MEDDE: MAXIMUM ENTROPY DEREGULARIZED DENSITY ESTIMATION

ROGER KOENKER

Abstract. The trouble with likelihood is not the likelihood itself, it's the log.

1. Introduction

Suppose that you would like to estimate a unimodal density. There are several exploratory inference approaches that could be employed, notably Cox (1966), Silverman (1981, 1983) and Hartigan and Hartigan (1985). More recently interest has focused on maximum likelihood estimation of log concave densities as described by Rufibach (2007), Walther (2009) and Cule and Samworth (2010), who offer a more direct approach to estimation of *strongly* unimodal densities as characterized by Ibragimov (1956). Weaker notions of unimodality have been explored in Koenker and Mizera (2010) and Han and Wellner (2016). In this note I would like to briefly describe some further experience with these weaker forms of unimodality and their relevance for both shape constrained estimation and norm constrained estimation of densities.

Our path leads us away from the well paved highway of maximum likelihood, but it is arguably more scenic. I will briefly review means of order ρ and their connection to classes of concave densities and Rényi entropy, and then illustrate their application to shape constrained and norm constrained density estimation. Further details about the computational methods are provided in Section 4.

2. Means of order ρ and a Hierarchy of Concave Densities

A natural hierarchy of concave functions can be built on the foundation of the weighted means of order ρ studied by Hardy, Littlewood and Pólya (1934). For any p in the unit simplex, $\mathcal{S} = \{p \in \mathbb{R}^n | p \geqslant 0, \sum p_i = 1\}$, let

$$M_\rho(a; p) = M_\rho(a_1, \cdots, a_n; p) = \Big(\sum_{i=1}^n p_i a_i^\rho \Big)^{1/\rho}, \quad \rho \neq 0,$$

with $M_0(a; p) = M_\rho(a_1, \cdots, a_n; p) = \prod_{i=1}^n a_i^{p_i}$ as a limit as $\rho \to 0$ The familiar arithmetic, geometric, and harmonic means correspond to ρ equal to 1, 0, and -1, respectively.

Version: November 12, 2016. The author would like to thank Ivan Mizera and Jon Wellner for very helpful conversations, and Fatih Guvenen for providing the kernel density estimate reconsidered in Example 2.2. All of the computational experience reported here was conducted in the R language with the package **REBayes**, Koenker and Gu (2016b). Code for all of the reported computation may be found in the vignette `medde.Rnw` that appears as part of that package.

Following Avriel (1972), a non-negative, real function f, defined on a convex set $C \subset \mathbb{R}^d$ is called ρ-*concave* if for any $x_0, x_1 \in C$, and $p \in [0,1]$,

$$f(p_0 x_0 + p_1 x_1) \geqslant M_\rho(f(x_0), f(x_1); p).$$

In this terminology log-concave functions are 0-concave, and concave functions are 1-concave. Since $M_\rho(a, p)$ is monotone increasing in ρ for $a \geqslant 0$ and any $p \in [0,1]$, it follows that if f is ρ-concave, then f is also ρ'-concave for any $\rho' < \rho$. Thus, concave functions are log-concave, but not vice-versa. In the limit $-\infty$-concave functions satisfy the condition

$$f(p_0 x_0 + p_1 x_1) \geqslant \min\{f(x_0), f(x_1)\},$$

so they are *quasi-concave*, and consequently so are all ρ-concave functions. Further details and motivation for ρ-concave densities can be found in Prékopa (1973), Borell (1975), and Dharmadhikari and Joag-Dev (1988).

2.1. **Estimation of log concave densities.** A probability density function, f, is called *log-concave* if $-\log f$ is a (proper) convex function on the support of f. Maximum likelihood estimation of log concave densities can be formulated as a convex optimization problem. Let $X = \{X_1, \cdots, X_n\}$ be a collection of data points in $\mathbb{R}^d$ such that the convex hull of X, $\mathcal{H}(X)$, has a nonempty interior in $\mathbb{R}^d$; such a configuration occurs with probability 1 if $n \geqslant d$ and the X_i behave like a random sample from f_0, a probability density with respect to the Lebesgue measure on $\mathbb{R}^d$. Setting $g = -\log f$, we can formulate the maximum likelihood problem as,

$$(\mathcal{P}_0) \qquad \min_g \Big\{ \sum_{i=1}^n g(X_i) \mid \int e^{-g}\, dx = 1, g \in \mathcal{K}(X) \Big\},$$

