


## ORIGINAL ARTICLE

WILEY

# Anomaly detection of time series correlations via a novel Lie group structure

Paul David<sup>1</sup>  | Weiqing Gu<sup>2</sup>

<sup>1</sup>Occidental College, Los Angeles, California, USA

<sup>2</sup>Harvey Mudd College, Claremont, California, USA

## Correspondence

Paul David, Occidental College, Los Angeles, CA, USA.

Email: [pdavid@oxy.edu](mailto:pdavid@oxy.edu)

We investigate a natural Lie group structure that correlations possess, interpreted as a special case of the quotient manifold structure for  $n \times n$  correlation matrices. The one-dimensional setting of the general theory gives rise to additional structure in the form of a natural associative multiplication of correlations making the space of correlations into a Lie group. We explicitly compute left-invariant vector fields, exponential and logarithmic mappings, and ultimately, a closed-form distance formula between correlations. The Lie group formalism is then applied to an application in anomalous correlation detection. The advantage of the Lie group method illustrated with time series data of stock prices.

## KEYWORDS

anomaly detection, correlations, Lie groups, Riemannian geometry

## 1 | INTRODUCTION

Correlations serve as one of the most commonly used metrics to discern the relationship between random variables. Their use is ubiquitous across the fields of statistics, economics and data science, among others, in part because of the computational ease and direct interpretability of linear relationships between data. Modern research into correlation matrices and their uses include estimation of correlation values (Fushiki, 2009; Muniz et al., 2021), cleaning correlation matrices for correcting errors (Bun et al., 2016, 2017) and a continued focus on correlation for financial markets (Pharasi et al., 2021). We introduce in this manuscript a new lens through which to view correlation values via a novel Lie group structure. While this may seem at the first glance to be a trivial characterization of correlation values (after all the set of correlations is just the interval  $(-1, 1)$ ), what we uncover is a natural Lie group structure derived from the manifold formalism of  $n \times n$  correlation matrices developed in David and Gu (2019). There are several implications that stem from this:

- The Lie group geometry imbues the set of correlations with a geometrically motivated metric so that there is a clear way to define distance between correlations other than a simple linear distance. The Lie group metric is *hyperbolic* in nature in that correlations become infinitely far away from other points as they approach  $\pm 1$ .
- The Lie algebra associated to the Lie group of correlations is actually diffeomorphic to the space of correlations via the exponential map. This artifact is utilized for the process of computing averages and standard deviations of sampled correlation values, allowing for a new interpretation of sampled correlations as statistical quantities used for analysis of real world data. We apply these ideas towards anomaly detection of recurrent times series data to identify instances in which random variables are highly correlated beyond normally observed behaviour. Furthermore, the methods of metastatistics of correlations are accomplished in a way that is intrinsic to the manifold  $\text{Corr}(2)$ .
- The methods presented here are a special case of the results presented in David and Gu (2019) and present a significant improvement in computational energy expended to compute averages of correlations as well as a distance function. Subsequent efforts have been made to simplify

This is an open access article under the terms of the [Creative Commons Attribution-NonCommercial-NoDerivs](https://creativecommons.org/licenses/by-nc-nd/4.0/) License, which permits use and distribution in any medium, provided the original work is properly cited, the use is non-commercial and no modifications or adaptations are made.

© 2022 The Authors. *Stat* published by John Wiley & Sons Ltd.

the quotient manifold formalism for correlations, notably in Thanwerdas and Pennec (2021) where the authors specified closed-form expressions for geodesics in  $\text{Corr}(n)$  in terms of horizontal and vertical distributions. Despite this significant improvement, a closed-form computation for distances between correlation matrices has not yet been found and is still relegated to an optimization problem. As a result averages of correlations are expressed as the minimizer of a learning problem defined on  $\text{Corr}(n)$  and is a highly expensive process to compute. In the present manuscript, though we merely concern ourselves with the one-dimensional case, it is arguably of greater interest to researchers who work with individual correlation values rather than large correlation matrices. In this instance, the learning procedure presented in David and Gu (2019) is bypassed entirely for the more favourable Lie group formalism which relies on direct computation of values.

