Embracing Statistical Challenges in the Information Technology Age

Bin Yu
Department of Statistics
University of California at Berkeley, CA
binyu@stat.berkeley.edu, www.stat.berkeley.edu/users/binyu

Abstract

Information Technology is creating an exciting time for statistics. In this article, we review the diverse sources of IT data in three clusters: IT core, IT systems, and IT fringe. The new data forms, huge data volumes, and high data speeds of IT are contrasted against the constraints on storage, transmission and computation to point to the challenges and opportunities. In particular, we describe the impacts of IT on a typical statistical investigation of data collection, data visualization, and model fitting, with an emphasis on computation and feature selection. Moreover, two research projects on network tomography and arctic cloud detection are used throughout the paper to bring the discussions to a concrete level.

1 Introduction

The Information Technology (IT) revolution has changed in fundamental ways how we live and work. More changes are on their way. Some are predictable and others surprising. The IT revolution has been centered around rapid advances in computer technologies. As a result, our data collection capabilities have increased tremendously over the last decade or so. These data are in diverse forms. The traditional numeric are still prevalent in science and engineering, while electronic documents (texts), sound files, images, and videos, and multi-media are typical IT data. Moreover, the ever-evolving computer technology is also changing how we communicate with each other and how we obtain information on topics from weather, to movies, to health information, and to scholarly papers.

All knowledge/information acquisition involves extraction and synthesis of useful information from data, interpreted in a broad sense. Statistics as a discipline has its primary role as assisting this knowledge/information acquisition process in a principled and scientific manner. IT data are massive or high-dimensional, no matter whether the forms are old (numeric) or new (text, images, videos, sound, and multi-media). Often IT data acquisition speed exceeds the speed of advances in storage, transmission and computation. The conflict of data explosion and (relative) constraints on storage, transmission, and computation gives rises to the many challenges and opportunities for statisticians. We believe that statistics could and should play a more significant role in knowledge/information acquisition for all fields in the IT age, provided that we identify fundamental

Report Documentation Page

Form Approved
OMB No. 0704-0188

1. REPORT DATE	2. REPORT TYPE	3. DATES COVERED
2006		**00-00-2006 to 00-00-2006**

4. TITLE AND SUBTITLE	5a. CONTRACT NUMBER
Embracing Statistical Challenges in the Information Technology Age	5b. GRANT NUMBER
	5c. PROGRAM ELEMENT NUMBER
6. AUTHOR(S)	5d. PROJECT NUMBER
	5e. TASK NUMBER
	5f. WORK UNIT NUMBER
7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES) **Department of Statistics,University of California at Berkeley,Berkeley,CA,94720**	8. PERFORMING ORGANIZATION REPORT NUMBER
9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)	10. SPONSOR/MONITOR'S ACRONYM(S)
	11. SPONSOR/MONITOR'S REPORT NUMBER(S)

12. DISTRIBUTION/AVAILABILITY STATEMENT
Approved for public release; distribution unlimited

13. SUPPLEMENTARY NOTES
The original document contains color images.

14. ABSTRACT

15. SUBJECT TERMS

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT **unclassified**	b. ABSTRACT **unclassified**	c. THIS PAGE **unclassified**		**24**	

problems to form a new core of statistical research and at the same time help solve data problems of subject matter importance.

In this paper we showcase the IT challenges and opportunities for statistics. Diverse IT areas are reviewed, and we share our IT research experience by giving examples and covering two projects of the author and co-workers to demonstrate the breath and variety of of IT problems and to ground the paper. The material covered reflects the author's, admittedly limited, research interests, even though we try to be broad.

The rest of the paper is organized as follows. Section 2 describes the varying forms of IT data in three clusters: IT core, IT systems, and IT fringe, and illustrates each cluster by concrete examples. The IT core areas provide new data forms: text, images, videos, and multi-media. The IT systems areas contain a mix of new and old forms. For example: the network management data are likely to be numeric, but in program debugging, the data are programs which are specialized texts. The IT fringe data are mainly numeric, but definitely often take the image form like in bioinformatics and remote sensing.

Section 3 examines the process of a statistical investigation impacted by the IT data demands. We emphasize the over-reaching iterative theme (Box, 1980 [4]) for any statistical investigation, and argue that this iterative theme becomes much more desired for IT data analysis. In particular, in Section 3.1, we cover issues related to data collection: data forms, difficulty of sampling, and modes of data collection (batch vs. on-line). Both modes are frequently necessary in an iterative fashion for solving a particular problem. To design a fast on-line algorithm, batch data analysis might be necessary to know what is the important features of the data to retain. Before batch data can be collected, data reduction or feature selection has to be carried out on-line to reduce the data volume so they can be stored. In Section 3.2, exploratory data analysis and data visualization are visited. We first point out both the difficulties of the conventional 2-dimensional plots for high-dimensional data and the opportunities given by the dynamic and multi-media data visualization of the IT age. It is again an iterative process between visualization and computationally feasible model fitting. With both the visualization and modeling stages enhanced by the IT technology and computation speed. Section 3.3 hits the center of a statistical investigation: model fitting. A distinctive characteristic of the IT model fitting is its reliance on prediction of the model through cross-validation or a hold-out test set. The key is the objective function to evaluate a model on the validation set and the emphasis on computational efficiency. Data reduction or feature selection is at the heart of this stage of statistical analysis. Sometimes we desire dimension reduction before hand to speed up the computation in the formal model fitting stage either on line or batch mode. This is where subject knowledge could help tremendously as illustrated by one of the two case studies in the paper: the arctic cloud detection problem. Other times, computational means have to be called upon to choose the features or construct the dimensional reduction as part of the model fitting process. The other case study is the most prominent Network Tomography problem, origin-destination (OD) traffic estimation. It demonstrates well how the two modes of data collection interact. The arctic cloud detection problem is used to illustrate all the steps of an IT statistical investigation; from data collection, to exploratory data analysis, to modeling.

Among the three factors of storage, transmission and computation, computation has had a

long interaction with statistics. Its integration into our systematic study of statistical methods is imperative for us to tackle IT and future statistical problems. Section 4 investigates the role of computation in statistical analysis and asks the question how much computation is necessary for a given statistical accuracy.

A summary and final remarks make up the last Section 5. In particular, we stress the necessity of theoretical work for IT problems with the understanding that we might need to embrace new mathematical concepts and structures such as graph theory and random matrix theory into our analysis tool box.

2 Clustering the diverse IT areas

This section describes the sources of IT data in terms of three clusters to give some organization on the origins of data which might explain the similarities of their statistical analysis. The clusters are made according to the data origin areas' relationships to the center of the IT revolution. Each subsection deals with one cluster with examples from the cluster. Some clusters have a higher demand on one of the three factors (storage, transmission, computation) while others require faster data reduction, but these requirements are intertwined and not separated.

2.1 The IT core areas

This cluster distinguishes itself by its new data forms: text, image, and video. Computer vision and image processing have images and videos as their primary subjects. A main problem in computer vision is object recognition where recent developments capitalize on IT computing power and building image analysis database by human experts. Image processing encompasses many tasks regarding images and videos such as compression, denoising, and enhancement. General purpose static image compression has passed its research stage to become JPEG standards, while special purpose image compression/processing is in much need due to the various image data forms in fields such as bioinformatics, medical research, astronomy, and remote sensing. Special properties of images have to be taken into account for these areas. Demands on movie/video over the internet are pushing video compression/processing research to combat transmission errors and coding delays. These fields are relatively familiar to statisticians so we choose to focus here on areas with mostly text data.

Information retrieval (e.g. web search), information extraction (e.g. title extraction from documents), natural language processing (e.g. speech recognition), and question answering (e.g. "what is the distance between Berkeley and San Francisco?") are products of the IT age. They are related and are outstanding examples of the first cluster, *IT core*. To help readers to follow up on these topics beyond this paper, we list below references and pointers to conferences. For IR and IE and NLP, read Manning and Schütze (1999) [31], Jurafsky and Martin (2000) [24], Manning et al (2007) [30]; for QA, read TREC Publications at http://trec.nist.gov/pubs.html For current developments in these areas, go to websites of the following conferences: StatNLP: Association for Computational Linguistics (ACL), North American ACL (NAACL), Empirical methods for NLP (EMNLP), European ACL (EACL), ICASSP, ICSLP, SIGIR, and WWW, and http://trec.nist.gov/pubs.html.

