Optimization on Manifolds: Gaussian Mixture Models

Thomas RENARD, Benoît MÜLLER

EPFL – MATH-512 – Project 1

April 2023 (last compiled October 4, 2023)

An exemple of sentence with a reference to

1 Gaussian Mixture Models

1.1 A Riemannian geometry for $\mathcal{P}_d$ and $\mathbb{R}^d \times \mathcal{P}_d$

1. Let's rewrite l:

$$\begin{aligned} l(\mu, \Sigma) &= -\sum_{i=1}^{n} \log(p_d(\mu, \Sigma; x_i)) \\ &= -\sum_{i=1}^{n} \log\left(\big(\det(\Sigma)(2\pi)^d \big)^{-1/2} \exp\big(-\frac{1}{2}(x_i-\mu)^\top \Sigma^{-1}(x_i-\mu) \big) \right) \\ &= -\sum_{i=1}^{n} \left(-\frac{1}{2} \log\big(\det(\Sigma)(2\pi)^d \big) - \frac{1}{2}(x_i-\mu)^\top \Sigma^{-1}(x_i-\mu) \right) \\ &= \frac{n}{2} \log\big(\det(\Sigma)(2\pi)^d \big) + \frac{1}{2} \sum_{i=1}^{n} (x_i-\mu)^\top \Sigma^{-1}(x_i-\mu) \end{aligned}$$

With respect to μ, l is $\|\Sigma\|_2$-strongly convex, so let's look for stationary points. With respect to Frobenius scalar product,

$$\nabla_{\hat{\mu}} l(\hat{\mu}, \Sigma) = \frac{1}{2} \sum_{i=1}^{n} \Sigma^{-1}(\hat{\mu} - x_i) = \frac{1}{2}\Sigma^{-1} \sum_{i=1}^{n} (\hat{\mu} - x_i) = \frac{1}{2}\Sigma^{-1}(n\hat{\mu} - \sum_{i=1}^{n} x_i) = 0$$

if and only if $\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} x_i$, and notice that actually this minimizer is unique and doesn't depends on Σ.

To find an optimal Σ, lets actually look for $A = \Sigma^{-1}$ since the expression simplifies to

$$\frac{n}{2} \log\big(\det(A)^{-1}(2\pi)^d \big) + \frac{1}{2} \sum_{i=1}^{n} (x_i-\mu)^\top A(x_i-\mu) = -\frac{n}{2} \log\big(\det(A) \big) + \frac{dn}{2} \log(2\pi) + \frac{1}{2} \sum_{i=1}^{n} (x_i-\mu)^\top A(x_i-\mu)$$

Since it is known that $A \mapsto -\det(A)^{1/d}$ is convex over $\mathcal{P}_d$ [*trouver une référence*], and $x \mapsto -\log(-x)$ is strictly convex and increasing, $-d\log(\det(A)^{1/d}) = -\log(\det(A))$ is strictly convex, and sumed with a linear term, l is convex in A. Lets look at stationary points. We can derive a formula of the derivative of det from the Laplace expansion $\det(A) = \sum_{k=1}^{n} A_{ik}\operatorname{cof}(A)_{ik}$:

$$\partial_{A_{ij}} \det(A) = \operatorname{cof}(A)_{ij},$$

so $\nabla \det = \operatorname{cof}$ and by chain rule with log,

$$\nabla(\log \circ \det)(A) = \det(A)^{-1}\operatorname{cof}(A) = A^{-\top} = A^{-1}.$$

For the linear term, wee see that for any vector y,

$$\nabla_{A_{ij}}(y^\top A y) = \nabla_{A_{ij}}(\sum_{i,j=1}^{n} y_i A_{ij} y_j) = y_i y_j = (yy^\top)_{ij},$$

so at the end we get

$$\nabla_A\left(-\frac{n}{2}\log\big(\det(A)\big)+\frac{dn}{2}\log(2\pi)+\frac{1}{2}\sum_{i=1}^{n}(x_i-\mu)^\top A(x_i-\mu)\right) = -\frac{n}{2}A^{-1}+\frac{1}{2}\sum_{i=1}^{n}(x_i-\mu)(x_i-\mu)^\top = 0.$$

If $\sum_{i=1}^{n}(x_i-\mu)(x_i-\mu)^\top$ is invertible (never the case when $n < d$), then the unique optimal is

$$\hat{\Sigma} = \frac{1}{n}\sum_{i=1}^{n}(x_i-\mu)(x_i-\mu)^\top.$$

2. The embedded manifold $\mathcal{M}$ is an open subset of its embedding euclidean space $\mathcal{E}$, so it is locally identifiable to $\mathcal{E}$, and all directions are in the tangent space, so $T_x\mathcal{M} = \mathcal{E} = \mathbb{R}^d \times \mathrm{Sym}_d$.

3. Let $f(\Theta) = \log p_d(\Theta; x)$ so we have

$$f(\mu,\Sigma) = \log p_d(\mu,\Sigma;x) = -\frac{1}{2}\log\det(\Sigma) - \frac{d}{2}\log(2*\pi) - \frac{1}{2}(\mu-x)^\top\Sigma^{-1}(\mu-x)$$

and let $(\Theta,\dot{\Theta}) \in T\mathcal{M}$. We first compute

$$\begin{aligned}\langle\nabla^2 f(\Theta)[\dot{\Theta}],\dot{\Theta}\rangle &= \frac{d^2}{dt^2}f(\Theta+t\dot{\Theta})|_{t=0}\\ &= \frac{d^2}{dt^2}f(\mu+t\dot{\mu},\Sigma+t\dot{\Sigma})|_{t=0}\end{aligned}$$

and will generalize for two different directions Θ after. We use the formulas for the derivatives of $\log \circ \log$ and the matrix inversion found in Question 1, and the product derivative formula:

$$
\begin{aligned}
f(\mu + t\dot{\mu}, \Sigma(t)) &= \log(\det(\Sigma + t\dot{\Sigma})) + d\log(2\pi) + (\mu(t) - x)^\top \Sigma(t)^{-1}(\mu(t) - x) \\
&\xmapsto{d/dt} \langle \Sigma(t)^{-1}, \dot{\Sigma} \rangle + \dot{\mu}^\top \Sigma(t)^{-1}(\mu(t) - x) + (\mu(t) - x)^\top \Sigma(t)^{-1}\dot{\Sigma}\Sigma(t)^{-1}(\mu(t) - x) + (\mu(t) - x)^\top \Sigma(t)^{-1}\dot{\mu} \\
&\xmapsto{d/dt} \langle \Sigma(t)^{-1}\dot{\Sigma}\Sigma(t)^{-1}, \dot{\Sigma} \rangle \\
&\quad + \dot{\mu}^\top \frac{d}{dt}\left(\Sigma(t)^{-1}\right)(\mu(t) - x) + \dot{\mu}^\top \Sigma(t)^{-1}\dot{\mu} \\
&\quad + \dot{\mu}^\top \Sigma(t)^{-1}\dot{\Sigma}\Sigma(t)^{-1}(\mu(t) - x) + (\mu(t) - x)^\top \frac{d}{dt}\left(\Sigma(t)^{-1}\dot{\Sigma}\Sigma(t)^{-1}\right)(\mu(t) - x) \\
&\quad + (\mu(t) - x)^\top \Sigma(t)^{-1}\dot{\Sigma}\Sigma(t)^{-1}\dot{\mu} \\
&\quad + \dot{\mu}^\top \Sigma(t)^{-1}\dot{\mu} \\
&\xmapsto{t=0} \langle \Sigma^{-1}\dot{\Sigma}\Sigma^{-1}, \dot{\Sigma} \rangle \\
&\quad + \dot{\mu}^\top \frac{d}{dt}\left(\Sigma(t)^{-1}\right)|_{t=0}(\mu - x) + \dot{\mu}^\top \Sigma^{-1}\dot{\mu} \\
&\quad + \dot{\mu}^\top \Sigma^{-1}\dot{\Sigma}\Sigma^{-1}(\mu - x) + (\mu - x)^\top \frac{d}{dt}\left(\Sigma(t)^{-1}\dot{\Sigma}\Sigma(t)^{-1}\right)|_{t=0}(\mu - x) + (\mu - x)^\top \Sigma^{-1}\dot{\Sigma}\Sigma^{-1}\dot{\mu} \\
&\quad + \dot{\mu}^\top \Sigma^{-1}\dot{\mu} \\
&\xmapsto{\mathbb{E}_x} \langle \Sigma^{-1}\dot{\Sigma}\Sigma^{-1}, \dot{\Sigma} \rangle + 2\dot{\mu}^\top \Sigma^{-1}\dot{\mu} + \mathbb{E}_x((\mu - x)^\top \frac{d}{dt}\left(\Sigma(t)^{-1}\dot{\Sigma}\Sigma(t)^{-1}\right)|_{t=0}(\mu - x))
\end{aligned}
$$

and we can rewrite

$$\langle \Sigma^{-1}\dot{\Sigma}\Sigma^{-1}, \dot{\Sigma} \rangle + 2\dot{\mu}^\top \Sigma^{-1}\dot{\mu} = \mathrm{Tr}(\Sigma^{-1}\dot{\Sigma}\Sigma^{-1}\dot{\Sigma}) + 2\dot{\mu}^\top \Sigma^{-1}\dot{\mu}.$$

Both formulas of the Fisher-Rao information metric are clearly symmetric bilinear maps, and using polarization identity, they are entirely determined by the quadratic map associated, which have been proven to be equal, so the whole bilinear maps are actually equal.

4. We start by showing that $\langle \cdot, \cdot \rangle_{FR}$ is a metric on $\mathcal{M} = \mathbb{R}^d \times \mathcal{P}_d$, and to do so, we will show that for any choice of $\Theta = (\mu, \Sigma) \in \mathcal{M}$, $\langle \cdot, \cdot \rangle_\Theta$ is an inner product on $T_\theta \mathcal{M}$.
We start by showing the symmetry. Let $\dot{\Theta}_1 = (\dot{\mu}_1, \dot{\Sigma}_1), \dot{\Theta}_2 = (\dot{\mu}_2, \dot{\Sigma}_2) \in T_\Theta \mathcal{M}$. We have :

$$
\begin{aligned}
\langle \dot{\Theta}_1, \dot{\Theta}_2 \rangle_\Theta &= Tr(\Sigma^{-1}\dot{\Sigma}_1\Sigma^{-1}\dot{\Sigma}_2) + 2\dot{\mu}_1^T\Sigma^{-1}\dot{\mu}_2 \\
&= Tr(\Sigma^{-1}\dot{\Sigma}_2\Sigma^{-1}\dot{\Sigma}_1) + 2\dot{\mu}_2^T\Sigma^{-1}\dot{\mu}_1 \\
&= \langle \dot{\Theta}_2, \dot{\Theta}_1 \rangle_\Theta
\end{aligned}
$$

where we used the permutation property of the trace since we are only working with symmetric matrices, and the fact that $\Sigma \in \mathcal{P}_d$, so Σ^{-1} is symmetric.
For the bi-linearity, we have $\forall a, b \in \mathbb{R}$, and $\dot{\Theta}_3 = (\dot{\mu}_3, \dot{\Sigma}_3) \in T_\Theta \mathcal{M}$:

$$
\begin{aligned}
\langle a\dot{\Theta}_1 + b\dot{\Theta}_2, \dot{\Theta}_3 \rangle_\Theta &= Tr(\Sigma^{-1}(a\dot{\Sigma}_1 + b\dot{\Sigma}_2)\Sigma^{-1}\dot{\Sigma}_3) + 2(a\dot{\mu}_1 + b\dot{\mu}_2)^T\Sigma^{-1}\dot{\mu}_3 \\
&= Tr(a\Sigma^{-1}\dot{\Sigma}_1\Sigma^{-1}\dot{\Sigma}_3 + b\Sigma^{-1}\dot{\Sigma}_2\Sigma^{-1}\dot{\Sigma}_3) + 2a\dot{\mu}_1^T\Sigma^{-1}\dot{\mu}_3 + 2b\dot{\mu}_2^T\Sigma^{-1}\dot{\mu}_3 \\
&= aTr(\Sigma^{-1}\dot{\Sigma}_1\Sigma^{-1}\dot{\Sigma}_3) + bTr(\Sigma^{-1}\dot{\Sigma}_2\Sigma^{-1}\dot{\Sigma}_3) + 2a\dot{\mu}_1^T\Sigma^{-1}\dot{\mu}_3 + 2b\dot{\mu}_2^T\Sigma^{-1}\dot{\mu}_3 \\
&= a(Tr(\Sigma^{-1}\dot{\Sigma}_1\Sigma^{-1}\dot{\Sigma}_3) + 2\dot{\mu}_1^T\Sigma^{-1}\dot{\mu}_3) + b(Tr(\Sigma^{-1}\dot{\Sigma}_2\Sigma^{-1}\dot{\Sigma}_3) + 2\dot{\mu}_2^T\Sigma^{-1}\dot{\mu}_3) \\
&= a\langle \dot{\Theta}_1, \dot{\Theta}_3 \rangle_\Theta + b\langle \dot{\Theta}_2, \dot{\Theta}_3 \rangle_\Theta
\end{aligned}
$$

where we used the linearity of the trace. Moreover, the linearity in the second argument follows by symmetry.
Now, for the positive-definiteness, if $\dot{\Theta} = (\dot{\mu}, \dot{\Sigma}) \in T_\Theta \mathcal{M}$ is non-zero, then observe that :

$$\begin{aligned}\langle \dot{\Theta}, \dot{\Theta} \rangle_\Theta &= Tr(\Sigma^{-1}\dot{\Sigma}\Sigma^{-1}\dot{\Sigma}) + 2\dot{\mu}^T\Sigma^{-1}\dot{\mu} \\ &= Tr((\Sigma^{-1}\dot{\Sigma})^T\Sigma^{-1}\dot{\Sigma}) + 2\dot{\mu}^T\Sigma^{-1}\dot{\mu}\end{aligned}$$

Now, since we already now from previous exercise sessions that the trace defines a standard inner product on matrix spaces, it follows that $Tr((\Sigma^{-1}\dot{\Sigma})^T\Sigma^{-1}\dot{\Sigma})$ is non-negative, and the same holds for $\dot{\mu}^T\Sigma^{-1}\dot{\mu}$, since $\Sigma \in \mathcal{P}_d$, and so is Σ^{-1}. Now, since $\dot{\Theta}$ is chosen to be non-zero, it follows, that either $\dot{\mu}$ or $\dot{\Sigma}$ is non-zero, and therefore, one of the term in the sum has to be strictly positive, therefore, $\langle \dot{\Theta}, \dot{\Theta} \rangle_\Theta > 0$ as wanted.
We now show that the FR metric on $\mathcal{M}$ is a riemannian metric. To do so, consider V, W two smooth vector fields on $\mathcal{M}$ and consider the map $F : \mathcal{M} \to \mathbb{R}$ defined by $F(\Theta) = \langle V(\Theta), W(\Theta) \rangle_\Theta$, for $\Theta = (\mu, \Sigma) \in \mathcal{M}$. Since both V and W are smooth vector fields on $\mathcal{M}$, by proposition 3.45 in the textbook, there exists smooth vector fields $\bar{V}, \bar{W}$ defined on open neighborhoods of $\mathcal{M}$, say U_1 and U_2 respectively, such that $V = \bar{V}|_\mathcal{M}$ and $W = \bar{W}|_\mathcal{M}$. Now observe that $U = U_1 \cap U_2 \supseteq \mathcal{M}$ is an open neighborhood of $\mathcal{M}$ in $\mathcal{E}$. We now define the map $\bar{F} : U \to \mathbb{R}$ defined by $\bar{F}(\Theta) = \langle \bar{V}(\Theta), \bar{W}(\Theta) \rangle_\Theta$, for $\Theta = (\mu, \Sigma) \in U$ and claim that this is a smooth extension of F on an open neighborhood U of $\mathcal{M}$ in $\mathcal{E}$.
Indeed, we start by decomposing $\bar{V} = (\dot{\mu}_{\bar{V}}, \dot{\Sigma}_{\bar{V}})$, so that $\bar{V}(\Theta) = (\dot{\mu}_{\bar{V}}(\Theta), \dot{\Sigma}_{\bar{V}}(\Theta)) \in T_\Theta\mathcal{M}, \forall \Theta \in \mathcal{M}$. Since $\bar{V}$ is smooth, it follows that both $\dot{\mu}_{\bar{V}}$ and $\dot{\Sigma}_{\bar{V}}$ are smooth maps. We analogously decompose $\bar{W} = (\dot{\mu}_{\bar{W}}, \dot{\Sigma}_{\bar{W}})$, and see that both components are smooth maps. Then, $\bar{F}(\Theta) = \langle \bar{V}(\Theta), \bar{W}(\Theta) \rangle_\Theta = Tr(\Sigma^{-1}\dot{\Sigma}_{\bar{V}}(\Theta)\Sigma^{-1}\dot{\Sigma}_{\bar{W}}(\Theta)) + 2\dot{\mu}_{\bar{V}}(\Theta)^T\Sigma^{-1}\dot{\mu}_{\bar{W}}(\Theta)$.
Observe first that the map $\Theta \mapsto \Sigma^{-1}$ is smooth since it is the composition of the projection map $\Theta \mapsto \Sigma$ with the inverse map $\Sigma \mapsto \Sigma^{-1}$, both maps being smooth, and smoothness being preserved by composition. Indeed, the first map is clearly smooth, and for the second, it will be shown in details in question 6 why it is smooth.
Now, from the section 4.7 in the texbook, by the product rule for differentials, the product of two maps from a manifold to matrix spaces such that the matrix multiplication is always well-defined, is smooth too. It follows that $\Theta \mapsto \Sigma^{-1}\dot{\Sigma}_{\bar{V}}(\Theta)\Sigma^{-1}\dot{\Sigma}_{\bar{W}}(\Theta)$ is smooth, as well as $\Theta \mapsto \dot{\mu}_{\bar{V}}(\Theta)^T\Sigma^{-1}\dot{\mu}_{\bar{W}}(\Theta)$. Moreover, the trace map is also smooth (it follows from the fact that the trace map is polynomial in the entries of the matrix given in input, and so if we consider its extension to a linear space containing the manifold on which the trace map is defined, its partial derivatives will all be C^∞, and therefore the map is smooth), therefore, the map $\Theta \mapsto Tr(\Sigma^{-1}\dot{\Sigma}_{\bar{V}}(\Theta)\Sigma^{-1}\dot{\Sigma}_{\bar{W}}(\Theta))$ is smooth, as a composition of smooth maps. It follows that $\bar{F}$ is a smooth map, as smoothness in preserved by linear combination. We can conclude that the FR metric is a Riemannian metric on $\mathcal{M}$.

