \infty.$$

(B.1) For each $\theta \in \Theta$ and $x \in \mathcal{X}$ let us define

$$h(\theta,\eta_1,\cdots,\eta_p,x) = E\{\varphi(Y,T;\eta_1,\cdots,\eta_p,\theta)\,|X = x\}.$$

Then

$$\sup_{\theta,\eta_1,\cdots,\eta_p,x} \left|D^{r_\eta} D^{s_\theta} D^{t_x} h(\theta,\eta_1,\cdots,\eta_p,x)\right| < \infty$$

for $2 \leq |r_\eta| \leq 4$, $|s_\theta| \leq 2$, $|t_x| \leq 1$ and $|r_\eta| + |s_\theta| + |t_x| \leq 4$.

(B.2) Let the vector $\bar{\eta}_\theta(x) = (\bar{\eta}_{1,\theta}(x^1),\cdots,\bar{\eta}_{p,\theta}(x^p))^T$ be the solution to

$$\frac{\partial}{\partial\eta} h(\theta,\eta_1,\cdots,\eta_p,x) = 0,$$

with respect to η for each fixed θ and x. $\bar{\eta}_\theta(x)$ is unique and for any constant $\epsilon > 0$ there exists another $\delta > 0$ such that

$$\sup_\theta \sup_x |\frac{\partial}{\partial\eta_j} h(\theta,\overline{\eta_\theta}(x),x)| \leq \delta$$

implies that

$$\sup_\theta \sup_x |\overline{\eta_{j,\theta}}(x) - \eta_{j,\theta}(x)| \leq \epsilon$$

for $j = 1,\cdots,p$.

(B.3) Let

$$\Delta_{\eta,\theta}^{r_\eta,s_\theta}(Y,T) = D^{r_\eta} D^{s_\theta} \varphi\,(Y,T;\eta_1,\cdots,\eta_p,\theta)$$

and let $f_\theta^{(r_\eta,s_\theta)}\,(y,t|x)$ denote the conditional density of $\Delta_{\eta,\theta}^{r_\eta,s_\theta}(Y,T)$ given $X=x$. Then

(i) $E\left(\sup_\eta \sup_\theta \left|\Delta_{\eta,\theta}^{r_\eta,s_\theta}(Y,T)\right|\right) < \infty$ for $|r_\eta| \leq 5$ and $|s_\theta| \leq 3$.

(ii) For some even integer $q \geq 10$ then $\sup_\eta \sup_\theta E\left\{\left|\Delta_{\eta,\theta}^{r_\eta,s_\theta}(Y,T)\right|^q\right\} < \infty$, for $|r_\eta| \leq 3$ and $|s_\theta| \leq 4$.

iii) $\sup_\eta \sup_\theta \sup_{y,x,t} \left|f_{\eta\theta}^{(r_\eta,s_\theta)}\,(y,t|x)\right| < \infty$ for $|r_\eta| \leq 4$ and $|s_\theta| \leq 3$

iv) $\sup_x |D^{t_x} p(x)| < \infty$ and $\sup_\eta \sup_\theta \sup_{y,x,t} |D^{t_x} f_{\eta\theta}\,(y,t|x)| < \infty$ for $|t_x| \leq m+2$.

v) $0 < \inf_x p(x) < \sup_x p(x) < \infty$.

(K.1) The kernel $K(\cdot)$ is a real valued function on $\mathbb{R}^d$ such that, it is compactly supported with $z = (z_1, z_2, \cdots, z_d)^T$, $z_i \in \mathbb{R}$

$$\int z_1^{i_1} \cdots z_d^{i_d} K\,(z_1, z_2, \cdots, z_d)\, dz_1 \cdots dz_d =$$

$$1 \quad \text{if} \quad i_1 = i_2 = \cdots = i_d = 0$$

$$0 \quad \text{if} \quad 0 < i_1 + i_2 + \cdots + i_d < m$$

$$\int |z|^i |K(z)| dz < \infty \quad \text{for} \quad i = 0 \quad \text{and} \quad i = m.$$

and

$$\sup_z |D^{t_z} K(z)| < \infty \qquad \text{for} \quad |t_z| \leq m+2.$$

(H.1) h_N is a sequence of constants satisfying $h_N = O_P\,(N^{-\alpha})$,

$$\frac{1}{4m} < \alpha < \frac{1}{4d}\frac{q-p-2}{2p+q+4}$$

such that $\frac{m}{d} > \frac{q-p-2}{2p+q+4}$.

(Q.1) Let $\mathcal{G}$ denote a compact subset of $\mathbb{R}$ such that $g(t,\eta_1(x^1),\cdots,\eta_p(x^p),\theta) \in \mathcal{G}$ for all $t \in \mathcal{T}$, $x^1 \in \mathcal{X}_{d_1},\cdots,x^p \in \mathcal{X}_{d_p},\eta_1 \in H_1,\cdots,\eta_p \in H_p$ and $\theta \in \Theta$. Then $\sup_{\mathcal{G}} V\,(g) < \infty$, $\inf_{\mathcal{G}} V\,(g) > 0$, $\sup_{\mathcal{G}} \Omega(g) < \infty$ and $\sup_{\mathcal{G}} \int^g \Omega(s) ds$, where

$$\Omega(g) = \int^g \frac{ds}{V(s)}.$$

(Q.2) For $p = 1,\cdots,3$ then $\frac{\partial^p V(g)}{\partial g^p}$ exists and it is bounded for all $g \in \mathcal{G}$.

(Q.3) The function $g(\cdot)$ is at least three times continuously differentiable bounded with respect all its arguments.

(B.1') The same as for (B.1) replacing $\varphi(\cdot)$ by $r(\cdot)$.

(B.2') The same as for (B.2) replacing $\varphi(\cdot)$ by $r(\cdot)$.

(B.3') The same as for (B.3) replacing $\varphi(\cdot)$ by $r(\cdot)$.

Assumption (A.1) is an identification condition. It is imposed over the likelihood function. Assumption (A.2) is a standard condition that allows for the interchange of the integration and differentiation operations. Assumptions (B.1)-(B.3) are needed to show that the proposed nonparametric estimator is an estimator of a least favorable curve. This is similar to the so called $N^{1/4}$-consistency condition (see Andrews, 1994). If the nonparametric estimator is a smooth function of the parametric part then the previous condition is equivalent to impose the asymptotic orthogonality condition between the parametric and the nonparametric estimators, and it is needed to show that the nonparametric estimator does not affect the asymptotic distribution of the parametric one. The estimators proposed by Klein and Spady (1993), Ichimura and Lee (1990) and Rodriguez-Póo, Sperlich, Fernández (1999) among others satisfy the so called $N^{1/4}$-consistency property. Assumption (K.1) is a standard bias reducing technique, and jointly with assumption (H.1) on the bandwidth it is needed to achieve the previous condition of $N^{1/4}$-consistency. Note finally that the bandwidth rates that are allowed in condition (H.1) are smaller than the optimal ones. This is also standard in semiparametric models and it is due to the effect that the bias of the nonparametric estimator presents in the asymptotic properties of the parametric part. Finally, assumptions (Q.1) to (Q.3) are regularity conditions needed in the quasi-likelihood framework mainly to guarantee that the quasi-likelihood function used in estimating $\eta_1, \cdots, \eta_p$ and θ has the properties of a likelihood function.

