Shape constrained estimation using nonnegative splines

Dávid Papp
Department of Industrial Engineering and Management Sciences,
Northwestern University, Evanston IL 60208, USA,
e-mail:dpapp@iems.northwestern.edu
and
Farid Alizadeh
Rutgers Center for Operations Research,
Rutgers University, New Brunswick NJ 08854, USA

May 31, 2012

Abstract

We consider the problem of nonparametric estimation of unknown smooth functions in the presence of restrictions on the shape of the estimator and on its support, using polynomial splines. We provide a general computational framework that treats these estimation problems in a unified manner, without the limitations of the existing methods. Applications of our approach include computing optimal spline estimators for regression, density estimation, and arrival rate estimation problems in the presence of various shape constraints. Our approach can also handle multiple simultaneous shape constraints. The approach is based on a characterization of nonnegative polynomials that leads to semidefinite programming (SDP) and second order cone programming (SOCP) formulations of the problems. These formulations extend and generalize a number of previous approaches in the literature, including those with piecewise linear and B-spline estimators. We also consider a simpler approach, in which nonnegative splines are approximated by splines whose pieces are polynomials with nonnegative coefficients in a nonnegative basis. A condition is presented to test whether a given nonnegative basis gives rise to a spline cone that is dense in the space of nonnegative continuous functions. The optimization models formulated in the paper are solvable with minimal running time using off-the-shelf software. We provide numerical illustrations for density estimation and regression problems. These examples show that the proposed approach requires minimal computational time, and that the estimators obtained using our approach often match and frequently outperform kernel methods and spline smoothing without shape constraints.

Keywords: splines, density estimation, regression, second order cone programming, semidefinite programming, nonnegative polynomials, Bernstein polynomials

1 Introduction

We consider spline estimation problems with one or more shape constraints on the estimator. We demonstrate that nonnegative, monotone, and convex polynomial splines admit characterizations that lead to optimization models solvable with minimal running time using readily available software. The methods proposed in this paper are applicable to a variety of problems; we concentrate on regression and density estimation problems with various shape constraints. In our numerical examples we focus on nonnegative regression, isotonic regression and smoothing, and on unconstrained density estimation. Further potential applications include the estimation of the arrival rate of a non-homogeneous Poisson process, and log-concave density estimation.

In each of the estimation problems considered in this paper the goal is to reconstruct a real-valued function f from finitely many observations (function values observed with noise, realizations of random variables, arrival times, etc.) under a collection of constraints on the shape of the function. Examples of such shape constraints include: (1) f be nonnegative over $[a, b]$, or more generally, its graph lie in a specific bounded region (defined, for example, by linear or polynomial inequalities); (2) f be monotone non-decreasing (or non-increasing); (3) f be convex (or concave). The function f is otherwise assumed to belong to some (possibly infinite dimensional) functional space $\mathbb{H}$. The optimal estimator shall minimize a given *loss function* over the set of shape constrained functions from $\mathbb{H}$. We propose a variant of the classical *method of sieves* to find the estimator via the solution of finite dimensional optimization problems.

Our approach is based on convex (conic) optimization techniques. More specifically, we show that *linear programming (LP)*, *second order cone programming (SOCP)*, *semidefinite programming (SDP)*, and optimization over cones of nonnegative, monotone, or convex/concave functions in functional linear spaces can be employed to solve fairly complicated shape constrained estimation problems in a conceptually more satisfying manner than existing approaches, without resorting to unnecessary approximations of the problem. Furthermore, our approach is flexible enough to handle a number of additional constraints, such as interpolation, betweenness, and multiple shape constraints over different intervals. While a lot of attention has been given to shape constrained optimization, and also to employing (convex) optimization models in statistical estimation, no systematic study has appeared that demonstrate both the theoretical soundness and computational efficiency of convex optimization in such a general setting.

In the remainder of this section we give a broad overview of the vast literature on shape constrained estimation, focusing on the existing numerical algorithms. In Section 2 we lay out a general model using sieves of cones for shape constrained estimation problems. In Section 3 we consider the special case of polynomial splines, and we show how SOCP and SDP approaches can be used effectively to solve these problems. We also consider another approach that has been rediscovered several times in the literature. This approach is based on optimization over splines with nonnegative coefficients over some nonnegative basis. A condition is given that helps decide whether a given choice of nonnegative basis is appropriate. Section 4 discusses applications of the developed theory to a number of shape constrained estimation problems. The corresponding numerical results are collected in Section 5, where the SDP/SOCP approach is compared to the nonnegative basis approach, unconstrained smoothing splines, and to kernel estimators in density estimation and regression problems.

1.1 Past work and our contribution

There is a vast literature on shape constrained estimation and learning problems, much of which are summarized in the following surveys: in (Delecroix and Thomas-Agnan, 2000) a survey of smoothing regression problems with shape constraints is presented; in the thesis (Meyer, 1996) algorithms for shape constrained regression and density estimation are considered; the text of Robertson et al. (1988) is a comprehensive survey of order restricted estimation problems with over 800 references; Turlach (2005) has a more recent review on shape constrained spline smoothing.

There has also been substantial research on non-parametric estimation via optimization. The work of de Montricher, Scott, Tapia and Thompson (de Montricher and Tapia, 1975; Scott, 1976; Scott et al., 1980; Thompson and Tapia, 1990), are representative. Nemirovskii et al. (1984, 1985) provide detailed analysis of the consistency of maximum likelihood estimators and rate of convergence. However, unlike the developments in the asymptotic analysis of these estimation problems, the majority of the algorithmic techniques used to date are relatively simple and ad-hoc, often specific to the (single) shape constraint involved in the problem, with little potential for generalization.

Perhaps because in most interesting function spaces the nonnegativity constraint is expressible only by an infinite collection of linear inequalities, there is a hesitation to tackle nonnegativity

(or other shape constraints) directly in these spaces. The following three quotes from well-known references are typical in the literature.
From (Ramsay, 1988):

> Attempting to impose monotonicity on polynomials quickly becomes unpleasant...

From (Thompson and Tapia, 1990) p. 103:

> The nonnegativity constraint [in density estimation] is, in general, impossible to enforce when working with continuous densities which are not piecewise linear.

From (Mammen and Thomas-Agnan, 1999):

> Constrained smoothing splines with infinitely many constraints (like $m^{(r)}(x) \geq 0$) for all x are difficult to compute...

In this paper we show that these concerns may not be all that warranted. In particular, in the case of univariate estimation problems, which is in the focus of our paper, we show that using the modeling and algorithmic tools offered by second order cone programming, semidefinite programming, and related conic optimization problems, we can easily tackle an array of shape constrained estimation problems, and find optimal nonnegative, monotone, and convex spline estimators.

In the remainder of this section we summarize some of the advantages of our method, most of which are not shared by earlier methods.

General scope. Techniques that do not involve optimization are usually designed to solve a specific problem, mostly involving a single shape constraint. They are computationally extremely efficient, but they also have limited scope. Methods based on numerical optimization can typically incorporate additional linear constraints without difficulty. This increases the types of constraints one may consider: periodicity and interpolation constraints are just two examples of such constraints: as they are expressible by linear equations and inequalities on the estimator, they can be added to any convex optimization model freely.

Several approaches have been proposed for shape constrained spline smoothing using splines of a specific degree, including (Hildreth, 1958) and (Brunk, 1958) (piecewise linear estimators); He and Shi (1996) (quadratic B-splines for monotone regression); and Dierckx (1980) and Turlach (2005) (cubic splines). All of these approaches exploit the fact that the considered shape constraints in the considered spaces is characterized by a finite collection of linear inequalities; hence, none of them generalizes to higher order splines than what they were developed for.

Our approach is based on a characterization of nonnegative polynomial splines of any degree that leads to computationally tractable optimization models of all of the considered estimation problems. This also takes care of our next concern:

Direct characterization of nonnegativity. If the estimator f is required to be nonnegative, then a simple way of imposing this constraint is to write $f = g^2$, or $f = \exp(g)$, and then turn the attention to g (Good and Gaskins, 1971). Of course, this is overly restrictive if nonnegativity is only required, say, over an interval. Moreover, linear constraints on f (such as periodicity and the requirement that f integrates to one) are transformed into non-linear equations on g, rendering the transformed optimization problems difficult to handle. Additionally, the underlying space of g may not be the same as that of f; in particular, g may not belong to an easily characterizable finite dimensional linear space even if f does.

Avoiding oversimplification and approximation. At the time of some of the earlier works on shape constrained estimation, semidefinite and second order cone programming methods were either not available or known, or were considered computationally too expensive, hence these studies often used oversimplified, inexact optimization models. Most of the models reviewed and proposed in the surveys (Robertson et al., 1988) and (Delecroix and Thomas-Agnan, 2000) and in the theses of Meyer (1996) and d'Aspremont (2004) only approximate shape constraints, and find optimal estimators in a proper subset of the functional cone of interest. These are usually sets of functions with nonnegative coefficients in a nonnegative basis. Ramsay's I-spline (Ramsay, 1988) is based on the same idea.

As opposed to the overly constrained approaches above, some impose shape restrictions only on a *finite* subset of the domain of the unknown function in order to obtain optimization models that are simpler than the ones desired to be solved. For instance, Delecroix et al. (1996), Mammen and Thomas-Agnan (1999), and (Villalobos and Wahba, 1987) consider polynomial splines with nonnegativity on the k-th derivative, but nonnegativity is imposed only on the knot points, or at finitely many evenly spaced points.

