Operators
and
Matrices
Volume 13, Number 3 (2019), 607–627 doi:10.7153/oam-2019-13-46

A RIEMANNIAN STRUCTURE FOR CORRELATION MATRICES

PAUL DAVID AND WEIQING GU

(*Communicated by Y. Nakatsukasa*)

Abstract. In this paper we present a new approach to viewing the set of non-degenerate correlation matrices $Corr(n)$ as a manifold and provide an optimization procedure using its newfound Riemannian structure. First we give a proof that $Corr(n)$ is a quotient submanifold of the symmetric positive-definite matrices $SPD(n)$ obtained via a Lie group action of positive diagonal matrices $Diag_{+}(n)$. With this structure $Corr(n)$ naturally inherits a Riemannian metric from $SPD(n)$ and therefore enables us to develop a Riemannian-based Newton's method on $Corr(n)$. We subsequently compare this Newton method to other optimization methods on $Corr(n)$.

1. Introduction

Correlations and correlation matrices have been a standard object of study in statistics and probability to measure the relationship between random variables, and as such are very crucial to modern data analysis. A number of modern research areas have utilized $SPD(n)$ and correlation matrices (not necessarily positive-definite) in a variety of applications including but not limited to diffusion tensor imaging [6, 2, 18, 21], statistics for modeling Gaussian distributions [14, 23], and their role in classification of data sets that occur on non-linear spaces [14, 8, 13, 10]. Of fundamental importance to the aforementioned research is to find efficient ways for averaging and optimizing nonlinear matrix-valued data. We let $SPD(n)$ be the set of symmetric positive-definite matrices of size n, and define $Corr(n) \subset SPD(n)$ to be the set of positive-definite correlation matrices (elements of $SPD(n)$ with unit diagonals). $SPD(n)$ happens to be an open and convex subset of the symmetric matrices $Symm(n)$, hence linear averaging is certainly possible on this subset. Many researchers however have found that utilizing various Riemannian manifold structures of $SPD(n)$ yield more accurate results in optimization and modeling. The work done in [2, 3] for instance show how distances between symmetric positive-definite matrices can be interpreted in terms of a log-Euclidean framework, in which $SPD(n)$ is shown to be isomorphic to the space of symmetric matrices $Symm(n) = T_I SPD(n)$ under a special binary operation defined on this space. The work of Moakher [17] gives a criterion for the barycenter of $SPD(n)$-valued observations, a result we subsequently use for a gradient descent procedure we use in Section 3.

Mathematics subject classification (2010): 53-04.

Keywords and phrases: Riemannian geometry, correlation, Newton's method, quotient manifold, optimization.

Our investigation of the manifold structure of $Corr(n)$ is the first, to our knowledge, to incorporate its relationship as a subset of the symmetric positive-definite matrices. It was already proven by Grubišić and Pietersz in [7] that the nondegenerate correlation matrices possess manifold structure, though this was merely a special case of a larger result they proved: Correlation matrices of fixed size and rank form a manifold. Our work in this article deviates from their construction in that we offer a new quotient and Riemannian structure for correlation matrices in terms of the symmetric positive-definite matrices $SPD(n)$. Utilizing the affine-invariant geometry of $SPD(n)$ we are able to develop a Newton optimization algorithm for the sum of mean-squared distances on $Corr(n)$. In Section 2 we give our proof of the quotient manifold structure of $Corr(n)$ in terms of $SPD(n)$. The main challenge following this was to define a meaningful Riemannian metric and distance on $Corr(n)$ that was reflective of its inherent quotient manifold structure. As a submanifold $Corr(n)$ certainly inherits the affine-invariant metric of $SPD(n)$ simply by restriction, however we discovered that this metric provides additional symmetry which was useful for computation. We observed that the Lie group acts isometrically with respect to the affine-invariant metric, hence using the theory of isometric Lie group actions developed in [11] the expressions for geodesics and distances between elements of $Corr(n)$ are simply modifications of those on $SPD(n)$. The main modification roughly involves the following procedure. Taking two elements $C_1, C_2 \in Corr(n)$ with C_1 as the starting point for the geodesic

1. View C_1 and C_2 as belonging to $SPD(n)$ and find the element $\widetilde{C_2}$ lying along the fiber over C_2 which minimizes the SPD-distance between C_1 and $\widetilde{C_2}$.

2. Find the SPD geodesic $\widetilde{\gamma}$ between C_1 and $\widetilde{C_2}$.

3. Obtain the $Corr$ geodesic γ simply by projecting the SPD geodesic back to $Corr(n)$ (i.e. $\gamma = \pi \circ \widetilde{\gamma}$ where $\pi : SPD(n) \to Corr(n)$ is the natural projection arising from the group action).

Having obtained the necessary Riemannian framework for $Corr(n)$, and hence geodesics in this space, we embarked on developing a Newton's algorithm for optimizing the sum of the squared distances between points in $Corr(n)$, the details of which comprise Section 3. We developed computations for approximating the Hessians as well as obtaining Riemannian distances between elements of $Corr(n)$, a task which relies entirely on the quotient structure and the fact that the action is isometric. Our Newton's method is the first to utilize the affine-invariant structure in this manner and is suggestive of a general way in which to perform optimization on arbitrary quotient manifolds. We lastly provide some numerical experiments in Section 4 to compare the performance of our algorithm to those proposed in [7], and investigate the validity of our affine-invariant structure for computation, as well as the differences between Euclidean and Riemannian means on $Corr(n)$ for various sizes of n. We offer a brief conclusion of our results in Section 5.

2. The manifold of correlation matrices

2.1. Manifold structure for Corr(n)

We present here a proof that $Corr(n)$ is a manifold.

THEOREM 1. *The Lie group $Diag_+(n)$ of diagonal matrices with positive entries acts smoothly, properly, and freely on $SPD(n)$ as a map $Diag_+(n) \times SPD(n) \to SPD(n)$ given by $(D,P) \to DPD$. Subsequently the quotient manifold $SPD(n)/Diag_+(n)$ resulting from this group action is a smooth manifold in which every element can be uniquely expressed by an element of $Corr(n)$. We therefore identify $Corr(n)$ with this quotient space itself, and observe that* $\dim Corr(n) = \dim SPD(n) - \dim Diag_+(n) = \frac{n(n-1)}{2}$.

Proof. Consider the group action $Diag_+(n) \times SPD(n) \to SPD(n)$ given by $\Lambda \cdot \Sigma = \Lambda\Sigma\Lambda$. It follows from the theory of group actions on manifolds that if this is a smooth, free, and proper action then the quotient space $SPD(n)/Diag_+(n)$ is a smooth manifold and the projection mapping a smooth submersion. We immediately have that the action is smooth since matrix multiplication is smooth. The action is free when $\Lambda\Sigma\Lambda = \Sigma$ implies that $\Lambda = I$. Observe that on the diagonal elements of Σ the action being free implies that $\lambda_{ii}^2 \sigma_{ii} = \sigma_{ii}$. Since the $\sigma_{ii} > 0$ we observe that $\lambda_{ii}^2 = 1$ implying $\lambda_{ii} = 1$ for each $i \in \{1,\dots,n\}$. Therefore $\Lambda = I$ and the action is free.

To show that the action is proper, we borrow a result from [15] that properness is equivalent to showing the following: If $\{\Sigma_k\}$ is a sequence in $SPD(n)$ such that $\Sigma_k \to \Sigma$, and $\{D_k\}$ is a sequence in $Diag_+(n)$ such that $D_k \cdot \Sigma_k \to Q \in SPD(n)$ then a subsequence of the $\{D_k\}$ converges in $Diag_+(n)$. If $\{\Sigma_k\}, \{D_k\}$ and Σ, Q all satisfy the above requirements, then we need only look at diagonal entries. The convergence of sequences of matrices in this case is equivalent to the convergence of real number sequences $\sigma_{ii}^{(k)} \left(d_{ii}^{(k)}\right)^2 \to q_{ii}$. One can easily show that $\left(d_{ii}^{(k)}\right)^2 \to \frac{q_{ii}}{\sigma_{ii}} > 0$, and since these values are all positive we have that the sequence $\left\{d_{ii}^{(k)}\right\}$ converges to a positive real number. Since this result holds for each value of i we have that the sequence $\{D_k\}$ converges in $Diag_+(n)$ and the action is proper.

We conclude that the quotient space $SPD(n)/Diag_+(n)$ is a smooth manifold of dimension $\frac{n(n+1)}{2} - n = \frac{n(n-1)}{2}$ with smooth projection mapping $SPD(n) \to SPD(n)/Diag_+(n)$. Lastly observe that we can represent each class $[\Sigma]$ in the quotient space uniquely by its representative $D_\Sigma \cdot \Sigma$ where $D_\Sigma = (I \circ \Sigma)^{-1/2}$ and $\circ$ is the Hadamard product. Therefore $D_\Sigma \cdot \Sigma$ is symmetric positive-definite with unit diagonal; a correlation matrix. Since the quotient manifold is in bijective correspondence with the correlation matrices we conclude that

$$Corr(n) = SPD(n)/Diag_+(n). \quad \square \tag{1}$$

2.2. Isometric group action and Riemannian structure

While the result that $Corr(n)$ exhibits this particular quotient manifold structure is meaningful in and of itself, this fact alone does not yield results that are suitable for algorithms and computation. In order for our results to have meaning in an algorithmic sense we need the notion of distance on $Corr(n)$ which can only be obtained from a Riemannian structure.

