Waseda University – Academia Sinica Data Science Workshop

◆ Lecture 3

Dr. Chun-houh Chen

Research Fellow & Director, Institute of Statistical Science, Academia Sinica, Taiwan

Research fields: Clustering Analysis, Data/Information Visualization, Exploratory Data Analysis (EDA), Matrix Visualization (MV)

Title: Matrix Visualization: New Generation of Exploratory Data Analysis

"*It is important to understand what you CAN DO before you learn to measure how WELL you seem to have DONE it*" (Exploratory Data Analysis: John Tukey, 1977). Data analysts and statistics practitioners nowadays are facing difficulties in understanding higher and higher dimensional data with more and more complex nature while conventional graphics/visualization tools do not answer the needs. It is statisticians' responsibility for coming up with graphics/visualization environment that can help users really understand what one CAN DO for complex data generated from modern techniques and sophisticated experiments.

Matrix visualization (MV) for continuous, binary, ordinal, and nominal data with various types of extensions provide users more comprehensive information embedded in complex high dimensional data than conventional EDA tools such as boxplot, scatterplot, with dimension reduction techniques such as principal component analysis and multiple correspondence analysis.

In this talk I'll summarize our works on creating MV environment for conducting statistical analyses and introducing statistical concepts into MV environment for visualizing more versatile and complex data structure. Many real world examples will be demonstrated in this talk for illustrating the strength of MV for visualizing all types of datasets collected from scientific experiments and social surveys.

◆ Lecture 4

Dr. I-Ping Tu

Research Fellow & Deputy Director, Institute of Statistical Science, Academia Sinica, Taiwan

Research fields: Statistical Analysis for Biological Data Analysis, Dimension Reduction, High-dimensional Data Analysis

Title: On Principal Component Analysis and Multilinear Principal Component Analysis

PCA is arguably the most popular dimension reduction method. It has at least four advantages to gain its popularity. First, its statistical concept is very intuitive that it searches new variables, through orthogonal linear transformation, to capture the most variation of the data. Second, PCA is very friendly to users that most statistical or mathematical packages offer PCA with a very simple function format. Third, it has broad applications as long as the data can be presented as a matrix

(variable vs sample) form. Many modern data analysis tools which target on large data set will adopt PCA as a built-in tool to reduce the dimension of data and thus facilitate the processing, where tSNE is a typical example. Forth, when data complexity goes beyond the limit of PCA, a modified version of PCA has good chance to be invented. For example, multilinear PCA (MPCA) has been developed for tensor structure data. In this talk, we will give a brief introduction to PCA and its extension to MPCA and discuss their computational complexity. Some examples will be presented to help the audience develop intuitive understanding of PCA and MPCA.

◆ Lecture 5

Dr. Su-Yun Huang

Research Fellow, Institute of Statistical Science, Academia Sinica, Taiwan

Research fields: Dimension Reduction, High-dimensional Data Analysis, Machine Learning, Robust Statistical Inference

Title: Kronecker envelope PCA and Its Application to Cryo-EM Image Data Processing

Multilinear principal component analysis (MPCA) is an extension of PCA to tensor (or array) data. It preserves the natural Kronecker product structure of observations when searching for principal components. The main advantage of preserving the Kronecker product structure is the parsimonious usage of parameters in specifying the principal component subspaces, which mitigates the adverse influence of high-dimensionality, and hence, leads to efficiency gain in estimation and prediction. Note that PCA will convert possibly correlated variables to uncorrelated ones. However, it is not the case for MPCA. In some applications, decorrelation is necessary. Hence, the Kronecker envelope PCA is introduced. In this talk, we will give a friendly introduction to the basic concept and technique for the Kronecker envelope PCA. The audience will see the rationale for its success from the statistical point of view. Application to cryo-EM image processing will be presented.