Department of Statistics - Statistics Seminar – Fall 2013

Statistics Seminars are on Mondays
Time: 12:00 - 1:00 PM
Location: Room 903, 1255 Amsterdam Avenue,
Tea and Coffee will be served before the seminar at 11:30 AM, Room 1025

9/9/2013

Speaker: Prof. Michael Eichler, Maastricht University

Title: "Causal inference from multivariate time series"

Abstract:

In time series analysis, inference about cause-effect relationships among multiple time series is commonly based on the concept of Granger causality, which exploits temporal structure to achieve causal ordering of dependent variables. One major and well known problem in the application of Granger causality for the identification of causal relationships is the possible presence of latent variables that affect the measured components and thus lead to so-called spurious causalities.

We present a new graphical approach for describing and analysing Granger-causal relationships in multivariate time series that are possibly affected by latent variables. It is based on mixed graphs in which directed edges represent direct influences among the variables while dashed edges---directed or undirected---indicate associations that are induced by latent variables. We show how such representations can be used for inductive causal learning from time series and discuss the underlying assumptions and their implications for causal learning. Finally we will discuss tetrad constraints in the time series context and how they can be exploited for causal inference.

9/16/2013

Speaker: Prof. Michael Stein, Ralph and Mary Otis Isham Professor
Department of Statistics and the College, and Master, Physical Sciences Collegiate Division, University of Chicago

"Likelihood Approximation for Large Environmental Data Sets"

Abstract:

This talk will discuss methods for fitting Gaussian process models to large environmental datasets, both spatial and spatial-temporal, with a particular focus on ungridded data. The talk will consider approximations to the likelihood, such as composite likelihood, that reduce computations, and methods from numerical analysis, such as preconditioned conjugate gradient, that can reduce the memory and calculations required to compute exact or approximate likelihoods. Comparisons to alternative approaches to reducing computations, such as low rank methods, will be given. The talk will conclude with some speculations on the use of various approximate likelihoods for frequentist and Bayesian inference.

Nitis Mukhopadhyay (will be giving *two talks)
Department of Statistics
University of Connecticut-Storrs

Talk 1
Monday, September 23, 2013
Time:12:00 - 1:00 PM
Location: Room 903 SSW

"On Determination of an Appropriate Pilot Sample Size"

Abstract:
Any multistage sampling strategy requires a practitioner to begin with an initial dataset with a pilot size m, an appropriately chosen number. Under purely sequential sampling, a choice of m may not be very difficult to come up with as long as m is reasonably small. It is so because here one moves forward step-by-step adaptively.
Under multi-stage sampling, on the other hand, especially under two-stage sampling, the choice of an appropriate pilot size m happens to be extremely crucial. In such situations, understandably m should not be too large or too small. Our common wisdom may dictate that, but what choice of m qualifies to be labeled not “too large” or “too small”, especially when the “optimal fixed sample size” remains unknown?
We shall explore some concrete ideas based on (i) large-sample approximations, (ii) Fisher information, and then contrast them. Illustrations will be given.

Talk 2
Monday, September 23, 2013
Time: 1:20 - 2:20 PM
Location: Room 903 SSW

"Any Magic Left in Teaching Probability and Inference? You Decide"

Abstract:
Lately, I have written rather extensively on various topics in probability, statistical inference, and linear models. Those research topics and ensuing publications originated primarily from teaching graduate level, and some undergraduate level, courses in statistics.
In this presentation, I will touch upon invariant tests and Rao-Blackwell-Lehmann-Scheffe type theorems, as well other interesting stuff with regard to Student’s t-distributions, correlations, independence, and multivariate normality. I may, however, change the order of topics as needed. While I may not be able to touch upon every single aspect with equal emphasis, I will highlight as many specific interesting problems as possible along with their origin and resolutions.
Teaching-related ideas may lead to research and I find this process very exciting. I hope to share such excitement and positive experiences with my audience. I am reasonably sure that this presentation would be accessible to students.

9/30/2013

Ronny Luss, IBM Research

Title: Efficient Regularized Isotonic Regression

Abstract: Isotonic regression is a nonparametric approach for fitting monotonic models to data that has been widely studied from both theoretical and practical perspectives. However, this approach encounters computational and statistical overfitting issues in higher dimensions. To address both concerns we present an algorithm, which we term Isotonic Recursive Partitioning (IRP), for isotonic regression based on recursively partitioning the covariate space through the solution of progressively smaller "best cut" subproblems. This creates a regularized sequence of isotonic models of increasing model complexity that converges to the global isotonic regression solution. Models along this sequence are often more accurate than the unregularized isotonic regression model because of the complexity control they offer. We quantify this complexity control through estimation of degrees of freedom along the path. Furthermore, we show that IRP for the classic l2 isotonic regression can be generalized to convex differentiable loss functions such as Huber's loss. In another direction, we regularize isotonic regression by the range of model predictions. This problem can be formulated as a lasso problem in the very high dimensional basis of uppersets in the covariate space. We show how this problem can be solved by a generalization of the LARS algorithm, and compare this regularization behavior with that of IRP. An example of modeling gene-gene interactions in human disease will conclude the talk. This is joint work with Saharon Rosset and Moni Shahar.

