The Canadian Journal of Statistics
Vol. 40, No. 1, 2012, Pages 190–206
La revue canadienne de statistique

Constrained penalized splines

Mary C. MEYER*

Colorado State University

Key words and phrases: Cone projection; convex regression; function estimation; model selection; monotone regression.

MSC 2010: Primary 62G08; secondary 62G05.

Abstract: The penalized spline is a popular method for function estimation when the assumption of "smoothness" is valid. In this paper, methods for estimation and inference are proposed using penalized splines under additional constraints of shape, such as monotonicity or convexity. The constrained penalized spline estimator is shown to have the same convergence rates as the corresponding unconstrained penalized spline, although in practice the squared error loss is typically smaller for the constrained versions. The penalty parameter may be chosen with generalized cross-validation, which also provides a method for determining if the shape restrictions hold. The method is not a formal hypothesis test, but is shown to have nice large-sample properties, and simulations show that it compares well with existing tests for monotonicity. Extensions to the partial linear model, the generalized regression model, and the varying coefficient model are given, and examples demonstrate the utility of the methods. *The Canadian Journal of Statistics* 40: 190–206; 2012 © 2012 Statistical Society of Canada

Résumé: Les splines pénalisées sont une méthode populaire en estimation fonctionnelle lorsque l'hypothèse de régularité est valide. Dans cet article, nous proposons des méthodes pour faire de l'estimation et de l'inférence en utilisant les splines pénalisées sous des contraintes de forme supplémentaires telles que la monotonicité ou la convexité. Nous montrons que l'estimateur basé sur des splines pénalisées contraintes a le même taux de convergence que celui basé sur les splines pénalisées non contraintes quoique, en pratique, la perte quadratique est habituellement inférieure pour les versions contraintes. Le paramètre de pénalité peut être choisi à l'aide de la validation croisée généralisée ce qui donne aussi une méthode pour déterminer si les contraintes de forme sont respectées. La méthode n'est pas un test d'hypothèses proprement dit, mais elle a de bonnes propriétés asymptotiques. Des simulations montrent aussi qu'elle se compare avantageusement aux tests de monotonicité déjà existants. Des généralisations aux modèles linéaires partiels, au modèle de régression généralisé ainsi qu'à celui des coefficients variables sont fournies et des exemples illustrent l'utilité de ces méthodes. *La revue canadienne de statistique* 40: 190–206; 2012 © 2012 Société statistique du Canada

1. INTRODUCTION AND BACKGROUND

Consider the problem of estimating a function f from a scatterplot of data $\{(x_i, y_i)\}$, $i = 1, \ldots, n$, using the model

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

where $x_i \in [0, 1]$ and the ϵ_i are *iid* random variables with mean zero and unit variance. Smoothing methods that are well established include moving averages, kernel and local polynomial regression, smoothing splines, regression splines, and penalized splines. For an overview of these methods, see Ruppert, Wand, & Carroll (2003). Several methods are available if f is assumed to

* *Author to whom correspondence may be addressed.*
E-mail: meyer@stat.colostate.edu

be monotone increasing as well as smooth. Friedman & Tibshirani (1984) smoothed the isotonic regression estimator using a running average. Utreras (1986) provided the characterization and convergence rates for the monotone smoothing spline estimator that minimizes

$$\sum_{i=1}^{n}[y_i - f(x_i)]^2 + \lambda \int_0^1 \left[f^{(q)}(x)\right]^2 \mathrm{d}x,$$

over the monotone nondecreasing functions in C^q. Tantiyaswasdikul & Woodroofe (1994) obtained the solution for $q = 1$. The estimators are difficult to obtain for $q \geq 2$, however, because imposing the constraints at the observations is not sufficient to guarantee the constraints hold everywhere. Additional knots must be placed between observations; a general algorithm for the exact placement is unknown (Wang & Li, 2008). The monotone regression spline was introduced by Ramsay (1988) and extended to convex and other shape restrictions by Meyer (2008). Mammen (1991) proposed a two-step method in which a kernel smoother is applied to the isotonic regression estimator, or the kernel smoother is "isotonized." Wood (1994) proposed a smoothing spline with a sufficient (but not necessary) condition for monotonicity. He & Shi (1998) proposed monotone B-spline smoothing with L_1 optimization, using linear programming to obtain the estimator. Ramsay (1998) considered a smooth estimator of a strictly monotone function. Mammen & Thomas-Agnan (1999) showed that the shape-constrained smoothing spline estimator is equivalent to a two-step procedure and provided convergence rates. Hall & Huang (2001) provided a monotone kernel regression estimator. Mammen et al. (2001) generalized the idea of projecting a smoothed estimator onto a constrained subset of the estimation space.

In this paper a shape-constrained penalized spline estimator is proposed. The estimate is obtained through a single weighted projection of the data onto a polyhedral convex cone. We assume that $f \in C^{p+1}[0,1]$ for some integer $p \geq 1$, and specify a set of knots $0 = t_1 < \cdots < t_k = 1$. A set of $m = k + p - 1$ basis functions $\delta_1(x), \ldots, \delta_m(x)$ is defined so that the functions span the space of smooth piecewise degree-p polynomial spline functions with the given knots. The corresponding basis vectors $\boldsymbol{\delta}_1, \ldots, \boldsymbol{\delta}_m$ where $\boldsymbol{\delta}_{ji} = \delta_j(x_i)$ span an m-dimensional subspace of $\Re^n$. If $\boldsymbol{B}$ is the $n \times m$ matrix whose columns are the $\boldsymbol{\delta}_j$ vectors and $\tilde{\boldsymbol{b}} = (\boldsymbol{B}^t\boldsymbol{B})^{-1}\boldsymbol{B}^t\boldsymbol{y}$, then $\tilde{f}(x) = \sum_{j=1}^{m} \tilde{b}_j \delta_j(x)$ is the unconstrained, unpenalized regression spline estimate of f.

There are many possibilities for the set of basis functions; the B-splines are a standard choice. These are defined so that each basis vector is orthogonal to all but a few of the others, because the basis functions are concentrated around their associated knots, and hence most of the supports are nonoverlapping. These are shown for $k = 5$ in Figure 1(a) for $p = 2$ and (b) for $p = 3$. The curves represent the basis functions and the dots mark the values of the basis vectors for $n = 50$ equally spaced x_i. See De Boor (2001) for a thorough treatment, including formulas for the spline basis functions. Other choices such as the truncated polynomial basis (see Ruppert, Wand, & Carroll, section 3.7) span the same space but the resulting design matrix might be highly collinear. The regression splines are simple, flexible, and parsimonious function estimators. Some large sample theory was derived by Zhou, Shen, & Wolfe (1998), for mild assumptions. They derived the asymptotic bias and variance of the regression splines, the asymptotic normality of the estimates, and showed that if the number of knots grows as $n^{1/(2p+3)}$, the optimal point-wise convergence rate $\tilde{f}(x) - f(x) = O_p(n^{-(p+1)/(2p+3)})$ is attained.

In practice, the splines are sensitive to knot number and placement. Penalized splines are more robust and the user choices are simplified to a single penalty parameter. The P-splines of Eilers & Marx (1996) use "many" knots equally spaced in (0, 1), and penalize the qth order differences

