

## **Manish S. Dalwani UC Denver Health Sciences**

“Brain default mode network identification using Independent Component Analysis.”

Univariate-based general linear model has been the most popular method of analyzing functional MRI (fMRI) data. However, multivariate techniques such as principal component analysis (PCA) and independent component analysis (ICA) have gained importance in the last decade and provided at times superior ways of analyzing fMRI data especially functional network connectivity. ICA is a data-driven multivariate technique that is based on assumptions of non-gaussianity and independence. Those assumptions work well with the modality of fMRI as the signals in fMRI are often super Gaussian (e.g. Laplace or logistic). Methods: We used ICA on a busy functional MRI decision-making task where we first determined and then examined the default mode network (DMN) in brain between patients (14-18 year old females with conduct and substance problems) and controls (females of similar ages). We evaluated whether ICA could identify the DMN components within groups and then compared the (identified) DMN between groups to see if there were patient-control differences. Results: Using ICA on the busy fMRI decision-making task, we were able to identify the DMN in both patients and controls and both groups showed differential DMN activity. Conclusions: ICA is superior to both GLM and PCA in deconvolving the DMN network in brain specifically in a rapid-event functional MRI design.

Co-authors: Susan Mikulich-Gilbertson, Joseph Sakai

---

---

## **D.J. Bays, Tyler Bonnett, and Paige Yankey – Marshall University, West Virginia**

“Why not Substitute? A Simulation Study of Left Censored Data”

It is common for environmental scientists to report collected data as "non-detect," which means a value is known only to be below a certain detection limit of the measuring device. In mainstream statistics, this type of data is known as left-censored data. In this talk, we use a variety of substitution methods, along with maximum likelihood estimation (MLE) and the Kaplan-Meier method to estimate summary statistics (primarily the mean) of left-censored data, and rank their effectiveness for the one detection limit scenario and the multiple detection limit scenario.

---

---

## **Susan Mikulich, University of Colorado Denver – Health Sciences**

### **“Testing Associations between Stochastic Parameters from Multivariate Generalized Linear Mixed Models”**

This talk will review methods for testing associations between stochastic parameters discussed in Mikulich-Gilbertson et al. (2015). Different types of outcomes (e.g. binary, count, continuous) can be simultaneously modeled with multivariate generalized linear mixed models by assuming: (1) same or different link functions, (2) same or different conditional distributions, and (3) conditional independence given random subject effects. Others have used this approach for determining simple associations between subject-specific parameters (e.g. correlations between slopes) and Mikulich-Gilbertson et al. (2015) demonstrate how more complex associations (e.g. partial regression coefficients between slopes adjusting for intercepts, time lags of maximum correlation) can be estimated. The paper reparameterizes the model to directly estimate coefficients and compare standard errors based on the inverse of the Hessian matrix with more usual standard errors approximated by the delta method. A mathematical proof demonstrates their equivalence when the gradient vector approaches zero. Via reparameterization, the significance of coefficients can be evaluated with likelihood ratio tests and Mikulich-Gilbertson et al. (2015) compare this approach with more usual Wald-type t tests and Fisher's z transformations. Simulations indicate that the delta method and inverse Hessian standard errors are nearly equivalent and consistently overestimate the true standard error. The paper concludes that only the likelihood ratio test based on the reparameterized model has an acceptable Type I error rate and is therefore recommended for testing associations between stochastic parameters.

---

---

## **Alec McQuilkin University of Colorado Denver**

### **“Estimating the Basic Reproduction Number of the Seasonal Flu”**

Each year, millions of individuals across the globe become infected with the seasonal flu. An important component of effective public health intervention is to predict the severity of current and future strains in order to best reduce the negative effects of the virus. The basic reproduction number of a disease represents the number of people that an infected individual can be expected to infect in a completely susceptible population. Estimates of this number can provide information on the potential impact of a given flu strain. Here, I use a Susceptible-Infected-Recovered model to estimate the basic reproduction number of the 2009 outbreak of the seasonal flu for the city of Denver, the state of Colorado, and the United States as a whole based on data from Google Flu Trends. Results suggest least squares estimates of 1.42, 1.43, and 1.51 (maximum likelihood estimates of 1.43, 1.47, and 1.56) for Denver, Colorado, and the United States respectively.

