BIG DATA AND HIGH DIMENSIONAL DATA ANALYSIS

B.L.S. Prakasa Rao

CR RAO ADVANCED INSTITUTE OF MATHEMATICS, STATISTICS AND COMPUTER SCIENCE (AIMSCS)
University of Hyderabad Campus
GACHIBOWLI, HYDERABAD 500046

e-mail:blsprao@gmail.com

BIG DATA

- Statistical Methods have been applied in sciences such as

  Biometrics
  Bio-pharmaceutical Science
  Epidemiology
  Physical and Engineering Sciences
  Environmental and Ecology

BIG DATA

- and in business sciences such as

  Business and Economic Statistics
  Quality and Productivity and Marketing

  and for policy making in

  Official Statistics
  Health Policy Statistics
  Defence and National security

  and in

  Social Sciences
  Survey Research

BIG DATA

- Statistics provides methodology for use in these diverse fields and helps to solve problems. An important and unique feature of being a statistician is that you may collaborate in any discipline and you may change fields of application depending on the need and demand. Statistics improves human welfare by its contributions to all fields. It serves society. Some statisticians are doing research or involved in teaching, others are applying the methodology or developing software tools for applying the tools of statistics.

BIG DATA

- The present century is the century of data. We are collecting and processing data of all kinds on scales unimaginable earlier. Examples of such data are internet traffic, financial tick-by-tick data and DNA Microarrays which feed data in large streams into scientific and business bases worldwide.

  What is BIG DATA?

BIG DATA

- BIG DATA is relentless. It is continuously generated on a massive scale. It is generated by online interactions among people, by transactions between people and systems and by sensor enabled instrumentation.

BIG DATA

- BIG DATA can be related, linked and integrated to provide highly detailed information. Such a detail makes it possible:

  For Banks to introduce individually tailored services

  For Health care providers to offer personalized medicines

  For Public safety departments to anticipate crime in targeted areas

  Big Data is creating new business and transforming traditional markets. Big Data is a challenge to statistical community.

BIG DATA

- BIG DATA is a class of data sets so large that it becomes difficult to process using the standard methods of data processing. The problems of such data include collection, storage,search, sharing, transfer, visualization and analysis. An important advantage of analysis of BIG DATA is the additional information that can be obtained from a single large set as opposed to separate smaller sets. BIG DATA allows correlations to be found, for instance, to spot business trends.

BIG DATA

- BIG DATA involves increasing volume, that is, amount of data, velocity, that is speed at which data is in and out, and variety, that is, range of data types and sources. It is high in volume, high in velocity and also high in variety. This requires new forms of processing for decision making. BIG DATA can make important contributions to international development as it gives a cost-effective way to improve decision making in areas such as health care, employment, economic productivity, resource management and natural disasters. There are of course concerns such as privacy with collection of such data.

BIG DATA

- BIG DATA has unique features that are not shared by the traditional data sets. It is characterized by massive sample size and high-dimensionality. Massive sample sizes allows us to discover hidden patterns associated with small sub-populations. Modeling intrinsic heterogeneity of BIG DATA requires sophisticated methods. High-dimensionality and BIG DATA have special features such as noise accumulation, spurious correlation and incidental endogenity.

BIG DATA

- BIG DATA is a combination of many data sources coming from different sub-populations. Each sub population may exhibit features which are not shared by others. A mixture model may be appropriate to model BIG DATA. Analyzing such data requires simultaneous estimation or testing of several parameters. Errors associated with estimation of these parameters accumulate when a decision rule or prediction rule depend on these parameters. Such a noise accumulation effect is severe due to high-dimension and might even dominate the true signal. This is handled by sparsity assumption in high-dimensional data.