The manuscript proceeds as follows: in Section 2 we summarize the work presented (David & Gu, 2019) and motivate the foundational observation giving rise to this novel Lie group structure. We derive a multiplication of correlation values giving rise to the Lie group structure, compute exponential and logarithmic maps and finally derive a closed form distance formula for correlations. In Section 3 we prescribe a general method of averaging Lie group-valued elements. Computations within Lie groups have proven useful in a number of applications including averaging of MRI voxels represented as  $3 \times 3$  symmetric positive-definite matrices (Arsigny et al., 2007; Lui, 2012; Moakher & Zéraï, 2010) as well as state estimation of stochastic processes (Muniz et al., 2021). This is then extended to adapting the definition of standard deviation to distances within the Lie group. Our ability to perform these computations stems from the global diffeomorphism that exists between the Lie group and its Lie algebra (the data are mapped to the Lie algebra, computations are performed there, and the results translated back to the Lie group). This gives us a way of defining intrinsic standard deviation windows within  $\text{Corr}(2)$ , something that is not possible if we are simply viewing  $\text{Corr}(2)$  as a subset of  $\mathbb{R}$ . These notions of averaging and standard deviation of correlation values are applied to a problem in anomaly detection and identifying instances of above average correlation between random variables. This method is illustrated between historical stock data and averaging correlation sequences for the same interval of time over various years (in our case, we average correlations of closing prices of two stocks for the first fiscal quarter over a sequence of years). We offer a brief summary of our findings in Section 4.

## 2 | LIE GROUP STRUCTURE FOR CORRELATIONS

### 2.1 | The quotient manifold geometry of $\text{Corr}(n)$

What we aim to demonstrate, namely, that correlations possess a natural Lie group structure, stems from a more general result presented in David and Gu (2019) where it was shown that correlations possess a quotient manifold geometry. Letting  $\text{SPD}(n)$  be the space of symmetric positive-definite matrices (interpreted as the space of non-degenerate  $n \times n$  covariances),  $\text{Diag}_+(n)$  the  $n \times n$  diagonal matrices with positive entries, and  $\text{Corr}(n)$  the space of  $n \times n$  non-degenerate correlations (elements of  $\text{SPD}(n)$  with unit diagonal), it was shown that  $\text{Corr}(n)$  is the quotient manifold

$$\text{Corr}(n) = \text{SPD}(n)/\text{Diag}_+(n) \quad (1)$$

where the equivalence relation is that generated by the action  $\text{Diag}_+(n) \times \text{SPD}(n) \rightarrow \text{SPD}(n)$  given by  $(D, P) \rightarrow DPD$ , where the action was shown to be smooth, free and proper. A standard theorem of classical differential geometry shows this and can be found in Lee (2012). We subsequently see that  $\dim \text{Corr}(n) = \frac{n(n-1)}{2}$ .

It was further shown that the affine-invariant metric of  $\text{SPD}(n)$  given by

$$\langle X, Y \rangle_P = \text{Tr}(P^{-1}XP^{-1}Y) \quad (2)$$

remains invariant under the action of the Lie group  $\text{Diag}_+(n)$ . Denoting the action by  $\Phi_D$  for each  $D \in \text{Diag}_+(n)$ , it is a straightforward exercise to show

$$\langle d(\Phi_D)_P(X), d(\Phi_D)_P(Y) \rangle_{\Phi_D(P)} = \langle X, Y \rangle_P. \quad (3)$$

A theorem from Huckemann et al. (2010) demonstrates that an isometric action of a Lie group gives rise to simplified computations for geodesics and distances in the quotient space. This computation takes the form of an optimization problem in the quotient. Recall the expressions for geodesics and distances in  $\text{SPD}(n)$  between elements  $P_1, P_2$

$$\begin{aligned} \gamma_{\text{SPD}}(t) &= P_1^{1/2} \text{Exp} \left[ t \text{Log} \left( P_1^{-1/2} P_2 P_1^{-1/2} \right) \right] P_1^{1/2} \\ d_{\text{SPD}}(P_1, P_2) &= \left\| \text{Log} \left( P_1^{-1/2} P_2 P_1^{-1/2} \right) \right\|_F \end{aligned} \quad (4)$$

where  $\|\cdot\|_F$  represents the Frobenius norm. The work presented in Huckemann et al. (2010) then shows we can compute the corresponding geodesics and distances as solutions to the optimization:

$$\begin{aligned} d_{\text{Corr}}(C_1, C_2) &= \inf_{D \in \text{Diag}_+(n)} d_{\text{SPD}}(C_1, DC_2D) \\ D^* &= \operatorname{arginf}_{D \in \text{Diag}_+(n)} d_{\text{SPD}}(C_1, DC_2D) \\ \gamma_{\text{Corr}}(t) &= \pi \left( C_1^{1/2} \operatorname{Exp} \left[ t \operatorname{Log} \left( C_1^{-1/2} D^* C_2 D^* C_1^{-1/2} \right) \right] C_1^{1/2} \right) \end{aligned} \quad (5)$$

where  $\pi: \text{SPD}(n) \rightarrow \text{Corr}(n)$  is the quotient map  $\pi(P) = D_P^{-1/2} P D_P^{-1/2}$  where  $D_P$  is the diagonal matrix containing the diagonal entries of  $P$ .