Since nonnegative splines can be effectively handled by convex optimization (SOCP/SDP) methods, such approximations are no longer necessary. Nonetheless, we explore this approach as well, and compare it to models that use the entire cone of nonnegative functions. In this paper we

examine estimating nonnegative polynomial splines generated by the Bernstein polynomial basis (rather than the more popular B-splines), and compare the results to the SDP/SOCP approach. We will justify our choice of this basis over the B-spline basis in Theorem 2.

Recently some authors have applied SDP and SOCP techniques to a few shape constrained estimation problems. Wang and Li (2008) used cubic splines for isotonic regression, and Fushiki et al. (2006) have used semidefinite programming constraints with log-likelihood objective function in parametric density estimation with nonnegative polynomials. Alizadeh et al. (2008) have used SOCP and SDP models to estimate the smooth arrival rate of nonhomogeneous Poisson process based on observed arrival rates using cubic splines. Our work is an attempt to provide a general framework for a large number of shape constrained estimation problems at a considerably higher level of generality than in the above studies.

2 Conic sieves and nonnegative functions

2.1 Estimation in general spaces

In the most general setup the goal of an estimation problem is to reconstruct an unknown (possibly multivariate) real-valued function $f \in \mathbb{H}$ in some space $\mathbb{H}$ (usually a reproducing kernel Hilbert space or a Sobolev space), based on finitely many *observations* $\mathbf{z}_1, \mathbf{z}_2, \ldots, \mathbf{z}_N$ drawn from a subset $D \subseteq \mathbb{R}^k$. Constraints on the shape of the function translate to the requirement that f belongs to a closed convex set $\mathcal{K} \subseteq \mathbb{H}$. (Only two of the commonly considered shape constraints are not convex: unimodality and logconcavity.) Finally, we seek a function f that minimizes a convex *loss functional* $L(\cdot|\mathbf{z}_1, \ldots, \mathbf{z}_N)$. Hence, a *shape constrained estimation problem* is an optimization problem of the form

$$\inf\{L(f|\mathbf{z}_1, \ldots, \mathbf{z}_N) \mid A_i(f|\mathbf{z}_1, \ldots, \mathbf{z}_N) \leq b_i \ (i = 1, \ldots, I); \quad f \in \mathcal{K}\}, \tag{1}$$

where $A_1, \ldots, A_I$ are $\mathbb{H} \to \mathbb{R}$ linear functionals, $b_1, \ldots, b_I$ are real numbers, and L and $\mathcal{K}$ are as explained above. A few remarks are in order.

1. Care has to be taken to ensure that the problem (1) has an optimal solution, meaning that the infimum is finite and is attained; this largely depends on the underlying space $\mathbb{H}$ and

the loss function L. When a smooth estimator is sought, the Sobolev space $\mathbb{W}_2^d([a,b])$ or a reproducing kernel Hilbert space is an appropriate choice for a number of commonly used loss functions (Kimeldorf and Wahba, 1971). Moreover, in certain shape constrained problems it is also known that the optimal solution belongs to a specific *finite dimensional* space, in particular to a space of polynomial splines with given knot points (Wahba, 1990, Chapter 1), (Eggermont and LaRiccia, 2001).

2. Constraints on the shape of f are expressed by the constraint $f \in \mathcal{K}$ (possibly together with some linear equations and inequalities). By the nature of shape constraints, $\mathcal{K}$ is usually a *convex, pointed cone*, that is, $0 \not\equiv f \in \mathcal{K}$ imply that $\lambda f \in \mathcal{K}$ for every $\lambda \geq 0$ but $-f \notin \mathcal{K}$. For example, $\mathcal{K}$ can be the cone of nonnegative, monotone non-decreasing, or convex functions in $\mathbb{H}$. Multiple shape constraints can be modeled by considering a cone $\mathcal{K}$ that is created by intersection, Cartesian product, or Minkowski sum of the cones modeling the individual shape constraints.

3. If $\mathbb{H}$ is finite dimensional (which happens, for instance, when parametric models or finite dimensional approximations of infinite dimensional problems are considered), $\mathcal{K}$ is often the cone of nonnegative functions in $\mathbb{H}$, which we denote by $\mathcal{P}$. Many other shape constraints, especially in the univariate case, can be reduced to this case. For example, derivatives of f may be sign-constrained, ensuring monotonicity, convexity or concavity of f.

2.2 Sieves and finite dimensional approximations

The term *sieve* was coined by Grenander (1981) for a sequence of subsets $S_1, S_2, \ldots, S_m, \ldots$, of some metric space of functions Ω containing the unknown function; it is required that $\bigcup S_m$ be dense in Ω. The main idea is that for any kind of estimation problem, instead of minimizing a given loss functional on the space Ω, we search in S_m for some m, which is an easier problem. The density requirement ensures that some function in some S_m is a good approximation of the function that is being estimated. As m increases, it is assumed that the "complexity" of the functions under consideration also increases. The appropriate value of m is usually determined by techniques such as cross-validation. Note that the idea is almost independent of the loss functional to be minimized. Thus, sieve methods have been applied to regression, density estimation, and

even arrival rate estimation.

In this paper, we follow a restricted version of the sieve method. In particular, we require that sets in the sieve sequence are all finite dimensional convex cones, and that the sequence is *nested* or *asymptotically nested*:

Definition 1. Let $\mathcal{K} \subseteq \Omega$ be a closed, pointed, convex cone. A sequence $\mathcal{K}_1, \mathcal{K}_2, \ldots$ of closed, pointed and convex cones satisfying $\mathcal{K}_m \subseteq \mathcal{K}$ is called a *conic sieve* if each $\mathcal{K}_m$ is finite dimensional, and $\bigcup \mathcal{K}_m$ is dense in $\mathcal{K}$. We say that the sieve is *nested* if $\mathcal{K}_1 \subseteq \mathcal{K}_2 \cdots \subseteq \mathcal{K}_m \subseteq \cdots$, and it is *asymptotically nested* if for each m the sequence has a nested infinite subsequence containing $\mathcal{K}_m$.

Prior information, including assumptions on the shape of the estimator are particularly useful when the number of samples is too small for the particular shape to be immediately clear based on the data alone. Therefore, we shall not be particularly concerned with the behavior of the estimators in the case when the number of samples reaches the asymptotic range. Nevertheless, we shall mention that existing results on the consistency of constrained estimators are applicable, and prove the consistency of the estimators proposed in the paper. The results of Geman and Hwang (1982) prove the consistency of the maximum likelihood estimators of the present paper. Dong and Wets (2000) consider a more general setup for density estimation, where the negative log-likelihood loss function can be replaced with other convex loss functions, including least-squares and penalized log-likelihood. (Simultaneously, they make a case for the use of constrained ML estimators instead of penalized ML estimators.)

There is one additional technical restriction that needs to be added to the conic version of the sieve method: in the optimization models obtained by replacing $\mathcal{K}$ with the finite dimensional approximation $\mathcal{K}_m$ in (1), an additional constraint $\|f\| \leq B_m$ on the norm of f must be added to keep the set of feasible f's compact. (Note that since the optimization takes place in a finite dimensional space, all norms are equivalent.) With this addition the theorems of (Geman and Hwang, 1982) and (Dong and Wets, 2000) are directly applicable to conic sieves: as long as the bound increases slowly enough with m, the resulting estimators are consistent. Such bounds are often *a priori* imposed on the problem. For instance, nonnegative splines (with given knot points) that integrate to one form a bounded set.

The most important examples of conic sieves for our purposes are the sieves of *polynomial splines*. Recall that a *univariate polynomial spline* f of *degree* (or *order*) n and *continuity* $\mathcal{C}^r$

($r \leq n-1$) is a real valued function on $[a,b] = [a_0, a_m]$ defined piecewise on the intervals $[a_i, a_{i+1}]$, $i = 0, \ldots, m-1$ with the following properties: (1) f is a polynomial of degree n over each interval $[a_i, a_{i+1}]$, $i = 0, \ldots, m-1$; (2) f has continuous derivatives up to order r over (a,b). The points a_i are called the *knot points* of the spline. Throughout the paper, the sequence $(a_0, \ldots, a_m)$ will be abbreviated as $\mathbf{a}$. The length $\max_{0 \leq i \leq m-1}(a_{i+1} - a_i)$ of the longest subinterval in the knot point sequence is called the *mesh size* of $\mathbf{a}$; it is denoted by $\|\mathbf{a}\|$. The linear space of all splines of degree n and knot point sequence $\mathbf{a}$ is denoted by $\mathcal{S}(n, \mathbf{a})$. In this paper we will always assume $r = n-1$. Schumaker (1981) has a more general definition, where existence of derivatives of different orders are required at different knot points. Wahba (1990) and others consider only *natural splines*, which are more restricted on the first and last subinterval of the domain. The methods proposed in this paper can be adapted to these definitions without any difficulty.

Suppose that the degree n is fixed, and $\mathcal{K}_m$ is chosen to be $\mathcal{S}(n, \mathbf{a}_m)$, where $\mathbf{a}_1, \mathbf{a}_2, \ldots$ is an infinite sequence of knot point sequences with mesh sizes approaching 0. Then by (Schumaker, 1981, chap. 6) we have that $\bigcup \mathcal{K}_m$ is dense in the Sobolev space $\mathbb{W}_2^n\big([a,b]\big)$, and hence $\bigcup(\mathcal{K}_m \cap \mathcal{P})$ is dense in $\mathbb{W}_2^n\big([a,b]\big) \cap \mathcal{P}$. (As before, $\mathcal{P}$ denotes the set of nonnegative functions.) The sequence $(\mathcal{K}_m)_{m=1,\ldots}$ is not necessarily nested, however it is not difficult to create a subclass that forms a nested or asymptotically nested sequence. For instance, we may start from one interval $[a,b]$, and then recursively add the midpoint of the rightmost longest interval to the set of knots. The most straightforward way to obtain an asymptotically nested sieve is to consider knot point sequences that subdivide $[a,b]$ uniformly.

