Nonparametric Estimation of Regression Functions under Restrictions on Partial Derivatives[*]

Arie Beresteanu
Duke University

Last version: June 2004, This version: November 2007

Abstract

Economic theory often provides us with qualitative information on the properties of the functions in a model but rarely indicates their explicit functional form. Among these properties one can find monotonicity, concavity and supermodularity, which involve restricting the sign of the regression's partial derivatives. This paper focuses on such restrictions and provides a sieve estimator based on nonparametric least squares. The estimator enjoys three main advantages: it can handle a variety of restrictions, separately or simultaneously; it is easy to implement; and its geometric interpretation highlights the small sample benefits from using prior information on the shape of the regression function. The last is achieved by evaluating the metric entropy of the space of shape-restricted functions. The small sample efficiency gains are approximated.

JEL Classifications: C14, C21.
Keywords: Nonparametric regression, Shape restricted estimation, Sieve method, B-spline wavelets, Metric entropy.

[*] I wish to thank Charles Manski, Rosa Matzkin, Rob Porter and Tom Severini for many helpful comments and discussions. I also benefited from discussions with Han Hong, Ron Gallant, Joel Horowitz, Gautam Tripathi, George Tauchen and Tim Bollerslev as well as from comments I received in numerous seminars and conferences. An early version of this paper was circulated under the title "Nonparametric estimation of supermodular regression functions". All remaining errors are mine.

1 Introduction

Shape restricted estimation involves finding an optimal approximation for a function of economic interest from a family of functions that share a common property or properties. The literature on shape restricted estimation, motivated by economic theory, considers a variety of possible properties including monotonicity, concavity, and supermodularity, each of which involves restricting the sign of the function's partial derivatives. This paper focuses on restrictions on partial derivatives in a regression model and provides a nonparametric sieve estimator that imposes shape restrictions on the regression function. The sieve estimator proposed here is able to take into account a combination of shape restrictions and obtains the optimal rate of convergence.

Incorporating monotonicity into the estimation of regression functions dates back to the literature on isotonic regression. An early exposition of this literature appears in Barlow, Bartholomew, Bremner & Brunk (1972) and later in Robertson, Wright & Dykstra (1988). Consistency of monotonic regression is proved in Hanson, Pledger & Wright (1973) and of concave regression in Hanson & Pledger (1976). Smoothed versions of the estimators in this early literature can be found in Mukerjee (1988) and Mammen (1991). Dykstra (1983) and Goldman & Ruud (1993) provide efficient algorithms to compute these estimators.

The literature on isotonic regression is based on determining the fitted values of the estimator on a finite set of points (usually the observed covariates) and uses a set of inequality constraints to impose restrictions on the value of the regression function at these points. The algorithms used to compute these estimators can be computationally intensive and involve a large set of inequality restrictions and require a special structure of the support.[1] Series estimators provide a convenient alternative to the isotonic regression literature. Gallant (1981, 1982) proposes the Fourier Flexible Form (FFF) estimator which is based on the trigonometric functions base. He identifies the set of restrictions on the coefficients of the FFF expansion that are sufficient to impose convexity. Monotonicity, however, cannot be easily imposed on the estimator. Gallant (1982) discusses additional restrictions like homogeneity

[1] Supermodularity (defined in the next section) requires lattice structure. Data coming from a continuous distribution will not have a lattice structure almost surely. For more discussion on this see Beresteanu (2001).

and homotheticity, which are not covered here. Gallant & Golub (1984) discuss also quasi-convexity for FFF estimators. Especially convenient series estimators are those based on B-splines. This is a local base of functions that produces a piece-wise polynomial spline (see Schumaker (1981) and Chui (1992)). He & Shi (1998) uses B-splines to form a least absolute deviations estimator under monotonicity constraints. Dole (1999) constructs an estimator which can be monotone and concave using least squares and is based on smoothing splines. Both papers deal with the one dimensional covariate case.[2]

This paper generalizes these methods to the family of restrictions on partial derivatives of a possibly multi-dimensional regression function. The estimator proposed in this paper is a series estimator using a B-spline wavelet base of functions. A grid of points is constructed on the support of the covariates. The estimator imposes restrictions on the values of the estimator at the grid points and then uses interpolation to compute the predicted values in any desired point on the support. This yields a finite number of constraints which are translated to linear inequality restrictions on the values of the coefficients of the B-spline wavelet functions. The grid structure solves the problem of imposing complex restrictions on the regression functions in a multidimensional setting.

Using shape restrictions in nonparametric estimation in economics is discussed in Matzkin (1994). She shows that in some models, where there is lack of identification, shape restrictions may have identifying power. This paper, however, focuses on the following nonparametric regression model

$$Y = f(X) + \varepsilon$$

where $E(\varepsilon|X) = 0$ *a.s.* and $f \in \Im$, a family of functions possessing a certain common property. Although identification of f is not an issue, prior information on the regression function is valuable. Section 3.2 shows that incorporating restrictions on partial derivatives in the estimation procedure does not yield a higher rate of convergence. In fact the shape restricted estimator achieves the optimal rate of convergence computed by Stone (1980) as the estimator that ignores the prior information would achieve. The benefits of shape-restricted

[2] Dole (1999) applies his method in a semi-linear setting where the nonparametric part is one dimensional and is restricted to be monotone and concave. I discuss the semi-linear model in Section 4.1.

estimation show in the small sample properties of the estimator. The expected distance between the estimator and the true regression function shrinks as the number of observations grows:

$$E\left|\left|\hat{f}_n - f\right|\right| \leq Cn^{-r}$$

where n is the number of observations and $\hat{f}_n$ is some optimal nonparametric estimator based on a sample of n observations. Section 3.2 shows that the constant C can be significantly reduced if the estimator takes into account the prior information. This will result in a better performance of the estimator in small samples. The estimator proposed in this paper is a solution to a quadratic programming problem. This formulation of the estimator yields a geometric interpretation of the estimation problem that allows us to quantify the reduction of the constant. Section 3.3 shows that these gains can be substantial. The Monte-Carlo experiment described in Section 5 supports these results.

The estimator proposed here enjoys two additional advantages. First, the estimator can handle a variety of different restrictions with a multidimensional covariate. Any restriction on the regression function that involves signing a partial derivative of any order can be treated by this method. It is also possible to allow a number of such restrictions to hold simultaneously. For example, a regression function can be constrained to be both monotone and supermodular using the same technique. Second, the estimator is easy to implement since it is a solution of a quadratic programing problem with linear inequality constraints. The quadratic programing problem is easy to set and exact formulas for the matrices and vectors in this quadratic problem are given.

For tractability, the discussion about multi-dimensional covariates uses the two dimensional case. All results can be extended to a higher dimensional case but not without a considerable technical effort. In shape-restricted estimation, the curse of dimensionality has an additional effect: the number of constraints, needed to assure that the estimator satisfies certain restrictions on partial derivatives, increases with dimensionality. An increase in the number of constraints in addition to the usual curse of dimensionality can make the problem computationally cumbersome.

The remainder of the paper is structured as follows. Section 2 lays the foundations for estimation under restrictions on partial derivatives. Section 3 describes the asymptotic properties of the proposed estimator and the gains from using prior information on the shape of the regression function. Section 4 discusses two extensions: (1) a semi-linear model with restrictions on the nonparametric part of the model and (2) testing restrictions on partial derivatives in a nonparametric context. Section 5 presents a short Monte Carlo study on the efficiency and rates of convergence of the estimators. Section 7 concludes. Appendix A explains how to construct the B-spline wavelet basis functions used in this paper. Appendix B provides proofs and technical notes and Appendix C summarizes the Monte Carlo results for Section 5.

2 Restrictions on Partial Derivatives and their Difference Analog

The objective of this section is to describe a regression estimator that takes into account a variety of assumptions on the shape of the regression function but does not use functional form assumptions. The focus of this paper is on shape restrictions where the signs of certain partial derivatives of the regression function are determined. Monotonicity, concavity and supermodularity are three examples of such restrictions on partial derivatives. This section focuses on the technical description of the estimator. A discussion on the asymptotic properties is left for the next section.

Consider the following regression model

$$Y = f(X) + \varepsilon, \tag{1}$$

where Y is a random variable, X is a random vector and ε is a random variable satisfying $E(\varepsilon|X) = 0$. The regression function $f(\cdot)$ is assumed to belong to a class of functions, $\Im$, that satisfies certain regulatory conditions. These conditions are further discussed in Section 3.

A least squares estimator of the regression function in (1) is based on the empirical analog of the expected value of a square loss function and is the solution to the following problem

$$\min_{f \in \Im} \frac{1}{n} \sum_{i=1}^{n} (y_i - f(x_i))^2 . \tag{2}$$

Often when the regression function is assumed to belong to a large class of functions,[3] the least squares estimator is inconsistent. The method of sieves, suggested by Genander (1981), proposes the following remedy. Construct a sequence of approximating spaces $\Im_1, \Im_2, ...$ called a sieve. When given a finite sample of size n, perform the optimization in (2) using $\Im_n$ instead of $\Im$:

$$\min_{f \in \Im_n} \frac{1}{n} \sum_{i=1}^{n} (y_i - \hat{f}(x_i))^2 . \tag{3}$$

Genander (1981) shows that controlling the rate at which the sieve sequence converges to $\Im$ provides a consistent estimator. Shen & Wong (1994) show that an appropriate choice of this rate leads to an estimator that achieves an optimal rate of convergence. The sieve method is also a convenient framework for taking into account shape restrictions on the estimator. This flexibility is demonstrated in this section.

Our task in this section is to build a sieve sequence that satisfies the desired shape restriction. We proceed in three steps. First, shape restrictions are translated into linear inequality constraints to be imposed on the values the estimator takes on a (pre-chosen) equidistant grid. Second, various interpolation schemes are discussed. Each interpolation scheme is represented as a series estimator where each interpolation scheme means a different choice of a function base for the series expansion. The outcome of this stage is a curve that is defined on the support of the covariates and that satisfies the shape restrictions at least on the equidistant grid. Finally, shape preserving interpolation and shape preserving function bases are given. This step assures that the estimator satisfies the required restrictions globally on the support.

The following notations are used throughout the paper. The support of the marginal distribution of the covariate X is $S = [0, 1]$ in the univariate case and $S = [0, 1]^2$ in the

[3] Masures for the size of functions spaces are described in Section 3.3.

bivariate case.[4] All vectors are row vectors. Let $\Im$ the space of all continuous functions $f : S \rightarrow \Re^d$. For any vector $v = (v_1, ..., v_m)$ with entries from S and a function $f : S \rightarrow \Re$ define $f(v) = (f(v_1), ..., f(v_m))$ and similarly for a matrix $A = (a_{ij})_{i=1..m,j=1..n}$ and $f : S \rightarrow \Re$, define $f(A) = (f(a_{ij}))_{i=1..m,j=1..n}$. Furthermore, for a vector v, $v \leq 0$ means coordinate wise and for a matrix A, $A \leq 0$ means cell wise. Finally, $\otimes$ is the Kronecker product for matrices and for a matrix A of size $k \times l$, $vec(A) = (a_{11}, ..., a_{1l}, ..., a_{k1}, ..., a_{kl})$.

2.1 Imposing shape restrictions on an equidistant grid

Properties like monotonicity, concavity and supermodularity can be defined for non-differentiable functions as well. In this section, I define and use a difference analogue of a partial derivative. For notational convenience as well as for practical reasons, the support of each covariate is assumed to be $[0, 1]$. In what follows, this support is divided into equal parts and the desired restrictions are imposed on this discrete set of points.

Definition 1 *A vector* $\Gamma_m = (\gamma_0, ..., \gamma_m)$ *of length* $m + 1$ *is called a* ***grid vector*** *on* $[0, 1]$ *if* $0 \leq \gamma_0 < \gamma_1 < ... < \gamma_m \leq 1$. *It is called an* ***equidistant grid vector*** *if also* $\gamma_i - \gamma_{i-1} = \gamma_j - \gamma_{j-1}$ *for all* $1 \leq i, j \leq m$. *We denote by* $\bar{\Gamma}_m$ *the equidistant grid vector* $\left(0, \frac{1}{m}, \frac{2}{m}, .., 1\right)$.

In this paper the regression function $f(X)$ in (1) is estimated via a series expansion based on local functions defined on a grid of points. In definition 1 and through out the paper when we refer to equidistant grid vectors, sub-index m sets the mesh of the grid on which our series of functions is defined and determines the accuracy of the approximations.

Monotonicity can be written in terms of non-negative first differences as follows.

Definition 2 *A function* $f \in \Im$ *has a* ***non-negative first difference*** *if for any grid vector* $\Gamma_1 = (\gamma_0, \gamma_1)$,

$$(-1, 1) \cdot f(\Gamma_1)' \geq 0\,.$$

Non-negative first difference is just another name for monotonicity: $v_0 \leq v_1 \Rightarrow f(v_1) - f(v_0) \geq 0$. Convexity constrains the second difference: $v_0 < v_1 < v_2 \Rightarrow [f(v_2) - f(v_1)] - [f(v_1) - f(v_0)] \geq 0$.

[4] Some departures from the $[0, 1]^2$ support are discussed in section 2.4.4.

Definition 3 *A function $f \in \Im$ has a* ***non-negative second difference*** *if for any equidistant grid vector $\Gamma_2 = (\gamma_0, \gamma_1, \gamma_2)$,*

$$(-1,1) \cdot \begin{pmatrix} -1 & 1 & 0 \\ 0 & -1 & 1 \end{pmatrix} \cdot f(\Gamma_2)' \geq 0\,.$$

We now generalize the difference analogue of partial derivatives to any order using differentiation matrices.

