Applied Topology and Geometry in Process Monitoring and Fault Detection

Alexander Smith

Department of Chemical Engineering and Material Science, University of Minnesota, Minneapolis, MN 55455

Abstract

Datasets are mathematical objects (e.g., point clouds, matrices, graphs, images, fields/functions) that have shape. This shape captures intrinsic characteristics of the data that are independent of the environment and methods used to obtain the data. These representations (e.g., graphs, manifolds, point clouds) also provide means for integrating domain knowledge that can aid in the development of efficient, interpretable models for data analysis. Topology and geometry are fields of mathematics that provide tools for the characterization and quantification of these data representations (shapes). In this work, we apply topological and geometrical methods in the areas of industrial process monitoring and fault detection. We discuss how data taken from industrial processes (e.g., time series, images) can be represented as a shape and how that shape can be analyzed through topological and geometrical tools such as the Euler characteristic (EC) and Riemannian geometry. We provide a brief overview of these methods and illustrate how exploiting the topology and geometry of data can provide improvements in data-centric tasks such as dimensionality reduction and classification in the context of process monitoring and image analysis.

Keywords

Topology, Geometry, Data Science, Process Systems

Datasets are mathematical objects and characterizing the shape (geometrical and topological features) of these objects reduces the dimensionality and complexity of the data while minimizing information loss, but is not always straightforward (Smith et al., 2021). Popular tools from statistics, linear algebra, and signal processing (e.g., moments, correlation functions, singular value decomposition, convolutions, Fourier analysis) do not directly characterize the shape of data objects; instead, such tools are used to characterize other types of features (e.g., variance and frequency content).

Topology is a branch of mathematics that provides powerful tools to characterize the shape of data objects. One such tool is the so-called Euler characteristic (EC). The EC is a descriptor that characterizes geometrical features of a topological space defined by a data object (Smith and Zavala, 2021). This characterization is accomplished by performing a decomposition of the space into a set of independent topological bases. This decomposition is similar in spirit to an eigen-decomposition of a matrix; here, the matrix object is decomposed into a set of independent basis vectors. The EC is a scalar integer quantity that is defined as the alternating sum of the rank of the topological bases. The EC is often combined with a transformation technique known as filtration to characterize the geometry of different objects such as matrices, images, fields/functions, and weighted graphs. This characterization is summarized in the form of what is called an EC curve which provides a direct approach to quantify the topology of data. Topological descriptors such as the EC offer advantages over statistical descriptors. Descriptors such as Moran's I, which measures spatial structure via spatial autocorrelation, or correlation matrices do not directly capture the global structure of the data (thus limiting the ability to characterize geometrical features) (Mantz et al., 2008). Higher-order statistical descriptors such as 2-point correlation functions, which have been employed in characterizing the structure of heterogeneous materials, are also limited at capturing spatial and morphological features of the data (especially if the data object is irregular) (Mantz et al., 2008).

Topology and geometry also allow us to study data that is governed by *non-Euclidean* geometry. The assumption that data lies in a Euclidean space is pervasive throughout science and engineering and is the basis of diverse data analysis techniques used in these domains. Making this blanket assumption, however, is not always appropriate and can affect the accuracy/interpretability of such techniques or even break fundamental physical laws. An important example in which assuming Euclidean geometry can lead to spurious results is in the analysis of symmetric positive definite (SPD) matrices (e.g., covariance/correlation matrices) (Smith et al., 2022; Moakher and Batchelor, 2006). SPD matrices lie on a high-dimensional space which is governed by Riemannian geometry (known as a Riemannian manifold). Standard techniques for the analysis of SPD matrices (e.g., PCA or basic matrix norms) do not take this property into consideration and can lead to misleading results. Specifically, the so-called swelling effect can occur when applying operations in Euclidean geometry to SPD matrices. This effect introduces spurious results by inflating the determinants of SPD matrices and can also distort the results of commonly used methods (Smith et al., 2022; Moakher and Batchelor, 2006; Pennec, 2006). Computing interpolations and averages of SPD matrices, which is key in understanding physical systems (e.g. Brownian motion), can also break physical conservation laws if performed under Euclidean geometry (Smith et al., 2022; Moakher and Batchelor, 2006; Pennec, 2006).

