Nonparametric Density Estimation via Exponential Epi-Splines: Fusion of Soft and Hard Information

Johannes O. Royset

Naval Postgraduate School, Monterey, Calif., USA

Roger J-B Wets

University of California, Davis, Calif., USA

April 29, 2013

Abstract

Relying on exponential epi-splines allows us to introduce a new methodology to estimate probability density functions. It allows us to marry hard information (observations) with soft information by which one usually means all non-data information one might have or suspect about the underlying stochastic phenomena. The article develops the theoretical foundations for this methodology: first the properties of epi-splines and exponential epi-splines and then goes on to obtain consistency and related asymptotics. Next, it provides a collection of examples of how soft information can be included in the formulation of the estimation problem and concludes with a number of experimental results that confirm, maybe better than the theoretical results, the potential of such an approach.

1 Introduction

From the very outset of statistical estimation theory, in accordance with the precepts laid down by Ronald Fisher in the 1920's, there has been a concern of obtaining 'best' estimates that would be based on *all* the information available to the statistician. With this goal in mind, but nonetheless taking into account the restrictions levied by computational limitations, many schemes have been suggested and promoted that invariably supply adequate solutions in specific instances but don't provide a comprehensive framework. In this article, we propose a methodology that overcomes the major obstacles of including all available information in the context of the nonparametric estimation of a density function. More specifically, we address the following problem: given a finite number of observations and possibly some information about the underlying random phenomenon, including implicitly that its distribution can be described by a density function, one has to find a 'best' estimate of this density function.

While density estimation has been studied extensively, see [25, 27, 40, 43, 34] for example, standard estimators, such as those based on kernels and smoothing splines, may perform poorly in applications with little data. One can hope to improve estimates in such situations by including additional information about a density such as its continuity, smoothness, unimodality, monotonicity, moments, and other characteristics. How to merge *hard information* (observations) and such additional information, to which we refer as *soft information*, has been, explicitly or implicitly, a primal motivation of developments in Statistics. To begin with restricting the search for a density function to a parametric class already sets forth, with no reservations, that the statistician has quite extensive soft information about the random event; computationally, it reduces the problem to finding a few parameters, usually, a quite manageable problem. Another well known approach has been to include "prior" information and this gave rise to Bayesian estimators [10], certainly a way to include soft information but still in a rather confined fashion.

Even in the realm of nonparametric estimation, attempts at including soft information in the formulation of the estimation problem are hardly new. Probably the most common approach, ignoring ad-hoc adjustments based on practical experience, is to rely on penalties and regularizations to obtain densities with desirable properties [13, 7, 20, 16, 37] and more recently [17, 18, 19]. While in principle many types of constraints in an estimation problem can be represented by penalty terms, the equivalence of such reformulations depends on the successful selection of multipliers and penalty parameters which is far from trivial in practice. In fact, poor selection of these multipliers and parameters may cause computational challenges due to ill-conditioning of the resulting optimization problem as well as significant deterioration of the quality of the resulting density estimate; see [8] for further discussion. There is also an extensive literature dealing with specific instances of soft information [42, 41, 40, 45, 9, 33, 12, 32, 14] and, in particular [28], [36] and [22, 21]. Recent studies of k-monotone densities include [2, 11, 3]. For examples of the estimation of shape restricted surfaces in other context we refer to [44, 26, 24] and the references therein.

Although we don't build on the theory and methods developed for either *M-estimators* or those that appeal to penalization to compel the estimates to satisfy certain additional requirement(s), we start from somewhat similar premises. Although optimization techniques are now used widely to solve specific statistical problems, a rather recent noteworthy instance is the Lasso procedure, cf. [4] for example, the connection, at the more fundamental level has only received sporadic attention, cf. [40, 45] for example and more recently [8, 6] and our work is in the same vein. We view a density estimation problem as one of finding a nonnegative function that sums to 1, in a function space to be determined. Whatever soft information we may have about this density is translated into restrictions (i.e., constraints) imposed on the choice of this function. Although the tenets of the approach would not be compromised if other standard criteria were selected, here we trust *maximum likelihood* to identify a best estimate and also analyze stability in terms of the Kullback-Leibler divergence. From a mathematical viewpoint our problem is thus a (specific) constrained infinite dimensional optimization problem[1]. Since closed form solutions to such problems are more than rare, one has to resort to finding a finite-dimensional approximating problem that is guaranteed to generate an *approximating solution* accompanied whenever possible with error bounds. This is the task devolved to *exponential epi-splines*, a composition of the exponential function

[1] In this framework, the 'true' density function becomes an optimal solution of a limiting problem obtained by letting the sample size tend to infinity. Variational analysis, in particular the approximation theory for optimization problems [29, Chapter 7] can be put upon, at this point to guide the analysis.

with an epi-spline: such functions are determined by a finite number of parameters and are dense, in exactly the desired approximating topology in an unusual rich class of probability density functions. The article provides the justification for such an approach. Although the approach and associated methodology is applicable in a variety of estimation contexts, and in particular joint probability functions, our focus in this article is on the estimation of the density for a single variable and in our numerical examples, we zoom in on the implications when only a small number of observations are available.

The paper proceeds in Section 2 by defining our density estimator and summarizing underlying approximation theory. Section 3 provides asymptotic and finite sample size analyses. Section 4 describes implementation of soft information, with numerical examples following in Section 5.

2 Exponential Epi-Spline Estimator

We search for density estimators within a class of functions that makes the estimation problem well defined and computational tractable. Foremost, we seek nonnegative functions and are therefore naturally led to a composition of an exponential function with a real-valued function. We would like the latter function to be defined by a finite number of parameters, such as in the case of piecewise polynomials, to enable finite-dimensional optimization over those parameters. Still, the class of functions should be sufficiently rich to capture, or at least approximate to an arbitrarily high accuracy, most densities encountered in practice. These factors lead us to the class of exponential epi-splines as defined in [30] and briefly described in this section for completeness.

Given a random variable X^0 with density h^0 and a sample[2] $X^1, X^2, ..., X^\nu$, we consider a density estimator h^ν of the form e^{-s^ν}, where $s^\nu : I\!\!R \to I\!\!R$ is an *epi-spline* determined by the maximum likelihood criterion[3] and constraints induced by soft information. Naturally, we refer to h^ν as an *exponential epi-spline estimator* of h^0. Under an independent sample distributed as X^0, if h^ν and h^0 are represented by the same family of continuous epi-splines, then, after possibly passing to a subsequence,

$$h^\nu \to h^0 \text{ uniformly with probabilty one,}$$

regardless of the type of soft information imposed as long as it doesn't exclude h^0. In this paper, we prove this result and, by means of the Kullback-Leibler divergence, expand it substantially to account for discontinuous densities, soft information that incorrectly eliminates h^0, and density that can only be approximated using epi-splines.

While utilizing a composition with the exponential function is analytically and computationally appealing as we see below, analogous developments with estimator $h^\nu = s^\nu$ directly, or with other compositions are also possible and could be advantageous for some criteria and applications. We focus the exposition on an exponential epi-spline estimator that vanishes outside a compact interval. Further complications arise when there is a need for estimating

[2] The sample may be independently generated from h^0, but much of the development holds without this assumption.

[3] Other options such as minimizing a least-square criterion are also possible and lead to developments along similar paths.

the tails of a density, due to the lack of sample points in the tails, and we defer that topic to another study.

2.1 Approximation Tools

We start by defining the central building block of our framework. A *basic epi-spline* is a function given in terms of an *order* $p \in I\!N_0 := \{0\} \cup I\!N$, where $I\!N := \{1, 2, ...\}$ and a *mesh*[4] $m = \{m_k\}_{k=0}^N$, with $m_{k-1} < m_k$, $k = 1, 2, ..., N$, that partitions its domain $[m_0, m_N]$. Extensions that deal with the whole real line and even higher dimensions are possible, but as already indicated not covered here. We refer to [30] for a thorough analysis of epi-splines and here simply review essential results.

Definition 1 (basic epi-spline and associated mesh). *A* (basic) epi-spline $s : [m_0, m_N] \subset I\!R \to I\!R$ *with* mesh $m = \{m_k\}_{k=0}^N$ *and* mesh-grade $|m| := \max_{1 \le k \le N}(m_k - m_{k-1})$ *is of* order $p \in I\!N_0$ *if*

(i) on each subinterval (m_{k-1}, m_k) *for* $k = 1, \ldots, N$, s *is polynomial of degree* p *and*

(ii) on m, s *is finite-valued.*

The family of all such epi-splines is denoted by e-spl$^p(m)$.

The set of *basic exponential epi-splines* corresponding to a family e-spl$^p(m)$ consists of finite positive functions defined on a compact interval that are especially convenient tools for density estimation; see again [30] for details.

Definition 2 (basic exponential epi-spline). *The family of* (basic) exponential epi-splines *of order* $p \in I\!N_0$ *with mesh* $m = \{m_k\}_{k=0}^N$, *denoted by* x-spl$^p(m)$, *consists of functions* $h : [m_0, m_N] \to I\!R$ *of the form* $h = e^{-s}$, *where* $s \in$ e-spl$^p(m)$.

Since this paper deals with basic epi-splines and exponential epi-splines exclusively, we systematically drop ‘basic’ from now on. It’s clear from the definition that every $s \in$ e-spl$^p(m)$, with mesh $m = \{m_k\}_{k=0}^N$, is uniquely defined by $(p + 2)N + 1$ parameters and thereby satisfies our requirement that the family of functions under consideration in the density estimation problem must be defined by a finite number of parameters. The family of epi-splines is also sufficiently rich to approximate a large class of functions with arbitrary accuracy. In fact, many common density functions are *exactly* represented on $[m_0, m_N]$. For example, a normal density is of the form e^{-s}, with $s \in$ e-spl$^2(m)$, on $[m_0, m_N]$ for any mesh. An exponential density is of the form e^{-s}, with $s \in$ e-spl$^1(m)$, on $[m_0, m_N]$ for any mesh with $m_0 = 0$. Even more densities, such as the lognormal and the Pareto, are exactly represented on a compact interval after a logarithmic transformation. As we make rigorous in the following paragraphs, exponential epi-splines also approximate on compact intervals essentially all densities we expect to encounter with arbitrarily high accuracy *including discontinuous densities.*

Approximations rely on the refinement of the mesh as made precise in the next definition.

[4] The mesh relates to *knots* for ‘classical’ splines, but we here prefer the term *mesh* as an epi-spline may not be continuous at these points. We also note that the mesh can be selected independently from a sample of observations.

Definition 3 (infinite refinement). *Given the interval $[l,u]$, one refers to a sequence of meshes $\{m^\nu\}_{\nu\in N}$, with $m^\nu = \big\{l = m_0^\nu, m_1^\nu, \dots, m_{N^\nu}^\nu = u\big\}$, as an* infinite refinement *if their mesh-grade $|m^\nu| \to 0$.*

Of course, all 'natural' meshes satisfy this property.

Using a suitable metric on the space of functions under considerations as describe below, we find that for any $p \in I\!N_0$ and $\{m^\nu\}_{\nu\in N}$, an infinite refinement of $[l,u]$,

> the continuous x-spl$^p(m^\nu), \nu \in I\!N,$ are dense in the set of functions of the form $f = e^{-s},$ with $s : [l,u] \to I\!R$ continuous.

We plan to go beyond continuous and bounded densities and therefore need an extension of this approximation result to a broader class of semicontinuous functions. We rely on the *epi-topology* and *hypo-topology* (sometimes called the Attouch-Wets topologies), which are reviewed here for completeness; see [29, Section 7.I] for details. For any $l < u \in I\!R$, we denote by lsc-fcns$\big([l,u]\big)$ the set of all lower semicontinuous (lsc) functions $f : [l,u] \to \overline{I\!R} := I\!R \cup \{-\infty, \infty\}$ excluding $f \equiv \infty$, i.e., with empty (effective) domain. For any two functions, f and g, in this space, the *epi-distance* $d\!l$, is defined by

$$d\!l(f,g) := \int_0^\infty d\!l_\rho(f,g) e^{-\rho} d\rho,$$

where

$$d\!l_\rho(f,g) := \max_{\|x\|\le\rho} |d(x, \text{epi } f) - d(x, \text{epi } g)| \quad \text{and} \quad d(x,S) := \inf_{y\in S} \|x - y\| \text{ for } S \subset I\!R^2,$$

with epi $f := \{(x,\beta) \in I\!R^2 \mid f(x) \le \beta\}$ being the *epigraph* of f and similarly for epi g; see Figure 1 for an illustration. When the metric is defined in terms of the epi-distance, it generates the *epi-topology* on lsc-fcns$([l,u])$: (lsc-fcns$([l,u]), d\!l$) is a Polish (complete separable metric) space [29, Theorem 7.58], [1, §5]. A sequence of functions f^ν in lsc-fcns$\big([l,u]\big)$ *epi-converge* to f if their epigraphs set-converge, i.e., in the sense of taking Painlevé-Kuratowski limits [29, §7.B], which by [29, Theorem 7.58] takes place if and only if $d\!l(f^\nu, f) \to 0$.

When dealing with upper semicontinuous (usc) functions, usc-fcns$\big([l,u]\big)$, now excluding the function $\equiv -\infty$, after observing that hypograph of a function f, hypo $f = \big\{(x,\beta) \,\big|\, f(x) \ge \beta\big\}$ is just a mirror image of the epigraph of $-f$, one can mimic the definitions and constructions described for lsc functions to set up the *hypo-distance* $d\!l_{\text{hypo}}(f,g) := d\!l(-f,-g)$, between any two functions f and g and generate the *hypo-topology* which again makes (usc-fcns$([l,u]), d\!l_{\text{hypo}}$) a Polish space. A sequence of functions f^ν *hypo-converge* to f if $-f^\nu$ epi-converge to $-f$. The relationship between epi- and hypo-convergence and other modes are convergence in the present context is examined below; see also [29, Chapters 4 & 7] for a broader treatment.

Since the supremum of an usc function on a compact set is attained, the consideration of usc densities naturally arises in applications where the subsequent use of the densities involve maximization, such as for the purpose of finding their modes. Similarly, lsc densities is the natural class to consider in the context of subsequent minimization. We next state the main approximation results for exponential epi-splines.

