Geometric Comparison of Popular Mixture Model Distances

Scott A. Mitchell samitch@sandia.gov
Computing Research
Sandia National Laboratories
P.O. Box 5800, MS1316
Albuquerque, NM 87185-1316, USA

Editor:

Abstract

Triangular discrimination, Jensen-Shannon divergence, and the square of the Hellinger distance, are popular distance functions for mixture models, and are known to be similar. Here we expound upon their equivalence in terms of their functional forms after transformations, factorizations, and series expansions, and in terms of the geometry of their contours. The ratio between these distances is nearly flat for modest ratios of point coordinates, up to about 4:1. Beyond that the functions increase at different rates. We include derivations of ratio bounds, and some new difference bounds. We provide some constructions that nearly achieve the worst-cases. These help us understand when the different functions would give different orderings to the distances between points.

Keywords: Mixture Models, Geometry, Distance Functions, Theory

1. Introduction

Mixture models are ubiquitous in statistics and their applications. Mixture models express quantities whose components are positive and sum to one. They conveniently express a discrete probability distribution for exclusion settings, where probabilities sum to one. They also express fractions of a whole, e.g. they frequently arise after normalization. They are geometrically equivalent to points lying on a regular simplex. See Appendix A for how a mixture model arises in one information application.

A distance function measures how close two points are to one another. In clustering applications, points that are close to each other based on this distance are grouped together. Nearest-neighbors often play a special role. For a given point, different distance functions may give different orderings to the other points, and different clusters may result.

Triangular discrimination, Jensen-Shannon divergence, and the square of the Hellinger distance, are popular distance functions. There are others, but we focus on these three because they are known to be similar. The literature and folklore contain some relations, but these provide limited insight for the following reasons. The prior focus is on the most extreme results, worst case bounds, the maximum and minimum ratio of one distance to another. These are often given as a list of algebraic inequalities, without proof or even hints at reasons why the inequalities hold. We are interested in understanding which sets of points give rise to these extremes, and what we should expect in intermediate cases. We are interested in the geometry of the mappings underlying the functions, and their series expansions. These provide insight into the form and relation of the functions across all cases. Factorizations provides a simplification and parameterization of the bounds. Section 4.7 provides some new bounds on the difference between functions, i.e. $D_1 - D_2$.

These results provide some underpinnings for answering the question, "In what situations does it matter which distance function you choose?" using first principles rather than anecdotal case studies. That is, we explore the class of points for which the different distances would give different

answers, e.g. to nearest-neighbor queries or constant-value contour constructions. One value of our exposition is the algebraic decomposition of the functions into products of functions of one variable, always valuable when working in high dimensions. Some of these bounds appear to be well known, but we hope this is a useful geometric description, systematic treatment, and parameterization of these bounds. Also we provide proofs from first principles that readers will find easy to reproduce, something that appears to be currently missing in the literature.

2. Model

2.1 Mixture-Model Geometry

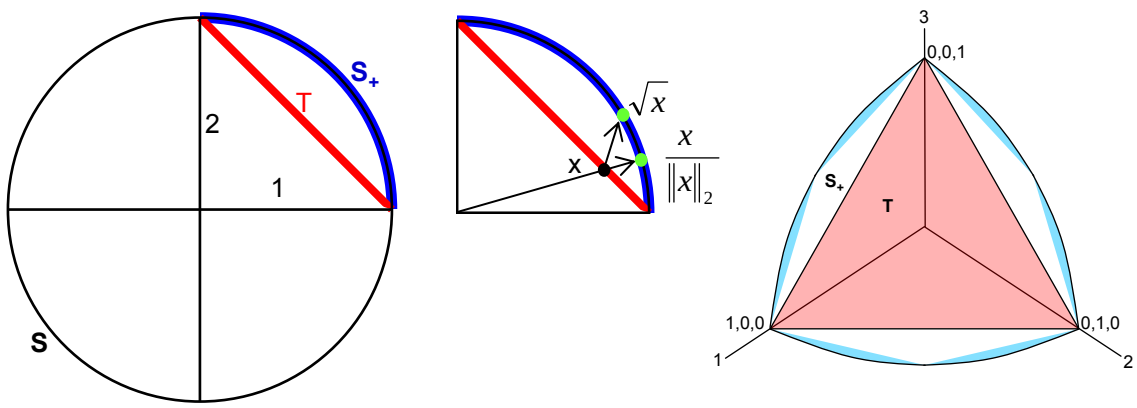

Figure 1: Left, the domain of mixture models is the simplex T, the unit sphere is S, and the non-negative part of unit sphere is S_+. This figure for two-dimensions with coordinate axes 1 and 2. Center, point x on T projected to S_+ under normalization (Euclidean) and square root (Hellinger). Right, three-dimensional simplex T and S_+ (cut-away).

Geometrically, mixture model points lie on a regular simplex T; see Figure 1. Algebraically, these are vectors with positive entries which sum to one. Let K denote the dimension of the model. Let x denote a data point, with x_k the kth coordinate of x. Then

$$T = \left\{ x : \sum_{k=1}^{K} x_k = 1, \text{ and } 1 \geq x_k \geq 0 \right\}.$$

T is a regular simplex in $\mathbb{R}^K$, the convex hull of the elementary vectors $e_k = \{x : x_k = 1, x_{j \neq k} = 0\}$, $\forall k \in [1, K]$.

For Hellinger and Euclidean (Cosine) distances, we map points from T to the unit K-sphere S, specifically the closed section S_+ of it bounded by the positive coordinate planes; see Figure 1. Since zero coordinates map to zero coordinates, all the vertices, edges, etc. of T map to the expected vertices, edges, etc. of S_+. That is, if we treat T and S_+ as simplicial complexes, with subsimplices T_I and S_{+I} with $x_{i \in I} = 0$ for all indicator sets of indices I, then both maps are isomorphisms from sub-simplex T_I to the expected sub-simplex S_{+I}.

Algebraic methods such as non-negative matrix factorization produces output on S_+. The range of some other algebraic methods, such as LSA after normalization, is all of S. The cosine similarity distance is naturally defined on all of S.

2.2 Distance Properties

We desire distances D that satisfy these *useful properties*:

0. **Unique Zero**: $D(x, y) \geq 0$, and $D(x, y) = 0$ if and only if $x = y$.
1. **Max 1**: $D \leq 1$ and $D(x, y) = 1$ for some $x, y \in T$.
2. **Symmetry**: $D(x, y) = D(y, x)$.
3. **Triangle Inequality**: $D(x, z) \leq D(x, y) + D(y, z)$.
4. **Orthogonal Max**: $D(x, y) = 1$ if $x \cdot y = 0$.

(Properties 0–3 are numbered as a reminder to their meaning.) These properties are desired for a variety of practical, theoretical, and historical reasons. Many of our distances satisfy all these except for **Triangle Inequality**. Properties **Unique Zero**, **Symmetry** and **Triangle Inequality** are required for a distance to be a metric. Property **Max 1** means we want distances to be bounded; we scale them to have a consistent maximum to facilitate comparisons. This is not required for metrics. Scaled distances are subscripted by s.

Orthogonal Max implies that the distances between points on disjoint sub-simplices of T are all equal. This is desirable from a mixture model perspective because such points are maximally independent, hence their distances should be the largest possible and equivalent.

For many ideas that originally emerged without some of these properties, the statistics community has developed versions which do. There are several interesting and popular pre-metrics that satisfy some of these. For example, Kullback-Leibler (Kullback and Leibler, 1951) lacks symmetry, but several versions of it have been "fixed."

2.3 Inter-Distance Properties

Two distances D and F have (are)

- **Bounded Difference**: if $c_1 \geq F(x, y) - D(x, y) \geq 0$ for some positive constant $c_1 < 1$.
- **Bounded Ratio**: if $F(x, y) \geq D(x, y) \geq c_2 F(x, y)$ for some positive constant c_2.
- **Order Preserving**: if $D(x, y) < D(x, z) \iff F(x, y) < F(x, z)$.

For our distances, one of them is always greater than the other, so considering the absolute value of the difference, i.e. $|F(x, y) - D(x, y)|$, provides no additional insight.

These properties are a way of relating one function to another. For example, **Order Preserving** functions will produce the same k-nearest neighbor clusterings, provided the analogous distance thresholds are picked. Cosine similarity interprets points as vectors from the origin, and measures the cosine of the angle between two of them. If points are first normalized to S, cosine similarity and Euclidean distance are **Order Preserving**, because the cosine of the angle and the chord length between the points are both monotonic in the angle.

If D and F satisfy **Max 1**, then **Bounded Ratio** implies **Bounded Difference** with $c_1 = 1 - c_2$, since $F - D = F(1 - D/F) \leq 1(1 - c_2)$. But in the following we often show a smaller constant c_1.

2.4 Distance Relation Summary

We define distances $\triangle_s$, JS, and H_s^2. We investigate them throughout the rest of the paper. Figure 2 summarizes their relationships. Motivated by the same desire for a common framework for comparison, Gibbs and Su (2002) provides a similar diagram.

Here $\triangle_s$ is scaled triangular discrimination, a variant of Chi-squared, χ^2. JS is Jensen-Shannon (a.k.a. half the Jeffreys Divergence), a form of Kullback-Leibler. H_s^2 is scaled and squared Hellinger,

$$\boxed{\Delta_s \quad \geq \quad JS \quad \geq \quad H_s^2}$$

$$\left\| \frac{(x+y)}{2} Q\left(\frac{x-y}{x+y}\right) \right\|_1 \qquad \left\| \frac{\max(x,y)}{2} Z\left(\frac{\min(x,y)}{\max(x,y)}\right) \right\|_1$$

$$\left\| \sum_{n=1}^{\infty} a_n \frac{(x-y)^{2n}}{(x+y)^{2n-1}} \right\|_1 \qquad \left\| \frac{\max(x,y)}{2} \sum_{n=2}^{\infty} b_n \left(1 - \frac{\min(x,y)}{\max(x,y)}\right)^n \right\|_1$$

Figure 2: Distance metric taxonomy. Given our scaling, the top line shows a strict ordering of the function values. Further, equality is achieved only at zero and one and we show non-trivial bounded ratio and difference. The bottom two lines show that each of the three functions can be factored into those four expressions, but with different Q and Z functions, and different a_n and b_n coefficients.

a variant of scaled Hellinger, H_s, and raw Hellinger, H. The inequalities denote componentwise inequality, plus bounded ratio and bounded difference. The equations below the box for $\triangle_s$, JS, and H_s^2 denote alternative functional forms derived from factorization and series expansions.

These satisfy all of our useful properties, except for **Triangle Inequality**. Raw H does satisfy **Triangle Inequality**. Figure 3 illustrates a few interesting examples of distance functions in three dimensions.

These three are *relative* distances, meaning they depend on the ratio of the pair of points' coordinates. Specifically, we show that each of the $\triangle_s$, JS, and H_s^2 distances (generically D) can be neatly factored into

$$D(x,y) = \left\| \frac{p}{2} Q(q) \right\|_1 = \left\| \frac{u}{2} Z(z) \right\|_1 = 1 - \left\| \frac{u}{2} W(z) \right\|_1 .$$

Here $p = x + y$ is *plus*, $d = x - y$ is *difference* and $q = d/p$ is *quotient*; also $u = \max(x,y)$, $v = \min(x,y)$, and $z = v/u$. Of course the Q, Z and W functions are different for each distance function D, so we will subscript them by the particular D. Throughout this paper all operations on vectors (e.g. d/p) are applied componentwise. Often the subscripts will be dropped on equations, usually this will still mean that the equality holds componentwise; instead we will explicitly mention it when equality only holds in the aggregate after taking norms. $\| \cdot \|_1$ denotes the standard L-1 vector norm, and is not a componentwise operation; and $| \cdot |$ denotes componentwise absolute value.

Moreover, we will show that all the Q are similar: componentwise

$$Q(q) = \sum_{n=1}^{\infty} a_n q^{2n}, 1 \geq a_n > 0, a_n \text{ rapidly decreasing, and } Q(0) = 0, Q(1) = 1.$$

All the Z are also similar: componentwise

$$Z(z) = \sum_{n=2}^{\infty} b_n (1-z)^n, 1 \geq b_n > 0, b_n \text{ decreasing, and } Z(0) = 1, Z(1) = 0.$$

$$Z(z) = 1 + z - W(z), \text{ with } Z \text{ and } W \text{ monotonic and } W(0) = 1, W(1) = 2.$$

For each of $\triangle_s$, JS, and H_s^2, the D, $\frac{p}{2}Q$, and $\frac{u}{2}Z$ functional forms are all componentwise equal; in contrast $D = 1 - \|\frac{u}{2}W\|_1$ only holds in aggregate after taking the 1-norm, i.e. $D_k \neq 1 - W_k$ nor $1/K - W_k$ in general.

In Section 3.3.2 we briefly contrast Hellinger to E_{us}, the Euclidean distance between mixture model points after they have been projected to the unit sphere.

Functions of the form $D_f(x, y) = \|xf(x/y)\|_1$ for convex functions f are known as f-divergences. They were largely developed by Csiszár (Csiszár, 1967). Dragomir (2001) provides many theorems about them, including noting that our family of functions are f-divergences. Jain and Srivastava (2007) provides some symmetric variants of f-divergence distances, including our triangular discrimination.

In particular, we have componentwise $D(x, y) = xD(1, x/y) = yD(y/x, 1)$, hence $Z(z) = D(1, z) = D(z, 1)$. Similarly we show $Q(q) = D(1 + q, 1 - q) = D(1 - q, 1 + q)$. We provide a simple geometric interpretation of these forms using similar triangles in Section 4.1, Figure 7.