For polynomial spline estimators the subscript m of $\mathcal{K}_m$ is the number of pieces of the spline. As we shall see, in the optimization models the estimator is represented by the coefficients of its polynomial pieces, and the upper bound B_m on the norm of the spline can be imposed by giving an upper and a lower bound on these coefficients. These only require adding linear inequalities to the optimization models.

In each spline space $\mathcal{S}(n, \mathbf{a})$ there may be a number of different convex cones that encode some type of shape constraint. Let us for the moment concentrate on the cone $\mathcal{P}^{[n,\mathbf{a}]}$ of nonnegative functions in $\mathcal{S}(n, \mathbf{a})$. This cone is the Cartesian product of nonnegative polynomials of degree n (where the i-th polynomial should be nonnegative on $[a_i, a_{i+1}]$), intersected with the linear space of $\mathcal{C}^r$ functions. Therefore, $\mathcal{P}^{[n,\mathbf{a}]}$ is a convex, but in general non-polyhedral, cone. One of the

main results of Section 3, and of the paper, is that problems of the form (1) with $\mathcal{K} = \mathcal{P}^{[n,\mathbf{a}]}$ can be solved efficiently for every n and $\mathbf{a}$. When considering sufficiently differentiable functions, many shape constraints mentioned in the Introduction reduce to the nonnegativity (or nonpositivity) of the derivatives. Hence, sieves of monotone non-increasing or non-decreasing, convex or concave functions using polynomial approximations can be defined and characterized analogously to the sieves of nonnegative functions; it is sufficient to consider nonnegativity as the "universal" shape constraint.

3 Representations of Nonnegative Polynomials and Splines

3.1 Representations involving Semidefinite and Second Order Cone constraints

In this section we summarize some well-known results about the characterization of nonnegative polynomials, and apply this theory to characterize nonnegative splines. For more details on nonnegative polynomials the reader is referred to (Karlin and Studden, 1966). As mentioned at the end of the previous section, this immediately gives rise to analogous characterizations of non-increasing, non-decreasing, concave, and convex splines, by simply imposing nonnegativity on the derivatives of the spline. A generalization of these results, and examples of other functional spaces where nonnegative functions have analogous characterizations, can be found in (Papp and Alizadeh, 2011).

We will use the following notations and conventions: vectors (resp., matrices) are typeset boldface, their components (resp., entries) are denoted with the corresponding lowercase italic character. Indexing of vectors and matrices starts from 0 rather than 1. For example, the $n+1$ dimensional row vector $\mathbf{p}$ could also be written as $(p_0, \ldots, p_n)$.

The inequality $\mathbf{X} \succcurlyeq 0$ denotes that $\mathbf{X}$ is a positive semidefinite real symmetric matrix. The *cone generated by* the set U, defined as $\{\alpha\mathbf{u} \colon \alpha \geq 0, \mathbf{u} \in U\}$ is denoted by $\operatorname{cone}(U)$; $\operatorname{int} S$ denotes the interior of the set S.

In this section, and throughout the paper, we assume (for notational simplicity) that unknown polynomials in optimization models are represented in the standard monomial basis, even though this basis is numerically rather poorly behaved. This means that we also identify the polynomial

function p with the coefficient vector $\mathbf{p}$ of the polynomial. There is no conceptual difficulty in modifying all theorems and algorithms in this paper so that they involve polynomials represented in any other basis, such as some orthogonal polynomial basis.

The main results on the representation of nonnegative polynomials over an interval are summarized in the following two theorems. We split the odd and even degree cases into separate propositions for better readability.

Proposition 1 (Odd degree case, Karlin and Studden 1966)**.** *Let $p = \sum_{i=0}^{n} p_i x^i$ be a polynomial of degree $n = 2k+1$, and $a < b$ be real numbers. Then $p(x) \geq 0$ for all $x \in [a,b]$ if and only if there exist symmetric $(k+1) \times (k+1)$ matrices $\mathbf{X} = (x_{ij})_{i,j=0}^{k}$ and $\mathbf{Y} = (y_{ij})_{i,j=0}^{k}$ satisfying $\mathbf{X} \succcurlyeq 0, \mathbf{Y} \succcurlyeq 0$, and*

$$p_\ell = \sum_{i+j=\ell} (-a x_{ij} + b y_{ij}) + \sum_{i+j=\ell-1} (x_{ij} - y_{ij}) \tag{2}$$

for all $\ell = 0, \ldots, 2k+1$.

Proposition 2 (Even degree case, Karlin and Studden 1966)**.** *Let $p = \sum_{i=0}^{n} p_i x^i$ be a polynomial of degree $n = 2k$, and $a < b$ be real numbers. Then $p(x) \geq 0$ for all $x \in [a,b]$ if and only if there exist a symmetric $(k+1) \times (k+1)$ matrix $\mathbf{X} = (x_{ij})_{i,j=0}^{k}$ and a symmetric $k \times k$ matrix $\mathbf{Y} = (y_{ij})_{i,j=0}^{k-1}$ satisfying $\mathbf{X} \succcurlyeq 0, \mathbf{Y} \succcurlyeq 0$, and*

$$p_\ell = \sum_{i+j=\ell} (x_{ij} - ab y_{ij}) + \sum_{i+j=\ell-1} (a+b) y_{ij} - \sum_{i+j=\ell-2} y_{ij} \tag{3}$$

for all $\ell = 0, \ldots, 2k$.

A useful property of quadratic and cubic polynomials nonnegative over an interval is that (following the above propositions) their characterizations involve only 2×2 positive semidefinite matrices. Positive semidefiniteness of 2×2 matrices can be translated to linear and quadratic (second order cone) constraints using the following, well-known fact:

Proposition 3. *The matrix $\left(\begin{smallmatrix} x_0 & x_1 \\ x_1 & x_2 \end{smallmatrix}\right)$ is positive semidefinite if and only if $(x_0 + x_2, x_0 - x_2, 2x_1)^\top \in \mathcal{Q}_3$, where $\mathcal{Q}_{k+1} = \left\{(z_0, \ldots, z_k) \colon z_0 \geq \left\|(z_1, \ldots z_k)^\top\right\|_2\right\}$ is the $(k+1)$-dimensional* second order cone *(or* Lorentz cone*).*

Constraints of the form $\mathbf{x} \in \mathcal{Q}_{k+1}$ are called *second order cone constraints*, while constraints of the form $\mathbf{X} \succcurlyeq 0$ are *semidefinite constraints*. Page limitations do not allow us to give a

comprehensive overview of second order cone programming (SOCP) and semidefinite programming (SDP), and deep knowledge of these fields is not necessary to apply the methods proposed in this paper, but a brief description is included in the Appendix, along with pointers to software available for the solution of SDPs and SOCPs. The reader is also encouraged to consult (Alizadeh and Goldfarb, 2003) for an accessible survey on SOCP and (Wolkowicz et al., 2000) for an in-depth review of many aspects of SDP.

The above characterization of nonnegative polynomials easily extends to a characterization of nonnegative splines over $[a_0, a_m]$: the nonnegativity of each polynomial piece is translated to a set of semidefinite constraints and linear equations, and the continuity of the derivatives translates to another finite set of linear equations.

The equalities in (2) and (3) suggest that we may run into serious numerical problems if the knot points of the spline are distributed unevenly, as the linear equations in the characterization will have coefficients that may differ by many orders of magnitude. This can be avoided by *scaling*: for each $i = 0, \ldots, m-1$ we apply an affine transformation on the ith polynomial piece of the spline that maps the interval $[a_i, a_{i+1}]$ to $[0, 1]$, and represent the spline S between the knot points a_i and a_{i+1} by the coefficients of the thus transformed polynomial $p^{(i)}$, rather than by the original coefficients. By way of formulas, the resulting *scaled representation* of the spline S is the following:

$$S(x) = p^{(i)}\left(\frac{x - a_i}{a_{i+1} - a_i}\right) = \sum_{k=0}^{n} p_k^{(i)} \Big(\frac{x - a_i}{a_{i+1} - a_i}\Big)^k \quad \forall x \in [a_i, a_{i+1}], \quad i = 0, \ldots, m-1, \qquad (4)$$

where each $p^{(i)}$ is a polynomial defined on $[0, 1]$, with coefficients $p_0^{(i)}, \ldots, p_n^{(i)}$ in the standard monomial basis. In the scaled representation of splines the nonnegativity and continuity constraints are entirely independent of the location of the knot points, as nonnegativity is expressible as $p^{(i)}(x) \geq 0$ for every $i = 0, \ldots, m-1$ and $x \in [0, 1]$, and the continuity of the spline is simply expressed as $p^{(i+1)}(0) = p^{(i)}(1) = 1$ for every $i = 0, \ldots, m-1$. The linear equations expressing the continuity of the higher order derivatives also depend only on the ratios of the distances between consecutive knot points $(a_{i+1} - a_i)/(a_i - a_{i-1})$.