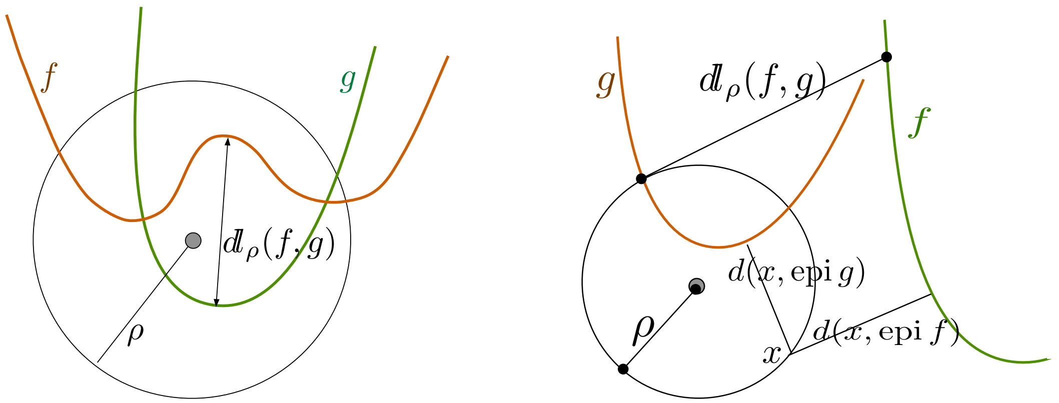


Figure 1: Examples of $d\!l_\rho(f,g)$ for epi f and epi g with different overlaps

Theorem 1 (lsc and usc dense approximations [30]). *For any $p \in \mathbb{N}_0$ and $\{m^\nu\}_{\nu \in N}$, an infinite refinement of $[l,u]$, under the hypo-topology,*

$$\left(\bigcup_{\nu \in N} \text{x-spl}^p(m^\nu)\right) \bigcap \text{usc-fcns}([l,u]) \textit{ is dense in } \{e^{-s} \mid s \in \text{lsc-fcns}([l,u])\}$$

and under the epi-topology,

$$\left(\bigcup_{\nu \in N} \text{x-spl}^p(m^\nu)\right) \bigcap \text{lsc-fcns}([l,u]) \textit{ is dense in } \{e^{-s} \mid s \in \text{usc-fcns}([l,u])\}.$$

We now turn to a convenient representation of exponential epi-splines, which plays an essential role in computations and analysis. Every $s \in \text{e-spl}^p(m)$, with $m = \{m_k\}_{k=0}^N$, is uniquely represented by an *epi-spline parameter*

$$r = (s_0, s_1, ..., s_N, a_1, a_2, ..., a_N), \quad s_k \in \mathbb{R},\ k = 0,1,...,N, \quad a_k \in \mathbb{R}^{p+1},\ k = 1,2,...,N,$$

such that for any $x \in [m_0, m_N]$,

$$s(x) = \langle c_{p,m}(x), r\rangle,$$

where the *basis function* $c_{p,m} : [m_0, m_N] \to \mathbb{R}^{(p+2)N+1}$ is defined by

$$c_{p,m}(x) := \begin{cases} (0_{N+1+(p+1)(k-1)}, 1, (x - m_{k-1}), (x-m_{k-1})^2, ..., (x-m_{k-1})^p, 0_{(p+1)(N-k)}) \\ \qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad \text{if } x \in (m_{k-1}, m_k),\ k = 1,2,...,N \\ (0_k, 1, 0_{N-k+(p+1)N}), \qquad\qquad\qquad\qquad\quad \text{if } x = m_k,\ k = 0,1,...,N, \end{cases}$$

with 0_k denoting the k-dimensional zero vector, $k \in \mathbb{N}$, and 0_0 being a term that is omitted. This representation of an epi-spline s lets the first $N+1$ components in the vector r be the values of s on m. The remaining $(p+1)N$ components are divided into N blocks of $(p+1)$-tuples, each of which gives the coefficients of the polynomial defining s on intervals of the form (m_{k-1}, m_k). Specifically, $a_k = (a_{k,0}, a_{k,1}, ..., a_{k,p})$ is such that

$$s(x) = \sum_{i=0}^{p} a_{k,i}(x - m_{k-1})^i, \text{ for } x \in (m_{k-1}, m_k),\ k = 1,2,...,N.$$

Since the first $N+1$ components of r determine the value of an epi-spline only on m, which consists of a finite number of points, we refer to the remaining $(p+1)N$ components of r as the *essential epi-spline parameter* and write $r = (r_{\text{mesh}}, r_{\text{ess}})$, with $r_{\text{mesh}} \in I\!R^{N+1}$ and $r_{\text{ess}} \in I\!R^{(p+1)N}$, to indicate this partition of r. Correspondingly, we let $c_{p,m} = (c_{\text{mesh}}, c_{\text{ess}})$.

It's clear that classical splines in their various forms are closely related to epi-splines. While classical splines are typically defined to posses a certain degree of smoothness and satisfy boundary conditions, epi-splines are more flexible; see [30] for a detailed comparison. As guided by soft information in a given application, we impose continuity, smoothness, and other condition as constraints in a variational formulation given in Subsection 2.2. This provides flexibility and facilitates the estimation of a wide range of densities under essentially any soft information as illustrated in Sections 4 and 5.

Since the value of a density at a finite number of points is immaterial for the characterization of the corresponding probability distribution, it may at first appear unnecessary to specify the value of an exponential epi-spline $e^{-\langle c_{p,m}(\cdot), r\rangle}$ on m. Instead of determining $r = (r_{\text{mesh}}, r_{\text{ess}})$, one could simply focus on r_{ess} and this is certainly the case for continuous exponential epi-splines. However, we would like to conveniently handle soft information about a discontinuous density on m, such as its value at a point in m, as well as the possibility of sample points taking values in m. The latter may occur by construction or when attempting to estimate a 'density' for a distribution that turns out to have atoms. Hence, a need arises for also considering r_{mesh} and we proceed with the more general framework.

The next result gives connections between various modes of convergence within the class of exponential epi-splines. As we see, convergence in the epi-spline parameter is equivalent to uniform convergence of the corresponding exponential epi-splines and, under a restriction to usc functions, also to convergence in the hypo-distance.

Theorem 2 (equivalent convergence [30]). *Suppose that $h^\nu, h^0 \in$ x-spl$^p(m)$, with $m = \{m_k\}_{k=0}^N$, $h^\nu = e^{-s^\nu} = e^{-\langle c_{p,m}(\cdot), r^\nu\rangle}$, and $h^0 = e^{-s^0} = e^{-\langle c_{p,m}(\cdot), r^0\rangle}$. Then, the following hold:*

$$r^\nu \to r^0 \iff h^\nu \to h^0 \text{ uniformly on } [m_0, m_N] \implies d\!l(-h^\nu, -h^0) \to 0 \iff d\!l(s^\nu, s^0) \to 0.$$

Moreover, if h^ν, h^0 are usc, then also

$$h^\nu \to h^0 \text{ uniformly on } [m_0, m_N] \Longleftarrow d\!l(-h^\nu, -h^0) \to 0.$$

We observe that since the hypo-distance doesn't distinguish between a function and its usc regularization (see Proposition 7.4 in [29]), uniform convergence can't generally be implied from hypo-convergence, even for exponential epi-splines.

In view of the preceding results, we find that exponential epi-splines are flexible approximation tools and proceed by letting them be the corner stone of a maximum likelihood estimator.

2.2 Maximum Likelihood Estimator

For p and $m = \{m_k\}_{k=0}^N$ given, we proceed by adopting a maximum likelihood criterion to determine an exponential epi-spline estimator $h^\nu = e^{-s^\nu} \in$ x-spl$^p(m)$ of a density h^0, with, of course, $\int_{m_0}^{m_N} h^\nu(x)dx = 1$. Let $X^1, X^2, \ldots, X^\nu$ be a sample, with $m_0 \le X^i \le m_N, i = 1, 2, ..., \nu$, almost surely. A realization of the sample is denoted by lower case. The epi-spline

s^ν is an optimal solution of a maximum likelihood problem or, as stated here equivalently, of the minimum negative log-likelihood problem

$$P^\nu_{p,m}: \quad s^\nu \in \operatorname*{argmin}_{s \in S^\nu} \frac{1}{\nu} \sum_{i=1}^{\nu} s(X^i) \ \text{ s.t. } \ \int_{m_0}^{m_N} e^{-s(x)} dx = 1,$$

with $S^\nu \subset \text{e-spl}^p(m)$ being a constraint set that accounts for soft information as elaborated in Sections 4 and 5, and 'argmin' denoting the set of optimal solutions. The set S^ν could be random, with realizations being subsets of $\text{e-spl}^p(m)$. However, both the random set and the realizations are denoted by S^ν as the meaning should be clear from the context. The ability to include almost every conceivable constraint in the formulation of S^ν provides significant flexibility for the statistician. The restriction to a sample taking values in $[m_0, m_N]$ eliminates pathological cases where the log-likelihood function is not defined (or if we let an epi-spline be identical to ∞ outside $[m_0, m_N]$, where the likelihood function is zero) regardless of the choice of exponential epi-spline in $\text{x-spl}^p(m)$.

We next deal with the issues of existence and uniqueness of the estimator and consider a computational convenient equivalent form of $P^\nu_{p,m}$ using the representation $s = \langle c_{p,m}(\cdot), r\rangle$. We denote by $R^\nu \subset I\!R^{(p+2)N+1}$ the set of epi-spline parameters corresponding to the set of epi-splines S^ν, i.e.,

$$R^\nu := \{r \in I\!R^{(p+2)N+1} \mid \langle c_{p,m}(\cdot), r\rangle \in S^\nu\}.$$

For example, if $S^\nu = \text{e-spl}^p(m)$, then $R^\nu = I\!R^{(p+2)N+1}$. When incorporating soft information, R^ν and S^ν become more restrictive as we see in Section 4. Again, we let both the random set and its realizations be denoted by R^ν. We also let

$$R^\nu_I := \left\{ r \in R^\nu \;\middle|\; \int_{m_0}^{m_N} e^{-\langle c_{p,m}(x), r\rangle} dx = 1 \right\}.$$

As stated next, $P^\nu_{p,m}$ is equivalent to the finite-dimensional problem

$$\bar{P}^\nu_{p,m}: \quad \min_{r \in R^\nu_I} \frac{1}{\nu} \sum_{i=1}^{\nu} \langle c_{p,m}(X^i), r\rangle.$$

A realization of $X^1, ..., X^\nu$ and S^ν generates a realization of $P^\nu_{p,m}$ and one of $\bar{P}^\nu_{p,m}$, which we refer to as being in *correspondence.*

Theorem 3 (computing estimate). *For $m = \{m_k\}_{k=0}^N$, the following holds for every corresponding realizations of $P^\nu_{p,m}$ and $\bar{P}^\nu_{p,m}$:*

(i) If $s^\nu \in \text{e-spl}^p(m)$ is optimal for $P^\nu_{p,m}$, then there exists an $r^\nu \in I\!R^{(p+2)N+1}$ optimal for $\bar{P}^\nu_{p,m}$ with $s^\nu = \langle c_{p,m}(\cdot), r^\nu\rangle$.

(ii) If $r^\nu \in I\!R^{(p+2)N+1}$ is optimal for $\bar{P}^\nu_{p,m}$, then $s^\nu = \langle c_{p,m}(\cdot), r^\nu\rangle$ is optimal for $P^\nu_{p,m}$ and the exponential epi-spline estimator

$$h^\nu(x) = \begin{cases} e^{-\langle c_{p,m}(x), r^\nu\rangle}, & x \in [m_0, m_N] \\ 0, & \textit{otherwise.} \end{cases}$$

(iii) If R_I^ν is nonempty and R^ν is compact, then $\bar{P}_{p,m}^\nu$ has an optimal solution.

Proof: The equivalence of $\bar{P}_{p,m}^\nu$ and $P_{p,m}^\nu$ follows directly from the representation $s = \langle c_{p,m}(\cdot), r\rangle$. The existence of an optimal solution of $\bar{P}_{p,m}^\nu$ follows trivially from the continuity of the involved functions and the compactness of R^ν. □

While the objective function in $\bar{P}_{p,m}^\nu$ is linear, R_I^ν may be nonconvex. Hence, $\bar{P}_{p,m}^\nu$ could possess local minimizers that are not globally optimal, increasing the complexity of solving the problem numerically. We see in Section 4 that R^ν is often a polyhedron, or at least convex. Hence, the main difficulty in $\bar{P}_{p,m}^\nu$ is associated with the integral constraint. However, under broad conditions stated next, that constraint can be relaxed.

Definition 4 *A realization of $\bar{P}_{p,m}^\nu$ is said to be* loosely constrained *if for every $r \in R^\nu$ with $\int_{m_0}^{m_N} e^{-\langle c_{p,m}(x),r\rangle}dx < 1$, there exists $r' \in R_I^\nu$ with $\sum_{i=1}^\nu \langle c_{p,m}(x^i), r' - r\rangle < 0$.*

The following Proposition 2 and Section 4 give examples of loosely constrained realizations. We give an immediate consequence next.

Proposition 1 *Suppose that a realization of $\bar{P}_{p,m}^\nu$ is loosely constrained. Then, that realization and the corresponding realization of the relaxed problem*

$$RP_{p,m}^\nu : \quad \min_{r\in R^\nu} \frac{1}{\nu}\sum_{i=1}^{\nu}\langle c_{p,m}(X^i), r\rangle \;\; \text{s.t.} \;\; \int_{m_0}^{m_N} e^{-\langle c_{p,m}(x),r\rangle}dx \le 1$$

have identical sets of optimal solutions. Moreover, if R^ν is convex, then $RP_{p,m}^\nu$ is a convex problem.

In Theorem 7 below we show that even beyond loosely constrained realizations, the consideration of $RP_{p,m}^\nu$ is justified. In view of the preceding discussion and results, it's clear that the exponential epi-spline estimator is computationally tractable by means of well-developed convex optimization algorithms in many practical situations and by means of nonlinear programming algorithms in even more situations. In some cases, for example when R^ν is polyhedral, some further computational benefits may arise from utilizing the following reformulation, which is valid under additional assumptions; see Section 4 for examples. The next result also gives a sufficient condition for a realization of $\bar{P}_{p,m}^\nu$ to be loosely constrained. We use the notation $1_{p,N}$ to indicate the $((p+2)N+1)$-dimensional vector consisting of zeros, except at entries 1 through $N+1$ as well as entries $N+2+(k-1)(p+1)$, $k = 1,2,...,N$, where it is unity.

Proposition 2 *A realization of $\bar{P}_{p,m}^\nu$ for which every $r \in R^\nu$ and $\beta \in I\!R$ satisfy $r + \beta 1_{p,N} \in R^\nu$, is loosely constrained and its set of optimal solutions is identical to that of the corresponding realization of the penalized problem*

$$PP_{p,m}^\nu : \quad \min_{r\in R^\nu} \frac{1}{\nu}\sum_{i=1}^{\nu}\langle c_{p,m}(X^i), r\rangle + \int_{m_0}^{m_N} e^{-\langle c_{p,m}(x),r\rangle}dx.$$

Moreover, if R^ν is convex, then $PP_{p,m}^\nu$ is a convex problem.

Proof: Consider corresponding realizations of $\bar{P}^\nu_{p,m}$ and $PP^\nu_{p,m}$ and let $r \in R^\nu$ satisfy $\int_{m_0}^{m_N} e^{-\langle c_{p,m}(x),r\rangle}dx = \gamma < 1$. For $r' = r + (\log\gamma)1_{p,N}$,

$$\int_{m_0}^{m_N} e^{-\langle c_{p,m}(x),r'\rangle}dx = \int_{m_0}^{m_N} e^{-\langle c_{p,m}(x),r\rangle - \log\gamma}dx = \frac{1}{\gamma}\int_{m_0}^{m_N} e^{-\langle c_{p,m}(x),r\rangle}dx = 1. \qquad (1)$$

Moreover,

$$\sum_{i=1}^{\nu}\langle c_{p,m}(x^i), r' - r\rangle = \sum_{i=1}^{\nu}\langle c_{p,m}(x^i), (\log\gamma)1_{p,N}\rangle = \nu\log\gamma < 0.$$

Since $r' \in R^\nu$ by assumption, the realization of $\bar{P}^\nu_{p,m}$ is loosely constrained by Definition 4.

We next consider the penalized problem. For any $r \in R^\nu$, let

$$f^\nu(r) = \frac{1}{\nu}\sum_{i=1}^{\nu}\langle c_{p,m}(x^i), r\rangle + \int_{m_0}^{m_N} e^{-\langle c_{p,m}(x),r\rangle}dx$$

and let $\hat{r} \in R^\nu$ be arbitrary. Since every epi-spline is piecewise polynomial and therefore integrates on $[m_0, m_N]$ to a finite number, there exists a $\gamma \in (0,\infty)$ such that $\int_{m_0}^{m_N} e^{-\langle c_{p,m}(x),\hat{r}\rangle}dx = \gamma$. By assumption, $\hat{r} + (\log\gamma)1_{p,N} \in R^\nu$ and, following the same argument as in (1),

$$\int_{m_0}^{m_N} e^{-\langle c_{p,m}(x),\hat{r}+(\log\gamma)1_{p,N}\rangle}dx = 1.$$

Consequently, $\hat{r} + (\log\gamma)1_{p,N}$ is feasible in the realization of $\bar{P}^\nu_{p,m}$. Suppose that r^ν is optimal for the realization of $\bar{P}^\nu_{p,m}$. It follows that r^ν also minimizes f^ν on R^ν_I because this problem deviates from the realization only with the constant one in the objective function. Using an argument similar to that of Lemma 2.3 in [14], we find that

$$\begin{aligned}
& f^\nu(\hat{r}) - f^\nu(r^\nu) \\
= \; & \frac{1}{\nu}\sum_{i=1}^{\nu}\langle c_{p,m}(x^i), \hat{r} + (\log\gamma)1_{p,N}\rangle - \log\gamma + 1 - 1 + \int_{m_0}^{m_N} e^{-\langle c_{p,m}(x),\hat{r}\rangle}dx - f^\nu(r^\nu) \\
= \; & f^\nu(\hat{r} + (\log\gamma)1_{p,N}) - \log\gamma - 1 + \gamma - f^\nu(r^\nu) \\
\geq \; & -\log\gamma - 1 + \gamma,
\end{aligned}$$

where the inequality follows from the fact that r^ν is optimal and $\hat{r} + (\log\gamma)1_{p,N}$ is feasible in the realization of $\bar{P}^\nu_{p,m}$. Since $-\log\gamma - 1 + \gamma > 0$ for $\gamma \in (0,\infty), \gamma \neq 1$, we find that every $r \in R^\nu$ with $\int_{m_0}^{m_N} e^{-\langle c_{p,m}(x),r\rangle}dx \neq 1$ has $f^\nu(r) > f^\nu(r^\nu)$ and consequently can't minimize f^ν on R^ν. The first conclusion then follows. Convexity of $PP^\nu_{p,m}$ follows directly from the convexity of the integral term. □

In general, one can't expect a unique optimal solution of a realization of $\bar{P}^\nu_{p,m}$, and consequently a unique exponential epi-spline estimate, due to the flexibility in the choice of values of the epi-spline on a mesh that isn't a subset of the sample realization $x^1, x^2, ..., x^\nu$. In fact, if the first $N+1$ components of the epi-spline parameter r are not constrained by R^ν, then there is an infinite number of optimal solutions whenever one exists. The next result shows that when these values are uniquely determined by the essential epi-spline parameter, uniqueness may still be achieved. Such a dependence on the essential epi-spline parameter is present, for example, in the case of continuous epi-splines.