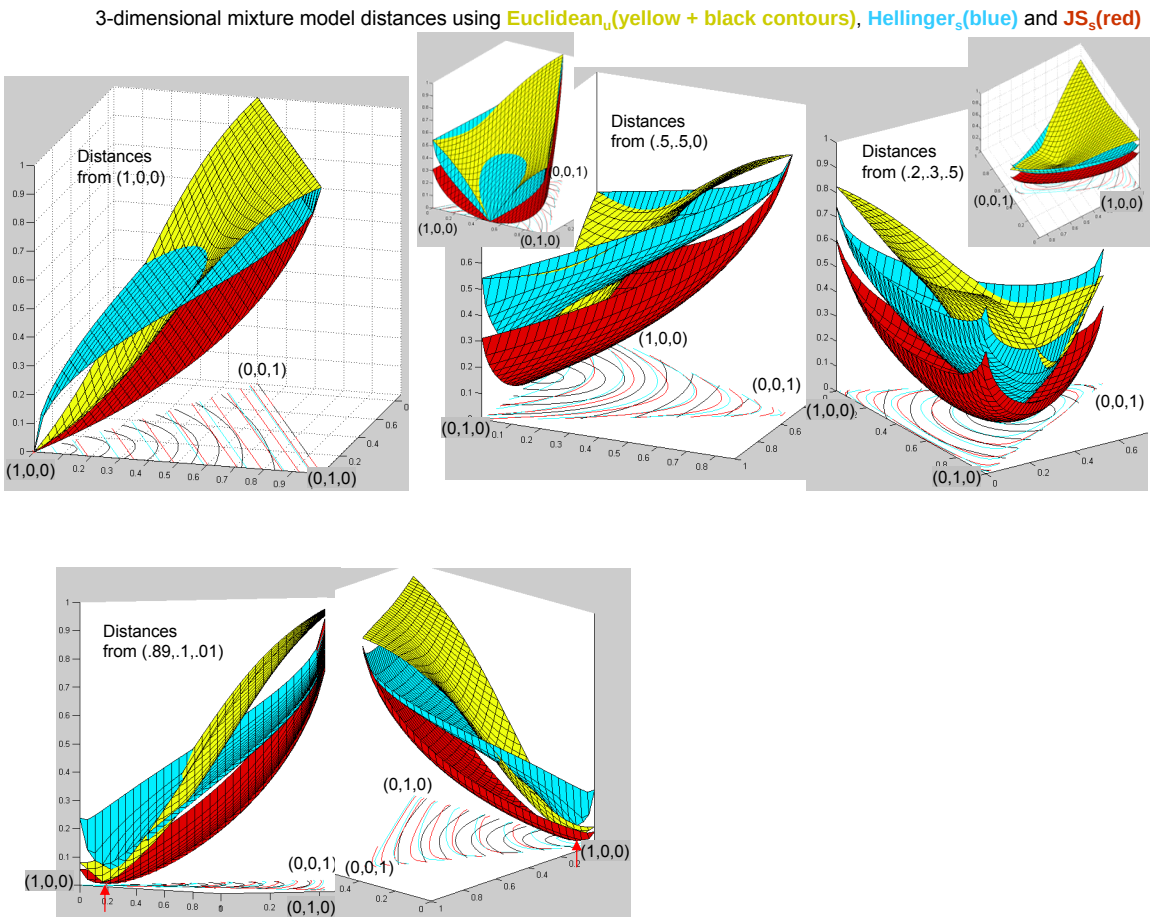

Figure 3: Comparison of E_{us}, Hellinger, and JS_s distances on 3d mixture models. Note the similarity between the contour lines for H_s and JS_s, and how they contrast with those of E_u in black. Bottom figures: the red arrow indicates the position of the point (0.89,.1,.01). Note the steep slope for H_s and JS_s as the line $\overline{(1, 0, 0), (0, 0, 1)}$ is approached, indicated by the blue and red "walls" on the edges of their graphs, and their contours curving sharply towards (0.89,.1,.01).

3. Distance Definitions

3.1 Triangular Discrimination, $\triangle_s$

The definition of the venerable Chi-Squared Test statistic (Pearson, 1900) is $\chi^2 = \sum \frac{(o-e)^2}{e}$ where o is the observed value and e is the expected value.

Most authors take o and e to be mixture model points, yielding $\chi^2(x,y) = \|(x-y)^2/y\|_1$. Alternatively, e could be taken to be the average of all of the points, or the simplex center, which would yield a univariate measure.

We fix the asymmetry to obtain the scaled triangular discrimination:

$$\triangle_s = \frac{1}{2}\sum_{k=1}^{K} \frac{(x_k - y_k)^2}{x_k + y_k} = \frac{1}{2}\left\| \frac{(x-y)^2}{x+y} \right\|_1 = \frac{1}{2}\left\| \frac{d^2}{p} \right\|_1 .$$

Another derivation of $\triangle_s$ is to assume mixture model points are taken from the same population, so the expected value is the average of the two points. That is $\triangle_s = \chi^2(x, (x+y)/2)$. For continuity the kth term of the sum is defined to be 0 if $x_k + y_k = 0$.

We use this simple measure because it turns out to fit in the same geometric family as Jensen-Shannon and Hellinger-squared.

$\triangle_s$ obviously satisfies properties **Unique Zero** and **Symmetry**. Any term where $y_k = 0$ reduces to x_k, so **Max 1** and **Orthogonal Max** hold.

But $\triangle_s$ is too convex, in the sense that $\triangle_s(x,y) \gg 2\,\triangle_s\,(x, x/2 + y/2)$, and does not satisfy **Triangle Inequality** even when restricted to mixture models. E.g. $x = [1,0], z = [0,1]$ and $y = [1/2, 1/2]$ has $\triangle_s(x,y) = \triangle_s(y,z) = 1/3$ and $\triangle_s(x,z) = 1$. Normalizing points so that they lie on the unit sphere S_+ first helps make the function less convex, but not enough: e.g. if $y = [1,1]/\sqrt{2}$ then $\triangle_s(x,y) = 0.379$.

3.2 Jensen-Shannon Divergence

One information theory approach to distance is based on entropy and divergence. The derivation starts with the Kullback-Leibler measure, KL, then modifies it for our useful properties, arriving at the Jensen-Shannon Divergence, JS. This is also known as one-half of the Jeffrey Divergence.

$$KL(x,y) = \left\| x \log_2(x/y) \right\|_1$$

KL is non-symmetric in x and y, which motivates $KL_{sym}(x,y) = \left\| x\log_2(x/y) + y\log_2(y/x) \right\|_1$. Despite $w \log w$ being reasonably well behaved near zero, having independent quantities inside the log's means $KL_{sym,k}$ is unbounded for $x_k = 0$ xor $y_k = 0$. Moreover, if $x_k = 0$, it doesn't matter what value $y_k > 0$ has, the k term always contributes the same amount, infinity. Indeed, it doesn't matter what any of the other $y_{j\neq k}$ or $x_{j\neq k}$ terms are! To "fix" this, we replace the denominator in the log's by the average of x and y:

$$JS = JS_s = \frac{1}{2}\left\| x\log_2 \frac{2x}{x+y} + y \log_2 \frac{2y}{x+y} \right\|_1 .$$

To make JS continuous $x_k \log_2 \frac{2x_k}{x_k+y_k} \equiv 0$ for $x_k = 0$, since $\lim_{w\to 0} w \log_2 w = 0$.

This is a measure of relative distance, the difference between small component values is accentuated non-linearly; see Figure 4 left. The constant factors inside the log's were chosen so that $JS_k = x_k/2$ for $y_k = 0$, which provides **Orthogonal Max** and **Max 1**. If $y_k = x_k$ then $\log_2(1) = 0$ which verifies **Unique Zero**. (It may not be obvious that $JS \geq 0$, but it is, and can be seen from some stronger results we prove later.) **Symmetry** holds by the symmetry of the functional form.

But JS does not satisfy **Triangle Inequality**, and is not fixed by normalizing the points to the sphere. This can be verified using the same easy points as for $\triangle_s$. Indeed, JS is even more convex than $\triangle_s$, as amplified in the following section.

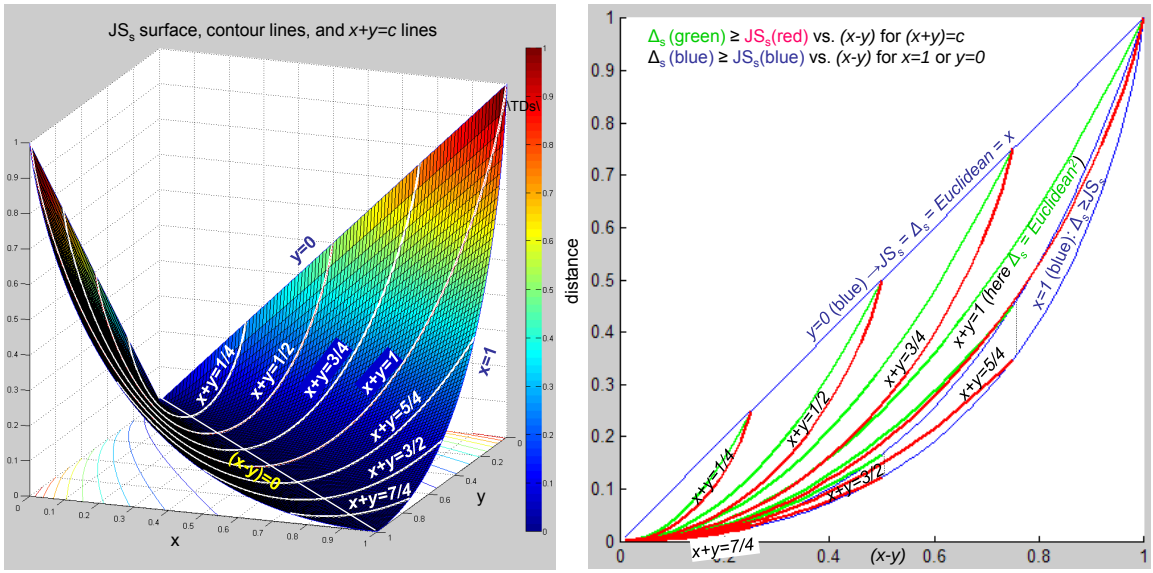

Figure 4: Left: one-dimensional JS. Right: comparison of the one-dimensional $\triangle_s$ and JS; for constant $x + y$ lines, $\triangle_s \geq JS$. Both plots range over the square. The square is the domain of each component when $K > 1$.

3.3 Hellinger Distance

3.3.1 Minkowski and Euclidean

The well-known order p Minkowski distances are $M^p(x, y) = \left(\sum |x_k - y_k|^p\right)^{1/p}$. Here $M_u^p(x, y) = M^p(x/||x||_p, y/||y||_p)$ measures the distance between the standard normalizations of x and y onto S. $M_u^2 = E_u(x, y)$ is Euclidean distance. Another common distance function, cosine similarity, is simply $E_u^2(x, y)$. Hellinger can be viewed as Euclidean distance after a peculiar geometric mapping.

3.3.2 Hellinger

Our Hellinger distance is a discrete form of the Hellinger integral (Hellinger, 1909) defined for more general spaces. Its use and form for our modern context was described by Blei and Lafferty (2007).

$$H^2 = \sum_{k=1}^{K} (\sqrt{x_k} - \sqrt{y_k})^2$$

H^2 means squaring after the sum, not componentwise: $H = \left(\sum_{k=1}^{K} (\sqrt{x_k} - \sqrt{y_k})^2\right)^{1/2}$.
We normalize H by a constant factor for property **Max 1**,

$$H_s = \frac{H}{\sqrt{2}}.$$

We observe that Hellinger (H) projects points from the simplex T to the spherical section S_+ using the componentwise square root transformation, then takes standard Euclidean distance, which is the chord length between transformed points. This is trivial but apparently not the way those using it for mixture models think about it. This also constitutes a simple geometric proof that H satisfies **Triangle Inequality**. The usual vector 2-norm normalization $x/\|x\|_2$ also takes points to

the sphere but is significantly different, e.g. it is a linear scaling of all components. When going to the sphere, Hellinger expands straight-line-distances near the boundary of T, where ratios of components are highest, whereas normalization modestly expands straight-line-distances near the center of T. Both map the same sub-simplices of T to the obvious subsimplices of S_+, and both maps agree at sub-simplex centers. Some bounds on the difference of these projections are known.

It is obvious that H satisfies properties **Unique Zero** and **Symmetry**. H_s satisfies **Orthogonal Max**. The algebraic argument is that, for orthogonal x and y, $x_k = 0$ iff $y_k \neq 0$, so $\sum_{k=1}^{K}(\sqrt{x_k} - \sqrt{y_k})^2 = \sum_{k=1}^{K} \sqrt{x_k}^2 + \sqrt{y_k}^2 = 2$. For a geometric argument, take $\sqrt{x}$ as the north pole of S, then an orthogonal $\sqrt{y}$ lies on the equator; all such point pairs are equidistant. Orthogonal x and y are in disjoint sub-simplices of T; the same holds for their projections onto S_+.