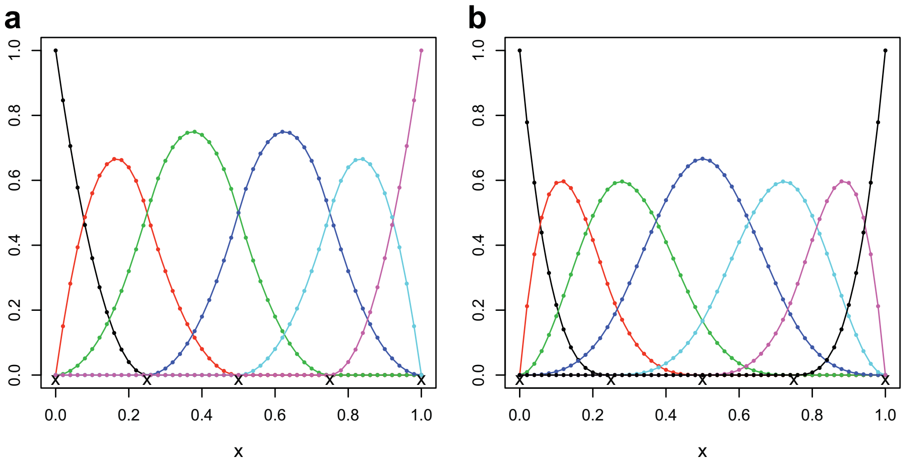

FIGURE 1: B-spline basis functions with $k = 5$ knots (marked with "×") equally spaced in (0, 1) (a) Piecewise quadratic; (b) piecewise cubic. [Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com]

in the sizes of the coefficients of adjacent B-splines. The penalized sum of squares is

$$\sum_{i=1}^{n}\left[y_i - \sum_{j=1}^{m} b_j\delta_j(x_i)\right]^2 + \lambda \sum_{j=q+1}^{m}(\Delta^q b_j)^2, \tag{2}$$

where $\Delta^1 b_j = b_j - b_{j-1}$ and $\Delta^q b_j = \Delta^{q-1}\Delta b_j$ for $q > 1$. The expression to minimize may be written in vector form:

$$\psi(\boldsymbol{b};\boldsymbol{y}) = \boldsymbol{b}^t(\boldsymbol{B}^t\boldsymbol{B} + \lambda\boldsymbol{D}^t\boldsymbol{D})\boldsymbol{b} - 2\boldsymbol{y}^t\boldsymbol{B}\boldsymbol{b}, \tag{3}$$

where $\boldsymbol{D}$ is the qth order difference matrix, and the coefficient estimate is found via a weighted projection onto a linear subspace: $\tilde{\boldsymbol{b}} = (\boldsymbol{B}^t\boldsymbol{B} + \lambda\boldsymbol{D}^t\boldsymbol{D})^{-1}\boldsymbol{B}^t\boldsymbol{y}$. The estimate of $\boldsymbol{\mu}$, where $\mu_i = f(x_i)$, is $\tilde{\boldsymbol{\mu}} = \boldsymbol{B}\tilde{\boldsymbol{b}}$, and the "effective degrees of freedom" (edf) of the model (see Hastie & Tibshirani 1990, chapter 5) is the trace of $\boldsymbol{B}(\boldsymbol{B}^t\boldsymbol{B} + \lambda\boldsymbol{D}^t\boldsymbol{D})^{-1}\boldsymbol{B}^t$. For $\lambda = 0$ this is m, the number of columns of $\boldsymbol{B}$, and as λ gets larger, the edf approaches $\min(p, q-1)$.

Claeskens, Krivobokova, & Opsomer (2009) investigated the asymptotic properties of the penalized splines. They defined the estimator in more generality as the minimizer of

$$\sum_{i=1}^{n}\left[y_i - \sum_{j=1}^{m} b_j\delta_j(x_i)\right]^2 + \lambda\int_0^1\left[\left\{\sum_{j=1}^{m} b_j\delta_j(x)\right\}^{(q)}\right]^2 \mathrm{d}x,$$

and noted that the simplification to (2) does not influence the asymptotic properties of the estimator. They demonstrate that the asymptotic behaviour of the penalized spline can be similar to that of the regression spline if the number of knots grows slowly, and similar to that of the smoothing spline for a faster-growing number of knots. In particular, the regression spline behaviour is seen when the number of knots grows at the rate $n^{1/(2p+3)}$ and $\lambda = O(n^{\gamma})$ where $\gamma \le n^{(p+2-q)/(2p+3)}$, where the degree of the spline is p and the degree of the penalty is q. For smoothing spline behaviour, the number of knots grows at the rate n^{ν}, where $\nu \ge 1/(2q+1)$, and $\lambda = O(n^{1/(2q+1)})$ such that $\lambda n^{2q-1} \to \infty$; then the rate $\hat{f}(x) - f(x) = O_p(n^{-q/(2q+1)})$ is attained. Li & Ruppert

(2008) showed that the asymptotic behaviour of penalized splines is similar to that of kernel estimators, and obtained asymptotic normality.

The penalty parameter may be chosen through generalized cross-validation (GCV) methods (see Ruppert, Wand, & Carroll, 2003, chapter 5), where the GCV for the fit $\tilde{\boldsymbol{\mu}}(\lambda)$ is

$$\mathrm{GCV}(\lambda)=\frac{\sum_{i=1}^{n}[y_i-\tilde{\mu}_i(\lambda)]^2}{(1-\mathrm{edf}/n)^2}. \tag{4}$$

In the next section, the method for estimating the constrained penalized splines is presented, and it is shown that the MSE is, with probability approaching one, less than or equal to that for the unconstrained penalized splines. Hence, the same convergence rate is attained under the same knot choices and model assumptions. In Section 3 the GCV is used to "test" for monotonicity or convexity. This is a natural and effective method for choosing between the constrained and unconstrained models, using a well-established criterion. Extensions to the partial linear model, the generalized regression model, and the varying coefficient model are considered in Section 4. Several examples of applications are given in Section 5, demonstrating the broad utility of the methods.

2. CONSTRAINED PENALIZED SPLINES

For the monotone case, quadratic B-splines are used. Let the $k \times m$ matrix $\boldsymbol{S}$ be defined by $S_{ij}=\delta'_j(t_i)$, the first derivatives of the basis functions at the knots. Then the linear combination $\sum_{j=1}^{m} b_j\delta_j(x)$ is nondecreasing if and only if the coefficient vector is in the set

$$\mathcal{C}=\{\boldsymbol{b}:\boldsymbol{S}\boldsymbol{b}\geq \boldsymbol{0}\}\subseteq \Re^m. \tag{5}$$

For the convex case, we use cubic splines and $S_{ij}=\delta''_j(t_i)$; then the linear combination is convex if and only if $\boldsymbol{S}\boldsymbol{b}\geq\boldsymbol{0}$. Note that $\boldsymbol{S}$ is $k\times(k+1)$ for the monotone case and $k\times(k+2)$ for the convex case, and both are full row-rank.

For degree-p spline functions, the $(p-1)$st derivative is linear, and hence positive if and only if it is nonnegative at the knots. Therefore, if cubic or higher degree splines are used for a monotone function, the constraints are no longer linear. See Meyer (2008) for more details about the cubic monotone spline estimator. Similarly, for splines with convexity constraints, nonlinear constraints would be necessary for fourth or higher degree splines. Of course, lower-degree splines may be used, if the true function is not assumed to have the required smoothness; see assumption (A1) below.

The extension to monotone *and* convex restrictions requires a $(k+1)\times(k+2)$ matrix S, containing the k rows from the convex constraints and an additional row to guarantee a positive slope at the left hand side of the interval. Variations such as decreasing and concave, or sigmoidal with known point of inflection require simple changes to the matrix $\boldsymbol{S}$. The set $\mathcal{C}$ is a convex polyhedral cone, and the problem of minimizing (3) for $\boldsymbol{b}\in\mathcal{C}$ is a quadratic programming problem. A routine such as the `R` function `qpsolve` might be used to find the coefficients of the basis functions.

Let $\boldsymbol{L}\boldsymbol{L}^t$ be the Cholesky decomposition of $\boldsymbol{B}^t\boldsymbol{B}+\lambda\boldsymbol{D}^t\boldsymbol{D}$, then define $\boldsymbol{\phi}=\boldsymbol{L}^t\boldsymbol{b}$ and $\boldsymbol{z}=\boldsymbol{L}^{-1}\boldsymbol{B}^t\boldsymbol{y}$, so that (3) can be written as $\|\ \boldsymbol{\phi}-\boldsymbol{z}\ \|^2=\sum_{i=1}^{n}(\phi_i-z_i)^2$ and the constraints are $\boldsymbol{A}\boldsymbol{\phi}\geq\boldsymbol{0}$, where $\boldsymbol{A}=\boldsymbol{S}(\boldsymbol{L}^t)^{-1}$. The minimizer $\hat{\boldsymbol{\phi}}$ is the projection of $\boldsymbol{z}$ onto the cone $\tilde{\mathcal{C}}=\{\boldsymbol{\phi}:\boldsymbol{A}\boldsymbol{\phi}\geq 0\}\subseteq\Re^m$, and subsequently $\hat{\boldsymbol{b}}=(\boldsymbol{L}^t)^{-1}\hat{\boldsymbol{\phi}}$. Some necessary background for cone projection is given next; more details and proofs of the following claims may be found in Silvapulle & Sen (2005, chapter 3).

