Brief Notes On Big Data

B.L.S. Prakasa Rao

Abstract - Without any doubt, the most discussed current trend in statistics is BIG DATA. Different people think of different things when they hear about Big Data. For statisticians, how to get usable information out of data bases that are so huge and complex that many of the traditional or classical methods cannot handle? For computer scientists, Big Data poses problems of data storage and management, communication and computation. For citizens, Big Data brings up questions of privacy and confidentiality. This brief notes gives a cursory look on ideas on several aspects connected with collection and analysis of Big Data. It is a compilation of ideas from different people, from various organizations and from different sources online. Our discussion does not cover computational aspects in analysis of Big Data.

◆

[Excerpted from C R RAO AIMSCS Lecture Notes Series]

1 What is BIG DATA?

(Fan et al. (2013))
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 equipment such as aerial sensing technologies (remote sensing), information-sensing mobile devices, wireless sensor networks etc.

Big Data is relatable. It can be related, linked and integrated to provide highly detailed information. Such a detail makes it possible, for instance, for banks to introduce individually tailored services and for health care providers to offer personalized medicines.

Big data is a class of data sets so large that it becomes difficult to process it using standard methods of data processing. The problems of such data include capture or collection, curation, storage, search, sharing, transfer, visualization and analysis. Big data is difficult to work with using most relational data base management systems, desktop statistics and visualization packages. Big Data usually includes data sets with size beyond the ability of commonly used software tools. When do we say that data is a Big Data? Is there a way of quantifying the data?

Advantage of studying Big Data is that additional information can be derived from anal- ysis of a single large set of related data, as compared to separate smaller sets with the same total amount of data, allowing correlations to be found. For instance, analysis of a large data in marketing a product will lead to information on business trend for that product. Big Data can make important contributions to international development. Analysis of Big Data leads to a cost-effective way to improve decision making in important areas such as health care, economic productivity, crime and security, natural disaster and resource management.

Large data sets are encountered in meteorology, genomics, biological and environmental research. They are also present in other areas such as internet search, finance and business informatics. Data sets are big as they are gathered using sensor technologies. There are also examples of Big Data in areas which we can call Big Science and in Science for research. These include“Large Hadron Collision Experiment” which represent about 150 million sensors delivering data at 40 million times per second.

There are nearly 600 million collisions per second. After filtering and not recording 99.999%, there are 100 collisions of interest per second. The Large Hadron collider experiment generates more than a petabyte (1000 trillion bytes) of data per year. Astronomical data collected by Sloan Digital Sky Survey (SDSS) is an example of Big Data. Decoding human genome which took ten years to process earlier can now be done in a week.

This is also an example of Big Data. Human genome data base is another example of a Big Data. A single human genome contains more than 3 billion base pairs. The 1000 Genomes pro ject has 200 terabytes (200 trillion bytes) of data. Human brain data is an example of a Big Data. A single human brain scan consists of data on more than 200,000 voxel locations which could be measured repeatedly at 300 time points.

For Government, Big Data is present for climate simulation and analysis and for national security areas. For private sector companies such as Flipkart and Amazon, Big Data comes up from millions of backend operations every day involving queries from customer transactions, from vendors etc.

Big Data sizes are a constantly moving target. It involves increasing volume (amount of data), velocity (speed of data in and out) and variety (range of data types and sources). Big Data are high volume, high velocity and/or high variety information assets. It requires new forms of processing to enable enhanced decision making, insight discovery and process optimization.

During the last fifteen years, several companies abroad are adopting to data-driven ap- proach to conduct more targeted services to reduce risks and to improve performance. They are implementing specialized data analytics to collect, store, manage and analyze large data sets. For example, available financial data sources include stock prices, currency and deriva- tive trades, transaction records, high-frequency trades, unstructured news and texts, consumer confidence and business sentiments from social media and internet among others. Analyzing these massive data sets help measuring firms risks as well as systemic risks. Anal- ysis of such data requires people who are familiar with sophisticated statistical techniques such as portfolio management, stock regulation, proprietary trading, financial consulting and risk management.

Big Data are of various types and sizes. Massive amounts of data are hidden in social net works such as Google, Face book, Linked In , You tube and Twitter. These data reveal numerous individual characteristics and have been exploited. Government or official statistics is a Big Data. There are new types of data now. These data are not numbers but they come in the form of a curve (function), image, shape or network. The data might be a ”Functional Data” which may be a time series with measurements of the blood oxygenation taken at a particular point and at different moments in time. Here the observed function is a sample from an infinite dimensional space since it involves knowing the oxidation at infinitely many instants. The data from e-commerce is of functional type, for instance, results of auctioning of a commodity/item during a day by an auctioning company. Another type of data include correlated random functions. For instance, the observed data at time t might be the region of the brain that is active at time t. Brain and neuroimaging data are typical examples of another type of functional data. These data is acquired to map the neuron activity of the human brain to find out how the human brain works. The next-generation functional data is not only a Big Data but complex.

