William Cipolli
16 N. Main St. Apt 6
Earlville, NY 13332
+1 (203) 848 5643
will@cipolli.com
www.cipolli.com

Research Statement

The ability to create large-scale, data-intensive applications is easier now than it ever has been, due to the proliferation of open-source programs, cloud-computing, affordable sensors and the ability to capture the unprecedented amount of data we, as humans, create every day.

Projects like Cisco's Internet of Things and IBM's Watson, display the power of high-dimensional data analysis power by leveraging large amounts of data from a variety of sources from millions of users. For example Cisco envisions fully connected lifestyles where our homes, cars, phones, and even our foods are connected to the web, constantly delivering data and IBM uses Watson to help summarize and uncover new knowledge in commercial, medical, and scientific fields.

While pursuing my PhD, I developed and implemented Bayesian non-parametric approaches to multiple testing, density estimation and a supervised learning classification technique. These theoretical contributions have been remarkably successful in applications to real data, as detailed in the Research Projects section below.

From my previous experiences and current interests, I see myself as uniquely qualified for doing research in Statistics and Data Science. I have a particular interest in doing interdisciplinary research with colleagues from Computer Science, Biology, Political Science, Economics, Business and other disciplines as these subject areas bring an important context to the complex theory.

My publications as a doctoral student at the University of South Carolina consistently included much computation and application to real data with practicality in mind. Particularly, the Bayesian Multiple Testing paper also included a Java application – scientists of all backgrounds are able to utilize the methodology without the steep learning curve of learning a new programming language.

My research stems from a fascination with the world; as our capability to capture data improves so must our methods to create meaning through data. In this spirit, I hope to advance my research, collaborate with colleagues and incite a similar level of curiosity in students as I move ahead in my career.

Authored Research

The Score Test for a Zero-inflated Bivariate Poisson Distribution (To be started Summer 2017)

Prediction and Feature Selection for High Dimensional Data through Smoothed Polya Trees(To be started Summer 2017)

Supervised Learning for High Dimensional Data through Smoothed Polya Trees via Givens Rotations (To be started Fall 2017)

Computationally Tractable Approximate and Smoothed Polya Trees for Accelerated Failure Time Models with Heteroskedastic Error (To be started Fall 2017)

Supervised Learning for High Dimensional Data through Smoothed Polya Trees (Submitted to Statistical Methods in Medical Research)

Many classification approaches make distributional assumptions about the data, most infamously the Gaussian distribution. We can create a more flexible approach by generalizing the Gaussian assumption by assuming the nonparametric multivariate Polya tree prior. The flexibility gained from relaxing the distributional assumptions from our analysis can prove paramount when even minor deviations from the distributional assumptions occur; the ability of the Polya tree approach to pick out even slight deviations could significantly improve classification over a model that is assumption-heavy. The outcomes obtained by this model can be compared to other methods including decision trees, k-nearest neighbor, naive Bayes, artificial neural networks, support vector machines, etc.

On The Distribution of Monochromatic Complete Subgraphs and Arithmetic Progressions (In progress)

We consider the distribution of the number of monochromatic complete subgraphs over edgewise 2-colorings of complete graphs as a type of statistical Ramsey theory. We also consider the analogous for monochromatic arithmetic progressions. We present convincing evidence that both distributions are very well-approximated by the family of Delaporte distributions.

The Score Test for a Bivariate Pareto Distribution (In progress)

Two score tests are proposed: one for testing independence based on Sankaran and Nair's bivariate Pareto distribution and one for testing whether Sankaran and Nair's parameterization reduces to the more popular bivariate Pareto distribution introduced by Lindley–Singpurwalla. The marginal distributions of both bivariate parameterizations are univariate Pareto II distributions, and the parameters of the bivariate distribution are estimated using numerical methods. The simulation studies show that both score tests maintain a significance level close to the nominal size. To check the efficiency of the derived score tests, the estimated significance level and power of the likelihood ratio and Wald tests are also compared. One real world data set is used to illustrate the application of both score tests.

