NISS

# Testing the monotonicity or convexity of a function using regression splines

Jianqiang C. Wang and Mary C. Meyer

Technical Report 177
November 2010

National Institute of Statistical Sciences
19 T.W. Alexander Drive
PO Box 14006
Research Triangle Park, NC
www.niss.org

*The Canadian Journal of Statistics*
Vol. 34, No. 4, 2010, Pages ???-???
*La revue canadienne de statistique*

# Testing the monotonicity or convexity of a function using regression splines

Jianqiang C WANG and Mary C MEYER

*Key words and phrases:* constrained estimation; convex; generalized regression; isotonic; monotone; partial linear model; shape restrictions.
*MSC 2010*: Primary 62G08; 62G10; secondary 62G09; 62G20.

*Abstract:* Methods for smoothed isotonic or convex regression are useful in many applications. Sometimes the shape assumptions constitute a-priori knowledge about the regression function, but often the shape is part of the research question. The authors propose tests for monotonicity and convexity using constrained and unconstrained regression splines. The tests have good large-sample properties and the small-sample behavior is illustrated through simulations. Extensions to the partial linear model and the generalized regression model are presented.

**Title in French: we can supply this**

*Résumé :* Methods for smoothed isotonic or convex regression are useful in many applications. Sometimes the shape assumptions constitute a-priori knowledge about the regression function, but often the shape is part of the research question. The authors propose tests for monotonicity and convexity using constrained and unconstrained regression splines. The tests have good large-sample properties and the small-sample behavior is illustrated through simulations. Extensions to the partial linear model and the generalized regression model are presented.

## 1. INTRODUCTION AND BACKGROUND

Nonparametric methods in function estimation are useful when a parametric form is not known. Methods for estimating a smooth regression function include regression splines, kernel smoothing, and smoothing splines. There are several comprehensive books on scatterplot smoothing and semi-parametric regression, including Hastie & Tibshirani (1999) and Ruppert, Wand & Carroll (2003). Shape restrictions may be incorporated if the regression function is known to be monotone or convex. Monotone regression splines were proposed by Ramsay (1988), and extended to other shape restrictions by Meyer (2008). Mammen (1991) considered a two-step estimator combining kernel smoothing and isotonization. The monotone smoothing splines were considered by Mammen and Thomas-Agnan (1999).

Sometimes the shape assumptions constitute a-priori knowledge, such as in the relationship of tree height with its diameter, or the effect of the size of a dose of toxin on an organism. Sometimes the shape assumptions constitute part of the research question: does the expected number of mates of a bullfrog increase with his size? Does willingness to volunteer for psychology studies increase with the degree of extroversion of the subject? Suppose that

$$y_i = f(x_i) + \sigma\epsilon_i, \quad i = 1, \ldots, n, \tag{1}$$

for $x_i \in [0,1]$ and the $\epsilon_i$ are *iid* with mean zero and unit variance. There is no information about a parametric form for $f$, but it can be assumed that it possesses some degree of smoothness, such as continuous with continuous first derivative. Polynomial regression splines can be used for a simple and flexible estimate of $f$. This estimator shares many of the nice properties of the ordinary least-squares regression, and the optimal convergence rates were shown by Zhou, Shen & Wolfe (1998) to be attained. Under Assumptions 1-3 of section 3, they derived the point-wise convergence rates

$$\hat{f}(x) - f(x) = O_p(n^{-(p+1)/(2p+3)})$$

for any $x \in [0,1]$, if the number of knots grows as $O(n^{1/(2p+3)})$. Extensions to shape-restricted versions use cone-projection ideas, and the constrained splines attain the same convergence rate as the unconstrained splines (Meyer 2008).

The classical test for monotonicity was presented in Robertson, Wright & Dykstra (1988), using the unsmoothed monotone regression. For the regression model (1), let $H_0 : f \equiv c$, $H_1 : f$ is increasing, and $H_2 : f$ is not non-decreasing over the range. They presented tests for $H_0$ versus $H_1$ and for $H_1$ versus $H_2$, using the unsmoothed monotone regression estimator for the $H_1$ case. The test $H_1$ versus $H_2$ can be constructed under the assumptions of smoothness, using the monotone regression spline for the $H_1$ fit and the corresponding unconstrained spline (with the same knots) for the alternative fit. The null distribution of the test statistic $B_{12} = (SSE_1 - SSE_2)/SSE_1$ is that of a mixture of beta random variables under $H_0$, but the test is quite biased for general functions in $H_1$. In particular, when the function is increasing over most of the range and decreasing only over a part, the $B_{12}$ test can have smaller power than the test size.

Kernel smoothing is a useful tool in exploring curve monotonicity or convexity. Chaudhuri & Marron (1999) introduced the "SiZer map" as a graphical tool for examining "zero crossings of estimated derivatives," in which the authors took a scale-space point of view of smoothing and the choice of bandwidth was blurred. Global inference on SiZer was improved by Hannig & Marron (2006) and simultaneous inference was used to improve the approximation of the SiZer distribution. Bowman, Jones & Gijbels (1998) used a "critical bandwidth" for the local linear estimate of the function, that is the smallest bandwidth for which the estimate is monotone. Using bootstrap methods, they determine a $p$-value for the monotonicity hypothesis. Hall & Heckman (2000) pointed out that the above test does not perform well when the true function has flat or nearly-flat spots, and proposed a test that estimates local slopes and approximates the distribution of the (weighted) minimum. Ghosal, Sen & van der Vaart (2000) proposed a test for monotonicity involving a locally weighted version of Kendall's tau statistic. Juditsky & Nemirovski (2002) considered the general problem of determining if a signal generating a Gaussian random process is contained in a convex cone.

In this paper we use the constrained and unconstrained regression spline estimators to test the null hypothesis that $f$ satisfies the shape restrictions (and is smooth) against the alternative that $f$ is smooth. The test uses an estimate of the distribution of the minimum slope of the spline estimator under the null hypothesis, and is presented in the next section; theoretical properties are given in section 3. The extension to the generalized regression model is in section 4 and to the partial linear model in section 5, and examples of data analyses are given in section 6. The finite-sample performance of our proposed test is demonstrated through a simulation study in section 7. Detailed proofs for the results in section 3 are found in the appendix.

## 2. TEST FOR SHAPE ASSUMPTIONS

Suppose the $x_i$'s in (1) are in the interval $[0,1]$, and choose knots $0 = t_1 < \ldots < t_k = 1$. The number of knots $k$ increases as the sample size increases, and the knot placement for each sample size $n$ originates from a deterministic rule, such as equally spaced or at equal $x$-quantiles. The degree-$p$ $B$-spline basis functions $B_{1p}(x), \ldots, B_{mp}(x)$, for $m = k + p - 1$, are piecewise polynomials that span the space of of degree-$p$ piecewise polynomials with the given knots. For more details and formulas, see de Boor (2001). Define the $m \times k$ matrix $\boldsymbol{S}$ of slopes at the knots by $\boldsymbol{S}_{jl} = B'_{jp}(t_l)$. If the spline function

$$g(x) = \sum_{j=1}^{m} b_j B_{jp}(x) \tag{2}$$

is non-decreasing, then $\boldsymbol{S}^t\boldsymbol{b} \geq \boldsymbol{0}$. Similarly, if $\boldsymbol{T}$ is the matrix of the second derivatives of the basis functions at the knots, i.e., $T_{jl} = B''_{jp}(t_l)$, then for convex spline functions (2), we have $\boldsymbol{T}^t\boldsymbol{b} \geq \boldsymbol{0}$. Note that for quadratic regression splines, $\boldsymbol{S}^t\boldsymbol{b} \geq \boldsymbol{0}$ if and only if the function (2) is non-decreasing, and similarly for cubic (and lower order) regression splines, $\boldsymbol{T}^t\boldsymbol{b} \geq \boldsymbol{0}$ is a necessary and sufficient condition for convexity. Throughout this paper, we will consider quadratic $B$-spline basis functions for the test of monotonicity and cubic $B$-splines for the test of convexity, but the use of higher-order splines is briefly described in section 3. In what follows, we present the monotonicity test, and the convexity test is similarly derived.

The unconstrained estimate of $f$ minimizes the least-squares criterion over $\boldsymbol{b} \in \Re^m$:

$$\hat{\boldsymbol{b}} = \arg\min_{\boldsymbol{b}} \sum_{i=1}^{n} \left[ y_i - \sum_{j=1}^{m} b_j B_{jp}(x_i) \right]^2 . \tag{3}$$

Let $\boldsymbol{\Delta}$ denote the $n \times m$ design matrix for which the $j$th spline basis vector $[B_{jp}(x_1), \cdots, B_{jp}(x_n)]^t$ is the $j$th column. The minimization criterion in (3) may be written in vector form

$$\psi(\boldsymbol{b}; \boldsymbol{y}) = \boldsymbol{b}^t(\boldsymbol{\Delta}^t\boldsymbol{\Delta})\boldsymbol{b} - 2\boldsymbol{y}^t\boldsymbol{\Delta}\boldsymbol{b}, \tag{4}$$

and the least-squares estimate of the coefficients for the unconstrained spline is $\hat{\boldsymbol{b}} = (\boldsymbol{\Delta}^t\boldsymbol{\Delta})^{-1}\boldsymbol{\Delta}^t\boldsymbol{y}$. The $\hat{\boldsymbol{b}}^*$ that minimizes $\psi(\boldsymbol{b}; \boldsymbol{y})$ under the monotonicity constraints $\boldsymbol{S}^t\boldsymbol{b} \geq \boldsymbol{0}$ (or the convexity constraints $\boldsymbol{T}^t\boldsymbol{b} \geq \boldsymbol{0}$) is found using standard quadratic programming routines such as given in Fraser and Massam (1989). Briefly, the constrained solution is a least-squares projection onto a linear space of (possibly) smaller dimension than $m$. The projection routine searches for this subspace in an efficient manner. The `R` code for performing constrained least-squares as well as carrying out our proposed test can be downloaded from the author's website at:

`http://www.stat.colostate.edu/~meyer/code.htm`.