To characterize the constrained solution and estimate its effective degrees of freedom, the dimension of the "face" of the cone containing the solution is determined as follows, for the case where $\boldsymbol{A}$ is $k \times m$ and full row-rank. First, the "edges" or "generators" of $\tilde{\mathcal{C}}$ are found. Let $\boldsymbol{e}_1, \ldots, \boldsymbol{e}_{m-k}$ span the null space of $\boldsymbol{A}$, and let $\tilde{\boldsymbol{A}}$ be the square, nonsingular matrix where the first k rows are the rows of $\boldsymbol{A}$ and the last rows are the $\boldsymbol{e}$ vectors. Then the first k columns of $\tilde{\boldsymbol{A}}^{-1}$ are the edges $\boldsymbol{\gamma}_1, \ldots, \boldsymbol{\gamma}_k$ of the cone, and the cone can be written as

$$\tilde{\mathcal{C}} = \left\{ \boldsymbol{\phi} : \boldsymbol{\phi} = \sum_{j=1}^{m-k} a_j \boldsymbol{e}_j + \sum_{i=1}^{k} c_j \boldsymbol{\gamma}_j, \quad \text{for} \quad c_j \geq 0, \quad j = 1, \ldots, k \right\}. \tag{6}$$

The projection of z onto $\tilde{\mathcal{C}}$ lands on a *face* of the cone, where the 2^k faces are indexed by the collection of sets $J \subseteq \{1, \ldots, k\}$, and are defined by

$$\mathcal{F}_J = \left\{ \boldsymbol{\phi} : \boldsymbol{\phi} = \sum_{j=1}^{m-k} a_j \boldsymbol{e}_j + \sum_{j \in J} c_j \boldsymbol{\gamma}_j, \quad \text{for} \quad c_j > 0, \quad j \in J \right\}.$$

Note that the interior of the cone $\tilde{\mathcal{C}}$ is a face with $J = \{1, \ldots, k\}$, and the linear space spanned by the $\boldsymbol{e}_j$ vectors is a face with $J = \emptyset$. Further, the faces partition $\tilde{\mathcal{C}}$. The cone projection algorithm determines the face $\mathcal{F}_J$ on which the projection falls; then the solution coincides with the ordinary least-squares projection onto the linear space spanned by the vectors $\boldsymbol{e}_1, \ldots, \boldsymbol{e}_{m-k}$ and the edges $\boldsymbol{\gamma}_j$ for $j \in J$.

To compute the edf associated with the constrained penalized fit, let $\boldsymbol{\Delta}_J$ be the matrix whose columns are the $\boldsymbol{e}$ vectors and the $\boldsymbol{\gamma}_j$ vectors for $j \in J$. Then the constrained estimate of $\boldsymbol{\mu}$ is $\hat{\boldsymbol{\mu}} = \boldsymbol{B}(\boldsymbol{L}^t)^{-1}\boldsymbol{\Delta}_J(\boldsymbol{\Delta}'_J\boldsymbol{\Delta}_J)^{-1}\boldsymbol{\Delta}'_J\boldsymbol{L}^{-1}\boldsymbol{B}^t\boldsymbol{y} =: \boldsymbol{P}_J\boldsymbol{y}$, and hence the edf is the trace of $\boldsymbol{P}_J$. Note that if $J = \{1, \ldots, k\}$, that is, all edges are used, then $\boldsymbol{\Delta}_J(\boldsymbol{\Delta}'_J\boldsymbol{\Delta}_J)^{-1}\boldsymbol{\Delta}'_J$ is the identity matrix, and the constrained edf is identical to the unconstrained edf. This happens when the unconstrained spline satisfies the constraints; otherwise, the edf for the constrained version is smaller. The edf for the constrained version is a random quantity with $k + 1$ possible values, the largest of which is that of the unconstrained version.

A cross-validation scheme for the constrained version uses

$$\text{CV}(\lambda) = \sum_{i=1}^{n} \left[y_i - \hat{f}_{(-i),\lambda}(x_i) \right]^2,$$

where $\hat{f}_{(-i),\lambda}$ is the estimated regression function using the edge vectors with the ith observation removed (the knots stay the same). If the index set J of edges used by this projection is the same as that using all the data, then

$$y_i - \hat{f}_{(-i),\lambda}(x_i) = \frac{y_i - \hat{f}_\lambda(x_i)}{1 - h_i},$$

where h_i is the ith diagonal element of the projection matrix for the face indexed by J and $\hat{f}_\lambda$ is the estimated regression function using all the data. Replacing h_i with the average over i gives the GCV in the form (4). Although the set J may vary when an observation is ignored, particularly for points of high leverage, the approximation is useful for the selection of λ and the method for validation of the shape assumptions that is given in the next section. Simulations using regression functions and sample sizes described at the end of this section suggest that the edges used to

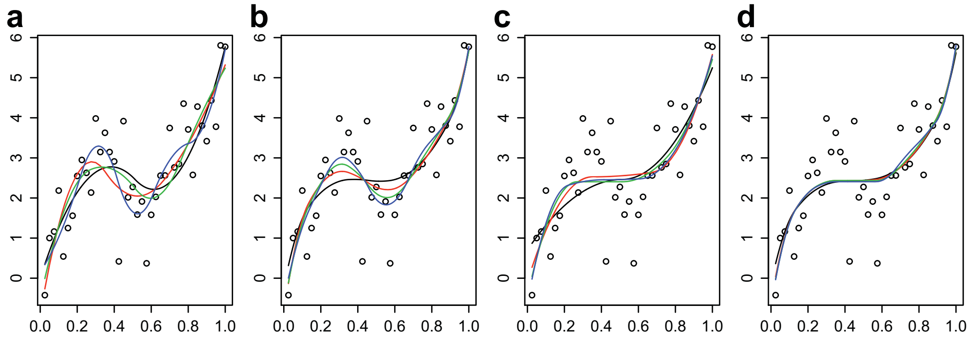

FIGURE 2: Spline fits to simulated data, using 4, 5, 6, and 7 edf. (a) Unconstrained, unpenalized (b) unconstrained, penalized, (c) constrained, unpenalized, and (d) constrained, penalized. [Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com]

form $\hat{f}$ are identical to those for $\hat{f}_{(-i)}$ for about 90% of values of i, with the others sets of edges being different by only one or two additions or subtractions.

The penalized regression spline fits to (1) where $cov(\boldsymbol{\epsilon}) = A$ for known positive-definite A is a simple extension of either the constrained or unconstrained model. The expression to be minimized is

$$\psi(\boldsymbol{b}; \boldsymbol{y}) = \boldsymbol{b}^t(\boldsymbol{B}^t\boldsymbol{A}^{-1}\boldsymbol{B} + \lambda\boldsymbol{D}^t\boldsymbol{D})\boldsymbol{b} - 2\boldsymbol{y}^t\boldsymbol{A}^{-1}\boldsymbol{B}\boldsymbol{b},$$

which can be accomplished as above with the Cholesky decomposition $\boldsymbol{L}\boldsymbol{L}^t = \boldsymbol{B}^t\boldsymbol{A}^{-1}\boldsymbol{B} + \lambda\boldsymbol{D}^t\boldsymbol{D}$ and $z = \boldsymbol{L}^{-1}\boldsymbol{B}\boldsymbol{A}^{-1}\boldsymbol{y}$. This extension is important for correlated-errors models, for heteroskedastic data, and for the iteratively re-weighted least-squares methods such as for generalized regression.

The estimators are more robust to knot choices when shape restrictions are employed, especially if convex or concave constraints are appropriate. The restriction to the cone allows for all the flexibility within the constraints, without allowing the "wiggling" associated with over-fitting in unconstrained nonparametric function estimation. To illustrate, a data set of size $n = 50$ is used to compare quadratic regression spline fits in Figure 2. In plot (a) the unconstrained, unpenalized spline fits with 3, 4, 5, and 6 equally spaced knots (having edf = 4, 5, 6, and 7) are the most variable. The penalized spline fits, each with 10 knots placed evenly in the span of x-values, are shown in plot (b), with the choices of penalty parameter that coincide with 4, 5, 6, and 7 edf. Next, suppose that the relationship must be monotone increasing. In plot (c) we see the unpenalized, constrained splines, again with 3, 4, 5, and 6 knots, and finally, in plot (d) the penalized constrained splines (with the same penalty parameters as in (b)) are seen to vary the least with edf, that is, are most robust to user choices. Quadratic splines are allowed inflection points only at the knots; the penalized version is more robust because the larger number of knots allows the data to choose where to bend, even at low edf.

