Statistica Sinica **12**(2002), 7-29

GENERALIZED ASSOCIATION PLOTS: INFORMATION VISUALIZATION VIA ITERATIVELY GENERATED CORRELATION MATRICES

Chun-Houh Chen

Academia Sinica, Taipei

Abstract: Given a p-dimensional proximity matrix $D_{p\times p}$, a sequence of correlation matrices, $\mathbf{R} = (R^{(1)}, R^{(2)}, \ldots)$, is iteratively formed from it. Here $R^{(1)}$ is the correlation matrix of the original proximity matrix D and $R^{(n)}$ is the correlation matrix of $R^{(n-1)}$, $n > 1$. This sequence was first introduced by McQuitty (1968), Breiger, Boorman and Arabie (1975) developed an algorithm, CONCOR, based on their rediscovery of its convergence. The sequence $\mathbf{R}$ often converges to a matrix $R^{(\infty)}$ whose elements are $+1$ or -1. This special pattern of $R^{(\infty)}$ partitions the p objects into two disjoint groups and so can be recursively applied to generate a divisive hierarchical clustering tree. While convergence is itself useful, we are more concerned with what happens before convergence. Prior to convergence, we note a rank reduction property with elliptical structure: when the rank of $R^{(n)}$ reaches two, the column vectors of $R^{(n)}$ fall on an ellipse in a two-dimensional subspace. The unique order of relative positions for the p points on the ellipse can be used to solve seriation problems such as the reordering of a Robinson matrix. A software package, Generalized Association Plots (GAP), is developed which utilizes computer graphics to retrieve important information hidden in the data or proximity matrices.

Key words and phrases: Data visualization, divisive clustering tree, latent structure, perfect symmetry, proximity matrices, seriation.

1. Introduction

Correlation matrices are among the most well-studied objects in statistics. But consider the following problem. For any p by p matrix D, define $\phi(D)$ to be the p by p matrix whose ijth entry equals the Pearson's correlation coefficient for the ith and the jth columns of D,

$$\frac{\sum_k (d_{ik} - \bar{d}_{i\cdot})(d_{jk} - \bar{d}_{j\cdot})}{\sqrt{\sum_k (d_{ik} - \bar{d}_{i\cdot})^2}\sqrt{\sum_k (d_{jk} - \bar{d}_{j\cdot})^2}},$$

where d_{ik} is the ikth entry of D and $\bar{d}_{i\cdot}$ denotes the mean $p^{-1}\sum_k d_{ik}$. What happens if we apply this correlation operator $\phi(\cdot)$ to a matrix D iteratively to

obtain the sequence, $R^{(1)} = \phi(D)$, and $R^{(n+1)} = \phi(R^{(n)})$, $n = 1, 2, \ldots$? Will the sequence converge? If so, to what does it converge? Surprisingly, such natural mathematical questions do not yet admit an easy solution in the literature. Our interest stems from the need for developing information visualization tools in analyzing data collected for studying the grouping structure among schizophrenic symptoms and patients (Lin, Chen, Hwu, Lin and Chen (1998)). In general, we can take D to be any proximity matrix; see Section 2.1 for a brief review. D is denoted by $R^{(0)}$ when D itself is a correlation matrix.

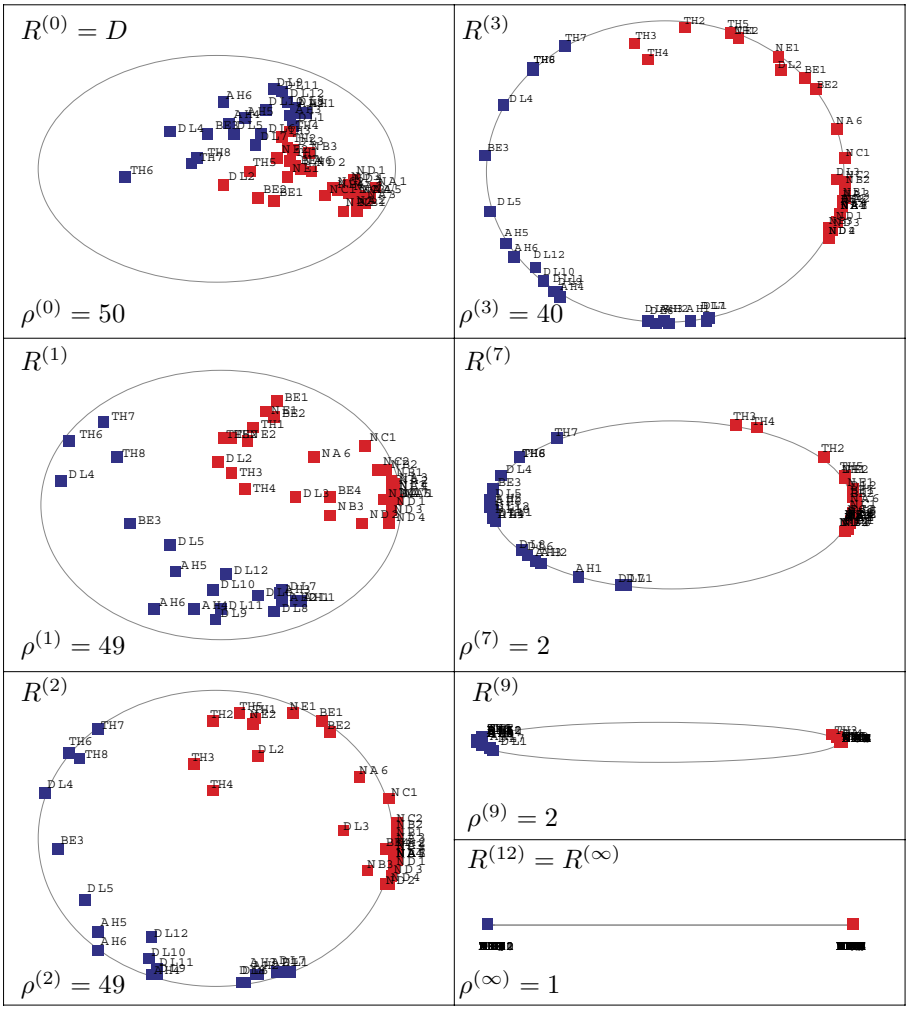

Figure 1. Plots for first two eigenvectors for selected correlation matrices in the converging sequence. ($\rho(n)$ is the rank of $R^{(n)}$).

One way of studying such a correlation matrix series is through visualization. To do so, for each matrix $R^{(n)}$, we project the p column vectors of $R^{(n)}$ onto the plane spanned by the first two eigenvectors of $R^{(n)}$. We did this for the sequence obtained by taking D to be the correlation matrix for the aforementioned schizophrenic data. Figure 1 gives the projections of $R^{(n)}$, for $n = 0, 1, 2, 3, 7, 9$, and 12. As we can see, clear elliptical clusters begin to form at step 3.

One of our goals in this paper is to provide a theoretical explanation for the formation of the elliptical clustering patterns which appear in all our real data and simulation studies. This is given in Sections 3 and 4. In Section 5, we further explore the clustering pattern discovered earlier, and use it as a new way to construct a seriation algorithm. For a brief review on the seriation problem, see Section 2.2. A software package, GAP (generalized association plots), utilizing these ideas is described in Section 6, and some concluding remarks are given in Section 7.

1.1. The psychosis disorder data

Our study is motivated by a data set from the Taiwan multidimensional psychopathological group research program (MPGRP) (Lin, Chen, Hwu, Lin and Chen, (1998)). The data set consisted of the Andreasen's positive and negative symptom scales (Andreasen (1983) and (1984)) of 95 first-time hospitalized psychosis disorder patients. Among the 95 patients, 69 patients were diagnosed as schizophrenic and 26 patients as bipolar disorder. The system of Andreasen's symptom scales include the Scale for Assessment of Positive Symptoms (SAPS) with 30 items, and the Scale for Assessment of Negative Symptoms (SANS) with 20 items. SAPS includes four subgroups: hallucinations (AH1-6), delusions (DL1-12), behavior (BE1-4) and thought disorder (TH1-8). SANS has five subgroups: expression (NA1-7), speech (NB1-4), hygiene (NC1-3), activity (ND1-4) and inattentiveness (NE1-2). The available data set has ninety-five subjects (patients) with fifty variables (symptoms). All the symptoms are recorded on a six-point scale (0-5). Complete SAPS and SANS tables are available on our web site.