5. Let $A \in \mathbb{R}^{d\times d}$ be invertible and $b \in \mathbb{R}^d$, and consider the map $\phi : \mathcal{M} = \mathbb{R}^d \times \mathcal{P}_d \to \mathcal{M}$ defined by $\phi(\mu, \Sigma) = (A\mu + b, A\Sigma A^T)$. Observe first that $\forall \Sigma \in \mathcal{P}_d, A\Sigma A^T \in \mathcal{P}_d$ since $\forall x \in \mathbb{R}^d$ non-zero, one has $x^T A\Sigma A^T x = (A^T x)\Sigma(A^T x) > 0$, since $A^T x \in \mathbb{R}^d$ is non-zero and $\Sigma \in \mathcal{P}_d$, and therefore the map ϕ is well-defined.
We first show that ϕ is bijective. To do so, we define the map $\phi^{-1} : \mathcal{M} \to \mathcal{M}$ such that $\phi^{-1}(\mu, \Sigma) = (A^{-1}(\mu - b), A^{-1}\Sigma A^{-T})$ which is well-defined since A is invertible, and with an identical argument as before, one has that $A^{-1}\Sigma A^{-T} \in \mathcal{P}_d$ whenever $\Sigma \in \mathcal{P}_d$. We now claim that ϕ^{-1} is the inverse map of

ϕ. Indeed, let $(\mu, \Sigma) \in \mathcal{M}$, then we have :

$$\begin{aligned}\phi^{-1} \circ \phi(\mu, \Sigma) &= \phi^{-1}(A\mu + b, A\Sigma A^T) \\ &= (A^{-1}(A\mu + b - b), A^{-1}(A\Sigma A^T)A^{-T}) \\ &= (\mu, \Sigma)\end{aligned}$$

and

$$\begin{aligned}\phi \circ \phi^{-1}(\mu, \Sigma) &= \phi(A^{-1}(\mu - b), A^{-1}\Sigma A^{-T}) \\ &= (A(A^{-1}(\mu - b)) + b, A(A^{-1}\Sigma A^{-T})A^T) \\ &= (\mu, \Sigma)\end{aligned}$$

We now argue that ϕ is a diffeomorphism. Indeed, ϕ is a smooth map, as one can consider the extension $\bar{\phi} : \mathbb{R}^d \times Sym_d \to \mathbb{R}^d \times Sym_d$ defined by $\bar{\phi}(\mu, \Sigma) = (A\mu + b, A\Sigma A^T)$, and see that since the left-hand side is only composed of matrix multiplications and additions, $\bar{\phi}$ is polynomial in the entries of its input, and therefore, as a map between linear spaces, its partial derivatives will all be C^∞, which implies that it is a smooth extension of ϕ. Now, we also have that ϕ^{-1} is smooth, as, similarly, one can consider the extension $\bar{\phi}^{-1} : \mathbb{R}^d \times Sym_d \to \mathbb{R}^d \times Sym_d$ defined by $\bar{\phi}^{-1}(\mu, \Sigma) = (A^{-1}(\mu - b), A^{-1}\Sigma A^{-T})$, and since A is invertible, and here agin for the same reason, it is polynomial in the entries of μ and Σ, and therefore, all its partial derivatives will be C^∞, and $\bar{\phi}^{-1}$ is a smooth extension of ϕ^{-1}.
We now want to verify that $\forall (\Theta, \dot{\Theta}_1), (\Theta, \dot{\Theta}_2) \in T\mathcal{M}$ we have $\langle \dot{\Theta}_1, \dot{\Theta}_2 \rangle_\Theta = \langle D\phi(\Theta)[\dot{\Theta}_1], D\phi(\Theta)[\dot{\Theta}_2] \rangle_{\phi(\Theta)}$
Now, decompose $\Theta = (\mu, \Sigma)$ and $\dot{\Theta}_i = (\dot{\mu}_i, \dot{\Sigma}_i), i = 1, 2$. We have for $i = 1, 2$:

$$\begin{aligned}D\phi(\Theta)[\dot{\Theta}_i] &= \lim_{t \to 0} \frac{\phi(\mu + t\dot{\mu}_i, \Sigma + t\dot{\Sigma}_i) - \phi(\mu, \Sigma)}{t} \\ &= \lim_{t \to 0} \frac{(A\mu + tA\dot{\mu}_i + b, A\Sigma A^T + tA\dot{\Sigma}_i A^T) - (A\mu + b, A\Sigma A^T)}{t} \\ &= (A\dot{\mu}_i, A\dot{\Sigma}_i A^T)\end{aligned}$$

Therefore, we can compute :

$$\begin{aligned}\langle D\phi(\Theta)[\dot{\Theta}_1], D\phi(\Theta)[\dot{\Theta}_2] \rangle_{\phi(\Theta)} &= Tr((A\Sigma A^T)^{-1} A\dot{\Sigma}_1 A^T (A\Sigma A^T)^{-1} A\dot{\Sigma}_2 A^T) + 2(A\dot{\mu}_1)^T (A\Sigma A^T)^{-1} (A\dot{\mu}_2) \\ &= Tr(A^{-T}\Sigma^{-1}A^{-1}A\dot{\Sigma}_1 A^T A^{-T}\Sigma^{-1}A^{-1}A\dot{\Sigma}_2 A^T) + 2\dot{\mu}_1^T A^T A^{-T}\Sigma^{-1}A^{-1}A\dot{\mu}_2 \\ &= Tr(A^{-T}\Sigma^{-1}\dot{\Sigma}_1\Sigma^{-1}\dot{\Sigma}_2 A^T) + 2\dot{\mu}_1^T\Sigma^{-1}\dot{\mu}_2 \\ &= Tr(\Sigma^{-1}\dot{\Sigma}_1\Sigma^{-1}\dot{\Sigma}_2 A^T A^{-T}) + 2\dot{\mu}_1^T\Sigma^{-1}\dot{\mu}_2 \\ &= Tr(\Sigma^{-1}\dot{\Sigma}_1\Sigma^{-1}\dot{\Sigma}_2) + 2\dot{\mu}_1^T\Sigma^{-1}\dot{\mu}_2 \\ &= \langle \dot{\Theta}_1, \dot{\Theta}_2 \rangle_\Theta\end{aligned}$$

where we used the fact that the trace is invariant under cyclic permutations.
Therefore, ϕ is an isometry of the Riemannian manifold $(\mathbb{R}^d \times \mathcal{P}_d, \langle \cdot, \cdot \rangle_{FR})$

6. Clearly, if we define $\phi : \mathcal{P}_d \to \mathcal{P}_d$ by $\phi(\Sigma) = \Sigma^{-1}$, then the inverse map $\phi^{-1} : \mathcal{P}_d \to \mathcal{P}_d$ exists, is well-defined (we recall that the inverse of a symmetric positive-definite matrix is a symmetric positive-definite matrix), and is just ϕ itself, showing that ϕ is bijective. We now argue that ϕ (and therefore its inverse too) is smooth. To do so, consider an extension $\bar{\phi} : GL_d(\mathbb{R}) \to GL_d(\mathbb{R})$, defined by $\bar{\phi}(\Sigma) = \Sigma^{-1}, \Sigma \in GL_d(\mathbb{R})$, where $GL_d(\mathbb{R})$ denotes the set of $d \times d$ matrices with real coefficients, that are invertible. Observe that $GL_d(\mathbb{R})$ is an open neighborhood of $\mathcal{P}_d$. Indeed, clearly, it contains

$\mathcal{P}_d$, and it is open since we have that $GL_d(\mathbb{R}) = \det^{-1}(\mathbb{R}\backslash\{0\})$ is the preimage of an open set by a continuous map (the determinant map), and is therefore open as well. Now, to show that $\bar{\phi}$ is a smooth map, we observe that $\forall \Sigma \in GL_d(\mathbb{R})$, we have $\Sigma^{-1} = \det(\Sigma)^{-1} adj(\Sigma)$, where $adj(\Sigma)$ denotes the adjucate matrix of Σ whose entries are polynomial in the coefficients of Σ. Since $\det(\Sigma)$ is also a polynomial in the coefficients of Σ, each entry of Σ^{-1} is a rational polynomial in the coefficients of Σ and therefore the map $\bar{\phi}$ admits C^∞ partial derivatives, and is therefore smooth, showing that it is a smooth extension of ϕ, and therefore ϕ is smooth. This way, we showed that ϕ is a diffeomorphism.
Recall now that we can identify the tangent spaces $T_\Sigma \mathcal{P}_d$ of $\mathcal{P}_d$ with Sym_d. We now want to show that $\forall (\Sigma, \dot{\Sigma}_1), (\Sigma, \dot{\Sigma}_2) \in \mathcal{P}_d \times Sym_d$ we have $\langle \dot{\Sigma}_1, \dot{\Sigma}_2 \rangle_\Sigma = \langle D\phi(\Sigma)[\dot{\Sigma}_1], D\phi(\Sigma)[\dot{\Sigma}_2] \rangle_{\phi(\Sigma)}$. To do so, we start by computing $D\phi(\Sigma)[\dot{\Sigma}_i]$ for $i = 1, 2$. The computation follows from the product rule. Indeed, let $\psi : \mathcal{P}_d \to \mathcal{P}_d$ be the identity map. It is clear that $(\phi \cdot \psi)(\Sigma) = \Sigma^{-1}\Sigma = I_d$. Therefore, we have :

$$\begin{aligned} 0_{d\times d} &= D(\phi \cdot \psi)(\Sigma)[\dot{\Sigma}_i] \\ &= D\phi(\Sigma)[\dot{\Sigma}_i]\psi(\Sigma) + \phi(\Sigma)D\psi(\Sigma)[\dot{\Sigma}_i] \\ &= D\phi(\Sigma)[\dot{\Sigma}_i]\Sigma + \Sigma^{-1}\dot{\Sigma}_i \\ \implies D\phi(\Sigma)[\dot{\Sigma}_i] &= -\Sigma^{-1}\dot{\Sigma}_i\Sigma^{-1} \end{aligned}$$

Where we used the fact that $0_{d\times d} = D(\phi \cdot \psi)(\Sigma)[\dot{\Sigma}_i]$ since $\phi \cdot \psi$ is a constant map, therefore its differential is zero, and $D\psi(\Sigma)[\dot{\Sigma}_i] = \dot{\Sigma}_i$, because ψ is the identity map. Therefore we have :

$$\begin{aligned} \langle D\phi(\Sigma)[\dot{\Sigma}_1], D\phi(\Sigma)[\dot{\Sigma}_2] \rangle_{\phi(\Sigma)} &= \langle -\Sigma^{-1}\dot{\Sigma}_1\Sigma^{-1}, -\Sigma^{-1}\dot{\Sigma}_2\Sigma^{-1} \rangle_{\Sigma^{-1}} \\ &= Tr((\Sigma^{-1})^{-1}(-\Sigma^{-1}\dot{\Sigma}_1\Sigma^{-1})(\Sigma^{-1})^{-1}(-\Sigma^{-1}\dot{\Sigma}_2\Sigma^{-1})) \\ &= Tr(\Sigma\Sigma^{-1}\dot{\Sigma}_1\Sigma^{-1}\Sigma\Sigma^{-1}\dot{\Sigma}_2\Sigma^{-1}) \\ &= Tr(\dot{\Sigma}_1\Sigma^{-1}\dot{\Sigma}_2\Sigma^{-1}) \\ &= Tr(\Sigma^{-1}\dot{\Sigma}_1\Sigma^{-1}\dot{\Sigma}_2) \\ &= \langle \dot{\Sigma}_1, \dot{\Sigma}_2 \rangle_\Sigma \end{aligned}$$

Showing that the map $\Sigma \mapsto \Sigma^{-1}$ is an isometry of the Riemannian manifold $(\mathcal{P}_d, \langle \cdot, \cdot \rangle_{FR})$.

7. We first show that $\mathcal{P}_{br,d+1}$ is an embedded submanifold of the manifold $\mathcal{P}_{d+1}$. To do so we consider the map $h : \mathcal{P}_{d+1} \to \mathbb{R}$ defined by $h(X) = Tr(EX) - 1 = [X]_{(d+1)\times(d+1)} - 1$, where E is defined to be the $(d+1)\times(d+1)$ matrix whose entries are all zeros, except the bottom right one, which is equal to 1. Observe that $\mathbb{R}$ is a linear space, and therefore a manifold, of dimension 1. Observe also that h is a linear map, and since we already argued that the trace map is smooth, we have that h is a smooth map too, as composition of smooth maps. We will use corollary 8.76 in the textbook to show that $\mathcal{P}_{br,d+1}$ is an embedded submanifold of the manifold $\mathcal{P}_{d+1}$. Indeed, we see that $\mathcal{P}_{br,d+1} = h^{-1}(0)$ is a non-empty level set. As for the differential of h we have for any $X \in \mathcal{P}_{d+1}, U \in Sym_{d+1}$:

$$\begin{aligned} Dh(X)[U] &= \lim_{t\to 0} \frac{h(X+tU) - h(X)}{t} \\ &= \lim_{t\to 0} \frac{tr(E(tX+U)) - 1 - tr(EX) + 1}{t} \\ &= tr(EU) \end{aligned}$$

where we used the linearity of the trace map. Clearly, whenever $X \in \mathcal{P}_{d+1}$, $Dh(X)$ is a linear map from Sym_{d+1} to $\mathbb{R}$, which is non-zero, so its image has to be whole $\mathbb{R}$ and therefore, it is of rank

$1 = \dim \mathbb{R}$. So we conclude by the Corollary 8.76 that $\mathcal{P}_{br,d+1}$ is an embedded submanifold of the manifold $\mathcal{P}_{d+1}$ and it has dimension :

$$\begin{aligned}\dim \mathcal{P}_{br,d+1} &= \dim \mathcal{P}_{d+1} - 1 \\ &= \frac{(d+1)^2 + d + 1}{2} - 1 \\ &= \frac{d(d+3)}{2}\end{aligned}$$

Indeed, recall that $\dim(T_X\mathcal{P}_{d+1}) = \dim \mathcal{P}_{d+1}$, but by the question 2, we can identify $T_X\mathcal{P}_{d+1}$ with Sym_{d+1}, which a linear space of dimension $\frac{(d+1)^2+d+1}{2}$. As for the tangent spaces, since h is a local defining function for $\mathcal{P}_{br,d+1}$, we have :

$$\begin{aligned}T_X\mathcal{P}_{br,d+1} &= \ker Dh(X) \\ &= \{A \in Sym_{d+1} | Dh(X)[A] = 0\} \\ &= \{A \in Sym_{d+1} | Tr(EA) = 0\} \\ &= \{A \in Sym_{d+1} | [A]_{(d+1)\times(d+1)} = 0\}\end{aligned}$$

Now, the fact that $\mathcal{P}_{br,d+1}$ is an embedded submanifold of Sym_{d+1} immediately follows from the question 2 of the first exercise in the exercise session 3, since we already know that $\mathcal{P}_{d+1}$ is an embedded submanifold of the linear space Sym_{d+1}, $\mathcal{P}_{br,d+1}$ is a subset of $\mathcal{P}_{d+1}$ defined as the 0 level set of smooth function $h : \mathcal{P}_{d+1} \to \mathbb{R}$, which admits differentials of constant rank $1, \forall X \in \mathcal{P}_{br,d+1}$.

