Accepted Manuscript

Improved data visualisation through multiple dissimilarity modelling

Iain Rice

PII: S0020-0255(16)30555-2
DOI: 10.1016/j.ins.2016.07.073
Reference: INS 12399

To appear in: *Information Sciences*

Received date: 9 May 2016
Revised date: 22 July 2016
Accepted date: 28 July 2016

Please cite this article as: Iain Rice, Improved data visualisation through multiple dissimilarity modelling, *Information Sciences* (2016), doi: 10.1016/j.ins.2016.07.073

Improved data visualisation through multiple dissimilarity modelling

Iain Rice

Aston University, UK.

Abstract

Popular dimension reduction and visualisation algorithms rely on the assumption that input dissimilarities are typically Euclidean, for instance Metric Multidimensional Scaling, t-distributed Stochastic Neighbour Embedding and the Gaussian Process Latent Variable Model. It is well known that this assumption does not hold for most datasets and often high-dimensional data sits upon a manifold of unknown global geometry. We present a method for improving the manifold charting process, coupled with Elastic MDS, such that we no longer assume that the manifold is Euclidean, or of any particular structure. We draw on the benefits of different dissimilarity measures allowing for the relative responsibilities, under a linear combination, to drive the visualisation process.

Keywords: `Dissimilarity`, `Visualisation`, `Multidimensional Scaling`, `Euclidean`

1. Introduction

In order to create humanly interpretable visualisations, popular algorithms which chart high dimensional data assume some global or local geometric structure. Methods such as the Sammon map [25], Stochastic Neighbour Embedding (SNE) [12] and its' derivatives, the Gaussian Process Latent Variable Model [13], the Generative Topographic Map (GTM) [5], Metric Multidimensional Scaling (MDS) and Curvilinear Component Analysis (CCA) [7] assume global Euclidean structure. The work of [27, 30, 29] removes the metric constraint in MDS and CCA using the Bregman divergence, however the Euclidean dissimilarity over inputs is retained. Typically observation manifolds are better characterised by Riemannian manifolds than the restrictive Euclidean spaces [2]. These manifolds can be assumed to have locally Euclidean structure, as in Locally Linear Embedding [24], Laplacian Eigenmaps [4], Riemannian Manifold Learning [18] and methods using geodesic distances based upon local Euclidean structure such as Isomap [31], the Geodesic Nonlinear Map [17] and Curvilinear Distance Analysis [16].

Email address: `i.rice@aston.ac.uk` (Iain Rice)

These methods relying on a local Euclidean structure, assuming smooth continuity between local charts. This assumption can lead to a poor approximation of the manifold chart when observations are complex non-Euclidean structures or have a fractal dimensionality. Further to this the local estimates of a chart in high dimensional space can be unreliable, particularly when data is sparse. These methods are also sensitive to the choice of the neighbourhood size parameter, potentially leading to false neighbourhoods in the mapping of data.

In particular we can consider the case of GTM which assumes the observations are distributed according to an isotropic Gaussian, akin to a hypershpere, which in high dimensions contains nearly all of its' mass in a thin shell sitting upon the surface (see [15] for an overview of this). This particular case highlights the need for a more thorough approach to characterising reliable dissimilarities over observations in the generation of visualisations.

The combination of different dissimilarity operators has recieved much interest in recent years, particularly in the area of multiple kernel learning. The use of these multiple kernels allows for measures inducing diverse topologies upon observations to warp high dimensional data, often with complex structure, in order to achieve better regression and classification performance, outlined in [6, 10], as well as in the field of manifold learning [1]. The tasks of regression and classification are supervised by nature in that true targets exist. In this paper we consider the unsupervised case where the targets are learned by minimisation of a mapping cost function, such that the data structure is preserved. One such case where a non-Euclidean dissimilarity measure was used to characterise observations is described in [20] where class label information was used as a descriptor to motivate the visualisation process. We present an extension to this work where we no longer rely on a single dissimilarity measure. Similar work was performed in [19] relying on linear discriminant analysis to form visualisations, however the learned mapping is linear and the optimisation method does not generalise to nonlinear mappings as in this paper.

This paper details a method for incorporating several dissimilarity measures into a visualisation framework. By learning a mixture representation for the observation space the learned visualisation is more interpretable and better spans the latent dimensions than the Euclidean counterpart. It was found in [15] that visualisation mappings which learn the position of latent points through optimisation of a non-convex cost function (such as MDS, SNE and CCA) perform better than those whose cost functions are convex (such as PCA, LLE and Isomap). SNE, CCA and their variants require tuning of the mapping parameters such as the perplexity and neighbourhood parameters in CCA and Culrvilinear Distance Analysis. The more traditional approaches of the Sammon map and MDS require no such tuning. Despite constructing a topographic mapping, the local focus of the Sammon mapping is improved by using Elastic MDS as shown in [27]. Further local focus can be achieved with other variants of MDS, however the optimisation of the resulting cost functions requires stochastic gradient descent and poor local minima are likely in the optimisation procedure. We therefore focus on the specific case of Elastic MDS to show the potential improvements in this paper, though the approach does generalise

to the other methods mentioned. Five standard datasets are analysed with the proposed mixture dissimilarity approach qualitatively compared to the standard MDS method. Following [33] quantitative quality measures typically used for visualisations are not appropriate so we rely on a visual comparison to show improvement.

2. The Learning Task

The method described in this paper is born from the desire for robust dimension reduction, however we focus on the particular case of visualisation where the mapped data is 2-dimensional. Elastic MDS [22] embeds a dataset, X, into a reduced dimensional space, $Y \in \mathbb{R}^V$. In the experiments of this paper we fix $V = 2$. X can be nonvectorial as all that is required for the projection to Y is a matrix of pairwise relative dissimilarities, $D_x(i, j)$, between observations X_i and X_j. The latent points corresponding to each observation, denoted $\boldsymbol{y}_i$, are learned through gradient descent of the cost function:

$$E = \sum_{i,j<i} \frac{\left(D_x(i,j) - D_y(i,j)\right)^2}{\left(D_x(i,j)\right)^2}, \tag{1}$$

where $D_y(i, j)$ denotes the dissimilarity between visualised points $\boldsymbol{y}_i$ and $\boldsymbol{y}_j$, taken as standard to be the Euclidean distance. The term Elastic MDS is due to the quadratic factor in the denominator of equation (1), as opposed to that of the Sammon map. This has the effect of forcing the mapping process to focus more on local than on global distance preservation. We naturally desire to observe a physically motivated clustering of observations in a visualisation without imposing this in the mapping procedure.

In order to remove the restrictive assumption that D_x consists of Euclidean distances only we treat the measure as a weighted combination of multiple separate dissimilarity measures:

$$D_x = \sum_l \alpha_l D_x^l$$

where α_l is the weight corresponding to the l-th dissimilarity measure. To construct a more flexible input dissimilarity matrix we utilise 15 dissimilarity measures for the experiments in this paper. Our approach is flexible such that additional measures can be included and removed if desired. Table 1 taken from [23] lists 14 of the dissimilarities and the final measure is the geodesic distance given by Dijkstra's algorithm [8] as in Isomap and other related algorithms. In this paper we restrict our discussion and experiments to the case where observations are vectorial. However, the approach of learning a mixture of input dissimilarities is generic and a trivial change of the measures of table 1 to other measures allows for analysis of other data structures. Typical non-vectorial examples include probability distributions, binary data, images, graphs and time series for instance. Dissimilarity measures specific to these applications are detailed in [23].

	Measure	Dissimilarity - $d(\mathbf{x}, \mathbf{y})$	M	E
1	Euclidean	$\sqrt{(\mathbf{x}-\mathbf{y})^T(\mathbf{x}-\mathbf{y})}$	Yes	Yes
2	Weighted Euclidean	$\sqrt{(\mathbf{x}-\mathbf{y})^T diag(w_i^2)(\mathbf{x}-\mathbf{y})}$	Yes	Yes
3	City block	$\sum_{i=1}^m \lvert x_i - y_i \rvert$	Yes	No
4	Max norm	$\max_i \lvert x_i - y_i \rvert$	Yes	No
5	Minkowski (l_p)	$(\sum_{i=1}^m \lvert x_i - y_i \rvert^p)^{\frac{1}{p}}$, $p \geq 1$, $p \neq 2$	Yes	No
6	Mahalanobis	$\sqrt{(\mathbf{x}-\mathbf{y})^T C^{-1}(\mathbf{x}-\mathbf{y})}$, C psd	Yes	Yes
7	Median distance	$\text{median}_i(\lvert x_i - y_i \rvert)$	No	No
8	Correlation based (D_{corr})	$\frac{1}{2}\left(1 - \frac{\mathbf{x}^T\mathbf{y}}{\lVert\mathbf{x}\rVert^2 + \lVert\mathbf{y}\rVert^2}\right)$	No	No
9	Correlation based (D_{corr2})	$\frac{1}{2}\left(1 - \frac{\mathbf{x}^T\mathbf{y}}{\lVert\mathbf{x}\rVert^2 + \lVert\mathbf{y}\rVert^2 - 2\mathbf{x}^T\mathbf{y}}\right)$	No	No
10	Cosine	$\frac{1}{2}\left(1 - \frac{\mathbf{x}^T\mathbf{y}}{\lVert\mathbf{x}\rVert\lVert\mathbf{y}\rVert}\right)$	No	No
11	Divergence	$\sqrt{\sum_{i=1}^n \frac{(x_i - y_i)^2}{(x_i + y_i)^2}}$	No	No
12	Bray and Curtis	$\frac{\sum_{i=1}^n \lvert x_i - y_i \rvert}{\sum_{i=1}^n x_i + y_i}$	No	No
13	Soergel	$\frac{\sum_{i=1}^n \lvert x_i - y_i \rvert}{\sum_{i=1}^n \max\{x_i, y_i\}}$	No	No
14	Ware and Hedges	$\sum_{i=1}^n \left(1 - \frac{\min\{x_i, y_i\}}{\max\{x_i, y_i\}}\right)$	No	No
15	Geodesic	$\delta(D_{Euc})$	No	No

Table 1: Numbered dissimilarity measures between vectors $\mathbf{x}, \mathbf{y}$ listing whether they are metric (M) and Euclidean (E).

