Non-Linear Dimensionality Reduction : A Comparative Performance Analysis

Olivier de Vel and Sofianto Li
Department of Computer Science,
James Cook University,
Townsville Q4811, Australia.
olivier@cs.jcu.edu.au

Danny Coomans
Dept of Mathematics and Statistics,
James Cook University,
Townsville Q4811, Australia.
Danny.Coomans@jcu.edu.au

Abstract

We present an analysis of the comparative performance of non-linear dimensionality reduction methods such as Non-Linear Mapping, Non-Metric Multidimensional Scaling and the Kohonen Self-Organising Feature Map for which data sets of different dimensions are used. To obtain comparative measures of how well the mapping is performed, Procrustes analysis, the Spearman rank correlation coefficient and the scatter-plot diagram are used. Results indicate that, in low dimensions, Non-Linear Mapping has the best performance especially when measured in terms of the Spearman rank correlation coefficient. The output from the Kohonen Self-Organising Feature Map is easier to interpret than the output from the other methods as it often provides a superior qualitative visual output. Also, the Kohonen Self-Organising Feature Map may outperform the other methods in a high-dimensional setting.

1 Introduction

In many applications, dimensionality reduction is used to explore a data set to try to obtain some insight into the nature of the phenomenon that produced the data. We are often interested in understanding the structural relationships that exist in the feature space, such as clusters or data point density discontinuities. Measures that could be used to reveal such structural relationships include the inter-point distance, the "shape" of the data distribution etc...Such an understanding is very domain-dependent and may require a deep understanding of the structures, causality etc...that may exist between the features. In some cases, the application may impose a constraint on the serialisation or ordering of the topological mapping as in, for example, chronological ordering in a time series or regression analysis [CLN91].

In many applications, one important objective of dimensionality reduction is that it preserves as much as possible the structural relationships that exist in the data set when performing the mapping from a high dimension to a lower one (usually two or three dimensions) while, at the same time, removing any redundancy in the data. For data characterised by low dimensions (i.e. where the number of features is small), the "important" or "interesting" relationships may be detected relatively easily by using pictorial methods such as histograms (one or two dimensions), scatter-plot diagrams, or even kinematic graphic techniques (such as small angle rigid body rotations in three dimensions). For a high-dimensional feature space, we have to resort to techniques that reduce or project the feature space into one, or more,

lower two- or three-dimensional representations. Such techniques must be able to effectively handle the "curse of dimensionality" (due to the fact that high-dimensional space is sparse for small-to-medium sample sizes) and be able to ignore redundant and noisy features.

The most popular dimensionality reduction techniques are linear transformations, such as principal components analysis (PCA) [Fuk90]. However, PCA and many other similar approaches assume a linear constraint of input space and therefore would not perform satisfactorily for non-linear constraints of input space common in high-dimensional data. Furthermore, the covariance structure of data does not necessarily relate to the clustering of data points. Some limited work has been undertaken on the comparative performance of PCA as a dimensionality reduction technique (e.g. [BD91] and [MHP94]).

More general approaches to dimensionality reduction are those of non-linear methods which do not impose any input space constraints. Such techniques include Non-Linear Mapping (NLM) [Sam69], Multidimensional Scaling (MDS) [BL87] and various clustering algorithms [Fuk90]. Several neural networks techniques have also been introduced more recently as in, for example, Kohonen Self-Organising Feature Map [Koh88] and the Back Propagation algorithm [Sau89]. However, neither the comparative performances nor the quality of the results produced by these methods have been sufficiently investigated.

In this paper, we present three non-linear dimensionality reduction techniques: Non-metric Multidimensional Scaling, Non-Linear Mapping and the Kohonen Self-Organising Feature Map (SOFM), and provide a comparative analysis of the performance of each technique. We briefly give the theoretical background for each method and the techniques used for evaluating the comparative performance of each method.

2 Dimensionality Reduction Methods

2.1 Kohonen Self-Organising Feature Map

The Kohonen SOFM algorithm [Koh88] attempts to produce a distorted, but topographically-organised, projection of the input space where the similarity of the input vectors is converted into a proximity relationship in the projection. The map will attempt to preserve the "important" similarity relationships present in the high-dimensional input space while, at the same time, factoring out any redundancy latent in the input features [LGZ93].

