ROYAL STATISTICAL SOCIETY
DATA | EVIDENCE | DECISIONS

Journal of the Royal Statistical Society
Statistical Methodology
Series B

J. R. Statist. Soc. B (2016)
78, *Part* 4, *pp.* 729–754

Generalized additive and index models with shape constraints

解决各种形状约束的问题，单调性，凸性函数等

Yining Chen and Richard J. Samworth

University of Cambridge, UK

[Received April 2014. Final revision July 2015]

Summary. We study generalized additive models, with shape restrictions (e.g. monotonicity, convexity and concavity) imposed on each component of the additive prediction function. We show that this framework facilitates a non-parametric estimator of each additive component, obtained by maximizing the likelihood. The procedure is free of tuning parameters and under mild conditions is proved to be uniformly consistent on compact intervals. More generally, our methodology can be applied to generalized additive index models. Here again, the procedure can be justified on theoretical grounds and, like the original algorithm, has highly competitive finite sample performance. Practical utility is illustrated through the use of these methods in the analysis of two real data sets. Our algorithms are publicly available in the R package scar, short for shape-constrained additive regression.

Keywords: Generalized additive models; Index models; Non-parametric maximum likelihood estimation; Shape constraints

1. Introduction

Generalized additive models (GAMs) (Hastie and Tibshirani, 1986, 1990; Wood, 2006) have become an extremely popular tool for modelling multivariate data. They are designed to enjoy the flexibility of non-parametric modelling while avoiding the curse of dimensionality (Stone, 1986). Mathematically, suppose that we observe pairs $(\mathbf{X}_1, Y_1), \ldots, (\mathbf{X}_n, Y_n)$, where $\mathbf{X}_i = (X_{i1}, \ldots, X_{id})^{\mathrm{T}}$ is the predictor and Y_i is the real-valued response, for $i = 1, \ldots, n$. A GAM relates the predictor and the mean response $\mu_i = \mathbb{E}(Y_i|\mathbf{X}_i)$ through

$$g(\mu_i) = f(\mathbf{X}_i) = \sum_{j=1}^{d} f_j(X_{ij}) + c,$$

where g is a specified link function, and where the response Y_i conditional on $\mathbf{X}_i$ follows an exponential family distribution. Here $c \in \mathbb{R}$ is the intercept term and, for every $j = 1, \ldots, d$, the additive component function $f_j : \mathbb{R} \to \mathbb{R}$ is assumed to satisfy the identifiability constraint $f_j(0) = 0$. Our aim is to estimate the additive components $f_1, \ldots, f_d$ together with the intercept c on the basis of the given observations. Standard estimators are based on penalized spline-based methods (e.g. Wood (2004, 2008)), and involve tuning parameters whose selection is not always straightforward, especially if different additive components have different levels of smoothness, or if individual components have non-homogeneous smoothness.

In this paper, we propose a new approach, motivated by the fact that the additive components of f often follow certain common shape constraints such as monotonicity, convexity or

Address for correspondence: Richard J. Samworth, Statistical Laboratory, Centre for Mathematical Sciences, University of Cambridge, Wilberforce Road, Cambridge, CB3 0WB, UK.
E-mail: r.samworth@statslab.cam.ac.uk

Table 1. Different shape constraints and their corresponding labels

Shape constraint	*Label*	*Shape constraint*	*Label*	*Shape constraint*	*Label*
Linear	1	Monotone increasing	2	Monotone decreasing	3
Convex	4	Convex increasing	5	Convex decreasing	6
Concave	7	Concave increasing	8	Concave decreasing	9

concavity. For instance, monotone regression techniques have been used in biology to search for gene–gene interactions (Luss *et al.*, 2012), and in medicine to study the expression of a leukaemia antigen as a function of white blood cell count and DNA index (Schell and Singh, 1997). Economic theory dictates that utility functions are increasing and concave (Matzkin, 1991) and that the cost function of a standard perfectly competitive firm is increasing and convex (Aït-Sahalia and Duarte, 2003), whereas production functions are often assumed to be concave (Varian, 1984). In finance, theory restricts call option prices to be convex and decreasing functions of the strike price (Aït-Sahalia and Duarte, 2003); in stochastic control, value functions are often assumed to be convex (Keshavarz *et al.*, 2011).

The full list of constraints that we consider is given in Table 1, with each assigned a numerical label to aid our exposition. By assuming that each of $f_1, \ldots, f_d$ satisfies one of these nine shape restrictions, we show in Section 2 that it is possible to derive a non-parametric maximum likelihood estimator, which requires no choice of tuning parameters and which can be computed by using fast convex optimization techniques. In theorem 1, we prove that, under mild regularity conditions, it is uniformly consistent on compact intervals.

More generally, as we describe in Section 3, our approach can be applied to generalized additive index models (GAIMs), in which the predictor and the mean response $\mu_i = \mathbb{E}(Y_i|\mathbf{X}_i)$ are related through

$$g(\mu_i) = f^{\mathrm{I}}(\mathbf{X}_i) = f_1(\boldsymbol{\alpha}_1^{\mathrm{T}}\mathbf{X}_i) + \ldots + f_m(\boldsymbol{\alpha}_m^{\mathrm{T}}\mathbf{X}_i) + c, \tag{1}$$

where the value of $m \in \mathbb{N}$ is assumed known, where g is a known link function, and where the response $Y_i|\mathbf{X}_i$ again follows an exponential family distribution. Here, $\boldsymbol{\alpha}_1, \ldots, \boldsymbol{\alpha}_m \in \mathbb{R}^d$ are called the *projection indices*, $f_1, \ldots, f_m : \mathbb{R} \to \mathbb{R}$ are called the *ridge functions* (or, sometimes, *additive components*) of f^{I}, and $c \in \mathbb{R}$ is the intercept. Such index models have also been widely applied, especially in the area of econometrics (Li and Racine, 2007). When g is the identity function, the model is also known as projection pursuit regression (Friedman and Stuetzle, 1981); when $m = 1$, the model reduces to the single-index model (Ichimura, 1993). As for additive models, in some applications it is natural to impose shape constraints on the ridge functions; for instance, Foster *et al.* (2013) argued in favour of the use of monotone single-index models for analysing certain randomized clinical trial data. In other cases, as pointed out by Xu *et al.* (2014), shape restrictions are attractive as tractable non-parametric relaxations of linear models. Recent work by Kim and Samworth (2014) has shown that shape-restricted inference without further assumptions can lead to slow rates of convergence in higher dimensions. The additive or index structure therefore becomes particularly attractive in conjunction with shape constraints as an attempt to evade the curse of dimensionality. In Section 3, we extend our methodology and theory to this setting, allowing us to estimate simultaneously the projection indices, the ridge functions and the intercept.

The challenge of computing our estimators is taken up in Section 4, where our algorithms are described in detail. In Section 5, we summarize the results of a thorough simulation study

designed to compare the finite sample properties of `scar` with several alternative procedures. We conclude in Section 6 with two applications of our methodology to real data sets concerning doctoral publications in biochemistry and the decathlon. All proofs, as well as various auxiliary results, are given in the on-line supplementary material.

This paper contributes to the larger literature of regression in the presence of shape constraints. In the univariate case, and with the identity link function, the properties of shape-constrained least squares procedures are well understood, especially for the problem of isotonic regression. See, for instance, Brunk (1958, 1970) and Barlow *et al.* (1972). For the problem of univariate convex regression, see Hanson and Pledger (1976), Groeneboom *et al.* (2001, 2008) and Guntuboyina and Sen (2013). These references cover consistency, local and global rates of convergence, and computational aspects of the estimator. Hall and Huang (2001) proposed an alternative approach to univariate monotone regression, based on perturbing a kernel estimator. Banerjee (2007) and Banerjee (2009) studied monotone regression models in which the conditional distribution of a response given a covariate is assumed to come from a regular parametric model and an exponential family respectively. There have been several papers studying additive isotonic regression, including Bacchetti (1989), Morton-Jones *et al.* (2000), Tutz and Leitenstorfer (2007), Cai and Dunson (2007), Mammen and Yu (2007), Brezger and Steiner (2008), Cheng (2009), Cheng *et al.* (2012), Fang and Meinshausen (2012), Rueda (2013) and Yu (2014). The recent work of Meyer (2013a) develops similar methodology (but not theory) to ours in the Gaussian, non-index setting. The problem of GAMs with shape restrictions was also recently studied by Pya and Wood (2015), who proposed a penalized spline method that is compared with ours in Section 5; in particular, they consider the same set of constraints as in this paper. Meyer *et al.* (2011) investigated a Bayesian spline-based approach to the problem of GAMs, with a focus on the isotonic case.

2. Generalized additive models with shape constraints

2.1. Background

Recall that the density function of a natural exponential family (EF) distribution with respect to a reference measure (either Lebesgue measure on $\mathbb{R}$ or counting measure on $\mathbb{Q}$) can be written in the form

$$f_Y(y;\mu,\phi) = h(y,\phi)\exp\left[\frac{y\,g(\mu) - B\{g(\mu)\}}{\phi}\right],$$

where $\mu \in \mathcal{M} \subseteq \mathbb{R}$ and $\phi \in \Phi \subseteq (0,\infty)$ are the mean and dispersion parameters respectively. To simplify our discussion, we restrict our attention to the most commonly used natural EF distributions, namely the Gaussian, gamma, Poisson and binomial families, and take g to be the canonical link function. Expressions for g and the (strictly convex) log-partition function B for the different exponential families can be found in Table 2. The corresponding distributions are denoted by $\mathrm{EF}_{g,B}(\mu,\phi)$, and we write $\mathrm{dom}(B) = \{\eta \in \mathbb{R} : B(\eta) < \infty\}$ for the domain of B. As a convention, for the binomial family, the response is scaled to take values in $\{0, 1/T, 2/T, \ldots, 1\}$ for some known $T \in \mathbb{N}$.

If $(\mathbf{X}_1, Y_1), \ldots, (\mathbf{X}_n, Y_n)$ are independent and identically distributed pairs taking values in $\mathbb{R}^d \times \mathbb{R}$, with $Y_i|\mathbf{X}_i \sim \mathrm{EF}_{g,B}[g^{-1}\{f(\mathbf{X}_i)\}, \phi]$ for some *prediction function* $f : \mathbb{R}^d \to \mathrm{dom}(B)$, then the (conditional) log-likelihood of f can be written as

$$\frac{1}{\phi}\sum_{i=1}^{n}[Y_i f(\mathbf{x}_i) - B\{f(\mathbf{x}_i)\}] + \sum_{i=1}^{n}\log\{h(Y_i,\phi)\}.$$

Table 2. Exponential family distributions, their corresponding canonical link functions, log-partition functions and mean and dispersion parameter spaces

Exponential family	$g(\mu)$	$B(\eta)$	$dom(B)$	$\mathcal{M}$	Φ
Gaussian	μ	$\eta^2/2$	$\mathbb{R}$	$\mathbb{R}$	$(0,\infty)$
Gamma	$-\mu^{-1}$	$-\log(-\eta)$	$(-\infty,0)$	$(0,\infty)$	$(0,\infty)$
Poisson	$\log(\mu)$	$\exp(\eta)$	$\mathbb{R}$	$(0,\infty)$	$\{\mathbf{1}\}$
Binomial	$\log\{\mu/(1-\mu)\}$	$\log\{1+\exp(\eta)\}$	$\mathbb{R}$	$(0,1)$	$\{1/T\}$

Since we are only interested in estimating f, it suffices to consider the *scaled partial log-likelihood*

$$\bar{\ell}_n(f) \equiv \bar{\ell}_n(f;(\mathbf{X}_1,Y_1),\ldots,(\mathbf{X}_n,Y_n)) := \frac{1}{n}\sum_{i=1}^{n}[Y_i f(\mathbf{X}_i) - B\{f(\mathbf{X}_i)\}] \equiv \frac{1}{n}\sum_{i=1}^{n}\ell_i(f), \tag{2}$$

say.

2.2. Maximum likelihood estimation under shape constraints

Let $\bar{\mathbb{R}} = \mathbb{R}\cup\{-\infty,\infty\}$ denote the extended real line. To guarantee the existence of our estimator, defined in expression (3) below, it turns out to be convenient to extend the definition of each ℓ_i (and therefore $\bar{\ell}_n$) to all $f:\mathbb{R}^d\to\bar{\mathbb{R}}$, which we do as follows.

(a) For the gamma family, if $f(\mathbf{X}_i)\geqslant 0$, then we take $\ell_i(f)=-\infty$.
(b) If $f(\mathbf{X}_i)=-\infty$, then we set $\ell_i(f)=\lim_{a\to-\infty}\{Y_i a - B(a)\}$. Similarly, if $f(\mathbf{X}_i)=\infty$ (in the Gaussian, Poisson or binomial setting), then we define $\ell_i(f)=\lim_{a\to\infty}\{Y_i a - B(a)\}$. Note that both limits always exist in $\bar{\mathbb{R}}$.

