THE FIELDS INSTITUTE

DISTINGUISHED LECTURE SERIES IN STATISTICAL SCIENCE

BIN YU

University of California at Berkeley

Stability

APRIL 23, 2015 AT 3:30 P.M. AT THE FIELDS INSTITUTE

Reproducibility is imperative for any scientific discovery. More often than not, modern scientific findings rely on statistical analysis of high-dimensional data. At a minimum, reproducibility manifests itself in stability of statistical results relative to "reasonable" perturbations to data and to the model used. Jacknife, bootstrap, and cross-validation are based on perturbations to data, while robust statistics methods deal with perturbations to models.

In this talk, a case is made for the importance of stability in statistics. Firstly, we motivate the necessity of stability of interpretable encoding models for movie reconstruction from brain fMRI signals. Secondly, we find strong evidence in the literature to demonstrate the central role of stability in statistical inference. Thirdly, a smoothing parameter selector based on estimation stability (ES), ES-CV, is proposed for Lasso, in order to bring stability to bear on cross-validation (CV). ES-CV is then utilized in the encoding models to reduce the number of predictors by 60% with almost no loss (1.3%) of prediction performance across over 2,000 voxels. Last, a novel "stability" argument is seen to drive new results that shed light on the intriguing interactions between sample to sample variability and heavier tail error distribution (e.g. double-exponential) in high dimensional regression models with p predictors and n independent samples. In particular, when p/n belongs to (0.3, 1) and error is double-exponential, the Least Squares (LS) is a better estimator than the Least Absolute Deviation (LAD) estimator.

This talk draws materials from papers with S. Nishimoto, A. T. Vu, T. Naselaris, Y. Benjamini, J. L. Gallant, with C. Lim, and with N. El Karoui, D. Bean, P. Bickel, and C. Lim.

The multi-facets of a data science project to answer: how are organs formed?

APRIL 24, 2015 AT 11 A.M. AT THE FIELDS INSTITUTE

Genome wide data reveal an intricate landscape where gene actions and interactions in diverse spatial areas are common both during development and in normal and abnormal tissues. Understanding local gene networks is thus key to developing treatments for human diseases. Given the size and complexity of recently available systematic spatial data, defining the biologically relevant spatial areas and modeling the corresponding local biological networks present an exciting and on-going challenge. It requires the integration of biology, statistics and computer science; that is, it requires data science.

In this talk, I present results from a current project co-led by biologist Erwin Frise from Lawrence Berkeley National Lab (LBNL) to answer the fundamental systems biology question in the talk title. My group (Siqi Wu, Antony Joseph, Karl Kumbier) collaborates with Dr. Erwin and other biologists (Ann Hommands) of Celniker's Lab at LBNL that generate the Drosophila spatial expression embryonic image data. We leverage our group's prior research experience from computational neuroscience to use appropriate ideas of statistical machine learning in order to create a novel image representation decomposing spatial data into building blocks (or principal patterns). These principal patterns provide an innovative and biologically meaningful approach for the interpretation and analysis of large complex spatial data. They are the basis for constructing local gene networks, and we have been able to reproduce almost all the links in the Nobel-prize winning (local) gap-gene network. In fact, Celniker's lab is running knock-out experiments to validate our predictions on gene-gene interactions. Moreover, to understand the decomposition algorithm of images, we have derived sufficient and almost necessary conditions for local identifiability of the algorithm in the noiseless and complete case. Finally, we are collaborating with Dr. Wei Xu from Tsinghua Univ to devise a scalable open software package to manage the acquisition and computation of imaged data, designed in a manner that will be usable by biologists and expandable by developers.

Bin Yu is Chancellor's Professor in the Departments of Statistics and of Electrical Engineering & Computer Science at the University of California at Berkeley. Her current research interests focus on statistics and machine learning theory, methodologies, and algorithms for solving high-dimensional data problems. Her group is engaged in interdisciplinary research with scientists from genomics, neuroscience, and remote sensing. She is Member of the U.S. National Academy of Sciences and Fellow of the American Academy of Arts and Sciences. She was a Guggenheim Fellow in 2006, an Invited Speaker at ICIAM in 2011, and the Tukey Memorial Lecturer of the Bernoulli Society in 2012. She was President of IMS (Institute of Mathematical Statistics) in 2013-2014.

For more information, please visit:
www.fields.utoronto.ca/programs/scientific/14-15/DLSS/

NSERC CRSNG

Ontario

THE FIELDS INSTITUTE FOR RESEARCH IN MATHEMATICAL SCIENCES
222 College Street, Second Floor, Toronto, Ontario, M5T 3J1 • www.fields.utoronto.ca • 416-348-9710