Preliminaries and notation

Let $PD(p)$ denote the set of positive definite $p \times p$ matrices, and consider the class

$$\mathcal{N}_p = \{N_p(\mu, \Sigma) | \mu \in \mathbb{R}^p,\ \Sigma \in PD(p)\}$$

of p-variate normal distributions with mean μ and covariance Σ.
As demonstrated in ... $\mathcal{N}_p$ can be considered as a Riemannian manifold, when equipped with the Fisher metric. The present paper describes an algorithm for determining the geodesic distance between two of these distributions.
The functional form of a geodesic curve between two multivariate normals are described in ... and has been intensively studied in ...
We start by reviewing notation and fundamental results.
Let $d_F((\mu_1, \Sigma_1), (\mu_2, \Sigma_2))$ denote the geodesic distance between two multivariate normals. Then the invariance of the Fisher metric under affine transformations means that

$$d_F((\mu_1, \Sigma_1), (\mu_2, \Sigma_2)) = d_F((0, I_p), (U(\mu_2 - \mu_1), U\Sigma_2 U^*)$$

where U is a square root of Σ_1^{-1} i.e. $UU^* = \Sigma_1^{-1}$, where U^* is the transpose of U. This means that we can limit the study to geodesics extending from $(0, I_p)$.
It turns out that describing the geodesics is a bit more straigthforward in terms of the socalled canonical parameters

$$\Delta = \Sigma^{-1} \text{ and } \delta = \Sigma^{-1}\mu$$

Our aim is to determine the distance

$$D(\delta, \Delta) = d_F((0, I_p), (\Delta^{-1}\delta, \Delta^{-1}))$$

Let $\{(\delta(t), \Delta(t)) | t \in \mathbb{R}\}$ denote a geodesic with $(\delta(0), \Delta(0)) = (0, I_p)$. Adopting the notation, where the "dot"-operator means differentiation wrt t, then as noted in ... the geodesic is a solution to the differential equation

$$\dot{\delta}(t) = B\delta(t) + \varepsilon(t)x \text{ and } \dot{\Delta}(t) = B\Delta(t) + x\delta(t)^* \tag{1}$$

where $\varepsilon(t) = 1 + \delta(t)^*\Delta(t)^{-1}\delta(t)$ and B is a symmetric $p \times p$ matrix giving the initial direction of $\Delta(t)$ and x a p-vector giving the initial direction of $\delta(t)$.
Now, define the matrix

$$A = \begin{Bmatrix} B & x & 0 \\ x^* & 0 & -x^* \\ 0 & -x & -B \end{Bmatrix}$$

Then as shown in ... the geodesic can be extracted from

$$\Lambda(t) = \exp(At) = \begin{Bmatrix} \Delta(t) & \delta(t) & \Phi(t) \\ \delta(t)^* & \varepsilon(t) & \gamma(t)^* \\ \Phi(t)^* & \gamma(t) & \Gamma(t) \end{Bmatrix} \tag{2}$$

Note that

$$\Lambda(-t) = \Lambda(t)^{-1} = \begin{Bmatrix} \Gamma(t) & \gamma(t) & \Phi(t)^* \\ \gamma(t)^* & \varepsilon(t) & \delta(t)^* \\ \Phi(t) & \delta(t) & \Delta(t) \end{Bmatrix} \tag{3}$$

which means that $(\gamma(t), \Gamma(t)) = (\delta(-t), \Delta(-t))$ is the normal distribution in the opposite direction $(-B, -x)$ with the same distance to $(0, I_p)$. The interpretation of $\Phi(t)$ is as yet unclear.
As shown in ... the distance from $(0, I_p)$ to $(\delta, \Delta) = (\delta(1), \Delta(1))$ is given by $D(\delta, \Delta) = \int_0^1 \sqrt{f(t)}dt$, where

$$f(t) = \frac{1}{2}\mathrm{tr}(\Delta(t)\dot{\Sigma}(t)\Delta(t)\dot{\Sigma}(t)) + \dot{\mu}(t)^*\Delta(t)\dot{\mu}(t)$$

