Multi-Site, Multi-Pollutant Atmospheric Data Analysis using Riemannian Geometry

Alexander Smith[a], Jinxi Hua[b], Benjamin de Foy[d], James J. Schauer[a,c], and Victor M. Zavala[a*]

[a]Department of Chemical and Biological Engineering

University of Wisconsin, Madison, WI, USA

[b]School of Architecture

Taiyuan University of Technology, Taiyuan, China

[c]Department of Civil and Environmental Engineering

University of Wisconsin, Madison, WI, USA

[d]Department of Earth and Atmospheric Sciences

Saint Louis University, St. Louis, MO, USA

Abstract

We demonstrate the benefits of using Riemannian geometry in the analysis of multi-site, multi-pollutant atmospheric monitoring data. Our approach uses covariance matrices to encode spatio-temporal variability and correlations of multiple pollutants at different sites and times. A key property of covariance matrices is that they lie on a Riemannian manifold and one can exploit this property to facilitate dimensionality reduction, outlier detection, and spatial interpolation. Specifically, the transformation of data using Reimannian geometry provides a better data surface for interpolation and assessment of outliers compared to traditional data analysis tools that assume Euclidean geometry. We demonstrate the utility of using Riemannian geometry by analyzing a full year of atmospheric monitoring data collected from 34 monitoring stations in Beijing, China.

Keywords: Data science, Atmospheric monitoring, Multivariate analysis, Dimensionality reduction, Outlier detection, Spatial interpolation.

*Corresponding Author: Engineering Hall, 1415 Engineering Drive, Madison WI 53706 (victor.zavala@wisc.edu)

1 Introduction

Air pollution damages human health and impacts the environment (e.g., climate change) [13, 20, 25]. A key factor in developing pollution mitigation policies, technological solutions, and improving public awareness is the monitoring and modeling of atmospheric pollutant behavior [33]. Air pollution is traditionally measured within spatially-distributed monitoring stations. These stations provide accurate measurements of multiple atmospheric pollutants (e.g., O_3, NO_x, $PM_{2.5}$) at high temporal resolution. Historically, air pollution research and policy has focused on the control of individual pollutants due to the complexities that arise in the analysis, modeling, and interpretation of multi-pollutant data [7]. However, the need for air quality management tools and methods that integrate multi-pollutant data has been recognized by government agencies, such as the Environmental Protection Agency (EPA), as being crucial in estimating health risks and environmental impacts of complex mixtures of air pollutants [7, 40, 9, 19]. Furthermore, dynamic relationships between different pollutants encoded in multi-pollutant measurement data can provide insight into the chemical and physical interactions between pollutants. For example, chemical interactions between NO_x and O_3 can be captured by observing temporal correlations between their atmospheric concentrations. For instance, a positive correlation between NO_x and O_3 is commonly present due to the formation of O_3 through the photolysis of NO_2, whereas a negative correlation suggests the depression of O_3 concentration due to NO_x titration [26].

A common approach to quantify and study spatio-temporal relationships between pollutants consists of encoding the multivariate time series data as covariance (correlation) matrices [35, 15, 39, 17]. This approach enables the use of powerful multivariate methods (e.g., principal component analysis) for dimensionality reduction [35, 39], modeling [15, 36, 14], source detection [22], and monitoring station performance [17]. However, these methods make the blanket assumption that data lies in a Euclidean space and this assumption can miss geometric structure contained in the data that might be relevant for analysis [32, 30, 29]. Furthermore, many of these analysis techniques are based on single pollutant monitoring or on a single monitoring location; specifically, they do not account for the dynamic relationships between different pollutants across varying monitoring sites. Our proposed framework extends these analysis techniques for use in multi-pollutant, multi-site monitoring .

A key observation that we exploit in our analysis framework is that covariance matrices lie on a Riemannian manifold [28, 23, 34]. A Riemannian manifold represents a space that is governed by *non-Euclidean* geometry, meaning that the space is curved [28]. A simple example of a Riemannian manifold is the surface of the Earth (i.e., a smooth sphere). For instance, if we were to plan a trip from antipodal points on the Earth and we ignored the geometry of the Earth's surface we would end up passing through the center of the Earth. However, if we apply the concepts of Riemannian geometry and account for the *curvature* of the Earth we would construct a path that is constrained to Earth's surface. This same concept can be applied in the analysis of covariance matrices derived from atmospheric data. Accounting for the curvature of the Riemannian manifold formed by the data allows us to compute relationships that respect the data's high-dimensional structure, rather than (incorrectly) assuming the data is governed by Euclidean geometry [1, 34]. This has been shown to improve outcomes of multivariate data analysis methods in tasks such as classification and dimensionality reduction [28, 2, 23, 18]. Consideration for the geometry of these matrices also prevents results of data analysis (e.g., spatial interpolation) that could be physically inconsistent by constraining results to the Riemannian manifold [28, 23]. For example, spatial interpolation through Euclidean geometry can result in matrices that have no physical meaning and have inflated variance [34]. However, if Riemannian geometry is assumed, the interpolated result is guaranteed to be a proper covariance matrix by constraining the interpolation to the matrix manifold.

The analysis of data through Riemannian geometry has been well studied and successfully applied in a broad range of scientific fields such as statistics, physics, neuroscience, medical image analysis, engineering, and computer vision, but has been minimally explored in the analysis of atmospheric data [2, 23, 8, 10, 18]. These applications demonstrate that analysis tools that assume the data lies in Euclidean geometry can be drastically improved using Riemannian geometry. The assumption of Euclidean geometry can introduce analysis errors, physically inconsistent results, and apply constraints to the data that do not reflect the physics and dynamics of the data, all of which are addressed through a Riemannian geometric approach. Therefore, in this work we provide a practical introduction to the mathematics of the Riemannian geometry of covariance matrices and a framework for application in atmospheric data analysis. We also demonstrate the need for the incorporation of Riemannian geometric approaches to atmospheric data analysis as many of the

current tools that are based on the assumption of Euclidean geometry can artificially bias analysis results due to the nature of the constraints created by this assumption. We apply the introduced methods in an analysis of real, multi-pollutant data taken from 34 air quality monitoring sites in Beijing, China [12]. The data is made available through the Beijing Municipal Monitoring Center (bjmemc.com.cn). For each site we record hourly concentrations of six atmospheric pollutants: CO, NO_2, O_3, PM_{10}, $PM_{2.5}$ and SO_2. For each of the 34 sites we obtain a stochastic multivariate time series. We can compute the pairwise covariance between each of the time series and construct a covariance matrix for each site. Our analysis is focused on quantifying and understanding the spatial and temporal behavior of these covariance matrices through dimensionality reduction and spatial interpolation. We demonstrate the benefits of incorporating Riemannian geometry into the analysis through comparisons with methods that (incorrectly) assume these matrices live in Euclidean space. All code and data required to reproduce all of the results in this paper can be found in https://github.com/zavalab/ML/tree/master/Atmospheric_Analysis.

2 Methods

This section will cover necessary definitions and properties of the Riemannian manifold of covariance matrices that are symmetric and positive definite (SPD). A more detailed understanding of these methods and mathematical background can be found in our previous work [28].

We first provide context for the methods developed in this section. Figure 1 illustrates the data pre-processing steps that are taken in order to analyze the multi-site, multi-component dataset. Each measurement site, represented as a red point in the map of Beijing, produces a multivariate time series which measures the concentration of six air pollutants: CO, NO_2, O_3, PM_{10}, $PM_{2.5}$ and SO_2. These values are recorded in hour intervals for a period of one year. Taken together, this data forms a multivariate time series that is inherently stochastic. Thus, for a given interval of time, we can use measurements recorded at each site to construct a covariance matrix representing the temporal dynamics of six pollutants at each site. Each of the 6 measured variables are represented as a univariate random variable x_i where $i = 1, 2, ..., 6$. We denote the observations of each signal at time $t = 1, 2, ..., m$ as $x_i(t) \in \mathbb{R}^m$. We center each of the measured signals $\hat{x}_i := x_i - \frac{1}{m}\sum_{t=1}^{m} x_i(t)$ and denote the collection of centered signals as a multivariate random vector $\mathbf{X} = (\hat{x}_1, ..., \hat{x}_n)$ where $n = 6$.

