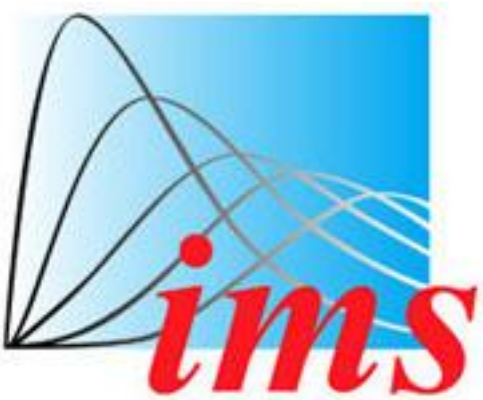

Statistical Science
2018, Vol. 33, No. 4, 595–614
https://doi.org/10.1214/18-STS671

A Framework for Estimation and Inference in Generalized Additive Models with Shape and Order Restrictions

Mary C. Meyer

Abstract. Methodology for the partial linear generalized additive model is presented, where components for continuous predictors may be modeled with shape-constrained regression splines, and components for ordinal predictors may have partial orderings. The estimated mean function is obtained through a projection (or iteratively reweighted projections) onto a polyhedral convex cone; this is key for formally derived inference procedures. Pointwise confidence bands and hypothesis tests for the individual components, as well as a model selection method, are proposed. These methods are available in the `R` package `cgam`.

Key words and phrases: Monotone, convex, partial linear, confidence interval.

1. BACKGROUND

Estimation and inference under shape and order constraints dates back to Hildreth (1954), who formulated maximum likelihood estimators for convex functions. Contributions to estimation and inference involving complete and partial orderings were made by Brunk (1955), Ayer et al. (1955), van Eeden (1956) and others. These early estimators were not smooth; the isotonic regression estimator is piecewise constant, and the convex regression estimator is piecewise linear. Isotonic smoothing spline estimators were proposed by Tantiyaswasdikul and Woodroofe (1994), and various ideas for shape-constrained kernel regression estimators were given by Mammen (1991), Hall and Huang (2001) and Du, Parmeter and Racine (2013).

Regression splines are attractive from an inferential perspective because the shape-constrained estimator can be expressed as a mixture of linear estimators. Monotone regression spline estimators were introduced by Ramsay (1988), who defined the I-splines, shown in Figure 1 in the left-hand plot. These quadratic spline basis functions have the property that for each knot (marked by the dotted vertical lines), exactly one basis function has nonzero slope. Hence, nonnegative coefficients comprise necessary and sufficient conditions for linear combinations of these basis functions to be nondecreasing. Any nondecreasing quadratic spline function can be expressed as a linear combination of the I-spline basis functions with nonnegative coefficients, plus an unconstrained intercept. Meyer (2008) defined the C-splines, shown in the middle panel of Figure 1. These cubic spline basis functions are constructed so that at each knot, exactly one has nonzero second derivative. Any convex cubic spline function can be expressed as a linear combination of the C-spline basis functions with nonnegative coefficients, plus an unconstrained linear function. Functions with combinations of monotonicity and convexity assumptions can be modeled with the C-splines as well. Meyer, Kim and Wang (2018) derived convergence rates for the constrained splines, under mild conditions, if the number of knots increases on the order of $n^{1/7}$ for the quadratic splines and on the order of $n^{1/9}$ for the cubic splines.

The third plot in Figure 1 shows increasing and concave least-squares fits to a scatterplot of height versus diameter of white spruce trees, available as the `whitespruce` data set in the `R` package

Mary C. Meyer is Professor, Statistics Department, Colorado State University, 212 Statistics Building, Fort Collins, Colorado 80523-1877, USA (e-mail: meyer@stat.colostate.edu).

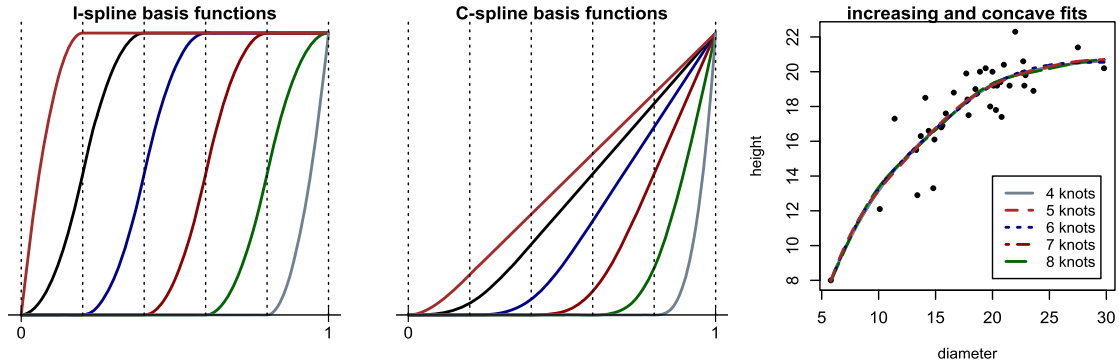

FIG. 1. *Basis functions for constrained splines, and example fits to a data set of* 36 *observations of height and diameter of white spruce trees to demonstrate robustness to knot choices.*

`ConSpline`. Constrained scatterplot smoothers are more robust to tuning parameters, compared to unconstrained smoothers, because of the shape restriction, increasing the number of knots does not necessarily lead to severe overfitting. The constraints prevent the wiggling typically associated with overfitting; hence, the increasing and concave fits to the white spruce data using 4–8 knots are almost identical.

With constrained regression splines, the least-squares estimator is found by projection onto a polyhedral convex cone, and for the generalized regression, through iteratively reweighted cone projections. A set $\mathcal{C}$ in $\mathbb{R}^n$ is a cone if for any $\boldsymbol{\eta} \in \mathcal{C}$, positive multiples of $\boldsymbol{\eta}$ are also in $\mathcal{C}$. The cone is polyhedral if it is finitely generated, that is, there is a finite set of vectors in the cone so that any vector in the cone can be expressed as a linear combination of these vectors, with linearly constrained coefficients.

The projection onto a polyhedral cone lands on a face of the cone, and coincides with the projection onto a linear space defined by this face. The projection of another realization of the response might land on a different face, so coincides with a projection onto another linear space. Thus the constrained estimator is a mixture of ordinary linear estimators, and inference methods use this important property. Meyer (2008) provided a test of constant versus increasing regression function, where the alternative is estimated with I-splines, and a test for linear versus convex regression function, where the alternative is estimated with C-splines. These have, under the normal errors assumption, a likelihood ratio test where the null distribution is that of a mixture of beta random variables. The partial linear constrained spline least-squares model was developed by Meyer (2018a), who derived inference methods for the linear term that also take advantage of the polyhedral cone formulation.

The additive isotonic model without smoothing was considered by Bacchetti (1989), with application by Morton-Jones et al. (2000) and additional work by Mammen and Yu (2007), Cheng (2009), Cheng, Zhao and Li (2012), Fang and Meinshausen (2012), Meyer (2013), Yu (2014), and Chen and Samworth (2016). The generalized additive model with smooth monotone components was fit with boosting techniques by Tutz and Leitenstorfer (2007), with Bayesian ideas by Meyer, Hackstadt and Hoeting (2011), and with penalized splines by Pya and Wood (2015). Villalobos and Wahba (1987) considered inequality-constrained thin-plate smoothing splines.

In this paper, we extend these ideas and give a framework for estimation and inference for the generalized additive partial linear model, where the components for continuous predictors are modeled with splines, and the components for ordinal predictors are modeled with partial orderings. The focus is on inference methods for the components. We consider generalized regression models for independent observations $Y_1, \ldots, Y_n$, where the probability distribution can be expressed as

$$p(y_i; \eta_i, \tau) = \exp\big[\{y_i \eta_i - b(\eta_i)\}\tau - c(y_i, \tau)\big], \quad i = 1, \ldots, n.$$

The family of distributions determine the functions b and c, and the parameter τ is related to the dispersion. For example, if Y_i is a binary response, we can define $b(\eta_i) = \log(1 + e^{\eta_i})$ to get the logistic model with $\mu_i = \mathrm{E}(Y_i) = e^{\eta_i}/(1 + e^{\eta_i})$. For the normal-errors model, η_i is the expected value of Y_i, $b(\eta_i) = \eta_i^2/2$, $\tau = 1/\sigma^2$, where σ^2 is the model variance. In any case, $\mu_i \equiv E(Y_i) = b'(\eta_i)$.

The vector $\boldsymbol{\eta} = (\eta_1, \ldots, \eta_n)^\top$ is determined by the predictor values. We consider predictor functions of the

form

$$\eta_i = f_1(t_{1i}) + \cdots + f_L(t_{Li}) + g_1(z_{1i}) + \cdots + g_R(z_{Ri}) + \mathbf{x}_i^\top \boldsymbol{\beta}, \tag{1.1}$$

where the predictors t_ℓ are treated as continuous and f_ℓ are functions to be estimated with shape-constrained splines, for $\ell = 1, \ldots, L$. The z_r predictors are ordinal and a partial ordering will be imposed on the components g_r, $r = 1, \ldots, R$. The parametrically modeled predictors are included in the vector $\mathbf{x}$.

In the next section, the f_ℓ and g_r components of the model will be formulated in terms of convex cones. The component cones and linear spaces are then combined to form a single convex cone. For the least-squares model, the estimate $\hat{\boldsymbol{\eta}}$ is simply the projection onto the single cone, and the component estimates are readily determined. For the generalized regression models, $\hat{\boldsymbol{\eta}}$ is estimated through iteratively reweighted cone projections. Hypothesis tests concerning the individual components are given in Section 3. Pointwise confidence intervals for $\boldsymbol{\eta}$ and the model components are given in Section 4 and shown through simulations to have good coverage and small lengths compared to other methods. A model selection method, where the shapes do not have to be specified a priori, is given in Section 5, and some discussion is provided in Section 6.

2. BUILDING THE MODEL CONE

2.1 Methodology

We consider three types of model components: constrained regression splines to model the effect of a continuous predictor, components for ordinal predictors with partial or complete orderings and components for parametrically modeled covariates such as nominal or linear predictors. The estimate of each model component is found in a polyhedral convex cone or a linear space. We combine the individual cones and linear spaces into a large cone so that the maximum likelihood estimator $\hat{\boldsymbol{\eta}}$ is found by projections onto a single cone, after which we sort out the individual component estimates.

To estimate f_ℓ with constrained regression splines, we define knots $\xi_{\ell,1} < \cdots < \xi_{\ell,K_\ell}$ where the $t_{\ell,i}$ values fall in $[\xi_{\ell,1}, \xi_{\ell,K_\ell}]$, $i = 1, \ldots, n$. If f_ℓ is assumed to be "increasing" or "decreasing," we define a set of K_ℓ quadratic I-spline basis functions with the property that at each knot, exactly one basis function has nonzero slope. We estimate f_ℓ as a linear combination of the basis functions plus an intercept; necessary and sufficient conditions for monotonicity are that the coefficients are nonnegative. If the constraint is "convex" or "concave" we define a set of K_ℓ cubic C-spline basis functions with the property that at each knot, exactly one basis function has nonzero second derivative. We model f as a linear combination of these basis functions with positive coefficients, plus an unconstrained multiple of t_ℓ as well as an intercept; these are necessary and sufficient conditions for convexity. If the constraints involve both monotonicity and convexity, such as "increasing and convex," we use the K_ℓ cubic C-spline basis functions and include t_ℓ as a basis function with a constrained coefficient. For a more detailed treatment of constrained regression splines, see Meyer (2008).

For each $\ell = 1, \ldots, L$, let the spline basis functions be $\delta_{\ell,1}(t), \ldots, \delta_{\ell,m_\ell}(t)$, where $m_\ell = K_\ell$ for monotone or convex splines and $m_\ell = K_\ell + 1$ for combinations of monotonicity and convexity. Let $\boldsymbol{\Delta}_\ell$ be an $n \times m_\ell$ matrix with $\boldsymbol{\Delta}_{\ell,ji} = \delta_{\ell,j}(t_i)$. The cone over which the sum of squared residuals is to be minimized, or the likelihood to be maximized, is

$$\mathcal{C}_\ell^{(1)} = \{\boldsymbol{\eta} \in \mathbb{R}^n : \boldsymbol{\eta} = \boldsymbol{\Delta}_\ell \boldsymbol{\alpha} + \mathbf{X}_{0,\ell}\boldsymbol{\alpha}_0, \text{ for } \boldsymbol{\alpha} \geq \mathbf{0}\},$$

where for constraints involving monotonicity, $\mathbf{X}_{0,\ell} = \mathbf{1} = (1, \ldots, 1)^\top$, and for convex or concave constraints, $\mathbf{X}_{0,\ell} = [\mathbf{1}|\mathbf{t}_\ell]$. (The intercepts when $L > 1$ will be combined later.)