8. Let $\phi : \mathbb{R}^d \times \mathcal{P}_d \to \mathcal{P}_{br,d+1}$ defined by $\phi(\mu, \Sigma) = \begin{pmatrix} \Sigma + \mu\mu^t & \mu \\ \mu^t & 1 \end{pmatrix}$. Now, define the map $\phi^{-1} : \mathcal{P}_{br,d+1} \to \mathbb{R}^d \times \mathcal{P}_d$ defined, for any $X = \begin{pmatrix} A & b \\ b^t & 1 \end{pmatrix} \in \mathcal{P}_{br,d+1}$ where $b \in \mathbb{R}^d$ and $A \in \mathcal{P}_d$ to ensure that X is indeed in $\mathcal{P}_{br,d+1}$, by $\phi^{-1}(X) = (b, A - bb^t)$. We have that this map is well defined, i.e, that $(b, A - bb^t) \in \mathbb{R}^d \times \mathcal{P}_d$. Indeed, $A - bb^t$ is symmetric since $(A - bb^t)^t = A^t - (bb^t)^t = A - bb^t$, and it is also positive-definite. Indeed, since X is positive-definite we have that $\forall z = (x, y) \in \mathbb{R}^{d+1}$ non-zero, where $x \in \mathbb{R}^d$ and $y \in \mathbb{R}$:

$$\begin{aligned} & z^t X z > 0 \\ \implies & x^t A x + x^t b y + y b^t x + y^2 > 0 \\ \implies & x^t A x + 2 y x^t b + y^2 > 0\end{aligned}$$

(where we used the developed expression for a quadratic map using block matrix representation) In particular, if we let $z = (x, -x^t b)$ for a non-zero $x \in \mathbb{R}^d$, then z is still non-zero, and therefore :

$$\begin{aligned} & x^t A x - 2(x^t b)^2 + (x^t b)^2 > 0 \\ \implies & x^t A x - (x^t b)^2 > 0 \\ \implies & x^t A x - x^t b b^t x > 0 \\ \implies & x^t (A - bb^t) x > 0\end{aligned}$$

showing that $A - bb^t$ is indeed positive-definite, and thus, ϕ^{-1} is well-defined.
We now show that ϕ^{-1} is the inverse map of ϕ, showing that ϕ is a bijection. To do so, let $(\mu, \Sigma) \in$

$\mathbb{R}^d \times \mathcal{P}_d$, then :

$$\begin{aligned}\phi^{-1} \circ \phi(\mu, \Sigma) &= \phi^{-1} \begin{pmatrix} \Sigma + \mu\mu^t & \mu \\ \mu^t & 1 \end{pmatrix} \\ &= (\mu, (\Sigma + \mu\mu^t) - \mu\mu^t) \\ &= (\mu, \Sigma)\end{aligned}$$

Now, let $X = \begin{pmatrix} A & b \\ b^t & 1 \end{pmatrix} \in \mathcal{P}_{br,d+1}$, then :

$$\begin{aligned}\phi \circ \phi^{-1}(X) &= \phi(b, A - bb^t) \\ &= \begin{pmatrix} (A - bb^t) + bb^t & b \\ b^t & 1 \end{pmatrix} \\ &= \begin{pmatrix} A & b \\ b^t & 1 \end{pmatrix} = X\end{aligned}$$

as wanted.
Now, we argue that ϕ is smooth. To do so, consider the extension $\bar{\phi} : \mathcal{M} = \mathbb{R}^d \times Sym_d \to \mathrm{Sym}_{d+1}$ defined by $\bar{\phi}(\mu, \Sigma) = \begin{pmatrix} \Sigma + \mu\mu^t & \mu \\ \mu^t & 1 \end{pmatrix}$, where we already know that $\mathcal{M}$ is a linear space containing $\mathbb{R}^d \times \mathcal{P}_d$, and similarly, Sym_{d+1} is a linear space containing $\mathcal{P}_{br,d+1}$ as a subset. Clearly, this extension is well-defined. Now, observe that the expression of $\bar{\phi}(\mu, \Sigma)$ is polynomial in the entries of the input (μ, Σ), and it follows that the partial derivatives of $\bar{\phi}$ will necessarily be C^∞, therefore, $\bar{\phi}$ is a smooth extension of ϕ, showing that ϕ is smooth.
Finally, we show that ϕ^{-1} is also a smooth map. To do so, consider the extension $\bar{\phi}^{-1} : \mathrm{Sym}_{d+1} \to \mathcal{M}$, between linear spaces as before, defined for any $X = \begin{pmatrix} A & b \\ b^t & 1 \end{pmatrix} \in Sym_{d+1}$ by $\bar{\phi}^{-1}(X) = (b, A - bb^t)$. Observe that whenever $X = \begin{pmatrix} A & b \\ b^t & 1 \end{pmatrix} \in Sym_{d+1}$, then it has to be that $b \in \mathbb{R}^d$ and $A \in Sym_d$, so that $\bar{\phi}^{-1}$ is a well-defined map (i.e $A - bb^t$ is indeed in Sym_d). Observe now that the expression of $\bar{\phi}(X)$ is polynomial in the entries of the input $X = \begin{pmatrix} A & b \\ b^t & 1 \end{pmatrix}$, and it follows that the partial derivatives of $\bar{\phi}^{-1}$ will necessarily be C^∞, therefore, $\bar{\phi}^{-1}$ is a smooth extension of ϕ^{-1}, showing that ϕ^{-1} is smooth as well.
In conclusion, we showed that the map $\phi : \mathbb{R}^d \times \mathcal{P}_d \to \mathcal{P}_{br,d+1}$ is indeed a diffeomorphism.

9. To show that the map ϕ is an isometry from $(\mathbb{R}^d \times \mathcal{P}_d, \langle \cdot, \cdot \rangle)_{FR}$ to $(\mathcal{P}_{br,d+1}, \langle \cdot, \cdot \rangle)_{FR}$, since by question 8 we already showed that ϕ is a diffeomorphism, we just need to show that $\forall (\Theta, \dot{\Theta}_1), (\Theta, \dot{\Theta}_2) \in \mathbb{R}^d \times \mathcal{P}_d \times \mathbb{R}^d \times Sym_d$ we have $\langle \dot{\Theta}_1, \dot{\Theta}_2 \rangle_\Theta = \langle D\phi(\Theta)[\dot{\Theta}_1], D\phi(\Theta)[\dot{\Theta}_2] \rangle_{\phi(\Theta)}$. Note that we actually decompose $\Theta = (\mu, \Sigma)$ and $\dot{\Theta}_i = (\dot{\mu}_i, \dot{\Sigma}_i), i = 1, 2$, where $\mu \in \mathbb{R}^d$, $\Sigma \in \mathcal{P}_d$, $\dot{\mu}_i \in \mathbb{R}^d$, $\dot{\Sigma}_i \in Sym_d$ for $i = 1, 2$.

We start by doing the following computation for $i = 1, 2$:

$$\begin{aligned}
D\phi(\Theta)[\dot{\Theta}_i] &= \lim_{t\to 0} \frac{\phi(\mu + t\dot{\mu}_i, \Sigma + t\dot{\Sigma}_i) - \phi(\mu, \Sigma)}{t} \\
&= \lim_{t\to 0} \frac{1}{t} \begin{pmatrix} \Sigma + t\dot{\Sigma}_i + (\mu + t\dot{\mu}_i)(\mu + t\dot{\mu}_i)^t & \mu + t\dot{\mu}_i \\ \mu + t\dot{\mu}_i^t & 1 \end{pmatrix} - \frac{1}{t} \begin{pmatrix} \Sigma + \mu\mu^t & \mu \\ \mu^t & 1 \end{pmatrix} \\
&= \lim_{t\to 0} \frac{1}{t} \begin{pmatrix} \Sigma + t\dot{\Sigma}_i + (\mu + t\dot{\mu}_i)(\mu + t\dot{\mu}_i)^t - \Sigma - \mu\mu^t & \mu + t\dot{\mu}_i - \mu \\ \mu + t\dot{\mu}_i^t - \mu^t & 1 - 1 \end{pmatrix} \\
&= \lim_{t\to 0} \frac{1}{t} \begin{pmatrix} t\dot{\Sigma}_i + t\dot{\mu}_i\mu^t + t\mu\dot{\mu}_i^t + t^2\dot{\mu}_i\dot{\mu}_i^t & t\dot{\mu}_i \\ t\dot{\mu}_i^t & 0 \end{pmatrix} \\
&= \begin{pmatrix} \dot{\Sigma}_i + \dot{\mu}_i\mu^t + \mu\dot{\mu}_i^t & \dot{\mu}_i \\ \dot{\mu}_i^t & 0 \end{pmatrix}
\end{aligned}$$

Now we can compute :

$$\langle D\phi(\Theta)[\dot{\Theta}_1], D\phi(\Theta)[\dot{\Theta}_2]\rangle_{\phi(\Theta)} =$$
$$\mathrm{Tr}\Big(\begin{pmatrix} \Sigma + \mu\mu^t & \mu \\ \mu^t & 1 \end{pmatrix}^{-1} \begin{pmatrix} \dot{\Sigma}_1 + \dot{\mu}_1\mu^t + \mu\dot{\mu}_1^t & \dot{\mu}_1 \\ \dot{\mu}_1^t & 0 \end{pmatrix} \begin{pmatrix} \Sigma + \mu\mu^t & \mu \\ \mu^t & 1 \end{pmatrix}^{-1} \begin{pmatrix} \dot{\Sigma}_2 + \dot{\mu}_2\mu^t + \mu\dot{\mu}_2^t & \dot{\mu}_2 \\ \dot{\mu}_2^t & 0 \end{pmatrix} \Big)$$

We start by giving a nice expression for $\begin{pmatrix} \Sigma + \mu\mu^t & \mu \\ \mu^t & 1 \end{pmatrix}^{-1}$ which follows from expressions of inverse block matrices :

$$\begin{aligned}
\begin{pmatrix} \Sigma + \mu\mu^t & \mu \\ \mu^t & 1 \end{pmatrix}^{-1} &= \begin{pmatrix} (\Sigma + \mu\mu^t - \mu\mu^t)^{-1} & -(\Sigma + \mu\mu^t - \mu\mu^t)^{-1}\mu \\ -\mu^t(\Sigma + \mu\mu^t - \mu\mu^t)^{-1} & 1 + \mu^t\Sigma^{-1}\mu \end{pmatrix} \\
&= \begin{pmatrix} \Sigma^{-1} & -\Sigma^{-1}\mu \\ -\mu^t\Sigma^{-1} & 1 + \mu^t\Sigma^{-1}\mu \end{pmatrix}
\end{aligned}$$

We now compute the following expression for $i = 1, 2$:

$$\begin{aligned}
\begin{pmatrix} \Sigma + \mu\mu^t & \mu \\ \mu^t & 1 \end{pmatrix}^{-1} \begin{pmatrix} \dot{\Sigma}_i + \dot{\mu}_i\mu^t + \mu\dot{\mu}_i^t & \dot{\mu}_i \\ \dot{\mu}_i^t & 0 \end{pmatrix} &= \begin{pmatrix} \Sigma^{-1} & -\Sigma^{-1}\mu \\ -\mu^t\Sigma^{-1} & 1 + \mu^t\Sigma^{-1}\mu \end{pmatrix} \begin{pmatrix} \dot{\Sigma}_i + \dot{\mu}_i\mu^t + \mu\dot{\mu}_i^t & \dot{\mu}_i \\ \dot{\mu}_i^t & 0 \end{pmatrix} \\
&= \begin{pmatrix} \Sigma^{-1}(\dot{\Sigma}_i + \dot{\mu}_i\mu^t + \mu\dot{\mu}_i^t) - \Sigma^{-1}\mu\dot{\mu}_i^t & \Sigma^{-1}\dot{\mu}_i \\ -\mu^t\Sigma^{-1}(\dot{\Sigma}_i + \dot{\mu}_i\mu^t + \mu\dot{\mu}_i^t) + (1 + \mu^t\Sigma^{-1}\mu)\dot{\mu}_i^t & -\mu^t\Sigma^{-1}\dot{\mu}_i \end{pmatrix} \\
&= \begin{pmatrix} \Sigma^{-1}(\dot{\Sigma}_i + \dot{\mu}_i\mu^t) & \Sigma^{-1}\dot{\mu}_i \\ -\mu^t\Sigma^{-1}(\dot{\Sigma}_i + \dot{\mu}_i\mu^t) + \dot{\mu}_i^t & -\mu^t\Sigma^{-1}\dot{\mu}_i \end{pmatrix}
\end{aligned}$$

Now we compute the product between these expressions for $i = 1, 2$ yielding :

$$\begin{pmatrix} \Sigma + \mu\mu^t & \mu \\ \mu^t & 1 \end{pmatrix}^{-1} \begin{pmatrix} \dot{\Sigma}_1 + \dot{\mu}_1\mu^t + \mu\dot{\mu}_1^t & \dot{\mu}_1 \\ \dot{\mu}_1^t & 0 \end{pmatrix} \begin{pmatrix} \Sigma + \mu\mu^t & \mu \\ \mu^t & 1 \end{pmatrix}^{-1} \begin{pmatrix} \dot{\Sigma}_2 + If we\dot{\mu}_2\mu^t + \mu\dot{\mu}_2^t & \dot{\mu}_2 \\ \dot{\mu}_2^t & 0 \end{pmatrix} =$$
$$\begin{pmatrix} A & * \\ * & b \end{pmatrix}$$

Where

$$\begin{aligned}
A &= \Sigma^{-1}(\dot{\Sigma}_1 + \dot{\mu}_1\mu^t)\Sigma^{-1}(\dot{\Sigma}_2 + \dot{\mu}_2\mu^t) + (\Sigma^{-1}\dot{\mu}_1)(-\mu^t\Sigma^{-1}(\dot{\Sigma}_2 + \dot{\mu}_2\mu^t) + \dot{\mu}_2^t) \\
b &= (-\mu^t\Sigma^{-1}(\dot{\Sigma}_1 + \dot{\mu}_1\mu^t) + \dot{\mu}_1)\Sigma^{-1}\dot{\mu}_2 + \mu^t\Sigma^{-1}\dot{\mu}_1\mu^t\Sigma^{-1}\dot{\mu}_2
\end{aligned}$$

And the entries denoted by $*$ were not computed since they will not be taken in account when computing the trace of this matrix, as we will only add the trace of the two diagonal blocks. We have after simplification :

$$\begin{pmatrix} A & * \\ * & b \end{pmatrix} = \begin{pmatrix} \Sigma^{-1}\dot{\Sigma}_1(\Sigma^{-1}\dot{\Sigma}_2 + \Sigma^{-1}\dot{\mu}_2\mu^t) + \Sigma^{-1}\dot{\mu}_1\dot{\mu}_2^t & * \\ * & -\mu^t\Sigma^{-1}\dot{\Sigma}_1\Sigma^{-1}\dot{\mu}_2 + \dot{\mu}_1^t\Sigma^{-1}\dot{\mu}_2 \end{pmatrix}$$

And we can finally compute :

$$\begin{aligned}
\langle D\phi(\Theta)[\dot{\Theta}_1], D\phi(\Theta)[\dot{\Theta}_2]\rangle_{\phi(\Theta)} &= Tr(A) + Tr(b) \\
&= Tr(\Sigma^{-1}\dot{\Sigma}_1(\Sigma^{-1}\dot{\Sigma}_2 + \Sigma^{-1}\dot{\mu}_2\mu^t) + \Sigma^{-1}\dot{\mu}_1\dot{\mu}_2^t) + Tr(-\mu^t\Sigma^{-1}\dot{\Sigma}_1\Sigma^{-1}\dot{\mu}_2 + \dot{\mu}_1^t\Sigma^{-1}\dot{\mu}_2) \\
&= Tr(\Sigma^{-1}\dot{\Sigma}_1\Sigma^{-1}\dot{\Sigma}_2) + Tr(\Sigma^{-1}\dot{\Sigma}_1\Sigma^{-1}\dot{\mu}_2\mu^t) + Tr(\Sigma^{-1}\dot{\mu}_1\dot{\mu}_2^t) - Tr(\mu^t\Sigma^{-1}\dot{\Sigma}_1\Sigma^{-1}\dot{\mu}_2) \\
&+Tr(\dot{\mu}_1^t\Sigma^{-1}\dot{\mu}_2) \\
&= Tr(\Sigma^{-1}\dot{\Sigma}_1\Sigma^{-1}\dot{\Sigma}_2) + Tr(\Sigma^{-1}\dot{\Sigma}_1\Sigma^{-1}\dot{\mu}_2\mu^t) + Tr(\dot{\mu}_2^t\Sigma^{-1}\dot{\mu}_1) - Tr(\Sigma^{-1}\dot{\Sigma}_1\Sigma^{-1}\dot{\mu}_2\mu^t) \\
&+Tr(\dot{\mu}_1^t\Sigma^{-1}\dot{\mu}_2) \\
&= Tr(\Sigma^{-1}\dot{\Sigma}_1\Sigma^{-1}\dot{\Sigma}_2) + \dot{\mu}_2^t\Sigma^{-1}\dot{\mu}_1 + \dot{\mu}_1^t\Sigma^{-1}\dot{\mu}_2 \\
&= Tr(\Sigma^{-1}\dot{\Sigma}_1\Sigma^{-1}\dot{\Sigma}_2) + 2\dot{\mu}_1^t\Sigma^{-1}\dot{\mu}_2 \\
&= \langle \dot{\Theta}_1, \dot{\Theta}_2\rangle_\Theta
\end{aligned}$$

Showing that ϕ is indeed an isometry from $(\mathbb{R}^d \times \mathcal{P}_d, \langle\cdot,\cdot\rangle_{FR})$ to $(\mathcal{P}_{br,d+1}, \langle\cdot,\cdot\rangle_{FR})$.

