Fusion of Hard and Soft Information in Nonparametric Density Estimation*

Johannes O. Royset

Department of Operations Research
Naval Postgraduate School
joroyset@nps.edu

Roger J-B Wets

Department of Mathematics
University of California, Davis
rjbwets@ucdavis.edu

Abstract. This article discusses univariate density estimation in situations when the sample (hard information) is supplemented by "soft" information about the random phenomenon. These situations arise broadly in operations research and management science where practical and computational reasons severely limit the sample size, but problem structure and past experiences could be brought in. In particular, density estimation is needed for generation of input densities to simulation and stochastic optimization models, in analysis of simulation output, and when instantiating probability models. We adopt a constrained maximum likelihood estimator that incorporates any, possibly random, soft information through an arbitrary collection of constraints. We illustrate the breadth of possibilities by discussing soft information about shape, support, continuity, smoothness, slope, location of modes, symmetry, density values, neighborhood of known density, moments, and distribution functions. The maximization takes place over spaces of extended real-valued semicontinuous functions and therefore allows us to consider essentially any conceivable density as well as convenient exponential transformations. The infinite dimensionality of the optimization problem is overcome by approximating splines tailored to these spaces. To facilitate the treatment of small samples, the construction of these splines is decoupled from the sample. We discuss existence and uniqueness of the estimator, examine consistency under increasing hard and soft information, and give rates of convergence. Numerical examples illustrate the value of soft information, the ability to generate a family of diverse densities, and the effect of misspecification of soft information.

Keywords: density estimation, data analytics, data fusion, epi-splines

AMS Classification: 62G07, 62G20, 62G35

Date: June 10, 2015

1 Introduction

It is recognized that statistical estimates can be improved greatly by including contextual information to supplement the information derived from data. We refer to the contextual information as *soft*

*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 the referees for insightful comments, Drs. R. Sood and D. Singham for carrying out a part of the numerical tests, and Prof. N. Sukumar for invigorating discussions.

Report Documentation Page

Form Approved
OMB No. 0704-0188

1. REPORT DATE	2. REPORT TYPE	3. DATES COVERED
10 JUN 2015		**00-00-2015 to 00-00-2015**

4. TITLE AND SUBTITLE	5a. CONTRACT NUMBER
Fusion of Hard and Soft Information in Nonparametric Density Estimation	5b. GRANT NUMBER
	5c. PROGRAM ELEMENT NUMBER
6. AUTHOR(S)	5d. PROJECT NUMBER
	5e. TASK NUMBER
	5f. WORK UNIT NUMBER
7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES) **Naval Postgraduate School,Department of Operations Research,Monterey,CA,93943**	8. PERFORMING ORGANIZATION REPORT NUMBER
9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)	10. SPONSOR/MONITOR'S ACRONYM(S)
	11. SPONSOR/MONITOR'S REPORT NUMBER(S)

12. DISTRIBUTION/AVAILABILITY STATEMENT
Approved for public release; distribution unlimited

13. SUPPLEMENTARY NOTES

14. ABSTRACT

This article discusses univariate density estimation in situations when the sample (hard information) is supplemented by ???soft??? information about the random phenomenon. These situations arise broadly in operations research and management science where practical and computational reasons severely limit the sample size, but problem structure and past experiences could be brought in. In particular, density estimation is needed for generation of input densities to simulation and stochastic optimization models, in analysis of simulation output, and when instantiating probability models. We adopt a constrained maximum likelihood estimator that incorporates any, possibly random, soft information through an arbitrary collection of constraints. We illustrate the breadth of possibilities by discussing soft information about shape, support, continuity, smoothness, slope, location of modes symmetry, density values, neighborhood of known density, moments, and distribution functions. The maximization takes place over spaces of extended real-valued semicontinuous functions and therefore allows us to consider essentially any conceivable density as well as convenient exponential transformations. The infinite dimensionality of the optimization problem is overcome by approximating splines tailored to these spaces. To facilitate the treatment of small samples, the construction of these splines is decoupled from the sample. We discuss existence and uniqueness of the estimator, examine consistency under increasing hard and soft information, and give rates of convergence. Numerical examples illustrate the value of soft information, the ability to generate a family of diverse densities, and the effect of misspecification of soft information.

15. SUBJECT TERMS

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT **unclassified**	b. ABSTRACT **unclassified**	c. THIS PAGE **unclassified**	**Same as Report (SAR)**	**39**	

information, in contrast to *hard information* derived from observations (data). In this article, we consider univariate probability density estimation exploiting, in concert, hard and soft information. Although our development, theoretical and numerical, makes no distinction based on sample size, not surprisingly, it is when the sample size is small that this fusion of hard and soft information plays a crucial role in producing quality estimates. We limit the scope to densities of random variables with distributions that are absolutely continuous with respect to the Lebesgue measure on a bounded interval.

The need for estimating probability density functions is prevalent across operations research and management science. For example, an essential step in simulation analysis and stochastic optimization is the generation of probability densities for input random variables; see for example [11, 27, 5]. Density estimation is also needed when populating probability models and when analyzing simulation output beyond their typical first and second moments. In all these situations, however, the sample available is typically extremely small due to practical and computational limitations. One is usually forced to restrict the attention to parametric families of densities. In this paper, we provide the theoretical foundations of an alternative approach that brings in soft information about problem structure and past experiences to obtain reasonable *nonparametric* density estimates even for very small sample sizes. The approach has been successfully applied in the context of simulation output analysis [65], uncertainty quantification [58], as well as estimation of errors in forecasts for commodity prices [74] and electricity demand [26]; see also [55].

A natural and widely studied approach to density estimation is to adopt an M-estimator with additional constraints to account for soft information. We continue this tradition by defining an estimator that is an optimal solution of a *constrained* maximum likelihood problem. An appealing property of such estimators is that for any sample size, an estimate is the best possible within the class of allowable functions according to the given criterion (likelihood).

We trace the consideration of soft information in terms of shape constraints at least back to [31, 32]. More recent studies of univariate log-concave densities include [35, 37, 71, 48, 23, 2], with computational comparisons in [60]; see also the review [72] and, in the case of multivariate densities, e.g., [12, 13]. Convexity and monotonicity restrictions are examined in [34, 46] and monotonicity, monotonicity and convexity, U-shape, as well as unimodality with known mode are studied in [47, 46]. Unimodal functions are also covered in [54, 36], with the former covering U-shape as well. Monotone, convex, and log-concave densities are dealt with in [6]. Studies of k-monotone densities include [3, 28, 4]. Densities given as monotone transformations of convex functions are examined in [61]. Convex formulation of a collection of shape restrictions is discussed in [49, 50]. We refer to the recent dissertation [22] and the discussion in [13] for a more comprehensive review and to [44] for the related context of shape-restricted regression.

Although these studies address important cases, there is no overarching framework that allows for a comprehensive description of soft information formulated by a large variety of constraints. Initial work in this direction is found in [73], which deals with parametric nonlinear least-squares regression subject to a finite number of smooth equality and inequality constraints. That paper examines the asymptotics of the least-squares estimator using the convergence theory of constrained optimization, specifically epi-convergence. In the context of constrained maximum likelihood estimation, [21] establishes consistency of an estimator through a functional law of large numbers and epi-convergence. The latter work is an immediate forerunner to the present paper.

Having adopted a nonparametric constrained maximum likelihood framework, we face technical challenges along two axes. First, one needs to deal with constrained optimization problems. Of course,

in principle, constraints can be handled through penalties and regularizations; see for example [30, 16, 43, 39, 64, 67] and more recently [25, 69, 40, 41, 45, 42, 7]. However, the equivalence and interpretations of such reformulations depends on the successful selection of multipliers and penalty parameters which is far from trivial in practice, especially in the case of multiple constraints. 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. Moreover, it becomes unclear in what sense, if any, an estimator is "best" when an otherwise natural criterion such as likelihood is mixed with nonzero penalty terms; see [21] for further discussion. It is also possible to devise specialized algorithms such as the iterative convex minorant algorithm [35, 37] to account for certain constraints or modify "unconstrained" estimators such as those based on kernels; [36] handles unimodality, [6] considers monotonicity, convexity, and log-concavity, and [15] aims to reduce the number of modes; see [75, 53] for computational tools. Again, it is unclear in what sense, if any, such estimates are "best" in the case of finite samples. Moreover, it is challenging to generalize these approaches to handle other types of soft information. We direct the reader to [68] and references therein for treatments of kernel estimators including a discussion of optimality.

The second challenge with a nonparametric constrained maximum likelihood framework is the infinite-dimensionality of the resulting optimization problem. Naturally, there is a computational need to consider families of approximating densities characterized by a finite number of parameters. The method of sieves [33, 29, 10] provides a framework for constructing, typically, finite-dimensional approximating subsets that are gradually refined as the sample size grows and that in the limit is dense in a function space of interest. However, difficulties arise from three directions. First, with our focus on small sample sizes, the linkage between sample size and sieves becomes untenable. Second, in order to allow for the possibility of discontinuous densities and exponential transformations, we choose as underlying space the extended real-valued lower or upper semicontinous functions, but neither is a linear space. Consequently, the mathematically inbred tendency to obtain a finite-dimensional approximation by relying on a well-chosen finite basis is problematic; see for example [18, 45] for such an approach based on splines. Third, despite progress towards handling shape restrictions on sieves (see for example [20, 19, 17, 49, 50]), there is no straightforward way of handling a comprehensive set of soft information.

In this paper, as in [21], we consider an arbitrarily constrained maximum likelihood estimator for densities. We appear to be the first to consider such general constraints (soft information) in the context of nonparametric density estimation. The soft information might even be random, i.e., the soft information may not be known a priori but is realized with the sample. We give concrete formulations of the constrained maximum likelihood problem in the case of soft information about support bounds, semicontinuity, continuity, smoothness, slope information and related quantities, monotonicity, log-concavity, unimodality, location of modes, symmetry, bounds on density values, neighborhood of known density, bounds on moments, and bounds on cumulative distribution functions. We allow for *any combination* of these, and essentially any other constraint too.

We overcome the technical difficulty caused by constraints through the theory of constrained optimization, specifically epi-convergence, and therefore avoid tuning parameters related to penalties and regularization. With the exception of the preliminary work [21], this paper is the first to utilize epi-convergence to analyze constrained density estimators. We overcome the difficulty of infinite dimensionality through the use of a new class of splines, epi-splines [57], which are highly flexible, allow for discontinuities, and enable convenient exponential transformations. Here, for the first time, the theoretical foundations for using epi-splines in density estimation are laid out. In contrast to sieves,

epi-splines can be constructed independently of the sample and therefore handles small sample sizes naturally. The precursor [21] relies on a finite approximation of $\mathcal{L}^2$ by Fourier coefficients. In this paper, we consider the spaces of extended real-valued semicontinuous functions, exponential transformations, and epi-spline approximations.

The reliance on epi-convergence and epi-splines allow us to view the constrained maximum likelihood problem as an approximation of a limiting optimization problem involving the actual probability density, correct soft information, and the full space of semicontinuous functions; we reference [52] for a related study in the context of regression utilizing graphical convergence. Consequently, we not only approximate a certain function space or deal with finite sample size, but study the approximation of the whole estimation process as formulated by the limiting optimization problem. The approach facilitates the examination of families of estimators such as those that are near-optimal solutions of a constrained maximum likelihood problem.

Our primary motivation is to obtain reasonable estimates in situations with little hard information and we provide a consistency result as soft information is refined, quantify finite sample errors, and present a small computational study to motivate the estimator in that regard. Still, we also establish consistency and quantify asymptotic rates, as hard information is refined, under general constraints.

We focus exclusively on univariate densities that vanish beyond a compact interval of the real line. Although most of the results extend to the unbounded case and higher dimensions, technical issues will then become prominent and obscure the treatment of arbitrary random constraints and the supporting epi-spline approximations. Moreover, with a small sample, tail behavior can only come in via soft information, which is easily handled by our framework but omitted here for simplicity; a few experimental results can be found in [66].

