Data Visualisation
CSE3020

Matrix Visualization

- Boxplots, Scatter plots, aided by interactive functionality, have provided statistical community with important graphical tools
- These tools uses varies techniques for reducing dimensions, are exploring the structure of the data when there are a moderate number of variables and the structure is not too complex.
- However, in Big Data era, the data generation is rapid and getting high-demensional datasets in many applications $\rightarrow$ the number of variables associated with the dataset can easily reach tens of thousands
- Dimension reduction tools oten become less effective when applied to the visual exploration of information structures embedded in high-dimensional datasets

Matrix Visualization

- matrix visualization, when integrated with computing, memory, and display technologies, has the potential to enable us to visually explore the structures that underlie massive and complex datasets
- Matrix Visualization is a graphical technique that can simultaneously explore the associations between thousands of subjects, variables, and their interactions, without needing to first reduce the dimensions of the data.
- Matrix visualization involves permuting the rows and columns of the raw data matrix using suitable seriation (reordering) algorithms, together with the corresponding proximity matrices.
- The permuted raw data matrix and two proximity matrices are then displayed as matrix maps via suitable color spectra, and the subject clusters, variable groups, and interactions embedded in the dataset can be extracted visually.

Matrix Visualization

- It was introduced by **Bertin** (1967)as a reorderable matrix for systematically presenting data structures and relationships.
- The color histogram of Wegman (1990) was the first color matrix visualization to be reported in the statistical literature. $\rightarrow$ extended idea of color histogram can be used for outlier detection.
- Some matrix visualization techniques were developed to explore only proximity matrices
  - Ling (1973) looked for factors of variables by examining relationships using a shaded correlation matrix;
  - Murdoch and Chow et.al used elliptical glyphs to represent large correlation matrices
- Important steps in matrix visualization:
  - The permutation (ordering) of the columns and rows of a data matrix
  - Proximity matrix for variables and samples

Matrix Visualization

- **The Basic Principle of Matrix Visualization**
  - We use GAP $\rightarrow$ generalized association plots to illustrate the basic principles of matrix visualization
  - The dataset has continuous data, using 6400 genes and 851 micro-array experiments in the published yeast expression database for visualization.
  - For the purposes of illustration, we have selected 15 samples and 30 genes across these samples
  - rows correspond to genes and columns to microarray experiments
  - roles played by rows and columns can be interchangeable $\rightarrow$ samples and variables are treated symmetrically

Matrix Visualization

- **The Basic Principle of Matrix Visualization**
  - The first step is to produce raw data matrix X_{30X15}.
  - The proximity matrix for the rows R_{30X30}
  - The proximity matrix for the columns C_{15X15}
  - The proximity matrix are calculated with user-specified similarity (or dissimilarity) measures.
  - The three matrices are then projected through suitable **color spectra** to construct corresponding matrix maps.
  - In this, each matrix entry (raw data or proximity measurement) is represented by a color dot.