2.2 Multidimensional Scaling

Multidimensional Scaling (MDS) is based on dissimilarity data which reflect the amount of dissimilarity between pairs of objects, events or concepts [Kru64]. Typically, the dissimilarity data are distances between all pairs of points in multidimensional space. Thus similar objects are close together and dissimilar objects are far apart. Consider a set of input patterns (also called *configuration*) $\{\mathbf{x}_i\}$ in n-dimensional space, with dissimilarity data which can be calculated from inter-point distances $\delta_{rs} = \| \mathbf{x}_r - \mathbf{x}_s \|$. The method attempts to find a configuration $\{\mathbf{y}_i\}$ in m-dimensional space, where $m < n$, with inter-point distances $d_{rs} = \| \mathbf{y}_r - \mathbf{y}_s \|$ such that $d_{rs} \approx \delta_{rs}$ for all r and s. In achieving $d_{rs} \approx \delta_{rs}\ \forall\ r$ and s, a transformation from $\{\mathbf{x}_i\}$ to $\{\mathbf{y}_i\}$ is effected. MDS models with such dissimilarity data, together with linear transformations, are known as *metric Multidimensional Scaling* [BL87]. Rather than using distance magnitudes, it is possible to preserve the rank ordering of d_{rs} to be the same as that of δ_{rs}. In this case, the search for the $\{\mathbf{y}_i\}$ configuration should satisfy :

$$d_{rs} \approx f(\delta_{rs}) \tag{1}$$

where f is a monotonically increasing function satisfying:

$$\delta_{r_1 s_1} < \delta_{r_2 s_2} \iff f(\delta_{r_1 s_1}) < f(\delta_{r_2 s_2}) \tag{2}$$

for some r_1, s_1 and r_2, s_2. This variant is referred to as *non-metric Multidimensional Scaling*.

Major operations on non-metric MDS are iterative calculations of the $\{\mathbf{y}_i\}$ configuration for which the monotonicity between d_{rs} and δ_{rs} must always hold. In order to ensure this condition, a measure of departure from monotonicity is defined so that adjustments on the $\{\mathbf{y}_i\}$ configuration can be made to improve the degree of monotonicity. This measure is called the *stress* value, S. The algorithm involves computing $\hat{d}_{rs}$ which minimises [Kru64]:

$$S^2 = \frac{\sum\sum_{r<s}(d_{rs} - \hat{d}_{rs})^2}{\sum\sum_{r<s} d_{rs}^2} \qquad (3)$$

while holding the monotone constraint:

$$\hat{d}_{r_1 s_1} \leq \hat{d}_{r_2 s_2} \leq \ldots \leq \hat{d}_{r_m s_m} \qquad (4)$$

Typically, $\hat{d}_{rs}$ are computed using a monotonic regression algorithm and the iterative calculation of the m-dimensional configuration $\{\mathbf{y}_i\}$ is achieved by using the gradient descent method.

2.3 Non-Linear Mapping

Non-Linear Mapping (NLM) is a similar concept to that of MDS and was introduced by Sammon in 1969 [Sam69]. Similar to metric MDS, NLM tries to preserve distances between points in the original data and the reduced dimensional data. The NLM method is also based on iterative calculations of the $\{\mathbf{y}_i\}$ configuration which minimises the error:

$$E = \frac{\sum\sum_{r<s} \frac{(\delta_{rs} - d_{rs})^2}{\delta_{rs}}}{\sum\sum_{r<s} \delta_{rs}} \qquad (5)$$

As in the case of MDS, gradient descent is a typical minimisation procedure employed. In all SOFM, non-metric MDS and NLM methods the gradient descent procedure may become trapped in local minima.

3 Comparative Performance Analysis Techniques

Two methods of comparing the performance of each dimensionality reduction technique are presented: Procrustes analysis and the Spearman rank correlation coefficient. A third, empirical method, the *scatter-plot*, is used to plot inter-point distances in the original data space versus distances in the reduced dimension data space. This enables the data correlation to be visualised as a two-dimensional scatter-plot.

3.1 Procrustes Analysis

The Procrustes analysis method is aimed at measuring how well the shapes of two data configurations match one another. The "shape" of a configuration is commonly understood to refer to the geometrical attributes that remain invariant when the configuration is subject to a rigid body transformation (translation and rotation) and dilatation.

Let the input data configuration, viewed as a geometrical figure in $\Re^n$, consist of p labeled points and represented by a $p \times n$ matrix $\mathbf{X}$. Similarly, define the output data configuration as a $p \times m$ matrix $\mathbf{Y}$ in $\Re^m$, where $m \leq n$. It is assumed that the output configuration $\mathbf{Y}$ is in the $\mathbf{X}$ subspace i.e. point i in $\mathbf{X}$, $\mathbf{X}^{(i)}$, corresponds to point i in the $\mathbf{Y}$ configuration, $\mathbf{Y}^{(i)}$. In an attempt to match configurations $\mathbf{X}$ and $\mathbf{Y}$, $n - m$ columns of zeros are added to the $\mathbf{Y}$ configuration to obtain $m = n$. A typical measure of the degree of coincidence between the two configurations is the sum of the square of distances between corresponding points (M^2), i.e.

$$M^2 = \sum_{i=1}^{p} ||\mathbf{X}^{(i)} - \mathbf{Y}^{(i)}||^2 \qquad (6)$$

The two configurations are first translated, rotated and dilatated to obtain the best possible fit. M^2 measures the "lack of fit", or *residual sum of squares*. Small values of M^2 result in a good match between two given shapes [Kar82].