The paper proceeds in §2 by defining the constrained maximum likelihood estimator, summarizing the underlying approximation theory, which is based on [57], and discussing existence and uniqueness. Section 3 exemplify the breadth of soft information that can be included and §4 provides consistency, asymptotics, and finite sample error results. A small collection of numerical examples are featured in §5. The paper is summarized in §6.

2 Exponential Epi-Spline Estimator

This section formulates a constrained maximum likelihood problem and presents a finite-dimensional approximation. We discuss existence, uniqueness, and computations. The section also includes the prerequisite approximation results.

2.1 Constrained Likelihood Maximization and Epi-Spline Approximations

We consider a random variable X, with $-\infty < l \leq X \leq u < \infty$ a.s. and a distribution that is absolutely continuous with respect to the Lebesgue measure, an iid sample $X^1, X^2, \ldots, X^n$, and a possibly random set F^n that accounts for soft information about the density of X; see §3 and §5 for concrete examples. Realizations of F^n are subsets of allowable functions on $[l, u]$. The randomly constrained maximum likelihood problem takes the form:

$$(P^n): \quad f^n \in \operatorname{argmax} \prod_{i=1}^{n} e^{-f(X^i)} \;\text{ such that }\; f \in F^n \subset \mathcal{F}, \int_l^u e^{-f(x)} dx = 1,$$

where "argmax" denotes the set of optimal solutions and $\mathcal{F}$ is the space of extended real-valued lower semicontinuous (lsc) functions $f : [l, u] \to \overline{I\!R} = I\!R \cup \{-\infty, \infty\}$ excluding $f \equiv \infty$, or alternatively the space of extended real-valued upper semicontinuous (usc) functions $f : [l, u] \to \overline{I\!R}$ now excluding $f \equiv -\infty$. The density estimator then takes the form

$$e^{-f^n(\cdot)}, \text{ with } f^n \text{ a solution of } (P^n).$$

These spaces of functions under considerations are large enough to capture essentially all densities on $[l, u]$ including, of course, those with discontinuities. Moreover, the ability to handle $f(x) = \infty$ ensures that $\exp(-f(x)) = 0$ and, therefore, the exponential transformation in (P^n) does not eliminate the possibility of vanishing densities at points in $[l, u]$. If $f(x) = -\infty$, then $\exp(-f(x)) = \infty$, which obviously can at most occur for x in a set of (Lebesgue) measure zero if $\exp(-f)$ is a density. The exponential transformation (see [30, 16] for early use of such transformations and [61] for a broader treatment) automatically ensures that $\exp(-f(\cdot))$ is nonnegative and explicit constraints for that purpose are redundant. In addition, some types of soft information are more easily formulated for f than for $h = \exp(-f(\cdot))$; see examples in §3. Since we approximate lsc (usc) functions by the piecewise polynomial epi-splines, to be discussed shortly, further motivation for the exponential transformation is provided by the fact that many of the common densities are indeed exponential transformations of polynomials. We observe that the lsc and usc functions are measurable and consequently the integral in (P^n) is well-defined, but possibly infinite.

It is clear that a solution f^n of (P^n) generates a density $\exp(-f^n(\cdot))$ that, regardless of the sample size, possesses the properties embedded in F^n, which presumably are the properties of the actual density (see Theorems 4.2 and 4.4 and Section 5.4 for a discussion of misspecification). Moreover, it will be a best possible density in terms of the maximum likelihood criterion and the set of allowable densities.

In view of the above formulation and discussion, we are unable to build on the extensive literature on sieves and follow a different path. We instead introduce a new class of functions called *exponential epi-splines* from which we can construct approximations *independently* of the sample. They allow us to substitute for the infinite-dimensional (P^n), a finite-dimensional problem, guaranteed to generate a solution that approximates, to any desired level of accuracy, a solution of (P^n).

We start by defining the central building block of our approximation framework; see [57] for details. 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* $m := \{m_k\}_{k=0}^N$, with $m_{k-1} < m_k$, $k = 1, 2, ..., N$, that partitions its domain $[m_0, m_N]$ in N open subintervals, where on each subinterval the basic epi-spline is a polynomial of degree p.

Our focus on small samples and the use of highly flexible candidate densities in (P^n) and its epi-spline-based approximations can easily lead to overfitting. This might give the impression that the mesh m will become an important tuning parameter. However, since tuning the mesh might be challenging and easily could have become the subject of arbitrary decisions, we take another approach. We recall that (P^n) is the actual problem of interest and the estimator is $\exp(-f^n)$, with f^n being one of its solutions. Since such a solution is not directly available, our effort is directed towards obtaining an approximation through a "discretization" of the space $\mathcal{F}$. The mesh should therefore be selected fine enough to allow epi-splines to adequately approximate the underlying space $\mathcal{F}$ of lsc (usc) functions, or possibly subsets of continuous or continuously differentiable functions, if such restrictions are warranted. With this perspective, it becomes F^n that needs to be appropriately defined to ensure that (P^n) has reasonable solutions that avoid overfitting, among other things. Since we allow for arbitrary constraints, there are usually several ways soft information can be brought in to ensure reasonable solutions, which

leads to flexibility for the analyst; see §5 for examples. In most of the paper, we therefore assume that the mesh m is fixed and sufficiently fine.

Obviously, basic epi-splines are structurally related to polynomial splines, widely used in engineering and statistical applications [70, 18, 45], as both are piecewise polynomial functions. However, basic epi-splines are more flexible, with continuity not required at mesh points, and they can approximate any extended real-valued semicontinuous function (see Theorem 2.4 below). The formal definition is stated next.

2.1 Definition (basic epi-spline and associated mesh). *A* (basic) epi-spline $s : [m_0, m_N] \subset \mathbb{R} \to \mathbb{R}$ *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 $p \in \mathbb{N}_0$ *if on each subinterval* (m_{k-1}, m_k) *for* $k = 1, \dots, N$, s *is polynomial of degree* p.

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

Exponential transformations of epi-splines result in exponential epi-splines:

2.2 Definition (basic exponential epi-spline). *The family of* (basic) exponential epi-splines *of order* $p \in \mathbb{N}_0$ *with mesh* $m = \{m_k\}_{k=0}^N$, *denoted by* x-spl$^p(m)$, *consists of functions* $h : [m_0, m_N] \to \mathbb{R}$ *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. The approximation of (P^n), relying on (exponential) epi-splines then takes the following form (after the customary switch to log-likelihood):

$$(P^n_{p,m}): \quad s^n \in \operatorname{argmin} \frac{1}{n}\sum_{i=1}^n s(X^i) \ \text{ such that } \ s \in S^n \subset \text{e-spl}^p(m), \int_{m_0}^{m_N} e^{-s(x)}dx = 1,$$

where S^n is the formulation and possibly approximation of soft information in terms of epi-splines. In this paper, we therefore examine

$$h^n := e^{-s^n(\cdot)}, \text{ with } s^n \text{ a solution of } (P^n_{p,m}),$$

which is our approximation of $\exp(-f^n)$. We refer to h^n as the exponential epi-spline estimator. Throughout the paper we make the assumption that the support $[l, u]$ of the true density is a subset of $[m_0, m_N]$.

It is 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[†]. Consequently, $(P^n_{p,m})$ is equivalent to a finite-dimensional optimization problem, usualy easily solved by standard algorithms. The next subsection provides the justification for approximating (P^n) by $(P^n_{p,m})$. We note that this approximation is carried out for computational reasons, as a means to overcome the infinite dimensionality of (P^n). The hard and soft information are considered fixed.

[†]There are N subintervals (m_{k-1}, m_k) each with a polynomial of degree p. This gives $N(p+1)$ parameters. In addition, there are $N + 1$ mesh points on which an epi-spline is freely defined (unless continuity is imposed) as motivated in §2.3, which leads to an additional $N + 1$ parameters. We note that the mesh m is fixed.

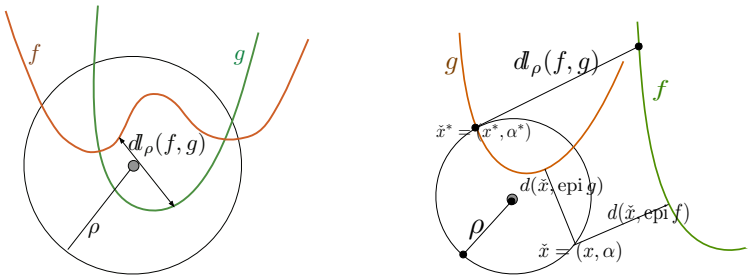


Figure 1: Examples of $d\!\!l_\rho(f,g)$ for epi f and epi g with different overlaps

2.2 Approximation Results

Approximations of extended real-valued semicontinuous functions by epi-splines rely on the refinement of the mesh as made precise in the next definition. This subsection is based on [57].

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

It is clear from classical spline theory that continuous functions can be approximated by polynomial splines. We need to go beyond continuous functions to extended real-valued semicontinuous functions. We rely on the *epi-topology* and *hypo-topology* (sometimes called the Attouch-Wets topologies), which are reviewed here for completeness; see [56, §7.I] for details. For any $l < u \in \mathbb{R}$, we denote by lsc-fcns$\big([l,u]\big)$ the set of all lsc functions $f : [l,u] \to \overline{\mathbb{R}}$ excluding $f \equiv \infty$. 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 $d\!\!l_\rho(f,g) := \max_{\|\check{x}\|\le\rho} |d(\check{x}, \operatorname{epi} f) - d(\check{x}, \operatorname{epi} g)|$ and $d(x,S) := \inf_{y\in S}\|x-y\|$ for $S \subset \mathbb{R}^2$, with $\|z\| := (\sum_i z_i^2)^{1/2}$ and $\operatorname{epi} f := \{\check{x} = (x,\alpha) \in \mathbb{R}^2 \mid f(x) \le \alpha\}$ 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 [56, 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 [56, §7.B], which by [56, Theorem 7.58] takes place if and only if $d\!\!l(f^\nu, f) \to 0$.

When dealing with usc functions, usc-fcns$\big([l,u]\big)$, now excluding the function $\equiv -\infty$, after observing that hypograph of a function f, hypo $f := \big\{(x,\alpha) \,\big|\, f(x) \ge \alpha\big\}$ 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 [56, Chapters 4 & 7] and [57] for broader treatments.

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 involves maximization, such as for the purpose of finding their modes. Similarly, the lsc densities is the natural class to consider in the context of subsequent minimization. We next state an approximation results for exponential

epi-splines.

2.4 Theorem (lsc and usc dense approximations [57]). *For any $p \in \mathbb{N}_0$ and $\{m^\nu\}_{\nu\in\mathbb{N}}$, an infinite refinement of $[l,u]$, under the hypo-topology,*

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

and under the epi-topology,

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

Consequently, for a sufficiently fine mesh and regardless of the order, exponential epi-splines provide arbitrarily accurate approximations of $\exp(-f(\cdot))$, $f \in \text{lsc-fcns}([l,u])$ and $f \in \text{usc-fcns}([l,u])$. In the remainder of the paper, we therefore mainly focus on $(P^n_{p,m})$ for fixed p and m.

2.3 Computations, Existence, and Uniqueness

We now turn to a convenient representation of epi-splines, also given in [57], which plays an essential role in computations and analysis. This leads to a finite-dimensional optimization problem for computing the estimator h^n, which we then analyze.

Every $s \in \text{e-spl}^p(m)$, with $m = \{m_k\}_{k=0}^N$, is uniquely represented by an *epi-spline parameter*

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

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

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

where $\langle z, z'\rangle := \sum_i z_i z'_i$ and $c_{p,m} : [m_0, m_N] \to \mathbb{R}^{(p+2)N+1}$ is defined by

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

with $x_k = x - m_{k-1}$ and $\vec{0}_k$ denoting the k-dimensional zero vector, $k \in \mathbb{N}$, and $\vec{0}_0$ 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 $c_{p,m} = (c_{\text{mesh}}, c_{\text{ess}})$.

