Econometric Theory, **19**, 2003, 1008–1039. Printed in the United States of America.
DOI: 10.1017/S0266466603196065

SEMIPARAMETRIC ESTIMATION OF SEPARABLE MODELS WITH POSSIBLY LIMITED DEPENDENT VARIABLES

Juan M. Rodríguez-Póo
Universidad de Cantabria

Stefan Sperlich
Universidad Carlos III de Madrid

Philippe Vieu
Université Paul Sabatier

In this paper we introduce a general method for estimating semiparametrically the different components in separable models. The family of separable models is quite popular in economic research because this structure offers clear interpretation, has straightforward economic consequences, and is often justified by theory. This family is also of statistical interest because it allows us to estimate high-dimensional complexity semiparametrically without running into the curse of dimensionality. We consider even the case when multiple indices appear in the objective function; thus we can estimate models that are typical in economic analysis, such as those that contain limited dependent variables. The idea of the new method is mainly based on a generalized profile likelihood approach. Although this requires some hypotheses on the conditional error distribution, it yields a quite general usable method with low computational costs but high accuracy even for small samples. We give estimation procedures and provide some asymptotic theory. Implementation is discussed; simulations and an application demonstrate its feasibility and good finite-sample behavior.

1. INTRODUCTION

Separability plays an extremely important role in economics and econometrics. As long ago as 1947 Leontief (1947a, 1947b) introduced definitions, interpre-

This research was financially supported by Dirección General de Investigación del Ministerio de Ciencia y Tecnología under research grants BEC2001-1121 and BEC2001-1270; Dirección General de Enseñanza Superior del Ministerio de Educación y Ciencia under Subprograma de estancias de investigadores españoles en centros de investigación españoles y extranjeros, ref. PR2000-0096; and by the Danish Social Science Research fund. We also thank M. Delgado, O. Linton, two anonymous referees, and all participants of the working group STAPH on functional statistics in Toulouse, the activities of which are available on line at http://www.lsp.ups-tlse.fr/Fp/Ferraty/staph.html. Address correspondence to: Juan M. Rodríguez-Póo, Departamento de Economia, Universidad de Cantabria, Avda. de los Castros s/n, 39005 Santander, Spain; e-mail: rodrigjm@unican.es.

0266-4666/03 $12.00

tations, and consequences for different levels of separability. In general, separability is motivated by the idea of two-stage decision making: Let $G(x)$ be a utility, demand, or production function with $x \in R^d$. Imagine that the regressors can be partitioned into aggregates or groups $x^j \subset x$ so that preferences within them can be described independently of the quantities of the others. Then we have subutility functions $\eta_j(x^j)$, and we can write $G(x) = F(\eta_1, \eta_2, \dots, \eta_p)$. Thus the decision is made first for each x^j and then for the resulting set of η_j via $F(\cdot)$. In household expenditure decisions, for example, we imagine first a budgeting for groups such as food, shelter, and entertainment and in the second step budgeting within each group. Alternatively, in production processes separability is characterized by the independence of the marginal rate of substitution between a group of inputs from changes in the level of another input: $(\partial/\partial x_k)(g^{(i)}/g^{(j)}) = 0$ or $g^{(j)}g^{(i,k)} - g^{(i)}g^{(j,k)} = 0$ with $g^{(i)} = \partial G/\partial x_i$, $g^{(i,k)} = \partial^2 G/\partial x_i \partial x_k$, $i, j, k = 1, \dots, d$. One speaks of *weak* separability when x_i, x_j are from the same subset of inputs but x_k is from a different one. Strong separability is given when x_i, x_j may also be from different subsets. The subsets are thought to be mutually exclusive and exhaustive. Regarding the consequences for the functional form of G, Goldman and Uzawa (1964) show that strong separability is equivalent to additivity $G(x) = F(\eta_1 + \cdots + \eta_p)$ whereas weak separability is equivalent to $G(x) = F(\eta_1, \dots, \eta_p)$, where η_s is a function of the elements x_k, $k = 1, \dots, d_s$ of subset s, $s = 1, \dots, p$ only. Blundell and Robin (2000) give a long discussion of the extension of this concept to latent separability, i.e., grouping commodities without having even weak separability.

There is an enormous amount of literature discussing separability for demand and utility functions (Deaton and Muellbauer, 1980) and for production functions (Denny and Fuss, 1977; Fuss, McFadden, and Mundlak, 1978). It is often considered in the context of problems of aggregation and substitution (Berndt and Christensen, 1973) but also for the specification of flexible functional forms and separated inferences. From a statistical point of view, Stone (1985, 1986) stresses flexibility, dimensionality, and interpretability. He proves that additive modeling can circumvent the curse of dimensionality, which in nonparametrics is of fundamental importance; moreover, it makes these methods feasible for higher dimensional problems. This actually carries over to the more general case when the impact function can be decomposed into lower dimensional functionals, as will be seen in this paper.

Nonparametric estimation methods for separable models have so far focused on additive models (see, e.g., the backfitting by Hastie and Tibshirani, 1990; Mammen, Linton, and Nielsen, 1999) and on the marginal impact of particular regressors applying marginal integration (see, e.g., Tjøstheim and Auestadt, 1994; Linton and Nielsen, 1995). Recently, Horowitz (2001) has presented a conditional moment estimator for generalized additive models with unknown link function but without partial parametric modeling and has discussed 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, with the only payment being some numerical deficiencies.

Unfortunately, none of the methods mentioned previously allow for more complex structural models such as multiple index models. For some approaches to non- or semiparametric estimation of some censored Tobit models we refer to Lewbel and Linton (2002) and Ai and Chen (2002). Both use moment estimators, and therefore they do not need to introduce distributional assumptions. The latter approach also considers semiparametric separable models but can only estimate the parametric part and is of a more theoretical nature.

In this paper we propose a semiparametric estimation method for separable models. Such models allow for a finite-dimensional parameter vector and also for nonparametric modeling of separable components. The unknown finite-dimensional parameter vector is incorporated in the most general way; it can be part of the link function or the error distribution, or it can be used for partial linear modeling or even for combining nonparametric components.

This estimator extends previous estimators in at least two directions: weak and latent separability is allowed, and limited dependent variables can also be taken into account. We also derive conditions such that the finite-dimensional parameter can be estimated with the parametric rate and each separable component with the rate according to its input variables.

Because we wanted a feasible, computationally nonintensive, but well performing procedure that allows us to derive some asymptotic theory we have chosen an estimation procedure that is based on a generalized profile likelihood approach. This method requires the assumption of a likelihood function for the model errors. Certainly, our method requires more distributional assumptions than conditional moment estimators. However, several of the econometric models considered can only be identified by specifying either the distribution or the functionals. Furthermore, likelihood methods apply straightforwardly to all of the problems mentioned and have good finite-sample performance. Note finally that errors in likelihood estimation caused by violation of distributional assumptions are less problematic than sometimes believed. Instead, for models typical in economic research, the errors due to a misspecification in the index functions are usually quite serious compared to the rather negligible ones caused by misspecification of the link. See, e.g., Fernandez and Rodríguez-Póo (1997). They show that switching from a generalized linear model (GLM) to a single-index model (SIM) with unknown link usually does not really change the final results, whereas modeling the index in a more flexible way changes them a great deal.

The rest of the paper is organized as follows. In Section 2 we introduce and explain the reasons behind the model. The estimator and its asymptotic properties are given in Section 3. In Section 4 we discuss more practical questions and illustrate finite-sample behavior by simulations. Section 5 concludes and gives further discussion. All proofs are postponed to the Appendix.

2. THE STATISTICAL MODEL AND MOTIVATION

Consider a model for the relationship between a dependent variable $Y \in \mathbb{R}$ and a set of explanatory variables (X, T), $X \in \mathcal{X}$, $T \in \mathcal{T}$, such that the conditional distribution of Y given X and T belongs to the following family of density functions indexed by θ and $\eta_1, \ldots, \eta_p$:

$$\{\ell(\bullet|T; \eta_1, \ldots, \eta_p; \theta) : \eta_1 \in H_1, \ldots, \eta_p \in H_p, \theta \in \Theta\}, \tag{1}$$

where Θ is assumed to be a compact subset of $\mathbb{R}^k$, $H_1, \ldots, H_p$ are, respectively, compact subsets in $\mathbb{R}$, and $\mathcal{X}$ and $\mathcal{T}$ are also assumed to be compact sets $\mathcal{X} \subset \mathbb{R}^d$, $\mathcal{T} \subset \mathbb{R}^k$.

The relationship between Y and (X, T) will be fully characterized by determining the values of θ and η_1, $\eta_2, \ldots, \eta_p$, where θ is an unknown parameter vector and η_j's are assumed to be unknown smooth functions of not necessarily disjoint subsets of X, $\eta_j : \mathcal{X}_{d_j} \to H_j$, that take values in a set

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

More precisely, $\eta_1 = \eta_1(x^1)$, $\eta_2 = \eta_2(x^2), \ldots, \eta_p = \eta_p(x^p)$, where the vectors $x^i \in \mathcal{X}_{d_i}$, $i = 1, \ldots, p$, may contain common elements.