10/21/2013

Speaker: Prof. Andrew Barron, Yale University

"Communication by Statistical Regression"

High rate reliable communication is a challenging statistical task. Fundamental limits were originally established by Shannon 65 years ago and some progress in practice has occured in the last 20 years. However, proof that a practical scheme is reliable at all rates below Shannon Capacity has been elusive. Here we discuss our development of Sparse Superposition Codes and their properties in reaching these objectives.

This is joint work with Antony Joseph and Sanghee Cho.

11/11/2013

Jiancheng Jiang, UNC Charlotte

Title: Modeling Multivariate Nonlinear Time

Abstract. Vector time series data widely exist in practice. For modeling such data, one would generally use multivariate models rather than univariate models. In this talk we propose multivariate functional-coefficient regression models with heteroscedasticity to fit vector time series data. Different estimation methods are employed to estimate the unknown coefficient matrices. Asymptotic normality of the proposed estimators is established. Several practical problems such as bandwidth selection are also considered. Simulations are conducted to show that the proposed estimation procedures well capture nonlinear structures of coefficients. A real dataset is used to illustrate the value of the proposed methodology. A number of open topics worthy of further study are given in a discussion section.

11/25/2013

Dr. Kenny Shirley

Title: Maximum Entropy Summary Trees

Abstract- We present a method for summarizing and visualizing large, tree-structured data. Many data sets can be represented by a rooted, node-weighted tree, such as a company organizational chart, clicks on webpages, flows to and from IP addresses, or hard disk file structures, for example, where the weights represent some attribute of interest for each node. If such a tree has thousands (or millions) of nodes, it is difficult to visualize on a single sheet or paper or computer screen. We define a way to aggregate the weights of a large, n-node tree into a smaller k-node “summary tree” (where k is something like 50 or 100), and we present a dynamic programming algorithm to compute the summary tree with maximum entropy among all summary trees of a given size, where the entropy of a node-weighted tree is defined as the entropy of the discrete probability distribution whose probabilities are the normalized node weights. We discuss and provide examples of how this algorithm produces useful visualizations (both static, and interactive visualizations using d3), and may also be optimal for certain kinds of data analysis tasks. The talk will be heavy on visualization techniques, but I will also spend some time discussing statistical issues related to hierarchical data.

This is joint work with Howard Karloff.

12/2/2013

Jose Zubizaretta, Columbia University Business School

"Using Mixed Integer Programming for Matching in Observational Studies: Effect of the 2010 Chilean Earthquake on Posttraumatic Stress"

Abstract:
Matching is a widely used method of adjustment for observed covariates in observational studies. With matching, one attempts to replicate a randomized experiment by finding matched groups that look alike in terms of their observed covariates, as if they were randomly assigned to treatment. However, most matching methods involve a considerable amount of guesswork because they do not target covariate balance directly. In a recent paper (Zubizarreta 2012), I describe a new matching method based on mixed integer programming that overcomes this issue and targets covariate balance directly. By either optimizing or constraining several measures of covariate imbalance simultaneously, this new method can directly balance univariate moments (such as means, variances, and skewness), multivariate moments (such as correlations), and statistics (such as the Kolmogorov-Smirnov statistic). Furthermore, while balancing several of these measures, it makes possible to match with fine balance for more than one nominal covariate, whereas most matching algorithms can finely balance only a single nominal covariate. By virtue of optimality, matching based on mixed integer programming also tells whether certain forms of covariate balance are feasible with the data at hand, and therefore whether the data supports certain causal comparisons. In this talk I will go over the basics of matching, and illustrate this

new method in an observational about the effect of the 2010 Chilean earthquake on post traumatic stress (Zubizarreta et al 2013). A new R package called mipmatch implements this method with a wide range of applications in the medical and social sciences.

References:

- Zubizarreta, J. R., (2012), “Using Mixed Integer Programming for Matching in an Observational Study of Acute Kidney Injury after Surgery,” Journal of the American Statistical Association, 107, 1360–1371.
- Zubizarreta, J. R., Cerda, M., and Rosenbaum, P. R. (2013), “Effect of the 2010 Chilean Earthquake on Posttraumatic Stress: Designing an Observational Study to be Less Sensitive to Unmeasured Biases,” Epidemiology, 24, 79–87.

12/9/2013

Jennifer Hill, NYU

Title: Assessing sensitivity to an unmeasured confounder in the presence of a nonlinear data generating process.

Abstract:

Researchers seeking answers to causal questions frequently are unable to perform randomized experiments due to ethical, legal, or logistical constraints. However observational studies typically require the untestable assumption that all confounders have been measured (the ignorability assumption). There is no reason to believe that this assumption generally holds in practice. One way to ascertain how much confidence we should have in these observational results is to quantify the extent to which our estimates might be different in the presence of an unobserved confounder. A range of methods exist to perform these types of sensitivity analyses but there are tradeoffs among them with regard to features such as their assumptions and interpretability. Methods that are focused on the calibration and interpretability of the sensitivity parameters often make unreasonable assumptions with regard to the parametric specification of the data generating process. We extend one such method to accommodate more flexible functional forms by embedding a nonparametric Bayesian algorithm (Bayesian Additive Regression Trees) within a more standard sensitivity analysis algorithm. Performance compared to close alternatives is assessed with regard to bias and RMSE and the functionality is illustrated using an applied example.