To investigate convergence rates, the mean squared error for the constrained penalized spline is compared to that for the unconstrained penalized spline. The assumptions are

(A1) The regression function f is in $C^{p+1}[0, 1]$,

(A2) the design points $x_1, \ldots, x_n$ follow a density $g(x) > 0$ on $[0, 1]$, and

(A3) the knots $t_1, \ldots, t_k$ have "bounded mesh ratio," that is, the ratio of the largest inter-knot interval to the smallest is bounded.

and *either*

(B1) The number of knots is $k \sim C_1 n^{1/(2p+3)}$, for a constant C_1, and $\lambda = O(n^{\gamma})$ with $\gamma \le (p + 2 - q)/(2p + 3)$, or

(B2) $k \sim C_2 n^{\nu}$, for a constant C_2, where $\nu \ge 1/(2q + 1)$ and $\lambda = O(n^{1/(2q+1)})$ such that $\lambda n^{2q-1} \to \infty$.

The Kuhn-Tucker conditions (see Silvapulle & Sen (2005), Appendix 1) for the projection $\hat{\boldsymbol{\phi}}$ minimizing $\| \boldsymbol{z} - \boldsymbol{\phi} \|^2$ over $\boldsymbol{\phi} \in \tilde{\mathcal{C}}$ are

$$(\boldsymbol{z} - \hat{\boldsymbol{\phi}})^t \hat{\boldsymbol{\phi}} = 0 \quad \text{and} \quad (\boldsymbol{z} - \hat{\boldsymbol{\phi}})^t \boldsymbol{\phi} \le \boldsymbol{0}, \quad \text{for all} \quad \boldsymbol{\phi} \in \tilde{\mathcal{C}}.$$

These are easily seen to be equivalent to

$$(\boldsymbol{y} - \hat{\boldsymbol{\mu}})^t \hat{\boldsymbol{\mu}} = \lambda \hat{\boldsymbol{b}}^t \boldsymbol{D}^t \boldsymbol{D} \hat{\boldsymbol{b}} \quad \text{and} \quad (\boldsymbol{y} - \hat{\boldsymbol{\mu}})^t \boldsymbol{\mu} \le \lambda \hat{\boldsymbol{b}}^t \boldsymbol{D}^t \boldsymbol{D} \boldsymbol{b}, \quad \text{for all} \quad \boldsymbol{\mu} = \boldsymbol{B}\boldsymbol{b}, \quad \boldsymbol{b} \in \mathcal{C}.$$

For the unconstrained estimator $\tilde{\boldsymbol{b}}$ and $\tilde{\boldsymbol{\mu}} = \boldsymbol{B}\tilde{\boldsymbol{b}}$, we have $(\boldsymbol{y} - \tilde{\boldsymbol{\mu}})^t \boldsymbol{\mu} = \lambda \tilde{\boldsymbol{b}} \boldsymbol{D}^t \boldsymbol{D} \boldsymbol{b}$, for all $\boldsymbol{\mu} = \boldsymbol{B}\boldsymbol{b}$, $\boldsymbol{b} \in \Re^m$. Let $\boldsymbol{b}_0$ be the minimizer of $\psi(\boldsymbol{b}; \boldsymbol{\mu})$ for ψ defined in (3), so that $\boldsymbol{\mu}_0 = \boldsymbol{B}\boldsymbol{b}_0$ is the penalized spline fit to the true function values, and suppose $\boldsymbol{S}\boldsymbol{b}_0 \ge \boldsymbol{0}$, so that the shape restrictions hold. Then

$$\begin{aligned}
\| \tilde{\boldsymbol{\mu}} - \boldsymbol{\mu}_0 \|^2 &= \| \tilde{\boldsymbol{\mu}} - \hat{\boldsymbol{\mu}} \|^2 + \| \hat{\boldsymbol{\mu}} - \boldsymbol{\mu}_0 \|^2 + 2(\tilde{\boldsymbol{\mu}} - \hat{\boldsymbol{\mu}})^t (\hat{\boldsymbol{\mu}} - \boldsymbol{\mu}_0) \\
&= \| \tilde{\boldsymbol{\mu}} - \hat{\boldsymbol{\mu}} \|^2 + \| \hat{\boldsymbol{\mu}} - \boldsymbol{\mu}_0 \|^2 - 2(\boldsymbol{y} - \tilde{\boldsymbol{\mu}})^t (\hat{\boldsymbol{\mu}} - \boldsymbol{\mu}_0) + 2(\boldsymbol{y} - \hat{\boldsymbol{\mu}})^t (\hat{\boldsymbol{\mu}} - \boldsymbol{\mu}_0) \\
&\ge \| \tilde{\boldsymbol{\mu}} - \hat{\boldsymbol{\mu}} \|^2 + \| \hat{\boldsymbol{\mu}} - \boldsymbol{\mu}_0 \|^2 + 2\lambda(\hat{\boldsymbol{b}} - \tilde{\boldsymbol{b}})^t \boldsymbol{D}^t \boldsymbol{D} (\hat{\boldsymbol{b}} - \boldsymbol{b}_0),
\end{aligned}$$

so

$$\| \tilde{\boldsymbol{\mu}} - \boldsymbol{\mu}_0 \|^2 - \| \hat{\boldsymbol{\mu}} - \boldsymbol{\mu}_0 \|^2 \ge \| \tilde{\boldsymbol{\mu}} - \hat{\boldsymbol{\mu}} \|^2 + 2\lambda(\hat{\boldsymbol{b}} - \tilde{\boldsymbol{b}})^t \boldsymbol{D}^t \boldsymbol{D} (\hat{\boldsymbol{b}} - \boldsymbol{b}_0).$$

The first term on the right is always positive, and the second term can be written as

$$\begin{aligned}
&2\lambda(\hat{\boldsymbol{b}} - \tilde{\boldsymbol{b}})^t \boldsymbol{D}^t \boldsymbol{D} (\hat{\boldsymbol{b}} - \tilde{\boldsymbol{b}}) + 2\lambda(\hat{\boldsymbol{b}} - \tilde{\boldsymbol{b}})^t \boldsymbol{D}^t \boldsymbol{D} (\tilde{\boldsymbol{b}} - \boldsymbol{b}_0) \\
&\quad \ge 2\lambda \, \| \boldsymbol{D}(\hat{\boldsymbol{b}} - \tilde{\boldsymbol{b}}) \| \left[\| \boldsymbol{D}(\hat{\boldsymbol{b}} - \tilde{\boldsymbol{b}}) \| - \| \boldsymbol{D}(\tilde{\boldsymbol{b}} - \boldsymbol{b}_0) \| \right].
\end{aligned}$$

Recall that D is a $q + 1$ banded matrix, and the values of the elements are constant as n grows. Suppose that $\| \tilde{\boldsymbol{b}} - \boldsymbol{b}_0 \|^2 / k = O_p(n^{-2r})$ for some $r > 0$, so there is a constant C such that $\| \boldsymbol{D}(\tilde{\boldsymbol{b}} - \boldsymbol{b}_0) \| \le C k^{1/2} n^{-r}$ with probability approaching one as n grows. If $\| \boldsymbol{D}(\tilde{\boldsymbol{b}} - \hat{\boldsymbol{b}}) \| > C k^{1/2} n^{-r}$ then the term in the brackets is positive and $\| \tilde{\boldsymbol{\mu}} - \boldsymbol{\mu}_0 \|^2 \ge \| \hat{\boldsymbol{\mu}} - \boldsymbol{\mu}_0 \|^2$. Otherwise, $\| \boldsymbol{D}(\tilde{\boldsymbol{b}} - \hat{\boldsymbol{b}}) \| \le C k^{1/2} n^{-r}$, and the constrained penalized spline attains the convergence rate of unconstrained spline, which is the optimal rate under (A1)–(A3) and either (B1) or (B2).

