NOTES ON THE IMPLEMENTATION OF RÉNYI PENALIZED DENSITY ESTIMATION

ROGER KOENKER

> medder, v. used in Bahamian dialect, mostly on the Family Islands like Eleuthera and Cat Island meaning "mess with" "get involved," "get entangled," "fool around," "bother:" "I don't like to medder up with all kinda people" "Don't medder with people (chirren)" "Why you think she medderin up in their business."
>
> [Urban Dictionary]

1. Introduction

Our medderin' about with maximum entropy de-regularized density estimation began with an exploration of total variation penalties for smoothing density estimates in Koenker and Mizera (2007). This led to subsequent work on shape constrained density estimation initially focused on log-concavity and eventually to weaker concavity penalties that required replacing the maximum likelihood objective by an alternative Rényi entropy criteria. From the beginning our implementations of these methods relied on the convex optimization software Mosek, initially within a Matlab interface, and more recently within the Rmosek interface to R. These notes were prepared mainly as an *aide memoire* for the transition of the software implementation of the function `medde` in the R package `REBayes` from Mosek V8 to V9. For additional details one can consult ApS (2018) and ApS (2019).

2. Theory

Koenker and Mizera (2010) began with the variational formulation of the log-concave MLE problem for given observations $X = \{X_1, \cdots, X_n\}$, with $X_i \in \mathbb{R}^d$:

$$(P_1) \qquad \min\Big\{ \frac{1}{n}\sum_{i=1}^{n} g(X_i) + \int e^{-g(x)}dx \;\Big|\; g \in \mathcal{K}(X) \Big\},$$

Version: March 27, 2025.

with $\mathcal{K}(X)$ denoting the set of closed convex functions on the convex hull, $\mathcal{H}(X)$, of X. A solution $\hat{g} : \mathcal{H}(X) \mapsto \mathbb{R}$ yields a density estimate $\hat{f}(x) = \exp(-\hat{g}(x))$ on $\mathcal{H}(X)$; the fact that this obviously positive quantity is a probability density estimate, that is, its integral is equal to one, is assured by the presence of the integral term in (P_1) Outside $\mathcal{H}(X)$, the solution $\hat{g}(x) = -\infty$, implying that $\hat{f}(x) = 0$. Interpreting (P_1) as a "primal" formulation in the context of convex programming, the associated "dual" problem is,

$$(D_1) \qquad \max\Big\{\int -f \log f dx \Big|\ f = \frac{d(\mathbb{Q}(X) - G)}{dx},\ G \in \mathcal{K}(X)^o\Big\},$$

where $\mathbb{Q}(X) = n^{-1} \sum_{i=1}^{n} \delta_{X_i}$ is the empirical probability measure,

$$\mathcal{K}(X)^o = \Big\{G \in \mathcal{C}^*(X) \ \Big|\ \int g\, dG \leq 0,\ g \in \mathcal{K}(X)\Big\}$$

is the polar cone associated with $\mathcal{K}(X)$, and $\mathcal{C}^*(X)$ denotes the set of (signed) Radon measures on $\mathcal{H}(X)$. The appearance of the Shannon entropy in the dual formulation (D_1) may be interpreted as the intention to find $\hat{f}$ closest in Kullback-Leibler divergence to the uniform distribution on $\mathcal{H}(X)$ subject to the concavity constraint.

For the problem (P_1), the solutions admit further characterization: $\hat{g}$ are piecewise linear on $\mathcal{H}(X)$, so estimated densities are piecewise exponential. This feature motivated a search for larger classes of quasi-concave densities that would accommodate heavier tails and more sharply peaked densities than the log concaves. Such classes are provided by s-concave functions. Loosely speaking, a function is called *s-concave*, for $s > 0$, if its s-th power is concave. More precisely, a non-negative, real function f, defined on a convex set $C \subset \mathbb{R}^d$ is s-concave, if there is a convex function g such that

$$f = \begin{cases} (-g)^{1/s} & \text{for } s > 0, \\ \mathrm{e}^{-g} & \text{for } s = 0, \\ g^{1/s} & \text{for } s < 0. \end{cases}$$

Note that log-concave functions are 0-concave, and concave functions are 1-concave; also, if f is s-concave, then f is also s'-concave for any $s' < s$. The limiting class of $-\infty$-concave, the union of all s-concave classes for all $s \in \mathbb{R}$, is the class of *quasi-concave* functions – functions with upper level sets convex. In the one-dimensional case, for $d = 1$, this class is identical with that of unimodal functions. In higher dimensions this equivalence no longer holds.