Proof of Theorem 1

The proof of this theorem is based on a generalization of Propositions 1 and 2 from Severini and Wong (1992), p. 1780. Assumptions (A.1) and (A.2) imply directly Conditions I and S from Severini and Wong (1992), pp. 1777 and 1778. Furthermore, for fixed θ, under (A.1), (A.2), (K.1) and (H.1) the estimator obtained as a solution of

$$(\hat{\eta}_{1,\theta}, \hat{\eta}_{2,\theta}, \cdots, \hat{\eta}_{p,\theta}) = \sup_{\eta_1 \in H_1, \cdots, \eta_p \in H_p} W(\eta_1, \cdots, \eta_p, \theta), \tag{12}$$

is an estimator of a least favorable curve. To see it, note that if $\hat{\eta}_\theta(x) = (\hat{\eta}_{1,\theta}(x), \cdots, \hat{\eta}_{p,\theta}(x))^T$ is the solution to (12) then

$$\sum_{i=1}^{N} \frac{\partial}{\partial \eta} \log \ell(Y_i, T_i; \hat{\eta}_{1,\theta}, \cdots, \hat{\eta}_{p,\theta}, \theta) K\left(\frac{x - X_i}{h}\right) = 0,$$

where we denote by $K\left(\frac{x-X_i}{h}\right)$ the d-variate kernel $K_1\left(\frac{x_0^1 - X_i^1}{h}\right) \times \cdots \times K_d\left(\frac{x_0^d - X_i^d}{h}\right)$. Furthermore

$$\sum_{i=1}^{N} \frac{\partial^2}{\partial \theta \partial \eta^T} \log \ell(Y_i, T_i; \hat{\eta}_{1,\theta}, \cdots, \hat{\eta}_{p,\theta}, \theta) K\left(\frac{x - X_i}{h}\right)$$
$$+ \sum_{i=1}^{N} \frac{\partial^2}{\partial \eta \partial \eta^T} \log \ell(Y_i, T_i; \hat{\eta}_{1,\theta}, \cdots, \hat{\eta}_{p,\theta}, \theta) K\left(\frac{x - X_i}{h}\right) \frac{\partial}{\partial \theta^T} \hat{\eta}_\theta(x) = 0.$$

Then, using the previous assumptions and the properties of the Watson-Naradaya smoother then

$$\frac{\partial}{\partial\theta^T}\hat{\eta}_\theta(x) \to_p -\left\{E\left[\frac{\partial^2}{\partial\eta\partial\eta^T}\varphi(Y,T;\eta_1(X^1),\cdots,\eta_p(X^p),\theta_0)|T=t,X=x\right]\right\}^{-1}$$
$$\times E\left[\frac{\partial^2}{\partial\theta\partial\eta^T}\varphi(Y,T;\eta_1(X^1),\cdots,\eta_p(X^p),\theta_0)|T=t,X=x\right]$$

and the estimator obtained in (12) is an estimator of a least favorable curve (This is Condition NP(b) from Severini and Wong, 1992; p. 1779). Condition NP(a) is obtained as follows. Let us denote

$$\hat{h}_N(\theta,\eta_1,\cdots,\eta_p,x) = \frac{G_{\eta,\theta}^{(r_\eta,s_\theta)}(x)}{\hat{f}(x)} = \frac{\frac{1}{Nh_N^d}\sum_i K\left(\frac{x-X_i}{h_N}\right)\Delta_{\eta,\theta}^{(r_\eta,s_\theta)}(Y_i,T_i)}{\frac{1}{Nh_N^d}\sum_i K\left(\frac{x-X_i}{h_N}\right)}.$$

Consider the case $r_\eta = s_\theta = 0$. Then, using the same approach as in the proof of Lemmas 5 and 8 from Severini and Wong (1992) it is possible to show that

$$\sup_{\eta_1,\cdots,\eta_p}\sup_\theta\sup_x\left|D^{t_x}G_{\eta,\theta}(x) - D^{t_x}h(\theta,\eta_1,\cdots,\eta_p,x)\right|$$
$$= O_p\left(h_N^m + N^{-\frac{q}{2(p+q+2)}}N^\gamma h_N^{-\left(|t_x|+d\left(\frac{2p+q+4}{p+q+2}\right)\right)}\right)$$

and

$$\sup_{\eta_1,\cdots,\eta_p}\sup_\theta\sup_x\left|D^{t_x}\hat{f}(x) - D^{t_x}f(x)\right| = O_p\left(h_N^m + N^{-\frac{q}{2(p+q+2)}}N^\gamma h_N^{-\left(|t_x|+d\left(\frac{2p+q+4}{p+q+2}\right)\right)}\right)$$

for some $\gamma > 0$. Use for the bandwidth the rate assumed in (H.1), then

$$\sup_{\eta_1,\cdots,\eta_p}\sup_\theta\sup_x\left|\hat{h}_N(\theta,\eta_1,\cdots,\eta_p,x) - h(\theta,\eta_1,\cdots,\eta_p,x)\right| = o_p\left(N^{-1/4}\right),$$

and

$$\sup_{\eta_1,\cdots,\eta_p}\sup_\theta\sup_x\left|D^{t_x}\hat{h}_N(\theta,\eta_1,\cdots,\eta_p,x) - D^{t_x}h(\theta,\eta_1,\cdots,\eta_p,x)\right| = o_p\left(N^{-1/4}h_N^{-|t_x|}\right).$$

The same can be done for $|r_\eta| > 0$ and $|s_\theta| > 0$ and then Conditions NP(a) from Severini and Wong (1992), pp. 1779, are verified. Since Conditions I, S and NP are verified, then Propositions 1 and 2 apply and the proof is done.

∎

Proof of Theorem 2

In order to simplify the proofs, j is fixed, and we can see that $\hat{\eta}_j$ is indeed such that

$$\widehat{\eta_j} = \arg\ \max\ W_j^*\left(\eta_j,\hat{\theta}_N\right),$$

where

$$W_j^*\left(\eta_j,\hat{\theta}_N\right)=\sum_{i=1}^{N}K_j\left(\frac{x_0^j-X_i^j}{h_j}\right)\log\ell\left(Y_i,T_i;\hat{\eta}_1(X_i^1),\cdots,\eta_j,\cdots,\hat{\eta}_p(X_i^p),\theta\right).$$

A Taylor expansion of $\varphi_j^{(1)}$ around the point $\eta_j^i=\eta_j(X_i^j)$ gives directly the existence of some $\bar{\eta}_j^i$ belonging between η_j and η_j^i and such that

$$\begin{aligned}&\varphi_j^{(1)}\left(y,t;\hat{\eta}_1^i,\cdots,\cdots,\eta_j^i,\cdots,\hat{\eta}_p^i,\hat{\theta}_N\right)=\\&\varphi_j^{(1)}\left(y,t;\hat{\eta}_1^i,\cdots,\hat{\eta}_{j-1}^i,\eta_j^i,\hat{\eta}_{j+1}^i,\cdots,\hat{\eta}_p^i,\hat{\theta}_N\right)\\&+\left(\eta_j-\eta_j^i\right)\varphi_j^{(2)}\left(y,t;\hat{\eta}_1^i,\cdots,\hat{\eta}_{j-1}^i,\bar{\eta}_j^i,\hat{\eta}_{j+1}^i,\cdots,\hat{\eta}_p^i,\hat{\theta}_N\right).\end{aligned}$$