Proposition 3 *Suppose that corresponding realizations of $\bar{P}^\nu_{p,m}$ and $RP^\nu_{p,m}$ have R^ν convex, $\{x^1,...,x^\nu\} \cap m = \emptyset$, and satisfy the condition:*

$$(r_{\text{mesh}}, r_{\text{ess}}), (r'_{\text{mesh}}, r'_{\text{ess}}) \in R^\nu, \text{ with } r_{\text{ess}} = r'_{\text{ess}}, \text{ implies } r_{\text{mesh}} = r'_{\text{mesh}}.$$

Then, the following hold:

(i) If an optimal solution r of the realization of $RP^\nu_{p,m}$ is in R^ν_I, then there are no other optimal solutions.

(ii) The realization of $PP^\nu_{p,m}$ has at most one optimal solution.

Proof: We start by showing strictly convexity of the integral term as a function of the essential epi-spline parameters. Given $m = \{m_k\}_{k=0}^N$, we define $\psi : I\!R^{(p+1)N} \to I\!R$ and $\phi : [m_0, m_N] \times I\!R^{(p+1)N} \to I\!R$ by

$$\psi(r_{\text{ess}}) = \int_{m_0}^{m_N} \phi(x, r_{\text{ess}})dx, \text{ with } \phi(x, r_{\text{ess}}) = e^{-\langle c_{\text{ess}}(x), r_{\text{ess}}\rangle}.$$

For all $x \in [m_0, m_N]$ and $r_{\text{ess}}, r'_{\text{ess}} \in I\!R^{(p+1)N}$, twice differentiation with respect to the second argument in ϕ gives that

$$\langle r'_{\text{ess}}, \nabla^2\phi(x, r_{\text{ess}})r'_{\text{ess}}\rangle = \langle c_{\text{ess}}(x), r'_{\text{ess}}\rangle^2 e^{-\langle c_{\text{ess}}(x), r_{\text{ess}}\rangle} \geq 0.$$

Suppose that $r'_{\text{ess}} \neq 0$. Then, there exists a $\hat{k} \in \{1, 2, ..., N\}$ such that $\langle c_{\text{ess}}(x), r'_{\text{ess}}\rangle$ is a polynomial in x for $x \in (m_{\hat{k}-1}, m_{\hat{k}})$ with not all coefficients zero. Hence, there exists a subset of $(m_{\hat{k}-1}, m_{\hat{k}})$ with positive Lebesgue measure on which $\langle c_{\text{ess}}(x), r'_{\text{ess}}\rangle \neq 0$ and

$$\int_{m_0}^{m_N} \langle c_{\text{ess}}(x), r'_{\text{ess}}\rangle^2 e^{-\langle c_{\text{ess}}(x), r_{\text{ess}}\rangle} dx > 0. \tag{2}$$

Since the dominated convergence theorem implies that the left-hand side of (2) equals $\langle r'_{\text{ess}}, \nabla^2\psi(r_{\text{ess}})r'_{\text{ess}}\rangle$, we find that ψ is strictly convex by the second-order condition for convexity.

We let $\tilde{\psi} = (1/\nu)\sum_{i=1}^\nu \langle c_{\text{ess}}(x^i), \cdot\rangle + \psi(\cdot)$, which is therefore also strictly convex.

We first consider (ii). Suppose for the sake of a contradiction that there exist $r = (r_{\text{mesh}}, r_{\text{ess}}) \neq r' = (r'_{\text{mesh}}, r'_{\text{ess}})$ that both are optimal for the realization of $PP^\nu_{p,m}$, with optimal value v^*. Since $\{x^1, ..., x^\nu\} \cap m = \emptyset$, the objective function in this problem depends only on the essential epi-spline parameter and, in fact, $\tilde{\psi}(r_{\text{ess}}) = \tilde{\psi}(r'_{\text{ess}}) = v^*$. We consider two cases.

a) Suppose that $r_{\text{ess}} = r'_{\text{ess}}$, but then $r_{\text{mesh}} = r'_{\text{mesh}}$ by assumption and we contradict the hypothesis that $r \neq r'$.

b) Suppose that $r_{\text{ess}} \neq r'_{\text{ess}}$. Since $\tilde{\psi}$ is strictly convex, there exists a unique minimizer r''_{ess} of $\tilde{\psi}$ over the convex hull of r_{ess} and r'_{ess}. Moreover, there exists an $\alpha \in (0, 1)$ such that $r''_{\text{ess}} = \alpha r_{\text{ess}} + (1 - \alpha)r'_{\text{ess}}$ and $\tilde{\psi}(r''_{\text{ess}}) < v^*$. By the convexity of R^ν, $r'' = (\alpha r_{\text{mesh}} + (1 - \alpha)r'_{\text{mesh}}, r''_{\text{ess}}) \in R^\nu$ and its objective function value in $PP^\nu_{p,m}$ is $\tilde{\psi}(r''_{\text{ess}}) < v^*$, which contradicts the optimality of v^*.

Second, we focus on (i). Suppose that $r = (r_{\text{mesh}}, r_{\text{ess}}) \in R^\nu_I$ is optimal for the realization of $RP^\nu_{p,m}$. We consider two cases.

a) Suppose that $\int_{m_0}^{m_N} e^{-\langle c_{p,m}(x), r'\rangle} dx \geq 1$ for all $r' \in R^\nu$. Then by strict convexity of ψ, there exists a unique minimizer r''_{ess} of ψ on $\{r'''_{\text{ess}} \in I\!R^{(p+1)N} \mid (r'''_{\text{mesh}}, r'''_{\text{ess}}) \in R^\nu$ for some $r'''_{\text{mesh}} \in I\!R^{N+1}\}$. However, $r''_{\text{ess}} = r_{\text{ess}}$ because $\psi(r_{\text{ess}}) = 1$. Another optimal solution for the realization of $RP^\nu_{p,m}$ would thus have essential epi-spline parameter identical to r_{ess}. However, by assumption, such a solution would then also be identical to r in the remaining components, which implies it coincides with r.

b) Suppose that there exists $r' \in R^\nu$ such that $\int_{m_0}^{m_N} e^{-\langle c_{p,m}(x), r'\rangle} dx < 1$. Then, the Slater constraint qualification is satisfied and there exists a multiplier $\lambda \geq 0$ such that the realization of $RP^\nu_{p,m}$ has the same set of optimal solutions as the problem

$$\min_{r \in R^\nu} \frac{1}{\nu} \sum_{i=1}^{\nu} \langle c_{p,m}(x^i), r\rangle + \lambda \int_{m_0}^{m_N} e^{-\langle c_{p,m}(x), r\rangle} dx. \tag{3}$$

Repeating the arguments that lead to (ii), with (3) in place of the realization of $PP^\nu_{p,m}$, shows that the there are no other optimal solutions of the realization of $RP^\nu_{p,m}$ than r. □

We end this section by observing that $\bar{P}^\nu_{p,m}$, $RP^\nu_{p,m}$, and $PP^\nu_{p,m}$ involve one-dimensional integrals, which, in practice, must be evaluated numerically. However, this fact introduces no significant difficulty as numerical integration is easily carried out with high accuracy in short computing time due to the smoothness of the integrand in each segment (m_{k-1}, m_k). Hence, assuming that R^ν is defined in terms of a finite number of smooth inequality and equality constraints, all these problems are tractable by standard nonlinear programming solvers and, in the case of convexity, powerful convex solvers.

3 Consistency, Asymptotics, and Error Bounds

We achieve consistency, asymptotics, and other results by viewing $\{P^\nu_{p,m}\}_{\nu=1}^\infty$, for given m and p, as a sequence of optimization problems that under quite general assumptions converges in some sense to a limiting optimization problem, whose optimal solution recovers a *true* density $h^0 \in$ x-spl$^p(m)$ of a random variable X^0, as the sample size $\nu \to \infty$. We note that the restriction to x-spl$^p(m)$ for given m and p is justified by Theorem 1, but we also discuss the consideration of densities beyond this broad class; see Theorem 5 below. Before defining the limiting problem, we recall the Kullback-Leibler divergence, which is closely related to the likelihood function.

Let $d_{KL}(h||g)$ denote the Kullback-Leibler divergence from a density h to a density g defined on $I\!R$, i.e.,

$$d_{KL}(h||g) := \int_{-\infty}^{\infty} h(x) \log \frac{h(x)}{g(x)} dx.$$

Here and below we make the standard interpretation that $\beta_1 \log \beta_1/\beta_2 = 0$ when $\beta_1 = 0$ regardless of the value of $\beta_2 \in I\!R$ and $\beta_1 \log \beta_1/\beta_2 = \infty$ when $\beta_1 > 0$ and $\beta_2 = 0$.

We define the 'approximation' of a density h by an exponential epi-spline as follows.

Definition 5 (Kullback-Leibler projection). *For any density h on $I\!R$ and family* e-spl$^p(m)$, $m = \{m_k\}_{k=0}^N$, *the* Kullback-Leibler projection *of h on* e-spl$^p(m)$ *is the set*

$$\mathcal{P}_{p,m}(h) := \operatorname*{argmin}_{s \in \text{e-spl}^p(m)} d_{KL}(h||e^{-s}) \text{ s.t. } \int_{m_0}^{m_N} e^{-s(x)} dx = 1. \tag{4}$$

If the minimization is further constrained by $s \in S \subset$ e-spl$^p(m)$, then we denote the set of optimal solutions by $\mathcal{P}^S_{p,m}(h)$ and refer to it as the Kullback-Leibler projection relative to S.

We see that $\mathcal{P}_{p,m}(h)$ is the set of epi-splines that gives the 'closest' exponential epi-spline densities to h in the sense of the Kullback-Leibler divergence. It is well known that $d_{KL}(h||g) \geq 0$ for all densities h and g, and that $d_{KL}(h||g) = 0$ if and only if $h = g$, except possibly on a set of Lebesgue measure zero. Hence, if a density $h = e^{-s} \in$ x-spl$^p(m)$, $m = \{m_k\}_{k=0}^N$, then $s \in \mathcal{P}_{p,m}(h)$ and all $\tilde{s} \in \mathcal{P}_{p,m}(h)$ are identical to s (Lebesgue) almost everywhere on $[m_0, m_N]$. Since s and $\tilde{s}$ are polynomials of order p on each open interval (m_{k-1}, m_k), $k = 1, 2, ..., N$, they must be identical possibly except on m.

Now suppose that $h^0 = e^{-s^0} \in$ x-spl$^p(m), m = \{m_k\}_{k=0}^N$, is the density of a random variable X^0, which we aim to estimate. Then, for any $s \in$ e-spl$^p(m)$,

$$d_{KL}(h^0||e^{-s}) = E\{\log h^0(X^0)\} + E\{s(X^0)\}.$$

Hence, there is a constant term (with respect to s) in the objective function of (4) that can be dropped and we reach the fact that every optimal solution of

$$P^0_{p,m} : \quad \min_{s \in \text{e-spl}^p(m)} E\{s(X^0)\} \text{ s.t. } \int_{m_0}^{m_N} e^{-s(x)} dx = 1 \tag{5}$$

is identical to s^0, except possibly on m. Consequently, if the family x-spl$^p(m)$ under consideration contains the true density h^0, then $P^0_{p,m}$ recovers h^0 or a member in its 'equivalence class.' In contrast to $P^\nu_{p,m}$, we refer to $P^0_{p,m}$ as the *true problem.* Intuitively, if $s^0 \in S^\nu$ and ν is large, the problem $P^\nu_{p,m}$ approximates the true problem in some sense and one would hope that the corresponding optimal solutions are close. We next formalize this observation, which implies strong consistency of the estimator $h^\nu = e^{-s^\nu}$ obtained from solving $P^\nu_{p,m}$.

Theorem 4 (consistency). *Suppose that the true density $h^0 = e^{-s^0}$, with $s^0 = \langle c_{p,m}(\cdot), r^0\rangle \in$ e-spl$^p(m)$ and $m = \{m_k\}_{k=0}^N$, $P^\nu_{p,m}$ is derived by independent sampling from h^0, and $\{s^\nu\}_{\nu=1}^\infty$ is a sequence of optimal solutions of $P^\nu_{p,m}$, with epi-spline parameters $\{r^\nu\}_{\nu=1}^\infty$.*

If $\lim R^\nu$ exists almost surely[5] *and is deterministic, then every accumulation point r^∞ of $\{r^\nu\}_{\nu=1}^\infty$ satisfies*

$$\langle c_{p,m}(\cdot), r^\infty\rangle \in \mathcal{P}^{S^\infty}_{p,m}(h^0) \text{ almost surely,}$$

where $S^\infty := \{s \in$ e-spl$^p(m) \mid s = \langle c_{p,m}(\cdot), r\rangle, r \in \lim R^\nu\}$.

Moreover, regardless of whether R^ν has a limit, if there exists a sequence $\{\hat{r}^\nu\}_{\nu=1}^\infty$, with $\hat{r}^\nu \in R^\nu$ for all ν, such that $\hat{r}^\nu \to r^0$ almost surely, then the following hold almost surely.

(i) The accumulation point r^∞ also satisfies $\langle c_{p,m}(\cdot), r^\infty\rangle \in \mathcal{P}_{p,m}(h^0)$.

(ii) The essential epi-spline parameter subvector of r^∞ is identical to the essential epi-spline parameter subvector of r^0.

(iii) If $r^\nu \to^K r^\infty$ along a subsequence K, then $\langle c_{p,m}(\cdot), r^\nu\rangle \to^K s^0$ and $e^{-\langle c_{p,m}(\cdot), r^\nu\rangle} \to^K h^0$ uniformly on $[m_0, m_N]$, possibly except on m.

[5] Limits of sets are here taken in the sense of Painlevé-Kuratowski [29, §7.B] and the probability space is that induced by $\{P^\nu_{p,m}\}_{\nu=1}^\infty$.

