ON THE BURES-WASSERSTEIN DISTANCE BETWEEN POSITIVE DEFINITE MATRICES

RAJENDRA BHATIA, TANVI JAIN, AND YONGDO LIM

ABSTRACT. The metric $d(A,B) = \left[\operatorname{tr} A + \operatorname{tr} B - 2\operatorname{tr}(A^{1/2}BA^{1/2})^{1/2}\right]^{1/2}$ on the manifold of $n \times n$ positive definite matrices arises in various optimisation problems, in quantum information and in the theory of optimal transport. It is also related to Riemannian geometry. In the first part of this paper we study this metric from the perspective of matrix analysis, simplifying and unifying various proofs. Then we develop a theory of a mean of two, and a barycentre of several, positive definite matrices with respect to this metric. We explain some recent work on a fixed point iteration for computing this Wasserstein barycentre. Our emphasis is on ideas natural to matrix analysis.

1. INTRODUCTION

Let $\mathbb{M}(n)$ be the space of $n\times n$ complex matrices, $\mathbb{H}(n)$ the real subspace of $\mathbb{M}(n)$ consisting of Hermitian matrices, and $\mathbb{P}(n)$ the subset of $\mathbb{H}(n)$ consisting of positive semi definite (psd) matrices. The Frobenius inner product on $\mathbb{M}(n)$ is defined as $\langle A, B\rangle = \operatorname{Re}\operatorname{tr} A^*B$, and the associated norm $\|A\|_2 = (\operatorname{tr} A^*A)^{1/2}$ is called the Frobenius norm. Every psd matrix A has a unique psd square root, which we denote by $A^{1/2}$. Given A, B in $\mathbb{P}(n)$ define $d(A,B)$ by the relation

$$d(A,B) = \left[\operatorname{tr} A + \operatorname{tr} B - 2\operatorname{tr}\,\left(A^{1/2}BA^{1/2}\right)^{1/2}\right]^{1/2}. \tag{1}$$

It turns out that $d(A,B)$ is a metric on the space $\mathbb{P}(n)$. This metric has been of interest in quantum information where it is called the *Bures distance,* and in statistics and the theory of optimal transport where it is called the *Wasserstein metric.* If A and B are diagonal matrices, then $d(A,B)$ reduces to the Hellinger distance between probability distributions and is related to the Rao-Fisher metric in information theory. The metric d is of interest in differential geometry, as it is the distance function corresponding to a Riemannian metric.

In this paper we explore some fundamental properties of this metric from the perspective of matrix analysis. This allows us to unify several known facts and to simplify their proofs, to point out new connections, to raise new questions and to answer some of them.

2010 *Mathematics Subject Classification.* 47A63, 47A64, 53C20, 53C22, 60D05, 81P45.
Key words and phrases. Positive definite matrices, Bures distance, Wasserstein metric, Riemannian metric, matrix mean, optimal transport, coupling problem, fidelity.

2. Some variational principles

The metric $d(A,B)$ and the quantity $(A^{1/2}BA^{1/2})^{1/2}$ occurring in it, both are related to solutions of extremal problems arising in different contexts.

Recall that a matrix A is psd if and only if it can be expressed as $A = MM^*$ for some $M \in \mathbb{M}(n)$. Another matrix N satisfies the relation $A = NN^*$ if and only if $N = MU$ for some unitary matrix U. One special matrix among all these is $A^{1/2}$. Let $U(n)$ stand for the group of all unitary matrices. Given a psd matrix A let $\mathcal{F}(A)$ be the set defined as

$$\begin{aligned} \mathcal{F}(A) &= \{M \in \mathbb{M}(n) : A = MM^*\} \\ &= \left\{A^{1/2}U : U \in U(n)\right\}. \end{aligned} \tag{2}$$

Theorem 1. *If $d(A,B)$ is defined as in (1), then*

$$\begin{aligned} d(A,B) &= \min_{\substack{M\in\mathcal{F}(A)\\ N\in\mathcal{F}(B)}} \|M-N\|_2 \\ &= \min_{U\in U(n)} \|A^{1/2} - B^{1/2}U\|_2. \end{aligned} \tag{3}$$

The minimum in the last expression in (3) is attained at a unitary U occurring in the polar decomposition of $B^{1/2}A^{1/2}$:

$$B^{1/2}A^{1/2} = U|B^{1/2}A^{1/2}| = U(A^{1/2}BA^{1/2})^{1/2}.$$

Proof. We have for every $U \in U(n)$

$$\begin{aligned} \|A^{1/2} - B^{1/2}U\|_2^2 &= \|A^{1/2}\|_2^2 + \|B^{1/2}\|_2^2 - \operatorname{tr}(A^{1/2}U^*B^{1/2} + A^{1/2}B^{1/2}U) \\ &= \operatorname{tr}A + \operatorname{tr}B - \operatorname{tr}(U^*B^{1/2}A^{1/2} + A^{1/2}B^{1/2}U). \end{aligned}$$

Hence,

$$\begin{aligned} &\min_{U\in U(n)} \|A^{1/2} - B^{1/2}U\|_2^2 \\ &\quad = \operatorname{tr}A + \operatorname{tr}B - \max_{U\in U(n)} \operatorname{tr}(U^*B^{1/2}A^{1/2} + A^{1/2}B^{1/2}U). \end{aligned} \tag{4}$$

To evaluate the maximum in (4) let $X = B^{1/2}A^{1/2}$. Then $|X| := (X^*X)^{1/2} = (A^{1/2}BA^{1/2})^{1/2}$. Let $X = VP$ be the polar decomposition of X, where $P = |X|$ and V is unitary. Then

$$\operatorname{tr}(U^*B^{1/2}A^{1/2} + A^{1/2}B^{1/2}U) = \operatorname{tr}(U^*X + X^*U) = \operatorname{tr}(U^*VP + PV^*U).$$

Putting $W = U^*V$, the last expression above can be written as $\operatorname{tr}(W + W^*)P$. Choosing a basis in which $W = \operatorname{diag}(e^{i\theta_1}, \cdots, e^{i\theta_n})$ we have

$$\operatorname{tr}(W+W^*)P = \sum_{j=1}^{n} (2\cos\theta_j)p_{jj}.$$

The maximum value of this is attained when $W = I$ and is equal to

$$\sum_{j=1}^{n} 2p_{jj} = 2\operatorname{tr}P = 2\operatorname{tr}|X|.$$

So, from (4) we have

$$\min_{U \in U(n)} \|A^{1/2} - B^{1/2}U\|_2^2 = \mathrm{tr}A + \mathrm{tr}B - 2\mathrm{tr}(A^{1/2}BA^{1/2})^{1/2}.$$

This shows the equality of the two extreme sides of (3). The expression in the middle is equal to this because of the unitary invariance of $\|\cdot\|_2$. In the course of the proof we saw that the minimum in (3) is attained when $W = U^*V = I$; i.e., when $U = V$, the polar factor for $X = B^{1/2}A^{1/2}$. ■

From the representations in (3) it is easy to see that $d(A,B)$ is indeed a metric. Obviously $d(A,B) \geq 0$. The compact sets $\mathcal{F}(A)$ and $\mathcal{F}(B)$ are disjoint unless $A = B$. So $d(A,B) = 0$ if and only if $A = B$. To prove the triangle inequality, note that for all psd matrices A, B, C, and unitaries U, V, we have

$$\begin{aligned} d(A,B) &\leq \|A^{1/2} - B^{1/2}U\|_2 \\ &\leq \|A^{1/2} - C^{1/2}V\|_2 + \|B^{1/2}U - C^{1/2}V\|_2 \\ &= \|A^{1/2} - C^{1/2}V\|_2 + \|B^{1/2} - C^{1/2}VU^*\|_2. \end{aligned}$$

Taking the minimum over all U, V, we see that $d(A,B) \leq d(A,C) + d(B,C)$. This proof is adopted from [17].

A well-known and important problem in factor analysis and in multidimensional scaling is the orthogonal Procrustes problem. This asks for the solution of the minimisation problem $\min \|A - BU\|_2$ where A and B are given matrices (not necessarily psd) and U varies over unitaries. The argument in Theorem 1 shows that the minimum is attained when U is the unitary polar factor of B^*A. In applications A and B represent multivariate data sets, and the problem is to ascertain whether they are equivalent up to a rotation. See [18] for a brief and [16] for an expansive discussion.

In the following remarks we point out some more connections between the Bures distance and some other classical problems in matrix analysis.

1. The expression in (1) is reminiscent of the matrix arithmetic-geometric mean inequality [11]. Indeed this inequality tells us that for any two psd matrices $|||A^{1/2}B^{1/2}||| \leq \frac{1}{2}|||A+B|||$ for every unitarily invariant norm. For the trace norm $\|\cdot\|_1$ the left hand side of this inequality is equal to $\mathrm{tr}(A^{1/2}BA^{1/2})^{1/2}$ and the right hand side to $\frac{1}{2}(\mathrm{tr}A + \mathrm{tr}B)$. That these two quantities are equal if and only if $A = B$ is one of the assertions included in the statement that $d(A,B)$ is a metric. (This has been known for the Schatten p-norms, $1 < p < \infty$ [21], and is false for the case $p = \infty$.)

2. Let $p = (p_1, \ldots, p_n)$ and $q = (q_1, \ldots, q_n)$ be nonnegative vectors, and let

$$\rho(p,q) = \left[\sum_{i=1}^{n} (\sqrt{p_i} - \sqrt{q_i})^2 \right]^{1/2}. \tag{5}$$

This is the l_2 norm distance between the square roots of the vectors p and q. If p and q are probability distributions (i.e., $\sum p_i =$

$\sum q_i = 1$), then $\rho(p,q)$ is called the *Hellinger distance*. In analogy, one could define a distance $\rho(A,B)$ on psd matrices by putting $\rho(A,B) = \|A^{1/2} - B^{1/2}\|_2$. When A and B commute, the distance $d(A,B)$ is equal to $\rho(A,B)$.

