Abstract. Given two samples from a Wishart distribution with same covariance matrix Σ as parameter, the distribution of some distances between them are estimated numerically. In the case of the Log-Euclidean distance the distribution appears to be independent from the parameter Σ, hence from its distribution it's possible to deduce a test for the equality of two covariance matrices.

1 Distances

Given the space of SDP matrices, let's consider the following distances between matrices, introduced by [Dryden et al, p.1112]:

Name	**Notation**	**Form**
Euclidean	$d_E(S_1, S_2)$	$\|\|S_1 - S_2\|\|$
Log-Euclidean	$d_L(S_1, S_2)$	$\|\|log(S_1) - log(S_2)\|\|$
Riemannian	$d_R(S_1, S_2)$	$\|\|log(S_1^{-1/2} S_2 S_1^{-1/2})\|\|$
Cholesky	$d_C(S_1, S_2)$	$\|\|chol(S_1) - chol(S_2)\|\|$
Root Euclidean	$d_H(S_1, S_2)$	$\|\|S_1^{1/2} - S_2^{1/2}\|\|$
Procrustes size-and-shape	$d_S(S_1, S_2)$	$inf_R \|\|L_1 - L_2 R\|\|$

Two other distances are not considered, because both their definitions depend on a parameter (β and α).

Let's consider the case:

$$\underline{x}_1, ..., \underline{x}_n \overset{iid}{\sim} N(\underline{0}_p, \Sigma)$$

$$\underline{y}_1, ..., \underline{y}_m \overset{iid}{\sim} N(\underline{0}_p, \Sigma)$$

then:

$$S_x := \sum_i \underline{x}_i \underline{x}_i^T \sim Wishart(\Sigma, n)$$

$$S_y := \sum_i \underline{y}_i \underline{y}_i^T \sim Wishart(\Sigma, m)$$

All the above distances $d(\cdot, \cdot)$ have distribution $F(n, m, p, \Sigma)$. The question is:

$$d(S_x, S_y) \overset{?}{\sim} F(n, m, p)$$

i.e. does its distribution depend only on the parameters n, m and p, or does it depend also on the parameter Σ?
This question is interesting because in most of the cases Σ is unknown, while instead the other parameters are known.

2 Simulation

In some cases the answer is negative. Let's take for instance the Euclidean distance. If you changed the parameter Σ to 2Σ, the distance would double:

$$d(2S_x, 2S_y) = ||2S_x - 2S_y|| = 2||S_x - S_y|| = 2d(S_x, S_y)$$

In some other cases, the answer is less trivial. Theoretically it's possible to compute the density of $d(S_x, S_y)$ and then answer the question. Indeed, [Anderson, p.255] gives an explicit formula of the Wishart distribution.
However, I rather prefer to get a hint of the answer with the following empirical method:

- n, m and p are chosen randomly and fixed.
- 10 different parameters Σ_i are picked randomly.
- For every Σ_i, the PDF of $d(S_x, S_y)$ is simulated picking 10000 samples of S_x and S_y from the distributions $Wishart(\Sigma_i, n)$ and $Wishart(\Sigma_i, m)$
- The 10 different PDFs are plotted and compared qualitatively
- If they look different from one another, then this is an empirical evidence that the distribution depends on the parameter Σ.

In the following plot 10 different Σ produced the shape of 10 different PDFs for $d_E(S_x, S_y)$, one for each colour. It's quite evident that they do not come from the same distribution [cfr. Github for the code].

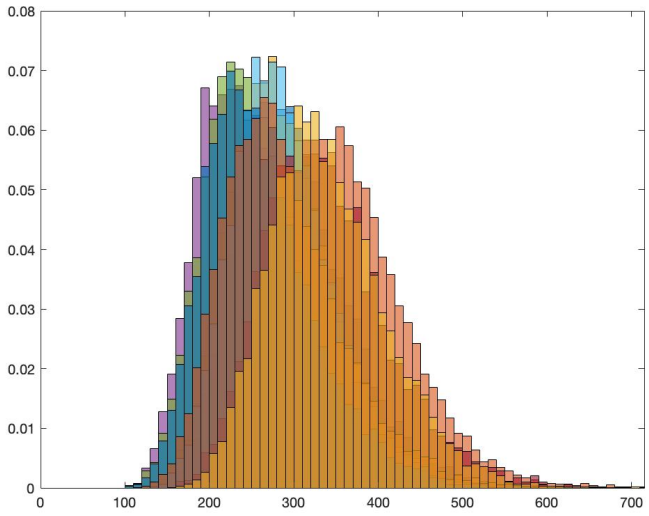

Fig. 1. $d_E(S_x, S_y)$

In the following plot 10 different Σ produced the shape of 10 different PDFs for $d_L(S_x, S_y)$. Here there is no qualitative evidence that the distributions are different, hence suggesting the hypotesis that the PDF for the Log-Euclidean distance does not depend on the parameter Σ.

