AN EXTENSION OF GRAND TOUR METHODS BASED ON ANDREWS CURVES

César García-Osorio
Departamento de Ingeniería Civil
Universidad de Burgos,
Burgos, Spain
email: cgosorio@ubu.es

Colin Fyfe
Applied Computational Intelligence Research Unit
University of Paisley,
Scotland, United Kingdom
email: colin.fyfe@paisley.ac.uk

ABSTRACT:
We review two methods of using Andrews curves to perform one and two dimensional grand tours. Then, we show our own method of using a variant of Andrews' curves to obtain a dynamic sequence of three dimensional projections of a data set and, therefore, a three dimensional grand tour. We illustrate this method using a real dataset. Finally, we propose a way to combine in the same representation the information obtained from the raw data, and from the principal component projection of the data, obtaining a more robust exploratory data analysis tool. The proposed idea can be extended to combine other views of the data.

KEY WORDS:
Visual Data Mining, Data Treatment and Visualization, Exploratory Data Analysis.

1 Introduction

The grand tour [2, 4] is a multivariate visualization method that consists in looking at the data from all points of view by presenting a continuous sequence of low dimensional projections. We can think about the grand tour as a generalisation of rotations in high-dimensional space. The dynamic nature of the grand tour give data analysts an extra dimensionality perception that helps them in the process of finding interesting structure in the data. From that point of view, the grand tour shares a common objective with exploratory pursuit techniques. In both cases the human ability for visual pattern recognition is exploited.

In [19], Wegman and Shen explain how to use a variant of Andrews' curves [1] to obtain a two dimensional grand tour (actually a pseudo grand tour, not a real grand tour since the vector is neither a unit vector nor does it exhaust all possible orientations of a one-dimensional vector; in the following we will not make such a distinction). We have proposed another extension [8] that lets us construct a three dimensional version of the grand tour. Here we illustrate that extension using a real dataset, and we show how we can combine, in the same grand tour, the information obtained from the raw data and the principal component projection of the raw data. As noted by Embrechts and Herberg [5], the original Andrews' curves are influenced by the order and scale of the variables; the method proposed here can be used also to combine two grand tours obtained with different order or scale of the variables of the same dataset.

2 Andrews Curves

Andrews [1] described his curves in 1972, early on in the computing era; it is an interesting observation that he thought it necessary to counsel "an output device with relatively high precision ... is required". Current standard PC hardware and software are quite sufficient for the purpose. The method is another way to attempt to visualise and hence to find structure in high dimensional data. Each data point $\mathbf{x} = (x_1, x_2, ..., x_n)$ defines a function

$$f_{\mathbf{x}}(t) = x_1/\sqrt{2} + x_2 \sin(t) + x_3 \cos(t) + x_4 \sin(2t) + x_5 \cos(2t) + ...$$

and this function is then plotted for $-\pi < t < \pi$. Thus each data point may be viewed as a line between $-\pi$ and π. The interesting thing is that this function representation preserves distances; that is close points will appear as close curves and distant points as distant curves. If there is structure in the data, it may be visible in the Andrews' Curves of the data at particular values of t. The curves themselves can be considered as the projections of each point into the vector

$$\mathbf{w} = \big(1/\sqrt{2}, \sin(t), \cos(t), \sin(2t), \cos(2t), ...\big)$$

Wegman and Shen [19] discuss the benefits of using a slightly different projection, namely that onto

$$\mathbf{w}_1 = \sqrt{2/n}\big(\sin(\lambda_1 t), \cos(\lambda_1 t), ..., \sin(\lambda_{\frac{n}{2}} t), \cos(\lambda_{\frac{n}{2}} t)\big)$$
$$\mathbf{w}_2 = \sqrt{2/n}\big(\cos(\lambda_1 t), -\sin(\lambda_1 t), ..., \cos(\lambda_{\frac{n}{2}} t), -\sin(\lambda_{\frac{n}{2}} t)\big)$$

