Matrix versions of the Hellinger distance

Tanvi Jain

Indian Statistical Institute Delhi

December 18, 2018

Hellinger distance

$p = (p_1, \ldots, p_n)$, $q = (q_1, \ldots, q_n)$: discrete probability distributions, i.e.

$$p_i, q_i \geq 0, \sum_{i=1}^{n} p_i = \sum_{i=1}^{n} q_i = 1.$$

Hellinger distance

$$\begin{aligned} d_H(p, q) &= \frac{1}{\sqrt{2}} \|\sqrt{p} - \sqrt{q}\|_2 \\ &= \left(\sum_{i=1}^{n} \frac{p_i + q_i}{2} - \sum_{i=1}^{n} \sqrt{p_i}\sqrt{q_i} \right)^{1/2}. \end{aligned}$$

$$d_H(p,q) = \left(\mathrm{tr}\mathcal{A}(p,q) - \mathrm{tr}\mathcal{G}(p,q)\right)^{1/2},$$

where $\mathrm{tr}\, p = \sum_{i=1}^{n} p_i$.

Hellinger distance: useful to quantify the similarity between two probability distributions.

Important in

- Probability
- Statistics
- Machine learning ...

Matrix/noncommutative/quantum version

Commutative	Noncommutative
p, q probability vectors	A, B density matrices i.e. $A, B \geq 0$ (positive semidefinite) $\mathrm{tr}A = \mathrm{tr}B = 1$.
$d_H(p, q) = \sqrt{\mathrm{tr}\mathcal{A}(p, q) - \mathrm{tr}\mathcal{G}(p, q)}$	$d_H(A, B) = \sqrt{\mathrm{tr}\mathcal{A}(A, B) - \mathrm{tr}\mathcal{G}(A, B)}$.

In this talk, we will work mainly with positive definite matrices.

$A, B > 0$ (positive definite).

There is only one possible arithmetic mean

$$\mathcal{A}(A, B) = \frac{A + B}{2}.$$

But geometric mean can have different meanings:

- $A^{1/2}B^{1/2}$ or $A^{1/4}B^{1/2}A^{1/4}$
- $(AB)^{1/2}$
- $\left(A^{1/2}BA^{1/2}\right)^{1/2}$
- $A\#B = A^{1/2}\left(A^{-1/2}BA^{-1/2}\right)^{1/2}A^{1/2}$
- $\exp\frac{\log A + \log B}{2}$

In this talk

Different versions of $\mathcal{G}(A, B)$ give different versions of the Hellinger distance on matrices.

We aim to study

1. The different Hellinger distances on positive matrices and their properties, esp. the convexity properties.
2. Barycentres with respect to these distances.

A straightforward generalization

$$
\begin{aligned}
d_1(A,B) &= \frac{1}{\sqrt{2}}\|A^{1/2} - B^{1/2}\|_2 \\
&= \sqrt{\mathrm{tr}\mathcal{A}(A,B) - \mathrm{tr}A^{1/2}B^{1/2}}.
\end{aligned}
$$

d_1 is a metric on $\mathbb{P}$ (set of all positive definite matrices).

$\mathcal{G}(A, B) = (AB)^{1/2}$

$A, B \geq 0,$

$AB \geq 0$ iff A and B commute.

All eigenvalues of AB are nonnegative.

There is a unique square root of AB that has nonnegative eigenvalues.

$$(AB)^{1/2} = A^{1/2}\left(A^{1/2}BA^{1/2}\right)^{1/2}A^{-1/2}$$

is that unique square root.

$(AB)^{1/2}$ and $\left(A^{1/2}BA^{1/2}\right)^{1/2}$ are similar.

Hence $\operatorname{tr}(AB)^{1/2} = \operatorname{tr}\left(A^{1/2}BA^{1/2}\right)^{1/2}.$

Second version

$$d_2(A,B) = \sqrt{\operatorname{tr}\mathcal{A}(A,B) - \operatorname{tr}\left(A^{1/2}BA^{1/2}\right)^{1/2}}.$$

d_2 is a metric on $\mathbb{P}$ much studied as the *Bures distance* in quantum information, and as the *Wasserstein distance* in optimal transport theory and statistics.

