A Metric for Covariance Matrices*

Wolfgang Förstner, Boudewijn Moonen
Institut für Photogrammetrie, Universität Bonn
Nussallee 15, D-53115 Bonn, e-mail: wf|bm@ipb.uni-bonn.de

Diese Sätze führen dahin, die Theorie der krummen Flächen aus einem neuen Gesichtspunkt zu betrachten, wo sich der Untersuchung ein weites noch ganz unangebautes Feld öffnet . . . so begreift man, dass zweierlei wesentlich verschiedene Relationen zu unterscheiden sind, theils nemlich solche, die eine bestimmte Form der Fläche im Raume voraussetzen, theils solche, welche von den verschiedenen Formen . . . unabhängig sind. Die letzteren sind es, wovon hier die Rede ist . . . man sieht aber leicht, dass eben dahin die Betrachtung der auf der Fläche construirten Figuren, . . . , die Verbindung der Punkte durch kürzeste Linien*] *u. dgl. gehört. Alle solche Untersuchungen müssen davon ausgehen, dass die Natur der krummen Fläche an sich durch den Ausdruck eines unbestimmten Linearelements in der Form* $\sqrt{(Edp^2 + 2Fdpdq + Gdq^2)}$ *gegeben ist . . .*

Carl Friedrich Gauss

Abstract

The paper presents a metric for positive definite covariance matrices. It is a natural expression involving traces and joint eigenvalues of the matrices. It is shown to be the distance coming from a canonical invariant Riemannian metric on the space $Sym^+(n, \mathbb{R})$ of real symmetric positive definite matrices

In contrast to known measures, collected e. g. in Grafarend 1972, the metric is invariant under affine transformations and inversion. It can be used for evaluating covariance matrices or for optimization of measurement designs.

Keywords: Covariance matrices, metric, Lie groups, Riemannian manifolds, exponential mapping, symmetric spaces

1 Background

The optimization of geodetic networks is a classical problem that has gained large attention in the 70s.

1972 E. W. Grafarend put together the current knowledge of network design, datum transformations and artificial covariance matrices using covariance functions in his classical monograph [5]; see also [6]. One critical part was the development of a suitable measure for comparing two covariance matrices. Grafarend listed a dozen measures. Assuming a completely isotropic network, represented by a unit matrix as covariance matrix, the measures depended on the eigenvalues of the covariance matrix.

*Quo vadis geodesia ...?, Festschrift for Erik W. Grafarend on the occasion of his 60th birthday, Eds..: F. Krumm, V. S. Schwarze, TR Dept. of Geodesy and Geoinformatics, Stuttgart University, 1999

*]emphasized by the authors

1983, 11 years later, at the Aalborg workshop on 'Survey Control Networks' Schmidt [13] used these measures for finding optimal networks. The visualization of the error ellipses for a single point, leading to the same deviation from an ideal covariance structure revealed deficiencies of these measures, as e. g. the trace of the eigenvalues of the covariance matrix as quality measure would allow a totally flat error ellipse to be as good as a circular ellipse, more even, as good as the flat error ellipse rotated by 90^0.

Based on an information theoretic point of view, where the information of a Gaussian variable increases with $\ln \sigma^2$, the first author guessed the squared sum $d^2 = \sum_i \ln^2 \lambda_i$ of the logarithms of the eigenvalues to be a better measure, as deviations in both directions would be punished the same amount if measured in percent, i. e. relative to the given variances. He formulated the conjecture that the distance measure d would be a metric. Only in case d would be a metric, comparing two covariance matrices $\mathbf{A}$ and $\mathbf{B}$ with covariance matrix $\mathbf{C}$ would allow to state one of the two to be better than the other. Extensive simulations by K. Ballein [2] substantiated this conjecture as no case was found where the triangle inequality was violated.

1995, 12 years later, taking up the problem within image processing, the first author proved the validity of the conjecture for 2×2-matrices [3]. For this case the measure already had been proposed by V. V. Kavrajski [8] for evaluating the isotropy of map projections. However, the proof could not be generalized to higher dimensions. Using classical results from linear algebra and differential geometry the second author proved the distance d to be a metric for general positive definite symmetric matrices. An extended proof can be found in [11].

This paper states the problem and presents the two proofs for 2×2-matrices and for the general case. Giving two proofs for $n = 2$ may be justified by the two very different approaches to the problem.

2 Motivation

Comparing covariance matrices is a basic task in mensuration design. The idea, going back to Baarda 1973 [1] is to compare the variances of arbitrary functions $f = \mathbf{e}^T\mathbf{x}$ on one hand determined with a given covariance matrix $\mathbf{C}$ and on the other hand determined with a reference or criterion matrix $\mathbf{H}$.

One requirement would be the variance $\sigma_f^{2(C)}$ of f when calculated with $\mathbf{C}$ to be always smaller than the variance $\sigma_f^{2(H)}$ of f when calculated with $\mathbf{H}$. This means:

$$\mathbf{e}^T\mathbf{C}\mathbf{e} \le \mathbf{e}^T\mathbf{H}\mathbf{e} \quad \text{for all } \mathbf{e} \neq \mathbf{0}$$

or the Raleigh ratio

$$0 \le \lambda(\mathbf{e}) = \frac{\mathbf{e}^T\mathbf{C}\mathbf{e}}{\mathbf{e}^T\mathbf{H}\mathbf{e}} \le 1 \quad \text{for all } \mathbf{e} \neq \mathbf{0}.$$

The maximum λ from $1/2\partial\lambda(\mathbf{e})/\partial\mathbf{e} = \mathbf{0} \leftrightarrow \lambda\mathbf{H}\mathbf{e} - \mathbf{C}\mathbf{e} = (\lambda\mathbf{H} - \mathbf{C})\mathbf{e} = \mathbf{0}$ results in the maximum eigenvalue $\lambda_{\max}(\mathbf{C}\mathbf{H}^{-1})$ from the generalized eigenvalue problem

$$|\lambda\mathbf{H} - \mathbf{C}| = 0\,, \tag{1}$$

Observe that $\lambda\mathbf{e}^T\mathbf{H}\mathbf{e} - \mathbf{e}^T\mathbf{C}\mathbf{e} = \mathbf{e}^T(\lambda\mathbf{H} - \mathbf{C})\mathbf{e} = 0$ for $\mathbf{e} \neq \mathbf{0}$ only is fulfilled if (1) holds. The eigenvalues of (1) are non-negative if the two matrices are positive semidefinite.

This suggests the eigenvalues of $\mathbf{C}\mathbf{H}^{-1}$ to capture the difference in form of $\mathbf{C}$ and $\mathbf{H}$ completely.

The requirement $\lambda_{\max} \le 1$ can be visualized by stating that the (error) ellipsoid $\mathbf{x}^T\mathbf{C}^{-1}\mathbf{x} = c$ lies completely within the (error) ellipsoid $\mathbf{x}^T\mathbf{H}^{-1}\mathbf{x} = c$.

The statistical interpretation of the ellipses results from the assumption, motivated by the principle of maximum entropy, that the stochastical variables are normally distributed, thus having density:

$$p(\mathbf{x}) = \frac{1}{\sqrt{(2\pi)^n \det \mathbf{\Sigma}}} \mathrm{e}^{-\frac{1}{2}\mathbf{x}^T \mathbf{\Sigma}^{-1} \mathbf{x}}$$

with covariance matrix $\mathbf{\Sigma}$. Isolines of constant density are ellipses with semiaxes proportional to the square roots of the eigenvalues of $\mathbf{\Sigma}$. The ratio

$$\lambda_{\max} = \max_{\mathbf{e}} \frac{\sigma_f^{2(C)}}{\sigma_f^{2(H)}}$$

thus gives the worst case for the ratio of the variances when calculated with the covariances $\mathbf{C}$ and $\mathbf{H}$ respectively.

