

Tendency Curves for Visual Clustering Assessment

Yingkang Hu

Georgia Southern University
Department of Mathematical Sciences
Statesboro, GA 30460-8093
USA
yhu@GeorgiaSouthern.edu

Richard J. Hathaway

Georgia Southern University
Department of Mathematical Sciences
Statesboro, GA 30460-8093
USA
rhathaway@GeorgiaSouthern.edu

Abstract: We improve the visual assessment of tendency (VAT) technique, which, developed by J.C. Bezdek, R.J. Hathaway and J.M. Huband, uses a visual approach to find the number of clusters in data. Instead of using square gray level images of dissimilarity matrices as in VAT, we further process the matrices and produce the *tendency curves*. Possible cluster structure will be shown as peak-valley patterns on the curves, which can be caught not only by human eyes but also by the computer. Our numerical experiments showed that the computer can catch cluster structures from the tendency curves even in cases where the visual outputs of VAT are virtually useless.

Key-Words: Clustering, similarity measures, data visualization, clustering tendency

1 Introduction

In clustering one partitions a set of objects $O = \{o_1, o_2, \dots, o_n\}$ into c self-similar subsets (clusters) based on available data and some well-defined measure of similarity. But before using a clustering method one has to decide whether there are meaningful clusters, and if so, how many are there. This is because all clustering algorithms will find any number (up to n) of clusters, even if no meaningful clusters exist. The process of choosing the number of clusters is called the *assessing of clustering tendency*. We refer the reader to Tukey [1] and Cleveland [2] for visual approaches in various data analysis problems, and to Jain and Dubes [3] and Everitt [4] for formal (statistics-based) and informal techniques for cluster tendency assessment. Recently the research on the *visual assessment of tendency* (VAT) technique has been quite active; see the original VAT paper by Bezdek and Hathaway [5], also see Bezdek, Hathaway and Huband [6], Hathaway, Bezdek and Huband [7], and Huband, Bezdek and Hathaway [8, 9].

The VAT algorithms apply on relational data, in which each *pair* of objects in O is represented by a relationship. Most of the time, the relationship between o_i and o_j is given by their dissimilarity R_{ij} (a distance or some other measure). These n^2 data items form a symmetric matrix $R = [R_{ij}]_{n \times n}$. If the data is given as a set $X = \{x_1, x_2, \dots, x_n\} \subset \mathbb{R}^s$, called *object data*,

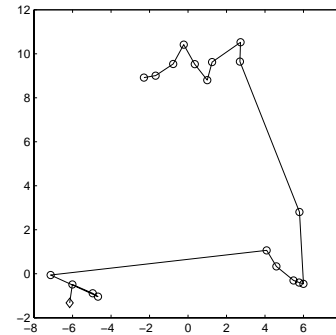


Figure 1: A data set X ordered by VAT

then R_{ij} can be computed as the distance between x_i and x_j measured by some norm or metric in \mathbb{R}^s . In this paper if the data is given as object data X , we will use as R_{ij} the square root of the Euclidean norm of $x_i - x_j$, that is, $R_{ij} = \sqrt{\|x_i - x_j\|_2}$. The VAT algorithms reorder (through indexing) the points so that points that are close to one another in the feature space will generally have similar indices. Their numeric output is an *ordered dissimilarity matrix* (ODM). We will still use the letter R for the ODM. It will not cause confusion since this is the only information on the data we are going to use. The ODM satisfies

$$0 \leq R_{ij} \leq 1, \quad R_{ij} = R_{ji} \quad \text{and} \quad R_{ii} = 0.$$

The largest element of R is 1 because the algorithms scale the elements of R .

The ODM is displayed as *ordered dissimilarity*

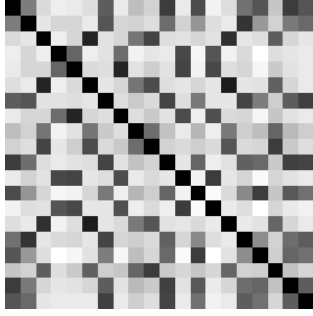
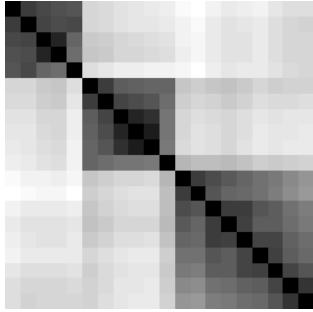

