Improved data visualisation through nonlinear dissimilarity modelling

Iain Rice

Aston University, UK.

Abstract

Inherent to state-of-the-art dimension reduction algorithms is the assumption that global distances between observations are Euclidean, despite the potential for altogether non-Euclidean data manifolds. We demonstrate that a non-Euclidean manifold chart can be approximated by implementing a universal approximator over a dictionary of dissimilarity measures, building on recent developments in the field. This approach is transferable across domains such that observations can be vectors, distributions, graphs and time series for instance. Our novel dissimilarity learning method is illustrated with four standard visualisation datasets showing the benefits over the linear dissimilarity learning approach.

Keywords: `Dissimilarity`, `Visualisation`, `Multidimensional Scaling`, `RBF Network`

1. Introduction

Dimension reduction algorithms used to generate visualisations of high dimensional data require a chart of observations which must follow a global or local structure. The Sammon map [43], Stochastic Neighbour Embedding (SNE) [19] and variants, the Gaussian Process Latent Variable Model (GPLVM) [21], Generative Topographic Map (GTM) [7], Metric Multidimensional Scaling (MDS) and Curvilinear Component Analysis (CCA) [11] assume global Euclidean structure. Bregman divergences generate mappings with non-metric multidimensional scaling in [45, 47, 46] however the use of the Euclidean distance, as in standard MDS, remains. In the case where the observed data is known to sit upon a non-Euclidean manifold it is typically assumed that local regions of the manifold are Euclidean. Algorithms such as Locally Linear Embedding [42], Laplacian Eigenmaps [6], Riemannian Manifold Learning [26] and methods using geodesic distances based upon local Euclidean structure such as Isomap [49], the Geodesic Nonlinear Map [25] and Curvilinear Distance Analysis [24] rely on

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

this property holding. Furthermore these algorithms require smooth continuity between local charts. This is known to not be the case where a manifold is, for instance, fractal or when observations are sparse and not true neighbours. As such the choice of local neighbourhood size parameters presents a challenge, causing the potential for short-circuits in neighbourhood graphs. The work of FINE [10] assumes that observations sit upon a statistical Riemannian manifold which is less restrictive than the Euclidean counterpart [3]. As such FINE uses an approximation to the Fisher Information metric to calculate local distances between observations, however each of the proposed approximations are not without limitations. In contrast the framework of [40] embeds non-Euclidean data onto a latent sphere of with calculated radius. This is in contrast to almost all other dimension reduction algorithms which do not restrict the structure of the latent space.

The latent variable models GTM and GPLVM assume that observations sit upon hyper-ellipses. In the GTM case this structure is treated as isotropic and as such suffers from the issues of hyper-spherical geometry (see [23] for details). The hyper-ellipse of GPLVM only permits dimensions between observations to be independent, a trait known to be false in many time series and image analysis domains for instance. These approaches are therefore incapable of constructing a reliable chart for complex datasets. In [23] it was demonstrated that dimension reduction algorithms relying on nonconvex optimisation of latent points, for instance MDS, CCA and GTM, perform superior to mappings using convex optimisation including PCA, LLE and Isomap.

An alternative approach is considered in [30, 31] for the task of pattern discovery in large datasets. Local affinity patterns are identified across patterns and observed dimensions to convey significant attributes. The test for significance involves a Euclidean thresholding scheme over the cleaned graph weight matrix. The highlighted local affinities should be anomalies or sources of information, allowing the user to focus on a small subset of a large collection of data. In contrast the result of information visualisation is to utilise all attributes of observations and present the visual map over all datapoints to a human for interpretation. Such a weighting matrix as used in [30] can however be integrated within several visualisation frameworks when the weighting function is specified.

The notion and impact of non-Euclidean pattern analysis is discussed at length in [37]. Despite the fact that there are many causes of non-Euclidean observations, frameworks to handle such datasets are still emerging and have not been widely adopted [12]. When the nature of an observed manifold is of unknown topology one naturally is unaware of the dissimilarity measure which charts the manifold. It is however possible to learn such a chart using a combination of a set of multiple dissimilarity measures, a dictionary. This is the approach of multiple kernel learning where kernels are combined linearly or nonlinearly in order to improve regression or classification performance (see [9, 15] for an overview). Multiple kernel learning has also been implemented in the field of manifold learning [2]. Multi-feature kernels were developed in [55] to learn features for facial recognition based on a dictionary approach and discriminant analysis rather than dimension reduction. This notion is developed further in

[1] where a sparse hierarchical dictionary based on Gaussian kernels is used for classification.

The tasks of regression and classification are by nature supervised. In this paper we consider the case of dimension reduction, in particular visualisation, which is unsupervised and as such mapping targets do not exist. The targets in this case are learned by minimisation of a mapping cost function such that neighbourhoods and the topological ordering of data is preserved. Non-Euclidean charts form the input to the visualisation framework of [27] relying on linear discriminant analysis. This linear approach does not generalise to non-linear mappings. Another linear projection is used in [56] to map data whose dissimilarities are specified by a probabilistic measure. This approach cannot suitably map nonlinear structures, but the proposed measure can be incorporated into the framework of this paper. A distance metric learning approach is proposed in [51] focussing on clustering by adapting a kernel to learn a dissimilarity measure rather than fixed combinations of kernels. The clustering of data through spectral construction of kernels is detailed in [4] with links to Laplacian Eigenmaps, however this by nature learns a local descriptor of data. Using adaptive metrics [17] present an analogue of PCA with the goal of intrinsic dimensionality estimation rather than performing dimension reduction. The variables of interest in data are learned in [39] in a linear fashion prior to dimension reduction, however the relative significance of these features compared to one-another is not retained. Linear combinations of kernels form the basis of the nonlinear dimension reduction performed in [53] however the kernels used are restricted to polynomial functions and no learning of dissimilarities upon a manifold is performed. The Canonical Correlation Analysis approach of [57] linearly combines separate local and global kernels to perform dimension reduction which for certain kernel choices will behave like Isomap. A far more expressive linear combination of multiple kernels is presented in [33] where the weighting is fixed prior to dimension reduction. The work of [14] creates an ensemble of different clustering partitions, which may be non-Euclidean by nature, allows for more accurate clustering and classification.

In this paper we present a method for learning a chart based on a dictionary of dissimilarity measures whilst simultaneously constructing a nonlinear mapping. In [41] a linear combination of dissimilarity measures was used in this way and it was shown that the quality and interpretability of visualisations improved when the chart is learned. This paper builds on this linear model by learning a nonlinear combination of dissimilarity measures using a universal approximator. In order to show the improvements of this nonlinear learning of dissimilarities we use Elastic MDS as in [41] to provide a benchmark for our experimental results, however our approach generalises to other visualisation algorithms. In order to demonstrate the impact of our approach we generate visualisations of four standard datasets with Elastic MDS and Isomap. We assess the quality of our results with a visual comparison of the mapped latent variables, however as discussed in [52] quantitative comparison with visual quality metrics are not appropriate for non-Euclidean mappings.

2. The Learning Task

The aim of this paper is to accurately estimate the chart of an observed manifold without assuming a particular metric, but by learning a mixture from a fixed dictionary of dissimilarities. As a precursor we build on the work of [41] and therefore focus on the case of Elastic MDS [32] to perform a comparative analysis of our approach. We assess the performance of the constructed chart through visual analysis of an embedding of a dataset. This embedding need not be Euclidean in terms of Witney's emedding theorem [54] as visualisation would only be possible here if the intrinsic dimensionality of a dataset were 3-dimensional or less. Elastic MDS generates an embedding of a dataset, X, with N observations by constructing a set of latent points, $Y \in \mathbb{R}^V$. As is typical for the task of visualisation we fix $V = 2$ in this paper, however our methods generalise trivially to other integer values for V.

A particular benefit of MDS methods is that X need not be vectorial or even explicitly known, it is only required that the matrix of pairwise dissimilarities $D_x(i,j)$ between observations X_i and X_j is given. The latent points $\boldsymbol{y}_i$ corresponding to observation X_i are learned through gradient descent of the Elastic MDS 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 the latent, visualised points $\boldsymbol{y}_i$ and $\boldsymbol{y}_j$. This measure is typically taken to be the Euclidean distance. Elastic MDS is distinct from the popular Sammon map due to the quadratic term in the denominator of equation (1), making the cost function more sensitive to local distances by stretching $D_x(i,j)$, hence the term elastic. This local focus naturally comes at the expense of global preservation.

For the case that X consists of vectorial observations, $\boldsymbol{x}_i$, it is typically assumed that $D_x(i,j)$ is the Euclidean distance in the literature. This measure is only appropriate in the cases where the observed manifold is Euclidean. In the Riemannian or non-Riemannian manifold cases this distance function will give an incorrect approximation of distance. On statistical Riemannian manifolds the natural distance measure is known to be the Fisher Information Metric, which is typically approximated [10] using other divergence measures. The aim of this paper is to approximate the distance between observations:

$$D_x(i,j) = f\left(X_i, X_j\right). \tag{2}$$

In [41] the function f is approximated using a linear combination of dissimilarity measures as a dictionary:

$$D_x(i,j) = \sum_{l=1}^{L} \alpha_l D^l(i,j), \tag{3}$$

where α_l is the weight corresponding to the l-th dissimilarity measure, constrained such that α_l sums to unity. The dictionary of L dissimilarity measures

is user specified and the weights were learned during the optimisation of the Elastic MDS cost function in equation (1). These weights were optimised using gradient descent over equation (3) with respect to each factor α_l in order to find the optimal representation achieving a global minima. The dictionary-based approach is suited to situations where the natural metric of the observed data is unknown. In the regression or classification setting it would typically be assumed that the measure generating a chart over observations X_i is that which achieves the highest predictive performance, however there is no guarantee that the measure which charts the manifold will be identified. For the unsupervised dimension reduction case we cannot identify a single measure in the same way. The use of a broad dictionary, containing for instance, weighted, unweighted, correlation and divergence measures for vectorial datasets allows us to circumvent this issue. To ensure the chart approximation was flexible 15 dissimilarity measures were used in [41] as a dictionary which are listed in table 1, taken from [36]. For the developments in this paper we utilise the same dictionary, noting that the methods discussed need not be restricted to vectorial observations $\boldsymbol{x}_i$. In the non-vectorial observation case such that X_i may be for instance a probability distribution, binary data, an image, a graph or a time series for instance, the technique generalises provided a similar dictionary to that of table 1 is provided (see [36] for examples).