In this work, we provide a brief overview of the EC and Riemannian geometry and their applications in the context of process monitoring and fault detection. Our applications are focuses on the Tennessee Eastman Process (TEP) and the MVTEC-AD dataset. The TEP is a chemical process where anomalies/faults are systematically introduced which shifts the relationships between the measured variables (Downs and Vogel, 1993). Covariance matrices encode these changing relationships and are then used to predict what type of anomaly the process is experiencing. The second application is in defect/anomaly detection of textiles taken from the MVTEC AD dataset (Bergmann et al., 2019). Here, the topology of textile images are analyzed for the presence of anomalies/defects (e.g., a cut or discoloration of the textile).

Graphs and Manifolds

We begin by defining a couple of fundamental topological/geometrical data representations: graphs and manifolds. A graph is a 2D topological object that consists of an ordered pair $G(V,E)$, where V represents a set of *vertices* and E represents a set of paired vertices known as *edges*. Edges represent relationships (connectivity) between vertices. We can characterize a graph by quantifying specific topological features such as the number of *cycles* and *connected components*. A cycle represents a path that traverses edges on a graph starting at a particular vertex v_j and ending at that same vertex v_j. A connected component is a subset of a graph $C(V_C,E_C) \subseteq G(V,E)$ in which any vertex $v_i \in V_C$ of the subgraph can reach any other vertex $v_j \in V_C$ by traversing edges of the subgraph $\{v_i,v_j\} \in E_C$, and is disconnected from all other subsets of the graph. In other words, the number of connected components is the number of connected partitions of a graph. Data can also be encoded in a graph object (in nodes and edges) using functions $f : V \to \mathbb{R}$ and $f : E \to \mathbb{R}$. Values attached to nodes or edges are typically called weights or features; as such, graphs that encode data are also known as weighted graphs.

Manifolds are also versatile topological data representations that can capture continuous forms of information in high-dimensional spaces. This contrasts with graph representations, which capture discrete characteristics of a data object (e.g., number of edges, nodes). A manifold $\mathcal{M}$ is a topological space that *locally* resembles a Euclidean space; this means that the neighborhood of a point $x \in \mathcal{U}$ in an n-dimensional manifold (with $\mathcal{U} \subseteq \mathcal{M}$) can be mapped to n-dimensional Euclidean space through a continuous, bijective function. These neighborhoods and associated mappings are also known as *charts*. For example, the surface of the Earth is a 2D manifold and we can map the curved surface of the Earth to a flat Euclidean plane (i.e., a 2D Euclidean space) using a chart in order to measure properties such as distances or areas. The general nature of manifolds allows them to represent a broad range of structures, shapes, and complex geometric objects. Manifolds can also have encoded data on them (e.g., Earth surface temperature), which is captured using a continuous function $f : \mathcal{M} \to \mathbb{R}$. In Figure 1, we present a manifold representation for a textile image. Here, the image is a 2D manifold and we define a continuous function that captures the pixel intensity at each point in the image.

The Euler Characteristic

Graph and manifold representations are able to capture both discrete and continuous information within a dataset and the data's topology can be directly quantified/summarized using a descriptor known as the Euler characteristic (Smith and Zavala, 2021). The EC is denoted as $\chi \in \mathbb{Z}$ and is mathematically defined as the alternating sum of the rank of topological bases for a given space known as Betti numbers $\beta_i \in \mathbb{Z}_+$, where $i \in \mathbb{Z}_+$ represents the dimensionality of the topological basis:

$$\chi := \sum_{i=0}^{n} (-1)^i \beta_i \tag{1}$$

Importantly, the topological bases of a space (e.g., connected components, holes, voids) are preserved under deformations such as stretching, twisting, and bending (are topological invariants). For any topological space of n-dimensions, there can only exist topological bases up to that given dimension.