The testing procedure is outlined here; the theoretical properties are detailed in the next section. Let $\boldsymbol{\theta}$ be the vector of values of the true function at the observed $x_i$; that is, $\theta_i = f(x_i)$, and let $\boldsymbol{\beta} = (\boldsymbol{\Delta}^t\boldsymbol{\Delta})^{-1}\boldsymbol{\Delta}^t\boldsymbol{\theta}$; then the expected value of $\hat{\boldsymbol{b}}$ is $\boldsymbol{\beta}$ and its covariance matrix is $(\boldsymbol{\Delta}^t\boldsymbol{\Delta})^{-1}\sigma^2$. The vector of slopes of the unconstrained estimate at the knots, $\boldsymbol{S}^t\hat{\boldsymbol{b}}$, has mean $\boldsymbol{S}^t\boldsymbol{\beta}$ and covariance $\boldsymbol{S}^t(\boldsymbol{\Delta}^t\boldsymbol{\Delta})^{-1}\boldsymbol{S}\sigma^2$. The proposed test of $H_1 : f'(x) \geq 0$ on $[0, 1]$ versus the unconstrained alternative $H_2$ is as follows:

1. Obtain the unconstrained minimizer $\hat{\boldsymbol{b}}$ of (4) and determine the minimum of the slopes at the knots, i.e., calculate $s_{\min} = \min(\boldsymbol{S}^t\hat{\boldsymbol{b}})$.

2. If $s_{\min}$ is non-negative, then do not reject $H_1$.

3. Otherwise, estimate the distribution of the minimum slope under the null hypothesis, using $\hat{\boldsymbol{b}}^*$ for $\boldsymbol{\beta}$ and estimating the model variance $\sigma^2$ using the unconstrained spline:

$$\hat{\sigma}^2 = \frac{1}{n-m} \parallel \boldsymbol{y} - \boldsymbol{\Delta}\hat{\boldsymbol{b}} \parallel^2 . \tag{5}$$

4. If $s_{min}$ is smaller than the estimated $\alpha$-level percentile, then we reject $H_1$ in favor of $H_2$.

**Null distribution of $s_{\min}$:** We propose two approaches for obtaining the null distribution. The first approach is an analytic method which requires normality assumption or using normal approximation and the second method is resampling-based.

To implement the first approach, let $P_{\beta,\sigma^2}(r) = P(s_{\min} \le r)$ for a scalar $r$, and note that under the assumption of normality,

$$P_{\boldsymbol{\beta},\sigma^2}(r) \quad = \quad 1 - \int \cdots \int_{\{\boldsymbol{z}|\boldsymbol{z}-r\mathbf{1}\ge\mathbf{0}\}} \phi(\boldsymbol{z}; \boldsymbol{S}^t\boldsymbol{\beta}, \boldsymbol{S}^t(\boldsymbol{\Delta}^t\boldsymbol{\Delta})^{-1}\boldsymbol{S}\sigma^2)d\boldsymbol{z},$$

where $\phi(\boldsymbol{z}; \boldsymbol{u}, \Sigma)$ denotes the multivariate normal density with mean vector $\boldsymbol{u}$ and variance-covariance matrix $\Sigma$. The inequality in the integral limits is component-wise. In constructing the test of monotonicity, we define

$$Q_\alpha = \inf\left\{r \mid P_{\boldsymbol{\beta}^*,\sigma^2}(r) \ge \alpha\right\},$$

where $\boldsymbol{\beta}^*$ minimizes $\psi(\boldsymbol{\beta}; \boldsymbol{\theta})$ over $\boldsymbol{S}^t\boldsymbol{\beta} \ge \mathbf{0}$. Note that $\boldsymbol{\beta}^*$ may differ from $\boldsymbol{\beta}$ if the underlying function is not monotone. Both $\boldsymbol{\beta}^*$ and $\sigma^2$ are unknown in practice, so we estimate $Q_\alpha$ by $\widehat{Q}_\alpha$ defined below

$$\widehat{Q}_\alpha = \inf\left\{r \mid P_{\hat{\boldsymbol{b}}^*,\hat{\sigma}^2}(r) \ge \alpha\right\},$$

where $\hat{\boldsymbol{b}}^*$ denotes the coefficients of the constrained spline estimate for $\boldsymbol{\beta}$, and $\hat{\sigma}^2$ denotes the estimated model variance. The null hypothesis is rejected if $s_{\min} = \min(\boldsymbol{S}^t\hat{\boldsymbol{b}}) \le \min(\widehat{Q}_\alpha, 0)$ . Alternatives to estimating the model variance by (5) include the difference-based estimator initially proposed by Rice (1984) and Gasser, Sroka & Jennen-Steinmetz (1986).

The second approach uses a bootstrap method to approximate the distribution of $s_{\min}$ under the null hypothesis. Resamples of size $n$ can be drawn from the distribution of $\{x_i\}$, denoted as $\{x_1^*, x_2^*, \cdots, x_n^*\}$. The resample could be drawn from the empirical distribution of $\{x_i\}$, or its smoothed counterpart, or could simply be the original sample if the $\{x_i\}$'s are considered fixed design points. Then resamples of $y$ are simulated using

$$y_i^* = \sum_{j=1}^{m} \hat{b}_j^* B_{jp}(x_i^*) + \epsilon_i^*,$$

where $\epsilon_i^*$ is a resample of the residuals from the constrained regression spline. We then fit the unconstrained regression spline model (with same knot placement) to each resampled data set $\{(x_1^*, y_1^*), (x_2^*, y_2^*), \cdots, (x_n^*, y_n^*)\}$, and calculate the minimum slope estimate $s_{\min}^*$ based on each bootstrap resample. The $\alpha$-level quantile of the distribution of $s_{\min}^*$ can be used as the critical value or we can compute a p-value based on this distribution.

## 3. THEORETICAL PROPERTIES

The following assumptions are used in the proofs.

**Assumption 1**. *The true regression function $f(x)$ is $p+1$-times continuously differentiable on $x \in [0,1]$ for testing monotonicity,* i.e., *$f(x) \in C^{p+1}[0,1]$; and for convexity, $f(x) \in C^{p+2}[0,1]$*

**Assumption 2**. *The number of knots $k$ is $O(n^{1/(2p+3)})$, and there exists constant $M > 0$, such that*

$$\limsup_{n\to\infty} \frac{\max_{1\le j\le k} |t_{j+1} - t_j|}{\min_{1\le j\le k} |t_{j+1} - t_j|} < M.$$

**Assumption 3**. *The distribution function of the $x_i$ is absolutely continuous on $[0, 1]$, with probability density function $m_X(x)$ bounded away from 0.*

The distribution of estimated coefficients $\hat{\boldsymbol{b}}$ and estimated slopes $\boldsymbol{S}^t\hat{\boldsymbol{b}}$ conditioning on the design points is exactly normal under the assumption of normal errors $\epsilon_i$, and is asymptotically normal under certain regularity conditions (Lemma A.8 of Huang, Wu & Zhou 2004):

PROPOSITION 1. *Assuming that the errors $\epsilon_i$ are iid normal random variables with mean 0 and variance $\sigma^2$, the vector of unconstrained estimates of slopes $\boldsymbol{S}^t\hat{\boldsymbol{b}}$ follows a multivariate normal distribution conditioning on the knot placement and covariate values, i.e.,*

$$(\boldsymbol{S}^t\hat{\boldsymbol{b}} \mid k; t_1, t_2, \cdots, t_k; x_1, x_2, \cdots, x_n) \sim N(\boldsymbol{S}^t\boldsymbol{\beta}, \boldsymbol{S}^t(\boldsymbol{\Delta}^t\boldsymbol{\Delta})^{-1}\boldsymbol{S}\sigma^2).$$

*As a result of Huang et al (2004), under Assumptions 1-3, $\boldsymbol{S}^t\hat{\boldsymbol{b}}$ is asymptotically normally distributed with mean vector $\boldsymbol{S}^t\boldsymbol{\beta}$ and covariance matrix $\boldsymbol{S}^t(\boldsymbol{\Delta}^t\boldsymbol{\Delta})^{-1}\boldsymbol{S}\sigma^2$ conditioning on the knot placement, thus*

$$P(s_{\min} \leq r \mid k; t_1, t_2, \cdots, t_k) \asymp P_{\boldsymbol{\beta},\sigma^2}(r),$$

*where the symbol $\asymp$ means asymptotic equivalence, i.e., $a_n \asymp b_n$ means that $\lim_{n\to\infty} a_n/b_n = 1$.*

We show that for strictly increasing regression functions, the probability of rejecting the null hypothesis goes to zero as the sample size grows, under some mild assumptions. For regression functions that are decreasing over an interval, the power of the test goes to one. The performance of the proposed test when the infimum of the derivative is exactly zero will be examined in the simulation experiment. Results for the test of convexity are similar.