1.2 Reparameterizing the problem (MLE1)

10. Consider $q_{d+1}(X; y) = \sqrt{2\pi}\exp(\frac{1}{2})p_{d+1}(0, X; y)$ where $y^t = [x^t, 1] \in \mathbb{R}^{d+1}$ and $X = \phi(\mu, \Sigma) = \begin{pmatrix} \Sigma + \mu\mu^t & \mu \\ \mu^t & 1 \end{pmatrix}$, we then have :

$$\begin{aligned}
q_{d+1}(X; y) &= \sqrt{2\pi}\exp(\frac{1}{2})\det(X)^{-\frac{1}{2}}(2\pi)^{-\frac{d+1}{2}}\exp(-\frac{1}{2}y^tX^{-1}y) \\
&= (2\pi)^{-\frac{d}{2}}\exp(\frac{1}{2})\det(\Sigma + \mu\mu^t - \mu\mu^t)^{-\frac{1}{2}}\exp(-\frac{1}{2}y^tX^{-1}y) \\
&= \det(\Sigma)^{-\frac{1}{2}}(2\pi)^{-\frac{d}{2}}\exp(\frac{1}{2}(1 - y^tX^{-1}y))
\end{aligned}$$

Using expressions for the determinant of block matrices. We now do the following computation :

$$\begin{aligned}
1 - y^tX^{-1}y &= 1 - [x^t, 1]\begin{pmatrix} \Sigma + \mu\mu^t & \mu \\ \mu^t & 1 \end{pmatrix}^{-1}\begin{pmatrix} x \\ 1 \end{pmatrix} \\
&= 1 - [x^t, 1]\begin{pmatrix} \Sigma^{-1} & -\Sigma^{-1}\mu \\ -\mu^t\Sigma^{-1} & 1 + \mu^t\Sigma^{-1}\mu \end{pmatrix}\begin{pmatrix} x \\ 1 \end{pmatrix} \\
&= 1 - [x^t, 1]\begin{pmatrix} \Sigma^{-1}x - \Sigma^{-1}\mu \\ -\mu^t\Sigma^{-1}x + 1 + \mu^t\Sigma^{-1}\mu \end{pmatrix} \\
&= 1 - x^t\Sigma^{-1}x + x^t\Sigma^{-1}\mu + \mu^t\Sigma^{-1}x - 1 - \mu^t\Sigma^{-1}\mu \\
&= -(x - \mu)^t\Sigma^{-1}(x - \mu)
\end{aligned}$$

as wanted, and therefore we get :

$$\begin{aligned}
q_{d+1}(X; y) &= \det(\Sigma)^{-\frac{1}{2}}(2\pi)^{-\frac{d}{2}}\exp(-\frac{1}{2}(x - \mu)^t\Sigma^{-1}(x - \mu)) \\
&= p_d(\mu, \Sigma; x)
\end{aligned}$$

11. In the case $k = 1$, we only have on weight, which has to be equal to 1. Therefore, our problem can be seen as minimizing $l(X) = -\sum_{i=1}^{n} \log(q(X; y_i))$ over $\mathcal{P}_{br,d+1}$. We transform our cost function to have :

$$\begin{aligned} l(X) &= -\sum_{i=1}^{n} \log(\sqrt{2\pi}\exp(\frac{1}{2}) \det(X)^{-\frac{1}{2}} (2\pi)^{-\frac{d+1}{2}} \exp(-\frac{1}{2} y_i^t X^{-1} y_i) \\ &= -\sum_{i=1}^{n} (\frac{1}{2} - \frac{d}{2}\log(2\pi) - \frac{1}{2}\log(\det(X)) - \frac{1}{2} y_i^t X^{-1} y_i) \\ &= -\frac{n}{2} + \frac{nd}{2}\log(2\pi) + \frac{n}{2}\log(\det(X)) + \sum_{i=1}^{n} \frac{1}{2} y_i^t X^{-1} y_i \end{aligned}$$

and here we clearly see that our optimization problem is equivalent to the following problem:

$$\min_{X \in \mathcal{P}_{br,d+1}} n\log(\det(X)) + \sum_{i=1}^{n} y_i^t X^{-1} y_i$$

From now on, since $\mathcal{P}_{br,d+1}$ is a subset of $\mathcal{P}_{d+1}$, all the arguments that we used in the question 1 can be applied here again, that is, we can again look at the expression when we replace X by A^{-1}, observe the strict convexity of the function, differentiate and equalize to zero, just as it was done in question 1, and get this time a unique critical point $\hat{X} = \frac{1}{n}\sum_{i=1}^{n} y_i y_i^t$. We indeed have $\hat{X} \in \mathcal{P}_{br,d+1}$ as it is clearly symmetric positive definite and additionally :

$$\begin{aligned} [\hat{X}]_{d+1,d+1} &= [\frac{1}{n}\sum_{i=1}^{n} y_i y_i^t]_{d+1,d+1} \\ &= \frac{1}{n}\sum_{i=1}^{n} [y_i y_i^t]_{d+1,d+1} \\ &= 1 \end{aligned}$$

since the d-th entry of all the y_i is equal to 1 by definition, therefore $[y_i y_i^t]_{d+1,d+1} = 1$. Moreover, this unique critical point has to be a global minimum. This observation follows from the fact that the problem (MLE2) has been obtained from the problem (MLE1) by reparameterizing the latter using a diffeomorphism, and in the case $k = 1$ we see that if we apply the diffeomorphism ϕ from question 8 to the global minimum obtained in question 1, we get our critical point for this question.

1.3 Tools for optimization on $\mathcal{P}_{d+1}$ and $\mathcal{P}_{br,d+1}$ with FR metric

12. To show that R is well defined, we need to show that $\forall (X, V) \in T\mathcal{P}_{br,d+1}$, we have that $R_X(V)$ exists and is in $\mathcal{P}_{br,d+1}$. So, consider $(X, V) \in T\mathcal{P}_{br,d+1}$ and observe first that since $\mathcal{P}_{br,d+1}$ is an embedded submanifold, and in particular a subset, of $\mathcal{P}_{d+1}$ (it was shown in question 7), it follows that $(X, V) \in T\mathcal{P}_{d+1}$, and therefore, it makes sense to compute $\tilde{R}_X(V)$. We now make the observation that for any $A \in \mathcal{P}_{d+1}$, the entries on the main diagonal of A cannot be non-positive. Indeed, assume on the contrary that the i-th entry of the diagonal of A is non-positive, then if we consider the canonical vector $e_i \in \mathbb{R}^{d+1}$ which has only zero entries except a 1 at the i-th entry, it follows that $e_i^t A e_i = [A]_{i,i} \leq 0$, which contradicts the fact that A is positive-definite. Therefore, $\frac{1}{[\tilde{R}_X(V)]_{d+1,d+1}} \in \mathbb{R}$ is always well defined since $[\tilde{R}_X(V)]_{d+1,d+1}$ can never be non-positive. Now, we see that since $\tilde{R}_X(V) \in \mathcal{P}_{d+1}$, and $[\tilde{R}_X(V)]_{d+1,d+1}$ can never be non-positive, it follows that $R_X(V) = \frac{\tilde{R}_X(V)}{[\tilde{R}_X(V)]_{d+1,d+1}}$ is a matrix in $\mathcal{P}_{d+1}$,

and moreover, $[R_X(V)]_{d+1,d+1} = \frac{[\tilde{R}_X(V)]_{d+1,d+1}}{[\tilde{R}_X(V)]_{d+1,d+1}} = 1$, and therefore, $R_X(V) \in \mathcal{P}_{br,d+1}$ as wanted.
We now argue that R is a smooth map. We see that since $\tilde{R}$ is a retraction, it is a smooth map, and also, we already argued in question 7 that the map $A \mapsto [A]_{d+1,d+1}$, where $A \in \mathcal{P}_{d+1}$ is a smooth map (actually, we showed that the map $A \mapsto [A]_{d+1,d+1} - 1$ is smooth, but it immediately follows that the map $A \mapsto [A]_{d+1,d+1}$ is smooth too). Therefore, the map $(X, V) \mapsto \frac{1}{[\tilde{R}_X(V)]_{d+1,d+1}}$ is smooth as it is the composition of the smooth maps $\tilde{R}$, $A \mapsto [A]_{d+1,d+1}$ and $x \mapsto \frac{1}{x}$ (we already know from analysis courses that the map $g : \mathbb{R}\backslash\{0\} \to \mathbb{R}\backslash\{0\}$ defined by $g(x) = \frac{1}{x}$ is C^∞ on $\mathbb{R}\backslash\{0\}$). It follows that R is a smooth map as a product of smooth maps (it is the product rule given in the lecture notes from week 3).
We are now showing the last required property for R to be a retraction on $\mathcal{P}_{br,d+1}$. To do so, for each $(X, V) \in T\mathcal{P}_{br,d+1}$, we consider the curve $c(t) = R_X(tV) = \frac{\tilde{R}_X(tV)}{[\tilde{R}_X(tV)]_{d+1,d+1}}$. Observe that since $\tilde{R}$ is a retraction, we can define the following curve $\tilde{c}(t) = \tilde{R}_X(tV)$ which necessarily has to satisfy $\tilde{c}(0) = X$ and $\tilde{c}'(0) = V$, and it follows that $c(t) = \frac{\tilde{c}(t)}{[\tilde{c}(t)]_{d+1,d+1}}$.
We immediately see that :

$$\begin{aligned} c(0) &= \frac{X}{[X]_{d+1,d+1}} \\ &= X \end{aligned}$$

since $X \in \mathcal{P}_{br,d+1}$ and therefore $[X]_{d+1,d+1} = 1$ by definition.
To show that $c'(0) = V$, we first define $f : \mathbb{R} \to \mathbb{R}$ by $f(t) = \frac{1}{[\tilde{c}(t)]_{d+1,d+1}}$ so that $c(t) = \tilde{c}(t)f(t)$, and then, by the product rule, we have:

$$\begin{aligned} c'(0) &= \tilde{c}'(0)f(0) + \tilde{c}(0)f'(0) \\ &= V\frac{1}{[X]_{d+1,d+1}} + Xf'(0) \\ &= V + Xf'(0) \end{aligned}$$

We now compute $f'(0)$. To do so, consider the inverse map $g : \mathbb{R}\backslash\{0\} \to \mathbb{R}\backslash\{0\}$ defined by $g(x) = \frac{1}{x}$, and the map $F : \mathcal{P}_{d+1} \to \mathbb{R}\backslash\{0\}$ defined by $F(A) = [A]_{d+1,d+1}$, and observe that $f = g \circ F \circ \tilde{c}$. Therefore, using the chain rule we have :

$$\begin{aligned} f'(0) &= (g \circ (F \circ \tilde{c}))'(0) \\ &= g'(F \circ \tilde{c}(0))(F \circ \tilde{c})'(0) \\ &= -\frac{1}{F \circ \tilde{c}(0)}(F \circ \tilde{c})'(0) \\ &= -\frac{1}{F(X)}(F \circ \tilde{c})'(0) \\ &= -\frac{1}{[X]_{d+1,d+1}}DF(X)[V] \\ &= -DF(X)[V] \\ &= -[V]_{d+1,d+1} \\ &= 0 \end{aligned}$$

Where we used the facts that $[X]_{d+1,d+1} = 1$ since $X \in \mathcal{P}_{br,d+1}$, and $[V]_{d+1,d+1} = 0$, since $V \in T_X\mathcal{P}_{br,d+1}$ (recall that the tangent spaces of $\mathcal{P}_{br,d+1}$ were defined in question 7 to be the set of all

symmetric $(d+1)\times(d+1)$ matrices that have their bottom-right entry to be equal to zero), and $(F\circ\tilde{c})'(0) = DF(X)[V]$ was given in the lecture notes of week 3. Therefore, $c'(0) = V$ as wanted. In conclusion, we showed that R is well defined and is a retraction for $\mathcal{P}_{br,d+1}$.

13. Consider the map $\tilde{R} : T\mathcal{P}_{d+1} \to \mathcal{P}_{d+1}$ defined by $\tilde{R}(X,V) = \tilde{R}_X(V) = X + V + \frac{1}{2}VX^{-1}V$. We start by showing that this map is well defined, that is, $\forall (X,V) \in \mathcal{P}_{d+1} \times Sym_{d+1} = T\mathcal{P}_{d+1}$, we have that $\tilde{R}_X(V) \in \mathcal{P}_{d+1}$. We start by showing that $\tilde{R}_X(V)$ is symmetric :

$$\begin{aligned}\tilde{R}_X(V)^t &= (X + V + \frac{1}{2}VX^{-1}V)^t \\ &= X^t + V^t + \frac{1}{2}V^tX^{-t}V^t \\ &= X + V + \frac{1}{2}VX^{-1}V \\ &= \tilde{R}_X(V)\end{aligned}$$

We now show that $\tilde{R}_X(V)$ is positive-definite. To do so, since $X \in \mathcal{P}_{p+1}$, we have (recall question 6) that $X^{-1} \in \mathcal{P}_{d+1}$ too, and therefore its square root $X^{-\frac{1}{2}}$ is well defined and is symmetric positive-definite too. Therefore we see that :

$$\begin{aligned}X^{-\frac{1}{2}}\tilde{R}_X(V)X^{-\frac{1}{2}} &= X^{-\frac{1}{2}}(X + V + \frac{1}{2}VX^{-1}V)X^{-\frac{1}{2}} \\ &= I + X^{-\frac{1}{2}}VX^{-\frac{1}{2}} + \frac{1}{2}X^{-\frac{1}{2}}VX^{-1}VX^{-\frac{1}{2}} \\ &= I + A + \frac{1}{2}A^2\end{aligned}$$

where we let $A = X^{-\frac{1}{2}}VX^{-\frac{1}{2}}$. Since A is clearly symmetric, we can diagonalize it as follows : $A = UDU^t$, where U is an orthogonal matrix, and D is diagonal. Therefore :

$$\begin{aligned}I + A + \frac{1}{2}A^2 &= UU^t + UDU^t + \frac{1}{2}UDU^tUDU^t \\ &= U(I + D + \frac{1}{2}D^2)U^t\end{aligned}$$

where $I + D + \frac{1}{2}D^2$ is a diagonal matrix. Observe that the map $x \mapsto 1 + x + \frac{1}{2}x^2$ (from $\mathbb{R}$ to $\mathbb{R}$) is strictly positive $\forall x \in \mathbb{R}$, and therefore all the entries of $I + D + \frac{1}{2}D^2$ are strictly positive, but then, since these entries are actually the eigenvalues of $X^{-\frac{1}{2}}\tilde{R}_X(V)X^{-\frac{1}{2}}$ (follows from the spectral theorem and the fact that this matrix is symmetric), we have that $X^{-\frac{1}{2}}\tilde{R}_X(V)X^{-\frac{1}{2}}$ is positive definite. Now since $X^{-\frac{1}{2}}$ has also to be non-singular, we have for any non zero $y \in \mathbb{R}^{d+1}$:

$$y^tX^{-\frac{1}{2}}\tilde{R}_X(V)X^{-\frac{1}{2}}y = (X^{-\frac{1}{2}}y)^t\tilde{R}_X(V)(X^{-\frac{1}{2}}y) > 0$$

and it follows that $\tilde{R}_X(V)$ must positive-definite, and therefore the map $\tilde{R}$ is well defined.
Now, we argue that $\tilde{R}$ is smooth. Recall that we already showed in question 6 that the inverse map $X \mapsto X^{-1}$ is a smooth map (in particular, X^{-1} is polynomial in the entries of X). Now if we look at the expression $\tilde{R}_X(V) = X + V + \frac{1}{2}VX^{-1}V$, we see that it is only composed of matrix multiplications and additions, and it contains the inverse of X, so we see that $\tilde{R}_X(V)$ is polynomial in the entries of (X,V), therefore, if we consider an extension of R from $Sym_{d+1} \times Sym_{d+1}$ to Sym_{d+1}, it will be a smooth extension (since all its partial derivatives will necessarily be C^∞) and therefore, R is a smooth

map.
Finally, for each $(X, V) \in T\mathcal{P}_{d+1}$, we consider the following curve : $c(t) = \tilde{R}_X(tV) = X + tV + \frac{t^2}{2}VX^{-1}V$. We clearly have $c(0) = X$, and then, $c'(t) = V + tVX^{-1}V$, which implies that $c'(0) = V$ as required.
In conclusion, we showed that $\tilde{R}$ is a retraction for $\mathcal{P}_{d+1}$ defined on all $T\mathcal{P}_{d+1}$.

14. Consider the map $\tilde{R} : T\mathcal{P}_{d+1} \to \mathcal{P}_{d+1}$ defined by $\tilde{R}(X, V) = \tilde{R}_X(V) = X + V + \frac{1}{2}VX^{-1}V$. By the question 13, this is a retraction for $\mathcal{P}_{d+1}$ defined on all $T\mathcal{P}_{d+1}$. Now define $R : T\mathcal{P}_{br,d+1} \to \mathcal{P}_{br,d+1}$ by $R_X(V) = \frac{\tilde{R}_X(V)}{[\tilde{R}_X(V)]_{d+1,d+1}}$. By question 12, we have that this map is a retraction for $\mathcal{P}_{br,d+1}$.