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, in the discontinuous case the situation is more subtle. Since we consider functions in lsc-fcns($[l, u]$), which are defined on the whole $[l, u]$, their approximations should also be defined on the whole $[l, u]$. In addition, we would like to handle soft information such as bounds on the values of a density estimate at particular points, including at m. Hence, we find it most convenient to consider the value of epi-splines at the mesh independently and proceed with the more general framework involving r_{mesh}.

We note that 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. Specifically, if $h^\nu, h^0 \in \text{x-spl}^p(m)$, 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 [57]:

$$\begin{aligned} r^\nu \to r^0 &\Longleftrightarrow h^\nu \to h^0 \text{ uniformly on } [m_0, m_N] \\ &\Longrightarrow d\!\!l(-h^\nu, -h^0) \to 0 \Longleftrightarrow d\!\!l(s^\nu, s^0) \to 0. \end{aligned}$$

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 does not distinguish between a function and its usc regularization (see Proposition 7.4 in [56]), uniform convergence cannot generally be implied from hypo-convergence, even for exponential epi-splines.

We next deal with existence and uniqueness of the estimator $h^n = \exp(-s^n(\cdot))$ and consider a computational convenient equivalent form of $(P^n_{p,m})$ using the representation $s = \langle c_{p,m}(\cdot), r\rangle$. We consider a realization of $(P^n_{p,m})$ with $X^1, ..., X^n$ replaced by observed values $x^1, \ldots, x^n$, and S^n is now a realization of the random constraint set. Since the meaning is clear from the context, we denote realizations also by $(P^n_{p,m})$. We let $R^n \subset I\!R^{(p+2)N+1}$ be the set of epi-spline parameters corresponding to the set of epi-splines S^n, i.e.,

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

for example, if $S^n = \text{e-spl}^p(m)$, then $R^n = I\!R^{(p+2)N+1}$. When incorporating soft information, R^n and S^n become more restrictive as we see in §3. We let both the random set and its realizations be denoted by R^n. We also let

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

As stated next, $(P^n_{p,m})$ is equivalent to the finite-dimensional problem:

$$(\tilde{P}^n_{p,m}): \quad \min_{r \in R^n_I} \frac{1}{n} \sum_{i=1}^n \langle c_{p,m}(x^i), r\rangle.$$

Clearly, a realization $x^1, \ldots x^n$ and S^n generates a realization $(P^n_{p,m})$ as well as a corresponding realization $(\tilde{P}^n_{p,m})$.

2.5 Theorem (computing estimates). *Given p and $m = \{m_k\}_{k=0}^N$, for every corresponding realizations $(P_{p,m}^n)$ and $(\tilde{P}_{p,m}^n)$, one has*

(i) If $s^n \in$ e-spl$^p(m)$ is optimal for $(P_{p,m}^n)$, then there exists an $r^n \in \mathbb{R}^{(p+2)N+1}$ optimal for $(\tilde{P}_{p,m}^n)$ with $s^n = \langle c_{p,m}(\cdot), r^n \rangle$.

(ii) If $r^n \in \mathbb{R}^{(p+2)N+1}$ is optimal for $(\tilde{P}_{p,m}^n)$, then $s^n = \langle c_{p,m}(\cdot), r^n \rangle$ is optimal for $(P_{p,m}^n)$ and the exponential epi-spline estimator

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

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

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

While the objective function in $(\tilde{P}_{p,m}^n)$ is linear, R_I^n may be nonconvex. Hence, $(\tilde{P}_{p,m}^n)$ could possess local minimizers that are not globally optimal, increasing the complexity of solving the problem numerically. We see in §3 that R^n is often a polyhedron or at least convex. Hence, the main difficulty in $(\tilde{P}_{p,m}^n)$ is associated with the integral constraint. However, under broad conditions stated next, that constraint can be relaxed as utilized in other contexts earlier (see for example [34]). These conditions essentially imply that r can be improved whenever the corresponding function integrates to a number less than one.

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

The following Proposition 2.8 and §3 provide examples of loosely constrained realizations. We give an immediate consequence next.

2.7 Proposition *Suppose that a realization $(\tilde{P}_{p,m}^n)$ is loosely constrained. Then, that realization and the corresponding relaxed problem*

$$(rlxP_{p,m}^n): \quad \min_{r \in R^n} \frac{1}{n} \sum_{i=1}^n \langle c_{p,m}(x^i), r \rangle \text{ such that } \int_{m_0}^{m_N} e^{-\langle c_{p,m}(x), r \rangle} dx \leq 1$$

have identical sets of optimal solutions. Moreover, if R^n is convex, then $(rlxP_{p,m}^n)$ is a convex problem.

In Theorem 4.7 below we show that even beyond loosely constrained realizations, the consideration of $(rlxP_{p,m}^n)$ is justified. In view of the preceding discussion and results, it is 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^n is polyhedral, some further computational benefits may arise from utilizing the following reformulation, which is valid under additional assumptions; see §3 for examples.

The next result also gives a sufficient condition for a realization $(\tilde{P}^n_{p,m})$ to be loosely constrained. We use the notation $\vec{1}_{p,N}$ 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.

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

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

Moreover, if R^n is convex, then $(pnlP^n_{p,m})$ is a convex problem.

Proof: We consider corresponding realizations $(\tilde{P}^n_{p,m})$ and $(pnlP^n_{p,m})$ and let $r \in R^n$ satisfy $\int_{m_0}^{m_N} e^{-\langle c_{p,m}(x), r\rangle}dx = \gamma < 1$. For $r' = r + (\log\gamma)\vec{1}_{p,N}$,

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

Moreover, $\sum_{i=1}^{n}\langle c_{p,m}(x^i), r'-r\rangle = \sum_{i=1}^{n}\langle c_{p,m}(x^i), (\log\gamma)\vec{1}_{p,N}\rangle = n\log\gamma < 0$. Since $r' \in R^n$ by assumption, $(\tilde{P}^n_{p,m})$ is loosely constrained by Definition 2.6.

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

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

and let $\hat{r} \in R^n$ 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, $\hat{r} + (\log\gamma)\vec{1}_{p,N} \in R^n$ and, following the same argument as in (1),

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

Consequently, $\hat{r} + (\log\gamma)\vec{1}_{p,N}$ is feasible in $(\tilde{P}^n_{p,m})$. Suppose that r^n is optimal for $(\tilde{P}^n_{p,m})$. It follows that r^n also minimizes f^n on R^n_I because this problem deviates from $(\tilde{P}^n_{p,m})$ only by the constant one in the objective function. Using an argument similar to that of Lemma 2.3 in [34], we find that

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

where the inequality follows from the fact that r^n is optimal and $\hat{r} + (\log\gamma)\vec{1}_{p,N}$ is feasible in $(\tilde{P}^n_{p,m})$. Since $-\log\gamma - 1 + \gamma > 0$ for $\gamma \in (0,\infty), \gamma \neq 1$, we find that every $r \in R^n$ with $\int_{m_0}^{m_N} e^{-\langle c_{p,m}(x), r\rangle}dx \neq 1$

has $f^n(r) > f^n(r^n)$ and consequently cannot minimize f^n on R^n. The first conclusion then follows. Convexity of $(pnlP^n_{p,m})$ follows directly from the convexity of the integral term. □

In general, one cannot expect a unique optimal solution of a realization $(\tilde{P}^n_{p,m})$, 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 is not a subset of the sample realization $x^1, x^2, ..., x^n$. In fact, if the first $N+1$ components of the epi-spline parameter r are not constrained by R^n, 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 manifest, for example, in the case of continuous epi-splines used when dealing with densities known to be continuous.

2.9 Proposition *Suppose that corresponding realizations $(\tilde{P}^n_{p,m})$ and $(rlxP^n_{p,m})$ have R^n convex, $\{x^1, ..., x^n\} \cap m = \emptyset$, and satisfy the condition:*

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

Then, the following hold:

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

(ii) The realization $(pnlP^n_{p,m})$ 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 $\varphi : [m_0, m_N] \times I\!R^{(p+1)N} \to I\!R$ by

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

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

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

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 $\tilde{\psi} = (1/n)\sum_{i=1}^n \langle c_{\text{ess}}(x^i), \cdot\rangle + \psi(\cdot)$, 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 $(pnlP^n_{p,m})$, with optimal value v^*. Since $\{x^1, ..., x^n\} \cap m = \emptyset$, 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 $r_{\rm ess} = r'_{\rm ess}$, but then $r_{\rm mesh} = r'_{\rm mesh}$ by assumption and we contradict the hypothesis that $r \neq r'$.

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

Second, we focus on (i). Suppose that $r = (r_{\rm mesh}, r_{\rm ess}) \in R^n_I$ is optimal for the realization $(rlxP^n_{p,m})$. 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 $r' \in R^n$. Then by strict convexity of ψ, there exists a unique minimizer $r''_{\rm ess}$ of ψ on $\{r'''_{\rm ess} \in I\!R^{(p+1)N} \mid (r'''_{\rm mesh}, r'''_{\rm ess}) \in R^n \text{ for some } r'''_{\rm mesh} \in I\!R^{N+1}\}$. However, $r''_{\rm ess} = r_{\rm ess}$ because $\psi(r_{\rm ess}) = 1$. Another optimal solution for the realization $(rlxP^n_{p,m})$ would thus have essential epi-spline parameter identical to $r_{\rm ess}$. 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 $r' \in R^n$ 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 $(rlxP^n_{p,m})$ has the same set of optimal solutions as the problem

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

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

3 Soft Information

We implement soft information about the density under consideration in the estimation problem $(\tilde{P}^n_{p,m})$ through the set R^n, which can be any, possibly random, subset of $I\!R^{(p+2)N+1}$. It is observed empirically and also illustrated in §5 that soft information tends to improve density estimates. In this section, we give a *soft consistency* theorem that, in part, explains these observations. We also give examples of constraints for specific instances of soft information. We start, however, with a convenient result regarding the Kullback-Leibler divergence.

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 $\beta_2 = 0$. An immediate consequence of the definition of the divergence is the following result, which facilitates formulation of certain soft information as well as theoretical results below.

3.1 Proposition *Suppose h and e^{-s} are densities with $s = \langle c_{p,m}(\cdot), r \rangle \in$ 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* $s' = \langle c_{p,m}(\cdot), r' \rangle \in$ 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 theorem is a direct consequence of Proposition 3.1.

3.2 Theorem (soft consistency). *If the true density* $h^0 = e^{-\langle c_{p,m}(\cdot), r^0 \rangle}$, *with* $r^0 \in R^n$ *and there exists a* $\rho > 0$ *such that* $\|r - r'\| \leq \rho$ *for all* $r, r' \in R^n$, *then the estimate* $h^{n,\rho}$ *obtained from solving a realization* $(\tilde{P}^n_{p,m})$ *satisfies*

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

Moreover, for any fixed n, *if* $\rho \to 0$, *then* $h^{n,\rho} \to h^0$ *uniformly on* $[m_0, m_N]$.

An effective strategy for improving exponential epi-spline estimates is therefore to reduce the size of R^n, of course, without eliminating the true epi-spline parameter.

We next illustrate the wide range of soft information that is easily included within the exponential epi-spline framework[‡].

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 bounds of the support of the true density, respectively. If these are unknown, conservative values could be used as our ability to approximate extended real-valued functions does not rule out the possibility of vanishing densities on $[m_0, m_N]$, even for the exponentially transformed kind. In practice, m_0 and m_N 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 and soft information about possible locations of discontinuities, for example, could lead to other choices. Consequently, the mesh is selected essentially independently of the sample and one should simply focus on having a mesh that is sufficiently fine to allow epi-splines to approximate the underlying functions with a sufficient accuracy. Of course, some restrain on mesh refinement might be imposed by computational considerations. The number of decision variables in the resulting optimization problem grows linearly in N.