BIG DATA

- High dimensionality brings in spurious correlation due to the fact that many uncorrelated random variables may have high sample correlation coefficient in high-dimensions. Spurious correlation in turn lead to wrong inferences. Sheer size of BIG DATA leads to incidental endogenity. It is the existence of correlation between variables "unintentionally" as well as due to "high-dimensionality". It is like finding two persons in a large group of people who look alike but have no genetic relationship or meeting an acquaintance by chance in a big city. Endogenity occurs as a result of selection bias, measurement errors and omitted variables.

High Dimensional Data

- In traditional statistical data analysis, we think of observations of instances of a particular phenomenon. These observations being vector of values measured on several variables such as blood pressure, weight, height etc. In traditional statistical methodology, we assumed many observations and a few selected variables which are well chosen to explain the phenomenon.

High Dimensional Data

- The trend today is towards more observations, but even more so, to a very large number of variables—automatic, systematic collection of a large amount of detailed information about each observation. The observations could be curves, images or movies so that a single observation has dimension in the thousands or millions or billions while only tens or hundreds of observations for study. This is HIGH DIMENSIONAL DATA.

  Classical methods are not designed to cope with this kind of growth of dimensionality of the observation vector.

High Dimensional Data

- Traditionally, data analysis is a part of the subject of statistics with its basics in probability theory, decision theory and analysis. New sources of data —such as from satellites— generate automatically huge volumes of data whose summarization called for a wide variety of data processing and analysis tools. For such data, traditional ideas of mathematical statistics such as hypothesis testing and confidence intervals *do not help*.

RECENT DATA TRENDS

- Over the last twenty years, data, data management and data processing have become important factors. Large investments have been made in gathering various data and in data processing mechanisms. The information technology industry is the fastest growing and most lucrative segment of the world economy and much of the growth occurs in the development, management and warehousing of streams of data for scientific, medical, engineering and commercial purposes. Examples of such data are the following:

RECENT DATA TRENDS

- **Biotech data:** Every body is aware of the progress made in getting data about human genome. The genome is indirectly related to protein function and the protein function is indirectly related to the cell function. At each step, more and more massive data are compiled.

  Financial data: Over the last twenty years, high frequency financial data has become available.

RECENT DATA TRENDS

- **Satellite imagery**: Providers of satellite imagery have available a vast data bases of such images. National remote sensing data is one such with applications of such imagery include natural resource discovery and agriculture.

  Consumer financial data: Every transaction we make on the web, whether a visit, a search or a purchase is being recorded, correlated, compiled in data bases and sold and resold as advertisers scramble to correlate consumer action with pockets of demand for various goods and services.

RECENT DATA TRENDS

- How to store such data? Through data matrix.

DATA MATRIX

- **DATA MATRIX:** A data matrix is a rectangular array with N rows and p columns, the rows giving different *observations* or *individuals* and the columns giving different *attributes* or *variables*.

  In a classical framework, N may be 25000 and $p = 100$.

  Let us see some recent examples indicating some applications of $N \times p$ data matrices.

DATA MATRIX

- (i) **Web-term document data:** Document retrieval by web searching is an important activity. One approach to document retrieval is the vector space model of information retrieval. Here one compiles *term-document matrices*, $N \times p$ arrays, where N is the number of documents which may be in millions and p the number of terms (words) is in tens of thousands and each entry in the array measures the frequency of occurrence of given terms in the given document in a suitable normalization. Each such request can be looked at as a vector of term frequencies.

DATA MATRIX

- (ii) **Sensor array data:** Consider a problem in neuroscience. An array of p sensors may be attached to the scalp with each sensor recording N observations over a period of seconds at a rate of ten thousand samples per second. One hopes to use such data to study the response of human neuronal system to external stimuli.

DATA MATRIX

- (iii)**Gene expression data:** Here we obtain data on the relative abundance of p genes in each of N different cell lines. The goal is to term which genes are associated with various diseases or other states associated with cell lines.

