Determining the Cause of Negative Dissimilarity Eigenvalues

Weiping Xu, Richard C. Wilson, and Edwin R. Hancock

Dept. of Computer Science, University of York, UK
{elizaxu,wilson,erh}@cs.york.ac.uk

Abstract. Pairwise dissimilarity representations are frequently used as an alternative to feature vectors in pattern recognition. One of the problems encountered in the analysis of such data, is that the dissimilarities are rarely Euclidean, and are sometimes non-metric too. As a result the objects associated with the dissimilarities can not be embedded into a Euclidean space without distortion. One way of gauging the extent of this problem is to compute the total mass associated with the negative eigenvalues of the dissimilarity matrix. However,this test does not reveal the origins of non-Euclidean or non-metric artefacts in the data. The aim in this paper is to provide simple empirical tests that can be used to determine the origins of the negative dissimilarity eigenvalues. We consider three sources of the negative dissimilarity eigenvalues, namely a) that the data resides on a manifold (here for simplicity we consider a sphere), b) that the objects may be extended and c) that there is Gaussian error. We develop three measures based on the non-metricity and the negative spectrum to characterize the possible causes of non-Euclidean data. We then experimentally test our measures on various real-world dissimilarity datasets.

Keywords: non-Euclidean pairwise data, metric, embedding.

1 Introduction

Pairwise dissimilarity representations offer a powerful alternative to vectorial or feature-based characterisations of objects. Specifically, they provide a natural way of capturing the relationships between objects that are not characterised by ordinal measurements or feature vectors [6]. One way to translate such data into a vector representation is to represent the similarity data using a kernel matrix, and to embed the data into a vector space using kernel principal components analysis. In this way a vector representation is obtained by projecting the dissimilarity data into a vector space of fixed dimension.

However, one of the problems with dissimilarity representations and their embeddings is that the distance measures can not be used to construct a Euclidean vector space if the underlying Gram matrix contains negative eigenvalues. If this is the case, then the data can not be embedded into a real-valued Euclidean space, and must instead be embedded into a complex valued or Krein space [5].

A. Berciano et al. (Eds.): CAIP 2011, Part I, LNCS 6854, pp. 589–597, 2011.

In order to analyse non-Euclidean dissimilarity data using traditional geometric machine learning or pattern recognition techniques, we must first attempt to rectify the data so as to minimize the non-Euclidean artifacts. Examples of translating similarities into vector representation include using only the positive definite subspace of the distances, adding a constant amount to the off diagonal elements, i.e. the constant shift embedding [4], or manifold embedding (e.g. the spherical embedding in [1]).

Each of these approaches is based on assumptions concerning the sources of the negative eigenvalues. The positive definite subspace embedding assumes that metric violations are an artifact of noise and that the distances in the negative sub-space do not carry any significant discriminative information. The manifold embedding assumes that the Euclidean violations are geodesic and that the data resides on a manifold. Recent studies [2,4] have showed that the negative eigenspace can contain valuable information. Moreover, Euclidean correction can lead to poor classification performance. Thus, before using any of the above approaches to attempt to rectify non-Euclidean data, it is advisable to analyze the underlying causes.

We model the distribution of non-Euclidean pairwise data in the following three situations: a) that the objects reside on the surface of a sphere (a simple manifold) and that the pairwise similarities are geodesic distances across the manifold, b) a non-metric dataset based on the distances between the surfaces of randomly positioned balls having different radii (Delft's balls data) and c) a noisy dataset with the Gaussian noise added to the distance between points in Euclidean space. By observing the spectrum of negative eigenvalues of the resulting Gram matrices and the additive constant required to render it metric, we identify three measures that can be used to characterise the above sources of negative eigenvalues. A variety of dissimilarity datasets are tested on the measures. Our analysis provides insight into the non-Euclidean behaviour of dissimilarity datasets and can be used to select appropriate embedding methods suited to the non-Euclidean data in hand.

Another secondary contribution of the paper is to develop a measure that assesses the contribution of each object to the mass of negative eigenvalues that provides further insight into the cause of non-Euclidean behavior. In this paper, we test a finer measure that assesses the contribution of each object to the mass of negative eigenvalues. In this way it is possible to determine whether the non-Euclidean artifacts are attributable to the dissimilarities associated with a few outlying objects or are uniformly distributed throughout the dataset.