The differential equation (1) is equivalent to

$$\dot{\mu}(t) = \Sigma(t)x \quad \text{and} \quad \Sigma(t)^{-1}\dot{\Sigma}(t) = -B - x\mu(t)^*$$

Inserting into f then yields

$$f(t) = \frac{1}{2}\mathrm{tr}(B^2) + \frac{1}{2}(\mu(t)^*x)^2 + \mu(t)^*Bx + x^*\Sigma(t)x$$

Now

$$\frac{d}{dt}f(t) = \mu(t)^*xx^*\Sigma(t)x + x^*\Sigma(t)Bx - x^*(\Sigma(t)B + \Sigma(t)x\mu(t)^*)x = 0$$

Hence $f(t) = f(0) = \frac{1}{2}\mathrm{tr}(B^2) + x^*x$ and

$$D(\delta, \Delta) = \sqrt{\frac{1}{2}\mathrm{tr}(B^2) + x^*x} = \frac{1}{2}\sqrt{\mathrm{tr}(A^2)} \tag{4}$$

Calculating geodesic distance

Equation (3) can be used to define a mapping taking (B, x) to $(\delta(1), \Delta(1))$. If we were able to give a "simple" expression of the inverse, then the distance is given by (4).
There does not seem to be a "simple" solution, so instead we shall develop an algorithm that approximates the solution.
So we are given (δ, Δ) and want to calculate $D(\delta, \Delta)$, which means that we need to determine (B, x) mapping to $(\delta(1), \Delta(1)) = (\delta, \Delta)$.
Let us first remark that the relation $\Lambda(-t)\Lambda(t) = I_p$ implies by (2) and (3) that

$$\Delta\Phi^* + \delta\delta^* + \Phi\Delta = 0$$

where we subsequently will suppress dependence on t.
Looking at this expression, it is tempting to guess the value of Φ:

$$\Phi_0 = -\frac{1}{2}\delta\delta^*\Delta^{-1}$$

Another implication of $\Lambda(-t)\Lambda(t) = I_p$ is

$$\Delta\gamma + \varepsilon\delta + \Phi\delta = 0$$

Letting $\Phi = \Phi_0$ and solving for γ yields

$$\gamma_0 = -(1 + \frac{1}{2}\tau)\Delta^{-1}\delta$$

where $\tau = \delta^*\Delta^{-1}\delta$. Finally, $\Lambda(-t)\Lambda(t) = I_p$ implies

$$\Delta\Gamma + \delta\gamma^* + \Phi^2 = I_p$$

Letting $(\gamma, \Phi) = (\gamma_0, \Phi_0)$ and solving for Γ yields

$$\Gamma_0 = \Delta^{-1} + (1 + \frac{1}{4}\tau)\Delta^{-1}\delta\delta^*\Delta^{-1}$$

In summary

$$\begin{aligned} \Phi_0 &= -\frac{1}{2}\delta\delta^*\Delta^{-1} \\ \gamma_0 &= -(1 + \frac{1}{2}\sigma^2)\Delta^{-1}\delta \\ \Gamma_0 &= \Delta^{-1} + (1 + \frac{1}{4}\tau)\Delta^{-1}\delta\delta^*\Delta^{-1} \end{aligned} \tag{5}$$