 Figure 2: Dissimilarity image *before* reordering X


Figure 3: ODI using the order in Fig 1.

image (ODI), which is the visual output of VAT. In ODI the gray level of pixel (i, j) is proportional to the value of R_{ij} : pure black if $R_{ij} = 0$ and pure white if $R_{ij} = 1$. The idea of VAT is shown in Fig.1–3. Fig.1 shows a scatterplot of a data set $X \subset \mathbb{R}^2$ of 20 points containing three well-defined clusters. Its original order is random, as in most applications. The dissimilarity image in Fig.2 contains no useful visual information about the cluster structure in X . The broken line in Fig.1 shows the new order of the data set X , with the diamond in the lower left corner representing the first point in the ordered data set. Fig.3 gives the corresponding ODI. Now the three clusters are represented by the three well-formed black blocks.

The VAT algorithms are certainly useful, but there is room for improvements. It seems to us that our eyes are not very sensitive to structures in gray level images. One example is given in Fig.4. There are three clusters in the data as we will show later. The clusters are not well separated, and the ODI from VAT does not reveal any sign of the existence of the structure.

The approach of this paper is to focus on changes in dissimilarities in the ODM, the numeric output of VAT that underlies its visual out-

put ODI. The results will be displayed as curves, which we call the *tendency curves*. The borders of clusters in the ODM (or blocks in the ODI) are shown as certain patterns in peaks and valleys on the tendency curves. The patterns can be caught not only by human eyes but also by the computer. It seems that the computer is more sensitive to these patterns on the curves than human eyes are to them or to the gray level patterns in the ODI. For example, the computer caught the three clusters in the data set that produced the virtually useless ODI in Fig.4.

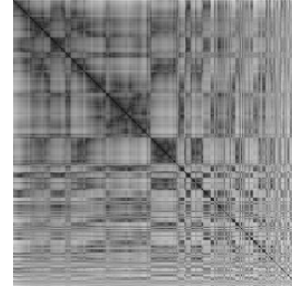


Figure 4: How many clusters are in this ODI?

2 Tendency Curves

Our approach is to catch possible diagonal blocks in the ordered dissimilarity matrix R by using various averages of distances, which are stored as vectors and displayed as curves. Let n be the number of points in the data, we define

$$m = 0.05n, \quad M = 5m, \quad w = 3m. \quad (1)$$

We restrict ourselves to the w -subdiagonal band (excluding the diagonal) of R , as shown in Fig.5. Let $\ell_i = \max(1, i - w)$, then the i -th “row-average” is defined by

$$r_1 = 0, \quad r_i = \frac{1}{i - \ell_i} \sum_{j=\ell_i}^{i-1} R_{ij}, \quad 2 \leq i \leq n. \quad (2)$$

In another word, each r_i is the average of the elements of row i in the w -band. The i -th m -row moving average is defined as the average of all elements in up to m rows above row i , inclusive, that fall in the w -band. This corresponds to the region between the two horizontal line segments in Fig.5. We also define the M -row moving average in almost the identical way except with m replaced by M . They will be referred to as the r -curve, the m -curve and the M -curve, respectively.

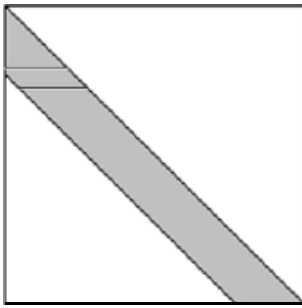


Figure 5: Sub-diagonal band of the ODM

The idea of the r -curve is simple. Just imagine a horizontal line, representing the current row in the program, moving downward in an ODI such as the one in Fig.3. When it moves out of one diagonal black block and into another, the r -curve should first show a peak because the numbers to the left of diagonal element R_{ii} will suddenly increase. It should drop back down rather quickly when the line moves well into the next black block. Therefore the border of two blocks should be represented as a peak on the r -curve if the clusters are well separated.

When the situation is less than ideal, there will be noise, which may destroy possible patterns on the r -curve. That is how the m -curve comes in, which often reveals the pattern beneath the noise. Since the VAT algorithms tend to order outliers near the end, so the m -curve tends to move up in the long run, which makes it hard for the program to identify peaks. That is why we introduce the M -curve, which shows long term trends of the r -curve. The difference of the m - and M -curves, which we call the d -curve, retains the shape of the m -curve but is more horizontal, basically lying on the horizontal axis. Furthermore, the M -curve changes more slowly than the m -curve, thus when moving from one block into another in the ODM, it will tend to be lower than the m -curve. As the result, the d -curve will show a valley, most likely below the horizontal axis, after a peak. It is the peak-valley, or high-low, patterns that signal the existence of cluster structures. This will become clear in our examples in the section that follows.

