High-Dimensional Data Analysis: The Curses and Blessings of Dimensionality - A Brief Summary

Jingqian Xu
Department of Statistics
University of Connecticut

October 2022

1 Introduction

1.1 Abstract

- The coming century is the century of data and high-dimensional data analysis will be important activity in the coming the 21st century.
- In comparison with the traditional statistical analysis, the trend today is towards more observations but even more radically larger numbers of variables.
- Since mathematicians attack the curse of dimensionality in specific ways, create many of most commonplace tools in everyday data analysis, they play key role in high-dimensional data analysis.
- Ongoing developments in high-dimensional data analysis may lead mathematicians to study new problems and that many of the problems of low dimensional data analysis are unsolved and are similar to problems in harmonic analysis which have only recently been attacked

1.2 Data Analysis Today

- Information Technology (IT) has developed rapidly in recent years and lead to the work way change of statisticians.
- The data analysis community today is distinct form the mathematical tradition in three different ways.
- The largest contribution of statistics to data analysis have been in the formalization of information technology for data analysis, in form of software packages, like packages in R and Python programming language.

2 Major Recent Data Trends

2.1 Data Trend and Extrapolation of Trends

- Huge investments have been made in various data gathering and data processing on profound fields.
- In biotech field, more and more massive Biotech databases will be compiled like genomics data to decode human genome and make it available to biological researchers who translate these data into an understanding of protein expression.
- In finance field, high frequency financial data becoming available lead to the availability of large scale historical data and data mining based on it.
- Widely use of satellite imagery in natural resource discovery and agriculture.
- Consumer financial data is more important for advertiser to cut cost and increase the effect of advertisements.

2.2 Data Supply and Demand

- Our society will more and more think of itself as a data-driven society.
- Open availability of data is another trend.
- The job like accessing the data, simulating the system and visualizing the system will be easier in the future.

3 Data Matrices

3.1 Application of Large Data Matrices

- There are a broad range of applications where people can have N by D data matrices.
- Application fields include document retrieval by web searching, study about the association between genes and various diseases, recommendation system using consumer preference data.

4 Data Analysis

- A number of fundamental tasks of data analysis include classification, regression, latent variables analysis and clustering.
- For Classification, one example is predicting bankruptcy of clients based on their historical bankruptcy data.
- For Regression, predicting the variability of exchange rates given recent exchange rates.
- For latent variables analysis, we hope to find the underlying latent variables that are responsible for essentially the structure of array X. Principal Component Analysis (PCA) is example of latent variables analysis, it aims to find the variables explain the most variability of responds. It reduces the dimension of high-dimension data.
- Clustering analysis is used in latent semantic indexing, and analysts seek arrangement of documents so that nearby documents are similar.

5 High-Dimensionality

- For many high dimensional data problem today, we don't know which variables to measure in advance.
- The main difference between classic and post-classical world of data analysis is that the number of variables now is larger than that of observations and this leads to malfunction of tools we use in classic world.
- In biotech research, relatively few patients with a given genetic disease is an example.

5.1 Curse of Dimensionality

- General speaking, the term "Curse of Dimensionality" describe how difficult it was to perform high-dimensional numerical integration.
- The issues include data quality, never enough data, noise accumulation and spurious correlation.
    1. Never enough data refers we never have enough data to cover every part of high-dimensional input space. In other words,the very slow rate of convergence in high dimensions is the ugly head of the curse of dimensionality.
    2. Boundary concentration means all the volume inside a hyper cubic region of input space lies closer to the boundary or surface of the hypercube as the number of dimensions grows larger
    3. Noise accumulates as the increase of dimensions.

5.2 Blessings of Dimensionality

- Increasing in dimensionality is often helpful to mathematical analysis.
- Concentration of measure and Dimension asymptotics are two blessings of high-dimensionality.
- Based on these two bleesings, model selection, asymptotics for principal components are examples showing how each of these blessings comes up in high-dimensional data analysis.

6 Conclusion

- Data analysis today is not an unsophisticated activity carried out by hand; it is much more ambitious and an intellectual force to be reckoned with.
- Being different from Tukey's implicit points, data analysis today has now reached a point of sophistication where it can challenge and enrich mathematical discourse.