Understanding How Choices Change Clusterings: Geometric Comparison of Popular Mixture Model Distances

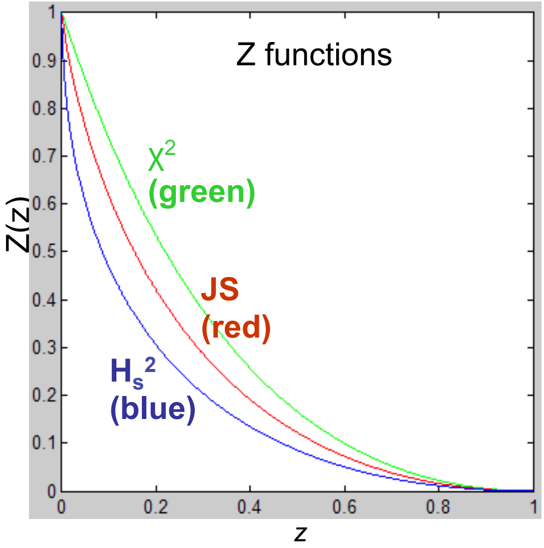
Scott A. Mitchell
http://www.cs.sandia.gov/~samitch

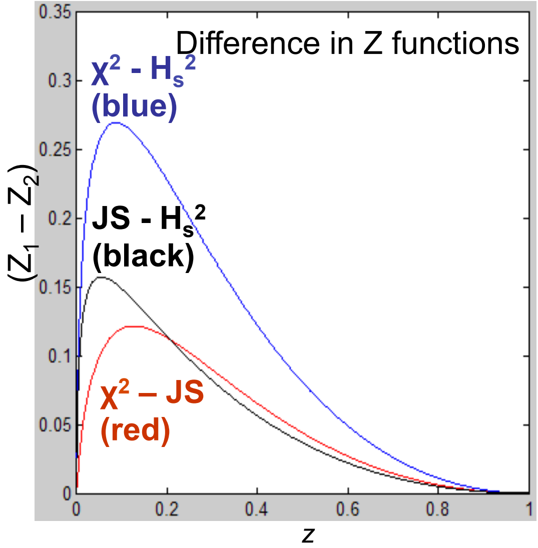
Computer Science and Informatics Department (1415)
Sandia National Laboratories

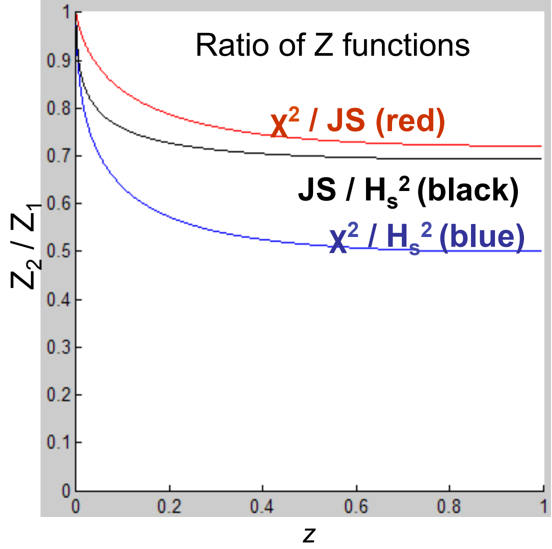
Sandia is a multiprogram laboratory operated by Sandia Corporation, a Lockheed Martin Company,
for the United States Department of Energy’s National Nuclear Security Administration
under contract DE-AC04-94AL85000.

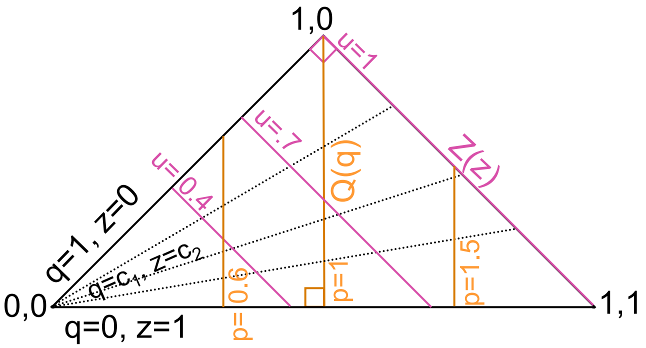
Preview Summary

$$\begin{tikzcd}
 & & & & \sqrt{x} & \frac{x}{\|x\|_2} \\
\left\|\frac{\max(x,y)}{2} Z\left(\frac{\min(x,y)}{\max(x,y)}\right)\right\|_1 & & & & C_{\sqrt{}} \arrow[d, phantom, "="] & C_u \arrow[d, phantom, "="] \\
 & \chi^2 \arrow[r, phantom, "\geq"] & JS \arrow[rr, phantom, "\geq"] & & H_s^2 \arrow[d, leftrightarrow, "\sqrt{\cdot}"', "\geq"] & E_s^2 \arrow[d, leftrightarrow, "\geq"', "\sqrt{\cdot}"] \\
\left\|\frac{(x+y)}{2} Q\left(\frac{x-y}{x+y}\right)\right\|_1 & & & & H_s \arrow[d, leftrightarrow, "\text{arcsin}"', "\leq"] & E_s \arrow[d, leftrightarrow, "\leq"', "\text{arcsin}"] \\
\left\|\sum_{n=1}^{\infty} a_n \frac{(x-y)^{2n}}{(x+y)^{2n-1}}\right\|_1 & & & & G_{\sqrt{}s} & G_{us}
\end{tikzcd}$$







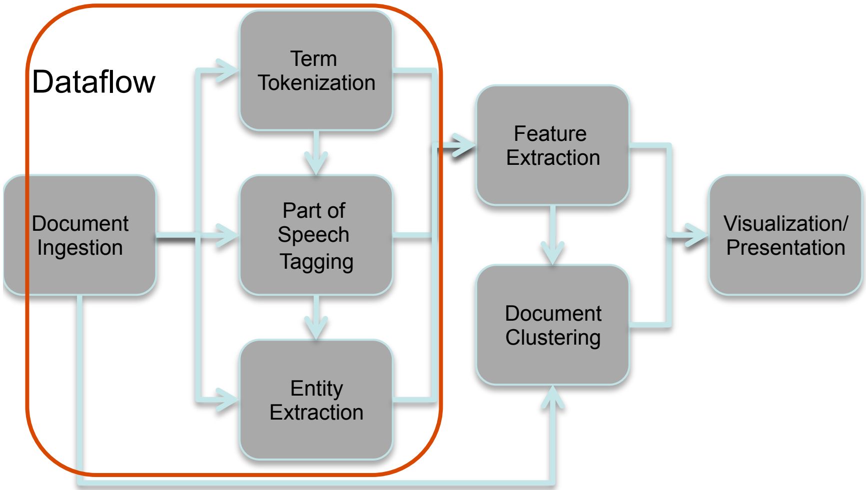


Outline

- Application context
- Distances
- 3d plots
- Algebraic reformulation as Z, Q
- Ratios
- Worst-case difference construction

Text Document Clustering at Sandia

Problem: given a pile of documents, categorize them
- so you only have to read one category
- to identify outliers not in any category



Dataflow

Document Ingestion

Term Tokenization

Part of Speech Tagging

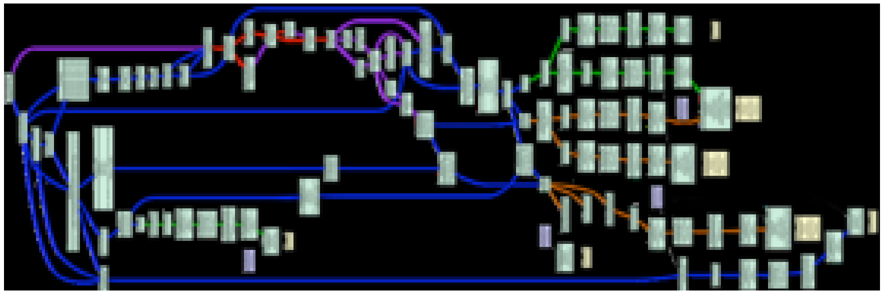
Entity Extraction

Feature Extraction

Document Clustering

Visualization/ Presentation

Example Software:
Prototype-2 from Network Grand Challenge LDRD



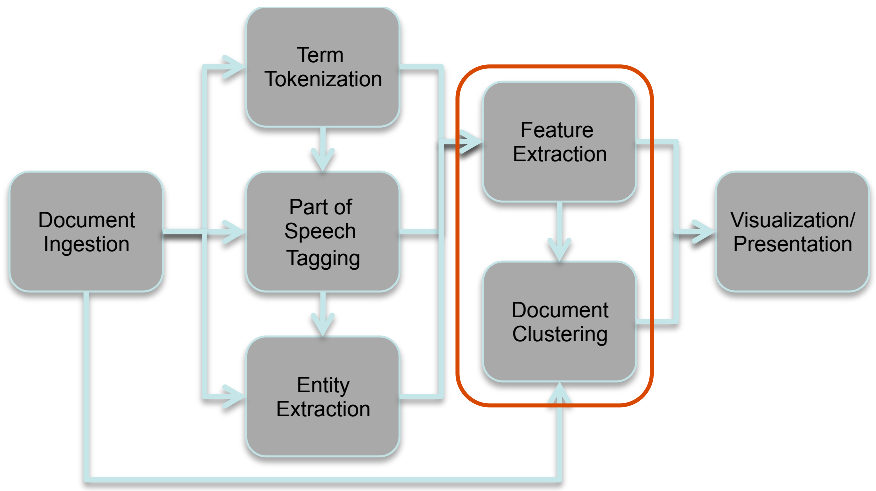
Example: How strange is my technical paper, the one I’m giving this talk on?

- Google Scholar search “mixture model geometry”, get 529,000 hits, save them all to local disk
- Throw out “the”, “and”, “however” to de-noise. Stem “geometrical”, “geometry”, to “geometry”.
- Identify grammar, to help answer “expository or framing?”
- Identify authors, institutions, to help identify relationships.

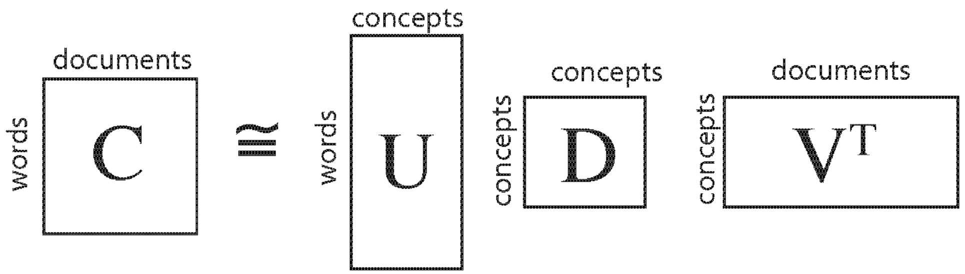
- Each document is a bag of “words”, unordered

Source Credit: Danny Dunlavy

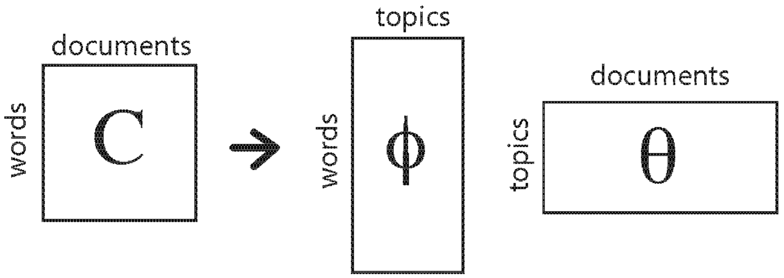
Reduce word-space to feature-space
e.g. 20,000 dimensions to 50

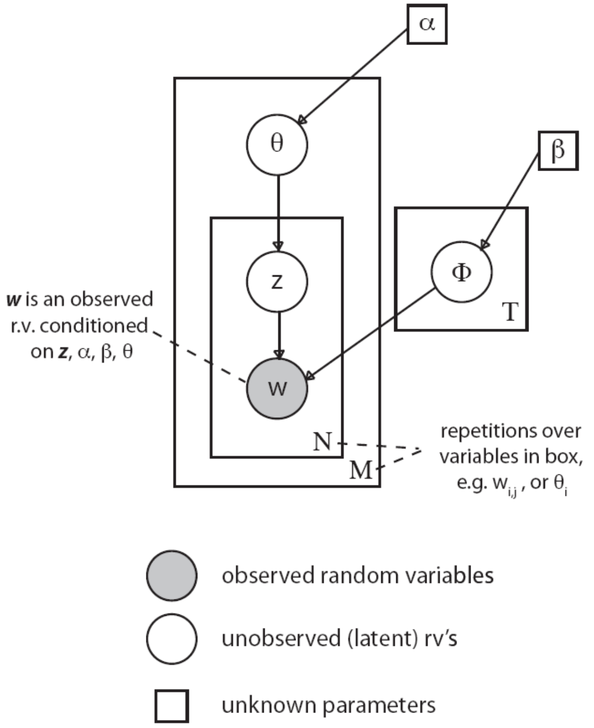


Latent Semantic Analysis

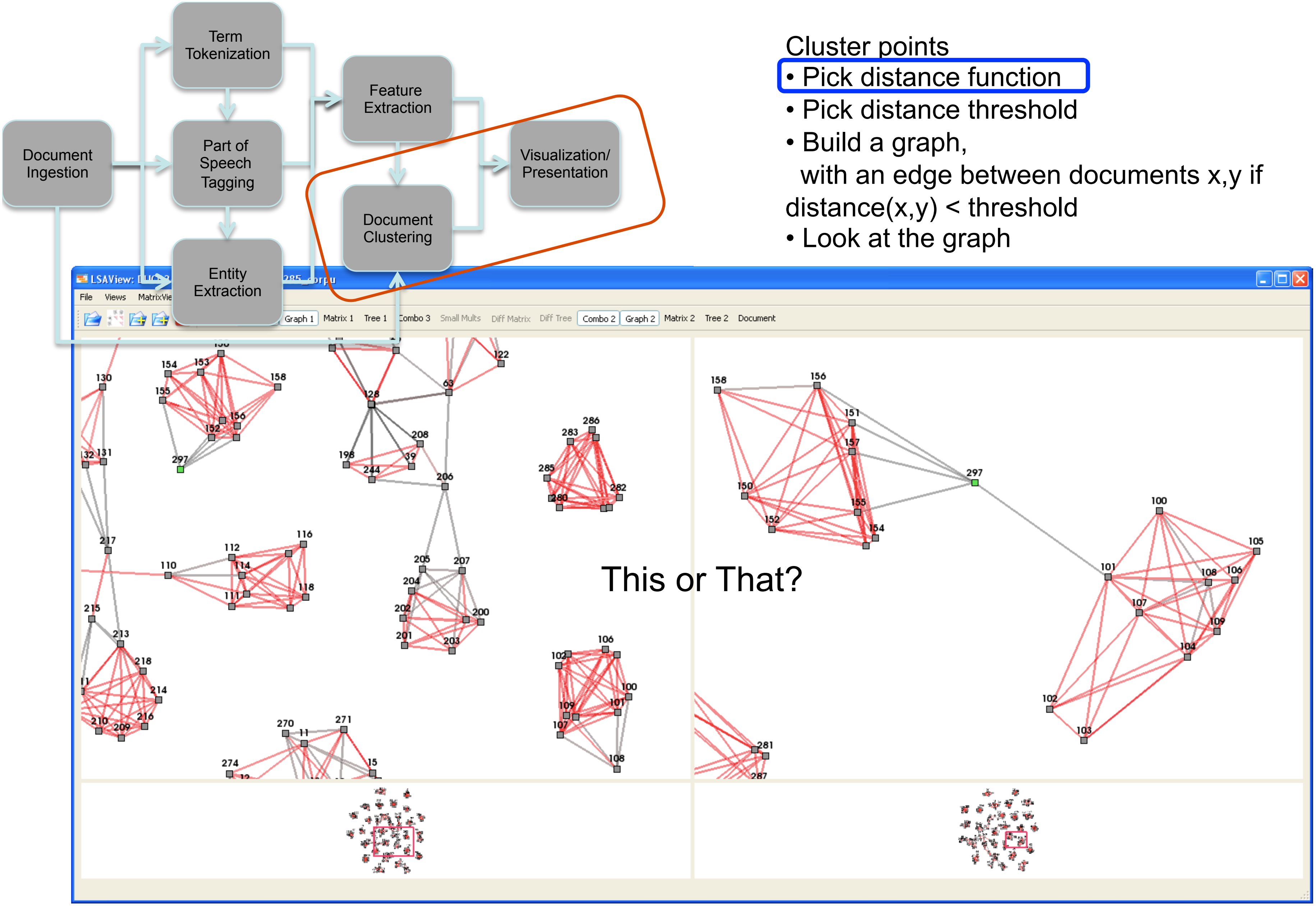


Probabilistic Latent Semantic Analysis





Source Credit: David G. Robinson



This or That?