Psychiatrists in the MPGRP are concerned about the grouping structure among the symptoms, the clustering structure of patients and the general behavior of patient-clusters on each symptom-group. They can be phrased as three multivariate analysis problems: (1) the linkage amongst n subject points in the p-dimensional space; (2) the linkage between p variable vectors in the n-dimensional space; (3) the interaction linkage between the sets of subjects and variables. Factor analysis and clustering methods are commonly applied to solve the first two problems but there is no general technique for studying the third problem.

2. Proximity Matrix Map and Seriation

2.1. The proximity matrix map

The concept of using points of different shading to represent proximity and raw data matrices is not new to researchers in the fields of taxonomy (Sneath and Sokal (1973)), seriation (Caraux (1984), Streng (1991), Minnotte and West (1998)), cluster analysis (Gale and Halperin (1984), Streng (1991)), statistical computing (Murdoch and Chow (1996)), multidimensional scaling (Chen and Chen (2000)), and gene expression (Wen et al. (1998), Lyer et al. (1999)). As an illustration, consider the Pearson correlation matrix of the fifty symptoms (Kendall's rank correlation and other possible association measurements produce similar results). This is the matrix D used in Figure 1. First a color spectrum (blue-red in Figure 2a) is selected. Then the proximity matrix is projected through the color spectrum to get a color proximity matrix map (Figure 2b). From Figure 2b, blocks of dark red points on the main diagonal can be easily located. Thus, thirty positive symptoms are divided into several small groups of symptoms; whereas, all the twenty negative symptoms form a more coherent cluster except symptoms NE1 and NE2 on the lower right corner. Murdoch and Chow (1996) cleverly used elliptical glyphs to represent correlations. However, their method is for displaying correlation matrices and not other types of proximities.

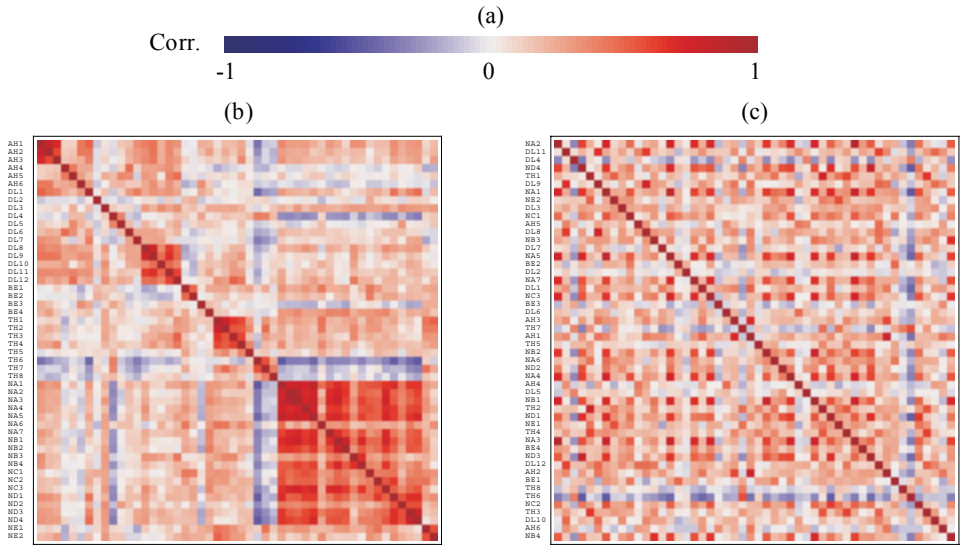

Figure 2. The colored correlation matrix maps for the fifty symptoms. (a) Blue-red color spectrum for correlation coefficients; (b) Correlation matrix map with the original SAPS/SANS order; (c) correlation matrix map with a random permutation.

2.2. The seriation problem

In Figure 2b grouping is transparent, but this is not always true for proximity matrix maps. When the fifty symptom's correlation map was constructed, the variables were already laid out in a prespecified order from Andreasen's symptom table. When the fifty symptoms are randomly permuted, as shown in Figure 2c, almost everything is lost: the structural patterns, the blocks, and the relationship between the groups. Note that permutation does not alter any numerical information in the original proximity matrix, but the corresponding map becomes useless. In order to recover the missing structure from Figure 2c, or even to get a better structure than Figure 2b, we need a seriation algorithm. Seriation is a data analytic tool for finding a permutation or ordering of a set of objects using a data matrix (symmetric or asymmetric). Hubert (1976) and Marcotorchino (1991) discuss seriation from the standpoints of problem setting, methodology and algorithms. Using Marcotorchino's (1991) notation, the initial matrix is denoted as T, with the set of objects and the set of variables denoted as I and J respectively. The basic principle of seriation is to find a reshaped matrix T' with a permutation of I, together with a permutation of J, to identify the embedded latent structure. When T is symmetric, we usually want T' to approximate a Robinson form (Robinson (1951)). A Robinson Matrix, $R = [r_{ij}]$, is a symmetric matrix such that $r_{ij} \leq r_{ik}$ if $j < k < i$ and $r_{ij} \geq r_{ik}$ if $i < j < k$. If rows and columns of a symmetric matrix T can be sorted such that it becomes a Robinson matrix, we call T *pre-Robinson*.

3. Properties Related to the Convergence Problem

Given a proximity matrix D, a sequence of sample Pearson correlation matrices is iteratively generated from D, $\boldsymbol{R} = (R^{(1)}, R^{(2)}, \ldots)$, where $R^{(n)} = \phi(R^{(n-1)})$, $n > 1$, and $R^{(1)} = \phi(D)$. Typically, this sequence of correlation matrices converges to a matrix $R^{(\infty)}$. Figure 3 illustrates the convergence well, using colored maps. The top-left map is that of Figure 2b. In fact it takes twelve iterations (only six of them are shown here) for this sequence of correlation matrices to converge to a limiting matrix $R^{(\infty)}$ in which all elements are plus or minus one. Our web page has complete set of maps for the converging sequence of correlation matrices.

3.1. The p-dimensional cube and cone

Correlation matrices can be visualized in an alternative way. The p column vectors of a matrix can be treated as p points in the cube $[-1, 1]^p$. A 3-dimensional example is shown in Figure 4a. Starting with the sample correlation matrix $\begin{bmatrix} 1 & 0.197 & 0.072 \\ 0.197 & 1 & -0.003 \\ 0.072 & -0.003 & 1 \end{bmatrix} = R^{(0)}$, it takes six iterations to converge to

$\begin{bmatrix} 1 & 1 & -1 \\ 1 & 1 & -1 \\ -1 & -1 & 1 \end{bmatrix} = R^{(\infty)}$. The three points of $R^{(n)}$ are shown as a_n, b_n, and c_n for $n = 1, \ldots, 6$. As can be seen, a_n and b_n move together toward $(1, 1, -1)$ while c_n goes toward $(-1, -1, 1)$.

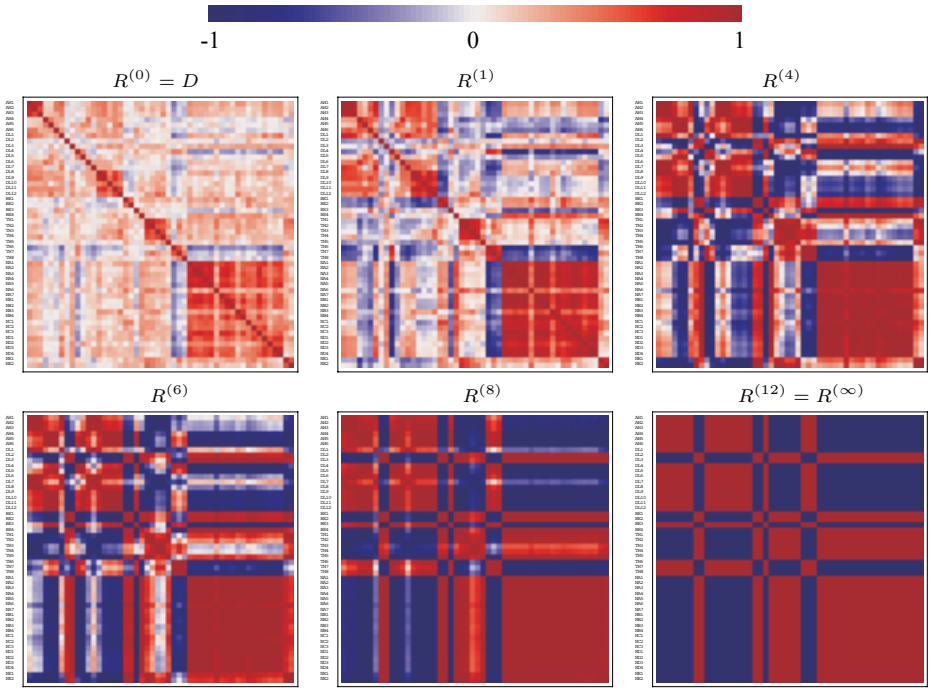

Figure 3. Maps for the converging sequence of correlation matrices at selected iterations (0,1,4,6,8,12) for the fifty symptoms.