For the sake of simplicity, we will consider X and T absolutely continuous random variables defined in a compact support. This is by no means necessary; we could also include discrete or even dummy variables, especially for T, replacing the densities by probabilities measures. For the variable X that will enter into the nonparametric estimation part we refer to Delgado and Mora (1995). They show that the impact of discrete variables can be handled nonparametrically in the same way and does not even affect the rate of convergence.

Note that our approach is fairly general and nests all separable models. In fact, the structure of the model will depend on the family of conditional densities into which it will be assumed to fall. That is, the functional form of $\ell(\cdot|T; \eta_1, \eta_2, \ldots, \eta_p; \theta)$ will be determined not only by the statistical assumptions but also by the restrictions introduced by economic theory.

To explain our considerations further we present one example from economics where separability and/or limited dependency is of essential interest. Gronau (1973) proposes a theory of how a housewife decides whether or not to work and how much to work. He assumes that the wage offered, W^o, is given to each housewife independently of the hours worked H. Then, given a W^o, a housewife maximizes her utility function $U(C, X^1, X^2)$ subject to $X = W^oH + V$ and $C + H = \mathit{Time}$, where C is time spent at home for child care, $X = (X^1, X^2)$ represents all other goods, *Time* is total available time, and V is other income. Thus, a housewife does not work if

$$\left[\frac{\partial U/\partial C}{\partial U/\partial X}\right]_{H=0} > W^o$$

and does work if the inequality is reversed. If she works, the hours of work H and the actual wage rate W must be such that

$$\left[\frac{\partial U/\partial C}{\partial U/\partial X}\right] = W.$$

Gronau calls the left-hand side of the preceding equation the housewife's value of time, i.e., the reservation wage, W^r. If the utility function $U(\cdot)$ is strongly separable in the bundles of goods X^1 and X^2, then Gronau's model could be statistically described as follows:

$$\begin{aligned}
&w_i^o = \eta_1(x_i^1) + u_{1i}, \qquad w_i^r = \eta_2(x_i^2) + u_{2i} - u_{1i},\\
&w_i = w_i^o \quad \text{if } w_i^o > w_i^r,\\
&w_i = 0 \quad \text{if } w_i^o \le w_i^r,\ i = 1,2,\dots,N,
\end{aligned} \tag{2}$$

where he assumes that the pair (u_{1i}, u_{2i}) are independent and identically distributed (i.i.d.) random variables generated from a bivariate Gaussian distribution with zero mean, variances σ_1^2, σ_2^2, and covariance σ_{12}.

Note that in this example we have considerably weakened Gronau's assumptions because we are allowing for $\eta_1(\cdot)$ and $\eta_2(\cdot)$ to be unknown smooth functions of the different variables instead of imposing, as he did, that w^o and w^r be writable as linear combinations of the independent variables. As already remarked in many studies, linearity is often imposed for the sake of simplicity in the specification of the statistical model but without justification from economic theory. Unfortunately, if our interest lies in weakening the restrictions that concern the index function, we cannot at the same time relax the assumptions about the conditional distribution of the error term. Therefore, for the sake of identification, we need to take the conditional density of the error term as known.

In model (2) we are interested in estimating the parameters $\theta = (\sigma_1^2, \sigma_2^2, \sigma_{12})$ and the unknown functions $\eta_1(\cdot)$ and $\eta_2(\cdot)$. Just to simplify further discussion we impose for the sake of simplicity $\sigma_{12} = 0$. This condition is used in Gronau (1973); however, as already remarked in Amemiya (1985), this is a rather unrealistic assumption, and in fact it is not necessary for estimation purposes. Then, the full likelihood based on the statistical model that is described in (2) is

$$\begin{aligned}
L = \prod_0 &\left[1 - F\left\{\frac{\eta_1(x_i^1) - \eta_2(x_i^2)}{\sigma_2}\right\}\right]\\
&\times \prod_1 F\left\{\frac{\eta_1(x_i^1) - \eta_2(x_i^2)}{\sigma_2}\right\}\sigma_1^{-1} f\left\{\frac{w_i - \eta_1(x_i^1)}{\sigma_1}\right\},
\end{aligned} \tag{3}$$

where F is the cumulative standard normal distribution and f is its density, $\prod_0$ is the product over all observations without jobs, and $\prod_1$ is the product overall with jobs. Note that (3) represents a multiple-index model where the nonparametric components are introduced into the structural model additively.

The assumption of strong separability of the utility function can be relaxed, e.g., by introducing weak separability between the bundles X^1 and X^2. Then Gronau's model can take the following alternative form:

$$\begin{aligned}
&w_i^o = \eta_1(x_i^1) + u_{1i}, \qquad w_i^r = \eta_1(x_i^1)\eta_2(x_i^2) + u_{2i} - u_{1i},\\
&w_i = w_i^o \quad \text{if } w_i^o > w_i^r,\\
&w_i = 0 \quad\ \ \text{if } w_i^o \leq w_y^r, \qquad i = 1,2,\ldots,N,
\end{aligned} \tag{4}$$

and the full likelihood expression for model (4) takes the form

$$\begin{aligned}
L = \prod_0 &\left[1 - F\left\{\frac{\eta_1(x_i^1)\{1-\eta_2(x_i^2)\}}{\sigma_2}\right\}\right]\\
&\times \prod_1 F\left\{\frac{\eta_1(x_i^1)\{1-\eta_2(x_i^2)\}}{\sigma_2}\right\}\sigma_1^{-1} f\left\{\frac{w_i - \eta_1(x_i^1)}{\sigma_1}\right\}.
\end{aligned} \tag{5}$$

In this case (5) represents a multiple-index model where the nonparametric components are introduced in a nonadditive manner. Certainly, a further possibility for relaxing the model would be to make use of latent separability, i.e., allowing for some regressors to appear in both X^1 and X^2.

Following this example, it is also possible to develop a structural model where the participation (binary choice) equation in (2) and (4) is replaced by a truncated Tobit model. This extension is based on Heckman's model (Heckman, 1974), which differs from Gronau's model in that the determination of hours worked, H, is included in the model. If we are only interested in the hours of work (H) we have the following truncated Tobit model:

$$H_i = \begin{cases} h(x_i, t_i) + u_i & \text{if } h(x_i, t_i) + u_i > 0\\ 0 & \text{otherwise,}\end{cases} \tag{6}$$

where X and T are characterized at the beginning of Section 2. Assume also that the error term, u, is normally distributed with zero mean and variance σ^2. Then it could be interesting to model $h(\cdot)$ in various ways: a partially linear model, a semiparametric partially additive model (strong separability), or weaker forms of separability such as weak or latent separability. Under the preceding assumptions the truncated Tobit model has the following likelihood function:

$$L = \prod_1 F\left(\frac{h(x_i, t_i)}{\sigma}\right)^{-1} \sigma^{-1} f\left(\frac{H_i - h(x_i, t_i)}{\sigma}\right). \tag{7}$$

Setting $\theta = (\gamma^T, \sigma)^T$ we again face one of the regression problems discussed. Compare also Section 4.

In the preceding examples, we have given reasons for working with a wide class of semiparametric structural econometric models that present estimation problems that are far from trivial and, to our knowledge, remain unsolved: the estimation of censored and truncated models under different types of separability restrictions. Certainly, in the much simpler case of a single regressor on $[0,1]$, Conditions I, S, and NP of Severini and Wong (1992) allow censored or truncated variables. Under the assumption of strong separability, alias additive models, a wide class of estimators has been already reviewed in Section 1. However, the assumption of separability is much weaker than additivity. It allows for any combination and interaction between the component functions, as can be realized in (4). The main advantage of the estimation procedure proposed in this paper is that it can handle a wide class of separable models jointly with any kind of nonparametric multi-index structure, as we have in (2) or (4).

3. THE ESTIMATOR

Suppose we have a sample of N independent replicates $\{(Y_i, X_i, T_i)\}_{i=1,\dots,N}$. Our goal will be to estimate jointly the parameter vector $\theta \in \Theta$ and the nuisance parameters $\eta_1, \eta_2, \dots, \eta_p$. If as stated in Section 1 the conditional density of Y given X and T belongs to the family of parametric functions represented in (1) we could write the following likelihood function:

$$L_n(\theta, \eta_1, \dots, \eta_p) = \sum_{i=1}^{N} \ell(Y_i | T_i; \eta_1, \dots, \eta_p; \theta) \tag{8}$$

and then maximize it over the parameters of interest. Estimation of a finite-dimensional parameter in the presence of an infinite-dimensional nuisance parameter has been considered by a number of authors. Here, we propose estimators that are based on a generalized profile likelihood approach (see Severini and Wong, 1992). The estimation procedure consists of approximating the likelihood function locally through a weighted likelihood approach developed in Staniswalis (1989). Furthermore, under certain hypotheses on the likelihood function we also develop an estimator that is based on maximizing a local quasi-likelihood function (see McCullagh and Nelder, 1989; Severini and Staniswalis, 1994). The main advantage of the method based on the quasi-likelihood function is that there is no need to assume knowledge of the conditional density function. However, its drawback is that it rules out the possibility of considering the multiple-index models that are typical in standard microeconomic analysis.