3. A psd matrix A with $\mathrm{tr}A = 1$ is called a *density matrix* or a *state*. The quantity

$$F(A,B) = \mathrm{tr}(A^{1/2}BA^{1/2})^{1/2} \tag{6}$$

is called the *fidelity* between two states A and B. In this case from (1) we see that

$$\frac{1}{2}d^2(A,B) = 1 - F(A,B). \tag{7}$$

In the quantum information theory literature it is customary to define the Bures distance between density matrices A and B as the quantity $\sqrt{1 - F(A,B)}$. This is just the distance (1) restricted to density matrices. An illuminating discussion of the Bures distance from the QIT perspective can be found in [5]. If $A = uu^*$ for some unit vector u, then A is called a *pure state*. In this case we have

$$F(A,B) = \langle u, Bu\rangle^{1/2}.$$

If both A and B are pure states given as $A = uu^*$, $B = vv^*$, then

$$F(A,B) = |\langle u,v\rangle|,$$

and

$$\frac{1}{2}d^2(A,B) = 1 - |\langle u,v\rangle|.$$

4. The Bures distance is related to a measure of separation between subspaces of $\mathbb{C}^n$. Let $\mathcal{M}$ and $\mathcal{N}$ be two l-dimensional subspaces of $\mathbb{C}^n$, and let P and Q be the orthogonal projections with ranges $\mathcal{M}$ and $\mathcal{N}$, respectively. Among all unitary operators on $\mathbb{C}^n$ that map $\mathcal{M}$ onto $\mathcal{N}$, there is a special one called a *direct rotation*. This unitary operator U can be represented in a particular orthonormal basis as

$$U = \begin{bmatrix} C & -S & O \\ S & C & O \\ O & O & I \end{bmatrix},$$

where C and S are nonnegative diagonal matrices. If $2l \le n$, then C and S are $l \times l$ matrices, and if $2l > n$, then they are $(n-l)\times(n-l)$ matrices. Further, $C^2 + S^2 = I$. The operator $\Theta(\mathcal{M},\mathcal{N}) = \arccos C$ is called the *angle operator* between $\mathcal{M}$ and $\mathcal{N}$. The diagonal entries of this diagonal operator are called the *canonical angles* between the spaces $\mathcal{M}$ and $\mathcal{N}$. It can be seen that the nonzero singular values of PQ are the nonzero diagonal entries of C. The direct rotation was used in [13] in connection with perturbation theory of eigenvectors. See also [6], Section VII.1 and [32] Chapter II, Section 4. The fidelity between projections P and Q is the sum of the cosines of the canonical angles between the spaces $\mathcal{M}$ and $\mathcal{N}$:

$$F(P,Q) = \mathrm{tr}(PQP)^{1/2} = \|PQ\|_1 = \sum c_j.$$

Here c_j are the diagonal entries of C if $2l \leq n$. In the case when $2l > n$, we take c_j to be the diagonal entries of C for $1 \leq j \leq n-l$ and take them to be 0 for $n-l < j \leq l$. They are thus the cosines of the canonical angles between $\mathcal{M}$ and $\mathcal{N}$. We have

$$\frac{1}{2}d^2(P,Q) = \sum (1-c_j).$$

The fidelity $F(A,B)$ is a quantity of great interest and it is useful to have more descriptions of it. Some variational characterisations of it are given below. We need some facts from the theory of geometric means. See Chapter 4 of [7].
Let A and B be positive definite matrices. Their geometric mean $A\#B$ is defined by the Pusz-Woronowicz formula [30]

$$A\#B = A^{1/2}(A^{-1/2}BA^{-1/2})^{1/2}A^{1/2}. \tag{8}$$

This mean is symmetric in A and B. It is the unique positive definite solution of the Riccati equation

$$XA^{-1}X = B. \tag{9}$$

The matrix AB has positive eigenvalues, and it has a unique square root $(AB)^{1/2}$ that has positive eigenvalues. The eigenvalues of BA are the same as those of AB. We have

$$A\#B = A(A^{-1}B)^{1/2} = (AB^{-1})^{1/2}B. \tag{10}$$

Another useful characterisation is

$$A\#B = \max\left\{X : \begin{bmatrix} A & X \\ X & B \end{bmatrix} \geq 0\right\}. \tag{11}$$

Here the maximum is with respect to the Loewner partial order; for Hermitian matrices X and Y we say $X \geq Y$ if $X - Y$ is psd. We recall also two necessary and sufficient conditions for the block matrix

$$\begin{bmatrix} A & X \\ X^* & B \end{bmatrix} \tag{12}$$

to be psd. The first says that the matrix (12) is psd if and only if

$$A \geq XB^{-1}X^*, \tag{13}$$

and the second that this is so if and only if there exists a contraction K (an operator with $\|K\| \leq 1$) such that

$$X = A^{1/2}KB^{1/2}. \tag{14}$$

See Chapter 1 of [7].

Theorem 2. *Let A and B be positive definite matrices. Then*

$$\text{(i)} \qquad F(A,B) = \min_{X>0} \frac{1}{2}\mathrm{tr}(AX+BX^{-1}). \tag{15}$$

$$\text{(ii)} \qquad F(A,B) = \min_{X>0} \sqrt{\mathrm{tr}(AX)\mathrm{tr}(BX^{-1})}. \tag{16}$$

$$\text{(iii)} \qquad F(A,B) = \max_{X>0} \left\{|\mathrm{tr}X| : A \geq XB^{-1}X^*\right\}. \tag{17}$$

Proof. (i) Consider the function $f(X) = \mathrm{tr}(AX+BX^{-1})$ defined on $\mathbb{P}(n)$. This is a convex function and its derivative $Df(X)$ is the linear map from $\mathbb{H}(n)$ into $\mathbb{R}$ given by the formula.

$$Df(X)(Y) = \mathrm{tr}(AY - BX^{-1}YX^{-1}), \ \ Y \in \mathbb{H}(n).$$

(See [6] pp.310 - 312.) So a point X_0 is a minimum for f if and only if

$$\mathrm{tr}(A - X_0^{-1}BX_0^{-1})Y = 0, \ \ Y \in \mathbb{H}(n).$$

This is so if and only if $A - X_0^{-1}BX_0^{-1} = 0$, or in other words $X_0AX_0 = B$. This is the Riccati equation (9). So, $X_0 = A^{-1}\#B$. We have then

$$\min_{X>0} f(X) = f(X_0) = \mathrm{tr}(A(A^{-1}\#B) + B(A\#B^{-1})).$$

Using (10), the right hand side of this equation can be expressed as

$$\begin{aligned}\mathrm{tr}(A\cdot A^{-1}(AB)^{1/2} + B(AB)^{1/2}B^{-1}) &= 2\mathrm{tr}(AB)^{1/2} \\ &= 2\mathrm{tr}(A^{1/2}BA^{1/2})^{1/2} = 2F(A,B).\end{aligned}$$

This proves (i).

(ii) In the proof of (i) above we have seen that at $X_0 = A^{-1}\#B$, we have

$$\mathrm{tr}AX_0 = \mathrm{tr}BX_0^{-1} = F(A,B).$$

So

$$\frac{\mathrm{tr}AX_0 + \mathrm{tr}BX_0^{-1}}{2} = \sqrt{\mathrm{tr}(AX_0)\mathrm{tr}(BX_0^{-1})}.$$

This proves (ii).

(iii) We have remarked earlier that

$$A \geq MB^{-1}M^* \Leftrightarrow \begin{bmatrix} A & M \\ M^* & B \end{bmatrix} \geq 0 \Leftrightarrow M = A^{1/2}KB^{1/2}$$

for some contraction K. By the Schwarz inequality we have

$$\begin{aligned}|\mathrm{tr}M| = |\mathrm{tr}(A^{1/2}KB^{1/2}| &\leq \|A^{1/2}K\|_2\|B^{1/2}\|_2 \\ &\leq \|A^{1/2}\|_2\|B^{1/2}\|_2 = \sqrt{\mathrm{tr}A\mathrm{tr}B}.\end{aligned}$$

If

$$\begin{bmatrix} A & M \\ M^* & B \end{bmatrix} \geq 0,$$

then for every $X > 0$ we have

$$\begin{aligned} 0 &\le \begin{bmatrix} X^{1/2} & O \\ O & X^{-1/2} \end{bmatrix} \begin{bmatrix} A & M \\ M^* & B \end{bmatrix} \begin{bmatrix} X^{1/2} & O \\ O & X^{-1/2} \end{bmatrix} \\ &= \begin{bmatrix} X^{1/2}AX^{1/2} & X^{1/2}MX^{-1/2} \\ X^{-1/2}M^*X^{1/2} & X^{-1/2}BX^{-1/2} \end{bmatrix}. \end{aligned}$$

Hence

$$|\mathrm{tr} X^{1/2}MX^{-1/2}| \le \sqrt{\mathrm{tr}(X^{1/2}AX^{1/2})\ \mathrm{tr}(X^{-1/2}BX^{-1/2})}.$$

In other words,

$$|\mathrm{tr} M| \le \sqrt{\mathrm{tr}(AX)\ \mathrm{tr}(BX^{-1})}.$$

This is true for all M satisfying the condition $A \ge MB^{-1}M^*$ and for all $X > 0$. So

$$\begin{aligned} \max\left\{|\mathrm{tr} M| : A \ge MB^{-1}M^*\right\} &\le \min_{X>0} \sqrt{\mathrm{tr}(AX)\mathrm{tr}(BX^{-1})} \\ &= F(A,B). \end{aligned} \tag{18}$$

Let $M = (AB)^{1/2} = A(A^{-1}\#B)$. Then

$$\begin{aligned} MB^{-1}M^* &= (AB)^{1/2}B^{-1}(BA)^{1/2} = B^{-1}B(AB)^{1/2}B^{-1}(BA)^{1/2} \\ &= B^{-1}(BA)^{1/2}(BA)^{1/2} = B^{-1}BA = A. \end{aligned}$$

So, the maximum on the left hand side of (18) is attained when $M = (AB)^{1/2}$ and it is equal to $\mathrm{tr}(A^{1/2}BA^{1/2})^{1/2} = F(A,B)$. This proves (iii). ■

Theorem 2 with different proofs can be found in [2, 35].

3. The Statistical distance

Let X, Y be complete separable metric spaces and let μ, ν be Borel probability measures on X and Y, respectively. Let $\Gamma(\mu,\nu)$ be the collection of probability measures γ on $X \times Y$ whose marginals are μ and ν. Let $c(x,y)$ be a nonnegative Borel measurable function on $X \times Y$. The *optimal transport problem* is the minimisation problem of finding

$$\inf_{\gamma \in \Gamma(\mu,\nu)} \int_{X \times Y} c(x,y)\, d\gamma(x,y).$$

(Here μ, ν are thought of as mass distributions, and $c(x,y)$ is the cost of moving a unit mass from X to Y. The problem is of moving one mass distribution to another at the least cost.)

