Statistica Sinica **28** (2018), 277-292
doi:https://doi.org/10.5705/ss.202016.0342

CONSTRAINED PARTIAL LINEAR REGRESSION SPLINES

Mary C. Meyer

Colorado State University

Abstract: The constrained partial linear model is fit using a single cone projection, without back-fitting. The cone formulation not only provides efficient computation, but also allows for derivation of convergence rates and inference methods. Conditions for simultaneous root-n convergence of the parameters and optimal convergence for the regression function are given. Hypothesis tests involving the nonlinear regression function, while controlling for the effects of the linear term, use a test statistic whose null distribution is that of a mixture-of-betas random variables, under the normal errors assumption. Inference involving the linear term uses approximate t and F distributions; simulations show these perform well compared to competitors.

Key words and phrases: Constrained estimation, convergence rates, hypothesis testing, isotonic, smoothing.

1. Introduction

Consider the regression model

$$Y = f(T) + \boldsymbol{X}^\top \boldsymbol{\beta} + \varepsilon, \tag{1.1}$$

where the random variable T takes values in $[0,1]$, $\boldsymbol{X} = (X_1, \ldots, X_p)^\top$ are parametrically modeled covariates, $\boldsymbol{\beta} \in \mathbb{R}^p$ is an unknown parameter vector, and ε is a mean-zero random error independent of $(\boldsymbol{X}, T)$. We observe independent realizations $(y_i, t_i, \boldsymbol{x}_i)$, $i = 1, \ldots, n$. The function f is to be estimated with shape-constrained regression splines, and the principle of least-squares is used to estimate f and $\boldsymbol{\beta}$ simultaneously with a cone projection. Interest is in convergence rates for estimates of $\boldsymbol{\beta}$ and f, and in inference for each component while controlling for the effect of the other.

For unconstrained spline estimation without the linear term, optimal rates of convergence were given by Stone (1980, 1982). If q is the order of the spline and $K = n^{1/(2q+1)}$ is the number of knots then, under mild conditions, the convergence rate $n^{-q/(2q+1)}$ is attained for the estimate $\tilde{f}(x_0)$, $x_0 \in [0,1]$. The

global rate of convergence is $\sum_{i=1}^{n}[\tilde{f}(x_i) - f(x_i)]^2/n = O_p(n^{-2q/(2q+1)})$. For the partial linear model, optimal rates for both parametric and non-parametric components can be obtained if the predictors are independent; see for example Heckman (1986). For correlated $\boldsymbol{X}$ and T, and smoothing spline estimation of f, Rice (1986) showed that the estimator for f must be under-smoothed so that the bias of $\hat{\boldsymbol{\beta}}$ gets smaller at a faster rate than its standard error. Similarly, Speckman (1988) showed that the kernel regression estimator must be under-smoothed to attain the root-n convergence of the parametric term, and similar results for the local polynomial partial linear model were attained by Opsomer and Ruppert (1999). For piece-wise polynomial estimation of f, Chen (1988) gave mild conditions under which the parametric and non-parametric terms might simultaneously have optimal convergence rates.

Huang (2002) considered the partial linear isotonic regression model without smoothing, and showed that root-n convergence of the linear term is attained if f is strictly increasing. The estimator is attained through "back-fitting," iterating between estimating f and $\boldsymbol{\beta}$ until a convergence criterion is attained. Smooth monotone partially linear estimation was considered by Hazelton and Turlach (2011), with a Bayesian approach. The partial linear model with constrained penalized splines was considered by Meyer (2012) and Pya and Wood (2015); the latter provided fitting and inference routines in the `R` package `scam`.

In the next section, we show that f and $\boldsymbol{\beta}$ can be estimated simultaneously with a single cone projection without back-fitting. The fitting routine is considerably faster, but more importantly, the cone formulation allows for the derivation of conditions for which the estimate $\hat{\boldsymbol{\beta}}$ has root-n convergence. We show in Section 3 that if f is strictly increasing for the monotonicity shape, or strictly convex for the convex shape, then root-n convergence for $\hat{\boldsymbol{\beta}}$ is attained with the number of knots necessary for the optimal convergence rate for $\hat{f}$, under mild assumptions. In Section 4, we formulate a test for f; for example constant versus increasing, or linear versus convex, while controlling for the effects of the linearly-modeled covariates. This test is exact under the normal errors assumption, and is demonstrated in an analysis of uncounted votes in the 2000 U.S. presidential election. In Section 5 the convergence results for $\boldsymbol{\beta}$ are used for inference; simulations show that the proposed test performs well compared to competitors. The methods in this paper, as well as the election data, are available in the `R` package `ConSpline`.

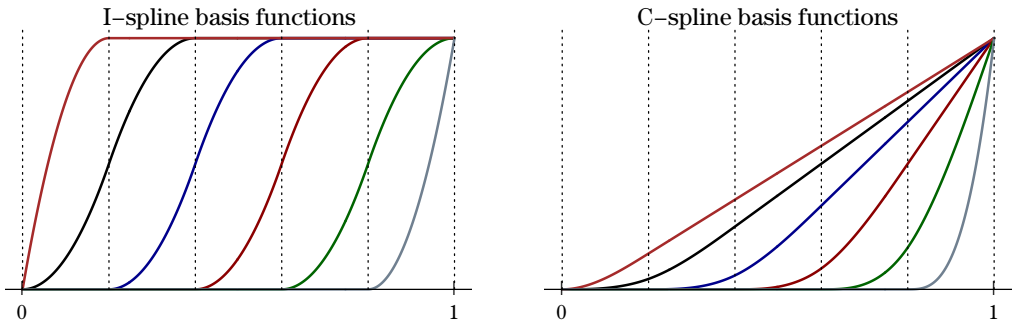

Figure 1. Spline basis functions, with the knots indicated by the dotted vertical lines.

2. Spline Estimation

We assume that f is smooth and has a shape, such as increasing or convex, and estimate f using polynomial splines of the appropriate degree for the shape assumption. For example, we use quadratic splines for monotonicity, because a quadratic spline function has piece-wise linear first derivatives, and hence is increasing if and only if it is increasing at the knots. The second derivative of cubic splines is piece-wise linear, so the spline function is convex if and only if the second derivative is non-negative at the knots. Combinations of monotonicity and convexity also use cubic splines; for details see Meyer (2008). The most popular formulation for spline basis functions is the B-splines of de Boor (2001); other formulations include M-splines, I-splines, and C-splines; see Ramsay (1988) and Meyer (2008). The I-splines and C-splines are specifically formulated for shape constraints and are shown in Figure 1.

The I-spline basis functions are piece-wise quadratic, and at each of the knots $\xi_1, \ldots, \xi_K$, there is exactly one basis function with non-zero slope. Each basis function is non-decreasing, so that a linear combination of I-spline basis functions is increasing if and only if the coefficients are positive. Any quadratic spline function is a linear combination of these basis functions, plus a constant. The C-spline basis functions are convex and piece-wise cubic, and at each knot, there is exactly one basis function with non-zero second derivative. Hence, a linear combination of C-spline basis functions is convex if and only if the coefficients are positive. Any cubic spline function is a linear combination of these basis functions, plus a constant, plus a multiple of the identity function.

