ELSEVIER

Journal of Statistical Planning and Inference 107 (2002) 289–306

journal of statistical planning and inference

www.elsevier.com/locate/jspi

Testing of monotonicity in parametric regression models

E. Doveh[a], A. Shapiro[b,*], P.D. Feigin[a]

[a] *Faculty of Industrial Engineering and Management, Technion-Israel Institute of Technology, Haifa 32000, Israel*

[b] *School of Industrial and Systems Engineering, Georgia Institute of Technology, Atlanta, GA 30332-0205, USA*

Abstract

In data analysis concerning the investigation of the relationship between a dependent variable Y and an independent variable X, one may wish to determine whether this relationship is monotone or not. This determination may be of interest in itself, or it may form part of a (nonparametric) regression analysis which relies on monotonicity of the true regression function. In this paper we generalize the test of positive correlation by proposing a test statistic for monotonicity based on fitting a parametric model, say a higher-order polynomial, to the data with and without the monotonicity constraint. The test statistic has an asymptotic chi-bar-squared distribution under the null hypothesis that the true regression function is on the boundary of the space of monotone functions. Based on the theoretical results, an algorithm is developed for evaluating significance of the test statistic, and it is shown to perform well in several null and nonnull settings. Extensions to fitting regression splines and to misspecified models are also briefly discussed.

Keywords: Monotone regression; Large samples theory; Testing monotonicity; Chi-bar-squared distributions; Misspecified models

1. Introduction

The general setting of this paper is that of the regression of scalar Y on scalar X over the interval $[a, b]$. In many practical situations, the researcher may be tempted to assume monotonicity of the regression function and subsequently to design an estimation procedure based on this assumption. The natural question is whether the available

* Corresponding author. Tel.: +1-404-894-6544; fax: +1-404-894-2301.

E-mail addresses: ierde01@ie.technion.ac.il (E. Doveh), ashapiro@isye.gatech.edu (A. Shapiro), paulf@ie.technion.ac.il (P.D. Feigin).

0378-3758/02/$ - see front matter
PII: S0378-3758(02)00259-8

data support this assumption. In other words, is it possible to test for monotonicity, statistically?

In this paper we approach that problem in a framework of a parametric setting, namely that of a multiple linear regression with given regressor functions; and with the true regression function being a member of the class of candidate fitting functions (i.e., with the model being correctly specified). A more general theory would cover the case of testing for monotonicity when the model is misspecified. This case will be referred to briefly in the sequel and would be a subject of further investigation.

There has been a fair amount of interest in the numerical and statistical aspects of monotonic regression in recent years. We mention some examples from the literature. In isotonic regression, the data are fitted by a function from an appropriate class of isotonic (or monotonic) functions. A thorough discussion of this method is presented in Barlow et al. (1972) and Robertson et al. (1988). Since no structural assumptions apart from monotonicity are made, the derived estimator is then a piecewise linear function which, of course, is not "smooth".

In contrast, the method of (regression) splines provides a smooth representation of the data by piecewise polynomial functions with a specified degree of differentiability (see, for example, a survey paper by Wegman and Wright (1983) and references therein). Practical and numerical aspects of smoothing random data by splines under monotonicity constraints have been studied in a number of publications (for example, Wright and Wegman, 1980; Utreras, 1985; Ramsay, 1988, 1998).

As mentioned above, we concentrate on the situation where the regression function $f(x)=E(Y|X=x)$ belongs to a known parametric family of continuously differentiable functions. We then address the problem of testing whether $f(x)$ is monotonically increasing (decreasing) on the given interval $[a,b]$. It will be shown that a test statistic, derived from the difference between a monotone and unrestricted least-squares fit, has asymptotically a chi-bar-squared distribution. This result will provide us with a way to approximate a quantitative measure of significance of the monotonicity constraints (cf. Hastie and Tibshirani, section "Why monotone?" in the discussion of Ramsay, 1988, p. 451). In a sense, the testing procedure presented in this article is a generalization of testing the sign of the slope of a simple linear regression curve.

In a recent paper Bowman et al. (1998) suggested a nonparametric approach to testing regression monotonicity. Their method is analogous to Silverman's test (1983) of multimodality in density estimation and is based on the size of a critical bandwidth which is required to force the estimated regression function to be monotone. Bootstrapping is used to calculate the null distribution of the test statistic. Our approach is different in several respects. It would be certainly of interest to compare these two approaches numerically.

2. Theoretical background

In this section, we describe our basic assumptions and derive some theoretical results. Consider the regression model

$$Y_i = f(X_i, \theta_0) + \varepsilon_i, \tag{2.1}$$

where $f:\mathbb{R}\times\mathbb{R}^k\to\mathbb{R}$ is a known function and (X_i,Y_i), $i=1,\dots,n$, are independent and identically distributed observations from the bivariate distribution of (X,Y). It is assumed that the model is such that X_i and the errors ε_i are independent, $\mathbb{E}(\varepsilon_i)=0$, $\mathrm{var}(\varepsilon_i)=\sigma^2$, $i=1,\dots,n$; and that the true value, θ_0, of the parameter vector θ belongs to a subset of $\mathbb{R}^k$.

Unless states otherwise, we assume that the function $f(x,\theta)$ is linear in θ. That is,

$$f(x,\theta):=\theta_1 g_1(x)+\cdots+\theta_k g_k(x)=\theta^{\mathrm{T}}g(x), \tag{2.2}$$

where $\theta=(\theta_1,\dots,\theta_k)^{\mathrm{T}}$ and $g(x)=(g_1(x),\dots,g_k(x))^{\mathrm{T}}$, and the functions $g_1(x),\dots,g_k(x)$ are supposed to be known. (We assume that vectors are written as columns, by A^{T} we denote the transpose of matrix (vector) A and use $g'(x)$ and $g''(x)$ to denote the vector of componentwise first- and second-order derivatives of $g(x)$, respectively.) Two important examples are: (i) polynomial regression with $g_j(x)=x^{j-1}$; $j=1,\dots,k$; and (ii) spline regression with $g_j(x)$ given by basis splines depending on chosen knots and a given order. Unless stated otherwise, we assume that the basis functions $g_j(x)$; $j=1,\dots,k$, are twice continuously differentiable.

Let us remark that the linearity assumption (2.2) is made for the sake of simplicity. The cases where $f(x,\theta)$ is nonlinear and smooth can be investigated in a similar way, we will discuss that later.

We are interested in testing monotonicity of the regression function $f(\cdot,\theta_0)$. Specifically, we consider testing whether this function is monotonically increasing (or rather nondecreasing) on a given interval $[a,b]$. Since the function $f(\cdot,\theta)$ is differentiable, it is nondecreasing on the interval $[a,b]$ if and only if the derivative $\partial f(x,\theta)/\partial x$ is nonnegative for all $x\in[a,b]$. Therefore, the monotonicity condition is equivalent to the statement that θ_0 belongs to the parameter set

$$\Theta:=\left\{\theta\in\mathbb{R}^k:\ \partial f(x,\theta)/\partial x\geqslant 0 \text{ for all } x\in[a,b]\right\}. \tag{2.3}$$

Consequently, we arrive at testing the null hypothesis H_0: $\theta_0\in\Theta$ against the unrestricted alternative H_1: $\theta_0\in\mathbb{R}^k$.

