Generalized association plots (GAP): Dimension free information visualization environment for multivariate data structure

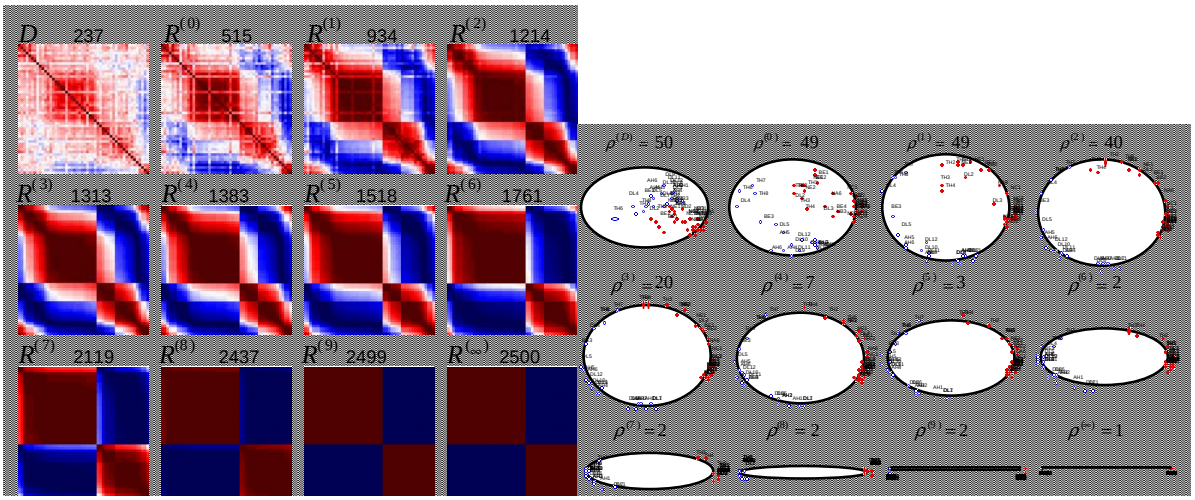
Chun-houh Chen, Shun-Chuan Chang, Yueh-Yun Chi, and Chih-Wen Ou-Young

Academia Sinica, Taiwan

Abstract: GAP is a dimension free visualization environment for multivariate data structure. Given a multivariate data set, GAP first compute the proximity matrices for variables as well as for subjects. Proper seriations (permutations) are searched for rearrange these two matrices to satisfy certain properties. Double sorted raw data matrix together with two sorted proximity matrices are then projected through appropriate color spectrums to create matrix maps. These three maps should be cross-examined to identify important information structure embedded in the raw data matrix. GAP has unique seriation algorithm for identifying permutations of matrices with better global sense and clustering algorithm for finding multiple clustering patterns co-exist in proximity matrices. Extensions have been added to GAP for analyzing more complicated formats such as categorical, longitudinal, and multiple-set structure.

1 CONVERGING SEQUENCE OF CORRELATION MATRICES

The computation kernel of GAP is a converging sequence of iteratively generated correlation matrices, Figure 1. This sequence of correlation matrices is used to dynamically identify seriation and clustering for a given proximity matrix.



(a) (b)

Figure 1. Converging Sequence of Iteratively Generated Correlation Matrices.
(a). Sorted Color Maps for the Correlation Matrices; (b). First Two Eigen-Vectors for the Correlation Matrices.

2 BASIC PRINCIPLE OF GAP

There are three major pieces of information contained in any multivariate data set with *n* subjects and *p* variables: 1. the linkage amongst n subject points in the p-dimensional space; 2. the

linkage between p variable vectors in the n-dimensional space; and 3. the interaction linkage between the sets of subjects and variables. We use a psychosis disorder data with 95 patients on 50 symptoms to illustrate the framework of a standard GAP analysis, in Figure 2. GAP integrates the following four major steps to extract and summarize information embedded in a multivariate data set.

2.1 Raw Data and Proximity Matrix Maps with Suitable Color Projection The raw data matrix is denoted as D_a in Figure 2a. A gray spectrum is applied to project numerical 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, also the correlation matrix is used as the proximity matrix. The diverging blue-red color scheme is used to represent the bi-directional property of the correlation coefficients.

