**William Cipolli**
*1000 Watermark Pl. Apt 111*
*Columbia, SC 29210*
*+1 (203) 848 5643*
*william.cipolli@gmail.com*
*http:/people.stat.sc.edu/Cipolli*

**Research Statement** November 08, 2015

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.

## Research Projects

### Supervised Learning for High Dimensional Data through Smoothed Polya Trees via Givens Rotations (In Progress)

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.

**Computationally Tractable Approximate and Smoothed Polya Trees (In Revision)**

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 Multiple Testing (In Revision)**

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 (Accepted)**

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.

## Professional Consulting

### 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 the data provided by David Asiamah is provided. The results show that only the treatment is significant in modeling the difference in test score. Further, it is shown that the visual and textual treatments lead to a larger positive change in score however there is not a significant difference the difference in test score across visual and textual treatments. The resulting model indicates that the 'treated' groups have a score increase of about 4.7 when compared to the control group. The statistical insignificance of other 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.