For ordinal predictors, we assume that a partial ordering is known. For each $r = 1, \ldots, R$, define $\boldsymbol{\eta}_r \in \mathbb{R}^n$ such that $\eta_{r,i} = g_r(z_{r,i})$; Meyer (2013) showed that the set of possible components for the ordinal predictor can be described as

$$\mathcal{C}_r^{(2)} = \{\boldsymbol{\eta} \in \mathbb{R}^n : \mathbf{A}_r \boldsymbol{\eta} \geq \mathbf{0} \text{ and } \mathbf{B}_r \boldsymbol{\eta} = \mathbf{0}\},$$

where the inequality constraints using $\mathbf{A}_r$ impose the partial ordering, and the equality constraints using $\mathbf{B}_r$ ensure that $\eta_{r,i} = \eta_{r,j}$ whenever $z_{r,i} = z_{r,j}$. Let $\mathcal{V}_r$ be the largest linear space contained in $\mathcal{C}_r^{(2)}$.

For a "toy" example, suppose a treatment variable has three levels, one of which is a placebo, and there are three subjects assigned to each group, so that $n = 9$. Suppose subjects 1, 2, 3 are in the placebo group, subjects 4, 5, 6 are given Treatment 1, and subjects 7, 8, 9 are given Treatment 2. The researchers want to assume that the two treatments have at least as great an effect as the placebo, without imposing an order on the two treatments (this is called a *tree ordering*). Here, the

constraint matrix $\mathbf{A}_r$ has two rows:

$$\mathbf{A}_r = \begin{pmatrix} -1 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ -1 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 \end{pmatrix},$$

and $\mathbf{B}_r$ ensures that the effect of this predictor is the same within each of the three groups:

$$\mathbf{B}_r = \begin{pmatrix} -1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ -1 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & -1 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & -1 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & -1 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & -1 & 0 & 1 \end{pmatrix}.$$

The linear space $\mathcal{V}_r$ contained in the cone is the one-dimensional space of the constant vectors.

Define $\mathbf{D}_r$ so that its columns are a basis for the linear space that is orthogonal both to $\mathcal{V}_r$, and to the space spanned by the rows of $\mathbf{B}_r$. Proposition 2.2 of Meyer (2013) determined that if $\mathbf{A}_r$ is full row rank, then defining $\mathbf{\Gamma}_r = \mathbf{D}_r(\mathbf{A}_r\mathbf{A}_r^\top)^{-1}$ we have

$$\mathcal{C}_r^{(2)} = \{\boldsymbol{\eta} \in \mathbb{R}^n : \boldsymbol{\eta} = \mathbf{v} + \mathbf{\Gamma}_r\boldsymbol{\gamma};\ \text{for } \boldsymbol{\gamma} \geq \mathbf{0}, \mathbf{v} \in \mathcal{V}_r\}.$$

For the above tree ordering example, $\mathbf{\Gamma}_r$ is a multiple of the matrix

$$\begin{pmatrix} -1 & -1 & -1 & -1 & -1 & -1 & 2 & 2 & 2 \\ -1 & -1 & -1 & 2 & 2 & 2 & -1 & -1 & -1 \end{pmatrix}^\top,$$

and the effect of the categorical variable is a constant plus positive multiples of the columns of $\mathbf{\Gamma}_r$.

Next, we combine the component cones and linear spaces into a single cone. For the parametrically modeled covariates, let the $n \times p$ matrix $\mathbf{X}$ have rows $\mathbf{x}_i$, $i = 1 \ldots, n$. The component cones may contain same linear spaces, so let the columns matrix $\mathbf{X}_c$ represent a basis for the Minkowski sum of all these $L + R$ linear spaces. For example, the column space of $\mathbf{X}_c$ contains the constant vectors and any $\mathbf{t}_\ell$ where f_ℓ is modeled as convex or concave. We estimate $\boldsymbol{\eta}$ to be in the large cone

$$\begin{aligned} \mathcal{C} = \{\boldsymbol{\eta} \in \mathbb{R}^n : \boldsymbol{\eta} &= \mathbf{\Delta}_1\boldsymbol{\alpha}_1 + \cdots + \mathbf{\Delta}_L\boldsymbol{\alpha}_L \\ &+ \mathbf{\Gamma}_1\boldsymbol{\gamma}_1 + \cdots + \mathbf{\Gamma}_R\boldsymbol{\gamma}_R + \mathbf{X}_c\boldsymbol{\beta}_c + \mathbf{X}\boldsymbol{\beta}, \\ &\text{where } \boldsymbol{\alpha}_\ell \geq \mathbf{0}, \ell = 1, \ldots, L, \\ &\text{and } \boldsymbol{\gamma}_r \geq \mathbf{0}, r = 1, \ldots, R\}. \end{aligned} \tag{2.1}$$

Create an overall $n \times m$ "design" matrix for the constrained components

$$\mathbf{\Delta} = [\mathbf{\Delta}_1|\cdots|\mathbf{\Delta}_L|\mathbf{\Gamma}_1|\cdots|\mathbf{\Gamma}_R] \tag{2.2}$$

as well as an $m \times 1$ constrained coefficient vector $\boldsymbol{\zeta}^\top = [\boldsymbol{\alpha}_1^\top, \ldots, \boldsymbol{\alpha}_L^\top, \boldsymbol{\gamma}_1^\top, \ldots, \boldsymbol{\gamma}_R^\top]$. We assume that the columns of $\mathbf{\Delta}$, $\mathbf{X}$ and $\mathbf{X}_c$ form a linearly independent set; otherwise, the component effects are not identifiable. Linear dependence in the columns results from highly or completely correlated predictors and/or poorly spaced knots. For example, if the knot spacing for t_ℓ is such that there are no observed $t_{\ell,i}$ between two knots, this may lead to linear dependence. This may be corrected by changing the knot spacing.

The cone (2.1) can be written as

$$\begin{aligned} \mathcal{C} = \{\boldsymbol{\eta} \in \mathbb{R}^n : \boldsymbol{\eta} &= \mathbf{\Delta}\boldsymbol{\zeta} + \mathbf{X}_c\boldsymbol{\beta}_c + \mathbf{X}\boldsymbol{\beta} \\ &\text{where } \boldsymbol{\zeta} \geq \mathbf{0}\}, \end{aligned} \tag{2.3}$$

and the projection of $\mathbf{y}$ onto $\mathcal{C}$ minimizes $\|\mathbf{y} - \boldsymbol{\eta}\|^2$ over $\boldsymbol{\eta} \in \mathcal{C}$. This projection lands on a *face* of the cone. Subsets $J \subseteq \{1, \ldots, m\}$ index the faces, defined as

$$\mathcal{F}_J = \{\boldsymbol{\eta} \in \mathbb{R}^n : \boldsymbol{\eta} = \mathbf{\Delta}_J\boldsymbol{\zeta} + \mathbf{X}_c\boldsymbol{\beta}_c + \mathbf{X}\boldsymbol{\beta} \text{ where } \boldsymbol{\zeta} \geq \mathbf{0}\},$$

and the columns of $\mathbf{\Delta}_J$ are the columns of $\mathbf{\Delta}$ that are indexed by J. If the projection $\hat{\boldsymbol{\eta}}$ of $\mathbf{y}$ onto the cone $\mathcal{C}$ lands on face $\mathcal{F}_J$, then $\hat{\boldsymbol{\eta}}$ coincides with the projection of $\mathbf{y}$ onto the linear space $\mathcal{L}_J$ that is spanned by the columns of $\mathbf{\Delta}_J$, $\mathbf{X}_c$, and $\mathbf{X}$ (see Proposition 4 of Meyer, 1999).

2.2 Computation

The function `coneB` in the package `coneproj` will determine the set J, through a sequence of "guesses" and ordinary least-squares projections. Then $\hat{\boldsymbol{\zeta}}$, $\hat{\boldsymbol{\beta}}_c$ and $\hat{\boldsymbol{\beta}}$ are determined as in ordinary least-squares regression with "design matrix" $[\mathbf{\Delta}_J|\mathbf{X}_c|\mathbf{X}]$. The estimates of the individual components are constructed from the elements of these estimated coefficient vectors.

Suppose we instead want to find $\hat{\boldsymbol{\eta}} \in \mathcal{C}$ to minimize $(\mathbf{y} - \boldsymbol{\eta})^\top\mathbf{W}(\mathbf{y} - \boldsymbol{\eta})$ for a positive definite $n \times n$ matrix $\mathbf{W}$; that is, we have a weighted regression model. This is equivalent to minimizing $\|\mathbf{W}^{1/2}\mathbf{y} - \mathbf{W}^{1/2}\mathbf{\Delta}\boldsymbol{\zeta} - \mathbf{W}^{1/2}\mathbf{X}_c\boldsymbol{\beta}_c - \mathbf{W}^{1/2}\mathbf{X}\boldsymbol{\beta}\|^2$, with the constraint $\boldsymbol{\zeta} \geq \mathbf{0}$. The above cone projection procedure can be followed with $\mathbf{W}^{1/2}\mathbf{y}$ in place of $\mathbf{y}$, $\mathbf{W}^{1/2}\mathbf{\Delta}$ in place of $\mathbf{\Delta}$, $\mathbf{W}^{1/2}\mathbf{X}_c$ in place of $\mathbf{X}_c$, and $\mathbf{W}^{1/2}\mathbf{X}$ in place of $\mathbf{X}$.

For the generalized regression model, recall that η is the predictor function rather than the expected response. We use a local scoring method to develop an iteratively reweighted cone projection. This is analogous to the iteratively reweighted least-squares fit to the GLM (McCullagh and Nelder, 1989, Section 2.5), but with a cone instead of a linear space. Let $\psi(\boldsymbol{\eta}) = -\sum_{i=1}^n \log p(y_i; \eta_i, \tau)$, and denote the gradient func-

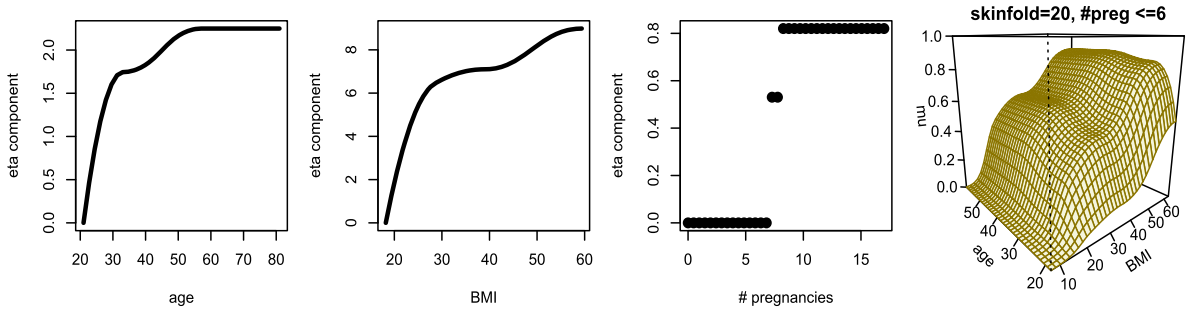

FIG. 2. *First three plots*: *estimated components of the log-odds of diabetes in Pima Indian women*, *for age*, *BMI and number of pregnancies*. *Final plot*: *surface representing estimated probabilities of diabetes in age and BMI*, *when skin fold is fixed at* 20 *and the number pregnancies is* 6 *or fewer.*

tion as $\nabla\psi(\boldsymbol{\eta})$ and the Hessian matrix as $\mathbf{H}(\boldsymbol{\eta})$. To minimize ψ over the cone $\mathcal{C}$:

1. Initialize $\boldsymbol{\eta}^{(0)} \in \mathcal{C}$.
2. At the kth iteration, evaluate $\nabla\psi(\boldsymbol{\eta}^{(k)})$ and $\mathbf{H}(\boldsymbol{\eta}^{(k)})$, and compute

$$\boldsymbol{\xi}^{(k)} = \boldsymbol{\eta}^{(k)} + \mathbf{H}(\boldsymbol{\eta}^{(k)})^{-1}\nabla\psi(\boldsymbol{\eta}^{(k)})$$

as well as $\mathbf{W}^{(k)} = -\mathbf{H}(\boldsymbol{\eta}^{(k)})$. Project $\boldsymbol{\xi}^{(k)}$ onto $\mathcal{C}$ using weights $\mathbf{W}^{(k)}$ to get $\boldsymbol{\eta}^{(k+1)}$.