For the experiments described in this paper the weighted Euclidean (measure 2) parameters were fixed to be the inverse of the sample mean vector and p in the Minkowski distance (measure 5) was fixed to be 1.2 to induce a metric between the city block and Euclidean measures. The weighting matrix, C, in the Mahalanobis distance (measure 6) is the data sample covariance matrix. These choices of parameters allow the measures to be invariant to the absolute scaling of the data in each of the observed dimensions such that a single dimension does not disproportionately affect the overall dissimilarity. The geodesic distance (measure 15) is performed over the Euclidean distance as is typical in the literature.

We propose an iterative learning scheme such that the α_l weights are optimised on a slower timescale to the visualised points, $\boldsymbol{y}_i$. The latent points can be learned using gradient descent over the standard cost function of equation (1) using scaled conjugate gradients, as in section 3, or a quasi-Newton approach

with the following gradients:

$$\frac{\partial E}{\partial \boldsymbol{y}_i} = 4 \sum_j \left(-\frac{1}{D_x(i,j) D_y(i,j)} + \frac{1}{(D_x(i,j))^2)} \right) (\boldsymbol{y}_i - \boldsymbol{y}_j) \,.$$

The weight parameters can also be learned through gradient descent of equation (1). Due to the requirement for normalisation of the weights we opt to learn $\tilde{\alpha}_l$ to recover the weights $\alpha_l = \tilde{\alpha}_l / \sum_m \tilde{\alpha}_m$ as follows:

$$\frac{\partial E}{\partial \tilde{\alpha}_l} = \frac{\partial E}{\partial D_x(i,j)} \frac{\partial D_x(i,j)}{\partial \tilde{\alpha}_l} = \left(\frac{-2(D_y(i,j))^2}{(D_x(i,j))^3} + \frac{2D_y(i,j)}{(D_x(i,j))^2} \right) \left(\frac{\partial D_x(i,j)}{\partial \tilde{\alpha}_l} \right).$$

Through use of the product rule we can write the final term as:

$$\frac{\partial}{\partial \tilde{\alpha}_l} \left[\sum_l \frac{\tilde{\alpha}_l}{\sum_m \tilde{\alpha}_m} D_x^l(i,j) \right] = \left(\frac{1}{\sum_m \tilde{\alpha}_m} D_x^l(i,j) - \frac{1}{(\sum_m \tilde{\alpha}_m)^2} \sum_l \tilde{\alpha}_l D_x^l(i,j) \right).$$

Combining these two parts we find the gradients with respect to the weights:

$$\frac{\partial E}{\partial \tilde{\alpha}_l} = \sum_{i,j} \left(\frac{-2(D_y(i,j))^2}{(D_x(i,j))^3} + \frac{2D_y(i,j)}{(D_x(i,j))^2} \right) \left(\frac{D_x^l(i,j)}{\sum_m \tilde{\alpha}_m} - \frac{\sum_l \tilde{\alpha}_l D_x^l(i,j)}{(\sum_m \tilde{\alpha}_m)^2} \right).$$

It should be noted that α_l can be negative as well as positive, and indeed greater than 1, allowing for a more flexible model than restricting α_l to between 0 and 1.

Since we are optimising over the MDS cost function of equation (1) we should note the obvious characteristic that a minimum occurs when $D_y(i,j)$ matches $D_x(i,j)$ as closely as possible, i.e. when the elements $D_x(i,j)$ can be reproduced in $D_y(i,j)$. In standard MDS the only way to achieve this is by manipulating the latent points, $\boldsymbol{y}_i$. Our approach allows for the modification of $D_x(i,j)$ in order to further minimise the mapping error. The learning task has therefore shifted from not only identifying the latent co-ordinates which minimise the equation (1), but which mixture of dissimilarity measures generates the $D_x(i,j)$ which is most re-producible by $D_y(i,j)$, subject to $D_y(i,j) = \|\boldsymbol{y}_i - \boldsymbol{y}_j\|_2$.

Naturally optimising α_l with respect to the mapping STRESS is one of several ways to determine the parameters. An alternative way to pose the problem occurs when we consider combining the mapping with a feed-forward network as in the NeuroScale [21] approach to information visualisation. In this setting the set of latent points are learned at each gradient step and the feed-forward mapping weights, $\boldsymbol{W}$, are found through the iterative shadow-targets approach [32]. When the network prototypes span the entire observation space this can be considered as the solution of:

$$\phi(D_x)\, \boldsymbol{W} = \boldsymbol{Y},$$

where the weight matrix $\boldsymbol{W}$ is learned by pseudoinverse and ϕ is a nonlinear basis function. Once the learning problem is posed this way it is clear that α_l

Algorithm 1 Pseudocode for multiple dissimilarity learning in Elastic MDS.

Require: Dissimilarities $D_x^l(i,j)$,
Initialise $\tilde{\alpha}_l$ randomly or by the method of section 3.5 and $\alpha_l = \tilde{\alpha}_l / \sum_m \tilde{\alpha}_m$,
Initialise $D_x = \sum_l \alpha_l D_x^l$
Initialise $\boldsymbol{y}_i$ by kernel PCA or randomly.
Calculate latent dissimilarities $D_y(i,j) = \|\boldsymbol{y}_i - \boldsymbol{y}_j\|_2$,
Calculate initial error E from equation (1),
for epochs = 1:Nepochs **do**
    for iter = 1:Niter **do**
        Calculate error gradients $\frac{\partial E}{\partial \boldsymbol{y}_i}$,
        Perform a gradient step for $\boldsymbol{y}_i$ with SCG to give $\bar{\boldsymbol{y}}_i$,
        Re-calculate latent dissimilarities $D_y(i,j) = \|\bar{\boldsymbol{y}}_i - \bar{\boldsymbol{y}}_j\|_2$,
        Re-calculate mapping error E,
        if error reduced **then**
            Update $\boldsymbol{y}_i \leftarrow \bar{\boldsymbol{y}}_i$,
        end if
    end for
    Perform a gradient step of $\tilde{\alpha}_l$ with $\frac{\partial E}{\partial \tilde{\alpha}_l}$ with SCG to give $\tilde{\alpha}_l'$,
    Re-calculate input dissimilarities $D_x = \sum_l \alpha_l' D_x^l$,
    Re-calculate mapping error E,
    if error reduced **then**
        Update $\tilde{\alpha}_l \leftarrow \tilde{\alpha}_l'$
    end if
end for

can be learned in one of the many ways from the field of multiple kernel learning (see [10, 34] for an overview). A natural way of using these multiple dissimilarity measures is the incorporation into kernel PCA [26] as this naturally operates on kernels (which may in fact be the weighted sum of a number of different kernels), however it is well known that PCA spreads data in the direction of maximal variance and is therefore an inferior method of generating reliable, structure-preserving mappings compared to Metric MDS.

In this context there are many different ways of optimising the mixing weights in both convex and non-convex ways to improve the regression performance. For the purpose of generating a fixed visualisation in this paper however we choose to learn α_l though the gradient descent procedure detailed above by minimising the MDS cost function as by design this will improve visualisation quality. The pseudocode for the algorithm used to obtain the results of section 3 is shown in algorithm 1. For the experiments performed for this paper Nepochs was fixed to 10 and Niter was fixed to 75, however with α_l initialised as outlined in section 3.5 Nepochs and Niter can be changed to 3 and 150 respectively reducing the training time whilst achieving the same visualisations. Scaled conjugate gradients (SCG) was used to update the latent positions, $\boldsymbol{y}_i$, and the input mixture weights, α_l.

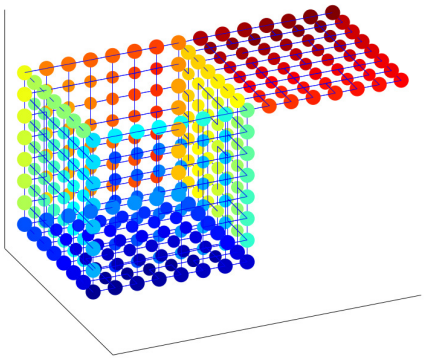


Figure 1: Open box dataset in 3-dimensional observation space. The structure contains an open top which typically poses problems for visualisation algorithms. The regular spacing between points in the box sides should be preserved in the visualisations.

3. Experimental Results

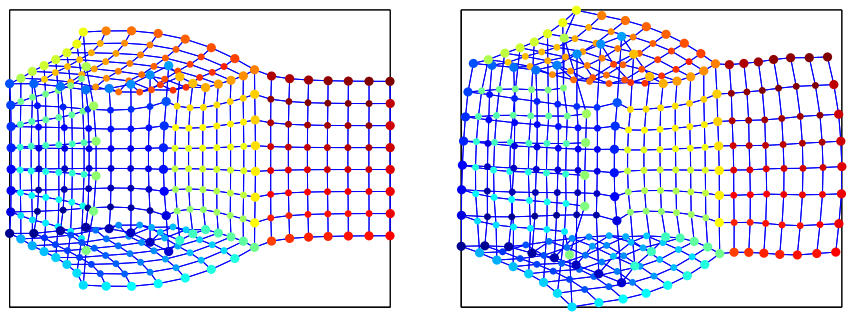
In order to compare the impact of learning the input dissimilarities with that of standard Elastic MDS we generate visualisations of five datasets using the standard version of Elastic MDS and the modified version proposed in this paper for comparison. Experiments were run five times, using the above iteration and epoch setups in algorithm 1, with random initialisations of the latent points to avoid poor local minima. In each case the models arrived at the same latent space indicating a minima of the stress measure was achieved.

