DSSV 2020
DATA SCIENCE, STATISTICS & VISUALISATION

July 29-31, 2020
SPEAKER TITLES/ABSTRACTS

Kohei Adachi
Osaka University, Japan

"Principal Component versus Factor Analyses with their Intermediate Procedure in Matrix Decomposition Formulation"

Principal component analysis (PCA) and factor analysis (FA) are time-honored dimension reduction methods and still indispensable tools now in statistical learning. Though PCA can be formulated in apparently different manners (e.g., Adachi, 2020), the PCA formulation in this talk is restricted to approximating a multivariate data matrix by the product of reduced-rank (RR) score and loading matrices. As PCA and FA can be applied to an identical data matrix, users may have the question which of PCA and FA should be used. Recent developments concerning the question are reviewed and a new procedure is proposed in this talk, which consists of three parts.

First, the matrix decomposition (MD) formulation of FA established recently is reviewed, in which FA is modeled as the sum of the above product of RR matrices and the matrix of the unique factors having one-to-one correspondences to variables. Here, each unique factor accounts for the variations in the corresponding variable remaining unexplained by the product of RR matrices (e.g., Adachi, 2019; Adachi & Trendafilov, 2018). In this formulation, FA differs from PCA only in that the former model has the unique factors.

Second, some inequalities contrasting the PCA and FA solutions for the same data set are reviewed, which are derived from the MD formulation of FA (Adachi & Trendafilov, 2019). The inequalities include the following two: [1] the sum of the squared loadings resulting in PCA □ the FA counterpart, and [2] the goodness-of-fit of FA □ that of PCA. The former [1] suggests that the users should use PCA who wish to obtain the loadings of larger magnitudes. On the other hand, [2] suggests that FA should be used if users wish to explain data better.

Inequality [2] follows from that the FA model has additional parameters, i.e., the unique factors, whose existence is not supposed in PCA. It allows us to consider a continuum from PCA (without unique factors) to FA (with them). Thus, I propose an intermediate procedure located midway on the continuum, in which the coefficients for the unique factors are sparsified so that some coefficients are estimated as zeros, i.e., the corresponding variables have not unique factors.

References:
Adachi, K. (2019). Factor analysis: Latent variable, matrix decomposition, and constrained uniqueness formulations. WIREs Computational Statistics, https://onlinelibrary.wiley.com/doi/abs/10.1002/wics.1458

Adachi, K. (2020). Matrix-based introduction to multivariate data analysis, Second edition. Wiley.

Adachi, K., & Trendafilov, N. T. (2018). Some mathematical properties of the matrix decomposition solution in factor analysis. Psychometrika, 83, 407–424.

Adachi, K., & Trendafilov, N. T. (2019). Some inequalities contrasting principal component and factor analyses solutions. Japanese Journal of Statistics and Data Science, 2, 31-47.

Andreas Alfons
Erasmus University

"Cellwise and Rowwise Robust Regression with Compositional Covariates"

We introduce a novel robust procedure to estimate a linear regression model with a real-valued response variable, and compositional and real-valued explanatory variables. The proposed procedure is designed to be robust against cellwise outliers (outliers in individual cells of the data matrix) and rowwise outliers (entire outlying observations). First, cellwise outliers are filtered and then imputed via rowwise robust multiple imputation. Afterwards, regression coefficient estimates are obtained via rowwise robust compositional regression. In simulations, the proposed procedure generally outperforms traditional rowwise-only robust regression estimators as well as more recently proposed cellwise robust regression methods. In an application to bio-environmental data, the proposed procedure leads to conclusions that are more aligned with established scientific knowledge in the field (compared to other regression methods).

Joint work with Nikola Stefelova, Javier Palarea-Albaladejo, Peter Filzmoser and Karel Hron

Chun-houh Chen
Academia Sinica

"Covariate-adjusted Heatmaps for Visualizing Biological Data via Correlation Decomposition"

Heatmap is a popular visualization technique in biology and related fields. In this study, we extend heatmaps within the framework of matrix visualization (MV) by incorporating a covariate adjustment process through the estimation of conditional correlations. MV can explore the embedded information structure of high-dimensional large-scale datasets effectively without dimension reduction. The benefit of the proposed covariate-adjusted heatmap is in the exploration of conditional association structures among the subjects or variables that cannot be done with conventional MV.

For adjustment of a discrete covariate, the conditional correlation is estimated by the within and between analysis. This procedure decomposes a correlation matrix into the within- and between-component matrices. The contribution of the covariate effects can then be assessed through the relative structure of the between-component to the original correlation matrix while the within-component acts as a residual. When a covariate is of continuous nature, the conditional correlation is equivalent to the partial correlation under the assumption of a joint normal distribution.

A test is then employed to identify the variable pairs which possess the most significant differences at varying levels of correlation before and after a covariate adjustment. In addition, a z-score significance map is constructed to visualize these results. A simulation and three biological datasets are employed to illustrate the power and versatility of our proposed method.

GAP is available to readers and is free to non-commercial applications. The installation instructions, the user's manual, and the detailed tutorials can be found at http://gap.stat.sinica.edu.tw/Software/GAP.

Authors: Han-Ming Wu, National Taipei University, New Taipei City 23741, Taiwan
Chun-houh Chen*, Academia Sinica, Taipei 11529, Taiwan

Hao Chen
University of California, Davis

"Change-point Analysis for Modern Data"

After observing snapshots of a network, can we tell if there has been a change in dynamics? After collecting spiking activities of thousands of neurons in the brain, how shall we extract meaningful information from the recording? We introduce a change-point analysis framework utilizing graphs representing the similarity among observations. This approach is non-parametric and can be applied to data when an informative similarity measure can be defined. Analytic approximations to the significance of the test statistics are derived to make the method fast applicable to long sequences. The method is illustrated through the analysis of the Neuropixels data.

