Moving heaven and earth: Distances between distributions

Suresh Venkatasubramanian

Abstract

This column comes in two parts. In the first, I discuss various ways of defining distances between distributions. In the second, Jeff Erickson (chair of the SoCG steering committee) discusses some matters related to the relationship between ACM and the Symposium on Computational Geometry.

1 Introduction

One of the most basic tests in statistics is the two-sample test: "Using samples drawn from two distributions, determine if they are the same or not". If the two distributions come from recognizable families, then the distribution of the difference in means can usually be characterized, giving a test with guaranteed confidence bounds (the two-sample t-test is an example of this).

If the distribution families are unknown, this is a tricky problem even if the two distributions are drawn from the same (finite) domain. In this case, the test statistic, instead of being a difference between means, is a *distance between distributions* [10]. In property testing for example, one might assume that both distributions are defined over $[1 \dots n]$, and the goal is to determine with few samples whether the ℓ_1 distance (also called the *total variation distance*) between the two distributions is small [4].

The problem remains the same even when the domain is a high dimensional space (as is common in data mining): given samples from distributions defined over some domain, compute some estimate of the distance between the distributions.

But what distance do we measure ? In this article, I'll explore a few of the ways in which people commonly try to compare distributions, and point out some surprising geometric connections between them that have been developed within the machine learning community and might not be well known to a TCS audience.

I'll also attempt a bolder claim: that we should revisit the default choice for comparing distributions over metric spaces, because there are other measures that serve the same purpose but are much easier to work with.

2 Information Geometry

The general area of building a geometry (and therefore distances) on distributions is called *information geometry*. This is a vast area that is covered comprehensively in the book by Amari and Nagaoka [1]. While it would take a whole other column to describe the ideas there, the basic concept is to view a parametrized family of distributions as a statistical manifold and use differential geometry to define distances between two distributions (points on the manifold). Some of the more popular distances between distributions that arise this way are the Bregman divergences (and the Kullback-Leibler divergence in particular), the Hellinger distance, and the so-called α-divergences.

Information geometry yields interesting distances. For example, the space of Gaussian distributions on the line is endowed with the geometry of hyperbolic space (with coordinates μ, σ), and the space of multinomial distributions yields the Kullback-Leibler divergence. However, all these distances are defined in terms of *parameters* of the underlying distributions. Thus, if you only have access to samples from the distributions, you're forced to estimate the parameters before you can compute the distance. For some families, this is easy. In the case of Gaussians, one can compute estimates for the mean and variance from the samples, and use these estimates to compute an estimated distance. With some work, we can even get confidence bounds on the distance produced in terms of the number of samples used.

What if we don't know the underlying distribution ? In that case, all we might possess are samples from the underlying domain. If the domain is finite, we might still be able to derive something meaningful. For example, a common technique used to compare distributions over $[1 \dots n]$ is to compute the normalized frequency counts $p_i = p(X = i)$ and $q_i = p(Y = i)$ for each distribution and compute some distance between the vectors p_i and q_i such as the Kullback-Leibler divergence or the total variation distance.

But what if the underlying domain is not $[1 \dots n]$? What if it's the line, or even the interval $[0, 1]$? You could always discretize the region into bins, but there are two problems with this. Firstly, the number of bins needed grows exponentially with the dimension of the space. Secondly, where does one place the bin boundaries ? As we can see in Figure 2, merely shifting the grid origin can yield two completely different frequency count vectors for the same sample !

The problem of course is the binary nature of a bin: points are either in or out. But a sample indicates some probability mass in a region, and we don't expect that mass to be isolated. So what we'd like to do instead is *use the geometry of the underlying domain* to determine when two samples might be "close" and therefore similar.