For comparison we also include visualisations computed by Isomap, both assuming a geodesic dissimilarity over Euclidean distances (as is standard in the literature) and a geodesic over the learned D_x by the Elastic MDS case for each dataset. In both cases the neighbourhood parameter, k, is fixed to the smallest neighbourhood size for which a connected graph is achieved.

3.1. Open Box Dataset

Firstly we consider the Open Box dataset [28] which is used as a benchmark comparison for linear and nonlinear visualisation algorithms in [15] containing an open top and uniformly sampled faces with 316 datapoints shown in figure 1.

The two visualisations in figure 2 are similar, due to the fact that the structure consists of six connected Euclidean planes. The modified inputs allow for



(a) (b)

Figure 2: Visualisations of the open box dataset with (a) standard Elastic MDS and (b) Elastic MDS with input learning.

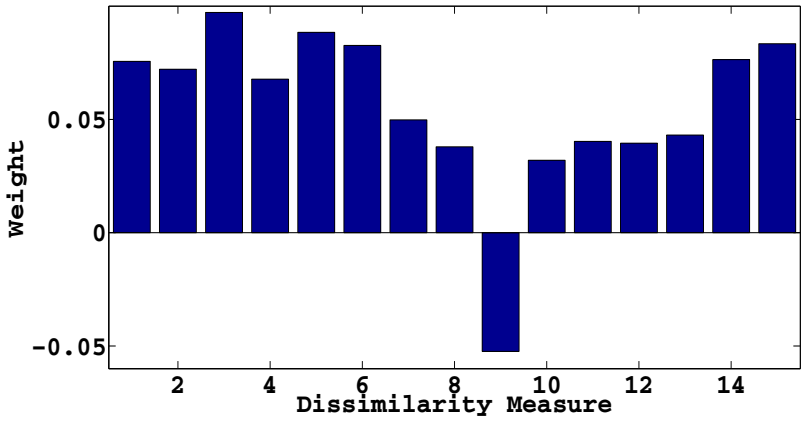
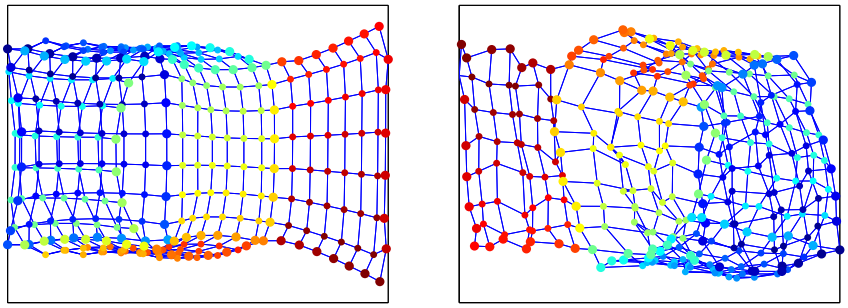


Figure 3: Weighting of the dissimilarity measures for the Open Box mapping.



(a) (b)

Figure 4: Visualisations of the open box dataset with (a) Isomap and (b) Isomap with learned input dissimilarities.

the visualisation in figure 2b to curve the lid and sides more than the standard visualisation in figure 2a, allowing for a further unfolding of the box. The weights, α_l, corresponding to each of the dissimilarity measures listed in table 1 are shown in figure 3. Interestingly despite the box being a Euclidean structure there is as much emphasis given to the Euclidean measures as for the non-Euclidean measures. The second correlation distance (measure 9) is given a negative weight, in part to cancel out the similar effects of the first correlation distance (measure 8) as the dissimilarities are constant for many points along each box face contrary to the other measures.

Figure 4 shows the visualisations of the Open box using Isomap with neighbourhood sizes $k = 4 and k = 14$ for the standard and learned input mappings respectively. When the learned input dissimilarities, D_x, are used as inputs the box is a deformed version of the standard version in figure 4a). The embedding generated by preservation of inner products in Isomap is not as effective as in the MDS case of figure 2b) where the box appears to be perturbed by random noise. In this case the visualisation generated by Isomap with a mixture input dissimilarity is inferior to the standard Isomap mapping. It should be noted that this structure is an entirely artifical dataset and the mappings generated in the following experiments are more representative of real-world observations.

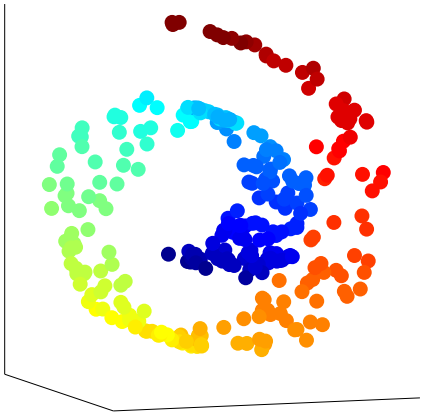


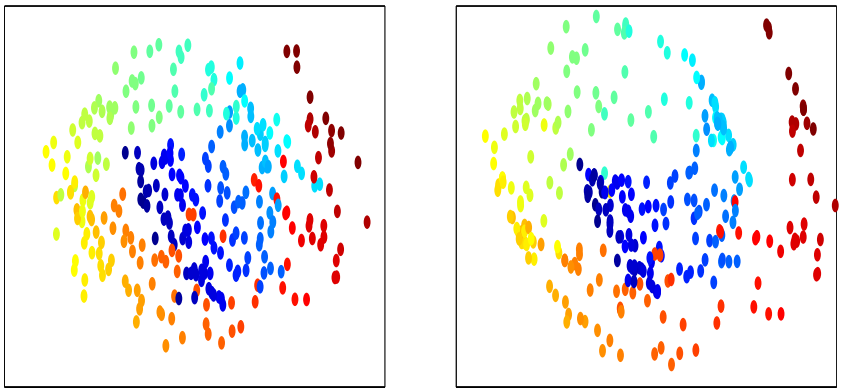
Figure 5: Swiss roll dataset in 3-dimensional observation space. The structure contains dark blue points at the centre (the left of the 2-dimensional latent rectangle) and dark red points at the exterior of the curve (the right of the 2-dimensional latent rectangle).

3.2. Swiss Roll

The second dataset used in this paper is the swiss roll, a benchmark for visualisation algorithms studied extensively in [15]. The 3-dimensinoal manifold consists of 300 points randomly sampled from a 2-dimensional rectangular grid mapped into the roll.

Figure 6 shows the two visualisations of the swiss roll generated by Elastic MDS. When the input dissimilarities are learned, achieving the visualisation of figure 6b), the degree of overlap between the blue and orange points is lessened, better preserving the local neighbourhood structures than the mapping of figure 6a). The learned weights for this toy example are shown in figure 7. The dominating dissimilarity measure is the Bray and Curtis, followed by the second correlation based measure. The normalisation terms in both of these measures have allowed the mapping to focus more on local neighbourhoods than the more separated points.

The visualisations of the swiss roll generated by Isomap are shown in figure 8 with neighbourhood sizes of $k = 6 and k = 10$ for the standard and multiple dissimilarity input methods respectively. The standard Isomap visualisation of figure 8a) attempts to unfold the structure, however the 2-dimensional rectangle is not recovered and many points are sat atop one-another. When the learned mixture of dissimilarities, as in figure 6b), is used with Isomap the latent representation of figure 8b) is obtained. This latent representation more closely matches the 2-dimensional rectangle than the Euclidean Isomap coun-



(a) (b)

Figure 6: Visualisations of the swiss roll dataset with (a) standard Elastic MDS and (b) Elastic MDS with learned input dissimilarities.

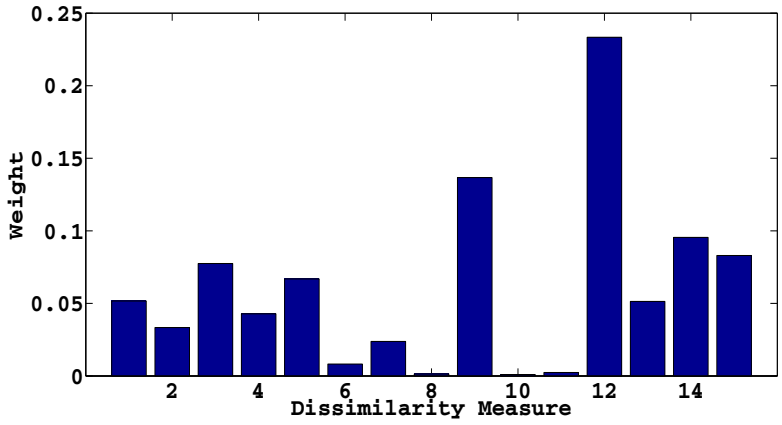
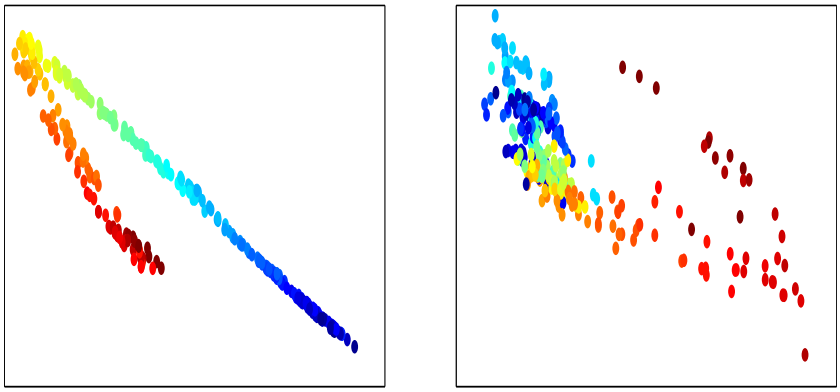


Figure 7: Weighting of the dissimilarity measures for the swiss roll mapping.