Definition 4 *A* ***differentiation matrix*** *of size p is a $p \times (p+1)$ matrix and is denoted by D_p and defined as*

$$D_p = \begin{pmatrix} -1 & 1 & 0 & 0 & \cdots & 0 \\ 0 & -1 & 1 & 0 & \cdots & 0 \\ \vdots & & \ddots & \ddots & & \vdots \\ 0 & \cdots & 0 & -1 & 1 & 0 \\ 0 & \cdots & 0 & 0 & -1 & 1 \end{pmatrix}_{p\times(p+1)}$$

Definition 5 *A function $f \in \Im$ has a* ***non-negative*** *p^{th}* ***difference*** *if for any equidistant grid vector $\Gamma_p = (\gamma_0, \gamma_1, ..., \gamma_p)$,*

$$D_1 \cdot ... \cdot D_p \cdot f(\Gamma_p)' \geq 0\,.$$

Constraining the least squares problem in (3) by $D_1 \cdot ... \cdot D_p \cdot f(\Gamma_p)' \geq 0$ for any possible grid vector Γ_p implies infinite number of constraints. These infinite number of restrictions are impossible to implement with a finite sample. We can circumvent this problem using a sieve estimator. Starting with monotonicity, consider an estimator $\hat{f}$ which is monotone on the equidistant grid $\bar{\Gamma}_m$:

$$\hat{f}(0) \leq \hat{f}(\frac{1}{m}) \leq ... \leq \hat{f}(1). \tag{4}$$

The above implies m inequality constraints to be imposed on the estimator. These restrictions can be written using matrix notations as $D_m \cdot \hat{f}(\bar{\Gamma}_m)' \geq 0$. Similarly, the non-negative second difference (i.e. convexity) implies the following $m-1$ restrictions on $\hat{f}$:

$$\hat{f}(\frac{1}{m}) - \hat{f}(0) \leq \hat{f}(\frac{2}{m}) - \hat{f}(\frac{1}{m}) \leq ... \leq \hat{f}(1) - \hat{f}(\frac{m-1}{m}), \tag{5}$$

or in matrix notations, $D_{m-1} \cdot D_m \cdot \hat{f}(\bar{\Gamma}_m)' \geq 0$. Using the same argument, restricting the p^{th} difference to be non-negative yields $m-p+1$ restrictions[5] that can be written as

$$D_{m-p+1} \cdot \ldots \cdot D_{m-1} \cdot D_m \cdot \hat{f}(\bar{\Gamma}_m)' \geq 0. \tag{6}$$

The two-dimensional case involves additional notations and definitions.

Definition 6 *A matrix $\Gamma_{(m_1,m_2)}$ of dimensions $(m_1+1)\times(m_2+1)$ is called an* ***grid matrix*** *based on the equidistant grid vectors $(\gamma_0, ..., \gamma_{m_1})$ and $(\delta_0, ..., \delta_{m_2})$, if*

$$\Gamma_{(m_1,m_2)} = \begin{pmatrix} (\gamma_0, \delta_0) & \cdots & (\gamma_0, \delta_{m_2}) \\ \vdots & & \vdots \\ (\gamma_{m_1}, \delta_0) & \cdots & (\gamma_{m_1}, \delta_{m_2}) \end{pmatrix}.$$

Let $\bar{\Gamma}_{(m_1,m_2)}$ denotes the ***equidistant grid matrix*** *which is based on the equidistant grid vectors $\bar{\Gamma}_{m_1}$ and $\bar{\Gamma}_{m_2}$.*

Definition 7 *A bivariate function f has a* ***non-negative*** $(1,1)^{th}$ ***difference*** *if for any grid matrix $\Gamma_{(1,1)}$*

$$[(-1,1)\otimes(-1,1)] \cdot vec\left(f(\Gamma_{(1,1)})\right)' \geq 0.$$

Functions with non-negative $(1,1)^{th}$ difference are called supermodular functions.[6] Finally, we extend the above definitions to mixed derivatives of order (p_1, p_2).

Definition 8 *A bivariate function f has a* ***non-negative*** $(p_1,p_2)^{th}$ ***difference*** *if for any equidistant grid matrix $\Gamma_{(p_1,p_2)}$*

$$[(D_1 \cdot ... \cdot D_{p_2}) \otimes (D_1 \cdot ... \cdot D_{p_1})] \cdot vec\left(f\left(\Gamma_{(p_1,p_2)}\right)\right)' \geq 0.$$

Following the arguments leading to (6), an estimator $\hat{f}$ has $(p_1,p_2)^{th}$ negative difference on an equidistant grid matrix $\bar{\Gamma}_{(m_1,m_2)}$ if

$$[(-D_{m_2-p_2+1} \cdot ... \cdot D_{m_2}) \otimes (D_{m_1-p_1+1} \cdot ... \cdot D_{m_1})] \cdot vec(f(\bar{\Gamma}_{(m_1,m_2)}))' \geq 0, \tag{7}$$

[5] Generally, we cannot impose restrictions on a derivative of order higher than the number of segments in the grid. Therefore, we require that $p \leq m$.

[6] Proof: $(-1,1)\otimes(-1,1) = (1,-1,-1,1)$, $vec(f(\Gamma_{(1,1)})) = (f(\gamma_0,\delta_0), f(\gamma_0,\delta_1), f(\gamma_1,\delta_0), f(\gamma_1,\delta_1))$. Therefore, $(-1,1)\otimes(-1,1)\cdot vec(f(\Gamma_{(1,1)}))' \geq 0$ implies $+f(\gamma_0,\delta_0) - f(\gamma_0,\delta_1) - f(\gamma_1,\delta_0) + f(\gamma_1,\delta_1) \geq 0$ which is equivalent to supermodularity of f.

where $D_0 = (1)$. Here (7) involves $(m_1 - p_1)(m_2 - p_2)$ inequalities.

The least squares estimator in (3) can be written as a quadratic programing problem with linear inequality constraints. Restricting the $(p_1, p_2)^{th}$ partial derivative (or its difference analogue) using a grid of mesh $m = (m_1, m_2)$ translates to the following problem:[7]

$$\begin{aligned} &\min_g ||y - Bg||_2 \\ s.t.\quad &A_m^p g \geq 0 \end{aligned} \tag{8}$$

where $g = vec\left(\hat{f}\left(\bar{\Gamma}_m\right)\right)$ is a $(m_1 + 1)(m_2 + 1) \times 1$ vector representing the values that the estimator takes on the grid matrix and

$$A_m^p = ((D_{m_2 - p_2 + 1} \cdot ... \cdot D_{m_2}) \otimes (-D_{m_1 - p_1 + 1} \cdot ... \cdot D_{m_1}))$$

sets the linear inequality constraints assuring that the estimator satisfies the shape restrictions on the grid. B is a $N \times (m_1 + 1)(m_2 + 1)$ matrix of weights depending on the observations and the interpolation scheme that we employ. The next section explains how to construct the weight matrix B.

2.2 Interpolation

A comparison between (3) and (8) reveals that Bg in (8) corresponds to the vector of predicted values on the sample points $(\hat{f}(x_1), ..., \hat{f}(x_n))$ in (3). In other words, if g represents the values that the estimator takes on the equidistant grid, then B represents the way in which these values are weighted to yield the predicted values of the estimator at the sample points $\{x_i\}$. These weights depend on how we interpolate the values that the estimator takes on the grid to the whole domain. The approach here is to represent the interpolation using a series estimator. Thus, different choices of basis functions represent different choices of interpolation schemes.

The series expansion implemented in this paper uses the normalized B-splines base.[8] Generally speaking, normalized B-splines are a base of local functions centered around the

[7] To simplify the discussion on the asymptotic properties of the estimators in subsequent sections, we use the Euclidean norm in (8). The Euclidean norm can be replaced with the absolute norm which leads to a constrained LAD estimator. If one is willing to asume a specific distribution for the error term in the regression function, a constrained maximum likelihood estimator is feasible here. Finally, weighted versions of these estimators can be considered as well.

[8] For a comprehensive discussion of B-splines and alternative series estimators see Chen (2007).

grid points. As is shown below, using this base yields piece-wise polynomial functions whose properties are controlled by the degree of smoothness of the functions composing the B-spline base and by the coefficients put on this base. In this section we look first at the second factor - the coefficients of the expansion. In order to simplify the discussion, in this section, I fix the base of normalized B-splines corresponding to piecewise linear splines defined on the equidistant grid vector $\bar{\Gamma}_m$. The role of the smoothness of the base functions and B-splines of higher degree of smoothness is discussed in the next section.

We denote the base of piece-wise linear functions on $\bar{\Gamma}_m$ by Ψ_m. Starting with the univariate case, let the base of functions be composed of the following $m+1$ functions

$$\Psi_m = \left[\psi_{m,0}, \psi_{m,1}, ..., \psi_{m,m}\right] \tag{9}$$

where

$$\psi_{m,j}(x) = \begin{cases} 1 - |mx - j| & x \in \left[\frac{j-1}{m}, \frac{j+1}{m}\right] \cap [0,1] \\ 0 & \text{otherwise} \end{cases} \tag{10}$$

for $j = 0, ..., m$. This is simply a series of triangular kernel functions (see Figure 1). The next section discusses other choices of base functions as well as the choice of m. The set of all possible (linear) expansions based on Ψ is[9]

$$\Im\left(\Psi_m\right) = \left\{ f\left(x\right) = \sum_{i=0}^{m} \theta_i \psi_i\left(x\right) : \theta \in \Re^{m+1} \right\} \tag{11}$$

where θ is $m+1$ column vector of coefficients.

In the two-dimensional case define $\Psi = vec(\Psi'_{m_1}\Psi_{m_2})$ to be the tensor product of the two bases Ψ_{m_1} and Ψ_{m_2}. Therefore, for $m = (m_1, m_2)$, the piecewise linear spline base in two dimensions is

$$\begin{aligned} &\Psi_m\left(x^1, x^2\right) = \\ &\quad [\psi_{m_1,0}\left(x^1\right)\psi_{m_2,0}(x^2), ..., \psi_{m_1,0}(x^1)\psi_{m_2,m_2}(x^2), \\ &\quad \psi_{m_1,1}(x^1)\psi_{m_2,0}(x^2), ..., \psi_{m_1,1}(x^1)\psi_{m_2,m_2}(x^2), \\ &\quad ... \\ &\quad \psi_{m_1,m_1}(x^1)\psi_{m_2,0}(x^2), ..., \psi_{m_1,m_1}(x^1)\psi_{m_2,m_2}(x^2)] \end{aligned} . \tag{12}$$

[9] Note that the sieve sequence $\{\Im_m(\Psi_m(x))\}_{m=1}^{\infty}$ is a non-nested sequence in the sense that $\Im_m(\Psi_m(x)) \nsubseteq \Im_{m+1}(\Psi_{m+1}(x))$.

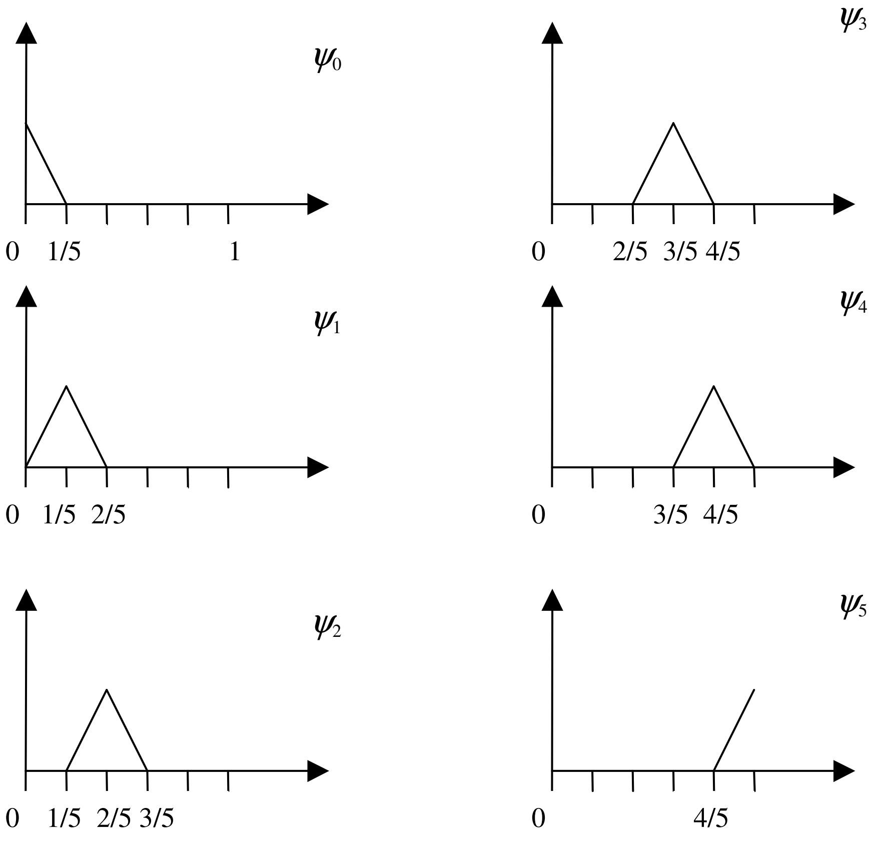

Figure 1: Ψ_5 corresponding to the piecewise-linear spline estimator

Meaning that Ψ_m is the $(m_1+1)(m_2+1)$ vector of base functions that consists of all the possible multiplications of functions from $\Psi_{m_1}(x^1)$ and from $\Psi_{m_2}(x^2)$. The set of all possible finite expansions is

$$\Im(\Psi_m) = \left\{ f\left(x^1,x^2\right) = \sum_{i=0}^{m_1}\sum_{j=1}^{m_2} \theta_{ij}\psi_{m_1,i}\left(x^1\right)\psi_{m_2,j}\left(x^2\right) : \theta \in \Re^{(m_1+1)(m_2+1)} \right\} \tag{13}$$

where θ is $(m_1+1)(m_2+1)$ column vector of coefficients

The least squares estimator based on the functional expansion Ψ_m is:

$$\begin{array}{rc} & \min\limits_{\theta}\frac{1}{N}\sum_{i=1}^{N}\left(y_i - \Psi_m(x_i^1,x_i^2)\theta\right)^2 \\ s.t. & \\ & A_m^p\Psi_m(vec(\bar{\Gamma}_m))\theta \geq 0 \end{array} \tag{14}$$

where $\bar{\Gamma}_m$ is the $(m_1+1)\times(m_2+1)$ equally spaced grid matrix and $\Psi_m(vec(\bar{\Gamma}_m))$ is the $(m_1+1)(m_2+1)\times(m_1+1)(m_2+1)$ matrix whose columns are the values of the vector of functions Ψ_m evaluated at the grid points. $\Psi_m(x_i^1,x_i^2)$ is the $N\times(m_1+1)(m_2+1)$ matrix of base functions evaluated at the observations and therefore $\Psi_m(x_i^1,x_i^2)\theta$ represents the expansion based on the base Ψ_m evaluated at the observation points. In other words $\hat{f}\left(x_i^1,x_i^2\right) = \Psi_m(x_i^1,x_i^2)\theta$.