with the λ_j linearly independent over the rationals; now the curves are not periodic and it makes sense to draw them for t outside the interval $(-\pi, \pi)$. Clearly, $(\mathbf{w}_1, \mathbf{w}_2)$ form a set of 2 orthonormal basis vectors. If we define $y_1 = \mathbf{w}_1^T \mathbf{x}$, $y_2 = \mathbf{w}_2^T \mathbf{x}$ then we have a two dimensional display on which to project $\mathbf{x}$ so that we can look for structure by

eye. Visually from this projection, we can identify clusters of points which are nearby and whose trajectories as we change t (i.e. as we move along the Andrews' Curves) keep close together. When we use these curves in this way we obtain a two dimensional grand tour [2] of the data (to be more precise we obtain what Wegman and Shen call a pseudo grand tour).

We have proposed in [8] an extension of this idea that makes use of three different orthogonal vectors so that now $\mathbf{w}_i = f(t,s)$ such as

$$\begin{aligned} y_1 = \mathbf{w}_1^T\mathbf{x} &\propto x_1\cos(\lambda_1 t)\cos(\mu_1 s) + \\ &\quad x_2\cos(\lambda_1 t)\sin(\mu_1 s) + x_3\sin(\lambda_1 t) + \ldots \\ y_2 = \mathbf{w}_2^T\mathbf{x} &\propto x_1\sin(\lambda_1 t)\cos(\mu_1 s) + \\ &\quad x_2\sin(\lambda_1 t)\sin(\mu_1 s) - x_3\cos(\lambda_1 t) + \ldots \\ y_3 = \mathbf{w}_3^T\mathbf{x} &\propto x_1\sin(\mu_1 s) - x_2\cos(\mu_1 s) + x_3 * 0 + \ldots \end{aligned}$$

where we have the implicit requirement that the number of terms in each expansion is a multiple of 3 rather than 2 as previously. Note that these equations really give three different groups of surfaces in 3D space but this gives a diagrammatic representation which is very difficult to understand. Thus we prefer to change t and s independently and view the movement of the groups of points through 3D space. We call the curves obtained when the value of t is fixed, *S-slices*, each corresponding to a particular slice of the surface with a specific t value. Similarly, we call *T-slices* the curves obtained when we fix the value of s. Figure 1 illustrates this for one group of surfaces. An alternative is to let $t = s$ and view the equations as a curve moving in 3D space. One interpretation of this is that the first component represents the (unit) tangent vector to the curve, the second the (unit) normal vector and the third the binormal vector so that these three vectors represent a natural local basis for that space.

The Andrews' curves have been utilized in fields as different as neurology [12], sociology [16], biology [13] and semiconductor manufacturing [15, 14]. Some of they uses include the detection of period and outliers in time series [6] or the visualization of learning in artificial neural networks [7]. Khattree and Naik [11] have suggested their utilization in robust design and in correspondence analysis.

2.1 Some properties

Some of the properties of our extension are:

- **Mean preservation.** The function corresponding to the mean of a set of n multivariate observations, is the pointwise mean of the functions corresponding to those observations:

$$f_{\bar{\mathbf{x}}}(t,s) = \frac{1}{n}\sum_{i=1}^{n} f_{\mathbf{x}_i}(t,s)$$

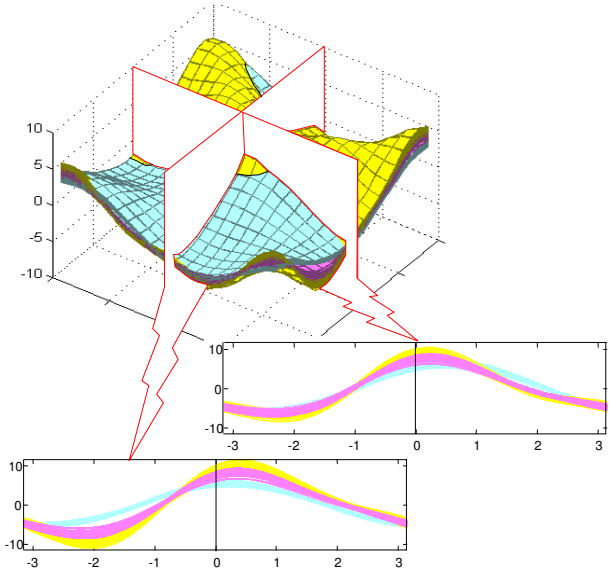

Figure 1. One of the surfaces and its T and S slices.

- **Distance preservation.** The distance between two functions, defined as the volume between the two surfaces defined by the functions:

$$||f_{\mathbf{x}}(t,s) - f_{\mathbf{y}}(t,s)|| = \int_{-\pi}^{\pi}\int_{-\pi}^{\pi}\big(f_{\mathbf{x}}(t,s) - f_{\mathbf{y}}(t,s)\big)^2 dt\, ds$$

  is proportional to the Euclidean distance between the corresponding points[1]:

$$||f_{\mathbf{x}}(t,s) - f_{\mathbf{y}}(t,s)|| = \pi^2||\mathbf{x} - \mathbf{y}||^2$$

- **Linear relationships.** If a point $\mathbf{y}$ lies on a line joining $\mathbf{x}$ and $\mathbf{z}$, then for all values of t and s, $f_{\mathbf{y}}(t,s)$ is between $f_{\mathbf{x}}(t,s)$ and $f_{\mathbf{z}}(t,s)$.

3 Using the surfaces

The use of our surfaces provides us with two different methods to interact with the dataset. One is the usual grand tour that the user can direct by changing the s or t values. The other one in the direct inspection of the slices of the surfaces. In the slices we can see groups of curves. If the groups of curves are bound close together that is a sign of clustering of the corresponding points. To confirm such cluster we can use the mechanism of brushing [3] to highlight the curves and check the closeness in the rest of the slice where we have made the selection or check that in others slices. As the curves are linked[2] with the points in the grand tour, we can also use the grand tour to check if all the selected curves/points belong to the same cluster.

[1] This property only holds for the first two groups of surfaces.

[2] *Linking*: A mechanisms for relating information in one plot to the information in another.

The data set used to illustrate the use of our surfaces is from a scientific study of various forms of algae some of which have been manually identified. Each sample is recorded as an 18 dimensional vector representing the magnitudes of various pigments. Some algae have been identified as belonging to specific classes which are numbered 1 to 9 (72 samples). Others are as yet unclassified and these are labelled 0 (46 samples). We have pre-processed the algae data set by centering the samples, i.e. subtracting the mean of all samples from each sample.

This data set has the advantage that we can check if algae which have been identified as belonging to one group do, in fact, remain together throughout multiple values of the parameters (see below) while simultaneously investigating whether the algae labelled 0 are possibly members of existing groups or come from quite distinct groups of algae. The use of the surfaces lets us identify almost all the clusters. Also it was possible to discover relations between some of the labelled clusters, new members of the existing clusters that were originally unlabelled, and the discovery of a new cluster within the unlabelled observations. We give two examples of the clustering process.

3.1 Example one

We can notice that around the value $t = -0.7$ in the first T-slice for $s = -0.28897$ (see Figure 2), there exists a possible cluster. After highlighting the curves and analyzing their evolution over different parts of our surfaces, we conclude that is possible to make a subdivision of this cluster into two distinct clusters. By checking the existing classification of the data reveals that this cluster corresponds to the classes 7 and 9. The proximity of the clusters suggests some kind of relation between the two groups of algae. These points and their associated curves are then removed from the display.