Cluster points
- Pick distance function
- Pick distance threshold
- Build a graph,
  with an edge between documents x,y if
distance(x,y) < threshold
- Look at the graph

What happens if I tweak one of the 50+ knobs on this Frankenstein? Why? Predictable changes? What does it all mean?

Need coherent research program. Distances are one puzzle piece.

Which Distance?

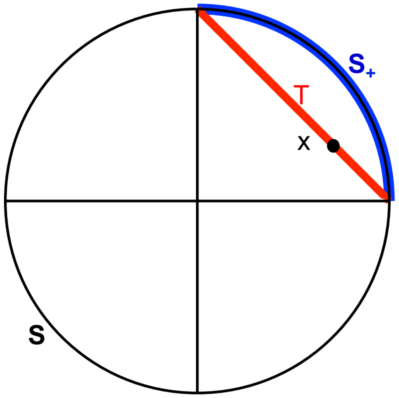
- Criteria
  - Reproduce ground truth?
    - What ground truth? Journal I sent my paper to? Expert opinion?
  - Stability of outcome?
    - stability ≠ accuracy
  - Use the one this application area always uses?
    - Maybe not such a bad idea: leverage insight, one knob at a time comparisons
  - Information theory?
    - “I think entropy is relevant and this distance measures it”
- Let’s try something else
  - Does it even matter? When do two distance functions give a different ordering to points?
  - Geometry and algebra

Generality

- “Documents” could be any pile of data
- “Words” could be any discrete categorical features you care about
- “Graphs” could have more structure: filtered simplicial complexes;
  or less: proximity to cluster center
- Any dimension K – typically > 50
  e.g. documents in 50-d concept space, or concepts in 20,000-d wordspace
- Applications
  - Cluster cyber-traffic based on header features, content analysis.

Specialization

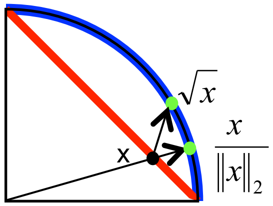
- Bivariate distances
  - Not convoluting document-in-conceptspace with concept-in-wordspace.
  - Not univariate measure of points.
  - Not univariate measure of partition as Graph Entropy (Berry, Phillips)
- Distances between points which are mixture models



$$\sum_{k=1}^{K} x_k = 1, \text{ and } 1 \geq x_k \geq 0.$$

contrast to LSA on S

sometimes distances project to positive part of sphere S_+



Outline

- Application context
- Distances
  - properties
- 3d plots
- Algebraic reformulation as Z, Q
- Ratios
- Worst-case difference construction

Distance Properties

100’s of distances to choose from.

Bivariate: D(x,y). Not D(x). Not D(partition).

Where variants exist, pick the ones with

0. **Unique Zero**: $D(x,y) = 0$ 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 $x \cdot y = 0$.

Most won’t satisfy 3.

Iter-Distance Properties

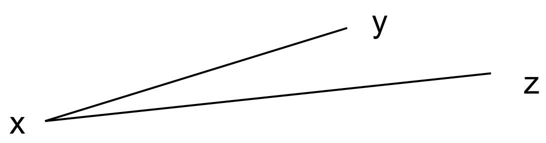
- What matters is ordering of points, level-set shape
  same level sets → same clusterings

Two distances D and F have (are)

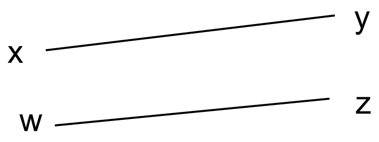
stronger

- **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 $D(x,y) < D(x,z) \iff F(x,y) < F(x,z)$.
- **Global Order Preserving**: if $D(x,y) < D(w,z) \iff F(x,y) < F(w,z)$.

- Stronger properties don't hold, but we'll see how they don't.
  - Max 1, Bounded Ratio c_2 → Bounded Difference $c_1 = 1-c_2$
    but we'll show smaller c_1



Local order preserving:
D_1 D_2 agree which of {y,z} is closer to x



Global order preserving:
D_1 D_2 agree which of (x,y) and (w,z) is smaller

Outline

- Application context
- Distances
  - properties
  - the distances
- 3d plots
- Algebraic reformulation as Z, Q
- Ratios
- Worst-case difference construction

Chi-Squared χ^2

- Base

$$\chi^2(x,y) = \sum_{k=1}^{K}\frac{(x_k-y_k)^2}{y_k} = \left\|\frac{(x-y)^2}{y}\right\|_1$$

- Problems
  - unbounded as y→0, no Max 1
  - unsymmetric

0. **Unique Zero**: $D(x,y)=0$ 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 $x\cdot y=0$.

- Fix (a.k.a Trianglar Discrimination)

χ^2(observed, expected) expected = midpoint if same distribution

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

$p\equiv x+y$

$d\equiv|x-y|$

- View: f(x,0) = x, const x+y, x, head-on, iso-curves
  - fig 1, 2, 3

Chi-Squared χ^2

- Componentwise
  Algebraic Properties

$$\chi^2 = \frac{1}{2}\frac{(x-y)^2}{(x+y)}$$

- Q
  p = x+y; d = x-y; q = d/p
  p in [0,2]; d,q in [0,1]

$$\chi^2 = \frac{p}{2}Q(q) = \frac{p}{2}q^2$$

- Z
  u = max(x,y); v = min(x,y); z = u/v
  u,v,z in [0,1]

  revisit figures 1,2,3

$$\chi^2 = \frac{u}{2}Z(z) = \frac{u}{2}\frac{(1-z)^2}{(1+z)} = \frac{u}{2}\left(1+z-\frac{4z}{1+z}\right)$$

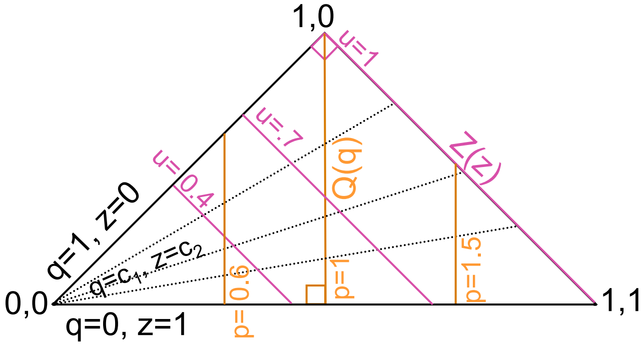
- Further study: f-divergence, Csiszár(1967), Dragomir(1980+) RGMIA

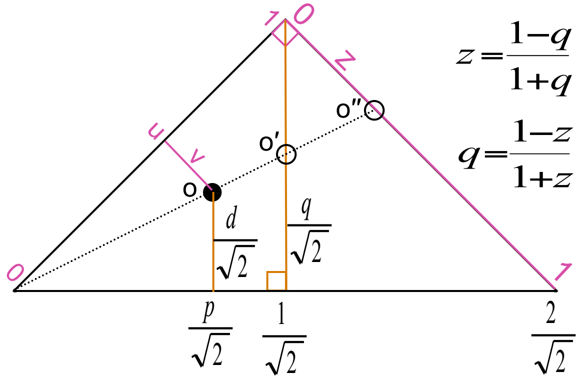
$$D_f(a,b) = \left\| af\left(\frac{b}{a}\right)\right\|_1$$

Chi-Squared χ^2

- Componentwise Geometric Interpretation

labels are lengths of segments, except points o, o′, o″

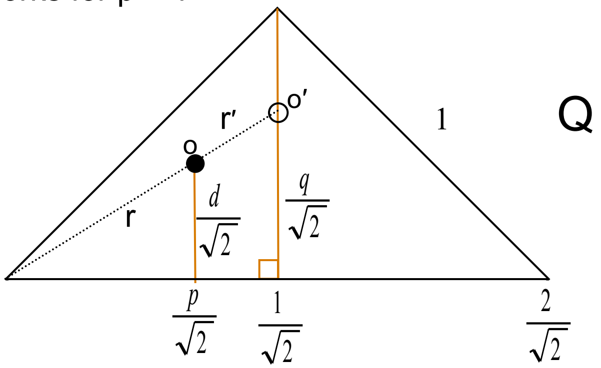




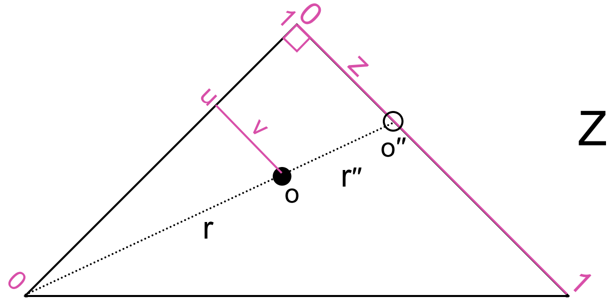
$$z=\frac{1-q}{1+q}$$

$$q=\frac{1-z}{1+z}$$

geometrically: D(•) = (r / r’) D(o) = (p/1) D(o) = p Q(q)/2
also works for p > 1

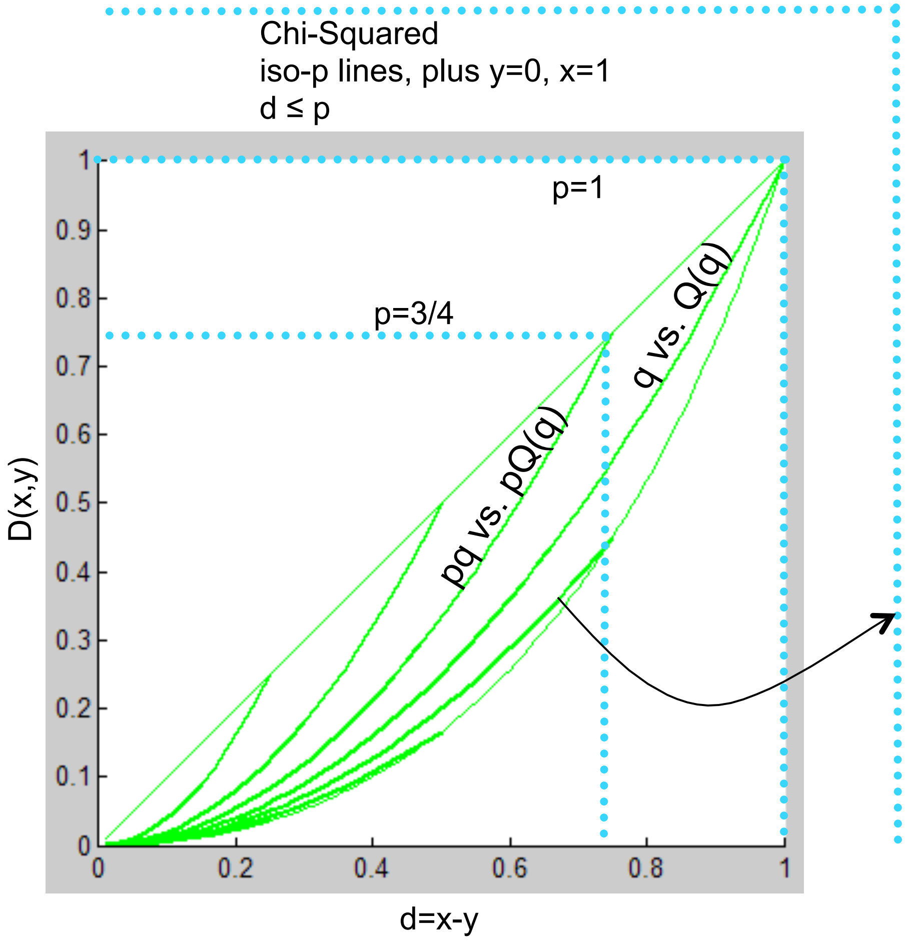


geometrically: D(•) = (s / s’) D(o) = (u / 1) D(o) = u/2 Z(z)



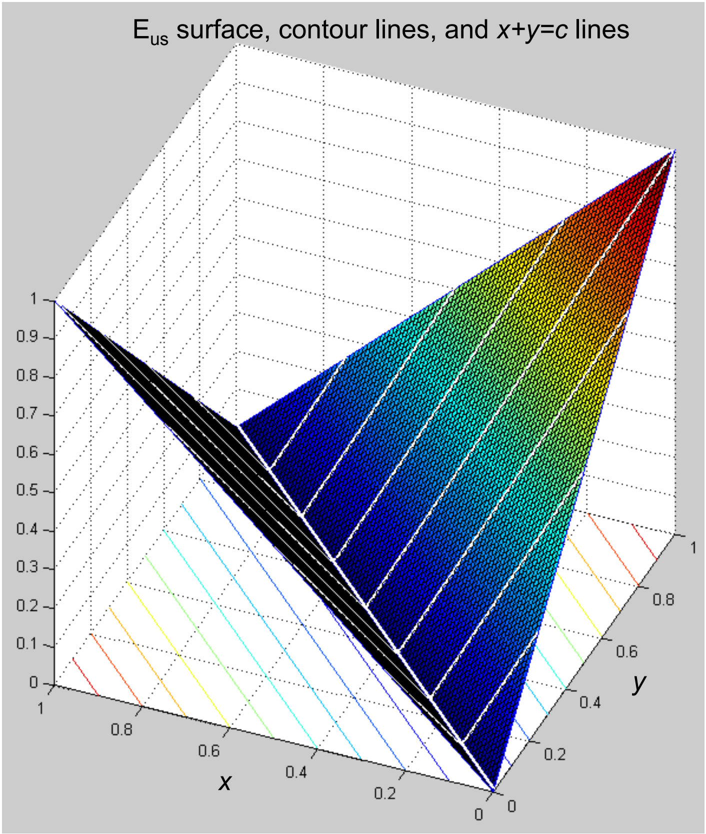
Chi-Squared χ^2

- Q curves Z same



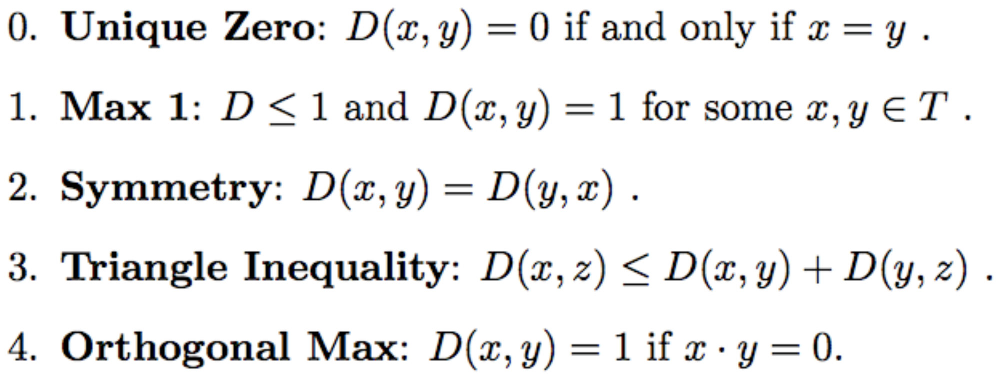
Chi-Squared
iso-p lines, plus y=0, x=1
$d \leq p$

Compare to Euclidean, function of d only, no p dependence



E_{us} surface, contour lines, and $x+y=c$ lines

Jenson-Shannon (same form as χ^2)



0. **Unique Zero**: $D(x,y) = 0$ 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 $x \cdot y = 0$.

Figure 4

- Kullback-Leibler

$$K(x,y) = \sum_{k=1}^{K} x_k \log_2\left(\frac{x_k}{y_k}\right)$$

  - measures entropy, information theoretic
  - non-symmetric
  - unbounded
- Jenson-Shannon Fix

$$JS(x,y) = \sum_{k=1}^{K}\left[x_k \log_2\left(\frac{2x_k}{y_k + x_k}\right) + y_k \log_2\left(\frac{2y_k}{y_k + x_k}\right)\right]$$

  - x=0…
  - x=y…
  - One of the terms can be negative, but JS ≥ 0, Unique Zero holds
  - [illegible]
    - stronger, factor as Chi-Squared $JS_s(x,y) = \left\| \frac{u}{2} Z_{JS}\left(\frac{v}{u}\right) \right\|_1 = \left\| \frac{p}{2} Q_{JS}\left(\frac{d}{p}\right) \right\|_1$

Hellinger

- Hellinger

$$H^2 = \sum_{k=1}^{K} \left(\sqrt{x_k} - \sqrt{y_k} \right)^2 = \left\| \left(\sqrt{x} - \sqrt{y} \right)^2 \right\|_1$$

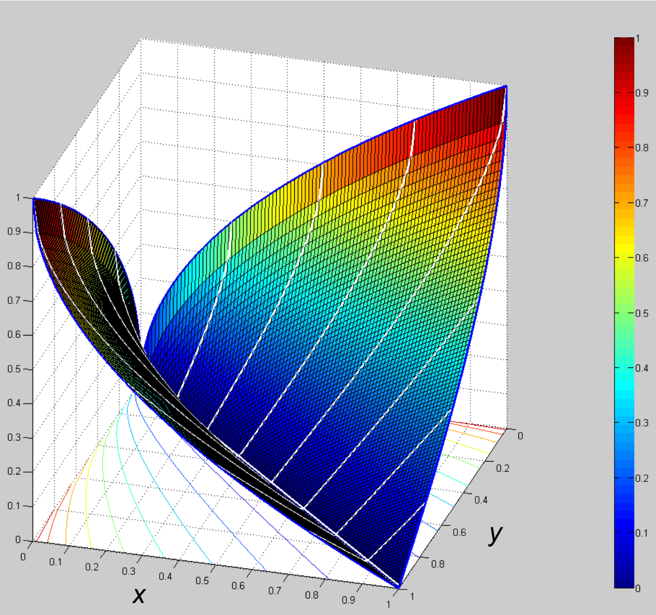
0. **Unique Zero**: $D(x, y) = 0$ if and only if $x = y$.
1. **Max 1**: $D \leq 1$ and $D(x, y) = 1$ for some $x, y \in T$.
2. **Symmetry**: $D(x, y) = D(y, x)$.
3. **Triangle Inequality**: $D(x, z) \leq D(x, y) + D(y, z)$.
4. **Orthogonal Max**: $D(x, y) = 1$ if $x \cdot y = 0$.

