OPTIMAL TRANSPORTATION

ROGER KOENKER

Abstract. A brief introduction to the multitudinous charms of optimal transportation.

1. Introduction

Since optimal transportation seems to be popping up all over the place in economics and many other fields, I thought it might be worthwhile to try to give a very brief introduction so you can at least nod and say, "oh, yeah, we did some of that with Koenker in his topics course."

2. The Classical Monge-Kantorovich Problem

We have a pile of sand, and a hole, both of volume $1m^3$. And we have a spoon. Our task is to transport the sand, with the spoon, from the pile to the hole. The cost of transport, for lack of any more imaginative idea, is $c(x, y) = |x - y|$. We must design a transport plan, ideally of the form $y = T(x)$ that describes how much sand we should take from each point x to be placed at each point y. In this form the problem was posed by Gaspard Monge in 1781, but despite some 19th century progress, no satisfactory solution emerged until the breakthrough of Kantorovich in 1941. Although Kantorovich's innovation and subsequent development of linear programming provided a basic apparatus for the solution of such problems, it is only within the last 20 years or so that the optimal transportation (OT) literature has exploded as researchers in a wide variety of fields recognized its applicability.

In economics the most immediate application is matching models: We have a pile of workers who want/need jobs in a hole of firms. Another prominent example is hedonic demand analysis in which a pile of consumers would like to acquire one of the differentiated products in a shopping hole. I won't venture into these topics, the elegant forthcoming monograph of Galichon (2016) provides extensive illustrations and references. Further mathematical details are provided by Villani (2003) and Villani (2008).

To make things more concrete let's consider a more explicit version of the Kantorovich problem. I'll adapt slightly a toy problem from Rolet et al. (2016), the paper that recently revived my interest in optimal transport. In this example our pile of sand is a histogram, randomly generated from a random sample from a Gaussian mixture distribution. Roughly speaking, the mixture is composed of three Gaussians with different locations and amounts of mass. More precisely, histograms are generated by the following R code.

Version: April 14, 2016.

```r
# Data Generation for Toy Example in Rolet, et al
DGP <- function(mu = c(-6,0,6), n = 1000, N = 100, vmu = 2){
    histo <- function(y, x)
        hist(y, x, plot = FALSE)$density
    x <- seq(-15, 15, len = 101)
    U <- matrix(runif(3 * N), N, 3)
    Y <- matrix(0, n, N)
    for(i in 1:N)
        Y[,i] <- sample(mu, n, prob = U[i,], replace = TRUE) +
            rnorm(1, sd = sqrt(2)) + rnorm(n)
    H <- apply(Y, 2, histo, x = x)
    H <- H/apply(H,2,sum)
    xm <- (x[-1] + x[-length(x)])/2
    M <- abs(outer(xm, xm, "-"))
    M <- M/median(M)
    list(x = x, H = H, M = M)
}
```

In our simplest exercise we generate two of these random histograms and then try to design a transport plan between them. The concise statement of the problem is

$$\min_{Z}\{\mathrm{Tr}M^\top Z \mid Z1 = p, Z^\top 1 = q\},$$

where p and q are m vectors representing the mass of the histogram bars for the source and destination histograms, respectively. The matrix M represents the distances between the pairs of histogram bars, i.e., $M_{ij} = |x_i - y_j|$. The trace formulation of the objective is simply the discrete analogue of the continuous formulation,

$$\min_{\pi \in \Pi(\mu,\nu)} \int c(x,y)d\pi(x,y),$$

where $\Pi(\mu,\nu) = \{\pi \in P(X \times Y) | \pi(A \times Y) = \mu(A), \pi(X \times B) = \nu(B)\}$ for all Borel sets A and B. The matrix Z in the discretized problem plays the role of π in the continuous formulation, indicating how much mass is to be transferred from points x to points y. Clearly, we have a linear program: a linear objective function subject to polyhedral constraints. In the matrix form it perhaps doesn't yet look quite like a proper LP, but if we write instead.