15. The euclidean gradient of $f : \mathcal{P}_{d+1} \to \mathbb{R}$ is the unique vector in Sym_{d+1} such that $\langle \nabla f(X), V\rangle_{\mathcal{E}} = Df(X)[V], \forall X \in \mathcal{P}_{d+1}, V \in Sym_{d+1}$ and the Riemannian gradient is the unique vector in Sym_{d+1} such that $\langle gradf(X), V\rangle_{FR} = Df(X)[V], \forall X \in \mathcal{P}_{d+1}, V \in Sym_{d+1}$. Therefore, we have :

$$\begin{aligned}
\langle \nabla f(X), V\rangle_{\mathcal{E}} &= \langle gradf(X), V\rangle_{FR}, \forall V \in Sym_{d+1} \\
\implies Tr(\nabla f(X)V) &= Tr(X^{-1}gradf(X)X^{-1}V), \forall V \in Sym_{d+1} \\
\implies Tr(\nabla f(X)V - X^{-1}gradf(X)X^{-1}V) &= 0, \forall V \in Sym_{d+1} \\
\implies Tr((\nabla f(X) - X^{-1}gradf(X)X^{-1})V) &= 0, \forall V \in Sym_{d+1} \\
\implies \langle \nabla f(X) - X^{-1}gradf(X)X^{-1}, V\rangle_{\mathcal{E}} &= 0, \forall V \in Sym_{d+1} \\
\implies \nabla f(X) - X^{-1}gradf(X)X^{-1} &= 0_{d+1\times d+1}, \\
\implies gradf(X) &= X\nabla f(X)X
\end{aligned}$$

where we used the positive-deifniteness of the inner product (observe that here we denoted for simplification $\mathcal{E} = \mathrm{Sym}_{d+1}$, which is different than the $\mathcal{E}$ given at the beginning of the project, and used the usual inner product on a matrix linear space).

16. Since $\mathcal{P}_{d+1}$ is a submanifold of the linear space Sym_{d+1}, together with the Riemannian metric $\langle\cdot,\cdot\rangle_{FR}$, and $\mathcal{P}_{br,d+1}$ is viewed as a Riemannian submanifold of $\mathcal{P}_{d+1}$, if we let $f : \mathcal{P}_{br,d+1} \to \mathbb{R}$ be a smooth function on $\mathcal{P}_{br,d+1}$ with a smooth extension $\bar{f}$ defined on an open subset of $\mathcal{P}_{d+1}$ which contains $\mathcal{P}_{br,d+1}$, then the Riemannian gradient of f in $(\mathcal{P}_{br,d+1}, \langle\cdot,\cdot\rangle_{FR})$, denoted $gradf$, is the orthogonal projection of the Riemannian gradient of $\bar{f}$ on $\mathcal{P}_{d+1}$, denoted $grad\bar{f}$ (recall from the previous question that $grad\bar{f}(X) = X\nabla\bar{f}(X)X, \forall X \in \mathcal{P}_{br,d+1}$ where $\nabla\bar{f}$ is the euclidean gradient of $\bar{f}$), that is, $gradf(X) = Proj_X(grad\bar{f}(X)), \forall X \in \mathcal{P}_{br,d+1}$. Here, $Proj_X : T_X\mathcal{P}_{d+1} \to T_X\mathcal{P}_{br,d+1}$ denote the orthogonal projection onto $T_X\mathcal{P}_{br,d+1}$, that is, $Proj_X$ is the unique linear map such that $Proj_X(V) \in T_X\mathcal{P}_{br,d+1}$ and $\langle Proj_X(V) - V, U\rangle_{FR} = 0$, $\forall V \in T_X\mathcal{P}_{d+1} = Sym_{d+1}$ and $\forall U \in T_X\mathcal{P}_{br,d+1}$. We claim that $Proj_X$ is defined as follows : $Proj_X(V) = V - \frac{[V]_{d+1,d+1}}{[X]^2_{d+1,d+1}}XEX$ where E is as before, the matrix with only zero entries, except a 1 at the bottom-right entry. We now show that this map has the properties required. We start by showing the linearity, so let $t \in \mathbb{R}, V, U \in Sym_{d+1}$, we have :

$$\begin{aligned}
Proj_X(tU + V) &= tU + V - \frac{[tU + V]_{d+1,d+1}}{[X]^2_{d+1,d+1}}XEX \\
&= tU + V - \frac{t[U]_{d+1,d+1} + [V]_{d+1,d+1}}{[X]^2_{d+1,d+1}}XEX \\
&= tU - t\frac{[U]_{d+1,d+1}}{[X]^2_{d+1,d+1}}XEX + V - \frac{[V]_{d+1,d+1}}{[X]^2_{d+1,d+1}}XEX \\
&= tProj_X(U) + Proj_X(V)
\end{aligned}$$

We now show that $\forall V \in Sym_{d+1}$ we have $Proj_X(V) \in T_X\mathcal{P}_{br,d+1}$, that is, $[Proj_X(V)]_{d+1,d+1} = 0$. Indeed, we have :

$$\begin{aligned}[Proj_X(V)]_{d+1,d+1} &= [V - \frac{[V]_{d+1,d+1}}{[X]^2_{d+1,d+1}} XEX]_{d+1,d+1} \\ &= [V]_{d+1,d+1} - \frac{[V]_{d+1,d+1}}{[X]^2_{d+1,d+1}} [XEX]_{d+1,d+1} \\ &= [V]_{d+1,d+1} - \frac{[V]_{d+1,d+1}}{[X]^2_{d+1,d+1}} [X]^2_{d+1,d+1} = 0\end{aligned}$$

where the fact that $[XEX]_{d+1,d+1} = [X]^2_{d+1,d+1}$ comes from the observation that EX is a matrix having its first d rows completely filled with zeros, and the last row is just the last row of X, so $[XEX]_{d+1,d+1}$ is equal to the inner product (the usual inner product on $\mathbb{R}^{d+1}$) between the $d+1$-th row of X and the $d+1$-th column of EX, the latter having only zero entries except $[X]_{d+1,d+1}$ at the last one, yielding $[XEX]_{d+1,d+1} = [X]^2_{d+1,d+1}$.
Finally we show that $\forall V \in \mathrm{Sym}_{d+1}$ and $\forall U \in T_X\mathcal{P}_{br,d+1}$ we have that $\langle Proj_X(V) - V, U\rangle_{FR} = 0$. Note that since $U \in T_X\mathcal{P}_{br,d+1}$, $[U]_{d+1,d+1} = 0$. Now we compute :

$$\begin{aligned}\langle Proj_X(V) - V, U\rangle_{FR} &= \langle -\frac{[V]_{d+1,d+1}}{[X]^2_{d+1,d+1}} XEX, U\rangle_X \\ &= -\frac{[V]_{d+1,d+1}}{[X]^2_{d+1,d+1}} \langle XEX, U\rangle_X \\ &= -\frac{[V]_{d+1,d+1}}{[X]^2_{d+1,d+1}} Tr(X^{-1}XEXX^{-1}U) \\ &= -\frac{[V]_{d+1,d+1}}{[X]^2_{d+1,d+1}} Tr(EU)\end{aligned}$$

But EU is a matrix whose first d rows are completely filled with zeros, and the last row is just the $d+1$-th row of U, and since $[U]_{d+1,d+1} = 0$, it follows that $Tr(EU) = 0$, and the conclusion follows.
Observe that whenever $X \in \mathcal{P}_{br,d+1}, [X]_{d+1,d+1} = 1$ and we can further simplify the expression of the projection and have $Proj_X(V) = V - [V]_{d+1,d+1}XEX$.
We can now compute the Riemannian gradient of f :

$$\begin{aligned}gradf(X) &= Proj_X(grad\bar{f}(X)) \\ &= Proj_X(X\nabla\bar{f}(X)X) \\ &= X\nabla\bar{f}(X)X - [X\nabla\bar{f}(X)X]_{d+1,d+1}XEX\end{aligned}$$

17. We implement all our functions and programs to be compatible with Manopt [1], and we write a manifold factory for $\mathcal{P}_{br,d+1}$ based on $\mathcal{P}_{d+1}$ wich correspond to `sympositivedefinitefactory.m` in Manopt. For the details of the code, it can be read in the Github repository Benoit-Muller/opti-manifolds-GMM [2], the manifold factory correspond to the file `spd_br_factory.m`. We make widely use of all the results prooved before.

1.4 Optimizing over the weights

18. We first show that the map $\psi : \mathbb{S}^{k-1} \to \Delta^{k-1}$ is well-defined. Let $u = (u_1, \ldots, u_k) \in \mathbb{S}^{k-1}$, then, $v = \psi(u) = u \odot u = (u_1^2, \ldots, u_k^2)$. We now see that $\sum_{j=1}^k v_j = \sum_{j=1}^k u_j^2 = \|u\|_2^2 = 1$ since $u \in \mathbb{S}^{k-1}$.

Moreover, $u_j^2 \geq 0, \forall j = 1, \ldots, k$, so $\psi(u) \in \Delta^{k-1}$ and ψ is well defined.
We now show that ψ is surjective. Let $u = (u_1, \ldots, u_k) \in \Delta^{k-1}$, that is, $\sum_{j=1}^k u_j = 1$ and $u_j \geq 0, \forall j = 1, \ldots, k$. Since $u_j \geq 0, \forall j = 1, \ldots, k$ it makes sense to define $v_j = \sqrt{u_j}, \forall j = 1, \ldots, k$, so that we can consider the vector $v = (v_1, \ldots, v_k)$. We have that $\|v\|_2^2 = \sum_{j=1}^k v_j^2 = \sum_{j=1}^k \sqrt{u_j}^2 = \sum_{j=1}^k u_j = 1$; so $v \in \mathbb{S}^{k-1}$ and clearly, $\psi(v) = v \odot v = (v_1^2, \ldots, v_k^2) = (u_1, \ldots, u_k) = u$ as wanted.
Finally, ψ is not injective. As an example of why it is not, consider the two distinct vectors $u = (0, \ldots, 0, 1), v = (0, \ldots, 0, -1) \in \mathbb{S}^{k-1}$, then $\psi(u) = (0, \ldots, 0, 1) = \psi(v)$.

19. Consider the set $U := \{w \in \mathbb{R}^k : w_j > 0, \forall j = 1, \ldots, k\}$, which is clearly an open neighborhood of Δ_+^{k-1} in $\mathbb{R}^k$. Let $h : U \to \mathbb{R}$ defined by $h(w) = (\sum_{j=1}^k w_j) - 1$. Clearly, $U \cap \Delta_+^{k-1} = \{w \in U : h(w) = 0\} = h^{-1}(0)$ and h is clearly smooth as it is polynomial in the entries of w. We now compute $\forall v \in \mathbb{R}^k$ and $\forall w \in \Delta_+^{k-1}$:

$$\begin{aligned} Dh(w)[v] &= \lim_{t \to 0} \frac{h(w + tv) - h(w)}{t} \\ &= \sum_{j=1}^k v_j \\ &= (1, \ldots, 1)v \\ \implies Dh(w) &= (1, \ldots, 1)^t \end{aligned}$$

And therefore, clearly, $Dh(w)$ is of rank 1 $\forall w \in \Delta_+^{k-1}$, showing that h is a local defining function for Δ_+^{k-1}. We then have by a theorem in the lecture notes the following expression for the tangent spaces of Δ_+^{k-1} :

$$\begin{aligned} T_w \Delta_+^{k-1} &= \ker Dh(w), \forall w \in \Delta_+^{k-1} \\ &= \{v \in \mathbb{R}^k : Dh(w)[v] = 0\} \\ &= \{v \in \mathbb{R}^k : \sum_{j=1}^k v_j = 0\} \end{aligned}$$

20. We first show that it defines a metric on Δ_+^{k-1}, that is, $\forall w \in \Delta_+^{k-1}$, $\langle \cdot, \cdot \rangle_w$ is an inner product for $T_w \Delta_+^{k-1}$. Observe first that $\langle \cdot, \cdot \rangle_w$ is well defined, since $w_j > 0, \forall j = 1, \ldots, k$, so that it makes sense to divide by w_j.
We first show the symmetry. Let $\dot{w}, \dot{w}' \in T_w \Delta_+^{k-1}$, we have :

$$\begin{aligned} \langle \dot{w}, \dot{w}' \rangle_w &= \frac{1}{4} \sum_{j=1}^k \frac{\dot{w}_j \dot{w}'_j}{w_j} \\ &= \frac{1}{4} \sum_{j=1}^k \frac{\dot{w}'_j \dot{w}_j}{w_j} \\ &= \langle \dot{w}', \dot{w} \rangle_w \end{aligned}$$

We now show the bilinearity. Let $\dot{w}, \dot{w}', \dot{w}'' \in T_w\Delta_+^{k-1}$ and $a, b \in \mathbb{R}$, we have :

$$\begin{aligned}
\langle a\dot{w} + b\dot{w}', \dot{w}''\rangle_w &= \frac{1}{4}\sum_{j=1}^{k} \frac{(a\dot{w}_j + b\dot{w}'_j)\dot{w}''_j}{w_j} \\
&= \frac{1}{4}\sum_{j=1}^{k} \frac{a\dot{w}_j\dot{w}''_j + b\dot{w}'_j\dot{w}''_j}{w_j} \\
&= a\frac{1}{4}\sum_{j=1}^{k} \frac{\dot{w}_j\dot{w}''_j}{w_j} + b\frac{1}{4}\sum_{j=1}^{k} \frac{\dot{w}'_j\dot{w}''_j}{w_j} \\
&= a\langle \dot{w}, \dot{w}''\rangle_w + b\langle \dot{w}', \dot{w}''\rangle_w
\end{aligned}$$

and the linearity in the second argument follows by symmetry.
Finally, we show the positive-definiteness. Let $\dot{w} \in T_w\Delta_+^{k-1}$ be a non-zero vector, then:

$$\langle \dot{w}, \dot{w}\rangle_w = \frac{1}{4}\sum_{j=1}^{k} \frac{\dot{w}_j^2}{w_j}$$

But since, $\dot{w}_j^2 \geq 0, \forall j = 1, \ldots, k$, and $\dot{w}$ is a non-zero vector, then at least one of the $\dot{w}_j$ is non-zero, so that $\dot{w}_j^2 > 0$ and $\langle \dot{w}, \dot{w}\rangle_w > 0$ as wanted.
We now show that $\langle \cdot, \cdot \rangle_w$ on Δ_+^{k-1} is a Riemannian metric. To do so, let V, W be two smooth vector fields on Δ_+^{k-1}, and define the map $F : \Delta_+^{k-1} \to \mathbb{R}$ by $F(w) = \langle V(w), W(w)\rangle_w$. We will show that F is a smooth function. . Since both V and W are smooth vector fields on Δ_+^{k-1}, by proposition 3.45 in the textbook, there exists smooth vector fields $\bar{V}$ and $\bar{W}$ defined on an open neighborhood of Δ_+^{k-1}, say U_1 and U_2 respectively, such that $V = \bar{V}|_{\Delta_+^{k-1}}$ and $W = \bar{W}|_{\Delta_+^{k-1}}$. We can assume that we can take U_1 and U_2 small enough so that both are open subsets of the set $\{w \in \mathbb{R}^k : w_j > 0, \forall j = 1, \ldots, k\}$. Observe that $U = U_1 \cap U_2$ is an open neighborhood of Δ_+^{k-1} in $\mathbb{R}^k$. We now define the map $\bar{F} : U \to \mathbb{R}$ defined by $F(w) = \frac{1}{4}\sum_{j=1}^{k} \frac{\bar{V}_j(w)\bar{W}_j(w)}{w_j}$, where we decomposed $\bar{V}(w) = (\bar{V}_1(w), \ldots, \bar{V}_k(w))$ and similarly for $\bar{W}$, so that $F = \bar{F}|_{\Delta_+^{k-1}}$. Note that since we $U \subseteq \{w \in \mathbb{R}^k : w_j > 0, \forall j = 1, \ldots, k\}$, it makes sense to divide by w_j and therefore $\bar{F}$ is a well defined map. Observe that since $\bar{V}$ is a smooth extension of V, we can view each $\bar{V}_j$ as a smooth map from U to $\mathbb{R}$, and the same goes for $\bar{W}$. It follows that the map $w \mapsto \bar{V}_j(w)\bar{W}_j(w)$ is smooth $\forall j = 1, \ldots, k$ as a product of smooth maps, and since $w_j > 0, \forall j = 1, \ldots, k$, the map $w \mapsto \frac{1}{w_j}$ is smooth too $\forall j = 1, \ldots, k$. Therefore the map $w \mapsto \frac{\bar{V}_j(w)\bar{W}_j(w)}{w_j}$ is smooth $\forall j = 1, \ldots, k$, and so $\bar{F}$ is smooth as a linear combination of smooth maps. In conslusion, $\bar{F}$ is a smooth extension of F and it follows that F is smooth, so that our metric is indeed a Riemannian metric on Δ_+^{k-1}.

21. We start by showing that $\psi : \mathcal{S}_+^{k-1} \to \Delta_+^{k-1}$ is a bijection. We start by showing that it is surjective. Indeed, let $u = (u_1, \ldots, u_k) \in \Delta_+^{k-1}$, that is, $\sum_{j=1}^{k} u_j = 1$ and $u_j > 0, \forall j = 1, \ldots, k$. Since $u_j > 0, \forall j = 1, \ldots, k$ it makes sense to define $v_j = \sqrt{u_j}, \forall j = 1, \ldots, k$, so that we can consider the vector $v = (v_1, \ldots, v_k)$. We have that $\|v\|_2^2 = \sum_{j=1}^{k} v_j^2 = \sum_{j=1}^{k} \sqrt{u_j}^2 = \sum_{j=1}^{k} u_j = 1$; so $v \in \mathbb{S}^{k-1}$ and clearly, $v_j > 0, \forall j = 1, \ldots, k$, so that $v \in \mathcal{S}_+^{k-1}$. Moreover, $\psi(v) = v \odot v = (v_1^2, \ldots, v_k^2) = (u_1, \ldots, u_k) = u$ as wanted.

