Conference on Predictive Inference and Its Applications
May 7 and 8, 2018
Iowa State University

Speaker Abstracts

Efficient Use of Electronic Health Records for Translational Research
Tianxi Cai
Harvard University

While clinical trials remain a critical source for studying disease risk, progression and treatment response, they have limitations including the generalizability of the study findings to the real world and the limited ability to test broader hypotheses. In recent years, due to the increasing adoption of electronic health records (EHR) and the linkage of EHR with specimen bio-repositories, large integrated EHR datasets now exist as a new source for translational research. These datasets open new opportunities for deriving real-word, data-driven prediction models of disease risk and progression as well as unbiased investigation of shared genetic etiology of multiple phenotypes. Yet, they also bring methodological challenges. For example, obtaining validated phenotype information, such as presence of a disease condition and treatment response, is a major bottleneck in EHR research, as it requires laborious medical record review. A valuable type of EHR data is narrative free-text data. Extracting accurate yet concise information from the narrative data via natural language processing is also challenging. In this talk, I'll discuss various statistical approaches to analyzing EHR data that illustrate both opportunities and challenges. These methods will be illustrated using EHR data from Partners Healthcare.

Learning About Learning
Ronald Christensen
University of New Mexico

This talk will use Support Vector Machines for binary data to motivate a discussion of the wider role of reproducing kernels in Statistical Learning. If time permits, comments on bagging and boosting will be included.

Predictive Inference from Replicated Network Data

David Dunson
Duke University

Replicated network data consist of repeated observations of relationships among a common set of nodes. In one motivating application, the nodes correspond to regions of interest (ROIs) in the human brain and the replicates to different individuals brains. In another application, the nodes correspond to players on a soccer team and the replicates to different soccer matches. As opposed to the enormous literature on methods for modeling and analysis of a single (e.g., social) network, there has been very little consideration of replicated network data. Our particular focus is on developing statistical methods for interpretable prediction inference from replicated networks. For example, we want to infer lower-dimensional structure in brain networks predictive of human traits or related to human exposures and we want to identify soccer passing network motifs predictive of goal scoring. After introducing replicated networks (RNs) and spatial RNs (SRNs), I propose some useful factorizations of the data that can be exploited to characterize the complex data in terms of simpler more interpretable pieces. This can be done to embed the networks into lower-dimensional Euclidean spaces, to identify modes of variability, and identify common motifs. Exploiting these representations, we develop efficient algorithms that can be implemented for large datasets and apply these algorithms to human brain connectome and soccer network data. The results show remarkable ability to predict human traits and exposures based on brain structure, while also inferring interesting substructure and motifs – also in the soccer application.

High Dimensional Predictive Inference
Ed George
University of Pennsylvania

Let X and Y be independent p-dimensional multivariate normal vectors with common unknown mean μ. Based on only observing $X = x$, we consider the problem of obtaining a predictive density $\hat{p}(y \mid x)$ for Y that is close to $p(y \mid \mu)$ as measured by expected Kullback-Leibler loss. This is the predictive version of the canonical problem of estimating μ under quadratic loss, and we see that a strikingly parallel theory exists for addressing it.

To begin with, a natural "straw man" procedure for this problem is the (formal) Bayes predictive density $\hat{p}_U(y \mid x)$ under the uniform prior $\pi_U(\mu) \equiv 1$, which is best invariant and minimax. It turns out that there are wide classes procedures that dominate $\hat{p}_U(y \mid x)$ including Bayes predictive densities under superharmonic priors. Indeed, any Bayes predictive density will be minimax if it is obtained by a prior yielding a marginal that is superharmonic or whose square root is superharmonic. For the characterization of admissible procedures for this problem, the class of all generalized Bayes rules here is seen to form a complete class, and easily interpretable conditions are seen to be sufficient for the admissibility of a formal Bayes rule.

Moving on to the multiple regression setting, our results are seen to extend naturally. Going further, we address the situation where there is model uncertainty and only an unknown subset of the predictors in A is thought to be potentially irrelevant. For this purpose, we develop multiple shrinkage predictive estimators along with general minimaxity conditions. Finally, we provide an explicit example of a minimax multiple shrinkage predictive estimator based on scaled harmonic priors. (This is joint work with Larry Brown, Feng Liang and Xinyi Xu).