Now let us consider the restricted ($\hat\theta=\hat\theta_n$) and the unrestricted ($\tilde\theta=\tilde\theta_n$) minimizers of the function

$$S(\theta):=\sum_{i=1}^{n}[Y_i-f(X_i,\theta)]^2. \tag{2.4}$$

That is, $\hat\theta:=\arg\min_{\theta\in\Theta}S(\theta)$ and $\tilde\theta:=\arg\min_{\theta\in\mathbb{R}^k}S(\theta)$. We consider the following test statistic:

$$T:=S(\hat\theta)-S(\tilde\theta)=\min_{\theta\in\Theta}S(\theta)-\min_{\theta\in\mathbb{R}^k}S(\theta). \tag{2.5}$$

The statistic T measures a discrepancy between the monotone and unrestricted least-squares fit of the data. Of course, if the errors ε_i have normal distribution $\mathrm{N}(0,\sigma^2)$, then the statistic T becomes the log-likelihood ratio test statistic for testing the null hypothesis H_0 against the alternative H_1. It is reasonable to reject the null hypothesis for large values of T. In order to calculate the corresponding significance levels, we evaluate now the asymptotic (large sample) distribution of T.

Note that the function $S(\theta)$ can be written in the form

$$S(\theta) = (Y - A\theta)^{\mathrm{T}}(Y - A\theta), \tag{2.6}$$

where A is the $n \times k$ matrix with (i,j)-element $a_{ij} := g_j(X_i)$; $i = 1,\ldots,n$; $j = 1,\ldots,k$; and Y is the $n \times 1$ vector $Y = (Y_1,\ldots,Y_n)^{\mathrm{T}}$. It is well known that the unrestricted least-squares estimator $\tilde{\theta}$ is given by

$$\tilde{\theta} = (A^{\mathrm{T}}A)^{-1}A^{\mathrm{T}}Y, \tag{2.7}$$

assuming, of course, that the matrix $A^{\mathrm{T}}A$ is nonsingular. Moreover, the statistic T can be written in the form

$$T = \min_{\theta \in \Theta} (\tilde{\theta} - \theta)^{\mathrm{T}}(A^{\mathrm{T}}A)(\tilde{\theta} - \theta). \tag{2.8}$$

This follows immediately from the well-known formula for orthogonal projections

$$\|Y - A\theta\|^2 = \|Y - A\tilde{\theta}\|^2 + \|A\tilde{\theta} - A\theta\|^2, \tag{2.9}$$

where $\|x\| = (x^{\mathrm{T}}x)^{1/2}$ denotes the Euclidean norm of a vector x.

Suppose that the expectations $\mathbb{E}[g_i(X)^2]$, $i = 1,\ldots,k$, are finite and consider the $k \times k$ matrix $\Omega = [\omega_{ij}]$ with

$$\omega_{ij} := \mathbb{E}\{g_i(X)g_j(X)\}, \quad i,j = 1,\ldots,k. \tag{2.10}$$

Then it follows by the Law of Large Numbers that $n^{-1}A^{\mathrm{T}}A$ converges with probability one to Ω. It also follows by the Central Limit Theorem that, under model (2.1)–(2.2), $n^{1/2}(\tilde{\theta}_n - \theta_0)$ converges in distribution to a random vector Z which is normally distributed, $\mathrm{N}(0, \sigma^2\Omega^{-1})$.

Suppose for the moment that $f(x,\theta)$ is a smooth (not necessarily linear in θ) function and consider the parameter set Θ given in (2.3). This parameter set is defined by an infinite number of constraints $q_x(\theta) \geqslant 0$, $x \in [a,b]$, where $q_x(\theta) := \partial f(x,\theta)/\partial x$. For $\theta \in \Theta$ consider the set

$$M(\theta) := \{x \in [a,b]\colon q_x(\theta) = 0\} = \left\{x \in [a,b]\colon \frac{\partial f(x,\theta)}{\partial x} = 0\right\} \tag{2.11}$$

corresponding to the active at θ constraints. Note that since $\theta \in \Theta$ it follows that if the set $M(\theta)$ is nonempty, then this set coincides with the set of minimizers of $\partial f(\cdot,\theta)/\partial x$ over the interval $[a,b]$. It is said that the extended Mangasarian–Fromovitz constraint qualification holds at $\theta_0 \in \Theta$ if there exists a vector $\xi \in \mathbb{R}^k$ such that

$$\xi^{\mathrm{T}}\nabla q_x(\theta_0) > 0 \quad \text{for all } x \in M(\theta_0). \tag{2.12}$$

(By $\nabla q_x(\theta) := (\partial q_x(\theta)/\partial\theta_1,\ldots,\partial q_x(\theta)/\partial\theta_k)^{\mathrm{T}}$ we denote the gradient vector of the function $q_x(\cdot)$ at θ.) Under the above constraint qualification the set Θ is approximated at the point θ_0 by the cone

$$C := \{\zeta \in \mathbb{R}^k\colon \zeta^{\mathrm{T}}\nabla q_x(\theta_0) \geqslant 0,\ x \in M(\theta_0)\}. \tag{2.13}$$

More precisely the following result holds (see, e.g., Bonnans and Shapiro, 2000, p. 66).

Proposition 2.1. *Suppose that $q_x(\theta)$ is differentiable in θ with $\nabla q_x(\theta)$ being continuous jointly in $x \in [a,b]$ and $\theta \in \mathbb{R}^k$, the set $M(\theta_0)$ is nonempty and the extended Mangasarian–Fromovitz constraint qualification holds at $\theta_0 \in \Theta$. Then the cone C, defined in* (2.13)*, has the following approximation properties*:

$$\inf_{\zeta \in C} \|(\theta - \theta_0) - \zeta\| = o(\|\theta - \theta_0\|), \quad \theta \in \Theta \tag{2.14}$$

and

$$\inf_{\theta \in \Theta} \|(\theta - \theta_0) - \zeta\| = o(\|\zeta\|), \quad \zeta \in C. \tag{2.15}$$

Suppose now that $f(x,\theta)$ is linear in θ and is given in form (2.2). Then

$$q_x(\theta) = \theta_1 \frac{dg_1(x)}{dx} + \cdots + \theta_k \frac{dg_k(x)}{dx} = \theta^{\mathrm{T}} g'(x) \tag{2.16}$$

and $\nabla q_x(\theta) = g'(x)$, and hence

$$C = \{\zeta \in \mathbb{R}^k : \zeta^{\mathrm{T}} g'(x) \geqslant 0,\ x \in M(\theta_0)\}. \tag{2.17}$$

Moreover, the set Θ is convex and the extended Mangasarian–Fromovitz constraint qualification is equivalent to the Slater condition: there exists a point θ^* such that $\partial f(x,\theta^*)/\partial x$ is greater than zero for all $x \in [a,b]$.

Under the Slater condition and if $M(\theta_0)$ is nonempty, then the cone C given in (2.17) coincides with the tangent cone to the set Θ at the point θ_0. Note that since Θ is convex, it follows that $\Theta - \theta_0$ is included in the tangent to Θ at θ_0 cone, and hence the property (2.14) trivially holds with the right-hand side of (2.14) being zero. Note also that if θ_0 is an interior point of Θ, then $C = \mathbb{R}^k$. However, we are particularly interested in the situations where θ_0 is a boundary point of Θ. Under the Slater condition, θ_0 lies on the boundary of Θ if and only if the set $M(\theta_0)$ is nonempty.