Semi-continuity, continuity and smoothness. It is straightforward to ensure usc, lsc, continuity, and various degrees of differentiability through linear constraints; see [58] for details. The inclusion of such constraint will keep a problem loosely constrained as the sufficient condition for being loosely constrained in Proposition 2.8 is satisfied.

We recall that the epi-spline parameter is of the forms

$$r = (s_0, s_1, ..., s_N, a_{1,0}, a_{1,1}, ..., a_{1,p}, a_{2,0}, a_{2,1}, ..., a_{2,p}, ..., a_{N,0}, a_{N,1}, ..., a_{N,p}),$$

where 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

[‡]Naturally, with the possibility of including incorrect soft information, there is a need for validation. Although important, we limit the discussion of this topic to Theorems 4.2, 4.4, and 4.7 as well as §5.4; see for example [63] and [9] for tests in related contexts.

(m_{k-1}, m_k), $k = 1, 2, ..., N$.

Slope information. The quantity $\int_{-\infty}^{\infty} h'(x)^2/h(x)dx$ 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," which also squares the denominator, results in a convex constraint. Specifically, if $h = e^{-\langle c_{p,m}(\cdot), r\rangle}$, then

$$\int_{-\infty}^{\infty} (h'(x)/h(x))^2 dx = \sum_{k=1}^{N} \int_{m_{k-1}}^{m_k} \left(\sum_{i=1}^{p} i a_{k,i} (x - m_{k-1})^{i-1} \right)^2 dx.$$

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

$$h''(x)/h(x) = -\sum_{i=2}^{p} i(i-1) a_{k,i} (x - m_{k-1})^{i-2} + \left(\sum_{i=1}^{p} i a_{k,i} (x - m_{k-1})^{i-1} \right)^2.$$

Upper and lower bounds on the first quantity result 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 $h'(x)/h(x)$ needs 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 2.8 is satisfied. We observe that constraints on $h'(x)/h(x)$ is an effective way of controlling the "tails" near m_0 and m_N.

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.

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, (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, the problem remains loosely constrained.

Unimodality and locations of modes. We implement soft information about unimodality of a continuous density by designating one mesh point $m_{k'}$ as the mode, and then constraining the density to be increasing and decreasing on $(m_{k'-1}, m_{k'})$ and $(m_{k'}, m_{k'+1})$, respectively, and nondecreasing on $[m_0, m_{k'})$ and nonincreasing on $(m_{k'}, m_N]$. Solving the resulting estimation problem gives a candidate density. The process is repeated for alternative mode locations m_k, $k = 0, 1, ..., N$, $k \neq k'$, and the density with the largest likelihood is retained as the estimate. The same result is obtained by solving a single augmented problem involving $N + 1$ binary variables. K-modality is achieved similarly by partitioning $[m_0, m_N]$ into K intervals, with each having a unimodal constraint. The process must be repeated for each partition of interest. To specify that certain m_k are modes is achieved by ensuring that the density is increasing and decreasing on (m_{k-1}, m_k) and (m_k, m_{k+1}), respectively.

Symmetry. We ensure symmetry by designating a point of symmetry m_k and then solving only for the upper half of the density on $[m_k, m_N]$, with trivial changes to the likelihood function and integral constraint. The process is repeated for each possible symmetry point. Again, auxiliary binary variables would obtain the same effect within one augumented formulation.

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

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

and

$$-\log h^{\min}(x) \geq s_k \geq -\log h^{\max}(x) \text{ for } x = m_k, k = 0, 1, ..., N.$$

While these constraints are linear, they do not satisfy the assumption of Proposition 2.8. However, if only the lower bound $h(x) \geq h^{\min}(x)$ is imposed, the resulting problem remains loosely constrained.

Kullback-Leibler divergence and the Bayesian paradigm. Proposition 3.1 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}, which for example could correspond to the posterior mean obtained through Bayes theorem. Then, a constraint

$$d_{KL}(h^{\text{ref}} || e^{-\langle c_{p,m}(\cdot), r\rangle}) \leq \varphi(n), \tag{4}$$

indeed ensures that the estimate h^n is within $\varphi(n)$ of h^{ref} as measured by the Kullback-Leibler divergence. If h^{ref} resulted from Bayes theorem, then these constraints allow for some flexibility to explore

densities near the one prescribed by a classical Bayesian approach. In view of Proposition 3.1, this constraint is linear in r and thus easily implementable. Here, $\varphi : \mathbb{N}_0 \to [0, \infty)$ is the *cognitive content* of the reference density h^{ref} and should satisfy $\varphi(0) = 0$, $\lim_{n\to\infty} \varphi(n) = \infty$, and be increasing since an increasing sample size should place gradually less emphasis on h^{ref}. Of course, if $\varphi(n) = 0$, then $(\tilde{P}^n_{m,p})$ simply returns h^{ref}, or a density that deviates at most on m. If $\varphi(n) = \infty$, then no information about the reference density is included. While technically not correct in the sense of classical Bayesian statistics, one can also view h^{ref} as a "prior" density and the resulting density h^n obtained from $(\tilde{P}^n_{m,p})$ as the "posterior" density. 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 §5.2 for an example.

Bounds on moments. Soft information may result in constraints on the j-th moment of the form $\mu_j^{\min} \leq \int_{m_0}^{m_N} x^j e^{-\langle c_{p,m}(x),r\rangle} dx \leq \mu_j^{\max}$, where $\mu_j^{\min}, \mu_j^{\max} \in \mathbb{R}$, $\mu_j^{\min} \leq \mu_j^{\max}$ are given constants. The right-most inequality results in a convex constraint in 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 $p^{\min}$ and the upper bound $p^{\max}$. This results in the two convex constraints $\int_{m_0}^{\gamma} e^{-\langle c_{p,m}(x),r\rangle} dx \leq p^{\max}$ and $\int_{\gamma}^{m_N} e^{-\langle c_{p,m}(x),r\rangle} dx \leq 1 - p^{\min}$.

4 Consistency, Asymptotics, and Error Bounds

Being concerned, from now on, with asymptotics, we again view $(P^n_{p,m})$ to be a random optimization problem, i.e.,

$$(P^n_{p,m}): \quad \min_{s\in S^n} \frac{1}{n}\sum_{i=1}^{n} s(X^i) \;\text{ such that }\; \int_{m_0}^{m_N} e^{-s(x)}\, dx = 1;$$

whose random elements are the variables $X^1, \ldots, X^n$ and the random set S^n; we still designate a solution by s^n which is now, itself, a random epi-spline. To achieve consistency, derive asymptotics and other results, we view $\{(P^n_{p,m})\}_{n=1}^{\infty}$, for given m and p, as a sequence of random optimization problems that under quite general assumptions converges in some sense to a limiting optimization problem, whose optimal solution recovers a *true* density $h^0 \in \text{x-spl}^p(m)$, as the sample size $n \to \infty$. We note that the restriction to x-spl$^p(m)$ for given m and p is justified by Theorem 2.4, but we also discuss the consideration of densities beyond this broad class; see Theorem 4.4 below.

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

4.1 Definition (Kullback-Leibler projection). *For any density h on $\mathbb{R}$ 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*

$$\mathcal{S}_{p,m}(h) := \operatorname*{argmin}_{s\in \text{e-spl}^p(m)} d_{KL}(h||e^{-s}) \text{ such that } \int_{m_0}^{m_N} e^{-s(x)} dx = 1. \tag{5}$$

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

We see that $\mathcal{S}_{p,m}(h)$ 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 $h = e^{-s} \in \text{x-spl}^p(m)$, $m = \{m_k\}_{k=0}^N$, then $s \in \mathcal{S}_{p,m}(h)$ and all $\tilde{s} \in \mathcal{S}_{p,m}(h)$ 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.

Suppose that $h^0 = e^{-s^0} \in \text{x-spl}^p(m), m = \{m_k\}_{k=0}^N$, is the density of a random variable X^0, which we aim to estimate. Then, for any $s \in \text{e-spl}^p(m)$, $d_{KL}(h^0||e^{-s}) = E\{\log h^0(X^0)\} + E\{s(X^0)\}$. Hence, there is a constant term (with respect to s) in the objective function of (5) that can be dropped and we reach the fact that every optimal solution of

$$(P^0_{p,m}): \quad \min_{s \in \text{e-spl}^p(m)} E\{s(X^0)\} \text{ such that } \int_{m_0}^{m_N} e^{-s(x)} dx = 1 \tag{6}$$

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 $(P^n_{p,m})$, we refer to $(P^0_{p,m})$ as the *true problem*. Intuitively, if $s^0 \in S^n$ and n is large, the problem $(P^n_{p,m})$ 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 $h^n = e^{-s^n}$ obtained from solving $(P^n_{p,m})$.

4.2 Theorem (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 $m = \{m_k\}_{k=0}^N$, $(P^n_{p,m})$ is derived by independent sampling from h^0, and $\{s^n\}_{n=1}^\infty$ is a sequence of optimal solutions of $(P^n_{p,m})$, with epi-spline parameters $\{r^n\}_{n=1}^\infty$.*

If $\lim R^n$ exists a.s.[§] and is deterministic, then every accumulation point r^∞ of $\{r^n\}_{n=1}^\infty$ satisfies

$$\langle c_{p,m}(\cdot), r^\infty \rangle \in \mathcal{S}^{S^\infty}_{p,m}(h^0) \text{ a.s.},$$

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

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

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

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

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

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

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

[§]Limits of sets are here taken in the sense of Painlevé-Kuratowski [56, §7.B] and the probability space is that induced by $\{(P^n_{p,m})\}_{n=1}^\infty$.

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

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

Since $R_I^\infty = \lim R_I^n$, 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 $\hat{r}^0 \in \mathbb{R}^{(p+2)N+1}$ is arbitrary. If $\hat{r}^0 \notin R_I^\infty$, then

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

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

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

Epi-convergence of $\langle (1/n) \sum_{i=1}^{n} c_{p,m}(X^i), \cdot \rangle + \delta_{R_I^n}$ to $\langle E\{c_{p,m}(X^0)\}, \cdot \rangle + \delta_{R_I^\infty}$ a.s. then follows by Proposition 7.2 in [56] and the first conclusions by Theorem 7.31 of [56] and the fact that $\hat{r} \in \arg\min_r \langle E\{c_{p,m}(X^0)\}, r\rangle + \delta_{R_I^\infty}$ if and only if $\langle c_{p,m}(\cdot), \hat{r}\rangle \in \mathcal{S}_{p,m}^{S^\infty}(h^0)$.

We next turn to the second part of the theorem. Since the additional assumption implies that R^n becomes arbitrary close to r^0 a.s., item (i) follows by a similar argument as above. Items (ii) and (iii) are conclusions from the discussion following Definition 4.1. □

The first part of Theorem 4.2 shows that regardless of the soft information, which may even *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 $\{(P_{p,m}^n)\}_{n=1}^\infty$ 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. We refer to [24, 14, 44] for related results on model misspecfication. The second part shows that if the true density is not excluded by the soft information, then $\{(P_{p,m}^n)\}_{n=1}^\infty$ 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 cannot 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 will 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.

4.3 Proposition *Suppose that $f^n : \mathbb{R} \to [0, \infty]$, $f^0 : \mathbb{R} \to [0, \infty]$ are Lebesgue integrable on every compact subset of $\mathbb{R}$ and $d(-f^n, -f^0) \to 0$. Then, for every compact set $X \subset \mathbb{R}$, $\int_X f^n(x)dx \to \int_X f^0(x)dx$.*

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

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

Proof: Following a similar argument as in Proposition 2.8, we see that the equality constraints in the problems defining $\mathcal{S}_{p,m}(h^n)$ and $\mathcal{S}_{p,m}(h^0)$ can be replaced by inequality. Consequently, every $s^n \in \mathcal{S}_{p,m}(h^n)$ is of the form $s^n = \langle c_{p,m}(\cdot), r^n\rangle$, with $r^n \in \operatorname{argmin}_r \psi^n(r) + \delta_I(r)$, where $\psi^n(r) := \langle \int_{m_0}^{m_N} c_{p,m}(x)h^n(x)dx, r\rangle$ 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 $s^0 \in \mathcal{S}_{p,m}(h^0)$ 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^n, but with h^n replaced by h^0. Clearly, $\psi^n + \delta_I$ and $\psi^0 + \delta_I$ are convex.