3. Check for convergence. If the required distance is "small," the iteration ends; otherwise increment k and go to (2).

In Step 3, we compare the distance between $\boldsymbol{\mu}^{(k)}$ and $\boldsymbol{\mu}^{(k+1)}$, rather than the distance between and $\boldsymbol{\eta}^{(k)}$ and $\boldsymbol{\eta}^{(k+1)}$, because for models such as binomial and Poisson, the latter can be unbounded. For example, if the binomial responses corresponding to the last knot interval are all successes, then the maximum likelihood spline estimate for the log-odds might be infinite at the endpoint. However, the estimate for the mean tends to one. For the iteratively reweighted algorithm, the cone projection routine is modified so that the first guess for J at iteration k is the J from iteration $k-1$; this makes the routine almost as fast as for the unconstrained fit.

To demonstrate the method, we consider the data set `pima` in the `R` package `MASS` with the following description: "A population of women who were at least 21 years old, of Pima Indian heritage and living near Phoenix, Arizona, was tested for diabetes according to World Health Organization criteria. The data were collected by the US National Institute of Diabetes and Digestive and Kidney Diseases." The response variable is whether or not the subject has diabetes, and three predictors are treated as continuous: skin-fold measurement, body mass index and age. We also have the number of pregnancies, which is treated as ordinal. We assume that we know a priori that the probability of diabetes is increasing in each of the predictors.

Three of the estimated log odds components are shown in Figure 2. For the spline fits, the default knot choice in `cgam` is used, which for $n = 530$ is nine equally spaced knots. The estimated components deviate substantially from linearity; in particular, for age and BMI, the log-odds increase rapidly at the beginning of the range before leveling off. The skin-fold variable does not seem to have much effect on the response, given the effects of the other predictors. Of course, we would like to determine the statistical significance of each of the components, and determine confidence bounds for the probability estimates. These inference methods are addressed in the next two sections.

3. HYPOTHESIS TESTING FOR MODEL COMPONENTS

3.1 Tests Concerning the Parametrically Modeled Components

For inference about the parameters $\boldsymbol{\beta}$, we start with the least-squares model. Suppose $\mathbf{y} = \boldsymbol{\eta} + \boldsymbol{\varepsilon}$, where the components of $\boldsymbol{\eta}$ are as in (1.1) and $\boldsymbol{\varepsilon}$ is a mean-zero random vector with covariance matrix $\sigma^2\mathbf{I}$. Suppose that the projection of $\mathbf{y}$ onto the cone $\mathcal{C}$ falls on the face $\mathcal{F}_J$. Meyer (2018b) showed that, if $\mathbf{P}_J$ represents the projection matrix for the space $\mathcal{L}_J$ spanned by the columns of $\boldsymbol{\Delta}_J$ and $\mathbf{X}_c$, then

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}^\top(\mathbf{I} - \mathbf{P}_J)\mathbf{X})^{-1}\mathbf{X}^\top(\mathbf{I} - \mathbf{P}_J)\mathbf{y}. \tag{3.1}$$

The same paper also gave conditions for root-n convergence of $\hat{\boldsymbol{\beta}}$, where Theorem 4 states that under mild

assumptions,

$$\sqrt{n}(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta})=\sqrt{n}\big[\mathbf{X}^\top(\mathbf{I}-\mathbf{P}_J)\mathbf{X}\big]^{-1}\mathbf{X}^\top(\mathbf{I}-\mathbf{P}_J)\boldsymbol{\varepsilon}+o_p(1).$$

Approximate t and F statistics are constructed from the estimated covariance matrix

$$\widehat{\mathrm{cov}}(\hat{\boldsymbol{\beta}})=\hat{\sigma}^2\big[\mathbf{X}^\top(\mathbf{I}-\mathbf{P}_J)\mathbf{X}\big]^{-1},$$

where $\hat{\sigma}^2$ is the model variance estimate given by Meyer and Woodroofe (2000):

$$\hat{\sigma}^2=\frac{\sum_{i=1}^n(y_i-\hat{\mu}_i)^2}{n-d}, \tag{3.2}$$

where $d=\min(1.5D,m)$ and D is the dimension of $\mathcal{L}_J$ for the observed J.

Huang (2002) and Cheng (2009) derived inference results for the linear parameters in the partial linear additive model without smoothing.

For weighted least-squares regression with positive-definite weight matrix $\mathbf{W}$, we transform into the unweighted case by substituting $\mathbf{W}^{1/2}\mathbf{y}$ for $\mathbf{y}$, $\mathbf{W}^{1/2}\boldsymbol{\Delta}$ for $\boldsymbol{\Delta}$, and $\mathbf{W}^{1/2}\mathbf{X}$ for $\mathbf{X}$. For the generalized regression model, the converged value of $\boldsymbol{\xi}$ is used in place of the response $\mathbf{y}$ in the weighted least-squares regression with the converged value of the weight vector.

Next, we use simulations to compare these hypothesis tests for the linear terms in the binomial model, coded in the `R` package `cgam` (Liao and Meyer, 2018), with those of `scam` (Pya and Wood, 2015) and the routine `gam` in the `mcgv` package. We use a continuous predictor t with values equally spaced in $[0,1]$, and a nominal covariate x with three levels. The model is $\eta_i=f(t_i)+\beta_0+\beta_1 I\{x_i=1\}+\beta_2 I\{x_i=2\}$, for $i=1,\ldots,n$. We imagine that we know a priori that f is smooth and increasing, so we model f with constrained I-splines. The null hypothesis is that there is no effect of the covariate x on the response (i.e., $H_0:\beta_1=\beta_2=0$ versus H_1: at least one of β_1 or β_2 is not zero), and an approximate F test for submodels is used. To construct this test statistic, we use the final values $\boldsymbol{\xi}$ and $\mathbf{W}$ from the iteratively reweighted least-squares procedure, and use the weighted sums of squared residuals from the the null (SSE_0) and alternative fits (SSE_1). Because we do not assume the model variance σ^2 is known, the test statistic is

$$F=\frac{(SSE_0-SSE_1)/2}{SSE_1/(n-df_1)},$$

where df_1 is the dimension of $\mathcal{L}_J$ for the alternative hypothesis cone projection. Under H_0, this has approximately an $F(2,n-df_1)$ distribution.

For the simulated data sets, the values of x are related to the values of t, so that $x=1$ is more likely for larger values of t, and $x=2$ is more likely for smaller values of t. Specifically, we generate independent uniform random numbers $u_1,\ldots,u_n$, and if $u_i\in(0,0.05+0.85t_i^4)$ we set $x_i=1$, if $u_i\in[0.4+0.55t_i^{1/4},1]$, we set $x_i=2$; otherwise $x_i=3$. For the simulations, the true log odds is either linear: $f(x)=4x-2$, or $f(t)=-1+60(t-1/2)_+^4+\beta_1 I\{x=1\}$, where $(\cdot)_+=\max(0,\cdot)$.

An example data set with the nonlinear log-odds, $\beta_1=-1$ and $n=200$ is shown in Figure 3, along with fits from the three methods. The constrained fits are

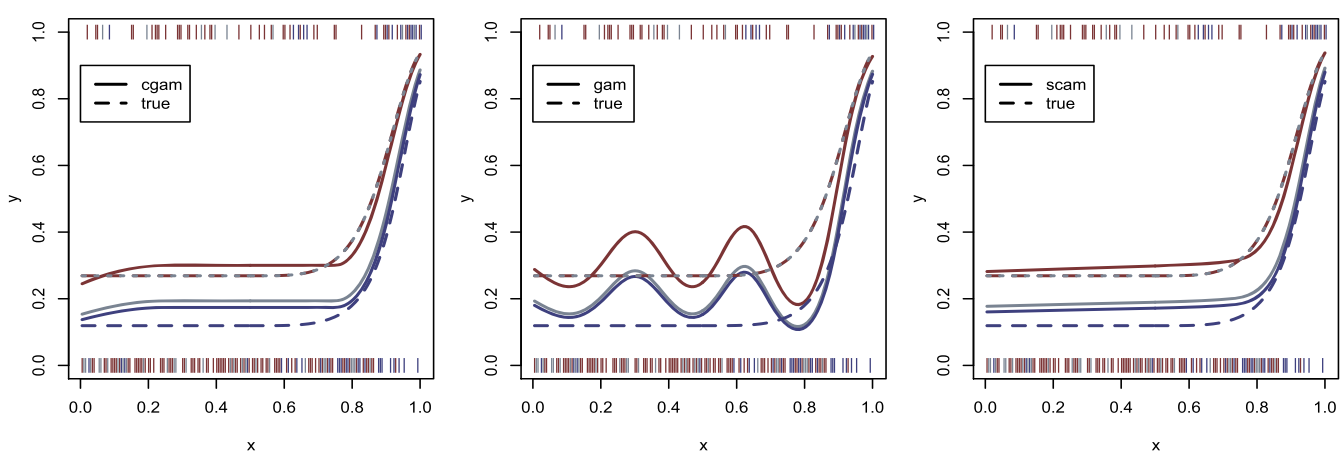

FIG. 3. *An example data set generated from a binomial model with log-odds function* $\eta_i=-1+60(t_i-1/2)_+^4-I\{x_i=1\}$. *Estimated probability functions are for three levels of the categorical predictor* x. *The* `cgam` *and* `scam` *fits are constrained to be increasing.*

TABLE 1

Proportions of 10,000 *data sets for which* $H_0 : \boldsymbol{\beta} = 0$ *is rejected, with test size* $\alpha = 0.05$. *The response is binomial with log-odds function* $\eta_i = f(t_i) + \beta_1 I\{x_i = 1\} + \beta_2 I\{x_i = 2\}$

β_1	n	$\eta(t) = 4x - 2$				$\eta(t) = -1 + 60(t_i - 1/2)_+^4$			
		gam (def.)	scam (def.)	scam (sp = 0)	cgam (def.)	gam (def.)	scam (def.)	scam (sp = 0)	cgam (def.)
0	200	0.042	0.050	0.041	0.051	0.050	0.146	0.044	0.053
0	400	0.050	0.050	0.048	0.053	0.050	0.089	0.053	0.052
0	800	0.052	0.050	0.047	0.049	0.047	0.083	0.052	0.050
−1	200	0.499	0.513	0.464	0.491	0.304	0.394	0.381	0.423
−1	400	0.813	0.837	0.800	0.808	0.651	0.746	0.730	0.743
−1	800	0.986	0.991	0.981	0.983	0.951	0.969	0.970	0.969

similar; the difference in the hypothesis test results is due to the testing method rather than the fit. We generated 10,000 data sets for each of the two functions, four sample sizes and two values of β_1, to compare test size and power, which are shown in Table 1. For the `scam` test, we use both the default tuning parameters, and the tuning parameters that "match" those of `cgam`. That is, we set the penalty parameter to zero, and use the same number of knots as the `cgam` default. We find that for the linear log-odds, `scam` with the default choices outperforms the other methods. However, for the nonlinear log-odds, the test size for `scam` is inflated if the default tuning parameters are used, especially for the smaller sample sizes, while the power for our method is consistently greater than that for `gam`. For the matching tuning parameters, the `cgam` and `scam` results are similar.

3.2 Testing $H_0 : \eta \in \mathcal{V}$ Versus $H_1 : \eta \in \mathcal{C}\backslash\mathcal{V}$

Next, we turn to hypothesis tests concerning the constrained components. Let $\mathcal{V}$ be the linear space spanned by the columns of $\mathbf{X}_c$ and $\mathbf{X}$; this is the largest linear space contained in the cone $\mathcal{C}$. The traditional test of $H_0 : \boldsymbol{\eta} \in \mathcal{V}$ versus $H_1 : \boldsymbol{\eta} \in \mathcal{C}\backslash\mathcal{V}$, for the normal-errors model, was first presented by Bartholomew (1959), in terms of the one-way ANOVA model, where the null hypothesis corresponds to constant means, versus the completely ordered alternative. A more general treatment of these one-sided tests was given by Raubertas, Lee and Nordheim (1986). Here, the null hypothesis is that none of constrained predictors is related to the response; that is, $\boldsymbol{\alpha}_1 = \cdots = \boldsymbol{\alpha}_L = \boldsymbol{\gamma}_1 = \cdots = \boldsymbol{\gamma}_R = 0$, and the alternative is that at least one of the constrained predictors has a nonzero coefficient. Sen and Meyer (2017) formulated a "double-cone" test against a linear model, where the alternative shape does not need to be specified.

We use the following notation. A random variable T has a mixture of chi-squared distributions with mixing parameters $\mathbf{p} \geq \mathbf{0}$, $\sum_{j=d_1}^{d_2} p_j = 1$, written $T \sim \bar{\chi}^2_{d_1:d_2}(\mathbf{p})$, if for $c > 0$,

$$\mathrm{P}(T \leq c) = \sum_{j=d_1}^{d_2} \mathrm{P}\big(\chi_j^2 \leq c\big) p_j,$$

where χ_j^2 is a chi-squared random variable with j degrees of freedom. Proofs of the following result are available in Raubertas, Lee and Nordheim (1986), Robertson, Wright and Dykstra (1988) and Meyer (2003).