IT core research is happening mostly outside the traditional statistics community in areas such as signal processing, machine learning, and artificial intelligence. Information Retrieval (IR) is the science and practice of indexing and searching data, especially in the text or other unstructured forms. A typical IR task could be searching for an image with horses in an image database, or searching for a document with a specific name on my computer.

Web search is something we all become to rely on more and more for seeking information on almost anything and everything. Searches for papers on a topic, papers with specific titles, showtimes and locations of a particular movie, mortgage rates, email addresses of colleagues, their telephone numbers, are just few examples of searches on Google. Web search is the hottest topic in IR, but its scale is gigantic and desires a huge amount of computation. First, the target of web search is moving: the content of a website is changing within a week for 30% or 40% of the websites (Fetterly et al, 2004 [15]). A crawler is the algorithm that a search engine uses to collect websites into its database to answer queries. Without the website population indexed, one can not easily carry out a random sampling to crawl. So the crawling results, or query database, might be biased. Worse yet, the content of a website is more than text, fonts vary in size, and images or videos could be part of it. The website data is therefore very unstructured and making its processing or data reduction/feature extraction difficult. We refer the readers to Henzinger (2000) [18] and Henzinger et al (2002) [19] for more details on data collection and algorithm issues related to web search.

Based on the websites collected in a database by a crawler, when a query is entered, the relevant websites will be found and ranked. This fuels a very active research area in machine learning: ranking function estimation. The ranking problem also occurs in other IT core areas such as machine translation where phonetic information is given as the query and a ranked list of characters are the results as needed in a typical Chinese typing software. Ranking function in web search usually depends on weighting websites and links among the sites, as in the PageRank algorithm used by Google (Brin and Page, 1998 [8]). When a weighting scheme is open to the public, however, opportunities arise for the so-called SEO's (search engine optimizers) to mislead the search engine to irrelevant websites of an SEO's customers. Search engines therefore have to outwit SEOs in their search and ranking strategies in addition to dealing with the fast-changing and growing world of websites.

Information extraction (IE) attempts to do more than IR. It would like to extract facts for users in electronic documents (in all languages), that is, it aims to use text as a demonstration of understanding. IE had already existed early in natural language processing, but its potential is boosted dramatically by the IT revolution. For example, Lloyds of London, a shipping company, performed an IE task with human analysts for hundreds of years. The company wants to know all the ship sinking incidents around the world and put the information in a database. The IE question in the IT age is whether we could replace human analysts by a computer algorithm automated to collect this information from data such as newspapers, web broadcasts, and government documents.

IR and IE, including web search, all involve a certain level of natural language processing (NLP), and there is a general question of how to incorporate linguistic rules in statistical or other quantitative approaches. The move in NLP seems now towards using the vast text data that a machine can handle to derive structures and various levels of linguistic generalizations. That is,

NLP is becoming very empirical or data driven. In the empirical approach, large copora are needed for implementation and evaluation. Often, creating an evaluation data set involves humans. A scoring system or loss function also has to be devised to reflect the human evaluation. Then we are in a domain of statistical or machine learning research.

Question answering (QA) takes open domain questions and searches over a collection of documents to find concise answers to the questions. It takes IR and IE further to deal with natural language sentence queries and return answers that need to be precise. Current QA systems could answer simple questions about facts like the one about the distance between San Francisco and Berkeley, but have difficulty answering complex questions requiring reasoning or analysis such as "What is the difference between a private university and a public university?".

In all three areas of IR, IE and QA, documents need to be represented by numeric forms before further actions. Programming skills are required to process these documents, and statistical thinking in natural language processing is necessary to keep the key information in the numeric form (or forming the feature vector) for downstream comparisons between documents. In addition, statistical modeling is often a must to relate the feature vector to the goal of the task as illustrated in the title extraction example next.

Now we describe an IE problem from Hu et al (2005) [21]. It is a title extraction task from PowerPoint documents. The documents are processed to become numeric features based on the understanding of how people usually make the title page of their power point documents. Specifically, they first get every line from the first slide of a PowerPoint document as a data unit by using the program office automation. Then they extract format and other information from all units. There are two kinds of features: format features and linguistic features. The format features include discrete variables (mostly indicator variables) on font size, boldface, alignment, empty neighboring unit, font size changes, alignment changes, and same paragraph. The linguistic features are indicator variables on positive word ("title", "subject", etc) and negative word ("To", "By", "Updated", etc), and discrete variables on word count in the unit (titles should not be too long) and ending characters (ending on special characters or not). It is important to remark that the feature vectors in this problem are mostly binary, common in other IT core text problems as well. This is advantageous from the computational point of view, especially when the feature vector is high dimensional.

The title extraction problem can then be formulated as a classification problem: for each unit, the response is 1 if it is part of the title and the feature vector is the predictor vector in the classification problem. Using training data with both response and predictor information, Hu et al (2005) is interested in predicting the title from the feature vector of a future PowerPoint document.

2.2 IT systems areas

Computer systems are the center of computer science and computer technology. Our second cluster consists of emerging areas of statistical research that tackle central computer system questions. We term this cluster *IT systems*. Examples include hardware (e.g. chip) design, software debugging (Biblit et al, 2005 [2]), and network tomography for computer network management. Computer system problems exist long before the IT revolution, but they were separated from statistical research.

Recently we have seen a surge of efforts to bring statistical methods to these problems as evidenced in the establishment of a new center, the Reliable, Adaptive and Distributed Systems Laboratory, in the computer science department at UC Berkeley. It is endowed by multi-million grants from Google, Microsoft, and Sun Microsystems. To quote the New York Times ("Three Technology Companies Join to Finance Research" by Peter DaSilva, Dec. 15, 2005): The new Berkeley center "will focus on the design of more dependable computing systems." And "the research focus of the new center will be to apply advances in the use of statistical techniques in machine learning to Web services - from maps to e-mail to online calendars - which have become an increasingly important part of the commercial Internet."

Case study: Network tomography

Since my two-year stint at Lucent Bell Labs from 1998-2000, I have been collaborating with network researchers at Bell Labs and Sprint Research Labs on problems of network tomography. Network tomography is a new area at the interface of computer science and statistics which uses less expensive measurement data on computer networks to estimate, via statistical means, expensive characteristics of a network in order to monitor the performance and plan for future. The term network tomography was invented by Vardi (1996) [42] to reflect the similarities between network tomography and medical tomography. See Coates et al (2003) [11] and Castro et al (2004) [10] for two reviews on Network tomography.

Vardi studied a particular network tomography problem, estimation of origin-destination (OD) traffic based on the inexpensive link data which are linear aggregations of the OD traffic according to the routing matrix. We generalized in Cao et al (2000) [9] Vardi's Poisson linear OD model to a Gaussian linear model with a mean-variance power relationship. We used real network data for the first time in the literature. Non-stationarity was dealt with by making an iid assumption on a moving window of about two hours of data. EM algorithm was employed to obtain the maximum likelihood estimates, but MLE was too expensive even for a network of 7 nodes and the algorithm complexity scales as E^5 where E is the number of edge nodes.

Fig.1 displays the Sprint European Network with nodes corresponding to PoPs (Point of Presence) or major metropolitan areas. There are 13 PoPs and 18 link traffic measurements can be collected through the inexpensive SNMP (Simple Network Management Protocol) at interfaces of the nodes.

Based on link vector time series $\{Y_t\}$ in R^{18}, we would like to estimate the OD vector time series $\{X_t\}$ in R^{169} by utilizing the linear network tomography model

$$Y_t = AX_t$$

where A is the 0-1 routing matrix.