For any shape vector $\mathbf{L}_d=(l_1,\ldots,l_d)^{\mathrm{T}}\in\{1,2,\ldots,9\}^d$, let $\mathcal{F}=\mathcal{F}^{\mathbf{L}_d}$ denote the set of functions $f:\mathbb{R}^d\to\mathbb{R}$ of the form

$$f(\mathbf{x}) = \sum_{j=1}^{d} f_j(x_j) + c$$

for $\mathbf{x}=(x_1,\ldots,x_d)^{\mathrm{T}}\in\mathbb{R}^d$, where, for every $j=1,\ldots,d$, $f_j:\mathbb{R}\to\mathbb{R}$ is a function obeying the shape restriction indicated by label l_j and satisfying $f_j(0)=0$, and where $c\in\mathbb{R}$. Whenever f has such a representation, we write $f\sim^{\mathcal{F}}(f_1,\ldots,f_d,c)$, and call $\mathbf{L}_d$ the *shape vector*. The pointwise closure of $\mathcal{F}$ is defined as

$$\mathrm{cl}(\mathcal{F}) = \{f:\mathbb{R}^d\to\bar{\mathbb{R}} \mid \exists\, f^1, f^2,\ldots\in\mathcal{F} \text{ such that } \lim_{k\to\infty} f^k(\mathbf{x}) = f(\mathbf{x}) \text{ for every } \mathbf{x}\in\mathbb{R}^d\}.$$

For a specified shape vector $\mathbf{L}_d$, we define the shape-constrained maximum likelihood estimator (SCMLE) as

$$\hat{f}_n \in \operatorname*{arg\,max}_{f\in\mathrm{cl}(\mathcal{F})} \bar{\ell}_n(f). \tag{3}$$

Our reason for maximizing over $\mathrm{cl}(\mathcal{F})$ rather than $\mathcal{F}$ in the definition of $\hat{f}_n$ is a technical convenience: as we see from proposition 1 below, it ensures that a maximizer always exists. This would be false in certain special cases if instead we only maximized over $\mathcal{F}$ (see example 1 in Appendix A for such an instance), though, from the paragraph immediately following theorem

1 below, we see that the distinction is not too important. Like many other shape-restricted regression estimators, $\hat{f}_n$ is not unique in general. However, as can be seen from the second part of proposition 1, the value of $\hat{f}_n$ is uniquely determined at $\mathbf{X}_1, \ldots, \mathbf{X}_n$.

Proposition 1. The set $\hat{S}_n := \arg\max_{f \in \mathrm{cl}(\mathcal{F})} \bar{\ell}_n(f)$ is non-empty. Moreover, all elements of $\hat{S}_n$ agree at $\mathbf{X}_1, \ldots, \mathbf{X}_n$.

In fact, as can be seen from the proof of proposition 1, if the EF distribution is Gaussian or gamma, then $\hat{S}_n \cap \mathcal{F} \neq \emptyset$. Whenever $\hat{S}_n \cap \mathcal{F} \neq \emptyset$, it contains an element for which each additive component is piecewise linear, and any solution obtained from our algorithm in Section 4.1 has this piecewise linear property.

2.3. Consistency of the shape-constrained maximum likelihood estimator

In this subsection, we show the consistency of $\hat{f}_n$ in a random-design setting. We shall impose the following assumptions.

Assumption 1. $(\mathbf{X}, Y), (\mathbf{X}_1, Y_1), (\mathbf{X}_2, Y_2), \ldots$ is a sequence of independent and identically distributed (IID) pairs taking values in $\mathbb{R}^d \times \mathbb{R}$.

Assumption 2. The random vector $\mathbf{X}$ has a Lebesgue density with support $\mathbb{R}^d$.

Assumption 3. Fix $\mathbf{L}_d \in \{1, 2, \ldots, 9\}^d$. Suppose that $Y|\mathbf{X} \sim \mathrm{EF}_{g,B}[g^{-1}\{f_0(\mathbf{X})\}, \phi_0]$, where $f_0 \in \mathcal{F}$ and $\phi_0 \in (0, \infty)$ denote the true prediction function and dispersion parameter respectively.

Assumption 4. f_0 is continuous on $\mathbb{R}^d$.

We are now in the position to state our main consistency result.

Theorem 1. Assume assumptions 1–4. Then, for every $a_0 \geqslant 0$,

$$\sup_{\hat{f}_n \in \hat{S}_n} \sup_{\mathbf{x} \in [-a_0, a_0]^d} |\hat{f}_n(\mathbf{x}) - f_0(\mathbf{x})| \to 0 \qquad \text{almost surely}$$

as $n \to \infty$.

Under assumption 3, we may write $f_0 \sim^{\mathcal{F}} (f_{0,1}, \ldots, f_{0,d}, c_0)$. When the assumptions of theorem 1 hold, in particular assumption 2, we see from the proof of theorem 1 that, for any $a_0 > 0$, with probability 1, for sufficiently large n, any $\hat{f}_n \in \hat{S}_n$ can be written in the form $\hat{f}_n(\mathbf{x}) = \sum_{j=1}^d \hat{f}_{n,j}(x_j) + \hat{c}_n$ for $\mathbf{x} = (x_1, \ldots, x_d)^{\mathrm{T}} \in [-a_0, a_0]^d$, where $\hat{f}_{n,j}$ satisfies the shape constraint l_j and $\hat{f}_{n,j}(0) = 0$ for each $j = 1, \ldots, d$.

We now turn to estimation of the additive components and the intercept. Recall that, whenever we write $f \sim^{\mathcal{F}} (f_1, \ldots, f_d, c)$, we insist that $f_j(0) = 0$ for all j, and refer to it as an identifiability condition. This is because, if we also had $f \sim^{\mathcal{F}} (\tilde{f}_1, \ldots, \tilde{f}_d, \tilde{c})$, then we would have

$$\sum_{j=1}^d f_j(x_j) + c = \sum_{j=1}^d \tilde{f}_j(x_j) + \tilde{c} \tag{4}$$

for all $(x_1, \ldots, x_d)^{\mathrm{T}} \in \mathbb{R}^d$. By considering this equation at $\mathbf{0}$ and at points of the form $\{(x_1, 0, \ldots, 0)^{\mathrm{T}} : x_1 \in \mathbb{R}\}, \ldots, \{(0, \ldots, 0, x_d)^{\mathrm{T}} : x_d \in \mathbb{R}\}$, we could then conclude that $\tilde{c} = c$ and $\tilde{f}_j = f_j$. Note, however, that, as observed by Meyer (2013a), if we only know that the equality (4) holds at a finite set of points $(x_{1,1}, \ldots, x_{1,d})^{\mathrm{T}}, \ldots, (x_{n,1}, \ldots, x_{n,d})^{\mathrm{T}} \in \mathbb{R}^d$, then we do not even necessarily know that $\tilde{f}_j(x_{i,j}) = f_j(x_{i,j})$ for all i, j. Nevertheless, the following corollary establishes the important fact that each additive component (as well as the intercept term) is estimated consistently by the SCMLE.

Corollary 1. Assume assumptions 1–4. Then, for any $a_0 \geqslant 0$,

$$\sup_{\hat{f}_n \in \hat{S}_n} \Big\{ \sum_{j=1}^{d} \sup_{x_j \in [-a_0, a_0]} |\hat{f}_{n,j}(x_j) - f_{0,j}(x_j)| + |\hat{c}_n - c_0| \Big\} \to 0 \qquad \text{almost surely}$$

as $n \to \infty$.

3. Generalized additive index models with shape constraints

3.1. The generalized additive index model and its identifiability

Recall that, in the GAIM, the real-valued response Y_i and the predictor $\mathbf{X}_i = (X_{i1}, \ldots, X_{id})^{\mathrm{T}}$ are related through equation (1), where g is a known link function, and where, conditionally on $\mathbf{X}_i$, the response Y_i has a known EF distribution with mean parameter $g^{-1}\{f^{\mathrm{I}}(\mathbf{X}_i)\}$ and dispersion parameter ϕ.

Let $\mathbf{A} = (\boldsymbol{\alpha}_1, \ldots, \boldsymbol{\alpha}_m)$ denote the $d \times m$ *index matrix*, where $m \leqslant d$, and let $f(\mathbf{z}) = \sum_{j=1}^{m} f_j(z_j) + c$ for $\mathbf{z} = (z_1, \ldots, z_m)^{\mathrm{T}} \in \mathbb{R}^m$, so the prediction function can be written as $f^{\mathrm{I}}(\mathbf{x}) = f(\mathbf{A}^{\mathrm{T}}\mathbf{x})$ for $\mathbf{x} = (x_1, \ldots, x_d)^{\mathrm{T}} \in \mathbb{R}^d$. As in Section 2, we impose shape constraints on the ridge functions by assuming that $f_j : \mathbb{R} \to \mathbb{R}$ satisfies the shape constraint with label $l_j \in \{1, 2, \ldots, 9\}$, for $j = 1, \ldots, m$, and consider the shape vector $\mathbf{L}_m \in \{1, \ldots, 9\}^m$ to be fixed throughout, so that in this section, as well as in the corresponding proofs and algorithm, $\mathcal{F} = \mathcal{F}^{\mathbf{L}_m}$.

The interesting question of the identifiability of GAIMs was settled recently by Yuan (2011). To discuss the issue of identifiability carefully, we first state the main result of Yuan (2011) and then relate it to our shape-constrained setting. It is convenient to say that $(\boldsymbol{\alpha}_1, \ldots, \boldsymbol{\alpha}_m, f_1, \ldots, f_m, c)$ satisfy the additive index model assumptions on $(-a, a)^d$ with $a \in (0, \infty]$ if the following conditions hold.

Condition 1.

(a) $f_1, \ldots, f_m : (-a, a) \to \mathbb{R}$ are non-zero functions with $f_j(0) = 0$ for $j = 1, \ldots, m$, and $c \in \mathbb{R}$.
(b) $\|\boldsymbol{\alpha}_j\|_1 = 1$ for $j = 1, \ldots, m$, where $\|\cdot\|_1$ denotes the ℓ_1-norm.
(c) The first non-zero entry of $\boldsymbol{\alpha}_j$ is positive for every j.
(d) There is at most one linear ridge function in $f_1, \ldots, f_m$; if f_j is linear, then $\boldsymbol{\alpha}_k^{\mathrm{T}} \boldsymbol{\alpha}_j = 0$ for every $k \neq j$.
(e) There is at most one quadratic ridge function in $f_1, \ldots, f_m$.
(f) $\mathbf{A} = (\boldsymbol{\alpha}_1, \ldots, \boldsymbol{\alpha}_m)$ has full column rank m.
(g) Each f_j is either continuous at 0 or is monotonic or bounded on a subinterval of $(-a, a)$. (This condition is not stated in Yuan (2011), but it is implicitly assumed in his lemma 2, which states that the only solutions of Cauchy's functional equation are linear.)

Theorem 2 (Yuan (2011), theorem 1). Assume that both $(\boldsymbol{\alpha}_1, \ldots, \boldsymbol{\alpha}_m, f_1, \ldots, f_m, c)$ and $(\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_q, g_1, \ldots, g_q, \tilde{c})$ satisfy the additive index model assumptions on $(-a, a)^d$, and that

$$\sum_{j=1}^{m} f_j(\boldsymbol{\alpha}_j^{\mathrm{T}}\mathbf{x}) + c = \sum_{\ell=1}^{q} g_\ell(\boldsymbol{\beta}_\ell^{\mathrm{T}}\mathbf{x}) + \tilde{c} \tag{5}$$

for all $\mathbf{x} \in (-a, a)^d$. Then $\tilde{c} = c$ and $q = m$ and there is a permutation π of $\{1, \ldots, m\}$ such that $\boldsymbol{\beta}_j = \boldsymbol{\alpha}_{\pi(j)}$ and $g_j = f_{\pi(j)}$.

Although theorem 2 requires several conditions, most of these are very natural to rule out trivial lack of identifiability problems. The most interesting conditions are 1(d) and 1(e). As

explained in equation (2.4) of Yuan (2011), even if there is only one linear function, further restrictions are required because

$$f_1(\boldsymbol{\alpha}_1^{\mathrm{T}}\mathbf{x}) + b_2\boldsymbol{\alpha}_2^{\mathrm{T}}\mathbf{x} = \{f_1(\boldsymbol{\alpha}_1^{\mathrm{T}}\mathbf{x}) - b_1\boldsymbol{\alpha}_1^{\mathrm{T}}\mathbf{x}\} + \|b_1\boldsymbol{\alpha}_1 + b_2\boldsymbol{\alpha}_2\|_1 \left(\frac{b_1\boldsymbol{\alpha}_1 + b_2\boldsymbol{\alpha}_2}{\|b_1\boldsymbol{\alpha}_1 + b_2\boldsymbol{\alpha}_2\|_1}\right)^{\mathrm{T}} \mathbf{x}$$

for all non-zero scalars b_1 and b_2. This is why we require $\boldsymbol{\alpha}_k^{\mathrm{T}}\boldsymbol{\alpha}_j = 0$ for every $k \neq j$ whenever f_j is linear. As shown in proposition 1 of Yuan (2011), condition 1(e) is necessary, because, if there are two quadratic ridge functions, then their corresponding projection indices are not identifiable. This fact is closely related to the identifiability of independent component analysis models (Eriksson and Koivunen, 2004; Samworth and Yuan, 2012).

In our setting, we say that $(\boldsymbol{\alpha}_1, \ldots, \boldsymbol{\alpha}_m, f_1, \ldots, f_m, c)$ satisfy the shape-constrained additive index model assumptions on $(-a, a)^d$ with shape vector $\mathbf{L}_m = (l_1, \ldots, l_m)^{\mathrm{T}}$ if the following condition 1(a′) and conditions 1(b)–1(f) hold.

Condition 1(a′). For $j = 1, \ldots, m$, $f_j : (-a, a) \to \mathbb{R}$ satisfies shape constraint l_j, is non-zero and satisfies $f_j(0) = 0$, and $c \in \mathbb{R}$.