Prediction of Complex Traits in Animal Breeding: From Henderson to Bayesian Machine Learning

Daniel Gianola
University of Wisconsin

Large amounts of genomic information, e.g., single nucleotide polymorphisms and sequences, are available in animals and plants. The situation is one in which there may be thousands of genotyped individuals (with potentially millions of predictor variables) and millions of un-genotyped animals, for which only pedigree is available. Animals may be phenotyped or un-phenotyped (e.g., bulls cannot produce milk yet). Animal breeders have exploited this type of data (and pedigrees) and used prediction methods that conform mainly to the strong assumption of additive inheritance, amply corroborated by theoretical and empirical evidence.

Charles Henderson set the foundations for prediction of complex traits in agriculture, albeit mainly under additive inheritance and assuming known variance-covariance structure. He introduced the best linear unbiased predictor (BLUP); his definition of unbiasedness, however, is not what most statisticians have in mind when estimating unknown quantities. BLUP is also a conditional posterior mean under Gaussian assumptions, a penalized maximum likelihood estimator, a reproducing kernel Hilbert spaces regression (RKHS) and a linear neural network. Most importantly, it has been useful, mainly because it is flexible (model can be amended easily) and it can be applied in cross-sectional, longitudinal, spatial and multiple-trait setting. Animal breeders have developed algorithms for solving hundreds of millions of equations iteratively and have found solutions for combining information from both genotyped and un-genotyped individuals. This flexibility is not shared by more modern and "sophisticated" (remember KISS) methodology.

Henderson's pedigree based BLUP has morphed into GBLUP (G: genomics): the mostly used method for prediction in animal breeding, now in the era of genomic selection. Breeders have also used a battery of Bayesian linear regression models based on a suite of priors on marker effects ranging from Gaussian-inverse Wishart to spike and slab models (the "Bayesian alphabet"). While Bayesian methods provide excellent prediction machines and sometimes outperform GBLUP, they must be used with care for inference; the "small n large p" problem may create a false illusion of effective learning when, in fact, it is the prior that makes the difference (remember Kempthorne!).

Some challenges posed by the growing ensemble of phenotypic and genomic data are not well met by additive (inheritance) linear predictors, mainly because of inability in coping with high-dimensional interactions among genes, as well as between genes and environmental variables. The conflict with knowledge from regulatory biology and systems approaches is obvious, as interactions and nonlinearity are pervasive. A sensible alternative is provided by machine learning approaches. So far, it appears that RKHS is the most stable prediction machine. Bayesian neural networks often fail (unless run using MCMC, impractical in animal breeding industries), and early evidence from application of hot "deep learning" is unconvincing. Unless stronger evidence is provided by empirical studies, it appears that Henderson's BLUP will continue dominating the prediction scene in animal breeding. Robinson (1991) was right: "that BLUP is good".

Model-Based Prediction in General and as Applied to the Outcomes of College Football Games

David Harville
Iowa State University

Success in model-based prediction depends critically on knowledge of the underlying phenomena and on being able to distinguish what is important from what is not. Typically, the data are observational in nature, giving rise to potential pitfalls of a kind that can be difficult to recognize and avoid. Performance in repeated application is likely to be important; at a minimum, this requires that the predictions be what Dawid referred to as well-calibrated.

For purposes of illustration, the prediction of the outcomes of college football games will be discussed. Some numerical results will be presented for the 2017 season. Among the issues encountered in this application is the lumpiness of the distribution of scores noted by Mosteller; this issue has been exacerbated by the introduction of overtime.

Maximizing the Usefulness of Statistical Classifiers for Two Populations

Daniel Jeske
University of California, Riverside

The usefulness of two-class statistical classifiers is limited when one or both of the conditional misclassification rates is unacceptably high. Incorporating a neutral zone region into the classifier provides a mechanism to refer ambiguous cases to follow-up where additional information might be obtained to clarify the classification decision. Through the use of the neutral zone region, the conditional misclassification rates can be controlled and the classifier becomes useful. An application to prostate cancer will be used to illustrate how neutral zone regions can extract utility from a potentially disappointing classifier that might otherwise be abandoned.