PROPOSITION 2. *Under Assumptions 1-3,*

1. *If $\inf_{x\in[0,1]} f'(x) \geq \epsilon_n > 0$ with $\epsilon_n = O(n^{-2/7+\nu})$ for some $\nu > 0$, then $P(s_{\min} \leq \min(\widehat{Q}_\alpha, 0)) = o(1)$, meaning that when the underlying function $f(x)$ is strictly increasing, the type I error rate of the proposed test converges to 0.*
2. *If $f'(x) = c < 0$, for some $x \in (0, 1)$, then $1 - P(s_{\min} \leq \min(\widehat{Q}_\alpha, 0)) = o(1)$, meaning that when the underlying function $f(x)$ decreases at at least one interior point (actually over an interval, by Assumption 1), the power of the proposed test converges to 1.*

Note that Proposition 2.1 implies that the test may be biased, although it is asymptotically unbiased. Specifically, for small samples with $f$ strictly increasing over most of the range but decreasing over a small interval, the power might be less than the target test size. However, the simulations show that the small sample performance of the test is good over a reasonable range of regression functions that are decreasing over only a portion of the range of $x$-values.

**Higher-order splines:** In cases where a quadratic spline fails to approximate the local curvature of the true regression function, we consider increasing the degree of the basis functions. In general, we consider a degree-p spline regression ($p \geq 3$ for the monotone case), where $f'(x) \geq 0$ over $[0, 1]$ is not necessarily equivalent to $\boldsymbol{S}^t\boldsymbol{\beta} \geq 0$, or in other words, simply checking the slopes at the knots is not enough for testing the null hypothesis. Instead, we consider

$$s_{\min} = \inf_x |\,\hat{g}'(x)|\,,$$

as our test statistic, where

$$\hat{g}'(x) = \sum_{j=1}^{m} \hat{b}_j B'_{jp}(x) = \sum_{j=1}^{m} \frac{p(\hat{b}_j - \hat{b}_{j-1})}{t_{j+1} - t_j} B_{j,p-1}(x)$$

denotes unconstrained estimate of the derivative $f'(x)$ using a degree-p spline basis and $\hat{b}_0 = 0$. The distribution of $s_{\min}$ can be derived given the distribution of $\hat{g}'(x)$. Under certain regularity conditions, the distribution of $\hat{g}'(x)$ converges to a nonstationary Gaussian process $G(x)$ defined on $[0, 1]$ with continuous path. The mean function of $G(x)$ is $\mu_G(x) = \mathrm{E}[G(x)] \asymp f'(x)$ (higher-order corrections to the mean function are also possible) and the covariance structure is

$$C_G(u, v) = \lim_{n\to\infty} \mathrm{Cov}(\hat{g}'(u), \hat{g}'(v)) = \lim_{n\to\infty} \sigma^2 [\boldsymbol{S}(u)]^t (\boldsymbol{\Delta}^t \boldsymbol{\Delta})^{-1} \boldsymbol{S}(v),$$

where $\boldsymbol{S}(u) = (B'_{1,p}(u), \cdots, B'_{m,p}(u))^t$ denotes the vector of derivatives of the basis functions evaluated at $u \in [0, 1]$.

The limiting Gaussian process is nonstationary, and an analytic expression of the distribution of the extrema of this random process is very difficult to obtain. We could use a numerical method to approximate the distribution of $s_{\min}$ under the null hypothesis. We propose to approximate the mean function $\mu_G(x)$ by the constrained estimate of $f'(x)$ using quadratic regression splines

$$\hat{\mu}^*_G(x) = \sum_{j=1}^{m} \hat{b}^*_j B'_{j2}(x_i) \tag{6}$$

and to estimate the covariance structure by $\widehat{C}(u, v) = \hat{\sigma}^2 [\boldsymbol{S}(u)]^t (\boldsymbol{\Delta}^t \boldsymbol{\Delta})^{-1} \boldsymbol{S}(v)$, where $\hat{\sigma}^2$ is defined by (5). More specifically, we can estimate the random process $G(x)$ defined on $x \in [0, 1]$ through simulations, i.e.,

$$\widehat{G}(x) = \hat{\mu}^*_G(x) + \hat{\sigma}[\boldsymbol{S}(x)]^t \boldsymbol{U}^{-1} \boldsymbol{Z},$$

where $\hat{\mu}^*_G(x)$ is defined by (6), $\boldsymbol{U}^t \boldsymbol{U}$ is the Cholesky decomposition of $\boldsymbol{\Delta}^t \boldsymbol{\Delta}$, and $\boldsymbol{Z}$ is an $m$-dimensional vector of standard normal random variables.

## 4. EXTENSION TO GENERALIZED REGRESSION MODELS

For the generalized regression model, the response $y_i$ has a distribution from an exponential family

$$m_Y(y; \theta, \phi) = \exp\{(y\theta - a(\theta))/\phi + c(y, \phi)\}.$$

The expected value and variance of $y$ are given by $\mu = E(y) = a'(\theta)$ and $\mathrm{Var}(y) = a''(\theta)\phi$. For example, the Bernoulli model uses $a(\theta) = \log(1 + e^\theta)$, $c(y, \phi) = 0$, and $\phi = 1$; and for the Poisson model, $a(\theta) = e^\theta$, $c(y, \phi) = -\log(y!)$ and $\phi = 1$. The conditional mean $\mu$ is related to the predictor variable $x$ through a link function $h(\mu) = \eta$, where $\eta$ is a smooth function of $x$. The link function for the Bernoulli model is $h(\mu) = \log(\mu/(1 - \mu))$, and for the Poisson model, $h(\mu) = \log(\mu)$. Note that if the link function is monotone, that $\eta$ is increasing in $x$ is a necessary and sufficient condition for $\mu$ to be increasing in $x$. The log-likelihood function

$$\ell(\boldsymbol{\theta}, \phi; \boldsymbol{y}) = \sum_{i=1}^{n} \left[ \frac{y_i\theta_i - a(\theta_i)}{\phi} + c(y_i, \phi) \right]$$

is to be maximized over a set of quadratic regression splines, and both constrained and unconstrained fits may be computed. The algorithm involves iteratively re-weighted projections, and follows the same ideas for the generalized linear model as found in McCullagh & Nelder (1989). Starting with $\boldsymbol{\eta}^0 \in \mathcal{C}$, the estimate $\boldsymbol{\eta}^{k+1}$ is obtained from $\boldsymbol{\eta}^k$ by constructing $\boldsymbol{z}$

$$z_i = \eta^k_i + (y_i - \mu^k_i) h'(\mu^k_i),$$

where $\mu^k_i = h^{-1}(\eta^k_i)$ and the derivative of the link function is evaluated at $\mu^k_i$. For the unconstrained splines, the weighted projection of $\boldsymbol{z}$ onto the space defined by $\boldsymbol{\Delta b}$ for $\boldsymbol{b} \in \Re^m$ is obtained with weight vector $\boldsymbol{w}$, where $1/w_i^k = h'(\mu_i^k)^2 V_k$, and $V_k$ is the variance function evaluated at $\mu_i^k$. For constrained splines, $\boldsymbol{z}$ is projected onto the cone containing $\boldsymbol{b}$ such that $\boldsymbol{S}^t\boldsymbol{b} \geq \boldsymbol{0}$, using the weights $\boldsymbol{w}$. This scheme can be shown to converge to the value of $\boldsymbol{\eta}$ that maximizes the likelihood.

As before, let $\hat{\boldsymbol{b}}$ be the MLE for the unconstrained version, and use $\hat{\boldsymbol{b}}^*$ for the constrained MLE. If $\boldsymbol{W}$ is the diagonal matrix containing the final $w_i$ values, then $\hat{\boldsymbol{b}}$ is approximately normal with mean $\boldsymbol{\beta}$ and covariance matrix

$$\mathrm{Cov}(\hat{\boldsymbol{b}}) \approx \phi(\boldsymbol{\Delta}^t\boldsymbol{W}\boldsymbol{\Delta})^{-1}.$$

The testing schemes in Section 2 may be implemented, using the covariance matrix for the unconstrained fit, and the mean function from the constrained fit in the estimation of the null distribution of $\hat{\boldsymbol{b}}$.

## 5. EXTENSION TO PARTIAL LINEAR MODELS

Accounting for covariates when investigating the relationship between a response variable and the predictor variable of main interest is important when there may be confounding variables or when the co-variates account for some of the variation in the response. Consider the following partial linear model,

$$y_i = f(x_i) + \boldsymbol{z}_i^t\boldsymbol{\gamma} + \sigma\epsilon_i, \quad i = 1, \ldots, n, \tag{7}$$

where $f(\cdot)$ is an unknown smooth function, $\boldsymbol{z}_i$ are known $q-$dimensional vectors, and $\boldsymbol{\gamma}$ is an unknown parameter vector. Various estimation approaches for the partial linear model can be found in Wahba (1984), Engle, Ganger, Rice & Weiss (1986), Speckman (1988) and Heckman (1986), among others. More recently, Xie & Huang (2009) considers SCAD-penalized estimation of model (7) as the dimension of covariates approaches infinity at a certain rate. In our analysis of the partial linear model, the number $q$ of covariates is fixed, and the interest is in whether the relationship between $E(y)$ and $x$ is monotone (or convex), after adjusting for the effect of the covariates $\boldsymbol{z}$.