Theorem 2.1. *Suppose that model* (2.1)–(2.2) *holds and that the matrix Ω is nonsingular. Then the statistic T converges in distribution to*

$$U := \min_{\zeta \in C} (Z - \zeta)^{\mathrm{T}} \Omega (Z - \zeta), \tag{2.18}$$

where Z is a normally distributed $N(0, \sigma^2 \Omega^{-1})$ *random vector and C is the cone tangent to Θ at θ_0.*

Proof. We have that

$$T = \min_{\theta \in \Theta} n(\tilde{\theta}_n - \theta)^{\mathrm{T}} [n^{-1}(A^{\mathrm{T}} A)](\tilde{\theta}_n - \theta).$$

Since $n^{-1} A^{\mathrm{T}} A$ converges in probability to Ω, it is then not difficult to show that

$$T = \min_{\theta \in \Theta} n(\tilde{\theta}_n - \theta)^{\mathrm{T}} \Omega (\tilde{\theta}_n - \theta) + o_p(1).$$

The proof can then be completed as in the proof of Shapiro (1985, Lemma 2.2) using the asymptotic normality of $n^{1/2}(\tilde{\theta}_n - \theta_0)$ and the approximating properties (2.14) and (2.15) of the tangent cone C. □

Recall that, under the Slater condition, the tangent cone to Θ at θ_0 can be written in the form (2.17). If the function $f(x,\theta)$ is smooth (not necessarily linear in θ), then it is possible to show by using the approximating properties (2.14) and (2.15) of the cone C, that under mild regularity conditions the asymptotic result of Theorem 2.1 still holds with the elements of the matrix Ω given by

$$\omega_{ij} := \mathbb{E}\left[\frac{\partial f(X,\theta_0)}{\partial \theta_i}\,\frac{\partial f(X,\theta_0)}{\partial \theta_j}\right].$$

It is also possible to show that the estimator $\hat{\theta}$ has an asymptotic distribution described in the following theorem (see Shapiro, 1989, Theorem 3.3).

Theorem 2.2. *Under the assumptions of Theorem* 2.1, $n^{1/2}(\hat{\theta}_n - \theta_0)$ *converges in distribution to the minimizer* $\bar{\zeta}(Z)$ *of the quadratic function* $(Z-\zeta)^{\mathrm{T}}\Omega(Z-\zeta)$ *over the cone* C.

It is known that the random variable U, given in (2.18), normalized by the coefficient σ^{-2}, has a chi-bar-squared distribution. That is

$$\Pr\{\sigma^{-2}U \geqslant c\} = \sum_{i=0}^{k} w_i \Pr\{\chi_i^2 \geqslant c\}, \tag{2.19}$$

where w_i are nonnegative weights, $w_0 + \cdots + w_k = 1$, and χ_i^2 is a chi-squared random variable with i degrees of freedom, $\chi_0^2 \equiv 0$. For a discussion of basic properties of chi-bar-squared variables and technical details, the reader is referred to Robertson et al. (1988) and Shapiro (1988). Theorem 2.1 shows that the corresponding result holds for the test statistic T asymptotically, that is,

$$\lim_{n\to\infty} \Pr\{\hat{\sigma}^{-2}T \geqslant c\} = \sum_{i=0}^{k} w_i \Pr\{\chi_i^2 \geqslant c\}, \tag{2.20}$$

where $\hat{\sigma}^2$ is a consistent estimator of σ^2. A consistent estimator of σ^2 can be obtained in a standard way, since $\mathbb{E}\{S(\tilde{\theta})\} = (n-k)\sigma^2$ we can take

$$\hat{\sigma}^2 := (n-k)^{-1} S(\tilde{\theta}). \tag{2.21}$$

Eq. (2.20) suggests that the asymptotic distribution of T depends on the weights w_i which in turn are functions of the matrix Ω and the cone C. Since the cone C is determined by the set $M(\theta_0)$ of active constraints (which is not known a priori), evaluation of the weights w_i may pose quite a delicate problem. Let us consider the situation where $M(\theta_0)$ has a finite number of points, say $M(\theta_0) = \{x_1^*, \ldots, x_\ell^*\}$. This, of course, always happens in the case of polynomial regression. Define vectors $b_i := g'(x_i^*)$ and matrix $B := [b_1, \ldots, b_\ell]$. Assuming that vectors $b_1, \ldots, b_\ell$, are linearly independent, the weights are given by

$$w_i = \begin{cases} w_{\ell-i}^*(\ell, B^{\mathrm{T}}\Omega^{-1}B), & i = 0, \ldots, \ell, \\ 0, & i = \ell+1, \ldots, k, \end{cases} \tag{2.22}$$

where $w_j^*(\ell, V)$, $j = 0, \ldots, \ell$, denote the weights corresponding to the chi-bar-squared distribution with covariance matrix V and the cone given by the positive orthant $\mathbb{R}_+^\ell$

(see Shapiro, 1988, Section 5, for details). In particular, for $\ell=1$ we have $w_0=w_1=\frac{1}{2}$, and for $\ell=2$ we have $w_0=(2\pi)^{-1}(\pi-\cos^{-1}\rho_{12})$, $w_1=\frac{1}{2}$ and $w_2=\frac{1}{2}-w_0$, where ρ_{12} is the $(1,2)$-element of the correlation matrix corresponding to the covariance matrix $V=B^{\mathrm{T}}\Omega^{-1}B$. There are also closed form expressions for the weights $w_j^*(\ell,V)$ for $\ell=3$ and $\ell=4$ (Kudô, 1963; Shapiro, 1985). For $\ell>4$, Kudô (1963) proposed a recurrence formula which is computationally workable for moderate values of ℓ (see Bohrer and Chow, 1978). It appears that, at least in the case of a polynomial regression, in practical applications the weights w_i should be calculated for small or moderate values of ℓ.

In principle, it is possible to evaluate the weights w_i, and hence, by using Eq. (2.20), to calculate (approximately) the critical values of the test statistic T, if consistent estimates of the matrices Ω and B are available. This leaves the practical problem of estimating the set $M(\theta_0)$. In the case of testing inequality linear constraints in linear regression, the least favorable distribution occurs when all inequality constraints are active (see Robertson and Wegman, 1978, Theorems 2.1, 2.2). The same holds here for local alternatives. In other words, from a conservative point of view one should be prepared for a worst situation and include all *suspicious* points in an estimator of the set $M(\theta_0)$.

3. Technical considerations

In trying to implement the test procedure based on the statistic T given in (2.5) or (2.8), there are several technical problems that must be resolved. In this section we describe an approach for resolving these technical problems. We assume that model (2.1)–(2.2) holds, and that the matrix Ω is nonsingular. We briefly discuss the case where the true regression function is not a member of the class of fitted models in Section 3.3.

3.1. Computing the test statistic

In order to compute the restricted estimate $\hat{\theta}$, and hence the test statistic T, one has to solve a semi-infinite (mathematical) programming (SIP) problem. An SIP problem involves optimization of an objective function in a *finite* number of variables, over a feasible region defined by an infinite number of constraints. One way of solving such problems is by using a simple discretization method—replacing the infinite set of constraints by a finite subset of them. This finite subset consists of those constraints corresponding to the values of x on a finite grid. The resulting problem is a quadratic programming problem that can be easily solved, for example, by using the QPROG/DQPROG procedures of the IMSL software library. Of course, the solution thus obtained is actually an approximation to the real solution. We have used this method in the simulations, to be described below, mainly because it was readily available and so far seems adequate for the task.

A somewhat more efficient discretization method, one that was originally proposed for a linear objective function by Hettich (1986), can be adapted to our problem. It is a more complicated, iterative scheme and since our main goal is to provide a useful

and relatively straightforward tool for testing the hypothesis of monotonicity, we have not attempted to implement it at this stage. A survey of other methods for solving SIP problems can be found in Fiacco and Kortanek (1983), Hettich and Kortanek (1993) and Goberna and Lopez (1998).