For this Sprint European network, the OD traffic was collected for two weeks over 10 min intervals through the computation intensive netflow software on Oracle routers of the network (Lakhina et al, 2004 [26]). It is thus feasible to validate any OD estimation algorithm over this two-week period. The link time series contains $36,288 = 6 \times 14 \times 14 \times 18$ average bytes per second

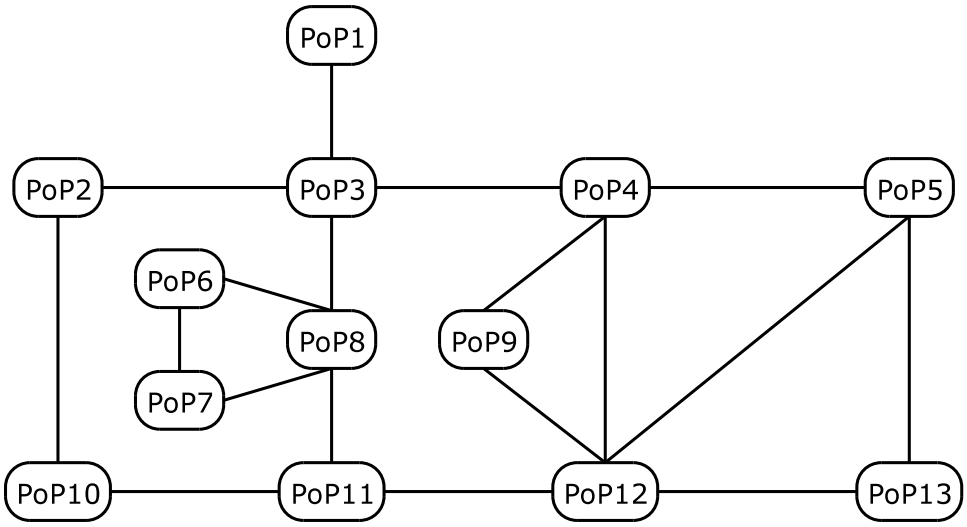


Figure 1: Sprint European Network with 13 PoPs and 18 links.

and it will be used to inversely estimate the OD time series of 340, 704 average bytes per second. Using the method in Zhang et al (2003) [45], AT&T has been retrieving OD information to decide on which links to bring down to maintain– links with light loads are primary targets so that network performance is not sacrificed much.

2.3 IT fringe areas

The third and last cluster contains all the other areas that are impacted by the IT revolution but had been using statistical methods in the pre-IT era. Examples are biology (bioinformatics), medicine, neuroscience, remote sensing (atmospheric science), environmental science, astronomy, particle physics, chemistry, banking, and finance. This cluster is termed *IT fringe*.

Most of these IT fringe areas have a long history of science or social science traditions that statisticians could reply on for much qualitative if not quantitative information. However, the high volume and high speed of data collection are providing challenging but valuable collaborative opportunities for us. Scientists need help to comprehend IT data because subject knowledge alone might not be sufficient. Statistical data analysis or modeling made in these areas forges new frontiers or new sciences themselves.

While it is impossible to enumerate areas from this cluster, we describe four non-random examples from astronomy, particle physics, complex computer mode evaluation, and remote sensing (and atmospheric science), respectively.

Enormous amounts of data are flooding astronomers from the next generation of sky surveys such as the 2 Micro All Sky survey (2MASS) and the Sloan Digital Sky Survey (SDSS) (cf. Welling and Derthick, 2001 [44], Jacob and Husman, 2001 [22]). From the SDSS website (www.sdss.org),

"Simply put, the Sloan Digital Sky Survey (SDSS) is the most ambitious astronomical survey ever undertaken. When completed, it will provide detailed optical images covering more than a quarter of the sky, and a 3-dimensional map of about a million galaxies and quasars. As the

survey progresses, the data are released to the scientific community and the general public in annual increments."

In a five year period, 2MASS and SDSS are expected to produce 550 gigabytes of reduced data and 1 terabyte of cutout images around each detected object. These volumes surpass humans' ability to study them even if we want to. The only alternative is to reply on computing power to sift them through.

In particle physics, gigantic experiments are undertaken to understand the most elementary ingredients of matter and their interactions. See Knuteson and Padley (2003) [25]. One of the experiments, the Compact Muon Solenoid (CMS) at CERN in Geneva, generates about 40 terabytes per second, which has to be reduced to about 10 terabytes per day in real time for later analysis.

Complex computer model simulation is used in wide-ranging areas from meteorology, wildfire control, transportation planning, to immune system function as evident in the workshop on this topic (cf. Berk et al (2003) [1]. Its impact is acutely felt in policy and decision making. One model covered by the workshop is MM5, or Mesoscale Model Version 5 in atmospheric science, a joint effort of National Center for Atmospheric Research (NCAR) and Penn State University. This model utilizes atmospheric observations as initial values, and solves partial differential equations regarding physical thermodynamic and micro-physical processes on a 3-dim grid. The question is how to evaluate or validate such a computer model which takes much computing power for one run. Many difficult issues have to be confronted and we refer interested readers to Berk et al (2003) for collective thoughts on steps necessary to integrate statistics more into this model simulation world.

We now describe the last example of this section from remote sensing or atmospheric science which does not use computer simulation models. The example will be used again in the next section.

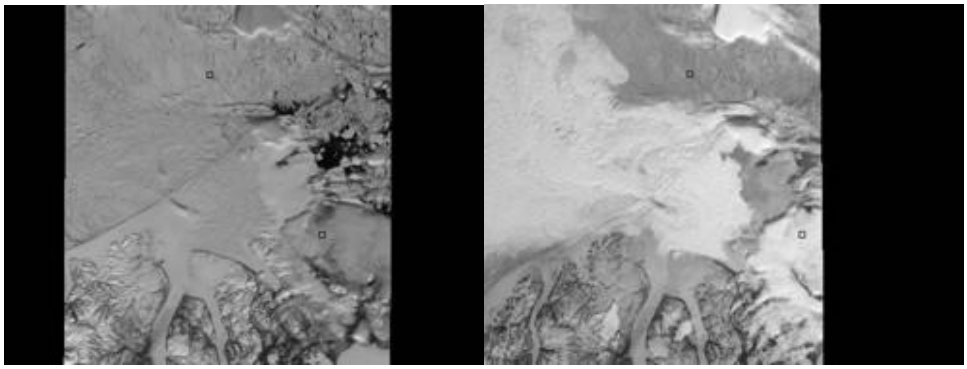
Case Study: Arctic Cloud Detection

Over the past three years, we have been working on a collaborative project of arctic cloud detection using Multi-angle Imaging SpectroRadiometer (MISR) satellite data. The material in this case study is taken from Shi et al (2004 [39], 2006 [40]).

Global climate models predict that the strongest dependences of surface temperatures on increasing atmospheric carbon dioxide levels will occur in the Arctic and this region's temperature increment can lead to global temperature increase. A systematical study of this relationship requires accurate global scale measurements, especially the cloud coverage, in arctic regions. Ascertaining the properties of clouds in the Arctic is a challenging problem because liquid and ice water cloud particles often have similar properties to the snow and ice particles that compose snow- and ice-covered surfaces. As a result, the amount of visible and infrared electromagnetic radiation emanating from clouds and snow- and ice-covered surfaces is often similar, which leads to problems in the detection of clouds over these surface types. Without accurate characterization of clouds over the Arctic we will not be able to assess their impact on the flow of solar and terrestrial electromagnetic radiation through the Arctic atmosphere and we will not be able to ascertain whether they are changing in ways that enhance or ameliorate future warming in the Arctic.

Multi-angle Imaging SpectroRadiometer (MISR) is a sensor aboard NASA's EOS satellite Terra

launched in 1999. It makes novel electromagnetic radiation measurements at 9 different viewing angles of 70.5^o (Df), 60^o (Cf), 45.6^o (Bf), and 26.1^o (Af) in the forward direction, 0.0^o (An) in the nadir direction and 26.1^o (Aa), 45.6^o (Ba), 60^o (Ca) and 70.5^o (Da) in the aft direction. There are also four bands (red, green, blue, and near-infrared (NIR)) of which all are collected in the 275 m $\times$ 257 m resolution intially. Due to the high data volume, all other bands except for the red are aggeragated to the coarser 1.1 km $\times$ 1.1 km resolution before their transmission to the base stations on earth. Fig. 2 shows the An- and Df-angle images of part of Greenland in the Arctic. Clearly, thin clouds not seen in the An image are shown clearly in the Df image, proving that angular information is useful for cloud detection.



An angle image Df angle image

Figure 2: MSIR An (nadir) and Df (most forward) angle images of Greenland, date?

Our goal in this project is to provide a cloud label to each pixel based on MISR's red band 9-viewing angle data (other bands have a coarser resolution and do not seem to contain more information on the cloud lable). For MISR operational purposes, we would like to have an on-line algorithm which outputs a label while data come in from the MISR sensor.