For the experiments described in this paper the parameters of the weighted Euclidean (measure 2) were fixed to be the inverse of the sample mean vector. The Minkowski distance (measure 5) parameter, p, 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) was the sample covariance matrix calculated from the observations. These dissimilarity measures are therefore invariant to the absolute scale of the data in each dimesion. This ensures 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.

It is well known that linear functions are not capable of universal approximation and that they are subject to adversarial examples [16] from an improper interpolation of the input space. We propose an extension to the approximation in equation (3) with a nonlinear interpolator. For this task we propose to use an RBF network [8] which is known to be a universal approximator of functions [38, 35, 44]. In particular we use an RBF network of the form:

$$(\boldsymbol{z}_i)^k = \sum_j \boldsymbol{W}_{jk} \phi \left((\boldsymbol{u}_i - \boldsymbol{v}_j)^T \boldsymbol{H}^{-1} (\boldsymbol{u}_i - \boldsymbol{v}_j) \right), \tag{4}$$

where $(\boldsymbol{z}_i)^k$ is the k-th dimension of output vector $\boldsymbol{z}_i$, $\boldsymbol{u}_i$ is the input vector, $\boldsymbol{W}_{jk}$ are the jk-th weights of the matrix $\boldsymbol{W}$, ϕ is a nonlinear basis function, $\boldsymbol{v}_j$ is the j-th network prototype and $\boldsymbol{H}$ is a diagonal matrix of weights to perform automatic relevence detection (originally derived in the artificial neural networking literature [29, 34]) learning the significant dimensions of $\boldsymbol{u}_i$. The mapping of equation (4) can be written in matrix form as $\boldsymbol{Z} = \Phi \boldsymbol{W}$ with the

	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).

elements of Φ given by:

$$\Phi_{ij} = \phi\left((\boldsymbol{u}_i - \boldsymbol{v}_j)^T \boldsymbol{H}^{-1}(\boldsymbol{u}_i - \boldsymbol{v}_j)\right). \tag{5}$$

In order for the RBF network to approximate the function of equation (2) we must explicitly specify the inputs and outputs. Denoting the i-th vector of dissimilarities in $D_x(i,j)$ by $\boldsymbol{d}_i$ we use the convention that the outputs of the RBF network should be $(\boldsymbol{d}_i)^j = (\boldsymbol{z}_i)^k$. The network inputs $\boldsymbol{u}_i$ should naturally relate to the individual L dissimilarity measures in the dictionary to be interpolated, each vector of which we denote $\boldsymbol{d}_i^l$ containing N elements $D^l(i,j)$ for fixed i. In the same way that dynamical systems form a delay embedding to approximate the intrinsic dimensionality of inputs through Taken's theorem [48, 20] we form an embedding of the dissimilarity spaces forming the inputs through the concatenation:

$$\boldsymbol{u}_i = [\boldsymbol{d}_i^1, \boldsymbol{d}_i^2, \ldots, \boldsymbol{d}_i^L],$$

and as such the dimensions of $\boldsymbol{u}_i$ will be $N \times L$. The network prototypes $\boldsymbol{v}_i$ are a subset of these observations which can be learned in the optimisation process which follows. Given the potential for the dimensionality of $\boldsymbol{u}_i$ to be very large it is naturally desirable to reduce this to avoid memory issues in the computational implementation of the process. In section 3 we demonstrate that the impact of choosing a subset N' of the N elements of each vector $\boldsymbol{d}_i^l$ does not reduce the visualisation quality and in fact assists in model regularisation. An additional tool for imposing sparsity in this model is through the manipulation of $\boldsymbol{H}$ such that insignificant dimensions of $\boldsymbol{u}_i$ can be 'switched off' as the elements of $\boldsymbol{H}$ approach zero. In [41] the responsibility of each of the dissimilarity measures used in the linear mixture is assessed through plotting the values of the linear weights, α_l, to identify which input dissimilarity measure is the most significant. Here however the matrix $\boldsymbol{H}$ allows us to analyse this by plotting the responsibilities, r_l:

$$\tilde{r}_l = \sum_{i \in [l-1 \times N, l \times N]} \boldsymbol{H}_{i,i},$$

$$r_l = \frac{\tilde{r}_l}{|\sum_l \tilde{r}_l|},$$

i.e. only summing the values in $\boldsymbol{H}$ relating to the l-th dissimilarity measure and calculating the responsibilities which must sum to ± 1. Aside from the potential for inducing sparsity it was found that the learning of $\boldsymbol{H}$ in the experiments of section 3 did not improve mappings compared to when $\boldsymbol{H}$ was fixed to the identity matrix. The decision to learn $\boldsymbol{H}$ will therefore depend on the necessity for a sparse solution. In the case where $\boldsymbol{H}$ is fixed such that $\boldsymbol{H} = \boldsymbol{I}$, where $\boldsymbol{I}$ is the identity matrix, then Φ can be fixed such that training time can be reduced.

In order to optimise the parameters of our model, $\{\boldsymbol{y}_i, D_x(i,j), \boldsymbol{W}, \boldsymbol{H}\}$, we propose an alternating iterative learning scheme. The parameters belonging to the RBF network, $\{\boldsymbol{W}, \boldsymbol{H}\}$, are typically learned through pseudoinverse of $\boldsymbol{W}$ with respect to the network output targets. The targets here do not exist and as such we choose to optimise these parameters in the same fashion as in NeuroScale with Shadow Targets [50].

As in standard Elastic MDS the visualised points are learned through gradient descent of the cost function in equation (1), which could be performed via scaled conjugate gradients (SCG). The gradients required are given by:

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

The RBF network parameters are global in nature and should therefore be optimised on a slower timescale than the latent points. From the Shadow Targets approach we require the gradients of the cost function of equation (1), with respect to the network outputs:

$$\frac{\partial E}{\partial D_x(i,j)} = \left(\frac{-2(D_y(i,j))^2}{(D_x(i,j))^3} + \frac{2 D_y(i,j)}{(D_x(i,j))^2} \right). \tag{6}$$

We then perform a gradient step over the $D_x(i,j)$ with a learning rate η:

$$\tilde{D}_x(i,j) = D_x(i,j) - \eta \frac{\partial E}{\partial D_x(i,j)}, \tag{7}$$

in order to update the RBF network outputs. The updating rule for $\boldsymbol{H}$ is given by:

$$\tilde{\boldsymbol{H}}_{k,k} = \boldsymbol{H}_{k,k} + \eta \left(\sum_{ij} (\tilde{D}_x - D_x) \boldsymbol{W}^T \Phi'(\boldsymbol{u}_i - \boldsymbol{v}_j) \boldsymbol{H}^{-1} (\boldsymbol{u}_i - \boldsymbol{v}_j)^T \right), \tag{8}$$

as is typical in the ARD setup. Here we use Φ' to denote the elementwise derivative of the basis function matrix Φ with respect to its' argument.Once $\tilde{\boldsymbol{H}}$ is calculated the basis function matrix Φ can be updated from equation 5 with the new $\boldsymbol{H}$ matrix to give $\tilde{\Phi}$. Following this the weight matrix $\boldsymbol{W}$ can be learned in the standard pseudoinverse method:

$$\tilde{\boldsymbol{W}} = \tilde{\Phi}^{\dagger} \tilde{D}_x.$$

As with [41] by attempting to minimise the Elastic MDS cost function of equation 1 we are attempting to construct a dissimilarity matrix $D_x(i,j)$ which can be accurately reproduced in $D_y(i,j)$. This is performed by both modifying $D_x(i,j)$ and the latent points $\boldsymbol{y}_i$ which are used to calculate $D_y(i,j)$.

In this context there are many different ways of optimising the RBF network parameters in either a convex or non-convex fashion to improve the regression performance. Since in this paper we are concerned with the creation of a fixed visualisation we choose to learn $\{\boldsymbol{W}, \boldsymbol{H}\}$ though the gradient descent procedure detailed above by matching the shadow targets of the MDS cost function as by design this will improve the quality of the generated visualisations. The pseudocode for the algorithm is presented in algorithm 1. In order to generate the visualisations of section 3 we fixed Nepochs to 300 and Niter to 20, however the models all converged to a minimum before reaching these levels. It is proposed that $\boldsymbol{H}$ be initialised to the identity matrix so as not to bias the dissimilarity weighting at the beginnning of training. We have found in the experimental results that initialising $\boldsymbol{W}$ in such a way that training time is minimised is best done by fixing $D_x(i,j)$ suitably using the linear input dissimilarity model of equation (3). Here α_l is set to the maximal eigenvalue of each dissimilarity measure as outlined in [41] to allow for $D_x(i,j)$ to be quickly initialised and $\boldsymbol{W}$ to be learned by pseudoinverse as above.

