A Graphical User Interface for the Exploratory Analysis of High-Dimensional Data Using ISOMAP

Wendy L. Martinez
Office of Naval Research, Arlington, Virginia 22217-5660
Angel R. Martinez
Naval Surface Warfare Center Dahlgren Division, Dahlgren, Virginia 22448

KEY WORDS: Nonlinear dimensionality reduction, ISOMAP, graphical exploratory data analysis, parallel coordinate plots, Andrews curves, scatterplot matrices, high-dimensional data

Abstract

The ISOMAP nonlinear dimensionality reduction method of Tenenbaum, de Silva and Langford, was originally implemented in MATLAB by the developers of the algorithm. One of the issues involved with ISOMAP is the need to determine the number of reduced dimensions that best represents the original data. For this purpose, Tenenbaum, de Silva and Langford provide a plot similar to the scree plot in principal component analysis, where the elbow in the curve represents an estimate of the intrinsic dimensionality. However, for many data sets, the elbow is sometimes difficult to see. Thus, it would be useful to have a Graphical User Interface (GUI) that allows one to explore the results of ISOMAP via scatterplots, parallel coordinates, Andrews curves, and other EDA methods to better understand the effects of dimensionality reduction and to determine the minimum number of dimensions to use. This paper is tutorial in nature and demonstrates a MATLAB GUI for the graphical exploratory analysis of the results of ISOMAP.

1. Introduction

Say we have a set of p-dimensional data, where each observation is of the form $(X_1, ..., X_p)$. Dimensionality reduction is the process of reducing the p dimensions to some number $d < p$. Techniques for accomplishing this include principal component analysis, factor analysis, and multidimensional scaling [Jackson, 1991; Cox & Cox, 2001]. Several recent nonlinear methods have been developed that have interesting properties. These are ISOMAP [Tenenbaum, et al., 2000], local linear embedding [Roweis & Saul, 2000], and Hessian eigenmaps [Donoho & Grimes, 2003]. We focus on the output from ISOMAP in this paper, but the ideas are appropriate for any method.

The motivation for this GUI arose from the work of Martinez [2002], where one of the goals was to cluster unstructured text documents. In this case, the true topic labels for the documents were available, so that information could be used in the process. In conducting this research, Martinez had to evaluate the results from over 100 experiments, where different parameters were changed (e.g., text pre-processing, measure of semantic similarity, k values, etc.) before dimensionality reduction using ISOMAP.

The results of these experiments had to be processed and the best set of reduced features extracted for further analysis (clustering and classification). Martinez used techniques from graphical exploratory data analysis to help in the selection. These techniques include scree plots of residuals, scatterplots, parallel coordinate plots and Andrews curves. His goal was to find a reduced set of features where the topics were readily visible in the reduced space. Processing the output from these experiments would be tedious using a command line interface, so a GUI was developed. The GUI tool described in this paper is an extension of the one in Martinez [2002].

In this paper, we first provide a brief description of the ISOMAP nonlinear dimensionality reduction procedure. This is followed by a discussion of the capabilities included in the GUI, along with screen shots showing some of the options. We conclude with a summary and some future directions.

2. ISOMAP

The goal of ISOMAP is to find a set of d-dimensional coordinates for data that lie on a manifold that is embedded in a p-dimensional space. These coordinates should preserve the topological structure of the data, meaning that Euclidean distances in the d-dimensional space should correspond to distances between the points along the manifold [Tenenbaum, et al., 2000]. The basic steps of the algorithm consist of the following: 1) construct a neighborhood graph using the interpoint distances, 2) calculate the graph distance (the smallest path between the points, where the length of the path is the sum of its edges), 3) apply multidimensional scaling using the geodesic or graph distances. The last step yields a lower dimensional embedding such that the neighborhood structure is preserved.

We illustrate this idea with a small example. We generated data along a surface, which is shown in Figure 1. Note that the color of the patches is matched to the height of the surface. While this is really a 3-D structure, we will reduce it to 2-D using ISOMAP. The points are shown in a 2-D scatterplot in Figure 2, where we can see that the neighborhood relationships are mostly preserved.

