Statistical Graphics and Visualization

中央研究院統計科學研究所2005秋季課程
Statistics and Machine Learning
01/06/2006

Han-Ming Wu
Chun-houh Chen
Institute of Statistical Science, Academia Sinica
中央研究院 統計科學研究所

Outlines

- **Introduction**
  - Graphical Methods
  - The Advantages of Graphical Approaches to Data Analysis
  - Graphical Perception
- **Graphical Displays for Univariate Data**
  - Histogram
  - Density Plot (Smoothed Histograms)
  - Quantile Plots
  - Box Plots, Dotplot
- **Graphical Displays for Bivariate Data**
  - 2D/3D Scatterplot, Scatterplot Matrix
  - Extensions of Scatterplots
  - Scatter plot and MA plot
  - Glyph Plot, Volcano Plot
  - Quantile-Quantile Plot (QQplot)
- **Graphical Displays for Multivariate Data**
  - Star Plot, Radar Plot
  - Chernoff Faces Plot
  - Parallel Plot
  - Andrews' Plot
- **High-dimensional Data : Dimension Reduction Techniques**
  - Biplot
  - Principal Component Analysis (PCA)
  - Sliced Inverse Regression (SIR)
  - Multidimensional Scaling (MDS)
  - Self-Organizing Maps (SOM)
- **High-dimensional Data: Dimension-free Visualization**
  - Dendrogram and Heatmap
  - Generalized Association Plots (GAP)
- **Visualizing Categorical Data**
  - Mosaic Display
  - Correspondence Analysis
  - Multiple Correspondence Analysis
  - Categorical Generalized Association Plots

Software for Statistical Graphics and Analysis
Data Desk, R, GGobi,GAP

Graphical Methods

- The ***purpose*** of statistical graphics is to provide visual representations of quantitative/qualitative information.
- As a methodological tool, statistical graphics comprise a set of strategies and techniques that provide the researchers with important insights about the data under examination and help guide the subsequent steps of the research process.

The objectives of Graphical Methods

- Statistical graphics are useful for exploring the contents of a data set.
  - Statistical graphics can be used to address questions about the variables in an analysis (e.g., what are the distributional shapes, ranges, typical values or unusual observations?).
- Statistical graphics are used to find structure in data.
  - Tukey (1977): look at the data and see what it seems to say.
- Graphical methods can be used for checking assumptions in statistical models.
  - Direct visual representations of a statistical model and the residuals from the model greatly facilitate the examination of assumptions.
- Statistical graphics are very useful for communicating the results of an analysis.

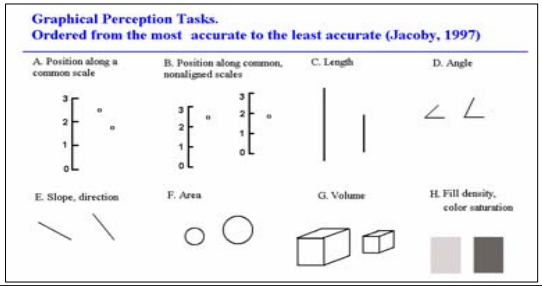
The Advantages of Graphical Approaches to Data Analysis

- Graphs provide useful summaries for large, complicated data sets.
- Graphs provide an effective means of downplaying the details of the data and emphasizing the important features.
- Graphical analysis facilitates greater interaction between the researcher and the data. Effective visual presentations highlight interesting and unusual aspects of the quantitative information under investigation.
- Graphs usually are superior for revealing patterns, trends, and relative quantities within data sets regardless of their size.

Graphical Perception

Human reception and comprehension of graphical information involves three fundamental perceptual task:

- **Detection**: the visual recognition of a geometric aspect that encodes a physical value. The basic information from the data must be discernible in the graph.
- **Assembly**: assembly is the process of discerning patterned regularities among the discrete elements of a graphical display.
- **Estimation**: Estimation is the visual assessment of the relative magnitudes of two or more quantitative physical values.



Graphs and Data/Information Visualization

What is Visualization?

- To visualize = to make visible, to transform into pictures.
- Making things/processes visible that are not directly accessible by the human eye.
- Transformation of an abstraction to a picture.
- Computer aided extraction and display of information from data.

Tegarden, D. P. (1999). Business Information Visualization. Communications of AIS 1, 1-38.

Data/Information Visualization

- Exploiting the human visual system to extract information from data.
- Provides an overview of complex data sets.
- Identifies structure, patterns, trends, anomalies, and relationships in data.
- Assists in identifying the areas of interest.

Visualization = Graphing + Fitting + Graphing

for Data for Model

The Iris Data (Anderson 1935; Fisher 1936)

no.	Sepal Length	Sepal Width	Petal Length	Petal Width	Species
1	5.1	3.5	1.4	0.2	setosa
2	4.9	3.0	1.4	0.2	setosa
3	4.7	3.2	1.3	0.2	setosa
4	4.6	3.1	1.5	0.2	setosa
5	5.0	3.6	1.4	0.2	setosa
...					
76	6.6	3.0	4.4	1.4	versicolor
...					
150	5.9	3.0	5.1	1.8	virginica

Images source: http://www.stat.auckland.ac.nz/~ihaka/120/Lectures/lecture27.pdf

- The iris data published by Fisher (1936) have been widely used for examples in discriminant analysis and cluster analysis.
- The sepal length, sepal width, petal length, and petal width are measured in centimeters on fifty iris specimens from each of three species, *Iris setosa*, *I. versicolor*, and *I. virginica*.

Graphical Displays for Univarate Data

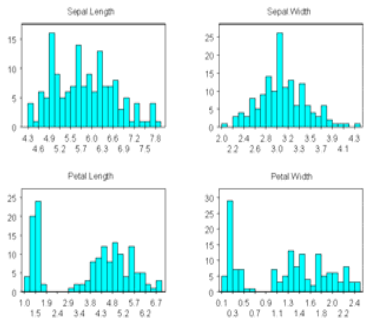
- Histogram
- Density Plot (Smoothed Histograms)
- Quantile Plots
- Box Plots

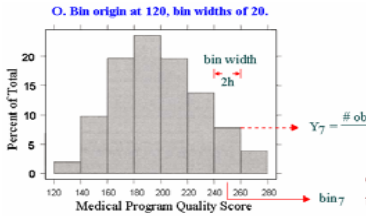
- ✓ Univariate graphs provide information about the distribution of observation on a single variable.
- ✓ A univariate graph is a model.
- ✓ The objective is to construct an abstraction that highlights the salient aspects of the data without distorting any feature or imposing undue assumptions.

Histogram

The histogram graphically shows:

1. center of the data (location)
2. spread of the data (scale)
3. skewness of the data
4. presence of outliers
5. presence of multiple modes in the data.





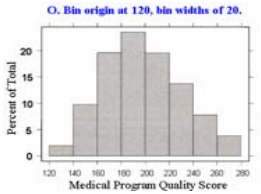
$$Y_7 = \frac{\#\text{ observations within bin}_7}{(2h)\,n}$$

bin$_7$

Changes in bin origin and bin widths affect the shape of the histogram

Histogram (conti.)

- 1/2h adjusts the height of each bar so that the total area enclosed by the entire histogram is 1.
- The area covered by each bar can be interpreted as the probability of an observation falling within that bar.



Disadvantage for displaying a variable's distribution:

- selection of origin of the bins.
- selection of bin widths.
- the very use of the bins is a distortion of information because any data variability within the bins cannot be displayed in the histogram. (no differentiation)

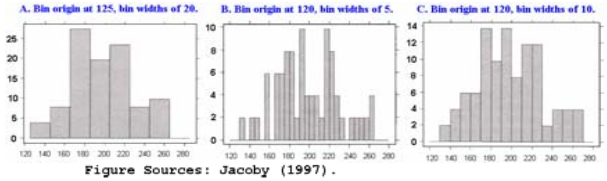
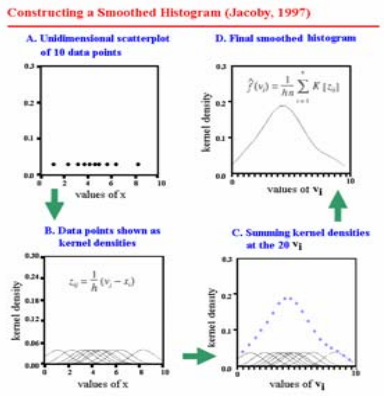


Figure Sources: Jacoby (1997).

Density plots (Smoothed Histograms)

- Smoothed histograms overcome some of the disadvantages caused by the arbitrary, discrete bins used in traditional histogram.
- The relative height of the smooth curve corresponds to the local density.
- The overall height is adjusted so that the total area under the curve is approximately equal to 1.
- The area under the curve between any two points along the horizontal scale can be interpreted as the probability that an observation falls within that interval of data values.

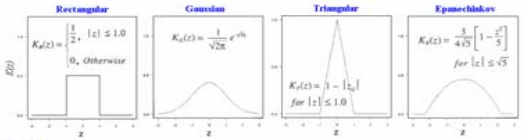


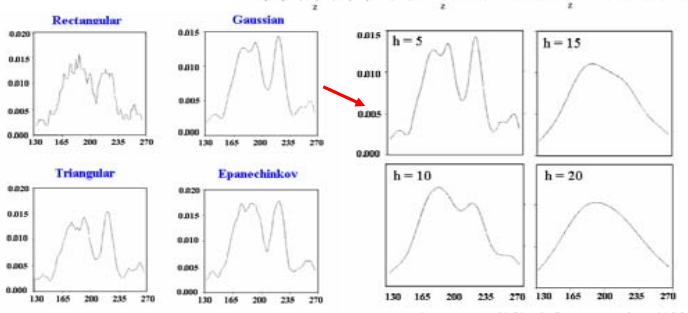
Density plots (conti.)

- Selection of kernels
- Selection of bandwidth

$$\hat{f}(v_i) = \frac{1}{hn}\sum_{i=1}^{n} K[z_{ij}]$$

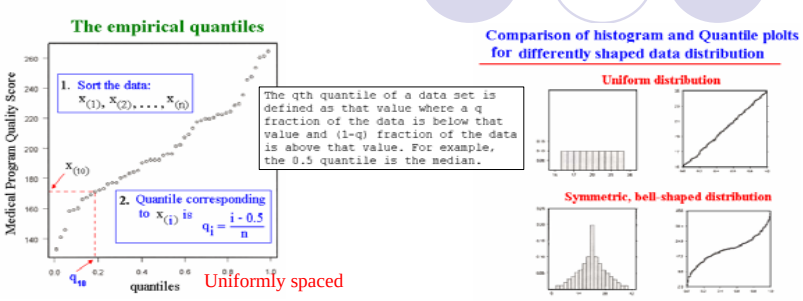
$$z_{ij} = \frac{1}{h}(v_j - x_i)$$





Figures modified from Jacoby (1997)

Quantile Plots



- 0.5 is subtracted from each i value to avoid extreme quantiles of exactly 0 or 1.
- The latter would cause problems if empirical quantiles were to be compared against quantiles derived from a theoretical asymptotic distribution such as the normal.
- This adjustment has no effect on the shape of any graphical display.

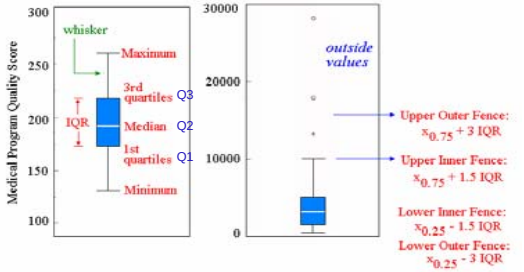
Figures modified from Jacoby (1997)

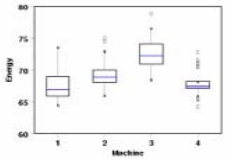
Box Plots

- Box plot (Tukey 1977, Chambers 1983) is an excellent tool for conveying location and variation information in data sets, particularly for detecting and illustrating location and variation changes between different groups of data.

The box plot can provide answers to the following questions:

- Is a factor significant?
  - Does the location differ between subgroups?
  - Does the variation differ between subgroups?
  - Are there any outliers?





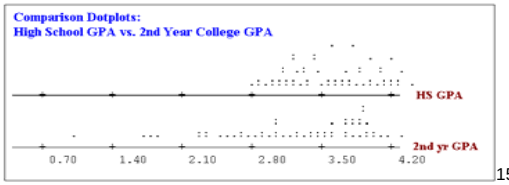
Further reading: http://www.itl.nist.gov/div898/handbook/eda/section3/boxplot.htm

Dotplot

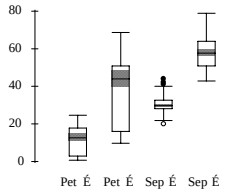
- A dotplot is an informal alternative to a histogram for displaying continuous data.
  - The sample values are plotted as dots along a horizontal axis.
  - The center and spread of dots can easily be identified along with outliers.
  - For moderately large samples, a dotplot diagram is also an effective method of examining the shape of the distribution.
- **Multidimensional dotplots** for more than one variable are also extremely useful for exploring multivariate data and are known as **scatterplots**.

Note:

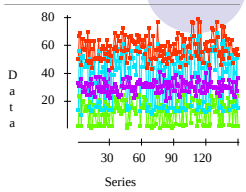
- The much longer left tail of 2nd year GPA which was apparent from the comparison boxplots shown earlier.
- The extremely low outlier for 2nd year GPA.



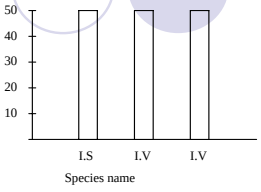
The Iris Data (Anderson 1935; Fisher 1936)



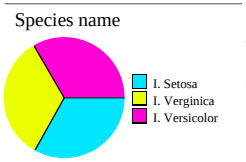
Box Plots



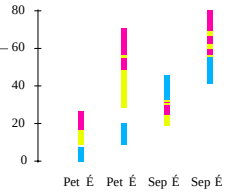
Line Plots



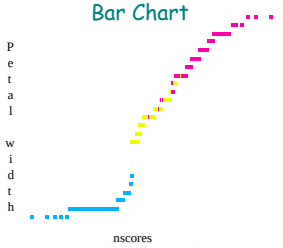
Bar Chart



Pie Chart



Dot Plots



Normal Prob. Plot

Stem and Leaf Plot

Consider the following data set, sorted in ascending order:

8, 13, 16, 25, 26, 29, 30, 32, 37, 38, 40, 41, 44, 47, 49, 51, 54, 55, 58, 61, 63, 67, 75, 78, 82, 86, 95

```
0 | 8
1 | 3 6
2 | 5 6 9
3 | 0 2 7 8
4 | 0 1 4 7 9
5 | 1 4 5 8
6 | 1 3 7
7 | 5 8
8 | 2 6
9 | 5
```

Stem and Leaf Plot Advantages

The stem and leaf plot essentially provides the same information as a histogram, with the following added benefits:

- The plot can be constructed quickly using pencil and paper.
- The values of each individual data point can be recovered from the plot.
- The data is arranged compactly since the stem is not repeated in multiple data points.

The stem and leaf plot offers information similar to that conveyed by a histogram, and easily can be constructed without a computer.

Graphical Displays for Bivariate Data

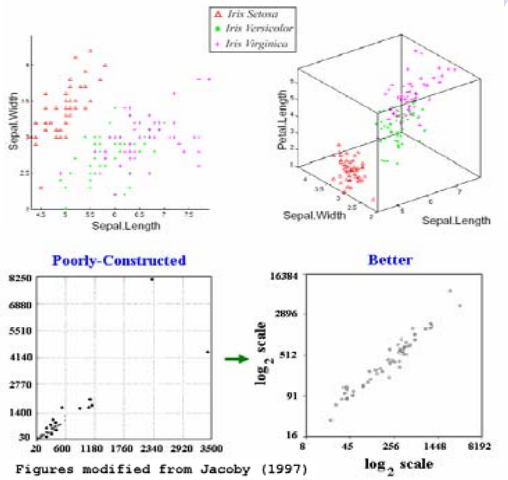
- 2D/3D Scatterplot
- Scatterplot Matrix
- Extensions of Scatterplots
- Scatter plot and MA plot
- Glyph Plot
- Volcano Plot
- Quantile-Quantile Plot (QQplot)

2D/3D Scatterplot

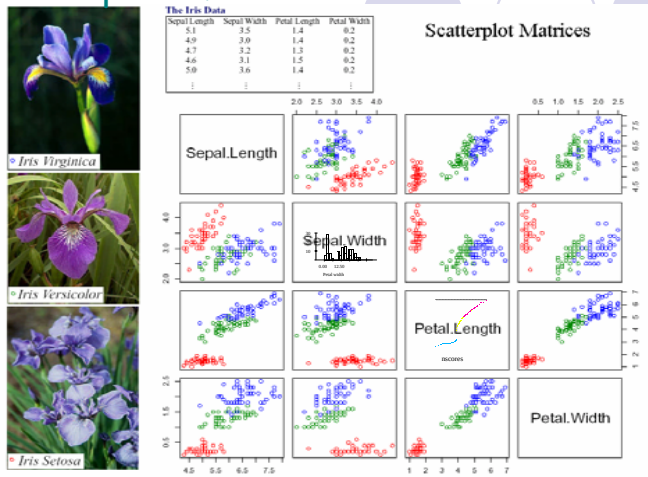
The scatterplot gives an idea of:

- Positive (negative) linear relationship
- Positive (negative) curved relationship
- Other relationships
- No relationship
- Visual evidence of outliers or suspicious observations.

NOTE:
Scatterplots are used only for quantitative variables (those that are comparable numerically).