Any function base other than Ψ_m can fit into the framework described above. However, the B-spline yielding the piecewise linear spline simplifies the estimator in (14). In this specific case $\Psi_m(vec(\bar{\Gamma}_m))$ is equal to the identity matrix and therefore θ represents the values that the estimator takes on the grid points. Thus, the set of constraints in (14) can be simply written as $A_m^p\theta' \geq 0$.

2.3 Shape preserving Interpolation

We are now ready to generalize the discussion on B-splines to smoother bases. Let $\Im^p$ be the set of functions in $\Im$ that satisfy p^{th} non-negative difference everywhere. Let $\Im^p(\Psi_m)$ be the set of functions satisfying the p^{th} non-negative difference restrictions on the equidistant grid $\bar{\Gamma}_m$ and use the interpolation scheme induced by the function basis Ψ_m. This section identifies a basis Ψ_m such that $\Im^p(\Psi_m) \subset \Im^p$. In other words, a basis Ψ_m which induces a shape preserving interpolation.

The basis of functions used in the previous two sections yields a piecewise linear approximation. A greater degree of smoothness can be achieved using B-splines based on polynomials of higher degree. These B-spline functions are defined in Appendix A (see also Schumaker (1981, section 4.4)). The degree of the B-spline function is denoted by l and is assumed to be an even number such that $l \geq 2$. Using B-splines of higher degree requires extending the grid points beyond the support of the function.

Definition 9 *We denote by* $\bar{\Gamma}^l_m = \left(\frac{-l/2+1}{m}, ...\frac{-1}{m}, 0, \frac{1}{m}, ..., \frac{m-1}{m}, 1, \frac{m+1}{m}, ..\frac{m+l/2-1}{m}\right)$ *the* ***extended equidistant grid vector*** *of length* $m + l - 1$. *The* ***extended equidistant grid matrix*** $\bar{\Gamma}^{(l_1,l_2)}_{(m_1,m_2)}$ *is based on the extended equidistant grid vectors* $\bar{\Gamma}^{l_1}_{m_1}$ *and* $\bar{\Gamma}^{l_2}_{m_2}$ *and has dimensions* $(m_1 + l_1 - 1) \times (m_2 + l_2 - 1)$.

Note that when $l = 2$ the extended equidistant grid vector becomes the regular equidistant grid vector, i.e. $\bar{\Gamma}^2_m = \bar{\Gamma}_m$.

Let Ψ^l_m be the basis functions of normalized B-splines of degree $l \geq 2$ centered around the extended equidistant grid $\bar{\Gamma}^l_m$. For example, the basis defined in (10) is a normalized B-splines of degree two and Figure 1 depicts Ψ^2_5. The following theorem specifies the conditions under which the basis Ψ^l_m is shape preserving.

Theorem 1 *For any* $m = (m_1, m_2)$ *and* $p = (p_1, p_2)$ *such that* $m_1 > p_1$ *and* $m_2 > p_2$. *Let* $\Im^p_m(\Psi^l_m) = \left\{\theta'\Psi_m : A^p_m\theta'\Psi^l_m(vec(\bar{\Gamma}^{l-2}_m)) \geq 0\right\}$, *if* $m_1 > l_1 \geq p_1$ *and* $m_2 > l_2 \geq p_2$ *then* $\Im^p_m\left(\Psi^l_m\right) \subset \Im^p$.

The proof appears in Appendix B.

The piece-wise linear spline based on Ψ^2_m is sufficient for imposing restrictions like monotonicity, concavity or supermodularity. The higher order B-spline bases are needed if one has to impose restrictions on derivatives of order higher than two or if one wants the estimator to be smoother than piece-wise linear. Section 3.2 shows that higher order B-splines are also needed in order to achieve higher rates of convergence if the regression function is known to belong to a class of highly differentiable functions (see Theorem 3 bellow).

2.4 Practical Implementation of the Estimator

This section discusses the computational aspects of finding the solution to the quadratic programming problem in (14). A unique solution to this problem is shown to exist. Combining several shape restrictions and extrapolation are discussed next.

2.4.1 Solvability

Suppose we are given a sample (y, x), where y is a $N \times 1$ vector, x is a $N \times k$ matrix of covariates with $k \in \{1, 2\}$. Fix m and l and let D be the number of grid points in the equidistant grid vector or matrix.[10] The quadratic programing problem in (14) can be written as

$$\begin{array}{c} \min_{\theta \in \Re^D} \theta' H \theta + 2 f \theta \\ s.t \\ R\theta \leq 0 \end{array} \tag{15}$$

where $f = -y'\Psi^l_m(x)$, $H = \Psi^l_m(x)'\Psi^l_m(x)$ and $R = -A^p_m \Psi^l_m(vec(\bar{\Gamma}^l_m))$. This quadratic programing problem has a unique solution if H is a positive definite matrix. This is true if $\Psi^l_m(x)$ is a full rank matrix. To satisfy this condition we need that the sample is sufficiently spread over the whole support set $[0, 1]^k$. In section 2.4.4 I discuss the case where $\Psi^l_m(x)$ is not a full rank matrix and suggest solutions. In the rest of this section I assume that H is indeed a positive definite matrix. Algorithms for solving the quadratic programing problem (15) are well known (e.g. Luenberger (1984, Chapter 14)). The problem, however, can be computationally intensive as D and the number of rows in R gets bigger. An algorithm for this problem when D and R are large is discussed in Goldman & Ruud (1993).

The set $\{\theta : R\theta \leq 0\}$ is an unbounded set. It can be useful to establish upper and lower bounds on the values that θ should take. The following theorem extends the results from Robertson et al. (1988) Theorems 1.3.1 to 1.3.4 and provides such bounds under some conditions on the constraints matrix R.

Theorem 2 *Let $G = \left\{\theta \in \Re^D : R\theta \leq 0\right\}$. If G is a* ***sub-lattice*** *(i.e. $\forall \theta_1, \theta_2 \in G$ also $\theta_1 \vee \theta_2 \in G$ and $\theta_1 \wedge \theta_2 \in G$ where $\vee$ and $\wedge$ are taken point-wise) that contains the point*

[10] $D = m + l - 1$ in the one dimensional case and $D = (m_1 + l_1 - 1)(m_2 + l_2 - 1)$ in the two dimensional case.

$(1, 1, .., 1)$ *then there is a unique solution* $\theta^* \in G$ *to (15) such that each coordinate of* θ^* *is bounded between* $\min(Y)$ *and* $\max(Y)$.

Theorem 2 is useful in two important cases. The first is under monotone restrictions. If two functions are monotone then so are the point-wise minimum and maximum of these two functions. The second case in which this is true is under supermodularity restrictions (i.e. when $p = (1, 1)$). In other words, under either monotonicity or supermodularity, the set of functions is a sub-lattice. It is also important to note that if the B-splines are of order 2, the coefficients vector θ coincides with the function itself, evaluated at the grid points. Therefore, a lattice structure of the functions implies that G is a lattice as well. This point becomes clearer in the following section.

2.4.2 Geometric Interpretation

The proof for Theorem 2 (in Appendix B) provides an interesting insight on the geometry of the problem. The set G can be looked at as the dual space for $\Im_m^p(\Psi_m^l)$. Formally, the dual space of a function space $\Im$ and a grid vector or matrix $\Gamma = (\gamma_1, ..., \gamma_D)$ is defined as

$$G(\Gamma, \Im) = \left\{\theta \in \Re^D : \exists f \in \Im \text{ such that } \theta = (f(\gamma_1), ..., f(\gamma_D))\right\}.$$

The proof for Theorem (2) uses the fact that the properties of $\Im$, namely the sub-lattice structure, are transferred to its dual space. Since the problem in (15) can be written as finding the closest point in G to the point $y = (y_1, ..., y_N)$ with respect to the Euclidean distance we can use the properties of G to draw conclusions on the solution θ^*. The dual space of $\Im_m^p$ is $G(\bar{\Gamma}_m, \Im_m^p) = \left\{\theta \in \Re^D : -A_m^p \Psi_m^l(vec(\bar{\Gamma}_m))\theta \leq 0\right\}$. This space is a polyhedron in $\Re^D$. This interpretation is useful when we discuss the small sample efficiency gained from using prior information on the shape of the regression function in section 3.3.

2.4.3 Combining several shape restrictions

The estimator in (14) takes into account only one shape restriction on the regression function. Situations where the regression function is assumed to satisfy more then one shape restriction are common. Beresteanu (2005) estimates a cost function assuming both monotonicity and submodularity[11] with respect to the outputs produced. In other words, we assume that $f \in \Im^{1,0} \cap \Im^{0,1} \cap -\Im^{1,1}$. The estimator in (14) can be easily modified to take into account a combination of several restrictions. For monotonicity and submodularity the constraints in (14) are written as

$$\begin{bmatrix} A_m^{1,0} \\ A_m^{0,1} \\ -A_m^{1,1} \end{bmatrix} \Psi_m \left(vec \left(\bar{\Gamma}_m^l \right) \right) \theta \geq 0.$$

Another restriction that can be added to the above shape restriction is a restriction on the variation of the function.[12] For example, say all the derivatives up to an order k are assumed to be bounded in the interval $[L, U]$. This implies $A^p g \geq L$ and $-A^p g \geq 0$ for all $p \leq k$ in the one dimensional covariate case. In the multidimensional case this implies $A^{p_1,p_2} g \geq L$ and $-A^{p_1,p_2} g \geq U$ for all integers p_1, p_2 such that $p_1 + p_2 \leq k$. The intersection of convex cones is a convex cone (Rockafellar (1970, Theorem 2.5)). Therefore, the discussion in Section 2.4.1 is relevant to the case where several restrictions are combined together as well.

2.4.4 Extrapolation

In the one dimensional covariate case the sample can always be linearly transformed to cover the interval $[0, 1]$. In the two dimensional case, however, we can have samples that do not cover the whole $[0, 1]^2$ box. Beresteanu (2005) examines the cost function of local telephone companies in the U.S. The cost function is expressed as a function of two outputs: local and toll calls. It is evident from Figure 2 that the support of the covariates $X =$(log local calls, log toll calls) is not the whole set $[0, 1]^2$. This means that for our choice of grid mesh m, one can find a box $\left[\frac{j_1}{m_1}, \frac{j_1+1}{m_1}\right] \times \left[\frac{j_2}{m_2}, \frac{j_2+1}{m_2}\right]$ such that no observation point is contained in it. The result is that the matrix H in (15) is not positive definite since it has a zero rows and columns in it. Furthermore, the constraints matrix R in (15) will have zero rows in it as well which puts no constrain on some of the parameters.

Whether empty boxes occur obviously depends on the mesh of the grid we choose. However, in some applications, as in the one depicted in Figure 2, any non trivial grid choice

[11] A function f is submodular if $-f$ is supermodular.

[12] This fits the restriction $\left| f^{(q)} \right| \leq L$ used in the next section to make sure that the function space is compact.

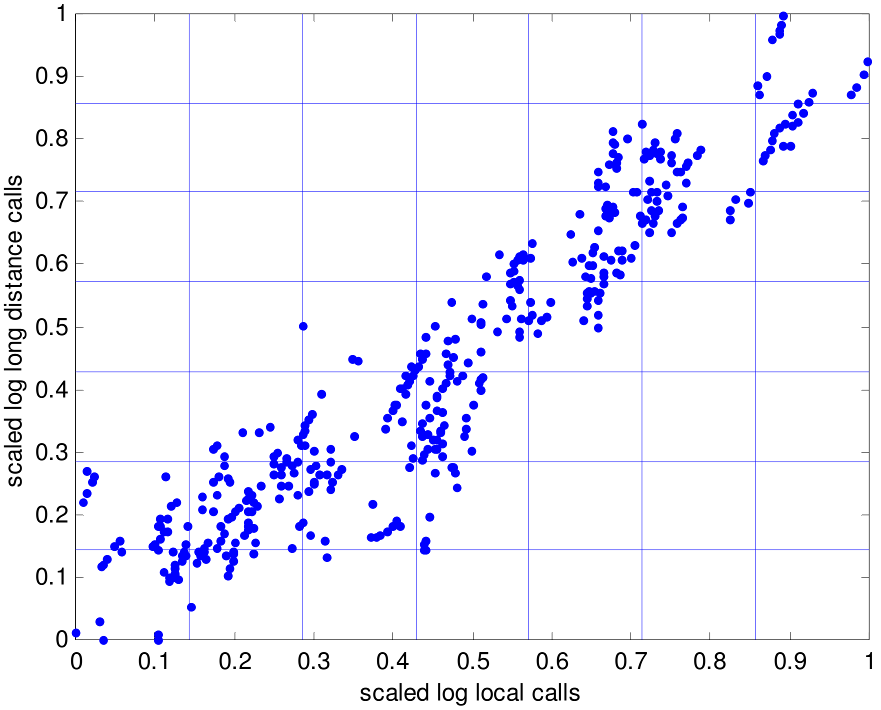

Figure 2: The support of outputs produced by local telephone companies and the estimation grid

will lead to zero rows in H and R of (15). Therefore, driving down the mesh of the grid will have no impact on the problem and will cause over-smoothing. An alternative solution is to redefine the support $[0,1]^2$ such that the problem will not occur. Beresteanu (2005) imposes monotonicity and submodularity only for boxes that contain at least one observation. According to this solution we use the same problem as in (8) but use only the functions in Ψ_m^l that are non zero for at least one observation in the sample. Let $\tilde{\Psi}_m^l = \left\{\psi \in \Psi_m^l : \psi\left(x_i\right) > 0 \text{ for some } x_i \text{ in the sample}\right\}$ and denote by $\tilde{\theta}$ the coefficients in θ that correspond to functions in $\tilde{\Psi}_m^l$. Denote by $\tilde{\Gamma}_m^{l-2}$ the points in $\bar{\Gamma}_m^{l-2}$ that correspond to corners of boxes that contain at least one observation and rewrite (15) using $\tilde{f} = -y'\tilde{\Psi}_m^l(x)$, $\tilde{H} = \tilde{\Psi}_m^l(x)'\tilde{\Psi}_m^l(x)$ and $\tilde{R} = -A_m^p\tilde{\Psi}_m^l(vec(\tilde{\Gamma}_m^{l-2}))$. This process of trimming out empty boxes and constraints gives a quadratic programing problem with a unique solution.