To compare the constrained and unconstrained fits for small samples, simulations were conducted for two choices of regression function, edf, and sample size. The x-values are equally spaced in (0, 1) and unit model variance is used. The function $f(x) = 1.5(2x - 1)^3$ is flat in the middle and steep on either end, while $f(x) = \log(2x + 0.1)$ is increasing rapidly at the left end of the interval, and flattens out at the right end. For the latter, monotone *and* concave constraints were also considered. The penalty parameters for the $n = 50$ case are chosen so that the edf values for the unconstrained estimator are 4 and 7, while the values for $n = 100$ are so that the edf is 5 and 8. In addition, we compute the constrained and unconstrained fits using the minimum-GCV value of the penalty parameter, where the range for $n = 50$ is so that the unconstrained edf varies from 3.5 to 8 by half-steps, and for $n = 100$, the range is 4.0–8.5. The minimum-GCV value of the penalty

TABLE 1: SASEL for constrained and unconstrained penalized splines. The model variance is unity, the x values are equally spaced in (0, 1), and $N = 10{,}000$ datasets per cell.

			$f(x) = 1.5(2x-1)^3$		$f(x) = \log(2x-0.1)$		
n	k	edf	Monotone	Unconstr	Mon-conc	Monotone	Unconstr
50	10	4	0.248	0.259	0.217	0.236	0.252
50	10	7	0.279	0.338	0.237	0.279	0.338
50	10	GCV	0.272	0.294	0.226	0.254	0.270
100	15	5	0.188	0.201	0.169	0.187	0.201
100	15	8	0.214	0.256	0.179	0.216	0.256
100	15	GCV	0.199	0.210	0.173	0.193	0.200

parameter is typically smaller for the constrained case, because the constrained estimator allows for a high degree of flexibility without introducing the "wiggling" that is typically associated with over-fitting in nonparametric regression. In particular, with the constrained estimator, one can decrease the penalty parameter without (necessarily) increasing the effective degrees of freedom of the fit, because the face of the cone on which the projection lands may be of the same small dimension as for the fit with larger penalty. The computations here and in the rest of the paper use $q = 3$, so that as λ increases, the fit tends toward a cubic polynomial. The square root of the average squared error loss (SASEL) for the rth data set was computed as

$$\mathrm{SASEL}_r = \left[\frac{1}{n}\sum_{i=1}^{n}(\hat{\mu}_{ri} - \mu_i)^2\right]^{1/2},$$

and the average over 10,000 data sets is reported in Table 1. In each case, the more stringent constraints provide smaller SASEL, with bigger improvements shown for larger edf.

The monotone penalized spline is also compared with the monotone B-spline of He & Shi (1998). That method uses quadratic B-splines, L_1 optimization, and linear programming; the number of knots was chosen so that the number of basis functions is equivalent to the edf given for the penalized spline. The regression function is $f = a(2x-1)^3$, with $a = 2$ for the "weak signal" case and $a = 5$ for the "strong signal" case. Because the L_1 optimization was used for the competing estimator, the absolute error loss

$$\mathrm{AEL}_r = \frac{1}{n}\sum_{i=1}^{n}|\hat{\mu}_{ri} - \mu_i|$$

was also computed for each data set. Further, the double-exponential errors were used in addition to normal errors. Table 2 shows the expected results that the constrained penalized spline (CPS) has smaller average SASEL and AEL for normal errors, and the He and Shi (HS) estimator tends to perform better for double-exponential errors, except for the smaller sample size with stronger signal. In general, the percentage improvement for the CPS over the HS for normal errors is substantially larger than the percent improvement for HS over the CPS for double-exponential errors, but the HS method could be considered to be more robust to outliers.

The HS method was coded using the R function lpcdd for the linear programming; there are $m + 2n$ variables with $m - 1 + 3n$ linear equality and inequality constraints. Although the linear programming might seem to be a simpler method than the quadratic programming for

TABLE 2: Average SASEL and AEL for constrained penalized splines (CPS) compared with the monotone B-splines of He and Shi (HS). The regression function is $a(2x-1)^3$ with $a=2$ for the "weak" signal and $a=5$ for the "strong," with two error distributions. The x values are equally spaced in (0, 1), and $N=10{,}000$ datasets per cell.

			Mean SASEL				Mean AEL			
			N(0,1)		dexp(1)		N(0,1)		dexp(1)	
Signal	n	edf	CPS	HS	CPS	HS	CPS	HS	CPS	HS
weak	20	5	0.389	0.465	0.506	0.504	0.311	0.375	0.407	0.388
weak	40	6	0.305	0.370	0.403	0.371	0.241	0.298	0.319	0.288
strong	20	5	0.436	0.538	0.577	0.595	0.348	0.432	0.457	0.456
strong	40	6	0.336	0.424	0.455	0.441	0.264	0.342	0.358	0.342

the penalized spline, the much larger dimension required for the linear programming version is such that the HS method took, on average, about four times longer to run than the CPS. Further, for the larger sample size of $n=40$, the routine did not always converge, and increasing the sample size or the number of knots caused the method to fail more often.

3. CHECK FOR VALIDITY OF SHAPE ASSUMPTIONS

The shape assumptions often fall into the category of *a priori* knowledge, but occasionally the research question might concern the shape. The classic formal hypothesis test of monotone versus nonmonotone regression function is presented in Robertson, Wright, & Dykstra (1988), which tests $H_0 : f$ is monotone versus $H_a : f$ is not monotone, using unsmoothed monotone regression. Their test statistic has the distribution of a mixture of beta random variables, if the true function is constant, but otherwise the test is biased. Bowman, Jones, & Gijbels (1998) formulate a test using kernel regression, where they define the "critical" bandwidth to be the smallest for which the fit is monotone. Using the errors from this fit, a bootstrap method is used for the p-value. The test proposed by Ghosal, Sen, & van der Vaart (2000) uses a locally-weighted Kendall's tau statistic to determine if there any decreasing portions of the regression function. Hall & Heckman (2000) compute a "running derivative" by fitting lines over portions of the data, and their test statistic is the minimum slope. Juditsky & Nemirovski (2002) formulate a general problem of testing whether the signal in a Gaussian random process is contained in a convex cone. Wang & Meyer (2011) propose a formal hypothesis test for monotone versus nonmonotone regression function using constrained and unconstrained splines, and a simulated distribution of a test statistic.

A simple model-selection method is proposed here to check for the validity of the shape assumptions, that is based on the generalized cross-validation (GCV) choices of the penalty parameters. We choose a set of possible penalty parameters, and compute the constrained and unconstrained penalized spline fits and the GCV for each penalty value. If the minimum GCV value for the constrained splines is smaller than that for the unconstrained splines, that is evidence that the shape restrictions hold. Conversely, if the smallest GCV value is for an unconstrained spline, then we conclude that there is evidence against the shape restriction. In the case of a tie, we decide in favour of the shape restrictions, because a tie often indicates that the unconstrained fit satisfies the constraints.

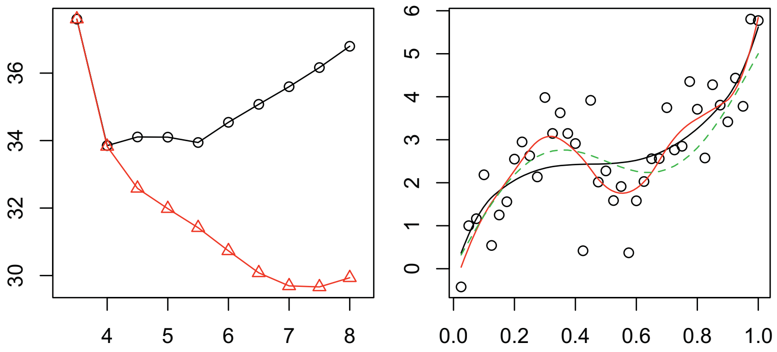

FIGURE 3: (a) The GCV values for the data set of Figure 2, where the circles are for the constrained version and the triangles are for the unconstrained splines. The method chooses the unconstrained version with edf $= 7.5$, shown in plot (b) along with the "best" constrained fit. The dashed curve is the true regression function. [Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com]

The procedure is illustrated in Figure 3, using a data set with $n = 50$ generated from the regression function shown as the dash curve in plot (b). Ten penalty parameters were used, chosen so that the edf values for the unconstrained fit are $3.5, 4.0, \ldots, 8.0$, and 10 knots equally spaced over the range of the MRI values. The GCV values are shown in plot (a), where the values for the constrained fit are shown as circles and those for the unconstrained fit are triangles. The winner is the unconstrained spline with penalty parameter chosen to coincide with edf $= 7.5$. The minimum-GCV constrained and unconstrained fits are shown in plot (b) along with the true regression function.