We now show that it is injective. Let $u = (u_1, \dots, u_k), v = (v_1, \dots, v_k) \in \mathcal{S}_+^{k-1}$ such that $\psi(u) = \psi(v)$, that is, $u_j^2 = v_j^2, \forall j = 1, \dots, k$. Now, since both u_j and v_j are strictly positive, the only solution for these equations is $u_j = v_j, \forall j = 1, \dots, k$, and therefore $u = v$. Therefore, since ψ is bijective, it necessarily admits an inverse map $\psi^{-1} : \Delta_+^{k-1} \to \mathcal{S}_+^{k-1}$, which is well defined, and clearly, it is defined by $\psi(u) = (\sqrt{u_1}, \dots, \sqrt{u_k}), \forall u \in \Delta_+^{k-1}$.
We now argue that both ψ and ψ^{-1} are smooth. For ψ, define the map $\bar{\psi} : \mathbb{R}^k \to \mathbb{R}^k$ by $\bar{\psi}(u) = (u_1^2, \dots, u_k^2)$ which is clearly a smooth extension of ψ since it is polynomial in the entries of $u \in \mathbb{R}^k$. Now, For ψ^{-1}, define the map $\bar{\psi}^{-1} : \mathbb{R}_{>0}^k \to \mathbb{R}_{>0}^k$, where $\mathbb{R}_{>0}^k = \{w \in \mathbb{R}^k : w_j > 0, \forall j = 1, \dots, k\}$ is an open neighborhood of Δ_+^{k-1}, by $\bar{\psi}^{-1}(u) = (\sqrt{u_1}, \dots, \sqrt{u_k})$, which is a smooth extension of ψ^{-1} as each of its coordinate function $u \mapsto \sqrt{u_j}$ is smooth $\forall j = 1, \dots, k$ on $\mathbb{R}_{>0}^k$ (as a composition of the smooth maps $u \mapsto u_j$ and $x \mapsto \sqrt{x}$ for $x \in \mathbb{R}$ and $x > 0$).
In conclusion, the entry-wise squaring map $\psi : \mathcal{S}_+^{k-1} \to \Delta_+^{k-1}$ is a diffeomorphism.

22. To show that $\psi : \mathcal{S}_+^{k-1} \to \Delta_+^{k-1}$ is an isometry from $\mathcal{S}_+^{k-1}$ with the usual Riemannian metric to Δ_+^{k-1} with the Riemannian metric previously defined, we need to show that $\langle u, v \rangle_{\mathcal{S}_+^{k-1}} = \langle D\psi(w)[u], D\psi(w)[v] \rangle_{\Delta_+^{k-1}}$, for all $(w, u), (w, v) \in T\mathcal{S}_+^{k-1}$. We first compute :

$$\begin{aligned} D\psi(w)[u] &= \lim_{t \to 0} \frac{\psi(w + tu) - \psi(w)}{t} \\ &= \lim_{t \to 0} \frac{1}{t}((w_1 + tu_1)^2 - w_1^2, \dots, (w_k + tu_k)^2 - w_k^2) \\ &= \lim_{t \to 0} \frac{1}{t}(2tw_1u_1 + t^2u_1^2, \dots, 2tw_ku_k + t^2u_k^2) \\ &= 2(w_1u_1, \dots, w_ku_k) \end{aligned}$$

and similarly, $D\psi(w)[v] = 2(w_1v_1, \dots, w_kv_k)$, and therefore :

$$\begin{aligned} \langle D\psi(w)[u], D\psi(w)[v] \rangle_{\Delta_+^{k-1}} &= \langle D\psi(w)[u], D\psi(w)[v] \rangle_{\psi(w)} \\ &= \frac{1}{4} \sum_{j=1}^{k} \frac{2w_ju_j 2w_jv_j}{w_j^2} \\ &= \frac{1}{4} \sum_{j=1}^{k} \frac{4w_j^2u_jv_j}{w_j^2} \\ &= \sum_{j=1}^{k} u_jv_j \\ &= \langle u, v \rangle_{\mathcal{S}_+^{k-1}} \end{aligned}$$

as wanted.

1.5 Computing Riemannian gradients of the negative log-likelihood

23. Let $\mathcal{M} = \mathbb{S}^{k-1} \times \prod_{j=1}^{k} \mathcal{P}_{br,d+1}$ and $l : \mathcal{M} \to \mathbb{R}$ the cost function defined in problem (MLE3). We start by recalling the Product Metric exercise in the exercise session 4 stating that we can turn the product of Riemannian manifolds into a Riemannian manifold by giving it the Riemannian product metric, (which, we briefly recall, consists of "adding" the metrics altogether), and for a smooth function f from this product Riemannian manifold to $\mathbb{R}$, its gradient can be seen as the vector whose

entries are the respective gradients of the functions obtained from f by fixing all the inputs in their respective manifold, except one. Our strategy to compute the Riemannian gradient of l will be to compute the gradient of its extension to an open set included in $\mathbb{R}^k \times \prod_{j=1}^k \mathcal{P}_{d+1}$, and use the usual projection on the tangent spaces of the sphere, and the projection defined in question 16 to project each component to, respectively, the tangent spaces of $\mathbb{S}^{k-1}$ and $\mathcal{P}_{br,d+1}$.
Let $(u, (X_j)_{j=1}^k) \in \mathcal{M}$, recall that the tangent space at $(u, (X_j)_{j=1}^k)$ is given by $T_u\mathbb{S}^{k-1} \times \prod_{j=1}^k T_{X_j}\mathcal{P}_{br,d+1}$. So, we let $(v, (V_j)_{j=1}^k) \in T_u\mathbb{S}^{k-1} \times \prod_{j=1}^k T_{X_j}\mathcal{P}_{br,d+1}$ be a tangent vector at $(u, (X_j)_{j=1}^k)$. We will compute first $Dl(u, (X_j)_{j=1}^k)[(v, (V_j)_{j=1}^k)]$ using the curve perspective. To do so, define the smooth curve $c(t) = (u, (X_j)_{j=1}^k) + t(v, (V_j)_{j=1}^k)$, we have that $Dl(u, (X_j)_{j=1}^k)[(v, (V_j)_{j=1}^k)] = \frac{d}{dt}(l \circ c)|_{t=0}$.
Now, to ease the notations, we define $u(t) = u + tv$, so that we can write $u_j(t) = u_j + tv_j, \forall j = 1, \dots, k$, and also $X_j(t) = X_j + tV_j, \forall j = 1, \dots, k$. We will also make an abuse of notation as we will use $\langle \cdot, \cdot \rangle$ to denote the usual inner product on the respective spaces (i.e, for $\mathbb{S}^{k-1}$ it is given by $\langle x, y \rangle = x^t y$, and for $\mathcal{P}_{br,d+1}$ by $\langle A, B \rangle = Tr(AB)$). We also denote by $\langle \cdot, \cdot \rangle_{FR}$ the FR-metric previously defined. We can now start the computations :

$$\begin{aligned}
\frac{d}{dt}(l \circ c)(t) &= -\frac{d}{dt}\sum_{i=1}^n \log(\sum_{j=1}^k u_j(t)^2 q(X_j(t); y_i)) \\
&= -\sum_{i=1}^n \frac{d}{dt}\log(\sum_{j=1}^k u_j(t)^2 q(X_j(t); y_i)) \\
&= -\sum_{i=1}^n \frac{\sum_{j=1}^k \frac{d}{dt}(u_j(t)^2 q(X_j(t); y_i))}{\sum_{l=1}^k u_l(t)^2 q(X_l(t); y_i)} \\
&= -\sum_{i=1}^n \frac{\sum_{j=1}^k 2u_j(t)v_j q(X_j(t); y_i) + u_j(t)^2 \frac{d}{dt}(q(X_j(t); y_i))}{\sum_{l=1}^k u_l(t)^2 q(X_l(t); y_i)}
\end{aligned}$$

Where :

$$\begin{aligned}
\frac{d}{dt}(q(X_j(t); y_i)) &= \sqrt{2\pi}\exp(\frac{1}{2})\frac{d}{dt}(p(0, X_j(t); y_i)) \\
&= \sqrt{2\pi}\exp(\frac{1}{2})(2\pi)^{-\frac{d+1}{2}}\frac{d}{dt}(\det(X_j(t))^{-\frac{1}{2}}\exp(-\frac{1}{2}y_i^t X_j(t)^{-1}y_i)) \\
&= \sqrt{2\pi}\exp(\frac{1}{2})(2\pi)^{-\frac{d+1}{2}}(\frac{d}{dt}(\det(X_j(t))^{-\frac{1}{2}})\exp(-\frac{1}{2}y_i^t X_j(t)^{-1}y_i) \\
&+ \det(X_j(t))^{-\frac{1}{2}}\frac{d}{dt}(\exp(-\frac{1}{2}y_i^t X_j(t)^{-1}y_i)))
\end{aligned}$$

We now compute :

$$\frac{d}{dt}(\det(X_j(t))^{-\frac{1}{2}}) = -\frac{1}{2}\det(X_j(t))^{-\frac{3}{2}}\frac{d}{dt}\det(X_j(t))$$

where (the result is taken from the wikipedia page of the determinant, but it follows from the multilinearity of the determinant) :

$$\frac{d}{dt}\det(X_j(t)) = \sum_{l=1}^{d+1}\det(X_{j,1}(t), \dots, X_{j,l-1}(t), \frac{d}{dt}X_{j,l}(t), X_{j,l+1}(t), \dots, X_{j,d+1}(t))$$

where $X_{j,l}(t) = X_{j,l} + tV_{j,l}$ and the additional index here is used to denote the l-th column of the corresponding matrix. It follows that :

$$\begin{aligned}
\frac{d}{dt}\det(X_j(t))|_{t=0} &= \sum_{l=1}^{d+1}\det(X_{j,1},\dots,X_{j,l-1},V_{j,l},X_{j,l+1},\dots,X_{j,d+1}) \\
\implies \frac{d}{dt}\det(X_j(t))^{-\frac{1}{2}}|_{t=0} &= -\frac{1}{2}\det(X_j)^{-\frac{1}{2}}\det(X_j^{-1})\sum_{l=1}^{d+1}\det(X_{j,1},\dots,X_{j,l-1},V_{j,l},X_{j,l+1},\dots,X_{j,d+1}) \\
&= -\frac{1}{2}\det(X_j)^{-\frac{1}{2}}\sum_{l=1}^{d+1}\det(X_j^{-1}(X_{j,1},\dots,X_{j,l-1},V_{j,l},X_{j,l+1},\dots,X_{j,d+1})) \\
&= -\frac{1}{2}\det(X_j)^{-\frac{1}{2}}\sum_{l=1}^{d+1}\det(diag(1,\dots,1,X_{j,l}^{-1}V_{j,l},1,\dots,1)) \\
&= -\frac{1}{2}\det(X_j)^{-\frac{1}{2}}\sum_{l=1}^{d+1}X_{j,l}^{-1}V_{j,l} \\
&= -\frac{1}{2}\det(X_j)^{-\frac{1}{2}}Tr(X_j^{-1}V_j) = -\frac{1}{2}\det(X_j)^{-\frac{1}{2}}\langle X_j^{-1},V_j\rangle
\end{aligned}$$

where *diag* is used to denote a diagonal matrix, whose main diagonal is given by the vector associated to *diag*. We now compute, using what we derived at the question 6:

$$\begin{aligned}
\frac{d}{dt}\exp(-\frac{1}{2}y_i^tX_j(t)^{-1}y_i) &= -\frac{1}{2}\frac{d}{dt}(y_i^tX_j(t)^{-1}y_i)\exp(-\frac{1}{2}y_i^tX_j(t)^{-1}y_i) \\
&= \frac{1}{2}y_i^tX_j(t)^{-1}\frac{d}{dt}(X_j(t))X_j(t)^{-1}y_i\exp(-\frac{1}{2}y_i^tX_j(t)^{-1}y_i) \\
&= \frac{1}{2}y_i^tX_j(t)^{-1}V_jX_j(t)^{-1}y_i\exp(-\frac{1}{2}y_i^tX_j(t)^{-1}y_i) \\
\implies \frac{d}{dt}\exp(-\frac{1}{2}y_i^tX_j(t)^{-1}y_i)|_{t=0} &= \frac{1}{2}y_i^tX_j^{-1}V_jX_j^{-1}y_i\exp(-\frac{1}{2}y_i^tX_j^{-1}y_i) \\
&= \frac{1}{2}Tr(y_i^tX_j^{-1}V_jX_j^{-1}y_i)\exp(-\frac{1}{2}y_i^tX_j^{-1}y_i) \\
&= \frac{1}{2}Tr(X_j^{-1}V_jX_j^{-1}y_iy_i^t)\exp(-\frac{1}{2}y_i^tX_j^{-1}y_i) \\
&= \frac{1}{2}\langle X_j^{-1}V_jX_j^{-1},y_iy_i^t\rangle\exp(-\frac{1}{2}y_i^tX_j^{-1}y_i)
\end{aligned}$$

So going all the way back to the different expressions we computed, and evaluating them at $t = 0$, in the end we get :

$$\begin{aligned}
\frac{d}{dt}(l\circ c)(0) &= -\sum_{i=1}^{n}\frac{\sum_{j=1}^{k}2u_jv_jq(X_j;y_i)+\frac{1}{2}u_j^2q(X_j;y_i)(-\langle X_j^{-1},V_j\rangle+\langle X_j^{-1}V_jX_j^{-1},y_iy_i^t\rangle)}{\sum_{l=1}^{k}u_l^2q(X_l;y_i)} \\
&= \sum_{j=1}^{k}v_j(-2u_j\sum_{i=1}^{n}\frac{q(X_j;y_i)}{\sum_{l=1}^{k}u_l^2q(X_l;y_i)})+\sum_{j=1}^{k}\sum_{i=1}^{n}\frac{\frac{1}{2}u_j^2q(X_j;y_i)(\langle X_j^{-1},V_j\rangle-\langle X_j^{-1}V_jX_j^{-1},y_iy_i^t\rangle)}{\sum_{l=1}^{k}u_l^2q(X_l;y_i)} \\
&= \langle v,-2u\odot w\rangle+\sum_{j=1}^{k}\langle V_j,\frac{1}{2}u_j^2\sum_{i=1}^{n}\frac{q(X_j;y_i)}{\sum_{l=1}^{k}u_l^2q(X_l;y_i)}(X_j-y_iy_i^t)\rangle_{X_j}
\end{aligned}$$

where we denote by w the vector whose entries are $w_j = \sum_{i=1}^{n} \frac{q(X_j; y_i)}{\sum_{l=1}^{k} u_l^2 q(X_l; y_i)}$, and by identification, we immediately get the gradient we wanted to compute at first :

$$(-2u \odot w, (\frac{1}{2} u_j^2 \sum_{i=1}^{n} \frac{q(X_j; y_i)}{\sum_{l=1}^{k} u_l^2 q(X_l; y_i)} (X_j - y_i y_i^t))_{j=1}^{k})$$

But, to get the Riemannian gradient of l, the last thing we need to do is to project this expression to the respective tangent spaces. To do so, for the first entry, we will return $Proj_u(-2u \odot w) = (I - uu^t)(-2u \odot w)$, and for the next k entries, namely the expressions of the form $\frac{1}{2} u_j^2 \sum_{i=1}^{n} \frac{q(X_j; y_i)}{\sum_{l=1}^{k} u_l^2 q(X_l; y_i)} (X_j - y_i y_i^t)$, we will use to projection defined in question 16, therefore, we will compute for every $j = 1, \ldots, k$:

$$Proj_{X_j}(\frac{1}{2} u_j^2 \sum_{i=1}^{n} \frac{q(X_j; y_i)}{\sum_{l=1}^{k} u_l^2 q(X_l; y_i)} (X_j - y_i y_i^t)) =$$

$$\frac{1}{2} u_j^2 \sum_{i=1}^{n} \frac{q(X_j; y_i)}{\sum_{l=1}^{k} u_l^2 q(X_l; y_i)} (X_j - y_i y_i^t) - [\frac{1}{2} u_j^2 \sum_{i=1}^{n} \frac{q(X_j; y_i)}{\sum_{l=1}^{k} u_l^2 q(X_l; y_i)} (X_j - y_i y_i^t)]_{d+1,d+1} X_j E X_j$$

and return for our Riemannian gradient the expression :

$$((I - uu^t)(-2u \odot w), (Proj_{X_j}(\frac{1}{2} u_j^2 \sum_{i=1}^{n} \frac{q(X_j; y_i)}{\sum_{l=1}^{k} u_l^2 q(X_l; y_i)} (X_j - y_i y_i^t)))_{j=1}^{k})$$

24. We recall some useful facts of complexity theory about matrix operations. First of all, the direct application of the mathematical definition of the product of an $n \times m$ matrix with an $m \times l$ matrix gives us a way to compute this matrix product in $O(nml)$ operations. For square matrices of size d, computing their determinant or their inverse has a cost of $O(d^3)$ operations. It can be shown using recurrence and block matrices formulas. Knowing that, we see that for computing $q(X_j; y_i)$ for arbitrary i and j, the most expensive operation is either computing the determinant, or the inverse, of X_j, which can be done in roughly $O(d^3)$, and all the other operations required are either multiplications by some scalar, or matrix-vector multiplications which are done in, respectively, $O(1)$ and $O(d^2)$ which we bound by $O(d^3)$. So computing $q(X_j; y_i)$ can be done in $O(d^3)$ operations, and computing all of them requires $O(nkd^3)$ operations as we have to do that for every $i = 1, \ldots, n$ and $j = 1, \ldots, k$. Therefore, computing $\sum_{l=1}^{k} u_l^2 q(X_l; y_i)$ requires $O(kd^3)$ operations $\forall i = 1, \ldots, n$, and so computing all of them requires $O(nkd^3)$ operations and we now have everything to compute w, which therefore can be done in $O(nkd^3)$ operations. We know observe that we can bound the cost of computing $(Proj_{X_j}(\frac{1}{2} u_j^2 \sum_{i=1}^{n} \frac{q(X_j; y_i)}{\sum_{l=1}^{k} u_l^2 q(X_l; y_i)} (X_j - y_i y_i^t)))_{j=1}^{k}$ by $O(nkd^3)$, observing that once we computed w, all the operations we do for computing each component are just matrix-matrix multiplications, done in $O(d^3)$, vector-vector multiplications, done in $O(d^2)$ operations, matrix addition, done in $O(d^2)$ operations, and scalar multiplication, done in $O(1)$ operations. Now, for the expression $(I - uu^t)(-2u \odot w)$, we see that computing $(-2u \odot w)$ can be done in $O(nkd^3)$ operations, as computing w is clearly the most expensive operation, and the other operations are done in $O(k)$ or $O(1)$ operations. Now, for the term $(I - uu^t)$, as it only consists of vector-vector multiplication and matrix addition but of size k this time, it is done in $O(k^2)$ operations, and therefore, computing $(I - uu^t)(-2u \odot w)$ is done in $O(k^2 + nkd^3)$ operations. In conclusion, computing the Riemannian gradient of f at a point in $\mathcal{M}$ requires roughly $O(k^2 + nkd^3)$ arithmetic operations, which can be upper bounded by $O(nk^2d^3)$.