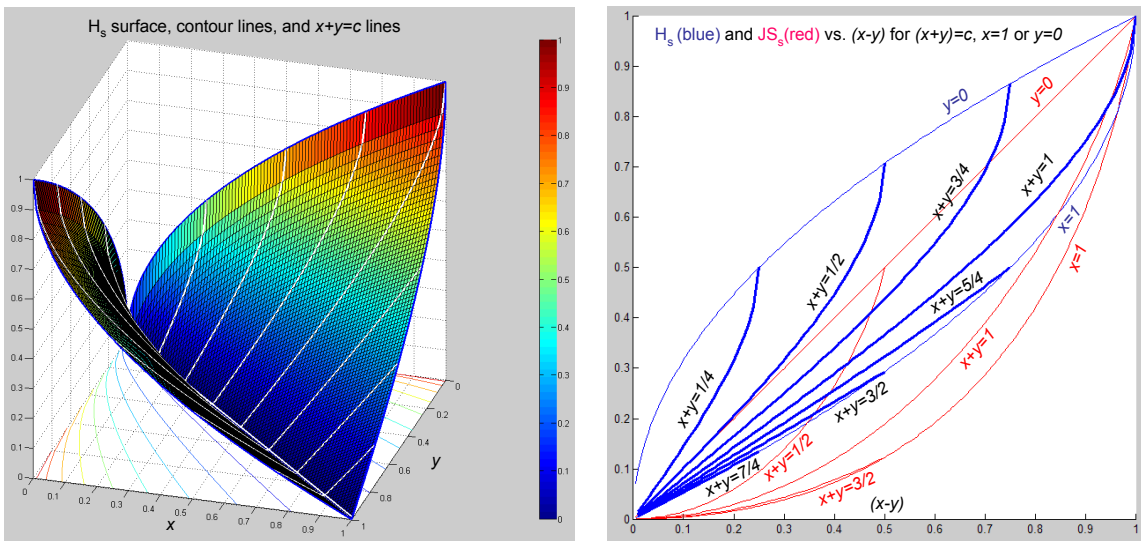

Figure 5: Left: one-dimensional H_s. Right: comparison of the one-dimensional H_s and JS_s. This provides little insight for higher dimensions, because Hellinger takes the square root after summing all components.

4. Comparisons

Recall $u = \max(x, y)$ and $v = \min(x, y)$ with max and min and other vector operations taken componentwise. Also $p = x + y = u + v$ and $d = |x - y| = u - v$ and $q = d/p$ and $z = v/u$. For components where $x = y = 0$ define $q = 1$ and $z = 0$. Note $u = (p + d)/2$ and $v = (p - d)/2$. For componentwise ranges we have $p \in [0, 2]$ and $d, q, u, v, x, y, z \in [0, 1]$. We have inequalities $p \geq d$ and $u \geq v$.

Theorem 1 (f-divergences) *For distances $\triangle_s, JS$ and H_s^2 (generically D) and $a \geq 0$, $D(ax, ay) = aD(x, y)$ componentwise. This implies componentwise $D(x, y) = \frac{u}{2}Z(z)$ and $D(x, y) = \frac{p}{2}Q(q)$ with $Z(z) = 2D(1, z)$ and $Q(q) = D(1 + q, 1 - q)$.*

Proof If $u = 0$ then $x = y = 0$ and $p = 0$. In this case it is trivial to check $D = 0$ (**Unique Zero**) for each of the three functions. So assume $u > 0$ and $p > 0$. Verifying $D(ax, ay) = aD(x, y)$ is a simple factorization exercise for each function. It implies $D(x, y) = xD(1, y/x) = yD(x/y, 1)$. Consider each component in turn, and, by symmetry, WLOG assume $x \geq y$. Then $xD(1, y/x) = uD(1, z)$ and $D(x, y) = (x+y)D(1/2+(x-y)/(2(x+y)), 1/2-(x-y)/(2(x+y))) = pD((1+q)/2, (1-q)/2)$.

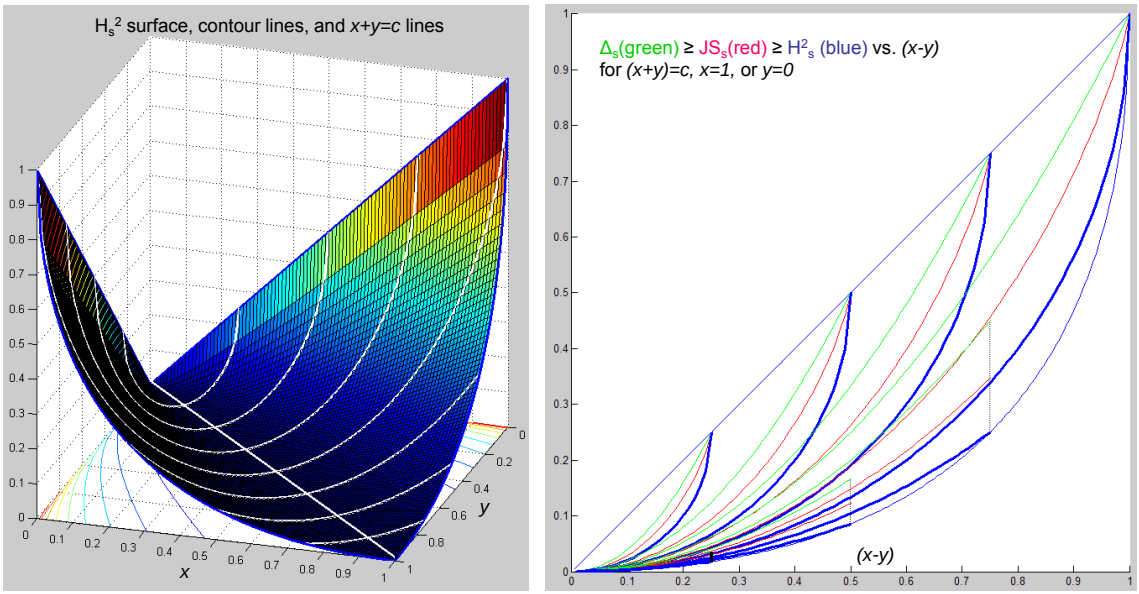

Figure 6: Left: one-dimensional H_s^2. Right: comparison of the one-dimensional $\triangle_s$, JS, and H_s^2. In Section 4.1 we show that the family of $\triangle_s$, JS, H_s^2 triples of curves are all linear scalings (and truncations) of a single triple of curves, the plots of the Q functions. In Section 4.3 we show that straight-lines from the origin to the $x = 1$ curve (the rightmost-edge of the left figure) map out the surface. This $x = 1$ curve is the lower envelop of the curves on the right, and is the Z function.

For our functions the restriction of the domain to the unit square can be ignored, so we can factor out the 1/2 which provides the compact expression $Q(q) = D(1 + q, 1 - q)$. ■

What this means geometrically is straight-lines from the origin to the $x = 1$ (or $y = 1$) curve map out the surface of the one-dimensional distance functions over the square; see Figure 7. The slope of each line is related to the value of the Z-function (slope $= Z(z)/\sqrt{1+z^2}$) or Q-function (slope $= Q(q)\sqrt{2}/\sqrt{1+q^2}$).

4.1 Functions of p and q

Here we examine the Q functions for each of $\triangle_s$, JS and H_s^2. Figure 8 illustrates various relationships between them.

Theorem 2 (Q-functions) *Componentwise*

$$\triangle_s = \frac{p}{2} Q_\triangle(q) \text{ and } JS = \frac{p}{2} Q_{JS}(q) \text{ and } H_s^2 = \frac{p}{2} Q_H(q)$$

where

$$Q_\triangle(q) = q^2$$

$$Q_{JS}(q) = \frac{1}{2}\left((1+q)\log_2(1+q) + (1-q)\log_2(1-q)\right)$$

$$Q_H(q) = 1 - \sqrt{1-q^2}$$

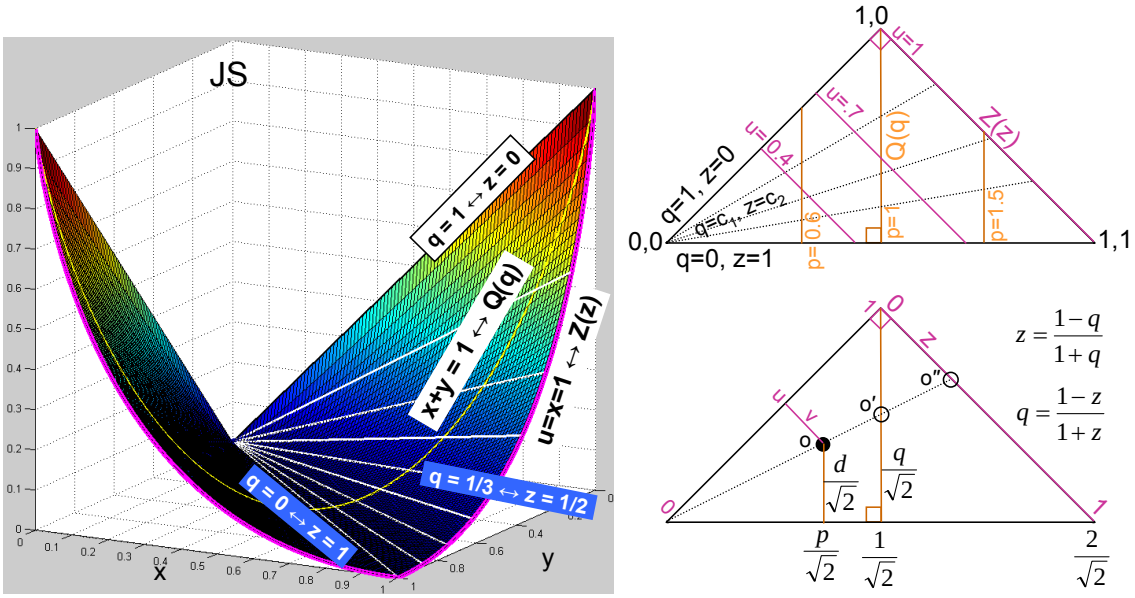

Figure 7: Left, graph of one-dimensional JS highlighting the iso-z = iso-q lines, and the Q- and Z-functions. Graphs for H_s^2 and $\triangle_s$ are similar. The geometric interpretation of $D(ax, ay) = aD(x, y)$ in Theorem 1 is that the constant-x/y curves are straight lines. Right top, planar view of the left graph, showing additional iso-u and iso-p lines. JS is a linear scaling of the Z (or Q) curve along these iso lines. Right bottom, the distance at a point o can be evaluated by translating it along its iso-z (iso-q) line to the Z (Q) curve. Geometrically, let $|\cdot|$ denote the straight-line distance from a point to the left $(0, 0)$ corner of the triangle in the plane. Then linear scaling along iso-q (iso-z) lines and similar triangles yields $D(o) = (|o|/|o'|)D(o') = (p/1)D(o') = (p/2)Q(q)$ and $D(o) = (|o|/|o''|)D(o'') = (u/1)D(o'') = (u/2)Z(z)$. (To clear up any confusion, the final 1/2 factor appears because the K-dimensional distances (D) have an extra 1/2 normalization factor that the 1-dimensional distances do not, and we defined Z and Q in terms of the 1-dimensional distances for illustration purposes in this figure.)

and, for convenience, we list

$$Q'_{\triangle} = 2q$$

$$Q'_{JS} = \frac{1}{2}\left(\log_2(1+q) - \log_2(1-q)\right)$$

$$Q'_H = \frac{q}{\sqrt{1-q^2}}$$

Component-wise equality holds in all of the above. This implies equality after taking the 1-norm: $\triangle_s = \|\frac{p}{2}Q_{\triangle}(q)\|_1$, $JS = \|\frac{p}{2}Q_{JS}(q)\|_1$ *and* $H_s^2 = \|\frac{p}{2}Q_H(q)\|_1$.

Proof

$Q_{\triangle}$ is trivial. From Theorem 1, $2Q_{JS} = 2JS(1+q, 1-q) = (1+q)\log_2(1+q) + (1-q)\log_2(1-q)$. $2Q'_{JS} = \log_2(1+q) + (1+q)/((1+q)\log 2) + (1-q)/((1-q)\log 2) - \log_2(1-q) = \log_2(1+q) - \log_2(1-q)$. From Theorem 1, $Q_H = \left(\sqrt{1+q} - \sqrt{1-q}\right)^2/2 = \left(1+q+1-q-2\sqrt{(1+q)(1-q)}\right)/2 = 1 - \sqrt{1-q^2}$.