Filtrations

Analysis of data represented as manifolds (or weighted graphs) requires an added processing step known as a *filtration*. A filtration quantifies the topology of *sublevel sets* of the manifold. Given an n-dimensional manifold $\mathcal{M}$ and a continuous function $f : \mathcal{M} \to \mathbb{R}$, a *sublevel set* of the manifold is defined as $\mathcal{M}_{k_i}$ that contains points $\{x \in \mathcal{M} : f(x) \leq k_i\}$, where $k_i \in \mathbb{R}$ represents our *filtration threshold*. Hence, we can construct nested sublevel sets at increasing filtration thresholds for the manifold (or graph):

$$\mathcal{M}_{k_1} \subseteq \mathcal{M}_{k_2} \subseteq ... \subseteq \mathcal{M}_{k_n} \subseteq \mathcal{M} \tag{2}$$

where $k_1 < k_2 < ... < k_n$ represent our filtration thresholds, and $\mathcal{M}$ represents the original manifold. We can measure/quantify the topology of these nested sublevel sets with the EC at each filtration threshold $\{\chi_1, \chi_2, ..., \chi_n\}$. We ultimately obtain an ordered pair of values $\{k_i, \chi_i\}$, which characterize the topology of the manifold and its associated function. The filtration of a weighted graph is conducted in an analogous manner (by eliminating nodes or edges in which the data is below a certain threshold value). We note that filtration operations are easy to conduct and are thus scalable. An example filtration of a textile image is found in Figure 1. Here, the textile image is treated as a manifold $\mathcal{M} \in \mathbb{R}^2$ with

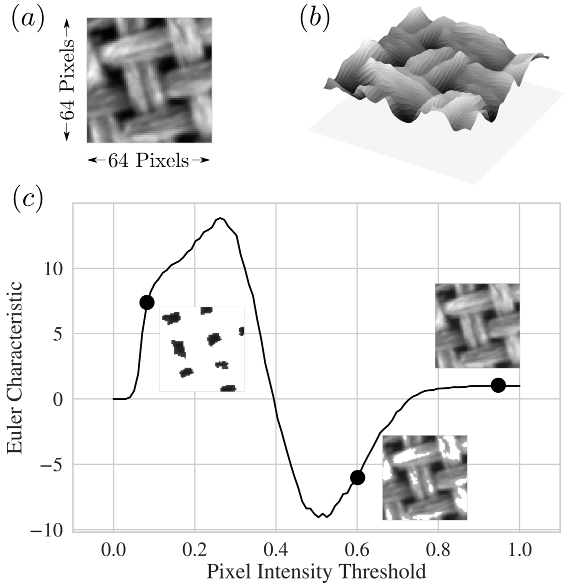

Figure 1: (a) Non-defective textile image. (b) Manifold representation of a textile image $\mathcal{M}$, here the 3rd dimension (z-axis) represents the value of the function $f:\mathcal{M}\to\mathbb{R}$. (c) EC curve constructed from the filtration of the manifold/function shown in (b), with sublevel sets $\mathcal{M}_{0.1},\mathcal{M}_{0.6},\mathcal{M}_{0.95}$. The EC is computed at sublevel sets $\mathcal{M}_{k_i}$ with increasing thresholds $k_i\in\mathbb{R}$.

a function $f:\mathcal{M}\to\mathbb{R}$ representing the pixel intensity at each point in the image. The graph represents the ordered pairs $\{k_i,\chi_i\}$ computed during the filtration.

Riemannian Geometry

We now shift our focus to a brief review of geometric methods of data analysis from the field of Riemannian geometry. In particular, we are focused on the applications of Riemannian geometry with respect to *symmetric*, *postive definite* (SPD) matrices. SPD matrices are widely used in the analysis of process data, and most commonly appear in the form of covariance/correlation matrices. The main message of this brief review is that SPD matrices lie on a Riemannian manifold and that important computations (e.g., matrix operations, summarizing statistics, classification, regression, and dimensionality reduction) can be performed by incorporating the geometry of this manifold. Respecting such properties can lead to important improvements in efficiency and interpretability.