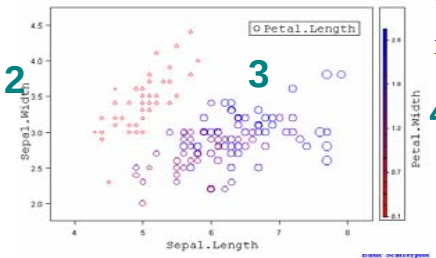


Scatterplot Matrix



Images source: http://www.stat.auckland.ac.nz/~ihaka/120/Lectures/lecture27.pdf

Extensions of Scatterplots

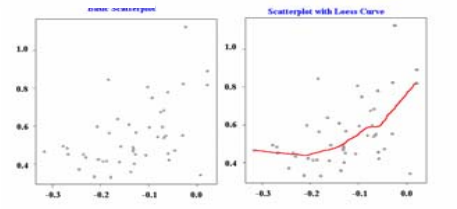


2 3 4

1

Bubble Color Scatterplot

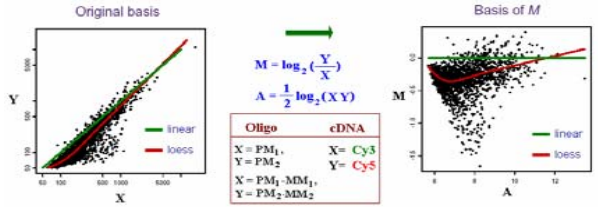
Scatterplot with a smoothing (e.g., loess) curve



Scatterplot for Gene Expression Data and MA plot

- **Features of scatter plot.**
  - the substantial correlation between the expression values in the two conditions being compared.
  - the preponderance of low-intensity values. (the majority of genes are expressed at only a low level, and relatively few genes are expressed at a high level)
- **Goals:** to identify genes that are differentially regulated between two experimental conditions.
- **Outliers in logarithm scale**
  - spreads the data from the lower left corner to a more centered distribution in which the prosperities of the data are easy to analyze.
  - easier to describe the fold regulation of genes using a log scale. In log2 space, the data points are symmetric about 0.

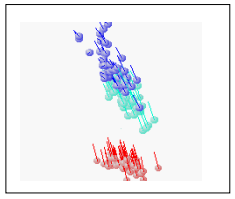
- **MA plots** can show the intensity-dependant ratio of raw microarray data.



Glyph Plot

- **Multiple-symbol**: usually one display panel simultaneously shows values of multiple variables that are represented by different shapes, sizes, colors, and locations of symbols (Tukey & Tukey, 1988).

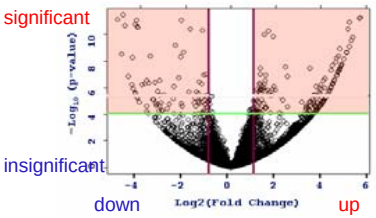
- A "**tail**" can also be added to each data point, in which the value of the fourth dimension is indicated by the angle and length of the tail.
- Since the data points represented by complex symbols are called "glyphs," this type of display is termed as a "glyph plot."



Glyph graph for Fisher Iris data

x: Petal length.
y: Sepal length.
z: Sepal width.
Angle: Petal width.
Color: three species of iris.

Volcano Plot



A volcano plot is a heuristic device that arranges genes along dimensions of biological and statistical significance.

A volcano plot is helpful in identifying significance and magnitude of change in expression of a set of genes between two conditions.

- A volcano plot displays the negative log of p-values from a t-test on one axis and the log2 of change between two conditions on the other axis on the scatterplot view.

- The researcher can then make judgments about the most promising candidates for follow-up studies, by trading off both these criteria by eye.

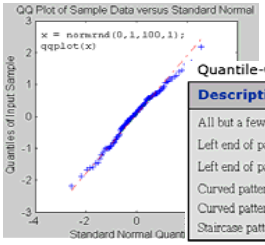
Quantile-Quantile Plot (QQplot)

Normal probability plot for graphical normality testing

- `qqplot(X)` displays a quantile-quantile plot of the sample quantiles of X versus theoretical quantiles from a normal distribution.
  If the distribution of X is normal, the plot will be close to linear.
- `qqplot(X,Y)` displays a quantile-quantile plot of two samples.
  If the samples do come from the same distribution, the plot will be linear.

```
The qth quantile of a data set is
defined as that value where a q
fraction of the data is below that
value and (1-q) fraction of the data
is above that value. For example,
the 0.5 quantile is the median.
```

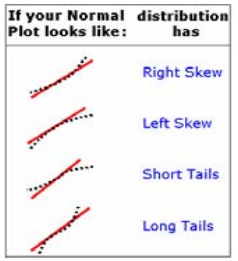
- If the quantiles of the theoretical and data distributions agree, the plotted points fall on or near the line $y = x$.
- If the theoretical and data distributions differ only in their location or scale, the points on the plot fall on or near the line $y = ax + b$. The slope a and intercept b are visual estimates of the scale and location parameters of the theoretical distribution.

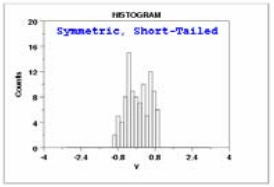


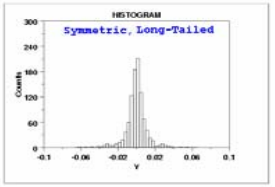
Quantile-Quantile Plot Diagnostics

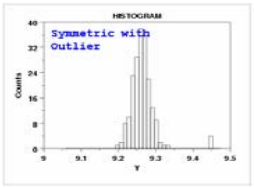
Description of Point Pattern	Possible Interpretation
All but a few points fall on a line	Outliers in the data
Left end of pattern is below the line; right end of pattern is above the line	Long tails at both ends of the data distribution
Left end of pattern is above the line; right end of pattern is below the line	Short tails at both ends of the data distribution
Curved pattern with slope increasing from left to right	Data distribution is skewed to the right
Curved pattern with slope decreasing from left to right	Data distribution is skewed to the left
Staircase pattern (plateaus and gaps)	Data have been rounded or are discrete

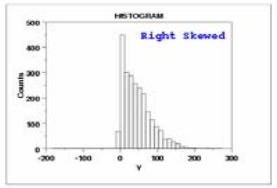
QQplot (conti.)

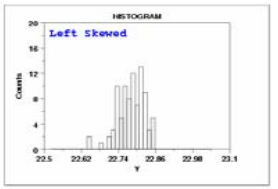












Graphical Displays for Multivariates Data

- Star Plot
- Radar Plot
- Chernoff Faces Plot
- Parallel Plot
- Andrews' Plot

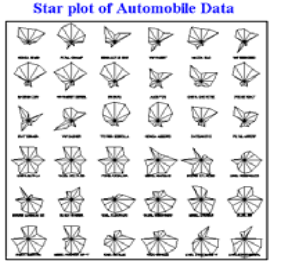
Star Plot

Variable assignment key for star plot

The variables at the sides and bottom are related to size;

The others relate to price and performance.

Star plot of Automobile Data



- The star plot (Chambers 1983) consists of a sequence of equi-angular spokes, called **radii**, with each spoke representing one of the variables.
  - The data length of a spoke is proportional to the magnitude of the variable for the data point relative to the maximum magnitude of the variable across all data points.
  - A line is drawn connecting the data values for each spoke.
  - This gives the plot a star-like appearance and the origin of the name of this plot.
- Typically, star plots are generated in a multi-plot format with many stars on each page and each star representing one observation.

The star plot can show:

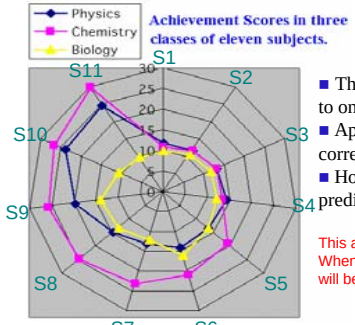
- What variables are dominant for a given observation?
- Which observations are most similar, i.e., are there clusters of observations?
- Are there outliers?

The primary weakness of star plot is that their effectiveness is limited to data sets with **less than a few hundred points**.

- Each star represents one car model.
- Each ray in the star is proportional to one variable.
- The dominant pattern is that the star symbols in the top rows have long rays on the top (good price and performance) and short rays on the bottom (small in size variables), but the reverse is generally true for the heaviest models in the bottom rows.

Radar Plot

- The idea of a radar plot is similar to that of the star plot.
- In a radar plot, the value of the measurement is also represented by radii stretching out from the center of a circle.
- However, here each radius stands for a subject instead of a variable.
- The subjects' response on each variable is displayed by points of different shapes, colors, or both.



Order of subjects?

- The graph shows the frequencies of data series relative to one another.
- Apparently, the biology scores are the lowest and are not correlated to neither physics nor chemistry scores.
- However, physics and chemistry scores are good predictors to each other.

This approach suffers the same shortcoming as the star graph. When there are too many variables and subjects, the data pattern will be concealed.

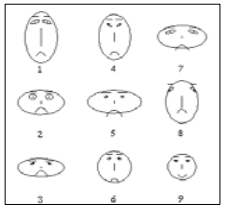
Source: ATKOSoft: Survey on Visualisation methods and software tools

Chernoff Faces Plot (Chernoff, 1973)

- Each facial feature denotes a particular variable.
  - For example, X1 can be associated with the size of the mouth,
  - X2 with the size of the nose,
  - X3 with the size of the eyes, and so on.
- The power of Chernoff face is its highly condensation of data and its interesting way of presentation.

A **major drawback** of Chernoff faces is that the subjective assignment of facial expressions to variables affects on the shape of the face.

- Chernoff and Rizvi (1975) found that the permutations of the assignment of features caused an error rate of as high as 25 for the task of classifying faces into groups.
- It means that classifying two faces as "fairly similar" is greatly influenced by the assignment of variables to specific features.
- Some researchers criticised that the symmetrical feature of Chernoff faces is redundant.
- Like the star graph, the power of showing multiple relationships in Chernoff faces are limited in a still mode.



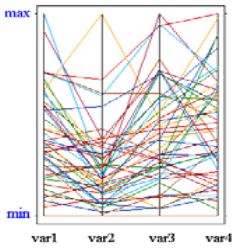
Source: ATKOSoft: Survey on Visualisation methods and software tools

Parallel Plot

- The Parallel Plot is a series of **parallel Dotplots** with lines linking the data values of each of the zones instead of individual symbols on each axis.
  - Variables to be visualized are represented by a series of parallel axes as replacement of the standard orthogonal approach.
  - Then, each row from the multidimensional dataset is represented as a series of unbroken line segments, which connect the parallel axes.

Parallel plot is useful to quickly identify interactions between variables:

- Clusters of observations with the similar lines across all axes.
- Direct relationship between a pair of variables appears in the plot as two axes connected by a series of parallel lines.
- Inverse relationship between two variables should be displayed as a series of lines, which cross each other.



Alfred Inselberg

1959

now

http://www.math.tau.ac.il/~aiisreal/

Andrews' Plot

- Data $\{x_{ij}: i=1,\ldots,n; j=1,\ldots,p\}$, (vector observations in p-dimensions so x_{ij} is the j^{th} element of the i^{th} observation).

$$f_{x_i}(t) = \frac{1}{\sqrt{2}}x_{i1} + x_{i2}\sin t + x_{i3}\cos t + x_{i4}\sin 2t + x_{i5}\cos 2t + \ldots + x_{ip}\begin{matrix}\sin \\ \cos\end{matrix}\left(\left[\frac{p}{2}\right]t\right)$$

Order of variables?

This maps p-dimensional data $\{x_i\}$ onto 1-dimensional $\{f_{x_i}(t)\}$ for any t.

If we plot $f_{x_i}(t)$ over $-\pi < t < \pi$ we obtain a 1-dimensional representation of the data

Properties:

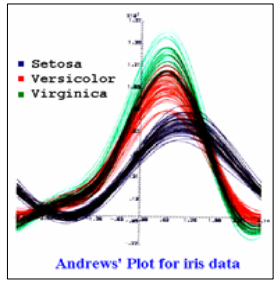
(i) preserves means,

$$f_{\bar{x}}(t) = \frac{1}{n}\sum_{i=1}^{n} f_{x_i}(t)$$

(ii) preserves distances;

$$\left\|f_{x_1}(t) - f_{x_2}(t)\right\|^2 = \int_{-\pi}^{\pi}(f_{x_1}(t) - f_{x_2}(t))^2 dt = \pi\sum_{j=1}^{p}(x_{1j} - x_{2j})^2$$

(iii) yields 1-dimensional views of the data: at $t=t_0$ we obtain the projection of the data onto the vector $f_1(t_0)=(1/\sqrt{2}, \sin t_0, \cos t_0, \sin 2t_0, \ldots.)'$



Andrews' Plot for iris data

Visualizing High-dimensional Data: dimension reduction techniques

- Biplot
- Principal Component Analysis (PCA)
- Sliced Inverse Regression (SIR)
- Multidimensional Scaling (MDS)
- Self-Organizing Maps (SOM)
- Unified Matrix Method

✓ Dimension reduction visualization is often adopted for presenting grouping structure for methods such as K-means.

The Biplot (Gabriel 1971, 1981; Gower & Hand, 1996)

The data matrix can be factored:

$$\mathbf{X} = \mathbf{AB}'$$

$\mathbf{X}_{n\times p}$: data matrix.
$\mathbf{A}_{n\times k}$: the coordinates for the n observations points along k rectangular axes.
$\mathbf{B}_{p\times k}$: the coordinates for the p variables along the same k axes.

To obtain $\mathbf{A}$ and $\mathbf{B}$, using Singular Value Decomposition (SVD)

$$\mathbf{X} = \mathbf{UDV}'$$

$\mathbf{A}_{[2]}$: the $n \times 2$ matrix of biplot coordinates for the observation points.
$\mathbf{B}_{[2]}$: the $p \times 2$ matrix of biplot coordinates for the variables.

$$\mathbf{A}_{[2]} = \mathbf{U}_{[2]}\mathbf{D}^c_{[2]}$$

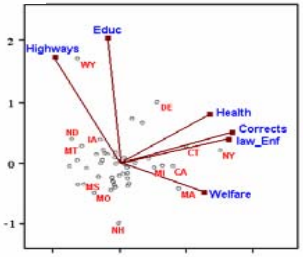
$$\mathbf{B}_{[2]} = \mathbf{V}_{[2]}\mathbf{D}^{1-c}_{[2]}$$

$\mathbf{U}_{[2]}$: the first two columns of $\mathbf{U}$.
$\mathbf{V}_{[2]}$: the first two columns of $\mathbf{V}$.
$\mathbf{D}_{[2]}$: the diagonal matrix formed by the first two singular values.

$$\mathbf{X}_{[2]} = \mathbf{A}_{[2]}\mathbf{B}'_{[2]}$$

Each row of $\mathbf{A}_{[2]}$ is plotted as a point in a two-axis coordinate system.
The rows of $\mathbf{B}_{[2]}$ are also plotted within the same space.
Goodness of fit measure R (s_r : singular values)

$$R = \frac{s_1^2 + s_2^2}{\sum_{r=1}^{p} s_r^2}$$

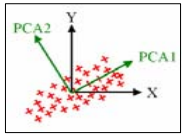


Biplot of 1992 State Policy Spending

0≤c≤1
c: weight for observation
1-c: weight for variables

Principal Component Analysis (PCA)

(Pearson 1901; Hotelling 1933; Jolliffe 2002)



The ith principal component of $\mathbf{X}$ is $\mathbf{X}'\mathbf{v}_i$, where $\mathbf{v}_i$ is the ith normalized eigenvector of $\Sigma_{\mathbf{x}}$ corresponding to the ith largest eigenvaules.

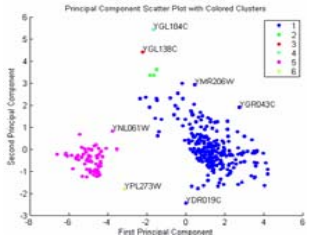
The PCA summaries the dispersion of data points as data cloud in a small number of major axes (principal components) of variation among the variables.

Goal: to reduce the dimensionality of the data matrix by finding the new variables (linear combinations of original variables).

Cumulative Sum of the Variances:

1	78.3719
2	89.2140
3	93.4357
4	96.0831
5	98.3283
6	99.3203
7	100.0000

This shows that almost 90% of the variance is accounted for by the first two principal components.



Yeast Microarray Data is from
DeRisi, JL, Iyer, VR, and Brown, PO.(1997). "Exploring the metabolic and genetic control of gene expression on a genomic scale"; Science, Oct 24;278(5338):680-6.

Sliced Inverse Regression (SIR)

- Li (1991) introduced the following model

$$y = f(\beta_1'\mathbf{x}, \cdots, \beta_K'\mathbf{x}, \epsilon).$$

- The β's are referred to an effective dimension-reduction (*e.d.r.*) direction.
- The estimated β's can be obtained by
  (1) first standardizing $\mathbf{x}$ and
  (2) then conduct a **weighted principal component analysis** for the weighted covariance matrix $\hat{V} = \sum_{h=1}^{H} \hat{p}_h \hat{m}_h \hat{m}_h'$, where H is the number of slices, $\hat{p}_h$ is the proportion of the y_i that falls in slice h, and $\hat{m}_h$ is the standardized sample mean in the slice h.

y	X_1	X_2	X_2
10	1	4	2
13	3	5	3
9	0	3	2
17	5	5	6
12	2	4	3
11	2	5	2

⇒

y	X_1	X_2	X_2
9	0	3	2
10	1	4	2
11	2	5	2
12	2	4	3
13	3	5	3
17	5	5	6

⇒

$\bar{x}_1$	0.5	3.5	2.0
$\bar{x}_2$	2.0	4.5	2.5
$\bar{x}_3$	4.0	5.0	4.5

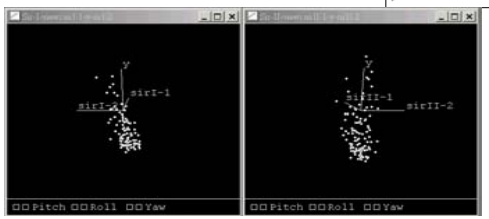
$\Rightarrow \Sigma_W \beta_i = \lambda_i \Sigma_X \beta_i$

Li, K. C. (1991). Sliced inverse regression for dimensional reduction (with discussion). *JASA* **86**, 316-342.

SIR (conti.)

- IDEA: under regular conditions, the centered inverse regression curve $E[\mathbf{x}|y] - E[\mathbf{x}]$ is contained in the linear subspace spanned by $\beta_k \Sigma_{XX}$ $(k = 1, \cdots, K)$, where Σ_{XX} denotes the covariance matrix of $\mathbf{x}$.
- CONDITION 3.1 (Li, 1991)
  For any b in R^p, the conditional expectation $E(b'\mathbf{x}|\beta_1'\mathbf{x}, \cdots, \beta_K'\mathbf{x})$ is linear in $\beta_1'\mathbf{x}, \cdots, \beta_K'\mathbf{x}$; that is, for some constants $c_0, c_1, \cdots, c_k$, $E(b'\mathbf{x}|\beta_1'\mathbf{x}, \cdots, \beta_K'\mathbf{x}) = c_0 + c_1\beta_1'\mathbf{x} + \cdots + c_k\beta_K'\mathbf{x}$.

```
(def x (g-normal 5 100))
(def err (normal-rand 100))
(def y (+ 5 (nth 0 x) (nth 1 x) (nth 2 x) err))
(def y-trans (** y 2))
(SIR-I-II-MODEL x y-trans)
```



Graphical Regression

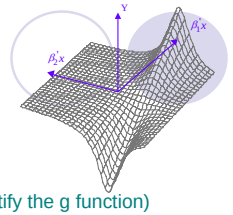
- **SIR II**
- **SAVE**
- **pHd**

see homework!