DATA MATRIX

- (iv)**Consumer preference data:** These are used to gather information about borrowing and shopping behaviour of customers and use this to modify presentation of information to users. Examples include recommendation system like Netflix. Each consumer is asked to rate 100 films; based on that rating, one consumer is compared to another customer with similar preferences and predictions are made of other movies which might be of interest to one consumer based on experiences of other customers who viewed or rated those movies. Here we have a rectangular array giving responses of N individuals on p movies with N possibly running into millions and p in the hundreds or thousands.

DATA MATRIX

- (v)**Consumer financial history:** Example of credit card transaction records on $N = 250,000$ consumers where several dozen variables p are available on each consumer. Here p may be in the hundreds.

IMPACT ON COMPUTING INFRASTRUCTURE

- Massive sample size of BIG DATA challenges the traditional computing infrastructure. BIG DATA is highly dynamic and impossible to store in a central data base. The basic approach to such data is to divide and analyze. The idea is to partition a large problem into more tractable subproblems. Each subproblem is attacked in parallel by different processing units. Results from sub-problems are combined to obtain the final result. When a large number of computers are connected to process large computing tasks, it is possible that some of them fail. Given the size of computing, we may want to distribute the work evenly between computers for a balanced work output. Designing very large scale, highly adaptive and fault-tolerant computing systems is a difficult problem.

IMPACT ON COMPUTING METHODS

- Analysis of BIG DATA leads to problems of large scale optimization. Optimization with a large number of variables is not only expensive but also suffers from slow numerical rates of convergence. Methods for implementation of large-scale non-smooth optimization procedures have to be investigated.

IMPACT ON COMPUTING METHODS

- It is often computationally infeasible to directly make inferences based on the raw data. To handle BIG DATA from both the statistical and the computational views, the idea of dimension reduction is an important step before start of processing of a BIG DATA.

HIGH DIMENSIONAL DATA ANALYSIS

- (A)**Classification:** In classification, one of the p variables is an indicator of class membership. For instance, in a consumer financial data base, most of the variables measure consumer payment history and one of the variables indicates whether the consumer has declared bankruptcy. The analyst would like to predict bankruptcy from credit history. Many approaches are suggested for classification ranging from identifying hyperplanes which partition the sample space in non-overlapping groups to k-nearest neighbour classification.

HIGH DIMENSIONAL DATA ANALYSIS

- (B)**Regression:** In regression setting, one of the p variables is a quantitative response variable. Examples include, for instance in financial database, the variability of exchange rates today given recent exchange rates.

HIGH DIMENSIONAL DATA ANALYSIS

- **Tools for regression analysis:**

HIGH DIMENSIONAL DATA ANALYSIS

- (i) Linear regression modeling:

$$X_{i1} = a_0 + a_2 X_{i2} + \cdots + a_p X_{ip} + Z_i$$

X_{i1} Response ; $X_{i2}, \ldots, X_{ip}$ predictors;

HIGH DIMENSIONAL DATA ANALYSIS

- (ii) Nonlinear regression modeling:

$$X_{i1} = f(X_{i2}, \ldots, X_{ip}) + Z_i.$$

HIGH DIMENSIONAL DATA ANALYSIS

- (iii) Latent variable analysis:

  In latent variable modeling, it is proposed that

$$X = AS$$

  where X is a vector-valued observable, S is a vector of unobserved latent variables and A is linear transformation converting one into the other. It is hoped that a few underlying latent variables are responsible for essentially the structure we see in the array X and, by uncovering those variables, we get important insights.

HIGH DIMENSIONAL DATA ANALYSIS

- Principle Component Analysis (PCA) is an example. Here one takes the covariance matrix C of the observables X, obtains the eigenvectors, which will be orthogonal, places them as columns in an orthogonal matrix U and defines

$$S = U'X.$$

HIGH DIMENSIONAL DATA ANALYSIS

- Here we have the latent variable form with $A = U$. This technique is widely used for data analysis in sciences, engineering and commercial applications. The mathematical reason for this approach is that the projection on the space spanned by the first k eigenvectors of C gives the best rank k approximation to the vector X in a mean square sense.