LEMMA 3.1. *Let* $\mathcal{C}$ *be a polyhedral cone where the largest linear space contained in the cone has dimension* d_1, *and the smallest linear space containing the cone has dimension* d_2. *If* $\boldsymbol{\varepsilon} \sim N(\mathbf{0}, \sigma^2\mathbf{I})$, *then*

$$\frac{1}{\sigma^2}\big\|\Pi(\boldsymbol{\varepsilon}|\mathcal{C})\big\|^2 \sim \bar{\chi}^2_{d_1:d_2}(\mathbf{p}),$$

where $\mathbf{p} = (p_{d_1}, \ldots, p_{d_2})$ *and* p_d *is the probability that the projection of* $\boldsymbol{\varepsilon}$ *onto* $\mathcal{C}$ *falls on a face where the dimension of the corresponding linear space is* d.

The intuition behind this lemma is as follows: The projection of $\boldsymbol{\varepsilon}$ onto the cone lands on a face $\mathcal{F}_J$, and coincides with the projection of $\boldsymbol{\varepsilon}$ onto $\mathcal{L}_J$. We know that $\|\Pi(\boldsymbol{\varepsilon}|\mathcal{L}_J)\|^2/\sigma^2 \sim \chi^2(\dim(\mathcal{L}_J))$; because another realization of the response might have a projection that lands on a different face, we have a mixture of chi-squared random variables.

First, we consider a linear space $\mathcal{V}_0 \subseteq \mathcal{V}$, and the test of $H_0 : \boldsymbol{\eta} \in \mathcal{V}_0$ versus $H_a : \boldsymbol{\eta} \in \mathcal{C} \backslash \mathcal{V}_0$. We start by defining a matrix $\mathbf{X}_0$ whose columns form a basis for $\mathcal{V}_0$, a projection matrix $\mathbf{P}_0$ for the linear space $\mathcal{V}_0$, and a matrix $\mathbf{X}_1$ whose columns form a basis for $\mathcal{V} \cap \mathcal{V}_0^{\perp}$, where $\mathcal{V}_0^{\perp}$ denotes the linear space orthogonal to $\mathcal{V}_0$. Then define $\mathbf{E} = (\mathbf{I} - \mathbf{P}_0)\boldsymbol{\Delta}$ and the cone can be written as

$$\mathcal{C} = \{\boldsymbol{\eta} \in \mathbb{R}^n : \boldsymbol{\eta} = \mathbf{E}\boldsymbol{\alpha} + \mathbf{X}_0\boldsymbol{\beta}_0 + \mathbf{X}_1\boldsymbol{\beta}_1, \text{ for } \boldsymbol{\alpha} \geq \mathbf{0}\}.$$

The projection of $\mathbf{y}$ onto $\mathcal{C}$ (the alternative hypothesis fit) can be written as

$$\hat{\boldsymbol{\eta}}_1 = \mathbf{E}_J\hat{\boldsymbol{\alpha}} + \mathbf{X}_0\hat{\boldsymbol{\beta}}_0 + \mathbf{X}_1\hat{\boldsymbol{\beta}}_1,$$

while the null hypothesis fit is $\hat{\boldsymbol{\eta}}_0 = \mathbf{X}_0\hat{\boldsymbol{\beta}}_0$ (the $\hat{\boldsymbol{\beta}}_0$ is the same for the null and alternative fits due to orthogonality).

Let $SSR_0 = \|\mathbf{y} - \hat{\boldsymbol{\eta}}_0\|^2$ be the sum of squared residuals under the null hypothesis fit, and let $SSR_1 = \|\mathbf{y} - \hat{\boldsymbol{\eta}}_1\|^2$ be the sum of squared residuals under the alternative fit. By orthogonality,

$$SSR_0 - SSR_1 = \|\mathbf{E}_J\hat{\boldsymbol{\alpha}} + \mathbf{X}_1\hat{\boldsymbol{\beta}}_1\|^2,$$

and $\mathbf{E}_J\hat{\boldsymbol{\alpha}} + \mathbf{X}_1\hat{\boldsymbol{\beta}}_1$ is the projection of $\mathbf{y}$ onto the cone

$$\mathcal{C}_1 = \{\boldsymbol{\eta} \in \mathbb{R}^n : \boldsymbol{\eta} = \mathbf{E}\boldsymbol{\alpha} + \mathbf{X}_1\boldsymbol{\beta}_1, \text{ for } \boldsymbol{\alpha} \geq \mathbf{0}\}.$$

Such an orthogonality argument is well understood in the linear model setting; it applies to cones as well. For, if a cone is contained in a linear space orthogonal to a second linear space, than each face of the cone is orthogonal to the second linear space. If two cones are contained in orthogonal subspaces, then each face of one cone is orthogonal to every face of the other cone.

Under $H_0 : \boldsymbol{\eta} \in \mathcal{V}_0$, the expected value of $\mathbf{y}$ is in a space orthogonal to $\mathcal{C}_1$, so we can apply Lemma 3.1 if the normal errors assumption holds. The likelihood ratio test statistic distribution is that of a mixture of chi-squared random variables:

$$T = \frac{1}{\sigma^2}(SSR_0 - SSR_1) \sim \bar{\chi}^2_{d_1:d_2}(\mathbf{p}),$$

where d_1 is the dimension of the largest linear space contained in $\mathcal{C}_1$, and d_2 is the dimension of the smallest linear space containing $\mathcal{C}_1$. The mixing distribution is determined through simulations where the response vector is generated as a normal random vector with mean zero and identity covariance matrix.

If σ^2 is unknown, we use a test statistic with a mixture-of-betas null distribution. Define

$$B = \frac{T}{T + SSR_1/\sigma^2} = \frac{SSR_0 - SSR_1}{SSR_0}$$

then for $c \in (0, 1)$,

$$\mathrm{P}(B \leq c) = \sum_{d=d_1}^{d_2} B\big(d/2, (n - d_1 - d)/2\big)p_d,$$

where $B(a, b)$ is a beta random variable with parameters a and b. The mixing parameter $\mathbf{p}$ is the same as for the known-variance case.

For the normal errors model, this test is exact, at least to the precision of the estimated mixing parameters. For the generalized model, the null hypothesis fit is obtained, and as in the test for linear effects, the final values of $\boldsymbol{\xi}$ and $\mathbf{W}$ in the iterative reweighting algorithm are treated as response vector and weights, respectively. The mixing parameters are obtained using the alternative hypothesis cone that is transformed by the weights. The test is not exact for the generalized model, but the following simulations demonstrate its good qualities.

To compare this test with the methods in `gam` and `scam`, we use the same simulation set-up as in the previous example, but we test $H_0 : f \equiv 0$ versus $H_1 : f$ is increasing. Specifically, the log-odds in the binomial model is assumed to be of the form $\eta(t_i, x_i) = f(t_i) + \beta_1 I\{x_i = 1\} + \beta_2 I\{x_i = 2\}$, but this time we fix $\beta_1 = 2$ and $\beta_2 = -1$. We use $f(t) \equiv 0$ to find the test size, and $f(t) = 30(t - 1/2)_+^4$ is used to compare the powers of the tests. To find the mixing parameters in `cgam`, 1000 replications are used. These results are presented in Table 2. The default tuning parameters in `scam` give conservative test sizes and lower power than the other methods, but the test sizes for the "matching" tuning parameters are inflated. The power for the `cgam` method, compared to the test in `gam`, is substantially higher, mostly due to comparing a one-sided test with a two-sided test.

3.3 Tests Concerning Individual Constrained Components

Next, we consider the case where there are several constrained components and we want to test the significance of a single component of the model. In this case, our null hypothesis is not described by a linear space; instead we have a test against a subcone. Suppose the null hypothesis is $H_0 : \boldsymbol{\alpha}_1 = \mathbf{0}$ versus $H_1 : \boldsymbol{\alpha}_1 \geq 0$. Define

$$\boldsymbol{\Delta}_0 = \big[\boldsymbol{\Delta}_2 | \cdots | \boldsymbol{\Delta}_L | \boldsymbol{\Gamma}_1 | \cdots | \boldsymbol{\Gamma}_R\big]$$

TABLE 2

Proportions of 10,000 *data sets for which* $H_0 : f(t) = 0$ *is rejected, versus* $H_1 : f(t)$ *is increasing, with test size* $\alpha = 0.05$. *The response is binomial with log-odds function* $\eta(t_i, x_i) = f(t_i) + \beta_1 I\{x_i = 1\} + \beta_2 I\{x_i = 2\}$, $i = 1, \ldots, n$

n	$f(t) \equiv 0$				$f(t) = 30(t - 1/2)_+^4$			
	gam (def.)	scam (def.)	scam (sp = 0)	cgam (def.)	gam (def.)	scam (def.)	scam (sp = 0)	cgam (def.)
100	0.040	0.037	0.142	0.042	0.090	0.084	0.074	0.194
200	0.048	0.033	0.129	0.043	0.166	0.156	0.157	0.334
400	0.053	0.035	0.125	0.048	0.328	0.321	0.365	0.590
800	0.053	0.029	0.123	0.048	0.660	0.644	0.760	0.879

and

$$\boldsymbol{\zeta}_0 = \begin{bmatrix} \boldsymbol{\alpha}_2 \\ \vdots \\ \boldsymbol{\alpha}_L \\ \boldsymbol{\gamma}_1 \\ \vdots \\ \boldsymbol{\gamma}_R \end{bmatrix},$$

so that we can write $\boldsymbol{\eta} = \boldsymbol{\Delta}_1\boldsymbol{\alpha}_1 + \boldsymbol{\Delta}_0\boldsymbol{\zeta}_0 + \mathbf{X}_c\boldsymbol{\beta}_c + \mathbf{X}\boldsymbol{\beta}$. Let $\mathbf{P}_v$ be the projection matrix for $\mathcal{V}$, the space spanned by the columns of $\mathbf{X}_c$ and $\mathbf{X}$. Let $\tilde{\boldsymbol{\Delta}}_1 = (\mathbf{I} - \mathbf{P}_v)\boldsymbol{\Delta}_1$ and $\tilde{\boldsymbol{\Delta}}_0 = (\mathbf{I} - \mathbf{P}_v)\boldsymbol{\Delta}_0$, so $\boldsymbol{\eta}$ can be written in the form $\tilde{\boldsymbol{\Delta}}_1\boldsymbol{\alpha}_1 + \tilde{\boldsymbol{\Delta}}_0\boldsymbol{\zeta}_0 + \mathbf{X}_c\boldsymbol{\beta}_c + \mathbf{X}\boldsymbol{\beta}$, with $\boldsymbol{\alpha}_1 \geq \mathbf{0}$ and $\boldsymbol{\alpha}_0 \geq \mathbf{0}$.

Let $\mathbf{P}_0$ be the projection matrix for the space spanned by the columns of $\tilde{\boldsymbol{\Delta}}_0$. Then

$$\begin{aligned} \tilde{\boldsymbol{\Delta}}_1\boldsymbol{\alpha}_1 + \tilde{\boldsymbol{\Delta}}\boldsymbol{\zeta}_0 &= (\mathbf{I} - \mathbf{P}_0)\tilde{\boldsymbol{\Delta}}_1\boldsymbol{\alpha}_1 + \mathbf{P}_0\tilde{\boldsymbol{\Delta}}_1\boldsymbol{\alpha}_1 + \tilde{\boldsymbol{\Delta}}_0\boldsymbol{\zeta}_0 \\ &= (\mathbf{I} - \mathbf{P}_0)\tilde{\boldsymbol{\Delta}}_1\boldsymbol{\alpha}_1 \\ &\quad + \tilde{\boldsymbol{\Delta}}_0\big[(\tilde{\boldsymbol{\Delta}}_0^\top\tilde{\boldsymbol{\Delta}}_0)^{-1}\tilde{\boldsymbol{\Delta}}_0^\top\tilde{\boldsymbol{\Delta}}_1\boldsymbol{\alpha}_1 + \boldsymbol{\zeta}_0\big] \\ &=: \mathbf{E}_1\boldsymbol{\alpha}_1 + \tilde{\boldsymbol{\Delta}}_0\mathbf{a}, \end{aligned}$$

and $\boldsymbol{\eta} = \mathbf{E}_1\boldsymbol{\alpha}_1 + \tilde{\boldsymbol{\Delta}}_0\mathbf{a} + \mathbf{X}_c\boldsymbol{\beta}_c + \mathbf{X}\boldsymbol{\beta}$. To estimate the coefficients $\boldsymbol{\beta}_c$ and $\boldsymbol{\beta}$, we project $\mathbf{y}$ onto the linear space $\mathcal{V}$. To estimate $\mathbf{a}$ and $\boldsymbol{\alpha}_1$, we minimize $\|\mathbf{y} - (\mathbf{E}_1\boldsymbol{\alpha}_1 + \tilde{\boldsymbol{\Delta}}_0\mathbf{a})\|^2$ subject to

$$\begin{bmatrix} \mathbf{I} & \mathbf{M} \\ \mathbf{0} & \mathbf{I} \end{bmatrix} \begin{pmatrix} \mathbf{a} \\ \boldsymbol{\alpha}_1 \end{pmatrix} \geq \mathbf{0},$$

where $\mathbf{M} = -(\tilde{\boldsymbol{\Delta}}_0^\top\tilde{\boldsymbol{\Delta}}_0)^{-1}\tilde{\boldsymbol{\Delta}}_0^\top\tilde{\boldsymbol{\Delta}}_1$. By orthogonality, the estimate of $\boldsymbol{\eta}$ is the sum of the two projections.