3 Numerical Examples

We give one group of examples in \mathbb{R}^2 so that we can use their scatterplots to show how well/poorly the clusters are separated. We also give the visual outputs (ODIs) of VAT for comparison. These sets are generated by choosing $\alpha = 8, 4, 3, 2, 1$ and 0 in the following settings: 2000 points (observa-

tions) are generated in three groups from multivariate normal distribution having mean vectors $\mu_1 = (0, \alpha\sqrt{6}/2)$, $\mu_2 = (-\alpha\sqrt{2}/2, 0)$ and $\mu_3 = (\alpha\sqrt{2}/2, 0)$. The probabilities for a point to fall into each of the three groups are 0.35, 0.4 and 0.25, respectively. The covariance matrices for all three groups are I_2 . Note that μ_1, μ_2 and μ_3 form an equilateral triangle of side length $\alpha\sqrt{2}$.

There are three well separated clusters for the case $\alpha = 8$, the ODI from VAT (not shown) has three black blocks on the diagonal with sharp borders. Fig.6 shows the data set for $\alpha = 4$, in which the three clusters are reasonably well separated. Our r -curve in Fig.8 (the one with “noise”) has two vertical rises and the m -curve (the solid curve going through the r -curve where it is relatively flat) has three peaks. Two of the three peaks are followed by valleys, corresponding to the two block borders in the ODI in Fig.7. The M -curve, the smoother, dash-dotted curve, is only interesting in its relative position with respect to the m -curve. That is, it is only useful in generating the d -curve, the difference of these two curves. The d -curve looks almost identical to the m -curve, also having three peaks and two valleys. The major difference is that it is in the lower part of the figure, around the horizontal axis. Note that

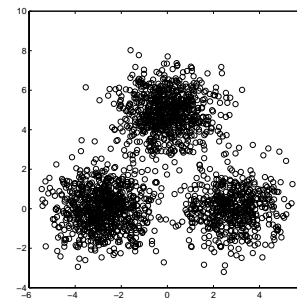


Figure 6: Three normally distributed clusters in \mathbb{R}^2 with $\alpha = 4$

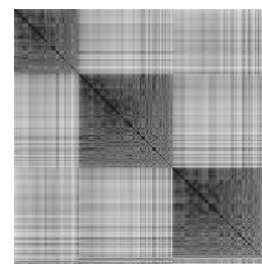


Figure 7: ODI from VAT, $\alpha = 4$

the wild oscillations near the end of the r -curve

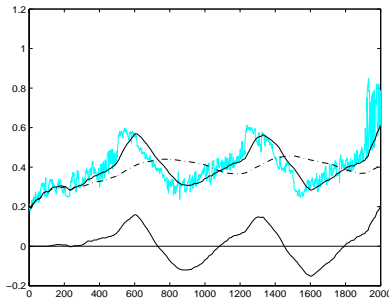


Figure 8: Tendency curves for $\alpha = 4$

bring up all other three curves, forming the third peak. This corresponds to the small region in the lower-right corner of the ODI, where there lacks pattern. Note also that no valley follows from the third rise or peak. This is understandable because a valley appears when the curve index (the horizontal variable of the graphs) runs into a cluster, shown as a block in ODI. Now we know what we should look for: vertical rises or peaks followed by valleys, or high-low patterns, on all the tendency curves maybe except the M -curve.

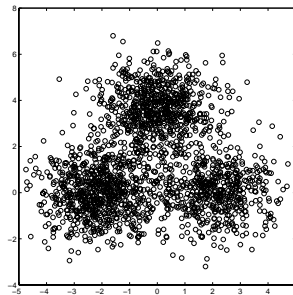


Figure 9: Three normally distributed clusters in \mathbb{R}^2 with $\alpha = 3$

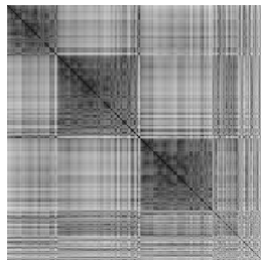


Figure 10: ODI from VAT, $\alpha = 3$

The case $\alpha = 3$ is given in Fig.9–11. One can still easily make out the three clusters in the scatterplot, but it is harder to tell to which cluster many points in the middle belong. It is ex-