The ISOMAP method was implemented in MATLAB by the creators and is available for download at http://isomap.stanford.edu/. The usual input to the ISOMAP function is the interpoint distance matrix, along with a value for k (number of nearest neighbors) or an epsilon neighborhood. The output from the function consists of a MATLAB structure with two fields (the coordinates in lower dimensional space for several values of d and an index of the embedded points), a vector of residuals, and a matrix for the neighborhood graph. We note that different lower-dimensional embeddings are obtained when the value of k or epsilon is changed.

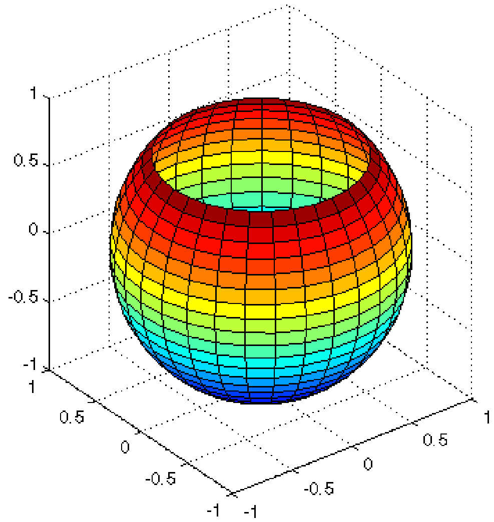


Figure 1. Points were randomly generated along this surface. The color is mapped to the height of the surface.

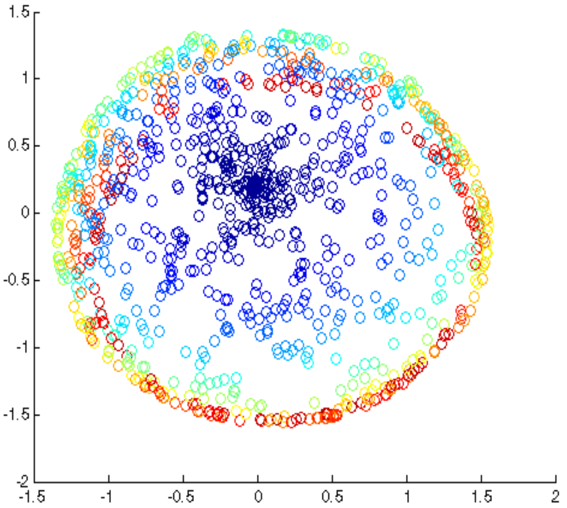


Figure 2. Here is the 2-D embedding from ISOMAP. The color of the points is the same as in Figure 1. Note that for the most part, points that are close together on the surface are near each other in the 2-D plot.