Because the space spanned by the columns of $\mathbf{E}_1$ is orthogonal to the space spanned by the columns of $\tilde{\boldsymbol{\Delta}}_0$ and also orthogonal to $\mathcal{V}$, we have

$$\begin{aligned} &\big\|\mathbf{y} - (\mathbf{E}_1\hat{\boldsymbol{\alpha}}_1 + \tilde{\boldsymbol{\Delta}}_0\hat{\mathbf{a}} + \mathbf{X}_c\hat{\boldsymbol{\beta}}_c + \mathbf{X}\hat{\boldsymbol{\beta}})\big\|^2 \\ &\quad = \big\|\mathbf{y} - (\tilde{\boldsymbol{\Delta}}_0\hat{\mathbf{a}} + \mathbf{X}_c\hat{\boldsymbol{\beta}}_c + \mathbf{X}\hat{\boldsymbol{\beta}})\big\|^2 + \|\mathbf{E}_1\hat{\boldsymbol{\alpha}}_1\|^2. \end{aligned}$$

Again by orthogonality, the last term in the above equation is the squared length of the projection of $\mathbf{y}$ onto the cone defined by the columns of $\mathbf{E}_1$. Therefore, by Lemma 3.1, if H_0 is true so that $\mathrm{E}(\mathbf{y})$ is orthogonal to the column space of $\mathbf{E}_1$, we have

$$T_1 = \frac{1}{\sigma^2}\|\mathbf{E}_1\hat{\boldsymbol{\alpha}}_1\|^2 \sim \bar{\chi}^2_{0:m_1}(\mathbf{p}),$$

where m_1 is the number of columns of $\mathbf{E}_1$. The mixing coefficients in $\mathbf{p}$ can be approximated through simulations to the desired precision, similar to the method for previous test.

If σ^2 is known, the test with statistic T_1 is exact under the normal errors assumption, with larger values supporting the alternative hypothesis. In the more usual case of unknown model variance, we can construct a mixture of betas test statistic. Let $\hat{\boldsymbol{\eta}}_1 = \mathbf{E}_1\hat{\boldsymbol{\alpha}}_1 + \tilde{\boldsymbol{\Delta}}_0\hat{\mathbf{a}} + \mathbf{X}_c\hat{\boldsymbol{\beta}}_c + \mathbf{X}\hat{\boldsymbol{\beta}}$, and consider

$$\begin{aligned} T_2 &= \frac{1}{\sigma^2}\|\mathbf{y} - \hat{\boldsymbol{\eta}}_1\|^2 \\ &= \frac{1}{\sigma^2}\big\|(\mathbf{I} - \mathbf{P}_J)\mathbf{y}\big\|^2 \\ &= \frac{1}{\sigma^2}\big\|(\mathbf{I} - \mathbf{P}_J)(\boldsymbol{\eta} + \boldsymbol{\varepsilon})\big\|^2, \end{aligned}$$

where $\mathbf{P}_J$ is the projection matrix for the face of $\mathcal{C}$ on which the projection of $\mathbf{y}$ lands. Then, if $\bar{\boldsymbol{\eta}} = \mathbf{P}_J\boldsymbol{\eta}$,

$$\begin{aligned} T_2 &= \frac{1}{\sigma^2}\big\|(\mathbf{I} - \mathbf{P}_J)\boldsymbol{\varepsilon}\big\|^2 \\ &\quad + \frac{1}{\sigma^2}\big[\|\boldsymbol{\eta} - \bar{\boldsymbol{\eta}}\|^2 + 2\boldsymbol{\varepsilon}^\top(\boldsymbol{\eta} - \bar{\boldsymbol{\eta}})\big] \\ &= \frac{1}{\sigma^2}\big\|(\mathbf{I} - \mathbf{P}_J)\boldsymbol{\varepsilon}\big\|^2 + o_p(n), \end{aligned}$$

by the approximation error rate for the spline functions.

The term $\|(\mathbf{I} - \mathbf{P}_J)\boldsymbol{\varepsilon}\|^2/\sigma^2$ has a $\bar{\chi}^2_{(n-m_1-m-d_v):(n-d_v)}$ distribution, where d_v is the dimension of $\mathcal{V}$. However,

the exact mixing distribution cannot be found through simulations. Instead, we approximate the distribution of this term with a chi-squared random variable with degrees of freedom $n - d_{\text{obs}}$, where d_{obs} is the observed dimension of the face of $\mathcal{C}$ on which the projection of $\mathbf{y}$ lands. If n is large compared to the dimension of the cone, this approximation is reasonable. The statistics T_1 and T_2 are independent by orthogonality of the spaces into which they project (the two cones are contained in orthogonal subspaces), so $T = T_1/(T_1 + T_2)$ has approximately a mixture-of-betas distribution. Specifically, for $c \in (0, 1)$,

$$\mathrm{P}(T \le c) \approx \sum_{d=0}^{m_1} \mathrm{P}\big(B\big(d/2, (n - d_{\text{obs}})/2\big) \le c\big) p_d.$$

To compare this test with the tests in `gam` and `scam`, we generated 10,000 data sets from the model $y_i = f_1(t_{1i}) + f_2(t_{2i}) + \varepsilon_i$, for $i = 1, \ldots, n$, where $f_1(t) = 40(t_1 - 1/2)_+^4$, and $(\cdot)_+ = \max(\cdot, 0)$. The ε_i are independent standard normal, but we do not assume the model variance is known. We considered the test of $H_0 : f_2$ is constant versus $H_1 : f_2$ is increasing, using increasing regression splines to estimate both terms. The t_{1i} values were generated as uniform on the unit interval, and the t_{2i} values were the t_{1i} values plus a normal random variable with standard deviation 1/4, then scaled to be in the unit interval. The average correlation between the predictors is about 0.76. Simulation results are in Table 3, for three sample sizes and two underlying functions. The test sizes for `gam` are slightly over the target 0.05, but getting closer to the target as n increases. The `scam` test sizes with the default tuning parameters are getting substantially smaller than the target, and with the "matching" tuning parameters the test size is inflated. The `cgam` test has appropriate size and the best power; its main advantage over `gam` being that it is a one-sided test.

For the subcone test in the generalized linear model, we first fit the null hypothesis model, and retain the final $\boldsymbol{\xi}$ as a response vector and $\mathbf{W}$ as the weights, from the iteratively reweighted cone projection algorithm. We perform the test with the transformed response and the transformed cones. For the diabetes data used in the Section 2 example, we can test for the statistical significance of the skin-fold predictor. As expected, we get a large p-value (0.71) and conclude that this predictor is not important for predicting diabetes, when BMI, age and number of pregnancies are controlled for. We proceed to test each of the three remaining predictors, while controlling for the effects of the other two. We get small p-values for BMI and age (3×10^{-7} and 2×10^{-6}, resp.) but for the number of pregnancies, $p = 0.10$.

4. CONFIDENCE AND PREDICTION INTERVALS

Buja, Hastie and Tibshirani (1989) provide an overview of linear scatterplot smoothers, where the least-squares estimator can be expressed as $\hat{\boldsymbol{\eta}} = \mathbf{S}\mathbf{y}$ for an $n \times n$ matrix $\mathbf{S}$ that does not depend on $\mathbf{y}$. For regression splines, $\mathbf{S}$ is a projection matrix for a linear subspace spanned by the spline basis functions, but for kernel, smoothing spline and penalized spline methods, $\mathbf{S}$ is not idempotent. The true $\boldsymbol{\eta}$ is not assumed to be in the space of possible estimators, so there is a bias component of the fit as well as a variance component. Reducing the amount of smoothing by increasing the number of knots or decreasing the bandwidth, smoothing parameter, or penalty parameter, will decrease the bias because the estimator can get closer to the true function, but this is at the expense of increasing the variance. The optimal convergence rate Stone (1980) is attained by balancing the bias and variance. Similarly, Huang (2001) characterized the bias as approximation error and the variance as estimation error, and noted that increasing the dimension of the estimation space

TABLE 3

Proportions of 10,000 *data sets for which* $H_0 : f_2 = 0$ *is rejected, with test size* $\alpha = 0.05$. *The predictors* t_1 *and* t_2 *are positively correlated, and* $f_1(t_1) = 40(t_1 - 1/2)_+^4$

n	$f_2(t_2) \equiv 0$				$f_2(t_2) = \sqrt{t_2}$			
	`gam` (def.)	`scam` (def.)	`scam` (sp = 0)	`cgam` (def.)	`gam` (def.)	`scam` (def.)	`scam` (sp = 0)	`cgam` (def.)
100	0.065	0.058	0.182	0.051	0.192	0.186	0.199	0.309
200	0.058	0.034	0.158	0.049	0.271	0.203	0.245	0.449
400	0.056	0.018	0.133	0.050	0.431	0.295	0.469	0.640

decreases approximation error but increases estimation error, and showed that setting the two error rates equal gave optimal rates of convergence.

The estimated covariance matrix for the fit, $\widehat{\mathrm{var}}(\hat{\boldsymbol{\eta}}) = \hat{\sigma}^2\mathbf{S}\mathbf{S}^\top$ can be used (with a suitable estimator $\hat{\sigma}^2$ for the model variance σ^2) to construct the confidence intervals as

$$\hat{\eta}_i \pm z_{\alpha/2}\big[\widehat{\mathrm{var}}(\hat{\boldsymbol{\eta}})\big]_{ii}^{1/2},$$

but because this does not take into account the bias, the coverage probability will be low unless the estimator is undersmoothed. Zhou, Shen and Wolfe (1998) showed that for regression splines, the knots must increase at a faster-than-optimal rate to achieve asymptotically the desired confidence level.

The necessity for undersmoothing the fit to get good coverage is seen in other methods of nonparametric function estimation. The scatterplot in Figure 4 was simulated from the dashed curve with independent standard normal errors. The upper left plot shows the fit and pointwise confidence intervals using the popular `gam` function in the `R` package `mgcv` (Wood, 2018) with the default parameters, and the coverage probabilities estimated from 10,000 such fits are the top right. The default parameters do not provide good coverage where the curve is steep. In the second row, a fit to the same scatterplot is shown with options `k=40, sp=0.01` which provide more flexibility so that the approximation error or bias is minimal. The coverage probabilities are good, but there is spurious wiggling in the fit and the confidence bands are wider. In the final two rows are examples of constrained fits and confidence interval simulations results for the method proposed here. In the case where there are a priori shape assumptions about the regression function f, constrained regression splines obviate the problem of bias-variance trade-off. Many knots can be used with monotone or convex regression splines, because the constraints do not allow the spurious wiggling. Both good coverage and smaller-length intervals can be attained, because with more knots the bias is dominated by the variance. The default number of knots in `cgam` for $n = 100$ is $K = 7$; for $n = 200$, we use $K = 8$, and for $n = 400$, $K = 9$.

The projection $\hat{\boldsymbol{\eta}}$ lands on a face $\mathcal{F}_J$ of the cone $\mathcal{C}$. If $\mathbf{P}_J$ denotes the projection matrix for the linear space $\mathcal{L}_J$, then

$$\hat{\boldsymbol{\eta}} = \sum_J \mathbf{P}_J\mathbf{y}I\{\mathbf{y} \in \mathcal{C}_J\},$$

where $\mathcal{C}_J \subset \mathbb{R}^n$ is the set of $\mathbf{y}$ whose projection onto $\mathcal{C}$ lands on $\mathcal{F}_J$. To construct the pointwise confidence intervals at the design points, we propose

$$\widehat{\mathrm{var}}(\hat{\boldsymbol{\eta}}) = \hat{\sigma}^2 \sum_J \mathbf{P}_J \hat{p}_J, \tag{4.1}$$

where $\hat{p}_J$ is the estimated probability that $\mathbf{y}$ is in $\mathcal{C}_J$, and $\hat{\sigma}^2$ is defined in (3.2). The $\hat{p}_J$ values are obtained by simulating many normal random vectors with mean $\hat{\boldsymbol{\eta}}$ and covariance matrix $\hat{\sigma}^2\mathbf{I}$, and recording the resulting sets J.