Shih-Hsiung Chou and **Phil Turk**
Atrium Health

"CURVE: a Web Application for In-Hospital Resource Forecasting During the COVID-19 Outbreak"

The emergence of COVID-19 has created an urgent threat to public health worldwide. With rapidly evolving demands on healthcare resources, it is imperative that healthcare systems have the ability to access real-time local data to predict, plan, and effectively manage resources. This study aimed to develop an interactive COVID-19 Utilization and Resource Visualization Engine (CURVE) web application as a data visualization tool to inform decision making and guide a large healthcare system's proactive pandemic response. CURVE was designed using R Shiny to display real-time features of healthcare utilization at Atrium Health (Charlotte, NC) with projections based upon local data for the COVID-19 pandemic. The CURVE app embeds three of our models – a Susceptible-Infected-Recovered model that incorporated social distancing and imperfect detection built on April 11 (SIR-D2), a Bayesian SIR model built on April 25 (eSIR), and an ARIMA model built on May 23. We examined these models' performance in terms of root mean square error (RMSE) for data collected from March 26 to June 22, 2020 and a 30-day out-of-sample forecast from the date of model deployment. The results showed that the ARIMA model provided the lowest RMSE in hospital, ICU, and ventilator census for both in-sample data and the 30-day out-of-sample forecast among the models. The CURVE assisted Atrium Health leadership with timely, actionable insights to guide decision-making during the COVID-19 pandemic by predicting utilization of hospital beds, ICU beds, and number of ventilators, and demonstrated its powerful, interactive interface that provides locally relevant, dynamic, timely information to guide health system decision making and pandemic preparedness.

Keywords: COVID-19, R Shiny, pandemic, forecasting, SIR model, ARIMA

Yining Chen
London School of Economics

"Jump or Kink: Super-efficiency in Segmented Linear Regression Break-point Estimation"

We consider the problem of segmented linear regression with the focus on estimating of the location of the break-point(s). Let n be the sample size, we show that the global minimax convergence rate for this problem in terms of the mean absolute error is O(n^{-1/3}). On the other hand, we demonstrate the construction of a super-efficient estimator that achieves the pointwise convergence rate of either O(n^{-1}) or O(n^{-1/2}) for every fixed parameter values, depending on whether the structural change is a jump or a kink. We discuss the implications and the potential remedy. We also illustrate this phenomenon in more complex settings.

Xiaoyue "Zoe" Cheng
University of Nebraska

"Visually Exploring Age-based Population Data over Time"

Demographers commonly use population pyramid plots to visualize the age and gender distribution. However, the static population pyramid can only show the distribution at a given time point. When we add the time domain to age-population data, it becomes a challenge to display the distribution over time. Another interesting question for visualization is whether the unusual population change can be explained by the impact of a sudden event or by some age preference or restrictions. In this talk we will employ dynamic and interactive graphics to visualize age-based population by time. Animated 2-d and interactive 3-d population pyramid plots are created by R using packages ggplot2, animation, shiny, and plotly. The application uses population data from United States and other countries, to track the population patterns and compare population generations, via the proposed graphical tool.

Glen Wright Colopy
Cenduit

"Personalized Inference Protects Patients and Science"

Patients' vital sign data is a golden example of how patient-specific data cries out for patient-specific inference. We will begin by discussing the obvious case, in which a patient's vital sign data are used to summarize their health status and likelihood of adverse events. We will show several scenarios where patient-specific inference identifies deteriorating physiology that would be missed by population-based methods. Next, we show how patient-specific inference isn't just useful to monitor a patient's health, but also the health of clinical trials. In particular, we describe how personalized patient inference can replace rules-based algorithmic monitoring of data quality.

David Dunson
Duke University

"Generalized Bayes for Probabilistic Uncertainty Quantification in Unsupervised Learning"

Loss-based optimization algorithms provide the most commonly used methods for unsupervised learning - focusing on inferring latent structure in data. Canonical examples include k-means and PCA. Bayesian alternatives to k-means focus on model-based clustering via mixture models, while

Bayes alternative to PCA focus on latent factor modeling. Both types of approaches are much less widely used than their loss-based competitors due to complexity of implementations and issues with brittleness and sensitivity to model misspecification. To implement a fully Bayesian analysis, we need a likelihood for everything and such likelihoods are very hard to specify accurately. The generalized Bayes framework using Gibbs posteriors provides a general alternative to likelihood-based Bayesian inferences that can be viewed as a coherent update of Bayesian beliefs. Here we develop general theory and methods showing how G-Bayes can be used to provide probabilistic Bayesian implementations of loss-based unsupervised learning algorithms such as k-means and PCA - providing uncertainty quantification and borrowing the best of both worlds from the loss and Bayes frameworks.

Joint work with Tommy Rigon & Amy Herring (clustering part) and Steven Winter (PCA part)

Dirk Eddelbuettel
University of Illinois at Urbana-Champaign

"Reliable Reproducible Research via Containers from the Rocker Project"

Use of containers as a standard computing tool has seen a rapid rise in popularity and deployment in recent years, both in industry and research. Containers are constructed from a simple declarative description of their content and behavior. They can be deployed anywhere the service (say, Docker, or Singularity, LXC, Podman,...) is provided thanks to a well-specified run-time interface.

The strong focus on standardization makes containers a very attractive proposition with implications for all phases of a computational research project: development, testing, and deployment. Moreover, it fits the needs of preproducible research really well. In this talk, we briefly describe Docker, then introduce the Rocker Project (Boettiger and Eddelbuettel, 2017; Nuest, Eddelbuettel, et al, 2020) which integrates R and Docker, and illustrate some use cases with a particular emphasis on reproducible research (Nuest, Sochat, et al 2020).