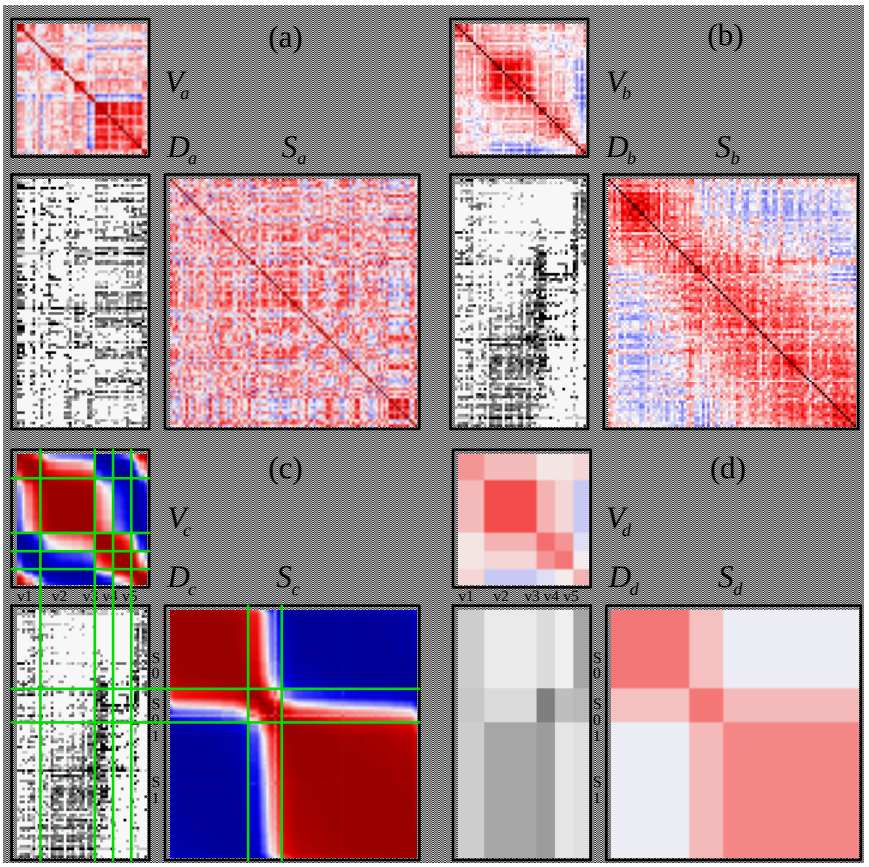


Figure 2. Standard GAP Procedure.

(a). Original Raw Data and Proximity Matrices Maps; (b). Sorted Data Matrix Map and Proximity Matrix Maps; (c). Clustering for Variables and Subjects; (d). Sufficient Graph with Three Multivariate Linkages.

2.2 The Sorted Matrix Maps with the Principle of Geometry The next step is to find proper seriations for V_a and S_a respectively. Seriation is a data analytic tool for finding a permutation or ordering of a set of objects using a data matrix. The seriations are then applied to arrange the two correlation matrices V_a and S_a into V_b and S_b in Figure 2b. 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 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.

2.3 Partitioned Matrix Maps with near Stationary Iterations After the seriations have been applied to the matrix maps, the potential groups of variable and clusters of subject are identified using dynamic clustering procedures (to be discussed in Section 2) as displayed in Figure 2c. The sorted proximity matrix maps for variables and subjects are then partitioned into squares on the diagonal for within-group mean correlation structure and rectangles off diagonal for between-group mean correlation relationship, Figure 2c. The double sorted raw data matrix map is also partitioned into rectangles to represent the mean interaction effect between each subject-cluster on every variable-group.

2.4 The Sufficient Graph with Three Multivariate Linkages In order to extract and summarize the visualized information in Figure 2b, we can further convert these matrix maps into a simplified version. Illustrated in Figure 2d are the mean-structure maps of the three matrices for raw data and proximities. These three mosaic-displays in Figure 2d contain the principal structural information embedded in the original data set. The mean function in Figure 2d can be replaced with any statistic for displaying desired information structure. We shall 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.

