UC Davis Statistical Sciences Symposium 2013: Analysis of Complex and Massive Data

Co-Sponsored by the MPS Deans Office of UC Davis, NSF via an RTG training grant, and the Graduate School of Management of UC Davis

April 13, 2013

Room 1147, Mathematical Sciences Building, University of California, Davis

08:30am-09:00am: Registration, Breakfast

Opening Session (Chair: Wolfgang Polonik)

09:00am-09:15am: Welcome

Winston Ko, Dean of Mathematical and Physical Sciences

Hans-Georg Müller, Chair of the Department of Statistics

09:15am-09:45am: Rachel Schutt (Johnson Research Labs and adjunct at Statistics, Columbia University)

09:45am-10:15am: Owen Carmichael (Neuroscience and Computer Science, UC Davis)

Break

Session 2 (Chair: Christiana Drake)

10:40am-11:10am: Christopher Genovese (Statistics, Carnegie Mellon University)

11:10am-11:40am: Xiaodan Fan (Statistics, Chinese University Hong Kong)

11:40am-12:10pm: Ken Joy (Computer Science, UC Davis)

Lunch

Session 3 (Chair: Debashis Paul)

01:40pm-02:10pm: Tony Tyson (Physics, UC Davis)

02:10pm-02:40pm: Anna Michalak (Global Ecology, Stanford University)

02:40pm-03:10pm: Paul Baines (Statistics, UC Davis)

Break

Closing Session (Chair: Alexander Aue)

03:35pm-04:05pm: Xiaotong Shen (Statistics, University of Minnesota)

04:05pm-04:35pm: Jie Peng (Statistics, UC Davis)

04:35pm-05:05pm: Fernando Perez (Neuroscience, UC Berkeley)

05:05pm-05:15pm: Closing Remarks

06:00pm-09:00pm: Reception at John Natsoulas Gallery (521 1st St.)

Abstracts:

Paul Baines (Statistics, UC Davis)

Statistics and Scalability: Parametrization, Algorithmic Efficiency and Parallelization

The history of statistics is as intertwined with the history of computation as almost any discipline. As we move into the age of "big data", statistics must again adapt to both the computational challenges of modern datasets and the computational opportunities afforded by modern computational resources. In this talk we will examine the interface between statistical modeling, algorithm development and programming implementation in the context of the EM algorithm. We show how the choice of parametrization for the EM algorithm has important implications for both the parallelizability of the computation as well as the theoretical convergence rate. By drawing on classical statistical principles such as sufficiency and ancillarity, and their connection to parallelizability, we discuss general strategies for designing scalable statistical computation.

Owen Carmichael (Neuroscience and Computer Science, UC Davis)

Estimating Brain Networks From Neuroimaging Data: The Wild West Era

In vivo brain imaging data is increasingly used to represent the human brain in terms of networks: dispersed groups of brain regions that show coordinated patterns of activity, glucose metabolism, growth or atrophy over time, and other scientifically relevant properties. Network analyses have provided new insights into the maturation and aging of the healthy brain, as well as the brain changes that underlie neurological disorders such as Alzheimer's disease; funding agencies are increasingly investing in large-scale projects such as the Human Connectome Project that seek to chart the structure and function of such brain networks. But the field currently has a chaotic "wild west" feel: divergent scientific results abound, a wide array of ad hoc quantitative techniques are employed, and there are few widely agreed upon guidelines for best computational and statistical practices. There is much room for statisticians to develop theoretically-sound estimators of brain network connectivity. I will give a brief overview of current approaches to assessing brain network connectivity and describe some current efforts to provide estimation techniques that have a firmer theoretical foundation. Joint work with Hans Mueller, Jane-Ling Wang, Debashis Paul, Jie Peng, and Luda Sakhanenko.