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. We also illustrate several points $\mathbf{P}_1,\mathbf{P}_2,\mathbf{P}_3\in\mathcal{P}(n)$, where each $\mathbf{P}_i$ represents an SPD covariance matrix.

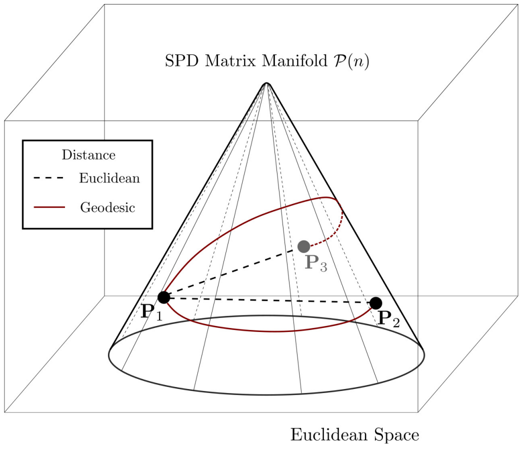

Figure 2: An illustration of a Riemannian manifold formed by SPD matrices $\mathcal{P}(n)$, with a set of points (matrices) $\mathbf{P}_1,\mathbf{P}_2,\mathbf{P}_3\in\mathcal{P}(n)$. A comparison of geodesic (red, solid line) and Euclidean distances (dashed, black line) are shown between $\mathbf{P}_1,\mathbf{P}_2$ and $\mathbf{P}_1,\mathbf{P}_3$.

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:

$$d_E(\mathbf{P}_i,\mathbf{P}_j):=||\mathbf{P}_i-\mathbf{P}_j||_F \tag{3}$$

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

$$d_G(\mathbf{P}_i,\mathbf{P}_j):=||\log(\mathbf{P}_i)-\log(\mathbf{P}_j)||_F \tag{4}$$

where $\log(\cdot)$ represents the matrix logarithm. 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):

$$d_E(\mathbf{P}_1,\mathbf{P}_2)>d_E(\mathbf{P}_1,\mathbf{P}_3) \tag{5}$$

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

$$d_G(\mathbf{P}_1,\mathbf{P}_3)>d_G(\mathbf{P}_1,\mathbf{P}_2) \tag{6}$$

Thus, accounting for the correct (non-Euclidean) geometry of the manifold provides a different answer than if Euclidean geometry is assumed. The impact of incorrectly assuming Euclidean geometry in the analysis of matrices is demonstrated in the TEP case study.

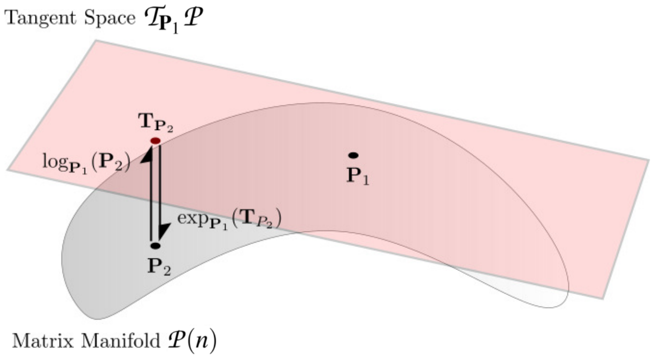
Tangent Space $\mathcal{T}_{\mathbf{P}_1}\mathcal{P}$
$\mathbf{T}_{\mathbf{P}_2}$
$\log_{\mathbf{P}_1}(\mathbf{P}_2)$
$\mathbf{P}_1$
$\exp_{\mathbf{P}_1}(\mathbf{T}_{P_2})$
$\mathbf{P}_2$
Matrix Manifold $\mathcal{P}(n)$

Figure 3: 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 $\log_{\mathbf{P}_1}(\mathbf{P}_2) \to \mathbf{T}_{\mathbf{P}_2}$ 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 $\exp_{\mathbf{P}_1}(\mathbf{T}_{\mathbf{P}_2}) \to \mathbf{P}_2$ 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 $\mathcal{T}_{\mathbf{P}_1}\mathcal{P}$. This mapped data can then be used directly in common data analysis methods such as principal component analysis.