One of the primary challenges to the above formalism is the efficient computation of optimal solutions  $D^*$ . In David and Gu (2019), a gradient descent procedure was utilized within the fibre space  $\text{Diag}_+(n)$  to find these points minimizing the distance. In the least squares distance problem addressed in David and Gu (2019) this type of computation is repeated for as many observations as there are in the dataset, incurring a large computational cost. Faster algorithms were presented in Grubišić and Pietersz (2005), which relied on a different manifold geometry of correlations stemming from the Cholesky decomposition of symmetric positive-definite matrices. Grubišić and Pietersz (2005) were then able to take advantage of the orthogonality constraints of Cholesky decompositions and directly apply the results of Edelman et al. (1998). While this presents a fast method of optimization for correlations, the approach taken in the present manuscript is to emphasize the intrinsic quotient manifold relationship between correlations and covariances via the affine-invariant geometry. As we will see in the next section, this interpretation of correlations presents many advantages including an intrinsic Lie group structure whose left-invariant distance is precisely that inherited from the affine-invariant metric.

## 2.2 | Intrinsic Lie group structure for correlations

The Lie group structure we discuss for correlations is limited to the  $2 \times 2$  case where we see explicitly that

$$\text{Corr}(2) = \left\{ \begin{bmatrix} 1 & x \\ x & 1 \end{bmatrix} : -1 < x < 1 \right\}. \quad (6)$$

While seemingly elementary, there is in fact a rich geometric structure for correlations which is the first to our knowledge to be exhibited in the research literature. In the work that follows we will adopt the notation

$$C(x) := \begin{bmatrix} 1 & x \\ x & 1 \end{bmatrix}. \quad (7)$$

Additionally, when the context is clear, we may also choose to abuse notation by referring directly to the value  $x$  of the correlation rather than the  $2 \times 2$  matrix  $C(x)$ . The key observation in this work is to note that the ordinary matrix multiplication of two elements of  $\text{Corr}(2)$  remains symmetric. Namely, we observe

$$C(x)C(y) = \begin{bmatrix} 1+xy & x+y \\ x+y & 1+xy \end{bmatrix}. \quad (8)$$

This observation is additionally illuminated by the fact that in the spectral decomposition of  $\text{Corr}(2)$ , elements are all factored with respect to the same orthogonal matrix. Noting that the result in Equation (8) is still positive-definite, we can simply follow this by the quotient map to obtain a product  $*$ :  $\text{Corr}(2) \times \text{Corr}(2) \rightarrow \text{Corr}(2)$

$$C(x) * C(y) = \pi(C(x)C(y)) = \begin{bmatrix} 1 & \frac{x+y}{1+xy} \\ \frac{x+y}{1+xy} & 1 \end{bmatrix}. \quad (9)$$

When the context is clear, we will refer to this product directly on the correlation values themselves as  $x * y = \frac{x+y}{1+xy}$ .

**Theorem 1.** Correlations, interpreted as the space  $\text{Corr}(2)$  possesses the structure of a commutative Lie group.

*Proof.* The product  $*$ :  $\text{Corr}(2) \times \text{Corr}(2) \rightarrow \text{Corr}(2)$  is seen to be smooth since it is the composition of ordinary matrix multiplication and the quotient map. It is easy to verify that  $C(0) = I$  is the group identity as well that  $C(x)^{-1} = C(-x)$ . The last property to show is associativity. One can verify that

$$(x * y) * z = x * (y * z) = \frac{x + y + z + xyz}{1 + xy + xz + yz}. \quad (10)$$

Commutativity should be apparent from the definition of the multiplication map.  $\square$

The next aspect of the group structure we investigate is the Lie algebra. It should be clear that via the tangent space representation that

$$\mathfrak{corr}(2) = T_0 \text{Corr}(2) = \left\{ \begin{bmatrix} 0 & s \\ s & 0 \end{bmatrix} : s \in \mathbb{R} \right\}. \quad (11)$$

At first glance the Lie algebra may appear devoid of an interesting structure, especially since  $\text{Corr}(2)$ , being one-dimensional, yields a trivial bracket operation in the Lie algebra. Where we instead discern a great deal of information is when we compute the left-invariant vector fields of  $\text{Corr}(2)$ . For ease of computation, we will define the matrix