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

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

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

   where the weighted likelihood is

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

   and $K(\cdot)$ is a d-variate kernel function and h the corresponding bandwidth. Note also that all estimators depend on θ. This step is repeated for all x_0 because these estimators are needed for the next part of the algorithm.
2. Given the previous estimates for the nonparametric part, we perform a simple maximum likelihood estimation for θ_0, i.e.,

   $$\hat{\theta}_N = \sup_{\theta \in \Theta} \sum_{i=1}^{N} \log \ell(Y_i, T_i; \hat{\eta}_{1,\theta}(X_i^1), \dots, \hat{\eta}_{p,\theta}(X_i^p), \theta), \tag{9}$$

   and set $\hat{\eta}_j = \hat{\eta}_{j,\hat{\theta}_N}$ for all $j = 1, 2, \dots, p$.
3. Now, with the estimators obtained in steps 1 and 2, we reestimate the nonparametric part as follows:

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

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

The first two steps are derived from the generalized profile likelihood approach proposed in Severini and Wong (1992) but extended to the case in which θ, $\eta_1, \dots, \eta_p$, and X are multidimensional. From these two steps we obtain a root-N-consistent, efficient estimator of θ_0, but unfortunately the nonparametric components are estimated with the problem of the curse of dimensionality.

The third step is included to reduce the curse of dimensionality (d is reduced to d_j) of the nonparametric estimators. Furthermore, it is also argued that this third step is oracle efficient, i.e., as efficient as the infeasible estimate that is based on knowing all components but the one of interest. This property has been discussed in the context of additive models (see Linton, 1997, 2000) and is extended here to weak and latent separable models.

Let us introduce some notation and assumptions that will be used in the remainder of the paper. Denote by $p(X)$ the marginal density of $X = (X^1, \dots, X^p)$ and by $p_j(X^j)$ the marginal density of X^j. Further, set $\sigma^2(x) = E[Y^2 | X = x]$, $\sigma_j^2(x^j) = E[Y^2 | X^j = x^j]$, and

$$\varphi(y, t; \eta_1, \dots, \eta_p, \theta) = \ln \ell(y, t; \eta_1, \dots, \eta_p, \theta), \tag{10}$$

$$F_j^{(l)}(y, t; \eta_1, \dots, \eta_p, \theta) = \frac{\partial}{\partial \eta_j^l} F(y, t; \eta_1, \dots, \eta_p, \theta) \qquad j = 1, \dots, p. \tag{11}$$

Here, $F_j(\cdot)$ can be, respectively, $\ell_j(\cdot)$, $\varphi_j(\cdot)$, or $r_j(\cdot)$. Let μ_x denote a k-vector of nonnegative integer constants. For such a vector we define

(i) $|\mu_x| = \sum_{j=1}^{k} \mu_j$, where $\mu_x = (\mu_1, \ldots, \mu_k)^T$,
(ii) for any function $a(x)$ on $\mathbb{R}^k$:

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

Now, we require that the family of density functions

$$\{\ell(\bullet | T; \eta_1, \ldots, \eta_p; \theta) : \eta_1 \in H_1, \ldots, \eta_p \in H_p, \theta \in \Theta\}$$

satisfy the following conditions:

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

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

$$\theta \in \Theta, \eta_1 \in H_1, \ldots, \eta_p \in H_p.$$

If $\theta \neq \theta_1$, then

$$\rho(\eta_1, \eta_2, \ldots, \eta_p, \theta) < \rho(\eta_1^+, \eta_2^+, \ldots, \eta_p^+, \theta_1).$$

Let $I_\theta(\eta_1, \eta_2, \ldots, \eta_p, \theta)$ denote the marginal Fisher information for θ in the parametric model, i.e.,

$$\begin{aligned} I_\theta(\eta_1, \eta_2, \ldots, \eta_p, \theta) \\ &= E\left[\frac{\partial}{\partial\theta} \varphi(Y, T; \eta, \theta) \frac{\partial}{\partial\theta^T} \varphi(Y, T; \eta, \theta)\right] \\ &\quad - E\left[\frac{\partial}{\partial\theta} \varphi(Y, T; \eta, \theta) \frac{\partial}{\partial\eta^T} \varphi(Y, T; \eta, \theta)\right] \\ &\quad \times E\left[\frac{\partial}{\partial\eta} \varphi(Y, T; \eta, \theta) \frac{\partial}{\partial\eta^T} \varphi(Y, T; \eta, \theta)\right]^{-1} \\ &\quad \times E\left[\frac{\partial}{\partial\eta} \varphi(Y, T; \eta, \theta) \frac{\partial}{\partial\theta^T} \varphi(Y, T; \eta, \theta)\right], \end{aligned}$$

where

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

and

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

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

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

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

exists for almost all Y and T. Further, assume that

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

Condition (A.1) is an identification condition. It is equivalent to Condition I of Severini and Wong (1992, p. 1777). Condition (A.2) is standard in likelihood-related problems, and it allows differentiation and integration to be interchanged when differentiating

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

This assumption is equivalent to Condition S of Severini and Wong (1992, p. 1777).

Next, we need to include some smoothness assumptions that are necessary because of the use of nonparametric smoothing methods.

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

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

Then

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

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

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

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

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

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

implies that

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

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

(B.3). We define

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

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) $E(\sup_\eta \sup_\theta |\Delta_{\eta,\theta}^{r_\eta, s_\theta}(Y,T)|) < \infty$ for $|r_\eta| \leq 5$ and $|s_\theta| \leq 3$,
(ii) for any even integer $q \geq 10$ it holds that $\sup_\eta \sup_\theta E\{|\Delta_{\eta,\theta}^{r_\eta, s_\theta}(Y,T)|^q\} < \infty$ for $|r_\eta| \leq 3$ and $|s_\theta| \leq 4$,
(iii) $\sup_\eta \sup_\theta \sup_{y,x,t} |f_{\eta\theta}^{(r_\eta, s_\theta)}(y,t|x)| < \infty$ 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) and $0 < \inf_x p(x) < \sup_x p(x) < \infty$.

These smoothness assumptions affect the rate of convergence of the nonparametric estimators of the different components $\eta_1, \ldots, \eta_p$. In fact, in the Appendix it is shown that (B.1)–(B.3) are sufficient conditions to prove that the nonparametric estimators $\hat{\eta}_j$, $j = 1, \ldots, p$, from Step I, have the following properties:

(i) the $N^{1/4}$-consistency condition (see Andrews, 1994)

$$\sup_{x_0^j \in \mathcal{X}_{d_j}} |\widehat{\hat{\eta}}_j(x_0^j) - \eta_j(x_0^j)| = o_p(N^{-1/4}),$$

(ii) and thus they are estimators of least favorable curves.

Property (i) provides the slowest rate of convergence that the nonparametric estimators must fulfill to allow $\hat{\theta}_N$ (in step 2) to achieve the $\sqrt{N}$ rate. The estimators proposed by Klein and Spady (1993) and Ichimura and Lee (1991), e.g., satisfy this property. Property (ii) imposes an asymptotic orthogonality condition between the parametric and the nonparametric estimators. This implies that the asymptotic distribution of $\hat{\theta}_N$ is not affected by the distribution of $\hat{\eta}_1, \ldots, \hat{\eta}_p$ and therefore the parametric estimator achieves the marginal Fisher information for θ. Properties (i) and (ii) are equivalent to what is called Condition NP in Severini and Wong (1992, p. 1779).

Finally, we also need to impose some conditions on $K(\cdot)$ and h_N. The main conditions are as follows.

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

$$\int z_1^{i_1} \dots z_d^{i_d} K(z_1, z_2, \dots, z_d)\, dz_1 \dots dz_d = \begin{cases} 1 & \text{if } i_1 = i_2 = \cdots = i_d = 0 \\ 0 & \text{if } 0 < i_1 + i_2 + \cdots + i_d < m, \end{cases}$$

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

and

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

(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 $m/d > (q - p - 2)/(2p + q + 4)$.

To have mean square error converging to zero, if $d \ge 4$, Assumption (H.1) implies bandwidth rates of order $\alpha < \frac{1}{4}$. Thus m must be larger than 2. By Assumption (K.1) this implies a higher order kernel. This is a standard bias reducing technique, and jointly with Assumption (H.1) on the bandwidth it is needed to achieve the preceding condition of $N^{1/4}$-consistency. Note finally that the bandwidth rates required in condition (H.1) are faster than the optimal ones. This is also standard in semiparametric models and is due to the fact that the bias of the nonparametric estimator would otherwise appear in the asymptotic properties of the parametric part.

Then, if $\hat{\theta}_N$ is the solution to optimization problem (9) in step 2, the following result is proved in the Appendix.

THEOREM 1. *Under Assumptions (A.1), (A.2), (B.1)–(B.3), (K.1), and (H.1), as N tends to infinity*

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

where

$$\begin{aligned} I_\theta(\eta_1, \eta_2, \dots, \eta_p, \theta) &= E\left[\frac{\partial}{\partial\theta} \varphi(Y, T; \eta, \theta) \frac{\partial}{\partial\theta^T} \varphi(Y, T; \eta, \theta)\right] \\ &\quad - E\left[\frac{\partial}{\partial\theta} \varphi(Y, T; \eta, \theta) \frac{\partial}{\partial\eta^T} \varphi(Y, T; \eta, \theta)\right] \\ &\quad \times E\left[\frac{\partial}{\partial\eta} \varphi(Y, T; \eta, \theta) \frac{\partial}{\partial\eta^T} \varphi(Y, T; \eta, \theta)\right]^{-1} \\ &\quad \times E\left[\frac{\partial}{\partial\eta} \varphi(Y, T; \eta, \theta) \frac{\partial}{\partial\theta^T} \varphi(Y, T; \eta, \theta)\right] \end{aligned}$$

and

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

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

As can be observed from this result, the semiparametric estimator achieves the semiparametric efficiency bound (see Chamberlain, 1992; Newey, 1990, 1994). Note that our model restrictions do not contain any information about a possible dependence structure between X and T. Further note that the asymptotic variance could be approximated with the aid of the Hessian matrix.