Note that condition 1(a′) ensures that condition 1(g) holds. It follows immediately from theorem 2 that if both $(\boldsymbol{\alpha}_1, \ldots, \boldsymbol{\alpha}_m, f_1, \ldots, f_m, c)$ and $(\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_q, g_1, \ldots, g_q, \tilde{c})$ satisfy the shape-constrained additive index model assumptions on $(-a, a)^d$ with shape vectors $(l_1, \ldots, l_m)^{\mathrm{T}}$ and $(l'_1, \ldots, l'_q)^{\mathrm{T}}$ respectively, and if equation (5) holds for all $\mathbf{x} \in (-a, a)^d$, then $\tilde{c} = c$, $q = m$ and there is a permutation π of $\{1, \ldots, m\}$ such that $\boldsymbol{\beta}_j = \boldsymbol{\alpha}_{\pi(j)}$ and $g_j = f_{\pi(j)}$. Thus, in this sense, if conditions 1(a′) and 1(b)–1(f) hold, then the GAIM is identifiable.

3.2. Generalized additive index model estimation

Let $\mathbf{A}_0 = (\boldsymbol{\alpha}_{0,1}, \ldots, \boldsymbol{\alpha}_{0,m})$ denote the true index matrix. For $\mathbf{x} = (x_1, \ldots, x_d)^{\mathrm{T}} \in \mathbb{R}^d$, let

$$f_0^{\mathrm{I}}(\mathbf{x}) = f_{0,1}(\boldsymbol{\alpha}_{0,1}^{\mathrm{T}}\mathbf{x}) + \ldots + f_{0,m}(\boldsymbol{\alpha}_{0,m}^{\mathrm{T}}\mathbf{x}) + c_0$$

be the true prediction function, and write $f_0(\mathbf{z}) = \Sigma_{j=1}^m f_{0,j}(z_j) + c_0$ for $\mathbf{z} = (z_1, \ldots, z_m)^{\mathrm{T}} \in \mathbb{R}^m$. Again we restrict our attention to the common EF distributions listed in Table 2 and take g to be the corresponding canonical link function. In the light of the identifiability discussion in Section 3.1, it makes sense to define the class of index matrices associated with a given shape vector $\mathbf{L}_m$ as

$$\mathcal{A} = \{\mathbf{A} = (\boldsymbol{\alpha}_1, \ldots, \boldsymbol{\alpha}_m) \in \mathbb{R}^{d \times m} | \mathbf{A} \text{ satisfies conditions 1(b) and 1(c), and, if } \exists\, k \in \{1, \ldots, m\} \text{ such that } l_k = 1, \text{then } \boldsymbol{\alpha}_j^{\mathrm{T}}\boldsymbol{\alpha}_k = 0 \text{ for every } j \neq k\}.$$

We can now consider the set of shape-constrained additive index functions given by

$$\mathcal{G} = \{f^{\mathrm{I}} : \mathbb{R}^d \to \mathbb{R} | f^{\mathrm{I}}(\mathbf{x}) = f(\mathbf{A}^{\mathrm{T}}\mathbf{x}), \text{ with } f \in \mathcal{F} \text{ and } \mathbf{A} \in \mathcal{A}\}.$$

By analogy with the approach that was adopted in Section 2, a natural idea is to seek to maximize the scaled partial log-likelihood $\bar{\ell}_n$ over the pointwise closure of $\mathcal{G}$. As part of this process, and writing $\bar{\ell}_n(f; \mathbf{A}) = \bar{\ell}_n\{f; (\mathbf{A}^{\mathrm{T}}\mathbf{X}_1, Y_1), \ldots, (\mathbf{A}^{\mathrm{T}}\mathbf{X}_n, Y_n)\}$ for the scaled partial *index log-likelihood*, we would like to find a $d \times m$ matrix in $\mathcal{A}$ that maximizes

$$\Lambda_n(\mathbf{A}) = \sup_{f \in \mathcal{F}} \bar{\ell}_n(f; \mathbf{A}), \tag{6}$$

where the dependence of $\Lambda_n(\cdot)$ on $\mathbf{L}_m$ and $(\mathbf{X}_1, Y_1), \ldots, (\mathbf{X}_n, Y_n)$ is suppressed for notational convenience. We argue, however, that this strategy has two drawbacks. First, if $m \geqslant 2$ and

$\mathbf{L}_m \notin \mathcal{L}_m := \{1,4,5,6\}^m \cup \{1,7,8,9\}^m$, then, in certain cases, maximizing $\Lambda_n(\mathbf{A})$ over $\mathcal{A}$ can lead to a perfect fit to the data; second, the function $\Lambda_n(\cdot)$ need not be upper semicontinuous, so we are not guaranteed that a maximizer exists. These phenomena are illustrated in examples 2 and 3 in Appendix A.

As a result, certain modifications are required for our shape-constrained approach to be successful in the context of GAIMs. To deal with the first issue when $\mathbf{L}_m \notin \mathcal{L}_m$, we optimize $\Lambda_n(\cdot)$ over the subset of matrices

$$\mathcal{A}^\delta = \{\mathbf{A} \in \mathcal{A} : \lambda_{\min}(\mathbf{A}^{\mathrm{T}}\mathbf{A}) \geqslant \delta\}$$

for some predetermined $\delta > 0$, where $\lambda_{\min}(\cdot)$ denotes the smallest eigenvalue of a non-negative definite matrix. Other strategies are also possible. For example, when $\mathbf{L}_m = (2,\ldots,2)^{\mathrm{T}}$, the 'perfect fit' phenomenon can be avoided by considering only matrices with non-negative entries (see Section 6.2 below).

To address the second issue, we shall show that, given $f_0^{\mathrm{I}} \in \mathcal{G}$ satisfying the identifiability conditions, to obtain a consistent estimator, it is sufficient to find $\tilde{f}_n^{\mathrm{I}}$ from the set

$$\begin{aligned}\tilde{S}_n \in \{ f^{\mathrm{I}} : \mathbb{R}^d \to \mathbb{R} \,|\, f^{\mathrm{I}}(\mathbf{x}) = f(\mathbf{A}^{\mathrm{T}}\mathbf{x}), \text{ with } f \in \mathcal{F}; \text{ if } \mathbf{L}_m \in \mathcal{L}_m \text{ or } m = 1, \text{ then } \mathbf{A} \in \mathcal{A}; \\ \text{otherwise, } \mathbf{A} \in \mathcal{A}^\delta; \bar{\ell}_n(f;\mathbf{A}) \geqslant \bar{\ell}_n(f_0;\mathbf{A}_0)\},\end{aligned} \tag{7}$$

for some $\delta \in (0, \lambda_{\min}(\mathbf{A}_0^{\mathrm{T}}\mathbf{A}_0)]$. We write $\tilde{f}_n^{\mathrm{I}}(\mathbf{x}) = \tilde{f}_n(\tilde{\mathbf{A}}_n^{\mathrm{T}}\mathbf{x})$, where $\tilde{\mathbf{A}}_n = (\tilde{\boldsymbol{\alpha}}_{n,1},\ldots,\tilde{\boldsymbol{\alpha}}_{n,m}) \in \mathcal{A}$ or $\mathcal{A}^\delta$ is the estimated index matrix and $\tilde{f}_n(\mathbf{z}) = \sum_{j=1}^m \tilde{f}_{n,j}(z_j) + \tilde{c}_n$ is the estimated additive function with $\tilde{f}_{n,j}$ satisfying the shape constraint l_j and $\tilde{f}_{n,j}(0) = 0$ for every $j = 1,\ldots,m$. We call $\tilde{f}_n^{\mathrm{I}}$ the *shape-constrained additive index estimator* (SCAIE), and write $\tilde{\mathbf{A}}_n$ and $\tilde{f}_{n,1},\ldots,\tilde{f}_{n,m}$ respectively for the corresponding estimators of the index matrix and ridge functions.

When there is a maximizer of the function $\Lambda_n(\cdot)$ over $\mathcal{A}$ or $\mathcal{A}^\delta$, the set $\tilde{S}_n$ is certainly non-empty; in fact, $\tilde{S}_n$ is always non-empty, in view of the following proposition and the argument below it.

Proposition 2. The function $\Lambda_n(\cdot)$ is lower semicontinuous.

If a maximizer of $\Lambda_n(\cdot)$ does not exist, there must exist some $\mathring{\mathbf{A}}_n$ such that $\Lambda_n(\mathring{\mathbf{A}}_n) > \Lambda_n(\mathbf{A}_0)$. It then follows from proposition 2 that

$$\liminf_{\mathbf{A} \to \mathring{\mathbf{A}}_n} \Lambda_n(\mathbf{A}) \geqslant \Lambda_n(\mathring{\mathbf{A}}_n) > \Lambda_n(\mathbf{A}_0) \geqslant \bar{\ell}_n(f_0;\mathbf{A}_0),$$

so any $\mathbf{A}$ sufficiently close to $\mathring{\mathbf{A}}_n$ yields a prediction function in $\tilde{S}_n$. A stochastic search algorithm can be employed to find such matrices; see Section 4.2 for details.

3.3. Consistency of shape-constrained additive index estimator

In this subsection, we show the consistency of $\tilde{f}_n^{\mathrm{I}}$ and $\tilde{\mathbf{A}}_n$ under a random-design setting. In addition to assumptions 1 and 2, we require the following conditions.

Condition 2. Fix $\mathbf{L}_m \in \{1,2,\ldots,9\}^m$. The true prediction function f_0^{I} and the corresponding index matrix $\mathbf{A}_0$ satisfy the shape-constrained additive index model conditions 1(a′) and 1(b)–1(f) on $\mathbb{R}^d$ with shape vector $\mathbf{L}_m$. In particular, if the ridge function $f_{0,j}$ is linear, then $l_j = 1$.

Condition 3. Suppose that $Y|\mathbf{X} \sim \mathrm{EF}_{g,B}[g^{-1}\{f_0^{\mathrm{I}}(\mathbf{X})\}, \phi_0]$, where $\phi_0 \in (0,\infty)$ is the true dispersion parameter.

Condition 4. f_0^{I} is continuous on $\mathbb{R}^d$.

Theorem 3. Assume assumptions 1 and 2 as well as conditions 2–4. Then, provided that $\delta \leqslant \lambda_{\min}(\mathbf{A}_0^{\mathrm{T}}\mathbf{A}_0)$ when $\mathbf{L}_m \notin \mathcal{L}_m$, we have for every $a_0 \geqslant 0$ that

$$\sup_{\tilde{f}_n^{\mathrm{I}} \in \tilde{S}_n} \sup_{\mathbf{x} \in [-a_0, a_0]^d} |\tilde{f}_n^{\mathrm{I}}(\mathbf{x}) - f_0^{\mathrm{I}}(\mathbf{x})| \to 0 \qquad \text{almost surely}$$

as $n \to \infty$.

To obtain consistency, whenever $\mathbf{L}_m \notin \mathcal{L}_m$, the practitioner requires an *a priori* assumption of a lower bound for $\lambda_{\min}(\mathbf{A}_0^{\mathrm{T}}\mathbf{A}_0)$, and this lower bound plays a role in the computation. However, it is quite natural to want projection indices not to be too highly correlated to aid interpretability. In practice, choosing δ too small can result in overfitting when n is small, but in our experience the method is relatively insensitive to quite a wide range of choices of δ.

Consistency of the estimated index matrix and the ridge functions is established in the next corollary. Some care, however, is required to define an appropriate notion of distance between the estimator and estimand. In particular, note that the ordering of the additive components is arbitrary. This means that we can only hope to estimate the set of projection indices, and not their ordering, so it is only by allowing a permutation of the components that we can guarantee that the estimated quantities are asymptotically close to their population counterparts. Nevertheless, each projection index has a corresponding ridge function, so in permuting the ordering we must ensure to apply the same permutation to both the projection indices and the ridge functions. Similarly, since we are also unable to estimate the zero entries of the index matrix exactly, we should also allow the sign of each column of the index matrix to be flipped. This discussion leads us to make the following definition: if $f^{\mathrm{I}}(\mathbf{x}) = f(\mathbf{A}^{\mathrm{T}}\mathbf{x})$ with $\mathbf{A} = (\boldsymbol{\alpha}_1, \ldots, \boldsymbol{\alpha}_m)$ and $f \sim^{\mathcal{F}} (f_1, \ldots, f_m, c)$, and $g^{\mathrm{I}}(\mathbf{x}) = g(\mathbf{B}^{\mathrm{T}}\mathbf{x})$ with $\mathbf{B} = (\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_m)$ and $g \sim^{\mathcal{F}} (g_1, \ldots, g_m, c')$, then for $a > 0$ we set

$$d_a(f^{\mathrm{I}}, g^{\mathrm{I}}) = \min_{\pi \in \mathcal{P}_m} \min_{\epsilon_1, \ldots, \epsilon_m \in \{-1,1\}} \sum_{j=1}^{m} \left\{ \|\epsilon_j \boldsymbol{\alpha}_{\pi(j)} - \boldsymbol{\beta}_j\|_1 + \sup_{z_j \in [-a,a]} |f_{\pi(j)}(\epsilon_j z_j) - g_j(z_j)| \right\} + |c - c'|,$$

where $\mathcal{P}_m$ denotes the set of permutations of $\{1, \ldots, m\}$.