For estimating the variance of $\hat{f}_1(t_1)+\cdots+\hat{f}_L(t_L)+\hat{g}_1(z_1)+\cdots+\hat{g}_R(z_R)+\mathbf{x}^\top\hat{\boldsymbol{\beta}}$ at arbitrary values of the predictors, we construct $\mathbf{w}$ to be a predictor vector as follows. Let $\mathbf{w}_c$ be a vector of length d_c, where d_c is the number of columns of $\mathbf{X}_c$, and the first element is one, and the other elements are t_ℓ, for ℓ such that f_ℓ is concave or convex. Let $\mathbf{b}_\ell = (b_{\ell,1}(x_\ell),\ldots,b_{\ell,m_\ell}(x_\ell))^\top$, where $b_{\ell,j}$ is the jth spline basis function defined for the predictor x_ℓ. Let $\mathbf{d}_r$ contain the row of $\boldsymbol{\Gamma}_r$ appropriate for the value z_r, and let $\mathbf{b} = [\mathbf{b}_1^\top,\ldots,\mathbf{b}_L^\top,\mathbf{d}_1^\top,\ldots,\mathbf{d}_R^\top]^\top$. Finally, $\mathbf{w} = (\mathbf{b}^\top,\mathbf{w}_c^\top,\mathbf{x}^\top)^\top$. Then $\hat{f}_1(t_1)+\cdots+\hat{f}_L(t_L)+\hat{g}_1(z_1)+\cdots+\hat{g}_R(z_R)+\mathbf{x}^\top\hat{\boldsymbol{\beta}} = \mathbf{w}^\top[\hat{\boldsymbol{\alpha}}|\hat{\boldsymbol{\beta}}_c|\hat{\boldsymbol{\beta}}]$, where $\hat{\boldsymbol{\alpha}}$, $\hat{\boldsymbol{\beta}}_c$ and $\hat{\boldsymbol{\beta}}$ are the coefficients for the projection of $\mathbf{y}$ onto the cone defined in (2.3).

Then $\widehat{\mathrm{var}}(\hat{\eta}(t_1,\ldots,t_L,z_1,\ldots,z_R,\mathbf{x})) = \hat{\sigma}^2\mathbf{w}^\top\hat{\mathbf{C}}\mathbf{w}$, where $\hat{\mathbf{C}} = \sum_J \mathbf{C}^{(J)}\hat{p}_J$ and $\mathbf{C}^{(J)}$ is an $M \times M$ matrix, where $M = m + d_c + p$ and $\mathbf{C}_J$ is constructed as follows. For $\mathbf{J} \in \{1,\ldots,m\}$, define $\boldsymbol{\Delta}_J$ to have the columns of $\boldsymbol{\Delta}$ (defined in (2.2)) that are indexed by the elements of J. Define $\mathbf{X}_J = [\boldsymbol{\Delta}_J|\mathbf{X}_c|\mathbf{X}]$, and let m_J be the number of elements in J. For $k,\ell \in \{1,\ldots,m\}$,

$$\mathbf{C}_{k,\ell}^{(J)} = \begin{cases} (\mathbf{X}_J^\top\mathbf{X}_J)^{-1}_{j_k,j_\ell} & \text{if } j_k\text{th element of } J \text{ is } k \\ & \text{and } j_\ell\text{th element of } J \text{ is } \ell, \\ 0 & \text{if } k \notin J \text{ or } \ell \notin J. \end{cases}$$

For $k \in \{1,\ldots,m\}$ and $\ell \in \{d_c+p+1,\ldots,M\}$,

$$\mathbf{C}_{k,\ell}^{(J)} = \begin{cases} (\mathbf{X}_J^\top\mathbf{X}_J)^{-1}_{j_k,\ell+m_j-m} & \text{if } j_k\text{th element of } J \text{ is } k, \\ 0 & \text{if } k \notin J. \end{cases}$$

For $k,\ell \in \{d_c+p+1,\ldots,M\}$,

$$\mathbf{C}_{k,\ell}^{(J)} = (\mathbf{X}_J^\top\mathbf{X}_J)^{-1}_{k+m_j-m,\ell+m_j-m}.$$

The proposed confidence interval for $\eta(t_1,\ldots,t_L,z_1,\ldots,z_R,\mathbf{x})$ is $\hat{f}_1(t_1)+\cdots+\hat{f}_L(t_L)+\hat{g}_1(z_1)+\cdots+\hat{g}_L(z_R)+\mathbf{z}^\top\hat{\boldsymbol{\beta}} \pm z_{\alpha/2}[\hat{\sigma}^2\mathbf{w}\hat{\mathbf{C}}\mathbf{w}^\top]^{1/2}$. This expression is consistent with (4.1); if $\mathbf{D} = [\boldsymbol{\Delta}|\mathbf{X}_0|\mathbf{Z}]$, then $\widehat{\mathrm{var}}(\hat{\boldsymbol{\eta}}) = \hat{\sigma}^2\mathbf{D}\hat{\mathbf{C}}\mathbf{D}^\top$.

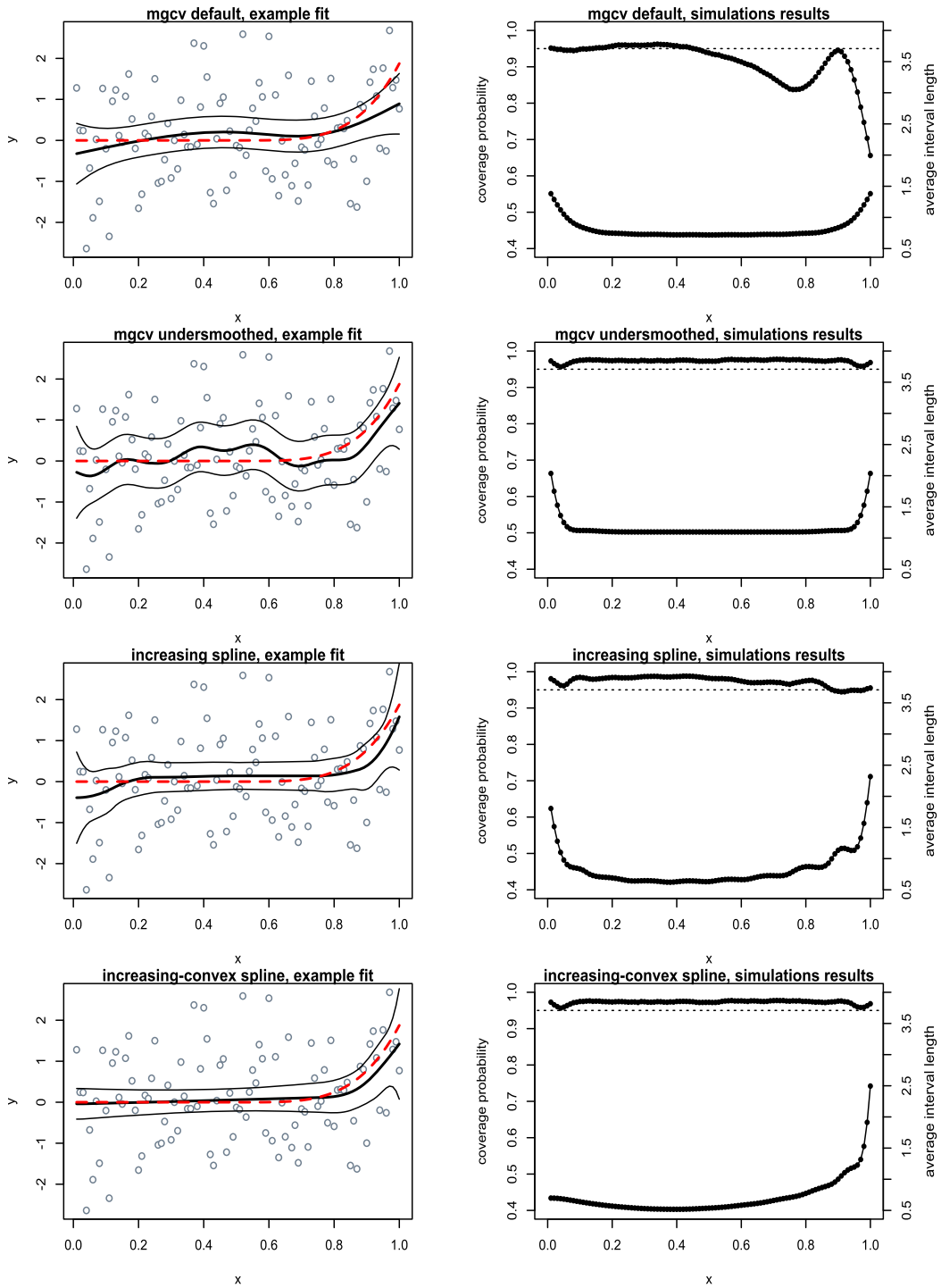

FIG. 4. *Example scatterplot smoothers with pointwise confidence bands*; *the first two rows use the* `gam` *function in the* `R` *package* `mgcv`. *The bottom two rows display the proposed constrained spline method. The data shown in the plots on the left were generated from the true regression function shown as the dashed curve with i.i.d. normal errors. The plots on the right show coverage probabilities* (*upper dots*, *left scale*) *and interval lengths* (*lower dots*, *right scale*).

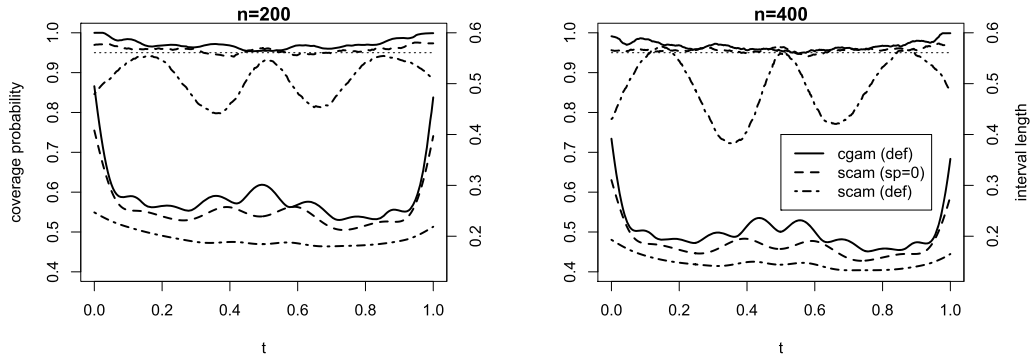

FIG. 5. *Coverage probabilities* (*top curves*, *left scale*) *and average interval lengths* (*bottom curves*, *right scale*) *from* 2000 *simulated data sets*.

For the generalized regression model, we use the terminal $\boldsymbol{\xi}$ and $\mathbf{W}$ in the iteratively reweighted projection, and transform both $\boldsymbol{\xi}$ and the cone $\mathcal{C}$ using the weights. We determine the confidence bands as above, within the transformed model, and transform back to $\boldsymbol{\eta}$ and $\boldsymbol{\mu}$.

To compare our method to that of `scam`, we ran simulations with the binomial model with a single continuous predictor. We use two sample sizes and the underlying true mean function

$$\mu(t) = 0.3 + \frac{0.5\exp(10t-5)}{1+\exp(10t-5)},$$

and again we use two sets of tuning parameters for the `scam` method. In Figure 5, the coverage probabilities and average interval lengths of the methods are compared. Although `scam` with the default tuning parameters produces the smallest intervals, the coverage is poor and does not improve with the larger sample size. The method with `sp=0` and the same number of knots as the default in `cgam` produces similar results to `cgam`.

Finally, we apply the method to the diabetes data set used for examples in the last two sections. Using BMI and age, the two predictors found to be significantly related to the probability of diabetes, we obtain 95% pointwise confidence bands for the probabilities for all values of BMI, and three values of age; these are shown in Figure 6.