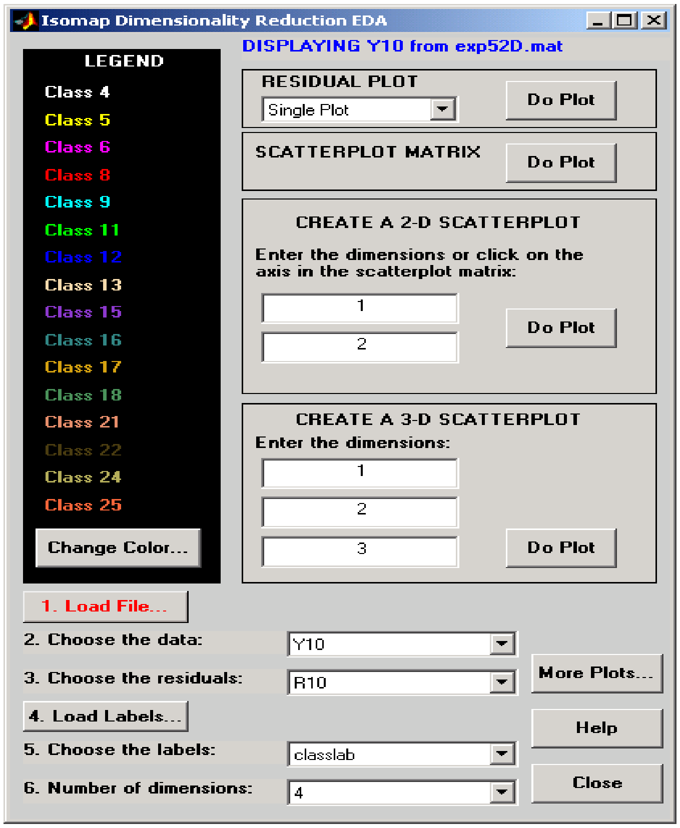
3. Graphical User Interface

As stated before, the purpose of this GUI tool is to provide a way of rapidly assessing the results of ISOMAP using graphical exploratory data analysis techniques [Wegman & Carr, 1993]. Of course, these methods are applicable to more areas than the one described here. In the case of ISOMAP, the analyst needs to answer such questions as: What is a good value for *k* and *d*? Are groups readily visible in the *d*-dimensional space?

The GUI tool has the following capabilities:

- Load ISOMAP output interactively
- Includes default colors for classes (if known)
- Change color for classes
- Various residual plots to determine *d*
- Scatterplots and scatterplot matrices
- Parallel coordinate plots
- Andrews curves
- Permutation tours

The GUI is activated by typing in `isomapeda` at the command line. A screen shot of the GUI is given in Figure 3, where a data set has already been loaded. The steps to load the data are given sequentially in the interface, and it assumes that all of the output from ISOMAP is saved in the same `.mat` file.



DISPLAYING Y10 from exp52D.mat

Class 4
Class 5
Class 6
Class 8
Class 9
Class 11
Class 12
Class 13
Class 15
Class 16
Class 17
Class 18
Class 21
Class 22
Class 24
Class 25

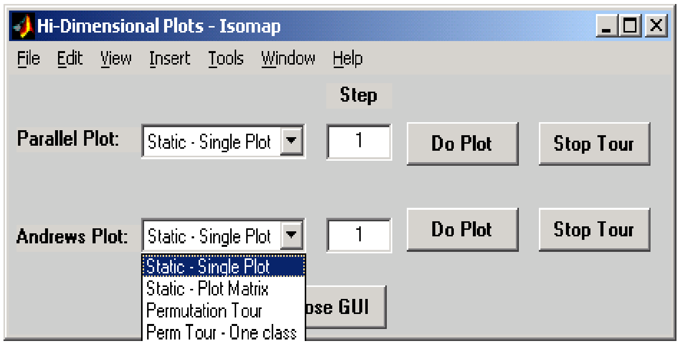
RESIDUAL PLOT: Single Plot

Enter the dimensions or click on the axis in the scatterplot matrix:
1
2

Enter the dimensions:
1
2
3

2. Choose the data: Y10
3. Choose the residuals: R10
5. Choose the labels: classlab
6. Number of dimensions: 4

Figure 3. Screen capture of the GUI tool after data has been loaded.



Step
Parallel Plot: Static - Single Plot 1
Andrews Plot: Static - Single Plot 1

Figure 4. Screen capture of the interface to the high-dimensional plots.

In this initial GUI interface, one can obtain the residual plots, a scatterplot matrix, a 2-D scatterplot and a 3-D scatterplot. A separate GUI interface is provided to construct the higher-dimensional plots such as Andrews curves and parallel coordinate plots. We now provide several examples of these plots using some data from Martinez [2002], recalling that in this case we are looking for groups or clusters in the lower dimensional space that correspond to topics. We see in Figure 3 that we have 16 topics or classes, each with a different color.

First, to help us determine the best value for d, we can make use of a scree-like plot [Jackson,1991]. This plots the residual variance as a function of the ISOMAP dimensionality, d. To determine the 'best' value for d, we look for an elbow in the curve. In this example, $d = 3$ looks like a good value. Other residual plots are available through the pop-up menu. The residual plot for our example is shown in Figure 5.