where $\mathcal{K}(X)$ denotes the class of closed convex functions on $\mathcal{H}(X) \subset \mathbb{R}^d$. As shown in Koenker and Mizera (2010) such problems have solutions that admit a finite dimensional characterization determined by the function values of $\hat{g}$ evaluated at the observed X_i with values of g elsewhere determined by linear interpolation. This primal problem has an equivalent dual formulation as,

$$(\mathcal{D}_0) \qquad \max_f \Big\{ -\int f \log f\, dy \mid f = (d(P_n - G))/dy, \quad G \in \mathcal{K}(X)^\circ \Big\},$$

where $\mathcal{K}(X)^\circ$ denotes the polar cone corresponding to $\mathcal{K}(X)$, that is,

$$\mathcal{K}(X)^\circ = \Big\{ G \in \mathcal{C}^*(X) \mid \int g\, dG \leqslant 0 \text{ for all } g \in \mathcal{K}(X) \Big\},$$

and $\mathcal{C}^*(X)$ denotes the space of (signed) Radon measures on $\mathcal{H}(X)$, its distinguished element is P_n, the empirical measure supported by the data points $\{X_i, i = 1, \cdots, X_n\}$.

It is a notable feature of the log concave MLE that it is tuning parameter free, well posed without any further need for regularization. This feature carries over to weaker forms of concave, shape constrained estimators we will consider next. It is hardly surprising in view of prior likelihood experience that our dual formulation involves maximizing Shannon entropy, since we are already well aware of the close connection to Kullback-Leibler

divergence. One potentially disturbing aspect of foregoing formulation is the finding that solutions, $\hat{g}$ must be piecewise linear, so the estimated density, $\hat{f}$ must be piecewise exponential, which when extrapolated into the tails implies sub-exponential tail behavior. This finding motivated consideration of weaker forms of concavity that permit heavier tail behavior as well as more peaked densities. A second, seemingly anomalous feature of the dual problem is the fact that G must be chosen to annihilate the jumps in P_n in order to produce a density, f, with respect to Lebesgue measure. Further details on this may be found in Koenker and Mizera (2010).

2.2. **Rényi likelihood and weaker forms of concave densities.** Given the appearance of Shannon entropy in the likelihood formulation of log concave density estimation, it is natural, or at least tempting, to consider the family of Rényi (1961) entropies,

$$R_\alpha(f) = (1-\alpha)^{-1} \log(\int f^\alpha(x)dx)$$

as a vehicle for estimating ρ-concave densities. Shannon is conveniently nested as $\alpha = 1$ in this family. We will focus attention on $\alpha < 1$, corresponding to $\rho = \alpha - 1 < 0$. Maximizing $R_\alpha(f)$ with respect to f is equivalent to maximizing, for fixed $\alpha < 1$,

$$\alpha^{-1} \int f^\alpha(x)dx.$$

This yields the new pairing of primal and dual problems:

$$(\mathcal{P}_\alpha) \qquad \min_g \Big\{ \sum_{i=1}^n g(X_i) + \frac{1}{\beta} \int g^\beta \, dx \mid g \in \mathcal{K}(X) \Big\},$$

and

$$(\mathcal{D}_\alpha) \qquad \max_f \Big\{ \frac{1}{\alpha} \int f^\alpha(y) \, dy \mid f = d(P_n - G)/dy, \quad G \in \mathcal{K}(X)^o \Big\},$$

with α and β conjugates in the usual sense that $\alpha^{-1} + \beta^{-1} = 1$.

This formulation weakens the unimodality constraint admitting a larger class of heavier tailed, more peaked densities; at the same time it modifies the fidelity criterion replacing log likelihood with a criterion based on Rényi entropy. Why not stick with log likelihood and just modify the constraint, as suggested by Seregin and Wellner (2010)? The pragmatic reason is that modifying both preserves an extremely convenient form of the convex optimization problem. This motivation is further elablorated in Koenker and Mizera (2010). From a more theoretical perspective weaker concavity requirements pose difficulties for the standard likelihood formulation, Doss and Wellner (2016) elaborate on these difficulties and demonstrate among many other things that when imposing concavity constraints with $\alpha < 0$ the MLE fails to exist.