Once log-concavity is imposed, maximizing log likelihood in (P_1) appears to be especially convenient, as it leads to a convex program with

the only nonlinearity arising from the integrability constraint. However, when weaker forms of concavity are considered, it proves more convenient to adapt the fitting criterion – in particular to retain the convexity of the optimization formulation. This was already apparent in an earlier work of Groeneboom, Jongbloed, and Wellner (2001) who employed least squares fitting rather than log-likelihood when imposing the stronger requirement of concavity of the density itself. While it is not really obvious how to adapt (P_1) to obtain a viable fitting formulation, the appearance of the Kullback-Leibler divergence in (D_1) suggests the possibility of replacing it by one of the abundant assortment of alternative divergences. Koenker and Mizera (2008, 2010) pointed out that for s-concave densities, this turns out to produce a lucky match. They proposed replacing the Shannon entropy in (D_1) by a variationally equivalent form of the Rényi entropy, a move that yielded a family of new dual and primal pairings,

$$(D_\alpha) \qquad \max\Big\{\frac{1}{\alpha}\int f^\alpha(y)\,dy \;\Big|\; f = \frac{d(\mathbb{Q}(X) - G)}{dy}, \quad G \in \mathcal{K}(X)^o\Big\},$$

and

$$(P_\alpha) \qquad \min\Big\{\sum_{i=1}^n g(X_i) + \frac{|1-\alpha|}{\alpha}\int g^\beta\,dx \;\Big|\; g \in \mathcal{K}(X)\Big\}.$$

The Rényi exponent α here corresponds to Avriel's $s = \alpha - 1$, and β is conjugate to α in the usual sense: $1/\alpha + 1/\beta = 1$.

Among the Rényi entropies, the ones enjoying particular connections to the existing literature are those with α being a multiple of $1/2$. Koenker and Mizera (2010) focused primarily on the log concave, $\alpha = 1$, case and the Hellinger, $\alpha = 1/2$, case; the latter imposes the weaker constraint that $-1/\sqrt{f}$ be concave. The implementation described here also allows us to venture into the netherworld of $\alpha \leq 0$. It should be noted that since $\alpha = 0.5$ already already subsumes tail behavior like Cauchy smaller α may be considered somewhat pathological. Sceptics, however, are encouraged to consider the examples in Koenker and Mizera (2019).

3. Praxis

Earlier implementations of these methods were developed in the standalone package `MeddeR`. However, a unification of the methods seemed desirable and has now been realized in a single function `medde` provided by the R package `REBayes` available on CRAN. These notes are intended as further documentation of this implementation for the convex optimization software Mosek. Because Mosek underwent a quite

dramatic transition from Version 8 to 9, we treat both implementations here. We will begin by treating the shape constrained case and then turn to norm constraints. Some connections to the NPMLE methods for mixture models will be described briefly in a final section.