The linear parameter vector $\boldsymbol{\gamma}$ can be estimated at the convergence rate for parametric estimation, as indicated by Speckman (1988), Chen (1988), Xie & Huang (2009), among others. The necessary conditions 1-4 from Chen (1988) can be translated into Assumption 3, and the following assumptions. First, the regression function $f(\cdot)$ satisfies

$$|f''(u) - f''(v)| \leq M_1|u - v|^{1/2},$$

for some constant $M_1 > 0$, and any $0 \leq u, v \leq 1$. Second, there exist $q \times q$ positive definite matrices $\Sigma_0$ and $\Sigma_1$ such that both $\mathrm{Cov}(\boldsymbol{z}_i \mid x_i = x) - \Sigma_0$ and $\Sigma_1 - \mathrm{Cov}(\boldsymbol{z}_i \mid x_i = x)$ are nonnegative definite for all $x \in [0, 1]$. Third, there exists a positive constant $M_2$ such that

$$|\mathrm{E}(\boldsymbol{z}_i \mid x_i = x)| \leq M_2$$

for all $x \in [0, 1]$. These assumptions guarantee that $\hat{\boldsymbol{\gamma}} - \boldsymbol{\gamma} = O_p(n^{-1/2})$, under Theorem 2 of Chen (1988), where $\hat{\boldsymbol{\gamma}}$ is defined as the minimizer of the following profile objective function (Xie & Huang 2009),

$$\hat{\boldsymbol{\gamma}} = \arg\min \left\|(\mathrm{I} - \boldsymbol{\Delta}(\boldsymbol{\Delta}^t\boldsymbol{\Delta})^{-1}\boldsymbol{\Delta}^t)(\boldsymbol{u} - \boldsymbol{Z}^t\boldsymbol{\gamma})\right\|^2,$$

where $\boldsymbol{u}$ is an $n \times 1$ vector of $u_i$, $\boldsymbol{Z} = (\boldsymbol{z}_1, \boldsymbol{z}_2, \cdots, \boldsymbol{z}_n)^t$, and I is the $m \times m$ identity matrix.

The values of $f$ at the $x_i$ will be estimated as $\boldsymbol{\Delta}\hat{\boldsymbol{b}}$, where the least-squares estimates $\hat{\boldsymbol{b}}$ and $\hat{\boldsymbol{\gamma}}$ are obtained simultaneously through a projection onto the linear space spanned by the columns of $\boldsymbol{\Delta}$ and the columns of $\boldsymbol{Z}$. The constrained estimators $\hat{\boldsymbol{b}}^*$ and $\hat{\boldsymbol{\gamma}}^*$ may be obtained through a single cone projection. Let $\boldsymbol{\alpha} = (\boldsymbol{\beta}^t, \boldsymbol{\gamma}^t)^t$ and let $\boldsymbol{X} = [\boldsymbol{\Delta}, \boldsymbol{Z}]$. The constraints on the parameter vector $\boldsymbol{\alpha}$ may be written as $\boldsymbol{S}^t\boldsymbol{\alpha} \geq \boldsymbol{0}$, where the last $q$ rows of $\boldsymbol{S}$ contain zeros, and the first $m$ rows are identical to the $\boldsymbol{S}$ matrix for the univariate case. Then the estimated slopes of $f$ at the knots are again $\boldsymbol{S}^t\hat{\boldsymbol{\alpha}}$, and the testing methods of section 2 can be directly applied.

## 6. EXAMPLES

We present three examples. The first uses weighted least-squares regression with a categorical covariate. The response variable for the second example is binary, and the third example uses Poisson regression. For each, the first approach to approximating the null distribution is used.

**Example 1: Uncounted votes in Georgia**

In the US presidential election of 2000, there was controversy about the high proportion of uncounted votes in Florida. The state of Georgia actually had a higher proportion of uncounted votes, but there was less national attention because the race was not close. The rate of uncounted votes were published on the secretary of state's web site, along with other information about the $n = 159$ Georgia counties. These data are currently posted on

`http://www.stat.colostate.edu/∼meyer/absvote.htm`.

Suppose interest is in determining if the rates of uncounted votes are, on average, larger for counties with higher proportions of registered voters who are African Americans. The percents of uncounted votes are shown in Figure 1(a), plotted against percent black voters, where each circle represents a county. The radius of the circle is scaled to reflect the number of ballots cast in the county.

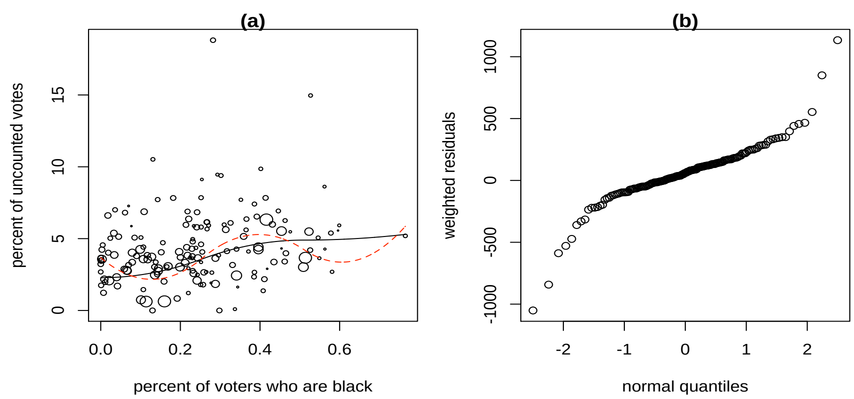


Figure 1: (a) Scatterplot of percent uncounted votes in the 2000 US presidential election, against percent of black voters for the 159 counties in the state of Georgia. The solid curve is the constrained fit, and the dashed curve is the unconstrained fit (both weighted by number of ballots). (b) Probability plot of the (weighted) residuals for the constrained fit.

The regression must be weighted by the number of ballots, as this is the denominator of the $y$ values. A weighted version of the constrained estimator is easily obtained with a transformation. Suppose $\text{cov}(\boldsymbol{\epsilon}) = \sigma^2\Sigma$; then the model $\boldsymbol{y} = \boldsymbol{\theta} + \boldsymbol{\epsilon}$ (where $\boldsymbol{\theta}$ is to be modeled as $\boldsymbol{\Delta b}$ with $\boldsymbol{S}^t\boldsymbol{b} \geq \boldsymbol{0}$) is transformed using the Cholesky decomposition $\boldsymbol{LL}^t$ of $\Sigma$. Multiplying the model through by $\boldsymbol{L}^{-1}$ gives $\boldsymbol{y}^* = \boldsymbol{\theta}^* + \boldsymbol{\epsilon}^*$, where $\boldsymbol{\theta}^*$ is to be modeled as $\boldsymbol{\Delta}^*\boldsymbol{b}$, $\boldsymbol{\Delta}^* = \boldsymbol{L}^{-1}\boldsymbol{\Delta}$, with the same constraint matrix.

The solid curve of Figure 1(a) represents the weighted constrained fit with four knots, and the dashed curve is the corresponding unconstrained fit. The hypothesis test is performed in the transformed model, to give a $p$-value of about 0.2, indicating that there is no evidence that the curve violates the monotonicity assumption. Now the constant versus increasing test of Meyer

(2008) can be performed (using the transformed data); the $p$-value is less than 0.001, showing that the percent of black voters is a strongly significant predictor of proportions of uncounted votes. Plot (b) of Figure 1 shows the sorted, weighted residuals against the normal quantiles. The curvature in the plot suggests that the errors are heavy-tailed, and perhaps the bootstrap approximation to the test is required. However, heavy-tailed residuals may also be explained by a missing source of variation in the model.

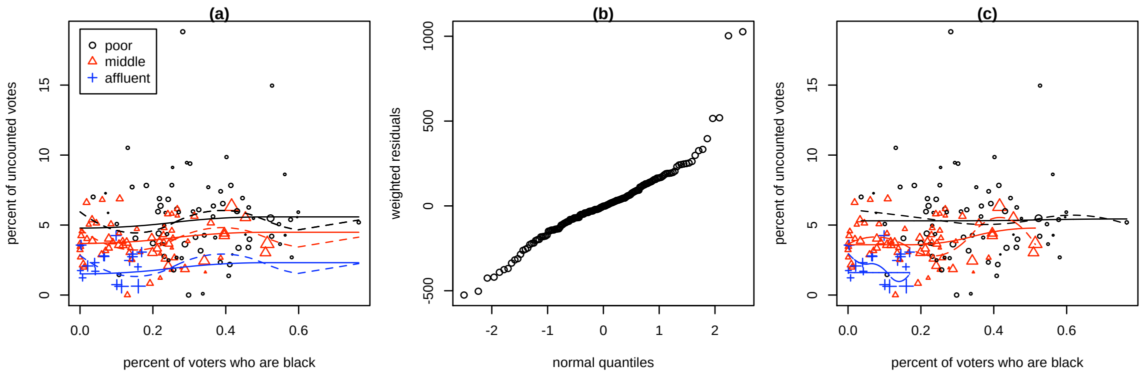


Figure 2: (a) Fits to the Georgia voting data set using the economic indicator co-variate. The solid curves are the constrained fit, and the dashed curves are the unconstrained fit. (b) Probability plot of the (weighted) residuals for the constrained fit. (c) Fits without the assumption of additivity.