HIGH DIMENSIONAL DATA ANALYSIS

- (C)**Clustering:** Here one seeks to arrange an unordered collection of objects in a fashion so that nearby objects are similar.

COMMENTS

- Huge resources are now invested on a global scale on the collection of massive data bases. In spite of large variety of phenomena that data can be measuring from astronomical to financial data, there are some standard forms that data often take and we can ask some standard questions on this data.

COMMENTS

- There is a large group of people working to answer practical problems based on real data. They do not use heavy mathematical machinery but use computer simulation.

COMMENTS

- For such data analysts:

  (i) there is no formal definition of the phenomenon in terms of a carefully studied mathematical model;

  (ii) there is no formal derivation of the proposed method suggesting that it is in some sense naturally associated with the phenomenon to be treated; and

  (iii) there is no formal analysis justifying the apparent improvement in simulation.

COMMENTS

- This is far from the tradition of mathematicians and of statisticians.

HIGH DIMENSIONALITY

- **High Dimensionality: We are now in the era of massive automatic data collection, systematically obtaining many measurements, not knowing which ones will be relevant to the phenomenon of interest. Our aim is to find the relevant variables from a large number of them.**

HIGH DIMENSIONALITY

- **This is different from the past where it was assumed that one was dealing with a few well chosen variables, for example, using scientific knowledge to measure just the right variables in advance. It has become much cheaper to gather data than to worry much about what data to gather.**

HIGH DIMENSIONALITY

- **The basic methodology which was used in classical statistical inference is no longer applicable. The classical theory was based on the assumption that $p < N$ and N tends to ∞. Many of the results assume that the observations are multivariate normal. All the results fail if $p > N$. In the high dimensional case, we might even have p tends to ∞ with N fixed.**

HIGH DIMENSIONALITY

- For many types of event we can think of, we have potentially a very large number of measurables quantifying that event but a relatively few instances of that event.

  Example : Many genes and relatively few patients with a given genetic disease.

HIGH DIMENSIONALITY

- (A)**Curse of dimensionality:** The phrase "curse of dimensionality" was introduced by Richard Bellman who developed the method of dynamic programming for study of some optimization problems. In optimization: if we must minimize a function f of d variables and we know that it is Lipschitz, that is,

$$|f(x) - f(y)| \leq C||x - y||, x, y, \in R^d,$$

then we need to order $(\frac{1}{\epsilon})^d$ evaluations on a grid in order to approximate the minimizer within error ϵ.

HIGH DIMENSIONALITY

- (B)**Blessings of dimensionality:**There are practical difficulties caused by increase in dimensionality but there are theoretical benefits due to probability theory. The regularity of having many "identical" dimensions over which one can "average" is a fundamental tool.

HIGH DIMENSIONALITY

- (i)**Dimension asymptotics:** One use is that we can obtain results on the phenomenon by letting the dimension go to infinity.

HIGH DIMENSIONALITY

- (ii)**Approach to continuum:** Many times we have high dimensional data because the underlying objects are really continuous space or continuous phenomena; there is an underlying curve or image that we are sampling such as in functional data analysis or image processing. Since the measured curves are continuous, there is an underlying compactness to the space of observed data which will be reflected by an approximate finite-dimensionality and an increasing simplicity of analysis for large d.

IMPACT ON DATA ANALYSIS

- **"Curse of dimensionality" in nonparametric estimation**:
  Suppose we have a data set with d variables and the first one is dependent on the others through a model of the form

$$X_{i1} = f(X_{i2}, \ldots, X_{id}) + \epsilon_i. \quad (1)$$

  Suppose that f is unknown and we are not willing to specify a specific model for f such as a linear model. Instead, we are willing to assume that f is a Lipschitz function of these variables and that ϵ_i are i.i.d. $N(0, 1)$ variables .