Tangent Spaces

The analysis of SPD 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 (Bhatia, 2009; Smith et al., 2022). This means that every point on the Riemannian manifold $\mathbf{P} \in \mathcal{P}(n)$ has an associated tangent space denoted as $\mathcal{T}_{\mathbf{P}}\mathcal{P} \in \mathcal{S}(n)$, constructed from the tangent vectors of all possible *curves* passing through the point on the manifold $\mathbf{P} \in \mathcal{P}(n)$ (Smith et al., 2022). We define a curve as a continuous function $\phi : [0,1] \to \mathcal{P}(n)$. 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 (Bhatia, 2009). 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 3. Figure 3 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. The logarithmic map takes a point ($\mathbf{P}_2$ in Figure 3) on the manifold and maps it to a point in the tangent space ($\mathbf{T}_{\mathbf{P}_2} \in \mathcal{T}_{\mathbf{P}_1}\mathcal{P}$ in Figure 3). For two points $\mathbf{P}_i, \mathbf{P}_j \in \mathcal{P}(n)$ we define the logarithmic map as:

$$\log_{\mathbf{P}_i}(\mathbf{P}_j) := \mathbf{P}_i^{1/2} \log\left(\mathbf{P}_i^{-1/2}\mathbf{P}_j\mathbf{P}_i^{-1/2}\right)\mathbf{P}_i^{1/2} \tag{7}$$

The exponential map is the inverse of the logarithmic map. It maps points from a tangent space ($\mathbf{T}_{\mathbf{P}_2}$ in Figure 3) back to the Riemannian manifold ($\mathbf{P}_2$ in Figure 3). 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:

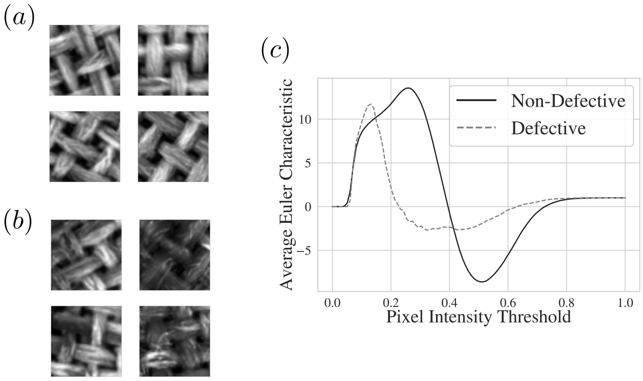

Figure 4: (a) Set of non-defective textile images. (b) Set of defective textile images. (c) The average EC curve computed for the group of defective and non-defective textiles. There is a distinct, quantifiable difference between the EC curves of the two groups. This leads to highly accurate classification of the textile images as either defective or non-defective.

$$\exp_{\mathbf{P}_i}(\mathbf{T}_{\mathbf{P}_j}) := \mathbf{P}_i^{1/2} \exp\left(\mathbf{P}_i^{-1/2}\mathbf{T}_{\mathbf{P}_j}\mathbf{P}_i^{-1/2}\right)\mathbf{P}_i^{1/2} \tag{8}$$

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). This transformation allows us to incorporate the geometry of the Riemannian manifold in multivariate data analysis methods that assume data lies in a vector space, such as PCA, classification, and regression.

Application: Textile Manufacturing

We now explore an application of the Euler characteristic in the analysis of manufacturing images taken from the production of textiles. The goal in this analysis is to identify whether or not a given textile image represents a defective (faulty) or non-defective textile. Example images of defective and non-defective textiles can be found in Figure 4.