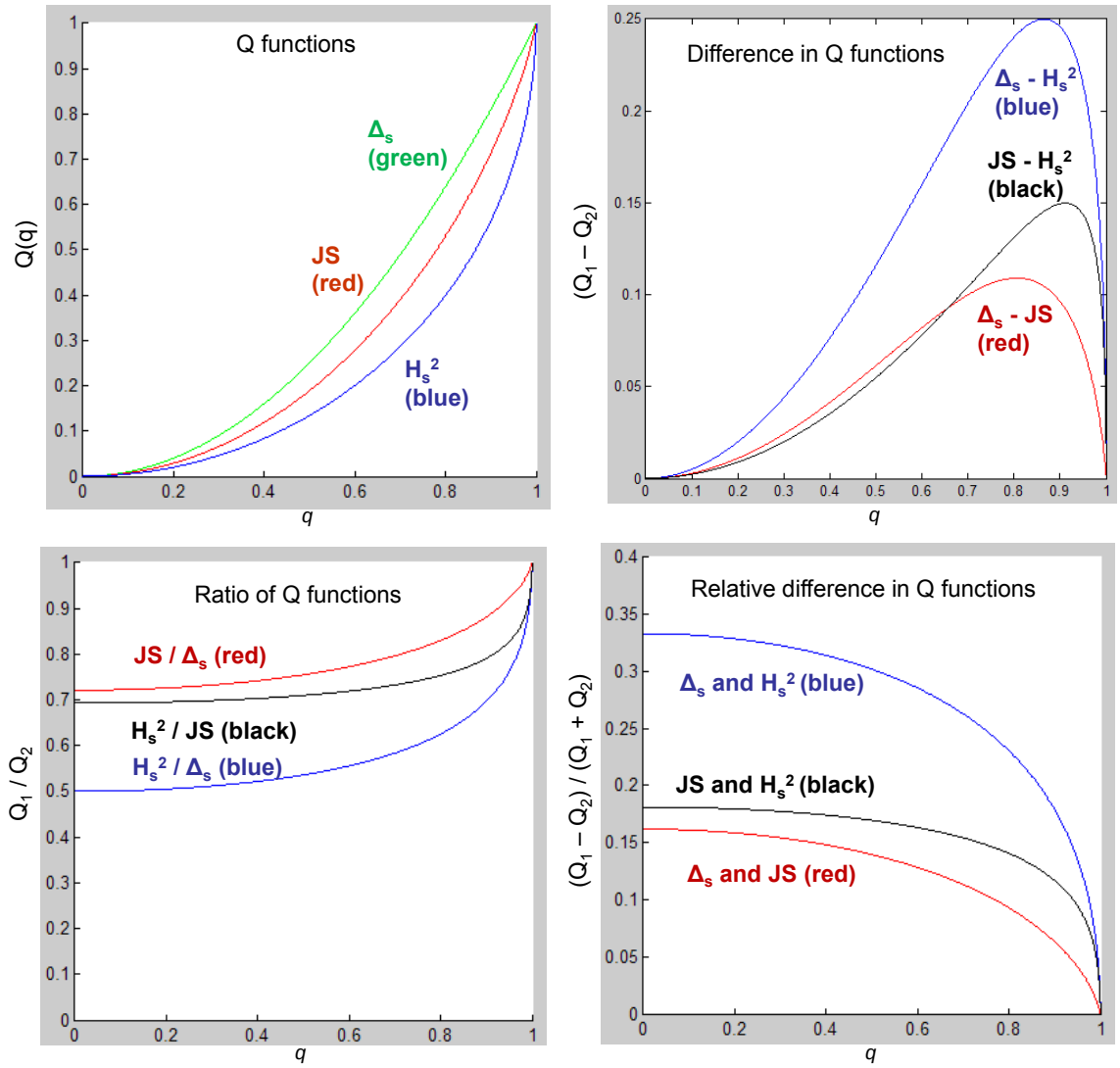

Figure 8: Graphs of relationships between the Q functions.

■

Remark 1 (linear scaling) *All the D vs. d for $p = p_c$ (constant) curves in Figure 4 and Figure 6 are linear $1/p_c$ scalings of the Q functions: $D(x, y)/p_c$ vs. $(x - y)/p_c \iff Q(q)$ vs. q. For $p_c > 1$ the functions are truncated at $q = 2/p_c - 1$.*

Proof This follows almost by definition: $(x - y)/p_c = d/p = q$ and $D/p_c = pQ/p_c = Q$. Note the $1/2$ factor is missing in the Q decomposition of D because the one-dimensional distance functions plotted in the figures are normalized without it. The curves for $p_c > 1$ are truncated at $d = 2-p_c \iff q = 2/p_c-1$ since $x \leq 1 \Rightarrow 1+y \geq x+y = p_c$ or $y \geq p_c-1$, so $x-y \leq x+1-p_c \leq 2-p_c$. ■

Remark 2 (geometric Q) *For a geometric interpretation see Figure 7. The Q curve and all its translated scalings that lie on the distance function are perpendicular to the $p = 1$ ($x + y = 1$) diagonal. In the figure, the following operations can be observed geometrically. Considering point coordinates, $o = ((p + d)/2, (p - d)/2)$ and $o' = ((1 + q)/2, (1 - q)/2) = o/p$. Hence $D(o) = pD(o') = (p/2)Q(q)$. In the same figure, the Z curve (section 4.3) and all its translated scalings are perpendicular to the $z = 0$ ($y = 0$ if $y < x$) axis.*

4.2 JS and H_s^2 via series in q

4.2.1 JS q-series

Using $\log_2(\cdot) = \ln(\cdot)/\ln(2)$, and the expansion $\ln(1 + r) = \sum_{n=1}^{\infty} \frac{(-1)^{n+1} r^n}{n}$, we get

$$2 \ln 2 Q_{JS} = (1 + q) \sum_{n=1}^{\infty} \frac{(-1)^{n+1}}{n} q^n - (1 - q) \sum_{n=1}^{\infty} \frac{1}{n} (q)^n$$

Recombining like powers of q

$$= \sum_{n=1}^{\infty} \frac{(-1)^{n+1} - 1}{n} q^n + \sum_{n=1}^{\infty} \frac{(-1)^{n+1} + 1}{n} q^{n+1}$$

$$= \sum_{n=1}^{\infty} \frac{(-1)^{n+1} - 1}{n} q^n + \sum_{m=2}^{\infty} \frac{(-1)^m + 1}{m - 1} q^m$$

$$= \sum_{n=2}^{\infty} \left(\frac{(-1)^{n+1} - 1}{n} + \frac{(-1)^n + 1}{n - 1} \right) q^n$$

$$= \sum_{n=2}^{\infty} \frac{(n - 1)(-1)^{n+1} - n + 1 + n(-1)^n + n}{n(n - 1)} q^n = \sum_{n=2}^{\infty} \frac{(-1)^{n+2} + 1}{n(n - 1)} q^n$$

The numerator is zero if n is odd and 2 if n is even. Retaining the even terms and re-indexing gives

$$= \sum_{n=1}^{\infty} \frac{1}{n(2n - 1)} q^{2n}$$

Thus

$$JS_k(x, y) = \frac{p}{2} \sum_{n=1}^{\infty} \frac{1}{n(2n - 1) 2 \ln 2} q^{2n} \tag{1}$$

$$= \left(\frac{1}{4 \ln 2} \right) pq^2 + \left(\frac{1}{24 \ln 2} \right) pq^4 + \left(\frac{1}{60 \ln 2} \right) pq^6 + \left(\frac{1}{112 \ln 2} \right) pq^8 + \left(\frac{1}{180 \ln 2} \right) pq^{10} + \cdots$$

$$\approx 0.361 pq^2 + 0.060 pq^4 + 0.024 pq^6 + 0.013 pq^8 + 0.008 pq^{10} + \cdots$$

Note the sum of the coefficients is 0.5 by **Max 1** and $\sum_k p_k = 2$, where $d = p$ for orthogonal x and y components.

Note that the leading term of the JS series expansion is the same as $\triangle_s$, up to a small constant factor. JS is an interesting mix of absolute and relative difference. Consider $pq^{2n} = dq^{2n-1}$, so in contrast to pure relative difference d/p, JS weights the relative difference more if the absolute difference is large;

4.2.2 H_s^2 q-series

Using the expansion $\sqrt{1+r} = \sum_{n=0}^{\infty} \frac{(-1)^n (2n)!}{(1-2n)(n!)^2 4^n} r^n$ with $r = -q^2$ gives

$$Q_H = 1 - \sum_{n=0}^{\infty} \frac{(-1)^n (2n)!}{(1-2n)(n!)^2 4^n} (-1)^n q^{2n} = \sum_{n=1}^{\infty} \frac{(2n)!}{(2n-1)(n!)^2 4^n} q^{2n}$$

Therefore componentwise

$$H_s^2 = \frac{p}{2} \sum_{n=1}^{\infty} \frac{(2n)!}{(2n-1)(n!)^2 4^n} q^{2n} \tag{2}$$

$$= \left(\frac{1}{4}\right) pq^2 + \left(\frac{1}{16}\right) pq^4 + \left(\frac{1}{32}\right) pq^6 + \left(\frac{5}{256}\right) pq^8 + \left(\frac{7}{512}\right) pq^{10} + \cdots$$

$$\approx 0.2500pq^2 + 0.0625pq^4 + 0.0312pq^6 + 0.0195pq^8 + 0.0137pq^{10} + \cdots$$

Note the coefficients of the larger powers are bigger than for the JS series, which is illustrated by the larger curvature in Figure 6 right.

$\triangle_s, JS$ and H_s^2 have similar behavior, but through different operations and of different order. $\triangle_s$ is a simple ratio of powers, JS uses $\log_2$, and H_s^2 uses $\sqrt{\ }$. If $y_k = 0$, then the kth component of $\triangle_s$, JS, and H_s^2 are all equal to $x_k/2$. The contours (iso-value lines) for all three have similar shape. See Figure 3.

4.3 Functions of u and z

Here we describe the Z functional forms for $\triangle_s$, JS and H_s^2. We also introduce $W(z)$ functions. Figure 9 illustrates various relationships between the Z-functions for different distances.

Theorem 3 (Z-functions) *Componentwise*

$$\triangle_s = \frac{u}{2} Z_\triangle(z) \text{ and } JS = \frac{u}{2} Z_{JS}(z) \text{ and } H_s^2 = \frac{u}{2} Z_H(z)$$

where

$$Z_\triangle(z) = \frac{(1-z)^2}{1+z} = 1 + z - \frac{4z}{1+z}$$

$$Z_{JS}(z) = 1 + z - (1+z)\log_2(1+z) + z\log_2 z$$

$$Z_H(z) = 1 + z - 2\sqrt{z}$$

and for convenience we list

$$Z'_\triangle = 1 - \frac{4}{(1+z)^2} = \frac{(z+3)(z-1)}{(1+z)^2}$$

$$Z'_{JS} = 1 - \log_2(1+z) + \log_2(z) = 1 - \log_2(1+z^{-1})$$

$$Z'_H = 1 - \frac{1}{\sqrt{z}}$$

As for the Q functions, all of the above holds componentwise and implies equality after taking 1-norms.

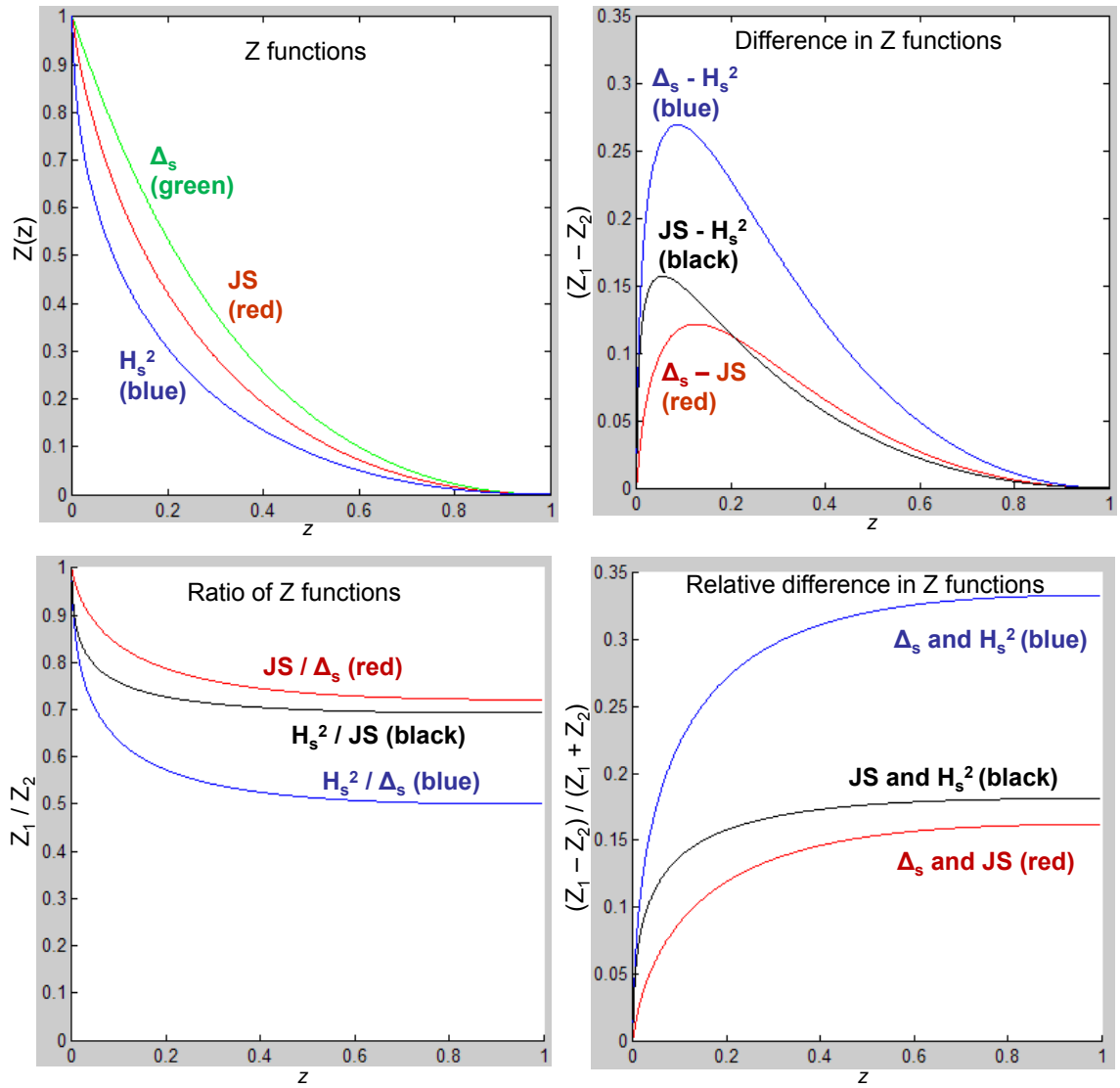

Figure 9: Graphs of relationships between the Z functions.