IMPACT ON DATA ANALYSIS

- How does the accuracy of the estimate depend on N, the number of observations in our data set?

IMPACT ON DATA ANALYSIS

- Let Θ be the class of functions f which are Lipchitz on $[0,1]^d$. It can be shown that

$$\sup_{f \in \Theta} E[\hat{f} - f(x)]^2 \geq CN^{-2/(2+d)}$$

(cf. Ibragimov and Khasminskii (1981)).

IMPACT ON DATA ANALYSIS

- This lower bound is non-asymptotic. How much data do we need in order to obtain an estimate of f accurate to within $\epsilon = .1$? We can answer this question by the above result and we need a very large sample and the sample size increases as the dimension d increases. This is the curse of dimensionality.

IMPACT ON DATA ANALYSIS

- Suppose we have a linear regression problem where there is a dependent variable X_{i1} which we want to model as a linear function of $X_{i2}, \ldots, X_{id}$ as in (1). However d is *very large* and let us suppose that we are in a situation where there are thought to be only a few relevant variables but we do not know which are relevant. If we leave out too many relevant variables in the model, we might get very poor performance. For this reason, statisticians , for a long time, considered model selection by searching among subsets of the possible explanatory variables trying to find just a few variables among the many which will adequately explain the dependent variable.

IMPACT ON DATA ANALYSIS

- Since 1970's, one approach to model selection was to optimize the complexity penalized form over subset models: minimize

$$RSS(Model) + \lambda \ \ (Modelcomplexity)$$

where RSS is the residual sum of squares of the residuals $X_{i1} - Model_{i1}$ and model complexity is the number of variables $X_{i1}, \ldots, X_{id}$ used in forming the model.

IMPACT ON DATA ANALYSIS

- One choice for $\lambda = 2\sigma^2$ where σ^2 is the assumed variance of the noise in (1). The idea is to impose a cost on large complex models. More recently another choice suggested for λ is $\lambda = 2\sigma^2 \log d$ (Johnstone (1998)). With logarithmic penalty function, one can mine the data to one's taste while controlling the risk of finding spurious structure.

IMPACT ON DATA ANALYSIS

- Asymptotics of principal component:
  This is another instance of "blessing of dimensionality"–
  -that results for higher dimensions are easier to derive than for moderate dimensions. Suppose

$$X^{(i)} = (X_{i1}, \ldots, X_{id}) \simeq N(0, \Gamma).$$

  We are interested in knowing whether $\Gamma = I$ or $\Gamma \# I$.

IMPACT ON DATA ANALYSIS

- This problem can be rephrased as $\lambda_1 = 1$ or $\lambda_1 > 1$ where λ_1 is the largest eigenvalue of C. It is natural to develop a test based on ℓ_1, the largest eigenvalue of the empirical covariance matrix

$$C = N^{-1} X'X.$$

IMPACT ON DATA ANALYSIS

- This leads to finding the null distribution of ℓ_1. Exact formula for this distribution is known but not useful in practice. Suppose d is large and we are in the setting of many dimensions and many variables.

IMPACT ON DATA ANALYSIS

- What is the behaviour of $C_{d,N}$ the largest eigenvalue of C as $\frac{d}{N} \to \beta$? This is a problem discussed in "Random Matrix Theory". Iain Johnstone obtained asymptotic results for the largest eigenvalue in the statistical setting.

IMPACT ON DATA ANALYSIS

- These results are found to be useful even for dimension $d = 6$. Hence, in some cases, we can obtain useful results for moderate dimensions from high-dimensional analysis.

LASSO ESTIMATOR

- Modeling high-dimensional data is challenging. For a continuous response variable Y, a simple but very useful approach is given by a linear model

$$Y_i = \sum_{j=1}^{p} \beta_j X_i^{(j)} + \epsilon_i, i = 1, \ldots, n$$

where $\epsilon_i, i = 1, \ldots, n$ are zero mean independent identically distributed errors. The unusual aspect of this model is p, the number of covariates is very large compared to n, the number of observations.