Proof: Since $X^0 \in [m_0, m_N]$ almost surely, $c_{p,m}(X^0)$ is a random vector with finite moments. By the law of large number $(1/\nu)\sum_{i=1}^{\nu} c_{p,m}(X^i) \to E\{c_{p,m}(X^0)\}$ almost surely. Let $\hat{r}^0 \in I\!R^{(p+2)N+1}$ be arbitrary. Then, for any sequence $\hat{r}^\nu \to \hat{r}^0$,

$$\left\langle \frac{1}{\nu}\sum_{i=1}^{\nu} c_{p,m}(X^i), \hat{r}^\nu \right\rangle \to \left\langle E\{c_{p,m}(X^0)\}, \hat{r}^0 \right\rangle \text{ almost surely.}$$

For any $R \subset I\!R^{(p+2)N+1}$, we define $\delta_R(r) := 0$ if $r \in R$ and $\delta_R(r) := \infty$ otherwise. Moreover, let $R_I^\infty = \{r \in \lim R^\nu \mid \int_{m_0}^{m_N} e^{-c_{p,m}(x),r\rangle} dx = 1\}$. If $\hat{r}^0 \in R_I^\infty$, then

$$\liminf \left\langle \frac{1}{\nu}\sum_{i=1}^{\nu} c_{p,m}(X^i), \hat{r}^\nu \right\rangle + \delta_{R_I^\nu}(\hat{r}^\nu) \geq \langle E\{c_{p,m}(X^0)\}, \hat{r}^0\rangle + \delta_{R_I^\infty}(\hat{r}^0) \text{ almost surely.}$$

Since $R_I^\infty = \lim R_I^\nu$, it is closed. Consequently, if $\hat{r}^0 \notin R_I^\infty$, then the previous inequality holds with infinity on both sides. Next, suppose that $\hat{r}^0 \in I\!R^{(p+2)N+1}$ is arbitrary. If $\hat{r}^0 \notin R_I^\infty$, then

$$\limsup \left\langle \frac{1}{\nu}\sum_{i=1}^{\nu} c_{p,m}(X^i), \hat{r}^\nu \right\rangle + \delta_{R_I^\nu}(\hat{r}^\nu) \leq \langle E\{c_{p,m}(X^0)\}, \hat{r}^0\rangle + \delta_{R_I^\infty}(\hat{r}^0) = \infty \text{ almost surely.}$$

If $\hat{r}^0 \in R_I^\infty$, then, since $R_I^\infty = \lim R_I^\nu$, there exists a sequence $\hat{r}^\nu \to \hat{r}^0$ with $\hat{r}^\nu \in R_I^\nu$ for all ν. Consequently,

$$\left\langle \frac{1}{\nu}\sum_{i=1}^{\nu} c_{p,m}(X^i), \hat{r}^\nu \right\rangle + \delta_{R_I^\nu}(\hat{r}^\nu) \to \langle E\{c_{p,m}(X^0)\}, \hat{r}^0\rangle + \delta_{R_I^\infty}(\hat{r}^0) \text{ almost surely.}$$

Almost sure epi-convergence of $\langle(1/\nu)\sum_{i=1}^{\nu} c_{p,m}(X^i), \cdot\rangle + \delta_{R_I^\nu}$ to $\langle E\{c_{p,m}(X^0)\}, \cdot\rangle + \delta_{R_I^\infty}$ then follows by Proposition 7.2 in [29] and the first conclusions by Theorem 7.31 of [29] and the fact that $\hat{r} \in \operatorname{argmin}_r \langle E\{c_{p,m}(X^0)\}, r\rangle + \delta_{R_I^\infty}$ if and only if $\langle c_{p,m}(\cdot), \hat{r}\rangle \in \mathcal{P}_{p,m}^{S\infty}(h^0)$.

We next turn to the second part of the theorem. Since the additional assumption implies that R^ν becomes arbitrary close to r^0 almost surely, item (i) follows by a similar argument as above. Items (ii) and (iii) are conclusions from the discussion following Definition 5. □

The first part of Theorem 4 shows that regardless of the soft information, which even may *exclude* the true density, the resulting exponential epi-splines tend to one that is as 'close' as possible to the true density under the given constraints as the sample size increases. Specifically, the epi-splines computed from $\{P_{p,m}^\nu\}_{\nu=1}^\infty$ tend to a point in the Kullback-Leibler projection, *relative* to the soft information constraint set, of the true density on the class of epi-splines under consideration. The second part shows that if the true density is not excluded by the soft information, then $\{P_{p,m}^\nu\}_{\nu=1}^\infty$ eventually yields the true density, or possibly a closely related one that deviates at most on m.

The preceding results deal with the case when the true density can be exactly represented by an exponential epi-spline. If the true density is outside the class under consideration, one can't expect to tend to the true density even if the sample size goes to infinity. However, as we see next, if two densities are close in the hypo-distance, then their Kullback-Leibler projections on e-spl$^p(m)$ must also be close in some sense. We'll see that this has a direct consequence on the quality of density estimates when the true density is outside the class of exponential epi-splines. Before the main theorem, we give an intermediate result.

Proposition 4 *Suppose that $f^\nu : I\!R \to [0,\infty]$, $f^0 : I\!R \to [0,\infty]$ are Lebesgue integrable on every compact subset of $I\!R$ and $d\!l(-f^\nu, -f^0) \to 0$. Then, for every compact set $X \subset I\!R$,*

$$\int_X f^\nu(x)dx \to \int_X f^0(x)dx.$$

Proof: The restrictions of f^ν and f^0 to X, denoted by f_X^ν and f_X^0, satisfy $d\!l(-f_X^\nu, -f_X^0) \to 0$. Consequently, $A_X^\nu = \{(x, x_0) \in X \times [0,\infty) \mid f_X^\nu(x) \geq x_0\} \to A_X^0 = \{(x, x_0) \in X \times [0,\infty) \mid f_X^0(x) \geq x_0\}$ in the Painlevé-Kuratowski sense. Since the Lebesgue measures of A_X^ν and A_X^0 are identical to $\int_X f^\nu(x)dx$ and $\int_X f^0(x)dx$, respectively, the conclusion follows. □

Theorem 5 (stability of Kullback-Leibler projection). *Suppose that densities h^ν, h^0 on $[l,u]$ satisfy $d\!l(-h^\nu, -h^0) \to 0$. If r^ν is such that $\langle c_{p,m}(\cdot), r^\nu\rangle \in \mathcal{P}_{p,m}(h^\nu)$ for $m = \{m_k\}_{k=0}^N$ with $m_0 = l$, $m_N = u$, then every accumulation point of $\{r^\nu\}_{\nu=1}^\infty$ is the epi-spline parameter of some $s^0 \in \mathcal{P}_{p,m}(h^0)$.*

Proof: Following a similar argument as in Proposition 2, we see that the equality constraints in the problems defining $\mathcal{P}_{p,m}(h^\nu)$ and $\mathcal{P}_{p,m}(h^0)$ can be replaced by inequality. Consequently, every $s^\nu \in \mathcal{P}_{p,m}(h^\nu)$ is of the form $s^\nu = \langle c_{p,m}(\cdot), r^\nu\rangle$, with $r^\nu \in \operatorname{argmin}_r \psi^\nu(r) + \delta_I(r)$, where

$$\psi^\nu(r) = \left\langle \int_{m_0}^{m_N} c_{p,m}(x)h^\nu(x)dx, r \right\rangle$$

and $\delta_I(r) = 0$ if $\int_{m_0}^{m_N} e^{-\langle c_{p,m}(x), r\rangle}dx \leq 1$ and $\delta_I(r) = \infty$ otherwise. Similarly, every $s^0 \in \mathcal{P}_{p,m}(h^0)$ is of the form $s^0 = \langle c_{p,m}(\cdot), r^0\rangle$, where r^0 is a minimizer of ψ^0 defined similar to ψ^ν, but with h^ν replaced by h^0. Clearly, $\psi^\nu + \delta_I$ and $\psi^0 + \delta_I$ are convex.

By Proposition 4, $\int_X h^\nu(x)dx \to \int_X h^0(x)dx$ for any compact set $X \subset [m_0, m_N]$. But since $c_{p,m}$ is piecewise polynomial and $[m_0, m_N]$ is a bounded interval, we also have that for any $k = 1, 2, ..., N$,

$$\int_{m_{k-1}}^{m_k} c_{p,m}(x)h^\nu(x)dx \to \int_{m_{k-1}}^{m_k} c_{p,m}(x)h^0(x)dx.$$

Hence, it follows by Proposition 7.2 and Theorem 7.53 in [29] that $\psi^\nu + \delta_I$ totally epi-converges to $\psi^0 + \delta_I$. The result then is a consequence of Corollary 7.55 in [29]. □

If we take the densities h^ν in Theorem 5 to be exponential epi-splines, possibly defined on increasingly fine meshes, Theorem 1 shows that these densities indeed can be made to approximate with arbitrary accuracy any lsc or usc density h^0 with appropriate choice of mesh. Consequently, the assumption of $d\!l(-h^\nu, -h^0) \to 0$ in Theorem 5 holds and, combined with Theorem 4, we find that for a fine mesh and a large sample size the resulting exponential epi-spline estimator is 'close' to the true density, even if that density is outside the class of exponential epi-splines.

'Convergence' in the Kullback-Leibler divergence is closely related to other modes of convergence. Before we make these connection clear, we state an immediate consequence of the definition of the divergence that is also useful when constructing the set R^ν describing soft information.

Proposition 5 *Suppose that h and e^{-s} are densities with $s = \langle c_{p,m}(\cdot), r\rangle \in \text{e-spl}^p(m)$, $m = \{m_k\}_{k=0}^N$. Then,*

$$d_{KL}(h||e^{-s}) = \left\langle \int_{m_0}^{m_N} c_{p,m}(x)h(x)dx, r \right\rangle + \int_{-\infty}^{\infty} (\log h(x))h(x)dx.$$

If in addition $h = e^{-s'}$ with $s' = \langle c_{p,m}(\cdot), r'\rangle \in \text{e-spl}^p(m)$, then

$$d_{KL}(h||e^{-s}) = \left\langle \int_{m_0}^{m_N} c_{p,m}(x)h(x)dx, r - r' \right\rangle.$$

The next connections complement Theorem 2.

Proposition 6 *Suppose that densities $h^\nu, h^0 \in \text{x-spl}^p(m)$, with $h^\nu = e^{-\langle c_{p,m}(\cdot),(r^\nu_{\text{mesh}}, r^\nu_{\text{ess}})\rangle}$ and $h^0 = e^{-\langle c_{p,m}(\cdot),(r^0_{\text{mesh}}, r^0_{\text{ess}})\rangle}$. Then,*

$$(r^\nu_{\text{mesh}}, r^\nu_{\text{ess}}) \to (r^0_{\text{mesh}}, r^0_{\text{ess}}) \Longrightarrow d_{KL}(h^0||h^\nu) \to 0 \Longleftrightarrow d_{KL}(h^\nu||h^0) \to 0 \Longrightarrow r^\nu_{\text{ess}} \to r^0_{\text{ess}}.$$

Proof: We let $r^\nu = (r^\nu_{\text{mesh}}, r^\nu_{\text{ess}})$ and $r^0 = (r^0_{\text{mesh}}, r^0_{\text{ess}})$.

The implication $r^\nu \to r^0 \Longrightarrow d_{KL}(h^0||h^\nu) \to 0$ follows directly from Proposition 5.

To show that $d_{KL}(h^0||h^\nu) \to 0 \Longrightarrow r^\nu_{\text{ess}} \to r^0_{\text{ess}}$ we observe that $d_{KL}(\cdot||\cdot) \geq 0$ and for any two densities f, g on $[m_0, m_N]$, $d_{KL}(f||g) = 0$ if and only if $f(x) = g(x)$ for Lebesgue almost every $x \in [m_0, m_N]$. We therefore consider the problem $\min_{r\in R} d_{KL}(h^0||e^{-\langle c_{p,m}(x),r\rangle})$, with $R = \{r \in I\!R^{(p+2)N+1} \mid \int_{m_0}^{m_N} e^{-\langle c_{p,m}(x),r\rangle}dx = 1\}$, where r^0 is a minimizer and in fact every minimizer must coincide with r^0_{ess} in its last $(p+1)N$ components. In view of Proposition 5, the objective function in this problem is linear and the single constraint is continuously differentiable. The first-order optimality condition for this problem and the fact that $\{r \in I\!R^{(p+2)N+1} | \int_{m_0}^{m_N} e^{-\langle c_{p,m}(x),r\rangle}dx \leq 1\}$ is a convex set imply that the hyperplane $W = \{r \in I\!R^{(p+2)N+1} | d_{KL}(h^0||e^{-\langle c_{p,m}(x),r\rangle}) = 0\}$ is a supporting hyperplane of R with r^0_{ess} being the only $(p+1)N$-dimensional vector r_{ess} that can be augmented by a $\beta \in I\!R^{N+1}$ such that $\{(\beta, r_{\text{ess}})\} = R \cap W$. Since $r^\nu \in R$ and for sufficiently large ν is arbitrarily close to W, we consequently reach the conclusion.

We realize that $d_{KL}(h^\nu||h^0) \to 0 \Longrightarrow d_{KL}(h^0||h^\nu) \to 0$ by establishing that $r^\nu_{\text{ess}} \to r^0_{\text{ess}}$ whenever $d_{KL}(h^\nu||h^0) \to 0$ using a similar argument as above and then use Proposition 5.

We find that $d_{KL}(h^0||h^\nu) \to 0 \Longrightarrow d_{KL}(h^\nu||h^0) \to 0$ by invoking that $d_{KL}(h^0||h^\nu) \to 0 \Longrightarrow r^\nu_{\text{ess}} \to r^0_{\text{ess}}$ and Proposition 5. □

Asymptotic normality of the distribution of the exponential epi-spline estimator and corresponding moments may also hold when we limit the scope to the essential epi-spline parameters. As we see from the discussion before Proposition 3, one can't expect a unique estimator — a prerequisite for asymptotic normality — unless the scope is limited in this manner[6]. This focus on the essential epi-spline parameter requires additional notation that we introduce next.

For any $r_{\text{ess}} \in I\!R^{(p+1)N}$, let[7]

$$H(r_{\text{ess}}) := \int_{m_0}^{m_N} \rangle c_{\text{ess}}(x), c_{\text{ess}}\langle e^{-\langle c_{\text{ess}}(x), r_{\text{ess}}\rangle}dx$$

[6] One could appeal to more sophisticated central limit theorems, such as those in [15], but additional conditions and machinery is required and would require us to stray too far from our main theme.

[7] Here we use $\rangle y, y\langle$ to denote the outer product $yy^\top$ for a column vector y.

be the Hessian of $\int_{m_0}^{m_N} e^{-\langle c_{\text{ess}}(x),\cdot\rangle}dx$ at r_{ess}. We also let Σ_{ess} be the variance-covariance matrix of $c_{\text{ess}}(X^0)$, with X^0 distributed by the true density h^0, and $\Sigma(r_{\text{ess}}) = H(r_{\text{ess}})^{-1}\Sigma_{\text{ess}}H(r_{\text{ess}})^{-1}$, where we note that $H(r_{\text{ess}})$ is nonsingular by the argument in the proof of Proposition 3. For notational convenience, we also let $\Sigma_k(r_{\text{ess}})$ be the $(p+1)\times(p+1)$ submatrix of $\Sigma(r_{\text{ess}})$ consisting of elements in columns $(k-1)(p+1)+1$, $(k-1)(p+1)+2$, ..., $(k-1)(p+1)+(p+1)$ and the corresponding rows in the latter matrix. These are the coefficients corresponding to interval (m_{k-1}, m_k). Moreover, let $r_{\text{ess},k}$ be the subvector of components $N+1+(k-1)(p+1)+1$, ..., $N+1+(k-1)(p+1)+(p+1)$ of r_{ess}, i.e., the parameters that define the epi-spline in (m_{k-1}, m_k), and the corresponding subvectors of c_{ess} are denoted by $c_{\text{ess},k}$. Finally, we let $\mu_j^0 := \int_{-\infty}^{\infty} x^j h^0(x)dx$ be the jth moment of the true density h^0, $\mathcal{N}(0,\Sigma)$ denote a zero-mean normal vector with variance-covariance matrix Σ, and $\to^d$ convergence in distribution. We are now ready to state an asymptotic result for an exponential epi-spline estimator, where we make the assumption that the soft information is 'clearly' correct, i.e., the true density corresponds to a point in the interior of the sets R^ν almost surely for sufficiently large ν.