To characterize the topology of the textile images, we first represent the images as manifolds $\mathcal{M}$ with an associated function $f : \mathcal{M} \to \mathbb{R}$ representing the pixel intensity at each point of a given image. We can now perform a filtration and quantify the EC at each point of the filtration, constructing an EC curve. This process is illustrated in Figure 1 for a non-defective textile image. Figure 4 also shows the average EC curves for the set of defective and non-defective textiles. There is an obvious difference between the mean EC curves of these two groups, allowing us to distinguish defective and non-defective textiles with high accuracy (98% testing set accuracy) using a simple linear classifier. We trained the linear classifier using 70 % of the dataset and tested the model on the remaining 30 %. These same results are not achievable when the images are not pre-processed with the EC. Figure 5 compares an application of Principal Component Analysis (PCA) to data that has been pre-processed with the EC and filtration against PCA applied directly to the images. PCA

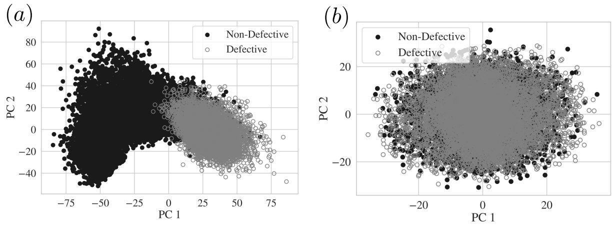

Figure 5: (a) PCA analysis of the EC curves derived from the set of defective and non-defective textiles. (b) PCA analysis of the images without pre-processing. PCA on the EC captures clustering of the data into two groups (defective, non-defective), whereas PCA applied to the images shows complete overlap of the two groups.

performed on the pre-processed data shows a distinct separation and clustering of the data into groups representing the defective and non-defective textile images. PCA performed directly on the images shows no separation in the data, and when a linear classifier is applied to the images the resulting classification accuracy is very low (54% testing set accuracy). This dramatic improvement in performance is most likely due to the invariance of the EC to various transformations of the data (e.g., translation, rotation), making it robust to the different rotations and placements of the textiles within the images.

Application: Tennessee Eastman Process

We focus on data obtained from a simulated industrial process known as the Tennessee Eastman Process (TEP) (Downs and Vogel, 1993). This dataset is a widely used benchmark dataset for testing and comparing various anomaly (i.e. fault) detection methods (Chiang et al., 2000). Figure 6 provides a high-level illustration of the process along with the multivariate time series data that is produced by the sensors monitoring the process. The process has a total of 52 measurements, 41 are process variables, 11 are manipulated variables. There are 20 different potential faults that can occur in the TEP; for further details see Appendix A in (Downs and Vogel, 1993).

To begin our analysis of the TEP, we must first pre-process the TEP data into multiple covariance matrices. To accomplish this, we represent each of the 52 measured variables as a univariate random variable x_i where $i = 1, 2, ..., 52$ and we denote the collection of signals as a multivariate random vector $\mathbf{X} = (x_1, ..., x_n)$ where $n = 52$. We denote the observations of each signal at time $t = 1, 2, .., m$ as $x_i(t) \in \mathbb{R}^m$. We use this representation to construct the sample covariance matrix for the process data $\mathbf{P} \in \mathcal{P}(n)$ as:

$$\mathbf{P} := \frac{1}{m-1}\mathbf{X}\mathbf{X}^T \tag{9}$$

The TEP dataset consists of multiple separate simulations of the process, both with and without faults. Thus, for each simulation we construct a sample covariance matrix $\mathbf{P}_i$. Our goal is to pair each simulation sample covariance matrix with the fault occurring in the simulation. We note that there are no obvious differences between the covariance matrices that would identify a given fault (Figure 6). These covariance matrices lie on the SPD manifold, and can be integrated into our geometric framework. In the analysis of a set of SPD matrices, we often need to identify a center point on the SPD manifold that will minimize the distortion of all geometric relationships between the matrices of the dataset when mapped to the tangent space. This matrix is the (Riemannian) geometric mean of the matrices. For a set of SPD matrices, the geometric mean is the matrix that minimizes the sum of squared geodesic distances to all other matrices in the set. We compute the geometric mean $\bar{\mathbf{P}}$ for our set of covariance matrices $\mathbf{P}_i \in \mathcal{P}(n)$, the matrices are then projected to the tangent space $\mathcal{T}_{\bar{\mathbf{P}}}\mathcal{P}$ centered at the geometric mean via the logarithmic mapping $\mathbf{T}_{\mathbf{P_i}} = \log_{\bar{\mathbf{P}}}(\mathbf{P}_i)$. The data is now projected into a vector space that retains the geometric characteristics of the SPD matrices with minimal distortion, and can be integrated in dimensionality reduction and classification algorithms to perform analysis.

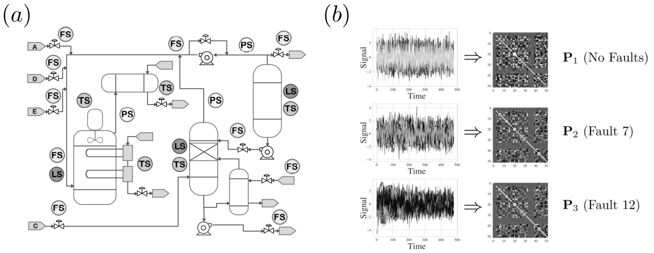

Figure 6: (a) Illustration of the Tennesee Eastman Process (TEP). (b) Representation of the multivariate time series derived from the TEP as covariance matrices.

Mapping the process data covariance matrices to the tangent space provides an avenue for the application of common dimensionality reduction techniques. Here, we apply PCA to the matrices mapped to the tangent (vector) space. PCA applied on the tangent space of the SPD manifold is commonly known as Principal Geodesic Analysis (PGA), as it identifies the geodesics that capture the most variance in the data (Smith et al., 2022). An example comparison of PGA versus PCA (directly on the covariance matrices) is presented in Figure 7. The simulations with no faults are colored in red and the faulty simulations are represented by different grayscale values. We can see that using only the first two components in PGA, we are able to perfectly separate the faulty and non-faulty simulations; on the other hand, when applying PCA directly on the covariance matrices we can see that there is significant overlap between the faulty and non-faulty simulations. The comparison of these projections demonstrates that capturing the geometry of the Riemannian manifold in the analysis of covariance matrices can improve the performance with minimal added complexity.

PGA analysis of the covariance matrices also reveals that there is definite clustering of the data with respect to the different fault types within the TEP dataset. This suggests that classification of the different fault types can be done directly using the tangent space of the SPD manifold. To investigate this we use a simple linear (ridge) classifier. We compare the prediction accuracy of the linear classifier using coefficients of the tangent space projected matrices versus the co-

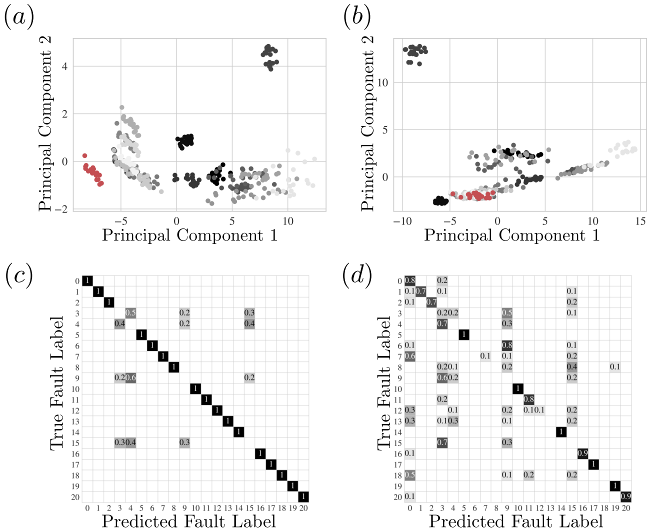

Figure 7: (a) PCA analysis of the covariance matrices mapped to the tangent space (non-Euclidean assumption). (b) PCA analysis applied directly to the covariance matrices (Euclidean assumption). (c) Confusion matrix for the classification of faults using the covariance matrices mapped to the tangent space. (d) Confusion matrix for the classification of faults using the covariance matrices directly.