3 Asymptotic Properties

A large body of literature is devoted to establishing the consistency of spline and series estimators and to the rate at which they converge. Most of this literature treats the case where no shape restrictions are imposed on the estimator. Therefore, before using any of the results from this literature, one has to make sure that they remain intact under shape restrictions. The results presented in this section show that the shape-restricted estimator described above is consistent and that it achieves the same optimal rate of convergence as in the unrestricted estimation problem. In the last part of this section I discuss the merits of using prior information on the shape of the regression function. A measure of the information embodied in the shape restrictions is constructed.

To avoid cluttering the discussion with complicated notations, I focus on the one dimensional covariate case. Let $\Im$ be the following set of functions

$$\Im = \left\{ f \in C^q(S) : \left|\left|f^{(j)}\right|\right|_{\text{sup}} < \infty, j = 0, ..., q, \left|f^{(q)}(x_1) - f^{(q)}(x_2)\right| \leq L \right\} \tag{16}$$

where q is an integer and L is a known scalar.[13] These assumptions on the parameter space $\Im$ assure that it is compact. As before, $\Im_m(\Psi_m^l)$ denotes the set of functions defined by normalized B-splines of degree l on the equidistant grid $\bar{\Gamma}_m$ without any shape restrictions and $\Im_m^p(\Psi_m^l)$ denotes the subset of functions with p^{th} nonnegative difference. Consistency, rate of convergence and optimal mesh selection depends on the result described below.

For a function $f \in \Im$ and for an arbitrary set of functions $\Im' \subset \Im$ we define $d(f, \Im')_\infty = \inf_{g \in \Im'} ||f - g||_\infty$ to be the distance between a function f and the functions space $\Im'$. Schumaker (1981, Theorem 6.27) shows that for $f \in \Im$, $d(f, \Im_m\left(\Psi_m^l\right)) \leq Cm^{-q}$ for some constant $C > 0$. The theorem below extends this result for $f \in \Im^p$ and $d(f, \Im_m^p(\Psi_m^l))$.

Theorem 3 *Let $f \in \Im^p$ and let $k = \min(l-2, p)$ then $d(f, \Im_m^p(\Psi_p^l)) \leq m^{-k}$ and $\exists \tilde{f} \in \Im^p$ such that $d(\tilde{f}, \Im_m^p(\Psi_p^l)) \geq \tilde{C} m^{-k}$ for some constant $0 < \tilde{C} < 1$.*

[13] This is assumed for simplicity. Shen & Wong (1994) discuss a more general case where the condition is $\left|f^{(q)}(x_1) - f^{(q)}(x_2)\right| \leq L\,||x_1 - x_2||^\alpha$ for $\alpha > 0$ and L is unknown. Unknown L requires using a bound l_n that increases with the number of observations when building a sieve. Shen & Wong (1994) discuss the rate at which l_n should grow. In case L is unknown, putting no restriction on $\left|f^{(q)}(x_1) - f^{(q)}(x_2)\right|$ can lead to a suboptimal rate of convergence (a simmilar claim appears in Birge & Massart (1993)).

Proof. *See appendix B.* ■

Theorem 3 show that the approximation error shrinks to zero as m goes to infinity. Moreover, the rate at which the approximation goes to infinity is exactly $k = \min(l-2, p)$. A function which is p times differentiable can be approximated in an order m^{-p} at most. However, this rate is not achieved unless a smooth enough base is used (i.e. we choose $l \geq p+2$). In the next sections Theorem 3 is used to investigate the behavior of

$$E \int \left(\hat{f}_n(x) - f(x)\right)^2 dP(x) \tag{17}$$

where the expectation is taken over the joint distribution of (X, Y) and $P(x)$ is the marginal distribution of the covariates.

3.1 Consistency

Consistency means that the expression in (17) converges to zero as n goes to infinity. This result was first proved in Genander (1981) and in Geman & Hwang (1982).[14] They treat general parameter spaces which include the shape restricted estimator discussed in this paper. To avoid introducing more notations and repeating much of the discussion in Genander (1981) and in Geman & Hwang (1982), the following arguments about the applicability of their consistency result to our case are made.

The first condition for consistency is that the sieve is dense in the parameter space $\Im$. This is proved in Theorem 3 above. Finally, as Geman & Hwang (1982) show, the least square criteria function satisfies the rest of the requirements in their Theorem 1. Hence, consistency applies in the constrained case.

3.2 Rate of Convergence

Nonparametric estimators suffer from low rates of convergence. The less assumed about the smoothness class of the regression function and the higher the dimension of the covariate

[14] Consistency of sieve estimators in general parameter spaces is brought in Theorem 1 in section 9.3 of Genander (1981) and in Theorem 1 in Geman & Hwang (1982).

space, the lower the convergence rate.[15] Higher rates of convergence can be achieved by either reducing the number of variables that are modeled nonparametrically (e.g. using a semiparametric model) or by assuming that the regression function is smoother (e.g. by assuming a specific functional form). Both approaches are, usually, ad-hoc and lack a solid economic justification. It is reasonable to ask if incorporating prior information on the shape of the regression function into the estimation can reduce the rate of convergence and by that serve as a more appealing solution to the problem of slow rates of convergence. The results presented below give a negative answer to this question. This result is in line with similar results for shape restricted estimation of density functions in Kiefer (1982).

An estimator $\hat{f}_n$, based on a sample of size n, achieves the rate of convergence r if

$$\lim_{n\to\infty} n^{2r} \sup_{f\in\Im} E \int_S \left[\hat{f}_n(x) - f(x)\right]^2 dP(x) \geq C \tag{18}$$

where S is the support of the covariates as before and C is some constant depending on $\Im$.[16] Stone (1980, 1982) showed that if the dimension of X is d and f has q continuous derivatives then $r = \frac{q}{2q+d}$. Shen & Wong (1994) discuss the rate of convergence of the B-spline sieve with no shape restrictions. They show that the convergence rate of this sieve estimator is the optimal one. In this section I show that these results follow to the constraint case. The first discussion on the impact of shape restrictions on the rate of convergence appears, to the best of my knowledge, in Kiefer (1982). He considers various shape restrictions that can be imposed in the nonparametric density estimation context. Kiefer (1982) shows that shape restrictions cannot improve the rate at which the estimator converges to the true parameter. The following theorem states the same result for the shape restricted estimator described in this paper.

[15] For a discussion on rates of convergence for unconstraints sieve estimators see Chen & Shen (1998). The discussion there allows weakly dependant data (see definitions in the source).

[16] If instead we are interested in the supremum norm of $\hat{f}_n - f$, then an additional $\log n$ term is needed and the rate of convergence is r if

$$\lim_{n\to\infty} (n/\log n)^{2r} \sup_{f\in\Im} E_f \sup_{x\in S} \left(\hat{f}_n(x) - f(x)\right)^2 \geq C.$$

Theorem 4 *Let $\Im$ and $\Im^p$ be defined as above. Then, under the assumptions in model 1 of Stone (1980), the optimal rate of convergence is $r = \frac{q}{2q+1}$ for both $\Im^p$ and $\Im$.*

The proof of this theorem appears in Appendix B and follows closely the proof of Theorem 1 in Stone (1980). The intuition of the proof is as follows. Let g_n be a sequence of infinitely differentiable functions with compact support $[0, x_n]$ such that $x_n = Kn^{-\frac{1}{2p+1}}$ for some positive K and that satisfy the Lipschitz condition in (16). Consider the following sequence of perturbations $f_n = f + \varepsilon g_n$, $\varepsilon > 0$. Stone computes the likelihood of distinguishing between f and f_n based on $\{X_i, Y_i\}_{i=1}^n$ and shows that it shrinks to zero at a rate n^{-r}. The proof for Theorem 4 imitates Stone's proof but makes sure that we choose ε small enough such that also $f_n \in \Im^p$.

Our next task is to compute the grid's mesh m that yields the optimal rate of convergence for our estimator. Two factors determine the rate of convergence of a sieve estimator. The first is the rate at which the sieve grows inside the target function space. This element is nonstochastic and depends only on the structure of the chosen sieve. The second is the stochastic element of the problem and depends on the data generating process. The following decomposition of the distance between the estimator and the true regression function demonstrates this argument.

$$\left|\left|\hat{f}_n - f\right|\right|_2 \leq \left|\left|\hat{f}_n - f_n^*\right|\right|_2 + ||f_n^* - f||_2 \tag{19}$$

where

$$\begin{aligned} \hat{f}_n &= \arg\min_{h \in \Im_n} \frac{1}{n} \sum_{i=1}^n (y_i - h(x_i))^2 \\ f_n^* &= \arg\min_{h \in \Im_n} \int (f(x) - h(x))^2 \, dP(x) \end{aligned}$$

and

$$||h||_2 = \left(\int_S h^2(x) dP(x) \right)^{\frac{1}{2}}.$$

The second element on the right hand side of (19) is nonstochastic and measures the distance between the sieve and the true function. This is the best approximation for f using $\Im_n$

instead of $\Im$. This term is called the bias of the sieve estimator. Equation (19) demonstrates the trade-off between bias and variance. If the rate at which the sieve grows is higher, the rate at which the bias reduces is faster. On the other hand, the faster the sieve grows, the slower the variance vanishes. Since the rate at which the estimator converges to the truth is the slower of the two, the optimal rate at which the sieve grows should balance between the "bias" and the "variance".

The rate at which the B-spline sieve grows in the target function space is m^{-q} and a choice of m that equalizes the rate at which the bias decays and the rate at which the variance decays is the optimal choice of mesh m that achieves the optimal rate of convergence. This choice of the optimal sieve mesh is

$$m = Cn^{\frac{1}{2q+1}}$$

for some constant C.[17]

3.3 Efficiency

Kiefer (1982) is also the first to suggest that the constant term in (18) can change as a result of using shape restrictions. Motivated by this conjecture, Birge (1987) shows how to compute a lower and upper bound for the constant for a certain family of unimodal density functions. The constant in (18) depends on the size of the function space measured by its metric entropy (defined below). More precisely, if the regression function comes from a rich family of functions, the constant that we can choose in (18) is large.

Computing the exact value of the minimal convergence constant is an extremely complicated task and so far only bounds on the value of the constant could be calculated for a few specific cases. These calculations are, unfortunately, intractable and case specific. The approach taken here is to evaluate the small sample efficiency gains in terms of reducing $E\left|\left|\hat{f}_n - f\right|\right|$. The discussion makes use of the dual spaces defined in Section 2.4.2.

To facilitate the discussion on efficiency, we formalize how to measure the size of a function space.

[17] The exact formulation of C is not computed here. A cross validation procedure can be used to find C.

Definition 10 *Let $(\mathcal{F}, ||\cdot||)$ be a normed space and $\varepsilon > 0$. A collection of balls of radius ε, $\mathcal{U}$, is called an **ε-covering** of $\mathcal{F}$ if $\cup_{U\in\mathcal{U}}U \supset \mathcal{F}$. $N(\varepsilon, \mathcal{F}, ||\cdot||)$ is called the **minimal ε-covering number** which is the minimal number of elements in an ε-covering.*

If $\mathcal{F}$ is a compact space with respect to the norm $||\cdot||$, then $N(\varepsilon, \mathcal{F}, ||\cdot||)$ is finite for each $\varepsilon > 0$.[18] The rate at which the covering number grows as $\varepsilon \to 0$ plays a significant role in the rate of convergence. Let $\mathcal{F}$ be a generic function space representing either $\Im$ or $\Im^p$ and $\mathcal{F}_1, \mathcal{F}_2, ...$ be the appropriate sieve based on the B-spline basis functions Ψ^l. Before we investigate the impact of $N(\varepsilon, \mathcal{F}, ||\cdot||)$ we look at the following lemma that bounds this covering number. Using the dual space of $\mathcal{F}$ we can write the metric entropy of $\mathcal{F}$ in terms of the Euclidean volume of the dual space. The last is often easier to compute.

Lemma 1 *Let $\mathcal{F}_K = \left\{f \in \Im\left(\Psi_m^l\right) : |f|_\infty \leq K\right\}$ for some $m > 0$ and $0 < l < m$. Let F_K be the dual space of $\mathcal{F}_K$ as defined in Section 2.4.2, then $N\left(\frac{\varepsilon}{2}, \mathcal{F}_K, ||\cdot||_2\right) \leq Vol(F_K)\cdot c_m^{-1}\cdot\varepsilon^{-(m+1)}$ where $Vol(A)$ is the Euclidean volume of set A in $\Re^m$ and c_m is the volume of a ball of radius 1 in $\Re^m$.*

The analysis of the impact of the minimal covering numbers on the constant hinges on the following error decomposition. By Lemma 10.1 in Gyorfi, Kohler, Krzyzak & Walk (2002),

$$\begin{aligned} & \int \left(\hat{f}_n(x) - f(x)\right)^2 dP(x) \\ = & \left\{E\left(\hat{f}_n(X) - Y\right)^2 - E\left(f(X) - Y\right)^2 - 2\frac{1}{n}\sum_{i=1}^n \left[\left(\hat{f}_n(x_i) - y_i\right)^2 - \left(f(x_i) - y_i)^2\right)\right]\right\} \\ & + \left\{2\frac{1}{n}\sum_{i=1}^n \left[\left(\hat{f}_n(x_i) - y_i\right)^2 - \left(f(x_i) - y_i)^2\right)\right]\right\} \\ = & A_n + B_n. \end{aligned}$$

This rewrites the distance between the estimator $\hat{f}_n$ and the true regression function f in terms of distances between the criteria functions. We can bound B_n from above:

$$B_n \leq 2 \inf_{g\in\Im_n} \int (g(x) - f(x))^2 \, dP(x).$$

[18] The sufficient condition for $N(\varepsilon, \mathcal{G}, ||\cdot||)$ to be finite is that $\mathcal{G}$ is totally bounded in the topology induced by the norm $||\cdot||_2$. For more discussion on covering numbers as well as for direct calculation of the covering numbers of some function sapces, see Kolmogorov & Tihomirov (1961).