Score Test for Independence Using the Normal Copula and Poisson Marginals (In progress)

An increasingly popular approach to model the dependence between random variables centers on the use of copula functions. A variety of bivariate copula families are explored, in particular the Guassian copula. Furthermore, a score test for testing independence of response variables is proposed for the specific case where the marginal distributions are known to be Poisson. The simulation study shows the test keeps the significance level close to the nominal one. Similarly, the estimated significance level and power of the likelihood ratio and Wald tests are also compared to show our test is numerically stable. A real world data set is used to demonstrate the application of our test.

The Score Test for a Bivariate Poisson Distribution (Submitted to Computational Statistics)

A score test for testing independence in Lakshminarayana's bivariate Poisson distribution is proposed. The marginal distributions of the bivariate model are the univariate Poisson distributions, and the parameters of the bivariate distribution are estimated using maximum likelihood methods. The simulation study shows that the score test maintains a significance level close to the nominal one. To check the efficiency of the derived score test, the estimated significance levels and powers of the Likelihood Ratio and Wald tests are also compared. A relevant data set is used to demonstrate the application of the bivariate Poisson model and score test with success.

The paradox of Group Mind "People in a group" Have More Mind than "a Group of People"(Accepted at Accepted at the Journal of Experimental Psychology: General)

Three studies examine how subtle shifts in framing can alter the mind perception of groups. Study 1 finds that people generally perceive groups to have less mind than individuals. However, Study 2 demonstrates that changing the framing of a group from "a group of people" to "people in a group," substantially increases mind perception—leading to comparable levels of mind between groups and individuals. Study 3 reveals that this change in framing influences people's sympathy for groups, an effect mediated by mind perception. We conclude that minor linguistic shifts can have big effects on how groups are perceived—with implications for mind perception and sympathy for mass suffering.

In Search of a Five-Star: The Centrality of Body Discourses in the Scouting of High School Football Athletes(Submitted to Sociology of Sport Journal)

This project explores the question of how scouting reports represent the body, skills, and physical characteristics of the subject. The hypothesis is that body discourses focus predominantly on physical characteristics of the athlete In other words, we predict that representations related to size and strength are more important than more football-specific characteristics. It is also hypothesized that scouting 'experts,' here conceptualized as analysts working for each major media outlet, are rather arbitrary with their subjected player analyses tend to be more important than more objective measurements such as listed weight, height, and calculated body mass index (BMI). To explore these questions, we have attempted to identify the most important predictors of high rated athletes – typically referred to in this context as "five-star" prospects.

Bifactor Models of the Strengths and Difficulties Questionnaire in a Large U.S. Community Sample (Presented at the annual meeting of the American Psychological Association - Denver)

Analysis of a data obtained in a CDC funded study under Dr. Kate Flory. This study analyzed 25-item screenings of children and adolescents ages 3-16. The best-fitting model was a new bifactor model where all factors are allowed to correlate which is an improvement on the bifactor model in existing research literature.

Computationally Tractable Approximate and Smoothed Polya Trees (Published in Statistics and Computing)

The second chapter after the introduction in my dissertation introduces a discrete approximation to the Polya tree prior that enjoys surprisingly simple and efficient conjugate updating. This approximation is illustrated via simulation and in two applied contexts: the implementation of a nonparametric meta-analysis involving studies on the relationship between alcohol consumption and breast cancer and random intercept Poisson regression for Ache armadillo hunting times. The discrete approximation Dirac measures are then replaced with Gaussian densities to provide a smoothed mixture of Polya trees that can be used in standard contexts; the smoothed approximation is illustrated on density estimation of the eruption times of the Old Faithful geyser.