Following the example introduced in Section 2, in Gronau's model with strong separability (see equation (2)), the nonparametric estimators for η_1 and η_2 are obtained by maximizing the smoothed log-likelihood version of (3) and then

$$\hat{\eta}_1=\frac{\frac{1}{Nh^d}\sum_{i=1}^{N}K\left(\frac{x_0-x_i}{h}\right)\mathbb{1}(w_i^o-w_i^r>0)w_i}{\frac{1}{Nh^d}\sum_{i=1}^{N}K\left(\frac{x_0-x_i}{h}\right)\mathbb{1}(w_i^o-w_i^r>0)},\tag{12}$$

$$\hat{\eta}_2=\hat{\eta}_1-\sigma_2F^{-1}\left(\frac{\frac{1}{Nh^d}\sum_{i=1}^{N}K\left(\frac{x_0-x_i}{h}\right)\mathbb{1}(w_i^o-w_i^r>0)}{\frac{1}{Nh^d}\sum_{i=1}^{N}K\left(\frac{x_0-x_i}{h}\right)}\right).\tag{13}$$

The finite-dimensional parameter σ_1^2 can be estimated by maximizing the unsmoothed log-likelihood version of (3),

$$\hat{\sigma}_1^2=\frac{\sum_{i=1}^{N}(w_i-\hat{\eta}_1(x_i))^2\mathbb{1}(w_i^o-w_i^r>0)}{\sum_{i=1}^{N}\mathbb{1}(w_i^o-w_i^r>0)},\tag{14}$$

whereas $\hat{\sigma}_2$ is a solution of an implicit formula expression.

Then, under the conditions stated in Theorem 1, we obtain the following result:

$$\sqrt{N}(\hat{\sigma}_1^2-\sigma_1^2)\to_d N\left(0,\sigma_1^4\left\{\frac{1-\gamma}{\gamma}\right\}\right)\tag{15}$$

and

$$\gamma=E\left[F\left(\frac{\eta_1(X^1)-\eta_2(X^2)}{\sigma_2}\right)\right].\tag{16}$$

If instead of using strong separability, alias additivity, in Gronau's model, we use weak separability as in equation (4), then the estimator in step 1 for η_1 is the same as in the additive case, but for η_2 we obtain the expression

$$\hat{\eta}_2 = 1 - \frac{\sigma_2}{\hat{\eta}_1} F^{-1}\left(\frac{\dfrac{1}{Nh^d}\displaystyle\sum_{i=1}^{N} K\left(\frac{x_0 - x_i}{h}\right) \mathbb{1}(w_i^o - w_i^r > 0)}{\dfrac{1}{Nh^d}\displaystyle\sum_{i=1}^{N} K\left(\frac{x_0 - x_i}{h}\right)} \right) \tag{17}$$

for $\eta_1(x_0) \neq 0$, $\forall x_0 \in \mathcal{X}$. Further expressions can be obtained as in the strongly separable case.

Next we study the asymptotic behavior of the nonparametric estimators obtained in step 3. To do this, we introduce the following additional assumptions.

(C.1). $Nh_j^{d_j} \to \infty$ and $Nh_j^{d_j+4} \to 0$, for $j = 1, \ldots, p$, 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 $j = 1, \ldots, p$ it holds that

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

Assumptions (C.1)–(C.3) are standard in nonparametric regression literature. For example, condition (C.1) makes variance and bias tend to zero when the sample size increases. The condition $Nh_j^{d_j+4} \to 0$ can be weakened by assuming only that $h_j \to 0$. Then, under the previous assumptions we obtain

$$\sup_{x_0^j \in \mathcal{X}_{d_j}} |\widehat{\widehat{\eta}}_j(x_0^j) - \eta_j(x_0^j)| = O_p\left(\sqrt{\frac{\log N}{Nh_j^{d_j}}}\right) + O(h_j^2).$$

Therefore, the asymptotic distribution of $\widehat{\widehat{\eta}}_j(x_0^j)$ will have a nonnegligible bias of order $O(Nh_j^{d_j+4})$. To correct the bias, two alternative procedures are available in the relevant literature. One is to approximate the distribution of $\sup_{x_0^j \in \mathcal{X}_{d_j}} |\widehat{\widehat{\eta}}_j(x_0^j) - \eta_j(x_0^j)|$, e.g., by a bootstrap or subsampling technique (see Section 5). The other approach is to assume that the bandwidth h_j tends to zero slightly faster than the optimal rate, i.e., $Nh_j^{d_j+4} \to 0$. We have chosen the second procedure, although we remark that at this rate of decay of the bandwidth, our nonparametric estimator will achieve a rate of convergence that is slightly slower than the optimal rate. Condition (C.3) is an identification condition that is similar in additive models under Gaussian errors to the backfitting algorithm (see Hastie and Tibshirani, 1990). Under these assumptions the following theorem holds.

THEOREM 2. *Under the conditions of Theorem 1 and Assumptions (C.1)–(C.3), for any* $j = 1, \ldots, p$

(i)

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

(ii)

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

where

$$V_j(\eta_j,\theta_0)=\int K_j^2(t)\,dt\,\frac{I_j(\eta_j,\theta_0)}{p_j(x_0^j)\{H_j(\eta_j,\theta_0)\}^2}, \tag{18}$$

$$H_j(\eta_j,\theta_0)=E[-\varphi_j^{(2)}(Y,T;\eta_1(X^1),\ldots,\eta_j(X^j),\ldots,\theta_0)|X^j=x_0^j],\ \textit{and} \tag{19}$$

$$I_j(\eta_j,\theta_0)=E[\varphi_j^{(1)}(Y,T;\eta_1(X^1),\ldots,\eta_j(X^j),\ldots,\eta_p(X^p),\theta_0)^2|X^j=x_0^j] \tag{20}$$

as N tends to infinity.

Remark. Note that under correct specification of the statistical model the information equality holds, $H_j(\eta_j,\theta_0)=I_j(\eta_j,\theta_0)$, and then the asymptotic variance in (18) collapses to

$$V_j(\eta_j,\theta_0)=\int K_j^2(t)\,dt\,\frac{1}{p_j(x_0^j)H_j(\eta_j,\theta_0)}.$$

As an example, in Gronau's model with strong separability (see equation (2)) under the conditions assumed in Theorem 2, the estimator for $\eta_1(x_0^1)$ obtained in step 3 has the closed form

$$\hat{\hat{\eta}}_j(x_0^1)=\frac{\dfrac{1}{Nh_1^{d_1}}\displaystyle\sum_{i=1}^N K_1\left(\frac{x_0^1-x_i^1}{h}\right)\mathbb{1}(w_i^o-w_i^r>0)w_i}{\dfrac{1}{Nh^{d_1}}\displaystyle\sum_{i=1}^N K_1\left(\frac{x_0^1-x_i^1}{h}\right)\mathbb{1}(w_i^o-w_i^r>0)} \tag{21}$$

and shows the following asymptotic behavior:

$$\sqrt{Nh_1^{d_1}}(\hat{\hat{\eta}}_1(x_0^1)-\eta_1(x_0^1))\rightarrow_d N\{0,V(x_0^1)\}$$

as N tends to infinity, where

$$V(x_0^1)=\frac{\displaystyle\int K_1^2(u)\,du\{\sigma_1^2+\eta_1^2(x_0^1)\}}{p_1(x_1^0)E\left\{F\left(\dfrac{\eta_1(X^1)-\eta_2(X^2)}{\sigma_2}\right)\middle|X^1=x_0^1\right\}}.$$

In the weakly separable case of equation (4), in step 3 we obtain the same estimator for η_1 as in the additive case. However, its asymptotic variance is different. In this case, it can be shown that

$$V^w(x_0^1) = \frac{\displaystyle\int K_1^2(u)\,du\{\sigma_1^2 + \eta_1^2(x_0^1)\}}{p_1(x_1^0)E\left\{F\left(\dfrac{\eta_1(X^1)\{1-\eta_2(X^2)\}}{\sigma_2}\right)\middle| X^1 = x_0^1\right\}}.$$

As can be noticed from the theorem, all nonparametric components are estimated at the fastest possible rate overcoming the curse of dimensionality. This agrees with the results found by Stone (1986) for additive models, but we remark that the same result is achieved here with a weaker restriction on separability. Also, as indicated before, we reach the same asymptotics as if the other components in the model were known. Finally, the (pointwise) asymptotic expressions could be estimated without too much extra calculation. Unfortunately, it is well known that in finite samples these estimates of asymptotic expressions are of little practical help. Instead, we again refer to Section 5.

In some contexts, the results obtained in Theorems 1 and 2, mainly consistency and Gaussian limiting distribution, still hold even if only part of the model is correctly specified. More precisely, by taking a one-parameter linear exponential family as the conditional distribution of Y given X, maximum likelihood estimates of the parameters of interest are consistent and asymptotically normal. The quasi-likelihood approach allows us to estimate models up to the identifiable (separable) components whenever they present a single-index structure.