Fidelity: $F(A,B) = \operatorname{tr}\left(A^{1/2}BA^{1/2}\right)^{1/2}.$

$$d_2(A,B) = \frac{1}{\sqrt{2}} \min_{U \in \mathbb{U}} \|A^{1/2} - B^{1/2}U\|_2.$$

$\mathcal{G}(A, B) = A\#B$

$$A\#B = A^{1/2}\left(A^{-1/2}BA^{-1/2}\right)^{1/2}A^{1/2}.$$

Introduced by Pusz and Woronowicz in 1975.

Most accepted definition of matrix geometric mean.

Has remarkable properties.

Important applications in diverse areas.

Third version

$$d_3(A, B) = \sqrt{\mathrm{tr}\mathcal{A}(A, B) - \mathrm{tr}A\#B}.$$

The log Euclidean mean

$$\mathcal{L}(A, B) = \exp \left(\frac{\log A + \log B}{2} \right).$$

Important in applications due to ease of computation.

The fourth version

$$d_4(A, B) = \sqrt{\mathrm{tr}\mathcal{A}(A, B) - \mathrm{tr}\mathcal{L}(A, B)}.$$

Are d_3 and d_4 metrics on $\mathbb{P}$?

Clearly, both are symmetric.

Comparing the different geometric means:

$$\operatorname{tr}(A\#B) \leq \operatorname{tr}\mathcal{L}(A,B) \leq \operatorname{tr}(A^{1/2}B^{1/2}) \leq \operatorname{tr}(AB)^{1/2}.$$

This gives

$$d_3^2(A,B) \geq d_4^2(A,B) \geq d_1^2(A,B) \geq d_2^2(A,B).$$

d_1 is a metric $\implies$
$d_3(A,B) = 0$ iff $A = B$, and $d_4(A,B) = 0$ iff $A = B$.

But d_3 and d_4 are *not metrics* as they do not satisfy the triangle inequality.

Let

$$A = \begin{bmatrix} 2 & 5 \\ 5 & 17 \end{bmatrix},\ B = \begin{bmatrix} 13 & 8 \\ 8 & 5 \end{bmatrix},\ C = \begin{bmatrix} 5 & 3 \\ 3 & 10 \end{bmatrix}.$$

Then $d_3(A, B) \approx 5.0347$ and
$d_3(A, C) + d_3(C, B) \approx 4.6768$.

Let

$$A = \begin{bmatrix} 4 & -7 \\ -7 & 13 \end{bmatrix}, B = \begin{bmatrix} 8 & -2 \\ -2 & 1 \end{bmatrix}, C = \begin{bmatrix} 5 & -4 \\ -4 & 5 \end{bmatrix}.$$

Then $d_4(A, B) \approx 3.3349$ and
$d_4(A, C) + d_4(C, B) \approx 3.3146$.

d_3 and d_4 are not metrics but their squares are divergences on $\mathbb{P}$

Divergence on $\mathbb{P}$

$\Phi : \mathbb{P} \times \mathbb{P} \to [0, \infty)$

(*i*) $\Phi(A, B) = 0$ if and only if $A = B$.

(*ii*) The first derivative $D\Phi$ with respect to the second variable vanishes on the diagonal; i.e.,

$$D\Phi(A, X)|_{X=A} = 0.$$

(*iii*) The second derivative $D^2\Phi$ is positive on the diagonal; i.e.,

$$D^2\Phi(A, X)|_{X=A}(Y, Y) \geq 0 \text{ for all Hermitian } Y.$$

Examples

- Square of the Euclidean distance: $\Phi(A, B) = \|A - B\|_2^2$.
- d_1^2 and d_2^2 are divergences.
- Umegaki relative entropy: $d(A\|B) = \operatorname{tr}(A \log A - \log B) - A - B))$.

Bregman divergence

$\varphi : \mathbb{R}_+ \to \mathbb{R}$ strictly convex and differentiable.

$$\tilde{\varphi}(A, B) = \mathrm{tr}\varphi(A) - \mathrm{tr}\varphi(B) - \mathrm{tr}\varphi'(B)(A - B).$$

When $\varphi(x) = x \log x - x$, $\tilde{\varphi}(A, B) = D(A \| B)$.