Instead of requiring the worst precision to be better than a specification one also could require the covariance matrix $\mathbf{C}$ to be closest to $\mathbf{H}$ in some sense. Let us for a moment assume $\mathbf{H} = \mathbf{I}$. Simple examples for measuring the difference in form of $\mathbf{C}$ compared to $\mathbf{I}$ are the trace

$$\operatorname{tr} \mathbf{C} = \sum_{i=1}^{n} \lambda_i(\mathbf{C}) \tag{2}$$

or the determinant

$$\det \mathbf{C} = \prod_{i=1}^{n} \lambda_i(\mathbf{C})\,. \tag{3}$$

These classical measures are invariant with respect to rotations (2) or affine transformations (3) of the coordinate system. Visualizing covariance matrices of equal trace or determinant can use the eigenvalues. Restricting to $n = 2$ in a 2D-coordinate system (λ_1, λ_2) covariance matrices of equal trace $\operatorname{tr}(\mathbf{C}) = c_{\mathrm{tr}}$ are characterized by the straight line $\lambda_1 = c_{\mathrm{tr}} - \lambda_2$ or $\sigma_1^2 = c_{\mathrm{tr}} - \sigma_2^2$. Covariance matrices of equal determinant $\det(\mathbf{C}) = c_{\det}$ are determined by the hyperbola $\lambda_1 = c_{\det}/\lambda_2$ or $\sigma_1^2 = c_{\det}/\sigma_2^2$. Obviously in both cases covariance matrices with very flat form of the corresponding error ellipse $\mathbf{e}^T \mathbf{C} \mathbf{e} = c$ are allowed. E. g., if one requires $c_{\mathrm{tr}} = 2$ then the pair $(0.02, 1.98)$ with a ratio of semiaxes $\sqrt{1.98/0.02} = 7$ is evaluated as being similar to the unit circle. The determinant measure is even more unfavourable. When requiring $c_{\det} = 1$ even a pair $(0.02, 50.0)$ with ratio of semiaxes 50 is called similar to the unit circle.

However, it would be desirable that the similarity between two covariance matrices reflects the deviation in variance in both directions according to the *ratio* of the variances. Thus deviations in variance by a factor f should be evaluated equally as a deviation by a factor $1/f$, of course a factor $f = 1$ indicating no difference. Thus other measures capturing the anisotropy, such as $(1 - \lambda_1)^2 + (1 - \lambda_2)^2$, not being invariant to inversion, cannot be used.

The conditions can be fulfilled by using the sum of the squared logarithms of the eigenvalues. Thus we propose the distance measure

$$\boxed{d(\mathbf{A}, \mathbf{B}) = \sqrt{\sum_{i=1}^{n} \ln^2 \lambda_i(\mathbf{A}, \mathbf{B})}} \tag{4}$$

between symmetric positive definite matrices $\mathbf{A}$ and $\mathbf{B}$, with the eigenvalues $\lambda_i(\mathbf{A}, \mathbf{B})$ from $|\lambda \mathbf{A} - \mathbf{B}| = 0$. The logarithm guarantees, that deviations are measured as factors, whereas the squaring guarantees factors f and $1/f$ being evaluated equally. Summing squares is done in close resemblance with the Euclidean metric.

This note wants to discuss the properties of d:

- d is invariant with respect to affine transformations of the coordinate system.
- d is invariant with respect to an inversion of the matrices.
- It is claimed that d is a *metric*. Thus

  (i) positivity: $d(\mathbf{A},\mathbf{B}) \geq 0$, and $d(\mathbf{A},\mathbf{B}) = 0$ only if $\mathbf{A} = \mathbf{B}$.

  (ii) symmetry: $d(\mathbf{A},\mathbf{B}) = d(\mathbf{B},\mathbf{A})$,

  (iii) triangle inequality: $d(\mathbf{A},\mathbf{B}) + d(\mathbf{A},\mathbf{C}) \geq d(\mathbf{B},\mathbf{C})$.

  The proof for $n = 2$ is given in the next Section. The proof for general n is sketched in the subsequent Sections 4 – 6.

3 Invariance Properties

3.1 Affine Transformations

Assume the $n \times n$ matrix $\mathbf{X}$ to be regular. Then the distance $d(\overline{\mathbf{A}},\overline{\mathbf{B}})$ of the transformed matrices

$$\overline{\mathbf{A}} = \mathbf{X}\mathbf{A}\mathbf{X}^T \quad \overline{\mathbf{B}} = \mathbf{X}\mathbf{B}\mathbf{X}^T$$

is invariant w. r. t. $\mathbf{X}$.

Proof: We immediately obtain:

$$\begin{aligned}
\lambda(\overline{\mathbf{A}},\overline{\mathbf{B}}) = \lambda(\mathbf{X}\mathbf{A}\mathbf{X}^T,\mathbf{X}\mathbf{B}\mathbf{X}^T) &= \lambda(\mathbf{X}\mathbf{A}\mathbf{X}^T(\mathbf{X}\mathbf{B}\mathbf{X}^T)^{-1}) \\
= \lambda(\mathbf{X}\mathbf{A}\mathbf{X}^T(\mathbf{X}^T)^{-1}\mathbf{B}^{-1}\mathbf{X}^{-1}) &= \lambda(\mathbf{X}\mathbf{A}\mathbf{B}^{-1}\mathbf{X}^{-1}) \\
= \lambda(\mathbf{A}\mathbf{B}^{-1}) &= \lambda(\mathbf{A},\mathbf{B})\,.
\end{aligned}$$

Comment: $\overline{\mathbf{A}}$ and $\overline{\mathbf{B}}$ can be interpreted as covariance matrices of $\mathbf{y} = \mathbf{X}\mathbf{x}$ in case $\mathbf{A}$ and $\mathbf{B}$ are the covariance matrices of $\mathbf{x}$. Changing coordinate system does not change the evaluation of covariance matrices. Obviously, this invariance only relies on the properties of the eigenvalues, and actually was the basis for Baarda's evaluation scheme using so-called S-transformations.

3.2 Inversion

The distance is invariant under inversion of the matrices.

Proof: We obtain

$$\begin{aligned}
d^2(\mathbf{A}^{-1},\mathbf{B}^{-1}) = d^2(\mathbf{A}^{-1}\mathbf{B}) &= \sum_{i=1}^{n} \left(\ln \lambda_i(\mathbf{A}^{-1}\mathbf{B})\right)^2 \\
= \sum_{i=1}^{n} \left(\ln[\lambda_i^{-1}(\mathbf{A}\mathbf{B}^{-1})]\right)^2 &= \sum_{i=1}^{n} \left(-\ln \lambda_i(\mathbf{A}\mathbf{B}^{-1})\right)^2 \\
= \sum_{i=1}^{n} \left(\ln \lambda_i(\mathbf{A}\mathbf{B}^{-1})\right)^2 &= d^2(\mathbf{A}\mathbf{B}^{-1}) \\
= d^2(\mathbf{A},\mathbf{B})\,.
\end{aligned}$$

Comment: $\mathbf{A}^{-1}$ and $\mathbf{B}^{-1}$ can be interpreted as weight matrices of $\mathbf{x}$ if one chooses $\sigma_o^2 = 1$. Here essential use is made of the property $\lambda(\mathbf{A}) = \lambda^{-1}(\mathbf{A}^{-1})$. The proof shows, that also individual *inversions* of eigenvalues do not change the value of distance measure, as required.