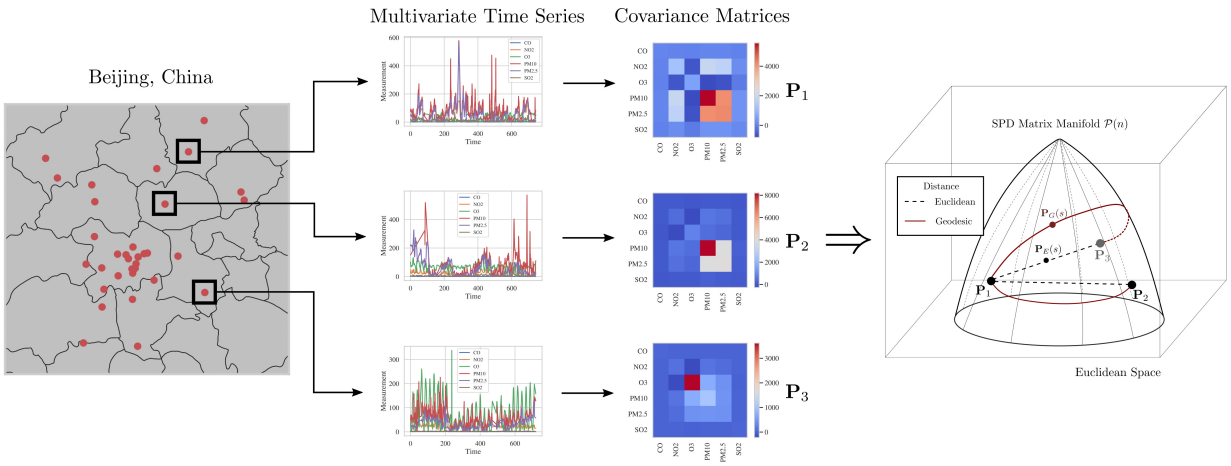


Figure 1: Illustration of the computational workflow used in the pre-processing and analysis of atmospheric data measured at multiple locations in Beijing, China. Multivariate measurements of six atmospheric pollutants: CO, NO_2, O_3,PM_{10}, $PM_{2.5}$ and SO_2 are taken over time at each monitoring site. These multivariate time series can be represented as covariance matrices, which are symmetric, positive definite (SPD). These matrices form a Riemannian manifold and this manifold can be leveraged to provide measures of similarity such as geodesic distances. Geodesic distances (red-solid lines) are constrained to the geometry of the overall data. Euclidean distances (dashed - black lines) do not capture this same information.

We then construct the sample covariance matrix for the multi-pollutant data $\mathbf{P} \in \mathcal{P}(n)$ as:

$$\mathbf{P} := \frac{1}{m-1}\mathbf{X}\mathbf{X}^T \tag{2.1}$$

Where $\mathcal{P}(n)$ represents the set of all $n \times n$ covariance matrices. Example covariance matrices are found in Figure 1. The covariance matrix $\mathbf{P}$ of dimension $n \times n$, where $n = 6$ in this example, quantifies both the variance of a given measured pollutant, and the covariance between each pollutant and all other pollutants. We can leverage these covariance matrices to quantify the dynamic relationships between pollutants at each site, and better understand how these dynamics change over time and space. The matrices $\mathbf{X}$ of size $n \times m$ contain the m time steps for n pollutants. The length of the time series m used in the matrices $\mathbf{X}$ and $\mathbf{P}$ is determined by the time period analyzed. It could be as short as 24 for the analysis of a single day but would more likely be a full month (720), year (8760) or more. An important aspect in the analysis of these covariance matrices is that they are *symmetric* and *positive definite* (SPD). A symmetric matrix is a matrix where the transpose of the matrix is exactly equal to the original matrix $\mathbf{P} = \mathbf{P}^T$. A positive definite matrix is a matrix where the eigenvalues of the matrix are all strictly greater than zero $\mathbf{P} \succ 0$. Matrices that meet these criterion have a special

geometric structure, known as a Riemannian manifold, which can be exploited to provide improvements in data centric tasks such as dimensionality reduction and spatial interpolation (e.g., Kriging).

The incorporation of Riemannian geometry into the analysis of covariance matrices requires few, computationally efficient, steps. These steps are outlined in this section, as well as section 2.1 - Tangent Spaces, and section 2.2 - Matrix Means. In summary, these steps allow us to map our covariance matrix data from the curved Riemannian manifold surface to a flat (Euclidean) vector space. This is similar to a map that represents a projection of the Earth's curved surface on a 2-dimensional plane. A 2-dimensional (flat) map of the earth allows us to easily compute angles and distances using common measurement devices (e.g., rulers, protractors). Similarly, mapping the curved Riemannian manifold (Earth's surface) onto a linear vector space (2-dimensional map) allows us to leverage common data analysis tools (e.g., PCA, Kriging).

We now focus on developing the mathematical notation and concepts needed to understand the Riemannian geometry of SPD matrices from a practical perspective. We denote the set of all $n \times n$ symmetric matrices as $\mathcal{S}(n) := \{\mathbf{S} \in \mathcal{M}(n), S^T = S\}$ where $\mathcal{M}(n)$ represents the space of all square $n \times n$ matrices. We then define the set of all $n \times n$ symmetric, positive definite matrices as $\mathcal{P}(n) := \{\mathbf{P} \in \mathcal{S}(n), u^T\mathbf{P}u > 0, \forall u \in \mathbb{R}^n\}$. The set $\mathcal{P}(n)$ represents our Riemannian manifold of covariance matrices. An illustrative representation of the $\mathcal{P}(n)$ manifold is found in Figure 2. The manifold $\mathcal{P}(n)$ is a curved, conic surface embedded in Euclidean space [38]. We also illustrate several points $\mathbf{P}_1, \mathbf{P}_2, \mathbf{P}_3 \in \mathcal{P}(n)$, that contain the covariance of the time series for each pollutant, for each site ($\mathbf{P}_1, \mathbf{P}_2, \mathbf{P}_3$ represent 3 of the 34 sites). Although the figure displays the SPD covariance matrices ($\mathbf{P}$) in three-dimensional space, they are actually in k-dimensional space where $k = n(n+1)/2$ and $n = 6$ is the number of pollutants in our measurement matrix $\mathbf{X}$ [23].

Figure 2 also demonstrates a simple distance analysis of these covariance matrices through two different methods: Euclidean and Geodesic. The Euclidean method assumes the matrices do not form a Riemannian manifold, and that they are governed by Euclidean geometry. We define the Euclidean distance between two matrices $\mathbf{P}_i, \mathbf{P}_j \in \mathcal{P}(n)$ as:

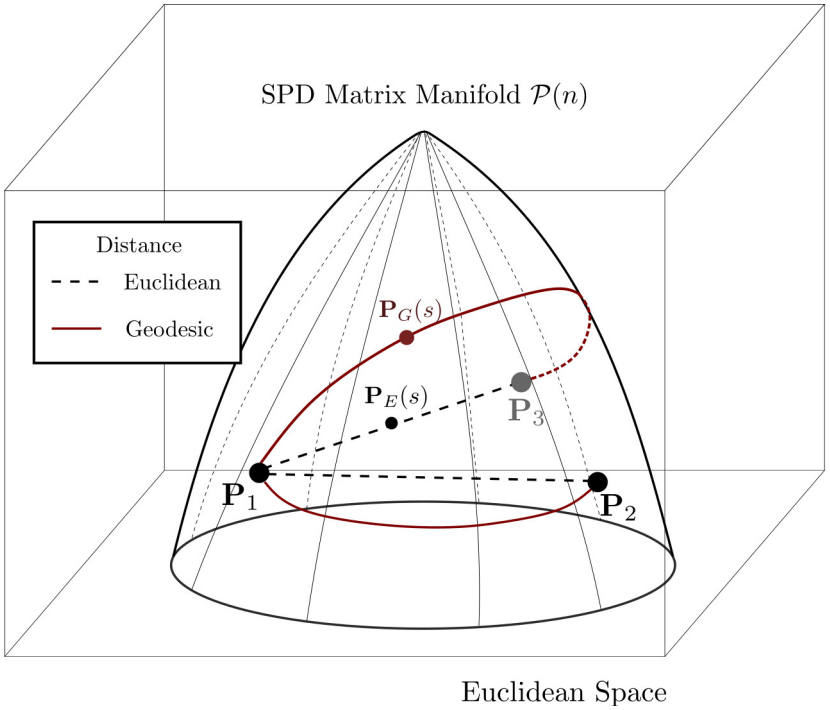


Figure 2: An illustration of a Riemannian manifold formed by SPD matrices $\mathcal{P}(n)$, with a set of points $\mathbf{P}_1, \mathbf{P}_2, \mathbf{P}_3 \in \mathcal{P}(n)$. A comparison of geodesic (red, solid line) and Euclidean distances (dashed, black line) are shown between $\mathbf{P}_1, \mathbf{P}_2$ and $\mathbf{P}_1, \mathbf{P}_3$ along with a comparison of a geodesic interpolation $\mathbf{P}(t)_G$ and Euclidean interpolation $\mathbf{P}(t)_E$ between points $\mathbf{P}_1, \mathbf{P}_3$. Geodesics and their distances/interpolations are constrained to the manifold $\mathcal{P}(n)$ and capture the inherent geometry of the data. Euclidean distances ignore the manifold geometry which can distort the perceived relationships between points on the manifold.