For observational data such as these, possible confounding effects should be accounted for in the model. The economic status of the county is available as a categorical variable, and the rates of uncounted votes varies as to whether the counties are affluent, middle status, or poor. Figure 2 shows the same data with the plot character indicating the economic status. Again, the constrained and unconstrained (weighted) fits are the solid and dashed curves, respectively, with the curves corresponding to the most affluent counties on the bottom and those for the poorest counties on the top. For this model we again can not reject the null hypothesis that the true curves are monotone ($p \approx 0.3$), and for the subsequent test of constant versus increasing, we can not reject the constant function. We conclude that the economic status of the county is a strong predictor of uncounted votes, and the original significant increase in uncounted votes with percent black voters was due to the higher prevalence of black voters in the poorer counties. The probability plot for the residuals is now much closer to what one would expect for normal errors.

Alternatively, we can perform the test without assuming additive effects; that is, allowing for interaction between economic status and percent black voters. Three sets of knots are chosen, one for each economic level, spanning the appropriate ranges, and three sets of spline basis functions are created. Defining three dummy variables for the levels, the model

$$y_i = f_1(x_i)d_{1i} + f_2(x_i)d_{2i} + f_3(x_i)d_{3i} + \sigma\epsilon_i$$

may be fit with and without monotonicity assumptions for the $f_j$. If $\boldsymbol{\Delta}_1$, $\boldsymbol{\Delta}_2$, and $\boldsymbol{\Delta}_3$ are the matrices of basis vectors used for the three regression functions, define $\boldsymbol{\Delta} = (\boldsymbol{\Delta}_1, \boldsymbol{\Delta}_2, \boldsymbol{\Delta}_3)$, and define a matrix $\boldsymbol{S}$ correspondingly, with slopes of the three sets of basis functions at the three sets of knots. Then the test of $H_0 : \boldsymbol{S}^t\boldsymbol{b} \geq \mathbf{0}$ can be performed, using the same ideas as outlined for a single function. The non-parallel fits to these data are shown in Figure 2(c); the $p$-value for the test is about .17, so we again conclude that the monotonicity assumption is not violated.

**Example 2: Volunteerism and extroversion**

Cowles & Davis (1987) published the results of a study examining the personality characteristics of undergraduates who express willingness to volunteer for psychology experiments, compared with those who do not. If some personalities are over-represented in the subjects for these experiments, the results could be biased. The data were discussed in Fox (1997), and are depicted in Figure 3, where the proportions of students who volunteer are plotted against a measure of extroversion. For each integer score of 2-23, the proportions of males (triangles) and females (circles) who volunteer are indicated, with the size of the plot character proportional to the denominator. In plot (a) we also see the probability curves (that for females being the higher of the two), estimated using standard logistic regression. The extroversion score is a significant predictor with $p < .0001$, and gender is significant with $p = .024$. The linear log odds mandates that the estimated function is monotone, but perhaps there is no clinical reason for fitting this particular probability function.

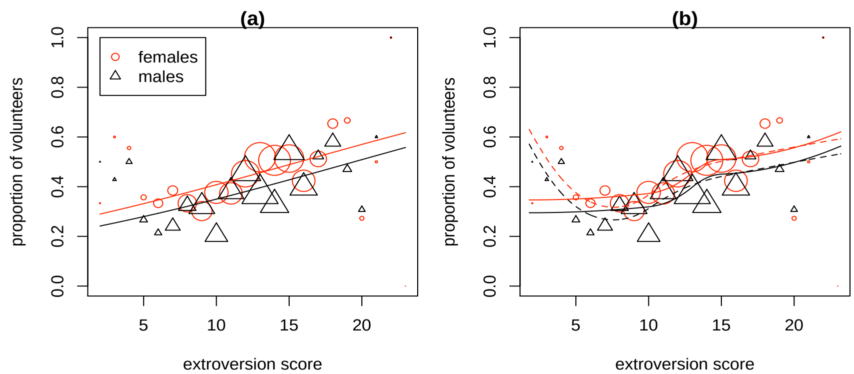


Figure 3: Estimates of the probability curve for volunteerism in psychological experiments as a function of extroversion and gender. (a) The standard logistic regression, and (b) the quadratic spline fits, where the solid curves are constrained to be monotone and the dashed are unconstrained.

Suppose instead that the researchers believe only that the relationship must be smooth, and the monotonicity is a research question. The constrained and unconstrained regression spline fits are shown in plot (b), where the gender effect is approximately the same as for the linear model. The unconstrained fit suggests that the subjects that scored very low on the extroversion scale are more likely to volunteer, but the size of the plot characters in this region indicate that this phenomenon is based on a smaller set of subjects. The $p$-value for the test of monotonicity is about .025, providing evidence that the true relationship is *not* monotone. The model with interaction between the predictors was also fit; it was quite similar so is not shown here.

**Example 3: Male toad size and number mates**

In this example we examine the relationship of the size of male toads and their number of mates, using data from a study by Arnold & Wade (1984) as discussed in Ramsey & Schafer (2002). Suppose a researcher conjectures that the number of mates is increasing with size, but would rather not impose a parametric form for the mean function. The data are shown in Figure 4, where there is some visual evidence of an increasing relationship. The dotted curve is the standard Poisson regression estimate, but it seems unlikely that the expected number mates should be increasing exponentially as assumed in this model. The dashed curve is the unconstrained spline estimate, which is decreasing in two places in the range of body sizes. The solid curve is the constrained estimate of the mean function. The $p$-value for the test of monotonicity is about .45, indicating there is no evidence against the researcher's conjecture of increasing relationship.

## 7. SIMULATIONS AND DISCUSSION

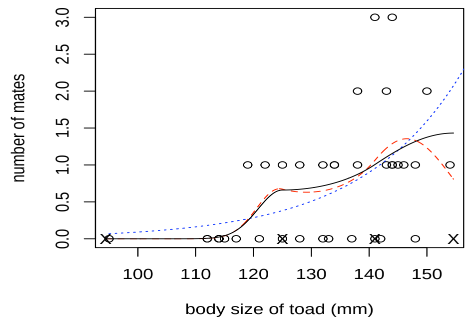


Figure 4: The number of mates of a male toad modeled as a function of body size. The solid curve is the monotone estimate of the mean function, the dashed curve is the unconstrained estimate, and the dotted curve is the estimate using a linear link function. The knots are marked with "X."

The performance of the proposed test is examined through extensive simulations. The proportion of times the null hypothesis of monotonicity is rejected using $\alpha = 0.05$ is computed under seven different regression functions with independent normal errors and $\sigma = 1$, each with five different sample sizes ($n = 25, 50, 100, 200$ and 400). We have implemented the proposed method with the first approach to estimating the distribution of $s_{\min}$. We have performed some preliminary simulations to compare variance estimation methods; these indicated that those of Rice (1984) and Gasser, Sroka, and Jennen-Steinmetz (1986) performed similarly to (5), hence we use the latter, simpler estimate. We used two pre-specified numbers of knots ($k = 4$ or $k = 6$), and in addition, we use a data-driven choice of $k$. In practice, one might use a method such as generalized cross validation (GCV) to choose $k$; see Ruppert *et al* (2003), chapter 5, for details. The range of $k$ values considered in the GCV-$k$ simulations is 3 to 8 for $n = 25$ and $n = 50$; 4-9 for $n = 100$, 5-10 for $n = 200$, and 6-11 for $n = 400$. When the test is implemented using the GCV-$k$, the test sizes are necessarily inflated. Additionally, we compare our results with two other methods: that of Bowman *et al* approach (referred to as "BJG") and of Ghosal *et al* approach (referred to as "GSV"). The BJG test is implemented via the "sm.monotonicity" function in the `R` "`sm`" package. The GSV test is coded in `R` with fixed bandwith of $h = 0.5n^{-1/5}$.

The observed predictor variable values and the knots are equally spaced on $[0, 1]$, and among the true regression functions, the first three are monotone on $[0, 1]$ whereas the rest have local dips. The regression function $f_a$ is taken from Bowman *et al* (1998), where

$$f_a(x) = 1 + x - a \exp[-50(x - 1/2)^2].$$

The "null" version has $a = 0.15$ and is strictly monotone over $[0, 1]$. The "mild" version uses $a = 0.25$ and has a small dip, while the "strong" version with $a = 0.45$ has a pronounced dip. The function $\mu_4$ is taken from Ghosal *e*t al (2000), defined as

$$\mu_4 = \begin{cases} 10(x-0.5)^3 - \exp\left(-100(x-0.25)^2\right), & \text{if } x < 0.5, \\ 0.1(x-0.5) - \exp\left(-100(x-0.25)^2\right), & \text{otherwise.} \end{cases}$$

Because the Bowman *et al* and Ghosal *e*t al papers used $\sigma = .1$, we use $10f_a$ and $10\mu_4$. We simulated $10,000$ datasets under each scenario; the results are summarized in Table 1.

For the constant function, the proportions of rejections decrease towards the target test size using our method under either fixed or adaptively chosen number of knots; however, the actual test size is unacceptably large for the GCV-$k$ when the sample size is $n = 25$. The BJG approach could not finish and is well known to break down for constant functions. For the GSV approach, the test size tends to become more conservative for larger $n$; an appropriately chosen bandwidth might alleviate this problem.