3.3 d is a Distance Measure

We show that d is a distance measure, thus the first two criteria for a metric hold in general.

ad 1 $d \geq 0$ is trivial from the definition of d, keeping in mind, that the eigenvalues all are positive. Proof of $d = 0 \leftrightarrow \mathbf{A} = \mathbf{B}$:

- $\leftarrow$: If $\mathbf{A} = \mathbf{B}$ then $d = 0$.
  PROOF: From $\lambda(\mathbf{A}\mathbf{B}^{-1}) = \lambda(\mathbf{I})$ follows $\lambda_i = 1$ for all i, thus $d = 0$.
- $\rightarrow$: If $d = 0$ then $\mathbf{A} = \mathbf{B}$.
  PROOF: From $d = 0$ follows $\lambda_i(\mathbf{A}\mathbf{B}^{-1}) = 1$ for all i, thus $\mathbf{A}\mathbf{B}^{-1} = \mathbf{I}$ from which $\mathbf{A} = \mathbf{B}$ follows.

ad 2 As $(\mathbf{A}\mathbf{B}^{-1})^{-1} = \mathbf{B}\mathbf{A}^{-1}$ symmetry follows from the inversion invariance.

3.4 Triangle Inequality

For d providing a metric on the symmetric positive definite matrices the triangle inequality must hold.

Assume three $n \times n$ matrices with the following structure:

- The first matrix is the unit matrix:
$$\mathbf{A} = \mathbf{I}\,.$$
- The second matrix is a diagonal matrix with entries e^{b_i} thus
$$\mathbf{B} = Diag(\mathrm{e}^{b_i})\,.$$
- The third matrix is a general matrix with eigenvalues e^{c_i} and modal matrix $\mathbf{R}$, thus
$$\mathbf{C} = \mathbf{R}Diag(\mathrm{e}^{c_i})\mathbf{R}^T\,.$$

This setup can be chosen without loss of generality, as $\mathbf{A}$ and $\mathbf{B}$ can be orthogonalized simultaneously [4].

The triangle inequality can be written in the following form and reveals three terms

$$s \doteq d(\mathbf{A}, \mathbf{B}) + d(\mathbf{A}, \mathbf{C}) - d(\mathbf{B}, \mathbf{C}) = d_c + d_b - d_a \geq 0\,. \tag{5}$$

The idea of the proof is the following:

(i) We first use the fact that d_b and d_c are independent on the rotation $\mathbf{R}$.
$$s(\mathbf{R}) = d_c + d_b - d_a(\mathbf{R})\,.$$

(ii) In case $\mathbf{R} = \mathbf{I}$ then the correctness of (5) results from the triangle inequality in $\mathbb{R}^n$. This even holds for any permutation $P(i)$ of the indices i of the eigenvalues λ_i of $\mathbf{B}\mathbf{C}^{-1}$. There exists a permutation P_{max} for which d_a is maximum, thus $s(\mathbf{R})$ is a minimum.

(iii) We then want to show, and this is the crucial part, that any rotation $\mathbf{R} \neq \mathbf{I}$ leads to a decrease of $d_a(\mathbf{R})$, thus to an increase of $s(\mathbf{R})$ keeping the triangle inequality to hold.

3.4.1 Distances d_c and d_b

The distances d_c and d_b are given by

$$d_c^2 = \sum_{i=1}^{n} b_i^2 \quad , \quad d_b^2 = \sum_{i=1}^{n} c_i^2 \,.$$

The special definition of the matrices $\mathbf{B}$ and $\mathbf{C}$ now shows to be useful. The last expression results from the fact that the eigenvalues of $\mathbf{C}\mathbf{A}^{-1} = \mathbf{C}$ are independent on rotations $\mathbf{R}$.

3.4.2 Triangle Inequality for No Rotation

In case of no rotations the eigenvalues of $\mathbf{B}\mathbf{C}^{-1}$ are $\mathrm{e}^{b_i}/\mathrm{e}^{c_i}$. Therefore the distance d_a yields

$$d_a^2 = \sum_{i=1}^{n} \left(\ln \frac{\mathrm{e}^{b_i}}{\mathrm{e}^{c_i}} \right)^2 = \sum_{i=1}^{n} (b_i - c_i)^2 \,.$$

With the vectors $\mathbf{b} = (b_i)$ and $\mathbf{c} = (c_i)$ the triangle inequality in $\mathbb{R}^n$ yields

$$|\mathbf{c}| + |\mathbf{b}| - |\mathbf{b} - \mathbf{c}| \geq 0$$

or

$$s = \sqrt{\sum_{i=1}^{n} c_i^2} + \sqrt{\sum_{i=1}^{n} b_i^2} - \sqrt{\sum_{i=1}^{n} (b_i - c_i)^2} \geq 0 \,.$$

holds.

For any permutation $P(i)$ we also get

$$s = \sqrt{\sum_{i=1}^{n} c_{P(i)}^2} + \sqrt{\sum_{i=1}^{n} b_i^2} - \sqrt{\sum_{i=1}^{n} (b_i - c_{P(i)})^2} \geq 0 \,. \tag{6}$$

which guarantees that there is a permutation $P_{max}(i)$ for which s in (6) is minimum.

3.4.3 d is Metric for 2×2-Matrices

We now want to show that the triangle inequality holds for 2×2 matrices. Thus we only need to show that $s(\mathbf{R}(\phi))$ is monotonous with ϕ in $[0..\pi/2]$, or equivalently that $d_a(\mathbf{R})$ is monotonous.

We assume (observe the change of notation in the entries b_i and c_i of the matrices)

$$\mathbf{B} = \begin{pmatrix} b_1 & 0 \\ 0 & b_2 \end{pmatrix} \quad , \quad b_1 > 0 \,,\; b_2 > 0 \,. \tag{7}$$

With

$$x = \sin\phi \tag{8}$$

the rotation $\mathbf{R}(x) = \mathbf{R}(\phi)$ is represented as

$$\mathbf{R}(x) = \begin{pmatrix} \sqrt{1-x^2} & x \\ -x & \sqrt{1-x^2} \end{pmatrix} ,$$

thus only values $x \in [0, 1]$ need to be investigated.