So this leads directly to

$$\frac{\partial W_j^*\left(\eta_j,\hat{\theta}_N\right)}{\partial\eta_j}=A_1\left(\hat{\theta}_N\right)+A_2(\eta_j,\hat{\theta}_N)+\left[A_3\left(\hat{\theta}_N\right)+A_4(\eta_j,\hat{\theta}_N)\right]\left(\eta_j-\eta_j^0\right),$$

where

$$A_1\left(\hat{\theta}_N\right)=\frac{\sum_{i=1}^{N}K_j\left(\frac{x_0^j-X_i^j}{h_j}\right)\varphi_j^{(1)}\left(y,t;\hat{\eta}_1^i,\cdots,\hat{\eta}_{j-1}^i,\eta_j^i,\hat{\eta}_{j+1}^i,\cdots,\hat{\eta}_p^i,\hat{\theta}_N\right)}{\sum_{i=1}^{N}K_j\left(\frac{x_0^j-X_i^j}{h_j}\right)},$$

$$A_2(\eta_j,\hat{\theta}_N)=\frac{\sum_{i=1}^{N}K_j\left(\frac{x_0^j-X_i^j}{h_j}\right)\left(\eta_j^0-\eta_j^i\right)\varphi_j^{(2)}\left(y,t;\hat{\eta}_1^i,\cdots,\hat{\eta}_{j-1}^i,\bar{\eta}_j^i,\hat{\eta}_{j+1}^i,\cdots,\hat{\eta}_p^i,\hat{\theta}_N\right)}{\sum_{i=1}^{N}K_j\left(\frac{x_0^j-X_i^j}{h_j}\right)},$$

$$A_3\left(\hat{\theta}_N\right)=\frac{\sum_{i=1}^{N}K_j\left(\frac{x_0^j-X_i^j}{h_j}\right)\varphi_j^{(2)}\left(y,t;\hat{\eta}_1^i,\cdots,\hat{\eta}_{j-1}^i,\eta_j^i,\hat{\eta}_{j+1}^i,\cdots,\hat{\eta}_p^i,\hat{\theta}_N\right)}{\sum_{i=1}^{N}K_j\left(\frac{x_0^j-X_i^j}{h_j}\right)},$$

and

$$\begin{aligned}&A_4(\eta_j,\hat{\theta}_N)=\\&\frac{1}{\sum_{i=1}^{N}K_j\left(\frac{x_0^j-X_i^j}{h_j}\right)}\sum_{i=1}^{N}K_j\left(\frac{x_0^j-X_i^j}{h_j}\right)\left[\varphi_j^{(2)}\left(y,t;\hat{\eta}_1^i,\cdots,\hat{\eta}_{j-1}^i,\bar{\eta}_j^i,\hat{\eta}_{j+1}^i,\cdots,\hat{\eta}_p^i,\hat{\theta}_N\right)\right.\\&\left.-\varphi_j^{(2)}\left(y,t;\hat{\eta}_1^i,\cdots,\hat{\eta}_{j-1}^i,\eta_j^i,\hat{\eta}_{j+1}^i,\cdots,\hat{\eta}_p^i,\hat{\theta}_N\right)\right].\end{aligned}$$

Now, we will study the asymptotics of the previous terms recalling that under the conditions established in Theorem 1, $\sqrt{N}(\hat{\theta}_N-\theta_0)=O_p(1)$ and $\sup_{x^j\in\mathcal{X}_{d_j}}|\hat{\eta}_j(x^j)-\eta_j(x^j)|=o_p\left(N^{-1/4}\right)$ for $j=1,\cdots,p$.

For the term A_1 note that by the mean value theorem then

$$\begin{aligned}&\tfrac{1}{Nh_j^{d_j}}\textstyle\sum_{i=1}^{N}K_j\left(\frac{x_0^j-X_i^j}{h_j}\right)\varphi_j^{(1)}\left(y,t;\hat{\eta}_1^i,\cdots,\hat{\eta}_{j-1}^i,\eta_j^i,\hat{\eta}_{j+1}^i,\cdots,\hat{\eta}_p^i,\hat{\theta}_N\right)=\\&\tfrac{1}{Nh_j^{d_j}}\textstyle\sum_{i=1}^{N}K_j\left(\frac{x_0^j-X_i^j}{h_j}\right)\varphi_j^{(1)}\left(y,t;\eta_1^i,\cdots,\eta_{j-1}^i,\eta_j^i,\eta_{j+1}^i,\cdots,\eta_p^i,\theta_0\right)+\tfrac{1}{Nh_j^{d_j}}\textstyle\sum_{i=1}^{N}\\&K_j\left(\frac{x_0^j-X_i^j}{h_j}\right)\textstyle\sum_{l=1}^{p}\frac{\partial}{\partial\eta_l}\varphi_j^{(1)}\left(y,t;\bar{\eta}_1^i,\cdots,\bar{\eta}_{j-1}^i,\eta_j^i,\bar{\eta}_{j+1}^i,\cdots,\bar{\eta}_p^i,\bar{\theta}_N\right)\left(\hat{\eta}_l(X_i^l)-\eta_l(X_i^l)\right)\\&+\tfrac{1}{Nh_j^{d_j}}\textstyle\sum_{i=1}^{N}K_j\left(\frac{x_0^j-X_i^j}{h_j}\right)\frac{\partial}{\partial\theta^T}\varphi_j^{(1)}\left(y,t;\bar{\eta}_1^i,\cdots,\bar{\eta}_{j-1}^i,\eta_j^i,\bar{\eta}_{j+1}^i,\cdots,\bar{\eta}_p^i,\bar{\theta}_N\right)\left(\hat{\theta}_N-\theta_0\right).\end{aligned}$$

Since $\frac{\partial}{\partial \eta_j}\varphi_j^{(1)}$ and $\frac{\partial}{\partial \theta^T}\varphi_j^{(1)}$ are absolutely continuous, then by using the root-N consistency of $\hat{\theta}_N$, the uniform properties of the nonparametric estimators and a strong law of large numbers we obtain,

$$A_1\left(\hat{\theta}_N\right) = A_1\left(\theta_0\right) + O_p\left(\frac{h_j^2}{\sqrt{N}} + \frac{1}{Nh_j^{d_j/2}}\right) + o_p\left(N^{-1/4}\right), \tag{13}$$

where

$$A_1\left(\theta_0\right) = \frac{1}{Nh_j^{d_j}}\sum_{i=1}^{N} K_j\left(\frac{x_0^j - X_i^j}{h_j}\right)\varphi_j^{(1)}\left(y,t;\eta_1^i,\cdots,\eta_{j-1}^i,\eta_j^i,\eta_{j+1}^i,\cdots,\eta_p^i,\theta_0\right).$$

Now, since $E\left[\varphi_j^{(1)}\left(Y,T;\eta_1(X^1),\cdots,\eta_j(X^j),\cdots,\eta_p(X^p),\theta_0\right)|X^j = x_0^j\right] = 0$, standard results on Watson-Nadaraya smoothers give us

$$\begin{aligned} A_1\left(\theta_0\right) &\to_p 0 \\ E\left(A_1\left(\theta_0\right)\right) &= O(h_j^2) \\ Var\left(A_1\left(\theta_0\right)\right) &= \frac{1}{Nh_j^{d_j}}\left[\int K_j^2(t)dtI_j(\eta_j^0,\theta_0)p_j^{-1}\left(x_0^j\right)\right] + o\left(\frac{1}{nh_j^{d_j}}\right). \end{aligned} \tag{14}$$

and

$$I_j(\eta_j^0,\theta_0) = E\left[\varphi_j^{(1)}\left(Y,T;\eta_1(X^1),\cdots,\eta_j(X^j),\cdots,\eta_p(X^p),\theta_0\right)^2\middle| X^j = x_0^j\right].$$