  - H satisfies Triangle Inequality, H^2 doesn't
  - x=0…
  - x=y…
  - $H^2(ax, ay) = aH^2(x, y)$
    - Factor as Chi-Squared, JS

$$H_s^2(x,y) = \left\| \frac{u}{2} Z_H\left(\frac{v}{u} \right) \right\|_1 = \left\| \frac{p}{2} Q_H\left(\frac{d}{p} \right) \right\|_1$$

- Performs well for certain applications (Kegelmeyer, Robinson favorites)

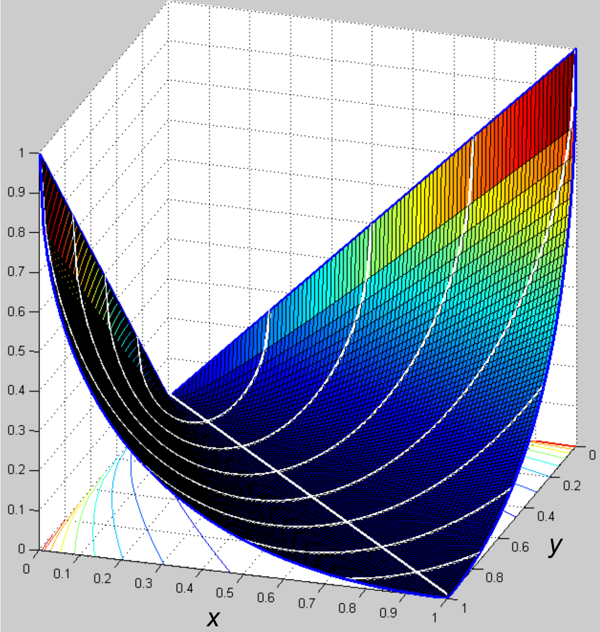
H_s surface, contour lines, and $x+y=c$ lines



H

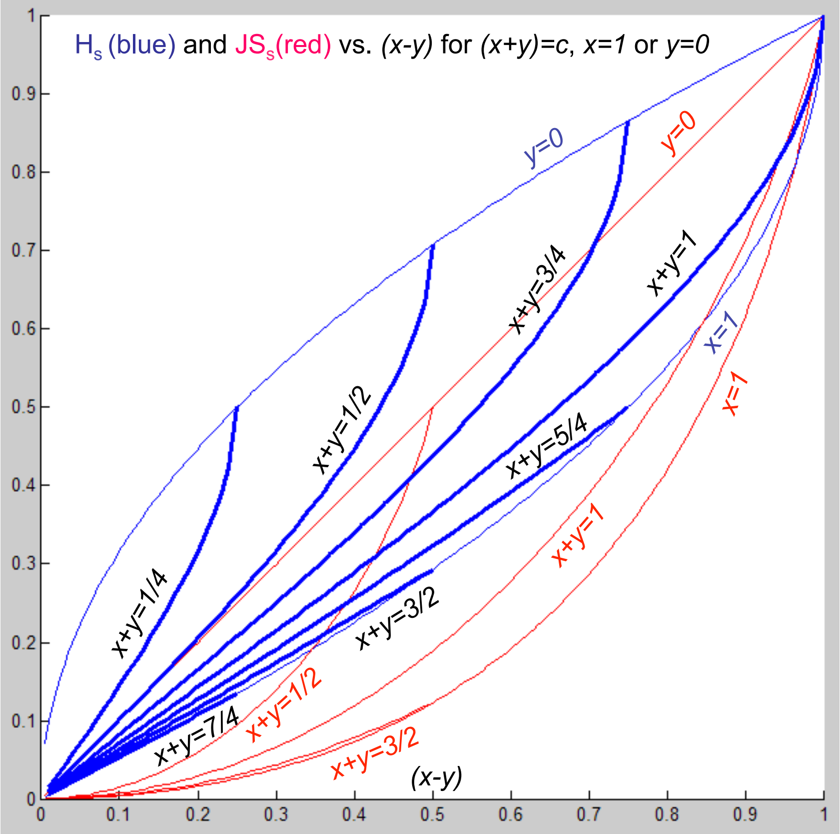
Figure 5

$C_{\sqrt{}}=H_s^2$ surface, contour lines, and $x+y=c$ lines

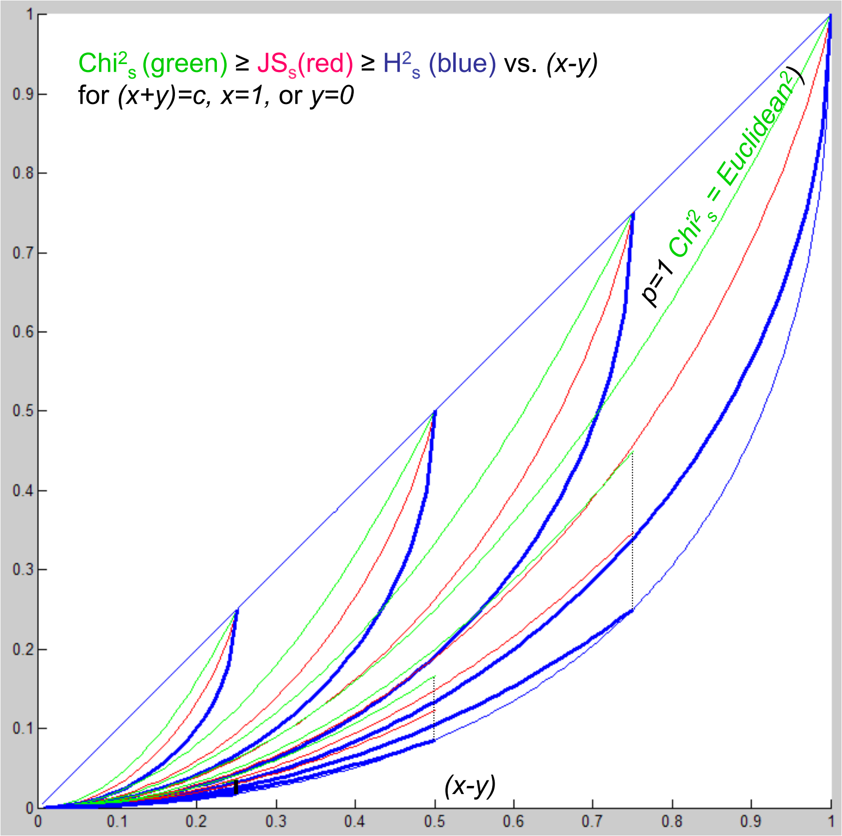


H^2

H and JS



1D Q plots, χ^2, JS, H^2



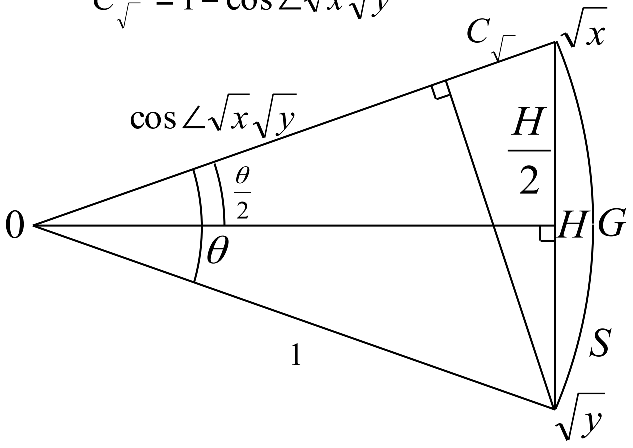
Hellinger and Cosine
Euclidean and Cosine

Add animation fig

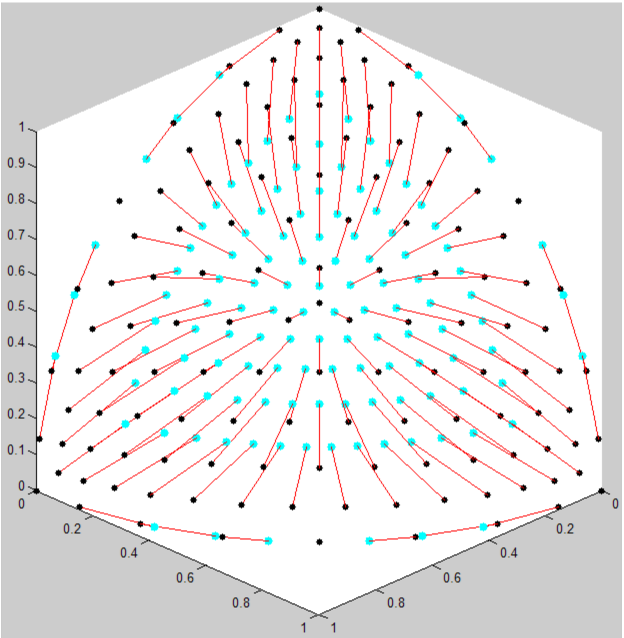
- Hellinger projects to sphere using square-root, then takes Euclidean distance

$$\sum x_k = 1 \Leftrightarrow \sum \sqrt{x_k}^2 = 1 \qquad H = \left\| \sqrt{x} - \sqrt{y} \right\|_2$$

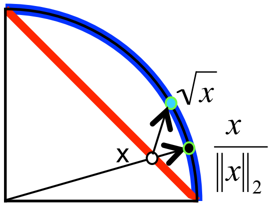
$$C_{\sqrt{}} = 1 - \cos\angle\sqrt{x}\sqrt{y}$$



$$\cos\theta = 1 - 2\sin^2\frac{\theta}{2} \Rightarrow C_{\sqrt{}} = \frac{H^2}{2} = H_s^2$$

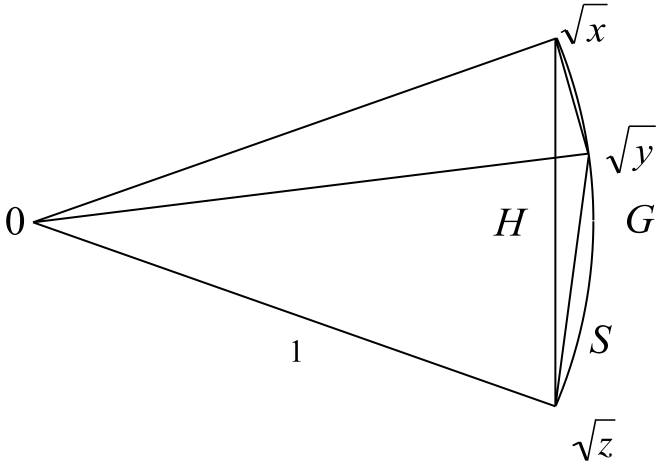


Hellinger and **Euclidean** projections to 3-sphere, and **difference** between them.

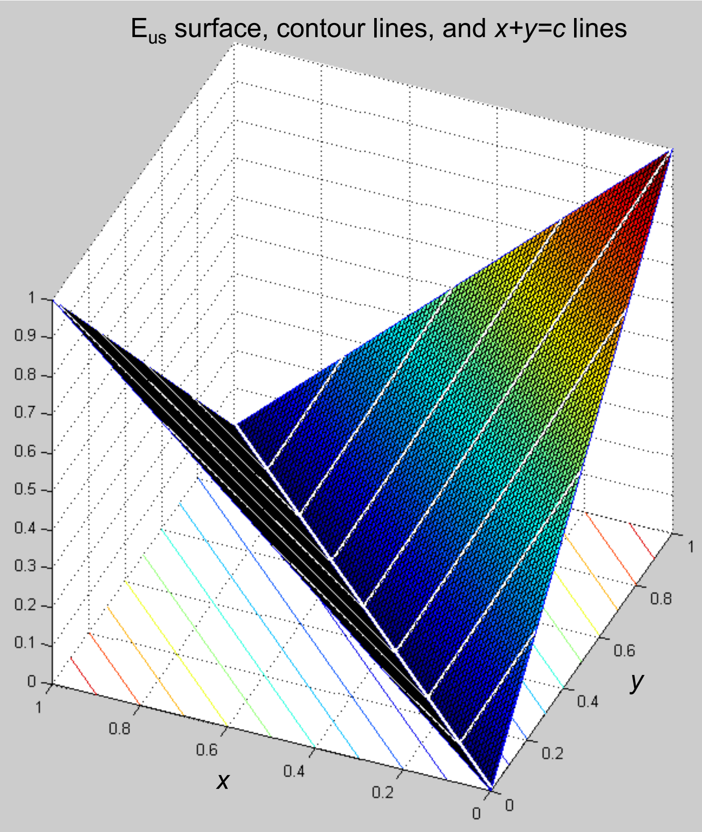


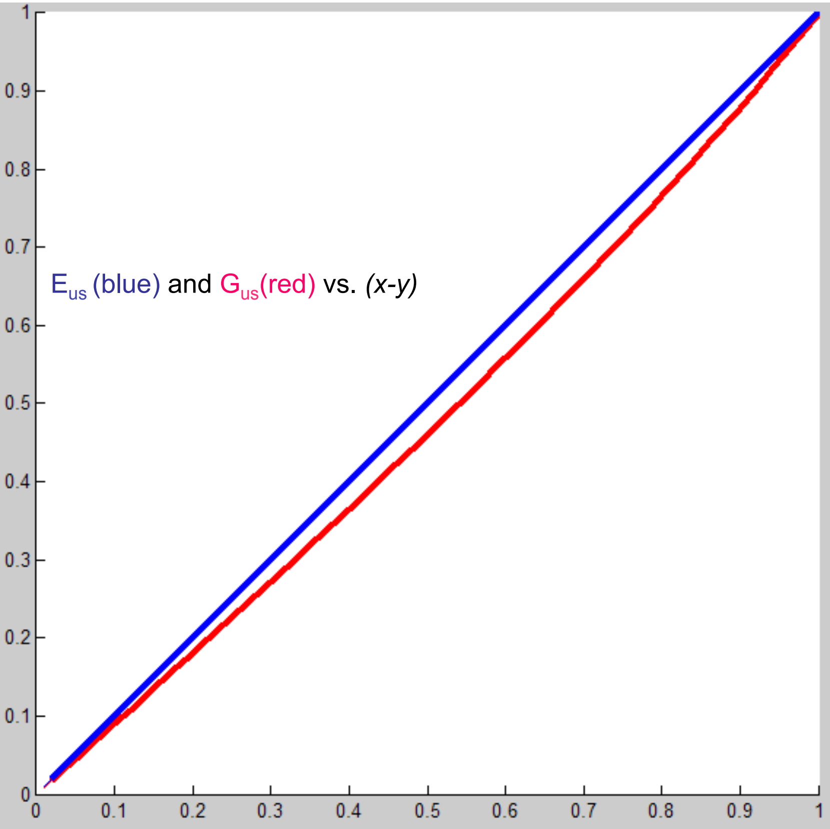
Geodesic Distance?