Sliced Inverse Regression (SIR)

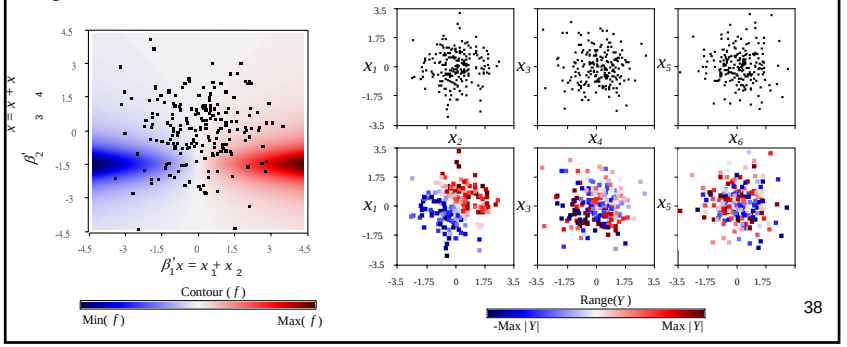
$\mathbf{x} = (x_1, x_2, x_3, x_4, x_5, x_6)$ (independent normal)

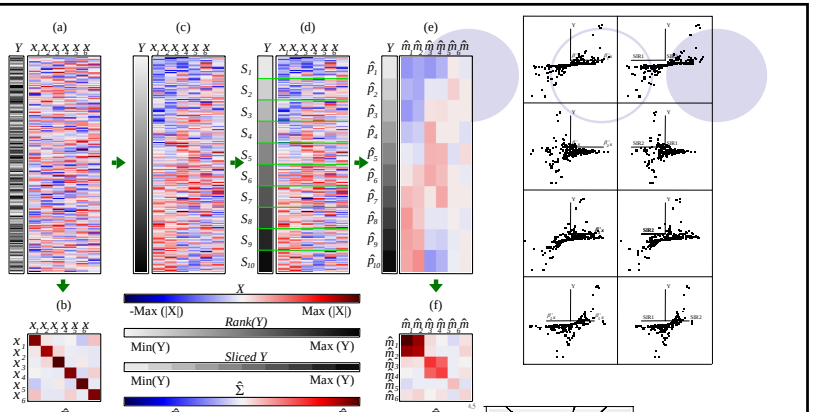
$$y = g(\beta_1'\mathbf{x}, \beta_2'\mathbf{x}, \varepsilon) = \frac{\beta_1'\mathbf{x}}{0.5 + (\beta_2'\mathbf{x} + 1.5)^2} + 0 \cdot \varepsilon$$



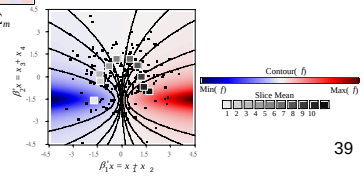
Given data (y, x), want to identify

$\beta_1 = (1,1,0,0,0,0)$ $\beta_2 = (0,0,1,1,0,0)$ (not to identify the g function)



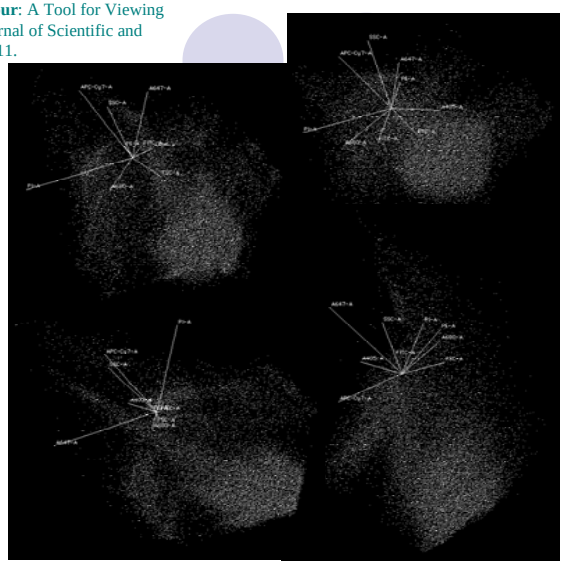


first vector	(-0.72 -0.68 -0.01 -0.01 0.07 0.01)
S.D.	(0.04 0.04 0.04 0.04 0.04 0.04)
ratio	(-17 .3 -16 .1 -0.2 -0.1 1.9 0.3)
second vector	(-0.01 0.08 -0.67 -0.72 0.05 -0.02)
S.D.	(0.09 0.10 0.09 0.09 0.09 0.09)
ratio	(-0.2 0.9 -7.6 -7.7 0.5 -0.2)
eigenvalues	(0.7 6 0.38 0.06 0.03 0.02 0.02)
p-values	(0.0 3.8E-7 0.62 NA NA NA)



Asimov, D. (1985). The **Grand Tour**: A Tool for Viewing Multidimensional Data. SIAM Journal of Scientific and Statistical Computing 6(1), 128 -- 11.

The grand tour of a multidimensional data set is a visualization technique for examining structure of high dimensional data using dynamic graphics. The idea, introduced in Asimov (1985) and Buja and Asimov (1986), is to capture, in some sense, the popular meaning of grand tour, that is, to look at a subject from all possible angles. In a data analytic setting, these authors propose projecting a d-dimensional data set into a dense set of the possible two-dimensional planes.



Multidimensional Scaling (MDS)

(Torgerson 1952; Cox and Cox 2001)

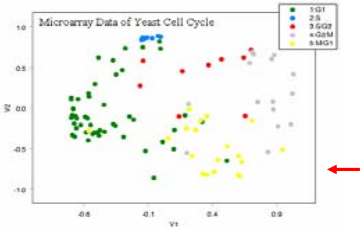
Classical MDS

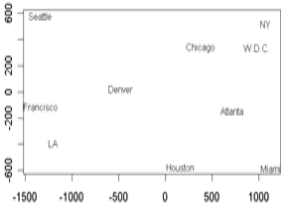
MDS takes a set of dissimilarities and returns a set of points such that the distances between the points are approximately equal to the dissimilarities.

Note that if the input-space distances are Euclidean, classical MDS is equivalent to PCA. (Mardia et al. 1979)

Analysis of Flying Mileages Between Ten U.S. Cities

0										Atlanta
587	0									Chicago
1212	920	0								Denver
701	940	879	0							Houston
1936	1745	831	1374	0						Los Angeles
604	1188	1726	968	2339	0					Miami
748	713	1631	1420	2451	1092	0				New York
2139	1858	949	1645	347	2594	2571	0			San Francisco
2182	1737	1021	1891	959	2734	2408	678	0		Seattle
543	597	1494	1220	2300	923	205	2442	2329	0	Washington D.C.



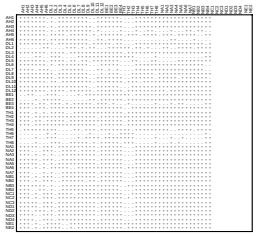


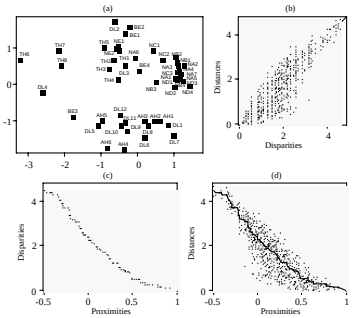
2D MDS configuration plot for 103 known genes

Conventional Multidimensional Scaling

input proximity matrix (of 50 conditions/genes/…)

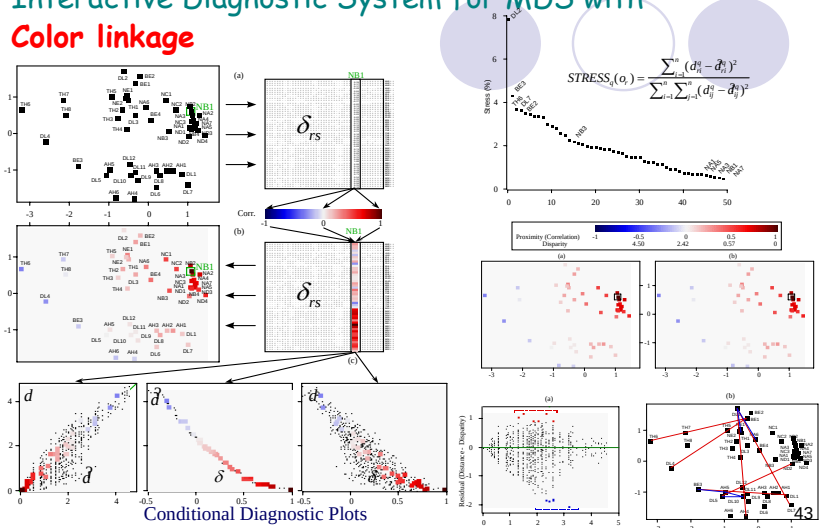
output configuration plots with overall diagnostic plots and residuals





$$STRESS_q(o_r) = \frac{\sum_{i=1}^{n}(d_{ri}^q - \hat{d}_{ri}^q)^2}{\sum_{i=1}^{n}\sum_{j=1}^{n}(d_{ij}^q - \hat{d}_{ij}^q)^2} = 20\%$$

Interactive Diagnostic System for MDS with Color linkage

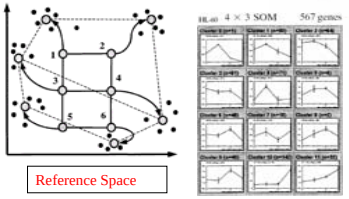


Conditional Diagnostic Plots

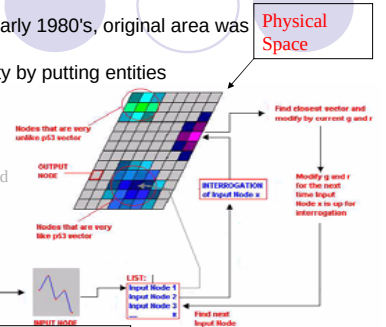
Self-Organizing Maps (SOM)

- SOMs were developed by Kohonen in the early 1980's, original area was in the area of speech recognition.
- ***Idea:*** Organise data on the basis of similarity by putting entities geometrically close to each other.
- SOM is unique in the sense that it combines both aspects. It can be used at the same time both to reduce the amount of data by clustering, and to construct a nonlinear projection of the data onto a low-dimensional display.

Figure modified from: *SCI*path



Reference Space



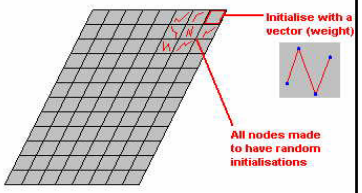
Reference Space

Macrophage Differentiation in HL-60 cells

Tamayo, P. et al. (1999). Interpreting patterns of gene expression with self-organizing maps: Methods and application to hematopoietic differentiation. *Proc Natl Acad Sci* 96:2907-2912.

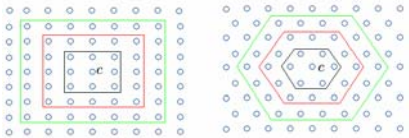
SOM - Initialization

Step 0: Initialize weights $\mathbf{w}_i(t)$.
Set topological neighborhood parameters $N_c(t)$.
Set learning rate parameters $\alpha(t)$ and $h_{ci}(t)$.



SOM initialization means to give each weight of the output node a random (or determined) vector value. *The dimensionality of the vector values put in* ***must match*** *the dimensionality of the raw data!* So if the raw data consists of 5 arrays, then the vectors must have 5 elements (dimensions).

Two examples of topological neighborhood.



$N_c(t_1) = 1, \quad N_c(t_2) = 2, \quad N_c(t_3) = 3, \quad t_1 < t_2 < t_3$

SOM Algorithm

Step 1: For each input vector $\mathbf{x}(t)$, do

BMU: Best matching Unit

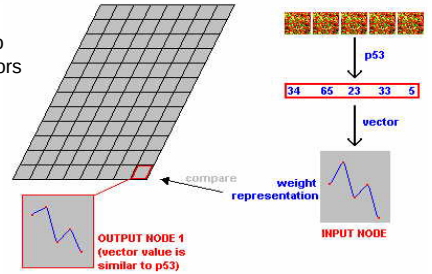
a. Finding a BMU: $\|\mathbf{x}(t) - \mathbf{w}_c(t)\| = \min_i \|\mathbf{x}(t) - \mathbf{w}_i(t)\|$

b. Learning process:

$$\mathbf{w}_i(t+1) = \begin{cases} \mathbf{w}_i(t) + h_{ci}(t)[\,\mathbf{x}(t) - \mathbf{w}_i(t)\,], & i \in N_c(t) \\ \mathbf{w}_i(t), & \text{o.w.} \end{cases}$$

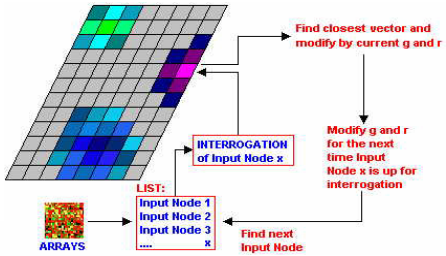
c. Go to the next unvisited input vector. If there are no unvisited input vector left then go back to the very first one and go to Step 2.

The SOM algorithm then goes on to interrogate the map for similar vectors



Figures source from: *SC*/path Home
http://www.ucl.ac.uk/oncology/MicroCore/tutorial.htm

Summary of SOM



Figures source from: *SC*/path Home
http://www.ucl.ac.uk/oncology/MicroCore/tutorial.htm

Step 0: Initialize weights $\mathbf{w}_i(t)$.
Set topological neighborhood parameters $N_c(t)$.
Set learning rate parameters $\alpha(t)$ and $h_{ci}(t)$.

Step 1: For each input vector $\mathbf{x}(t)$, do

a. Finding a BMU: $\|\mathbf{x}(t) - \mathbf{w}_c(t)\| = \min_i \|\mathbf{x}(t) - \mathbf{w}_i(t)\|$

b. Learning process:

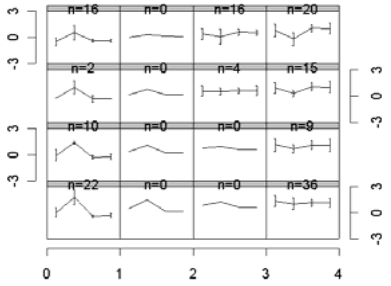
$$\mathbf{w}_i(t+1) = \begin{cases} \mathbf{w}_i(t) + h_{ci}(t)[\,\mathbf{x}(t) - \mathbf{w}_i(t)\,], & i \in N_c(t) \\ \mathbf{w}_i(t), & \text{o.w.} \end{cases}$$

c. Go to the next unvisited input vector. If there are no unvisited input vector left then go back to the very first one and go to Step 2.

Step 2: Incrementally decrease the learning rate and the neighborhood size, and repeat Step 1.

Step 3: Keep doing Steps 1 and 2 for a sufficient number of iterations.

SOM: iris example



Software: R: The som Package
http://cran.r-project.org/src/contrib/som_0.2-7.tar.gz

```
> iris.data.n <- normalize(iris.data, byrow=F)
> iris.som <- som(iris.data.n, xdim=4, ydim=4, topol="rect", neigh="gaussian")
> plot.som(iris.som)
```

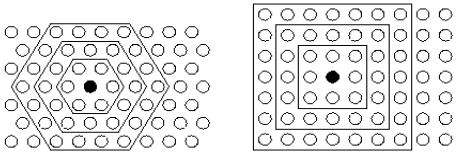
Self Organizing Map (SOM)

SOM(Self-Organizing Maps)是由 T. Kohonen 於 1980 年所提出。SOM 的基本原理源於大腦結構的特性，因爲大腦具有相同功能的腦細胞會聚集在一起的特性，例如：大腦中有專司味覺、視覺等的區塊。SOM 就是模擬這種個性，當 learning process 完畢後，其輸出單元相鄰近者會具有像似的連結加權值。

The SOM algorithm

SOM 是將高維的 input data space $\Re^m$ 映射到一個低維的 array of node，我們稱這低維的 space 爲 physical space，而稱高維的 $\Re^m$ 爲 weight space。對於 physical space 中的每一個 node i，都連接著一個 m 維的 reference vector $w_i = [w_{i1}, \ldots, w_{im}]^T \in \Re^m$。

在 physical space 裡形成格子狀的 array 可以爲矩形、六角形甚至不規則形，如圖一。而在 physical space 裡的所有 node 會與相鄰的 nodes 連結形成 *neighborhood*。



(a) Hexagonal grid (b) Rectangular grid

學習過程：

1. 計算 input vector 與各 reference vector 的距離:
   每次從 input data 載入一個 input vector $x \in \Re^m$，計算與每一個 reference vector w_i 的距離。
2. 找出優勝單元:
   距離最短的 reference vector w_c 稱作優勝單元(winner or best-matching node)。
   $$\|x - w_c\| = \min_i \{\|x - w_i\|\}$$
3. 調整 reference vector:
   $$w_i(t+1) = w_i(t) + h_{ci}(t)[x(t) - w_i(t)], \qquad i \in N_c(t)$$
   $$w_i(t+1) = w_i(t), \qquad i \notin N_c(t)$$
   $$h_{ci} = \alpha(t) \cdot \exp\left(-\frac{\|r_c - r_i\|^2}{2\sigma^2(t)}\right), \qquad 0 < \alpha(t) < 1, \qquad r_c, r_i \in \text{physical space}$$
   其中 $h_{ci}(t)$稱爲 *neighborhood function*，在此我們採用常用的 Gaussian function。而 $0 < \alpha(t) < 1$ 爲 learning-rate factor，且 $\sigma(t)$爲 width of the kernel 相當於 radius of $N_c(t)$。 $\alpha(t)$和 $\sigma(t)$對 t 來講都爲絕對遞減函數。

   並不是只有優勝單元 w_c 需要修正，而是 neighborhood $N_c(t)$裡的所有 reference vector 都需要修正。$N_c(t)$裡的 reference vector 離 w_c 越遠，則使得 $h_{ci}(t)$越小，也就使得 reference vector 的修正值也越小。
4. 對所有 input data 重複步驟 1 到 3 稱爲一個學習循環，每執行一個學習循環，將 radius of $N_c(t)$縮小一次，縮小至 0 就不再縮小下去。
5. 重複步驟 1 到 4 直到收斂，或執行一定次數的學習循環，即完成學習過程。

Possible parameters used in SOM analysis:

1. grid dimension: 1D, 2D, 3D, …
2. grid shape: in 2D –> Hexagon, Rectangle, …
3. # of node: in 2D_Rect -> 4x6, 5x5, 3x8, …
4. kernel shape: Gaussian, Biweightr, Epanechnikov, Triangular, Unif.
5. kernel width.
6. learning rate α(t)
7. neighborhood size: radius of N_c(t)
8. initial locations of reference vectors: random, use input vector.
9. order of input vectors: X_i
10. ways of learning. # of iteration.