An important special case of this problem is the following. Let $X = Y = \mathbb{C}^n$, let μ, ν have finite second moments, and let $c(x,y) = \|x-y\|^2$. In this case it can be shown that the quantity

$$d_W(\mu,\nu) = \inf_{\gamma \in \Gamma(\mu,\nu)} \left(\int_{\mathbb{R}^n \times \mathbb{R}^n} \|x-y\|^2\, d\gamma(x,y) \right)^{1/2} \tag{19}$$

defines a metric, which is called the 2 -*Wasserstein distance* between μ and ν. The integral on the right hand side of (19) is also written as $E\|x-y\|^2$, where E stands for expectation.

In the most important special case of Gaussian measures, the distance d_W coincides with the Bures distance, and this is explained below.

Let x and y be random vectors with values in $\mathbb{C}^n$, each having zero mean, and with covariance matrices A and B, respectively. This last statement means that

$$A=[E(\overline{x}_i x_j)]\,, \quad B=[E(\overline{y}_i y_j)]\,. \tag{20}$$

We want to find x and y for which $E\|x-y\|^2$ is minimal.

The covariance matrix of the vector (x,y) is

$$C=\begin{bmatrix} [E(\overline{x}_i x_j)] & [E(\overline{x}_i y_j)] \\ \\ [E(\overline{y}_i x_j)] & [E(\overline{y}_i y_j)] \end{bmatrix} = \begin{bmatrix} A & M \\ M^* & B \end{bmatrix}. \tag{21}$$

Our problem is to minimise

$$\begin{aligned} E\|x-y\|^2 &= E\left(\sum_{i=1}^n (|x_i|^2+|y_i|^2-2\mathrm{Re}\ \overline{x}_i y_i)\right) \\ &= \sum_{i=1}^n E(|x_i|^2+|y_i|^2-2\mathrm{Re}\ \overline{x}_i y_i) \\ &= \mathrm{tr}(A+B)-2\mathrm{Re}\ (\mathrm{tr}\ M). \end{aligned} \tag{22}$$

This is the problem of finding

$$\max\left\{|\mathrm{tr}M| : C=\begin{bmatrix} A & M \\ M^* & B \end{bmatrix} \geq 0\right\}. \tag{23}$$

(As we vary x and y over all vectors with covariance matrices A and B, the covariance matrix of (x,y) varies over all psd matrices of the form in (21).) By Theorem 2(iii) the value of the maximum in (23) is $F(A,B)$. So

$$\begin{aligned} \min\ E\|x-y\|^2 &= \mathrm{tr}(A+B)-2\mathrm{tr}(A^{1/2}BA^{1/2})^{1/2} \\ &= d^2(A,B) \end{aligned}$$

where $d(A,B)$ is as defined in (1).

Let x be a vector with mean 0 and covariance matrix A. Then for any $T\in\mathbb{M}(n)$ we have

$$\begin{aligned} E(\langle x,Tx\rangle) &= E\left(\sum_{i,j} t_{ij}\ \overline{x_i}\ x_j\right) = \sum_{i,j} t_{ij}\ E(\overline{x_i}\ x_j) \\ &= \sum_{i,j} t_{ij}\ a_{ij} = \mathrm{tr}\ TA. \end{aligned}$$

Hence,

$$\begin{aligned} E\|x - Tx\|^2 &= E(\|x\|^2 + \|Tx\|^2 \ - \ 2\mathrm{Re}\ \langle x, Tx\rangle) \\ &= \mathrm{tr} A + \mathrm{tr}\ T^*TA \ - 2\mathrm{Re}\ \mathrm{tr}\ TA \\ &= \mathrm{tr}\ A \ + \mathrm{tr}\ TAT^* - 2\mathrm{Re}\ \mathrm{tr}\ A^{1/2}TA^{1/2}. \end{aligned}$$

If we choose $T = A^{-1}\#B$, then from (8) we see that $\mathrm{tr}\ A^{1/2}\, TA^{1/2} = \mathrm{tr}\ (A^{1/2}BA^{1/2})^{1/2}$, and from (9) that $\mathrm{tr}\ TAT = \mathrm{tr}\ B$. Thus, for this choice of T, we have

$$\begin{aligned} E\|x - \ Tx\|^2 &= \mathrm{tr}\ (A+B) - 2\,\mathrm{tr}\ (A^{1/2}BA^{1/2})^{1/2} \\ &= d^2(A, B). \end{aligned}$$

Thus the problem

$$\min\ E\|x - y\|^2$$

where x, y are vectors with mean zero and covariance matrices A and B, respectively, has as its solution the pairs (x, y), where x is any vector and $y = Tx$, with $T = A^{-1}\#B$. The matrix T is called the *optimal transport plan*, or the *optimal transport map,* from A to B.

Let x be a vector with covariance matrix A, and let $y = Tx$. Then

$$\begin{aligned} E(\overline{y}_i y_j) &= E \sum_{k,l} t_{ik}\ t_{kl}\ \overline{x}_k\ x_l \\ &= \sum_{k,l} t_{ik}\ t_{kl}\ a_{kl} \ = \ (TAT)_{ij}. \end{aligned}$$

If T is the optimal transport map from A to B, then $TAT = B$. This shows that the covariance matrix of the vector y is B.

The results in this section were proved by Olkin and Pukelsheim [29] and by Dowson and Landau [14]. The authoritative reference for optimal transport theory is [36]. An interesting article explaining connections between optimal transport and Riemannian geometry is [4].

4. Riemannian geometry

The Bures-Wasserstein distance corresponds to a Riemannian metric, and that is explained now.

From now on we consider positive definite (i.e., nonsingular psd) matrices. We continue to use the notation $\mathbb{P}(n)$ for the set of all such matrices. This is an open subset of the real vector space $\mathbb{H}(n)$. Let $GL(n)$ be the set of all nonsingular matrices. This is an open subset of $\mathbb{M}(n)$. Both $GL(n)$ and $\mathbb{P}(n)$ are viewed here as differentiable manifolds.

Let $\pi : GL(n) \to \mathbb{P}(n)$ be the map defined as $\pi(M) = MM^*$. This is a differentiable map, and its derivative $D\pi(M)$ at any point M is a linear map from $\mathbb{M}(n)$ to $\mathbb{H}(n)$. The action of this map is

$$D\pi(M)(Z) = ZM^* + MZ^*, \quad Z \in \mathbb{M}(n). \tag{24}$$

The kernel of this map is

$$\begin{aligned}\ker D\pi(M) &= \{Z : ZM^* + MZ^* = 0\} \\ &= \{Z : ZM^* \text{ is skew-Hermitian}\} \\ &= \Big\{Z = KM^{*-1} : K \text{ skew-Hermitian}\Big\}. \qquad (25)\end{aligned}$$

The orthogonal complement of this space with respect to the Frobenius inner product can be readily computed. A matrix X is in this orthogonal complement, if and only if we have for all skew-Hermitian matrices K

$$0 = \langle X, KM^{*-1}\rangle = \mathrm{Re}\,\mathrm{tr} X^* K M^{*-1} = \mathrm{Re}\,\mathrm{tr} M^{*-1} X^* K.$$

This happens if and only if $M^{*-1}X^*$ is Hermitian; i.e. XM^{-1} is Hermitian. Thus

$$(\ker D\pi(M))^{\perp} = \{X = HM : H \in \mathbb{H}(n)\}. \qquad (26)$$

So, we have a direct sum decomposition of the tangent space $T_M GL(n) = \mathbb{M}(n)$ as

$$\begin{aligned}T_M\ GL(n) &= \ker D\pi(M) \oplus (\ker D\pi(M))^{\perp} \\ &= \mathcal{V}_M \oplus \mathcal{H}_M. \qquad (27)\end{aligned}$$

The spaces $\mathcal{V}_M$ and $\mathcal{H}_M$, given by (25) and (26) are, respectively, called the *vertical space* and the *horizontal space* at M (for the map π).

At this stage we recall two theorems from Riemannian geometry. Let $(\mathcal{M}, g)$ and $(\mathcal{N}, h)$ be Riemannian manifolds with Riemannian metrics g and h. A differentiable map $\pi : \mathcal{M} \to \mathcal{N}$ is said to be a *smooth submersion* if its differential $D\pi(m) : T_m\mathcal{M} \to T_{\pi(m)}\mathcal{N}$ is surjective at every point m. Let $T_m\mathcal{M} = \mathcal{V}_m \oplus \mathcal{H}_m$ be a decomposition of $T_m\mathcal{M}$ into vertical and horizontal spaces. Then π is called a *Riemannian submersion* if it is a smooth submersion and the map $D\pi(m) : \mathcal{H}_m \to T_{\pi(m)}\mathcal{N}$ is isometric for all m.

Theorem 3. *Let $(\mathcal{M}, g)$ be a Riemannian manifold. Let G be a compact Lie group of isometries of $(\mathcal{M}, g)$ acting freely on $\mathcal{M}$. Let $\mathcal{N} = \mathcal{M}/G$ and let $\pi : \mathcal{M} \to \mathcal{N}$ be the quotient map. Then there exists a unique Riemannian metric h on $\mathcal{N}$ for which $\pi : (\mathcal{M}, g) \to (\mathcal{N}, h)$ is a Riemannian submersion.*

Theorem 4. *Let $(\mathcal{M}, g)$ and $(\mathcal{N}, h)$ be Riemannian manifolds and $\pi : (\mathcal{M}, g) \to (\mathcal{N}, h)$ a Riemannian submersion. Let γ be a geodesic in $(\mathcal{M}, g)$ such that $\gamma'(0)$ is horizontal. Then*

(i) *$\gamma'(t)$ is horizontal for all t.*
(ii) *$\pi \circ \gamma$ is a geodesic in $(\mathcal{N}, h)$ of the same length as γ.*

See [15].