$$E := \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}. \quad (12)$$

**Theorem 2.** The left-invariant vector fields  $S \in \mathfrak{corr}(2)$  are of the form

$$S_x = s(1 - x^2)E. \quad (13)$$

*Proof.* Specifying an element  $S_0 \in T_0 \text{Corr}(2)$ , and letting  $s$  be the off-diagonal element of  $S_0$ , we can then extend  $S_0$  to a left-invariant vector field  $S \in \mathcal{X}(\text{Corr}(2))$  where  $S_x = (L_x)_*(S_0)$  where  $L_x : \text{Corr}(2) \rightarrow \text{Corr}(2)$  is left multiplication  $L_x(y) = x * y$  and  $(L_x)_* : T_y \text{Corr}(2) \rightarrow T_{x*y} \text{Corr}(2)$  is its pushforward. Given a curve  $\gamma : (-\epsilon, \epsilon) \rightarrow \text{Corr}(2)$  where  $\gamma(0) = 0$  and  $\gamma'(0) = s$  we compute the pushforward as

$$s_x = (L_x)_*(s) = \left. \frac{d}{dt} \right|_{t=0} L_x \gamma(t) = \left. \frac{d}{dt} \right|_{t=0} \frac{x + \gamma(t)}{1 + x\gamma(t)} = s(1 - x^2).$$

$\square$

To make the relationship between the Lie algebra and the Lie group concrete, we would like to derive the exponential map  $\text{Exp} : \mathfrak{corr}(2) \rightarrow \text{Corr}(2)$  and potentially recognize if there exists an inverse to this mapping. One should carefully note that we cannot simply adopt the matrix exponential in this case. Indeed, the matrix exponential would have worked in the case that ordinary matrix multiplication gave rise to the group structure. In order to properly derive an exponential map, we need to recall that in the general setting, given a Lie group  $G$  with Lie algebra  $\mathfrak{g}$ , the exponential map  $\text{Exp} : \mathfrak{g} \rightarrow G$  is defined as

$$\text{Exp}(X) = \gamma_X(1) \quad (14)$$

where  $\gamma_X(t)$  is the maximal integral curve with respect to  $X$  (i.e., a curve of maximal domain with  $\dot{\gamma}(t) = X_{\gamma(t)}$  for all  $t$ ). It is precisely from this abstract form of the definition that we uncover precisely the hyperbolic nature of correlations.

**Corollary 1.** The exponential map  $\text{Exp} : \mathfrak{corr}(2) \rightarrow \text{Corr}(2)$  is given by

$$\text{Exp}(sE) = I + \tanh(s)E. \quad (15)$$

By the invertibility of  $\tanh : \mathbb{R} \rightarrow (-1, 1)$ , we similarly obtain a globally defined logarithmic map  $\text{Log} : \text{Corr}(2) \rightarrow \mathfrak{corr}(2)$

$$\text{Log}(I + xE) = \frac{1}{2} \ln \left( \frac{1+x}{1-x} \right) E. \quad (16)$$

*Proof.* The definition of the exponential map arises directly from computing the maximal integral curves of the left-invariant vector fields. The condition that  $\dot{\gamma}(t) = S_{\gamma(t)}$  gives rise to the ODE

$$\dot{\gamma} = s(1 - \gamma^2). \quad (17)$$

Solving this differential equation is an elementary exercise. With  $\gamma(0) = 0$ , we obtain  $\gamma(t) = \tanh(st)$ . The definition of the logarithm given in Equation (16) can be obtained by inverting  $\tanh(s)$ . In the matrix representation of this group we find

$$\begin{aligned} \text{Exp}\left(\begin{bmatrix} 0 & s \\ s & 0 \end{bmatrix}\right) &= \begin{bmatrix} 1 & \tanh(s) \\ \tanh(s) & 1 \end{bmatrix} \\ \text{Log}\left(\begin{bmatrix} 1 & x \\ x & 1 \end{bmatrix}\right) &= \begin{bmatrix} 0 & \frac{1}{2} \ln\left(\frac{1+x}{1-x}\right) \\ \frac{1}{2} \ln\left(\frac{1+x}{1-x}\right) & 0 \end{bmatrix}. \end{aligned} \quad (18)$$

□

A couple of points are of interest here. First, it is remarkable to see that the group multiplication is in fact an expression of a hyperbolic trigonometric identity.<sup>1</sup> As we see from the differential geometry literature, the maximal integral curves of left-invariant vector fields are precisely the one-parameter subgroups. We verify both of these points via the computation<sup>2</sup>

$$\tanh(s) * \tanh(k) = \frac{\tanh(s) + \tanh(k)}{1 + \tanh(s)\tanh(k)} = \tanh(s+k). \quad (19)$$