The function $f(x) = 15(x-1/4)_+^2$ is flat over $[0, 1/4]$, and the approximation $\boldsymbol{\Delta\beta}$ to the true function values $\boldsymbol{\theta}$ is actually decreasing on $[0, 1/8]$ for $k = 4$ knots, as shown in plot (a) of Figure 5. The thick grey curve is the true regression function, while the thin black curve is the piecewise quadratic spline approximation with $k = 4$. This decrease leads to a test size that is too large, and indeed increases with $n$ if $k$ stays fixed at four. However, when the number of knots $k$ increases to 6, these proportions are close to nominal test size. The GCV method tends to choose smaller $k$, hence the test size is unacceptably large. The approximation $\boldsymbol{\Delta\beta}$ for $k = 6$ is shown in plot (b) of Figure 5, and an example data set with $n = 100$ and $k = 6$ is known in plot (c). The constrained fit is the solid curve, the unconstrained is the dashed curve, and the true function is shown as the dotted curve. The null hypothesis is not rejected. The GSV method is again more conservative as the test size increases. The BJG approach can be implemented but rejects too often even for the largest sample size. The function $f(x) = 15(x-1/4)^2$ is similar to $f(x) = 15(x-1/4)_+^2$ but decreasing over $[0, 1/4]$. In plot (d) of Figure 5, an example data set using $f(x) = 15(x-1/4)^2$ is shown; here the null hypothesis is (correctly) rejected, with p-value of 0.009. The largest rejection rates are for methods with inflated test sizes for $f(x) = 15(x-1/4)_+^2$, and for each method, the rejection rate grows with $n$ as expected.

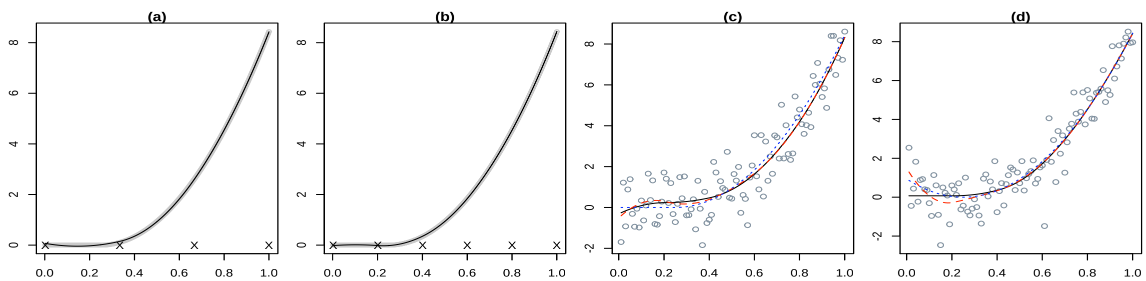


Figure 5: For the function $f(x) = 15(x-1/4)_+^2$, the approximation $\boldsymbol{\Delta\beta}$ to $\boldsymbol{\theta}$ for $k = 4$ and $k = 6$ are shown in plots (a) and (b); the knots are marked with "X." Plots (c) and (d) show example data sets with $n = 100$. The constrained fits as the solid curves, the dashed fits as the unconstrained curves, and the true functions are shown as the dotted curves. The null hypothesis of monotonicity is accepted for the data in (c), and rejected in (d).

For the "null" version of $f_a$, our test sizes are smaller than the target for fixed $k$, reflecting Proposition 2(2). The test size is too large for $n = 25$ when the number of knots are adaptively chosen, but otherwise less than the target. The BJG and GSV methods also have test sizes that seem to go to zero as $n$ increases. For the "mild" and "strong" versions of $f_a$, our methods perform roughly similarly to the BJG method. The GSV method does not seem to find the "mild" dip for $f_{.25}$, even for the largest sample size, but performs well for $f_{.45}$.

For the seventh function $\mu_4$, the proposed approach successfully rejects the null hypothesis with high probability for all three variations except when $n = 25$ and 50 where the number of knots is fixed at 6. However, the fits to the data with small $k$ are "incorrect" as shown in Figure 6. For $k = 4$ knots, the closest function in the estimation space, i.e., $\boldsymbol{\Delta\beta}$, is shown as the dashed curve. The fit to a data set with $n = 200$ is shown as the solid curve, and the true function is the dotted curve. Neither $k = 4$ nor $k = 6$ provides enough flexibility for fit the steep dip, although for $k = 4$, the decrease at the left end of the approximated curve is steep enough so that the null hypothesis is rejected for all 10,000 data sets of this size. For $k = 6$, the approximated curve is actually increasing over the steep dip, with a decline in the flat part of the true curve. Eight or ten knots will provide the appropriate flexibility; for these choices, the rejection rate is high even

Table 1: Simulations to compute test size and power for the test of monotone versus non-monotone regression function. Entries are the proportions of data sets for which the null hypothesis is rejected, for seven choices of underlying regression functions and five sample sizes. The model variance is unity and the target test size is $\alpha = 0.05$.

| | | percent rejection | | | | |
|---|---|---|---|---|---|---|
| $\mu$ | $n$ | $k=4$ | $k=6$ | GCV-$k$ | BJG | GSV |
| const | 25 | .072 | .078 | .162 | * | .104 |
| | 50 | .062 | .063 | .065 | * | .057 |
| | 100 | .060 | .062 | .063 | * | .039 |
| | 200 | .061 | .060 | .056 | * | .034 |
| | 400 | .057 | .057 | .057 | * | .031 |
| $15(x-1/4)_+^2$ | 25 | .087 | .062 | .221 | .316 | .026 |
| | 50 | .092 | .051 | .236 | .231 | .013 |
| | 100 | .103 | .043 | .125 | .297 | .007 |
| | 200 | .132 | .047 | .082 | .379 | .004 |
| | 400 | .170 | .045 | .073 | .153 | .006 |
| $10f_a$ (null) | 25 | .018 | .027 | .090 | .036 | 1e-4 |
| | 50 | .011 | .018 | .044 | .018 | 0 |
| | 100 | .004 | .015 | .038 | .016 | 0 |
| | 200 | 7e-4 | .018 | .038 | .024 | 0 |
| | 400 | 1e-4 | .032 | .047 | .003 | 0 |
| $10f_a$ (mild) | 25 | .068 | .047 | .158 | .067 | 6e-4 |
| | 50 | .096 | .081 | .151 | .080 | 2e-4 |
| | 100 | .165 | .229 | .252 | .166 | .001 |
| | 200 | .332 | .574 | .474 | .397 | .005 |
| | 400 | .635 | .905 | .723 | .748 | .032 |
| $10f_a$ (strong) | 25 | .336 | .196 | .404 | .171 | .090 |
| | 50 | .749 | .538 | .591 | .298 | .858 |
| | 100 | .991 | .933 | .838 | .698 | .856 |
| | 200 | 1 | 1 | .931 | .981 | .998 |
| | 400 | 1 | 1 | .949 | 1 | 1 |
| $15(x-1/4)^2$ | 25 | .198 | .111 | .381 | .584 | .071 |
| | 50 | .313 | .147 | .543 | .744 | .071 |
| | 100 | .502 | .200 | .453 | .900 | .115 |
| | 200 | .751 | .308 | .427 | .979 | .207 |
| | 400 | .952 | .469 | .435 | .998 | .422 |
| $10\mu_4$ | 25 | .600 | .052 | .902 | .297 | .027 |
| | 50 | .952 | .015 | .908 | .317 | .011 |
| | 100 | 1 | .103 | .970 | .288 | .019 |
| | 200 | 1 | .851 | 1 | .239 | .937 |
| | 400 | 1 | 1 | 1 | .166 | 1 |

at $n = 25$. The GCV-$k$ method correctly chooses a large number of knots for a high proportion of the data sets. The BJG test has small power, and the power seems to decrease as $n$ increases. The GSV method performs comparably to our proposed method only under the larger sample sizes.

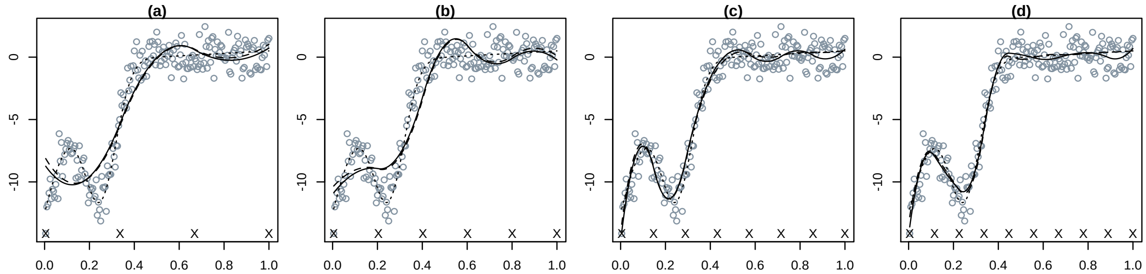


Figure 6: Fits to a data set generated using $10\mu_4$ and $n = 200$, for (a) $k = 4$, (b) $k = 6$, (c) $k = 8$, and (d) $k = 10$. The fit is the solid curve, $10\mu_4$ is the dotted curve, and the dashed curve represents $\boldsymbol{\Delta\beta}$, the closest function in the estimation space to the true function.