3. Experimental Results

It was established in [41] that learning a linear mixture of input dissimilarities improved the visualisation quality when compared to standard Elastic MDS relying on Euclidean distances as inputs. Using this as a benchmark we test the impovements of the nonlinear mixture of learned input dissimilarities using

Algorithm 1 Pseudocode for multiple dissimilarity learning in Elastic MDS.

Require: Dissimilarities $D_x^l(i,j)$,
Initialise $D_x = \sum_l \alpha_l D_x^l$ with α given by the maximal eigenvalue of each dissimilarity matrix.
Initialise $\boldsymbol{H} = \boldsymbol{I}$, calculate Φ from equation (5),
Initialise $\boldsymbol{W} = \Phi^\dagger D_x$,$D_x = \Phi \boldsymbol{W}$,
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 $\overline{\boldsymbol{y}}_i$,
        Re-calculate latent dissimilarities $D_y(i,j) = \|\overline{\boldsymbol{y}}_i - \overline{\boldsymbol{y}}_j\|_2$,
        Re-calculate mapping error E,
        if error reduced **then**
            Update $\boldsymbol{y}_i \leftarrow \overline{\boldsymbol{y}}_i$,
        end if
    end for
    Perform a gradient step using equation (6) to give $\tilde{D}_x = D_x - \eta \frac{\partial E}{\partial D_x}$,
    Perform a gradient step using equation (8) to give $\tilde{\boldsymbol{H}}$,
    Calculate $\tilde{\Phi}$ from equation (5),
    Update weights $\tilde{\boldsymbol{W}} = \tilde{\Phi}^\dagger D_x$,
    Re-calculate input dissimilarities $D_x = \tilde{\Phi}\tilde{\boldsymbol{W}}$,
    Re-calculate mapping error E,
    if error reduced **then**
        Update $\boldsymbol{H} \leftarrow \tilde{\boldsymbol{H}}$,$\boldsymbol{W} \leftarrow \tilde{\boldsymbol{W}}$, $\Phi \leftarrow \tilde{\Phi}$.
    end if
end for

the framework of algorithm 1 on four standard datasets. Experiments have been repeated five times with randomly perturbed initialisations. In each case the mappings converged to the same latent space indicating a global minimum has been achieved. We also present visualisations generated using Isomap with a geodesic dissimilarity measure over both a linear and nonlinear mixture of input dissimilarities. The neighbourhood size, k, is set such that a connected neighbourhood graph is achieved. For each experiment the following parameter settings of algorithm 1 were used:

- Nepochs = 150 iterations,
- Niter = 20 iterations optimised with scaled conjugate gradients and a threshold parameter $1e^{-4}$ resulting in early stopping of the gradient calculation,
- Learning parameter $\eta = 0.8$ which is increased by a factor of 1.2 upon

successfull gradient steps and decreased by a factor of 0.64 otherwise.

For each experiment we detail the training time for each section of algorithm 1 as these will depend on model parameterisation. The first category is naturally the construction of $D_x^l(i,j)$ as this will depend on the specified dictionary. Secondly we detail the initialisation of D_x and Φ,$\boldsymbol{W}$ stages corresponding to algorithm 1 steps 1 and 2-3 resepctively. Finally we report the computation time for the training of the latent space mapping. All experiments were performed using unoptimised code in Matlab on a core i5 1.7Ghz quadcore computer. These results are compared to the time required to generate a standard Elastic MDS mapping, with justification for the time differences, in section 3.5.

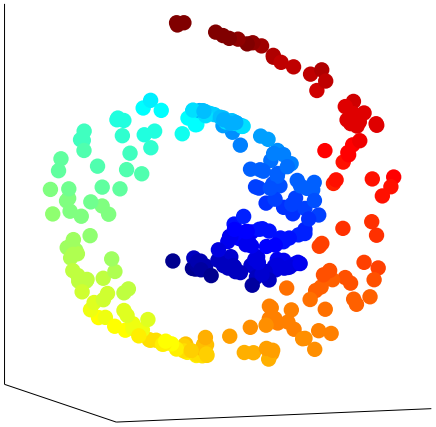


Figure 1: The swiss roll dataset in 3-dimensional observation space. The structure contains dark blue points at the centre (the left of the 2-dimensional latent rectangle) and dark red points at the exterior of the curve (the right of the 2-dimensional latent rectangle).

3.1. Swiss Roll

Firstly we analyse the artificial swiss roll dataset used extensively in [23] to test nonlinear dimension reduction algorithms. The 3-dimensional observed structure consists of 300 points randomly sampled from a 2-dimensional rectangular grid mapped into the roll.

Figures 2 and 3 show three visualisations of the swiss roll generated by Elastic MDS. For reference we use Elastic MDS based on a standard Euclidean distance input as a benchmark in figure 2. When the learned input dissimilarities are linear, shown in figure 3a), there is a greater level of overlap between the blue and orange points than when the learned input dissimilarities are nonlinear in figure 3b). The spread of points within each local region of the roll is also lower in the nonlinear extension. As such the unnormalised mapping stress for the linear mixture in figure 3a) is 143,320 compared to 11,649 for the nonlinear extension in figure 3b). This is due to the removal of restrictions in the linear case allowing for a more flexible dissimilarity matrix, $D_x(i, j)$, to be constructed in the nonlinear model. Figure 4 shows the responsibilities for each of the dissimilarity measures in table 1. In this artificial dataset the model utilises all measures in the mapping.

The Isomap visualisations are shown in figure 5 with linear mixtures of input dissimilarities in figure 5a) and the nonlinear extension in figure 5b). The neighbourhood parameters were $k = 10$ for both experiments, achieving a connected graph. The nonlinear learned structure in figure 5b) appears to form

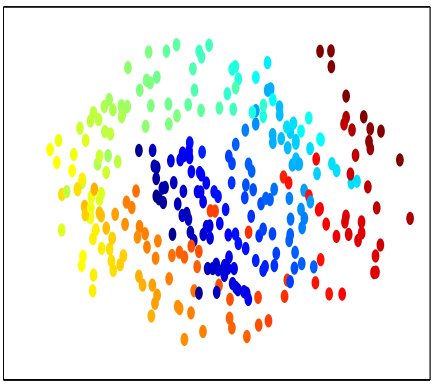
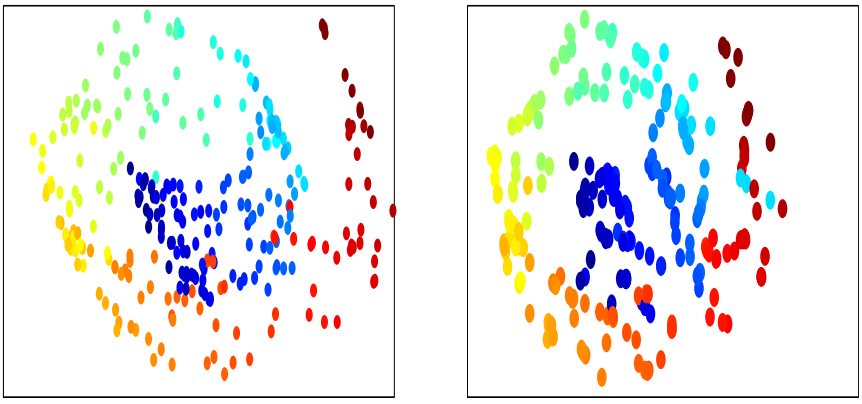


Figure 2: Visualisation of the swiss roll dataset in 2-dimensional latent space assuming Euclidean distance inputs. There is a significant overlap of dark blue points with orange points and minimal separation between the red and light-blue points.



(a) (b)

Figure 3: Visualisations of the swiss roll dataset with (a) linear input dissimilarities Elastic MDS and (b) nonlinear input dissimilarities in Elastic MDS.

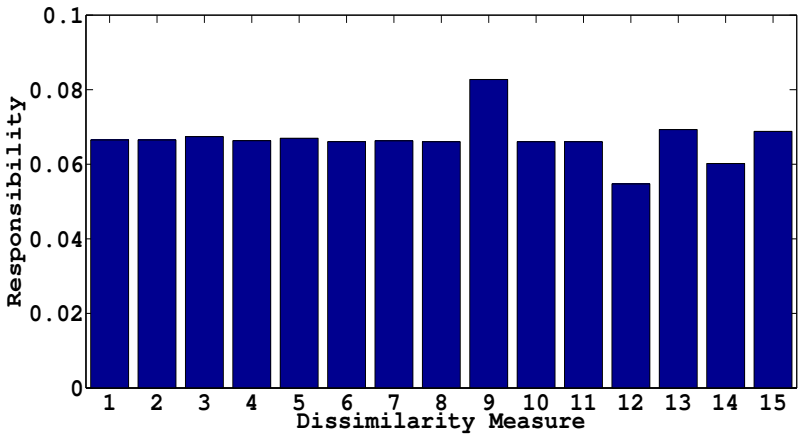
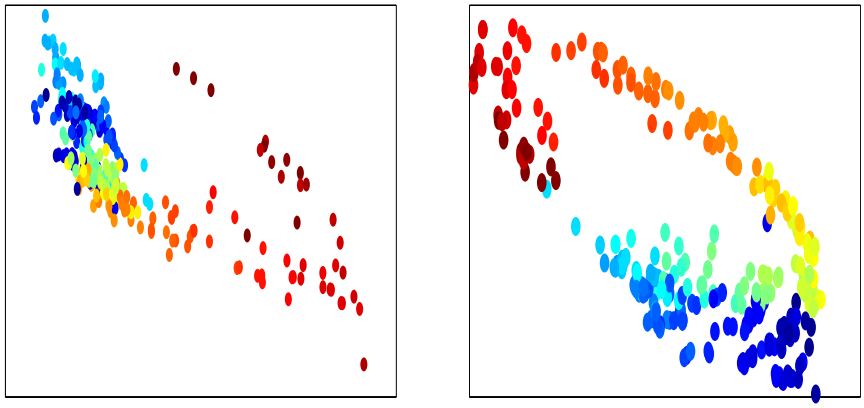


Figure 4: Responsibilities of the dissimilarity measures for the swiss roll mapping.