If we define

$$\Lambda_0 = \begin{Bmatrix} \Delta & \delta & \Phi_0 \\ \delta & \varepsilon & \gamma_0^* \\ \Phi_0^* & \gamma_0 & \Gamma_0 \end{Bmatrix} \tag{6}$$

we may define a "distance" by

$$D_0(\delta, \Delta) = \frac{1}{2}\sqrt{\mathrm{tr}(\log(\Lambda_0))} \tag{7}$$

where log is the principal matrix logarithm.
In general this is not the geodesic distance. However if δ is an eigenvector of Δ, then D_0 may be verified to be the geodesic distance.
Adopting the notation in ... - except that B is substituted by $-B$ -we define G to be a symmetric square root of $B^2 + 2xx^*$ and let G^- denote a generalized inverse of G. If x is an eigenvector of B, then by an orthogonal rotation we may assume that B is diagonal with diagonal elements $b_1, \ldots, b_p$ and that only the first coordinate x_1 of x is different from zero. This means that G is diagonal with elements $g_1 = \sqrt{2x_1^2 + b_1^2}, g_2 = |b_2|, \ldots, g_p = |b_p|$). In this case Theorem 1

in ... means that if $g_1 > 0$ then

$$\begin{aligned}
\Delta_{11} &= 1 + \frac{1}{2}(1 + \frac{b_1^2}{g_1^2})(\cosh(g_1) - 1) + \frac{b_1}{g_1}\sinh(g_1) \\
\Delta_{ii} &= \exp(g_i) \quad i > 1 \\
\Delta_{ij} &= 0 \quad i \neq j \\
\delta_1 &= \frac{x_1 b_1}{g_1^2}(\cosh(g_1) - 1) + \frac{x_1}{g_1}\sinh(g_1) \\
\delta_i &= 0 \quad i > 1 \\
\Phi_{11} &= -\frac{1}{2}\frac{x_1^2}{g_1^2}(\cosh(g_1) - 1) \\
\Phi_{ij} &= 0 \quad (i,j) \neq (1,1)
\end{aligned} \tag{8}$$

where sinh and cosh are hyperbolic sine and cosine. We note that x being an eigenvector of B is equivalent to δ being an eigenvector of Δ. Tedious calculations reveal that $-2\Delta_{11}\Phi_{11} = \delta_1^2$, which means that $\Phi = -\frac{1}{2}\Delta^{-1}\delta\delta^* = -\frac{1}{2}\delta\delta^*\Delta^{-1}$,i.e. $\Phi = \Phi_0$ and hence $(\gamma, \Gamma) = (\gamma_0, \Gamma_0)$. Remarking that $\Gamma = \Delta(-B, -x)$ we obtain from ... that

$$\frac{1}{2}(\Delta + \Gamma) - \Phi = \cosh(G)$$

where cosh is matrix hyperbolic cosine. Hence

$$\begin{gathered}
G = \operatorname{acosh}(\frac{1}{2}(\Delta + \Delta^{-1}) + \frac{1}{2}[(1 + \frac{1}{4}\tau)\Delta^{-1}\delta\delta^*\Delta^{-1} + \delta\delta^*\Delta^{-1}]) \\
D(\delta, \Delta) = \sqrt{\frac{1}{2}\operatorname{tr}(G^2)}
\end{gathered} \tag{9}$$

As an example, consider the distance between (μ_1, Σ) and $(\mu_2, \sigma^2\Sigma)$, i.e. the covariances are proportional.
In this case $\delta = U(\mu_2 - \mu_1)$, $\Delta = \sigma^2 I_p$, where U is a square root of Σ. By a rotation we may assume $\delta^* = (\|\delta\|, 0, \ldots, 0)$ which makes G diagonal with

$$\begin{aligned}
G_{11} &= \operatorname{acosh}(\frac{1}{2}(\sigma^2 + \sigma^{-2} + (1 + \sigma^{-2} + \frac{1}{4}\sigma^{-2}\|\delta\|^2)\|\delta\|^2) \\
G_{ii} &= \operatorname{acosh}(\frac{1}{2}(\sigma^2 + \sigma^{-2}) \quad i > 1
\end{aligned}$$

where

$$\begin{gathered}
\|\delta\|^2 = (\mu_2 - \mu_1)^*\Sigma^{-1}(\mu_2 - \mu_1) \\
D(\delta, \Delta) = \sqrt{\frac{1}{2}(G_{11}^2 + (p-1)G_{22}^2)}
\end{gathered}$$