Corollary 2. Assume assumptions 1 and 2 and conditions 2–4. Then, provided that $\delta \leqslant \lambda_{\min}(\mathbf{A}_0^{\mathrm{T}}\mathbf{A}_0)$ when $\mathbf{L}_m \notin \mathcal{L}_m$, we have for every $a_0 \geqslant 0$ that

$$\sup_{\tilde{f}_n^{\mathrm{I}} \in \tilde{S}_n} d_{a_0}(\tilde{f}_n^{\mathrm{I}}, f_0^{\mathrm{I}}) \to 0 \qquad \text{almost surely}$$

as $n \to \infty$.

4. Computational aspects

4.1. Computation of shape-constrained maximum likelihood estimator

Throughout this subsection, we fix the shape vector $\mathbf{L}_d = (l_1, \ldots, l_d)^{\mathrm{T}}$, the EF distribution and the values of the observations, and present an algorithm for computing the SCMLE described in Section 2. The algorithm is quite different from backfitting algorithms that are commonly used for fitting (generalized) additive models (Breiman and Friedman, 1985; Buja *et al.*, 1989; Mammen and Park, 2006; Yu *et al.*, 2008), but we found it to be somewhat faster in practice for our purposes.

Our aim is to reformulate the problem as a convex program in terms of basis functions and to apply an active set algorithm (Nocedal and Wright, 2006). Such algorithms have recently become

Table 3. Pseudocode of the active set algorithm for computing the SCMLE

Step 1
Initialization—outer loop: sort $\{X_i\}_{i=1}^n$ co-ordinate by co-ordinate; define the initial working set as $\mathcal{S}_1 = \{(0, j) | j \in \{1, \ldots, d_1\}\} \cup \{(1, j) | l_j \in \{4, 7\}\}$; in addition, define the set of potential elements as

$$\mathcal{S} = \{(i, j) : i = 1, \ldots, n, j = d_1 + 1, \ldots, d\};$$

set the iteration count $k = 1$

Step 2
Initialization—inner loop: if $k > 1$, set $\mathbf{w}^* = \mathbf{w}^{(k-1)}$

Step 3
Unrestricted GLM: solve the following unrestricted GLM problem by using iteratively reweighted least squares (IRLS):

$$\frac{1}{n} \sum_{h=1}^{n} \Big[Y_h \Big\{ \sum_{(i,j) \in \mathcal{S}_k} w_{ij} g_{ij}(X_{hj}) + w_{00} \Big\} - B \Big\{ \sum_{(i,j) \in \mathcal{S}_k} w_{ij} g_{ij}(X_{hj}) + w_{00} \Big\} \Big],$$

where, for $k > 1$, $\mathbf{w}^*$ is used as a warm start; store its solution in $\mathbf{w}^{(k)}$ (with zero weights for the elements outside $\mathcal{S}_k$)

Step 4
Working set refinement: if $k = 1$ or if $w_{ij} > 0$ for every $(i, j) \in \mathcal{S}_k \backslash \mathcal{S}_1$, go to *step 5*; otherwise, define respectively the moving ratio p and the set of elements to drop as

$$p = \min_{\substack{(i,j) \in \mathcal{S}_k \backslash \mathcal{S}_1: \\ w_{ij} \leqslant 0}} \frac{w^*_{ij}}{w^*_{ij} - w_{ij}}, \quad \mathcal{S}_- = \left\{ (i, j) : (i, j) \in \mathcal{S}_k \backslash \mathcal{S}_1, w_{ij} \leqslant 0, \frac{w^*_{ij}}{w^*_{ij} - w_{ij}} = p \right\},$$

set $\mathcal{S}_k := \mathcal{S}_k \backslash \mathcal{S}_-$, overwrite $\mathbf{w}^*$ by $\mathbf{w}^* := (1 - p)\mathbf{w}^* + p\mathbf{w}^{(k)}$ and go to *step 3*

Step 5
Derivative evaluation: for every $(i, j) \in \mathcal{S}$, compute

$$D^{(k)}_{i,j} = \frac{\partial \psi_n}{\partial w_{ij}}(\mathbf{w}^{(k)})$$

Step 6
Working set enlargement: write $\mathcal{S}_+ = \arg\max_{(i,j) \in \mathcal{S}} D^{(k)}_{i,j}$ for the enlargement set, with maximum $D^{(k)} = \max_{(i,j) \in \mathcal{S}} D^{(k)}_{i,j}$; if $D^{(k)} \leqslant 0$ (or some other criteria are met if the EF distribution is non-Gaussian, e.g. $D^{(k)} < \epsilon_{\mathrm{IRLS}}$ for some predetermined small $\epsilon_{\mathrm{IRLS}} > 0$), *stop* the algorithm and go to *step 7*; otherwise, pick any single-element subset $\mathcal{S}^*_+ \subseteq \mathcal{S}_+$, let $\mathcal{S}_{k+1} = \mathcal{S}_k \cup \mathcal{S}^*_+$, set $k := k + 1$ and go back to *step 2*

Step 7
Output: for every $j = 1, \ldots, d$, set $\hat{f}_{n,j}(x_j) = \Sigma_{\{i:(i,j) \in \mathcal{S}_k\}} w^{(k)}_{ij} g_{ij}(x_j)$; take $\hat{c}_n = w^{(k)}_{00}$; finally, return the SCMLE as $\hat{f}_n(\mathbf{x}) = \Sigma_{j=1}^d \hat{f}_{n,j}(x_j) + \hat{c}_n$

popular for computing various shape-constrained estimators. For instance, Groeneboom *et al.* (2008) used a version, which they called the 'support reduction algorithm' in the one-dimensional convex regression setting; Dümbgen and Rufibach (2011) applied another variant to compute the univariate log-concave maximum likelihood density estimator. Recently, Meyer (2013b) developed a 'hinge' algorithm for quadratic programming, which can also be viewed as a variant of the active set algorithm.

Without loss of generality, we assume in what follows that only the first d_1 components ($d_1 \leqslant d$) of f_0 are linear, i.e. $l_1 = \ldots = l_{d_1} = 1$ and $(l_{d_1+1}, \ldots, l_d)^{\mathrm{T}} \in \{2, \ldots, 9\}^{d-d_1}$. Furthermore, we assume that the order statistics $\{X_{(i),j}\}_{i=1}^n$ of $\{X_{ij}\}_{i=1}^n$ are distinct for every $j = d - d_1 + 1, \ldots, d$. For $\mathbf{x} = (x_1, \ldots, x_d)^{\mathrm{T}} \in \mathbb{R}^d$, define the basis functions $g_{0j}(x_j) = x_j$ for $j = 1, \ldots, d_1$ and, for $i = 1, \ldots, n$,

$$g_{ij}(x_j) = \begin{cases} \mathbb{1}_{\{X_{(i),j} \leqslant x_j\}} - \mathbb{1}_{\{X_{(i),j} \leqslant 0\}}, & \text{if } l_j = 2, \\ \mathbb{1}_{\{x_j < X_{(i),j}\}} - \mathbb{1}_{\{0 < X_{(i),j}\}}, & \text{if } l_j = 3, \\ (x_j - X_{(i),j})\mathbb{1}_{\{X_{(i),j} \leqslant x_j\}} + X_{(i),j}\mathbb{1}_{\{X_{(i),j} \leqslant 0\}}, & \text{if } l_j = 4 \text{ or } l_j = 5, \\ (X_{(i),j} - x_j)\mathbb{1}_{\{x_j \leqslant X_{(i),j}\}} - X_{(i),j}\mathbb{1}_{\{0 \leqslant X_{(i),j}\}}, & \text{if } l_j = 6, \\ (X_{(i),j} - x_j)\mathbb{1}_{\{X_{(i),j} \leqslant x_j\}} - X_{(i),j}\mathbb{1}_{\{X_{(i),j} \leqslant 0\}}, & \text{if } l_j = 7 \text{ or } l_j = 9, \\ (x_j - X_{(i),j})\mathbb{1}_{\{x_j \leqslant X_{(i),j}\}} + X_{(i),j}\mathbb{1}_{\{0 \leqslant X_{(i),j}\}}, & \text{if } l_j = 8. \end{cases}$$

Note that all the basis functions given above are zero at the origin. Let $\mathcal{W}$ denote the set of weight vectors

$$\mathbf{w} = (w_{00}, w_{01}, \ldots, w_{0d_1}, w_{1(d_1+1)}, \ldots, w_{n(d_1+1)}, \ldots, w_{1d}, \ldots, w_{nd})^{\mathrm{T}} \in \mathbb{R}^{n(d-d_1)+d_1+1}$$

satisfying

$$\begin{aligned} w_{ij} &\geqslant 0, \qquad \text{for every } i = 1, \ldots, n \text{ and every } j \text{ with } l_j \in \{2, 3, 5, 6, 8, 9\}, \\ w_{ij} &\geqslant 0, \qquad \text{for every } i = 2, \ldots, n \text{ and every } j \text{ with } l_j \in \{4, 7\}. \end{aligned}$$

To compute the SCMLE, it suffices to consider prediction functions of the form

$$f^{\mathbf{w}}(\mathbf{x}) = w_{00} + \sum_{j=1}^{d_1} w_{0j} g_{0j}(x_j) + \sum_{j=d_1+1}^{d} \sum_{i=1}^{n} w_{ij} g_{ij}(x_j)$$

subject to $\mathbf{w} \in \mathcal{W}$. Our optimization problem can then be reformulated as maximizing

$$\psi_n(\mathbf{w}) = \bar{\ell}_n\{f^{\mathbf{w}}; (\mathbf{X}_1, Y_1), \ldots, (\mathbf{X}_n, Y_n)\}$$

over $\mathbf{w} \in \mathcal{W}$. Note that ψ_n is a concave (but not necessarily strictly concave) function. Since

$$\sup_{\mathbf{w} \in \mathcal{W}} \bar{\ell}_n(f^{\mathbf{w}}) = \bar{\ell}_n(\hat{f}_n),$$

our goal here is to find a sequence $(\mathbf{w}^{(k)})$ such that $\psi_n(\mathbf{w}^{(k)}) \to \sup_{\mathbf{w} \in \mathcal{W}} \bar{\ell}_n(f^{\mathbf{w}})$ as $k \to \infty$. In Table 3, we give the pseudocode for our active set algorithm for finding the SCMLE, which is implemented in the R package `scar` (Chen and Samworth, 2014).

We outline below some implementation details.

(a) Iteratively reweighted least squares (IRLS): step 3 solves an unrestricted GLM problem by applying IRLS. Since the canonical link function is used here, IRLS is simply the Newton–Raphson method. If the EF distribution is Gaussian, then IRLS gives the exact solution of the problem in just one iteration. Otherwise, there is no closed form expression for the solution, so a threshold ϵ_{IRLS} must be picked to serve as part of the stopping criterion. Note that here IRLS can be replaced by other methods that solve GLM problems, though we found that IRLS offers competitive timing performance.

(b) *Fast computation of the derivatives*: although step 5 appears at first sight to require $O(n^2 d)$ operations, it can actually be completed with only $O(nd)$ operations by exploiting some nice recurrence relations. Define the 'nominal' residuals at the kth iteration by

$$r_i^{(k)} = Y_i - \mu_i^{(k)}, \qquad \text{for } i = 1, \ldots, n,$$

where $\mu_i^{(k)} = g^{-1}\{f^{\mathbf{w}^{(k)}}(\mathbf{X}_i)\}$ are the fitted mean values at the kth iteration. Then

$$\frac{\partial \psi_n}{\partial w_{ij}}(\mathbf{w}^{(k)}) = \frac{1}{n} \sum_{u=1}^{n} r_u^{(k)} g_{ij}(X_{uj}).$$

Table 4. Pseudocode of the stochastic search algorithm for computing the SCAIE

Step 1
Initialization: let N denote the total number of stochastic searches; set $k=1$

Step 2
Draw random matrices: draw a $d \times m$ random matrix $\mathbf{A}^k$ by initially choosing the entries to be IID $N(0,1)$ random variables; for each column of $\mathbf{A}^k$, if there is a $j \in \{1,\ldots,m\}$ such that $l_j = 1$, subtract its projection to the jth column of $\mathbf{A}^k$ so that condition 1(d) is satisfied; then normalize each column so that conditions 1(b) and 1(c) are satisfied

Step 3
Rejection sampling: if $\mathbf{L}_m \notin \mathcal{L}_m$ and $\lambda_{\min}\{(\mathbf{A}^k)^{\mathrm{T}}\mathbf{A}^k\} < \delta$, then go back to *step 2*; otherwise, if $k < N$, set $k := k+1$ and go to *step 2*

Step 4
Evaluation of Λ_n: for every $k = 1,\ldots,N$, compute $\Lambda_n(A^k)$ using the active set algorithm described in Table 3

Step 5
Index matrix estimation—1: let $\mathbf{A}^* \in \arg\max_{1 \leqslant k \leqslant N} \Lambda_n(\mathbf{A}^k)$; set $\tilde{A}_n = \mathbf{A}^*$

Step 6
Index matrix estimation—2 (optional): treat $\mathbf{A}^*$ as a warm start and apply another optimization strategy to find $\mathbf{A}^{**}$ in a neighbourhood of $\mathbf{A}^*$ such that $\Lambda_n(\mathbf{A}^{**}) > \Lambda_n(\mathbf{A}^*)$; if such $\mathbf{A}^{**}$ can be found, set $\tilde{\mathbf{A}}_n = \mathbf{A}^{**}$

Step 7
Output: use the active set algorithm described in Table 3 to find

$$\tilde{f}_n \in \arg\max_{f \in \mathrm{cl}(\mathcal{F})} \bar{\ell}_n(f; \tilde{\mathbf{A}}_n);$$

finally, output the SCAIE as $\tilde{f}_n^{\mathrm{I}}(\mathbf{x}) = \tilde{f}_n(\tilde{\mathbf{A}}_n^{\mathrm{T}}\mathbf{x})$

For simplicity, we suppress henceforth the superscript k. Now fix j and reorder the pairs (r_i, X_{ij}) as $(r_{(1)}, X_{(1),j}), \ldots, (r_{(n)}, X_{(n),j})$ such that $X_{(1),j} \leqslant \ldots \leqslant X_{(n),j}$ (note that this is performed in step 1). Furthermore, define

$$R_{i,j} = \begin{cases} \sum_{u=1}^{i} r_{(u)}, & \text{if } l_j \in \{2,4,5,6\}, \\ -\sum_{u=1}^{i} r_{(u)}, & \text{if } l_j \in \{3,7,8,9\}, \end{cases}$$

for $i = 1,\ldots,n$, where we suppress the explicit dependence of $r_{(u)}$ on j in the notation. We have $R_{n,j} = 0$ because of the presence of the intercept w_{00}. The following recurrence relations can be derived by simple calculation.