Each member of the family has a density that can be written as

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

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

$$E(Y|X = x, T = t) = g(t, \eta_1(x^1), \dots, \eta_p(x^p), \theta_0), \text{ and}$$

$$V(Y|X = x, T = t) = \sigma^2 V_0(g(t, \eta_1(x^1), \dots, \eta_p(x^p), \theta_0)),$$

where $g(\cdot)$ is a known function. Note that again heteroskedasticity is included but the variance function V_0 must depend on T and X through the index $g(t, \eta_1(x^1), \dots, \eta_p(x^p), \theta_0)$. In this case, it is possible in steps 1–3 to replace the log-likelihood $\log \ell(\cdot)$ by the quasi-likelihood function $r(\cdot, g(t, \eta_1(x^1), \dots, \eta_p(x^p), \theta_0))$ defined as

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

where $V(\cdot)$ is a known function. If we replace the true distribution by the quasi-likelihood function in all steps of our estimation procedure, then, as shown in Rodríguez-Póo, Sperlich, and Vieu (2000), the preceding results still hold. However, as might be expected, there is an efficiency loss if the specification is not equivalent to the real distribution.

To show the main results of this section we need to make the following assumptions.

(Q.1). Let $\mathcal{G}$ denote a compact subset of $I\!R$ such that $g(t,\eta_1(x^1),\ldots,\eta_p(x^p),\theta) \in \mathcal{G}$ for all $t \in \mathcal{T}$, $x^1 \in \mathcal{X}_{d_1},\ldots,x^p \in \mathcal{X}_{d_p}$,$\eta_1 \in H_1,\ldots,\eta_p \in H_p$, 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 $p = 1,\ldots,3$ then $\partial^p V(g)/\partial g^p$ exists and is bounded for all $g \in \mathcal{G}$.

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

(B.1′). As for (B.1) replacing $\varphi(\cdot)$ by $r(\cdot)$.

(B.2′). As for (B.2) replacing $\varphi(\cdot)$ by $r(\cdot)$.

(B.3′). As for (B.3) replacing $\varphi(\cdot)$ by $r(\cdot)$.

(C.3′). For all $j = 1,\ldots,p$

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

Assumptions (Q.1)–(Q.3) are regularity conditions needed in the quasi-likelihood framework mainly to guarantee that the quasi-likelihood function used in estimating $\eta_1,\ldots,\eta_p$ and θ has the properties of a likelihood function. Condition (C.3′) is the same as the identification condition (C.3) but now in terms of the quasi-likelihood function.

THEOREM 3. *Under Assumptions (B.1′)–(B.3′), (K.1), (H.1), and (Q.1)–(Q.3), as N tends to infinity*

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

where

$$\begin{aligned} J_\theta(\eta_1, \eta_2, \dots, \eta_p, \theta) &= E\left[\frac{\partial}{\partial\theta} g(T;\eta,\theta_0)\frac{\partial}{\partial\theta^T} g(T;\eta,\theta_0)\right] \\ &\quad - E\left[\frac{\partial}{\partial\theta} g(T;\eta,\theta_0)\frac{\partial}{\partial\eta^T} g(T;\eta,\theta_0)\right] \\ &\quad \times E\left[\frac{\partial}{\partial\eta} g(T;\eta,\theta_0)\frac{\partial}{\partial\eta^T} g(T;\eta,\theta_0)\right]^{-1} \\ &\quad \times E\left[\frac{\partial}{\partial\eta} g(T;\eta,\theta_0)\frac{\partial}{\partial\theta^T} g(T;\eta,\theta_0)\right], \end{aligned}$$

with

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

and

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

As can be seen now, if the model is correctly specified both criterion functions coincide, and therefore we obtain the efficient estimator that was shown in Theorem 1.

The behavior of the estimator obtained in step 3 is given in the following theorem.

THEOREM 4. *Under the conditions of Theorem 3 and Assumptions (C.1), (C.2), and (C.3′), we have for any* $j = 1, \dots, p$

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

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

where

$$
\begin{aligned}
V_j(\eta_j,\theta_0) = {} & \frac{\int K_j^2(t)\,dt}{p_j(x_0^j)} \\
& \times \Bigg(E\Bigg[\Bigg\{ \frac{Y - g(t,\eta_1(X^1),\ldots,\eta_p(X^p),\theta_0)}{V(g(t,\eta_1(X^1),\ldots,\eta_p(X^p),\theta_0))}\,\frac{\partial}{\partial\eta_j} \\
& \quad \times g(T;\eta_1(X^1),\ldots,\eta_j(X^j),\ldots,\eta_p(X^p),\theta_0)\Bigg\}^2 \Bigg| X^j = x^j\Bigg] \Bigg/ \\
& \quad E\Bigg[\frac{1}{V_0(g(t,\eta_1(X^1),\ldots,\eta_p(X^p),\theta_0))} \times \frac{\partial}{\partial\eta_j} \\
& \quad \times g(T;\eta_1(X^1),\ldots,\eta_j(X^j),\ldots,\eta_p(X^p),\theta_0)^2 | X^j = x^j\Bigg]^2\Bigg)^{1/2},
\end{aligned}
$$

as N tends to infinity.

From Theorem 4 we remark on two important issues: the estimator is oracle efficient, i.e., as efficient as the infeasible estimate that is based on knowing all components except the one of interest. Moreover, this estimator avoids the curse of dimensionality. Linton (2000) proposes a two-step estimator for the additive components in a generalized additive (i.e., strongly separable) nonparametric regression model. In our approach, if we adopt a generalized additive model, the asymptotic variance given in Theorem 4 collapses to the variance given in Linton (2000, Theorem 1, p. 506). This result was expected because, for the strongly separable case, to obtain the estimators for the additive components Linton proposes a local linear approximation, whereas we propose a local constant one. As is well known (see Fan, 1992), the difference between the two types of approximation is in the bias, but the variance stays the same. Both estimators avoid the curse of dimensionality, but Linton's does not consider the weak and latent separability case.

4. COMPUTATIONAL ASPECTS AND SIMULATIONS

In this section we deal with some questions that are relevant in practice and illustrate the numerical performance of our estimators by some simulations. After a bandwidth discussion we come back to the model examples from Section 2, beginning with the censored Tobit model. Second, we consider a truncated Tobit and weak separability. In that example we also discuss the problem when the nonparametric part of the model of interest cannot be completely identified in step 1. For those cases we suggest a feasible, well performing algorithm, for which we also present a small simulation study.

Unfortunately, the choice of bandwidth or the corresponding smoothing parameters in spline-, wavelet-, etc., smoothing is still an open problem in multidimensional regression and testing. From a mathematical point of view there exists an optimal bandwidth in the sense that the mean squared error is minimized. Even though for our estimators plug-in methods and cross validation are both possible in theory the inexactness of the asymptotic expressions in finite samples and computational burdens renders these approaches impractical. On the other hand, one has to admit that in our context of estimating flexible functional forms non- and semiparametric methods are explorative tools, maybe even just for finding the proper parametric model. Looking at matters in that way, an increase in the bandwidth simply corresponds to a decrease in the degrees of freedom, unfortunately allowing an increasing bias toward constant functionals. An extension of our approach to a local linear smoother would relax this to a bias toward linear functionals (cf., e.g., Lejeune, 1985). But this is left open for future research. Note that these considerations do not hold for the problem of testing parametric vs. nonparametric functionals. There, the optimal bandwidth is the one that guarantees the correct significance level, which is usually a different bandwidth from the mean squared error minimizing one. As a consequence of all this, the empirical researcher should apply various bandwidths, always depending on their "tolerance" against smoothness or wiggliness. This holds especially for the choice of the h_j's in step 3.

In step 1, bandwidth h_N has to be chosen appropriately to yield good estimates for the parametric part of the model. This can be handled by considering the nonparametric part as nuisance parameters. Then, as can also be seen from theory (e.g., in condition (H.1)), one should undersmooth in step 1. All we need there for the nonparametric part is sufficient smoothness to reach convergence in the maximization algorithm. Therefore, in practice we choose bandwidths close to the smallest ones that still yield numerical convergence in steps 1 and 2. In applications on real data some additional weighting or trimming could be useful for points at the boundary or outliers.

How can a possibly unwanted double effect of h_N on step 3 be avoided? We speak of double effect because the final estimates $\widehat{\widehat{\eta}}_j$ are affected not only by the estimate of θ but also by the preestimates of the η_j, $j = 1, \dots, p$. As indicated in Section 3, the asymptotic results in Theorems 2 and 4 do not change if we use consistent preestimates different from those proposed. So one could, e.g., repeat step 3 without changing the theoretical results. This means in practice that by only one iteration in step 3, one can limit the impact of h_N (on the final estimates of the η_j) to the impact caused by $\hat{\theta}_{h_N}$. This allows us to consider the bandwidth choices in steps 1 and 3 separately.

4.1. Example 1—Censored Tobit

As mentioned previously, we come back to the censored Tobit models introduced in Section 2. We first consider the strong separable one; i.e., we generate data from

$$w_i^0 = \eta_1(x_i^1) + u_i, \qquad w_i^r = \eta_2(x_i^2) + v_i, \tag{22}$$

with $\eta_1(x) = \eta_2(x) = 2.0 + 2.0/(\exp\{-10x\} + 1)$ and u_i, v_i being independently drawn from $N\{0,1\}$ and observing $w_i = w_i^0 \, \mathbb{1}\{w_i^0 > w_i^r\}$.