By Proposition 4.3, $\int_X h^n(x)dx \to \int_X h^0(x)dx$ 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 $k = 1, 2, ..., N$,

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

Hence, it follows by Proposition 7.2 and Theorem 7.53 in [56] that $\psi^n + \delta_I$ totally epi-converges to $\psi^0 + \delta_I$. The result then is a consequence of Corollary 7.55 in [56]. □

If we take the densities h^n in Theorem 4.4 to be exponential epi-splines, possibly defined on increasingly fine meshes, Theorem 2.4 shows that these densities indeed can be made to approximate with arbitrary accuracy any lsc or usc density h^0 with appropriate choice of the mesh. Consequently, the assumption of $d\!\!\!/(-h^n, -h^0) \to 0$ in Theorem 4.4 holds and, combined with Theorem 4.2, 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 as stated next.

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

$$r^n \to r^0 \Longrightarrow d_{KL}(h^0||h^n) \to 0 \Longleftrightarrow d_{KL}(h^n||h^0) \to 0 \Longrightarrow r^n_{\text{ess}} \to r^0_{\text{ess}}.$$

Proof: We let $r^n = (r^n_{\text{mesh}}, r^n_{\text{ess}})$ and $r^0 = (r^0_{\text{mesh}}, r^0_{\text{ess}})$. The implication $r^n \to r^0 \Longrightarrow d_{KL}(h^0||h^n) \to 0$ follows directly from Proposition 3.1.

To show that $d_{KL}(h^0||h^n) \to 0 \Longrightarrow r^n_{\text{ess}} \to r^0_{\text{ess}}$ 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 3.1, 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 convex 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 $r^n \in R$ and for sufficiently large n is arbitrarily close to W, we reach the desired conclusion.

We realize that $d_{KL}(h^n||h^0) \to 0 \Longrightarrow d_{KL}(h^0||h^n) \to 0$ by establishing that $r^n_{\text{ess}} \to r^0_{\text{ess}}$ whenever $d_{KL}(h^n||h^0) \to 0$ using a similar argument as above and then use Proposition 3.1.

We find that $d_{KL}(h^0||h^n) \to 0 \Longrightarrow d_{KL}(h^n||h^0) \to 0$ by invoking that $d_{KL}(h^0||h^n) \to 0 \Longrightarrow r^n_{\text{ess}} \to r^0_{\text{ess}}$ and Proposition 3.1. □

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 2.9, one cannot expect a unique estimator — a prerequisite for asymptotic normality — unless the scope is limited in this manner[¶]. 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[‖] $H(r_{\text{ess}}) := \int_{m_0}^{m_N} \rangle c_{\text{ess}}(x), c_{\text{ess}} \langle e^{-\langle c_{\text{ess}}(x), r_{\text{ess}}\rangle} dx$ be the Hessian of the function $\int_{m_0}^{m_N} e^{-\langle c_{\text{ess}}(x),\cdot\rangle} dx$ 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 2.9. 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 $\mu^0_j := \int_{-\infty}^{\infty} x^j h^0(x) dx$ 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^n a.s. for sufficiently large n. Moreover, we assume that the true density is an exponential epi-spline. Although this might at first appear restrictive, Theorem 2.4 shows that such densities can approximate to an arbitrary level of accuracy essentially any density.

4.6 Theorem (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 the (set) inner limit of the R^n a.s. If $(P^n_{p,m})$ is obtained by independent sampling from h^0 and $\{s^n\}_{n=1}^{\infty}$ is a sequence of optimal solutions of $(P^n_{p,m})$, with epi-spline parameters $\{r^n = (r^n_{\text{mesh}}, r^n_{\text{ess}})\}_{n=1}^{\infty}$, and $h^n = e^{-\langle c_{p,m}(\cdot), r^n\rangle}$ for all n, then the following hold:*

(i)

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

[¶]One could appeal to more sophisticated central limit theorems, such as those in [38], but additional conditions and machinery is required and would require us to stray too far from our main theme.

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

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

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

(iii) For $j \in \mathbb{N}$, and setting $w = \int_{m_0}^{m_N} x^j c_{\text{ess}}(x)e^{-\langle c_{p,m}(x), r^0\rangle}dx$, the moment estimator $\mu_j^n = \int_{m_0}^{m_N} x^j e^{-\langle c_{p,m}(x), r^n\rangle}dx$ satisfies

$$n^{1/2}(\mu_j^n - \mu_j^0) \to^d \mathcal{N}(0, \langle w, \Sigma(r^0_{\text{ess}})w\rangle).$$

Proof: The law of large number gives that the objective function in $(P^n_{p,m})$ converges uniformly on compact sets to that of $(P^0_{p,m})$ a.s. 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 the (set) inner limit of the R^n a.s. Consequently, since $(P^0_{p,m})$ does not involve a restriction S^n, the set of optimal solutions of $(P^n_{p,m})$ coincides with those of the relaxation of $(P^n_{p,m})$ with S^n replaced by e-spl$^p(m)$ for sufficiently large n. Let

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

where X^1, X^2, ..., X^n is the sample from h^0. We deduce from Propositions 2.8 and 2.9 that (P^n_{ess}) and the relaxation of $(P^n_{p,m})$ have unique optimal solutions a.s. and that they are equivalent in the sense that they generate the same essential epi-spline parameter. Consequently, for sufficiently large n, the optimal solution of (P^n_{ess}) is r^n_{ess} a.s.

Let X^0 be a random variable with density h^0 and

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

We deduce from Propositions 2.8 and 2.9 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^n_{ess}) are strictly convex and unconstrained a.s., their unique optimal solutions are equivalently characterized as the zeros of the objective function gradients. Since these gradients converge uniformly on $\mathbb{R}^{(p+1)N}$ a.s. by the law of large numbers, and the corresponding Hessians are identical and positive definite, item (i) follows directly from Theorem 4 of [51]. The next items follow by a direct application of a Delta Theorem; see, for example, §7.2.7 in [62]. □

Although Theorem 4.6 provides rates of convergence, it excludes the possibility of soft information in R^n influencing the estimates for large samples and, in addition, 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 $(rlxP^n_{p,m})$, now viewed as a random optimization problem, which for any $\varepsilon \geq 0$ are defined as

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

where the optimal value of $(rlxP^n_{p,m})$ is

$$V^n := \inf_{r\in R^n} \frac{1}{n}\sum_{i=1}^{n}\langle c_{p,m}(X^i), r\rangle \text{ such that } \int_{m_0}^{m_N} e^{-\langle c_{p,m}(x),r\rangle}dx \le 1.$$

The statements below deal with the difference between the true density h^0 and $h^n_\varepsilon := e^{-\langle c_{p,m}(\cdot),r^n_\varepsilon\rangle}$, with $r^n_\varepsilon \in \mathcal{R}^n_\varepsilon$ for $\varepsilon > 0$. In fact, a numerical method for solving a realization of $(rlxP^n_{p,m})$ generates an element of $\mathcal{R}^n_\varepsilon$, and consequently also an estimate h^n_ε, as such methods utilize finite precision and various tolerances. Also let $\rho I\!B := \{y \mid \|y\| \le \rho\}$ in any Euclidean space, $\Delta_{p,m} := \max_{l=0,1,...,p}|m|^l$, and $d(x,S) := \inf_{y\in S}\|x-y\|$ for $x \in I\!R^k, S \subset I\!R^k$.

4.7 Theorem (finite sample error). *Suppose that the true density $h^0 \in$ x-spl$^p(m)$, $m = \{m_k\}_{k=0}^N$, with epi-spline parameter r^0, and $(rlxP^n_{p,m})$ is derived by independent sampling from h^0, has a nonempty feasible set a.s., and R^n is closed and convex a.s. For any $\alpha > 0$, $\varepsilon > 0$, $\rho > \max\{-V^n, d(r^0, \mathcal{R}^n_0)\}$, and some $h^n_\varepsilon = e^{-\langle c_{p,m}(\cdot),r^n_\varepsilon\rangle}$, $r^n_\varepsilon \in \mathcal{R}^n_\varepsilon$,*

$$d(r^0, \mathcal{R}^n_\varepsilon) > K \text{ and } d_{KL}(h^0||h^n_\varepsilon) > \left\|\int_{m_0}^{m_N} c_{p,m}(x)h^0(x)dx\right\| K,$$

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

$$K := \left(1 + \frac{4\rho}{\varepsilon}\right)\left[\alpha(\rho + \|r^0\|)\sqrt{(p+1)N} + \left(1 + \Delta_{p,m}\sqrt{(p+2)N+1}\right)d(r^0, R^n)\right].$$

Proof: Let X^0 be a random variable with density h^0 and $X^1, X^2, ..., X^n$ be the sample that generates $(P^n_{p,m})$. 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, $(1/n)\sum_{i=1}^n c^j_{p,m}(X^i) = 0$ a.s. 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 $j = N+2, N+3, ..., (p+2)N+1$, $c^j_{p,m}(X^0) \in [0, \Delta_{p,m}]$ a.s. and by Hoeffding's inequality,

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

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

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

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

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

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

with at least probability $1 - 2(p+1)Ne^{-2n(\alpha/\Delta_{p,m})^2}$. Using this fact, Example 7.62 of [56] gives that with the same probability,

$$d\!l_\rho^+(\varphi^n, \varphi^{0,n}) \leq \alpha(\rho + \|r^0\|)\sqrt{(p+1)N} + \left(1 + \Delta_{p,m}\sqrt{(p+2)N+1}\right) d(r^0, R^n),$$

where $d\!l_\rho^+$ is closely related to $d\!l_\rho$; see §7.I in [56]. Then, from Theorem 7.69 in [56], we deduce the first result after realizing that r^0 is an ε-optimal solution of $\min \varphi^{0,n}$, where the additional factor $1 + 4\rho/\varepsilon$ arises from that theorem. Proposition 3.1 yields the second conclusion. □

Theorem 4.7 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 $n\alpha^2$ is large. The second source is caused by $d(r^0, R^n)$, the distance between the true epi-spline parameter and the constraint set R^n. Of course, if only appropriate soft information is included, then $r^0 \in R^n$ and $d(r^0, R^n) = 0$. Otherwise, incorrect specification of soft information induces a "bias" in the density estimator. We note that Theorem 4.7 provides additional support for considering $(rlxP_{p,m}^n)$ also for instances which are not loosely constrained. Even in such cases, $(rlxP_{p,m}^n)$ is guaranteed to generate a density near the true density.

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

4.8 Corollary *For sufficiently large n, suppose that the assumptions of Theorem 4.7 hold and* $d(r^0, R^n) = 0$ *a.s. Then,*

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

Proof: Theorem 4.7 and the fact that $d(r^0, R^n) = 0$ imply that for sufficiently large n

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

where $K' = \left\| \int_{m_0}^{m_N} c_{p,m}(x) h^0(x) dx \right\| (1 + 4\rho/\varepsilon)(\rho + \|r^0\|)\sqrt{(p+1)N}$. We let $\zeta > 0$ and couple α and n such that $\zeta = 2(p+1)Ne^{-2n(\alpha/\Delta_{p,m})^2}$, i.e., $n = -\Delta_{m,p}^2 \log(\zeta/2(p+1)N)/(2\alpha^2)$. Consequently, $\text{Prob}(n^{1/2} d_{KL}(h^0 || h_\varepsilon^n) > \beta) \leq \zeta$, 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 $n^{-1/2}$ is obtained for the exponential epi-spline estimator even if soft information is "active."