With the diagonal matrix $Diag(c_1, c_2)$, containing the positive eigenvalues

$$c_1 > 0, \quad c_2 > 0\,, \tag{9}$$

this leads to the general matrix

$$\mathbf{C} = \mathbf{R}\begin{pmatrix} c_1 & 0 \\ 0 & c_2 \end{pmatrix}\mathbf{R}^T = \begin{pmatrix} c_1 x^2 + c_2(1-x^2) & -x\sqrt{1-x^2}(c_2 - c_1) \\ -x\sqrt{1-x^2}(c_2 - c_1) & c_1(1-x^2) + c_2 x^2 \end{pmatrix}.$$

The eigenvalues of $\mathbf{CB}^{-1}$ are (from Maple)

$$\lambda_1 = \frac{u(x) + \sqrt{v(x)}}{2b_1 b_2} \geq 0 \qquad , \qquad \lambda_2 = \frac{u(x) - \sqrt{v(x)}}{2b_1 b_2} \geq 0$$

with the discriminant

$$v(x) = u(x)^2 - w \geq 0$$

and

$$u(x) = (b_1\, c_1 + c_2\, b_2)(1-x^2) + (b_1\, c_2 + b_2\, c_1)x^2 \geq 0\,, \tag{10}$$

$$w = 4 b_1 b_2 c_1 c_2 \geq 0\,, \tag{11}$$

the last inequality holding due to (7)(9). The distance

$$d_a(x) = \sqrt{\ln^2 \lambda_1(x) + \ln^2 \lambda_2(x)}\,,$$

which is dependent on x, has first derivative

$$\begin{aligned}\frac{\partial d_a(x)}{\partial x} &= 2\frac{x(b_2 - b_1)(c_2 - c_1)}{v(x)\, d_a(x)}\left(\ln\frac{u(x) - \sqrt{v(x)}}{2b_1 b_2} - \ln\frac{u(x) + \sqrt{v(x)}}{2b_1 b_2}\right) \\ &= 2\frac{x(b_2 - b_1)(c_2 - c_1)\ln\dfrac{u(x) - \sqrt{v(x)}}{u(x) + \sqrt{v(x)}}}{v(x)\, d_a(x)}\,. \end{aligned} \tag{12}$$

For fixed b_1, b_2, c_1 and c_2 this expression does not change sign in $x \in [0,1]$. This is because the discriminant $v(x) = u^2(x) - w(x)$ (cf. (3.4.3)) is always positive, due to

$$\begin{aligned} v(0) &= (b_1 c_1 - b_2 c_2)^2 \geq 0 \\ v(1) &= (b_2 c_1 - b_1 c_2)^2 \geq 0 \\ \frac{\partial v(x)}{\partial x} &= -4x(b_2 - b_1)(c_2 - c_1)\, u(x) \end{aligned}$$

with $u(x) \geq 0$ (cf. (10)) thus $v(x)$ being monotonous. Furthermore, $v(x)$ is always smaller than u^2 (cf. (11), (3.4.3)), thus the logarithmic expression always negative. As the triangle equation is fulfilled at the extremes of the interval $[0,1]$ it is fulfilled for all x, thus for all rotations.

Comment: When substituting $x = \sin\phi$ (cf. (8)) the first derivative (12) is of the form $\partial d_a(\phi)/\partial\phi = \sin\phi\, f(\phi)$ with a symmetric function $f(\phi) = f(-\phi)$. Thus the derivative is skew symmetric w. r. t. $(0,0)$, indication d_a to be symmetric $d_a(\phi) = d_a(-\phi)$, which is to be expected.

3.4.4 d is Claimed to be a Metric for $n \times n$-Matrices

The proof of the metric properties of d for 2×2 matrices suggests that in the general case of $n \times n$ matrices any rotation away from the worst permutation of the indices (cf. (6)) results in an increase of the value s. The proof for the case $n = 2$ can be used to show, that, starting with the worst permutation of the indices, any single rotation around one of the axes leads to a monotonous change of s. Therefore, for proving the case of general n, there would have to be shown that any combination of *two* rotations away from the worst permutation leads to a monotonous change of s allowing to reach any permutation by a rotation while increasing $s(\mathbf{R})$. Completing this line of proof has not been achieved so far.

4 Restating the Problem

In der Kürze liegt die Würze.

DEUTSCHER VOLKSMUND

Let

$$M(n,\mathbb{R}) := \{\, \mathbf{A} = (a_{ij}) \ \mid 1 \leq i,j \leq n \ , \ a_{ij} \in \mathbb{R} \,\}$$

be the space of real $n \times n$-matrices, and let

$$S^+ := Sym^+(n,\mathbb{R}) := \left\{ \mathbf{A} \in M(n,\mathbb{R}) \ \middle|\ \mathbf{A} = \mathbf{A}^T \ , \ \mathbf{A} > 0 \right\}$$

be the subspace of real, symmetric, positive definite matrices. Recall that any symmetric matrix $\mathbf{A}$ can be substituted into a function $f : \mathbb{R} \longrightarrow \mathbb{R}$ which gives a symmetric matrix $f(\mathbf{A})$ commuting with all matrices commuting with $\mathbf{A}$. In particular, a symmetric matrix $\mathbf{A}$ has an exponential $\exp(\mathbf{A})$, and a symmetric positive definite matrix A has a logarithm $\ln(\mathbf{A})$, and these assignments are inverse to each other. A symmetric positive definite matrix $\mathbf{A}$ also has a unique square root $\sqrt{\mathbf{A}}$ which is of the same type. Define, for $\mathbf{A},\mathbf{B} \in Sym^+(n,\mathbb{R})$, their distance $d(\mathbf{A},\mathbf{B}) \geq 0$ by

$$\boxed{d^2(\mathbf{A},\mathbf{B}) := \operatorname{tr}\left(\ln^2\left(\sqrt{\mathbf{A}^{-1}}\mathbf{B}\sqrt{\mathbf{A}^{-1}}\right)\right)}\,, \tag{13}$$

where tr denotes the trace. In particular, this shows that

$$d(\mathbf{A},\mathbf{B}) \geq 0 \qquad , \qquad d(\mathbf{A},\mathbf{B}) = 0 \iff \mathbf{A} = \mathbf{B}\,.$$

In more down-to-earth terms:

$$d(\mathbf{A},\mathbf{B}) = \sqrt{\sum_{i=1}^{n} \ln^2 \lambda_i(\mathbf{A},\mathbf{B})}, \tag{14}$$

where $\lambda_1(\mathbf{A},\mathbf{B}),\ldots,\lambda_n(\mathbf{A},\mathbf{B})$ are the *joint eigenvalues* of $\mathbf{A}$ and $\mathbf{B}$, i.e. the roots of the equation

$$\det(\lambda\mathbf{A} - \mathbf{B}) = 0.$$

This is the proposal of [3], i.e. of equation (4) above. (To see why these two definitions coincide, note that

$$\lambda\mathbf{A} - \mathbf{B} = \sqrt{\mathbf{A}}\,(\lambda\mathbf{E} - \sqrt{\mathbf{A}^{-1}}\mathbf{B}\sqrt{\mathbf{A}^{-1}})\,\sqrt{\mathbf{A}}\,,$$

so that the joint eigenvalues $\lambda_i(\mathbf{A}, \mathbf{B})$ are just the eigenvalues of the real symmetric positive definite matrix $\sqrt{\mathbf{A}^{-1}}\mathbf{B}\sqrt{\mathbf{A}^{-1}}$; in particular, they are positive real numbers and so the definition (14) makes sense.) The equation (14) shows that d is invariant under congruence transformations with $\mathbf{X} \in GL(n, \mathbb{R})$, where $GL(n, \mathbb{R})$ is the group of regular linear transformations of $\mathbb{R}^n$:

$$\forall\, \mathbf{X} \in GL(n, \mathbb{R}) : \quad d(\mathbf{A}, \mathbf{B}) = d(\mathbf{X}\mathbf{A}\mathbf{X}^T, \mathbf{X}\mathbf{B}\mathbf{X}^T) \tag{15}$$

(since $\det(\lambda\mathbf{A} - \mathbf{B})$ and $\det\big(\mathbf{X}(\lambda\mathbf{A} - \mathbf{B})\mathbf{X}^T\big)$ have the same roots); this is not easily seen from definition (13). It also shows that

$$d(\mathbf{A}, \mathbf{B}) = d(\mathbf{B}, \mathbf{A}) \ , \ d(\mathbf{A}, \mathbf{B}) = d(\mathbf{A}^{-1}, \mathbf{B}^{-1}).$$