Keywords: R, Docker, Rocker, Reproducible Research
References:
Carl Boettiger and Dirk Eddelbuettel. "An Introduction to Rocker: Docker Containers for R", The R, 2017, 9:2, doi:10.32614/RJ-2017-065,

Daniel Nüst and Dirk Eddelbuettel and Dom Bennett and Robrecht Cannoodt and Dav Clark and Gergely Daroczi and Mark Edmondson and Colin Fay and Ellis Hughes and Lars Kjeldgaard and Sean Lopp and Ben Marwick and Heather Nolis and Jacqueline Nolis and Hong Ooi and Karthik Ram and Noam Ross and Lori Shepherd and Péter Sólymos and Tyson Lee Swetnam and Nitesh Turaga and Charlotte Van Petegem and Jason Williams and Craig Willis and Nan Xiao. "The: Packages and Applications for Containerization with R", Pre-print, arXiv:2001.10641 (forthcoming in The R Journal).

Daniel Nüst and Vanessa Sochat and Ben Marwick and Stephen Eglen and Tim Head and Tony Hirst and Benjamin Evans. "Ten Simple Rules for Writing Dockerfiles for Reproducible Data Science." OSF Preprints. 17 April 2020. doi:10.31219/osf.io/fsd7t.

Robert Gramacy
Virginia Tech University

"Replication or Exploration? Sequential Design for Stochastic Simulation Experiments"

We investigate the merits of replication, and provide methods that search for optimal designs (including replicates), in the context of noisy computer simulation experiments. We first show that replication offers the potential to be beneficial from both design and computational perspectives, in the context of Gaussian process surrogate modeling. We then develop a lookahead based sequential design scheme that can determine if a new run should be at an existing input location (i.e., replicate) or at a new one (explore). When paired with a newly developed heteroskedastic Gaussian process model, our dynamic design scheme facilitates learning of signal and noise relationships which can vary throughout the input space. We show that it does so efficiently, on both computational and statistical grounds. In addition to illustrative synthetic examples, we demonstrate performance on two challenging real-data simulation experiments, from inventory management and epidemiology.

Xan Gregg
SAS Institute

"Understanding Smoothers through Interactive Examples"

Smoothers provide a powerful way to cut through the noise and focus on trends in data. However, each smoothing technique brings underlying assumptions and can produce poor results when the assumptions are strained. We will review the utility of smoothers from a data visualization perspective and explore their parameters and their weaknesses. Moving average, spline and loess are covered for the case of a single input variable. A fix for a loess artifact is proposed.

Patrick J.F. Groenen and **Michael Greenacre**
Erasmus University Rotterdam and Universitat Pompeu Fabra

"Interpretable Kernels for Explainable AI"

The use of kernels in for example support vector machines and kernel ridge regression is a powerful tool in estimating nonlinear predictions. What these methods share is that, instead of the original $n \times p$ matrix of predictor variables, a mapping of the rows into a high dimensional feature space is done, a ridge penalty term on the corresponding weights is used, and the solution is obtained through solving a dual problem. One major drawback of the use of kernels is that the interpretation in terms of the original predictor variables is lost. In this paper, we argue that in the case of a wide $n \times p$ matrix of predictor variables (with $p \geq n$), the kernel solution can be re-expressed in terms of a linear combination of the original matrix of predictor variables and a ridge penalty that involves a special metric. Consequently, the solution can be interpreted in the usual manner as a weighted linear combination of the predictor variables. In the case $p < n$, we discuss a least-squares approximation of the kernel matrix that still allows the interpretation in terms of a linear combination.

Dorit Hammerling
Colorado School of Mines

"Contained Chaos: Ensemble Consistency Testing for the Community Earth System Model"

State-of-the-science climate models are valuable tools for understanding past and present climates, and are particularly vital for addressing otherwise intractable questions about future climate. Given the societal relevance, maintaining model confidence is critical. The many complex processes characterizing the Earth System lead to inherently chaotic behavior within these models, meaning that output can be highly sensitive to seemingly minor changes to the hardware or software stack. This sensitivity makes defining "correctness" of model output separately from bit-reproducibility a practical necessity. To this end, we have developed a statistical testing framework that utilizes an ensemble of established simulations paired with principal component analysis to perform hypothesis testing for model correctness. This test is already implemented and provides valuable feedback to model developers and users. We are now exploring the impact of ensemble size on this test, specifically as it relates to systematic errors in estimating the unknown correlation matrix of the ensemble simulations, which is a critical step in the testing process. To estimate this correlation matrix with sufficient accuracy in the current framework, a large ensemble of simulations is required, which is both computationally and time intensive to generate. We will provide an overview of our findings regarding the impact of ensemble size on the test, then present a number of potential solutions for reducing the number of ensemble simulations required to achieve acceptable test results.

Jim Harner[1], **Chris Grant**[2], and **Mark Lilback**[2]
[1]West Virginia University and [2]Rc2ai

"Reproducible Computing and Reporting in a Complex Software Environment"

Statistical projects consisting of code-based text, data analyses, visualizations, and simulations can be "containerized," creating a transparent reproduction of the computational workflow and results. With the advent of container platforms such as Docker and their integration with cloud technologies and standalone operating systems, entire environments can be archived, providing later access to the exact data, methods, platforms, and configurations used to generate the computations and reports. Generally, all processes should be automated by Dockerfiles, makefiles, and shell scripts, along with a version control system to ensure reproducibility.

This talk extends the single container model to multiple containers and Docker applications, which are needed for complex software environments. The base image, 'rcompute,' extends Rocker's 'verse' based on Ubuntu (https://github.com/rocker-org/rocker-versioned2) by adding components required for big data and streaming data computations, including drivers for the PostgreSQL database and Spark, supporting R packages, and various Linux libraries and applications. This base container supports Git version control and connectivity to GitHub; it allows R Markdown documents to be edited by RStudio Server or Vim. By itself, 'rcompute' supports single node execution for R, Python, Spark and database access, which is sufficient for reproducible reports, but not at scale.