Example. In Koenker and Mizera (2010) we stressed the (Hellinger) case that $\alpha = 1/2$, so $\beta = -1$, and $g = -1/\sqrt{f}$. Since g is constrained to be concave we conclude that the estimated density, f is ρ-concave for $\rho = -1/2$, a class that includes all of the Student t densities with degrees of freedom, $\nu \geqslant 1$, as well as all the log concaves. To illustrate the

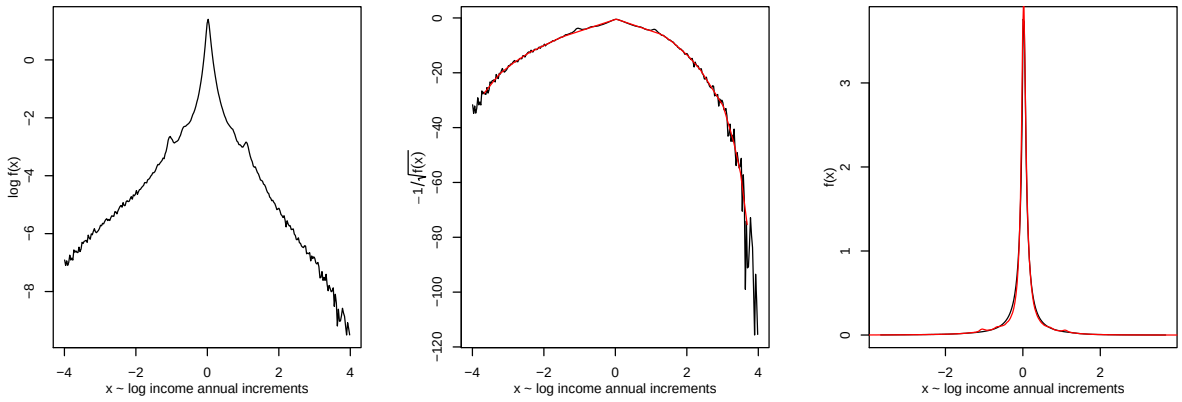

FIGURE 1. Density estimation of annual increments in log income for U.S. individuals over the period 1994-2013. The left panel of the figure reproduces a plot of the logarithm of a kernel density estimate from Guvenen et al. (2016) showing that annual income increments are clearly not log concave. However the middle panel showing $-1/\sqrt{f}$ does appear to be nicely concave and is fit remarkably well by the Rényi procedure with $\alpha = 1/2$.

applicability of this Hellinger criterion within the Rényi family, we reconsider a density estimation problem arising in econometrics. Guvenen et al. (2016) have estimated models of income dynamics using very large (10 percent) samples of U.S. Social Security records linked to W2 data. Their work reveals quite extreme tail behavior in annual log income increments. In the left panel of Figure 1 we reproduce the Guvenen et al plot of a conventional kernel density estimate of the log density of annual increments of log income based on their sample. Clearly, this density is *not* log-concave, however when we plot instead $-1/\sqrt{f}$ we see that concavity looks extremely plausible. When the Rényi estimate is superimposed in red, it fits almost perfectly.

Permitting Cauchy tail behavior may be regarded as sufficiently indulgent for most statistical purposes, but the next example illustrates that more extreme Rényi fitting criteria with $\alpha < 1/2$ is sometimes needed to accommodate sharp peaks in the target density.

Example. We reconsider the rotational velocity of stars data considered previously in Koenker and Mizera (2010). The data was taken originally from Hoffleit and Warren (1991) and is available from the R package **REBayes**. Figure 2 illustrates a histogram of the 3806 positive rotational velocities from the original sample of 3933. After dropping the 127 zero velocity observations, the histogram looks plausibly unimodal and we superimpose three distinct Rényi shape constrained estimates. The Hellinger ($\alpha = 1/2$) estimate is clearly incapable of capturing the sharp peak around $x = 18$, and even the fit for $\alpha = 0$ fails to do so. But pressing further, we see that setting $\alpha = -1$ provides an excellent fit by constraining $-1/f^2$ to be concave. Inquiring minds may wonder whether we could go