Predicting the Emotions of Tennis Players from Single-Camera Video

Stephanie Kovalchik
Victoria University and Tennis Australia

Mentality has been one of the most elusive concepts in quantitative sport performance analysis. In this talk, I present a framework for capturing emotional information for tennis players from match broadcasts. The framework yields predictions for seven emotional states relevant to sport: anxiety, anger, annoyance, dejection, elation, focus, and fired up. Two feature sets are included in the model training for each emotion in sport: the Facial Action Coding System and 17 facial action units. Multiple prediction approaches were trained and tested using these features using a labeled dataset of 1,700 facial images of professional male and female tennis players extracted from 505 match videos. We applied the prediction models to establish emotional profiles for the Big 4 (Roger Federer, Rafael Nadal, Andy Murray, and Novak Djokovic) at the 2017 Australian Open. Rafael Nadal exhibited the most anxiety of the four players (32%, 95% CI 29 to 35%), while Roger Federer was the only player whose predominant state was neutral (24%, 95% CI 21 to 27%). All players except for Roger Federer showed significant emotional reactions to the outcomes of points. Further, several emotional states of Rafael Nadal and Novak Djokovic were significantly predictive of their chances of winning the next point. Our predictive framework for extracting emotional data from single-camera video in professional tennis shows the feasibility of bringing the quantitative study of the inner game into sports performance analysis.

Accounting for Uncertainties in Predictive Inference

Jing Lei
Carnegie Mellon University

When prediction is made from complex data using more advanced methods, adequately accounting for the model and sample uncertainty is necessary for reliable inference. I will introduce a new predictive inference framework based on a generic tool that converts any point estimator to an interval predictor. The resulting prediction bands have valid average coverage under essentially no assumptions, while retaining the optimality of the initial point estimator under standard assumptions. Secondly, I will discuss how to make better use of cross-validation for improved bias-variance trade-off, by adopting a hypothesis testing framework. Both methods are applicable to a very wide range of prediction algorithms.

Prediction with Confidence – a General Framework for Predictive Inference

Regina Liu
Rutgers University

We propose a general framework for prediction in which a prediction is in the form of a distribution function, called ‘predictive distribution function’. This predictive distribution function is well suited to prescribing the notion of confidence under the frequentist interpretation, and it can provide meaningful answers for prediction-related questions. A general approach under this framework is formulated and illustrated using the so-called confidence distributions (CDs). This CD-based prediction approach inherits many desirable properties of CD, including its capacity to serve as a common platform for directly connecting the existing procedures of predictive inference in Bayesian, fiducial and frequentist paradigms. We discuss the theory underlying the CD-based predictive distribution and related efficiency and optimality issues. We also propose a simple yet broadly applicable Monte-Carlo algorithm for implementing the proposed approach. This algorithm together with the proposed definition and associate theoretical development provide a comprehensive statistical inference framework for prediction. Finally, the approach is demonstrated by simulation studies and a real project on predicting the incoming volume of application submissions to a government agency. The latter shows the applicability of the proposed approach to dependent data settings.

This is joint work with Jieli Shen, Goldman Sachs, and Minge Xie, Rutgers University.

Prediction in Sports
Carl Morris
Harvard University

Predictive inference dovetails neatly with sports analysis, these being two key interests of David Harvilles career. Sports analytics have become a booming field with the onset of the internet, with fast computation, and with awareness of publications like the best seller “Moneyball”. Almost all professional teams now realize they will win more often if they develop and use better statistics to help assess each players impact on winning, and also to discover better game-time tactics and strategies.

Sports data now are widely available on-line for the major sports. These data are collected and organized carefully, usually arising from excellent experimental designs for balanced team schedules. They are supported by consistently applied rules across years. This talk considers predictive inferences conducted via models and methods checked predictively.

For baseball data, topics include viewing future performance levels as outcomes of random effects in multilevel models. Predicting future batting averages from previous averages on a group of players and comparing these estimates with actual outcomes is a well-studied example. Such multilevel models predict the sophomore slump phenomenon for the previous years rookie of the year in terms of regression toward the mean. Baseballs Markov model provides a valuable basis for considering and evaluating optimal tactics.

