Notes on KWDual[1] for Mosek 9
Extended to Cover Rényi Objectives

Roger Koenker

1. Introduction

This is basically an *aide memoire* for the conversion of the original Mosek implementation of the Kiefer-Wolfowitz nonparametric maximum likelihood estimator for mixture models, now extended to permit a modified objective based on Rényi entropies.

2. The classical MLE formulation

The V8 implementation of KWDual directly implemented an additive formulation of the MLE objective function:

$$\min\{\sum w_i \log(\nu_i) \mid 0 \leq A\nu \leq d,\ \nu \in \mathcal{S}^n\}$$

where, to preserve the minimization sense, the $w_i = -1$. In V9 a new formulation is required to move the nonlinearity of the log objective into conic constraints. To accomplish this we need to replace the explicit log terms with the auxiliary variables, $t \in \mathbb{R}^n$, and then link the ν_i's with the t_i's via the exponential cone constraints, $t \leq \log \nu_i,\ i = 1, \cdots, n$. In Mosek cookbook notation this is written as $(\nu, 1, t) \in \mathcal{K}$. The canonical exponential cone in Mosek speak is $\mathcal{K} = \{x \in \mathbb{R}^3 | x_0 \geq x_1 \exp(x_2/x_1, x_0, x_1 \geq 0\}$, so $t \leq \log \nu_i,\ i = 1, \cdots, n$ becomes $\nu \geq \exp(t)$.

We now have $2n$ variables, so the objective function becomes linear in the t, as $w^\top t$, and we need to augment the A matrix as well to kill the t contribution, and we have as before $\nu \in \mathbb{R}^n_+$, but t lives in all of $\mathbb{R}^n$. This leaves the cone constraints. For this I was just extrapolating somewhat from example AFFCO2 in the Rmosek manual. We need to impose the constraints, for $i = 1, \cdots, n$,

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

These constraints can be stacked with the following R code:

```
P$F <- sparseMatrix(c(seq(1,3*n, by = 3), seq(3, 3*n, by = 3)),
       c(1:n, (n+1):(2*n)), x = rep(1,2*n))
P$g <- rep(c(0,1,0), n)
P$cones <- matrix(list("PEXP", 3, NULL), nrow = 3, ncol = n)
rownames(P$cones) <- c("type", "dim", "conepar")
```

This exploits the simplest of the storage schemes in the Matrix package in which one just specifies the row and column indices of the matrix and then the entries as a triple.

The F matrix in this case looks like this for $n = 6$,

[1]For obscure historical reasons the dual formulation referred to in the title is referred to as a primal formulation here, and vice-versa

```
 [1,] 1 . . . . . . . . . . .
 [2,] . . . . . . . . . . . .
 [3,] . . . . . . 1 . . . . .
 [4,] . 1 . . . . . . . . . .
 [5,] . . . . . . . . . . . .
 [6,] . . . . . . . 1 . . . .
 [7,] . . 1 . . . . . . . . .
 [8,] . . . . . . . . . . . .
 [9,] . . . . . . . . 1 . . .
[10,] . . . 1 . . . . . . . .
[11,] . . . . . . . . . . . .
[12,] . . . . . . . . . 1 . .
[13,] . . . . 1 . . . . . . .
[14,] . . . . . . . . . . . .
[15,] . . . . . . . . . . 1 .
[16,] . . . . . 1 . . . . . .
[17,] . . . . . . . . . . . .
[18,] . . . . . . . . . . . 1
```

3. Rényi Alternatives

To explore alternatives to the MLE based on Rényi entropies we need to modify the cone constraints. Here we build on the framework of Koenker and Mizera (2018) for estimating families of concave densities. In that context we began with maximum likelihood estimation of log concave densities as solutions to a primal problem,

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

with $\mathcal{K}(X)$ denoting the set of closed convex functions on the convex hull, $\mathcal{H}(X)$, of the observed sample X. Here $g(x) = -\log f(x)$, so the second term in the objective function represents a Lagrangian term imposing an integrability constraint on the estimated density, f with implicit Lagrange multiplier one. This primal problem has dual formulation,

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

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

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

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 desire to find $\hat{f}$ closest in the Kullback-Leibler divergence to the uniform distribution on $\mathcal{H}(X)$ subject to the concavity constraint.

Replacing Shannon entropy in (D_1) by a variationally equivalent form of the Rényi entropy, yields new pairs of dual and primal problems:

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

and

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

Here β is conjugate to α in the usual sense: $1/\alpha + 1/\beta = 1$. Special provision for $\alpha \in \{0, 1\}$ must obviously be made; we have already considered $\alpha = 1$, MLE case which corresponds to e^{-g}, and we will now consider other cases, beginning with $\alpha = 0$.