The functions d_3^2 and d_4^2

Theorem

d_3^2 and d_4^2 are divergences.

We shall denote d_i^2 by Φ_i, $i = 1, \ldots, 4$.

The maps $X \mapsto A\#X$ and Φ_3

Let $g(X) = A\#X$.

For $X > 0$ $X^{1/2} = \frac{1}{\sqrt{2}} + \frac{1}{\pi}\int_0^\infty \left(\frac{\lambda}{\lambda^2+1} - (\lambda + X)^{-1}\right)\lambda^{1/2}\mathrm{d}\lambda.$

$$Dg(X)(Y) = \int_0^\infty (\lambda + XA^{-1})^{-1} Y (\lambda + A^{-1}X)^{-1} \mathrm{d}\nu(\lambda),$$

where $\mathrm{d}\nu(\lambda) = \frac{1}{\pi}\lambda^{1/2}\mathrm{d}\lambda.$

Then

$$D\Phi_3(A, A) = 0,$$

and

$$D^2\Phi_3(A, A)(Y, Y) = \frac{1}{4}\operatorname{tr} YA^{-1}Y.$$

Thus Φ_3 is a divergence.

Convexity and joint convexity

$f : \mathbb{P} \to \mathbb{P}$ or $\mathbb{R}_+$ is *convex* if for all $X, Y \in \mathbb{P}$ and for all $0 < t < 1$

$$f((1-t)X + tY) \leq (1-t)f(X) + tf(Y).$$

f is *strictly convex* if the two sides are equal only if $X = Y$.

$f : \mathbb{P} \times \mathbb{P} \to \mathbb{P}$ or $\mathbb{R}_+$ is *jointly convex* if for all $X_1, X_2, Y_1, Y_2 \in \mathbb{P}$ and for all $0 < t < 1$

$$f((1-t)X_1 + tY_1, (1-t)X_2 + tY_2) \leq (1-t)f(X_1, X_2) + tf(Y_1, Y_2).$$

Convexity of Φ_3

Convexity of Φ_3

Φ_3 is jointly convex, and strictly convex in each of its variables separately.

$(A, B) \mapsto A\#B$ is jointly concave.

(A basic fact in the theory of geometric mean)

This implies $(A, B) \mapsto \mathrm{tr} A\#B$ is jointly concave and hence Φ_3 is jointly convex.

$f : \mathbb{R}_+ \to \mathbb{R}_+$ continuous.

Define $\hat{f} : \mathbb{P} \to \mathbb{R}$ as

$$\hat{f}(X) = \mathrm{tr} f(X).$$

f is concave (strictly concave) iff $\hat{f}$ is so.

$t \mapsto t^{1/2}$ is strictly concave. Hence $X \mapsto \mathrm{tr} X^{1/2}$ is strictly concave.

This implies $X \mapsto \mathrm{tr} A \# X$ is strictly concave.

Coming to Φ_4

$$\Phi_3(A, B) \geq \Phi_4(A, B) \geq \Phi_1(A, B).$$

We also know that

$$\Phi_3(A, A) = \Phi_4(A, A) = \Phi_1(A, A) = 0,$$

and

$$D\Phi_1(A, A) = D\Phi_3(A, A) = 0.$$

These together imply

$$D\Phi_4(A, A) = 0.$$

We have seen that Φ_4 satisfies the first two conditions for being a divergence.

divergence

Φ_4 is a divergence on $\mathbb{P}$.

Third condition: a consequence of the convexity of Φ_4.

We establish a nice connection of Φ_4 with the relative entropy.

A digression

Barycentres with respect to the divergence Φ

$A_1, \ldots, A_m \in \mathbb{P}$.

Problem 1: Find Y_0 in $\mathbb{P}$ that minimizes

$$\sum_{j=1}^{m} \frac{1}{m} \Phi(A_j, Y).$$

Problem 2: Find X_0 in $\mathbb{P}$ that minimizes

$$\sum_{j=1}^{m} \frac{1}{m} \Phi(X, A_j).$$

The minimizers in Problems 1 and 2 need not exist and need not be unique.