- Suggest Geodesic distance on sphere G over $\frac{x}{\|x\|_2}$ and $\sqrt{x}$
  - More convex than H (or E), barely satisfies triangle inequality
  - G(x,z)=G(x,y)+G(y,z) for y=λx+(1-λ)z
    strict inequality for other y
- C, E, G are global order preserving, over both $\frac{x}{\|x\|_2}$ and $\sqrt{x}$



Euclidean and Geodesic (norm)

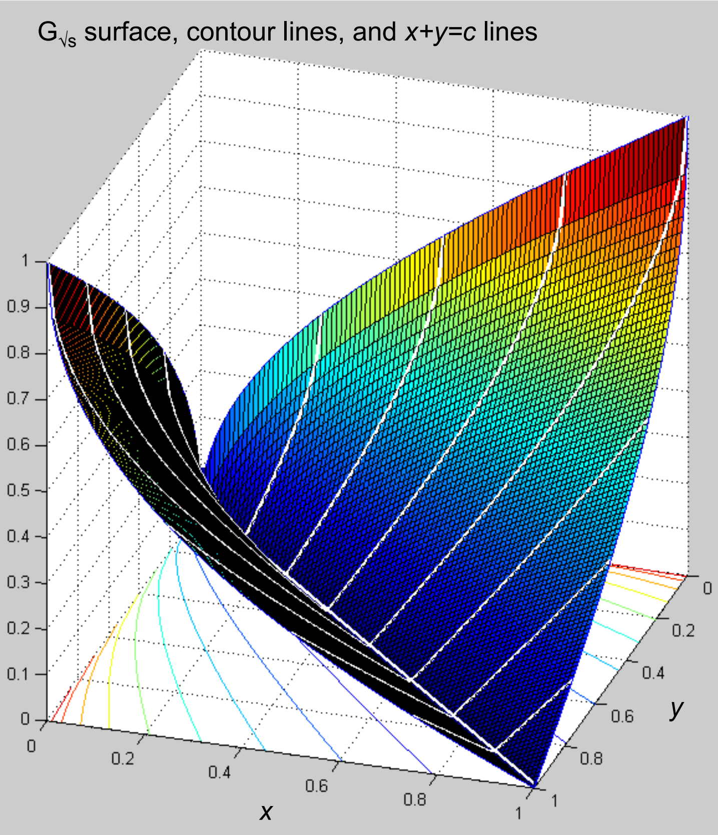


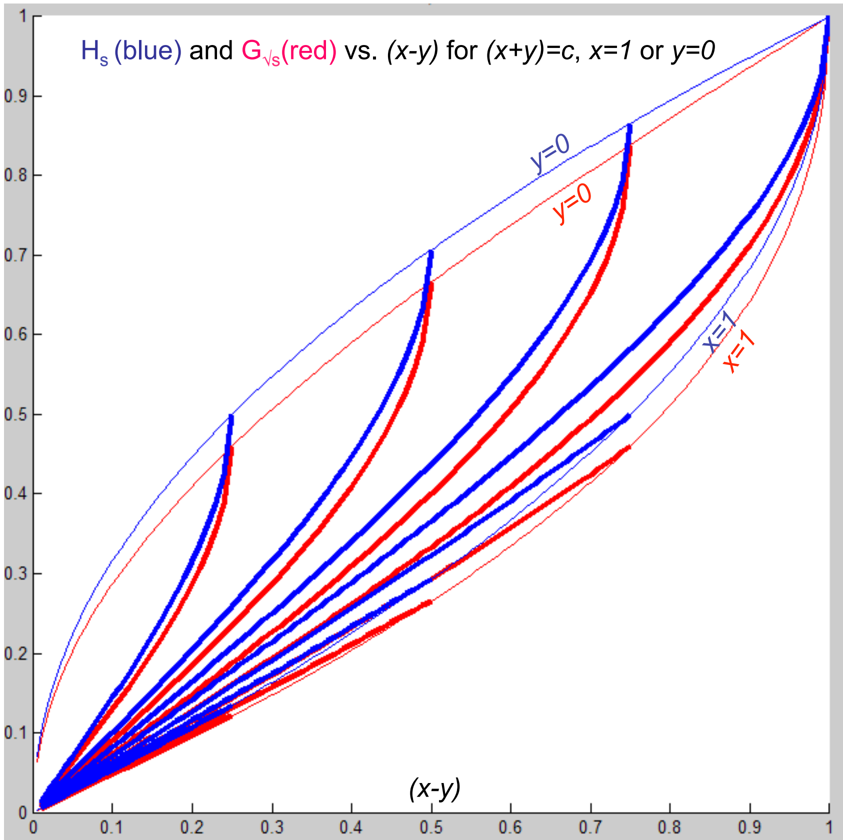


(x-y)

Hellinger and Geodesic (root)

left is visually indistinguishable from Hs, so skip it



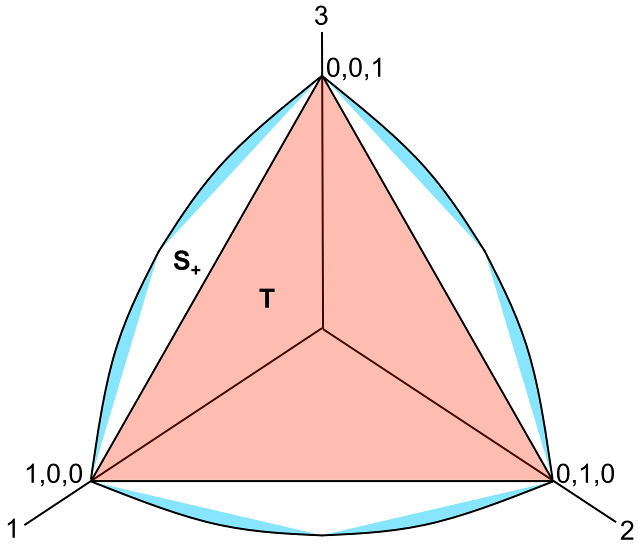


Outline

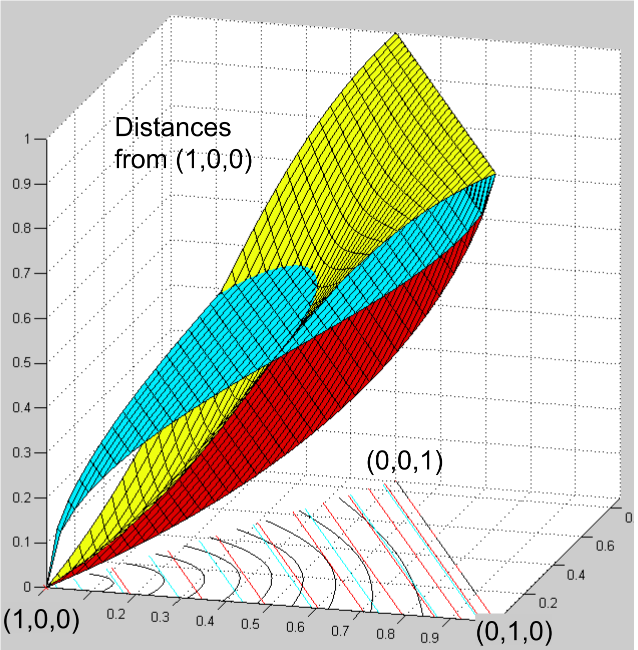
- Application context
- Distances
- 3d plots
- Algebraic reformulation as Z, Q
- Ratios
- Worst-case difference construction

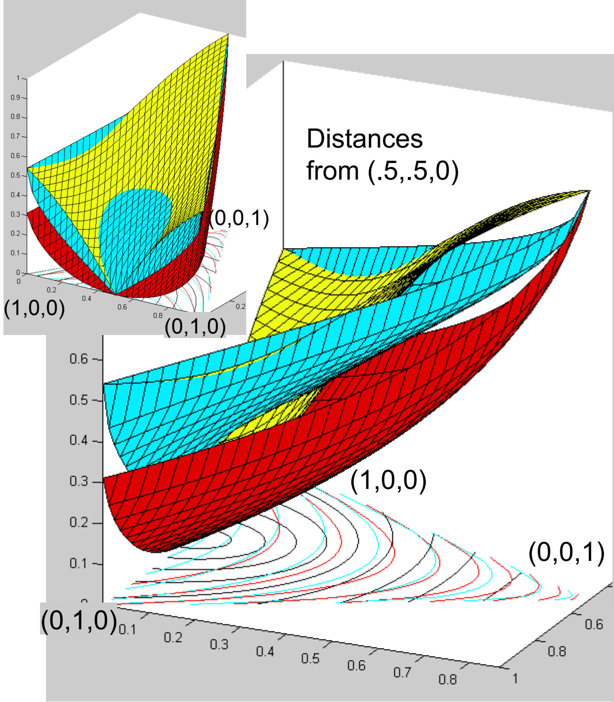
3d plots for selected x points

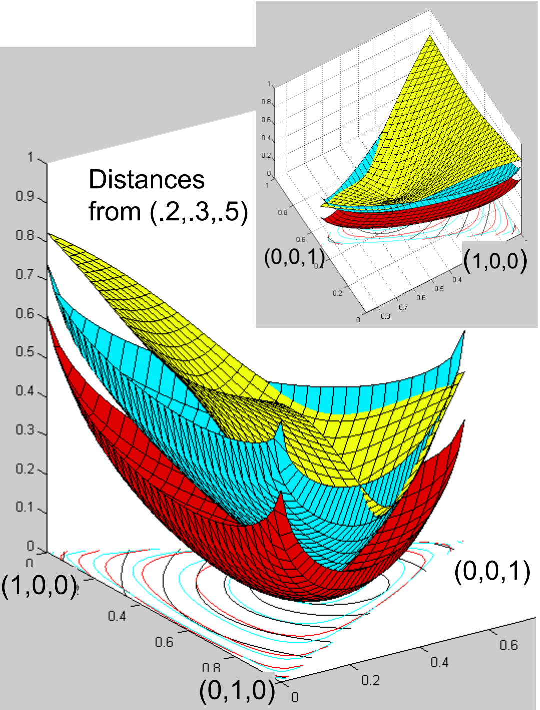
- x=[1,0,0]
  - fig 6: H^2, JS, Chi^2
  - fig 7: H, H^2, E shortcomings
    - No ortho-max for E
  - fig 8: H, H^2, E_u
- X=[1,1,0]/2
  - fig 9: H^2, JS, Chi^2
  - fig 10: H, H^2, E_u
- X=[1,1,1]/3
  - fig 11: H^2, JS, Chi^2
  - fig 12: H, H^2, E_u
- X=[0.2 0.3 0.5]
  - fig 13: H^2, JS, Chi^2
  - fig 14: H, H^2, E_u
- X=[0.89, 0.1 0.01] - sharp upturns
  - fig 15: H^2, JS, Chi^2
  - fig 16: H, H^2, E_u

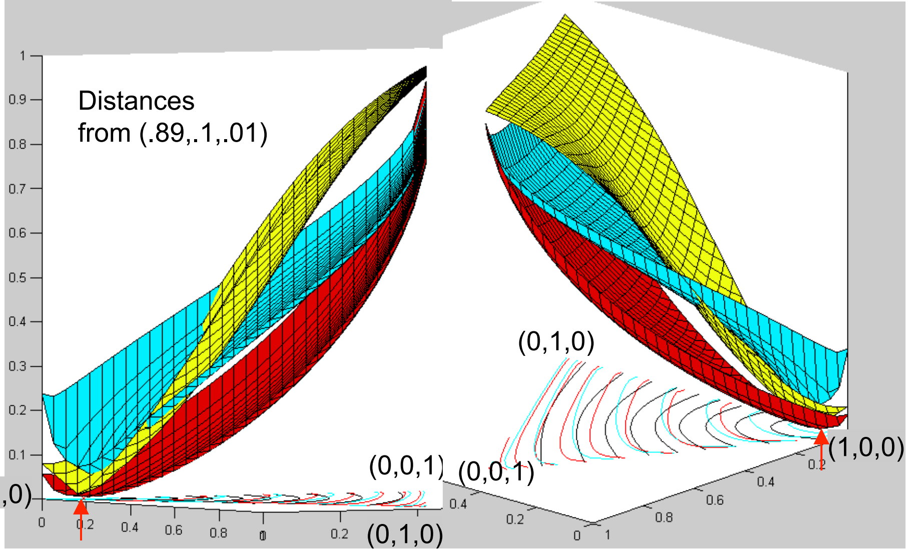


3-dimensional mixture model distances using **Euclidean$_u$(yellow + black contours)**, **Hellinger$_s$(blue)** and **JS$_s$(red)**









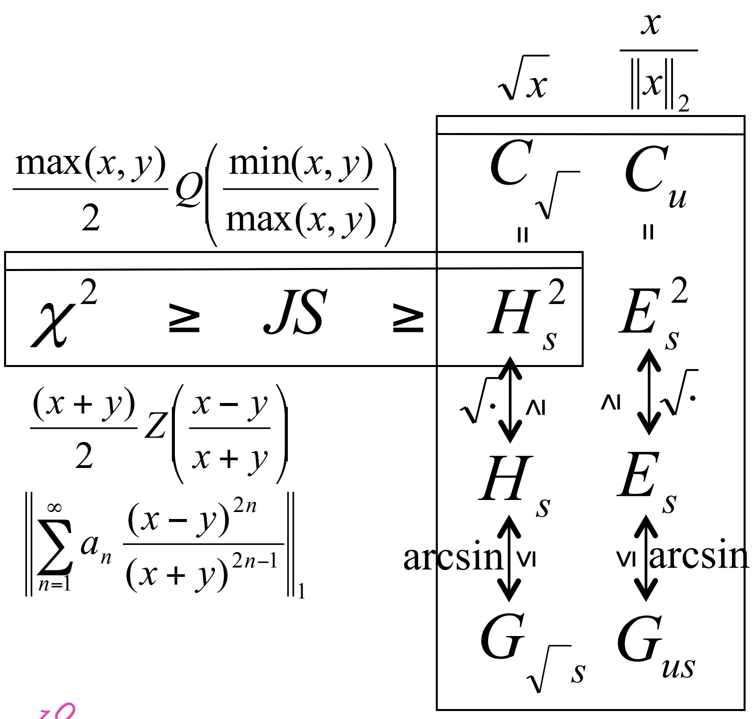
Static backup

Outline

- Application context
- Distances
- 3d plots
- Algebraic reformulation as Z, Q, W
- Ratios
- Worst-case difference construction

Algebraic Forms, Q, Z for X, JS, H2

- Q fn
- Q series
- Z fn



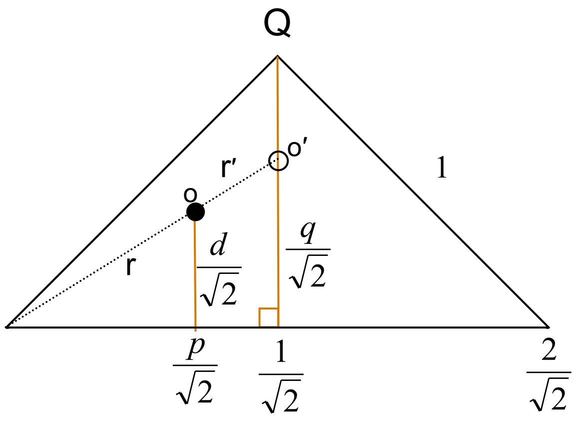
$$\frac{\max(x,y)}{2}\,Q\left(\frac{\min(x,y)}{\max(x,y)}\right)$$

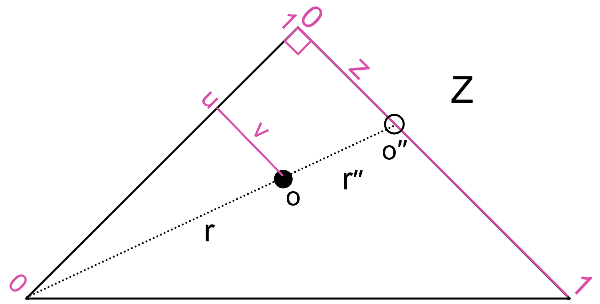
$$\boxed{\chi^2 \quad \geq \quad JS \quad \geq \quad H_s^2}$$

$$\frac{(x+y)}{2}\,Z\left(\frac{x-y}{x+y}\right)$$

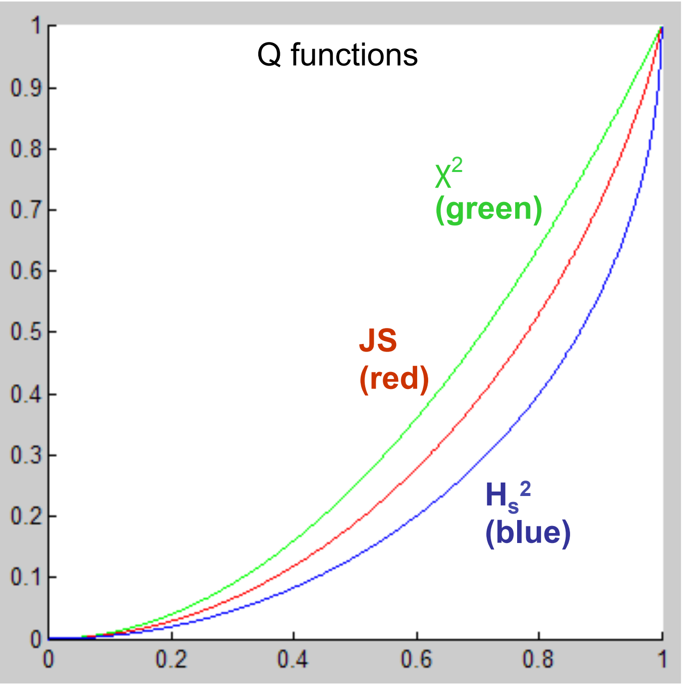
$$\left\|\sum_{n=1}^{\infty} a_n \frac{(x-y)^{2n}}{(x+y)^{2n-1}}\right\|_1$$

$$\begin{tikzcd}
\sqrt{x} & \frac{x}{\|x\|_2} \\
C_{\sqrt{}} \arrow[d, equal] & C_u \arrow[d, equal] \\
H_s^2 \arrow[d, leftrightarrow, "\sqrt{\cdot}"', "\geq"] & E_s^2 \arrow[d, leftrightarrow, "\geq"', "\sqrt{\cdot}"] \\
H_s \arrow[d, leftrightarrow, "\arcsin"', "\leq"] & E_s \arrow[d, leftrightarrow, "\leq"', "\arcsin"] \\
G_{\sqrt{}s} & G_{us}
\end{tikzcd}$$