3 EXTENSIONS OF GAP

Section 2 introduced a typical GAP procedure for a data set with continuous variables measured at a single time point. We have developed several modules of GAP for handling many possible variants of data structure.

3.1 Categorical GAP Figure 3 displays four matrix maps with different data type. Categorical GAP with dual scaling can be used for information visualization for binary and nominal data matrices.

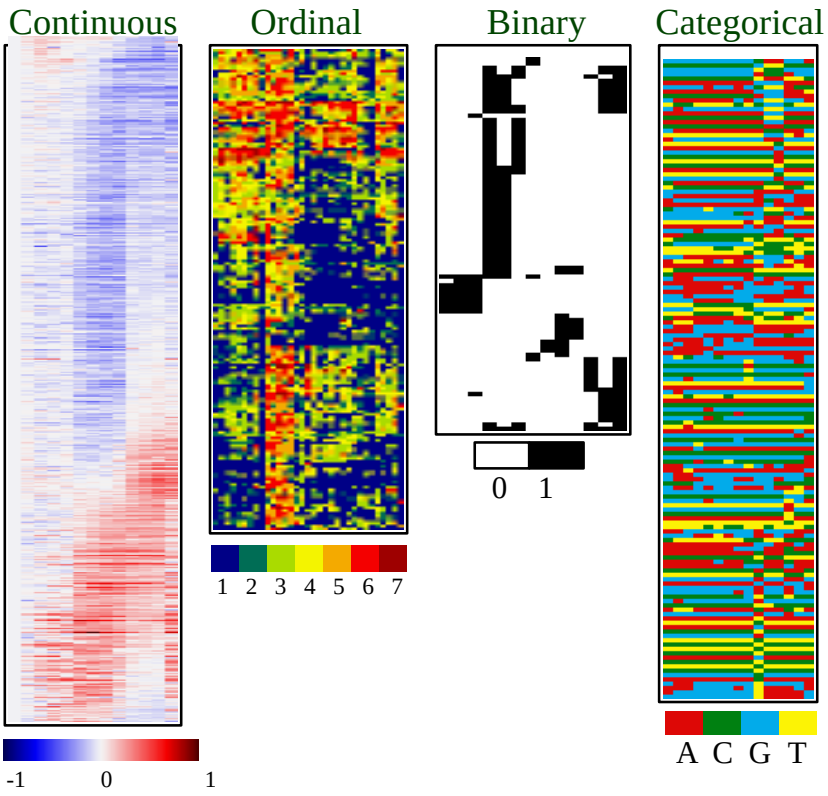


Figure 3. Information Visualization for Data Matrices with Different Data Types.

3.2 Longitudinal GAP It is possible that data profile may be measured at multiple time points as illustrated in Figure 4. When the data profile is observed at multiple time points, we have an extra piece of linkage for subjects and variables over time. This extra linkage makes the visualization of longitudinal multivariate data profiles a much harder statistical challenge than

the single profile C edure. Longitudinal GAP is created for visua a profiles with longitudinal structu