Here, we have chosen the same functional form for η_1 and η_2 to compare them afterward. One might expect that the information lacking (by not observing w^r) would worsen the estimation of η_2 and σ_v compared to η_1, σ_u. We implement a procedure that maximizes the likelihood:

$$L = \prod_0 \left[1 - F\left\{ \frac{\eta_1(x_i^1) - \eta_2(x_i^2)}{\sqrt{\sigma_u^2 + \sigma_v^2}} \right\} \right] \times \prod_1 F\left\{ \frac{w_i - \eta_2(x_i^2)}{\sigma_v} \right\} \sigma_u^{-1} f\left\{ \frac{w_i - \eta_1(x_i^1)}{\sigma_u} \right\},$$

as did Gronau (1974), where F is the cumulative standard normal distribution and f its density.

As we know that correlation between the covariates worsens the estimation results for estimators such as marginal integration and backfitting (see Sperlich, Linton, and Härdle, 1999) sometimes by a great deal, we draw $N = 500$ observations $(x^1, x^2)^T$ from $N\{0, \Sigma\}$ with variances of 1.0 and covariances $\sigma_{12} = 0.3$. Further, for the simulation it is more convenient to have all data inside a given interval; therefore we transform each variable by $x \to 0.55(2.4 \arctan(x)/\pi + 1.0) - 0.55$, thus projecting all observations into the interval $[-0.55, 0.55]$. Then the functions η_1, η_2 are estimated on a grid of 30 points from -0.5 to 0.5.

In step 1 we try bandwidths $h_N = 2.2\sigma_x$, $2.4\sigma_x$, $2.6\sigma_x$, where σ_x indicates the vector of the standard deviations of x_1, x_2. Here, σ_x is close to the smallest bandwidth without running into numerical problems. Condition (H.1) requires higher order kernels in step 1. As q in (H.1) can be quite large, a kernel of order $m = 4$ is sufficient, and we choose the "optimal" fourth-order kernel (see Lejeune, 1985). Mean and standard deviation of $(\hat{\sigma}_u, \hat{\sigma}_v)$ after performing 250 simulation runs for each bandwidth h_N are displayed in Table 1.

In step 3 higher order kernels are no longer, necessary and we choose the quartic one, which is of order 2. As bandwidths we try $h := (h_1, h_2)^T = 0.8\sigma_x$, $(h_1, h_2)^T = 1.0\sigma_x$, and $1.2\sigma_x$. Note that this leads to nine different estimates (three different h_N times three different h) for η_1, η_2 from which in Figure 1

TABLE 1. Means and standard deviations (in parentheses) of the estimates resulting from steps 1 and 2 after 250 simulation runs with model (22)

h_N	$2.2\sigma_x$	$2.4\sigma_x$	$2.6\sigma_x$
$\hat{\sigma}_u$	1.0387 (0.1457)	1.0242 (0.1209)	1.0370 (0.1552)
$\hat{\sigma}_v$	0.8510 (0.2603)	0.8560 (0.2651)	0.8723 (0.2890)

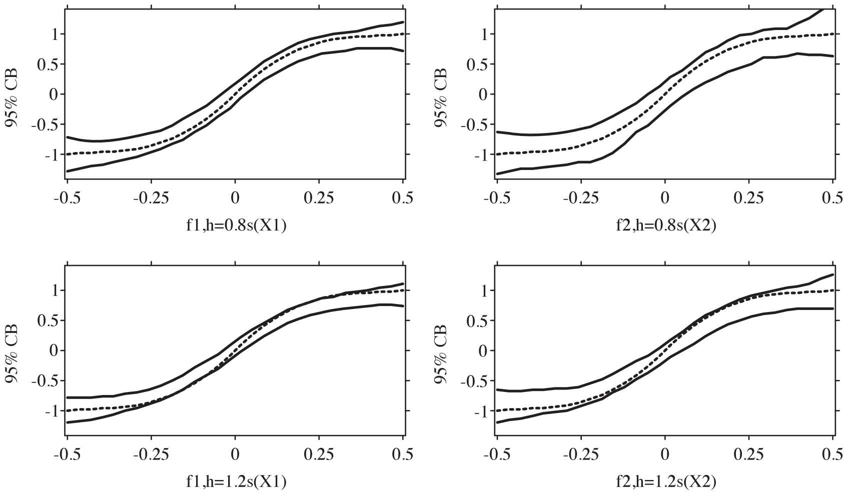

FIGURE 1. 95% pointwise confidence bands of $\hat{\hat{\eta}}_1$ (left), $\hat{\hat{\eta}}_2$ (right) from model (22) for bandwidths $h^T = 0.8\sigma_x$ in the upper row, and $h^T = 1.2\sigma_x$ in the lower one. The construction of the confidence band (c.b.) is based on 250 replications with $h_N = 2.4\sigma_x$. Dashed lines represent the data generating functions.

are given some selected results when using $h_N = 2.4\sigma_x$ in step 1. The results are displayed in terms of 95% pointwise confidence bands (without bias correction) calculated from 250 simulation runs. The dashed lines are the data generating functions.

First, as expected we discover clearly the bias and boundary effects, especially when choosing a large h in step 3. Further, one can see how the underestimation of σ_v leads to confidence bands for η_2 much wider than for η_1. Finally, as a result of the "information loss" of not observing w^r but w^o, η_2 is clearly estimated with less exactness than η_1. Nevertheless, despite the complexity of the model and highly correlated regressors the real shapes can be detected by our procedure quite well.

We now consider an example extending the estimation problem in three directions: weak separability, higher dimensions, and nonidentifiability in step 1.

4.2. Example 2—Truncated Tobit

Let us turn to a problem we so far have not explicitly discussed. Consider a truncated Tobit model as introduced in Section 2:

$$y_i = \begin{cases} t_i^T\gamma + \eta(x_i) + u_i & \text{if } t_i^T\gamma + \eta(x_i) + u_i > 0 \\ 0 & \text{otherwise,} \end{cases} \tag{23}$$

with $t_i \in I\!R^2$, $x_i \in I\!R^5$, $u_i \sim N(0,\sigma_1^2)$, and

$$\eta(x) = \eta_1(x_1) + \eta_2(x_2) + \eta_3(x_3) + \eta_3(x_3)\eta_4(x_4).$$

Recalling the likelihood function, the problem we face here is that in steps 1 and 2 we cannot uniquely identify $\eta_1, \eta_2, \dots, \eta_4$ but only γ, σ_u, and η. The estimation algorithm we suggest in that case is as follows: first proceed to steps 1 and 2 for γ, σ_u, and η. Then in step 3 solve a simultaneous equation system for the smoothed likelihoods of $\widehat{\widehat{\eta}}_1(x_i^1)$ up to $\widehat{\widehat{\eta}}_4(x_i^4)$ making use of the estimates obtained in steps 1 and 2.

A computationally feasible way is an iteration in step 3 over the four likelihoods for the $\widehat{\widehat{\eta}}_j$ as we have already suggested in the context of bandwidth discussion. The major difference here is that as a result of the identification problem in step 1, we do not have all necessary consistent preestimates at hand as demanded in Theorems 2 and 4. Although it is known that the derivation of a closed asymptotic theory for such an algorithm is impossible, it is a most intuitive procedure, known from operations research to converge under some regularity (smoothness) conditions and providing the right thing, i.e., the estimates presented in our theorems. The preestimates lacking in step 3 would be replaced by appropriately flexible parametric estimates. In our simulation study we use simply linear preestimation for η_1 to η_3 and start the iteration over η_4. Obvious extensions of our method that would allow identification but also the derivation of asymptotic theory are discussed in the next section.

The data generating process is model (23) with $\eta_1(x) = 1.0$, $\eta_2(x) = 2x^2 - 1$, $\eta_3(x) = \sin(2x)$, $\eta_4(x) = x$, $\gamma = (-1.5, 2.0)^T$, and $\sigma_u = 1.0$. The regressors are drawn independently with $t_1, t_2 \sim U[0,2]$ and $x_j \sim U[-1,1]$ for $j = 1,\dots,4$. We draw about 700 observations to end up always with $N = 600$ nontruncated $(y > 0.0)$ observations. For the estimation we apply the identification condition $E[\eta_2] = E[\eta_3] = E[\eta_4] = 0$ even though this is quite restrictive for η_4. However, simulations without fixing the expectation of η_4 lead to identification problems for η_3 and η_4.

Again we have to choose a higher order kernel in step 1 (compare condition (H.1)). Note that because of the identification problem p is just equal to 1 in steps 1 and 2. With q large, a kernel of order $m = 6$ is sufficient to fulfill (H.1). We choose the "optimal" sixth-order kernel (see Lejeune, 1985) but the quartic one in step 3. Our bandwidths are $h_N = 4.0\sigma_x$ and $h = (0.6, 0.6, 0.6, 0.9)^T\sigma_x$, respectively.