5 The results

One then has

Theorem 1. *The map d defines a distance on the space $Sym^+(n, \mathbb{R})$, i.e there holds*

(i) Positivity: $d(\mathbf{A}, \mathbf{B}) \geq 0$, and $d(\mathbf{A}, \mathbf{B}) = 0 \iff \mathbf{A} = \mathbf{B}$

(ii) Symmetry: $d(\mathbf{A}, \mathbf{B}) = d(\mathbf{B}, \mathbf{A})$

(iii) Triangle inequality: $d(\mathbf{A}, \mathbf{C}) \leq d(\mathbf{A}, \mathbf{B}) + d(\mathbf{B}, \mathbf{C})$

for all $\mathbf{A}, \mathbf{B}, \mathbf{C} \in Sym^+(n, \mathbb{R})$. Moreover, d has the following invariances:

(iv) It is invariant under congruence transformations, i.e.

$$d(\mathbf{A}, \mathbf{B}) = d(\mathbf{X}\mathbf{A}\mathbf{X}^T, \mathbf{X}\mathbf{B}\mathbf{X}^T)$$

for all $\mathbf{A}, \mathbf{B} \in Sym^+(n, \mathbb{R})$, $\mathbf{X} \in GL(n, \mathbb{R})$

(v) It is invariant under inversion, i.e.

$$d(\mathbf{A}, \mathbf{B}) = d(\mathbf{A}^{-1}, \mathbf{B}^{-1})$$

for all $\mathbf{A}, \mathbf{B} \in Sym^+(n, \mathbb{R})$

The same conclusions hold for the space $SSym(n, \mathbb{R})$ of real symmetric positive definite matrices of determinant one, when one replaces the general linear group $GL(n, \mathbb{R})$ with the special linear group $SL(n, \mathbb{R})$, the $n \times n$–matrices of determinant one, and the space of real symmetric matrices $Sym(n, \mathbb{R})$ with the space $Sym_0(n, \mathbb{R})$ of real symmetric traceless matrices.

Remark 1. We use here the terminology “distance” in contrast to the standard terminology “metric” in order to avoid confusion with the notion of “Riemannian metric”, which is going to play a rôle soon.

The case $n = 2$ is already interesting; see Remark 2 below.

All the properties except property (iv), the triangle inequality, are more or less obvious from the definition (see above), but the triangle inequality is not. In fact, the theorem will be the consequence of a more general theorem as follows.

The most important geometric way distances arise is as associated distances to Riemannian metrics on manifolds; the Riemannian metric, as an infinitesimal description of length is used to define the length of paths by integration, and the distance between two points then arises as the greatest lower bound on the length of paths joining the two points. More precisely, if M

is a differentiable manifold (in what follows, "differentiable" will always mean "infinitely many times differentiable", i.e. of class $\mathcal{C}^\infty$), a *Riemannian metric* is the assignment to any $p \in M$ of a Euclidean scalar product $\langle - \mid - \rangle_p$ in the tangent space T_pM depending differentiably on p. Technically, it is a differentiable positive definite section of the second symmetric power S^2T^*M of the cotangent bundle, or a positive definite symmetric 2-tensor. In classical terms, it is given in local coordinates (U, x) as the "square of the line element" or "first fundamental form"

$$\mathrm{d}s^2 = g_{ij}(x)\mathrm{d}x^i\mathrm{d}x^j \tag{16}$$

(Einstein summation convention: repeated lower and upper indices are summed over). Here the g_{ij} are differentiable functions (the *metric coefficients)* subjected to the transformation rule

$$g_{ij}(x) = g_{kl}(y(x))\frac{\partial y^k}{\partial x^i}\frac{\partial y^l}{\partial x^j}\,.$$

A differentiable manifold together with a Riemannian metric is called a *Riemannian manifold.*

Given a piecewise differentiable path $c : [a, b] \longrightarrow M$ in M, its *length* $L\,[c]$ is defined to be

$$L\,[c] := \int_a^b \|\dot{c}(t)\|_{c(t)}\,\mathrm{d}t\,,$$

where for $X \in T_pM$ we have $\|X\|_p := \sqrt{\langle X \mid X\rangle_p}$, the Euclidean norm associated to the scalar product in T_pM given by the Riemannian metric. In local coordinates

$$L\,[c] = \int_a^b \sqrt{g_{ij}(c(t))\,\dot{c}^i(t)\dot{c}^j(t)}\,\mathrm{d}t\,.$$

Given p, $q \in M$, the distance $d(p, q)$ associated to a given Riemannian metric then is defined to be

$$d(p, q) := \inf_c\, L\,[c]\;, \tag{17}$$

the infimum running over all piecewise differentiable paths c joining p to q. This defines indeed a distance:

Proposition. *The distance defined by* (17) *on a connected Riemannian manifold is a metric in the sense of metric spaces, i.e. defines a map* $d : M \times M \longrightarrow \mathbb{R}$ *satisfying*

(i) $d(p, q) \geq 0$*,* $d(p, q) = 0 \iff p = q$

(ii) $d(p, q) = d(q, p)$

(iii) $d(p, r) \leq d(p, q) + d(q, r)$ *.*

An indication of proof will be given in the next section.

So the central issue here is the fact that a Riemannian metric is the differential substrate of a distance and, in turn, defines a distance by integration. This is the most important way of constructing distances, which is the fundamental discovery of Gauss and Riemann. In our case, this paradigm is realized in the following way.

The space $Sym^+(n, \mathbb{R})$ is a differentiable manifold of dimension $n(n+1)/2$, more specifically, it is an open cone in the vector space

$$Sym(n, \mathbb{R}) := \left\{\, \mathbf{A} \in M(n, \mathbb{R}) \;\middle|\; \mathbf{A} = \mathbf{A}^T \,\right\}$$

of all real symmetric $n \times n$-matrices. Thus the tangent space $T_{\mathbf{A}}Sym^+(n,\mathbb{R})$ to $Sym^+(n,\mathbb{R})$ at a point $\mathbf{A} \in Sym^+(n,\mathbb{R})$ is just given as

$$T_{\mathbf{A}}Sym^+(n,\mathbb{R}) = Sym(n,\mathbb{R}).$$

The tangent space $T_{\mathbf{A}}SSym^+(n,\mathbb{R})$ to $SSym(n,\mathbb{R})$ at a point $\mathbf{A} \in SSym(n,\mathbb{R})$ is just given as

$$T_{\mathbf{A}}SSym(n,\mathbb{R}) = Sym_0(n,\mathbb{R}) := \{\, \mathbf{X} \in Sym(n,\mathbb{R}) \mid \operatorname{tr}(\mathbf{X}) = 0 \,\},$$

the space of traceless symmetric matrices.

Now note that there is a natural action of $GL(n,\mathbb{R})$ on S^+, namely, as already referred to above, by congruence transformations: $\mathbf{X} \in GL(n,\mathbb{R})$ acts via $\mathbf{A} \mapsto \mathbf{X}\mathbf{A}\mathbf{X}^T$. If one regards $\mathbf{A}$ as the matrix corresponding to a bilinear form with respect to a given basis, this action represents a change of basis. This action is transitive, and the isotropy subgroup at $\mathbf{E} \in S^+$ is just the orthogonal group $O(n)$:

$$\{\, \mathbf{X} \in GL(n,\mathbb{R}) \;\big|\; \mathbf{X}\mathbf{E}\mathbf{X}^T = \mathbf{E} \,\} = \{\, \mathbf{X} \in GL(n,\mathbb{R}) \;\big|\; \mathbf{X}\mathbf{X}^T = \mathbf{E} \,\} = O(n),$$

and so S^+ can be identified with the homogeneous space $GL(n,\mathbb{R})/O(n)$ upon which $GL(n,\mathbb{R})$ acts by left translations (its geometric significance being that its points parametrize the possible scalar products in $\mathbb{R}^n$) [†] .