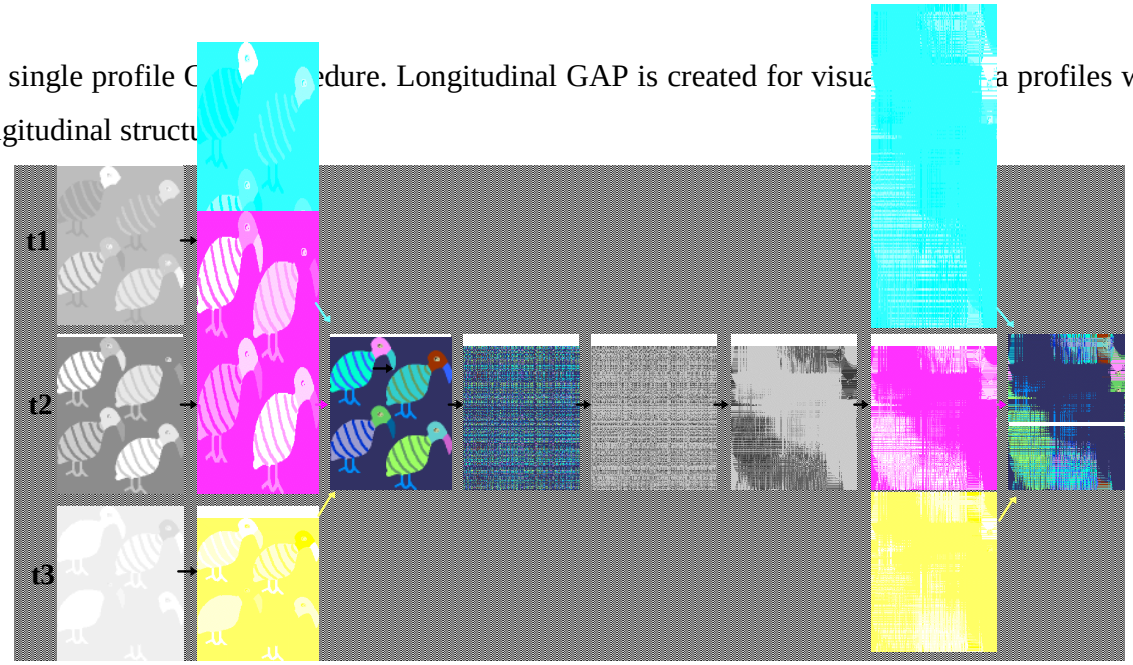


Figure 4. Composed Color Profile with Seriation for Data Matrices at Multiple Time Points

3.3 Canonical GAP Given two data sets with different sets of variables for the same subjects, we would like to explore the relationship with Canonical GAP, Figure 5.

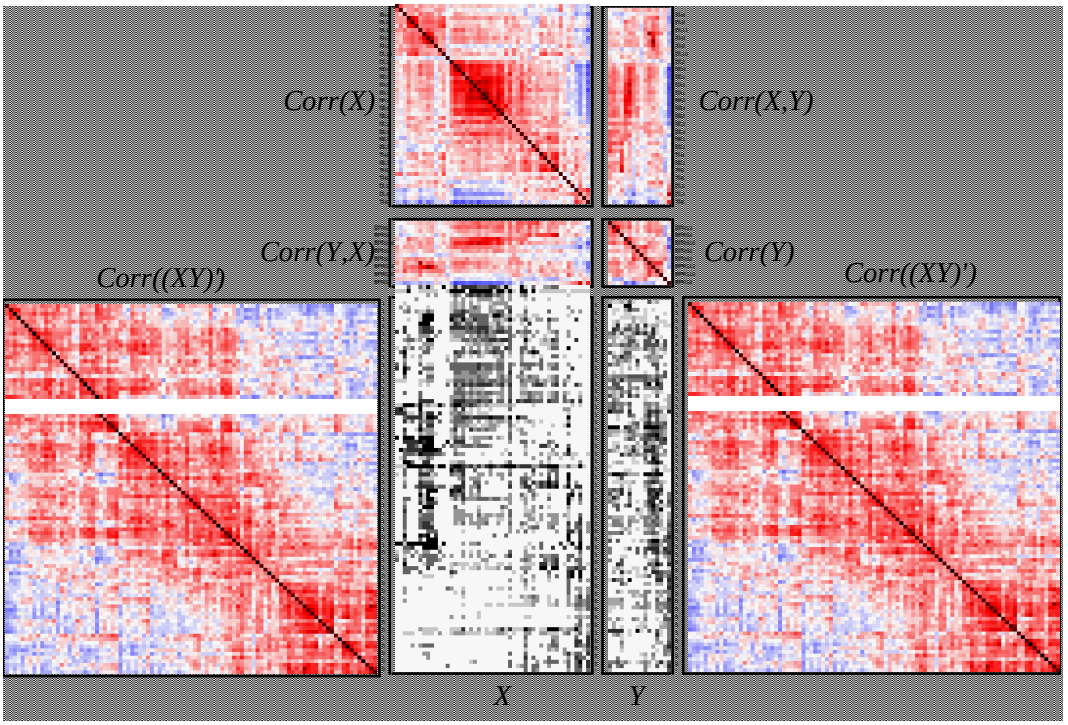


Figure 5. Information Visualization for Comparing Two Sets of Variables with Canonical GAP.