$$d_E(\mathbf{P}_i, \mathbf{P}_j) := ||\mathbf{P}_i - \mathbf{P}_j||_F \tag{2.2}$$

where $||\cdot||_F$ is the Frobenius norm. The geodesic method computes a *geodesic distance*. The geodesic distance is the length of a *geodesic* between two points on a Riemannian manifold [23]. A geodesic is the shortest path between two points on a Riemannian manifold that accounts for the curvature and shape of the manifold $\mathcal{P}(n)$ [28]. For SPD covariance matrices a geodesic distance can be defined as [28, 23]:

$$d_G(\mathbf{P}_i, \mathbf{P}_j) := ||\log(\mathbf{P}_i) - \log(\mathbf{P}_j)||_F \tag{2.3}$$

where $\log(\cdot)$ represents the matrix logarithm, commonly used in the analysis of matrix manifolds [28, 23]. In an analysis of our matrices $\mathbf{P}_1, \mathbf{P}_2, \mathbf{P}_3 \in \mathcal{P}(n)$ through the Euclidean distance we obtain the following result (illustrated in Figure 2):

$$d_E(\mathbf{P}_1, \mathbf{P}_2) > d_E(\mathbf{P}_1, \mathbf{P}_3) \tag{2.4}$$

Whereas if we perform the analysis using the geodesic distance we obtain the opposite result (illustrated in Figure 2):

$$d_G(\mathbf{P}_1, \mathbf{P}_3) > d_G(\mathbf{P}_1, \mathbf{P}_2) \tag{2.5}$$

The geodesic distance captures the correct relationship between these matrices because the distance accounts for the geometry of the manifold $\mathcal{P}(n)$, which the Euclidean distance ignores. We explore the practical importance of incorporating the manifold geometry through examples in dimensionality reduction and spatial interpolation in Section 3. Another property of geodesic distances that is helpful in the analysis of real world data is *congruence invariance*. Congruence invariance means that any $n \times n$ invertable matrix $\mathbf{X}$ applied to a set of covariance matrices $\mathbf{P}_i \in \mathcal{P}(n)$ does not impact

the geodesic distance between the two matrices:

$$d_g(\mathbf{X}^T\mathbf{P}_i\mathbf{X}, \mathbf{X}^T\mathbf{P}_j\mathbf{X}) = d_g(\mathbf{P}_i, \mathbf{P}_j) \tag{2.6}$$

Operations such as scaling and normalization, which can be reformulated as invertable matrices, have no impact on the geodesic distance between matrices. Thus, the geometry of the geodesic distance is robust to many issues that arise in real data sets (e.g., scaling of variables). This is further explored in Section 3.

As mentioned previously, the geodesic distance reflects the length of the shortest path between two points on a Riemannian manifold [28]. Geodesics are particularly useful in the interpolation and averaging of covariance matrices, which is often needed in algorithms for spatial interpolation (e.g. Kriging). An illustrative comparison of a geodesic versus a Euclidean method for interpolation is shown in Figure 2. Assuming Euclidean geometry, an interpolation is given by:

$$\mathbf{P}_E(s) := \mathbf{P}_1(1-s) + \mathbf{P}_3(s) \tag{2.7}$$

where $s \in [0,1]$ and $\mathbf{P}_E(0) = \mathbf{P}_1$, $\mathbf{P}_E(1) = \mathbf{P}_3$. Unfortunately, as seen in Figure 2, this interpolation does not follow a geodesic path. This can result in covariance matrices that have inflated variance and other potential issues [1]. However, if we consider the geometry of the manifold we can construct an interpolation via the geodesic [28, 23]:

$$\mathbf{P}_G(s) := \mathbf{P}_1^{1/2}\left(\mathbf{P}_1^{-1/2}\mathbf{P}_3\mathbf{P}_1^{-1/2}\right)^s\mathbf{P}_1^{1/2} \tag{2.8}$$

where $s \in [0,1]$, $\mathbf{P}_G(0) = \mathbf{P}_1$, $\mathbf{P}_G(1) = \mathbf{P}_3$, and $\mathbf{P}_i = \mathbf{P}_i^{1/2}\mathbf{P}_i^{1/2}$. Further details around the derivation of the geodesic can be found in our previous work [28]. Leveraging the geodesic between matrices $\mathbf{P}_1$ and $\mathbf{P}_3$ ensures that the interpolated matrix $\mathbf{P}_G(s) \in \mathcal{P}(n)$ is a symmetric, positive definite covariance matrix [34, 24]. These interpolation methods are the basis for constructing means (e.g.,

mid-point between multiple matrices). They are also used in spatial interpolation where the behavior of the pollutants (i.e. the covariance matrix) at a new spatial location is estimated by a weighted average of the covariance matrices at neighboring locations. In contrast with a linear interpolation of the measurements themselves, a geodesic interpolation would seek to follow the underlying dynamics of the system. Thus, developing averages or spatially interpolating values is highly dependent on the data geometry and is demonstrated through case studies in section 3. We note that the geodesic distance selected here, based on the matrix logarithm, is one of many approaches to constructing geodesics on the manifold of SPD matrices. These methods include decompositions of SPD matrices, such as the Cholesky decomposition or the matrix square root, and geometric methods such as Procrustes analysis; a detailed overview of these methods can be found in the following reference [8]. The logarithm based metric was chosen for the analysis of atmospheric data due to the ability of the metric to combat potentially adverse effects in data analysis such as swelling (described in the supplementary information) and its ease of interpretability [23].

2.1 Tangent Spaces

The analysis of covariance matrices through Riemannian manifolds provides numerous benefits, however many multivariate methods for tasks such as dimensionality reduction (e.g., principal component analysis) require data that lies in a (linear) vector space. The non-linear (curved) nature of the Riemannian manifold of SPD matrices requires that we map the structure of the manifold to a linear space. Fortunately, Riemannian manifolds are *differentiable* manifolds [16]. This means that every point on the Riemannian manifold $\mathbf{P} \in \mathcal{P}(n)$ has an associated tangent space denoted as $T_{\mathbf{P}}P \in \mathcal{S}(n)$, constructed from the tangent vectors of all possible *curves* passing through the point on the manifold $\mathbf{P} \in \mathcal{P}(n)$ [4, 28, 16]. We define a curve as a continuous function $\phi : [0, 1] \rightarrow \mathcal{P}(n)$. The tangent space at any point $\mathbf{P} \in \mathcal{P}(n)$ represents a (linear) vector space that is of the same dimension as the Riemannian manifold [3]. The tangent space encodes the geometry of the Riemannian manifold and can be used directly in common multivariate analysis methods.

An illustration of the tangent space $\mathcal{T}_{\mathbf{P}_1}\mathcal{P}$ constructed at a point $\mathbf{P}_1 \in \mathcal{P}(n)$ is found in Figure 3a. Figure 3a also illustrates two functions that connect the tangent space $\mathcal{T}_{\mathbf{P}_1}\mathcal{P}$ to the Riemannian manifold. These functions are known as the *logarithmic* map and the *exponential* map. For example,

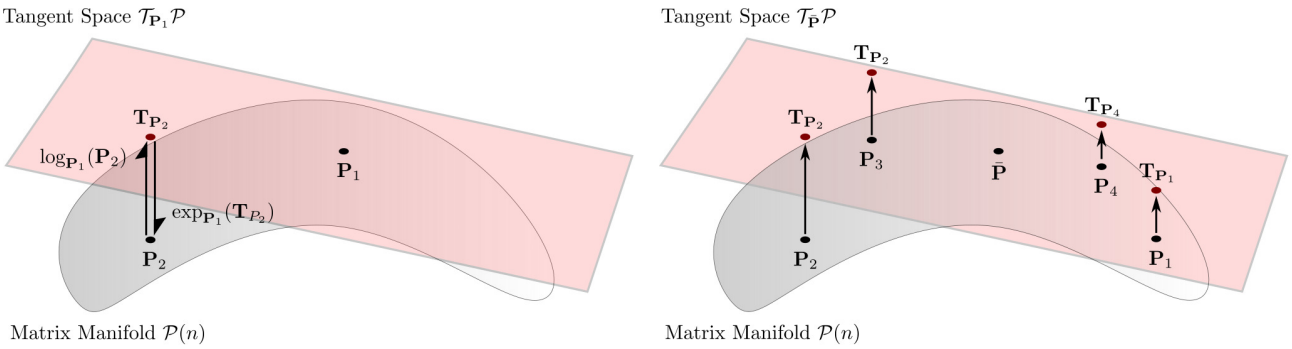


Figure 3: (a) Illustration of the tangent space $\mathcal{T}_{\mathbf{P}_1}\mathcal{P}$ formed at the point $\mathbf{P}_1 \in \mathcal{P}(n)$. The tangent space represents a (linear) vector space that is of the same dimension as the manifold $\mathcal{P}(n)$ and intersects the manifold at $\mathbf{P}_1$. The logarithmic map $\log_{\mathbf{P}_1}(\mathbf{P}_2) \rightarrow \mathbf{T}_{\mathbf{P}_2}$ takes point $\mathbf{P}_2 \in \mathcal{P}(n)$ to the point $\mathbf{T}_{\mathbf{P}_2} \in \mathcal{T}_{\mathbf{P}_1}\mathcal{P}$. The exponential map $\exp_{\mathbf{P}_1}(\mathbf{T}_{\mathbf{P}_2}) \rightarrow \mathbf{P}_2$ does the inverse. The exponential and logarithmic mappings allow us to map our covariance matrix data from the curved manifold space $\mathcal{P}(n)$ to a vector space $\mathcal{T}_{\mathbf{P}_1}\mathcal{P}$. (b) Illustration of the geometric mean of a set of covariance matrices $\mathbf{P}_1, \mathbf{P}_2, \mathbf{P}_3, \mathbf{P}_4 \in \mathcal{P}(n)$. The geometric mean is the point $\bar{\mathbf{P}}$ that minimizes the geodesic distance to all other data points on the manifold $\mathcal{P}(n)$. The logarithmic map $\log_{\bar{\mathbf{P}}}(\cdot) : \mathcal{P}(n) \rightarrow \mathcal{T}_{\bar{\mathbf{P}}}\mathcal{P}$ is used to map the covariance matrices from the curved manifold $\mathcal{P}(n)$ to the linear tangent space $\mathcal{T}_{\bar{\mathbf{P}}}\mathcal{P}$ while minimizing the potential geometric distortion from the mapping. This mapped data can then be used directly in common data analysis methods such as principal component analysis.