(a) (b)

Figure 8: Visualisations of the swiss roll dataset with (a) standard Isomap and (b) Isomap with mixture of input dissimilarities.

terpart. Neighbourhoods are better preserved in this visualisation, however the Elastic MDS mapping of figure 6b) provides a human observer with a clearer understanding of the observed manifold.

3.3. MNist Dataset

The next dataset we consider is a subset of the MNist greyscale digits dataset [14] containing 50 '0's, '1's and '6's. The images are 28×28 pixels and therefore analysed as 784-dimensional vectors. Of the three classes it is clear that the '6's share both the circular structure of the '0's and the straight line of the '1's, sitting between the two classes in the observation space. Figure 10a) shows the visualisation generated by standard Elastic MDS with the images plotted atop the latent points, $\boldsymbol{y}_i$. There is a clear separation of the '1's from the remaining classes, however there are '0's which have clearly been removed from their neighbourhoods in the mapping process and are then surrounded by '6's and vice versa. When the input dissimilarities are varied we obtain the mapping of figure 10b). The visualisation space is more continuous, resembling a filled circle with the class of '1's on the right hand side and now fully connected to the class of '6's in the centre with the '0's on the left side. The individual '6's and '0's at the top of the space are the more deformed with unusual slants and disconnected circles in the '0's. In addition to this the '1's located in this region are more slanted than the rest of the class and contain curves which do not match the standard caligraphy of the digit. The '0's which are placed in the centre of the space amid the group of '6's now share more in common with their latent neighbours, unlike those of figure 10a) such as the thick lined and non-circular '0's. The latent dimensions in figure 10b) clearly span the

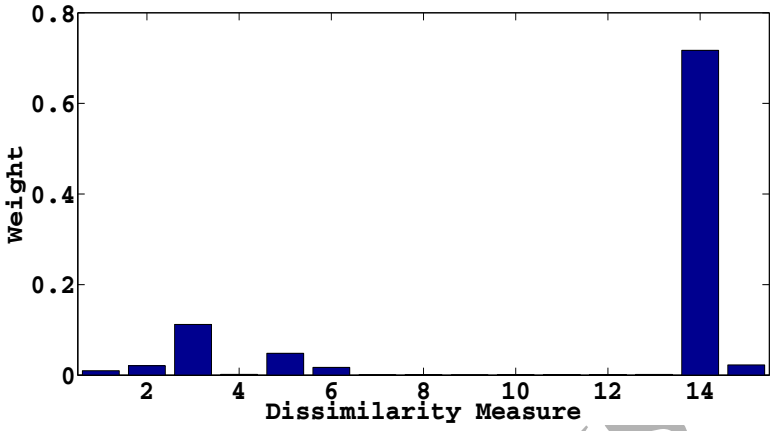
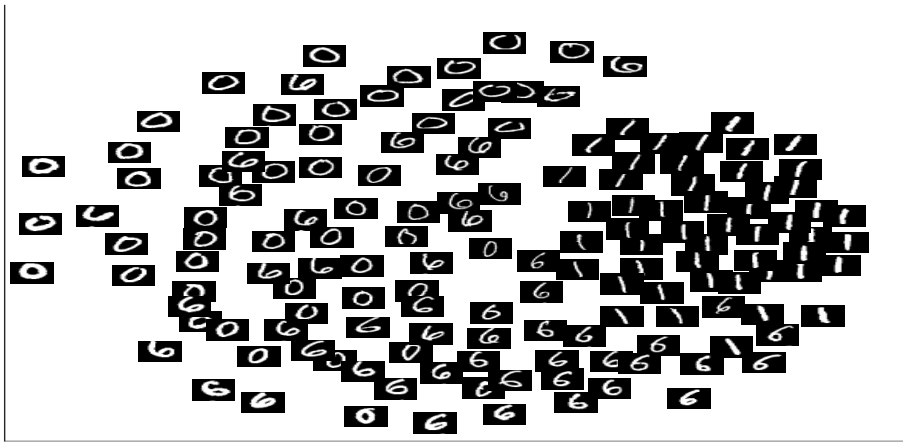


Figure 9: Weighting of the dissimilarity measures for the MNist mapping.

digit thickness in the y-axis (from thin to thick to thin) and digit orientation in the $y = x$-axis. The three digits covered in this experiment are estimated to possess dimensionality between 7 and 12 [11] so there are naturally more latent dimensions than can be conveyed than for $Y \in \mathbb{R}^2$, however a visualisation should present the information of these variables in an interpretable way. The visualisation of figure 10a) has not clearly conveyed either of these important latent features. The learning of the input dissimilarity has allowed for a more visually appealing and intuitive mapping than the standard case.

The weights allocated to each of the dissimilarity measures are shown in figure 9 where the Ware and Hedges dissimilarity is the most dominant, combined largely with the City block and Minkowski distances. On first inspection it would appear this is likely caused by the fact that the minimum and maximum measures in the Ware and Hedges dissimilarity measure in each region are a sufficient statistic to characterise the largely black and white images. This, however, is not the only possible cause and will be discussed further in section 3.5.

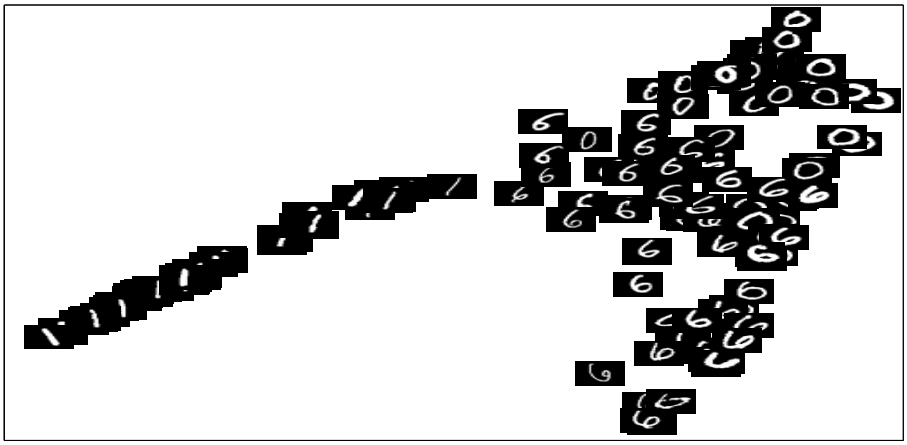
Figure 11 shows the visualisations of the MNist dataset using Isomap with neighbourhood sizes $k = 15 and k = 28$ for the standard and multiple dissimilarity methods respectively. In the standard Isomap visualisation of figure 11a) the class of '1's are fixed to a linear structure with left bends appearing on the left of the cluster and '1's with a right curve at the top on the right. There is however no clear separation between thick and thin '1's. The remaining two classes overlap with no clear distinction or structure between bold characters. On the other hand the visualisation of figure 11b) with a mixture of input dissimilarities concentrates the bold '6's and '0's in the region connecting the two clusters. The class of '1's is also more structured than in the standard Isomap case. The mapping here is not as informative as either of the Elastic MDS cases, however it is clear that a learned input dissimilarity framework generates a more



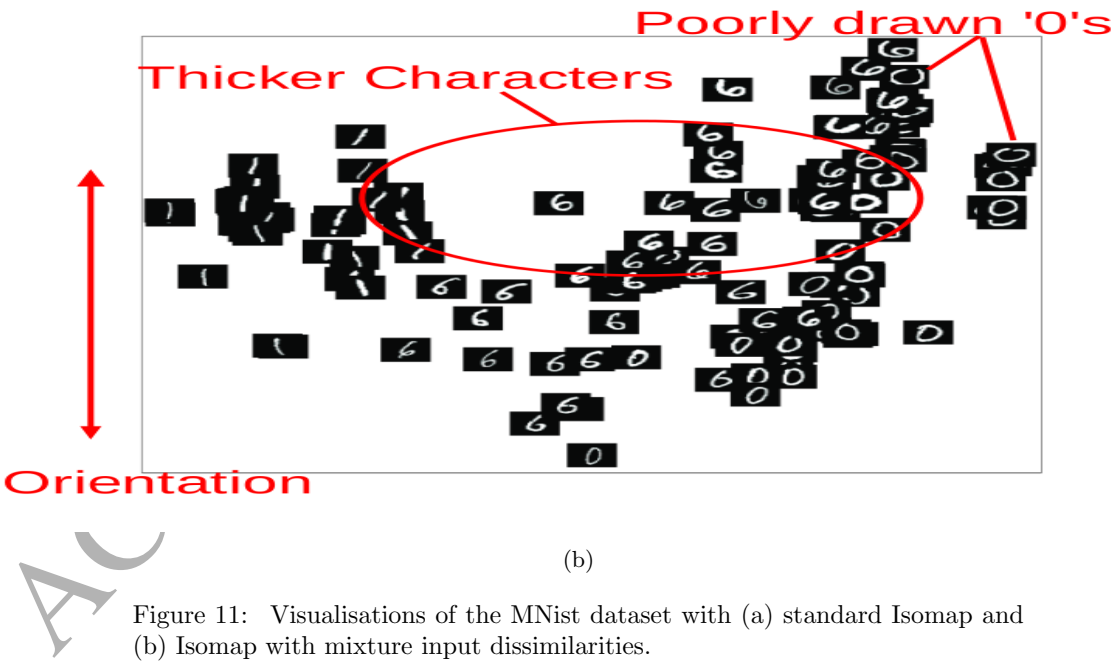
(a)

(b)

Figure 10: Visualisations of the MNist dataset with (a) standard Elastic MDS and (b) Elastic MDS with input learning.