The Riemannian structure of the ambient manifold $SPD(n)$ has been well-studied and has been used for computations in a variety of contexts. In particular, authors have identified the affine-invariant structure for computing means [17], computer visualization [18, 19], statistics [16], DT MRI [21], and kernel dictionary learning [9] to name a few. In addition to this, another popular metric given by the Log-Euclidean framework realizes $SPD(n)$ as a manifold diffeomorphic to the vector space of symmetric matrices using the exponential map as a diffeomorphism between the spaces, and the matrix logarithm as its smooth inverse. The Log-Euclidean framework has been found to be useful again in diffusion tensor MRI [2, 20], kernel SVM [12], with Arsigny et al. uncovering a novel Lie group and vector space structure on $SPD(n)$ using this framework in [3].

For our present research, we find that the affine-invariant structure of $SPD(n)$ leads us to a natural Riemannian structure for $Corr(n)$. While we can consider the Riemannian structure of $Corr(n)$ induced by the affine-invariant metric of $SPD(n)$ given by

$$\langle A,B\rangle_P \;=\; \operatorname{Tr}\left[P^{-1}AP^{-1}B\right] \tag{2}$$

simply via the restriction of this metric to the tangent space at each point, we find that the Lie group action of $Diag_+(n)$ plays a crucial role in our understanding of this structure. In particular, we find that $Diag_+(n)$ acts isometrically on $SPD(n)$, leaving the value of the metric unchanged along the fibers of each point. Defining the map $\Phi_D : SPD(n) \to SPD(n)$ for each $D \in Diag_+(n)$ as $\Phi_D(P) = DPD$, we can easily verify that $d\left(\Phi_D\right)_P(A) = DAD$. We subsequently find that the pushforward along fibers satisfies

$$\langle d\left(\Phi_D\right)_P(A),\, d\left(\Phi_D\right)_P(B)\rangle_{\Phi_D(P)} \;=\; \langle A,B\rangle_P\,.$$

In [11] Huckemann et al. proved many useful results on the Riemannian structure of manifolds obtained from isometric Lie group actions. The main result needed for our present paper is the fact that the geodesic connecting two points in the quotient, can be expressed as the geodesic in the ambient manifold from the starting point to *an optimal representative of the end point, lying on the fiber over the desired endpoint.*

To make this concrete, given $C_1, C_2 \in Corr(n)$, the geodesic and corresponding distance in $SPD(n)$ connecting these two points are given by the following:

$$\begin{aligned} \gamma_{SPD}(t) &= C_1^{1/2}\operatorname{Exp}\left[t\operatorname{Log}\left(C_1^{-1/2}C_2C_1^{-1/2}\right)\right]C_1^{1/2} \\ d_{SPD}^2\left(C_1,C_2\right) &= \left|\left|\operatorname{Log}\left(C_1^{-1/2}C_2C_1^{-1/2}\right)\right|\right|_F^2. \end{aligned}$$

In order to adapt this Riemannian structure to $Corr(n)$ we need to find the optimal representative of C_2 with respect to the starting point C_1. This is done by finding the unique element $\widetilde{C}_2$ in the fiber $\pi^{-1}(C_2)$ which minimizes the SPD-distance between C_1 and $\widetilde{C}_2$. This can be found by writing

$$\begin{aligned} d^2_{Corr}(C_1,C_2) &= \inf_{D\in Diag_+(n)} d^2_{SPD}(C_1,DC_2D) && (3)\\ D^* &= \operatorname*{arg\,inf}_{D\in Diag_+(n)} d^2_{SPD}(C_1,DC_2D) && (4)\\ \widetilde{C}_2 &= D^*C_2D^*. && (5) \end{aligned}$$

As we will see in Section 3.3 we will generally approximate solutions to equation 4 using a Riemannian gradient descent on $Diag_+(n)$. We then find that the corresponding geodesic can be taken as the projection of the SPD-geodesic connecting C_1 and $\widetilde{C}_2$:

$$\gamma_{Corr}(t) \;=\; \pi\left(C_1^{1/2}\operatorname{Exp}\left[t\operatorname{Log}\left(C_1^{-1/2}\widetilde{C}_2C_1^{-1/2}\right)\right]C_1^{1/2}\right). \tag{6}$$

2.3. Adaptability to complex correlations

We briefly mention here that the preceeding formulation can just as well be adapted to complex-valued correlations by looking at a similar action on the set of Hermitian positive-definite matrices which we denote $HPD(n)$. Following the work of [1], we know that $HPD(n)$ is a homogeneous space equivalent to $HPD(n) = GL_n(\mathbb{C})/U(n)$, with canonical tangent space $T_IHPD(n) = Herm(n)$, the set of Hermitian matrices. An affine-invariant Riemannian metric is similarly given by

$$\langle A,B\rangle_P \;=\; \operatorname{Tr}\left[P^{-1}AP^{-1}B\right] \tag{7}$$

where remarkably the right hand-side of equation 7 is real-valued, which can easily be demonstrated by noting that $P^{-1}AP^{-1}$ and B are both Hermitian and therefore admit decompositions of the form $X+iY$, where X is real symmetric and Y is real skew-symmetric. Theorem 1 can easily be adapted by considering the action $Diag_+(n)\times HPD(n)\to HPD(n)$ by $(D,H)\to DHD$. The proof is identical once we acknowledge that elements of $HPD(n)$ must have real positive diagonal entries, and thus we obtain that complex correlations possess a quotient structure

$$Corr(n,\mathbb{C}) \;=\; HPD(n)/Diag_+(n). \tag{8}$$

The Lie group $Diag_+(n)$ similarly acts isometrically on $HPD(n)$, and therefore the expressions for geodesics and distances follow similarly using the results of [11]. The one discrepancy here is simply acknowledging the dimensionality differences between the real and complex cases. Viewing $HPD(n)$ as a real manifold we find that

$$\begin{aligned} \dim HPD(n) = \dim GL_n(\mathbb{C}) - \dim U(n) &\;=\; 2n^2-n^2 \;=\; n^2\\ \dim Corr(n,\mathbb{C}) = \dim HPD(n) - \dim Diag_+(n) &\;=\; n^2-n \;=\; n(n-1). \end{aligned}$$

3. Newton's method on SPD(n) and Corr(n)

We present here a Newton's gradient descent algorithm which seeks to minimize the mean-squared distances of $SPD(n)$ and $Corr(n)$-valued observations with respect to the affine-invariant distance. Given observations $P_{(1)},\dots,P_{(N)} \in SPD(n)$, the objective function on $SPD(n)$ we will incorporate is

$$F_{SPD}(P) \;=\; \frac{1}{2}\sum_{i=1}^{N} d^2_{SPD}\left(P_{(i)},P\right) \;=\; \frac{1}{2}\sum_{i=1}^{N}\left|\left|\mathrm{Log}\left(P_{(i)}^{-1/2}PP_{(i)}^{-1/2}\right)\right|\right|_F^2. \tag{9}$$

We will similarly discuss the same objective function applied to $Corr(n)$ where given observations $C_{(1)},\dots,C_{(N)} \in Corr(n)$ we define

$$F_{Corr}(C) \;=\; \frac{1}{2}\sum_{i=1}^{N} d^2_{Corr}\left(C_{(i)},C\right) \;=\; \frac{1}{2}\sum_{i=1}^{N}\underset{D_i\in Diag_+(n)}{\arg\inf}\left|\left|\mathrm{Log}\left(C_{(i)}^{-1/2}D_iCD_iC_{(i)}^{-1/2}\right)\right|\right|_F^2. \tag{10}$$

For the sake of computing the gradients and Hessians of these similar functions, we initially focus our attention on equation 9. The reason for this is that in equation 9 we seek to optimize P, while in equation 10 we seek to optimize C. It is therefore desirable to express our Newton's method for working in $SPD(n)$, and then to adapt our results when we are restricted to working $Corr(n)$, where Newton's method again occurs in the ambient space $SPD(n)$, but with the added steps of finding optimal points along fibers (before the update) and then projection back to $Corr(n)$ (after the update).