Thus, B_n is the approximation error of the sieve element $\Im_n$. We first claim that B_n is unaffected by shape restrictions. To see this, consider the case where f belongs to $\Im^p$ and the regression is estimated once under the restriction that $f \in \Im^p$ and once without this restriction. Under both estimators B_n converges at the same rate and with the same constant to zero. This is because $\Im^p \subset \Im$ and thus the infimum in B_n is achieved by picking the best function from $\Im^p$. Therefore, efficiency gain, if it exists, should come from the first element, A_n, which represents the estimation variance. The rate at which A_n converges to zero is discussed in a few sources and under various (alternative) sets of assumptions. They all, however, share the same general result.

$$P\left[A_n > \delta\right] \leq c_1 \exp\left(-c_2 n \delta\right)$$

for some constants c_1,c_2 and for $\delta > \delta_n^*$ where δ^* is such that

$$\delta_n^* \geq n^{-\frac{1}{2}} \left[\int_0^{\sqrt{\delta_n^*}} \sqrt{\log N\left(\varepsilon, \mathcal{F}_{K,n}, ||\cdot||_2\right) d\varepsilon}\right] \tag{20}$$

where $\mathcal{F}_{K,n}$ is the n^{th} element in the sieve sequence approximating $\mathcal{F}_K$. This relation between the metric entropy of the sieve element and the rate at which the estimator converges to the true function appears in a number of sources. Van de Geer (2000) and Van der Vaart & Wellner (1996) demonstrate this relation under the assumptions that the error term in (1) is sub-Gausian and provide somewhat weaker results for error term with a thicker tail distribution. Gyorfi et al. (2002) achieve a similar result under the assumption that $E\left(f^2(X)\right) \leq \tilde{K} E(f(X))$ for some constant $\tilde{K}$. Van der Vaart & Wellner (1996) also show that the requirement of sub-Gausian error term can be lifted if the criteria is absolute deviation instead of least squares. Condition (20) is used in these sources to build a sieve sequence that achieves the optimal rate of convergence reported in the previous section. We will not repeat this discussion here and instead focus on the small sample efficiency gains from shape restrictions.

Since $E\left[A_n\right] = \int_0^\infty P\left(A_n > u\right) du$ we have,

$$E\left[A_n\right] \leq \delta^* + \frac{c_1}{c_2 \cdot n} \exp\left(-c_2 n \delta^*\right). \tag{21}$$

Table 1: Evaluating qualitative restrictions on the regression function

	No assumptions	Supermodularity	Monotonicity	Supermodularity and monotonicity
Volume	1	$1.095 \cdot 10^{-2}$	$1.157 \cdot 10^{-4}$	$2.756 \cdot 10^{-6}$
δ^*	0.05106	0.04170	0.03258	0.02540
efficiency ratio	1	1.22	1.57	2.01
# of restrictions	0	4	12	16

The second term in (21) is negligible and we should choose δ^* such that the rates of convergence of $E\left[A_n\right]$ is the optimal one. The smallest δ^* that can be chosen is such that (20) is satisfied with an equality. As we suggested before, this metric entropy is hard to compute. We use the upper bound on the metric entropy set in Lemma (1) to write (20) as

$$\delta^* = n^{-\frac{1}{2}} \left[\int_0^{\sqrt{\delta^*}} \sqrt{\log\left[Vol(F_{K,n}) \cdot c_{m+1}^{-1} \cdot \varepsilon^{-(m+1)}\right]} d\varepsilon \right]. \tag{22}$$

A numerical solution for δ^* using (22) can be computed if $Vol(F_K)$ can be computed. We turn to this task next.

To illustrate the usage of (22) we look at the quadratic programing problem in (8) with $\Im = C([0,1]^2)$ using a grid with a mesh $m = (2,2)$ and $K = 1$ (i.e. $|f| \leq 1$). The dual space is a subset of $\Re^9$ whose volume depends on the restrictions imposed on the estimator. It is easy to see that with no shape restrictions the volume of the dual space is 1. Adding the assumption of supermodularity turns the dual space to a polyhedral in $\Re^9$ and reduces the volume of the dual space and thus its metric entropy. Lemma 1 associates the metric entropy of the sieve $\Im_m$ to the metric entropy of its dual space. Table 1 reports the volumes of the polyhedrals given different set of assumptions as well as δ^* resulting from (22) and the ration between the δ^* of the restricted models and that of the unrestricted model.[19] We can see that the most substantial restriction is monotonicity in terms of volume reduction. Section 5 compliments the discussion here with a Monte-Carlo experiment.

[19] The volumes of the polyhedrons were computed using a program for polytope volume computation from http://www.math.uni-augsburg.de/~enge/ written by Andreas Enge. The website includes documentation and a Unix code. The number of observations used to compute δ^* in Table 1 is 400 and $c_9 = \frac{32}{945}\pi^4$.

4 Extensions

4.1 Semiparametric models

Nonparametric estimators suffer from the curse of dimensionality. Apart from slow rates of convergence, a large dimension of the covariates vector can make the nonparametric estimator infeasible for small samples. Implementing the estimator described in Section 2 requires setting a grid of points for a highly dimensional covariate vector and also building a restriction matrix that takes into account the various multi-dimensional restrictions. A convenient solution to the curse of dimensionality are the semiparametric methods.

Donald & Newey (1994) consider the following semi-linear model,

$$Y = X'\beta + g(Z) + \varepsilon.$$

A shape restricted estimator in this case is an estimator where $g \in \Im^p$. Donald & Newey (1994) prove that if $p = 0$ (i.e. no shape restrictions) then an estimator for g using B-splines yields a consistent estimator both for g and β and the rate of convergence of the estimator for β is $\sqrt{n}$. When $p > 0$ (i.e. with shape restrictions on g) we need to make sure that the assumptions in their theorems still hold.

We maintain assumptions 1 and 2 in Donald & Newey (1994) and assume that $g \in \Im^p$. Theorem 3 here shows that the approximation power of the B-spline estimator stays intact under shape restriction. Therefore, under the assumptions of Theorems 1 and 2 in Donald & Newey (1994), the shape restricted estimator produces a $\sqrt{n}$-consistent and estimator for β. Moreover, the asymptotic distribution of $\hat{\beta}$ is

$$(23) \qquad \left(A_n^{-1} B_n A_n^{-1}\right)^{-1/2} \sqrt{n}\left(\hat{\beta} - \beta\right) \to N(0, I_q)$$

where q is the dimension of X, $A_n = \frac{1}{n}\sum_{i=1}^{n} u_i u_i'$, $B_n = \frac{1}{n}\sum_{i=1}^{n}\left(\varepsilon_i^2 u_i u_i'\right)$ and $u_i = x_i - E(x_i|z_i)$. A related result appears in Tripathi (2000). He shows that the efficiency of the estimator for β, represented by its asymptotic variance, cannot be improved upon by using an estimator that involves restrictions like $g \in \Im^p$ for $p = 1$ or $p = 2$ (i.e. monotonicity and concavity).[20] Meaning that the asymptotic variance of $\hat{\beta}$ when restrictions on $g(\cdot)$ are taken into account is the same as the asymptotic variance when monotonicity or concavity are not taken into account. The variance presented in (23) supports this argument since it does not depend on the restrictions imposed on $g(\cdot)$.

4.2 Testing

Testing models in a nonparametric environment has received increasing attention in recent years. Most of the literature covers the case of testing parametric models against a nonparametric alternative. These procedures suffers from two main disadvantages. The first is that the null hypothesis is that the regression function belongs to a finite dimensional parameter space. The second is that the nonparametric alternative is totally unspecified. In this section I review a possible extension of the literature based on the above estimator. This testing procedure is based on Hong & White (1995).

The general testing problem is formalized as follows:

$$\begin{aligned} H_0 &: f \in \Im_0 \\ H_1 &: f \in \Im \backslash \Im_0 \end{aligned}$$

where $\Im_0 \subset \Im$.

Hong & White (1995) suggest a testing procedure for the case of parametric $\Im_0$ and nonparametric $\Im$. The testing procedure is based on a sieve estimator for the model under H_1. A sieve sequence $\Im_{1,1}, \Im_{1,2}, ...$ is built such that $\Im_{1,i}$ is a parametric subset of $\Im$. $\hat{f}_1$ is replaced with $\hat{f}_{1,n}$, which is the estimator under $H_{1,n} : f \in \Im_{1,n}$.[21]

A natural extension of Hong & White (1995) for the case where the null is also nonparametric can be based on the following procedure. Build a sieve $\Im_{0,1}, \Im_{0,2}, ...$ that approximates the null space $\Im_0$ in addition to $\Im_{1,1}, \Im_{1,2}, ...$ that approximates the alternative space $\Im$. The

[20] Significant efficiency gains do exist when g is known to be homogeneous. The difference between homogeneity and the restrictions discussed in this paper is that homogeneity reduces the dimensionality of the function g whereas restrictions on partial derivatives do not.

[21] Wooldridge (1992) suggests an alternative procedure using the theory of non-nested tests developed for parametric tests by Davidson & MacKinnon (1981). In his test the sieve is constructed such that $\Im_0$ is not nested in $\Im_{1,n}$.

estimation method proposed in this paper enjoys the advantage that $\Im_{0,i}$ and $\Im_{1,i}$ are parametric and thus the estimation under the null and under the alternative is easy to implement. Furthermore, this testing procedure can take into account maintained assumptions on the model and does not require that the alternative is totally unspecified. In other words, the researcher can assume that the regression function is monotone and test for, say, concavity. In this case the null hypothesis is the set of monotone and concave functions where the alternative includes the monotone (but non-concave) functions. The rate of convergence, the power of the test and many other technical aspects of this problem are yet to be worked out.

A commonly used loss function is based on the L_2 distance between two functions. The test considered here is based on the following statistic,

$$T_n = \frac{\frac{1}{n}\sum_{i=1}^{n}\left(\hat{f}_{0,n}(x_i) - y_i^2\right)^2 - \frac{1}{n}\sum_{i=1}^{n}\left(\hat{f}_{1,n}(x_i) - y_i^2\right)^2}{\frac{1}{n}\sum_{i=1}^{n}\left(\hat{f}_{0,n}(x_i) - y_i^2\right)^2} \tag{24}$$

where $\hat{f}_{0,n}$ is the restricted estimator under the null based on n observations and $\hat{f}_{1,n}$ is the unrestricted estimator from the space $\Im$.

Example 1 *In ? two alternative models are considered. The null hypothesis is that the expected total cost as a function of the outputs, local and toll calls, is both submodular and monotone. The alternative is that the expected total cost is just monotone. The estimated models are described in Figure 3. The piecewise linear estimator with grid $m = (6,6)$ was used. A testing procedure compares the sum of residuals from both estimators using a bootstrap method to compute a confidence interval for this statistic.*

5 Monte Carlo Study

Four B-spline estimators are considered: unrestricted regression, monotone regression, supermodular regression and monotone and supermodular regression. The performance of these estimators for a sample of 400observations is compared through a Monte-Carlo study. The models used in this study are reported in Table 2. All functions are defined on the set $[0,1]^2$ and are both supermodular and monotone. These models are estimated using the four estimators for various joint distributions of the covariates and distributions of the error term.

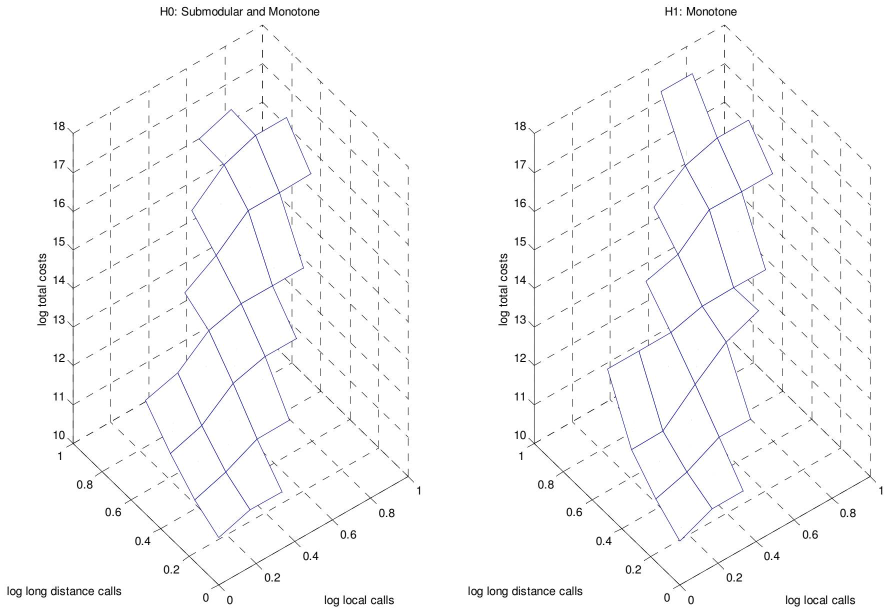

Figure 3: Piecewise-linear estimation of the expected total cost of Local Exchange Companies (LECs) as a function of local and long distance calls

Table 2: Properties of the functions participating in the Monte Carlo study

Function	Properties
$f_1(x_1,x_2)=\min(x_1,x_2)$	Continuous but not differentiable on the 45° line.
$f_2(x_1,x_2)=\begin{cases}(x_1-\frac{1}{2})(x_2-\frac{1}{2})+x_1x_2 & \text{if } x_1\geq\frac{1}{2} \text{ and } x_2\geq\frac{1}{2}\\ x_1x_2 & \text{otherwise}\end{cases}$	Continuous and differentiable but the derivatives are not continuous.
$f_3(x_1,x_2)=x_1^{1/3}x_2^{2/3}$	Infinitely differentiable on $(0,1)^2$ derivatives are not defined on the axis.