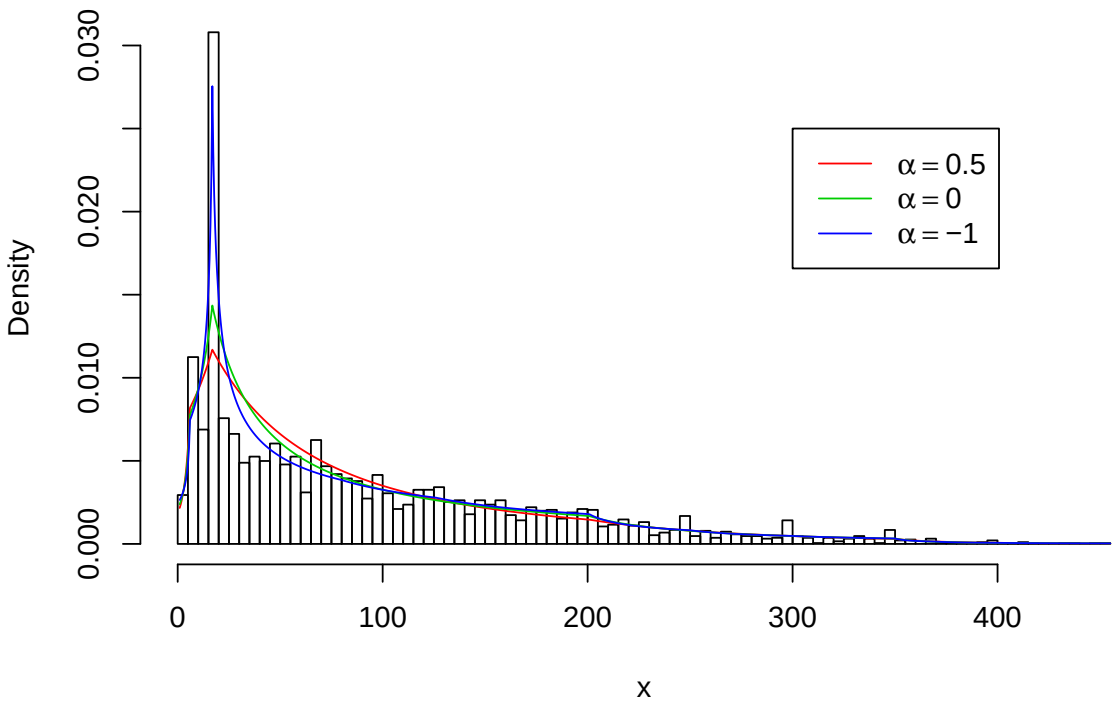

FIGURE 2. Rotational velocity of stars with three quasi concave shape constrained density estimates using the Rényi likelihood.

even further. Current numerical methods enable one to go only a tiny bit further: with $\alpha = -1.1$ one gets a fit that matches the height of the tallest histogram bar almost exactly, but venturing beyond this point fails to yield a convergent solution to the fitting algorithm. ■

3. Rényi likelihood and norm constrained density estimation

Although our original intent for the Rényi likelihood was strictly pragmatic – to maintain the convexity of the optimization problem underlying the estimation while maintaining weaker forms of the concavity constraint – I would now like to briefly consider its use in norm constrained settings where the objective of penalization is smoothness of the estimated density rather than shape constraint.

There is a long tradition of norm penalized nonparametric maximum likelihood estimation of densities. Perhaps the earliest example is Good (1971) who proposed the penalty,

$$J(f) = \int (\sqrt{f}')^2 dx,$$

which shrinks the estimated density toward densities with smaller Fisher information for location. The deeper rationale for this form of shrinkage remains obscure, and most of the subsequent literature has instead focused on penalizing derivatives of $\log f$, with the familiar cubic smoothing spline penalty,

$$J(f) = \int (\log f'')^2 dx,$$

receiving most of the attention. Silverman (1982) proposed penalizing the squared L_2 norm of the *third* derivative of $\log f$ as a means of shrinking toward the Gaussian density.

Squared L_2 norm penalties are ideal for smoothly varying densities, but they abhor sharp bends and kinks, so there has been some interest in exploring total variation penalization as a way to expand the scope of penalty methods. The taut-string methods of Davies and Kovac (2001, 2004) penalize total variation of the density itself. Koenker and Mizera (2007) describe some experience with penalties of the form,

$$J(f) = \int |\log f''| dx,$$

that penalize the total variation of the first derivative of $\log f$. In the spirit of Silverman (1982) the next example illustrates penalization of the total variation of the third derivative of $\log f$, again with the intent of shrinking toward the Gaussian, but in a manner somewhat more tolerant of abrupt changes in the derivatives than with Silverman's squared L_2 norm.