After 500 replications we get for $\hat{\theta}^T = (\hat{\gamma}_1, \hat{\gamma}_2, \hat{\sigma}_u)$ in the mean (−1.5051, 2.0115, 1.1105) with standard deviations (0.09733, 0.10479, 0.059). Figure 2 gives the 99% pointwise confidence bands for the estimators of η_1 to η_4, again without bias correction and based on 500 replications. The dashed lines are the data generating functions.

In Figure 2 we detect slight boundary effects, and certainly the bands for the functional η_1, i.e., the constant, are quite wide because we estimate it without

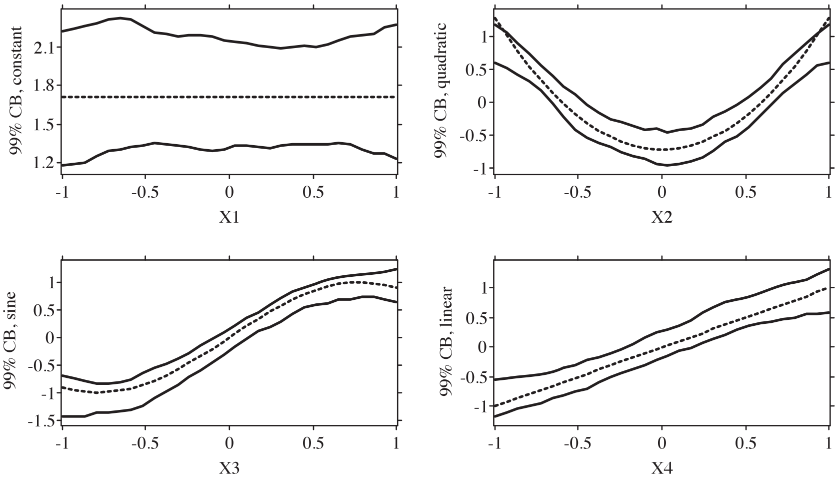

FIGURE 2. 99% pointwise confidence bands for all function estimates $\widehat{\widehat{\eta}}_1$ to $\widehat{\widehat{\eta}}_4$ from the upper left to the lower right, based on 500 replications. Dashed lines are the data generating functions.

fixing the mean $E[\eta_1]$. But otherwise the estimation procedure works quite well and always detects the correct functional form even in this high-dimensional, complex model with only $N = 600$ observations.

5. CONCLUSION AND EXTENSIONS

In this paper we present a new method for estimating semiparametric models with nonparametric separable components and/or limited dependent variables. These models are rather standard in economics literature. In Section 2 we show how our procedure can take into account many complex regression systems as Tobit models. This, to our knowledge, is the first method to date that allows semiparametric modeling of such complex structures. Aside from the theoretical consequences, shown in Section 3, the use of maximum likelihood techniques allows us to identify the semiparametric model. Moreover, it makes the estimator feasible in the small data sets typical in empirical research. This is demonstrated in Section 4 together with a detailed discussion of problems related to application.

However, for the future it will be necessary to give some attention to the problem of testing the distribution assumption. Let us emphasize again that for the sake of identification one cannot model nonparametrically both the distribution function and the index functionals. Therefore, in our context the distribution assumptions cannot be relaxed more but have to be tested afterward. In the case of SIM, i.e., models of the form $E[Y|X] = G\{\beta^T t + \eta(x)\}$, where

$\eta(\cdot)$ is nonparametric, one can test the specification of $G(\cdot)$ as proposed in the bootstrap paper by Härdle, Huet, Mammen, and Sperlich (2000). Approximate χ^2 tests could perhaps also be applied on the residuals.

The latter paper also gives many interesting guidelines with respect to bootstrap inferences in our context. The construction of confidence bands (or intervals) and the various specification tests developed there can be applied. However, making use of our estimators, the implementation to the models considered here is not immediate. Certainly the theory has yet to be properly developed.

An extension of this method to dependent data is straightforward. A careful check of the proof reveals that the same statements we made in Theorems 1–4 can be made for time series data with some strong mixing conditions. However, for the sake of transparency in the ideas of methods and proof we have restricted ourselves to the independent case and instead refer to Bosq (1998) for an idea as to how the proofs need to be modified to fit the dependent case.

An extension for which further theory is not necessary is the following. Recall the identification problem discussed in Section 4. Nonparametric functions, say, η_1 and η_2, which always occur jointly in the index or indices, and in the same additive or multiplicative way, e.g., $\eta_1 + \eta_2$, can never be identified separately in step 1 of our procedure. Instead, one would restrict oneself, in steps 1 and 2, to the estimation of the identifiable parameters and functions. That is, one would estimate $\eta_{1,2} = \eta_1 + \eta_2$. To estimate η_1, η_2 efficiently, in step 3, it is sufficient to have any consistent preestimator for η_1 or η_2. It is well known (see, e.g., Linton, 2000; Härdle et al., 2000) that these can be obtained by marginal integration (here, by integrating over $\hat{\eta}_{1,2}$), and the necessary theory carries directly over to our case. Note also that marginal integration applies not only to additivity but also to many more combinations (see Linton and Nielsen, 1995; Sperlich, Tjøstheim, and Yang, 2002; Pinske, 2000).

If one were to apply marginal integration in general, before step 3, then our paper overlaps for special cases with existing papers. For instance for single-indexed generalized additive partial linear models, our paper includes then the estimators presented in Härdle et al. (2000) but adding step 3 to yield full efficiency. For the same model Linton (2000) proposes a procedure, also using marginal integration, where he yields "efficiency" by doing a single-iteration backfit in step 3. He speaks as we do of efficiency in the sense of reaching the bounds of an oracle estimator for which the nuisance components are known. See also our references to that paper in Section 3, where we compare the final estimators in detail.

REFERENCES

Ai, C. & X. Chen (2002) Efficient estimation of models with conditional moment restrictions containing unknown functions. *Econometrica*.

Amemiya, T. (1985) *Advanced Econometrics*. Cambridge: Harvard University Press.

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

Berndt, E.R. & L.R. Christensen (1973) The internal structure of functional relationships: Separability, substitution, and aggregation. *Review of Economic Studies* 40, 403–410.

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

Bosq, D. (1998) *Nonparametric Statistics for Stochastic Processes*. Lecture Notes in Statistics 110. New York: Springer-Verlag.

Chamberlain, G. (1992) Efficiency bounds for semiparametric regression. *Econometrica* 60, 567–596.

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

Delgado, M.A. & J. Mora (1995) Nonparametric and semiparametric estimation with discrete regressors. *Econometrica* 63, 1477–1484.

Denny, M. & M. Fuss (1977) The use of approximation analysis to test for separability and the existence of consistent aggregates. *American Economic Review* 67, 404–418.

Fan, J. (1992) Design-adaptive nonparametric regression. *Journal of the American Statistical Association* 87, 998–1004.

Fernández, A.I. & J.M. Rodríguez-Poó (1997) Estimation and specifications testing in female labor participation models: Parametric and semiparametric methods. *Econometric Reviews* 16, 229–248.

Fuss, M., D. McFadden, & Y. Mundlak (1978) A survey of functional forms in the economic analysis of production. In M. Fuss & D. McFadden (eds.), *Production Economics: A Dual Approach to Theory and Applications*, vol. 1, pp. 219–268.

Goldman, S.M. & H. Uzawa (1964) A note on separability in demand analysis. *Econometrica* 32, 387–398.

Gronau, R. (1973) The effects of children on the housewife's value of time. *Journal of Political Economy* 81, S168–S199.

Härdle, W., S. Huet, E. Mammen, & S. Sperlich (2000) Bootstrap Inference in Semiparametric Generalized Additive Models. Working paper 00–70, Universidad Carlos III de Madrid, Spain.

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

Heckman, J. (1974) Shadow prices, market wages, and labor supply. *Econometrica* 42, 679–694.

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

Ichimura, H. & L.F. Lee (1991) Semiparametric estimation of multiple index models. In W.A. Barnett, J. Powell, & G. Tauchen (eds.), *Nonparametric and Semiparametric Methods in Econometrics and Statistics*, pp. 3–50. New York: Cambridge University Press.

Klein, R.W. & R.H. Spady (1993) An efficient semiparametric estimator for discrete choice models. *Econometrica* 61, 387–421.

Lejeune, M. (1985) Estimation non-paramétrique par noyaux: Régression polynomiale mobile. *Revue de Statistique Appliquée* 33, 43–67.

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. & O. Linton (2002) Nonparametric censored and truncated regression. *Econometrica* 70, 765–779.

Linton, O.B. (1997) Efficient estimation of additive nonparametric regression models. *Biometrika* 84, 469–474.

Linton, O.B. (2000) Efficient estimation of generalized additive nonparametric regression models. *Econometric Theory* 16, 502–523.

Linton, O.B. & J.P. Nielsen (1995) A kernel method of estimating structured nonparametric regression based on marginal integration. *Biometrika* 82, 93–101.

Mammen, E., O. Linton, & 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. & J.A. Nelder (1989) *Generalized Linear Models*. London: Chapman and Hall.

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.

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

Rodríguez-Póo, J.M., S. Sperlich, & P. Vieu (2000) Semiparametric Estimation of Weak and Strong Separable Models. Discussion paper 00-69, Statistics and Econometrics Series, Universidad Carlos III, Madrid.

Serfling, T. (1980) *Approximation Theorems of Mathematical Statistics*. New York: Wiley.