Although this method is not a formal hypothesis test, we can show that the monotone versus nonmonotone procedure has nice large-sample properties if the number of knots does not grow too quickly, for $p = 2$ and $q = 3$. The following theorem is proved in the appendix.

Theorem 1. *Assume $(A1)$–$(A3)$ and either $(B1)$ or $(B2)$. If the true regression function is strictly increasing, so that $f'(x) \geq \xi > 0$ for $x \in [0, 1]$, and $k = O(n^{2/7-\eta})$ for some $\eta \in (0, 1/7]$, then the probability that the method chooses the constrained version goes to one as the sample size increases. On the other hand, if the true regression function is decreasing over an interval, so that $f'(x) \leq -\xi < 0$ for $x \in [a, b]$ where $0 \leq a < b \leq 1$, the probability that the method chooses the unconstrained version goes to one.*

TABLE 3: Simulations for determination of monotonicity, where the entries are percent of simulated data sets for which the method makes the correct determination. The model variance is unity, $N = 10{,}000$ datasets per cell.

	Proposed GCV method			Wang & Meyer test		
f	$n = 50$	$n = 100$	$n = 200$	$n = 50$	$n = 100$	$n = 200$
$10f_{0.15}$	95.2	96.2	97.4	95.6	96.2	96.2
$10f_{0.25}$	20.3	33.1	54.1	15.1	25.2	47.4
$10f_{0.45}$	94.0	99.9	100	59.1	83.8	93.1
$3\mu_4$	34.1	54.7	89.4	25.3	30.7	60.6

The small-sample performance of the procedure is illustrated and compared to other methods through simulations. The percents of times the correct choice is made are provided in Table 3, for four choices of regression function and three sample sizes, where 10,000 data sets were generated for each combination, and model variance is $\sigma^2 = 1$. For the smallest sample size $n = 50$, the 10 penalty parameters are chosen so that the edf for the unconstrained fit ranges from 3.5 to 8. For $n = 100$, the range is 4 to 8.5, and for $n = 200$, the range is 4.5 to 9. The regression functions are chosen from the Bowman, Jones, & Gijbels (1998) paper and the Ghosal, Sen, & van der Vaart (2000) paper. For the former,

$$f_a(x) = 1 + x - a \exp\left\{ -50(x-0.5)^2 \right\},$$

for $a = 0.15, 0.25$, and 0.45. Note that $f_{0.15}$ is increasing over the range [0, 1], $f_{0.25}$ has a mild dip, and $f_{0.45}$ has a more pronounced dip. That paper uses $\sigma = 0.1$ for their simulations; we use $\sigma = 1$ and $10 f_a$. Using a test size of $\alpha = 0.05$, their rates of correct choice for $a = 0.15$ are 99.2% and 99.6%, for $n = 50$ and $n = 100$, respectively. For $a = 0.25$, their rates are 10.0% and 17.4% correct choices, and for $a = 0.45$, their rates are 54.4% and 87.4%. Ghosal, Sen, & van der Vaart (2000) defined

$$\mu_4 = \begin{cases} 10(x-0.5)^3 - \exp\{-100(x-0.25)^2\} & \text{for } x \in [0, 0.5) \\ 0.1(x-0.5) - \exp\{-100(x-0.25)^2\}, & \text{for } x \in [0.5, 1], \end{cases}$$

which has a very sharp dip followed by a flat section. In that paper they use $\sigma = 0.1$; here we use $\sigma = 1$ and $3\mu_4$. The signal-to-noise ratio for our simulations is over three times smaller, but their rates of correct choices are lower: 2.6, 8.0, and 75.7, for $n = 50$, 100, and 200.

The method proposed here consistently has higher "power" than the competitors, for examples that were chosen for those methods, and also consistently "wins" compared to the Wang and Meyer test results shown in Table 3. The advantage is that the new method does not require the choice of a single penalty parameter or bandwidth, but allows the flexibility of the fit to be chosen by the data. Therefore, the sharp dip of μ_4 and the more regular variation of f_a are fit equally well. The smaller rates of correct choices for $10 f_{0.25}$ reflect the fact that the regression function is increasing over most of the range, with a very small dip in the middle that is hard to detect. This is analogous to the formal hypothesis testing situation where there is lower power when the truth is "close to" the null hypothesis.

Alternatively, the criterion $\mathrm{AIC} = n \log(\mathrm{SSE}/n) + 2\mathrm{edf}$ or $\mathrm{BIC} = n \log(\mathrm{SSE}/n) + \mathrm{edf} \log(n)$ may be used for model selection. The AIC will choose the unconstrained model more often than the GCV, and the BIC will choose the constrained model more often.

4. EXTENSIONS

The partial linear model is a simple extension. An arbitrary number of covariates to be modeled parametrically can be included in the model, so that

$$y_i = \boldsymbol{z}_i' \boldsymbol{a} + f(x_i) + \sigma \epsilon_i, \tag{7}$$

for $z_i \in \Re^r$, where again the ϵ_i are uncorrelated random variables with mean zero and unit variance. Let the $n \times r$ matrix $\boldsymbol{Z}$ be the design matrix for the covariates, and suppose that its columns and the $\boldsymbol{x}$ vector form a linearly independent set, and further suppose that the $\mathbf{1}$ vector is not in the column space of $\boldsymbol{Z}$. Estimates of $E(\boldsymbol{y})$ will be of the form $\boldsymbol{Za} + \boldsymbol{Bb} = \boldsymbol{Wc}$, where the first r columns of $\boldsymbol{W}$ are the covariate vectors and the last m are the B-spline basis vectors. The constraint $\boldsymbol{Sc} \geq \mathbf{0}$ is again imposed, where now the first r columns of the $k \times (r+m)$ matrix $\boldsymbol{S}$ contain zeros. The cone projection proceeds as described in Section 2, except now the null space of $\boldsymbol{A}$ is of dimension $m + r - k$. The ability to add covariates to a model is important if there is a possible confounding variable or to account for another source of variation.

For the generalized regression problem, the responses y_i are independent observations from a distribution written in the form of an exponential family:

$$f(y_i) = \exp\left\{[y_i\theta_i - b(\theta_i)]/\phi^2 - c(y_i, \tau)\right\},$$

where the specifications of b and c determine the sub-family of models. Common examples are $b(\theta) = \log(1 + e^\theta)$ for the Bernoulli and $b(\theta) = \exp(\theta)$ for the Poisson model. The vector $\boldsymbol{\mu}$ is defined as $E(\boldsymbol{y})$, and $\mu_i = b'(\theta_i)$. The variance of y_i is $b''(\theta_i)\tau$, and the variance function is written in terms of the mean as $\text{var}(y_i) = V(\mu_i)$. The mean vector is related to a design matrix of predictor variables through a link function $g(\mu_i) = \eta_i$; for the generalized *linear* model, $\eta_i = \boldsymbol{x}_i'\boldsymbol{\beta}$, where $\boldsymbol{x}_i$ is the ith row of the design matrix and $\boldsymbol{\beta}$ is the parameter vector to be estimated. Here, we let $\eta_i = f(x_i) + \boldsymbol{z}_i'\boldsymbol{\alpha}$ where f has known shape and degree of smoothness, and $\boldsymbol{z}_i$ is an r-vector of co-variates. It is often natural to impose monotonicity restrictions on the mean function; for example, the probability of cancer in a lab rat might be increasing with amount of carcinogen administered, or the expected count of water bird nests at a lake might be known to be decreasing with amount of boat traffic. If the link function is one-to-one, this is equivalent to constraining the function component f to be monotone.