3.1. Computing gradients

The computation of the gradient for $SPD(n)$ follows immediately from the work of Moakher [17] where it was shown for the objective function F_{SPD}

$$\nabla F_{SPD}(P) \;=\; P\sum_{i=1}^{N}\mathrm{Log}\left(P_{(i)}^{-1}P\right). \tag{11}$$

3.2. Approximating Hessians

We focus again on the techniques developed in [17] in order to approximate the Hessian to F_{SPD}. Fundamental to this is the notion of approximating our expression for the gradient above using a Taylor expansion. It is helpful for this analysis to define

$$f_i(P) \;=\; \frac{1}{2}\left|\left|\mathrm{Log}\left(P_{(i)}^{-1/2}PP_{(i)}^{-1/2}\right)\right|\right|_F^2 \;=\; \frac{1}{2}\mathrm{Tr}\left[\mathrm{Log}^2\left(P_{(i)}^{-1/2}PP_{(i)}^{-1/2}\right)\right]$$

and to observe that

$$F_{SPD}(P) = \sum_{i=1}^{N} f_i(P) \qquad \nabla F_{SPD}|_P = \sum_{i=1}^{N}\nabla f_i|_P \qquad \mathrm{Hess}\, F_{SPD}|_P = \sum_{i=1}^{N}\mathrm{Hess}\, f_i|_P. \tag{12}$$

To proceed with approximating the Hessian, we borrow a key lemma from [17], which demonstrated that given a smooth matrix-valued function $X(t)$ of a real variable t such that $X^{-1}(t)$ exists for all t, we have that

$$\frac{d}{dt}\frac{1}{2}\operatorname{Tr}\left[\operatorname{Log}^2(X(t))\right] \;=\; \operatorname{Tr}\left[(\operatorname{Log}X(t))X^{-1}(t)\dot{X}(t)\right]. \tag{13}$$

To find an explicit approximation to the Hessian, consider the application Hess $f_i(\Delta,\Delta)|_P$ with tangent vector Δ. We compute the Hessian in coordinates by considering the $SPD(n)$-valued curve

$$X(t) \;=\; P_{(i)}^{-1/2}P^{1/2}\operatorname{Exp}\left(tP^{-1/2}\Delta P^{-1/2}\right)P^{1/2}P_{(i)}^{-1/2} \;=\; P_{(i)}^{-1/2}P(t)P_{(i)}^{-1/2}$$

where $P(t)$ is the SPD geodesic satisfying $P(0)=P$ and $\dot{P}(0)=\Delta$. A straightforward, albeit tedious, computation demonstrates that

$$\frac{d}{dt}\frac{1}{2}\operatorname{Tr}\left[\operatorname{Log}^2(X(t))\right] \;=\; \operatorname{Tr}\left[\operatorname{Log}\left(P_{(i)}^{-1}P^{1/2}\operatorname{Exp}\left(tP^{-1/2}\Delta P^{-1/2}\right)P^{1/2}\right)P^{-1}\Delta\right] \tag{14}$$

where we note a liberal use of the identity $A\operatorname{Log}(B)A^{-1}=\operatorname{Log}\left(ABA^{-1}\right)$, which can easily be proven taking the series expansion of the logarithm into account. In computing the Hessian at the point $P\in SPD(n)$, we need to take another derivative of the expression above and then evaluate it at $t=0$. For symmetric matrices, the logarithm function admits a closed-form expression using the spectral decomposition of its argument. We however chose to avoid this method in computing the Hessian since this would necessitate a chain. Instead, we considered the Taylor expansion of the Logarithm function, which for algorithmic purposes yields terms expressed in ordinary matrix multiplication and tensor products. For our purposes, we truncated our expansion to third order, and obtained the approximate expression

$$\begin{aligned}\operatorname{Log}(A) &= \sum_{k=1}^{\infty}\frac{(-1)^{k+1}}{k}(A-I)^k \approx A-I-\frac{1}{2}(A^2-2A+I)+\frac{1}{3}(A^3-3A^2+3A-I)\\ &= \frac{1}{3}A^3-\frac{3}{2}A^2+3A-\frac{11}{6}I.\end{aligned}$$

Taking $A=P_{(i)}^{-1}P^{1/2}\operatorname{Exp}\left(tP^{-1/2}\Delta P^{-1/2}\right)P^{1/2}$, we compute the derivatives of each term as

$$\begin{aligned}\left.\frac{d}{dt}A\right|_{t=0} &= \left.\frac{d}{dt}P_{(i)}^{-1}P^{1/2}\operatorname{Exp}\left(tP^{-1/2}\Delta P^{-1/2}\right)P^{1/2}\right|_{t=0} = P_{(i)}^{-1}\Delta\\ \left.\frac{d}{dt}A^2\right|_{t=0} &= \left.A\frac{dA}{dt}+\frac{dA}{dt}A\right|_{t=0} = P_{(i)}^{-1}PP_{(i)}^{-1}\Delta+P_{(i)}^{-1}\Delta P_{(i)}^{-1}P\\ \left.\frac{d}{dt}A^3\right|_{t=0} &= \left.A\frac{dA^2}{dt}+\frac{dA}{dt}A^2\right|_{t=0} = P_{(i)}^{-1}P\left[P_{(i)}^{-1}PP_{(i)}^{-1}\Delta+P_{(i)}^{-1}\Delta P_{(i)}^{-1}P\right]+P_{(i)}^{-1}\Delta P_{(i)}^{-1}PP_{(i)}^{-1}P.\end{aligned}$$

Our full expression for the second derivative is therefore given by the approximation

$$\begin{aligned}
&\frac{d^2}{dt^2}\frac{1}{2}\operatorname{Tr}\left[\operatorname{Log}^2(X(t))\right]\Big|_{t=0} \qquad (15)\\
&\approx \frac{1}{3}\operatorname{Tr}\left[P_{(i)}^{-1}PP_{(i)}^{-1}PP_{(i)}^{-1}\Delta P^{-1}\Delta + 2P_{(i)}^{-1}PP_{(i)}^{-1}\Delta P_{(i)}^{-1}\Delta\right]\\
&\quad -\frac{3}{2}\operatorname{Tr}\left[P_{(i)}^{-1}PP_{(i)}^{-1}\Delta P^{-1}\Delta + P_{(i)}^{-1}\Delta P_{(i)}^{-1}\Delta\right] + 3\operatorname{Tr}\left[P_{(i)}^{-1}\Delta P^{-1}\Delta\right].
\end{aligned}$$

It is important to note that we can find the off-diagonal terms using the following equation

$$\begin{aligned}
\operatorname{Hess} f_i(\Delta_1,\Delta_2) = {}& \frac{1}{6}\operatorname{Tr}\left[P_{(i)}^{-1}\left(PP_{(i)}^{-1}\right)^2\Delta_1 P^{-1}\Delta_2\right] + \frac{1}{6}\operatorname{Tr}\left[P_{(i)}^{-1}\left(PP_{(i)}^{-1}\right)^2\Delta_2 P^{-1}\Delta_1\right]\\
&+\frac{1}{3}\operatorname{Tr}\left[P_{(i)}^{-1}PP_{(i)}^{-1}\Delta_1 P_{(i)}^{-1}\Delta_2\right] + \frac{1}{3}\operatorname{Tr}\left[P_{(i)}^{-1}PP_{(i)}^{-1}\Delta_2 P_{(i)}^{-1}\Delta_1\right]\\
&-\frac{3}{4}\operatorname{Tr}\left[P_{(i)}^{-1}PP_{(i)}^{-1}\Delta_1 P^{-1}\Delta_2\right] - \frac{3}{4}\operatorname{Tr}\left[P_{(i)}^{-1}PP_{(i)}^{-1}\Delta_2 P^{-1}\Delta_1\right]\\
&-\frac{3}{4}\operatorname{Tr}\left[P_{(i)}^{-1}\Delta_1 P_{(i)}^{-1}\Delta_2\right] - \frac{3}{4}\operatorname{Tr}\left[P_{(i)}^{-1}\Delta_2 P_{(i)}^{-1}\Delta_1\right]\\
&+\frac{3}{2}\operatorname{Tr}\left[P_{(i)}^{-1}\Delta_1 P^{-1}\Delta_2\right] + \frac{3}{2}\operatorname{Tr}\left[P_{(i)}^{-1}\Delta_2 P^{-1}\Delta_1\right].
\end{aligned}$$

The derivation for this comes from the polarization identity which we borrow from [5], and leave the details of the computation in the Appendix A. In Newton's method we will need to invert the Hessian of our objective function, hence it is necessary to express our equation in terms of a standalone expression. Observing the tensor relation

$$(\operatorname{vec} A)^T(B\otimes C)(\operatorname{vec} D) \;=\; \operatorname{Tr}\left(DB^TA^TC\right) \;=\; \operatorname{Tr}\left(B^TA^TCD\right),$$

the fact that all of the matrices that we're considering are symmetric, and recalling the affine-invariant metric, we compute the following for each term in the above Hessian approximation:

$$\begin{aligned}
\operatorname{Hess} f_i|_P \approx {}& \frac{1}{6}\left[\left(PP_{(i)}^{-1}\right)^3\otimes I + I\otimes\left(P_{(i)}^{-1}P\right)^3\right] \qquad (16)\\
&+\frac{1}{3}\left[\left(PP_{(i)}^{-1}\right)^2\otimes P_{(i)}^{-1}P + PP_{(i)}^{-1}\otimes\left(P_{(i)}^{-1}P\right)^2\right]\\
&-\frac{3}{4}\left[\left(PP_{(i)}^{-1}\right)^2\otimes I + I\otimes\left(P_{(i)}^{-1}P\right)^2\right]\\
&+\frac{3}{2}\left[PP_{(i)}^{-1}\otimes I + I\otimes P_{(i)}^{-1}P - PP_{(i)}^{-1}\otimes P_{(i)}^{-1}P\right].
\end{aligned}$$

We obtain our full approximation for the Hessian as

$$\operatorname{Hess} F_{SPD}|_P \;\approx\; \sum_{i=1}^{N}\operatorname{Hess} f_i. \qquad (17)$$

The process of truncating the Taylor approximation of the matrix logarithm is admittedly an ad hoc procedure, one which may be viewed as problematic, especially if the argument is far away from the identity. In our algorithm we actually include a criterion as to whether we include the component Hess f_i in our overall Hessian. Recall that the matrix logarithm $\mathrm{Log}(A)$ is absolutely convergent in the case that $||A-I||_F < 1$. We indeed impose this condition and discard terms that do not satisfy this.