$$\min_{z}\{c^\top z | Az = b, z \geq 0\},$$

where $c = vec(m)$, $z = vec(Z)$, and

$$A = \begin{bmatrix} 1^\top \otimes I \\ I \otimes 1^\top \end{bmatrix}, \quad b = \begin{bmatrix} p \\ q \end{bmatrix},$$

everything looks more conventional and can be easily inserted into a standard LP optimizer. The next code chunk illustrates how this looks in R using the Rmosek interface to Mosek. This approach is fine for small problems like our toy example when we have only 100 histogram bins, so z has only 10,000 elements. However, larger problems require different computational strategies.

```
require(Rmosek)
## Loading required package:  Rmosek
## Loading required package:  Matrix
D <- DGP(N = 2)
x <- D$x
H <- D$H
par(mfrow = c(1,3))
for(i in 1:2){
    main <- paste("Histogram", i)
    plot(stepfun(x[-c(1,100)], H[,i]), main = main, do.points = FALSE)
}
n <- nrow(D$H)
f <- c(D$M)
one <- matrix(1, 1, n)
A <- rbind(kronecker(one, Diagonal(n)),kronecker(Diagonal(n), one))
b <- c(D$H)
Aq <- NA
bq <- NA
lb <- rep(0, n^2)
ub <- rep(Inf, n^2)
P <- mosek_lptoprob(f,Aq,bq,A,b,lb,ub)
r <- mosek(P, opts = list(verbose = 0))
X <- matrix(r$sol$itr$xx,n,n) * 1000
xm <- (x[-1] + x[-length(x)])/2
contour(xm, xm, X, nlevels = 10, main = "Optimal Transport Plan")
```

In Figure 1 we illustrate one realization of our toy problem. There are two histograms and the rightmost panel depicts contours of the optimal transport plan. It is useful to try rerunning the example several times to get a feeling for how the optimal plan adapts to the variety of the histograms. The contour plot is obviously poor, a better visualization device is `rgl.surface` from the **rgl** package. This can be used to rotate the 3d plot as will be demonstrated in class.

The Monge-Kantorovich problem is really just the tip of the iceberg for current applications of OT methods. I'll conclude with two somewhat more elaborate examples to illustrate the scope of these methods. Another way to think about the Kantorovich problem is that it defines a distance between two distributions. When the distributions are just histograms on the real line as above there is a nice way to characterize this distance. If we plot the two cdfs, the distance is just the L_1 distance between the two cdfs. This distance has a variety of names in the literature including (in honor of Monge) earth-movers distance, Mallows distance and Wasserstein distance. The latter name seems to be the new

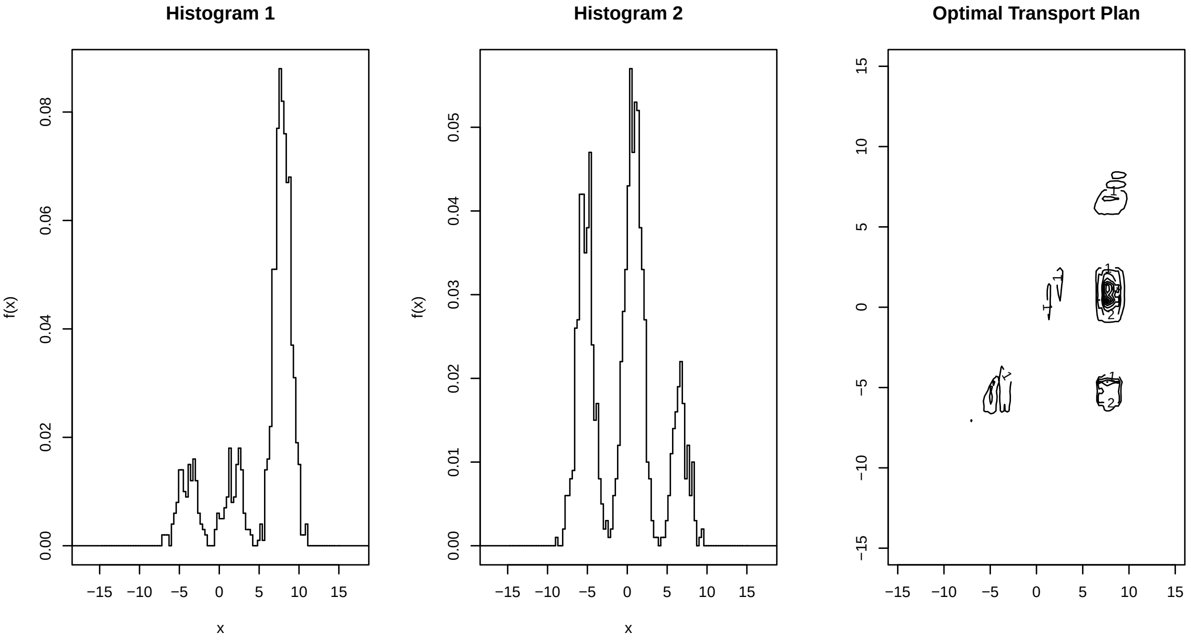

FIGURE 1. Random Histograms and Optimal Transport Plan

standard. By a somewhat obvious geometric argument one can also see that this is equivalent to the L_1 distance between the two quantile functions, providing a nice connection to quantile treatment effects.

