Jensen Divergence Based SPD Matrix Means and Applications

Frank Nielsen[1], Meizhu Liu[2], Xiaojing Ye[3] and Baba C. Vemuri[4]
[1] *Sony Computer Science Laboratories, Inc., Tokyo, Japan.*
[2] *Siemens Corporate Research, Princeton, NJ, USA.*
[3] *School of Mathematics, Georgia Tech, Atlanta, GA, USA.*
[4] *Center for Vision Graphics and Medical Imaging, UF, Gainesville, FL, USA.*
E-mails:Frank.Nielsen@acm.org, meizhu.liu@siemens.com, xiaojing.ye@math.gatech.edu, vemuri@cise.ufl.edu

Abstract

Finding mean of matrices becomes increasingly important in modern signal processing problems that involve matrix-valued images. In this paper, we define the mean for a set of symmetric positive definite (SPD) matrices based on information-theoretic divergences as the unique minimizer of the averaged divergences, and compare it with the means computed using the Riemannian and Log-Euclidean metrics. For the class of divergences induced by the convexity gap of a matrix functional, we present a fast iterative concave-convex optimization scheme with guaranteed convergence to efficiently approximate those divergence-based means.

1. Introduction

A recent trend in image processing is to consider *matrix-valued* images, where each pixel of the image is represented as a matrix of coefficients instead of a traditional intensity value. Typical applications include diffusion magnetic resonance image analysis [15], radar signal processing [16], elasticity tensors [7] in mechanical engineering, and structure tensors [3, 14, 13] in computer vision.

Due to the use of matrix based images, the conventional intensity-based image processing toolbox (e.g., inpainting, interpolation, segmentation etc.) needs to be extended to matrix-valued images. In this paper, we consider calculating the mean of matrices that is required for example in interpolation and clustering.

The mean of matrices is in general defined as follows: Given a collection of SPD matrices $\{M_1, ..., M_n\} \subset \mathrm{Sym}^\star_+(d)$, where $\mathrm{Sym}^\star_+(d)$ represents the set of $d \times d$ SPD matrices. The mean $\bar{M}$ is defined as

$$\bar{M} = \arg \min_{M \in \mathrm{Sym}^\star_+(d)} \frac{1}{n} \sum_{i=1}^{n} D(M_i, M)^2, \qquad (1)$$

where D is a distance function. Different distance functions give different means. For instance, if D is the Fröbenius norm induced distance, i.e., $D(P,Q) = \|P - Q\|_F^2$, then $\bar{M}$ becomes the arithmetic matrix mean, and $\bar{M} = \frac{1}{n}\sum_{i=1}^n M_i$. However, the arithmetic matrix mean is not robust to outliers, and it may have a determinant larger than the input which is physically not plausible in many applications [1]. The Log-Euclidean (LE) distance is defined as $D(P,Q) = \|\log Q - \log P\|_F$, where $\log M$ is the principal logarithm of matrix M. In [1], Arsigny et al. showed that the *LE mean* inherits a vector space structure, and has a closed-form $\bar{M}_{\mathrm{LE}} = \exp[(\sum_i \log M_i)/n]$. The Riemannian distance is defined as $D(P,Q) = [\mathrm{tr}(\log^2(P^{-1}Q))]^{1/2}$ and the mean is shown to be the *unique* matrix $\bar{M}_R$ satisfying $\sum_{i=1}^n \log(M_i^{-1}\bar{M}_R) = 0$, which has a closed-form solution when $n = 2$. For $n > 2$, Fiori et al. proposed an optimization scheme to approximate the mean [8].

In [5], Ando et al. summarized ten properties for a "good" matrix mean. Bathia and Holbrook [2] investigated properties of Riemannian matrix means. Bini and Iannazzo [4] recently proposed another geometric matrix mean definition that satisfies most but not all of the ten Ando-Li-Mathias properties.

In this work, we study the SPD mean with respect to a *non-metric* distance function, called a *divergence*. A divergence may not be symmetric nor satisfy the triangle inequality as regular metrics.