The MISR red-band data is 9 dimensional per pixel and there are about 3 million pixels per image block per angle. When we started our investigation, the MISR operational algorithm, Stereo Derived Cloud Mask (SDCM), uses the red-band data and retrieves cloud height based on matching of clouds in images of different angles. Cloud mask is obtained by thresholding the cloud height based on the terrain height. The matching step is computationally expensive and error-prone. The MISR team was open about the problems and encouraged us to collaborate.

In all the three examples of the IT fringe cluster, there was a problem of on-line data reduction because of the sheer volume of raw data. We will address this issue in the next section.

3 Challenges and opportunities of the IT age for Statistics

Despite of the diverse origins of IT data sets as described in the previous section, the IT problems share much in common in terms of statistical information extraction. That is, the high dimensionality (both sample size and/or feature dimension) and high-speed of data collection and their tensions with storage, transmission and computation are the key ingredients to consider in a sta-

tistical investigation. On one hand, this commonality forms the basis for systematic studies of IT problems within statistics. On the other hand, subject knowledge still plays an important role for subject matter often suggests meaningful data reduction/feature extraction and dictates the storage, transmission, and computation limitations. In this section, we go through the steps in a statistical investigation to illustrate the impact of IT on data collection, exploratory analysis, and model fitting. Considerations of storage, transmission, computation are contrasted against the high-dimensionality and high speed of IT data in all three steps to suggest potential innovations of an IT statistical investigation. Throughout this section, the iterative nature of a statistical investigation is flashed out to demonstrate the inseparatability of its many components.

3.1 Data collection

IT data come in huge volumes and high rates, as much as terabytes in seconds as in particle physics, and the date rate will only increase. Moreover, the data forms vary. One major new form is text which is the raw data for problems in IR, IE, QA and NLP in the IT core cluster. Other new forms include images and videos. For web search, there is a mixture of these forms as raw data. Manipulating and converting these new forms or types of data require specialized softwares and/or programming skills. In some sense, converting these new types into traditional vectors in Euclidean spaces is a form of data compression and possibly lossy. The question is whether the useful statistical information for a particular task is retained by this mapping or compression.

Before data collection, one has to decide on what data unit to collect. Often sampling is used to reduce the cost or data volume. In web search, sampling is a major problem for crawlers as pointed out by Henzinger et al (2002) [19] . For network monitoring data collection, Monila et al (2005) [33] propose intelligent sampling methods based on hashing to reduce data volume. Sampling is another form of possibly lossy data compression and thus runs also the risk of losing essential information for downstream statistical analysis.

In terms of data collection mode, there are two kinds. One is the traditional batch-mode where data are transmitted and stored on a storage device after taking consideration of transmission bandwidth and storage space limits. Statisticians have access (no matter how slowly) to all the data available for the statistical quest of useful information. The other mode is on-line (or streamed data) where data have to be discarded along the way and analysis is done while the data are coming in. The second mode is an IT phenomenon. Obviously, both modes can be part of the same statistical investigation. Batch mode analysis might be necessary to know what features or data reduction should be conducted on-line to extract most useful information. An on-line data reduction such as downsampling or aggregation might be necessary for the data to meet the storage limitation for batch mode analysis. Recall that the MISR sensor collects data in the 275m $\times$ 275m resolution for all four bands, but aggregates the three non-red bands into the 1.1 km $\times$ 1.1 km resolution before transmitting the data down to the earth. Other times, more sophiscated quantization schemes are called for to keep more relevant statistical information to meet the transmission constraints. For example, Braverman et al (2003) [6] use vector quantization for remote sensing data size reduction so that users of NASA data sets can download them over the internet while maintaining essential statistical informatione. In Jörnsten et al (2003) [23], both lossless and lossy compression

schemes are designed to preserve statistical information in a microarray image. When statistical model is parametric, there is a field called multi-terminal information theory that addresses the issue of optimal lossy compression by individual data collection stations to estimate the parameter depending on all the data. However, since it is not always clear what is the downstream statistical analysis, the lossy compression schemes have to retain general information about the high precision data.

The other mode of statistical analysis is on-line or streamed data mode. The special issue of *J. Computational and Graphical Statistics* on streamed data in 2002 gives a good collection of examples of statistical research in this area. We will refer to some of the articles there in the subsequent text. IEEE transactions contain many more articles under the name of on-line data analysis. On-line and streamed data are not exactly the same, to be precise. The former emphasizes the algorithmic development without worrying about the data handling aspects, while the latter deals with the two aspects together at the face of a never-ending data stream.

For this mode, data can not be stored in its entirety and analysis has to be conducted as data come in. There are many interesting examples of work on streamed data as the special issue shows, but it is difficult to conduct a systematic study of this mode of analysis. The difficulty lies in the fact that different examples have different hardware and computation constraints as shown in Hoar et al (2003) [20] and Knuteson and Padley (2003) [25]. The question is whether we could distill a mathematical framework to encompass different considerations (data storage, computation and statistical) to allow a meaningful and relevant analysis.

The emergence of sensor networks adds one more consideration into the streamed data analysis: data transmission. Sensor networks are self-networked small devices that are engineered to collaborate with each other and collect information concerning the environment around them. Their flexibility greatly extends our ability to monitor and control the physical environments from remote locations. Their applications range from seismic, natural environmental monitoring, to industrial quality control, and to military usages. However, the sensors are constrained by the battery power whose major consumer is communication (data transmission), and then to a much lesser extent computation. So far sensor networks research is dominated by researchers from computer science and electrical engineering, but it provides one of the ideal platforms for us to integrate statistical analysis with computation, data compression and transmission because the overwriting power constraint forces us to consider all the players in the same framework to maximize the utility of the limited battery energy. It would be interesting to try to devise a framework to encompass compression, transmission and statistical analysis to answer optimality questions. Special case analysis does exist in the literature. Nguyen et al (2005) [34] study the interaction of classification rules based on 1-bit quantizations at iid sensors and give related references.

Network Tomography Case Study (continued)

Most of the works on the OD estimation problems are in the batch mode (e.g. Vardi, 1996 [42], Cao et al, 2000 [9], Medina et al, 2002 [32], Liang and Yu, 2003 [29], Zhang et al, 2003 [45]). Liang and Yu (2003) [29] employed the Gaussian model with a mean-variance relationship and sped up

the Maximum likelihood estimation in Cao et al (2000) [9] from E^5 to $E^{3.5}$ (E is the number of edge-nodes) by applying maximum pseudo-likelihood (MPL) . We compared MPL with Zhang et al [45]'s method on the Sprint European Network data set and found a relative error rate around 30% for large OD traffic values for both methods (with MPL being slightly better). Very recently Liang et al (2006) [28] present an on-line approach, PamTram (PArtial Measurement of TRAffic Matrices), which bypasses the expensive parameter fitting based on the Gaussian model from the link vector. It also uses partial information on directly collected OD traffic as in Soule et al [41]. PamTram uses one (or a couple) randomly chosen OD pair measurement all the time while Soule et al [41] collects all the OD traffic over a certain period of time. It is made clear in Liang et al [28] that the OD collection overhead is much less for PamTram and one randomly selection OD pair. PamTram replies on Iterative Proportional Fitting (IPF), which was used in Cao et al (2000), and the Gaussian model on the OD vector. It is computationally light-weight and uses very little overhead OD measurement. The relative error is reduced from 30% to 5% or 7 % which is below the 10% relative error upper bound necessary for Tier-III carriers like Sprint to use for OD traffic estimation. See Fig. 3 for two sample OD traffic time series plot, with the true traffic, and the PamTram estimates with various selection schemes to choose the OD flow to measure (see Liang et al, 2006 [28], for more details).