Theorem 6 (asymptotics). *Suppose that the true density $h^0 = e^{-s^0} \in$ x-spl$^p(m)$, with $m = \{m_k\}_{k=0}^N$, $s^0 = \langle c_{p,m}(\cdot), r^0\rangle$, and $r^0 = (r^0_{\text{mesh}}, r^0_{\text{ess}})$ is in the interior of* $\liminf R^\nu$ *almost surely. If $P^\nu_{p,m}$ is derived by independent sampling from h^0 and $\{s^\nu\}_{\nu=1}^\infty$ is a sequence of optimal solutions of $P^\nu_{p,m}$, with epi-spline parameters $\{r^\nu = (r^\nu_{\text{mesh}}, r^\nu_{\text{ess}})\}_{\nu=1}^\infty$, and $h^\nu = e^{-\langle c_{p,m}(\cdot),r^\nu\rangle}$ for all ν, then the following hold:*

(i)

$$\nu^{1/2}(r^\nu_{\text{ess}} - r^0_{\text{ess}}) \to^d \mathcal{N}(0, \Sigma(r^0_{\text{ess}}))$$

(ii) For $x \in (m_{k-1}, m_k)$, $k = 1, 2, ..., N$,

$$\nu^{1/2}(h^\nu(x) - h^0(x)) \to^d \mathcal{N}\left(0, e^{-2\langle c_{\text{ess},k}(x), r_{\text{ess},k}\rangle}\langle c_{\text{ess},k}(x), \Sigma_k(r^0_{\text{ess}})c_{\text{ess},k}(x)\rangle\right).$$

(iii) For $j \in I\!N$, the moment estimator $\mu_j^\nu := \int_{m_0}^{m_N} x^j e^{-\langle c_{p,m}(x),r^\nu\rangle}dx$ satisfies

$$\nu^{1/2}(\mu_j^\nu - \mu_j^0) \to^d \mathcal{N}(0, \langle w, \Sigma(r^0_{\text{ess}})w\rangle), \text{ where } w = \int_{m_0}^{m_N} x^j c_{\text{ess}}(x)e^{-\langle c_{p,m}(x),r^0\rangle}dx.$$

Proof: The law of large number gives that the objective function in $P^\nu_{p,m}$ converges uniformly on compact sets to that of $P^0_{p,m}$ almost surely. We recall that $\langle c_{p,m}(\cdot), r^0\rangle$ is an optimal solution of $P^0_{p,m}$ and, by assumption, r^0 is also in the interior of $\liminf R^\nu$ almost surely. Consequently, since $P^0_{p,m}$ doesn't involve a restriction S^ν, the set of optimal solutions of $P^\nu_{p,m}$ coincides with those of the relaxation of $P^\nu_{p,m}$ with S^ν replaced by e-spl$^p(m)$ for sufficiently large ν. Let

$$P^\nu_{\text{ess}}: \quad \min_{r_{\text{ess}}\in R^{(p+1)N}} \frac{1}{\nu}\sum_{i=1}^{\nu}\langle c_{\text{ess}}(X^i), r_{\text{ess}}\rangle + \int_{m_0}^{m_N} e^{-\langle c_{\text{ess}}(x),r_{\text{ess}}\rangle}dx,$$

where X^1, X^2, ..., X^ν is the sample from h^0. We deduce from Propositions 2 and 3 that P^ν_{ess} and the relaxation of $P^\nu_{p,m}$ have unique optimal solutions almost surely and that they are equivalent in the sense that they generate the same essential epi-spline parameter. Consequently, for sufficiently large ν, the optimal solution of P^ν_{ess} is r^ν_{ess} almost surely.

We let X^0 be a random variable with density h^0 and

$$P^0_{\text{ess}}: \quad \min_{r_{\text{ess}} \in R^{(p+1)N}} E\{\langle c_{\text{ess}}(X^0), r_{\text{ess}}\rangle\} + \int_{m_0}^{m_N} e^{-\langle c_{\text{ess}}(x), r_{\text{ess}}\rangle} dx.$$

We deduce from Propositions 2 and 3 that an optimal solution of this problem is unique and coincides with the essential epi-spline parameter r^0_{ess} of h^0.

Since P^0_{ess} and P^ν_{ess} are strictly convex and unconstrained almost surely, their unique optimal solutions are equivalently characterized as zeros of the objective function gradients. Since these gradients converge uniformly on $I\!R^{(p+1)N}$ almost surely by the law of large numbers, and the corresponding Hessians are identical and positive definite, item (i) follows directly from Theorem 4 of [23]. The next items follow by a direct application of a Delta Theorem; see for example Section 7.2.7 in [35]. □

While Theorem 6 provides rates of convergence, it excludes the effect of soft information given by R^ν and deals only with the essential epi-spline parameter. We end the section by examining errors for a finite sample size under relaxed assumptions, which leads to another rate of convergence result. However, the treatment requires us to focus on *ϵ-optimal solutions* of $RP^\nu_{p,m}$, which for any $\epsilon \geq 0$ are defined as

$$\mathcal{R}^\nu_\epsilon := \left\{ r \in R^\nu \;\middle|\; \frac{1}{\nu}\sum_{i=1}^{\nu} \langle c_{p,m}(X^i), r\rangle \leq V^\nu + \epsilon, \; \int_{m_0}^{m_N} e^{-\langle c_{p,m}(x), r\rangle} dx \leq 1 \right\},$$

where the optimal value of $RP^\nu_{p,m}$ is

$$V^\nu := \inf_{r \in R^\nu} E\{\langle c_{p,m}(X^0), r\rangle\} \text{ s.t. } \int_{m_0}^{m_N} e^{-\langle c_{p,m}(x), r\rangle} dx \leq 1.$$

The statement below deals with the difference between the true density h^0 and $h^\nu_\epsilon = e^{-\langle c_{p,m}(\cdot), r^\nu_\epsilon\rangle}$, with $r^\nu_\epsilon \in \mathcal{R}^\nu_\epsilon$ for $\epsilon > 0$. The latter density is, in fact, what is generated by numerical methods for solving $RP^\nu_{p,m}$ as such methods utilize finite precision and various tolerances. Also let $\rho I\!B := \{y \mid \|y\| \leq \rho\}$ in any Euclidian space and $\Delta_{p,m} := \max_{l=0,1,\dots,p} |m|^l$.

Theorem 7 (finite sample error). *Suppose that the true density $h^0 \in \text{x-spl}^p(m)$, $m = \{m_k\}_{k=0}^N$, with epi-spline parameter r^0, $RP^\nu_{p,m}$ is derived by independent sampling from h^0, and is also feasible with a closed and convex R^ν almost surely. For any $\alpha > 0$, $\epsilon > 0$, and $\rho > \max\{-V^\nu, d(r^0, \mathcal{R}^\nu_0)\}$,*

$$d(r^0, \mathcal{R}^\nu_\epsilon) > K$$

$$d_{KL}(h^0 || h^\nu_\epsilon) > \left\| \int_{m_0}^{m_N} c_{p,m}(x) h^0(x) dx \right\| K, \textit{ for } h^\nu_\epsilon = e^{-\langle c_{p,m}(\cdot), r^\nu_\epsilon\rangle}, r^\nu_\epsilon \in \mathcal{R}^\nu_\epsilon,$$

with probability at most $2(p+1)Ne^{-2\nu(\alpha/\Delta_{p,m})^2}$, where[8]

$$K = (1 + 4\rho/\epsilon)[\alpha\sqrt{(p+1)N}(\rho + \|r^0\|) + (1 + \sqrt{(p+2)N+1}\Delta_{p,m}) d(r^0, R^\nu)].$$

[8] Here, $d(x, S) := \inf_{y \in S} \|x - y\|$ for $x \in I\!R^n, S \subset I\!R^n$.

Proof: Let X^0 be a random variable with density h^0 and $X^1, X^2, ..., X^\nu$ be the sample that generates $P^\nu_{p,m}$. We denote by $c^j_{p,m}(X^0)$ the components of $c_{p,m}(X^0)$, $j = 1, 2, ..., (p+2)N+1$. For $j = 1, 2, ..., N+1$, $c^j_{p,m}(X^0) = 1$ if $X^0 = m_{j-1}$ and $c^j_{p,m}(X^0) = 0$ otherwise. Consequently, $E\{c^j_{p,m}(X^0)\} = 0$ and, likewise, $(1/\nu)\sum_{i=1}^\nu c^j_{p,m}(X^i) = 0$ almost surely. For $j = N+1+(p+1)(k-1)+l+1$, $l = 0, 1, ..., p$, $k = 1, 2, ..., N$, $c^j_{p,m}(X^0) = (X^0 - m_{k-1})^l$ if $X^0 \in (m_{k-1}, m_k)$ and $c^j_{p,m}(X^0) = 0$ otherwise. Consequently, for $j = N+2, N+3, ..., (p+2)N+1$, $c^j_{p,m}(X^0) \in [0, \Delta_{p,m}]$ almost surely and by Hoeffding's inequality,

$$P\left(\left|\frac{1}{\nu}\sum_{i=1}^{\nu} c^j_{p,m}(X^i) - E\{c^j_{p,m}(X^0)\}\right| \geq \alpha\right) \leq 2e^{-2\nu(\alpha/\Delta_{p,N})^2}$$

for every $\alpha \geq 0$. Moreover, Boole's inequality gives, when taking advantage of the zero error for $j = 1, ..., N+1$, that

$$P\left(\bigcup_{j=1}^{(p+2)N+1}\left\{\left|\frac{1}{\nu}\sum_{i=1}^{\nu} c^j_{p,m}(X^i) - E\{c^j_{p,m}(X^0)\}\right| \geq \alpha\right\}\right) \leq 2(p+1)Ne^{-2\nu(\alpha/\Delta_{p,m})^2}.$$

Let $\phi^\nu : \mathbb{R}^{(p+2)N+1} \to \overline{\mathbb{R}}$ be defined by $\phi^\nu = (1/\nu)\sum_{i=1}^\nu \langle c_{p,m}(X^i), \cdot\rangle + \delta^\nu(\cdot)$ where $\delta^\nu(r) = 0$ if $r \in R^\nu$ and $\int_{m_0}^{m_N} e^{-\langle c_{p,m}(x), r\rangle}dx \leq 1$, and $\delta^\nu(r) = \infty$ otherwise. Let $\phi^{0,\nu} : \mathbb{R}^{(p+2)N+1} \to \overline{\mathbb{R}}$ be defined by $\phi^{0,\nu} = E\{\langle c_{p,m}(X^0), \cdot\rangle\} + \delta^{0,\nu}(\cdot)$ where $\delta^{0,\nu}(r) = 0$ if $\int_{m_0}^{m_N} e^{-\langle c_{p,m}(x), r\rangle}dx \leq 1$ and r is in the convex hull of R^ν and r^0, and $\delta^{0,\nu}(r) = \infty$ otherwise.

In view of the preceding results and definitions, for $r - r^0 \in \rho\mathbb{B}$, with $\rho \in (0, \infty)$,

$$\left|(1/\nu)\sum_{i=1}^{\nu}\langle c_{p,m}(X^i), r\rangle - E\{\langle c_{p,m}(X^0), r\rangle\}\right| \leq \alpha\sqrt{(p+1)N}(\rho + \|r^0\|)$$

with at least probability $1 - 2(p+1)Ne^{-2\nu(\alpha/\Delta_{p,m})^2}$. Using this fact, Example 7.62 of [29] gives that with the same probability,

$$\widehat{d\!l}^{\,+}_{\rho}(\phi^\nu, \phi^{0,\nu}) \leq \alpha\sqrt{(p+1)N}(\rho + \|r^0\|) + (1 + \sqrt{(p+2)N+1}\Delta_{p,m})d(r^0, R^\nu),$$

where $\widehat{d\!l}^{\,+}_{\rho}$ is closely related to $d\!l_\rho$; see Section 7.I in [29]. Then, from Theorem 7.69 in [29], we deduce the first result after realizing that r^0 is an ϵ-optimal solution of $\min\phi^{0,\nu}$, where the additional factor $1 + 4\rho/\epsilon$ arises from that theorem. Proposition 5 yields the second conclusion. □

Theorem 7 shows that there are two sources of error in the estimation process corresponding to the two parts of K. The first source is sampling error, represented by the term involving α, which can be made small by selecting a small α and this error is only exceeded with a small probability if ν/α^2 is large. The second source is caused by $d(r^0, R^\nu)$, the distance between the true epi-spline parameter and the constraint set R^ν. Of course, if only appropriate soft information is included, then $r^0 \in R^\nu$ and $d(r^0, R^\nu) = 0$. Otherwise, incorrect specification of soft information induces a 'bias' in the density estimator. We also note that Theorem 7 provides additional support for considering $RP^\nu_{p,m}$ also for instances which are not loosely constrained. Even in such cases, $RP^\nu_{p,m}$ is guaranteed to generate a density near the true density.

We recall the notion of 'bounded in probability.' For a sequence of random variables $\{Y^\nu\}_{\nu=1}^\infty$, we write $Y^\nu = O_p(1)$ when for any $\zeta > 0$, there exists a $\beta \geq 0$ such that $\text{Prob}(|Y^\nu| > \beta) \leq \zeta$ for all ν.

Corollary 1 *For sufficiently large ν, suppose that the assumptions of Theorem 7 hold and $d(r^0, R^\nu) = 0$ almost surely. Then,*

$$\nu^{1/2} d_{KL}(h^0||h_\epsilon^\nu) = O_p(1) \textit{ for any } h_\epsilon^\nu = e^{-\langle c_{p,m}(\cdot), r_\epsilon^\nu\rangle}, r_\epsilon^\nu \in \mathcal{R}_\epsilon^\nu.$$

Proof: Theorem 7 and the fact that $d(r^0, R^\nu) = 0$ imply that

$$\text{Prob}(\nu^{1/2} d_{KL}(h^0||h_\epsilon^\nu) > K'\alpha\nu^{1/2}) \leq 2(p+1)Ne^{-2\nu(\alpha/\Delta_{p,m})^2}$$

for sufficiently large ν, where $K' = \left\|\int_{m_0}^{m_N} c_{p,m}(x)h^0(x)dx\right\| (1+4\rho/\epsilon)\sqrt{(p+1)N}(\rho+\|r^0\|)$. We let $\zeta > 0$ and couple α and ν such that $\zeta = 2(p+1)Ne^{-2\nu(\alpha/\Delta_{p,m})^2}$, i.e., $\nu = -\Delta_{m,p}^2 \log(\zeta/2(p+1)N)/(2\alpha^2)$. Conseqently,

$$\text{Prob}(\nu^{1/2} d_{KL}(h^0||h_\epsilon^\nu) > \beta) \leq \zeta,$$

where $\beta = K'(-\Delta_{m,p}^2 \log(\zeta/2(p+1)N)/2)^{1/2}$ and the conclusion follows. □

In view of the preceding result, we see that the canonical rate of $\nu^{-1/2}$ is obtained for the exponential epi-spline estimator even if soft information is 'active.'

4 Soft Information

We implement soft information about the density under consideration in the estimation problem $\bar{P}_{p,m}^\nu$ through the set R^ν. Intuitively, we expect that soft information may improve density estimates, which we also see empirically in Section 5. In fact, if the true density $h^0 = e^{-\langle c_{p,m}(\cdot), r^0\rangle}$, with $r^0 \in R^\nu$ and there exists a $\rho > 0$ such that $\|r - r'\| \leq \rho$ for all $r, r' \in R^\nu$ almost surely, then in view of Proposition 5

$$d_{KL}(h^0||h^\nu) \leq \left\|\int_{m_0}^{m_N} c_{p,m}(x)h^0(x)dx\right\| \rho.$$

Consequently, an effective strategy for improving exponential epi-spline estimates would be to reduce the size of R^ν, of course, without eliminating the true epi-spline parameter. Naturally, with the inclusion of questionable soft information, there is a need for validation. While important, we omit a discussion of this topic; see for example [36] and [5] for tests in related contexts.

We next consider examples of R^ν. It is straightforward to translate many important types of soft information into constraints on the epi-spline parameters

$$r = (s_0, s_1, ..., s_N, a_{1,0}, a_{1,1}, ..., a_{1,p}, a_{2,0}, a_{2,1}, ..., a_{2,p}, ..., a_{N,0}, a_{N,1}, ..., a_{N,p}) \in I\!R^{(p+2)N+1},$$

where we recall that the first $N + 1$ components specify the value of the epi-spline at the mesh points m_0, m_1, ..., m_N and the remaining N blocks of $p + 1$ components give the polynomial of order p in each interval (m_{k-1}, m_k), $k = 1, 2, ..., N$. In fact, many types of soft information simply result in linear and convex constraints as we see below.