As an example, the complete list of constraints that characterize a nonnegative cubic spline of continuity $\mathcal{C}^2$, with knot points $a_0, \ldots, a_m$ is provided in Theorem 3 in the supplementary material (Appendix B).

3.2 Polyhedral cones of splines

In this section we examine polyhedral approximations of cones of nonnegative splines. We concentrate on models of the following form: we fix a basis $U = \{u_0, \ldots, u_n\}$ of degree n polynomials nonnegative over $[0, 1]$, and then consider splines whose coefficients $p_k^{(i)}$ in the scaled representation (4) are all nonnegative. Let $\mathcal{P}(U, \mathbf{a})$ denote the set of such splines with knot point sequence $\mathbf{a}$. These are clearly a subset of all nonnegative splines.

As mentioned in the Introduction, the approach of approximating nonnegative splines by functions with nonnegative coefficients in a nonnegative basis (henceforth called the *nonnegative basis method*) is certainly not new. But to our knowledge there has not been any systematic analysis of what bases may or may not be used in such an approach. The following simple example shows that this question is of prime importance. Consider, in our notation, the basis $U = \{1, x, \ldots, x^n\}$ of degree n polynomials. This is a nonnegative basis over $[0, 1]$. Since these functions are also monotone non-decreasing, it is immediate that for every knot point sequence $\mathbf{a}$, the cone $\mathcal{P}(U, \mathbf{a})$ consists only of monotone non-decreasing functions. Therefore, optimization over $\mathcal{P}(U, \mathbf{a})$ will not yield useful estimators if a general (possibly decreasing) nonnegative estimator is sought. As we shall see below, using the Bernstein polynomial basis, defined by $u_i(x) = \binom{n}{i} x^i (1-x)^{n-i}, (i = 0, \ldots, n)$, in place of the monomial basis is a theoretically sound choice.

Our first result in this section is a sufficient condition for a sequence of polyhedral spline approximations to form a sieve. Recall that the mesh size of the knot point sequence $\mathbf{a}$ is denoted by $\|\mathbf{a}\|$.

Theorem 1. *Consider a basis $U = \{u_0, \ldots, u_n\}$ of polynomials of degree $n \geq 1$ such that each u_i is nonnegative over $[0, 1]$, and assume that $1 \in \operatorname{int} \operatorname{cone}(U)$, where 1 denotes the constant one polynomial. Furthermore, let $\{\mathbf{a}_i\}$ be an asymptotically nested sequence of knot point sequences in $[0, 1]$ satisfying $\lim_{i\to\infty} \|\mathbf{a}_i\| = 0$. Then the set $\bigcup_i \mathcal{P}(U, \mathbf{a}_i)$ is a dense subcone of $\mathcal{P} \cap C([0, 1])$, the cone of nonnegative functions over $[0, 1]$.*

See Appendix C in the supplementary material for the proof.

As Bernstein polynomials of degree n sum to the constant polynomial 1, we can construct, for every n, a sieve that consists of polyhedral cones of n times differentiable polynomial splines.

Corollary 1. *For every n, the cone of polynomial splines of degree n whose pieces have nonnegative*

weights in the Bernstein polynomial basis is a dense polyhedral subcone of nonnegative continuous functions over $[a, b]$ *consisting entirely of* $n - 1$ *times differentiable functions.*

Henceforth we shall call this subset of nonnegative splines *piecewise Bernstein polynomial splines.*

Our last observation about polyhedral sets of splines is about *B-splines.* B-splines are particularly popular in the approximation and engineering literature because of their excellent theoretical and computational properties. However, as it has been recently shown, cones generated by B-splines are proper subcones of piecewise Bernstein polynomial splines.

Theorem 2 (Papp 2011, Thm. 3.6)**.** *For every positive integer* n *and knot point sequence* $\mathbf{a}$*, the cone of functions generated by B-splines of degree* n *with knot points* $\mathbf{a}$ *is a subset of the cone of piecewise Bernstein polynomial splines of the same degree, with knot points* $\mathbf{a}$*. For* $n \geq 2$ *this containment is strict.*

Note that if the knot points and the degrees are fixed, piecewise Bernstein polynomial splines and B-splines have the same degrees of freedom. Therefore, piecewise Bernstein polynomial splines provide a better approximation of nonnegative splines at no additional cost. Hence, we do not consider B-splines in this paper any further.

3.3 Knot point selection

In each of the spline models above we have assumed that a fixed sequence of knot points $a_0, \ldots, a_m$ is given. Finding the best selection of knot points is a central, but very difficult, problem. Ideally, we would make the knot points variables, and optimize over them as well as the coefficients of the polynomials, but this would result in an intractable, non-convex optimization problem.

A common practice is to use evenly spaced knot points. While this is a very crude method, it is also very simple, and it fits the conic sieve framework discussed in Section 2, as splines with evenly spaced knot points give rise to an asymptotically nested sieve. In our numerical examples we used this method.

Another possibility is to place knot points at the data points. A theoretical result supporting this idea is that the optimal solutions to certain estimation problems (least squares regression with a penalty term for smoothing) are natural cubic splines with knots at the data points (Wahba,

1990). We can start with a trivial subdivision (with two knot points, one at each endpoint of the domain), and add knot points one by one, at each step subdividing one of the intervals with the largest number of data points in it. This gives rise to a nested sieve.

In either case the optimal number of knot points can be found by common model selection procedures: validation on a test set (if there is one), cross-validation or k-folding (if there is no separate test set), or by using some information criterion, such as the Akaike or the Bayesian Information Criterion (AIC and BIC, respectively); see (Burnham and Anderson, 2004). All models proposed in this paper are solvable quickly enough that even using leave-one-out cross-validation is computationally feasible. In the computational experiments, when the knot point sequences were not nested, but asymptotically nested, we used leave-one-out cross-validation. Whenever we used nested sieves, and the objective function had no smoothing penalty, we used AIC_c for model selection for simplicity.

4 Optimization models for shape constrained estimation

This section reviews a number of estimation problems that are solvable with the techniques proposed in the paper. Numerical illustrations of these applications can be found in Section 5.

4.1 Nonparametric regression of a nonnegative function

In nonnegative regression our goal is to estimate a function f based on data $\mathbf{z}_i = (x_i, y_i)$ $i = 1, \ldots, N$, assumed to come from the model $y_i = f(x_i) + \varepsilon_i,\ i = 1, \ldots, N$, where ε_i are independent, identically distributed random variables with mean zero, and the function f is assumed to belong to a class of nonnegative functions $\mathcal{F}$. The goodness-of-fit of f to the data is measured by some loss functional of the form $L(f|\mathbf{z}_1, \ldots \mathbf{z}_N) = d(f) + s(f)$, where the term $d(f)$ measures the distance of the function values $f(x_i)$ and y_i, while $s(f)$ is a penalty (or smoothing) term that penalizes "rough" or overly complex solutions.

A common choice for $d(f)$ is the *residual sum of squares* $d(f) = \sum_{i=1}^{N}(f(x_i) - y_i)^2$. The smoothing term $s(f)$ may be omitted. If present, it is typically chosen to be $\int |f'|$, $\int |f''|$, or $\int (f'')^2$. If f is estimated with nonnegative splines of degree three or less, then all the above choices of $s(f) + d(f)$ lead to optimization models with only linear and second order cone constraints.

Further possible constraints include periodicity and interpolation constraints, which can be modeled by adding linear equations to the set of constraints. This does not change the difficulty of the optimization models.

4.2 Monotone, Convex, and Concave Regression

Suppose the estimator of the unknown function f belongs to a linear space of twice differentiable functions whose first and second derivatives lie in a space where nonnegativity is easily characterized. Then any combination of monotonicity, concavity, and convexity constraints on the estimator can be added to the optimization models without any difficulty, as these constraints reduce to sign constraints on the derivatives. When using cubic splines, convexity and concavity can be expressed by finitely many linear constraints, since f'' is piecewise linear.

In the presence of multiple shape constraints further simplifications may be possible, especially when splines of low degree are used. For example, a concave function f is nonnegative over $[a, b]$ if and only if $f(a) \geq 0$ and $f(b) \geq 0$. Consequently, in the presence of the concavity constraint the nonnegativity constraint simplifies to two linear inequalities.

4.3 Unconstrained Density Estimation

One can formulate the estimation of a probability density function (*pdf*) from a finite set of independent samples $\{X_1, \ldots, X_n\}$ as an optimization problem involving nonnegative functions.

A pdf must be nonnegative and integrate to one. We assume that the pdf to be estimated has finite support, say $[a_0, a_m]$, and that it is continuous, therefore it can be approximated by nonnegative polynomial splines of a fixed degree. When using a spline model, the condition that the pdf should integrate to one simplifies to a linear constraint, since the integral of a polynomial on a given interval is a linear function of the coefficients of the polynomial. For example, a cubic spline model can be constructed by adding the constraint $\sum_{i=0}^{m-1} \sum_{j=0}^{3} \frac{a_{i+1}-a_i}{j+1} p_j^{(i)} = 1$ to the characterization of nonnegative cubic splines provided in Theorem 3 of the Appendix.

Finally, the objective function needs to be determined. The most common and straightforward approach is *maximum likelihood* estimation. If the unknown pdf is denoted by f, this amounts to maximizing the likelihood function $\prod_{i=1}^{n} f(X_i)$. This objective function will cause numerical problems, and furthermore is not necessarily a concave function, which makes its maximization

difficult. Instead, we will use the negative log-likelihood function $-\sum_{i=1}^{n} \log f(X_i)$ as the loss function, which is convex if f is a polynomial spline of a given knot sequence. This way we obtain a convex optimization model with only linear and second order cone or semidefinite constraints, depending on the degree of the spline. (The importance of the optimization model being convex is that local optima are global, making the optimization considerably easier.) It is also important to note that by constraining f to be a polynomial spline, the above maximum likelihood optimization problem is always well-defined: it has an optimal solution for every fixed set of knot points.