3.1. **Mosek 8.** The separable convex optimization `scopt` formalism of Mosek 8 allows additive objective functions with nonlinear components that can take the following types: "ent" $fx\log(x)$; "exp" fe^{gx+h}; "log" $f\log(gx+h)$ and "pow" $f(x+h)^g$, where f, g, h are specified constants. For our dual formulation these terms appear only in the objective function not in the constraints, so they are represented in the Mosek formulation in the "opro" matrix of dimension 5 by p and type "list" with rows containing respectively: "type", the index j of the coordinate of x, and the corresponding elements, f, g, and h. Thus, for example for the log-concave case with $\alpha = 1$ we would have the matrix:

$$\begin{bmatrix} "ent" & "ent" & \cdots & "ent" \\ 1 & 2 & \cdots & p \\ -1 & -1 & \cdots & -1 \\ 0 & 0 & \cdots & 0 \\ 0 & 0 & \cdots & 0 \end{bmatrix}.$$

Whereas for $\alpha = 0$ we have

$$\begin{bmatrix} "log" & "log" & \cdots & "log" \\ 1 & 2 & \cdots & p \\ 1 & 1 & \cdots & 1 \\ 1 & 1 & \cdots & 1 \\ 0 & 0 & \cdots & 0 \end{bmatrix}.$$

and for other α we have

$$\begin{bmatrix} "pow" & "pow" & \cdots & "pow" \\ 1 & 2 & \cdots & p \\ -\mathrm{sgn}(\beta) & -\mathrm{sgn}(\beta) & \cdots & -\mathrm{sgn}(\beta) \\ \alpha & \alpha & \cdots & \alpha \\ 0 & 0 & \cdots & 0 \end{bmatrix}.$$

where, as usual, $1/\alpha + 1/\beta = 1$. In contrast, the primal versions for $\alpha \in \{0, 1\}$ are

$$\begin{bmatrix} "exp" & "exp" & \cdots & "exp" \\ 1 & 2 & \cdots & p \\ 1 & 1 & \cdots & 1 \\ -1 & -1 & \cdots & -1 \\ 0 & 0 & \cdots & 0 \end{bmatrix},$$

and

$$\begin{bmatrix} "log" & "log" & \cdots & "log" \\ 1 & 2 & \cdots & p \\ -1 & -1 & \cdots & -1 \\ 1 & 1 & \cdots & 1 \\ 0 & 0 & \cdots & 0 \end{bmatrix}.$$

For other α the primal form is: check this!!

$$\begin{bmatrix} "pow" & "pow" & \cdots & "pow" \\ 1 & 2 & \cdots & p \\ \text{sgn}(\beta) & \text{sgn}(\beta) & \cdots & \text{sgn}(\beta) \\ \beta & \beta & \cdots & \beta \\ 0 & 0 & \cdots & 0 \end{bmatrix}.$$

3.2. **Mosek 9.** For Mosek 9 a major revision occurred that replaced the `scopt` formulation with an alternative scheme that allowed users greater flexibility, but required them to express nonlinear components of the objective function in terms of convex cone constraints. Now there are two essential types of cone constraints: exponential cones and power cones.

The canonical exponential cone is,

$$\mathcal{K}_e = \{x \in R^3 | x_1 \geq x_2 \exp(x_3/x_2), x_1, x_2 \geq 0\},$$

or equivalently,

$$\mathcal{K}_e = \{x \in R^3 | x_3 \leq x_2 \log(x_1/x_2), x_1, x_2 \geq 0\}.$$

Thus, if we introduce auxiliary variables, $t_1, \cdots, t_p$, and require that,

$$\begin{pmatrix} e_i^\top & 0 \\ 0 & 0 \\ 0 & e_i^\top \end{pmatrix} \begin{pmatrix} x \\ t \end{pmatrix} + \begin{pmatrix} 0 \\ 1 \\ 0 \end{pmatrix} \in \mathcal{K}_e$$

we have effectively imposed that $\log(x_i) \geq t_i$, so if we can now replace our nonlinear objective function with a linear one in the auxiliary t variables. This is precisely what is required for our $\alpha = 0$ case. If instead, we require that

$$\begin{pmatrix} 0 & 0 \\ e_i^\top & 0 \\ 0 & e_i^\top \end{pmatrix} \begin{pmatrix} x \\ t \end{pmatrix} + \begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix} \in \mathcal{K}_e$$

we have imposed that $t_i \leq x_i \log(1/x_i) = -x_i \log(x_i)$. This corresponds to our primal problem with $\alpha = 1$. For other settings of α we must rely on the Mosek implementation of power cones.