Q fn



- Recall

$$\chi^2 = \left\| \frac{1}{2} \frac{(x-y)^2}{(x+y)} \right\|_1$$

$$JS = \left\| x \log_2\left(\frac{2x}{x+y}\right) + y \log_2\left(\frac{2y}{x+y}\right) \right\|_1$$

$$H^2 = \left\| \left(\sqrt{x} - \sqrt{y}\right)^2 \right\|_1$$

- Q fn

$$D_f(x,y) = \left\| \frac{p}{2} Q_f\left(\frac{d}{p}\right) \right\|_1$$

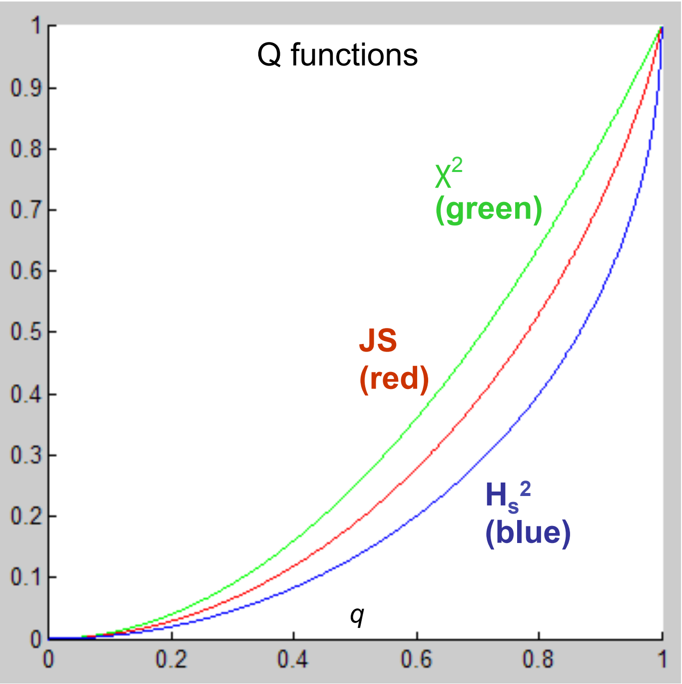
  - p=x+y   d=|x-y| q=d/p

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

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

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

Q fn series



- $\chi^2 = \frac{1}{2} p q^2$
-

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

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

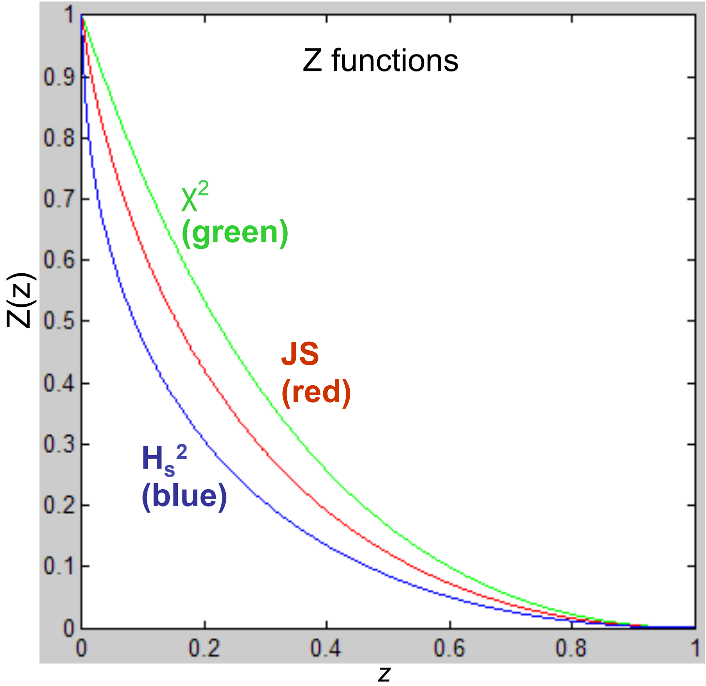
$$\approx 0.361pq^2 + 0.060pq^4 + 0.024pq^6 + 0.013pq^8 + 0.008pq^{10} + \cdots$$

-

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

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

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



Z fn

- Recall

$$\chi^2 = \left\| \frac{1}{2}\frac{(x-y)^2}{(x+y)} \right\|_1$$

$$JS = \left\| x\log_2\left(\frac{2x}{x+y}\right) + y\log_2\left(\frac{2y}{x+y}\right) \right\|_1$$

$$H^2 = \left\| \left(\sqrt{x} - \sqrt{y}\right)^2 \right\|_1$$

- Z fn

$$D_f(x,y) = \left\| \frac{u}{2} Z_f\left(\frac{v}{u}\right) \right\|_1$$

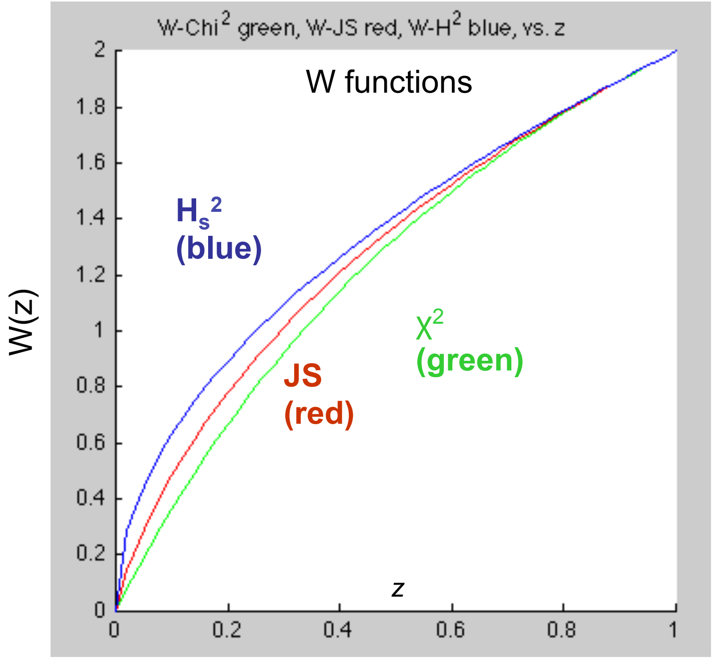
u=max(x,y)   v=min(x,y)   z=v/u

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

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

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

W fn



- Z_f =1+z+… for all f

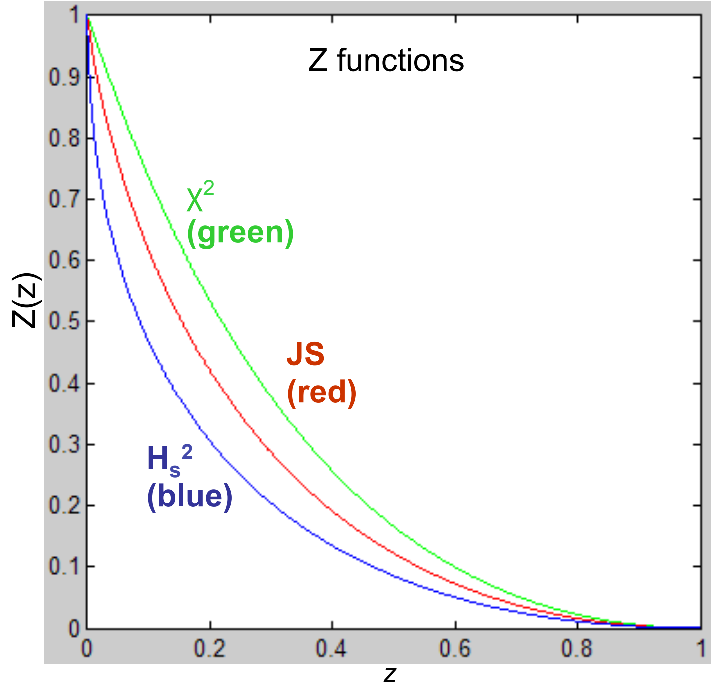
  u(1+z) = u+v and $||u+v||_1$ = 2

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

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

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

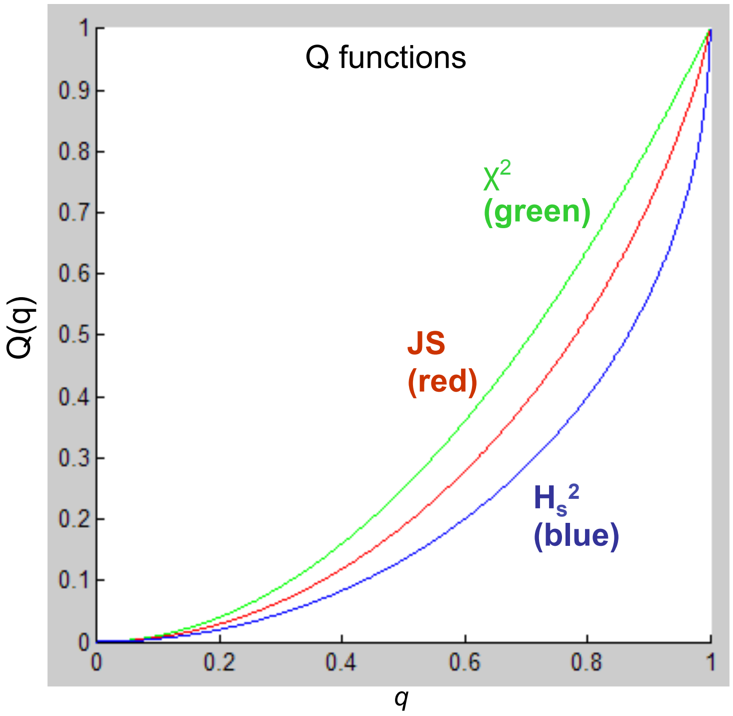
Not componentwise equality

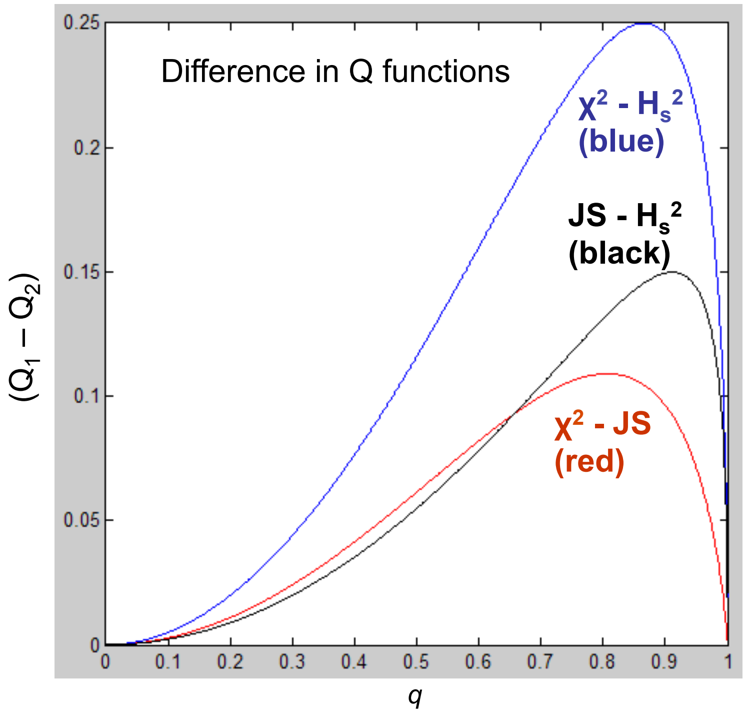


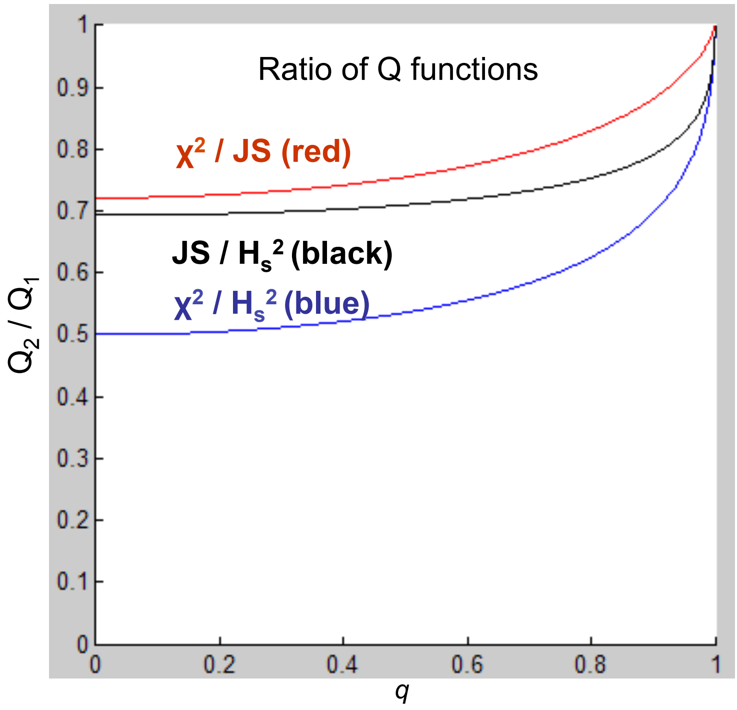
Outline

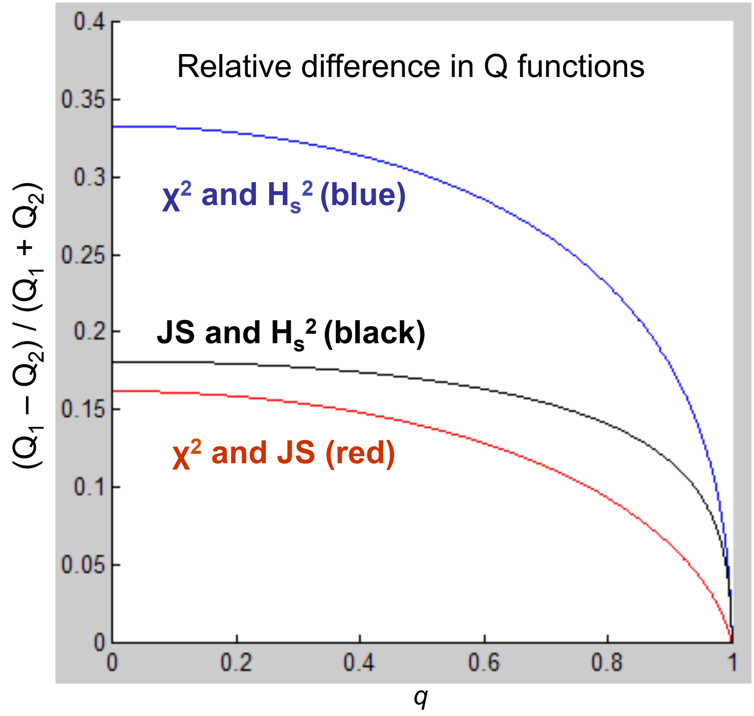
- Application context
- Distances
- 3d plots
- Algebraic reformulation as Z, Q, W
- Ratios
- Worst-case difference construction

Q plots

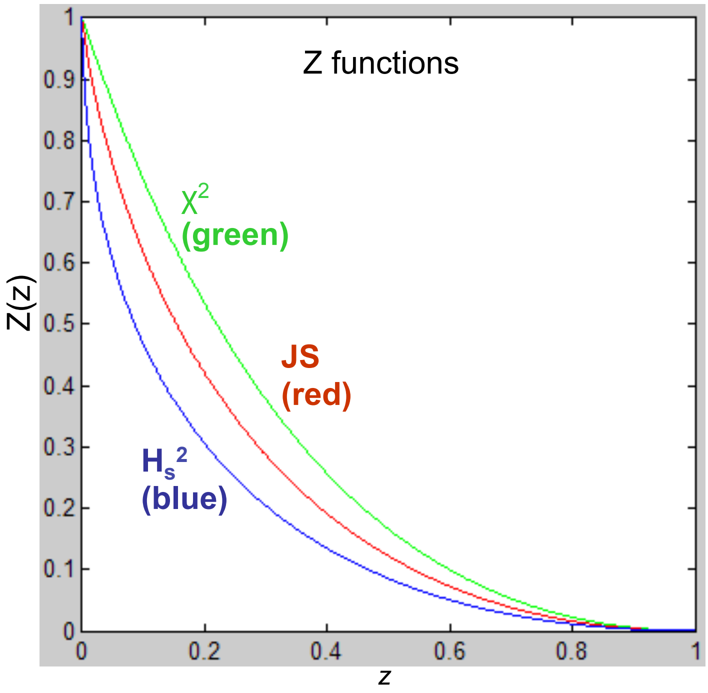

Q functions
χ^2
(green)
JS
(red)
H_s^2
(blue)
Q(q)
q
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1
0
0.2
0.4
0.6
0.8
1