Bayesian Nonparametric Multiple Testing (Published in Computational Statistics & Data Analysis)

Multiple testing, or multiplicity, problems often require testing several means with the assumption that we will reject infrequently, as motivated by the need to analyze DNA microarray data. The goal is to keep the combined rate of false discoveries and non-discoveries as small as possible. We propose a discrete approximation to a Polya tree prior that enjoys fast, conjugate updating, centered at the usual Gaussian distribution, thus generalizing Scott and Berger (2006) to a nonparametric setting. This new technique and the advantages of our approach are demonstrated using extensive simulation and data analysis accompanied by a Java web application. The numerical studies demonstrate that our new procedure shows promising FDR and estimation of key values in the mixture model with very reasonable computational speed.

A Comparison of Three Diagnostic Tests for Diagnosis of Carpal Tunnel Syndrome Using Latent Class Analysis (Published in The Journal of Bone and Surgery)

The current reference standard for carpal tunnel syndrome is under debate. Recent studies have demonstrated similar diagnostic accuracy between ultrasound and nerve conduction studies. Given the lack of a universally accepted reference standard for carpal tunnel syndrome, latent class analysis is an established method to determine the true diagnostic accuracy of these tests. The purpose of this study is to determine sensitivity and specificity of ultrasound (US), nerve conduction studies (NCS), and CTS-6 for diagnosis of carpal tunnel syndrome (CTS) using latent class analysis.

We used latent class models to estimate the sensitivity and specificity of these three diagnostic tests in the absence of a gold-standard. All models were fitted in WinBUGS, including identifiable models regression carpal tunnel syndrome status on age and gender.

Automated Essay Grading (Technical Report)

The William and Flora Hewlett Foundation have sponsored a Kaggle.com competition, whose goal is to find an algorithm that can accurately grade short answer questions. Features I use from these observations include: the number of words, punctuation, words per grammar, sentences, unique words, stop words, misspelled words, as well as the presence of a semicolon, popular stemmed words from highly scored essays, and the presence of popular stemmed words from lowly scored essays. We consider a survey of methods including ANN, SVM, Naïve Bayes, Logistic, and Random Forest evaluated by cross validation – focusing on the artificial neural network approach whose results fall within the guidelines of 53% to 81% for exact predictions and near the guideline of 97% to 100% for exact or adjacent predictions, provided by Mark Shermis and Jill Burstein (2003).

Mathematical and Computational Analysis of Cancer Cell Lineage Models (Technical Report)

Cancer stem cells (CSCs) have been identified in primary breast cancer tissues and cell lines. The CSC population varies widely among cancerous tissues and cell lines, and is often associated with aggressiveness of breast cancer. Despite of intensive research, how the CSC population is regulated within a tumor is still not well understood so far. In this paper, we present a mathematical model to explore the growth kinetics of CSC population. Our mathematical modeling suggests that there exist non-linear growth kinetics of CSCs and negative feedback mechanisms to control the balance between the population of CSCs and that of non-stem cancer cells. To better simulate the dynamic changes in cancer cell populations, we first propose that terminal differentiated cancer cells have negative feedback regulations on the self-renewal probability and division rate of CSCs and/or progenitor cells. This novel control mechanism capitalizes on emerging evidences in literature, and correlates nicely with recent findings in the literature on how to achieve the equilibrium.

Publications

W. Cipolli, M. Dascalu, and A. Robertson, "On the distribution of monochromatic complete subgraphs and arithmetic progressions." In Progress, 2017.

R. Bower, J. Hussey, J. Zhang, J. Quattro, M. Muhling, W. Cipolli, and J. Hardin, "The score test for independence of two marginal Poisson variables." Submitted to Computational Statistics, 2017.

R. Bower, J. Hussey, J. Zhang, J. Quattro, W. Cipolli, and J. Hardin, "Multiple score tests for a bivariate Pareto distribution." In Progress, 2017.

R. Bower, J. Hussey, J. Zhang, J. Quattro, W. Cipolli, and J. Hardin, "A copula approach for testing independence using Poisson cumulative distribution functions." In Progress, 2017.

E. Cooley, B. Payne, W. Cipolli, C. Cameron, A. Berger, and K. Gray, "The paradox of group mind: "people in a group" have more mind than "a group of people"." Accepted at the Journal of Experimental Psychology: General., 2017.