Basis vectors $\boldsymbol{z}_1, \ldots, \boldsymbol{z}_m$ in $\mathbb{R}^n$ can be defined, where z_{ji} is the jth basis function evaluated at t_i, $i = 1, \ldots, n$ and $j = 1, \ldots, m$. If $\boldsymbol{Z}$ is an $n \times m$ matrix with columns $\boldsymbol{z}_1, \ldots, \boldsymbol{z}_m$, and the rows of $\boldsymbol{W}$ are $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n$, then we can estimate

$\boldsymbol{\mu}$, where $\mu_i = E(y_i)$, as $\boldsymbol{Z\alpha} + \boldsymbol{W}_0\boldsymbol{\beta}_0 + \boldsymbol{W\beta}$, with the shape constraints holding if and only if $\boldsymbol{\alpha} \geq \boldsymbol{0}$. For the monotone assumption, $m = K$, $\boldsymbol{Z}$ contains the I-spline basis vectors, and $\boldsymbol{W}_0 = \boldsymbol{1} = (1, \ldots, 1)^\top$. For convex constraints, $m = K$, $\boldsymbol{Z}$ contains the C-spline basis vectors, and $\boldsymbol{W}_0 = [\boldsymbol{1}|\boldsymbol{t}]$, where $\boldsymbol{t} = (t_1, \ldots, t_n)^\top$. For increasing and convex constraints, $m = K+1$, $\boldsymbol{Z}$ contains the C-spline basis vectors and $\boldsymbol{t}$, and $\boldsymbol{W}_0 = \boldsymbol{1}$. In each case, the number of columns of $\boldsymbol{W}$, $\boldsymbol{W}_0$, and $\boldsymbol{Z}$ sum to $K + q - 2 + p$, where q is the order of the splines ($q = 3$ for I-splines and $q = 4$ for C-splines). For any shape assumption, we assume the columns of $\boldsymbol{Z}$, $\boldsymbol{W}_0$, and $\boldsymbol{W}$ together form a linearly independent set.

The set

$$\mathcal{C} = \{\boldsymbol{\mu} \in \mathbb{R}^n : \boldsymbol{\mu} = \boldsymbol{Z\alpha} + \boldsymbol{W}_0\boldsymbol{\beta}_0 + \boldsymbol{W\beta}, \text{ where } \boldsymbol{\alpha} \geq \boldsymbol{0}\},$$

is a closed convex cone, where a set $\mathcal{C}$ is a "cone" if for any $\boldsymbol{\mu} \in \mathcal{C}$, positive multiples of $\boldsymbol{\mu}$ are also in $\mathcal{C}$. The least-squares estimator for $\boldsymbol{\mu}$ is the projection of $\boldsymbol{y}$ onto $\mathcal{C}$. For a comprehensive treatment of cones and cone projection, see Rockafellar (1970) or Silvapulle and Sen (2005).

Let $\mathcal{V}$ be the linear space spanned by the columns of $\boldsymbol{W}_0$, and $\boldsymbol{W}$; this is the largest linear space contained in the cone. Define $\Omega = \mathcal{C} \cap \mathcal{V}^\perp$, where $\mathcal{V}^\perp$ is the linear space orthogonal to $\mathcal{V}$. Then Ω is a cone that does not contain a linear space of dimension larger than zero, and hence has a set of edges that are unique up to positive multiplicative constants. An "edge" is defined as a vector in the cone that is not a linear combination with positive coefficients of other non-zero vectors in the cone. The proof of the following is in the appendix.

Theorem 1. *The edges and generators of Ω are the columns $\boldsymbol{\delta}_1, \ldots, \boldsymbol{\delta}_m$ of $\boldsymbol{\Delta} = (\boldsymbol{I} - \boldsymbol{P}_\mathcal{V})\boldsymbol{Z}$, where $\boldsymbol{P}_\mathcal{V}$ is the projection matrix for $\mathcal{V}$,*

$$\Omega = \left\{ \boldsymbol{\theta} \in I\!R^n : \boldsymbol{\theta} = \sum_{j=1}^m \alpha_j \boldsymbol{\delta}_j, \text{ where } \alpha_j \geq 0, \ j = 1, \ldots, m \right\}.$$

There are 2^m faces of Ω, indexed by subsets $J \subseteq \{1, \ldots, m\}$, where

$$\mathcal{F}_J = \left\{ \boldsymbol{\theta} \in \mathbb{R}^n : \boldsymbol{\theta} = \sum_{j \in J} \alpha_j \boldsymbol{\delta}_j, \text{ where } \alpha_j > 0, \ j \in J \right\}.$$

The faces partition the cone Ω, so that the projection of $\boldsymbol{y}$ onto Ω lands on one of the faces. The set of $\boldsymbol{y} \in \mathbb{R}^n$ that land on $\mathcal{F}_J$ is itself a convex cone $\mathcal{C}_J$. Because Ω and $\mathcal{V}$ are orthogonal, the projection of $\boldsymbol{y}$ onto $\mathcal{C}$ is the sum of the projections of $\boldsymbol{y}$ onto Ω and $\mathcal{V}$.

The cone projection algorithm of Meyer (2013) (available as the function `coneB` in the `R` package `coneproj`) determines a subset of the edges, say the columns of $\boldsymbol{\Delta}_J$, so that the projection $\hat{\boldsymbol{\theta}}$ of $\boldsymbol{y}$ onto Ω coincides with the projection of $\boldsymbol{y}$ onto the linear space spanned by the subset of those edges, $\hat{\boldsymbol{\theta}} = \boldsymbol{\Delta}_J(\boldsymbol{\Delta}_J^\top\boldsymbol{\Delta}_J)^{-1}\boldsymbol{\Delta}_J^\top\boldsymbol{y}$. Then $\hat{\boldsymbol{\mu}} = \hat{\boldsymbol{\theta}} + \hat{\boldsymbol{v}}$, where $\hat{\boldsymbol{v}}$ is the projection of $\boldsymbol{y}$ onto $\mathcal{V}$. Let the columns of $\boldsymbol{Z}_J$ be the subset of columns of $\boldsymbol{Z}$ indexed by J, let $\boldsymbol{Z}_+ = [\boldsymbol{Z}_J|\boldsymbol{W}_0]$, and let $\boldsymbol{P}_J$ be the projection matrix for the column space of $\boldsymbol{Z}_+$. Let P_w be the projection matrix for the space spanned by the columns of $\boldsymbol{W}$ and let $\hat{\boldsymbol{\alpha}}_+^\top = [\hat{\boldsymbol{\alpha}}^\top, \hat{\boldsymbol{\beta}}_0^\top]$, where $\hat{\boldsymbol{\alpha}}$, $\hat{\boldsymbol{\beta}}_0$ and $\hat{\boldsymbol{\beta}}$ are the least-squares estimators obtained through the cone projection.

Lemma 1. *For $\boldsymbol{y} \in \mathcal{C}_J$, the least-squares estimate of the parameter vector $\boldsymbol{\beta}$ is given by*

$$\hat{\boldsymbol{\beta}} = [\boldsymbol{W}^\top(\boldsymbol{I} - \boldsymbol{P}_J)\boldsymbol{W}]^{-1}\boldsymbol{W}^\top(\boldsymbol{I} - \boldsymbol{P}_J)\boldsymbol{y}, \tag{2.1}$$

where $\boldsymbol{P}_J$ is the projection matrix for the linear space spanned by the columns of $\boldsymbol{Z}_J$ and $\boldsymbol{W}_0$. Further, $\hat{\boldsymbol{\alpha}}_+ = (\boldsymbol{Z}_+^\top(\boldsymbol{I} - \boldsymbol{P}_w)\boldsymbol{Z}_+)^{-1}\boldsymbol{Z}_+^\top(\boldsymbol{I} - \boldsymbol{P}_x)\boldsymbol{y}$.