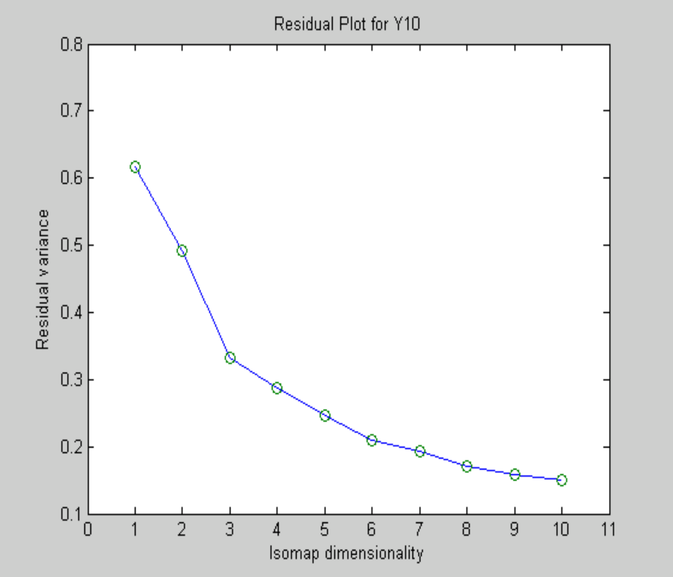


Figure 5. Residual plot for the ISOMAP embeddings.

We provide the capability of constructing 2-D and 3-D scatterplots. The 2-D scatterplot is shown in Figure 6, and the 3-D scatterplot is given in Figure 7. These plots can be rotated using the MATLAB rotation toolbar button, if better views are needed for discovering structure or groups. We can also construct a scatterplot matrix, which is shown in Figure 8. One of the options in MATLAB for this type of plot shows histograms of the individual dimensions along the diagonal elements of the plot matrix. Thus, a histogram for d_1 is shown in the upper left corner, d_2 is in the middle and d_3 is in the lower right. The scatterplot matrix provided with this GUI tool has an additional capability. The user can click on one of the scatterplots, and the corresponding plot will appear in its own figure window. This is useful when the user wants to see greater detail.

We now move on to the higher-dimensional plots that can be constructed with this interface. As stated before, these include parallel coordinate plots and Andrews curves. For more information on these types of plots, please see Wegman and Carr [1993].

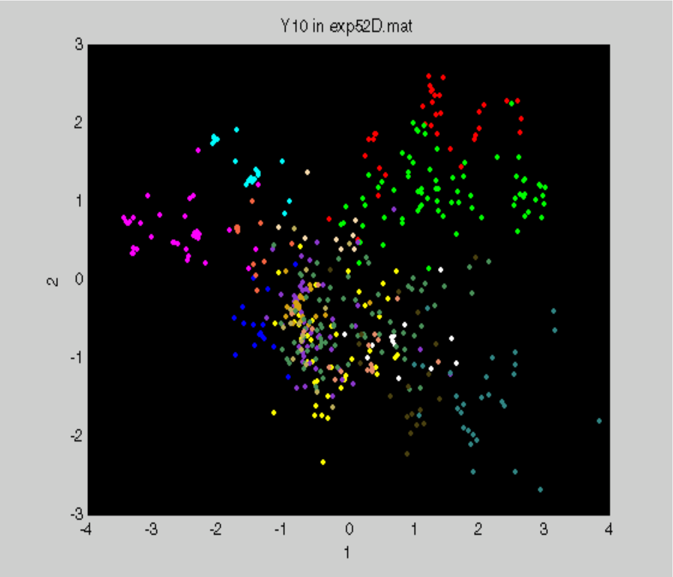


Figure 6. 2-D scatterplot.