(i) For $l_j \in \{2,3\}$, we have $D_{1,j} = 0$ and $nD_{i,j} = -R_{i-1,j}$ for $i = 2,\ldots,n$.

(ii) For $l_j \in \{4,5,7,9\}$, the initial condition is $D_{n,j} = 0$, and

$$nD_{i,j} = nD_{i+1,j} - R_{i,j}(X_{(i+1),j} - X_{(i),j}), \qquad \text{for } i = n-1,\ldots,1.$$

(iii) For $l_j \in \{6,8\}$, the initial condition is $D_{1,j} = 0$, and

$$nD_{i,j} = nD_{i-1,j} + R_{i-1,j}(X_{(i),j} - X_{(i-1),j}), \qquad \text{for } i = 2,\ldots,n.$$

Therefore, the complexity of step 5 in our implementation is $O(nd)$.

(c) *Convergence*: if the EF distribution is Gaussian, then it follows from theorem 1 of Groeneboom *et al.* (2008) that our algorithm converges to the optimal solution after finitely many iterations. In general, the convergence of this active set strategy depends on the following two aspects.
   (i) *Convergence of IRLS*: the convergence of the Newton–Raphson method in step 3 depends on the starting values. It is not guaranteed without step size optimization; see Jørgensen (1983). However, starting from the second iteration, each subsequent IRLS step is performed by starting from the previous well-approximated solution, which typically makes the method work well.
   (ii) *Accuracy of IRLS*: if IRLS gives the *exact* solution every time, then $\psi_n(\mathbf{w}^{(k)})$ increases at each iteration. In particular, one can show that, at the kth iteration, the new element $\mathcal{S}_+^*$ added into the working set in step 6 will remain in the working set $\mathcal{S}_{k+1}$ after the $(k+1)$th iteration. However, since IRLS returns only an approximate solution, there is no guarantee that the above-mentioned phenomenon continues to hold. One way to resolve this issue is to reduce the tolerance ϵ_{IRLS} if $\psi_n(\mathbf{w}^{(k)}) \leqslant \psi_n(\mathbf{w}^{(k-1)})$, and to redo the computations for both the previous and the current iteration.

Here we terminate our algorithm in step 6 if either $\psi_n(\mathbf{w}^{(k)})$ is non-increasing or $D^{(k)} < \epsilon_{\mathrm{IRLS}}$. In our numerical work, we did not encounter convergence problems, even outside the Gaussian setting.

4.2. Computation of shape-constrained additive index estimator

The computation of the SCAIE can be divided into two parts.

(a) For a given $\mathbf{A}$, find $f \in \mathrm{cl}(\mathcal{F})$ that maximizes $\bar{l}_n(f; \mathbf{A})$ by using the algorithm in Table 3 but with $\mathbf{A}^{\mathrm{T}}\mathbf{X}_i$ replacing $\mathbf{X}_i$. Denote the corresponding maximum value by $\Lambda_n(\mathbf{A})$.
(b) For a given lower semicontinuous function Λ_n on $\mathcal{A}$ or $\mathcal{A}^\delta$ as appropriate, find a maximizing sequence $(\mathbf{A}^k)$ in this set.

The second part of this algorithm solves a finite dimensional optimization problem. Possible strategies include the differential evolution method (Price *et al.*, 2005; Dümbgen *et al.*, 2011) or a stochastic search strategy (Dümbgen *et al.*, 2013) described below. In Table 4, we give the pseudocode for computing the SCAIE. We note that step 4 of the stochastic search algorithm is parallelizable.

5. Simulation study

To analyse the empirical performance of the SCMLE and SCAIE, we ran a simulation study focusing on the running time and the predictive performance. Throughout this section, we took $\epsilon_{\mathrm{IRLS}} = 10^{-8}$.

5.1. Generalized additive models with shape restrictions

For each data set, we took $\mathbf{X}_1, \ldots, \mathbf{X}_n \sim^{\mathrm{IID}} U[-1, 1]^d$. The following three problems were considered.

(a) In problem 1, $d = 4$. We set $\mathbf{L}_4 = (4, 4, 4, 4)^{\mathrm{T}}$ and $f_0(\mathbf{x}) = |x_1| + |x_2| + |x_3|^3 + |x_4|^3$.
(b) In problem 2, $d = 4$. We set $\mathbf{L}_4 = (5, 5, 5, 5)^{\mathrm{T}}$ and

$$f_0(\mathbf{x}) = x_1 \mathbb{1}_{\{x_1 \geqslant 0\}} + x_2 \mathbb{1}_{\{x_2 \geqslant 0\}} + x_3^3 \mathbb{1}_{\{x_3 \geqslant 0\}} + x_4^3 \mathbb{1}_{\{x_4 \geqslant 0\}}.$$

(c) In problem 3, $d = 8$. We set $\mathbf{L}_8 = (4, 4, 4, 4, 5, 5, 5, 5)^{\mathrm{T}}$ and

$$f_0(\mathbf{x}) = |x_1| + |x_2| + |x_3|^3 + |x_4|^3 + x_5 \mathbb{1}_{\{x_5 \geqslant 0\}} + x_6 \mathbb{1}_{\{x_6 \geqslant 0\}} + x_7^3 \mathbb{1}_{\{x_7 \geqslant 0\}} + x_8^3 \mathbb{1}_{\{x_8 \geqslant 0\}}.$$

For each of these three problems, we considered three types of EF distribution.

(i) *Gaussian*: for $i = 1, \ldots, n$, conditionally on $\mathbf{X}_i$, draw $Y_i \sim N\{f_0(\mathbf{X}_i), 0.5^2\}$ independently.
(ii) *Poisson*: for $i = 1, \ldots, n$, conditionally on $\mathbf{X}_i$, draw $Y_i \sim \mathrm{Pois}[g^{-1}\{f_0(\mathbf{X}_i)\}]$ independently, where $g(\mu) = \log(\mu)$.
(iii) *Binomial*: for $i = 1, \ldots, n$, draw T_i (independently of $\mathbf{X}_1, \ldots, \mathbf{X}_n$) from a uniform distribution on $\{11, 12, \ldots, 20\}$, and then, conditionally on $\mathbf{X}_i$ and T_i, draw $Y_i \sim T_i^{-1}\mathrm{Bin}[T_i, g^{-1}\{f(\mathbf{x}_i)\}]$ independently, where $g(\mu) = \log\{\mu/(1-\mu)\}$.

All the component functions are convex, so f_0 is convex. This allows us to compare our method with other shape-restricted methods in the Gaussian setting. In the binomial setting, we considered the EF with different dispersion parameters since T_i here can take different values. The new partial log-likelihood can be viewed as a weighted version of the original partial log-likelihood, where this feature can be easily incorporated in the SCMLE. Regardless of the EF distributions, problem 3 represents a more challenging (higher dimensional) problem.

In the Gaussian setting, we compared the performance of the SCMLE with shape-constrained additive models, SCAM (Pya and Wood, 2015), GAMs with integrated smoothness estimation, GAMIS (Wood, 2004), multivariate adaptive regression splines with maximum interaction degree equal to 1, MARS (Friedman, 1991), regression trees, Tree (Breiman *et al.*, 1984), convex adaptive partitioning, CAP (Hannah and Dunson, 2013) and multivariate convex regression, MCR (Lim and Glynn, 2012; Seijo and Sen, 2011). Some of these methods are not designed to deal with non-identity link functions, so in the Poisson and binomial settings we compared the SCMLE with only SCAM and GAMIS.

SCAM can be viewed as a shape-restricted version of GAMIS. It is a spline-based method, and is implemented in the R package `scam` (Pya, 2012). GAMIS is implemented in the R package `mgcv` (Wood, 2012), whereas MARS can be founded in the R package `mda` (Hastie *et al.*, 2011). The method of regression trees is implemented in the R package `tree` (Ripley, 2012), and CAP is implemented in MATLAB by Hannah and Dunson (2013). We implemented MCR in MATLAB using the `interior-point-convex` solver. Default settings were used for all the competitors mentioned above.

For different sample sizes $n = 200, 500, 1000, 2000, 5000$, we ran all the methods on 50 randomly generated data sets. Our numerical experiments were carried out on standard 32-bit desktops with 1.8 GHz central processor units. Each method was given at most 1 h per data set. Beyond this limit, the run was forced to stop and the corresponding results were omitted. Tables 1 and 2 in the on-line supplementary material provide the average running time of different methods per training data set. The SCMLE method takes roughly 1–2 s with a sample size of $n = 1000$ and 4–8 additive components, which is unsurprisingly slower than Tree or MARS, but it is typically faster than other shape-constrained methods such as SCAM and MCR. Note that MCR is particularly slow compared with the other methods and becomes computationally infeasible for $n \geqslant 1000$.

To study the empirical performance of the SCMLE, we drew 10^5 covariates independently from $U[-0.98, 0.98]^d$ and estimated the mean integrated squared error MISE, namely $\mathbb{E}\{\int_{[-0.98,0.98]^d}(\hat{f}_n - f_0)^2\}$, using Monte Carlo integration. Estimated MISEs are given in Tables 5 and 6. For every setting that we considered, the SCMLE method performs better than Tree, CAP and MCR. This is largely because these three estimators do not take into account the additive structure. In particular, MCR suffers severely from its boundary behaviour.

Table 5. Estimated MISEs in the Gaussian setting for problems 1–3†

Method	*MISEs for the following values of n:*				
	n = 200	*n = 500*	*n = 1000*	*n = 2000*	*n = 5000*
Problem 1					
SCMLE	0.41	*0.17*	*0.085*	*0.044*	*0.021*
$SCAM_{10}$	0.41	0.25	0.16	0.13	0.079
$SCAM_{20}$	*0.40*	0.21	0.12	0.052	0.024
GAMIS	0.41	0.18	0.095	0.049	0.024
MARS	0.54	0.25	0.14	0.087	0.044
Tree	3.7	2.8	2.5	2.3	2.3
CAP	3.2	1.7	0.91	0.55	0.28
MCR	200	8400	—	—	—
Problem 2					
SCMLE	0.27	*0.10*	*0.053*	*0.028*	*0.012*
$SCAM_{10}$	*0.26*	0.11	0.058	0.032	0.016
$SCAM_{20}$	0.27	0.11	0.055	0.030	0.013
GAMIS	0.36	0.15	0.079	0.041	0.019
MARS	0.42	0.18	0.087	0.050	0.021
Tree	2.0	1.3	1.1	1.0	0.97
CAP	1.3	0.74	0.42	0.25	0.15
MCR	9400	15000	—	—	—
Problem 3					
SCMLE	11	*3.8*	*2.1*	*1.1*	*0.48*
$SCAM_{10}$	*9.3*	4.9	3.7	2.7	2.4
$SCAM_{20}$	9.4	4.7	3.0	1.9	1.1
GAMIS	11	4.6	2.5	1.4	0.63
MARS	14	6.6	4.9	3.6	3.1
Tree	120	94	87	81	80
CAP	93	72	51	39	30
MCR	170	1500	—	—	—

†The lowest MISE-values are in italics.

The performance of SCAM depends on the choice of a tuning parameter k that controls the number of B-spline basis functions for each component function. The default choice in the `scam` package is $k = 10$, though we also experimented with $k = 20$ and $k = 30$. The picture is somewhat mixed: in some settings, moving from $k = 10$ to $k = 20$ resulted in improvements for larger sample sizes, whereas in others the results were very similar, or even resulted in a deterioration in performance. Moving from $k = 20$ to $k = 30$ made much smaller differences. A drawback of increasing k is that the computation quickly became very burdensome, and in fact sometimes went beyond our '1 h per data set' cut-off for $k = 30$ when $n = 5000$. The $k = 10$ and $k = 20$ versions of SCAM are denoted as $SCAM_{10}$ and $SCAM_{20}$ respectively in Tables 5 and 6.

We found that SCAM and GAMIS occasionally offer slightly better performance than the SCMLE method when n is small. This is also mainly caused by the boundary behaviour of the SCMLE and is alleviated as the number of observations n increases. In fact, in each of the problems considered, the SCMLE method enjoys better predictive performance than the other methods for $n \geqslant 500$. The SCMLE appears to offer particular advantages when the true signal exhibits inhomogeneous smoothness, since it can regularize in a locally adaptive way, whereas both SCAM and GAMIS rely on a single level of regularization throughout the covariate space.