(a)



(b)

Figure 11: Visualisations of the MNist dataset with (a) standard Isomap and (b) Isomap with mixture input dissimilarities.

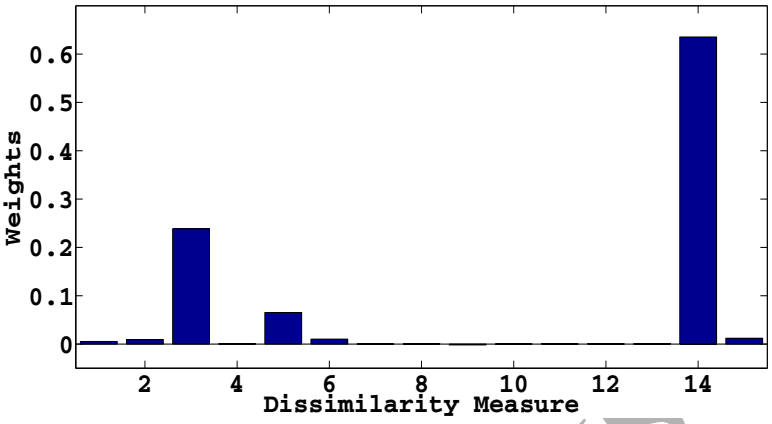


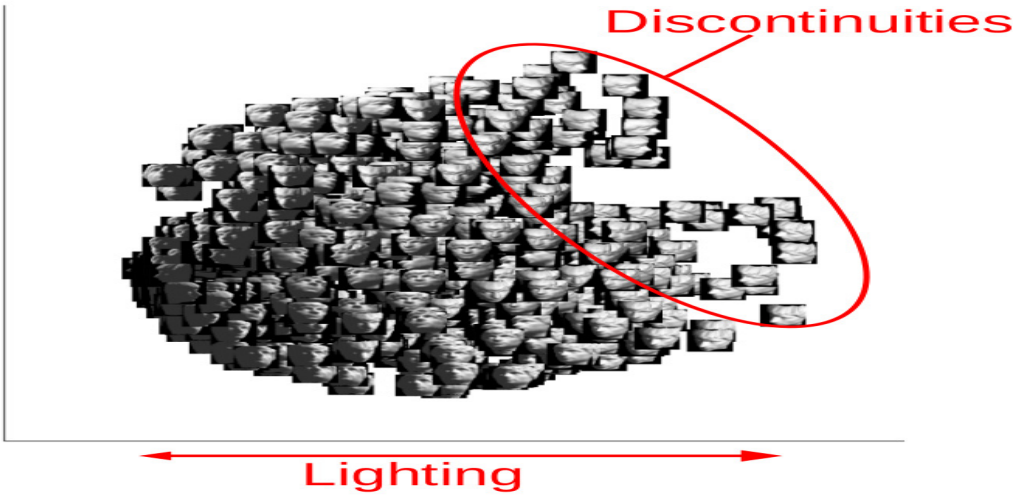
Figure 12: Weighting of the dissimilarity measures for the faces mapping.

intuitive and interpretable representation.

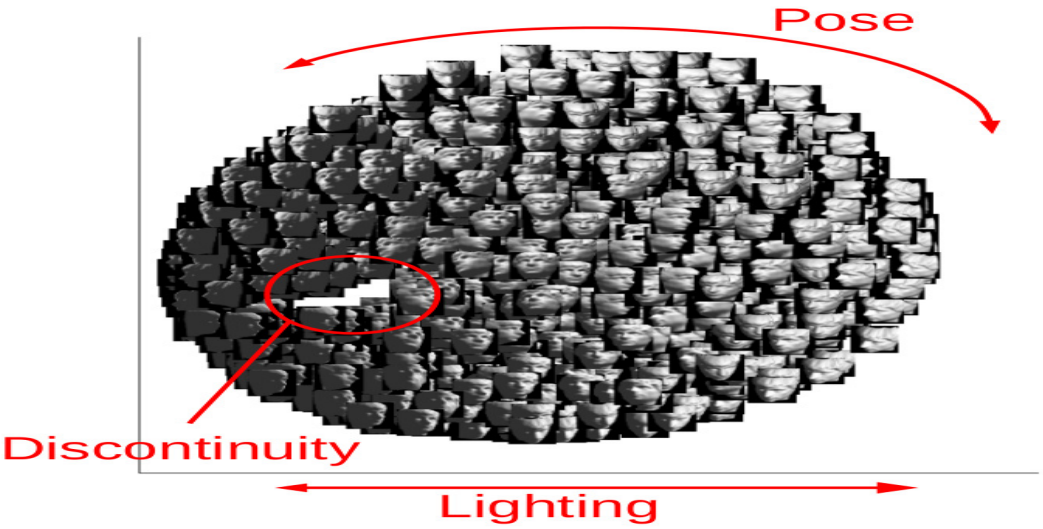
3.4. Artificial Faces Dataset

The fouth experiment is the artificial faces dataset [31] originally demonstrated with Isomap. There are 698 images with varying light levels and camera orientation on the 64×64 pixel images, therefore treated as 4096-dimensional observation vectors. Grouping in the latent space should be based upon both factors which make up the dataset in a continuous fashion as the variations between neighbourhing images in the observation space are small. Figure 13a) shows the standard Elastic MDS visualisation space where the images are ordered from low light levels on the left to higher light levels on the right. The relative clustering of images with similar camera orientations is not well preserved as the neighbourhoods are broken, with many clear outliers. In addition to this the mapping is not continuous as would be expected, with a series of similar images removed from the main cluster on the right side. On the other hand the visualisation generated with learned input dissimilarities in figure 13b) is a more continuous mapping. Moving from left to right in the visualisation space again marks a change in the light levels however the microclusters of camera orientation are more intuitive now, with higher camera angles associated with the top and right-most areas of the visualisation space and the more central orientations in the centre of the latent space. The optimised weights are shown in figure 12. As with the MNist mapping the most dominant weight is allocated to the Ware and Hedges dissimilarity, however more weight is given to the City block and Minkowski dissimilarities.

Figure 14 contains the visualisations generated by Isomap with neighbourhood sizes $k = 8$ and $k = 16$ for the standard and multiple dissimilarity approaches respectively. For the standard Euclidean distances as inputs, shown in figure 14a), the latent representation contains discontinuities at the top right



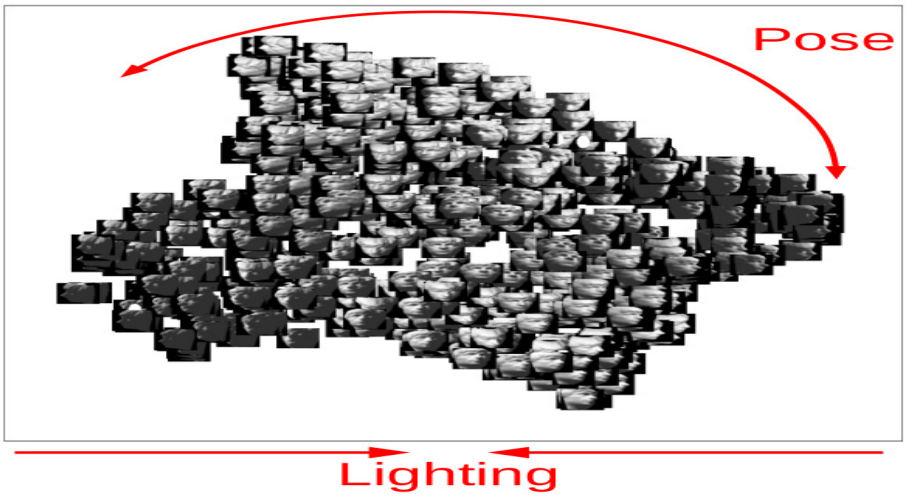
(a)



(b)

Figure 13: Visualisations of the artificial faces dataset with (a) standard Elastic MDS and (b) Elastic MDS with input learning.

(a)



(b)

Figure 14: Visualisations of the artificial faces dataset with (a) standard Isomap and (b) Isomap with learned inputs.