Let us return to our setup now. $GL(n)$ is a Riemannian manifold with the metric induced by the Frobenius inner product. The group $U(n)$ is a compact Lie group of isometries for this metric. The quotient space $GL(n)/U(n)$ is

$\mathbb{P}(n)$. The metric inherited by the quotient space $\mathbb{P}(n)$ is (upto a constant factor) exactly the one given in Theorem 1; i.e.,

$$\min\|A^{1/2} - B^{1/2}U\|_2 = d(A, B). \tag{28}$$

The map $\pi(M) = MM^*$ is a smooth submersion, as is evident from (26). By Theorem 3 there is a unique Riemannian metric on $\mathbb{P}(n)$ (for each point A of $\mathbb{P}(n)$ an inner product $\langle\cdot,\cdot\rangle_A$ on the tangent space $T_A\mathbb{P}(n) = \mathbb{H}(n)$) for which π is a Riemannian submersion. To find this inner product we proceed as follows. Let $A = MM^*$. We want the map $D\pi(M) : \mathcal{H}_M \to T_A\mathbb{P}(n) = \mathbb{H}(n)$ to be an isometry. The inner product between two elements HM and KM in the horizonal space $\mathcal{H}_M$ is $\langle HM, KM\rangle = \operatorname{Re}\operatorname{tr} KMM^*H = \operatorname{Re}\operatorname{tr} KAH$. By (24) we have $D\pi(M)(HM) = HMM^* + MM^*H = HA + AH$. So for $D\pi(M)$ to be an isometry the inner product $\langle\cdot,\cdot\rangle_A$ on $T_A\mathbb{P}(n) = \mathbb{H}(n)$ must be given by

$$\langle HA + AH, KA + AK\rangle_A = \operatorname{Re}\operatorname{tr} KAH. \tag{29}$$

Let Y be any element of $\mathbb{H}(n)$. Then there exists a unique $H \in \mathbb{H}(n)$ such that

$$HA + AH = Y. \tag{30}$$

Indeed, in an orthonormal basis in which $A = \operatorname{diag}(\alpha_1, \ldots, \alpha_n)$ the equation (30) is satisfied by the matrix H with entries

$$h_{ij} = \frac{y_{ij}}{\alpha_i + \alpha_j}. \tag{31}$$

Let Z be another element of $\mathbb{H}(n)$. Then the matrix $k_{ij} = z_{ij}/(\alpha_i + \alpha_j)$ satisfies the equation $KA + AK = Z$. So from (29) and (30) we get

$$\langle Y, Z\rangle_A = \sum_{i,j} \alpha_i \frac{\operatorname{Re}\overline{y}_{ji} z_{ji}}{(\alpha_i + \alpha_j)^2}. \tag{32}$$

To sum up, we have proved the following.

Theorem 5. *For each $A \in \mathbb{P}(n)$ let $\langle Y, Z\rangle_A$ be the inner product on $\mathbb{H}(n)$ given by* (32)*. This gives a Riemannian metric on the manifold $\mathbb{P}(n)$, the distance function corresponding to which coincides with* (28)*.*

Figure 1 is a schematic representation of the Riemannian submersion in Theorem 5.

Next, we obtain a formula for the geodesic joining A and B in $\mathbb{P}(n)$. Let U be the unitary polar factor of $B^{1/2}A^{1/2}$; i.e.,

$$B^{1/2}A^{1/2} = U|B^{1/2}A^{1/2}| = U(A^{1/2}BA^{1/2})^{1/2}, \tag{33}$$

and let

$$Z(t) = (1-t)A^{1/2} + tB^{1/2}U, \quad 0 \le t \le 1. \tag{34}$$

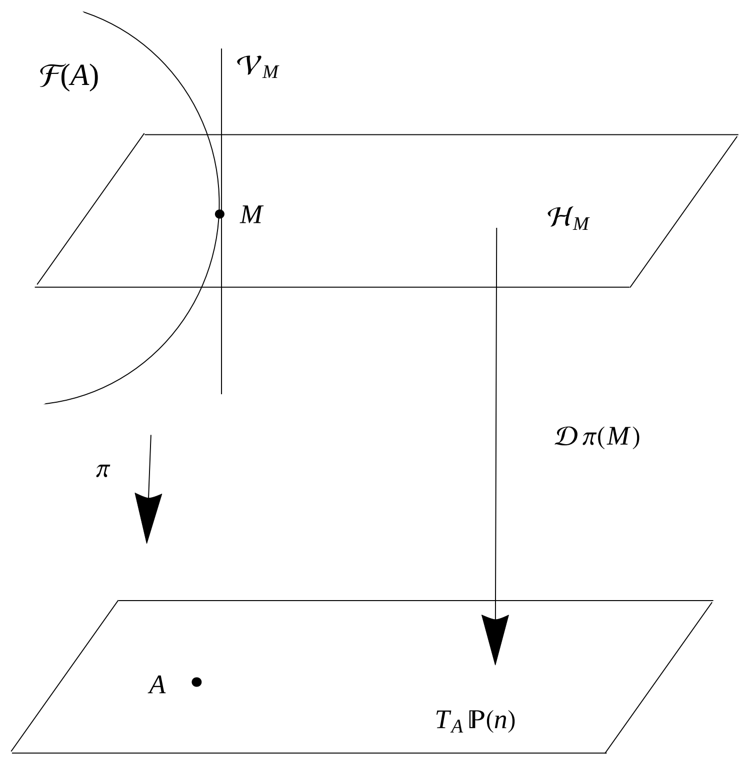

FIGURE 1.

From (33) we have

$$\begin{aligned} U &= B^{1/2}A^{1/2}(A^{1/2}BA^{1/2})^{-1/2} \\ &= B^{1/2}A^{1/2}(A^{1/2}BA^{1/2})^{-1/2}A^{-1/2}A^{1/2} \\ &= B^{1/2}(AB)^{-1/2}A^{1/2} \\ &= B^{-1/2}B(B^{-1}A^{-1})^{1/2}A^{1/2} \\ &= B^{-1/2}(B\#A^{-1})A^{1/2}. \end{aligned} \tag{35}$$

So, the equation (34) can be written as

$$Z(t) = (1-t)A^{1/2} + t(A^{-1}\#B)A^{1/2}, \quad 0 \le t \le 1. \tag{36}$$

We have

$$Z(0) = A^{1/2},\ Z(1) = (A^{-1}\#B)A^{1/2} = B^{1/2}U, \tag{37}$$

and

$$Z'(t) = B^{1/2}U - A^{1/2} = (A^{-1}\#B - I)A^{1/2}, \ \ 0 \le t \le 1. \tag{38}$$

Note that

$$Z(t) = ((1-t)I + t(A^{-1}\#B))A^{1/2},$$

being a product of two positive definite matrices is in $GL(n)$. Being a straight line segment, it is a geodesic. From (26) and (38) we see that $Z'(0)$ is in the horizontal space $H_{A^{1/2}}$. So, by Theorem 4 $\gamma(t) = \pi(Z(t))$ is a geodesic in the space $\mathbb{P}(n)$ with respect to the Riemannian metric (32). From (37) we see that

$$\gamma(0) = \pi(Z(0)) = A, \quad \text{and}$$
$$\gamma(1) = \pi(Z(1)) = Z(1)Z(1)^* = B.$$

Thus $\gamma(t)$ is a geodesic joining A and B. An explicit expression for $\gamma(t)$ can be obtained by using (34) and (35). We have

$$\begin{aligned}\gamma(t) &= Z(t)Z(t)^* \\ &= (1-t)^2 A + t^2 B + t(1-t)\big[A^{1/2}U^*B^{1/2} + B^{1/2}UA^{1/2}\big] \\ &= (1-t)^2 A + t^2 B + t(1-t)\big[A(A^{-1}\#B) + (A^{-1}\#B)A\big] \\ &= (1-t)^2 A + t^2 B + t(1-t)\big[(AB)^{1/2} + (BA)^{1/2}\big]. \qquad (39)\end{aligned}$$

Theorem 4 tells us that the length L_γ of the geodesic in $\mathbb{P}(n)$ is equal to the length L_Z in $GL(n)$. The latter is the length of the straight line segment joining $A^{1/2}$ and $B^{1/2}U$. So, from Theorem 1 we have

$$L_\gamma = \|A^{1/2} - B^{1/2}U\|_2 = d(A,B).$$

We started with the distance $d(A,B)$ on $\mathbb{P}(n)$ and used Theorems 3 and 4 to show that this distance corresponds to a Riemannian metric given by (32). If, to begin with, we are provided with the metric (32) at each point A, then starting from it we can obtain the distance function $d(A,B)$.

At the beginning of this section we introduced the vertical and horizontal spaces at a point M of $GL(n)$. A curve $\widetilde{\gamma}$ in $GL(n)$ is called *horizontal* if for each t the tangent vector $\widetilde{\gamma}'(t)$ is in the horizontal space $\mathcal{H}_{\widetilde{\gamma}(t)}$. From (26) we see that $\widetilde{\gamma}$ is horizontal if and only if there exists a Hermitian matrix $H(t)$ such that

$$\widetilde{\gamma}'(t) = H(t)\widetilde{\gamma}(t), \quad 0 \le t \le 1. \qquad (40)$$

Let

$$\gamma(t) = \widetilde{\gamma}(t)\widetilde{\gamma}(t)^*. \qquad (41)$$

Then γ is a curve in $\mathbb{P}(n)$. Differentiating the relation (41) and then using (40) we see that

$$\gamma'(t) = \gamma(t)H(t) + H(t)\gamma(t). \qquad (42)$$

If γ is any curve in $\mathbb{P}(n)$, then a curve $\widetilde{\gamma}$ in $GL(n)$ is said to be a *horizontal lift* of γ if $\widetilde{\gamma}$ is horizontal and the relation (41) is satisfied. Every curve γ in $\mathbb{P}(n)$ has a unique horizontal lift $\widetilde{\gamma}$ that satisfies the condition $\widetilde{\gamma}(0)\ \widetilde{\gamma}(0)^* = \gamma(0)$. This can be seen as follows. Given $\gamma(t)$ let $H(t)$ be the unique solution of the Sylvester equation (42). From the smoothness of γ it follows that $H(t)$ is continous. Let M be a point of $GL(n)$ such that $MM^* = \gamma(0)$. The initial value problem $X'(t) = H(t)X(t)$, $X(0) = M$, has a unique solution. Call this $\widetilde{\gamma}(t)$. We have seen above that this curve is a horizontal lift of $\gamma(t)$.