These as well as the results are reported in Appendix C. The four estimators were compared based on four criteria: fit at $X=(0,0)$, fit at $X=\left(0,\frac{1}{3}\right)$, fit at $X=\left(\frac{1}{2},\frac{1}{2}\right)$ and over all fit (L_2-distance from the true function). This choice allows examination of local performance on the boundaries of the support and in an interior point and global performance of the estimators. Tables 3, 4 and 5 in Appendix C report the 95% intervals built from these Monte-Carlo experiments.

The results demonstrate the claim that using (correct) prior information on the properties of the regression function improves the small sample performance of the estimator. For example, consider the first model where $Y=\min\left(X_1,X_2\right)+\varepsilon,\ X_1,X_2$ are independent and uniformly distributed on $[0,1]^2$ and $\varepsilon\tilde{}N(0,1)$. In this case $E\left(Y|X=\left(\frac{1}{2},\frac{1}{2}\right)\right)=\frac{1}{2}$. The unrestricted estimator reports 95% of the estimators in the interval $[-0.5236,1.6829]$. Adding the assumption that the regression is supermodular reduces this interval to $[-0.0211,0.8819]$. Assuming monotonicity (but not supermodularity) reduces the interval to $[0.1658,0.6310]$. The monotone regression increases the accuracy in compare to the unrestricted regression by more than the supermodular regression does. Combining both assumptions reduces the interval to $[0.1685,0.6304]$. These results support the claim in Section 3.3 and the results in Table 1.

The next Monte Carlo experiment examines the rate of convergence achieved by the estimators proposed in Section 2.4. Using the models and estimators described above, I

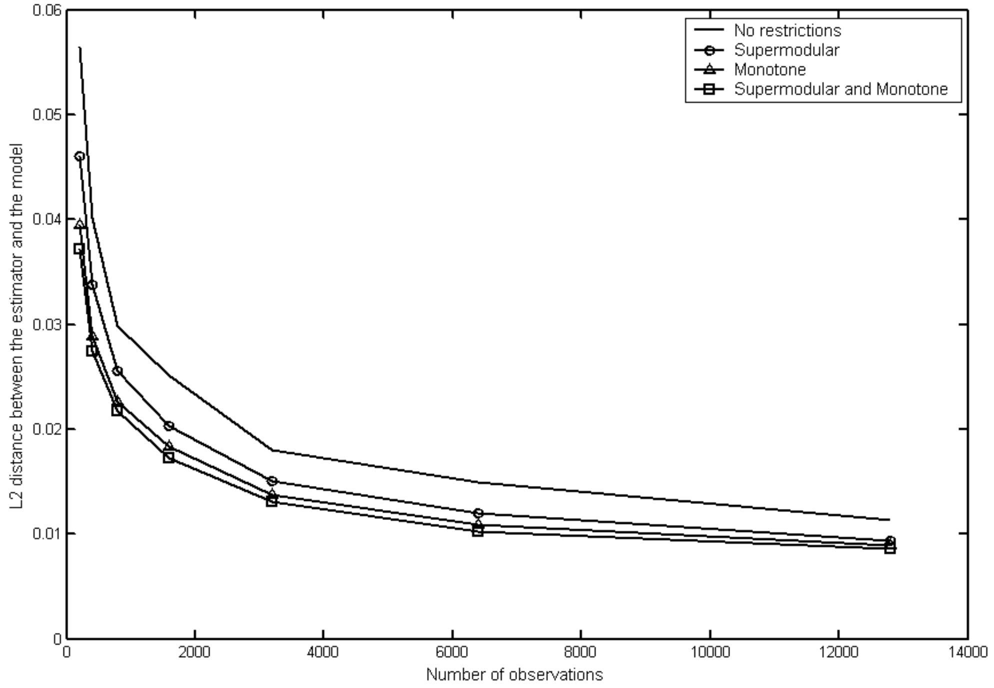

Figure 4: L_2 distance of the estimator from the true model

change the number of observations and measure the L_2 distance between the estimators and the true models. For each number of observation, 250 Monte Carlo experiments were preformed. Figure 4 describes one of the Monte Carlo experiments for $y = \min(x_1, x_2) + \varepsilon$ where $\varepsilon \sim N(0,1)$.[22] The L_2 distance in Figure 4 is the mean of the L_2 distance achieved in the 250 experiments performed. The rate of convergence computed from this Monte Carlo experiment is approximately $\frac{1}{3}$. This rate of convergence corresponds to setting $p = 2$ in Section 3.2. Figure 5 reports the mean of the ratio between the L_2 distance of the restricted versus the unrestricted estimators. It is clear that the gains increase the more relevant restrictions are used. The magnitude of these gains is similar to the gains suggested in Table 1.

[22] Experiments performed on the other models showed the same results.

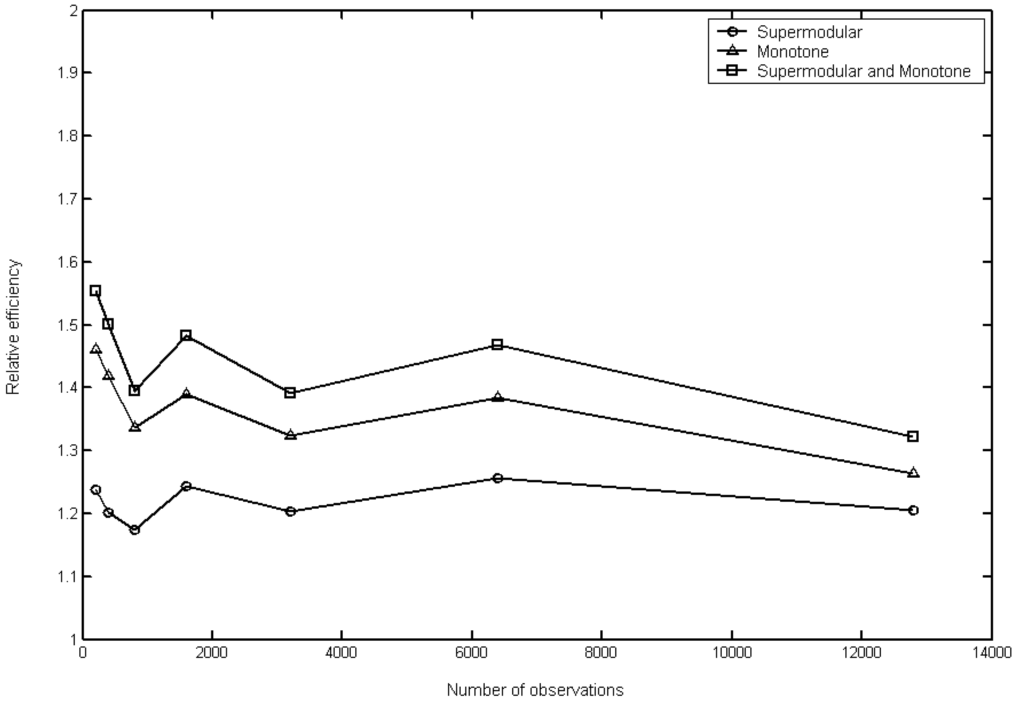

Figure 5: Efficiency gains from shape restrictions

6 Empirical Illustration

To be added

7 Concluding Remarks

A nonparametric shape restricted estimator based on a sieve method is described in this paper. The benefits of the sieve framework are as follows. First, various assumptions like monotonicity and supermodularity are easily incorporated into the estimator. Variety of shape restrictions can be treated using the same framework. Second, the estimator is computed using quadratic programming with linear inequalities constraints. Therefore, this estimator is easy to implement and each additional shape restriction results in additional linear inequality constraints added to the problem. A shape-restricted semi-linear version of the model is discussed and adds to the applicability of the shape-restricted estimator in empirical work. Finally, the geometric interpretation of the estimator provides an interesting insight about the small sample efficiency gains resulting from using the information on the shape of the regression function.

Appendix

A Normalized B-splines

The B-spline basis defined here is based on Schumaker (1981, Chapter 4). Normalized B-splines of degree l are piecewise polynomials of degree $l-1$. It is common to use even integers for l. It is convenient to define the normalized B-splines using a kernel function centered around zero. Let $K^l(x)$ be the following kernel function

$$K^l(x) = \begin{cases} \sum_{j=0}^{l} \frac{(-1)^j}{(l-1)!} \binom{l}{j} (x + \frac{l}{2} - j)_+^{l-1} & x \in [-\frac{l}{2}, \frac{l}{2}] \\ 0 & otherwise \end{cases}$$

where

$$(x - x_0)_+^a = \begin{cases} (x - x_0)^a & x \geq x_0 \\ 0 & otherwise \end{cases} .$$

To normalize $K^l(x)$ to the interval $\left[\frac{i-\frac{l}{2}}{m}, \frac{i+\frac{l}{2}}{m}\right]$ we define

$$\psi_{m,i}^l(x) = K^l(mx - i)$$

for any integer i. The basis functions in Ψ_m^l are centered around the extended equidistant grid vector $\bar{\Gamma}_m^l$ and consist of the following $m+l-1$ function:

$$\Psi_m^l = \{\psi_{m,i}^l(x)\}_{i=-l/2+1}^{m+l/2-1}.$$

The following are useful properties of the normalized B-spline basis:

1. **Unit division**: $\sum_{i=-\infty}^{\infty} \psi_{m,i}^l(x) \equiv 1$ for any positive integers $m \geq 1$ and $l \geq 2$.

2. **Differentiability**: $\psi_{m,i}^l$ is $l-1$ times differentiable for each i and the $l-1^{th}$ derivative is undefined everywhere except at a subset of the grid points $\bar{\Gamma}_m^l$ on which $\psi_{m,i}^l$ takes a value different than zero.

3. **Symmetry**: $\frac{d^k \psi_{m,i}^l}{dx^k}(\frac{i}{m} - c) = (-1)^k \frac{d^k \psi_{m,i}^l}{dx^k}(\frac{i}{m} + c)$ for all integer i, real number $c > 0$ and integer k such that $0 \leq k \leq l-1$.

B Proofs and Technical Notes

B.1 Proof of Theorem 1

Proof. The proof is given for the one dimensional case. The two dimensional case is analogues to the one dimensional case but requires additional notation and thus is omitted. We prove the claim by induction on p.

Let $m > l \geq p$ be three integers such that l is an even number. In what follows $\bar{\Gamma}^l_m$ is the extended equidistant grid on $[0, 1]$ as in definition (9) and Ψ^l_m is the base of normalized B-spline wavelets of order l centered around the equidistant grid vector points $\bar{\Gamma}^l_m$ as defined in Appendix A. Let $f \in \Im^p_m \left(\Psi^l_m\right)$. We proceed using induction.

Assume $p = 1$ and $l \geq 2$, and let $\Gamma_2 = (\gamma_0, \gamma_1)$ be a grid vector on $[0, 1]$. First consider the case where $l = 2$ and $\hat{f} = \sum_{i=0}^{m} \theta_i \psi_{m,i}$. Assume that $\gamma_0 \in \left[\frac{j_0}{m}, \frac{j_0+1}{m}\right)$ and $\gamma_1 \in \left[\frac{j_1}{m}, \frac{j_1+1}{m}\right)$ where $j_0 \leq j_1$. Since $l = 2$, $\hat{f}$ is a piecewise linear function and $\hat{f}\left(\gamma_0\right) = \theta_{j_0} w_0 + \theta_{j_0+1}(1 - w_0)$ and $\hat{f}\left(\gamma_1\right) = \theta_{j_1} w_1 + \theta_{j_1+1}(1 - w_1)$ where $w_0 = m\gamma_0 - j_0$ and $w_1 = m\gamma_1 - j_1$. If $j_0 = j_1$ then it has to be that $w_0 < w_1$ and thus $\hat{f}\left(\gamma_0\right) < \hat{f}\left(\gamma_1\right)$. If $j_0 < j_1$ then $\theta_{j_0} \leq \theta_{j_1}$ and $\theta_{j_0+1} \leq \theta_{j_1+1}$ and therefore $\hat{f}\left(\gamma_0\right)$ is a weighted average of two numbers which are smaller than the two numbers that $\hat{f}\left(\gamma_1\right)$ is and average of. Now consider the case where $l > 2$. $\hat{f}$ is differentiable everywhere and therefore by the mean value theorem there is $\tilde{\gamma} \in [\gamma_0, \gamma_1]$ such that $\hat{f}\left(\gamma_1\right) - \hat{f}\left(\gamma_0\right) = \hat{f}'\left(\tilde{\gamma}\right)\left(\gamma_1 - \gamma_0\right)$. since $\gamma_1 \geq \gamma_0$ the sign of $\hat{f}\left(\gamma_1\right) - \hat{f}\left(\gamma_0\right)$ is equal to the sign of $\hat{f}'\left(\tilde{\gamma}\right)$. There exist an integer $\tilde{j}$ such that $\frac{\tilde{j}}{m} \leq \tilde{\gamma} < \frac{\tilde{j}+1}{m}$. From properties 1 and 2 of the normalized B-spline in Appendix A it is clear that $\sum_{i=-\infty}^{\tilde{j}} \frac{d\psi^l_{m,i}}{dx}(\tilde{\gamma}) = -\sum_{i=\tilde{j}+1}^{\infty} \frac{d\psi^l_{m,i}}{dx}(\tilde{\gamma})$ by taking the derivative of property 1 with respect to x and evaluate it at $x = \gamma$. From property 3 and the construction of the normalized B-splines $\frac{d\psi^l_{m,i}}{dx}(\tilde{\gamma}) \leq 0$ for $i \leq \tilde{j}$ and $\frac{d\psi^l_{m,i}}{dx}(\tilde{\gamma}) \geq 0$ for $i > \tilde{j}$. When $p = 1$ $A^p_m \theta \geq 0$ implies that $\theta_i \leq \theta_{i+1}$. Therefore, $\hat{f}'(\tilde{\gamma}) = \sum_{i=-\infty}^{\tilde{j}} \theta_i \frac{d\psi^l_{m,i}}{dx}(\tilde{\gamma}) + \sum_{i=\tilde{j}+1}^{\infty} \theta_i \frac{d\psi^l_{m,i}}{dx}(\tilde{\gamma}) \leq 0$. This concludes the proof for $p = 1$.