3.2. Determining the roots of the derivative

In order to derive an appropriate reference distribution, one has to decide how many roots the derivative function has, and if there is more than one root, to evaluate where these roots are located. This is the problem of estimating the set $M(\theta_0)$.

Let us assume that the true value θ_0 of the parameter vector lies on the boundary of the set Θ, and that the set $M(\theta_0)$ is finite. This means that we are testing for lack of monotonicity with respect to a null model that has only a finite number of stationary points in the open interval (a,b). (Moreover, we may also wish to add a priori information such as that "there are no stationary points within a "small" distance δ from the endpoints of $[a,b]$".) The problem is then to identify the elements of $M(\theta_0)$. First, we consider conditions under which it is theoretically possible to estimate these stationary points of $f(x,\theta_0)$. Second, we consider a procedure to decide whether to include a candidate minimizer of $\tilde{\theta}_n^{\mathrm{T}} g'(x)$ as a point in $M(\theta_0)$.

Under the null hypothesis, the derivative $\partial f(x,\theta_0)/\partial x$ is nonnegative for all $x \in [a,b]$ and hence the roots of this derivative are also its (global) minimizers on this interval. Let us consider the set $L(\theta)$ of *local* minimizers of the function $\partial f(\cdot,\theta)/\partial x$ on the interval (a,b). Note that by the standard optimality conditions, every point x of the set $L(\theta)$ satisfies the equation

$$\partial^2 f(x,\theta)/\partial x^2 = 0. \tag{3.1}$$

It follows that $M(\theta_0) \subset L(\theta_0)$, where the inclusion is strict if $\partial f(\cdot,\theta_0)/\partial x$ possesses local minimizers which are not global. This suggests that the elements of $L(\tilde{\theta})$ (or $L(\hat{\theta})$) are natural candidates for estimators of the elements of $M(\theta_0)$. In this respect the following result if useful.

Proposition 3.1. *Suppose that*: (i) *the functions* $g_1(x),\dots,g_k(x)$ *are three times continuously differentiable*, (ii) $L(\theta_0)$ *contains a finite number of points*; *say* $L(\theta_0) = \{x_1^*,\dots,x_m^*\}$, (iii) *the derivatives* $\partial^3 f(x_i^*,\theta_0)/\partial x^3$; $i = 1,\dots,m$, *are positive.*

Then: (a) *For all* θ *sufficiently close to* θ_0, *the set* $L(\theta)$ *contains the same number of points, say* $L(\theta) = \{\bar{x}_1(\theta),\dots,\bar{x}_m(\theta)\}$, (b) $\bar{x}_i(\theta)$ *tend to* $\bar{x}_i(\theta_0) = x_i^*$, $i = 1,\dots,m$, *as* $\theta \to \theta_0$, (c) $\bar{x}_i(\theta)$ *are differentiable at* θ_0 *and the corresponding gradients are given by*

$$\nabla \bar{x}_i(\theta_0) = -[\theta_0^{\mathrm{T}} g'''(x_i^*)]^{-1} g''(x_i^*), \quad i = 1,\dots,m. \tag{3.2}$$

Results (a)–(c) of Proposition 3.1 follow easily from the Implicit Function Theorem, applied to Eq. (3.1) at the points $(x_1^*,\theta_0),\dots,(x_m^*,\theta_0)$, and from the fact that $\partial^2 f(x,\theta)/\partial x^2$ tends to $\partial^2 f(x,\theta_0)/\partial x^2$ as $\theta \to \theta_0$ uniformly in $x \in [a,b]$.

In some situations, as for example in the case of cubic splines, the first assumption of Proposition 3.1, concerning three times continuous differentiability of $g_i(x)$, is too strong. This assumption can be relaxed by employing the Implicit Function Theorem due to Clarke (1983, pp. 255–256). Suppose that the functions $g_i(x)$; $i = 1,\ldots,k$, are twice differentiable and that the corresponding second derivatives $g_i''(x)$ are Lipschitz continuous on $[a,b]$. Consider the function $\partial^2 f(\cdot,\theta_0)/\partial x^2 = \theta_0^{\mathrm{T}} g''(\cdot)$. Its generalized gradient $\partial(\theta_0^{\mathrm{T}} g''(x))$, at x, is defined as the convex hull of all limits of the form $\lim_{n\to\infty} \theta_0^{\mathrm{T}} g'''(x_n)$ such that $x_n \to x$ and such that the derivatives $g_i'''(x_n)$, $i = 1,\ldots,k$, exist. Note that since $g_i''(x)$ are Lipschitz continuous, their derivatives $g_i'''(x)$ exist for all $x \in [a,b]$ except possibly on a set of Lebesgue measure zero. The following proposition is then a refinement of Proposition 3.1.

Proposition 3.2. *Suppose that*: (i) *the functions* $g_1(x),\ldots,g_k(x)$ *are twice differentiable and that their second-order derivatives* $g_i''(x)$ *are Lipschitz continuous on* $[a,b]$, (ii) $L(\theta_0)$ *contains a finite number of points, say* $L(\theta_0) = \{x_1^*,\ldots,x_m^*\}$, (iii) *the generalized gradients* $\partial(\theta_0^{\mathrm{T}} g''(x_i^*))$, $i = 1,\ldots,m$, *do not contain zero.*

Then assertions (a) *and* (b) *of Proposition* 3.1 *hold.*

These alternative assumptions of Proposition 3.2 hold for cubic splines if their second-order derivatives, which are piecewise linear functions, have no a flat piece to the left or to the right of each x_i^*.

Since $\tilde{\theta}$ and $\hat{\theta}$ are consistent estimators of θ_0, we have that under the assumptions of Proposition 3.1 (or Proposition 3.2), $L(\tilde{\theta})$ and $L(\hat{\theta})$ provide consistent estimators of the set $L(\theta_0)$. However, the inclusion $M(\theta_0) \subset L(\theta_0)$ can be strict and consequently we still face the problem of selecting elements of the set $L(\tilde{\theta})$ (or $L(\hat{\theta})$) as estimators of the corresponding elements of $M(\theta_0)$. It is natural to take only those elements of $L(\tilde{\theta})$ (or $L(\hat{\theta})$) for which $\partial f(x,\tilde{\theta})/\partial x$ (or $\partial f(x,\hat{\theta})/\partial x$) are not "too much" larger than zero. In order to proceed, let us consider the function

$$\phi_I(\theta) := \min_{x\in I} \theta^{\mathrm{T}} g'(x), \tag{3.3}$$

where I is a closed subinterval of (a,b), and use the following result.

Proposition 3.3. *Suppose that*: (i) *the interval* I *contains one element, say* x_j^*, *of* $L(\theta_0)$, (ii) *the functions* $g_1(x),\ldots,g_k(x)$ *are twice differentiable.*

Then $n^{1/2}(\phi_I(\tilde{\theta}_n) - \phi_I(\theta_0))$ *converges in distribution to a normal random variable with zero mean and variance.*

$$\sigma^2 [g'(x_j^*)]^{\mathrm{T}} \Omega^{-1} [g'(x_j^*)]. \tag{3.4}$$

Proof. By a theorem of Danskin (1967), we have that the min-function $\phi_I(\theta)$ is differentiable at θ_0 and its gradient $\nabla\phi_I(\theta_0)$ is equal to $g'(x_j^*)$. It follows that

$$\phi_I(\theta) - \phi_I(\theta_0) = (\theta - \theta_0)^{\mathrm{T}} g'(x_j^*) + \mathrm{o}(\|\theta - \theta_0\|). \tag{3.5}$$

The proof can be completed using the asymptotic normality of $n^{1/2}(\tilde{\theta}_n - \theta_0)$. □

Moreover, we have that $\phi_I(\theta_0)$ is zero if x_j^* belongs to the set $M(\theta_0)$ and is greater than zero otherwise. This result leads us to the possibility of testing whether an element $x_j^* = \bar{x}_j(\theta_0)$ of $L(\theta_0)$ estimated by $\bar{x}_j(\tilde{\theta}_n)$ belongs to the set $M(\theta_0)$ or not.