(a) (b)

Figure 5: Swiss roll dataset visualisations with Isomap based on (a) linear and (b) nonlinear input dissimilarities.

a latent loop, however there is a discontinuity at the join, whilst imposing a lower level of overlap between neighbourhoods than the linear input dissimilarity model in figure 5a).

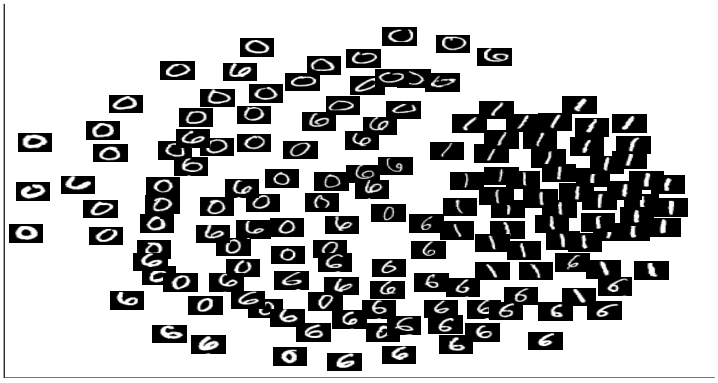


Figure 6: MNist mapping with Elastic MDS based on Euclidean distances as inputs. The latent variables of digit thickness and curvature are not mapped across the latent space, nor are incorrectly drawn digits clustered appropriately.

3.2. MNist Dataset

The second experiment presented in this paper is a subset of the MNist dataset [22]. We focus on 150 patterns with 50 '0's, '1's and '6's such that $N = 150$. The handwritten digit images are 28×28 pixels and are therefore treated as 784-dimensional observation vectors. A natural latent similarity is expected between the '0' and '6' classes which both posess a circular structure. Similarly we expect the '1's and '0's to be separated since the straight line in the '1's should not be seen in the joined '0's. For reference we show the visualisation based on Euclidean distances as inputs in figure 6. Figures 8a) and 8b) contain the visualisations of the MNist dataset with Elastic MDS based on linear and nonlinear input dissimilarity learning respectively. In both cases the topological ordering of points is similar moving from '0's on the left through '6's to '1's on the right of the latent space with bold face characters are located centrally. Slanted '1's are situated at the top of the latent spaces and the latent variable of orientation in '6's moves from left to right slants along the y-axis. The digits with poor caligraphy, particularly the '1's with bends and discontinuous '0's and '6's, are located at the top of the visualisation. The noticable difference between these two mappings is that the latent points, $\boldsymbol{y}_i$, are located upon an approximately uniform circular grid in figure 8b). On the other hand there is more clustering in the edges of the latent circular structure in figure 8a). The nonlinear learning of $D_x(i, j)$ has allowed the Elastic MDS cost function to be reduced from 20,565,002 in the linear case to 1,509 when the dissimilarity measure of equation 2 is better approximated.

The responsibilities, r_l, allocated to each dissimilarity measure in the training phase are shown in figure 7. Unlike in the linear input dissimilarity case of [41] where the Ware and Hedges dissimilarity measure was most significant

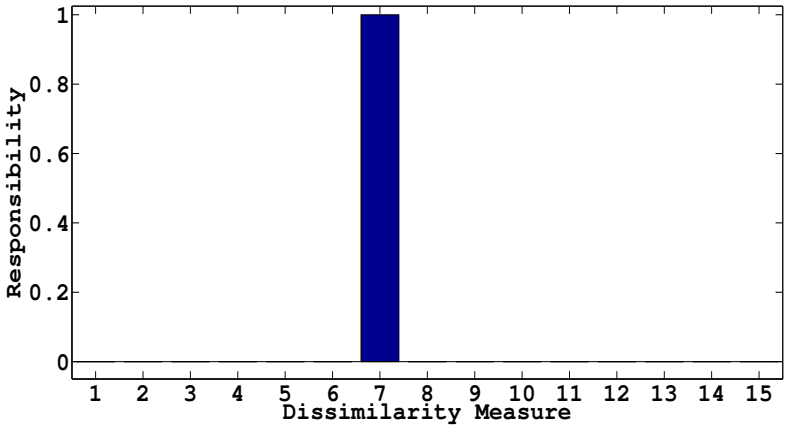
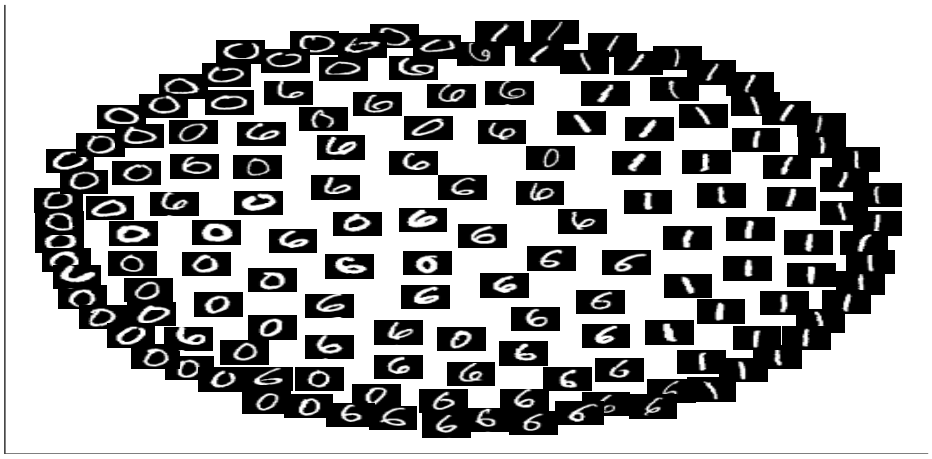


Figure 7: Responsibilities of the dissimilarity measures for the MNist mapping.

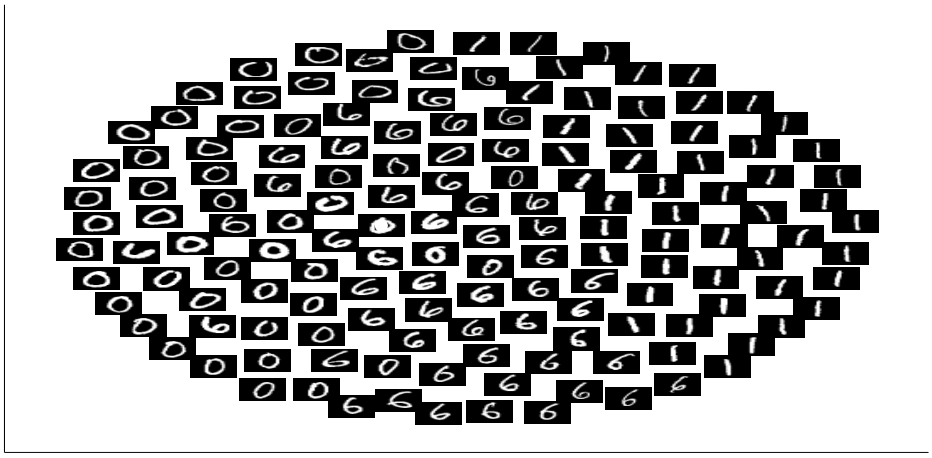
in mappings, here the median distance is sufficient to allow nonlinear interpolation of the latent variables. This would allow for a sparse solution in this case as a single dissimilarity measure accounts for nearly all of the responsibility in the mapping. No link between the responsibilities and the dissimilarity measures, $D^l(i,j)$ has been identified and since RBF networks perform a many-to-one mapping it is not expected that r_l will have a physical meaning beyond automatic relevence detection.

The visualisations of the MNist dataset usign Isomap on the learned dissimilarity measures are shown in figure 9. In the linear case a neighbourhood size of $k = 28$ was required however for $D_x(i,j)$ constructed using the method proposed in this paper a neighbourhood size of $k = 3$ was sufficient for the graph to be fully connected. The linear learned input dissimilarity visualisation of figure 9a) separates the class of '1's from the other two classes and does not separate out the poorly drawn digits. On the other hand the mapping of figure 9b) bears more resemblence to that of figure 8b) generating an approximately uniform filled circle with bold face characters at the central regions and disconnected '0's at the top. This visualisation is clearly more informative than the linear case of figure 9a).

The intrinsic dimensionality of these digits is estimated to be between 7 and 12 [18] and therefore all physical latent variables cannot be accounted for in a 2-dimensional latent mappings. In spite of this a successful visualisation must ensure the pivotal latent variables are interpretable, as the visualisations with nonlinear input dissimilarity learning do.