The second image, 'rpsql', provides PostgreSQL database services, the most powerful open-source database available. Although 'rpsql' can be run as a standalone container, it is generally used in conjunction with 'rcompute.' As such, a Docker application was constructed based on both 'rcompute' and 'rpsql' using 'docker-compose.' Docker compose specifies dependencies among the images and ports so that the running containers can work seamlessly together and can communicate.

The third component 'rspark' is a Docker application comprising a small standalone Spark cluster, composed of a master and worker images. The principal R package connecting 'rcompute' and

‘rspark’ is ‘sparklyr,’ which provides a ‘dplyr' interface to Spark. Typically, the ‘rcompute’ application, including ‘rpsql’, is run on an edge node, e.g., a laptop, and the ‘rspark’ application is run in the cloud, e.g., AWS. Future work will extend ‘rspark’ to a Kubernetes (k8s) Spark cluster for scaling out.

The images underlying the Docker containers discussed here can be frozen at any time on Docker Hub and thus reproducible computing and reporting is possible even at scale. This talk and the LaTeX version of the abstract, together with links to the repositories for the ‘rcompute’ and ‘rspark’ Docker applications, can be found here: https://github.com/jharner/DSSV2020rspark

Keywords: Reproducible report, Git, Docker, Rocker, R Markdown, PostgreSQL, Spark

Soren Harner and **Jim Harner**
Permaling, LLC and West Virginia University

“Harnessing Big Data and Machine Learning with Arrow Data Frames in R and Python”

Data frames have become an essential tool for data scientists, providing a convenient and intuitive tool chain for manipulating structured data in R and Python. Data frames combine the flexibility of imperative code with the declarative power of SQL. Nonetheless, current implementations have shortcomings related to scalability, interoperability, and semantics. In this talk, we examine how Apache Arrow addresses these challenges by providing a standard for storing, transporting, and operating on tables of data in a columnar memory format. These improvements will usher in new tools that make big data analytics and large-scale machine learning accessible to more people. We look at the layers of Arrow C++ API, and how they abstract datasets, query engines, and data frames, exposed through Python and R bindings. We then examine how Arrow enables interoperability and performance in advanced ML and analytics applications, including Stoic, a big data spreadsheet, and Nvidia RAPIDS, a suite of GPU accelerated data science and ML libraries.

Keywords: Apache Arrow, Data Frame, Machine Learning, GPU Acceleration, SIMD, Columnar Data

Heike Hofmann
Iowa State University

“Visualizing Elections in the U.S.”

In the run-up for the U.S. presidential elections in November 2020, we will be inundated with information about the candidates' every move. Pollsters are ramping up their game, and the news is full of predictions and speculations about the outcome of the Presidential elections. We will be using statistical charts to represent the state of affairs. The U.S. has a rich history of representing election information using graphics. Building on these, we present different ways of visualizing U.S. election results and polls, and provide the means for you to choose and pick your favorites. The visualizations discussed are (mostly) available as part of an R package available on github: https://github.com/heike/electionViz/

Co-authors: Kiegan Rice, Susan Vander Plas

Inge Koch
University of Western Australia

"Principal Components for High-Dimensional and Directional Data"

Principal component analysis (PCA) is a widespread tool for selecting a smaller number of dimensions and key features in multivariate and high-dimensional data. More recently a number of variants of PCA have been developed including sparse PCA for high-dimensional data and robust PCA. In this talk we focus on PCA developments for multivariate and high-dimensional directional random vectors and data which have been transformed to live on the surface of the d-dimensional sphere. These random vectors are also known as special signs. For directional random vectors we review robust covariance related matrices, including the sign and rank covariance matrices, and we present theoretical results of these and relate their relationships to the canonical population covariance matrix.

For random vectors and data from the elliptic distribution we point out relationships between these robust population covariance matrices and their sample counterparts. For non-elliptic data, much less is known at the population level and the sample level about behaviour of these various covariance matrices. We begin with sample versions of the robust covariance matrices, and show the relationships between them and between sample and corresponding population quantities.

We complement these comparisons with calculations based on real data and simulated data ranging from multivariate Gaussian and skew-normal to bimodal and data with high kurtosis and outliers. For such data we study the behaviour of the first few eigenvectors and calculate the closeness of eigenvectors arising from different robust covariance matrices. For simulated data we also calculate their closeness to the eigenvector of the population covariance matrix for a range of dimensions as the sample size increases. Our findings show that kurtosis is a key feature which affects the closeness of the sample eigenvectors to those of the population and we suggest criteria based on the amount of kurtosis which may provide a guide to choosing the `best' sample covariance to use for particular datasets.

Eun-Kyung Lee
Ewha Womans University

"Tree-Structured Models using Projection Pursuit Method and their Explanation"

In this talk, we propose a new tree-structured regression method using a projection pursuit approach. It extends the projection pursuit classification tree to the regression problem. The main advantage of the projection pursuit regression tree is the exploration of the independent variable space in each range of the dependent variable. Also, it keeps all the main properties of the projection pursuit classification tree. To improve the predictability, the projection pursuit regression tree provides several methods to assign values in the final node. With this development, we can easily explore the data space for the piecewise regression and find a better model with good predictability. We also apply explainable artificial intelligence techniques to this regression tree and propose a way to examine the predictions.