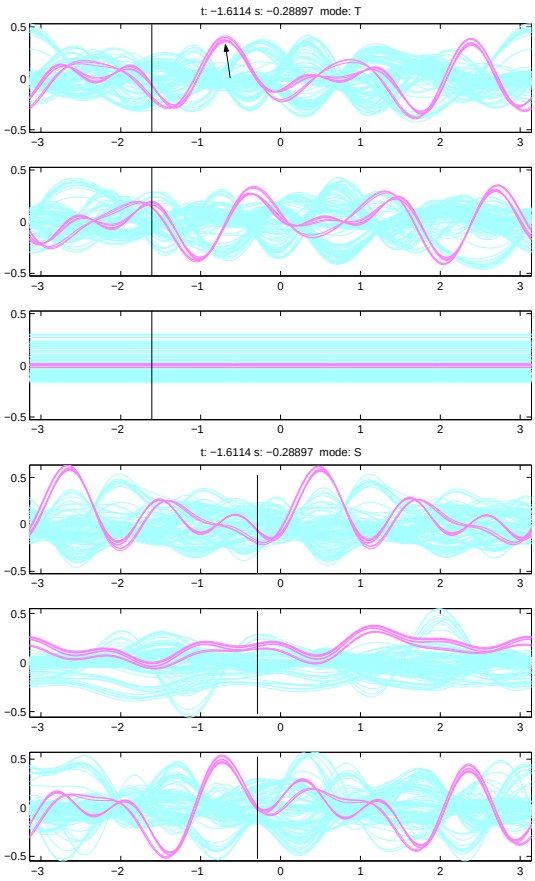

Figure 2. Upper: the T-slices at $s = -0.28897$ (some curves are highlighted). Lower: the S-slices at $t = -1.6114$ (with the same highlighted curves). The arrow in the top T-slice shows the place which motivated the cluster identification. After brushing those curves we can see that the clustering remains in the other slices. Other places could have been used to brush this cluster, for example around the values $s = -2.5$ or $s = 0.5$ in the top S-slice.

3.2 Example two

In the third S-slice (that does not depend on the value of t), around the value $s = -2.5$ one can clearly identify another candidate cluster and indeed, we find that all the selected points belong to class 4. After investigating other slices it was decided to include in the cluster an additional point. As one can see in Figure 3, this inclusion is quite reasonable since the shape of the curve is very similar to the other curves in the cluster. This is an interesting finding, since this curve corresponds to a point that is not classified in the original data set (i.e. is labelled 0). The grand tour in which the points have been highlighted reinforces the conclusion that we have identified a previously unclassified data point and have been able to classify it as class 4 with a reasonable degree of confidence.