3.3. Optimizing along fibers

The above analysis can be applied equally well in $SPD(n)$ and will ultimately be used for Newton's method on $Corr(n)$. However, before we continue we need to specify how to optimize along the fibers over $Corr(n)$-valued elements since this is crucial in accurately representing Riemannian distances between points of $Corr(n)$. We recall again from [11] that in order to appropriately find the distance between elements C_1, C_2 of a quotient manifold we need to find the element between C_1 and the fiber over C_2 which minimizes the overall distance[1]. Recall again that the distance between $C_1, C_2 \in Corr(n)$ is given by

$$d^2_{Corr}(C_1,C_2) = \inf_{D\in Diag_+(n)} d^2_{SPD}(C_1, DC_2D) = \inf_{D\in Diag_+(n)} \mathrm{Tr}\left[\mathrm{Log}^2\left(C_1^{-1/2} DC_2DC_1^{-1/2}\right)\right]$$

where we note that by symmetry we can just as well fix C_2 and then optimize over the fiber of C_1. For our purposes, we intend to minimize the distance between an iterate C_t of our algorithm between all of the observations $C_{(1)},\dots,C_{(N)}$, hence we want to arrange our algorithm so that we are always keeping our iterate fixed and then optimizing along the fibers of our observations. In this way, we guarantee that we are updating our itererated point appropriately. If we were to optimize along the fiber of our iterate, we would end up with drastically different optimal points, not yielding a consistent base point.

An important note here is to recall the quantity D^* given in equation 4. We acknowledge that fiber optimization is relegated to a learning problem on the Lie group $Diag_+(n)$. Even though $Diag_+(n)$ is convex, the fact that it is noncompact leaves us without a guarantee for convergence. However as we have seen in practice, the convergence of our algorithm is quite reliable and has not posed any issues with respect to optimization along $Diag_+(n)$. In finding the optimal point, we employ a simpler gradient descent in the Lie group $Diag_+(n)$ with respect to the objective function

$$g_i(D) \;=\; \frac{1}{2} d^2_{SPD}(C_1, DC_2D). \tag{18}$$

In computing the gradient we use equation 13, the same result as given in [17]. In this case we take

$$X(t) \;=\; C_1^{-1/2} S(t) C_1^{-1/2} \;=\; C_1^{-1/2} \gamma(t) C_2 \gamma(t) C_1^{-1/2}$$

[1] Here the distance is taken in the ambient manifold. In our case, this would be $SPD(n)$.

where $\gamma(t) = D^{1/2}\operatorname{Exp}\left(tD^{-1/2}\Delta D^{-1/2}\right)D^{1/2}$ and $S(t) = \gamma(t)C_2\gamma(t)$. Here γ is a geodesic in $Diag_+(n)$ with respect to the same affine-invariant metric. In order to compute the gradient of the objective function we compute

$$\begin{aligned}
&\frac{d}{dt}\frac{1}{2}d^2_{SPD}(C_1, DC_2D)\Big|_{t=0} \\
&= \frac{d}{dt}\frac{1}{2}\operatorname{Tr}\left[\operatorname{Log}^2\left(C_1^{-1/2}S(t)C_1^{-1/2}\right)\right]\Big|_{t=0} \\
&= \operatorname{Tr}\left[\operatorname{Log}\left(C_1^{-1/2}S(0)C_1^{-1/2}\right)\left(C_1^{1/2}D^{-1}C_2^{-1}D^{-1}C_1^{1/2}\right)C_1^{-1/2}\left(\Delta C_2D + DC_2\Delta\right)C_1^{-1/2}\right] \\
&= \operatorname{Tr}\left[\operatorname{Log}\left(C_1^{-1}DC_2D\right)\left(D^{-1}C_2^{-1}D^{-1}\right)\left(\Delta C_2D + DC_2\Delta\right)\right] \\
&= \operatorname{Tr}\left[\operatorname{Log}\left(C_2DC_1^{-1}D\right)D^{-1}\Delta + \operatorname{Log}\left(C_1^{-1}DC_2D\right)D^{-1}\Delta\right].
\end{aligned}$$

Interpreting the above quantity in light of the affine-invariant metric in equation 2 applied to $Diag_+(n)$, we find that

$$\nabla g_i(D) \;=\; I \circ D\left[\operatorname{Log}\left(C_2DC_1^{-1}D\right) + \operatorname{Log}\left(C_1^{-1}DC_2D\right)\right] \tag{19}$$

where we take the Hadamard product with respect to I to guarantee this quantity is restricted to diagonal matrices. For computational convenience we can actually simplify this further taking the symmetrization operator $\operatorname{Sym}(A) = \frac{1}{2}(A + A^T)$ and the fact that due to the Taylor series expansion for the logarithm we know that $A\operatorname{Log}(B)A^{-1} = \operatorname{Log}(ABA^{-1})$. We find that

$$\begin{aligned}
&D\operatorname{Log}\left(C_2DC_1^{-1}D\right) + D\operatorname{Log}\left(C_1^{-1}DC_2D\right) \\
&= D\operatorname{Log}\left(C_2DC_1^{-1}D\right) + D\operatorname{Log}\left(C_1^{-1}DC_2D\right)D^{-1}D \\
&= D\operatorname{Log}\left(C_2DC_1^{-1}D\right) + \operatorname{Log}\left(DC_1^{-1}DC_2\right)D \\
&= D\operatorname{Log}\left(C_2DC_1^{-1}D\right) + \left[D\operatorname{Log}\left(C_2DC_1^{-1}D\right)\right]^T = 2\operatorname{Sym}\left[D\operatorname{Log}\left(C_2DC_1^{-1}D\right)\right].
\end{aligned}$$

Hence we further simplify

$$\nabla g_i(D) \;=\; I \circ 2\operatorname{Sym}\left[D\operatorname{Log}\left(C_2DC_1^{-1}D\right)\right]. \tag{20}$$

In order to minimize the objective function 18 we can follow a simple gradient descent algorithm using a stepsize $\delta > 0$ by the following iterative steps

$$\begin{aligned}
\Delta_t &= I \circ 2\operatorname{Sym}\left[D_t\operatorname{Log}\left(C_2D_tC_1^{-1}D_t\right)\right] \\
D_{t+1} &= D_t^{1/2}\operatorname{Exp}\left(-\delta D_t^{-1/2}\Delta_tD_t^{-1/2}\right)D_t^{1/2} = D_t\operatorname{Exp}\left(-\delta D_t^{-1}\Delta_t\right)
\end{aligned}$$

until a desired stopping criterion is reached. We note that in the step rule for D_{t+1} that because all of the elements commute we can avoid finding matrix square-roots to ease computation. Once we find an optimal Lie group element $D^* \in Diag_+(n)$, as a result of minimizing $g(D)$, we write our optimal fiber element over C_2 as $\widetilde{C}_2 = D^*C_2D^*$.

3.4. Newton's method

In writing out a geodesic Newton's method for $SPD(n)$, we see that our update rules are given by

$$
\begin{aligned}
H_t &= -\left(\operatorname{Hess} F_{SPD}\right)^{-1}\left(\nabla F_{SPD}(P_t)\right) \\
P_{t+1} &= P_t^{1/2}\operatorname{Exp}\left(P_t^{-1/2}H_tP_t^{-1/2}\right)P_t^{1/2}
\end{aligned}
$$

where ∇F_{SPD} and $\operatorname{Hess} F_{SPD}$ are given by equations 11 and 17, respectively. While this works fine for optimization in $SPD(n)$, we need to modify these updates when taken in the context of $Corr(n)$ for the following reasons:

1. To accurately reflect distance between elements of $Corr(n)$, we need to find optimal points $\widetilde{C}_{(i)}$ along the fibers of our observations $C_{(i)}$, and measure the distance of these optimal points to our current iterate C_t.

2. Because our optimal points $\widetilde{C}_{(i)}$ will in general belong to $SPD(n)$ but not $Corr(n)$, Newton's method will then be carried out in $SPD(n)$ using the above update rules, treating our current iterate $C_t \in SPD(n)$.

3. We determine a new update P_{t+1} using the above update rules, but because P_{t+1} will in general not lie in $Corr(n)$, we project this element down and define $C_{t+1} = \pi(P_{t+1})$.

We summarize our proposed algorithms for $SPD(n)$ and $Corr(n)$ in algorithms 1 and 2.