2 Characterising the Causes of non-Euclidean Data

In this paper we are concerned with the sources of non-Euclidean data. Our overall aim is to identify the causes of a given set of non-Euclidean dissimilarity data so as to find out suitable correction methods to make them more Euclidean.

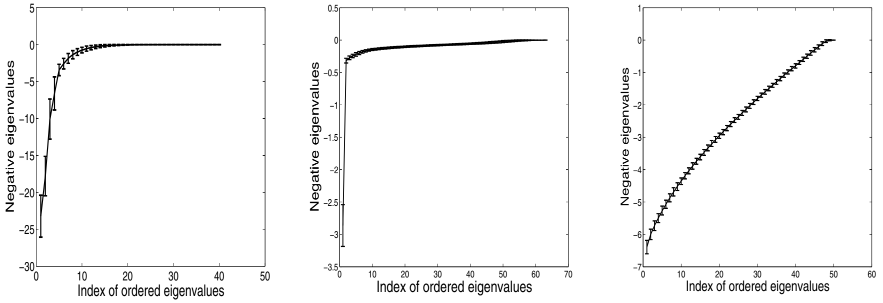
2.1 The Causes of non-Euclidean Data

We begin by identifying three reasons for non-Euclidean behaviour [2].

Manifold. If the data points reside on a curved manifold, then the distances between them are intrinsically non-Euclidean (but still metric). This is one possible source of non-Euclidean distances. Here we model such data as points on the surface of a sphere, a simple surface where distances are easy to compute. It is simple to simulate patches with various degrees of curvature that depart from Euclidean behavior by changing the curvature of the patch. The dissimilarity measurements on the sphere are metric but non-Euclidean.

Extended objects. If objects are not point-like but rather are extended in space, then the distances between them are measured between the closest points on their surface. As a result the distances will be non-Euclidean and possibly non-metric. Delft's balls data [2] is a typical example. Randomly positioned balls are generated with varying radius. The pairwise dissimilarities are the surface distances between the balls. As a result only the pairwise distances between balls with zero radius are Euclidean. It is also simple to modify the degree of non-Euclidean behaviour by adjusting the radii of the balls.

Gaussian noise. The final source is Gaussian noise added to the original Euclidean dissimilarities. This will generate data that is both non-Euclidean and non-metric.



(a) on the sphere (b) balls data (c) Gaussian noise

Fig. 1. The negative eigenvalues of the resulting Gram matrix of 100 points on the sphere, from extended objects and Gaussian noise as a function to the index of ordered negative eigenvalues

2.2 Negative Spectrum

We study with the three simple modes of the occurrence of non-Euclidean pairwise data. The Gram matrix of non-Euclidean dissimilarity data is indefinite, i.e. it has negative eigenvalues. One way to gauge the degree to which a pairwise distance matrix exhibits non-Euclidean artefacts is to analyse the properties of its centralised Gram matrix. For an $N \times N$ symmetric pairwise dissimilarity matrix D with the pairwise distance as elements, the centralized Gram matrix $G = -\frac{1}{2}JD^2J$,where $J = I - \frac{1}{N}11^T$ is the centering matrix and 1 is the all-ones vector of length N. The degree to which the distance matrix departs from being Euclidean can be measured by using the relative mass of negative eigenvalues or "negative eigenfraction " $F_{eigS} = \sum_{\lambda_i<0} |\lambda_i| / \sum_{i=1}^{N} |\lambda_i|$ [3]. This measure is zero when the distances are Euclidean and increases as the distance becomes increasingly non-Euclidean.

We commence by examining the negative spectrum of the Gram matrix under the three models. Figure 1 shows the non-Euclidean dissimilarities from the sphere and balls data-sets have spectrum which contain a strong negative component, with a concentration towards the low end of the spectrum. The non-Euclidean dissimilarities from Gaussian noise have a more slowly decreasing negative spectrum. Each of the negative spectrum appear to follow an exponential decay. Thus the slope and the intercept from an exponential fit should be able to discriminate at least the Guassian noise model from the remaining two models. An exponential curve of the form $y = ae^{bx}$ is fitted to the data, with b the slope and a the intercept. These two parameters are used as measures to characterise the negative spectrum.