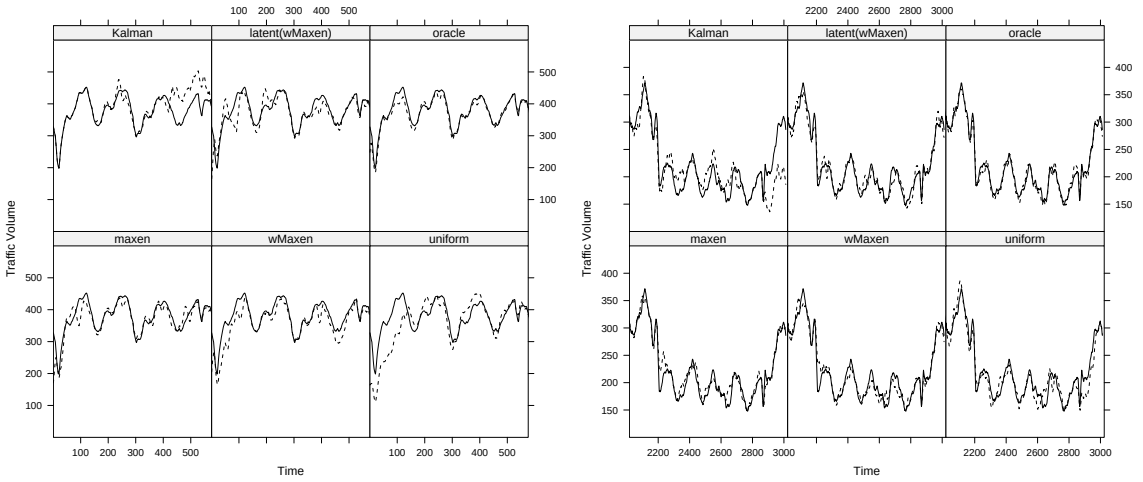


Figure 3: Two sample OD traffic flows. latent(wMaxen), maxen, wMaxen and uniform are different random selection schemes to select the one OD flow to measure in PamTram. Oracle uses the best OD flow to reduce the error based on the knowledge of all the OD flows.

In this case study, the batch mode analysis was imperative for the later on-line PamTram. The Gaussian model was validated by quite a few batch model analysis (e.g. Cao et al [9], Zhang et al [45], Lakhina et al [26]) and the IPF was first used for the batch model analysis in Cao et al [9].

Due to the huge data volume, even the batch mode poses formidable challenges and unforeseen opportunities for statisticians in the IT era. These issues will be addressed in detail in the sections to follow.

3.2 Exploratory data analysis

Since the advocacy and teaching of Tukey on Exploratory Data Analysis (EDA) in the 70's, EDA has been part of statistician's routine analysis of data. The main tools of EDA are summary statistics and 2-dim visual displays such as histogram, box plots, scatter plots and time series plots. For high-dimensional data, 2-dim visualization abilities become quite limited for understanding the complex structures in them, even though efforts within the statistics community have been made to accommodate the high dimension through parallel plots and selective projections of data as in xgobi (cran.r-project.org/doc/packages/xgobi.pdf) and crystal-vision (http://www.crystalvision.tv/). For other communities (e.g. machine learning and signal processing) dealing with similar high dimensional data, EDA is not part of their eduction so not always carried out before formal algorithmic analysis or modeling.

For IT core areas such as information retrieval and information extraction, the original form of data are often texts, not always well formulated or structured. The obvious is to map these data into the numeric form so to use traditional summaries and EDA visualization tools. New summary and visualization tools seem necessary for text and other IT forms of data.

For high dimensional and high speed IT data, summary statistics can be computationally expensive to collect. Calculating simple summaries such as mean and variance requires going through all the data units and therefore is costly if the data units are in millions. Random sampling could be used, but not effective when the data are sparse as in many machine language translation problems.

There are two reasons for us to be cautiously optimistic about data visualization in the IT age. Both are brought about by the IT revolution itself. The first reason is that there are more tools available to represent data beyond the conventional 2-dim forum. Many advances going beyond the 2-dim plots are being made by people outside the statistics community, visual artists in this case. It is encouraging to see that some of them are actually specilaizing in visualizing scientific data as in the works of Ben Fry of MIT (http://acg.media.mit.edu/people/fry/). Of course, statisticians (and collaborators) are also getting on top of data representation using multimedia programming and visualization tools. For example, Coates et al (2003) [12] and Wegman and Marchette (2003) [43] visualize network data, Hansen and Ruben represent web data by music (sounds) and images (videos) in the much acclaimed statistician-artist collaboration Listening Post (http://www.earstudio.com/), and Welling and Derthick (2001) [44] and Jacob and Husman (2001) [22] design special visualization tools to display large-scale sky surveys.

The second reason is that there are meaningful structures in even high-dimensional data. If we find these structures, then the high-dimensional data can be reduced to low dimensions for visualization. With the increasing computing power in the IT age, we could use algorithmic means or statistical machine learning methods to help search for these structures at a speed impossible before (cf. Section 3.3 below). That is, data visualization and model fitting ought to be conducted iteratively. Seeing suggests models to fit and model fits gives data to see. This is similar to what we do in residual analysis for regression models, but residual plots are replaced with multi-media data representation and regression models are replaced by more general methods. Admittedly, this is easier said than done.

For the arctic cloud detection problem, it was not trivial to read MISR data into matlab and

access each pixel, due to the huge volume of data ($9 \times 2084 \times 1536 = 28,311,552$ or 28 millions raw observations for one block of 9 viewing angle red images). It took 2 minutes just to calculate simple summary statistics and physical features for a one-angle image block ($2048 \times 1536 = 3{,}145{,}728$ or 3 millions). We had to program a gui interface in matlab and pre-compute all the simple statistics needed before we could sit down and look at them.

To make the summarizing and visualizing large and possibly non numeric data part of our routine data analysis requires rethinking of our undergraduate and graduate curriculums. It is necessary to introduce and keep up with new programming language and data structure developments in computer science. The question is whether we should send our students to take computer science courses or teach a course by our own faculty. There are pros and cons of both approaches, but the agreement is that our students need to acquire these skills and our departments need to get on the computational science wagon in the form of new centers sprouting all over the universities in the country.

3.3 Model fitting: the role of feature extraction

A distinctive characteristic of current IT statistical research and practice is its reliance on test or validation data, or predictive performance (cf. Breiman, 2001 [7]). This is a consequence of the availability of large IT data in many areas with responses or labels. At the core of this predictive paradigm lies the loss or objective or scoring function to fit a model and to evaluate a fitted model or procedure against the validation data. Computation efficiency is also a must for all IT problems.

Even though IT data sets are massive, understanding can be obtained only through data reduction, one way or the other. IT data sets are high dimensional in both sample size and predictor dimension. Random sampling is a straightforward way to reduce the data size, but other methods might preserve better statistical information. Fast computation and data reduction have a relationship of chickens and eggs. Efficient computation is feasible if data reduction or feature extraction reduces the size and dimension of the data; on the other hand, fast computation is frequently desired to extract features or reduce the dimension of the data. Computation and feature extraction are tangled together and aid each other.

3.3.1 Feature selection via subject knowledge

Data reduction for statistical inference goes back at least to sufficient statistic which is a beautiful and practical concept in parametric inference. It still has a significant role to play because parametric models suggested by science or empirically proven are still the best even if we have massive data sets – its simplicity or sparsity is the best data reduction.

Building on existing atmospheric science knowledge, we found ourselves in a parametric situation with the arctic cloud detection project. In a nutshell, our approach is based on three physically meaningful features, thresholding rules on the features, and then using the thresholding labels to carry out Fisher's Quadratic Discriminate Analysis (QDA) for the final labels. In retrospect, our approach is simple (and hence very fast to be implemented on-line), but it took more than three years and many iterations of data analysis by us and feedbacks and working together with the

scientists of the MISR team. Subject knowledge played an important role, or we used the most powerful computer – human brain.

For a given pixel, we now define the following three features (cf. Shi et al, 2004):

1. SD, the standard deviation of the nadir (An) angle data over a square patch of 8×8 pixels centered at the given pixel;

2. Corr, the average correlation between the nadir (An) angle and the two angles, Af and Bf, on the same 8×8 patch;

3. NDAI, the index of forward scattering, developed by MISR team member Nolin and co-workers (cf. Nolin et al, 2002):

$$NDAI = \frac{X_{Df} - X_{An}}{X_{Df} - X_{An}},$$

where X_{Df} and X_{An} are the Df and An measurements at the given pixel.