General predictive inference issues are considered. These include postdictive inferences, non-linear forecasts, and choosing estimates versus distributions as measures of predictive outcomes.

Getting Beyond the Mean in Predictive Inference
Hal Stern
University of California-Irvine

Predictive modeling in many domains attempts to draw inferences about the parameters characterizing performance of the set of units of interest. Examples include predictions of team strength in sports, animal breeding potential, and risk of disease incidence. Often units are ranked based on the posterior mean of the parameters or a related estimator. This can be misleading though in settings where there is considerable variation in the information available for different units. Alternative estimands can be useful in this case. Examples are provided in the settings described above.

Charles Roy Henderson: Farm Boy, Athlete, Scientist

Dale Van Vleck
University of Nebraska-Lincoln

Charles Henderson has had more impact on animal agriculture than any other person through universal use of his Mixed Model Equations to obtain predictors of genetic merit. Those equations also made possible obtaining BLUE and BLUP for many other applications. They provide the foundation for many statistical analyses packages. The presentation will include a brief biography and comments on steps along the unusual path that eventually led to his influence on statistical estimation and prediction. His early life and academic and athletic achievements in Iowa and the interim of 10 years after college to his PhD program in animal breeding and statistics at Iowa State College beginning in 1946 will be summarized. His scientific career began at age of 37 in 1948 as an associate professor of animal husbandry at Cornell University. His contributions to statistical analyses and especially to genetic prediction began immediately and include: Hendersons Methods 1, 2 and 3 for variance component estimation, his Mixed Model Equations to obtain BLUE and BLUP, and his remarkable algorithm to obtain easily the inverse of the genetic relationship matrix which is needed to be able to use information from all relatives for genetic prediction. His contributions should not be forgotten.

Computational Challenges in High-dimensional Prediction

Martin Wainwright
University of California-Berkeley

Prediction with large-scale data sets poses challenges that are both statistical and computational, and this talk provides two vignettes at this interface. In the first part, we discuss the use of randomized dimensionality reduction techniques, also known as sketching, for quickly obtaining approximate solutions to large-scale optimization problems. We demonstrate randomized algorithms that operate on highly compressed versions of the original data set, yet are guaranteed (with overwhelming probability) to have predictive performance as good as a method operating on the full data set. In the second vignette, we discuss the use of methods based on boosting and early stopping to fit predictive models. We show how early stopping acts as an algorithmic form of regularization, and provide data-dependent rules for when to stop. Based on joint work with Mert Pilanci (Univ. Michigan), Yuting Wei (UC Berkeley), Yun Yang (Univ. Florida), Fanny Yang (UC Berkeley).

Interaction Selection: Its Past, Present, and Future

Hao Helen Zhang
University of Arizona

In regression and classification problems, interaction selection plays an important role to improve model prediction and interpretability. The topic of interaction selection has recently drawn much attention in the literature, partially due to its wide applications in identifying important gene-gene and gene-environmental effects in biological and medical research. In this talk, we will first start with a comprehensive review on basic principles for interaction selection, classical methods, and their properties. Then we will discuss unique and enormous challenges associated with interaction selection for high dimensional data, from both computational and theoretical perspectives. A variety of penalty-based interaction selection methods for data of moderate dimensions will be reviewed. Finally, we present the recently works on scalable methods of interaction selection and screening for ultra-high dimensional settings. In particular, a new hierarchy-preserving regularization solution path algorithm will be introduced, along with some new theoretical and empirical results.

Spatial Prediction: A Graphical Overview

Dale Zimmerman
University of Iowa

One important application of predictive inference is to spatial prediction, i.e., to a setting where the observations and the predictands are indexed by their locations in a common physical space (often, but not always, Euclidean). Such a setting lends itself nicely to the graphical display (using maps over that space) of several important concepts of predictive inference. In this talk, we give an overview of spatial best linear unbiased prediction (BLUP) – also known as kriging – and we display several interesting concepts associated with it (the "screening effect," "perfect interpolation," and others). We highlight the impact that the spatial correlation function has on BLUP and the impact that estimating such a function has on empirical BLUP, and we consider the consequences of these for good spatial sampling design for prediction.