25. We implement the function in the file `loglikelyhood.m`, and we test it in `questions.m` a cell `Question25`. The output on the Command Window:

```
--- Question 25 ---
M.exp should ideally (but does not have to) be a function handle.
M.log should ideally (but does not have to) be a function handle.
M.pairmean should ideally (but does not have to) be a function handle.
Random tangent vector norm: 1 (should be 1).
norm(v - v)_x = 0 (should be 0).
<u, v>_x = -0.0832719, <v, u>_x = -0.0832719, difference = -4.16334e-17 (should ...
    be 0).
<au+bv, z>_x = -0.031316, a<u, z>_x + b<v, z>_x = -0.031316, difference = ...
    -4.16334e-17 (should be 0).
Norm of tangent vector minus its projection to tangent space: 0 (should be zero).
Couldn't check exp and dist.
Unless otherwise stated, M.vec seems to return real column vectors, as intended.
Checking mat/vec are inverse pairs: 0, 0 (should be two zeros).
Checking if vec is linear: 0 (should be zero).
M.vecmatareisometries says false.
If true, this should be zero: -0.018449.
Testing M.dim() (works best when dimension is small):
    If this number is machine-precision zero, then M.dim() may be too large: ...
        0.00107348
    If this number is not machine-precision zero, then M.dim() may be too small: ...
        5.19833e-16
It is recommended also to call checkretraction.
The slope should be 2. It appears to be: 2.00012.
If it is far from 2, then directional derivatives might be erroneous.
The residual should be 0, or very close. Residual: 7.94411e-15.
If it is far from 0, then the gradient is not in the tangent space.
In certain cases (e.g., hyperbolicfactory), the tangency test is inconclusive.
```

and we also give in Figure 1 the plot returnend by checkgradient.

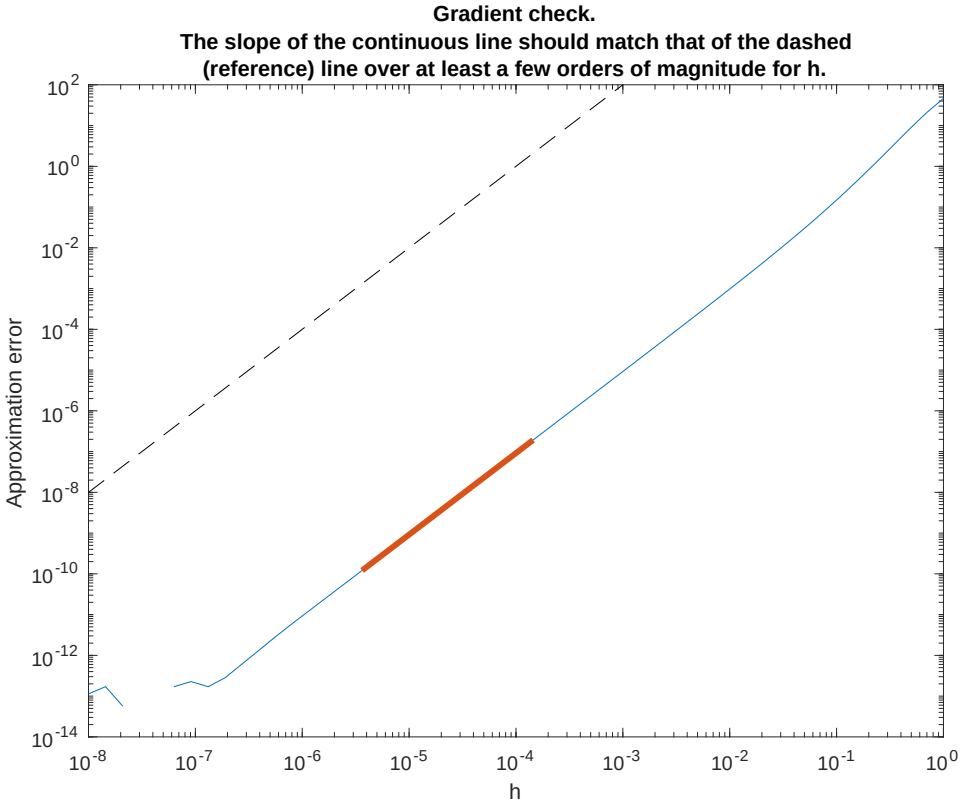

Figure 1: Plot returned by the function checkgradient

2 Generating Data

3 Algorithms

3.1 A performance measure

26. We start by showing that the Hellinger distance is always between 0 and 1. To do so, consider some arbitrary $(\mu_1, \Sigma_1), (\mu_2, \Sigma_2) \in \mathbb{R}^d \times \mathcal{P}_d$. Our first observation is that $\Sigma_{1/2} \in \mathcal{P}_d$ since both Σ_1 and Σ_2 are symmetric positive-definite. It follows that $\Sigma_{1/2}^{-1} \in \mathcal{P}_d$ and $\det(\Sigma_{1/2})$ is non-negative. So we see that the expression $\frac{\det(\Sigma_1)^{1/4}\det(\Sigma_2)^{1/4}}{\det(\Sigma_{1/2})^{1/2}} \exp(-\frac{1}{8}(\mu_1-\mu_2)^t\Sigma_{1/2}^{-1}(\mu_1-\mu_2))$ is always non-negative. Since H is a distance that we assume to be well-defined for every $(\mu_1, \Sigma_1), (\mu_2, \Sigma_2) \in \mathbb{R}^d \times \mathcal{P}_d$, it has to be that $1 - \frac{\det(\Sigma_1)^{1/4}\det(\Sigma_2)^{1/4}}{\det(\Sigma_{1/2})^{1/2}} \exp(-\frac{1}{8}(\mu_1-\mu_2)^t\Sigma_{1/2}^{-1}(\mu_1-\mu_2)) \geq 0 \implies 0 \leq \frac{\det(\Sigma_1)^{1/4}\det(\Sigma_2)^{1/4}}{\det(\Sigma_{1/2})^{1/2}} \exp(-\frac{1}{8}(\mu_1-\mu_2)^t\Sigma_{1/2}^{-1}(\mu_1-\mu_2)) \leq 1$, and it follows that the Helliger distance is always between 0 and 1.
So, if we consider arbitrary parameters $\Theta = (w, (\mu_j, \Sigma_j)_{j=1}^k)$ and some arbitrary ground truth $\Theta^* = (w^*, (\mu_j^*, \Sigma_j^*)_{j=1}^k)$, we first observe that $Err(\Theta, \Theta^*) \geq 0$ as we only sum positive terms, whatever permutation we consider. Now to see that $Err(\Theta, \Theta^*) \leq 2$, we use the fact we showed for the Hellinger distance and see that for any permutation $\sigma \in S(k)$ we have:

$$\begin{aligned}
\sum_{j=1}^{k}(w_j^* H((\mu_{\sigma(j)}, \Sigma_{\sigma(j)}), (\mu_j^*, \Sigma_j^*)) + \frac{1}{2}|w_{\sigma(j)} - w_j^*|) &\leq \sum_{j=1}^{k}(w_j^* + \frac{1}{2}|w_{\sigma(j)} - w_j^*|) \\
&\leq \sum_{j=1}^{k}(w_j^* + \frac{1}{2}(|w_{\sigma(j)}| + |w_j^*|)) \\
&= \sum_{j=1}^{k} w_j^* + \frac{1}{2}\sum_{j=1}^{k} w_{\sigma(j)} + \frac{1}{2}\sum_{j=1}^{k} w_j^* \\
&= 1 + \frac{1}{2} + \frac{1}{2} = 2
\end{aligned}$$

27. If $Err(\Theta, \Theta^*) = 0$ then it does not necessarily imply that $\Theta = \Theta^*$. To illustrate this, we provide an example for the case $k = 2$. Consider a ground truth $\Theta^* = (w^*, (\mu_1^*, \Sigma_1^*), (\mu_2^*, \Sigma_2^*))$, and define the parameters $\Theta = (w, (\mu_1, \Sigma_1), (\mu_2, \Sigma_2))$, where $w_1 = w_2^*$, $w_2 = w_1^*$, $\mu_1 = \mu_2^*$, $\mu_2 = \mu_1^*$, $\Sigma_1 = \Sigma_2^*$ and $\Sigma_2 = \Sigma_1^*$. Clearly, except for few special cases, we have that $\Theta \neq \Theta^*$, but if we consider the permutation $\sigma = (12)$, we have that :

$$\sum_{j=1}^{2}(w_j^* H((\mu_{\sigma(j)}, \Sigma_{\sigma(j)}), (\mu_j^*, \Sigma_j^*)) + \frac{1}{2}|w_{\sigma(j)} - w_j^*|) = \sum_{j=1}^{2}(w_j^* H((\mu_j^*, \Sigma_j^*), (\mu_j^*, \Sigma_j^*)) + \frac{1}{2}|w_j^* - w_j^*|) = 0$$

Showing that $Err(\Theta, \Theta^*) = 0$.
We can generalize this result and see that if $Err(\Theta, \Theta^*) = 0$, most of the time it implies that the estimated parameters are essentially the same as the ground truth, but up to a permutation of its entries, i.e $P(\Theta, \cdot) = P(\Theta^*, \cdot)$.

28. The function Err is implemented in `Err.m`, and is tested in `questions.m` at cell `Question 28`. The output on the Command Window is:

```
               Question 28
    Computation of the total variation distance
    First example
    true     : 0.4467
    computed : 0.446651
    Second example
    true     : 1.1228
    computed : 1.122767
```

3.2 Riemannian gradient descent

29. The Riemannian gradient descent is implemented in `RGD.m`, and is tested in `questions.m` at cell `Question 29`. We give the output on the Command Window for a toy test we did to check if correctly implemented RGD:

```
               Question 29
         iter: 918
     gradnorm: 0.0021
         time: 47.3291
         cost: 328.4698
        alpha: 0.0039
```

3.3 Riemannian conjugate gradient descent

30. See the code given with the report for the implementations.

3.4 Experiments

31. We give in Figure 2 and 3 the plots asked for this question (we used 1000 samples for this question). What we observe is that both methods tend to converge to the same optimal solution, but CGD does it in a significantly smaller number of iterations (RGD does around 4 times more iterations compared to CGD). We additionally observe that CGD terminates because the gradient norm goes below the tolerance, meanwhile RGD terminates because it ran longer than the maximum time allowed.

```
    --- Question 31 ---
    Average running times with k=1, d=2, n=1000, tolgradnorm=0.001000 :
       riemanian GD  : 10.071422 +- 0.046414
       congugated GD : 0.439973 +- 0.098148
```

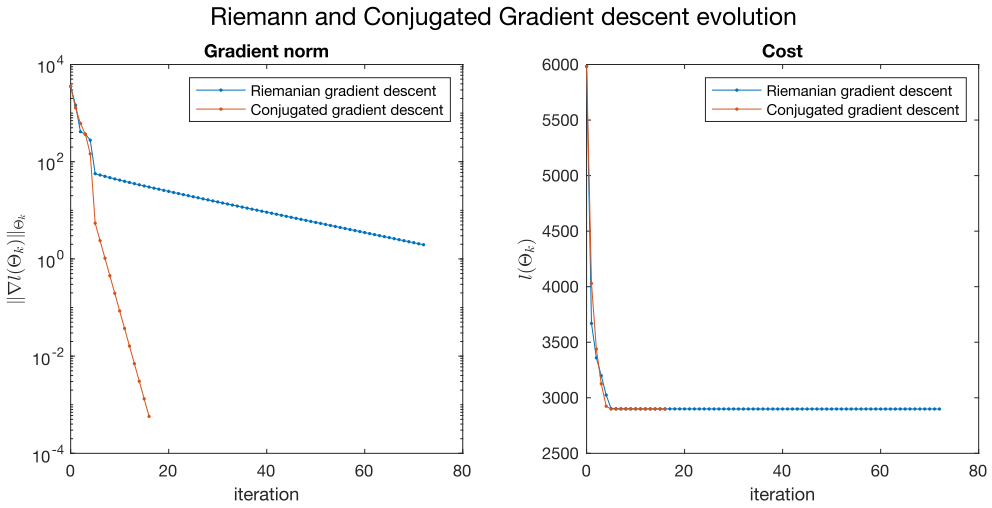

Figure 2: Plots of the norm of the gradient (left) and the negative log likelihood (right) as functions of the iteration number for both RGD and RCG

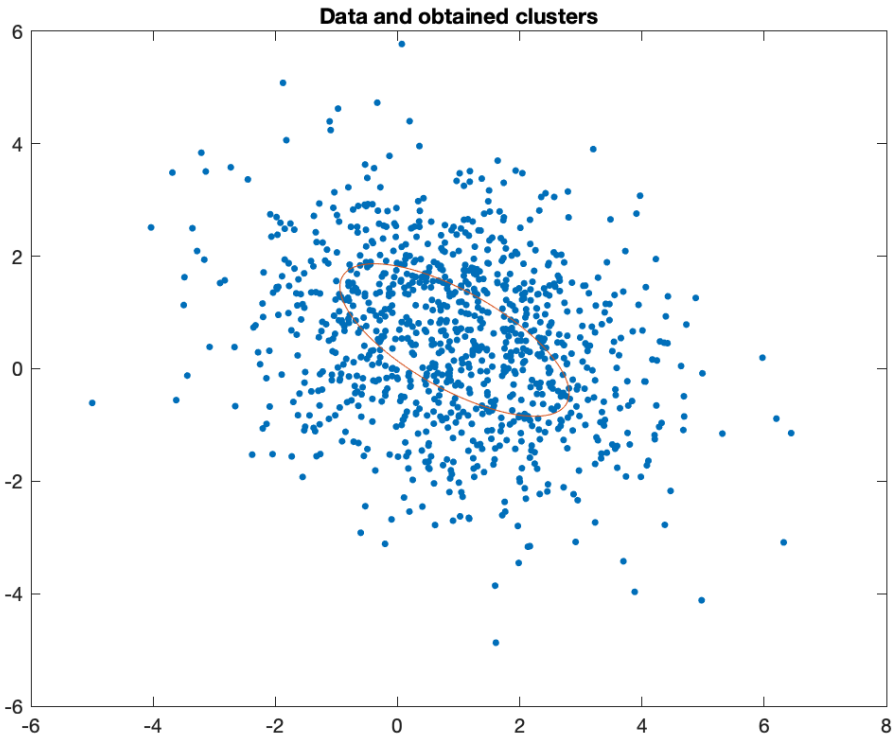

Figure 3: Scatterplot of the data points and the solution found by the algorithm (orange ellipsoid)

0

32. Output on the Command Window:

```
--- Question 32 ---
Average running times with k=2, d=2, n=1000, tolgradnorm=0.001000 :
   riemanian GD  : 9.268081 +- 1.953445
   congugated GD : 6.454152 +- 1.682247
```

Due to lack of time we were not able to run all the experiments asked, so we give here the plots for the experiment we were able to do. Note that we run our algorithms for 1000 data points. What we see is that CGD tends to perform better in precision and in running time than RGD. Empirically, the time to compute each iteration should be quadratic in k, if we recall that the cost of computing the gradient is roughly $O(nk^2d^3)$. We also noticed during our tests that many times, we were not able to find clusters, and got a message telling us that it was due to the fact that covariance matrices were not positive definite. Figure 6 was created out of one of these cases. It seems that our algorithms can converge to a point that lies outside of our manifolds. We were not able to determine if it comes from our code, or if it is something that can happen when optimizing over manifolds.