Support bounds and mesh. The choice of mesh $m = \{m_k\}_{k=0}^N$ accounts for support bounds and m_0 and m_N should, ideally, correspond to the lower and upper bound of the support of the true density, respectively. If these are unknown, the values can be selected such that the observed sample is well within $[m_0, m_N]$. The mesh is often selected to be uniform, but the methodology offers much flexibility. For example, if discontinuities and intervals with steep slopes can be anticipated, other choices may be preferred. We note, however, that under the assumption that a true density is in x-spl$^p(m)$ for some m and p, there is no imperative need for refining the mesh beyond m as the sample size increases.

Continuity. We ensure that an exponential epi-spline estimate is usc by the constraints

$$s_{k-1} \leq a_{k,0}, \quad s_k \leq \sum_{i=0}^{p} a_{k,i}(m_k - m_{k-1})^i, \quad k = 1, 2, ..., N.$$

Identical constraints with the inequalities reversed ensure lsc of the exponential epi-spline. Continuity would require the same constraints with equality. Of course, by omitting some of these constraints, one has the ability to ensure continuity on parts of m. All of these constraints are linear. Moreover, their inclusion will keep a problem loosely constrained as the sufficient condition for being loosely constrained in Proposition 2 is satisfied.

Smoothness. We restrict the search to r-times continuously differentiable densities, with $r \leq p$, by imposing the conditions for continuity and the linear constraints

$$\sum_{i=j}^{p} \prod_{l=0}^{j-1} (i - l) a_{k,i}(m_k - m_{k-1})^{i-j} = a_{k+1,j}, \quad k = 1, 2, ..., N-1, j = 1, 2, ..., r.$$

Higher order smoothness is automatically achieved if these constraints are imposed with $r = p$. Again, selective implementation of these constraints could be a useful tool in practice. Again, the inclusion of these constraints will keep a problem loosely constrained as the sufficient condition for being loosely constrained in Proposition 2 is satisfied.

Fisher information and related quantities. The Fisher information $\int_{-\infty}^{\infty} h'(x)^2/h(x)dx$ of a density h is a 'measure of smoothness' that is easily expressed in terms of the epi-spline parameter, but upper and lower bounds on this expression result in undesirable nonconvex constraints. However, an alternative 'normalization' results in a convex constraint. Specifically, if $h = e^{-\langle c_{p,m}(\cdot), r\rangle}$, then

$$\int_{-\infty}^{\infty} (h'(x)/h(x))^2 dx = \int_{m_0}^{m_N} \left\langle c'_{p,m}(x), r\right\rangle^2 dx = \sum_{k=1}^{N} \int_{m_{k-1}}^{m_k} \left(\sum_{i=1}^{p} i a_{k,i}(x - m_{k-1})^{i-1} \right)^2 dx.$$

An upper bound on this quantity results is a convex constraint. In some application, one may also seek bounds at $x \in (m_{k-1}, m_k)$ by restricting

$$h'(x)/h(x) = -\langle c'_{p,m}(x), r\rangle = -\sum_{i=1}^{p} i a_{k,i}(x - m_{k-1})^{i-1}$$

and/or

$$\begin{aligned} h''(x)/h(x) &= -\langle c''_{p,m}(x), r\rangle + \langle c'_{p,m}(x), r\rangle^2 \\ &= -\sum_{i=2}^{p} i(i-1)a_{k,i}(x-m_{k-1})^{i-2} + \left(\sum_{i=1}^{p} i a_{k,i}(x-m_{k-1})^{i-1}\right)^2. \end{aligned}$$

Upper and lower bounds on the first quantity results in linear constraints and upper bounds on the second quantity gives a quadratic convex constraint. The constraints could be imposed at any number of values of x, but we note that if $p = 2$ and the density is log-concave, as describe below, and continuously differentiable, then lower bounds on $h'(x)/h(x)$ at m_1, m_2, ..., m_N suffices to ensure that the constraints are satisfied for all $x \in [m_0, m_N]$. Similarly, an upper bound on $h'(x)/h(x)$ need only be imposed at m_0, m_1, ..., m_{N-1}. The inclusion of the pointwise constraints keep a problem loosely constrained as the sufficient condition for being loosely constrained in Proposition 2 is satisfied.

Monotonicity. We achieve a nondecreasing (nonincreasing) density by imposing nonnegativity (nonpositivity) on $h'(x)/h(x)$ for all $x \in (m_{k-1}, m_k), k = 1, 2, ..., N$ as well as

$$s_{k-1} \geq (\leq) a_{k,0}, \quad s_k \leq (\geq) \sum_{i=0}^{p} a_{k,i}(m_k - m_{k-1})^i, \quad k = 1, 2, ..., N.$$

Again, simplifications arise, for example, if $p = 2$ and the density is log-concave. Then, it suffices to impose that $a_{k,1} + 2a_{k,2}(m_k - m_{k-1}) \leq 0$ $(a_{k,1} \geq 0)$, $k = 1, 2, ..., N$. Again, a problem remains loosely constrained after the inclusion of these constraints.

Unimodality and Log-Concavity. We recall that $h = e^{-\langle c_{p,m}(\cdot), r\rangle}$ is log-concave if and only if $\langle c_{p,m}(\cdot), r\rangle$ is convex. This condition is ensured if $\langle c_{p,m}(\cdot), r\rangle$ is (i) continuous (see above), (ii) for $k = 1, 2, ..., N-1$, its left derivatives at m_k is no larger than its right derivative, i.e.,

$$\sum_{i=1}^{p} i a_{k,i}(m_k - m_{k-1})^{i-1} \leq a_{k+1,1}, \quad k = 1, 2, ..., N-1,$$

and (iii) on each (m_{k-1}, m_k), $k = 1, 2, ..., N$, $\langle c_{p,m}(\cdot), r\rangle$ is convex, i.e.,

$$\sum_{i=2}^{p} i(i-1)a_{k,i}(x - m_{k-1})^{i-2} \geq 0, \quad k = 1, 2, ..., N, x \in (m_{k-1}, m_k).$$

Here, the obvious interpretations are required when $p = 0, 1$. The latter condition simplifies to $a_{k,2} \geq 0$, $k = 1, 2, ..., N$, when $p = 2$. Hence, in that case, the condition of log-concavity requires only a finite number of linear constraints. Again, a problem remains loosely constrained. Since log-concavity implies unimodality, the preceding constraints are also sufficient to ensure unimodality of the resulting exponential epi-spline density.

Bounds on density values. It is straightforward to impose pointwise upper and lower bounds $u(x)$ and $l(x)$ on the value of $h(x) = e^{-\langle c_{p,m}(x), r\rangle}$, with $0 < l(x) \leq u(x) < \infty$. It suffices to set

$$-\log l(x) \geq \sum_{i=0}^{p} a_{k,i}(x - m_{k-1})^i \geq -\log u(x) \text{ for } x \in (m_{k-1}, m_k)$$

and

$$-\log l(x) \geq s_k \geq -\log u(x) \text{ for } x = m_k, k = 0, 1, ..., N.$$

While these constraints are linear, they don't satisfy the assumption of Proposition 2. However, if only the lower bound $h(x) \geq l(x)$ is imposed, the problem instance remains loosely constrained.

Kullback-Leibler divergence and the Bayesian paradigm. Proposition 5 provides a convenient form of implementing soft information about a reference density h^{ref}. In a Bayesian-like paradigm, suppose that we seek a density that is 'near' h^{ref}. Then, a constraint

$$d_{KL}(h^{\text{ref}} || e^{-\langle c_{p,m}(\cdot), r\rangle}) \leq \phi(\nu), \tag{6}$$

indeed ensures that the estimate h^ν is within $\phi(\nu)$ of h^{ref} as measured by the Kullback-Leibler divergence. In view of Proposition 5, this constraint is linear in r and thus easily implementable. Here, $\phi : I\!N_0 \to [0, \infty)$ is the *cognitive content* of the reference density h^{ref} and should satisfy $\phi(0) = 0$, $\lim_{\nu\to\infty} \phi(\nu) = \infty$, and be increasing since an increasing sample size should place gradually less emphasis on h^{ref}. Of course, if $\phi(\nu) = 0$, then $\bar{P}^\nu_{m,p}$ simply returns h^{ref}, or a density that deviates at most on m. If $\phi(\nu) = \infty$, then no information about the reference density is included. While technically not correct in the sense of classical Bayesian statistics, one can view h^{ref} as a 'prior' density and the resulting density h^ν obtained from $\bar{P}^\nu_{m,p}$ as the 'posterior' density. (An alternative to a constraint on the Kullback-Leibler divergence would be to constrain a norm between h^{ref} and $e^{-\langle c_{p,m}(\cdot), r\rangle}$, such as the mean squared error with respect to h^{ref}. However, such constraints would be nonconvex.) Of course, a constraint $d_{KL}(h^{\text{ref}} || e^{-\langle c_{p,m}(\cdot), r\rangle}) \geq \kappa$, for some $\kappa > 0$ is also easily implementable, and could be relevant in contexts where a 'diversity' of densities is sought. For example, one may be concerned with the validity of the soft information imposed in an initial estimate of a density and seek a set of alternative densities that are some distance away from the original estimate; see Section 5.2 for an example.

Bounds on moments. Soft information may result in constraints on the j-th moment:

$$l \leq \int_{m_0}^{m_N} x^j e^{-\langle c_{p,m}(x), r\rangle} dx \leq u, \tag{7}$$

where $l, u \in I\!R$, $l \leq u$ are given constants. The right-most inequality results in a convex constraint on r, while the left-most in a nonconvex constraint.

Bounds on cumulative distribution functions. Suppose that the cumulative distribution function of $h = e^{-\langle c_{p,m}(\cdot), r\rangle}$ at $\gamma \in [m_0, m_N]$ must lie between the lower bound l and the upper bound u. This results in the two convex constraints

$$\int_{m_0}^{\gamma} e^{-\langle c_{p,m}(x), r\rangle} dx \leq u \text{ and } \int_{\gamma}^{m_N} e^{-\langle c_{p,m}(x), r\rangle} dx \leq 1 - l.$$

5 Numerical Examples

We illustrate the exponential epi-spline estimators through a series of examples. While we don't attempt a comprehensive comparison across methods, we also compute kernel estimates using 'ksdensity' in Matlab, with the default normal kernel. The estimation problems are solved by 'fmincon' in Matlab versions 7.10.0. In all cases, we use epi-splines of order 2 and if there is no soft information about support bounds, we set m_0 (m_N) to two sample-estimated standard errors below (above) the smallest (largest) sample point. The Gauss-Legendre quadrature rule with 20 points evaluates the integrals over each segment (m_{k-1}, m_k). We often assess the quality of an estimate h^ν of a density h^0 by the mean-square error (MSE) $\int_{-\infty}^{\infty}(h^\nu(x) - h^0(x))^2 h^0(x)dx$.

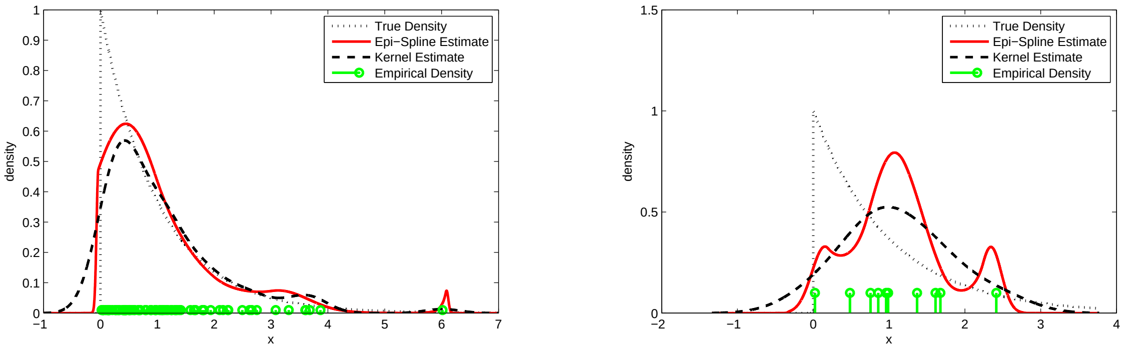
The section starts with showing that soft information can dramatically improve density estimates, both qualitatively and quantitatively. We proceed by discussing the Kullback-Leibler divergence, the challenging situation with discontinuities, as well as incorrect soft information. The section ends with average statistics over hundreds of replications. For additional numerical results we refer to [39, 31, 38].

5.1 Value of Soft Information

We consider the exponential density with parameter $\lambda = 1$ and show some typical results; see the end of the section and [39, 38] for average results over thousands of replications. Throughout this subsection, we use $N = 10$ and assume that the exponential epi-spline estimates are continuously differentiable. For moderately large sample sizes, both kernel and exponential epi-spline estimates capture the essence of the exponential density, though the nonnegative support is violated; see Figure 2(a) where $\nu = 100$ and the true density is the dotted black curve, the exponential epi-spline estimate is the solid red curve, the kernel estimate is the dashed black curve, and the green stems show the sample points. The MSE for the exponential epi-spline and kernel estimates are 0.0309 and 0.0515, respectively. While additional soft information improves the exponential epi-spline estimate, we provide no further details and instead turn to the more challenging situation with a sample size of $\nu = 10$. Figure 2(b) shows corresponding estimates in this case, where the MSE worsens to 0.1432 (exponential epi-spline) and 0.1285 (kernel). Neither the exponential epi-spline nor the kernel estimate resemble qualitatively the exponential density. However, additional soft information carries promise to improve the situation.

Figure 3(a) shows the estimates for the same sample as in Figure 2(b), but with nonnegative support also incorporated. While the estimates now have correct lower support bound, substantial oscilations in density values cause the MSE to increase to 0.2765 (exponential epi-spline) and 0.3273 (kernel). We note that the kernel estimate reaches well above 4.5 near zero, though the plot is truncated for the sake of clarity. There is no systematic way of incorporating further soft information in the kernel estimate. However, it is straightforward to ensure a log-concave exponential epi-spline estimate as shown in Figure 3(b). The exponential epi-spline estimate improves visually and the MSE reduces to 0.1144. In this and the following plots, the kernel estimate of Figure 3(a) is reproduced for the sake of comparison.

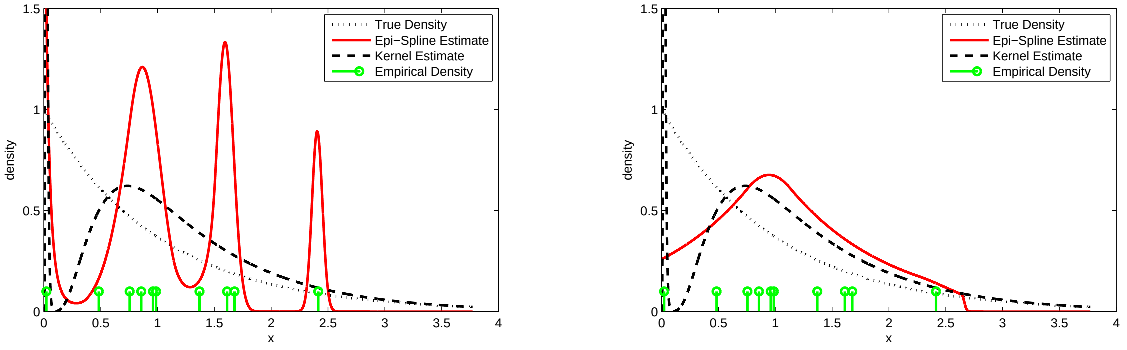
Further soft information improves the exponential epi-spline estimates. Figure 4(a) shows the visually improved exponential epi-spline estimate when we also assume a nonincreasing density. The MSE improves substantially to 0.0470. The exponential epi-spline



(a)

(b)

Figure 2: Exponential Example: Exponential epi-spline and kernel estimates for $\nu = 100$ (a) and $\nu = 10$ (b).