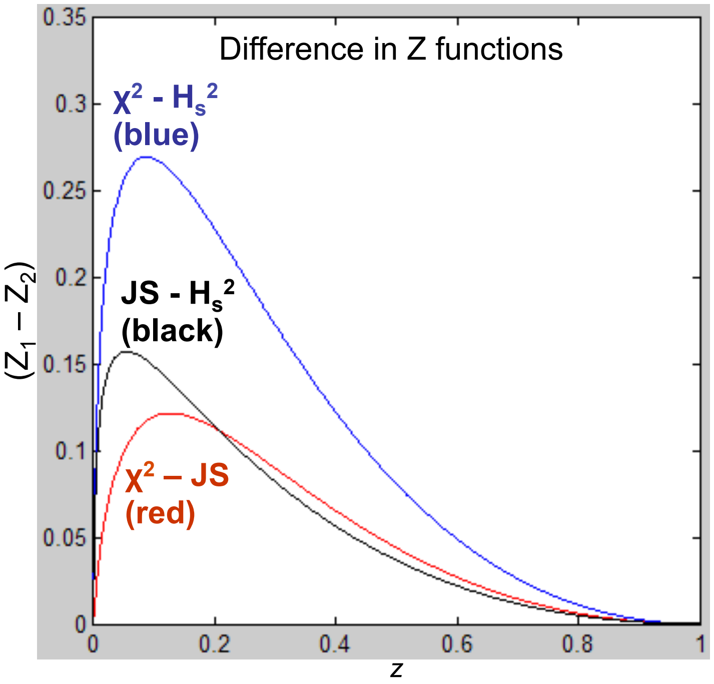

Difference in Q functions
χ^2 - H_s^2
(blue)
JS - H_s^2
(black)
χ^2 - JS
(red)
$(Q_1 - Q_2)$
q
0
0.05
0.1
0.15
0.2
0.25
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1

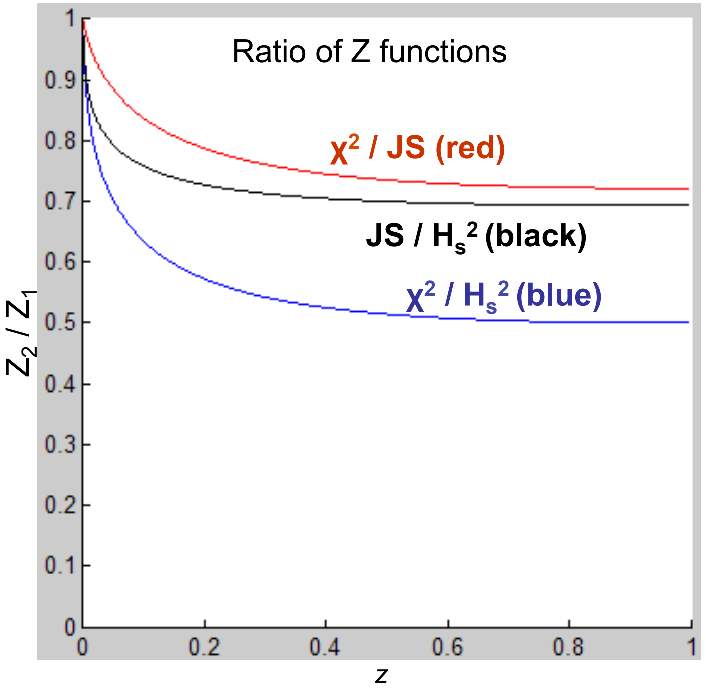

Ratio of Q functions
χ^2 / JS (red)
JS / H_s^2 (black)
χ^2 / H_s^2 (blue)
Q_2 / Q_1
q
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1
0
0.2
0.4
0.6
0.8
1

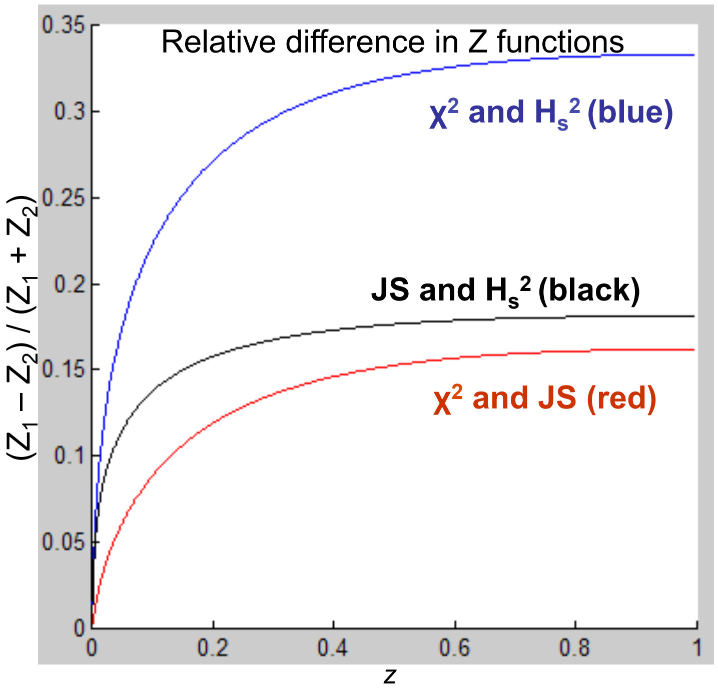

Relative difference in Q functions
χ^2 and H_s^2 (blue)
JS and H_s^2 (black)
χ^2 and JS (red)
$(Q_1 - Q_2) / (Q_1 + Q_2)$
q
0
0.05
0.1
0.15
0.2
0.25
0.3
0.35
0.4
0
0.2
0.4
0.6
0.8
1

Z plots









related by
q = (1-z)/(1+z)
p=(u/2)(1+z)

Selected theorems

- Z-increasing <≈> Q-decreasing
- Z, Q monotonic
- Z_1/Z_2, Q_1/Q_2 ratios monotonic
- Z, Q differences 1 unique max, 1 inflection pt

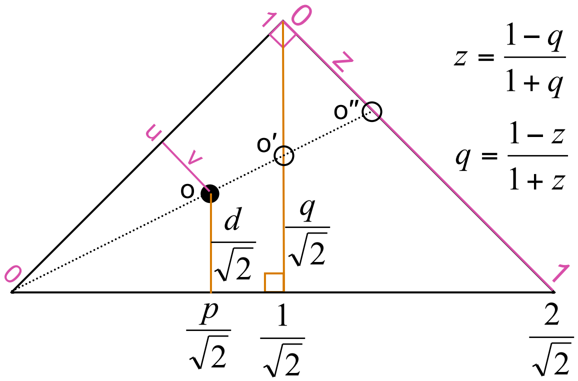
Z-Q Equivalence

Theorem 8 (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 9. *Z decreasing $\Rightarrow$ Q increasing; also Z decreasing $\Leftarrow$* $Q'(q) > \frac{1}{1+q}Q(q) \geq 0$.



- Z_f actually are decreasing (lots of algebra, derivates, L'Hopital's rule)

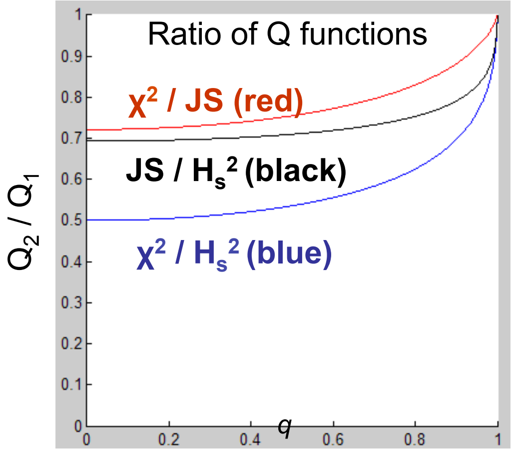
Z-Q Ratios

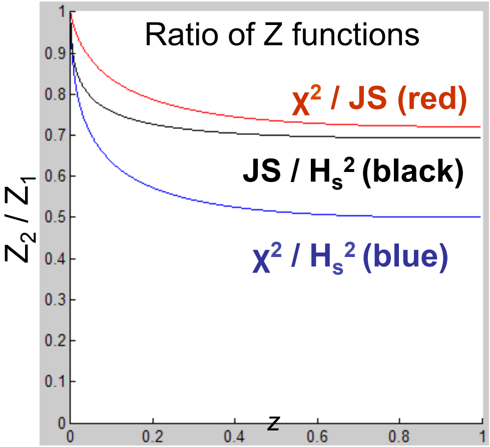
Thm: $\frac{Z_1}{Z_2}$ decreasing $\Leftrightarrow$ $\frac{Q_1}{Q_2}$ increasing $\max \frac{Z_1}{Z_2} = \max \frac{Q_1}{Q_2}$ $\min \frac{Z_1}{Z_2} = \min \frac{Q_1}{Q_2}$

Proof from componentwise

$$\frac{D_1}{D_2} = \frac{Z_1}{Z_2}(z) = \frac{Q_1}{Q_2}(q) \qquad q = \frac{1-z}{1+z} \text{ and } z = \frac{1-q}{1+q}$$

	Q^*		q^*	Z^*		z^*
$R_{\chi/H}$	$1/2 =$	$.500$	0	$1/2 =$	$.500$	1
$R_{\chi/JS}$	$1/2 \log 2 >$	$.721$	0	$1/2 \log 2 >$	$.721$	1
$R_{JS/H}$	$\log 2 >$	$.693$	0	$\log 2 >$	$.693$	1
	max is 1 at $q = 1$ and $z = 0$					





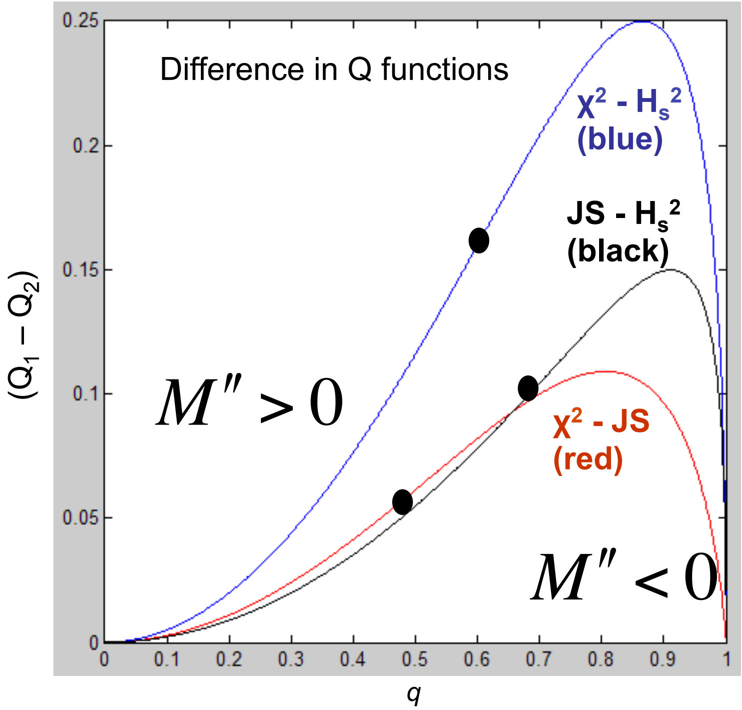
Proof from leading term of
series, or
directly from functional forms

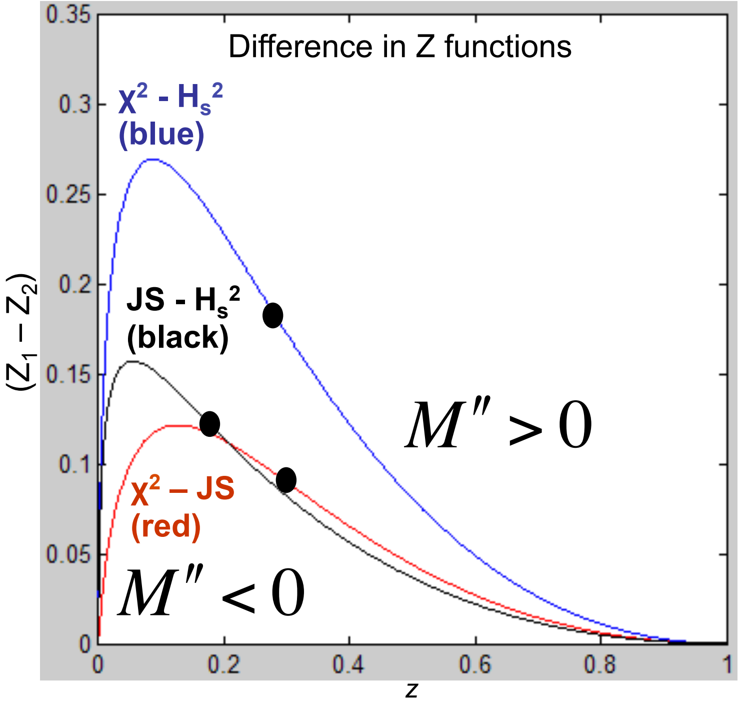
Key feature is large flat section.
This appears new.

Z-Q Differences

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

No closed form except upper left



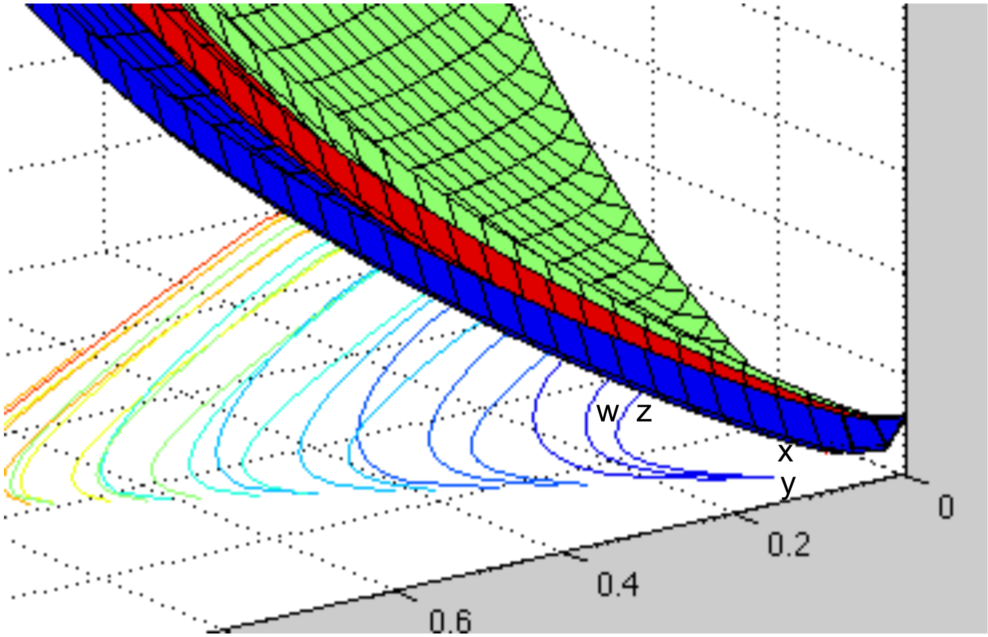


Outline

- Application context
- Distances
- 3d plots
- Algebraic reformulation as Z, Q, W
- Ratios
- Worst-case difference construction

Contours

- The contours looked similar?
- Not local-order preserving
  - Exploits different ratios for small z vs. moderate z
  - x=[0.89, 0.1, 0.01]
  - y=[0.9, 0, 0.01]
  - z=[0.65, 0.35, 0]
  - w=[0.6, 0.4, 0]



$$\chi^2(x,y) < \chi^2(x,z) \text{ but } JS(x,y) > JS(x,z) \text{ and } H_s^2(x,y) > H_s^2(x,z)$$

$$JS(x,y) < JS(x,w) \text{ but } H_s^2(x,y) > H_s^2(x,w)$$

Worst-Case Construction

- Find points x,y,z, such that $D_1(x,y)=D_1(x,z)$ but D_2 gives the most different answer

