European Journal of Operational Research 247 (2015) 532–547

Contents lists available at ScienceDirect

European Journal of Operational Research

journal homepage: www.elsevier.com/locate/ejor

Stochastics and Statistics

Fusion of hard and soft information in nonparametric density estimation

Johannes O. Royset [a,*], Roger J-B Wets [b]

[a] Department of Operations Research, Naval Postgraduate School, 1411 Cunningham Rd, Monterey, CA 93943, United States
[b] Department of Mathematics, University of California, Davis, United States

ARTICLE INFO

Article history:
Received 1 December 2014
Accepted 12 June 2015
Available online 18 June 2015

Keywords:
density estimation
data analytics
data fusion
epi-splines

ABSTRACT

This paper 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.

Published by Elsevier B.V.

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 information*, in contrast to *hard information* derived from observations (data). In this paper, 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 Barton, Nelson, and Xie (2010); Chick (2001); Freimer and Schruben (2002). 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 Singham, Royset, and Wets (2013), uncertainty quantification Royset, Sukumar, and Wets (2013), as well as estimation of errors in forecasts for commodity prices Wets and Rios (Under review) and electricity demand Feng, Gade, Ryan, Watson, Wets, and Woodruff (2013); see also Rios, Wets, and Woodruff (Under review).

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.

* Corresponding author. Tel.: 1 831 656 2578, Fax: 1 831 656 2595
E-mail addresses: joroyset@nps.edu (J.O. Royset), rjbwets@ucdavis.edu (R. J-B Wets).

http://dx.doi.org/10.1016/j.ejor.2015.06.034
0377-2217/Published by Elsevier B.V.

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 Grenander (1956a), 1956b). More recent studies of univariate log-concave densities include Balabdaoui, Rufiback, and Wellner (2009); Dumbgen and Rufibach (2009); Groenenboom and Wellner (1992); Jongbloed (1998); Pal, Woodroofe, and Meyer (2007); Walther (2002), with computational comparisons in Rufiback (2007); see also the review Walther (2009) and, in the case of multivariate densities, e.g., Cule, Samworth, and Stewart (2010a), 2010b). Convexity and monotonicity restrictions are examined in Groenenboom, Jongbloed, and Wellner (2001); Meyer (2012b) and monotonicity, monotonicity and convexity, U-shape, as well as unimodality with known mode are studied in Meyer (2012b); Meyer and Habtzghib (2011). Unimodal functions are also covered in Hall and Kang (2005); Reboul (2005), with the former covering U-shape as well. Monotone, convex, and log-concave densities are dealt with in Birke (2009). Studies of k-monotone densities include Balabdaoui and Wellner (2007), 2010); Gao and Wellner (2009). Densities given as monotone transformations of convex functions are examined in Seregin and Wellner (2010). Convex formulation of a collection of shape restrictions is discussed in Papp (2011); Papp and Alizadeh (2014). We refer to the recent dissertation Doss (2013) and the discussion in Cule, Samworth, and Stewart (2010b) for a more comprehensive review and to Lim and Glynn (2012) 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 Wang (1996), 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, Dong and Wets (2007) 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 Good and Gaskin (1971); Klonias (1982); Leonard (1978); de Montricher, Tapia, and Thompson (1975); Silverman (1982); Thompson and Tapia (1990) and more recently Bühlmann and van de Geer (2011); Eggermont and LaRiccia (2001); Koenker and Mizera (2006), 2008), 2010); Meyer (2012a); Turlach (2005). 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 Dong and Wets (2007) for further discussion. It is also possible to devise specialized algorithms such as the iterative convex minorant algorithm Groenenboom and Wellner (1992); Jongbloed (1998) to account for certain constraints or modify "unconstrained" estimators such as those based on kernels; Hall and Kang (2005) handles unimodality, Birke (2009) considers monotonicity, convexity, and log-concavity, and Davies and Kovac (2004) aims to reduce the number of modes; see Racine (2015); Wolters (2012) 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 Tsybakov (2009) 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 Chen (2007); Geman and Hwang (1982); Grenander (1981) 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 Delecroix and Thomas-Agnan (2000); Meyer (2012a) for such an approach based on splines. Third, despite progress towards handling shape restrictions on sieves (see for example Dechevsky and Penev (1997); DeVore (1977a), 1977b); Papp (2011); Papp and Alizadeh (2014)), there is no straightforward way of handling a comprehensive set of soft information.

In this paper, as in Dong and Wets (2007), 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 Dong and Wets (2007), 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 Royset and Wets (2014), 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 Dong and Wets (2007) 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 refer Pflug and Wets (2013) 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