Tianxi Li
University of Virginia

"Linear Regression and its Inference on Noisy Network-linked Data"

Linear regression on a set of observations linked by a network has been an essential tool in modeling the relationship between response and covariates with additional network data. Despite its wide range of applications in many areas, such as social sciences and health-related research, the problem has not been well-studied in statistics so far. Previous methods either lack of inference tools or rely on restrictive assumptions on social effects, and usually treat the network structure as precisely observed, which is too good to be true in many problems. We propose a linear regression model with nonparametric social effects. Our model does not assume the relational data or network structure to be accurately observed; thus, our method can be provably robust to a certain level of perturbation of the network structure. We establish a full set of computationally efficient asymptotic inference tools under a general requirement of the perturbation and then study the robustness of our method in the specific setting when the perturbation is from random network models. We discover a phase-transition phenomenon of inference validity concerning the network density when no prior knowledge about the network model is available, while also show the significant improvement achieved by knowing the network model. A by-product of our analysis is a rate-optimal concentration bound about subspace projection that may be of independent interest. We conduct extensive simulation studies to verify our theoretical observations and demonstrate the advantage of our method compared to a few benchmarks under different data-generating models. The method is then applied to adolescent network data to study the gender difference in social activities.

Xinyi Li
SAMSI

"Sparse Learning and Structure Identification for Ultra-High-Dimensional Image-on-Scalar Regression"

This paper considers high-dimensional image-on-scalar regression, where the spatial heterogeneity of covariate effects on imaging responses is investigated via a flexible partially linear spatially varying coefficient model. To tackle the challenges of spatial smoothing over the imaging response's complex domain consisting of regions of interest, we approximate the spatially varying coefficient functions via bivariate spline functions over triangulation. We first study estimation when the active constant coefficients and varying coefficient functions are known in advance. We then further develop a unified approach for simultaneous sparse learning and model structure identification in the presence of ultra-high-dimensional covariates. Our method can identify zero, nonzero constant and spatially varying components correctly and efficiently. The estimators of constant coefficients and varying coefficient functions are consistent and asymptotically normal for constant coefficient estimators. The method is evaluated by Monte Carlo simulation studies and applied to a dataset provided by the Alzheimer's Disease Neuroimaging Initiative.

Sugnet Lubbe[1] and Peter Filzmoser[2]
[1]Department of Statistics and Actuarial Science, Stellenbosch University, South Africa
[2]Institute of Statistics and Mathematical Methods in Economics, Vienna University of Technology, Austria

"Comparison of Zero Replacement Strategies for Compositional Data with Large Numbers of Zeros"

Compositional data analysis needs special attention since the data carry relative information. A multivariate set of samples consisting of D parts, "live" in the $(D-1)$-dimensional simplex. In order to perform standard statistical analyses, the data is typically transformed to the Euclidean

space by a set of logratios. Zero observations are problematic, either as denominator in the ratio, or as numerator in the log function. To overcome this problem, zeros are replaced by some positive value. There is a growing literature on strategies to replace zero values with plausible values that do not alter the structure of the data set.

When analysing high throughput sequencing data, the observed values are in the form of counts of the number of genetic sequences observed in each sample. These count tables are typically sparse with in excess of 90% of the cells containing zeros. Different strategies of zero replacement are compared with respect to the effect on the correlation structure between the parts (variables) and the effect on distances between samples. Many methods are completely unable to deal with such a large proportion of zeros in the data. Some of the best performing strategies for moderate proportions of zeros do not perform very well with large proportions of zeros. However, simple replacement strategies seem to be able to cope with large proportions of zeros while still preserving the data structure in a reasonable way.

Keywords: compositional data analysis, logratio analysis, zero replacement strategies

Javier Luraschi
RStudio

"Training ImageNet using TensorFlow and R"

ImageNet is an image database organized according to the WordNet hierarchy and designed for use in visual object recognition software research. Annual competitions, like ILSVRC, provide researchers a fertile ground to compete and compare algorithms for visual recognition tasks. In 2012 a deep convolution network named AlexNet achieved a 16% error classification rate which was a significant improvement from the original previous years. Today, deep learning has emerged as a new discipline supporting various neural network architectures, including convolutions, and ImageNet still remains an important dataset for image classification tasks. However, it's still challenging to train models against ImageNet since it often requires hundreds of GPUs to train under an hour and distributed computing techniques as well.

This talk will present new advancements in R packages that allow us to train models using ImageNet across multiple GPUs using tools like TensorFlow, Spark, and Keras. The ease of use of R and it's rich ecosystem of packages enables new interesting applications and makes this technology more accessible to the R community. You will learn what you need to train Distributed Deep Learning models from R with help of cloud computing and a few additional R packages. It is our hope that this talk will encourage the R community to train and participate in large scale image recognition challenges, and to be equipped to solve similar large unstructured data problems that many organizations face today.

Pulong Ma
SAMSI

"Multifidelity Computer Model Emulation with High-Dimensional Output: An Application to Storm Surge"

Hurricane-driven storm surge is one of the most deadly and costly natural disasters, making precise quantification of the surge hazard of great importance. Inference of such systems is done through physics-based computer models of the process. Such surge simulators can be implemented with a

wide range of fidelity levels, with computational burdens varying by several orders of magnitude due to the nature of the system. The danger posed by surge makes greater fidelity highly desirable, however such models and their high-volume output tend to come at great computational cost, which can make detailed study of coastal flood hazards prohibitive. These needs make the development of an emulator combining high-dimensional output from multiple complex computer models with different fidelity levels important. We propose a parallel partial autoregressive cokriging model to predict highly-accurate storm surges in a computationally efficient way over a large spatial domain. This emulator has the capability of predicting storm surges as accurately as a high-fidelity computer model given any storm characteristics and allows accurate assessment of the hazards from storm surges over a large spatial domain.

Kelci Miclaus
JMP Life Sciences

"The Role of Visualization in Translational and Clinical Research"

Current debates around p-value thresholds question the application and validity of reporting "statistically significant" findings for replicable research. Genomic, translational and clinical research particularly can be vulnerable to such issues as biometrics teams must balance communicating statistical analysis results of complex biological systems while testing a high dimension of endpoints. While reporting p-values has pitfalls and interpretation has been notoriously variable, statistical significance (such as a p-value threshold) plays an essential role in translational studies using genomic, other biological, or clinical endpoints. For example, without thresholds, we cannot appropriately apply multiple testing correction methodology. In this presentation, we will focus on driving better practices by incorporation of visualization with analyses instead of relying on the communication of a statistical result alone. Appropriate graphical display of the results of a statistical analysis can provide solutions to balancing the statistical signals against clinical/biological meaning as well as drive clearer communication.