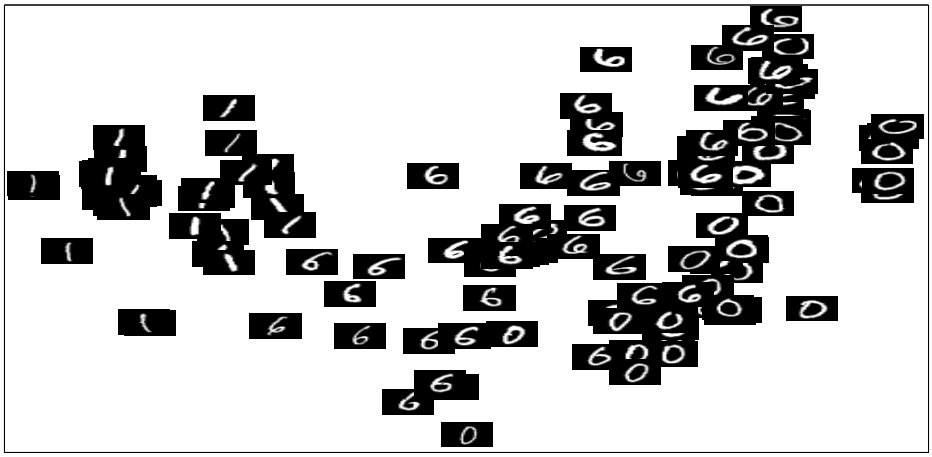


(a)

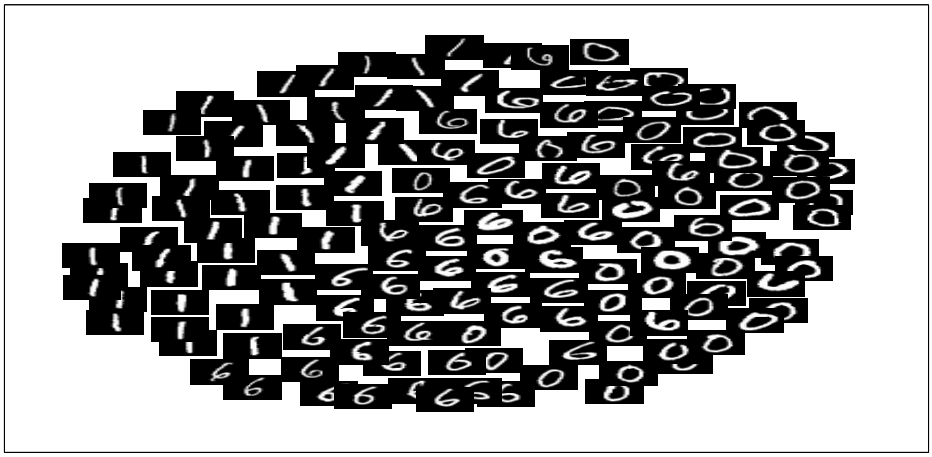


(b)

Figure 8: Visualisations of the MNist dataset using Elastic MDS with (a) linear and (b) nonlinear input dissimilarity learning.



(a)



(b)

Figure 9: Visualisations of the MNist dataset using Isomap with (a) linear and (b) nonlinear input dissimilarity learning.

3.3. Artificial Faces Dataset

The next dataset analysed in this paper is the artificial faces dataset [49] which was demonstrated with Isomap. The 698 observations consist of 64×64 pixel images, treated as 4096-dimensional vectors, where the face is subject to differing levels of light and camera orientation. These two latent variables should be clearly represented in any generated visualisations. Figure 10 shows the visualisation based on standard Elastic MDS for reference.

As with the MNist dataset the visualisations generated using Elastic MDS with linear and nonlinear dissimilarity learning, shown in figures 12a) and 12b) respectively, are very similar in terms of the topological ordering and neighbourhoods of points. The latent variable of lumination is mapped from dark on the left to light on the right with orientation changing as the latent points move around the circular structure. The nonlinear learned model does not posess the discontinuity present in the linear equivalent in figure 12a). The mapping stress for the linear model is 89,465,000 compared to only 32,533 using the method developed in this paper. The nonlinear model is naturally more flexible than the linear counterpart, allowing for instance the latent points to be situated further from one-another. One way of assessing this is to analyse the range of values in the two latent dimensions under each model. For the standard Elastic MDS case with Euclidean input dissimilarities the latent space is contained within the rectangle $\{(-380.05, -617.72), (426.90, 628.84)\}$. The Elastic MDS model with a linear mixture of input dissimilarities is contained within the larger rectangle $\{(-1,788.1 - 1,803.1), (1,886.3, 1,614.3)\}$. When the linear restrictions on the approximation of the natural dissimilarity over observations are removed in the nonlinear learned input dissimilarity model, the visualised points are contained within the much larger rectangle $\{(-5,017.9, -5,033.3), (5,051.3, 4,947.1)\}$. The larger spacing between latent points in the nonlinear model of figure 12b) allows for the reduction in stress when compared to the linear counterpart in figure 12a).

The responsibilities for each of the input dissimilarities, $D^l(i,j)$, are shown in figure 11. For the artificial faces dataset the dominant features are shared between 7 different dissimilarity measures, allowing a sparsity level of over 53%. As with the MNist dataset the median distance (measure 7) is significant in the generation of the visualisation.

The Isomap visualisations of the faces dataset with linear and nonlinear input dissimilarity learning are shown in figures 13a) and 13b) respectively. Both cases required a neighbourhood size of $k = 16$ to create a connected graph. For the linear case the latent space appears to peel the two regions of darker faces apart based on the difference in orientation. On the other hand when D_x is learned nonlinearly the latent space is more continuous moving from dark images on the left to lighter images on the right, as in figure 12a). The latent variable of orientation is again mapped around the oval structure on the more continuous latent structure. The approach of this paper therefore generates a more intuitive visualisation of this dataset also than that of [41].

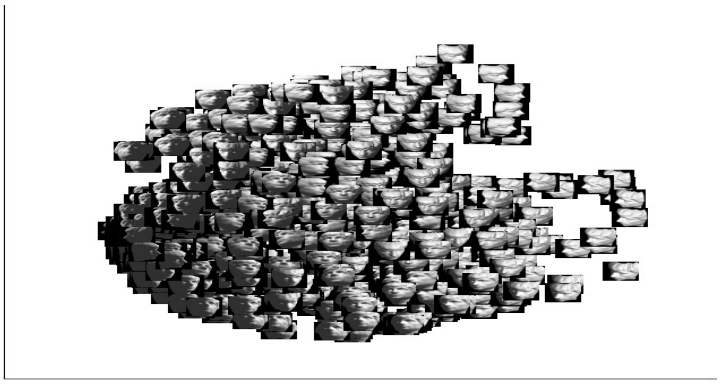


Figure 10: Visualisation of the Faces dataset using Elastic MDS based on Euclidean input distances. There are clear discontinuities on the right side of the latent space and outliers on the upper left which have clearly been removed from their neighbours.

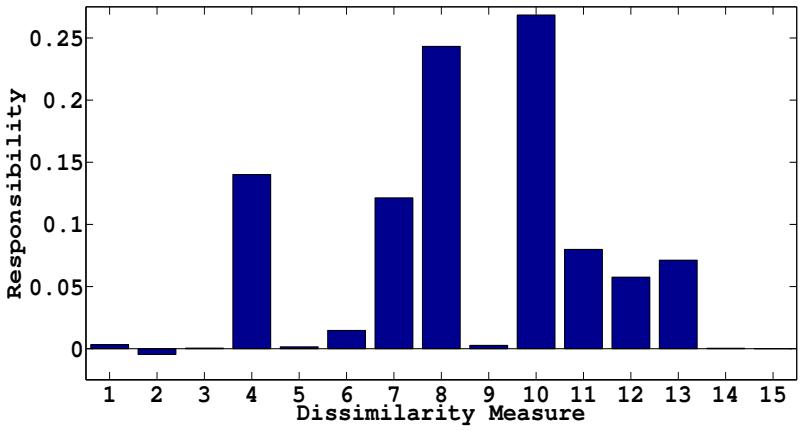
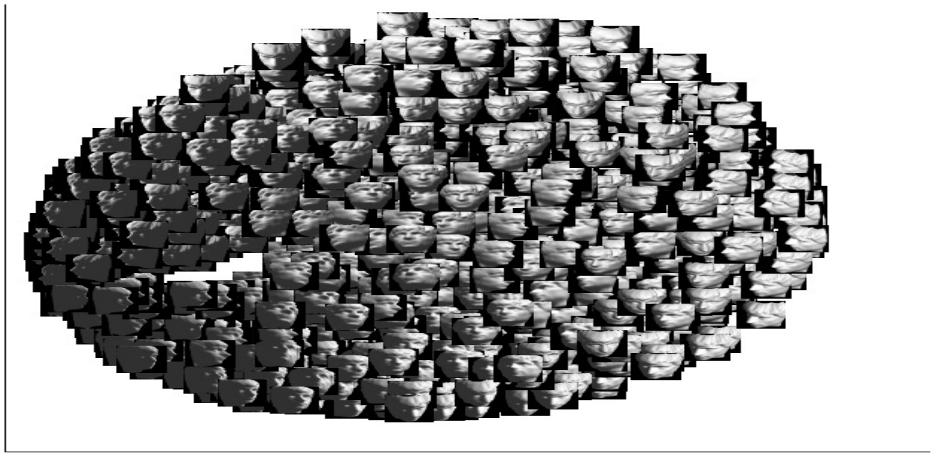
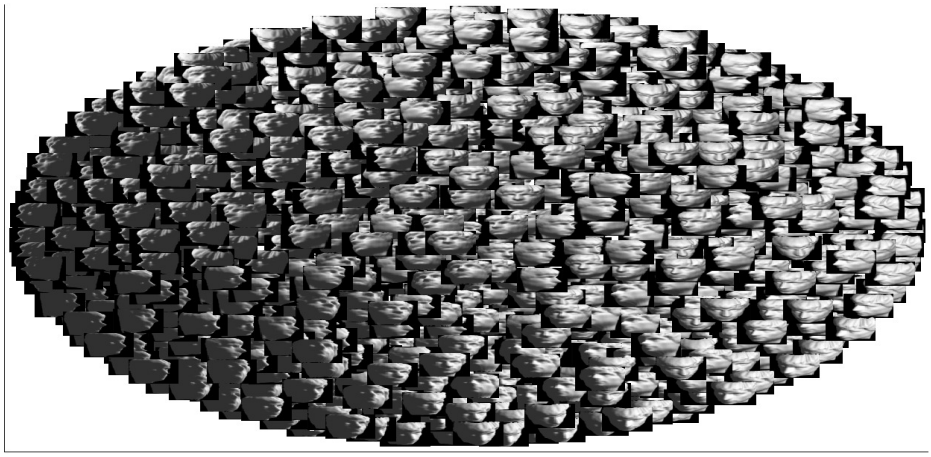


Figure 11: Responsibilities of the dissimilarity measures for the faces mapping.