5. SHAPE AND MODEL SELECTION

For a model with many predictors, some of them correlated, the shapes might not be known a priori. In fact, the scientific question may concern the shapes. Is the probability of a cure decreasing in age, while controlling for the effects of other predictors? Does the strength of a filament increase with conductivity, given the thickness? A model selection method may be considered where predictors are chosen along with the shapes of the relationships. The cone information criterion (CIC) given by Meyer (2013) was derived as an

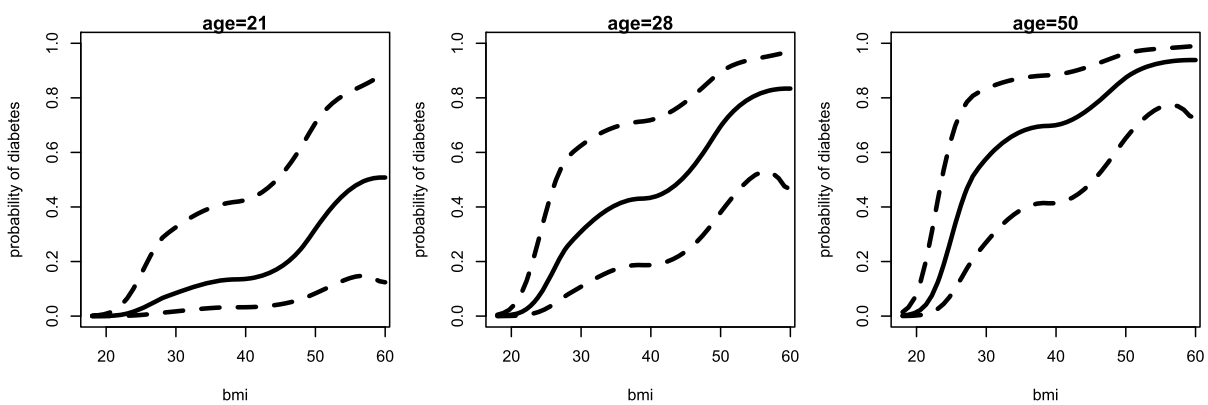

FIG. 6. *Estimated probability of diabetes in Pima Indian women, at three ages and all levels of BMI* (*solid curves*). *The* 95% *pointwise confidence bands for the probabilities are shown as dashed curves*.

estimate of the predicted squared error, and is given by

$$\text{CIC} = \psi(\boldsymbol{\eta}) + \log\left(\frac{2(E_0(D))}{n - 1.5E_0(D)} + 1\right),$$

where ψ is the negative log-likelihood function and $E_0(D)$ is the expected null degrees of freedom of the model. The second term of the CIC penalizes the complexity of the model, and hence it would be inappropriate to use observed degrees of freedom. The dimension of the observed face tends to be larger when the predictors are more strongly related to the response, and in the case where none of the constrained predictors are important, the observed face could simply be the linear space contained in the cone. Using observed face would result in larger penalties for better models. To determine the expected null degrees of freedom, simulations are necessary for each model in the list of possible models, where the response is generated from the appropriate model with no predictor effects. The observed degrees of freedom for each simulated response will be between d_1, the largest linear space contained in the cone, and d_2, the smallest linear space that contains the cone. The average of these observed degrees of freedom is averaged over many responses to get $E_0(D)$.

Several model or variable selection methods have been proposed involving isotonic or ordering restrictions, where the relationships are assumed to be monotone and variables can be chosen or not. Anraku (1999) proposed an information criterion specific to the order-restricted models, and Zhao and Peng (2002) used this for an isotonic dose-response problem, to determine at which levels the probability of a success increases. Peddada et al. (2003) proposed a method for selecting and clustering genes in gene expression data, based on traditional order-restricted models. Rueda (2013) proposed a variable selection method with unsmoothed isotonic regression, based on the observed degrees of freedom of the fit. A method similar to that proposed here, where shape is chosen from several different options, is found in Moisen et al. (2016) and is currently used for monitoring forest disturbances via Landsat signals. In that application, different forest disturbances produce different shapes in the Landsat time series.

To demonstrate the proposed CIC method, we look at the data set `trade.union` in the `R` package `SemiPar`. The response variable is the logarithm of the wages, for $n = 532$ workers in 1985. Three predictors are modeled with splines: age, years of experience and years of education. Each of the three spline functions can take any of the nine following shapes: flat, increasing, decreasing, convex, concave or any of the four combinations of monotonicity and convexity. There are seven nominal predictors: sex, occupation, whether or not the worker is in a union, whether the worker is married, whether the worker is in the south, race of worker and sector (manufacturing, construction, other). With the constrained splines, we can choose the shape: if each of the three "continuous" variables can be either flat, increasing, decreasing, convex, concave or one of the four combinations of monotonicity and convexity, we have $9^3 \times 2^7 = 93{,}312$ models. For each model, the calculation of $E_0(D)$ involves simulations from the null model, so the fitting of all models takes about two days. For this and larger data sets, we can use a genetic algorithm to search for the best model.

We define *phenotypes* as strings having a one-to-one map with the possible models. Each element of the string describes one of the variables, where the shapes are coded with numbers such as "1 = increasing," "2 = decreasing," etc., coding each of the eight shapes, and reserving the code "0" to mean "not in the model." For the nominal predictors, 1 means "in the model" and 0 means "not in the model." For example, the string $(5, 1, 4, 1, 1, 0, 0, 1, 0, 0)$ represents a model where $\boldsymbol{\eta}$ is increasing and convex in x_1, increasing in x_2 and concave in x_3. The remaining digits represent the nominal covariates.

To begin the genetic algorithm, we make an initial population of a few hundred phenotypes. For each phenotype, we can compute its *fitness*, that is, the negative CIC value for this model with the observed data. Then we let the population evolve. In each generation, there are the following steps. For the mutation step, we randomly choose a small number of phenotypes in the population, and for each, we randomly choose a gene and randomly change it to one of the allowed values. For the reproduction step, we randomly choose two members of the population. We randomly choose half the genes of one and combine these with the remaining genes of the other, to get an offspring. One of the old members of the population is replaced with the offspring. For the reproduction step, the more fit members of the population must be more likely to reproduce. To do this, tournament methods are common: choose two pairs and have each pair "fight" where the more fit are more likely to win; then the winners get to reproduce. Another way is to randomly assign genders to the phenotypes: all females but only the most fit males will

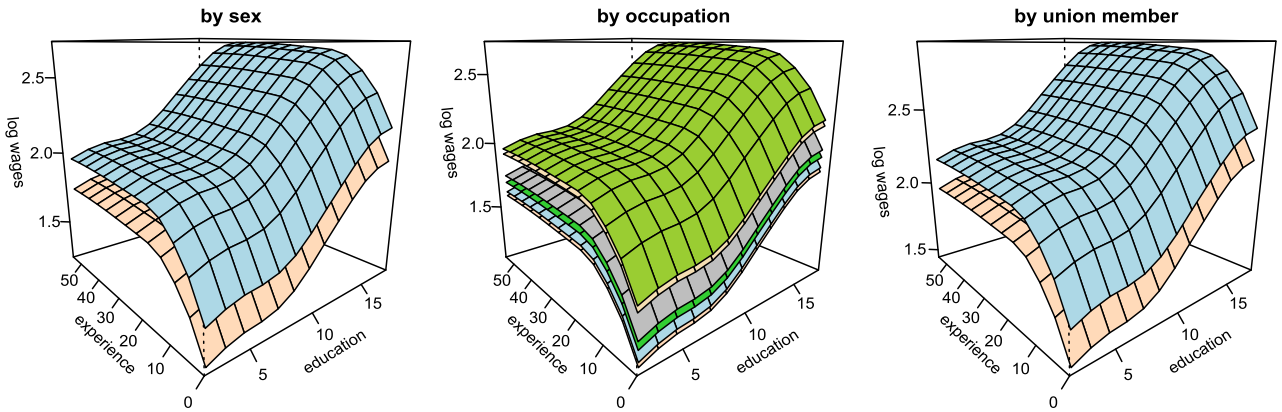

FIG. 7. *Best model fits for the* `trade.union` *data set.*

reproduce. By mimicking "real" evolution, the genetic algorithm converges to the best model.

There is no proof that a genetic algorithm always produces the optimal phenotype. The standard method for checking validity of the answer is to use another starting population and run the algorithm again. If a different answer is obtained, a larger starting population is needed. For the `trade.union` data, the genetic algorithm reliably gives the correct answer (minimizing the CIC over all models) with a population of 400 phenotypes and 10% mutation. The `ShapeSelect` function in the `R` package `cgam` allows the user to choose population size and mutation rate, as well as the set of shapes for each variable.

The genetic algorithm finds the best model for the trade union data in about half an hour. It chooses two of the continuous predictors, so that the fit to wages is increasing in years of education, increasing and concave in years of experience and flat in age. Three nominal variables chosen are sex, occupation and union membership. The fits for this model are shown in Figure 7, where the surface is shown with the effects of each of the three nominal predictors. It would be difficult to determine a priori some parametric regression shapes that would be appropriate, so nonparametric methods are valuable in this situation. However, fits that are only "smooth" are difficult to interpret. The set of possible shapes provides a rich field and minimal assumptions; it is appropriate if there is confidence that the true components of the relationship do not "wiggle." The results are interpretable in the context of the problem and provide information about the nature of the relationships.

6. DISCUSSION

Interest is growing for estimation and inference procedures where a minimum of assumptions is required. Linear models are convenient and inference methods are well known, but serious problems with model misspecification arise from making strong, invalid assumptions. Consider the data in Figure 8, where $y_i = f(x_i) + \varepsilon_i$ and $f(x) = 5e^{10x-5}/(1+e^{10x-5})$ with i.i.d. standard normal ε_i. A categorical covariate was also simulated, strongly related to x as shown in the plot. The scatterplot might suggest a linear relationship with different intercepts for the levels of the covariate. The results of the linear model, however, erroneously suggest that the covariate is a significant predictor, and ordinary residual analysis might not detect departures from the model. If we assume only that the relationship between y and x is smooth and increasing, the fit on the right in Figure 8 is obtained, with a large p-value for the covariate effect. In addition, the fitted function is quite close to the true mean. Especially when there are many predictors that are related to each other, parametric relationships are difficult to determine. Strong parametric assumptions that may be unjustified will lead to making the wrong conclusions, and the probability of making the wrong conclusion actually increases with the sample size. Nonparametric function estimation methods make only vague assumptions about the components, such as "smooth" and "increasing." Minimizing the assumptions will minimize the model misspecification errors. Using shape constraints as well as smoothing, either as a priori assumptions or through shape and variable selection, leads to flexible fits that are interpretable in the context of the problem.

If unsmoothed shape-constrained estimators are used for continuous predictors, the dimension of the small-

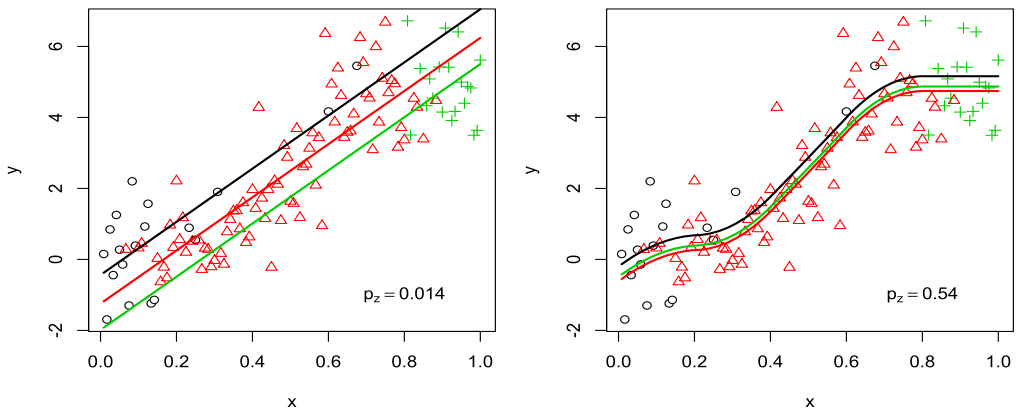

FIG. 8. *Data simulated from a sigmoidal regression function, with linear and spline fits. The value of the categorical covariate is indicated by the plot character.*

est linear space containing the cone can be as large as n, the sample size. Dimension reduction through regression splines is an important advantage as well as having a more satisfactory estimator when the smoothness assumption is valid. Modeling continuous predictors with regression splines allows a much smaller model dimension, which in turn allows for the inference methods such as the subcone test. The choice of knots, crucial for unconstrained splines, is rendered less important by the robustness of the constrained fits to knot choices. We can choose a "generous" number of knots for each component, because constraints disallow the "wiggling" usually associated with over-fitting in scatterplot smoothers.

For the `scam` methods, Pya and Wood (2015) implemented a Bayesian approach to obtain approximately normal distributions for the spline basis coefficients, attributed to Wahba (1983) and Silverman (1985). The methods coded in `cgam` use the cone formulation with formally derived inference methods, specifically for estimation with constraints.