3.2 Spearman Rank Correlation Coefficient

The Spearman rank correlation coefficient is a method which measures how much the rank-

ing of two groups of data agree with one another. Using the distance data calculated from scatter-plot diagrams (a graph of the inter-point distances in the original and reduced data spaces), the coefficient provides a measure of the correlation between the original data set and the reduced dimension data set.

The Spearman rank correlation coefficient, Γ, is defined as follows [Hay81]:

$$\Gamma = 1 - \frac{6 \sum_{l=1}^{N} D_l^2}{N(N^2 - 1)} \qquad (7)$$

where D_l is the scalar difference between each element of two ranked vectors l and N is the number of data. A strong correlation is indicated by a value close to 1, 0 meaning no correlation at all, and -1 meaning irrelevance.

4 Performance Results and Discussion

Experiments are undertaken on three data sets; (i) thyroid data, (ii) glass data, and (iii) fish data. The dimension of these data sets is 5, 10 and 12, respectively; the number of input pattern vectors is 100, 50 and 34, respectively, and the number of classes is 3, 4 and 3, respectively. The thyroid and glass data sets include outliers. All experiments produce two-dimensional data from the high dimensional input data, and the Euclidean distance is used as a standard distance measure in all techniques. All input data are transformed using one or more of the following: the Z-transform, the log transform, mean-centering and double-centering. For the SOFM, the algorithm is run several times with different parameters (size of feature map, number of training iterations etc...), for each data set, to achieve optimal performance (minimal over-fitting and maximum output resolution). A typical map size was 10×10, and the number of iterations was 80,000. Scatterplots are generated from the output produced by each reduction method and the comparative performance is evaluated by using Procrustes analysis and the Spearman rank correlation coefficient. More details of the experiments and optimisation procedure are given in [LdVC94] .

Results of the experiments are summarised in Figures 1a) and 1 b). For completeness, results for the linear PCA method are also included. As to be expected, the performance of the PCA techniques degrades rapidly as the dimensionality of the data increases.

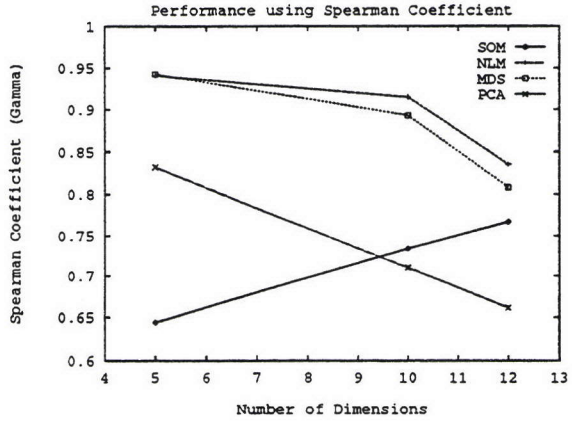
NLM and, to a lesser extent, non-metric MDS appear to have an overall best performance in terms of the Spearman coefficient, whereas non-metric MDS has the best performance in terms of Procrustes analysis, particularly for low input dimensions. Non-metric MDS has a better performance in terms of reducing the residual sum of squares and preserving the shape of the data point distribution, whereas NLM is superior in terms of the rank correlation between the input and reduced data sets.

The SOFM algorithm generates output maps with better visualisation and interpretation capabilities; data points are more evenly distributed in the output space with the classes often clearly distinguished. However, in general, the performance of the SOFM algorithm is inferior to the other non-linear dimensionality reduction techniques in low dimensional input spaces. Some of the structure in the data point distribution is lost in the mapping process. Also, the SOFM algorithm is not always sensitive to outliers (particularly in the case of the thyroid data set), whereas outliers are easily identifiable when using NLM or MDS.

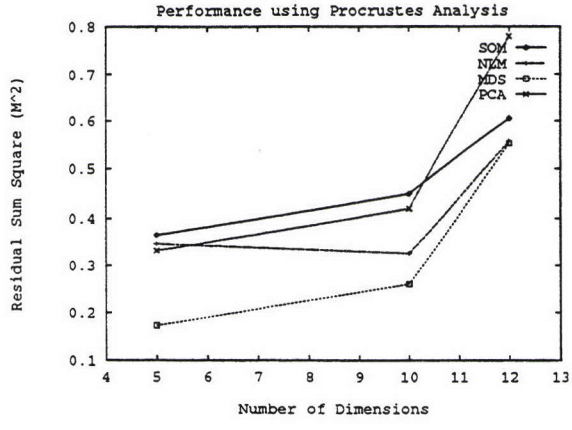
As the dimensionality of the input space increases, the relative performance of the SOFM algorithm improves significantly (particularly when using the Spearman rank correlation coefficient as the performance measure, where the value of the coefficient actually increases with dimensionality) and the SOFM algorithm may actually outperform both NLM and non-metric MDS when the dimension of input data is high. This suggests that the SOFM algorithm would seem to avoid the curse of dimensionality better than the other non-linear methods. Higher dimensionality may also improve the convergence

rate of the SOFM algorithm, although this was not verified (see, however, [MSRK91]).