5 Numerical Examples

We illustrate the exponential epi-spline estimator through a series of examples using a freely available Matlab toolbox [59] that relies on the fmincon solver (Matlab 7.10.0); see also [8] for a corresponding

R toolbox. The focus is on showing the effect of including various sources of soft information in the context of small sample sizes. §5.1 shows estimates of an exponential density using 10 observation and an increasing collection of soft information. §5.2 provides an alternative to the Bayesian paradigm and demonstrates how a diverse family of densities can be generated. §5.3 examines the probability density of customer time-in-service for a modified M/M/1 queue. §5.4 shows the effect of incorrect soft information. The section ends with §5.5, where soft information about moments is examined for increasing sample sizes. It is beyond the scope of the paper to include a comprehensive comparison with alternative density estimators, which in any case have difficulties with incorporating an arbitrary set of soft information. Occasionally, we simply contrast with kernel estimates using "ksdensity" in Matlab, a Gaussian kernel, and default bandwidths, which are optimized in some sense for the normal densities. These estimates can possibly be improved with better bandwidth and kernel choices. 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, and use uniform meshes. The set R^n always includes the loose constraints $-1000 \leq r \leq 1000$. The Gauss-Legendre quadrature rule with 20 points evaluates the integrals over each segment (m_{k-1}, m_k) with high accuracy. We often assess the quality of an estimate h^n of a density h^0 by the mean-square error (MSE) $\int_{-\infty}^{\infty}(h^n(x) - h^0(x))^2 h^0(x)dx$. For additional numerical results we refer to [66, 58, 65, 57].

5.1 Value of Soft Information

We illustrate the effect of soft information in a simple example. For a true exponential density with parameter $\lambda = 1$ (dotted black curves in Figure 2) and a sample of size 10 (see green stems), Figure 2 shows our exponential epi-spline estimates (solid red curves) under two classes of soft information: (a) continuously differentiable, nonnegatively supported, and log-concave density and (b) also nonincreasing density and a relative bound on the slope. The soft information about relative slope amounts to letting the quantity $h^{n\prime}(x)/h^n(x)$ be in the interval $[-1, 0]$. We observe that the exponential density h^0 with parameter $\lambda = 1$ has $h^{0\prime}(x)/h^0(x) = -1$ for all $x \geq 0$.

For comparison, a kernel estimate, incorporating information about a nonnegative support, is displayed by dashed black curves. The exponential epi-spline estimates are obtained using a mesh with $N = 10$. In Figure 2(a), MSE is 0.1144 and 0.3273 for exponential epi-spline and kernel estimates, respectively. The kernel estimate reaches well above 4.5 near zero, though the plots are truncated for the sake of clarity. Figure 2(b) shows the visually improved exponential epi-spline estimate with a reduced MSE of 0.0416. The exponential epi-spline estimates miss 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 estimate is both qualitatively and quantitatively close to the true density elsewhere. The ability to incorporate various kinds of soft information along the lines illustrated here offers the analyst a valuable tool for exploring assumptions and their consequences. One can attempt to improve the kernel estimate using various bandwidth as well as truncation (see for example [68, p.19]). The effect on bandwidth in the kernel estimate is illustrated in Figure 3(a), where the case with default bandwidth (given in Figure 2(a)) is supplemented by estimates using bandwidth 0.15, 0.2625, 0.375, 0.4875, and 0.6. The combination of a nonnegative support and few data points make it nontrivial to select an appropriate bandwidth and the estimates remain mostly qualitatively similar. In Figure 3(b) we remove the requirement of a nonnegative support. Again, the choice of bandwidth appears challenging. However, truncation and renormalization of the portion of the density estimates to the left

of the origin improves the situation; see the dashed blue lines in Figures 2(a), 2(b), and 3(a) that give the resulting estimate when truncating the (default) density estimate in Figure 3(b) (black line).

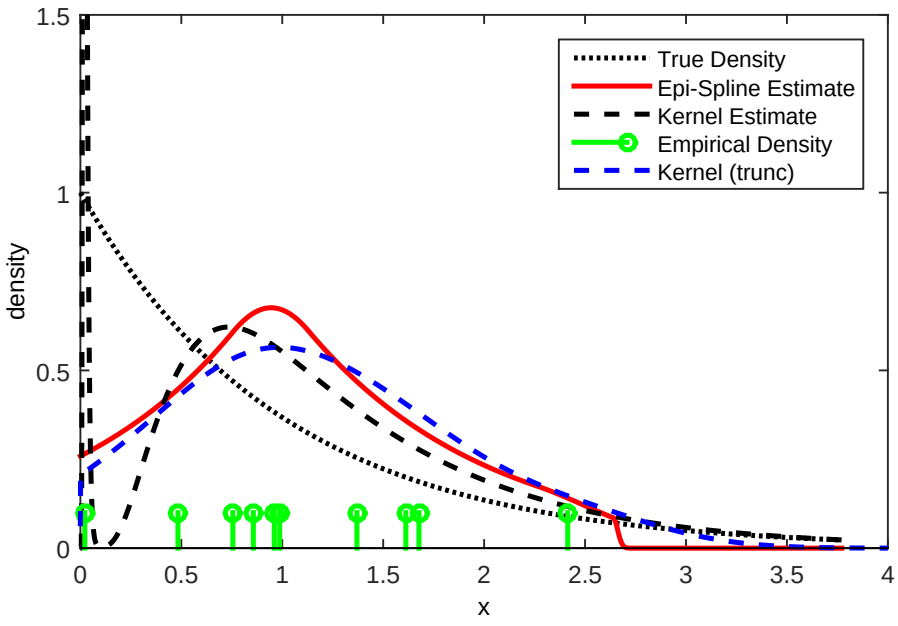
5.2 Kullback-Leibler Divergence and the Bayesian Paradigm

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 ($\tilde{P}^n_{m,p}$) to a neighborhood of a reference density h^{ref} using (4). 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 4. We set $N = 10$ and restrict the search to continuously differentiable densities. If no emphasis is placed on the reference density, i.e., $\varphi(10) = \infty$ in (4), then we obtain the exponential epi-spline estimate marked with the red dotted line in Figure 4. As proximity to the reference density is enforced more vigorously by setting $\varphi(10) = 1$, 0.1, and 0.01, we obtain the dashdot, dashed, and solid lines, respectively, in Figure 4. The Kullback-Leibler divergence constraints dampen the oscillations caused by the sample by a degree determined by $\varphi(10)$, which in practice should be selected based on the confidence in the correctness of the reference density.

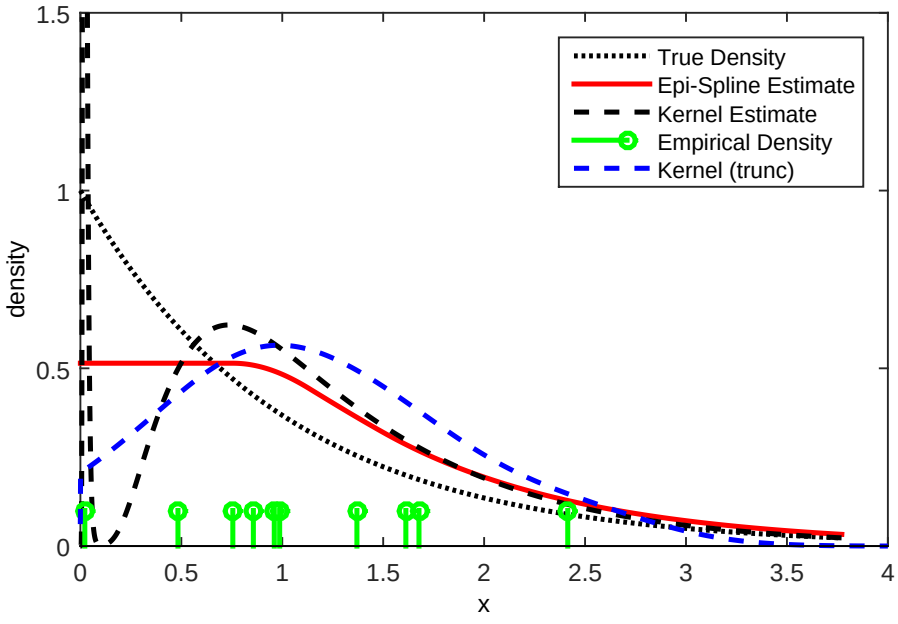
A related situation arises when an analyst would like to generate multiple densities that span a range of possibilities, for example to account in some manner for questionable soft information. For example, when the estimated density is to be used as input in further simulation and optimization, it may 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 §5.1. Suppose that the second density generated there (see Figure 2(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 (4) 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 2(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 likelihood function of the sample. Figure 5 shows the results with the solid red line and dotted black line showing the original estimate and true density as in Figure 2(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 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 Queueing Model Output

Significant challenges arise when the density to be estimated is discontinuous. We illustrate this situation here by an example taken from [65]; see [57] for additional examples. Suppose that the random variable of interest is the customer time-in-service of a modified M/M/1 queue with arrival rate $\lambda = 1$ and service rate $\mu = 1.5$, but where 50% of customers who enter the system are held at a separate station for two time units. Obviously, the true density is an equal mixture of the probability density of the customer time-in-service without a separate station (an exponential density) and the same density shifted to the right by two time units, yielding a discontinuous density. Using a sample of size $n = 100$, we aim to recover this density using a lower semicontinuous exponential epi-spline estimate with $N = 10$.

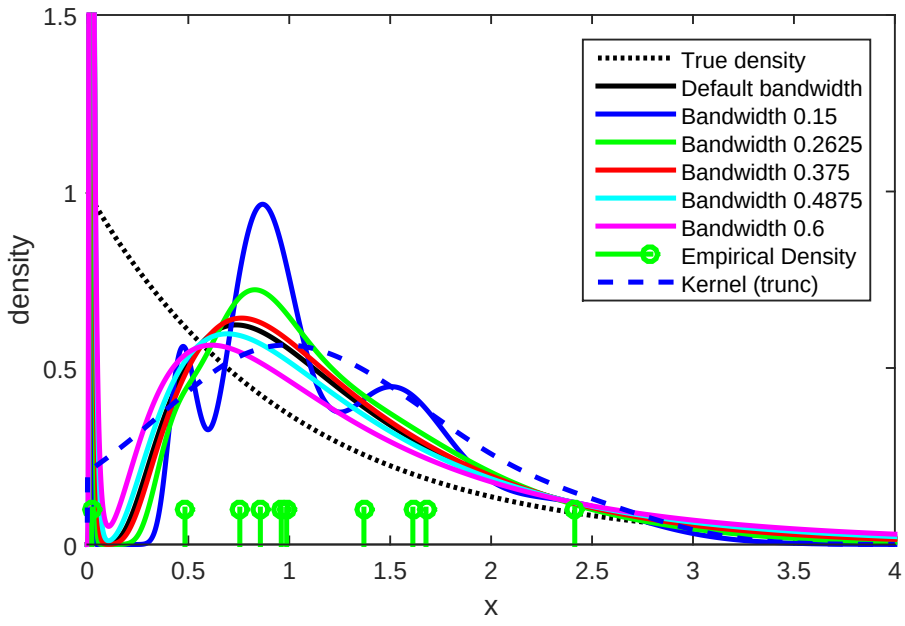


(a)

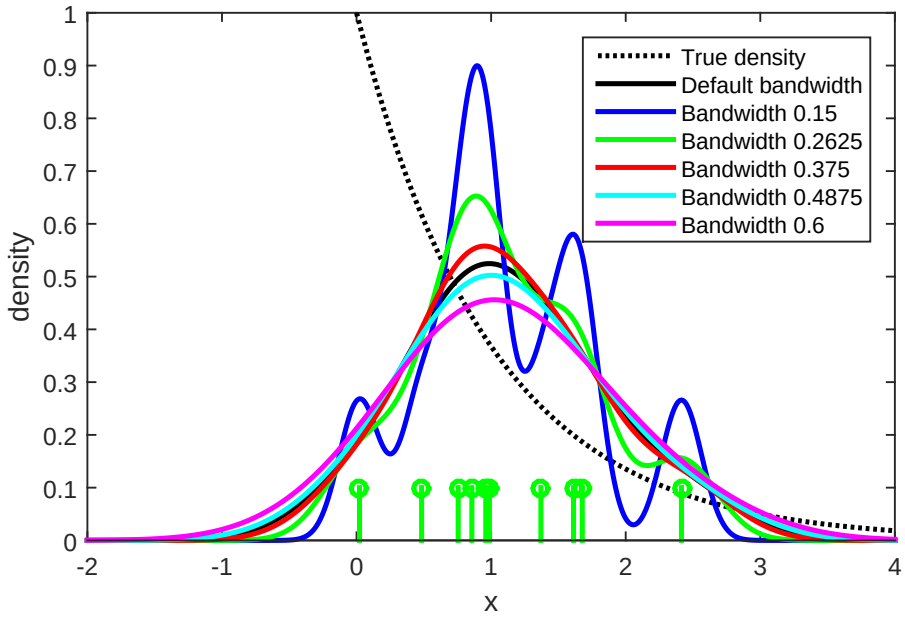


(b)

Figure 2: Exponential example: (a) continuously differentiable, nonnegative support, and log-concave, (b) also nonincreasing and relative bounds on slope.