Matrix Visualization

- The simple matrix visualization for assumed dataset

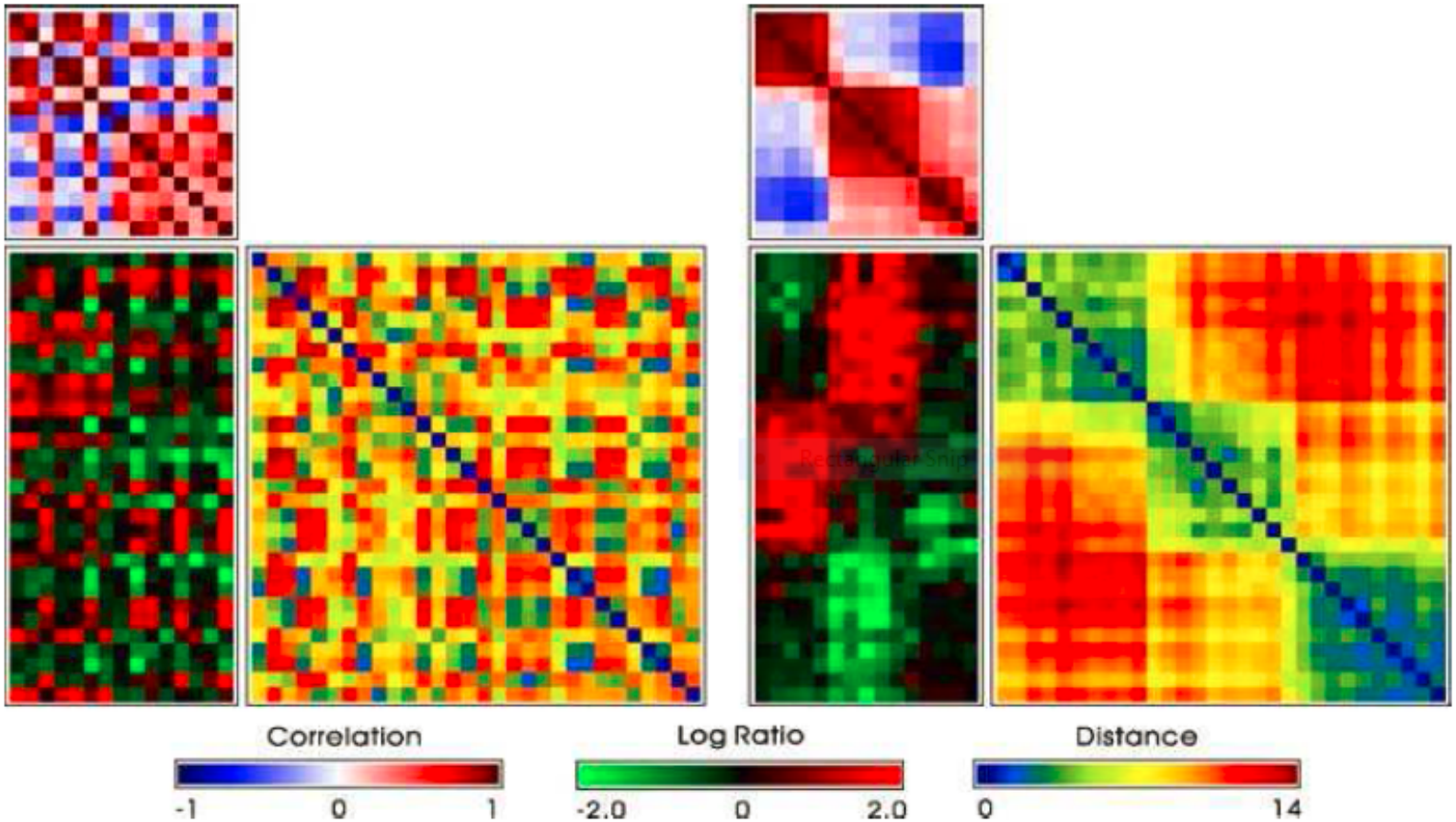

Figure 15.1. [This figure also appears in the color insert.] *Left*: unsorted data matrix (log ratio gene expression) map with two proximity matrix (Pearson correlation for arrays and Euclidean distance for genes) maps for Dataset 1. *Right*: application of elliptical seriations to the three matrix maps on the left panel

- **The Basic Principle of Matrix Visualization**
  - The left panel in the above figure shows the raw data matrix of $\log_2$ transformed ratios of expressions coded by a bidirectional green-black-red spectrum.
  - Pearson correlations for between-array relations coded by a bidirectional blue-white-red spectrum
  - Euclidean distances for between-gene relations coded by a unidirectional rainbow spectrum.
  - In the raw data matrix map, a red dot in the ijth position of the map for X_{30X15} means that ith gene at the jth array is relatively **up**.
  - In the raw data matrix map, a green dot in the ijth position of the map for X_{30X15} means that ith gene at the jth array is relatively **down**.

Matrix Visualization

- **The Basic Principle of Matrix Visualization**
  - A black dot stands for a relatively non-differentially expressed gene/array combination.
  - A red point in the ijth position of the C_{15X15} matrix map represents a positive correlation between arrays i and j.
  - A blue point in the ijth position of the C_{15X15} matrix map represents a negative correlation between arrays i and j.
  - Darker (lighter) intensities of color stand for stronger absolute correlation coefficients
  - white dots represent no correlations
  - A blue point in the ijth position of the R_{30X30} matrix map represents a relatively small distance between genes i and j.
  - A red point in the ijth position of the R_{30X30} matrix map represents a relatively large distance between genes i and j.
  - Yellow dot represents a median distance

Matrix Visualization

- **Data Transformation**
  - It may be necessary to apply transformations such as log, zero mean, unit variance or normalization to the raw data before the data map construction or proximity matrices calculated
  - It gives the meaningful visual representation of the data structure
- **Selection of Proximity Measures**
  - Proximity matrices have two major functions:
  - 1. To serve as the direct visual representation of the relationships among variables and between samples;
  - 2. To serve as the medium used to reorder the variables and samples for better visualization of the three matrix maps.
  - The selection of proximity measures in matrix visualization plays a more important role than it does in numerical or modeling analyses.