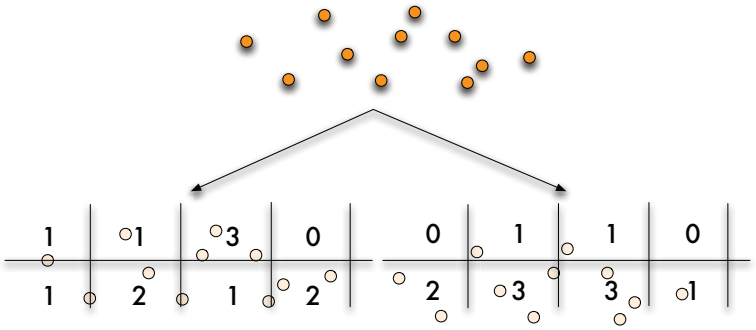

Figure 1: Inconsistent binning yields different histograms

3 Comparing clouds of objects

Let us now move to a different problem. In the previous section, we started off with the desire to compare distributions, and ended up trying to incorporate the geometry of the underlying domain into the calculation. Suppose we needed to do the opposite ?

Let's say we have a collection of objects that inhabit a metric space (X, d). A very natural operation is to ask how we might compare *distributions* over the metric space. One such example that comes up in practice is comparing two color spectrums. Suppose I have two images with each pixel representing a point in some color space. I can build a frequency histogram for each image where each entry is the number of pixels having a particular color value. The color space itself might have an underlying metric [9][1]. How do I compare these histograms ?

One idea that we might try is to define a metric over distributions by asking for the expected distance between points sampled from the two distributions. But with this approach the distance from a distribution to itself would be nonzero !

Once again, we're presented with the problem of defining a distance between two distributions whose underlying domain is endowed with metric structure.

4 The earthmover distance

The earthmover distance [22] (or EMD) is one solution to this problem. It is a very natural approach to defining a distance between distributions on a metric, so much so that it's been defined over and over again ever since it was first proposed by Monge in 1781 [17, 16]. Possibly a more appropriate name for the distance might be the Monge-Kantorovich-Wasserstein-Mallows-Gini-D'All Aglio distance [23].

[1]Many thanks to Allison Sekular [25] for providing a wealth of background information and pointers on color space metrics.

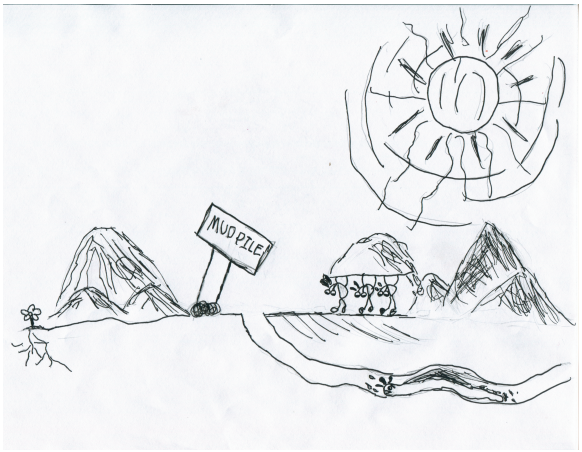

Figure 2: A rendering of the earthmover distance by the author's 7 year old son.

Formally, the earthmover distance is defined as follows: given distributions p and q defined over a metric space (X, d), minimize the quantity

$$d_{\text{EMD}}(p, q) = \min_W \int_X W(x, x')d(x, x')$$

with the constraint that the *transport function* W has marginals p, q:

$$\int_X W(x, y)dy = p(x)$$
$$\int_X W(x, y)dx = q(y)$$

A physical interpretation of the distance (which also gives it its name[2]) is that p and q are like piles of dirt placed at different locations in X, and the goal is to move dirt around to make p look like q. The effort involved in carrying a pile of dirt of mass $W(x, y)$ from x to y is $W(x, y)d(x, y)$, and no dirt is created or lost in the process[3].

5 Integral Probability Metrics

While the earthmover distance (or more generally the transportation distance) was first defined and studied in the context of optimal transportation, it turns out that it's just one

[2]Rubner et al. attribute [22] the choice of name to a 1994 communication from Jorge Stolfi. It is admittedly the most picturesque way of describing the distance.