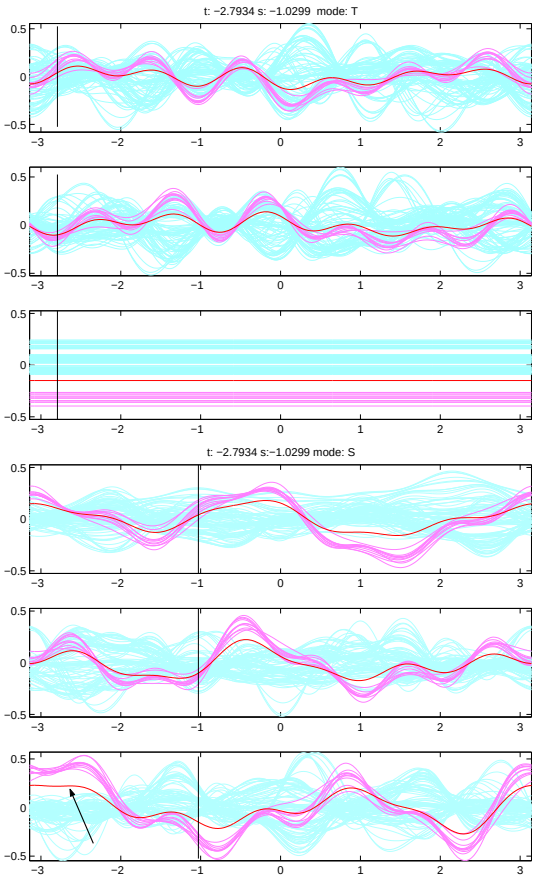

Figure 3. The initial highlighted curves of cluster four, plus one additional curve with a similar behaviour. Upper: the T-slices at $s = -1.0299$. Lower: the S-slices at $t = -2.7934$ (the arrow indicates the place which motivates the initial selection). The point associated with the additional curve was unlabelled in the data set. Here we see that, although being an outlier within that cluster, it has been possible its identification as an additional cluster member.

4 Combining Grand Tours

For each value of t and s our method gives three different values for each observation in the dataset. These values can be used as coordinates of a 3D projection of the dataset. As we change the t and s values we get a 3D grand tour. However these three values can be used in other ways. One suggestion is just to use two of these values as the coordinates of a 2D projection and the other one as a parameter to determine the size of the points used to represent the observations. This is, in some sense, also a 3D grand tour. Watching this grand tour we can also identify regularities in the data, this time as consequence of the proximity of the points and the similarity in their size.

But now the idea is to combine whatever of the previous grand tours with another one. Once the points are positioned in a 3D projection, or in a 2D projection (with the adequate size), we can use the values obtained from another group of surfaces to change other parameters of the representation. The obvious choice, is to use *three* values as the color components of the points. After a normalization step, we can get three values between 0 and 255 that we can use as the three values of a RGB colouring scheme. The points that are close according to the new grand tour will have similar colours. With this method we can save part of the work of brushing since the points and curves can be auto-coloured using the values obtained from the other grand tour. This use of our surfaces is in some way related to the image grand tour [20, 17], but instead of using multispectral images we use an arbitrary dataset.

The user now can change independently the t and s values of one grand tour to get the positions of the points, or can change the values of the grand tour used to get the color components, or can choose to change both simultaneously. That is, we can see an animation of the projection of the points moving in space, or we can see a static image of points changing theirs colours, or we can see both changes at the same time. This last method is the one used to obtain the grand tours whose snapshots appear in Figures 4 to 6. In Figure 6 we can see a snapshot of a grand tour we can call multidimensional, since we can see simultaneously the projection for all the t values as we change the s value. As a second grand tour we have used the principal component projection of the original dataset. An animation of this and others grand tours of the some dataset can be obtained in the web page http://cis.paisley.ac.uk/fyfe-ci0/cgo/VIIP2004/.

One of the major drawbacks of Andrews' curves is that, although they preserve the distance and the mean, their shapes depend on the order of the coefficients. Embrechts and Herzberg [5] suggest trying different orders for the coefficients. The combination proposed here can speed up the exploratory data analysis. We can combine in the same device the information that come from different ordinations or, as it happens in Figures 4 to 6, the information that come from a particular order of the raw data and from the principal component projection of the data.

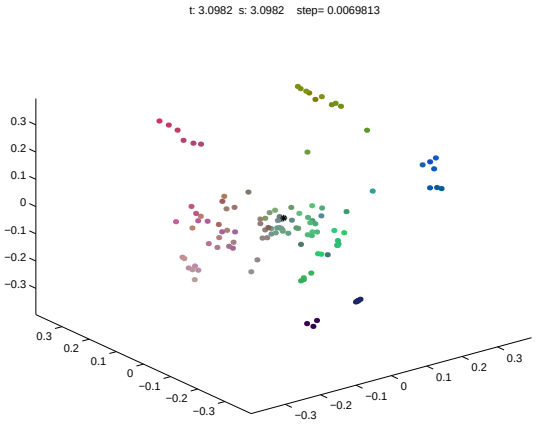

Figure 4. A snapshot of a 3D grand tour combined with the auto-coloring of the points (the full animation can be downloaded from http://cis.paisley.ac.uk/fyfe-ci0/cgo/VIIP2004/GT3D.avi). The position of the points is obtained from the surfaces constructed using the raw data, the color comes from the surfaces constructed using the principal component projection of the data.