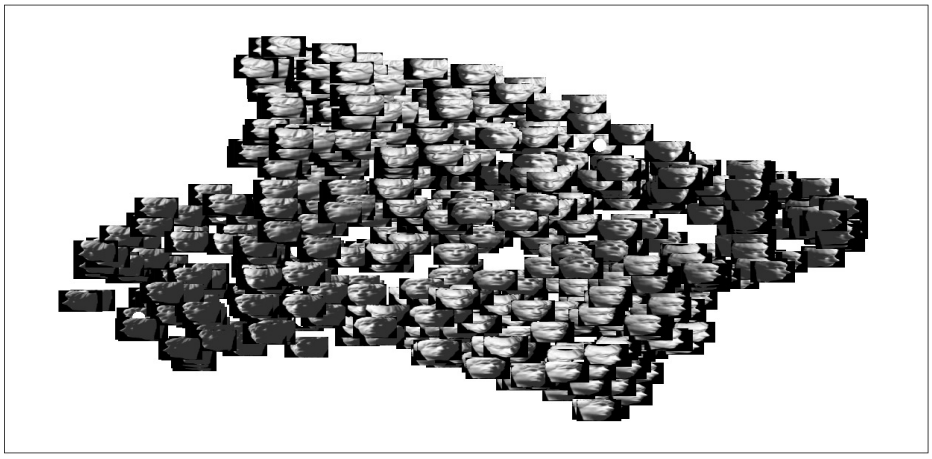


(a)

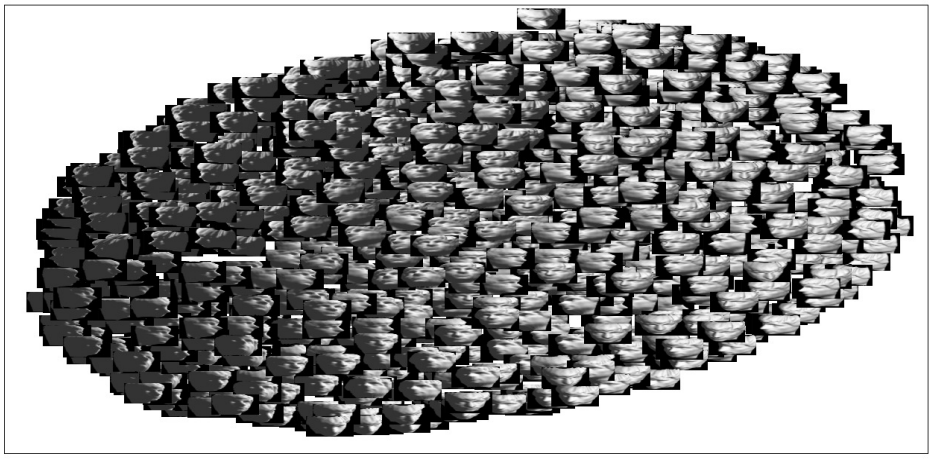


(b)

Figure 12: Visualisations of the Faces dataset using Elastic MDS with (a) linear and (b) nonlinear input dissimilarity learning.



(a)



(b)

Figure 13: Visualisations of the MNist dataset using Isomap with (a) linear and (b) nonlinear input dissimilarity learning.

3.4. Caltech101 Images Dataset

Our final experiment considers visualisation of the Caltech101 images dataset [13]. A subset of the images is used for analysis containing 52 aeroplanes, 52 Bonsai trees and 52 dollar bills. We first generate a SURF bag-of-words [5] in which the 156 observed images are mapped to 500-dimensional vectors as input to the Elastic MDS and Isomap algorithms. The visualisation based on standard Euclidean Elastic MDS is shown in figure 14.

The visualisations using linear and nonlinear input dissimilarity learning are shown in figures 16 and 17 respectively. As with the previous two datasets the topological ordering and neighbourhood structures are very similar, grouping dollar bills on the right and bonsai trees at the top of the latent space. Aeroplanes with grass present on the runways are placed in close proximity and darker bonsai tree images are mapped on the left side of the circular structure. In both cases the dollar bill and bonsai tree with red backgrounds are mapped as neighbours, but the bonsai tree in the lower right of the linear case with the white background is correctly placed among the other bonsai trees in the visualisation of figure 17. In addition to this the bonsai tree and dollar bill with brown wood backgrounds are neighbours in the nonlinear model visualisation of figure 17, placed at the upper left region, whereas in the linear equivalent they are separated. Further to this the latent points, $\boldsymbol{y}_i$, of the method proposed in this paper are again placed approximately uniformly across a latent circle with less clustered edges as in the linear case of figure 16. The mapping stress in the linear input dissimilarity case was 1,244,400 which was far higher than that of the nonlinear case, 1,628.

The responsibilities allocated to each dissimilarity measure are presented in figure 15. As with the MNist case the median distance (measure 7) is allocated a large responsibility, however here the weight is negative. This dominant feature will allow for a level of sparsity and therefore a more efficient model to be trained.

The Isomap visualisations of the Caltech101 dataset are shown in figures 18 and 19 for the linear and nonlinear model cases respectively. For the linear input dissimilarity case of figure 18, generated with a neighbourhood size of $k = 68$, there are three distinct clusters with some dollar bills and aeroplanes being incorrectly placed in the neighbourhood of the bonsai tree cluster. On the other hand the visualisation when the input dissimilarities are learned in a nonlinear fashion, generated with a small neighbourhood size of only $k = 5$, resembles the Elastic MDS cases of figures 16 and 17 placing dollar bills on the right with a smooth transition from aeroplanes with grass present to the similarly coloured bonsai trees. The dollar bill and bonsai tree with red backgrounds are placed close together, as are the bonsai tree and dollar bill with the brown wooden backgrounds at the top of the elliptical structure. Further to this the latent space is again approximately uniform across the ellipse, with latent points almost as regularly spaced as in the nonlinear Elastic MDS case of figure 17. This visualisation is clearly an improvement on the linear input dissimilarity model and therefore also better than the standard Elastic MDS and Isomap cases for generating visualisations.

Figure 14: Visualisation of the Caltech101 dataset using Elastic MDS with Euclidean input distances. There are notable issues with this representation such as the cluster of dollar bills on the left side which should be with the remainder of dollar bills on the right side. In addition an outlying dollar bill is contained within the region of aeroplanes on the lower left of the space.

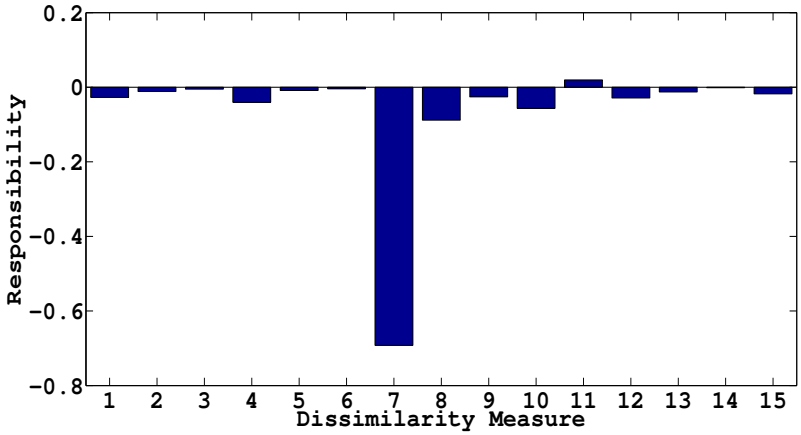


Figure 15: Weighting of the dissimilarity measures for the Caltech101 mapping.

Figure 16: Visualisation of the Caltech101 dataset using Elastic MDS with linear input dissimilarity learning.

Figure 17: Visualisation of the Caltech101 dataset using Elastic MDS with nonlinear input dissimilarity learning.

Figure 18: Visualisation of the Caltech101 dataset using Isomap with linear input dissimilarity learning.

Figure 19: Visualisation of the Caltech101 dataset using Isomap with nonlinear input dissimilarity learning.