A further observation shows that the points at each iteration are confined on three of the six surfaces of the cube. These surfaces form a 3-sided "cone" with the vertex at the intersection point, $(1, 1, 1)$, of the three planes. At early iterations (iterations 0 to 1 in this example), each point moves toward its own corner (the corner with all coordinates equal to -1 except for the point itself). At intermediate iterations, the centering and product steps in calculating the correlation coefficient force columns with similar pattern to attract each other and move toward the corner with simultaneous ones on these coordinates. Several groups may form at an intermediate stage. At the final iteration, only two groups survive and these two groups of points are at one of the $2^{p-1} - 1$ pairs of opposite corners with 1 and -1 on opposite coordinates. The converging paths for 20 of the 50 symptoms projected onto the 3-dimensional cube of columns NB1, DL9,

and TH3 are displayed in Figure 4b. Behavior similar to that in the simulation can be seen in the converging paths.

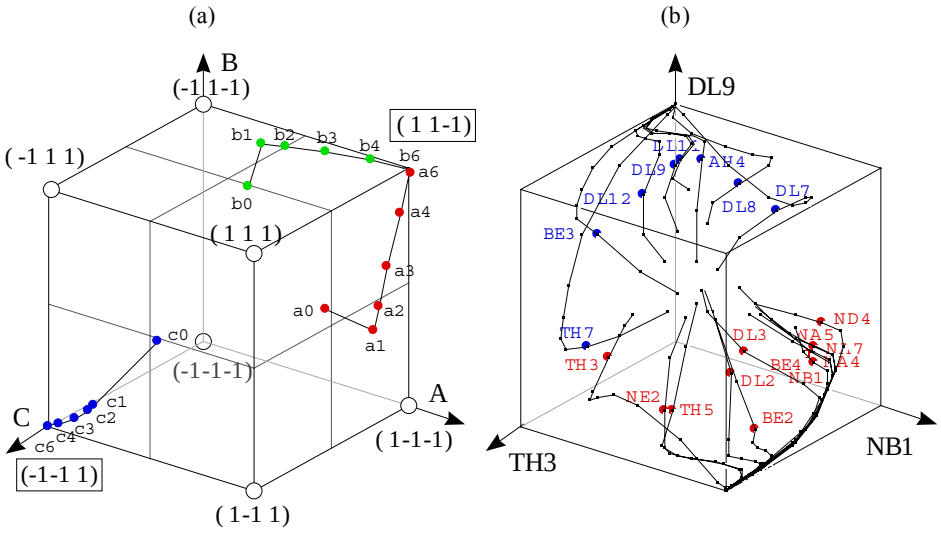

Figure 4. The Converging paths of the columns on the p-dimensional cube. (a) Simulation study with 3 columns on a 3-dimensional cube; (b) Converging paths for twenty of the fifty symptoms projected onto the 3-dimensional cube of (NB1, DL9, and TH3).

3.2. Will the sequence converge?

The iteratively formed correlation matrix sequence for the fifty-symptom example converges to the limiting matrix $R^{(\infty)}$ with positive and negative ones. Is convergence guaranteed? If so, does the limit matrix $R^{(\infty)}$ contain only positive and negative ones? The first answer seems to be Yes according to computer output for all examples we have studied, including extensive computer simulations of correlation matrices with various dimension and correlation structure. Unfortunately, we cannot prove it yet. What we know at this moment are some weaker facts.

It is easy to verify that the mapping ϕ defined in Section 1 is continuous at any correlation matrix R, except when R is degenerate (a correlation matrix R is non-degenerate if all diagonal entries are one and no column of R equals $1 = (1, \ldots, 1)^T$, and is degenerate otherwise). Since the set of all correlation matrices is compact and the mapping is differentiable, it can be speculated that some fixed point theorem can be used to prove convergence. This is non-trivial due to the existence of chaotic mapping.

In an unpublished article, Kruskal (1977) attempts to prove convergence but he needs complicated conditions on $R^{(0)}$. A related problem is to find the

stationary points of $\phi : \phi(R) = R$. It is easy to see that if R consists of $+1$ or -1, then R is a stationary point. Are they the only ones? The answer is no: Figure 5 in Section 4.2 gives all stationary points for $p = 3$, and 4.

3.3. The decreasing sequence of ranks

While convergence is important in itself, we are more concerned about what happens before convergence. We find an interesting rank reduction property with elliptical structure and later explore this for seriation and clustering purposes. At each iteration, groups become more coherent while individual columns gradually lose their identity. The high dimensional structure collapses to a lower dimensional one each time two or more columns move close enough to a common corner. We have the following observation.

Lemma 3.1. *The ranks of* $\{R^{(0)}, R^{(1)}, R^{(2)}, \ldots\}, \{\rho^{(0)}, \rho^{(1)}, \rho^{(2)}, \ldots\}$, *form a non-increasing sequence.*

Proof.

$$\begin{aligned}\rho^{(n+1)} &= \rho(\phi(R^{(n)}) = \rho(Cor(R^{(n)})) = \rho(Cov(R^{(n)})) \\ &= \rho\Big(\frac{1}{p}\Big(R^{(n)} - \frac{11^T}{p}R^{(n)}\Big)^T\Big(R^{(n)} - \frac{11^T}{p}R^{(n)}\Big)\Big) = \rho\Big(R^{(n)}\Big(I - \frac{11^T}{p}\Big)\Big).\end{aligned}$$

Since $\rho(AB) \leq \min[\rho(A), \rho(B)]$ and $\rho(AB) \geq \rho(A)+\rho(B)-p$ (Sylvester's Law of Nullity), we have $\rho^{(n+1)} \leq \min(p-1, \rho^{(n)})$ and $\rho^{(n+1)} \geq (p-1)+\rho^{(n)}-p = \rho^{(n)}-1$. This completes the proof.

Two facts emerge from the proof: if $R^{(0)}$ is of full rank ($\rho^{(0)} = p$), the rank of $R^{(1)}$ is $p-1$; the rank is reduced by at most 1 at each iteration.

For our numerical results, rank is calculated as the number of eigenvalues greater than $\varepsilon = e^{-13}$ so different ε's could result in slightly different rank lists. For the case study in Figure 1, it takes twelve iterations for the sequence to converge to $R^{(12)} = R^{(\infty)}$. The ranks are $\{50, 49, 49, 40, 20, 7, 3, 2, 2, 2, 2, 2, 1\}$. Theoretically this cannot occur since the reduction of rank is at most one at each iteration and only the first iteration is guaranteed to reduce the rank. It is possible that the ranks for the rest of the matrices in the sequence are 49, but the structure of the column space spanned by $R^{(n)}$ becomes flatter and flatter and eventually collapses to a lower dimensional structure. Besides, while the numerical rank decreases from p to 1, the sum of squares of the eigenvalues (equal to the sum of squares of all p^2 elements in the correlation matrix) increases to p^2 at $R^{(\infty)}$. That is, the variation becomes more and more concentrated on the leading eigenvectors. For example, the sums of squares of the eigenvalues for the particular sequence of matrices are (237.0, 514.6, 933.7, 1214.4, 1313.0, 1383.1, 1517.7, 1761.0, 2119.1, 2436.5, 2499.1, 2500.0, 2500.0), Figure 9. However, this

increasing trend may not be true at early iterations for a highly homogeneous proximity matrix D.

3.4. The elliptical structure theorem

In this section, we investigate the mechanism for the formation of elliptical structure as in Figure 1.

Theorem 3.1. *Given a full rank correlation matrix R, all p column (row) vectors of $R, R_i, i = 1, \ldots, p$, fall on the p-dimensional ellipsoid generated by the kernel of R^{-1}, the inverse of R.*

Proof. Observe that the right hand side of the equality $RR^{-1}R = R$ is a correlation matrix, hence all diagonal elements equal one. That is $\text{diag}(RR^{-1}R) = \text{diag}(R) = (1, 1, \ldots, 1)$, which leads to $(R_i)^T R^{-1}(R_i) = 1, i = 1, \ldots, p$, thus completing the proof.