Proof : Defining $\boldsymbol{U} = [\boldsymbol{Z}_+|\boldsymbol{W}]$ and using the formula for the inverse of block matrices, we have

$$\begin{aligned}
(\boldsymbol{U}^\top\boldsymbol{U})^{-1} &= \begin{bmatrix} \boldsymbol{Z}_+^\top\boldsymbol{Z}_+ & \boldsymbol{Z}_+^\top\boldsymbol{W} \\ \boldsymbol{W}^\top\boldsymbol{Z}_+ & \boldsymbol{W}^\top\boldsymbol{W} \end{bmatrix}^{-1} \\
&= \begin{bmatrix} (\boldsymbol{Z}_+^\top(\boldsymbol{I} - \boldsymbol{P}_w)\boldsymbol{Z}_+)^{-1} & -(\boldsymbol{Z}_+^\top(\boldsymbol{I} - \boldsymbol{P}_w)\boldsymbol{Z}_+)^{-1} \\ & \boldsymbol{Z}_+^\top\boldsymbol{W}(\boldsymbol{W}^\top\boldsymbol{W})^{-1} \\ -(\boldsymbol{W}^\top(\boldsymbol{I} - \boldsymbol{P}_J)\boldsymbol{W})^{-1} & (\boldsymbol{W}^\top(\boldsymbol{I} - \boldsymbol{P}_J)\boldsymbol{W})^{-1} \\ \boldsymbol{W}^\top\boldsymbol{Z}_+(\boldsymbol{Z}_+^\top\boldsymbol{Z}_+)^{-1} & \end{bmatrix}, \\
\begin{bmatrix} \hat{\boldsymbol{\alpha}}_+ \\ \hat{\boldsymbol{\beta}} \end{bmatrix} &= (\boldsymbol{U}^\top\boldsymbol{U})^{-1}\boldsymbol{U}^\top\boldsymbol{y} \\
&= \begin{bmatrix} (\boldsymbol{Z}_+^\top(\boldsymbol{I} - \boldsymbol{P}_w)\boldsymbol{Z}_+)^{-1}\boldsymbol{Z}_+^\top\boldsymbol{y} \\ -(\boldsymbol{Z}_+^\top(\boldsymbol{I} - \boldsymbol{P}_x)\boldsymbol{Z}_+)^{-1}\boldsymbol{Z}_+^\top\boldsymbol{W}(\boldsymbol{W}^\top\boldsymbol{W})^{-1}\boldsymbol{W}^\top\boldsymbol{y} \\ -(\boldsymbol{W}^\top(\boldsymbol{I} - \boldsymbol{P}_J)\boldsymbol{W})^{-1}\boldsymbol{W}^\top\boldsymbol{Z}_+(\boldsymbol{Z}_+^\top\boldsymbol{Z}_+)^{-1}\boldsymbol{Z}_+^\top\boldsymbol{y} \\ +(\boldsymbol{W}^\top(\boldsymbol{I} - \boldsymbol{P}_J)\boldsymbol{W})^{-1}\boldsymbol{W}^\top\boldsymbol{y} \end{bmatrix} \\
&= \begin{bmatrix} (\boldsymbol{Z}_+^\top(\boldsymbol{I} - \boldsymbol{P}_w)\boldsymbol{Z}_+)^{-1}\boldsymbol{Z}_+^\top(\boldsymbol{I} - \boldsymbol{P}_w)\boldsymbol{y} \\ (\boldsymbol{W}^\top(\boldsymbol{I} - \boldsymbol{P}_J)\boldsymbol{W})^{-1}\boldsymbol{W}^\top(\boldsymbol{I} - \boldsymbol{P}_J)\boldsymbol{y}. \end{bmatrix}.
\end{aligned}$$

If $\mathrm{cov}(\boldsymbol{\varepsilon}) = \sigma^2\boldsymbol{A}$ for $\boldsymbol{A}$ known and positive definite, the transformation to

the uncorrelated case is straightforward. For $\boldsymbol{A} = \boldsymbol{L}\boldsymbol{L}^\top$ with lower-triangular $\boldsymbol{L}$, we can write $\tilde{\boldsymbol{y}} = \tilde{\boldsymbol{Z}}\boldsymbol{\alpha} + \tilde{\boldsymbol{W}}_0\boldsymbol{\beta}_0 + \tilde{\boldsymbol{W}}\boldsymbol{\beta} + \tilde{\boldsymbol{\varepsilon}}$, where $\tilde{\boldsymbol{y}} = \boldsymbol{L}^{-1}\boldsymbol{y}$, $\tilde{\boldsymbol{Z}} = \boldsymbol{L}^{-1}\boldsymbol{Z}$, etc., and $\mathrm{cov}(\tilde{\boldsymbol{\varepsilon}}) = \sigma^2\boldsymbol{I}$, the identity matrix. The parameters are estimated as above using the transformed model, and the inference of Sections 4 and 5 take place in the transformed model as well. Therefore, without loss of generality, we take $\mathrm{cov}(\boldsymbol{\varepsilon}) = \sigma^2\boldsymbol{I}$.

3. Rates of Convergence

In this section we use K_n for the number of knots, and J_n for the set of edge indices used in the cone projection, to emphasize that these are increasing with n. The derivation of the convergence rate for the unconstrained estimator of $\boldsymbol{\mu} = \mathrm{E}(\boldsymbol{y})$ follows the ideas of Huang (2001) and earlier papers referenced therein, and is outlined here. The first four assumptions are standard.

(A1) Let G be a continuous cdf with support $[0,1]$, such that the density g is bounded away from zero and infinity, and assume that for any $c \in [0,1]$, the proportion of t_i, $i = 1,\ldots,n$, such that $t_i \le c$, converges to $G(c)$ as n increases.

(A2) Assume $f \in C^q[0,1]$.

For monotone f we use quadratic splines ($q = 3$) and for convex f, cubic splines are used ($q = 4$).

(A3) Suppose the number K_n of knots grows with n, and there is an $M > 0$ such that $K_n M^{-1} \le \xi_{j+1} - \xi_j \le K_n^{-1} M$.

(A4) The errors ε_i have common variance σ^2 and bounded fourth moments.

(A5) For the monotone assumption we have $f'(t) \ge \epsilon > 0$ on $[0,1]$, and for the convex assumption, $f''(t) \ge \epsilon > 0$ on $[0,1]$. For the increasing and convex assumption, constraints hold strictly if $f''(t) \ge \epsilon > 0$ on $[0,1]$ and $f'(0) \ge \epsilon > 0$.