Let $\tilde{x}_n$ be a (local) minimizer of $\tilde{\theta}_n^{\mathrm{T}} g'(\cdot)$. Consider the statistic $Q_n := n^{1/2}\,\tilde{\theta}_n^{\mathrm{T}} g'(\tilde{x}_n)$. If $\tilde{x}_n$ is an estimator of a root x_0 of $\theta_0^{\mathrm{T}} g'(x)$, then by (3.5) the following approximate distributional result should hold:

$$Q_n \Rightarrow \mathrm{N}(0, \sigma^2 [g'(x_0)]^{\mathrm{T}} \Omega^{-1} [g'(x_0)]), \tag{3.6}$$

where "$\Rightarrow$" denotes convergence in distribution. If we find that Q_n is too large, then we would reject the hypothesis that $\tilde{x}_n$ is an estimator of an element of $M(\theta_0)$. In practice, we use $n^{-1}A^{\mathrm{T}}A$ instead of $\Omega, \hat{\sigma}^2$ instead of σ^2, and $\tilde{x}_n$ instead of x_0, when computing the z-score z_n based on the approximation in (3.6). Note that the univariate central limit result holds for this z-score. In the simulation study described below, we used a test size of $\alpha = 5\%$ for deciding whether each local minimum of $\tilde{\theta}_n^{\mathrm{T}} g'(x)$ should be included in $M(\theta_0)$.

Finally, we mention that under the assumptions of Proposition 3.1, it follows from (3.2) that

$$n^{1/2}[\tilde{x}_j(\tilde{\theta}) - x_j^*] \Rightarrow \mathrm{N}(0, \sigma^2 [\theta_0^{\mathrm{T}} g'''(x_j^*)]^{-2} [g''(x_j^*)]^{\mathrm{T}} [g''(x_j^*)]). \tag{3.7}$$

If we remove the assumption that "there are no stationary points within a positive distance δ from the endpoints of $[a, b]$", then we have to take into account a possibility that local minima of $\partial f(x, \theta_0)/\partial x$ can occur at one or both of the endpoints. Let us suppose for example that $\partial f(x, \theta_0)/\partial x$ has a local minimum at a (the discussion for b is similar), then $\partial f(a, \theta_0)/\partial x \geqslant 0$. We are interested in the case where $\partial f(a, \theta_0)/\partial x = 0$. If further $\partial^2 f(a, \theta_0)/\partial x^2 > 0$ (the equivalent condition at b is $\partial^2 f(b, \theta_0)/\partial x^2 < 0$) then, from smoothness assumptions on $f(x, \theta)$ we find that there is an interval $[a, \zeta]$ where $\partial^2 f(x, \theta_0)/\partial x^2$ is nonnegative and hence a is the only minimal point in it. If assumption (ii) of Proposition 3.3 holds, then since $\tilde{\theta}_n \to \theta_0$ and because of the continuity of the min-function ϕ_I, we have that

$$\min_{x \in I} \{\partial^2 f(x, \tilde{\theta}_n)/\partial x^2\} \to \min_{x \in I} \{\partial^2 f(x, \theta_0)/\partial x^2\} \tag{3.8}$$

with probability one as $n \to \infty$. Since $\partial^2 f(x, \theta_0)/\partial x^2$ is greater than zero for all x in I, we obtain that for n large enough $\partial f(x, \tilde{\theta}_n)/\partial x$ will be monotone on I, so that a will be the minimal point in it. If it turns out that $\partial^2 f(a, \tilde{\theta}_n)/\partial x^2$ is positive (or that $\partial^2 f(b, \tilde{\theta}_n)/\partial x^2$ is negative) we will test whether a (or b) is a point where $\partial f(x, \theta_0)/\partial x$ is zero.

Finally, having estimated $M(\theta_0)$ by $\tilde{M}$, say, we can approximate the reference distribution for the statistic T using the chi-bar-squared distribution based on $\tilde{M}$. Based on this approximate reference distribution, we can then compute an approximate p-value for the observed value of T.

3.3. Misspecified models

In this section, we discuss the following situation of misspecified models. Suppose that the data is analyzed according to model (2.1) while the true model is

$$Y_i = h(X_i) + \varepsilon_i, \quad i = 1, \ldots, n, \tag{3.9}$$

where $h(x)$ is a real-valued measurable function, such that $\mathbb{E}\{[h(X)]^2\}$ is finite and Y_i, X_i and ε_i satisfy the assumptions specified in Section 2. If the function $h(x)$ does not belong to the parametric family $f(x,\theta)$, then the model is said to be misspecified.

Consider the minimizer θ^* of the function

$$\mathbb{E}\{[h(X) - f(X,\theta)]^2\} = \mathbb{E}\{[h(X)]^2\} - 2\theta^{\mathrm{T}}\mathbb{E}\{h(X)g(X)\} + \theta^{\mathrm{T}}\Omega\,\theta. \tag{3.10}$$

It should be noted that if there is a unique θ_0 such that $h(x) = f(x,\theta_0)$, then obviously $\theta^* = \theta_0$. It is possible to show (e.g., White, 1980) that the unconstrained least-squares estimator $\tilde{\theta}$, given in (2.7), converges in probability (with probability one) to θ^* and that $n^{1/2}(\tilde{\theta} - \theta^*)$ converges in distribution to $\mathrm{N}(0, \Omega^{-1}V(\theta^*)\Omega^{-1})$, where

$$V(\theta^*) := \mathbb{E}\{[h(X) - g(X)^{\mathrm{T}}\theta^* + \varepsilon]^2 g(X)g(X)^{\mathrm{T}}\}.$$

Since $V(\theta^*) = \sigma^2\Omega + \Psi(\theta^*)$, where

$$\Psi(\theta^*) := \mathbb{E}\left\{[h(X) - g(X)^{\mathrm{T}}\theta^*]^2 g(X)g(X)^{\mathrm{T}}\right\},$$

we obtain that

$$n^{1/2}(\tilde{\theta} - \theta^*) \Rightarrow \mathrm{N}(0, \sigma^2\Omega^{-1} + \Omega^{-1}\Psi(\theta^*)\Omega^{-1}). \tag{3.11}$$

If $h(x)$ is close to $f(x,\theta^*)$, then (under mild regularity conditions) the elements of $\Psi(\theta^*)$ will be small. For example, if

$$[h(x) - g(x)^{\mathrm{T}}\theta^*]^2 \leqslant \frac{\varepsilon}{\max_{j\in\{1,\ldots,k\}}\omega_{jj}}$$

for all x in the range of X, then $|\psi_{ij}| \leqslant \varepsilon$ for $i,j = 1,\ldots,k$, where $|\psi_{ij}|$ is the ijth element of $\Psi(\theta^*)$. Since all the elements of Ω^{-1} are finite, for small enough ε all the elements of $\Omega^{-1}\Psi(\theta^*)\Omega^{-1}$ are also small. In this case, the bias caused by considering the variance of the asymptotic distribution of $n^{1/2}(\tilde{\theta} - \theta^*)$ as $\sigma^2\Omega^{-1}$ is small. Moreover, in case of misspecification, $\hat{\sigma}^2$ as defined in (2.21), converges in probability (with probability one) to

$$\sigma^2(\theta^*) := \mathbb{E}\{[h(X) - f(X,\theta^*)]^2\} + \sigma^2$$

(White, 1980). However, if the closeness condition between $h(x)$ and $f(x,\theta^*)$ holds, then

$$\sigma^2(\theta^*) - \sigma^2 \leqslant \frac{\varepsilon}{\max_{j\in\{1,\ldots,k\}}\omega_{jj}}.$$

In that case, σ^2 is estimated with only a small bias and hence $\hat{\sigma}^2(n^{-1}A^{\mathrm{T}}A)^{-1}$ is a good estimator, in the sense of consistency, of the variance of the asymptotic distribution of $n^{1/2}(\tilde{\theta} - \theta^*)$.