The length of the curve γ is defined as

$$L_\gamma = \int_0^1 \langle\ \gamma'(t),\ \ \gamma'(t)\ \rangle_{\gamma(t)}^{1/2}\ dt.$$

If the inner product in the integrand is defined by (29) and $H(t)$ by (42), then this gives

$$L_\gamma = \int_0^1 (\mathrm{tr}\ H(t)\ \gamma(t)\ H(t))^{1/2}\ dt.$$

Using (40) and (41) we obtain from this

$$L_\gamma = \int_0^1 \langle\ \widetilde{\gamma}'(t),\ \ \widetilde{\gamma}'(t)\ \rangle^{1/2}\ dt = \int_0^1 \|\widetilde{\gamma}'(t)\|_2\ dt.$$

This is the length of the curve $\widetilde{\gamma}$ with respect to the Euclidean distance, and cannot be smaller than the straight line distance. So

$$L_\gamma \geq\ \|\widetilde{\gamma}(0)\ -\ \widetilde{\gamma}(1)\ \|_2.$$

If $\gamma(0) = A$ and $\gamma(1) = B$, then $\widetilde{\gamma}(0)$ and $\widetilde{\gamma}(1)$ are points in $\mathcal{F}(A)$ and $\mathcal{F}(B)$, respectively. So, by Theorem 1

$$L_\gamma \geq\ d(A,B).$$

Earlier we have seen a curve for which the two sides of this inequality are equal. Thus the metric (32) leads to the distance function $d(A,B)$ by a direct computation.

The material in this section is based on [5, 20, 33, 34]. Takatsu [33] also discusses the metric geometry of the spaces of psd matrices of rank k, $1 \leq k \leq n$. A very interesting research paper by K. Modin [27] dicusses the connections between optimal transport, geometry and matrix decompositions.

5. The Wasserstein mean

There is another standard metric on $\mathbb{P}(n)$ which has been extensively studied. In this the inner product on the tangent space $T_A\mathbb{P}(n) = \mathbb{H}(n)$ is given by

$$\langle Y, Z\rangle_A = \mathrm{tr}\, A^{-1}YA^{-1}Z, \tag{43}$$

and the associated distance function is

$$\delta(A,B) = \|\log\, A^{-1/2}BA^{-1/2}\|_2. \tag{44}$$

Any two points A, B of $\mathbb{P}(n)$ can be joined by a unique geodesic with respect to this metric, and a natural parametrisation for this geodesic is

$$A\#_t B = A^{1/2}(A^{-1/2}BA^{-1/2})^t A^{1/2}, \quad 0 \leq t \leq 1. \tag{45}$$

The geometric mean $A\#B$ defined in (8) is evidently the midpoint of this geodesic; i.e.,

$$A\#B = A\#_{1/2}B.$$

The metric (44) has lots of isometries. We have

$$\delta(XAX^*, XBX^*) = \delta(A,B) \text{ for all } X \in GL(n), \tag{46}$$

and

$$\delta(A^{-1}, B^{-1}) = \delta(A,B) \text{ for all } A, B. \tag{47}$$

This bestows upon the geometric mean $A\#B$ several interesting and useful properties, and the object is much used in operator theory, quantum mechanics, electrical networks, elasticity, image processing, etc. The collection [28] has several articles on the theory, computation, and applications of this mean and its multivariable version.

It is natural to ask what properties the " mean" with respect to the distance (1) might have. Let us adopt the notation $A \diamond_t B$ for the geodesic $\gamma(t)$ given in (39). The midpoint of this is

$$A \diamond B = \frac{1}{4}(A + B + (AB)^{1/2} + (BA)^{1/2}). \tag{48}$$

We call this the *Wasserstein mean* of A and B. The relations

$$(AB)^{1/2} = A(A^{-1}\#B) = A^{1/2}(A^{1/2}BA^{1/2})^{1/2}A^{-1/2}, \tag{49}$$

will be used in the following discussion.

For the Bures-Wasserstein distance (1) only a very restrictive version of (46) is true: we have $d(UAU^*, UBU^*) = d(A,B)$ provided U is unitary. The analogue of (47) is not valid for d. So the Wasserstein mean does not have many of the interesting properties of the mean $A\#B$. The following theorem is, therefore, surprising. Recall the operator version of the harmonic-geometric-arithmetic mean inequality. This says

$$\left(\frac{A^{-1} + B^{-1}}{2}\right)^{-1} \le A\#B \le \frac{A+B}{2}. \tag{50}$$

The second inequality in (50) can be extended as

$$A\#_t B \le (1-t)A + tB, \quad 0 \le t \le 1. \tag{51}$$

This has an analogue for the Wasserstein mean:

Theorem 6. *For all positive definite matrices A and B we have*

$$A \diamond_t B \le (1-t)A + tB, \quad 0 \le t \le 1. \tag{52}$$

Proof. Using the equations (39) and (49) we have

$$\begin{aligned} A \diamond_t B &= \gamma(t) \\ &= (1-t)^2 A + t^2 B \\ &\quad + t(1-t)\left[A^{1/2}(A^{1/2}BA^{1/2})^{1/2}A^{-1/2} + A^{-1/2}(A^{1/2}BA^{1/2})^{1/2}A^{1/2}\right] \\ &= A^{-1/2}\big[(1-t)^2A^2 + t^2 A^{1/2}BA^{1/2} \\ &\quad t(1-t)\{A(A^{1/2}BA^{1/2})^{1/2} + (A^{1/2}BA^{1/2})^{1/2}A\}\big]A^{-1/2} \\ &= A^{-1/2}\big[(1-t)A + t(A^{1/2}BA^{1/2})^{1/2}\big]^2 A^{-1/2}. \end{aligned}$$

The map $f(X) = X^2$ is matrix convex; i.e., for all Hermitian matrices X and Y we have

$$[(1-t)X + tY]^2 \leq (1-t)X^2 + tY^2, \quad 0 \leq t \leq 1.$$

Hence,

$$\begin{aligned} A \diamond_t B &\leq A^{-1/2}\left[(1-t)A^2 + tA^{1/2}BA^{1/2}\right]A^{-1/2} \\ &= (1-t)A + tB. \end{aligned}$$

This proves the inequality (52). ∎

Another instructive proof of Theorem 6 goes as follows. Using the inequality

$$0 \leq A^{-1/2}(A - (A^{1/2}BA^{1/2})^{1/2})^2 A^{-1/2},$$

and (49) we obtain

$$(AB)^{1/2} + (BA)^{1/2} \leq A + B. \tag{53}$$

So, from (48) we have

$$A \diamond B \leq \frac{1}{2}(A+B). \tag{54}$$

Since $A \diamond_t B = \gamma(t)$ is the natural parametrisation of the geodesic joining A and B, we have

$$(A \diamond_s B) \diamond_u (A \diamond_t B) = A \diamond_v B,$$

where $v = (1-u)s + ut$ for all s, t, u in $[0,1]$. Using this we can obtain from (54) the inequality (52) for all dyadic rational values of t. By continuity it holds for all $0 \leq t \leq 1$.

Theorem 6 may lead us to expect that the inequality

$$\left(\frac{A^{-1} + B^{-1}}{2}\right)^{-1} \leq A \diamond B, \tag{55}$$

might also be true. However, this is not always the case.

If A and B are two positive definite matrices such that $A \leq B$, then it follows from Theorem 6 that $A \diamond B \leq B$. However, it is not necessary that $A \leq A \diamond B$. If we choose

$$A = \begin{bmatrix} 1 & 1 \\ 1 & 2 \end{bmatrix}, \quad B = \begin{bmatrix} 3 & 1 \\ 1 & 2 \end{bmatrix},$$

then

$$A \diamond B \approx \begin{bmatrix} 1.8495 & 1.0449 \\ 1.0449 & 1.9857 \end{bmatrix},$$

and $A \nleq A \diamond B$.

This example also shows that $A \diamond B$ is not monotone with respect to the partial order $\leq$; i.e. if $A \leq A'$, then it is not necessary that $A \diamond B \leq A' \diamond B$.

6. The Wasserstein barycentre

Let $A_1,\ldots,A_m$ be elements of $\mathbb{P}(n)$ and let $w=(w_1,\ldots,w_m)$ be a weight vector ; i.e., $w_j>0$ and $\Sigma\, w_j=1$. Consider the minimisation problem

$$\min_{X>0}\ \sum_{j=1}^{m} w_j d^2(X,A_j). \tag{56}$$

This problem was first considered by Knott and Smith [22] as a multivariable generalisation of the work of Olkin and Pukelsheim discussed in Section 3 above. Agueh and Carlier [1] studied the general problem of determining the barycentre of several probability measures on $\mathbb{R}^n$. The special case of Gaussian measures is the problem (56). The general problem has been studied as a part of the the *multimarginal transport problem* or the m *-coupling problem.*

Theorem 6.1 of [1] says that the problem (56) has a unique solution. The proof of uniqueness in [1], that draws on the earlier discussion of the general case, relies on tools from nonsmooth analysis, convex duality and the theory of optimal transport. In the spirit of this paper we now provide another proof using simple ideas from matrix analysis.

The minimiser in (56) is called the *Wasserstein barycentre* of $A_1\ldots,A_m$ with weights $w_1,\ldots,w_m$. This is the positive definite matrix

$$\Omega(w;A_1,\ldots,A_m)=\underset{X>0}{\operatorname{argmin}}\ \sum_{j=1}^{m} w_j d^2(X,A_j). \tag{57}$$

Using the definition (1) we see that the objective function in (57) is $f(X)$, where

$$f(X)=\sum_{j=1}^{m}\ w_j\ \operatorname{tr}A_j+\sum_{j=1}^{m}\ w_j\ \operatorname{tr}(X-2(A_j^{1/2}XA_j^{1/2})^{1/2}). \tag{58}$$

This is a differentiable function on the convex cone $\mathbb{P}(n)$. We will calculate the derivative of f, and show that there is a point in $\mathbb{P}(n)$ at which is vanishes. This local minimum for f will be a (unique) global minimum if f is a (strictly) convex function. From (58) it is clear that to prove strict convexity of f it is enough to establish strict concavity of the function $h(X)=\operatorname{tr}\,X^{1/2}$. This is our next theorem.