Lemma 2. *(Theorem* 6.25 *of* Schumaker (2007)*) For a function f satisfying* (A2)*, there is a function f_0 in the linear space of spline functions with K_n knots satisfying* (A3)*, and constants $C_r > 0$ not depending on K_n, such that for $r = 0,1,\ldots,q-1$,*

$$\sup_{t\in[0,1]} |f_0^{(r)}(t) - f^{(r)}(t)| \le C_r K_n^{-(q-r)}.$$

Define $\bar{\boldsymbol{\mu}}$ to be the projection of $\boldsymbol{\mu}$ onto the linear space spanned by the columns of $[\boldsymbol{Z}|\boldsymbol{W}_0|\boldsymbol{W}]$ and define $\boldsymbol{\theta} \in \mathbb{R}^n$ so that $\theta_i = f(t_i)$. If $\boldsymbol{\theta}_0 \in \mathbb{R}^n$ is such that $\theta_{0i} = f_0(t_i)$, and $\mu_{0i} = \theta_{0i} + \boldsymbol{x}_i^\top \boldsymbol{\beta}$, then

$$\frac{1}{n}\sum_{i=1}^{n}(\bar{\mu}_i - \mu_i)^2 \le \frac{1}{n}\sum_{i=1}^{n}(\mu_{0i} - \mu_i)^2 = \frac{1}{n}\sum_{i=1}^{n}(\theta_{0i} - \theta_i)^2 = O(K_n^{-2q}) \tag{3.1}$$

by Lemma 2 with $r = 0$. Let $\tilde{\boldsymbol{\mu}}$ be the projection of $\boldsymbol{y}$ onto the linear space spanned by the columns of $[\boldsymbol{Z}|\boldsymbol{W}_0|\boldsymbol{W}]$. If $\boldsymbol{P}_L$ is the projection matrix for this space, then

$$\sum_{i=1}^{n}(\tilde{\mu}_i - \bar{\mu}_i)^2 = \sum_{i=1}^{n}(\boldsymbol{P}_L\boldsymbol{\varepsilon})_i^2 = O_p(K_n)$$

because the dimension of the linear space is on the order of K_n. For, if $\boldsymbol{u}_1, \ldots, \boldsymbol{u}_k$ forms an orthonormal basis for the linear space of spline functions, then $\boldsymbol{P}_L\boldsymbol{\varepsilon} = \sum_{j=1}^{k} a_j \boldsymbol{u}_j$, where $a_j = \boldsymbol{u}_j^\top \boldsymbol{\varepsilon}$. Then $\sum_{i=1}^{n}(\boldsymbol{P}_L\boldsymbol{\varepsilon})_i^2 = \sum_{j=1}^{k} a_j^2$, and because the errors have finite fourth moments, each term in the sum is bounded in probability.

Therefore, we have the global convergence rate

$$\frac{1}{n}\sum_{i=1}^{n}(\tilde{\mu}_i - \mu_i)^2 = O_p(K_n^{-2q} + K_n n^{-1}),$$

which is minimized when $K_n = O(n^{1/(2q+1)})$.

Theorem 2. *The constrained estimator attains the same rate as the unconstrained estimator:*

$$\frac{1}{n}\sum_{i=1}^{n}(\hat{\mu}_i - \mu_i)^2 = O_p(n^{-2q/(2q+1)}).$$

Proof: Using (A4) and Lemma 2, we find a $\boldsymbol{\mu}_0 \in \mathcal{C}$ that is close to $\boldsymbol{\mu}$. For the monotone case, there is a quadratic spline function f_0 such that $\sup_{t\in[0,1]} |f_0'(t) - f'(t)| \le C_1 n^{-2/7}$. For large enough n, $C_1 n^{-2/7} < \epsilon$, so f_0 is increasing. Let $\boldsymbol{\theta}_0$ be the vector corresponding to f_0, hence $\boldsymbol{\mu}_0 = \boldsymbol{\theta}_0 + \boldsymbol{W}\boldsymbol{\beta} \in \mathcal{C}$. For the convex case, there is a cubic spline function f_0 such that $\sup_{t\in[0,1]} |f_0''(t) - f''(t)| \le C_2 n^{-2/9}$. For large enough n, $C_2 n^{-2/9} < \epsilon$, so f_0 is convex. Using the notation $\|\boldsymbol{a}\|^2 = \boldsymbol{a}^\top \boldsymbol{a}$, we have

$$\begin{aligned}\|\tilde{\boldsymbol{\mu}} - \boldsymbol{\mu}_0\|^2 - \|\hat{\boldsymbol{\mu}} - \boldsymbol{\mu}_0\|^2 &= \|\tilde{\boldsymbol{\mu}} - \hat{\boldsymbol{\mu}}\|^2 + 2(\tilde{\boldsymbol{\mu}} - \hat{\boldsymbol{\mu}})^\top(\hat{\boldsymbol{\mu}} - \boldsymbol{\mu}_0) \\ &= \|\tilde{\boldsymbol{\mu}} - \hat{\boldsymbol{\mu}}\|^2 - 2(\boldsymbol{y} - \tilde{\boldsymbol{\mu}})^\top(\hat{\boldsymbol{\mu}} - \boldsymbol{\mu}_0) + 2(\boldsymbol{y} - \hat{\boldsymbol{\mu}})^\top(\hat{\boldsymbol{\mu}} - \boldsymbol{\mu}_0).\end{aligned}$$

The second term in the last expression is zero by orthogonality of the projection with anything in the space. Because $\boldsymbol{\mu}_0 \in \mathcal{C}$, the third term is positive by

the Karush-Kuhn-Tucker conditions (see Silvapulle and Sen (2005), Proposition 3.12.3). Therefore, $\|\hat{\boldsymbol{\mu}} - \boldsymbol{\mu}_0\|^2 \leq \|\tilde{\boldsymbol{\mu}} - \boldsymbol{\mu}_0\|^2$. Then

$$\|\hat{\boldsymbol{\mu}} - \boldsymbol{\mu}\| \leq \|\hat{\boldsymbol{\mu}} - \boldsymbol{\mu}_0\| + \|\boldsymbol{\mu}_0 - \boldsymbol{\mu}\| \leq \|\tilde{\boldsymbol{\mu}} - \boldsymbol{\mu}_0\| + \|\boldsymbol{\mu}_0 - \boldsymbol{\mu}\| \leq \|\tilde{\boldsymbol{\mu}} - \boldsymbol{\mu}\| + 2\|\boldsymbol{\mu}_0 - \boldsymbol{\mu}\|,$$

which is sufficiently small, and hence the constrained estimator attains the same rate as the unconstrained estimator.

Next, we consider conditions under which $\hat{\boldsymbol{\beta}}$ can attain a root-n convergence rate, using methods similar to those in Chen (1988). Take $h_j(t) = \mathrm{E}(X_j|T = t)$, $j = 1, \ldots, p$, and $\Sigma_t = \mathrm{cov}(\boldsymbol{X}|T = t)$, with $\boldsymbol{X} = (X_1, \ldots, X_p)^\top$. Assumption (A5) ensures that the matrix $\boldsymbol{W}^\top(\boldsymbol{I} - \boldsymbol{P}_{J_n})\boldsymbol{W}$ is non-singular for $J_n = \{1, \ldots, m\}$, and hence for any subset J_n – see the proof of Lemma 2 of Chen (1988).

(A5) There exist positive definite matrices Σ_0 and Σ_1 such that both $\Sigma_t - \Sigma_0$ and $\Sigma_1 - \Sigma_t$ are positive definite, for all $t \in [0, 1]$.

(A6) The functions $h_j(t)$ have q continuous derivatives.

The following corollary to Theorem 2 is proved in the Appendix.