APPENDIX: R CODE FOR THE SIMULATIONS

TABLE 1
`cgam`

```
nloop=10000
n=800
x=0:(n-1)/(n-1)
eta0=-1+60*(x-1/2)^4;eta0[x<1/2]=-1
pr1=0.05+0.85*x^4
pr2=0.4+0.55*x^(1/4)
pval=1:nloop
for(iloop in 1:nloop){
   z=1:n*0+3
   u=runif(n)
   z[u<pr1]=1;z[u>pr2]=2
   eta=eta0
#  eta[z==1]=eta[z==1]-1  (uncomment for power)
   mu=exp(eta)/(1+exp(eta))
   y=1:n*0
   y[runif(n)<mu]=1
   ans=cgam(y~s.incr(x)+factor(z),family="binomial")
   pval[iloop]=anova(ans)$coefficients1[3]
}
sum(pval<=0.05)/iloop
```

TABLE 1
scam

```
nloop=10000
n=100
x=0:(n-1)/(n-1)
eta0=-1+60*(x-1/2)^4;eta0[x<1/2]=-1
pr1=0.05+0.85*x^4
pr2=0.4+0.55*x^(1/4)
pval=1:nloop
for(iloop in 1:nloop){
   z=1:n*0+3
   u=runif(n)
   z[u<pr1]=1;z[u>pr2]=2
   eta=eta0
#  eta[z==1]=eta[z==1]-1
   mu=exp(eta)/(1+exp(eta))
   y=1:n*0
   y[runif(n)<mu]=1
#  ans=scam(y~s(x,bs='mpi',k=10)+factor(z),family="binomial",sp=0) #(match)
   ans=scam(y~s(x,bs='mpi')+factor(z),family="binomial")  ## (default)
   pval[iloop]=anova(ans)$pTerms.pv
}
sum(pval<=0.05)/iloop
```

TABLE 2
cgam

```
nloop=10000
n=800
x=0:(n-1)/(n-1)
eta0=30*(x-1/2)^4;eta0[x<1/2]=0
#eta0=1:n*0
pr1=0.05+0.85*x^4
pr2=0.4+0.55*x^(1/4)
pval=1:nloop
for(iloop in 1:nloop){
   z=1:n*0+3
   u=runif(n)
   z[u<pr1]=1
   z[u>pr2]=2
   eta=eta0
   eta[z==1]=eta[z==1]+2
   eta[z==2]=eta[z==2]-1
   mu=exp(eta)/(1+exp(eta))
   y=1:n*0
   y[runif(n)<mu]=1
   ans=cgam(y~s.incr(x)+factor(z),family="binomial",nsim=0)
   pval[iloop]=anova(ans)$coefficients2[3]
}
hist(as.numeric(pval))
sum(pval<0.05)/nloop
```

TABLE 2
scam

```
nloop=10000
n=800
x=0:(n-1)/(n-1)
#eta0=30*(x-1/2)^4;eta0[x<1/2]=0
eta0=1:n*0
pr1=0.05+0.85*x^4
pr2=0.4+0.55*x^(1/4)
pval=1:nloop
for(iloop in 1:nloop){
   z=1:n*0+3
   u=runif(n)
   z[u<pr1]=1
   z[u>pr2]=2
   eta=eta0
   eta[z==1]=eta[z==1]+2
   eta[z==2]=eta[z==2]-1
   mu=exp(eta)/(1+exp(eta))
   y=1:n*0
   y[runif(n)<mu]=1
   ans=scam(y~s(x,bs='mpi',k=10)+factor(z),family="binomial",sp=0)  ## match
#  ans=scam(y~s(x,bs='mpi')+factor(z),family="binomial")   ## default
   pval[iloop]=summary(ans)$s.pv
}
sum(pval[1:nloop]<0.05,na.rm=TRUE)/(nloop-sum(is.na(pval)))
```

TABLE 3
cgam

```
nloop=10000
n=100
x1=0:(n-1)/(n-1)
x2=x1+rnorm(n)/2
x2=(x2-min(x2))/(max(x2)-min(x2))
pval=1:nloop
eta1=40*(x1-1/2)^4;eta1[x1<1/2]=0
#eta2=1:n*0   ## null
eta2=sqrt(x2)   ## alternative
for(iloop in 1:nloop){
   y=eta1+eta2+rnorm(n)
   ans=cgam(y~s.incr(x1)+s.incr(x2),nsim=0)
   pval[iloop]=anova(ans)$coefficients2[2,3]
}
```

TABLE 3
scam

```
nloop=10000
n=100
x1=0:(n-1)/(n-1)
x2=x1+rnorm(n)/2
x2=(x2-min(x2))/(max(x2)-min(x2))
pval=1:nloop
eta1=40*(x1-1/2)^4;eta1[x1<1/2]=0
eta2=1:n*0   ## null
#eta2=sqrt(x2)   ## alternative
for(iloop in 1:nloop){
   y=eta1+eta2+rnorm(n)
#  ans=cgam(y~s.incr(x1)+s.incr(x2),nsim=0)
   ans=scam(y~s(x1,bs='mpi',k=8)+s(x2,bs='mpi',k=8),sp=c(0,0))
   pval[iloop]=summary(ans)$s.pv[2]
}
sum(pval<0.05)/nloop
```

ACKNOWLEDGMENTS

Partially supported by an NSF DMS grant. This work was partially funded by NSF-MMS-1533804. The author would like to express great appreciation for Dr. Xiyue Liao, who maintains the `cgam` package on CRAN.

REFERENCES

ANRAKU, K. (1999). An information criterion for parameters under a simple order restriction. *Biometrika* **86** 141–152. MR1688078

AYER, M., BRUNK, H. D., EWING, G. M., REID, W. T. and SILVERMAN, E. (1955). An empirical distribution function for sampling with incomplete information. *Ann. Math. Stat.* **26** 641–647. MR0073895

BACCHETTI, P. (1989). Additive isotonic models. *J. Amer. Statist. Assoc.* **84** 289–294. MR0999691

BARTHOLOMEW, D. J. (1959). A test of homogeneity for ordered alternatives. *Biometrika* **46** 36–48. MR0104312

BRUNK, H. D. (1955). Maximum likelihood estimates of monotone parameters. *Ann. Math. Stat.* **26** 607–616. MR0073894

BUJA, A., HASTIE, T. and TIBSHIRANI, R. (1989). Linear smoothers and additive models. *Ann. Statist.* **17** 453–555. MR0994249

CHEN, Y. and SAMWORTH, R. J. (2016). Generalized additive and index models with shape constraints. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **78** 729–754. MR3534348

CHENG, G. (2009). Semiparametric additive isotonic regression. *J. Statist. Plann. Inference* **139** 1980–1991. MR2497554

CHENG, G., ZHAO, Y. and LI, B. (2012). Empirical likelihood inferences for the semiparametric additive isotonic regression. *J. Multivariate Anal.* **112** 172–182. MR2957294

DU, P., PARMETER, C. F. and RACINE, J. S. (2013). Nonparametric kernel regression with multiple predictors and multiple shape constraints. *Statist. Sinica* **23** 1347–1371. MR3114717

FANG, Z. and MEINSHAUSEN, N. (2012). LASSO isotone for high-dimensional additive isotonic regression. *J. Comput. Graph. Statist.* **21** 72–91. MR2913357

HALL, P. and HUANG, L.-S. (2001). Nonparametric kernel regression subject to monotonicity constraints. *Ann. Statist.* **29** 624–647. MR1865334

HILDRETH, C. (1954). Point estimates of ordinates of concave functions. *J. Amer. Statist. Assoc.* **49** 598–619. MR0065093

HUANG, J. Z. (2001). Concave extended linear modeling: A theoretical synthesis. *Statist. Sinica* **11** 173–197. MR1820005

HUANG, J. (2002). A note on estimating a partly linear model under monotonicity constraints. *J. Statist. Plann. Inference* **107** 343–351. MR1927773

LIAO, X. and MEYER, M. C. (2018). cgam: Constrained generalized additive model.

MAMMEN, E. (1991). Estimating a smooth monotone regression function. *Ann. Statist.* **19** 724–740. MR1105841

MAMMEN, E. and YU, K. (2007). Additive isotone regression. In *Asymptotics: Particles, Processes and Inverse Problems. Institute of Mathematical Statistics Lecture Notes—Monograph Series* **55** 179–195. IMS, Beachwood, OH. MR2459939

MCCULLAGH, P. and NELDER, J. A. (1989). *Generalized Linear Models*, 2nd ed. *Monographs on Statistics and Applied Probability*. CRC Press, London. MR3223057

MEYER, M. C. (1999). An extension of the mixed primal-dual bases algorithm to the case of more constraints than dimensions. *J. Statist. Plann. Inference* **81** 13–31. MR1718405

MEYER, M. C. (2003). A test for linear versus convex regression function using shape-restricted regression. *Biometrika* **90** 223–232. MR1966562

MEYER, M. C. (2008). Inference using shape-restricted regression splines. *Ann. Appl. Stat.* **2** 1013–1033. MR2516802

MEYER, M. C. (2013). Semi-parametric additive constrained regression. *J. Nonparametr. Stat.* **25** 715–730. MR3174293

MEYER, M. C. (2018a). Constrained partial linear regression splines. *Statist. Sinica* **28** 277–292. MR3752261

MEYER, M. C. (2018b). Estimation and inference for regression surfaces using shape-constrained splines. Unpublished manuscript.

MEYER, M. C., HACKSTADT, A. J. and HOETING, J. A. (2011). Bayesian estimation and inference for generalised partial linear models using shape-restricted splines. *J. Nonparametr. Stat.* **23** 867–884. MR2854243

MEYER, M. C., KIM, S.-Y. and WANG, H. (2018). Convergence rates for constrained regression splines. *J. Statist. Plann. Inference* **193** 179–188. MR3713471

MEYER, M. and WOODROOFE, M. (2000). On the degrees of freedom in shape-restricted regression. *Ann. Statist.* **28** 1083–1104. MR1810920

MOISEN, G. G., MEYER, M. C., SCHROEDER, T., LIAO, X., SCHLEEWEIS, K., FREEMAN, E. and TONEY, C. (2016). Shape selection in landsat time series: A tool for monitoring forest dynamics. *Glob. Change Biol.* **22** 3518–3528.

MORTON-JONES, T., DIGGLE, P., PARKER, L., DICKINSON, H. and BINKS, K. (2000). Additive isotonic regression models in epidemiology. *Stat. Med.* **19** 849–859.

PEDDADA, S. D., LOBENHOFER, E. K., LI, L., AFSHARI, C. A., WEINBERG, C. R. and UMBACK, D. M. (2003). Gene selection and clustering for time-course and dose–response microarray experiments using order-restricted inference. *Bioinformatics* **19** 834–841.

PYA, N. and WOOD, S. N. (2015). Shape constrained additive models. *Stat. Comput.* **25** 543–559. MR3334416

RAMSAY, J. O. (1988). Monotone regression splines in action. *Statist. Sci.* **3** 425–461.

RAUBERTAS, R. F., LEE, C.-I. C. and NORDHEIM, E. V. (1986). Hypothesis tests for normal means constrained by linear inequalities. *Comm. Statist. Theory Methods* **15** 2809–2833. MR0855765

ROBERTSON, T., WRIGHT, F. T. and DYKSTRA, R. L. (1988). *Order Restricted Statistical Inference. Wiley Series in Probability and Mathematical Statistics: Probability and Mathematical Statistics*. Wiley, Chichester. MR0961262

RUEDA, C. (2013). Degrees of freedom and model selection in semiparametric additive monotone regression. *J. Multivariate Anal.* **117** 88–99. MR3053536

SEN, B. and MEYER, M. (2017). Testing against a linear regression model using ideas from shape-restricted estimation. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **79** 423–448. MR3611753

SILVERMAN, B. W. (1985). Some aspects of the spline smoothing approach to nonparametric regression curve fitting. *J. Roy. Statist. Soc. Ser. B* **47** 1–52. MR0805063

STONE, C. J. (1980). Optimal rates of convergence for nonparametric estimators. *Ann. Statist.* **8** 1348–1360. MR0594650

TANTIYASWASDIKUL, C. and WOODROOFE, M. B. (1994). Isotonic smoothing splines under sequential designs. *J. Statist. Plann. Inference* **38** 75–87. MR1256849

TUTZ, G. and LEITENSTORFER, F. (2007). Generalized smooth monotonic regression in additive modeling. *J. Comput. Graph. Statist.* **16** 165–188. MR2345751

VAN EEDEN, C. (1956). Maximum likelihood estimation of ordered probabilities. *Indag. Math. (N.S.)* **18** 444–455. MR0083859

VILLALOBOS, M. and WAHBA, G. (1987). Inequality-constrained multivariate smoothing splines with application to the estimation of posterior probabilities. *J. Amer. Statist. Assoc.* **82** 239–248. MR0883352

WAHBA, G. (1983). Bayesian "confidence intervals" for the cross-validated smoothing spline. *J. Roy. Statist. Soc. Ser. B* **45** 133–150. MR0701084

WOOD, S. N. (2018). mgcv: Mixed GAM computation vehicle with automatic smoothness estimation.

YU, K. (2014). On partial linear additive isotonic regression. *J. Korean Statist. Soc.* **43** 11–17. MR3173232

ZHAO, L. and PENG, L. (2002). Model selection under order restriction. *Statist. Probab. Lett.* **57** 301–306. MR1914007

ZHOU, S., SHEN, X. and WOLFE, D. A. (1998). Local asymptotics for regression splines and confidence regions. *Ann. Statist.* **26** 1760–1782. MR1673277