We now assume that for $p - 1$ the claim in the theorem holds. Let $\Gamma_p = \left(\gamma_0, ..., \gamma_p\right)$ be an arbitrary grid vector on $[0, 1]$ and $f \in \Im^p_m \left(\Psi^l_m\right)$. Using the mean value theorem $A^p_p f(\Gamma_p) = A^{p-1}_{p-1} D_p f(\Gamma_p) \geq A^{p-1}_{p-1} f'(\tilde{\Gamma}_{p-1})\delta$ where $\delta = \min_{i=1..p} \gamma_i - \gamma_{i-1}$ and $\tilde{\Gamma}_{p-1} = (\tilde{\gamma}_0, ..., \tilde{\gamma}_{p-1})$ such that $\tilde{\gamma}_j \in [\gamma_j, \gamma_{j+1}]$. Since $\delta > 0$, $A^p_p f(\Gamma_p) \geq 0$ if $A^{p-1}_{p-1} f'(\tilde{\Gamma}_{p-1}) \geq 0$. Now $f' =$

$\sum_{i=-l/2+1}^{m+l/2-1}\theta_i\frac{d\psi^l_{m,i}}{dx}$ and $\frac{d\psi^l_{m,i}}{dx}$ satisfies conditions 1-3 above (when condition 1 is satisfied with equality to zero instead of one). Therefore, using the induction step we conclude that the claim in the theorem is satisfied for p. ■

B.2 Proof of Theorem 2

Definition 11 *Let $K \subset \Re^d$ be compact and $\Gamma = \{\gamma_1, ..., \gamma_D\} \subset K$ finite and let $\Im \subset C(K)$ be a sub class of continuous functions from K to $\Re$. Then $G(\Gamma, \Im) = \{u \in \Re^D | \exists g \in \Im$ and $\gamma \in \Gamma$ such that $u = g(\gamma)\}$ is called **the dual space** of $(\Gamma, \Im)$.*

Proof: Fix m and l and for notational convenience denote $G = G(\bar{\Gamma}_m, \Im^l_m)$ and $\Psi = \Psi^l_m(x)$. From the properties of $\Im^l_m$ it is clear that G is a non-empty, close in the topology induced by the distance function $d(u, v) = \frac{1}{D}\sum_{i=1}^{D}(u_i - v_i)^2$ and a convex cone. Define $\overline{y} = \max_{i=1...D} y_i$ and $\underline{y} = \min_{i=1...D} y_i$. Since $(1, 1, ..., 1) \in G$ also $(\overline{y}, \overline{y}, ..., \overline{y})$ and $\left(\underline{y}, \underline{y}, ..., \underline{y}\right)$ are in G. Assume that $\theta^* \in G$ solves (15) and define $\theta^{**} = \theta^* \vee \left(\underline{y}, \underline{y}, ..., \underline{y}\right)$. From the fact that G is a sub-lattice we know that $\theta^{**} \in G$. Assume that there exist a coordinate $\tilde{\imath}$ such that $\theta^*_{\tilde{\imath}} < \theta^{**}_{\tilde{\imath}}$. The matrix Ψ has only non-negative entries and thus

$$||\Psi\theta^{**} - y|| < ||\Psi\theta^* - y||$$

a contradiction to the fact that θ^*solves (15). The argument that $\theta^*_i \leq \overline{y}$ is proved in a similar way using $\theta^{***} = \theta^* \wedge (\overline{y}, \overline{y}, ..., \overline{y})$.

B.3 Proof of Theorem 3

Proof. Let $\bar{\Gamma}_m$ be the equidistant grid $(0, \frac{1}{m}, ..., 1)$. For any function $f \in \Im^p$, the values it takes on $\bar{\Gamma}_m$ satisfy $A^m_p f(\bar{\Gamma}_m)' \leq 0$. The points $f(\bar{\Gamma}_m)$ on the grid $\bar{\Gamma}_m$ can be interpolated using the base Ψ^l_m as explained in Section 2.3. We denote this function by f_m. Theorem 1 assures that $f_m \in \Im^p_m(\Psi^l_m) \subset \Im^p$. f and f_m are $k = \min(l-2, p)$ times continuously differentiable. Since the k-th derivative of f_m is continuous, then for each $x \in [\frac{j}{m}, \frac{j+1}{m}]$ there is ξ_x in this interval such that $\partial^k f_m(x) = \partial^k f(\xi_x)$. Thus, $\left|\partial^k f(x) - \partial^k f_m(x)\right| = \left|\partial^k f(x) - \partial^k f(\xi_x)\right| \leq \omega(\partial^k f, \frac{1}{m})_{L_\infty[\frac{j}{m}, \frac{j+1}{m}]}$ where ∂^k represents the k-th derivative operator and $\omega(\varphi,\delta)_{L_\infty[I]} = \sup_{x_1,x_2\in I,|x_1-x_2|\le h} |\varphi(x_1)-\varphi(x_2)|$ is the modulus of continuity. For $f \in \Im$, $\omega(\partial^k f, \frac{1}{m})_{L_\infty[\frac{j}{m},\frac{j+1}{m}]} \le m^{-k}$. To show the lower bound take the following $\tilde{f} \in \Im^p_{2m}\left(\Psi^l_{2m}\right)$, $\tilde{f} = \sum_{i=-l/2}^{2m+l/2} a_i \psi^l_{2m,i}$ where $a_{-l/2} = 0$ and $a_i = 1$ for $i = \frac{-l}{2}+1,...,m+\frac{l}{2}$. This function has positive differences up to order l. $d(\tilde{f}, \Im^p_m) \ge \tilde{C}\omega(\partial^k f, \frac{1}{m})_{L_\infty[\frac{j}{m},\frac{j+1}{m}]}$ following the same arguments in Schumaker (1981, Theorem 6.16). The constant $\tilde{C}$ depends on the order of splines used. ■

B.4 Proof of Theorem 4

Proof. Let $\theta \in \Im^p$ be the true regression function $\theta(x) = E(Y|X=x)$. Let $\gamma = 1/(2p+1)$ and let ψ be an infinitely differentiable function with compact support such that $\psi(0) > 0$ and such that Lipschitz condition holds for any $|\alpha| \le p-1$. Define the perturbation sequence as follows. For $\delta \in (0,1]$ and a constant $M > 0$ define $g_n(x) = \delta M^p n^{-\gamma p}\psi(M^{-1}n^\gamma x)$ and $\theta_n = \theta + \varepsilon g_n$. By choosing ε small enough we make sure that δ and M are chosen such that we have also $\theta_n \in \Theta_i$. From this point the proof follows exactly the steps in Stone (1980, 1982). ■

B.5 Proof of Lemma 1

This proof builds on the proof for Lemma 9.3 in Gyorfi et al. (2002). Let $\{f_1,...,f_n\}$ be a $\frac{\varepsilon}{2}$-net in $\mathcal{F}_K$ under $||\cdot||$ and let $\psi_1,...,\psi_D$ be a basis for $\mathcal{F}_K$. For any $(a_1,...,a_D)$ and $(b_1,...,b_D)$, vectors of real numbers, $||\sum_i a_i\psi_i - \sum_i b_i\psi_i|| = (a-b)'\Psi(a-b)$, where $\Psi = \left(\langle\psi_i,\psi_j\rangle\right)_{i,j=1..D}$. Ψ is positive semidefinite such that $\Psi = \Psi^{\frac{1}{2}}\Psi^{\frac{1}{2}}$. Therefore, $||\sum_i a_i\psi_i - \sum_i b_i\psi_i|| = \left|\left|(a-b)'\Psi^{\frac{1}{2}}\right|\right|$. Since $f_i \in \mathcal{F}_K$, $f_i = \sum_j a^i_j\psi_j$ for some vector $a^i = \left(a^i_1,...,a^i_D\right)$. Thus, $\left|\left|a^i\Psi^{\frac{1}{2}} - a^j\Psi^{\frac{1}{2}}\right|\right| = ||f_i - f_j|| \le \varepsilon$ and $a^1\Psi^{\frac{1}{2}},...,a^D\Psi^{\frac{1}{2}}$ is an ε-net in F_K of size $n = N\left(\frac{\varepsilon}{2},\mathcal{F}_K,||\cdot||\right)$. Let c_D be the volume of a ball of radius 1 in $\Re^D$, then $n \cdot c_D \cdot \varepsilon^D = Vol(F_K)$ and the conclusion follows.

C Monte Carlo Study - Results

Each model was estimated 1000 times for each design. 95% confidence intervals are calculated and reported in Tables 3, 4 and 5. The regression models that were estimated are described bellow. The mesh size used for the B-spline estimation is $m_1, m_2 = 6$.

$$\begin{aligned}
Model\ 1 &: \quad Y = \min(X_1, X_2) + \varepsilon \\
Model\ 2 &: \quad Y = X_1 X_2 + I_{\{X_1 \geq \frac{1}{2}, X_2 \geq \frac{1}{2}\}} \cdot (X_1 - \frac{1}{2})(X_2 - \frac{1}{2}) + \varepsilon \\
Model\ 3 &: \quad Y = X_1^{\frac{1}{3}} X_2^{\frac{2}{3}} + \varepsilon\ .
\end{aligned}$$

The fit of each estimation method is checked according to four criteria:

1. The fit at the corner of the domain $(x_1, x_2) = (0, 0)$.
2. The fit on a boundary point of the domain $(x_1, x_2) = (\frac{1}{3}, 0)$.
3. The fit in an interior point of the domain $(x_1, x_2) = (\frac{1}{2}, \frac{1}{2})$.
4. The l_2-distance between $\hat{f}$ and f over the whole domain.

The distributions of the design points used in this study are either:

1. X_1 and X_2 are uniformly and independently distributed on $[0, 1]^2$.
2. $X_1 \sim U[0, 1]$ and $X_2 | X_1 \sim U[\frac{X_1}{2}, \frac{X_1+1}{2}]$.

The two distributions of errors used are:

1. $\varepsilon \sim N(0, 1)$.
2. $\varepsilon \sim U[-1, 1]$.

Table 3: Monte Carlo results - Model 1

$Y = \min(X_1, X_2) + \varepsilon,\ N = 400$				
		$X_1, X_2 \sim U[0,1],\ \varepsilon \sim N(0,1)$		
	Unrestricted estimator	Supermodularity	Monotonicity	Supermodularity and monotonicity
$f(0,0) = 0$	$[-3.12, 2.83]$	$[-0.58, 2.09]$	$[-1.85, 0.10]$	$[-0.67, 0.14]$
$f(\frac{1}{3}, 0) = 0$	$[-1.55, 1.54]$	$[-0.54, 0.74]$	$[-0.55, 0.23]$	$[-0.41, 0.25]$
$f(\frac{1}{2}, \frac{1}{2}) = \frac{1}{2}$	$[-0.56, 1.64]$	$[-0.00, 0.89]$	$[0.16, 0.64]$	$[0.20, 0.65]$
L_2-norm	$[0.278, 0.416]$	$[0.145, 0.282]$	$[0.097, 0.225]$	$[0.084, 0.207]$
		$X_1 \sim U[0,1],\ X_2 \vert X_1 \sim U[\frac{X_1}{2}, \frac{X_1+1}{2}],\ \varepsilon \sim N(0,1)$		
	Unrestricted estimator	Supermodularity	Monotonicity	Supermodularity and monotonicity
$f(0,0) = 0$	$[-2.31, 2.31]$	$[-0.57, 2.00]$	$[-1.83, 0.10]$	$[-0.67, 0.14]$
$f(\frac{1}{3}, 0) = 0$	$[-1.17, 1.24]$	$[-0.55, 0.64]$	$[-0.46, 0.25]$	$[-0.41, 0.26]$
$f(\frac{1}{2}, \frac{1}{2}) = \frac{1}{2}$	$[-0.52, 1.49]$	$[0.01, 0.88]$	$[0.15, 0.65]$	$[0.18, 0.64]$
L_2-norm	$[0.378, 0.619]$	$[0.310, 0.509]$	$[0.306, 0.503]$	$[0.293, 0.489]$
		$X_1, X_2 \sim U[0,1],\ \varepsilon \sim U[0,1]$		
	Unrestricted estimator	Supermodularity	Monotonicity	Supermodularity and monotonicity
$f(0,0) = 0$	$[-1.64, 1.63]$	$[-0.39, 1.29]$	$[-1.06, 0.07]$	$[-0.44, 0.09]$
$f(\frac{1}{3}, 0) = 0$	$[-0.93, 0.96]$	$[-0.34, 0.47]$	$[-0.32, 0.16]$	$[-0.26, 0.17]$
$f(\frac{1}{2}, \frac{1}{2}) = \frac{1}{2}$	$[-0.08, 1.12]$	$[0.16, 0.71]$	$[0.25, 0.61]$	$[0.25, 0.61]$
L_2-norm	$[0.163, 0.241]$	$[0.091, 0.167]$	$[0.067, 0.136]$	$[0.060, 0.127]$
		$X_1 \sim U[0,1],\ X_2 \vert X_1 \sim U[\frac{X_1}{2}, \frac{X_1+1}{2}],\ \varepsilon \sim U[0,1]$		
	Unrestricted estimator	Supermodularity	Monotonicity	Supermodularity and monotonicity
$f(0,0) = 0$	$[-1.19, 1.24]$	$[-0.38, 1.17]$	$[-1.05, 0.06]$	$[-0.46, 0.10]$
$f(\frac{1}{3}, 0) = 0$	$[-1.06, 1.00]$	$[-0.33, 0.49]$	$[-0.44, 0.17]$	$[-0.29, 0.18]$
$f(\frac{1}{2}, \frac{1}{2}) = \frac{1}{2}$	$[0.06, 1.04]$	$[0.19, 0.68]$	$[0.25, 0.59]$	$[0.27, 0.60]$
L_2-norm	$[0.336, 0.466]$	$[0.304, 0.414]$	$[0.314, 0.422]$	$[0.310, 0.412]$