In general, a *homogeneous space* is a differentiable manifold with a *transitive* action of a Lie group G, whence it has a representation as a quotient $M = G/H$ with H a closed subgroup of G. The most natural Riemannian metrics in this case are then those for which the group G acts by isometries, or, in other words, which are invariant under the action of G; e.g. the classical geometries – the Euclidean, the elliptic, and the hyperbolic geometry – arise in this manner. Looking out for these metrics, Theorem 1 will be a consequence of the following theorem :

Theorem 2. *(i) The Riemannian metrics g on $Sym^+(n,\mathbb{R})$ invariant under congruence transformations with matrices $\mathbf{X} \in GL(n,\mathbb{R})$ are in one-to-one-correspondence with positive definite quadratic forms Q on $T_{\mathbf{E}}Sym^+(n,\mathbb{R}) = Sym(n,\mathbb{R})$ invariant under conjugation with orthogonal matrices, the correspondence being given by*

$$g_{\mathbf{A}}(\mathbf{X},\mathbf{Y}) \;=\; \mathbf{B}(\sqrt{\mathbf{A}^{-1}}\mathbf{X}\sqrt{\mathbf{A}^{-1}}, \sqrt{\mathbf{A}^{-1}}\mathbf{Y}\sqrt{\mathbf{A}^{-1}}),$$

where $\mathbf{A} \in Sym^+(n,\mathbb{R})$, $\mathbf{X},\mathbf{Y} \in Sym(n,\mathbb{R}) = T_{\mathbf{A}}Sym^+(n,\mathbb{R})$, and $\mathbf{B}$ is the symmetric positive bilinear form corresponding to Q

(ii) The corresponding distance d_Q is invariant under congruence transformations and inversion, i.e. satisfies

$$d_Q(\mathbf{A},\mathbf{B}) = d_Q(\mathbf{X}\mathbf{A}\mathbf{X}^T, \mathbf{X}\mathbf{B}\mathbf{X}^T)$$

and

$$d_Q(\mathbf{A},\mathbf{B}) = d_Q(\mathbf{A}^{-1},\mathbf{B}^{-1})$$

for all $\mathbf{A},\mathbf{B} \in Sym^+(n,\mathbb{R})$, $\mathbf{X} \in GL(n,\mathbb{R})$, and is given by

$$d_Q^2(\mathbf{A},\mathbf{B}) = \frac{1}{4}Q\big(\ln\big(\sqrt{\mathbf{A}^{-1}}\mathbf{B}\sqrt{\mathbf{A}^{-1}}\big)\big)\,. \tag{18}$$

[†] A concrete map $p : G \longrightarrow S^+$ achieving this identification is given by $p(\mathbf{A}) := \mathbf{A}\mathbf{A}^T$; it is surjective and satisfies $p(\mathbf{X}\mathbf{A}) = \mathbf{X}p(\mathbf{A})\mathbf{X}^T$. The fact that p is an identification map is then equivalent to the *polar decomposition*: any regular matrix can be uniquely written as the product of a positive definite symmetric matrix and an orthogonal matrix. This generalizes the representation $z = re^{i\varphi}$ of a nonzero complex number z.

(iii) In particular, the distance in Theorem 1 is given by the invariant Riemannian metric corresponding to the canonical non–degenerate bilinear form ‡]

$$\forall\, \mathbf{X}, \mathbf{Y} \in M(n, \mathbb{R}) : \quad B(\mathbf{X}, \mathbf{Y}) \;:=\; 4\, \mathrm{tr}\,(\mathbf{XY})$$

on $M(n,\mathbb{R})$, restricted to $Sym(n,\mathbb{R}) = T_{\mathbf{E}}Sym(n,\mathbb{R})^+$, i.e. to the quadratic form

$$\forall\, \mathbf{X} \in Sym(n, \mathbb{R}) : \quad Q(\mathbf{X}) := 4\; \mathrm{tr}\left(\mathbf{X}^2\right)$$

on $Sym(n,\mathbb{R})$. As a Riemannian metric it is, in classical notation,

$$\mathrm{d}s^2 = 4\; \mathrm{tr}\left(\left(\sqrt{\mathbf{X}^{-1}}\,\mathrm{d}\mathbf{X}\,\sqrt{\mathbf{X}^{-1}}\right)^2\right) = 4\; \mathrm{tr}\left(\left(\mathbf{X}^{-1}\,\mathrm{d}\mathbf{X}\right)^2\right) \tag{19}$$

where $\mathbf{X} = (X_{ij})$ is the matrix of the natural coordinates on $Sym^+(n,\mathbb{R})$ and $\mathrm{d}\mathbf{X} = (dX_{ij})$ is a matrix of differentials.

The same conclusions hold for the space $SSym(n,\mathbb{R})$ of real symmetric positive definite matrices of determinant one, when one replaces the general linear group $GL(n,\mathbb{R})$ with the special linear group $SL(n,\mathbb{R})$, the $n \times n$–matrices of determinant one, and the space of real symmetric matrices $Sym(n,\mathbb{R})$ with the space $Sym_0(n,\mathbb{R})$ of real symmetric traceless matrices.

Remark 2. Although the expression (19) appears to be explicit in the coordinates, it seems to be of no use for analyzing the properties of the corresponding Riemannian metric, since the operations of inverting, squaring, and taking the trace gives, in the general case, untractable expressions. In particular, it apparently is of no help in deriving the expression (14) for the associated distance by direct elementary means.

There is, however, one interesting case where it can be checked to give a very classical expression; this is the case $n = 2$. In this case, one has

$$SSym^+(2,\mathbb{R}) = \left\{ \begin{pmatrix} x & y \\ y & z \end{pmatrix} \;\middle|\; x, y, z \in \mathbb{R}\,,\; xz - y^2 = 1\,,\; x > 0 \right\}$$

a hyperboloid in 3-space. This is classically known as a candidate for a model of the hyperbolic plane.In fact, in this case, one may show by explicit computation that the metric (19) restricts to the classical hyperbolic metric, and that the corresponding distance just gives one of the classical formulas for the hyperbolic distance. For details, see [11].

Of course, the next question is which invariant metrics there are. Also this question can be answered:

Addendum. *(i) The positive definite quadratic forms Q on $Sym(n,\mathbb{R})$ invariant under conjugation with orthogonal matrices are of the form*

$$Q(\mathbf{X}) = \alpha\; \mathrm{tr}\left(\mathbf{X}^2\right) + \beta\big(\; \mathrm{tr}\,(\mathbf{X})\;\big)^2, \qquad \alpha > 0, \beta > -\frac{\alpha}{n}$$

(ii) The positive definite quadratic forms Q on $Sym_0(n,\mathbb{R})$ invariant under conjugation with orthogonal matrices are unique up to a positive scalar and hence of the form

$$Q(\mathbf{X}) = \alpha\; \mathrm{tr}\left(\mathbf{X}^2\right), \qquad \alpha > 0.$$

In particular, the Riemannian metric (19) corresponds to the case $\alpha = 1$, $\beta = 0$. Since from the point of this classification all these metrics stand on an equal footing, it would be interesting to know by which naturality requirements this choice can be singled out.