Algorithm 1 Newton's method on $SPD(n)$

```
Require: Observables $P_{(1)},\dots,P_{(N)}$, initial point $P_0$
$t = 0$
while Stopping criterion not met do:
    $G_t = P_t \sum_{i=1}^{N} \operatorname{Log}\left(P_{(i)}^{-1}P_t\right)$
    $H_t = \operatorname{Hess} F_{SPD}|_{P_t}$ (given in equation 17)
    $V_t = -H_t^{-1}(\operatorname{vec}(G_t))$ (result is $n^2 \times 1$)
    Reshape $V_t$ to be $n \times n$
    $P_{t+1} = P_t^{1/2}\operatorname{Exp}\left(P_t^{-1/2}V_tP_t^{-1/2}\right)P_t^{1/2}$
    $t \to t+1$
end while
```

We compare this with the augmented steps for Newton's method on $Corr(n)$ below:

Algorithm 2 Newton's method on $Corr(n)$

```
Require: Observables $C_{(1)},\dots,C_{(N)}$, initial point $C_0$, stepsize $\delta > 0$
 1: $t = 0$
 2: while Stopping criterion not met do:
 3:     for $i = 1,\dots,N$ do:
Require:        Initial point $D_0$
 4:         $k = 0$
 5:         while Stopping criterion not met do:
 6:             $\Delta_k = I \circ 2\,\mathrm{Sym}\left[D_k \,\mathrm{Log}\left(C_{(i)} D_k C_t^{-1} D_k\right)\right]$
 7:             $D_{k+1} = D_k \,\mathrm{Exp}\left(-\delta D_k^{-1} \Delta_k\right)$
 8:             $k \to k+1$
 9:         end while
10:         $\widetilde{C}_{(i)} = D_{k_{\max}} C_{(i)} D_{k_{\max}}$.
11:     end for
12:     $G_t = C_t \sum_{i=1}^{N} \mathrm{Log}\left(\widetilde{C}_{(i)}^{-1} C_t\right)$
13:     $H_t = \mathrm{Hess}\, F_{SPD}|_{C_t}$ (given in equation 17)
14:     $V_t = -H_t^{-1}(\mathrm{vec}(G_t))$ (result is $n^2 \times 1$)
15:     Reshape $V_t$ to be $n \times n$
16:     $P_{t+1} = C_t^{1/2} \,\mathrm{Exp}\left(C_t^{-1/2} V_t C_t^{-1/2}\right) C_t^{1/2}$
17:     Project back to $Corr(n)$ with $C_{t+1} = \pi(P_{t+1}) = (I \circ P_{t+1})^{-1/2} P_{t+1} (I \circ P_{t+1})^{-1/2}$
18:     $t \to t+1$
19: end while
```

We comment here that while we focus specifically on the cases on $SPD(n)$ and its quotient $Corr(n)$, the Newton's method employed here is indicative of a way to conduct Newton's method on quotient manifolds obtained via isometric Lie group actions. An additional note is in regards to the gradient expression in line 12 of algorithm 2. Here we are using the gradient expression for mean-squared distance on $SPD(n)$ since after we lift our matrices along fibers, we then update our position in $SPD(n)$ and then project back to $Corr(n)$.

4. Numerical experiments

In this section we discuss the performance of our proposed algorithm against other algorithms on the correlation manifold. First we analyze the difference between the mean-squared distance of randomly sampled points using different Riemannian structures. Specifically we compare Riemannian mean and Euclidean mean of randomly sampled correlations and give the distribution of their distances for different sizes of matrices. We next have a discussion of convergence by looking at the rate of convergence of our algorithm compared to others. We compare the log relative residual of the gradient steps to other algorithms for various size of matrix and number of correlations to minimize.

4.1. Affine-invariant mean vs. Euclidean mean

Since we have developed an algorithm on $Corr(n)$ for minimization of mean-squared distances, it is reasonable for us to consider in what sense that the minimizer of the objective function stated in equation 10 deviates from the Euclidean mean. While it is not clear the exact relationship between these quantities then we at least seek to quantify the distances between these points.

Given an observed set of correlations $C_{(1)},\dots,C_{(N)}$, we can find both the linear average given by

$$\overline{C}^{Lin} = \frac{1}{N}\sum_{i=1}^{N} C_{(i)}$$

as well as our Riemannian averaging given by

$$\overline{C}^{Riem} = \underset{C\in Corr(n)}{\arg\inf}\ F_{Corr}(C)$$

where again F_{Corr} is the function in 10. We lastly compute the distance between these points on the manifold by computing

$$dist = d_{Corr}\left(\overline{C}^{Lin}, \overline{C}^{Riem}\right).$$

We repeat this sequence of computations 500 times each on $Corr(2), Corr(5), Corr(10)$, and $Corr(12)$. Similarly we also consider a fixed matrix size and evaluate the changes in the distributions as we increased the sample size. Both histograms can be viewed in figure 1. These histograms can give us a sense in how much the affine-invariant structure for $Corr(n)$ deviates from the Euclidean structure we can impose on $Corr(n)$. The histograms indicate a well-defined distribution showing the deviation of the Riemannian mean from the Euclidean mean. In addition, there appears to be an increase in this deviation as we increase the size of the matrix. This phenomena may give an indication as to how the curvature of $Corr(n)$ changes with an increasing n, but we leave these considerations for future research.

4.2. Convergence and run times

Our Newton algorithm that we proposed is not the first Riemannian optimization method proposed on $Corr(n)$. In [7] Grubišić and Pietersz provided gradient descent and Newton algorithms on $Corr(n)$ using a different structure: the manifold of Cholesky factors. The work presented therein was mainly focused on the problem of rank reduction of correlation matrices and their optimization procedure involved defining a Riemannian manifold of Cholesky factors for each size n of the correlation matrix and each desired size d for the rank of said matrix. The formalism they presented was strongly motivated by the works [5, 4] where Edelman et al. defined efficient and robust optimization methods for the Grassmann and Stiefel manifolds. Adapting the geometric optimization procedures of the Grassmann and Stiefel manifolds yielded efficient methods for correlation optimization of the Correlation manifold.

As a way to compare the convergence properties of our proposed method compared to theirs we look at several of the methods utilized in [7, 22, 24]. Specifically we look at

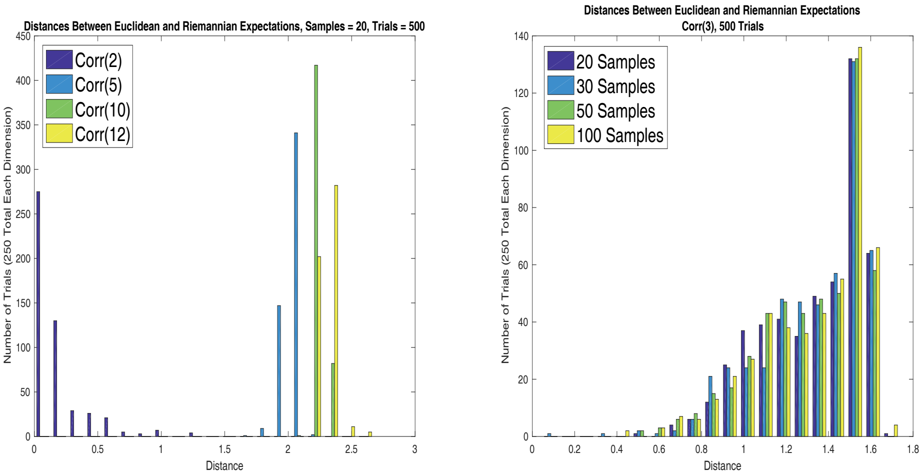

Figure 1: **Left**: Histogram of distances between Euclidean and Riemannian means for sets of randomly intialized correlations of $Corr(2), Corr(5), Corr(10)$, and $Corr(12)$. There are 500 trials for each size of matrix and each trial we take the mean of 20 correlations. We can see that as the dimension increases, so does the peak mean distance for the distribution. **Right**: Histogram of distances between Euclidean and Riemannian means in $Corr(3)$ for samplings sets of sizes 20, 30, 50, and 100. We can discern here a well-defined distribution which arises independent of the sample size. At the present moment our best understanding of the relationship between these two means is given statistically.

1. Newton's method on Cholesky factors,

2. Fletcher-Reeves optimization, and

3. Polak-Ribière optimization.

In order to offer an accurate comparison between these methods we would ideally need to utilize the same objective function accross optimization algorithms. Unfortunately the affine-invariant structure that we have developed for $Corr(n)$ does not yet offer flexibility for objective functions other than the mean-squared Riemannian distance. In our future work we intend to broaden the applicability of this Riemannian structure to other optimization problems on $Corr(n)$. In order to provide a meaningful comparison of the convergence properties of our algorithm we compare the Riemannian mean-squared optimization problem to that of the Euclidean mean-squared optimization problem, where given a sequence of observed correlations $C_{(1)}, \ldots, C_{(N)}$ we seek to

$$\text{minimize } F_{Lin}(C) \;=\; \frac{1}{2}\sum_{i=1}^{N} ||C - C_{(i)}||_F^2 \qquad \text{such that} \qquad C \in Corr(n). \tag{21}$$

To summarize: given sampled correlations $C_{(1)},\dots,C_{(N)}$ we minimize the objective function in equation 10 using the affine-invariant Newton method, and minimize equation 21 using the Cholesky Newton, Cholesky Fletcher-Reeves, and Cholesky Polak-Ribière algorithms. To compare the performance of each of these we look at the relative log-residuals of the gradient vectors at each iteration of our algorithms given by the quantity $\ln(||\nabla F(C_i)||/||\nabla F(C_1)||)$ where the norm of the tangent vectors is the same as the distance used for the corresponding optimization procedure (i.e. the Frobenius norm is used when we minimize equation 21 and the affine-invariant norm is used to minimize equation 10). We provide a few samples below in figure 2. What we find overall is that the affine-invariant Newton method has a relatively slow rate of convergence compared to the other methods.