in Figure 3a the logarithmic map takes $\mathbf{P}_2$ on the Riemannian manifold and maps it to a point $\mathbf{T}_{\mathbf{P}_2}$ in the tangent space $\mathcal{T}_{\mathbf{P}_1}\mathcal{P}$ centered at $\mathbf{P}_1$. For two points $\mathbf{P}_i, \mathbf{P}_j \in \mathcal{P}(n)$ we define the logarithmic map as [28, 4]:

$$\log_{\mathbf{P}_i}(\mathbf{P}_j) := \mathbf{P}_i^{1/2} \log\left(\mathbf{P}_i^{-1/2}\mathbf{P}_j\mathbf{P}_i^{-1/2}\right)\mathbf{P}_i^{1/2} \tag{2.9}$$

The exponential map is the inverse of the logarithmic map. It maps points from a tangent space ($\mathbf{T}_{\mathbf{P}_2}$ in Figure 3a) back to the Riemannian manifold ($\mathbf{P}_2$ in Figure 3a). For points $\mathbf{P}_i \in \mathcal{P}(n)$ and $\mathbf{T}_{\mathbf{P}_j} \in \mathcal{T}_{\mathbf{P}_i}\mathcal{P}$, we define the exponential map as:

$$\exp_{\mathbf{P}_i}(\mathbf{T}_{\mathbf{P}_j}) := \mathbf{P}_i^{1/2} \exp\left(\mathbf{P}_i^{-1/2}\mathbf{T}_{\mathbf{P}_j}\mathbf{P}_i^{-1/2}\right)\mathbf{P}_i^{1/2} \tag{2.10}$$

The exponential and logarithmic map allow for the curved Riemannian manifold of SPD covariance matrices to be mapped directly to a (linear) vector space (i.e., the tangent space). Mapping the

data to the linear (flat) vector space allows us to apply common data analysis techniques while encoding the manifold geometry. Thus, linear methods such as PCA and Kriging applied to this mapped data will automatically incorporate the geometry of the Riemannian manifold. Benefits of this are further explored in Section 3. For those readers interested, a detailed derivation of the logarithmic and exponential mapping can be found in the following references [28, 4].

2.2 Matrix Means

In the analysis of real datasets where there are many matrices being analyzed it is not obvious which point on the manifold should be chosen as the basis for the tangent space. The basis point is often chosen to be the *geometric mean* of the set of all matrices in the dataset $\mathbf{P}_i \in \mathcal{P}(n)$. The geometric mean is the point $\bar{\mathbf{P}} \in \mathcal{P}(n)$ that minimizes the geodesic distance to all other data points on the Riemannian manifold [28, 23, 3]:

$$\bar{\mathbf{P}} := \underset{\mathbf{A}}{\operatorname{argmin}} \sum_{i=1}^{n} ||\log(\mathbf{A}) - \log(\mathbf{P}_i)||_F \tag{2.11}$$

Once the geometric mean is identified, the set of matrices in the dataset can then be mapped from the Riemannian manifold to the tangent space through the logarithmic map. An illustration of this process is found in Figure 3b. Here, the geometric mean $\bar{\mathbf{P}}$ of matrices $\mathbf{P}_1, \mathbf{P}_2, \mathbf{P}_3, \mathbf{P}_4 \in \mathcal{P}(n)$ is computed. The matrices are then mapped to the tangent space through the logarithmic map $\log_{\bar{\mathbf{P}}}(\mathbf{P}_i) = \mathbf{T}_{\mathbf{P}_i}$. Once completed, the matrices have been mapped to a linear space and can be used in multivariate analysis methods while minimizing the potential geometric distortion of the logarithmic mapping [28, 3].

2.3 Spatial Interpolation

Consideration for the Riemannian geometry of covariance matrices is crucial when performing spatial interpolation of covariance matrix values [24]. As shown in Figure 2, interpolation between matrices with an assumption of Euclidean geometry (Equation 2.7) can result in matrices that do not lie on the covariance matrix manifold $\mathcal{P}(n)$. Whereas interpolation through geodesics (Equation 2.8) provides assurance that the resulting interpolated matrix will lie on the Riemannian manifold [24].

Euclidean interpolation of covariance matrices can also cause a *swelling effect* on the interpolated matrices [1, 28]. The swelling effect causes an increase in the generalized variance (i.e., determinant) of the interpolated covariance matrices. This introduces a spurious increase in the variance of the atmospheric pollutant data dynamics creating results that are not physically consistent. An example of this effect is shown in the supplementary information. This is not the case with geodesic interpolation which reflects a natural evolution of the generalized variance during interpolation. This problem is amplified when attempting to perform a spatial interpolation where the matrices at multiple spatial locations are averaged (in a weighted manner) to predict the matrix values at a new (unmeasured) spatial location. This will result in spurious increases in (co)variance that would lead to false conclusions about potential sources/dynamics of pollution at these unmeasured locations.

Another aspect of spatial interpolation methods is the modeling of spatial dependence (spatial autocorrelation) between observed data points. This is often modeled through the use of the empirical variogram [5]. Given a set of $i \in \mathbb{Z}_+$ sample covariance matrices $\mathbf{P}_i \in \mathcal{P}(n)$ measured at i spatial locations $s_i \in \mathbb{R}^2$ we compute a Euclidean or geodesic distance between each sample covariance matrix. We also compute the spatial lag distance between each location $||s_i - s_j||_2 \in \mathbb{R}$, where the s_i, s_j are latitude and longitude coordinates. The covariance matrix distances are binned over spatial lag distance and averaged within each bin. This is a simple method for understanding spatial autocorrelation between the covariance matrices. For the geodesic distance method we compute the empirical variogram as:

$$\hat{\gamma}(h \pm \delta)_G := \frac{1}{2|N(h \pm \delta)|} \sum_{(s_i, s_j) \in N(h \pm \delta)} ||\log(\mathbf{P}_i) - \log(\mathbf{P}_j)||_F \tag{2.12}$$

where $h, \delta \in \mathbb{R}$ represents the spatial lag bin center and width, $N(h \pm \delta) := \{(s_i, s_j) : ||s_i - s_j||_2 \in h \pm \delta\}$ which represents pairs of sites (s_i, s_j) with spatial distance $h - \delta \leq ||s_i - s_j||_2 \leq h + \delta$, and $|N(h \pm \delta)| \in \mathbb{Z}$ represents number of sites pairs contained in $N(h \pm \delta)$. For the Euclidean distance method the empirical variogram is given as:

$$\hat{\gamma}(h \pm \delta)_E := \frac{1}{2|N(h \pm \delta)|} \sum_{(s_i, s_j) \in N(h \pm \delta)} ||\mathbf{P}_i - \mathbf{P}_j||_F \tag{2.13}$$

We compare the performance of variograms constructed using a geodesic distance and a Euclidean distance on the Beijing atmospheric data in Section 3.

3 Application to Multi-Site, Multi-Pollutant Data

This section focuses on the application of Riemannian geometry in the analysis of multi-site, multi-pollutant time series data collected from 34 sites in Beijing, China [12]. The data is collected at 1-hour intervals and spans the entire calendar year of 2019. We apply and compare dimensionality reduction and spatial interpolation methods using Riemannian geometry to those that assume that data lies in Euclidean space.