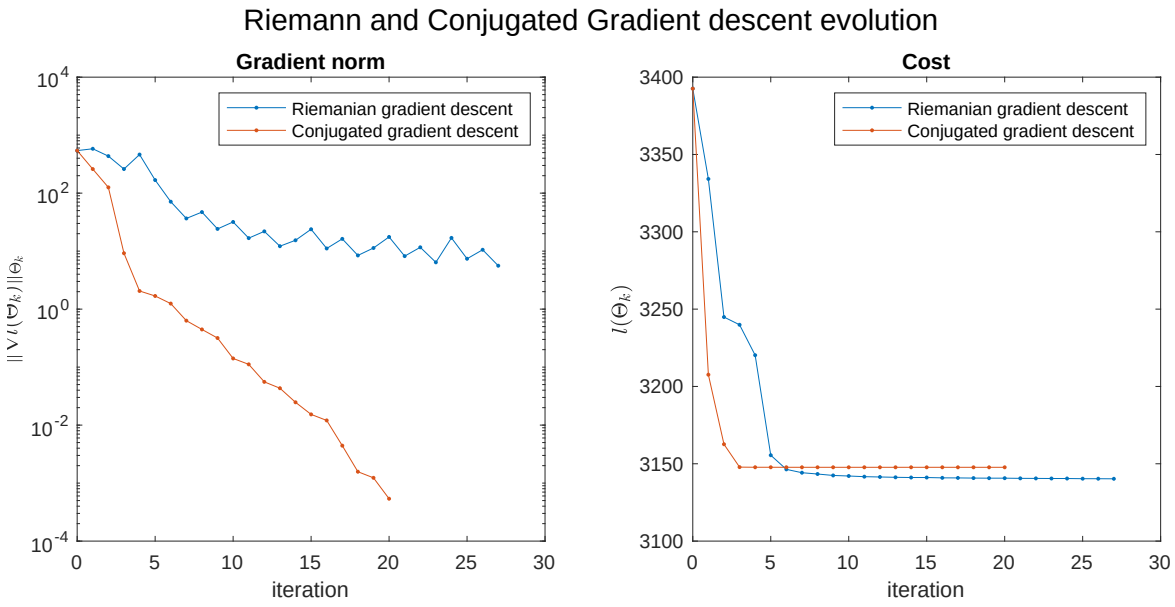

Figure 4: Plots of the norm of the gradient (left) and the negative log likelihood (right) as functions of the iteration number for both RGD and RCG for k = 2

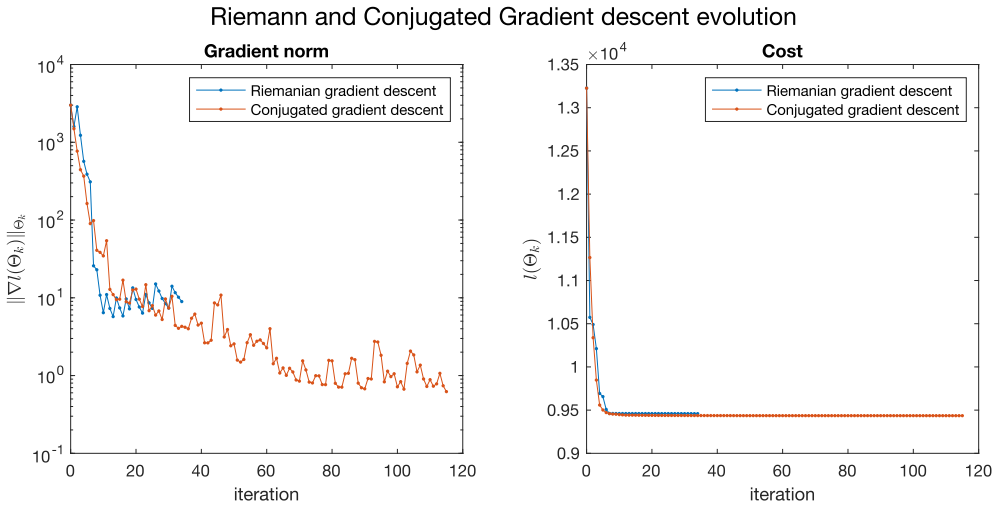

Figure 5: Plots of the norm of the gradient (left) and the negative log likelihood (right) as functions of the iteration number for both RGD and RCG for k = 5

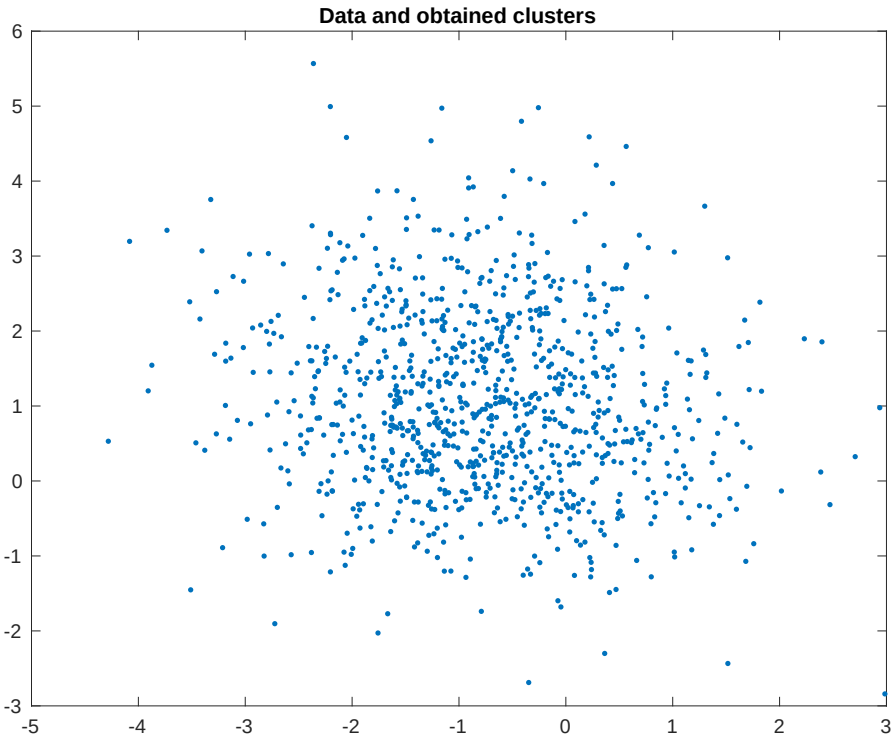

Figure 6: Scatterplot of the data points and the solution found by the algorithm (orange ellipsoid) for k = 2

33. Output on the Command Window:

```
--- Question 33 ---
c)
iter 1/10
iter 2/10
iter 3/10
iter 4/10
iter 5/10
iter 6/10
iter 7/10
iter 8/10
iter 9/10
iter 10/10
Average running times with k=2, d=5, n=1000, tolgradnorm=0.001000 :
   riemanian GD  : 9.587538 +- 1.773939
   congugated GD : 7.593906 +- 3.984803
g)
iter 1/30
Unrecognized field name "w".
Error in Err>Errp (line 16)
```

Due to lack of time we were not able to conduct these experiments, and therefore we can't really say anything about it. But, we could guess that augmenting d should augment the running time, and we also suspect that it could have some impact on the precision for large enough values of d, as we think about the "curse of dimensionality", an issue that very often occurs when doing Machine Learning.

34. We recall again that the cost of computing the gradient is roughly $O(nk^2d^3)$, therefore we can expect the time to scale at a power of 3 with d. Unfortunately we were not able to conduct these experiments properly.

35.

3.5 A comparison between different geometries on $\mathbb{R}_d \times \mathcal{P}_d$

36. We recall that the metric on $\mathbb{R}^d$ is given by : $\langle u, v\rangle_x = u^t v, \forall (x,u),(x,v) \in \mathbb{R}^d \times \mathbb{R}^d$, and the metric on $\mathcal{P}_d$ be given by $\langle A, B\rangle_X = Tr(X^{-1}AX^{-1}B), \forall (X,A),(X,B) \in \mathcal{P}_d \times Sym_d$. Now, using the problem in the exercise session about product Riemannian metric, we have that the metric metric on $\mathcal{M} = \mathbb{R}^d \times \mathcal{P}_d$ is given by :

$$\langle \dot{\Theta}_1, \dot{\Theta}_2\rangle_\Theta = Tr(\Sigma^{-1}\dot{\Sigma}_1\Sigma^{-1}\dot{\Sigma}_2) + \dot{\mu}_1^t\dot{\mu}_2$$

where $\Theta = (\mu, \Sigma) \in \mathbb{R}^d \times \mathcal{P}_d$, and $\dot{\Theta}_i = (\dot{\mu}_i, \dot{\Sigma}_i) \in T_\Theta\mathcal{M} = \mathbb{R}^d \times Sym_d$. The only difference with the FR metric we previously defined is that here we have a component of the form $\dot{\mu}_1^t\dot{\mu}_2$ instead of $2\dot{\mu}_1^t\Sigma^{-1}\dot{\mu}_2$, suggesting that in the FR metric we used a different inner product on $\mathbb{R}^d$. We assume that this difference will have an impact when we will run the Riemannian nonlinear conjugate gradient algorithm in question 38.

37. Our cost function is defined by $l(u, (\mu_j, \Sigma_j)_{j=1}^k) = -\sum_{i=1}^n \log(\sum_{j=1}^k u_j^2 p_d(\mu_j, \Sigma_j; x_i))$ for $(u, (\mu_j, \Sigma_j)_{j=1}^k) \in \mathcal{S}^{k-1} \times \prod_{j=1}^k (\mathbb{R}^d \times \mathcal{P}_d)$. We will follow the same steps as for question 23. Due to lack of time, some steps of the computation are omitted, but when they do it is because they are analogous to what has been done in question 23. We let $(v, (\dot{\mu}_j, \dot{\Sigma}_j)_{j=1}^k) \in T_u\mathcal{S}^{k-1} \times \prod_{j=1}^k (\mathbb{R}^d \times Sym_d)$ and define the

smooth curve $c(t) = (u, (\mu_j, \Sigma_j)_{j=1}^k) + t(v, (\dot{\mu}_j, \dot{\Sigma}_j)_{j=1}^k)$, and define to ease the notations $u(t) = u + tv$, $u_j = u_j + tv_j, \forall j = 1, \dots, k$, $\mu_j(t) = \mu_j + t\dot{\mu}_j, \forall j = 1, \dots, k$ and $\Sigma_j(t) = \Sigma_j + t\dot{\Sigma}_j, \forall j = 1, \dots, k$. We now have that :

$$\frac{d}{dt}(l \circ c)(t) = -\sum_{i=1}^{n} \frac{\sum_{j=1}^k 2u_j(t)v_j p_d(\mu_j(t), \Sigma_j(t); x_i) + u_j(t)^2 \frac{d}{dt}(p_d(\mu_j(t), \Sigma_j(t); x_i))}{\sum_{l=1}^k u_l(t)^2 p_d(\mu_j(t), \Sigma_j(t); x_i)}$$

We now compute :

$$\frac{d}{dt}(p_d(\mu_j(t), \Sigma_j(t); x_i)) = (2\pi)^{-\frac{d}{2}}(\frac{d}{dt}(\det(\Sigma_j(t))^{-\frac{1}{2}}) \exp(-\frac{1}{2}(x_i - \mu_j(t))^t \Sigma_j(t)^{-1}(x_i - \mu_j(t))) + \det(\Sigma_j(t))^{-\frac{1}{2}} \frac{d}{dt}(\exp(-\frac{1}{2}(x_i - \mu_j(t))^t \Sigma_j(t)^{-1}(x_i - \mu_j(t))))$$

Now, following the exact same steps as in question 23 we immediately get that :

$$\frac{d}{dt}(\det(\Sigma_j(t))^{-\frac{1}{2}})|_{t=0} = -\frac{1}{2}\det(\Sigma_j)^{-1/2} Tr(\Sigma_j^{-1}\dot{\Sigma}_j)$$

Now, to differentiate the exponential, we will basically follow what has been done in question 23, with the exception that we will use the product rule. We skip the details of the computation, as they are tedious, and immediately give the result evaluated at $t = 0$:

$$\frac{d}{dt}(\exp(-\frac{1}{2}(x_i - \mu_j(t))^t \Sigma_j(t)^{-1}(x_i - \mu_j(t))))|_{t=0} = -\frac{1}{2}(-(x_i - \mu_j)^t \Sigma_j^{-1}\dot{\Sigma}_j\Sigma_j^{-1}(x_i - \mu_j) + 2\dot{\mu}_j^t\Sigma_j^{-1}(\mu_j - x_i)) \exp(-\frac{1}{2}(x_i - \mu_j)^t\Sigma_j^{-1}(x_i - \mu_j))$$

We then obtain, after evaluating to $t = 0$ and reformulating the obtained expression to make the inner product on $\mathbb{R}^d$ and the FR metric appear :

$$\frac{d}{dt}(p_d(\mu_j(t), \Sigma_j(t); x_i))|_{t=0} = p_d(\mu_j, \Sigma_j; x_i)(-\frac{1}{2}\langle\dot{\Sigma}_j, \Sigma_j - (x_i - \mu_j)(x_i - \mu_j)^t\rangle_{\Sigma_j} - \langle\dot{\mu}_j, \Sigma_j^{-1}(\mu_j - x_i)\rangle)$$

Now, when we plug it in the primary expression we wanted to compute, and evaluate to $t = 0$, after many simplifications we get to the following expression (where $\langle\cdot,\cdot\rangle_{\mathbb{R}^d \times \mathcal{P}_d}$ is the metric we defined in the previous question, and $\langle\cdot,\cdot\rangle$ is the regular inner product on $\mathbb{R}^d$):

$$\frac{d}{dt}(l \circ c)(0) = \langle v, -2u \odot w\rangle + \sum_{j=1}^k \langle(\dot{\mu}_j, \dot{\Sigma}_j), \frac{u_j^2}{2}\sum_{i=1}^n \frac{p_d(\mu_j, \Sigma_j; x_i)}{\sum_{l=1}^k u_l^2 p_d(\mu_l, \Sigma_l; x_i)}(2\Sigma_j^{-1}(\mu_j - x_i), \Sigma_j - (x_i - \mu_j)(x_i - \mu_j)^t)\rangle_{\mathbb{R}^d \times \mathcal{P}_d}$$

where we denote by w the vector whose entries are $w_j = \sum_{i=1}^n \frac{p_d(\mu_j, \Sigma_j; x_i)}{\sum_{l=1}^k u_l^2 p_d(\mu_l, \Sigma_l; x_i)}$, and by identification, we immediately get the gradient we wanted to compute at first :

$$((I - uu^t)(-2u \odot w), (\frac{u_j^2}{2}\sum_{i=1}^n \frac{p_d(\mu_j, \Sigma_j; x_i)}{\sum_{l=1}^k u_l^2 p_d(\mu_l, \Sigma_l; x_i)}(2\Sigma_j^{-1}(\mu_j - x_i), \Sigma_j - (x_i - \mu_j)(x_i - \mu_j)^t))_{j=1}^k)$$

where we only projected the first component to the right tangent space, because we observe that all the other k components of the form $\frac{u_j^2}{2}\sum_{i=1}^n \frac{p_d(\mu_j, \Sigma_j; x_i)}{\sum_{l=1}^k u_l^2 p_d(\mu_l, \Sigma_l; x_i)}(2\Sigma_j^{-1}(\mu_j - x_i), \Sigma_j - (x_i - \mu_j)(x_i - \mu_j)^t)$ are already in their respective tangent spaces.

38.

3.6 Incorporating additional information

39. If we know a priori that the mean vectors μ_j^* of all the Gaussians lie on a given embedded submanifold of $\mathbb{R}^d$, call it $\mathcal{N}$, we could try to exploit the geometric properties of $\mathcal{N}$, that is, give it a riemannian metric and instead of solving the optimization problem (MLE4) on $\mathcal{S}^{k-1} \times \prod_{j=1}^{k} (\mathbb{R}^d \times \mathcal{P}_d)$, we would solve it on $\mathcal{S}^{k-1} \times \prod_{j=1}^{k} (\mathcal{N} \times \mathcal{P}_d)$ (note that the expression for the gradient of the cost function may differ from what we computed in question 37). Knowing this additional geometric constraint, it could possibly be that when we run our optimization algorithms on this new formulation of the problem, we would gain in efficiency or/and precision.

References

[1] Nicolas Boumal. *An introduction to optimization on smooth manifolds*. Cambridge University Press, 2023. DOI: 10.1017/9781009166164. URL: https://www.nicolasboumal.net/book.

[2] Benoît Müller and Thomas Renard. *opti-manifolds-GMM*. Version v1. 2023. URL: https://github.com/Benoit-Muller/opti-manifolds-GMM.