Table 4: Monte Carlo results - Model 2

$Y = X_1X_2 + I_{\{X_1\geq\frac{1}{2},X_2\geq\frac{1}{2}\}}\cdot(X_1-\frac{1}{2})(X_2-\frac{1}{2})+\varepsilon$				
$X_1, X_2 \sim U[0,1]$, $\varepsilon \sim N(0,1)$, $N = 400$				
	Unrestricted estimator	Supermodularity	Monotonicity	Supermodularity and monotonicity
$f(0,0) = 0$	$[-2.63, 3.10]$	$[-0.55, 2.11]$	$[-1.69, 0.05]$	$[-0.61, 0.08]$
$f(\frac{1}{3}, 0) = 0$	$[-1.72, 1.72]$	$[-0.58, 0.70]$	$[-0.64, 0.15]$	$[-0.47, 0.18]$
$f(\frac{1}{2}, \frac{1}{2}) = \frac{1}{4}$	$[-0.83, 1.29]$	$[-0.22, 0.70]$	$[0.00, 0.50]$	$[0.02, 0.49]$
L_2-norm	$[0.280, 0.42]$	$[0.144, 0.287]$	$[0.09, 0.221]$	$[0.082, 0.205]$
$X_1 \sim U[0,1]$, $X_2\|X_1 \sim U[\frac{X_1}{2}, \frac{X_1+1}{2}]$, $\varepsilon \sim N(0,1)$, $N = 400$				
	Unrestricted estimator	Supermodularity	Monotonicity	Supermodularity and monotonicity
$f(0,0) = 0$	$[-2.93, 3.42]$	$[-0.10, 2.35]$	$[-1.60, 0.31]$	$[-0.20, 0.36]$
$f(\frac{1}{3}, 0) = 0$	$[-1.62, 2.09]$	$[-0.24, 1.10]$	$[-0.31, 0.38]$	$[-0.13, 0.38]$
$f(\frac{1}{2}, \frac{1}{2}) = \frac{1}{4}$	$[-0.73, 1.36]$	$[-0.12, 0.75]$	$[0.17, 0.47]$	$[0.17, 0.46]$
L_2-norm	$[0.351, 0.616]$	$[0.225, 0.461]$	$[0.149, 0.275]$	$[0.152, 0.279]$
$X_1, X_2 \sim U[0,1]$, $\varepsilon \sim U[0,1]$, $N = 400$				
	Unrestricted estimator	Supermodularity	Monotonicity	Supermodularity and monotonicity
$f(0,0) = 0$	$[-1.60, 1.542]$	$[-0.32, 1.28]$	$[-1.10, 0.05]$	$[-0.38, 0.07]$
$f(\frac{1}{3}, 0) = 0$	$[-0.93, 0.88]$	$[-0.37, 0.42]$	$[-0.34, 0.11]$	$[-0.28, 0.12]$
$f(\frac{1}{2}, \frac{1}{2}) = \frac{1}{4}$	$[-0.32, 0.90]$	$[-0.04, 0.50]$	$[0.06, 0.43]$	$[0.07, 0.42]$
L_2-norm	$[0.163, 0.241]$	$[0.088, 0.164]$	$[0.067, 0.138]$	$[0.060, 0.125]$
$X_1 \sim U[0,1]$, $X_2\|X_1 \sim U[\frac{X_1}{2}, \frac{X_1+1}{2}]$, $\varepsilon \sim U[0,1]$, $N = 400$				
	Unrestricted estimator	Supermodularity	Monotonicity	Supermodularity and monotonicity
$f(0,0) = 0$	$[-1.56, 2.10]$	$[0.04, 1.33]$	$[-0.65, 0.32]$	$[0.02, 0.34]$
$f(\frac{1}{3}, 0) = 0$	$[-0.60, 1.30]$	$[-0.01, 0.77]$	$[-0.04, 0.36]$	$[0.07, 0.36]$
$f(\frac{1}{2}, \frac{1}{2}) = \frac{1}{4}$	$[-0.32, 0.98]$	$[0.05, 0.57]$	$[0.22, 0.40]$	$[0.23, 0.40]$
L_2-norm	$[0.263, 0.439]$	$[0.209, 0.363]$	$[0.171, 0.273]$	$[0.177, 0.277]$

Table 5: Monte Carlo results - Model 3

$Y = X_1^{\frac{1}{3}} X_2^{\frac{2}{3}} + \varepsilon$				
$X_1, X_2 \sim U[0,1]$, $\varepsilon \sim N(0,1)$, $N = 400$				
	Unrestricted estimator	Supermodularity	Monotonicity	Supermodularity and monotonicity
$f(0,0) = 0$	$[-2.50, 2.82]$	$[-0.56, 1.86]$	$[-1.73, 0.16]$	$[-0.59, 0.21]$
$f(\frac{1}{3}, 0) = 0$	$[-1.61, 1.56]$	$[-0.54, 0.75]$	$[-0.54, 0.30]$	$[-0.41, 0.30]$
$f(\frac{1}{2}, \frac{1}{2}) = \frac{1}{2}$	$[-0.59, 1.59]$	$[-0.06, 0.95]$	$[0.24, 0.72]$	$[0.25, 0.71]$
L_2-norm	$[0.282, 0.416]$	$[0.139, 0.282]$	$[0.091, 0.220]$	$[0.079, 0.201]$
$X_1 \sim U[0,1]$, $X_2 \vert X_1 \sim U[\frac{X_1}{2}, \frac{X_1+1}{2}]$, $\varepsilon \sim N(0,1)$, $N = 400$				
	Unrestricted estimator	Supermodularity	Monotonicity	Supermodularity and monotonicity
$f(0,0) = 0$	$[-1.84, 2.67]$	$[-0.02, 1.81]$	$[-0.84, 0.50]$	$[-0.06, 0.53]$
$f(\frac{1}{3}, 0) = 0$	$[-0.64, 1.67]$	$[-0.05, 1.13]$	$[-0.04, 0.55]$	$[0.10, 0.55]$
$f(\frac{1}{2}, \frac{1}{2}) = \frac{1}{2}$	$[-0.51, 1.53]$	$[0.06, 0.93]$	$[0.35, 0.63]$	$[0.36, 0.62]$
L_2-norm	$[0.330, 0.599]$	$[0.188, 0.411]$	$[0.103, 0.223]$	$[0.114, 0.231]$
$X_1, X_2 \sim U[0,1]$, $\varepsilon \sim U[0,1]$, $N = 400$				
	Unrestricted estimator	Supermodularity	Monotonicity	Supermodularity and monotonicity
$f(0,0) = 0$	$[-1.72, 1.50]$	$[-0.37, 1.09]$	$[-1.07, 0.12]$	$[-0.41, 0.16]$
$f(\frac{1}{3}, 0) = 0$	$[-1.02, 1.02]$	$[-0.32, 0.43]$	$[-0.32, 0.21]$	$[-0.23, 0.22]$
$f(\frac{1}{2}, \frac{1}{2}) = \frac{1}{2}$	$[-0.10, 1.15]$	$[-0.23, 0.76]$	$[-0.31, 0.67]$	$[-0.34, 0.67]$
L_2-norm	$[0.163, 0.241]$	$[0.081, 0.166]$	$[0.066, 0.136]$	$[0.056, 0.127]$
$X_1 \sim U[0,1]$, $X_2 \vert X_1 \sim U[\frac{X_1}{2}, \frac{X_1+1}{2}]$, $\varepsilon \sim U[0,1]$, $N = 400$				
	Unrestricted estimator	Supermodularity	Monotonicity	Supermodularity and monotonicity
$f(0,0) = 0$	$[-2.16, 3.18]$	$[0.22, 1.72]$	$[-0.75, 0.49]$	$[0.15, 0.51]$
$f(\frac{1}{3}, 0) = 0$	$[-0.39, 1.35]$	$[0.16, 0.95]$	$[0.05, 0.52]$	$[0.21, 0.53]$
$f(\frac{1}{2}, \frac{1}{2}) = \frac{1}{2}$	$[-0.26, 1.19]$	$[0.21, 0.77]$	$[0.41, 0.57]$	$[0.41, 0.57]$
L_2-norm	$[0.241, 0.408]$	$[0.183, 0.334]$	$[0.129, 0.221]$	$[0.139, 0.225]$

References

Barlow, R., Bartholomew, D., Bremner, J. & Brunk, H. (1972). *Statistical Inference under Order Restrictions*, John Wiley and Sons, London.

Beresteanu, A. (2001). *Nonparametric Estimation of Supermodular Functions with Application to the Telecommunication Industry*, PhD thesis, Northwestern University.

Beresteanu, A. (2005). Nonparametric analysis of cost complementarities in the telecommunications industry, *RAND journal of Economics* **Winter**: 870–889.

Birge, L. (1987). Estimating a density under order restrictions: Nonasymptotic minimax risk, *Annals of Statistics* **15**(3): 995–1012.

Birge, L. & Massart, P. (1993). Rates of convergence for minimum contrast estimators, *Probability Theory and Related Fields* **97**: 113–150.

Chen, X. (2007). Large sample sieve estimation of semi-nonparametric models, *in* J. J. Heckman & E. E. Leamer (eds), *Handbook of Econometrics Volume 6B*, Elsevier North Holland.

Chen, X. & Shen, X. (1998). Sieve extermum estimates for weakly dependent data, *Econometrica* **66**(2): 289–314.

Chui, C. (1992). *An Introduction to Wavelets*, Academic Press.

Davidson, R. & MacKinnon, J. G. (1981). Several tests for model specification in the presence of alternative hypotheses, *Econometrica* **49**(3): 781–793.

Dole, D. (1999). CoSmo: A constrained scatterplot smoother for estimating convex, monotonic transformations, *Journal of Business and Economic Statistics* **17**(4): 444–455.

Donald, S. G. & Newey, W. K. (1994). Series estimation of semilinear models, *Journal of Multivariate Analysis* **50**: 30–40.

Dykstra, R. (1983). An algorithm for restricted least squares regression, *Journal of the American Statistical Association* **78**: 837–842.

Gallant, R. (1981). On the bias in flexible functional forms and an essential unbiased form : The fourier flexible form, *Journal of Econometrics* **15**: 211–245.

Gallant, R. (1982). Unbiased determination of production technologies, *Journal of Econometrics* **20**: 285–323.

Gallant, R. A. & Golub, G. H. (1984). Imposing curvature restrictions on flexible functional forms, *Journal of Econometrics* **26**: 295–321.

Geman, S. & Hwang, C.-R. (1982). Nonparametric maximum likelihood estimation by the method of sieves, *Annals of Statistics* **10**(2): 401–414.

Genander, U. (1981). *Abstract Inference*, Wiley Series in Probability, John Wiley and Sons.

Goldman, S. & Ruud, P. (1993). Nonparametric multivariate regression subject to constraint, *mimeo University of California Berkley* .

Gyorfi, L., Kohler, M., Krzyzak, A. & Walk, H. (2002). *Adistribution Free Theory of Nonparametric Regression*, Springer Series in Statistics, Springer-Verlag New York, Inc.

Hanson, D. L., Pledger, G. & Wright, F. T. (1973). On consistency in monotonic regression, *Annals of Statistics* **1**(3): 401–421.

Hanson, D. & Pledger, G. (1976). Consistency in concave regression, *Annals of Statistics* **4**(6): 1038–1050.

He, X. & Shi, P. (1998). Monotone b-spline smoothing, *Journal of the American statistical Association* **93**(442): 643–650.

Hong, Y. & White, H. (1995). Consistent specification testing via nonparametric series regression, *Econometrica* **63**(5): 1133–1159.

Kiefer, J. (1982). Optimum rates for non-parametric density and regression estimates under order restrictions, *in* K. G., P. R. Krishnaiah & J. K. Ghosh (eds), *Statistics and Probability: Essays in Honor of C. R. Rao*, North-holland, Amsterdam, pp. 419–428.

Kolmogorov, A. & Tihomirov, V. (1961). Epsilon - entropy and epsilon-capacity of sets in functional spaces, *Americam Mathematical Society Translations, Series 2* **17**: 277–364.

Luenberger, D. G. (1984). *Linear and Nonlinear Programing*, second edn, Addison-Wsley Publishing Company.

Mammen, E. (1991). Estimating a smooth monotone regression function, *Annals of statistics* **19**(2): 724–740.

Matzkin, R. (1994). Restrictions of economic theory in nonparametric methods, *Handbook of Econometrics IV*, Vol. 4, North-Holland, chapter 42, pp. 2524–2558.

Mukerjee, H. (1988). Monotone nonparametric regression, *Annals of Statistics* **16**(2): 741–750.

Robertson, T., Wright, F. & Dykstra, R. (1988). *Order Restricted Statistical Inference*, Wiley Series in Probability and Mathematical Statistics, John Wiley and Sons.

Rockafellar, R. (1970). *Convex Analysis*, Princeton University Press.

Schumaker, L. L. (1981). *Spline Functions: Basic Theory*, John Wiley and Sons.

Shen, X. & Wong, W. H. (1994). Convergence rate of sieve estimates, *The Annals of Statistics* **22**(2): 580–615.

Stone, C. (1980). Optimal rate of convergence for nonparametric estimators, *The Annals of Statistics* **8**(6): 1348–1360.

Stone, C. (1982). Optimal global rates of convergence for nonparametric regression, *Annals of Statistics* **10**(4): 1040–1053.

Tripathi, G. (2000). Local semiparametric efficiency bounds under shape restrictions, *Econometric Theory* **16**: 729–739.

Van de Geer, S. (2000). *Applications of Empirical Process Theory*, Cambridge University Press.

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

Wooldridge, J. M. (1992). A test for functional form against nonparametric alternatives, *Econometric Theory* **8**: 452–475.