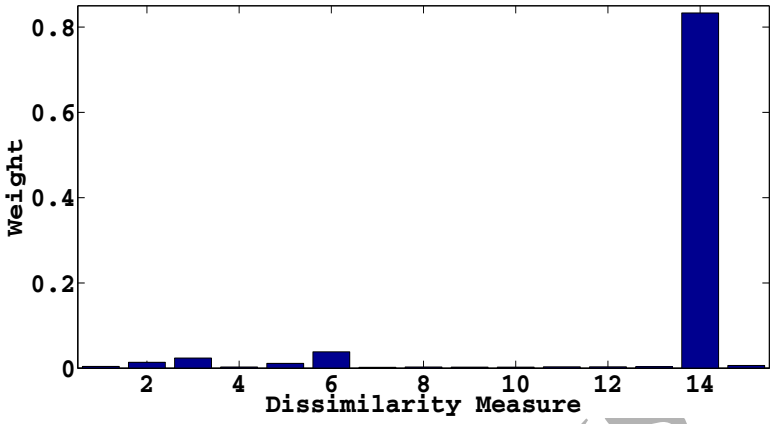


Figure 15: Weighting of the dissimilarity measures for the Caltech101 mapping.

and left corners. The latent variable of pose is in general preserved with front facing images in the centre of the visualisation, however the lighting levels in the images are not so clearly mapped. When the mixture of input dissimilarities is used as input to Isomap of figure 14b), the latent variable of lighting is mapped from dark to light at the centre of the latent space. The orientation of the face in the images moves from left-facing to right facing across the x-axis. As with the Elastic MDS case this representation is more visually appealing.

3.5. Caltech101 Images Dataset

The final dataset we use in this paper is the Caltech101 computer vision dataset [9]. This consists of multiple images from 101 different categories. We focus on a subset of these images, using 52 aeroplanes, 52 Bonsai trees and 52 dollar bills for our visualisations. Following the creation of a SURF bag-of-words [3] analysis we map the 156 feature vectors (500 features from the bag were used for each image) into a 2-dimensional visualisation space.

Figure 16 shows the latent space with the corresponding image plotted atop the latent point, $\boldsymbol{y}_i$ as with the previous two datasets. The space appears filled, however there is a cluster of aeroplanes at the bottom of the visualisation, with Bonsai trees at the top and left side apart from a small number of outliers. The dollar bills are split into a main cluster on the right side and a smaller cluster which sits directly between the aeroplanes and Bonsai trees where there is a clear region of overlap between all three classes.

The visualisation generated with learned input dissimilarities is shown in figure 17. As with the previous two datasets it is clear that the mapping generates a more continuous latent space again resembling a filled circle. Similarly to figure 16 the aeroplanes are clustered at the bottom with Bonsai trees at the top but the class overlap has been reduced with only one aeroplane incorrectly

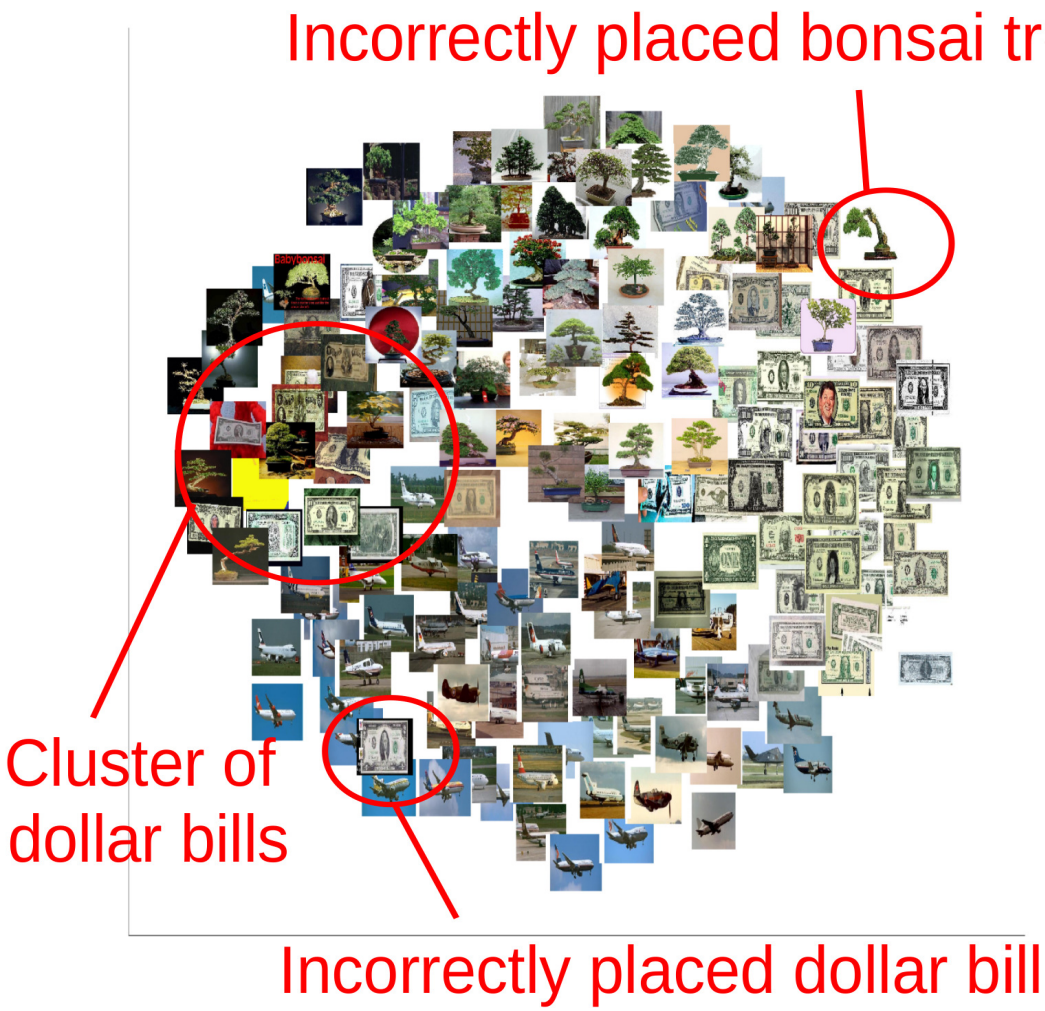


Figure 16: Visualisation of the Caltech101 dataset with standard Elastic MDS.

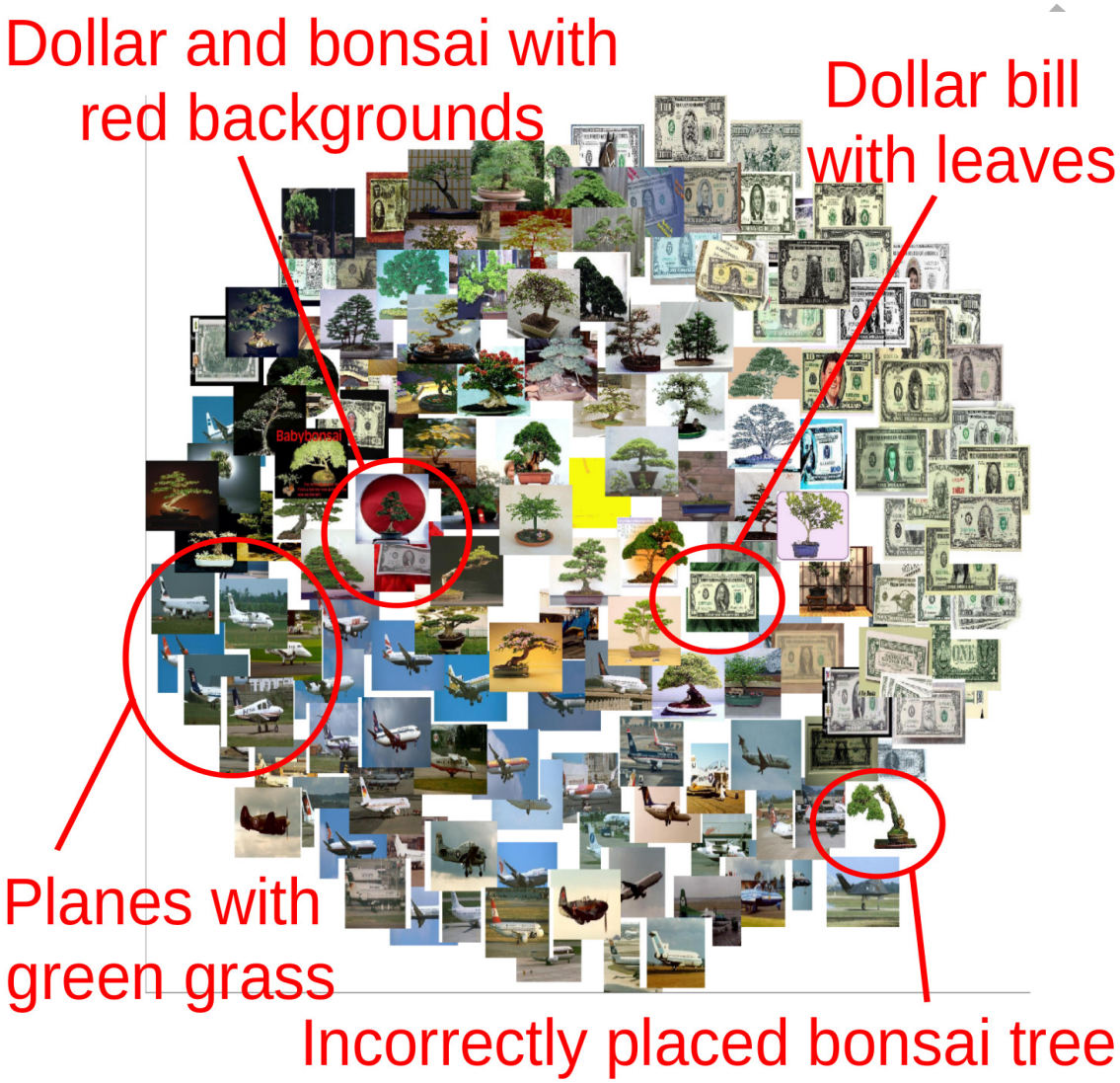


Figure 17: Visualisation of the Caltech101 dataset with learned inputs.

placed at the left side and a Bonsai tree surrounded by aeroplanes at the bottom right side. Through the mixture of different dissimilarities the mapping has pulled the dollar bill images to the right side with a small number of outliers. These outliers are more intuitive now, for instance the dollar bill and Bonsai tree with dark red backgrounds are now placed in close proximity. Further to this the aeroplane images with grass are placed close to the Bonsai tree images, a trait which is not present in figure 16.