4.4 Unimodal Density Estimation

Further constraints may be added to the density estimation problem of the previous section for *unimodal* density estimation. If the mode is known, we can place one of the knot points on the mode, and then add constraints that the spline is increasing from the first knot point to the mode, and decreasing from the mode to the last knot point.

If the mode is unknown, we have a non-convex problem: the set of unimodal functions is not convex simply because the sum of two unimodal functions is not necessarily unimodal. In this situation an approximate solution to the problem can be found by solving a sequence of optimization models, each with a different mode, and comparing the optimal solutions that correspond to the different modes. Finding the exact solution this way is still nontrivial, as the maximum likelihood function is not a unimodal function of the mode.

5 Numerical Examples

In this section we have compiled results of a number of numerical experiments in which the SOCP-based sieves of cubic and quartic splines (with uniform, asymptotically nested knot points) were compared to piecewise Bernstein polynomial spline models, to kernel methods, to smoothing splines, and in one case to parametric models. Owing to the large number of models and problems, as well as page limitations, these results are included only as numerical illustrations, which demonstrate the wide applicability and computational feasibility of the models of Sections 3 and 4. A comprehensive empirical comparison of all the available methods in each of the shape constrained estimation problems covered by our framework would require, and perhaps merits,

a separate paper. Several more results, along with more details on the experiments below, are included in (Papp, 2011, sec. 3.5).

The optimization models described in Sections 3 and 4 were implemented using the AMPL modeling language, and were solved using the nonlinear solvers KNITRO version 5.1 (Nocedal and Waltz, 2003) or CPLEX version 12.3 (CPLEX, 2011).

5.1 Density estimation

Tests were conducted to compare our methods to a wide range of kernel methods for density estimation. The choice of benchmark distributions, as well as the experimental design is based on (Eggermont and LaRiccia, 2001, chap. 8); the distributions are shown on Fig. 1. The probability density functions of the benchmark distributions are:

$$
\begin{aligned}
f_1(x) &= \frac{9}{10}\phi_{1/2}(x-5) + \frac{1}{10}\phi_{1/2}(x-7), & f_2(x) &= \phi_1(x-5),\\
f_3(x) &= \frac{1}{5}U([3,8]), & f_4(x) &= \frac{1}{5}\psi_{1.4,2.6}\left(\frac{1}{5}(x-0.3)\right),\\
f_5(x) &= \frac{1}{4}\phi_{9/5}(x-6) + \frac{4}{5}\phi_{1/10}(x-2), & f_6(x) &= \frac{1}{2}\phi_{1/2}(x-3.5) + \frac{1}{2}\phi_{1/2}(x-6.5),
\end{aligned}
$$

where $\phi_\sigma(x)$ is the pdf of the normal distribution with mean zero and standard deviation σ, $U([a,b])$ is the pdf of the uniform distribution on $[a,b]$, and $\psi_{\alpha,\beta}$ is the pdf of the Beta density with parameters α and β. (See Fig. 1).

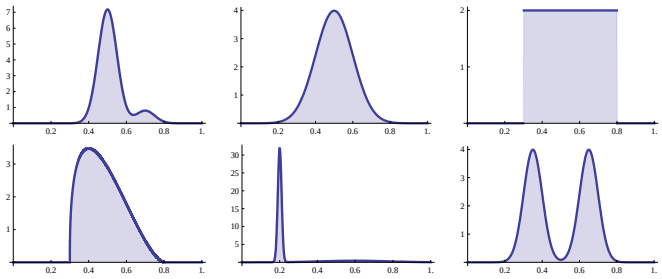

Figure 1: Benchmark distributions from Eggermont and LaRiccia (2001).

For each benchmark density, random samples of size 100 were generated. Then the optimal cubic spline densities were determined using the Bernstein polynomial based and the SOCP based methods outlined in Sections 3.1 and 3.2, and compared to the kernel estimates of Eggermont and LaRiccia (2001), which employ the Epanechnikov kernel and the normal density kernel, and

thirteen bandwidth selection methods. These bandwidth selection methods include the "optimal method", which simply determines the bandwidth that minimizes the L_1 error. This is clearly not a rational method, as it requires the knowledge of the estimated pdf, but it serves as a perfect benchmark, as it is an upper bound on the performance of all possible bandwidth selection methods. The process was repeated 100 times, the statistics of the L_1 distances of the estimators and the true pdfs are reported in Fig. 2.

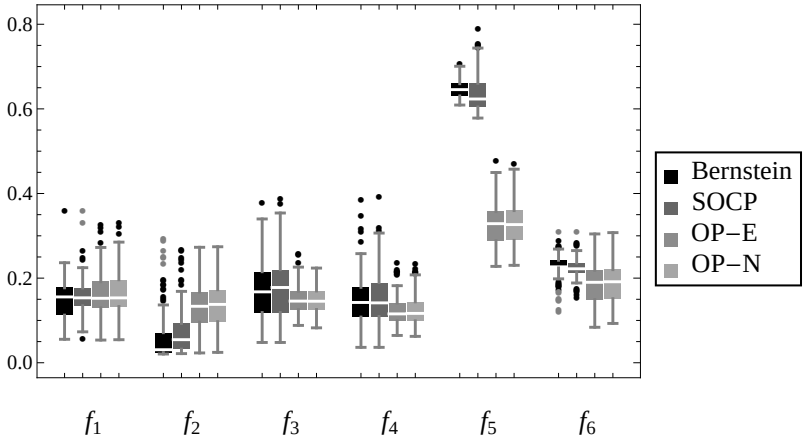

Figure 2: Boxplots showing the median, the range, and the inter-quartile range of the L_1 errors of "optimal" kernel and spline estimates in the density estimation benchmarks; dots mark outliers. OP-E and OP-N are the lower bounds on the errors of the best possible kernel estimates using the Epanechnikov and normal density kernels; Bernstein and SOCP are actual errors from an untuned implementation of the proposed methods.

It is safe to conclude that in three examples both the Bernstein polynomial-based and the SOCP methods give comparable results to the upper bounds of best possible kernel methods. There is significant difference between only in the results with f_2 (favoring Bernstein and SOCP; Mann–Whitney test p-value $< 10^{-10}$), and with f_5 and f_6 (favoring kernel methods; p-values $< 10^{-9}$). The results are easy to interpret: the less smooth the estimated pdf is, the higher the disadvantage of cubic splines to the kernel methods, and for very smooth pdfs the spline estimators clearly outperform kernel methods.

The difference between Bernstein polynomial-based and the SOCP methods is mostly insignificant, but SOCP did better in the least smooth examples, f_5 and f_6 (Mann–Whitney test p-values $1.4 \cdot 10^{-4}$ and 0.023, respectively.) The bounds OP-E and OP-N are insignificantly different from

each other (p-value > 0.749 for each test function).

5.2 Isotonic and convex/concave regression

5.2.1 Monotone regression – shape constraint versus smoothing penalty

In this section we outline an experiment we used to compare the effect of imposing shape restrictions on the estimator to the effect of using a smoothing penalty.

We simulated data using the model $Y = f(X) + \varepsilon$, where f is a smooth function given by

$$f(x) = 5 + \sum_{i=1}^{4} \operatorname{erf}(15i(x - i/5)) \qquad x \in [0, 1], \tag{5}$$

where $\operatorname{erf}(x) = \frac{2}{\sqrt{\pi}} \int_0^x e^{-x^2}$ is the error function, and ε is normally distributed with mean 0. The standard deviation σ of ε was varied in different experiments. The function f was chosen so that it is increasing, yet it has a number of essentially flat sections, as well as strictly increasing sections of various slopes. As a result, this function is likely to expose the shortcomings of regression methods that do not include monotonicity as a constraint in their model. With its step-function-like behavior, this function is also expected to expose the oscillation problems that often arise in estimation with polynomials.

Random samples of size 100 were drawn uniformly from the interval $[0, 1]$. Optimal cubic spline estimators were selected from the unconstrained and monotone increasing cubic splines with up to 100 uniformly placed knot points; the final number of knot points was selected by cross-validation.

We also considered smoothing splines: cubic splines that minimize the penalized residual sum of squares objective function $d(f) + s(f) = \sum_{i=1}^{N} (f(x_i) - y_i)^2 + \lambda \int_0^1 (f''(x))^2 \, \mathrm{d}x$, where $\lambda > 0$ is the smoothing parameter. The optimal number of knot points and the optimal λ were both determined by cross-validation. We considered both unconstrained and monotone increasing splines with smoothing penalty.

We compared the resulting four estimators by measuring the L_1 and L_2 distances of the function f and the estimators. This process was repeated 100 times with two different noise levels: $\sigma = 0.15$ and $\sigma = 0.3$. Boxplots of the L_1 errors are shown in Fig. 3; the plots corresponding to the L_2 distances are similar.