LASSO ESTIMATOR

- Since $p > n$, the ordinary least squares estimator is not unique and will heavily overfit the data. Tibshirani (1996) proposed an estimator called LASSO (Least Absolute Shrinkage and Selection Operator). It has become very popular for high-dimensional data analysis.

DIRECTION OF PROGRESS

- **DIRECTION OF PROGRESS:** High dimensional analysis are leading to challenges in both mathematics and statistics. New subjects such as high-dimensional approximate linear algebra and high dimensional approximation theory using bases such as wavelets are being developed to face such challenges.

DATA SCIENTISTS

- **DATA SCIENTISTS:** DATA SCIENTISTS are people who draw insights from large amount of data. They are expected to have expertise in statistical modeling and machine learning; specialized programming skills and a solid grasp of business environment they are working in. Data science is an interdisciplinary blend of statistical, mathematical and computational sciences.

e-COMMERCE

- **e-COMMERCE:** Electronic commerce (e-Commerce) has experienced an extreme surge of popularity in recent years. It transformed the economy, eliminated borders, opened the door to many innovations and created new ways in which consumers and business interact. E-commerce transactions include activities such as online buying, selling or investing; e-book stores , e-grocers, web-based reservation system and ticket purchasing, online banking etc.

e-COMMERCE

- **e-COMMERCE:** Due to the availability of massive amounts of data, empirical research is thriving. E-commerce data arrives with many new statistical issues and challenges, starting from data collection to data exploration and to analysis and modeling. e-commerce data is usually very large in both dimensionality and size. It is often a combination of longitudinal and cross-sectional data. Most of it is observational. Data privacy protection is a major issue for e-commerce.

e-COMMERCE

- **e-COMMERCE:** There is a fairly new branch of statistics called "Functional Data Analysis" (FDA) which is being used for analyzing the data obtained via e-commerce. Here the objects of interest are a set of curves or shapes and in general a set of functional observations. In classical statistics, the interest is around data values or data vectors. e-commerce is creating new challenges.

e-COMMERCE

- **e-COMMERCE:** The first step in any FDA consists of recovering the underlying functional object from the observed data. This involves choosing a suitable family of basis functions to reduce the dimensionality of the problem which in turn depends on

  (i) the nature of the data;

  (ii) the level of smoothness that the application warrants;

  (iii) the aspects of the data we want to study;

  (iv) the size of the data; and

  (v) the type of analysis we plan to perform.

STATISTICIANS

- **STATISTICIANS:** Statistical thinking sets us the statisticians apart from others in three ways:

  (1) Like others, we look for features in large data - but we guard against FALSE DISCOVERY, BIAS AND CONFOUNDING;

  (2) Like others, we build statistical models that explain, predict and forecast- but we qualify their use with MEASURES OF UNCERTAINTY;

  (3) Like others, we work with available data- but we also design studies to produce data that have RIGHT INFORMATION CONTENT

  That is the difference between us the STATISTICIANS and them DATA MINERS/DATA SCIENTISTS.

STATISTICIANS

- **STATISTICIANS:** In his recent address to ASA, its former president Robert Rodriguez said "Young statisticians will need to have continuing education in statistical theory, methodology and applications. They will need to know about new data, new problems and new computational technology. They need to develop skills in collaboration, communication and leadership. It is necessary to have a broad view of what is being accomplished as a statistician."

STATISTICIANS

- **STATISTICIANS:** "It is more than designing an experiment and analyzing data. Ideas should transform business, influence the decision making in formulating public policy. The future of statistics as a profession is unlimited as we grow the opportunity for statistics to meet the needs of a data dependent society".

  This is true for Indian Statisticians too!!!

- THANK YOU