If $p = 1$, i.e. the univariate case, this reduces to $\frac{\sqrt{2}}{2}G_{11}$, where

$$G_{11} = \operatorname{acosh}(\frac{1}{2}(\frac{\sigma_1^2}{\sigma_2^2} + \frac{\sigma_2^2}{\sigma_1^2} + (\mu_1 - \mu_2)^2(\sigma_1^{-2} + \sigma_2^{-2}) + \frac{1}{4}(\mu_1 - \mu_2)^4\sigma_1^{-2}\sigma_2^{-2}))$$

which may be rewritten to give the distance

$$D((\mu_1,\sigma_1^2),(\mu_2,\sigma_2^2)) = \frac{\sqrt{2}}{2}\text{acosh}(\frac{1}{2}(\frac{(\mu_1-\mu_2)^2}{2\sigma_1\sigma_2}+\frac{\sigma_1}{\sigma_2}+\frac{\sigma_2}{\sigma_1})^2-1) \tag{10}$$

Now consider the situation where we have two univariate normal samples of size n and want to test

$$H_0: \quad (\mu_1,\sigma_1) = (\mu_2,\sigma_2)$$

If the hat-operator means maximum likelihood estimation, then it may be verified that the likelihood ratio statistic(LRT) fullfils

$$D((\hat{\mu}_1,\hat{\sigma}_1),(\hat{\mu}_2,\hat{\sigma}_2)) = \frac{\sqrt{2}}{2}\text{acosh}(2LRT((\hat{\mu}_1,\hat{\sigma}_1),(\hat{\mu}_2,\hat{\sigma}_2))^{-\frac{2}{n}}-1)$$

i.e. the geodesic distance is in this case equivalent to the likelihood ratio statistic. Subsequently we assume wlog that Δ is diagonal and define $D=\sqrt{\Delta}$, $d=D^{-1}\delta$ and

$$T = \begin{Bmatrix} D & d & Z \\ 0 & 1 & -d^*D^{-1} \\ 0 & 0 & D^{-1} \end{Bmatrix} \tag{11}$$

which shall fullfil that

$$T^{-1} = \begin{Bmatrix} D^{-1} & -D^{-1}d & Z^* \\ 0 & 1 & d^* \\ 0 & 0 & D \end{Bmatrix}$$

Then for a suitable choise of Z we have that $T^*T = \Lambda(1)$.
The relation $TT^{-1} = I_p$ means that

$$DZ^* + dd^* + ZD = 0$$

so that if $D = \text{diag}(D_1,\ldots,D_p)$ and $d^* = (d_1,\ldots,d_p)$ then

$$D_iZ_{ji} + d_id_j + D_jZ_{ij} = 0$$

Hence

$$\begin{aligned} Z_{ii} &= -\frac{d_i^2}{2D_i} \quad i=1,\ldots,p \\ Z_{ji} &= -\frac{d_id_j + D_jZ_{ij}}{D_i} \quad 1\le i<j\le p \end{aligned} \tag{12}$$

i.e the diagonal of Z is known and the lower triangular part is determined from the upper triangular part.
Let y denote the upper triangular part of Z and consider Z as a function $Z(y)$ of y as defined by (12). We define $T(y)$ by inserting $Z(y)$ into (11).
Now, let

$$A(y) = log(T(y)^*T(y)) = \begin{Bmatrix} B(y) & x(y) & C(y) \\ x(y)^* & 0 & -x(y)^* \\ C(y)^* & -x(y) & -B(y) \end{Bmatrix}$$

If $C(y) = 0$ then the distance is given by $\frac{1}{2}\sqrt{\mathrm{tr}(A(y)^2)}$. So we aim at finding y such that the function

$$f(y) = \mathrm{tr}(C(y)^2)$$

is zero.
To this end we implement an algorihm for minimizing f, i.e. determining a value of y such that $f(y) = 0$.