[3]There are other variants where the total mass of p and q may be different from each other: we do not consider them here.

member of a large class of metrics on distributions that have been studied extensively in statistics. These are called the *integral probability metrics* [18][4].

One way of thinking about how to compare two distributions is by integrating them. We can define a class of *test functions*, and integrate the distributions with respect to each member of the class, taking the maximum difference between the resulting values. Formally, let $\mathcal{F}$ be a class of integrable functions. We can then define a distance between distributions as

$$d_{\mathcal{F}}(p,q) = \sup_{f \in \mathcal{F}} \left| \int f dp - \int f dq \right|$$

If this seems a little mysterious, consider the following geometric analog. Suppose I want to define a distance between two points in the plane. I decide to use the inner product on $\mathbb{R}^2$ for this purpose. I fix a set of directions S, and then define the distance

$$d_S(p,q) = \sup_{\mathbf{v} \in S} |\langle v, p \rangle - \langle v, q \rangle|$$

Choosing different S yields different distances. For example, if $S = \{(0,1),(1,0)\}$, then $d_S(p,q) = \|p-q\|_\infty$. If $S = \{(1/\sqrt{2}, 1/\sqrt{2}), (1/\sqrt{2}, -1/\sqrt{2})\}$, then $d_S(p,q) = \|p-q\|_1$. Of course if S is the set of all directions in the plane, then $d_S(p,q) = \|p-q\|_2$.

Similarly, choosing different function classes $\mathcal{F}$ yields different distances between distributions, and that's where things start getting interesting. Let $\|f\|_\infty = \sup_{x \in X} |f(x)|$ and let $\mathcal{F}_{TV} = \{f \mid \|f\|_\infty \leq 1\}$. Then $d_{TV}(p,q) = \|p-q\|_1$ (the total variation distance). Setting $\mathcal{F} = \{\mathbf{1}_{(-\infty,t]}\}$ instead yields the *Kolmogorov distance* between distributions, which is the max norm of the difference between their cumulative distributions.

The earthmover distance can also be written as an integral probability metric. Let $\|f\|_L \triangleq \sup\{|f(x) - f(y)|/d(x,y), x \neq y \in X\}$ be the *Lipschitz semi-norm* of a real-valued f. It is a consequence of a celebrated result by Kantorovich and Rubinstein [14] that the earthmover distance can be written as $d_{\mathcal{F}}$ where $\mathcal{F} = \{f \mid \|f\|_L \leq 1\}$. For finite domains this relation follows from linear programming duality.

5.1 Kernels

Of the distances described above, the earthmover distance is the only one that actually incorporates a metric structure on the base space X. Suppose we endow X with a little more structure, namely a *kernel*. Kernels have a long and illustrious history in machine learning, and it would be folly to try and give a complete description of kernels here. I'll point you to Hal Daumé's excellent "From Zero to RKHS in twelve steps" [8] for a quick and yet comprehensive overview of the mathematical foundations of kernels, and will limit myself to definitions here.

[4] Earlier work by Zolotarev [32] had referred to them as *probability metrics with a ζ-structure.*

A (positive definite) kernel $K : X \times X \to \mathbb{R}$ is a bounded function with the property that for all n and all sequences $x_1, x_2, \ldots, x_n \in X$, the matrix $\mathbf{K} = \{K(x_i, x_j)\}$ is positive definite. What makes kernels interesting is that we can associate with K a feature map $\Phi_K : X \to \mathcal{H}$ from X to a special Hilbert space (called a *reproducing kernel Hilbert space*) with the property that

$$K(x, y) = \langle \Phi_K(x), \Phi_K(y) \rangle_{\mathcal{H}}$$

where $\langle \cdot, \cdot \rangle_{\mathcal{H}}$ is the inner product associated with $\mathcal{H}$.