Corollary 1. *The components do not converge more slowly than* $\hat{\boldsymbol{\mu}}$*,*

$$\frac{1}{n}\sum_{i=1}^{n}(\hat{\theta}_i - \theta_i)^2 = O_p(n^{-2q/(2q+1)}) \;\; \textit{and} \;\; \frac{1}{n}\sum_{i=1}^{n}[(\boldsymbol{W}\hat{\boldsymbol{\beta}})_i - (\boldsymbol{W}\boldsymbol{\beta})_i]^2 = O_p(n^{-2q/(2q+1)}).$$

Take $\eta_j = X_j - h_j(T)$, and let $\boldsymbol{\eta}_j \in \mathbb{R}^n$ be such that $\eta_{ji} = x_{ji} - h_j(t_i)$, and $\boldsymbol{h}_j \in \mathbb{R}^n$ where the ith element of $\boldsymbol{h}_j$ is $h_j(t_i)$. If

$$\sigma_{j\ell} = \mathrm{cov}(X_j - h_j(T), X_\ell - h_\ell(T))$$

then by the law of large numbers, $\boldsymbol{\eta}_j^\top\boldsymbol{\eta}_\ell/n \xrightarrow{p} \sigma_{j\ell}$. Define the matrix Σ with elements $\sigma_{j\ell}$.

Theorem 3. *Under* (A1)-(A6), $\boldsymbol{W}^\top(\boldsymbol{I} - \boldsymbol{P}_{J_n})\boldsymbol{W}/n \xrightarrow{p} \Sigma$.

Proof: The j, ℓth element of $\boldsymbol{W}^\top(\boldsymbol{I} - \boldsymbol{P}_{J_n})\boldsymbol{W}$ is $[(\boldsymbol{I} - \boldsymbol{P}_{J_n})\boldsymbol{x}_j]^\top[(\boldsymbol{I} - \boldsymbol{P}_{J_n})\boldsymbol{x}_\ell]$, by the idempotence of projection matrices, and $(\boldsymbol{I} - \boldsymbol{P}_{J_n})\boldsymbol{x}_j = (\boldsymbol{I} - \boldsymbol{P}_{J_n})(\boldsymbol{h}_j + \boldsymbol{\eta}_j)$. The term $(\boldsymbol{I} - \boldsymbol{P}_{J_n})\boldsymbol{h}_j$ is the residual of the spline fit to $\boldsymbol{h}_j$ and, by (A6), this is sufficiently small if the space spanned by the cone edges indexed by J_n is sufficiently "rich." In particular, the *knots* indexed by J_n must satisfy (A3). The following lemma is proved in the Appendix:

Lemma 3. *Let* A_n *be the event that* $j, j+1 \notin J_n$ *for some* $j = 1, \ldots, K_n - 1$. *Then* $\lim_{n\to\infty} P(A_n) = 0$.

If J_n contains indices for at least every other knot, then assumption (A3) holds for the sequence of knots indexed by J_n. That is, we have bounded mesh ratio for the knots "used" in the constrained case. Then $\|(\boldsymbol{I} - \boldsymbol{P}_{J_n})\boldsymbol{h}_j\|^2 = O_p(n^{1/(2q+1)})$. Using $\boldsymbol{a}^\top\boldsymbol{b} \leq \|\boldsymbol{a}\|\|\boldsymbol{b}\|$, we have $[(\boldsymbol{I}-\boldsymbol{P}_{J_n})\boldsymbol{h}_j]^\top[(\boldsymbol{I}-\boldsymbol{P}_{J_n})\boldsymbol{h}_\ell] = o_p(n)$ and $\boldsymbol{\eta}_j^\top(\boldsymbol{I}-\boldsymbol{P}_{J_n})\boldsymbol{h}_\ell = o_p(n)$ by Chebyshev's inequality. Finally, $\boldsymbol{\eta}_j^\top(\boldsymbol{I}-\boldsymbol{P}_{J_n})\boldsymbol{\eta}_\ell = \boldsymbol{\eta}_j^\top\boldsymbol{\eta}_\ell + o_p(n)$, and the j,ℓth element of $\boldsymbol{W}^\top(\boldsymbol{I} - \boldsymbol{P}_{J_n})\boldsymbol{W}$ is $\boldsymbol{\eta}_j^\top\boldsymbol{\eta}_\ell + o_p(n)$.

Theorem 4. *Under (A1)-(A6), we have*

$$\sqrt{n}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) = \sqrt{n}[\boldsymbol{W}^\top(\boldsymbol{I} - \boldsymbol{P}_{J_n})\boldsymbol{W}]^{-1}\boldsymbol{W}^\top(\boldsymbol{I} - \boldsymbol{P}_{J_n})\boldsymbol{\varepsilon} + o_p(1).$$

Proof: From (2.1), we have

$$\begin{aligned}\hat{\boldsymbol{\beta}} - \boldsymbol{\beta} &= [\boldsymbol{W}^\top(\boldsymbol{I} - \boldsymbol{P}_{J_n})\boldsymbol{W}]^{-1}\boldsymbol{W}^\top(\boldsymbol{I} - \boldsymbol{P}_{J_n})(\boldsymbol{\theta} + \boldsymbol{\varepsilon}) \\ &= [\boldsymbol{W}^\top(\boldsymbol{I}-\boldsymbol{P}_{J_n})\boldsymbol{W}]^{-1}\boldsymbol{W}^\top(\boldsymbol{I}-\boldsymbol{P}_{J_n})\boldsymbol{\theta}+[\boldsymbol{W}^\top(\boldsymbol{I}-\boldsymbol{P}_{J_n})\boldsymbol{W}]^{-1}\boldsymbol{W}^\top(\boldsymbol{I}-\boldsymbol{P}_{J_n})\boldsymbol{\varepsilon},\end{aligned}$$

so interest is in showing that the first term in the last expression is negligible compared to the second term. The jth element of $\boldsymbol{W}^\top(\boldsymbol{I} - \boldsymbol{P}_{J_n})\boldsymbol{\theta}$ is $\boldsymbol{x}_j^\top(\boldsymbol{I} - \boldsymbol{P}_{J_n})\boldsymbol{\theta} = (\boldsymbol{h}_j + \boldsymbol{\eta}_j)^\top(\boldsymbol{I} - \boldsymbol{P}_{J_n})\boldsymbol{\theta}$. Recalling $\|(\boldsymbol{I} - \boldsymbol{P}_{J_n})\boldsymbol{\theta}\|^2 = O_p(n^{1/(2q+1)})$, we have $\boldsymbol{\eta}_j^\top(\boldsymbol{I}-\boldsymbol{P}_{J_n})\boldsymbol{\theta} = o_p(n^{1/2})$ and $\boldsymbol{h}_j^\top(\boldsymbol{I}-\boldsymbol{P}_{J_n})\boldsymbol{\theta} = [(\boldsymbol{I}-\boldsymbol{P}_{J_n})\boldsymbol{h}_j]^\top[(\boldsymbol{I}-\boldsymbol{P}_{J_n})\boldsymbol{\theta}] = o_p(n^{1/2})$.