3.1. **Rényi $\alpha = 0$.** In the primal formulation $\alpha = 0$ replaces e^{-g} by $\log g$. This allows us to maintain the exponential cone formulation except that now the roles of t and ν are reversed and we have instead of (1),

$$(2) \qquad \begin{pmatrix} 0 & e_i^\top \\ 0 & 0 \\ e_i^\top & 0 \end{pmatrix} \begin{pmatrix} \nu \\ t \end{pmatrix} + \begin{pmatrix} 0 \\ 1 \\ 0 \end{pmatrix} \in \mathcal{K}$$

implemented in R with,

```
P$F <- sparseMatrix(c(seq(3,3*n, by = 3), seq(1, 3*n, by = 3)),
       c(1:n, (n+1):(2*n)), x = rep(1,2*n))
P$g <- rep(c(0,1,0), n)
P$cones <- matrix(list("PEXP", 3, NULL), nrow = 3, ncol = n)
rownames(P$cones) <- c("type", "dim", "conepar")
```

3.2. **Rényi $\alpha \in (0, 1)$.** The remaining Rényi formulations all require Mosek V9 "power cones:" in dimension 3 the canonical power cone takes the form,

$$\mathcal{K} = \{x \in \mathbb{R}^3 | x_0^\alpha x_1^{1-\alpha} \geq |x_2|, x_0, x_1 \geq 0\}.$$

We first consider $\alpha \in (0, 1)$ which includes the important Hellinger case $\alpha = 1/2$. Again we introduce auxiliary variables, $t \in \mathbb{R}^n$, and would like to impose the condition $\nu^\beta \geq t$ which can be written as $\nu^\alpha t^{1-\alpha} \geq 1$, and implemented in R as,

```
P$F <- sparseMatrix(c(seq(1, 3 * n, by = 3), seq(3, 3 * n, by = 3)),
       c(1:n, (n + 1):(2 * n )), x = rep(1, 2 * n))
P$g <- rep(c(0, 1, 0), n)
P$cones <- matrix(list("PPOW", 3, c(alpha, 1 - alpha)), nrow = 3, ncol = n)
rownames(P$cones) <- c("type", "dim", "conepar")
```

3.3. **Rényi $\alpha > 1$.** Pearson fidelity, $\alpha = 2$ is the primary case of interest when $\alpha > 1$. If we write $\gamma = 1/\alpha$, then we can implement the power cone constraint $\nu^\gamma t^{1-\gamma} \geq 1$ by replacing α by $1/\alpha$.

3.4. **Rényi $\alpha < 0$.** Finally, in the netherworld of $\alpha < 0$ we can replace α by $1/(1-\alpha)$.

TABLE 1. MSE Performance of Rényi alternatives setting 1

n	1	2	0.5	0	-0.5
50	0.991	0.936	0.936	0.895	0.928
100	0.927	0.907	0.907	0.895	0.915
200	0.880	0.879	0.879	0.906	0.897
400	0.866	0.884	0.884	0.947	0.910
500	0.841	0.861	0.861	0.934	0.889
1000	0.842	0.873	0.873	0.968	0.904
2000	0.832	0.869	0.869	0.979	0.902

TABLE 2. MSE Performance of Rényi alternatives setting 2

n	1	2	0.5	0	-0.5
50	0.553	0.609	0.609	0.671	0.668
100	0.484	0.570	0.570	0.739	0.654
200	0.482	0.592	0.592	0.840	0.689
400	0.460	0.591	0.591	0.901	0.695
1000	0.447	0.580	0.580	0.941	0.688
2000	0.452	0.590	0.590	0.973	0.698

4. SIMULATIONS

To compare performance of the Rényi alternatives to the NPMLE we considered two distinct simulation settings. In the first $Y_i = \mu_i + u_i$ with $\mu_i \sim U[5, 15]$, $u_i \sim \mathcal{N}(0, 1)$ and several sample sizes. In each case we compute mean squared error of the posterior mean predictions of the μ_i's, and report the results based on 200 replications in Table 1.

The second setting is the same except that $\mu_i \in \{0, 3\}$ with equal probability. Again MSE based on 200 replications are reported in Table 2. Although there is some evidence that for small sample sizes, the MLE is bested by other Rényi alternatives for the Uniform setting, the discrete setting shows the NPMLE to be totally dominant.

REFERENCES

KOENKER, R., AND I. MIZERA (2018): "Shape Constrained Density Estimation Via Penalized Rényi Divergence," *Statistical Science*, 33, 510–526.