The geometric significance of this is immense: any space that admits a kernel can be "straightened out" by lifting into a (potentially infinite-dimensional) space that "looks Euclidean".

With that intuition in mind, let us return to integral probability metrics, and define the class $\mathcal{F}_k = \{f \mid \|f\|_{\mathcal{H}} \leq 1\}$. Recall that $\|f\|_{\mathcal{H}} = \int f(x) f(y) k(x, y) dx dy$. Note that instead of using the Lipschitz semi-norm defined in terms of the underlying metric, we're using a Hilbert space norm defined in terms of the underlying kernel.

As worked out by Gretton et al. [11], something very interesting happens when we try to do the maximization. The resulting expression for d_K is a closed form formula with no dependence on $\mathcal{F}_k$ at all ! Let $\mu_p = \int \Phi(x) dp \in \mathcal{H}_k$. Then

$$d_K(p, q) = \|\mu(p) - \mu(q)\|$$

There is a simple expression for d_K that does not require the feature map Φ. Noting that $\|\mu(p) - \mu(q)\|^2 = \|\mu(p)\|^2 + \|\mu(q)\|^2 - 2\langle \mu(p), \mu(q) \rangle$, we get

$$d_K^2(p, q) = \int K(x, y) p(x) p(y) dx dy + \int K(x, y) q(x) q(y) dx dy - 2 \int K(x, y) p(x) q(y) dx dy$$

One way to get a handle on the kernel distance is to examine the form it takes for specific kernels. Let's take the easiest kernel of all: $K(x, y) = \mathbf{1}_{x=y}$. Note that for any sequence of points $x_1, \ldots, x_n$, the resulting matrix $\mathbf{K}$ is the identity matrix, and so this kernel is positive definite.

Suppose our distributions p, q consist of the uniform distribution over finite sets A, B respectively. Then $p(x) = \frac{\mathbf{1}_{x \in A}}{|A|}$ and $q(x) = \frac{\mathbf{1}_{x \in B}}{|B|}$. Substituting the expressions in the above equation (and squaring both sides for convenience), we get

$$d_K^2(p, q) = \frac{A \Delta B}{|A||B|}$$

which is a scaled version of the Jaccard distance between the two sets. The intuition is that a kernel function $K(x, y)$ *blurs* the sharp threshold of equality in set intersections. For more on the kernel distance (and its generalization to distances between geometric measures) you can look at the article that Jeff Phillips and I wrote [20].

5.2 A computational perspective

Applying an algorithmic lens to these distances yields a number of interesting questions. How hard is it to compute the distance given a collection of samples ? What is the geometric nature of the resulting metric (does it embed in nice spaces, for example) ? How easy is it to approximate or estimate over large data sets ?

It is here that the kernel distance proves to be more effective than the earthmover distance. I'll discuss these points one by one.

Computing the distance. Computing the earthmover distance is what is called the *assignment problem*, and it can be solved using the standard Hungarian[5] primal-dual algorithm. If the original matrix of distances is $n \times n$, the Hungarian algorithm can be implemented to run in time $O(n^3)$.

This assumes the underlying metric is a black box accessed only via distance queries. The more we know about the underlying metric, the easier it gets to compute the EMD. For example, if the distributions are defined on $\mathbb{R}$ with $d(x, y) = |x - y|$, then the EMD reduces to computing the ℓ_1 distance between the two distributions and can be computed in linear time.

Because of the cost of computing the earthmover distance exactly, much effort has gone into finding fast approximations. These typically work by embedding the distance approximately into a space like ℓ_1 where distances can be computed cheaply, and I'll say more about that in what follows. For the case of point sets in a low-dimensional Euclidean space, Indyk gave a constant-factor approximation running in near-linear time [12].

The kernel distance by its definition is a sum over pairs of points and can be computed in $O(n^2)$ time. This is an "easy" $O(n^2)$, in that the individual terms can be computed independently and added – there just happen to be $\Omega(n^2)$ of them. As with the earthmover distance, the best approximations of the kernel distance are obtained by using embeddings, which I discuss next.