4. Inference for the Nonlinear Term

Assuming normal errors, an exact test for $H_0 : \boldsymbol{\mu} \in \mathcal{V}$ versus $H_1 : \boldsymbol{\mu} \in \mathcal{C}$ is available. For example, we can test constant versus increasing f, or linear versus convex f, while controlling for covariates. If SSR_0 is the sum of squared residuals under H_0 and SSR_1 is the sum of squared residuals under H_1, then large values of

$$B_{01} = \frac{SSR_0 - SSR_1}{SSR_0}$$

support the alternative hypothesis. A standard result in cone projection is that the null distribution of B_{01} is that of a mixture of Beta random variables. Specifically, when H_0 is true,

$$\mathrm{P}\left(B_{01} \leq a\right) = \sum_{d=0}^{m} \mathrm{P}\left[B\left(\frac{d}{2}, \frac{(n-d)}{2}\right) \leq a\right] p_d,$$

where $B(a, b)$ is a Beta(a, b) random variable, and $p_0, \ldots, p_m$ are mixing probabilities readily obtained through simulations to the desired precision. See Robertson, Wright and Dykstra (1988) for details concerning the constant versus increasing test, without smoothing or covariates. Meyer (2003) gave a similar result for the linear versus convex test, again without smoothing or covariates. The gen-

eral test for $\mathcal{V}$ versus $\mathcal{C}$ was given by Raubertas, Lee and Nordheim (1986). This extension to the partial linear model allows for a test against a constant or linear f, while controlling for possibly confounding covariates.

To demonstrate the test, we look at voting data from the 2000 U.S. presidential election, in the state of Georgia, using the percent of uncounted votes as a response variable and asking whether the percent of uncounted votes is increasing in the percent of black voters registered in the county. The data are shown in Figure 2; for more details see Meyer (2002). Because the response is a proportion, the variance is inversely proportional to the number of ballots, so this can be used as a weight. The test of constant versus increasing regression function, without accounting for covariates, gives p-values less than 10^{-9}, for numbers of knots between 5 and 9, indicating that percent of uncounted votes is significantly higher in counties with more black voters. With $K = 7$ knots (the default in `ConSpline`), the coefficient of determination is $R^2 = 0.28$. When we control for the method of voting, the p-value continues to be small (about 10^{-7}), and $R^2 = 0.38$. In addition we find that the method is significant, with punch cards having a significantly higher proportion of uncounted votes compared with OS-PC (optical scan, precinct count) and the lever, OS-CC (optical scan, central count), while paper ballots are not significantly different from punch cards. When we include an economics covariate, it is highly significant that the richer counties have lower proportions of uncounted votes, and poorer counties have more. Here $R^2 = 0.53$ and $p = 0.88$ for the test of constant versus increasing proportion of uncounted votes with the number of black voters. The proportion of black voters is strongly confounded with the economics variable, showing the need for inclusion of covariates in this test. The voting method is also quite confounded with economic status – after the latter variable is controlled for, we find that the lever method and both optical scan methods are associated with a significantly lower proportion of uncounted votes, compared to the punch card method.

5. Inference for the Linear Term

The result (2.1) gives $\text{cov}(\hat{\boldsymbol{\beta}}) \approx [\boldsymbol{X}^\top(\boldsymbol{I} - \boldsymbol{P}_J)\boldsymbol{X}]^{-1}\sigma^2$ for $\boldsymbol{y} \in \mathcal{C}_J$; this is used for inference concerning the linear term. A variance estimate $\hat{\sigma}^2$ typically uses the sum of squared residuals divided by the residual degrees of freedom. However for cone regression, Meyer and Woodroofe (2000) showed that this tends to underestimate the variance, with

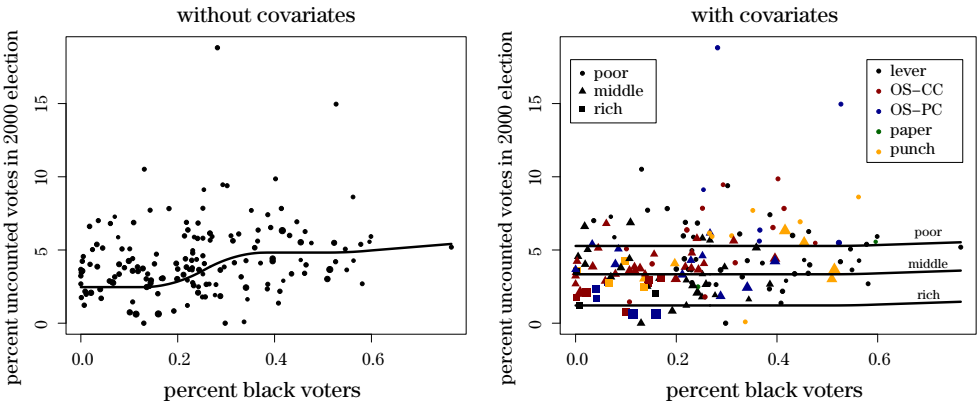

Figure 2. Percent of uncounted votes in the 2000 U.S. presidential election is plotted against percent of black voters, for the $n = 159$ counties in Georgia, with the size of the plot character proportional to the log of the number of ballots cast in the county. In the second plot, the shape is the economic status of the county, and the shading represents the type of voting system used by the county. The curves are the estimated trends for the lever method and the three economic status levels.

$$n - 2(\mathrm{E}(D) + d_0 + p) \le \frac{\mathrm{E}(SSR)}{\sigma^2} \le n - (\mathrm{E}(D) + d_0 + p),$$

where D is the size of J and d_0 is the number of columns of $\boldsymbol{X}_0$. This suggests that a reasonable estimator of σ^2 is

$$\hat{\sigma}^2 = \frac{SSR}{n - c(d + d_0 + p)},$$

where $c \in (1, 2)$ and d is the observed size of J. For unsmoothed monotone regression, Meyer and Woodroofe (2000) demonstrated that c is about 1.5, but there is empirical evidence that c should be smaller for convex regression and for splines. In the spline case, if $c > 1$, then $c(d + p)$ may often be larger than $m + d_0 + p$, the largest possible degrees of freedom. For a conservative choice, $edf = \max(1.2(d + d_0 + p), m + d_0 + p)$, and $\hat{\sigma}^2 = SSR/(n - edf)$. Then t or F statistics can be used to test hypotheses about $\boldsymbol{\beta}$, with edf for the model degrees of freedom and $\hat{\mathrm{cov}}(\hat{\boldsymbol{\beta}}) = [\boldsymbol{X}^\top(\boldsymbol{I} - \boldsymbol{P}_J)\boldsymbol{X}]^{-1}\hat{\sigma}^2$.

For a simple demonstration, we did simulations to compare the size and power of the test $H_0 : \beta = 0$ versus $H_1 : \beta \neq 0$, where $y_i = f(t_i) + \beta x_i + \varepsilon_i$, $i = 1, \ldots, n$, x_i with values in $\{0, 1\}$, and the ε_i independent standard normal. The t values were equally spaced in $[0, 1]$, and the probability that $x_i = 1$ was $\max[t^4, (1-t)^4]$, so $x = 1$ was more likely when t was close to zero or close to one, and $x = 0$ was more likely for middle-range t values. The default knots choices in ConSpline were used: $K = \lceil 2n^{1/2(q+1)} \rceil + q$ equally-spaced knots.

Table 1. Proportions of simulated data sets for which $H_0 : \beta = 0$ was rejected. The proposed method is labeled "CS_m" when using increasing constraints, and "CS_{mc}" for increasing and convex constraints. The numbers of knots were 6, 7, and 8, for $n = 100$, $n = 200$, and $n = 400$, respectively. For CS and F-test, $N = 100{,}000$ data sets were generated, but for `scam`, $N = 10{,}000$.