---

---

---

---

**Harrison Fell and Peter Maniloff**

Colorado School of Mines

“Negative leakage in subnational climate policy”

Leakage, or the relocation of polluting activities from regions where they are regulated to regions where they are not, is a common prediction before passage of environmental regulations. However, ex poste statistical analysis rarely shows evidence of leakage. In 2009, ten northeastern states enacted a cap-and-trade system for greenhouse gas emissions from electricity generation. This paper shows that the cap-and-trade system has lead to substantial leakage of electricity generation from capped states to their uncapped neighbors. However, since generation in the uncapped states is less emissions-intensive, this has lead to an overall decrease in emissions. We discuss why we find evidence of leakage where prior researchers have not.

---

---

**Stanley Xu, PhD**

Head of Biostatistics/Investigator

Institute for Health Research

“Signal detection of adverse events with imperfect confirmation rates in vaccine safety studies using self-controlled case series design”

The Vaccine Safety Datalink project captures electronic health record data including vaccinations and medically-attended adverse events on 8.8 million enrollees annually from participating managed care organizations in the US. While the automated vaccination data are generally of high quality, a presumptive adverse event based on diagnosis codes in automated health care data may not be true (misclassification). Consequently, analyses using automated health care data can generate false positive results, where an association between the vaccine and outcome is incorrectly identified, as well as false negative findings, where a true association or signal is missed. We developed novel conditional Poisson regression models and fixed effects models that accommodate misclassification of adverse event outcome for self-controlled case series design. We conducted simulation studies to evaluate their performance in signal detection in vaccine safety hypotheses generating (screening) studies. We also re-analyzed four previously-identified signals in a recent vaccine safety study using the newly proposed models. Our simulation studies demonstrated that i) outcome misclassification resulted in both false positive and false negative signals in screening studies; ii) the newly proposed models reduced both the rates of false positive and false negative signals. In re-analyses of four previously-identified signals using the novel statistical models, the incidence rate ratio estimates and statistical significances were similar to those using conventional models and including only

medical record review confirmed cases.

---

---

**Kristin Brom – Colorado State University**

“Dynamic occupancy models for explicit colonization processes”

Multi-season occupancy models are used in ecology to map a species' distribution and its changes over time. The models connect a distribution and its changes to environmental covariates to infer on the patterns we see. However, few versions of the model relate the changes to the occupancies of neighboring sites or patches. I will briefly describe what an occupancy model is and a Bayesian version of the model. Then I will describe the dynamic components of the model and how its diffusion-based equations cause the distribution to change over time. The model is applied to the common myna, a bird that is invasive in South Africa and whose range continues to expand in recent years.

---

---

**Kayla Williamson – University of Wyoming**

“Visualizing bias in Microbiome data”

Using data from the Virginia Commonwealth University's Human Microbiome Project, I consider visualizing data that were collected to assess bias in measuring proportions of different bacteria from samples. A positive controls experiment was conducted to study the bias that accumulates from DNA extraction, PCR amplification and sequencing and taxonomic classification. While real microbiome data has very high dimension, on the order of 200-300 taxa, in the experiment, there were 7-50 taxa observed. Both data are highly sparse.

---

---

**Ekaterina Smirnova – University of Wyoming**

“Multiway data integration: pregnant women vaginal microbiome case study”

Recent advances in various “omics” technologies allow for comprehensive examination of microbial communities, together with other data such as those obtained on lipids and cytokines, along with usual clinical (covariate) data, all collected on the same individual. A characteristic of these studies is that the data are often quite sparse but collected on a large number of variables.

Eigendecomposition-based methods provide tools for not only representing data in lower dimensions, but also enable systematic integration and comparison of multi-omic data sets. In particular, using data from a vaginal microbiome project, we address multiple co-inertia analysis methods for integrating three types of ``omics" data, namely, microbial data obtained from 16S sequencing, data on some cytokines and some lipids collected on the same women.