This, in fact, shows that Exp and Log in this context are group isomorphisms  $(\text{Corr}(2), *) \cong (\mathbb{R}, +)$ . Second, this characterization bears similarity to a Lie group structure imposed on the space  $\text{SPD}(n)$ . Arsigny et al. (2007) defined the following product on  $\text{SPD}(n)$

$$P \odot Q = \text{Exp}(\text{Log}(P) + \text{Log}(Q)) \quad (20)$$

called the *Log-Euclidean product*, where here  $\text{Exp} : \text{Symm}(n) \rightarrow \text{SPD}(n)$  and its inverse  $\text{Log} : \text{SPD}(n) \rightarrow \text{Symm}(n)$  are the ordinary matrix exponential and logarithm functions, both of which admit simple computations via the spectral theorem for symmetric matrices  $\text{Symm}(n)$ . The similarities between  $\text{Corr}(2)$  and  $\text{SPD}(n)$  is that these are both manifolds globally isomorphic to their canonical tangent spaces. In the case of Arsigny et al. (2007) the group structure imposed from the global diffeomorphism, and in the present case the isomorphism derived from a particular multiplication map. Following the formalism of Arsigny et al., we could in fact extend this isomorphism of groups to be an isomorphism of vector spaces by introducing the scalar product  $\lambda \otimes C = \text{Exp}(\lambda \text{Log}(C))$  for  $\lambda \in \mathbb{R}$ —an artifact made possible by the diffeomorphism  $\text{corr}(2) \cong \text{Corr}(2)$ . This article leaves unanswered the question of whether  $\text{Corr}(n)$  in general admits a group structure, but the similarity between these examples is striking and worthy of further investigation.

We are finally able to derive a closed form expression for the distance between correlation values by defining the left-invariant metric

$$\langle S, K \rangle_x = \langle (L_x^{-1})_*(S), (L_x^{-1})_*(K) \rangle_0 \quad (21)$$

where  $\langle S, K \rangle_0 = \text{Tr}(SK)$ .

**Corollary 2.** The distance function derived from the left-invariant metric of  $\text{Corr}(2)$  is given by

$$d_{\text{Corr}}(C(x), C(y)) = \sqrt{2} \left| \tanh^{-1}(x) - \tanh^{-1}(y) \right|. \quad (22)$$

*Proof.* We can interpret the distance between two points in a manifold as the length of the tangent vector of the geodesic connecting the two points. In a Lie group, we can utilize the multiplication map so that the distance from  $x$  to  $y$  is the same as the distance from the identity  $e$  to  $y^{-1}x$ . When a logarithm is defined between the spaces, we can then utilize the logarithm to compute the distance. In the case of  $\text{Corr}(2)$ ,

<sup>1</sup>This is in fact also an ordinary trigonometric identity.

<sup>2</sup>One might argue however that this characterization is somewhat redundant; the manifold  $\text{Corr}(2)$  is one-dimensional, hence there is in some sense only a single one-parameter subgroup—namely the whole group itself. Nevertheless, the hyperbolic nature of the group product is a novel characterization that is worthy of comment.

$$d_{\text{Corr}}(x, y) = d_{\text{Corr}}(0, y^{-1} * x) = \|\text{Log}(y^{-1} * x)\|_F.$$

It can be shown through some straightforward algebra that the off-diagonal terms are given by

$$\tanh^{-1}(y^{-1} * x) = \tanh^{-1}(x) - \tanh^{-1}(y).$$

The factor  $\sqrt{2}$  appears as a result of applying the Frobenius norm for the matrix  $\|A\|_F = \sqrt{\text{Tr}(A^T A)}$ .  $\square$

The discovery of a closed-form distance function of correlations in Equation (22) is a significant improvement from the algorithm presented in David and Gu (2019), at least in the one-dimensional setting, where the general approach to computing Riemannian correlation distances was through a descent algorithm on the group  $\text{Diag}_+(n)$ .

### 3 | ANOMALY DETECTION VIA LIE GROUP STATISTICS

#### 3.1 | Statistical computations in Lie groups

The main advantage of this newfound Lie group structure we wish to convey in this manuscript is the possibility of performing statistical analyses of Lie group-valued data with the standard computations of mean and standard deviation. In an arbitrary manifold, computation of the barycenter for a random sampling of points can only be defined via an optimization problem. In such cases, optimization is generally accomplished via a descent algorithm which incurs significant computational cost.

Lie groups by contrast are in fact naturally equipped to accommodate computations of barycenter and standard deviation neighbourhoods via the exponential and logarithmic maps. Given a Lie group  $G$  with corresponding Lie algebra  $\mathfrak{g}$ , and samples  $g_1, \dots, g_N \in G$ , we can compute their barycenter<sup>3</sup>  $\bar{g}$  as

$$\bar{g} := \text{Exp}\left(\frac{1}{N} \sum_{i=1}^N \text{Log}(g_i)\right). \quad (23)$$

Note that the summation occurs among elements of the Lie algebra which is a vector space. The corresponding standard deviation  $\sigma_G$  can then be computed as

$$\sigma_G := \sqrt{\frac{1}{N-1} \sum_{i=1}^N d_G^2(\bar{g}, g_i)} \quad (24)$$

where  $d_G : G \times G \rightarrow \mathbb{R}$  is the left-invariant distance derived from the left-invariant metric. This naturally gives rise to a neighbourhood of standard deviation around the barycenter of points within the Lie group. We recognize  $U_{\sigma_G}(\bar{g})$  to be the open submanifold of  $G$  consisting of those point  $x$  such that  $d_G(\bar{g}, x) < \sigma_G$ . In the analysis that follows any observation  $g_i$  not belonging to this set can be deemed **anomalous with precision**  $\sigma_G$ . Notice that we can extend this definition to consider  $n$ -standard deviations by considering the set  $U_{n\sigma_G}(\bar{g})$ .