Reference

[1] Kohonen, T., Self-Organizating Maps, New York : Springer-Verlag, 1997.
[2] Freeman, James A and Skapura, David M. Neural networks: algorithms, application, and programming techniques. 1992.
[3] http://www.cis.hut.fi/projects/somtoolbox/documentation/somalg.shtml
[4] http://davis.wpi.edu/~matt/courses/soms/
[5] http://www.cis.hut.fi/~tho/thesis/
[6] http://www.dcs.napier.ac.uk/hci/martin/msc/node16.html

Visualizing High-dimensional Data: dimension-free visualization

- Dendrogram and HeatMap
- Generalized Association Plots (GAP)

✓ Dimension free data visualization can explore the overall data structure without any dimension reducing procedure.
✓ The computing power, memory size, and display ability of modern computers have provided dimension free visualization a brand new platform.

Dendrogram (Kaufman and Rousseeuw, 1990)

- **Hierarchical Clustering**

  Example: Agglomerative algorithm + Average linkage clustering

	a	b	c	d	e
a	0	2	6	10	9
b		0	5	9	8
c			0	4	5
d				0	3
e					0

$D(\{a,b\},\{c\}) = \frac{1}{2}[D(a,c)+D(b,c)] = \frac{1}{2}(6+5) = 5.5$

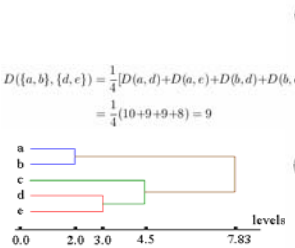
	{a, b}	c	d	e
{a, b}	0	5.5	9.5	8.5
c		0	4	5
d			0	3
e				0

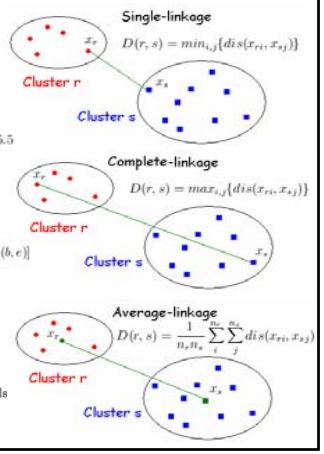
$D(\{a,b\},\{d,e\}) = \frac{1}{4}[D(a,d)+D(a,e)+D(b,d)+D(b,e)]$

$= \frac{1}{4}(10+9+9+8) = 9$

	{a,b}	c	{d, e}
{a, b}	0	5.5	9.0
c		0	4.5
{d, e}			0

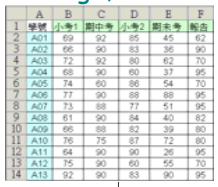
	{a, b}	{c, d, e}
{a, b}	0	7.83
{c, d, e}		0



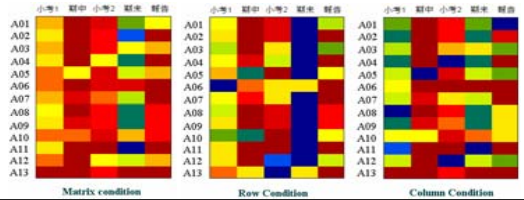


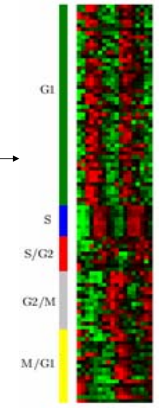
Heat Map
(Data Image, Matrix Visualization)

Microarray Data of Yeast Cell C)



- Synchronized by alpha factor arrest method (Spellman et al. 1998; Chu et al. 1998)
- 103 known genes: every 7 minutes and totally 18 time points.





Lab 309 for Information Visualization


Kao, Chiun-How


Chang, David


Lin, Chien-Ru


Ouyoung, Chih-Wen


Tzeng, Sheng Li


Ho,Meng-Ru


Wu, Yi-Chen


Dr. Wu, Han-Ming


Chung, Oliver


Tien, Yin-Jing

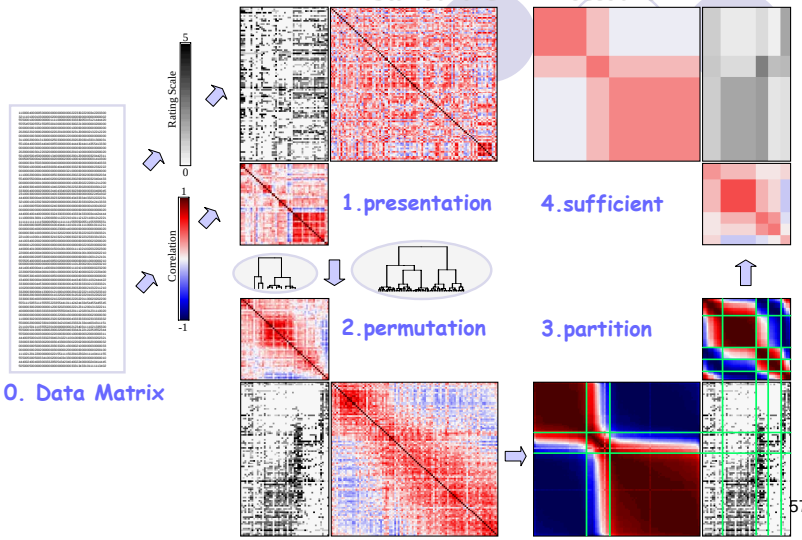

Yeh, Tzu-Chun

We generalize the concept of MV into :

1. presentation of raw data matrix & computation/presentation of proximity matrices
2. permutation of proximity & raw data matrices
3. partitioning of proximity & raw data matrices
4. sufficient statistical graph

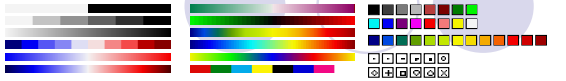
Chun-Houh Chen
"Generalized association plots (GAP): information visualization via iteratively formed correlation matrices,"
***Statistica Sinica* 12 (2002), 7-29**

A Standard GAP Procedure

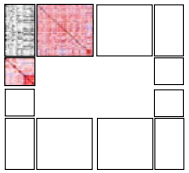


57

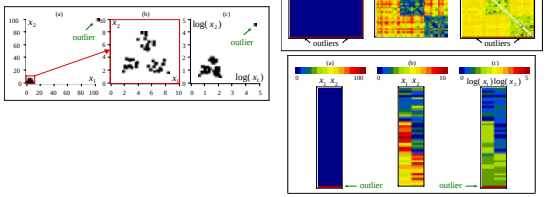
1. Selection of suitable color spectrum



1.presentation



2. Transformation/Standardization of data? ("Resolution" of a Statistical Graph)



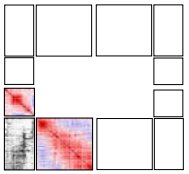
3. Selection of proximities for variable/sample correlation/covariance/distance/ . . .

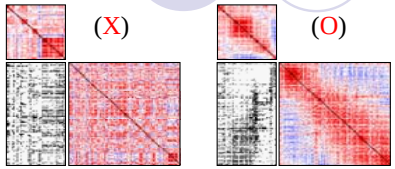
58

1. Relativity of a statistical graph (Chen 2002)

Concept of placing similar (different) objects at closer (distant) positions

2.permutation

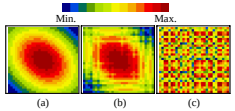




(Friendly & Kwan (2003) used effect-ordered data display, for ordering information in general visual displays.)

2. Need evaluation criteria

* Global Structure => Robinson Form



Robinson (a)
pre-Robinson (b & c)

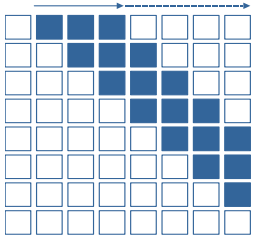
* Local behavior => Minimum Span Loss

59

Combination of Anti-Robinson Measure & Minimal Span Loss

(Local) w = 1 2 3 • • • n-1 (Global)

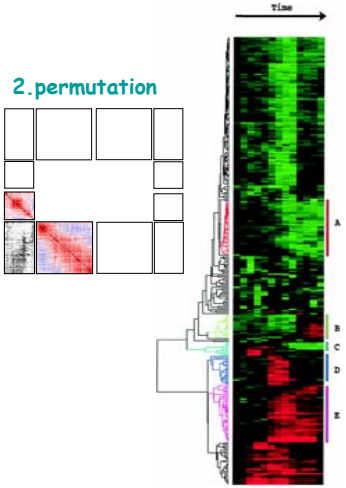
Open a band along the diagonal with width (0 <w < n) and compute the anti-Robinson measurements within that band



60

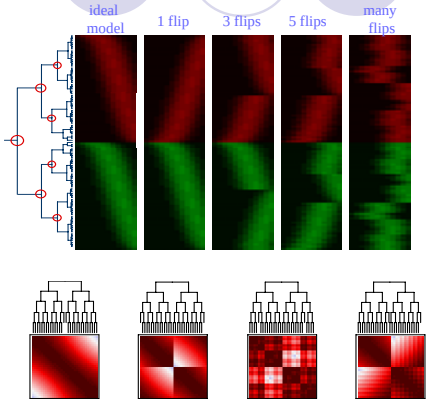
3. Tree seriation & flipping of intermediate nodes

2.permutation



Eisen et al. (1998)

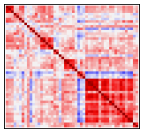
Different Seriations (Ordering of Terminal Nodes or Leaves) Generated from Identical Tree Structure



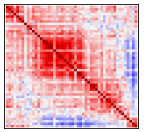
=> external and internal references

Converging sequence of iteratively formed correlation matrices

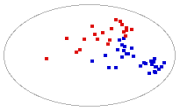
correlation map (original order)



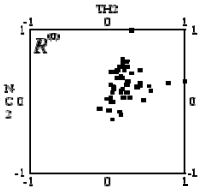
correlation map (ellipse order)



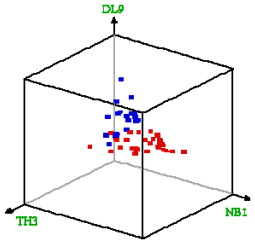
first 2 eigenvectors



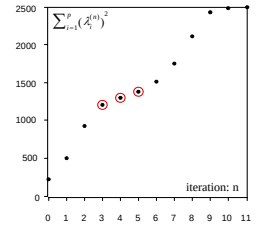
2D projection



3D projection

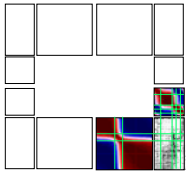


Sum squared eigenvalues (sum squared correlations)

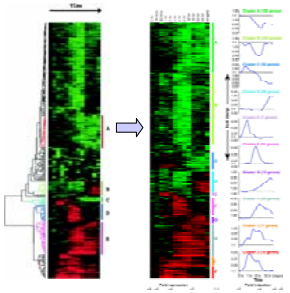


Selection of suitable partitioning (clustering) algorithm

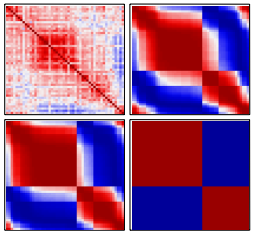
3.partition



1. Dendrogram of a clustering tree

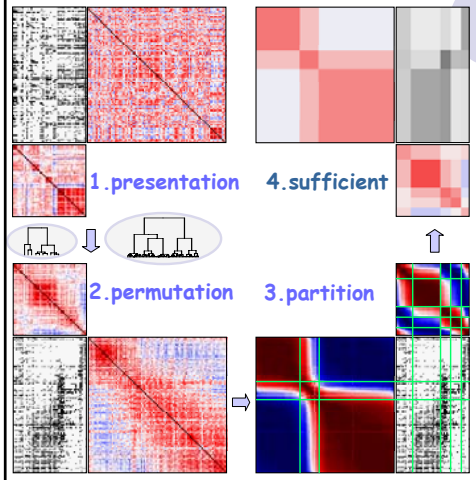


2. GAP converging process



3. K-means clustering

4. Sufficient Statistical Graph (Chen 2002)

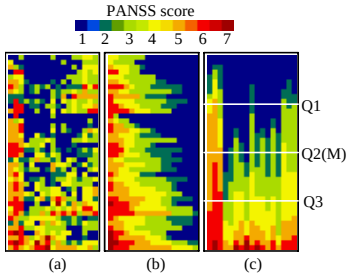


to **summarize** information embedded in the raw data matrix and proximity matrices with a **simplified version of MV.**

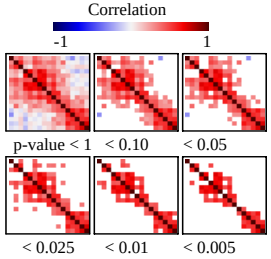
Summary statistic: mean/median/variance . . .

Generalization and flexibility of MV

1. Sediment MV

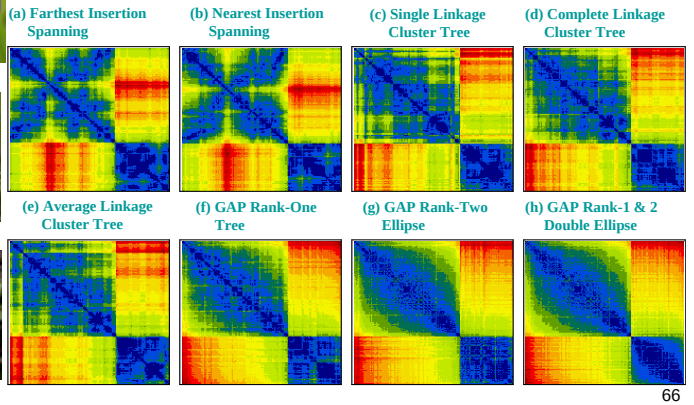


2. Sectional MV



Local versus Global Criteria.

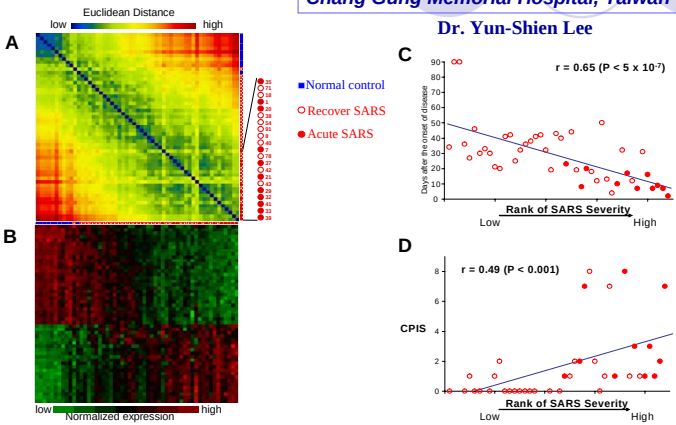
Comparison of Seriation Results of 8 Sorting Techniques on the Euclidean Distance Matrix of 150 Iris Flowers (Fisher 1936)



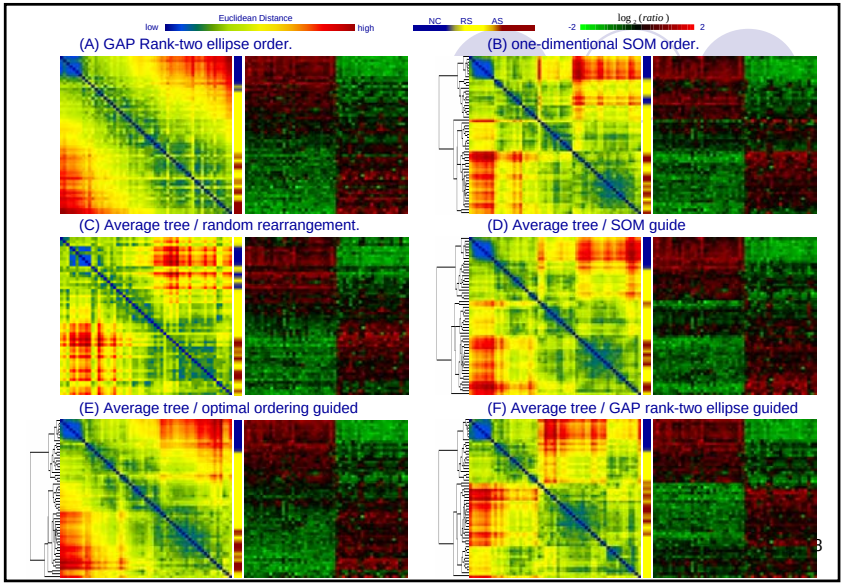
Human Response to SARS-CoV by Enhancing Innate Immunity and Depressing Adaptive Immunity as Shown by Gene Expression Profiles in Peripheral Blood