		$f_1(t) = t$			$f_2(t) = 4(t-1/2)^2$			$f_3(t) = 20(t-2/3)_+^2$			
n	β	CS_m	lin	$\texttt{scam}_m$	CS_c	quad	$\texttt{scam}_c$	CS_{mc}	lin	quad	$\texttt{scam}_{mc}$
100	0	0.046	0.050	0.060	0.050	0.050	0.056	0.048	0.389	0.050	0.067
200	0	0.047	0.050	0.060	0.050	0.050	0.053	0.048	0.650	0.050	0.055
400	0	0.047	0.050	0.056	0.050	0.050	0.051	0.050	0.911	0.049	0.051
100	$-1/4$	0.173	0.223	0.227	0.192	0.170	0.176	0.159	0.081	0.164	0.155
200	$-1/4$	0.308	0.403	0.386	0.326	0.296	0.305	0.295	0.108	0.289	0.302
400	$-1/4$	0.541	0.680	0.648	0.557	0.526	0.537	0.572	0.164	0.510	0.530

For $f_1(t) = t$, we compared the proposed test with monotone increasing shape constraints with the test available in `scam`, also using monotone constraints. We also compared with an exact t-test using a parametric model with $f(t) = t$. The results are in Table 1, where the proposed test is labeled CS with a subscript indicating the shape. For $f_2 = 4(x - 1/2)^2$, we compared the proposed test with convex assumptions to the corresponding test in `scam`, and to an exact t test where $f(t)$ was assumed to be quadratic in t. For both the monotone and convex cases, the proposed test had correct or conservative size, while the `scam` test had a slightly inflated test size. The `scam` test had higher power than the proposed test for the monotone case, but the proposed test had higher power for the convex case.

The function $f_3(t) = 20(t-2/3)_+^2$ was chosen to "fool" the t-test with linear assumptions. Because the value of x_i was more likely to be one for t-values in the middle of the range, the linear model chose a negative value for β, to "pull up" the middle observations into a line. Therefore the test size was quite large, and worse for larger n. Although the quadratic model was not fooled, if we choose a different pattern for the probability that $x_i = 1$, such as higher probabilities at the right end only, this test will also have an inflated size. Although (A4) does not hold for f_3, the results for the proposed test are still acceptable, because we use a large number of knots. Because the constrained splines are robust to knot choices (Meyer (2008)), we can choose a number of knots that would result in under-smoothing, if the spline function were unconstrained.

6. Discussion

The proposed estimation and inference methods are available in the `R` package `ConSpline`, where the shape choices are increasing, decreasing, convex or concave, with all four combinations of monotonicity and convexity. The user may specify the knots and choose $c \in [1, 2]$, the adjustment constant for the variance estimation and degrees of freedom. The simulations and example from the previous sections used the default choices.

Our proof that the linear term can attain root-n convergence simultaneously with optimal convergence rate of the regression function uses the assumption that the constraints hold strictly. The root-n convergence can alternatively be attained by under-smoothing, or having the knots grow at a faster rate than is optimal for the estimation of f. For unconstrained splines, choosing a larger number of knots may result in unacceptably "wiggly" function estimates, but the constrained case obviates wiggling. Therefore, the rate at which the knots grow is not a practical consideration, and in `ConSpline`, the default choices are larger numbers of knots than would be reasonable for the unconstrained case. As a result, the test size tends to be conservative rather than inflated, without sacrificing power.

The cone projection methods are computationally efficient. The function `conspline` was called 100,000 times with $n = 400$ for the simulations in Table 1. This took 16 minutes using a Mac Powerbook with a 2.8 GHz processor. In contrast, the 10,000 calls to `scam` took over 25 minutes on the same machine. Simulations to compare mean squared error of `scam` fits with that of the proposed fits were performed for a variety of regression functions, sample sizes, and model variances; no appreciable advantages were found for either method.

Acknowledgment

This research was partially supported by an NSF-DMS grant. The thoughtful comments and suggestions of two referees are much appreciated and substantially improved this work.

Appendix

A. Proofs

Proof of Theorem 1*:* Checking that Ω is a convex cone is straight-forward. Suppose $\boldsymbol{\theta}_1$ and $\boldsymbol{\theta}_2$ are non-zero vectors in Ω such that $\boldsymbol{\theta}_1 + \boldsymbol{\theta}_2 = \boldsymbol{\delta}_j$, the jth

column of $\boldsymbol{\Delta}$. Because $\boldsymbol{\theta}_1$ and $\boldsymbol{\theta}_2$ are in $\mathcal{C}$, we can write $\boldsymbol{\theta}_1 = \boldsymbol{Z}\boldsymbol{\alpha}_1+\boldsymbol{X}_0\boldsymbol{\beta}_{01}+\boldsymbol{X}\boldsymbol{\beta}_1$, and $\boldsymbol{\theta}_2 = \boldsymbol{Z}\boldsymbol{\alpha}_2 + \boldsymbol{X}_0\boldsymbol{\beta}_{02} + \boldsymbol{X}\boldsymbol{\beta}_2$, and we can write $\boldsymbol{Z} = \boldsymbol{P}_{\mathcal{V}}\boldsymbol{Z} + \boldsymbol{\Delta}$. Then $\boldsymbol{\theta}_j = \boldsymbol{\Delta}\boldsymbol{\alpha}_j + \boldsymbol{v}_j$, where $\boldsymbol{v}_j \in \mathcal{V}$. Because $\boldsymbol{\theta}_1$ and $\boldsymbol{\theta}_2$ are in $\mathcal{V}^\perp$, $\boldsymbol{v}_1 = \boldsymbol{v}_2 = 0$. Then $\boldsymbol{\theta}_1 + \boldsymbol{\theta}_2 = \boldsymbol{\Delta}(\boldsymbol{\alpha}_1 + \boldsymbol{\alpha}_2) = \boldsymbol{\Delta}\boldsymbol{e}_j$, where $\boldsymbol{e}_j$ is a vector with all zeros but the jth element is one. Because a basis for $\mathcal{V}$ and columns of $\boldsymbol{Z}$ form a linearly independent set, the columns of $\boldsymbol{\Delta}$ are linearly independent, and we must have $\boldsymbol{\alpha}_1 + \boldsymbol{\alpha}_2 = \boldsymbol{e}_j$. But the elements of $\boldsymbol{\alpha}_1$ and $\boldsymbol{\alpha}_2$ are non-negative, so all elements but the jth must be zero. Therefore, $\boldsymbol{\theta}_1$ and $\boldsymbol{\theta}_2$ are collinear and $\boldsymbol{\delta}_j$ is an edge.

Proof of the Corollary: Suppose $\boldsymbol{a}_n$ and $\boldsymbol{b}_n$ are random vectors taking values in $\mathbb{R}^n$. If $\|\boldsymbol{a}_n + \boldsymbol{b}_n\|$ converges to zero in probability at a rate faster than $\|\boldsymbol{a}_n\|$ converges to zero, then $\|\boldsymbol{a}_n + \boldsymbol{b}_n\|/|\boldsymbol{a}_n\| \xrightarrow{p} 0$. We can write $\boldsymbol{b}_n$ as the sum of its projection onto $\boldsymbol{a}_n$ and the residual of this projection:

$$\boldsymbol{b}_n = \frac{\boldsymbol{a}_n^\top \boldsymbol{b}_n}{\|\boldsymbol{a}_n\|^2}\boldsymbol{a}_n + \boldsymbol{r}_n,$$

so that

$$\|\boldsymbol{a}_n + \boldsymbol{b}_n\|^2 = \left[1 + \frac{\boldsymbol{a}_n^\top \boldsymbol{b}_n}{\|\boldsymbol{b}_n\|^2}\right]^2 \|\boldsymbol{a}_n\|^2 + \|\boldsymbol{r}_n\|^2$$