When standard Isomap, with neighbourhood size $k = 25$, is used to generate a representation of the Caltech dataset (figure 18) the aeroplane and Bonsai tree classes are separated, but the dollar bill images are not. Some dollar bills are clustered closer to the Bonsai tree class than other similar images from the same class. On the other hand the multiple dissimilarity representation, with neighbourhood size $k = 68$, of figure 19 separates the main dollar bill cluster from the other two classes. In addition to this the images of planes with similar backgrounds, particularly the colour of the sky and clouds, are neighbours unlike the mapping of figure 18. The latent representations found through Isomap both include discontinuities and do not map certain features, such as the Bonsai tree and dollar bill with red backgrounds, in close proximity to one-another. As with the previous datasets we find that Isomap generates inferior visualisations to that of Elastic MDS, however the quality does improve when mixtures of dissimilarity measures are incorporated.

The relative weights allocated to each of the input dissimilarity measures is shown in figure 15. As with the previous two datasets the most dominant measure is the Ware and Hedges dissimilarity. For those mappings there seemed an intuitive reason as to why the datasets cluster naturally based on minimum and maximum values in greyscale images, however the inputs here are not greyscale. Instead they are high dimensional feature vector descriptors of the true images.

In order to gain understanding as to why the weights were allocated in this way we can analyse the maximal eigenvalues of each of the dissimilarity measures. For the case of data covariance matrices it is well known that the eigenvalues define the spread in each principal axis. When we consider a dissimilarity matrix this maps data from the observation space to the dissimilarity space [23]. The principal eigenvalue defines the spread of data in the principal axis of the dissimilarity space. It is therefore clear that greater spread in the input dissimilarity space (used by Elastic MDS and Isomap), and therefore less clustering and overlap, lends the input space better to recreation in a latent Euclidean space, reducing the mapping cost function with respect to the weights α_l. The maximal eigenvalues (normalised to maximum unity) shown in figure 20 correspond to the dominant weights allocated to each of the five datasets visualised in this section. These relative eigenvalues cannot be used to fix α_l prior to training in this framework, since the maximal eigenvalue for dissimilarity measures such as those in table 1 will always be positive which does not match the case of the Open Box dataset. Using normalised eigenvalues is however a more efficient method for initialising α_l than using a random scheme. Once the visualisation space was constructed from eigenvalue-initialised weights we analysed the relative deviations from initial to trained weights. The largest

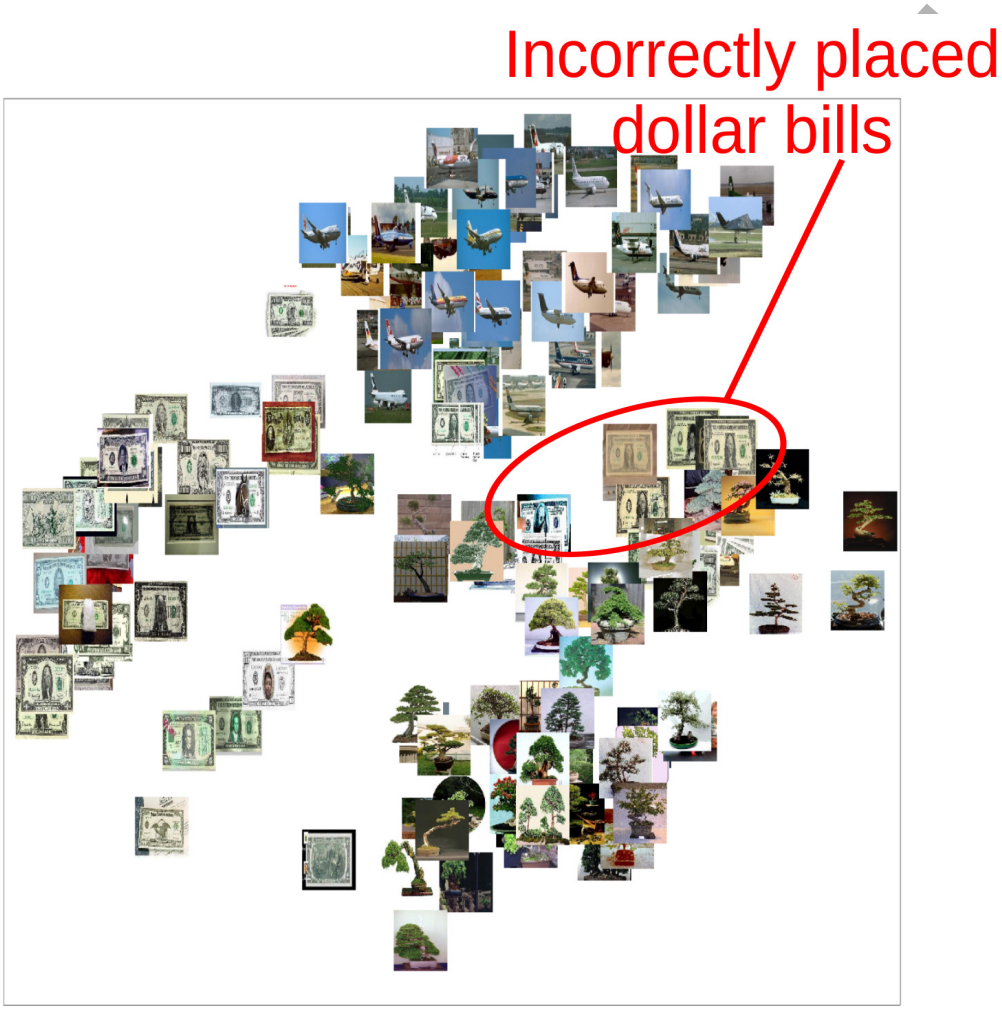


Figure 18: Isomap Visualisation of the Caltech101 dataset.

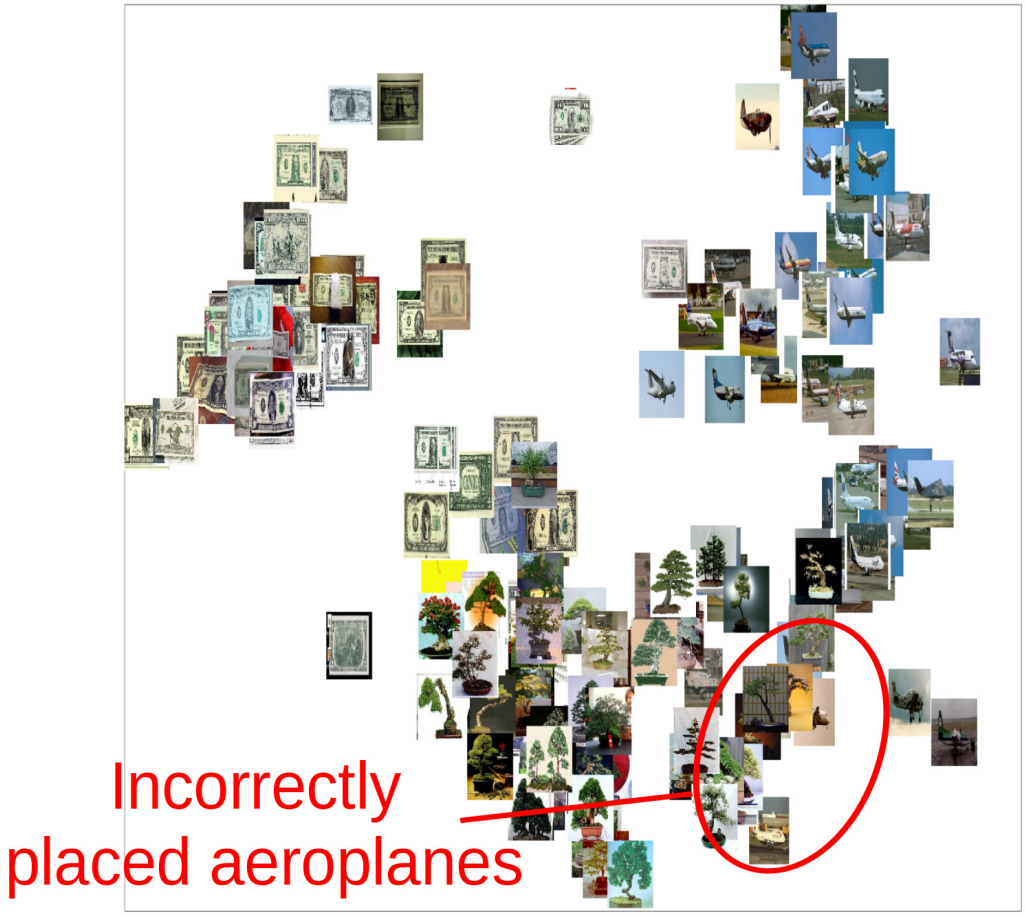


Figure 19: Isomap Visualisation of the Caltech101 dataset with learned inputs.