For both next terms, with the same arguments we arrive at

$$A_2(\eta_j,\hat{\theta}_N) = A_2(\eta_j,\theta_0) + O_p\left(\frac{h_j^2}{\sqrt{N}} + \frac{1}{Nh_j^{d_j/2}}\right) + o_p\left(N^{-1/4}\right) \tag{15}$$

where

$$A_2(\eta_j,\theta_0) = \frac{\sum_{i=1}^N K_j\left(\frac{x_0^j - X_i^j}{h_j}\right)\left(\eta_j^0 - \eta_j^i\right)\varphi_j^{(2)}\left(y,t;\eta_1^i,\cdots,\eta_{j-1}^i,\eta_j^i,\eta_{j+1}^i,\cdots,\eta_p^i,\theta_0\right)}{\sum_{i=1}^N K_j\left(\frac{x_0^j - X_i^j}{h_j}\right)},$$

with

$$\begin{aligned} A_2(\eta,\theta_0) &\to_p 0 \\ E\left(A_2(\eta,\theta_0)\right) &= o(h_j^2) \\ Var\left(A_2(\eta,\theta_0)\right) &= o\left(Var(A_1\left(\theta_0\right))\right), \end{aligned} \tag{16}$$

and

$$\begin{aligned} A_3\left(\hat{\theta}_N\right) &= E\left[\varphi_j^{(2)}\left(Y,T;\eta_1(X^1),\cdots,\eta_j(X^j),\cdots,\theta_0\right)|X^j = x_0^j\right] \\ &\quad + O_p\left(\frac{1}{\sqrt{N}}\right) + o_p\left(h_j^2 + \frac{1}{\sqrt{Nh_j^{d_j}}}\right) + o_p(n^{-1/4}). \end{aligned}$$

But now, we remark that

$$\begin{aligned} & E\left[\varphi_j^{(2)}\left(Y,T;\eta_1(X^1),\cdots,\eta_j(X^j),\cdots,\theta_0\right)|X^j=x_0^j\right]= \\ & =E\left[\tfrac{\partial^2}{\partial\eta_j^2}\varphi_j\left(Y,T;\eta_1(X^1),\cdots,\eta_j(X^j),\cdots,\theta_0\right)|X^j=x_0^j\right] \\ & =-E\left[\tfrac{\partial}{\partial\eta_j}\varphi_j\left(Y,T;\eta_1(X^1),\cdots,\eta_j(X^j),\cdots,\theta_0\right)^2|X^j=x_0^j\right], \end{aligned}$$

what finally leads to

$$A_3\left(\hat{\theta}_N\right)=-I_j(\eta_j^0,\theta_0)+o_p(1). \tag{17}$$

For the term $A_4(\eta_j,\hat{\theta}_N)$, it can be dealt with by using various arguments. Indeed, the absolute continuity condition on $\varphi_j^{(2)}$ leads directly to

$$A_4(\eta_j,\hat{\theta}_N)\to 0, \tag{18}$$

in probability. This convergence is uniform over η_j and θ (since both η_j and θ belong to some compact and so the continuity of $\varphi_j^{(2)}$ is indeed uniform).
The proof of **i)** is closed as follows. Let us denote

$$Z=\sqrt{Nh_j^{d_j}}\left(\hat{\eta}_j-\eta_j^0\right). \tag{19}$$

By applying (13) at point $\eta_j=\hat{\eta}_j$, we arrive at

$$Z_j=\sqrt{Nh_j^{d_j}}\left(\frac{-A_1\left(\hat{\theta}_N\right)-A_2\left(\hat{\eta}_j,\hat{\theta}_N\right)}{A_3\left(\hat{\theta}_N\right)+A_4\left(\hat{\eta}_j,\hat{\theta}_N\right)}\right).$$

Using (15) we have that

$$\sqrt{Nh_j^{d_j}}A_2\left(\hat{\eta}_j,\hat{\theta}_N\right)=\sqrt{Nh_j^{d_j}}A_2\left(\hat{\eta}_j,\theta_0\right)+O_p\left(h_j^{2+\frac{d_j}{2}}+\frac{1}{\sqrt{N}}\right).$$

Moreover, by expression (16)

$$\sqrt{Nh_j^{d_j}}E\left[A_2\left(\hat{\eta}_j,\theta_0\right)\right]=o\left(\sqrt{Nh_j^{d_j}h_j^4}\right)$$

and finally, because of condition (C.1) on the bandwidth we have that Z has asymptotically the same distribution as

$$\sqrt{Nh_j^{d_j}}\frac{-A_1\left(\hat{\theta}_N\right)}{A_3\left(\hat{\theta}_N\right)+A_4\left(\hat{\eta}_j,\hat{\theta}_N\right)}.$$

Apply now (13), (17) and (18) and remark that thus Z has the same distribution as $\sqrt{Nh_j^{d_j}}\frac{A_1(\theta_0)}{I_j(\eta_j^0,\theta_0)}$.
On the other hand the Lindeberg-Feller theorem together with (14) leads to

$$\sqrt{Nh_j^{d_j}}A_1\left(\theta_0\right)\longrightarrow_d N\left(0,\sqrt{\frac{\int K_j^2(t)dt}{p_j(x_0^j)I_j(\eta_j^0,\theta_0)}}\right). \tag{20}$$

In order to close the proof of the first part of the theorem, note that we have by continuity of the function $V_j(\eta_j)$ and because of Theorem 2 (ii) we have

$$\frac{V_j\left(\hat{\eta}_j\right)}{V_j\left(\eta_j^0\right)} \longrightarrow_p 1. \tag{21}$$

Finally, because of Slutsky's theorem, (21) and (20) are enough to prove the result of Theorem 2 (i).

In order to show **ii)**, if in place of using the Lindeberg-Feller theorem as we did to prove **i)**, we use Bernstein's type inequality (see Serfling, 1980; p. 95) we get immediately for $A_1(\theta_0)$ the following expression

$$A_1(\theta_0) = O_p\left(\sqrt{\frac{logN}{Nh_j^{d_j}}}\right).$$

Writing now S in the form

$$S = \left(\frac{-A_1\left(\hat{\theta}_N\right) - A_2\left(\hat{\eta}_j, \hat{\theta}_N\right)}{A_3\left(\hat{\theta}_N\right) + A_4\left(\hat{\eta}_j, \hat{\theta}_N\right)}\right),$$

and using (15), (16), (17) and (18) to treat the terms A_2, A_3 and A_4, we get directly

$$S = O_p\left(\sqrt{\frac{logN}{Nh_j^{d_j}}}\right). \tag{22}$$

Finally, (19) and (22) are enough to finish the proof of part **ii)** of the theorem.

■

Proof of Theorem 3

The proof of this result is again based on Propositions 1 and 2 from Severini and Wong (1992). Assumptions (Q.1)-(Q.3) are regularity conditions that imply Conditions I and S from Severini and Wong (1992), pp. 1777 and 1778. This can be shown following the same lines as Severini and Staniswalis (1994), p. 511. To finish the proof we need to show that conditions NP are verified. In order to do so, we will show that the estimator obtained as a solution of

$$\left(\hat{\eta}_{1,\theta}, \hat{\eta}_{2,\theta}, \cdots, \hat{\eta}_{p,\theta}\right) = \sup_{\eta_1 \in H_1, \cdots, \eta_p \in H_p} W^*\left(\eta_1, \cdots, \eta_p, \theta\right),$$

where

$$W^*\left(\eta_1, \cdots, \eta_p, \theta\right) =$$
$$\sum_{i=1}^N K_1\left(\frac{x_0^1 - X_i^1}{h_1}\right) \times \cdots \times K_d\left(\frac{x_0^d - X_i^d}{h_d}\right) r\left(Y_i, g\left(T_i, \eta_1, \cdots, \eta_p, \theta\right)\right),$$

for fixed θ is an estimator of a least favorable curve. To see it, note that if $\hat{\eta}_\theta(x) = (\hat{\eta}_{1,\theta}(x), \cdots, \hat{\eta}_{p,\theta}(x))^T$ is the solution to the previous maximization problem then