3.1 Dimensionality Reduction

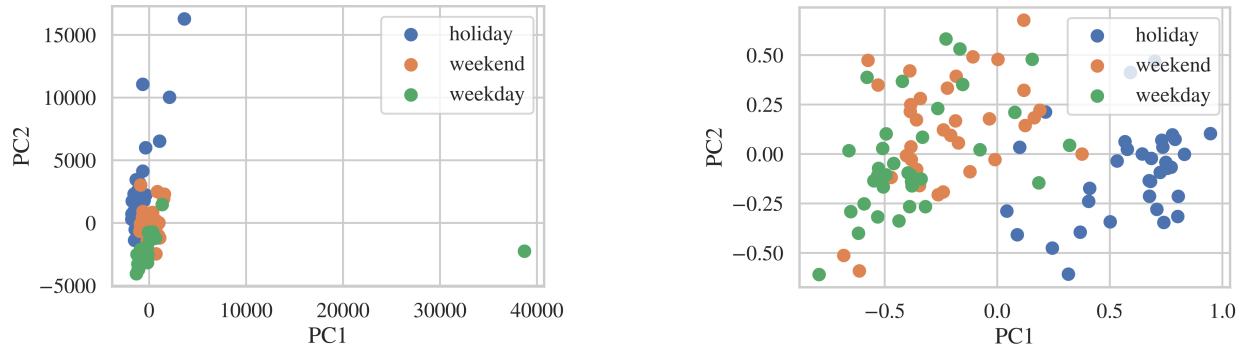
Dimensionality reduction helps facilitate both visualization of data and subsequent processing of data for other data-centric tasks such as classification, regression, and outlier detection [37, 30, 32, 27, 31]. A common method in dimensionality reduction for atmospheric data is principal component analysis (PCA), which will be the focus of this section [35, 39, 17]. Our analysis is based on covariance matrices constructed from multi-pollutant time series data at each of the 34 sites in Beijing. As discussed previously, each site measures 6 pollutants: CO, NO_2, O_3, PM_{10}, $PM_{2.5}$ and SO_2. We represent each of the 6 measured variables as a centered, univariate random variable x_i where $i = 1, 2, ..., 6$ and we denote the collection of signals as a multivariate random vector $\mathbf{X} = (x_1, ..., x_n)$ where $n = 6$. Thus, we obtain multivariate random vectors $\mathbf{X}_j$ for each site $j = 1, 2, ..., 34$. We first analyze the multi-pollutant data collected during weekdays, weekends, and holidays for each of the 34 sites. For each site, we obtain 3 covariance matrices, one representing dynamics observed during the weekday, one during the weekend, and a final during holidays for the year. We denote the observations of each signal at time $t = 1, 2, ..., m$ as $x_i(t) \in \mathbb{R}^m$. The value of m will change for weekdays $m = 6936$ (hours), weekends $m = 1128$ (hours), and holidays $m = 696$ (hours). The covariance matrix size is only dependent on the number of measured variables, which in this case is the 6 pollutants. Thus, we

can directly compare these three timeframes even though they have a varying number of time samples. We use these measurements to construct the sample covariance matrix for the multi-pollutant data $\mathbf{P}_j \in \mathcal{P}(n)$ for each site $j = 1, 2, ..., 34$ as:

$$\mathbf{P}_j := \frac{1}{m-1}\mathbf{X}_j\mathbf{X}_j^T \tag{3.14}$$

We first perform PCA without consideration for the Riemannian geometry of our covariance matrices (assuming a Euclidean space). Here, each covariance matrix $\mathbf{P}_j \in \mathcal{P}(n)$ can be can be vectorized $P_j := \text{vec}(\mathbf{P}_j) \in \mathbb{R}^{n^2}$, where $n = 6$. We can then construct a matrix $\mathbf{M} = \left[P_1^T, P_2^T, ..., P_l^T\right]^T \in \mathbb{R}^{l\times 36}$, where $l = 34 * 3$. The matrix $\mathbf{M}$ contains the 3 covariance matrices for each of the 34 sites. Each row in the matrix $\mathbf{M}$ has the 36 values of the 6×6 covariance matrix $\mathbf{P}_j$, and there is a row for each site and for each type of day leading to 34×3 rows. We then perform a singular value decomposition of the matrix $\mathbf{M}$ and project the data onto the leading eigenvectors. The eigenvectors represent basis directions in Euclidean space that capture the most variance (information) in the data. We can project our data onto these eigenvector basis so that we can view a low-dimensional representation of our data that captures the most information. This is the basis of Principal Component Analysis (PCA). The results of this analysis for the first two leading eigenvectors (directions) is found in Figure 4. In this Euclidean approach we perform the same type of analysis but do not project the covariance matrices to the tangent space at the geometric mean. Thus, we are analyzing the covariance matrices without any pre-processing. The results of this Euclidean based approach are shown in Figure 4.

We can then perform the proposed Riemannian geometric method for the data. This method is often referred to as Principal Geodesic Analysis (PGA) and was first introduced in the areas of shape analysis and medical image analysis [11, 10] . We note that since the introduction of these methods, there has been a deep exploration of dimensionality reduction methods applied to manifold data, but this method is chosen for its simplicity and interpretability [18] . To perform our analysis, we begin by first identifying the geometric mean of the matrices $\bar{\mathbf{P}}$, and then mapping the data from the Riemannian manifold $\mathcal{P}(n)$ to the tangent space at the geometric mean $\mathcal{T}_{\bar{\mathbf{P}}}\mathcal{P}$ through the logarithmic map $\log_{\bar{\mathbf{P}}}(\mathbf{P}_j) \rightarrow \mathbf{T}_{\mathbf{P}_j}$, where $\mathbf{T}_{\mathbf{P}_j}$ represents our data points that are mapped to the linear (flat) vector space. With the matrices now in a (linear) vector space they can be analyzed



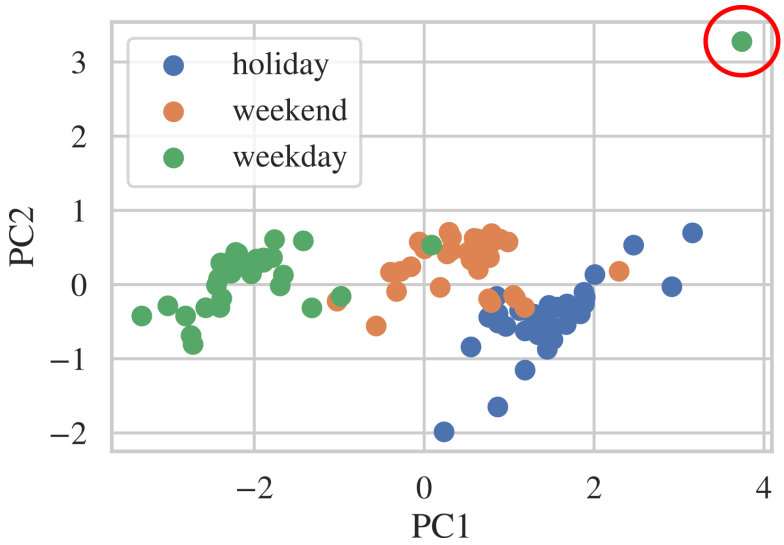
(a) Euclidean PCA

(b) Normalized Euclidean PCA

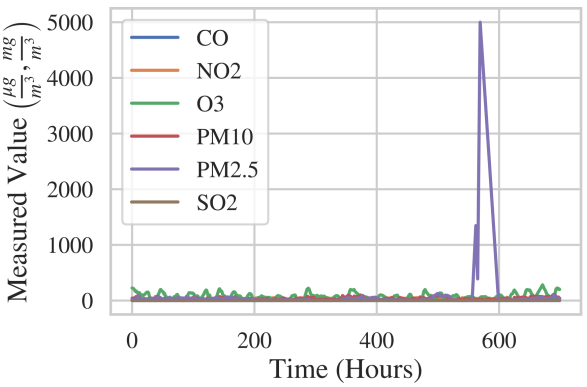
Figure 4: Dimensionality reduction for the multi-pollutant covariance matrices derived from holiday, weekend, and weekday data assuming Euclidean geometry. (a) Principal component analysis where the covariance matrices have been assumed to be governed by Euclidean geometry. The Euclidean PCA method is dominated by outliers, is impacted by the scale of the variables, and does not show clear separation between the three sample groups. (b) An attempt to normalize the data results in a structure that no longer captures outliers and also no longer shows clear clustering of the data.

through principal component analysis (PCA). Here, each mapped covariance matrix $\mathbf{T}_{\mathbf{P}_j} \in \mathcal{T}_{\bar{\mathbf{P}}}\mathcal{P}$ can be can be vectorized $T_{P_j} := \text{vec}(\mathbf{T}_{\mathbf{P}_j}) \in \mathbb{R}^{n^2}$, where $n = 6$. We can then construct a matrix $\mathbf{M}_{\mathcal{T}} = \left[T_{P_1}^T, T_{P_2}^T, ..., T_{P_l}^T\right]^T \in \mathbb{R}^{l \times 36}$, where $l = 34 * 3$. The matrix $\mathbf{M}_{\mathcal{T}}$ contains the 3 transformed covariance matrices for each of the 34 sites. We then perform a singular value decomposition of the matrix $\mathbf{M}_{\mathcal{T}}$ and project the data onto the first two leading eigenvectors. Here, the eigenvectors now represent geodesic directions on the Riemannian SPD manifold. The results of this analysis for the first two leading eigenvectors (directions) is found in Figure 5. We compare these results to the previous analysis that assumes the covariance matrices are governed by Euclidean, rather than Riemannian, geometry (Figure 4).