Xiaodan Fan (Statistics, Chinese University Hong Kong)

Model-based Integration of Heterogeneous Datasets for Clustering

In the new era of big data, we often have multiple datasets measured from different angles for the same set of objects with the goal of elucidating the complicate grouping structures of these objects. Traditional clustering algorithms are designed for either vector data or relational data alone, therefore incapable of efficient integration of these two kinds of datasets measured from different angles. We aim to provide a principled clustering framework with such data integration function. Hierarchical models together with a Bayesian computing approach are introduced for this task.

Christopher Genovese (Statistics, Carnegie Mellon University)

Manifold Surrogates and Ridge Pursuit

Spatial data and high-dimensional data, such as collections of images, often contain high-density regions that concentrate around lower dimensional structure. In many cases, these structures are well-modeled by smooth manifolds, or collections of manifolds. For example, the distribution of matter in the universe at large scales forms a web of intersecting clusters (0-dimensional manifolds), filaments (1-dimensional manifolds), and walls (2-dimensional manifolds), and the shape and distribution of these structures have cosmological implications.

I will discuss new theory and methods for the problem of estimating manifolds (and collections of manifolds) from noisy data in the embedding space. The noise distribution has a dramatic effect on the performance (e.g., minimax rates) of estimators that is related to but distinct from what happens in measurement-error problems. Some variants of the problem are “hard” in the sense that no estimator can achieve a practically useful level of performance. I will show that in the “hard” case, it is possible to achieve accurate estimators for a suitable surrogate of the unknown manifold that captures many of the key features of the object. And I will describe efficient methods for estimating surrogates, characterizing “hyper-ridges” in many dimensions, and assessing uncertainties in the results.

Ken Joy (Computer Science, UC Davis)

Inverse Problems in Large-Scale Scientific Visualization

Visualization plays a powerful role in addressing many forms of information uncertainty. Many visualization problems can be classified as inverse problems, where one must produce visualizations that maximize the amount of information, under the constraints of preferring a given solution when many solutions fit the data equally. These inverse problems can frequently be solved by applying Bayesian methods. In this talk, we present the application of Bayesian methods to two distinct inverse problems: finding the material interface between multiple materials in fluid flow applications, and discovering uncertainty in numerical ensemble forecasting. These problems represent a class of problems that are typical in large-scale scientific visualization.

Anna M. Michalak (Global Ecology, Carnegie Institution for Science, Stanford)

Big data meets big models in carbon cycle science: Spatiotemporal tools for constraining the CO2 budget from atmospheric observations

Predicting future changes to the global carbon cycle (and therefore climate) and quantifying anthropogenic emissions of carbon dioxide both require an understanding of net carbon sources and sinks, and their variability, across a variety of spatial and temporal scales. This need highlights the importance of understanding the spatial and temporal scale-dependence of parameters controlling this variability, and developing methods for using data collected at multiple scales to infer carbon fluxes.

This presentation will describe ongoing work examining the carbon cycle from an atmospheric perspective in three major areas: spatiotemporal inverse modeling tools for characterizing uptake and emissions of carbon dioxide at the Earth surface using atmospheric observations of CO2; regression approaches for understanding the scale-dependence of processes controlling carbon flux variability; and mapping tools for massive datasets used for obtaining global atmospheric CO2 distributions from satellite observations. Common challenges across these applications include: the multiscale and nonstationary space-time variability of the signal; the lack of direct observations of CO2 exchange at the Earth surface at the scales of highest interest; the diffusive nature of atmospheric transport, which links the unobserved CO2 exchange at the Earth surface with observations of atmospheric CO2; the large to massive data volumes; and the large to massive size of the state space.

Jie Peng (Statistics, UC Davis)

High-Dimension Gaussian Graphical Model Building with Re-Sampling based Methods

Regularization techniques are widely used for tackling high-dimension-low-sample-size problems. Yet, finding the right amount of regularization is challenging, especially in the unsupervised setting, where traditional methods such as BIC or cross-validation often result in too many false positives. In this talk, we first introduce Gaussian graphical models (GGMs) and its inference under the high-dimension regime. We then propose a resampling based method to directly control the false discovery rates (FDRs) of edge detection. The idea is to fit a mixture distribution for the selection frequencies and then estimate the FDRs. The proposed method is illustrated through numerical examples.