2.3 Non-metricity

A distance measure is considered to be non-metric if it is either non-symmetric, negative or violates the triangle inequality. A dissimilarity matrix rarely satisfies the triangle inequality, but is usually positive [4]. Thus the violation of the triangle inequality is considered when measuring non-metricity. A constant $C = \max_{i,j,k} |d_{ij} + d_{ik} - d_{jk}|$ is computed and added to the off-diagonal elements of the dissimilarity matrix so as to increase the amount of data that satisfies triangle equality [3]. If C is zero, the pairwise dissimilarity is considered to be metric. Moreover, the dissimilarity values over the sphere are metric. Thus a, b and C can be used as three measures to identify the three modeled sources of non-Euclidean behavior.

2.4 Object's Contribution to the non-Euclidean Behaviour of Dissimilarities

If the non-Euclidean artefacts are created solely by the set of distances to a few outlying objects which are incorrectly placed, then it is possible to restore the data to a Euclidean state by editing these objects from the dataset. Based on this idea the notion of measuring the contribution of each object to the negative eigenfraction of a dissimilarity matrix is introduced.That is, the fraction given by the ratio of the sum of the negative distances originating from an individual object to each of the remaining objects, divided by the total.

The points can be embedded in Krein space as follows $Y = \sqrt{\Lambda}\Phi^T$ where Λ is the diagonal matrix with the ordered eigenvalues of centralised Gram matrix as elements and Φ is the eigenvector matrix with the ordered eigenvectors as columns. When the centered Gram matrix has negative eigenvalues then those dimensions of the embedding associated with negative eigenvalues are represented by imaginary numbers, and those associated with positive eigenvalues by real numbers. In other words, the data are embedded into a pseudo Euclidean or Krein space [5]. Under the embedding,the coordinate vector of point j is $y_j = (\sqrt{\lambda_1}\Phi_{1j}, ..., \sqrt{\lambda_i}\Phi_{ij}, \sqrt{\lambda_N}\Phi_{Nj})^T$. The contribution to the squared distance between two points k and e is

$$d_{ke}^2 = \sum_i (y_k(i) - y_e(i))^2 = \sum_i \lambda_i(\phi_{ik} - \phi_{ie})^2$$

The sum of negative squared distances and the sum of positive squared distances from point k to all the remaining points are:

$$d_{k-}^2 = \sum_{\lambda_i<0} \lambda_i \sum_{e\neq k} (\phi_{ik} - \phi_{ie})^2, \quad d_{k+}^2 = \sum_{\lambda_i>0} \lambda_i \sum_{e\neq k} (\phi_{ik} - \phi_{ie})^2$$

Thus the fraction of negative squared distances from point k is

$$f_{pneig} = \frac{|d_{k-}^2|}{|d_{k-}^2| + |d_{k+}^2|}$$

This measure is zero for all objects (or points) when the distances are Euclidean and non-zero for outlier objects. Thus the measure can be useful to identify whether the non-Euclidean is caused by the second sources.

3 Experiments

To model distances sampled from a manifold, we commence with 100 points uniformly distributed on the surface of a 3D sphere with unit radius. The spherical coordinates of an object are $x = (r\sin\theta\cos\phi, r\sin\theta\sin\phi, r\cos\theta)^T$ where r is the radius of the sphere, θ is the elevation angle($[0, \pi]$) and $\phi([0, 2\pi])$ is azimuth angle. The pairwise geodesic distances are computed as the lengths of great circle arcs between pairs of objects. We can use the tangent space projection and increase the radius or change the range of the elevation angle to control the extent to which the patches deviate from a Euclidean surface, i.e. the degree of non-Euclideanness in the dissimilarity matrix. In total 100 initial configurations of points are used.

To model the extended objects, we pick 100 randomly positioned points in a 7D hypercube with length 100, and we take each point as the center of a ball with radius $r(r \geq 0)$. The balls do not overlap. The pairwise distance is the Euclidean distances between the centers of two balls minus the radii of the two balls. We regard the balls with radius greater than 0 as non-Euclidean balls. We vary the fraction of non-Euclidean balls, and take the fraction to be 0.1, 0,3, 0.5, 0.7 or 0.9 in our experiments. The radii of the non-Euclidean balls are 2, 3 or 4. We also generate 100 balls with uniformly distributed radii ranging from 0 to 4.

To model the Gaussian noise, we commence with 100 randomly positioned points in a 3D Euclidean space and calculate the Euclidean dissimilarity matrix. Then we add Gaussian noise with zero mean and various values of standard deviation to the off-diagonal elements of the dissimilarity matrix to generate a non-Euclidean dissimilarity matrix. The value of the standard deviation of the Gaussian noise is 0.1, 0.3, 0.5, 0.7 and 0.9.