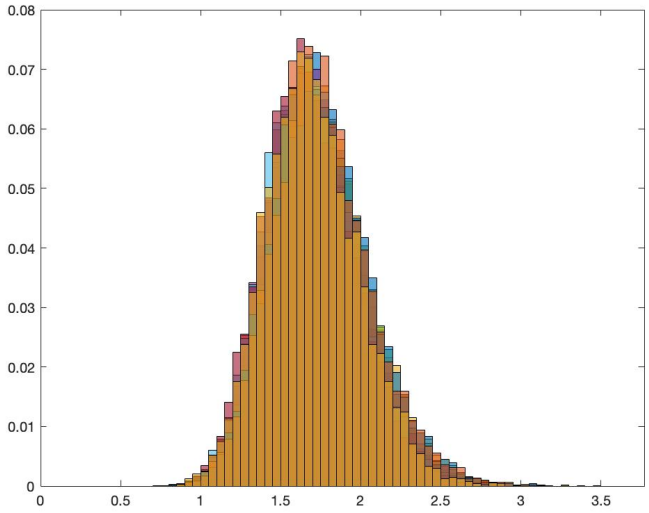

Fig. 2. $d_L(S_x, S_y)$

If the simulation is replicated for every distance (cfr. Appendix for details), the result is the following:

Name	**Notation**	**Form**	**Independence from Σ**
Euclidean	$d_E(S_1, S_2)$	$\|\|S_1 - S_2\|\|$	NO
Log-Euclidean	$d_L(S_1, S_2)$	$\|\|log(S_1) - log(S_2)\|\|$	YES
Riemannian	$d_R(S_1, S_2)$	$\|\|log(S_1^{-1/2} S_2 S_1^{-1/2})\|\|$	YES
Cholesky	$d_C(S_1, S_2)$	$\|\|chol(S_1) - chol(S_2)\|\|$	NO
Root Euclidean	$d_H(S_1, S_2)$	$\|\|S_1^{1/2} - S_2^{1/2}\|\|$	NO
Procrustes size-and-shape	$d_S(S_1, S_2)$	$inf_R \|\|L_1 - L_2 R\|\|$	NO

The fact that the PDFs of d_E, d_C, d_H, d_S depend on the parameter Σ is not surprising. Analytically it's not difficult to prove this dependence. There should be at least a scale factor, e.g. such as $1/tr(S_x + S_y)$, to have some PDF without dependence from Σ.
Instead the results for d_L and d_R are quite surprising, because they are definitely not straightforward.

From now on, I'll focus on d_L and I conjecture that its PDF does not depend on Σ.

3 Test for equivalence of two covariance matrices

Given the **unproved assumption** that the PDF of d_L does not depend on the parameter Σ, it's possible to estimate it with the following simulation. Since all Σ are equivalent, the identity matrix I_p is chosen.

$$\underline{x}_1, ..., \underline{x}_n \overset{iid}{\sim} N(\underline{0}_p, I_p)$$

$$\underline{y}_1, ..., \underline{y}_m \overset{iid}{\sim} N(\underline{0}_p, I_p)$$

which is equivalent to:

$$x_{ij}, y_{ij} \overset{iid}{\sim} N(0, 1)$$

It's quite easy from this to simulate $S_x := \sum_i \underline{x}_i \underline{x}_i^T$ and $S_y := \sum_i \underline{y}_i \underline{y}_i^T$, and so the distribution of $d_L(S_x, S_y)$.
[Anderson,p.253] provides the explicit distribution for $[S_x]_{ij}$ and $[S_y]_{ij}$, but I found it not efficient during the simulation.