Table 6. Estimated MISEs in the Poisson and binomial settings for problems 1–3†

Model	*Method*	*MISEs for the following values of n:*				
		$n=200$	$n=500$	$n=1000$	$n=2000$	$n=5000$
Problem 1						
Poisson	SCMLE	0.34	*0.13*	*0.067*	*0.038*	*0.017*
	$SCAM_{10}$	0.34	0.21	0.14	0.11	0.069
	$SCAM_{20}$	*0.33*	0.17	0.082	0.041	0.019
	GAMIS	*0.33*	0.14	0.078	0.043	0.021
Binomial	SCMLE	0.93	*0.28*	*0.15*	*0.079*	*0.037*
	$SCAM_{10}$	*0.50*	0.32	0.27	0.24	0.22
	$SCAM_{20}$	*0.50*	0.32	0.26	0.24	0.21
	GAMIS	0.64	0.28	0.15	0.085	0.040
Problem 2						
Poisson	SCMLE	0.44	*0.14*	*0.079*	*0.042*	*0.019*
	$SCAM_{10}$	*0.38*	0.18	0.092	0.047	0.024
	$SCAM_{20}$	0.46	0.24	0.15	0.086	0.042
	GAMIS	0.51	0.21	0.12	0.064	0.030
Binomial	SCMLE	*0.36*	*0.13*	*0.065*	*0.036*	*0.016*
	$SCAM_{10}$	0.45	0.23	0.14	0.072	0.025
	$SCAM_{20}$	0.46	0.25	0.13	0.065	0.018
	GAMIS	0.45	0.17	0.090	0.054	0.024
Problem 3						
Poisson	SCMLE	*4.4*	*1.5*	*0.75*	*0.41*	*0.18*
	$SCAM_{10}$	5.4	3.4	2.5	2.1	1.7
	$SCAM_{20}$	6.0	4.4	4.0	3.8	2.7
	GAMIS	4.7	1.9	0.98	0.57	0.28
Binomial	SCMLE	41	*11*	*5.7*	*3.0*	*1.3*
	$SCAM_{10}$	24	17	14	13	12
	$SCAM_{20}$	*23*	15	13	11	7.5
	GAMIS	25	12	6.3	3.5	1.6

†The lowest MISE-values are in italics.

Finally, we note that in certain other shape-constrained estimation problems where boundary effects are also observed, such as log-concave density estimation, it is possible to construct a fully automatic smoothed estimate to alleviate these issues (Dümbgen and Rufibach, 2009; Cule *et al.*, 2010; Chen and Samworth, 2013). However, in this instance, it seems that the most obvious remedy for boundary effects for small sample sizes would be to impose an additional constraint on the Lipschitz constant of each convex or concave component, and an upper and lower bound on each monotone component. Although feasible to implement, it seems difficult to give practical advice for the choice of these tuning parameters, and we do not pursue this issue further here.

5.2. Generalized additive index models with shape restrictions

In our comparisons of different estimators in GAIMs, we focused on the Gaussian case to facilitate comparisons with other methods. We took $\mathbf{X}_1, \ldots, \mathbf{X}_n \overset{\text{IID}}{\sim} U[-1,1]^d$ for each data set and considered the following two problems.

Table 7. Estimated MISEs in problems 4 and 5†

Method	*MISEs for the following values of n:*				
	$n=200$	$n=500$	$n=1000$	$n=2000$	$n=5000$
Problem 4					
SCAIE	*0.26*	*0.074*	*0.038*	*0.019*	*0.008*
SSI	0.88	0.48	0.31	0.21	—
PPR	0.68	0.42	0.28	0.20	0.15
MARS	0.63	0.44	0.24	0.18	0.14
Tree	1.9	0.74	0.43	0.41	0.41
CAP	0.35	0.14	0.081	0.056	0.016
MCR	2.5×10^3	3.5×10^4	—	—	—
Problem 5					
SCAIE	*0.078*	*0.030*	*0.016*	*0.008*	*0.005*
PPR	0.14	0.055	0.027	0.015	0.010
MARS	0.081	0.034	0.018	0.010	0.006
Tree	0.37	0.24	0.27	0.31	0.31

†The lowest MISE-values are in italics.

(a) In problem 4, $d=4$ and $m=1$. We set $L_1=4$ and $f_0^{\mathrm{I}}(\mathbf{x})=|0.25x_1+0.25x_2+0.25x_3+0.25x_4|$.
(b) In problem 5, $d=2$ and $m=2$. We set $\mathbf{L}_2=(4,7)^{\mathrm{T}}$ and $f_0^{\mathrm{I}}(\mathbf{x})=(0.5x_1+0.5x_2)^2-|0.5x_1-0.5x_2|^3$.

In both problems, conditionally on $\mathbf{X}_i$, we drew independently $Y_i \sim N\{f_0^{\mathrm{I}}(\mathbf{X}_i), 0.5^2\}$ for $i=1,\ldots,n$. We compared the performance of our SCAIE with projection pursuit regression, PPR (Friedman and Stuetzle, 1981), multivariate adaptive regression splines with maximum two interaction degrees, MARS, and Tree. In addition, in problem 4, we also considered the semiparametric single-index method SSI (Ichimura, 1993), CAP and MCR. SSI was implemented in the R package `np` (Hayfield and Racine, 2013). The SCAIE was computed by using the algorithm illustrated in Table 4. We picked the total number of stochastic searches to be $N=100$. Because problem 4 is a single-index problem (i.e. $m=1$), there is no need to supply δ. In problem 5, we chose $\delta=0.1$. We considered sample sizes $n=$ 200, 500, 1000, 2000, 5000.

Table 3 in the on-line supplementary material gives the average running time of different methods per training data set. Although SCAIE is slower than PPR, MARS and Tree, its computation can be accomplished within 10–20 s when $n=1000$. As SSI adopts a leave-one-out cross-validation strategy, it is typically considerably slower than the SCAIE method.

Estimated MISEs of different estimators over $[-0.98, 0.98]^d$ are given in Table 7 based on 50 randomly generated data sets. In both problem 4 and problem 5, we see that SCAIE outperforms its competitors for all the sample sizes that we considered. It should, of course, be noted that SSI, PPR, MARS and Tree do not enforce the shape constraints, whereas MARS, Tree, CAP and MCR do not take into account the additive index structure.

In the index setting, it is also of interest to compare the performance of those methods that directly estimate the index matrix. We therefore estimated root-mean squared errors RMSE, given by $\sqrt{\mathbb{E}(\|\tilde{\boldsymbol{\alpha}}_{n,1}-\boldsymbol{\alpha}_{0,1}\|_2^2)}$ in problem 4, where $\boldsymbol{\alpha}_{0,1}=(0.25,0.25,0.25,0.25)^{\mathrm{T}}$. For problem 5, we estimated mean errors in Amari distance ρ, defined by Amari *et al.* (1996) as

Table 8. Distance between the estimated index matrix and the truth†

Method	*Results for the following values of n:*				
	$n=200$	$n=500$	$n=1000$	$n=2000$	$n=5000$
Problem 4					
SCAIE	*0.23*	*0.10*	*0.056*	*0.038*	*0.024*
SSI	0.68	0.62	0.60	0.49	—
PPR	0.58	0.60	0.54	0.48	0.45
Problem 5					
SCAIE	*0.22*	*0.14*	*0.090*	*0.062*	*0.045*
PPR	0.26	0.21	0.14	0.10	0.067

†RMSEs were estimated in problem 4, whereas the mean Amari errors were estimated in problem 5. The lowest distances are in italics.

$$\rho(\tilde{\mathbf{A}}_n, \mathbf{A}_0) = \frac{1}{2d}\sum_{i=1}^{d}\left(\frac{\sum_{j=1}^{d}|C_{ij}|}{\max_{1\leqslant j\leqslant d}|C_{ij}|} - 1\right) + \frac{1}{2d}\sum_{j=1}^{d}\left(\frac{\sum_{i=1}^{d}|C_{ij}|}{\max_{1\leqslant i\leqslant d}|C_{ij}|} - 1\right),$$

where $C_{ij} = (\tilde{\mathbf{A}}_n \mathbf{A}_0^{-1})_{ij}$ and

$$\mathbf{A}_0 = \begin{pmatrix} 0.5 & 0.5 \\ 0.5 & -0.5 \end{pmatrix}.$$

This distance measure is invariant to permutation and takes values in $[0, d-1]$. Results obtained for the SCAIE and, where applicable, SSI and PPR are displayed in Table 8. For both problems, the SCAIE method performs better in these senses than both SSI and PPR in terms of estimating the projection indices.

6. Real data examples

In this section, we apply our estimators in two real data examples. In the first, we study doctoral publications in biochemistry and fit a generalized (Poisson) additive model with concavity constraints, whereas in the second we use an additive index model with monotonicity constraints to study javelin performance in the decathlon.

6.1. Doctoral publications in biochemistry

The scientific productivity of a doctoral student may depend on many factors, including some or all of the number of young children that they have, the productivity of the supervisor, their gender and marital status. Long (1990) studied this topic, focusing on the gender difference; see also Long (1997). The data set is available in the R package `AER` (Kleiber and Zeileis, 2013) and contains $n=915$ observations. Here we model the number of articles written by the ith doctoral student in the last 3 years of their research as a Poisson random variable with mean μ_i, where

$$\log(\mu_i) = f_1(\text{kids}_i) + f_2(\text{mentor}_i) + a_3\,\text{gender}_i + a_4\,\text{married}_i + c,$$

Table 9. Estimates obtained from the SCMLE, SCAM and GAMIS on the doctoral publications data set

Method	$\hat{f}_{n,1}(0)$	$\hat{f}_{n,1}(1)$	$\hat{f}_{n,1}(2)$	$\hat{f}_{n,1}(3)$	$\hat{a}_{n,3}$	$\hat{a}_{n,4}$
SCMLE	0	−0.110	−0.284	−0.816	−0.218	0.126
SCAM	0	−0.136	−0.303	−0.770	−0.224	0.152
GAMIS	0	−0.134	−0.301	−0.784	−0.226	0.157

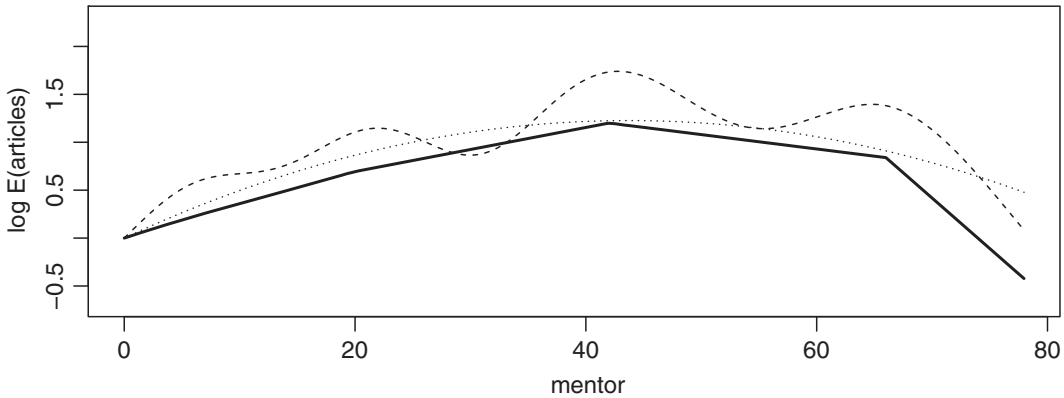

Fig. 1. Different estimates of f_2: ———, SCMLE; · · · · · ·, SCAM; - - - - - -, GAMIS

for $i = 1, \ldots, n$, where kids_i and mentor_i are respectively the number of that student's children who are less than 6 years old, and the number of papers published by that student's supervisor during the same period of time. Both gender_i and married_i are factors taking values 0 and 1, where 1 indicates 'female' and 'married' respectively. In the original data set, there is an extra continuous variable that measures the prestige of the graduate programme. We chose to drop this variable in our example because

(a) its values were determined quite subjectively and
(b) including this variable does not seem to improve the predictive power in the above settings.

To apply the SCMLE, we assume that f_1 is a concave and monotone decreasing function, whereas f_2 is a concave function. The main estimates obtained from the SCMLE are summarized in Table 9 and Fig. 1. Outputs from SCAM and GAMIS are also reported for comparison. We see that, with the exception of $\hat{f}_{n,2}$, estimates obtained from these methods are relatively close. Note that, in Fig. 1, the GAMIS estimate of f_2 displays local fluctuations that might be more difficult to interpret than the estimates obtained by using the SCMLE and SCAM methods.

Finally, we examine the prediction power of the different methods via cross-validation. Here we randomly split the data set into training (70%) and validation (30%) subsets. For each split, we compute estimates by using only the training set and assess their predictive accuracy in terms of the root-mean-square prediction error RMSPE on the validation set. The RMSPEs reported in Table 10 are averages over 500 splits. Our findings suggest that, although comparable with SCAM, the SCMLE offers slight improvements over GAMIS and Tree for this data set.

Table 10. Estimated prediction errors of the SCMLE, SCAM, GAMIS and Tree on the doctoral publications data set†

Method	*RMSPE*
SCMLE	*1.822*
SCAM	1.823
GAMIS	1.838
Tree	1.890

†The smallest RMSPE is in italics.