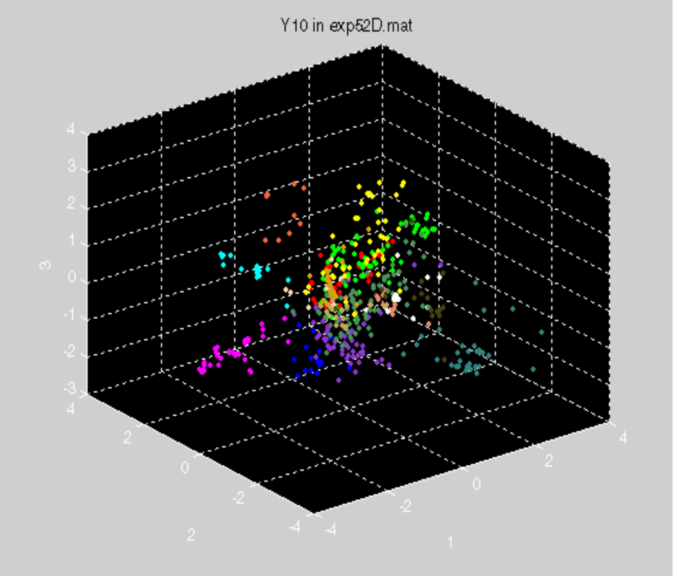


Figure 7. 3-D scatterplot.

A parallel coordinates plot showing all 16 color-coded topics is shown in Figure 9. We do not get a lot of information from this plot, because of overplottting. We have a similar situation with the Andrews curves plot shown in Figure 10. To help alleviate the situation and to facilitate exploring the results of ISOMAP, we also provide a plot matrix, where each plot shows the observations for one of the topics. The plots can either be parallel coordinates or Andrews curves.

An example of the parallel coordinate plot matrix is shown in Figure 11 at the end of the paper. Several things can be noted in this plot. First, we see that several of the topics (4, 5, 6, 8, 9) look distinctive and different from others. On the other hand, we see that some of the topics look very similar (21 and 22) or are rather incoherent pattern (15 and 18). So, we might conclude that some groups or topics are separable from others, but some might be difficult to distinguish. A plot matrix of Andrews curves is given in Figure 12 at the end of the paper. We see the same type of situation here. Some groups look like cohesive and separable topics, while others do not.

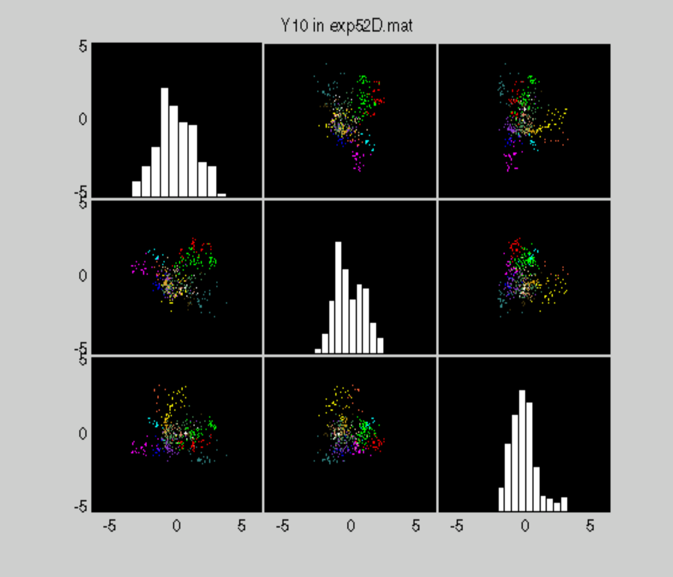


Figure 8. Scatterplot matrix for 3 dimensions.

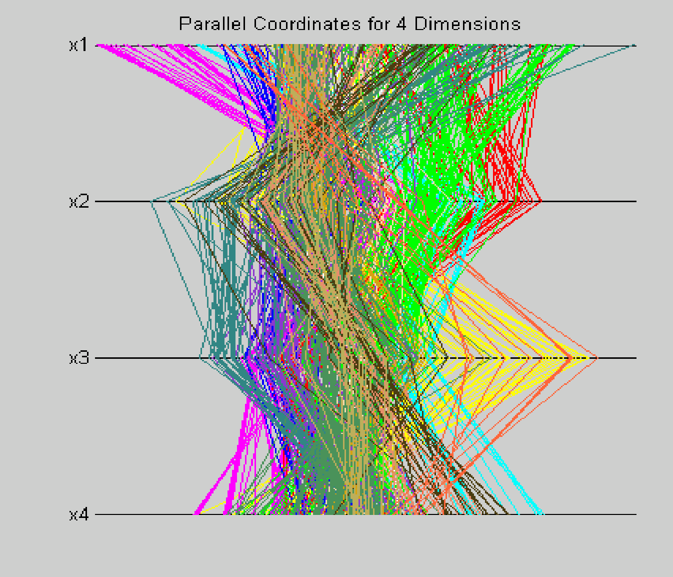


Figure 9. Parallel coordinate plot showing all 16 topics.