The geometry of the metric. Embedding a distance in a geometry is often more important than merely computing it. If I can construct a mapping from a metric space to a vector space endowed with some metric structure, then I can find averages of points, cluster them, process general near neighbor queries, and so on. It may not always be possible to do this embedding isometrically, so we often talk about the *distortion* of an embedding: the maximum "stretch" or "shrink" incurred by any particular pair of points when transformed from the source to target metric space.

There's been a considerable amount of work on embedding the earthmover distance. In the most general case (when the underlying metric is arbitrary), Charikar [6] showed

[5] One wonders when we should start calling it the German algorithm given that Jacobi apparently invented it [16] well before König and Egerváry, who in turn developed it before Harold Kuhn.

how to use LSH-like methods to embed the EMD into the underlying metric space with distortion $O(\log n \log\log n)$. Here n is both the size of the underlying metric space as well as the size of the support for the distributions. Building on this, Indyk and Thaper showed how to embed the ℓ_2-EMD into ℓ_1^d with distortion $O(d)$. Unfortunately, this turned out to be the limit of what can be done in general: Khot and Naor [15], using Fourier techniques, showed that this bound is tight[6]. In the face of this bound, researchers have looked at embeddings when the input space is limited [2], as well as embeddings into more general spaces [31]. While I won't go into it in detail, *sketching* has also provided a way to embed the EMD into a geometry. Formally, sketching maps distributions to a space in which an estimate of the EMD can be computed easily. The sketches themselves are small and can often be computed using streaming methods; however, the estimation might not compute a metric and so this is not a *metric* embedding. There has also been work on embedding the EMD into a wavelet-based weighted ℓ_1 space[26, 7]: the resulting distance is within a constant factor of the EMD and can be computed efficiently using wavelet transforms.

For the kernel distance however, the story is much friendlier. From the definition, we see that the kernel distance maps isometrically into a (possibly infinite-dimensional) Hilbert space. This ℓ_2-like nature of the kernel distance makes estimating it significantly easier. In particular, as long as we can extract the eigenspace of the kernel function, we can construct a high-quality embedding into a finite-dimensional ℓ_2 space by projecting the eigenspace (akin to a singular value decomposition).

For the kernels that are typically used in practice, like the Gaussian kernel $k(x,y) = \exp(-\|x-y\|^2/\sigma^2)$, the eigenspace can be characterized using tools from harmonic analysis. Using a well-known theorem of Bochner [5] characterizing positive definite functions as Fourier transforms of probability measures, Rahimi and Recht [21] showed that *translation-invariant* kernels (i.e kernels of the form $K(x,y) = k(x-y)$) can be embedded in a finite dimensional ℓ_2 space of dimension $1/\varepsilon^2 \log(1/\delta)$ with error ε with probability $1-\delta$.

$$\langle f(x), f(y)\rangle - K(x,y) \leq \varepsilon$$

This is of course reminiscent of the Johnson-Lindenstrauss Lemma and is proved using similar techniques. Intuitively, the feature map lifts the kernel to the appropriate infinite dimensional Hilbert space on which the "standard" J-L technique can then be applied.

Once the kernel function itself can be approximated (and note that this is stronger than merely approximating the distance), there are variety of approaches to estimating the distance, including the use of core sets to compute small-sized sketches of the distributions that embed linearly in the approximate kernel space [13]. All of this yields near-linear time $1+\varepsilon$-approximation algorithms for computing the kernel distance, or *constant time* algorithms for estimating the distance after a near-linear amount of preprocessing.

[6] Naor and Schectman[19] showed later that even for points on the $n \times n$ grid, there is a lower bound of $\sqrt{\log n}$.

5.3 A Statistical Perspective

We've looked at integral probability metrics, and the EMD and kernel distance in particular, from the perspective of computing them. But what does the answer mean ?