$$\sum_{i=1}^{N} \frac{Y_i - g(T_i; \hat{\eta}_{1,\theta}, \cdots, \hat{\eta}_{p,\theta}, \theta)}{V\left(g(T_i; \hat{\eta}_{1,\theta}, \cdots, \hat{\eta}_{p,\theta}, \theta)\right)} \frac{\partial}{\partial \eta} g(T_i; \hat{\eta}_{1,\theta}, \cdots, \hat{\eta}_{p,\theta}, \theta) K\left(\frac{x - X_i}{h}\right) = 0,$$

where we denote by $K\left(\frac{x-X_i}{h}\right)$ the d-variate kernel $K_1\left(\frac{x_0^1 - X_i^1}{h}\right) \times \cdots \times K_d\left(\frac{x_0^d - X_i^d}{h}\right)$. Furthermore

$$\sum_{i=1}^{N} R_{1i}\left(T_i; \hat{\eta}_\theta, \theta\right) K\left(\frac{x - X_i}{h}\right) + \sum_{i=1}^{N} R_{2i}\left(T_i; \hat{\eta}_\theta, \theta\right) K\left(\frac{x - X_i}{h}\right) \frac{\partial}{\partial \theta^T} \hat{\eta}_\theta(x) = 0.$$

where

$$R_{1i}\left(T_i; \hat{\eta}_\theta, \theta\right) = \left(\frac{1}{V(g\left(T_i; \hat{\eta}_\theta, \theta\right))} + \frac{Y_i - g\left(T_i; \hat{\eta}_\theta, \theta\right)}{V(g\left(T_i; \hat{\eta}_\theta, \theta\right))^2}\right) \frac{\partial}{\partial \theta} g\left(T_i; \hat{\eta}_\theta, \theta\right) \frac{\partial}{\partial \eta^T} g\left(T_i; \hat{\eta}_\theta, \theta\right)$$
$$- \frac{Y_i - g\left(T_i; \hat{\eta}_\theta, \theta\right)}{V\left(g\left(T_i; \hat{\eta}_\theta, \theta\right)\right)} \frac{\partial^2}{\partial \theta \eta^T} g\left(T_i; \hat{\eta}_\theta, \theta\right)$$

$$R_{2i}\left(T_i; \hat{\eta}_\theta, \theta\right) = \left(\frac{1}{V((T_i; \hat{\eta}_\theta, \theta))} + \frac{Y_i - g\left(T_i; \hat{\eta}_\theta, \theta\right)}{V(g\left(T_i; \hat{\eta}_\theta, \theta\right))^2}\right) \frac{\partial}{\partial \eta} g\left(T_i; \hat{\eta}_\theta, \theta\right) \frac{\partial}{\partial \eta^T} g\left(T_i; \hat{\eta}_\theta, \theta\right)$$
$$- \frac{Y_i - g\left(T_i; \hat{\eta}_\theta, \theta\right)}{V\left(g\left(T_i; \hat{\eta}_\theta, \theta\right)\right)} \frac{\partial^2}{\partial \eta \eta^T} g\left(T_i; \eta, \hat{\eta}_\theta\right)$$

Then, using the previous assumptions and the properties of the Watson-Naradaya smoother then

$$\frac{\partial}{\partial \theta^T} \hat{\eta}_\theta(x) \rightarrow_p$$
$$\left\{E\left[\frac{\partial}{\partial \eta} g\left(T; \eta_1(X^1), \cdots, \eta_p(X^p), \theta_0\right) \frac{\partial}{\partial \eta^T} g\left(T; \eta_1(X^1), \cdots, \eta_p(X^p), \theta_0\right) | T = t, X = x\right]\right\}^{-1}$$
$$\times E\left[\frac{\partial}{\partial \theta} g\left(T; \eta_1(X^1), \cdots, \eta_p(X^p), \theta_0\right) \frac{\partial}{\partial \eta^T} g\left(T; \eta_1(X^1), \cdots, \eta_p(X^p), \theta_0\right) | T = t, X = x\right]$$

Therefore (see Lemma 1, Severini and Wong, 1992; p.1778) $\frac{\partial}{\partial \theta^T} \hat{\eta}_\theta(x)$ is an estimator of a least favorable curve (This is Condition NP(b) from Severini and Wong, 1992; p. 1779). Condition NP(a) is shown in the same way as in the proof of Theorem 1, and therefore the result is shown.

∎

Finally, the proof of Theorem 4 follows the same lines as in the proof of Theorem 2 by replacing the log-likelihood by the quasi-likelihood function.

Appendix II

Computational Remarks: The Newton-Raphson algorithm for Truncated Variable Model

First, remember the fully parametric case. The Maximum Likelihood Function for the truncated variable (Tobit 1) case is

$$\mathcal{L} = \sum_i -\frac{\ln 2\pi}{2} - \frac{\ln \sigma^2}{2} - \frac{\{y_i - \gamma^T x_i\}^2}{2\sigma^2} - \ln\left\{1 - F\left(\frac{-\gamma^T x_i}{\sigma}\right)\right\}$$

where the sum $\sum_i$ only runs over observations $Y_i > 0$.

The derivatives are as follows:

$$\begin{aligned}
\frac{\partial \mathcal{L}}{\partial \gamma} &= \frac{1}{\sigma}\sum_{i=1}^{n}\left[\frac{y_i - \gamma^T x_i}{\sigma} - \frac{f(\frac{-\gamma^T x_i}{\sigma})}{1 - F(\frac{-\gamma^T x_i}{\sigma})}\right] x_i \\
\frac{\partial^2 \mathcal{L}}{\partial \gamma^2} &= \frac{-1}{\sigma^2}\sum_{i=1}^{n}\left[1 - \frac{\gamma^T x_i \sigma^{-1} f(-\frac{\gamma^T x_i}{\sigma})}{1 - F(\frac{-\gamma^T x_i}{\sigma})} - \frac{f^2(\frac{-\gamma^T x_i}{\sigma})}{\{1 - F(\frac{-\gamma^T x_i}{\sigma})\}^2}\right] x_i {x_i}^T.
\end{aligned}$$

Some more calculation is needed to get $\partial\mathcal{L}/\partial\sigma$ and $\partial^2\mathcal{L}/\partial\sigma^2$:

$$\begin{aligned}
\frac{\partial \mathcal{L}}{\partial \sigma} &= \frac{-1}{\sigma}\sum_{i=1}^{n}\left[1 - \left(\frac{y_i - \gamma^T x_i}{\sigma}\right)^2 - \frac{\gamma^T x_i \sigma^{-1} f(\frac{-\gamma^T x_i}{\sigma})}{1 - F(\frac{-\gamma^T x_i}{\sigma})}\right] \\
\frac{\partial^2 \mathcal{L}}{\partial \sigma^2} &= \frac{-1}{\sigma^2}\sum_{i=1}^{n}\left[-1 + \frac{3(y_i - \gamma^T x_i)^2}{\sigma^2} + \frac{2}{\sigma}\left\{\frac{\gamma^T x_i f(\frac{-\gamma^T x_i}{\sigma})}{1 - F(\frac{-\gamma^T x_i}{\sigma})}\right\}\right. \\
&\quad \left. + \frac{f(\frac{-\gamma^T x_i}{\sigma})(-\gamma^T x_i)^3 \sigma^{-3}}{\{1 - F(\frac{-\gamma^T x_i}{\sigma})\}} - \frac{f^2(\frac{-\gamma^T x_i}{\sigma})(-\gamma^T x_i)^2 \sigma^{-2}}{\{1 - F(\frac{-\gamma^T x_i}{\sigma})\}^2}\right]
\end{aligned}$$