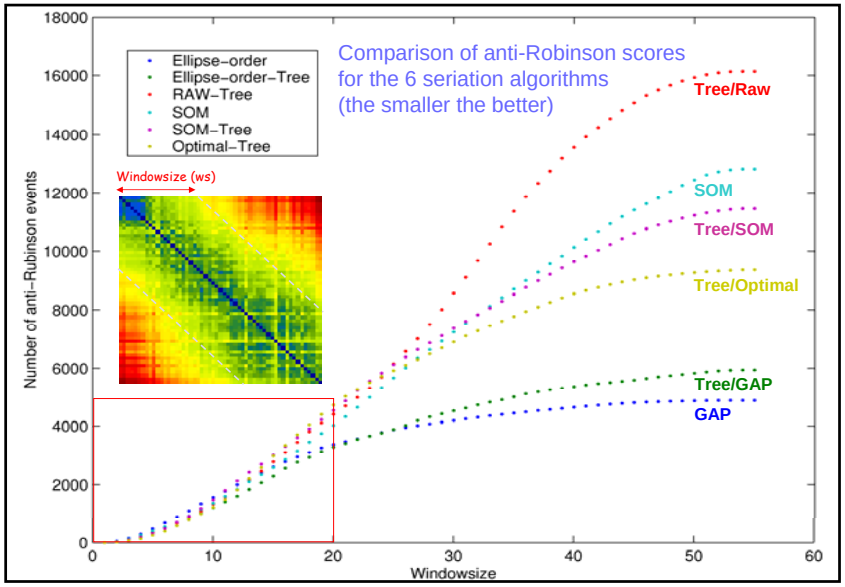
Chang Gung Memorial Hospital, Taiwan

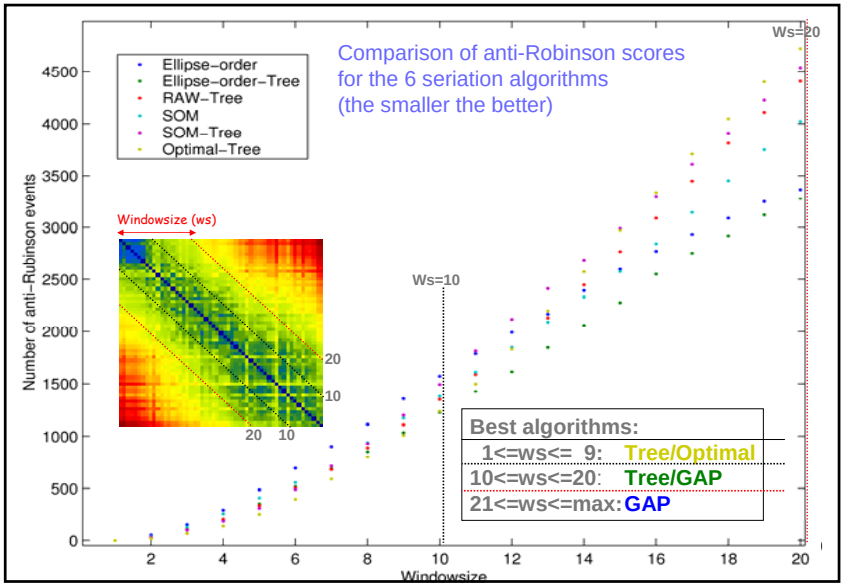
Dr. Yun-Shien Lee

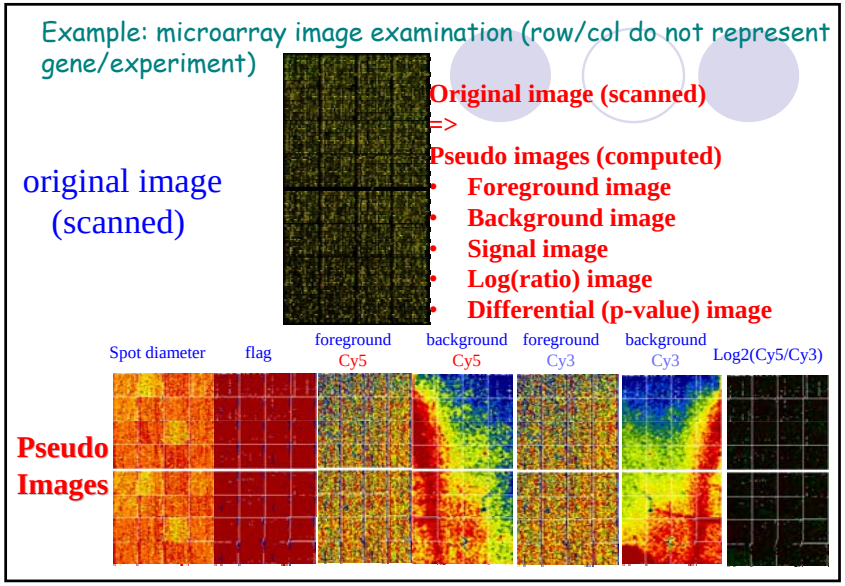


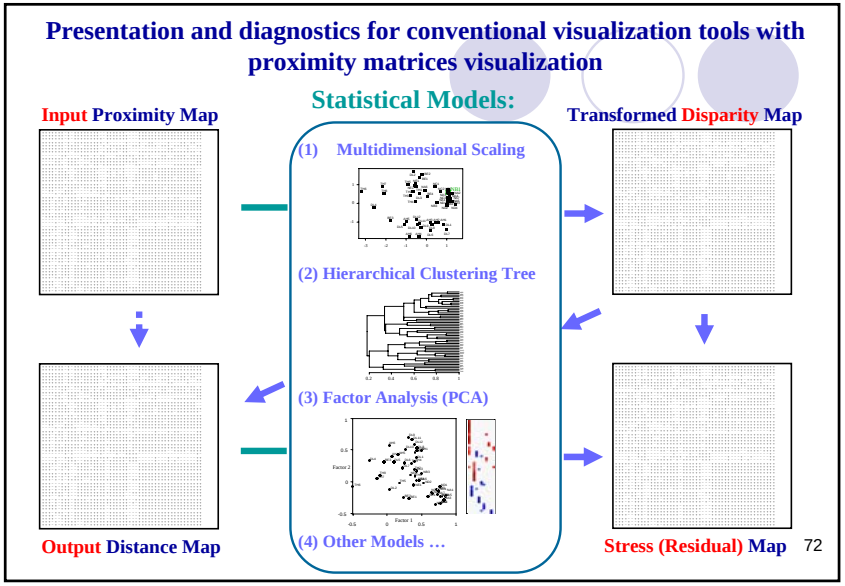
Severe Acute Respiratory Syndrome



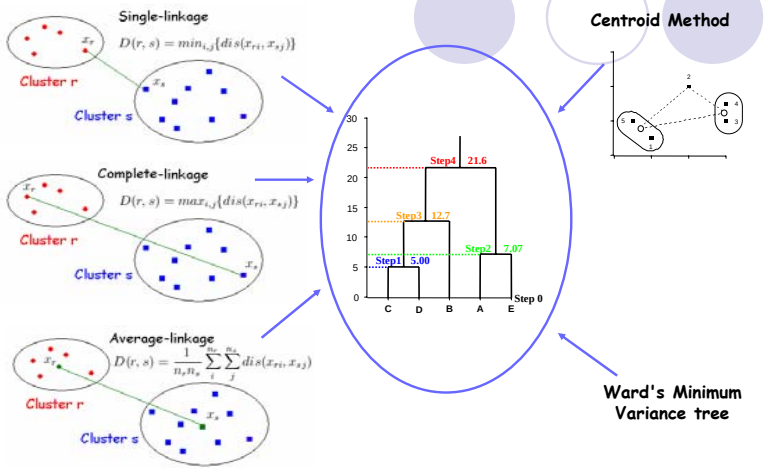

Ellipse–order
Ellipse–order–Tree
RAW–Tree
SOM
SOM–Tree
Optimal–Tree
Comparison of anti-Robinson scores
for the 6 seriation algorithms
(the smaller the better)
Windowsize (ws)
Tree/Raw
SOM
Tree/SOM
Tree/Optimal
Tree/GAP
GAP
Number of anti–Rubinson events
Windowsize


Ellipse–order
Ellipse–order–Tree
RAW–Tree
SOM
SOM–Tree
Optimal–Tree
Comparison of anti-Robinson scores
for the 6 seriation algorithms
(the smaller the better)
Windowsize (ws)
Ws=20
Ws=10
Best algorithms:
1<=ws<= 9: Tree/Optimal
10<=ws<=20: Tree/GAP
21<=ws<=max: GAP
Number of anti–Rubinson events
Windowsize


Example: microarray image examination (row/col do not represent gene/experiment)
original image (scanned)
Original image (scanned)
=>
Pseudo images (computed)
Foreground image
Background image
Signal image
Log(ratio) image
Differential (p-value) image
Spot diameter
flag
foreground Cy5
background Cy5
foreground Cy3
background Cy3
Log2(Cy5/Cy3)
Pseudo Images


Presentation and diagnostics for conventional visualization tools with proximity matrices visualization
Statistical Models:
Input Proximity Map
Transformed Disparity Map
(1) Multidimensional Scaling
(2) Hierarchical Clustering Tree
(3) Factor Analysis (PCA)
(4) Other Models …
Output Distance Map
Stress (Residual) Map
72

Matrix Visualization for Diagnosing Hierarchical Clustering



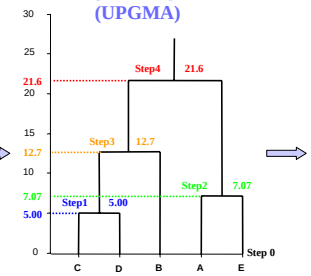
Matrix Visualization for Diagnosing Hierarchical Clustering

A dendrogram is displayed to represent the information in the input proximity matrix – a **dimension reduction** procedure – an **information reduce** process

Input **Proximity** Matrix

	A	B	C	D	E
A	0.00				
B	18.0	0.00			
C	20.6	14.1	0.00		
D	22.4	11.2	5.00	0.00	
E	7.07	18.0	25.0	25.5	0.00

Average Linkage **Tree** (UPGMA)



Output **Distance** Matrix (**Cophenetic Matrix**) from UPGMA Tree (**information reduced**)

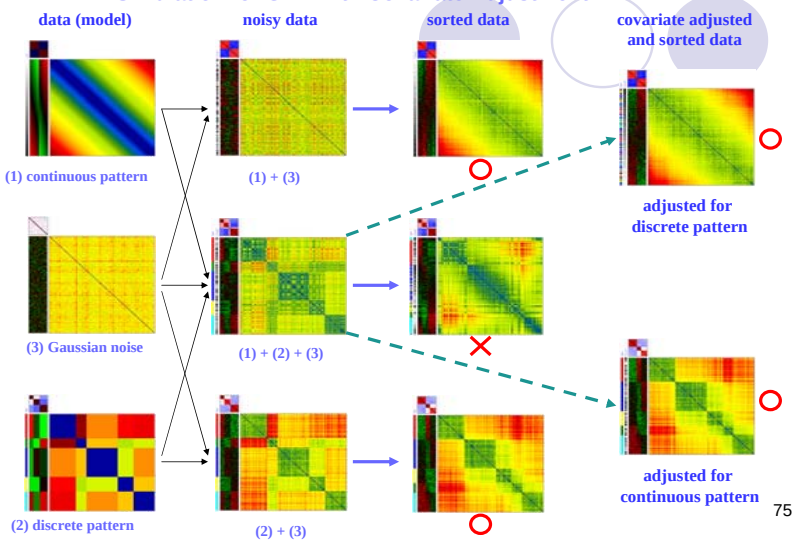
	A	B	C	D	E
A	0.00				
B	21.6	0.00			
C	21.6	12.7	0.00		
D	21.6	12.7	5.00	0.00	
E	7.07	21.6	21.6	21.6	0.00

Input Proximity / Output Distance

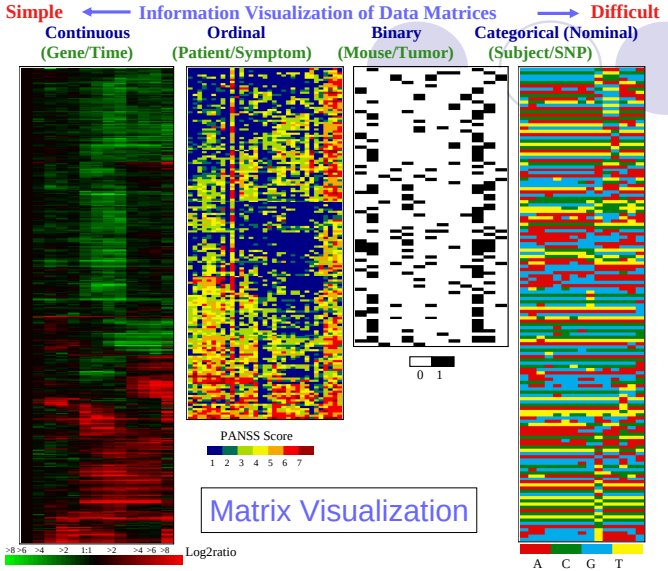
Cell	(B,A)	(C,A)	(D,A)	(E,A)	(C,B)	(D,B)	(E,B)	(D,C)	(E,C)	(E,D)
X	18.0	20.6	22.4	7.07	14.1	11.2	18.0	5.0	25.0	25.5
Y	21.6	21.6	21.6	7.07	12.7	12.7	21.6	5.0	21.6	21.6

The **Cophenetic** Correlation Coefficient
= Correlation Coefficient (x, y)
=R(x, y)
=0.93 (*goodness of fit*)

Simulation for GAP with Covariate Adjustment



GAP works only with continuous data?



77

Matrix visualization of binary data (GAP approach)

78

Similarity Measure for Binary Data

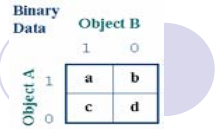
A binary variable has only two states: 0 or 1, where 0 means that the variable is absent, and 1 means that it is present.

Difference between symmetric and asymmetric binary variables

- A binary variable is symmetric if both of its states are equally valuable and carry the same weight, that is, there is no preference on which outcome should be coded as 0 or 1.
- A binary variable is asymmetric if the outcomes of the states are not equally important, such as the positive and negative outcomes of a disease test. By convention, we shall code the most important outcome, which is usually the rarest one, by 1 (e.g., HIV positive), and the other by 0 (e.g., HIV negative). Therefore, such binary variables are often considered “monary" (as if having one state).

79

Similarity measurements for Binary Data



similarity	equation	metric		Euclidean	
		$1-S$	$\sqrt{1-S}$	$1-S$	$\sqrt{1-S}$
Kulczynski	a/(b+c)	Y	Y	N	Y
Rao	a/(a+b+c+d)	Y	Y	N	Y
Jaccard	a/(a+b+c)	Y	Y	N	Y
simple match	(a+d)/(a+b+c+d)	Y	Y	N	Y
Sneath	a/(a+2b+2c)	Y	Y	N	Y
Rogers	(a+d)/(a+2b+2c+d)	Y	Y	N	Y
Hamman	(a+d-b-c)/(a+b+c+d)	Y	Y	N	Y
Phi	$\frac{ad-bc}{\sqrt{(a+b)(a+c)(d+b)(d+c)}}$	N	Y	N	Y
Yule	(ad-bc)/(ad+bc)	N	N	N	N

80

Binary GAP Example I

CGMIM *Online* http://www.bccrc.ca/ccr/CGMIM/

CGMIM performs automated text-mining of OMIM to identify genetically-related cancers

Online Mendelian In Man (OMIM) is a computerized database of information about genes and heritable traits in human populations

OMIM is maintained on the Internet by the National Center for Biotechnology Information at the US National Institutes of Health

CGMIM considers 21 anatomic sites based on the major cancers identified by the National Cancer Institute of Canada

CGMIM compares each OMIM entry name and alternative name with a list of gene names assigned by HUGO (HUman Genome Organization).

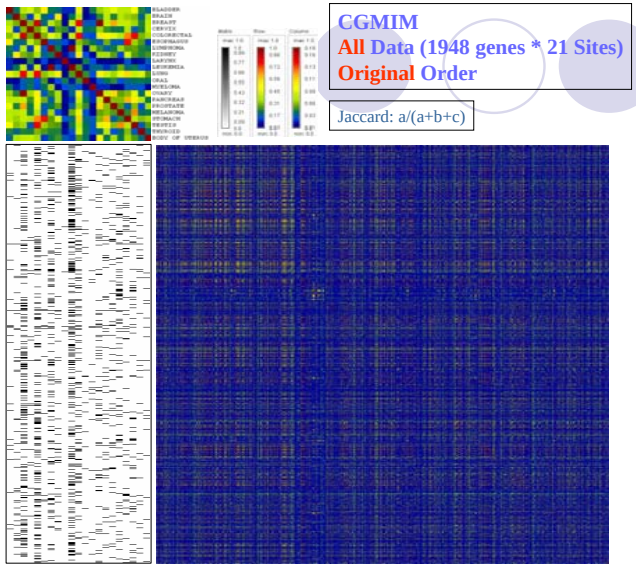
CGMIM produces the number of genes for which an OMIM entry mentions each pair of cancers, as well as a ratio of the observed and expected number of genes for the combination

BC Cancer Agency
CARE & RESEARCH
An agency of the Provincial Health Services Authority

21 Cancer Sites

BLADDER
BRAIN
BREAST
CERVIX
COLORECTAL
ESOPHAGUS
LYMPHOMA
KIDNEY
LARYNX
LEUKEMIA
LUNG
ORAL
MYELOMA
OVARY
PANCREAS
PROSTATE
MELANOMA
STOMACH
TESTIS
THYROID
BODY OF UTERUS

1948 Related Genes



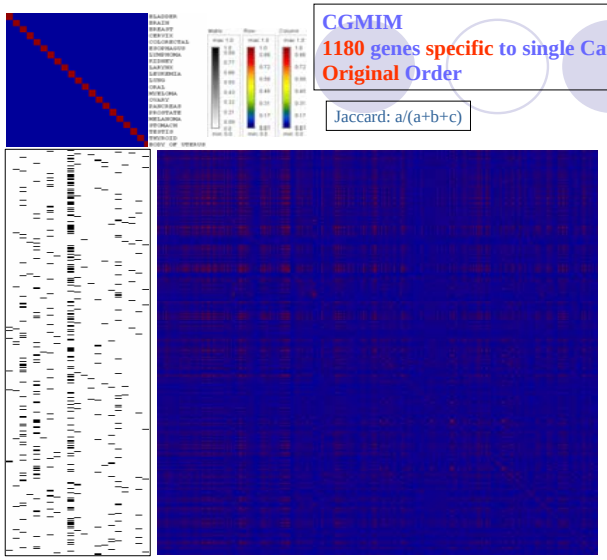
CGMIM
All Data (1948 genes * 21 Sites)
Original Order