Proof We use $Z(z) = 2D(1, z)$ from Theorem 1.

For $\triangle_s$ we have $2\,\triangle_s\,(1, z) = (1-z)^2/(1+z) = \left((1+z)^2 - 4z\right)/(1+z) = 1 + z - 4z/(1+z)$. Also $Z'_{\triangle} = (-2(1-z)(1+z) - (1-z)^2)/(1+z)^2 = (z^2 + 2z - 3)/(z+1)^2 = (z+3)(z-1)/(z+1)^2$.

$$2JS(1, z) = \log_2 \frac{2}{1+z} + z \log_2 \frac{2z}{1+z} = (1+z) \log_2 \frac{2}{1+z} + z \log_2 z = 1 + z - (1+z) \log_2(1+z) + z \log_2 z.$$

Also $Z'_{JS} = 1 - (1+z)/((1+z)(\log 2)) - \log_2(1+z) + z/(z \log 2) + \log_2 z = 1 - \log_2(1+z) + \log_2(z)$ $= 1 + \log_2(z/(1+z)) = 1 - \log_2((1+z)/z) = 1 - \log_2(1+z^{-1})$.

For H_s^2, we have $2H_s^2(1, z) = (1 - \sqrt{z})^2 = 1 + z - 2\sqrt{z}$. And Z'_H is trivial. ■

The leading $(u/2)(1+z) = (x+y)/2$ terms always sum to 1 over all the components of mixture models, so we have a concise expression of how much each of these measures is less than 1.

Corollary 4

$$\triangle_s = 1 - \|\frac{u}{2}W_\triangle(z)\|_1 = 1 - \|\frac{u}{2}\frac{4z}{1+z}\|_1$$

$$JS = 1 - \|\frac{u}{2}W_{JS}(z)\|_1 = 1 - \|\frac{u}{2}(z\log_2 z - (1+z)\log_2(1+z))\|_1$$

$$H_s^2 = 1 - \|\frac{u}{2}W_H(z)\|_1 = 1 - \|\frac{u}{2}(2\sqrt{z})\|_1$$

The leftmost equality is not componentwise equality. E.g. $H_{sk}^2 \neq 1/k - u_k\sqrt{z_k}$ *in general.*

Proof In this proof we use subscripts k to emphasize that keeping track of individual components is important. $\|\frac{u}{2}Z(z)\|_1 = \sum_{k=1}^K |\frac{u_k}{2}(1+z_k - W(z_k))|$. We first note that $1+z_k \geq W(z_k) \geq 0$ so we can remove the absolute value sign. The argument for this is that each component of the original distance functions is non-negative (since the distance functions are distances over all of $\mathbb{R}_+^K$ and not just mixture models) and equal to the Z functions. Each of the W are non-negative. Thus we may drop the absolute values and separate the sum into two giving $\|\frac{u}{2}Z(z)\|_1 = \sum_{k=1}^K \frac{u_k}{2}(1+z_k) - \sum_{k=1}^K W(z_k)$. The first sum is 1 because it is merely $\sum_{k=1}^K (u_k+v_k)/2 = \sum_{k=1}^K (x_k+y_k)/2$ and our domain is mixture models. ■

4.4 $\triangle_s, JS$ and H_s^2 via series in z

Here we provide series expansions for our functions in z, about the point $z=1$. For each we define $r = 1-z$, and each series contains integer powers of r starting with 2. We make use of $1-r=z$, $2-r = z+1$, and $1 \geq r \geq 0$.

4.4.1 $\triangle_s$ z-series

$$Z_\triangle = \frac{(1-z)^2}{1+z} = r^2\left(\frac{1}{2-r}\right) = \frac{r^2}{2}\left(\frac{1}{1-r/2}\right) = \sum_{n=0}^\infty \frac{r^{n+2}}{2^{n+1}} = \sum_{n=2}^\infty \frac{r^n}{2^{n-1}}$$

Thus

$$\triangle_s(x,y) = u\sum_{n=2}^\infty \frac{r^n}{2^n} \tag{3}$$

$$= \left(\frac{1}{4}\right)ur^2 + \left(\frac{1}{8}\right)ur^3 + \left(\frac{1}{16}\right)ur^4 + \left(\frac{1}{32}\right)ur^5 + \left(\frac{1}{64}\right)ur^6 + \cdots$$

$$= 0.25ur^2 + 0.125ur^3 + 0.0625ur^4 + 0.03125ur^5 + 0.015625ur^6 + \cdots$$

4.4.2 JS z-series

Z_{JS} has two log terms:

$$z\log z = (1-r)\log_2(1-r) = -\frac{1-r}{\ln 2}\sum_{n=1}^\infty \frac{r^n}{n} \text{ and}$$

$$(1+z)\log(1+z) = (2-r)\log_2(2-r) = \frac{(2-r)}{\ln 2}(\ln 2 + \ln(1-r/2)) = 2 - r + \frac{2-r}{\ln 2}\ln(1-r/2)$$

$$\text{where } \ln(1-r/2) = -\sum_{n=1}^\infty \frac{r^n}{n2^n}.$$

The $2-r$ in the second term cancels the leading $2-r$ (i.e. $1+z$) in Z_{JS}, yielding

$$Z_{JS} = \frac{2-r}{\ln 2}\sum_{n=1}^{\infty}\frac{r^n}{n2^n} - \frac{1-r}{\ln 2}\sum_{n=1}^{\infty}\frac{r^n}{n} = \frac{1}{\ln 2}\sum_{n=1}^{\infty}\left(\frac{r^n}{n2^{n-1}} - \frac{r^n}{n}\right) + \frac{1}{\ln 2}\sum_{n=1}^{\infty}\left(\frac{-r^{n+1}}{n2^n} + \frac{r^{n+1}}{n}\right).$$

The first sum is zero for $n=1$. Letting $m=n+1$ in the second sum we have

$$= \frac{1}{\ln 2}\sum_{n=2}^{\infty}\frac{r^n}{n}\left(\frac{1}{2^{n-1}}-1\right) + \frac{1}{\ln 2}\sum_{m=2}^{\infty}\frac{-r^m}{m-1}\left(1-\frac{1}{2^{m-1}}\right) = \frac{1}{\ln 2}\sum_{n=2}^{\infty}r^n\left(1-\frac{1}{2^{n-1}}\right)\left(\frac{1}{n-1}-\frac{1}{n}\right)$$

Thus

$$JS(x,y) = \frac{u}{2\ln 2}\sum_{n=2}^{\infty}r^n\left(1-\frac{1}{2^{n-1}}\right)\frac{1}{n(n-1)} \tag{4}$$

$$= \left(\frac{1}{8\ln 2}\right)ur^2 + \left(\frac{3}{48\ln 2}\right)ur^3 + \left(\frac{7}{192\ln 2}\right)ur^4 + \left(\frac{15}{640\ln 2}\right)ur^5 + \left(\frac{31}{1920\ln 2}\right)ur^6 + \cdots$$

$$\approx 0.1803ur^2 + 0.0902ur^3 + 0.0526ur^4 + 0.0338ur^5 + 0.0233ur^6 + \cdots$$

4.4.3 H_s^2 z-series

$$Z_H = 1+z-2\sqrt{z} = 2-r-2\sqrt{1-r} = 2-r-2\left(1-\frac{r}{2}+\sum_{n=2}^{\infty}\frac{r^n}{n!}\prod_{m=1}^{n}(m-3/2)\right) = -2\sum_{n=2}^{\infty}r^n\prod_{m=1}^{n}\frac{m-3/2}{m}$$

Thus

$$H_s^2(x,y) = -u\sum_{n=2}^{\infty}r^n\prod_{m=1}^{n}\frac{m-3/2}{m} \tag{5}$$

$$= \left(\frac{1}{8}\right)ur^2 + \left(\frac{1}{16}\right)ur^3 + \left(\frac{5}{128}\right)ur^4 + \left(\frac{7}{256}\right)ur^5 + \left(\frac{63}{3072}\right)ur^6 + \cdots$$

$$\approx 0.125ur^2 + 0.0625ur^3 + 0.0391ur^4 + 0.0273ur^5 + 0.0205ur^6 + \cdots$$

This series converges slowly for r close to 1, i.e. z near 0.

4.5 Z and Q Equivalence and Analysis

The different forms Z, Q, and W merely provide convenient alternatives for intuition, proofs, and perhaps applications. The Z and Q functions are very similar in form, as can be seen from the plots. Algebraically they are related in the following way. Componentwise equality with the original distance function means $\frac{p}{2}Q(q) = \frac{u}{2}Z(z)$. Since $q=(1-z)/(1+z)$ and $p=u(1+z)$, also $z=(1-q)/(1+q)$ and $u=p(1+q)/2$, we have the following theorem.

Theorem 5 (Z-Q-same)

$$Z(z) = (1+z)Q\left(\frac{1-z}{1+z}\right)$$

$$Q(q) = \frac{1+q}{2}Z\left(\frac{1-q}{1+q}\right)$$

Corollary 6 *Z decreasing $\Rightarrow$ Q increasing; also Z decreasing $\Leftarrow$ $Q'(q) > \frac{1}{1+q}Q(q) \geq 0$.*

Proof Z decreasing means $\forall z_1 < z_2, Z(z_1) > Z(z_2)$. Since $Z(z) = (1+z)Q((1-z)/(1+z))$ we have

$$Q\left(\frac{1-z_1}{1+z_1}\right) > \frac{1+z_2}{1+z_1}Q\left(\frac{1-z_2}{1+z_2}\right) > Q\left(\frac{1-z_2}{1+z_2}\right)$$

where $q_1 = \frac{1-z_1}{1+z_1} > \frac{1-z_2}{1+z_2} = q_2$. Since the mapping between z and q is a continuous isomorphism this inequality holds for arbitrary $q_1 > q_2$. For the other direction, Z decreasing $\iff$ $Q(q_1) > \frac{1+z_2}{1+z_1}Q(q_2) = \frac{1+q_1}{1+q_2}Q(q_2)$. We manipulate this inequality to get it into derivative form,

$$\frac{Q(q_1)-Q(q_2)}{q_1-q_2} > \left(\frac{1+q_1}{1+q_2}-1\right)Q(q_2)/(q_1-q_2) = \left(\frac{q_1-q_2}{1+q_2}\right)Q(q_2)/(q_1-q_2) = \frac{1}{1+q_2}Q(q_2).$$

This holds always if it holds in the limit as $q_1 \to q_2$, or $Q'(q) > \frac{1}{1+q}Q(q)$. ■

The stronger requirement for Q' is necessary; e.g. $Q = 1+q/2$ implies $Z = (1+z)(1+(1-z)/(2(1+z))) = 3/2 + z/2$, so here is an example where both Q and Z are increasing. Geometrically what is happening is that a constant z (or q) ray from the origin first intersects the $p = 1$ line, then the $u = 1$ line. Recall D is rising linearly along this ray. For smaller values of q, the p and u lines are farther apart, so D increases more. For example, a flat $Q(q) = 1$ function implies an increasing $Z(z) = 1 + z$ function.

Theorem 5 and Corollary 6 hold generically for functions with $D(ax, ay) = aD(x, y)$. We now turn to our particular functions, and show that the decreasing/increasing conditions hold in the half-open interval $q \in (0, 1]$ (or $z \in [0, 1)$), and they are flat at the excluded end point, i.e. $Q'(0) = 0$ and $Z'(1) = 0$.

Theorem 7 (Z-decreasing) *$Z_\triangle(z), Z_{JS}(z)$, and $Z_H(z)$ are all strictly decreasing in $[0, 1)$, with zero derivative at $z = 1$. Note $Z'_\triangle(0) = -3$, but $Z'_{JS}(0) = -\infty$ and $Z'_H(0) = -\infty$.*

Proof We can check $Z' < 0$ and the values at 0 and 1 directly from the formulas: $Z'_\triangle = \frac{(z+3)(z-1)}{(1+z)^2}$, all factors positive except $z-1$; $Z'_{JS} = 1-\log_2(1+z^{-1}) < 0 \iff 2 < 1+z^{-1}$; and $Z'_H = 1-z^{-1/2}$. All these check out for $0 \le z < 1$. ■

Corollary 8 (Q-increasing) *$Q_\triangle(z), Q_{JS}(z)$, and $Q_H(z)$ are all strictly increasing in $(0, 1]$, with zero derivative at $q = 0$. Note $Q'_\triangle(1) = 2$, but $Q'_{JS}(1) = \infty$ and $Q'_H(1) = \infty$.*

Proof Increasing in $(0, 1]$ follows from Corollary 6. Derivative values at 0 and 1 can be checked using $Q'_\triangle = 2q$. $2Q'_{JS} = \log_2(1+q) - \log_2(1-q)$, and $\log_2(1-q) < 0$ for $q > 0$. $Q'_H = q/\sqrt{1-q^2}$. ■

For our three functions, a more complicated but straight-forward alternative is to show Q is increasing then check the stronger derivative conditions from Corollary 6.