Finally, the mixed derivative

$$\frac{\partial^2 \mathcal{L}}{\partial \gamma \partial \sigma} = \frac{-1}{\sigma^2}\sum_i\left[\frac{2(y_i - \gamma^T x_i)}{\sigma} + \frac{\left\{\left(\frac{-\gamma^T x_i}{\sigma}\right)^2 - 1\right\} f(\frac{-\gamma^T x_i}{\sigma})}{\{1 - F(\frac{-\gamma^T x_i}{\sigma})\}} + \frac{\gamma^T x_i f^2(\frac{-\gamma^T x_i}{\sigma})}{\sigma\{1 - F(\frac{-\gamma^T x_i}{\sigma})\}^2}\right] x_i$$

Semiparametric with multivariate nonparametric part

We write down first the expressions for a multivariate nonparametric part $\eta(x)$. To derive the expressions for the components η_j is straight forward but certainly depend on the particular model specification. Set as before $K_h = \prod_j K_{h_j}$ and $\eta^i = \eta(x_i)$. Let x be the continuous, t

the discrete variables. The Maximum Likelihood Function for the truncated variable (Tobit 1) case is

$$\mathcal{L} = \sum_i -\frac{\ln 2\pi}{2} - \frac{\ln \sigma^2}{2} - \frac{\{y_i - \gamma^T t_i - \eta^i\}^2}{2\sigma^2} - \ln\left\{1 - F\left(\frac{-\gamma^T t_i - \eta^i}{\sigma}\right)\right\}$$

The local, smoothed Likelihood for η^j is

$$\mathcal{L}^S = \sum_i \left[-\frac{\ln 2\pi}{2} - \frac{\ln \sigma^2}{2} - \frac{\{y_i - \gamma^T t_i - \eta^j\}^2}{2\sigma^2} - \ln\left\{1 - F\left(\frac{-\gamma^T t_i - \eta^j}{\sigma}\right)\right\}\right] K_h(x_i - x_j)$$

where F is the cumulated standard normal probability function, and f we will use for the standard normal density. Note that for $p = 1$, this $\mathcal{L}^S$ is equivalent to the smoothed likelihood we called $W(\cdot)$ in Section 3. For the ease of notation set $u_{kl} = (-\gamma^T t_k - \eta^l)\sigma^{-1}$

We start with calculating $\partial\mathcal{L}^S/\partial\eta^j$ and $\partial^2\mathcal{L}^S/\partial(\eta^j)^2$.

$$\frac{\partial \mathcal{L}^S}{\partial \eta_j} = \frac{1}{\sigma}\sum_{i=1}^n \left[\frac{Y_i}{\sigma} + u_{ij} - \frac{f(u_{ij})}{1 - F(u_{ij})}\right] K_h(x_j - x_i), \tag{23}$$

$$\frac{\partial^2 \mathcal{L}^S}{\partial(\eta^j)^2} = \frac{-1}{\sigma^2}\sum_{i=1}^n \left[1 + \frac{u_{ij} f(u_{ij})}{1 - F(u_{ij})} - \frac{f^2(u_{ij})}{\{1 - F(u_{ij})\}^2}\right] K_h(x_j - x_i). \tag{24}$$

Next, we calculate $\partial\mathcal{L}/\partial\gamma$ and $\partial^2\mathcal{L}/\partial\gamma^2$. Set $\tilde{t}_{ij} = t_i + \delta\eta(x_j)/\delta\gamma$, then

$$\frac{\partial \mathcal{L}}{\partial \gamma} = \frac{1}{\sigma}\sum_{i=1}^n \left[\frac{y_i}{\sigma} + u_{ii} - \frac{f(u_{ii})}{1 - F(u_{ii})}\right] \tilde{t}_{ii}. \tag{25}$$

For the Hessian matrix we neglect the dependency of $\tilde{t}_{ii}$ on γ and get

$$\frac{\partial^2 \mathcal{L}}{\partial \gamma^2} = \frac{-1}{\sigma^2}\sum_{i=1}^n \left[1 + \frac{f(u_{ii})u_{ii}}{1 - F(u_{ii})} - \frac{f^2(u_{ii})}{\{1 - F(u_{ii})\}^2}\right] \tilde{t}_{ii}\tilde{t}_{ii}^T.$$

A little bit more complicated is to get $\partial\mathcal{L}/\partial\sigma$ and $\partial^2\mathcal{L}/\partial\sigma^2$; set $\eta_\sigma^j = \partial\eta^j/\partial\sigma$:

$$\frac{\partial \mathcal{L}}{\partial \sigma} = \frac{-1}{\sigma}\sum_{i=1}^n \left[1 - \left(\frac{y_i}{\sigma} + u_{ii}\right)^2 - \left\{\frac{y_i}{\sigma} + u_{ii}\right\}\eta_i^\sigma + \frac{f(u_{ii})(u_{ii} + \eta_\sigma^i)}{1 - F(u_{ii})}\right]. \tag{26}$$

For the Hessian matrix we again neglect the dependency of η_σ^j on σ and so get with $B_n(u_{ii})$ from (26) and get

$$\begin{aligned}\frac{\partial^2 \mathcal{L}}{\partial \sigma^2} = \frac{-1}{\sigma^2}\sum_{i=1}^n \Bigg[&-1 + 3\left\{\frac{y_i}{\sigma} + u_{ii}\right\}^2 + 4\left\{\frac{y_i}{\sigma} + u_{ii}\right\}\eta_\sigma^i + \left(\eta_\sigma^i\right)^2 \\ &- \frac{2f(u_{ii})(u_{ii} + \eta_\sigma^i)}{1 - F(u_{ii})} + \frac{f(u_{ii})u_{ii}(u_{ii} + \eta_\sigma^i)^2}{1 - F(u_{ii})} - \frac{f^2(u_{ii})(u_{ii} + \eta_\sigma^i)^2}{\{1 - F(u_{ii})\}^2}\Bigg],\end{aligned}$$

The question is how to get $\eta_\gamma^j = \partial\eta^j/\partial\gamma$. For the likelihood maximizing η^j expression (23) is equal to zero. First we derive it with respect to γ:

$$\begin{aligned}\frac{\partial^2\mathcal{L}^S}{\partial\eta^j\partial\gamma} &= -\sigma^{-2}\sum_{i=1}^{n}\left[1+\frac{f(u_{ij})u_{ij}}{1-F(u_{ij})}-\left\{\frac{f(u_{ij})}{1-F(u_{ij})}\right\}^2\right]K_h(x_j-x_i)\tilde{t}_{ij}=0\\ \Longleftrightarrow\ \eta_\gamma^j &= \frac{-\sum_{i=1}^{n}\left[1+\frac{f(u_{ij})u_{ij}}{1-F(u_{ij})}-\frac{f^2(u_{ij})}{\{1-F(u_{ij})\}^2}\right]K_h(x_j-x_i)t_i}{\sum_{i=1}^{n}\left[1+\frac{f(u_{ij})u_{ij}}{1-F(u_{ij})}-\frac{f^2(u_{ij})}{\{1-F(u_{ij})\}^2}\right]K_h(x_j-x_i)}.\end{aligned}$$