In general only the original proximity matrix D is of full rank, all subsequent correlation matrices $\{R^{(n)}\}_{n=1}^{\infty}$ have ranks smaller than p. We can then substitute R^{-1} in Theorem 3.1 with the generalized inverse R^{-}. If the correlation matrix is not orthogonal, usually the case, we have a rotated version of Theorem 3.1.

Corollary 3.1. *Let the p-dimensional correlation matrix R have rank $k, k < p$. Consider the spectral decomposition $RQ = Q\Lambda$, where Λ is a k-dimensional diagonal matrix with non-zero eigenvalues $(\lambda_1, \lambda_2, \ldots, \lambda_k)$ on the diagonal (λ_is not necessarily distinct) so that Q contains the $p \times 1$ eigenvectors $Q_1, Q_2, \ldots, Q_k$. All the principal components of $R, (RQ)_i, i = 1, \ldots, p$, fall on the k-dimensional ellipsoid generated by the kernel of Λ^{-1}.*

Usually when one deals with the ellipsoid generated from the quadratic form of a positive definite matrix like the correlation matrix, it is Corollary 3.2 that is of interest and not Corollary 3.1.

Corollary 3.2. *With the same setup as in Corollary* 3.1*, all p rows of $Q, (Q^T)_i$, $i = 1, \ldots, p$, fall on the k-dimensional ellipsoid generated by the kernel of Λ.*

We have now shown that each correlation matrix in the sequence $\{R^{(n)}\}_{n=0}^{\infty}$ has an exact $\rho^{(n)}$-dimensional ellipsoid embedded in it. Each time the rank decreases, the ellipsoid collapses to a lower dimensional one.

4. The General Converging Patterns

With different types of structure embedded in the proximity matrix D there are various types of $R^{(\infty)}$ that occur, and two that are major: non-symmetry and symmetry.

4.1. The rank-one non-symmetry converged matrix

For practical statistical data analyses, only one kind of $R^{(\infty)}$ can occur. That is the rank-one correlation matrix with all elements equal to plus or minus one. Here the ellipsoid has dimension one and all p vectors fall on two points. The grouping pattern of the positive and negative ones in $R^{(\infty)}$ can be used to split the p variables (objects) into two groups. Such partition has some nice simulation results with the splitting criterion, $\min_{\Gamma(p_1,p_2)} \sum_{j=1}^{p} | \sum_{i=1}^{p_1}(d_{ij} - \bar{d_{j\cdot}}) - \sum_{i=1}^{p_2}(d_{ij} - \bar{d_{j\cdot}})|$, where $\Gamma(p_1, p_2)$ stands for all possible splitting of p objects into groups of p_1 and p_2.

Example 4.1. Five hundred sets of 20 bivariate uniform (0,1) observations are generated. The pair-wise 2-dimensional Euclidean distances for the 20 points are calculated as the D matrix. For each set, all 2^{20-1}=524,288 possible partitions are compared with the splitting correlation result and the frequencies of the number of partitions that performed better than the proposed method is calculated. The correlation split method finds the best partition among all the 524,288 possible partitions in about 60% (298/500) of the simulations. In more than 90% (456/500) of simulations, this correlation split stands at the first to sixth place among all the 524,288 possible combinations. The worst case is a 446-th order, which stands at 99.9149322 percentile. Section 5.1 shows how this splitting rule can be recursively applied to the proximity matrix to grow a divisive hierarchical clustering tree.

4.2. The symmetrical converging structure

The general form of $R^{(\infty)}$ in the case of symmetry is

$$\begin{bmatrix} 1 & -1/(p-1) & \cdots & -1/(p-1) \\ -1/(p-1) & 1 & \cdots & -1/(p-1) \\ \vdots & \vdots & \ddots & \vdots \\ -1/(p-1) & -1/(p-1) & \cdots & 1 \end{bmatrix}.$$

It need not be that all $C(p, 2)$ pairs of measurement are identical, but there still exist structures wherein the p points cannot be divided into two groups. Figure 5 summarizes all the possible symmetry and non-symmetry structures with their corresponding limits for $p = 3$ and $p = 4$. The number of possible limits is an increasing function of p.

In general there are only three types of columns in a converged pattern matrix: (i) columns with only plus and minus ones in a rank-1 matrix, patterns 3(3), 4(5), and 4(6) for example; (ii) columns with summation of elements equal to zero (this type of matrix can be further divided into two subtypes- one for

symmetry, patterns 3(1) and 4(1) for instance; the other for a circular matrix, patterns 3(1), 4(2) and part of 4(3)); (iii) columns for symmetry center(s) with zeroes to all other points, column B in pattern 3(2), column D in pattern 4(3), and columns B/C in pattern 4(4) for example. Section 5.4 looks at possible applications of symmetry patterns to study crystallographic structure.

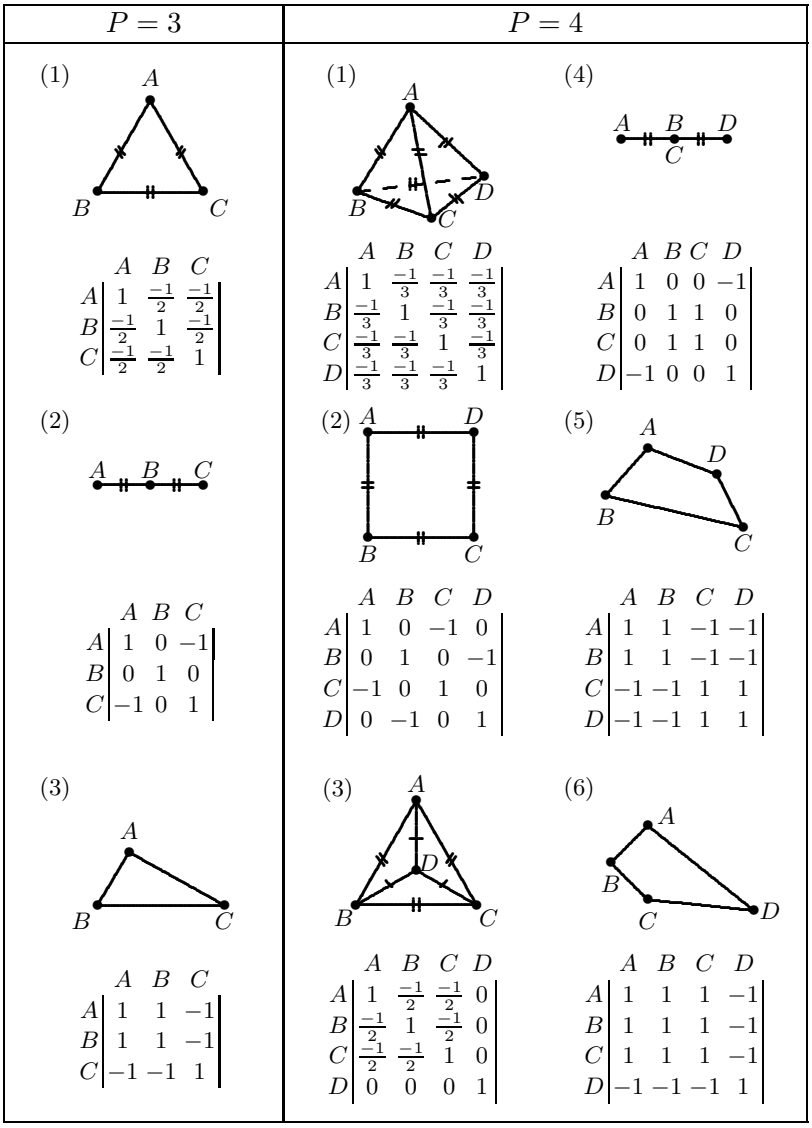

Figure 5. All possible symmetry and non-symmetry structures with the corresponding converged pattern matrices for $p = 3$ and 4. Segments with identical // or / signs have equal lengths while segments with // signs are longer than segments with / signs.

5. Applications

In previous sections, we described convergence properties for the sequence of iteratively formed correlation matrices. Various applications of these properties are discussed in this section.

5.1. The hierarchical divisive clustering tree with rank one splitting rule

Without perfect symmetry in the proximity matrix, one divides the p objects (variables or subjects) into two groups. We can recursively apply this correlation splitting rule to form a hierarchical divisive clustering tree. The clustering tree for the correlation proximity matrix of the fifty symptoms is illustrated in Figure 6a. This hierarchical divisive clustering tree with the correlation-splitting rule was the major topic of studies conducted by McQuitty (1968) and Breiger, Boorman, and Arabie (1975).

Figure 2c is reconstructed as Figure 6b, using the permutation of fifty symptoms from the order of terminal nodes of this clustering tree with a scheme to flip the two branches at each intermediate node. The permutation method is the first type of seriation developed here and is similar to the study by Gale and Halperin (1984).