Aaron J. Molstad
University of Florida

"Insights and Algorithms for the Multivariate Square-root Lasso"

We study the multivariate square-root lasso, a method for fitting the multivariate response (i.e., multi-task) linear regression model with dependent errors. This estimator minimizes the nuclear norm of the residual matrix plus a convex penalty. Unlike some existing methods for multivariate response linear regression, which require explicit estimates of the error covariance matrix or its inverse, the multivariate square-root lasso criterion implicitly adapts to dependent errors and is convex. To justify the use of this estimator, we establish an error bound which illustrates that like the univariate square-root lasso, the multivariate square-root lasso is pivotal with respect to the unknown error covariance matrix. Based on our theory, we propose a simple tuning approach which requires fitting the model for only a single value of the tuning parameter, e.g., does not require cross-validation. We propose two algorithms to compute the estimator: a prox-linear alternating direction method of multipliers algorithm, and an accelerated first order algorithm which can be applied in certain cases. In both simulation studies and a genomic data application, we show that the multivariate square-root lasso can outperform more computationally intensive methods which estimate both the regression coefficient matrix and error precision matrix.

Katherine E. Moore and **Kenneth S. Berenhaut**

Wake Forest University

"Communities in Data"

Harnessing a social perspective of alignment and conflict, we introduce an informative and broad approach for revealing underlying relational network structure inherent in a collection of perceived distances (between individuals in an arbitrary space). The approach can be valuable for hotspot and outlier detection, clustering, high-dimensional data perception, smoothing, classification, and in instances where near neighbor or density-based approaches are employed. Results are obtained in a straightforward manner without extraneous inputs and the method does not involve a search over an underlying parameter space. As a consequence of this perspective, we introduce a novel concept of cohesion and a resulting notion of communities in data.

John Nardini
SAMSI

"Learning Differential Equation Models for Noisy Biological Data"

We consider the problem of learning the dynamics governing a noisy dataset using sparse regression methods, also known as equation learning. The math biology field presents several current challenges for these methods. Namely, the data collection process can significantly corrupt the observation of data in a heteroscedastic manner, only a small number of time samples may be available for sampling, and the model patterns may drastically change due to inter-sample heterogeneity. We investigate the performance of equation learning methods in these challenging regimes and suggest steps that can be taken during the data collection process to increase their success when current methods fail. We focus on the progression of glioblastoma multiforme as a case-study in this talk, but the methods discussed are relevant to many other areas, including ecology and developmental biology.

Matey Neykov
Carnegie Mellon University

"High Temperature Structure Detection in Ferromagnets"

This talk focuses on structure detection problems in high temperature ferromagnetic (positive interaction only) Ising models. The goal is to distinguish whether the underlying graph is empty, i.e., the model consists of independent Rademacher variables, versus the alternative that the underlying graph contains a subgraph of a certain structure. We give matching upper and lower minimax bounds under which testing this problem is possible/impossible respectively. On the computational front, under a conjecture of the computational hardness of sparse principal component analysis, we prove that, unless the signal is strong enough, there are no polynomial time tests which are capable of testing this problem.

Jason Poulos
SAMSI

"Retrospective Causal Prediction via Elapsed-Time and Propensity-Weighted Matrix Completion, with an Evaluation of the Effect of European Integration on Labour Market Outcomes"

We propose a method of retrospective counterfactual prediction in panel data settings where there exists units exposed to treatment after an initial time period (later-treated), units always exposed to treatment (always-treated), but none that are never exposed to treatment (never-treated). We employ a matrix completion estimator to predict counterfactual outcomes of the later-treated during the pre-treatment period using information from the observed outcomes of both later-treated and always-treated. A possible complication of this approach is that during the post-treatment period, we observe outcomes for always-treated and never-treated in the same calendar time but the elapsed time since the treatment implementation is different for the two groups. We address this problem by weighting the objective function by propensity scores scaled by elapsed time since treatment. Our methodology is motivated by studying the effect of the visa policy of the Schengen Area and the elimination of work permits within the European Union on the labour market outcomes of treated border regions.

Authors: Andrea Albanese, Luxembourg Institute of Socio-Economic Research (LISER) ; Fan Li, Duke University; Andrea Mercatanti, Bank of Italy; Jason Poulos, Duke University and SAMSI

Brian Lee Yung Rowe
Pez.AI

"Achieving Practical Reproducibility with Transparency and Accessibility"

With the rise of data science and machine learning, code has become methodology. Reproducing and verifying results thus requires both running and reading code. This talk argues that transparent and accessible process automation is the key to effective reproducible science. Transparent code is well organized and well documented. Accessible code is easy to run and has few esoteric dependencies. By using a few standard tools and adhering to some basic software development techniques, most data science workflows can achieve practical reproducibility with minimal effort.

Keywords: reproducible science, methodology, automation, software development

Cynthia Rudin
Duke University

"Seeing into Data and Models"

When you look carefully at data and models, you never know what you might find. In this talk, I will present several stories that provide interesting lessons for modern design of models, the ways we look into them, and sacrifices we might need to make when we try to look at high dimensional data. In particular:

Point 1: Interpretable Models don't need to be simple, particularly if you can co-design interpretable models with their visualizations. I will give two examples of machine learning models that are complex, but that are also interpretable because the important parts of their reasoning processes can be visualized: These examples are (1) an entry into the FICO Explainable Machine Learning Challenge, and (2) a deep neural network (ProtoPNet) that uses cased-based reasoning to explains how it classifies an image.