Sample plots of some of the optimal estimators are shown in Fig. 4. These plots are typical in

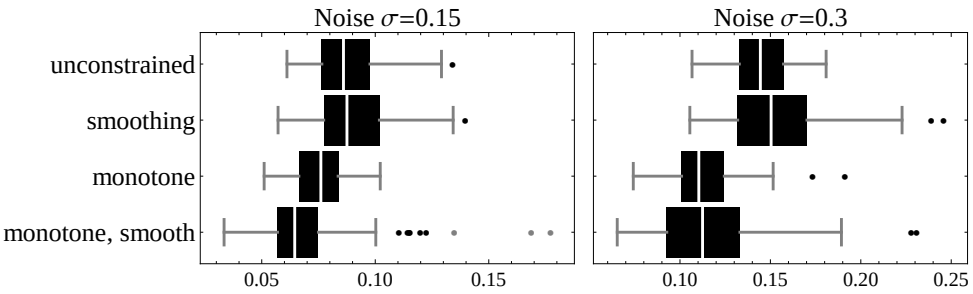

Figure 3: Comparison of the unconstrained and the monotone spline estimators, with and without smoothing penalty, for the regression curves of a dataset simulated using the model (5) with noise levels $\sigma = 0.15$ and $\sigma = 0.3$. The boxplots show the median, the inter-quartile range, and the range of the L_1 distances between the estimators and the true regression function; dots mark outliers. Imposing monotonicity on the estimator improves the quality of the estimation significantly, both for unconstrained and smooth estimators. The smoothing penalty does not have the same effect.

that the unconstrained splines generally showed considerably more oscillation than the monotone estimators (which, of course, cannot oscillate).

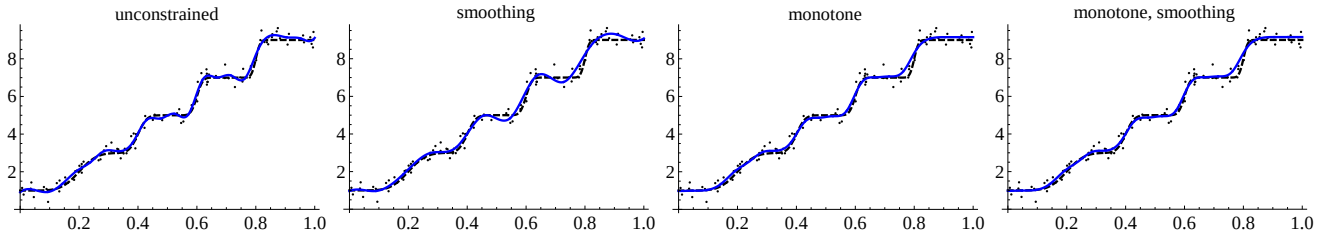

Figure 4: Unconstrained and monotone spline estimators for the regression curve of a simulated dataset. The function (5), to be estimated, is shown in black dashed line, the solid blue curves are the estimators. Left to right: unconstrained cubic spline, unconstrained smoothing spline, monotone increasing spline, monotone increasing smoothing spline. Smoothing reduces, but does not eliminate oscillation; in the presence of the monotonicity constraint smoothing does not yield noticeable improvement.

It is instructive to compare the distribution of the number of knot points of the four spline varieties considered; Fig. 5 shows the empirical distribution for the $\sigma = 0.15$ case. It is obvious from the figure that unlike the shape constraint, smoothing markedly increases both the number of knot points and the variability of the number of knot points of the optimal spline estimators.

The same trend was observed for the $\sigma = 0.3$ noise level. (Figure omitted.)

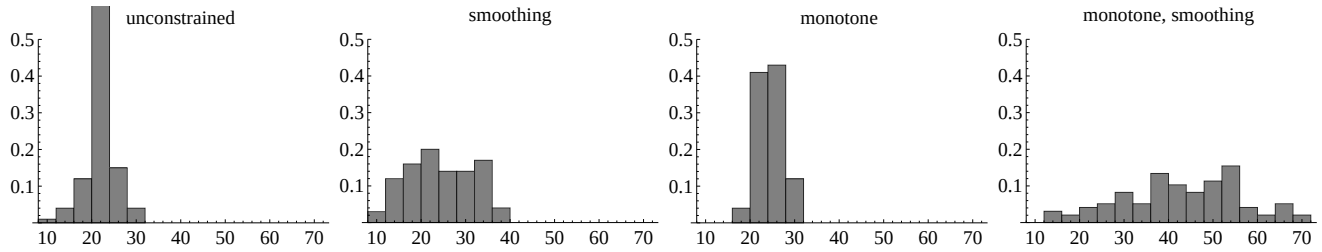

Figure 5: The empirical distribution of the number of knot points of the optimal spline estimators. Left to right: unconstrained cubic spline, unconstrained smoothing spline, monotone increasing spline, monotone increasing smoothing spline. Smoothing markedly increases the number of knot points and the variability in the number of knot points.

It is also interesting to note that unconstrained smoothing splines did not yield monotone estimators in any of the 200 experiments.

5.2.2 Mixed shape constraints and higher degree splines

In this section we outline an experiment we used to compare the effect of imposing multiple shape restrictions on the estimator, and the effect of increasing the degree of the spline estimator.

We simulated noisy data using the model $Y = f(X) + \varepsilon$, where $f(x) = \frac{1}{1+e^{-10x}}$, $x \in [0, 1]$, and ε is normally distributed with mean 0 and standard deviation 0.2. This function was chosen so that the function has a nearly linear increasing, and also a long, nearly horizontal part on the domain – this way it is likely that explicit monotonicity and concavity constraints will be required for a good quality fit.

Random samples of size 50 were drawn uniformly from the interval $[0, 1]$. As a baseline for the evaluation of the quality of the estimators, for each sample the least-squares optimal model from the one-parameter family $f_b(x) = \frac{1}{1+e^{-bx}} + \varepsilon$ and the two-parameter family $f_{a,b}(x) = \frac{1}{a+e^{-bx}} + \varepsilon$ were computed. We compared these models to different shape constrained polynomial spline models. These were obtained similarly to the spline estimators of the previous section, except that in this example we compared splines of different degrees and parametric models rather than comparing smoothing splines to splines without smoothing. The comparisons were made by measuring the L_1 and L_2 distances of the function f and the estimators. This process was repeated 100 times.

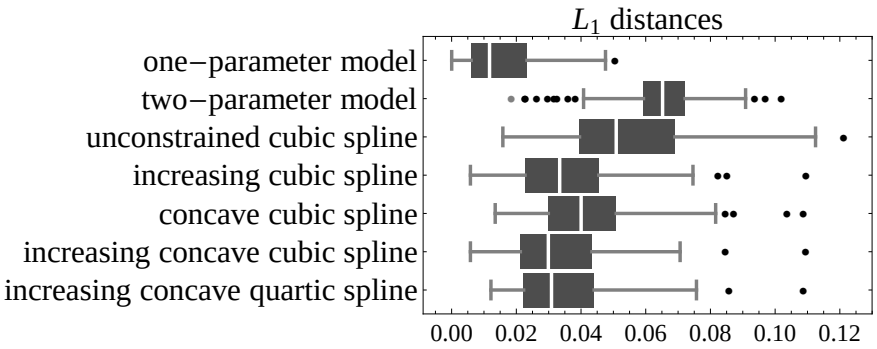

Figure 6: Comparison of parametric and spline estimators for the regression curves of a simulated dataset. The parametric models are best least-squares fits from a one- and a two-parameter family containing the true function. The remaining estimators are nonnegative splines with different combinations of additional shape constraints imposed on them. The boxplot shows the median, the inter-quartile range, and the range of the L_1 distances between the estimators and the true regression function; dots mark outliers. Each added shape restriction improves the quality of the estimation, but increasing the degree of the polynomial pieces does not help.

Boxplots of the L_1 errors are shown in Fig. 6; the plots corresponding to the L_2 distances are similar.

It is immediate from Fig. 6 that the estimators benefit from imposing either shape constraint. Both the increasing and the concave cubic splines are significantly better estimates than the (otherwise unconstrained) nonnegative splines, and the concave increasing spline is significantly better than the concave only estimator (Mann–Whitney test p-values $< 7 \cdot 10^{-5}$ for each pair); the increasing concave spline is also better than the increasing only estimator, but the difference is not significant (p-value $= 0.15$). Quartic splines, on the other hand, did not yield better results than cubic ones in this example. Notice that the spline estimators outperform even the two-parametric model. The optimal increasing concave spline usually had 5 or 6 effective parameters. As expected, the one-parameter model proved to be hard to match.

6 Discussion

Two optimization-based approaches using conic sieves of shape constrained polynomial splines were discussed, and compared to each other and to unconstrained smoothing splines and kernel

methods in various applications of shape constrained estimation problems. We proposed a novel approach, based on SDP and SOCP, and revisited the popular nonnegative basis approach. The choice of basis in the nonnegative basis approach is critical. We presented a condition to test whether a given basis is appropriate, and proved that piecewise Bernstein polynomial splines satisfy this condition.

Theoretically, the SDP/SOCP approach, which requires solving optimization problems with semidefinite and second order conic constraints, is clearly superior to the basic nonnegative basis approach (which employs only linear constraints to represent polyhedral cones of splines), as it allows us to optimize precisely over the set that we want: the set of nonnegative splines. Using polyhedral sieves (following the nonnegative basis approach) results in simpler, linearly constrained models, but in most cases it only allows us to optimize over a strictly smaller cone than necessary.

In some instances the two approaches give very similar results, but we have also found examples where the SOCP approach is superior in practice, too. We exhibited cases when the two approaches are provably equivalent; these are problems with multiple shape constraints on low-degree splines.