Second we derive this expression with respect to σ:

$$\begin{aligned}&\frac{\partial^2\mathcal{L}^S}{\partial\eta_j\partial\sigma}=\\ &\tfrac{-1}{\sigma^2}\textstyle\sum_{i=1}^{n}\left[\eta_\sigma^j+2\left(\frac{Y_i}{\sigma}+u_{ij}\right)+\frac{f(u_{ij})\{u_{ij}^2+u_{ij}\eta_\sigma^j-1\}}{1-F(u_{ij})}-\frac{f^2(u_{ij})\{u_{ij}+\eta_\sigma^j\}}{\{1-F(u_{ij})\}^2}\right]K_h(x_j-x_i),\\ &\Longleftrightarrow\ \eta_\sigma^j=\frac{-\sum_{i=1}^{n}\left[2\left(\frac{y_i}{\sigma}+u_{ij}\right)+\frac{f(u_{ij})\{u_{ij}^2-1\}}{1-F(u_{ij})}-\frac{f^2(u_{ij})u_{ij}}{\{1-F(u_{ij})\}^2}\right]K_h(x_j-x_i)}{\sum_{i=1}^{n}\left[1+\frac{f(u_{ij})u_{ij}}{1-F(u_{ij})}-\frac{f^2(u_{ij})}{\{1-F(u_{ij})\}^2}\right]K_h(x_j-x_i)}.\end{aligned}$$

Finally we need the mixed derivatives $\partial\mathcal{L}/\partial\gamma\partial\sigma$, respectively $\partial\mathcal{L}/\partial\sigma\partial\gamma$.

$$\frac{\partial\mathcal{L}}{\partial\gamma\partial\sigma}=\frac{-1}{\sigma^2}\sum_{i=1}^{n}\left[\eta_\sigma^i+2\left(\frac{y_i}{\sigma}+u_{ii}\right)+\frac{f(u_{ii})\{u_{ii}^2+u_{ii}\eta_\sigma^i-1\}}{1-F(u_{ii})}-\frac{f^2(u_{ij})(\eta_\sigma^i+u_{ii})}{\{1-F(u_{ii})\}^2}\right]\tilde{t}_{ii}$$

Here, we have neglected the dependency of η_σ^i on γ and the dependency of η_γ^i on σ.

The Hessian matrix for $\mathcal{L}^S$ is simply given by (24) and the one for $\mathcal{L}$ is given by

$$H_{\mathcal{L}}=\begin{pmatrix}\frac{\partial^2\mathcal{L}}{\partial\gamma^2} & \frac{\partial^2\mathcal{L}}{\partial\gamma\partial\sigma}\\ \frac{\partial^2\mathcal{L}}{\partial\sigma\partial\gamma} & \frac{\partial^2\mathcal{L}}{\partial\sigma^2}\end{pmatrix}.$$

For nonparametric functions when parameters are known

We first have to specify the considered model structure: we consider the two models

$$\eta(t)=\eta_1(x_1)+\eta_2(x_2)+\eta_3(x_3)+\eta_4(x_4)+\eta_5(x_5)+\eta_6(x_6)$$

which is the pure additive one, and

$$\eta(t)=\eta_1(x_1)+\eta_2(x_2)+\eta_3(x_3)+\eta_4(x_4)+\eta_5(x_5)\{1+\eta_6(x_6)\}$$

both with the identification conditions $E[\eta_j(X_j)]=0$ for $j=1,2,3,4,6$, i.e. except for η_5. Consequently, a possibly existing constant is going into η_5 for the additive and into $\eta_5(1+\eta_6)$ for the other case.

Let us only consider the second model and calculate first the derivatives for the components. As above, set $\eta^{ij} = \eta(x_{j1}, x_{i\underline{1}})$ where $x_{i\underline{1}}$ means all dimensions of vector x_i except the first and $v_{ij} = \frac{-\gamma^T t_i - \eta^{ij}}{\sigma}$:

$$\begin{aligned}
\frac{\partial \mathcal{L}_j^S}{\partial \eta_1^j} &= \frac{1}{\sigma} \sum_i \left\{ \frac{y_i}{\sigma} + v_{ij} - \frac{f(v_{ij})}{1 - F(v_{ij})} \right\} K(x_{j1} - x_{i1}) \\
\frac{\partial^2 \mathcal{L}_j^S}{\partial (\eta_1^j)^2} &= \frac{-1}{\sigma^2} \sum_i \left[1 + \frac{v_{ij} f(v_{ij})}{1 - F(v_{ij})} - \frac{f^2(v_{ij})}{\{1 - F(v_{ij})\}^2} \right] K(x_{j1} - x_{i1})
\end{aligned}$$

Certainly, for η_2, η_3, η_4 we get the same. Note, that, following straightly the notation of the proof of Theorem 2 you should take $v_{ij} = \frac{-\gamma^T t_i - \eta^{jj}}{\sigma}$, but intuitively it is clear, this was confirmed by simulations, that the numerical performance is much better when letting run the nuisance components over i.

For η_5 it is only slightly different, resulting in

$$\begin{aligned}
\frac{\partial \mathcal{L}_5^S}{\partial \eta_5^j} &= \frac{1}{\sigma} \sum_i \left\{ \frac{y_i}{\sigma} + v_{ij} - \frac{f(v_{ij})}{1 - F(v_{ij})} \right\} (1 + \eta_6^i) K(x_{j5} - x_{i5}) \\
\frac{\partial^2 \mathcal{L}_5^S}{\partial (\eta_5^j)^2} &= \frac{-1}{\sigma^2} \sum_i \left[1 + \frac{v_{ij} f(v_{ij})}{1 - F(v_{ij})} - \frac{f^2(v_{ij})}{\{1 - F(v_{ij})\}^2} \right] (1 + \eta_6^i)^2 K(x_{j5} - x_{i5})
\end{aligned}$$

where still $v_{ij} = \frac{-\gamma^T t_i - \eta^{ij}}{\sigma}$ but now $\eta^{ij} = \eta(x_{j5}, x_{i\underline{5}})$ where $x_{i\underline{5}}$ are all dimensions of vector x_i except the fifth.

Much more difficult for η_6 with notation $v_{ij} = \frac{-\gamma^T x_i - \eta^{ij}}{\sigma}$ where now $\eta^{ij} = \eta(x_{j6}, x_{i\underline{6}})$ and $x_{i\underline{6}}$ are all dimensions of vector x_i except the fourth:

$$\begin{aligned}
\frac{\partial \mathcal{L}_6^S}{\partial \eta_6^j} &= \frac{1}{\sigma} \sum_i \eta_5^i \left[\left\{ \frac{y_i}{\sigma} + v_{ij} \right\} - \frac{f(v_{ij})}{\{1 - F(v_{ij})\}} \right] K(x_{j6} - x_{i6}) \\
\frac{\partial^2 \mathcal{L}_6^S}{\partial (\eta_6^j)^2} &= \frac{-1}{\sigma^2} \sum_i (\eta_5^i)^2 \left[1 + \frac{v_{ij} f(v_{ij})}{1 - F(v_{ij})} - \frac{f^2(v_{ij})}{\{1 - F(v_{ij})\}^2} \right] K(x_{j6} - x_{i6})
\end{aligned}$$

References

Ai, C. and X.Chen (1999) Efficient Estimation of Models with Conditional Moment Restrictions Containing Unknown Functions. *Under revision in Econometrica*

AMEMIYA, T.(1985) *Advanced Econometrics.* Blackwell Pu. Co.

ANDREWS, D.W.K. (1994) Asymptotics for semiparametric econometric models via stochastic equicontinuity. *Econometrica*, **62**: 43-72.

ANDREWS, D.W.K. AND Y.-J.WHANG (1990) Additive interactive regression models: circumvention of the curse of dimensionality. *Econometric Theory*, **6**: 466-479.

BERNDT, E.R. AND L.R.CHRISTENSEN (1973) The internal Structure of Functional Relationships: Separability, Substitution, and Aggregation. *Review of Economic Studies,* **40:** 403-410.

BLUNDELL, R. AND J.-M.ROBIN (2000) Latent Separability: Grouping Goods without Weak Separability. *Econometrica,* **68:** 53-84.