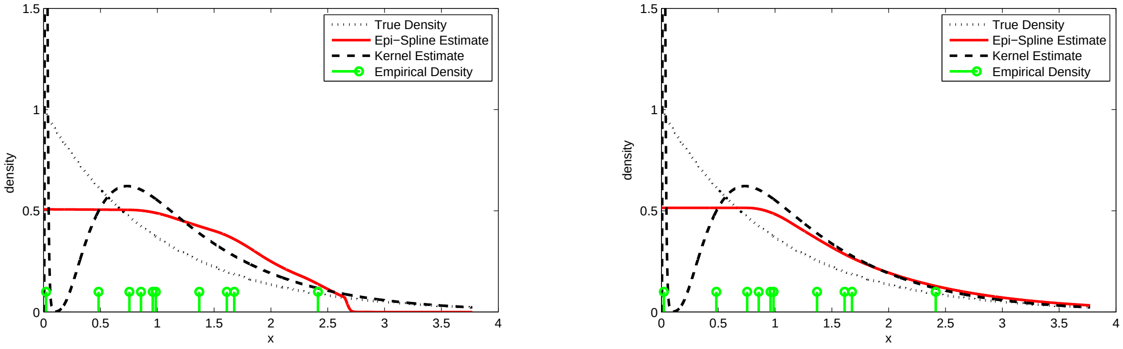


(a)

(b)

Figure 3: Exponential Example: Exponential epi-spline and kernel estimates for $\nu = 10$, with nonnegative support (a) and also log-concavity (b).

estimate drops off quickly in the first segment (m_{k-1}, m_k) after the last sample point as expected due to the maximum likelihood objective. Soft information that the 'pointwise Fisher' quantity $h^{\nu\prime}(x)/h^{\nu}(x)$ must lie in the interval $[-1, 0]$ remedies this deficiency. We observe that the exponential density h^0 with parameter $\lambda = 1$ has $h^{0\prime}(x)/h^0(x) = -1$ for all $x \geq 0$. The MSE of the exponential epi-spline improves to 0.0416 mainly due to improved tail estimate; see Figure 4(b). The resulting exponential epi-spline misses the density peak at zero, but the present sample provides few indications about such a peak and its capture will naturally be difficult. Still, the exponential epi-spline is both qualitatively and quantitatively close to the true density elsewhere. Lowering the upper bound on the pointwise Fisher quantify improves the estimate further, with a nearly perfect estimate (not depicted) when $h^{\nu\prime}(x)/h^{\nu}(x) = -1$ is required for all $x \geq 0$. The ability to incorporate various kinds of soft information along the lines illustrated here offers the statistician a valuable tool for exploring assumptions and their consequences.



(a)

(b)

Figure 4: Exponential Example: Exponential epi-spline and kernel estimates for $\nu = 10$, with also nonincreasing (a) and also $h'(x)/h(x) \in [-1, 0]$ (b) soft information.

5.2 Kullback-Leibler Divergence and the Bayesian Paradigm

As described in Section 4, our framework provides an alternative to traditional Bayesian updating. In addition to the inclusion of numerous types of soft information—which can be viewed as 'prior' information—we may also directly restrict $\bar{P}^{\nu}_{m,p}$ to a neighborhood of a reference density h^{ref} using (6). To illustrate the framework, consider a reference (prior) density that is standard normal and a sample consisting of 10 points from the same density; see Figure 5. We set $N = 10$ and restrict the search to continuously differentiable densities. If no emphasis is placed on the reference density, i.e., $\phi(10) = \infty$ in (6), then we obtain the exponential epi-spline estimate marked with the red dotted line in Figure 5. As proximity to the reference density is enforced more vigorously by setting $\phi(10) = 1$, 0.1, and 0.01, we obtain the dashdot, dashed, and solid lines, respectively, in Figure 5. The Kullback-Leibler divergence constraints dampen the variability caused by the sample by a degree determined by $\phi(10)$, which in practice should be selected based on the confidence in the correctness of the reference density.

A related situation arises when a statistician would like to generate multiple densities that span a range of possibilities, for example to account in some manner for questionable soft information. For example, when the estimated density is to be used as input in further simulation and optimization, it may be prudent to consider a set of densities and possibly let planning be based on the worst density in some sense. We illustrate this situation by returning to the exponential example of Section 5.1. Suppose that the last density generated there (see Figure 4(b)) is considered plausible, but we would like to also generate relevant alternatives. Retaining a restriction to continuously differentiable, nonincreasing, and nonnegatively supported densities, we construct three alternatives by imposing (6) with $\leq$ replaced by $\geq$ and right-hand side $0.1, 0.01$, and 0.001, and h^{ref} being the original estimate in Figure 4(b). Consequently, we determine densities that are at least certain 'distances' away from the original estimate in the sense of Kullback-Leibler divergence, while still maximizing the log-likelihood function of the sample. Figure 6 shows the results with the solid red line and dotted black line showing the original estimate and true density as in Figure 4(b). The alternative densities are depicted with dashed, dot-dashed, and dotted red lines for right-hand sides of $0.001, 0.01$, and 0.1, respectively. We observe that even though based on only

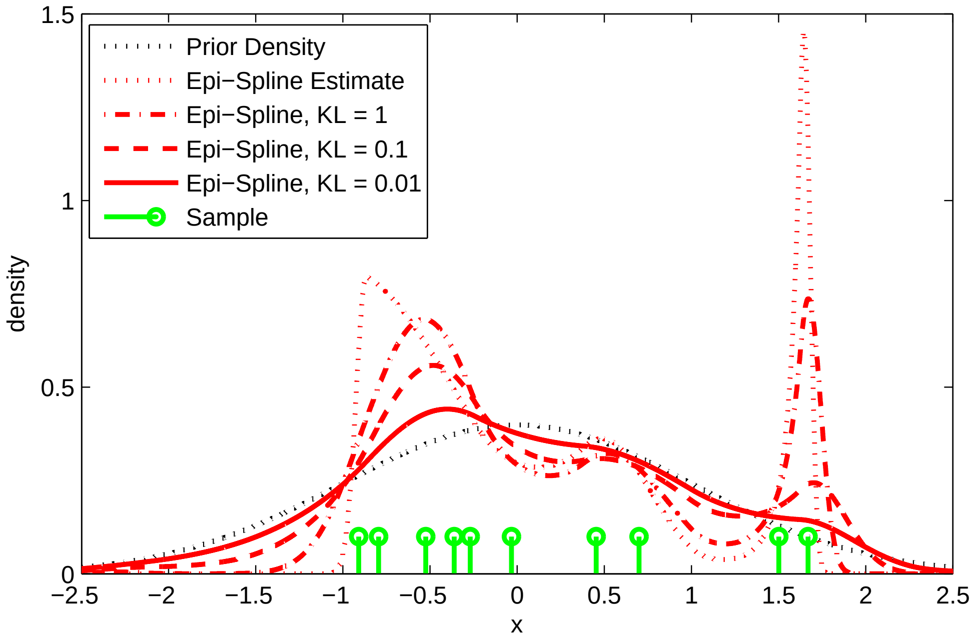


Figure 5: Normal Example: Kullback-Leibler divergence constraint.

10 sample points, the original together with the alternative densities provide a 'diversified' set of densities near the true density well suited as input for further studies.

5.3 Estimation of a Discontinuous Density

Significant challenges arise when the density to be estimated is discontinuous. We illustrate this situation by considering the 'uniform mixture' density

$$h^0(x) = 2 \text{ if } x = (0.1(k-1), 0.1(k-1)+0.05),\ k = 1,2,...,10, \text{ and } h^0(x) = 0 \text{ otherwise.}$$

Figure 7(a) shows the density (dotted black line) together with a kernel estimate (dashed black line) based on a sample of size 1000. Clearly, this kernel estimate is unable to capture the discontinuities in the mixture density. Other kernel estimators may improve the situation, but the selection of kernel base and bandwidth is generally difficult a prior. An exploration of such parameters is beyond the scope of the paper. We compute an exponential epi-spline estimate using $N = 50$ segments. It is natural to select a large number of segments when one is suspicious that the density might be discontinuities and we want to ensure a sufficiently flexible epi-spline. We also enforce the lsc constraints. However, with no sample points coinciding with m, and no other mesh related constraints, these constraints only influence the density estimate on the mesh and therefore are essentially superfluous. Finally, we let the pointwise Fisher quantity $h^{\nu\prime}(x)/h^\nu(x) \in [-1,1]$ for $x \in (m_{k-1}, m_k)$, $k = 1,2,...,N$. With the large number of segments and the possibility for discontinuities, this restriction improves the accuracy only marginally but ensure visually more accessible plots. Figure 7(a) shows the resulting exponential epi-spline estimate (solid red line). The MSE is 0.5724 in contrast with 1.1702 for the kernel estimate. We see that the exponential epi-spline estimate

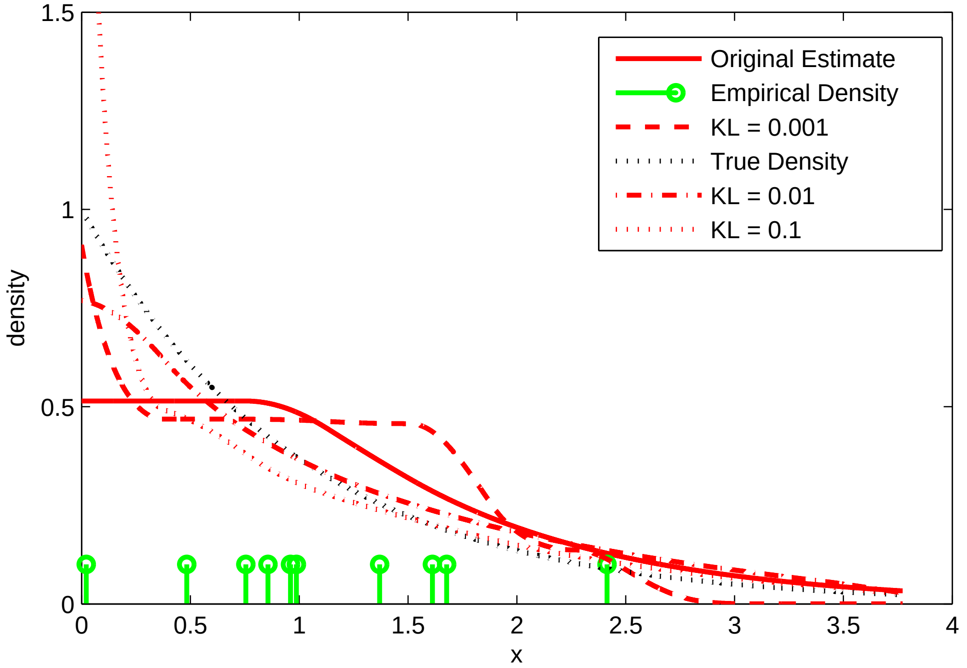


Figure 6: Exponential Example: Diversification through Kullback-Leibler divergence.

captures fairly well the mixture density, even though we don't provide any soft information about the support. As predicted by Theorem 1, results improve as N increases to 100; see Figure 7(b). The exponential epi-spline (solid red line) now tracks the mixture density to a large degree obtaining a MSE of 0.3518. Of course, the kernel estimate remains unchanged, but is included in Figure 7(b) for the sake of comparison.

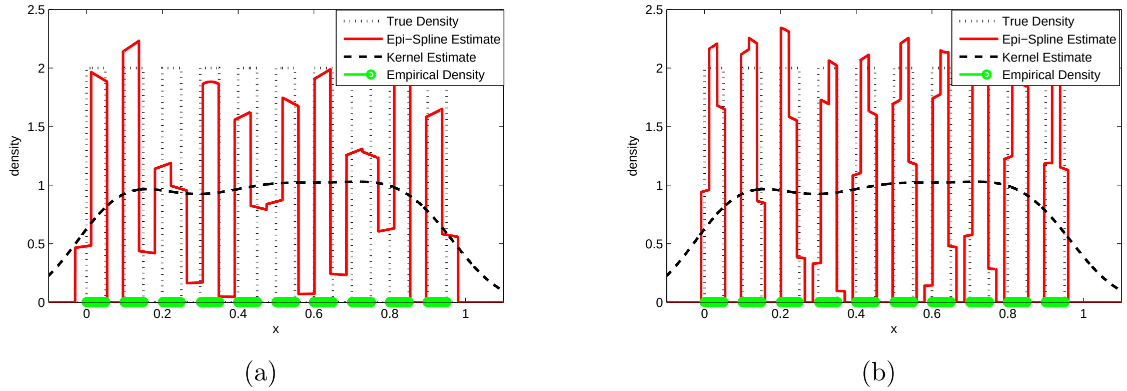
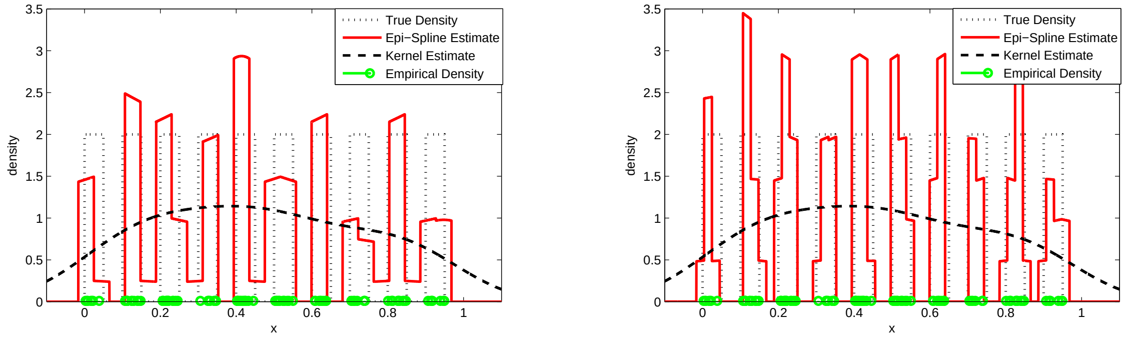


Figure 7: Uniform Mixture Example: Sample size 1000, and $N = 50$ (a) and $N = 100$ (b).

We repeat the calculations for $\nu = 100$ and show the results in Figure 8. Again, the kernel estimate (dashed black line) is unable to capture the discontinuities in the mixture density. The exponential epi-spline for $N = 50$ (Figure 8(a)) and $N = 100$ (Figure 8(b)) qualitatively reflect the mixture density to a significant degree.



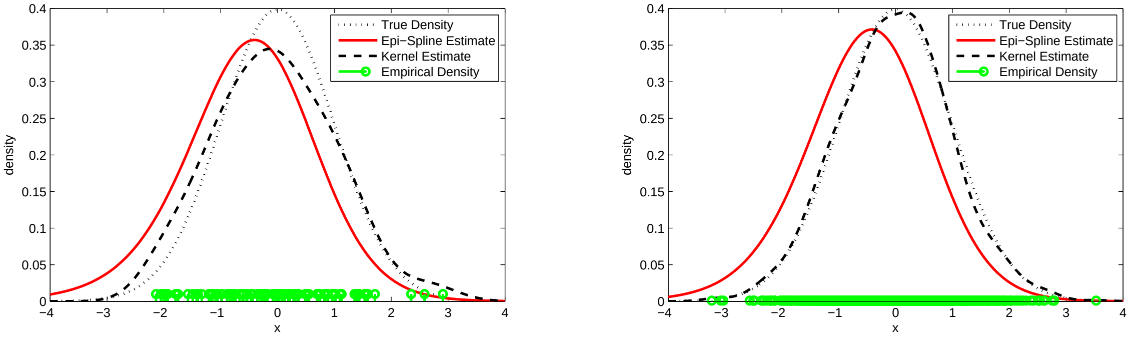
(a)

(b)

Figure 8: Uniform Mixture Example: Sample size 100, and $N = 50$ (a) and $N = 100$ (b).

5.4 Incorrect Soft Information

As given by Theorem 4, optimal solutions of $P_{p,m}^{\nu}$ tend to a point in the Kullback-Leibler projection of the true density h^0 relative to the set constructed by the soft information as the sample size grows. Consequently, in the presence of *incorrect* soft information that excludes h^0, we achieve the density 'nearest' to h^0 within the set of densities satisfying the (incorrect) soft information. We illustrate this situation by considering a standard normal density and its exponential epi-splines estimates based on $N = 10$. We adopt soft information about continuous differentiability and log-concavity. In addition, we impose the incorrect constraint that the expected value must be no larger than -0.5. Figure 9(a) shows the resulting exponential epi-spline estimate (solid red line) and the kernel estimate (dashed black line) for $\nu = 100$. Figure 9(b) displays the corresponding results for $\nu = 1000$. We observe that while the kernel estimator benefits from the larger sample size and obtains a nearly perfect estimate for $\nu = 1000$, the unfortunate expectation constraint on the exponential epi-spline prevents it from approaching the true density. However, we obtain a 'normal-looking' density with a shifted mean of -0.5.