by orthogonality of $\boldsymbol{a}_n$ and $\boldsymbol{r}_n$. By assumption we have $\|\boldsymbol{r}_n\|^2/\|\boldsymbol{a}_n\|^2 \xrightarrow{p} 0$ and $\boldsymbol{a}_n^\top \boldsymbol{b}_n/\|\boldsymbol{b}_n\|^2 \xrightarrow{p} -1$; then $\boldsymbol{a}_n = -\boldsymbol{b}_n+\boldsymbol{s}_n$, where $\boldsymbol{s}_n/\|\boldsymbol{b}_n\|^2 \xrightarrow{p} 0$. Now if $\boldsymbol{a}_n = \hat{\boldsymbol{\theta}}-\boldsymbol{\theta}$ and $\boldsymbol{b}_n = \boldsymbol{W}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})$, then $\boldsymbol{a}_n + \boldsymbol{b}_n = \hat{\boldsymbol{\mu}} - \boldsymbol{\mu}$. Therefore, if $\hat{\boldsymbol{\mu}} - \boldsymbol{\mu}$ converges to zero faster than $\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}$, then $\hat{\boldsymbol{\theta}} - \boldsymbol{\theta} = -\boldsymbol{W}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})$ plus a negligible residual; this is contradicted by (A5).

Proof of Lemma 3*:* By (A1),

$$\frac{1}{n}\sum_{i=1}^{n}\left[\hat{f}(t_i) - f(t_i)\right]^2 \asymp \int_0^1 \left[\hat{f}(t) - f(t)\right]^2 g(t)dt,$$

where $a_n \asymp b_n$ means $a_n/b_n \xrightarrow{p} 1$. Let $g_0 > 0$ be the smallest value of $g(t)$ on $[0,1]$. By (A3), there is a constant c such that $\inf_{j=1,\ldots,K}(\xi_{j+1} - \xi_j) \geq c/K$ for all sufficiently large n.

For the monotone case and quadratic splines, if $j, j+1 \notin J$, then $\hat{f}'(\xi_j) = \hat{f}'(\xi_{j+1}) = 0$, and $\hat{f}$ is constant in $I_j = [\xi_j, \xi_{j+1}]$.

$$\int_0^1 \left[\hat{f}(t) - f(t)\right]^2 g(t)dt \geq g_0 \int_{\xi_j}^{\xi_{j+1}} \left[\hat{f}(t) - f(t)\right]^2 dt,$$

and because $f'(t) \geq \epsilon > 0$, the integral on the right is minimized when $f'(t) = \epsilon$

over $[\xi_j, \xi_{j+1}]$, so that

$$\int_{\xi_j}^{\xi_{j+1}} \left[\hat{f}(t) - f(t)\right]^2 dt, \geq \epsilon^2 \int_{\xi_j}^{\xi_{j+1}} \left[t - \frac{\xi_j + \xi_{j+1}}{2}\right]^2 dt$$
$$= \frac{\epsilon^2}{12}(\xi_{j+1} - \xi_j)^3 \geq \frac{\epsilon^2}{12} cK^{-3}.$$

By the Corollary to Theorem 2 this is too large, and the probability of the event that $j, j+1 \notin J$ for some $j = 1, \ldots, K-1$, goes to zero.

For the convex case with cubic splines, if $j, j+1 \notin J$, then $\hat{f}''(\xi_j) = \hat{f}''(\xi_{j+1}) = 0$, and $\hat{f}$ is linear in $I_j = [\xi_j, \xi_{j+1}]$. Then the integral $\int_{\xi_j}^{\xi_{j+1}} [\hat{f}(t) - f(t)]^2 dt$ is minimized when $f''(t) = \epsilon$. It is straight-forward to show that for a parabola $p(t)$ with $p''(t) = \epsilon$, and linear $\ell(t)$, then

$$\int_a^b (p(t) - \ell(t)) dt \geq \frac{\epsilon^2 \delta^5}{180},$$

where $\delta = b - a$. Hence if $j, j+1 \notin J$, we have

$$\int_{\xi_j}^{\xi_{j+1}} \left[\hat{f}(t) - f(t)\right]^2 g(t) dt \geq \frac{g_0 \epsilon^2 (\xi_{j+1} - \xi_j)^5}{180} \geq \frac{g_0 \epsilon^2 c K^{-5}}{180}.$$

Again by the Corollary to Theorem 2, the probability of the event that $j, j+1 \notin J$ for some $j = 1, \ldots, K-1$, must go to zero.

References

Chen, H. (1988). Convergence rates for parametric components in a partly linear model. *Annals of Statistics* **16**, 136–146.

de Boor, C. (2001). *A Practical Guide to Splines* (Revised ed.). Springer-Verlag, New York.

Hazelton, M. L. and Turlach, B. A. (2011). Semiparametric regression with shape-constrained penalized splines. *Computational Statistics and Data Analysis* **55**, 2871–2879.

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

Huang, J. (2002). A note on estimating a partly linear model under monotonicity constraints. *Journal of Statistical Planning and Inference* **107**, 343–351.

Huang, J. Z. (2001). Concave extended linear modeling: a theoretical synthesis. *Statistica Sinica* **11**, 173–197.

Meyer, M. C. (2002). Uncounted votes: Does voting equipment matter? *Chance Magazine* **15**(4), 33–38.

Meyer, M. C. (2003). A test for linear vs convex regression function using shape-restricted regression. *Biometrika* **90**(1), 223–232.

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

Meyer, M. C. (2012). Constrained penalized splines. *Canadian Journal of Statistics* **40**(1), 190–206.

Meyer, M. C. (2013). A simple new algorithm for quadratic programming with applications in statistics. *Communications in Statistics* **42**(5), 1126–1139.

Meyer, M. C. and Woodroofe, M. (2000). On the degrees of freedom in shape-restricted regression. *Ann. Statist.* **28**, 1083–1104.

Opsomer, J. D. and Ruppert, D. (1999). A root-n consistent estimator for semiparametric additive modelling. *Journal of Computational and Graphical Statistics* **8**, 715–732.

Pya, N. and Wood, S. N. (2015). Shape constrained additive models. *Statistics and Computing* **25**, 543–559.

Ramsay, J. O. (1988). Monotone regression splines in action. *Statistical Science* **3**(4), 425–461.

Raubertas, R. F., Lee, C.-I. C. and Nordheim, E. V. (1986). Hypothesis tests for normals means constrained by linear inequalities. *Communications in Statistics – Theory and Methods* **15**(9), 2809–2833.

Rice, J. A. (1986). Convergence rates for partially splined models. *Statistics and Probability Letters* **4**, 203–208.

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

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

Schumaker, L. L. (2007). *Spline Functions: Basic Theory*. Cambridge Mathematical Library.

Silvapulle, M. J. and Sen, P. (2005). *Constrained Statistical Inference*. John Wiley & Sons.

Speckman, P. E. (1988). Regression analysis for partially linear models. *Journal of the Royal Statistical Society, Series B* **50**, 413–436.

Stone, C. J. (1980). Optimal rates of convergence for nonparametric estimators. *Annals of Statistics* **8**, 1348–1360.

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

Colorado State University, 212 Statistics Building, Fort Collins, 80523-1877, USA.

E-mail: meyer@stat.colostate.edu

(Received July 2016; accepted December 2016)