In applications one can apply the above formalism for anomaly detection specifically in the case for correlations. Replacing  $\text{Exp}$  and  $\text{Log}$  for  $\tanh$  and  $\tanh^{-1}$ , respectively, as well as the expressions for  $d_{\text{Corr}}$  given in Equation (22) we can find the average and standard deviation of correlation samples  $c_1, \dots, c_N$  as

$$\bar{c} = \tanh\left(\frac{1}{N} \sum_{i=1}^N \tanh^{-1}(c_i)\right) \quad \sigma_{\text{Corr}} = \sqrt{\frac{2}{N-1} \sum_{i=1}^N \left(\tanh^{-1}(\bar{c}) - \tanh^{-1}(c_i)\right)^2}. \quad (25)$$

Often researchers are interested when random variables possess strong correlation. Detection of strong correlation events can be automated using the above formalism for Lie group statistics. It is worth discussing before proceeding further however why such a Lie group formalism is

<sup>3</sup>The key assumption necessary in this computation is that the logarithmic map  $\text{Log} : G \rightarrow \mathfrak{g}$  is defined at the observations in question. While this may not be possible for arbitrary Lie groups, this analysis remains valid for the present context since  $\text{Exp}$  and  $\text{Log}$  are global diffeomorphisms for  $\text{Corr}(2)$ .

desirable over a simple linear averaging.  $\text{Corr}(2)$  can quite simply be considered as a subinterval of  $\mathbb{R}$ , and linear averaging always yields a value belonging to  $\text{Corr}(2)$ . This is unfortunately however where the analysis ends because the concept of standard deviation is non-viable in the Euclidean context.  $\text{Corr}(2)$  as a set is the interval  $(-1,1)$ , thus to consider standard deviations can vary wildly yielding windows that go beyond the interval  $(-1,1)$ . Using the Lie group formalism, this issue does not arise since all computations are self-contained and guarantee inclusion in  $\text{Corr}(2)$ . The Lie group computations are also more strongly motivated since the inherent structure of  $\text{Corr}(2)$  as a Lie group is derived rather naturally from basic computations of matrix analysis, whereas the mere artifact of  $\text{Corr}(2)$  being a subset of  $\mathbb{R}$  is not sufficient criteria for using a Euclidean metric.  $\text{Corr}(2)$  after all is one-dimensional and bounded so therefore does not possess vector space structure under ordinary scalar addition.<sup>4</sup>

### 3.2 | Anomaly detection for stock time series

We demonstrate these concepts with anomaly detection of stock correlation time series data taken over various intervals. Assume we have two stocks  $X$  and  $Y$  each sampled with  $K$  separate time series labelled  $x^{(1)}, \dots, x^{(K)}$  and  $y^{(1)}, \dots, y^{(K)}$  and suppose each time series has  $N$  times  $t_1, \dots, t_N$  in which they are sampled. For instance, the quantity  $x^{(j)}(t_i)$  would represent the price of stock  $X$  in the  $j$ th series at time  $t_i$ . To illustrate, the upper index can be used to indicate the year while the time index  $t_i$  can be used to represent specific trading days within a given year. In this vein, the following analysis can be used for pattern recognition of stock behaviours over recurring time intervals.

To translate individual stock prices into correlations we need to select a collection of price values over which to compute the correlation. For simplicity, we select an integer  $s \geq 2$  to serve as a window size. We then define the correlation value  $c_{XY}^{(j)}(t_i)$  to be the Pearson correlation computed from the pairings<sup>5</sup>

$$(x^{(j)}(t_i), y^{(j)}(t_i)), (x^{(j)}(t_{i+1}), y^{(j)}(t_{i+1})), \dots, (x^{(j)}(t_{i+s}), y^{(j)}(t_{i+s})).$$

The goal is then to compute an average correlation series  $\bar{c}_{XY} : \{t_1, \dots, t_{N-s+1}\} \rightarrow \text{Corr}(2)$ , which is obtained from the Lie group averaging of each series lined up in every time interval. Thus, at time  $t_i$  we compute

$$\bar{c}_{XY}(t_i) = \tanh\left(\frac{1}{K} \sum_{j=1}^K \tanh^{-1}(c_{XY}^{(j)}(t_i))\right). \quad (26)$$