Next, Figure 6b is compared with Figure 2b to see if Figure 6b has recovered the original structure embedded in Figure 2b from Figure 2c. It is seen that Figure 6b has even more structural pattern than does Figure 2b. There are five major groups along the main diagonal, the negative symptoms, the thought process symptoms, the hallucination symptoms, the delusion symptoms, and the mania symptoms.

5.2. The rank-two ellipse seriation technique

When the sequence reaches an iteration with rank two, the p objects fall on an ellipse and have unique relative positions on the ellipse. There are p possible cuts. The order on the two-dimensional ellipse can be combined with the one-dimensional split to find two orders with the cuts at the two gaps between the two converged groups. The elliptic seriation with the sorted correlation map is given in Figure 6c and d. The symptom order in Figure 6d is different from that in Figure 6b, but the major grouping patterns are identical.

Given a *pre-Robinson* matrix, the correlation matrix at the first iteration shows a perfect half-ellipse structure with all p vectors falling on half of the two-dimensional ellipse (see the web page for an example). It is possible to combine the rank-1 splitting rule and the rank-2 elliptical seriation to form a hybrid seriation for data sets with clustering structure. That is, one performs a separate rank-2 seriation on each split sub-matrix.

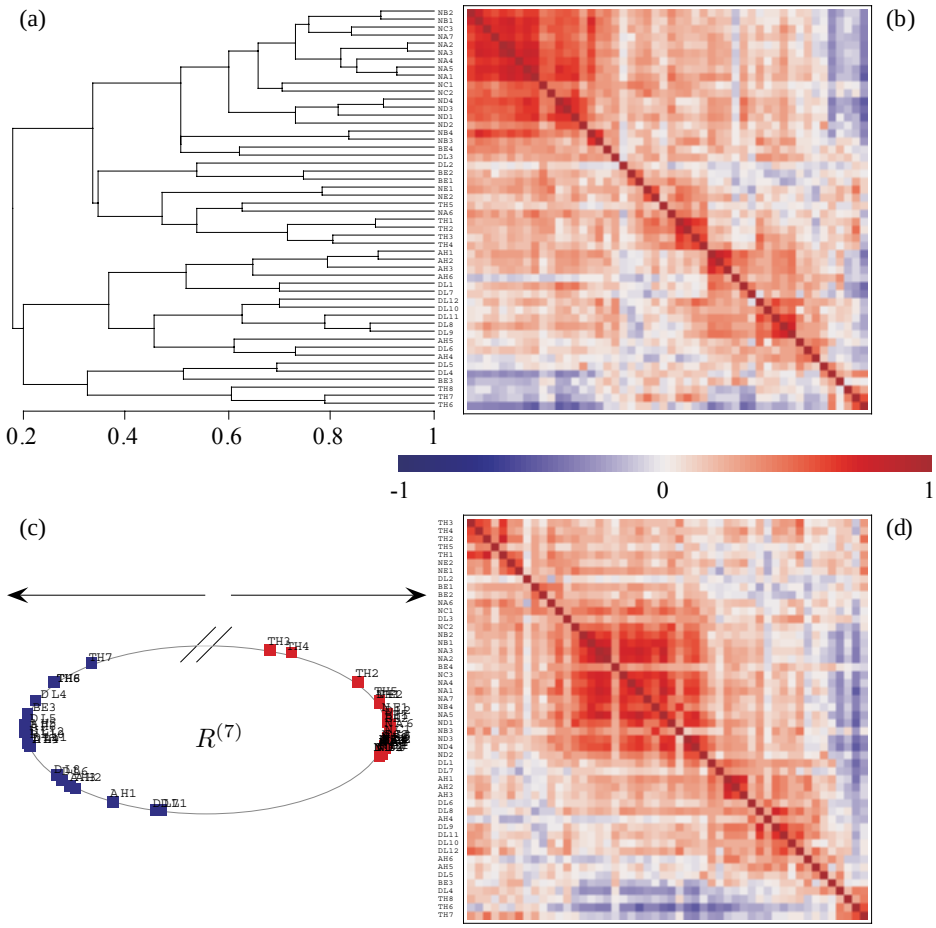

Figure 6. Seriation methods and the reconstructed correlation maps. (a) Divisive clustering tree with the rank-1 splitting rule; (b) Correlation map sorted by the tree seriation in (a); (c) Rank-two ellipse seriation at $R^{(7)}$; (d) Correlation map sorted by the ellipse seriation in (c).

5.3. Comparison of seriation algorithms using Iris data

In this section, Iris data (Fisher 1936) is used to compare the performance of the proposed seriations with several commonly used sorting algorithms. The target proximity matrix is the Euclidean distance matrix of the 150 iris flowers on four variables. Two conventional seriation algorithm sets, (a) farthest and nearest insertion spanning tour and (b) single, complete, and average linkage clustering trees, are compared with the proposed set (c) rank-2 ellipse, rank-1 tree, and rank-1 and 2 double ellipse. We present only the result for the best algorithm

from each of the two conventional sets, farthest insertion spanning, and average linkage tree, with our proposed rank-1 tree and rank-1 and 2 double ellipse. The conventional algorithms are discussed in Minnotte and West (1998). The four permuted distance maps are displayed in Figure 7. A clear near Robinson pattern can be identified in Figure 7d, using the proposed rank-1 and 2 algorithm. On the other hand Figure 7a with farthest insertion spanning algorithm only approximates the Robinson pattern in local regions. The performance of the average clustering tree in Figure 7b stands in between the other two. Our web page has result for comparison of all eight algorithms.

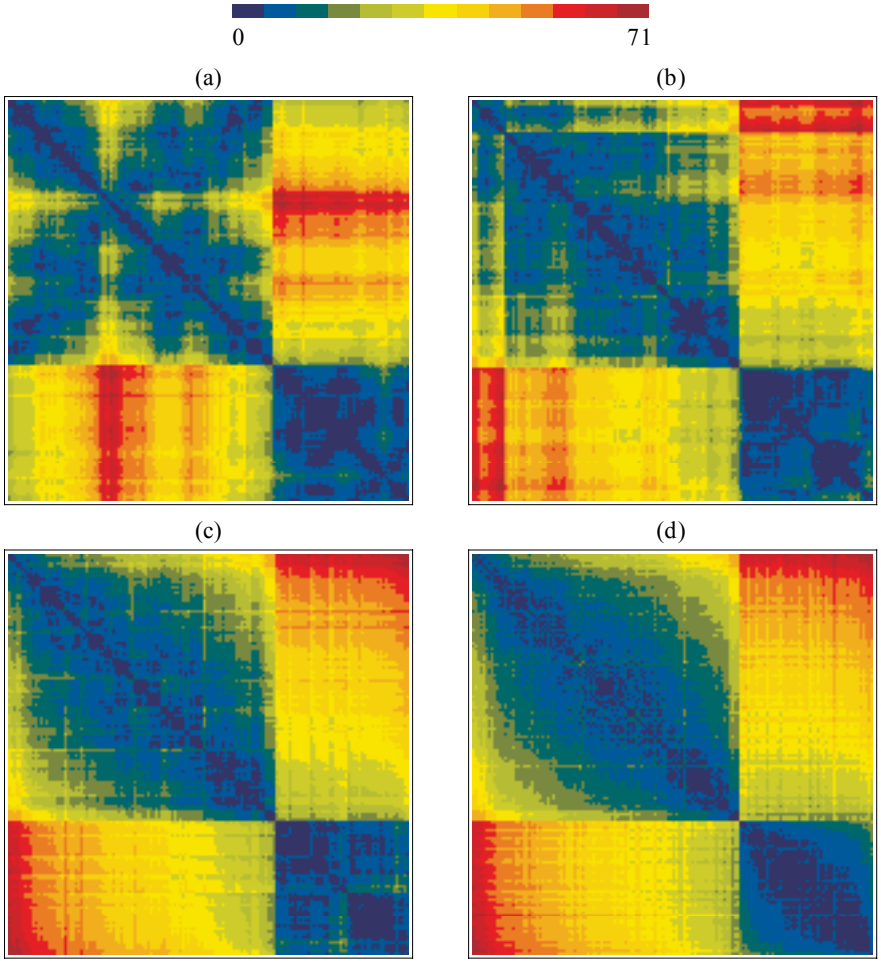

Figure 7. Permuted euclidean distance maps for iris data with conventional and proposed seriation algorithms. (a) Farthest insertion spanning tour; (b) Average linkage clustering; (c) Rank-1 tree; (d) Rank-1 and 2 double-ellipse.