Our labeling rule takes a different angle at the cloud detection problem from the MISR operational algorithm, SDCM. Instead of the expensive and inaccurate retrieving cloud height and thresholding for cloud pixels, we search for for surface or ice/snow pixels that do not move. By exclusion our method gives the cloud pixels. Precisely, our thresholding method, enhanced linear correlation matching (ELCM), is as follows:

1. If $SD < 2$, the pixel is surface or snow/ice.

2. When $SD \geq 2$, if Corr > 0.8 and $NDAI < T_{NDAI}$, the pixel is also labeled surface or snow/ice.

The two thresholds 2 and 0.8 on SD and Corr are found by taking into account the level of instrumental noise and through trial and error. The threshold T_{NAI} on NDAI is obtained adaptively by running an EM algorithm to fit a mixture of two Gaussians and use the dip in the middle of the mixture density (cf. Shi et al, 2004). The first inequality in ELCM on SD finds the surface frozen ice pixels (e.g. the fork-looking frozen river in Fig. 2). The second and third inequalities ascertain that the surface pixel under consideration has a texture signature captured by a high Corr and is not covered by thin clouds because NDAI is low (not in the white region in the Df image of Fig. 2). Shi et al (2006) show that the three features are much more robust against or can adapt to location and weather condition changes than the radiances.

Feeding ELCM labels into QDA gives a curved class boundary which better captures the class oundaries as shown in the 3-dim plot in Fig. 4 (corresponding to subimages in Fig. 2). The resulted ELCM-QDA algorithm does not require any human input. It provides substantial improvements over existing MISR operational algorithms as shown in Shi et al (2006) through an extensive testing of the algorithm against expert labels over 60 blocks of data (about 55,000 pixels per block with expert labels). To be precise, the improvement is from 72.99 % to 91.8 % on classification rate and from 27to 100% on coverage (that is, the percentage of pixels that a label is given) over the MISR operational algorithm: Stereo Derived Cloud Mask (SDCM).

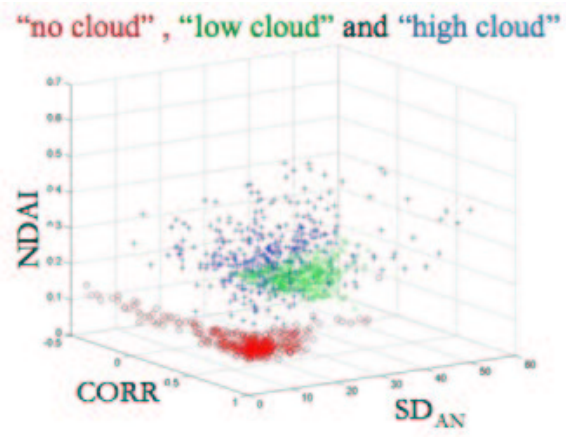


Figure 4: (b) 3-dim plot of the three features for small regions of the images in Fig. 2 with red, green and blue points as labeled above the plot.

3.3.2 Automatic feature extraction

Atmospheric science is a mature and established discipline so that we could take advantage of the knowledge and experience because of our access to the MISR scientists in the arctic cloud detection project. Things are different in the areas of the IT core and systems clusters in Section 2 where the IT data explosion is forging new frontiers and disciplines. Here the existing knowledege is either scarce or not adequate. For example, in natural language processing, linguistic theories are useful, but for good practical performance of NLP systems, it is necessary to combine them with large validation databases which are products of the IT age. These areas are fertile grounds for automatic feature selection.

Automatic feature extraction has traditionally been carried out via model selection in the statistics community. Implicitly, model selection schemes make use of a penalized likelihood function. To be precise, suppose we have n observations $Z_i = (Y_i, X_i)_{i=1}^n$, where $X_i \in R^p$ is a collection of feature vectors and Y_i is the associated response. Let $L(Z_i, \beta)$ denote a loss function (e.g. negative log-likelihood) for a model connecting X_i and Y_i that depends on a parameter vector $\beta \in R^p$. In this context, we select a model by specifying a value of β. Through a penalty $T(\beta)$ we can impose desirable characteristics like sparsity. We choose β to minimize the penalized least squares

$$\sum_{i=1}^{n} L(Z_i; \beta) + \lambda T(\beta), \tag{1}$$

where $\lambda \geq 0$ is a regularization parameter to be chosen based on data. If we select

$$T(\beta) = T_0(\beta) = \|\beta\|_0 = \sum_{i=1}^{p} I_{\{\beta_i \neq 0\}},$$

the so-called l_0 penalty, we are explicitly discouraging models that include a large number of predictors, or, non-zero values of β. In other words, with the l_0 penalty, we are conducting model

selection. While this choice of penalty makes intuitive sense, identifying the value of β that minimizes the penalized least squares involves an expensive combinatorial search. As an alternative, Tibshirani (1996) proposes the Lasso to get a simultaneously sparse and regularized solution. In (1) we set the l_1 penalty as

$$T(\beta) = T_1(\beta) = \|\beta\|_1 = \sum_{i=1}^{p} |\beta_i|,$$

and refer to the minimizer of the penalized least squares as the the Lasso estimate. For theoretical results regarding Lasso, see Fu and Knight (2000), Meinhausen and Bühlmann (2005), Meinhausen (2005), and Zhao and Yu (2006). Compared to the conditions for model selection criteria such as BIC and MDL to be consistent in the classical setting, Lasso needs more conditions on the design matrix (Zhao and Yu, 2006 [47]) – a price to pay for the computational speed-up. Or do we have to pay? Is there a better computational alternative than Lasso which can be consistent under the same conditions as BIC or MDL?

Efficient algorithms have been developed to solve the Lasso problems corresponding to different λ's. In the L^2 loss case, a fast algorithm is given by Osborne et al (2000a [35], 2000b [36]) and Efron et al (2004) [13] that uses the fact that Lasso's path is piecewise linear. It is called LARS by Efron et al (2004) [13]. The BLasso algorithm in Zhao and Yu (2004) [46] can handle any convex loss function with a convex penalty. For non-convex penalties, computation gets harder (cf. Fan and Li, 2001 [14]).

Next we compare automatic feature selection via $L1$ penalty with subject-knowledge based feature section in the arctic cloud problem.

Case Study: Arctic Cloud Detection (continued)

In the arctic cloud classification problem, we use the expert labels as training and testing data to compare the hand-crafted three physically meaningful features (SD, Corr, NDAI) with raw MISR measurements (we use only the five most relevant radiances). We apply both logistic regression (no feature selection) and L_1 penalized logistic regression (automatic feature selection). For the automatic selection, we use an algorithm developed by my Ph.D. student G. Rocha for a class, LR-LARS (LR for Logistic Regression), which is an extension of of the LARS algorithm for Lasso (Efron et. al, 2004 [13]) to the logistic regression case for tracing the regularization path. To a certain extent, it is related to the iteratively reweighted least squares (IRLS) procedure. The algorithm is started at the fully regularized solution (for which only the intercept is allowed to differ from zero). At each step, the Hessian and the gradient of the logistic loss at the current estimate are used to estimate the next λ at which a variable will come into the model or be dropped from the model. Newton steps are then used to compute the optimum at this new value of the regularization parameter, using the current solution as a starting point. While computing this new estimate, the active set is kept fixed. After the new solution is computed, the gradient is used to check whether the active set should have been changed. If that is the case, the new regularization parameter is reset to a value between the previous one and the current one. That allows for the points at which variables come into the model and go out of the model to be tracked more carefully. To a

certain degree, this can be compared to the algorithm proposed by Rosset et al (2004) [37] with the difference that here the step size is constantly adjusted from step to another. The algorithm was implemented in MatLAB and is similar to the one implemented in R by Park and Hastie (Park, M.Y. and Hastie, 2005, see http://www.stanford.edu/ mypark/).

We take MISR orbits 12791, 13257, and 13490 from May 14, June 1 and June 15, 2002, and carry out L_1-penalized logistic regression using LR-LARS with L in (1) as the log-likelihood of Y given x where Y is the label of snow/ice vs. cloud, and x is the predictor vector. For the predictor vector, we take the hand-crafted features (SD, CORR, NDAI) in one case, and the five radiances at angles Df, Cf, Bf, Af, An in the other case when running LR-LARS. The labels are from an expert, Professor Eugene Clothiaux from Department of Meteorology at Penn State University, who is familiar with the geographic features of the locations of these three orbits. He based his labels on viewing both MISR and hyperspectral MODIS images where MODIS is the major sensor aboard Terra. He labeled 70,917, 82,148, and 54,996 pixels for orbits 12791, 13257, and 13490, respectively. The labeled pixels for each orbit is divided randomly into a training and a testing set of equal size.