Looking at the Bregman divergence on $\mathbb{R}_+$:

$\varphi : \mathbb{R}_+ \to \mathbb{R}$ strictly convex, differentiable.

Associated Bregman divergence on $\mathbb{R}_+$:

$$\Phi(x, y) = \varphi(x) - \varphi(y) - \varphi'(y)(x - y).$$

Φ is strictly convex in x but need not be convex in y.

Solutions of Problem 1

$$\operatorname*{argmin}_{y \in \mathbb{R}_+} \sum_{j=1}^{m} \frac{1}{m} \Phi(a_j, y)$$

is the arithmetic mean $\sum\limits_{j=1}^{m} \frac{1}{m} a_j$.

This is the *characteristic property* of Bregman divergences.

Solution of Problem 2

$$\operatorname*{argmin}_{x\in\mathbb{R}_+} \sum_{j=1}^{m} \frac{1}{m}\Phi(x, a_j)$$

is$\varphi'^{-1}\left(\sum_{j=1}^{m} \frac{1}{m}\phi'(a_j)\right)$.

Minimization problems on matrix Bregman divergences

$\varphi : \mathbb{R}_+ \to \mathbb{R}$ strictly convex and differentiable.

$\tilde{\varphi}$ associated Bregman divergence on $\mathbb{P}$.

Solution to Problem 1 is the *arithmetic mean*.

Solution to Problem 2 is the matrix

$$(\varphi'^{-1}\left(\sum_{j=1}^{m} \frac{1}{m}\varphi'(A_j)\right)$$

A special case

$\varphi(x) = x \log x - x.$

$$\tilde{\varphi}(X, Y) = \operatorname{tr}\left(X(\log X - \log Y) - (X - Y)\right).$$

For $A_1, \ldots, A_m$ in $\mathbb{P}$ the minimizer for

$$\sum_{j=1}^{m} \frac{1}{m} \tilde{\varphi}(X, A_j))$$

is the *log Euclidean mean*

$$\mathcal{L}(A_1, \ldots, A_m) = \exp\left(\sum_{j=1}^{m} \frac{1}{m} \log A_j\right).$$

Computing the *variance,* i.e., the minimum value of the objective function:

$$\begin{aligned}\sigma_{\tilde{\varphi}}^2 &= \frac{1}{m}\sum_{j=1}^{m}\tilde{\varphi}(\mathcal{L}, A_j)\\ &= \frac{1}{m}\sum_{j=1}^{m}[\mathrm{tr}\mathcal{L}(\log\mathcal{L} - \log A_j) - \mathrm{tr}(\mathcal{L} - A_j)]\\ &= \frac{1}{m}\mathrm{tr}\left\{\sum_{j=1}^{m}\left[\mathcal{L}\left(\frac{1}{m}\sum_{k=1}^{m}\log A_k - \log A_j\right) - (\mathcal{L} - A_j)\right]\right\}\\ &= -\mathrm{tr}\mathcal{L} + \frac{1}{m}\mathrm{tr}\sum_{j=1}^{m}A_j.\end{aligned}$$

Thus

$$\sigma_{\tilde{\varphi}}^2 = \mathrm{tr}\mathcal{A}(A_1, \ldots, A_m) - \mathrm{tr}\mathcal{L}(A_1, \ldots, A_m).$$

In particular, the divergence Φ_4 can be characterized as the minimum value

$$\Phi_4(A, B) = \min_{X>0} \left[\tilde{\varphi}(X, A) + \tilde{\varphi}(X, B)\right],$$

where $\tilde{\varphi}(X, Y) = \mathrm{tr}\left(X(\log X - \log Y) - (X - Y)\right).$

Convexity of Φ_4

Let $f(x, y)$ be a jointly convex function which is strictly convex in each of its variables separately. Suppose for each a, b

$$g(a, b) = \min_x \left[f(x, a) + f(x, b) \right],$$

exists. Then the function $g(a, b)$ is jointly convex, and is strictly convex in each of the variables separately.

$\tilde{\varphi}$ is jointly convex and strictly convex in each of its variables.

Convexity of Φ_4

Φ_4 is jointly convex and strictly convex in each of its variables separately.
(Taking $f(X, Y) = \tilde{\varphi}(X, Y)$, we have $g(A, B) = \Phi_4(A, B)$).