For a numerical comparison, three anti-Robinson loss functions (Streng, (1978)) are calculated for each permuted matrix, $D = [d_{ij}]$, for the amount of deviation from a Robinson form with distance-type poximity:

$$AR(i) = \sum_{i=1}^{p} \Big[\sum_{j<k<i} I(d_{ij} < d_{ik}) + \sum_{i<j<k} I(d_{ij} > d_{ik}) \Big],$$

$$AR(s) = \sum_{i=1}^{p} \Big[\sum_{j<k<i} I(d_{ij} < d_{ik}) \cdot |d_{ij} - d_{ik}| + \sum_{i<j<k} I(d_{ij} > d_{ik}) \cdot |d_{ij} - d_{ik}| \Big],$$

$$AR(w) = \sum_{i=1}^{p} \Big[\sum_{j<k<i} I(d_{ij} < d_{ik}) |j-k| |d_{ij} - d_{ik}| + \sum_{i<j<k} I(d_{ij} > d_{ik}) |j-k| |d_{ij} - d_{ik}| \Big].$$

$AR(i)$ counts only the number of anti-Robinson events in the permuted matrix; $AR(s)$ sums the absolute value of anti-Robinson deviations; $AR(w)$ is a weighted version of $AR(s)$ penalized by the difference of column indices of the two entries. The results are summarized in Table 1. Clearly the proposed algorithms outperforms conventional methods by a significant margin.

The last column in Table 1 displays the amount of minimal span loss function for each permuted matrix, $MS = \sum_{i=1}^{p-1} d_{i,i+1}$. This is the object minimized in a minimal spanning algorithm such as the traveling salesman problem. The idea is to find a shortest path through all data points. The major concern is on the optimization for local structures only, different from the search for global structure in the Robinson setup. If only the local pattern is of concern then the conventional methods are better than the proposed algorithms, but the difference is not as significant as that in the Robinson setup.

Table 1. Anti-Robinson deviations for the permuted distance matrices for iris data.

Seriation Algorithm \ Loss Fun.	$AR(i)$	$AR(s)$	$AR(w)$	MS
Farthest Insertion Spanning	339,392	2,391,228.7	75,265,472.0	530.5
Average Linkage Clustering	148,950	381,679.8	4,576,913.2	558.8
Rank-1 Tree	86,367	166,953.6	1,613,008.1	625.5
Rank-1&2 Double Ellipse	83,217	146,115.5	1,602,892.1	789.5

5.4. The perfect symmetric structure in proximity matrix and the crystallographic structure

For a proximity matrix D with some forms of perfect symmetry, the convergence is to a matrix with a simpler structure that can be used to describe the symmetry pattern in D. We usually do not encounter proximity matrices with a perfect symmetric structure in a medical study of the sort looked at here, but one can search for a proximity matrix with a symmetric structure in physics, chemistry or molecular biology. An example is available on our web site.

5.5. The elliptical structure with the eigenvalue decomposition of the correlation matrices

We proved in Section 3.4 that every correlation matrix $R^{(n)}$ has all column vectors falling on the $\rho^{(n)}$-dimensional ellipsoid generated by the inverse or generalized inverse matrix of $R^{(n)}$. It is impossible for us to display the complete ellipsoid structure for each correlation matrix in the sequence unless it reaches the rank-two ellipse status. Instead, the leading two eigenvectors with their corresponding two-dimensional ellipse is plotted in Figure 1 for each iteration. Each such plot can only explain $(\lambda_1 + \lambda_2)/p \times 100\%$ of the total variation. If a data point falls exactly on the ellipse, then these two eigenvectors carry 100% information for that point at that iteration. This is seen in the negative symptom group in Figure 1 at $R^{(2)}$ for example. When a data point falls well inside the ellipse, the information for that point must be contained in the rest of the eigenvectors, see TH4 in Figure 1 at $R^{(2)}$. At $R^{(7)}$, when the rank equals to two, all fifty columns fall exactly on the ellipse and the plot carries 100% information for that iteration. From $R^{(8)}$ to $R^{(\infty)}$, these points move toward the two vertices along the curve of the ellipse.

In order to compare the sequence of eigenvector plots to conventional dimension reduction methods, we performed an exploratory factor analysis and a 2-dimensional non-metric multidimensional scaling analysis. While $R^{(0)}$ of Figure 1 mimics the pattern of the first two factor loadings by definition, the relative positions of points in $R^{(1)}$ is similar to the MDS configuration plot. The figures for factor analysis and MDS are available on our web site.

5.6. The sorted colored maps for the converging sequence of correlation matrices

It is of interest to visualize the formation of the clusters step by step when investigating a clustering problem. Both the eigenvector-plot and the colored map of the converging sequence provide users with this kind of information. There are some observations to be noted in this.

Fifty symptoms are arranged using the rank-2 elliptical seriation before plotting the sequence of correlation maps in Figure 8. For the original correlation map (top-left) in Figure 8, each individual column keeps its own characteristics, although several potential groups are forming at this early stage. From $R^{(0)}$ to $R^{(\infty)}$ is a dynamic grouping process. The problem is how to utilize the visual information in the sequence of eigenvector-plots and correlation maps for studying this process.

There is no rigorous rule but we do have the following suggestions. It is important to pay attention to both the within-group structure and the between-group difference while searching for a clustering pattern. For example, in Figure

8 at $R^{(4)}$, it may be seen that the within-group (main diagonal) correlations are high and homogeneous (dark red) while the between-group (off diagonal) structures all come with sharp edges. Usually this mature status can be identified with the help of the plot of the summations of squared eigenvalues at each iteration, as in Figure 9.

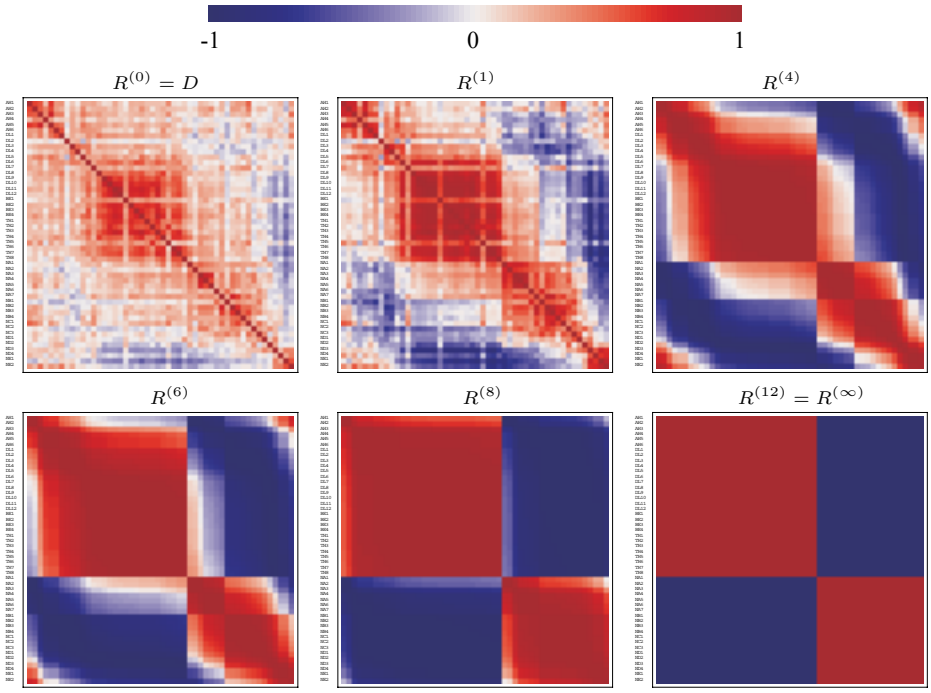

Figure 8. Sorted (by the rank-2 ellipse order) maps for the converging sequence of correlation matrices at selected iterations (0, 1, 4, 6, 8, 12) for the fifty symptoms.

The sequence of summations of squared eigenvalues generally has an increasing trend. We look for the iterations where this trend of increasing slows down (excluding the iterations just before convergence). At these iterations individual coefficients do not change much, which means the centering and product steps have no effect on them. This occurs when the process reaches a near stationary status where the formation of groups is mature. It is similar to the trap of a perfect symmetry structure. In Figure 9, this occurs at iterations 4 and 5. Started from iteration 7, this balanced status is broken and the process converges to the two winners and is trapped there. There are five major groups with minor substructure in Figure 8 at $R^{(4)}$. They are named V1 (TH3-BE2) for thought

disorder symptoms, V2 (NA6-ND2) for negative symptoms, V3 (DL1-DL8) for auditory hallucination symptoms, V4 (AH4-AH5) for loss of ego boundary symptoms, and V5 (DL5-TH7) for mania symptoms, respectively.