Theorem 7. *The map* $h(X)=\operatorname{tr}\,X^{1/2}$ *from* $\mathbb{P}(n)$ *into* $(0,\infty)$ *is strictly concave; i.e., if* X *and* Y *are two distinct elemens of* $\mathbb{P}(n)$ *and* α,β *are positive numbers with* $\alpha+\beta=1$, *then*

$$h(\alpha X+\beta Y)>\alpha h(X)+\beta h(Y). \tag{59}$$

Proof. It is well-known that $X\longmapsto X^{1/2}$ is an operator concave function. See Chapter V of [6]. So, we have

$$(\alpha X+\beta Y)^{1/2}\geq \alpha X^{1/2}+\beta Y^{1/2},$$

and hence

$$\operatorname{tr}(\alpha X+\beta Y)^{1/2}\geq \alpha\ \operatorname{tr}X^{1/2}+\beta\ \operatorname{tr}\,Y^{1/2}.$$

We have to show that in this last inequality the two sides cannot be equal if $X \neq Y$. Suppose

$$\operatorname{tr}\left[(\alpha X+\beta Y)^{1/2}-(\alpha X^{1/2}+\beta Y^{1/2})\right]=0.$$

The matrix inside the square brackets is positive semidefinite. So, its trace can be zero only if

$$(\alpha X+\beta Y)^{1/2}=\alpha X^{1/2}+\beta Y^{1/2}.$$

Square both sides, and then use the relations $\alpha-\alpha^2=\beta-\beta^2=\alpha\beta$, to obtain

$$\alpha\beta(X+Y-X^{1/2}Y^{1/2}-Y^{1/2}X^{1/2})=0.$$

Since $\alpha\beta\neq 0$, , this gives

$$(X^{1/2}-Y^{1/2})^2=0,$$

and hence $X^{1/2}=Y^{1/2}$, and $X=Y$. ■

Now we show that f does have a minimum in $\mathbb{P}(n)$ by evaluating the derivative $Df(X)$ and equating that to zero. A convenient summary of facts about matrix differential calculus can be found in Chapter X of [6].

The nonlinear term in (58) is $g(X)=(A^{1/2}XA^{1/2})^{1/2}$. We evaluate $Dg(X)$ from first principles. The derivative of the function $\Psi(A)=A^2$ is the linear map $D\Psi(A)$ defined as $D\Psi(A)(X)=AX+XA$. The function $\varphi(A)=A^{1/2}$ on $\mathbb{P}(n)$ is the inverse of Ψ. Hence $D\varphi(A)=[D\Psi(\varphi(A))]^{-1}=\left[D\Psi(A^{1/2})\right]^{-1}$. Thus $D\varphi(A)$ is the inverse of the linear map $X\longmapsto A^{1/2}X+XA^{1/2}$. By well known facts about the Sylvester matrix equation (see [6] or [12]) this inverse is given by the formula

$$D\varphi(A)(X)=\int_0^\infty e^{-tA^{1/2}}X\,e^{-tA^{1/2}}\,dt.$$

Let $\lambda(X)=A^{1/2}XA^{1/2}$. Then g is the composite $\varphi\circ\lambda$. So, by the chain rule of differentiation,

$$\begin{aligned}
Dg(X)(Y) &= (D\varphi(\lambda(X))\circ D\lambda(X))(Y)\\
&= D\varphi(A^{1/2}XA^{1/2})(A^{1/2}\,Y\,A^{1/2})\\
&= \int_0^\infty e^{-t(A^{1/2}X\,A^{1/2})^{1/2}}(A^{1/2}YA^{1/2})\,e^{-t(A^{1/2}XA^{1/2})^{1/2}}\,dt.
\end{aligned}$$

Taking traces, and using the cyclicity of trace, we get

$$\begin{aligned}\text{tr } Dg(X)(Y) &= \int_0^\infty (\text{tr } A^{1/2} e^{-2t(A^{1/2}X\ A^{1/2})^{1/2}} A^{1/2} Y)\ dt \\ &= \text{tr } A^{1/2} (\int_0^\infty e^{-2t(A^{1/2}X\ A^{1/2})^{1/2}}\ dt) A^{1/2} Y.\end{aligned}$$

The last integral above is equal to $\frac{1}{2}(A^{1/2}XA^{1/2})^{-1/2}$. (Use the fact that $\int_0^\infty e^{-t\alpha}\ dt = \frac{1}{\alpha}$ for every $\alpha > 0$.) Hence

$$\begin{aligned}\text{tr}Dg(X)(Y) &= \frac{1}{2}\text{tr}A^{1/2} \left(A^{-1/2}X^{-1}A^{-1/2}\right)^{1/2} A^{1/2}\ Y \\ &= \frac{1}{2}(A\#X^{-1})Y.\end{aligned}$$

So, from (58) we see that

$$\begin{aligned}Df(X)(Y) &= \sum_{j=1}^m w_j \text{tr}(Y - (A_j\#X^{-1})Y) \\ &= \text{tr}\left(I - \sum_{j=1}^m w_j \left(A_j\#X^{-1}\right)\right) Y.\end{aligned}$$

Thus $Df(X) = 0$ if and only if

$$I = \sum_{j=1}^m w_j (A_j\#X^{-1}). \tag{60}$$

This is equivalent to saying

$$X = \sum_{j=1}^m w_j (X^{1/2}A_jX^{1/2})^{1/2}. \tag{61}$$

Finally, we show that there exists a point X in $\mathbb{P}(n)$ that satisfies the equation (61). Indeed, if $\alpha I \le A_j \le \beta I$, for all $1 \le j \le m$, then this X belongs to the compact convex set $\mathcal{K} = \{X \in \mathbb{P}(n) : \alpha I \le X \le \beta I\}$. To see this consider the function

$$F(X) = \sum_{j=1}^m w_j (X^{1/2}A_jX^{1/2})^{1/2}.$$

Then note that $(X^{1/2}A_jX^{1/2})^{1/2} \le (\beta X)^{1/2} \le \beta I$ for all $X \in \mathcal{K}$. By the same reasoning $(X^{1/2}A_jX^{1/2})^{1/2} \ge \alpha I$ for all $X \in \mathcal{K}$. This shows that F maps $\mathcal{K}$ into itself. By Brouwer's fixed point theorem there exists a point X in $\mathcal{K}$ such that $F(X) = X$. This X is a solution of the equation (61).

We have proved the following theorem first obtained in [1], building upon the earlier work in [22] and [31]

Theorem 8. *The minimisation problem* (57) *has a unique solution which is also the solution of the nonlinear matrix equation* (61).

We do not know how to obtain the solution of (61) in an explicit form. In the special case $m = 2$, with $A_1 = A$, $A_2 = B$, and $(w_1, w_2) = (1-t, t)$, the equation (60) reduces to

$$I = (1-t)(A\#X^{-1}) + t(B\#X^{-1}). \tag{62}$$

The solution to this equation is

$$X = \operatorname{argmin}\left[(1-t)d^2(X,A) + td^2(X,B)\right].$$

By the definition of geodesics with respect to the metric d, such an X is the unique point on the geodesic segment joining A and B at distance $td(A,B)$ from A. In other words

$$X = A \diamond_t B = (1-t)^2 A + t^2 B + t(1-t)\left[(AB)^{1/2} + (BA)^{1/2}\right]. \tag{63}$$

The equation (61) can be used to obtain some important order properties of the Wasserstein barycentre. The next theorem is a multivariable analogue of Theorem 6.

Theorem 9. *Let* $A_1, \ldots, A_m$ *be positive definite matrices and let* $w = (w_1, \ldots, w_m)$ *be a weight vector. Then*

$$\Omega(w; A_1, \ldots, A_m) \le \sum_{j=1}^m w_j A_j. \tag{64}$$

Proof. The matrix $\Omega = \Omega(w; A_1, \ldots, A_m)$ obeys the relation

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

Square both sides and then use the fact that the function $f(A) = A^2$ is matrix convex. This gives

$$\Omega^2 \le \sum_{j=1}^m w_j \Omega^{1/2} A_j \Omega^{1/2} = \Omega^{1/2}\left(\sum_{j=1}^m w_j A_j\right)\Omega^{1/2}.$$

The inequality (64) follows from this. ∎

Theorem 9 is much stronger than the known inequality $\operatorname{tr}\Omega \le \operatorname{tr}\sum w_j A_j$, which has been proved in [3]. (See the last inequality in Theorem 4.2 there.)

7. The m-Coupling Problem

We explain briefly how the Wasserstein barycentre is useful in solving the several variable version of the problem considered in Section 3.

Let $x_1, \ldots, x_m$ be random vectors in $\mathbb{C}^n$, each having zero mean, and with covariance matrices $A_1, \ldots, A_m$. We are asked to find a tuple $(x_1, \ldots, x_m)$ that solves the minimisation problem

$$\min\ E \sum_{1<j} \|x_i - x_j\|^2. \tag{65}$$

This is the same problem as the one of maximising $E\|\sum x_j\|^2$. A little more generally, we consider the problem

$$\max\ E\| \sum_{j=1}^{m} w_j\ x_j \|^2, \tag{66}$$

where $w_1, \ldots, w_m$ are given weights.