Theorem 9 (Q-increasing-alt) *$Q_\triangle(z), Q_{JS}(z)$, and $Q_H(z)$ are all strictly increasing in $(0, 1]$, with zero derivative at $q = 0$. Note $Q'_\triangle(1) = 2$, but $Q'_{JS}(1) = \infty$ and $Q'_H(1) = \infty$.*

Proof Positive derivatives follow directly from the formulas. $Q'_\triangle = 2q$. $2Q'_{JS} = \log_2(1+q) - \log_2(1-q)/2$, and $\log_2(1-q) < 0$ for $q > 0$. $Q'_H = q/\sqrt{1-q^2}$. ■

Corollary 10 (Z-decreasing-alt) *$Z_\triangle(z), Z_{JS}(z)$, and $Z_H(z)$ are all strictly decreasing in $[0, 1)$, with zero derivative at $z = 1$. Note $Z'_\triangle(0) = -3$, but $Z'_{JS}(0) = -\infty$ and $Z'_H(0) = -\infty$.*

Proof Relying on Theorem 9 we check the conditions of Corollary 6. Recall $Q_\triangle = q^2$ and $Q'_\triangle = 2q$. Then $Q'_\triangle > Q_\triangle/(1+q) \iff 2q > q^2/(1+q) \iff 2q(1+q) > q^2 \iff q \neq 0$ and $2 > q$. Recall $Q_{JS} = \big((1+q)\log_2(1+q) + (1-q)\log_2(1-q)\big)/2$ and $Q'_{JS} = (\log_2(1+q) - \log_2(1-q))/2$. Then $Q'_{JS} > Q_{JS}/(1+q) \iff \log_2(1+q) - \log_2(1-q) > \log_2(1+q) + \frac{1-q}{1+q}\log_2(1-q) \iff 0 > \left(\frac{1-q}{1+q} + 1\right)\log_2(1-q)$. The first factor is positive and the second is negative for $q < 1$. Recall $Q_H(q) = 1 - \sqrt{1-q^2}$ and $Q'_H = q/\sqrt{1-q^2}$. Then $Q'_H > Q_H/(1+q) \iff q(1+q) > (1 - \sqrt{1-q^2})\sqrt{1-q^2} = \sqrt{1-q^2} - 1 + q^2 \iff 1 + q > \sqrt{1-q^2}$, which is true for $q > 0$ since the left side is increasing and the right is decreasing.

Values at 0 and 1 can be checked by recalling $Z'_\triangle = \frac{(z+3)(z-1)}{(z+1)^2}$, $Z'_{JS} = 1 - \log_2(1+z^{-1})$, and $Z'_H = 1 - z^{-1/2}$. ■

Theorem 11 $\frac{Z_1}{Z_2}$ *decreasing* $\iff$ $\frac{Q_1}{Q_2}$ *increasing. Moreover* $\max Z_1/Z_2 = \max Q_1/Q_2$ *and* $\min Z_1/Z_2 = \min Q_1/Q_2$

Proof Componentwise equality implies $\frac{D_1}{D_2} = \frac{Z_1}{Z_2}(z) = \frac{Q_1}{Q_2}\left(q = \frac{1-z}{1+z}\right)$ and max and min are preserved at $z = 0$ and $z = 1$ (where $q = 1$ and $q = 0$). ■

We next describe some bounds limiting how much these functions vary from one another. Then we give some examples where these bounds are nearly achieved.

4.6 Ratios of $\triangle_s$, JS and H_s^2

We start with describing linear bounds between the functions. Tighter bounds apply in a variety of situations. Many of these linear bounds are already known. For example the following lower bounds are stated in Jain and Srivastava (2007) without reference or proof. We hope providing straightforward descriptions and simple proofs here are helpful. In addition, the parameterization of the ratios by q and z, their monotonicity in these parameters, and geometrically describing their curves appears novel. Figure 10 summarizes the results of this section.

	Q^*		q^*	Z^*		z^*
$H_s^2/\triangle_s$	$1/2 =$	.500	0	$1/2 =$	.500	1
$JS/\triangle_s$	$1/2\log 2 >$	.721	0	$1/2\log 2 >$	.721	1
H_s^2/JS	$\log 2 >$	.693	0	$\log 2 >$	.693	1
max is 1 at $q = 1$ and $z = 0$						

Figure 10: Q^* and Z^* are the infimums of ratios between Q and Z functions, and q^* and z^* are the limit points where this is achieved. These results are exact.

Bounds on the ratio of Q (or Z) functions implies bounds on the ratio of actual distance functions D. If $1 \geq Q_1/Q_2 \geq a$ componentwise, then $1 \geq \max_k \frac{D_{1,k}}{D_{2,k}} \geq \frac{\|D_1\|_1}{\|D_2\|_1} \geq \min_k \frac{D_{1,k}}{D_{2,k}} = a$.

Theorem 12 (Ratio bounds $JS/\triangle_s$, $H_s^2/\triangle_s$, and H_s^2/JS)

$$1 \geq \frac{H_s^2}{\triangle_s} \geq 0.5$$

$$1 \geq \frac{JS}{\triangle_s} \geq \frac{1}{2\log 2}, \textit{ note } 1/2\log 2 > 0.721.$$

$$1 \geq \frac{H_s^2}{JS} \geq \log 2, \textit{ note } \log 2 > 0.693.$$

The maximum ratio of 1 *is achieved exactly when* $x \cdot y = 0$*, and the minimum ratio is approached as* $x \to y$*.*

Moreover $\frac{Z_H}{Z_\triangle}(z)$*,* $\frac{Z_{JS}}{Z_\triangle}(z)$ *and* $\frac{Z_H}{Z_{JS}}(z)$ *are decreasing* $\iff$ $\frac{Q_H}{Q_\triangle}(q)$*,* $\frac{Q_{JS}}{Q_\triangle}(q)$ *and* $\frac{Q_H}{Q_{JS}}(q)$ *are increasing.*

Proof Recall componentwise $\frac{D_1}{D_2}(x,y) = \frac{Z_1}{Z_2}(z) = \frac{Q_1}{Q_2}(q)$ so an upper or lower bound in the ratio of a component bounds the ratio of the one-norms of all components.

$JS/\triangle_s = 1/2 \ln 2 + q^2/12 \ln 2 + q^4/30 \ln 2 + \cdots$. This is obviously an increasing function of q, with value $1/2 \log 2$ at $q = 0$. At $q = 1$, recall the sum of the terms is 1 by property **Max 1** for JS. The same argument holds for $H_s^2/\triangle_s = 1/2 + q^2/8 + q^4 16 + \cdots$.

One can also prove limits on $Z_H/Z_\triangle$ directly without recourse to series. $Z_H/Z_\triangle = (1+z)(1-\sqrt{z})^2/(1-z)^2$. Let $w = \sqrt{z}$ and note $(1-w^2) = (1+w)(1-w)$. So $Z_H/Z_\triangle = (1+w^2)(1-w)^2/(1-w^2)^2 = (1+w^2)/(1+w)^2 = f(w)$. And $f'(w) = \big(2w(1+w)^2 - (1+w^2)2(1+w)\big)/(1+w)^4 = 2(w-1)/(1+w)^3$. This is < 0 for $w < 1$.

For H_s^2/JS, one might be tempted to consider the series expansions as well, but proving monotonicity from the two series is not so obvious.

It is easier for us to turn to the Z functions:

$$R(z) = \frac{Z_H}{Z_{JS}} = \frac{(1+z-2\sqrt{z})}{(1+z-\log_2(1+z) - z\log_2(1+z^{-1}))}.$$

We first evaluate R at its limits and then show it is decreasing.

We already know $Z_H(0) = Z_{JS}(0) = 1$ so $R(0) = 1$.

For $R(1)$, switching to Q and using the series expansions, after factoring out pq^2, the first terms give $\frac{Q_H}{Q_{JS}}(0) = \frac{1/2}{1/2 \ln 2} = \ln 2$. Some readers may find it instructive to consider that the leading q^2 terms also informs us of how many derivatives are required in a direct argument: $\lim_{z\to 1} \frac{Z_H}{Z_{JS}} = \frac{1+1-2}{1+1-1-1} = \frac{0}{0}$. Invoking L'Hôpital's rule we have $\lim_{z\to 1} \frac{Z'_H}{Z'_{JS}} = \frac{1-z^{-1/2}}{1-\log_2(1+z^{-1})} = \frac{0}{0}$. So invoking it again we get $\lim_{z\to 1} \frac{Z''_H}{Z''_{JS}} = \frac{\frac{1}{2}z^{-3/2}}{\frac{z^{-2}}{\ln 2(1+z^{-1})}} = \frac{1/2}{1/(2\ln 2)} = \ln 2$.

We now show that R is decreasing through repeated differentiation and checking values at $z = 0$ and 1 to eventually show that all the derivatives have the correct sign.

Using $1 + z^{-1} = (1+z)/z$ we rewrite $Z_{JS} = 1 + z - (1+z)\log_2(1+z) + z\log_2 z$.

$R'(z) = (1-z^{-1/2})(1+z-(1+z)\log_2(1+z)+z\log_2 z)-(1+z-2z^{1/2})(1-\log_2(1+z)+\log_2 z)/Z_{JS}^2$. Ignoring the positive denominator, we expand and cancel $1+z-(1+z)\log_2(1+z)$ terms to get $\operatorname{sgn}(R'(z)) = \operatorname{sgn}(R'_1(z))$, where $R'_1(z) = z\log_2 z - z^{-1/2} - z^{1/2} + z^{-1/2}(1+z)\log_2(1+z) - z^{1/2}\log_2 z - (1+z)\log_2(z) + 2z^{1/2} - 2z^{1/2}\log_2(1+z) + 2z^{1/2}\log_2 z$. Combining like log terms, and noting $z^{1/2} - z^{-1/2} = z^{-1/2}(z-1)$ we have $\operatorname{sgn}(R'(z)) = (z^{1/2}-1)\log_2 z + z^{-1/2}(z-1)(1-\log_2(1+z))$. We change variables with $w = \sqrt{z}$ yielding $R'(w) = 2(w-1)\log_2(w) + w^{-1}(w^2-1)(1-\log_2(w^2+1)$. Since we already know $R(w=1)$ and $R(w=0)$ we can restrict to $w \in (0,1)$. Since $w^2 - 1 = (w+1)(w-1)$, multiplying by $w/(w-1) < 0$ gives $\operatorname{sgn}(R'(w)) = -\operatorname{sgn}(R'_2(w))$ where $R'_2(w) = 2w\log_2(w) + (w+1)(1-\log_2(w^2+1))$.

Our goal is now to show $R'_2(w) \geq 0$. Note $R'_2(0) = 0 + 0 + 1 = 1$ and $R'_2(1) = -2 + 0 + 2 = 0$. So it suffices to show R'_2 is monotonic, i.e. $R''_2 \leq 0$.

$R''_2 = (1 + 2/\log 2) - 2w(w+1)/((1+w^2)\log 2) + \log_2(w^2/(1+w^2))$ after simplification. Note $R''_2(w \to 0) = -\infty$ and $R''_2(1) = 1 + 2/\log 2 - 4/2\log 2 + \log_2(1/2) = 0$. So it again suffices to show R''_2 is monotonic, i.e. $R'''_2 \geq 0$.

$R'''_2 = 2(w^2 - 2w - 1)/((1+w^2)^2\log 2) + 2/(w(1+w^2)\log 2)$. Since we are only interested in the sign, we drop the common positive $2/((1+w^2)\log 2)$ factor and simplify, $\operatorname{sgn}(R'''_2) = \operatorname{sgn}(R'''_3) = (w^3 - w^2 - w + 1)/(w(1+w^2))$. Dropping the new positive denominator we get $\operatorname{sgn}(R'''_2) = \operatorname{sgn}(R'''_4) =$

$w^3 - w^2 - w + 1 = (1-w)(1-w^2) \geq 0$, which was the goal. Going back up the chain of derivatives shows that each is monotonic and always has the correct sign. ■

Note that the relative difference bounds and curves in Figure 9 and Figure 8 follows directly from the ratio bounds.

4.7 Differences between $\triangle_s$, JS and H_s^2

Recall from Section 2.3 that bounds on ratios implies bounds on differences, namely $1 \geq d_1/d_2 \geq c \Rightarrow (1-c) \geq d_2 - d_1 \geq 0$. But here we prove tighter bounds on differences for the Z and Q functions. See Figure 8 and Figure 9. Figure 11 summarizes the results of this section.

	Q^*	q^*	Z^*	z^*	R bound
$F_\triangle - F_H$	$1/4 =$ .250	$\sqrt{3}/2 \approx$ .866	.270	.087	.5
$F_\triangle - F_{JS}$	.110	.807	.122	.127	.279
$F_{JS} - F_H$	.150	.912	.158	.055	.307
	min is 0 at $q = 0$, $q = 1$, $z = 0$, and $z = 1$				

Figure 11: Q^* and Z^* are the maximum difference between components of Q and Z functions, q^* and z^* are points near where this is achieved, and "R bound" is the weak bound on Q^* and Z^* implied by the ratio bounds. Except for the upper left, the results are not exact, but the actual maximum difference is provably $\leq Q^*$ ($\leq Z^*$).

Bounds on the difference in Q functions implies bounds on the difference in actual distance functions D. Componentwise, given $(Q_1 - Q_2) < a$, then since the difference between distances is a linear scaling of differences in Q, we have that componentwise $0 \leq D_1 - D_2 \leq \frac{p}{2}(Q_1 - Q_2) \leq ap/2$. Taking the 1-norm and noting that for our functions componentwise $D_1 - D_2 \geq 0$, we have $0 \leq \|D_1\|_1 - \|D_2\|_1 = \|D_1 - D_2\|_1 \leq a\|p/2\|_1 = a$.