The behavior is consistent and the log-resolution of tangent vectors decreases with each step as the various experiments show. To give a sense of quickly these algorithms run we recorded the run times of these algorithms over various dimensions and number of samples. We show the histograms for run times in figure 3.

5. Conclusion

We first presented here a proof that the correlation matrices form a quotient submanifold of $SPD(n)$, next developed a Riemannian-based Newton's algorithm for these matrices, and finally demonstrated the efficacy of this new optimization procedure with numerical experiments showing its convergence properties as well as comparing Euclidean and Riemannian means. Our work gives a new geometric characterization of correlation matrices which has a direct impact on our understanding of statistics and machine learning. The quotient structure of $Corr(n)$ proven is distinct from the other Riemannian structures defined on it since it inherits a highly structured Riemannian metric from $SPD(n)$. The affine-invariant metric of $SPD(n)$ remains unchanged with respect to the action of $Diag_+(n)$, thus we are able to utilize the theory of isometric Lie group actions to efficiently obtain expressions for geodesics in $Corr(n)$. We obtain these expressions by first finding optimal points along the fiber of each correlation, follow the $SPD(n)$-geodesic, and then project the result back to $Corr(n)$. Because of ubiquity of correlation matrices in applied mathematics, this newfound Riemannian structure for $Corr(n)$ has implications for probability, statistics, and methods for data analysis and machine learning.

A. Off-diagonal terms of the approximate Hessian

Here we derive the off-diagonal terms of the approximate Hessian given in 15. Here we use the polarization identity for quadratic forms Q

$$Q(\Delta_1,\Delta_2) = \frac{1}{4}\left[Q(\Delta_1+\Delta_2,\Delta_1+\Delta_2) - Q(\Delta_1-\Delta_2,\Delta_1-\Delta_2)\right] \tag{22}$$

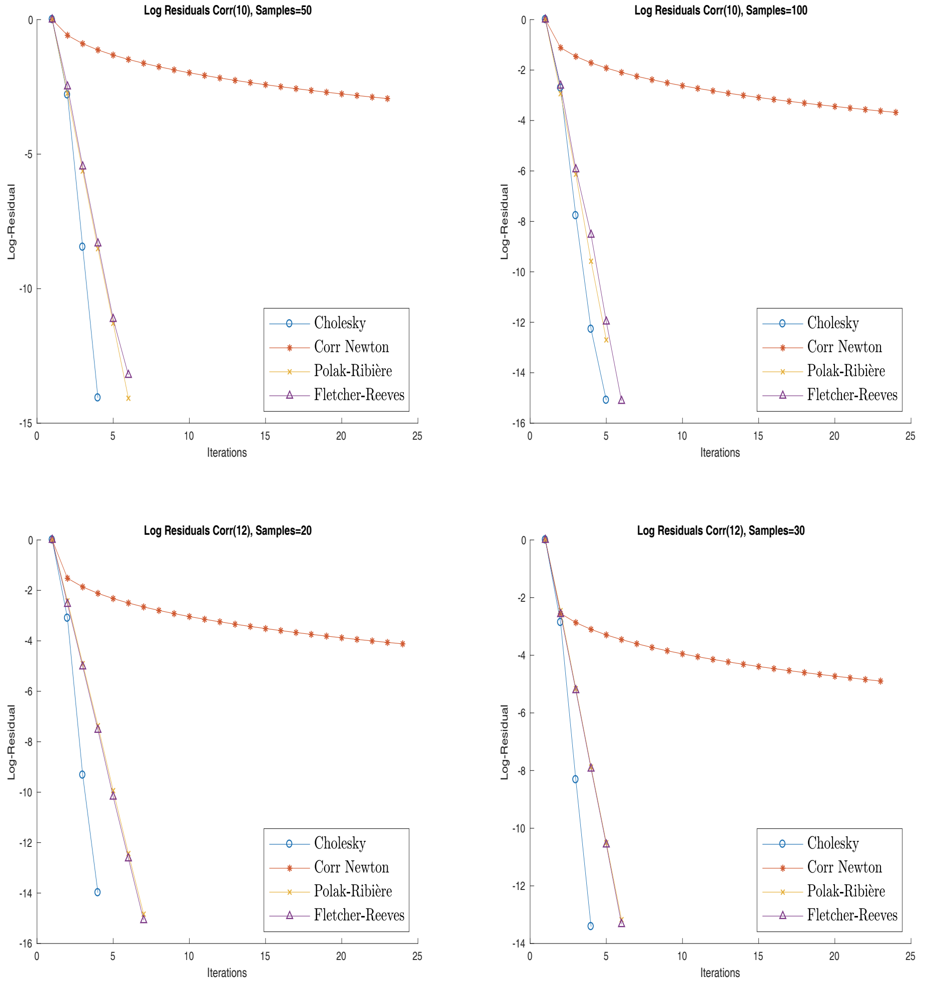

Figure 2: Relative Log-residuals for various matrix size and number of samples. The curves are given by $\ln(||\nabla F(C_i)||/||\nabla F(C_1)||)$ at iterate i. The upper left chart uses $Corr(10)$ averaging 50 samples. The upper left uses $Corr(10)$ averaging 100 samples. The lower left chart uses $Corr(12)$ averaging 20 samples. The lower right uses $Corr(12)$ averaging 30 samples.

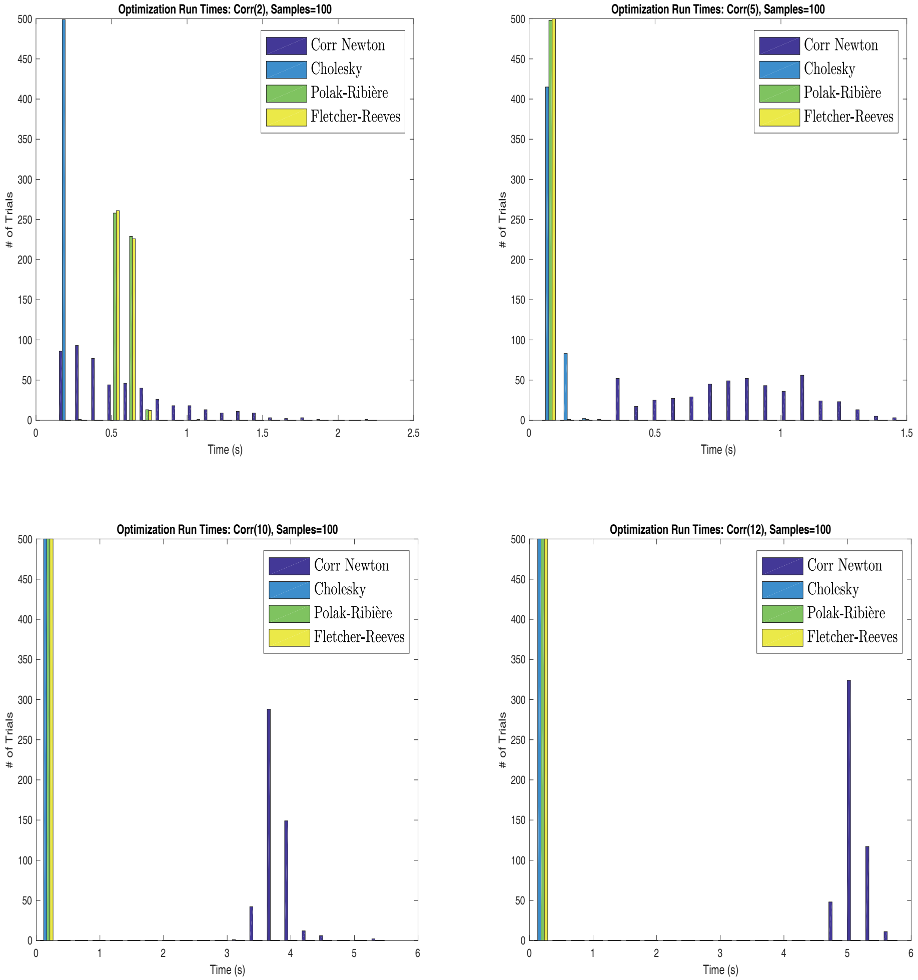

Figure 3: Run time histograms for matrices of size $n = 2, 5, 10, 12$ and 100 samples each. 500 trials were performed for each choice of matrix size and sample size. As the matrix dimension increases the run time distribution for the Corr-Newton method shifts to the right. Though the Corr-Newton method is measurably slower than the other algorithms, it still performs quickly overall.