This table of simulation results illustrates the double difficulties of a hypothesis testing with a smoothing method, and with a null hypothesis that is not represented by a point but instead covers an entire orthant of $\mathbb{R}^m$. When the true function lies away from the boundary, the test sizes may be too small, opening the door for biasedness. We have illustrated these difficulties with the problematic regression functions $(x - 1/4)_+^2$ and $\mu_4$. The performance of the test also depends on the user-defined parameters. For the quadratic regression function, the proposed method performs best with the smallest number of knots, because the approximated function is equal to the true function. However, when the true function is very "wiggly" a larger number of knots must be chosen to get a good approximation. If the data are allowed to choose, the test size becomes inflated. Ideally, the user has an *a priori* idea of how variable the regression function might be, and chooses $k$ accordingly. If the GCV-$k$ method is implemented, the user be aware of the inflated test size, and we do not recommend the GCV-$k$ for the smaller sample sizes. We have shown that our methods have good large-sample properties, and the behavior in small samples is comparatively good overall.

Further, the largest sample size $n = 400$ with $k = 6$ took less than 1.5 minutes of real time for all 10,000 simulations, using a Mac Powerbook processor, whereas its two competitors (BJG and GSV) are much slower (both take more than a day to complete). The shorter computation time and the simplicity of the test add to the advantage of our proposed testing procedures.

## APPENDIX

In the appendix, we present details for proving Proposition 2. The technical details mainly address the test of monotonicity ($p = 2$). The underlying theory for testing convexity ($p = 3$) is similar and will not be duplicated.

In the following, the $B_j$ basis functions depend on the set of knots, which gets larger with $n$. Similarly, $m = m_n$, $\boldsymbol{b} = \boldsymbol{b}_n$, and $\boldsymbol{t} = \boldsymbol{t}_n$, but for simplicity of presentation the subscripts are not used. In the proofs to come, we will use an important result from Lemma 5.4 of Zhou & Wolfe (2000), which says that there exist positive constant $K_L$ and $K_U$ that,

$$K_L \leq n^{4/7} a_{ii} \leq K_U \tag{8}$$

where $a_{ii}$ is the $(i, i)$-th element of $\boldsymbol{S}^t(\boldsymbol{\Delta}^t\boldsymbol{\Delta})^{-1}\boldsymbol{S}$.

LEMMA 1. *Under Assumptions 1-3, as the sample size $n \to \infty$,* *1.*

$$|\hat{\boldsymbol{b}} - \boldsymbol{b}|_\infty = O_p(n^{-3/7+\delta}),$$

*where* $\boldsymbol{b}$ *is the best approximation to the underlying regression function in the least-squares sense, i.e.,*

$$\boldsymbol{b} = \arg\min_{b \in \Re^m} \int \left( f(x) - \sum_{j=1}^{m} b_j B_{j2}(x) \right)^2 dx.$$

*2. For any* $\delta > 0$*,*

$$n^{2/7-\delta} \sup_x |\hat{g}'(x) - f'(x)| = O_p(1),$$

*where* $\hat{g}(x) = \sum_{j=1}^{m} \hat{b}_j B_{j2}(x)$ *and* $\hat{g}'(x) = \sum_{j=1}^{m} \hat{b}_j B'_{j2}(x)$.

*Proof:* Huang *et al* (2004) has shown that $|\mathrm{E}\hat{\boldsymbol{b}} - \boldsymbol{b}|_\infty = O_p(n^{-3/7})$, under the stated assumptions in Lemma A.11. Now let us demonstrate that $|\hat{\boldsymbol{b}} - \mathrm{E}\hat{\boldsymbol{b}}|_\infty = O_p(n^{-3/7+\delta})$, for some $\delta > 0$. Starting with

$$P\left(n^{3/7-\delta} \max_j \left|\hat{b}_j - \mathrm{E}\hat{b}_j\right| > \epsilon\right) \le \sum_{j=1}^{5} P\left(n^{3/7-\delta} \max_{l=0,\ldots,\lfloor m/5-1 \rfloor} \left|\hat{b}_{5l+j} - \mathrm{E}\hat{b}_{5l+j}\right| > \epsilon\right),$$

it then suffices to show that $P\left(n^{3/7-\delta} \max_{l=1,\ldots,\lfloor m/5 \rfloor} \left|\hat{b}_{5l} - \mathrm{E}\hat{b}_{5l}\right| > \epsilon\right) \to 0$. Here we exploit the fact that the quadratic $B$-spline basis functions are orthogonal if the subscripts are more than four removed, so that $\hat{b}_{5l}$ and $\hat{b}_{5(l+1)}$ are independent random variables. Let $\sigma^2_{5l}$ denote $Var(\hat{b}_{5l})$. Since the number of knots $k$ is $O(n^{1/7})$, the number of observations that contribute to the estimation of each coefficient $b_i$ is in the order of $n^{6/7}$. Thus $\sigma^2_{5l} = O(n^{-6/7})$ (see also Lemma A.9 of Huang *et al* 2004), and hence

$$\begin{aligned}
P(n^{3/7-\delta} \max_l \left|\hat{b}_{5l} - \mathrm{E}\hat{b}_{5l}\right| > \epsilon) &= 1 - P(n^{3/7-\delta} \max_l \left|\hat{b}_{5l} - \mathrm{E}\hat{b}_{5l}\right| \le \epsilon) \\
&= 1 - \prod_{l=1}^{\lfloor m/5 \rfloor} P(n^{3/7-\delta} \left|\hat{b}_{5l} - \mathrm{E}\hat{b}_{5l}\right| \le \epsilon) \\
&\asymp 1 - \prod_{l=1}^{\lfloor m/5 \rfloor} \left[2\Phi\left(\frac{\epsilon\, n^{\delta-3/7}}{\sigma_{5l}}\right) - 1\right] \\
&\asymp 1 - \left[2\Phi(c_1 \epsilon\, n^{\delta}) - 1\right]^{n^{1/7}} \\
&= 1 - \left[1 - 2\int_{c_1 \epsilon\, n^\delta}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-t^2/2} dt\right]^{n^{1/7}} \\
&\asymp 1 - \exp\left(-2n^{1/7} \int_{c_1 \epsilon\, n^\delta}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-t^2/2} dt\right),
\end{aligned}$$

where $c_1$ is a constant, and the last asymptotic equivalence follows from $\lim_{n\to\infty} \left(1 + \frac{x}{n}\right)^n = e^x$. The last expression converges to 0 as

$$n^{1/7} \int_{c_1 \epsilon\, n^\delta}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-t^2/2} dt \le n^{1/7} \sqrt{\frac{2}{\pi}} \frac{e^{-\epsilon^2 n^{2\delta}}}{c_1 \epsilon n^\delta} \to 0. \qquad (9)$$

Thus we have shown that $|\hat{\boldsymbol{b}} - \mathrm{E}\hat{\boldsymbol{b}}|_\infty = O_p(n^{-3/7+\delta})$, and the result easily follows from triangle inequality.

For part 2, the theory of spline approximation in Zhou & Wolfe (2000) gives that

$$n^{2/7} \sup_x |g'(x) - f'(x)| = O(1),$$

where $g'(x) = \sum_{j=1}^{m} \boldsymbol{\beta}_j B'_{j2}(x)$. Then we need to show that $n^{2/7-\delta} \sup_x |\hat{g}'(x) - g'(x)| = O_p(1)$ to complete the proof. It follows from de Boor (2001) that

$$\hat{g}'(x) - g'(x) = \sum_{j=2}^{m} \frac{\hat{b}_j - b_j - (\hat{b}_{j-1} - b_{j-1})}{t_{j+1} - t_j} 2B_{j1}(x), \tag{10}$$

in which only two terms in the summation are nonzero for degree-1 B-splines. Assumption 2 guarantees that there exists $k_1 > 0$ such that $\sup_{j=2,\cdots,k} \left| \frac{1}{t_{j+1}-t_j} \right| < k_1 n^{1/7}$, so by (10),

$$\sup_x |\hat{g}'(x) - g'(x)| \le 8k_2 n^{1/7} \sup_{x,j} |B_{j1}(x)| \cdot |\hat{\boldsymbol{b}} - \boldsymbol{b}|_\infty = O_p(n^{-2/7+\delta}),$$

and the proof is complete.

*Proof of Proposition 2.* We will prove the two statements separately. In both proofs, we use the inequality that, $|\min_{x\in T} \hat{g}'(x) - \min_{x\in T} f'(x)| \le \sup_{x\in T} |\hat{g}'(x) - f'(x)|$. To rationalize, we take $\hat{g}'(x_1) = \min_{x\in T} \hat{g}'(x)$ and $f'(x_2) = \min_{x\in T} f'(x)$, then

$$\hat{g}'(x_1) - f'(x_1) \le \min_{x\in T} \hat{g}'(x) - \min_{x\in T} f'(x) \le \hat{g}'(x_2) - f'(x_2),$$

where both limits have absolute value no greater than $\sup_{x\in T} |\hat{g}'(x) - f'(x)|$.

1. Let $T = \{t_0, t_1, \cdots, t_{k+1}\}$, then

$$\begin{aligned}
P(s_{\min} \le \min(\widehat{Q}_\alpha, 0)) &\le P(s_{\min} \le 0) \\
&= P\left(\min_{x\in T} \hat{g}'(x) - \min_{x\in T} f'(x) \le -\min_{x\in T} f'(x)\right) \\
&\le P(|\min_{x\in T} \hat{g}'(x) - \min_{x\in T} f'(x)| \ge \epsilon_n) \\
&\le P(\sup_{x\in T} |\hat{g}'(x) - f'(x)| \ge \epsilon_n) \\
&\le P(n^{2/7-\delta} \sup_{x\in T} |\hat{g}'(x) - f'(x)| \ge n^{2/7-\delta}\epsilon_n) \\
&= o(1),
\end{aligned}$$

for $\delta < \nu$ with $\nu$ defined in Proposition 2(1). The last equation follows from Lemma 1 (2).