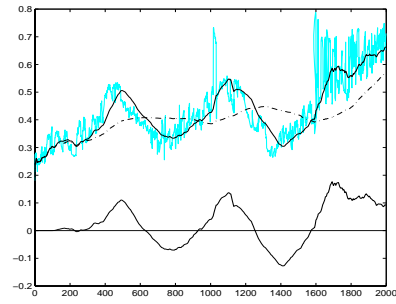


Figure 11: Tendency curves for $\alpha = 3$

pected that every visual method will have difficulties with them, as evidenced by the lower right corner of the ODI, and the oscillations on the last one fifth of the r -curve. The oscillations bring up the r - and m -curves, but not the d -curve. The d -curve remains almost the same as that in the previous case, except the third peak becomes larger and decreases moderately without forming a valley. The two high-low patterns on the m - and d -curves show the existence of three clusters. As we have said earlier that it is a valley on the m -curve and, especially, the d -curve that signals the beginning of a new cluster.

We hope by now the reader can see the purpose of the d -curve. Both the m - and M -curves in Fig.11 go up with wild oscillations, but the d -curve always stays low, lying near the horizontal axis. Unlike the other three curves, its values never get too high or too low. This enables us to detect the beginnings of new blocks in an ODM by catching the high-lows on the d -curve. When the d -curve hits a ceiling, set as 0.04, and then a floor, set as 0, the program reports one new cluster. The ceiling and floor values are satisfied by all cases in our numerical experiments where the clusters are reasonably, sometimes only barely, separated. If we lower the ceiling and raise the floor, we would be able to catch some of the less separated clusters we know we have missed, but it would also increase the chance of “catching” false clusters. We do not like the idea of tuning parameters to particular examples. We will stick to the same ceiling and floor values throughout this paper. In fact, *we do not recommend changing the suggested values of the parameters in our program*, that is, the values for the ceiling and floor set here, and those for m , M and w given in (1).

The situation in the case $\alpha = 2$, shown in Fig.12–14, really deteriorates. One can barely make out the three clusters in Fig.12 that are supposed to be there; the ODI in Fig.13 is a mess. In fact, this is the same ODI as the one in Fig.4, put

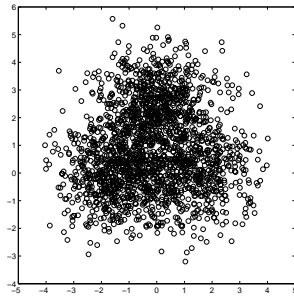


Figure 12: Three normally distributed clusters in \mathbb{R}^2 with $\alpha = 2$

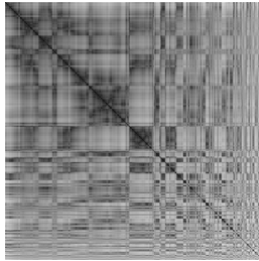


Figure 13: ODI from VAT, $\alpha = 2$

here again for side-by-side comparison with the scatterplot and the tendency curves. The tendency curves in Fig.14, however, pick up cluster structure from the ODM. The d -curve has several high-lows, with two of them large enough to hit both the ceiling and floor, whose peaks are near 600 and 1000 marks on the horizontal axis, respectively. This example clearly shows that our tendency curves generated from the ODM are more sensitive than the raw block structure in the graphical display (ODI) of the same ODM. The largest advantage of the tendency curves is probably the quantization which enables the computer, not only human eyes, to catch possible patterns.

When α goes down to zero, the cluster structure disappears. The scatterplots for $\alpha = 0$

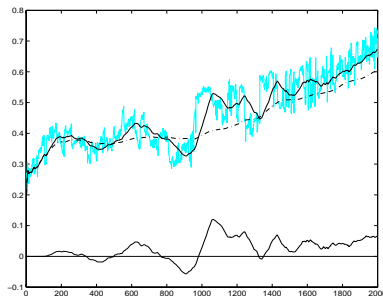


Figure 14: Tendency curves for $\alpha = 2$

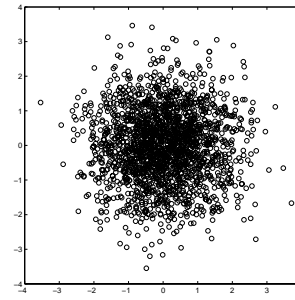


Figure 15: Three normally distributed clusters in \mathbb{R}^2 with $\alpha = 0$ (combining into one)