We first expand the first term in the expression above:

$$
\begin{aligned}
&\operatorname{Hess} f_i(\Delta_1+\Delta_2,\Delta_1+\Delta_2)\\
={}&\frac{1}{3}\operatorname{Tr}\left[P_{(i)}^{-1}PP_{(i)}^{-1}PP_{(i)}^{-1}(\Delta_1+\Delta_2)P^{-1}(\Delta_1+\Delta_2)+2P_{(i)}^{-1}PP_{(i)}^{-1}(\Delta_1+\Delta_2)P_{(i)}^{-1}(\Delta_1+\Delta_2)\right]\\
&-\frac{3}{2}\operatorname{Tr}\left[P_{(i)}^{-1}PP_{(i)}^{-1}(\Delta_1+\Delta_2)P^{-1}(\Delta_1+\Delta_2)+P_{(i)}^{-1}(\Delta_1+\Delta_2)P_{(i)}^{-1}(\Delta_1+\Delta_2)\right]\\
&+3\operatorname{Tr}\left[P_{(i)}^{-1}(\Delta_1+\Delta_2)P^{-1}(\Delta_1+\Delta_2)\right]\\
={}&\frac{1}{3}\operatorname{Tr}\left[P_{(i)}^{-1}\left(PP_{(i)}^{-1}\right)^2\Delta_1P^{-1}\Delta_1\right]+\frac{1}{3}\operatorname{Tr}\left[P_{(i)}^{-1}\left(PP_{(i)}^{-1}\right)^2\Delta_1P^{-1}\Delta_2\right]\\
&+\frac{1}{3}\operatorname{Tr}\left[P_{(i)}^{-1}\left(PP_{(i)}^{-1}\right)^2\Delta_2P^{-1}\Delta_1\right]+\frac{1}{3}\operatorname{Tr}\left[P_{(i)}^{-1}\left(PP_{(i)}^{-1}\right)^2\Delta_2P^{-1}\Delta_2\right]\\
&+\frac{2}{3}\operatorname{Tr}\left[P_{(i)}^{-1}PP_{(i)}^{-1}\Delta_1P_{(i)}^{-1}\Delta_1\right]+\frac{2}{3}\operatorname{Tr}\left[P_{(i)}^{-1}PP_{(i)}^{-1}\Delta_1P_{(i)}^{-1}\Delta_2\right]\\
&+\frac{2}{3}\operatorname{Tr}\left[P_{(i)}^{-1}PP_{(i)}^{-1}\Delta_2P_{(i)}^{-1}\Delta_1\right]+\frac{2}{3}\operatorname{Tr}\left[P_{(i)}^{-1}PP_{(i)}^{-1}\Delta_2P_{(i)}^{-1}\Delta_2\right]\\
&-\frac{3}{2}\operatorname{Tr}\left[P_{(i)}^{-1}PP_{(i)}^{-1}\Delta_1P^{-1}\Delta_1\right]-\frac{3}{2}\operatorname{Tr}\left[P_{(i)}^{-1}PP_{(i)}^{-1}\Delta_1P^{-1}\Delta_2\right]\\
&-\frac{3}{2}\operatorname{Tr}\left[P_{(i)}^{-1}PP_{(i)}^{-1}\Delta_2P^{-1}\Delta_1\right]-\frac{3}{2}\operatorname{Tr}\left[P_{(i)}^{-1}PP_{(i)}^{-1}\Delta_2P^{-1}\Delta_2\right]\\
&-\frac{3}{2}\operatorname{Tr}\left[P_{(i)}^{-1}\Delta_1P_{(i)}^{-1}\Delta_1\right]-\frac{3}{2}\operatorname{Tr}\left[P_{(i)}^{-1}\Delta_1P_{(i)}^{-1}\Delta_2\right]\\
&-\frac{3}{2}\operatorname{Tr}\left[P_{(i)}^{-1}\Delta_2P_{(i)}^{-1}\Delta_1\right]-\frac{3}{2}\operatorname{Tr}\left[P_{(i)}^{-1}\Delta_2P_{(i)}^{-1}\Delta_2\right]\\
&+3\operatorname{Tr}\left[P_{(i)}^{-1}\Delta_1P^{-1}\Delta_1\right]+3\operatorname{Tr}\left[P_{(i)}^{-1}\Delta_1P^{-1}\Delta_2\right]\\
&+3\operatorname{Tr}\left[P_{(i)}^{-1}\Delta_2P^{-1}\Delta_1\right]+3\operatorname{Tr}\left[P_{(i)}^{-1}\Delta_2P^{-1}\Delta_2\right].
\end{aligned}
$$

Next we expand the second term:

$$
\begin{aligned}
&\operatorname{Hess} f_i(\Delta_1-\Delta_2,\Delta_1-\Delta_2)\\
={}&\frac{1}{3}\operatorname{Tr}\left[P_{(i)}^{-1}PP_{(i)}^{-1}PP_{(i)}^{-1}(\Delta_1-\Delta_2)P^{-1}(\Delta_1-\Delta_2)+2P_{(i)}^{-1}PP_{(i)}^{-1}(\Delta_1-\Delta_2)P_{(i)}^{-1}(\Delta_1-\Delta_2)\right]\\
&-\frac{3}{2}\operatorname{Tr}\left[P_{(i)}^{-1}PP_{(i)}^{-1}(\Delta_1-\Delta_2)P^{-1}(\Delta_1-\Delta_2)+P_{(i)}^{-1}(\Delta_1-\Delta_2)P_{(i)}^{-1}(\Delta_1-\Delta_2)\right]\\
&+3\operatorname{Tr}\left[P_{(i)}^{-1}(\Delta_1-\Delta_2)P^{-1}(\Delta_1-\Delta_2)\right]\\
={}&\frac{1}{3}\operatorname{Tr}\left[P_{(i)}^{-1}\left(PP_{(i)}^{-1}\right)^2\Delta_1P^{-1}\Delta_1\right]-\frac{1}{3}\operatorname{Tr}\left[P_{(i)}^{-1}\left(PP_{(i)}^{-1}\right)^2\Delta_1P^{-1}\Delta_2\right]\\
&-\frac{1}{3}\operatorname{Tr}\left[P_{(i)}^{-1}\left(PP_{(i)}^{-1}\right)^2\Delta_2P^{-1}\Delta_1\right]+\frac{1}{3}\operatorname{Tr}\left[P_{(i)}^{-1}\left(PP_{(i)}^{-1}\right)^2\Delta_2P^{-1}\Delta_2\right]\\
&+\frac{2}{3}\operatorname{Tr}\left[P_{(i)}^{-1}PP_{(i)}^{-1}\Delta_1P_{(i)}^{-1}\Delta_1\right]-\frac{2}{3}\operatorname{Tr}\left[P_{(i)}^{-1}PP_{(i)}^{-1}\Delta_1P_{(i)}^{-1}\Delta_2\right]
\end{aligned}
$$

$$
\begin{aligned}
&-\frac{2}{3}\operatorname{Tr}\left[P_{(i)}^{-1}PP_{(i)}^{-1}\Delta_2P_{(i)}^{-1}\Delta_1\right]+\frac{2}{3}\operatorname{Tr}\left[P_{(i)}^{-1}PP_{(i)}^{-1}\Delta_2P_{(i)}^{-1}\Delta_2\right]\\
&-\frac{3}{2}\operatorname{Tr}\left[P_{(i)}^{-1}PP_{(i)}^{-1}\Delta_1P^{-1}\Delta_1\right]+\frac{3}{2}\operatorname{Tr}\left[P_{(i)}^{-1}PP_{(i)}^{-1}\Delta_1P^{-1}\Delta_2\right]\\
&+\frac{3}{2}\operatorname{Tr}\left[P_{(i)}^{-1}PP_{(i)}^{-1}\Delta_2P^{-1}\Delta_1\right]-\frac{3}{2}\operatorname{Tr}\left[P_{(i)}^{-1}PP_{(i)}^{-1}\Delta_2P^{-1}\Delta_2\right]\\
&-\frac{3}{2}\operatorname{Tr}\left[P_{(i)}^{-1}\Delta_1P_{(i)}^{-1}\Delta_1\right]+\frac{3}{2}\operatorname{Tr}\left[P_{(i)}^{-1}\Delta_1P_{(i)}^{-1}\Delta_2\right]\\
&+\frac{3}{2}\operatorname{Tr}\left[P_{(i)}^{-1}\Delta_2P_{(i)}^{-1}\Delta_1\right]-\frac{3}{2}\operatorname{Tr}\left[P_{(i)}^{-1}\Delta_2P_{(i)}^{-1}\Delta_2\right]\\
&+3\operatorname{Tr}\left[P_{(i)}^{-1}\Delta_1P^{-1}\Delta_1\right]-3\operatorname{Tr}\left[P_{(i)}^{-1}\Delta_1P^{-1}\Delta_2\right]\\
&-3\operatorname{Tr}\left[P_{(i)}^{-1}\Delta_2P^{-1}\Delta_1\right]+3\operatorname{Tr}\left[P_{(i)}^{-1}\Delta_2P^{-1}\Delta_2\right].
\end{aligned}
$$