The algorithm involves iteratively re-weighted quadratic programming problems, following 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 a "data vector" $\boldsymbol{u}^k$ where

$$u_i^k = \eta_i^k + \left(y_i - \mu_i^k\right)\left(\frac{\mathrm{d}\eta}{\mathrm{d}\mu}\right)_{ki},$$

where $\mu_i^k = g^{-1}(\eta_i^k)$ and the derivative of the link function is evaluated at μ_i^k. Weights w_i are defined by $1/w_i^k = (\mathrm{d}\eta/\mathrm{d}\mu)_k^2 V_k$. The estimate $\boldsymbol{\eta}^{k+1}$ is obtained as the weighted penalized constrained regression spline with data $\boldsymbol{u}^k$; that is, the weighted projection of $\boldsymbol{u}^k$ onto the cone defined by the study design. The algorithm is guaranteed to converge to the penalized maximum likelihood estimator; the proof is similar to the convergence proof in Meyer & Woodroofe (2004). The GCV may be used as check for the validity of shape assumptions, where minus two times the log likelihood is used in place of the sum of squared residuals, and the edf for the constrained version is computed according to the penalty parameter and the dimension of the cone face containing the solution.

The varying coefficient linear regression model was introduced by Hastie & Tibshirani (1993). A simple version is

$$y_i = \beta_0(u_i) + \beta_1(u_i)x_i + \epsilon_i,$$

where the functions β_0 and β_1 are assumed to be smooth. Here we impose a shape such as monotonicity or convexity on either or both coefficient functions. We define basis functions $\delta_j(u)$, $j = 1, \ldots, m$, over the range of u values, and create the matrix $\boldsymbol{B}$ with columns $\boldsymbol{\delta}_1, \ldots, \boldsymbol{\delta}_m, \boldsymbol{\delta}_1\boldsymbol{x}, \ldots, \boldsymbol{\delta}_m\boldsymbol{x}$. The $2m \times 2m$ penalty matrix $\boldsymbol{D}$ is "block" diagonal with identical $m \times m$ qth-order difference matrices. The constraint matrix $\boldsymbol{S}$ is constructed according to the desired shapes for β_0 and β_1. Then the estimation and inference methods of the previous sections can be implemented.

5. EXAMPLES AND DISCUSSION

The British coal mine accident data are used to illustrate the method of validation of the monotonicity assumption for count data. These data were used in the Eilers & Marx (1996) paper on penalized splines. Counts of serious mine accidents per year from 1850 to 1961 are shown in Figure 4(a). Suppose a historian believes that the risk of accident is decreasing over the time span, and the spike in the 1930s can be explained by random chance. Is there evidence in the data to contradict this belief? That is, does the unconstrained spline explain significantly more of the variation than the constrained spline? Constrained and unconstrained penalized quadratic spline fits were computed for penalty parameters chosen so that the effective degrees of freedom for the unconstrained spline increases from 3.5 to 7.5 by half-steps. The GCV values are shown in plot (b); the minimum value is attained by the constrained version, so we find no evidence that the historian is incorrect.

For an example of the varying coefficient model, consider data from a lawsuit from the early 1990s. The United Auto Workers sought pay equity adjustments for civil service jobs in the state of Michigan (Killingsworth, 2002), claiming that women were systematically underpaid by the state based on "comparable worth" arguments. Both parties agreed to the appointment of a committee to assign points to civil service job categories, meant to represent "worth." The data set then consists of three variables for each of 176 categories in Michigan civil service jobs: the points assigned by the committee, the percent of women in the job category, and the maximum wages (in dollars per hour) for the category. The data are shown in Figure 5(a) and (b), with wages plotted against each of the two predictors. The wages appear to be a decreasing function of percent of women employees, but perhaps the job categories with more women have, on average, lower points. As seen in plot (b), wages rise sharply with increasing points, with a curvature that could be accounted for if the slope and intercept of the linear relationship depends on the percent women. The model is $y_i = \beta_0(u_i) + \beta_1(u_i)x_i + \epsilon_i$, where y_i is wages for the ith job category, x_i is the points, and u_i is percent women. Interest is in whether the slope function is decreasing with percent of women in the job category. The function β_0 is unconstrained.

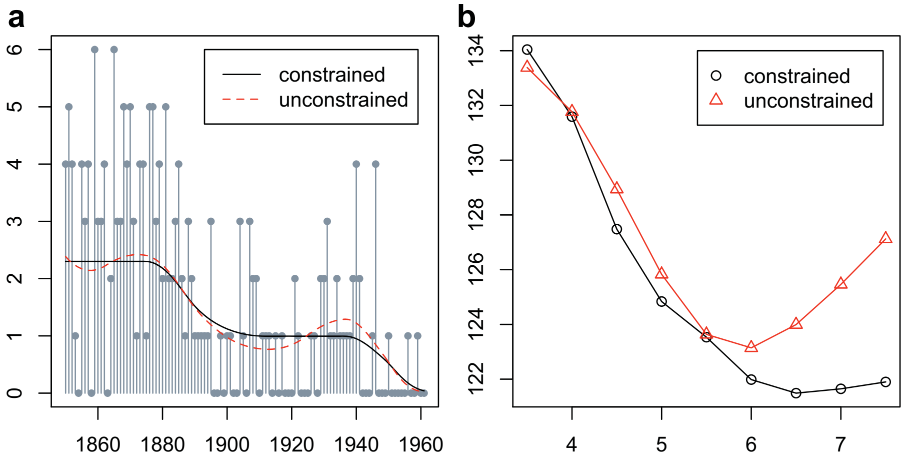

FIGURE 4: Analysis of British coal mine accident data. (a) Counts of accidents per year, along with estimated mean functions. The dashed curve is the penalized quadratic spline with third-order penalties, with penalty parameter chosen so that the effective degrees of freedom of the model is 6.5. The solid curve is the constrained version with the same penalty parameter. (b) Generalized cross-validation values for nine values of penalty parameter, chosen so that the unconstrained spline fits have edf ranging from 3.5 to 7.5. [Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com]

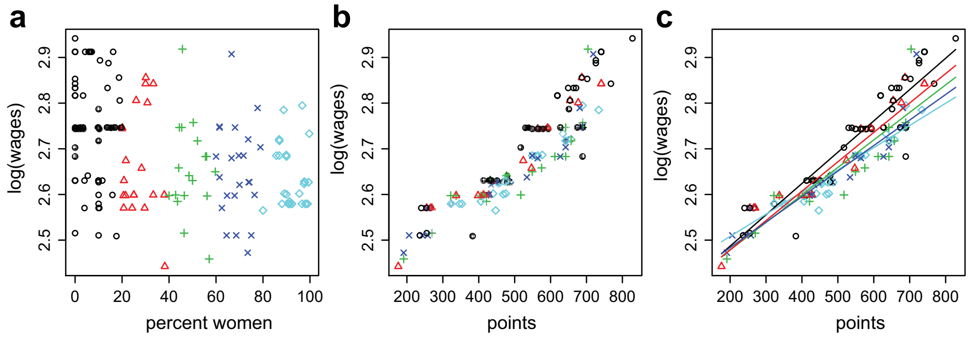

FIGURE 5: (a) Wages versus percent women for UAW data; (b) wages versus points; (c) fit to varying coefficient model for 0%, 30%, 50%, 70%, and 100% women. [Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com]

Penalty parameters that result in between 6 and 12.5 effective degrees of freedom for the unconstrained fit are chosen, and the GCV scores for the constrained and unconstrained fits are shown in Figure 6(a). The dotted line near the top indicates the GCV for the constant-slope model with varying intercept, indicating that the varying-slope model provides a better fit. The constrained and unconstrained fits coincide for the seven largest penalty parameters, and the GCV for the unconstrained fit is increasing past where they diverge. The minimum GCV fits for the intercept and slope functions are shown in plots (b) and (c). Finally, example fits to the scatterplot are shown in Figure 5(c), where the lines correspond to 0%, 30%, 50%, 70%, and 100% women, from the top to the bottom on the right.

In conclusion, we have found that penalized spline fits are more robust to user-defined choices when shape constraints are imposed. The constrained, penalized splines are computed via a weighted projection onto a polyhedral convex cone, which is a subset of the model space for the unconstrained penalized splines. All the flexibility of the penalized splines are retained, within the bounds of the shape restrictions. The convergence rates for the constrained versions are the same as for the unconstrained versions, but in practice, the constrained fits tend to have smaller squared error loss than the unconstrained fits, as well as more robustness to penalty parameter choice. Further, the method is fast and reliable: the constrained penalized spline estimates for 10,000 data sets with $n = 100$ as in Table 1 take less than 8 minutes to compute on a Mac Powerbook with a 3.06 GHz processor.