The optimization models obtained from both approaches are easily solvable in a fraction of a second using readily available software. This was demonstrated using a few examples with simulated data, we refer to the thesis of the first author for more detailed examples. From the viewpoint of computational feasibility these approaches are competitive with kernel methods and even with closed-form formulae, which are only available for very restricted special cases. We found that the nonnegative spline estimators match, and frequently outperform state-of-the-art kernel estimators in density estimation problems.

We shall underline that the validity and algorithmic efficiency of the proposed method relied only on the fact that the set of nonnegative polynomial splines admits a characterization using only linear and semidefinite constraints, which made it easy to optimize all commonly used loss functions over them. This characterization of nonnegative polynomials is an immediate consequence of a classic result that nonnegative univariate polynomials can be written as sums of squares of polynomials. Hence, the method can be applied verbatim in other spaces of functions where nonnegative functions have a similar "sum of squares" characterization. An example is trigonometric polynomials: nonnegative trigonometric polynomials have a characterization analogous to nonnegative polynomials. Ramifications of this observation, along with many more examples,

including certain families of rational functions and exponential families can be found in (Papp and Alizadeh, 2011) and (Papp, 2011). The application of the same approach in multivariate estimation problems requires further study.

SUPPLEMENTAL MATERIALS

Proofs and some other technical details have been moved to the supplementary materials as Appendices — **see them in the separately submitted file appendices.pdf**.

Second order cone programming (SOCP) and semidefinite programming (SDP) A supplementary section with basic information on SOCP and SDP.

An SOCP characterization of cubic splines Theorem and proof.

Proof of Theorem 1 Proof.

References

Alizadeh, F., Eckstein, J., Noyan, N., and Rudolf, G. (2008). Arrival rate approximation by nonnegative cubic splines. *Operations Research*, 56:140–156.

Alizadeh, F. and Goldfarb, D. (2003). Second-order cone programming. *Mathematical Programming Series B*, 95:3–51.

Brunk, H. D. (1958). On the Estimation of Parameters Restricted by Inequalities. *Ann. Math. Statist.*, 26:607–454.

Burnham, K. P. and Anderson, D. (2004). Multimodel Inference – Understanding AIC and BIC in Model Selection. *Sociological Methods & Research*, 33(2):261–304.

Cormen, T. H., Leiserson, C. E., Rivest, R. L., and Stein, C. (2009). *Introduction to Algorithms*. MIT Press, Cambridge, MA.

CPLEX (2011). *IBM ILOG CPLEX Optimization Studio V12.3 Documentation*. IBM Corp.

d'Aspremont, A. (2004). *Shape Constrained Optimization, with Applications in Finance and Engineering*. PhD thesis, Stanford University.

de Montricher, G. F. and Tapia, R. A. and Thompson, J. R. (1975). Nonparametric Maximum Likelihood Estimation of Probability Densities by Penalty Function Methods. *The Annals of Statistics*, 3(6):1329–1348.

Delecroix, M., Simioni, M., and Thomas-Agnan, C. (1996). Functional Estimation Under Shape Constraints. *J. Nonparametr. Statist.*, 6(1):69–89.

Delecroix, M. and Thomas-Agnan, C. (2000). Spline and kernel regression under shape restrictions. Wiley.

Dierckx, P. H. (1980). An algorithm for for cubic spline fitting with convexity constraints. *Computing*, 24:349–371.

Dong, M. X. and Wets, R. J.-B. (2000). Estimating density functions: a constrained likelihood approach. *Nonparametric Statistics*, 12:549–595.

Eggermont, P. and LaRiccia, V. (2001). *Maximum Penalized Likelihood Estimation, Volume I: Density Estimation*. Springer-Verlag, New York.

Fushiki, T., Horiuchi, S., and Tsuchiya, T. (2006). A maximum likelihood approach to density estimation with semidefinite programming. *Neural Computation*, 18:2777–2812.

Geman, S. and Hwang, C.-R. (1982). Nonparametric maximum likelihood estimation by the method of sieves. *The Annals of Statistics*, 10(2):401–414.

Good, I. J. and Gaskins, R. A. (1971). Nonparametric Roughness Penalties for Probability Densities. *Biometrica*, 58:255–277.

Grant, M. and Boyd, S. (2007). *CVX: Matlab software for disciplined convex programming (web page and software)*. Stanford university. Available from `http://stanford.edu/~boyd/cvx`.

Grenander, U. (1981). *Abstract Inference*. John Wiley and Sons.

He, X. and Shi, P. (1996). Monotone B-spline Smoothing. *Journal of the American Statistical Association*, 93:643–650.

Hildreth, C. (1958). Point Estimates of Ordinates of concave Functions. *J. Amer. Statist. Assoc.*, 49:598–619.

Karlin, S. and Studden, W. J. (1966). *Tchebycheff Systems, with Applications in Analysis and Statistics*. Wiley Interscience Publishers.

Kimeldorf, G. and Wahba, G. (1971). Some results on Tchebycheffian spline functions. *Journal of Mathematical Analysis and Applications*, 33(1):82–95.

Mammen, E. and Thomas-Agnan, C. (1999). Smoothing Splines and Shape Restrictions. *Scandinavian Journal of Statistics*, 26(2):239–252.

Meyer, M. (1996). *Shape Restricted Inference with Applications to Nonparametric Regression,Smooth Nonparametric Function Estimation, and Density Estimation*. PhD thesis, University of Michigan, Ann Arbor.

Nemirovskii, A. S., Polyak, B. T., and Tsybakov, A. B. (1984). Signal processing by the nonparametric maximum likelihood method. *Probl. Peredachi Inf.*, 20(3):29–46.

Nemirovskii, A. S., Polyak, B. T., and Tsybakov, A. B. (1985). Convergence rate of nonparametric estimates of maximum-likelihood type. *Probl. Peredachi Inf.*, 21(4):17–33.

Nocedal, J. and Waltz, R. A. (2003). KNITRO user's manual. Technical Report OTC 2003/05, Optimization Technology Center, Northwestern University, Evanston, IL.

Papp, D. (2011). *Optimization models for shape-constrained function estimation problems involving nonnegative polynomials and their restrictions*. PhD thesis, Rutgers University, New Brunswick, NJ, USA.

Papp, D. and Alizadeh, F. (2011). Semidefinite characterization of sum-of-squares cones in algebras. Technical Report RRR 14–2011, Rutgers Center for Operations Research.

Ramsay, J. O. (1988). Monotone Regression Splines in Action. *Statistical Science*, 3(4):426–461.

Robertson, T., Wright, F., and Dykstra, R. (1988). *Order restricted statistical inference*. Wiley, Chichester.

Schumaker, L. L. (1981). *Spline Functions: Basic Theory*. John Wiley and Sons.

Scott, D. W. (1976). *Nonparametric probability density estimation by optimization theoretic techniques*. PhD thesis, Rice University, Houston, TX.

Scott, D. W., Tapia, R. A., and Thompson, J. R. (1980). Nonparametric probability density estimation by discrete maximum penalized-likelihood criteria. *Ann. Statist.*, 8(4):820–832.

Sturm, J. F. (2001). *Using SeDuMi 1.02, a Matlab Toolbox for Optimization over Symmetric Cones (Updated for Version 1.05)*. Available from `http://sedumi.ie.lehigh.edu/`.

Thompson, J. R. and Tapia, R. A. (1990). *Nonparametric Function Estimation, Modeling, and Simulation*. SIAM, Philadelphia, PA.

Turlach, B. A. (2005). Shape constrained smoothing using smoothing splines. *Computational Statistcs*, 20(1):81–103.

Villalobos, M. and Wahba, G. (1987). Inequality-constrained multivariate smoothing splines with applications to the estimation of posterior probabilities. *J. Amer. Statist. Assoc.*, 82(397):239–248.

Wahba, G. (1990). *Spline Models for Observational Data*. SIAM.

Wang, X. and Li, F. (2008). Isotonic Smoothing Spline Regression. *Journal of Computational and Graphical Statistics*, 17(1):21–37.

Wolkowicz, H., Saigal, R., and Vandenberghe, L., editors (2000). *Handbook of Semidefinite Programming: Theory, Algorithms, and Applications*, volume 27. Kluwer, Norwell, MA.

A Second order cone programming (SOCP) and semidefinite programming (SDP)

Semidefinite programming is a generalization of the well-known linear optimization (or linear programming) problem. A *semidefinite program* (or SDP for short) is the abstract problem of finding the optimum of a linear function subject to the constraint that an affine combination of matrices is positive semidefinite. In other words, it is an optimization problem of the form

$$\begin{aligned} &\underset{\mathbf{x}\in\mathbb{R}^n}{\text{minimize}} \quad \sum_i c_i x_i \\ &\text{subject to} \quad \mathbf{A}_0 + \sum_{i=1}^{n} \mathbf{A}_i x_i \succcurlyeq \mathbf{0}, \end{aligned}$$

where the vector $\mathbf{c} \in \mathbb{R}^n$ and the $m \times m$ symmetric matrices $\mathbf{A}_i$, $(i = 0, \ldots, n)$ are given; x_i denotes the ith component of the vector $\mathbf{x}$; these are the variables. To simplify presentation, we shall now assume that the sought minimum exists (the infimum is attained); this indeed holds for all the SDPs considered in this paper.