2. When $\inf_x f'(x) < 0$, we can argue that $\widehat{Q}_\alpha < 0$ with probability tending to one, thus

$$
\begin{aligned}
& P\left(s_{\min} \le \min(\widehat{Q}_\alpha, 0)\right) \\
= \quad & P\left(\min_{x\in T} \hat{g}'(x) - \min_{x\in T} f'(x) \le \min(\widehat{Q}_\alpha, 0) - \min_{x\in T} f'(x))\right) \\
\ge \quad & P\left(|\min_{x\in T} \hat{g}'(x) - \min_{x\in T} f'(x)| \le \min(\widehat{Q}_\alpha, 0) - \min_{x\in T} f'(x))\right) \\
\ge \quad & P\left(\sup_{x\in T} |\hat{g}'(x) - f'(x)| \le |c| + \min(\widehat{Q}_\alpha, 0)\right) \\
\asymp \quad & P\left(\sup_{x\in T} |\hat{g}'(x) - f'(x)| \le |c| + \inf\left\{ r \,\middle|\, 1 - \int \cdots \int_{\{\boldsymbol{z}|\boldsymbol{z}-r\mathbf{1}\ge\mathbf{0}\}} \phi(\boldsymbol{z}; \boldsymbol{S}^t\hat{\boldsymbol{b}}^*, \boldsymbol{S}^t(\boldsymbol{\Delta}^t\boldsymbol{\Delta})^{-1}\boldsymbol{S}\sigma^2)d\boldsymbol{z} \ge \alpha \right\}\right) \\
\ge \quad & P\left(\sup_{x\in T} |\hat{g}'(x) - f'(x)| \le |c| + \inf\left\{ r \mid P_{\mathbf{0},\sigma^2}(r) \ge \alpha \right\}\right) \\
\ge \quad & P\left(\sup_{x\in T} |\hat{g}'(x) - f'(x)| \le |c| + \inf\{ r \mid \sum_{i=1}^k \Phi(r; 0, a_{ii}\sigma^2) \ge \alpha \}\right),
\end{aligned}
$$

where $a_{ii}$ denotes the $(i,i)^{\text{th}}$ entry of $\boldsymbol{S}^t(\boldsymbol{\Delta}^t\boldsymbol{\Delta})^{-1}\boldsymbol{S}$ . It suffices to show that $\inf\left\{ r \mid \sum_{i=1}^k \Phi(r; 0, a_{ii}\sigma^2) \ge \alpha \right\} \xrightarrow{p} 0$, as $n^{2/7-\delta}\sup_{x\in T}|\hat{g}'(x) - f'(x)| = O_p(1)$. Here,

$$
\begin{aligned}
0 \ge \inf\left\{ r \,\middle|\, \sum_{i=1}^k \Phi(r; 0, a_{ii}\sigma^2) \ge \alpha \right\} \quad & \ge \quad \inf\left\{ r \,\middle|\, C_3\Phi(r; 0, \sup_i a_{ii}\sigma^2) \ge \frac{\alpha}{k} \right\} \\
& = \quad \sup_i a_{ii}\sigma^2\Phi^{-1}\left(\frac{\alpha}{C_3 k}\right) \xrightarrow{p} 0
\end{aligned}
$$

as $\sup_i |a_{ii}| = O(n^{-4/7})$, and $\Phi^{-1}\left(\frac{\alpha}{C_3 k}\right)$ grows at log rate, where $C_3$ is a positive constant. The second inequality holds since we can find a positive constant $C_3$ such that $\sum_{i=1}^k \Phi(r; 0, a_{ii}\sigma^2) \le C_3 k\Phi(r; 0, \sup_i a_{ii}\sigma^2)$, and both terms are monotone functions of $r$.

## ACKNOWLEDGEMENTS

The work of Mary Meyer was partially supported by the National Science Foundation DMS grant. The main work of the paper was conducted when Jianqiang Wang was visiting Colorado State University, and current work of Jianqiang Wang is supported by the collaborative program between the National Institute of Statistical Sciences (NISS) and the National Agricultural Statistics Service (NASS). The authors would like to thank Subhashis Ghosal for sharing his code for the GSV test. The detailed and insightful comments from the Associate Editor and three referees are gratefully acknowledged.

## REFERENCES

S. Arnold, & M. Wade (1984). On the measurement of natural and sexual selection. *Evolution 38*(4), 720–734.

A. W. Bowman, M. C. Jones, & I. Gijbels (1998). Testing monotonicity of regression. *Journal of Computational and Graphical Statistics 7*, 489–500.

P. Chaudhuri, & J. S. Marron (1999). Sizer for exploration of structures in curves. *Journal of the American Statistical Association 94*, 807–823.

P. Chen (1988). An integrated formulation for selecting the most probable multinomial cell. *Annals of the Institute of Statistical Mathematics 40*, 615–625.

M. Cowles, & C. Davis (1987). The subject matter of psychology: Volunteers. *British Journal of Social Psychology 26*, 97–102.

C. de Boor (2001). *A Practical Guide to Splines.* Springer-Verlag.

R. F.Engle, C. W. J. Granger, J. Rice, & A. Weiss (1986). Semiparametric estimates of the relation between weather and electricity sales. *Journal of the American Statistical Association 81*, 310–320.

J. Fox (1997). *Applied Regression Analysis, Linear Models, and Related Methods.* Sage Publications Inc.

D. A. S.Fraser, & H. Massam (1989). A mixed primal-dual bases algorithm for regression under inequality constraints. application to convex regression. *Scandinavian Journal of Statistics 26*, 65–74.

T. Gasser, L. Sroka, & C. Jennen-Steinmetz (1986). Residual variance and residual pattern in nonlinear regression. *Biometrika 73*, 625–633.

S. Ghosal, A. Sen, & A. W. van der Vaart (2000). Testing monotonicity of regression. *The Annals of Statistics 28*(4), 1054–1082.

P. Hall, & N. E. Heckman (2000). Testing for monotonicity of a regression mean by calibrating for linear functions. *The Annals of Statistics 28*(1), 20–39.

J. Hannig, & J. S. Marron (2006). Advanced distribution theory for sizer. *Journal of the American Statistical Association 101*, 484–499.

T. Hastie, & R. Tibshirani (1999). *Generalized Additive Models.* Chapman & Hall Ltd.

N. E. Heckman (1986). Spline smoothing in a partly linear model. *Journal of the Royal Statistical Society, Series B: Methodological 48*, 244–248.

J. Z. Huang, C. O. Wu, & L. Zhou (2004). Polynomial spline estimation and inference for varying coefficient models with longitudinal data. *Statistica Sinica 14*, 763–788.

A. Juditsky, & A. Nemirovski (2002). On nonparametric tests of positivity/monotonicity/convexity. *The Annals of Statistics 30*(2), 498–527.

M. R. Leadbetter, H. Rootzen & G. Lindgren (1983). *Extremes and Related Properties of Random Sequences and Processes.* Springer-Verlag Inc.

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

E. Mammen, & C. Thomas-Agnan (1999). Smoothing splines and shape restrictions. *Scandinavian Journal of Statistics 26*, 239–252.

M. C. Meyer(2008). Inference using shape-restricted regression splines. *The Annals of Applied Statistics 2*(3), 1013–1033.

J. O. Ramsay(1988). Monotone regression splines in action (C/R: P442-461). *Statistical Science 3*, 425–441.

F. L. Ramsey, & D. W. Schafer (2002). *The Statistical Sleuth: a Course in Methods of Data Analysis.* Duxbury Press.

J. Rice(1984). Bandwidth choice for nonparametric regression. *The Annals of Statistics 12*, 1215–1230.

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

D. Ruppert, M. P. Wand, and R. J. Carroll (2003). *Semiparametric Regression.* Cambridge Series in Statistical and Probabilistic Mathematics.

P. Speckman (1988). Kernel smoothing in partial linear models. *Journal of the Royal Statistical Society, Series B: Methodological 50*, 413–436.

G. Wahba (1984). Partial spline models for the semiparametric estimation of functions of several variables. In *Statistical Analyses for Time Series, Jan-US Joint Seminar*, pp. 319–329. Institute of Statistical Mathematics, Tokyo.

H. Xie, & J. Huang (2009). SCAD-penalized regression in high-dimensional partially linear models. *The Annals of Statistics 37*, 673–696.

S. Zhou, X. Shen, & D. A. Wolfe (1998). Local asymptotics for regression splines and confidence regions. *The Annals of Statistics 26*(5), 1760–1782.

S. Zhou, & D. A. Wolfe (2000). On derivative estimation in spline regression. *Statistica Sinica 10*(1), 93–108.

---

*Received ???*
*Accepted ???*

Jianqiang WANG: qqwjq9916@gmail.com
*Hewlett-Packard Labs*
*Palo Alto, CA*
*U.S.A.,94304*

Mary MEYER: meyer@stat.colostate.edu
*Department of Statistics, Colorado State University*
*Fort Collins, CO*
*U.S.A.,80523*