Examples include the following: (1) Aramiki,E; Maskawa, S. and Morita, M. (2011) used the data from Twitter to predict influenza epidemic; (2) Bollen, J., Mao, H. and Zeng, X. (2011) used the data from Twitter to predict stock market trends.

Social media and internet contains massive amounts of information on the consumer preferences leading to information on the economic indicators, business cycles and political attitudes of the society.

Analyzing large amount of economic and financial data is a difficult issue. One important tool for such analysis is the usual vector auto-regressive model involving generally at most ten variables and the number of parameters grows quadratically with the size of the model. Now a days econometricians need to analyze multivariate time series with more than hundreds of variables.

Incorporating all these variables lead to over-fitting and bad prediction. One solution is to incorporate sparsity assumption. Another example, where a large number of variables might be present, is in portfolio optimization and risk management. Here the problem is estimating the covariance and inverse covariance matrices of the returns of the assets in the portfolio. If we have 1000 stocks to be managed, then there will be 500500 covariance parameters to be estimated. Even if we could estimate individual parameters, the total error in estimation can be large (Pourahmadi: Modern methods in Covariance Estimation with High-Dimensional Data (2013), Wiley, New York).

2 Some issues with Big Data

(cf. Fokoue (2015); Buelens et al. (2014))

(i) Batch data against incremental data production: Big Data is delivered generally in a sequential and incremental manner leading to online learning methods. Online algorithms have the important advantage that the data does not have to be stored in memory. All that is required is in the storage of the built model at the given time in the sense that the stored model is akin to the underlying model. If the sample size n is very large, the data cannot fit into the computer memory and one can consider building a learning method that receives the data sequentially or incrementally rather than trying to load the complete data set into memory. This can be termed as sequentialization. Sequentialization is useful for streaming data and for massive data that is too large to be loaded into memory all at once.

(ii) Missing values and Imputation schemes: In most of the cases of massive data, it is quite common to be faced with missing values. One should check at first whether they are missing systematically, that is in a pattern, or if they are missing at random and the rate at which they are missing. Three approaches are suggested to take care of this problem: (a) Deletion which consists of deleting all the rows in the Data matrix that contain any missing values ; (b) central imputation which consists of filling the missing cells of the Data matrix with central tendencies like mean, mode or median; and (c) Model-based imputation methods such as EM-algorithm.

(iii) Inherent lack of structure and importance of preprocessing: Most of the Big Data is unstructured and needs preprocessing. With the inherently unstructured data like text data, the preprocessing of data leads to data matrices, whose entries are frequencies of terms in the case of text data, that contain too many zeroes leading to the sparsity problem. The sparsity problem in turn leads to modeling issues.

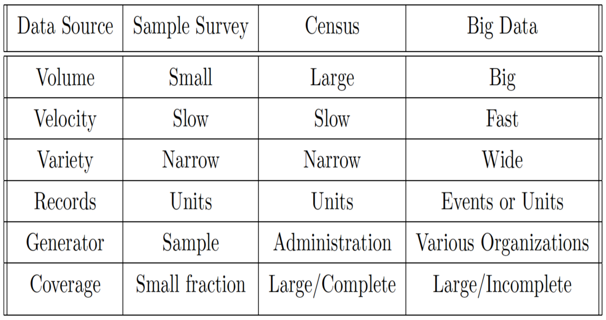
(iv) Homogeneity versus heterogeneity: There are massive data sets which have input space homogeneous, that is, all the variables are of the same type. Examples of such data include audio processing, video processing and image processing. There are other types of Big Data where the input space consists of variables of different types. Such types of data arise in business, marketing and social sciences where the variables can be categorical, ordi- nal, interval, count and real-valued.

(v) Differences in measurement: It is generally observed that the variables involved are measured on different scales leading to modeling problems. One way to take care of this problem is to perform transformations that pro ject the variables onto the same scale. This is done either by standardization which leads all the variables to have mean zero and variance one or by unitization which consists in transform the variables so that the support for all of them is the unit interval [0,1].

(vi) Selection bias and quality: When Big Data are discussed in relation to official statis- tics, one point of criticism is that Big Data are collected by mechanisms unrelated to prob- ability sampling and are therefore not suitable for production of official statistics. This is mainly because Big Data sets are not representative of a population of interest. In other words, they are selective by nature and therefore yield biased results. When a data set be- comes available through some mechanism other than random sampling, there is no guarantee what so ever that the data is representative unless the coverage is full. When considering the use of Big Data for official statistics, an assessment of selectivity has to be conducted. How does one assess selectivity of Big Data?

(vii) No clarity of target population: Another problem of Big Data dealing with official statistics is that many Big data sources contain records of events not necessarily directly associated with statistical units such as household, persons or enterprizes. Big Data is often a by-product of some process not primarily aimed at data collection. Analysis of Big Data is data-driven and not hypothesis-driven. For Big Data, the coverage is large but incomplete and selective. It may be unclear what the relevant target population is.