Similarly, a *second order cone program* (or SOCP for short) is an optimization problem

$$\begin{aligned} &\underset{\mathbf{x}\in\mathbb{R}^n}{\text{minimize}} \quad \sum_i c_i x_i \\ &\text{subject to} \quad \mathbf{a}_0 + \sum_{i=1}^{n} \mathbf{a}_i x_i \in \mathcal{Q}_{k+1}, \end{aligned}$$

with given vectors $\mathbf{c} \in \mathbb{R}^n$ and $\mathbf{a}_i$, $(i = 0, \ldots, n)$. Recall that $\mathcal{Q}_{k+1} = \left\{(z_0, \ldots, z_k) \colon z_0 \geq \left\|(z_1, \ldots z_k)^\top\right\|_2\right\}$ is the $(k+1)$-dimensional *second order cone*.

The constraints that appear in the above problems are called *semidefinite constraints* and *second order cone constraints* (the former are sometimes also referred to as (linear matrix inequalities).

It can be shown that SOCPs are special cases of SDPs, and in turn, linear programming is a special case of SOCP. This also implies that SOCP and SDP constraints can be mixed in the same optimization problem, the resulting problem is also an SDP.

The above formulations are considered "standard form" SDPs and SOCPs, but other, seemingly more general optimization problems that can be converted to the above form are also called SDPs and SOCPs. With this terminology, one of the key observations used in this paper can be

summarized as follows: The constraint that a piecewise polynomial spline with given knot points and degree is nonnegative over a given interval can be expressed as the constraints of an SDP, and if the spline degree is at most three, this SDP simplifies to an SOCP.

A.1 Remarks on software

Several efficient algorithms for the solution of second order cone programs and semidefinite programs are reviewed in (Alizadeh and Goldfarb, 2003) and (Wolkowicz et al., 2000); these references also provide a general introduction to these areas of optimization. The theoretical complexity of optimization problems involving only linear and second order cone constraints is the same as that of linear optimization (Alizadeh and Goldfarb, 2003). This means that the models presented in this paper are *polynomial time solvable*, which is the standard abstraction of the "efficient" solution of "tractable" problems (Cormen et al., 2009, sec. 34.1).

Linear constraints, second order cone constraints, and semidefinite constraints can be handled by some nonlinear optimization and modeling software, such as CVX (Grant and Boyd, 2007) and SeDuMi (Sturm, 2001) very effectively. However, currently none of these software can handle arbitrary convex, nonlinear objective function, which is necessary for some applications, such as maximum likelihood density estimation.

Second order cone constraints can also be handled by most state-of-the-art convex optimization software, which can solve problems with arbitrary convex objective functions. While carrying out the numerical experiments of this paper, we found the software KNITRO (Nocedal and Waltz, 2003), and CPLEX (CPLEX, 2011) especially effective and useful. The models of this paper can be solved in a fraction of a second without any numerical issues by the software mentioned above. Modeling languages (such as AMPL) also allow the user to input the mathematical optimization models of this paper, such as the representation of cubic splines shown in Theorem 3 in the Appendix, almost *verbatim* in the solvers. Only minimal effort is required from the user to obtain the estimators.

B An SOCP characterization of cubic splines

Theorem 3. *The coefficients $p_k^{(i)}$, $i = 0, \ldots, m-1$, $k = 0, \ldots, 3$ in (4) represent a nonnegative cubic spline over $[a_0, a_m]$ if and only if there exist real numbers $x_\ell^{(i)}, y_\ell^{(i)}$, $i = 0, \ldots, m-1$, $\ell = 0, 1, 2$ satisfying the following system of equations and inequalities for all $i = 0, \ldots, m-1$.*

$$p_0^{(i)} = y_0^{(i)} \tag{6a}$$

$$p_1^{(i)} = 2y_1^{(i)} + x_0^{(i)} - y_0^{(i)} \tag{6b}$$

$$p_2^{(i)} = y_2^{(i)} + 2x_1^{(i)} - 2y_1^{(i)} \tag{6c}$$

$$p_3^{(i)} = x_2^{(i)} - y_2^{(i)} \tag{6d}$$

$$(x_0^{(i)} + x_2^{(i)}, x_0^{(i)} - x_2^{(i)}, 2x_1^{(i)})^\top \in \mathcal{Q}_3 \tag{6e}$$

$$(y_0^{(i)} + y_2^{(i)}, y_0^{(i)} - y_2^{(i)}, 2y_1^{(i)})^\top \in \mathcal{Q}_3 \tag{6f}$$

$$p_0^{(i+1)} = \sum_{j=0}^{3} p_j^{(i)} \tag{6g}$$

$$\frac{1}{a_{i+2} - a_{i+1}} p_1^{(i+1)} = \sum_{j=1}^{3} \frac{j}{a_{i+1} - a_i} p_j^{(i)} \tag{6h}$$

$$\frac{2}{(a_{i+2} - a_{i+1})^2} p_2^{(i+1)} = \sum_{j=2}^{3} \frac{j(j-1)}{(a_{i+1} - a_i)^2} p_j^{(i)} \tag{6i}$$

Proof. Eqs. (6a)–(6f) come from Proposition 1, using Proposition 3 to translate the 2×2 semidefinite constraints to second order cone constraints. Eqs. (6g)–(6i) express the continuity of the derivatives up to order two. □

C Proof of Theorem 1

We start the proof by showing that for every polynomial p of degree n, strictly positive over $[0, 1]$, there exist nonnegative constants C_i such that $p + C_i \in \mathcal{P}(U, \mathbf{a}_i)$ for every i, and $\lim C_i = 0$.

Fix i, and consider two adjacent knot points a_k and a_{k+1} from the knot point sequence $\mathbf{a}_i$. The polynomial p can be represented as a piecewise polynomial spline of degree n with knot point sequence $\mathbf{a}_i$; its scaled representation (4) has $p^{(k)}(x) = p((a_{k+1} - a_k)x + a_k)$. Collecting terms in

the standard basis, we have

$$p^{(k)}(x) = p((a_{k+1} - a_k)x + a_k) = p(a_k) + \sum_{i=1}^{n} q_i^{(k)} x^i$$

with some $q_i^{(k)} = \mathcal{O}((a_{k+1} - a_k)^i)$, $i = 1, \ldots, n$. By assumption, $p(a_k) > 0$, because p is strictly positive on $[0, 1]$. All other coefficients $q_i^{(k)}$ are of order $\mathcal{O}(a_{k+1} - a_k)$. By the assumption on U, $\sum_{j=0}^{n} \alpha_j u_j \equiv p(a_k)$ for some positive $\alpha_0, \ldots, \alpha_n$. Now, if we express $p^{(k)}$ in the basis U: $p^{(k)} = \sum p_j^{(k)} u_j$, we have that $p_j^{(k)} = \alpha_j - \delta_j^{(k)}$ with $|\delta_j^{(k)}| = \mathcal{O}(a_{k+1} - a_k)$, consequently $p^{(k)} + p(a_k) \max_j(|\delta_j^{(k)}|/\alpha_j)$ has positive coefficients in the basis U. Applying the same argument for every k, we obtain that $p + C_i \in \mathcal{P}(U, \mathbf{a}_i)$ for

$$C_i = \max_{k: a_k \in \mathbf{a}_i} \big(p(a_k) \max_j(|\delta_j^{(k)}|/\alpha_j)\big).$$

Finally, as $|\delta_j^{(k)}| = \mathcal{O}(a_{k+1} - a_k)$ and p is bounded, $C_i \to 0$ as $\|\mathbf{a}_i\| \to 0$.

The same argument can be used prove that for every spline, positive over $[0, 1]$, with knot point sequence $\mathbf{a}$, and for every sequence $\{\mathbf{a}_i\}$ consisting of subdivisions of $\mathbf{a}$ satisfying $\lim \|\mathbf{a}_i\| = 0$, there exist nonnegative constants C_i such that $s + C_i \in \mathcal{P}(U, \mathbf{a}_i)$ for every i, and $\lim C_i = 0$.

Consequently, $\bigcup_i \mathcal{P}(U, \mathbf{a}_i)$ is a dense subset of nonnegative splines of degree n.

Finally, let us consider an arbitrary nonnegative function $f \in C([a, b])$. By the approximation theory of splines (see for example (Schumaker, 1981, Theorem 6.27)), for every n there exists a constant $M_n > 0$, depending only on n, but not on f, such that for every knot point sequence $\mathbf{a}$ there exists a (not sign-constrained) spline $s \in \mathcal{S}(n, \mathbf{a})$ satisfying

$$\|f - s\|_\infty \leq M_n \omega_n(f, \|\mathbf{a}\|), \tag{7}$$

where ω_n is the nth modulus of smoothness of f in $L^\infty([a, b])$, satisfying $\lim_{t \searrow 0} \omega_n(f, t) = 0$ for every $n \geq 1$ provided that f is continuous on $[a, b]$. Let ε denote the right-hand side of (7). Because f is nonnegative, the spline $s' = s + \varepsilon$ is nonnegative, and it satisfies

$$\|f - s'\|_\infty \leq 2M_n \omega_n(f, \|\mathbf{a}\|).$$

Hence, there are nonnegative splines $s_i \in \mathcal{P} \cap \mathcal{S}(n, \mathbf{a}_i)$ satisfying $\lim \|f - s_i\|_\infty = 0$.

We already saw that if $\lim \|\mathbf{a}_i\| = 0$, and $\{\mathbf{a}_i\}$ is asymptotically nested, then every nonnegative spline with knot point sequence $\mathbf{a}_i$ can be approximated with arbitrarily small positive error by some spline in $\mathcal{S}(U, \mathbf{a}_j)$ with a sufficiently high j. By the above argument the same holds for f.