These measures are intended to compare distributions, and are estimated from samples. An estimator for a quantity is *consistent* if in the limit as the number of samples goes to infinity, the estimator converges in probability to that quantity. As part of a long investigation into the statistical properties of integral probability metrics, Sriperumbudur et al. [27] showed that the EMD, the kernel distance and another IPM called the Dudley metric were *strongly consistent*: namely that the convergence happened almost surely, rather than just in probability. Moreover, they also show that the rate of convergence for the kernel distance is *independent of the dimension*, in contrast to the convergence rate for the EMD.

The kernel distance also bears a strong resemblance to another estimate of distance between distributions that has become popular in recent years. The *energy distance*[28, 29, 3] was defined as a high-dimensional test of equality between distributions, and the *distance covariance* [30] was defined subsequently as a way to test independence of random vectors. Sejdinovic et al. showed [24] that these distances (appropriately parametrized) are equivalent to the kernel distance, thus establishing another link between integral probability metrics and other methods for testing distance between distributions. The fact that kernels can be defined on many different kinds of structured data allows these measures to be applied in a variety of settings, instead of just in $\mathbb{R}^d$.

6 Conclusions

The choice of distance function when comparing geometric structures is a modeling choice, guided by the shape of the geometry you want to induce, as well as the kinds of computation you want to perform. It's not hard to define a distance function that seems to capture the shape of the space you're dealing with, but that can't really be computed very well. At that point, you have to decide whether your choice is really crucial or not.

The argument presented here is not that the earthmover distance is a bad choice for comparing distributions: far from it ! It is merely that the use of the earthmover distance is a choice, and in particular it is the choice of one member of a large family of distances, and there are other members with equally attractive properties. While one potential limitation of distances like the kernel distance is that they need more structure on the underlying space, this is not a real restriction in many settings. For many common kinds of data, there are well known kernels that can be used as the basis for defining a distance over distributions.

7 SoCG News

This section is written by Jeff Erickson, who is the chair of the Symposium on Computational Geometry (SoCG) steering committee. ACM (and SIGACT) and SoCG have had some "relationship troubles" in the past few years, and Jeff's note summarizes the history of the relationship and the current state of affairs.

Since its inception in 1985, the Symposium on Computational Geometry (SOCG) has been affiliated with the Association for Computing Machinery, through its special interest groups SIGACT and SIGGRAPH. Over the past three years, the SOCG community has been considering leaving ACM and organizing the conference independently. The community has already voted on this issue twice, both times with a majority of votes favoring at least partial independence from ACM, but in light of recent developments within ACM, the steering committee has agreed to hold a third and final vote this fall.

Although there are several interrelated issues driving the vote, the key issue is our ability to organize SOCG as an "in-cooperation" conference outside the United States, rather than a "sponsored" conference. In-cooperation conferences are financially independent; they do not receive financial backing or conference services from ACM, but they also avoid the administrative and financial overhead of ACM sponsorship. There is general agreement that in-cooperation status is both easier and cheaper than ACM sponsorship when SOCG is organized outside the US.

Unlike most ACM-sponsored conferences, SOCG is regularly held outside North America; the computational geometry community has significant representation on at least five continents. In the last ten years, the conference has been held five times in the US, and once each in Italy, Denmark, France, South Korea, and Brazil. SOCG 2014 is being organized in Kyoto, Japan, and SOCG 2015 will be organized in Eindhoven, in the Netherlands.

SOCG has been organized twice as an in-cooperation conference, first in Korea in 2007, and again in Denmark in 2009. In both cases, the organizers were able to secure enough local support to keep registration costs low. In 2011, when SOCG was held in Paris, the local organizers again applied for "in-cooperation" status, with the approval of the steering committee. The request was again approved by the SIGACT and SIGGRAPH executive boards, but denied by ACM, who stated that in-cooperation status was no longer an option for SOCG. The unexpected requirement of ACM sponsorship created a significant administrative burden and higher registration fees.