Given the following assumption:

$$\underline{x}_1, ..., \underline{x}_n \overset{iid}{\sim} N(\underline{0}_p, \Sigma_1)$$

$$\underline{y}_1, ..., \underline{y}_m \overset{iid}{\sim} N(\underline{0}_p, \Sigma_2)$$

and the test $\Sigma_1 = \Sigma_2$ vs $\Sigma_1 \neq \Sigma_2$, which can be rewritten in the following way:

$$H_0 : d_L(\Sigma_1, \Sigma_2) = 0$$
$$H_1 : d_L(\Sigma_1, \Sigma_2) > 0$$

the test statistic $d_L(S_1, S_2)$ has known distribution $F(n, m, p)$ above achieved through simulation.
Therefore it's possible to compute the p-value for the test.
After writing the code for the test, I ran two different simulations [cfr. Github] multiple times, first with:

$$\Sigma_1 = \Sigma_2 = \begin{bmatrix} 8 & 1 & -1 & 1 \\ 1 & 6 & 1 & -2 \\ -1 & 1 & 6 & 1 \\ 1 & -2 & 1 & 7 \end{bmatrix}$$

and then with same Σ_1 and

$$\Sigma_2 = \begin{bmatrix} 10 & -1 & 0 & 1 \\ -1 & 6 & 0 & 0 \\ 0 & 0 & 8 & -1 \\ 1 & 0 & -1 & 6 \end{bmatrix}$$

In the first case the p-value should be high (H_0 is true), while in the second case it should be low (H_0 is false).
Here are two examples of the output:

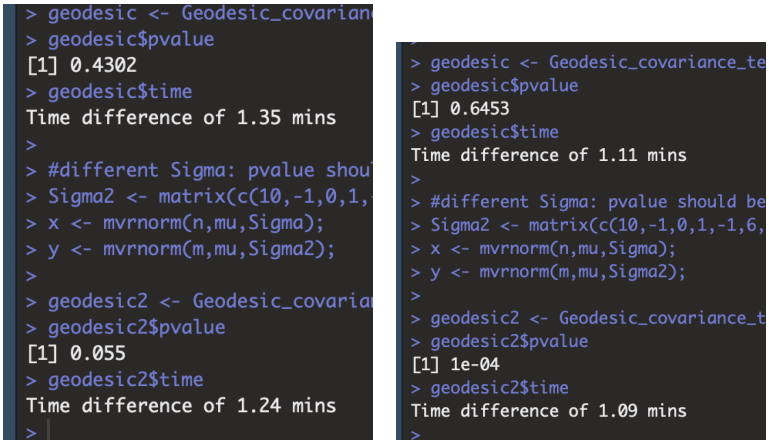

Fig. 3. The output of each test is followed by the computation time.

I ran other simulations of the test, with different values of n and m. Here are some comments on the results:

- The test seems to give significant results most of the times.
- If n and m are small, the mass of the PDF is distributed in a very wide range. The p-value is always high, hence the test is quite ineffective.
- It takes more than one minute to compute the distribution of d_L, but I think it can be improved writing a better code than mine.

4 Unsolved questions

- How stable is the test above introduced?
- Is it possible to prove that the distributions for $d_L(S_x, S_y)$ and $d_R(S_x, S_y)$ are independent from the parameter Σ?
- If so, is it possible to compute their PDF analytically?
- Are there any scale factors for the other distances, such that the new quantity is pivotal w.r.t. parameter Σ?
- Do the other two unmentioned distances depend on the parameter Σ?

5 References

– **Dryden et al.**: Dryden, Ian L.; Koloydenko, Alexey; Zhou, Diwei. Non-Euclidean statistics for covariance matrices, with applications to diffusion tensor imaging. Ann. Appl. Stat. 3 (2009), no. 3, 1102–1123. doi:10.1214/09-AOAS249. https://projecteuclid.org/euclid.aoas/1254773280
– **Anderson**: Anderson, T. W. (2003). An Introduction to Multivariate Statistical Analysis (3rd ed.). Hoboken, N. J.: Wiley Interscience. p. 259. ISBN 0-471-36091-0.
– **Github repository**: https://github.com/paolozapp/covariance-matrices