- **Selection of Proximity Measures**
  - Pearson correlation can be serves as the measure of proximity between variables.
  - Euclidean distance can be serves as the measure of proximity between samples
  - Spearman's rank correlation and Kendall's tau coefficients are used to measure proximity between variables in non-parametric (non-linear relations)
  - Isomap distance can be used to measure the distance between samples in non-linear relations

Matrix Visualization

- **Color Spectrum**
  - The selection of an appropriate color spectrum can be critical
  - It is user-dependent in visualization
  - It plays major role for information extraction from data and proximity matrices.
  - The selection of a suitable color spectrum should focus on the capacity to express **numerical** nature individually and globally in the matrices

- **Color Spectrum**

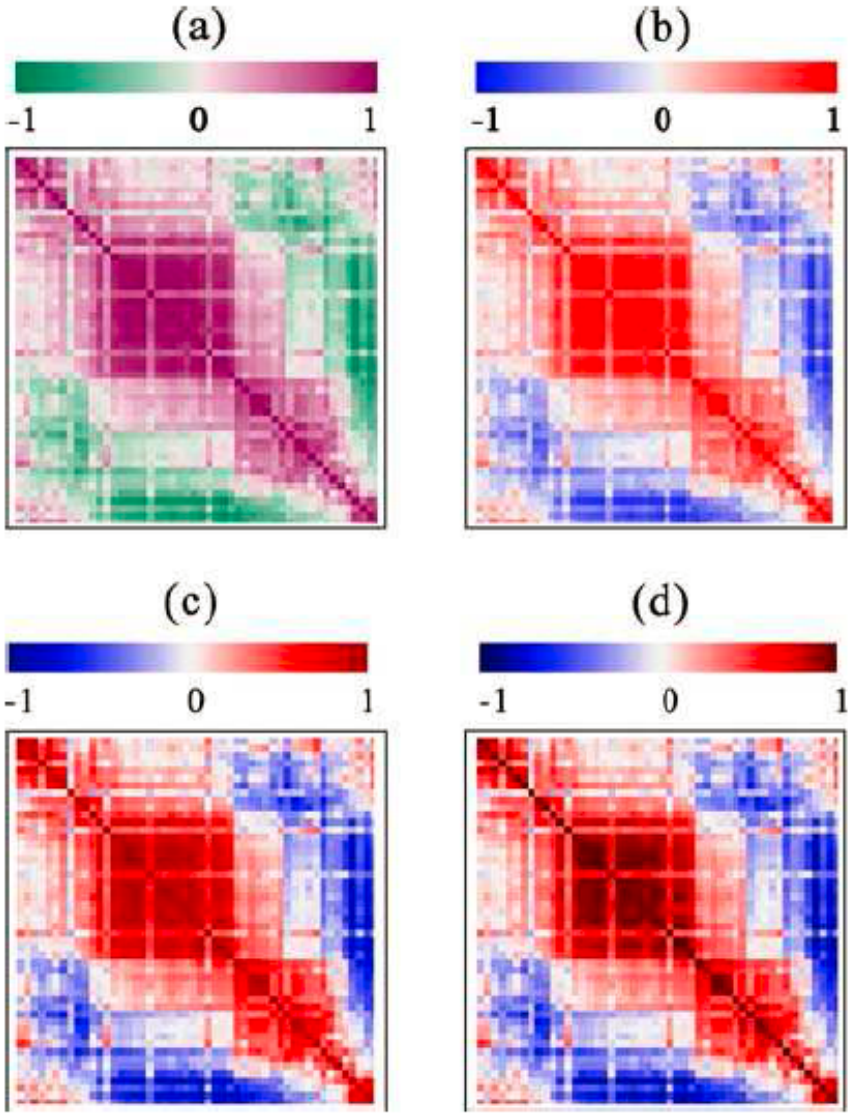

Figure 15.2. [This figure also appears in the color insert.] Four color spectra applied to the same