Table 11. Estimated index loadings by SCAIE and SCAIE_s

Method	$\hat{a}_{n,11}$	$\hat{a}_{n,21}$	$\hat{a}_{n,31}$	$\hat{a}_{n,41}$	$\hat{a}_{n,12}$	$\hat{a}_{n,22}$	$\hat{a}_{n,32}$	$\hat{a}_{n,42}$
SCAIE	0.222	0.173	0.262	0.343	0.522	0.457	0.006	0.015
SCAIE_s	0.140	0.320	0.235	0.305	0.536	0.464	0	0

6.2. Javelin throw

In this section, we consider the problem of predicting a decathlete's javelin performance from their performances in the other decathlon disciplines. Our data set consists of decathlon athletes who scored at least 6500 points in at least one athletic competition in 2012 and scored points in every event there. To avoid data dependence, we include only one performance from each athlete, namely their 2012 personal best performance (over the whole decathlon). The data set, which consists of $n = 614$ observations, is available in the R package `scar` (Chen and Samworth, 2014). For simplicity, we only select events (apart from the javelin) that directly reflect the athlete's ability in throwing and short distance running, namely the shot put, discus, 100 m race and 110 m hurdles race. We fit the following additive index model:

$$\begin{aligned}\text{javelin}_i = {} & f_1(A_{11}\text{shot}_i + A_{21}\text{discus}_i + A_{31}100\text{m}_i + A_{41}110\text{m}_i) \\ & + f_2(A_{12}\text{shot}_i + A_{22}\text{discus}_i + A_{32}100\text{m}_i + A_{42}110\text{m}_i) + c + \epsilon_i,\end{aligned}$$

for $i = 1, \ldots, 614$, where $\epsilon_i \overset{\text{IID}}{\sim} N(0, \sigma^2)$, and where javelin_i, shot_i, discus_i, 100m_i and 110m_i represent the corresponding decathlon event scores for the ith athlete. For the SCAIE, we assume that both f_1 and f_2 are monotone increasing, and we also assume that $A_{11}, \ldots, A_{41}, A_{12}, \ldots, A_{42}$ are non-negative. This slightly restricted version of the SCAIE aids interpretability of the indices, and prevents the 'perfect fit' phenomenon (see Section 3.2), so no choice of δ is required.

Table 11 gives the estimated index loadings by SCAIE. We observe that the first projection index can be viewed as the general athleticism associated with the athlete, whereas the second can be interpreted as a measure of throwing ability. Note that, when using the SCAIE method, $\hat{A}_{n,32}$ and $\hat{A}_{n,42}$ are relatively small. To simplify our model further, and to seek improvement in the prediction power, we therefore considered forcing these entries to be exactly 0 in the optimization steps of the SCAIE method. This sparse version is denoted as SCAIE_s. Its estimated index loadings are also reported in Table 11.

To compare the performance of our methods with PPR, MARS with maximum two degrees of interaction and Tree, we again estimated the prediction power (in terms of RMSPE) via 500

Table 12. Estimated prediction errors of SCAIE, SCAIE_s, PPR, MARS and Tree on the javelin data set†

Method	*RMSPE*
SCAIE	81.276
SCAIE_s	*80.976*
PPR	82.898
MARS	82.915
Tree	85.085

†The smallest RMSPE is in italics.

repetitions of 70%–30% random splits into training–test sets. The corresponding RMSPEs are reported in Table 12. We see that both SCAIE and SCAIE_s outperform their competitors in this particular data set. It is also interesting to note that SCAIE_s has a slightly lower RMSPE than the SCAIE, suggesting that the simpler (sparser) model might be preferred for prediction here.

7. Extensions and outlook

In this paper, we have developed methodology and theory for fitting GAMs with shape constraints on the additive components. Despite the non-parametric nature of the problem, our approach has the attractive feature that there are no tuning parameters to choose, and moreover it can be extended to handle an index structure. The algorithms that we have developed are fast to compute and are publicly available in the R package `scar`. We now describe various possible extensions of our theoretical result on the consistency of the SCMLE (theorem 1), and we conclude with a more general discussion of remaining challenges and possible future directions.

As a generalization of the Gaussian shape-constrained additive model, consider the setting where the IID pairs $(\mathbf{X}_1, Y_1), \ldots, (\mathbf{X}_n, Y_n)$ satisfy

$$Y_i = f_0(\mathbf{X}_i) + \epsilon_i,$$

for $i = 1, \ldots, n$, where $\mathbb{E}(\epsilon_i | \mathbf{X}_i) = 0$ and $\text{var}(\epsilon_i | \mathbf{X}_i) = \phi_0$. In this case, we can define the shape-constrained least squares estimator (SCLSE) by

$$\tilde{f}_n \in \underset{f \in \text{cl}(\mathcal{F})}{\arg\min} \frac{1}{n} \sum_{i=1}^{n} \{Y_i - f_0(\mathbf{X}_i)\}^2,$$

and note that the SCLSE coincides with the SCMLE when $\epsilon_i | \mathbf{X}_i \sim N(0, \phi_0)$. Theorem 1 can be extended to show the consistency of the SCLSE; see the remarks following the proof of theorem 1 in the on-line supplementary material.

Several of the other assumptions of theorem 1 can be weakened at the expense of lengthening the proof still further. First, in assumption 2, we can instead assume that the support $\text{supp}(\mathbf{X})$ of the covariates is a convex subset of $\mathbb{R}^d$ with positive Lebesgue measure. In that case, it can be concluded that the SCMLE $\hat{f}_n$ converges uniformly to f_0 almost surely on any compact subset contained in the interior of $\text{supp}(\mathbf{X})$. In fact, with some minor modifications, our proof can also be generalized to situations where some components of $\mathbf{X}$ are discrete. Second, consistency under a weaker L^1-norm on $[-a_0, a_0]^d$ can be established without assumption 4. In addition,

instead of assuming a single dispersion parameter ϕ_0 as done here, we can take $\phi_{ni} = \phi_0/w_{ni}$ for $i = 1, \ldots, n$, where w_{ni} are known, positive weights (this is frequently needed in practice in the binomial setting). In that case, the new partial log-likelihood can be viewed as a weighted version of the original partial likelihood. Consistency of the SCMLE can be established provided that $\liminf_{n\to\infty} \min_i w_{ni}/\max_i w_{ni} > 0$.

In another direction, our work could potentially be extended to non-canonical link functions, though this would require further technical conditions. For example, proposition 1 does not necessarily hold for every monotone link function g, i.e. the SCMLE $\hat{f}_n$ may not be unique at $\mathbf{X}_1, \ldots, \mathbf{X}_n$. See Table 1 of Wedderburn (1976) for a summary of the uniqueness of the maximum likelihood estimator for various link functions and EF distributions. Moreover, a different algorithm would be needed here because the objective function is not necessarily concave for a general link function.

There remain several outstanding theoretical and methodological challenges. On the theory side, it is of great interest to understand both the local and the global rates of convergence of the least squares and maximum likelihood estimators in shape-constrained additive models. The only prior works in this direction with which we are familiar are Mammen and Yu (2007), Cheng (2009) and Yu (2014) on local rates. Mammen and Yu (2007) studied the simplest setting of least squares estimators $\hat{f}_1, \ldots, \hat{f}_d$ based on independent observations from the monotone regression model

$$Y_i = f_1(X_{i1}) + \ldots + f_d(X_{id}) + \epsilon_i,$$

$i = 1, \ldots, n$, where ϵ_i is assumed to have a subexponential distribution. Under regularity conditions, including an assumption that the design points are supported on $[0, 1]^d$ and each f_j is strictly increasing and has a bounded derivative on $[0, 1]$, they showed that $n^{1/3}\{\hat{f}_1(x_1) - f_1(x_1)\}$ has a non-degenerate limiting distribution at all points $x_1 \in (0, 1)$ with $f_1'(x_1) > 0$. This shows that, under these conditions, the least squares estimator evades the curse of dimensionality that one typically observes in multivariate shape-constrained regression problems and is the optimal rate for monotone regression. It is natural to conjecture that, under appropriate conditions, the least squares estimator would also achieve the optimal rate of $O_p(n^{-2/5})$ for convex or concave components. Cheng (2009) and Yu (2014) extended the work of Mammen and Yu (2007) to a setting where the expected response is an additive function that can be decomposed as a sum of linear and monotone components. The behaviour of our estimators under model misspecification is another intriguing topic that warrants further research.

On the methodological side, in addition to studying other distributional families and shape constraints, it would be desirable to extend our methodology to handle settings with large numbers of covariates under an assumption that most of these covariates have a negligible effect on the response. In this direction, Fang and Meinshausen (2012) studied the special case of a high dimensional additive model with isotonic restrictions on the additive components. We suggest that the additive structure is an attractive way of pushing ideas of shape-constrained estimation into high dimensional settings.

Acknowledgements

We thank Mary Meyer for providing us with her manuscript Meyer (2013a) before its publication. We thank the Joint Editor, Associate Editor and two referees for their constructive suggestions, which helped to improve the paper. Both authors are supported by the second author's Engineering and Physical Sciences Research Fellowship EP/J017213/1.

Appendix A: Examples

In this appendix, we give examples to illustrate certain phenomena that are described in the main text.

A.1. Example 1

In this example, we show that the SCMLE need not exist if we were to maximize over $\mathcal{F}$ rather than $\mathrm{cl}(\mathcal{F})$ in expression (3), as claimed in Section 2.2. Suppose that $d=1$ and that $Y_i|X_i \sim \mathrm{Bin}\{1, p(X_i)\}$, where

$$\log\left\{\frac{p(X_i)}{1-p(X_i)}\right\} = c + f(X_i)$$

and where f is monotone increasing. If there exists $x^* \in \mathbb{R}$ such that $Y_i = 0$ whenever $X_i < x^*$, and $Y_i = 1$ whenever $X_i > x^*$, then we claim that there is no maximizer of the scaled partial log-likelihood (2). To see this, note that $\ell_i(f) = Y_i f(X_i) - \log[1+\exp\{f(X_i)\}]$, which is strictly increasing in $f(X_i)$ if $Y_i = 1$, and strictly decreasing in $f(X_i)$ if $Y_i = 0$. Thus, any maximizing sequence (f^k) in $\mathcal{F}$ must satisfy $f^k(x) \to -\infty$ for $x < x^*$ and $f^k(x) \to \infty$ for $x > x^*$. But the pointwise limit of such a sequence does not belong to $\mathcal{F}$.

A.2. Example 2

In this example, we show how maximizing the function $\Lambda_n(\cdot)$ in expression (6) over $A \in \mathcal{A}$ can result in a perfect fit to the data (or 'saturated solution'), as claimed in Section 3.2. Consider the Gaussian family with the identity link function. Suppose that we have data $(\mathbf{X}_1, Y_1), \ldots, (\mathbf{X}_n, Y_n)$ with $\mathbf{X}_i = (X_{i1}, X_{i2})^{\mathrm{T}}$ (so $d = m = 2$) and $\mathbf{L}_2 = (2,3)^{\mathrm{T}}$. We assume here for simplicity that $X_{11} < \ldots < X_{n1}$. It is possible to find an increasing function f_1, a decreasing function f_2 (with $f_1(0) = f_2(0) = 0$) and a constant c such that $f_1(X_{i1}) + f_2(X_{i1}) + c = Y_i$ for every $i = 1, \ldots, n$. Now pick ϵ such that

$$0 < \epsilon < \min\left\{\frac{1}{2}, \frac{\min\limits_{1 \leqslant i < n}(X_{i+1,1} - X_{i1})}{4(\max\limits_{1 \leqslant i \leqslant n}|X_{i2}| + 1)}\right\},$$

and let

$$\mathbf{A} = (\boldsymbol{\alpha}_1, \boldsymbol{\alpha}_2) = \begin{pmatrix} 1 & 1-\epsilon \\ 0 & \epsilon \end{pmatrix}.$$

It can be checked that $\{\boldsymbol{\alpha}_2^{\mathrm{T}}\mathbf{X}_i\}_{i=1}^n$ is a strictly increasing sequence, so one can find a decreasing function f_2^* such that $f_2^*(\boldsymbol{\alpha}_2^{\mathrm{T}}\mathbf{X}_i) = f_2(X_{i1})$ for every $i = 1, \ldots, n$. Consequently, by taking $\hat{f}^{\mathrm{I}}(\mathbf{x}) = f_1(\mathbf{A}^{\mathrm{T}}\mathbf{x}) + f_2^*(\mathbf{A}^{\mathrm{T}}\mathbf{x}) + c$, we can ensure that $\hat{f}^{\mathrm{I}}(\mathbf{X}_i) = Y_i$ for every $i = 1, \ldots, n$.

We remark that this 'perfect fit' phenomenon is quite general. Actually, one can show (via simple modifications of the above example) that it can happen whenever $m \geqslant 2$ and $\mathbf{L}_m \notin \mathcal{L}_m$, where $\mathcal{L}_m = \{1,4,5,6\}^m \cup \{1,7,8,9\}^m$.

A.3. Example 3

We now show that the function $\Lambda_n(\cdot)$ need not be upper semicontinuous, as was also claimed in Section 3.2. Again consider the Gaussian family with the identity link function. Take $d = m = 2$ and $\mathbf{L}_2 = (2,3)^{\mathrm{T}}$. Assume that there are $n = 4$ observations, namely $\mathbf{X}_1 = (0,0)^{\mathrm{T}}$, $\mathbf{X}_2 = (0,1)^{\mathrm{T}}$, $\mathbf{X}_3 = (1,0)^{\mathrm{T}}$, $\mathbf{X}_4 = (1,1)^{\mathrm{T}}$, $Y_1 = Y_2 = Y_3 = 0$ and $Y_4 = 1$. If we take

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

then it can be shown that $\Lambda_n(A) = 3/32$ by fitting $\hat{f}^{\mathrm{I}}(\mathbf{X}_1) = -\frac{1}{4}$, $\hat{f}^{\mathrm{I}}(\mathbf{X}_2) = \hat{f}^{\mathrm{I}}(\mathbf{X}_3) = \frac{1}{4}$ and $\hat{f}^{\mathrm{I}}(\mathbf{X}_4) = \frac{3}{4}$. However, for any sufficiently small $\epsilon > 0$, if we define

$$\mathbf{A}_\epsilon = \begin{pmatrix} 1-\epsilon & \epsilon \\ \epsilon & 1-\epsilon \end{pmatrix},$$

then we can take $\hat{f}^{\mathrm{I}}(\mathbf{X}_i) = Y_i$ for $i = 1, \ldots, 4$, so that $\Lambda_n(\mathbf{A}_\epsilon) = \frac{1}{8} > \Lambda_n(\mathbf{A})$.