2. Divergences from Jensen convexity gaps

Let $(PQ)_\lambda$ denote the linear interpolant $(1-\lambda)P + \lambda Q$ for $\lambda \in (0,1)$. From the (open cone) convexity of

the domain of Sym_+^*, it follows that

$$\forall P, Q \in \mathrm{Sym}_+^*, (PQ)_\lambda \in \mathrm{Sym}_+^*. \quad (2)$$

We build a *family* of skewed divergences from a strictly convex generator $F : \mathrm{Sym}_+^* \to \mathbb{R}^+$ as follows:

$$J_F^{(\alpha)}(P, Q) = (F(P)F(Q))_\alpha - F((PQ)_\alpha), \quad (3)$$

for $0 < \alpha < 1$. $J_F^{(\alpha)} \geq 0$ and $J_F^{(\alpha)} = 0$ iff $P = Q$.

Common convex matrix generators are

- $F(X) = \mathrm{tr}(X^T X)$ (quadratic matrix entropy),
- $F(X) = -\log\det X$ (matrix Burg entropy),
- $F(X) = \mathrm{tr}(X \log X - X)$ (von Neumann entropy).

In particular, the symmetric Burbea-Rao divergence [6] is obtained by choosing $\alpha = \frac{1}{2}$, i.e.,

$$\mathrm{BR}_F(P, Q) = \frac{F(P) + F(Q)}{2} - F\left(\frac{P+Q}{2}\right) \geq 0.$$

Choosing $F(X) = \mathrm{tr}(X \log X - X)$, we get the *Jensen-von Neumann* divergence, the matrix counterpart of the celebrated Jensen-Shannon divergence. An interesting property is that asymptotic skew Jensen divergences are equivalent to Bregman divergences [12]:

$$B_F(P, Q) = \lim_{\alpha \to 0} \frac{1}{\alpha} J_F^{(\alpha)}(P, Q),$$
$$B_F(Q, P) = \lim_{\alpha \to 1} \frac{1}{1-\alpha} J_F^{(\alpha)}(P, Q), \text{ with}$$
$$B_F(P, Q) = F(P) - F(Q) - \langle P - Q, \nabla F(Q) \rangle,$$

where $\langle X, Y \rangle = \mathrm{tr}(XY)$ is the matrix inner product. The von Neuman divergence

$$D_{\mathrm{vN}}(P, Q) = \mathrm{tr}(P(\log P - \log Q) - P + Q) \quad (4)$$

belongs to a broader parametric family of matrix divergences:

$$D_\alpha(P, Q) = \frac{4}{1-\alpha^2} \mathrm{tr}\left(\frac{1-\alpha}{2}P + \frac{1+\alpha}{2}Q - P^{\frac{1-\alpha}{2}} Q^{\frac{1+\alpha}{2}}\right),$$

with $D_{\mathrm{vN}}(P, Q) = \lim_{\alpha \to 1} D_\alpha(P, Q)$, and $D_{\mathrm{vN}}(P, Q) = \lim_{\alpha \to -1} D_\alpha(P, Q)$.

3. Concave-convex minimization for Jensen-based matrix means

By definition, the divergence-based (right-sided) means on a set of SPD matrices $\{M_1, ..., M_n\}$, are obtained by minimizing the average distortion measure:

$$l(X) = \frac{1}{n} \sum_{i=1}^n J_F^{(\alpha)}(M_i, X). \quad (5)$$

Note the left-sided mean can be calculated as a right sided-mean for parameter $\alpha' = 1 - \alpha$. The matrix mean is solved according to

$$\bar{M} = \arg_{X \in \mathrm{Sym}_+^*} \min l(X). \quad (6)$$

Removing the constant terms independent of X in $l(X)$, we get an *equivalent* optimization problem,