‡]the famous Cartan-Killing-form of Lie group theory

6 The proofs

To put this result into proper perspective and to cut a long story short, let us very briefly summarize why Theorem 2, and consequently Theorem 1, are true. First, however, we indicate a proof of the Proposition above, since it is on this Proposition that our approach to the triangle equality for the distance defined by (14) is based.

The fact that $d(p,q) \geq 0$ and the symmetry of d are immediate from the definitions. There remains to show $d(p,q) = 0 \implies p = q$ and the triangle inequality.

For given $p \in M$, choose a coordinate neighbourhood $U \cong \mathbb{R}^n$ around p such that p corresponds to $0 \in \mathbb{R}^n$. We then have the expression (16) for the given metric in U. Moreover, we have in U the standard Euclidean metric

$$\mathrm{d}s_E^2 := \delta_{ij}\mathrm{d}x^i\mathrm{d}x^j = \sum_{i=1}^{n}(dx^i)^2 .$$

Let $\|-\|$ denote the norm belonging to the given Riemannian metric in U and $|-|$ the norm given by the standard Euclidean metric. For $r > 0$ let

$$\overline{\mathbb{B}}(p;r) := \{\, x \in \mathbb{R}^n \;\mid |x| \leq r \,\}$$

be the standard closed ball with radius r around $p = 0$, and

$$\mathbb{S}(p;r) := \{\, x \in \mathbb{R}^n \;\mid |x| = r \,\}$$

its boundary, the sphere of radius r around $p \in \mathbb{R}^n$.

As a continuous function $U \times \mathbb{R}^n \longrightarrow \mathbb{R}$ the norm $\|-\|$ takes its minimum $m > 0$ and its maximum $M > 0$ on the compact set $\overline{\mathbb{B}}(p;1) \times \mathbb{S}(p;1)$. It follows that we have

$$\forall\, q \in \overline{\mathbb{B}}(p;1)\,,\; X \in \mathbb{R}^n : \quad m|X| \leq \|X\|_q \leq M|X|$$

by homogeneity of the norm, and so by integrating and taking the infimum

$$\forall\, q \in \overline{\mathbb{B}}(p;1) : \quad m d_E(p,q) \leq d(p,q) \leq M d_E(p,q) \tag{20}$$

where $d_E(p,q) = |q-p|$ is the Euclidean distance. If $q \notin \overline{\mathbb{B}}(p;1)$, then any path c joining p to q meets the boundary $\mathbb{S}(p;1)$ in some point r, from which follows $L\,[c] \geq L\,[c'] \geq d(p,r) \geq m$ – where c' denotes the part of c joining p to r for the first time, say – whence $d(p,q) \geq m$. In other words, if $d(p,q) < m$ we have $q \in \overline{\mathbb{B}}(p;1)$, where we can apply (20). If now $d(p,q) = 0$, then surely $d(p,q) < m$, and then by (20) $m d_E(p,q) \leq d(p,q) = 0$, whence $d_E(p,q) = 0$, which implies $p = q$.

For the triangle inequality, let c be a path joining p to q and d a path joining q to r. Let $c * d$ be the composite path joining p to r. Then $L\,[c*d] = L\,[c] + L\,[d]$. Taking the infimum on the left hand side over all paths joining p to r gives $d(p,r) \leq L\,[c] + L\,[d]$. Taking on the right hand side first the infimum over all paths joining p to q and subsequently over all the paths joining q to r then gives $d(p,r) \leq d(p,q) + d(q,r)$, which is the triangle inequality.

Remark 3. In particular, (20) shows that the metric topology induced by the distance d on a connected Riemannian manifold coincides with the given manifold topology.

Now to the proof of Theorem 2. Recall the terminology of [10], Chapter X: Let G be a Lie group with Lie algebra $\mathfrak{g}$, $H \subseteq G$ a closed Lie subgroup corresponding to the Lie subalgebra $\mathfrak{h} \subseteq \mathfrak{g}$. Let M be the homogeneous space $M = G/H$. Then G operates as a symmetry group on M by left translations. M has the distinguished point $o = eH = H$ corresponding to the coset of the unit element $e \in G$ with tangent space $T_oM = \mathfrak{g}/\mathfrak{h}$. This homogeneous space is called *reductive* if $\mathfrak{g}$ splits as a direct vector space sum $\mathfrak{g} = \mathfrak{h} \oplus \mathfrak{m}$ for a linear subspace $\mathfrak{m} \subseteq \mathfrak{g}$ such that $\mathfrak{m}$ is invariant under the adjoint action $\mathrm{Ad} : H \longrightarrow GL(\mathfrak{g})$. Then canonically $T_oM = \mathfrak{m}$. In our situation, $G = GL(n, \mathbb{R})$, $H = O(n)$. Then $\mathfrak{g} = M(n, \mathbb{R})$, the full $n \times n$-matrices, and $\mathfrak{h} = Asym(n, \mathbb{R})$, the antisymmetric matrices. As is well known,

$$M(n, \mathbb{R}) = Asym(n, \mathbb{R}) \oplus Sym(n, \mathbb{R})\,,$$

since any matrix $\mathbf{X}$ splits into the sum of its antisymmetric and symmetric part via

$$X = \frac{X - X^T}{2} + \frac{X + X^T}{2}\,.$$

The adjoint action of $\mathbf{O} \in O(n)$ on $M(n, \mathbb{R})$ is given by $\mathbf{X} \mapsto \mathbf{OXO}^{-1} = \mathbf{OXO}^T$ and clearly preserves $Sym(n, \mathbb{R})$. So $Sym^+(n, \mathbb{R})$ is a reductive homogeneous space.

We now have the following facts from the general theory:

a) On a reductive homogeneous space there is a distinguished connection invariant under the action of G, called the *natural torsion free connection* in [10]. It is uniquely characterized by the following properties ([10], Chapter X, Theorem 2.10)

- It is G–invariant
- Its geodesics through $o \in M$ are the orbits of o under the one-parameter subgroups of G, i.e. of the form $t \mapsto \exp(tX) \cdot o$ for some $X \in \mathfrak{g}$, where $\exp : \mathfrak{g} \longrightarrow G$ is the exponential mapping of Lie group theory
- It is torsion free

In particular, with this connection M becomes an *affine locally symmetric space*, i.e. the geodesic symmetries at a point of M given by inflection in the geodesics locally preserve the connection (*loc. cit* Chapter XI, Theorem 1.1). If M is simply connected, M is even an *affine symmetric space*, i.e. the geodesic symmetries extend to globally defined transformations of M preserving the connection (*loc. cit.*, Chapter XI, Theorem 1.2). By homogeneity, these are determined by the geodesic symmetry s at o. In our case $M = Sym^+(n, \mathbb{R})$, M is even contractible, hence simply connected, and so with the natural torsion free connection an affine symmetric space. We have $o = \mathbf{E}$, the $n \times n$ unit matrix. For $G = GL(n, \mathbb{R})$, the exponential mapping of Lie group theory is given by the "naive matrix exponential" $e^{\mathbf{X}} = \sum_{k=0}^{\infty} t^k \mathbf{X}^k / k!$. So the geodesics are $t \mapsto \exp(t\mathbf{X})\,\mathbf{E}\,\exp(t\mathbf{X})^T = e^{2t\mathbf{X}}$, where $\mathbf{X} \in Sym(n, \mathbb{R})$, and s is given by $s(\mathbf{X}) = \mathbf{X}^{-1}$.