(viii) Comparison of data sources: Let us look at a comparison of different data sources for official statistics as compared to Big Data. (Ref: Buelens et al. (2014))



Data Source	Sample Survey	Census	Big Data
Volume	Small	Large	Big
Velocity	Slow	Slow	Fast
Variety	Narrow	Narrow	Wide
Records	Units	Units	Events or Units
Generator	Sample	Administration	Various Organizations
Coverage	Small fraction	Large/Complete	Large/Incomplete

For Big Data dealing with the official statistics, there are no approaches developed till now to measure the errors or to check the quality. It is clear that bias due to selectivity has role to play in the accounting of Big Data.

(ix) Use of Big Data in official statistics:

(a) Big Data can be the single source of data for the production of some statistic about a population of interest. Assessing selectivity of the data is important. Correcting for selectiv- ity can some times be achieved by choosing suitable method of model-based inference (Leo Breiman (2001), Statistical Science, 16, 199-231). These methods are aimed at predicting values for missing/unobserved units. The results will be biased if specific sub-populations are missing from the Big Data set.

(b) Big Data set can be used as auxiliary data set in a procedure mainly based on a sam- ple survey. The possible gain of such an application for the sample survey is likely reduction in sample size and the associated cost. Using small area models, the Big Data can be used as a predictor for survey based measurement.

(c) Big Data mechanism can be used as a data collection strategy for sample surveys.

(d) Big Data may be used irrespective of selectivity issues as a preliminary survey. Find- ings obtained from Big Data can be further checked and investigated through sample surveys.

3 Computing issues for Big Data

(Fan et al. (2013))

As was mentioned earlier, the massive or very large sample size of Big data is a challenge for traditional computing infrastructure. Big Data is highly dynamic and not feasible or possible to store in a centralized database. The fundamental approach to store and process such data is to ”divide and conquer”. The idea is to partition a large problem into more tractable and independent sub-problems. Each subproblem is tackled in parallel by different processing units. Results from individual sub-problems are then combined to get the final result. ”Hadoop” is an example of basic software and programming infrastructure for Big Data processing. ”MapReduce” is a programming model for processing large data sets in a parallel fashion. ”Cloud Computing” is suitable for storing and processing of Big Data. We are not presenting the problems involved in storage and computation connected with Big Data in this brief notes.

References:

Aramiki, E., Maskawa, S., and Morita, M. (2011) Twitter catches the flu: Detecting in- fluenza epidemics using twitter. In Proceedings of the Conference on Empirical Methods in Natural Language Processing, pp. 1568-1576.

Benjamini, Y. and Hochberg, Y. (1995) Controlling the false discovery rate: a practical and powerful approach to multiple testing. Journal of the Royal Statistical Society, Series B, 57, 289-300.

Bollen, J., Mao, H., and Zeng, X. (2011) Twitter mood predicts the stock market. Journal of Computational Science, 2, 1-8.

Buelens, B., Daas, P., Burger, J., Puts, M. and van den Brakel, J. (2014) Selectivity of Big Data, Discussion Paper, Statistics Netherlands.

Fan Jianqing, Han Fang and Liu Han (2013) Challenges of Big Data analytics, arXiv:1308.1479v1 [stat.ML] 7 Aug 2013.

Fokoue, E. (2015) A taxonomy of Big Data for optimal predictive machine learning and data mining, arXiv.1501.0060v1 [stat.ML] 3 Jan 2015.

Leak, J. (2014) “Why big data is in trouble; they forgot about applied statistics”, ”“Simply Statistics”, May 7, 2014.

Pourahmadi, M. (2013) Modern Methods to Covariance Estimation with High-Dimensional Data, Wiley, New York.

Tibshirani, R. (1996) Regression analysis and selection via the Lasso, Journal of the Royal Statistical Society, Series B, 58, 267-288.

“Current trends and future challenges in statistics: Big Data” Statistics and Science: A Report of the London Workshop on future of the Statistical Sciences (2014), pp. 20-25.

Bhagavatula Lakshmi Surya Prakasa Rao won the Shanti Swarup Bhatnagar Prize for Science and Technology in Mathematical Science in 1982 and the Outstanding Alumni award from Michigan State University. He worked at the Indian Institute of Technology, Kanpur in the beginning of his career and later moved to Indian Statistical Institute, New Delhi. He was a Distinguished Scientist and Director of Indian Statistical Institute, Kolkata from 1992 to 1995. He also held visiting professorship at the University of California, Berkeley, University of Illinois, University of Wisconsin, Purdue University, University of California, Davis and University of Iowa. He was the Jawaharlal Nehru, and Dr. Homi J. Bhabha Chair Professor at the University of Hyderabad in 2006-08, and 2008–12, respectively. He is currently an Emeritus Professor at the Indian Statistical Institute and the Ramanujan Chair Professor at CR Rao Advanced Institute of Mathematics, Statistics and Computer Science in University of Hyderabad campus.