---

---

### **S. Huzurbazar - University of Wyoming**

“Analysis of geochronological data when ages are measured with error”

Data collected to study the formation of oceanic crust comes with interesting statistical problems. The geochronological data consist of ages of specific minerals, namely zircons, where the zircons are embedded in samples obtained from cores of the ocean crust. The ages are measured with error, and the processing of the samples yields fairly good estimates of the measurement errors. I model the age distribution of the zircons, and hence the samples using data collected by my collaborators, B. John and R. Pettiette in Geology and Geophysics at the Univ of Wyoming. While traditional treatment of similar data models it from the point of view of overdispersion, the specifics of these data lead us to cluster the samples when the ages are measured with error.

---

---

### **Kraig Thomas - University of Colorado Denver**

“Incorporating Relatedness in Gene Based Case-Only Analysis of Mendelian Traits”

Genome sequencing, a technology that provides complete genetic information within a focused area or genome-wide, can identify mutations related to simple (i.e. Mendelian) or complex traits. Statistical methods exist for detecting association between a gene-region and a disease, including a case-only test that uses only diseased individuals without the need for a matched set of unaffected individuals. In this project, we extend the case-only method to incorporate related subjects. This is motivated by the observation that rare Mendelian diseases often appear and are thus studied in families, but the existing case-only method used to study rare mutations within a gene region uses only unrelated subjects. The case-only test statistic is the number of subjects who have a mutation within a certain gene. As we compare the observed number of subjects to the expected number of subjects, the test statistic has a binomial distribution. The expected number of subjects having a mutation within a certain gene is estimated from the large number of genes across the genome (~20000) most of which we expect not to be associated with disease status. To incorporate the relatedness framework to the case-only model, we consider related pairs of subjects, and calculate the total number of pairs who share a mutation within a gene. We consider the effective number of mutations of a related pair: for instance, a pair of siblings shares approximately half of their genetic information. A pair of siblings contributes less unique genetic information than a pair of unrelated subjects, and thus, their effective number of mutations is less than two (i.e. the number of mutations contributed by a pair of unrelated subjects). To extend the

case only analysis we replace the number of subjects with one mutation with the effective number of pairs of related subjects who share a mutation, yielding a new test statistic with a binomial distribution. Future work includes theoretical and empirical verification of the combined model.

---

---

**Zachary D. Weller , Jennifer A. Hoeting**

“Testing Second Order Properties Of Spatial Data”

One of the most important aspects of modeling spatial data is appropriately specifying the second order properties of the random field. A practitioner working with spatial data is presented a number of choices regarding the structure of the dependence between observations. If the data do not appear to be dependent over space, spatial models may be abandoned. If the data do exhibit spatial dependence, decisions and assumptions often need to be made regarding isotropy and stationarity of the covariance function of the generating process. Misspecification of the second order properties could lead to misleading inferences, such as inaccurate predictions and parameter estimates. In a fashion similar to checking assumptions for simple linear regression by looking at residual plots, a researcher may use a set of informal graphical diagnostics, such as directional and regional sample variograms, to decide on an appropriate covariance model. These graphical techniques can be difficult to assess, open to subjective interpretations, and misleading. Objective hypothesis tests of the various assumptions may be more desirable. To this end, a number of tests of spatial dependence structure have been developed in both the spatial and spectral representations of random fields; however, most of the tests rely on (potentially restrictive) assumptions themselves and are not readily available in software. We provide an overview of common possible assumptions on the second order properties of random fields and re-view the various hypothesis test procedures available to check those assumptions. We also provide a summary of key test properties and gives insights on important considerations in choosing a test.

---

---

**Seth E. Spielman**

Asst. Professor of Geography

University of Colorado

[www.sethspielman.org](http://www.sethspielman.org) | [seth.spielman@colorado.edu](mailto:seth.spielman@colorado.edu)

“Improving the Quality of American Community Survey Estimates.”

This talk describes a method to improve the efficiency of small area demographic and economic estimates from the American Community Survey (ACS). Small area estimates from the ACS are terribly imprecise however it is possible to improve the precision of estimates via spatial (or temporal) aggregation. However, geographic aggregation of tracts can lead to significant

information loss and a dilution of spatial patterns. The method presented here is a novel spatial optimization algorithm that improves small area estimates by intelligently grouping small areas while minimizing information loss and preserving spatial patterns. The algorithm allows the post-processing of public use data to a user specified level of precision.