Andrews curves and parallel coordinate plots are sensitive to the ordering of the variables. In the case of Andrews curves, the initial variables have more effect on the shape of the curves, and parallel coordinates show pairwise relationships only. To alleviate this situation, we provide a permutation tour. These tours run through all permutations of the variables and replots them after each permutation. This is not the smart permutation tour of Wegman [1990], where the minimum number of permutations needed to cover all pairs of variables is outlined (along with a procedure for obtaining them).

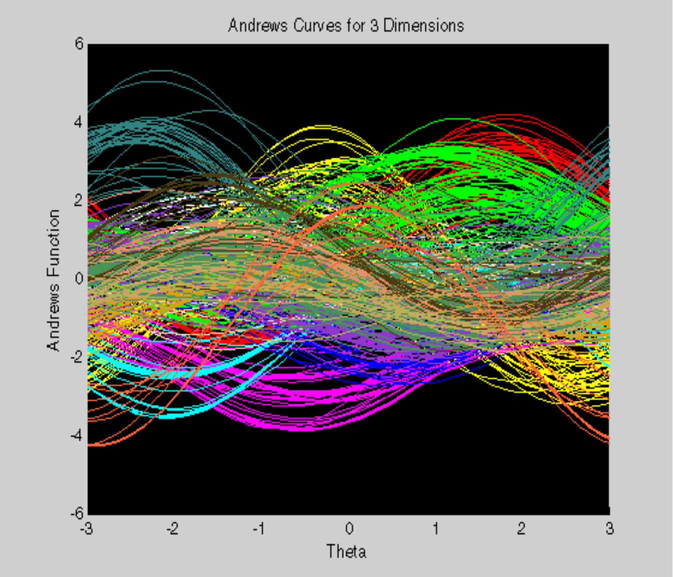


Figure 10. Andrews curves showing all 16 topics.

4. Summary

The GUI tool presented in this paper is a work in progress. We would like to implement linking and brushing [Wegman and Carr, 1993] for those applications where the true class labels are not known. We also have plans to implement the grand tour, both 2-D and *d*-dimensional [Asimov, 1985; Wegman, 1991].

The ideas and visualization capabilities included in the GUI tool are not unique to this application or to the output from ISOMAP. The tool is available for download at the Carnegie Mellon STATLIB website: http://lib.stat.cmu.edu/.

References

Asimov, D. 1985. "The grand tour: a tool for viewing multidimensional data," *SIAM Journal of Scientific and Statistical Computing*, 6, pp. 128-143.

Cox, T. F. & M. A. Cox. 2001. *Multidimensional Scaling*, Chapman & Hall.

Donoho, D. L. & C. Grimes. 2003. "Hessian eigenmaps: Locally linear embedding techniques for high-dimensional data," *Proceedings of the National Academy of Science*, 100, pp. 5591-5596.

Jackson, J. E. 1991. *A User's Guide to Principal Components*, John Wiley and Sons, New York.

Martinez, A. R. 2002. *A Framework for the Representation of Semantics*, Ph.D. Dissertation, George Mason University.

Roweis, S. T. & L. K. Saul. 2000. "Nonlinear dimensionality reduction by locally linear embedding," *Science*, 290, pp. 2323-2326.

Tenenbaum, J. B., V. deSilva & J. C. Langford. 2000. "A global geometric framework for nonlinear dimensionality reduction," *Science*, 290, pp. 2319 - 2323.

Wegman, E. J. 1990. "Hyperdimensional data analysis using parallel coordinates," *Journal of the American Statistical Association*, 85, pp. 664-675.

Wegman, E. J. 1991. "The grand tour in *k*-dimensions," *Computing Science and Statistics*, pp. 127-136.