Experiment	$D_x^l(i,j)$	$D_x(i,j)$	Φ, $\boldsymbol{W}$	Map	Standard EMDS
Swiss Roll	18.86	0.47	0.83	52.95	44.26
MNist	26.25	0.15	0.13	441.83	25.73
Faces	9737.13	5.20	14.76	659.78	47.00
Caltec	16.29	0.16	0.15	356.33	27.63

Table 2: Time in seconds to calculate the l components of $D_x^l(i,j)$, the initialisation of $D_x(i,j)$, the initialisation of the RBF network parameters Φ & $\boldsymbol{W}$ and the learning stage compared to that of standard Elastic MDS.

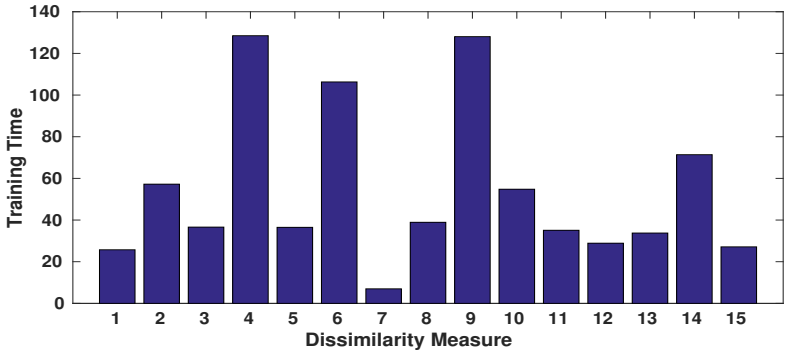


Figure 20: Time required to train a standard Elastic MDS mapping with input dissimilarity measure, $D_x(i,j)$, given by each of the measures taken from the proposed dictionary of table 1.

3.5. Computational Considerations

The time required to compute each of the above mappings is shown in table 2. The high cost for calulation of $D_x^l(i,j)$ for the Artificial Faces dataset is due to the high dimensionality (4096) of the observations. Naturally the experiments of our proposed method are more computationally expensive than the standard Elastic MDS with Euclidean input dissimilarities as we propose to learn not only a set of latent points, but the input dissimilarities also. This increase in time is worst for the MNist mapping with a slow down factor of over 17 times. Although this may seem a serious barrier to implementation we do not consider this an issue because we are incorporating a dictionary of different measures. As an alternative to our proposed learning approach a mapping should be computed with each dissimilarity measure in the dictionary in order to approximate a chart over the manifold. For the MNist dataset the time required to compute an Elastic MDS mapping with each of the 15 measures from the dictionary used in this paper is shown in figure 20. In total it would require 815.87 seconds of computation to generate the 15 mappings, following which these would require human interpretation, which is in excess of the 441.83 seconds required for our proposed approach.

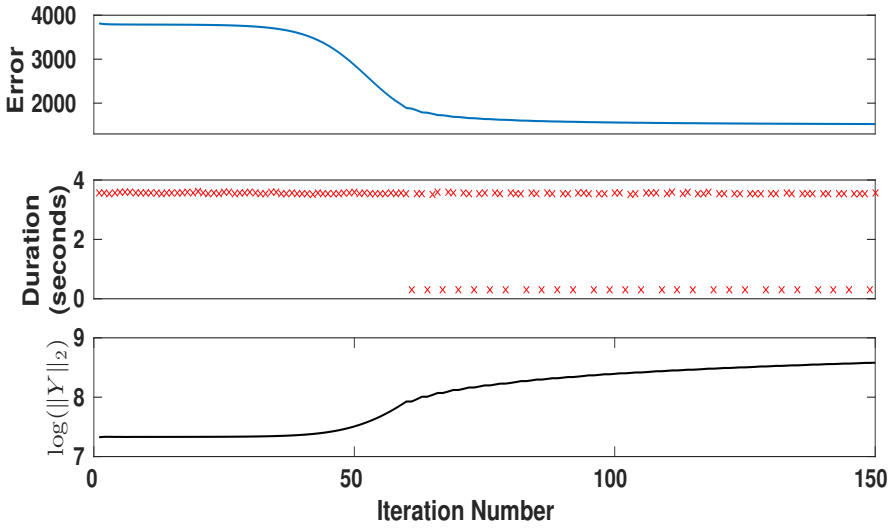


Figure 21: Plot showing error (top), duration per iteration (middle) and magnitude of latent points (bottom) during the training process. It is clear that the mapping focuses firstly on the minimisation of the error through movement of the latent points whilst retaining the same magnitude of latent points during the first 45 iterations. Beyond this point the mapping expands the latent space, increasing $\|\boldsymbol{Y}\|_2$, to cause a rapid decrease in error. During this period the relative positioning of the points does not change significantly. This is obvious through the lower training time requirement per iteration step, caused by the clipping parameter preventing miniscule gradient modifications.

When the mapping is trained in the approach of algorithm 1 we find an interesting two-stage development of the latent spaces. Figure 21 shows the changes in error, duration and latent space magnitude over time. There is a clear initial phase where the latent points are repositioned to minimise the cost function with the size of the latent space remaining unchanged post-initialisation. Following this phase the points $\boldsymbol{y}_i$ undergo an expansion and steps 8 to 16 of algorithm 1 are not required at each step.

3.6. Mapping Architecture

For the MNist visualisation we ran the experiment allowing for different configurations of the RBF network used to construct $D_x(i, j)$. Firstly we investigate the significance of the number of network prototypes, $\boldsymbol{v}_j$ used. Figure 22 plots the weight magnitude against iteration number for five different configurations of network where the percentage indicates the proportion of vectors $\boldsymbol{u}_i$ used as prototypes $\boldsymbol{v}_j$. The network clearly self-regularises preventing overtraining, as in NeuroScale [28], such that using more centres results in a lower weight

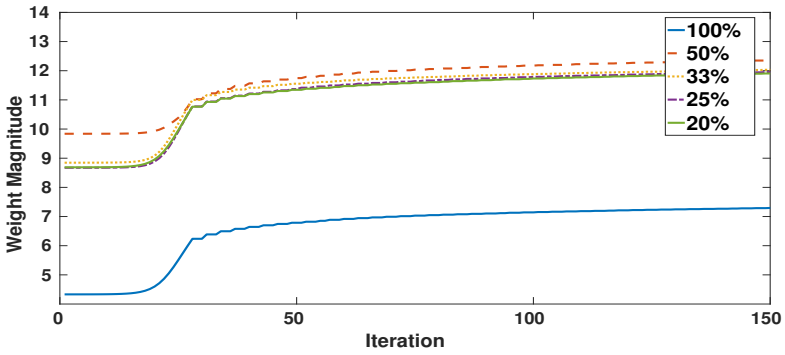


Figure 22: Weight magnitude (logarithmic scale) plotted against iteration number for the MNist dataset. Each curve corresponds to a different setup of RBF network where the percentage of datapoints $\boldsymbol{u}_i$ are used as network prototypes $\boldsymbol{v}_j$ where in the 100% case all inputs are used as network centres. In the other four cases the selected prototypes were chosen randomly.

Prototypes used	Mapping Stress
100%	**1,509**
50%	1,617
33%	1,716
25%	1,751
20%	1,764

Table 3: Mapping stress for different proportions of the RBF network inputs used as network prototypes.

magnitude. Furthermore, as shown in table 3, the mapping stress also reduces as the number of centres increases.

A further experiment we performed was to limit the size of the RBF network inputs, $\boldsymbol{u}_i$. We create a reduced dimensionality input vector $\tilde{\boldsymbol{u}}_i = [\tilde{\boldsymbol{d}}_i^1, \tilde{\boldsymbol{d}}_i^2, \ldots, \tilde{\boldsymbol{d}}_i^L]$ where $\tilde{\boldsymbol{d}}_i^1$ contains M elements with $M \leq N$. We investigated the impact on the MNist mapping when $\tilde{\boldsymbol{d}}_i$ contains 100%, 50%, 33%, 25% and 20% of the elements in $\boldsymbol{d}_i$. The mapping stress for each of these cases was constant at 1,509, however the weight magnitude varies in each case. Figure 23 presents the weight magnitude at during training for each of the setups. The highest weight magnitude, attributed to overtraining, is found in the case where $M = N$ at 100%, with the lowest achieved when $M = 0.2N$.

The optimal configuration for the above experiments was to fix $\boldsymbol{H} = \boldsymbol{I}$, using a large prototype set $\{\boldsymbol{v}_j\}_{j=1:N} = \{\tilde{\boldsymbol{u}}_i\}_{i=1:N}$, and a low dimensional input vector $\tilde{\boldsymbol{u}}_i$ with $M = 0.2N$.

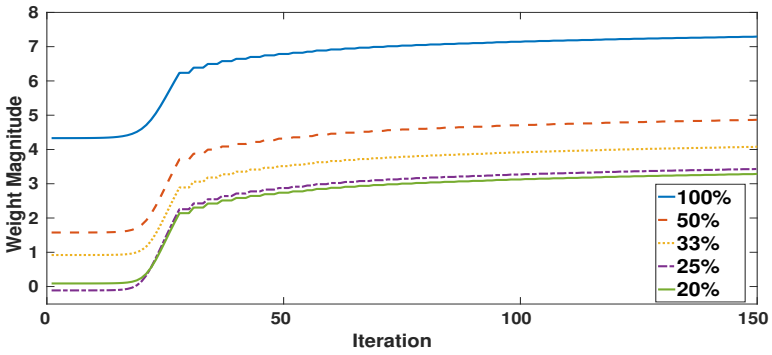


Figure 23: Weight magnitude (logarithmic scale) plotted against iteration number for the MNist dataset. Each curve corresponds to a different setup of RBF network where the percentage of dissimilarity vectors in $\tilde{\boldsymbol{u}}_i$ is varied. The 100% case corresponds to $\tilde{\boldsymbol{u}}_i = \boldsymbol{u}_i$ whereas the 20% case occurs when $\tilde{\boldsymbol{u}}_i$ only contains 20% of the elements in $\boldsymbol{u}_i$.