Jaccard: a/(a+b+c)

21 Cancer Sites

BLADDER
BRAIN
BREAST
CERVIX
COLORECTAL
ESOPHAGUS
LYMPHOMA
KIDNEY
LARYNX
LEUKEMIA
LUNG
ORAL
MYELOMA
OVARY
PANCREAS
PROSTATE
MELANOMA
STOMACH
TESTIS
THYROID
BODY OF UTERUS

1180 Related Genes



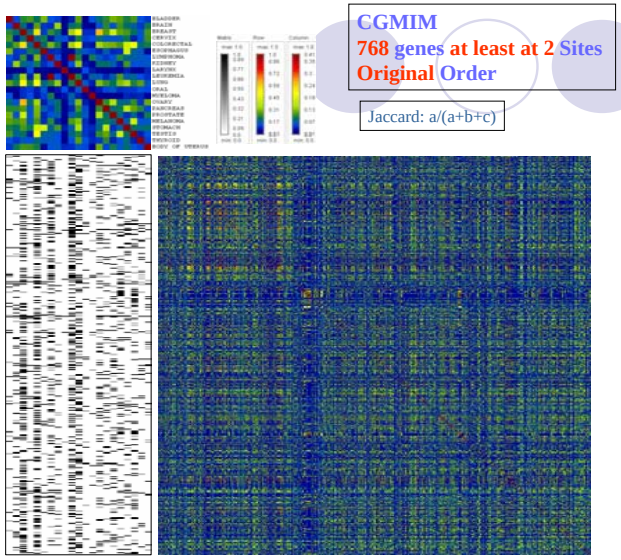
CGMIM
1180 genes specific to single Cancer
Original Order

Jaccard: a/(a+b+c)

21 Cancer Sites

BLADDER
BRAIN
BREAST
CERVIX
COLORECTAL
ESOPHAGUS
LYMPHOMA
KIDNEY
LARYNX
LEUKEMIA
LUNG
ORAL
MYELOMA
OVARY
PANCREAS
PROSTATE
MELANOMA
STOMACH
TESTIS
THYROID
BODY OF UTERUS

768 Related Genes



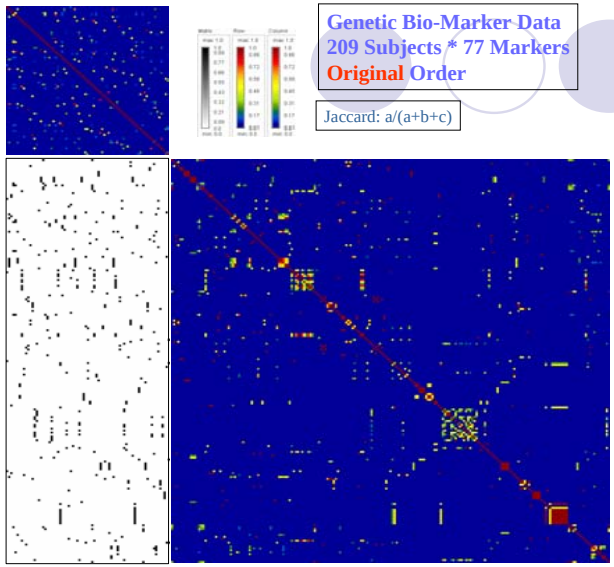
CGMIM
768 genes at least at 2 Sites
Original Order

Jaccard: a/(a+b+c)

Binary GAP Example II

Data Set:
Genetic Bio-Marker Data
209 Subjects * 77 Markers

Issues:
1. Dimensional too high –
Variables are not correlated
2. Single-linkage vs Complete-linkage



Visualizing Categorical Data

- Mosaic Display
- Correspondence Analysis

✓ "Categorical data consists of variables whose values comprise a set of discrete categories. Such data require different statistical and graphical methods than commonly used for quantitative data." Michael Friendly said so.

Mosiac Displays for Two-way Tables

- The mosaic display, proposed by Hartigan & Kleiner (1981) and extended in Friendly (1994a), represents the counts in a contingency table directly by tiles whose size is proportional to the cell frequency.

Eye Color	BLACK	BROWN	RED	BLOND	Total
Brown	68	119	26	7	220
Blue	20	84	17	94	215
Hazel	15	54	14	10	93
Green	5	29	14	16	64
Total	108	286	71	127	n 592

Question:
how to understand the nature of the association between hair and eye color.

The Pearson X2 for these data is 138.3 with 9 degrees of freedom, indicating substantial departure from independence.

Marginal proportions

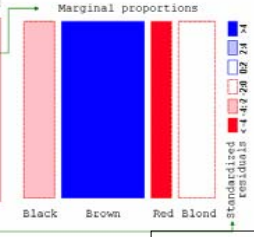
Black	Brown	Red	Blond
0.1824	0.4831	0.1199	0.2145

Fitted frequencies

Black	Brown	Red	Blond
148.00	148.00	148.00	148.00

Standardized Pearson residuals

Black	Brown	Red	Blond
-3.29	11.34	-6.33	-1.73



Size: visualization for data

Color: visualization for model

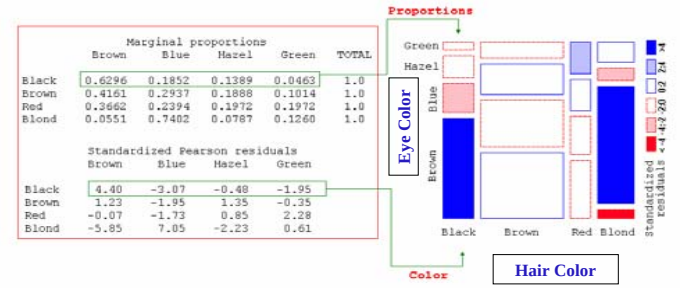
Reference: http://www.math.yorku.ca/SCS/Online/mosaics/about.html

Hair Color

Mosiac Displays (conti.)

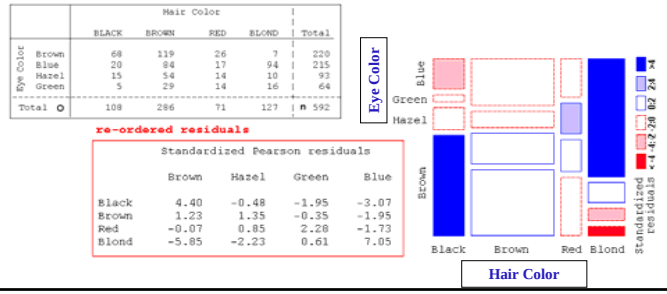
Interpretation

- To interpret the association between Hair Color and Eye Color, consider the pattern of positive (Blue) and negative (Red) tiles in the mosaic display.
  - **Positive values** as showing cells whose observed frequency is substantially greater than would be found under independence;
  - **Negative values** indicate cells which occur less often than under independence.



Mosiac Displays (conti.)

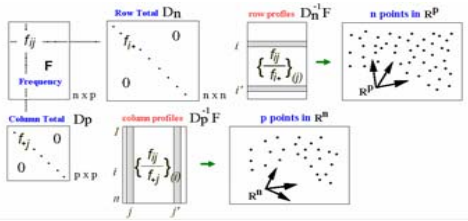
- This interpretation is enhanced by reordering the rows or columns of the two-way table so that the residuals have an opposite corner pattern of signs.
- Thus, the mosaic shows that the association between Hair and Eye color is essentially that
  - people with dark hair tend to have dark eyes,
  - those with light hair tend to have light eyes
  - people with red hair do not quite fit this pattern



Simple Correspondance Analysis (CA)

- Correspondence Analysis = PCA for categorical variables.
- Correspondence analysis is designed to analyze simple two-way and multi-way tables containing some measure of correspondence between the rows and columns.
- The results allow one to explore the structure of categorical variables included in the table.
- In a typical correspondence analysis, a crosstabulation table of frequencies is first standardized, so that the relative frequencies across all cells sum to 1.0.
- Correspondence analysis can also be viewed as finding the best simultaneous representation of two data structure (rows and columns of a data matrix)

This technique finds scores for the row and column categories on a small number of dimensions which account for the greatest proportion of the chi^2 for association between the row and column categories, just as principal components account for maximum variance.



Correspondance Analysis (conti.)

- Row points for the disciplines, Column points for the years.
- The anthropology degree and the engineering degree are far from each other because their profiles are different, mathematics degree is near the engineering degree because their profiles are similar.
- Each year point represents the profile of that year across the various disciplines.
- **Note that** the positions of two sets of points with respect to each other are not directly comparable and should be interpreted with caution.

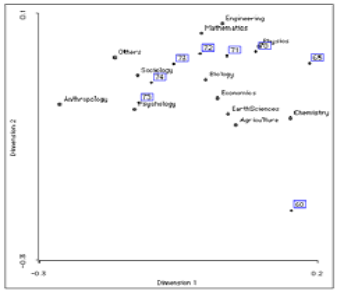
Interpretation

- Each discipline point will lie in the neighborhood of the year in which the discipline's profile is prominent.
- There are relatively more agriculture, earth science and chemistry degrees in 1960, while the trend from 1965 to 1975 appears to be away from the physical sciences towards the social sciences.
- The points such as earth sciences and economics lie within the parabolic configuration of the years points; this implies that the profiles of these disciplines are higher than average in the early and later years.

Science Doctorates in the USA, 1960-1975

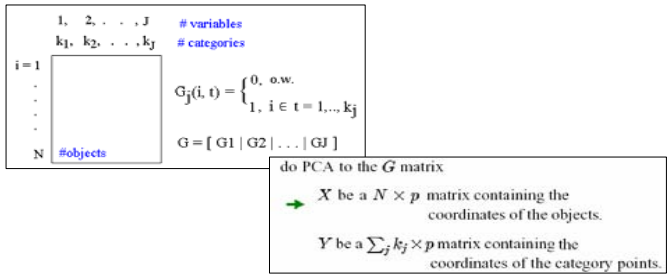
Discipline/Year	1960	1965	1970	1971	1972	1973	1974	1975
Engineering	794	2073	3432	3495	3475	3338	3144	2959
Mathematics	291	685	1222	1236	1281	1222	1196	1149
Physics	530	1046	1655	1740	1635	1590	134	1293
Chemistry	1078	1444	2234	2204	2011	1849	1792	1762
Earth Sciences	253	375	511	550	580	577	570	556
Biology	1245	1963	3360	3633	3580	3636	3473	3498
Agriculture	414	576	803	900	855	853	830	904
Psychology	772	954	1888	2116	2262	2444	2587	2749
Sociology	162	239	504	583	638	599	645	680
Economics	341	538	826	791	863	907	833	867
Anthropology	69	82	217	240	260	324	381	385
Others	314	502	1079	1392	1500	1609	1531	1550

The multidimensional time series on the number of science doctorates conferred in the USA from 1960 to 1975 (Greenacre, 1984).



Multiple Correspondance Analysis (Homogeneity Analysis)

- Multiple Correspondence Analysis (MCA) is known as homogeneity analysis, or dual scaling, or reciprocal averaging.
- The general idea of homogeneity analysis is to make a joint plot in p-space of all objects (or individuals) and the categories of all variables.
- Objects close to the categories they fall in and categories close to objects belonging in them



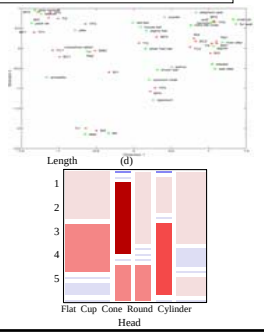
Matrix visualization of categorical data (GAP approach)

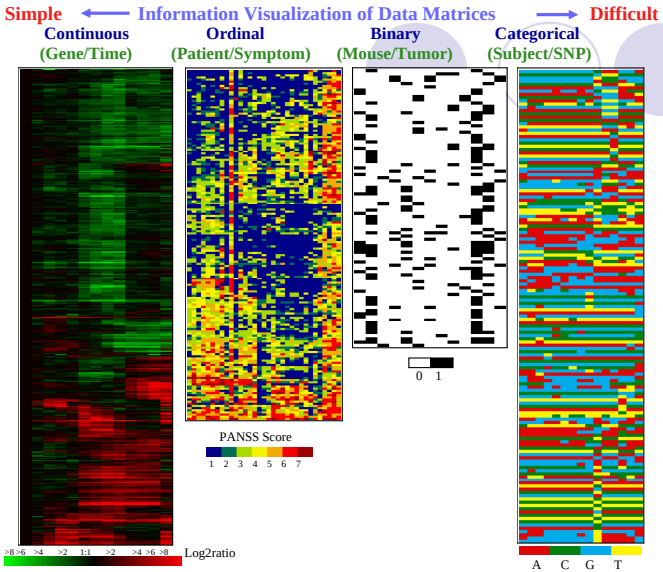
Multiple Correspondence Analysis (MCA)

Mosaic Display (Hartigan and Kleiner, Friendly, Spence, Theus and Lauer)

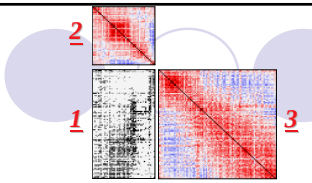
Categorical PCP

…





Three Major Components (1/2/3) for Constructing a GAP-Type Display



	Continuous	Categorical
1. Color Coding	Gray-Scale Rainbow Spectrum Blue-Red Dye ….	Binary 0/1: (white/black) OK (or not OK)? True nominal: red/green/blue/cyan/magenta/yellow/ black ? Nucleotide: A C G T Amino Acid: ???
2. Proximity for var.	Correlation Covariance polychoric correlation	• χ^2 type measurements?
3. Proximity for sub.	Euclidean Distance Manhattan Distance Correlation …	• Matching proportion?

Is there a natural way of taking care of all 3 problems?

Solution: Dual Scaling/Homogeneity Analysis/MCA

Early Works:

Richardson & Kuder (1933)
Hirschfeld (1935)
Horst (1935)
Edgerton & Kolbe (1936)
Hotelling (1936)
Wilks (1938)
Fisher (1940)
Maung (1941)
Guttman (1941, 1946)
Hayashi (1950, 1952)
Bock (1956, 1960)

Major Groups:

Hayashi school (1950-)
Benzecri school (1960-)
Gifi group (1967-)
de Leeuw & others
Toronto group (1969-)
Nishisato & others

Aliases:

Method of Reciprocal Average
Simultaneous Linear Regression
Appropriate Scoring, Additive Scoring
Hayashi's Theory of Quantification
Principal Component Analysis of Qualitative Data
Optimal Scaling
Analyse Factorielle des Correspondances
Homogeneity Analysis
Correspondence Analysis
Correspondence Factor Analysis
Basic Structure Content Scaling
Dual Scaling
Descriptive Multivariate Analysis
Nonlinear Multivariate Analysis

Review Works:

Nishisato, S. (1996), "Gleaning in the field of dual scaling," *Psychometrika*, **61**, 559-599.
Michailidis, and de Leeuw, J. (1998), "The Gifi System of Descriptive Multivariate Analysis," *Statistical Science*, **13**, 307-336.

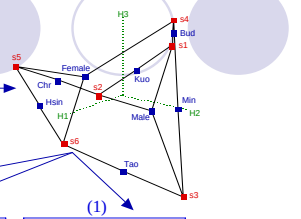
Concept of Categorical GAP with Gifi-Homals

Toy Data Set

Subject	Gender	Reli	Poli
s1	Male	Bud	Kuo
s2	Male	Chr	Kuo
s3	Male	Tao	Min
s4	Female	Bud	Min
s5	Female	Chr	Hsin
s6	Female	Tao	Hsin

1	1	1
1	2	1
1	3	2
2	1	2
2	2	3
2	3	3

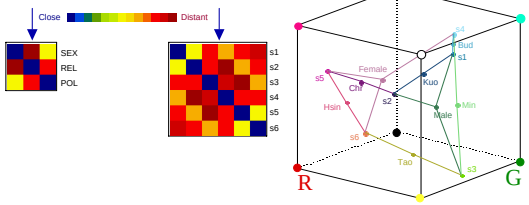
Obtain the Homals' 3 Dimensional Dual Space Solution



(3) Compute the Proximity for 2 Variables as the Sum of Weighted 3D Euclidean Distance between Corresponding Categories for the 2 Variables from the Homals' 3 Dimensional Dual Space

(2) Compute the Proximities for 2 Subjects as the 3D Euclidean Distances for the 2 subjects from the Homals' 3 Dimensional Dual Space

(1) Scale the Homals' 3 Dimensional Dual Space into the RGB Cube

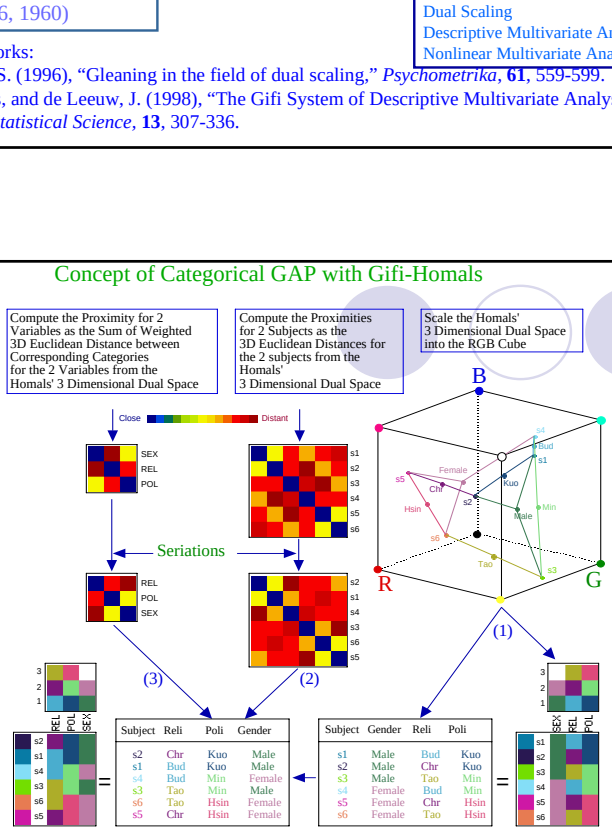


Concept of Categorical GAP with Gifi-Homals

(3) Compute the Proximity for 2 Variables as the Sum of Weighted 3D Euclidean Distance between Corresponding Categories for the 2 Variables from the Homals' 3 Dimensional Dual Space

(2) Compute the Proximities for 2 Subjects as the 3D Euclidean Distances for the 2 subjects from the Homals' 3 Dimensional Dual Space

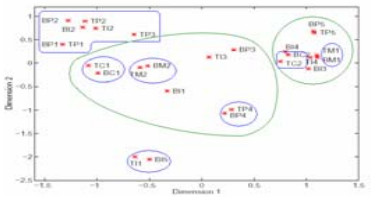
(1) Scale the Homals' 3 Dimensional Dual Space into the RGB Cube



Homogeneity Analysis (conti.)