Wegman, E. J. and D. B. Carr. 1993. "Statistical graphics and visualization," with D. B. Carr, in *Handbook of Statistics 9: Computational Statistics*, (Rao, C. R., ed.), Amsterdam: North Holland, pp. 857-958.

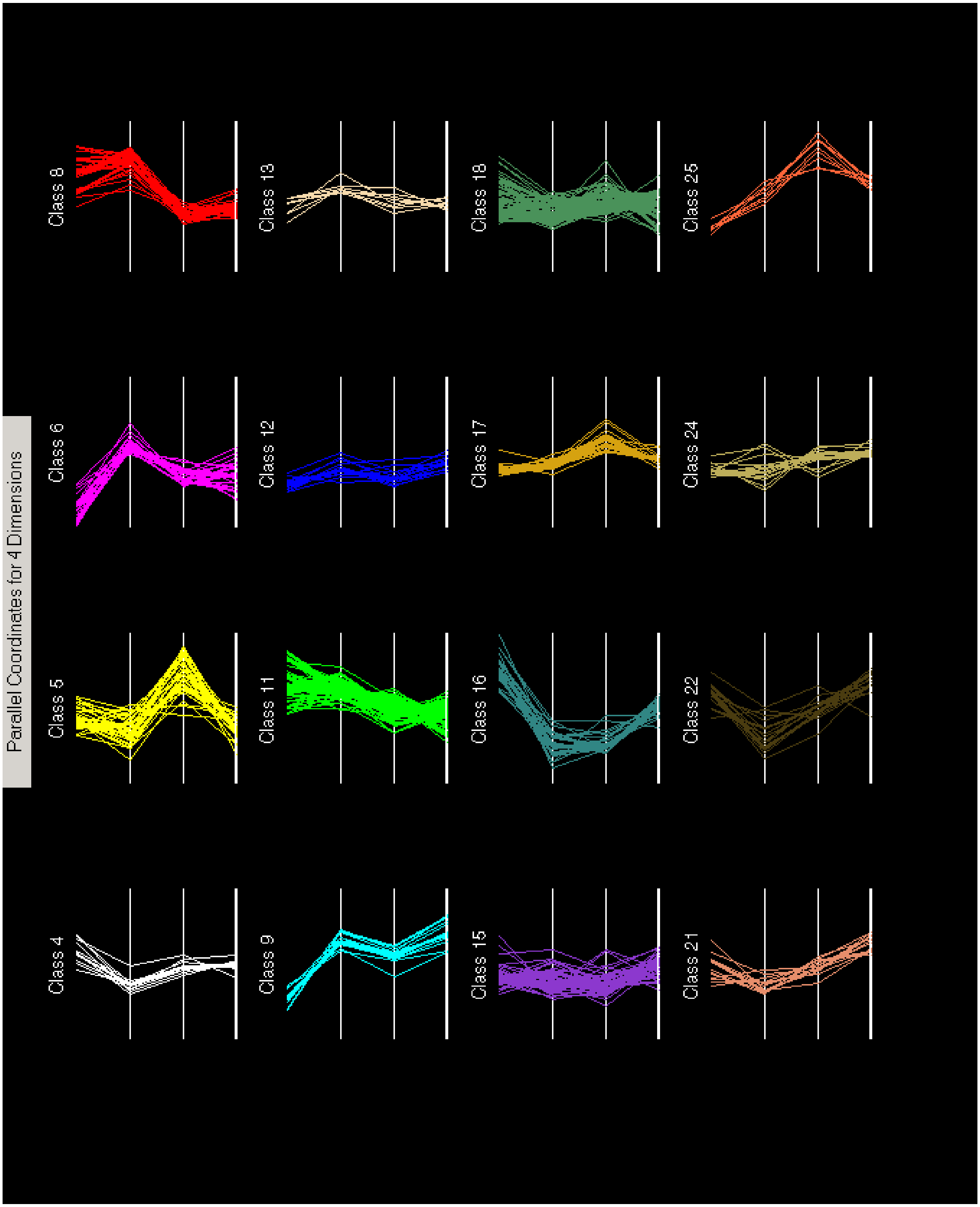


Figure 11. Plot matrix showing parallel coordinate plots for each topic.