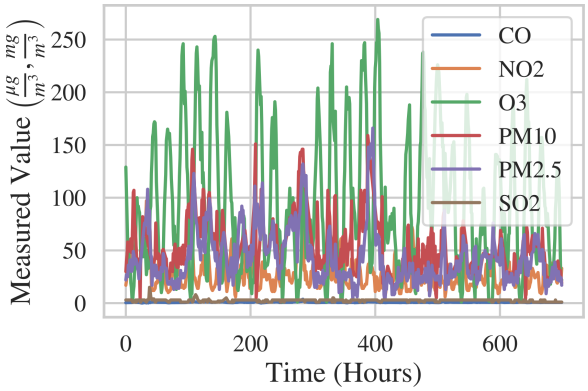
The output of PCA that considers the Riemannian geometry of the covariance matrices provides a much clearer result that demonstrates clustering of the behavior of the sites into weekday, weekend, and holiday groupings (Figure 5a). The Euclidean method does not capture this same information and is distorted (Figure 4a). One of the reasons for this distortion is the need for normalization or scaling of the data prior to PCA. This is not a challenge for the Riemannian approach due to the previously described property of *congruence invariance*. Operations such as re-scaling and normalization, which can be represented algebraically as invertable square matrices, have no impact on the geodesic distance between the matrices [2]. Therefore, there is no need to select specific scaling or normalization strategies when applying our geometry based analysis of the data, it is done naturally through



(a) Principal Geodesic Analysis



(b) Outlier Data

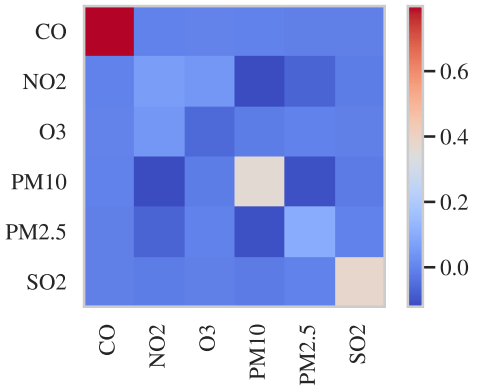


(c) Normal Data

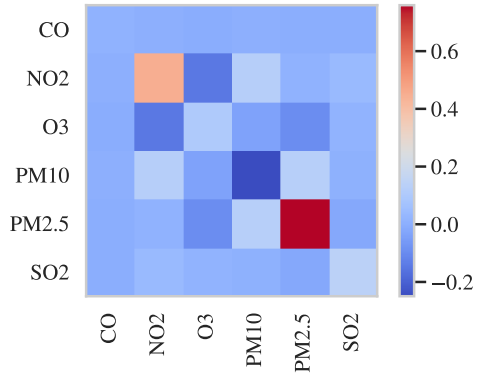
Figure 5: Dimensionality reduction for the multi-pollutant covariance matrices derived from holiday, weekend, and weekday data across the 34 sites in Beijing, China. (a) Principal component analysis of covariance matrices that have been mapped to the tangent space at the geometric mean $\mathcal{T}_{\bar{\mathbf{P}}}\mathcal{P}$. We observe clustering of the data into holiday, weekend, and weekday groups. We also note an outlier in the data, circled in red. (b,c) A comparison of the outlier point time series data (b) versus data taken from the same point hours earlier (c). The outlier point shows an abnormal spike in $PM_{2.5}$ when compared to normal measurements shown in (c). We note that CO measurements are in mg/m3.

the geometry of the manifold. However, this is not true when ignoring the data geometry, making Euclidean methods susceptible to the chosen framework (or lack of) for normalization/scaling (Figure 4). We can attempt to account for the distortion of the data in the Euclidean dimensionality reduction method by normalizing the data (e.g., constructing correlation matrices). The results of Euclidean PCA using normalized data is found in Figure 4b. We see that the distortion due to outliers/scaling has been eliminated, but there is minimal separation of the data into holiday, weekend, and weekday groups. We also see that a true outlier, identified in Figure 5, is no longer clearly identified. The inherent robustness of the Riemannian geometric methods (e.g., the property of congruence invariance) allows us to identify true outliers in the data while also capturing the clear clustering and structure of the data.

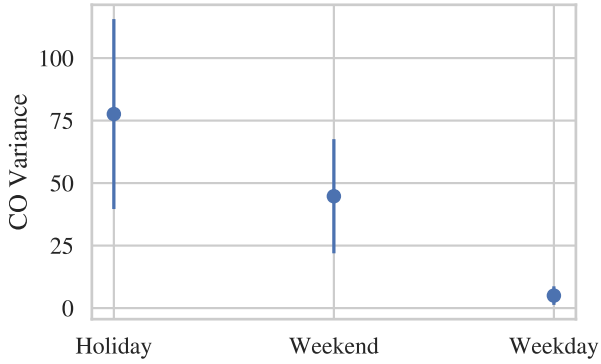
We can further explore the results of this dimensionality reduction through an analysis of the leading principal component eigenvectors. The principal component eigenvectors for the first PC and second PC are found in Figure 6. The eigenvectors represent basis directions in the Riemannian SPD manifold that capture the most variance (information) in the data. These directions can help us understand what variance and covariance values are the main causes of clustering for holidays, weekends, and weekdays and what causes their differences. For example, we see that holiday, weekend, and weekday behavior are stratified along the first principal component. The eigenvector (direction) associated with the first PC has a large positive weighting on the variance of CO (Figure 6a). This suggests that holidays and weekends result in a higher level of variance in CO when compared with weekdays, with holidays causing the highest increase. We can confirm this result by plotting the variance in CO measured during the holidays, weekends, and weekdays, which we show in Figure 6c. The high variance of CO during holidays is probably due to the large difference in CO emissions between the heating season and the non heating season. The CO emission in non heating seasons is small, regardless of holidays. During the heating season, there is a significant increase of indoor heating during holidays. This causes higher CO emissions than non holidays, as shown in previous work by Hua and co-workers [12]. We can also explore the separation of the data along the second eigenvector (direction). The second eigenvector is shown in Figure 6b. We can observe that there is a large negative weighting on the dynamical interactions (covariance) between NO_2 and O_3, suggesting a difference between weekdays, weekends and holidays based upon the dynamics of



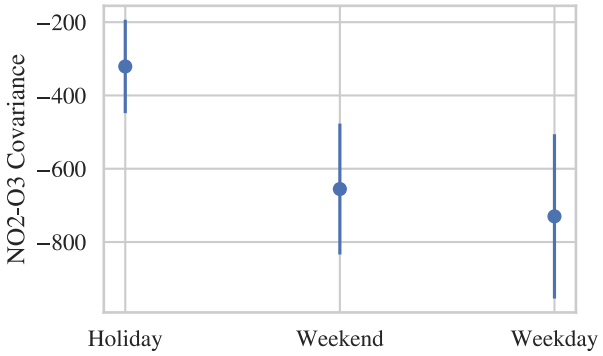
(a) Principal Component 1 Eigenvector



(b) Principal Component 2 Eigenvector



(c) CO Variance Analysis



(d) NO_2 - O_3 Covariance Analysis

Figure 6: The first two leading eigenvectors (basis directions) from the geodesic PCA method. The eigenvectors help understand the clustering and stratification of the data into weekday, weekend, and holiday clusters (Figure 5a). (a) The first eigenvector (basis direction) accounts for separation on the x-axis of Figure 5a. (b) The second eigenvector accounts for separation on the y-axis of Figure 5a. (c) Quantification of changes in CO variance during holiday, weekends, and weekdays. Each point represents the mean variance value, with the error bars representing a standard deviation from the mean. CO variance was identified as a factor that changes significantly between holidays, weekends, and weekdays through the first principal component eigenvector. (d) Quantification of changes in NO_2 - O_3 covariance during holidays, weekends, and weekdays. Each point represents the mean covariance value, with the error bars representing a standard deviation from the mean. Changes in NO_2 - O_3 covariance were identified as changing significantly between holidays, weekends, and weekdays through the second principal component eigenvector.

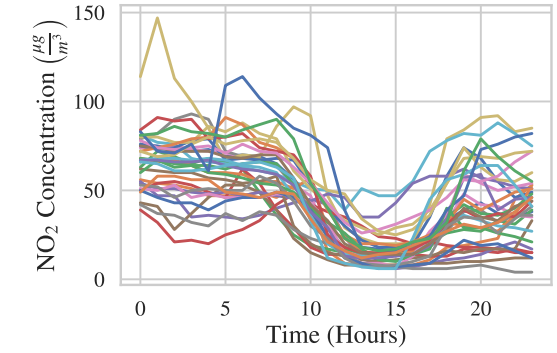
these two pollutants. Figure 6d shows the changes in the NO_2 and O_3 covariance during weekdays, weekends, and holidays, here we see a much larger inverse correlation between NO_2 and O_3 during weekdays and weekends in comparison with holidays. These analysis can help identify differences in the temporal dynamics of multi-component, multi-site air pollutant data.