To ensure the results are comparable over the dissimilarity data in various ranges and scales, all of the dissimilarity metrics are scaled such that the average dissimilarity is unity. We calculate the negative eigenvalues of each dissimilarity matrix and fit the average negative spectrum by an exponential curve to obtain the slope b, the intercept a and the average metric constant C. The whole process is repeated for a sample sizes of 500 and 1000 points.

Figure 2 shows the slope b as a function of the metric constant value C from the non-Euclidean dissimilarities on the sphere, the "balls" data and Gaussian noise. As the negative spectrum of the Gram matrix from the Euclidean points with Gaussian noise appears to be in a flat and linear in shape, so the value of slope b is very small with a value around -0.04. For the dissimilarities from the extended objects, the negative spectrum has a very sharp decreasing negative tail (just few significant negative eigenvalue), so the value for the slope b has a larger magnitude. Comparing the points on sphere and the ball data, there are several negative eigenvalues in the tail and the decrease is less sharp. This may explain why the slope of the non-Euclidean dissimilarities on the sphere is intermediate between that of the Gaussian noise and the non-Euclidean balls data. Another interesting finding is that the number of objects is not correlated with the slope, especially for points on the sphere and Gaussian noise. In terms of the parameters the three sources of negative eigenvalues are well separated from each other.

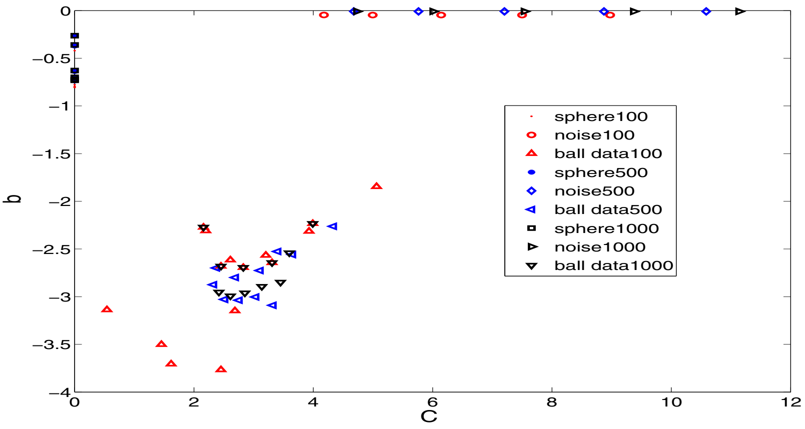
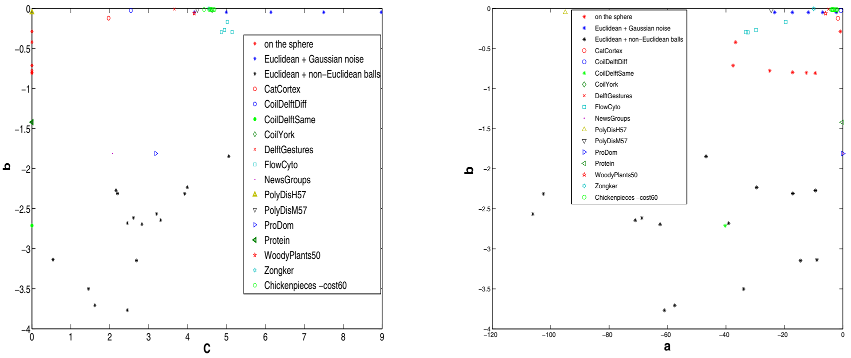


Fig. 2. The artificial non-Euclidean dissimilarity data caused by the manifold the data resides on, the extended objects and the Gaussian noise

We therefore use the above models to analyze a set of public domain dissimilarity data provided by the EU SIMBAD project consortium [2]. The Catcortex dataset contains dissimilarities based on the connection strengths between 65 cortical areas of the cat brain from four regions. CoilDelftDiff, CoilDelftSame and CoilYork are three dissimilariy datasets extracted from feature points detected in the COIL image database computed using different graph edit distances. FlowCyto contains four histogram dissimilarities for samples of breast cancer tissue. Newsgroups contains dissimilarities for messages in four classes of newsgroups. PolyDisH57 and PolyDisM57 are the dissimilarites of randomly generated polygons based on the standard and the modified Hausdorff distance. Protein contains the dissimilarities of protein sequences based on an evolutionary measure of distance. Woodyplants50 contains the shape dissimilarities between leaves of woody plants. Zongker contains the dissimilarities between handwritten digits based on deformable templates. Chickenpieces-cost60 contains 7 dissimilarity matrices from a weighed edit distance.