Mammals Dentition Example

The data for this example are taken from Hartigan (1975) (also discussed in Michailidis and De Leeuw,1999). Dental characteristics are used in the classification of 66 different kinds of mammals. Mammals' teeth are divided into four groups: incisors, canines, premolars, and molars.

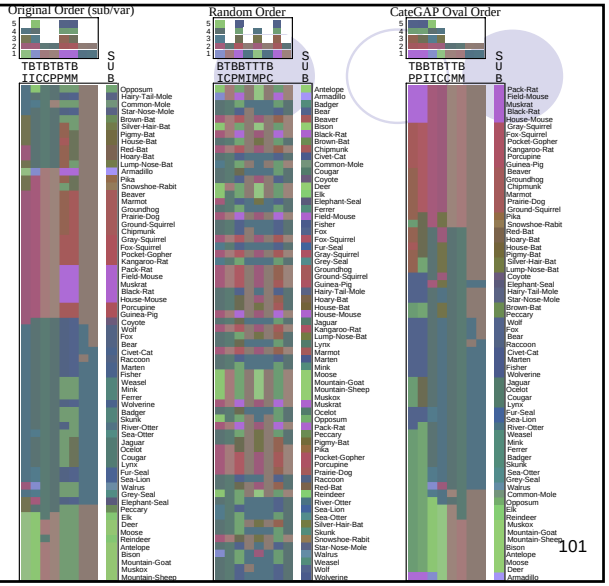


Category quantifications of the variables in the mammals dentition example

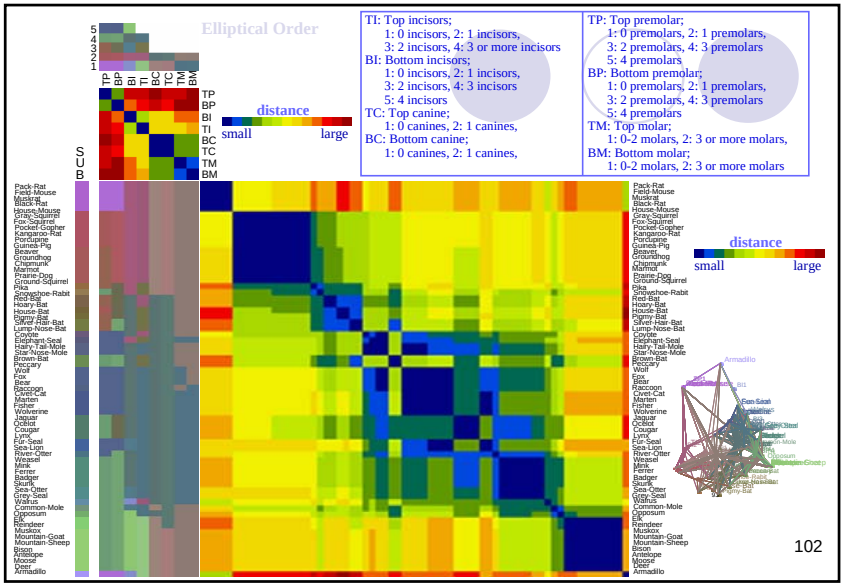
Description for Variables

TI: Top incisors,
1: 0 incisors, 2: 1 incisors,
3: 2 incisors, 4: 3 or more incisors
BI: Bottom incisors,
1: 0 incisors, 2: 1 incisors,
3: 2 incisors, 4: 3 incisors
5: 4 incisors
TC: Top canine,
1: 0 canines, 2: 1 canines,
BC: Bottom canine,
1: 0 canines, 2: 1 canines,
TP: Top premolar,
1: 0 premolars, 2: 1 premolars,
3: 2 premolars, 4: 3 premolars
5: 4 premolars
BP: Bottom premolar,
1: 0 premolars, 2: 1 premolars,
3: 2 premolars, 4: 3 premolars
5: 4 premolars
TM: Top molar,
1: 0-2 molars, 2: 3 or more molars,
BM: Bottom molar,
1: 0-2 molars, 2: 3 or more molars

Categorical Generalized Association Plots



101

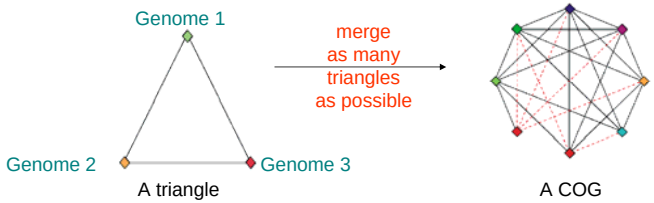


102

Matrix Visualization for Phylogenetic Profiles

COG

- Cluster of Orthologous Groups of proteins.
- Each COG consists of individual orthologous proteins or groups of paralogs from at least 3 lineages and thus corresponds to an ancient conserved domain.
- The proteins that comprise each COG are assumed to have evolved from an ancestral protein, and are therefore either orthologs or paralogs.

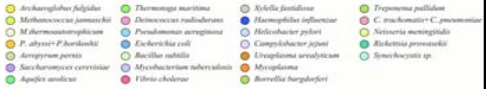


http://www.ncbi.nlm.nih.gov/COG/

103

Conventional Analysis of COG data

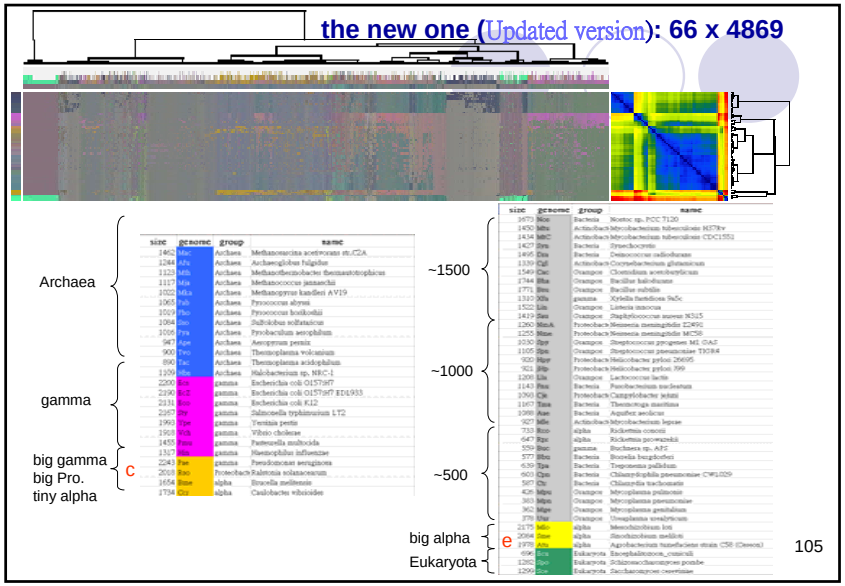
2. principal component analysis (PCA) of binary status in the phylogenetic patterns



1. co-occurrences of genomes among COGs

3. clustering tree of proteins within each COG

4

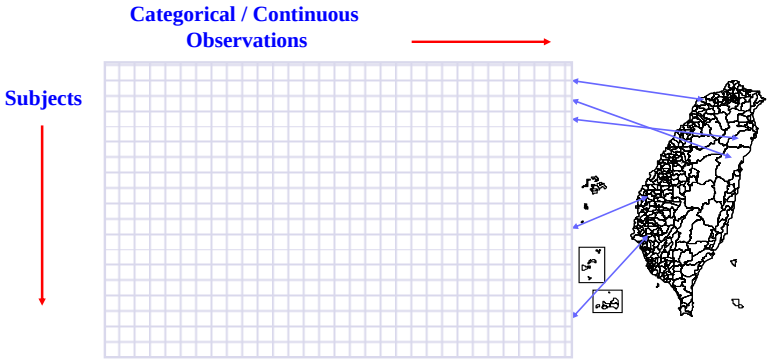


105

Matrix visualization with Geographical Information

106

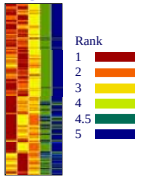
Cartography Coloring Scheme with Categorical GAP (CartoGAP) - 1



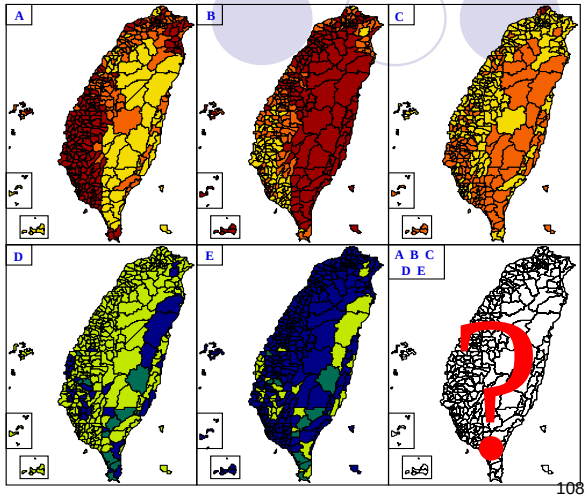
107

Cartography Coloring Scheme with Categorical GAP (CartoGAP) - 2

Data:
Ranks of
5 Candidates
(A B C D E) on
360 Townships
Taiwan Presidential Election 2000



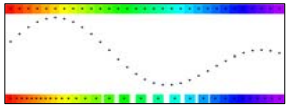
Is it possible to visualize information structure for all 5 candidates in a single MAP?



108

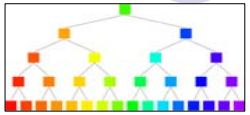
White, D, and Sifneos, J. C. (2002). "Regression Tree Cartography." JCGS, 11, 3 (2002).

Equally spaced rainbow spectrum



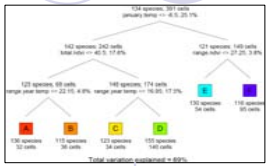
Weighted curve for resampling

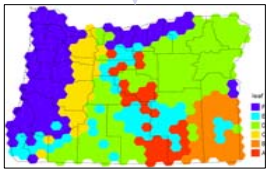
Unequally spaced rainbow spectrum



Equally spaced rainbow spectrum through the partitioning of **terminal nodes of a tree** dendrogram

Example:



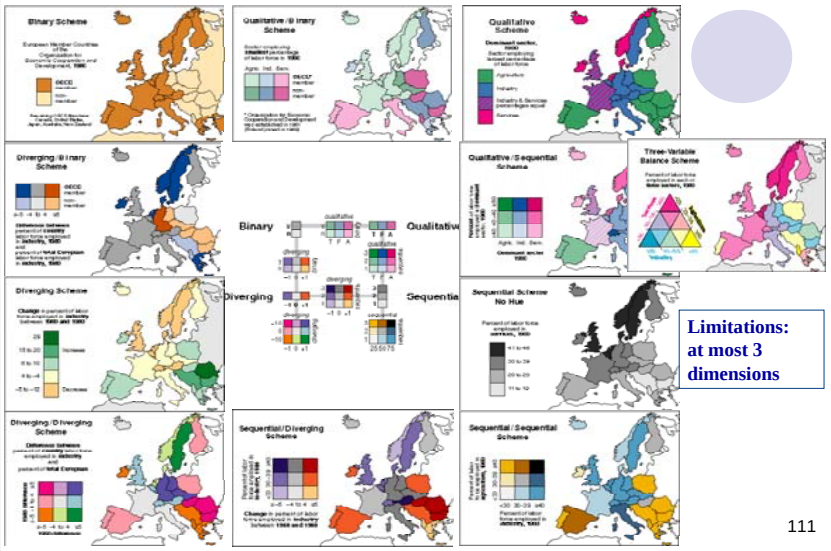


Limitations:

1. **Univariate response variable (Y)**
2. **Unidimensional color spectrum**
3. **Flips of intermediate nodes.**
4. **Categorical variable? (Univariate)**

Brewer, C. A. (1994). "Color Use Guidelines for Mapping and Visualization," Chapter 7 (pp. 123-147) in Visualization in Modern Cartography, edited by A.M. MacEachren and D.R.F. Taylor, Elsevier Science, Tarrytown, NY. (Also 1999 JSM Invited Talk)

- Binary Color Schemes
- Qualitative Binary Color Schemes
- Qualitative Color Schemes
- Diverging Binary and Diverging Sequential Color Schemes
- Qualitative Sequential Color Schemes
- Diverging Color Schemes
- Diverging Sequential Color Schemes
- Sequential Color Schemes
- Diverging Diverging Color Schemes

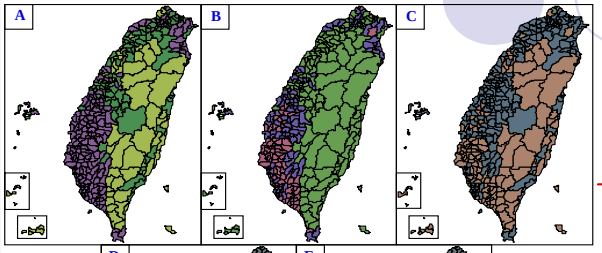


Limitations: at most 3 dimensions

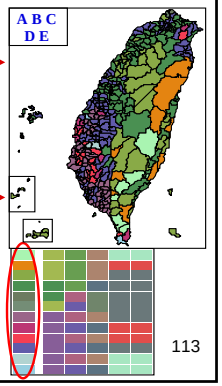
Matrix visualization with Geographical Information (GAP approach)

Cartography Coloring Scheme with Categorical GAP (CartoGAP) - 6

(F). CateGAP Color Map for Each Individual Variable (Candidate)

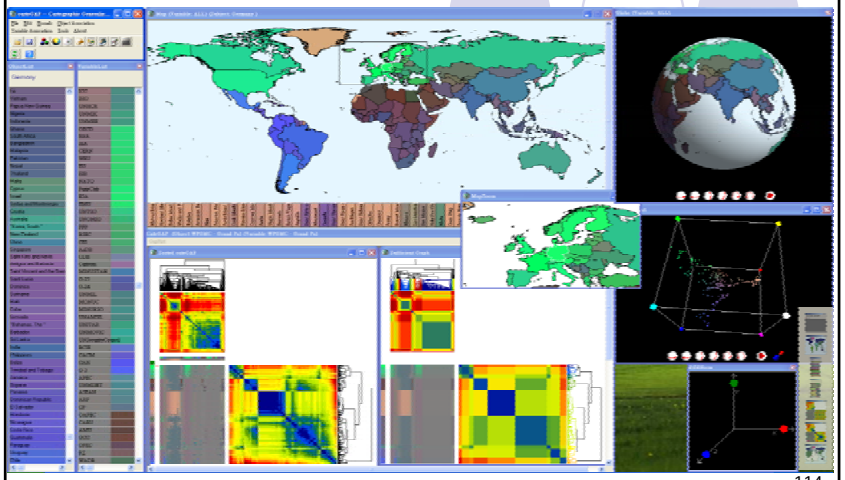


(G). Final Single CateGAP Cartography Color MAP for Complete Information Visualization

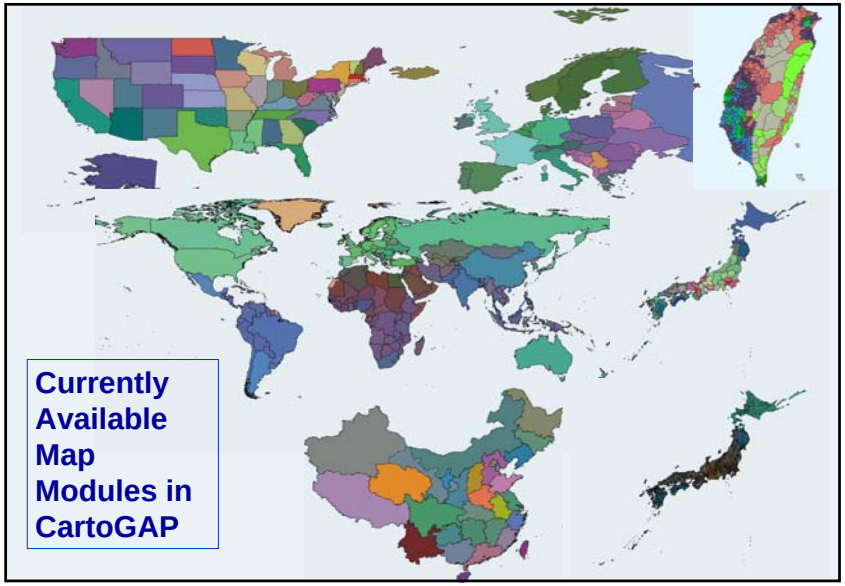


113

Cartography GAP



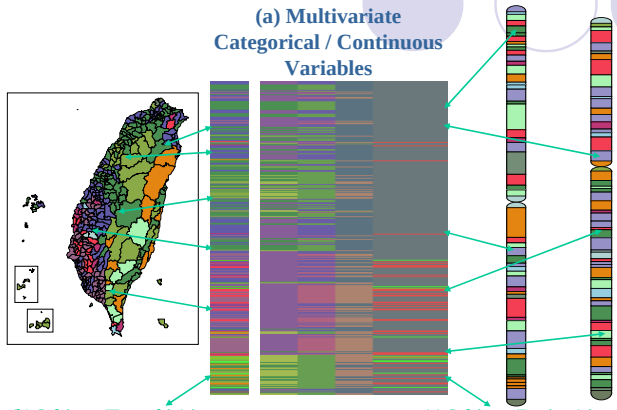
114



Currently Available Map Modules in CartoGAP

Cartography GAP (CartoGAP) for Chromosone Map

(a) Multivariate Categorical / Continuous Variables



(b) Subjects (Township) in A Geographical Map

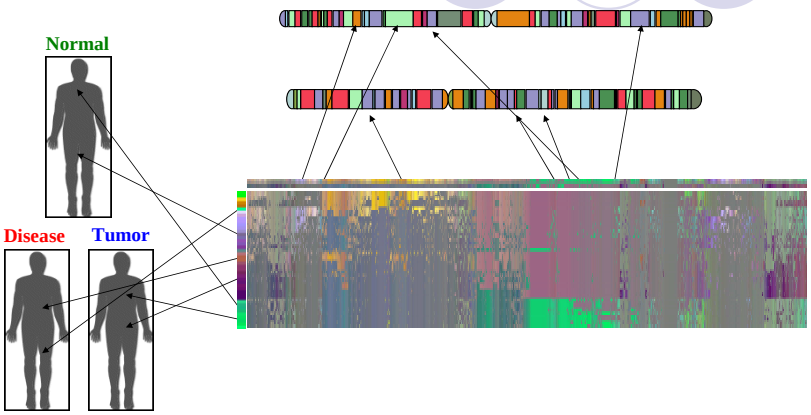
(c) Subjects (Regions) in A Chromosome Map