D. M. Silva, W. Cipolli, and R. Bower, "In search of a five-star: The centrality of body discourses in the scouting of high school football athletes," 2017.

W. Cipolli and T. Hanson, "Supervised learning for high dimensional data through smoothed Polya trees." Submitted to Statistical Methods in Medical Research., 2017.

K. Flory, B. A. Bell, K. Burgess, E. R. Siceloff, W. Cipolli, and R. Bower, "Bifactor models of the strengths and difficulties questionnaire in a large U.S. community sample," Presented at the annual meeting of the American Psychological Association, Denver, CO, 2016.

W. Cipolli and T. Hanson, "Computationally tractable approximate and smoothed Polya trees," *Statistics and Computing*, pp. 1–13, April 2016.

W. Cipolli, T. Hanson, and A. McLain, "Bayesian nonparametric multiple testing," *Computational Statistics and Data Analysis*, vol. 101, pp. 64–79, September 2016.

J. Fowler, W. Cipolli, and T. Hanson, "A comparison of three diagnostic tests for diagnosis of carpal tunnel syndrome using latent class analysis," *Journal of Bone & Joint Surgery*, vol. 97, pp. 1958–1961, December 2015.

Professional Consulting

The Othering of Muslims: Discourses of Radicalization in the New York Times, 1969–2014

Through the lens Edward Said's Orientalism and various perspectives within the othering paradigm the emergence and transformation of radicalization discourses in the news media is analyzed. Employing discourse analysis of 607 New York Times articles from 1969 to 2014, this article demonstrates that radicalization discourses are not new but are the result of complex sociolinguistic and historical developments that cannot be reduced to dominant contemporary understandings of the concept or to singular events or crises. The news articles were then compared to 850 government documents, speeches, and other official communications. The analysis of the data indicates that media conceptualizations of radicalization, which once denoted political and economic differences, have now shifted to overwhelmingly focus on Islam. As such, radicalization discourse now evokes the construct radicalization as symbolic marker of conflict between the West and the East. This work advances the established notion that the news media employ strategic discursive strategies that contribute to conceptual distinctions that are used to construct Muslims as an "alien other" to the West.

Judging Appropriateness of SUP Using Patient Demographic Data

This methodology explains how to record the dataset, fit a model and interpret the effects on the inappropriateness of Stress Ulcer Prophylaxis (SUP) using patient demographic data and other pertinent information. The analysis provided uses an imitated dataset - the SAS code and analysis produced are easy to read and to extend beyond the following variables: age, weight, gender, race, whether they have medical history of specialty care, prior use of concomitant antibiotics, prior use of proton pump inhibitor or H2 antagonist, severity of illness . When or if other variables are added, the statistical approach could stay the same and the added variables would be interpreted similarly.

Veteran Affairs Resource Utilization by Patients with HFpEF

Patients with heart failure with a preserved ejection fraction (HFpEF) represent a large portion of the heart failure population. The goal of the research is to identify differences in healthcare resource utilization by patients with HFpEF who are managed in the primary care setting compared to those receiving care in cardiology clinics. We document a statistical approach for a mock dataset created to closely match the description provided. The results shown are from the mock dataset and therefore aren't meaningful – they are only shown to describe the methodology that can be followed when the data becomes available to the client. The methodology explains how to fit a model and interpret the effects on the duration of stay using their age, gender, race, type of clinic, other conditions and smoking status. When other variables are added, the statistical approach could stay the same and the added variables would be interpreted similarly.

Attitude towards Homelessness Questionnaire

A statistical approach for modeling the data from surveys, which intimate sentiments towards homeless people, was provided. It is shown that the visual and textual treatments in the study lead to a slightly positive change in score, however this difference was not significant across visual and textual treatments. The statistical insignificance of controls previously shown to be relevant in past studies is another result of interest – this sample did not hold the same biases as the past experiment allowing for further exploration on the causes for such a difference.