The corresponding standard deviation  $\sigma_{XY} : \{t_1, \dots, t_{N-s+1}\} \rightarrow \mathbb{R}$  is given at time  $t_i$  by

$$\sigma_{XY}(t_i) = \sqrt{\frac{2}{N-1} \sum_{j=1}^K \left( \tanh^{-1}(\bar{c}_{XY}(t_i)) - \tanh^{-1}(c_{XY}^{(j)}(t_i)) \right)^2}.$$

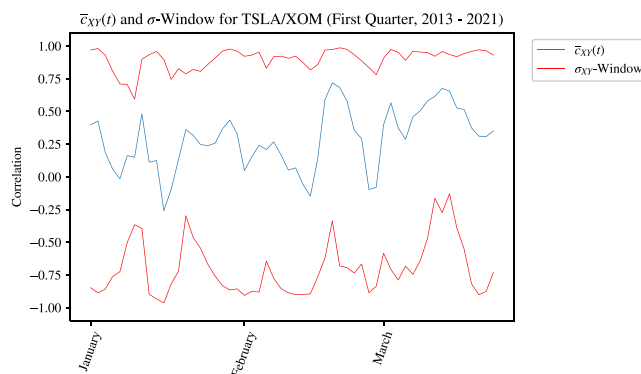
The interpretation of the standard deviation as a distance yields two additional time series representing the lower and upper bounds for a standard deviation window centred around the curve  $\bar{c}_{XY}$ . We will label these  $c_{XY}^{high}$  and  $c_{XY}^{low}$  and these can each be computed at time  $t_i$  as

$$\begin{aligned} c_{XY}^{low}(t_i) &= \tanh\left(\tanh^{-1}(\bar{c}_{XY}(t_i)) - \sigma_{XY}(t_i)\right) \\ c_{XY}^{high}(t_i) &= \tanh\left(\tanh^{-1}(\bar{c}_{XY}(t_i)) + \sigma_{XY}(t_i)\right). \end{aligned}$$

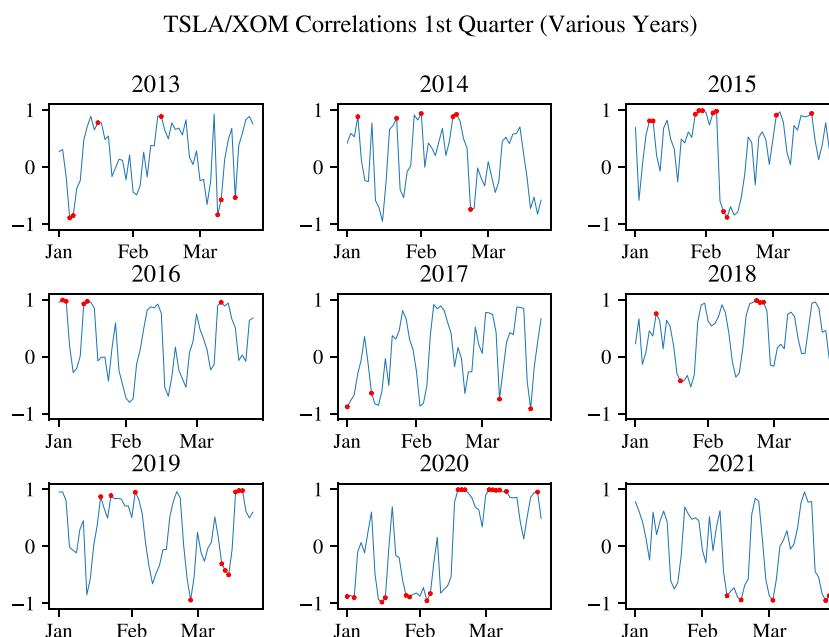
Note that the computations are finding the correlation values precisely one standard deviation below and above the average  $\bar{c}_{XY}$  at each time instance  $t_i$ . To illustrate these concepts concretely, we apply this method to historical stock data. We chose to compare Tesla (TSLA) with Exxon (XOM) in the first quarter across the years 2013–2021. Figure 1 shows the average curve across the first quarter along with the standard deviation window. Figure 2 shows the correlation curves for each individual year with anomalous correlation values highlighted.

<sup>4</sup>Note that there is in fact vector space structure under the Lie group formalism.

<sup>5</sup>Note that if each series has  $N$  time values, this necessarily makes the series of correlations only possess  $N - s + 1$  values per series since the last  $s$  prices have windows of size  $< s$  following them.