$$l'(X) = \alpha F(X) - \sum_{i=1}^n F((1-\alpha)M_i + \alpha X). \quad (7)$$

This loss function $l'(X) = A(X) + B(X)$ is a sum of a convex function $A(X) = \alpha F(X)$ plus a concave function $B(X) = -\sum_{i=1}^n F((1-\alpha)M_i + \alpha X)$. It follows that we can apply the concave-convex procedure [17] to get the following iterative scheme: We start from an initial estimate C_0 of the mean (say, the arithmetic mean $C_0 = \frac{1}{n}\sum_{i=1}^n M_i$), and update iteratively the current mean C_t using the concave-convex procedure (CCCP) optimization step [17] (that does not require to set up a learning rate):

$$\nabla A(C_{t+1}) = -\nabla B(C_t), \quad (8)$$

and get

$$C_{t+1} = (\nabla F)^{-1}\left(\sum_{i=1}^n \nabla F((1-\alpha)M_i + \alpha C_t)\right).$$

This iterative scheme is *guaranteed* to converge to a minimizer [17], and avoids to tune a learning step parameter as it is customary in gradient descent methods.

3.1. Matrix α-log-det divergence

When the convex generator is $F(X) = -\log\det X$, it gives us the α-log-det divergence, for $\alpha \in (-1, 1)$:

$$J_{LD}^{(\alpha)}(X, Y) = \frac{4}{1-\alpha^2}\left(\frac{1-\alpha}{2}F(X) + \frac{1+\alpha}{2}F(Y) - F\left(\frac{1-\alpha}{2}X + \frac{1+\alpha}{2}Y\right)\right)$$

The matrix mean of $\{M_1, ..., M_n\}$ is defined as the minimizer of the following optimization problem:

$$\bar{M}_\alpha = \arg\min_{X \in \mathrm{Sym}_+^*} \frac{1}{n}\sum_{i=1}^n J_{LD}^{(\alpha)}(X, M_i). \quad (9)$$

This can be solved by removing all terms independent of X, and applying the concave-convex procedure. We initialize $C_0 = \frac{1}{n}\sum_{i=1}^n M_i$ and update iteratively using the CCCP rule [17]

$$C_{t+1} = \left(\sum_{i=1}^n \frac{1}{n}\left(\frac{1-\alpha}{2}C_t + \frac{1+\alpha}{2}M_i\right)^{-1}\right)^{-1}.$$

Note that we can swap arguments in the α-log-det divergence by turning α into $-\alpha$:

$$J_{LD}^{(\alpha)}(X,Y) = J_{LD}^{(-\alpha)}(Y,X) \tag{10}$$

Furthermore, the α-log-det divergence is invariant under inversion and invertible transformations, i.e.,

$$J_{LD}^{(\alpha)}(X,Y) = J_{LD}^{(\alpha)}(X^{-1},Y^{-1}),$$
$$J_{LD}^{(\alpha)}(CXC^T,CYC^T) = J_{LD}^{(\alpha)}(X,Y), \forall C \in \mathrm{GL}(d),$$

where $\mathrm{GL}(d)$ is the set of invertible transformations. These properties are very important in many applications [15].

3.2. Symmetrized matrix α-log-det divergence

The symmetrized matrix α-log-det divergence is

$$sJ_{LD}^{(\alpha)}(X,Y) = \frac{1}{2}\left(J_{LD}^{(\alpha)}(X,Y) + J_{LD}^{(\alpha)}(Y,X)\right).$$

With initialization $C_0 = \frac{1}{n}\sum_{i=1}^n M_i$, the mean can also be solved using the updating CCCP rule,

$$C_{t+1} = (\nabla F)^{-1}\Big(\sum_{i=1}^n \frac{1}{n}(1-\alpha)\nabla F(\alpha M_i + (1-\alpha)C_t) + \alpha\nabla F(\alpha C_t + (1-\alpha)M_i)\Big)\Big).$$

4. Experiments

We have implemented the Jensen-based matrix concave-convex iteration algorithm in Java™ using the JAMA[1] matrix package. Our open source implementation is readily available[2] for reproducible research. We evaluated our method on both synthetic dataset and real shape dataset.

4.1. Synthetic dataset

To get an SPD matrix M, we randomly draw a lower triangle matrix L and let $M = LL^T$. Table 1 reports the gradients and inverse gradients for several commonly used convex generators .