Example. In Figure 3 we illustrate a histogram based on 500 standard Gaussian observations, and superimpose two fitted densities estimated by penalizaed maximum likelihood as solutions to

$$\min_f \Big\{ -\sum_{i=1}^{n} \log f(X_i) + \lambda \int |\log f'''| dx,$$

for two choices of λ. For λ sufficiently large solutions to this problem conform to the parametric Gaussian MLE since the penalty forces the solution to take a Gaussian shape, but does not constrain the location or scale of the estimated density. For smaller λ we obtain a more oscillatory estimate than conforms more closely to the vagaries of the histogram. ■

Penalizing total variation of $\log f''$ as in Figure 3 raises the question: What about other Rényi exponents for $\alpha \neq 1$? Penalizing $\log f''$ is implicitly presuming sub-exponential tail behavior that may be better controlled by weaker Rényi penalties. To explore this conjecture we consider a in the next example estimating a mixture of three lognormals.

■

Example. Figure 4 illustrates a histogram based on 500 observations from the mixture of lognormals density depicted in red. I have used this density for several years in class to illustrate how difficult it can be to choose an effective bandwidth for conventional kernel density estimation. A bandwidth sufficiently small to distinguish the two left-most modes is almost surely incapable of producing a smooth fit to the upper mode. Logspline methods

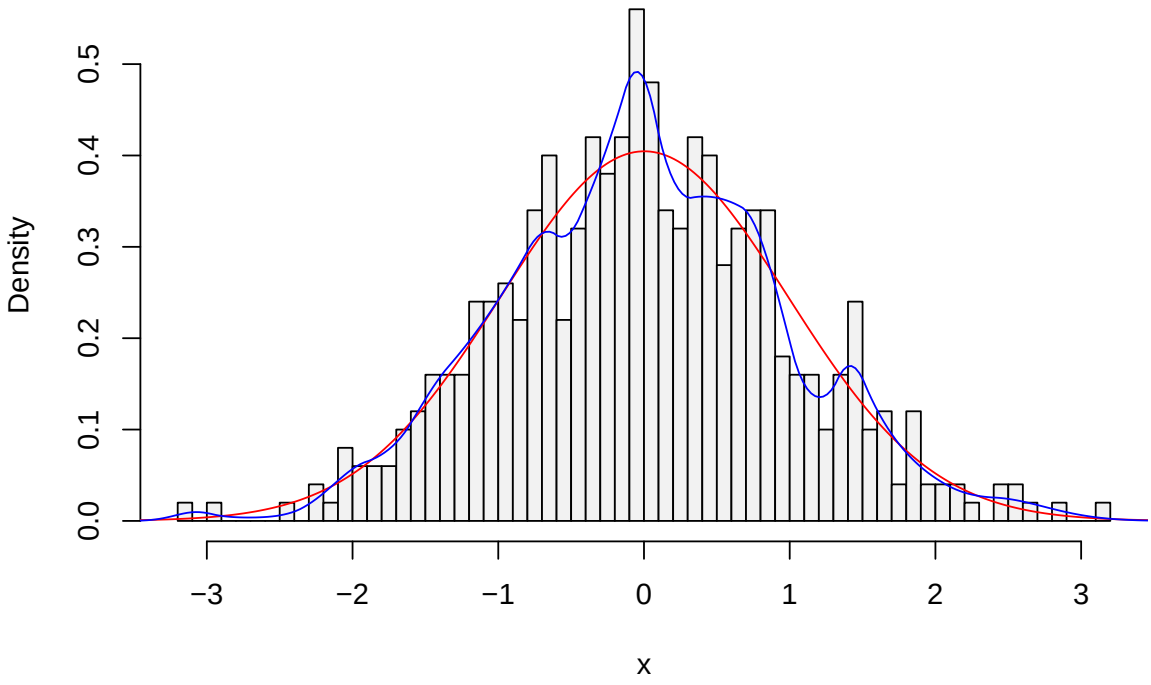

FIGURE 3. Gaussian histogram based on 500 observations and two penalized maximum likelihood estimates with total variation norm penalty and $\lambda \in \{0.5 \times 10^{-3}, 0.5 \times 10^{-6}\}$.

as proposed by Kooperberg and Stone (1991) perform much better in such cases, but they can be sensitive to knot selection strategies. The methods under consideration here are allied more closely to the smoothing spline literature, and thereby circumvent the knot selection task, but in so doing have introduced new knobs to turn and buttons to push. Not only do we need to choose the familiar λ, there is now a choice of the order of the derivative in the penalty, and the Rényi exponent, α, determining the transformation of the density. I would argue that these choices are more easily adapted to particular applications, but others may disagree. From a Bayesian perspective, however, it seems indisputable that more diversity in the class of tractable prior specifications is desirable.