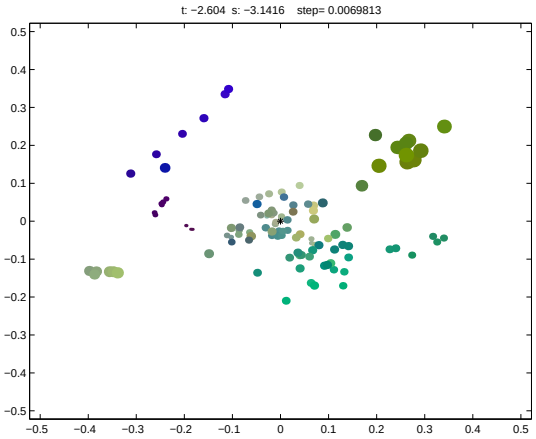

Figure 5. A snapshot of a 2D/size grand tour combined with the auto-coloring of the points (the full animation can be downloaded from http://cis.paisley.ac.uk/fyfe-ci0/cgo/VIIP2004/GT2D.avi). The position and size of the points come from the raw data, the color from the principal component projection of the data.

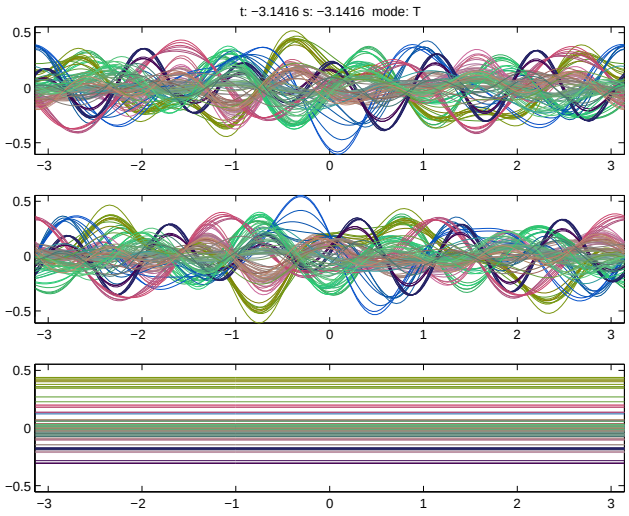

Figure 6. A snapshot of a n-D grand tour combined with the *auto-coloring* of the curves obtained from the principal components (the full animation can be downloaded from http://cis.paisley.ac.uk/fyfe-ci0/cgo/VIIP2004/GTnD.avi). This can save us having to brush some of the clusters.

5 Conclusion

We have shown that the proposed method can be used in two quite different but complementary ways:

1. The first allows us to walk along the curves using either of two parameters and find groups of curves which remain as a group for all possible values of the parameters.

2. The second allows us to use the human capacity for identifying structure in moving three dimensional displays, something for which our evolution in a three dimensional visual environment has created excellent pattern matchers.

The combined use of the new display with techniques such as brushing and linking, have enable us to identify most of the labelled clusters of a real dataset, but also allowed us to classify some of the unlabelled samples, and also to discover new clusters and identify sub-clusters in one labelled cluster.

The combination of the information that comes from two different views of the data reinforces the visualization method. The problems with high volume datasets (bad data-ink ratio and data density of a graphic [18]) can be addressed using an interface level with the curves, as for example Self Organizing Maps (see [10]). That is we can select a group of neurons in a SOM to draw only the points associated with the selected neurons. The use of our surfaces have let us "*visualize*" even the feature spaces defined by kernel methods [9]. All of the above suggest that these visualization tools are very promising.

References

[1] D. F. Andrews. Plots of High Dimensional Data. *Biometrics*, 28:125–136, 1972.