We illustrate another application of dimensionality reduction in the analysis of multi-site, rather than multi-pollutant, dynamics. Here, we measure the dynamics of NO_2 at each site over the entire year. For a given site we obtain a centered, univariate stochastic time series $x_i \in \mathbb{R}^m$, where $i = 1, 2, ..., 34$ representing the 34 sites, and $m = 24 * 365$ because NO_2 is measured in hourly intervals over an entire year. For each site i we split the time series data into subsets $x_i^h \in \mathbb{R}^p$ where $p = 365$ and $h = 1, 2, ..., 24$ representing the hours of the day. We take each subset and form a multi-variate time series matrix $\mathbf{X} = [x_1^h, x_2^h, ..., x_{34}^h]$ for each hour of the day $h = 1, 2, ..., 24$ (Figure 7a). We can then construct a covariance matrix from this data as:

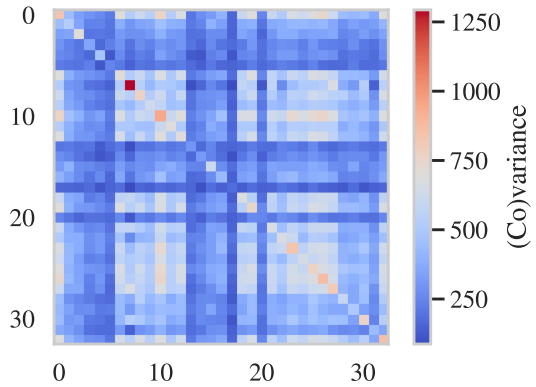
$$\mathbf{P}^h := \frac{1}{365 - 1}\mathbf{X}\mathbf{X}^T. \tag{3.15}$$

which results in a total of 24 covariance matrices of shape $\mathbf{P}_i^h \in \mathbb{R}^{n^2}$, where $n = 34$ represents the 34 measurement locations (Figure 7b). Thus, in this analysis we are focused on the dynamics and relationships between multiple sites, rather than multiple pollutants. To summarize, we now have a total of 24 covariance matrices representing each hour of the day. These covariance matrices are of size 34×34, representing the measurements of NO_2 at the 34 different sites.

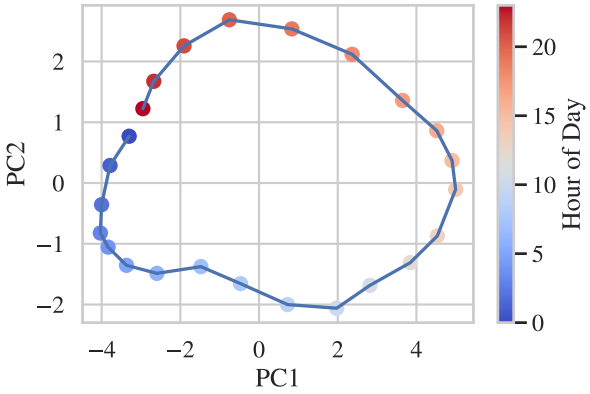
We then follow the same procedure as the previous example and perform dimensionality reduction using the PGA method. The results of the Riemannian geometric analysis are found in Figure 7. Using the Riemannian geometric approach, we see a data structure that indicates a clear cyclic behavior in NO_2 dynamics throughout the day. We can also understand what sites are impacting the dynamics of NO_2 by observing the values of the leading eigenvector associated with the first principal component (e.g., X-axis in Figure 7c). We visualize this in Figure 7d where we color each of the 34 sites with the coefficients of the leading eigenvector associated with the variance of each site (i.e., the values on the diagonal of the covariance matrix). From this analysis we can see which sites have a larger influence on the behavior of NO_2 dynamics during the day (positive coefficients -



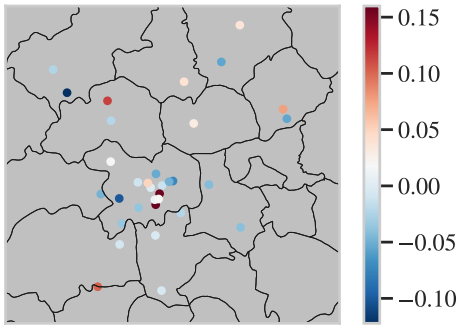
(a) NO_2 Measurements



(b) NO_2 Covariance matrix



(c) Geodesic PCA



(d) Coefficients of the leading eigenvector

Figure 7: (a) Multivariate time series constructed from the measurements of NO_2 concentration across all 34 Beijing, China sites during a period of 24 hours. (b) A 34×34 covariance matrix constructed from the multivariate time series in (a). The off-diagonal elements represent dynamic relationships in NO_2 between the different sites, and the diagonal values represent the variance of NO_2 at each individual site. (c) Illustration of the clearly cyclic behavior of NO_2 dynamics observed during each hour of the day across the different Beijing sites. Here, each point represents a projection one of the 24 covariance matrices onto the leading eigenvectors. (d) An analysis of the coefficients of the leading eigenvector of PCA. We find that certain sites have a large influence on the behavior of NO_2 dynamics during the day (positive coefficients - red color) and sites that have a larger influence during the night (negative coefficients - blue color)

red color) and those that have a larger influence during the night (negative coefficients - blue color). Interestingly, we see that some sites that are spatially close are dominated by NO_2 dynamics during different times of the day. This is most likely due to the activities associated with that particular area such as business/tourism (daytime dynamics) versus residential (nighttime dynamics).

3.2 Spatial Interpolation

As mentioned previously, a critical aspect of spatial interpolation methods is the modeling of spatial dependence. This is often modeled through the use of the empirical variogram [5]. We compare the performance of variograms constructed using a geodesic distance and a Euclidean distance defined in Section 2 by observing the spatial correlation between covariance matrices derived from each site during weekdays. We again represent each of the 6 measured variables as a univariate random variable x_j where $j = 1, 2, ..., 6$ and we denote the collection of signals as a multivariate random vector $\mathbf{X} = (x_1, ..., x_n)$ where $n = 6$. We denote the observations of each signal at time $t = 1, 2, .., m$ as $x_j(t) \in \mathbb{R}^m$. We use this representation to construct the sample covariance matrix for the multi-pollutant data $\mathbf{P}_i \in \mathcal{P}(n)$ at each site i. In this example we analyze measurements on weekdays during the hours of 8:00 AM to 3:00 PM at each site. Thus, we obtain one covariance matrix $\mathbf{P}_i$ for each site $i = 1, 2, ..., 34$. We can now compute the empirical geodesic variogram $\hat{\gamma}(h \pm \delta)_G$ and the Euclidean variogram $\hat{\gamma}(h \pm \delta)_E$ for this set of covariance matrices. These two variograms are found in Figure 8.

The variogram constructed with the geodesic distance reveals a clear relationship between the covariance matrix values and the spatial distance between each site, showing a structure that could be fit with a known variogram model (e.g., Gaussian). The variogram constructed with a Euclidean distance does not show a clear behavior that fits known variogram models [6]. We note that for the Euclidean comparison we normalize the data and use the correlation, rather than covariance, matrix to eliminate large effects from outliers. The incorporation of the Riemannian manifold geometry can reveal spatial relationships between the covariance matrices that are missed when Euclidean geometry is assumed.

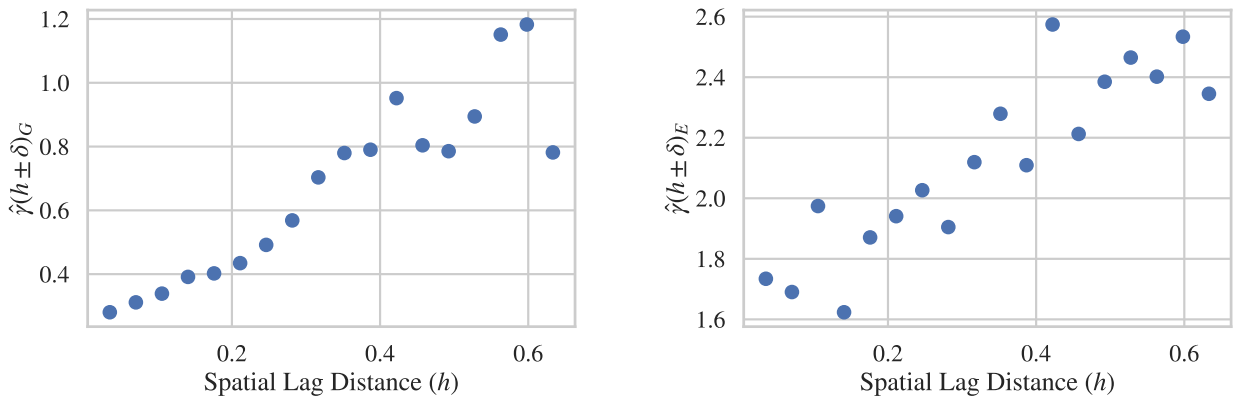


Figure 8: Empirical variograms constructed using geodesic and Euclidean distances during working hours (8:00 am to 3:00 PM) of weekdays throughout the year. (a) The variogram constructed using geodesic distances $\hat{\gamma}(h \pm \delta)_G$ reveals behavior that could be characterized through a simple variogram model (e.g., Gaussian model) , whereas the (b) Euclidean variogram shows almost no structure. The incorporation of the manifold geometry into the analysis of spatial depence can reveal information that is missed if Euclidean geometry is assumed.