116

Ideal Situation

Tissue Maps

Gene Maps



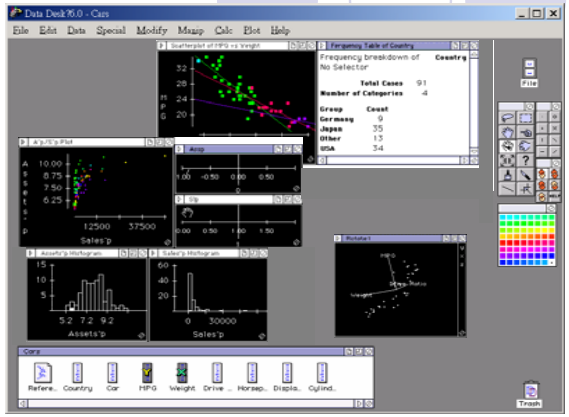
Software

- ***Freeware/Shareware***
  - ***Data Desk*** http://www.datadesk.com/products/data_analysis/datadesk/
  - ***R*** http://www.r-project.org/
  - ***GGobi*** http://www.ggobi.org/
  - ***GAPLite*** http://gap.stat.sinica.edu.tw/Software/index.htm
  - Vista http://forrest.psych.unc.edu/research/index.html
- ***Commercial***
  - Splus http://www.insightful.com/
  - SPSS http://www.spss.com/
  - Statistica http://www.statsoftinc.com/
  - Stata http://www.stata.com/

Data Desk v6.2

Features

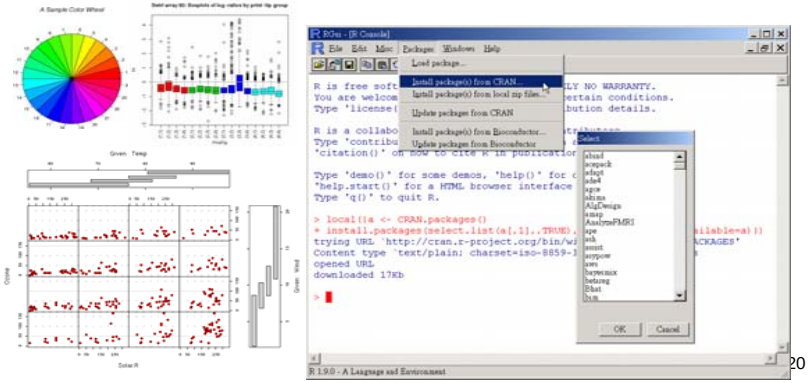
Direct Interface
Drag and Drop
Dynamic Graphics
Easy to Use
Excel Link
Fast and Lean
Interactive Analyses
Linked Plots
Template Files



Data Description, Inc.
http://www.datadesk.com/
Data Desk
http://www.datadesk.com/products/data_analysis/datadesk/
Data Desk Multimedia Tour
http://www.datadesk.com/products/mediadx/custom/lessonbook/nyheart.shtml

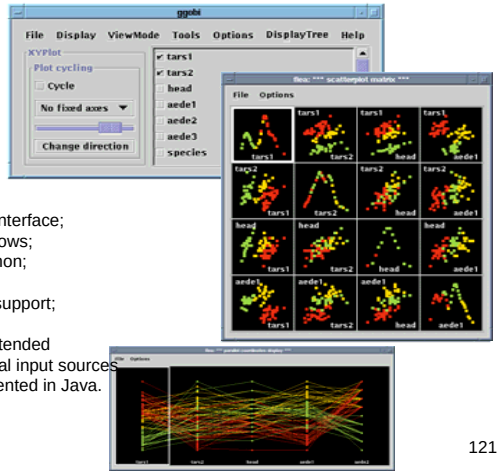
The R Project for Statistical Computing v1.9.0

- R is a language and environment for statistical computing and graphics.
- It is a GNU project which is similar to the S language and environment.
- R provides a wide variety of statistical (linear and nonlinear modelling, classical statistical tests, time-series analysis, classification, clustering, ...) and graphical techniques, and is highly extensible. http://www.r-project.org/



GGobi v1.0-1-beta

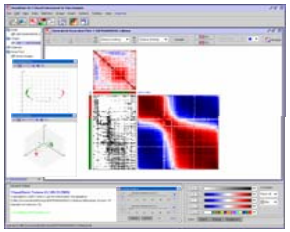
GGobi Data Visualization System
http://www.ggobi.org/
GGobi is a data visualization system for viewing high-dimensional data and is the next edition of xgobi.



Features:
- A new, simpler and more modern interface;
- Better portability to Microsoft Windows;
- Direct access from R Perl and Python;
- New input format using XML;
- Database (Postgres and MySQL) support;
- Works as a Gnumeric plugin;
- Plugin mechanism for providing extended functionality and support for additional input sources and formats. Plugins can be implemented in Java.

Software

GAP



(by Dr. Hank Wu)

CateGAP

CartoGAP

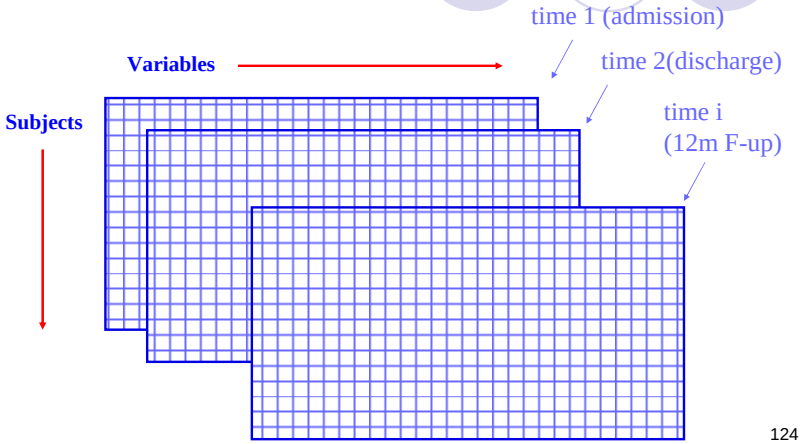
http://gap.stat.sinica.edu.tw/

Future directions for GAP_MV:

1. MV for longitudinal multivariate data
2. MV for multi-conditioned multivariate data
3. MV with dependent variable
4. MV with dependent (clustered) structure
5. MV for mixed data
6. MV for huge data set
7. MV for data with missing values
8. MV for color-blind people
9. MV for time series data
10. MV with nonlinear proximity measurement

These problems exist in not only MV but in all general statistical graphics and visualization

1. MV for longitudinal multivariate data



2. MV for multi-conditioned multivariate data

Set X
(CPT)
Raw Scores / Dprime / LnBeta

Set Y
(BCDS)
A / B / C / D …

Set Z
NPT
WAIS-R / WMS-R / Trail-A / Trail-B

Same Set of Subjects

(endo)phenotypes (endo)genotypes environtypes

125

3. MV with dependent variable(s)

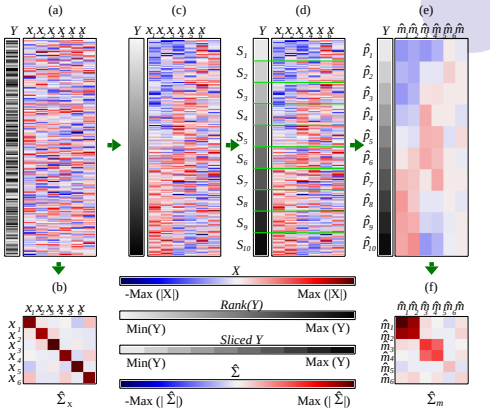


Figure 4. Matrix map of the raw data matrix (Y, $\mathbf{x}$) with a PCA analysis and the SIR algorithm. (a). original (unsorted) matrix map; (b). sample covariance matrix of $\mathbf{x}$ in (a), $\hat{\Sigma}_x$; (c). sorted (by rank of Y) map; (d). sliced sorted map; (e) map for sliced mean matrix, $\hat{m}$; (f). sample covariance matrix of sliced mean matrix in (e), $\hat{\Sigma}_m$.

MV for a regression context with dependent variables is similar but not identical to MV with adjusting covariates.

Sliced inverse regression (SIR) Li (1991) is a natural staring point.

126

4. MV with dependent (clustered) structure

Family 1
Father1
Mother1
Sib11
Sib12

Family 2
Father2
Mother2
Sib21
Sib22

Family k
Fatherk
Motherk
Sibk1
Sibk2

How to:

Compute proximity
for individuals?
for clusters?

Display matrices
for individuals?
for clusters?

127

5. MV for mixed data

Same Set of Subjects

Continuous Categorical

1. Calculation of proximity matrices for variables and subjects

General similarity coefficients Gower (1971)

General weighted two-way dissimilarity coefficients introduced Cox and Cox (2000)

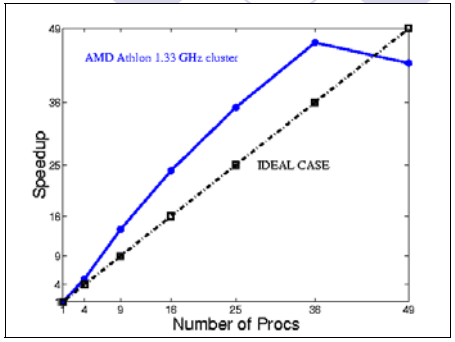
2. Color coding for a data matrix with mixed data is a more difficult task.

128

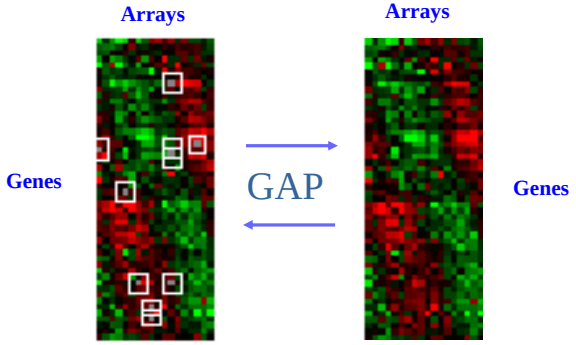
6. MV for huge data set

Problems

1. **Computation**
   - sampling
   - distributed
   - parallel
2. **Visualization**
   - sampling
   - smoothing
   - sequential



7. DMV for data with missing values



8. MV for Color Blind people

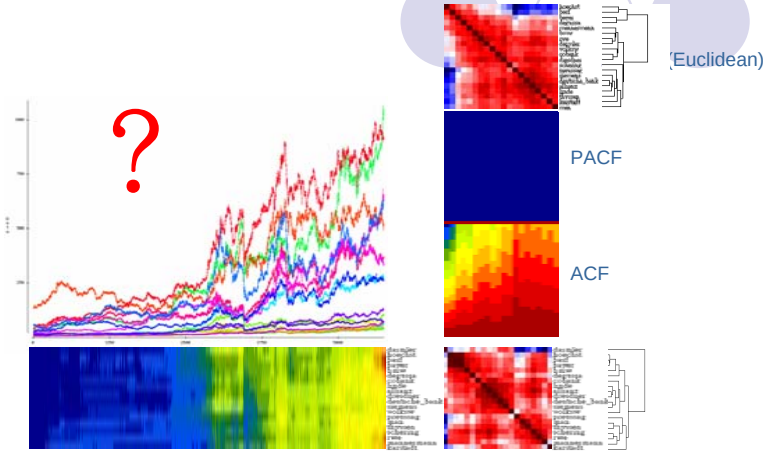
Vischeck
http://www.vischeck.com/examples/

Types of color blind
Monochromacy
Dichromacy
Protanopia and deuteranopia
Hereditary tritanopia
Anomalous Trichromacy

To act passively to prevent from using color systems that are difficult for color blind people to understand.
or
To work actively in assisting people with visual impairments to have better visualization of data/information.

"I believe there are more mathematics/statistics blind people than color blind people"

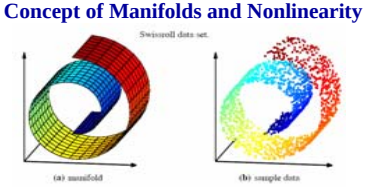
9. MV for multiple time series data

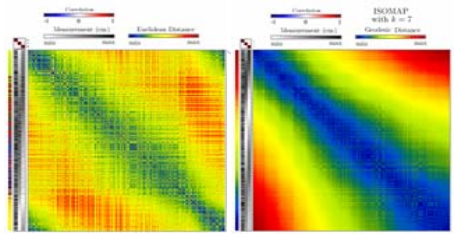


Data provided by Professor WOLFGANG HÄRDLE

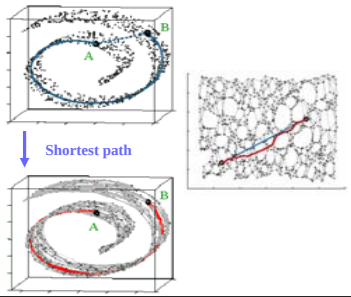
10. MV with nonlinear proximity measurement

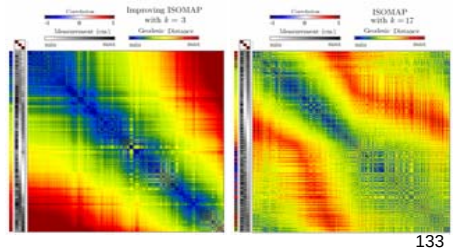
Concept of Manifolds and Nonlinearity





Isometric Mapping (isomap)





Reference

- Dr. Alexander Strehl: http://www.lans.ece.utexas.edu/~strehl/
- Michael Friendly's Home Page:http://www.math.yorku.ca/SCS/friendly.html
- Statistics and Statistical Graphics Resources http://www.math.yorku.ca/SCS/StatResource.html
- Gallery of Data Visualization: The Best and Worst of Statistical Graphics http://www.math.yorku.ca/SCS/Gallery/
- Statistical Graphics for Multivariate Data http://www.math.yorku.ca/SCS/sugi/sugi16-paper.html
- Michael Friendly, SCS Short Course: Exploratory and Graphical Methods of Data Analysis http://www.math.yorku.ca/SCS/Courses/eda/
- Chambers, J. M., Cleveland, W. S., Kleiner, B., & Tukey, P. A. (1983). Graphical Methods for Data Analysis. Belmont, CA: Wadsworth.
- Chen, C. H. (2002), Generalized Association Plots: Information Visualization via Iteratively Generated Correlation Matrices, Statistica Sinica, 12, 7-29.
- Cox, T. F. and Cox, M.A.A. (2001), Multidimensional Scaling, London: Chapman & Hall.
- Jacoby, William G. (1998). Statistical graphics for visualizing multivariate data. Thousand Oaks, Calif. : Sage Publications.
- Jacoby, William G(1997). Statistical graphics for univariate and bivariate data. Thousand Oaks, Calif. : Sage Publications.
- Kohonen, T. (2001), Self-Organizing Maps, Berlin: Springer.
- Li, K. C. (1991) . Sliced inverse regression for dimension reduction with discussion. Journal of the American Statistical Association 86, 316- 342.
- Li, K. C. (1992). On principal Hessian directions for data visualization and dimension reduction: another application of Stein's lemma," Journal of the American Statistical Association 87, 1025-1039.
- Minnotte , M. C. and West, R. W., (1999), "The Data Image: a Tool for Exploring High Dimensional Data Sets,". 1998 Proceedings of the ASA Section on Statistical Graphics, in press.
- Schmid, Calvin Fisher, (1992). Statistical graphics : design principles and practices. Malabar, FL : Krieger.

Handbook of Computational Statistics (Volume III)
Data Visualization

Chun-houh Chen, Wolfgang Härdle, and Antony Unwin (eds)
Springer-Verlag, Heidelberg

HBCSC for Editors | | HBCSC for Authors Last Updated: 2005/05/21

CSC Editorial Board



Chun-houh Chen
Institute of Statistical Science
Academia Sinica, Taiwan
cchen@stat.sinica.edu.tw
http://gap.stat.sinica.edu.tw

Wolfgang Härdle
Institut for Statistik und Econometrie
Humboldt-Universität zu Berlin
haerdle@wiwi.hu-berlin.de
http://ise.wiwi.hu-berlin.de

Antony Unwin
Institut for Mathematik
Universitat Augsburg
Antony.Unwin@Math.Uni-Augsburg.DE
http://stats.math.uni-augsburg.de/~unwin

CSC Conference
Workshop on Data and Information Visualization 2006
Aug. 24~25, 2006
at Berlin

Previous Handbooks

CSA
Handbook of Computational Statistics (Volume I)
Concepts and Methods
Gentle, James E.; Härdle, Wolfgang; Mori, Yuichi (Eds.)
2004, XII, 1070 p. 236 illus., Hardcover
ISBN: 3-540-40464-3
Language: English
Publisher: Springer-Verlag New York

CSB
Handbook of Computational Statistics (Volume II)
Partial Least Squares
Vincenzo Esposito Vinzi (managing editor)
Wynne W. Chin, Joerg Henseler, Huiwen Wang (Eds.)

PLS05 Conference
Barcelona, Spain
September 7-9, 2005

Designed and Maintained by Han-Ming (Hank) Wu
Institute of Statistical Science, Academia Sinica, Taiwan
hmwu@stat.sinica.edu.tw
http://www.sinica.edu.tw/~hmwu/

Homework

*** Register at the following page for GAP:**
http://gap.stat.sinica.edu.tw/Software/GAP/RegisterPage/index.htm
*** Download the GAP package**
*** Report 2 bugs and make 3 suggestions**