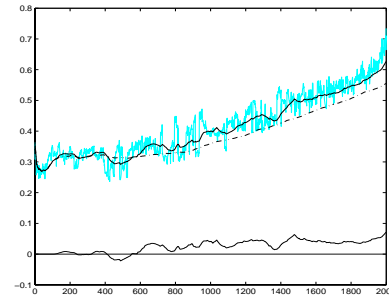


Figure 16: Tendency curves for $\alpha = 0$

(Fig.15) and $\alpha = 1$ (not shown) are almost identical, showing a single cluster in the center. The tendency curves for both cases (Fig.16 and 17) have no high-lows large enough to hit the ceiling then the floor, which is the way they should be. Note that while all other three curves go up when moving to the right, the d -curve, the difference of the m - and M -curves, stays horizontal, which is, again, the reason we introduced it.

We also tested our program on many other data sets, including small data sets containing 120 points in \mathbb{R}^4 . It worked equally well. We also tested two examples in Figures 12 and 13 of Bezdek and Hathaway [5], where the points are regularly arranged, on a rectangular grid, and along a pair of concentric circles, respectively. Be-

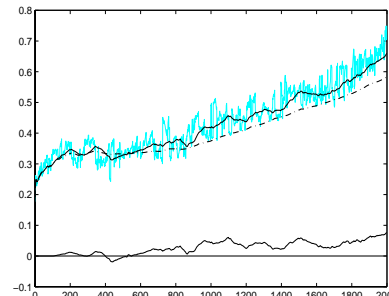


Figure 17: Tendency curves for $\alpha = 1$

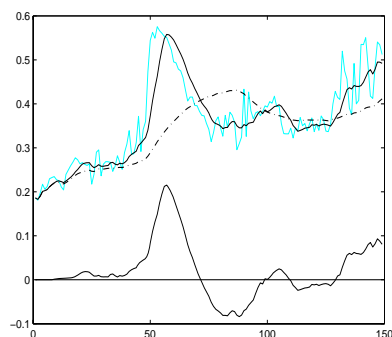


Figure 18: Tendency curves for the Iris data

cause we could only have speculated from ODI images produced from the sets before applying our program, it was to our great, and pleasant, surprise that the program worked seamlessly on them, accurately reported the number of clusters that exist. What we want to emphasize is that *we did all this without ever having to modify any suggested parameter values!* These tests will be reported in a forthcoming paper.

It is almost a sacred ritual that everybody tries the Iris data in a paper on clustering, so we conclude ours by trying our program on it. The tendency curves are given in Fig.18, and the computer caught the large high-low on the left and ignored the small one on the right, and correctly reported two clusters.

References:

- [1] J.W. Tukey, Exploratory Data Analysis. Reading, MA: Addison-Wesley, 1977.
- [2] W.S. Cleveland, Visualizing Data. Summit, NJ: Hobart Press, 1993.
- [3] A.K. Jain and R.C. Dubes, Algorithms for Clustering Data. Englewood Cliffs, NJ: Prentice-Hall, 1988.
- [4] B.S. Everitt, Graphical Techniques for Multivariate Data. New York, NY: North Holland, 1978.
- [5] J.C. Bezdek and R.J. Hathaway, VAT: A tool for visual assessment of (cluster) tendency. Proc. IJCNN 2002. IEEE Press, Piscataway, NJ, 2002, pp.2225-2230.
- [6] J.C. Bezdek, R.J. Hathaway and J.M. Huband, Visual Assessment of Clustering Tendency for Rectangular Dissimilarity Matrices, IEEE Trans. on Fuzzy Systems, **15** (2007) 890-903
- [7] R. J. Hathaway, J. C. Bezdek and J. M. Huband, Scalable visual assessment of cluster

tendency for large data sets, Pattern Recognition, **39** (2006) 1315-1324.

- [8] J. M. Huband, J.C. Bezdek and R.J. Hathaway, Revised visual assessment of (cluster) tendency (reVAT). Proc. North American Fuzzy Information Processing Society (NAFIPS), IEEE, Banff, Canada, 2004, pp.101-104.
- [9] J. M. Huband, J.C. Bezdek and R.J. Hathaway, bigVAT: Visual assessment of cluster tendency for large data set. PATTERN RECOGNITION, **38** (2005) 1875-1886.