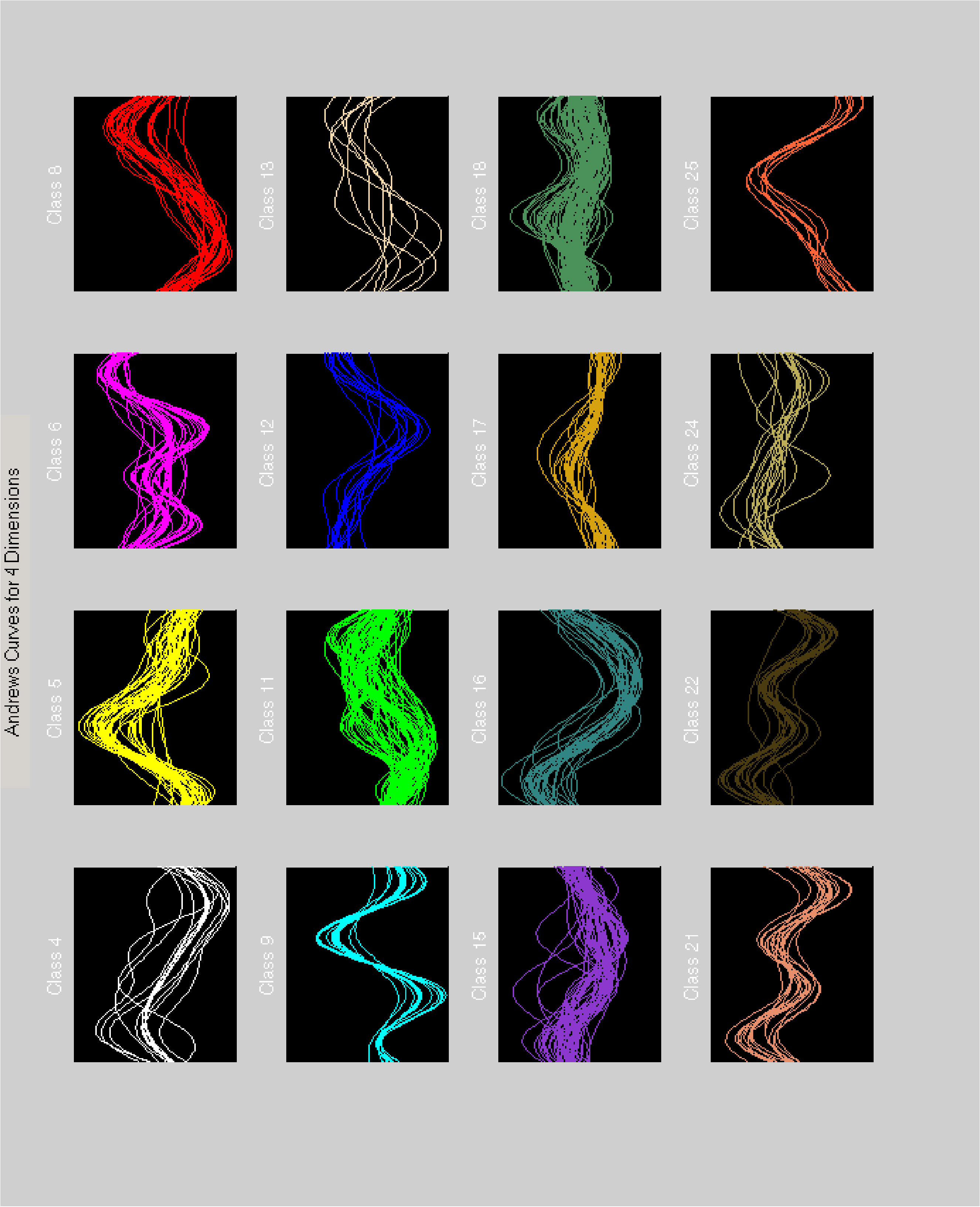


Figure 12. Plot matrix showing Andrews curves for each topic.