In response, the community held two electronic votes through the computational geometry community mailing list, which has more than 800 subscribers. The first vote in November 2011 offered three options; there were 29 votes to stay with ACM, 48 votes to retain ACM as a publisher but organize independently, and 47 votes to become completely independent and publish proceedings through LIPIcs. At SOCG 2012, the steering committee reported that ACM would not consider a publisher-only relationship, so a second vote was held in November 2011, again using the community mailing list. In the

second poll, there were 36 votes to stay with ACM and 50 votes to leave ACM.

When the results of the vote were announced, several prominent ACM members objected, in part because of the low voter turnout, and in part because ACM representatives were not given an opportunity to respond to statements provided to voters by the steering committee. Thanks to the efforts of SIGACT chair Paul Beame, ACM has agreed not to object to good-faith proposals for in-cooperation status outside the United States; in particular, SOCG 2014 will be organized in-cooperation.

In light of these legitimate criticisms and policy changes, the steering committee has agreed to hold a third and final vote, again with two options: stay with ACM, or leave ACM. The vote itself will take place in October 2013, with the results announced in early November. All members of the computational geometry community are strongly encouraged to participate.

To ensure that everyone has an opportunity to discuss the issues surrounding the vote, I have set up a discussion blog at `http://makingsocg.wordpress.com/`. In the weeks leading up to the vote, I will post summaries of most of the relevant issues, but I hope that most of the content will come from the research community and other stakeholders. In particular, I will invite posts and comments from past local organizers of SOCG; representatives of ACM, SIGACT, and SIGGRAPH (hopefully expressing their official positions); and representatives of other conferences, both with and without ACM affiliation. If you are interested in contributing, or if you have other questions about the vote, please send me email at jeffe@cs.uiuc.edu.

References

[1] S.-I. Amari and H. Nagaoka. *Methods of information geometry*. Oxford University Press., 2000.

[2] A. Andoni, P. Indyk, and R. Krauthgamer. Earth mover distance over high-dimensional spaces. In *Proceedings of the nineteenth annual ACM-SIAM symposium on Discrete algorithms*, SODA '08, pages 343–352, Philadelphia, PA, USA, 2008. Society for Industrial and Applied Mathematics.

[3] L. Baringhaus and C. Franz. On a new multivariate two-sample test. *Journal of Multivariate Analysis*, 88(1):190–206, 2004.

[4] T. Batu, L. Fortnow, R. Rubinfeld, W. D. Smith, and P. White. Testing that distributions are close. In *Foundations of Computer Science, 2000. Proceedings. 41st Annual Symposium on*, pages 259–269. IEEE, 2000.

[5] S. Bochner. Monotone funktionen, stieltjes integrale und harmonische analyse. *Math. Annal.*, 108:378–410, 1933.

[6] M. S. Charikar. Similarity estimation techniques from rounding algorithms. In *Proceedings of the thiry-fourth annual ACM symposium on Theory of computing*, STOC '02, pages 380–388, New York, NY, USA, 2002. ACM.

[7] R. R. Coifman and W. E. Leeb. Earth mover's distance and equivalent metrics for spaces with hierarchical partition trees. Technical Report YALEU/DCS/TR-1482, Yale University, 2013.

[8] H. Daumé. From zero to reproducing kernel hilbert spaces in twelve pages or less. `http://pub.hal3.name/daume04rkhs.ps`, 2004.

[9] R. Ennis and Q. Zaidi. Geometrical structure of perceptual color space is affine. *Journal of Vision*, 13(9):295, 2013.

[10] A. L. Gibbs and F. E. Su. On choosing and bounding probability metrics. *International Statistical Review*, 70(3):419–435, 2002.

[11] A. Gretton, K. Borgwardt, M. J. Rasch, B. Scholkopf, and A. J. Smola. A kernel method for the two-sample problem. In *NIPS*, pages 513–520, 2007.