Examining Figure 4 we see that the $\alpha = 1$ maximum likelihood estimate is a bit too smooth, failing to find the second mode, whereas the $\alpha = 0$ solution is too enthusiastic about the fitting the first mode, but at least does distinguish the second mode. Both methods produce an excellent fit to the third mode, almost indistinguishable from the true density. ■

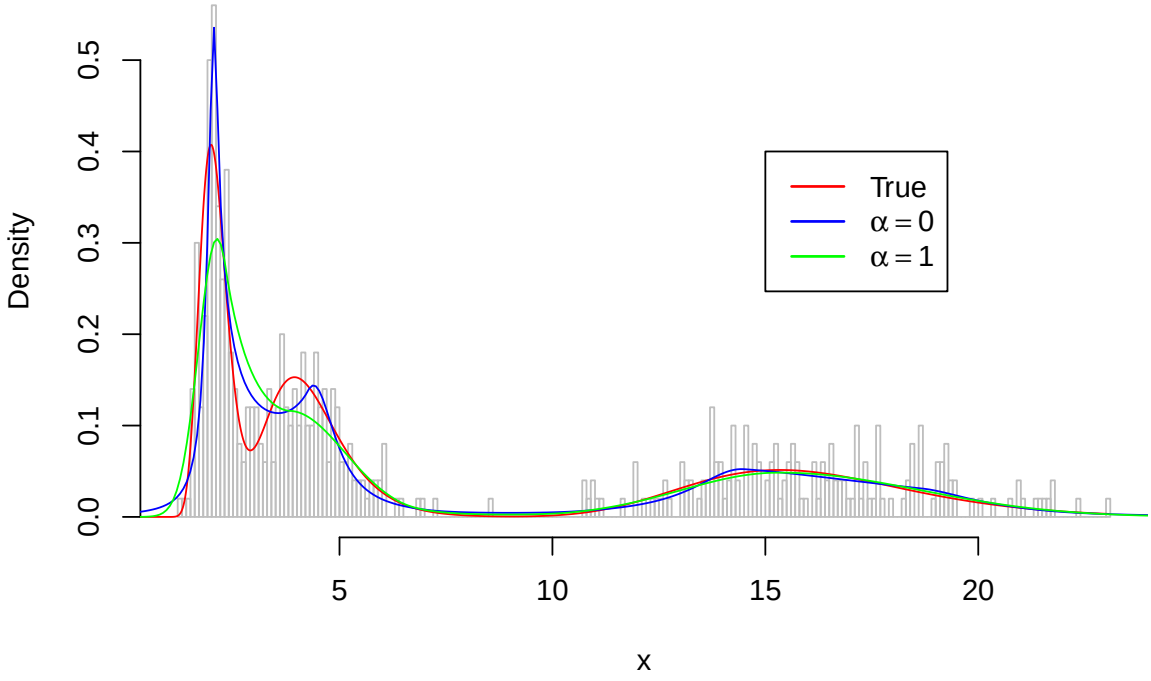

FIGURE 4. Mixture of three lognormals histogram and two Rényi likelihood estimates with total variation (L_1 norm) penalty with $\alpha \in \{0, 1\}$.

4. Discrete Implementation Details

The discrete formulation of the variational problems described above lead to extremely efficient algorithms that exploit modern interior point methods for convex optimization. All of our computational results were carried out with the function `medde` from the package **REBayes** for the R language and available from the CRAN website, `https://cran.r-project.org`. This package relies in turn on the Andersen (2010) optimization system and its Friberg (2012) interface for the R language. The current implementation in `medde` is restricted to univariate densities as we will do here as well. Koenker and Mizera (2010) describes some extensions to bivariate settings. Most of the other functionality of the **REBayes** is devoted to empirical Bayes methods and described in Koenker and Gu (2016a).