References

Aït-Sahalia, Y. and Duarte, J. (2003) Nonparametric option pricing under shape restrictions. *J. Econmetr.*, **116**, 9–47.

Amari, S., Cichocki, A. and Yang, H. (1996) A new learning algorithm for blind signal separation. In *Advances in Neural Information Processing Systems*, pp. 757–763. Cambridge: MIT Press.

Bacchetti, P. (1989) Additive isotonic models. *J. Am. Statist. Ass.*, **84**, 289–294.

Banerjee, M. (2007) Likelihood based inference for monotone response models. *Ann. Statist.*, **35**, 931–956.

Banerjee, M. (2009) Inference in exponential family regression models under certain shape constraints. In *Advances in Multivariate Statistical Methods* (ed. A. SenGupta), vol. 4, pp. 249–272. Singapore: World Scientific Publishing.

Barlow, R. E., Bartholomew, D. J., Bremner, J. M. and Brunk, H. D. (1972) *Statistical Inference under Order Restrictions*. New York: Wiley.

Breiman, L. and Friedman, J. H. (1985) Estimating optimal transformations for multiple regression and correlation. *J. Am. Statist. Ass.*, **80**, 580–598.

Breiman, L., Friedman, J. H., Olshen, R. A. and Stone, C. J. (1984) *Classification and Regression Trees*. Belmont: Wadsworth.

Brezger, A. and Steiner, W. J. (2008) Monotonic regression based on Bayesian P-splines: an application to estimating price response functions from store-level scanner data. *J. Bus. Econ. Statist.*, **26**, 90–104.

Brunk, H. D. (1958) On the estimation of parameters restricted by inequalities. *Ann. Math. Statist.*, **29**, 437–454.

Brunk, H. D. (1970) Estimation of isotonic regression. In *Nonparametric Techniques in Statistical Inference*, pp. 177–195. Cambridge: Cambridge University Press.

Buja, A., Hastie, T. and Tibshirani, R. (1989) Linear smoothers and additive models. *Ann. Statist.*, **17**, 453–555.

Cai, B. and Dunson, D. B. (2007) Bayesian multivariate isotonic regression splines. *J. Am. Statist. Ass.*, **102**, 1158–1171.

Chen, Y. and Samworth, R. J. (2013) Smoothed log-concave maximum likelihood estimation with applications. *Statist. Sin.*, **23**, 1373–1398.

Chen, Y. and Samworth, R. J. (2014) scar: shape-constrained additive regression: a maximum likelihood approach. *R Package Version 0.2-0.* Centre for Mathematical Sciences, University of Cambridge, Cambridge. (Available from `http://cran.r-project.org/web/packages/scar/`.)

Cheng, G. (2009) Semiparametric additive isotonic regression. *J. Statist. Planng Inf.*, **139**, 1980–1991.

Cheng, G., Zhao, Y. and Li, B. (2012) Empirical likelihood inference for the semiparametric additive isotonic regression. *J. Multiv. Anal.*, **112**, 172–182.

Cule, M., Samworth, R. and Stewart, M. (2010) Maximum likelihood estimation of a multi-dimensional log-concave density (with discussion). *J. R. Statist. Soc.* B, **72**, 545–607.

Dümbgen, L. and Rufibach, K. (2009) Maximum likelihood estimation of a log-concave density and its distribution function: basic properties and uniform consistency. *Bernoulli*, **15**, 40–68.

Dümbgen, L. and Rufibach, K. (2011) logcondens: computations related to univariate log-concave density estimation. *J. Statist. Softwr.*, **39**, 1–28.

Dümbgen, L., Samworth, R. and Schuhmacher, D. (2011) Approximation by log-concave distributions with applications to regression. *Ann. Statist.*, **39**, 702–730.

Dümbgen, L., Samworth, R. J. and Schuhmacher, D. (2013) Stochastic search for semiparametric linear regression models. In *From Probability to Statistics and Back: High-dimensional Models and Processes—a Festschrift in Honor of Jon A. Wellner* (eds M. Banerjee, F. Bunea, J. Huang, V. Koltchinskii and M. H. Matthews), pp. 78–90. Beachwood: Institute of Mathematical Statistics.

Eriksson, J. and Koivunen, V. (2004) Identifiability, separability and uniqueness of linear ICA models. *IEEE Sign. Process. Lett.*, **11**, 601–604.

Fang, Z. and Meinshausen, N. (2012) LASSO isotone for high-dimensional additive isotonic regression. *J. Computnl Graph. Statist.*, **21**, 72–91.

Foster, J. C., Taylor, J. M. G. and Nan, B. (2013) Variable selection in monotone single-index models via the adaptive LASSO. *Statist. Med.*, **32**, 3944–3954.

Friedman, J. H. (1991) Multivariate adaptive regression splines. *Ann. Statist.*, **19**, 1–67.

Friedman, J. H. and Stuetzle, W. (1981) Projection pursuit regression. *J. Am. Statist. Ass.*, **76**, 817–823.

Groeneboom, P., Jongbloed, G. and Wellner, J. A. (2001) Estimation of a convex function: characterizations and asymptotic theory. *Ann. Statist.*, **29**, 1653–1698.

Groeneboom, P., Jongbloed, G. and Wellner, J. A. (2008) The support reduction algorithm for computing nonparametric function estimates in mixture models. *Scand. J. Statist.*, **35**, 385–399.

Guntuboyina, A. and Sen, B. (2013) Global risk bounds and adaptation in univariate convex regression. *Probab. Theor. Reltd Flds*, to be published.

Hall, P. and Huang, L. S. (2001) Nonparametric kernel regression subject to monotonicity constraints. *Ann. Statist.*, **29**, 624–647.

Hannah, L. A. and Dunson, D. B. (2013) Multivariate convex regression with adaptive partitioning. *J. Mach. Learn. Res.*, **14**, 3261–3294.

Hanson, D. L. and Pledger, G. (1976) Consistency in concave regression. *Ann. Statist.*, **4**, 1038–1050.
Hastie, T. and Tibshirani, R. (1986) Generalized additive models (with discussion). *Statist. Sci.*, **1**, 297–318.
Hastie, T. and Tibshirani, R. (1990) *Generalized Additive Models*. London: Chapman and Hall.
Hastie, T., Tibshirani, R., Leisch, F., Hornik, K. and Ripley, B. D. (2011) mda: mixture and flexible discriminant analysis. *R Package Version 0.4-2*. Stanford University, Stanford. (Available from `http://cran.r-project.org/web/packages/mda/`.)
Hayfield, T. and Racine, J. S. (2013) np: nonparametric kernel smoothing methods for mixed data types. *R Package Version 0.50-1*. Eidgenössiche Technische Hochschule Zürich, Zürich. (Available from `http://cran.r-project.org/web/packages/np/`.)
Ichimura, H. (1993) Semiparametric least squares (SLS) and weighted SLS estimation of single-index models. *J. Econmetr.*, **58**, 71–120.
Jørgensen, B. (1983) Maximum likelihood estimation and large-sample inference for generalized linear and nonlinear regression models. *Biometrika*, **70**, 19–28.
Keshavarz, A., Wang, Y. and Boyd, S. (2011) Imputing a convex objective function. In *Proc. Int. Symp. Intelligent Control, Denver*, pp. 613–619. New York: Institute of Electrical and Electronics Engineers.
Kim, A. K. H. and Samworth, R. J. (2014) Global rates of convergence in log-concave density estimation. University of Cambridge, Cambridge. (Available from `http://arxiv.org/abs/1404.2298`.)
Kleiber, C. and Zeileis, A. (2013) AER: applied econometrics with R. *R Package Version 1.2-0*. Universität Basel, Basel. (Available from `http://cran.r-project.org/web/packages/AER/`.)
Li, Q. and Racine, J. S. (2007) *Nonparametric Econometrics: Theory and Practice*. Princeton: Princeton University Press.
Lim, E. and Glynn, P. W. (2012) Consistency of multidimensional convex regression. *Oper. Res.*, **60**, 196–208.
Long, J. S. (1990) The origins of sex differences in science. *Socl Forces*, **68**, 1297–1316.
Long, J. S. (1997) *Regression Models for Categorical and Limited Dependent Variables*. Thousand Oaks: Sage.
Luss, R., Rosset, S. and Shahar, M. (2012) Efficient regularized isotonic regression with application to genegene interaction search. *Ann. Appl. Statist.*, **6**, 253–283.
Mammen, E. and Park, B. U. (2006) A simple smooth backfitting method for additive models. *Ann. Statist.*, **34**, 2252–2271.
Mammen, E. and Yu, K. (2007) Additive isotone regression. *IMS Lect. Notes*, **55**, 179–195.
Matzkin, R. L. (1991) Semiparametric estimation of monotone and concave utility functions for polychotomous choice models. *Econometrica*, **59**, 1315–1327.
Meyer, M. C. (2013a) Semi-parametric additive constrained regression. *J. Nonparam. Statist.*, **25**, 715–743.
Meyer, M. C. (2013b) A simple new algorithm for quadratic programming with applications in statistics. *Communs Statist.*, **42**, 1126–1139.
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. Nonparam. Statist.*, **23**, 867–884.
Morton-Jones, T., Diggle, P., Parker, L., Dickinson, H. O. and Binks, K. (2000) Additive isotonic regression models in epidemiology. *Statist. Med.*, **19**, 849–859.
Nocedal, J., and Wright, S. J. (2006) *Numerical Optimization*, 2nd edn. New York: Springer.
Price, K., Storn, R. and Lampinen, J. (2005) *Differential Evolution: a Practical Approach to Global Optimization*. Berlin: Springer.
Pya, N. (2012) scam: shape constrained additive models. *R Package Version 1.1-5*. University of Bath, Bath. (Available from `http://cran.r-project.org/web/packages/scam/`.)
Pya, N. and Wood, S. N. (2015) Shape constrained additive models. *Statist. Comput.*, **25**, 543–559.
Ripley, B. D. (2012) tree: classification and regression trees. *R Package Version 1.0-33*. Department of Statistics, University of Oxford, Oxford. (Available from `http://cran.r-project.org/web/packages/tree/`.)
Rueda, C. (2013) Degrees of freedom and model selection in semiparametric additive monotone regression. *J. Multiv. Anal.*, **117**, 88–99.
Samworth, R. J. and Yuan, M. (2012) Independent component analysis via nonparametric maximum likelihood estimation. *Ann. Statist.*, **40**, 2973–3002.
Schell, M. J. and Singh, B. (1997) The reduced monotonic regression method. *J. Am. Statist. Ass.*, **92**, 128–135.
Seijo, E. and Sen, B. (2011) Nonparametric least squares estimation of a multivariate convex regression function. *Ann. Statist.*, **39**, 1633–1657.
Stone, C. (1986) The dimensionality reduction principle for generalized additive models. *Ann. Statist.*, **14**, 590–606.
Tutz, G. and Leitenstorfer, F. (2007) Generalized smooth monotonic regression in additive modeling. *J. Computnl Graph. Statist.*, **16**, 165–188.
Varian, H. R. (1984) The nonparametric approach to production analysis. *Econometrica*, **52**, 579–597.
Wedderburn, R. W. M. (1976) On the existence and uniqueness of the maximum likelihood estimates for certain generalized linear models. *Biometrika*, **63**, 27–32.
Wood, S. N. (2004) Stable and efficient multiple smoothing parameter estimation for generalized additive models. *J. Am. Statist. Ass.*, **99**, 673–686.
Wood, S. N. (2006) *Generalized Additive Models: an Introduction with R*. Boca Raton: Chapman and Hall.

Wood, S. N. (2008) Fast stable direct fitting and smoothness selection for generalized additive models. *J. R. Statist. Soc.* B, **70**, 495–518.
Wood, S. N. (2012) mgcv: mixed GAM computation vehicle with GCV/AIC/REML smoothness estimation. *R Package Version 1.7-22*. University of Bath, Bath. (Available from `http://cran.r-project.org/web/packages/mgcv/`.)
Xu, M., Chen, M. and Lafferty, J. (2014) Faithful variable screening for high-dimensional convex regression. Carnegie Mellon University, Pittsburgh. (Available from `http://arxiv.org/abs/1411.1805`.)
Yu, K. (2014) On partial linear additive isotonic regresison. *J. Kor. Statist. Soc.*, **43**, 11–17.
Yu, K., Park, B. U. and Mammen, E. (2008) Smooth backfitting in generalized additive models. *Ann. Statist.*, **36**, 228–260.
Yuan, M. (2011) On the identifiability of additive index models. *Statist. Sin.*, **21**, 1901–1911.

Supporting information
Additional 'supporting information' may be found in the on-line version of this article:

'Online supplementary material for "Generalized additive and index models with shape constraints"'.