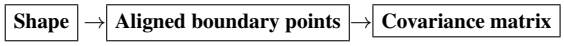
The Log-Euclidean-based, Riemannian-based and divergence-based methods all report the identity matrix for the mean of M with M^{-1}. We observed that our divergence-based algorithm converges fast to a unique global minimum in practice for the Jensen-von Neumann divergence: 10 iterations are enough to get a 0.1%-error-approximation to the minimum (linear convergence). As the dimension grows, the computational bottleneck is to calculate the eigendecomposition of the matrix for performing the log/exp matrix operations required for computing ∇F and $(\nabla F)^{-1}$. Indeed, eigendecomposition of d-dimensional square matrices requires roughly cubic time with a naive implementation.

Entropy	$F(X)$	∇F	$(\nabla F)^{-1}$
Quadratic	$\frac{1}{2}\mathrm{tr}(XX^T)$	X	X
log-det	$-\log\det X$	$-X^{-1}$	$-X^{-1}$
von Neum.	$\mathrm{tr}(X\log X - X)$	$\log X$	$\exp X$

Table 1. Gradients and inverse gradients of several convex matrix generators.

4.2. Shape clustering

Shape clustering is an important step for shape retrieval in a large database. Shape clustering enables hierarchical shape retrieval which is more efficient than brute force shape retrieval. We evaluated Jensen divergence based clustering on the MPEG-7 database [9], which consists of 70 different categories with 20 shapes per category, for a total of 1400 shapes. For each shape, we first extract its boundary points, align them using affine transformation, and then use the covariance matrix, which is an SPD matrix, of the aligned boundary points to represent this shape [11]. The SPD matrix is also the covariance matrix of the the Gaussian distribution estimated from the boundary points. The above process is portrayed using the flow chart shown below



The hard clustering algorithm [10, 11] is used to perform clustering. The clustering accuracy is measured according to a method proposed in [11], which is the optimal number of categories per cluster (denoted by $|S|^*$, $|S|$ represents the cardinality of S, i.e., the number of categories in S), divided by the average number of categories in each cluster (denoted by $\mathrm{Avg}(|S|)$). For example, if there are 10 clusters $\{S_i\}_{i=1}^{10}$, with an average of 140 shapes per cluster, and thus, $|S|^* = 140/20 = 7$; $\mathrm{Avg}(|S|) = \frac{\sum_{i=1}^{10}|S_i|}{10}$. The clustering accuracy describes the accuracy of separation of different categories. The optimal clustering accuracy is 1. Fig-

[1] http://math.nist.gov/javanumerics/jama/
[2] www.informationgeometry.org/SPD/

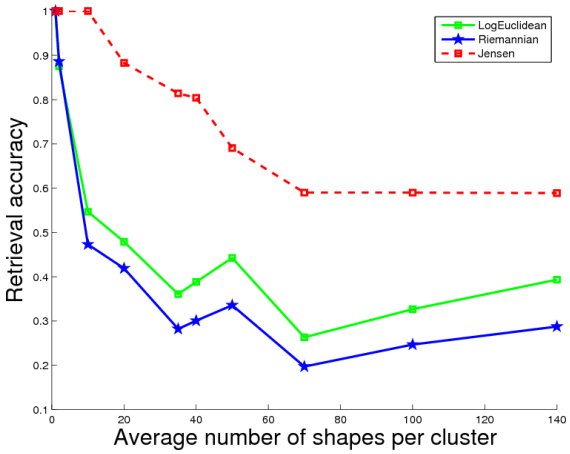


Figure 1. Shape clustering using Riemannian, LE, and Jensen divergences.