Suppose now that the point θ^* belongs to the set Θ given in (2.3). This means that the function $f(\cdot,\theta^*)$ is monotonically nondecreasing on the interval $[a,b]$. If $h(x)$

is increasing on $[a,b]$ and is also close to $f(\cdot,\theta^*)$, then testing the monotonicity hypothesis using the misspecified model will give correct results. It follows, in the case that $\theta^* \in \Theta$, that the statistic T, given in (2.5), converges in distribution to

$$U := \min_{\zeta \in C^*} (Z-\zeta)^{\mathrm{T}} \Omega (Z-\zeta), \tag{3.12}$$

where Z is a normally distributed $\mathrm{N}(0, \sigma^2\Omega^{-1} + \Omega^{-1}\Psi(\theta^*)\Omega^{-1})$ random vector and C^* is the cone tangent to Ω at θ^*. Since

$$(n^{-1}A^{\mathrm{T}}A)^{-1} n^{-1} \sum_{i=1}^{n} [(Y_i - \tilde{\theta}^{\mathrm{T}} g(X_i))^2 g(X_i) g(X_i)^{\mathrm{T}}] (n^{-1}A^{\mathrm{T}}A)^{-1}$$

converges with probability one to $\Omega^{-1}V(\theta^*)\Omega^{-1}$, the variance of Z can be estimated for large enough n.

If $\theta^* \notin \Theta$, which means that $f(\cdot,\theta^*)$ is not monotonically nondecreasing, then $T \to \infty$ as $n \to \infty$ and we will reject H_0 for large enough n. If $h(x)$ is also not monotonically nondecreasing and is close to $f(\cdot,\theta^*)$, then also in this case using the misspecified model will usually give correct results. It should be noted that it is possible that $h(x)$ is not monotone while $f(\cdot,\theta^*)$ is, or vice versa. If the class G, defined as $\{\theta^{\mathrm{T}} g(x)\colon \theta \in \mathbb{R}^k\}$, is chosen so that it is flexible enough, then this problem should only arise in pathological cases.

Now we discuss a special case of misspecification, the linear case that was experimented with in the simulations. Suppose that the true model is

$$Y_i = \theta^{\mathrm{T}} g(X_i) + \theta_{(1)}^{\mathrm{T}} g_{(1)}(X_i) + \varepsilon_i, \tag{3.13}$$

where

$$\theta := (\theta_1, \ldots, \theta_k)^{\mathrm{T}}, \quad g(x) := (g_1(x), \ldots, g_k(x))^{\mathrm{T}},$$

$$\theta_{(1)} := (\theta_{k+1}, \ldots, \theta_l)^{\mathrm{T}}, \quad g_{(1)}(x) := (g_{k+1}(x), \ldots, g_l(x))^{\mathrm{T}}$$

and functions $g_j(x)$, $j = 1, \ldots, l$, and variables X_i, Y_i, ε_i, $i = 1, \ldots, n$, obey the same assumptions as in Section 2. But, instead of fitting model (3.13), the misspecified model (2.1) is fitted. In this case $\tilde{\theta}$ converges with probability one to $\theta + \Omega^{-1}\Omega_{01}\theta_{(1)}$ and $\hat{\sigma}^2$ converges with probability one to $\sigma^2 + \theta_{(1)}^{\mathrm{T}}(\Omega_{11} - \Omega_{01}^{\mathrm{T}}\Omega^{-1}\Omega_{01})\theta_{(1)}$, where

- Ω is as above;
- Ω_{11} is an $(l-k) \times (l-k)$ matrix with $\Omega_{11}(i,j) := \mathbb{E}\{g_{k+i}(X) g_{k+j}(X)\}$ and $i,j = 1, \ldots, l-k$; is assumed to be nonsingular; and
- Ω_{01} is a $k \times (l-k)$ matrix where $\Omega_{01}(i,j) := \mathbb{E}\{g_i(X) g_{k+j}(X)\}$, $i = 1, \ldots, k$; $j = 1, \ldots, l-k$.

Define $(\theta_M^*)^{\mathrm{T}} := ((\theta^*)^{\mathrm{T}}, \bar{0})$, where $\bar{0}$ is a row vector of $l-k$ zeros. It can be shown that θ^* is the element of $\mathbb{R}^k$ for which $(\theta_M^*)^{\mathrm{T}}$ is closest to $\theta_0 = (\theta, \theta_{(1)})^{\mathrm{T}}$ in the sense that θ_M^* is the argmin of the following mathematical programming problem

$$\min_{\eta \in \mathbb{R}^l} (\eta - \theta_0)^{\mathrm{T}} \Omega_M (\eta - \theta_0) \quad \text{subject to } \eta_{(1)} = 0, \tag{3.14}$$

where Ω_M is an $l \times l$ matrix with (i,j) element $\Omega_M(i,j) = \mathbb{E}\{g_i(X) g_j(X)\}$, $i,j = 1, \ldots, l$; and $\eta_{(1)}$ consists of the last $(l-k)$ elements of η.

Furthermore, $n^{-1}T \to (\theta_M^* - \bar{\theta}^*)^{\mathrm{T}}\Omega_M(\theta_M^* - \bar{\theta}^*)$ with probability one as $n \to \infty$, where $\bar{\theta}^*$ is the optimal solution of

$$\min_{\vartheta \in \Theta_M} (\theta_M^* - \vartheta)^{\mathrm{T}}\Omega_M(\theta_M^* - \vartheta) \tag{3.15}$$

with

$$\Theta_M := \{\eta \in \mathbb{R}^k : (g'(x)^{\mathrm{T}}, g'_{(1)}(x)^{\mathrm{T}})\eta \geqslant 0 \text{ for all } x \in [a,b],\ \eta_{(1)} = 0\}.$$

Alternatively, $\bar{\theta}^*$ can also be shown to be the optimal solution of the problem

$$\min_{\vartheta \in \Theta_M} (\vartheta - \theta_0)^{\mathrm{T}}\Omega_M(\vartheta - \theta_0). \tag{3.16}$$

Consider now the case where X is not a random variable and its values are predetermined systematically. If we further assume that they are chosen in such a way that

$$\text{(i)} \quad n^{-1}A^{\mathrm{T}}A \to \Omega, \tag{3.17}$$

$$\text{(ii)} \quad n^{1/2}[n^{-1}A^{\mathrm{T}}A_{(1)} - n^{-1}(A^{\mathrm{T}}A)\Omega^{-1}\Omega_{01}] \to 0 \quad \text{as } n \to \infty \tag{3.18}$$

and the previous assumptions about the errors ε_i hold, then we obtain again that $n^{1/2}(\tilde{\theta} - \theta^*)$ converges in distribution to $\mathrm{N}(0, \sigma^2\Omega^{-1})$ (using the same technique as in Section 4.2 of Dhrymes, 1970). Properties (3.17) and (3.18) hold if the x grid is uniformly spaced on $[a,b]$. If σ^2 is known then this case is reduced to the case of Theorem 2.1. But if it is not known and we use the biased estimator $\hat{\sigma}^2$, as defined in (2.21), then $T/\hat{\sigma}^2$ will tend to be smaller than T/σ^2 and the test will have a smaller size than its nominal values.