4 Discussion and Conclusions

The analysis of multi-pollutant atmospheric data at multiple sites presents a unique challenge for the atmospheric sciences community. There is a need for the development of methods that are capable of handling this complex spatio-temporal data in order to better understand the short and long-term impacts of atmospheric pollutants on the environment, climate, and human health. We have demonstrated a set of methods for understanding and quantifying the spatial and temporal dynamics and relationships in multi-pollutant data through covariance matrices and Riemannian geometry. Covariance matrices form a Riemannian manifold. We leverage the geometry of this manifold in various data-centric tasks such as dimensionality reduction and spatial interpolation. We also compare these methods to ones that (incorrectly) assume a Euclidean geometry for the covariance matrices, which have been employed in previous studies [35, 14, 22]. The Euclidean geometry assumption leads to degradation in the performance of the proposed data-centric tasks (e.g., dimensionality reduction, spatial interpolation). This degradation is partly due to the need for normalization and scaling of the data (which the Riemannian geometric methods do not require) and the introduction of spurious variations in the data due to effects such as swelling. In future work we aim to study the full impact of Riemannian geometry on spatial completion algorithms such as Kriging and how Riemannian geom-

etry might be included in more advanced spatio-temporal analysis methods and for systems where the data may be non-stationary which has been explored recently [24, 21]. We also note that while we have focused on SPD covariance matrices the ideas of Riemannian geometry extend to other areas of statistics (e.g., information geometry), different types of matrices (e.g., matrices of fixed rank), and data that is known to lie on a Riemannian manifold (e.g., the surface of a sphere) that may arise in the analysis of atmospheric datasets.

4.1 Acknowledgement

The authors acknowledge funding from the U.S. National Science Foundation (NSF) under BIGDATA grant IIS-1837812.

4.2 Supplementary Information Available

All code and data needed to reproduce the results can be found in https://github.com/zavalab/ML/tree/master/Atmospheric_Analysis.

References

[1] V. Arsigny, P. Fillard, X. Pennec, and N. Ayache. Geometric means in a novel vector space structure on symmetric positive-definite matrices. *SIAM journal on matrix analysis and applications*, 29(1):328–347, 2007.

[2] A. Barachant and M. Congedo. A plug&play p300 bci using information geometry. *arXiv preprint arXiv:1409.0107*, 2014.

[3] R. Bhatia. Positive definite matrices. In *Positive Definite Matrices*. Princeton university press, 2009.

[4] R. Bhatia and J. Holbrook. Riemannian geometry and matrix geometric means. *Linear algebra and its applications*, 413(2-3):594–618, 2006.

[5] N. Cressie. *Statistics for spatial data*. John Wiley & Sons, 2015.

[6] N. Cressie and D. M. Hawkins. Robust estimation of the variogram: I. *Journal of the international Association for Mathematical Geology*, 12(2):115–125, 1980.

[7] F. Dominici, R. D. Peng, C. D. Barr, and M. L. Bell. Protecting human health from air pollution: shifting from a single-pollutant to a multi-pollutant approach. *Epidemiology (Cambridge, Mass.)*, 21(2):187, 2010.

[8] I. L. Dryden, A. Koloydenko, and D. Zhou. Non-euclidean statistics for covariance matrices, with applications to diffusion tensor imaging. 2009.

[9] U. EPA. The multi-pollutant report: Technical concepts and examples, 2007.

[10] P. T. Fletcher and S. Joshi. Principal geodesic analysis on symmetric spaces: Statistics of diffusion tensors. In *Computer Vision and Mathematical Methods in Medical and Biomedical Image Analysis: ECCV 2004 Workshops CVAMIA and MMBIA, Prague, Czech Republic, May 15, 2004, Revised Selected Papers*, pages 87–98. Springer, 2004.

[11] P. T. Fletcher, C. Lu, S. M. Pizer, and S. Joshi. Principal geodesic analysis for the study of nonlinear statistics of shape. *IEEE transactions on medical imaging*, 23(8):995–1005, 2004.

[12] J. Hua, Y. Zhang, B. de Foy, X. Mei, J. Shang, and C. Feng. Competing pm2. 5 and no2 holiday effects in the beijing area vary locally due to differences in residential coal burning and traffic patterns. *Science of The Total Environment*, 750:141575, 2021.

[13] M. Kampa and E. Castanas. Human health effects of air pollution. *Environmental pollution*, 151(2):362–367, 2008.

[14] Y. Kim, M. Kim, J. Lim, J. T. Kim, and C. Yoo. Predictive monitoring and diagnosis of periodic air pollution in a subway station. *Journal of hazardous materials*, 183(1-3):448–459, 2010.

[15] A. Kumar and P. Goyal. Forecasting of air quality index in delhi using neural network based on principal component analysis. *Pure and Applied Geophysics*, 170(4):711–722, 2013.

[16] J. M. Lee. *Riemannian manifolds: an introduction to curvature*, volume 176. Springer Science & Business Media, 2006.

[17] W.-Z. Lu, H.-D. He, and L.-y. Dong. Performance assessment of air quality monitoring networks using principal component analysis and cluster analysis. *Building and Environment*, 46(3):577–583, 2011.

[18] J. S. Marron and I. L. Dryden. *Object oriented data analysis*. CRC Press, 2021.

[19] J. L. Mauderly. The national environmental respiratory center (nerc) experiment in multi-pollutant air quality health research: I. background, experimental strategy and critique. *Inhalation Toxicology*, 26(11):643–650, 2014.

[20] H. Mayer. Air pollution in cities. *Atmospheric environment*, 33(24-25):4029–4037, 1999.

[21] A. Menafoglio, D. Pigoli, and P. Secchi. Kriging riemannian data via random domain decompositions. *Journal of Computational and Graphical Statistics*, 30(3):709–727, 2021.

[22] I. Ozga, N. Ghedini, C. Giosuè, C. Sabbioni, F. Tittarelli, and A. Bonazza. Assessment of air pollutant sources in the deposit on monuments by multivariate analysis. *Science of the total environment*, 490:776–784, 2014.

[23] X. Pennec, P. Fillard, and N. Ayache. A riemannian framework for tensor computing. *International Journal of computer vision*, 66(1):41–66, 2006.

[24] D. Pigoli, A. Menafoglio, and P. Secchi. Kriging prediction for manifold-valued random fields. *Journal of Multivariate Analysis*, 145:117–131, 2016.

[25] V. Ramanathan and Y. Feng. Air pollution, greenhouse gases and climate change: Global and regional perspectives. *Atmospheric environment*, 43(1):37–50, 2009.

[26] S. Sillman and D. He. Some theoretical results concerning o3-nox-voc chemistry and nox-voc indicators. *Journal of Geophysical Research: Atmospheres*, 107(D22):ACH–26, 2002.

[27] A. Smith, A. Keane, J. A. Dumesic, G. W. Huber, and V. M. Zavala. A machine learning framework for the analysis and prediction of catalytic activity from experimental data. *Applied Catalysis B: Environmental*, 263:118257, 2020.

[28] A. Smith, B. Laubach, I. Castillo, and V. M. Zavala. Data analysis using riemannian geometry and applications to chemical engineering. *arXiv preprint arXiv:2203.12471*, 2022.

[29] A. Smith, S. Runde, A. Chew, A. Kelkar, U. Maheshwari, R. Van Lehn, and V. Zavala. Topological analysis of molecular dynamics simulations using the euler characteristic. 2022.

[30] A. Smith and V. M. Zavala. The euler characteristic: A general topological descriptor for complex data. *Computers & Chemical Engineering*, 154:107463, 2021.

[31] A. D. Smith, N. Abbott, and V. M. Zavala. Convolutional network analysis of optical micrographs for liquid crystal sensors. *The Journal of Physical Chemistry C*, 124(28):15152–15161, 2020.

[32] A. D. Smith, P. Dłotko, and V. M. Zavala. Topological data analysis: concepts, computation, and applications in chemical engineering. *Computers & Chemical Engineering*, 146:107202, 2021.

[33] E. G. Snyder, T. H. Watkins, P. A. Solomon, E. D. Thoma, R. W. Williams, G. S. Hagler, D. Shelow, D. A. Hindin, V. J. Kilaru, and P. W. Preuss. The changing paradigm of air pollution monitoring. *Environmental science & technology*, 47(20):11369–11377, 2013.

[34] S. Sommer, F. Lauze, S. Hauberg, and M. Nielsen. Manifold valued statistics, exact principal geodesic analysis and the effect of linear approximations. In *European conference on computer vision*, pages 43–56. Springer, 2010.

[35] M. Statheropoulos, N. Vassiliadis, and A. Pappa. Principal component and canonical correlation analysis for examining air pollution and meteorological data. *Atmospheric environment*, 32(6):1087–1095, 1998.

[36] A. D. Syafei, A. Fujiwara, and J. Zhang. Prediction model of air pollutant levels using linear model with component analysis. *International Journal of Environmental Science and Development*, 6(7):519, 2015.

[37] L. Van Der Maaten, E. Postma, J. Van den Herik, et al. Dimensionality reduction: a comparative. *J Mach Learn Res*, 10(66-71):13, 2009.

[38] K. You and H.-J. Park. Re-visiting riemannian geometry of symmetric positive definite matrices for the analysis of functional connectivity. *NeuroImage*, 225:117464, 2021.

[39] T.-Y. Yu and L.-F. W. Chang. Selection of the scenarios of ozone pollution at southern taiwan area utilizing principal component analysis. *Atmospheric environment*, 34(26):4499–4509, 2000.

[40] A. Zanobetti, E. Austin, B. A. Coull, J. Schwartz, and P. Koutrakis. Health effects of multi-pollutant profiles. *Environment international*, 71:13–19, 2014.