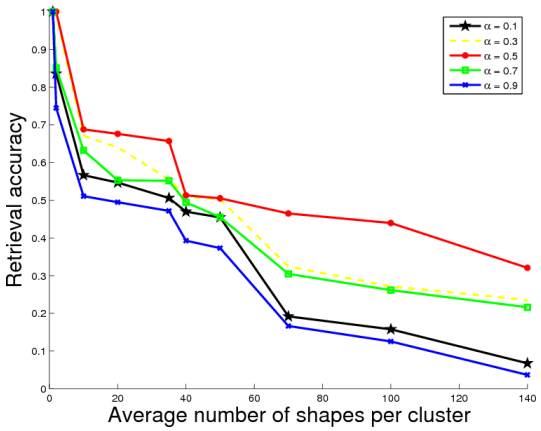


Figure 2. Clustering using symmetrized matrix α-log-det divergence for various α's.

ure 1 compares the clustering accuracy of using Log-Euclidean, Riemannian and our proposed Jensen divergence. The parameter α is set to be one which maximizes the clustering accuracy. In this experiment, the result achieves the best when $\alpha \simeq 0.4$ (this means that the center has more weight than each single element in the cluster). The results show that Jensen divergence enables much higher clustering accuracy, implying substantial capability to distinguish shapes from different categories.

We also used the symmetrized matrix α-log-det divergence to do clustering. By changing the α, we get different clustering accuracy, which is shown in Figure 2. The results illustrate that when $\alpha = 0.5$, the clustering achieves better accuracy.

5. Concluding remarks

We introduced divergence-based matrix means as minimizers of average divergences. We consider the class of matrix divergences induced by a convex functional, and described a novel efficient concave-convex iteration method to compute those means. The divergence-based mean depends on a convex matrix functional which may be tuned according to specific application domains.

References

[1] V. Arsigny, P. Fillard, X. Pennec, and N. Ayache. Geometric means in a novel vector space structure on symmetric positive-definite matrices. *SIAM J. Matrix Anal. Appl.*, 29(1):328–347, 2007.

[2] R. Bhatia and J. Holbrook. Riemannian geometry and matrix geometric means. *Linear Algebra and its Applications*, 413(2-3):594–618, 2006.

[3] J. Bigün and et al. Multidimensional orientation estimation with applications to texture analysis and optical flow. *IEEE TPAMI*, 13(8):775–790, 1991.

[4] D. Bini and B. Iannazzo. A note on computing matrix geometric means. *Adv. Comp. Math.*, 2010.

[5] D. A. Bini, B. Meini, and F. Poloni. An effective matrix geometric mean satisfying the Ando-Li-Mathias properties. *Math. Com.*, 79(269):437–452, 2010.

[6] J. Burbea and C. R. Rao. On the convexity of some divergence measures based on entropy functions. *IEEE TIT*, 28(3):489–495, 1982.

[7] S. C. Cowin and G. Yang. Averaging anisotropic elastic constant data. *J. Elasticity*, 46:151–180, 1997.

[8] S. Fiori. Learning the Fréchet mean over the manifold of symmetric positive-definite matrices. *Cognitive Computation*, 1:279–291, 2009.

[9] L. J. Latecki and et al. Shape descriptors for non-rigid shapes with a single closed contour. *IEEE CVPR*, 2000.

[10] M. Liu and et al. Total Bregman divergence and its applications to shape retrieval. *IEEE CVPR*, 2010.

[11] M. Liu and et al. Shape retrieval using hierarchical total Bregman soft clustering. *IEEE TPAMI*, 2012.

[12] F. Nielsen and S. Boltz The Burbea-Rao and Bhattacharyya Centroids. *IEEE TIT* 57(8), 2011.

[13] F. Nielsen and et al. Divergence based means of symmetric positive definite matrices. Matrix Information Geometry, ISBN 978-3-642-30231-2, Springer, 2012.

[14] O. Tuzel and et al. Human detection via classification on Riemannian manifolds. *IEEE TPAMI*, 2008.

[15] B. C. Vemuri and et al. Total Bregman divergence and its applications to DTI analysis. *IEEE TMI*, 2011.

[16] Y. Wang and C. Han. Polsar image segmentation by mean shift clustering in the tensor space. *IEEE Geo. Remote Sensing Letters*, 36(6):798–806, 2010.

[17] A. L. Yuille and A. Rangarajan. The concave-convex procedure. *Neural Computation*, 15(4):915–936, 2003.