[2] D. Asimov. The Grand Tour: a Tool for Viewing Multidimensional Data. *SIAM Journal on Scientific and Statistical Computing*, 6(1):128–143, 1985.

[3] Richard A. Becker and William S. Cleveland. Brushing Scatterplots. *Technometrics*, 29:127–142, 1987.

[4] A. Buja and D. Asimov. Grand Tour Methods: an Outline. In *Computer Science and Statistics: Proceedings of the Seventeenth Symposium on the Interface*, pages 63–67, Amsterdam: North Holland, 1985. D. Allen.

[5] Paul Embrechts and Agnes M. Herzberg. Variations of Andrews' Plots. *International Statistical Review*, 59(2):175–194, 1991.

[6] Paul Embrechts, Agnes M. Herzberg, and Allen C.K. Ng. An Investigation of Andrews' Plots to Detect Period and Outliers in Time Series Data. *Communications in Statistics – Simulation and Computation*, 15(4):1027–1051, 1986.

[7] Marcus Gallagher. *Multi-layer Perceptron Error Surfaces: Visualization, Structure and Modelling*. PhD thesis, Dept. Computer Science and Electrical Engineering, University of Queensland, 2000.

[8] C. García-Osorio and C. Fyfe. An Extension of Andrews Curves for Data Analysis. In *Emergent Solutions for the Information and Knowledge Economy (X SIGEF Congress)*, 2003.

[9] C. García-Osorio and C. Fyfe. Visualisation in High Dimensional Feature Spaces. In *International Workshop on Practical Applications of Agents and Multiagents Systems (IWPAAMS 2003)*, 2003.

[10] C. García-Osorio, J. Maudes, and C. Fyfe. Using Andrews Curves for Clustering and Sub-clustering Self-Organizing Maps. In *European Symposium on Artificial Neural Networks (ESANN 2004)*, pages 477–482. d-side publications, 2004.

[11] Ravindra Khattree and Dayanand N. Naik. Andrews Plots for Multivariate Data: some new Suggestions and Applications. *Journal of Statistical Planning and Inference*, 100:411–425, 2002.

[12] James A. Kokiol and Werner Hacke. A Bivariate Version of Aandrews Plots. *IEEE Transactions on Biomedical Engineering*, 38(12):1271–1274, 1991.

[13] John Frank Murphy. *Methods for Collection and Processing of Gene Expression Data*. PhD thesis, California Institute of Technology, Pasadena, California 91125, 2003.

[14] E. A. Rietman and N. Layadi. A Study on $\mathbb{R}^m \rightarrow \mathbb{R}^1$ Maps: Application to a 0.16-μm Via Etch Process Endpoint. *IEEE Transactions on Semiconductor Manufacturing*, 13(4):457–468, 2000.

[15] E. A. Rietman, J. T. C. Lee, and N. Layadi. Dynamic Images of Plasma Processes: Use of Fourier Blobs for Endpoint Detection during Plasma Etching of Patterned Wafers. *Journal of Vacuum Science and Technology*, 16(3):1449–1453, 1998.

[16] Neil H. Spencer. Investigating Data with Andrews Plots. *Social Science Computer Review*, 21(2):244–249, 2003.

[17] J. Symanzik, E. J. Wegman, A. J. Braverman, and Q. Luo. New Applications of the Image Grand Tour. *Computing Science and Statistics*, 34:500–512, 2002.

[18] Edward R. Tufte. *The Visual Display of Quantitative Information*. Graphics Press, Cheshire, CT, 2 edition, 2001.

[19] E. J. Wegman and J. Shen. Three-dimensional Andrews Plots and the Grand Tour. *Computing Science and Statistics*, 25:284–288, 1993.

[20] Edward J. Wegman, Wendy L. Poston, and Jeffrey L. Solka. Image Grand Tour. Technical Report TR 150, The Center for Computational Statistics, ftp://www.galaxy.gmu.edu/pub/papers/Image_Tour.pdf, April 1998.