Let $\Omega = \Omega(w; A_1, \ldots, A_m)$ and let $R_j = \Omega^{-1} \# A_j$, $1 \leq j \leq m$. Let $z = \sum w_j x_j$. Then $\langle z, z \rangle = \sum w_j \langle z, x_j \rangle$. If T is any positive definite matrix, and x, y any two vectors, then using the Schwarz inequality and the arithmetic-geometric mean inequality, we see that

$$\begin{aligned} \langle x, y \rangle &= \langle T^{1/2}x,\, T^{-1/2}y \rangle\ \leq \|T^{1/2}x\|\, \|T^{-1/2}y\| \\ &= \langle x, Tx \rangle^{1/2}\ \langle y, T^{-1}y \rangle^{1/2} \\ &\leq \frac{1}{2}\left[\langle x, Tx \rangle\ +\ \langle y, T^{-1}y \rangle \right]. \end{aligned} \tag{67}$$

Hence,

$$\langle z, z \rangle\ \leq \frac{1}{2}\left[\sum_{j=1}^{m}\ w_j\ \langle z, R_j z \rangle + \sum_{j=1}^{m}\ w_j\ \langle x_j, R_j^{-1} x_j \rangle \right].$$

From (60) we know that $\sum_{j=1}^{m}\ w_j\, R_j = I$. So, the inequality above yields

$$\langle z, z \rangle\ \leq \sum_{j=1}^{m}\ w_j\ \langle x_j, R_j^{-1} x_j \rangle.$$

Thus

$$E\|z\|^2\ \leq\ \sum_{j=1}^{m}\ w_j\ E\langle x_j, R_j^{-1} x_j \rangle.$$

Since x_j has covariance matrix A_j, this gives

$$\begin{aligned} E\|z\|^2 &\le \sum_{j=1}^m w_j \text{ tr } R_j^{-1}A_j \\ &= \sum_{j=1}^m w_j \text{ tr } A_j^{1/2}R_j^{-1}A_j^{1/2} \\ &= \sum_{j=1}^m w_j \text{ tr } A_j^{1/2}(\Omega\#A_j^{-1})A_j^{1/2} \\ &= \sum_{j=1}^m w_j \text{ tr } (A_j^{1/2}\Omega A_j^{1/2})\#I \\ &= \text{tr} \sum_{j=1}^m w_j \ (A_j^{1/2}\Omega A_j^{1/2})^{1/2} \\ &= \text{tr } \Omega. \end{aligned} \tag{68}$$

Note that both the inequalities in (67) are equalties if $y = Tx$. Hence, there is equality at the first step in (68) if $z = R_j x_j$ for $1 \le j \le m$. This can be achieved by choosing x_1 arbitrarily and then putting $x_j = R_j R_1^{-1} x_1$ for $2 \le j \le m$.

To sum up, we have shown that the problem (66) has the solution

$$\max E\|\sum_{j=1}^m w_j \ x_j\|^2 = \text{tr } \Omega(w; A_1, \ldots, A_m). \tag{69}$$

The maximum is attained at every m-tuple

$$(x_1, \ R_2R_1^{-1} \ x_1, \ R_3R_1^{-1} \ x_1, \ldots, R_mR_1^{-1} \ x_1), \tag{70}$$

where x_1 is chosen arbitrarily subject to the given conditions that it has mean 0 and covariance matrix A_1. Note that, then we have

$$\begin{aligned} \sum_{j=1}^m w_j x_j &= w_1 x_1 + \sum_{j=2}^m w_j R_j R_1^{-1} \ x_1 \\ &= \sum_{j=1}^m w_j R_j R_1^{-1} \ x_1 = R_1^{-1} \ x_1, \end{aligned} \tag{71}$$

the last equality being a consequence of the fact that $\sum\limits_{j=1}^m w_j R_j = I$. The maps $R_j R_1^{-1}$ are said to provide an optimal coupling between $x_1, \ldots, x_m$ that occur as a solution of (66).

Many of the ideas presented in Sections 6 and 7 go back to the paper of Knott and Smith [22]. Among other things, the matrix equation (61), that a solution to the minimisation problem (57) must satisfy, is derived there. However, questions about the existence and uniqueness of solutions of this equation are not settled in this paper. The existence was established by Ruschendorf and Uckelmann in [31], and the uniqueness by Agueh and Carlier in [1]. The elegant argument using Brouwer's fixed point theorem to establish the existence of a solution occurs in [1], and we have adopted it verbatim. Our proof of uniqueness is different, and uses ideas more familiar in matrix analysis. We must add that the problem studied in [1] is the more general problem of the barycentre of measures. The matrix case that we are discussing corresponds to the special Gaussian measures.

8. Computing the Barycentre

Whereas for two matrices A and B their barycentre is given by an explicit formula (51), no such formula is known in the case of three or more matrices. We know only that Ω is the unique solution of the equation (60), or equivalently of (61). The latter suggests that it may be possible to compute Ω by a fixed point iteration. Such an iteration has been developed in a very interesting recent paper [3]. In this section we explain the main ideas of this paper, restricting ourselves to matrix analytic techniques, and simplifying some proofs.

Throughout this section $A_1, \ldots, A_m$ are given positive definite matrices and $w = (w_1, \ldots, w_m)$ a given set of weights. For each $A \in \mathbb{P}(n)$ let

$$H_j(A) = A^{-1} \# A_j \ , \qquad 1 \leq j \leq m, \tag{72}$$

$$H(A) = \sum_{j=1}^{m} w_j \ H_j(A), \tag{73}$$

$$K(A) = A^{-1/2} \left(\sum_{j=1}^{m} w_j \ (A^{1/2} A_j A^{1/2})^{1/2} \right)^2 A^{-1/2}. \tag{74}$$

We note that

$$K(A) = H(A) A H(A). \tag{75}$$

Also, note that

$$\begin{aligned} A^{-1} \# K(A) &= A^{-1/2} \left(A^{1/2} \ K(A) A^{1/2} \right)^{1/2} A^{-1/2} \\ &= A^{-1/2} \left(\sum_{j=1}^{m} w_j \ (A^{1/2} A_j A^{1/2})^{1/2} \right) A^{-1/2} \\ &= \sum w_j \ H_j(A) \ = H(A). \end{aligned} \tag{76}$$

Equations (72) and (76) say that $H_j(A)$ and $H(A)$ are the optimal transport maps from A to A_j and to $K(A)$, respectively. We define the *variance* of A as

$$V(A) = \sum_{j=1}^{m} w_j \, d^2(A, A_j). \tag{77}$$

The following variance inequality is a rephrasing in our context of Proposition 3.3 in [3].

Theorem 10. *For every positive definite matrix A we have*

$$V(A) \geq V(K(A)) \ + \ d^2(A, K(A)). \tag{78}$$

Proof. Let $y_1, \ldots, y_m$ be vectors in $\mathbb{C}^n$ and let $\overline{y} = \sum_{j=1}^{m} w_j y_j$ be their weighted arithmetic mean. Then for every $x \in \mathbb{C}^n$ we have

$$\sum_{j=1}^{m} w_j \ \|x - y_j\|^2 \ = \ \sum_{j=1}^{m} w_j \ \|\overline{y} - y_j\|^2 + \|x - \overline{y}\|^2. \tag{79}$$

(This is the variance equality in Euclidean space that (78) mimics. The Euclidean distance is replaced by the metric d, the points y_j by the matrices A_j, the mean $\overline{y}$ by $K(A)$, and we have an inequality in place of equality.)

Choose a vector x in $\mathbb{C}^n$ with mean 0 and covariance matrix A. For $1 \leq j \leq m$, let $y_j = H_j(A)x$, we have from the results in Section 3

$$d^2(A, A_j) \ = \ E \ \|x - H_j(A)x\|^2 \ = \ E \ \|x - y_j\|^2.$$

Hence,

$$V(A) = \sum_{j=1}^{m} w_j \ E\|x - y_j\|^2. \tag{80}$$

Similarly, since H(A) is the optimal transport map from A to $K(A)$, we have

$$d^2(A, K(A)) \ = \ E \ \|x - H(A)x\|^2.$$

But $H(A)x \ = \ \sum w_j \, H_j(A)x \ = \ \sum w_j y_j \ = \overline{y}$. So,

$$d^2(A, K(A)) \ = \ E \ \|x - \overline{y}\|^2. \tag{81}$$

Since $H(A)$ is the transport map from A to $K(A)$ and x has covariance matrix A, it follows that $\overline{y}$ has $K(A)$ as its covariance matrix. Hence

$$E \ \|\overline{y} - y_j\|^2 \ \geq d^2(K(A), A_j), \quad 1 \leq j \leq m. \tag{82}$$

The relations (79)-(82) put together lead to the inequality (78). ∎

Remark. Using the definition of the variance $V(A)$ and of the metric $d(A, B)$ it can be seen that the inequality (78) is equivalent to the trace inequality

$$\sum_{j=1}^{m} w_j \ \mathrm{tr} \ (A_j^{1/2} \ K(A) \ A_j^{1/2})^{1/2} \ \geq \ \mathrm{tr} \ K(A). \tag{83}$$

It might be very difficult to prove this using the usual matrix analysis arguments. The very special case $A = I$ of (83) says

$$\mathrm{tr} \ \sum \ w_j \left(A_j^{1/2} \left(\sum w_j A_j^{1/2} \right)^2 \ A_j^{1/2} \right)^{1/2} \ \geq \ \mathrm{tr} \left(\sum \ w_j A_j^{1/2} \right)^2. \tag{84}$$

From the inequality (IX.11) on page 258 of [6] we have

$$\begin{aligned}\operatorname{tr}\left(A_j^{1/2}\left(\sum\ w_j\,A_j^{1/2}\right)^2 A_j^{1/2}\right)^{1/2} &\geq \operatorname{tr}\,A_j^{1/4}\left(\sum\ w_jA_j^{1/2}\right)A_j^{1/4}.\\ &= \operatorname{tr}\,A_j^{1/2}\left(\sum\ w_jA_j^{1/2}\right).\end{aligned}$$

The inequality (84) follows from this. So, even the special case $A = I$ of (83) needs rather intricate arguments. Results proved in the context of optimal transport could thus add to the tools used in deriving matrix inequalities.

The next theorem is the main result (Theorem 4.2) of [3]. Some steps in the proof have been simplified.

Theorem 11. *Let* S_o *be any positive definite matrix and for* $n \geq 0$ *define* $S_{n+1} = K(S_n)$, *where* K *is the map defined in* (74). *Then*

$$\begin{aligned}&\text{(i)} \qquad \lim_{n\to\infty}\ S_n = \Omega\\ &\text{(ii)} \qquad \operatorname{tr} S_n \leq \operatorname{tr} S_{n+1} \leq \operatorname{tr}\Omega \text{ for all } n \geq 1.\end{aligned}$$

Proof. By definition

$$S_{n+1} = S_n^{-1/2}\left(\sum_{j=1}^m w_j\left(S_n^{1/2}A_j\ S_n^{1/2}\right)^{1/2}\right)^2 S_n^{-1/2}.$$

The square function is matrix convex. Hence,

$$\begin{aligned}S_{n+1} &\leq S_n^{-1/2}\left(\sum_{j=1}^m w_j\,S_n^{1/2}A_j\,S_n^{1/2}\right)S_n^{-1/2}.\\ &= \sum_{j=1}^m w_j\,A_j.\end{aligned}$$

Thus the sequence $\{S_n\}$ is a bounded sequence in $\mathbb{P}(n)$. Hence it has a subsequence converging to a limit S. By the variance inequality (78), $V(S_n) \geq V(S_{n+1})$ for all n. So $\{V(S_n)\}$ is a decreasing sequence of positive numbers. Hence it converges. We must have $\lim V(S_n) = V(S)$. Since K is a continuous function, this implies $\lim V(K(S_n)) = V(K(S))$. But $K(S_n) = S_{n+1}$. So, $V(K(S)) = V(S)$. Hence, using the variance inequality (78), we have $d^2(S, K(S)) = 0$. This means $S = K(S)$. From the definition of $K(S)$ in (74), this is possible if and only if $S = \Omega(w;\, A_1, \ldots, A_m)$. This proves part (i).