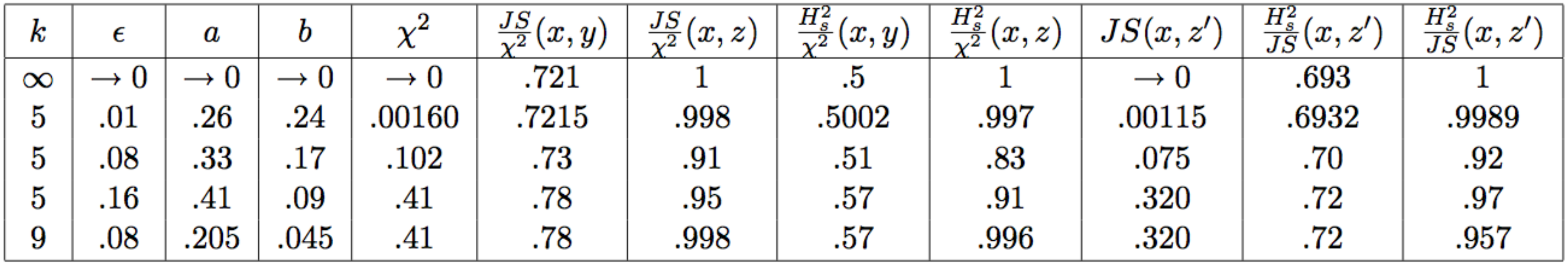
$$x = (a,a,\ldots,a,b,b\ldots,b,0)$$
$$y = \left(b,b,\ldots,b,a,a,\ldots,a,0\right)$$
$$z = \left(x_1,x_2,\ldots x_j,c,0,0,\ldots 0,d\right)$$
$$\text{where a} = \frac{1}{\text{k - 1}} + \varepsilon, \text{b} = \frac{1}{\text{k - 1}} - \varepsilon,$$
$$d,c,j : D_1(x,y) = D_1(x,z)$$

$$k \to \infty, \varepsilon \to 0 \text{ gives } \frac{D_2}{D_1}(x,y) \to \min, \frac{D_2}{D_1}(x,z) \to 1$$

- Relies on
  - Large K dimension
  - Zero component to get small z ratios
  - Moderately similar components to get moderate z ratios
    - Implies Distance is small

Near worst-case

- Doesn’t have to be very extreme



k	ϵ	a	b	χ^2	$\frac{JS}{\chi^2}(x,y)$	$\frac{JS}{\chi^2}(x,z)$	$\frac{H_s^2}{\chi^2}(x,y)$	$\frac{H_s^2}{\chi^2}(x,z)$	$JS(x,z')$	$\frac{H_s^2}{JS}(x,z')$	$\frac{H_s^2}{JS}(x,z')$
∞	$\to 0$	$\to 0$	$\to 0$	$\to 0$	.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	.996	.320	.72	.957

- Still relies on a zero component, but small dim k, large epsilon

$$x = (a,a,\ldots,a,b,b\ldots,b,0)$$

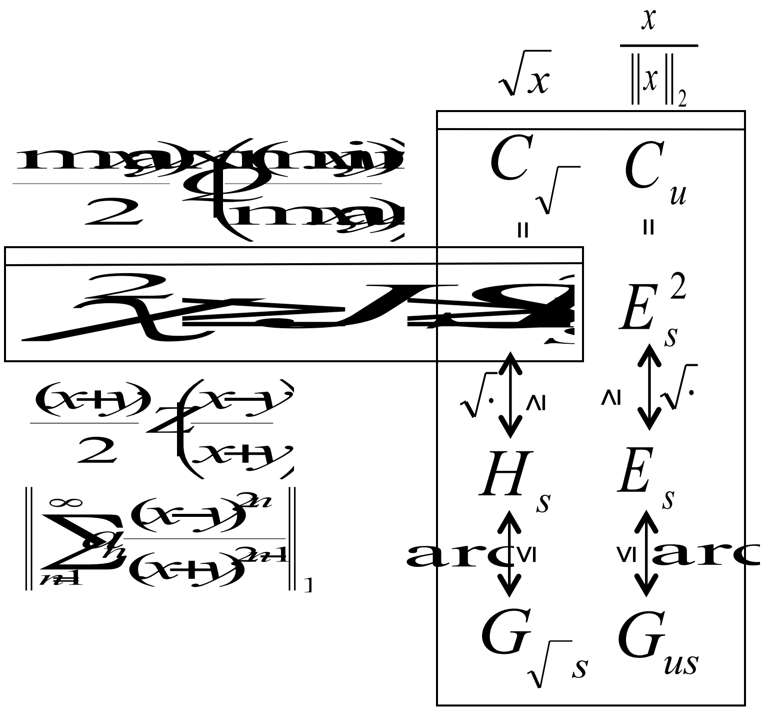
$$y = \left(b,b,\ldots,b,a,a,\ldots,a,0\right)$$

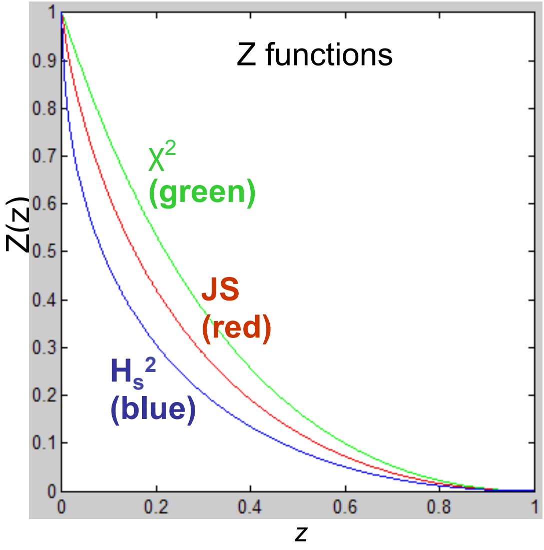
$$z = \left(x_1,x_2,\ldots x_j,c,0,0,\ldots 0,d\right)$$

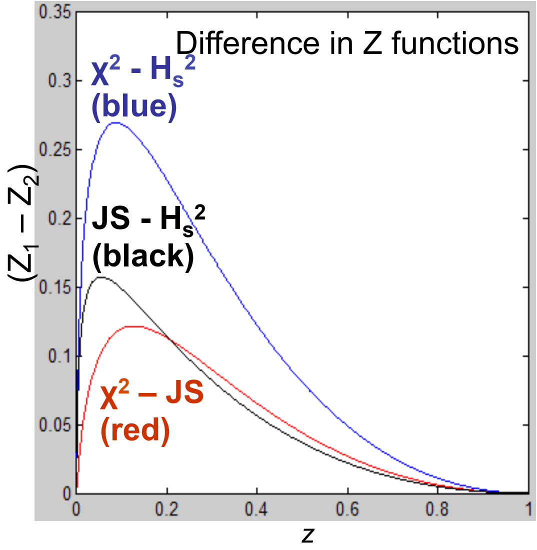
$$\text{where a} = \frac{1}{\text{k - 1}} + \varepsilon, \text{b} = \frac{1}{\text{k - 1}} - \varepsilon,$$

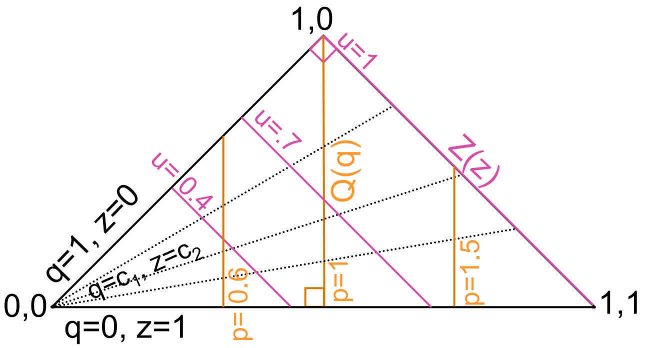
$$d,c,j : D_1(x,y) = D_1(x,z)$$

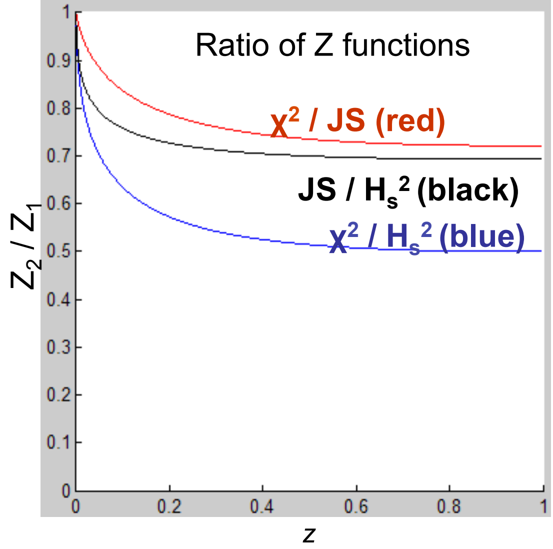
Main Observations





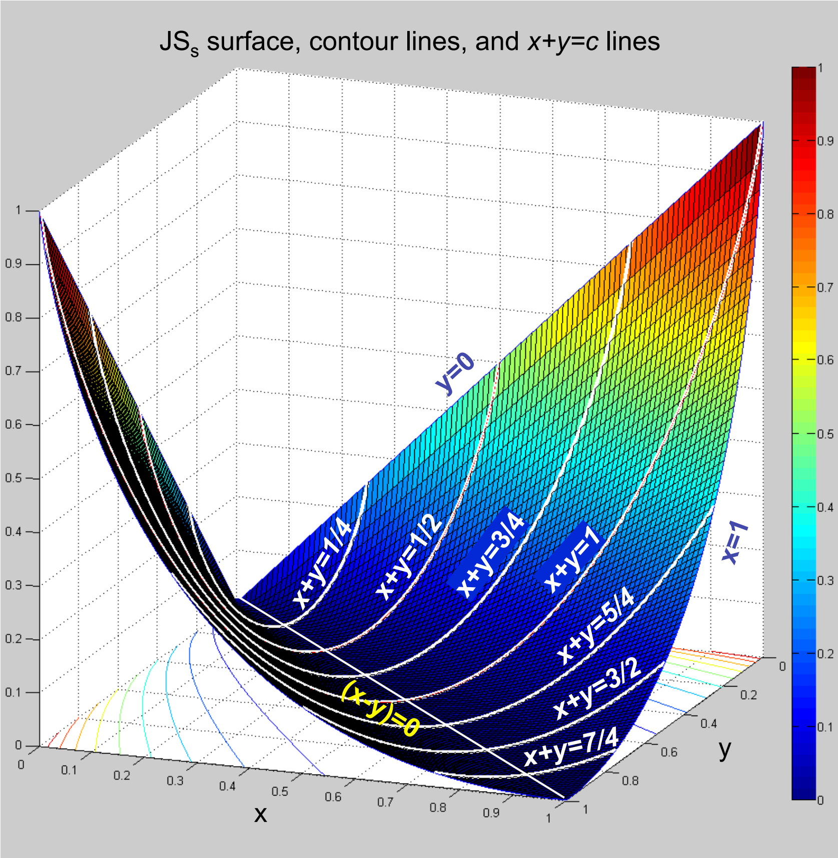




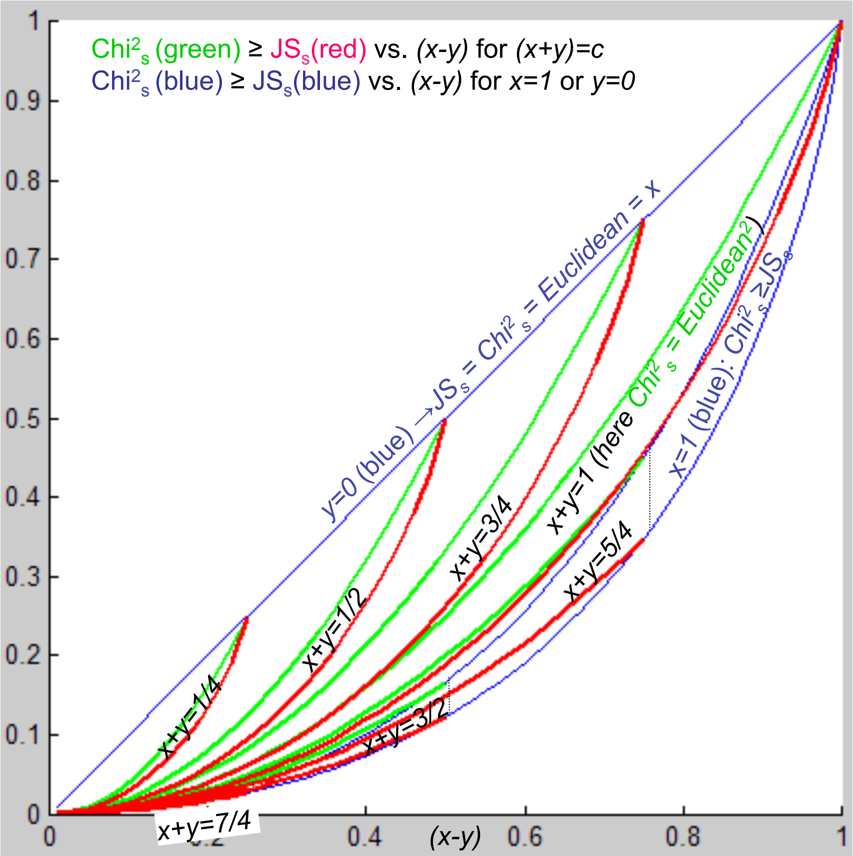


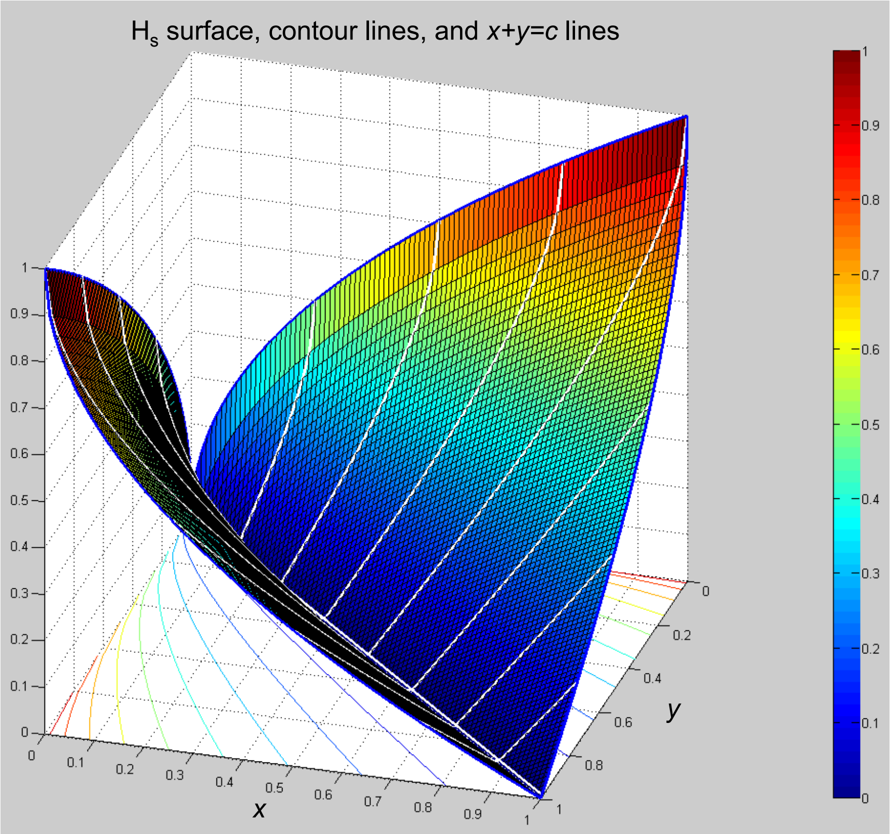
Conclusion

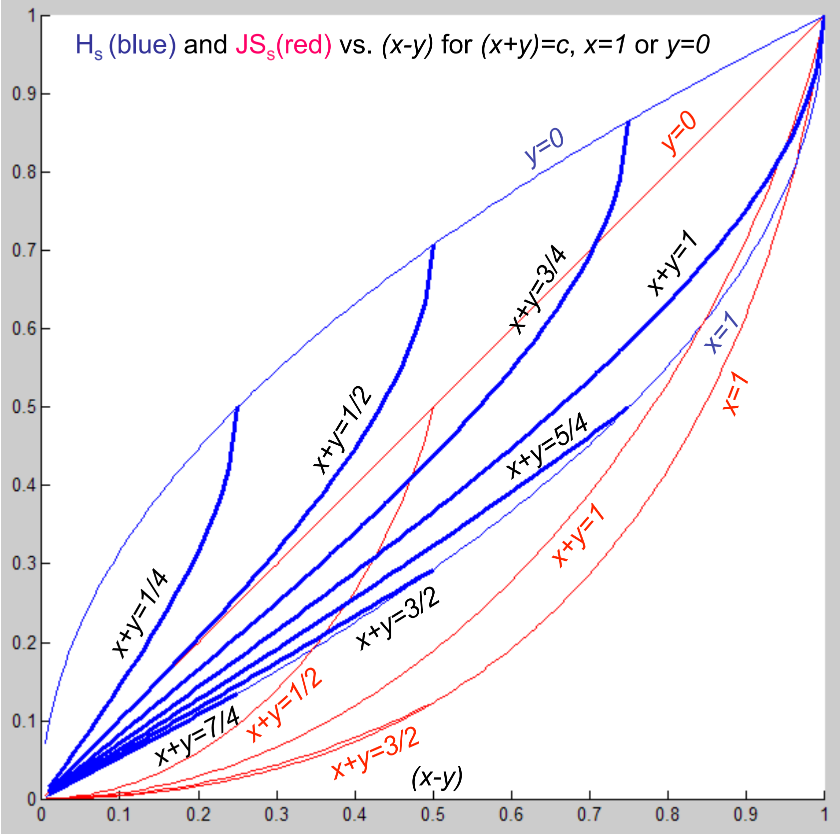
- Insight into what distances do differently
- What's new / novel?
  - Geometric analysis, pictures!
  - Ratio monotonicity, difference analysis
  - Algebraic limits probably known, but references hard to chase
- Future
  - Clustering case study on real data
  - Hellinger square-root projection study
  - Other families of distances
    - Compound distances, Earth Mover's
    - Partition/cluster metrics
  - Affect of norms other than 1-norm (triangle inequality)?
  - SNL needs program in understanding effects of text-analysis pipeline knobs.