4. Simulation results

The simulation study reported below is somewhat limited, and serves only to demonstrate the efficacy of the methodology described above in a specific setting. Given its feasibility in this basic setting, we justify trying to verify its applicability to more complicated settings. Some modifications of the methodology are already being contemplated based on some of the results described below. We will give only a brief summary of the simulation results. A detailed report and analysis of simulation results are given in Doveh et al. (1990).

The limited setting is that of three data sets. One is generated from a cubic polynomial regression and the other two from a fifth-order polynomial regression; all with normal (additive) errors. The models fitted are third-, fourth- and fifth-order polynomials. The domain of the regression is $[0,1]$.

Various (IMSL-based) random number generators were used in the preliminary stages in order to double-check the results obtained. No significant difference between generators was found.

Table 1
Comparing the size of the monotonicity test with the nominal size. ($q = 0, R = 0, 5000$ simulations)

m	σ	n	$\alpha = 0.01$					$\alpha = 0.05$				
			$p = 3$	$p = 4$		$p = 5$		$p = 3$	$p = 4$		$p = 5$	
			ENE	*E*	*N*	*E*	*N*	*ENE*	*E*	*N*	*E*	*N*
1	2	18	0.015	0.024	0.024	0.019	0.028	0.061	0.088	0.090	0.068	0.094
1	2	32	0.015	0.020	0.020	0.015	0.021	0.058	0.070	0.070	0.055	0.065
1	2	100	0.010	0.012	0.012	0.012	0.013	0.053	0.060	0.060	0.058	0.058
2	1.75	18	0.025	0.035	0.046	0.035	0.060	0.076	0.106	0.144	0.116	0.172
2	1.75	32	0.011	0.019	0.023	0.012	0.027	0.057	0.076	0.095	0.060	0.011
2	1.75	100	0.011	0.018	0.019	0.007	0.016	0.055	0.070	0.076	0.036	0.071
3	1.25	64	0.013	0.016	0.019	0.014	0.014	0.52	0.058	0.071	0.064	0.064
3	1.25	100	0.012	0.014	0.016	0.016	0.016	0.047	0.054	0.064	0.062	0.062
3	12.5	150	0.009	0.011	0.012	0.012	0.012	0.046	0.050	0.055	0.055	0.055

The simulation results were based on three sets of data. The first set is generated from the model

$$y_i = \frac{640}{3}(x_i - 0.5)^3 - qx_i + e_i, \tag{4.1}$$

the second set is generated from the model

$$y_i = \tfrac{2}{30}(x_i - 9)^5 - \tfrac{2}{6}(x_i - 9)^4(x_i - 0.5) + \tfrac{2}{3}(x_i - 9)^3(x_i - 0.5)^2 + 2960 - qx_i + e_i \tag{4.2}$$

and the third set is generated from the model

$$y_i = 100(x_i - 0.72)^5 - 500(x_i - 0.72)^4(x_i - 0.25) + 1000(x_i - 0.72)^3(x_i - 0.25)^2 - qx_i + e_i, \tag{4.3}$$

where x_i, $i = 1, \ldots, n$, is a (pseudo-) random sample from the uniform distribution U[0, 1], and e_i, $i = 1, \ldots, n$, is a (pseudo-) random sample from the $N(0, \sigma^2)$ distribution. If $q = 0$ then the regression function $f(x)$ is on the boundary of Θ defined for polynomial regression of order three or higher. Power calculations are performed by setting $q > 0$, which corresponds to a single negative minimum of the derivative of the true regression function in the first two models (at $x = 0.5$) and two negative minima of the derivative (at $x = 0.25$ and 0.72) in the third model.

The statistics are reported on the basis of 5000 simulations for each case studied and reported in Table 1; and on the basis of 500 simulations for each case studied and reported in Table 2. A case is defined by setting the following parameters;

- m, the model number (where $m = 1$ is for model (4.1), $m = 2$ is for model (4.2) and $m = 3$ is for model (4.3));

Table 2
Some power estimates—$\alpha = 0.05$

m	n	σ	q	R	$p=3$ *ENE*	$p=4$		$p=5$	
						E	N	E	N
1	18	1.0	10.00	42.43	1.00	0.98	0.98	0.79	0.80
1	18	1.0	1.67	7.07	0.21	0.23	0.23	0.14	0.16
1	18	2.0	20.00	42.43	1.00	0.99	1.00	0.90	0.91
1	18	2.0	3.33	7.07	0.21	0.25	0.25	0.13	0.19
1	100	1.0	4.24	42.43	1.00	1.00	1.00	0.84	0.84
1	100	1.0	0.71	7.07	0.20	0.21	0.21	0.13	0.13
1	100	2.0	8.48	42.43	1.00	1.00	1.00	0.84	0.84
1	100	2.0	1.41	7.07	0.21	0.22	0.22	0.14	0.14
2	18	0.50	1.77	15	0.27	0.27	0.27	0.09	0.13
2	18	0.50	4.12	35	0.97	0.96	0.97	0.76	0.79
2	18	1.50	5.30	15	0.31	0.31	0.35	0.23	0.29
2	18	1.50	12.37	35	1.00	1.00	1.00	1.00	1.00
2	100	0.50	0.75	15	0.32	0.33	0.33	0.10	0.10
2	100	0.50	1.75	35	0.99	0.99	0.99	0.74	0.74
2	100	1.50	2.25	15	0.31	0.35	0.35	0.09	0.13
2	100	1.50	5.25	35	0.99	0.99	0.99	0.90	0.91
3	100	0.50	0.75	15	0.23	0.24	0.25	0.29	0.29
3	100	0.050	1.75	35	0.66	0.66	0.69	0.80	0.80
3	100	0.75	1.12	15	0.29	0.31	0.33	0.30	0.30
3	100	0.75	2.62	35	0.87	0.85	0.87	0.84	0.84
3	100	1.25	1.88	15	0.40	0.41	0.46	0.32	0.32
3	100	1.25	4.37	35	0.95	0.94	0.96	0.92	0.92
3	150	0.50	0.60	15	0.22	0.23	0.24	0.30	0.30
3	150	0.50	1.43	35	0.69	0.69	0.70	0.80	0.80
3	150	0.75	0.90	15	0.33	0.33	0.33	0.30	0.30
3	150	0.75	2.14	35	0.86	0.86	0.87	0.83	0.83
3	150	1.25	1.50	15	0.44	0.42	0.45	0.32	0.32
3	150	1.25	3.57	35	0.93	0.93	0.94	0.90	0.90

- p, the order of the *fitted* polynomial;
- q, the degree of nonmonotonicity of the true regression function;
- n, the sample size;
- σ, the standard deviation of the error distribution.