For univariate densities convexity of piecewise linear functions can be enforced by imposing linear inequality constraints on the set of function values $\gamma_i = g(\xi_i)$ at selected points $\xi_1, \xi_2, \ldots, \xi_m$. We typically choose these ξ_i's on an equally spaced grid of a few hundred points, and the convex cone constraint can then be expressed as $D\gamma \geqslant 0$ for a

tridiagonal matrix D when the penalization is imposed on second derivatives as in the case of our concavity constraints, and quindiagonal in the case of third derivative constraints. By default in `medde` we choose $m = 300$, with the ξ_i's support extending slightly beyond the empirical support of the observations.

As described in Koenker and Mizera (2010) the primal formulation of the shape-constrained problem takes the discrete form,

$$\{w^\top L\gamma + s^\top \Psi(\gamma) | D\gamma \geqslant 0\} = \min! \tag{P}$$

where $\Psi(\gamma)$ denotes an m-vector with typical element $\psi(g(\xi_i)) = \psi(\gamma_i)$, L is an "evaluation operator" which either selects the elements from γ, or performs the appropriate linear interpolation from the neighboring ones, so that $L\gamma$ denotes the n-vector with typical element, $g(X_i)$, and w is an n-vector of observation weights, typically $w_i \equiv 1/n$. The matrix D is the discrete derivative operator that constrains the fitted function to lie in the convex cone $\mathcal{K}(X)$. The vector s denotes weights that impose the integrability constraint on the fitted density. As long as the grid is sufficiently fine in univariate settings elements of s can be averages of the adjacent spacings between the ξ_i's.

Associated with the primal problem (P) is the dual problem,

$$\{-s^\top \Psi^*(-\phi) \mid S\phi = -w^\top L + D^\top \eta, \phi \geqslant 0, D^\top \eta \geqslant 0\} = \max! \tag{D}$$

Here, η is an m-vector of dual variables and ϕ is an m-vector of function values representing the density evaluated at the ξ_i's, and $S = \operatorname{diag}(s)$. The vector Ψ^* is the convex conjugate of Ψ defined coordinate-wise with typical element $\psi^*(y) = \sup_x\{yx - \psi(x)\}$. Problems (P) and (D) are strongly dual in the sense of the following result, which may viewed as the discrete counterpart of Theorem 2 of Koenker and Mizera (2010).

For $\Psi(x)$ with typical element $\psi(x) = e^{-x}$ we have Ψ^* with elements $\psi^*(y) = -y \log y + y$, so the dual problem corresponding to maximum likelihood can be interpreted as maximizing the Shannon entropy of the estimated density subject to the constraints appearing in (D). Since g was interpreted in (P) as $\log f$ this result justifies our interpretation of solutions of (D) as densities provided that they satisfy our integrability condition. This is easily verified and thus justifies the implicit Lagrange multiplier of one on the integrability constraint in (P). Then solutions ϕ of (D) satisfy $s^\top \phi = 1$ and $\phi \geqslant 0$. as shown in Proposition 2 of Koenker and Mizera (2010). The crucial element of the proof of this proposition is that the differencing operator D annihilates the constant vector and therefore the result extends immediately to other norm-type penalties as well as to the other entropy objectives that we have discussed. Indeed, since the second difference operator representing our convexity constraint annihilates any affine function it follows by the same argument that the mean of the estimated density also coincides with the sample mean of the observed X_i's.

For penalties with Rényi exponents less than one, the dual formulation takes $\psi(y) = y^\alpha$ except of course for $\alpha = 0$ for which $\psi(y) = \log y$. To implement the total variation regularization rather than the concavity constraint, the L_1 constraint on $D\gamma$ in the primal becomes an L_∞ constraint in the dual, so in the dual formulation we simply constrain

$\|D^\top \eta\|_\infty \leqslant \lambda$, and similarly for total variation (L_1 norm) constraints on higher order derivatives.

Code to reproduce each of the figures appearing above is available from the author on request. Readers are cautioned that although all of the computational problems described above are strictly convex and therefore possess a unique solution, extreme choices of the penalty parameters can stress even the excellent optimization software provided by Mosek. In Example 2.2 we have seen that attempts to push the Rényi α parameter much below -1, cause difficulty. In Example 3 a choice of λ somewhat larger than those reported here also causes trouble. Fortunately, it is relatively easy to find values of these parameters that are within an empirically sensible range.

5. Conclusion

Density estimation by penalty methods is one of those [IJ] Good ideas of the 1970's that has matured rather slowly. Fortunately, recent developments in convex optimization have greatly expanded the menu of possible penalties, and there are promising opportunities for embedding these methods into more complex semi-parametric analyses.