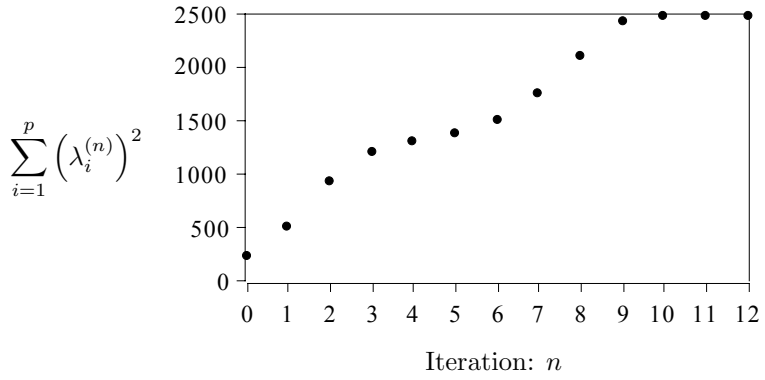

Figure 9. Plot of the sequence of summations of squared eigenvalues for each iteration for the fifty symptoms.

6. More Features of GAP

We have shown that the converging sequence has useful properties that can be applied in different areas of multivariate statistical analysis. In this section, We use the psychosis disorder data with 95 patients on 50 symptoms to illustrate the framework of a complete GAP analysis, in Figure 10. GAP integrates the following four major steps to extract and summarize information embedded in a multivariate data set with n subjects and p variables.

6.1. Raw data and proximity matrix maps with suitable color projection

The raw data matrix is denoted as D_a. A gray spectrum is applied to project ordinal numbers into gray dots with different intensities. The correlation matrix is calculated as the proximity matrix V_a for the 50 symptoms. For the 95 patients, the correlation matrix is also used as the proximity matrix S_a (We tried the Euclidean (standardized) distance as well). The diverging blue-red color scheme is used to represent the bi-directional property of the correlation coefficients. For a data profile with various variable scales, variables can be transformed (standardized) and projected through a suitable color spectrum to represent the characteristics of the scales. The covariance matrix with the Euclidean (standardized) distance matrix can also be calculated as V_a and S_a.

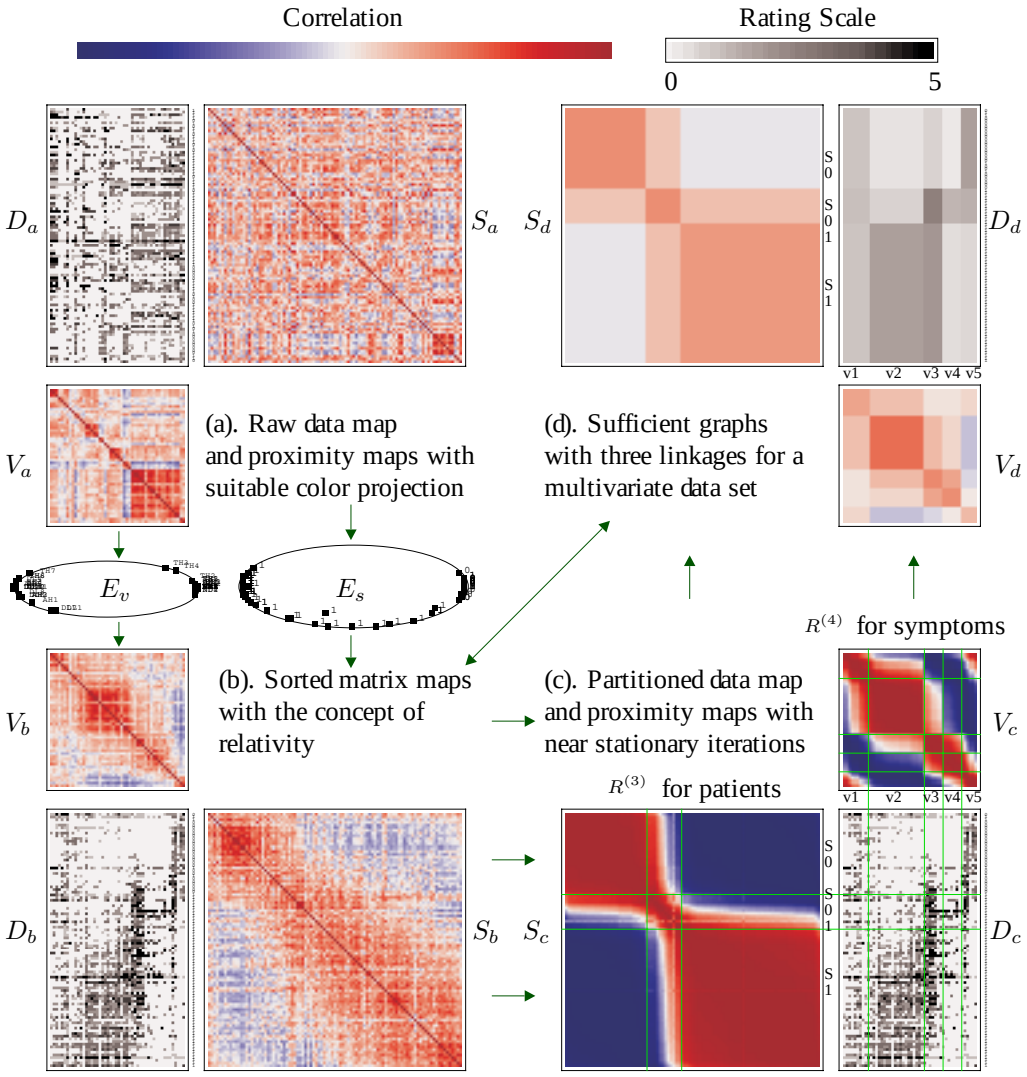

Figure 10. Complete GAP procedure for the psychosis disorder data set with ninety-five patients and fifty symptoms.

6.2. Sorted matrix maps with the concept of relativity

The next step is to form the sequences of correlation matrices for V_a and S_a to identify the ellipses E_V and E_S at iterations 7 and 5 respectively. The elliptical seriations for the patients and the symptoms are then applied to arrange the two correlation matrices V_a and S_a into V_b and S_b. The same seriations are also used to reshape the raw data matrix D_a into D_b. The difference between

V_a and V_b is not much since V_a is already grouped by the SAPS and SANS symptom tables. However, there is a dramatic change from S_a to S_b since the patients are admitted in a random order. There is a clear latent structure in D_b. A band of dark gray dots moves from the upper right corner to the lower left corner. Since the seriations for D_b are identical to those for V_b and S_b, these three maps are closely related to each other and should be cross-examined to find the information embedded in the raw data matrix and the two proximity matrices. We have succeeded in using the geometric information captured by the converging ellipse to guide the matrix sorting process so that similar objects are placed closer to each other.

We shall call this concept of placing similar (distinct) objects at positions close to (far away from) each other in a plot for representing the association structure the concept of relativity of a statistical graph.

6.3. Partitioned matrix maps with near stationary iterations

In Section 5.6, Figure 8 at $R^{(4)}$ partitions the correlation matrix of the fifty symptoms into five major groups. In this section we take a look at the possible patient-clusters and the general behavior of patient-clusters on symptom-groups. It seems that there is no clear patient-cluster structure in S_b except the negative between-group correlations in the off-diagonal area. It takes nine iterations for S_b to converge and to split all 95 patients into two groups. The first group is a mixed group of 26 bipolar-disorder patients with 12 schizophrenia patients, the second is a pure group with 57 schizophrenia patients.

In Figure 10c at $R^{(3)}(= S_c)$, a coherent group in the upper left corner is easily identified. This group, to be denoted as S0, consists of all 26 bipolar-disorder patients and only 4 schizophrenia patients. At the lower right corner there is a large group, S1, of pure schizophrenia patients, but the structure is not as tight as that of S0. Between S1 and S0 is a group of pure schizophrenia patients, but the between group relationships for this group with S1 and S0 are about equal. We use S01 to denote this group of patients.

We then plot the two-way sorted raw data map with the sorted correlation maps for patients at $R^{(3)}$ and for symptoms at $R^(4)$ attached to it, in Figure 10c. The green lines represent the partitions for symptom groups and for patient-clusters. The general behavior of patient-clusters on the symptom-groups can be easily identified in D_c.