(a)



(b)

Figure 3: Exponential example: Kernel estimates using varying bandwidth and (a) nonnegative support and (b) unbounded support.

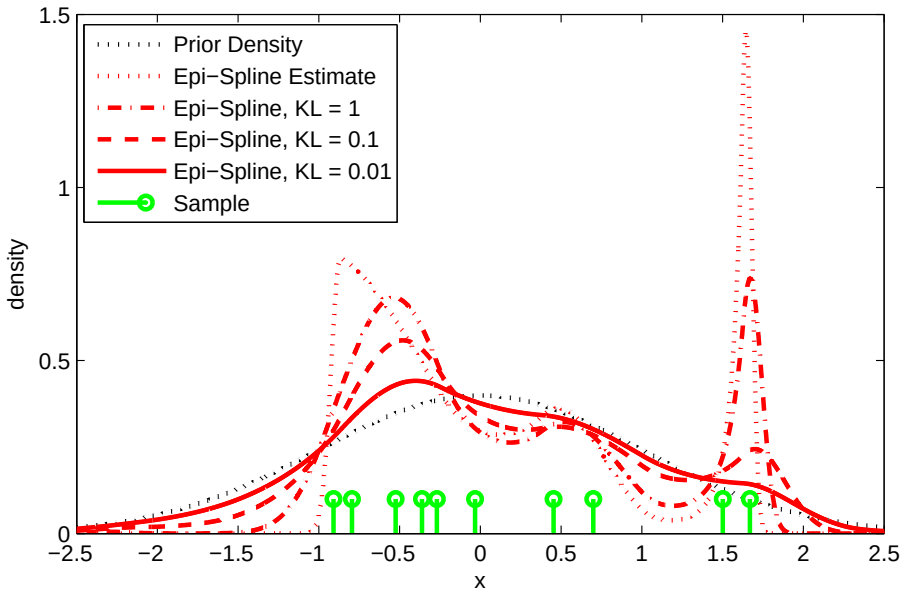


Figure 4: Normal Example: Kullback-Leibler divergence constraint.

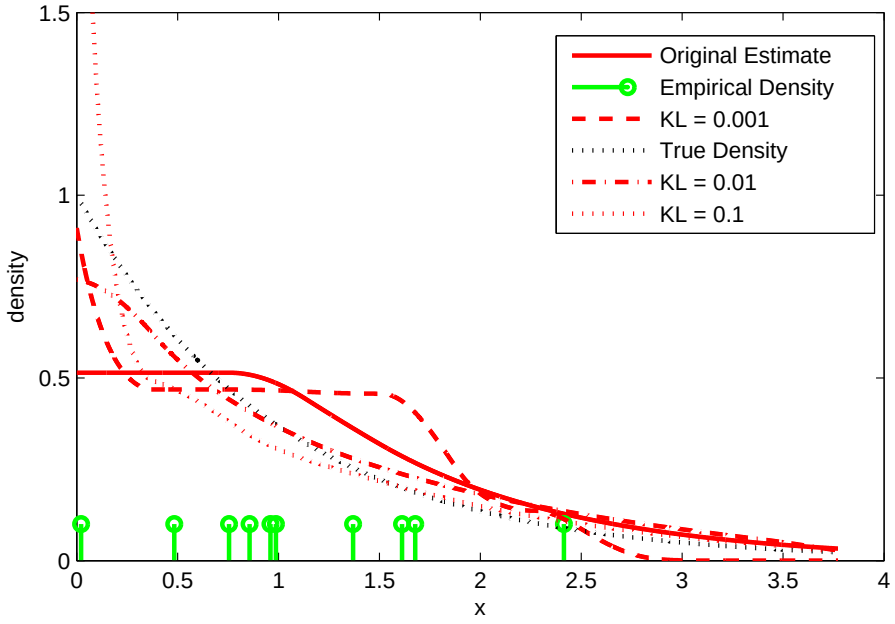


Figure 5: Exponential Example: Diversification through Kullback-Leibler divergence.

Table 1 shows aggregated results across 100 meta-replications for a variety of soft information. The first row of results shows MSE under no additional information beyond lower semicontinuity and bounds on the second-order derivatives. The second row corresponds to a restriction of the slope to be in the interval $[-4, 0]$ and the third row assumes a nonnegative support. The last row incorporates bounds on the slope, nonnegative support, and log-concavity of the upper tail. We show that the average MSE (second column) decreases with increasing soft information and mostly also the standard deviation of the MSE (third column).

Information	Average MSE	Standard Deviation
no additional info.	0.0045	0.0020
slope	0.0040	0.0016
lower support bound (lb)	0.0040	0.0021
slope, lb, and tail	0.0030	0.0017

Table 1: MSE of customer time-in-service for queueing model with various levels of soft information.

Figure 6 shows an instance corresponding to the last row in Table 1. The MSE of the exponential epi-spline (red line) estimate is 0.0016. The exponential epi-spline estimate captures the essence of the true density (dotted line) rather well.

5.4 Incorrect Soft Information

As given by Theorem 4.2, optimal solutions of $(P^n_{p,m})$ 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 7(a) shows the resulting exponential epi-spline estimate (solid red line) and the kernel estimate (dashed black line) for $n = 100$. Figure 7(b) displays the corresponding results for $n = 1000$. We observe that while the kernel estimator benefits from the larger sample size and obtains a nearly perfect estimate for $n = 1000$, 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.

5.5 Moment Information

We end the section by presenting a summary of results over a range of sample sizes for a normal density with zero mean and standard deviation of two. 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 that amounts to continuous differentiability, log-concavity, and bounds on first and second moments that ensure estimates with moments within 20% of their correct values.

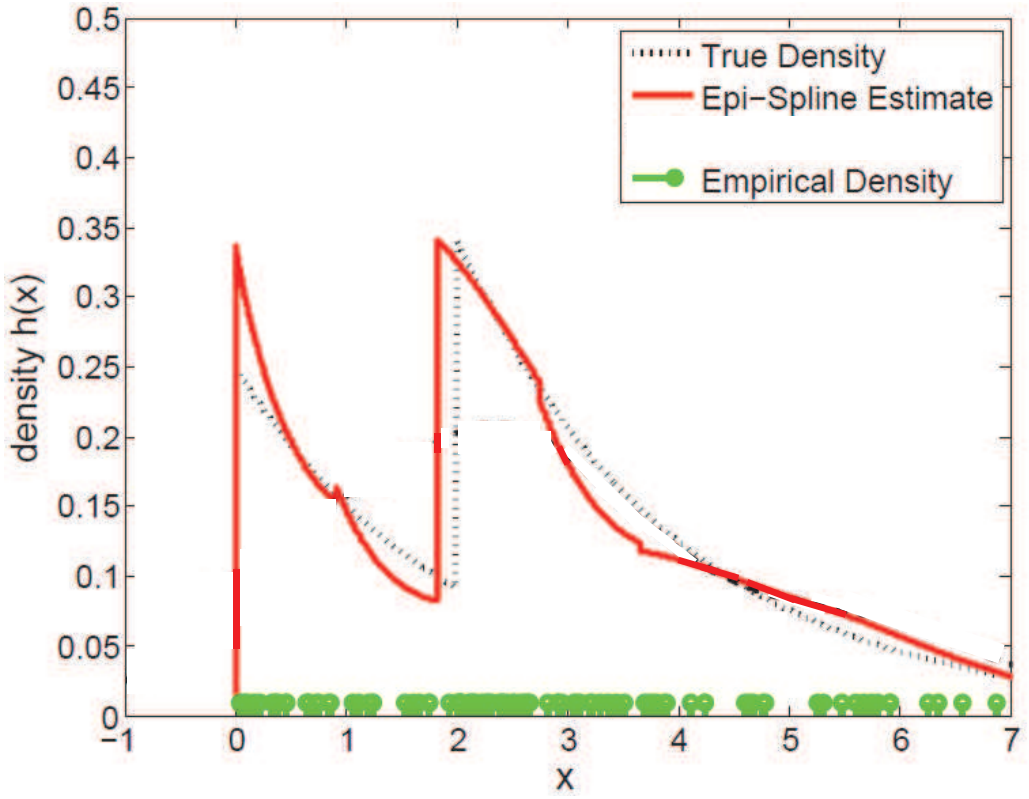
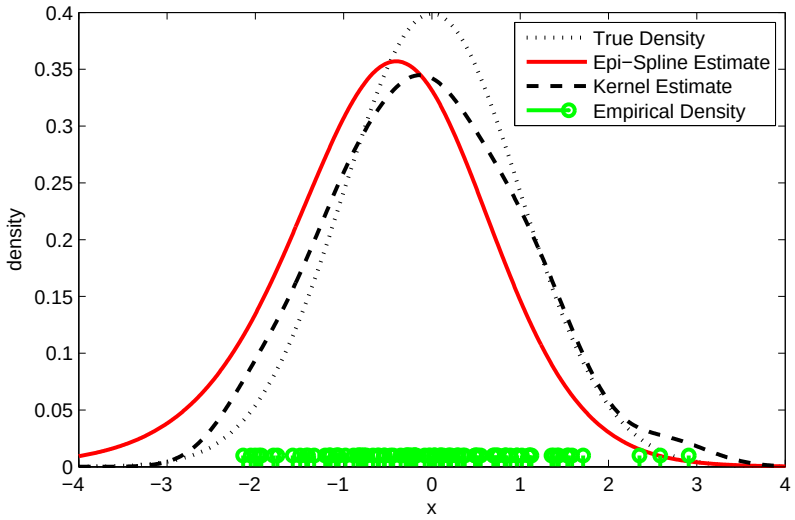
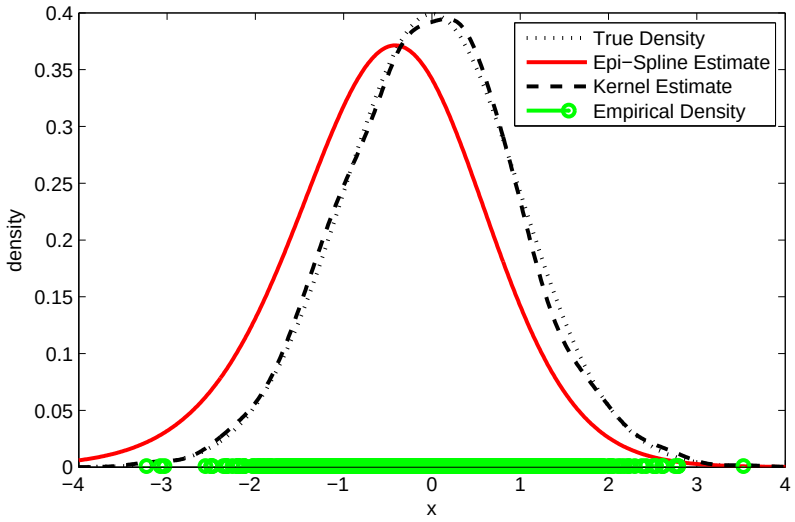


Figure 6: Customer time-in-service density.