Point 2: When visualizing complex high dimensional data in lower dimensional spaces, there is a tradeoff between local and global structure preservation. Dimension reduction techniques, such as t-SNE, UMAP, and LargeVis, can be used as visualization tools for high-dimensional data. These

methods aim to preserve as much structure as possible when transforming data from high-dimensions to low dimensions, but there is a fundamental tradeoff between the preservation of local and global structure. I will discuss some insights into how this tradeoff works, and present a new approach that has advantages in this tradeoff.
Point 3: We know that data reflects society, but after you look at it and realize what is there, it's important to keep discussion about its future uses open. We know that there are many gender/racial biases within essentially all of our datasets. The English language is gender/racially biased, so using any large text corpus as a dataset is biased. Almost all datasets containing images of the natural world are in some way biased, whether it is towards western societies or towards certain racial or cultural groups. I will give an example of where an algorithm we created for an art project - with an interactive interface - allowed users to reveal a bias in a standard software tool in computer vision, leading to a recent charged discussion in the AI community. As we design more interactive data exploration tools and see more of these biases, I would like to encourage our community to leave doors open for discussion on how to solve these problems together.
Papers I will discuss include:
This Looks Like That: Deep Learning for Interpretable Image Recognition. NeurIPS, 2019.
Chaofan Chen, Oscar Li, Alina Barnett, Jonathan Su, Cynthia Rudin
https://arxiv.org/abs/1806.10574

An Interpretable Model with Globally Consistent Explanations for Credit Risk. NIPS 2018 Workshop on Challenges and Opportunities for AI in Financial Services: the Impact of Fairness, Explainability, Accuracy, and Privacy, 2018.
Chaofan Chen, Kangcheng Lin, Cynthia Rudin, Yaron Shaposhnik, Sijia Wang, and Tong Wang
https://arxiv.org/abs/1811.12615

PULSE: Self-Supervised Photo Upsampling via Latent Space Exploration of Generative Models. CVPR, 2020. Sachit Menon, Alexandru Damian, Shijia Hu, Nikhil Ravi, Cynthia Rudin
https://arxiv.org/abs/2003.03808

Deborshee Sen
SAMSI

"Bayesian Neural Networks and Dimensionality Reduction"

In conducting non-linear dimensionality reduction and feature learning, it is common to suppose that the data lie near a lower-dimensional manifold. One class of model-based approaches for such problems includes latent variables in an unknown non-linear regression function; this includes Gaussian process latent variable models and variational auto-encoders (VAEs) as special cases. VAEs are artificial neural networks (ANNs) that employ approximations to make the computation tractable; however, current implementations lack adequate uncertainty quantification in estimating the unknown density and the lower-dimensional subspace, and can be unstable and lack reproducibility in practice. We attempt to solve this problem by deploying Markov chain Monte Carlo sampling algorithms for Bayesian inference in ANN models with latent variables. We address issues of identifiability by imposing constraints on the ANN parameters as well as by using anchor points. This is demonstrated on simulated and real data examples.

Susan VanderPlas
University of Nebraska-Lincoln

"Perception and Visual Communication in a Global Pandemic"

In the early months of 2020, as the novel coronavirus spread around the globe, we all turned to graphics and data visualizations in order to make sense of the unfolding catastrophe. What were the latest case counts? How many people were in ICUs? What were epidemiological models predicting the case load would be in a month?

In response, journalists, academics, and amateurs generated an astonishing amount of visualizations. Novel graphical forms and approaches to the data appeared amid the more traditional maps and case count charts. This creativity was in part a result of the challenges of depicting data which was exponentially increasing, from countries and states with vast differences in population size and different dates of initial infection, resulting from an assortment of testing strategies and public health interventions. Some approaches distilled complex information down into very simple (but imprecise) representations, while others provided incredibly detailed data that obscured the messy real-world situation with precise numbers and ratios. These graphics and trade offs highlight how little we know about graphical perception and visual numeracy, and how important it is to understand the impact of graphical design choices when communicating scientific information in a visual domain. Using COVID-related graphics, this presentation will examine what we know, what we think we know, and what we still need to explore to create useful, accurate, and informative statistical graphics.

Giuseppe Vinci
University of Notre Dame

Graph Quilting: Graphical Model Selection from Partially Observed Covariances

Estimating a conditional dependence graph is a seemingly impossible task when several pairs of variables have no joint observation. Recovering the edges of the graph in such settings requires one to infer conditional dependencies between variables with no empirical observation of their covariation. This largely unexplored statistical problem arises in several important situations, such as in neuroimaging where, because of technology limitations, the joint activities of several pairs of neurons remain unobserved. We call this statistical challenge the "Graph Quilting problem". In the Gaussian graphical model, the unavailability of parts of the covariance matrix translates into the nonidentifiability of the precision matrix, which specifies the graph. However, we demonstrate that, under mild conditions, it is possible to correctly identify not only the edges connecting the observed pairs of nodes, but also a superset of those connecting the variables that are never observed jointly. We perform the latter task by devising a novel technique that we call the "Recursive-Complement" algorithm. We propose an l1-regularized graph estimator based on partially observed sample covariances, and establish its rates of convergence in high-dimensions. We illustrate the methodology using synthetic data, as well as data obtained from in vivo calcium imaging of ten thousand neurons in mouse visual cortex. Finally, we discuss applications to genomics and other fields.

Tengyao Wang
University College, London

"High-Dimensional, Multiscale Online Changepoint Detection"

We introduce a new method for high-dimensional, online changepoint detection in settings where a p-variate Gaussian data stream may undergo a change in mean. The procedure works by performing likelihood ratio tests against simple alternatives of different scales in each coordinate,

and then aggregating test statistics across scales and coordinates. The algorithm is online in the sense that its worst-case computational complexity per new observation, namely O(p^2 log(ep)), is independent of the number of previous observations; in practice, it may even be significantly faster than this. We prove that the patience, or average run length under the null, of our procedure is at least at the desired nominal level, and provide guarantees on its response delay under the alternative that depend on the sparsity of the vector of mean change. Simulations confirm the practical effectiveness of our proposal.