BURDA, M., W.HÄRDLE, M.MÜLLER AND A.WERWATZ (1998) Semiparametric Analysis of German East-West Migration Intentions: Facts and Theory. *Journal of Applied Econometrics,* **13:** 525-541.

CHAMBERLAIN, G. (1992) Efficiency bounds for semiparametric regression. *Econometrica,* **60**: 567-596.

DEATON, A. AND J.MUELLBAUER (1980) *Economics and Consumer Behavior.* Cambridge University Press: Cambridge.

DELGADO, M.A. AND J.MORA (1995) Nonparametric and Semiparametric Estimation with Discrete Regressors. *Econometrica,* **63**: 1477-1484

DENNY, M. AND M.FUSS (1977) The Use of Approximation Analysis to Test for Separability and the Existence of Consistent Aggregates. *The American Economic Review,* **67:** 404-418.

DIEWERT, W.E. AND T.J.WALES (1997) Flexible Functional Forms and Tests for Heterogeneous Separability. *Journal of Econometrics,* **67:** 259-302.

FERNÁNDEZ, A.I. AND J.M.RODRÍGUEZ-POÓ (1997) Estimation and Specifications Testing in Female Labor Participation Models: Parametric and Semiparametric Methods. *Econometric Reviews,* **16(2):** 229-248.

FUSS, M., D.MCFADDEN AND Y.MUNDLAK (1978) A survey of functional forms in the economic analysis of production. *In Fuss, M. and McFadden, D. (eds.),* Production Economics: A Dual Approach to Theory and Applications, **1:** 219-268.

GOLDMAN, S.M. AND H.UZAWA (1964) A Note on Separability in Demand Analysis. *Econometrica,* **32:** 387-398.

HASTIE, T.J. AND R.J.TIBSHIRANI (1990) Generalized Additive Models. *Chapman and Hall: London.*

HÄRDLE, W., E.MAMMEN AND I.PROENCA (2000) A Bootstrap Test for Single Index Models. *Preprint, University Heidelberg, Germany*

HÄRDLE, W., S.HUET, E.MAMMEN AND S.SPERLICH (1999) Semiparametric additive indices for binary response and generalized additive models. *Preprint, Carlos III de Madrid, Spain*

HÄRDLE, W. , S.SPERLICH AND V.SPOKOINY (1997) Component Analysis for Additive Models. *Discussion Paper 52, SFB 373 Berlin, Germany*

HECKMAN, J. (1979) Sample Selection Bias as a Specification Error. *Econometrica,* **47**: 153-161.

HECKMAN, J. AND B.SINGER (1984) A Method for Minimizing the Impact of Distributional Assumptions in Econometric Models for Duration Data. *Econometrica,* **52:** 271-320.

HOLST, E. AND J.SCHUPP (1991) Frauenerwerbstätigkeit in den neuen und alten Bundesländern - Befunde des Sozio-ökonomischen Panels. *Discussion Paper 37, DIW, Berlin*

HOLST, E. AND J.SCHUPP (1994) Erwerbsbeteiligung und Erwerbsorientierung von Frauen in West- Ostdeutschland 1990-1993. *Discussion Paper 90, DIW, Berlin*

HOROWITZ, J. (1999) Nonparametric Estimation of a Generalized Additive Model with an Unknown Link Function. *Preprint, Iowa State University*

ICHIMURA, H. AND L. F. LEE (1990) Semiparametric estimation of multiple index models, *Nonparametric and Semiparametric Methods in Econometrics and Statistics*, ed. by W. A. Barnett, J.L. Powell, and G. Tauchen. New York: Cambridge University Press.

KEMPE, W. (1997) Das Arbeitsangebot verheirateter Frauen in den neuen und alten Bundesländern - Eine semiparametrische Regressionsanalyse. *Discussion Paper 3, SFB 373 Berlin, Germany*

KLEIN, R.W. AND R. H. SPADY (1993) An efficient semiparametric estimator for discrete choice models. *Econometrica,* **61:** 387-421.

LEONTIEF, W. (1947a) Introduction to a theory of the internal structure of functional relationships. *Econometrica,* **15:** 361-373.

LEONTIEF, W. (1947b) A note to the interrelation of subsets of independent variables of a continuous function with continuous first derivatives. *Bulletin of the American Mathematical Society,* **53:** 343-350.

LEWBEL, A. AND O.LINTON (2000) Nonparametric Censored and Truncated Regression. *Preprint, London School of Economics*

LINTON, O.B. AND NIELSEN, J.P. (1995) A kernel method of estimating structured nonparametric regression based on marginal integration. *Biometrika,* **82**: 93-101.

MAMMEN, E., O.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.

McCULLAGH, P. AND J.A.NELDER (1989) *Generalized Linear Models*, London: Chapman and Hall.

MERZ, J. (1990) Female Labor Supply: Labor Force Participation, Market Wage Rate and Working Hours of Married and Unmarried Woman in the Federal Republic of Germany. *Jahrbücher für Nationalökonomie und Statistik,* **207:** 240-270.

NEWEY, W.K. (1990) Semiparametric Efficiency Bounds, *Journal of Applied Econometrics*, **5**: 99-135.

NEWEY, W.K. (1994) The Asymptotic Variance of Semiparametric Estimators. *Econometrica*, **62**: 1349-1382.

NEWEY, W.K. (1995) Convergence rates for series estimators. *In Maddala, G.S., Phillips, P.C.B., Srinavsan, T.N. (eds.)*, Statistical Methods of Economics and Quantitative Economics: Essays in Honor of C.R. Rao. Blackwell, Cambridge, 254-275.

NEWEY, W.K., J.L.POWELL AND F.VELLA (1999) Nonparametric Estimation of Triangular Simultaneous Equations Models. *Econometrica*, **67:** 565-604.

NIELSEN, J.P. AND O.B.LINTON (1997) An optimization interpretation of integration and backfitting estimators for separable nonparametric models. *Journal of the Royal Statistical Society, Series B,* **60:** 217-222.

PINSKE, J. (2000) Feasible Multivariate Nonparametric Regression Estimation Using Weak Separability. *Preprint, University of British Columbia, Canada*

RODRIGUEZ-PÓO, J.M., S.SPERLICH AND A.I.FERNÁNDEZ (1999) Semiparametric Three Step Estimation Methods in Labor Supply Models. *Working Paper 99-83 (32), Carlos III de Madrid, Spain*

SERFLING, R.J. (1980) *Approximation Theorems of Mathematical Statistics.* Wiley.

SEVERINI, T.A. AND J.G.STANISWALIS (1994) Quasi-likelihood estimation in semiparametric models. *Journal of the American Statistical Association* **89:** 501-511.

SEVERINI, T.A. AND W.W.WONG (1992) Profile Likelihood and Conditionally Parametric Models. *The Annals of Statistics,* **4:** 1768-1802.

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.

SPERLICH, S., D.TJØSTHEIM AND L.YANG (1999) Nonparametric Estimation and Testing of Interaction in Additive Models. *Working Paper 99-85 (34), Carlos III de Madrid*

STANISWALIS, J.G. (1989) The kernel estimate of a regression function in likelihood-based models. *Journal of the American Statistical Association*, **84:** 276-283.

STONE, C.J. (1985) Additive regression and other nonparametric models. *Annals of Statistics,* **13:** 689-705.

STONE, C.J. (1986) The dimensionality reduction principle for generalized additive models. *Annals of Statistics,* **14:** 590-606.

TJØSTHEIM, D. AND B.H.AUESTAD (1994) Nonparametric identification of nonlinear time series: projections. *Journal of the American Statistical Association,* **89:** 1398-1409.

VIEU, P. (1991) Quadratic Errors for Nonparametric Estimators under Dependence. *Journal of Multivariate Analysis,* **93:** 324-347.