All dimensionality reduction methods, particularly the NLM and non-metric MDS, were shown to be sensitive to the presence of outliers. The presence of an outlier produced an unbalanced and distorted output map. This is as expected, given the sensitivity of the Euclidean metric to distance. Careful preprocessing (which was undertaken in the experiments reported here) must be implemented to remove all outliers and ensure robust dimensionality reduction.



(a)



(b)

Figure 1: *Comparative performance measured in terms of a) Spearman coefficient, and b) Procrustes analysis.*

5 Conclusions

The use of Procrustes analysis and the Spearman rank correlation coefficient in comparing and contrasting non-linear dimensionality reduction methods was described. Both non-metric MDS and Non-Linear Mapping perform relatively well for low dimensional data. Non-metric MDS has a better performance in terms of minimising the residual sum of squares and preserving the shape of the data point distribution. NLM is superior to non-metric MDS in terms of the rank correlation between the input and reduced data sets. The performance of both non-metric MDS and NLM degrade quite rapidly as the dimensionality increases.

The Self-Organising Feature Map generates a more superior visual output than the other two methods in that it is relatively easy to interpret and the data points are well-clustered. However, in low dimensions, the performance of the SOFM algorithm is inferior to that of NLM and non-metric MDS. Unlike NLM and MDS, the performance of the SOFM algorithm increases with dimensionality (when performance is measured in terms of the Spearman rank correlation coefficient) and may outperform non-metric MDS and NLM for high-dimensional data. It would seem that the SOFM algorithm avoids the curse of dimensionality better than the other non-linear methods.

Further work using large scale simulations will need to undertaken to establish more rigorously the superiority of the SOFM in a high-dimensional context and its ability to ignore noisy and information-poor features.

References

[BD91] F. Blayo and P. Demartines. "Data analysis: how to compare Kohonen neural networks to other techniques". *International Joint Conference on Neural Networks*, 1:469–476, 1991.

[BL87] I. Borg and J. Lingoes. *Multidimensional Similarity Structure*

Analysis. Springer-Verlag, New York, 1987.

[CLN91] V. Cherkassky and H. Lari-Najafi. "Constrained topological mapping for nonparametric regression analysis". *Neural Networks*, 4:27–40, 1991.

[Fuk90] K. Fukunaga. *Introduction to Statistical Pattern Recognition*. Academic Press, Inc., New York, 1990.

[Hay81] W.L. Hays. *Statistics*. CBS College Publishing, New York, 1981.

[Kar82] M.J. Karson. *Multivariate Statistical Methods : An Introduction*. The Iowa State University Press, Ames, Iowa, 1982.

[Koh88] T. Kohonen. *Self-organization and Associative Memory*. Springer-Verlag, Heidelberg, 1988.

[Kru64] J.B. Kruskal. "Nonmetric multidimensional scaling: a numerical method". *Psychometrika*, 29:115–129, 1964.

[LdVC94] S. Li, O. de Vel, and D. Coomans. *Comparative performance analysis of non-linear dimensionality reduction methods*. Technical report, Department of Computer Science, James Cook University, Townsville, Australia, 1994.

[LGZ93] X. Li, J. Gasteiger, and J. Zupan. "On the topology distortion in self-organizing feature maps". *Biological Cybernetics*, 70:189–198, 1993.

[MHP94] F. Murtagh and M. Hernández-Pajares. *The Kohonen self-organizing map method: An assessment*. Technical report, Dept de Matemàtica Aplicada i Telemàtica, Universitat Politècnica de Catalunya, Barcelona, Spain, 1994.

[MSRK91] W.J. Melssen, J.R.M. Smits, G.H. Rolf, and G. Kateman. *Two-dimensional mapping of infrared spectra using a parallel implemented self-organising feature map*. Technical report, University of Nijmegen, 1991.

[Sam69] J.W. Sammon. "A nonlinear mapping for data structure analysis". *IEEE Transactions on Computers*, 18:401–409, 1969.

[Sau89] E. Saund. "Dimensionality-reduction using connectionist networks". *IEEE Transaction on Pattern Analysis and Machine Intelligence*, 11:304–314, 1989.