6.4. The sufficient graph with three multivariate linkages

In order to extract and summarize the information in Figure 10b, we can put these matrix maps into a simplified version with the partitions in Figure

10c. Illustrated in Figure 10d are the mean-structure maps of the three matrices for raw data and proximities. Original proximity matrices for variables and subjects are represented by squares with different mean intensities on the diagonal for within-group structure, and rectangles off-diagonal for between-group relationship. The double sorted raw data matrix map is also represented by rectangles with various mean-gray intensities to express the interaction effect between each subject-cluster on every variable group. These three mosaic-displays in Figure 10d contain the principal structural information embedded in the original data set. The mean function in Figure 10d can be replaced with any statistic for displaying desired information structure. We name these three mosaic displays the sufficient graph for a multivariate data set. The sufficient graph is then used to answer the three multivariate problems raised by the psychiatrist. Fifty symptoms are divided into five symptom-groups with different within- and between-group structure. Ninety-five patients are also grouped into three clusters. The general behavior of these three patient clusters on each of the five symptom groups can now be easily comprehended. One can always go back to consult the three original sorted matrix maps (Figure 10b) for fear of losing too much information.

7. Discussion

Many useful properties of the converging sequence of iteratively computed correlation matrices given a proximity matrix have been introduced. Eigenvector structures of correlation matrices in earlier iterations mimic the effects in dimension reduction techniques such as factor analysis and multidimensional scaling. Near stationary iterations with the sorted colored maps can be employed to identify structural (clustering) information embedded in the data. A rank-two iteration finds the Robinson seriation in the proximity matrix while the converged rank-one structure splits a proximity matrix into two sets with the divisive clustering tree and the rank-one tree seriation. A non-rank-one converged pattern matrix can also be used to study the symmetry pattern that exists in the proximity matrix. With the aid of eigenvector projections with ellipse and correlation matrix maps, the convergent process is a powerful and dynamic visualization environment for the many faces of high-dimensional statistical data analyses.

The original purpose of this study was to investigate the general behavior of patient-clusters on symptom-groups for the psychosis disorder data set. Instead of using many of the available multivariate analysis methods, we have used the generalized association plots (GAP) for information visualization.

Our goal is partially accomplished. Through careful examination of the double sorted raw data matrix map with the sufficient graphs and the sequence of correlation maps for both the symptoms and the patients, we gain understanding

of the symptom groups and the patient clusters. However, this is only the beginning of an effort to understand the whole process of development of psychosis disorder disease. The fifty SAPS and SANS symptoms used in this study are only a part of the many rating scales in the MPGRP project. The complete data base comes with different rating scales (nominal, ordinal, and continuous) at different time points, with biological background information and genetic marker profile of each patient in the study. It is an extremely difficult challenge to develop multiphase longitudinal and categorical versions of GAP to help in understanding this kind of large-scale study.

Acknowledgements

Some of the figures in this article are either reduced in size or partially presented due to limited space available. Readers are encouraged to browse the full and complete set of figures with animations at http://gap.stat.sinica.edu.tw/GAP/Introduction/. The current version of GAP is written in XLISP-STAT (Tierney, 1990). We are converting it to C++ version. Support for this research was provided in part by the National Science Council (NSC 88-2118-M-001-017) and by National Health Research Institute (DOH 87- HR-306, 88-HR-82), R.O.C. The author thanks Su-Yun Huang and Ker-Chau Li for many helpful discussions, Hai-Gwo Hwu for providing the data set analyzed in this article. The author is also grateful to Donald Ylvisaker and Jan de Leeuw for valuable suggestions. Special thanks go to an anonymous referee for providing thoughtful suggestions when this article was previously reviewed by another journal.

References

Andreasen, N. C. (1983). The Scale for the Assessment of Negative Symptoms (SANS). University of Iowa, Iowa City, IA.

Andreasen, N. C. (1984). The Scale for the Assessment of Positive Symptoms (SANS). University of Iowa, Iowa City, IA.

Andreasen, N. C., Arndt, S., Alliger, R., Miller, D. and Flaum, M. (1995). Symptoms of schizophrenia: methods, meanings, and mechanisms. *Arch. of Gen. Psych.* **52**, 341-351.

Breiger, R. L., Boorman, S. A. and Arabie, P. (1975). An algorithm for clustering relational data with applications to social network analysis and comparison with multidimensional scaling. *J. Math. Psych.* **12**, 328-383.

Caraux, G. (1984). Rearrangement and visual representation of matrices with numerical data: an iterative algorithm (French). *Rev. Statist. Appl.* **32**, 5-23.

Chen, C, H. and Chen, J. A. (2000). Interactive diagnostic plots for multidimensional scaling with applications in psychosis disorder data analysis. *Statist. Sinica* **10**, 665-691.

Fisher, R. A. (1936). The use of multiple measurements in taxonomic problems. *Ann. Eugen.* **7**, 179-188.

Gale, N., Halperin, C. W. and Costanzo, C. M. (1984). Unclassed matrix shading and optimal ordering in hierarchical cluster analysis. *J. Classification* **1**, 75-92.

Hubert, L. (1976). Seriation using asymmetric proximity measures. *British J. Math. Statist. Psych.* **29**, 32-52.

Kruskal, B. J. (1977). A theorem about CONCOR. Unpublished manuscript.

Lin, A. S., Chen, C. H., Hwu H. G., Lin H. N. and Chen J. A. (1998). Psychopathological dimensions in Schizophrenia: a correlational approach to items of the SANS and SAPS. *Psychiatry Research* **77**, 121-130.

Marcotorchino, F. (1991). Seriation problems: an overview. *Applied Stochastic Models Data Anal.* **7**, 139-151.

Lyer, V. R., Eisen, M. B., Ross, D. T., Schuler, G., Moore, T., Lee, J. C. F., Trent, J. M., Staudt, L. M., Hudson, J. Jr., Boguski, M. S., Lashkari, D., Shalon, D., Botstein, D. and Brown, P. O. (1999). The transcriptional program in the response of human fibroblasts to serum. *Science* **283**, 83-87.

McFarlane, M. and Young, F. W. (1994). Graphical Sensitivity Analysis for Multidimensional Scaling. *J. Comp. and Graph. Statist.* **3**, 23-33.

McQuitty, L. L. (1968). Multiple clusters, types, and dimensions from iterative intercolumnar correlational analysis. *Multivariate Behavioral Research* **3**, 465-477.

Minas, I. H., Stuart, G. W., Klimidis, S., Jackson, H. J., Singh, B. S. and Copolov, D. L. (1992). Positive and negative symptoms in the psychoses: Multidimensional scaling of SAPS and SANS items. *Schizophrenic Research* **8**, 143-156.

Minnotte, M. and West, R. W. (1998). The Data Image: A tool for exploring high dimensional data sets. 1998 *Proceedings of the Section on Statistical Graphics*, American Statistical Association.

Murdoch, D. J. and Chow, E. D. (1996). A graphical display of large correlation matrices. *Statist. Comput.* **50**, 178-180.

Robinson, W. S. (1951). A method for chronologically ordering archaeological deposits. *Amer. Antiquity* **16**, 293- 301.

Sokal, R. R. and Sneath, P. H. (1963). *Principles of Numerical Taxonomy.* Freeman, San Francisco.

Streng, R. (1991). Classification and seriation by iterative reordering of a data matrix. In *Classification, Data Analysis, and Knowledge Organization: Models and Methods with Applications* (Edited by H. H. Bock and P. Ihm), 121-130. Springer-Verlag, New York.

Stuart, G. W., Malone, V., Currie, J., Klimidis, S. and Minas, I. H. (1995). Positive and negative symptoms in neuroleptic-free psychotic inpatients. *Schizophrenic Research* **16**, 175-188.

Takane, Y., Young, F. W. and de Leeuw, J. (1977). Nonmetric individual differences multidimensional scaling: an alternating least squares method with optimal scaling features. *Psychometrika* **42**, 7-67.

Tierney, L. (1990), *Lisp-Stat: An Object-Oriented Environment for Statistical Computing and Dynamic Graphics.* John Wiley, New York.

Wen, X. L., Fuhrman, S., Michaels, G. S., Carr, D. B., Smith, S., Baker, J. L. and Somogyi, R. (1998). Large-scale Temporal gene expression mapping of central nervous system development. *Proceedings of The National Academy of Science* **95**, 334-339.

Institute of Statistical Science, Academia Sinica, Taipei 115, Taiwan.

E-mail: cchen@stat.sinica.edu.tw

(Received October 2000; accepted June 2001)