For each case the statistics reported are the proportion of simulations, out of 5000 (or 500), for which the hypothesis of monotonicity is rejected based on obtaining (approximate) p-values less than 5% or 1%. The calculations may take into account the possibility of zeros at endpoints (denoted E), or may reject that possibility (denoted N). If the results in both cases (E and N) were identical, only one column was given for that case (denoted *ENE*). All the results are rounded to the nearest thousandth.

It was conjectured that the power should be a function of $R=q\sqrt{n}/\sigma$, as would be the case for the detection of a nonzero expectation based on an ordinary average, and this

is what the results show for the cases where the true function had only one minimum in the regression domain and a cubic polynomial was fitted. For this reason, R is also given in the tables. The conjecture does not hold in the cases where $m = 3$. We note that in every simulation run we used IMSL routine ZPORC/DZPORC for determining the zeros of a polynomial and if it failed to find any root then the ZPLRC/DZPLRC routine of IMSL was used for the same purpose. The simulations in case of two minima (model (4.3)) demanded more data than in case of one minimum (models (4.1) and (4.2)) since otherwise the two minima are not discovered.

From Table 1 we see that in the main, the estimated and nominal sizes of the tests are in good agreement, at least for moderate and large sample sizes. For this basic setting the method seems to be valid as far as test size is concerned. From Table 1, for the cases where $m = 1$ and $p > 3$, demonstrating the influence of overfitting a polynomial, we see that the estimated sizes are higher than the nominal ones. From the same table, cases with $m = 2$ and $m = 3$, demonstrating the influence of underfitting a polynomial, we see that nominal and estimated sizes are closer for third-order polynomials than for fourth- and even fifth-order polynomials (which is the true model). It should be noted that this happens not only in case of fifth-order models with only one minimum of the derivative, which has the same nature as a third-order polynomial in the domain of regression ($m = 2$), but also in the case of two minima in the domain ($m = 3$). In all the cases above the estimated sizes taking the ends into account are much closer to the nominal sizes than without them. Since we are usually interested in fixed size tests, it is recommended in this setting to take the ends into account.

The power calculations are also quite encouraging, and show that the procedure has the potential of providing a general method for testing for monotonicity. It turns out that except for very low σ, fitting the third-order polynomial gave usually better power results than the fifth-order polynomial even when the latter was the true model. Although fitting fourth-order polynomials shows good power results, it is not recommended because of its size problems, and since it does not have the overall nature of an ascending function.

In a preliminary attempt to extend the investigation beyond the basic setting described in (4.1)–(4.3), we also considered a true regression function of zig-zag type defined as follows:

$$f(x) = \begin{cases} \alpha(x-a), & x \leqslant a, \\ -\delta\dfrac{(x-a)}{(b-a)}, & a \leqslant x \leqslant b, \\ \beta(x-b) - \delta, & x \geqslant b. \end{cases} \tag{4.4}$$

In this case, when fitting the $p = 5$ model, there was a marked problem of local minima near *both* endpoints. Here, of course, we are dealing with the situation of a misspecified model, and one has to investigate how the approximating fifth-order polynomial to the regression function $f(x)$ of (4.4) really behaves (see the comments in Section 3.3). It will be interesting to check how the use of splines will improve the situation for this case. Another option would be to simply extend the order of the fitted

polynomial to allow for the additional flexibility required to mimic $f(x)$ adequately. All these issues will be considered in the continuation of the current research.

5. Concluding remarks

We suggested in this paper an approach to testing monotonicity of a regression function. Our approach is parametric and is based on the assumption that there exists a parametric family which can "sufficiently well" represent the data. The assumption that the model is linear in the parameters is not essential and is made for the sake of simplicity. A simulation study seems to indicate that the method is not sensitive to small misspecifications of the employed model. Our approach is very different from the nonparametric approach suggested in Bowman et al. (1998). In particular, the critical values of the considered test statistic are calculated (or rather approximated) analytically and a bootstrapping is not involved. In any case it would be interesting to compare the two methods numerically for various data sets.

References

Barlow, R.E., Bartholomew, D.J., Bremner, J.M., Brunk, H.D., 1972. Statistical Inference Under Order Restrictions. The Theory and Applications of Isotonic Regression. Wiley, New York.

Bohrer, R., Chow, W., 1978. Weights for one-sided multivariate inference. Appl. Statist. 27, 100–104.

Bonnans, J.F., Shapiro, A., 2000. Perturbation Analysis of Optimization Problems. Springer, New York.

Bowman, A.W., Jones, M.C., Gijbels, I., 1998. Testing monotonicity of regression. J. Comput. Graph. Statist., in preparation.

Clarke, F.H., 1983. Optimization and Nonsmooth Analysis. Wiley, New York.

Danskin, J.M., 1967. The Theory of Max–Min and its Applications to Weapon Allocation Problems. . Econometrics and Operations Research, Vol. 5. Springer, Berlin, New York.

Dhrymes, P.J., 1970. Econometrics, Statistical Foundations and Applications. Harper and Row, New York.

Doveh, E., Shapiro, A., Feigin, P.D., 1990. Testing of monotonicity in regression models. Mimeograph Series, Operations Research, Statistics and Economics, Vol. 401. Technion, Israel.

Fiacco, A.V., Kortanek, K.O. (Eds.) 1983. Semi-Infinite Programming and Applications, Lecture Notes in Economics and Mathematical Systems, Vol. 215. Springer, New York.

Goberna, M.A., Lopez, M.A., 1998. Linear Semi-Infinite Optimization. Wiley, Chichester.

Hettich, R., 1986. An implementation of a discretization method for SIP. Math. Prog. 34, 354–361.

Hettich, R., Kortanek, K.O., 1993. Semi-infinite programming: theory, methods, and applications SIAN Rev. 35, 380–429.

Kudô, A., 1963. A multivariate analogue of the one-sided test. Biometrika 50, 403–418.

Ramsay, J.O., 1988. Monotone regression splines in action (with discussion) Statist. Sci. 3, 425–461.

Ramsay, J.O., 1998. Estimating smooth monotone functions. J. Roy. Statist. Soc. Ser. B 60, 365–375.

Robertson, T., Wegman, E.J., 1978. Likelihood ratio tests for order restrictions in exponential families. Ann. Statist. 6, 485–505.

Robertson, T., Wright, F.T., Dykstra, R.L., 1988. Order Restricted Statistical Inference. Wiley, Chichester.

Shapiro, A., 1985. Asymptotic distribution of test statistics in the analysis of moment structures under inequality constraints. Biometrika 72, 133–144.

Shapiro, A., 1988. Towards a unified theory of inequality constrained testing in multivariate analysis. Int. Statist. Rev. 56, 49–62.

Shapiro, A., 1989. Asymptotic properties of statistical estimators in stochastic programming. Ann. Statist. 17, 841–858.

Silverman, B.W., 1983. Some properties of a test for multimodality based on kernel density estimates. In: Kingman, J.F.C., Reuter, G.E.H. (Eds.), Probability, Statistics and Analysis. Cambridge University Press, Cambridge, pp. 248–259.

Utreras, F.I., 1985. Smoothing noisy data under monotonicity constraints—existence, characterization and convergence rates. Numer. Math. 47, 611–625.

Wegman, E.J., Wright, I.W., 1983. Splines in statistics. J. Amer. Statist. Assoc. 78, 351–365.

White, H., 1980. Using least squares to approximate unknown regression functions. Int. Econom. Rev. 21, 149–170.

Wright, I.W., Wegman, E.J., 1980. Isotonic, convex and related splines. Ann. Statist. 8, 1023–1035.