For all the three orbits, LR-LARS selected all the three hand-crafted features; in two of the three orbits, LR-LARS selected three radiances out of 5 radiances. We also run logistic regression without any penalty in both cases, and the results are comparable with the penalized logistic regression in both cases. With or without penalties, the fitted logistic regression model based on handcrafted features give better testing error rates than the fitted model on the radiances for all three orbits as shown in Table 1. These results demonstrate a clear advantage of subject-knowledge based features over raw MISR radiances, even thought the automatic feature selection is providing decent results. We do not find any classification rate advantage of feature selection via LR-LARS over no-selection since the training sample size is huge (over 20,000 for each orbit) and the total number of features is only 5. Nevertheless, we are pleasantly surprised that 3 features are actually selected by LR-LARS for two of the three orbits.

Table 1: Classification error rates on test sets

	Orbit 12791		Orbit 132572		Orbit 13490	
	LR-LARS	LR	LR-LARS	LR	LR-LARS	LR
NDAI, SD, CORR	0.0682	0.0683	0.0394	0.0389	0.1555	0.1577
Radiances @ DF,CF,BF,AF,AN	0.1434	0.1485	0.0480	0.0476	0.2113	0.2031

3.4 How much computation is "minimally sufficient" for fitting a statistical model?

As argued in the previous section, data reduction and computation are closely related and inseparable at a practical level. At a theoretical level, in Kolmogorov's algorithmic complexity theory, the complexity of a data string is the length of the shortest program on a universal computer (cf. Li and Vitanyi, 1997 [27]) and hence the shortest program also serves as the reduction or model

for the data string. Thus data reduction and computation are one! Unfortunately, K-complexity is not always computable, that is, one may not find the shortest program for a particular data string within a finite amount of time. In a statistical investigation, we rely on scientific computing that is different from computation on a universal computer, although reconciliation might be possible (Blum et al, 1998 [3]).

The recent theoretical and algorithmic developments for solving Lasso efficiently are part of machine learning research now expanding from computer science to statistics. Machine learning is at the frontier of Statistics because its serious considerations of computation in statistical inference. Computation has entered statistics much earlier than data compression and transmission. Looking back, we might view the development of computation in statistics into three phases. The first phase was pre-computer where we depended on closed-form solutions. The second phase uses computer, but not in an integrated manner: we would design a statistical method and then worry about how to comput it later. Often calling a numerical optimization routine was the solution and we relied on the routine to decide on how the numerical convergence was decided, that is, convergence parameters were tuned for numerical reasons and the optimization routine was used as a black-box by statisticians. The third phase is the IT phase where data volume is so gigantic that procedures without computational considerations while in design might not be implementable. Machine learning methods and Markov Chain Monte Carlo algorithms are examples.

Two machine learning methodologies stand out: one is boosting (Freund and Schapire, 1997 [16], Hastie et al, 2001 [17]) and the other support vector machines (cf. Scholkopf and Smola, 2002, [38]). They both enjoy impressive empirical performances and now with some theoretical understanding from both the machine learning and statistics communities. The current view of boosting is that it fits an additive model via gradient descent or its variant algorithm to minimize an objective function $L(Z, \beta)$. It is stopped early by monitoring the generalization or prediction error of the fitted model either estimated by cross-validation or assessed over a proper test set. That is, the minimization is a "pretense" – we are really interested in the solutions along the way to the minimum, not the minimum itself, and prepared to stop early. This way numerical convergence is not important at all, but the prediction performance is. For support vector machines, computation is also the main focus via the "kernel trick". An implicit Reproducing Hilbert Space is induced by a kernel function and in this space, a linear model is fitted. However, all the computation is conveniently done via the kernel function.

There is something very novel about boosting (and fitting neural networks): the computation parameter, the number of iterations, serves also as a regularization parameter in statistical estimation. BLasso has a similar property. It is a component-wise gradient descent algorithm with a fixed step to minimize the generalized Lasso loss (convex loss and penalty functions in (1)) simultaneously for different values of λ's. It shares much similarity with boosting when a component-wise gradient descent algorithm is used, or the forward stagewise regression (FSR) as called by Efron et al (2004). That is, Blasso has a forward step just as in FSR, but with a backward step added to make sure the combined penalty in (1) is minimized not just the loss function part which is the aim of boosting. Moreover, Blasso solves a sequence of optimization problems corresponding to different λ's similar to the Barrier method in optimization (Boyd and Vandenberghe, 2004 [5]).

The coupling of computation and regularization in boosting and BLasso is reminiscent of the sameness of computation and modeling in K-complexity theory. Obviously statistical model fitting uses scientific computing, but statistical computation is special. Even in the parametric case, there is a well-known result that only one Newton step is needed to make a $\sqrt{n}$-consistent estimator efficient. That is, since our objective function is a random quantity, we do not need convergence of the minimization algorithm to get a statistically satisfying solution as shown in boosting. In nonparametric methods such as boosting, neural nets and BLasso, early stopping before convergence saves computation and regularizes the fitting procedure and hence results in a better statistical model. Again, computation and model fitting seem to be working in the same direction – less computation and better statistical accuracy. These facts indicate the intimate relationship between computation and model fitting. They prompt us to ask the following question:

Is there a minimal amount of computation needed for a certain statistical accuracy?

It is not clear whether this question is answerable because fast algorithms in scientific computation often rely on closed form equations or relationships derived through analytical means. Analytical calculations are infinite precision, while scientific computations are finite precision. Nevertheless, we believe it is a very interesting intellectual questions to ask and the pursuit of the answer to this question could lead to useful practical consequences for modeling IT data.

3.5 Concluding remarks

In this article, we discussed the exciting challenges and opportunities that the IT revolution is bringing us. We first reviewed three clusters of IT problems, IT core, IT systems, and IT fringe to point out the diverse forms of IT data: old (numeric) and new (text, images, videos, and multi-media), all of which are high-dimensional and could come in high speed. These IT data characteristics impact a statistical investigation through its data collection, exploratory data analysis, and model fitting stages. New data forms require more programming skills than conventional. Sampling has to get intelligent to extract representative information from the vast and often sparse population of IT data. Data collection could be in batch or on-line modes. Exploratory data analysis on one hand faces the tough task of visualizing high dimensional data, and on the other hand is aided by the multi-media representation of data and the faster computation to extract low dimension features to visualize. The formal model fitting phase often relies on data reduction or feature extraction that is closely connected to computation. Subject matter still helps feature selection just as in classical statistics as demonstrated in the arctic cloud detection problem; while computationally feasible automatic feature selection can be implemented through methods such as Lasso. By broadening the scope of methods from Lasso to machine learning, especially boosting, we made a case for systematic investigations of connections and interplays between computation and statistical model fitting.

Last but not least, we believe that theoretical analysis are in need to facilitate IT data modeling since data points in high dimensional spaces could have properties that we do not really understand yet. However, it is a non-trivial task to formulate a relevant analytical problem whose solution would

shed light on high-dimensional data modeling. We are optimistic that such theoretical results would emerge as we have recently seen in the eigen results of large random matrices.

4 Acknowledgements

This work is partially supported by NSF Grant DMS-03036508 and ARO Grant W911NF-05-1-0104. The author is very grateful for stimulating discussions with and pointers to references by John Rice, Dan Klein, Mike Jordan, David Purdy, Hang Li, and Yunhua Hu. The author would also like to thank Guilherme Rocha for obtaining, and Tao Shi for assisting, the L_1-penalized logistic regression and logistic regression results on the arctic cloud detection problem. Finally, we would like to thank Sprint Labs and the NASA Landley Research Center Atmospheric Science Data Center for providing data in the two case studies, and thank the MISR team members for invaluable discussions for the cloud project.

References

[1] R. A. Berk, P. Bickel, K. Campbell, R. Fovelli, S. Keller-McNutty, E. Kelly, R. Linn, B. Park, A. Perelson, N. Rouphail, J. Sacks, and F. Scheonberg. Workshop on statistical approaches for the evaluation of complex computer models. *Statist. Sci.*, 17(2):173–192, August 2002.