Wenjia Wang
SAMSI

"Uncertainty Quantification for Bayesian Optimization"

Bayesian optimization is a class of global optimization techniques. It regards the underlying objective function as a realization of a Gaussian process. Although the outputs of Bayesian optimization are random according to the Gaussian process assumption, quantification of this uncertainty is rarely studied in the literature. In this talk, I will talk about our recent work to assess the output uncertainty of Bayesian optimization algorithms, in terms of constructing confidence regions of the maximum point or value of the objective function. These regions can be computed efficiently, and their confidence levels are guaranteed by newly developed uniform error bounds for sequential Gaussian process regression. Our theory provides a unified uncertainty quantification framework for all existing sequential sampling policies and stopping criteria.

Adalbert F.X. Wilhelm
Jacobs University

"Visual Story Telling of Covid-19: A Case Study"

Covid-19 has created a world-wide interest in understanding the dynamics of a pandemic. Hence, numerous media outlets as well as researchers have produced comprehensive data visualizations to illustrate the relevant trends and figures. In this presentation, we will look at an elective choice of Covid-19 data visualizations to evaluate and discuss currently established visualization tools in their capacity to provide a communication channel both within the data science team and also between data analyst, domain experts and a more general interested audience. While there is no fixed catalogue of evaluation criteria for data visualizations we will try to provide an overview on the different core aspects of visualization evaluation and their competing principles.

Guohui Wu
SAS Institute

"Location Matters: Estimating Spatial Regression Models with Large Spatial Weight Matrices Using SAS® Econometrics"

Spatial regression models have been widely used to address spatial dependence in data, making sure that models correctly approximate the reality that they are modeling. By taking neighboring effects into account, spatial regression models enable us to understand if and how a certain event in one location is influenced by events in nearby locations. Despite their widespread use, these models can be time-consuming to estimate because of the challenges posed by the ever-increasing size of the spatial data sets they must account for. In this talk, I introduce a variety of spatial regression models

and a dedicated SAS® Econometrics procedure for spatial regression analysis. Examples are provided to demonstrate the usage and big data capability of our SAS Econometrics tool.

Jason Xu
Duke University

"A Proximal Distance Algorithm for Likelihood-Based Sparse Covariance Estimation"

We consider the task of estimating a covariance matrix under a patternless sparsity assumption. In contrast to existing approaches based on thresholding or shrinkage penalties, we propose a likelihood-based method that regularizes the distance from the covariance estimate to a symmetric sparsity set. This formulation avoids unwanted shrinkage induced by more common norm penalties, and we show that the resulting non-convex problem can be solved via a sequence of smooth, unconstrained problems. We develop a proximal distance algorithm from the principle of majorization-minimization; the resulting algorithm executes rapidly, gracefully handles settings where the number of parameters exceeds the number of cases, yields a positive definite solution, and enjoys desirable convergence properties. We demonstrate the merits of our approach via simulation studies and on cell signaling and international migration data.

Ming Yuan
Columbia University

"Information Based Complexity of High Dimensional Sparse Functions"

We investigate the optimal sample complexity of recovering a general high dimensional smooth and sparse function based on point queries. Our result provides a precise characterization of the potential loss, or lack thereof, in information when restricting to point queries as opposed to the more general linear queries, as well as the benefit of randomization and adaption. In addition, we also developed a general framework for function approximation to mitigate the curse of dimensionality that can also be easily adapted to incorporate further structure such as lower order interactions, leading to sample complexities better than those obtained earlier in the literature.

Ruda Zhang
SAMSI

"Normal-bundle Bootstrap"

Probabilistic models of data sets often exhibit salient geometric structure. Such a phenomenon is summed up in the manifold distribution hypothesis, and can be exploited in probabilistic learning. Here we present normal-bundle bootstrap, a method that generates new data which preserve the geometric structure of a given data set. Inspired by algorithms for manifold learning and concepts in differential geometry, our method decomposes the underlying probability measure into a marginalized measure on a learned data manifold and conditional measures on the normal spaces. The algorithm estimates the data manifold as a density ridge, and constructs new data by bootstrapping projection vectors and adding them to the ridge. We apply our method to the inference of density ridge and related statistics, and data augmentation to reduce overfitting.

Mikhail Zhelonkin
Erasmus University

“Probabilistic Forecasting of Binary Outcomes in the Presence of Outliers”

The problem of forecasting of binary outcomes is of prominent importance in various fields including Economics, Management, Finance, Marketing and Medicine, to mention a few. For instance, it can be a default of a company, a click on the online advertisement, or an occurrence of a disease. The traditional approach is to use classification methods, which can be seen as the point forecasts. However, from the perspective of a decision maker, it is valuable to have a probability forecast. The traditional benchmark parametric models, e.g., logistic regression, are unstable in the presence of outliers and data contamination. The alternative machine learning methods are often biased and require recalibration what makes them hardly interpretable. In this paper, we show that logistic regression estimated by robust methods is a viable alternative. Using the influence functions approach we show that the robustly fitted logistic regression delivers well-calibrated forecasts and that the loss of efficiency is negligible.

Mu Zhu
University of Waterloo

“Some Statistical Applications of Generative Neural Networks”

We present some examples of how the field of statistics can benefit from the Deep Learning movement. First, using generative neural networks (GNNs), we are now able to produce quasi Monte Carlo samples from "almost any" copula model. For example, we can do this even for mixtures of copulas with singular components. Second, using GNNs, we can now model and, most importantly, forecast multivariate time series without having to restrict ourselves to using only a few parametric copula families to describe the underlying multivariate dependence. We have empirical evidence that a better dependence model does indeed translate into better forecasts. (This is joint work with Marius Hofert and Avinash Prasad.