3.4 Cartograph GAP When each subject in a multivariate categorical data set is affiliated to a district in a map, Cartography GAP extents the power of Categorical GAP to systematically identify suitable color scheme for coloring districts in the map, Figure 6.

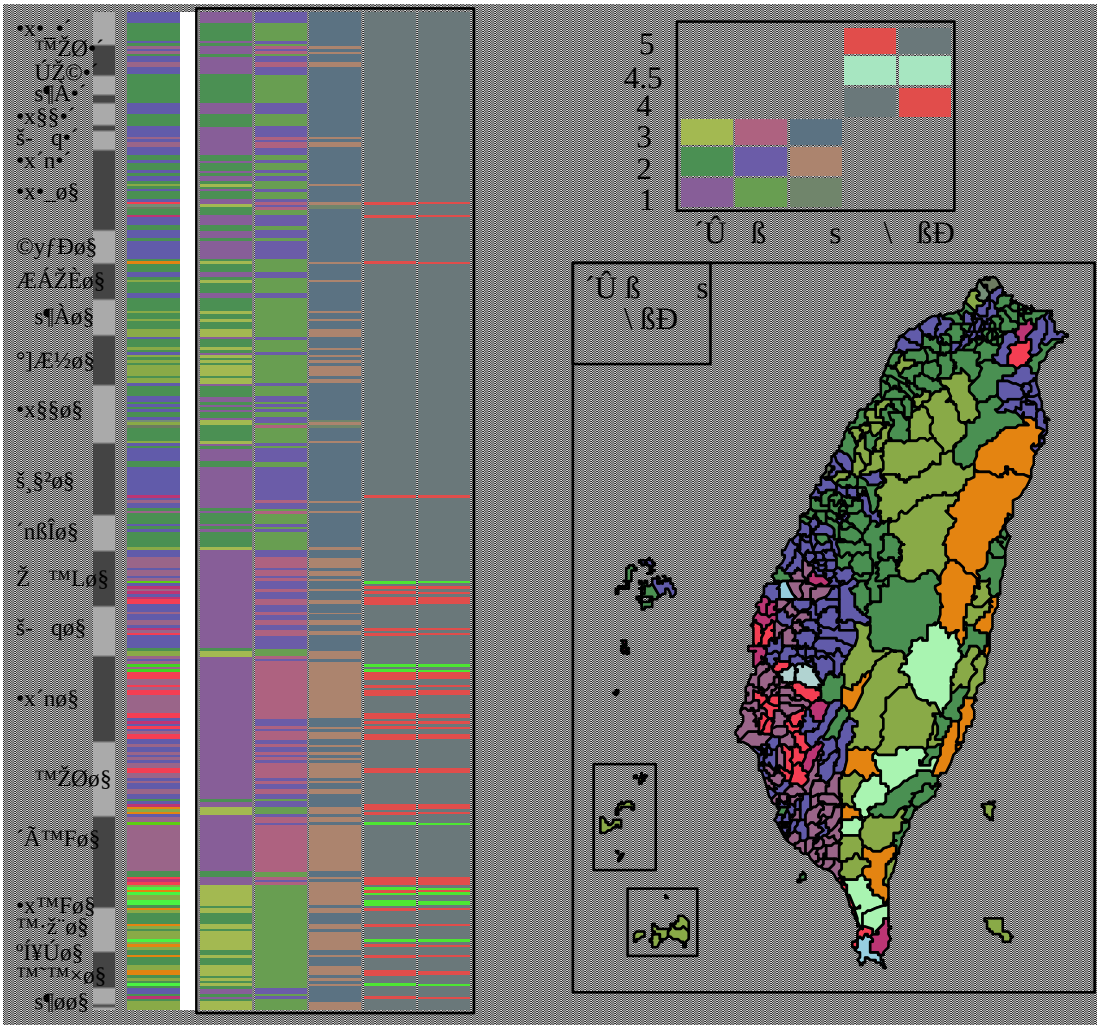


Figure 6. Cartography GAP for Taiwan Presidential Election 2000.