b) The Riemannian metrics g on M invariant under the action of G are in one-to-one-correspondence with positive definite quadratic forms Q on $\mathfrak{m}$ invariant under the adjoint action of H (*loc. cit, Chapter X*, Corollary 3.2), the correspondence being given by

$$\forall\, X \in T_oM = \mathfrak{m} : \quad g_o(X, X) = Q(X)\,.$$

This is intuitively obvious, since we can translate o to any point of M by operating on it with an element $g \in G$.

c) All G-invariant Riemannian metrics on M (there may be none) have the natural torsion free connection as their Levi-Cività connection (*loc. cit*, Chapter XI, Theorem 3.3). In particular, such a metric makes M into a Riemannian (locally) symmetric space, i.e. the geodesic symmetries are isometries, and the exponential map of Riemannian Geometry at o, $\mathrm{Exp}_o : T_oM = \mathfrak{m} \longrightarrow M$ is given by the exponential map of Lie group theory for G:

$$\forall\, X \in \mathfrak{m} : \quad \mathrm{Exp}_o(tX) = \exp(tX) \cdot o\,.$$

Collecting these results, we now can come to terms with formula (4). First we see that Part (i) of Theorem 2 is a standard result in the theory of homogeneous spaces. Furthermore, S^+, being a Riemannian symmetric space with the metric (19), is complete (*loc. cit*, Chapter XI, Theorem 6.4), the exponential mapping $\mathrm{Exp}_{\mathbf{E}}$ of Riemannian geometry is related to the exponential mapping $\exp : \mathrm{S} \longrightarrow \mathrm{S}^+$, $S = T_{\mathbf{E}}S^+$ from Lie theory and the matrix exponential $e^{\mathbf{X}}$ via $\mathrm{Exp}_{\mathbf{E}}(\mathbf{X}) = \exp(2\mathbf{X}) = e^{2\mathbf{X}}$ and is a diffeomorphism [§] .

Having reached this point, here is the showdown. Since, by general theory, the Riemannian exponential mapping is a radial isometry, we get for the square of the distance d_Q:

$$d_Q^2(\mathbf{A}, \mathbf{B}) = d_Q^2(\mathbf{E}, \sqrt{\mathbf{A}^{-1}}\mathbf{B}\sqrt{\mathbf{A}^{-1}})$$

since d_Q is invariant under congruences by (15) ,

$$= Q(\frac{1}{2}\exp^{-1}(\sqrt{\mathbf{A}^{-1}}\mathbf{B}\sqrt{\mathbf{A}^{-1}}))$$

since $\mathrm{Exp}_{\mathbf{E}}$ is a radial isometry ,

$$= \frac{1}{4}Q(\ln(\sqrt{\mathbf{A}^{-1}}\mathbf{B}\sqrt{\mathbf{A}^{-1}})) \, ,$$

and this is just equation (18). In particular, from this one directly reads off that the distance is invariant under inversion, as claimed. Of course, the invariances in question are for the particular case corresponding to (14) read off easily from the classical form (19) of the Riemannian metric. On the other hand, we see that the invariance under inversion comes from the structural facts that S^+ is a symmetric space, and that the geodesic symmetry at E, which on general grounds must be an isometry, is just given by matrix inversion (see a) above).

One should add that these arguments are general and pertain to the situation of a symmetric space of the non–compact type; for this, see [11].

The representation of the orthogonal group $O(n)$ on the symmetric matrices by conjugation is not irreducible, but decomposes as

$$Sym(n, \mathbb{R}) = Sym_0(n, \mathbb{R}) \oplus \mathfrak{d}(n, \mathbb{R}), \tag{21}$$

where $\mathfrak{d}(n, \mathbb{R})$ are the scalar diagonal matrices. It is easy to see that both summands are invariant under conjugation with orthogonal matrices, and it can be shown that both parts are irreducible representations of $O(n)$. From this it is standard to derive the Addendum. In the geometric framework of symmetric spaces, this describes the decomposition of the holonomy representation and correspondingly the canonical DE RHAM–decomposition

$$Sym^+(n, \mathbb{R}) \cong SSym(n, \mathbb{R}) \times \mathbb{R}^+$$

of the symmetric space $Sym(n, \mathbb{R})$ into irreducible factors. This is a direct product of Riemannian manifolds, i.e. the metric on the product is just the product of the metrics on the individual factors, that is given by the Pythagorean description. Thus it suffices to classify the invariant metrics on the individual factors, which accounts for the Addendum.

Thus, it transpires that the theorems above follow from the basics of Lie group theory and Differential Geometry and so should be clear to the experts. The main results upon which it is based appeared originally in the literature in [12]. All in all, it follows in a quite straightforward

[§] The fact that the naive matrix exponential is a diffeomorphism, whence S^+ is complete, can be seen by elementary means in the case under consideration. The main point is that it coincides with the exponential mapping coming from Riemannian Geometry (up to scaling with a factor of 2).

manner from the albeit rather elaborated machinery of modern Differential Geometry and the theory of symmetric spaces. In conclusion, it might therefore be still interesting to give a more elementary derivation of the result, as was done above in the case $n = 2$.

As a general reference for Differential Geometry and the theory of symmetric spaces I recommend [9], [10] (which, however, make quite a terse reading). A detailed exposition [11] covering all the necessary prerequisites is under construction; the purpose of this paper is to introduce the non–experts to all the basic notions of Differential Geometry and to expand the brief arguments just sketched.

References

[1] Baarda, W. (1973): *S-Transformations and Criterion Matrices*, Band 5 der Reihe 1. Netherlands Geodetic Commission, 1973.

[2] Ballein, K., 1985, Untersuchung der Dreiecksungleichung beim Vergleich von Kovarianzmatrizen, persönliche Mitteilung, 1985

[3] Förstner, W., A Metric for Comparing Symmetric Positive Definite Matrices, *Note, August 1995*

[4] Fukunaga, K., *Introduction to Statistical Pattern Recognition*, Academic Press, 1972

[5] Grafarend, E. W. (1972): *Genauigkeitsmasse geodätischer Netze.* DGK A 73, Bayerische Akademie der Wissenschaften, München, 1972.

[6] Grafarend, E. W. (1974): Optimization of geodetic networks. *Bolletino di geodesia e Scienze Affini*, 33:351–406, 1974.

[7] Grafarend, E. W. and Niermann, A. (1994): Beste echte Zylinderabbildungen, *Kartographische Nachrichten*, 3:103–107, 1984.

[8] Kavrajski, V. V., *Ausgewählte Werke*, Bd. I: Allgemeine Theorie der kartographischen Abbildungen, Bd. 2. Kegel- und Zylinderabbildungen (russ.), GVSMP, Moskau, 1958, zitiert in [7]

[9] Kobayashi, S. & Nomizu, K., *Foundations of Differential Geometry*, Vol. I, Interscience Publishers 1963

[10] Kobayashi, S. & Nomizu, K., *Foundations of Differential Geometry*, Vol. II, Interscience Publishers 1969

[11] Moonen, B, *Notes on Differential Geometry*, `ftp://ftp.uni-bonn.cd/pub/staff/moonen/symmc.ps.gz`

[12] Nomizu, K., Invariant affine connections on homogeneous spaces, *Amer. J. Math, 76 (1954)*, 33-65

[13] Schmitt, G. (1983): Optimization of Control Networks – State of the Art. In: Borre, K.; Welsch, W. M. (Eds.), *Proc. Survey Control Networks*, pages 373–380, Aalborg University Centre, 1983.