The same argument applies for Z. (The only difference is $b\|u/2\|_1 \leq b$, not necessarily equality.) So whichever constant is smaller provides a tighter bound. In our case, the Q bound is always smaller, which is always the case f-divergences. Geometrically, the bound is the maximum difference in height for a constant q (or z) ray from the origin traveling on two different D functions; see Figure 7. For f-divergences these rays monotonically and linearly spread vertically as they travel, and they pass the $p = 1$ curve defining the Q functions before they hit the $u = 1$ curve defining the Z functions.

Define $MQ_{12} \equiv Q_1 - Q_2$ and $MZ_{12} \equiv Z_1 - Z_2$. Except in one case, we are unable to provide a closed form solution to a bound for these M functions. Instead, we provide numeric proofs of the following form. We show analytically that $M'' < 0$ in an open interval, and $M'' > 0$ in the open complement of that interval, with a single point at their shared boundary where $M'' = 0$. Hence there is a unique maximum in the interval. We find a bound on that maximum numerically. We find two points to either side of the maximum, and the ordinate of the intersection of their tangent lines bounds the maximum Q^* (Z^*). We compute this maximum, accounting for possible round-off error, and report it as a "provable" maximum value. The abscissa of the intersection point provides an approximation to the point q^* (z^*) at which the true maximum occurs. We try to get points close to q^* through a binary search. Although other search and bounding means are undoubtedly more efficient, the functions are well behaved enough that this approach suffices.

Theorem 13 $(Q_\triangle - Q_H) \leq 1/4$, *with equality at* $q = \sqrt{3}/2$.

Proof $MQ_{\triangle H} = q^2 - 1 + \sqrt{1-q^2} = \sqrt{1-q^2}(1-\sqrt{1-q^2})$. Let $w = \sqrt{1-q^2}$, we have $MQ_{\triangle H} = w(1-w)$ and $MQ'_{\triangle H} = 1-2w$. The maximum occurs at $w = 1/2 \iff q = \sqrt{3}/2$ and has value $1/4$. ■

Lemma 14 *$(Q_\triangle - Q_{JS})'' < 0$ for $q \in (q_0, 1)$ and > 0 for $q \in (0, q_0)$ with $q_0 \approx 0.53$.*

Proof $MQ'_{\triangle J} = 2q - (\log_2(1+q) - \log_2(1-q))/2$. $MQ''_{\triangle J} = 2 - (1/(1+q) + 1/(1-q))/(2\log 2)$. Therefore $MQ''_{\triangle J} < 0 \iff 4\log 2 < 1/(1+q) + 1/(1-q) = 2/(1-q^2) \iff q > (1 - 1/2\log 2)^{1/2} = q_0 \approx 0.53$. ■

Lemma 15 *$(Q_{JS} - Q_H)'' < 0$ for $q \in (q_0, 1)$ and > 0 for $q \in (0, q_0)$ with $q_0 \approx 0.72$.*

Proof $MQ'_{JH} = (\log_2(1+q) - \log_2(1-q))/2 - q/\sqrt{1-q^2}$. $MQ''_{JH} = (1/(1+q) + 1/(1-q))/(2\log 2) - (1-q^2)^{3/2} = (1-q^2)^{-1}/\log 2 - (1-q^2)^{3/2}$. Therefore $MQ''_{JH} < 0 \iff \sqrt{1-q^2} < \log 2 \iff q > \sqrt{1 - \log^2 2} = q_0 \approx 0.72$. ■

Lemma 16 *$(Z_\triangle - Z_H)'' < 0$ for $z \in (0, z_0)$ and > 0 for $z \in (z_0, 1)$ with $z_0 \approx 0.24$.*

Proof $MZ'_{\triangle H} = -4/(1+z)^2 + z^{-1/2}$. $MZ''_{\triangle H} = 8/(1+z)^3 - z^{-3/2}/2$. Therefore $\mathrm{sgn}(MZ''_{\triangle H}) < 0 \iff 16z^{3/2} < (1+z)^3 \iff 2\sqrt[3]{2}\sqrt{z} < 1 + z$. Letting $w = \sqrt{z}$ and solving via the quadratic formula we have $\iff w < \sqrt[3]{2} - \sqrt{\sqrt[3]{4} - 1} \Leftarrow z < 0.24$. ■

Lemma 17 *$(Z_\triangle - Z_{JS})'' < 0$ for $z \in (0, z_0)$ and > 0 for $z \in (z_0, 1)$ with $z_0 \approx 0.31$.*

Proof $MZ_{\triangle J}\prime = -4/(1+z)^2 + \log_2(1+z^{-1})$ and $MZ_{\triangle J}\prime\prime = 8/(1+z)^3 - 1/(z(z+1)\log 2)$. Therefore $MZ_{\triangle J}\prime\prime < 0 \iff z^2 + (2 - 8\log 2)z + 1 > 0$. Solving via the quadratic formula with $b = (2 - 8\log 2)$ we have $\iff z < b/ - \sqrt{b^2/4 - 1} = z_0 \approx 0.31$ ■

Lemma 18 *$(Z_{JS} - Z_H)'' < 0$ for $z \in (0, z_0)$ and > 0 for $z \in (z_0, 1)$ with $z_0 \approx 0.16$.*

Proof

From Theorem 3 we have $MZ_{JH} = (Z_{JS} - Z_H) = -\log_2(1+z) - z\log_2(1+z^{-1}) + 2\sqrt{z}$ and $MZ'_{JH}(z) = -\log_2(1+z^{-1}) + z^{-1/2}$.

Log functions (base 2 and of $1+z$) can cross square-root functions multiple times, so it is very helpful to recourse to MZ''_{JH} to avoid these difficulties.

$MZ''_{JH} = \frac{1}{z(1+z)\log 2} - \frac{1}{2z^{3/2}}$. Note $\frac{1}{z(1+z)\log 2} - \frac{1}{2z^{3/2}} > 0 \iff \frac{1}{z(1+z)\log 2} > \frac{1}{2z^{3/2}} \iff z(1+z)\log 2 < 2z^{3/2} \iff (1+z)\log 2 < 2z^{1/2}$. Since $z > 0$, we may square both sides, $\iff (1+z)^2 < \frac{4}{\log^2 2}z \iff z^2 + 2z - \frac{4}{\log^2 2}z + 1 < 0$. Note $\mathrm{sgn}(f'') = \mathrm{sgn}(-z^2 + 2(\frac{2}{\log^2 2} - 1)z - 1)$. Using the quadratic formula the zeros are $2c - 1 \pm 2\sqrt{c^2 - c}$ where $c = 1/\ln^2 2 = \{z_0, z_1\} \approx \{0.162, 6.163\}$. Thus MZ_{JH} has a single inflection point in $(0,1)$ at z_0. It is easy to verify numerically that $MZ''_{JH}(z : z < z_0) < 0$ and $MZ''_{JH}(z : 1 \geq z > z_0) > 0$. ■

4.8 Contours of $\triangle_s$, JS and H_s^2

4.8.1 Not Order Preserving Examples.

Measures $\triangle_s$, JS, H_s^2 (hence H_s) are not order preserving, yet their contours (constant-value sets) are similar.

Examples of order being switched can be generated by exploiting the different curvatures in Figure 3 bottom or the different coefficients in the series expansions. For $x = [0.89,\ 0.10,\ 0.01]$, $y = [0.9,\ 0,\ 0.1]$, and $z = [0.65,\ 0.35,\ 0]$, we have $(\triangle_s(x,y) = 0.087) < (\triangle_s(x,z) = 0.093)$ but $(JS(x,y) = 0.081) > (JS(x,z) = 0.072)$ and $(H_s^2(x,y) = 0.073) > (H_s^2(x,z) = 0.052)$. Changing z to $[0.6,\ 0.4,\ 0]$, we have $(JS(x,y) = 0.081) < (JS(x,z) = 0.095)$ but $(H_s^2(x,y) = 0.073) > (H_s^2(x,z) = 0.069)$.

Recall none of $\triangle_s$, JS, or H_s^2 obey the triangle inequality.

4.8.2 Worst-Case Contour Construction

Here we demonstrate where these ratio bounds may be nearly achieved. As before, equality is achieved when all functions are 1, at $x \cdot y = 0$.

The following construction nearly achieves the extremes of the inequality bounds for a single contour. Let $x = (a, a, ...a, b, b, ...b, 0)$ where $a = 1/(k-1) + \epsilon$ and $b = 1/(k-1) - \epsilon$. Let $y = (b, b, ...b, a, a, ...a, 0)$ and $z = (x_1, x_2, ...x_j, c, 0, 0, ...0, d)$ where j, c, and d are chosen so that $D_1(x,y) = D_1(x,z)$. Then as $k \to \infty$ and $\epsilon \to 0$ we have $D_2(x,y)/D_1(x,y) \to$ the least possible and $D_2(x,z)/D_1(x,z) \to 1$. Figure 12 illustrates trends. This example relies on several things; if any of these do not hold, tighter bounds are possible. First, it relies on the dimension k being large, and (second) the availability of zero components in x. Third, it relies on $x_i \approx y_i$ and hence (fourth) D_1 being small.

However, one can get fairly close to this worst case without being very extreme, as observed from the large flat section of the z-ratio curves in Figure 9, as long as we keep the availability of zero (or near-zero) components to provide points where the ratio is near 1.

k	ϵ	a	b	$\triangle_s$ (x,y)	$\frac{JS}{\triangle_s}$ (x,y)	$\frac{JS}{\triangle_s}$ (x,z)	$\frac{H_s^2}{\triangle_s}$ (x,y)	$\frac{H_s^2}{\triangle_s}$ (x,z)	JS (x,z')	$\frac{H_s^2}{JS}$ (x,y)	$\frac{H_s^2}{JS}$ (x,z')
∞	$\to 0$	$\to 0$	$\to 0$	$\to 0$	.721	1	.5	1	$\to 0$	.693	1
5	.01	.26	.24	.00160	.7215	.998	.5002	.997	.00115	.6932	.9989
5	.08	.33	.17	.102	.73	.91	.51	.83	.075	.70	.92
5	.16	.41	.09	.41	.78	.95	.57	.91	.320	.72	.97
9	.08	.205	.045	.41	.78	.998	.57	.997	.320	.72	.957

Figure 12: Near worst-case ratio constructions for contours.

For example, keeping only the second condition, choosing $k = 5$, $\epsilon = 0.08$, gives $x = (.33, .33, .17, .17, 0)$, $y = (.17, .17, .33, .33, 0)$, and $z = (.33, .33, .042, .128)$. Here $\triangle_s(x,y) = \triangle_s(x,z) = .102$ and $JS(x,y) = 0.075, JS(x,z) = .094$ and $H_s^2(x,y) = .053, H_s^2(x,z) = .085$ so $JS(x,y)/\triangle_s(x,y) = .73 \approx 0.721$, $JS(x,z)/\triangle_s(x,z) = .91 \approx 1$, $H_s^2(x,y)/\triangle_s(x,y) = .51 \approx 0.5$, and $H_s^2(x,z)/\triangle_s(x,z) = .83 \approx 1$.

Changing $z = (.33, .33, .17, .060, .110)$ gives $JS(x,y) = JS(x,z) = .075$ and $H_s^2(x,y) = .053, H_s^2(x,z) = .069$ so $H_s^2(x,y)/JS(x,y) = .70 \approx 0.693$, and $H_s^2(x,z)/JS(x,z) = .92 \approx 1$. Figure 12 describes some other variations we have computed. The last three columns describe extreme ratios between Hellinger-squared and Jensen-Shannon, for the same point y and a new point z' equidistant from x under Jensen-Shannon, $z' : JS(x,y) = JS(x,z')$.

5. Conclusions

We hope that our organization of properties is illuminating for those using distance functions over mixture models, and will inspire further geometric analysis. The proofs are detailed so as to be easily reproducible, which we thought would be useful given our attempts to find them in the literature and the difficulty of combining log and square-root functions and various powers.

We have given an algebraic and geometric comparison of the components of $\triangle_s$, JS, and H_s^2. We have factored these functions into more easily comparable forms, in the process illuminating their dependence and behavior on features of the points. We have provided theoretical bounds on componentwise ratios and differences, and provided concrete examples that nearly achieve the ratio bounds. However, much work remains.

We have provided linear bounds on ratios $JS/\triangle_s$, etc., but the functional forms suggest it is possible to derive tighter nonlinear bounds. For example, the similarity of contours $\triangle_s = c_1$ and $JS = c_2$ might depend on c_1 and c_2 and be closer together than the linear bounds enforce. Our constructions show that getting close to the worst case ratios is fairly easy if some components are near zero. It would be interesting to explore worst-case constructions where none of the components are near-zero. We have not yet tried worst-case difference constructions.

We wish to explore further the geometric properties of the square root transformation in Hellinger. Using the same types of arguments for $\triangle_s$, JS, and H_s^2, we speculate that we could develop similar bounds between Hellinger, Euclidean, and two Geodesic distances. Other bivariate distances such as the Jaccard index and Tanimoto coefficient are worth analyzing. We wish to explore these types of comparisons for the multi-stage distances (e.g. Rubner et al. (1998)) and univariate distances (e.g. Meilă (2003)).