By the definition of $H(A)$ in (73) we have for every A

$$A^{1/2}\ H(A)A^{1/2} \ = \ \sum_{j=1}^{m} w_j\ A^{1/2}\ (A^{-1}\#A_j)A^{1/2} \ = \ \sum_{j=1}^{m} w_j\ (A^{1/2}\ A_j\, A^{1/2})^{1/2},$$

and hence,

$$\sum_{j=1}^{m} w_j\ d^2(A, A_j) \ = \ \text{tr}A \ + \ \sum_{j=1}^{m} w_j \text{tr}\, A_j \ - \ 2\text{tr}A^{1/2}\, H(A)A^{1/2}.$$

From this we can see that

$$V(S_n) - V(S_{n+1}) = \text{tr}\ S_n - \text{tr}\ S_{n+1} \ - \ 2\text{tr}\ S_n^{1/2}\ H(S_n)S_n^{1/2} \ + \ 2\text{tr}\ S_{n+1}^{1/2}\ H(S_{n+1})S_{n+1}^{1/2}, \qquad (85)$$

and

$$d^2(S_n, S_{n+1}) = \text{tr}\ S_n + \text{tr}\ S_{n+1} - 2\text{tr}\ S_n^{1/2}\ H(S_n)S_n^{1/2}. \qquad (86)$$

The variance inequality (78) together with these two relations gives

$$\text{tr}\ S_{n+1} \ \leq \ \text{tr}\ S_{n+1}^{1/2}\ H(S_{n+1}^{1/2})\ S_{n+1}^{1/2}. \qquad (87)$$

From (86) and (87) we obtain

$$\begin{aligned} 0 \ &\leq \ d^2\ (S_{n+1}, S_{n+2}) \\ &= \ \text{tr}\ S_{n+1} + \text{tr}\ S_{n+2} \ - \ 2\text{tr}\ S_{n+1}^{1/2} H(S_{n+1})\ S_{n+1}^{1/2} \\ &\leq \ \text{tr}\ S_{n+2} \ - \ \text{tr}\ S_{n+1}^{1/2}\ H(S_{n+1})\ S_{n+1}^{1/2}. \end{aligned} \qquad (88)$$

Finally, from (87) and (88) we see that

$$\text{tr}\ S_{n+1} \leq \text{tr}\ S_{n+2}.$$

That proves (ii). ∎

9. REMARKS

The geometric mean $A\#B$ has played a crucial role at several places in this paper. This is the midpoint of the geodesic joining A and B with the Riemannian metric δ defined in (46) and (47). The barycentre of m matrices $A_1, \ldots, A_m$ with weights $w_1, \ldots, w_m$ with respect to this metric is defined as

$$G(w; A_1, \ldots, A_m) \ = \ \underset{X>0}{\text{argmin}} \sum_{j=1}^{m} w_j \delta^2(X, A_j).$$

This has been an object of intense study in recent years. See [7] [8] [9] [10] [19] [23] [24] [25] [26]. A natural question, from the perspective of matrix analysis, would be to find comparisons between the two means G and Ω.

Another classical family of means, called the *power means* is defined as

$$Q_t(w; A_1, \ldots, A_m) = \left(\sum_{j=1}^{m} w_j A_j^t \right)^{1/t}, \ \ t > 0.$$

These play an important role in analysis. When $t = \frac{1}{2}$, we have

$$Q_{1/2}(w; A_1, \ldots, A_m) = \left(\sum_{j=1}^{m} w_j A_j^{1/2} \right)^2 .$$

In the special case when $A_1, \ldots, A_m$ are commuting matrices, the Wasserstein mean Ω and the mean $Q_{1/2}$ coincide. If we let

$$\rho(A, B) \;=\; \|A^{1/2} - B^{1/2}\|_2 \;=\; \left[\mathrm{tr}A + \mathrm{tr}B - 2\mathrm{tr}A^{1/2}B^{1/2} \right]^{1/2},$$

then

$$Q_{1/2}(w; A_1, \ldots, A_m) = \operatorname*{argmin}_{X>0} \sum_{j=1}^{m} w_j \rho^2(X, A_j)$$

It is natural to ask for comparisons between the means $Q_{1/2}$ and Ω.

These problems are studied in our forthcoming papers.

Acknowledgement. The work of R. Bhatia is supported by a J. C. Bose National Fellowship and of T. Jain by a SERB Women Excellence Award. The work of Y. Lim is supported by the National Research Foundation of Korea (NRF) grant founded by the Korea government (MEST) (No. 2015R1A3A2031159).

References

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

[2] P. M. Alberti, *A note on the transition probability over* C^* *-algebras*, Lett. Math. Phys. 7 (1983), 25-32.

[3] P.C. Alvarez-Esteban, E.del Barrio, J.A. Cuesta-Albertos and C. Matran, *A fixed point approach to Barycenters in Wasserstein spaces*, J. Math.Anal. Appl. 441(2016 744-762.

[4] F. Barbaresco, *Information geometry of covariance matrix: Cartan-Siegel homogeneous bounded domains, Mostow/Berger fibration and Frećhet median*, in Matrix Information Geometry, eds. F. Nielsen and R. Bhatia, Springer, (2013), 199-256.

[5] I. Bengtsson and K. Zyczkowski, *Geometry of Quantum States: An Introduction to Quantum Entanglement*, Cambridge University Press, 2006.

[6] R. Bhatia, *Matrix Analysis,* Springer, 1997.

[7] R. Bhatia, *Positive Definite Matrices,* Princeton University Press, 2007.

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

[9] R. Bhatia and J. Holbrook, *Riemannian geometry and matrix geometric means*, Linear Algebra Appl. 413 (2006), 594-618.

[10] R. Bhatia and R. L. Karandikar, *Monotonicity of the matrix geometric mean*, Math. Ann. 353 (2012), 1453-1467.

[11] R. Bhatia and F. Kittaneh, *On the singular values of a product of operators*, SIAM J. Matrix Anal. Appl. 11 (1990), 272-277.

[12] R. Bhatia and P. Rosenthal, *How and why to solve the operator equation* $AX - XB = Y$, Bull. London Math. Soc. 29 (1997), 1-21.

[13] C. Davis and W. M. Kahan, *The rotation of eigenvectors by a perturbation III*, SIAM J. Numer. Anal. 7 (1970), 1-46.

[14] D. Dowson and B. Landau, *The Fréchet distance between multivariate normal distributions*, J. Math. Anal. 12 (1982), 450-455.

[15] S. Gallot, D. Hulin and J. Lafontaine, *Riemannian Geometry*, Springer, 2004.

[16] J. C. Gower and G. B. Dijksterhuis, *Procrustes Problems*, Oxford University Press 2004.

[17] M. Hayashi, *Quantum Information*, Springer, 2006.

[18] N. Higham, *Functions of Matrices: Theory and Computation*, SIAM 2008.

[19] J. Holbrook, *No dice: a deterministic approach to the Cartan centroid*, J. Ramanujan Math. Soc. 27 (2012), 509-521.

[20] A. Jencová, *Geodesic distances on density matrices*, J. Math. Phys. 45 (2004), 1787-1794.

[21] F. Kittaneh, *On the convexity of the Heinz means*, Integr. Eqn. Oper. Theory 68 (2010), 519-527.

[22] M. Knott and C. Smith, *On the optimal mapping of distributions*, J. Optim. Theory Appl. 43 (1984), 39-49.

[23] J. Lawson and Y. Lim, *Monotonic properties of the least squares mean*, Math. Ann. 351 (2011), 267-279.

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

[25] Y. Lim and M. Palfia, *Weighted deterministic walks for the least squares mean on Hadamard spaces*, Bull. London Math. Soc. 46 (2014), 561-570.

[26] M. Moakher, *A differential geometric approach to the geometric mean of symmetric positive-definite matrices*, SIAM J. Matrix Anal. Appl. 26 (2005), 735-747.

[27] K. Modin, *Geometry of matrix decompositions seen through optimal transport and information geometry*, J. Geom. Mech. 9 (2017), 335-390.

[28] F. Nielsen and R. Bhatia, eds., *Matrix Information Geometry,* Springer, 2013.

[29] I. Olkin and F. Pukelsheim, *The distance between two random vectors with given dispersion matrices*, Linear Algebra Appl. 48 (1982), 257-263.

[30] W. Pusz and S. L. Woronowicz, *Functional calculus for sesquilinear forms and the purification map*, Rep. Math. Phys. 8 (1975), 159-170.

[31] L. Ruschendorf and L. Uckelmann, *On the n -coupling problem*, J. Multivariate Anal. 81(2002) 242-258

[32] G. W. Stewart and J. G. Sun, *Matrix Perturbation Theory*, Academic Press, 1990.

[33] A. Takatsu, *Wasserstein geometry of Gaussian measures*, Osaka J. Math. 48 (2011), 1005-1026.

[34] A. Uhlmann, *Density operators as an arena for differential geometry*, Rep. Math. Phys. 33 (1993), 255-263.

[35] A. Uhlmann, *Transition probability (fidelity) and its relatives*, Found. Phys. 41 (2011), 288-298.

[36] C. Villani, *Optimal Transport: Old and New*, Springer, 2008.

December 4, 2017

Ashoka University, Sonepat,, Haryana, 131029, India
E-mail address: `rajendra.bhatia@ashoka.edu.in`

Indian Statistical Institute, New Delhi 110016, India
E-mail address: `tanvi@isid.ac.in`

Department of Mathematics, Sungkyunkwan University, Suwon 440-746, Korea
E-mail address: `ylim@skku.edu`