**FIGURE 1** Average correlation sequence and accompanying standard deviation window for Tesla (TSLA) and Exxon (XOM) closing prices between 1 January and 31 March averaged over the years 2013–2021. We take  $s = 5$  as the window for constructing correlations. Note that the asymmetry between the lower and upper bounds of the standard deviation window is due precisely to the hyperbolic geometry of  $\text{Corr}(2)$



**FIGURE 2** Correlations in the first quarter for Tesla (TSLA) and Exxon (XOM) for years 2013–2021. We mark correlation instances that are outside a standard deviation of the mean curve  $\bar{c}_{XY}(t)$  for each year

## 4 | CONCLUSION

In this paper, we have discerned a new understanding of correlations via a novel Lie group structure. This structure is derived from the quotient manifold geometry of  $n \times n$  correlation matrices and furthermore deepens our ability to use correlations as a computational tool for discerning distributions. In particular, the Lie group structure allows for statistical computations of correlation samples in a way that is geometrically motivated and guarantees inclusion in the set of correlations. This is an improvement over simple linear averaging which has dubious meaning when attempting to look at standard deviations within a sample of correlations. We demonstrated the effectiveness of the Lie group formalism by finding anomalous correlations of stock time series data and finding instances of extreme correlation of stock prices.

This new understanding of correlations via a Lie group structure adds additional perspectives on how we can use correlations as a statistical tool for data science applications. Further directions to consider using this Lie group structure include stochastic state estimation of correlation values as well as distinguishing distributions via the newly found distance formula in Equation (22).

## DATA AVAILABILITY STATEMENT

The data that support the findings of this study are available in Yahoo Finance at <https://finance.yahoo.com/>. These data were derived from the following resources available in the public domain: - Yahoo Finance, <https://finance.yahoo.com/>.



## ORCID

Paul David  <https://orcid.org/0000-0002-1513-0399>

## REFERENCES

- Arsigny, V., Fillard, P., Pennec, X., & Ayache, N. (2007). Geometric means in a novel vector space structure on symmetric positive-definite matrices. *SIAM Journal on Matrix Analysis and Applications*, 29(1), 328347.
- Bun, J., Bouchard, J.-P., & Potters, M. (2016). Cleaning correlation matrices. Risk: Risk Management, Derivatives, Regulation, <https://www.cfm.fr/assets/ResearchPapers/Cleaning-Correlation-Matrices.pdf>
- Bun, J., Bouchard, J.-P., & Potters, M. (2017). Cleaning large correlation matrices: Tools from random matrix theory. *Physics Reports*, 666, 1109.
- David, P., & Gu, W. (2019). A Riemannian quotient structure for correlation matrices. *Elemath Operators and Matrices*, 11(1), 607–627.
- Edelman, A., Arias, T. A., & Smith, S. T. (1998). The geometry of algorithms with orthogonality constraints. *SIAM Journal on Matrix Analysis and Applications*, 20(2), 303–353. <https://doi.org/10.1137/S0895479895290954>
- Fushiki, T. (2009). Estimation of positive semidefinite correlation matrices by using convex quadratic semidefinite programming. *Neural Computation*, 21(7), 20282048.
- Grubišić, I., & Pietersz, R. (2005). Efficient rank reduction of correlation matrices. SSRN Electronic Journal.
- Huckemann, S., Hotz, T., & Munk, A. (2010). Intrinsic shape analysis: Geodesic pca for Riemannian manifold modulo isometric Lie group actions. *Statistica Sinica*, <http://www3.stat.sinica.edu.tw/sstest/oldpdf/A20n11.pdf>
- Lee, J. M. (2012). *Introduction to smooth manifolds*: Springer.
- Lui, Y. M. (2012). Advances in matrix manifolds for computer vision. *Image and Vision Computing*, 30(6-7), 380388.
- Moakher, M., & Zéraï, M. (2010). The Riemannian geometry of the space of positive-definite matrices and its application to the regularization of positive-definite matrix-valued data. *Journal of Mathematical Imaging and Vision*, 40(2), 171187.
- Muniz, M., Ehrhardt, M., & Günther, M. (2021). Approximating correlation matrices using stochastic lie group methods. *Mathematics*, 9(1), 94. <https://www.mdpi.com/2227-7390/9/1/94>
- Pharasi, H. K., Sadhukhan, S., Majari, P., Chakraborti, A., & Seligman, T. H. (2021). Dynamics of the market states in the space of correlation matrices with applications to financial markets.
- Thanwerdas, Y., & Pennec, X. (2021). Geodesics of the Quotient Affine Metrics on Full-Rank Correlation Matrices. ArXiv.org.

**How to cite this article:** David, P., & Gu, W. (2022). Anomaly detection of time series correlations via a novel Lie group structure. *Stat*, 11(1), e494. <https://doi.org/10.1002/sta4.494>