Barycentres with respect to the divergences Φ_j

Consider the functions:

$$\Psi_i(X) = \sum_{j=1}^{m} \frac{1}{m} \Phi_i(X, A_j) \;\; X \in \mathbb{P}.$$

Does there exist an X_j that minimizes $\Psi_j(X)$?

Is this X_j unique, if it exists?

If f is a convex function on an open convex set,
then a critical point of f is the global minimum of f.

If f is strictly convex, then f can have at most one such critical point.

$X \mapsto \Psi_i(X)$ is strictly convex on $\mathbb{P}$.

This reduces the problem of finding the minimizer to that of computing the critical point.

The barycentre with respect to Φ_1 is the *classical* $1/2$*-power mean*.

$$Q_{1/2} = \left(\sum_{j=1}^{m} \frac{1}{m} A_j^{1/2} \right)^2 .$$

For Φ_2 is the barycentre is the *Wasserstein mean*. This is the unique solution of the matrix equation

$$X = \sum_{j=1}^{m} \frac{1}{m} (X^{1/2} A_j X^{1/2})^{1/2}.$$

Has major applications in optimal transport, statistics, quantum information and other areas.

An observation

In both cases the barycentre is the unique X_i that satisfies the matrix equation

$$X = \sum_{j=1}^{m} \frac{1}{m} \mathcal{G}_i(X, A_j),$$

where $\mathcal{G}_1(X, A) = X^{1/4}A^{1/2}X^{1/4}$ and $\mathcal{G}_2(A, X) = \left(X^{1/2}AX^{1/2}\right)^{1/2}$.

With some work we can see that the same holds for the other two as well, i.e., the barycentres for Φ_3 and Φ_4 are the unique matrices that satisfy the respective matrix equations

$$X = \sum_{j=1}^{m} X \# A_j,$$

(Lim-Palfia power mean important in the study of geometric means.) and

$$X = \sum_{j=1}^{m} \frac{1}{m} \mathcal{L}(X, A_j).$$

References

(1) R. Bhatia, S. Gaubert and T. Jain, *Matrix versions of the Hellinger distance*, submitted.

(2) S. Amari, *Information Geometry and its Applications*, Springer (Tokyo), 2016.

(3) A. Banerjee, S. Merugu, I. S. Dhillon and J. Ghosh, *Clustering with Bregman divergences*, J. Mach. Learn. Res. 6 (2005), 1705-1749.

(4) H. Bauschke and J. M. Borwein, *Joint and separate convexity of the Bregman distance*, Stud. Comput. Math. 8 (2001), 23-36.

(5) F. Nielsen and R. Bhatia, eds., *Matrix Information Geometry*, Springer, 2013.

(6) R. Bhatia, *The Riemannian mean of positive matrices*, in Matrix Information Geometry, eds. F. Nielsen and R. Bhatia, Springer, (2013), 35-51.

(7) R. Bhatia, T. Jain and Y. Lim , *On the Bures-Wasserstein distance between positive definite matrices*, Expos. Math., to appear.

(8) M. Agueh and G. Carlier, *Barycenters in the Wasserstein space*, SIAM J. Math. Anal. Appl. 43 (2011), 904-924.

(9) V. Arsigny, P. Fillard, X. Pennec and N. Ayache, *Geometric means in a novel vector space structure on symmetric positive-definite matrices*, SIAM J. Math. Anal. Appl. 29 (2007), 328-347.

(10) R. Bhatia, T. Jain and Y. Lim, *Strong convexity of sandwiched entropies and related optimization problems*, Rev. Math. Phys. 30 (2018), 1850014.

(11) Y. Lim and M. Palfia, *Matrix power means and the Karcher mean*, J. Funct. Anal. 262 (2012), 1498-1514.

(12) I. S. Dhillon and J. A. Tropp, *Matrix nearness problems with Bregman divergences*, SIAM J. Matrix Anal. Appl. 29 (2004), 1120-1146.

(13) A. Jencova and M. B. Ruskai, *A unified treatment of convexity of relative entropy and related trace functions with conditions for equality*, Rev. Math. Phys. 22 (2010), 1099-1121.