Severini, T.A. & J.G. Staniswalis (1994) Quasi-likelihood estimation in semiparametric models. *Journal of the American Statistical Association* 89, 501–511.

Severini, T.A. & W.W. Wong (1992) Profile likelihood and conditionally parametric models. *Annals of Statistics* 4, 1768–1802.

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

Sperlich, S., D. Tjøstheim, & L. Yang (2002) Nonparametric estimation and testing of interaction in additive models. *Econometric Theory* 18, 197–251.

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. & B.H. Auestad (1994) Nonparametric identification of nonlinear time series: Projections. *Journal of the American Statistical Association* 89, 1398–1409.

APPENDIX: PROOF OF THE MAIN RESULTS

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, 1778). Furthermore, for fixed θ, under (A.1), (A.2), (K.1), and (H.1) the estimator obtained as a solution of

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

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

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

Furthermore

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

Then, using the preceding assumptions and the properties of the Watson–Nadaraya smoother

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

and the estimator obtained in (A.1) is an estimator of a least favorable curve (Severini and Wong, 1992, p. 1779, Condition NP(b)). Condition NP(a) is obtained as follows. Let us denote

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

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) one can show that

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

and

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

for some $\gamma > 0$. For the bandwidth use the rate assumed in (H.1); then

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

and

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

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

Proof of Theorem 2. To simplify the proofs, j is fixed, and we can see that $\hat{\hat{\eta}}_j$ is indeed such that

$$\hat{\hat{\eta}}_j = \arg\max W_j^*(\eta_j,\hat{\theta}_N),$$

where

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

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, such that

$$\begin{aligned}
&\varphi_j^{(1)}(Y_i, T_i; \hat{\eta}_1^i, \ldots, \ldots, \eta_j, \ldots, \hat{\eta}_p^i, \hat{\theta}_N) \\
&\quad = \varphi_j^{(1)}(Y_i, T_i; \hat{\eta}_1^i, \ldots, \hat{\eta}_{j-1}^i, \eta_j^i, \hat{\eta}_{j+1}^i, \ldots, \hat{\eta}_p^i, \hat{\theta}_N) \\
&\qquad + (\eta_j - \eta_j^i)\varphi_j^{(2)}(Y_i, T_i; \hat{\eta}_1^i, \ldots, \hat{\eta}_{j-1}^i, \bar{\eta}_j^i, \hat{\eta}_{j+1}^i, \ldots, \hat{\eta}_p^i, \hat{\theta}_N).
\end{aligned}$$

So this leads directly to

$$\frac{\partial W_j^*(\eta_j, \hat{\theta}_N)}{\partial \eta_j} = A_1(\hat{\theta}_N) + A_2(\eta_j, \hat{\theta}_N) + [A_3(\hat{\theta}_N) + A_4(\eta_j, \hat{\theta}_N)](\eta_j - \eta_j^0), \tag{A.2}$$

where

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

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

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

and

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

Now we will study the asymptotics of the preceding terms, recalling that under the conditions established in Theorem 1, $\sqrt{N}(\hat{\theta}_N - \theta_0) = O_p(1)$ and $\sup_{x^j \in \mathcal{X}_{d_j}} |\hat{\eta}_j(x^j) - \eta_j(x^j)| = o_p(N^{-1/4})$ for $j = 1, \ldots, p$. For the term A_1 note that by the mean value theorem

$$\begin{aligned}
&\frac{1}{Nh_j^{d_j}} \sum_{i=1}^{N} K_j\left(\frac{x_0^j - X_i^j}{h_j}\right) \varphi_j^{(1)}(Y_i, T_i; \hat{\eta}_1^i, \ldots, \hat{\eta}_{j-1}^i, \eta_j^i, \hat{\eta}_{j+1}^i, \ldots, \hat{\eta}_p^i, \hat{\theta}_N) \\
&\quad = \frac{1}{Nh_j^{d_j}} \sum_{i=1}^{N} K_j\left(\frac{x_0^j - X_i^j}{h_j}\right) \varphi_j^{(1)}(Y_i, T_i; \eta_1^i, \ldots, \eta_{j-1}^i, \eta_j^i, \eta_{j+1}^i, \ldots, \eta_p^i, \theta_0) \\
&\qquad + \frac{1}{Nh_j^{d_j}} \sum_{i=1}^{N} K_j\left(\frac{x_0^j - X_i^j}{h_j}\right) \sum_{l \neq j}^{p} \frac{\partial}{\partial \eta_l} \varphi_l^{(1)}(Y_i, T_i; \bar{\eta}_1^i, \ldots, \bar{\eta}_{j-1}^i, \eta_l^i, \bar{\eta}_{l+1}^i, \ldots, \bar{\eta}_p^i, \bar{\theta}_N) \\
&\qquad \times (\hat{\eta}_l(X_i^l) - \eta_l(X_i^l)) + \frac{1}{Nh_j^{d_j}} \sum_{i=1}^{N} K_j\left(\frac{x_0^j - X_i^j}{h_j}\right) \frac{\partial}{\partial \theta^T} \\
&\qquad \times \varphi_j^{(1)}(Y_i, T_i; \bar{\eta}_1^i, \ldots, \bar{\eta}_{j-1}^i, \eta_j^i, \bar{\eta}_{j+1}^i, \ldots, \bar{\eta}_p^i, \bar{\theta}_N)(\hat{\theta}_N - \theta_0).
\end{aligned}$$

By using the following results from Theorem 1:

$$\sup_{x_0^j \in \mathcal{X}_{d_j}} |\hat{\hat{\eta}}_j(x_0^j) - \eta_j(x_0^j)| = O_p(N^{-1/4}),$$

$$\hat{\theta}_N = \theta_0 + O_p\left(\frac{1}{\sqrt{N}}\right)$$

and a strong law of large numbers we obtain

$$A_1(\hat{\theta}_N) = A_1(\theta_0) + o_p(N^{-1/4}) + O_p\left(\frac{h_j^2}{\sqrt{N}} + \frac{1}{Nh_j^{d_j/2}}\right), \tag{A.3}$$

where

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

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

$$\begin{aligned}
A_1(\theta_0) &\rightarrow_p 0, \\
E(A_1(\theta_0)) &= O(h_j^2), \\
\operatorname{Var}(A_1(\theta_0)) &= \frac{1}{Nh_j^{d_j}}\left[\int K_j^2(t)\, dt I_j(\eta_j^0, \theta_0) p_j^{-1}(x_0^j)\right] + o\left(\frac{1}{nh_j^{d_j}}\right),
\end{aligned} \tag{A.4}$$

and

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

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

$$A_2(\eta_j,\hat{\theta}_N) = A_2(\eta_j,\theta_0) + o_p(N^{-1/4}) + O_p\left(\frac{h_j^2}{\sqrt{N}} + \frac{1}{Nh_j^{d_j/2}}\right), \tag{A.5}$$

where

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

with

$$\begin{aligned} A_2(\eta,\theta_0) &\to_p 0,\\ E(A_2(\eta,\theta_0)) &= o(h_j^2),\\ \mathrm{Var}(A_2(\eta,\theta_0)) &= o(\mathrm{Var}(A_1(\theta_0))), \end{aligned} \tag{A.6}$$

and

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

which finally leads to

$$A_3(\hat{\theta}_N) = -H_j(\eta_j^0,\theta_0) + o_p(1), \tag{A.7}$$

where

$$H_j(\eta_j^0,\theta_0) = E[-\varphi_j^{(2)}(Y,T;\eta_1(X^1),\ldots,\eta_j(X^j),\ldots,\theta_0)|X^j = x_0^j].$$

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

$$A_4(\eta_j,\hat{\theta}_N) \to 0 \tag{A.8}$$

in probability. This convergence is uniform over η_j and θ (because 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 completed as follows. Let us denote

$$Z = \sqrt{Nh_j^{d_j}}(\hat{\eta}_j - \eta_j^0). \tag{A.9}$$

By applying (A.2) at point $\eta_j = \hat{\eta}_j$, we arrive at

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

Using (A.5) we have that

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

Moreover, by expression (A.6)

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

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

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

Now apply (A.3), (A.7), and (A.8) and remark that thus Z has the same distribution as $\sqrt{Nh_j^{d_j}}A_1(\theta_0)/H_j(\eta_j^0,\theta_0)$. On the other hand the Lindeberg–Feller theorem together with (A.4) leads to

$$\sqrt{Nh_j^{d_j}}A_1(\theta_0) \to_d N\left(0, \sqrt{\int K_j^2(t)\,dt\,\frac{I_j(\eta_j^0,\theta_0)}{p_j(x_0^j)}}\right). \qquad \textbf{(A.10)}$$

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

$$\frac{V_j(\widehat{\hat{\eta}}_j)}{V_j(\eta_j^0)} \to_p 1. \qquad \textbf{(A.11)}$$

Finally, because of Slutsky's theorem, (A.11) and (A.10) are enough to prove the result of Theorem 2(i).

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 immediately get the following expression for $A_1(\theta_0)$:

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

Writing S now in the form

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

and using (A.5), (A.6), (A.7), and (A.8) to treat the terms A_2, A_3, and A_4, we get directly

$$S = O_p\left(\sqrt{\frac{\log N}{Nh_j^{d_j}}}\right). \qquad \textbf{(A.12)}$$

Finally, (A.9) and (A.12) are enough to complete the proof of part (ii) of the theorem.