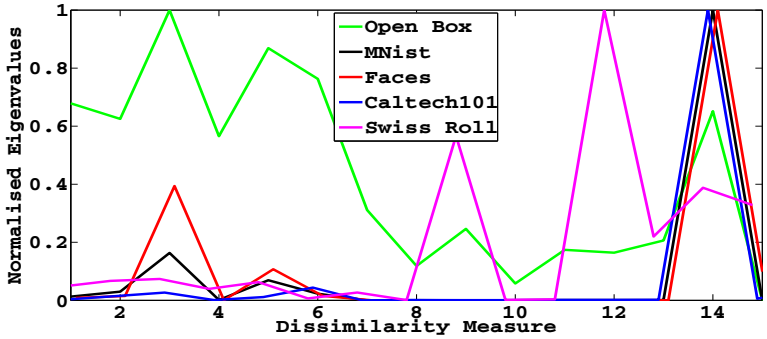


Figure 20: Normalised eigenvalues (to maximum unity) of dissimilarity matrices for each of the five mappings. The relative magnitude of the eigenvalues highly correlates with the weight allocated to each of the input dissimilarities, D^l.

deviations, $\max_l |\overline{\lambda}_l - \alpha_l|$, were 0.16 for the Open Box, 0.09 for the swiss roll, 0.01 for the MNist and faces datasets and 0.03 for the Caltech101 images subset. For best initialisation of the mixing weights we propose to use the relative normalised eigenvalues, i.e. $\alpha_l = \lambda_l / \sum_m \lambda_m$. Fixing the weights to the relative maximal eigenvalues in a re-run of the Open Box experiment produced a sub-optimal visualisation with higher stress. This shows that weight learning through the gradient descent procedure of section 2 is required to obtain better data representations, despite there often only being a minor deviation in the weight parameters. This finding allows for faster training of the mapping than when α_l is initialised randomly as would be typically done. The training time, post construction of the multiple dissimilarity matrices which make up D_x, is almost identical to that of standard Elastic MDS.

4. Conclusion

This paper has presented a novel way of generating projective visualisations through varying the input dissimilarities used to chart the observed manifold. The linear combination of different distance measures allows for multiple dissimilarities to be considered, a particular benefit when the geometry of the observation manifold is unknown. The combination of dissimilarity measures has links to the current research in multiple kernel learning, but in particular our experiments have shown the relative weights which minimise the cost function of Elastic MDS are linked to the normalised maximal eigenvalues of the individual dissimilarity matrices. This particular finding will allow for the dissimilarity weights to be reliably initialised at the start of the mapping to avoid potential local minima issues.

The changes resulting from a modified input dissimilarity were illustrated with five datasets. For the real-world high-dimensional datasets the mapping proposed in this paper generated a much more intuitive visualisation, preserving

the class structure which is known to humans but was not used in the mapping process. Further research will assess the impact that learning mixtures of input dissimilarities has on other mappings such as Curvilinear Component Analysis, Stochastic Neighbour Embedding and Bregman MDS. The work in this paper has focused on the case of vectorial observations, however replacing the measures of table 1 with other dissimilarities allows for trivial extensions to binary, time series, graph or probability distribution observation spaces.

References

[1] M. Ali, Z. Chahooki, and N.M. Charkari. Shape classification by manifold learning in multiple observation spaces. *Information Sciences*, 262:46 – 61, 2014.

[2] S. Amari. *Differential-geometrical methods in statistics*. Lecture Notes in Statistics. Springer-Verlag, 1985.

[3] Herbert Bay, Tinne Tuytelaars, and Luc Gool. Surf: Speeded up robust features. In Aleš Leonardis, Horst Bischof, and Axel Pinz, editors, *Computer Vision – ECCV 2006: 9th European Conference on Computer Vision, Graz, Austria, May 7-13, 2006. Proceedings, Part I*, pages 404–417. Springer Berlin Heidelberg, 2006.

[4] M. Belkin and P. Niyogi. Laplacian eigenmaps for dimensionality reduction and data representation. *Neural Computation*, 15:1373–1396, 2002.

[5] C. M. Bishop, M. Svensen, and C. K. I. Williams. GTM: The generative topographic mapping. *Neural Computation*, 10(1):215–234, 1998.

[6] C. Campbell and Y. Ying. *Learning with Support Vector Machines*, volume 5 of *Synthesis Lectures on Artificial Intelligence and Machine Learning*. Morgan & Claypool Publishers, February 2011.

[7] P. Demartines and J. Herault. Curvilinear component analysis: a self-organizing neural network for nonlinear mapping of data sets. *Neural Networks, IEEE Transactions on*, 8(1):148–154, Jan 1997.

[8] E.W. Dijkstra. A note on two problems in connection with graphs. *Numerical Mathematics*, 1:269–271, 1959.

[9] L. Fei-Fei, R. Fergus, and P. Perona. Learning generative visual models from few training examples: an incremental bayesian approach tested on 101 object categories. In *IEEE. CVPR 2004, Workshop on Generative-Model Based Vision.*, 2004.

[10] M. Gönen and E. Alpaydin. Multiple kernel learning algorithms. *J. Mach. Learn. Res.*, 12:2211–2268, July 2011. ISSN 1532-4435.

[11] M. Hein and J.Y. Audibert. Intrinsic dimensionality estimation of submanifolds in rd. In *Proceedings of the 22Nd International Conference on Machine Learning*, ICML '05, pages 289–296, 2005.

[12] G. Hinton and S. Roweis. Stochastic neighbor embedding. *Advances in neural information processing systems*, 15:833–840, 2003.

[13] N.D. Lawrence. Gaussian process latent variable models for visualisation of high dimensional data. In *Advances in Neural Information Processing Systems 16*, pages 329–336, December 2003.

[14] Y. LeCun, C. Cortes, and C.J.C. Burges. The MNIST database. http://yann.lecun.com/exdb/mnist/.

[15] J. A. Lee and M. Verleysen. *Nonlinear Dimensionality Reduction*. Springer Publishing Company, Incorporated, 1st edition, 2007. ISBN 0387393501, 9780387393506.

[16] J.A. Lee, A. Lendasse, N. Donckers, and M. Verleysen. A robust nonlinear projection method. In M. Verleysen, editor, *Proceedings of ESANN 2000, 8th European Symposium on Artificial Neural Networks*, pages 13–20. D-Facto public., Bruges, Belgium, April 2000.

[17] J.A. Lee, A. Lendasse, and M. Verleysen. Curvilinear distance analysis versus Isomap. In Michel Verleysen, editor, *Proceedings of ESANN 2002, 10th European Symposium on Artificial Neural Networks*, pages 185–192, 2002. ISBN 2-930307-02-1.

[18] T. Lin, H. Zha, and S. Lee. Riemannian manifold learning for nonlinear dimensionality reduction. In *Proceedings of the 9th European Conference on Computer Vision - Volume Part I*, ECCV'06, pages 44–55, Berlin, Heidelberg, 2006. Springer-Verlag. ISBN 3-540-33832-2, 978-3-540-33832-1.

[19] Y. Y. Lin, T. L. Liu, and C. S. Fuh. Multiple kernel learning for dimensionality reduction. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 33(6):1147–1160, 2011.

[20] D. Lowe. Novel 'topographic' nonlinear feature extraction using radial basis functions for concentration coding in the 'artificial nose'. In *Artificial Neural Networks, 1993., Third International Conference on*, pages 95–99, May 1993.

[21] D. Lowe and M.E. Tipping. Neuroscale: Novel topographic feature extraction using RBF networks. In M.C. Mozer, M.I. Jordan, and T. Petsche, editors, *Advances in Neural Information Processing Systems 9*, pages 543–549. MIT Press, 1997.

[22] V.E. McGee. The multidimensional scaling of "elastic" distances. *British Journal of Mathematical and Statistical Psychology*, 19:181 – 196, 1966.

[23] E. Pekalska and R.P.W. Duin. *The Dissimilarity Representation for Pattern Recognition: Foundations And Applications (Machine Perception and Artificial Intelligence)*. World Scientific Publishing Co., Inc., River Edge, NJ, USA, 2005. ISBN 9812565302.

[24] S.T. Roweis and L.K. Saul. Nonlinear dimensionality reduction by locally linear embedding. *Science*, 290:2323–2326, 2000.

[25] J. W. Sammon. A nonlinear mapping for data structure analysis. *IEEE Trans. Comput.*, 18(5):401–409, May 1969. ISSN 0018-9340.

[26] B. Schölkopf, A. Smola, and K.R. Müller. Nonlinear Component Analysis as a Kernel Eigenvalue Problem. *Neural Computation*, 10(5):1299–1319, 1998.

[27] J. Sun. *Extending Metric Multidimensional Scaling with Bregman Divergences*. PhD thesis, University of the West of Scotland, 2011.

[28] J. Sun. Open box dataset, 2012. URL `http://cis.uws.ac.uk/research/JigangSun/index.html`.

[29] J. Sun, M. Crowe, and C. Fyfe. Extending metric multidimensional scaling with Bregman divergences. *Pattern Recognition*, 44(5):1137 – 1154, 2011.

[30] J. Sun, M. Crowe, and C. Fyfe. Incorporating visualisation quality measures to curvilinear component analysis. *Information Sciences*, 223:75 – 101, 2013.

[31] J. B. Tenenbaum, V. Silva, and J. C. Langford. A Global Geometric Framework for Nonlinear Dimensionality Reduction. *Science*, 290(5500): 2319–2323, 2000.

[32] M.E. Tipping and D. Lowe. Shadow targets: A novel algorithm for topographic projections by radial basis functions. *NeuroComputing*, 19:211–222, 1997.

[33] X. Wang and C. Fyfe. Applyling Bregman divergences to the Neuroscale algorithm. In *Proceedings of 11th UK Workshop on Computational Intelligence*, pages 178–183, 2011.

[34] Z. Wang, Q. Fan, S. Ke, and D. Gao. Structural multiple empirical kernel learning. *Information Sciences*, 301:124 – 140, 2015.