[12] P. Indyk. A near linear time constant factor approximation for euclidean bichromatic matching (cost). In *Proceedings of the eighteenth annual ACM-SIAM symposium on discrete algorithms*, pages 39–42. Society for Industrial and Applied Mathematics, 2007.

[13] S. Joshi, R. V. Kommaraji, J. M. Phillips, and S. Venkatasubramanian. Comparing distributions and shapes using the kernel distance. In *Proceedings of the twenty-seventh annual symposium on Computational geometry*, SoCG '11, pages 47–56, New York, NY, USA, 2011. ACM.

[14] L. V. Kantorovich and G. Rubinstein. On a space of completely additive functions. *Vestnik Leningrad. Univ*, 13(7):52–59, 1958.

[15] S. Khot and A. Naor. Nonembeddability theorems via fourier analysis. In *Foundations of Computer Science, 2005. FOCS 2005. 46th Annual IEEE Symposium on*, pages 101–110, 2005.

[16] S. Martello. Jenö Egerváry: from the origins of the hungarian algorithm to satellite communication. *Central European Journal of Operations Research*, 18(1):47–58, 2010.

[17] G. Monge. Mémoire sur la théorie des déblais et des remblais. In *Histoire de l'Académie Royale des Sciences de Paris, avec les Mémoire de Mathématique et de Physique pour la même année*, pages 666–704. Paris, 1781.

[18] A. Müller. Integral probability metrics and their generating classes of functions. *Advances in Applied Probability*, pages 429–443, 1997.

[19] A. Naor and G. Schechtman. Planar earthmover is not in l_1. *SIAM J. Comput.*, 37(3):804–826, 2007.

[20] J. M. Phillips and S. Venkatasubramanian. A gentle introduction to the kernel distance. *arXiv preprint arXiv:1103.1625*, 2011.

[21] A. Rahimi and B. Recht. Random features for large-scale kernel machines. In J. C. Platt, D. Koller, Y. Singer, and S. T. Roweis, editors, *NIPS*. Curran Associates, Inc., 2007.

[22] Y. Rubner, C. Tomasi, and L. J. Guibas. The earth mover's distance as a metric for image retrieval. *International Journal of Computer Vision*, 40(2):99–121, 2000.

[23] L. Rüschendorf. Wasserstein metric. In M. Hazewinkel, editor, *Encyclopedia of Mathematics*. Kluwer Academic Publishers, 1998.

[24] D. Sejdinovic, B. Sriperumbudur, A. Gretton, and K. Fukumizu. Equivalence of distance-based and rkhs-based statistics in hypothesis testing. *arXiv preprint arXiv:1207.6076*, 2012.

[25] A. Sekular. On choices of color metrics. `https://plus.google.com/112165457714968997350/posts/QE6RoyEiX4T`.

[26] S. Shirdhonkar and D. Jacobs. Approximate earth mover's distance in linear time. In *Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on*, pages 1–8, 2008.

[27] B. K. Sriperumbudur, A. Gretton, K. Fukumizu, B. Schölkopf, and G. R. Lanckriet. Hilbert space embeddings and metrics on probability measures. *The Journal of Machine Learning Research*, 99:1517–1561, 2010.

[28] G. J. Székely and M. L. Rizzo. Testing for equal distributions in high dimension. *InterStat*, 2004.

[29] G. J. Székely and M. L. Rizzo. A new test for multivariate normality. *Journal of Multivariate Analysis*, 93(1):58–80, 2005.

[30] G. J. Székely and M. L. Rizzo. Brownian distance covariance. *The annals of applied statistics*, pages 1236–1265, 2009.

[31] E. Verbin and Q. Zhang. Rademacher-sketch: a dimensionality-reducing embedding for sum-product norms, with an application to earth-mover distance. In *Proc. 39th ICALP*, pages 834–845. Springer-Verlag, 2012.

[32] V. M. Zolotarev. Probability metrics. *Theory of Probability & Its Applications*, 28(2):278–302, 1984.