The really remarkable aspect of all this is that it extends to higher dimensions in a natural way provided we have a way to measure distance between objects. For example for 2d histograms we can use Euclidean distance between bins, and the problem setup is identical. This is the beginning of an exciting story of using OT to define various notions of multivariate quantiles.

3. Wasserstein Interpolation and Mixtures

Rather than pursue this line of inquiry, I'd like to describe some recent work on using Wasserstein distance for interpolation and mixture modeling. First let's consider interpolation. For this it is useful to consider a slightly generalized form of the OT problem regularized by an entropy penalty,

$$W_\gamma(p,q) = \min_{Z \in \mathcal{Z}(p,q)} \{\mathrm{Tr} M^\top Z - \gamma \eta(Z)\},$$

where as before $\mathcal{Z}(p,q) = \{Z \in \mathbb{R}_+^{mm} | Z1 = p, Z^\top 1 = q\}$, and $\eta(Z) = \sum z_{ij} \log z_{ij}$ is a (Shannon) entropy penalty design to regularize the problem, i.e., to produce a somewhat better behaved Z^*. For $\gamma = 0$ we are obviously back to the vanilla Kantorovich problem. Now to simplify the notation, define, $H_q(p) = W_\gamma(p,q)$ and suppose that we have a bunch of q_i's and we would like to find a p that solves,

$$\min_{p \in \mathcal{S}} \sum \lambda_k H_{q_k}(p) \tag{1}$$

where $\mathcal{S}$ denotes the unit simplex. This looks like quite a formidable task since it appears to require many OT solutions for each trial value of p. Fortunately, some convenient duality tricks can be brought to bear, and solutions are quite easily computed. So what is it good for? Frankly, I'm not quite sure at this point, but it is certainly true that we are in the midst of a much more flexible framework for multivariate analysis than we have seen before when we were mired in the Gaussian swamp. In the Gaussian setting we know how to average vectors, possibly weighting by inverse covariance matrices as in 508, but what if we wanted to average sometime more exotic like histograms, or like cows, and ducks and donuts?

For triangles we may recall that there is a notion of the centroid or barycenter, the family of weighted Wasserstein distances obtained by solving (1), provides a way to average, or interpolate, quite arbitrary objects, again provided that we have a meaningful notion of distance and a reasonable regularization strategy.

As a final example I'd like to briefly describe a mixture problem that employs OT ideas to finding components of quite general mixture problems. To fix ideas at the outset, suppose we have a big-ish collection of histograms, like those we began the lecture with. We suspect that they are all generated from a small-ish dictionary of basic components. In our artificial we know this to be true – all of then are built from three Gaussian component distributions. Can we discover this without imposing any a priori knowledge about Gaussian likelihoods, as we have earlier in the Kiefer-Wolfowitz setup. This is the subject of Rolet et al. (2016)

As with the barycenter problem we rely on regularized Wasserstein distances, but the trick is that rather than finding an average we try to estimate a factor model, that is we try to solve,

$$\min_{\Lambda, D} \sum H_{q_k}(D\lambda_k) + P(\Lambda, D),$$

where $q_i, i = 1, ...n$ are our data in the form, say, of histograms. The matrices D and Λ satisfy our usual constraints, so $D\lambda_k \in \mathcal{S}$ for all column vectors λ_k of Λ. What about the penalty term? Rolet et al. (2016) use our already old favorite Shannon entropy with distinct multipliers for $\eta(D)$ and $\eta(\Lambda)$, but it seems that the game is open to lots of other choices, provided we maintain the convexity of the problem. At this point we should hasten to add that the full problem is no longer convex, however it is biconvex, i.e., convex in D given fixed Λ, and vice versa. Implementation details remain a bit obscure, but seem to involve nothing out of the ordinary realm of convex analysis. And of course, there is nothing really special about histograms, the same techniques can be adapted to a host of other problems. Rolet et al. (2016) mention work on facial recognition and text classification.

References

Galichon A. 2016. *Optimal Transport Methods in Economics.* Princeton U. Press.

Rolet A, Cuturi M, Peyré G. 2016. Fast dictionary learning with a smoothed wasserstein loss. *Proceedings of the 19th International Conference on Artificial Intelligence and Statistics* .

Villani C. 2003. *Topics in optimal transportation.* 58. American Mathematical Society.

Villani C. 2008. *Optimal transport: old and new*, volume 338. Springer.