Many aspects remain to be explored. We have elementary Fisher consistency results from Koenker and Mizera (2010) and some rate and limiting distributional results from Han and Wellner (2016), and others, but there are many interesting theoretical questions. It would be nice to know more about multivariate extensions. Little is known about choice of the Rényi α, can it be estimated in a reasonable way? If only we could divert some energy away from kernel methods, maybe some progress could be made in one or more of these directions.

References

Andersen ED. 2010. The Mosek optimization tools manual, version 6.0. Available from `http://www.mosek.com`.

Avriel M. 1972. r-convex functions. *Math. Programming* **2**: 309–323.

Borell C. 1975. Convex set functions in d-space. *Periodica Math. Hungarica* **6**: 111–136.

Cox D. 1966. Notes on the analysis of mixed frequency distributions. *The British Journal of Mathematical and Statistical Psychology* **19**: 39–47.

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

Davies PL, Kovac A. 2001. Local extremes, runs, strings and multiresolution. *The Annals of Statistics* **29**: 1–65.

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

Dharmadhikari S, Joag-Dev K. 1988. *Unimodality, Convexity and Applications.* Boston: Academic Press.

Doss CR, Wellner JA. 2016. Global rates of convergence of the mles of log-concave and s-concave densities. *Annals of Statistics* **44**: 954–981.

Friberg HA. 2012. Users guide to the R-to-Mosek interface. Available from `http://rmosek.r-forge.r-project.org`.

Good IJ. 1971. A nonparametric roughness penalty for probability densities. *Nature* **229**: 29–30.

Guvenen F, Karahan F, Ozkan S, Song J. 2016. What do data on millions of U.S. workers reveal about life-cycle earnings dynamics? Federal Reserve Bank of New York Staff Reports.

Han Q, Wellner JA. 2016. Approximation and estimation of s-concave densities via rÃ©nyi divergences. *Annals of Statistics* **44**: 1332–1359.

Hardy GH, Littlewood JE, Pólya G. 1934. *Inequalities*. London: Cambridge U. Press.

Hartigan J, Hartigan P. 1985. The dip test of unimodality. *Annals of Statistics* **13**: 70–84.

Hoffleit D, Warren WH. 1991. *The Bright Star Catalog (5th ed.)*. New Haven: Yale University Observatory.

Ibragimov IA. 1956. On the composition of unimodal distributions. *Theory of Probability and its Applications* **1**: 255–260.

Koenker R, Gu J. 2016a. REBayes: An R package for empirical Bayes mixture methods. *Journal of Statistical Software* Available at: `https://CRAN.R-project.org/package=REBayes`.

Koenker R, Gu J. 2016b. *REBayes: Empirical Bayes Estimation and Inference in R*. Available from `https://www.r-project.org/package=REBayes`.

Koenker R, Mizera I. 2007. Density estimation by total variation regularization. In Nair V (ed.) *Advances in Statistical Modeling and Inference: Essays in Honor of Kjell Doksum*. World Scientific, 613–633.

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

Kooperberg C, Stone CJ. 1991. A study of logspline density estimation. *Computational Statistics and Data Analysis* **12**: 327–347.

Prékopa A. 1973. On logarithmic concave measures and functions. *Acta Sci. Math. (Szeged)* **32**: 335–343.

Rényi A. 1961. On measures of information and entropy. In *Proceedings of the 4th Berkeley Symposium on Mathematics, Statistics and Probability*. University of California Press, 547–561.

Rufibach K. 2007. Computing maximum likelihood estimators of a log-concave density function. *J. Statistical Computation and Simulation* **77**: 561–574.

Seregin A, Wellner JA. 2010. Nonparametric estimation of multivariate convex-transformed densities. *Annals of Statistics* **38**: 3751–3781.

Silverman B. 1981. Using kernel density estimates to investigate multimodality. *Journal of the Royal Statistical Society (B)* **43**: 97–99.

Silverman B. 1983. Some properties of a test for multimodality based on kernel density estimates. In Kingman J, Reuter G (eds.) *Probability, Statistics and Analysis*. Cambridge University Press, 248–259.

Silverman BW. 1982. On the estimation of a probability density function by the maximum penalized likelihood method. *Ann. Statist.* **10**: 795–810.

Walther G. 2009. Inference and modeling with log-concave distributions. *Statistical Science* **24**: 319–327.