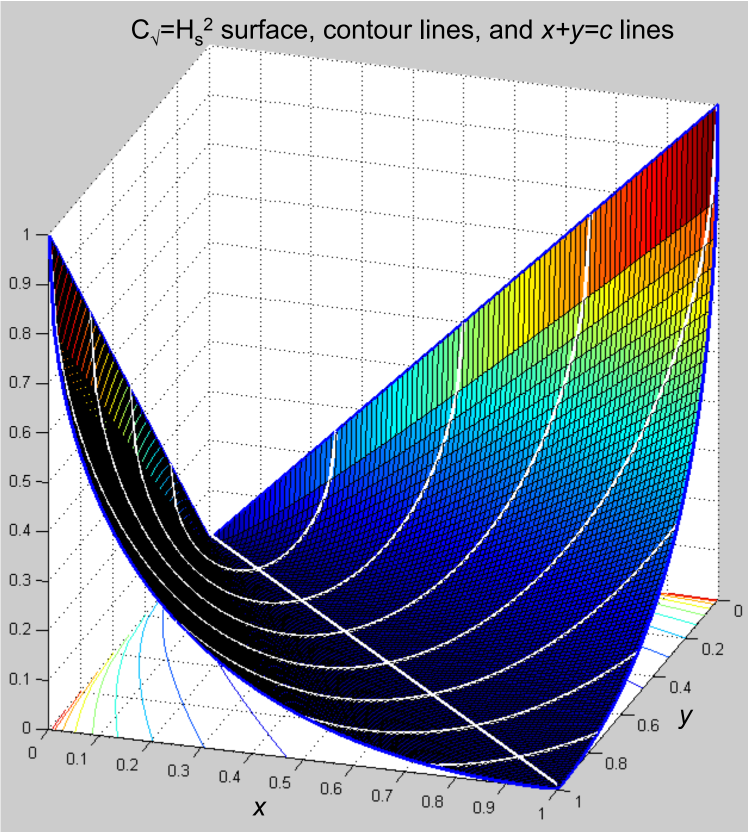


JS$_s$ surface, contour lines, and $x+y=c$ lines

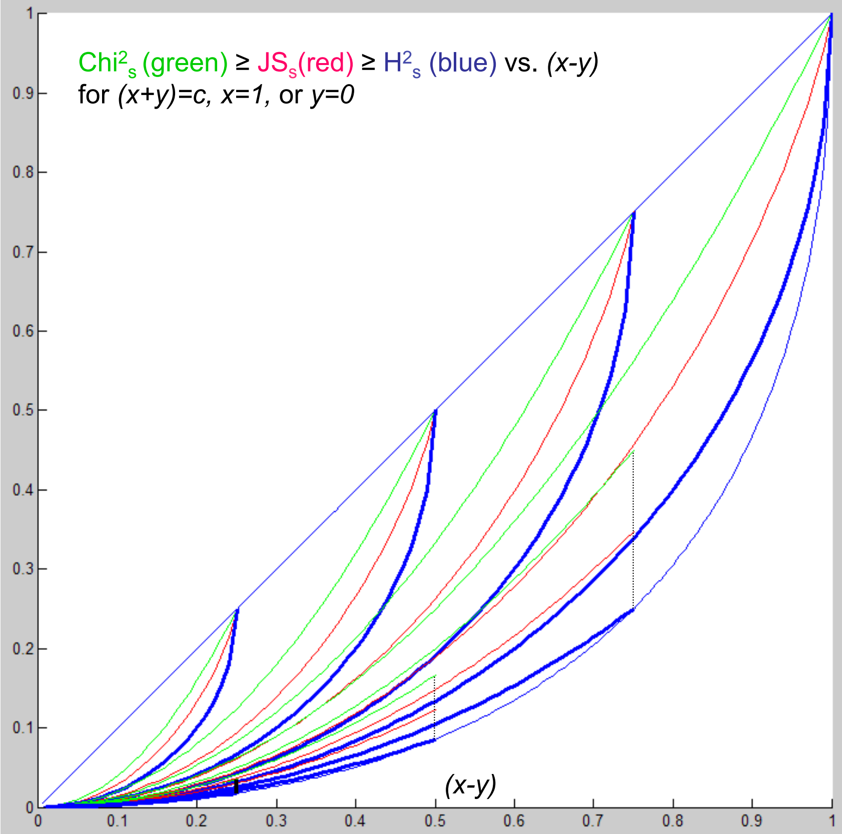



H$_s$ surface, contour lines, and $x+y=c$ lines
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1
x
y
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1


H$_s$ (blue) and JS$_s$(red) vs. $(x-y)$ for $(x+y)=c$, $x=1$ or $y=0$
y=0
y=0
x+y=1/4
x+y=1/2
x+y=3/4
x+y=1
x+y=5/4
x+y=3/2
x+y=7/4
x=1
x=1
x+y=1
x+y=1/2
x+y=3/2
(x-y)
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1

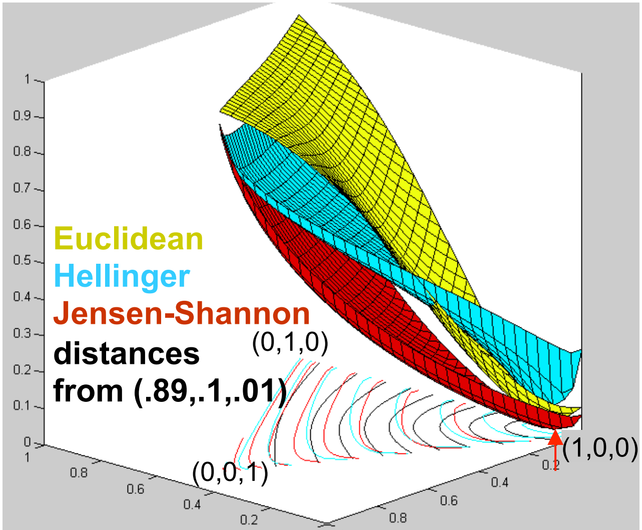


$C_{\sqrt{}}=H_s^2$ surface, contour lines, and *x+y=c* lines

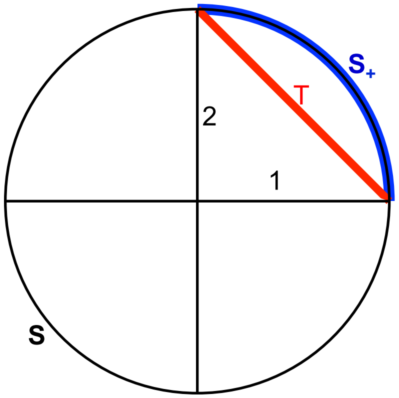


Chi^2_s (green) ≥ JS$_s$(red) ≥ H^2_s (blue) vs. *(x-y)*
for *(x+y)=c, x=1,* or *y=0*

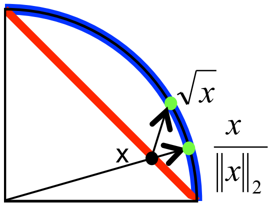
for late-start summary quad chart



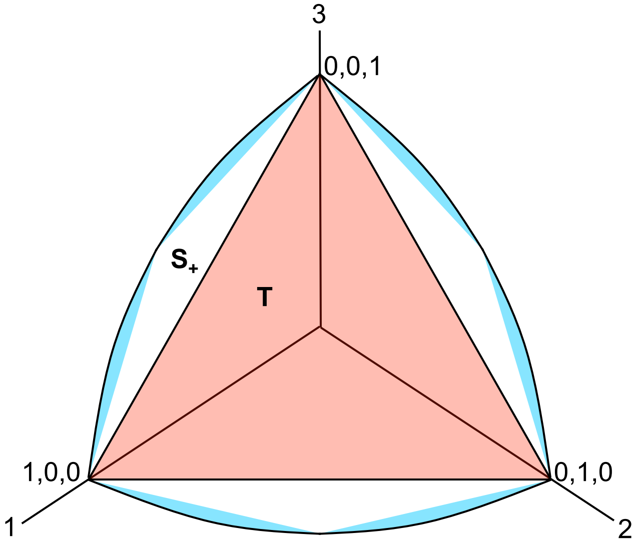
Sharp upturn on lower right shows **H & JS** emphasize high ratios of point coordinates.

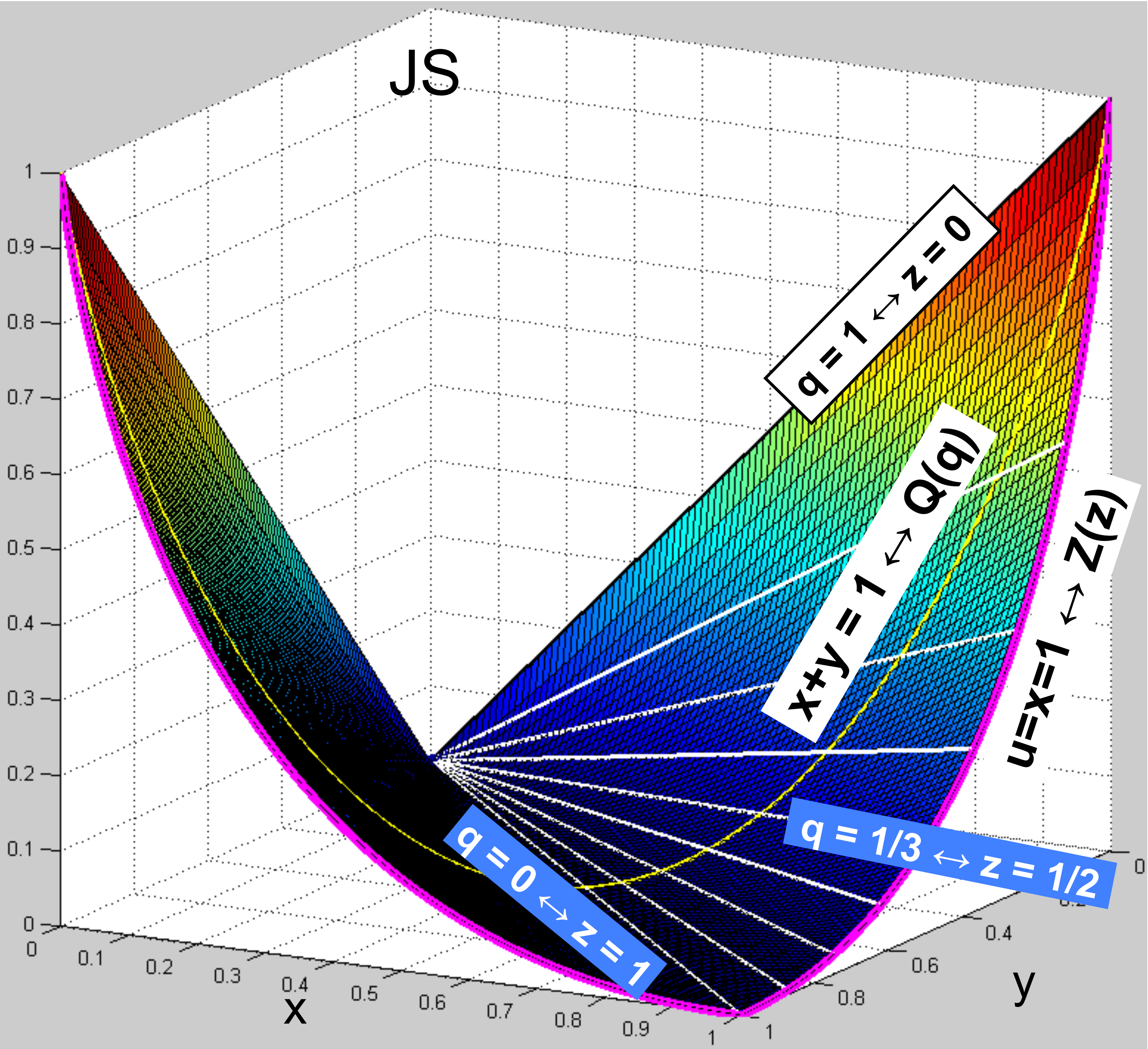


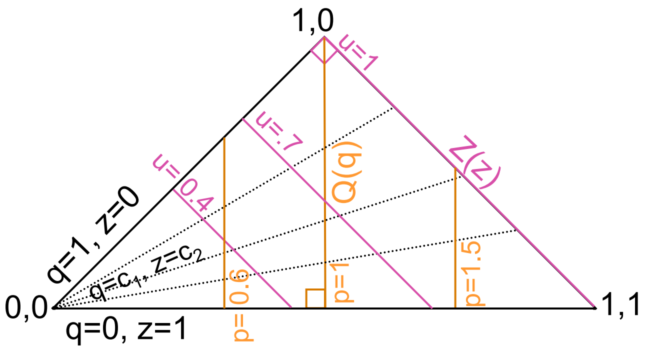
Two-dimensional mixture models T, unit sphere S, and positive part of unit sphere S_+. Coordinate axis 1 and 2.

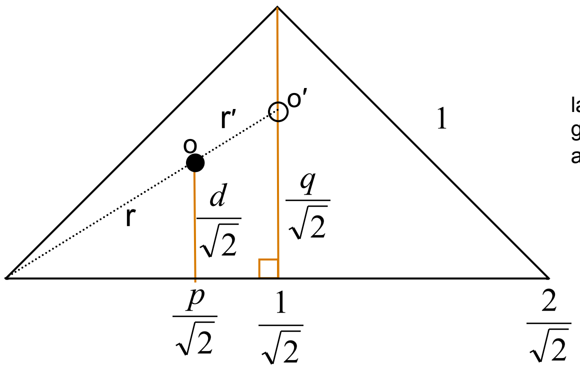


Point x on T projected to S_+ under normalization and square-root.

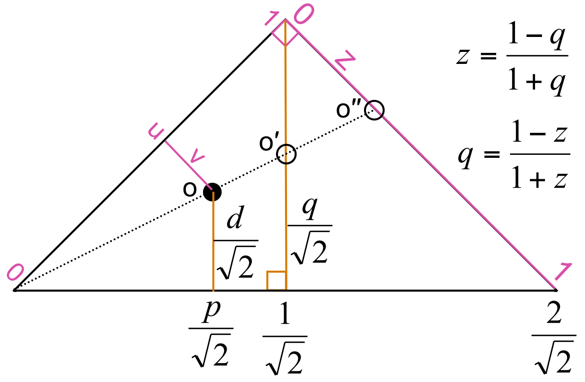

3
0,0,1
S+
T
1,0,0
0,1,0
1
2


JS
q = 1 ↔ z = 0
x+y = 1 ↔ Q(q)
u=x=1 ↔ Z(z)
q = 0 ↔ z = 1
q = 1/3 ↔ z = 1/2
1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1
x
0
0.4
0.6
0.8
1
y



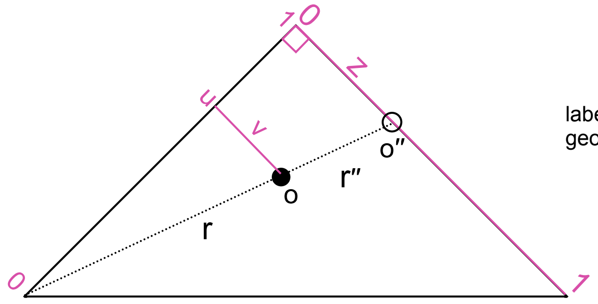


labels are lengths of segments, except points o, o’
geometrically: D(●) = (r / r’) D(o) = (p/1) D(o) = p Q(q)/2
also works for p > 1



$$z = \frac{1-q}{1+q}$$

$$q = \frac{1-z}{1+z}$$



labels are lengths of segments, except points o, o’
geometrically: D(●) = (s / s’) D(o) = (u / 1) D(o) = u/2 Z(z)