Acknowledgments

Thanks to David Robinson for providing some introductory material. Thanks to David Robinson, Andy Wilson, and Philip Kegelmeyer (all Sandia) for helping me understand LDA and mixture models and the state of clustering validation. Thanks to Brett Bader (Sandia) for help with LSA. And thanks to them all for promoting these approaches for cybersecurity and non-proliferation at Sandia, and helping me understand those applications. Thanks to Prof. George Michailidis (U. Michigan) and Joel Vaughan for helping me understand the context of this work in the statistics community. Thanks to Prof. Vageli Coutsias (U. New Mexico) for help with the Z series expansions.

Sandia is a multiprogram laboratory operated by Sandia Corporation, a Lockheed Martin Company, for the United States Department of Energy's National Nuclear Security Administration under Contract DE-AC04-94-AL85000.

References

Peter Barnum. Similarity and difference: Advanced perception course presentation. `http://www.cs.cmu.edu/~efros/courses/AP06/.../06-07-presentation.ppt`, 2006.

Michael W. Berry, Susan T. Dumais, and Gavin W. O'Brien. Using linear algebra for intelligent information retrieval. *SIAM Review*, 37(4):573–595, 1995. doi: 10.1137/1037127. URL `http://link.aip.org/link/?SIR/37/573/1`.

D. M. Blei and J. D. Lafferty. A correlated topic model of *Science*. *The Annals of Applied Statistics*, 1:17–35, 2007.

Imre Csiszár. Information-type measure of difference of probability distributions and indirect observations. *Studia Sci. Math. Hungar.*, 2:299–318, 1967.

Sever Dragomir. *Inequalities for Csiszár f-Divergences in Information Theory.* Research Group in Mathematical Inequalities and Applications, 2001.

A. L. Gibbs and F. E. Su. On choosing and bounding probability metrics. *International Statistical Review*, 70:419–435, 2002. doi: 10.1111/j.1751-5823.2002.tb00178.x.

James Hafner, Harpreet S. Sawhney, Will Equitz, Myron Flickner, and Wayne Niblack. Efficient color histogram indexing for quadratic form distance functions. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 17:729–736, 1995. ISSN 0162-8828. doi: http://doi.ieeecomputersociety.org/10.1109/34.391417.

E. Hellinger. Neue begrndung der theorie quadratischer formen von unendlichvielen vernderlichen. *Reine Angew. Math.*, 136:210–271, 1909.

T. Hofmann. Probabilistic latent semantic analysis. *Uncertainty in Artificial Intelligence, UAI'99 Stockholm*, pages 289–296, 1999.

K. C. Jain and Amit Srivastava. On symmetric information divergence measures of Csiszar's f-divergence class. *Journal of Applied Mathematics, Statistics and Informatics (JAMSI)*, 3(1): 85–99, 2007.

S. Kullback and R.A. Leibler. On information and sufficiency. *Annals of Mathematical Statistics*, 22(1):79–86, 1951. doi: http://dx.doi.org/10.1214/aoms/1177729694.

E. Levina and P. Bickel. The earth mover's distance is the Mallows distance: some insights from statistics. *Proceedings. Eighth IEEE International Conference on Computer Vision. ICCV 2001.*, 2:251–256, 2001.

Marina Meilă. Comparing clusterings by the variation of information. *Learning Theory and Kernel Machines*, pages 173–187, 2003.

Karl Pearson. On the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling. *Philosophical Magazine*, 50:157–175, 1900.

S. Peleg, M. Werman, and H. Rom. A unified approach to the change of resolution: Space and gray-level. *IEEE Trans. Pattern Anal. Mach. Intell.*, 11(7):739–742, 1989. ISSN 0162-8828. doi: http://dx.doi.org/10.1109/34.192468.

David G. Robinson. Statistical language analysis for automatic exfiltration event detection, SAND2010-2179. Technical report, Sandia National Laboratories, 2010.

Yossi Rubner, Carlo Tomasi, and Leonidas J. Guibas. The earth mover's distance as a metric for image retrieval. Technical report, Stanford University, Stanford, CA, USA, 1998.

Yossi Rubner, Jan Puzicha, Carlo Tomasi, and Joachim M. Buhmann. Empirical evaluation of dissimilarity measures for color and texture. *Computer Vision and Image Understanding*, 84: 25–43, 2001. doi: 10.1006/cviu.2001.0934. URL `http://www.idealibrary.com`.

Xiaojun Wan. A novel document similarity measure based on earth mover's distance. *Inf. Sci.*, 177 (18):3718–3730, 2007. ISSN 0020-0255. doi: http://dx.doi.org/10.1016/j.ins.2007.02.045.

Afra Zomorodian and Gunnar Carlsson. Computing persistent homology. In *SCG '04: Proceedings of the Twentieth Annual Symposium on Computational Geometry*, pages 347–356, New York, NY, USA, 2004. ACM. ISBN 1-58113-885-7. doi: http://doi.acm.org/10.1145/997817.997870.

Appendix A. Data Model and Application Context

One interesting source of mixture models is the analysis of a corpus of text documents using statistical techniques. For example, one might consider the corpus of math and computer science papers from the last five years, and be interested in seeing how well this paper (the one you are reading now!) clusters with the machine learning literature, or whether it is an outlier as the author suspects.

In order to answer such a question, one selects an appropriate model (an art full of choices) and then uses a mathematical computer program to turn documents into data points in some space: Latent Dirichlet Allocation and Latent Semantic Analysis are common choices. Next the points are clustered. (In some contexts the output of LDA is considered clustered already by the largest topic component.) But in order to cluster points (i.e. documents), some notion of *distance* between points is required.

However, which of the many distance functions should you choose? Current practice is that the distance function is chosen by some combination of four criteria. First, do you reproduce the ground truth? This is only possible if ground truth is available and trustworthy. E.g. you might consider the journal that a paper was published in as the ground truth of what cluster it belongs in. Unfortunately this confounds the choice of distance function with the choices of methods and other parameters. Second, you consider the stability of the outcome as in cross validation. Are similar clusterings produced when some data are withheld, or the distance threshold is varied, etc.? Third, you pick the distance function that has been historically used for your application domain. Despite the obvious shortcomings, this facilitates evaluating new work, and leverages the insight your application community has built up about your distance function. Fourth, you pick the distance function based on information theory, the idea that the distance is measuring something relevant such as entropy. This often coincides with historical application practice. This paper ignores all of the above (very reasonable) criteria, and instead considers complementary and foundational first-principles geometric and algebraic comparisons.

The computational geometry community has not historically focused on statistical distances, except for image analysis (Rubner et al., 1998). Even there, the study is usually based on evaluating the outcome as described above by the four criteria.

A.1 Other Uses of Distances

Clustering is a zero-dimensional structure, and distances can also be used to produce more nuanced structures. Examples of higher dimensional discrete structures include building a graph: connect two vertices with an edge if their distance is less than some threshold. Also building a simplicial complex, adding a simplex spanning points contained in a ball as in the Čech complex. One can build a whole family of discrete structures, filtered by an increasing distance threshold, as in persistent homology (Zomorodian and Carlsson, 2004) or alpha-complexes.

A.2 Other Types of Distances

There are other distance types for collections of mixture models that are outside the scope of this paper, but are worth mentioning.

Point-to-point distances can be conflated with some notion of distance (other than orthogonality) between the coordinate axes or histogram bins, called "cross-bin similarity." This is natural in the setting of LDA, where each coordinate represents a topic, and the topics themselves reside as mixture model points in a high-dimensional word space with meaningful distances between topics. Wan (2007) has used the Earth Mover's Distance in exactly this way for combining document similarity with sub-topic similarity. The Earth Mover's Distance (Peleg et al., 1989) a.k.a. Mallows Distance (Levina and Bickel, 2001) forms a certain product of these distances after solving a linear program. The Quadratic form (Hafner et al., 1995) is an alternative using a different product, without the

linear program. Rubner et al. (2001) compares nine distances, including some cross-bin similarities, in the application context of image comparisons; Barnum (2006) provides a nice slide summary.

Combining distances also arises in the setting where each point represents a structured histogram, humans have selected the bins, and the meaning of the bins of the histogram are more or less related. For example, in cybersecurity, one could build a histogram of features of packet headers. One might want to assign the bin for "day of the week" to have a smaller distance to the bin for "time of day" than the bin for "packet size."

Meilă Divergence (Meilă, 2003) measures the similarity between partitions based on entropy and mutual information. That is, it is useful to compare the quality of different clusterings, in contrast to whatever distance and method was used to create the clusterings in the first place.

Univariate measures (i.e. for single points) have their uses as well. For example in community detection, an entropy measure of a sub-graph may help one decide whether it is a community or should be further subdivided, and it may not matter how the entropy of two disjoint subgraphs compare.

A.3 Model Generation

Recall the problem of determining the relationship of this paper (document) to others in a corpus of journal papers (documents). A document is considered to be composed of a collection of words: a bag of words, where word order and grammar are ignored. Much art is devoted to selecting the words to keep. For example, one might throw away common words like "the." One might retain just the stem of words, obviously helpful for ignoring tense, but also emphasizing word roots and meaning by treating "weighting," "unweighted," and "weightier," all as "weight." The retained words in the bag are then weighted to produce entries in a document-word matrix C. Weighting is also an art, e.g. weights equal to frequency of occurrence are not as distinguishing as weights equal to entropy of occurrence. These approaches have proven very effective, despite the obvious information loss.

A.4 Statistical Model, LDA

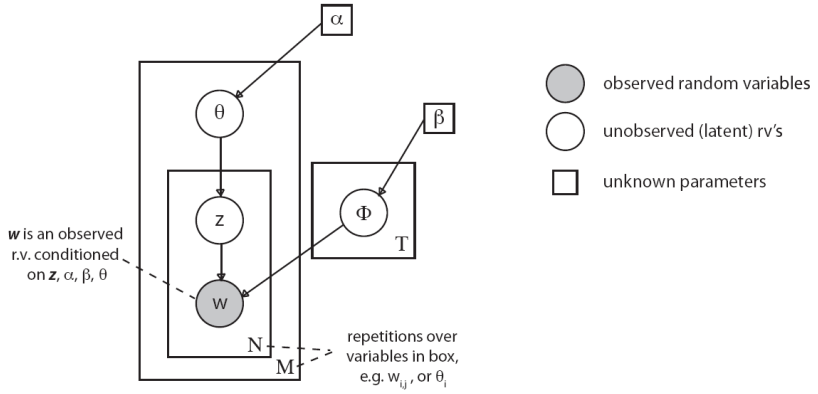

Figure 13: LDA implied Bayesian hierarchical structure. Theta and Phi are from Dirichlet distributions parameterized by multivariate alpha and beta: $\theta \sim Dir(\alpha)$, $\phi \sim Dir(\beta)$. The goal is to discover the unobserved circled quantities from the observed ones. Figure courtesy of Robinson (2010).

Latent Dirichlet Allocation (LDA) takes this document-word matrix C (corpus) and produces a topic-word matrix ϕ (which we will ignore) and a document-topic matrix θ (our data); see "Probabilistic Latent Semantic Analysis" in Figure 14 and Robinson (2010). Each document-column of

θ is a mixture of topics, the contribution of each topic to that document. The matrix product $\phi\theta$ gives the probability distribution over the vocabulary. This is in contrast to an approximation of the word counts.

LDA assumes a hidden generative model. The topics are hidden variables. The "true" underlying hidden model for each document is assumed to be a sequence of topics, of length equal to the number of words in that document. Topics may be repeated in this sequence. Each topic instance in the sequence randomly generated one of its words and contributed it to the document; these words are the observed data (Hofmann, 1999). (Despite the document word order being ignored, the correlations of the probability model recapture some word context.) This gives a hierarchical Bayesian framework; see Figure 13. Further, the model assumes that θ and ϕ come from a Dirichlet distribution, hence the name LDA.

Given this LDA model, a statistical computer program is used to take the observed words and discover (estimate) the hidden topics, both their presence in the documents θ and the probabilities with which they generated each word ϕ. Approximate inference methods for LDA is an active area of research with a large literature. Gibbs sampling is a popular way to accomplish the estimation process.

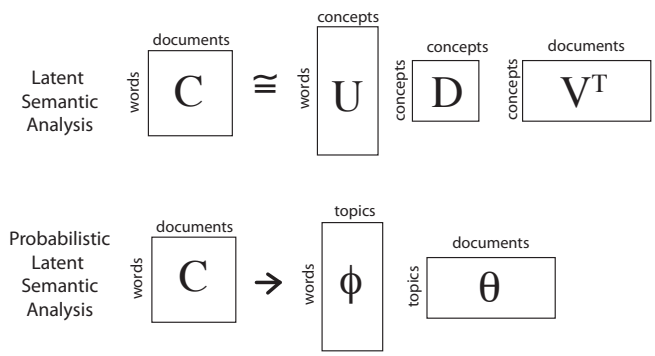

Figure 14: LDA (pLSA) and LSA derivations. Figure courtesy of Robinson (2010).

A.5 Algebraic Model, LSA

Other methods also produce geometrically similar output. The linear-algebra (non-statistical) technique Latent Semantic Analysis (LSA) (Berry et al., 1995) starts with similar input, a corpus of documents. A document-by-word (weighted) incidence matrix C (corpus) is formed as before. A singular value decomposition (SVD) produces several matrices; see Figure 14. Here C is well-approximated by the matrix products. For the word by concept matrix, columns are orthogonal concept vectors in word-space, with both positive and negative entries; in contrast, the topics from LDA are not orthogonal and have only non-negative entries. For the document by concept matrix, each row is the (positive and negative) coordinates of a document in concept-space. Documents in topic-space are often compared using cosine similarity. This measures the angle at the origin between points.