[2] B. Biblit, M. Naik, A. X. Zheng, A. Aiken, and M. Jordan. Scalable statistical bug isolation. *ACM SIGPLAN 2005 Conference on Programming Language Design and Implementation (PLDI 2005)*, 2005.

[3] L. Blum, F. Cucker, M. Shub, and S. Smale. Complexity and real computation. *New York: Springer*, 1998.

[4] G. E. P. Box. Sampling and bayes' inference in scientific modeling and robustness (with discussion). *J. Roy. Statist. Soc. Ser. A*, (143):383–430, 1980.

[5] S. P. Boyd and L. Vandenberghe. Convex optimization. *Cambridge, UK; New York: Cambridge*, 2004.

[6] A. Braverman, E. Fetzer, A. Eldering, S. Nittel, and K. Leung. Semi-streaming quantization for remote sensing data. *J. Computational and Graphical Statist.*, 12(4):759–780, 2003.

[7] L. Breiman. Statistical modeling: two cultures. *Statist. Sci.*, (16):199–231, 2001.

[8] S. Brin and L. Page. The anatomy of a large-scale hypertextual web search engine. 1998. Proceedings of the Seventh World Wide Web Conference.

[9] J. Cao, D. Davis, S. Vander Wiel, and B. Yu. Time-varying network tomography: router link data. *Journal of American Statistics Association*, 95:1063–1075, 2000.

[10] R. Castro, M. Coates, G. Liang, R. Nowak, and B. Yu. Network tomography: recent developments. *Statist. Sci.*, (19(3)):499–517, 2005.

[11] M. Coates, A. Hero, R. Nowak, and B. Yu. Internet tomography. *Signal Processing Magazine*, 19(3):47–65, 2002.

[12] C. Cortes, D. Pregibon, and C. Volinsky. Computational methods for dynamic graphics. *J. Computational and Graphical Statist.*, 12(4):950–970, 2003.

[13] B. Efron, T. Hastie, and R. Tibshirani. Least angle regression. *Annals of Statistics*, 32:407–499, 2004.

[14] J. Fan and R.Z. Li. Variable selection via nonconcave penalized likelihood and its oracle propertie. *Journal of American Statistical Association*, (32):407–499, 2001.

[15] D. Fetterly, M. Manasse, M. Najork, and J. L. Wiener. A large-scale study of the evolution of web pages. *Software Practice and Experience*, 1(1):1–27, 2004.

[16] Y. Freund and R. Schapire. A decision-theoretic generalization of on-line learning and an application to boosting. *J. Comput. Syst. Sci.*, (55(1)):199–139, 1997.

[17] T. Hastie, R. Tibshirani, and J.H. Friedman. *The Elements of Statistical Learning: Data Mining, Inference and Prediction*. Springer Verlag, 2001.

[18] M. R. Henzinger. Algorithmic challenges in web search engines. *Speech and Language Processing*, (Prentice Hall), 2000.

[19] M. R. Henzinger, R. Motwani, and C. Silverstein. Challenges in web search engines. 2002. ACM 18th International Joint Conference on Artificial Intelligence.

[20] T. J. Hoar, D. Milliff, R. F. Nychka, C. K. Wikle, and L. M. Berliner. Winds from a bayesian hierarchical model: computation for atmosphere - ocean research. *J. Computational and Graphical Statist.*, 12(4):781–807, 2003.

[21] Y. Hu, H. Li, Y. Cao, D. Meyerzon, and Q. Zheng. Automatic extraction of titles from general documents using machine learning. *JCDL'05, June 7-11, Denver, Colorado*, 2005.

[22] J. C. Jacob and L. E. Husman. Large-scale visualization of digital sky surveys. *Virtual observations of the future, Astronomical Society of the Pacific Conference Series*, 225(291-296), 2001.

[23] R. Jörsten, W. Wang, B. Yu, and K. Ramchandran. Microarray image compression: Sloco and the effects of information loss. *Signal Processing Journal*, 83:859–869, 2003.

[24] D. Jurafsky and J. H. Martin. *Speech recognition and Language Processing*, (Prentice Hall), 2000.

[25] B. Knuteson and P. Padley. Statistical challenges with massive datasets in partical physics. *J. Computational and Graphical Statist.*, 12(4):808–828, 2003.

[26] A. Lakhina, K. Papagiannaki, M. Crovella, C. Diot, E. Kolaczyk, and N. Taft. Structural analysis of network traffic flows. In *ACM Sigmetrics*, June 2004.

[27] M. Li and P. M. B. Vitanyi. An introduction to kolmogorov complexity and its applications. *New York: Springer*, 1997.

[28] G. Liang, N. Taft, and B. Yu. A fast lightweight approach to origin-destination ip traffic estimation using partial measurements. *Joint Issue of IEEE Trans. Information Theory and IEEE Trans. Networks on Data Networks*, (to appear), 2006.

[29] G. Liang and B. Yu. Maximum pseudo-likelihood estimation in network tomography. *IEEE Transactions on Signal Processing*, 51(8):2043–2053, August 2003.

[30] C. Manning, P. Raghavan, and H. Schütze. *Introduction to Information Retrieval*, (in preparation):http://www–csli.stanford.edu/ schuetze/information–retrieval–book.html, 2007.

[31] C. Manning and H. Schütze. *Foundations of Statistical Natural Languate Processing*, (MIT Press, Cambridge, MA), 1999.

[32] A. Medina, C. Fraleigh, N. Taft, S. Bhattacharyya, and C. Diot. A taxonomy of ip traffic matrices. In *SPIE ITCOM: Scalability and Traffic Control in IP Networks II*, Boston, USA, August 2002.

[33] M. Molina, S. Niccolini, and N. G. Duffield. Comparative experimental study of hash functions applied to packet sampling. *ITC-19,*, (Beijing), 2005.

[34] X. Nguyen, M. J. Wainwright, and M. I. Jordan. On divergences, surrogate loss functions and decentralized detection. *Technical Report 695, Department of Statistics, UC Berkeley*, 2005.

[35] M.R. Osborne, B. Presnell, and Turlach B.A. A new approach to variable selection in least squares problems. *Journal of Numerical Analysis*, (20(3)):389–403, 2000a.

[36] M.R. Osborne, B. Presnell, and Turlach B.A. On the lasso and its dual. *Journal of Computational and Graphical Statistics*, (9(2)):319–337, 2000b.

[37] S. Rosset, J. Zhu, and T. Hastie. Boosting as a regularized path to a maximum margin classifier. *Journal of Machine Learning Research*, 5:941–973, 2004.

[38] B. Scholkopf and A. J. Smola. Learning with kernels: support vector machines, regularization, optimization, and beyond. *Cambridge, Mass: MIT Press*, 2002.

[39] T Shi, B. Yu, E. Clothiaux, and A. Braverman. Cloud detection over ice and snow using misr data. *Technical Report 663, Statistics Department, UC Berkeley*, 2004.

[40] T Shi, B. Yu, E. Clothiaux, and A. Braverman. Arctic cloud detection using multi-angle satellite misr data. *in preparation*, 2006.

[41] A. Soule, A. Lakhina, N. Taft, K. Papagiannaki, K. Salamatian, A. Nucci, M. Crovella, and C. Diot. Traffice matrices: Balancing measurement, modeling and inference. In *ACM Sigmetrics*, June 2005.

[42] Y. Vardi. Network tomography: Estimating source-destination traffic intensities from link data. *Journal of the American Statistical Association*, 91:365–377, 1996.

[43] E. J. Wegman and D. J. Marchette. On some techniques for streaming data: a case study of internet packet headers. *J. Computational and Graphical Statist.*, 12(4):893–914, 2003.

[44] J. Welling and M. Dearthick. Visualization of large multi-dimensional datasets. *Virtual observations of the future, Astronomical Society of the Pacific Conference Series*, 225(284–290), 2001.

[45] Y. Zhang, M. Roughan, C. Lund, and D. Donoho. An information-theoretic approach to traffic matrix estimation. In *ACM SIGCOMM*, 2003.

[46] P. Zhao and B. Yu. Boosted lasso. *Technical Report, Statistics Department, UC Berkeley*, 2004.

[47] P. Zhao and B. Yu. On model selection consistency of lasso. *Technical Report, Statistics Department, UC Berkeley*, 2006.