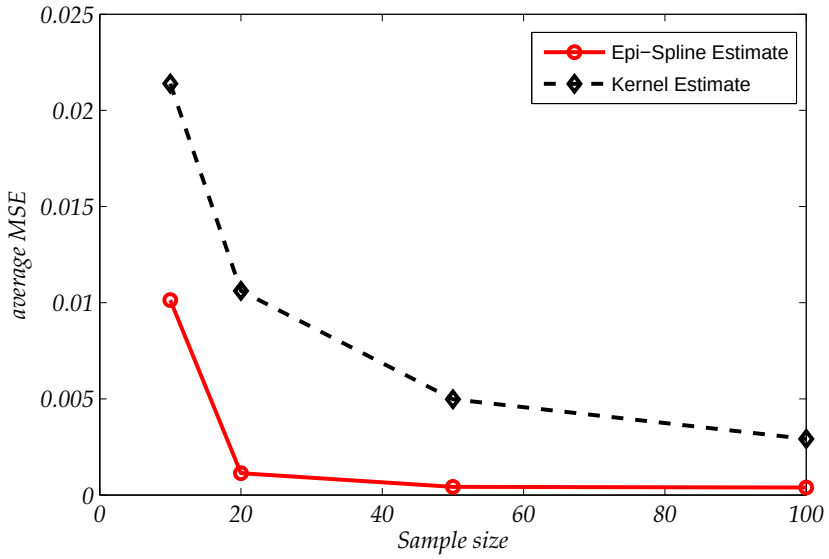


(a)

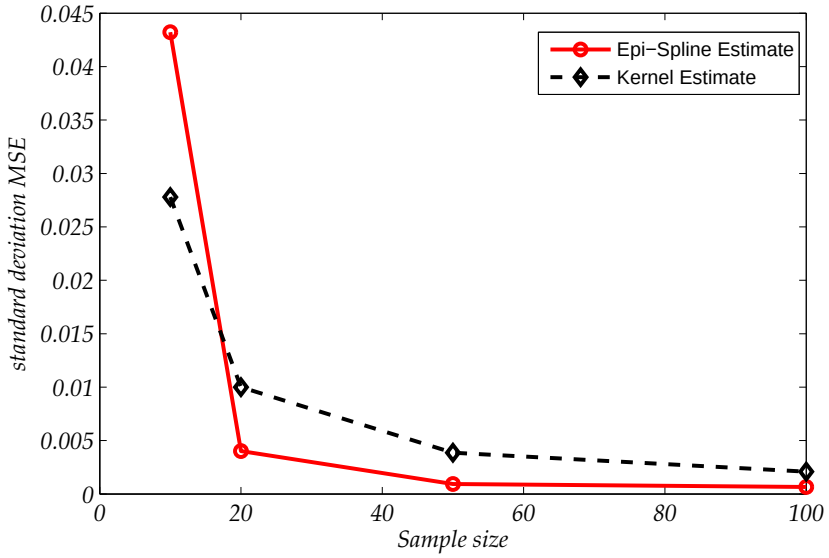


(b)

Figure 7: Normal Example: Estimates for $n = 100$ (a) and $n = 1000$ (b) with incorrect constraint $\int_{m_0}^{m_N} xh^n(x)dx \leq -0.5$.



(a)



(b)

Figure 8: Average Normal Example: Average (a) and standard deviation (b) of MSE for exponential epi-spline and kernel estimators for a range of sample sizes.

Figure 8 gives the corresponding average and standard deviation of the MSE for a range of sample sizes. We see that the exponential epi-splines estimates result in smaller MSE, on average. However, the advantage decreases as the sample size grows as expected.

6 Conclusions

We have developed a constrained maximum likelihood estimator that incorporates any soft information that might be available and therefore offers substantial flexibility for practitioners. In particular in situations with few (hard) observations, soft information can be brought in and reasonable estimates can be achieved with as little as 10 sample points. In simple but illustrative examples of estimating exponential, normal, and mixture of exponential distributions, we construct new estimates under a variety of soft information not commonly considered. The estimator requires the solution of an infinite-dimensional optimization problem, which we carry out approximately utilizing exponential epi-splines. The justification stems from the fact that exponential epi-splines can approximate to an arbitrary level of accuracy practically any density. We show that optimization over exponential epi-splines often reduces to convex programming. Our theoretical development establishes consistency, asymptotic normality, and finite sample error of order $O(n^{-1/2})$ under the assumption that the true density is an exponential epi-spline.

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, K. Rufiback, and J. A. Wellner. Limit distribution theory for maximum likelihood estimation of a log-concave density. *Annals of Statistics*, 37:1299–1331, 2009.

[3] 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.

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

[5] R.R. Barton, B.L. Nelson, and W. Xie. A framework for input uncertainty analysis. In *Proceedings of the 2002 Winter Simulation Conference*, 2010.

[6] M. Birke. Shape constrained kernel density estimation. *Journal of Statistical Planning and Inference*, 139:2851–2862, 2009.

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

[8] S. Buttrey, J.O. Royset, and R. Wets. XSPL estimator: An R toolbox. http://faculty.nps.edu/joroyset/XSPL.html, 2014.

[9] 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.

[10] X. Chen. Large sample sieve estimation of semi-nonparametric models. In *Handbook of Econometric*, pages 5549–5632. 2007. Volume 6B, Chapter 76.

[11] S. E. Chick. Input distribution selection for simulation experiments: Accounting for input uncertainty. *Operations Research*, 49:744–758, 2001.

[12] M. Cule, R.J. Samworth, and M. Stewart. Maximum likelihood estimation of a multi-dimensional log-concave density. *Journal of the Royal Statistical Society Series B*, 72:545–600, 2010.

[13] M. Cule, R.J. Samworth, and M. Stewart. Rejoinder to maximum likelihood estimation of a multi-dimensional log-concave density. *Journal of the Royal Statistical Society Series B*, 72:600–607, 2010.

[14] M. L. Cule and R. J. Samworth. Theoretical properties of the log-concave maximum likelihood estimator of a multidimensional density. *Electronic J. Statistics*, 4:254–270, 2010.

[15] P. Davies and A. Kovac. Densities, spectral densities and modality. *The Annals of Statistics*, 32:1093–1136, 2004.

[16] 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.

[17] L. Dechevsky and S. Penev. On shape-preserving probabilistic wavelet approximators. *Stochastic Analysis and Applications*, 15:187–215, 1997.

[18] M. Delecroix and C. Thomas-Agnan. Spline and kernel regression under shape restrictions. In M. G. Schimek, editor, *Smoothing and Regression*, pages 109–134. Wiley, 2000.

[19] R.A. DeVore. Monotone approximation by polynomials. *SIAM Journal on Mathematical Analysis*, 8:906–921, 1977.

[20] R.A. DeVore. Monotone approximation by splines. *SIAM Journal on Mathematical Analysis*, 8:891–905, 1977.

[21] 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.

[22] C.R. Doss. *Shape-Constrained Inference for Concave-Transformed Densities and their Modes*. Phd dissertation, University of Washington, 2013.

[23] L. Dumbgen and K. Rufibach. Maximum likelihood estimation of a log-concave density and its distribution function: Basic properties and uniform consistency. *Bernoulli*, 15(1):40–68, 2009.

[24] L. Dumbgen, R. J. Samworth, and D. Schuhmacher. Approximation by log-concave distributions with applications to regression. *Annals of Statistics*, 39:702–730, 2011.

[25] P.B. Eggermont and V.N. LaRiccia. *Maximum Penalized Likelihood Estimation, Volume I: Density Estimation*. Springer, 2001.

[26] Y. Feng, D. Gade, S. Ryan, J.-P. Watson, R. Wets, and D. Woodruff. A new approximation method for generating day-ahead load scenarios. In *2013 IEEE Power & Energy Society General Meeting*. IEEE, 2013.

[27] M. Freimer and L. W. Schruben. Collecting data and estimating parameters for input distributions. In *Proceedings of the 2002 Winter Simulation Conference*, 2002.

[28] 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.

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

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

[31] U. Grenander. On the theory of mortality measurement. I. *Skandinavisk Aktuarietidskrift*, 39:70–96, 1956.

[32] U. Grenander. On the theory of mortality measurement. II. *Skandinavisk Aktuarietidskrift*, 39:125–153, 1956.

[33] U. Grenander. *Abstract Inference*. Wiley, 1981.

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

[35] P. Groenenboom and J.A. Wellner. *Information bounds and nonparametric maximum likelihood estimation*. Birkhauser, Basel, 1992.

[36] P. Hall and K.-H. Kang. Unimodal kernel density estimation by data sharpening. *Statistica Sinica*, 15:73–98, 2005.

[37] G. Jongbloed. The iterative convex minorant algorithm for nonparametric estimation. *Journal of Computational and Graphical Statistics*, 7:310–321, 1998.

[38] 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.

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

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

[41] 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.

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

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

[44] E. Lim and P. W. Glynn. Consistency of multidimensional convex regression. *Operations Research*, 60:196–208, 2012.

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

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

[47] M. Meyer and D. Habtzghib. Nonparametric estimation of density and hazard rate functions with shape restrictions. *Journal of Nonparametric Statistics*, 23(2):455–470, 2011.

[48] J. Kumar Pal, M. Woodroofe, and M. Meyer. Estimating a polya frequency function. In R. Liu, W. Strawderman, and C.-H. Zhang, editors, *Complex datasets and inverse problems*, pages 239–249. Beachwood, OH, 2007. IMS Lecture Notes Monogr. Ser., volume 54.

[49] D. Papp. *Estimation problems involving nonnegative polynomials and their restrictions*. PhD dissertation, Rutgers University, 2011.

[50] D. Papp and F. Alizadeh. Shape constrained estimations using nonnegative splines. *Journal of Computational and Graphical Statistics*, 23(1):211–231, 2014.

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

[52] G. H. Pflug and R. J-B Wets. Shape restricted nonparametric regression with overall noisy measurements. *Journal of Nonparametric Statistics*, 25:323–338, 2013.

[53] J.S. Racine. Computational tools for nonparametric estimation. https://www.economics.mcmaster.ca/people/racinej, 2015.

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

[55] I. Rios, R. Wets, and D. Woodruff. Multi-period forecasting and scenarios generation with limited data. *Computational Management Science*, in review.

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

[57] J. O. Royset and R. J-B Wets. On function identification problems. Naval Postgraduate School, Monterey, California, 2014.

[58] 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.

[59] J.O. Royset and R. Wets. XSPL estimator in matlab. http://faculty.nps.edu/joroyset/XSPL.html, 2013.

[60] K. Rufiback. Computing maximum likelihood estimators of a log-concave density function. *Journal of Statistical Computation and Simulation*, 77:561–574, 2007.

[61] A. Seregin and J. A. Wellner. Nonparametric estimation of multivariate convex-transformed densities. *Annals of Statistics*, 38(6):3751–3781, 2010.

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

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

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

[65] D. Singham, J.O. Royset, and R. J-B Wets. Density estimation of simulation output using exponential epi-splines. In *Proceedings of the 2013 Winter Simulation Conference*, 2013.

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

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

[68] A.B. Tsybakov. *Introduction to Nonparametric Estimation*. Springer, 2009.

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

[70] G. Wahba. *Spline Models for Observational Data*. SIAM, 1990.

[71] G. Walther. Detecting the presence of mixing with multiscale maximum likelihood. *Journal of the American Statistical Association*, 97:508–513, 2002.

[72] G. Walther. Inference and modeling with log-concave distributions. *Statistical Science*, 24(3):319–327, 2009.

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

[74] R. Wets and I. Rios. Modeling and estimating commodity prices: copper prices. *Mathematics of Finance and Economics*, in review.

[75] M.A. Wolters. A greedy algorithm for unimodal kernel density estimation by data sharpening. *Journal of Statistical Software*, 47(6):1–26, 2012.