(a)

(b)

Fig. 3. (a)The slope b as a function of the metric constant C; (b)The slope b as a function of the intercept a

The left and right plots in Figure 3 respectively show the slope b and the intercept a as a function of the metric constant value C, for the artificial samples of 100 objects. The plots indicate that the non-Euclidean behaviour of Defilt-Gestures, PolyDisM57, Woodyplants50, Zongker, Chicken pieces, Catcortex and FlowCyto are likely to arise from Gaussian noise. On the other hand, the non-Euclidean behaviour of the Newgroups, ProDom and DelftSame datasets is likely to arise the non-Euclidean distances of a few outlying objects. We are unsure about the origin of the negative eigenvalues for the Protein and PolyDisH57 datasets. For PolyDisH57 the cause may be a combination of data residing on a manifold and the Gaussian noise. For the Protein dataset it may be a combination of data on the manifold and extended objects.

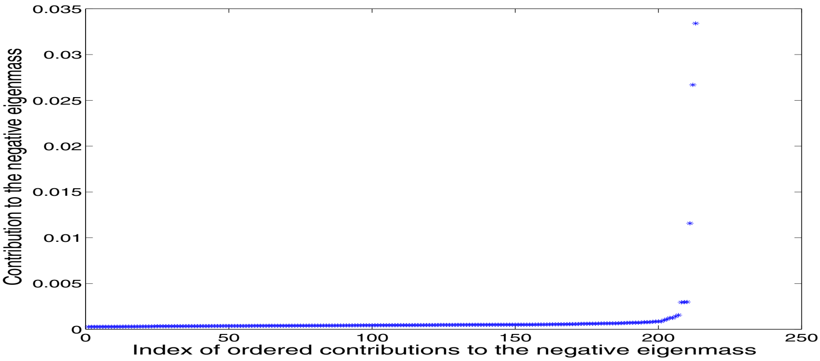


Fig. 4. Sorted each object's contribution to the negative eigenvalues at Protein

We plot the individual contribution to the negative eigenmass for the Protein dataset in Figure 4. This shows that the negative eigenvalues are caused by the non-Euclidean distances of just a few objects. The protein data is almost Euclidean with a very small negative eigenfraction value of 0.001. We have explored the effect of applying a leave one out nearest neighbor classifier to the dataset. When we edit out the effect of the outlier objects distances by adding a constant to the squared distances to the remaining objects, we obtain only a slightly smaller error rate of 0.47% compared to 1.9% for the original distances.

4 Conclusion

This paper discusses three possible sources of non-Euclidean behavior in dissimilarity data. We present three measures for analysing and determining the causes of negative eigenvalues in a non-Euclidean dissimilarity matrix. The three measures are based on distribution of the negative eigenvalues, and allow us to determine if the case is a) that data resides on a manifold, b)that the objects may be extended and c) that there is Gaussian noise.

Acknowledgments. Edwin R. Hancock is supported by the EU FET project SIMBAD and by a Royal Society Wolfson Research Merit Award.

References

1. Wilson, R., Hancock, E.: Spherical embedding and classification. In: SSPR (2010)
2. Duin, R.P.W., Pkekalska, E.: Non-Euclidean dissimilarities causes and informativeness. In: SSPR, pp. 324–333 (2010)
3. Pekalska, E., Harol, A., Duin, R., Spillmann, B., Bunke, H.: Non-Euclidean or Nonmetric Measures Can Be Informative. In: SSPR, pp. 871–880 (2006)

4. Lauba, J., Rothb, V., Buhmannb, J.M., Mllera, K.-R.: On the information and representation of non-Euclidean pairwise data. Pattern Recognition, 1815–1826 (2006)
5. Goldfarb, L.: A new approach to pattern recognition. Progress in Pattern Recognition, 241–402 (1985)
6. Sanfeliu, A., Fu, K.-S.: A Distance measure between attributed relational graphs for pattern recognition. IEEE Transactions on Systems, Man, and Cybernetics, 353–362 (1983)