The canonical three variable power cone is

$$\mathcal{K}_\alpha = \{x \in R^3 | x_1^\alpha x_2^{1-\alpha} \geq |x_2|,\ x_1, x_2 \geq 0\}.$$

Such cones can be employed to formulate our problem for $\alpha \notin \{0, 1\}$. Three cases are considered separately. For $\alpha \in (0, 1)$ we can simply use:

$$\begin{pmatrix} e_i^\top & 0 \\ 0 & 0 \\ 0 & e_i^\top \end{pmatrix} \begin{pmatrix} x \\ t \end{pmatrix} + \begin{pmatrix} 0 \\ 1 \\ 0 \end{pmatrix} \in \mathcal{K}_\alpha$$

which implies that $x_i^\alpha \geq |t_i|$. For $\alpha > 1$ we simply flip the role of x and t and replace α by its reciprocal. And for $\alpha < 0$ we impose,

$$\begin{pmatrix} e_i^\top & 0 \\ 0 & e_i^\top \\ 0 & 0 \end{pmatrix} \begin{pmatrix} x \\ t \end{pmatrix} + \begin{pmatrix} 0 \\ 0 \\ 1 \end{pmatrix} \in \mathcal{K}_\alpha$$

which implies that $t_i \geq x_i^\alpha$. In all cases we scale the summands of the now linear objective by the factor, $-\text{sgn}(\beta)$

4. Odds and Ends

The foregoing discussion was intentionally focused on transition from nonlinear objective functions in Mosek 8 to the reformulation of in terms of cone constraints in which the parameter α plays a crucial role. This has been motivated primarily by the family of concavity shape constraints. It is worth noting the flexibility of the Rényi fitting criteria can also be applied to the norm constrained estimation of densities. The parameter `lambda` in the `medde` function when it is negative imposes some form concavity restriction determined by the specification of `alpha`, when it is positive it controls the degree of smoothing imposed by the total variation penalty on some transformation of the density. When `alpha` is 1, this transformation is logarithmic.

Finally, it should be noted that `medde` also provides an opportunity to control the degree of the differential operator determining the constraints. Thus far, we have implicitly assumed that `Dorder` took the value 1, implying either a concavity constraint, or a norm constraint on the total variation of the first derivative of the fitted density. However, one can also set `Dorder = 0`, which imposes monotonicity on the estimated density when `lambda` is negative, or penalizes total variation of a transformation of the density when it is positive. When `Dorder = 2` one can impose TV smoothing on the second derivative of the transformed density as illustrated in the `demo(Silverman)`. This effectively imposes an L_1 penalty on the third derivative of the log density.

REFERENCES

APS, M. (2018): *Rmosek Release 8.1.47*.

——— (2019): *Rmosek Release 9.0.89*.

GROENEBOOM, P., G. JONGBLOED, AND J. A. WELLNER (2001): "Estimation of a Convex Function: Characterizations and Asymptotic Theory," *Annals of Statistics*, 29(6), 1653–1698.

KOENKER, R., AND I. MIZERA (2007): "Density estimation by total variation regularization," in *Advances in statistical modeling and inference, Essays in honor of Kjell A. Doksum*, ed. by V. Nair, pp. 613–633. World Scientific, Singapore.

——— (2008): "Primal and dual formulations relevant for the numerical estimation of a probability density via regularization," in *Tatra Mountains Mathematical Publications*, ed. by A. Pázman, J. Volaufová, and V. Witkovský, vol. 39, pp. 255–264. Slovak Academy of Sciences, Proceedings of the conference ProbaStat '06 held in Smolenice, Slovakia, June 5-9, 2006.

——— (2010): "Quasi-concave density estimation," *Annals of Statistics*, 38(5), 2998–3027.

——— (2019): "Shape Constrained Density Estimation Via Penalized Rényi Divergence," *Statistical Science*, 33, 510–526.