Subtracting one term from the other and dividing by 4 yield the Hessian on cross diagonal terms:

$$
\begin{aligned}
\operatorname{Hess} f_i(\Delta_1,\Delta_2) = {}&\frac{1}{6}\operatorname{Tr}\left[P_{(i)}^{-1}\left(PP_{(i)}^{-1}\right)^2\Delta_1P^{-1}\Delta_2\right]+\frac{1}{6}\operatorname{Tr}\left[P_{(i)}^{-1}\left(PP_{(i)}^{-1}\right)^2\Delta_2P^{-1}\Delta_1\right]\\
&+\frac{1}{3}\operatorname{Tr}\left[P_{(i)}^{-1}PP_{(i)}^{-1}\Delta_1P_{(i)}^{-1}\Delta_2\right]+\frac{1}{3}\operatorname{Tr}\left[P_{(i)}^{-1}PP_{(i)}^{-1}\Delta_2P_{(i)}^{-1}\Delta_1\right]\\
&-\frac{3}{4}\operatorname{Tr}\left[P_{(i)}^{-1}PP_{(i)}^{-1}\Delta_1P^{-1}\Delta_2\right]-\frac{3}{4}\operatorname{Tr}\left[P_{(i)}^{-1}PP_{(i)}^{-1}\Delta_2P^{-1}\Delta_1\right]\\
&-\frac{3}{4}\operatorname{Tr}\left[P_{(i)}^{-1}\Delta_1P_{(i)}^{-1}\Delta_2\right]-\frac{3}{4}\operatorname{Tr}\left[P_{(i)}^{-1}\Delta_2P_{(i)}^{-1}\Delta_1\right]\\
&+\frac{3}{2}\operatorname{Tr}\left[P_{(i)}^{-1}\Delta_1P^{-1}\Delta_2\right]+\frac{3}{2}\operatorname{Tr}\left[P_{(i)}^{-1}\Delta_2P^{-1}\Delta_1\right].
\end{aligned}
$$

In order to realize this as a Riemannian Hessian we have to take the affine-invariant metric into account. Considering the first term, we find

$$
\begin{aligned}
\frac{1}{6}\operatorname{Tr}\left[P_{(i)}^{-1}\left(PP_{(i)}^{-1}\right)^2\Delta_1P^{-1}\Delta_2\right] &= \frac{1}{6}\operatorname{Tr}\left[P_{(i)}^{-1}\left(PP_{(i)}^{-1}\right)^2PP^{-1}\Delta_1P^{-1}\Delta_2\right]\\
&= \frac{1}{6}\operatorname{Tr}\left[\left(P_{(i)}^{-1}P\right)^3P^{-1}\Delta_1P^{-1}\Delta_2\right]\\
&= \frac{1}{6}\operatorname{vec}\left(\Delta_1P^{-1}\right)^T\left(\left(PP_{(i)}^{-1}\right)^3\otimes I\right)\operatorname{vec}\left(P^{-1}\Delta_2\right)\\
&\to \frac{1}{6}\left(\left(PP_{(i)}^{-1}\right)^3\otimes I\right)
\end{aligned}
$$

where the last term is what we extract in finding a stand-alone expression for the Hessian. We omit the computations for the other terms since the steps are identical. Note that if $\Delta_i \in T_PSPD(n)$ then $P^{-1}\Delta_i$ will be symmetric.

REFERENCES

[1] KHALED ALYANI, MARCO CONGEDO, AND MAHER MOAKHER, *Diagonality measures of Hermitian positive-definite matrices with application to the approximate joint diagonalization problem*, Linear Algebra and its Applications, 528:290–320, 2017.

[2] VINCENT ARSIGNY, PIERRE FILLARD, XAVIER PENNEC AND NICHOLAS AYACHE, *Fast and simple calculus on tensors in the log-euclidean framework*, Lecture Notes in Computer Science Medical Image Computing and Computer-Assisted Intervention - MICCAI 2005, page 115–122, 2005.

[3] VINCENT ARSIGNY, PIERRE FILLARD, XAVIER PENNEC, AND NICHOLAS AYACHE, *Geometric means in a novel vector space structure on symmetric positive-definite matrices*, SIAM Journal on Matrix Analysis and Applications, 29(1):328–347, 2007.

[4] Z. BAI, G. SLEIJPEN, H. VAN DER VORST, R. LIPPERT, AND A. EDELMAN, *Nonlinear eigenvalue problems with orthogonality constraints*, Templates for the Solution of Algebraic Eigenvalue Problems, page 281–314, 2000.

[5] ALAN EDELMAN, TOMÁS A. ARIAS, AND STEVEN T. SMITH, *The geometry of algorithms with orthogonality constraints*, SIAM Journal on Matrix Analysis and Applications, 20(2):303–353, 1998.

[6] WOLFGANG FÖRSTNER AND BOUDEWIJN MOONEN, *A metric for covariance matrices*, Geodesy-The Challenge of the 3rd Millennium, page 299–309, 2003.

[7] IGOR GRUBIŠIĆ AND RAOUL PIETERSZ, *Efficient rank reduction of correlation matrices*, SSRN Electronic Journal, 2005.

[8] MEHRTASH T. HARANDI, RICHARD HARTLEY, BRIAN LOVELL, AND CONRAD SANDERSON, *Sparse coding on symmetric positive definite manifolds using Bregman divergences*, IEEE Transactions on Neural Networks and Learning Systems, 27(6):1294–1306, 2016.

[9] MEHRTASH T. HARANDI, CONRAD SANDERSON, RICHARD HARTLEY, AND BRIAN C. LOVELL, *Sparse coding and dictionary learning for symmetric positive definite matrices: A kernel approach*, Computer Vision - ECCV 2012 Lecture Notes in Computer Science, page 216–229, 2012.

[10] INBAL HOREV, FLORIAN YGER, AND MASAHI SUGIYAMA, *Geometry-aware principal component analysis for symmetric positive definite matrices*, 2015 ACML Conference Proceedings, 2015.

[11] STEPHAN HUCKEMANN, THOMAS HOTZ AND AXEL MUNK, *Intrinsic shape analysis: Geodesic pca for Riemannian manifold modulo isometric Lie group actions*, Statistica Sinica.

[12] SADEEP JAYASUMANA, RICHARD HARTLEY, MATHIEU SALZMANN, HONGDONG LI, AND MEHRTASH HARANDI, *Kernel methods on the Riemannian manifold of symmetric positive definite matrices*, 2013 IEEE Conference on Computer Vision and Pattern Recognition, 2013.

[13] TAKOUA KEFI, RIADH KSANTINI, MOHAMED BÉCHA KAÂNICHE, AND ADEL BOUHOULA, *A novel incremental covariance-guided one-class support vector machine*, Machine Learning and Knowledge Discovery in Databases Lecture Notes in Computer Science, page 17–32, 2016.

[14] HYUNWOO J. KIM, NAGESH ADLURU, BARBARA B. BENDLIN, STERLING C. JOHNSON, BABA C. VEMURI AND VIKAS SINGH, *Canonical correlation analysis on spd(n) manifolds* Riemannian Computing in Computer Vision, page 69–100, 2016.

[15] J. M LEE, *Introduction to Smooth Manifolds*, Springer, 2012.

[16] CHRISTOPHE LENGLET, MIKAËL ROUSSON, RACHID DERICHE AND OLIVIER FAUGERAS, *Statistics on the manifold of multivariate normal distributions: Theory and application to diffusion tensor MRI processing*, Journal of Mathematical Imaging and Vision, 25(3):423–444, 2006.

[17] MAHER MOAKHER, *A differential geometric approach to the geometric mean of symmetric positive-definite matrices*, SIAM Journal on Matrix Analysis and Applications, 26(3):735–747, 2005.

[18] MAHER MOAKHER AND PHILIPP G. BATCHELOR, *Symmetric positive-definite matrices: From geometry to applications and visualization*, Mathematics and Visualization Visualization and Processing of Tensor Fields, page 285–298, 2006.

[19] MAHER MOAKHER AND MOURAD ZÉRAÏ, *The Riemannian geometry of the space of positive-definite matrices and its application to the regularization of positive-definite matrix-valued data*, Journal of Mathematical Imaging and Vision, 40(2):171–187, 2010.

[20] XAVIER PENNEC, *Statistical computing on manifolds: From Riemannian geometry to computational anatomy*, Emerging Trends in Visual Computing Lecture Notes in Computer Science, page 347–386, 2009.

[21] XAVIER PENNEC, PIERRE FILLARD, AND NICHOLAS AYACHE, *A Riemannian framework for tensor computing*, International Journal of Computer Vision, 66(1):41–66, 2006.

[22] WOLFGANG RING AND BENEDIKT WIRTH, *Optimization methods on Riemannian manifolds and their application to shape space*, SIAM Journal on Optimization, 22(2):596–627, 2012.

[23] SALEM SAID, LIONEL BOMBRUN, YANNICK BERTHOUMIEU, AND JONATHAN H. MANTON, *Riemannian Gaussian distributions on the space of symmetric positive definite matrices*, IEEE Transactions on Information Theory, 63(4):2153–2170, 2017.

[24] HIROYUKI SATO AND TOSHIHIRO IWAI, *A new, globally convergent Riemannian conjugate gradient method*, Optimization, 64(4):1011–1031, 2013.

(Received March 22, 2018)

Paul David
Claremont Graduate University

Weiqing Gu
Harvey Mudd College