4. Conclusion

In this paper we have presented a novel method for generating a nonlinear chart over observations whilst simultaneously constructing a visualisation with Elastic MDS. The linear framework of [41] was extended to allow for nonlinear interpolation of the induced dissimilarity space. For each of the four visualisations presented in this paper the learned visualisations using the nonlinear input dissimilarity improved on the linear cases in terms of both mapping stress and the topological ordering of latent points. We have shown that the nonlinear chart can be combined with an ARD weight scheme to allow for sparse solutions, increasing efficiency. Further to this we have discussed the architecture of RBF network allowing for optimum visualisation performance while preventing overtraining.

Further research will test the extent to which the benefits of linear and nonlinear chart learning processes translate to other mappings such as Curvilinear Component Analysis, Stochastic Neighbour Embedding and Bregman MDS. Further to this we will investigate dissimilarity measures which form a robust dictionary to visualise non-vectorial observations such as probability distributions and graphs.

References

[1] V. Abrol, P. Sharma, and A. K. Sao. Greedy dictionary learning for kernel sparse representation based classifier. *Pattern Recognition Letters*, 78: 64 – 69, 2016. ISSN 0167-8655. doi: http://dx.doi.org/10.1016/j.patrec.2016.04.014. URL `http://www.sciencedirect.com/science/article/pii/S0167865516300666`.

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

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

[4] M. S. Baghshah, F. Afsari, S. B. Shouraki, and E. Eslami. Scalable semi-supervised clustering by spectral kernel learning. *Pattern Recognition Letters*, 45:161 – 171, 2014. ISSN 0167-8655. doi: http://dx.doi.org/10.1016/j.patrec.2014.02.014. URL `http://www.sciencedirect.com/science/article/pii/S0167865514000555`.

[5] H. Bay, T. Tuytelaars, and L. 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.

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

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

[8] D. Broomhead and D. Lowe. Multivariable functional interpolation and adaptive networks. *Complex Systems*, 2(3):321 – 355, 1988.

[9] 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.

[10] K. M Carter, R. Raich, W. G. Finn, and A. O. Hero III. Fine: Fisher information nonparametric embedding. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 31(11):2093–2098, November 2009.

[11] 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.

[12] R. P. W. Duin, E. Pekalska, and M. Loog. *Non-Euclidean Dissimilarities: Causes, Embedding and Informativeness*, pages 13–44. Springer London, London, 2013. ISBN 978-1-4471-5628-4. doi: 10.1007/978-1-4471-5628-4_2. URL `http://dx.doi.org/10.1007/978-1-4471-5628-4_2`.

[13] 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.

[14] A. L. N. Fred, A. Lourenço, H. Aidos, Samuel Rota B., N. Rebagliati, M. A. T. Figueiredo, and M. Pelillo. *Learning Similarities from Examples Under the Evidence Accumulation Clustering Paradigm*, pages 85–117. Springer London, London, 2013. ISBN 978-1-4471-5628-4. doi: 10.1007/978-1-4471-5628-4_5. URL `http://dx.doi.org/10.1007/978-1-4471-5628-4_5`.

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

[16] I. J. Goodfellow, J. Shlens, and C. Szegedy. Explaining and harnessing adversarial examples. In *International Conference on Learning Representations*, number 3, pages 29–41, 2015.

[17] L.A. Gottlieb, A. Kontorovich, and R. Krauthgamer. Adaptive metric dimensionality reduction. *Theoretical Computer Science*, 620:105 – 118, 2016. ISSN 0304-3975. doi: http://dx.doi.org/10.1016/j.tcs.2015.10.040. URL `http://www.sciencedirect.com/science/article/pii/S0304397515009469`. Algorithmic Learning Theory.

[18] 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.

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

[20] H. Kantz and T. Schreiber. *Nonlinear Time Series Analysis*. Cambridge University Press, 2 edition, 2004.

[21] 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.

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

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

[24] 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.

[25] 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.

[26] 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.

[27] 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.

[28] 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.

[29] D.J.C. Mackay. *Bayesian methods for backprop networks*, chapter 6, pages 211–254. Models of Neural Networks, III. Springer, 1994.

[30] A. Marinoni and P. Gamba. An efficient approach for local affinity pattern detection in remotely sensed big data. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 8(10):4622–4633, Oct 2015. ISSN 1939-1404. doi: 10.1109/JSTARS.2015.2485401.

[31] A. Marinoni and P. Gamba. Unsupervised data driven feature extraction by means of mutual information maximization. *IEEE Transactions on Computational Imaging*, 3(2):243–253, June 2017. ISSN 2333-9403. doi: 10.1109/TCI.2017.2669731.

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

[33] A. Nazarpour and P. Adibi. Two-stage multiple kernel learning for supervised dimensionality reduction. *Pattern Recognition*, 48(5):1854 – 1862, 2015. ISSN 0031-3203. doi: http://dx.doi.org/10.1016/j.patcog.2014.12.001. URL `http://www.sciencedirect.com/science/article/pii/S0031320314004890`.

[34] R.M. Neal. *Bayesian Learning for Neural Networks*. Number 118 in Lecture Notes in Statistics. Springer, 1996.

[35] J. Park and I.W. Sandberg. Universal approximation using radial-basis-function networks. *Neural Computation*, 3:246–257, 1991.

[36] 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.

[37] M. Pelillo, editor. *Similarity-Based Pattern Analysis and Recognition*. Advances in Computer Vision and Pattern Recognition. Springer, 2013. ISBN 978-1-4471-5627-7. doi: 10.1007/978-1-4471-5628-4. URL `https://doi.org/10.1007/978-1-4471-5628-4`.

[38] M.J.D Powell. *The Theory of Radial Basis Function Approximation*, volume 2 of *Advances in Numerical Analysis*, chapter 3. Oxford Science Publications, 1990.

[39] H. M. Rekabdarkolaee, E. Boone, and Q. Wang. Robust estimation and variable selection in sufficient dimension reduction. *Computational Statistics & Data Analysis*, 108:146 – 157, 2017. ISSN 0167-9473. doi: http://dx.doi.org/10.1016/j.csda.2016.11.007. URL `http://www.sciencedirect.com/science/article/pii/S0167947316302717`.

[40] P. Ren, F. Aziz, L. Han, E. Xu, R. C. Wilson, and E. R. Hancock. Geometricity and embedding. In *Similarity-Based Pattern Analysis and Recognition*, pages 121–155. 2013. doi: 10.1007/978-1-4471-5628-4_6. URL `https://doi.org/10.1007/978-1-4471-5628-4_6`.

[41] I. Rice. Improved data visualisation through multiple dissimilarity modelling. *Information Sciences*, 370-371:288–302, 8 2016. ISSN 0020-0255. doi: 10.1016/j.ins.2016.07.073.

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

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

[44] I.W. Sandberg. Gaussian radial basis functions and inner product spaces. *Circuit Systems Signal Processing*, 20(6):635–642, 2001.

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

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

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

[48] F. Takens. *Detecting strange attractors in turbulence*, volume 898 of *Dynamical Systems and Turbulence, Lecture Notes in Mathematics*, pages 366–381. Springer-Verlag, 1981.

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

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

[51] J. Wang, Z. Deng, K.S. Choi, Y. Jiang, X. Luo, F.L. Chung, and S. Wang. Distance metric learning for soft subspace clustering in composite kernel space. *Pattern Recognition*, 52:113 – 134, 2016. ISSN 0031-3203. doi: http://dx.doi.org/10.1016/j.patcog.2015.10.018. URL `http://www.sciencedirect.com/science/article/pii/S0031320315003970`.

[52] 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.

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

[54] H. Whitney. The self-intersections of a smooth n-manifold in $2n$-space. *The Annals of Mathematics, Second Series*, 45(2):220–246, April 1944.

[55] X. Wu, Q. Li, L. Xu, K. Chen, and L. Yao. Multi-feature kernel discriminant dictionary learning for face recognition. *Pattern Recognition*, 66:404 – 411, 2017. ISSN 0031-3203. doi: http://dx.doi.org/10.1016/j.patcog.2016.12.001. URL `http://www.sciencedirect.com/science/article/pii/S0031320316303880`.

[56] D. Zhang, Q. Zhu, and D. Zhang. Multi-modal dimensionality reduction using effective distance. *Neurocomputing*, 2017. ISSN 0925-2312. doi: http://dx.doi.org/10.1016/j.neucom.2016.07.075. URL `http://www.sciencedirect.com/science/article/pii/S0925231217302461`.

[57] X. Zhu, Z. Huang, H. T. Shen, J. Cheng, and C. Xu. Dimensionality reduction by mixed kernel canonical correlation analysis. *Pattern Recognition*, 45(8):3003 – 3016, 2012. ISSN 0031-3203. doi: http://dx.doi.org/10.1016/j.patcog.2012.02.007. URL `http://www.sciencedirect.com/science/article/pii/S0031320312000696`.

Author Biography

Iain is a postdoctoral research associate as Aston University applying machine learning and dimension reduction in the healthcare and defence domains. He received his PhD from Aston University in 2015 with the thesis title 'Probabilistic Topographic Information Visualisation'.