6 Appendix

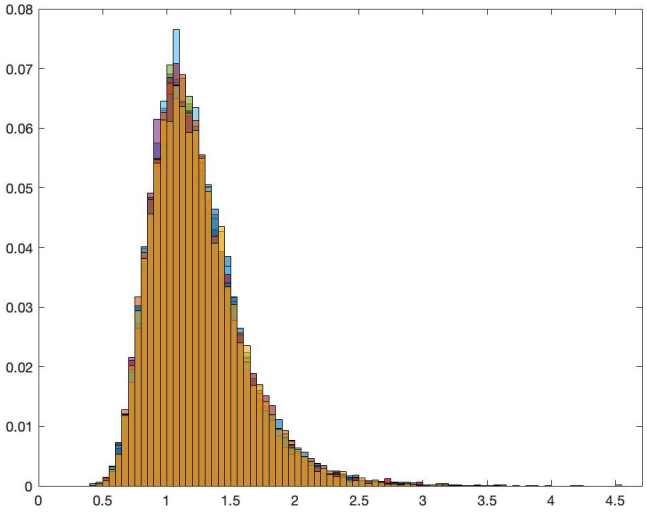

Fig. 4. $d_R(S_x, S_y)$

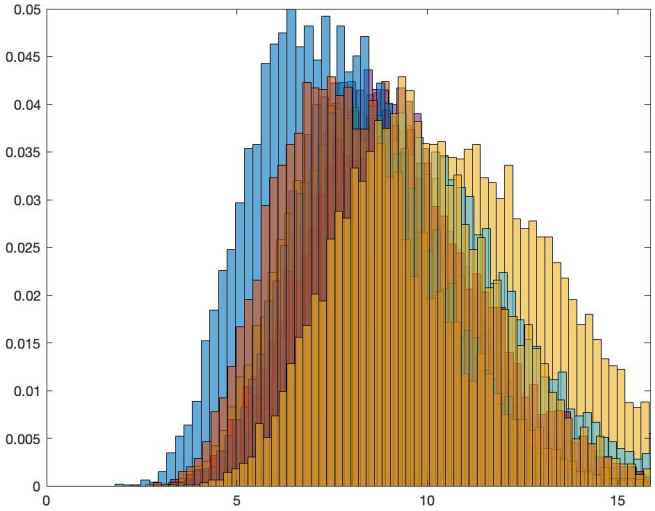

Fig. 5. $d_C(S_x, S_y)$

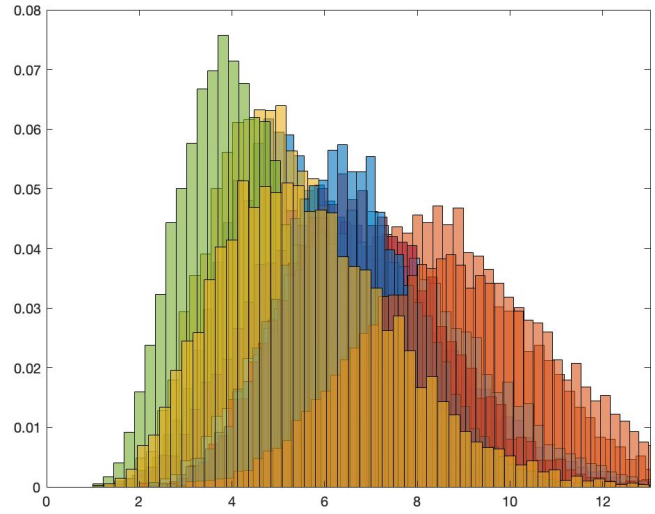

Fig. 6. $d_H(S_x, S_y)$

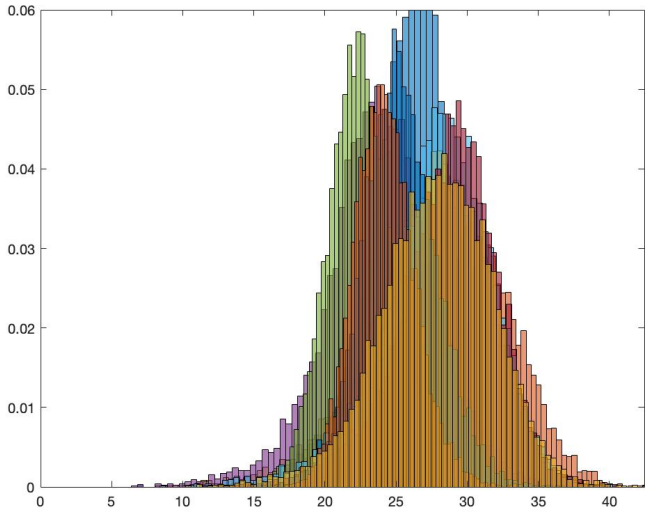

Fig. 7. $d_S(S_x, S_y)$