efficients of the non-transformed covariance matrices as input. In the analysis, we perform a simple train-test split of the data, where 30% of the data is used for testing and 70% of the data is use for training. Figure 7 illustrates the dramatic increase in accuracy when the model incorporates geometric information, which is reflected in the normalized confusion matrices. Here, a value $x \in [0,1]$ on the diagonal indicates an accuracy of $x * 100\%$ when classifying a particular fault. All values in the off diagonal (e.g., row i, column j, where $i \neq j$) represent the percentage of covariance matrices associated with fault i that have been incorrectly labeled as experiencing fault j. When the SPD manifold is accounted for via the tangent space projection, there is perfect classification of the data (with the exception of faults $3, 4, 9$ and 15). When the manifold geometry is ignored, there are few instances where high classification accuracy is achieved. The faults $3, 4, 9$ and 15 have been shown in prior work to be difficult to classify (Smith et al., 2022). We also note that these faults are only misclassified within their group (are never classified as having no fault), which suggests that there is limited quantifiable difference in the covariance matrices for these faults. The inclusion of more information around these particular faults may correct this issue and further increase accuracy.

Conclusion

Process systems data (e.g., images, time series) have complex topology and geometry which encodes information that is missed by common data analysis methods. Extraction and quantification of the information encoded in the shape of data can lead to improvements in data-centric tasks that are applied to the data (e.g., PCA, classification, regression). Correctly characterizing the topology and geometry of data can also provide robustness to various transformations of data (e.g., translations, rotations) and can greatly simplify the models needed to analyze data effectively. We demonstrated these benefits through applications in textile manufacturing (image analysis) and in the fault detection for chemical processes (Tennessee Eastman Process). This work has focused on topology and geometry in fault detection and process monitoring, but we note that relevance of these mathematical frameworks extends beyond this area of application. For example, in future work we aim to incorporate geometry in optimization and control to relax certain constraints that are already enforced by the underlying system geometry (e.g., optimization on the SPD manifold ensures that all solutions are SPD).

References

Bergmann, P., M. Fauser, D. Sattlegger, and C. Steger (2019). MVTec AD–A comprehensive real-world dataset for unsupervised anomaly detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9592–9600.

Bhatia, R. (2009). *Positive definite matrices*. Princeton university press.

Chiang, L. H., E. L. Russell, and R. D. Braatz (2000). Fault diagnosis in chemical processes using Fisher discriminant analysis, discriminant partial least squares, and principal component analysis. *Chemometrics and intelligent laboratory systems 50*(2), 243–252.

Downs, J. J. and E. F. Vogel (1993). A plant-wide industrial process control problem. *Computers & chemical engineering 17*(3), 245–255.

Mantz, H., K. Jacobs, and K. Mecke (2008). Utilizing minkowski functionals for image analysis: a marching square algorithm. *Journal of Statistical Mechanics: Theory and Experiment 2008*(12), P12015.

Moakher, M. and P. G. Batchelor (2006). Symmetric positive-definite matrices: From geometry to applications and visualization. In *Visualization and Processing of Tensor Fields*, pp. 285–298. Springer.

Pennec, X. (2006). Intrinsic statistics on Riemannian manifolds: Basic tools for geometric measurements. *Journal of Mathematical Imaging and Vision 25*(1), 127–154.

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

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

Smith, A. D. and V. M. Zavala (2021). The Euler characteristic: A general topological descriptor for complex data. *Computers & Chemical Engineering 154*, 107463.