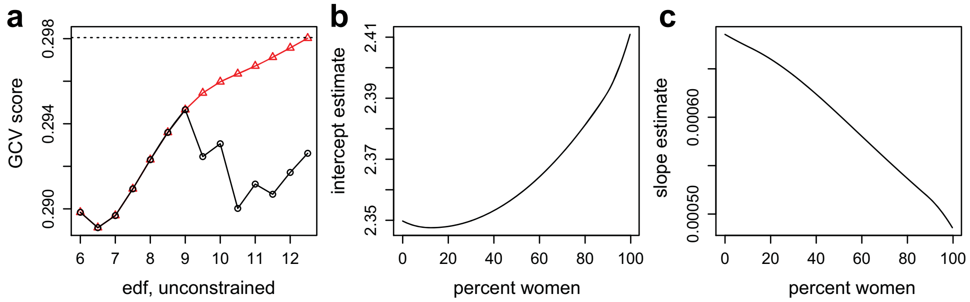

FIGURE 6: (a) The GCV values for the constrained and unconstrained slope functions are identical for the first seven values, and the smallest GCV occurs for an edf of 6.5. (b) The estimated intercept function for edf = 6.5, and (c) the estimated slope function. [Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com]

The GCV method of choosing the penalty parameter provides a simple way to determine if the shape assumptions are valid; this is not a formal hypothesis test but has been shown to have good large-sample properties and compares well to three competing methods. A straight-forward extension to the additive models has been given, where the fit is obtained through a single cone projection; no back-fitting or two-step procedures are necessary.

The ability to account for co-variates in data analysis is important, to check for possible confounding effects and to model another source of variation. The generalized regression model and the varying-coefficient model are also simple extensions, and co-variates are easily included in the model. User-friendly R-code for methods presented in this paper can be found at `http://www.stat.colostate.edu/~meyer/penspl.htm`.

ACKNOWLEDGEMENTS

This work was supported by a DMS grant from the National Science Foundation. The author is grateful for the helpful comments and suggestions of two referees and an associate editor.

BIBLIOGRAPHY

Bowman, A. W., Jones, M. C., & Gijbels, I. (1998). Testing monotonicity of regression. *Journal of Computational and Graphical Statistics*, 7(4), 489–500.

de Boor, C. (2001). *A Practical Guide to Splines*, revised edition, Springer, New York.

Claeskens, G., Krivobokova, T., & Opsomer, J. D. (2009). Asymptotic properties of penalized spline estimators. *Biometrika*, 96(3), 529–544.

Eilers, P. H. C. & Marx, B. D. (1996). Flexible smoothing with B-splines and penalites. *Statistical Science*, 11(2), 89–121.

Friedman, J. & Tibshirani, R. (1984). The monotone smoothing of scatterplots, *Technometrics*, 26(3), 243–250.

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

Hall, P. & Huang, L. S. (2001). Nonparametric kernel regression subject to monotonicity constraints. *Annals of Statistics*, 29(3), 624–647.

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

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

Hastie, T. J. & Tibshirani, R. J. (1993). Varying coefficient models. *Journal of the Royal Statistical Society, Series B*, 55(4), 757–796.

He, X., & Shi, P. (1998). Monotone *B*-spline smoothing. *Journal of the American Statistical Society*, 93(442), 643–650.

Huang, J. Z. (1998). Projection estimation in multiple regression with application to functional ANOVA models. *The Annals of Statistics*, 26(1), 242–272.

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

Killingsworth, M. R. (2002). Comparable worth and pay equity: Recent developments in the United States. *Canadian Public Policy*, 28, 171–186.

Li, Y. & Ruppert, D. (2008). On the asymptotics of penalized splines. *Biometrika*, 25(2), 415–436.

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

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

Mammen, E., Marron, J. S., Turlach, B. A., & Wand, M. P. (2001). A general projection framework for constrained smoothing. *Statistical Science*, 16(3), 232–248.

McCullagh, P. & Nelder, J. A. (1989). *Generalized Linear Models*, 2nd ed., Chapman & Hall, New York.

Meyer, M. C. & Woodroofe, M. (2004). Estimation of a unimodal density using shape restrictions, *Canadian Journal of Statistics*, 32(1), 85–100.

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

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

Ramsay, J. O. (1998). Estimating smooth monotone functions. *Journal of the Royal Statistical Society, Series B*, 60(2), 365–375.

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

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

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

Tantiyaswasdikul, C. & Woodroofe, M. (1994). Isotonic smoothing splines under sequential designs. *Journal of Statistical Planning and Inference*, 38, 75–88.

Utreras, F. I. (1985). Smoothing noisy data under monotonicity constraints: Existence, characterization, and convergence rates. *Numerishe Mathematik*, 47, 611–625.

Wang, X. & Li, F. (2008). Isotonic smoothing spline regression. *Journal of Computational and Graphical Statistics*, 17(1), 21–37.

Wang, J. C. & Meyer, M. C. (2011). Testing the monotonicity or convexity of a function using regression splines. *Canadian Journal of Statistics*, 39(1), 89–107.

Wood, S. N. (1994). Monotonic smoothing splines fitted by cross validation. *SIAM Journal of Scientific Computing*, 15(5), 1126–1133.

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

APPENDIX

Proof of Theorem 1. It is straightforward to show: ■

Lemma 1. *There is a constant* $c_1 \in (0, 1)$ *such that for any parabolas* g_1 *and* g_2 *on any interval* (a, b)*, where* $g_1'(a) \geq \xi$ *and* $g_2'(a) \leq 0$*, then*

$$\int_a^b \left[g_1(x) - g_2(x)\right]^2 \mathrm{d}x \geq c_1 \xi^2 (b-a)^3.$$

Let $\bar{f}_n$ minimize $\int_0^1 [f(x) - \bar{f}_n(x)]^2 \mathrm{d}x$ over the linear space of spline functions. Huang (1998) showed that $\sup_{x\in[0,1]} |\bar{f}_n(x) - f(x)| = O(k^{-3})$, under (A1)–(A3). This leads to:

Lemma 2. *If* $f'(x) \geq \xi > 0$*, then* $\sup_{x\in(0,1)} \bar{f}_n'(x) > \xi/2$ *for large enough n.*

Proof. Because $\bar{f}_n'$ is piecewise linear, its minimum value over [0, 1] occurs at a knot t_l, for some $l = 1, \ldots, k$. Then for some $\eta \in [t_l, t_{l+1}]$,

$$\begin{aligned}\bar{f}_n(t_{l+1}) - f(t_{l+1}) = \bar{f}_n(t_l) - f(t_l) + \left[\bar{f}_n'(t_l) - f'(t_l)\right](t_{l+1} - t_l) \\ + \frac{1}{2}\left[\bar{f}_n''(\eta) - f''(\eta)\right](t_{l+1} - t_l)^2.\end{aligned}$$

By bounded mesh ratio, $k(t_{l+1} - t_l)$ is bounded above and below, and f'' and $\bar{f}_n''$ are bounded, so we must have $\bar{f}_n'(x) \geq \xi/2$ for large enough n. ■

Proof of theorem. We assume $q = 3$ so that $\int_0^1 [\tilde{f}(x) - \bar{f}(x)]^2 dx = O_p(n^{-6/7})$ for either (B1) or (B2), by Claeskens, Krivobokova, & Opsomer (2009). If there is a j such that $\tilde{f}'(t_j) \leq 0$, then

$$\int_0^1 \left[\tilde{f}(x) - \bar{f}(x)\right]^2 dx \geq \int_{t_j}^{t_{j+1}} \left[\tilde{f}(x) - \bar{f}(x)\right]^2 dx$$
$$\geq c_1 \xi^2 (t_{j+1} - t_j)^3.$$

where c_1 is the constant from Lemma 1. This event has probability approaching zero if the number of knots grows as $n^{2/7-\eta}$ for some $\eta \in (0, 1/7]$.

For the second part of the theorem, we note that for the unconstrained fit, SSE/n is a consistent estimator of σ^2, but for the constrained fit, we can show that SSE/n converges to $\sigma^2 + c_2$, where $c_2 \geq (b - a)^2 \xi/4$. The limit of the numerator of the GCV, divided by n, is larger by c_2 for the constrained version, while the denominators both tend to one. The probability that the GCV for the unconstrained fit is smaller (and hence the unconstrained version is chosen) goes to one. ■

Received 17 January 2011
Accepted 4 October 2011