(a)

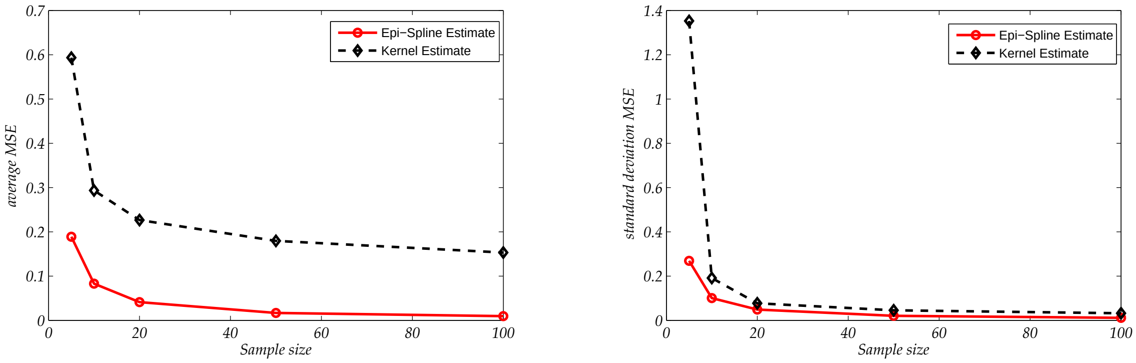
(b)

Figure 9: Normal Example: Estimates for $\nu = 100$ (a) and $\nu = 1000$ (b) with incorrect constraint $\int_{m_0}^{m_N} x h^{\nu}(x) dx \leq -0.5$.

5.5 Average Performance

We end the section by presenting a series of aggregate results using the exponential, normal, uniform, and pareto densities as well as a range of sample sizes. For each density and sample size, we carry out 104 meta-replications and compute average and standard deviation of the resulting MSE for both an exponential epi-spline estimate and a kernel estimate. We use $N = 20$ and soft information about a continuously differentiable density, with additional soft information implemented depending on the density.

We first consider the exponential density with parameter $\lambda = 0.5$. Using nonnegativity and nonincreasing soft information, we obtain for a range of sample sizes the average and standard deviation MSE results of Figure 10(a) and Figure 10(b), respectively. We find that the exponential epi-spline estimates result in substantially smaller MSE, on average, compare to those of the kernel estimate. However, the rate of convergence of the MSE appears to be the same for the two estimators.



(a)

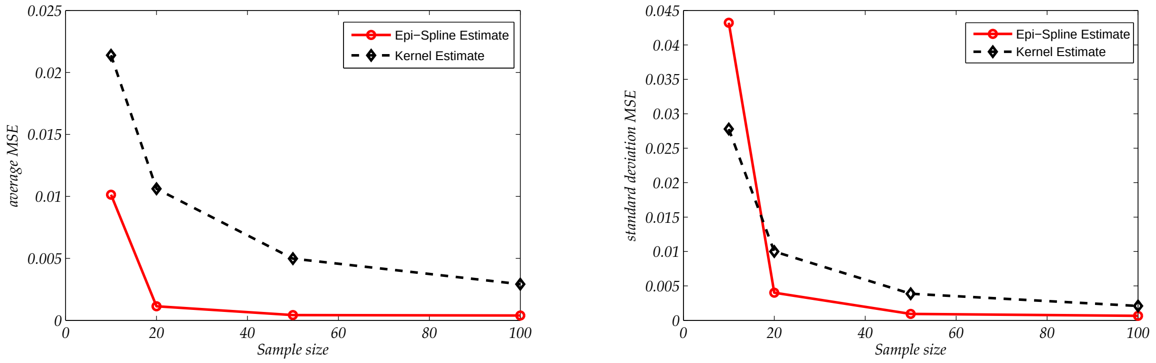
(b)

Figure 10: Exponential Example: Average (a) and standard deviation (b) of MSE for exponential epi-spline and kernel estimators for a range of sample sizes.

We second consider a normal density with zero mean and standard deviation of two. We compute exponential epi-splines assuming log-concavity and first and second moments being within 20% of their correct values. Figure 11 gives the corresponding average and standard deviation of the MSE for a range of sample sizes. Again, we see that the exponential epi-splines estimates result in smaller MSE, on average. However, the advantage vanishes as the sample size grows.

Third, we consider a uniform density on $[-1, 1]$. We compute exponential epi-spline estimates with soft information about the support bounds as well as log-concavity. The kernel estimates also make use of the information about support bounds. Again, the exponential epi-spline estimates result in smaller average MSE for all sample sizes examined; see Figure 12. In this case, the ratio of average MSE from the exponential epi-splines to that from the kernel estimates decreases substantially as the sample size increase up to 100.

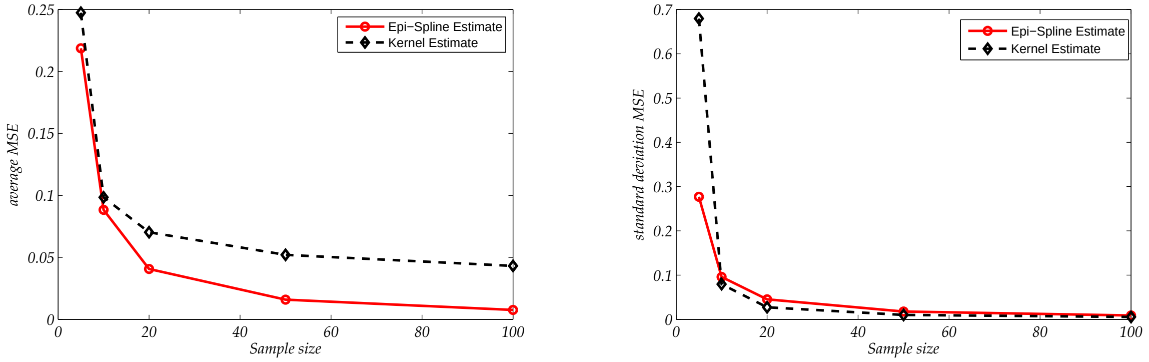
Fourth, we consider the Pareto density with shape parameter $k = 3$ and location parameter $\theta = 1$. We again assume log-concavity and the correct support bounds. Figure 13 shows the average and standard deviations of the MSE, with average MSE for the exponential epi-spline estimates substantially smaller than those of the kernel estimates. Across the examples, the standard deviations for the two methods are comparable.



(a)

(b)

Figure 11: Normal Example: Average (a) and standard deviation (b) of MSE for exponential epi-spline and kernel estimators for a range of sample sizes.



(a)

(b)

Figure 12: Uniform Example: Average (a) and standard deviation (b) of MSE for exponential epi-spline and kernel estimators for a range of sample sizes.

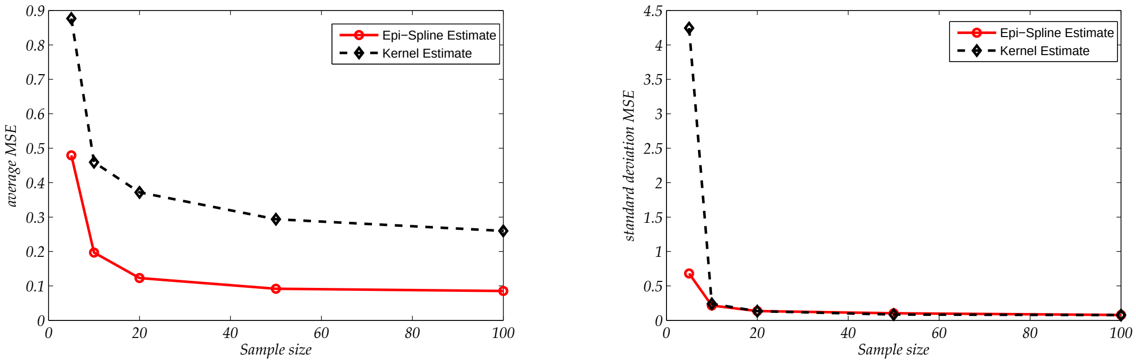
Acknowledgements. This material is based upon work supported in part by the U.S. Army Research Laboratory and the U.S. Army Research Office under grant numbers 00101-80683, W911NF-10-1-0246 and W911NF-12-1-0273. The authors thank Dr. Ritesh Sood for carrying out a part of the numerical tests as well as Prof. N. Sukumar for invigorating discussions.

References

[1] H. Attouch, R. Lucchetti, and R. Wets. The topology of the ρ-Hausdorff distance. *Annali di Matematica pura ed applicata*, CLX:303–320, 1991.

[2] F. Balabdaoui and J. A. Wellner. Estimation of a k-monotone density: limit distribution theory and the spline connection. *Annals of Statistics*, 35(6), 2007.

[3] F. Balabdaoui and J. A. Wellner. Estimation of a k-monotone density: characterizations, consistency and minimax lower bounds. *Statistica Neerlandica*, 64(1), 2010.



(a) (b)

Figure 13: Pareto Example: Average (a) and standard deviation (b) of MSE for exponential epi-spline and kernel estimators for a range of sample sizes.

[4] P. Bühlmann and S. van de Geer. *Statistics for High-Dimensional Data, Methods, Theory and Applications*. Springer, 2011.

[5] R. J. Carroll, A. Delaigle, and P. Hall. Testing and estimating shape-constrained nonparametric density and regression in the presence of measurement error. *Journal of the American Statistical Association*, 106(493):191–202, 2011.

[6] M. Casey and R. J-B Wets. Density estimation: exploiting non-data information. University of California, Davis, 2010.

[7] G. M. de Montricher, R.A. Tapia, and J.R. Thompson. Nonparametric maximum likelihood estimation of probability densities by penalty function method. *Annals of Statistics*, 3:1329–1348, 1975.

[8] M. X. Dong and R. J-B Wets. Estimating density functions: a constrained maximum likelihood approach. *Journal of Nonparametric Statistics*, 12(4):549–595, 2007.

[9] J. Dupacova. Epi-consistency in restricted regression models - the case of a general convex fitting function. *Computational Statistics and Data Analysis*, 14:417–425, 1992.

[10] T. S. Ferguson. Bayesian density estimation by mixtures of normal distributions. In M. H. Rizvi et al., editor, *Recent Advances in Statistics*. Acadmic Press, 1983.

[11] F. Gao and J. A. Wellner. On the rate of convergence of the maximum likelihood estimator of a k-monotone density. *Science in China Series A: Mathematics*, 52(7), 2009.

[12] C. J. Geyer. On the asymptotics of constrained M-estimation. *The Annals of Statistics*, 22:1993–2010, 1994.

[13] I. J. Good and R. A. Gaskin. Nonparametric roughness penalties for probability densities. *Biometrika*, 58:255–277, 1971.

[14] P. Groenenboom, G. Jongbloed, and J.A. Wellner. Estimation of a convex function: characterizations and asymptotic theory. *Annals of Statistics*, 29, 2001.

[15] A.J. King and R.T. Rockafellar. Asymptotic theory for solution of genaralized M-estimation and stochastic programming. Technical Report WP-90-76, International Institute for Applied Systems Analysis, Laxenburg, Austria, 1990.

[16] V. K. Klonias. Consistency of two nonparametric maximum penalized likelihood estimators of the probability density function. *Annals of Statistics*, 10:811–824, 1982.

[17] R. Koenker and I. Mizera. Density estimation by total variation regularization. In *A Festschrift for Kjell Doksum*. World Scientific, Singapore, 2006.

[18] R. Koenker and I. Mizera. Primal and dual formulations relevant for the numerical estimation of a density function via regularization. In A. Pázman, J. Volaufová, and V. Witkovský, editors, *Proceedings of the Conference ProbStat '06*, volume 38. Tatra Mountain Mathematical Publications, 2008.

[19] R. Koenker and I Mizera. Quasi-concave density estimation. *Annals of Statistics*, 38:2998–3027, 2010.

[20] T. Leonard. Density estimation, stochastic processes and prior information. *Journal of the Royal Statistical Society*, B40:113–146, 1978.

[21] M. Meyer. Constrained penalized splines. *Canadian Journal of Satistics*, 40:190–206, 2012.

[22] M. Meyer. Nonparametric estimation of a smooth density with shape restrictions. *Statistica Sinica*, 22:681–701, 2012.

[23] R. Pasupathy. On choosing parameters in retrospective-approximation algorithms for stochastic root finding and simulation optimization. *Operations Research*, 58:889–901, 2010.

[24] G. H. Pflug and R. J-B Wets. Shape restricted nonparametric regression with overall noisy measurements. *Journal of Nonparametric Statistics*, to appear, 2013.

[25] B. L. S. Prakasa Rao. *Nonparametric Functional Estimation.* Academic Press, New York, NY, 1983.

[26] L. Reboul. Estimation of a function under shape restrictions. applications to reliability. *The Annals of Statistics*, 33:1330–1356, 2005.

[27] P. Revesz. Density estimation. In P. Krishnaiah and P.K. Sen, editors, *Handbook of Statistics*, pages 531–549. North Holland, Amsterdam, Netherlands, 1984.

[28] T. Robertson, F. Wright, and R. Dykstra. *Order Restricted Statistical Inference.* Wiley, New York, NY, 1988.

[29] R. T. Rockafellar and R. J-B. Wets. *Variational analysis.* Springer, New York, NY, 1998.

[30] J. O. Royset and R. J-B Wets. Epi-splines and exponential epi-splines: Pliable approximation tools. Naval Postgraduate School, Monterey, California, 2013.

[31] J.O. Royset, N. Sukumar, and R. J-B Wets. Uncertainty quantification using exponential epi-splines. In *Proceedings of the International Conference on Structural Safety and Reliability*, 2013.

[32] T. Rychlik. Error reduction in density estimation under shape restrictions. *The Canadian Journal of Statistics*, 27:607–622, 1999.

[33] F. J. Samaniego and D. M. Reneau. Towards a reconciliation ofthe bayesian and frequentist approaches to point estimation. *Journal of the American Statistical Association*, 89:947–957, 1994.

[34] D. W. Scott. *Multivariate Density Estimation*. Wiley, New York, NY, 1983.

[35] A. Shapiro, D. Dentcheva, and A. Ruszczynski. *Lectures on Stochastic Programming: Modeling and Theory*. SIAM, Philadelpha, PA, 2009.

[36] M. Silvapulle and P. Sen. *Constrained Statistical Inference*. Wiley Series in Probability and Statistics. Wiley, New York, NY, 2005.

[37] B. W. Silverman. On the estimation of a probability density function by the maximum penalized likelihood method. *Annals of Statistics*, 10:795–810, 1982.

[38] D. Singham, J.O. Royset, and R. J-B Wets. Density estimation of simulation output using exponential epi-splines. Naval Postgraduate School, Monterey, CA, 2013.

[39] R. Sood and R. Wets. Information fusion. http://www.math.ucdavis.edu/ prop01, 2011.

[40] J. R. Thompson and R. A. Tapia. *Nonparametric Function Estimation, Modeling, and Simulation*. SIAM Publishers, Philadelphia, PA, 1990.

[41] S. Van de Geer. A new approach to least squares estimation. Center for Mathematics and Computer Science, Amsterdam, 1987.

[42] G. Wahba. Data-based optimal smoothing of orthogonal series density estimates. *Annals of Statistics*, 9, 1981.

[43] G. Wahba. *Spline Models for Observation Data*. SIAM Publishers, Philadelpha, PA, 1990.

[44] J. Wang. Asymptotics of least-squares estimators for constrained nonlinear regression. *Annals of Statistics*, 24(3):1316–1326, 1996.

[45] R. J-B Wets. Constrained estimation: consistency and asymptotics. *Applied Stochastic Models and Data Analysis*, 7:17–32, 1991.