Fernando Perez (Henry H. Wheeler Jr. Brain Imaging Center, UC Berkeley)

IPython: tools for the entire lifecycle of research computing

The IPython project (http://ipython.org) provides a rich architecture for interactive computing with:

- Terminal-based and graphical interactive consoles.
- A web-based Notebook system with support for code, text, mathematical expressions, inline plots and other rich media.
- Easy to use, high performance tools for parallel computing.

While the focus of the project is Python, its architecture is designed in a language-agnostic way to facilitate interactive computing in any language. This allows users to mix Python with R, Octave, Julia, Ruby, Perl, Bash and more.

In this talk, I will show how IPython supports all stages in the lifecycle of a scientific idea: individual exploration, collaborative development, large-scale production using parallel resources, publication and education. In particular, the IPython Notebook provides an environment for "literate computing" with a tight integration of narrative and computation. These Notebooks are stored an open document format that provides an "executable paper": notebooks can be version controlled, exported to HTML or PDF for publication, and used for teaching.

Rachel Schutt (Johnson Research Labs and adjunct at Statistics, Columbia University)

The Challenges to Statistics Posed by Data Science

Over the last few years, the job title "Data Scientist" has emerged in Industry in companies such as Facebook, Linked In and Google. While the Data Science phenomenon may have come from industry, universities are now teaching Data Science courses, and beginning to open institutes and academic programs in Big Data and Data Science. Thus the distinction between Statistics and Data Science is not simply expressive of the differences between academia and industry, because now we have both Statistics and Data Science in academic institutions. This talk will explore the challenges to traditional Statistics departments posed by Data Science from two perspectives, institutional and research-wise. (1) Isn't Statistics the Science of Data? (2) Some suggestions for open research problems that statisticians could explore by applying classical statistical techniques (sampling, experimental design, causal modeling) in the context of massive data sets and new types of data being generated in technology companies such as Google.

Xiaotong Shen (Statistics, University of Minnesota)

Personalized information filtering

Personalized information filtering extracts the information specifically relevant to a user, based on the opinions of users who think alike or the content of the items that a specific user prefers. In this talk, we discuss latent models to utilize additional user-specific and content-specific predictors, for personalized prediction. In particular, we factorize a user-over-item preference matrix into a product of two matrices, each having the same rank as the original matrix. On this basis, we seek a sparsest latent factorization from a class of overcomplete factorizations, possibly with a high percentage of missing values. A likelihood approach is discussed, with an emphasis towards scalable computation. Examples will be given to contrast with popular methods for collaborative filtering and contented-based filtering. This work is joint with Y. Zhu and C. Ye.

Tony Tyson (Physics, UC Davis)

Data analysis challenges and opportunities with the LSST sky survey

The history of astronomy has taught us repeatedly that there are surprises whenever we view the sky in a new way. With an unprecedented combination of sky coverage, cadence, and depth, the LSST makes it possible to attack high-priority scientific questions that are far beyond the reach of any existing or planned facility. Much of LSST's science will rely on the statistical precision obtainable with billions of objects. For the first time, the sky will be surveyed deep and fast, opening a new window on a universe of faint moving and distant exploding objects, as well as exploring the physics of dark energy with eight types of probes. Thirty TB of data will be produced nightly, and over 2 million alerts will be issued nightly within 60 seconds of a transient detection. Over 20 trillion photometric measurements will be made of 20 billion objects, populating a database of high dimensionality. Mining these data quickly and efficiently for the known knowns and the unknown unknowns presents unprecedented opportunities as well as object classification algorithm challenges.